



**Trinity College Dublin**

Coláiste na Tríonóide, Baile Átha Cliath

The University of Dublin

School of Psychology

# Endogenous and Exogenous Factors Shaping Visual Development in Early Infancy

Anna Kravchenko

January 2024

A dissertation submitted in fulfilment  
of the requirements for the degree of  
School of Psychology

Supervised by Prof. Rhodri Cusack

# Declaration

I hereby declare that this dissertation is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at <http://www.tcd.ie/calendar>.

A handwritten signature in blue ink, consisting of stylized, overlapping loops and a long horizontal stroke extending to the right.

Signed:

Date: 30.03.2023

# Abstract

The human mind is capable of forming elaborate concepts and models of the world and manipulating them so that it can adapt to complex, rapidly changing environments. In contrast with that, machine learning methods and artificial neural networks (ANNs) only perform well within narrow task constraints and are poor generalists. Infants and preschool children are capable of learning things that have been shown to be challenging even for cutting edge computational methods, especially when it comes to scarce or unbalanced sets, distorted data and the ability to adapt to new situations.

This difference is often attributed to the more complex neural architecture of the brain in comparison to ANNs. However, another potential factor contributing to these differences is the structure of the learning process itself. ANNs are often trained in a static way, with batches of examples randomly sampled from the whole training set throughout. In contrast, an infant's experience is dynamic and non-stationary on many timescales. Consider the learning of visual classes. Some factors are exogenous. In the real world, during any short period, some visual classes are much more common than others, while over time the distribution of different classes fluctuates, depending on the context the infant finds themselves in. Other factors are endogenous. For example, over the first few months of life, the eyes undergo rapid development and visual experience changes dramatically. Initially, visual acuity is low and input almost monochromatic, but in a few months it become sharp and richly coloured. Furthermore, the infant is not a passive viewer. Following their curiosity, they can move their eyes and apply selective attention, to allow them to shape their own curriculum. These rich trajectories during learning will change the nature of the underlying representations and may play a valuable role creating robust and generalisable concepts.

Our goal is to characterise these factors and study how they shape infant learning and change ANN performance. We investigated one exogenous factor - the statistics of categories observed by infants in their early years, and two endogenous factors - the impact of visual maturation on perceptual preferences, and the limitations of self-driven learning.

Our results suggest that ML algorithms may benefit from better training procedures and better design of their curriculum.

# Acknowledgements

I would like to express my sincere gratitude to everyone who has supported and contributed to the completion of this thesis.

I would like to thank my supervisor Rhodri Cusack for his guidance, mentorship, and constant support throughout the entire research process. His insightful feedback, constructive criticism, and encouragement have been critical in shaping this thesis.

I would also like to thank Aine Dineen, who worked tirelessly on the low-acuity vision project under my supervision. Her hard work, dedication, and commitment have made this project a success.

I am grateful to Cliona O'Doherty for generously providing the family movie dataset for the statistical learning chapter. Without her contribution, this study would not have been possible.

I am also indebted to Anna Truzzi and Lorijn Zaadnoordijk for their valuable contributions to this work, including their time, expertise, and feedback.

Finally, I would like to express my appreciation to my family and friends for their unwavering support and encouragement throughout my academic journey. Their love, patience, and understanding have been a source of motivation and inspiration.

Thank you all for your help and support.



# Contents

|          |                                                                                                                             |           |
|----------|-----------------------------------------------------------------------------------------------------------------------------|-----------|
| <b>1</b> | <b>Introduction</b>                                                                                                         | <b>1</b>  |
| 1.1      | The flexibility of the human brain . . . . .                                                                                | 1         |
| 1.2      | Neural mechanisms of human vision . . . . .                                                                                 | 3         |
| 1.3      | Visual development in infancy . . . . .                                                                                     | 5         |
| 1.3.1    | Biological development . . . . .                                                                                            | 6         |
| 1.3.2    | Learning the properties of the world . . . . .                                                                              | 7         |
| 1.4      | Learning the properties of objects . . . . .                                                                                | 9         |
| 1.5      | Cognitive modelling and parallels between human learning and AI . . . .                                                     | 10        |
| 1.6      | Artificial neural networks . . . . .                                                                                        | 12        |
| 1.7      | The 'deep learning revolution', convolutional neural networks and the possibility of human-like behaviour in ANNs . . . . . | 13        |
| 1.8      | The importance of curriculum for artificial neural networks . . . . .                                                       | 16        |
| 1.9      | Curiosity as a strategy of creating curriculum . . . . .                                                                    | 19        |
| 1.10     | The conflict between self-created and fixed curriculum. . . . .                                                             | 23        |
| 1.11     | Curriculum learning in humans . . . . .                                                                                     | 26        |
| 1.12     | Learning universals, early development and the importance of slow brain maturation . . . . .                                | 27        |
| 1.13     | Objective and methods . . . . .                                                                                             | 29        |
| <b>2</b> | <b>The effect of statistical properties of the world on category learning</b>                                               | <b>30</b> |
| 2.1      | Introduction . . . . .                                                                                                      | 30        |
| 2.1.1    | Order of semantic categories acquisition . . . . .                                                                          | 30        |

|          |                                                                         |           |
|----------|-------------------------------------------------------------------------|-----------|
| 2.1.2    | Area specificity . . . . .                                              | 30        |
| 2.1.3    | Potential factors driving area specificity . . . . .                    | 33        |
| 2.2      | Challenges and methods . . . . .                                        | 33        |
| 2.2.1    | Choosing a dataset . . . . .                                            | 33        |
| 2.2.2    | Dataset tagging . . . . .                                               | 34        |
| 2.2.3    | Metrics . . . . .                                                       | 36        |
| 2.2.4    | Statistical analysis . . . . .                                          | 37        |
| 2.3      | Experiment 1: professional movies dataset, automatic tagging . . . . .  | 38        |
| 2.3.1    | Data and methods . . . . .                                              | 38        |
| 2.3.2    | Results . . . . .                                                       | 40        |
| 2.4      | Experiment 2: infant experiences dataset, automatic tagging . . . . .   | 43        |
| 2.4.1    | Data and methods . . . . .                                              | 43        |
| 2.4.2    | Results . . . . .                                                       | 43        |
| 2.5      | Experiment 3: Professional movies dataset, manual tagging . . . . .     | 45        |
| 2.5.1    | Data and methods . . . . .                                              | 45        |
| 2.5.2    | Results . . . . .                                                       | 47        |
| 2.6      | Follow-up analysis . . . . .                                            | 48        |
| 2.6.1    | Comparison of sets . . . . .                                            | 48        |
| 2.6.2    | The impact of category level. . . . .                                   | 50        |
| 2.6.3    | Statistics of the world and category acquisition order. . . . .         | 50        |
| 2.7      | Discussion . . . . .                                                    | 53        |
| <b>3</b> | <b>Benefits of low initial acuity</b>                                   | <b>56</b> |
| 3.1      | Introduction . . . . .                                                  | 56        |
| 3.2      | Methods . . . . .                                                       | 61        |
| 3.2.1    | Model . . . . .                                                         | 61        |
| 3.2.2    | Training set . . . . .                                                  | 61        |
| 3.2.3    | Experiment 1: Modelling poor visual acuity with Gaussian blur . . . . . | 62        |
| 3.2.4    | Experiment 1: Training trajectories . . . . .                           | 62        |
| 3.2.5    | Experiment 1: Test set . . . . .                                        | 63        |

|          |                                                                       |            |
|----------|-----------------------------------------------------------------------|------------|
| 3.2.6    | Experiment 1: Quantifying shape bias . . . . .                        | 63         |
| 3.2.7    | Experiment 2: Impact of category type on shape bias . . . . .         | 64         |
| 3.2.8    | Experiment 3: Effect of shape bias on category separability . . . . . | 65         |
| 3.2.9    | Experiment 4: Robustness against distortions . . . . .                | 67         |
| 3.3      | Results . . . . .                                                     | 69         |
| 3.3.1    | Experiment 1: Effect of initial low acuity on shape bias . . . . .    | 69         |
| 3.3.2    | Experiment 2: Effect of category type on shape bias . . . . .         | 71         |
| 3.3.3    | Experiment 3: Effect of shape bias on category separability . . . . . | 73         |
| 3.3.4    | Experiment 4: Robustness to noise . . . . .                           | 74         |
| 3.4      | Discussion . . . . .                                                  | 77         |
| <b>4</b> | <b>Curriculum learning</b>                                            | <b>79</b>  |
| 4.1      | Introduction . . . . .                                                | 79         |
| 4.2      | Methods . . . . .                                                     | 82         |
| 4.2.1    | Data and tasks . . . . .                                              | 82         |
| 4.2.2    | Model . . . . .                                                       | 84         |
| 4.2.3    | Training procedure . . . . .                                          | 86         |
| 4.2.4    | Metrics to predict an optimised switching moment . . . . .            | 88         |
| 4.2.5    | Examining task representations . . . . .                              | 89         |
| 4.3      | Results . . . . .                                                     | 90         |
| 4.3.1    | Testing a range of curricula . . . . .                                | 90         |
| 4.3.2    | Metrics to predict the optimised switching moment . . . . .           | 93         |
| 4.3.3    | Using digit addition for Task 2 . . . . .                             | 101        |
| 4.3.4    | Examining task representations . . . . .                              | 102        |
| 4.4      | Discussion . . . . .                                                  | 104        |
| <b>5</b> | <b>General Discussion</b>                                             | <b>106</b> |
| 5.1      | Summary . . . . .                                                     | 106        |
| 5.2      | Benefits of curriculum . . . . .                                      | 110        |
| 5.3      | Directions of future research . . . . .                               | 112        |

|           |                                                                          |            |
|-----------|--------------------------------------------------------------------------|------------|
| 5.3.1     | Expanding existing experiments . . . . .                                 | 112        |
| 5.3.2     | Further goals . . . . .                                                  | 114        |
| 5.3.3     | Broad implications . . . . .                                             | 115        |
| 5.3.4     | Conclusion . . . . .                                                     | 117        |
| <b>A1</b> | <b>Statistical properties of the world influencing category learning</b> | <b>142</b> |
| A1.1      | Movies used in the experiment . . . . .                                  | 142        |
| A1.2      | Examples of categories with low/high tf-idf . . . . .                    | 145        |
| A1.3      | Ethical Approval . . . . .                                               | 147        |
| <b>A2</b> | <b>Low initial acuity</b>                                                | <b>149</b> |
| A2.1      | Finding optimal kernel size . . . . .                                    | 149        |
| A2.2      | Performance of networks with only low-acuity training . . . . .          | 149        |
| <b>A3</b> | <b>Curriculum learning</b>                                               | <b>151</b> |
| A3.1      | Effects of the number of of epochs. . . . .                              | 151        |
| A3.2      | Catastrophic forgetting. . . . .                                         | 152        |

# 1 Introduction

## 1.1 The flexibility of the human brain

The human brain is remarkable in its ability to adapt to the ever-changing environment. However, despite this incredible plasticity, there are considerable similarities in brain structure and function across different individuals [Glasser et al., 2016]. This raises the question of how many of these similarities are determined by biological factors versus environmental factors.

The nativist and empiricist schools of thought are two competing perspectives on the foundations of cognition. These perspectives differ in their views on the innate structures and mechanisms that underlie human cognition, as well as the role of experience and learning in shaping cognitive abilities. The ongoing debate between these two schools has shaped the development of cognitive models and AI systems in many ways.

The nativist school posits that certain cognitive abilities and structures are innate, meaning that they are present from birth or develop independently of specific experiences. Nativists argue that the human brain is equipped with specialized mechanisms and structures that enable the acquisition and processing of knowledge. This view is exemplified by Noam Chomsky's theory of universal grammar [Chomsky, 1959], which posits that humans possess an innate capacity for language acquisition. In the context of AI, nativist-inspired approaches often emphasize the importance of incorporating pre-specified knowledge or cognitive architectures into AI systems to simulate human-like intelligence.

On the other hand, empiricist theories posit that cognition arises from experience and sensory input, with knowledge being acquired incrementally through a process of learning and adaptation. Empiricists argue that the mind is as close as possible to blank slate or "tabula rasa" at birth [Elman, 1996], with cognitive development being driven by exposure to various stimuli and the formation of associations between them. In the realm of cognitive modeling and AI, empiricist perspectives usually rely on connectionist models. These models draw inspiration from the neural architecture of the human brain, seeking to emulate its capacity for learning and adaptation by focusing on the statistical structure of the world.

While the nativist and empiricist schools of thought have historically been portrayed as mutually exclusive (which is not entirely true, a broad range of stances exists in the middle ground [Spelke, 2000, Johnson, 2011]), recent advances in cognitive science and AI research have demonstrated the potential of more nuanced, integrated perspectives.

In this thesis we aim to investigate one aspect that can provide patterns of brain and behavioral organization that are consistent across individuals, while not being, in the strict sense of the term, innate. Specifically, we will explore the role of curriculum and how learning experiences can be shaped by the inherent properties of the natural world and their interplay with biological maturation.

Using computational models for such exploration paves the way to understanding the universals of learning that apply to both artificial intelligence and human learning. Discovery of such universals would not only accelerate advancements in the field of AI, enabling the creation of machines that can learn more naturally and efficiently but it would also deepen our understanding of the foundational elements that drive early human cognitive development, allowing for more insightful theories and studies.

One of the best-studied aspects of brain development is visual learning. Vision is the primary sense that humans use to navigate and interact with the world, and as such, it has been extensively researched and modelled. From the early stages of visual perception

development the brain undergoes remarkable changes in response to visual experience. Understanding the mechanisms underlying visual learning can provide insights into how the brain adapts to different types of learning and how we can enhance our learning abilities. Therefore, this thesis will focus on studying the neural basis of visual learning, with the aim of shedding light on the broader question of brain plasticity and the relative contribution of biological versus environmental factors.

## 1.2 Neural mechanisms of human vision

To understand the role of exogenous factors in visual system development we must first look into how it is organized in the brain. Over the course of evolution humans have developed a sophisticated visual system, highly adapted to their particular needs and environment.

Initial visual processing begins as early as in the retina—light-sensitive nerve tissue at the back of the eye that is responsible for converting light into electrical signals in the brain. It is composed of 10 distinct layers of tissue, each with different functions. Due to its organization the retina is capable of extracting visual features, such as colour, contrast, and motion, while mapping the outside world [Masland, 2012]. It is also involved in a number of other important functions, including the control of the pupil in response to changes in light levels, as well as local computation of visual stimuli [Vlasiuk and Asari, 2021].

When light hits the retina, it stimulates the layer of photoreceptor cells, which are divided into two types: rods and cones. Rods are responsible for detecting light and dark, and are most sensitive in low light conditions. Cones, on the other hand, are responsible for detecting color and are most sensitive in bright light. The retina contains about 120 million rods and 6 million cones. These cells then create an electric signal that is conveyed through interneurons of the retina first to bipolar and then to the retinal ganglion cells (RGCs). The RGCs are located on the inner surface of the retina, where they project visual information to the brain via their axons, which make up the optic

nerve [Curcio and Allen, 1990].

The optic nerve is primarily connected to the lateral geniculate nucleus (LGN)—a structure located in the thalamus, a part of the brain that plays a key role in integrating sensory information [Brodal, 2010]. In humans six distinctive layers are present in the LGN, each of which receives input from a specific type of ganglion cell in the retina, alternating inputs from each of the two eyes. The inner two layers of LGN, (1 and 2) are comprised of magnocellular cells and collect information required for the perception of motion, depth, and differences in brightness. The outer four layers, (3,4,5 and 6), are comprised of parvocellular cells which are responsible for the perception of color and form. Magnocellular cells are more sensitive to low spatial frequencies and are crucial for registering changes in luminance, detecting edges and performing visual search tasks. Parvocellular cells have greater spatial resolution, but lower temporal resolution, than the magnocellular cells, allowing perception of fine details.

The axons that leave the LGN go to the primary visual area of the cortex (area V1), the part of the brain that is responsible for higher-level visual processing [Brodal, 2010].

There is strong evidence that this visual pathway is segregated into two distinct streams—ventral and dorsal. It is theorized that the pathways are segregated in function in the following way:

- The dorsal stream, also known as the "where" pathway, is responsible for processing spatial information, such as the location and movement of objects in the visual field. It relies on information from the magnocellular layers of LGN. The processing performed by the dorsal stream starts with purely visual functions in the V1 and the projects to the posterior parietal cortex, the area that plays an significant role in movement planning, spatial reasoning, and attention.
- The dorsal stream is interconnected with the parallel ventral stream, also known as the "what" pathway, that is responsible for processing more complex visual information, such as the color, shape, and identity of objects. This stream gets its main input from the parvocellular layer of the LGN. From primary visual cortex, it then continues



through V2 and V4 to areas of the inferior temporal lobe and then sends the information to the inferotemporal cortex (IT), the primary region of the brain involved in object recognition which is located in the temporal lobe of the brain.. Each visual area contains a full representation of the visual field, providing increasingly more complex neuronal tuning.

Neuronal tuning is the ability of a neuron to respond selectively to specific features or stimuli in the visual field. For example, neurons in V1, have been found to have specific tuning to changes in contrast necessary for edge detection, while V2 is tuned to such visual characteristics as orientation, spatial frequency, size, and color, and V4 to geometrical shapes. A neuron further along the ventral stream, for example in the IT, may fire only when a specific face appears in its receptive field.

Having described the two-stream hypothesis in detail, we, however, find it important to mention that contemporary approaches based on recent experimental work offer a more complex approach than a simple separation of function into two-streams. In reality is a considerable interaction between object location and object features processing is more likely. In their 2010 review McIntosh and Schenk [Schenk and McIntosh, 2010] suggest approaching this model not as a formal hypothesis, but rather as a set of heuristics to guide experiment and theory.

For the goals of this thesis we will focus on the ventral stream, due to the successes of replicating it in artificial models [Yamins et al., 2014]. However, even the mechanisms governing the development of this single area are incredibly complex, this complexity stemming from an interplay of biological factors and the process of infants adjusting to their environment after birth.

### **1.3 Visual development in infancy**

Human visual development is a complex and dynamic process that begins at birth and continues throughout the first few years of life, allowing infants to see and perceive the world around them with increasing clarity. It is reliant both on the biological factors such

as retina and brain areas maturation and the necessity to learn about the nature of the environment.

### 1.3.1 Biological development

One of the earliest and most important developments is the ability to perceive simple perceptual details of stimuli, such as acuity and color.

Visual acuity is defined as a person's ability to recognize fine details with precision. At birth, infants have relatively poor acuity compared to adults, with an estimated acuity of 20/400 [Leat et al., 2009]. This means that they are only able to see objects clearly if they are located about 20 inches or closer to their face. This would be considered legally blind in most countries.

This change can be attributed to two factors. Anatomical studies [Yuodelis and Hendrickson, 1986, Banks and Bennett, 1988] have shown that at birth the newborn's cone photoreceptors of central retina reach over  $7.5 \mu$  in diameter, only shrinking to the normal adult width of  $2 \mu$  in about 3 years. As retina matures and the photoreceptors reach their adult size, the cone density in this region increases from 18 photoreceptors per  $100 \mu$  to 42 photoreceptors per  $100 \mu$ . This low number of photoreceptors per retinal area contributes to the initial poor acuity.

A second important factor is that at that stage cortical neurons have relatively sparse connections. The first six months of development witnesses a growth spurt of connections, i.e. while neural density stays roughly the same, connections between axons and dendrites become much richer [Atkinson, 1984].

For the same reasons color perception in infants is initially limited as well and tends to be more sensitive to contrasts between colours rather than the colors themselves [Skelton et al., 2022].

As infants grow and develop, their acuity improves significantly. By the age of 3 to 4 months, infants typically have an acuity of 20/40, which is similar to that of an adult.

They also become better able to distinguish between different colors and shades.

To evaluate acuity and colour perception in infants scientists use visual preference looking tasks and visual evoked potentials (VEPs). Visual preference looking tasks involve presenting infants with stimuli of different sizes or levels of detail and tracking their gaze patterns to determine which stimuli they prefer to look at. VEPs involve measuring the brain's electrical activity in response to visual stimuli, which can be used to assess the brain's response to different levels of detail in the stimuli.

### **1.3.2 Learning the properties of the world**

The ability to perceive objects is also reliant on learning and adjusting to the environment and the nature of the objects present in that environment.

This is achieved through neural coding [Simoncelli and Olshausen, 2001, Simoncelli, 2003], the adaptation of neurons to the statistical properties of the signals to which they are exposed and establishing a relationship between the stimulus and the individual or ensemble neuronal responses. This happens through both evolutionary and developmental processes, coding efficiency improving with age.

As an example, while newborns can detect edges and motion, and exhibit some capacity for depth perception [Slater, 1995, Slater et al., 2010], they do not yet perceive objects the same as adults. They have not yet learned that the world consists of separate objects at different distance which may overlap and have hidden parts.

This is resolved by developing object selectivity, or the ability to selectively attend to and process information about specific objects. It is an important aspect of visual development and it allows infants to focus on and learn about the objects and people in their environment. Object selectivity begins to emerge in infants at around 3 to 4 months of age and continues to develop throughout the first year of life.

One of the key features of object selectivity in infancy is the development of object permanence, or the understanding that objects continue to exist even when they are not

in sight. This understanding begins to emerge at around 4 to 5 months of age.

Another important aspect of object selectivity is the ability to recognize faces. This ability, known as face selectivity, begins to develop early in life and continues to improve throughout the first few years. Studies have shown that infants as young as two months old can distinguish between different faces and prefer to look at familiar faces over unfamiliar ones [Pascalis and Slater, 2003], they also have a strong preference for looking at human faces over other objects, such as inanimate objects or scrambled faces [Johnson et al., 1991]. Some existing data suggests that a predisposition for faces is even observed in newborns (see [Simion and Giorgio, 2015] for a detailed review), others even demonstrate that face selectivity is not disrupted in congenitally blind patients [Murty et al., 2020, van den Hurk et al., 2017]. Whether face selectivity is innate or at least partially acquired through experience remains a topic of debate and further research is needed to fully understand the interplay between innate predispositions and experiential factors.

Another important aspect of developing vision is categorization and the ability to recognize specific types of objects, such as animals. Existing studies demonstrated that this ability undergoes rapid development in the first year of life [Mandler and McDonough, 1993], with infants being able to categorize objects as natural or artificial based on their perceptual features as early as at 3-4 months [Behl-Chadha, 1996] and show elements of perceptual-based categorization at 2 months old [Quinn et al., 2000].

To measure object selectivity in infants several methods are used, namely visual preference looking tasks, habituation and dishabituation tasks, and eye tracking. Visual preference looking tasks involve presenting infants with stimuli of different objects and tracking their gaze patterns to determine which stimuli they prefer to look at. Habituation and dishabituation tasks involve presenting infants with a series of stimuli and measuring their attention and responsiveness to the stimuli over time. Eye tracking involves measuring the movements of the eyes to determine where infants are looking and how long they are looking at specific objects.

## 1.4 Learning the properties of objects

An important part of learning from experience is gathering statistics of the world and taking note of patterns of co-occurrences. This is one of the foundational principles of both human cognition and artificial intelligence. A field of cognitive science called statistical learning (SL) explores such processes. Studying SL allows us the opportunity to better understand the possible strategies of learning, as well as the impact of biological constraints imposed on learning by brain organisation and maturation. Thus, providing an avenue for potential insights into the commonalities between machine and human learning paradigms.

Initially statistical learning emerged as a tool for studying language development and processing and its roots can be traced back to [Esper, 1925] (see [Christiansen, 2018] for a full review and [Perruchet and Pacton, 2006] for comparison of conceptual frameworks) where it was called implicit statistical learning. However, the modern paradigm of “statistical learning” was introduced by Saffran, Aslin, and Newport [Saffran et al., 1996], who, using an artificial grammar showed that 8-months infants are sensitive to transitional probabilities (TPs) in speech, using them to discover word boundaries.

Since then these findings have been reproduced with many variations, including switching to a visual paradigm, where regularities were presented as spatial locations rather than temporal sequences [Fiser and Aslin, 2001], replacing transitional probabilities with probabilities of co-occurrence. [Slemmer et al., 2010] habituated 2-, 5-, and 8-month-old infants to a continuous stream of colored shapes, discovering that at each age, infants demonstrated a significant novelty preference for the random sequence.

In many studies learning of visual statistics was explored (e.g., [Kirkham et al., 2002, Siegelman and Frost, 2015, Turk-Browne et al., 2005]). However, it is worth noting that detecting transitional probabilities is more complex than detecting co-occurrence frequencies because transitional probabilities entail predicting relations among items. Instead of focusing on adjacent regularities like AB, researchers have examined sequences

in the form of  $AxB$ , where  $x$  represents a randomly chosen stimulus [Gómez, 2002, Newport and Aslin, 2004, Onnis et al., 2015]. Participants have been exposed to two sets of regularities rather than just one, either within [Gebhart et al., 2009, Karuza et al., 2017] or between e.g., [Emberson et al., 2011, Mitchel and Weiss, 2011, Weiss et al., 2015] modalities.

As switching to another modalities illustrates, a major appeal of SL lies in its ability to offer a foundational framework for understanding the processes of generalization, discrimination, and category formation. In essence, SL has the potential to explain a wide range of cognitive processes.

And while initially the approach grew out of attempts to model infant word segmentation from continuous speech, SL has been shown to be a robust model for various foundational non-linguistic mechanisms, including vision, audition, motor planning, event processing, reading, speech perception, language acquisition, semantic memory, and social cognition. As a majority of cognitive functions inherently entail the learning of regularities, SL appears to be fundamental for understanding most domains of cognition, including vision [Simoncelli, 2003, Ruderman, 1994, Torralba and Oliva, 2003].

## **1.5 Cognitive modelling and parallels between human learning and AI**

The paradigm of statistical learning opens a possibility of comparing human learning on artificial intelligence, as many AI models also rely on learning statistical regularities.

Computational/AI models have been widely used to investigate various cognitive processes, including perception, learning, problem solving, and decision making. They have proved themselves particularly useful for capturing early developmental phenomena such as, for example, category learning and the already mentioned statistical learning (e.g., [Althaus and Mareschal, 2014, Colung and Smith, 2003, Gliozzi et al., 2009, Mareschal et al., 2000, Mareschal and Thomas, 2007, Munakata and McClelland, 2003, Rogers and McClel-

land, 2008, Westermann and Mareschal, 2004, Westermann and Mareschal, 2012, Westermann and Mareschal, 2014]).

Many AI models are also specifically designed to resemble different aspects of human learning, such as imitating neurons in human brains in artificial neural networks or arrange their personalized learning material through reinforcement learning.

Mostly they can be divided into two large group. One approach is symbolic modeling, which involves representing knowledge in the form of symbols and rules, and using algorithms to manipulate these symbols to solve problems. Another approach is connectionist modeling, which involves simulating the processing of information by networks of interconnected neurons or artificial neural networks.

In this work we will focus on connectionist models of visual learning, as they are based on the insights from architecture of human ventral stream and have been shown to perform at human-like levels and exhibit many human-like behaviours on tests of object recognition. Training these models with slight variations in hyperparameters or learning input may provide important insights to the trajectories of infant vision development. Please see section 1.6 for a detailed discussion of this topic.

Modelling infants can also provide insights into designing better AI. Machine learning greatly succeeded mimicking adult learners with their access to a vast amount of prior knowledge, however, even state-of-the-art AI struggle with some of the things that come easily for children, and addressing infants learning strategies could be a route towards a new generation of learning algorithms [Zaadnoordijk et al., 2022].

However, a major challenge in cognitive modeling is to develop models that are both realistic and tractable, meaning that they accurately capture the complexity of human cognition while still being computationally feasible. Another challenge is to validate the models, or to demonstrate that they accurately capture the behaviors and processes that they are intended to represent. Solutions to these challenges are model-specific and require introducing our model first, therefore this topic will be addressed in more detail in chapter 1.7.

## 1.6 Artificial neural networks

While cognitive models are not restricted to artificial neural networks (ANNs), for the purposes of this thesis we will focus our attention on ANNs, as, being inspired by brain organization, they are currently the closest we have to humans. In addition, current state-of-the-art models have also been shown to represent objects using the same visual primitives as derived by the human visual cortex ([Yamins et al., 2014], please see section 1.6 for more detail).

ANNs are an artificial simplification of neurons in the brain, created specifically to replicate human connectome and capable of dealing with fuzzy and unlabeled data. They are made up of a large number of interconnected processing nodes, or artificial neurons. These neurons are organized into layers, with the input layer receiving the raw data and the output layer producing the final output of the network. There may be one or more hidden layers in between the input and output layers, depending on the complexity of the network.

Each neuron in the network receives input from other neurons, processes that input using a set of weights and a transfer function, and then passes the processed data on to other neurons in the next layer. The weights determine the strength of the connections between neurons and can be adjusted during the training process in order to optimize the network's performance. The transfer function is a mathematical operation that determines how the input is transformed as it passes through the neuron. Some common transfer functions include the sigmoid function and the rectified linear unit (ReLU) function.

In order to analyze and classify data, the input is passed through the input layer and then flows through the hidden layers, if present, until it reaches the output layer. As the data flows through the network, the neurons in each layer process and transform the data, passing it on to the next layer until the final output is produced.

Each layer of the neural network can be seen as a transformation mapping the data set from the original/previous layer feature space into a learned feature space. Same as in



the human brain, objects having certain features tend to consistently activate the same clusters of neurons. Such a cluster of neurons activated in a layer in response to a feature is called a feature representation in that layer.

Through the process of training, the weights of the connections between the neurons are adjusted in order to optimize the network's ability to interpret and classify the data. This is typically done using an algorithm called backpropagation, which involves adjusting the weights based on the error between the predicted output of the network and the actual output.

Depending on the architecture, the hyperparameters and the dataset the representations created by the network can vary drastically, and this leads to different ways of defining categories and clustering the dataset, some of which may be more efficient than the others, as well as more human-like.

## **1.7 The 'deep learning revolution', convolutional neural networks and the possibility of human-like behaviour in ANNs**

The field of artificial neural networks experienced a renaissance in the 2010s due to a number of factors, including advances in computational power and the availability of large datasets.

One of the key factors that contributed to the resurgence of ANNs was the increased computational power and efficiency of modern hardware, particularly graphics processing units (GPUs). This allowed for a better implementation of the already mentioned backpropagation algorithm (introduced in its modern form by [Rumelhart et al., 1986]), which was previously less viable due to computational constraints, in turn allowing to effectively train 'deep' neural networks stacking multiple hidden layers.

At the time they were mostly used for speech recognition but gradually the usage ex-

panded to other domains. In 2012, Ciresan and colleagues [Cireşan et al., 2012] showed using a deep neural network on GPU can significantly improve many vision benchmark records. In October 2012, a similar system by Krizhevsky et al. called AlexNet won the large-scale ImageNet visual recognition challenge [Krizhevsky et al., 2012].

Prior to AlexNet, the state-of-the-art in object recognition was achieved using shallow, hand-designed features. However, AlexNet demonstrated that a deep neural network, trained end-to-end on the ILSVRC dataset, could outperform these traditional approaches by a significant margin.

One of the key innovations in AlexNet was the use of a convolutional neural network model (CNN), that is particularly well suited for image recognition tasks, as it was initially developed as a simplified model of visual cortex [Fukushima, 1980, Hubel and Wiesel, 1962].

Since the introduction of AlexNet, there have been many other advances in the field of visual object recognition. These include the development of deeper and more powerful CNN architectures, such as VGG, ResNet, and Inception, as well as the use of attention mechanisms and transformer-based models to improve the ability of models to process and classify complex scenes.

In this work we will be focusing on CNNs specifically, as their behaviours are well-documented, and recent advances have shown that they indeed can be seen as a suitable model for human visual cortex: the more effective a CNN model is the more likely it is using human-like representations [Yamins et al., 2014].

"They also exhibit some human-like traits, such as having critical periods of learning. Many animals have fixed and crucial time periods during their early development when they are able to learn things they won't be able to learn later if such a period is missed. [Thomas and Johnson, 2006] demonstrated that critical periods are not restricted to biological systems, and can appear in ANNs due to a variety of causes. One such cause is self-terminated plasticity; where the critical period is experience-driven, created by a system driving itself into a representational state where it is no longer responsive due

to the lack of representational overlap between new and old experiences [Achille et al., 2019] demonstrated a milder case of such reduction of plasticity in CNNs, where learning did not necessarily stop altogether but representations were altered due to fundamental constraints arising from learning dynamics."

The existence of self-termination of plasticity is a discovery with a lot of implications for cognitive science, as previously critical periods were assumed to be biologically-driven. One influential model of critical periods in development is the "biological clock" model proposed by Takao Hensch [Hensch, 2004]. According to this model, the brain has a "biological clock" that regulates the timing of critical periods, opening and closing them at specific times during development. This clock is thought to be controlled by a network of genes that are activated at specific times during development, leading to changes in the brain's plasticity and sensitivity to stimuli.

Achille's result suggests that critical might have informational rather than biological nature. This can serve a good example of how the study of ANNs can potentially provided insight into human development,

However, there are also some notable differences between human learning and AI. One key difference is the speed at which they can learn and adapt. While humans can take a long time to learn and adapt to new situations, AI models can often learn and adapt much faster, thanks to their ability to process large amounts of data quickly. Another difference is the ability to generalize. Humans are generally good at applying what we have learned in one situation to similar situations, while AI models may struggle with this unless specifically trained to do so.

On some accounts CNNs have also been shown to behave very differently from humans, the 'bag of features' problem being the most prominent [Sinz et al., 2019]: CNNs tend to overfocus on small local features unlike humans who rely more on shapes to categorize objects [Malhotra et al., 2022]. However, as [Geirhos et al., 2019] demonstrates, to an extent this can be compensated with a training curriculum. Chapter 3 of this thesis will be dedicated to this particular topic.

CNNs are also susceptible to adversarial attacks [Malhotra et al., 2022]: small edits to the stimuli that are imperceptible to humans but completely change the classification outcome of a CNN.

"A good comprehensive overview of the suitability of CNNs as human learning models can be found in [Wichmann and Geirhos, 2023]. According to the authors, they are to be regarded as promising—but not yet adequate—models of human object recognition abilities, with many similarities with human vision being superficial.

However, when it comes to the topic of discovering statistical regularities and learning universals, as long as the differences and limitations are taken into account, CNNs are reasonably suitable for investigating the development of the visual system in infants. The feasibility of the model can be addressed by introducing a metric for comparing behaviour of the two learning systems."

The easiest way would be to compare output performance, checking whether the model and human experiment participants put the same objects into the same categories, what kind of errors they make and how often do they make them, and how quickly do they learn.

Another way is to analyze the internal representations learned by machine learning models and, via neuroimaging, to compare them to the internal representations used by humans to perform the same task. By comparing the structure and patterns in these representations, researchers can gain insights into how the models and humans are solving the task and how their approaches differ.

## **1.8 The importance of curriculum for artificial neural networks**

It is important to mention that multiple experiments, starting from Elman's classic work [Elman, 1993], have demonstrated that architecture is not everything that matters in an ANN and that, just as humans, ANNs can also benefit from more elaborate

approaches to training.

The simplest examples would be pre-training and transfer learning that allow to remove overfitting and create better feature representations.

Pre-training refers to the process of training a machine learning model on a large dataset before using it to perform a specific task. The idea behind pre-training is to initialize the model with good general-purpose representations of the data, which can then be fine-tuned for the specific task at hand.

Pre-training has become increasingly popular in recent years, particularly in the field of natural language processing (NLP). In NLP, large language models such as BERT and GPT-3 are first pre-trained on vast amounts of text data and are then fine-tuned for specific domains or tasks. These pre-trained models have achieved state-of-the-art results on many NLP benchmarks.

Pre-training has several benefits. One benefit is that it can help to reduce the amount of labeled data needed to train a model for a specific task. It can also improve the generalization performance of the model, meaning that it is able to perform well on a wider range of tasks.

Transfer learning refers to the process of applying knowledge and skills learned in one context to a new, related context. In the context of machine learning, transfer learning refers to the use of a pre-trained model as the starting point for a new model, rather than training a model from scratch.

Transfer learning can be useful when there is a shortage of labeled data for a specific task, or when the data distribution for the new task is significantly different from the data used to train the pre-trained model. By using a pre-trained model as a starting point, the model can take advantage of the knowledge and features learned from the previous task, and use this knowledge to perform better on the new task. Transfer learning has been successful in a number of applications, including natural language processing, computer vision, and speech recognition.

Even more complex curriculum approaches have been developed to tackle a variety of problems. Starting with more representative examples, and gradually increasing in diversity, improved convergence speed and helped the network to focus on relevant features while it was forming representations [Bengio et al., 2009]. Using auxiliary tasks has proven to be effective in improving generalization across tasks and robustness against various types of noise [Zeiler and Fergus, 2014, Mordan et al., 2018, Zhang et al., 2014]. Recent experiments with ANN training datasets have also demonstrated that some ANN limitations that are considered an inherent property of ANN architectures, such as, for example, a bias towards local processing, can be reduced by adjusting the dataset [Geirhos et al., 2019] A more detailed literature review of this topic can be found in Chapter 1.

Multiple-task curricula also help tackle catastrophic forgetting [Bell and Lawrence, 2021]: the tendency of an artificial neural network to completely forget previously learned information upon learning new information. This is especially important for naturalistic learning scenarios, where differences between human and AI learning capabilities have been shown to be the most prominent.

Also, as has already been mentioned, deep neural networks also exhibit critical periods during which a temporary stimulus deficit can significantly impair their learning [Achille et al., 2019]. Similar to humans, it might be important to teach them some things early in the training process.

The main problem with using a curriculum, however, is that labelled data is scarce, creating curriculum is time-consuming and guessing an optimal one is not guaranteed. In the case of cognitive modelling, Another human-like approach comes to the rescue: active learning. The ability of a network to pick its curriculum for itself, based on the data available in a single run.

It has shown good results both in the form of reinforcement learning and other approaches, such as attention models and developmental robotics - a subfield of robotics that focuses on the study of how robots can learn and develop intelligent behavior through

experience, rather than being pre-programmed with specific tasks or skills.

However, although there are many proof-of-principle papers and good results in controlled experimental settings [Graves et al., 2017, Oudeyer and Kaplan, 2009], it is not actually used in practice in any major applications. It seems there is still much to learn and implement and the obvious direction would be to mimic learning patterns typical humans.

In human development literature such self-driven strategies would be known as curiosity-driven learning. However, even for humans it brings up a variety of other problems and questions.

## **1.9 Curiosity as a strategy of creating curriculum**

Living organisms are driven by a wide variety of motivations, pushing them, for example, to maintain a certain energy level. Some animals possess higher order motivations that push them towards exploration, engagement in new behaviours and playful activities and desire to learn about things that may not be related to the current goal. This trait is especially prominent in humans and there exists a considerable body of literature demonstrating why it is essential and investigating the processes behind it [White, 1959, Berlyne, 1954, Berlyne, 1960, Deci and Ryan, 1985, Ryan and Deci, 2000, Csikszentmihalyi, 1991].

This type of motivation has received the name "intrinsic motivation" [Ryan and Deci, 2000]. What exactly makes some activities more fun than others is still a subject of debate, as well as the reasons why humans or other living beings evolved to have this trait.

In 'commonsense psychology' and some neurocognitive literature this trait is often called curiosity, however these two terms, strictly speaking, are not mutually interchangeable. Such things as aesthetic perception also fall under the category of intrinsic motivation, while not being considered a type of curiosity. Some researchers [Dubey and Griffiths,

2020, Kidd and Hayden, 2015, Murayama et al., 2019] have also advocated for viewing curiosity as any type of information-seeking while being agnostic to whether it is driven by intrinsic or extrinsic motivations. For the goals of this thesis we will use the term curiosity as referring to intrinsic motivation in information-seeking behavior.

Attempts to conceptualise curiosity led to a theory that the most "interesting" situations were considered those of intermediate level of familiarity/complexity [Berlyne, 1954, Kidd et al., 2012], where complexity is defined in terms of the learner's current set of mental representations. These theories predict that learners will preferentially select stimuli of an intermediate level of complexity - those neither overly simple (already encoded into their representational space) nor overly complex (too different from existing representations).

A variety of models of this behaviour exists, usually applied in developmental robotics - a subfield of robotics inspired by human and animal learning that focuses on the study of how robots can learn and develop intelligent behavior through experience, rather than being pre-programmed with specific tasks or skills.

The exact mechanisms or whether curiosity is a single mechanism or a series of heuristics are still subjects of debate. For example, such metrics as learning progress and minimising surprise have been proposed. A good overview of existing metrics can be found in [Oudeyer et al., 2016].

Broadly summarizing, in developmental robotics the majority of robotic "curiosity" systems consist of two components. The first one learns to predict the sensorimotor consequences when a given action is executed in a given context. The second one, called a "meta-learning machine", learns to predict errors that first component will make in its predictions. The output of the meta-learning machine is then used to decide the potential "interestingness" of a situation. Existing approaches to making this decision can be divided into two broad groups, based on the metrics they use to conceptualize interestingness.

1. Surprise based/novelty based



Early works demonstrated good results implementing novelty-seeking algorithms [Thrun, 1995, Huang and Sebastian, 2014, Barto et al., 2004]. In computational terms it can be modelled by maximizing prediction error - events and objects that violate the expectations the most are considered the most "interesting". The action that a robot chooses at each step is the one for which its meta-learning machine predicts the largest error in prediction of the action's consequences. This is in line with dopamine "prediction error" studies, novelty-seeking behavior related to trait-curiosity in adults [Oudeyer et al., 2016] and Gopnik's theory of curiosity in infants [Gopnik et al., 2017].

However, these results were achieved under narrow deterministic conditions and these approaches also do not take into account expected gain. It means that a robot using a surprise-oriented method will be stuck on white noise or situations for which the causal variables can not be observed, or, more generally, by situations which are too complex for its learning algorithms, thus continuously rendering its predictions useless [Oudeyer and Kaplan, 2009, Burda et al., 2018].

These approaches are also unlikely to be plausible models of infant learning, since children exist in a rich heterogeneous non-deterministic environment. Surprise maximization would risk pushing the infant towards white noise or complex, unlearnable things, which would hinder their development.

To account for this problem in Itti and Baldi introduced a metrics called "Bayesian Surprise" [Itti and Baldi, 2009] and, using a computational model, demonstrated how it could predict patterns of human visual attention better than previous approaches. This notion of surprise takes previously learned data into account and is different from Shannon/Boltzmann's entropy-based notion of surprise, protecting the model from unlearnable high entropy data. Only data with learnable statistical or algorithmic regularities is considered interesting by this model. In some ways it is closer to the progress-based methods in the next section.

## 2. Progress maximization

Other researchers have treated curiosity as a strategy of maximizing learning in a given

context. In other words, the intrinsic reward is based on how much the environment model changed after the new experience. These models are better at dealing with unlearnable situations, as they use prediction error less directly. The 'value' of potential scenarios is assessed using the discrepancy between the anticipated mean error rate of predictions in the near future. There exists a variety of these, ranging from information compression [Schmidhuber, 2009] to progress in prediction [Oudeyer et al., 2016] or minimizing discrepancies between stimuli and internal representations and adding the notion of learning plasticity [Twomey and Westermann, 2017a]. Taking into account previous learning and expected gain, these models are more effective in non-deterministic environments. These approaches usually align with the Goldilocks principle, picking events of intermediate novelty or complexity.

However, the basic implementations have only been tested in spaces in which only one kind of activity is possible, for example, categorizing images or learning to grip things using visual cues. When encountering various kinds of activities and the need to switch between various tasks [Oudeyer and Kaplan, 2009], measures of the evolution of performance can become incomprehensible. Adding a module that categorizes tasks or situations into types and/or a measure of the similarity of situations fixes the problem for robots [Oudeyer and Kaplan, 2009], however, this has not been compared to child behavior yet.

### 3. Other approaches

There also exist approaches that go beyond pure intrinsic motivation or do not entirely fit behavioral data, but should be taken into consideration as well.

#### - maximizing competence

This approach was proposed by [Oudeyer and Kaplan, 2009] and is based on the measures of competence that an agent has for achieving self-determined goals. While not entirely intrinsic, it goes in line with the psychological theories described in section 2, especially the need for autonomy and competence emphasised by Deci and Ryan [Deci and Ryan, 1985]. The concept of challenge and the measure of difficulty are central

here, and the “interestingness of a possible” chain of action is evaluated. Autonomous Mental Development (AMD) can be viewed as an example of such model [Weng et al., 2001].

- morphological complexity (correlations of the object parts, other types of complexity)

This approach is focused on mathematical of the sensory-motor input, irregardless of what the cognitive model predicts. Complexity of the stimuli seems to have a significant effect on choosing previously unfamiliar stimuli [Kidd et al., 2012], and can interfere with metrics of information gain in partially familiar settings. For example, choosing between an unfamiliar simple stimulus and a more familiar complex one may yield different results from different algorithms and can be a possible way to distinguish between them in behavioral experiments. In the context of developmental robotics and the choice of sensory channels the usage of morphological properties can push the system towards either tracking or exploratory activity, depending on whether synchronicity (picking correlating channels) or variance (picking channels with high variance) are used [Kaplan and Oudeyer, 2007].

Curiosity-driven learning in humans has been linked to better learning outcomes [Gruber et al., 2014, Gruber and Ranganath, 2019, Ellaway, 2014, ?, Malcom, 2019, Chen et al., 2022a] and applying a variety of curiosity-like metrics in ANNs have been shown to facilitate learning in artificial models as well [Schmidhuber, 1991, Thrun, 1995, Twomey and Westermann, 2017b, Poli et al., 2022]. Therefore, curiosity-driven learning can be hypothesised to be a strategy of creating an optimal curriculum.

## **1.10 The conflict between self-created and fixed curriculum.**

However, interaction/the choice between fixed curriculum and active learning is quite intricate. In human learning literature, there exists a long-standing conflict between

curiosity-driven learning versus a fixed curriculum [Bonawitz et al., 2011]. As has already been mentioned, curiosity has been demonstrated to enhance learning in many circumstances, both by improving memory [Gruber and Ranganath, 2019] and by helping students to learn important background information by deviating from a fixed curriculum [Ellaway, 2014, ?, Malcom, 2019] and helping to close knowledge gaps acting as a part of wider range of metacognitive abilities [Chen et al., 2022b]. However, formal education also has been shown to improve IQ [Ritchie and Tucker-Drob, 2018] and external reward has been shown to suppress intrinsic motivation [Deci, 1971].

This conflict had been somewhat resolved by the concept of zone of proximal development (ZPD), developed by Russian psychologist Lev Vygotsky [Vygotsky, 1978] to describe the range of skills and knowledge that a learner is able to acquire with the help of a more knowledgeable or skilled person. According to Vygotsky, the ZPD is an important factor in the learning process, as it represents the gap between what a learner is able to do independently and what they are able to do with the support of someone else.

The ZPD is often used in the field of education to guide the design of learning activities and to identify the most appropriate level of support for learners. By providing learners with tasks that are within their ZPD, educators can help them to learn new skills and knowledge while still feeling challenged and motivated. It can be seen as a strategy of combining both curiosity and curriculum learning and it has been shown to be beneficial for children's development of executive function, language acquisition, and academic abilities [Bernier et al., 2010, Carlson, 2003, Hammond et al., 2012, Hughes and Ensor, 2009, Hughes et al., 2013].

One way that educators can support learners in the ZPD is through scaffolding, which involves providing learners with the necessary support to complete a task or activity. Scaffolding can take many forms, such as providing verbal prompts, demonstrating a skill, or breaking a task down into smaller steps. By gradually removing the support as the learner becomes more proficient, educators can help learners to develop their skills

and eventually complete tasks independently.

Another form it could take is guided play [Weisberg et al., 2016], an effective approach to early childhood education that combines the benefits of both free play and structured learning, emphasizing the role of adults in providing support and guidance to children's natural exploration and discovery.

However with all the vast body of literature that exists on scaffolding, there is still no formal definition of it. Specifically, it's hard to separate the cases when it is absolutely essential and when it's just a facilitating mechanism that speeds up the process but isn't really necessary. In many works there's also a lack of clarity about which part of it is simply emotional support (helping a learner with their anxiety to allow them to better focus on a task) and which is providing better learning tools, as well as how to distinguish the situations where these better tools are needed.

If we go back to modelling, curiosity-driven algorithms have been shown to encounter problems when dealing with completely unpredictable stochastic data [Burda et al., 2018] and in developmental robotics they often fail when the need to switch between a variety of tasks arises [Oudeyer and Kaplan, 2009]. This has been attributed to imperfections in models and metrics not capturing curiosity correctly, however it is also possible that our models may be encountering the same problems that human children do. That's why a mix of curiosity and curriculum created by scaffolding is efficient and that's why conflicting evidence exists about intrinsic and extrinsic motivation. Specifying a learning curriculum for children (whether by parenting or by the process of biological maturation of the brain) could be a way of bypassing similar problems. Human children evidently cannot always manage to learn on their own and, perhaps, we should not yet expect our artificial models to surpass them.

To properly explore this topic we need to formalise a class of scenarios where curiosity fails to help with learning. Networks can provide a starting point for this: we can check where networks fail to learn, relying on curiosity alone, and then search for similar patterns in humans.

Another thing of importance is that the topic of early infant education curriculum and the intricacies of combining curriculum and self-driven learning has only recently gained attention, with researchers' focus still mostly being pre-school children [Wood and Hedges, 2016]. Same refers to curiosity: whether curiosity changes during various stages of development is currently unknown, as there exist little to no studies comparing children and adult curiosity [Gopnik et al., 2017, Pelz and Kidd, 2020]. This is especially important in the light of new findings about our earlier years of life.

## 1.11 Curriculum learning in humans

Despite all the wonders of human mind, humans often need external guidance to learn especially complex things. While it is fairly obvious that we should not teach differential equations to students before teaching addition, the finer details of building curricula are still often subject to debate.

One field where it is especially lacking is the role of curriculum in infant development. The effects of a study curriculum and correct learning patterns are well-studied in relation to higher order cognitive functions and children from the preschool age or older, and the word "curriculum" is not normally applied to younger infants. However, their learning experiences are heavily structured by their surroundings and biology: the examples include but are not limited to the process of brain maturation (e.g. limiting working memory and therefore the information an infant can focus on), other aspects of our biology such as visual system maturation, and the general statistical structure of the world around us.

Caregivers can also inadvertently create a curriculum for infants due to culturally-learned behaviour such as, for example, infant-directed speech, also known as "baby talk" or "motherese". It's a type of speech that is used when adults communicate with infants. It is characterized by a number of distinctive features, including a higher pitch, exaggerated intonation, and slower tempo compared to adult-directed speech.

Infant-directed speech is thought to be beneficial for language development in infants, as

it helps to capture their attention and make language more salient and easier to process. It has been found to be more effective at engaging infants and promoting language learning compared to adult-directed speech.

Apart from visual maturation such as changes in visual acuity biologically-causes curricula can even stem from side effects of embodiment, such as changes in vision from supine to crawling to walking (Smith et al).

All these things impose learning patterns that have been shown to be both beneficial for developing humans and reproducible in artificial models, improving their performance as well. This might lead to a lot of implications, taking into account how important are infants' early experiences for their consecutive life.

## **1.12 Learning universals, early development and the importance of slow brain maturation**

The desire to move away from labelled and carefully curated datasets has sparked a renewed interest in unsupervised learning, and mimicking strategies from infant and child development is seen as one of the possible inspirations [Zaadnoordijk et al., 2022]. This is as children are a perfect example of general intelligence, which the field of AI is struggling to reproduce, while lacking more specific higher-level skills in which AI models excel.

In addition to the benefits for AI development, it might also provide particularly valuable insights into infant development. Pre-verbal infants are notoriously hard to study, and modelling infant developmental trajectories with AI might help us to discover the "fingerprints" of certain learning strategies, facilitating easier behavioural testing of these strategies. It is also a good way to discover how environmental constraints influence learning and how environmental factors interact with internal biological constraints.

One of the most critical stages of infant development and learning happens before they reach the age of five. The first five years of life are crucial to children's health, wellbe-

ing, and the overall trajectory of their lives in a variety of ways [Shonkoff and Phillips, 2000]. The National Scientific Council on the Developing Child [on the Developing Child, 2007] emphasizes that early experiences, including structured learning opportunities, shape brain architecture and have long-lasting effects on children's cognitive, social, and emotional development.

The first two years are especially interesting, since from an outside perspective it seems as if not a lot of things happen: e.g., babies usually start walking around 10 and 18 months old, don't learn their first word until 12 months old and only start to use short phrases by the end of the 2nd year of their life. Until recently, it was considered that human children are born with an underdeveloped brain: a constraint imposed on our species by bipedalism which limits pelvis size. However, recent discoveries demonstrate that our brains are surprisingly well-developed from an early age [Cusack et al., 2018] and that slow maturation of skills exists for some other reason.

Some studies suggest that even slow biological maturation of infants' cognitive abilities leads to better learning [Elman, 1993, Gopnik et al., 2017, French et al., 2002, Cusack et al., 2018]. Comparing humans to other animal species also shows species with a more developed social structure, such as, for example, corvids, also tend to have longer weaning periods [Faust et al., 2020], which perhaps provide an environment for learning communication skills.

This suggests that the initial human infants' helplessness could be a crucial hidden learning period, a process of learning how to learn. Therefore, the way their "curriculum" is structured (albeit unintentionally by biological constraints, the world around us and interaction with caretakers) can play an immense role in laying the foundation for their future cognitive development. Early differences in learning setting might snowball into significant problems or advantages.

Using examples of early visual development, classic CNN tasks and models of active learning, in this work we aim to investigate how such curricula can emerge both in infancy and later years and how they can help overcome typical learning challenges.



## 1.13 Objective and methods

In this thesis we are trying to outline the cases where a human or an AI learner would need external help and structure and the ways in which this structure might be achieved in a natural environment, be it by the statistics of natural world, biological constraints or self-driven learning.

Rather than focusing on specific biological processes we want to search for universal principles governing learning that will likely be shared between biological and artificial neural networks at least to some extent.

In this work we investigate several types of learning constraints with the use of computational modelling. We are using examples from visual development and a variety of convolutional neural networks (CNN) due to their ability to mimic the human visual cortex.

In Chapter 2 we will look into how statistical properties of the world create a natural curricula and how it can potentially influence the formation of categories, resulting in a different speed of learning and representational organisation.

In Chapter 3 we will look into another type of curriculum where the gradual introduction of new information shifts the focus of the learner towards specific features and representations. In humans this corresponds to perceptual maturation of the visual system.

In chapter 4 we will investigate the limitations of self-created curriculum and self-directed learning, one of the main things that differentiates humans from ANNs.

The conclusion discusses the consequences of these findings both for theories of infant development and AI.

## 2 The effect of statistical properties of the world on category learning

### 2.1 Introduction

#### 2.1.1 Order of semantic categories acquisition

When children expand their vocabulary in their first years of life, their acquisition of semantic categories is unbalanced: some categories take priority for children, in that they are learned earlier and include more words [Rescorla, 1981, Rakison et al., 2003]. While there are significant individual differences in the rate and timing of language acquisition, the ordering of certain categories exhibits substantial consistency for different children and across different cultures [Braginsky et al., 2019]. Words, referring to sounds, games and routines, and body parts are all over-represented. Categories such as people, animals, vehicles, food, and clothing all are more variable across cultures. Small household items and furniture show less cultural variability [Braginsky et al., 2019].

#### 2.1.2 Area specificity

Existing research also suggests that humans have stereotypical brain locations for representing certain categories of stimuli. For example, neurological evidence strongly implicates a dedicated machinery for face processing in the human brain [Tsao and Living-

stone, 2008, Kanwisher and Yovel, 2006]. Specifically, the fusiform face area (FFA), a region located in the lateral ventral temporal cortex, demonstrates selective activation in response to faces, even when they are presented in degraded or inverted form [Kanwisher and Yovel, 2006]. Additionally, the extrastriate body area (EBA), situated more dorsally in the lateral temporal cortex, is responsible for the selective activation in response to bodies and plays a crucial role in processing body posture, movement, and body parts [Downing and Peelen, 2011, van Koningsbruggen et al., 2013]. Another category-specific region is parahippocampal place area (PPA) which is located in the medial ventral temporal lobe. It is selectively activated by scenes, such as images of places and buildings, and it is thought to be involved in scene recognition and navigation [Epstein et al., 1999].

A broader scale organisation of the ventral temporal lobe was proposed by Konkle and Caramazza [Konkle and Caramazza, 2013] Using fMRI they found evidence for two major dimensions that humans use to categorise objects: size and animacy. Object size drives differential responses only in the object domain, not the animate domain, yielding a tripartite distinction in the space of object representation.

Certain other categories, such as, for example, household objects have been shown to be processed in a distributed manner, where visual information is not processed by a single region, but rather by a network of regions. However, Huth et al. [Huth et al., 2012] demonstrated that the representation of thousands of object and action categories throughout the human brain can be effectively described by a continuous semantic space, encompassing both perceptual objects and action maps. In other words, nearby areas of the cortex tend to respond to similar categories.

Interestingly, those semantic maps were also found to be consistent across different human participants. When the experimenters employed computational models to decode the neural activity patterns within these regions, they were able to reconstruct the visual stimuli that the brain was processing [Huth et al., 2012].

In infants, one of the earliest distinctions learned is between animate and inanimate. At

10 months of age, infants are able to utilise animacy cues in order to focus their attention [Anderson et al., 2018]. One potential explanation for the emergence of stereotypical organisation in the visual cortex is the presence of "proto-organisation." A study by Kamps and Dilks [Kamps et al., 2020] suggests that certain regions of the visual cortex may have a predisposition to learn about specific types of visual stimuli, such as visually low spatial frequency and foveal stimuli in the fusiform face area (FFA). Through typical experiences, the FFA may become specialised for processing faces. It has also been shown that visual experience is not necessary for the development of face-selectivity in the lateral fusiform gyrus [Murty et al., 2020], suggesting a role for long-range connectivity in face-selectivity. The idea that connectivity may shape the development of selectivity is further supported by research from Cabral and Cusack [Cabral et al., 2022], who found characteristic patterns of connectivity for regions that will become object-selective even in neonates.

Another study [van den Hurk et al., 2017] has also demonstrated that the functional cortical layout in blind individuals shows remarkable similarity to the well-documented layout observed in sighted controls, at least for the distinction between faces, scenes, body parts, and objects. This suggests that visual functional brain organization does not fully rely on visual input, however, the question remains open for more detailed category structure and for the origin of such distinction.

The aforementioned tripartite structure is of particular interest due to the fact preference zones for the tripartite distinction are duplicated on the ventral and lateral surface of the brain which exhibit differential neural organisation. Such a duplication persevering across differently organised structures suggests that the categories' representational structure places some informative constraints on the nature of the underlying cognitive architecture. There are some objective properties in the data that prompt such a distinction.

### **2.1.3 Potential factors driving area specificity**

Several perceptual properties, including spatial frequency [Kamps et al., 2020], eccentricity [Todorović, 2009], and statistics of the visual data [Fausey et al., 2016], might promote in specific cortical regions. A possible factor, which we plan to investigate in our study, is occurrence frequency and contextual constraints of objects. If there are differences in the statistical properties of each category they could influence the organisation of representations in the visual cortex.

In comparison to other perceptual features, this is underexplored. Prior research has explored occurrence frequency for faces [Fausey et al., 2016], however, no study has yet characterized the relative frequency of more broad object categories in naturalistic experiences.

This task is notably challenging, primarily due to the significant amount of data required to obtain a representative sample. The available datasets concerning infants' early experiences are scarce, and even when provided with adequate video recordings, the process of tagging images is effort-intensive. The use of an image recognition artificial intelligence (AI) for labelling, also known as machine-tagging, could be a potential solution. Therefore, our secondary objective is to assess the feasibility of machine-tagging and the appropriateness of utilising existing video datasets for such purposes.

## **2.2 Challenges and methods**

### **2.2.1 Choosing a dataset**

A substantial challenge is acquiring an appropriate dataset. There exist notable differences between the experiences of children and adults. In an ideal setting we would want to continuously monitor infants' interactions with the world from birth, involving as many infants as possible, potentially hundreds.

Given the current limitations of available technologies and the scope of the current study,

this ambitious undertaking was not feasible, particularly under the scope of the present study. This leaves us with a task of evaluating the adequacy of the accessible data to address our research objectives.

The SayCAM dataset [Sullivan et al., 2021] has proven to be invaluable for describing and analysing infants the learning environment of infants. To create it infants aged 6–32 months wore a head-mounted camera for an approximate duration of two hours per week, over a period of roughly two-and-a-half years. The resulted in a large, naturalistic, longitudinal dataset of videos from the infant and child perspective, which is highly representative of naturalistic experiences. However, all these videos were filmed indoors, and only 3 children participated in the creation of this corpus. This set might not provide enough variation. Therefore, to investigate the effect of the limited number of people and contexts, a much broader dataset was created by utilising professional movies, which are plentiful, of high quality, and display considerable variability. 100 movies were chosen from IMDb (Internet Movie Database) American and European movies with high rating were selected, with the main criteria of containing everyday life objects and scenes (i.e. not displaying space or medieval fantasy scenes or prehistoric fauna).

Our research aims to address whether a movie dataset would be suitable for making inferences about infant experience by conducting a comparative analysis of the SayCAM and professional movie datasets.

## **2.2.2 Dataset tagging**

In this study we apply both human and automatic tagging of datasets. 'Machine' or 'automatic' tagging involves using an image recognition algorithm to automatically identify and label objects and scenes within images.

Automatic tagging of images has several benefits over human tagging, including speed, scalability, and consistency. Machines can process large amounts of data in a short amount of time, making it easier to create a large dataset. Additionally, machine tagging is not affected by subjective biases, which can result in tags that are more reliable and

more consistent across the dataset. However, machines may struggle with ambiguity and context, leading to inaccurate tags. In contrast, human tagging can provide more accurate and nuanced tags, but it is a slower and more labor-intensive process.

One potential problem is that automatic tagging prioritises a different level of detail. To assess what is a scene, we have to use classes defined by word labels. These can be done at different levels, such as basic, subordinate, and superordinate categories. While collecting object statistics with the help of automatic image recognition methods seems promising, it is unclear to what extent automatic methods focus on the same things as humans, e.g. whether they share human preference for basic-level categories [Rosch et al., 1976], such as preferring “dog” over “collie” or “canine”.

Another problem could arise if the effect of figure/ground in the case of human tagging and the absence of such in automatic tagging. Humans instinctively perceive objects as either being in the foreground or the background [McLuhan, 1970] and are more likely to pay more attention to the ‘figure’ and not mention an object they consider to be a ‘background’. Unless intentionally designed to do so, machine learning architectures do not share this tendency. It is hard to tell to what extents this distinction influences learning, as a large part of human visual perception is not conscious.

### **Tag post-processing**

Both machine- and human-made tags were reviewed by the researcher and each object was manually assigned several superordinate categories/category types.

As our starting point we used the tripartite structure (animate beings, small inanimate objects, large inanimate objects), as it represents a very basic distinction present in the brain and can be assessed even in the datasets with a small amount of data points.

We also used more specific categories, such as separating humans from other animals, breaking down small inanimate objects into categories such as ‘food’, ‘tools’, ‘clothing’ and large inanimate objects into ‘vehicles’, ‘furniture’, etc.

This is especially useful to control machine tagging quality, as some of these categories

are hypothesised to attract more attention from human participants than the others due to their non-perceptual properties (ability to move, utility). It might be an important point of difference between machine- and human-made tags, although indirectly this effect may be reproduced machine algorithms via training process or unaccounted linguistic effects.

We also tagged whether each label belonged to a superordinate, subordinate or basic category.

A full list of categories can be found on Github ([link](#)).

### 2.2.3 Metrics

When evaluating object statistics, an important factor to take in consideration is object frequency but it could also be important how context-specific an object is. Some objects or animals only appear in specific contexts while others are simply rare, which could affect, for example, how interesting an object is for an infant.

Therefore to evaluate the regularities in data we used two metrics. The first was frequency, another one was a tf-idf metric, commonly used in search engines and information retrieval to estimate term specificity.

Tf-idf, an abbreviation for 'term frequency-inverse document frequency', is a scalar statistic that was designed to reflect how important a word is to a document in a collection or corpus. This statistic increases proportionally to the frequency with which a word appears in the document and inversely to the number of documents in the corpus that contain the word, which helps filter out the words that appear more frequently in general. A normalised version of term-frequency weight was used (Equation 2.1).

$$tfidf(t) = \frac{tfidf_{max}}{df \log(tfidf)/\log(df)} \quad (2.1)$$

We also adapted tf-idf to the video dataset, treating video segments of the set as "documents" for the automatically tagged sets and separate screenshots for the human-



tagged sets. For movies we also analysed the effect of the video segment size to correct for imperfections of the scene segmentation procedure.

We also investigated the effect of level of categories used, using WordNet, analysing the dependency of frequency on the max level of hypernyms available for a synset. While imperfect, this approach does provide some insight into impact of category level on object frequency.

## **2.2.4 Statistical analysis**

To analyse the data we used non-parametric statistical methods, as the distributions proved to be far from normal. We used the Kruskal-Wallis test and Mood's median to compare the mean and the median of distributions, as both are insensitive to outliers.

These statistical tests are similar to the one-way analysis of variance (ANOVA) test, but they does not assume that the data is normally distributed, or that the variances are equal across groups. Such tests are preferable when the assumptions of parametric tests, such as ANOVA, are violated. Mood's median test in particular, has been shown especially useful when the variances across the groups are unequal.

The Kruskal-Wallis test is a non-parametric statistical method used to compare independent groups that are not normally distributed, that is usually applied to 3 or more groups, however pairwise comparison is viable as well. It is similar to the one-way analysis of variance (ANOVA) test, but it does not require the data to meet the assumption of normality. Instead, this test ranks the data and compares the sum of ranks between the groups. The null hypothesis is that there is no significant difference between the groups, while the alternative hypothesis is that at least one group has a different median than the others. The test statistic is calculated based on the ranks and the number of observations in each group, and its significance is assessed using the chi-square distribution. The Kruskal-Wallis test is a useful tool when working non-normal data or when the assumption of homogeneity of variance is violated.

The Mood's median test, also known as the Brown-Mood median test, is a non-parametric statistical method used to compare two independent groups of data that are not normally distributed. It is an alternative to the independent samples t-test, which requires that the data meet the assumption of normality. The Mood's median test compares the medians of the two groups rather than their means, making it robust to outliers and non-normal distributions. The test involves ranking the data, calculating the difference between the group medians, and testing whether this difference is significant using a chi-square distribution. The null hypothesis of the test is that there is no significant difference between the medians of the two groups, while the alternative hypothesis is that there is a significant difference. It is useful when comparing two independent groups with non-normal data, particularly when the sample sizes are small or the data is heavily skewed.

Due to the small set size, we also used bootstrapping for the manually tagged set to smooth out some effects when plotting confidence intervals.

All data and methods can be found on Github ([repository link](#)).

## **2.3 Experiment 1: professional movies dataset, automatic tagging**

### **2.3.1 Data and methods**

For our first experiment we used the professional movies dataset. It consisted of 100 movies by a variety of directors with an average duration of 1.58 h, sampled at 1 Hz. All movies were of realistic genres (i.e. not containing space scenes or prehistoric fauna or other objects an infant is unlikely to encounter in their daily life), rated P12 and not displaying any scenes that would be stressful or disturbing to the participants. A full list can be found in Appendix 1. The movies were then automatically tagged using Amazon's Rekognition service,

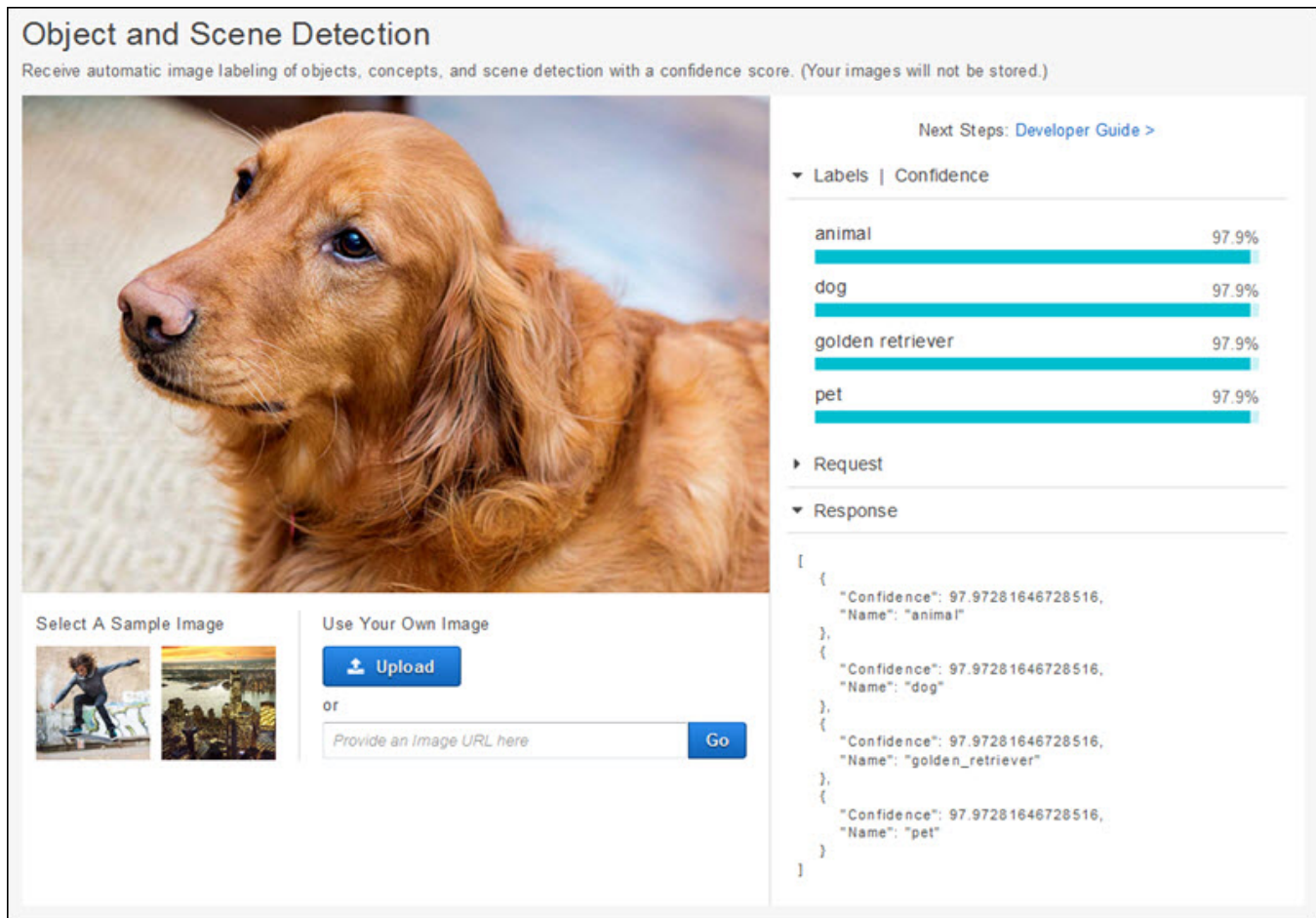


Figure 2.1: An example of tagging performed by Rekognition. Confidence score indicates the level of certainty with which the tag was assigned.

Amazon Rekognition is a cloud-based computer vision service that uses a CNN-based deep learning algorithm to identify objects and scenes within the image. Based on the objects identified, Amazon Rekognition assigns relevant tags or labels to the image. These tags can include labels such as "person," "dog," "mountain," "beach," "car," and many others, depending on the objects and scenes present. Each tag assigned by Amazon Rekognition comes with a confidence score, which indicates the level of certainty with which the tag was assigned. The tagged image is then returned along with the assigned tags and confidence scores.

Snapshots at 200 ms intervals were tagged from the 158.4 hr of videos, resulting in 2851200 data points. For our tasks we used only labels with confidence score larger than 90%. Outliers mentioned only once and categories describing actions instead of

objects ('running', 'people having conversation') were dropped from the analysis.

For each category discovered by Rekognition, tf-idf score and frequency were calculated.

As it is hard to automatically detect the boundary of a narrative segment, for the tf-idf metric the set was split into segments of fixed length. We investigated the effect of segments of size on the category distribution using segments with 1000 (approximately 3 min), 3000 (approximately 9 min), 10000 (approximately 30 min) and 30000 (approximately 1 h 30 min) data points.

### 2.3.2 Results

Tagging the set resulted in a list of approx 2000 categories (a full list can be found at Github).

As can be seen in Figure 2.2 the influence of video segment size on the distribution of categories can be considered negligible, with discernible effects only evident for exceptionally infrequent categories. Consequently, a segment size of 9 minutes, comprising 3000 datapoints, was chosen for subsequent analysis to ensure a balance of smoothness and scene detailisation. This resulted in a set of 950 segments.

For segments of 3000 data points, frequent, rare non-specific, and rare context-specific terms could all be identified in the dataset (Fig. 2.3).

It is evident from the visualisation presented in Figure 2.4 that the tripartite categorisation system yields three distinctly separate clusters. Tf-idf allowed us to separate rare non-specific from rare context-specific items (Fig. 2.3), however, when subjected to statistical analysis (as presented in table 2.1), pure frequency turned out more predictive of category than context specificity. Both Kruskal-Wallis test and Mood's median showed significant statistical difference between frequencies of animate vs. large, animate vs. small and small vs. large comparison ( $p < 0.001$  in all cases). Conversely, there was significant statistical differences ( $p < 0.05$ ) in tf-idf scores between animate and inanimate

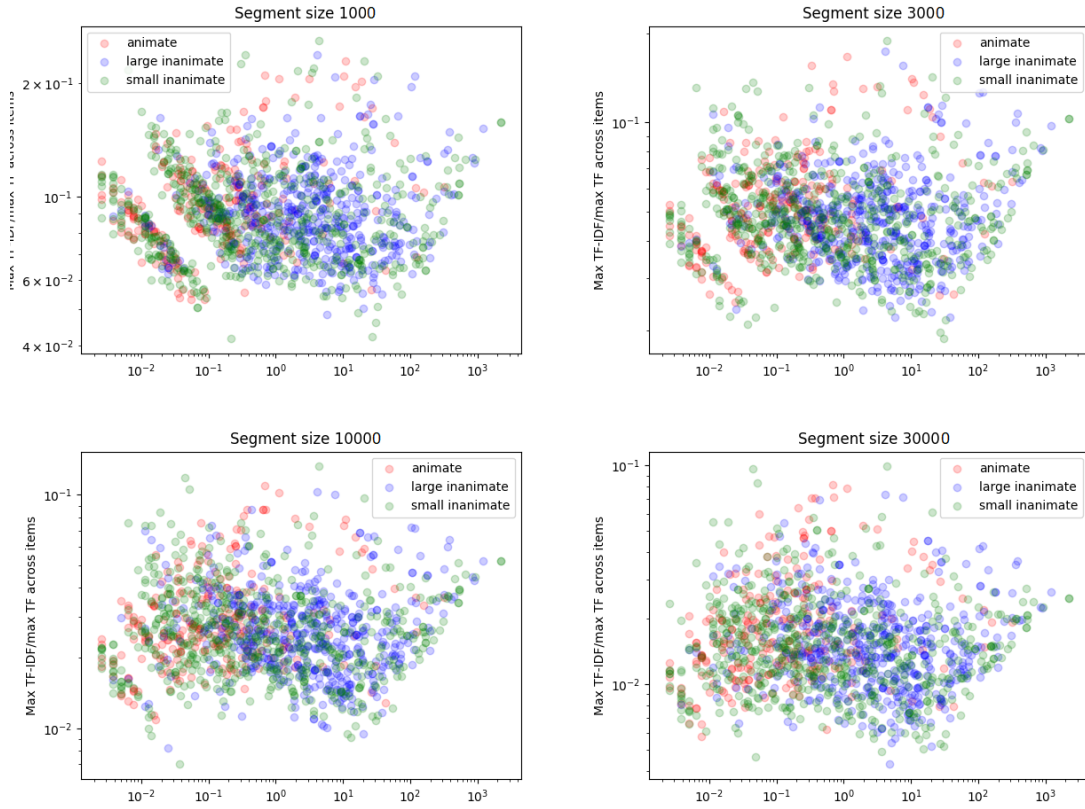


Figure 2.2: The impact of the video segment size on the category distribution. Difference was noticeable only for exceptionally rare categories.

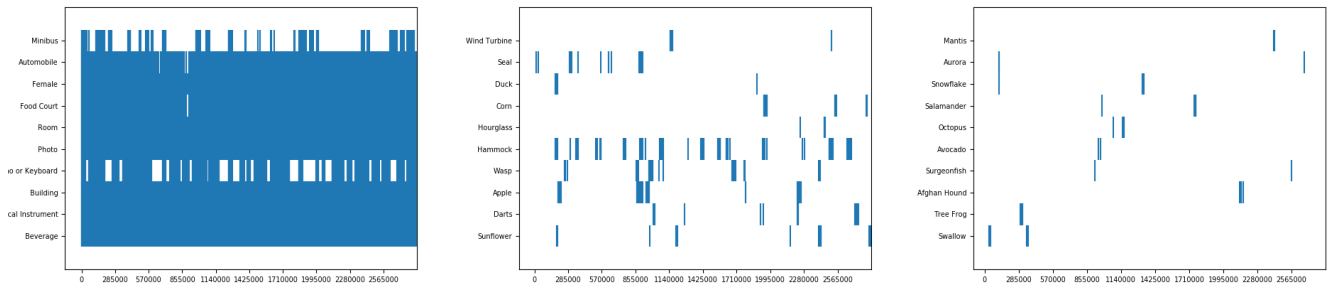


Figure 2.3: Occurrences of entities in the segments of the movie corpus. From left to right: frequent, rare non-specific (low frequency, low tfidf), rare context-specific (low frequency, high tfidf).

objects but not between large and small inanimate ( $p > 0.1$  for both tests). Therefore frequency proved to be the sole parameter that enabled differentiation between large and small inanimate categories. Both the idf and frequency scores demonstrated the ability to differentiate between animate and inanimate categories.

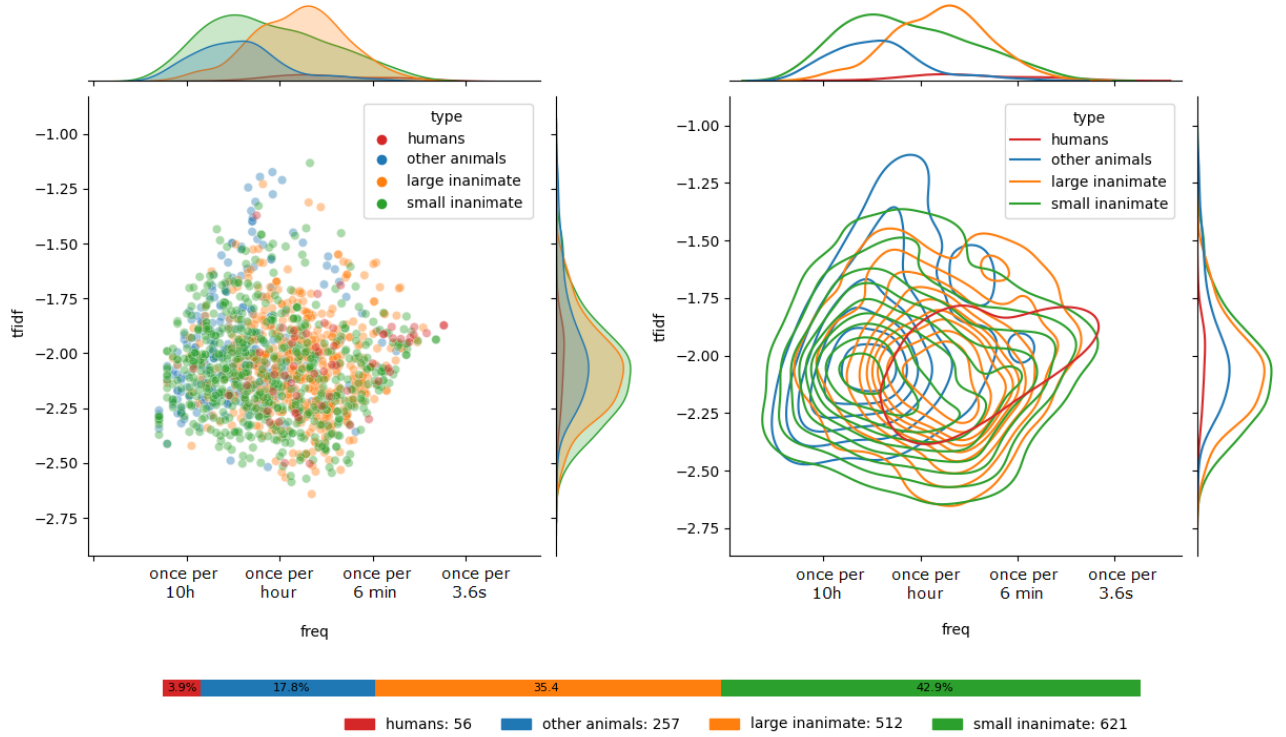


Figure 2.4: Tf-idf (in logarithmic scale) and frequencies for the Rekognition tagged professional movie set. Segment size = 3000 data points (9min). For scaling reasons  $\log_{10}(tf - idf)$  is displayed on the y axis.

| metric | category         | Family Movies<br>(Rekognition tagged) |               |
|--------|------------------|---------------------------------------|---------------|
|        |                  | Kruskal-Wallis                        | Mood's median |
| tf-idf | Animate vs large | 0.00029                               | 0.01344       |
| tf-idf | Animate vs small | <0.001                                | 0.0006        |
| tf-idf | Large vs small   | 0.1636                                | 0.3563        |
| freq   | Animate vs large | <0.001                                | <0.001        |
| freq   | Animate vs small | <0.001                                | <0.001        |
| freq   | Large vs small   | <0.001                                | <0.001        |

Table 2.1: Comparison for the family movies dataset, the table displays the p-values associated with the various statistics. Only the frequency metric was capable of distinguishing between large and small inanimate categories. On the other hand, both tf-idf and frequency scores were found to be effective in distinguishing between animate and inanimate categories.

## 2.4 Experiment 2: infant experiences dataset, automatic tagging

### 2.4.1 Data and methods

In our second experiment, we used the SayCAM dataset [Sullivan et al., 2021] to investigate whether the statistics of objects observed by infants was similar to professional movies.

SAYCam is large, longitudinal audiovisual dataset recorded from the infant’s perspective. It consists of recording from 3 children aged 6-32 months who wore a head-mounted camera for approximately 2 h per week, over the course of approximately two and a half years.

It totals to 460 recording sessions, a recording session lasting approximately 60 to 80 min, each split into multiple video files, with a maximum single-file duration of 30 min.

The same tagging procedure as in Experiment 1 was employed. For the tfidf calculation the set was split into segments of 3000 data points (9 min). The statistic of categories derived from it was then compared with results of Experiment 1, allowing to evaluate category distribution under identical tagging conditions.

### 2.4.2 Results

Tagging the SAYCam set resulted in a set of 14700 data points. As can be seen on Figure 2.5 tripartite categories still form three notably different clusters, however, in comparison to the professional movies dataset, the clusters are less easily separable visual.

Same as in the previous experiment, both Kruskal-Wallis test and Mood’s median showed significant statistical difference between frequencies of animate vs. large, animate vs. small and small vs. large comparison, with  $p < 0.001$  in all cases (Table 2.2). Tf-idf score proved completely irrelevant in this case even for the animate/inanimate distinction. It is

unclear whether this effect can be attributed to differences in the statistics of categories or the limitations of dataset: i.e. fewer data points, improper segmentation or small amount of infant participants leading to insufficient variation in data.

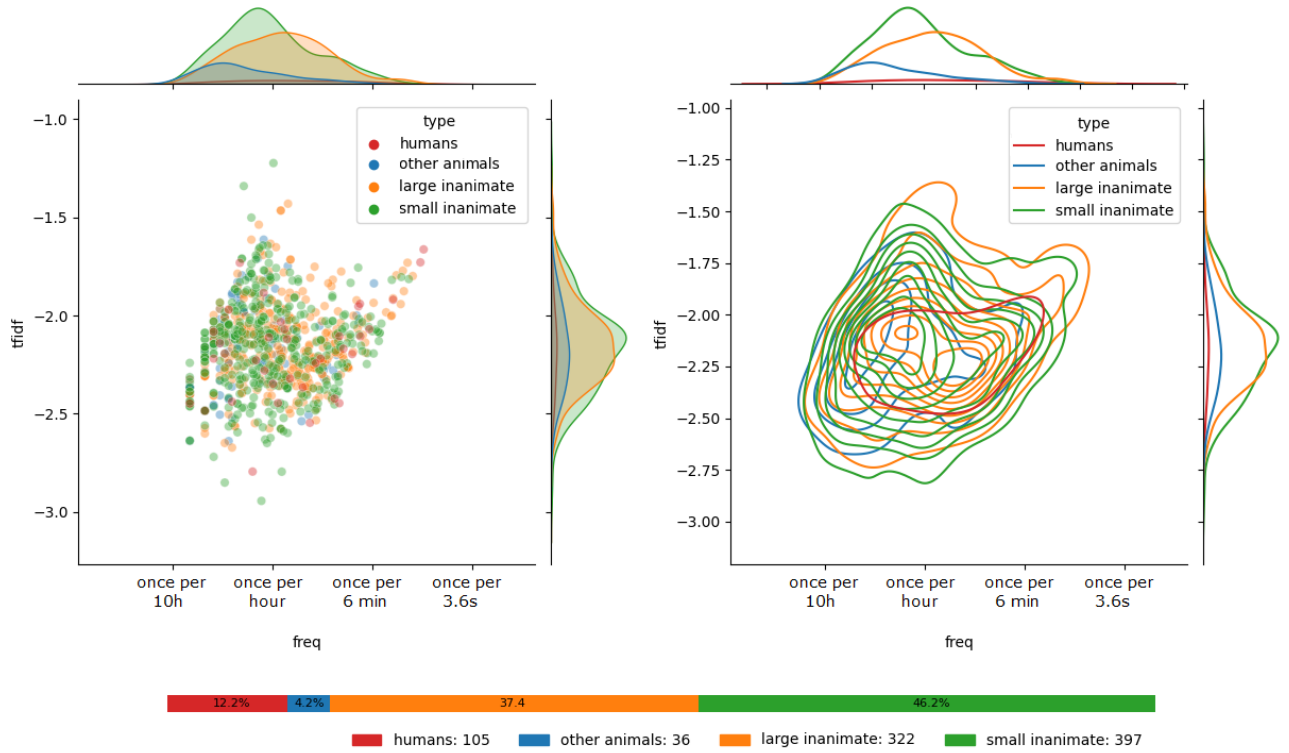


Figure 2.5: Tf-idf and frequencies for the Rekognition tagged SAYCAM set, segment size = 3000 data points (9 min). For scaling reasons  $\log_{10}(\text{tfidf})$  is displayed on the y axis.

| metric | category         | SayCAM         |               |
|--------|------------------|----------------|---------------|
|        |                  | Kruskal-Wallis | Mood's median |
| tf-idf | Animate vs large | 0.31022        | 0.6731        |
| tf-idf | Animate vs small | 0.6048         | 0.2724        |
| tf-idf | Large vs small   | 0.5273         | 0.7328        |
| freq   | Animate vs large | <0.001         | <0.001        |
| freq   | Animate vs small | <0.001         | 0.0005        |
| freq   | Large vs small   | 0.0004         | <0.001        |

Table 2.2: Statistic comparisons for the SAYCAM dataset, the table displays the p-values. Tf-idf proved ineffective when distinguishing between any of the categories, frequency emerged as a more effective metric for this purpose.

When plotting frequency against tf-idf, the automatically tagged professional movies set and SayCAM set provided almost identical pictures (Figures 2.4 and 2.5). This suggests that even despite small number of participants SayCAM set may be sufficient to analyse



the statistics of the categories encountered by an infant. These results also demonstrate that movies are surprisingly accurate in capturing the statistics of the world.

## **2.5 Experiment 3: Professional movies dataset, manual tagging**

### **2.5.1 Data and methods**

Our third experiment again utilised the professional movies set, however, this time it was manually tagged. We compared two types of tagging for this tag: automatic (as done in Experiment 1) and human.

For manual tagging Amazon Mechanical Turk (MTurk) was utilised. Mturk is a crowdsourcing platform created by Amazon in 2005 that enables businesses and individuals to outsource tasks to a large pool of remote workers.

All the participants accessed the data virtually through the MTurk website on an electronic device of their preference. All participants were English-speaking but not necessarily native speakers. There were no restrictions on age and gender. In accordance to MTurk rules, all participants resided in the US.

Prior to the study ethics approval was acquired from the Trinity College Dublin Ethics Committee. The details can be found in Appendix 1.

After confirming their consent, each participant was presented with a series consisting of 100 movie screenshots (1 screenshot per movie) and 5 additional hand-drawn sketches used for filtering out bots and low-effort answers, and had to name what they saw by typing into a response box. No priming stimuli were presented. The study was self-paced and there was no time limit on each answer.

200 people participated in the study, each participant saw one of 40 possible sequences. Each sequence was presented exactly to 5 people and contained a different set of screenshots.

Each screenshot belonged to of the frames previously tagged by Amazon Rekognition to allow analysing the level of categories utilized by machine algorithms and human participants.

To filter out bots 5 CAPTCHA style images were added to each sequence and randomly presented. CAPTCHA stands for "Completely Automated Public Turing test to tell Computers and Humans Apart". It is a type of challenge-response test that is used to determine whether the user is a human or a machine. As automatic recognition software experiences difficulties, recognizing hand-drawn pictures, in this case images that would filter out bots consisted of pencil sketches of 5 categories (fish, tree, pencil, sheep, leaf) and labels with possible spelling variations to account for potential categorization inconsistency (such as 'sheep'/'goat' or 'leaf'/'maple leaf') or identifiable typos. Participants failing to correctly identify the sketch were filtered out automatically.

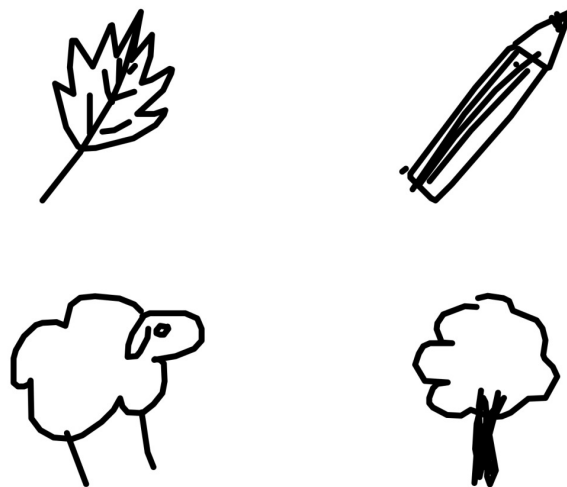


Figure 2.6: Examples of pictures to filter out bots. Automatic recognition software experiences difficulties, recognising hand-drawn sketches, while the answer is obvious for human participants.

To preclude spam answers on the online platform, results were also manually assessed. Any answers showing low effort (e.g. always entering the same object name, entering nonsense characters, answering with so distant a result as to indicate the participant was not paying attention) were discarded.

For discarded replies the sequence was ran again for another participant to ensure equal

amount of participants viewing each picture.

## 2.5.2 Results

Tagging resulted in a set of 40000 data points. As can be seen from Fig.2.7, in the manually-tagged dataset clusters are harder to separate visually. The 'vertical line' pattern observed in the picture for non-frequent entities is an artifact of a small sample size and is comprised of objects that only appeared no more than 3 times in the set. Those only appearing once turned out to be mostly typos ('aplle' instead of 'apple') and were excluded from further analysis.

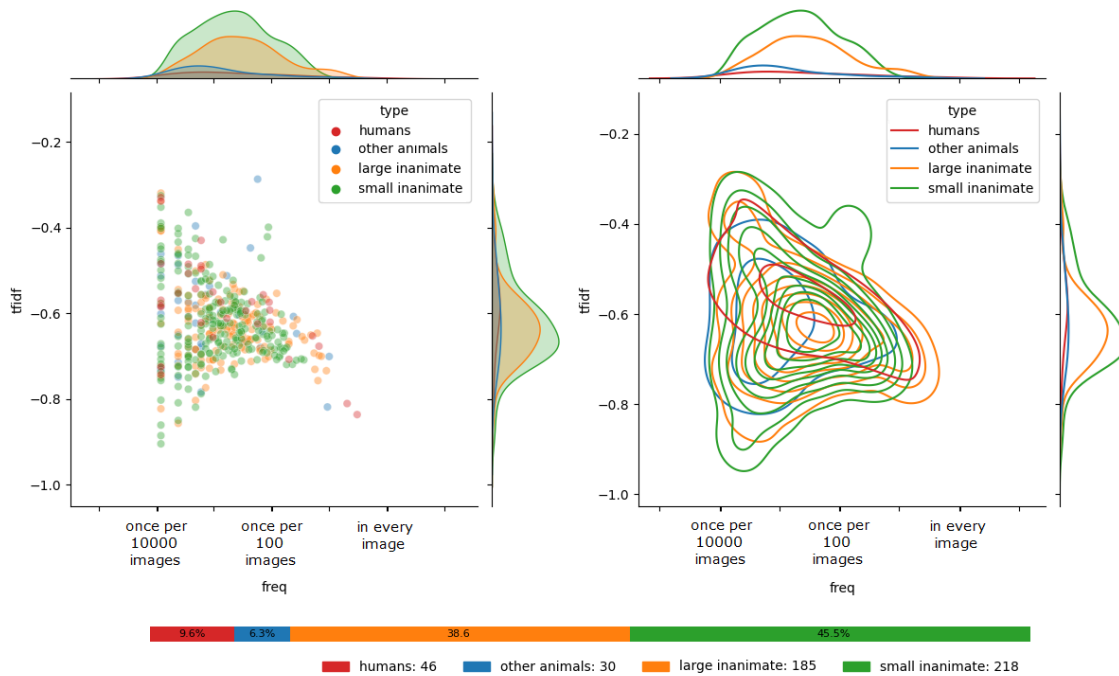


Figure 2.7: Tf-idf and frequencies for the manually tagged professional movies set, segment size = 1500 data points (15 min). For scaling reasons  $\log_{10}(tfidf)$  is displayed on the y axis.

Statistical analysis demonstrated that frequency is still a good predictor of category type (Table 2.3), even on a human-tagged dataset of much smaller scale. In this case, only Mood's median allowed to distinguish between the two sets ( $p < 0.01$  for animate vs. both inanimate and  $p < 0.05$  between small and large objects).

| metric        | <i>category</i>  | Movies (MTurk) |               |
|---------------|------------------|----------------|---------------|
|               |                  | Kruskal-Wallis | Mood's median |
| <b>tf-idf</b> | Animate vs large | 0.0009         | 0.1296        |
| <b>tf-idf</b> | Animate vs small | 0.0001         | 0.0742        |
| <b>tf-idf</b> | Large vs small   | 0.6819         | 0.4100        |
| <b>freq</b>   | Animate vs large | 0.0312         | 0.0002        |
| <b>freq</b>   | Animate vs small | 0.2098         | 0.0004        |
| <b>freq</b>   | Large vs small   | 0.2334         | 0.0282        |

Table 2.3: Statistic comparisons for the manually tagged professional movies dataset, the table displays the p-values. Once again, Tf-idf proved ineffective when distinguishing between any of the categories, frequency proving to be a more effective metric.

## 2.6 Follow-up analysis

### 2.6.1 Comparison of sets

Bootstrapping across categories was used to calculate confidence intervals. Confidence intervals, while not contradicting the initial theory, were inconclusive (2.8), however, statistical analysis shows that frequency can be used to differentiate between these three types of categories. In all sets there appear to be significant difference between frequency of animate, small inanimate and large inanimate objects.

In the human-tagged movie set the size of effect for small and large inanimate objects was much smaller and it is unclear whether such a significant difference in representation can be driven by small differences in occurrence. However, this lack of difference can also be attributed to the small size of the set and further investigation is required.

This discrepancy can also be partially explained by an artefact of tagging: humans paying more attention to objects and not to context.

It is also worth noting that while frequency had a significant effect on learning outcome and could act as a predictor for the animate/inanimate distinction, tf-idf was not as effective for predicting the same things. It is possible that this metric required a more precise scene segmentation, however, analysis of the impact of segment size appears to be insignificant (2.2).

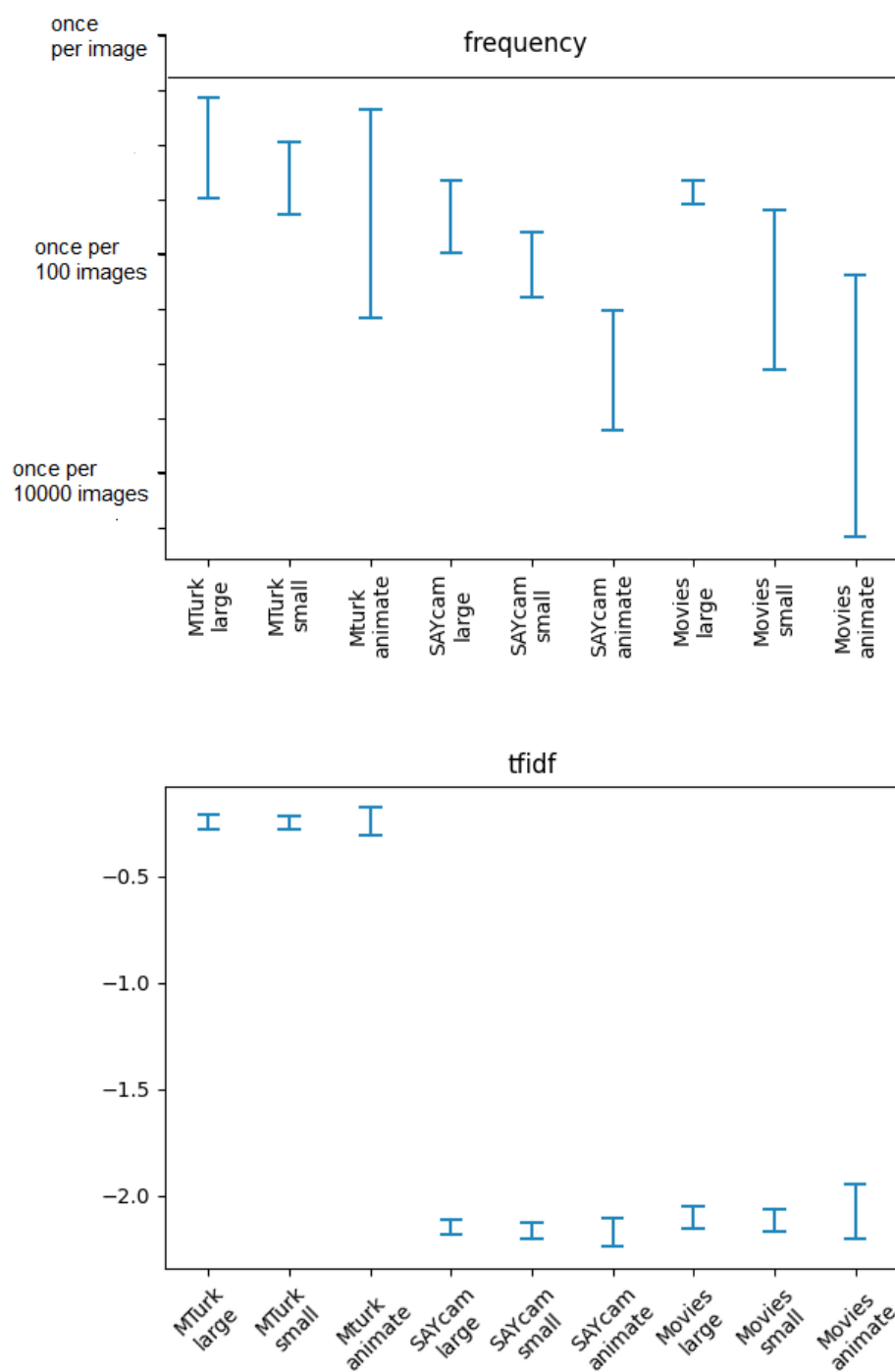


Figure 2.8: Confidence intervals and mean for small inanimate, large inanimate, animate, bootstrapping for all 3 sets. For scaling reasons  $\log_{10}(tfidf)$  is displayed on the y axis.

This suggests that either a better metric is needed or that context or term specificity are not relevant for humans in this particular case.

## 2.6.2 The impact of category level.

Some of the differences between manual and automatic tagging could be attributed to human preference for broad, basic categories and machine preference for more specific, subordinate categories.

To explore whether category level may be predicted by term's frequency we used WordNet levels of hierarchy. While there is no explicit "level number" in WordNet entries, it is possible to get an approximation by going up from the synset position and counting how many hypernyms are between it and the top of the tree. This is a crude estimate, however, at the moment there are no datasets allowing a better one.

Visual representations of levels against frequencies (Figure 2.9) suggests that the effect of level on the statistical properties of the category is minor. This may be attributed to machine tagging methods adding superordinate categories to the label pool rather than replacing the basic ones.

Regressing frequency against category level also showed no significant correlation, p value consistently being  $> 0.05$ .

## 2.6.3 Statistics of the world and category acquisition order.

Out of the three sets, only the automatically tagged professional movies set turned out large enough to allow looking into distribution of more specific categories (Figure 2.10). However, as previous experiments demonstrate, it can be seen as a viable replacement of more a human-tagged everyday scenes set.

As had been mentioned in the introduction such categories as animals, vehicles, food, and clothing all are more variable across cultures, while small household items and furniture show less cultural variability [Braginsky et al., 2019].

Furniture and small household items indeed demonstrated a higher frequency compared to other object types (Figure 2.10), which may account for the simultaneous acquisition of these categories by infants from diverse cultural backgrounds. However, the clothing

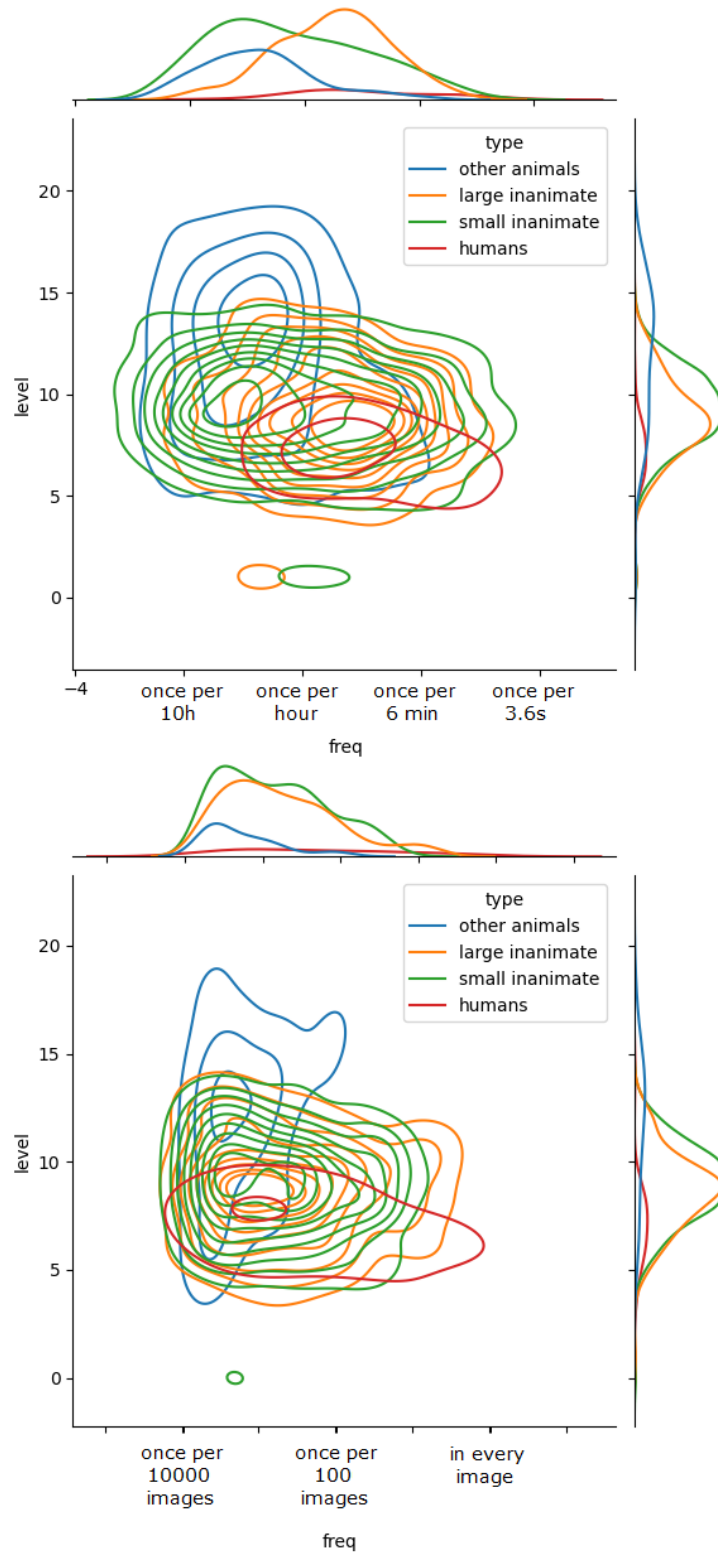


Figure 2.9: The effect of level on the statistical properties of the category was minor. From top to bottom: automatically tagged professional movies set, manually tagged professional movie set.

category, despite its comparable prevalence, does not exhibit the same learnability, and the pace of acquisition varies across cultures. It remains unclear whether this discrepancy can be ascribed to differences in human and AI perception (e.g., an infant's attention to worn vs stored clothing) or the increased difficulty in identifying clothing as a unified category due to greater perceptual variability.

It is crucial to acknowledge that these films pertained to Western culture, and the distribution of household objects and other statistics may differ in other regions, which would make these results invalid. Further studies on sets from different cultures should be conducted.

In conclusion, the statistical properties of the environment may contribute to certain consistencies in category learnability; however, this does not rule out the influence of other factors, such as perceptual properties and function, on the learning process.



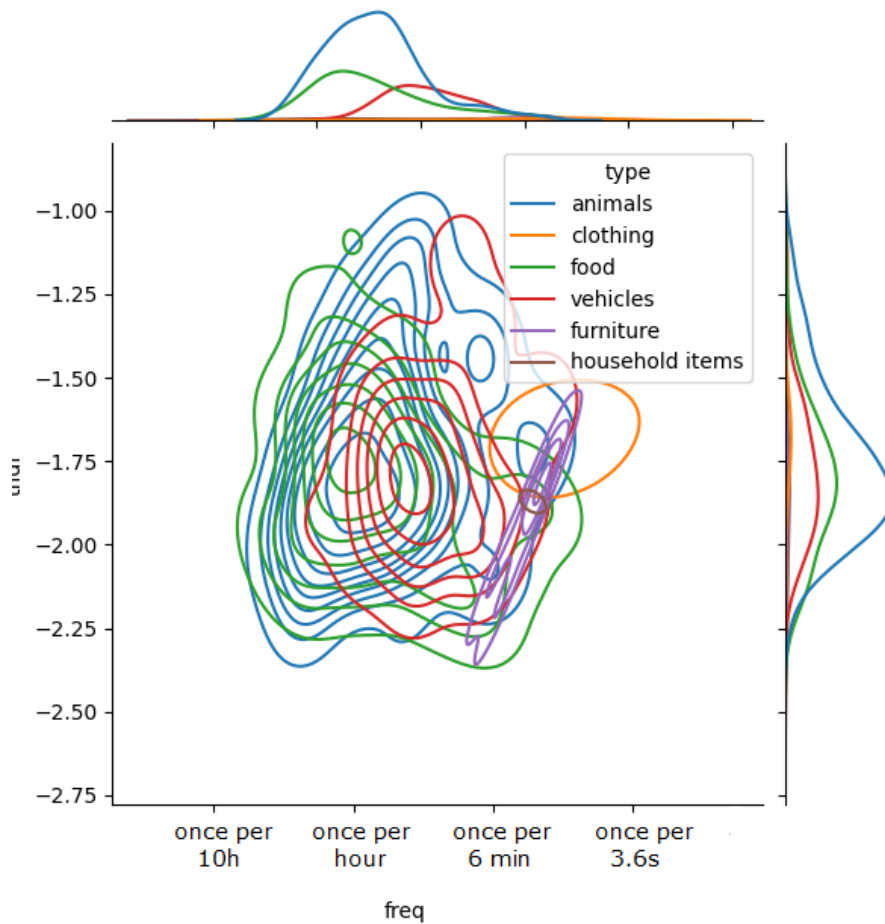


Figure 2.10: Tf-idf and frequency of more specific categories. Automatically tagged professional movies set.

## 2.7 Discussion

This study has demonstrated a clear distinction in the statistical properties of animate and inanimate entities encountered by both adults and infants in their daily lives.

The animate/inanimate distinction was the most prominent and it is possible that these differences render specific areas of the visual cortex more receptive to acquiring knowledge of particular category types.

Some differences in the statistics of large and small inanimate things can be observed too but it is unclear if it's enough to drive such a significant variance in representations.

Due to the fact that for animate objects the effect is much larger it is also possible

that, while large and small inanimate objects are naturally organised by size, for animate objects this pattern of organisation is disrupted due to importance of the statistical properties of the category.

Context specificity seemed to provide no effect on learning, appeared to have no discernible impact on learning, indicating that infants may rely on alternative mechanisms, or that a more refined approach is required, such as for example, evaluating neural representations inside the category.

We have also demonstrated the viability of automatic tagging and the possible pitfalls a researcher may encounter. Depending on the topic of the study, in most cases automatic tagging may be safely used. However, for studies involving comparison of basic and superordinate categories at least some level of comparison would be required.

Our study has also revealed that a limited number of participants did not significantly constrain the diversity and the statistics of categories present in the SAYCam dataset. Additionally, we demonstrated that professional movies sufficiently capture infant experience in the absence of dedicated infant datasets, making it suitable for research on basic categories. However, it would be good to extend this in future work. The movie datasets will have been biased, in that movie directors intend to convey a narrative, and so focus on the some things at the exclusion of others. As a result, the distribution of animate and inanimate objects in such videos might significantly differ from those observed in naturalistic settings. Although it could be argued that these days children are also educated by the TV screen as much as by the outside world, videos of this type may not accurately reflect the infants' visual experiences. A significantly larger dataset, the PLAY set (Adolph et al., 2019), is presently under development, although at the time of writing, not yet available. We anticipate that the findings of the present study will provide valuable insight for future projects based on it.

One notable constraint of this study was the limited quantity of manually labelled data that was accessible.

Apart from gathering more data, another possible solution could involve using other ap-

proaches to dealing with data limitations, for example bootstrapping pairwise differences instead of the set itself.

Another issue pertained to the imprecise segmentation of scenes during automated tagging, potentially impeding the efficacy of the tf-idf scoring method.

Future investigations may involve analyzing a more extensive corpus of infant experience data to study subordinate-level categories in more detail and utilizing either manual or AI-driven scene segmentations to more effectively assess the relevance of context.

It would could be informative to train a set of artificial networks with varying learning rates or other hyperparameters to explore the changes in the speed of learning category types with certain statistical properties.

An additional promising avenue for future research involves comparing our present findings with fMRI data, as well as exploring the impact of frequency and context specificity on neural representations of categories.

## 3 Benefits of low initial acuity

### 3.1 Introduction

Human infants are born with relatively poor vision. A measure of the sharpness of vision is called visual acuity and is typically expressed in terms of cycles per degree (CPD), which measures the number of cycles of a sine wave that an individual can discern per degree of visual angle. For a human eye with excellent acuity, the maximum theoretical resolution is 50 CPD, however, infants are born with 4 CPD, which makes them unable to recognise small details with precision during their first months of life; in most countries this would be considered legally blind.

Due to a combination of factors, visual acuity improves rapidly between birth and age 6 months, while only a marginal improvement is observed between the ages of 6 and 12 months [Mayer and Dobson, 1982, Norcia and Tyler, 1985]. One of the aspects contributing to that is the continued development of cone cells. Research conducted by Banks and Bennett [Banks and Bennett, 1988] demonstrated that the cone photoreceptors in the central retina of newborns have a diameter exceeding 7.5 microns, only reducing to the typical adult width of 2 microns over a period of approximately three years. As the retina matures and photoreceptors attain their adult size, cone density in this region rises from 18 photoreceptors per 100 microns to 42 photoreceptors per 100 microns. The reduced number of receptors in infants consequently results in diminished resolution and compromised acuity.

In parallel with retinal development, there is an increase in connectivity between cortical

neurons in the period after birth. It was shown first in primates [Rakic, 1981, Lund et al., 1993, Goldman-Rakic, 1987, Yeterian et al., 2012] and then in human infants [Huttenlocher et al., 1982] that there is a rapid increase in synaptic density following birth. At birth, infant monkeys' exhibit synaptic density approximately equivalent to the densities found in adult monkey brains. The number of connections continues to increase rapidly, peaking at the age of 3 months in the visual area (the peak stage varies from region to region). At that age the peak densities are twice those seen in adult monkeys. This pattern persists until 3 years of age, the age of sexual maturity. At age 3, synaptic densities begin to rapidly decrease, finally stabilising at adult levels at age 4 to 5 years.

In 1982 Huttenlocher reported a similar profile of changes in synaptic densities over the lifespan in the human visual cortex. Again, at birth, synaptic densities in this brain area are close to adult levels. In most parts of the brain densities increase rapidly between 2 and 4 months of age and peak between 8 and 12 months of age at levels 60% higher than those seen in adults. In the visual area, the process takes longer, stabilising at adult levels only at 11 years of age.

To summarise, the first year of development witnesses a great elaboration of synaptic connections. The full effect of this is unclear, however, it is likely that the quality of learning during that period would be crucial for later functioning. Slow development of retina hinders the ability of infants to orient in the world and learn fine details during this critical time, which seems suboptimal. However, recent findings suggest that this slow maturation of vision may in fact be an adaptive mechanism, encouraging the development of the ability to perceive global features rather than local details.

Studies have shown that patients with congenital cataracts who have their sight restored as early as two months of age experience deficits in spatial integration that persist into adulthood. These impairments include difficulties in processing configural face information and a reduced ability to perceive global form, movement and illusory contours [Vogelsang et al., 2018, Ellemberg et al., 2000, Lewis et al., 2002, McKyton et al.,

2015, Robbins et al., 2011]. Children treated for congenital cataracts have been shown to be able to perform face discrimination tasks where the patterns differ in local features, but not configural changes [Grand et al., 2001, Grand et al., 2004]. Therefore it is reasonable to assume compromised spatial integration in such cases. [Vogelsang et al., 2018] suggested that the initial blurry experience of typically developing infants is adaptive, allowing them to see the "big picture" and develop the representations that support configural face processing. To explain the deficits following cataracts, Vogel-sang proposed the high initial acuity (HIA) hypothesis, which attributes these deficits to the immediate availability of high spatial frequency information in the visual environment from the onset of sight thus triggering local processing bias compared to typically developing infants.

A similar bias is displayed by Convolutional Neural Networks [McKyton et al., 2015, Brendel and Bethge, 2019], which have been shown to be otherwise viable models of the ventral visual stream [Zhuang et al., 2019, Yamins et al., 2014]. Until recently, the consensus in the computer science community was that this preference could be attributed to the architectural limitations of the existing CNN models. However, a recent study [Geirhos et al., 2019] demonstrates that a network can be pushed towards global feature preference by altering the training set, improving performance and robustness to noise in the process.

Along with the data from patients with cataracts, this suggests that human preference for global features emerges at least in part from the learning curriculum rather than the intrinsic architecture of the visual cortex itself. Furthermore, these findings imply that the human learning trajectory observed during infancy can be replicated in a CNN through appropriate modifications to the training set, improving performance of the network as a side effect. In [Vogelsang et al., 2018], this hypothesis was confirmed for stimuli depicting faces. The methods used in the experiment were later subject to some criticism [Katzhender and Weinshall, 2019], which has called into question the strength of the results obtained. The study also did not account for the fact that cataracts allow for residual vision, which brings up the question whether a specific level of low acuity

or the dynamics of acuity change (e.g., abrupt vs. gradual) might also have an effect. Despite this, the study remains an important contribution to the field and has provided valuable insights into the impact of curriculum during early visual learning.

It is also important to note that in humans it may be hard to disentangle the generic effects of initial low acuity from the effects of specific face-specific neural machinery, as face recognition may only partially be relying on global processing. In a child with congenital cataract, the physiological mechanisms of neural maturation continue to proceed despite the lack of input from retina, and other critical periods may play into effect. It is therefore possible that a stronger effect may be observed in other domains, where global processing plays a bigger role.

Here we suggest a domain-independent approach, using another way to evaluate the HIA hypothesis, based on the shape vs texture bias, which is well-researched both in developmental and computer science. This also allows us to evaluate the benefits of this learning trajectory when encountering a variety of image distortions.

Perception of shape and texture can be seen as excellent examples of global and local processing. Recognising shapes requires the ability to integrate individual pieces of information across expansive areas, with configural information and overarching form taking precedence in recognition. Texture, on the opposite end of the spectrum, requires analysing minute patterns and grains on the object's surface.

Geirhos and colleagues [Geirhos et al., 2019] developed a paradigm to quantify the tendency of a network to classify an image based on its global shape or local details, texture. Neural Style Transfer [Gatys et al., 2016] takes a texture image and applies it to the shape of another image 3.2. Using this technique, cue conflict stimuli were created which possess incongruent texture and shape, such as a cat with an elephant texture. These stimuli are presented to a trained CNN and the proportion of correct answers that correspond to the shape of the stimulus is designated the shape bias of the network. To push the network towards shape bias Neural Style Transfer was also used, albeit in a different way: a variety of distortions are applied to the training set, eliminating local

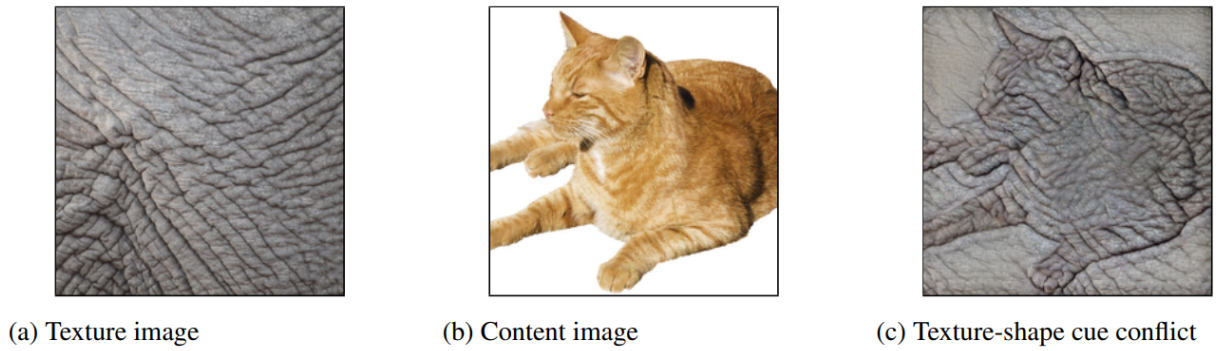


Figure 3.1: Cue conflict stimulus generated using Style Transfer



Figure 3.2: Visualisation of Stylized-ImageNet, created by R.Geirhos by applying style transfer to ImageNet images. After applying style transfer, local texture cues are no longer highly predictive of the target class, while the global shape tends to be retained.

texture cues.

However, the style transfer method is time- and computational resources-consuming and probably impossible for a young baby that does not have mature texture and shape representations. In this work we tested whether the same result can be achieved with a naturalistic trajectory imitating the gradual increase in acuity during infants' first months of age, and whether shape bias influences categorisation precision and robustness to noise.

Apart from performance, we also examined how the strength of shape bias influences the categorization process, e.g., by lowering representational differences between subcate-



gories of the same basic categories such as dog and cat breeds, preserving the difference between basic categories. We have also investigated whether the change in shape bias during training is more prominent for animate and inanimate categories, as it has been shown that animacy perceptual cues can shift an infant’s shape bias [Twomey and Westermann, 2017b].

## 3.2 Methods

### 3.2.1 Model

A supervised CNN architecture, ResNet-50 [He et al., 2015] was trained on the ImageNet dataset [Deng et al., 2009] for a total of 60 epochs, using a standard iterative training process. The training was implemented using the Python programming language version 3.7.7, utilizing the open-source deep learning library Pytorch version 1.7.1, with CUDA version 10.0.130. Batch normalization was implemented during the training process, with a batch size of 256. Prior to training, images were pre-processed by taking a random crop of size 224 by 224 pixels and randomly flipping horizontally. This pre-processing step is a common practice to reduce overfitting of the network’s weights to the training dataset [Srivastava et al., 2014]. Additionally, centering was performed by subtracting the mean value of each pixel across the training dataset from each pixel in a given image, this was performed across all color channels to obtain raw RGB pixel values [He et al., 2015]. Kernel size of 21 by 21 pixels in the first convolutional layer was used to fully utilize the effects of blur. Effect of smaller size on shape bias and receptive fields can be found in Appendix (A3.1).

### 3.2.2 Training set

The network was trained on the ImageNet dataset [Deng et al., 2009], a collection of 1.2 million images organised into one thousand labelled categories called synsets. Colour information was not removed from the ImageNet dataset. At early learning stages Gaussian blur was implemented in the data loader to imitate low initial acuity.

For testing precision the normal ImageNet validation set was used, and for evaluating shape bias a set of chimera stimuli was created (see section 3.2.5).

### **3.2.3 Experiment 1: Modelling poor visual acuity with Gaussian blur**

To model reduced visual acuity, a widely accepted method is to apply a Gaussian blur to an image, as it acts as a low-pass filter, selectively removing high spatial frequencies. To apply the Gaussian blur to an image, the matrix or kernel representing the Gaussian function is convolved over the image. In practice, a 9 by 9 pixel kernel is used, as the effect of increasing the kernel size further is negligible. The value of Sigma  $\sigma$  was set to 0, 2, 4, or 6 to create the desired level of blur for the training regimes. The Gaussian blur was applied to the image by using the Pytorch GaussianBlur function in the data loader. It's important to note that the relationship between the visual acuity and the sigma value is not exact as it would also require knowledge of image size and viewing distance, and therefore the conversion is not absolute.

### **3.2.4 Experiment 1: Training trajectories**

The ResNet-50 network was trained in a series of trajectories in order to establish a baseline and determine the optimal amount and length of removal of high spatial frequency information from the images. To establish a baseline, a ResNet-50 network was trained using only high resolution images. As a control, networks were also trained entirely on low resolution images that had been subjected to Gaussian blur with a value of sigma of 6, 4 or 2.

To investigate the effect of removing different amounts of high spatial frequency information, training was split into two phases of equal length, each lasting 30 epochs. The first phase, referred to as the pretraining phase, consisted of low resolution images that had been subjected to Gaussian blur with a value of sigma of 6, 4 or 2. The second phase, epochs 31-60, consisted of high resolution images. To emulate slowing of synaptic

connections growth learning rate was changed from 0.01 to 0.001 after epoch 30

The best performing training trajectory, as determined by the highest shape bias, was chosen based on the optimal amount and length of blur as determined from the previous analyses. This network is referred to as the **optimal network/trajectory**.

To investigate the effect of a gradual increase in the availability of high spatial frequency information, the network was trained with a progressively less blurry trajectory that involved dividing the 60 training epochs into 4 blocks of 15 epochs, which contained images with Gaussian blur applied with a value of sigma of 6, 4, 2, and 0, respectively.

To check whether the learning trajectory was important or only the presence of blurry images and to further separate the effects of learning rate the network was also trained on a 'scrambled' dataset comprised of presented blurred and original images from the best performing trajectory presented in a random order. In the rest of the chapter we will refer to it as 'random' set. Learning rate for these two alternative sets change remained the same as in previous experiments.

An average of 3 runs was computed for each trajectory to reduce stochastic variation between single runs.

### 3.2.5 Experiment 1: Test set

Networks were tested on the independent sample of ImageNet not included in the training set, called the validation set. There are 50 validation images corresponding to each of the 1000 ImageNet synsets, a total of 50,000 validation images. All of the validation images are high resolution, control networks trained on low resolution images alone were further tested on the validation set which had Gaussian blur with a value of sigma equivalent to the training images applied.

### 3.2.6 Experiment 1: Quantifying shape bias

As was outlined in the introduction, Geirhos and colleagues [Geirhos et al., 2019] developed a paradigm to quantify the tendency of a neural network to categorize an image

based on its overall shape or local details. This was accomplished through the creation of stimuli featuring incongruent texture and shape, derived from applying a texture image to the shape of another image. These stimuli are presented to a trained CNN and the proportion of correct answers that correspond to the shape of the stimulus is designated the shape bias of the network.

The hierarchical structure of WordNet was utilized to map ImageNet synsets to 16 broad categories, including airplane, bear, bicycle, bird, boat, bottle, car, cat, chair, clock, dog, elephant, keyboard, knife, oven, and truck. Cue conflict stimuli were generated using style transfer techniques [Gatys et al., 2016] by applying texture images to stimuli with a white background, leading to an incongruence between the texture and shape of the image, resulting in a cue conflict. Ten shape images from each of the 16 categories were placed on a white background and three texture images were sampled from each category to create five instances of each shape with a given texture, resulting in 80 texturized images for each shape, a total of 1280 stimuli. Stimuli with congruent texture and shape were eliminated, leaving 1200 stimuli.

The shape bias of a network was calculated by determining the predicted probability of each synset and summing the probabilities of each synset in each category, with the category with the highest cumulative probability being considered the network's choice. The change in shape bias from baseline was calculated by subtracting the shape bias of the network from a baseline, which was taken as the shape bias of the equivalent network trained on all high-resolution images. The percentage increase in the shape bias was calculated by dividing the change in shape bias from baseline by the baseline shape bias and multiplying the result by 100.

### **3.2.7 Experiment 2: Impact of category type on shape bias**

In human visual cortex, category representations are organized along two major cognitive dimensions of object representation: animacy and real-world size [Konkle and Caramazza, 2013]. It has been shown that real-world size drives differential responses only in the

inanimate object domain, not the animate domain, yielding a tripartite distinction in the space of object representation: specifically, cortical zones with distinct response preferences for big objects, all animals, and small objects. The preference zones are duplicated on the ventral and lateral surface of the brain. Such a duplication indicates the possibility of objective differences between categories that, together with constraints of the underlying cognitive architecture, create the distinction described above. These difference can potentially skew the shape bias during learning.

It has also been demonstrated that animacy perceptual cues can shift an infant's shape bias [Twomey and Westermann, 2017a, Davidson et al., 2017]: children demonstrate a strong use of the shape bias with categorization instructions (cue phrases such as 'same kind', 'goes with'), possibly viewing shape as another type of cue. The shape bias is also more robust in the inanimate condition, regardless of type of instruction.

In an effort to discern if a prospective shape bias is influenced by the type of stimulus, the sixteen categories were divided into three subgroups: animals, large inanimate objects, and small inanimate objects. The shape bias for each subgroup was calculated for both the baseline network and the optimal trajectory network, and the magnitude of the difference between the optimal and baseline trajectories, as well as the corresponding percentage change, were determined.

### **3.2.8 Experiment 3: Effect of shape bias on category separability**

Category acquisition can be impacted by the perceptual distance between objects belonging to the same or different categories. In some cases between-category distance can be smaller than in-category distance, leading to cases of confusion. For example, it has been shown that due to different amount of variation in breeds infants at some point of their development consider cats a subtype of dogs [Quinn et al., 1993]. Quinn and colleagues have demonstrated that the categorical representations for dogs were shown to be strongly influenced by the variability of breeds, as defined by the adult typicality

ratings. They have also shown that this effect can be corrected by modifying category variability, i.e. picking dog breeds with less variation or less typical cat breeds. French and colleagues [French et al., 2002] have also demonstrated that completely removing high spatial frequency with Gabor filters could increase the difference between objects from two different categories without altering the variability of data.

Just as Gabor filters, Gaussian blur can remove some of the details and shift the saliency of the features, therefore it is possible that category distance will change as well with different levels of shape bias. While infants do exhibit the cats/dogs confuion despite the preference for global processing, we do not expect significant categorization changes, however, as well-studied examples cat and dog breeds provide useful data for evaluating the effect of shape bias on categorization and perceptual distance between instances of basic and subordinate categories.

Basic categories is a term used to refer to clusters in data that are the most obvious for humans, they are the broadest and most inclusive categories in a particular categorization system. They are the quickest people name, are often characterized by clear and necessary features, and they have a low level of within-category variability [Rosch et al., 1976]. Subordinate categories, on the other hand, are more specific categories that are nested within basic categories. They are typically defined by more specific features and have a higher level of within-category variability. In our examples, "dog" would be a basic category while "Siberian huskey" and "Shiba Inu" would be considered subordinate categories.

To estimate that effect, we used the Oxford pet dataset. The dataset comprises 37 pet categories, each containing approximately 200 images, exhibiting substantial variations in scale, pose, and lighting. Every image is accompanied by ground truth annotations indicating breed. 12 breeds were randomly picked from each class, 5 images per breed, all images were reshaped to fit the network test format, colour information was preserved. Networks with different pre-training trajectories were then tested on these images, representational dissimilarity matrix built for features in the output layer of the network,

to check whether different dog breeds become closer together in representational space and between category groups (dogs vs cats) further apart after blurry training.

A Representational Dissimilarity Matrix (RDM) is a mathematical representation of the similarity or dissimilarity between neural representations of stimuli or mental states [Kriegeskorte, 2008]. It is a square matrix where each element represents the dissimilarity between two different neural representations and is usually used in the field of cognitive neuroscience and machine learning to study how neural representations are structured and how they relate to behavior and cognition. RDMs are commonly used to compare and contrast neural responses across different brain regions, conditions, and species, as well as comparing AI representations with neural data or different instances of AI models between themselves. They are also used to investigate the relationships between the representations and aspects of the stimuli.

### **3.2.9 Experiment 4: Robustness against distortions**

The robustness of convolutional neural networks has been identified as a critical limitation in their practical applicability [Szegedy et al., 2013]. As we mentioned in the introduction, Geirhos' style transfer approach [Geirhos et al., 2019] demonstrated this effect can be attributed to the reliance on local patterns rather than global shape and structure of the image, however, the style transfer method of training requires substantial computational resources. It is possible that a simpler method would suffice, as long as it can shift the network's bias towards global processing.

For models with different training trajectories, we evaluated how model accuracy degrades if images are distorted by uniform noise, contrast changes, high- and low-pass filtering after greyscale conversion (without preserving colour information).

Uniform noise is random noise that has a uniform probability distribution over a certain range of values. In other words, any value within the range is equally likely to occur. Depending on the amount of noise added and type of data uniform noise augmentation has been shown to produce varying effects on classification accuracy: normally it hinders

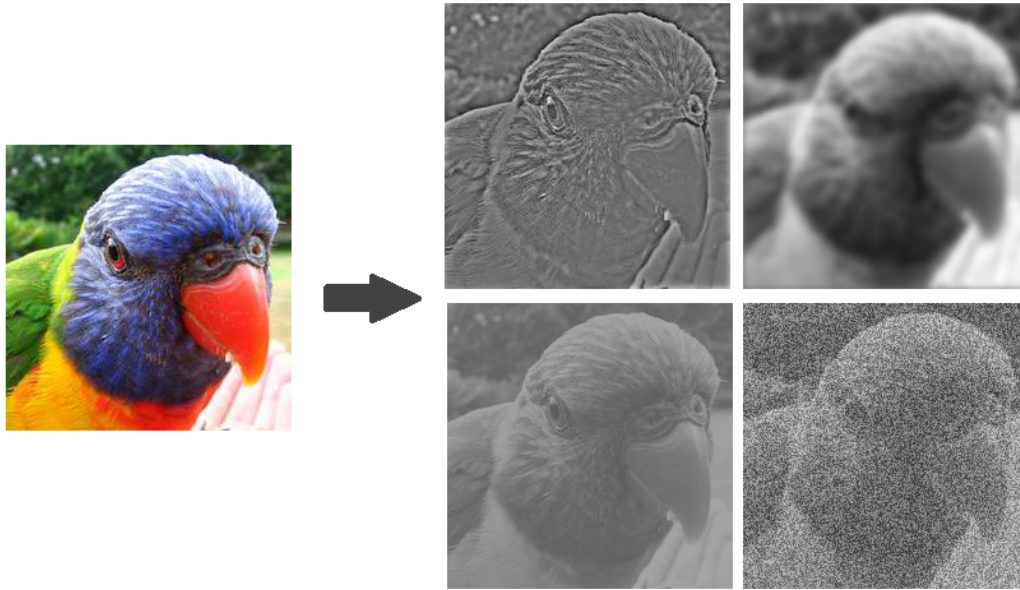


Figure 3.3: Examples of distortions applied.

performance [Lau Y., 2021], however in some cases of small well-controlled noise it can help to make the CNN more robust to other types of noise and variations in the input data, such as random changes in illumination or small translations.

Contrast change refers to a change in the difference between the darkest and lightest parts of an image. Adjusting contrast can affect border detection by the CNN, negatively impacting classification performance.

A high pass filter is a type of filter that allows high frequency signals to pass through while blocking or attenuating low frequency signals. The high pass filter can be used to remove the low frequency components of a signal, leaving only the high frequency components, which contain the important details and high-frequency information. In contrast to that, low pass filter is a type of filter that allows low frequency signals to pass through, removing the high frequency components.

To generate the noisy test set filters created by Robert Geirhos for his paper were applied to the ImageNet validation set (Figure 3.3). The implementation can be found at Github.

Performance for each training trajectory was then evaluated for each noise type separately, average performance of 3 different models of the same trajectory was calcu-



lated.

## 3.3 Results

### 3.3.1 Experiment 1: Effect of initial low acuity on shape bias

#### Baseline performance and shape bias

In order to establish a benchmark for the performance of the supervised ResNet-50 architecture, the top 1 and top 5 accuracy of the model were evaluated when trained on high resolution images from the ImageNet dataset. At epoch 60, the top 1 accuracy was found to be 70.754% and the top 5 accuracy was 90.250% (Figure 3.4).

At epoch 60, baseline shape bias was 20.50%, meaning that of the 795 of the 1200 cue conflict stimuli that were correctly recognised, 20.50% of correct choices were congruent with the stimulus shape while the remainder identified the image based on its texture.

For networks trained entirely on low resolution images, classification accuracy remained very low throughout training, with top 1 accuracy of less than 5%. This poor performance was anticipated as the validation set consists only of high resolution images. Networks trained entirely on low resolution images were additionally tested on a validation set which consisted of images that had been blurred with a Gaussian blur with a value of sigma corresponding to that of the training images of the network. Results can be found in Appendix 2.

#### Performance of networks with low-to-high acuity training

We found that the shape bias increased considerably following the pretraining on blurry images (Figure 3.5). It was most pronounced immediately after transitioning from low to high-resolution images, which progressively declined as the training progressed. Scrambling the dataset and presenting some of the blurred images later in the training process

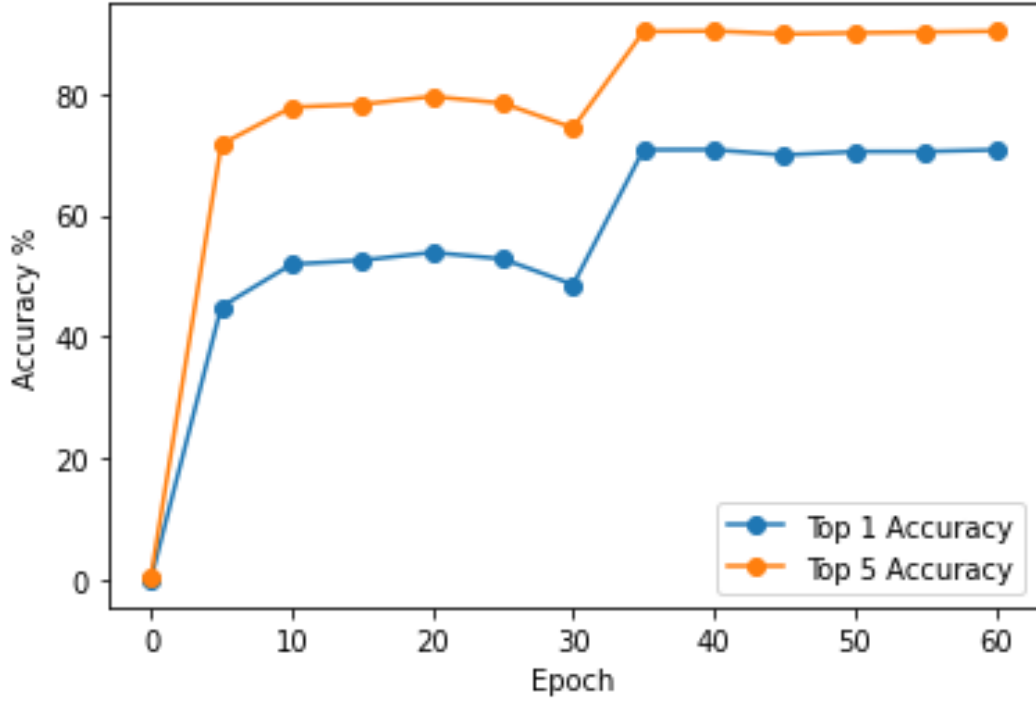


Figure 3.4: Training on unaltered ImageNet set.

also resulted in larger shape bias, indicating some level of forgetting being displayed by the network. Nevertheless, the change in shape bias remained significantly above the baseline, even after training on an equal amount of high-resolution images (Figure 3.6).

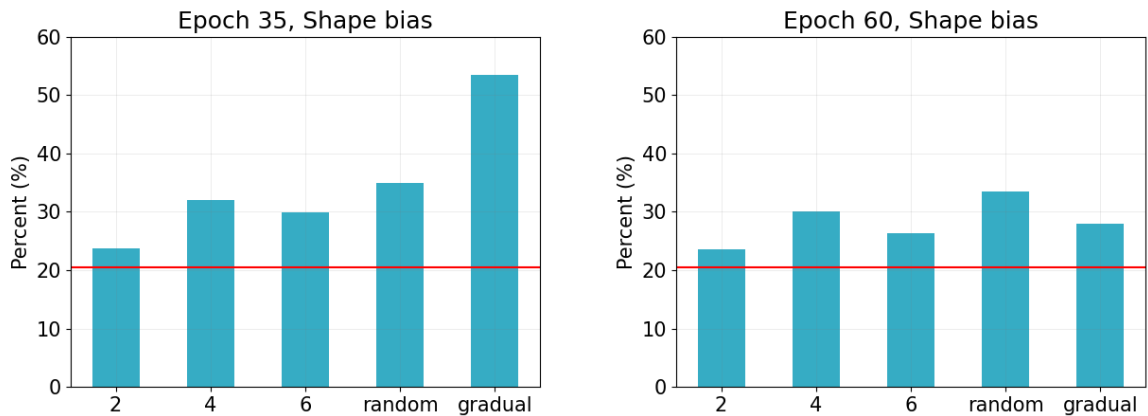


Figure 3.5: Change in shape bias with respect to training, average of 3 runs for trajectories. Red horizontal line denotes the baseline shape bias for a network trained on high resolution images only.

All networks pre-trained on low-resolution images exhibited minor decrease in top 1 accuracy compared to the baseline. The reduction ranged from a 3.314% decrease for

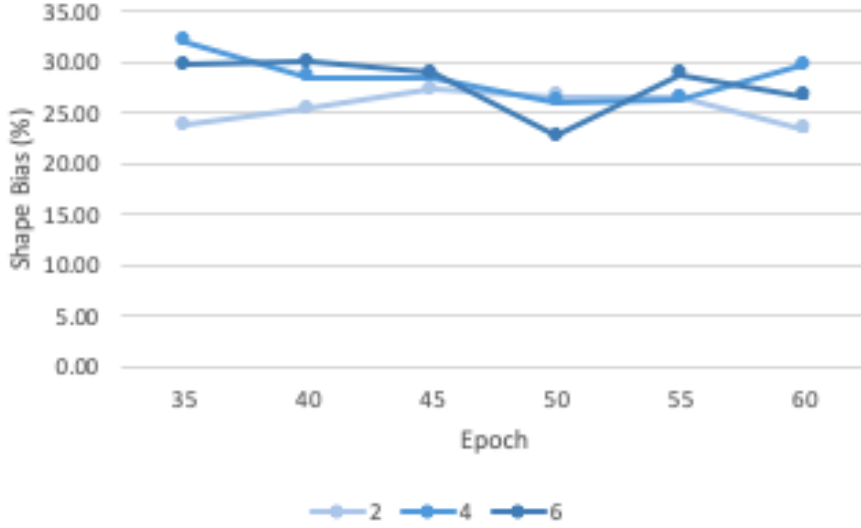


Figure 3.6: Change in shape bias with respect to training, individual runs for blur levels of  $\sigma$  2, 4 and 6. The highest final shape bias is observed for blur level with  $\sigma=4$ .

the network pre-trained on images blurred with sigma of 6 for 30 epochs, to a 1.783% decrease for the network pre-trained on images blurred with sigma of 2 for 30 epochs (Figure 3.7).

The biggest shape bias was observed at  $\sigma=4$  (Figure 3.5), however, the best performance was achieved at  $\sigma=2$ . It is possible that a filter of limited strength allowed both to induce shape bias and preserve features of intermediate size.

### 3.3.2 Experiment 2: Effect of category type on shape bias

The breakdown of shape bias into animate, inanimate large and inanimate small groupings for the high resolution network and the network which was pretrained on Gaussian blurred images with a  $\sigma$  of 4 for 30 epochs is presented in 3.8. The results indicate that for each of these networks, shape bias is larger for inanimate objects than animate objects. Within the inanimate category, the shape bias for inanimate small objects is always greater than that of inanimate large objects. This hierarchy is consistent across all networks tested, with inanimate large objects always demonstrating the strongest shape bias and animate objects always demonstrating the weakest. The increase in magnitude of shape bias observed when the network is pretrained on blurry images does not appear

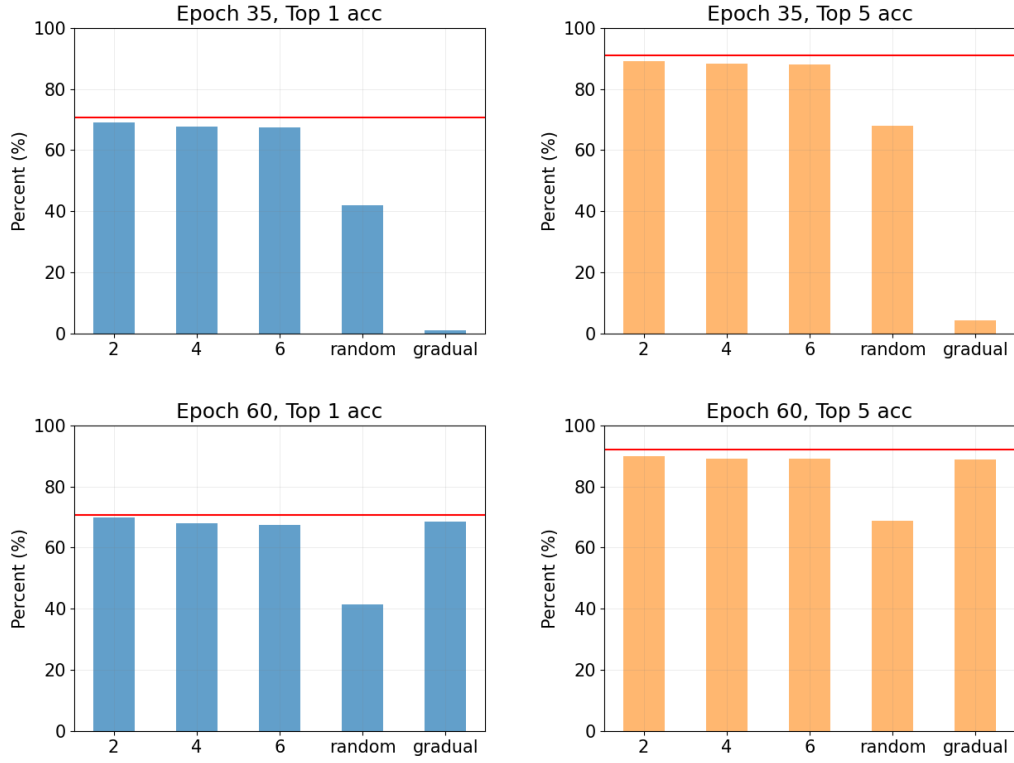


Figure 3.7: Change in shape bias with respect to training, average of 3 runs for trajectories. Red horizontal line denotes the baseline performance for a network trained on high resolution images only.

to be driven by an increase in shape bias in any one particular category, but rather an increase that is relatively evenly distributed across all groupings.

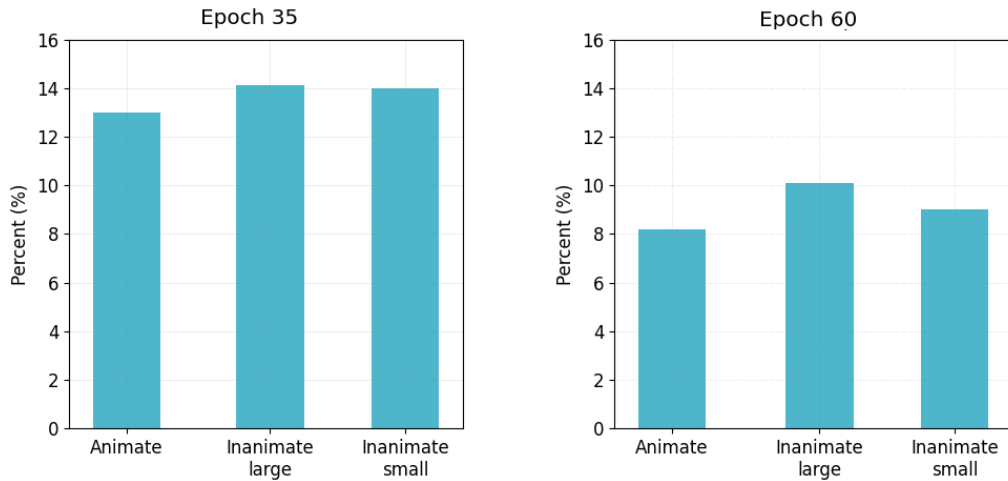


Figure 3.8: Difference in shape bias by category for optimal network of  $\sigma=4$  (difference from baseline). The magnitude of shape biases increased relatively evenly across categories, preserving the original distribution.

### 3.3.3 Experiment 3: Effect of shape bias on category separability

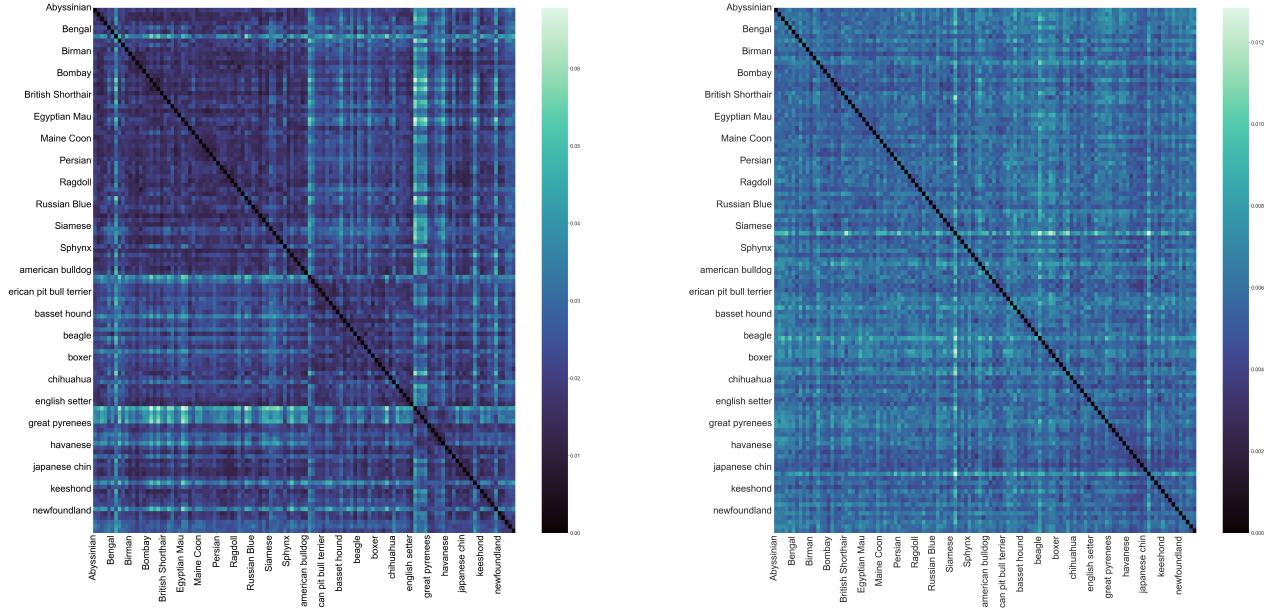


Figure 3.9: Representational Dissimilarity Matrices for cats and dogs for instances of networks pretrained with  $\sigma=0$  and  $\sigma=4$ .

Increasing shape bias by reducing the range of the spatial frequencies available to network during training decreases the variance distribution between all the categories (Figure 3.9), both subordinate and basic ones. As ResNet-50 is normally trained on more epochs (300 epochs on average while in these experiments we only used 60), it is possible that the current amount of training was insufficient for representations to converge, therefore these results are inconclusive.

An alternative hypothesis is that initiating the training process with blurred images may not yield the desired outcome with more labels consisting of subordinate categories instead of basic ones, as employed in the previous related study [French et al., 2002]. Another study has demonstrated that training on superordinate categories [Ahn et al., 2021] enhances the shape bias of networks trained on high-resolution images. Consequently, it is plausible that both basic category labels and low-resolution data are essential for achieving the intended effect.

In this case, training on a set with simpler, less detailed labels would have produced a different effect.

### 3.3.4 Experiment 4: Robustness to noise

Learning with low initial acuity proved beneficial for certain types of noise 3.10, however, the benefits were not in direct correlation with the shape bias itself.

While trajectory with  $\sigma = 2$  levels of blur provided the smallest increase shape bias among the attempted training regimes, it resulted in improved performance across all types of noise with the exception of high pass filter. It's followed by the optimal trajectory of  $\sigma = 4$  and gradual transition from  $\sigma = 6$ .

All trajectories with initial low acuity proved most useful in filtering out uniform noise and discerning shape in low contrast images in comparison to training only on a high-resolution dataset. Performance on the set with high pass filter applied (leaving only the fine details for the network to rely on) dropped significantly, the opposite being observed for the low pass filter.

For the optimal trajectory of  $\sigma = 4$  images presenting blurry and non-blurry images in random order during training led to both better performance and the final network being more resistant to noise.

Irregardless of the severity of noise, performance of networks with a larger shape bias remained higher for uniform noise, low-contrast and low-pass images than of the baseline network trained on pure ImageNet (Figure 3.10).

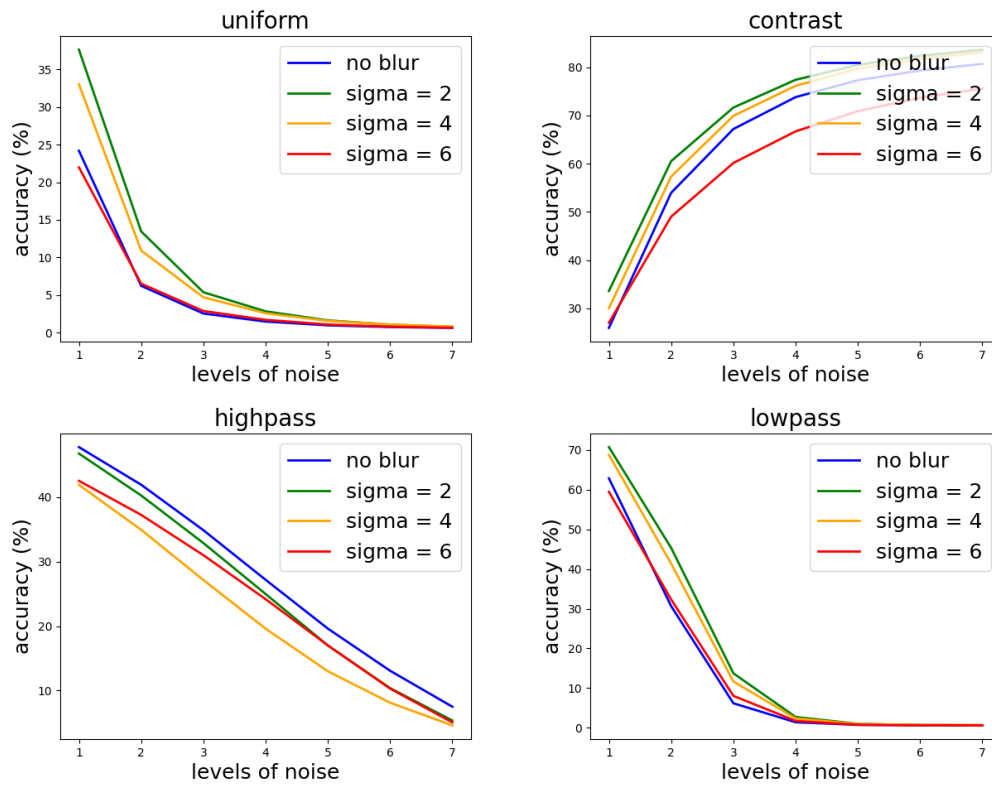


Figure 3.10: Classification accuracy on parametrically distorted images with regard to the amount of initial blur and the level of distortion. ResNet-50 trained with an initial blurry start is more robust towards distortions than the same network trained on ImageNet.

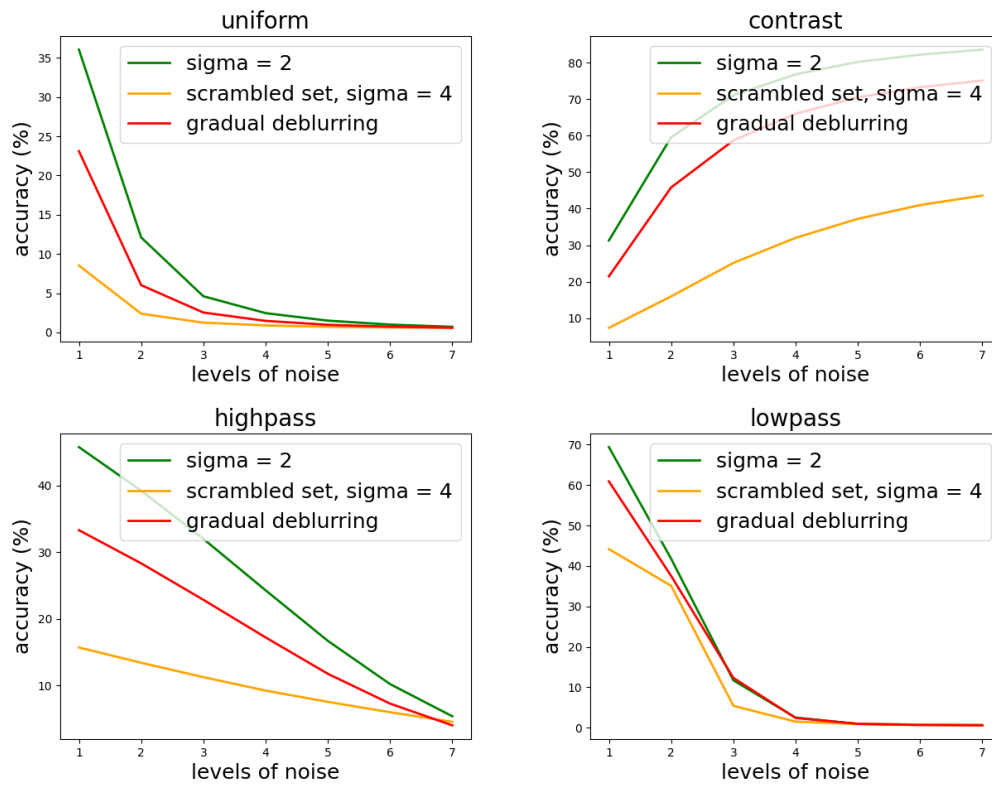


Figure 3.11: Comparing the highest-performing trajectory for no blur, initial blur ( $\sigma=2$ ) with gradual deblurring and random order blur ( $\sigma=4$ ). Low levels of blur turned out to be the most beneficial among all training trajectories.



### 3.4 Discussion

Our research investigated the influence of pretraining on low-resolution images on the development of artificial neural network's visual processing systems. Our findings suggest that pretraining on low-resolution images enhances the network's sensitivity to global image shape. This suggests that visual acuity limitations during early development may significantly contribute to the expansion of spatial processing in the ventral stream.

An increased shape bias was observed following pretraining on blurry images, with the highest increase immediately following the switch from low to high resolution images with a progressive decline as training progressed. Scrambling the dataset and presenting some of the blurred images later along the training process also produced larger shape bias, indicating some level of forgetting. However, the change in shape bias remained significantly above the base even after training on equal amount of sharp images.

That this one simple biologically inspired change to the training procedure was sufficient to increase the shape bias of the network provides support for the theory that the bias towards using local features to classify an image that is frequently observed in CNNs is not strictly determined by the architecture of a network but rather is facilitated by the nature of training.

Our findings also revealed variations in shape bias across different categories. Both for networks trained on high-resolution images and networks pretrained on low-resolution images, shape bias was higher for inanimate objects than animate objects, and higher for inanimate large objects than inanimate small objects. This is in line with behavioral evidence that, despite generally demonstrating a shape bias, children are more likely to label a perceived animate object based on texture than shape, whereas for perceived inanimate objects, they are more likely to label them based on shape [Twomey and Westermann, 2017b].

Robustness to noise was found to be not directly correlated with shape bias. While trajectory with  $\sigma = 2$  levels of blur provided the smallest increase shape bias among

the attempted training regimes, it resulted in the most improvements across most types of noise, the effect of training trajectories that induced larger shape bias was not as prominent.

In addition to that, compared to style transfer images, improvements to robustness to noise was not substantial and no improvement in performance was observed. This result is consistent with an earlier study [Mummadi et al., 2021], in which contour images were used to train the network. Higher shape bias also did not always correspond to higher robustness to noise. This suggests that more factors and types of visual variation which were present in the original Geirhos' experiment and absent in subsequent works that are required to facilitate human-like robustness to noise. Additionally, due to differences in training procedures it cannot be directly compared to the style transfer method and it is possible that for some types of noise (uniform, low pass) it may achieve similar performance. Further experiments with variations in colour, shape distortion and a variety of learning regimes are required.

In a different study employing low-pass filter to simulate initial low acuity [Jinsi et al., 2023] it has been shown that such an approach can even increase network performance on the original set, but only when blur decreases linearly over time. Some of the results, including overall drop in performance, lower performance for higher levels of blur could be attributed to the non-linear nature of chosen deblurring strategy, this suggests the network could benefit from a more complex regime of training. Additionally, it offers an explanation for the significantly impaired accuracy for the training trajectory where blurred images were presented randomly.

To conclude, the findings indicate that shape bias alone does not account for the performance improvements and should instead be regarded as a secondary effect of the more robust representations developed by the network.

## 4 Curriculum learning

We've acknowledged the value of a curriculum in specific scenarios, but it's also crucial to determine when it becomes an absolute necessity. Humans often compensate for the lack of a formal curricula by engaging in self-driven curious learning. As discussed in the introduction, this approach has resulted in enhanced learning outcomes in many computational models as well. But it is not a one-size-fits-all solution for all learning challenges.

To develop effective human-centered curricula or scaffolding approaches, as well as more advanced artificial models, it is important to delve into the potential properties of tasks that cannot be learned without a fixed curriculum. Let's take a closer look at an example of such a task to better understand the specific conditions that require structured learning.

### 4.1 Introduction

Human education is a refined enterprise. Those around a child adapt their behaviour in many ways to help the child's learning. In preschool and school, teachers follow carefully structured curricula, designed to introduce foundational concepts before more advanced ones. Pedagogy is a science in itself and we dedicate no small amount of effort to finding out how to teach children in an optimal way. Some studies suggest that even slow biological maturation of infants' cognitive abilities leads to better learning [Elman, 1993, Gopnik et al., 2017, French et al., 2002, ?].

Curriculum learning experiments, starting from Elman’s classic work [Elman, 1993], have found that artificial neural networks (ANNs) can also benefit greatly from elaborate approaches to training. Starting with more representative examples, and gradually increasing in diversity, improved convergence speed and helped the network to focus on relevant features while it was forming representations [Bengio et al., 2009]. Using auxiliary tasks has proven to be effective in improving generalization across tasks and robustness against various types of noise [Zeiler and Fergus, 2014, Mordan et al., 2018, Zhang et al., 2014]. Recent experiments with ANN training datasets have also demonstrated that some ANN limitations that are considered an inherent property of ANN architectures, such as, for example, a bias towards local processing, can be removed by adjusting the dataset [Geirhos et al., 2019].

When training a network to perform multiple tasks a correctly chosen curriculum has been shown to push the network towards either more specialized or more flexible representations without having to modify the architecture [Yang et al., 2019], affecting the emergence of alternative neural mechanisms and allowing the researchers to choose between higher performance for learned tasks at the costs of flexibility or a better performance over a wide range of tasks at the expense of a lower performance for the learned tasks.

Decomposing a complex task into a sequence of source tasks with gradually increasing complexity has been shown to greatly improve performance for certain types of reinforcement learning [Manela and Biess, 2022].

Curriculum learning has also been proven especially useful when dealing with continual learning. Modern machine learning have shown exceptional results when trained on fixed datasets and stationary environments but these models often fail to emulate the robustness and efficiency of human learning in a non-stationary world [McCloskey and Cohen, 1989, Ratcliff, 1990]. One of the most known cases is catastrophic forgetting — the rapid performance degradation on earlier learned tasks that occurs when dealing with highly non-stationary data — but ANN models more broadly under-perform when

presented with changing or incremental data regimes [Hadsell et al., 2020].

A solution proposed for this problem is using a curriculum, including using auxiliary tasks [Javed and White, 2019], pre-training [McRae, 1993, Mehta et al., 2021] and active learning [Parisi et al., 2019, Shi et al., 2020].

The idea of active learning (also known as curiosity-driven learning and automated curriculum learning) initially stemmed from addressing a problem that using curricula introduces: designing an effective curriculum can be hard as there are many free parameters. It therefore requires time and effort to create and it may not generalise to other training sets. Instead of manually creating curricula, an attractive alternative is that the network, by monitoring its learning, could automatically choose its own curriculum — automated curriculum learning. This approach is appealing as it makes design simpler and could allow networks to adapt to new datasets and continuous streams of data.

Graves et al [Graves et al., 2017] showed that automated curriculum learning can indeed significantly accelerate progress and increase accuracy by proceeding from easier to harder tasks. However, the proposed method requires a suitable progress metric used to indicate when to switch tasks. The best metric varied for different sets of tasks, which raises the question under which conditions actively choosing a curriculum for learning is possible, and whether it is achievable to identify the best progress signal in advance, without retraining the network for every potential metric. This is a particular concern as in a related field, curiosity-driven algorithms have been shown to encounter pathological cases, such as becoming fixated on learning from completely unpredictable stochastic data [Burda et al., 2018] and in developmental robotics they can fail when the need to switch between a variety of tasks arises [Oudeyer and Kaplan, 2009].

The literature on human learning is also relevant. In education there has long existed a conflict between curiosity-driven learning and fixed curriculum learning [Bonawitz et al., 2011]. Curiosity has been demonstrated to enhance learning, both by improving memory [Gruber et al., 2014, Gruber and Ranganath, 2019] and by helping students to learn important background information by deviating from a fixed curriculum [Malcom, 2019].

Children and infants are drawn to things that have just the right "Goldilocks" level of novelty [Berlyne, 1954, Kidd et al., 2012], which has been shown to be a strategy of creating an optimized learning curriculum that is also applicable in machine learning [Twomey and Westermann, 2017a]. External incentives have been shown to suppress curiosity-driven intrinsic motivation in some cases and harm learning [Jovanovic and Matejevic, 2014]. However, at the same time, in human learning, formal schooling, which to a large extent uses a fixed curriculum is highly effective, increasing the intelligence of learners by 1-5 IQ points per year [Ritchie and Tucker-Drob, 2018]. When dealing with difficult tasks humans often need at least some degree of guidance, especially in infancy, when their knowledge of the world is limited. Machines are not the only ones who struggle with "pure" active learning.

The need for guidance is further supported by the Vygotskian concept of scaffolding [Wood et al., 1976]. By providing support in tasks that are just above the student's skill level and breaking tasks into small manageable chunks a competent teacher can allow them to learn content they would not have been able to process on their own [Bransford et al., 2000]. Our hypothesis is that ANNs may require very much the same treatment.

The goal of this paper is to investigate the conditions under which active learning is possible and the optimal metrics to guide it. To address this, we created a network and a task for which curriculum learning was important, evaluated a variety of training regimes and measured whether an optimal regime could be discovered by the network itself.

## 4.2 Methods

### 4.2.1 Data and tasks

Our goal was to design a task set that would benefit from training with a curriculum, but that was sufficiently simple to allow exploration of final performance across many

training scenarios. The curriculum of our network consisted of two tasks. Task 1 was a simple supporting task of recognizing single-digit numbers. There was then one of two more complex main tasks. Task 2a required the network to discriminate whether the sum of two digits in the picture was odd or even. Task 2b (discussed further in section 2.3.2) required the network to calculate the sum of two digits in the picture.

Task 2a could be conceptualised as two tasks that we would expect the network to be able to learn in isolation: an odd/even classification for each of the digits (which should be of a similar difficulty to digit recognition and would require the network to create single digit representations); and XOR to resolve the digit sum - *even* if both single digits are either odd or even, and *odd* if one is odd and the other even. While the network should be able to learn each of these tasks in isolation, when they are presented together, we predicted the network would fail to learn. Each digit is equally distributed among the odd-even digit-sum classes and so the error signal would be orthogonal to the presence of any given digit. We therefore predicted the network would be unable to learn the single digits (or the initial odd-even classification). Put another way, during training the weight changes each batch will cancel each other out, which would make it difficult or potentially even impossible to learn the image-to-single digit mapping required for good task performance. We therefore predicted that pre-training on Task 1 would be a valuable stage towards optimizing performance on Task 2a, as learning the digit representations beforehand eliminates this problem.

More generally, we hypothesized that such hierarchical learning scenarios will be encountered when dealing with tasks that require higher-level abstract information (such as learning language or mathematical equations) or where sub-goals seemingly require moving further away from the final goal (the Tower of Hanoi game). However, some abstract tasks may not have this property. To investigate this, we explored a variant, Task 2b, where the network had to learn the sum of digits in the picture instead of whether that sum was odd or even. Our prediction was that, although this is also a multi-digit task building upon single-digit representations with an even greater number of potential answers, it would be more easily learnable as the error signal for the multi-digit task was

not orthogonal to the single-digit classes. We hypothesized that the Task 2b error signal will thus be able guide effective learning of the single digits.

We evaluated a pre-training curriculum, where Task 1 always preceded Task 2, with varying degrees of training on each.

The training datasets were based on the MNIST dataset [LeCun et al., 2010]: a set of 60000 28x28 greyscale images of handwritten digits, centered and normalized. In both tasks, the network was trained in a supervised way (Fig. 4.1). For the “simple” Task 1, digits were scattered on a white background 84x84 squares to ensure learning spatially invariant representations (Note the images were padded and not resized, and so scale invariant training was not needed). The network was required to recognize single digits. For Task 2, we used MNIST dataset to create 55 classes of double digit stimuli, comprising every combination of two unordered digits. The digits were again scattered on the 84x84 white background and were constrained so that they never overlapped. Both sets consisted of 30000 examples each. The test set consisted of 10000 double-digit images labeled for performing an odd-even task.

## 4.2.2 Model

To best explore the benefits of curricula, we used a simple ANN, with two convolutional layers (kernel size = 3, ReLU activation, average pooling) followed by a densely connected layer of 10 nodes, and two dense layers of 2 nodes (Fig. 4.1, centre column). During Task 1, the loss function was calculated from the first fully connected layer with 10 nodes. During Task 2 the loss function was calculated at the final layer. This simple network did not perform at ceiling on the target task, allowing us to distinguish the relative strengths of different curricula. There are architectures that could allow better performance in Task 2a/b, irrespective of the training regime. However, to provide sensitivity to the training manipulations, it was important that the network did not perform at ceiling. We therefore intentionally chose network architecture and hyperparameters that did not perform at ceiling on the task (see S1 Fig.1, in the Supporting information section).



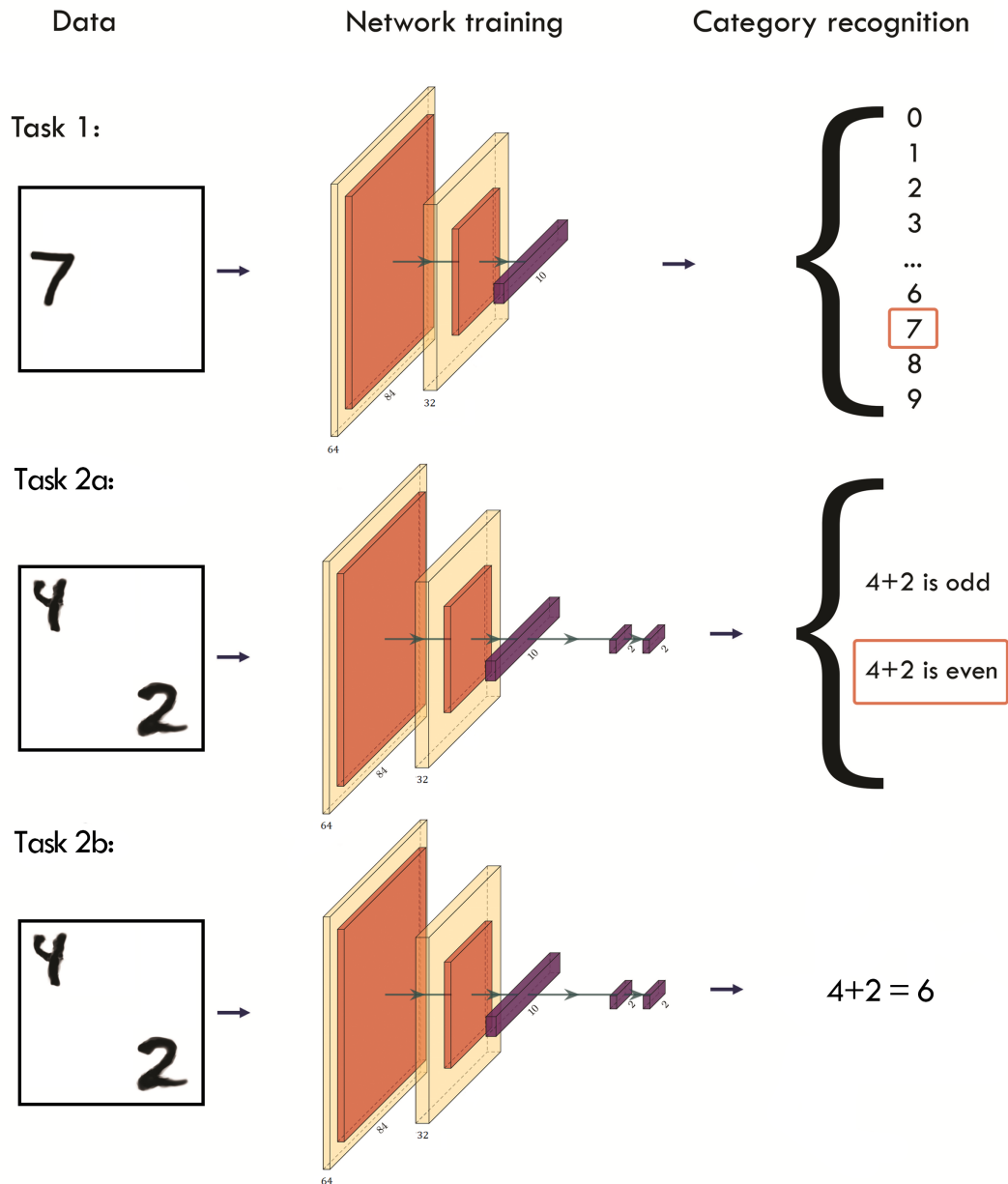


Figure 4.1: **Architecture and tasks** Task 1: Single digit recognition task. Task 2a: Sum of digit odd/even recognition task. Task 2b: Sum calculation task. In all tasks digits were scattered on a white background to ensure learning spatially invariant representations.

To investigate the generalisation of switching metric performance across networks, we also tested a network variant that was similar, except that it had three rather than two convolutional layers.

The models were implemented using python 3.5 and Tensorflow 2.0, and ran on an Amazon Web Services g3s.xlarge machine (4 virtual CPU, 30.5 GiB RAM, Tesla M60 GPU). The code can be found at [https://github.com/ankravchenko/curriculum\\_learning](https://github.com/ankravchenko/curriculum_learning).

If a metric for curriculum learning is going to be useful, it has to be informative within a single instance of training a network. We therefore evaluated instance-to-instance variation, by training each model 8 times with random starting weights, to obtain a mean and standard deviation of performance accuracy and switching metrics. For calculating these values and further statistical analysis the Python SciPy library was used (version 1.7).

### 4.2.3 Training procedure

#### Comparing static and continual learning

Both the Task 1 and Task 2 stimuli were constructed from the 30000 MNIST digits, with 3000 examples for each of the digits 0-9 and for 250 writers [LeCun et al., 2010].

For static pre-training, a subset of  $n * 1250$  digits was sampled without replacement from the 30000 digits, where  $n = [0.24]$  determined the quantity of pre-training. In each of three epochs, the entire subset was presented in random order in batches of 32.

For continual pre-training,  $n$  non-overlapping subsets of 1250 exemplars were sampled without replacement from the pool of 30000 MNIST digits. These subsets will have on average sampled the digits and the writers in a balanced way, but differed in their specific exemplars. Three epochs of training were conducted on the first subset, then three epochs on the second, and so on, until  $n$  subsets had been used. This "exemplar incremental" design reflects a weaker form of non-stationarity than is often present in the environment, but it still affected network learning substantially.

Task 2 training then followed. For comparability, the same procedure was used for static and continual learning. 24 non-overlapping subsets of 1250 exemplars were sampled without replacement from a pool of 30000 images. Three epochs of training were conducted on the first subset, then three epochs on the second, and so on, until  $n$  subsets had been used.

At no point in the training process were the weights frozen.

### Testing the orthogonality hypothesis

To test our prediction about the effect of the orthogonality of the error signal, we also included a variation of the second task, Task 2b. The image stimuli of double digits remained the same, the labels, however, were changed to a sum of digits present on the picture.

To ensure that the labels were not orthogonal to the component digits, we provided the network with a "magnitude style" label encoding:

$$l_i = \begin{cases} 1, & \text{if } i \leq X \\ 0, & \text{if } i > X \end{cases}$$

where  $l_i = [120]$  is the output of neuron  $i$ , and  $X$  is the sum of the two digits.

As the labels were no longer "hot one", the softmax layer was removed and the loss function was altered from classification accuracy to the mean absolute error.

Since we were only interested in learnability, for Task 2 the training procedure was also simplified. We used static training and investigated three different amounts of pre training: for 30000, 15000 and 0 images of Task 1, with training on the full set of Task 2 after that.

#### 4.2.4 Metrics to predict an optimised switching moment

We then investigated under what conditions the network itself could determine the optimal moment for switching between tasks using only the data available during a single training run. Here we define the optimal switching moment as the amount of pre-training on Task 1 needed for the network’s mean final performance on Task 2 (averaged across 8 runs) to reach the top decile of the accuracy observed over all regimes of training.

In the existing literature a variety of metrics have been used to help the learner to determine when to switch between task stages. Oudeyer [Oudeyer and Kaplan, 2009] called them adaptive intrinsic measures and categorised them into competence-based, predictive and information-theoretic types. Graves [Graves et al., 2017] added complexity gain measures (effective information encoded in the network weights) to the list. Based on this, we considered four kinds of promising metrics:

- **Performance.** A popular approach in curriculum learning is to define a performance threshold on the pre-training task for advancement to the next task [Zaremba and Sutskever, 2014]. We therefore investigated if there was a threshold, consistent across scenarios, to define the optimal switching level.
- **Competence.** A typical competence measure would be to evaluate initial Task 2 performance (performance after the first training set slice of 1250 examples on Task 2 after pre-training on Task 1) after a given amount of pre-training.
- **Prediction gain.** In the literature on human learning, intrinsic motivation is viewed as a system that generates rewards when predictions improve over time. Such a system will try to maximize learning/prediction progress, i.e., to decrease of prediction errors. To capture this, the gradients of performance on Task 1 and Task 2 were captured by the change in performance across a step of 1250 trials as measures of learning progress to be used for determining the switch time.
- **Information gain.** Information theory has also been used in automated curriculum learning. Building on the work of [Achille et al., 2019], we hypothesized

that the peak in Fisher Information (FI) of weights, which corresponds to a phase transition in learning, might be an effective signal that could guide the switch in learning. We evaluated the Fisher Information for the weights calculated with the loss functions from both Task 1 and Task 2. Both the FI peak and the post-peak drop could be viable points of switching, therefore to determine the optimal point we calculated Pearson correlation coefficient between FI value at the moment of switching and the final performance on Task 2 at each  $n \times 1250$  examples step of pre-training on Task 1 using information from 8 instances of our network. For each 1250-example step of pre-training on Task 1 we also compared FI for the possible next steps: another 1250-example slice of training on Task 1 and first 1250-example slice of Task 2, as greater information gain from a different task could also indicate a viable switching point.

#### 4.2.5 Examining task representations

To investigate how representations developed in the network during pre-training phase, we used t-SNE analysis to visualise activations in the first dense layer of the network at various stages of pre-training on Task 1 in response to selected diagnostic classes, using 100 examples of each class. The single digits 1, 3 and 8, and their double-digit combinations (11, 13, etc) were used as diagnostic classes, selected on the basis of 3's similarity to 8 and dissimilarity of 1 to both of the former.

T-SNE (t-Distributed Stochastic Neighbor Embedding) is a powerful non-linear dimensionality reduction technique widely used for clustering analysis in machine learning and data science. It allows to visualize high-dimensional data by mapping it into a low-dimensional space while preserving its local structures. The t-SNE algorithm starts by defining a probability distribution over pairs of high-dimensional objects that is then transformed into a probability distribution over pairs of low-dimensional points. The algorithm then minimizes the divergence between these two distributions using a gradient descent method.

Implementation of t-SNE from sklearn python library was applied using 2 dimension components, 300 iterations, with perplexity to 40 and verbosity level set to 1. No other dimensionality reduction methods were used before feeding the data to t-SNE.

We also displayed activations for ten random examples of each image class to explore individual variation in embeddings.

## 4.3 Results

### 4.3.1 Testing a range of curricula

Our first goal was to investigate if the network benefited from training on Task 1 first. Both for continual and static learning, final performance on Task 2 depended on the amount of pre-training on Task 1 (Fig. 4.2), with Task 2 unlearnable without some initial training on Task 1. This shows that this task structure was effective in requiring a curriculum. For continual learning, too much pre-training could also be detrimental (Fig. 4.2a, Fig 4.3). Thus, the network benefited from having a curriculum, and the amount of training on Task 1 was important.

It is worth noting that simply splitting the curriculum in two tasks without learning them in the right order did not solve the problem. Learning the XOR task first was not beneficial either: the performance was still no better than random (Fig. 4.4).

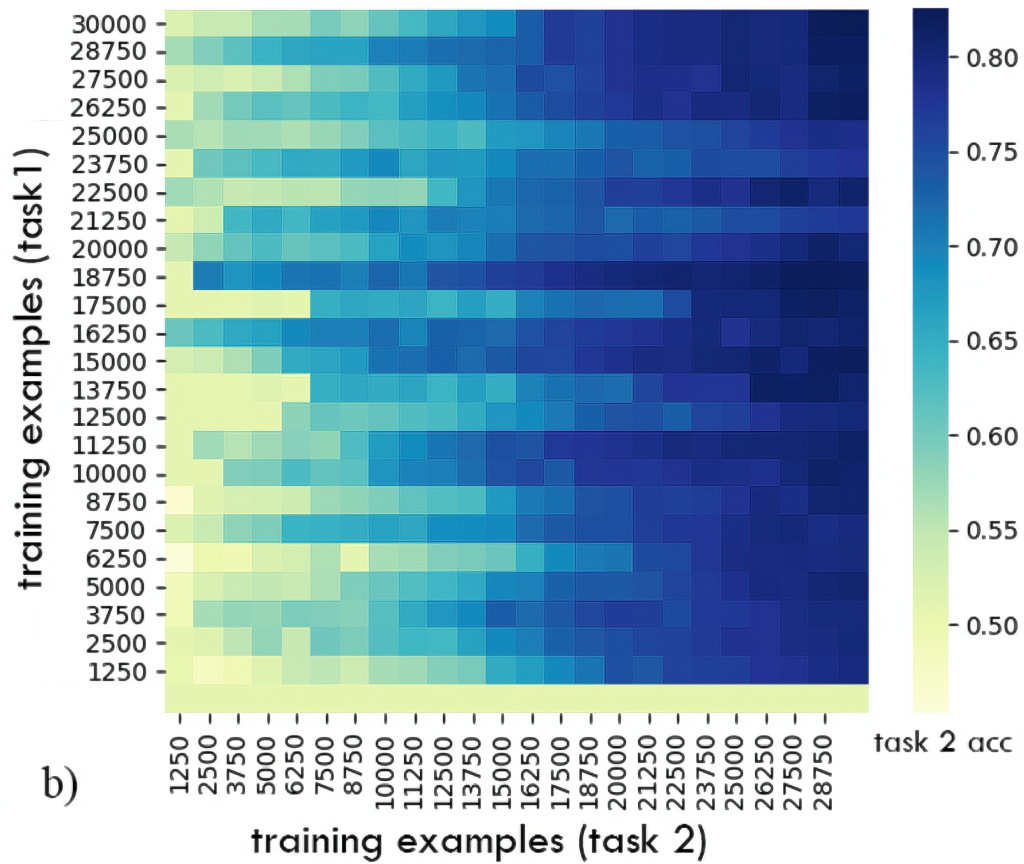
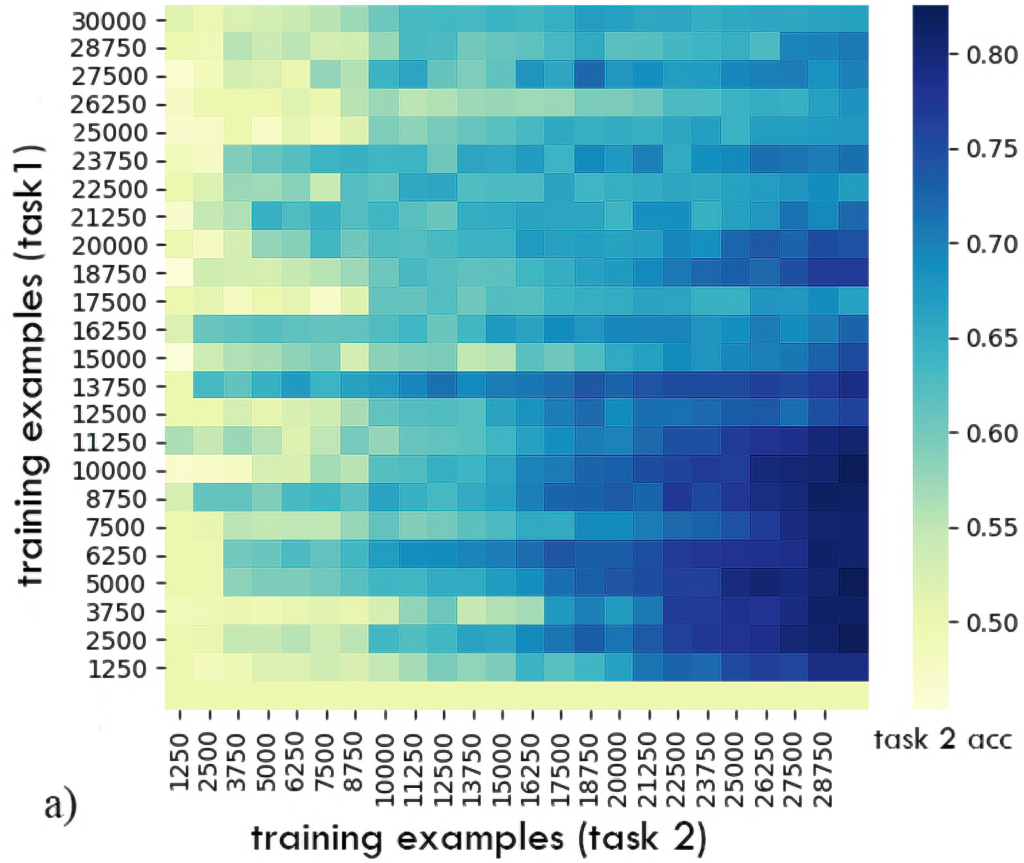


Figure 4.2: **Task 2 performance for continual (a) and static learning (b) of Task 1.** The average of 8 runs is shown. With no Task 1 pre-training (bottom row of either panel), Task 2 performance was at chance irrespective of the quantity of Task 2 training. This confirms the task structure was effective in requiring a curriculum. With continual learning, too much Task 1 pre-training was also detrimental to final Task 2 performance. This could be due to overfitting and accumulation of error when dealing with a sequence of tasks.

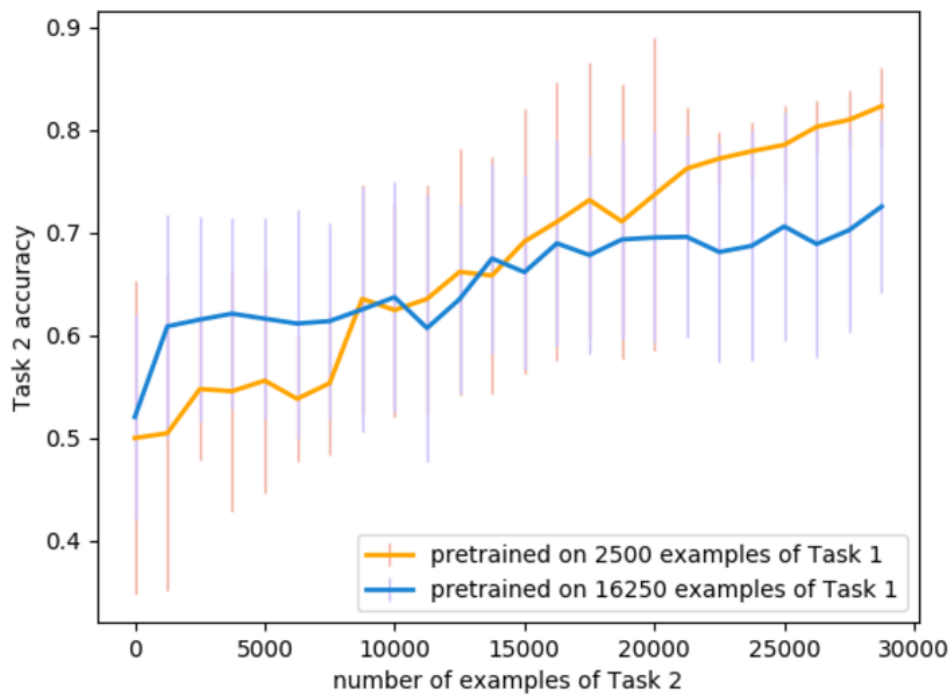


Figure 4.3: **Comparison of learning trajectories for Task 2 after different amount of pre-training.** a) During continual learning, in single runs the effect of too much pre-training on Task 1 on Task 2 performance did not emerge until later in the Task 2 training.



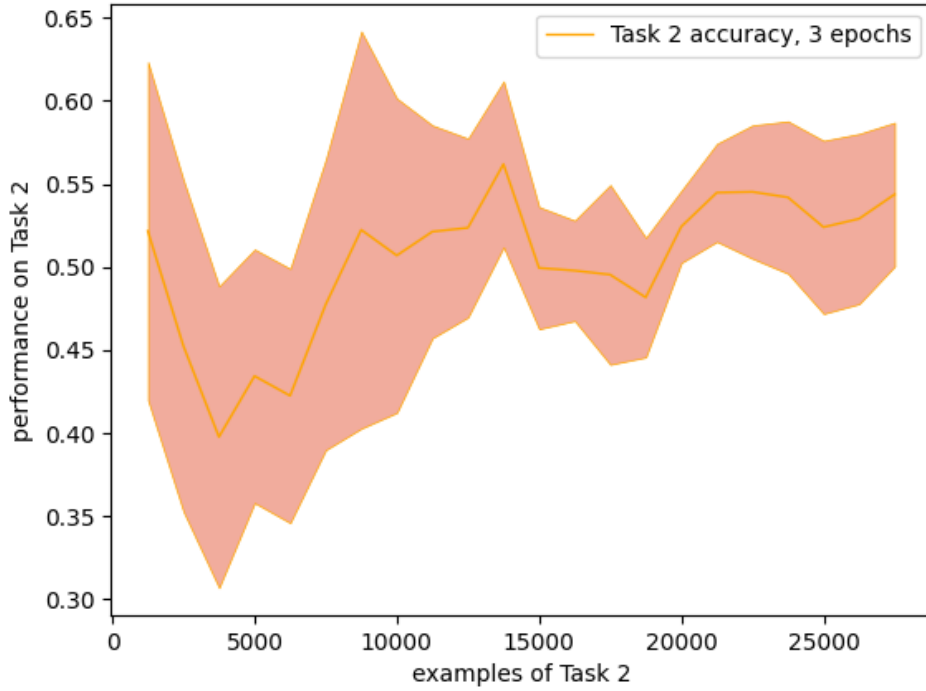


Figure 4.4: **Learning Task 2 first followed by Task 1.** Learning the XOR task first did not lead to an improvement in performance.

### 4.3.2 Metrics to predict the optimised switching moment

To optimise the curriculum, we evaluated four types of metric that could be used by the network for task-switching. We discuss each in turn.

- **Performance.** When tested on the original 2 convolutional-layer network with a standard learning rate performance on Task 1 seemed a possible indicator of the optimal switching moment during continual learning (Fig. 4.5a, orange). To test if the critical threshold of performance generalised across scenarios, we repeated training and analysis for a variant on this ANN, with three instead of two convolutional layers. The Task 1 performance level that signalled the optimal switching moment was not consistent across these two networks, ruling out this as an effective metric (Fig. 4.5). The same was observed during stationary learning: as can be seen at Fig. 4.7, the red line marking to the optimal switching moment does not correspond to any visual cues or irregularities in any of the metrics.

- **Competence.** Initial competence on Task 2 showed no consistent change as a function of the quantity of Task 1 training (Fig. 4.5b, blue), and so this will not be useful as a metric. Furthermore, even after a little training, initial competence on Task 2 was not informative, because of pre-training sometimes yielding delayed benefits that were not apparent until considerable Task 2 training (Fig. 4.3).
- **Prediction gain.** Prediction gain metrics did not provide a possible signal for switching in our tasks either. While at some points of training Task 2 "learning progress" exceeded Task 1 learning progress, it was indistinguishable from noise (Fig. 4.5 1b). Even if we were to attempt to use learning progress as a possible measure, as it showed substantial run-to-run variation and in many cases it would switch too early.
- **Information gain.** For Task 2 there was also no statistically significant difference between Fisher information (FI) of weights at optimal and suboptimal points of switching (Table 4.1). This can be explained by the requirement to learn higher-level features to perform Task 2, and is in line with [Achille et al., 2019] finding that high-level features are not affected by critical periods.

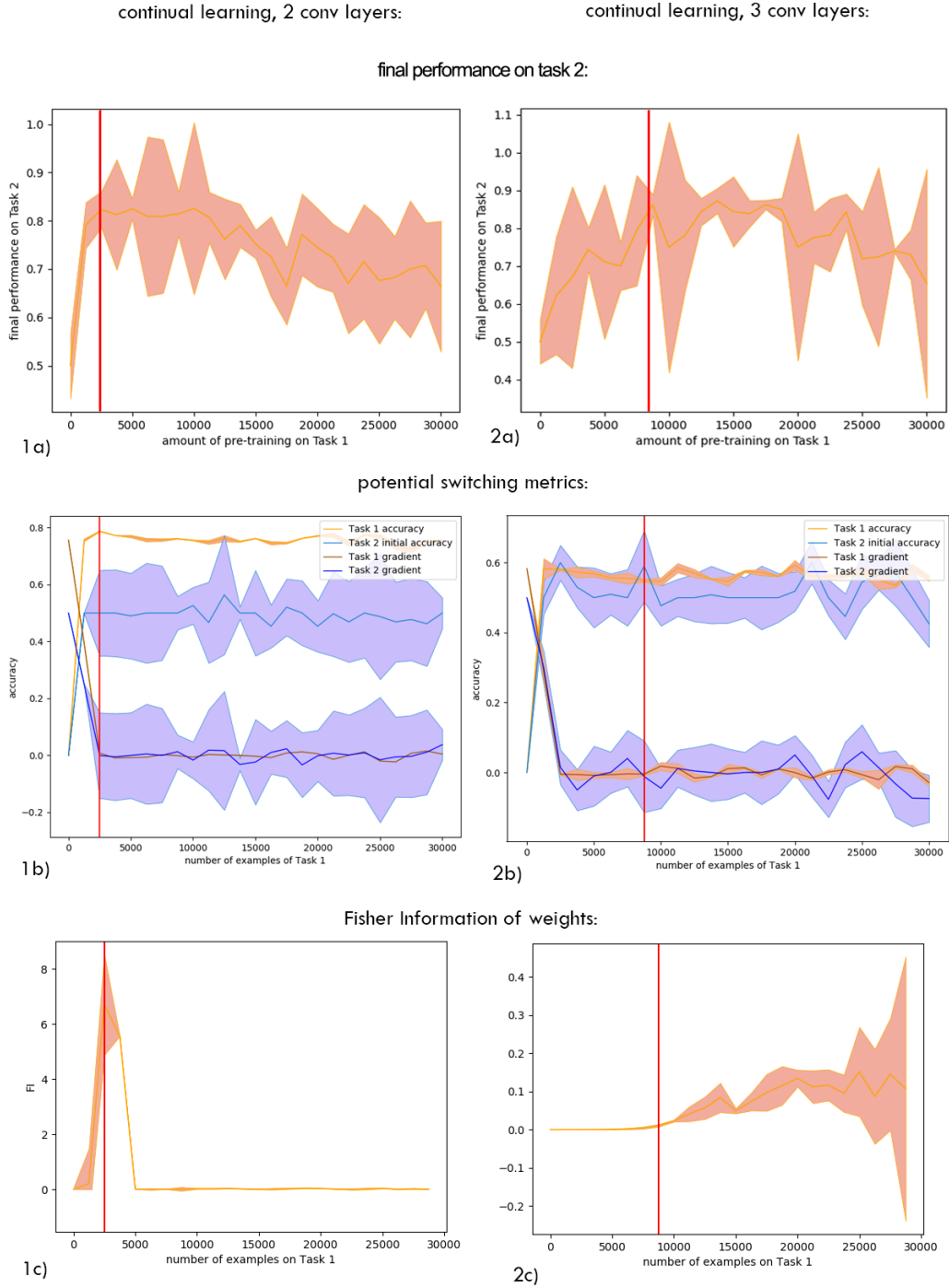
The FI of weights in Task 1 was partially effective at predicting the beginning of switching period in our original network. To explore whether this decrease in weight-change generalised to other cases of learning, and whether this effect may arise due to switching from feature memorizing to feature reorganization described by Achille [Achille et al., 2019], we calculated FI for Task 1 for a deeper network and for the original network trained on smaller slices. As can be seen on Fig. 4.5 2c and Fig. 4.8, FI in these tasks showed no correlation with final performance only peaking after the optimal switching window in these cases (Fig. 4.9). FI for the next step of training on Task 1 also remained significantly greater than FI after switching to Task 2 (Fig. 4.10). This shows that FI is not a reliable indicator to guide curriculum learning.

To sum up, we found that neither main nor supporting task performance or their gradi-

| Amount of pre-training | <i>FI in Task 2</i> |                |
|------------------------|---------------------|----------------|
|                        | mean                | stdev          |
| <b>2500</b>            | 0.001124378395      | 0.002429203634 |
| <b>5000</b>            | 0.0005223868013     | 0.002110925266 |
| <b>10000</b>           | 0.001173843067      | 0.004293080181 |
| <b>15000</b>           | 0.0008320442903     | 0.001386499149 |

Table 4.1: **The Fisher Information of weights for the original 2-conv network with respect to the Task 2 loss function during first 1250 examples of Task 2 training.**

ents, nor Fisher information of weights provided a signal that could guide active learning when training a network in a continual way. These results provide no evidence that the optimal learning strategy indeed can be derived from information available to the network.



**Figure 4.5: Comparison of potential switching metrics during continual learning.** Network versions with 2 and 3 convolutional layers were compared. On all figures red vertical lines indicate the optimal switching window. a) Mean of 8 runs and standard deviation bands for final performance on Task 2 with respect to the amount on pre-training on Task 1. b) Neither auxiliary nor primary task, nor their gradients provided a signal that could guide active learning. c) The mean of the Fisher Information of weights with respect to the loss function for Task 1 across 8 runs. The bars show one standard deviation. Fisher Information peak behaved differently in a deeper network with 3 convolutional layers, also not providing a signal to guide switching (mean of 12 runs, red vertical lines indicate the optimal switching window).

continual learning, 3 conv layers:

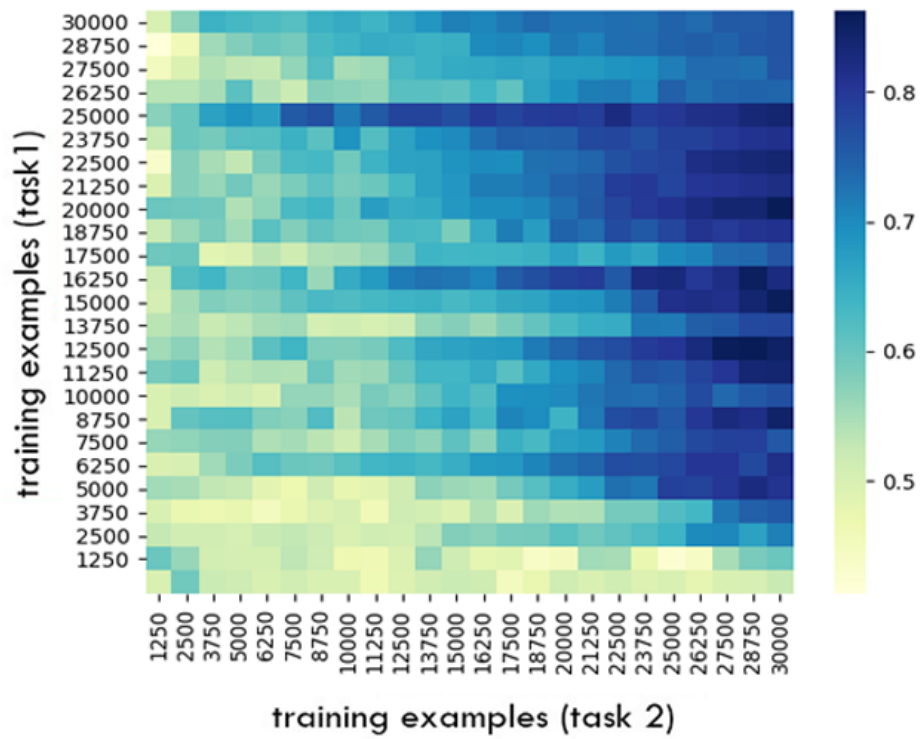


Figure 4.6: **Using 3 convolutional layers.** Performance on task 2 for a deeper network with respect to amount of training on Task 1 and Task 2.

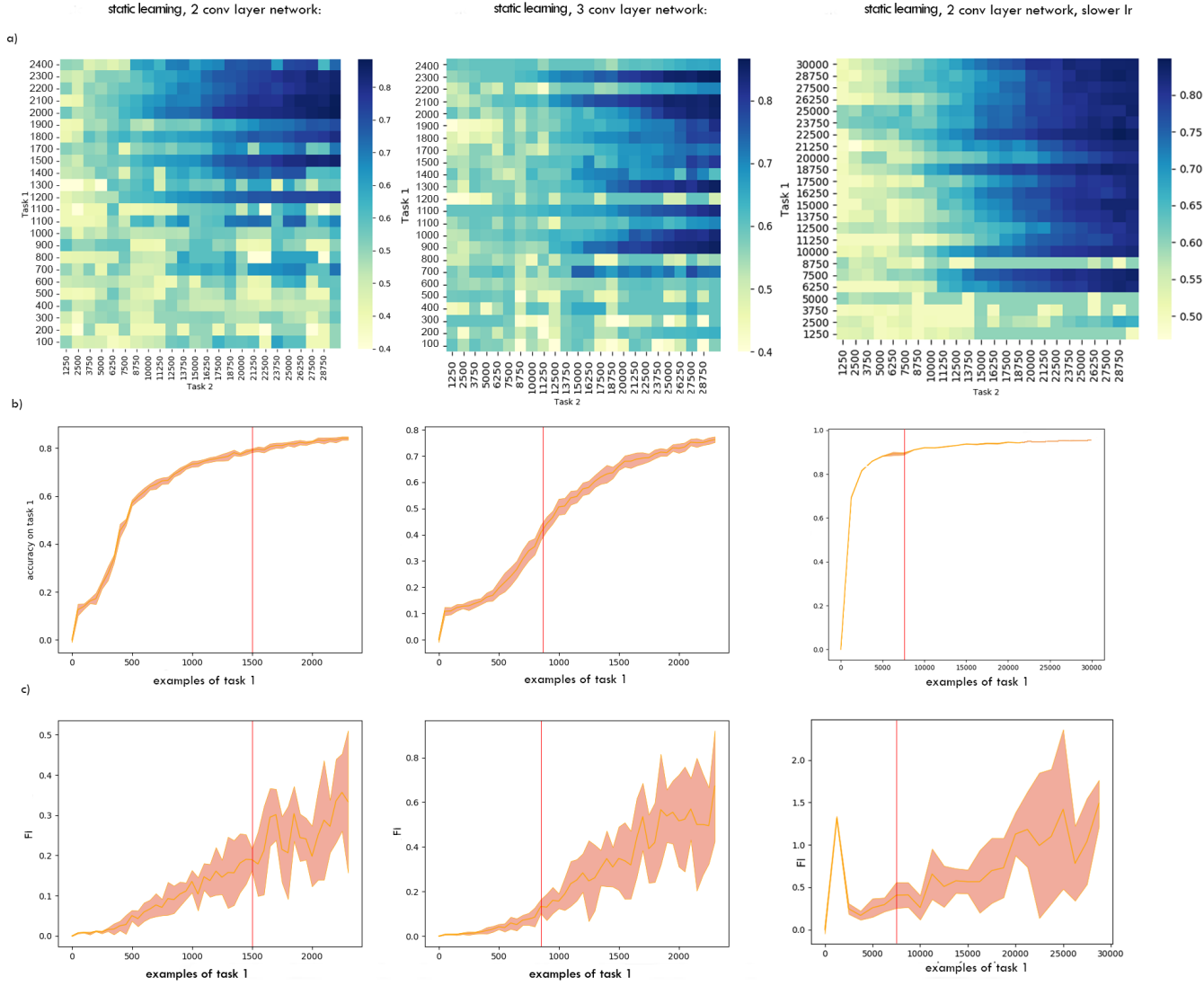


Figure 4.7: **Comparison of potential switching metrics during static learning.** The original 2-convolutional layer network, a deeper 3-convolutional layer network and the original 2-convolutional layer network with a slower learning rate of 0.0005 were compared. a) Performance on Task 2 for a with respect to amount of training on Task 1 and Task 2. b) Mean of 8 runs and standard deviation bands for performance on Task 1. c) Mean of 8 runs and standard deviation bands for Fisher Information.

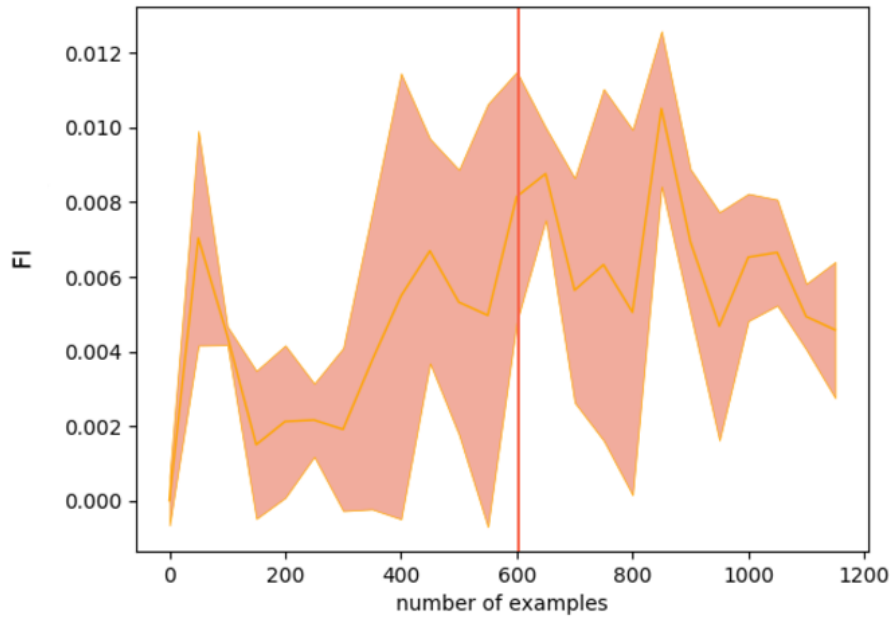
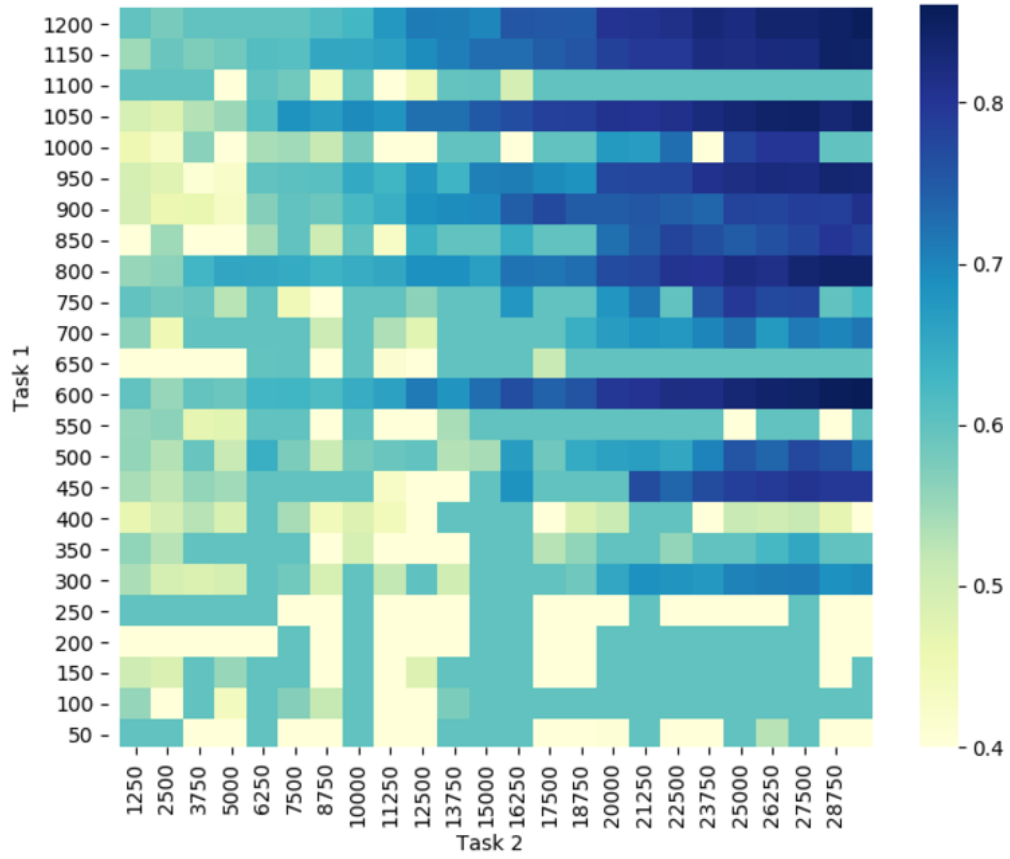


Figure 4.8: **Early stages of continual learning.** FI (mean of 6 runs and standard deviation) and final performance for smaller slices during the early stages of continual learning.

Figure 4.9: **Predicting final performance on Task 2 with FI.** Data from 8 runs, each point corresponds to the FI value for Task 1 at the moment of switching and the final performance on Task 2 for a single learning run. Correlation between final performance on Task 2 and Fisher Information with various degrees of training on Task 1: for continual learning:  $r=0.115$ ,  $p=0.21$ , for static learning:  $r=0.22$ ,  $p=0.007$ . In this case FI can only predict 1.3% and 4.8% of the Task 2 final performance.

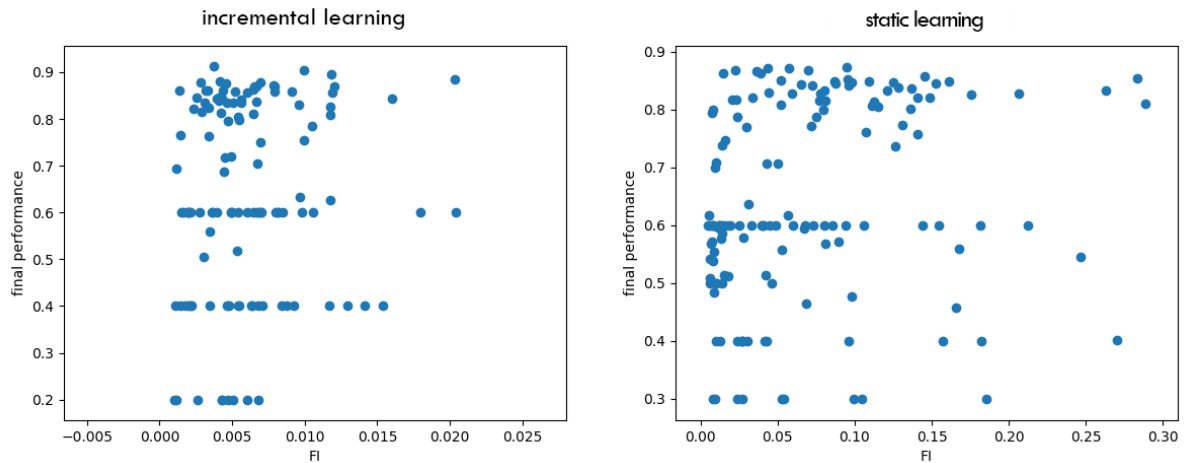


Figure 4.10: **Using FI to predict the gains from switching to another task on the next step.** FI for training on the next slice of Task 1 set and the first slice of Task 2 set after a given amount of pre-training on Task 1. After any amount of pre-training, information gain for continuing learning on Task 1 was significantly greater than for switching to the first slice of Task 2 set.)



### 4.3.3 Using digit addition for Task 2

In the odd-even task, the error signal was orthogonal to any single-image class, providing equal reward for learning and 'unlearning' it. Every digit was distributed equally in the odd and even digit-sum classes, and so learning weight changes after each batch of examples were cancelling each other.

An addition Task 2b, in which the network must report the sum of two digits in an input image, is in many ways similar to the odd/even task: two digits are presented; a mapping from an image to single representation would be valuable; and the output is derived from a combination of the two digits. However, in contrast to the odd/even task, it was possible to design it in a way so that the classes in Task 2b were not orthogonal to single digits. Our prediction, therefore, was that it would not require pre-training.

We found it could be learned without pre-training and differences in performance for varying amount of pre-training did not exceed the standard deviation between runs (Fig. 4.11). This supports the interpretation that it is the relationship between the training labels and the desired intermediate representation (here, categorisation of single digits) that determines the need for pre-training.

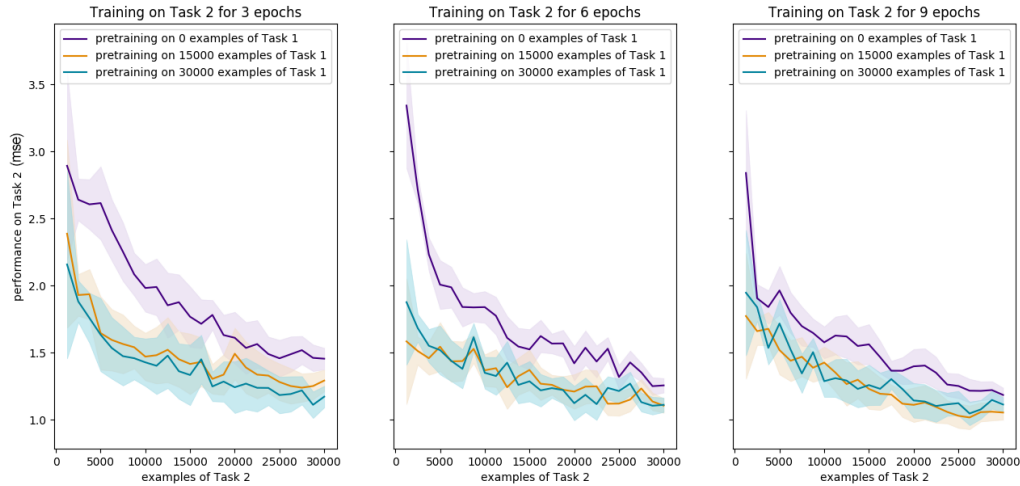


Figure 4.11: **Comparison of learning trajectories for the addition task with respect to various amount of pre-training.** Mean of 4 runs and standard deviation bands for performance on Task 2 (mean square error is used instead of accuracy for this task).

#### 4.3.4 Examining task representations

As performance for the addition task shows, these are likely to be dependent on the nature of the task. To understand how the representations developed in the network, we used t-SNE to see the manifold distances between a selection of digits. Fig.4.12 shows that examples consisting of two same digits proved to be the most problematic to represent correctly. This suggests that active learning could be harder to perform on tasks with low inter-class distance.

The two tasks themselves were also orthogonal. Task 1, while providing the network with useful representations, did not give it any information about odd and even sums, and Task 2a did not provide a useful signal for the network to be able to "catch up" with learning separate digits if the representations were not already formed.

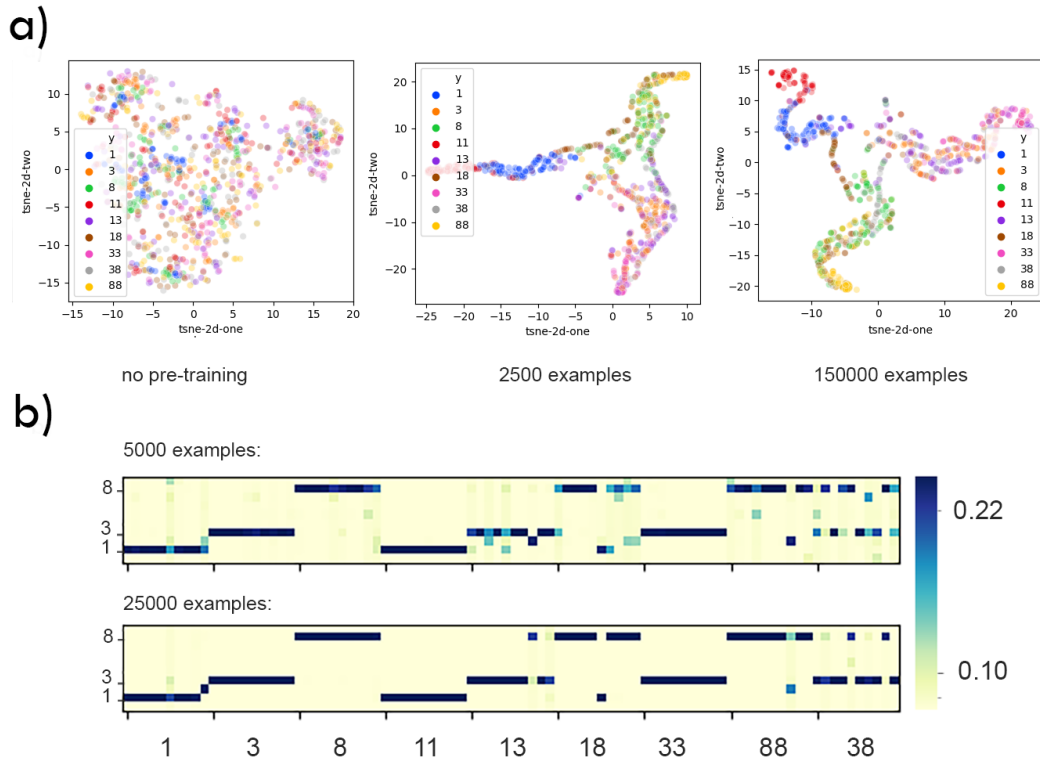


Figure 4.12: **Evolution of task representations with training.** a) t-SNE analysis of activations in the first dense layer of the CNN during the continual learning process. Representations of single digits are formed in the first few slices of the training set and change little with overtraining. The model was also able to discriminate repeated digits from single digits (e.g. "11" from "1"). However, the representation of images comprising two different numbers appears to be converging on the centre of the space: potentially, the "average" response to the two individual digits. b) First dense layer activation for training on 5000 and 25000 examples of Task 1. The x-axis shows the image classes presented to the network. Ten examples of each image class were shown. The y-axis labels the 10 neurons at the first dense layer, ordered by digit preference. Representations were stable over time. Single digits were reliably recognised, irrespective of the degree of training. However, when encountering stimuli consisting of two different digits it seems to only represent one of them, and this was exacerbated by overtraining.

## 4.4 Discussion

This study highlights a limitation of automated curriculum learning in a simple task context. It may not be a viable strategy for tasks in which the benefits of the chosen curriculum is not apparent until much later. Such tasks would require an external curriculum to be learned correctly.

Due to the limitations of architecture imposed by the original study goal, modelling true non-stationarity and catastrophic forgetting was not attempted (S2 section of the Supporting information can be referred to for more detail). However, even the simple case of balanced class representation provided enough challenge to the current network when learning in a naturalistic continual setting.

We tested four diverse types of measures, based on those that have proven successful in previous work. We found no evidence that any of the measures could indicate when sufficient pre-training had been obtained, suggesting that active learning may not be possible for this task. However, it should be noted that our search of measures was not exhaustive and it remains possible that there is some statistic that would allow active learning in this case. A further consideration is that even if there is some measure that would have worked in this task, it may not be possible to know in advance which measure, and at what threshold, learning should transition. Active learning would not be useful, if identifying its hyperparameters is more difficult than determining the optimal curriculum through an exhaustive search.

We speculate that active learning will be particularly difficult when the loss function of the pre-training and final task are orthogonal, so requiring quite different directions in gradient descent. In such tasks learning one level, while providing useful representations for future learning, will not teach the network anything about other levels. This should be especially noticeable in case of specific kinds of multitask curricula. Consistent with this, we found that a addition task sharing many similar demands (Task 2b) did not benefit from pre-training as much as the odd/even digit sum task (Task 2a).

Ruder's overview [Ruder, 2017] of multi-task learning approaches provides a summary of task relations. Simplifying his classification, we suggest that mutually assisting tasks (related or even adversarial tasks) may benefit from active learning. When tasks share features the model may have enough information to know when to switch. However, if one task provides a foundation for learning the other, but not the other way around, switching may prove problematic.

"In pedagogical practice it is accepted that there are benefits to an overarching framework to the fixed curriculum, but that within this active learning can be valuable [Bonawitz et al., 2011]. This is supported by the Vygotskian concept of scaffolding, in which teachers support learners to achieve the next level by challenging them to the optimal degree. Thus, keeping learners in their "zone of proximal development", while still relying on curious exploratory behaviour to some extent [Wood et al., 1976]. It is possible that CNNs and other computational models could benefit from similar frameworks. However, this requires further investigation of the benefits of interplay between curricula and active learning."

## 5 General Discussion

### 5.1 Summary

The work in this thesis aimed to investigate the how the process of infant learning may be structured by different environmental and biological factors in the infant's early years, creating a naturally-emerging curricula. The research focused on the effects of the infant's environment, the maturation of the brain and visual system, and of infants' innate curiosity.

When we think of the term 'curriculum' we usually turn to the classic more narrow definition of it. The importance of a well-designed curriculum for preschool and primary school children is well-studied and cannot be overstated [Burchinal et al., 2008, Weiland and Yoshikawa, 2013, Weiland et al., 2013]. Studies have consistently shown that a good curriculum positively impacts children's cognitive, social, and emotional development, providing children with age-appropriate learning opportunities, which help them develop critical thinking, creativity, and problem-solving skills. For older children formal schooling have been shown to help bridge the achievement gap between children from different socioeconomic backgrounds [Huang and Sebastian, 2014] and to improve general intelligence: for every year of education, students gain on average one to five IQ points [Ritchie and Tucker-Drob, 2018].

For obvious reasons there appear to be few works on the topic of curriculum in infancy, despite the period being essential for development and laying out the foundation for future learning [Shonkoff and Phillips, 2000, on the Developing Child, 2007]. One thing

modern researchers can agree on is the fact it should be based on principles of responsive and relationship-based care [Gonzalez-Mena and Widmeyer Eyer, 2017, Weisberg et al., 2016].

However, we think it is important to discuss the existence of less formal, naturally-emerging curricula: the way infants' experiences are structured not deliberately by their caretakers but rather by the inherent properties of the world itself. As mentioned earlier, it is not normally thought of as a 'curriculum', as we tend to associate curriculum with formal schooling.

However, once we try to mimic infant learning with AI modelling, this underscores the structured nature of infant learning that is influenced by both biological and environmental factors and unequivocally structured and consistent among numerous individuals. Consequently, it is reasonable to classify these naturally-emerging curricula through the lens of curriculum learning. Adopting this perspective may provide insights into the old nature vs nurture debate: deciding which part of our talents, behaviour and development patterns is biologically predetermined. Examining the naturally-occurring curricula that infants are exposed to may prove instrumental in disentangling innate characteristics from those that are acquired through learning, yet remain nearly-universal due to the consistency of the properties of the world.

Such naturally-occurring curricula can emerge in several different ways: statistics of the environment, the maturation of perceptual systems, and for older children - through self-directed learning. These types of curricula can also interact and with one level feeding into the next.

In this work we have investigated possible ways of curricula emergence, using visual learning as an example and computational modelling as our main method. We focussed upon the visual cortex as a target for modelling, as it is well-studied both behaviourally and neurally, and the suitability of convolutional neural networks as models of visual cortex has been investigated in detail.

In the introduction in Chapter 1 we motivated the validity of this approach, reviewing

the development of visual system, and the suitability of convolutional neural networks (CNNs) as models of the human visual cortex. In addition to that have also analyzed the current state of research on curriculum learning, both in human and artificial models, as well as the concept of curiosity as a self-driven mechanism for curriculum generation.

In the experimental chapters we studied examples of possible origins of infant curricula, such as statistics of the environment, the physiological development of sensory systems, and the viability of strategies used to guide attention.

In Chapter 2 we investigated how category acquisition in infancy is influenced by the statistics of the world. When children expand their vocabulary in their first years of life, they tend to learn object categories in a certain order. The precedence of some superordinate groupings of categories, such as animate vs. inanimate is consistent across different languages. The animate/inanimate distinction in particular also corresponds to systematic large-scale structures in the visual cortex. Current theories explaining it mostly focus on the perceptual properties of the objects, social interaction, and linguistic cues. However, statistics of the world, such as, for example, frequency and the context within which an object occurs, also undoubtedly plays an important role. In this study we investigated the amount of information one can derive from such statistics, using video datasets of everyday life scenes. To do this, we needed to evaluate automatic methods for tagging (validating them against human tagging in a smaller sample). We also investigated generality across dataset by using three different sources.

In Chapter 3 we demonstrated that slow maturation of visual system may, in fact, be beneficial for visual learning, inducing a bias for global processing. Infants are born with poor visual acuity, incapable of perceiving fine details during their first months of life. While an apparent disadvantage, it may be adaptive by encouraging the learning of the “big picture” - global rather than local features. Congenitally blind infants who have sight restored as early as 2 months of age, later experience characteristic deficits in spatial integration, including deficits in configural face processing and a reduced capacity to



perceive global form. The high initial acuity (HIA) hypothesis [Vogelsang et al., 2018] proposes that these deficits arise from characteristics of the visual environment from the onset of sight, specifically the near immediate availability of high spatial frequency information, unavailable to a newborn. CNNs, which have been shown to be otherwise viable models of the ventral visual stream, similarly demonstrate a tendency to use local image features for image classification. Until recently, the consensus in the computer science community was that this preference could be attributed to the architectural limitations of the existing CNN models. However, a recent study [Geirhos et al., 2019] demonstrates that a network can be pushed towards shape or texture bias by adding a certain type of distortions the training set. Better performance on the task of face recognition has also been displayed by a network partially trained on blurry images [Vogelsang et al., 2018]. We extended these works, addressing their key limitations and using a naturalistic learning trajectory mimicking infant vision maturation to train the network on domain-independent data. We then evaluated its shape bias (i.e. the tendency to rely on global features for visual classifications). The results suggest that such a simple change to the training process is sufficient to trigger a preference for global processing, also impacting robustness to noise and intercategory perceptual distance for non-basic categories.

Curiosity-driven learning as an important aspect of infant learning and in the fields of cognitive modelling and developmental robotics curiosity is seen as a strategy of choosing the model's own training data that would optimize learning outcomes [Oudeyer et al., 2016]. In a sense it can be viewed as a way of creating own curriculum. in Chapter 4 we investigated possible limitations of creating own curriculum through self-driven curious learning. Models of curiosity has been shown to improve learning for AI, some results also being consistent with human category learning and sequence learning [Twomey and Westermann, 2017a, Poli et al., 2022], which suggest that, to some extent, artificial neural networks can be viable models of human curiosity.

It has been shown that in many applications, artificial neural networks are best trained for a task by following a curriculum, in which simpler concepts are learned before more

complex ones. This curriculum can be hand-crafted by the engineer or optimised like other hyperparameters, by evaluating many curricula. However, this is computationally intensive and the hyperparameters are unlikely to generalise to new datasets. An attractive alternative, demonstrated in influential prior works, is that the network could choose its own curriculum by monitoring its learning. This would be particularly beneficial for continual learning, in which the network must learn from an environment that is changing over time, relevant both to practical applications and in the modelling of human development. In this experiment we tested the generality of this approach using a proof-of-principle model, training a network on two sequential tasks under static and continual conditions, and investigated both the benefits of a curriculum and the handicap induced by continuous learning. We found that in some cases even in this simple scenario the a network is often unable to choose the optimal curriculum, as the benefits are sometimes only be apparent with hindsight, at the end of training.

## 5.2 Benefits of curriculum

To sum up these experiments, we have shown that the environment could play a significant role in shaping learning both for humans and artificial models. Specifically, we have demonstrated how statistics of the environment and the process of visual maturation can create a curriculum, as well as the need for a structured curriculum in certain situations.

It is important to note that, while the impact of structured learning normally researched for more complex topics suitable for at least pre-school schildren, low-level processes also benefit from curriculum.

Despite consistent evidence of importance of curriculum in human and animal studies [Lawrence, 1952, Baker and Osgood, 1954, Elio and Anderson, 1984, Wilson et al., 2019], recent literature on curriculum learning in computer science presents conflicting evidence regarding its effectiveness, and necessity (see [Saglietti et al., 2022] for review and analysis). Some simulation studies reach the conclusion that curricula are

moderately effective or even ineffective in most cases. Experiments in a variety of domains [Hacohen and Weinshall, 2019, Pentina et al., 2014] have demonstrated that speed and generalisation improvements from curricula are generally underwhelming. According to Wu and colleagues recent extensive empirical analysis, curriculum yields minimal benefits on standard datasets [Wu et al., 2020].

One of the suggestions for this discrepancy is the differences in memory and synaptic consolidation strategies [Saglietti et al., 2022]. Based on our findings and the existing conflict between structured and curiosity-driven learning in human subjects, we argue that the efficacy of a curriculum is also contingent upon the type of task. The question is then not whether curriculum is beneficial at all but rather what are the situations in which it improves learning and when it does not.

Inferring from the results of our experiments and existing studies we can suggest the following benefits of curriculum:

- **Learning in a less superficial way.** In works that question the usefulness of the curricula ANNs have been trained on the CIFAR dataset and similar perceptual tasks [Wu et al., 2020, Hacohen and Weinshall, 2019, Pentina et al., 2014], whereas the utility of curriculum has been proven in such tasks as language learning [Elman, 1993, Nagatsuka et al., 2021] and solving mathematical equations [Polu et al., 2023]. This could be a potential avenue for future research.

It is possible that a structured curriculum is more beneficial for the development of higher-order cognitive functions and learning language-like patterns and other types of highly-structured information. It can potentially help prioritize the "importance" of a feature over perceptual salience. However, further experiments are required to test this.

- **Speeding up learning.** Larger networks have been shown to need more time to converge and the learning process may be shortened by a curriculum [Bengio et al., 2009], curriculum also being beneficial in noisy settings [Li et al., 2021] and when a sparse signal is embedded in many irrelevant dimensions of variation [Saglietti et al.,

2022]. While it makes little difference in a lab experiment, overcoming noise and time limits is important in naturalistic settings. As effective as they are, modern state-of-the-art models, such as ChatGPT and PALM are trained on several orders of magnitude more data than a child ever receives [Leiter et al., 2023].

- **Adapting generic architecture to specific tasks.** It has been shown that the design and implementation of a curriculum can influence the degree of specialization or generalization exhibited by a neural network: When training a network to perform multiple tasks a correctly chosen curriculum has been shown to push the network towards either more specialized or more flexible representations without having to modify the architecture [Yang et al., 2019]. This can be used to help the architecture adapt to a knowledge domain.

Prolonged weaning periods may be an important mechanism for achieving adaptability in animals. This is because a longer weaning period allows for increased learning and exploration, which can enhance an animal's ability to adjust to different environments, 'outsourcing' part of the adaptation to parents instead of evolutionary selection. Some researchers interpret it as a trait that is particularly useful for social learning [Faust et al., 2020].

## 5.3 Directions of future research

### 5.3.1 Expanding existing experiments

There are many ways in which current experiments may be improved and expanded.

#### **Learning the statistics of the world**

- In current work we have investigated the statistics of visual world and confirmed the existence of clusters resembling tripartite distinction, however, comparison with fMRI data remains to be conducted. It is also possible that, while not visible for broad clusters, non-stationarity or context specificity of some categories may still affect their cortical representations in a more subtle way.

- It would be interesting to explore how network architecture (such as varying learning rate or network size) corresponds to the learnability of classes with certain statistical properties. If there exists a gradient in the ventral visual stream circuitry that leads to certain areas being tuned to particular statistics (e.g., preferred frequency) then the statistical differences could lead to the stereotypical organisation of the ventral visual stream.
- A more in-depth investigation of more specific categories can also be conducted once the PLAY dataset is available.

### **Benefits of low initial acuity**

Apart from visual acuity, the role of gradual colour vision development and variations in shape/angle also needs to be investigated. Infants experience significant variation of light conditions when learning about objects, which influences the colour scheme of objects, they also eagerly introduce variation to their own visual input once they develop sufficient motor skills and pay great attention to movement [Dusing and Harbourne, 2010].

The effects of low initial acuity on movement perception also have not been sufficiently studied.

Another potential development of the low initial acuity studies is to explore its benefits for resisting classic adversarial attacks, as those heavily rely on local processing bias of artificial neural networks.

Yet another simple improvement would be combining style transfer with low initial acuity training, to check whether performance improvements achieved by Geirhos [Geirhos et al., 2019] can be further enhanced.

### **The limitations of curiosity-driven learning**

The current curiosity experiment could be expanded to more realistic stimuli, allowing to draw parallel with the classic experimental paradigms designed for studying curiosity in infants and older children. The classic Plunkett's model [Mather and Plunkett, 2011,

Twomey and Westermann, 2017a] of category specificity with combinations of features described by different mathematical regularities could be a good starting point.

Testing a proper 'curiosity' model instead of only curiosity metrics, allowing the network to select its own stimuli could also lead to new results.

Using more realistic stimuli would also allow to compare the model's behaviour to human participants and possibly to compare adult and infant curiosity.

### **5.3.2 Further goals**

Along the course of this thesis, a search for foundational models capturing the essential features necessary to understand the underlying principles, have become a significant direction in AI research. Learning in the helpless period of infancy may provide researchers with potential ideas for developing better computational models [Zaadnoordijk et al., 2022].

Existing body of literature suggests that curriculum learning may benefit artificial models just as much as it benefits humans.

#### **Helping AI 'understand' data**

As previously stated, curricula may serve as a mechanism for guiding learners towards a more in-depth approach to learning, potentially mitigating the common issue of shallow data comprehension among networks.

It is possible that curricula could enhance the capacity of networks prioritize meaningful features over salient ones and focus on interconnected elements, resulting in more human-like behaviour. More experiments are required to test that hypothesis.

Interaction of curiosity with curriculum is of particular interest in this kind of tasks, as the tasks humans normally require a curriculum for could also be tasks where it is difficult to discover meaningful features without external help.

#### **- Redefining scaffolding**

When researching scaffolding, it is important to separate the emotional component (e.g. reduced anxiety due to the help of a caretaker) from the actual curriculum required for the task, as well as determine the exact type of curriculum required. Modelling provides a good starting point for that. Existing works suggest that multiple types of scaffolding exist [Rogoff, 1990, Belland et al., 2013, Van Lier, 2004], and their effectiveness in particular cases may be tested with a variety of artificial curricula.

Experiments with artificial models can be seen as a way of predicting prospective directions of investigation of when and how do humans benefit from curriculum. They can also provide a convenient tool for separating purely informational aspects of scaffolding from emotional benefits of studying with a teacher.

We believe that in this thesis, by segregating different strategies of creating naturalistic curricula (statistical properties of the dataset, modifying perceptual properties and using sequential tasks) and demonstrating what kind of benefits they entail, we have provided a foundation for recognizing naturally-occurring curricula, better identifying zones of proximal development, and, potentially, for building better curricula.

### **5.3.3 Broad implications**

We have shown the curriculum to be a fundamental element not just in education but in early infant learning, serving as a tool for directing learning towards specific goals.

However, the scope of this thesis only concerns visual learning. These effects may differ in other modalities or during multimodal learning.

Multiple senses might provide complementary information, for example during speech processing humans have been shown to rely both on sound and on visual cues such as lip movement [McGurk and MacDonald, 1976]. "This has the potential to both facilitate and hinder learning, however, the existing evidence is contradictory. It has been shown to improve word segmentation [Lusk and Mitchel, 2016] and learning under certain circumstances [Bernstein et al., 2013, Mitchel et al., 2022]. However, studies also

exist which show that obstruction of visual cues it does not influence infants' learning progress [Çetinçelik et al., 2023] also exist. The validity of the visual cue conflict effect outside of lab setting is also currently a question of study [Engen et al., 2022].

Another implication of the importance of early curricula is the potential to provide help to infants, for whom some aspects of natural curricula have been disrupted.

Although at the moment these results are insufficient to be used as guidance in clinical practice, this accentuates the importance of creating rich learning environments during infancy and the potential helpfulness of customized curricula for neurodivergent children, which can facilitate the attainment of developmentally appropriate trajectories.

For example, it has been shown to be effective for social learning in autistic children [Chrysaitis and Seriès, 2023]. In the case of vocabulary acquisition (which, in many cases, is a good indicator of overall pace of cognitive development), healthy infants who generated more variable visual object images through manual object manipulation at 15 months of age experienced greater vocabulary growth over the next six months [Slone et al., 2019]. This can be connected to hindered developmental progress observed in some preterm infants who are more susceptible to motor impairment [Moreira et al., 2014] and can be implied that implementing specific interventions could potentially facilitate the attainment of developmental milestones.

We believe that many other cases may be found with rigorous enough investigation. Some of the neurodivergency symptoms may be amplified by feedback loops and could be compensated early with the right interventions.

There are also implications for curiosity modelling. It is possible that current demands made of curiosity models in developmental robotics are beyond what even humans are capable. For example, existing curiosity algorithms tend to focus on white noise, prioritizing it over everything else, while humans do not exhibit this behaviour. This inability to deal with pure randomness is seen as imperfection of the algorithms while it is highly possible that humans might be trained to cancel out noise during the development of our visual system.



Better knowledge of interaction between curiosity and previous knowledge, which curriculum implies, might also allow to compare adult and infant curiosity from an informational perspective.

Whether curiosity changes during various stages of development is currently unknown. There exists little to no studies comparing children and adult curiosity [Gopnik et al., 2017]. Alison Gopnik observed that younger learners sometimes outperform older ones on learning tasks showing more flexibility. Her theory suggests that childhood involves a shift from a highly exploratory mode in infancy to an adult strategy of exploiting the skills one has developed [Gopnik et al., 2017].

This effect can be attributed both to biological reasons such as greater neuronal plasticity or immaturity of certain zones in infants and the shift of effectiveness in strategies learning with acquiring more information about the world.

#### **5.3.4 Conclusion**

In conclusion, the significance of curriculum, particularly in the context of early development, cannot be overstated.

In this work we have shown how curriculum can arise from the environment and biological maturation of the brain and perceptual systems. We have also provided an example of a case where external curricula is necessary as the learner will fail to find an optimal one through self-driven learning.

We believe these findings highlight the potential of another aspect of infant development to serve as a source of inspiration and insight for the advancement of AI.

# Bibliography

- [Achille et al., 2019] Achille, A., Rovere, M., and Soatto, S. (2019). Critical Learning Periods in Deep Neural Networks. *arXiv:1711.08856 [cs, q-bio, stat]*. arXiv:1711.08856.
- [Ahn et al., 2021] Ahn, S., Zelinsky, G. J., and Lupyan, G. (2021). Use of superordinate labels yields more robust and human-like visual representations in convolutional neural networks. *Journal of Vision*, 21(13):13.
- [Althaus and Mareschal, 2014] Althaus, N. and Mareschal, D. (2014). Labels direct infants' attention to commonalities during novel category learning. *PLoS ONE*, 9(7):e99670.
- [Anderson et al., 2018] Anderson, N., Meagher, K., Welder, A., and Graham, S. A. (2018). Animacy cues facilitate 10-month-olds' categorization of novel objects with similar insides. *PLoS ONE*, 13(11):e0207800.
- [Atkinson, 1984] Atkinson, J. (1984). Human visual development over the first 6 months of life. *Human Neurobiology*, pages 61–74.
- [Baker and Osgood, 1954] Baker, R. A. and Osgood, S. W. (1954). Discrimination transfer along a pitch continuum. *Journal of Experimental Psychology*, 48(4):241.
- [Banks and Bennett, 1988] Banks, M. S. and Bennett, P. J. (1988). Optical and photoreceptor immaturities limit the spatial and chromatic vision of human neonates. *Journal of the Optical Society of America A*, 5(12):2059.

- [Barto et al., 2004] Barto, A. G., Mahadevan, S., and Modi, N. R. (2004). Intrinsic and extrinsic motivation. *Adaptive Behavior*, 11(2):79–113.
- [Behl-Chadha, 1996] Behl-Chadha, G. (1996). Basic-level and superordinate-like categorical representations in early infancy. *Cognition*, 60(2):105–141.
- [Bell and Lawrence, 2021] Bell, S. J. and Lawrence, N. D. (2021). Behavioral experiments for understanding catastrophic forgetting.
- [Belland et al., 2013] Belland, B., Kim, C., and Hannafin, M. (2013). Scaffolding and motivation in technology-enhanced learning environments: A meta-analysis. *Instructional Science*, 41(3):657–688.
- [Bengio et al., 2009] Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 41–48.
- [Berlyne, 1954] Berlyne, D. E. (1954). A theory of human curiosity. *British Journal of Psychology. General Section*, 45(3):180–191.
- [Berlyne, 1960] Berlyne, D. E. (1960). *Conflict, arousal, and curiosity*. McGraw-Hill Book Company.
- [Bernier et al., 2010] Bernier, A., Carlson, S. M., and Whipple, N. (2010). From external regulation to self-regulation: Early parenting precursors of young children’s executive functioning. *Child Development*, 81(1):326–339.
- [Bernstein et al., 2013] Bernstein, L. E., Auer, E. T., Eberhardt, S. P., and Jiang, J. (2013). Auditory perceptual learning for speech perception can be enhanced by audiovisual training. *Frontiers in Neuroscience*, 7.
- [Bonawitz et al., 2011] Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., and Schulz, L. (2011). The Double-edged Sword of Pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, 120(3):322–330.

- [Braginsky et al., 2019] Braginsky, M., Yurovsky, D., Marchman, V. A., and Frank, M. C. (2019). Consistency and Variability in Children’s Word Learning Across Languages. *Open Mind*, 3:52–67.
- [Bransford et al., 2000] Bransford, J., Brown, A., and Cocking, R. (2000). *How People Learn: How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. National Academies Press.
- [Brendel and Bethge, 2019] Brendel, W. and Bethge, M. (2019). Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. *arXiv:1904.00760 [cs, stat]*. arXiv: 1904.00760.
- [Brodal, 2010] Brodal, P. (2010). *The central nervous system: Structure and function (4th ed.)*. Oxford University Press.
- [Burchinal et al., 2008] Burchinal, M., Howes, C., Pianta, R., Bryant, D., Early, D., Clifford, R., and Barbarin, O. (2008). Predicting child outcomes at the end of kindergarten from the quality of pre-kindergarten teacher–child interactions and instruction. *Applied Developmental Science*, 12(3):140–153.
- [Burda et al., 2018] Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., and Efros, A. A. (2018). Large-Scale Study of Curiosity-Driven Learning. *arXiv:1808.04355 [cs, stat]*. arXiv: 1808.04355.
- [Cabral et al., 2022] Cabral, L., Zubiaurre-Elorza, L., Wild, C. J., Linke, A., and Cusack, R. (2022). Anatomical correlates of category-selective visual regions have distinctive signatures of connectivity in neonates. *Developmental Cognitive Neuroscience*, 58:101179.
- [Carlson, 2003] Carlson, S. M. (2003). Developmentally sensitive measures of executive function in preschool children. *Developmental Neuropsychology*, 23(3):413–429.
- [Çetinçelik et al., 2023] Çetinçelik, M., Rowland, C. F., and Snijders, T. M. (2023). Does the speaker's eye gaze facilitate infants' word segmentation from continuous speech? an ERP study. *Developmental Science*.

- [Chen et al., 2022a] Chen, X., Twomey, K. E., and Westermann, G. (2022a). Curiosity enhances incidental object encoding in 8-month-old infants. *Journal of Experimental Child Psychology*, 223:105508.
- [Chen et al., 2022b] Chen, X., Twomey, K. E., and Westermann, G. (2022b). The role of metacognitive abilities and curiosity in learning.
- [Chomsky, 1959] Chomsky, N. (1959). Review of b. f. skinner's verbal behavior. *Language*, 35(1):26–58.
- [Christiansen, 2018] Christiansen, M. H. (2018). Implicit statistical learning: A tale of two literatures. *Topics in Cognitive Science*, 11(3):468–481.
- [Chrysaitis and Seriès, 2023] Chrysaitis, N. A. and Seriès, P. (2023). 10 years of bayesian theories of autism: A comprehensive review. *Neuroscience & Biobehavioral Reviews*, 145:105022.
- [Cireşan et al., 2012] Cireşan, D., Meier, U., and Schmidhuber, J. (2012). Multi-column deep neural networks for image classification.
- [Colung and Smith, 2003] Colung, E. and Smith, L. B. (2003). The emergence of abstract ideas: evidence from networks and babies. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1435):1205–1214.
- [Csikszentmihalyi, 1991] Csikszentmihalyi, M. (1991). *Flow: The Psychology of Optimal Experience*. Harper Perennial, New York, NY.
- [Curcio and Allen, 1990] Curcio, C. A. and Allen, K. A. (1990). Topography of ganglion cells in human retina. *The Journal of Comparative Neurology*, 300(1):5–25.
- [Cusack et al., 2018] Cusack, R., Wild, C. J., Zubiaurre-Elorza, L., and Linke, A. C. (2018). Why does language not emerge until the second year? *Hearing Research*, 366:75–81.
- [Davidson et al., 2017] Davidson, D., Rainey, V., Vanegas, S., and Hilvert, E. (2017). The effects of type of instruction, animacy cues, and dimensionality of objects on the

- shape bias in 3- to 6-year-old children. *Infant and Child Development*, 27.
- [Deci, 1971] Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology*, 18(1):105–115.
- [Deci and Ryan, 1985] Deci, E. L. and Ryan, R. M. (1985). *Intrinsic Motivation and Self-Determination in Human Behavior*. Springer US.
- [Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.
- [Downing and Peelen, 2011] Downing, P. E. and Peelen, M. V. (2011). The role of occipitotemporal body-selective regions in person perception. *Cognitive Neuroscience*, 2(3-4):186–203.
- [Dubey and Griffiths, 2020] Dubey, R. and Griffiths, T. L. (2020). Reconciling novelty and complexity through a rational analysis of curiosity. *Psychological Review*, 127(3):455–476.
- [Dusing and Harbourne, 2010] Dusing, S. C. and Harbourne, R. T. (2010). Variability in postural control during infancy: Implications for development, assessment, and intervention. *Physical Therapy*, 90(12):1838–1849.
- [Elio and Anderson, 1984] Elio, R. and Anderson, J. (1984). The effects of information order and learning mode on schema abstraction. *Memory Cognition*, 12:20–30.
- [Ellaway, 2014] Ellaway, R. H. (2014). When I say ... epistemic curiosity. *Medical Education*, 48(2):113–114.
- [Elleberg et al., 2000] Elleberg, D., Lewis, T. L., Maurer, D., and Brent, H. P. (2000). Influence of monocular deprivation during infancy on the later development of spatial and temporal vision. *Vision Research*, 40(23):3283–3295.
- [Elman, 1993] Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*, 48(1):71–99.

- [Elman, 1996] Elman, J. L. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. MIT press.
- [Emberson et al., 2011] Emberson, L. L., Conway, C. M., and Christiansen, M. H. (2011). Timing is everything: Changes in presentation rate have opposite effects on auditory and visual implicit statistical learning. *Quarterly Journal of Experimental Psychology*, 64(5):1021–1040.
- [Engen et al., 2022] Engen, K. J. V., Dey, A., Sommers, M. S., and Peelle, J. E. (2022). Audiovisual speech perception: Moving beyond McGurk. *The Journal of the Acoustical Society of America*, 152(6):3216–3225.
- [Epstein et al., 1999] Epstein, R., Harris, A., Stanley, D., and Kanwisher, N. (1999). The parahippocampal place area. *Neuron*, 23(1):115–125.
- [Esper, 1925] Esper, E. A. (1925). *A technique for the experimental investigation of associative interference in artificial linguistic material*. Number 1-4. Linguistic Society of America.
- [Fausey et al., 2016] Fausey, C. M., Jayaraman, S., and Smith, L. B. (2016). From faces to hands: Changing visual input in the first two years. *Cognition*, 152:101–107.
- [Faust et al., 2020] Faust, K. M., Carouso-Peck, S., Elson, M. R., and Goldstein, M. H. (2020). The origins of social knowledge in altricial species. *Annual Review of Developmental Psychology*, 2(1):225–246.
- [Fiser and Aslin, 2001] Fiser, J. and Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, 12(6):499–504.
- [French et al., 2002] French, R. M., Mermillod, M., Chauvin, A., Quinn, P. C., and Mareschal, D. (2002). The Importance of Starting Blurry: Simulating Improved Basic-Level Category Learning in Infants Due to Weak Visual Acuity. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 24(24).

- [Fukushima, 1980] Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202.
- [Gatys et al., 2016] Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [Gebhart et al., 2009] Gebhart, A. L., Newport, E. L., and Aslin, R. N. (2009). Statistical learning of adjacent and nonadjacent dependencies among nonlinguistic sounds. *Psychonomic Bulletin & Review*, 16(3):486–490.
- [Geirhos et al., 2019] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv:1811.12231 [cs, q-bio, stat]*. arXiv: 1811.12231.
- [Glasser et al., 2016] Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., and Essen, D. C. V. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178.
- [Gliozzi et al., 2009] Gliozzi, V., Mayor, J., Hu, J.-F., and Plunkett, K. (2009). Labels as features (not names) for infant categorization: A neurocomputational approach. *Cognitive Science*, 33(4):709–738.
- [Goldman-Rakic, 1987] Goldman-Rakic, P. S. (1987). Development of cortical circuitry and cognitive function. *Child Development*, 58(3):601.
- [Gómez, 2002] Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, 13(5):431–436.
- [Gonzalez-Mena and Widmeyer Eyer, 2017] Gonzalez-Mena, J. and Widmeyer Eyer, D. (2017). *Infants Toddlers and Caregivers: A Curriculum of Respectful Responsive*



*Relationship-Based Care and Education*. McGraw-Hill Higher Education, New York, NY, 11th edition.

- [Gopnik et al., 2017] Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., Aboody, R., Fung, H., and Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences*, 114(30):7892–7899.
- [Grand et al., 2001] Grand, R. L., Mondloch, C. J., Maurer, D., and Brent, H. P. (2001). Early visual experience and face processing. *Nature*, 410(6831):890–890.
- [Grand et al., 2004] Grand, R. L., Mondloch, C. J., Maurer, D., and Brent, H. P. (2004). Impairment in holistic face processing following early visual deprivation. *Psychological Science*, 15(11):762–768.
- [Graves et al., 2017] Graves, A., Bellemare, M. G., Menick, J., Munos, R., and Kavukcuoglu, K. (2017). Automated Curriculum Learning for Neural Networks. *arXiv:1704.03003 [cs]*. arXiv: 1704.03003.
- [Gruber et al., 2014] Gruber, M., Gelman, B., and Ranganath, C. (2014). States of Curiosity Modulate Hippocampus-Dependent Learning via the Dopaminergic Circuit. *Neuron*, 84(2):486–496.
- [Gruber and Ranganath, 2019] Gruber, M. J. and Ranganath, C. (2019). How Curiosity Enhances Hippocampus-Dependent Memory: The Prediction, Appraisal, Curiosity, and Exploration (PACE) Framework. *Trends in Cognitive Sciences*, 23(12):1014–1025.
- [Hacohen and Weinshall, 2019] Hacohen, G. and Weinshall, D. (2019). On the power of curriculum learning in training deep networks.
- [Hadsell et al., 2020] Hadsell, R., Rao, D., Rusu, A., and Pascanu, R. (2020). Embracing change: Continual learning in deep neural networks. *Trends in Cognitive Sciences*, 24:1028–1040.

- [Hammond et al., 2012] Hammond, S. I., Muller, U., Carpendale, J. I. M., Bibok, M. B., and Liebermann-Finestone, D. (2012). The effects of parental scaffolding on toddlers' executive function. *Child Development*, 83(3):896–910.
- [He et al., 2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.
- [Hensch, 2004] Hensch, T. K. (2004). Critical period regulation. *Annual Review of Neuroscience*, 27:549–579.
- [Huang and Sebastian, 2014] Huang, H. and Sebastian, J. (2014). The role of schools in bridging within-school achievement gaps based on socioeconomic status: a cross-national comparative study. *Compare: A Journal of Comparative and International Education*, 45(4):501–525.
- [Hubel and Wiesel, 1962] Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154.
- [Hughes and Ensor, 2009] Hughes, C. and Ensor, R. (2009). Executive function and theory of mind: Predictive relations from ages 2 to 4. *Developmental Psychology*, 45(2):352–364.
- [Hughes et al., 2013] Hughes, C., Roman, G., Hart, S., and Ensor, R. (2013). Does executive functioning predict socioeconomic status and language development in preschoolers? *Journal of Experimental Child Psychology*, 116(1):105–119.
- [Huth et al., 2012] Huth, A. G., Nishimoto, S., Vu, A. T., and Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224.
- [Huttenlocher et al., 1982] Huttenlocher, P. R., de Courten, C., Garey, L. J., and Van der Loos, H. (1982). Synaptogenesis in human visual cortex—evidence for synapse elimination during normal development. *Neurosci. Lett.*, 33(3):247–252.

- [Itti and Baldi, 2009] Itti, L. and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306.
- [Javed and White, 2019] Javed, K. and White, M. (2019). Meta-learning representations for continual learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- [Jinsi et al., 2023] Jinsi, O., Henderson, M. M., and Tarr, M. J. (2023). Early experience with low-pass filtered images facilitates visual category learning in a neural network model. *PLoS ONE*, 18(1):e0280145.
- [Johnson, 2011] Johnson, M. H. (2011). Interactive specialization: A domain-general framework for human functional brain development? *Developmental Cognitive Neuroscience*, 1(1):7–21.
- [Johnson et al., 1991] Johnson, M. H., Dziurawiec, S., Ellis, H., and Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, 40(1-2):1–19.
- [Jovanovic and Matejevic, 2014] Jovanovic, D. and Matejevic, M. (2014). Relationship between rewards and intrinsic motivation for learning – researches review. *Procedia - Social and Behavioral Sciences*, 149:456–460.
- [Kamps et al., 2020] Kamps, F. S., Hendrix, C. L., Brennan, P. A., and Dilks, D. D. (2020). Connectivity at the origins of domain specificity in the cortical face and place networks. *Proceedings of the National Academy of Sciences*, 117(11):6163–6169.
- [Kanwisher and Yovel, 2006] Kanwisher, N. and Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1476):2109–2128.
- [Kaplan and Oudeyer, 2007] Kaplan, F. and Oudeyer, P.-Y. (2007). In search of the neural circuits of intrinsic motivation. *Frontiers in Neuroscience*, 1:225–236.

- [Karuza et al., 2017] Karuza, E. A., Emberson, L. L., Roser, M. E., Cole, D., Aslin, R. N., and Fiser, J. (2017). Neural signatures of spatial statistical learning: Characterizing the extraction of structure from complex visual scenes. *Journal of Cognitive Neuroscience*, 29(12):1963–1976.
- [Katzhender and Weinshall, 2019] Katzhender, G. and Weinshall, D. (2019). Potential upside of high initial visual acuity? *Proceedings of the National Academy of Sciences*, 116(38):18765–18766.
- [Kidd and Hayden, 2015] Kidd, C. and Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, 88(3):449–460.
- [Kidd et al., 2012] Kidd, C., Piantadosi, S. T., and Aslin, R. N. (2012). The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS ONE*, 7(5):e36399.
- [Kirkham et al., 2002] Kirkham, N. Z., Slemmer, J. A., and Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, 83(2):B35–B42.
- [Konkle and Caramazza, 2013] Konkle, T. and Caramazza, A. (2013). Tripartite organization of the ventral stream by animacy and object size. *Journal of Neuroscience*, 33(25):10235–10242.
- [Kriegeskorte, 2008] Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*.
- [Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.
- [Lau Y., 2021] Lau Y., Sim W., C. K. N. Y. S. Z. (2021). Understanding how noise affects the accuracy of CNN image classification. *Journal of Applied Technology and Innovation*, 5(2).

- [Lawrence, 1952] Lawrence, D. H. (1952). The transfer of a discrimination along a continuum. *Journal of Comparative and Physiological Psychology*, 45(6):511.
- [Leat et al., 2009] Leat, S. J., Yadav, N. K., and Irving, E. L. (2009). Development of visual acuity and contrast sensitivity in children. *Journal of Optometry*, 2(1):19–26.
- [LeCun et al., 2010] LeCun, Y., Cortes, C., and Burges, C. J. (2010). Mnist handwritten digit database. *ATT Labs [Online]*. Available: <http://yann.lecun.com/exdb/mnist>.
- [Leiter et al., 2023] Leiter, C., Zhang, R., Chen, Y., Belouadi, J., Larionov, D., Fresen, V., and Eger, S. (2023). Chatgpt: A meta-analysis after 2.5 months.
- [Lewis et al., 2002] Lewis, T. L., Ellemberg, D., Maurer, D., Wilkinson, F., Wilson, H. R., Dirks, M., and Brent, H. P. (2002). Sensitivity to global form in glass patterns after early visual deprivation in humans. *Vision Research*, 42(8):939–948.
- [Li et al., 2021] Li, J., Xiong, C., and Hoi, S. C. (2021). Learning from noisy data with robust representation learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE.
- [Lund et al., 1993] Lund, J. S., Yoshioka, T., and Levitt, J. B. (1993). Comparison of intrinsic connectivity in different areas of macaque monkey cerebral cortex. *Cerebral Cortex*, 3(2):148–162.
- [Lusk and Mitchel, 2016] Lusk, L. G. and Mitchel, A. D. (2016). Differential gaze patterns on eyes and mouth during audiovisual speech segmentation. *Frontiers in Psychology*, 7.
- [Malcom, 2019] Malcom, D. R. (2019). Curiosity and Knowledge are Vital Components of Curricular Reform. *American Journal of Pharmaceutical Education*, 83(1).
- [Malhotra et al., 2022] Malhotra, G., Dujmović, M., and Bowers, J. S. (2022). Feature blindness: A challenge for understanding and modelling visual object recognition. *PLoS Computational Biology*, 18(5):e1009572.

- [Mandler and McDonough, 1993] Mandler, J. M. and McDonough, L. (1993). Concept formation in infancy. *Cognitive Development*, 8(3):291–318.
- [Manela and Biess, 2022] Manela, B. and Biess, A. (2022). Curriculum learning with hindsight experience replay for sequential object manipulation tasks. *Neural Networks*, 145:260–270.
- [Mareschal et al., 2000] Mareschal, D., French, R. M., and Quinn, P. C. (2000). A connectionist account of asymmetric category learning in early infancy. *Developmental Psychology*, 36(5):635–645.
- [Mareschal and Thomas, 2007] Mareschal, D. and Thomas, M. S. C. (2007). Computational modeling in developmental psychology. *IEEE Transactions on Evolutionary Computation*, 11(2):137–150.
- [Masland, 2012] Masland, R. H. (2012). The neuronal organization of the retina. *Neuron*, 76(2):266–280.
- [Mather and Plunkett, 2011] Mather, E. and Plunkett, K. (2011). Same items, different order: Effects of temporal variability on infant categorization. *Cognition*, 119:438–447.
- [Mayer and Dobson, 1982] Mayer, D. L. and Dobson, V. (1982). Visual acuity development in infants and young children, as assessed by operant preferential looking. *Vision Research*, 22(9):1141–1151.
- [McCloskey and Cohen, 1989] McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, pages 109–165. Elsevier.
- [McGurk and MacDonald, 1976] McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588):746–748.
- [McKyton et al., 2015] McKyton, A., Ben-Zion, I., Doron, R., and Zohary, E. (2015). The limits of shape recognition following late emergence from blindness. *Current*

*Biology*, 25(18):2373–2378.

- [McLuhan, 1970] McLuhan, M. (1970). *Culture is Our Business*. McGraw-Hill.
- [McRae, 1993] McRae, K. (1993). Catastrophic interference is eliminated in pretrained networks.
- [Mehta et al., 2021] Mehta, S. V., Patil, D., Chandar, S., and Strubell, E. (2021). An empirical investigation of the role of pre-training in lifelong learning. In *ICML CL Workshop*.
- [Mitchel et al., 2022] Mitchel, A. D., Lusk, L. G., Wellington, I., and Mook, A. T. (2022). Segmenting speech by mouth: The role of oral prosodic cues for visual speech segmentation. *Language and Speech*, page 002383092211376.
- [Mitchel and Weiss, 2011] Mitchel, A. D. and Weiss, D. J. (2011). Learning across senses: Cross-modal effects in multisensory statistical learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(5):1081–1091.
- [Mordan et al., 2018] Mordan, T., THOME, N., Henaff, G., and Cord, M. (2018). Revisiting Multi-Task Learning with ROCK: a Deep Residual Auxiliary Block for Visual Detection. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 1310–1322. Curran Associates, Inc.
- [Moreira et al., 2014] Moreira, R. S., Magalhães, L. C., and Alves, C. R. (2014). Effect of preterm birth on motor development, behavior, and school performance of school-age children: a systematic review. *Jornal de Pediatria*, 90(2):119–134.
- [Mummadi et al., 2021] Mummadi, C. K., Subramaniam, R., Hutmacher, R., Vitay, J., Fischer, V., and Metzen, J. H. (2021). Does enhanced shape bias improve neural network robustness to common corruptions?
- [Munakata and McClelland, 2003] Munakata, Y. and McClelland, J. L. (2003). Connectionist models of development. *Developmental Science*, 6(4):413–429.

- [Murayama et al., 2019] Murayama, K., FitzGibbon, L., and Sakaki, M. (2019). Process account of curiosity and interest: A reward-learning perspective. *Educational Psychology Review*, 31(4):875–895.
- [Murty et al., 2020] Murty, N. A. R., Teng, S., Beeler, D., Mynick, A., Oliva, A., and Kanwisher, N. (2020). Visual experience is not necessary for the development of face-selectivity in the lateral fusiform gyrus. *Proceedings of the National Academy of Sciences*, 117(37):23011–23020.
- [Nagatsuka et al., 2021] Nagatsuka, K., Broni-Bediako, C., and Atsumi, M. (2021). Pre-training a BERT with curriculum learning by increasing block-size of input text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 989–996, Held Online. INCOMA Ltd.
- [Newport and Aslin, 2004] Newport, E. L. and Aslin, R. N. (2004). Learning at a distance i. statistical learning of non-adjacent dependencies. *Cognitive Psychology*, 48(2):127–162.
- [Norcia and Tyler, 1985] Norcia, A. M. and Tyler, C. W. (1985). Infant VEP acuity measurements: Analysis of individual differences and measurement error. *Electroencephalography and Clinical Neurophysiology*, 61(5):359–369.
- [on the Developing Child, 2007] on the Developing Child, N. S. C. (2007). The timing and quality of early experiences combine to shape brain architecture: Working paper no. 5. Retrieved from <http://developingchild.harvard.edu>.
- [Onnis et al., 2015] Onnis, L., Destrebecqz, A., Christiansen, M. H., Chater, N., and Cleeremans, A. (2015). Implicit learning of non-adjacent dependencies. In *Implicit and Explicit Learning of Languages*, pages 213–246. John Benjamins Publishing Company.
- [Oudeyer et al., 2016] Oudeyer, P.-Y., Gottlieb, J., and Lopes, M. (2016). Intrinsic motivation, curiosity, and learning. In *Motivation - Theory, Neurobiology and Applications*, pages 257–284. Elsevier.



- [Oudeyer and Kaplan, 2009] Oudeyer, P.-Y. and Kaplan, F. (2009). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics*, 1.
- [Parisi et al., 2019] Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71.
- [Pascalis and Slater, 2003] Pascalis, O. and Slater, A. (2003). The development of face processing in infancy and early childhood: current perspectives. In *Developmental Review*, volume 23, pages 229–246. Elsevier.
- [Pelz and Kidd, 2020] Pelz, M. and Kidd, C. (2020). The elaboration of exploratory play. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1803):20190503.
- [Pentina et al., 2014] Pentina, A., Sharmanska, V., and Lampert, C. H. (2014). Curriculum Learning of Multiple Tasks. *arXiv:1412.1353 [cs, stat]*. arXiv: 1412.1353.
- [Perruchet and Pacton, 2006] Perruchet, P. and Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, 10(5):233–238.
- [Poli et al., 2022] Poli, F., Meyer, M., Mars, R. B., and Hunnius, S. (2022). Contributions of expected learning progress and perceptual novelty to curiosity-driven exploration. *Cognition*, 225:105119.
- [Polu et al., 2023] Polu, S., Han, J. M., Zheng, K., Baksys, M., Babuschkin, I., and Sutskever, I. (2023). Formal mathematics statement curriculum learning. In *The Eleventh International Conference on Learning Representations*.
- [Quinn et al., 1993] Quinn, P. C., Eimas, P. D., and Rosenkrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception*, 22(4):463–475.

- [Quinn et al., 2000] Quinn, P. C., Johnson, M. H., Mareschal, D., Rakison, D. H., and Younger, B. A. (2000). Understanding early categorization: One process or two? *Infancy*, 1(1):111–122.
- [Rakic, 1981] Rakic, P. (1981). Development of visual centers in the primate brain depends on binocular competition before birth. *Science*, 214(4523):928–931.
- [Rakison et al., 2003] Rakison, D. H., Oakes, L. M., et al. (2003). *Early category and concept development: Making sense of the blooming, buzzing confusion*. Oxford University Press.
- [Ratcliff, 1990] Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, 97(2):285–308.
- [Rescorla, 1981] Rescorla, L. A. (1981). Category development in early language. *Journal of child language*, 8(2):225–238.
- [Ritchie and Tucker-Drob, 2018] Ritchie, S. J. and Tucker-Drob, E. M. (2018). How Much Does Education Improve Intelligence? A Meta-Analysis. *Psychological Science*, 29(8):1358–1369.
- [Robbins et al., 2011] Robbins, R. A., Shergill, Y., Maurer, D., and Lewis, T. L. (2011). Development of sensitivity to spacing versus feature changes in pictures of houses: Evidence for slow development of a general spacing detection mechanism? *Journal of Experimental Child Psychology*, 109(3):371–382.
- [Rogers and McClelland, 2008] Rogers, T. T. and McClelland, J. L. (2008). Précis of semantic cognition: A parallel distributed processing approach/i. *Behavioral and Brain Sciences*, 31(6):689–714.
- [Rogoff, 1990] Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context*. Oxford University Press.

- [Rosch et al., 1976] Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439.
- [Ruder, 2017] Ruder, S. (2017). An overview of multi-task learning in deep neural networks.
- [Ruderman, 1994] Ruderman, D. L. (1994). The statistics of natural images. *Network: Computation in Neural Systems*, 5(4):517–548.
- [Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.
- [Ryan and Deci, 2000] Ryan, R. M. and Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68–78.
- [Saffran et al., 1996] Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.
- [Saglietti et al., 2022] Saglietti, L., Mannelli, S. S., and Saxe, A. (2022). An analytical theory of curriculum learning in teacher-student networks, journal = Journal of Statistical Mechanics: Theory and Experiment. 2022(11):114014.
- [Schenk and McIntosh, 2010] Schenk, T. and McIntosh, R. D. (2010). Do we have independent visual streams for perception and action? *Cognitive Neuroscience*, 1(1):52–62.
- [Schmidhuber, 1991] Schmidhuber, J. (1991). *Curiosity and exploration*, pages 17–37. Cognitive Systems.
- [Schmidhuber, 2009] Schmidhuber, J. (2009). Simple algorithmic theory of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. *Journal of SICE*, 48:21–32.

- [Shi et al., 2020] Shi, M., Zhang, T., and Zeng, Y. (2020). A curiosity-based learning method for spiking neural networks. *Frontiers in Computational Neuroscience*, 14.
- [Shonkoff and Phillips, 2000] Shonkoff, J. P. and Phillips, D. A. (2000). *From neurons to neighborhoods: The science of early childhood development*. National Academies Press.
- [Siegelman and Frost, 2015] Siegelman, N. and Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language*, 81:105–120.
- [Simion and Giorgio, 2015] Simion, F. and Giorgio, E. D. (2015). Face perception and processing in early infancy: inborn predispositions and developmental changes. *Frontiers in Psychology*, 6.
- [Simoncelli, 2003] Simoncelli, E. P. (2003). Vision and the statistics of the visual environment. *Current Opinion in Neurobiology*, 13(2):144–149.
- [Simoncelli and Olshausen, 2001] Simoncelli, E. P. and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216.
- [Sinz et al., 2019] Sinz, F. H., Pitkow, X., Reimer, J., Bethge, M., and Tolias, A. S. (2019). Engineering a less artificial intelligence. *Neuron*, 103(6):967–979.
- [Skelton et al., 2022] Skelton, A. E., Maule, J., and Franklin, A. (2022). Infant color perception: Insight into perceptual development. *Child Development Perspectives*, 16(2):90–95.
- [Slater et al., 2010] Slater, A., Riddell, P., Quinn, P. C., Pascalis, O., Lee, K., and Kelly, D. J. (2010). Visual perception. In *The Wiley-Blackwell Handbook of Infant Development*, pages 40–80. Wiley-Blackwell.
- [Slater, 1995] Slater, A. M. (1995). Visual perception and memory at birth.

- [Slemmer et al., 2010] Slemmer, J. A., Kirkham, N. Z., and Johnson, S. P. (2010). Visual statistical learning in infancy. *Journal of Vision*, 1(3):25–25.
- [Slone et al., 2019] Slone, L. K., Smith, L. B., and Yu, C. (2019). Self-generated variability in object images predicts vocabulary growth. *Developmental Science*, 22(6).
- [Spelke, 2000] Spelke, E. S. (2000). Core knowledge. *American Psychologist*, 55(11):1233–1243.
- [Srivastava et al., 2014] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.
- [Sullivan et al., 2021] Sullivan, J., Mei, M., Perfors, A., Wojcik, E., and Frank, M. C. (2021). SAYCam: A large, longitudinal audiovisual dataset recorded from the infant’s perspective. *Open Mind*, 5:20–29.
- [Szegedy et al., 2013] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. In *International Conference on Learning Representations*.
- [Thomas and Johnson, 2006] Thomas, M. S. C. and Johnson, M. H. (2006). The computational modeling of sensitive periods. *Developmental Psychobiology*, 48(4):337–344.
- [Thrun, 1995] Thrun, S. (1995). *Exploration in active learning*, pages 381–384. MIT Press.
- [Todorović, 2009] Todorović, D. (2009). The effect of face eccentricity on the perception of gaze direction. *Perception*, 38(1):109–132.
- [Torralba and Oliva, 2003] Torralba, A. and Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, 14(3):391–412.
- [Tsao and Livingstone, 2008] Tsao, D. Y. and Livingstone, M. S. (2008). Mechanisms of face perception. *Annual Review of Neuroscience*, 31(1):411–437.

- [Turk-Browne et al., 2005] Turk-Browne, N. B., Jungé, J. A., and Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General*, 134(4):552–564.
- [Twomey and Westermann, 2017a] Twomey, K. E. and Westermann, G. (2017a). Curiosity-based learning in infants: a neurocomputational approach. *Developmental Science*, 21(4):e12629.
- [Twomey and Westermann, 2017b] Twomey, K. E. and Westermann, G. (2017b). Learned labels shape pre-speech infants' object representations. *Infancy*, 23(1):61–73.
- [van den Hurk et al., 2017] van den Hurk, J., Baelen, M. V., and de Beeck, H. P. O. (2017). Development of visual category selectivity in ventral visual cortex does not require visual experience. *Proceedings of the National Academy of Sciences*, 114(22).
- [van Koningsbruggen et al., 2013] van Koningsbruggen, M. G., Peelen, M. V., and Downing, P. E. (2013). A causal role for the extrastriate body area in detecting people in real-world scenes. *The Journal of Neuroscience*, 33(16):7003–7010.
- [Van Lier, 2004] Van Lier, L. (2004). The ecology and semiotics of language learning: A sociocultural perspective. *The ecology of language learning*, 4:285–296.
- [Vlasiuk and Asari, 2021] Vlasiuk, A. and Asari, H. (2021). Feedback from retinal ganglion cells to the inner retina. *PLoS ONE*, 16(7):e0254611.
- [Vogelsang et al., 2018] Vogelsang, L., Gilad-Gutnick, S., Ehrenberg, E., Yonas, A., Diamond, S., Held, R., and Sinha, P. (2018). Potential downside of high initial visual acuity. *Proceedings of the National Academy of Sciences*, 115(44):11333–11338.
- [Vygotsky, 1978] Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- [Weiland et al., 2013] Weiland, C., Ulvestad, K., Sachs, J., and Yoshikawa, H. (2013). Associations between classroom quality and children's vocabulary and executive func-

- tion skills in an urban public prekindergarten program. *Early Childhood Research Quarterly*, 28(2):199–209.
- [Weiland and Yoshikawa, 2013] Weiland, C. and Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development*, 84(6):2112–2130.
- [Weisberg et al., 2016] Weisberg, D. S., Hirsh-Pasek, K., Golinkoff, R. M., Kittredge, A. K., and Klahr, D. (2016). Guided play: Principles and practices. *Current Directions in Psychological Science*, 25(3):177–182.
- [Weiss et al., 2015] Weiss, D. J., Poepsel, T., and Gerfen, C. (2015). Tracking multiple inputs. In *Implicit and Explicit Learning of Languages*, pages 167–190. John Benjamins Publishing Company.
- [Weng et al., 2001] Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., and Thelen, E. (2001). Autonomous mental development by robots and animals. *Science*, 291(5504):599–600.
- [Westermann and Mareschal, 2004] Westermann, G. and Mareschal, D. (2004). From parts to wholes: Mechanisms of development in infant visual object processing. *Infancy*, 5(2):131–151.
- [Westermann and Mareschal, 2012] Westermann, G. and Mareschal, D. (2012). Mechanisms of developmental change in infant categorization. *Cognitive Development*, 27(4):367–382.
- [Westermann and Mareschal, 2014] Westermann, G. and Mareschal, D. (2014). From perceptual to language-mediated categorization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1634):20120391.
- [White, 1959] White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological Review*, 66(5):297–333.

- [Wichmann and Geirhos, 2023] Wichmann, F. A. and Geirhos, R. (2023). Are deep neural networks adequate behavioural models of human visual perception?
- [Wilson et al., 2019] Wilson, R. C., Shenhav, A., Straccia, M., and Cohen, J. D. (2019). The eighty five percent rule for optimal learning. *Nature Communications*, 10(1):4646.
- [Wood et al., 1976] Wood, D., Bruner, J. S., and Ross, G. (1976). The Role of Tutoring in Problem Solving\*. *Journal of Child Psychology and Psychiatry*, 17(2):89–100.
- [Wood and Hedges, 2016] Wood, E. and Hedges, H. (2016). Curriculum in early childhood education: critical questions about content, coherence, and control. *The Curriculum Journal*, 27(3):387–405.
- [Wu et al., 2020] Wu, X., Dyer, E., and Neyshabur, B. (2020). When do curricula work? In *International Conference on Learning Representations*.
- [Yamins et al., 2014] Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624.
- [Yang et al., 2019] Yang, G. R., Cole, M. W., and Rajan, K. (2019). How to study the neural mechanisms of multiple tasks. *Current Opinion in Behavioral Sciences*, 29:134–143.
- [Yeterian et al., 2012] Yeterian, E. H., Pandya, D. N., Tomaiuolo, F., and Petrides, M. (2012). The cortical connectivity of the prefrontal cortex in the monkey brain. *Cortex*, 48(1):58–81.
- [Yuodelis and Hendrickson, 1986] Yuodelis, C. and Hendrickson, A. (1986). A qualitative and quantitative analysis of the human fovea during development. *Vision Research*, 26(6):847–855.
- [Zaadnoordijk et al., 2022] Zaadnoordijk, L., Besold, T. R., and Cusack, R. (2022). Lessons from infant learning for unsupervised machine learning. *Nature Machine*



*Intelligence*, 4(6):510–520.

- [Zaremba and Sutskever, 2014] Zaremba, W. and Sutskever, I. (2014). Learning to execute.
- [Zeiler and Fergus, 2014] Zeiler, M. D. and Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 818–833, Cham. Springer International Publishing.
- [Zhang et al., 2014] Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2014). Facial Landmark Detection by Deep Multi-task Learning. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 94–108, Cham. Springer International Publishing.
- [Zhuang et al., 2019] Zhuang, C., Yan, S., Nayebi, A., and Yamins, D. (2019). Self-supervised neural network models of higher visual cortex development. In *2019 Conference on Cognitive Computational Neuroscience*. Cognitive Computational Neuroscience.

# A1 Statistical properties of the world influencing category learning

## A1.1 Movies used in the experiment

Taxi Driver (1976), directed by Martin Scorsese

A Date for Mad Mary (2016), directed by Darren Thornton

The Notebook (2004), directed by Nick Cassavetes

13 Going on 30 (2004), directed by Gary Winick

Bridget Jones: The Edge of Reason (2004), directed by Beeban Kidron

Step Up 2: The Streets (2008), directed by Jon M. Chu

Bend It Like Beckham (2002), directed by Gurinder Chadha

Black Swan (2010), directed by Darren Aronofsky

The Bourne Ultimatum (2007), directed by Paul Greengrass

Annie Hall (1977), directed by Woody Allen

The Van (1996), directed by Stephen Frears

Milk (2008), directed by Gus Van Sant

Clueless (1995), directed by Amy Heckerling

Quantum of Solace (2008), directed by Marc Forster

Jane Eyre (2011), directed by Cary Fukunaga

Of Time and the City (2008), directed by Terence Davies

Step Up (2006), directed by Anne Fletcher

Ocean's Twelve (2004), directed by Steven Soderbergh

Adaptation (2002), directed by Spike Jonze

When We Were Kings (1996), directed by Leon Gast  
Fight Club (1999), directed by David Fincher  
Working Girl (1988), directed by Mike Nichols  
Wedding Crashers (2005), directed by David Dobkin  
Ocean's Thirteen (2007), directed by Steven Soderbergh  
Brooklyn (2015), directed by John Crowley  
The Da Vinci Code (2006), directed by Ron Howard  
Metamorphosis (2012), directed by Chris Swanton  
Paris, Texas (1984), directed by Wim Wenders  
The Old Man and the Sea (1958), directed by John Sturges  
Blood Diamond (2006), directed by Edward Zwick  
Casino Royale (2006), directed by Martin Campbell  
A Passage to India (1984), directed by David Lean  
Mamma Mia! (2008), directed by Phyllida Lloyd  
45 Years (2015), directed by Andrew Haigh  
Marley & Me (2008), directed by David Frankel  
I, Daniel Blake (2016), directed by Ken Loach  
Catfish (2010), directed by Ariel Schulman and Henry Joost  
Little Miss Sunshine (2006), directed by Jonathan Dayton and Valerie Faris  
West Side Story (1961), directed by Robert Wise and Jerome Robbins  
To Kill a Mockingbird (1962), directed by Robert Mulligan  
Chariots of Fire (1981), directed by Hugh Hudson  
A Little Princess (1995), directed by Alfonso Cuarón  
The Tempest (2010), directed by Julie Taymor  
The Break-Up (2006), directed by Peyton Reed  
Amélie (2001), directed by Jean-Pierre Jeunet  
The Social Network (2010), directed by David Fincher  
Moonlight (2016), directed by Barry Jenkins  
Youth (2015), directed by Paolo Sorrentino  
The Little Rascals (1994), directed by Penelope Spheeris  
In Bruges (2008), directed by Martin McDonagh  
Memento (2000), directed by Christopher Nolan

Moulin Rouge! (2001), directed by Baz Luhrmann  
Belle Epoque (1992), directed by Fernando Trueba  
The Bourne Supremacy (2004), directed by Paul Greengrass  
Angus (1995), directed by Patrick Read Johnson  
The Bourne Identity (2002), directed by Doug Liman  
Muriel's Wedding (1994), directed by P.J. Hogan  
High School Musical (2006), directed by Kenny Ortega  
Pretty Woman (1990), directed by Garry Marshall  
Pluto (2012), directed by Shin Su-won  
Layer Cake (2004), directed by Matthew Vaughn  
Fame (1980), directed by Alan Parker  
Grease (1978), directed by Randal Kleiser  
Gone Girl (2014), directed by David Fincher  
Pulp Fiction (1994), directed by Quentin Tarantino  
The Italian Job (1969), directed by Peter Collinson  
The Caiman (2006), directed by Nanni Moretti  
Desperately Seeking Susan (1985), directed by Susan Seidelman  
P.S. I Love You (2007), directed by Richard LaGravenese  
Hairspray (2007), directed by Adam Shankman  
The Lion in Winter (1968), directed by Anthony Harvey  
Forrest Gump (1994), directed by Robert Zemeckis  
Bridget Jones's Diary (2001), directed by Sharon Maguire  
High School Musical (2006), directed by Kenny Ortega  
Hannah and Her Sisters (1986), directed by Woody Allen  
The Pyjama Girl Case (1977), directed by Flavio Mogherini  
The Proposal (2009), directed by Anne Fletcher  
Sing Street (2016), directed by John Carney  
Tangerine (2015), directed by Sean Baker  
Three Billboards Outside Ebbing, Missouri (2017), directed by Martin McDonagh  
The Secret Garden (1993), directed by Agnieszka Holland  
Mean Girls (2004), directed by Mark Waters  
Synecdoche, New York (2008), directed by Charlie Kaufman

Tomboy (2011), directed by Céline Sciamma  
Family (2018), directed by Laura Steinel  
The Truman Show (1998), directed by Peter Weir  
The Guard (2011), directed by John Michael McDonagh  
Steel Magnolias (1989), directed by Herbert Ross  
The Blind Side (2009), directed by John Lee Hancock  
Ocean's Eleven (2001), directed by Steven Soderbergh  
Romy and Michele's High School Reunion (1997), directed by David Mirkin  
Inception (2010), directed by Christopher Nolan  
Carol (2015), directed by Todd Haynes  
Spectre (2015), directed by Sam Mendes  
The Breakfast Club (1985), directed by John Hughes  
The King's Speech (2010), directed by Tom Hooper  
Exodus: Gods and Kings (2014), directed by Ridley Scott  
Manchester by the Sea (2016), directed by Kenneth Lonergan  
27 Dresses (2008), directed by Anne Fletcher  
Strange Occurrences in a Small Irish Village (2016), directed by Aoife Kelleher

## A1.2 Examples of categories with low/high tf-idf

### Low tf-idf

Human 0.0037611930063829423  
Person 0.0037611930063829423  
Apparel 0.00476986067428842  
Clothing 0.00476986067428842  
Face 0.006418157141992042  
Indoors 0.006464493825701375  
Room 0.007678588772509666  
Furniture 0.008126871268895624  
Hair 0.008489913113420188  
Chair 0.008849104204284964  
Coat 0.009339087399891526  
People 0.009936079360272129

Man 0.010016086391835497  
Home Decor 0.010205167841456586  
Female 0.01025476307203354  
Accessory 0.010261305624797056  
Accessories 0.010261330677057394  
Pub 0.010396506096838447  
Cushion 0.010426368409340048  
Photo 0.010434915619629878  
Photography 0.010434915619629878  
Crowd 0.010564866516449986  
Restaurant 0.010771969405520308  
Food 0.010848654778090569  
Bar Counter 0.010869364631198213  
Jacket 0.010918130676116277  
Electronics 0.011082670191040712  
Table 0.011133786474876978  
Screen 0.011223612570446078  
Blanket 0.011519387715256355  
Door 0.011721795668089492

### **High tf-idf**

Yak 0.4642857142857143  
Lotion 0.46545454545454545  
Swordfish 0.46621621621621623  
Sea Lion 0.4675324675324675  
Squash 0.46875  
Eskimo Dog 0.472636815920398  
Sheet Music 0.47368421052631576  
Rocking Chair 0.4766081871345029  
Onion 0.48  
Deodorant 0.48148148148148145  
Canary 0.4838709677419355  
Dung Beetle 0.49333333333333335

Pickle 0.5  
Peacock 0.5  
Dolphin 0.5026178010471204  
Seal 0.5087719298245614  
Bow Tie 0.5111111111111111  
Wind Turbine 0.5185185185185185  
Relish 0.5238095238095238  
Rooster 0.5294117647058824  
Swallow 0.5357142857142857  
Tree Frog 0.5384615384615384  
Corn 0.5517241379310345  
Stopsign 0.5555555555555556  
Crossword Puzzle 0.56  
Duck 0.5625  
Hourglass 0.5714285714285714  
Screw 0.6  
Snail 0.6153846153846154  
Comb 0.6829268292682927  
Collar 0.75  
Bison 0.7777777777777778  
Wand 0.8  
Honeycomb 0.8  
Bongo 0.8  
Snowflake 0.8666666666666667  
Garlic 0.8888888888888888  
Oatmeal 0.9090909090909091

## A1.3 Ethical Approval



Coláiste na Tríonóide, Baile Átha Cliath  
Trinity College Dublin

Ollscoil Átha Cliath | The University of Dublin

F.A.O. Anna Kravchenko

**Approval ID: SPREC022022-04**

**School of Psychology Research Ethics Committee**

21<sup>st</sup> March 2022

Dear Anna,

The School of Psychology Research Ethics Committee has reviewed your application entitled "Labelling images to gather the statistics of the visual world", and I am pleased to inform you that it was approved.

Please note that you will be required to submit a completed **Project Annual Report Form** on each anniversary of this approval, until such time as the research is complete and the thesis is submitted. The form is available for download from the Ethics section of the School website.

Adverse events associated with the conduct of this research must be reported immediately to the Chair of the Ethics Committee.

Yours sincerely,

Richard Carson  
Chair,  
School of Psychology Research Ethics Committee

**Scoil na Siceolaíochta**

Dámh na nEolaíochtaí Sóisialta agus Daonna,  
Áras an Phiarsaigh,  
Coláiste na Tríonóide,  
Baile Átha Cliath,  
Ollscoil Átha Cliath,  
Baile Átha Cliath 2, Éire.

**School of Psychology**

Faculty of Arts, Humanities and Social Sciences,  
Trinity College Dublin,  
The University of Dublin,  
Dublin 2, Ireland.

+353 1 896 1886  
psychology@tcd.ie  
www.tcd.ie/psychology



## A2 Low initial acuity

### A2.1 Finding optimal kernel size

Increasing the size of the convolutional kernel in layer 1 led to a slight decrease in top 1 and top 5 performance but a slight increase in the shape bias of the networks. A 15x15 pixel kernel resulted in an unusually high shape bias of 28.09%, warranting further investigation. To determine the effect of increased kernel size on trajectories that include a pretraining phase with high spatial frequency information removed, a network was trained with a 21x21 pixel kernel in the first convolutional layer. At epoch 35, this network achieved the highest shape bias observed so far, 33.60%. At epoch 60, the shape bias was 29.45%, which was marginally lower than the equivalent network with a 7x7 pixel kernel. Despite minor differences, the utility of the 21x21 pixel kernel for effective receptive field visualization was not significantly offset by the results observed with the 7x7 pixel kernel.

### A2.2 Performance of networks with only low-acuity training

Networks trained entirely on low-resolution images had low classification accuracy throughout training, with top 1 accuracy below 5%. This was expected since the validation set only contained high-resolution images. These networks were also tested on a validation set of images blurred with a Gaussian blur corresponding to the training images. At epoch 60, classification performance on low-resolution images was about 5% below that

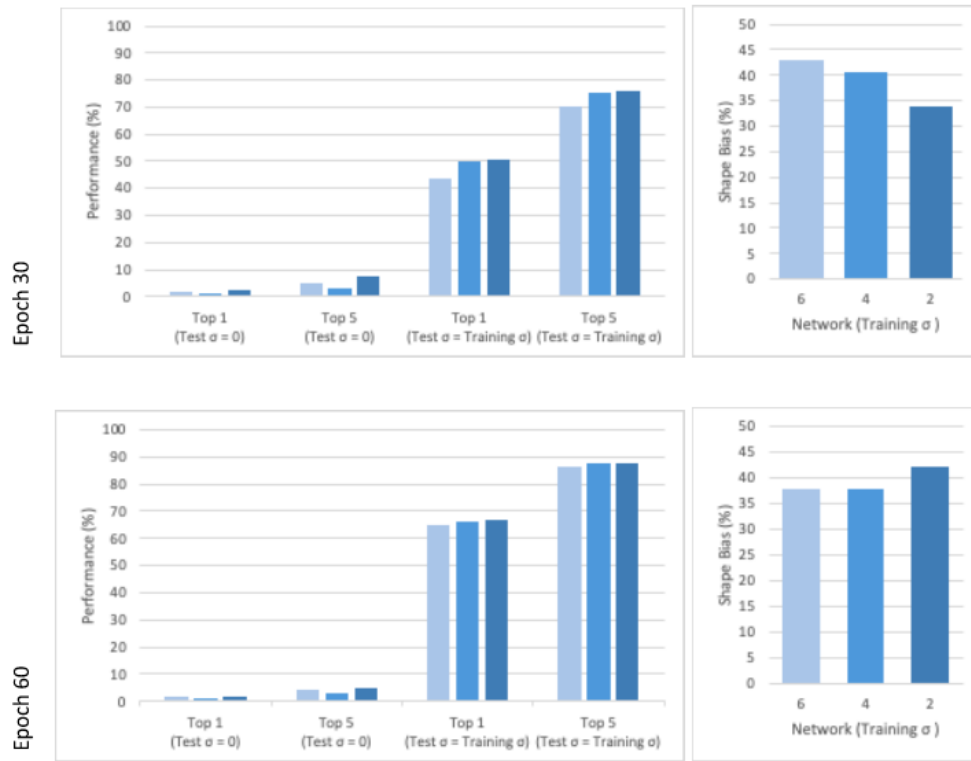


Figure A2.1: **Performance and shape bias of networks trained only on low resolution stimuli.**

of the network trained and tested on high-resolution images. To determine shape bias in networks trained on low-resolution images, Gaussian blur equivalent to the training dataset was applied to the cue conflict stimuli. As anticipated, shape bias was higher than in high-resolution images. Interestingly, the highest shape bias at epoch 60 was observed in the network trained on low-resolution images with the least high spatial frequency information removed, those with a Gaussian blur with sigma of 2.

## A3 Curriculum learning

### A3.1 Effects of the number of of epochs.

Doubling the amount of epochs made the task learnable, even if performance wasn't as good as with the addition of pre-training on Task 1.

With the original amount of epochs performance on Task 2 without pre-training averages to random with respect to some noise. In some learning cases there was a pattern of cycling between 40% and 60% accuracy (we attribute this effect to classes where the same digit repeats twice, the result then depends whether the network associated the visual features of this digit with the odd or even class). Again, in multiple runs this averages to 50% accuracy .

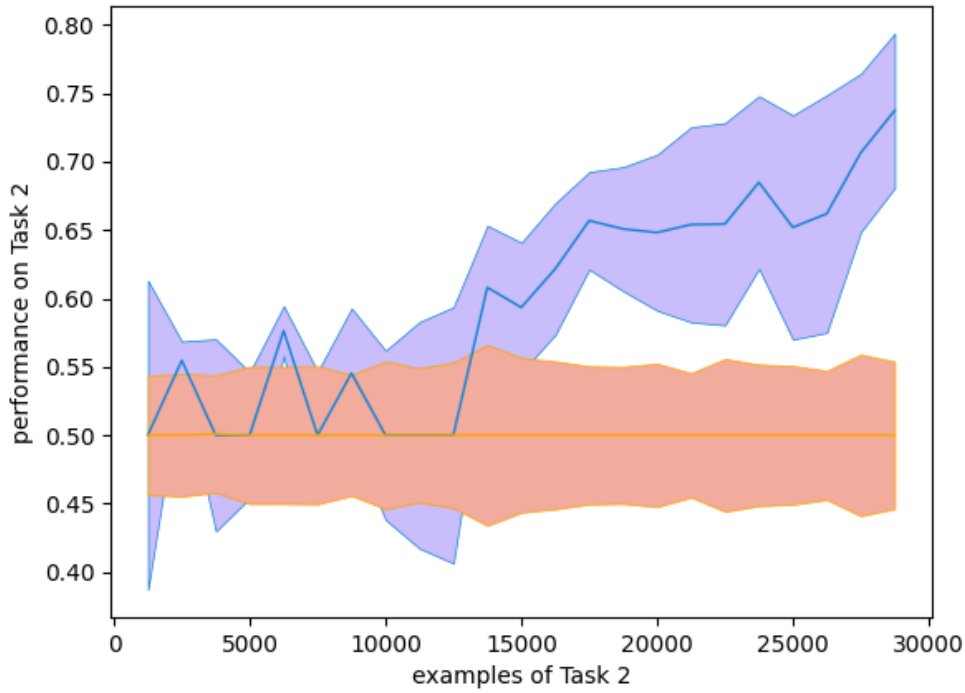


Figure A3.1: Increasing number of epochs.

## A3.2 Catastrophic forgetting.

Class-incremental learning during pre-training learned to a catastrophic forgetting scenario that couldn't have been solved simply with switching at the correct time.

Due to the limitations of architecture imposed by the original study goal, ours isn't a suitable model to deal with true non-stationarity and catastrophic forgetting. Investigating the interplay between curriculum and active learning will be the direction of our future work.

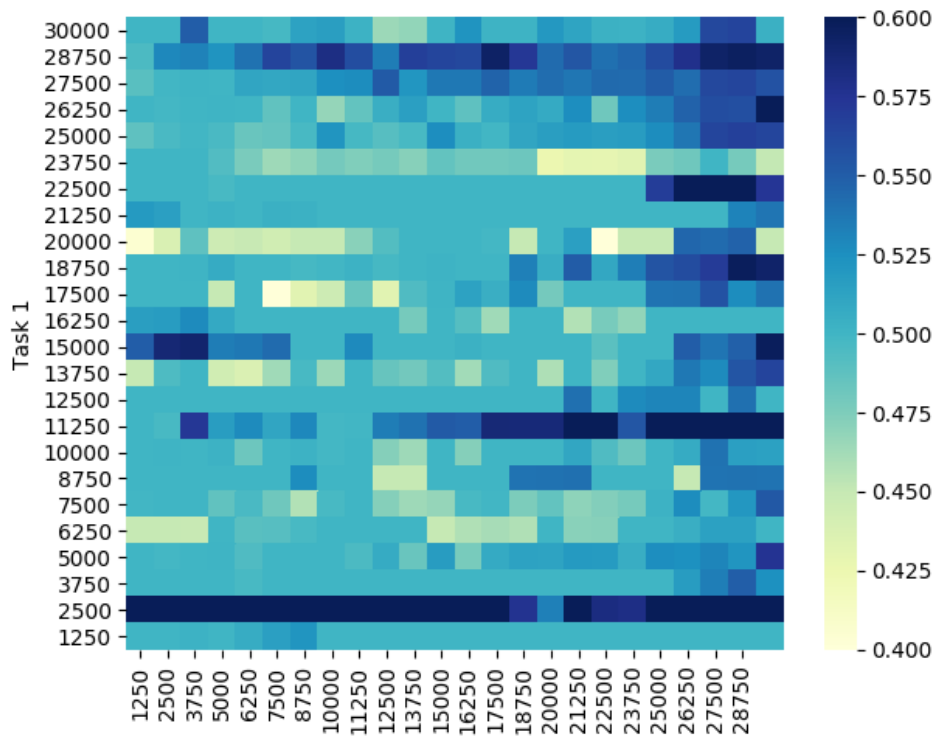


Figure A3.2: Learning trajectories during class-incremental training.