arXiv:1706.08606v2 [stat.ML] 29 Jun 2017

Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study

Samuel Ritter [*1] **David G.T. Barrett** [*1] **Adam Santoro** [1] **Matt M. Botvinick** [1]

Abstract

Deep neural networks (DNNs) have achieved unprecedented performance on a wide range of complex tasks, rapidly outpacing our understanding of the nature of their solutions. This has caused a recent surge of interest in methods for rendering modern neural systems more interpretable. In this work, we propose to address the interpretability problem in modern DNNs using the rich history of problem descriptions, theories and experimental methods developed by cognitive psychologists to study the human mind. To explore the potential value of these tools, we chose a well-established analysis from developmental psychology that explains how children learn word labels for objects, and applied that analysis to DNNs. Using datasets of stimuli inspired by the original cognitive psychology experiments, we find that state-of-the-art one shot learning models trained on ImageNet exhibit a similar bias to that observed in humans: they prefer to categorize objects according to shape rather than color. The magnitude of this shape bias varies greatly among architecturally identical, but differently seeded models, and even fluctuates within seeds throughout training, despite nearly equivalent classification performance. These results demonstrate the capability of tools from cognitive psychology for exposing hidden computational properties of DNNs, while concurrently providing us with a computational model for human word learning.

1. Introduction

During the last half-decade deep learning has significantly improved performance on a variety of tasks (for a review, see LeCun et al. (2015)). However, deep neural network (DNN) solutions remain poorly understood, leaving many to think of these models as black boxes, and to question whether they can be understood at all (Bornstein, 2016; Lipton, 2016). This opacity obstructs both basic research seeking to improve these models, and applications of these models to real world problems (Caruana et al., 2015).

Recent pushes have aimed to better understand DNNs: tailor-made loss functions and architectures produce more interpretable features (Higgins et al., 2016; Raposo et al., 2017) while output-behavior analyses unveil previously opaque operations of these networks (Karpathy et al., 2015). Parallel to this work, neuroscience-inspired methods such as activation visualization (Li et al., 2015), ablation analysis (Zeiler & Fergus, 2014) and activation maximization (Yosinski et al., 2015) have also been applied.

Altogether, this line of research developed a set of promising tools for understanding DNNs, each paper producing a glimmer of insight. Here, we propose another tool for the kit, leveraging methods inspired not by neuroscience, but instead by psychology. Cognitive psychologists have long wrestled with the problem of understanding another opaque intelligent system: the human mind. We contend that the search for a better understanding of DNNs may profit from the rich heritage of problem descriptions, theories, and experimental tools developed in cognitive psychology. To test this belief, we performed a proof-of-concept study on state-of-the-art DNNs that solve a particularly challenging task: one-shot word learning. Specifically, we investigate Matching Networks (MNs) (Vinyals et al., 2016), which have state-of-the-art one-shot learning performance on ImageNet and we investigate an Inception Baseline model (Szegedy et al., 2015a).

Following the approach used in cognitive psychology, we began by hypothesizing an inductive bias our model may use to solve a word learning task. Research in developmental psychology shows that when learning new words, humans tend to assign the same name to similarly shaped

[*]Equal contribution [1]DeepMind, London, UK. Correspondence to: Samuel Ritter <ritters@google.com>, David G.T. Barrett <barrett@google.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

items rather than to items with similar color, texture, or size. To test the hypothesis that our DNNs discover this same "shape bias", we probed our models using datasets and an experimental setup based on the original shape bias studies (Landau et al., 1988).

Our results are as follows: 1) Inception networks trained on ImageNet do indeed display a strong shape bias. 2) There is high variance in the bias between Inception networks initialized with different random seeds, demonstrating that otherwise identical networks converge to qualitatively different solutions. 3) MNs also have a strong shape bias, and this bias closely mimics the bias of the Inception model that provides input to the MN. 4) By emulating the shape bias observed in children, these models provide a candidate computational account for human one-shot word learning. Altogether, these results show that the technique of testing hypothesized biases using probe datasets can yield both expected and surprising insights about solutions discovered by trained DNNs.

1.1. Related Work: Cognitive Modeling with Neural Networks

The use of behavioral probes to understand neural network function has been extensively applied within psychology itself, where neural networks have been employed effectively as models of human cognitive function (Rumelhart et al., 1988; Plaut et al., 1996; Rogers & McClelland, 2004; Mareschal et al., 2000). In contrast, in the present work we are advocating for the application of behavioral probes along with associated theories and hypotheses from cognitive psychology to address the interpretability problem in modern deep networks.

In spite of the widespread adoption of deep learning methods in recent years, to our knowledge, work applying behavioral probes to DNNs in machine learning for this purpose has been quite limited; we only are aware of Zoran et al. (2015) and Goodfellow et al. (2009), who used psychophysics-like experiments to better understand image processing models.

2. Inductive Biases, Statistical Learners and Probe Datasets

Before we delve into the specifics of the shape bias and one-shot word learning, we will describe our approach in the general context of inductive biases, probe datasets, and statistical learning. Suppose we have some data $\{y_i, x_i\}_{i=1}^N$ where $y_i = f(x_i)$. Our goal is to build a model of the data $g(.)$ to optimize some loss function L measuring the disparity between y and $g(x)$, e.g., $L = \sum_i ||y_i - g(x_i)||^2$. Perhaps this data x is images of ImageNet objects to be classified, images and histology of tumors to be classified as benign or malignant (Kourou et al., 2015), or medical history and vital measurements to be classified according to likely pneumonia outcomes (Caruana et al., 2015).

A statistical learner such as a DNN will minimize L by discovering properties of the input x that are predictive of the labels y. These discovered predictive properties are, in effect, the properties of x for which the *trained* model has an inductive bias. Examples of such properties include the shape of ImageNet objects, the number of nodes of a tumor, or a particular constellation of blood test values that often precedes an exacerbation of pneumonia symptoms.

Critically, in real-world datasets such as these, the discovered properties are unlikely to correspond to a single feature of the input x; instead they correspond to complex conjunctions of those features. We could describe one of these properties using a function $h(x)$, which, for example, returns the shape of the focal object given an ImageNet image, or the number of nodes given a scan of tumor. Indeed, one way to articulate the difficulty in understanding DNNs is to say that we often can't intuitively describe these conjunctions of features $h(x)$; although we often have numerical representations in intermediate DNN layers, they're often too arcane for us to interpret.

We advocate for addressing this problem using the following hypothesis-driven approach: First, propose a property $h_p(x)$ that the model may be using. Critically, it's not necessary that $h_p(x)$ be a function that can be evaluated using an automated method. Instead, the intention is that $h_p(x)$ is a function that humans (e.g. ML researchers and practitioners) can intuitively evaluate. $h_p(x)$ should be a property that is believed to be relevant to the problem, such as object shape or number of tumor nodes.

After proposing a property, the next step is to generate predictions about how the model should behave when given various inputs, if in fact it uses a bias with respect to the property $h_p(x)$. Then, construct and carry out an experiment wherein those predictions are tested. In order to execute such an experiment, it typically will be necessary to craft a set of probe examples x that cover a relevant portion of the range of $h_p(x)$, for example a variety of object shapes. The results of this experiment will either support or fail to support the hypothesis that the model uses $h_p(x)$ to solve the task. This process can be especially valuable in situations where there is little or no training data available in important regions of the input space, and a practitioner needs to know how the trained model will behave in that region.

Psychologists have developed a repertoire of such hypotheses and experiments in their effort to understand the human mind. Here we explore the application of one of these theory-experiment pairs to state of the art one-shot learning

models. We will begin by describing the historical backdrop for the human one-shot word learning experiments that we will then apply to our DNNs.

3. The problem of word learning; the solution of inductive biases

Discussions of one-shot word learning in the psychological literature inevitably begin with the philosopher W.V.O. Quine, who broke this problem down and described one of its most computationally challenging components: there are an enormous number of tenable hypotheses that a learner can use to explain a single observed example. To make this point, Quine penned his now-famous parable of the field linguist who has gone to visit a culture whose language is entirely different from our own (Quine, 1960). The linguist is trying to learn some words from a helpful native, when a rabbit runs past. The native declares "gavagai", and the linguist is left to infer the meaning of this new word. Quine points out that the linguist is faced with an abundance of possible inferences, including that "gavagai" refers to rabbits, animals, white things, that specific rabbit, or "undetached parts of rabbits". Quine argues that indeed there is an infinity of possible inferences to be made, and uses this conclusion to bolster the assertion that meaning itself cannot be defined in terms of internal mental events[1].

Contrary to Quine's intentions, when this example was introduced to the developmental psychology community by Macnamara (1972), it spurred them not to give up on the idea of internal meaning, but instead to posit and test for cognitive biases that enable children to eliminate broad swaths of the hypothesis space (Bloom, 2000). A variety of hypothesis-eliminating biases were then proposed including the whole object bias, by which children assume that a word refers to an entire object and not its components (Markman, 1990); the taxonomic bias, by which children assume a word refers to the basic level category an object belongs to (Markman & Hutchinson, 1984); the mutual exclusivity bias, by which children assume that a word only refers to one object category (Markman & Wachtel, 1988); the shape bias, with which we are concerned here (Landau et al., 1988); and a variety of others (Bloom, 2000). These biases were tested empirically in experiments wherein children or adults were given an object (or picture of an object) along with a novel name, then were asked whether the name should apply to various other objects.

Taken as a whole, this work yielded a computational level (Marr, 1982) account of word learning whereby people make use of biases to eliminate unlikely hypotheses when inferring the meaning of new words. Other contrasting and complementary approaches to explaining word learning exist in the psychological literature, including association learning (Regier, 1996; Colunga & Smith, 2005) and Bayesian inference (Xu & Tenenbaum, 2007). We leave the application of these theories to deep learning models to future work, and focus on determining what insight can be gained by applying a hypothesis elimination theory and methodology.

We begin the present work with the knowledge that part of the hypothesis elimination theory is correct: the models surely use some kind of inductive biases since they are statistical learning machines that successfully model the mapping between images and object labels. However, several questions remain open. What predictive properties did our DNNs find? Do all of them find the same properties? Are any of those properties interpretable to humans? Are they the same properties that children use? How do these biases change over the course of training?

To address these questions, we carry out experiments analogous to those of Landau et al. (1988). This enables us to test whether the shape bias – a human interpretable feature used by children when learning language – is visible in the behavior of MNs and Inception networks. Furthermore we are able to test whether these two models, as well as different instances of each of them, display the same bias. In the next section we will describe in detail the one-shot word learning problem, and the MNs and Inception networks we use to solve it.

4. One-shot word learning models and training

4.1. One-shot word learning task

The one-shot word learning task is to label a novel data example $\hat{x}$ (e.g. a novel probe image) with a novel class label $\hat{y}$ (e.g. a new word) after only a single example. More specifically, given a support set $S = \{(x_i, y_i) : i \in [1, k]\}$, of images x_i and their associated labels y_i, and an unlabelled probe image $\hat{x}$, the one-shot learning task is to identify the true label of the probe image $\hat{y}$ from the support set labels $\{y_i : i \in [1, k]\}$:

$$\hat{y} = \arg\max_y P(y|\hat{x}, S). \quad (1)$$

We assume that the image labels y_i are represented using a one-hot encoding and that $P(y|\hat{x}, S)$ is parameterised by a DNN, allowing us to leverage the ability of deep networks to learn powerful representations.

4.2. Inception: baseline one-shot learning model

In our simplest *baseline* one-shot architecture, a probe image $\hat{x}$ is given the label of the nearest neighbour from the

[1]Unlike Quine, we use a pragmatic definition of meaning - a human or model understands the meaning of a word if they assign that word to new instances of objects in the correct category.

support set:

$$\hat{y} = y$$
$$(x, y) = \arg \min_{(x_i, y_i) \in S} d(h(x_i), h(\hat{x})) \quad (2)$$

where d is a distance function. The function h is parameterised by Inception – one of the best performing ImageNet classification models (Szegedy et al., 2015a). Specifically, h returns features from the last layer (the softmax input) of a pre-trained Inception classifier, where the Inception classifier is trained using rms-prop, as described in Szegedy et al. (2015b), section 8. With these features as input and cosine distance as the distance function, the classifier in equation 2 achieves 87.6% accuracy on one-shot classification on the ImageNet dataset (Vinyals et al., 2016). Henceforth, we call the Inception classifier together with the nearest-neighbor component the Inception Baseline (IB) model.

4.3. Matching Nets model architecture and training

We also investigate a state-of-the-art one-shot learning architecture called *Matching Nets* (MN) (Vinyals et al., 2016). MNs are a fully differentiable neural network architecture with state-of-the-art one shot learning performance on ImageNet (93.2% one-shot labelling accuracy).

MNs are trained to assign label $\hat{y}$ to probe image $\hat{x}$ according to equation 1 using an attention mechanism a acting on image embeddings stored in the support set S:

$$a(\hat{x}, x_i) = \frac{e^{d(f(\hat{x},S),g(x_i,S))}}{\sum_j e^{d(f(\hat{x},S),g(x_j,S))}}, \quad (3)$$

where d is a cosine distance and where f and g provide context-dependent embeddings of $\hat{x}$ and x_i (with context S). The embedding $g(x_i, S)$ is a bi-directional LSTM (Hochreiter & Schmidhuber, 1997) with the support set S provided as an input sequence. The embedding $f(\hat{x}, S)$ is an LSTM with a read-attention mechanism operating over the entire embedded support set. The input to the LSTM is given by the penultimate layer features of a pre-trained deep convolutional network, specifically Inception, as in our baseline IB model described above (Szegedy et al., 2015a).

The training procedure for the one-shot learning task is critical if we want MNs to classify a probe image $\hat{x}$ after viewing only a single example of this new image class in its support set (Hochreiter et al., 2001; Santoro et al., 2016).

To train MNs we proceed as follows: (1) At each step of training, the model is given a small support set of images and associated labels. In addition to the support set, the model is fed an unlabelled probe image $\hat{x}$; (2) The model parameters are then updated to improve classification accuracy of the probe image $\hat{x}$ given the support set. Parameters are updated using stochastic gradient descent with a learning rate of 0.1; (3) After each update, the labels $\{y_i : i \in [1, k]\}$ in the training set are randomly re-assigned to new image classes (the label indices are randomly permuted, but the image labels are not changed). This is a critical step. It prevents MNs from learning a consistent mapping between a category and a label. Usually, in classification, this is what we want, but in one-shot learning we want to train our model for classification after viewing a single in-class example from the support set. Formally, our objective function is:

$$L = E_{C \sim T} \left[E_{S \sim C, B \sim C} \left[\sum_{(x,y) \in B} \log P(y|x, S) \right] \right] \quad (4)$$

where T is the set of all possible labelings of our classes, S is a support set sampled with a class labelling $C \sim T$ and B is a batch of probe images and labels, also with the same randomly chosen class labelling as the support set.

Next we will describe the probe datasets we used to test for the shape bias in the IB and MNs after ImageNet training.

5. Data for bias discovery

5.1. Cognitive Psychology Probe Data

The Cognitive Psychology Probe Data (CogPsyc data) that we use consists of 150 images of objects (Figure 1). The images are arranged in triples consisting of a probe image, a shape-match image (that matches the probe in colour but not shape), and a color-match image (that matches the probe in shape but not colour). In the dataset there are 10 triples, each shown on 5 different backgrounds, giving a total of 50 triples.[2]

The images were generously provided by cognitive psychologist Linda Smith. The images are photographs of stimuli used previously in shape bias experiments conducted in the Cognitive Development Lab at Indiana University. The potentially confounding variables of background content and object size are controlled in this dataset.

5.2. Probe Data from the wild

We have also assembled a *real-world* dataset consisting of 90 images of objects (30 triples) collected using Google Image Search. Again, the images are arranged in triples consisting of a probe, a shape-match and a colour-match. For the probe image, we chose images of real objects that are unlikely to appear in standard image datasets such as ImageNet. In this way, our data contains the irregularity

[2] The CogPsyc dataset is available at http://www.indiana.edu/~cogdev/SB_testsets.html

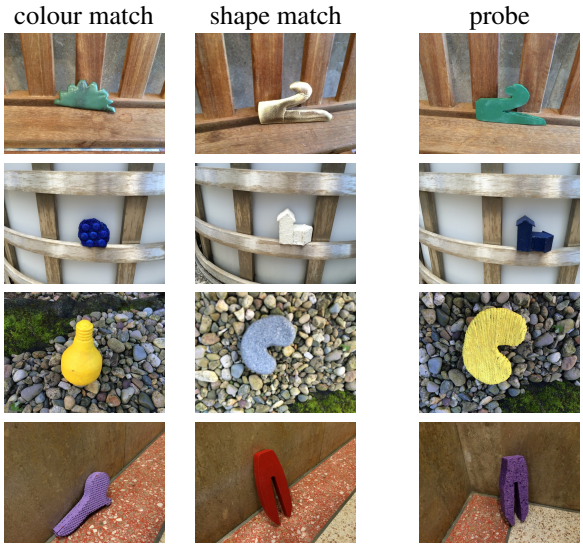

Figure 1. Example images from the Cognitive Psychology Dataset (see section 5). The data consists of image triples (rows), each containing a *colour match* image (left column), a *shape match* image (middle column) and a *probe* image (right column). We use these triples to calculate the shape bias by reporting the proportion of times that a model assigns the shape match image class to the probe image. This dataset was supplied by cognitive psychologist Linda Smith, and was designed to control for object size and background.

of the real world while also probing our models' properties outside of the image space covered in our training data. For the shape-match image, we chose an object with a similar shape (but with a very different colour), and for the colour-match image, we chose an object with a similar colour (but with a very different shape). For example, one triple consists of a silver tuning fork as the probe, a silver guitar capo as the colour match, and a black tuning fork as the shape match. Each photo in the dataset contains a single object on a white background.

We collected this data to strengthen our confidence in the results obtained for the CogPsych dataset and to demonstrate the ease with which such probe datasets can be constructed. One of the authors crafted this dataset solely using Google Image Search in the span of roughly two days' work. Our results with this dataset, especially the fact that the bias pattern over time matches the results from the well established CogPsych dataset, support the contention that DNN practitioners can collect effective probe datasets with minimal time expenditure using readily available tools.

6. Results

6.1. Shape bias in the Inception Baseline Model

First, we measured the shape bias in IB: we used a pre-trained Inception classifier (with 94% top-5 accuracy) to provide features for our nearest-neighbour one-shot classifier, and probed the model using the CogPsyc dataset. Specifically, for a given probe image $\hat{x}$, we loaded the shape-match image x_s and corresponding label y_s, along with the colour-match image x_c and corresponding label y_c into memory, as the support set $S = \{(x_s, y_s), (x_c, y_c)\}$. We then calculated $\hat{y}$ using Equation 2. Our model assigned either y_c or y_s to the probe image. To estimate the shape bias B_s, we calculated the proportion of shape labels assigned to the probe:

$$B_s = E(\delta(\hat{y} - y_s)), \tag{5}$$

where E is an expectation across probe images and δ is the Dirac delta function.

We ran all IB experiments using both Euclidean and cosine distance as the distance function. We found that the results for the two distance functions were qualitatively similar, so we only report results for Euclidean distance.

We found the shape bias of IB to be $B_s = 0.68$. Similarly, the shape bias of IB using our real-world dataset was $B_s = 0.97$. Together, these results strongly suggest that IB trained on ImageNet has a stronger bias towards shape than colour.

Note that, as expected, the shape bias of this model is qualitatively similar across datasets while being quantitatively different - largely because the datasets themselves are quite different. Indeed, the datasets were chosen to be quite different so that we could explore a broad space of possibilities. In particular, our CogPsyc dataset backgrounds have much larger variability than our real-world dataset backgrounds, and our real-world dataset objects have much greater variability than the CogPsyc dataset objects.

6.2. Shape bias in the Matching Nets Model

Next, we probed the MNs using a similar procedure. We used the IB trained in the previous section to provide the input features for the MN as described in section 4.3. Then, following the training procedure outlined in section 4.3 we trained MNs for one-shot word learning on ImageNet, achieving state-of-the-art performance, as reported in (Vinyals et al., 2016). Then, repeating the analysis above, we found that MNs have a shape of bias $B_s = 0.7$ using our CogPsyc dataset and a bias of $B_s = 1$ using the real-world dataset. It is interesting to note that these bias values are very similar to the IB bias values.

6.3. Shape bias statistics: within models and across models

The observation of a shape bias immediately raises some important questions. In particular: (1) Does this bias depend on the initial values of the parameters in our model?

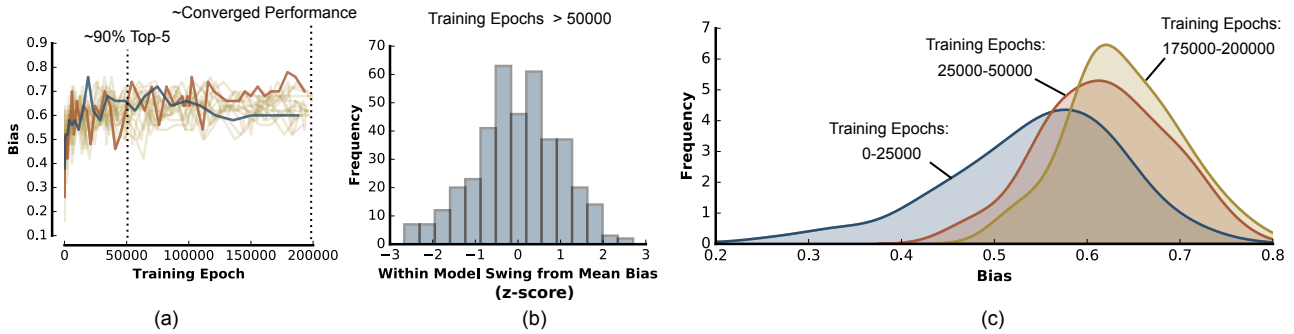

Figure 2. Shape bias across models with different initialization seeds, and within models during training calculated using the CogPsyc dataset. (a) The shape bias B_s of 15 Inception models is calculated throughout training (yellow lines). A strong shape bias emerges across all models. A bias value $B_s > 0.5$ indicates a shape bias and $B_s < 0.5$ indicates a colour bias. Two examples are highlighted here (blue and red lines) for clarity. (b) The shape bias fluctuates strongly within models during training by up to three standard deviations. (c) The distribution of bias values, calculated at the start (blue), middle (red) and end (yellow) of training. Bias variability is high at the start and end of training. Here, these distributions are calculated using kernel density estimates from all shape bias measurements from all models within the indicated window.

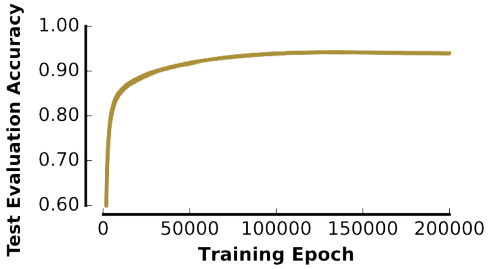

Figure 3. Classification accuracy of all 15 Inception models evaluated on a test set during training on ImageNet (same models as in Figure 2) . All 15 Inception network seeds achieve near identical test accuracy (overlapping yellow lines).

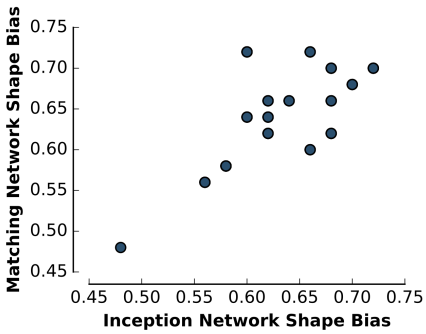

Figure 4. Scatter plot showing Matching Network (MN) bias as a function of Inception bias. Each MN receives input through an Inception model. Each point in this scatter plot is the bias of a MN and the bias of the Inception model providing input to that particular MN. In total, the bias values of 45 MN models are plotted (some dots are overlapping).

(2) Does the size of the shape bias depend on model performance? (3) When does shape bias emerge during training - before model convergence or afterwards? (4) How does shape bias compare between models, and within models?

To answer these questions, we extended the shape bias analysis described above to calculate the shape bias in a population of IB models and in a population of MN models with different random initialization (Figs. 2 and 5).

(1) We first calculated the dependence of shape bias on the initialization of IB (Fig. 2). Surprisingly, we observed a strong variability, depending on the initialization. For the CogPsyc dataset, the average shape bias was $\overline{B}_s = 0.628$ with standard deviation $\sigma_{B_s} = 0.049$ at the end of training and for the real-world dataset the average shape bias was $\overline{B}_s = 0.958$ with $\sigma_{B_s} = 0.037$.

(2) Next, we calculated the dependence of shape bias on model performance. For the CogPsych dataset, the correlation between bias and classification accuracy was $\rho = 0.15$, with $t_{n=15} = 0.55$, $p_{one\_tail} = 0.29$, and for the *real-world* dataset, the correlation was $\rho = -0.06$ with $t_{n=15} = -0.22$, $p_{one\_tail} = 0.42$. Therefore, fluctuations in the bias cannot be accounted for by fluctuations in classification accuracy. This is not surprising, because the classification accuracy of all models was similar at the end of training, while the shape bias was variable. This demonstrates that models can have variable behaviour along important dimensions (e.g., bias) while having the same performance measured by another (e.g., accuracy).

(3) Next we explored the emergence of the shape bias during training (Fig. 2a,c; Fig. 5a,c). At the start of training, the average shape bias of these models was $\overline{B}_s =$

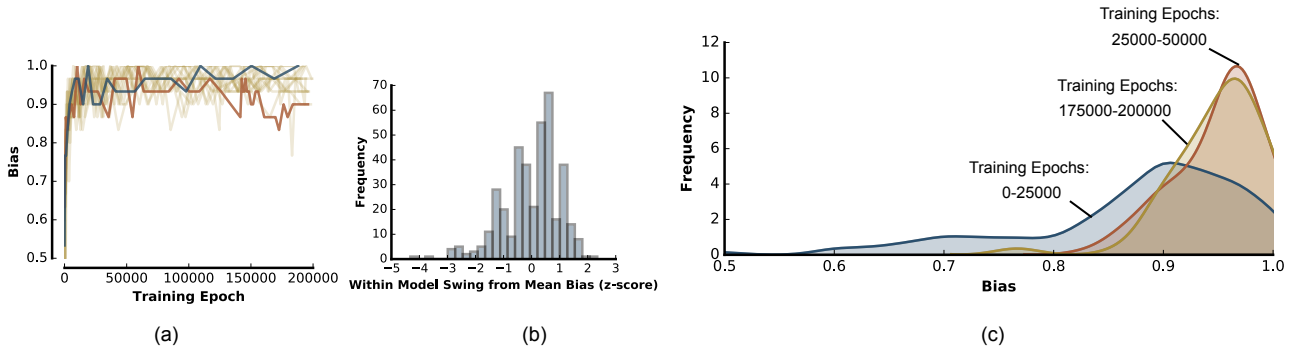

Figure 5. Shape bias across models with different initialization seeds, and within models during training calculated using the real-world dataset. (a) The shape bias B_s of 15 Inception models is calculated throughout training (yellow lines). A strong shape bias emerges across all models. Two examples are highlighted here (blue and red lines) for clarity. (b) The shape bias fluctuates strongly within models during training. (c) The distribution of bias values, calculated at the start (blue), middle (red) and end (yellow) of training. Bias variability is high at the start and end of training.

0.448 with standard deviation $\sigma_{B_s} = 0.0835$ on the CogPsyc dataset and $\overline{B}_s = 0.593$ with $\sigma_{B_s} = 0.073$ on the real-world dataset. We observe that a shape bias began to emerge very early during training, long before convergence.

(4) Finally, we compare shape bias within models during training, and between models at the end of training. During training, the shape bias within IB fluctuates significantly (Fig. 2 b; Fig. 5b). In contrast, the shape bias does not fluctuate during training of the MN. Instead, the MN model inherits its shape bias characteristics at the start of training from the IB that provides it with input embeddings (Fig. 4) and this shape-bias remains constant throughout training. Moreover, there is no evidence that the MN and corresponding IB bias values are different from each other (paired t-test, $p = 0.167$). Note that we do not fine-tune the Inception model providing input while training the MN. We do this so that we can observe the shape-bias properties of the MN independent of the IB model properties.

7. Discussion

7.1. A shape bias case study

Our psychology-inspired approach to understanding DNNs produced a number of insights. Firstly, we found that both IB and MNs trained on ImageNet display a strong shape bias. This is an important result for practitioners who routinely use these models - especially for applications where it is known *a priori* that colour is more important than shape. As an illustrative example, if a practitioner planned to build a one-shot fruit classification system, they should proceed with caution if they plan to use pre-trained ImageNet models like Inception and MNs because fruit are often defined according to colour features rather than shape. In applications where a shape bias is desirable (as is more often the case than not), this result provides reassurance that the models are behaving sensibly in the presence of ambiguity.

The second surprising finding was the large variability in shape bias, both within models during training and across models, depending on the randomly chosen initialisation of our model. This variability can arise because our models are not being explicitly optimised for shape biased categorisation. This is an important result because it shows that not all models are created equally - some models will have a stronger preference for shape than others, even though they are architecturally identical and have almost identical classification accuracy.

Our third finding – that MNs retain the shape bias statistics of the downstream Inception network – demonstrates the possibility for biases to propagate across model components. In this case, the shape bias propagates from the Inception model through to the MN memory modules. This result is yet another cautionary observation; when combining multiple modules together, we must be aware of contamination by unknown properties across modules. Indeed, a bias that is benign in one module might only have a detrimental effect when combined later with other modules.

A natural question immediately arises from these results - how can we remove an unwanted bias or induce a desirable bias? The biases under consideration are properties of an architecture and dataset synthesized together by an optimization procedure. As such, the observation of a shape-bias is partly a result of the statistics of natural image-labellings as captured in the ImageNet dataset, and partly a result of the architecture attempting to extract these statistics. Therefore, on discovering an unwanted bias, a practitioner can either attempt to change the model architecture

to explicitly prevent the bias from emerging, or, they can attempt to manipulate the training data. If neither of these are possible - for example, if the appropriate data manipulation is too expensive, or, if the bias cannot be easily suppressed in the architecture, it may be possible to do zero-th order optimization of the models. For example, one may perform post-hoc model selection either using early stopping or by selecting a suitable model from the set of initial seeds.

An important caveat to note is that behavioral tools often do not provide insight into the neural mechanisms. In our case, the DNN mechanism whereby model parameters and input images interact to give rise to a shape bias have not been elucidated, nor did we expect this to happen. Indeed, just as cognitive psychology often does for neuroscience, our new computational level insights can provide a starting point for research at the mechanistic level. For example, in future work it would be interesting to use gradient-based visualization or neuron ablation techniques to augment the current results by identifying the mechanisms underlying the shape bias. The convergence of evidence from such introspective methods with the current behavioral method would create a richer account of these models' solutions to the one-shot word learning problem.

7.2. Modelling human word learning

There have been previous attempts to model human word learning in the cognitive science literature (Colunga & Smith, 2005; Xu & Tenenbaum, 2007; Schilling et al., 2012; Mayor & Plunkett, 2010). However, none of these models are capable of one-shot word learning on the scale of real-world images. Because MNs both solve the task at scale and emulate hallmark experimental findings, we propose MNs as a computational-level account of human one-shot word learning. Another feature of our results supports this contention: in our model the shape bias increases dramatically early in training (Fig. 2a); similarly, humans show the shape bias much more strongly as adults than as children, and older children show the bias more strongly than younger children (Landau et al., 1988).

As a good cognitive model should, our DNNs make testable predictions about word-learning in humans. Specifically, the current results predict that the shape bias should vary across subjects as well as within a subject over the course of development. They also predict that for humans with adult-level one-shot word learning abilities, there should be no correlation between shape bias magnitude and one-shot-word learning capability.

Another promising direction for future cognitive research would be to probe MNs for additional biases in order to predict novel computational properties in humans. Probing a model in this way is much faster than running human behavioural experiments, so a wider range of hypotheses for human word learning may be rapidly tested.

7.3. Cognitive Psychology for Deep Neural Networks

Through the one-shot learning case study, we demonstrated the utility of leveraging techniques from cognitive psychology for understanding the computational properties of DNNs. There is a wide ranging literature in cognitive psychology describing techniques for probing a spectrum of behaviours in humans. Our work here leads the way to the study of *artificial cognitive psychology* - the application of these techniques to better understand DNNs.

For example, it would be useful to apply work from the massive literature on episodic memory (Tulving, 1985) to the recent flurry of episodic memory architectures (Blundell et al., 2016; Graves et al., 2016), and to apply techniques from the semantic cognition literature (Lamberts & Shanks, 2013) to recent models of concept formation (Higgins et al., 2016; Gregor et al., 2016; Raposo et al., 2017). More generally, the rich psychological literature will become increasingly useful for understanding deep reinforcement learning agents as they learn to solve increasingly complex tasks.

8. Conclusion

In this work, we have demonstrated how techniques from cognitive psychology can be leveraged to help us better understand DNNs. As a case study, we measured the shape bias in two powerful yet poorly understood DNNs - Inception and MNs. Our analysis revealed previously unknown properties of these models. More generally, our work leads the way for future exploration of DNNs using the rich body of techniques developed in cognitive psychology.

Acknowledgements

We would like to thank Linda Smith and Charlotte Wozniak for providing the Cognitive Psychology probe dataset; Charles Blundell for reviewing our paper prior to submission; Oriol Vinyals, Daan Wierstra, Peter Dayan, Daniel Zoran, Ian Osband and Karen Simonyan for helpful discussions; James Besley for legal assistance; and the DeepMind team for support.

References

Bloom, Paul. *How children learn the meanings of words*. MIT press Cambridge, MA, 2000.

Blundell, Charles, Uria, Benigno, Pritzel, Alexander, Li, Yazhe, Ruderman, Avraham, Leibo, Joel Z, Rae, Jack, Wierstra, Daan, and Hassabis, Demis. Model-free episodic control. *arXiv preprint arXiv:1606.04460*, 2016.

Bornstein, Aaron. Is artificial intelligence permanently inscrutable? Despite new biology-like tools, some insist interpretation is impossible. *Nautilus*, 2016.

Caruana, Rich, Lou, Yin, Gehrke, Johannes, Koch, Paul, Sturm, Marc, and Elhadad, Noemie. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1721–1730. ACM, 2015.

Colunga, Eliana and Smith, Linda B. From the lexicon to expectations about kinds: a role for associative learning. *Psychological review*, 112(2):347, 2005.

Goodfellow, Ian, Lee, Honglak, Le, Quoc V, Saxe, Andrew, and Ng, Andrew Y. Measuring invariances in deep networks. In *Advances in neural information processing systems*, pp. 646–654, 2009.

Graves, Alex, Wayne, Greg, Reynolds, Malcolm, Harley, Tim, Danihelka, Ivo, Grabska-Barwińska, Agnieszka, Colmenarejo, Sergio Gómez, Grefenstette, Edward, Ramalho, Tiago, Agapiou, John, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, 2016.

Gregor, Karol, Besse, Frederic, Rezende, Danilo Jimenez, Danihelka, Ivo, and Wierstra, Daan. Towards conceptual compression. In *Advances In Neural Information Processing Systems*, pp. 3549–3557, 2016.

Higgins, Irina, Matthey, Loic, Glorot, Xavier, Pal, Arka, Uria, Benigno, Blundell, Charles, Mohamed, Shakir, and Lerchner, Alexander. Early visual concept learning with unsupervised deep learning. *arXiv preprint arXiv:1606.05579*, 2016.

Hochreiter, Sepp and Schmidhuber, Jürgen. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Hochreiter, Sepp, Younger, A Steven, and Conwell, Peter R. Learning to learn using gradient descent. In *International Conference on Artificial Neural Networks*, pp. 87–94, 2001.

Karpathy, Andrej, Johnson, Justin, and Fei-Fei, Li. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*, 2015.

Kourou, Konstantina, Exarchos, Themis P, Exarchos, Konstantinos P, Karamouzis, Michalis V, and Fotiadis, Dimitrios I. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

Lamberts, Koen and Shanks, David. *Knowledge Concepts and Categories*. Psychology Press, 2013.

Landau, Barbara, Smith, Linda B, and Jones, Susan S. The importance of shape in early lexical learning. *Cognitive development*, 3(3):299–321, 1988.

LeCun, Yann, Bengio, Yoshua, and Hinton, Geoffrey. Deep learning. *Nature*, 521(7553):436–444, 2015.

Li, Jiwei, Chen, Xinlei, Hovy, Eduard, and Jurafsky, Dan. Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066*, 2015.

Lipton, Zachary C. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

Macnamara, John. Cognitive basis of language learning in infants. *Psychological review*, 79(1):1, 1972.

Mareschal, Denis, French, Robert M, and Quinn, Paul C. A connectionist account of asymmetric category learning in early infancy. *Developmental psychology*, 2000.

Markman, Ellen M. Constraints children place on word meanings. *Cognitive Science*, 14(1):57–77, 1990.

Markman, Ellen M and Hutchinson, Jean E. Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive psychology*, 16(1): 1–27, 1984.

Markman, Ellen M and Wachtel, Gwyn F. Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology*, 20(2):121–157, 1988.

Marr, David. Vision: A computational investigation into the human representation and processing of visual information, henry holt and co. *Inc., New York, NY*, 2:4–2, 1982.

Mayor, Julien and Plunkett, Kim. A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychological review*, 117(1):1, 2010.

Plaut, David C, McClelland, James L, Seidenberg, Mark S, and Patterson, Karalyn. Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological review*, 103(1):56, 1996.

Quine, Willard Van Orman. *Word and object*. MIT press, 1960.

Raposo, David, Santoro, Adam, Barrett, David G.T., Pascanu, Razvan, Lillicrap, Timothy, and Battaglia, Peter. Discovering objects and their relations from entangled scene representations. *arXiv preprint arXiv:1702.05068*, 2017.

Regier, Terry. *The human semantic potential: Spatial language and constrained connectionism*. MIT Press, 1996.

Rogers, Timothy T and McClelland, James L. *Semantic cognition: A parallel distributed processing approach*. MIT press, 2004.

Rumelhart, David E, McClelland, James L, Group, PDP Research, et al. *Parallel distributed processing*, volume 1. IEEE, 1988.

Santoro, Adam, Bartunov, Sergey, Botvinick, Matthew, Wierstra, Daan, and Lillicrap, Timothy. Meta-learning with memory-augmented neural networks. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 1842–1850, 2016.

Schilling, Savannah M, Sims, Clare E, and Colunga, Eliana. Taking development seriously: Modeling the interactions in the emergence of different word learning biases. In *CogSci*, 2012.

Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott, Anguelov, Dragomir, Erhan, Dumitru, Vanhoucke, Vincent, and Rabinovich, Andrew. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015a.

Szegedy, Christian, Vanhoucke, Vincent, Ioffe, Sergey, Shlens, Jonathon, and Wojna, Zbigniew. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567*, 2015b.

Tulving, Endel. Elements of episodic memory. 1985.

Vinyals, Oriol, Blundell, Charles, Lillicrap, Timothy, Kavukcuoglu, Koray, and Wierstra, Daan. Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080*, 2016.

Xu, Fei and Tenenbaum, Joshua B. Word learning as bayesian inference. *Psychological review*, 114(2):245, 2007.

Yosinski, Jason, Clune, Jeff, Nguyen, Anh, Fuchs, Thomas, and Lipson, Hod. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

Zeiler, Matthew D and Fergus, Rob. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833, 2014.

Zoran, Daniel, Isola, Phillip, Krishnan, Dilip, and Freeman, William T. Learning ordinal relationships for mid-level vision. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 388–396, 2015.