Deep neural networks have learnt to do an amazing array of tasks - from recognising and reasoning about objects in images to playing Atari and Go at super-human levels. As these tasks and network architectures become more complex, the solutions that neural networks learn become more difficult to understand. This is known as the ‘black-box’ problem, and it is becoming increasingly important as neural networks are used in more and more real world applications.

At DeepMind, we are working to expand the toolkit for understanding and interpreting these systems. In our latest paper, recently accepted at ICML, we proposed a new approach to this problem that employs methods from cognitive psychology to understand deep neural networks. Cognitive psychology measures behaviour to infer mechanisms of cognition, and contains a vast literature detailing such mechanisms, along with experiments for verifying them. As our neural networks approach human level performance on specific tasks, methods from cognitive psychology are becoming increasingly relevant to the black-box problem.

To demonstrate this point, our paper reports a case study where we used an experiment designed to elucidate human cognition to help us...
understand how deep networks solve an image classification task.

Our results showed that behaviours observed by cognitive psychologists in humans are also displayed by these deep networks. Further, the results revealed useful and surprising insights about how the networks solve the classification task. More generally, the success of the case study demonstrated the potential of using cognitive psychology to understand deep learning systems.

**Measuring the Shape Bias in One-shot Word Learning models**

In our case study, we considered how children recognise and label objects - a rich area of study in developmental cognitive psychology. The ability of children to guess the meaning of a word from a single example - so-called ‘one-shot word learning’ - happens with such ease that it is tempting to think it is a simple process. However, a classic thought experiment from the philosopher Willard Van Orman Quine illustrates just how complex this really is:

A field linguist has gone to visit a culture whose language is entirely different from our own. The linguist is trying to learn some words from a helpful native speaker, when a rabbit scurries by. The native speaker declares “gavagai”, and the linguist is left to infer the meaning of this new word. 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”. There is an infinity of possible inferences to be made. How are people able to choose the correct one?

Fifty years later, we are confronted with the same question about deep neural networks that can do one-shot learning. Consider the **Matching Network**, a neural network developed by our colleagues at DeepMind. This model uses recent advances in attention and memory to achieve state-of-the-art performance classifying ImageNet images using only a single example from a class. However, we do not know what assumptions the network is making to classify these images.

To shed light on this, we looked to the work of developmental psychologists (1-4) who found evidence that children find the correct inferences by applying inductive biases to eliminate many of the incorrect inferences. Such biases include:

- **whole object bias**, by which children assume that a word refers to an entire object and not its components (eliminating Quine’s concern about undetached rabbit parts)
- **taxonomic bias**, by which children assume that a word refers to the basic level category an object belongs to (quelling Quine’s fears that all animals might be chosen as the meaning of “rabbit”)
- **shape bias**, by which children assume the meaning of a noun is based on object shape rather than colour or texture (relieving Quine’s anxiety that all white things might be assigned as the meaning of “rabbit”)

We chose to measure the shape bias of our neural networks because there is a particularly large body of work studying this bias in humans.

Examples of the stimuli that we use to measure shape bias in our deep networks. These images were generously supplied by Linda Smith from the Cognitive Development Lab at Indiana University.

The classic shape bias experiment that we adopted proceeds as follows: we present our deep networks with images of three objects: a probe object, a shape-match object (which is similar to the probe in shape but not in colour), and a colour-match object (which is similar to the probe in colour but not in shape). We then measure the shape bias as the proportion of times that the probe image is assigned the same label as the shape-match image instead of the colour-match image.
We used images of objects used in human experiments in the Cognitive Development Lab at Indiana University.

Schematic of our cognitive psychology experiment with the matching network. The matching network matches the probe image (left) to either image ‘A’ (top, middle) or image ‘B’ (top, right). The output (bottom right) depends on the strength of the shape bias in the matching network.

We tried this experiment with our deep networks (Matching Networks and an Inception baseline model) and found that - like humans - our networks have a strong bias towards object shape rather than colour or texture. In other words, they have a ‘shape bias’.

This suggests that Matching Networks and the Inception classifier use an inductive bias for shape to eliminate incorrect hypotheses, giving us a clear insight into how these networks solve the one-shot word learning problem.

The observation of shape bias wasn’t our only interesting finding:

We observed that the shape bias emerges gradually over the course of early training in our networks. This is reminiscent of the emergence of shape bias in humans: young children show smaller shape bias than older children, and adults show the largest bias (2).

We found that there are different levels of bias in our networks depending on the random seed used for initialisation and training. This taught us that we must use a large sample of trained models to draw valid conclusions when experimenting with deep learning systems, just as psychologists have learnt not to make a conclusion based on a single subject.

We found that networks achieve the same one shot learning performance even when the shape bias is very different, demonstrating that different networks can find a variety of equally effective solutions to a complex problem.

The discovery of this previously unrecognised bias in standard neural network architectures illustrates the potential of using artificial cognitive psychology for interpreting neural network solutions. In other domains, insights from the episodic memory literature may be useful for understanding episodic memory architectures, and techniques from the semantic cognition literature may be useful for understanding recent models of concept formation. The psychological literature is rich in these and other areas, giving us powerful new tools to address the ‘black box’ problem and to more deeply understand the behaviour of our neural networks.

This work was done by Sam Ritter*, David G.T. Barrett*, Adam Santoro and Matt M. Botvinick

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

References


