Recurrent networks can recycle neural resources to flexibly trade speed for accuracy in visual recognition

Courtney J Spoerer (courtney.spoerer@mrc-cbu.cam.ac.uk)
MRC Cognition and Brain Sciences Unit, University of Cambridge,
Cambridge, United Kingdom

Tim C Kietzmann (tim.kietzmann@mrc-cbu.cam.ac.uk)
MRC Cognition and Brain Sciences Unit, University of Cambridge,
Cambridge, United Kingdom

Nikolaus Kriegeskorte (nk2765@columbia.edu)
Department of Psychology, Department of Neuroscience, Department of Electrical Engineering,
Zuckerman Mind Brain Behavior Institute, Columbia University,
New York, NY, USA

Abstract

Deep feedforward neural network models of vision dominate in both computational neuroscience and engineering. However, the primate visual system contains abundant recurrent connections. Recurrent signal flow enables recycling of limited computational resources over time, and so might boost the performance of a physically finite brain. In particular, recurrence could improve performance in vision tasks. Here we find that recurrent convolutional networks outperform feedforward convolutional networks matched in their number of parameters in large-scale visual recognition tasks. Moreover, recurrent networks can trade off accuracy for speed, balancing the cost of error against the cost of a delayed response (and the cost of greater energy consumption). We terminate recurrent computation once the output probability distribution has concentrated beyond a predefined entropy threshold. Trained by backpropagation through time, recurrent convolutional networks resemble the primate visual system in terms of their speed-accuracy trade-off behaviour. These results suggest that recurrent models are preferable to feedforward models of vision, both in terms of their performance at vision tasks and their ability to explain biological vision.

Keywords: deep learning; recurrence; visual processing; object recognition; speed-accuracy trade-off

Introduction

Neural networks have a long history as models of biological vision and the recent success of deep neural networks (DNNs) in computer vision has led to a renewed interest in neural network models within neuroscience (Kriegeskorte, 2015; Yamins & DiCarlo, 2016; Kietzmann, McClure, & Kriegeskorte, 2019). Contemporary deep neural networks not only perform better in machine learning challenges but also provide better predictions of neural and behavioural data than previous, shallower models (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Güçlü & van Gerven, 2015).

In terms of computational mechanisms, artificial DNNs diverge largely from their biological counterpart. While some degree of abstraction is necessary when modelling complex systems such as the brain, it is important to understand which features of biology are essential to the computations as reflected in task performance (Kietzmann, McClure, & Kriegeskorte, 2019).

One area that has received particular interest within machine learning and neuroscience has been the role of recurrence. Although core object recognition has typically been viewed as a feedforward process in primates, it is known from neuroanatomy that the visual system is highly recurrent (Felleman & Van Essen, 1991; Sporns & Zwi, 2004). Functional evidence also indicates that recurrent computations are utilised during object recognition (Freiwald & Tsao, 2010; Kietzmann, Spoerer, et al., 2019; Kar, Kubilius, Schmidt, Issa, & DiCarlo, 2019).

Results

We trained a range of deep convolutional neural networks on two large-scale visual object-recognition tasks, ImageNet (Russakovsky et al., 2015) and ecoset (Mehrer, Kietzmann, & Kriegeskorte, 2017). The networks trained included a feedforward network, referred to as B (bottom-up only, layers: 7, feature maps: [96, 128, 192, 256, 512, 1024, 2048]; kernel size: [7, 5, 3, 3, 3, 3, 1]), and a recurrent network, referred to as BL, with bottom-up and lateral recurrent connections (recurrent connections within a layer). We focus our investigation on lateral connections, which constitute a form of recurrence that is ubiquitous in biological visual systems and proved more powerful than top-down recurrent connections on simple tasks in earlier work (Spoerer, McClure, & Kriegeskorte, 2017).

The recurrent networks are implemented by unrolling the computational graph for a finite number of time steps. The model is trained to produce a readout at each time step, which predicts the category of the object present in the image. We defined the prediction of the model as the average of the category readout across all time steps, referred to as cumulative readout hereafter.

As the addition of recurrent connections adds more parameters to the models, we use three larger feedforward architectures as control. The first of these architectures (B-K) uses a larger kernel sizes ([11, 7, 5, 5, 5, 5, 3]). Second, we included control models with a larger number of features in each layer (B-F features: [192, 256, 384, 512, 1024, 2048, 4096]). Finally, we trained a deeper feedforward network (B-D), approximately matching the number of parameters to BL by doubling the number of layers (feature maps: [96, 96, 128, 128, 192, 192, 256, 256, 512, 512, 1024, 1024, 2048, 2048]; kernel sizes: [7, 7, 5, 5, 3, 3, 3, 3, 3, 3, 3, 3, 1, 1]).

Recurrent networks outperform parameter-matched feedforward models

The recurrent models performed best on both large scale object recognition data sets, outperforming both the baseline feedforward model, B, and the parameter-matched controls (Table 1). BL showed a performance benefit of over 1.5 percentage points relative to the best feedforward model, B-D, on both tasks.

Table 1: **Accuracies on held-out data and number of parameters for each model**

models	ImageNet	ecoset	parameters
B (baseline)	58.42%	64.25%	11.0 million
B-K	56.46%	62.81%	39.8 million
B-F	60.34%	66.54%	40.0 million
B-D	62.68%	68.36%	28.9 million
BL (recurrent)	**64.37%**	**69.98%**	28.9 million

The number of parameters are calculated for ImageNet models, ecoset models have slightly fewer parameters due to the fewer categories in the final readout layer.

Both B-D (deeper network) and B-F (more feature maps) outperformed the baseline model, B. B-K had a worse test accuracy than the baseline model, suggesting that the increase in kernel size in our models lead to overfitting. Pairwise McNemar tests (Dietterich, 1998) showed all differences in model performance to be significant ($p \leq 0.05$, Bonferroni corrected).

Single recurrent models span speed-accuracy trade-offs of multiple feedforward models

Next, we compared the computational efficiency of feedforward and recurrent networks by measuring the accuracy as a function of the number of floating-point operations. The number of floating-point operations of a model reflects the energy cost, which might be related to the metabolic cost in a biological system. A feedforward model has a fixed computational cost, whereas a recurrent model can flexibly terminate computations when confidence passes a threshold, trading off accuracy for speed.

For the recurrent models, we used cumulative readouts with entropy thresholding. The network runs until the entropy of its cumulative readout falls below a predefined threshold. The final cumulative readout is then taken as the network's prediction. This effectively uses an internal estimate of the networks' confidence in the decision and terminates once a desired confidence level is reached. Entropy thresholding closely corresponds to theories of biological decision making, where evidence is accumulated until it reaches a bound (Gold & Shadlen, 2007).

When comparing the recurrent models to feedforward models we see a remarkable correspondence between the two classes of architecture (Fig. 1): The accuracy of the recurrent models as a function of the computational cost passes through the points describing the feedforward control models. This means that the different architectures yield similar accuracy for a given computational budget. However, the computational costs and accuracies of the feedforward models are fixed, whereas the recurrent models can be left to compute longer so as to achieve higher accuracies.

These results suggest that recurrent models perform similarly to feedforward models when matching the number of floating-point operations. This is surprising given that recurrent networks operate under the additional constraint of having to use their weights across multiple time steps. The graceful degradation of performance of recurrent models when the computational cost is limited may depend on training with a loss function that rewards rapid convergence to an accurate output.

Overall our results suggest that we can use a single recurrent network to span the space of speed-accuracy trade-offs covered by multiple feedforward networks. Furthermore, using the same network we can achieve a higher performance than all of the parameter-matched feedforward networks by running more recurrent computations.

Network reaction times predict human recognition uncertainty

Recurrent connections endow a model with temporal dynamics. If the recurrent computations in a model match those of the human brain during object recognition, then model behaviour should be predictive of human behaviour. For example, images that require the model to perform more extended recurrent computations for accurate recognition should be more challenging also for humans.

To test this hypothesis we used data from an object categorisation task where humans had to categorise 1,500 greyscale images as either animate or inanimate (Eberhardt, Cader, & Serre, 2016). To quantify the extent to which images that were more consistently recognised by humans were more rapidly recognised by the models, we computed a decision uncertainty index D based on the proportion correct, PC, across humans. D was defined as $0.5 - |0.5 - PC|$. This metric is largest when humans are most inconsistent in their decision making (if $PC = 0.5$ then $D = 0.5$), and it is smallest when all

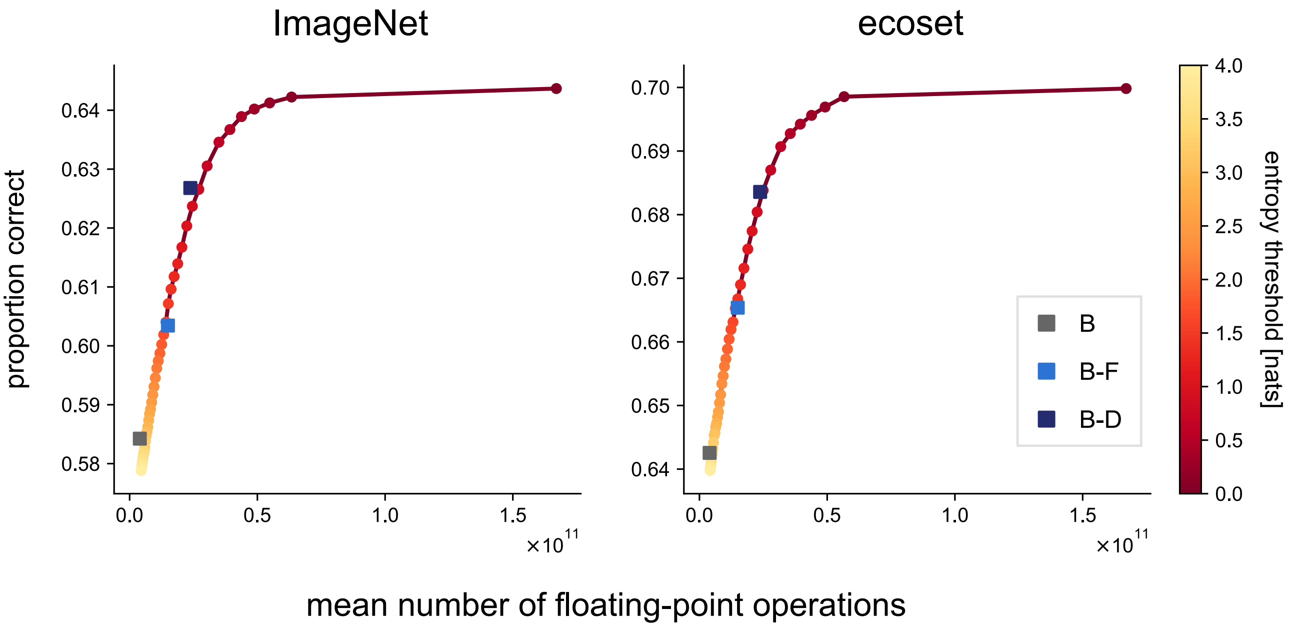

Figure 1: **Relationship between computational cost and performance.** The recurrent models are assessed using a range of entropy thresholds, with the computational cost corresponding to the mean number of floating-point operations used across the test set to reach the given entropy threshold. The computational cost for feedforward models is the number of floating-point operations in a single pass through the model. Performance is assessed based on held-out data.

decisions across trials are the same (if $PC = 1.0$ or $PC = 0.0$ then $D = 0.0$).

We fitted ImageNet and ecoset trained models to these human data and tested them using cross-validation across images. Network reaction times were extracted by training an additional readout for the animacy discrimination task and fitting an entropy threshold to maximise the correlation with human uncertainty. We then tested the fitted models by predicting human uncertainty for different images via crossvalidation (using Spearman correlation to measure prediction accuracy). As a control, we ran the same fitting procedure using a network with randomly initialised weights. As an additional control, we shuffled the images within each category before fitting the entropy thresholds and recomputing the network reaction times.

Our results show that reaction times obtained from recurrent networks significantly predicted human decision uncertainty (Fig. 2). Furthermore, both networks outperformed a randomly initialised network that was fitted using the same procedure (two-tailed paired permutation test, $p < 0.01$). Overall, images for which our recurrent networks took longer to converge were less consistently recognised by humans.

Conclusions

The results described here show that recurrent architectures can outperform parameter-matched feedforward control models on naturalistic vision tasks. We also demonstrated that a single recurrent network can span the space of speed-accuracy tradeoffs covered by multiple feedforward models. Not only did the recurrent architectures have a practical benefit, but they were also able to predict human object recognition behaviour.

The work described here adds to a growing body of research into RCNNs as models of object recognition (Liang & Hu, 2015; Liao & Poggio, 2016; Spoerer et al., 2017; Nayebi et al., 2018; Kubilius et al., 2018; Kietzmann, Spoerer, et al., 2019). These models provide us with a white box, a vision system that can be observed from input to behavioural response. By understanding how these models perform object recognition we might shed some light on the elusive role of recurrent processing within biological vision.

Acknowledgments

We thank Sven Eberhardt, Jonah Cader and Thomas Serre for sharing their behavioural data. This project has received funding from the European Unions Horizon 2020 Programme for Research and Innovation under the Specific Grant Agreement No. 720270 and 785907 (Human Brain Project SGA1 and SGA2), the NVIDIA GPU Grant Program, and the German Science Foundation (DFG).

References

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895–1923.

Eberhardt, S., Cader, J. G., & Serre, T. (2016). How deep is the feature analysis underlying rapid visual categorization? In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, &

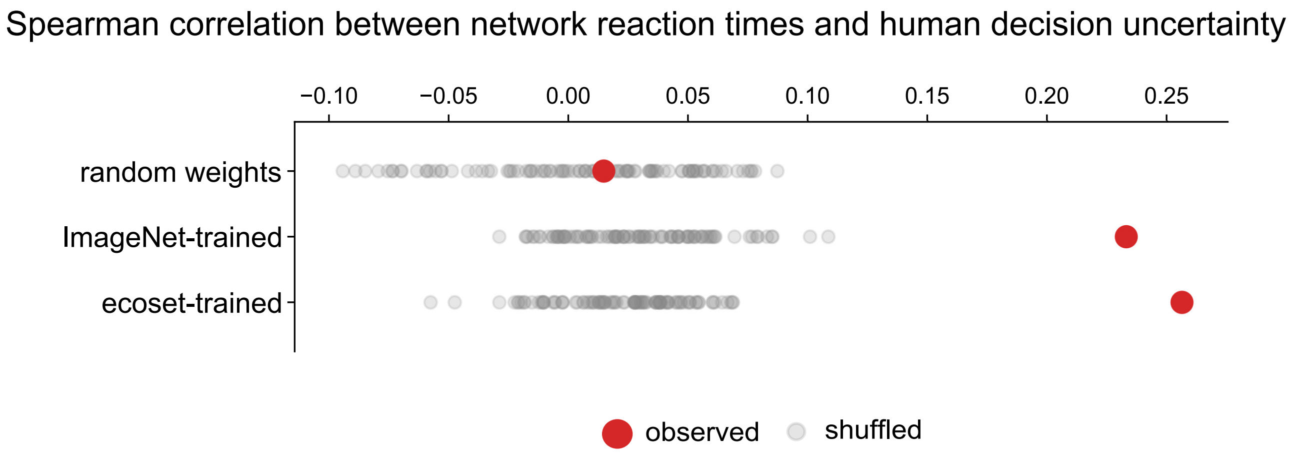

Figure 2: **Model reaction times are shorter for images that humans are uncertain about.** Spearman correlations between network reaction times and human decision uncertainty (red) alongside correlations obtained when timecourses from trained readouts were randomly shuffled within categories (grey).

R. Garnett (Eds.), *Advances in neural information processing systems 29* (pp. 1100–1108). Curran Associates, Inc.

Felleman, D. J., & Van Essen, D. C. (1991, 01). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*(1), 1-47.

Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, *330*(6005), 845–851.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*(1), 535–574. (PMID: 17600525)

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Kar, K., Kubilius, J., Schmidt, K. M., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral streams execution of core object recognition behavior. *Nature Neuroscience*.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLOS Computational Biology*, *10*(11), 1–29.

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2019). Deep neural networks in computational neuroscience. *Oxford Research Encyclopedia of Neuroscience*.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence required to capture the dynamic computations of the human ventral visual stream. *arXiv preprint arXiv:1903.05946*.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*(1), 417–446.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L., & DiCarlo, J. J. (2018). Cornet: Modeling the neural mechanisms of core object recognition. *bioRxiv*.

Liang, M., & Hu, X. (2015). Recurrent convolutional neural network for object recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3367–3375). Boston, MA, USA.

Liao, Q., & Poggio, T. (2016). Bridging the gaps between residual learning, recurrent neural networks and visual cortex. *arXiv preprint arXiv:1604.03640*.

Mehrer, J., Kietzmann, T. C., & Kriegeskorte, N. (2017). Deep neural networks trained on ecologically relevant categories better explain human it. In *Conference on cognitive computational neuroscience.* New York, NY, USA.

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., ... Yamins, D. L. (2018). Task-driven convolutional recurrent models of the visual system. In *Advances in neural information processing systems* (pp. 5290–5301).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... others (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology*, *8*, 1551.

Sporns, O., & Zwi, J. D. (2004). The small world of the cerebral cortex. *Neuroinformatics*, *2*, 145–162.

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*, 356–365.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & 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.