Recurrent Attentional Selection Can Explain Flexible Trading of Accuracy and Energy in Biological Vision

Eivinas Butkus (eb3407@columbia.edu)
Department of Psychology, Columbia University
New York, NY 10027, USA

Zhuofan Ying (zy2559@columbia.edu)
Department of Psychology, Columbia University
New York, NY 10027, USA

Peiyu Chen (pc3092@columbia.edu)
Department of Psychology, Columbia University
New York, NY 10027, USA

Nikolaus Kriegeskorte (nk2765@columbia.edu)
Departments of Psychology, Neuroscience and Electrical Engineering,
Zuckerman Mind Brain Behavior Institute, Columbia University
New York, NY 10027, USA

Abstract

Biological vision is energetically costly. Visual attention may save energy by selecting, on the basis of a cursory initial analysis, the features and locations that deserve scrutiny. Here we investigate this idea using recurrent convolutional neural network models with graded attentional selection, implemented as multiplicative gain on features (*what gain*) and locations (*where gain*). The task for both humans and models was to determine the class (*what*) and location (*where*) of a handwritten digit among letters. Humans viewed brief presentations of such cluttered images and also rated the *difficulty* of each search. Models were trained with a loss encouraging high accuracy in both the what and the where task and low energy use. We found that models with attention achieved the best (Pareto-optimal) combinations of energy and accuracy. In contrast to models with no penalty on energy use, models optimized with an intermediate energy cost term consistently had a higher correlation across images between model energy use and human difficulty judgments. Finally, models that included feature-based attention (*what gain*) better explained human difficulty judgments. Our work demonstrates the importance of resource costs for understanding the computational mechanisms of biological vision.

Keywords: attention; vision; energy; recurrent neural networks

Introduction

Biological visual systems are energetically costly (Wong-Riley, 2010). Visual attention may save energy by steering scrutiny to aspects of the image that matter given current task demands and computational state. Here we evaluated this idea using recurrent convolutional neural networks with adaptive multiplicative gain on feature maps ("what gain"), locations ("where gain"), both ("what & where gain"), or neither ("no gain"). We trained these networks to maximize accuracy on a digit detection task while minimizing energy use (defined as the sum of activations in the convolutional layers). The networks had noise added to the activations of the convolutional layers, creating a trade-off, where high signal relative to the noise incurred high energy costs.

Related work. Spoerer et al. (2020) explain time and accuracy trade-offs in vision using recurrent convolutional neural networks without an explicit attentional mechanism (top-down gain modulation) as used here. They define energy cost as the number of floating-point operations and do not optimize networks for low energy. We use the sum of neural firing rates across space and time (a biologically plausible notion of energy usage) as a term in the cost function. Młynarski and Tkačik (2022) show how feedback signals may adapt the sensory code to use fewer spikes. However, their model is limited to simple visual tasks where computing the full posterior distribution is tractable. Konkle and Alvarez (2023) implement a feature-based gain mechanism that captures a few classic top-down neural signatures of category-based attention and increases AlexNet's alignment with brain representations. Our work is complementary, focusing more on how attention mechanisms may improve energy efficiency of biological vision.

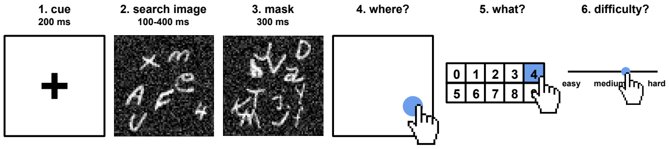

Figure 1: *Human behavioral task.*

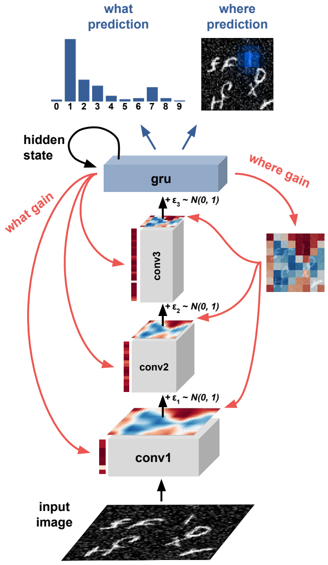

Figure 2: *Model architecture.*

Task

The task for humans and models was to detect an MNIST (LeCun, 1998) handwritten digit among handwritten letters sampled from Extended MNIST (Cohen et al., 2017) (Fig. 1). For each of the 400 64x64 images, we uniformly sampled the level of pixel-wise Gaussian noise (SD: 0-0.2) and the number of distractors (1-8). Presentation times were sampled from {100, 200, 300, 400}ms and each search image was accompanied by a backward masking image with similar statistics. We asked subjects to report by mouse click (1) *where* the digit was on a blank square, (2) *what* class it belonged to, and (3) the *difficulty* level of that particular search on a continuous scale "easy-medium-hard". Each subject received 10 training trials with feedback for *what* and *where* guesses, but there was no feedback in the experiment. The task was performed by 20 subjects on Prolific.

Model

All models are recurrent convolutional neural networks making 4 passes on the static search image (Fig. 2). The base *no gain* model consists of 3 convolutional layers followed by a gated recurrent unit (GRU) layer that keeps a hidden state across passes. On each pass, the digit class ("what prediction") and location ("where prediction") is linearly read out from the GRU output.

Multiplicative gain. Models compute gain maps linearly from the hidden state of the GRU. The *what gain* model has gain on features (channels), *where gain* model has gain on spatial locations, and *what & where gain* model has both. Gain maps are broadcast and multiplied element-wise with pre-activations in convolutional layers. A feature is attended, thus, to the extent that it has significant what and where gain. Suppression of half the feature maps and half the locations would reduce the attended features to one fourth.

Energy use. Motivated by the idea that a neural spike requires a quantum of energy, we interpreted activations of our rate-coded models as neural firing rates and defined the energy loss L_{energy} on a perceptual trial as the sum of all convolutional activations across the 4 passes. We prevented the models from scaling activations down arbitrarily to save energy by adding standard normal noise (after applying gain). The models are thus required to use substantial activations (signal relative to noise) to transmit significant information, while being optimized to use energy efficiently. The final loss encourages high accuracy in both the what and the where task and low energy use: $L = L_{\text{what}} + L_{\text{where}} + \lambda_{\text{energy}} \times L_{\text{energy}}$. We used 12 different weights for the energy term $\lambda_{\text{energy}} \in [0.0, 0.3]$ and trained 10 instances for each model-and-energy-weight combination.

Results

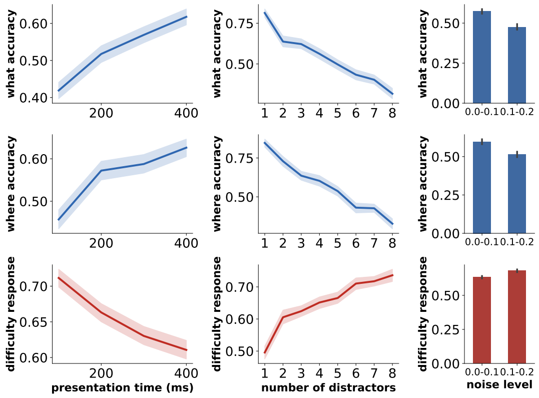

Figure 3: *Human behavioral results* (N=20). Shaded areas and error bars represent standard error of the mean.

Human behavioral results (Fig. 3). Accuracy in both what and where tasks increased with presentation time, and decreased with the number of distractors and noise level. Difficulty judgments showed the opposite pattern.

Model accuracy (Fig. 4A). When $\lambda_{\text{energy}} = 0$, models with gain have a higher accuracy in both what and where tasks. So attention is beneficial even without a constraint on energy.

Model accuracy-energy trade-off (Fig. 4B). Models with attention achieve Pareto-optimal combinations of what accuracy, where accuracy and energy use.

Model-human correlations (Fig. 5). We computed image-level Spearman rank correlations between model and human errors on what and where tasks, as well as model energy use and human difficulty judgements as a function of energy weight λ_{energy} (Fig. 5A). We then selected λ_{energy} values that resulted in most human-aligned models (indicated by a star) and ran paired bootstrap difference tests with Bonferroni correction for multiple comparisons (Fig. 5B). We found no statistically significant differences in human what or where error prediction between the attentional and non-attentional models trained with maximally human-aligned λ_{energy}. However, we found that non-zero weight on the energy cost $\lambda_{\text{energy}} > 0.0$ helps *all models* better explain human difficulty judgments (Fig. 5A, bottom). Moreover, models that included feature-based attention (*what gain* and *what & where gain*) had a significantly higher correlation between energy use and human difficulty judgements than *no gain* model (Fig. 5B, bottom).

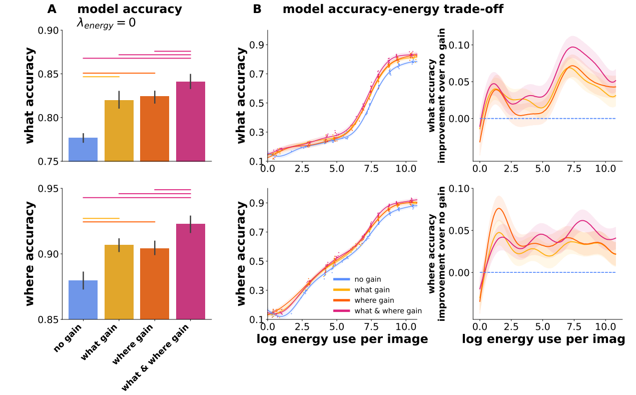

Figure 4: *Model results.* **A**. Error bars represent standard error and colored horizontal lines indicate significant differences between the models. **B**. Lines and shaded areas represent predictions (mean and SD) from a fitted Gaussian process.

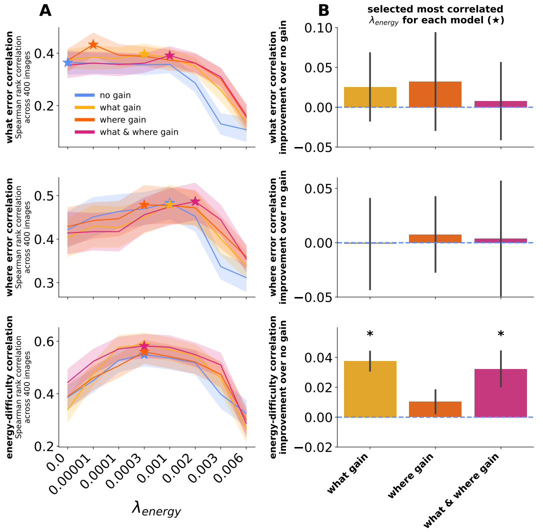

Figure 5: *Model-human correlations.* We computed Spearman rank correlations across 400 images between model and human what and where errors, as well as between model energy use and human difficulty ratings. Shaded area and error bars represent 95% confidence intervals bootstrapped across subjects and model instances.

Conclusion

We show how energy costs can be optimized along with task performance in neural network models of biological vision and demonstrate that attentional mechanisms can boost performance, save energy, and help explain human behavior and task difficulty judgments.

Acknowledgments

This work was supported in part by the National Science Foundation under Grant No. 1948004 to NK.

References

Cohen, G., Afshar, S., Tapson, J., & van Schaik, A. (2017, March). *EMNIST: an extension of MNIST to handwritten letters.* arXiv. Retrieved 2024-04-16, from (arXiv:1702.05373 [cs])

Konkle, T., & Alvarez, G. (2023). Cognitive Steering in Deep Neural Networks via Long-Range Modulatory Feedback Connections.

LeCun, Y. (1998). The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

Młynarski, W., & Tkačik, G. (2022, December). Efficient coding theory of dynamic attentional modulation. *PLOS Biology*, *20*(12), e3001889. Retrieved 2023-05-19, from doi: 10.1371/journal.pbio.3001889

Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I., & Kriegeskorte, N. (2020, October). Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLOS Computational Biology*, *16*(10), e1008215. Retrieved 2023-07-10, from doi: 10.1371/journal.pcbi.1008215

Wong-Riley, M. (2010, July). Energy metabolism of the visual system. *Eye and Brain*, 99. Retrieved from doi: 10.2147/EB.S9078