Predictive remapping and allocentric coding as consequences of energy efficiency in recurrent neural network models of active vision

Thomas Nortmann (tnortmann@uni-osnabrueck.de)
Institute of Cognitive Science, University of Osnabrück, 49090 Osnabrück, Germany

Philip Sulewski (psulewski@uni-osnabrueck.de)
Institute of Cognitive Science, University of Osnabrück, 49090 Osnabrück, Germany

Tim C. Kietzmann (tim.kietzmann@uni-osnabrueck.de)
Institute of Cognitive Science, University of Osnabrück, 49090 Osnabrück, Germany

Abstract

Despite moving our eyes from one location to another, our perception of the world is stable - an aspect thought to rely on predictive computations that use efference copies to predict the upcoming foveal input. Are these complex computations genetically hard-coded, or can they emerge from simpler principles? Here we consider the organism's limited energy budget as a potential origin. We expose a recurrent neural network to sequences of fixation patches and saccadic efference copies, training the model to minimise energy consumption (preactivation). We show that targeted inhibitory predictive remapping emerges from this energy efficiency optimization alone. As furthermore demonstrated, this computation relies on the model's learned ability to re-code egocentric eye-coordinates into an allocentric (image-centric) reference frame. Together, our findings suggest that both allocentric coding and predictive remapping can emerge from energy efficiency constraints during active vision, demonstrating how complex neural computations can arise from simple physical principles.

Keywords: predictive remapping; energy efficiency; recurrence; visual stability; natural scenes; eye-movements; neuroconnectionism

Method

Dataset

Natural scenes were sourced from the MS-COCO dataset (Lin et al., 2015) with human-like fixation sequences generated via the DeepGaze III model (Kümmerer et al., 2022). Each input sequence included seven fixations. A training set of 48,236 images was selected with an additional test set of 2,051 images. For each image, 10 different fixation sequences were generated. The original greyscaled scenes had a size of 256 × 256 pixels; the fixation crops were selected to be 128 × 128 pixels.

Model and Training

The model architecture consisted of a fully connected RNN with two hidden layers (2048 units each) and lateral connections. Input consisted of 128×128 pixel fixation crops with corresponding efference copies (Δ x, Δ y coordinates). Following Ali et al. (2022), input drive was fixed (non-learnable) to prevent the network from ignoring input to save energy.

The RNN was trained to minimise metabolic energy consumption using mean absolute preactivation as the loss function (see Fig 1B) where $\mathcal{L}$ represents the energy efficiency loss, N is the number of units, and $|\text{preactivation}_{i,t}|$ denotes the absolute preactivation value of unit i at timestep t.

Results

Energy efficiency drives predictive remapping

Energy-optimized RNNs developed targeted inhibitory predictive remapping, significantly outperforming control conditions including average crop luminance, location-specific average crop, previous fixation control, shuffled fixation sequences, and models without efference copies (all $p < .001$, see Fig 2A,B). Even with 56% smaller fixation crops reducing adjacent overlap, networks maintained predictive capabilities above all controls.

Top-down feedback was learned to be inhibitory (μ = -0.39, 99% CI = [-0.78, -0.09], n = 86142) and spatially specific to saccadic target locations rather than global inhibition. The model's internal drive aligned with ideal inhibition patterns, showing smooth yet targeted predictions matching expected visual input at upcoming fixation locations. Performance improved when recent fixations were spatially proximate, indicating spatial memory formation across multiple saccades.

Allocentric coding enables predictive computations

Networks spontaneously learned allocentric coding, achieving high accuracy in decoding absolute fixation coordinates from relative efference copy sequences (x-coordinate: R^2 = 0.91; y-coordinate: R^2 = 0.93). This transformation from egocentric to allocentric reference frames occurred without explicit supervision, relying on sparse units (approx. 0.5% of units).

Targeted lesioning of allocentric units (n = 20, top 0.5%) eliminated predictive capabilities. Lesioned networks reverted to using current rather than predicted visual input for inhibition. The correlation with ideal inhibition dropped dramatically (intact: r = .46, $p < .001$; lesioned: r = .05, $p < .001$, see Fig 2E). Instead, lesioned models aligned with current crop inhibition patterns (r = .58, $p < .001$, see Fig 2E). Critically, lesioned networks could no longer outperform the no-efference-copy control ($p > 0.999$, see Fig 2D), demonstrating functional necessity of allocentric coding for predictive remapping. Random lesioning of equivalent units produced no performance degradation.

Discussion

We demonstrated that sophisticated visual stability mechanisms could emerge from basic energy constraints without specialized genetic programming. The spontaneous development of allocentric coding mirrors spatial navigation research (Banino et al., 2018), suggesting shared computational principles between visual exploration and navigation systems. Our findings extend predictive coding frameworks by showing how complex spatial transformations arise from simple physical constraints. This work provides testable neuroscientific predictions: allocentric coding neurons should disproportionately impact visual stability, and neural navigation/visual stability mechanisms should overlap significantly. These results contribute to understanding how fundamental physical constraints drive sophisticated neural computations, offering a potential solution to the hard binding problem (Cavanagh et al., 2010) through energy efficiency rather than complex genetic architectures.

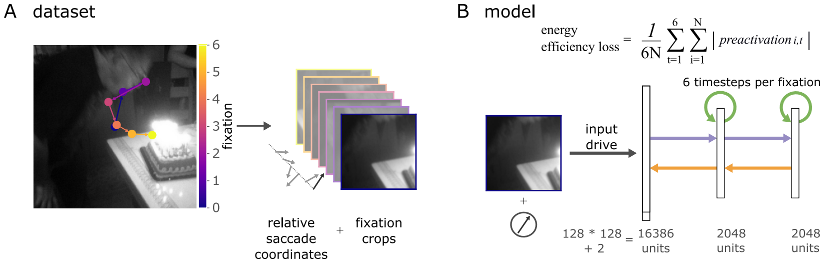
A dataset
6
5
4
3
2
1
0
fixation
relative saccade coordinates
+
fixation crops
B model
energy efficiency loss = $\frac{1}{6N}\sum_{t=1}^{6}\sum_{i=1}^{N}|preactivation_{i,t}|$
6 timesteps per fixation
input drive
+
128 * 128 + 2 = 16386 units
2048 units
2048 units

Figure 1: **Minimizing preactivation in response to human-like saccade sequences. (A)** Fixation crop sequence generation using DeepGaze III fixations on MS-COCO images. **(B)** Model architecture and loss function.

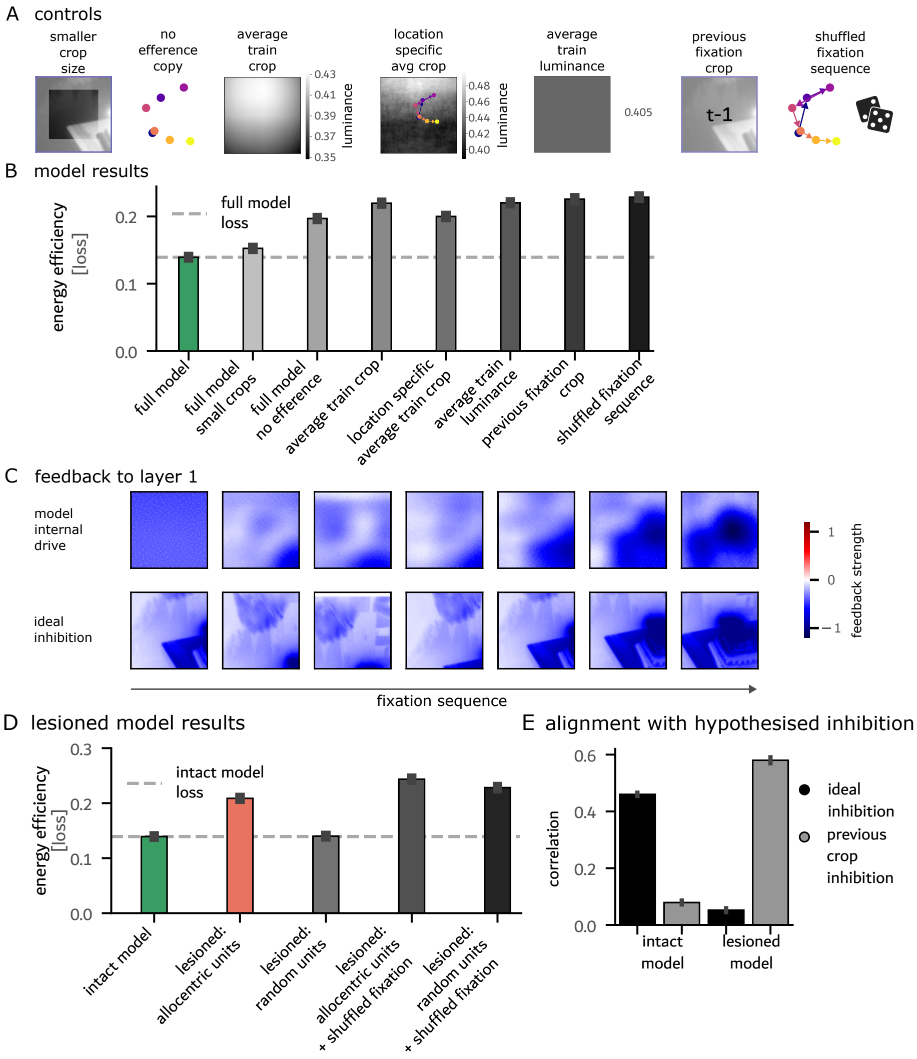

Figure 2: **Inhibitory predictive remapping with allocentric reference frame as a consequence of energy efficiency. (A)** Illustration of the control conditions: training with smaller crops (56 % smaller), training without efference copies, location specific average crop, average crop, average luminance value, previous fixation crop and testing with shuffled fixation sequences. **(B)** Energy efficiency loss in the RNN and control conditions. **(C)** Example of the RNN's internal feedback to layer 1 (upper row), together with the ideal inhibition (bottom row). While smooth, the inhibitory patterns align with the ideal inhibition. **(D)** Lesioning allocentric units led to a significant increase in energy consumption. **(E)** Correlation of matrices for observed and hypothesised prediction patterns. The lesioned model aligns with the current crop; the intact model aligns with the ideal prediction of the future input.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 456666331. P.S. and T.C.K. were supported by the ERC Starting Grant TIME (101039524). P.S. was supported by the Federal Ministry of Education and Research (BMBF; Studienstiftung des deutschen Volkes) and the Max Planck Society (MPG; Max-Planck-School of Cognition), and Deutsche Forschungsgemeinschaft (DFG; German Research Foundation) - GRK 2340. The authors thank Dr. Varun Kapoor for assisting with the dataset creation.

References

Ali, A., Ahmad, N., De Groot, E., Johannes Van Gerven, M. A., & Kietzmann, T. C. (2022). Predictive coding is a consequence of energy efficiency in recurrent neural networks. *Patterns*, *3*(12), 100639. doi: 10.1016/j.patter.2022.100639

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., . . . Kumaran, D. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*, *557*(7705), 429–433. doi: 10.1038/s41586-018-0102-6

Cavanagh, P., Hunt, A. R., Afraz, A., & Rolfs, M. (2010). Visual stability based on remapping of attention pointers. *Trends in Cognitive Sciences*, *14*(4), 147–153. doi: 10.1016/j.tics.2010.01.007

Kümmerer, M., Bethge, M., & Wallis, T. S. A. (2022). DeepGaze III: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*, *22*(5), 7. doi: 10.1167/jov.22.5.7

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., . . . Dollár, P. (2015). *Microsoft COCO: Common Objects in Context.* doi: 10.48550/arXiv.1405.0312