Sparks of cognitive flexibility: self-guided context inference for flexible stimulus-response mapping by attentional routing

Rowan P. Sommers[†], Sushrut Thorat[†], Daniel Anthes, Tim C. Kietzmann
{rsommers, sthorat, danthes, tkietzma}@uos.de
Institute of Cognitive Science, University of Osnabrück, Germany; [†]Equal contribution.

Abstract

Flexible cognition entails a rapid adaptation of stimulus-response mappings. Standard neural networks struggle in tasks requiring rapid remapping. Here, we propose the Wisconsin Neural Network (WiNN), which generalizes fast-and-slow learning to real-world tasks demanding flexible behavior, using adjustable context states that guide attention in a pretrained convolutional neural network. We evaluate WiNN on a variant of the Wisconsin Card Sorting Task, revealing several markers of cognitive flexibility: (i) WiNN autonomously infers underlying rules, (ii) requires fewer examples than control models reliant on large-scale parameter updates, and (iii) can perform rule inference solely via context-state adjustments. This approach offers a path toward context-sensitive models that retain knowledge while rapidly adapting to complex, rule-based tasks.

Keywords: cognitive flexibility; attention; routing; context inference; continual learning; compositionality; neuroconnectionism

Model and learning dynamics

WiNN consists of four elements: 1) A four-layer convolutional network, pretrained on MiniEcoset (Thorat et al., 2023) and frozen thereafter, a stand-in for an adult ventral stream (Yamins & DiCarlo, 2016); 2) A linear readout converts the backbone's global-pooled activity into a binary judgment: does the image satisfy the active rule or not? 3) A context state which is internally updated to learn the current rule. 4) A multiplicative attention matrix broadcasts the current context state to every neuron in every convolutional layer, attending features relevant to the current rule context (Lindsay & Miller, 2018; Singer et al., 2024).

Whenever WiNN errs, learning proceeds in two stages, inspired by Hummos (2023). An inner loop performs up to 100 gradient steps on only the context vector (learning rate = 10^{-2}), shifting attention just far enough to correct the decision. A single outer step then updates the attention and readout (add-on) weights at a much smaller rate (10^{-4}), leaving the backbone unchanged. Separate optimizers maintain the strict fast–slow separation, so perceptual features remain reusable while context and the add-on weights absorb task drift.

Experimental protocol

WiNN tackles an image-based analogue of the Wisconsin Card Sorting Test. In each block, of 800 unique stimuli drawn from the 3D-Shapes dataset, a hidden single-factor rule—floor colour, wall colour, object colour or object shape—governs the "yes" class. To construct the various blocks, three values per factor are sampled, yielding 12 blocks in total. This sampling is repeated 10 times to construct as many experiments. A held-out validation set gauges generalization, while an extended context-inference set tests whether WiNN can adapt by moving its context state alone. That set includes (i) seen single-factor rules, (ii) unseen rules which contain unseen values of seen factors, and (iii) novel compositional rules formed by conjoining two seen single-factor rules.

Five control models were used as comparisons: 3 pretrained CNNs that get similar/different data diets and learning rates as WiNN: Either the CNN gets no repetitions of the same image (CNN No Repeat), or it gets repetitions with either 10^{-2} (CNN 100x Repeat) or 10^{-4} (CNN 100x Repeat Slow) learning rates - analogous to the inner-loop context state update. In addition, we used two versions of WiNN, one with a random backbone and one with frozen attention and readout weights.

Results

Results are shown in Figure. 1.

Efficient rule discovery After convergence, WiNN reaches the 90 % validation-accuracy threshold in far fewer images than any control model, demonstrating quicker extraction of the latent rule while keeping generalization high. Control models that update thousands of backbone weights either lag substantially or, when driven by large learning rates, catastrophically over-fit to the block in progress.

better generalization Although some control models eventually top 90 % on the training stream, their validation accuracy remains several percentage points lower, signaling residual over-fitting that WiNN largely avoids.

Context-only remapping With all synaptic weights frozen, WiNN's inner loop still brings performance on previously seen rules close to ceiling, and lifts accuracy on unseen simple and compositional rules well above chance. Over successive sequences, context-only accuracy on novel compositional rules even climbs steadily, showing that attention and readout layers have learned abstractions that can be recombined on the fly. Ablations confirm that freezing attention and readout parameters cripples this ability. This underscores the synergy between stable features, learned gain control and a malleable context state.

Conclusion

By combining rapid context inference with slow, broadly-useful weight updates, WiNN reconciles two goals often viewed as incompatible in deep learning: instant behavioural remapping and long-term knowledge retention. The model therefore supplies tangible "sparks of cognitive flexibility"—swift adaptation to new or hidden rules, sample-efficient learning, resistance to catastrophic forgetting and compositional generalization. Beyond its engineering value, WiNN represents a successful computational model of how thalamocortical and fronto-parietal circuits might gate context in the brain. Extending the same principles to richer memory stores, language-conditioned context vectors or hierarchical/serial attention could push artificial agents closer to the breadth of human adaptive behavior while furnishing new, mechanistic questions for systems neuroscience.

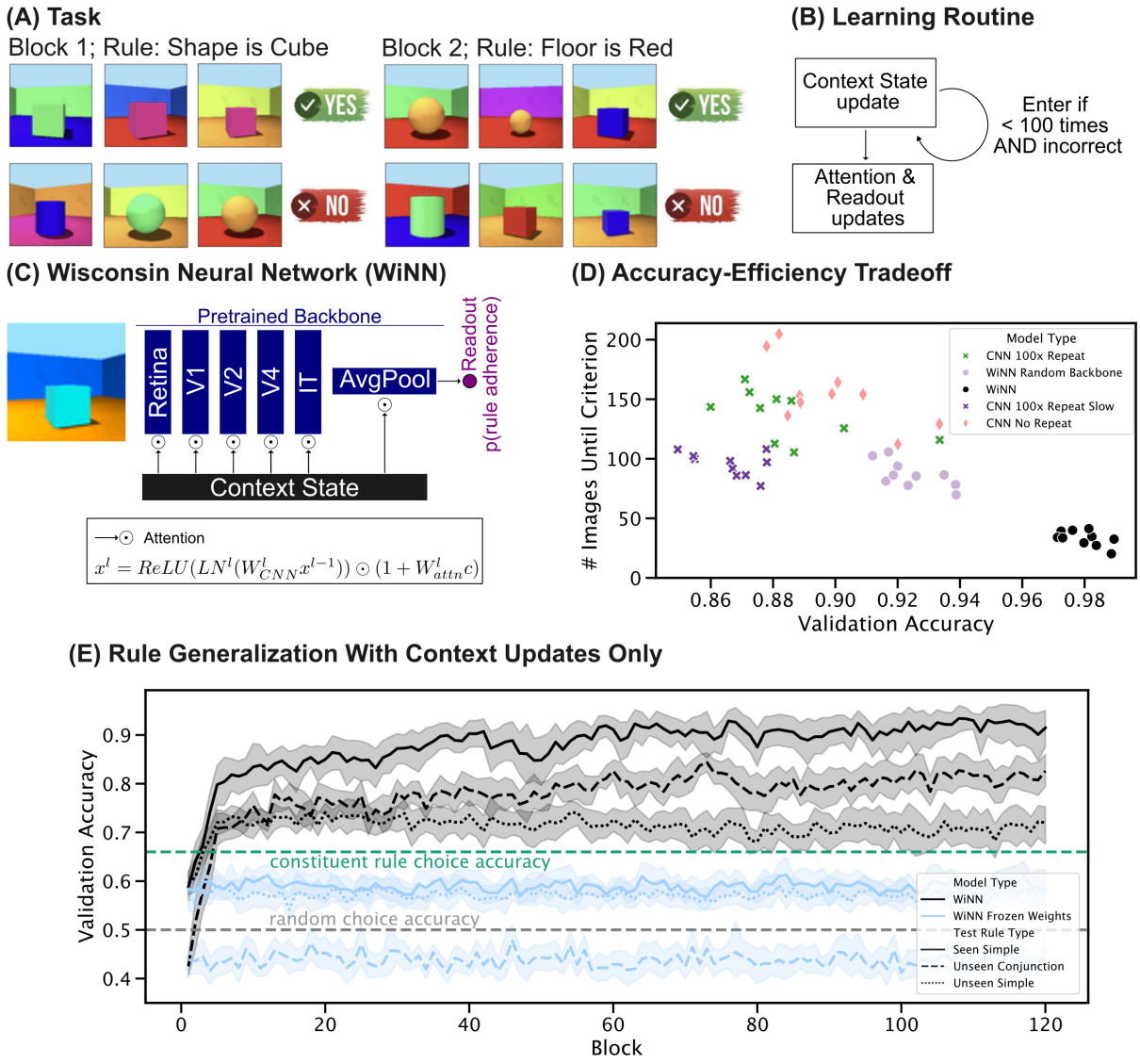

Figure 1: **(A)** During an experiment, images are presented one at a time, and in blocks. Each block has a hidden rule. For each image, the task is to decide whether it adheres to the hidden rule. After 800 images, a block ends and the hidden rule switches. Importantly, there is no indication of a rule switch other than altered feedback. **(B)** Whenever WiNN produces an error, its context state is updated iteratively to remap the network's response. After 100 updates or if the response is correct, a single update is applied to the attention and readout parameters. The backbone is kept frozen. **(C)** The Wisconsin Neural Network (WiNN) is built for flexible rule inference over complex image streams. A pretrained convolutional neural network (CNN) maps the image to a response that is modulated by the inferred context. The bottom panel specifies how the attention weights map the context state c modulation onto the l^{th} CNN layer. **(D)** WiNN can infer rules better than the control models: it can better generalize the seen stimulus-response mapping to unseen stimuli. WiNN can infer rules faster than the control models: it needs to see fewer stimuli to be accurate on subsequent stimuli. In contrast, control CNNs suffer from an accuracy-efficiency tradeoff. In addition, pretraining the backbone is important for WiNN's efficient rule inference capability. **(E)** WiNN can infer previously seen simple rules solely with context state updates. WiNN can also infer unseen simple rules and compositions of seen rules, although not perfectly. This suggests that the learned context state/attention/readout mappings to and from the backbone are general enough to extend to unseen rules, and especially to compositions of seen rules. Moreover, the inference of rules becomes better through the experiment, suggesting meta-learning. WiNN with frozen attention/readout weights perform worse, indicating that good generalization cannot be achieved by updating the context state alone and that appropriate attention/readout weights are necessary for rule generalization.

Acknowledgments

The project was financed by the funds of the research training group “Computational Cognition” (GRK2340) provided by the Deutsche Forschungsgemeinschaft (DFG), Germany and the European Union (ERC, TIME, Project 101039524).

References

Hummos, A. (2023). Thalamus: a brain-inspired algorithm for biologically-plausible continual learning and disentangled representations. In *The eleventh international conference on learning representations.*

Lindsay, G. W., & Miller, K. D. (2018). How biological attention mechanisms improve task performance in a large-scale visual system model. *ELife*, *7*, e38105.

Singer, J. J., Cichy, R. M., Kietzmann, T. C., & Thorat, S. (2024). Contrasting computational models of task-dependent readout from the ventral visual stream. In *Conference on cognitive computational neuroscience.*

Thorat, S., Doerig, A., & Kietzmann, T. C. (2023). Characterising representation dynamics in recurrent neural networks for object recognition. In *Conference on cognitive computational neuroscience.*

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356–365.