Epiphenomenal representations of abstract rules in a connectionist model of the Delayed Match to Sample task

Badr AlKhamissi Sony AI
Muhammad ElNokrashy Microsoft
Akshay Srinivasan Sony AI
Zeb Kurth-Nelson DeepMind
Sam Ritter DeepMind

Summary

In the human brain, some individual neurons respond selectively to abstract variables, invariant to sensory grounding [1]. Similar units also appear in artificial networks trained on cognitive tasks [2]. It is often implicitly assumed that the emergence of such interpretable neurons plays a key role in the behavior of the trained network. Here we show that this is not necessarily the case. We train a biologically inspired artificial agent comprising two key components—a recurrent network and an associative memory—on a canonical rule-based neuroscience task [3], and observe the emergence of brain-like rule representations in the recurrent network. Crucially, however, we find that these representations are not used to guide behavior at test time—ablating these units has minimal impact on performance. However, we find that ablating other units in the recurrent network can severely degrade performance. These results call into question the assumption that observing representations in a brain region along with performance degradation when that region is lesioned is sufficient to infer that those representations cause the animal's behavior in the task. These results point the way toward further modeling and animal experiments that may improve our understanding of epiphenomenality in the brain.

Methods

The *Delayed Matching to Sample* (**DMS**) task was originally designed to study the neural basis of the PFC in representing abstract rules by recording single-neuron activities in macaque monkeys [1]. It involves the agent comparing two images presented in succession, separated by a delay period. The agent must then indicate that they are the same or different according to the rule in effect. The **match** abstract rule asks the agent to answer `Yes` if both stimuli are *identical* and `No` otherwise, while the **non-match** abstract rule requires the opposite—`Yes` if both stimuli are *different*. The rule to be applied is signified by a cue presented alongside the sample at the beginning of each trial. Figure 1 shows the structure of the experiment with the expected outcomes. To reach the correct decision at response time, the agent needs to maintain a representation of both the abstract rule and the sample image across the delay period. It then applies the rule on both the remembered image and the test image when that is presented. Further, to demonstrate an understanding of the two abstract rules, the agent needs to be able to apply them on novel stimuli that it has not experienced during training. The images presented to the agent are sampled from the **Stimuli** dataset [4]. It contains $2,400$ unique objects, 400 of which are held-out for testing. Similar to [3], two distinct cues are used for each abstract rule. In the following section, we replicate the behavioral and neural results of [3] by observing the activations of single-units in an **LSTM** (at the $\boldsymbol{h}_t$ boundary), instead of recording the firing-rate of single-neurons in the PFC. Our architecture is similar to the **A2C LSTM** proposed by [5], with the exception that we augment it with an episodic memory in a similar manner to [6].

Results

Behavioral The results in Table 1 are the aggregate of 30 different seeds. To test the generalization of the trained models, we evaluate them on 400 randomly sampled images from different datasets as well as randomly generated pixels. Note that those models were trained only on the 2k images from the *Stimuli* dataset. Here, each image was tested against the four possible cues, each of which had the identical and a different image sampled for the testing phase, making for a total of $400 \times 4 \times 2 = 3200$ trials.

Dataset	Mean ± SEM	Max	Min
Stimuli	98.15% ± 0.2	99.53%	96.34%
MNIST	95.62% ± 0.5	98.81%	85.72%
CIFAR10	97.49% ± 0.2	99.16%	93.72%
Noise	69.23% ± 1.5	91.66%	51.56%

Table 1: Test time performance. SEM is standard error of the mean.

Neural In the original experiment, [3] identified single-neurons that are selective to one of the two abstract rules by showing that they exhibit greater activity during trials that correspond to one of the rules regardless of the cue presented or which object was remembered. Here, we call those **rule** units. The first panel in Figure 1 *(Right Top)* shows the activations of one *rule* unit we found in the LSTM, averaged across 400 held-out trials. The main feature that characterizes *rule* units is the clear separation between the `match` and `non-match` activations during the delay period regardless of which cue was used to signify the rule. On the other hand, the second panel in Figure 1 *(Right Top)* shows another type of unit identified in the LSTM that codes for the answer (i.e. either `Yes` or `No`). We call these **answer** units. They can be classified visually by looking at the test period when the second sample is presented to the agent—a clear bifurcation emerges that shows the separation between the `Yes` and `No` answers.

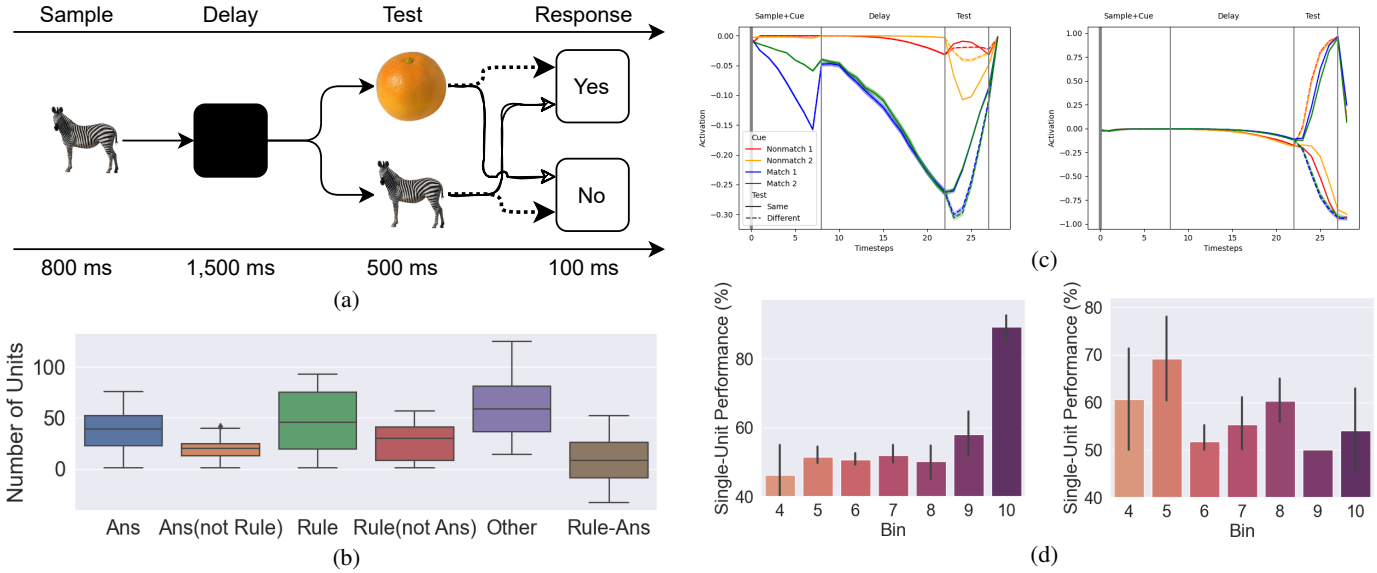

Figure 1: **(a)**: Flowchart of the four scenarios of the DMS task. Sketch arrows answer the `match` rule; dotted arrows answer the `non-match` rule. **(b)**: Boxplot shows the statistics of each unit type. **(c)**: Activations of two units, each averaged across the 400 testing images for the four cue–truth pairs. Cues 1 & 2 of each rule are distinct. **(d)**: Average *task* performance in systems where all but one unit were dropped; chosen unit was shown to lie in a certain accuracy bin (of 10% intervals) for each unit type—*answer* (Left) and *rule* (Right).

Analysis

To systematically classify the type of each unit in the LSTM (128 units), we fit a linear model on the unit's activations at a particular timestep for the classification of each of the *answer* and *rule* unit types. The accuracy of that linear model (e.g. how well it separates the `match` vs `non-match` activations) dictates how much it satisfies each of the types (*rule* vs *answer*). To understand the effect of each type of unit, we ablate the units that belong to a certain type across all timesteps in a trial. Concretely, Figure 2 shows the performance when ablating the top $x\%$ from each unit type. It is clear that the *answer* units are the ones that drive performance whereas ablating the *rule* units is similar to dropping the same number of randomly selected units. We then ask whether a single *answer* unit is enough to achieve high performance on its own (i.e. dropping all other units in the LSTM on all timesteps) since it already encodes the answer. That was indeed the case—there were some models that had *answer* units that achieved a high accuracy ($\sim 99\%$) on their own. In Figure 1 *(Right Bottom)* we show the average single-unit performance as a function of its corresponding accuracy bin of only one model as an example. It can be seen that *answer* units that belong to the high end of the spectrum achieve higher performance (i.e. are already relied on by the **policy network**), whereas it does not matter for the *rule* units what bin they are in.

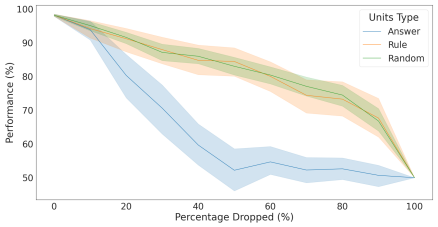

Figure 2: Performance when dropping the top $x\%$ of units per unit type on all timesteps.

Conclusion

We show that the *rule* units previously focused on in research on the *DMS* task are not *necessarily* used during inference time by the policy network. Future research should address the behavior of these units during training and their role in the emergence of *answer* units—a unit type which *is* utilized by the policy network. Notably, these results put into question the belief that observing a neuron that encodes some task variable should imply a causal link between that task variable and the resultant behavior of the agent.

References

[1] Farshad Mansouri, David Freedman, and Mark Buckley. Emergence of abstract rules in the primate brain. *Nature reviews. Neuroscience*, 21, 09 2020.

[2] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. https://distill.pub/2021/multimodal-neurons.

[3] Jonathan Wallis, Kathleen Anderson, and Earl Miller. Single neurons in prefrontal cortex encode abstract rules. *Nature*, 411:953–6, 07 2001.

[4] Timothy Brady, Talia Konkle, George Alvarez, and Aude Oliva. Visual long-term memory has a massive capacity for object details. *Proceedings of the National Academy of Sciences of the United States of America*, 105:14325–9, 10 2008.

[5] Jane X. Wang, Zeb Kurth-Nelson, Hubert Soyer, Joel Z. Leibo, Dhruva Tirumala, Rémi Munos, Charles Blundell, Dharshan Kumaran, and Matt M. Botvinick. Learning to reinforcement learn. *ArXiv*, abs/1611.05763, 2016.

[6] Sam Ritter, Ryan Faulkner, Laurent Sartran, Adam Santoro, Matt Botvinick, and David Raposo. Rapid Task-Solving in Novel Environments. page 16, 2020.