Multitasking leads to Generalizable Disentangled Representations in RNNs

Pantelis Vafidis (pvafeidi@caltech.edu)
Computation & Neural Systems, California Institute of Technology
1200 E California Blvd, Pasadena, CA 91125 USA

Aman Bhargava (abhargav@caltech.edu)
Computation & Neural Systems, California Institute of Technology
1200 E California Blvd, Pasadena, CA 91125 USA

Antonio Rangel (arangel@caltech.edu)
Humanities and Social Sciences, California Institute of Technology
1200 E California Blvd, Pasadena, CA 91125 USA

Abstract

Abstract, or disentangled, representations are a promising mathematical framework for efficient and effective generalization in both biological and artificial systems. We investigate abstract representations in the context of multi-task classification over noisy evidence streams – a canonical decision-making neuroscience paradigm. We derive theoretical bounds that guarantee the emergence of disentangled representations in the latent state of any optimal multi-task classifier, when the number of tasks exceeds the dimensionality of the state space. Turning to simulations, we confirm that RNNs trained on multi-task classification learn disentangled representations in the form of continuous attractors, and zero-shot generalize out-of-distribution (OOD). We demonstrate the flexibility of the abstract RNN representations across various decision boundary geometries and in tasks requiring classification confidence estimation. Closely relating to representations found in humans and animals during decision-making and spatial reasoning tasks, our framework suggests a general principle for the formation of cognitive maps that organize knowledge to enable flexible generalization in biological and artificial systems alike.

Keywords: RNNs; representation learning; disentanglement; multi-task learning; zero-shot generalization; continuous attractors; evidence accumulation; NeuroAI; cognitive maps

Introduction

Humans and animals can decide between previously unencountered options effortlessly (Lake, Ullman, Tenenbaum, & Gershman, 2016; Bongioanni et al., 2021). One mechanism for such OOD generalization is to first decompose a stimulus to its attributes. For example, imagine you are at a grocery store, trying to pick a mango. You might have never seen a mango before, but you can still infer whether one is ripe or not, based on your experience with other fruit, e.g. bananas. Crucially, such inferences are automatic when the brain's representation of these stimuli is compositional, each axis of the representation corresponding to an attribute; training a linear decoder (i.e. a downstream neuron) to differentiate ripe vs. unripe bananas zero-shot generalizes to mangos (Fig. 1a).

Such representations have been coined abstract (Saez, Rigotti, Ostojic, Fusi, & Salzman, 2015; Bernardi et al., 2020), or (in the ML literature) disentangled (Higgins et al., 2017), and have been linked to OOD generalization. Johnston and Fusi (2023) showed that feedforward neural networks develop abstract representations when trained to multitask. However, real-world decisions evolve dynamically over time, as the decision-maker collects information about attributes of available options (e.g. caloric content, nutrients etc. for food) and integrates this information towards a final decision (Krajbich, Armel, & Rangel, 2010). Therefore, to accommodate for the crucial ability to dynamically update evolving beliefs over time for decision-making, we train recurrent neural networks (RNNs) to multitask canonical evidence accumulation tasks.

Results

Model and Setup

In our setting, RNNs receive two noisy evidence streams $\mathbf{X}(t) \in \mathbb{R}^2$ (although we extend to higher input dimensionality D later), i.e. $\mathbf{X}(t) = \mathbf{x}^* + \sigma\,\mathcal{N}(\mathbf{0}, \mathbb{I})$, where $\mathbf{x}^* = [x_1, x_2]^T$ is the true evidence in a certain trial ($x_i \sim \textit{Uniform}(-0.5, 0.5)$) and σ is the input noise standard deviation, and a fixation input triggering the final decision (Fig. 1c). x_1 and x_2 correspond to different options under consideration or different attributes of the same item. The inputs are passed through a static encoder which non-linearly mixes them and increases their dimensionality, akin to a perceptual system. The target output $\mathbf{y}(\mathbf{x}^*)$ is a vector of +1s and -1s, depending on whether $\mathbf{x}^*$ is above or below each of the classification lines (Fig. 1b). For example, if x_1 is food and x_2 water reward, lines of different slopes correspond to preference of one or the other depending on the animal's internal state. We train the network for all tasks simultaneously with backpropagation-through-time (BPTT).

Multitasking leads to disentangled representations

Fig. 1d shows the top 3 PCs of RNN activity for a network trained on 24 tasks. Each trial is a line, while color saturation indicates time in the trial. Trials start from the center and move outwards, according to the location in state space $\mathbf{x}^*$ for this trial corresponds to. To map the representation space to the state space, we color the last timepoint in each trial (squares)

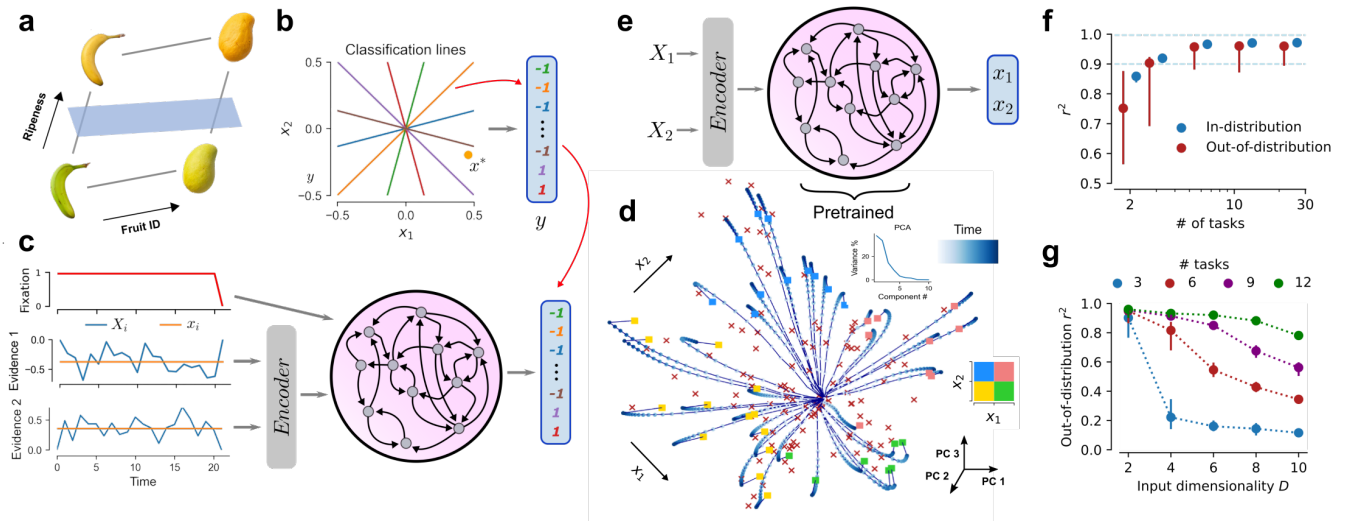

Figure 1: (a) An abstract representation. (b) Data generating process. Each line is a task. (c) Multitasking RNN architecture. (d) Representation learned by the RNN (top 3 PCs, capturing 85% of the total variance). (e)-(f) Zero-shot OOD generalization setting and performance. (g) Simulations validate our theory: OOD generalization is achieved when $N_{task} \geq D$.

according to the quadrant the trial was drawn from. We find that the network learns a 2D continuous attractor (red x's) that provides a disentangled representation of the 2D state space. This implies that the network can store and update an estimate of $\mathbf{x}^*$ in short-term memory, from which it can presumably generalize to any other task involving these variables.

To test that, we keep network weights fixed and train a linear decoder to output $\mathbf{x}^*$ at the end of the trial (Fig. 1e). We perform OOD 4-fold crossvalidation, i.e. train the decoder on 3 out of 4 quadrants and test on the remaining quadrant. We also evaluate in-distribution (ID) performance by training the decoder in all quadrants. We find that ID performance (as quantified by r^2) increases with the number of tasks, and the OOD generalization gap decreases until OOD and ID generalization performance are almost identical (Fig. 1f); therefore the network has learned an abstract representation that generalizes OOD in a zero-shot fashion. The same conclusions hold when training RNNs for free reaction-time tasks where they have to report confidence in their decisions (Krajbich et al., 2010), and for non-linear tasks (data not shown).

A theory of OOD generalization

We sought to understand the properties of optimal multi-classifiers in the paradigm illustrated in Fig. 1b,c. We denote the set of classification estimates as $\hat{\mathbf{Y}}$, a vector of Bernoulli random variables. We prove (not included here for brevity) that **any optimal multi-task classifier** with i.i.d. noisy inputs $\mathbf{X}(1), \ldots, \mathbf{X}(t)$ **implicitly estimates the ground truth coordinate $\mathbf{x}^*$ in its latent state $\mathbf{Z}(t)$ in a disentangled format**.

$$\mathbf{x}^* \rightarrow \{\mathbf{X}(t)\} \rightarrow \mathbf{Z}(t) \rightarrow \hat{\mathbf{Y}}(t) \tag{1}$$

Specifically, we prove that $\mathbf{Z}(t)$ is an abstract representation of $\mathbf{x}^*$ as long as the number of classification tasks N_{task} exceeds the dimensionality of the state space D. This result holds for *any* system that performs optimal multi-task classification with a latent variable separating the inputs from the outputs (e.g., RNNs, Bayesian filters, etc.), **regardless of the internal dynamics of the latent state**.

To test our theory, we run RNN simulations increasing D (i.e. adding more noisy inputs to Fig. 1c), while varying N_{task}. Fig. 1g shows OOD generalization performance for various combinations of D and N_{task}. We observe that performance is bad when the $N_{task} < D$, but it increases when $N_{task} \geq D$. This increase is more gradual for higher D, which is in line with remarks by Johnston and Fusi (2023) that it is easier to learn abstract representations for high D. Overall, these findings confirm our theory that in order to learn representations that generalize OOD, N_{task} should exceed D. This result is remarkable, especially for high D, because it goes against our intuition that N_{task} should scale exponentially with D to adequately fill up the space in order to provide enough information to localize $\mathbf{x}^*$; instead it need only scale linearly.

Discussion

We here show that multitasking readily leads to representations that can OOD generalize to any downstream task involving the same variables, and develop a theory that explains why. So far, the workhorse model for decision-making has been context-dependent computation, where a single task is carried out at a time and task identity is cued to the RNN by a one-hot vector (Mante, Sussillo, Shenoy, & Newsome, 2013). However, context-dependent decision making results in collapsed representations that utilize separate parts of the state space for different tasks (Yang & Wang, 2020), scaling badly with the number of tasks (linearly) and of variables (exponentially). Such inefficiency could be detrimental for brains, which

need to pack a lot of computation within a large yet limited neural substrate. Multitasking results in compact, state-space efficient representations that can be used for any downstream task, and scale linearly with the number of variables. Both types of representations are likely to be found in the brain.

Our findings closely relate to representations found in monkeys during novel inferential choices (Bongioanni et al., 2021), similar orthogonal representations found in humans (Flesch, Nagy, Saxe, & Summerfield, 2022), and to the problem of path-integration where non-abstract 2D continuous attractors are learned (Sorscher, Mel, Ocko, Giocomo, & Ganguli, 2023).

Overall, these findings shed light in the conditions under which biological and artificial systems alike develop representations that generalize well: they do so when there is enough pressure from many tasks that involve the same latent variables. If on the other hand only a single task is to be accomplished, a system is more likely to rely on input-output mappings, overfitting to the task. Apart from understanding learning in brains, we hope this work will inspire the development of deep learning systems with OOD generalization in mind.

Acknowledgments

Supported by Onassis Foundation Scholarship (PV) and NOMIS Distinguished Scientist and Scholar Award (AR).

References

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020, November). The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, *183*(4), 954–967.e21. Retrieved from https://doi.org/10.1016/j.cell.2020.09.031 doi: 10.1016/j.cell.2020.09.031

Bongioanni, A., Folloni, D., Verhagen, L., Sallet, J., Klein-Flügge, M. C., & Rushworth, M. F. S. (2021, January). Activation and disruption of a neural mechanism for novel choice in monkeys. *Nature*, *591*(7849), 270–274. Retrieved from https://doi.org/10.1038/s41586-020-03115-5 doi: 10.1038/s41586-020-03115-5

Flesch, T., Nagy, D. G., Saxe, A., & Summerfield, C. (2022). *Modelling continual learning in humans with hebbian context gating and exponentially decaying task signals.* arXiv. Retrieved from https://arxiv.org/abs/2203.11560 doi: 10.48550/ARXIV.2203.11560

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., ... Lerchner, A. (2017). beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations.* Retrieved from https://openreview.net/forum?id=Sy2fzU9gl

Johnston, W. J., & Fusi, S. (2023, February). Abstract representations emerge naturally in neural networks trained to perform multiple tasks. *Nature Communications*, *14*(1). Retrieved from https://doi.org/10.1038/s41467-023-36583-0 doi: 10.1038/s41467-023-36583-0

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298. Retrieved from https://doi.org/10.1038/nn.2635 doi: 10.1038/nn.2635

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016, November). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*. Retrieved from https://doi.org/10.1017/s0140525x16001837 doi: 10.1017/s0140525x16001837

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013, November). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), 78–84. Retrieved from https://doi.org/10.1038/nature12742 doi: 10.1038/nature12742

Saez, A., Rigotti, M., Ostojic, S., Fusi, S., & Salzman, C. (2015, August). Abstract context representations in primate amygdala and prefrontal cortex. *Neuron*, *87*(4), 869–881. Retrieved from https://doi.org/10.1016/j.neuron.2015.07.024 doi: 10.1016/j.neuron.2015.07.024

Sorscher, B., Mel, G. C., Ocko, S. A., Giocomo, L. M., & Ganguli, S. (2023, January). A unified theory for the computational and mechanistic origins of grid cells. *Neuron*, *111*(1), 121–137.e13. Retrieved from https://doi.org/10.1016/j.neuron.2022.10.003 doi: 10.1016/j.neuron.2022.10.003

Yang, G. R., & Wang, X.-J. (2020, September). Artificial neural networks for neuroscientists: A primer. *Neuron*, *107*(6), 1048–1070. Retrieved from https://doi.org/10.1016/j.neuron.2020.09.005 doi: 10.1016/j.neuron.2020.09.005