# Cognitive abstractions for data-efficient learning, systematic generalization, and latent factor inference.

**Felipe del Rio (fidelrio@uc.cl)**
Computer Science Department, Universidad Catolica
Santiago, Chile

**Eugenio Herrera (eugenio.herrera@cenia.cl)**
Centro Nacional de Inteligencia Artificial
Santiago, Chile

**Julio Hurtado (julio.hurtado@warwick.ac.uk)**
CAMaCS, University of Warwick
Coventry, UK

**Alvaro Soto (asoto@ing.puc.cl)**
Computer Science Department, Universidad Catolica
Santiago, Chile

**Ali Hummos* (ahummos@MIT.edu)**
Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology
Cambridge, MA, USA

**Cristian B. Calderon* (cristian.buc@cenia.cl)**
Centro Nacional de Inteligencia Artificial
Santiago, Chile

*Co-senior authors

## Abstract

**Humans excel at generalization. Recent studies suggest that the latter emerges from systematic compositionality: The ability to adapt to (or understand) novel contexts based on the flexible recombination of previously learned (sub)concepts (or "cognitive abstractions"). Here, we propose a framework to study the effects of cognitive abstractions by leveraging standard generative neural networks. As predicted by empirical human studies, neural networks with cognitive abstractions, learn faster and generate systematic out-of-distribution (OOD). Moreover, we show that these cognitive abstractions can be used to infer the underlying latent factor that generate the images. Our framework can be used to analyze the representational basis that allow for the emergence of these properties.**

**Keywords:** Compositional learning; systematic generalization; auto-encoder; gradient-based inference.

Imagine wearing a red coat, a green hat, yellow pants, a red nose, and huge shoes. Aside from looking like a clown, your ability to generate that image emerges from your systematic generalization skills (Lake et al., 2017). Such skills are based on compositional learning: encoding concepts into their sub-components and systematically recombining these sub-components to make sense of novel contexts or generate novel behavior to flexibly adapt to these contexts (Calderon et al., 2022; Lake & Baroni, 2023).

Compositional learning has been shown to be useful in reinforcement learning settings (Lehnert et al., 2020; Liu & Frank, 2022). In the context of neural networks, compositional learning is suggested to depend on the neural geometry of cognitive representations (Fusi et al., 2016; Ito et al., 2022). However, how these representations are recombined in order to adapt to novel contexts remains unclear. Recent work suggests that flexible control (or recombination) of these representations may be subtended by thalamo-cortical projections that flexibly gate processing in prefrontal areas (Halassa & Sherman, 2019). Such mechanistic implementations in neural networks have been shown to produce robust context-dependent task performance (Flesch et al., 2022), avoid catastrophic forgetting in continual learning (Hummos, 2023), promote task structure generalization (Collins & Frank, 2013), and balance the trade-off between shared versus separated task representations (Verbeke & Verguts, 2022).

In this work, we present a framework to study the effects of providing cognitive abstractions to neural networks, and test the emerging properties associated with this addition. We propose **C**ognitive **A**bstraction-**BA**sed compositional **GE**neralization (CABAGE), a framework inspired by gating functions of the thalamus. As predicted by human learning studies (Dekker et al., 2022; Lake et al., 2017), we show that

these abstractions generate several interesting properties, in an image generation setting. Moreover, CABAGE provides a natural solution to infer cognitive abstractions via gradient-based inference (GBI; see below). In the remainder of the manuscript we: (i) provide a formal description of CABAGE, (ii) show that it allows for faster learning, (iii) OOD image generation and compositional recombination, (iv) and can be used to infer the latent factors generating the image.

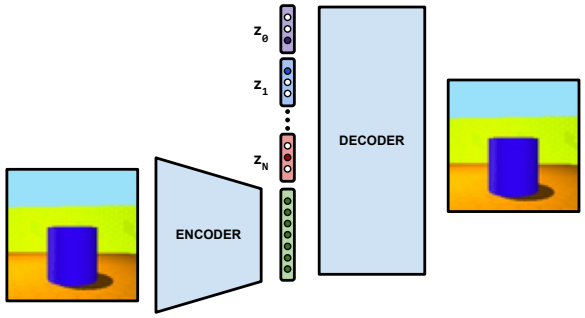


Figure 1: CABAGE network structure: We provide an (variational) auto-encoder with an extended cognitive abstraction embedding vector $(z_0, z_1, ..., z_N)$ which encodes the latent factors used to generate the images in the dataset, concatenated with the usual latent embedding (green encoding).

**CABAGE formalization.** We set up CABAGE within the (V)AE framework (Kingma & Welling, 2022). We pad an additional cognitive abstraction layer to the usual latent embedding of (V)AEs (Fig. 1)[1]. This padding is built via the concatenation of one-hot encoding representations of each latent factor $(z_0, z_1, ..., z_N)$. During training, the network is presented with an image while simultaneously activating the nodes (in the cognitive abstraction layer) representing the ground truth factors that generated the image. The goal for the network is to minimize the pixel-wise mean-squared error (MSE) reconstruction loss:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (1)$$

We train the networks to reconstruct images from the 3D-Shapes dataset (Burgess & Kim, 2018). Images of this dataset are generated from 6 ground truth-independent latent factors. Such a dataset allows us to divide the dataset in training trials, test trials, and OOD trials (i.e., combination of latent factors that has never been seen by the networks), for this work we follow the compositional splits provided by Schott et al., 2022.

**CABAGE learns better and faster.** As shown in figure 2, CABAGE displays steeper learning curves (indicative of faster learning) compared to traditional (V)AEs. Note that this is true even in cases where the dimensionality of the latent embedding is equated between both network types. Moreover, the MSE loss reaches a lower level in CABAGE, hence the quality of the reconstruction is higher.

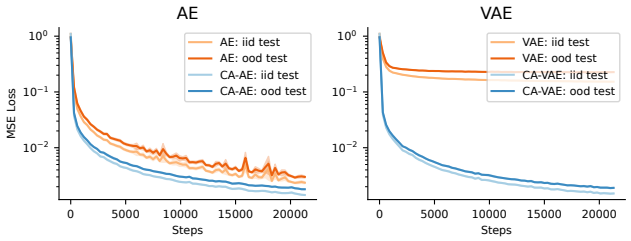


Figure 2: A. Learning dynamics for AE (left) and VAE (right) with (blue) and without (red) cognitive abstractions (CA) on iid and ood examples.

**CABAGE produces OOD and systematic generalizations.** Figure 2 further displays the ability of CABAGE to produce OOD generalization. The darker shade lines reflect the MSE loss of trials that combine latent factors never seen before by the network. Compared with traditional (V)AEs, CABAGE can generate OOD images with a lower MSE loss. Moreover, figure 3A displays examples of OOD image generation starting from a noisy picture (figure 3A), or even in the absence of any input from the latent embedding, i.e. the latent embedding is clipped and activation to the decoder is only generated by activating the cognitive abstraction layer (figure 3B). Importantly, figure 3C also shows that setting the cognitive abstraction node values to states that combine two latent factors generates images that systematically combine these factors. Here, we show an example where we co-activate the cube and spherocylinder cognitive abstractions, thereby generating an "in-between" geometric shape.

**CABAGE allows quick latent factor inference.** An important property of CABAGE is its ability to rapidly infer the latent factors generating a given image. To do so, we perform gradient-based inference (GBI), a one-iteration inference process. The latter entails taking the reconstruction loss gradients at the cognitive abstraction layer (i.e., gradients are back-propagated up to that layer), normalizing them within each vector representing the distinct factors, and selecting the node with the highest probability. Figure 4 shows that GBI can classify the latent factors generating the image much higher than chance-level both for in-distribution and out-of-distribution trials.

**Conclusions.** We propose a framework to study the addition of cognitive abstractions in neural networks. As predicted by human learning, we show that these abstractions generate fast learning, systematic generalization, and quick latent factor inference. Moreover, they can be recombined for flexible adaptation (or image generation in this case). CABAGE allows for future work to explore the representations that underlie such properties.

[1] A formal description of the network architectures can be found here: https://anonymous.4open.science/r/gradient_based_inference-A2DD/

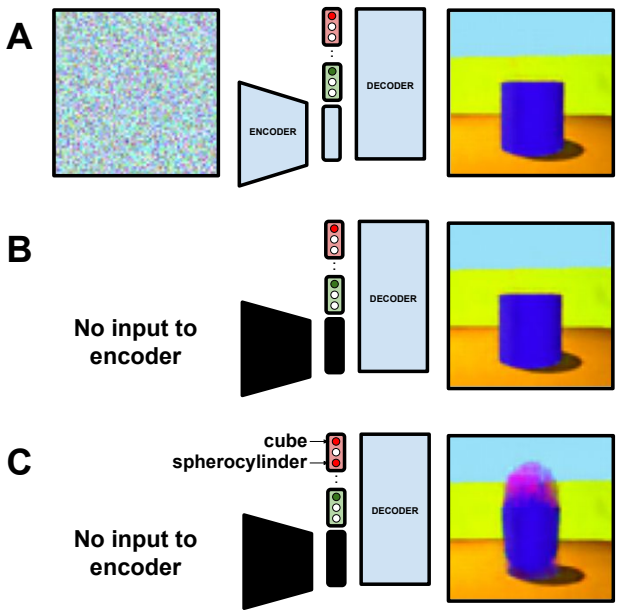


Figure 3: A. OOD generalization from a noisy input image. B. OOD generalization only from the cognitive abstraction layer. C. Systematic generalization by recombining cognitive abstractions.

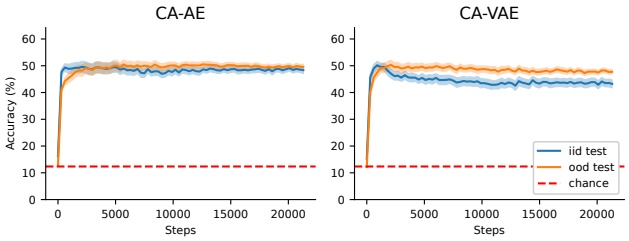


Figure 4: Gradient-based inference (GBI) of latent factors for CA-AE (left) and CA-VAE (right). The distinct lines depict the accuracy of GBI as a function of learning steps.

## Acknowledgments

This work was funded by *will be added after double blind review process.*

## References

Burgess, C., & Kim, H. (2018). 3d shapes dataset.

Calderon, C. B., Verguts, T., & Frank, M. J. (2022). Thunderstruck: The acdc model of flexible sequences and rhythms in recurrent neural circuits. *PLoS Computational Biology*, *18*(2), e1009854.

Collins, A. G., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological review*, *120*(1), 190.

Dekker, R. B., Otto, F., & Summerfield, C. (2022). Curriculum learning for human compositional generalization. *Proceedings of the National Academy of Sciences*, *119*(41), e2205582119.

Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., & Summerfield, C. (2022). Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, *110*(7), 1258–1270.

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: High dimensionality for higher cognition. *Current opinion in neurobiology*, *37*, 66–74.

Halassa, M. M., & Sherman, S. M. (2019). Thalamocortical circuit motifs: A general framework. *Neuron*, *103*(5), 762–770.

Hummos, A. (2023). Thalamus: A brain-inspired algorithm for biologically-plausible continual learning and disentangled representations. *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=6orC5MvgPBK

Ito, T., Klinger, T., Schultz, D., Murray, J., Cole, M., & Rigotti, M. (2022). Compositional generalization through abstract representations in human and artificial neural networks. *Advances in Neural Information Processing Systems*, *35*, 32225–32239.

Kingma, D. P., & Welling, M. (2022). Auto-encoding variational bayes.

Lake, B. M., & Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. *Nature*, *623*(7985), 115–121.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*, e253.

Lehnert, L., Littman, M. L., & Frank, M. J. (2020). Reward-predictive representations generalize across tasks in reinforcement learning. *PLoS computational biology*, *16*(10), e1008317.

Liu, R. G., & Frank, M. J. (2022). Hierarchical clustering optimizes the tradeoff between compositionality and expressivity of task structures for flexible reinforcement learning. *Artificial intelligence*, *312*, 103770.

Schott, L., Kügelgen, J. V., Träuble, F., Gehler, P. V., Russell, C., Bethge, M., Schölkopf, B., Locatello, F., & Brendel, W. (2022). Visual representation learning does not generalize strongly within the same domain. *International Conference on Learning Representations*. https://openreview.net/forum?id=9RUHPladgh

Verbeke, P., & Verguts, T. (2022). Using top-down modulation to optimally balance shared versus separated task representations. *Neural networks*, *146*, 256–271.