“Conscious” Multi-Modal Perceptual Learning for Grounded Simulation-Based Cognition

Sean Kugele (seankugele@gmail.com) and Stan Franklin (franklin.stan@gmail.com)
Department of Computer Science and Institute for Intelligent Systems, University of Memphis
Memphis, TN 38152 USA

Abstract

Barsalou (1999) presented a simulation-based theory of grounded cognition called Perceptual Symbol Systems. According to this theory, a fully functional conceptual system can be implemented using only modal representations (aka perceptual symbols) and simulations. While the theory has gained considerable neuroscientific and experimental support, there is an urgent need for computational accounts that flesh out the theory. The current paper explores one approach for implementing these computational foundations. We present an implementation of perceptual symbols, simulators, simulation-based perception, and “conscious” multi-modal perceptual learning based on generative neural networks, called β-variational autoencoders, combined with LIDA, a biologically-inspired cognitive architecture. We show that our implementation satisfies many of the properties attributed to perceptual symbol systems, and provides a solid foundation for future computational work in perception, categorization, and simulation-based cognition.

Keywords: perceptual symbol systems; multi-modal perception; mental simulation; LIDA; unsupervised machine learning

Introduction

Barsalou (1999) argued that “cognition is inherently perceptual,” using similar mental representations and processes as perception. He demonstrated that a “fully functional conceptual system” could, in theory, be implemented using *modal representations* (representations grounded[1] in sensory and motor systems) and *modal simulations* (the reenactment of previously learned perceptual and motor states). He referred to the resulting architecture as a perceptual symbol system (PSS).

Barsalou’s approach relies heavily on generative processes called *simulators* that collectively form the basis of an individual’s conceptual system. “A concept is equivalent to a simulator” according to Barsalou’s theory, and once individuals can simulate an object, entity, or event accurately and reliably, they can be said to “understand” it. Based on this, Barsalou concluded that “the primary goal of human learning is to establish simulators.”

While there is a growing body of neuroscientific and experimental support for PSSs, computational mechanisms are needed to further validate the theory. Barsalou (2009) stated, “Perhaps the most pressing issue surrounding this area of work is the lack of well-specified computational accounts. Our understanding of simulators, simulations, situated conceptualizations and pattern completion inference would be much deeper if computational accounts specified the underlying mechanisms.”

The goal of this paper is to explore one approach for implementing the computational foundations of a PSS. The present work focuses on perceptual symbols, simulators, simulation-based perception, and “conscious” multi-modal perceptual learning. Our implementation combines generative neural networks, called β-variational autoencoders (Higgins, et al., 2017), with LIDA (Franklin, et al., 2016), a biologically-inspired cognitive architecture. We will argue that our approach satisfies many of the properties attributed to perceptual symbol systems, and provides a solid foundation for future work in perception, categorization, and simulation-based cognition. We believe that continued research in this direction will lead to theoretical advances in both PSSs and LIDA, and may inspire similar approaches in other cognitive architectures and computational frameworks.

Background

In this section, we review the core components of a perceptual symbol system, as outlined in Barsalou (1999), namely, perceptual symbols, simulators, and simulations. We will also provide a brief introduction to LIDA and variational autoencoders (VAEs).

Perceptual Symbols

Barsalou (1999) argued that the patterns of activation occurring in sensory and motor systems during perception and action can be learned into long-term memory, albeit in a partial and attenuated form. If later recalled (i.e., reactivated), these perceptual representations, which he called *perceptual symbols*, can signify entities, objects, and events in the world.

Perceptual symbols are (1) **modal**, grounded in modality-specific sensory and motor representations, (2) **analogical**, sharing properties with, and likely bearing some structural resemblance to, their originating perceptual states, (3) **not complete recordings of perceptual states**, reflecting only their most salient or important aspects, (4) **dynamic** (i.e., their reactivations are sensitive to differences in context and changes in nearby regions of long-term memory resulting in variable reconstructions), and (5) **componential**, representing a conjunction of independently activatable feature dimensions (e.g., shape, orientation, and color).

[1] See (Harnad, 1990) for more information on the meaning of “grounding” and “the symbol grounding problem.”

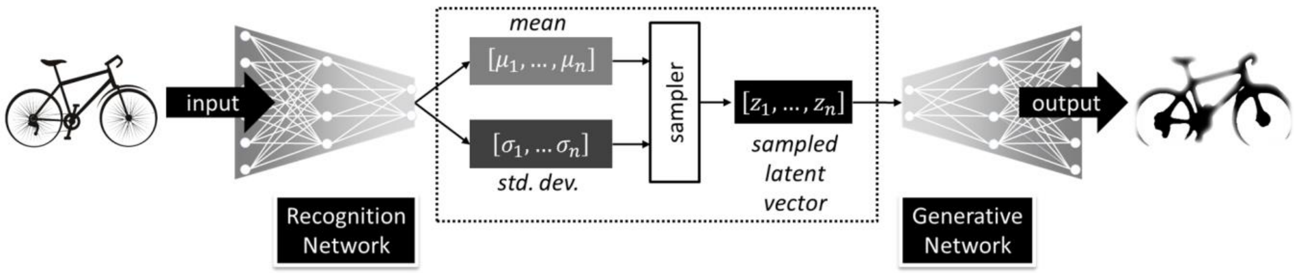

Figure 1: Depiction of a variational autoencoder (VAE).

Simulators and Simulations

Perceptual symbols become integrated into *simulators*, which construct *simulations* of their associated concepts. Simulators encompass "the knowledge and accompanying processes that allow an individual to represent some kind of entity or event adequately" (Barsalou, 1999). An individual can be said to "understand" a concept once a simulator is learned that can adequately simulate that concept. Simulations are typically preconscious[2] representations that are referred to as *mental images* if/when they become conscious.

LIDA

Learning Intelligent Decision[3] Agent (LIDA) (see Franklin et al., 2016) is a biologically-inspired cognitive architecture that provides a comprehensive theory and model of minds (both biological and artificial). LIDA also implements, and fleshes out, many psychological theories, including the Global Workspace Theory (GWT) of consciousness (Baars, 1988), making it ideal for the modeling of "preconscious simulations" and "conscious mental imagery."

Cognition occurs in LIDA over a series of *cognitive cycles*, where "cognition" in this context refers loosely to the sum total of an agent's mental activities, including, but not limited to, perception, long-term memory recall, situational understanding, attention, and action selection and execution. Each cognitive cycle can be conceptually divided into three phases: understanding, attention, and "action and learning."

During the *understanding phase*, modality-specific sensory stimuli (from the environment via an agent's sensors) are encoded into the activation of low-level features in Sensory Memory (SM). These, in turn, activate object, entity, and event representations in Perceptual Associative Memory (PAM). Sufficiently activated representations in PAM are instantiated as "percepts" in the preconscious workspace (p-Workspace). Content in the p-Workspace can also "cue" long-term memory (that is, activate long-term memory representations via associative links) causing their instantiation and integration into the p-Workspace. Specialized processors called structure building codelets (SBCs) monitor the content in the p-Workspace, and may construct complex representations that facilitate an agent's current situational understanding.

During the *attention phase*, other specialized processors called attention codelets (ACs) advocate for the salience of preconscious content in the p-Workspace. Based on their individual concerns (for example, situational or goal relevance, novelty, surprise, etc.), ACs identify preconscious content of interest to them, and collaborate with other "like-minded" ACs (i.e., those that are also interested in the same, or related, content) to form coalitions. Coalitions compete in the Global Workspace (GW), and the winning coalition's content is included in the global "conscious" broadcast. The content contained in the global broadcast is received by all LIDA modules, initiating the "action and learning" phase.

During the *action and learning phase*, module-specific learning mechanisms can create, or update, representations stored in each module. All "significant learning" in LIDA is mediated by the global broadcast, proceeding only from "conscious" content. This position is a direct consequence of LIDA's commitment to the Conscious Learning Hypothesis from GWT. (For brevity, we omit a summary of the action-related portions of the "action and learning phase," as it is not needed to understand the partial implementation of LIDA presented in this paper.)

Variational Autoencoders (VAEs)

Variational autoencoders (VAEs) (Kingma & Welling, 2014) are connectionist (neural network) architectures composed of a *recognition network* and a *generative network* (see Figure 1). Data are fed into the recognition network, which learns to generate probability distributions (e.g., Gaussians) that (hopefully) characterize the most important features of those data. *Latent vectors* are then sampled from these probability distributions and fed into the generative network, which learns to construct likenesses of the original inputs from the latent vectors.

VAEs learn by unsupervised learning, that is, from unlabeled data such as images, sounds, or other forms of uncategorized sensory stimuli. They achieve this (in part) by attempting to minimize discrepancies between the recognition network's inputs and the generative network's attempted reconstructions of those inputs (i.e., its

[2] We follow the Franklin et al. (2016) convention of using the term "preconscious" (instead of unconscious) to denote non-conscious representations that have the *potential* to become conscious.

[3] For historical reasons, this word was previously "distribution." It was later changed.

reconstruction error). β-VAEs (Higgins, et al., 2017) augment the standard VAE loss function with a penalty coefficient (β) that encourages *disentangled latent representations*, but are otherwise identical to "vanilla" VAEs. Unlike the "entangled" representations that are typically learned by standard VAEs, disentangled representations can be decomposed into subcomponents that represent distinct generative features (e.g., brightness, position, and size). These subcomponents can be manipulated to selectively control aspects of the generative process.

General Approach and Implementation

In this section, we describe our "proof-of-concept" computational implementations of perceptual symbols, simulators, and "conscious" multi-modal perceptual learning. Our approach is based on a partial implementation of the LIDA cognitive model (see Figure 2) focused on the understanding phase of LIDA's cognitive cycle. We use β-VAEs to implement Sensory Memory (SM) and "simulator" structure-building codelets (SBCs). We implement Perceptual Associative Memory (PAM) as a content-addressable "activation graph" that exhibits simple perceptual priming and typicality effects (based on the frequency, recency, and overall "strength" of "conscious" experiences). Perceptual symbols are implemented as subgraphs of this activation graph. All learning is limited to "conscious" content in accordance with the Conscious Learning Hypothesis.

In the subsections that follow, we first detail our implementations of relevant LIDA modules and processes, perceptual symbols, and simulators. We then describe how these interact to implement bottom-up perception and "conscious" multi-modal perceptual learning.

Component Implementations

Sensory Memory (SM) We implement SM using a set of modality-specific β-VAE recognition networks—one per sensory modality (see Figure 3). Incoming sensory stimuli initiate their feed-forward activation, resulting in the generation of modal probability distributions. These probability distributions are represented as vectors of means ($\vec{\mu}$) and std. deviations ($\vec{\sigma}$) that are approximately Gaussian after "enough training." Modal probability distributions are the basis for several key capabilities including the activation of modal representations in PAM and simulation.

Perceptual Associative Memory (PAM) We implement PAM using a directed, hierarchical, "activation graph." Each node in our graph has two parameters: a *current activation* (representing its current situational relevance) and a *base-level activation* (representing its historical frequency, recency, and "strength" in "conscious" broadcasts). We refer to the sum of the current and base-level activations simply as a node's activation. Activation can propagate between nodes over directed links, resulting in increased current activation in the targeted nodes.

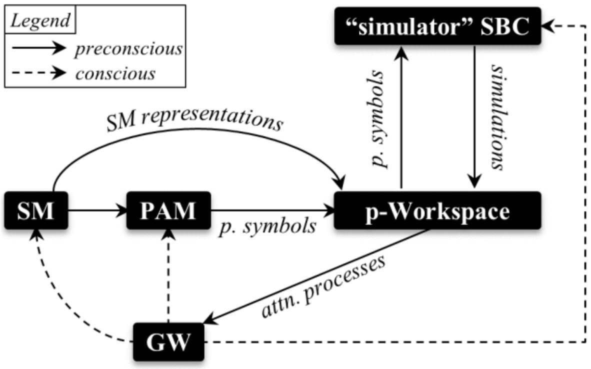

Figure 2: Our (partial) implementation of the LIDA cognitive model focused on the understanding phase. Relevant LIDA modules and processes are depicted, as well as mental representations, such as perceptual symbols (abbreviated as "p. symbols").

We differentiate between two types of PAM nodes: *primitive* and *non-primitive feature detectors*. Primitive feature detectors receive activation exclusively from SM, whereas non-primitive feature detectors can receive activation from representations in both SM and PAM. Each primitive feature detector, in our implementation, is associated with a modality indicator and a modal probability distribution. Non-primitive feature detectors, on the other hand, have neither of these attributes, as they are (typically) multi-modal, and receive all of their current activation from other PAM nodes over directed activation links.

Perceptual Symbols We implement perceptual symbols as the combination of a uniquely-assigned, non-primitive feature detector (representing some, potentially multi-modal, perceptual experience) combined with a set of modality-specific, primitive feature detectors connected to it over directed links. SM can activate these primitive feature detectors (e.g., during bottom-up perception), and part of this activation can then propagate to linked, non-primitive feature detector(s)[4]. If, as a result, the activation of a non-primitive feature detector becomes greater than an *instantiation threshold*, the entire perceptual symbol is instantiated into LIDA's p-Workspace as part of a percept. We implement this instantiation operation as the insertion of a *reference*[5] to the perceptual symbol in the p-Workspace.

[4] A primitive feature detector can be (and frequently is) associated with multiple non-primitive feature detectors.

[5] We use *references* (not *copies*), so that perceptual symbols and their instantiations can share the same parameter values.

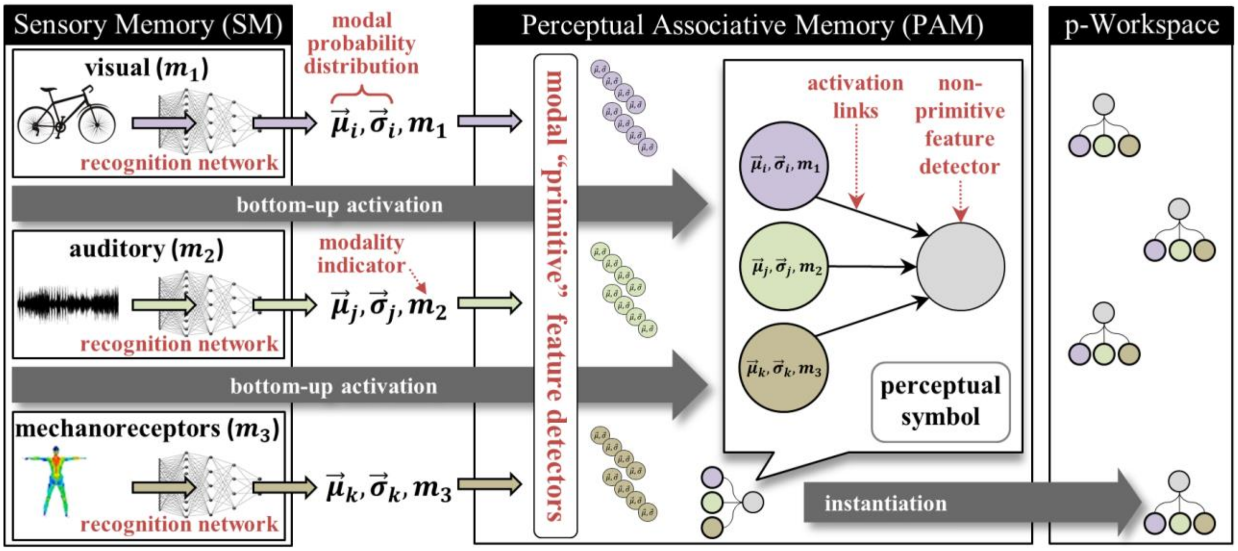
Sensory Memory (SM)
modal probability distribution
visual (m_1)
recognition network
$\vec{\mu}_i, \vec{\sigma}_i, m_1$
bottom-up activation
auditory (m_2)
modality indicator
recognition network
$\vec{\mu}_j, \vec{\sigma}_j, m_2$
bottom-up activation
mechanoreceptors (m_3)
recognition network
$\vec{\mu}_k, \vec{\sigma}_k, m_3$
Perceptual Associative Memory (PAM)
modal "primitive" feature detectors
activation links
non-primitive feature detector
$\vec{\mu}_i, \vec{\sigma}_i, m_1$
$\vec{\mu}_j, \vec{\sigma}_j, m_2$
$\vec{\mu}_k, \vec{\sigma}_k, m_3$
perceptual symbol
instantiation
p-Workspace

Figure 3: Bottom-up perception. Modality-specific recognition networks (in SM) generate modal probability distributions, which activate primitive feature detectors (in PAM). Activation spreads through PAM's activation graph over directed links, activating perceptual symbols. If they receive enough activation, perceptual symbols are instantiated into the p-Workspace.

Simulators and Simulation Barsalou (1999) defined simulators as the combination of knowledge, and generative processes that use that knowledge, to construct simulations. For our proof-of-concept implementation, these are implemented as perceptual symbols and "simulator" SBCs, respectively. Future work will extend our knowledge representations to more complex "conceptual" representations, such as frames (see Barsalou, 1999).

Our simulator SBC is composed of a stochastic "sampler" and a set of modality-specific generative networks. The sampler generates a latent vector from each modal probability distribution associated with a perceptual symbol, which then activates the generative networks creating a set of modal simulations. The SBC associates these simulations with their corresponding perceptual symbol in the p-Workspace.

Attention Our attentional processes consist of two ACs and a coalition-forming process. The first AC selects perceptual symbols based on their activation, and the second their *reconstruction error*[6]. A high reconstruction error indicates that an agent lacks the experience to adequately simulate an object, entity, or event, and can be interpreted as an indication of "surprise."

Whenever both ACs advocate for the same perceptual symbol, our coalition-forming process creates a single coalition containing that perceptual symbol and adds it to the GW; otherwise, two (competing) coalitions are added. When a competition is "triggered" in the GW, content from the coalition with the highest activation is included in the global (conscious) broadcast. In the LIDA conceptual model, a coalition's activation can be based on many factors (see Franklin et al., 2016); however, for our simple implementation, we base it solely on the activations of the selected perceptual symbols and the magnitudes of their associated reconstruction errors.

Activation Decay and Structural Pruning Decay is applied to the current and base-level activations of PAM nodes following each conscious broadcast, with current activations decaying much more rapidly than base-level activations. Perceptual symbols are pruned from PAM's activation graph if their base-level activations are less than the PAM *removal threshold*. References to perceptual symbols are removed from the p-Workspace when their activation crosses below the instantiation threshold.

Perception (Bottom-Up)

We define bottom-up perception (see Figure 3) as a feed-forward process that begins with the arrival of sensory stimuli in SM and ends with a global (conscious) broadcast. Each step of this process is described below.

(1) Bottom-up perception begins when sensory stimuli activate SM's modality-specific β-VAE recognition networks, resulting in the generation of modal probability distributions.

[6] The reconstruction error can only be calculated after the simulator SBC has constructed modal simulations for a perceptual symbol; as a result, our ACs are constrained to select from the subset of perceptual symbols in the p-Workspace which have associated modal simulations.

(2) SM updates the current activation of primitive feature detectors (in PAM) based on their cosine similarity with the probability distributions generated in step (1).

(3) *In parallel to step 2*, SM's probability distributions are encapsulated in new nodes that are sent to the p-Workspace along with the multi-modal sensory inputs that generated them.

(4) Base-level and current activations combine (in PAM) to determine the activation of each primitive feature detector. Activation propagates over directed links to connected non-primitive feature detectors associated with perceptual symbols. Perceptual symbols with activations over the instantiation threshold are instantiated into the p-Workspace.

(5) A simulator SBC continually scans the p-Workspace looking for perceptual symbols *without associated simulations*. For each such perceptual symbol, the simulator SBC constructs a set of modal simulations, which it associates with the perceptual symbol.

(6) ACs scan the p-Workspace looking for perceptual symbols *with associated simulations*, and selects from among these based on their own interests (e.g., activation or reconstruction error). Selected perceptual symbols are sent to the coalition forming process, which constructs coalition(s) and sends them to the GW.

(7) The GW conducts an activation-based, winner-take-all competition among the coalitions, and globally broadcasts the winning coalition's content.

"Conscious" Perceptual Learning

Learning Perceptual Symbols New perceptual symbols are constructed by an SBC in the p-Workspace. This SBC scans the p-Workspace for groups of unattached modal nodes from SM (created during step (3) of the bottom-up perceptual process described earlier). If found, the SBC creates a new node that will function as the perceptual symbol's non-primitive feature detector, and directed links to it from each unattached modal node in a group. These nodes will later function as primitive feature detectors. The originating multi-modal sensory stimuli (i.e., the inputs to the SM recognition networks) are also associated with this new structure. If/when this "proto-perceptual symbol" is attended to by ACs and consciously broadcast to PAM, it will be learned into PAM as a new perceptual symbol.

Updating Base-Level Activation When PAM receives a "conscious" broadcast, it increases the base-level activation of each node in its activation graph that was present in the broadcast. The magnitude of this increase is based on the "strength" of the conscious broadcast (i.e., the activation of the winning coalition).

Updating β-VAE Parameters Following the "conscious" broadcast of a perceptual symbol, the recognition and generative networks are updated based on the β-VAE loss using stochastic gradient descent. The calculation of the loss function requires the original stimuli, their corresponding simulations, and the modal probability distributions associated with that perceptual symbol.

Evaluation

The viability of our approach depends on whether our SM representations (i.e., modal probability distributions) have several properties. First, the (cosine) similarity between two SM representations must serve as a reliable proxy for the degree of resemblance between their corresponding sensory stimuli. We refer to this as the *property of analogical representations*. Second, our SM representations must capture enough information about their originating sensory stimuli to enable the construction of simulations that "sufficiently" resemble those stimuli. We refer to this as the *sufficiency of generative representations*.

In the remainder of this section, we describe a series of experiments that demonstrate the feasibility of learning SM representations (and, by extension, perceptual symbols) that satisfy these properties. For brevity, we focus on a single sensory modality; however, the same approach can be easily extended to multiple modalities by confirming these properties independently[7] for each modality.

Experimental Setup

We trained a β-VAE with a convolutional architecture (see Krizhevsky, Sutskever, & Hinton, 2012) on Fashion MNIST (Xiao, Rasul, & Vollgraf, 2017)—a well-known data set containing $70,000$ grayscale images (28×28 pixels each) from ten different categories of "fashion products." Training consisted of five training epochs[8] over the data set's 60,000 "training" images. Our β-VAE had 538,529 parameters (i.e., weights and biases), our latent vectors ($\vec{z}$) had 128 dimensions, and we used a β value of 1.2. We calculated the cosine similarity (δ) for two probability distributions over their means ($\vec{\mu}$), and the current activation (α_c) using the sigmoidal function

$$\alpha_c(\delta) = \frac{1}{1 + e^{(-15\delta+10)}} \quad .$$

All demonstrations that follow are based on the data set's 10,000 "test" images, which were unseen during training.

Analogical Representations

We randomly selected 250 images (25 per object class) and generated their modal probability distributions by feed-

[7] This follows from the fact that our SM representations are only (directly) used to (1) activate modality-specific primitive feature detectors in PAM and (2) construct unimodal simulations. And, that multi-modal representations (e.g., perceptual symbols) are only constructed via their association.

[8] 5 epochs of training took approximately 20 seconds on a single mid-range GPU. Additional epochs yielded only modest improvements, and were deemed unnecessary for the present demonstrations.

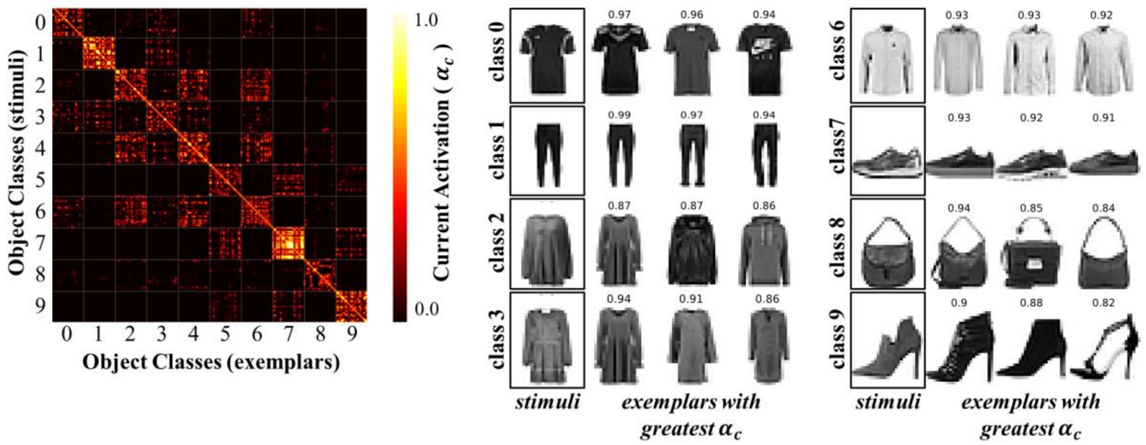

Figure 4: Current activations (α_c) for 250 randomly selected images (25 per object class). A heatmap (left) depicts the α_c for each pair of images. Example stimuli and their most activated exemplars are also shown.

forward activation of the β-VAE's recognition network. α_c was calculated for each pair of probability distributions, resulting in a 250×250 matrix, which we plotted as a heatmap (see Figure 4, Left) with rows/columns sorted by object class. We found that α_c was much higher (on average) for stimuli/exemplar pairs of the same class. The exemplars receiving the highest α_c were from the same class as the stimuli in over 70% of cases (see Figure 4, Right for examples). Inter-class activation generally occurred as the result of confusion between highly similar object classes (e.g., object classes 0, 2, 4, and 6, which represented "t-shirts," "pullovers," "coats," and "shirts," respectively). These observations are consistent with the hypothesis that our implementation satisfies the property of analogical representations.

Sufficiency of Generative Representations

We will claim that our implementation has learned generative representations that *suffice* if they allow the creation of modal simulations that are *recognizable by the implementation*. In other words, if we generate a simulation for a shirt, shoe, or bag, we want the system's perceptual processes to reactivate perceptual symbols for shirt-like, shoe-like, and bag-like objects, respectively. This criterion is very different from the usual focus in the machine learning community that typically rates generative quality in terms of human judgments. From our perspective, the simulations may look like random noise to a human, so long as they activate perceptual representations if and only if they "sufficiently resemble" those simulations.

To demonstrate that our implementation possesses this property, we generated modal simulations for the same randomly selected images, and used these to activate the β-VAEs recognition network *as if they were incoming sensory stimuli*, generating modal probability distributions, and calculated the pairwise α_c as before. The resulting α_c heatmap (not shown) looked very similar to the heatmap shown in Figure 4 (Left) with only slightly more inter-class noise. This strongly suggests that our recognition networks recognize the "gist" of the objects depicted in modal simulations.

Related Work

While there have been numerous attempts at implementing portions of a PSS, few have attempted to systematically build a PSS from the ground up based on first principles. Many implementations attempt to address topics of high theoretical interest, such as abstract concepts and language, without a firm implementation of PSS's basic components. While these are worthwhile pursuits, we believe they are premature. We briefly survey a few noteworthy attempts at more general PSS, and simulation-based, implementations, and contrast them with our approach.

Joyce, Richards, Cangelosi, and Coventry (2003) implemented a connectionist, computational model based on a recurrent neural network architecture that they call the Connectionist Perceptual Symbol System Network (CPSSN), and they applied it to labelled video sequences. The authors claimed that CPSSN is a mechanism for implementing perceptual symbols, and that it contains "categorical information summarising the event/episode." A major drawback of this, and most other *purely* connectionist approaches, is that the representations tend to be buried within the network's hidden units, limiting their ability to support compositionality and other cognitive processes.

Perlovsky and Illin (2012) argued that computational accounts of PSS require new mathematical frameworks that are "different from traditional artificial intelligence, pattern recognition, or connectionist methods," and they propose the use of Dynamic Logic (DL) for that purpose. They experimentally show that DL can implement object/situation representations and recognition, and may be capable of supporting multiple modalities; however, the connections between DL's operations and PSS are somewhat speculative, and would benefit from additional (property-based) analysis. It's also not clear whether DL will be able to model all of PSS's components, as the authors' hope.

Shanahan (2006) proposed a cognitive architecture that implements internal simulations, analogical representations via "topographically organized maps of neurons," and portions of GWT. While he didn't intend his architecture to be an implementation of a PSS, it's one of the few simulation-based cognitive architectures, and well worth mentioning. Shanahan's architecture combines a low-level, *reactive*, behavioral system with a higher-level, *predictive* system that "simulates" action outcomes. These simulations can elicit affective responses (i.e., "feelings") and guide action selection. Unlike the simulations in our implementation, Shanahan's do not seem to support conceptual representations or operations. Instead, they function similarly to models in model-based reinforcement learning that are used for "state-space planning" (Sutton & Barto, 2018).

Previous work on LIDA has explored symbol grounding using modular composite representation (MCR) vectors (Snaider & Franklin, 2014; Agrawal, Franklin, & Snaider, 2018). MCR vectors can be grounded and analogical, but they are not generative. Therefore, it's unclear how to use them as the representational basis for a PSS.

Discussion

Our goals were to establish a computational foundation for PSS, simulation-based perception, and "conscious" multi-modal perceptual learning. We believe that we have made progress towards these goals. Our initial experiments suggest that our SM representations, and, by extension, our perceptual symbols, satisfy both the *property of analogical representations* and *sufficiency of generative representations*, paving the way for future work in perception, categorization, and simulation-based cognition. Furthermore, our perceptual symbols exhibit many of the properties attributed to them by Barsalou (1999). They are **multi-modal**, **analogical** (since they satisfy the property of analogical representations), **not complete recordings**, **dynamic**, and **componential** (since they are based on disentangled latent representations).

Our approach differentiates itself from many previous attempts at implementing a PSS based on (1) its generality (i.e., it's not application-specific or focused on a single theoretical concern), (2) our explicit identification and intentional construction of each fundamental PSS component (as opposed to post-hoc attributions), (3) our systematic attempt at analyzing the properties of said components, and (4) our integration with a well-developed, agent architecture (LIDA). An additional strength of this approach is that it leverages theoretically sound, and experimentally proven, generative connectionist networks (i.e., VAEs), rather than trying to "reinvent the wheel."

Future work will explore resemblance-based generalization for concept learning, context-dependent and multi-part mental simulations, and the development of simulation-based, analogical reasoning processes. We will also complete our implementation of LIDA's cognitive cycle, incorporating an action phase with motor simulations.

Acknowledgments

We would like to thank Dr. Pulin Agrawal and Dr. Lynn Kugele for reviewing an earlier draft of this manuscript, and the CogSci 2020 reviewers for their excellent feedback.

References

Agrawal, P., Franklin, S., & Snaider, J. (2018). Sensory memory for grounded representations in a cognitive architecture. In *Proceedings of the Sixth Annual Conference on Advances in Cognitive Systems (ACS Poster Collection)* (pp. 1-18).

Baars, B. J. (1988). *A Cognitive Theory of Consciousness.* Cambridge University Press.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and brain sciences, 22*(4), 577-660.

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1521), 1281-1289.

Franklin, S., Madl, T., Strain, S., Faghihi, U., Dong, D., Kugele, S., . . . Chen, S. (2016). A LIDA cognitive model tutorial. *Biologically Inspired Cognitive Architectures, 16,* 105-130.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena, 42*, 335-346.

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. *ICLR, 2*(5), 6.

Joyce, D., Richards, L., Cangelosi, A., & Coventry, K. R. (2003). On the foundations of perceptual symbol systems: Specifying embodied representations via connectionism. In *The logic of cognitive systems: Proceedings of the Fifth International Conference on Cognitive Modeling* (pp. 147-152).

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR).*

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

Perlovsky, L., & Ilin, R. (2012). Mathematical model of embodied symbols: cognition and perceptual symbol system. *Journal of Behavioral and Brain Science, 2*(2), 195-220.

Shanahan, M. (2006). A cognitive architecture that combines internal simulation with a global workspace. *Consciousness and cognition, 15*(2), 433-449.

Snaider, J., & Franklin, S. (2014). Modular composite representation. *Cognitive Computation, 6*(3), 510-527.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT press.

Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.