Primitive Linguistic Compositionality in a Hebbian Neural Network

George Flint
georgeflint@berkeley.edu
Cognitive Science
University of California, Berkeley

Anna A. Ivanova[1]
a.ivanova@gatech.edu
School of Psychology
Georgia Institute of Technology

Abstract

Humans have a powerful ability to generate novel compositional representations. For example, imagining a *pink banana* requires compositional mappings between signifiers *pink* and *banana* and the perceptual referents of these signifiers. This essential cognitive faculty remains challenging to model in a biologically plausible way. Here, we present a model that implements signifier-referent compositional associations using Hebbian associative learning. The model satisfies the following constraints: (1) once associated, both signific and referential inputs can activate the shared representation, and (2) when signific and referential inputs are compositional, the model should generalize to novel compositional combinations. When trained on MNIST, the model successfully learns to associate number labels with corresponding images. On colored MNIST, the model learns signific-referential associations for both digits and colors, with somewhat successful generalization to new digit-color combinations. This work serves as a proof of concept for biologically plausible models of signifier-referent association.

Keywords: biological plausibility, deep learning, Hebbian learning, compositionality, language

Code: https://github.com/roccoflint/PLCHNN

Introduction

Language endows humans with an ability to evoke mental representations via signifiers—information, such as signs or words, used to refer to other information (de Saussure, 1959). To effectively use signifiers, the mind should have established a two-way connection: a referent (such as a perceived image) should evoke a representation linked to the appropriate signifier, and a signifier should enable easy access to the representation associated with the referent. A compositional version of these faculties might employ compositions of signifiers to classify or imagine compositional representations. For example, when presented with a composition of signifiers such as *pink banana*, the corresponding representational composition is activated, without the need to present a pink banana in real life. The presence and upshots of the power of signific-referential association, enhanced by the compositional power of signifiers, are well noted throughout the literature (Humboldt, 1836; Hockett, 1960, and many others). The challenge of compositional association is well known in the cognitive science world, with many academic works tackling the challenge (Piantadosi et al., 2016; Lake & Baroni, 2023; Russin et al., 2024; Zhou et al., 2023, and others). With the advent of vision-language models, such as DALL-E, some have claimed that the compositionality challenge has been solved (a famous example is DALL-E successfully generating images of "avocado chair"[2]). However, subsequent exploration has shown that the compositional abilities of vision-language models are still far from perfect (Leivada et al., 2023). The majority of computational models exploring vision-language compositionality use mainstream supervised deep learning architectures. Perhaps a simpler approach, in line with traditional philosophical proposals of the use of signifiers, would instead entail an associational approach, whereby a signifier and a referent become associated with one another during learning. However, the modeling approaches so far have been hindered by the lack of easy-to-use implementations of biologically plausible learning mechanisms, such as Hebbian learning.

In this work, we propose a theory on the mechanistic core of these faculties, rooted in a biologically plausible perspective—one of associative learning described by Hebbian dynamics (Hebb, 1949). We posit two minimal applications of the Hebbian mechanism: (1) that by associating signifiers with referents, either modality alone can activate a shared representation, and subsequently activate the omitted modality. (For example, signific input *banana* activates a shared representation, which then activates referential activations to some degree, producing an image.) We will refer to this mechanism as signific-referential association (SRA). And (2) that when signific and referential inputs are compositional, that a compositional structure to their associations will emerge. A robust SRA mechanism should support reliable classificatory and imaginative faculties, and a robust compositional SRA mechanism should support compositional generalization to the faculties.

We thus evaluate two hypotheses: (1) that a model using biologically plausible associative learning rules can achieve a robust signific-referential association, and (2) that the same model, given compositional structure to signific and referential data, can achieve robust signific-referential composition.

Approach

We develop a biologically plausible model that learns to link signifiers with their referents in an associative, bottom-up way. This model is unsupervised in that there is no error cor-

[1] Senior author.

[2] https://openai.com/index/dall-e/

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

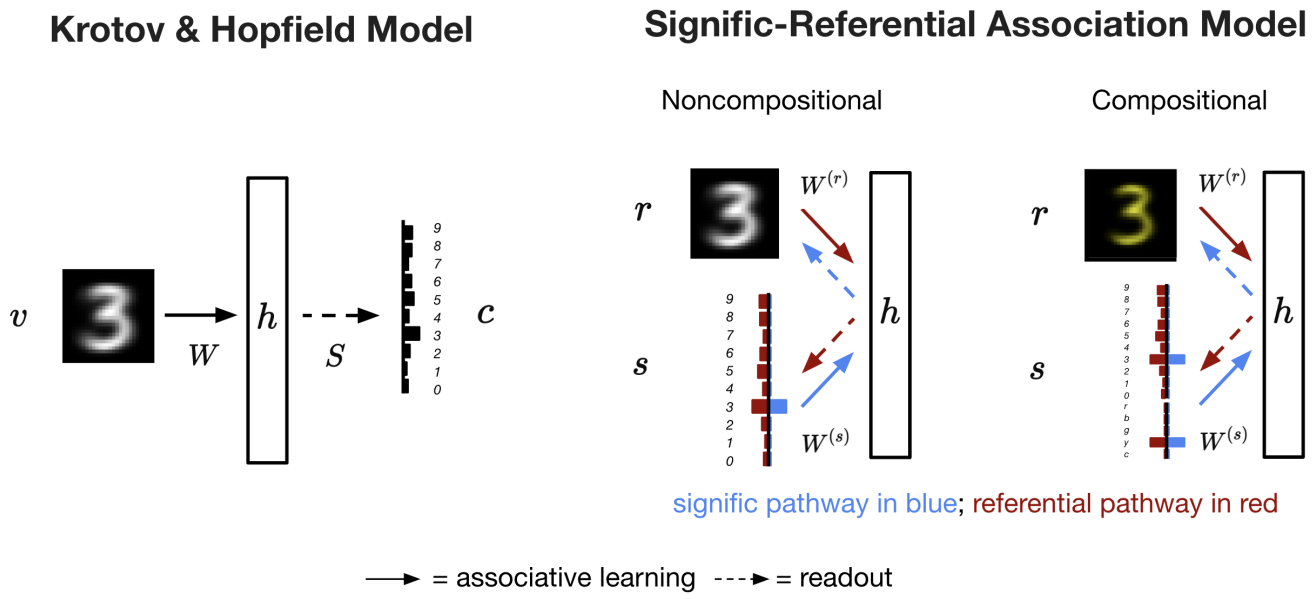

Figure 1: Diagrams of the Krotov-Hopfield (KH) model (2019) and our signific-referential association (SRA) model. In the KH model, referential image input v is passed to the hidden layer h, with weights W adjusted via Hebbian learning. Then, a supervised classifier S is trained to classify images into signifier classes c. The SRA model employs Hebbian learning with two distinct pathways: referential (operating on referential input r with Hebbian weights W^r) and signific (operating on signific input s with Hebbian weights W^s).

rection signal that it learns from (although the correct labels are provided as part of the bottom-up input). During test, the model might receive either signific or referential input; during training it receives both at the same time.

The base Hebbian model We build upon a model of Hebbian learning from Krotov & Hopfield, 2019 (henceforth KH). KH develop Hebbian learning rules with homeostatic plasticity dynamics for unsupervised feature learning. They offer a full dynamical model and a fast implementation, where the dynamical evolution of hidden unit activities are replaced with with a ranking heuristic for efficiency. Its behavior is governed as follows: For an input vector x_i, the activation h_j of a neuron j is calculated with

$$h_j = \sum_i \text{sign}(W_{ji})|W_{ji}|^{p-2}x_i. \tag{1}$$

These activations are used in the Hebbian learning rule:

$$\Delta\tilde{W}_{ji} = g(h_j)x_i - g(h_j)h_jW_{ji}, \tag{2}$$

with the gating function $g(h_j)$ defined as:

$$g(h_j) = \begin{cases} 1, & \text{if } h_j \text{ ranked highest} \\ -\Delta, & \text{if } h_j \text{ ranked } k\text{-th highest} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Δ is the anti-Hebbian strength. Learning rates decay linearly over epochs, and weights are initialized from a normal distribution: $W_{ji} \sim \mathcal{N}(0, \sigma^2)$. The learning rate decays linearly over epochs to 0. The hyperparameters p, k, and Δ are set–or learned–separate from the learning of the dynamical model. Weights are then multiplied by the learning rate, which decays linearly over epochs to 0, and normalized by dividing by the maximum update:

$$\Delta W_{ji} = \eta \frac{\Delta\tilde{W}_{ji}}{\max(\max_{j,i}(\Delta\tilde{W}_{ji}), \varepsilon)} \tag{4}$$

with $\varepsilon > 0$.

Differences in the signific pathway During experimental development, we found that the difference in dimensionalities between signific and referential inputs complicated learning. To address this, we adapt learning rules by treating hidden units as input and label activations as output, and use a Lebesgue norm parameter particular to the signific pathway.

Signific-referential association We develop upon the architecture of the KH model with the addition of signific neurons to the input layer, fully connected to the hidden layer—the "signific pathway" (Figure 1). We develop upon the learning rules of the model with the addition of a coefficient to the hyperparameter p in the signific pathway, on the conjecture that small-dimensional categorical data requires different sparsity constraints than high-dimensional visual features. We conduct training by providing signific and referential input simultaneously.

Readout strategies In the KH model, readout is enabled by a supervised (backpropagation-based) layer appended on top of the Hebbian model. In our model, signifiers are instead provided as inputs; we make use of our two distinct input pathways for probing representations learned by the model.

We develop two novel readout methods particular to our model: (1) **"referential-to-signific" (RtS)**, which tests propagation of referential input through the hidden layer into signific activations as a classification task, and (2) **"signific-to-referential" (StR)** which tests propagation of signific input through the hidden layer into referential activations as an imagination task. We compare our RtS accuracy to that of a supervised decoder connected to the hidden layer.

Experiment 1: Signific-Referential Association

Here, we test the hypothesis that an associative Hebbian model trained on pairs of signifiers and referents (e.g., images accompanied by a corresponding label) will learn to successfully associate the two, such that (a) when presented with a novel referent, the model's internal representation will enable easy readout of the corresponding signifier (RtS readout), and (b) when presented with a signifier, the model's internal representation will enable easy reconstruction of the prototypical referent for that signifier (StR readout). We train our model on the MNIST dataset (LeCun et al., 1998), compute RtS classification performance against a supervised decoder, and compute StR reconstruction performance using averaged images from each digit class as a ground truth.

Method

Data For referential input, we use the MNIST dataset loaded in the standard uint8 form, where each pixel is represented as an integer between 0 and 255, with 28 by 28 pixel images flattened into 784-dimensional vectors. The vectors are normalized to the [0, 1] range. For signific input, we use one-hot vectors to encode the class of a given MNIST image. As MNIST has 10 classes, signific vectors are 10-dimensional. MNIST image vectors are paired with the corresponding one-hot encoding vectors for their class during dataloading.

Model We use two sets of weights connected to the hidden layer h: referential weights $W^{(r)}$ and signific weights $W^{(s)}$. Given that our model uses two input pathways, we use a boolean hyperparameter "pathway interaction" which controls whether to add signific activations to referential activations before applying the competitive activation function g. Weights are initialized from the standard normal distribution, as in KH. Hidden layer activations, weight updates, and learning dynamics are computed in accordance with the fast implementation of KH, with the exception of a coefficient to p in signific weights $W^{(s)}$. We introduce this coefficient (as a controllable hyperparameter) on the conjecture that that small-dimensional categorical data requires different sparsity constraints than high-dimensional visual features.

Training At each training iteration, the model receives (1) the MNIST image vector ("referential input" or r) and (2) the one-hot encoding vector for the MNIST image class ("signific input" or s). If pathway interaction is enabled, weights are updated according to the activation function g ranking both referential and signific activations r and s. As in the fast implementation of KH, training proceeds in minibatches for a preset number of epochs with learning rate decaying linearly over epochs.

Readout After training, we conduct RtS and StR readout, in which input from one pathway propagates to the hidden layer, and then through the other pathway to reconstructed "input" activations. Propagation uses the p based nonlinearity for both pathways, though with the coefficient to p in the signific pathway. During readout, weights remain fixed.

During RtS readout, the model receives only referential activations r, first propagating through referential weights $W^{(r)}$ to the hidden layer h. Activations in the hidden layer subsequently propagate through signific weights $W^{(s)}$ to signific activations s, producing a classificatory probability distribution over the 10 classes. The predicted class is determined via an argmax over these final signific activations. Classificatory efficacy is measured by accuracy of prediction.

RtS readout is compared to a linear classifier ("decoder") where the hidden layer activations h for each referential input r are classified using a Moore-Penrose pseudoinverse approach. The weight matrix W is computed as $W = X^{+}Y$, where X represents the matrix of hidden activations and Y the one-hot encoded labels. The predicted class is determined via is performed via an argmax operation over XW.

During StR readout, the model receives only signific activations s, first propagating through signific weights $W^{(s)}$ to the hidden layer h. Activations in the hidden layer subsequently propagate through referential weights $W^{(s)}$ to referential activations r, producing a reconstructed image. Imaginative efficacy is measured by structural similarity index (SSIM) measure to an image which is the average of all dataset images for a given class.

Optuna optimization We employed Optuna to optimize our model's hyperparameters across 20 trials using Tree-structured Parzen Estimators (TPE) with single objective loss function composed of RtS and StR readout metrics. Specifically, loss is defined with

$$L = -\ln(1 - \|C - I\|_F + \varepsilon) - \ln(\|S - I\|_F + \varepsilon) \qquad (5)$$

where C is the true class-normalized confusion matrix from RtS classification, S is the similarity matrix for StR readout, I is the identity matrix, N is the number of classes, and $\varepsilon = 10^{-10}$ is a small constant to avoid $\ln(0)$. The Frobenius norm (FN), defined as $\|X\|_F = \sqrt{\sum_{i,j} X_{ij}^2}$, is used to measure matrix differences. We use FN difference from the identity matrix rather than raw accuracy or similarity to discourage misclassifications or reconstructional local minima where all

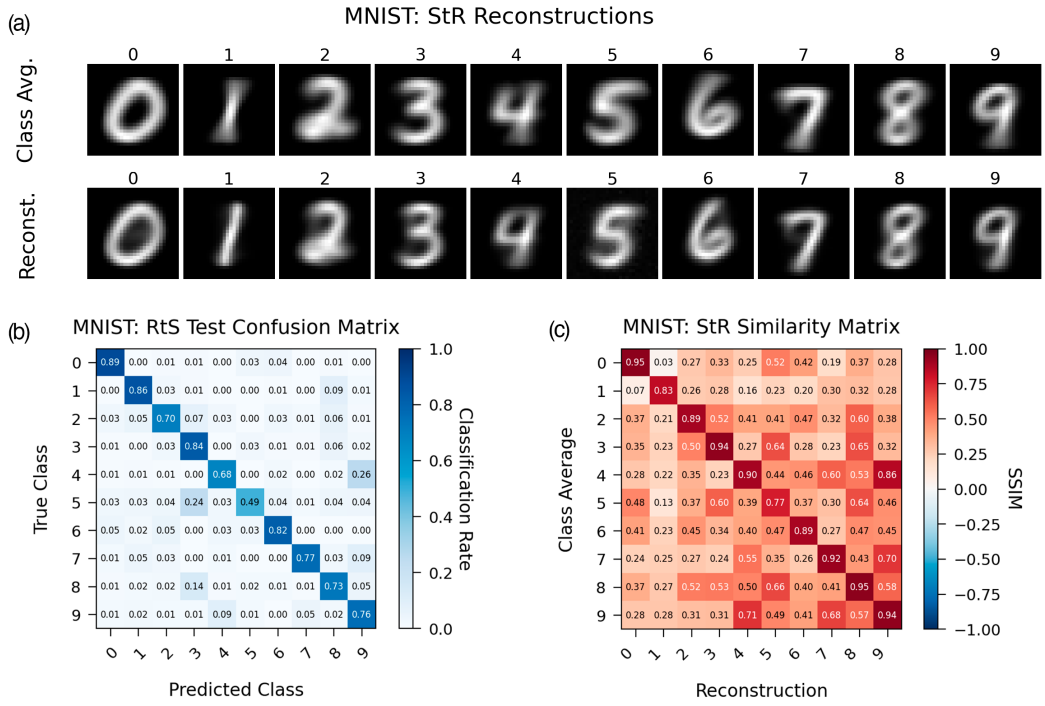

Figure 2: Modeling results for Experiment 1 (MNIST). (a) Signific-to-referential (StR) reconstructions of referential input provided signific input only; class averages images for each MNIST digit. (b) True class-normalized referential-to-signific (RtS) confusion matrix for MNIST classification. (c) Structural Similarity Index Measure (SSIM) values between StR reconstructions and MNIST class average images.

reconstructions are similar to all class averages.

The optimization search space included the hidden layer size $|h|$ (100 to 324 neurons), exponent p (2.0 to 4.0), ranking parameter k (1 to 7), anti-Hebbian strength Δ (0.2 to 0.6), signific p coefficient (1.0 to 4.0), the number of epochs (10 to 60), minibatch size (64 to 128), initial learning rate (0.01 to 0.03 on a log scale), and the boolean parameter for pathway interaction. Optimization begins with initial parameters $|h| = 324$, signific p multiplier $= 1.0$, and pathway interaction enabled, along with $p = 3$, $k = 7$, $\Delta = 0.4$, number of epochs $= 20$, minibatch size $= 100$, and learning rate $= 0.04$, as presented in KH.

Results

Optuna optimization Optuna optimization produced hyperparameters $|h| = 279$, $p = 2.7149$, $k = 2$, $\Delta = 0.2847$, signific p multiplier $= 3.7437$, pathway interaction disabled, number of epochs $= 44$, minibatch size $= 115$, and learning rate $= 0.0240$. Hyperparameter importances to the signific-referential association experiment are listed in Table 1.

Readout RtS readout shows a classificatory accuracy of 75.29% on training and 75.87% on testing. Decoder readout shows an accuracy of 84.86% on training and 85.59% on testing. Per-class confusion data for RtS classification is available in Figure 2. StR readout, measured by the SSIM between a given reconstruction and the class average image, shows an on-diagonal mean of 0.898 and an off-diagonal mean of 0.3868. Per-class similarity data for StR reconstruction is available in Figure 2.

Hyperparameter	Importance
Δ	0.4246
Learning rate	0.2463
Pathway interaction	0.1246
p	0.0801
Number of epochs	0.0600
k	0.0319
Batch size	0.0146
$\|h\|$	0.0099
Signific p coefficient	0.0079

Table 1: Hyperparameter importance values for the signific-referential association experiment estimated using Functional ANOVA in Optuna optimization.

Experiment 2: Signific-Referential Composition

Here, we test the hypothesis that a model, using associative learning rules and having undergone sufficient training on compositional signific and referential input distributions, will achieve classificatory and imaginative faculties which generalize to novel compositions. Specifically, we train our model

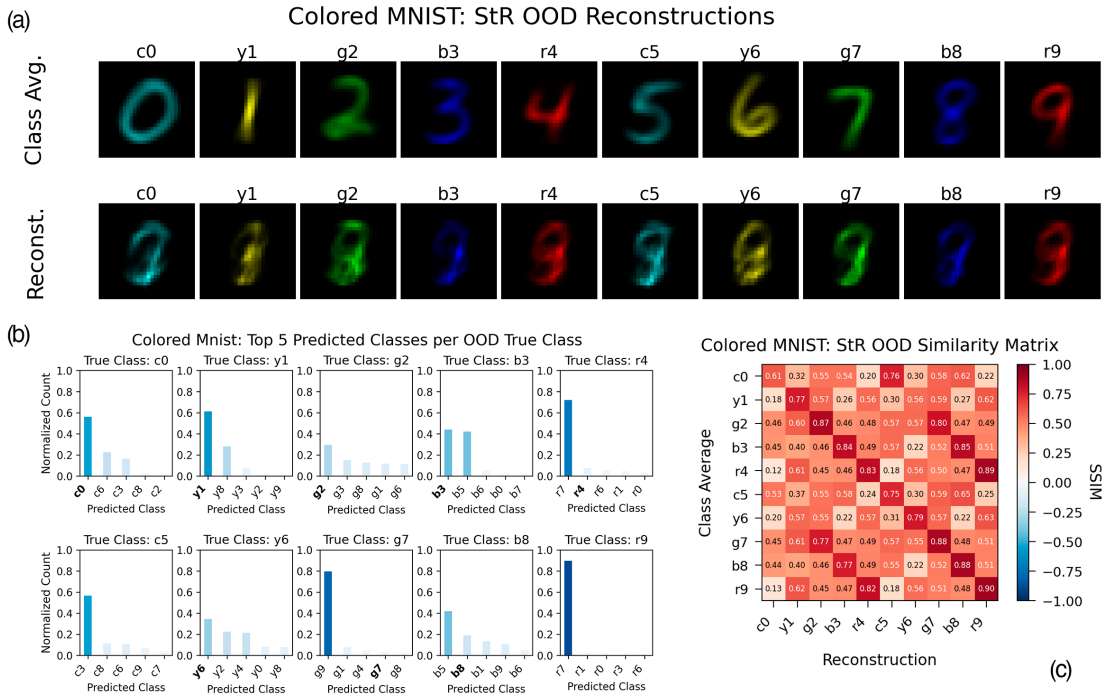

Figure 3: Modeling results for Experiment 2 (colored MNIST). (a) Signific-to-referential (StR) reconstructions of OOD referential input classes provided signific input only; class averages for each color-digit class. In-distribution (ID) composition reconstructions and class averages are not shown. (b) Top 5 referential-to-signific (RtS) predicted classes for each OOD true class. (c) SSIM values between StR reconstructions of OOD digit-color compositions and digit-color class averages.

on a generated variant of the MNIST dataset which applies colors to image instances ("Colored MNIST"), compute classificatin accuracy with RtS readout, and compute reconstruction validity with StR readout. We conduct readout on both in-distribution (ID) and out-of-distribution (OOD) compositional input.

Method

We employ the exact same method used in Experiment 1 with only those modified components identified below:

Data Referential input consists of the MNIST dataset just as in Experiment 1, but with n color channels for each pixel in addition to the intensity channel. A given image is assigned a color's intensity channel, where all values in those channels mirror the grayscale intensity channel. Images thus have no inherent RGB structure to them–an RGB visualization function is applied for reconstructions. Sample images from each digit class are separated into n partitions, with all samples of each partition being assigned a color value. Signific input consists of two-hot vectors, with 10 components encoding digit class and n components encoding color class. A selection of digit-color compositions (20%) are withheld from input distributions. We use $n = 5$ in this experiment.

Readout RtS readout occurs just as in Experiment 1, but uses two separate argmax operations over signific activations s: one over digit-encoding components, and one over color-encoding components. StR readout occurs just as in Experiment 1, but uses two-hot signific input s.

Optuna optimization The number of trials and hyperparameter search space mirror those of Experiment 1. However, here we optimize using Non-dominated Sorting Genetic Algorithm II with a multi-objective loss. The RtS loss is the log-based FN difference between the true class-normalized confusion matrix and the identity matrix, while the StR loss is the log of the diagonal mean of the reconstruction-class average similarity matrix. We do not apply FN difference from the identity for StR, as it resulted in challenging gradients, likely due to the prevalence of color-based similarities. Each metric has two losses: one over the full matrix and one for OOD entries, for four total objectives.

Results

Optuna optimization Optuna optimization produced hyperparameters $|h| = 107$, $p = 3.2728$, $k = 3$, $\Delta = 0.4034$, signific p multiplier $= 3.7227$, pathway interaction disabled, number of epochs $= 48$, minibatch size $= 78$, and learning rate $= 0.0112$. Hyperparameter importances to the signific-referential association experiment are listed in Table 2.

Readout RtS readout shows a classificatory accuracy of 70.06% on ID compositions and 26.50% on OOD composi-

Hyperparameter	Importance
Δ	0.4150
k	0.1709
Batch size	0.1212
Learning rate	0.0761
p	0.0679
Signific p multiplier	0.0566
Number of epochs	0.0401
$\|h\|$	0.0317
Pathway interaction	0.0204

Table 2: Hyperparameter importance values for signific-referential composition experiment estimated using Functional ANOVA in Optuna optimization.

Metric	RtS	Decoder
Overall Accuracy	**61.35%**	72.94%
ID Accuracy	**70.06%**	84.09%
OOD Accuracy	**26.50%**	28.35%

Table 3: Classificatory accuracy for RtS and Decoder readouts on ID and OOD compositions. RtS performance lags 16.69% and 6.53% behind the decoder in ID and OOD composition classification respectively. Color accuracies, for both readout methods on both ID and OOD samples, were 100%, leaving digit accuracy as the bottleneck.

tions. Decoder readout shows an accuracy of 84.09% on ID composition classifications and 0.2835 on OOD composition classifications. Digit- and color-specific accuracies are also available in Table 3. Per-class confusion data for RtS classification on OOD compositions is available in 3. StR readout shows a mean SSIM to class averages of 0.9303 on ID composition reconstructions, 0.8786 on OOD composition reconstructions, and an off diagonal mean of 0.4335.

Discussion

In this work, we extend the unsupervised Hebbian learning framework of Krotov & Hopfield (2019) by introducing a signific pathway, enabling bidirectional associations between signifiers (labels) and their referents (images) in a fully unsupervised and biologically plausible manner. We introduce a novel training paradigm that associates signific and referential pathways, along with bidirectional readout methods to evaluate information flow between them. Having conducted associative training with our model and collected readout data, we evaluate our hypotheses on abilities we should expect to see with the model: (1) that our model should achieve a faculty for "classification" (activation of appropriate signific activations given referential input) and "imagination" (activation of appropriate referential activations given signific input), and (2) that, when trained on compositional data, these faculties should generalize to novel compositions not seen during training. Having undergone signific-referential association training on pairs of MNIST images and class labels, our model achieves significant classificatory accuracy (75-76%) on MNIST images, lagging just behind that of a decoder (84-85%). Given signific input, our model also achieves high reconstructional similarities to corresponding class averages (0.898) compared to non-corresponding class averages (0.3868). Qualitatively, visual similarities between class reconstructions and class averages is in line with observed behavior of Oja's rule–a Hebbian-based rule with homeostatic constraints which the KH-style rule is based on–which aligns learned representations with principal components of the input distribution (Oja, 1982). These results are in accordance with our first hypothesis. Having undergone the same training on pairs of MNIST images in color and class labels encoding both digit and color class, the model achieves significant classificatory accuracy for in-distribution (70%) compositions. Out-of-distribution accuracy (26.50%), while lower than in-distribution accuracy, significantly exceeds chance levels, and legs behind decoder performance by only 1.85%. Given signific input, our model also achieves high reconstructional similarities to corresponding class averages for in-distribution compositions (0.9303) and out-of-distribution compositions (0.8786) compared to non-corresponding class averages (0.4335). These results are in accordance with our second hypothesis.

Limitations and future directions We limit our exploration to a primitive sort of compositionality–we do not posit any theoretical perspectives on nor investigate with our model any mechanistic underpinnings of order-dependent, hierarchical, or syntactic compositional faculties. Furthermore, the compositional structure of our data is quite simple. And, at that, the model's proficiency on these tasks, especially Experiment 2, was modest. Open questions remain as to how to scale the capabilities of these learning rules and dynamics. We suggest our model as a potentially fruitful perspective through which to explore compositional faculties from a biologically plausible perspective.

Conclusion

We have proposed a theory on the mechanistic core of classificatory and imaginative faculties enabled by the use of signifiers, rooted in the biologically plausible perspective of associative learning described by Hebbian dynamics. We posit two minimal applications of the Hebbian mechanism: the association of signifiers and referents to shared representations (signific-referential association) and a compositional structure to those associations, supporting compositional generalization (signific-referential compositionality). In our model, we observe the emergence of these faculties, using these mechanisms, under Hebbian-like associative learning rules. Using readout techniques developed for our model, we find compelling results for a proof of concept of the theory. We suggest the theory and model as grounds for future exploration of the mechanistic core of primitive cognitive linguistic faculties.

Acknowledgments

Anna Ivanova acknowledges funding support from the School of Psychology, Georgia Institute of Technology

References

de Saussure, F. (1959). *Course in general linguistics*. New York: Philosophical Library. (Translated by Wade Baskin)

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. John Wiley & Sons.

Hockett, C. F. (1960). The origin of speech. *Scientific American*, *203*, 88–96.

Humboldt, W. v. (1836). *Über die verschiedenheit des menschlichen sprachbaues und ihren einfluss auf die geistige entwicklung des menschengeschlechts*. Druckerei der Königlichen Akademie der Wissenschaften.

Krotov, D., & Hopfield, J. J. (2019). Unsupervised learning by competing hidden units. *Proceedings of the National Academy of Sciences*, *116*(16), 7723–7731. doi: 10.1073/pnas.1820458116

Lake, B. M., & Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. *Nature*, *623*(7985), 115–121. Retrieved from https://www.nature.com/articles/s41586-023-06668-3 doi: 10.1038/s41586-023-06668-3

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Leivada, E., Murphy, E., & Marcus, G. (2023). Dall· e 2 fails to reliably capture common syntactic processes. *Social Sciences & Humanities Open*, *8*(1), 100648.

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, *15*(3), 267–273.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, *123*(4), 392–424. doi: 10.1037/a0039980

Russin, J., McGrath, S. W., Williams, D. J., & Elber-Dorozko, L. (2024). From frege to chatgpt: Compositionality in language, cognition, and deep neural networks. *arXiv*, *2405.15164*. Retrieved from https://arxiv.org/abs/2405.15164

Zhou, Y., Feinman, R., & Lake, B. M. (2023). Compositional diversity in visual concept learning. *Cognition*, *244*, 105711. Retrieved from https://arxiv.org/abs/2305.19374 doi: 10.1016/j.cognition.2023.105711