Toward hierarchical compositionality with shallow hierarchical networks

Francisco M. López (lopez@fias.uni-frankfurt.de)
Frankfurt Institute for Advanced Studies
and
Xidian-FIAS International Joint Research Center
Ruth-Moufang-Straße 1, 60438 Frankfurt am Main

Judith Massmann (massmann@fias.uni-frankfurt.de)
Frankfurt Institute for Advanced Studies
Ruth-Moufang-Straße 1, 60438 Frankfurt am Main

Jochen Triesch (triesch@fias.uni-frankfurt.de)
Frankfurt Institute for Advanced Studies
Ruth-Moufang-Straße 1, 60438 Frankfurt am Main

Abstract

Unlike conventional deep neural networks, the human brain has a myriad of direct connections from subcortical nuclei to all cortical areas. These enable fast information transfer and facilitate hierarchical compositionality but have not yet been explored in artificial systems. In this work, we present the Shallow Hierarchical Artificial Neural Network (SHANN), a novel brain-inspired architecture with shallow connections from the input to all the hierarchical processing layers. We show that SHANNs can outperform shallow and deep networks in reconstruction and classification tasks. Initial explorations reveal that SHANNs use hierarchical compositionality to combine information from different levels of abstraction.

Keywords: Deep neural networks; Hierarchical processing; Compositionality; Shallow hierarchical networks

Introduction

Deep neural networks (DNNs) with tens or hundreds of hidden layers achieve super-human performance in classification tasks (LeCun, Bengio, & Hinton, 2015). However, the human brain remains the most energy-efficient neural system (Yu & Yu, 2017) and attains semantic knowledge in fewer abstraction levels (Lindsay, 2021). As such, the brain should serve as inspiration to machine learning researchers.

In this work, we explore a novel brain-inspired architecture for artificial neural networks with shallow and hierarchical connections. Recent work has shown that hierarchical cortical processing (Felleman & Van Essen, 1991) could be overly simplistic in the light of the abundant forward and feedback connections from subcortical nuclei to all areas of the cortex (Tervo et al., 2016). This leads to a shallow interpretation of the brain (Suzuki, Pennartz, & Aru, 2023), where the cortical hierarchies are orthogonal to cortical-subcortical connections. As way of example, the traditional interpretation of the ventral visual processing in sequential stages from the retina to the lateral geniculate nucleus (LGN) of the thalamus, visual cortex areas V1, V2, V4, and finally inferior temporal lobe (IT), must be challenged due to direct connections from LGN to deeper visual cortex and IT (Siegle et al., 2021).

Disentangling depth and hierarchy

The main argument behind our proposal is that the high performance achieved by DNNs originates in the hierarchical abstractions that emerge in the deepest layers, regardless of their actual depth. To facilitate the distinction, we define two properties for a layer in a network. First, we define the *depth* of a layer as the minimum number of forward passes from the input through other layers required to reach it. Second, we define the *hierarchy* of a layer as the maximum possible number of forward passes from the input through other layers to reach it. In conventional DNNs, depth and hierarchy are equivalent. Alternative architectures that use skip connections to solve the vanishing gradient problem reduce depth while maintaining hierarchy, although the disentanglement is lost when adding, e.g. ResNets (He, Zhang, Ren, & Sun, 2015), or aggregating, e.g. DenseNets (Huang, Liu, Van Der Maaten, & Weinberger, 2016), copies of the values emerging from different depths.

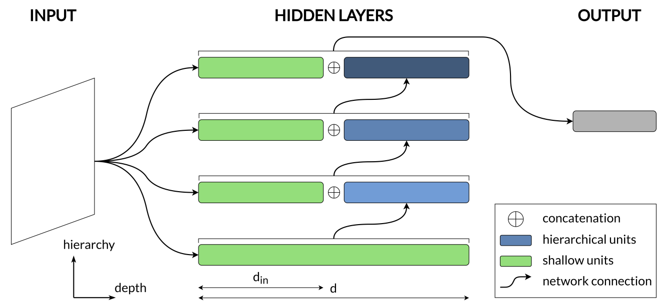

Figure 1: Architecture of a SHANN with 4 hidden layers. Each layer is the concatenation of shallow connections from the input and hierarchical connections from the previous layer.

Shallow Hierarchical Networks

In line with the interpretation of a shallow brain, we introduce a new architecture which we call the Shallow Hierarchical Artificial Neural Network (SHANN). Its main innovation is the inclusion of direct connections from the input to all hidden layers. Therefore, SHANNs have a shallow structure with multiple hierarchical levels (see Fig. 1). For a hidden layer j of size d, the first d_{in} values are obtained directly from the input and the remaining $d - d_{in}$ are obtained from connections of the lower hidden layer. Conventional DNN connections from layer $j-1$ to layer j are described by:

$$x_j = f^{(d)}(x_{j-1}|\theta_{j-1}; b_{j-1}; a), \tag{1}$$

where x_j is the activity of layer j and f is the result of an affine transformation with weight matrix θ and biases b followed by a non-linear activation function a. On the other hand, a SHANN's connections are given by:

$$x_j = f^{(d_{in})}(x_{in}|\theta_{in,j}; b_{in,j}; a) \oplus f^{(d-d_{in})}(x_{j-1}|\theta_{j-1}; b_{j-1}; a)], \tag{2}$$

where $\oplus$ is the concatenation operator. The first term describes the shallow connections directly from the input x_{in}, and the second term describes the hierarchical connections from the immediate lower level of the network. The output of a SHANN is derived directly from the highest level in the hierarchy following Eq. 1.

In Fig. 1 we show an example of a SHANN consisting of 4 hidden layers with increasing hierarchy but constant depth of 1. Like conventional DNNs, SHANNs can achieve high levels of abstraction in more hierarchical layers. However, the additional direct connections from the input to all hidden layers can facilitate faster training and transfer of information. The concatenation of connections derived from different levels provides the highest layer with a hierarchical compositionality that can be exploited by the output.

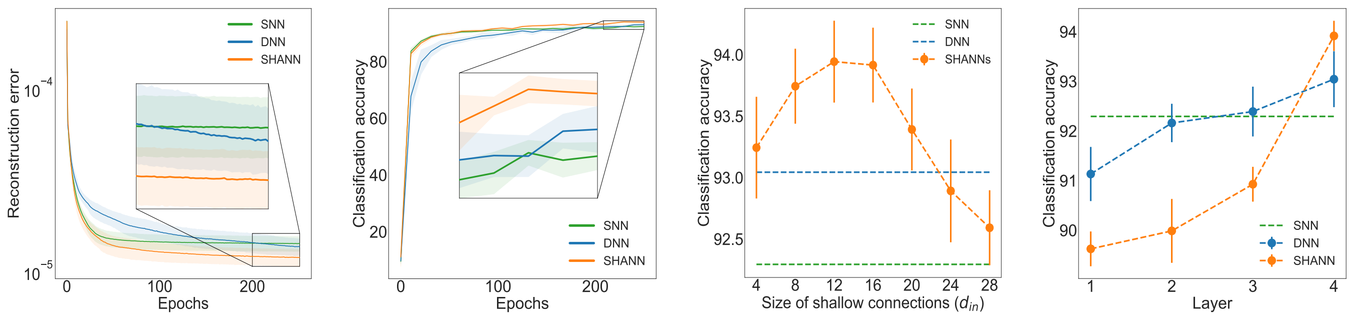

Figure 2: Comparison of SHANNs with a SNN and a DNN. *Left*: Reconstruction errors during training. *Center-Left*: Classification accuracies during training. *Center-Right*: Classification accuracies for different sizes of shallow connections. *Right*: Classification accuracies from representations learned at different layers.

Experiments

Comparison with Deep and Shallow Networks

We hypothesize that SHANNs can simultaneously achieve the fast learning of Shallow Neural Networks (SNNs) as well as the high performance of DNNs. To test this hypothesis, we evaluate the performance of SHANNs with 4 hidden layers, as depicted in Fig. 1, in comparison to a SNN and a DNN. The DNN also has 4 hidden layers but only with the hierarchical connections of Eq. 1. The SNN has a single hidden layer. We use the MNIST dataset, with the images normalized and flattened into vectors of length 784. The hidden layers of all models have length $d = 32$.

First, we train the three models to encode the MNIST digits using unsupervised learning with an additional single-layer decoder and the mean squared error of the reconstructions as a loss function. The network weights are subsequently frozen and used to train linear classifiers to predict the labels of the images. Results are shown in Fig 2. We see that a SHANN with $d_{in} = 16$ outperforms both SNNs and DNNs. We find the best classification results using SHANNs with $d_{in} = 12$ and $d_{in} = 16$, i.e. around half the size of the hidden layers. All SHANNs with up to $d_{in} = 20$ outperform the DNN. The training curves also reveal that while both the DNN and the SHANN have more trainable parameters than the SNN, only the latter learns as fast. Therefore, we conclude that access to shallow connections in a hierarchical architecture accelerates learning and leads to better representations.

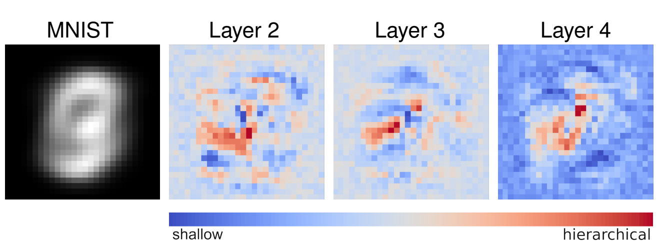

Figure 3: Average of all MNIST test images and feature importances in shallow and hierarchical connections.

Hierarchical compositionality

Second, we evaluate the hierarchical compositionality of the trained SHANN by analyzing the classification accuracies from the representations learned at different layers of the network. The resulting plot is shown in Fig. 2. While the SHANN outperforms the DNN and SNN using the representations in layer 4, the accuracy for lower levels of the hierarchy is below that of a SNN as well as the lower levels of a DNN. Therefore, we find that SHANNs perform hierarchical compositionality by distributing the representations of the inputs across the shallow and hierarchical connections in the last layer.

To better understand this phenomenon, we evaluate the feature importances for the shallow and hierarchical connections. We analyze how the image pixels contribute to the activation of the different neurons of a layer. Here, we compute the average feature importances using all images in the test dataset for layers 2, 3, and 4. For each, we compute the averages for the shallow and hierarchical connections, and plot the differences between the two in Fig. 3. Blue areas, indicative of more prominent shallow connections, are visible on the borders of the images, whereas the center of the images is more relevant for the hierarchical connections. This is consistent with hierarchical compositionality, whereby the shallow connections encode regions scarsely informative, whereas the hierarchical connections encode the most informative region, i.e. the center of the digit.

Conclusions

SHANNs achieve shallow-like fast learning while outperforming both SNNs and DNNs. In future work, we will explore how SHANNs can be extended to other classes of neural networks, in particular with convolutional layers, and perform additional evaluations with more complex tasks where shallow connections alone are insufficient. We believe the shallow hierarchical architecture can result in valuable improvements to artificial neural networks and help bridge that gap between neuroscience and machine learning. Nevertheless, there are many other types of brain computations missing in artificial neural networks that could improve performance even further, such as cortical-subcortical loops and local learning rules.

Acknowledgments

This research was supported by "The Adaptive Mind" funded by the Hessian Ministry of Higher Education, Research, Science and the Arts, Germany, and by the German Research Foundation (DFG) Project numbers 520617944, 520223571 ("Sensing LOOPS"). JT was supported by the Johanna Quandt foundation.

References

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, *1*(1), 1–47.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2016). Densely connected convolutional networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4700–4708).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, *33*(10), 2017–2031.

Siegle, J. H., Jia, X., Durand, S., Gale, S., Bennett, C., Graddis, N., ... others (2021). Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, *592*(7852), 86–92.

Suzuki, M., Pennartz, C. M., & Aru, J. (2023). How deep is the brain? the shallow brain hypothesis. *Nature Reviews Neuroscience*, *24*(12), 778–791.

Tervo, D. G. R., Hwang, B.-Y., Viswanathan, S., Gaj, T., Lavzin, M., Ritola, K. D., ... others (2016). A designer aav variant permits efficient retrograde access to projection neurons. *Neuron*, *92*(2), 372–382.

Yu, L., & Yu, Y. (2017). Energy-efficient neural information processing in individual neurons and neuronal networks. *Journal of Neuroscience Research*, *95*(11), 2253–2266.