Toward Structural Similarities between the Brain and Neural Networks

MinSung Cho (nninept@kau.kr)
Department of Artificial Intelligence, Korea Aerospace University
76, Hanggongdaehak-ro, Deogyang-gu, Goyang-si, Republic of Korea

Jay Hoon Jung (jhjung@kau.ac.kr)
Department of Artificial Intelligence, Korea Aerospace University
76, Hanggongdaehak-ro, Deogyang-gu, Goyang-si, Republic of Korea

Abstract

The emergence of artificial neural networks has brought significant advancements to the modeling of intelligence. However, there are still notable differences between their functioning and the structural operations of the brain. From a macroscopic perspective, the brain processes information received from various sensory organs and uses this input to make decisions associated with stored memories. In contrast, artificial neural networks perform cognition and decision-making simultaneously, a structure that reduces the model's explainability and complicates the implementation of structural divisions. This research presents a framework that *distinguishes* between cognition and decision structures within artificial neural networks. Experimental analysis using a simplified model is conducted to showcase this distinction.

Keywords: neural network; learning; cognitive system; memory

Introduction

With the advancement of artificial neural networks, research in biological-inspired neural networks has continued to progress. For example, Spiking Neural Networks (SNNs) (Gordleeva et al., 2021), which aim to replicate the functionality of biological neurons, are regarded as the next generation of neural networks, despite their limited training approaches that do not rely on gradients. There are also attempts to reduce not only low-level differences, such as the functionality of neurons, but also high-level differences in structural segmentation, such as cognition, decision-making, and memory, between current neural networks and the brain. Recently, significant achievements have been made in research aimed at implementing memory in networks, using episodic memory in reinforcement learning (Lin, Zhao, Yang, & Lintao, 2018) or working memory in Convolutional Neural Networks (CNNs) (Yang et al., 2023).

Current neural networks assign the role of memory to weight parameters simultaneously, akin to equating synapses with memory themselves. In reality, the brain's actual memory is formed by the organization of multiple synapses, and the brain's cognitive components are distinguishable from its memory components. Therefore, this paper proposes a framework for structurally partitioning neural networks into cognitive modules and long-term memory (LTM), and conducts experiments with a simple model to demonstrate how the explainability of the model for different classes can be enhanced.

Approach

Pattern Recognition

We chose CNN as the approach for pattern recognition, as they allow us to capture spatial information in images. However, instead of having recognition output from a single CNN model, we designed it to pass through multiple parallel convolutional layers to combine information. This was done in consideration of the memory structure to be discussed later, aiming to evenly learn class-specific features at each layer.

For an image I and a kernel K, the convolution operation can be expressed as follows:

$$(I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n)$$

If both the norms of I and K are 1, the upper bound of $(I * K)(i,j)$ is as follows:

$$(I * K)(i,j) \leq 1$$

The maximum value of the convolution occurs when the kernel and the patch of the image share identical values, indicating the presence of the same pattern. If the given image is normalized, determining the scale of the maximum value depends on the kernel used. In CNN, the convolution operation is carried out as many times as the number of kernels, same as the number of output channels c. Thus, for each feature map h in the output, by extracting the maximum value, a feature vector z is obtained. This vector represents the strength of specific patterns of kernels found in the image.

$$z_c = relu(max(h_c))$$

Memory Module

The brain recognizes individual objects by combining specific patterns within images. The role of the memory module, mimicking this process, is to map combinations of vector elements from the feature vector z obtained in the convolutional layer into each class. The definition of the memory module M follows $M \in \{x | x \geq 0\}^{n_{cls} \times c}$, where n_{cls} is number of classes.

$$M'_{kj} = \begin{cases} 0 & \tanh(M_{kj}) \leq 0.5 \\ 1 & \tanh(M_{kj}) \geq 0.5 \end{cases}$$

Here, l_i represents the i-th layer. Using the binarized M', we conduct matrix multiplication with z.

$$l(x) = z \cdot (M')^T$$

The classification model $L(X) : X \rightarrow R^{n_{cls}}$ can be represented as $L(x) = \sum_i^N l_i(x)$

Memory Training

The optimization of memory follows a biological intuition. The important features stored in the memory M^i of each i-th layer are reinforced through λ when they have a significant influence on the output values, while features with less impact are gradually eliminated from memory through τ. Memory receives positive feedback only when successful predictions are made for each data point; otherwise, it receives negative feedback.

$$M^i_{kj} = \begin{cases} M^i_{kj} + \lambda & \text{if } j = argmax(z^i)) \\ M^i_{kj} - \tau & else \end{cases}$$

Experiments

1. Model Design

We designed a simple model to test the framework. We utilized the MNIST dataset and employed a total of 8 layers. To reduce the non-linearity of each convolutional layer, we fixed the depth of all convolutional layers to 1. Given the shallow nature of the convolution, it becomes challenging to capture global features at each layer. Therefore, we applied resizing to the data depending on the layer. Since we maintained a fixed channel count of 8 for all convolutional layers, the size of the memory M for each layer remained consistent at (10,8). For the memory, we set λ to 0.3 and τ to 0.008, and we ensured that the memory was updated after each batch calculation.

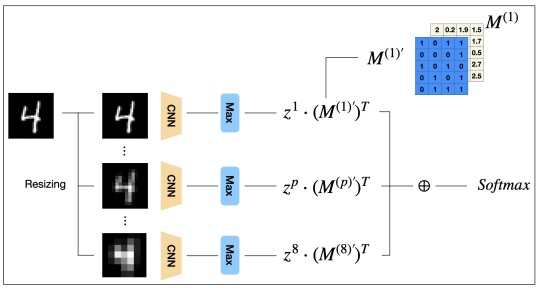

Figure 1: Structure of the model for MNIST classification

2. Model Analysis

To assess whether the trained convolutional layers adequately learned features, we utilized heatmaps. In our model, when the elements of z are larger, it indicates that the corresponding convolution kernel's pattern strongly appears in the image. Therefore, we positioned the kernel where the maximum value occurred in the image and resized it to match the original image size.

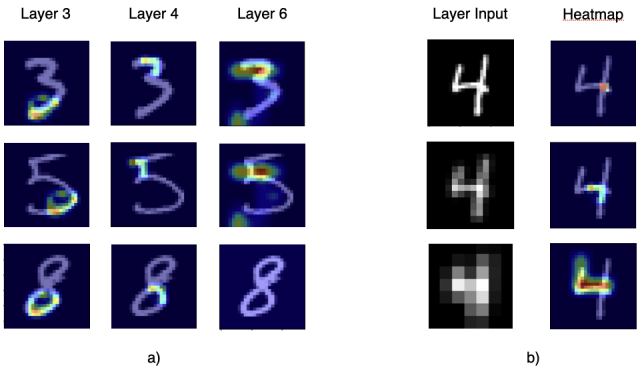

Figure 2: a) Classes 3, 5, and 8 heatmap for Layers 3, 4, and 6. b) Heatmap based on pooled data for Class 4.

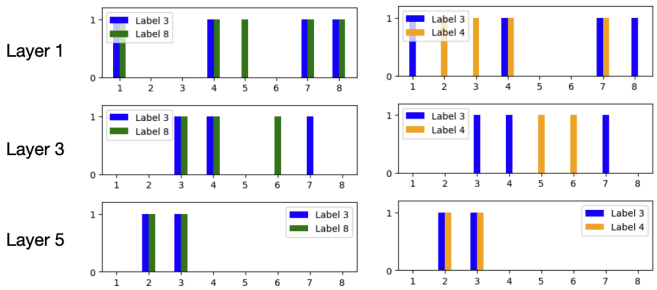

Figure 3: Momory compare between class 3 vs class 8 and class 3 vs class 4. The memories were represented after undergoing binarization

In Figure 2, **a)** Each heatmap displays the channel weights with the maximum values in the convolution results. It can be observed that for class 8, unlike in Layer 6 where no information was extracted, for classes 3 and 5, the maximum values are obtained from common channels. **b)** In this heatmap, as resizing progresses, more global features are captured. As seen in the heatmaps, the kernels of each channel were able to learn patterns that could be *intuitively accepted* by humans.

Also we examined the class-specific memories from some layers. Classes 3 and 8 exhibit relatively similar shapes, and accordingly, their memories have a similar distribution, as shown in the Figure 3. However, the learned memories for classes 3 and 4 are notably different. In layer 5, all classes—3, 4, and 8—have learned the same pattern.

Conclusion

Through this research, we propose a structural change in neural networks that separates recognition from decision-making. We anticipate that this approach will enhance the interpretability of the model and reduce the gap between the functioning of the brain and neural networks. Furthermore, we expect it to lay the groundwork for a more diverse range of research in biological-inspired neural networks.

References

Gordleeva, S. Y., Tsybina, Y. A., Krivonosov, M. I., Ivanchenko, M. V., Zaikin, A. A., Kazantsev, V. B., & Gorban, A. N. (2021). Modeling working memory in a spiking neuron network accompanied by astrocytes. *Frontiers in Cellular Neuroscience*, *15*, 631485.

Lin, Z., Zhao, T., Yang, G., & Lintao, Z. (2018, 07). Episodic memory deep q-networks. In (p. 2433-2439). doi: 10.24963/ijcai.2018/337

Yang, H., Chen, B., Hu, J., Huang, T., Geng, J., & Tang, L. (2023). Modeling working memory using convolutional neural networks for knowledge tracing. In *Advanced intelligent computing technology and applications: 19th international conference, icic 2023, zhengzhou, china, august 10–13, 2023, proceedings, part ii* (p. 137–148). Berlin, Heidelberg: Springer-Verlag. doi: 10.1007/978-981-99-4742-3$_1$1