Neural Dynamics and Deep Learning

Connecting text and images using predictive coding

1 Introduction

This document delves into the intersection of neural dynamics and deep learning, focusing on predictive coding—a framework for understanding how the brain processes information. It explores the construction of a shared representation space through the integration of modality embeddings, utilizing a dataset of basic colored shapes and text labels. The work assesses the impact of model architecture on learning efficiency and the potential of models to capture and relate intricate patterns and structures in data.

> What is the influence of architecture on predictive coding network in cross modal reconstruction?

2 Predictive coding

Our perception largely depends on what we expect to perceive. We expect that coffee is hot, so we take the first sips slowly or as dark thunderclouds gather in the sky, you anticipate that rain is comming. Indeed, it would be inefficient and very costly for our brain to just passively encode the massive stream of sensory information. This intuition stands behind the predictive coding model, that came in the late 90s [4]. A few years later, it turned out, that predictive coding naturally emerges as a model of perception under a unified brain theory making predictive coding an maximization (or free-energy minimization) schema for hierarchical Gaussian generative models [5].

Predictive coding is a theoretical framework used to explain how the brain processes information, particularly in the context of perception. It suggests that the brain constantly generates predictions about sensory inputs and compares these predictions with actual sensory information to minimize prediction errors. Predictive coding often incorporates hierarchical models, where higher-level predictions influence lower-level predictions

Any computational model would need to satisfy the following constraints to be considered biologically plausible:

- Local computation: A neuron performs computations only on the basis of the activity of its input neurons and synaptic weights associated with these inputs (rather than information encoded in other parts of the circuit).
- Local plasticity: Synaptic plasticity is only based on the activity of pre-synaptic and post-synaptic neurons.

The model of Rao and Ballard (1999) [4] fully satisfied these constraints making it suitable to use as a baseline model.

3 Method

We experiment with model architecture in terms of model depth and width size within the layers in order to better understand the limits of predictive coding when applied in this setting.

3.1 Dataset

The challenge of integrating modality embeddings to establish a unified representation space is considerable. Our primary objective is to learn a fundamental mapping task. For this purpose, we utilize a dataset of basic colored shapes[1] paired with their respective textual encoded labels.

Ideally, word embeddings would be employed due to their capacity to encapsulate relationships between words, which is crucial for understanding language and interpreting the world. Our goal is to achieve a commutative embedding property, where the combination of embeddings for words like 'red' and 'apple' aligns closely with the embedding for 'red apple'.

[1] https://github.com/AdityaDutt/MultiColor-Shapes-Database/tree/main

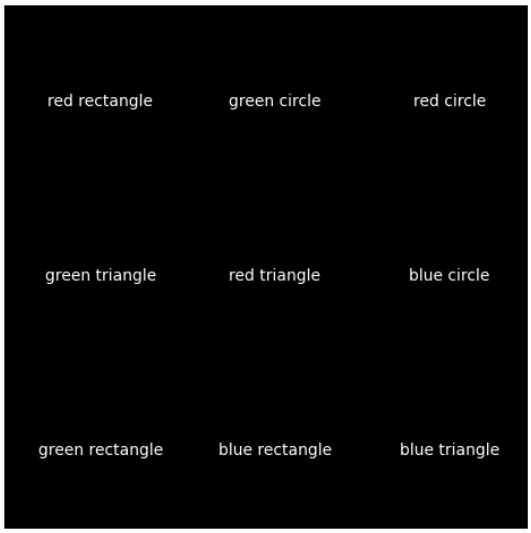

Figure 1: Text embedding classes (a)

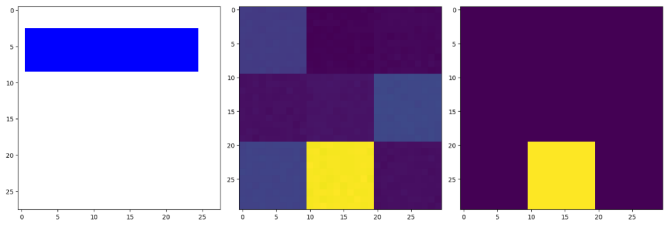

Figure 2: Given modality, Image2text reconstruction, true label

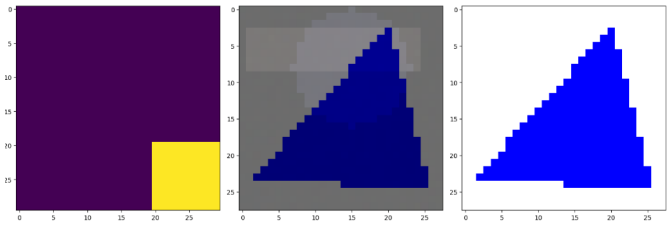

Figure 3: Given modality, Text2image reconstruction, true label

However, while such features are straightforward to learn and reconstruct in a unimodal context, correlating these features with other modalities and different features in other embeddings presents a significant challenge. This complexity hinders their direct applicability in our setting. Consequently, we opt for constructing embeddings using label encoding that corresponds to specific classes.

Our aim is to ensure that visually similar embeddings correspond closely in semantic meaning with text. This alignment, however, is complicated by the entropy inherent in the textual descriptions or classes. For instance, a term like 'top blue circle' could correspond to multiple visual representations, introducing more randomness and thereby increasing the difficulty of the shared learning task. High entropy in the text or class descriptions, indicating greater variability in representation, thus poses a significant challenge in achieving a cohesive and effective shared learning model.

3.2 Model

Our model aims to learn a shared representation that encapsulates the relationship between the averages of different classes, denoted as $class_i$ and $class_j$. The specific labeling of these classes is crucial for distinguishing between attributes, such as size differences in rectangles. In simpler terms, the model is tasked with reconstructing entities like a *blue circle*. However, without a detailed class description, the model tends to default to the average representation of the specified class. This phenomenon and its implications are further illustrated in Appendix B, particularly in Figure 8.

The model's architecture features an all-to-all connection between two layers. This involves flattening the input image and establishing a fully connected network, where each pixel is connected to a corresponding pixel in the subsequent layer. The initial modality receives a flattened 2352-dimensional input (dimensions 28x28x3 from an image), which is coupled with a flattened 30x30-dimensional *text embedding* input. This configuration enables the model to process multi-modal data.

3.2.1 Inference Process

The inference process in our model begins by setting the firing rate of the neurons in the first layer to match the input modality. This step initiates the propagation of the signal through the network. During this signal transmission, the neuron firing rates, represented as r, are updated at each layer. These updates are guided by the learned synaptic weights denoted as U.

This sequential updating continues layer by layer, until the signal reaches the final layer, where we obtain the final representation, r_h. The entire process of signal propagation and the transformation of the input through the layers of the network can be visualized in Figure 6. This figure illustrates how the input is processed and transformed as it progresses through the network, leading to the final inferred representation in the context of the predictive coding framework.

3.2.2 Reconstruction Process

The reconstruction process in our network commences at the top layer. Here, we can propagate the error downwards, updating the layer weights (U) and successively improving our reconstruction at each level. This reconstruction can occur in two primary forms:

Unimodal Reconstruction: In this method, we first infer an image by encoding an input into a latent representation. Following this, we reconstruct the output along the same unimodal path. This approach maintains consistency in modality, essentially performing an 'image-to-image' or 'text-to-text' reconstruction.

Cross-Modal Reconstruction: This method parallels unimodal reconstruction in its initial steps, wherein we encode an input into the shared representation, r. However, during the decoding phase, a different modality is chosen for reconstruction. This strategy enables transformations such as *Text2Image* and *Image2Text*, as depicted in figures 2 and 3. It allows the network to convert input from one modality to an output in another, enhancing the model's versatility.

3.3 Metrics

To assess the effectiveness of our model, particularly in the context of Text2Image or Image2Text reconstructions, we employ cosine similarity as the primary metric. Cosine similarity, distinct from Mean Squared Error (MSE), measures the cosine of the angle between two vectors. This metric is especially relevant in the field of image comparison, as it focuses on the similarity in the direction of image vectors, irrespective of their magnitude. The utility of cosine similarity lies in its ability to evaluate patterns or structures in images, prioritizing these elements over exact pixel-by-pixel intensity matches. As outlined by Pirlo (2013) [3], this approach is particularly effective in scenarios where the comparison of image structures and relative positions is more critical than the intensity of individual pixels. By inherently normalizing the image vectors, cosine similarity demonstrates robustness against variations in lighting, color intensity, and other factors that might alter the scale of pixel values.

4 Results

Our investigation centered on the impact of model architecture, specifically examining the influence of increased depth and width in model layers.

In conventional deep learning models, a deeper architecture can encounter challenges in signal propagation through successive layers, which may hinder learning. Figure 4, where we illustrate the effect of adding intermediate layers at half the size of their predecessors, we see mostly see an upwards trends in model performance.

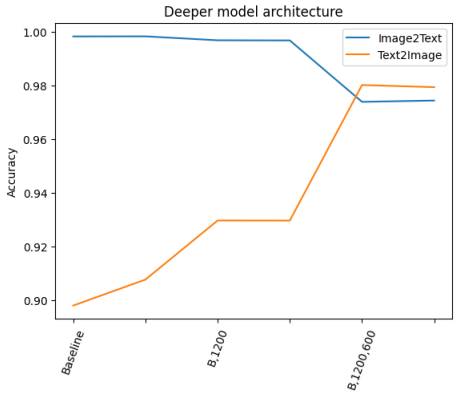

Figure 4: Gradually adding intermediate layers with half of the size.

The data indicates that as the model depth is increased, there is an observable trade-off in accuracy between the two unimodal reconstructions—Image2Text and Text2Image. These reconstructions appear to compete for representation within the shared latent space. The graph in Figure 4 presents this trend, with deeper architectures influencing the performance of each reconstruction type distinctly.

Adding to this complexity is the biological implausibility of indiscriminately widening intermediate layers. Hierarchical structures observed in the brain suggest that higher-level features are composed of combined lower-level features, as discussed by Friston (2008) [1]. This implies that an effective model should not merely expand layer width but should also emulate the hierarchical organization found in biological neural networks.

Although we see no significant difference in model performance, we notice that *Text2Image* performance drops when reducing the layer to be smaller than the input size of (2352).

5 Conclusion

The model's ability to grasp the general concepts of colors and shapes is contingent upon the descriptiveness of the labels, as evidenced in Figure 2. When labels are sufficiently detailed, the model not

Figure 5: Gradually increasing the size of the intermediate layer of both unimodals with a fixed joint size

only identifies the correct class—such as a *blue rectangle*—but also recognizes other classes sharing similar shapes or colors, suggesting a broader conceptual understanding. Conversely, if labels lack consistency in defining shape, color, and position, the model may default to an average representation of the class, leading to suboptimal reconstructions in text-to-image tasks. Model width and size don't seem to be the limiting factor when in comes to cross modal reconstruction.

6 Future work

- Integrating relative positional data, such as the orientation of shapes (left, right, top, bottom), could enhance the model's understanding of spatial context.
- Investigating the weight distribution in deeper and wider models could shed light on the slower decrease in Mean Squared Error (MSE) loss and the corresponding impact on reconstruction quality. Exploring architectural innovations, such as the implementation of residual connections described in He et al. (2016) [2], or sparse lateral connections to facilitate earlier integration of modalities, may improve the formation of shared representations.
- Testing the model's ability to rapidly assimilate new shapes and colors through few-shot learning could provide insights into its proficiency in relating and integrating new concepts.
- Expanding the model to include GNNs could leverage the advantages of graph structures, which are not confined to feed-forward models. GNNs could offer benefits such as:
  - Local Updating Rule: Mimicking the brain's neuronal updates, GNNs revise node representations using neighboring node information.
  - Efficiency in Learning: Their potential efficiency in learning relational and hierarchical structures may mirror the brain's own approach to understanding the world.
  - Hierarchical Processing: GNN architectures that process information hierarchically could align with the brain's processing under predictive coding theories.
- Word embedding alreayd capture the following relation:

$$emb(king) - emb(man) = emb(queen) - emb(female)$$

  So using contrastive learning in combination with word embedding might work to capture relations and features between modalities.
- Drawing inspiration from human-in-the-loop methodologies or reinforcement learning (RL), where acquisition functions guide model improvement in response to incorrect predictions and large errors, could enhance the model's adaptive learning capabilities.

References

[1] Karl Friston. "Hierarchical models in the brain". In: *PLoS computational biology* 4.11 (2008), e1000211.

[2] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[3] Giuseppe Pirlo and Donato Impedovo. "Cosine similarity for analysis and verification of static signatures". In: *IET biometrics* 2.4 (2013), pp. 151–158.

[4] Rajesh PN Rao and Dana H Ballard. "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects". In: *Nature neuroscience* 2.1 (1999), pp. 79–87.

[5] Tommaso Salvatori et al. "Brain-inspired computational intelligence via predictive coding". In: *arXiv preprint arXiv:2308.07870* (2023).

A Appendix

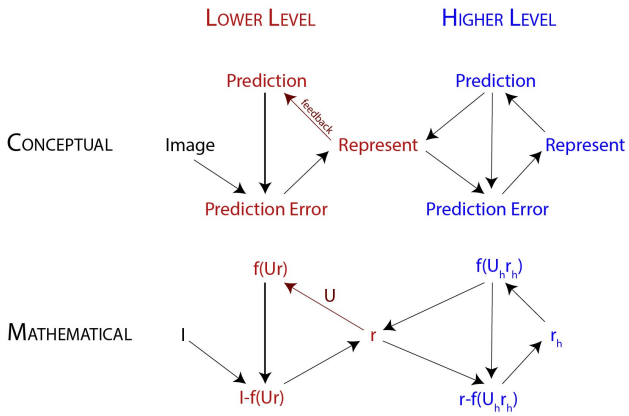

Figure 6: Updated illustration of Rao and Ballard

B Reconstructions

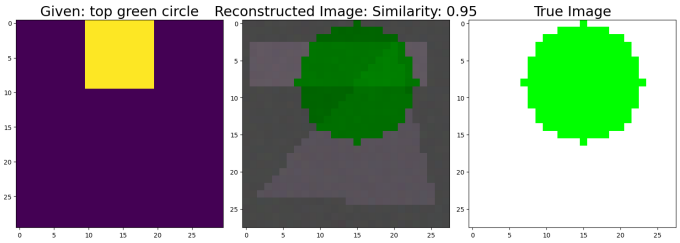

Figure 7: Cross modal reconstruction of the embedding 'top green circle'

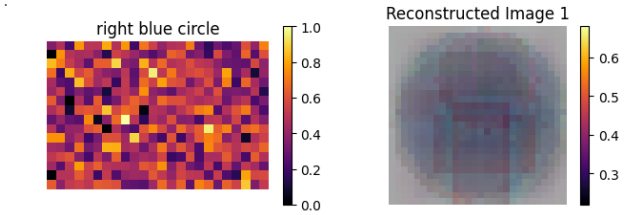

Figure 8: Cross modal reconstrucion of the avg blue circle in the dataset