Representational subspaces with different levels of abstraction in transformers

Brian S. Robinson (brian.robinson@jhuapl.edu)
Johns Hopkins University Applied Physics Laboratory
Laurel, MD 20723, United States

Colin Conwell (colinconwell@jh.edu)
Department of Cognitive Science, Johns Hopkins University
Baltimore, MD 21218, United States

Michael F. Bonner (mfbonner@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
Baltimore, MD 21218, United States

Abstract

A widespread assumption in analyzing the representations of artificial neural networks (ANNs) and the brain is that neurons in the same ANN layer or cortical region have a shared level of abstraction. In this work, by analyzing the learned LayerNorm weights across a range of transformer networks, we find evidence for distinct subspaces in the network dimensions. In an in-depth analysis for a single vision transformer, we find three representational subspaces within each layer that can be identified by LayerNorm weights. In comparisons to human fMRI representations, we find distinct properties of these subspaces with two of the subspaces demonstrating higher representational similarity to early and late regions of the cortical visual hierarchy. These findings show that analyses of hierarchical feature processing in ANNs need to consider the role of subspaces with distinct representational properties.

Keywords: Vision Transformers; Neural Network Representations; Models of Human Vision; fMRI; DNN; Visual Cortex

Introduction

In both the brain and ANNs, a foundational principle is that successive layers of feedforward processing correspond to the conversion of low-level sensory features into abstract high-level features (DiCarlo, Zoccolan, & Rust, 2012). Furthermore, a widespread assumption in analyzing ANNs is that representations in the same ANN layer have a shared level of abstraction (Yamins et al., 2014). Modern transformer-based neural networks (Vaswani et al., 2017; Dosovitskiy et al., 2020), however, have a flexible system design which does not include an explicit inductive bias for hierarchical feature processing in successive layers (as is the case in widely studied convolutional neural networks). Here we examine whether the representations of vision transformers diverge from the conventional hierarchy of representational abstraction.

Results

Analysis of Network Representations We used the learned weights from the layer normalization components in vision and language transformer models as a tool to analyze the properties of representations within individual network layers (Fig. 1). Layer normalization standardizes the input embeddings and applies learnable weight and bias vectors across all dimensions. Notably, these weight vectors are shared across tokens and can be used as a tool to investigate shared properties across network dimensions.

We found that in early layers of highly trained transformer models, a considerable portion of the network's dimensions effectively ignore or "squash" the input by having near-zero layer normalization weights. This was especially prominent in more complex models like ViT-CLIP for vision and Llama-2 for language, where around 45% and 97% of dimensions respectively were effectively squashed in early layers.

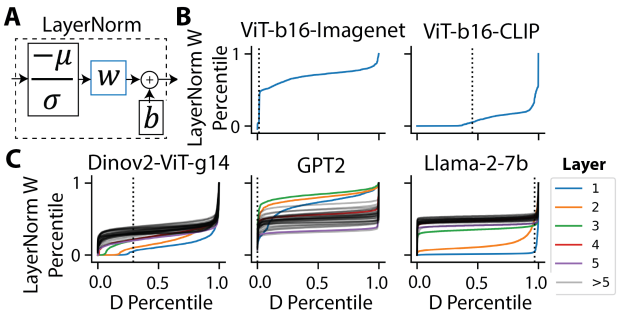

Figure 1: Distribution of LayerNorm weights reveal near-zero incorporation of network inputs for considerable portions of network dimensions in early layers of highly trained transformer networks. A. LayerNorm has learnable weights, w, used for analysis. B. LayerNorm weight distributions for vision transformer networks with pre-normalization. C. LayerNorm weight distributions per layer in vision (Dinov2) and language models (GPT2, Llama-2). Dotted lines denote D percentile at 5% of maximum LayerNorm weight.

Functional Roles of Dimensions Taking a closer look at the ViT-CLIP vision model, we identified three distinct clusters of dimensions based on their properties (Fig. 2):

- Position-Dominated: These dimensions were heavily influenced by the positional embeddings rather than the actual image input.

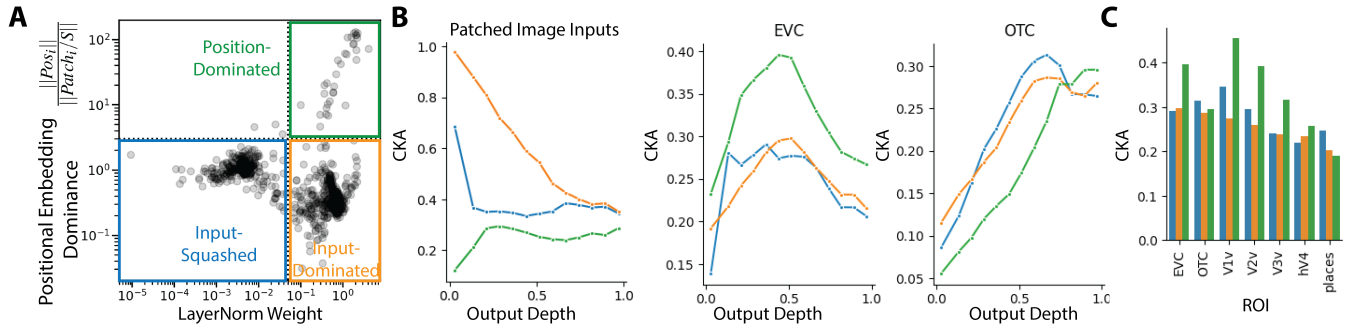

Figure 2: For ViT-CLIP, network dimensions with distinct properties have distinct representational similarity to early and late processing regions of interest in fMRI responses to natural images. A. Network dimensions can be divided into clusters that are Position-Dominated, Input-Dominated, and Input-Squashed, which are investigated separately for representational similarity with linear centered kernel alignment in B-C. B. Representational similarity across layers in the ViT-CLIP network with centered kernel alignment to patched image inputs as well as fMRI responses to early visual cortex and occipital temporal cortex. C. Maximum CKA scores across layers for additional regions of interest.

- Input-Squashed: These dimensions effectively ignored the input via near-zero normalization weights. Their similarity to the image inputs drastically decreased after the first layer and remained low in subsequent layers.
- Input-Dominated: The remaining dimensions were primarily driven by the image input embeddings. These dimensions started with high similarity to the image inputs, and this similarity gradually decreased across layers. This aligns with the typical hierarchical processing of visual features.

Relationship to Brain Representations We next analyzed how these clusters of dimensions mapped onto brain representations from fMRI data of humans viewing natural images, and observed distinct patterns of representational similarity to low- and higher-level regions of visual cortex. We use the DeepJuice codebase (Conwell, Prince, Kay, Alvarez, & Konkle, 2023) to perform representational similarity analysis with linear centered kernel alignment (CKA) (Kornblith, Norouzi, Lee, & Hinton, 2019) on flattened network embeddings to the image inputs of the network and to the fMRI stimulus evoked responses of a representative subject of the natural scenes dataset (Allen et al., 2022). Most strikingly, we found that the position-dominated dimensions were highly similar to EVC and had a relatively lower similarity to OTC. We also found that the Input-Squashed dimensions had a slightly higher similarity to OTC relative to the Input-Dominated dimensions, consistent with the interpretation of these Input-Squashed dimensions as encoding more abstract scene information.

Discussion

Our results reveal that within the same network layer, there can be distinct subspaces of dimensions that encode different levels of visual abstraction in parallel. This challenges the common assumption that all units within a layer operate at the same level of the representational abstraction.

The presence of distinct subspaces based on near-zero LayerNorm weights was observed in multiple transformers. This suggests it may be a general computational strategy of modern neural networks that should be accounted for when analyzing the inner workings of these models and their relationship to the brain.

Acknowledgments

This research was supported by internal funding from the Johns Hopkins University Applied Physics Laboratory.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... Kay, K. (2022, January). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), 116–126.

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2023). What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines. *bioRxiv*.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Similarity of neural network representations revisited. In *International conference on machine learning* (pp. 3519–3529).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619–8624.