Probing Vision Transformer Attention: Image Statistics and Patch Similarity

Aaron Philip[1], Xu Pan[2], & Odelia Schwartz[2]

[1]*Department of Physics and Astronomy, Michigan State University, East Lansing, MI;* [2]*Department of Computer Science, University of Miami, Coral Gables, FL*

Abstract

Background: Recent advances in AI and deep learning, and specifically the transformer architecture, have had dramatic influence on natural language processing [1] (e.g., ChatGPT). The transformer has also been influential in computer vision, forming the basis of most state-of-the-art image recognition algorithms since the Visual Transformer (ViT) was introduced 2 years ago [2]. While Convolutional Neural Networks (CNNs) have been shown to capture properties of visual neural processing in the brain, transformers operate in fundamentally different ways and are not well understood. The self-attention mechanism in the transformer is suggested to be an analogy to associative memory, capturing similarity between embedded patches corresponding to the key and query [3]. However, motivated from studies of visual processing in the brain, biological neurons suppress inputs that are visually similar and homogenous across space, and therefore highlight stimuli that are salient and stand out from their visual surroundings. But there has been limited work probing principled, fundamental properties of the attention for large, trained ViT models.

Method/Results: We quantify the statistical dependencies between embedded image patches in ViTs, and find that the mutual information is higher for later layer neurons and drops off faster with spatial distance (except the first layer which has significantly lower mutual information). We further interpret pairwise patch attention scores based on similarity, which is obtained from the angle between patch embeddings. We find that attention is higher in early layers between similar patches, while later layer attention highlight dissimilarity. This suggests that transformers may partially capture visual salience, which we propose to test with structured visual stimuli in future work.

Keywords: Deep Learning, Image Classification, Vision Transformers.

Attention & ViT

Fundamental to the Transformer architecture is the attention mechanism. Attention allows a model to consider several "tokens" that represent different parts of medium such as a sentence or an image and compare them to one another. A single head attention (SHA) computation uses learnable weight and bias matrices indicated by the **W**s and **B**s. Note that D_k is the inherent dimensionality of the tokens:

$$\mathbf{q_i} = (\mathbf{x_i}\mathbf{W^Q} + \mathbf{B^Q}),$$
$$\mathbf{k_i} = (\mathbf{x_i}\mathbf{W^K} + \mathbf{B^K}),$$
$$\mathbf{v_i} = (\mathbf{x_i}\mathbf{W^V} + \mathbf{B^V}).$$

$$\mathbf{A_{i,j}} = \text{softmax}(\frac{\mathbf{q_i}\mathbf{k_j^T}}{\sqrt{D_k}}),$$
$$\text{SHA}(\mathbf{x_i}) = \sum_{j=0}^{j=N} \mathbf{A_{i,j}}\mathbf{v_j}$$

Multi-Head Attention (MHA) is a simple extension of single-head attention where each head learns different weight and biases to enrich the diversity of learned codomains. In the Vision Transformer, each token is a patch of the image.

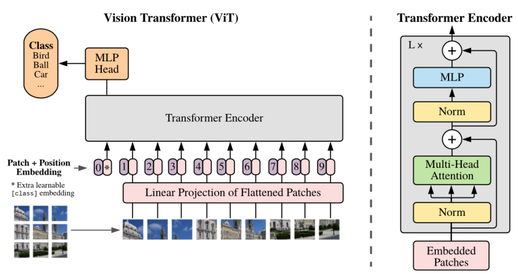

Figure 1: ViT Architecture [1]

Statistical Analysis

We ultimately defined a neuron in a similar manner to those of CNNs. Namely, a neuron in a MHA layer could be specified by its channel number and the corresponding spatial location in the image. To perform statistical analysis, we activated and recorded the responses of all neurons in the network by running inference over a random subset of the ImageNet1K dataset.

Joint Distributions

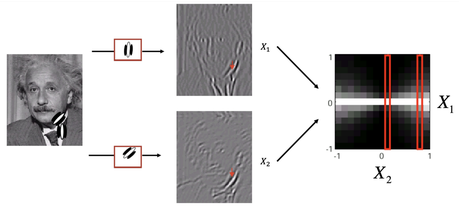

Figure 2: Example of statistically dependent neurons. We use bowtie plots to visualize their conditional dependencies.

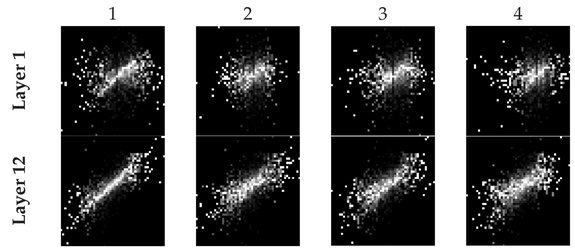

Figure 3: Joint conditional statistics of spatially central ViT neurons and their spatial neighbors at increasing distance.

Mutual Information

$$I(X;Y) = \iint P_{X,Y}(x,y) \log\left(\frac{P_{X,Y}(x,y)}{P_X(x)\cdot P_Y(y)}\right) dx\, dy \quad (1)$$

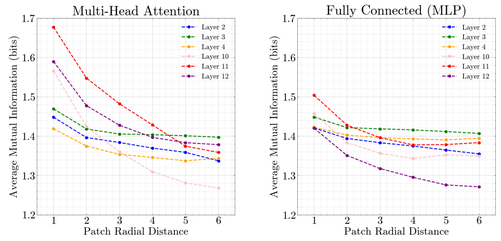

Figure 4: Average mutual information (Eq. 1) between a spatially central neuron and all other neurons in MLP and MHA sublayers. Note that Layer 1 was left out of both due to having considerably lower [0.8,1.0] range.

Main Conclusions

- The first layer (not shown) neurons had considerably lower average mutual information than the other layers. High layers had higher average mutual information than lower layers for nearby patches but fell off more rapidly at increasing distance.
- Relative patch position has clear correlation to neuron mutual informations. This supports findings in [4], [5].
- Most neuron activations followed a normal distribution with first layer neurons' were more narrow with a higher kurtosis (not pictured).

Similarity or Dissimilarity?

Tokens $\mathbf{x_i}, \mathbf{y_i}$ directly interact with one another through the term $\mathbf{x_i}\mathbf{W^Q}\mathbf{W^{K\,T}}\mathbf{y_i^T}$. We decomposed the term into: $\mathbf{W^Q}\mathbf{W^{K\,T}} = \mathbf{M_{sym}} + \mathbf{M_{asym}}$
By decomposing the matrix product into a sum of symmetric and antisymmetric matrices, we studied their relative spectra $\lambda(\mathbf{M_{sym}}), \lambda(\mathbf{M_{asym}})$. We propose an analogy to neural suppression and facilitation seen in neuroscience studies. We found that $\mathbf{M_{sym}}$ was largely positive definite in every MHA layer, implying that its role is to indicate similarity between two patches. $\mathbf{M_{asym}}$ by definition has imaginary eigenvalues, but the coefficient of the imaginary part is the only relevant magnitude to consider and comes in pairs ($\pm\lambda_k$).

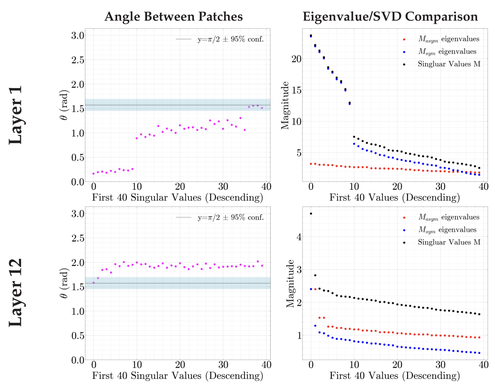

Figure 5: Eigenvalues and singular values of the key-query product. Note that the eigenvalues shown of the antisymmetric matrix also have a paired negative eigenvalue.

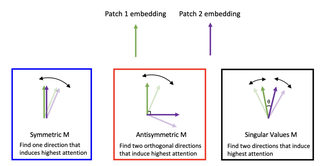

Figure 6: Intuition behind Fig. 5 left column.

Figure 7: See full collection of plots at the QR code link.

Main Conclusions

- The embedding angle was dominated by values higher than 90 degrees in later layers. This also corresponded to key-query products with eigenvalues of the antisymmetric matrix being larger than eigenvalues of the symmetric matrix.
- The singular value decay of most layers is relatively slow indicating a high rank matrix. However, the last layer has dominating principal singular values, indicating that it could scaled down.

Future Work

- We intend to further confirm the relation we find with the angle measure as indicating similarity/dissimilarity with well-chosen stimuli that align closely with the eigenvectors we identified.
- Having studied the properties of well-parameterized ViT, we want to explore constrained training [6] to reduce data quantity requirements.
- The code written enables easy interaction with model weights and activation maps, facilitating a broader range of Transformer studies in other domains such as natural language processing, as well as specific endeavors such as transfer learning or image statistics.

Acknowledgements

I would like to thank my mentors Xu Pan and Professor Odelia Schwartz for their guidance and mentorship. This material is based upon work supported by the National Science Foundation under Grant No. CNS-1949972. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Ashish Vaswani et. al. Attention is all you need, 2017.

[2] Alexey Dosovitskiy et. al. An image is worth 16x16 words: Transformers for image recognition at scale. 2021.

[3] Paria Mehrani and John K. Tsotsos. Self-attention in vision transformers performs perceptual grouping, not attention, 2023.

[4] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks?, 2022.

[5] Van-Anh Nguyen, Khanh Pham Dinh, Long Tung Vuong, Thanh-Toan Do, Quan Hung Tran, Dinh Phung, and Trung Le. Vision transformer visualization: What neurons tell and how neurons behave?, 2022.

[6] M. Raissi, P. Perdikaris, and G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.