Are ViTs as global as we think? - Assessing model locality for brain-model mapping

Fangrui Huang (fangruih@stanford.edu)
Computer Science Department, 290 Jane Stanford Way
Stanford, California 94305 United States of America

Klemen Kotar (klemenk@stanford.edu)
Computer Science Department, 290 Jane Stanford Way
Stanford, California 94305 United States of America

Wanhee Lee (wanhee@stanford.edu)
Applied Physics Department, 290 Jane Stanford Way
Stanford, California 94305 United States of America

Rosa Cao (rosacao@stanford.edu)
Philosophy Department, 290 Jane Stanford Way
Stanford, California 94305 United States of America

Daniel Yamins (yamins@stanford.edu)
Computer Science Department, 290 Jane Stanford Way
Stanford, California 94305 United States of America

Abstract

Recent explorations into the neural predictivity of Vision Transformer (ViT) models have shown remarkable similarities to traditional Convolutional Neural Networks (CNNs) in predicting neural responses within the visual cortex. This juxtaposition raises intriguing questions about the underlying architectural similarities and differences between these two model types, particularly in the context of spatial locality. Our study investigates the locality of receptive fields within ViTs compared to CNNs, employing a novel methodological approach that adjusts for differences in layer resolutions and total number of layers across models. Our findings suggest that despite ViTs' global connectivity potential through attention mechanisms, they exhibit a strong bias towards local processing akin to CNNs, particularly after training. This convergence in locality patterns may explain their similar effectiveness in neural predictivity, providing new insights into how transformative architectures process visual information and their neurophysiological parallels.

Keywords: AI; neuroscience; brain-model mapping; interpretability

Introduction

It has recently been observed that Vision Transformer (ViT)(Dosovitskiy et al., 2020) models have similar ability to predict neural responses in visual corte as Convolutional Neural Networks (CNNs)(Conwell, Prince, Kay, Alvarez, & Konkle, 2022). This observation is potentially surprising, because transformers and CNNs seem at first glance like very different architectures. To understand this, it is useful if we first think through what make Vision Transformers (ViTs) and CNNs similar and different. The three core neurophysiological observations of the vision cortex are that it is (1) arranged in a hierarchy of areas, (2) roughly speaking, the input-output transfer function computed in each area of the hierarchy is composed of small number of simple but computationally universal linear-nonlinear operations, and (3) these operations are largely spatially local, leading to a gradual increase in receptive field size across the visual hierarchy. CNNs build in a version of all three of these features, by virtue of being (1) feedforward, (2) composed (largely) of Linear-ReLU blocks and residual connections, and (3) applying these operations convolutionally with small kernel sizes.

ViTs, on the other hand, build in the first two of these core features. As for the first feature – hierarchy, ViTs use a similar feedforward pattern to the CNNs, and typically have about the same number of layers as CNNs. As for the second core feature – simple but universal computational primitives – the transformer blocks inside each ViT layer are composed of a standard Linear-Nonlinear MLP pathway, complemented by a multiplicative attention pathway. This general class of operations is similar in a coarse sense to the operations inside a CNNs, in that they are composed of a small number of computationally-universal primitives, albeit with a different micro-architecture expressing different specific inductive biases. However, it is with regard to the third feature – locality

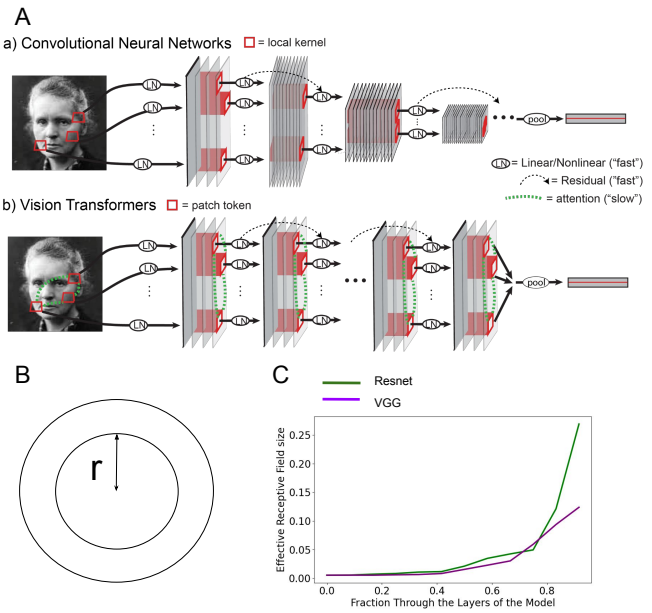

Figure 1: **A)** Differences of CNN and ViT in how they process information. **B)** Quantification of receptive field. **C)** Example for quantification of receptive field.

– where it is less clear whether how CNNs and ViTs compare. Unlike with convolution where a limited-size kernel is applied equivalently at each spatial location, in the ViT, the transformer attention pathway can be spatially non-local, with potential connections between units that are spatially widely separated. On the other hand, the typical tokenization used with ViTs – namely, small (e.g. 8x8 or 16x16 pixel) spatial patches – is inherently spatially local, a fact that is further emphasized by the spatial encoding attached to most ViTs. Moreover, the MLP pathway of the transformer block presents a natural "fast pathway" by which each of the spatially-local positional patches is preferentially connected to the equivalent token at the next layer, establishing a natural spatial grid with 1-1 correspondences between spatially equivalent positions across all layers. Finally, even though the attention weights of the ViT can in principle be arbitrarily global, they could learn through training to be become substantially more local. Thus, it is an empirical question of how local ViTs (both trained and untrained) are – and thus, whether they organically possess (or learn to possess) the same brain-like features of CNNs. To the extent that ViTs are local, it is perhaps less surprising that they have been found similar to CNNs in their neural predictivity capacities.

Methods

Here, we investigate the question by quantifying and comparing between the receptive field of models in order to assess locality.

Receptive field quantification For any hierarchical network, we compute receptive fields of each layer in the models using the absolute value of the gradient of input pixels from

central part of the each layer. Since the resolution of various layers is different across models, we choose the central part of each layer to have the same relative size across all layers in all models. For each layer i, the receptive field is represented as

$$\frac{\delta \sum(Y_{central})}{\delta M}$$

where Y is the output of layer i; M is the receptive field map(input pixels). Then we summarize the locality of each receptive field map M by:

$$\frac{1}{M_{max}} \int_0^R r E_r[M] \, dr$$

where $E_r[M]$ is the expected value of the receptive field at radius r. (Figure 1)

Models In this analysis we compare the results for:
(a) several convolutional neural networks (ResNet50, VGG) and vision transformers (ViT, DINOv1)
(b) trained and untrained networks. (ViT is trained on ImageNet-21k. All other models are trained on ImageNet-1k.)
(c) two distinct computation pathways of the vision transformer architecture. Vision transformers contain a "fast pathway" where a residual connection is directly added to the output of each layer introducing a strong local bias to the exact corresponding patch in the input space. They also contain a "slow pathway" where the attention mechanism introduces global connections between input patches. In order to separate the "fast pathway" from the "slow pathway" we separate each attention block into 2 micro-stages: one right after the calculation of the self-attention mechanism("slow pathway"); and the other after adding the residual("fast pathway"). While we can not fully disentangle the effects of the two pathways, this approach allows us to analyze their dominant characteristics.

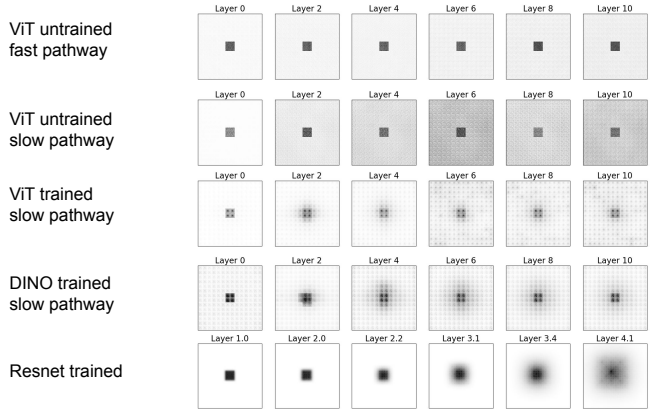

Figure 2: Receptive fields of trained and untrained ViTs and CNN across layers.

Results

Inherent Bias Toward Locality in ViTs Though ViTs are thought to be global because of their attention mechanism, they actually have a strong local bias. The embeddings of both trained and untrained ViTs are heavily biased toward their corresponding patches in the input space as shown in figure 2. This results from the fast feedforward residual pathway predominantly influencing the receptive fields, suggesting that non-local attention weights play a lesser role. This finding implies that the intrinsic design of ViTs favors local interactions.

Locality Across Sub-modules Within Attention Blocks In order to separate out the fast feedforward pathway from slow pathway, we analyze receptive fields for sub-modules within each attention blocks. One of the specific micro-stages within the untrained ViT ("attention-pre-residual" sublayer) appears substantially more global than other layers. This is because it is influenced less by the fast residual pathway. Apart from the central patches, the untrained ViT emphasizes the input space uniformly. With training, this global effect is reduced(figure 2). The trained ViT and DINO "slow pathway" emphasizes pixels surrounding the central parts which means it learns to have more local receptive fields even if it has the potential to have global receptive field. This is also summarized in figure 3. The receptive field size of "slow pathway" is large for untrained ViT and DINO and it becomes substantially more local after training.

Convergence of Receptive Field Patterns with CNNs Upon Training Trained ViTs have a pattern of increase in receptive field size that is reasonably similar to CNNs. In other words, training has the effect of adjusting the attention weights to allow the ViT layers to compute somewhat more local and CNN-like functions of its inputs.

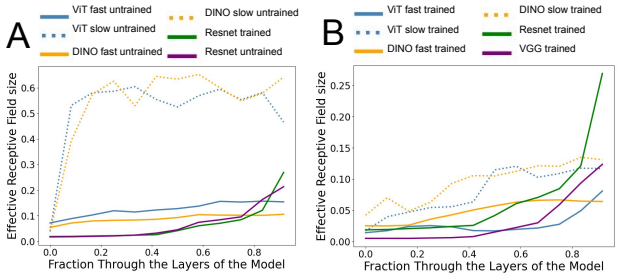

Figure 3: **Changes in sizes of receptive fields A)** The "slow pathway" of an untrained ViT has a large receptive field, while the "fast pathway" has a smaller receptive field. **B)** After training the receptive field of "slow pathway" becomes substantially smaller. The overall trend of the receptive fields of trained ViTs is similar to that of CNNs.

Conclusion

Taken together, we develop a method for fairly comparing receptive fields between layers across different model architectures and quantify their receptive fields. Our results suggest that although ViTs have an inherent locality bias and a weaker global connectivity path, which is invariant to locality at initialization, they modify their receptive field during the course of training to roughly resemble that of a CNN. Since these findings suggest and ViTs automatically, learn a receptive field bias similar to the one hand designed in the CNN architecture it might not be surprising that they have similar predictability for neural responses.

References

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 9650–9660).

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2022). What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv*, 2022–03.

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*.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356–365.

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.