Do Vision Transformers See Like Convolutional Neural Networks?

Maithra Raghu[1] **Thomas Unterthiner**[1] **Simon Kornblith**[1] **Chiyuan Zhang**[1] **Alexey Dosovitskiy**[1]

[1]*Google Research, Brain Team*

Abstract

Convolutional neural networks (CNNs) have so far been the de-facto model for visual data. Recent work has shown that (Vision) Transformer models (ViT) can achieve comparable or even superior performance on image classification tasks. This raises a central question: *how are Vision Transformers solving these tasks*? Are they acting like convolutional networks, or learning entirely different visual representations? Analyzing the internal representation structure of ViTs and CNNs on image classification benchmarks, we find striking differences between the two architectures, such as ViT having more uniform representations across all layers. We explore how these differences arise, finding crucial roles played by self-attention, which enables early aggregation of global information, and ViT residual connections, which strongly propagate features from lower to higher layers. We study the ramifications for spatial localization, demonstrating ViTs successfully preserve input spatial information, with noticeable effects from different classification methods. Finally, we study the effect of (pretraining) dataset scale on intermediate features and transfer learning, and conclude with a discussion on connections to new architectures such as the MLP-Mixer.

1 Introduction

Over the past several years, the successes of deep learning on visual tasks has critically relied on convolutional neural networks [20, 16]. This is largely due to the powerful inductive bias of spatial equivariance encoded by convolutional layers, which have been key to learning general purpose visual representations for easy transfer and strong performance. Remarkably however, recent work has demonstrated that Transformer neural networks are capable of equal or superior performance on image classification tasks at large scale [14]. These Vision Transformers (ViT) operate almost *identically* to Transformers used in language [13], using self-attention, rather than convolution, to aggregate information across locations. This is in contrast with a large body of prior work, which has focused on more explicitly incorporating image-specific inductive biases [30, 9, 4]

This breakthrough highlights a fundamental question: *how* are Vision Transformers solving these image based tasks? Do they act like convolutions, learning the same inductive biases from scratch? Or are they developing novel task representations? And what is the role of scale in learning these representations? In this paper, we study these questions, uncovering insights about key differences between ViTs and CNNs. Specifically, our contributions are

- We investigate the internal representation structure of ViTs and CNNs, finding striking differences between the two models, such as ViT having more uniform representations, with greater similarity between lower and higher layers.
- Analyzing how local/global spatial information is utilised, we find ViT incorporates more global information than ResNet at lower layers, leading to quantitatively different features.

- Nevertheless, we find that incorporating local information at lower layers remains vital, with large-scale pre-training data helping early attention layers learn to do this
- We study the uniform internal structure of ViT, finding that skip connections in ViT are even more influential than in ResNets, having strong effects on performance and representation similarity.
- Motivated by potential future uses in object detection, we examine how well input spatial information is preserved, finding connections between spatial localization and methods of classification.
- We study the effects of dataset scale on transfer learning, with a linear probes study revealing its importance for high quality intermediate representations.

2 Related Work

Developing non-convolutional neural networks to tackle computer vision tasks, particularly Transformer neural networks [44] has been an active area of research. Prior works have looked at *local* multiheaded self-attention, drawing from the structure of convolutional receptive fields [30, 36], directly combining CNNs with self-attention [4, 2, 46] or applying Transformers to smaller-size images [6, 9]. In comparison to these, the Vision Transformer [14] performs even less modification to the Transformer architecture, making it especially interesting to compare to CNNs. Since its development, there has also been very recent work analyzing aspects of ViT, particularly robustness [3, 31, 28] and effects of self-supervision [5, 7]. Other recent related work has looked at designing hybrid ViT-CNN models [49, 11], drawing on structural differences between the models. Comparison between Transformers and CNNs are also recently studied in the text domain [41].

Our work focuses on the representational structure of ViTs. To study ViT representations, we draw on techniques from neural network representation similarity, which allow the quantitative comparisons of representations within and across neural networks [17, 34, 26, 19]. These techniques have been very successful in providing insights on properties of different vision architectures [29, 22, 18], representation structure in language models [48, 25, 47, 21], dynamics of training methods [33, 24] and domain specific model behavior [27, 35, 38]. We also apply *linear probes* in our study, which has been shown to be useful to analyze the learned representations in both vision [1] and text [8, 32, 45] models.

3 Background and Experimental Setup

Our goal is to understand whether there are differences in the way ViTs represent and solve image tasks compared to CNNs. Based on the results of Dosovitskiy et al. [14], we take a representative set of CNN and ViT models — ResNet50x1, ResNet152x2, ViT-B/32, ViT-B/16, ViT-L/16 and ViT-H/14. Unless otherwise specified, models are trained on the JFT-300M dataset [40], although we also investigate models trained on the ImageNet ILSVRC 2012 dataset [12, 37] and standard transfer learning benchmarks [50, 14]. We use a variety of analysis methods to study the layer representations of these models, gaining many insights into how these models function. We provide further details of the experimental setting in Appendix A.

Representation Similarity and CKA (Centered Kernel Alignment): Analyzing (hidden) layer representations of neural networks is challenging because their features are distributed across a large number of neurons. This distributed aspect also makes it difficult to meaningfully compare representations across neural networks. Centered kernel alignment (CKA) [17, 10] addresses these challenges, enabling quantitative comparisons of representations within and across networks. Specifically, CKA takes as input $\mathbf{X} \in \mathbb{R}^{m \times p_1}$ and $\mathbf{Y} \in \mathbb{R}^{m \times p_2}$ which are representations (activation matrices), of two layers, with p_1 and p_2 neurons respectively, evaluated on the same m examples. Letting $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^\top$ and $\boldsymbol{L} = \boldsymbol{Y}\boldsymbol{Y}^\top$ denote the Gram matrices for the two layers (which measures the similarity of a pair of datapoints according to layer representations) CKA computes:

$$\mathrm{CKA}(\boldsymbol{K}, \boldsymbol{L}) = \frac{\mathrm{HSIC}(\boldsymbol{K}, \boldsymbol{L})}{\sqrt{\mathrm{HSIC}(\boldsymbol{K}, \boldsymbol{K})\mathrm{HSIC}(\boldsymbol{L}, \boldsymbol{L})}}, \tag{1}$$

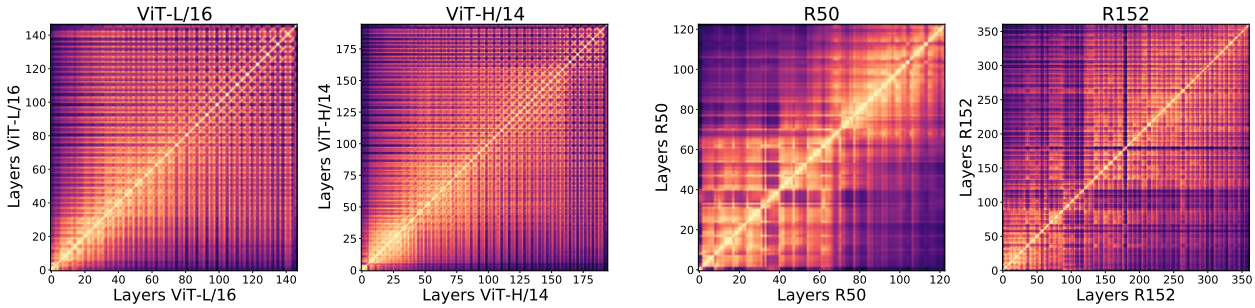


Figure 1: Representation structure of ViTs and convolutional networks show significant differences, with ViTs having highly similar representations throughout the model, while the ResNet models show much lower similarity between lower and higher layers. We plot CKA similarities between all pairs of layers across different model architectures. The results are shown as a heatmap, with the x and y axes indexing the layers from input to output. We observe that ViTs have relatively uniform layer similarity structure, with a clear grid-like pattern and large similarity between lower and higher layers. By contrast, the ResNet models show clear stages in similarity structure, with smaller similarity scores between lower and higher layers.

where HSIC is the Hilbert-Schmidt independence criterion [15]. Given the centering matrix $\boldsymbol{H} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\mathsf{T}$ and the centered Gram matrices $\boldsymbol{K}' = \boldsymbol{HKH}$ and $\boldsymbol{L}' = \boldsymbol{HLH}$, $\text{HSIC}(\boldsymbol{K}, \boldsymbol{L}) = \text{vec}(\boldsymbol{K}') \cdot \text{vec}(\boldsymbol{L}')/(m-1)^2$, the similarity between these centered Gram matrices. CKA is invariant to orthogonal transformation of representations (including permutation of neurons), and the normalization term ensures invariance to isotropic scaling. These properties enable meaningful comparison and analysis of neural network hidden representations. To work at scale with our models and tasks, we approximate the unbiased estimator of HSIC [39] using minibatches, as suggested in [29].

4 Representation Structure of ViTs and Convolutional Networks

We begin our investigation by using CKA to study the internal representation structure of each model. How are representations propagated within the two architectures, and are there signs of functional differences? To answer these questions, we take every pair of layers $\boldsymbol{X}, \boldsymbol{Y}$ within a model and compute their CKA similarity. Note that we take representations not only from outputs of ViT/ResNet blocks, but also from intermediate layers, such as normalization layers and the hidden activations inside a ViT MLP. Figure 1 shows the results as a heatmap, for multiple ViTs and ResNets. We observe clear differences between the internal representation structure between the two model architectures: (1) ViTs show a much more uniform similarity structure, with a clear grid like structure (2) lower and higher layers in ViT show much greater similarity than in the ResNet, where similarity is divided into different (lower/higher) stages.

We also perform cross-model comparisons, where we take all layers $\boldsymbol{X}$ from ViT and compare to all layers $\boldsymbol{Y}$ from ResNet. We observe (Figure 2) that the lower half of 60 ResNet layers are similar to approximately the lowest quarter of ViT layers. In particular, many more lower layers in the ResNet are needed to compute similar representations to the lower layers of ViT. The top half of the ResNet is approximately similar to the next third of the ViT layers. The final third of ViT layers is less similar to all ResNet layers, likely because this set of layers mainly manipulates the CLS token representation, further studied in Section 6.

Taken together, these results suggest that (i) ViT lower layers compute representations in a different way to lower layers in the ResNet, (ii) ViT also more strongly propagates representations between lower and higher layers (iii) the highest layers of ViT have quite different representations to ResNet.

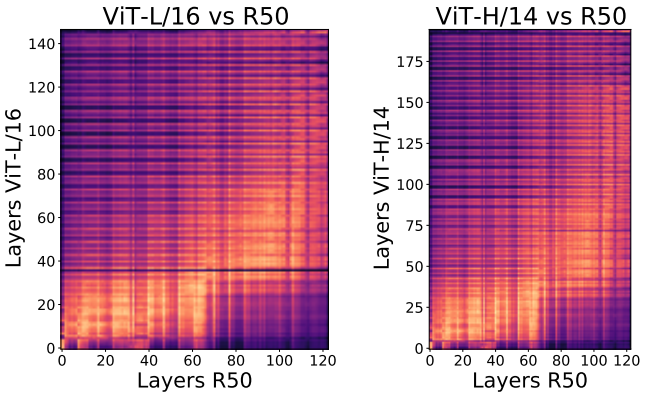


Figure 2: Cross model CKA heatmap between ViT and ResNet illustrate that a larger number of lower layers in the ResNet are similar to a smaller set of the lowest ViT layers. We compute a CKA heatmap comparing all layers of ViT to all layers of ResNet, for two different ViT models. We observe that the lower half of ResNet layers are similar to around the lowest quarter of ViT layers. The remaining half of the ResNet is similar to approximately the next third of ViT layers, with the highest ViT layers dissimilar to lower and higher ResNet layers.

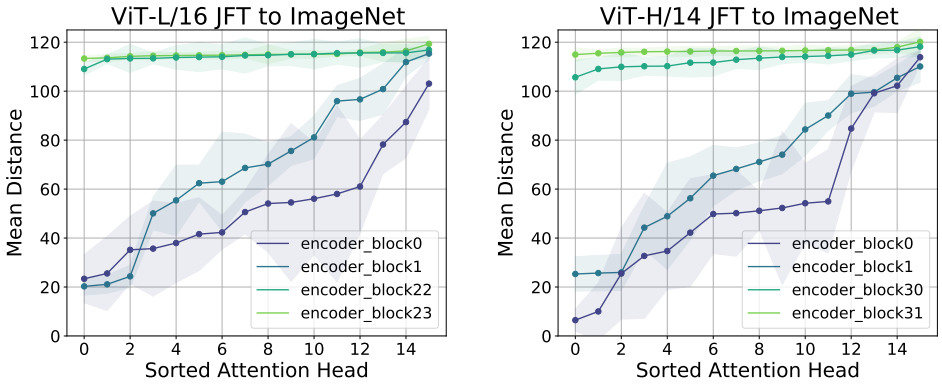


Figure 3: Plotting attention head mean distances shows lower ViT layers attend both locally and globally, while higher layers primarily incorporate global information. For each attention head, we compute the pixel distance it attends to, weighted by the attention weights, and then average over 5000 datapoints to get an average attention head distance. We plot the heads sorted by their average attention distance for the two lowest and two highest layers in the ViT, observing that the lower layers attend both locally and globally, while the higher layers attend entirely globally.

5 Local and Global Information in Layer Representations

In the previous section, we observed much greater similarity between lower and higher layers in ViT, and we also saw that ResNet required more lower layers to compute similar representations to a smaller set of ViT lower layers. In this section, we explore one possible reason for this difference: the difference in the ability to incorporate global information between the two models. How much global information is aggregated by early self-attention layers in ViT? Are there noticeable resulting differences to the features of CNNs, which have fixed, local receptive fields in early layers? In studying these questions, we demonstrate the influence of global representations and a surprising connection between scale and self-attention distances.

Analyzing Attention Distances: We start by analyzing ViT self-attention layers, which are the mechanism for ViT to aggregate information from other spatial locations, and structurally very different to the fixed receptive field sizes of CNNs. Each self-attention layer comprises multiple self-attention heads, and for each head we can compute the average distance between the query patch position and the locations it attends to. This reveals how much local vs global information each self-attention layer is aggregating for the representation. Specifically, we weight the pixel distances by the attention weights for each attention head and average over 5000 datapoints, with results shown in Figure 3. In agreement with Dosovitskiy et al. [14], we observe that even in the lowest layers of ViT, self-attention layers have a mix of local heads (small distances) and global

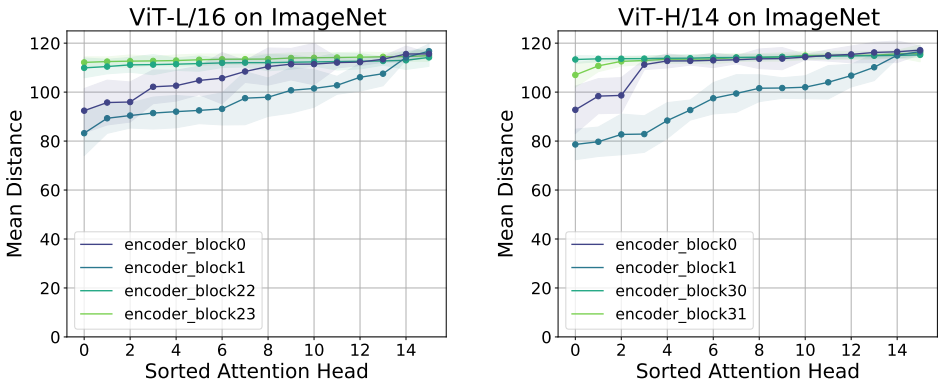


Figure 4: With less training data, lower attention layers do not learn to attend locally. Comparing the results to Figure 3, we see that training only on ImageNet leads to the lower layers not learning to attend more locally. These models also perform much worse when only trained on ImageNet, suggesting that incorporating local features (which is hardcoded into CNNs) may be important for strong performance. (See also Figure C.5.)

heads (large distances). This is in contrast to CNNs, which are hardcoded to attend only locally in the lower layers. At higher layers, all self-attention heads are global.

Interestingly, we see a clear effect of scale on attention. In Figure 4, we look at attention distances when training *only* on ImageNet (no large-scale pre-training), which leads to much lower performance in ViT-L/16 and ViT-H/14 [14]. Comparing to Figure 3, we see that with not enough data, ViT *does not learn to attend locally* in earlier layers. Together, this suggests that using local information early on for image tasks (which is hardcoded into CNN architectures) is important for strong performance.

Does access to global information result in different features? The results of Figure 3 demonstrate that ViTs have access to more global information than CNNs in their lower layers. But does this result in different learned features? As an interventional test, we take subsets of the ViT attention heads from the first encoder block, ranging from the subset corresponding to the most local attention heads to a subset of the representation corresponding to the most global attention heads. We then compute CKA similarity between these subsets and the lower layer representations of ResNet.

The results, shown in Figure 5, which plot the mean distance for each subset against CKA similarity, clearly show a monotonic decrease in similarity as mean attention distance grows, demonstrating that access to more global information also leads to quantitatively different features than computed by the local receptive fields in the lower layers of the ResNet.

Effective Receptive Fields: We conclude by computing *effective receptive fields* [23] for both ResNets and ViTs, with results in Figure 6 and Appendix C. We observe that lower layer effective receptive fields for ViT are indeed larger than in ResNets, and while ResNet effective receptive fields grow gradually, ViT receptive fields become much more global midway through the network. ViT receptive fields also show strong dependence on their center patch due to their strong residual connections, studied in the next section. As we show in Appendix C, in attention sublayers, receptive fields taken before the residual connection show far less dependence on this central patch.

6 Representation Propagation through Skip Connections

The results of the previous section demonstrate that ViTs learn different representations to ResNets in lower layers due to access to global information, which explains some of the differences in representation structure

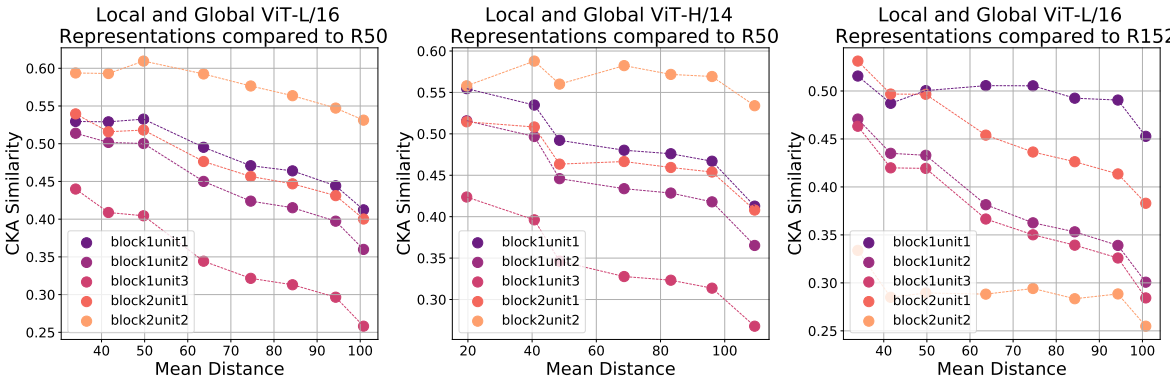


Figure 5: Lower layer representations of ResNet are most similar to representations corresponding to local attention heads of ViT. We take subsets of ViT attention heads in the first encoder block, ranging from the most locally attending heads (smallest mean distance) to the most global heads (largest mean distance). We then compute CKA similarity between these subsets and lower layer representations in the ResNet. We observe that lower ResNet layers are most similar to the features learned by local attention heads of ViT, and decrease monotonically in similarity as more global information is incorporated, demonstrating that the global heads learn quantitatively different features.

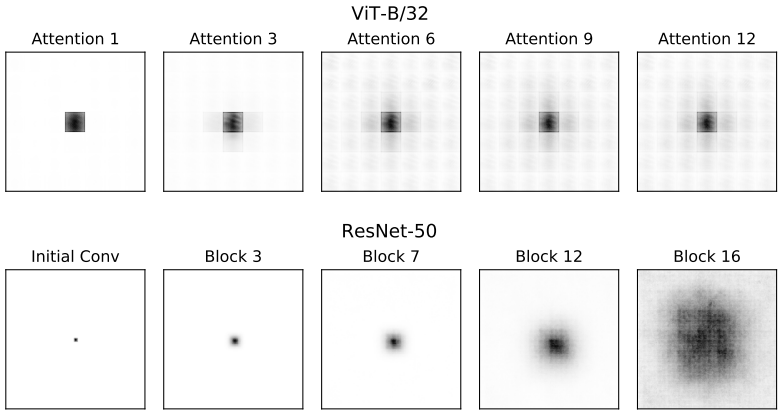


Figure 6: ResNet effective receptive fields are highly local and grow gradually; ViT effective receptive fields shift from local to global. We measure the effective receptive field of different layers as the absolute value of the gradient of the center location of the feature map (taken after residual connections) with respect to the input. Results are averaged across all channels in each map for 32 randomly-selected images.

observed in Section 4. However, the highly uniform nature of ViT representations (Figure 1) also suggests lower representations are faithfully propagated to higher layers. But how does this happen? In this section, we explore the role of skip connections in representation propagation across ViTs and ResNets, discovering ViT skip connections are highly influential, with a clear phase transition from preserving the CLS (class) token representation (in lower layers) to spatial token representations (in higher layers).

Like Transformers, ViTs contain *skip* (aka *identity* or *shortcut*) connections throughout, which are added on after the (i) self-attention layer, and (ii) MLP layer. To study their effect, we plot the norm ratio $||z_i||/||f(z_i)||$ where z_i is the hidden representation of the ith layer coming from the skip connection, and $f(z_i)$ is the transformation of z_i from the *long branch* (i.e. MLP or self-attention.)

The results are in Figure 7 (with additional cosine similarity analysis in Figure E.2.) The heatmap on the left shows $||z_i||/||f(z_i)||$ for different token representations. We observe a striking phase transition: in the first half of the network, the CLS token (token 0) representation is primarily propagated by the skip connection branch (high norm ratio), while the spatial token representations have a large contribution coming from the long branch (lower norm ratio). Strikingly, in the second half of the network, this is reversed.

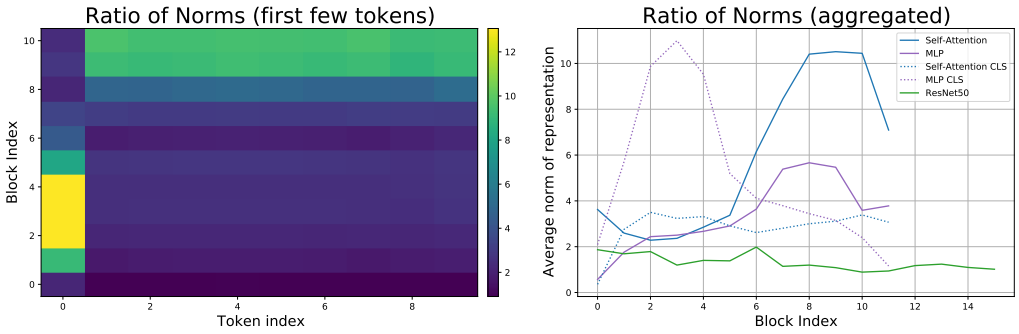


Figure 7: Most information in ViT passes through skip connections. Comparison of representation norms between the skip-connection (identity) and the long branch for ViT-B/16 trained on ImageNet and a ResNet. For ViT, we show the CLS token separately from the rest of the representation. (left) shows the ratios separated for the first few tokens (token 0 is CLS), (right) shows averages over all tokens.

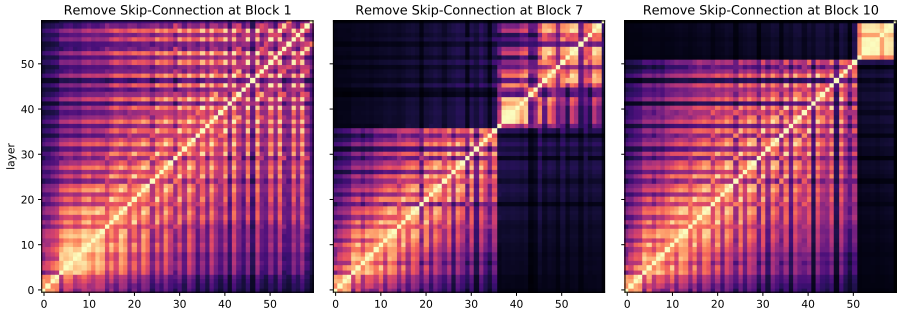


Figure 8: ViT models trained without any skip connections in block i show very little representation similarity between layers before/after block i. We train several ViT models without any skip connections at block i for varying i to interventionally test the effect on representation structure. For middle blocks without skip connections, we observe a performance drop of 4%. We also observe that removing a skip connection at block i partitions similar representations to before/after block i — this demonstrates the importance of skip connections in ViT's standard uniform representation structure.

The right pane, which has line plots of these norm ratios across ResNet50, the ViT CLS token and the ViT spatial tokens additionally demonstrates that skip connection is much more influential in ViT compared to ResNet: we observe much higher norm ratios for ViT throughout, along with the phase transition from CLS to spatial token propagation (shown for the MLP and self-attention layers.)

ViT Representation Structure without Skip Connections: The norm ratio results strongly suggest that skip connections play a key role in the representational structure of ViT. To test this interventionally, we train ViT models with skip connections removed in block i for varying i, and plot the CKA representation heatmap. The results, in Figure 8, illustrate that removing the skip connections in a block partitions the layer representations on either side. (We note a performance drop of 4% when removing skip connections from middle blocks.) This demonstrates the importance of representations being propagated by skip connections for the uniform similarity structure of ViT in Figure 1.

7 Spatial Information and Localization

The results so far, on the role of self-attention in aggregating spatial information in ViTs, and skip-connections faithfully propagating representations to higher layers, suggest an important followup question: how well can ViTs perform *spatial localization*? Specifically, is spatial information from the input preserved in the higher

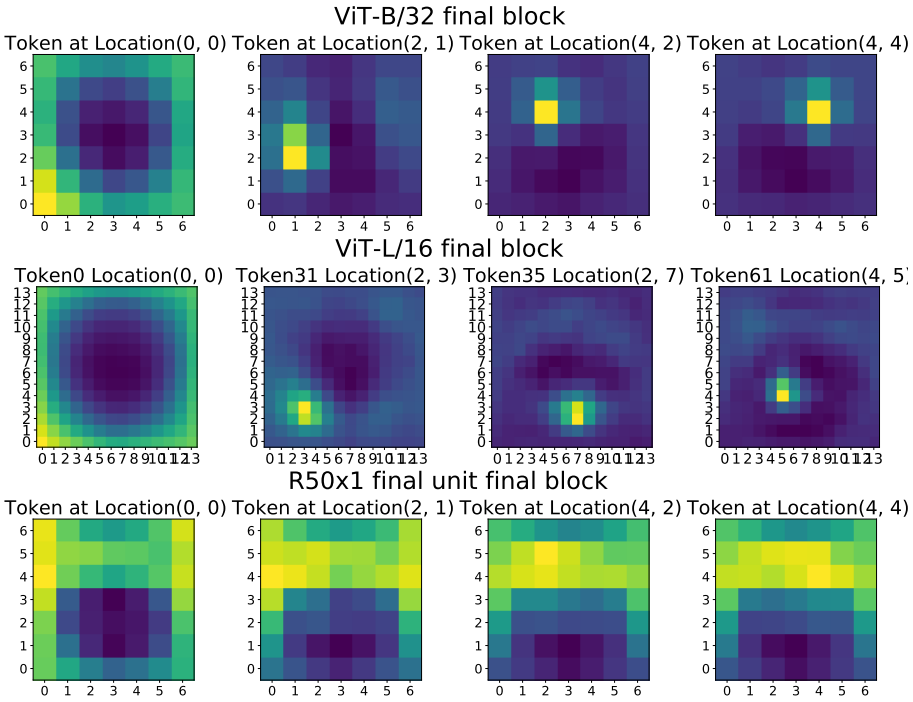


Figure 9: Higher layers of ViT maintain spatial location information more faithfully than ResNets. Each heatmap plot shows the CKA similarity between a single token representation in final block of the model and the input images, which are divided into non-overlapping patches. We observe that ViT tokens have strongest similarity to their corresponding spatial location in the image, but tokens corresponding to spatial locations at the edge of the image (e.g. token 0) additionally show similarity to other edge positions. This demonstrates that spatial information from the input is preserved even at the final layer of ViT. By contrast, ResNet "tokens" (features at a specific spatial location) are much less spatially discriminative, showing comparable similarity across a broad set of input spatial locations. See Appendix for additional layers and results.

layers of ViT? And how does it compare in this aspect to ResNet? An affirmative answer to this is crucial for uses of ViT beyond classification, such as object detection.

We begin by comparing token representations in the higher layers of ViT and ResNet to those of input patches. Recall that ViT tokens have a corresponding input patch, and thus a corresponding input spatial location. For ResNet, we define a token representation to be all the convolutional channels at a particular spatial location. This also gives it a corresponding input spatial location. We can then take a token representation and compute its CKA score with input image patches at different locations. The results are illustrated for different tokens (with their spatial locations labelled) in Figure 9.

For ViT, we observe that tokens corresponding to locations at the edge of the image are similar to edge image patches, but tokens corresponding to interior locations are well localized, with their representations being most similar to the corresponding image patch. By contrast, for ResNet, we see significantly weaker localization (though Figure D.3 shows improvements for earlier layers.)

One factor influencing this clear difference between architectures is that ResNet is trained to classify with a global average pooling step, while ViT has a separate classification (CLS) token. To examine this further, we test a ViT architecture trained with global average pooling (GAP) for localization (see Appendix A for training details). The results, shown in Figure 10, demonstrate that global average pooling does indeed reduce localization in the higher layers. More results in Appendix Section D.

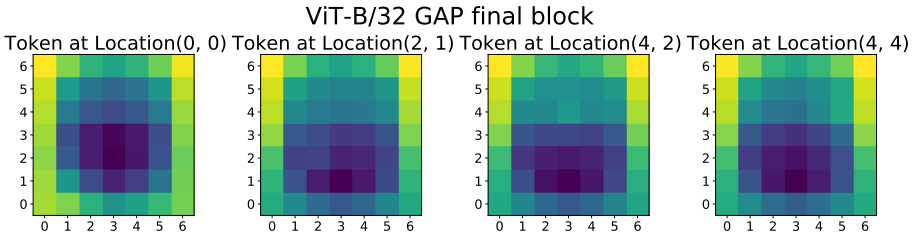
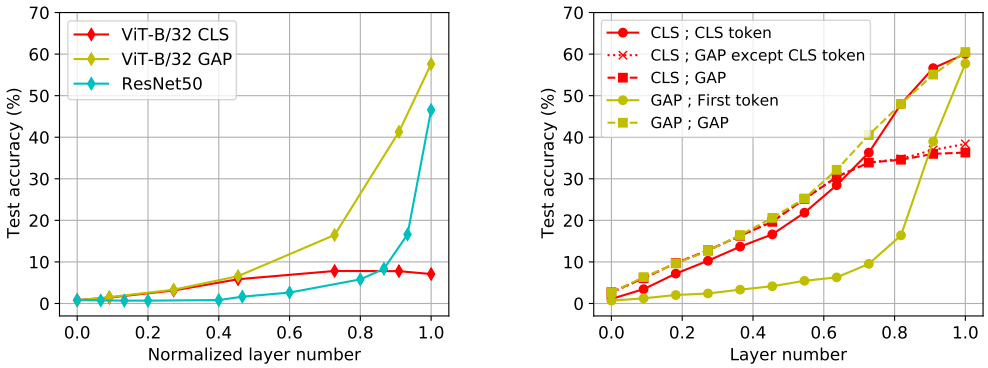


Figure 10: When trained with global average pooling (GAP) instead of a CLS token, ViTs show less clear localization (compare Figure 9). We plot the same CKA heatmap between a token and different input images patches as in Figure 9, but for a ViT model trained with global average pooling (like ResNet) instead of a CLS token. We observe significantly less localization.



(a) Individual token evaluation

(b) CLS vs GAP models

Figure 11: Spatial localization experiments with linear probes. We train linear classifiers on 10-shot ImageNet classification from the representations extracted from different layers of ViT-B/32 models. We then plot the accuracy of the probe versus the (normalized) layer number. **Left:** We train a classifier on each token separately and report the average accuracy over all tokens (excluding the CLS token for the ViT CLS model.) **Right:** Comparison of ViT models pre-trained with a classification token or with global average pooling (GAP) and then evaluated with different ways of aggregating the token representations.

Localization and Linear Probe Classification: The previous results have looked at localization through direct comparison of each token with input patches. To complete the picture, we look at using each token separately to perform *classification* with linear probes. We do this across different layers of the model, training linear probes to classify image label with closed-form few-shot linear regression similar to Dosovitskiy et al. [14] (details in Appendix A). Results are in Figure 11, with further results in Appendix F. The left pane shows average accuracy of classifiers trained on individual tokens, where we see that ResNet50 and ViT with GAP model tokens perform well at higher layers, while in the standard ViT trained with a CLS token the spatial tokens do poorly – likely because their representations remain spatially localized at higher layers, which makes global classification challenging. Supporting this are results on the right pane, which shows that a single token from the ViT-GAP model achieves comparable accuracy in the highest layer to all tokens pooled together. With the results of Figure 9, this suggests all higher layer tokens in GAP models learn similar (global) representations.

8 Effects of Scale on Transfer Learning

Motivated by the results of Dosovitskiy et al. [14] that demonstrate the importance of dataset scale for high performing ViTs, and our earlier result (Figure 4) on needing scale for local attention, we perform a study of the effect of dataset scale on representations in transfer learning.

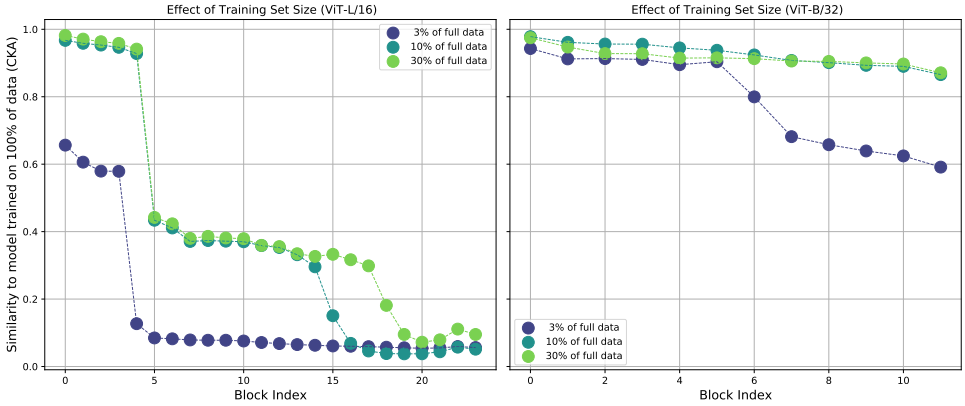
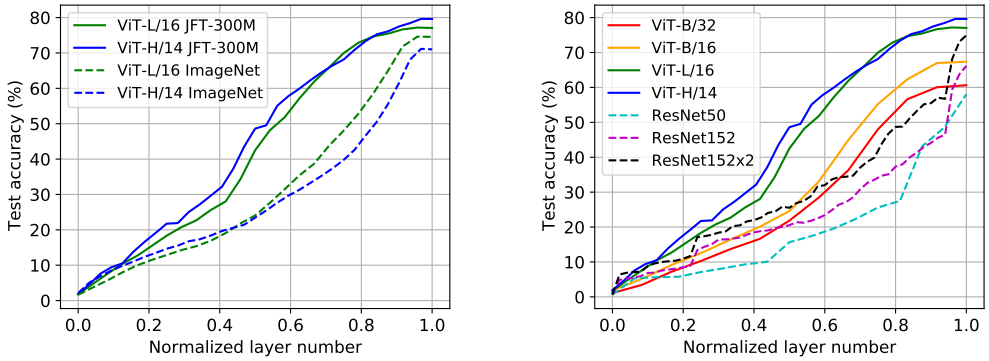


Figure 12: Measuring similarity of representations learned with varying amounts of data shows the importance of large datasets for higher layers and larger model representations. We compute the similarity of representations at each block for ViT models that have been trained on smaller subsets of the data to a model that has been trained on the full data on ViT-L/16 (left) and ViT-B/32 (right). We observe that while lower layer representations have high similarity even with 10% of the data, higher layers and larger models require significantly more data to learn similar representations.



(a) JFT-300M vs ImageNet pre-training

(b) ViTs vs ResNets

Figure 13: Experiments with linear probes. We train linear classifiers on 10-shot ImageNet classification from the aggregated representations of different layers of different models. We then plot the accuracy of the probe versus the (normalized) layer number. **Left:** Comparison of ViTs pre-trained on JFT-300M or ImageNet and evaluated with linear probes on Imagenet. **Right:** Comparison of ViT and ResNet models trained JFT-300m, evaluated with linear probes on ImageNet.

We begin by studying the effect on representations as the JFT-300M pretraining dataset size is varied. Figure 12 illustrates the results on ViT-B/32 and ViT-L/16. Even with 3% of the entire dataset, lower layer representations are very similar to the model trained on the whole dataset, but higher layers require larger amounts of pretraining data to learn the same representations as at large data scale, especially with the large model size. In Section G, we study how much representations change in finetuning, finding heterogeneity over datasets.

We next look at dataset size effect on the larger ViT-L/16 and ViT-H/14 models. Specifically, in the left pane of Figure 13, we train linear classifer probes on ImageNet classes for models pretrained on JFT-300M vs models only pretrained on ImageNet. We observe the JFT-300M pretained models achieve much higher accuracies even with middle layer representations, with a 30% gap in absolute accuracy to the models pretrained only on ImageNet. This suggests that for larger models, the larger dataset is especially helpful in learning high quality intermediate representations. This conclusion is further supported by the results of the right pane of Figure 13,

which shows linear probes on different ResNet and ViT models, all pretrained on JFT-300M. We again see the larger ViT models learn much stronger intermediate representations than the ResNets. Additional linear probes experiments in Section F demonstrate this same conclusion for transfer to CIFAR-10 and CIFAR-100.

9 Discussion

Given the central role of convolutional neural networks in computer vision breakthroughs, it is remarkable that Transformer architectures (almost *identical* to those used in language) are capable of similar performance. This raises fundamental questions on whether these architectures work in the same way as CNNs. Drawing on representational similarity techniques, we find surprisingly clear differences in the features and internal structures of ViTs and CNNs. An analysis of self-attention and the strength of skip connections demonstrates the role of earlier global features and strong representation propagation in ViTs for these differences, while also revealing that some CNN properties, e.g. local information aggregation at lower layers, are important to ViTs, being learned from scratch at scale. We examine the potential for ViTs to be used beyond classification through a study of spatial localization, discovering ViTs with CLS tokens show strong preservation of spatial information — promising for future uses in object detection. Finally, we investigate the effect of scale for transfer learning, finding larger ViT models develop significantly stronger intermediate representations through larger pretraining datasets. These results are also very pertinent to understanding MLP-based architectures for vision proposed by concurrent work [42, 43], further discussed in Section H, and together answer central questions on differences between ViTs and CNNs, and suggest new directions for future study.

References

[1] G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

[2] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3286–3295, 2019.

[3] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, and A. Veit. Understanding robustness of transformers for image classification. *arXiv preprint arXiv:2103.14586*, 2021.

[4] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[5] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

[6] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020.

[7] X. Chen, S. Xie, and K. He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

[8] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *ACL*, 2018.

[9] J.-B. Cordonnier, A. Loukas, and M. Jaggi. On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*, 2019.

[10] C. Cortes, M. Mohri, and A. Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13(1):795–828, 2012.

[11] S. d'Ascoli, H. Touvron, M. Leavitt, A. Morcos, G. Biroli, and L. Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. *arXiv preprint arXiv:2103.10697*, 2021.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[15] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, A. J. Smola, et al. A kernel statistical test of independence. In *Nips*, volume 20, pages 585–592. Citeseer, 2007.

[16] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby. Big transfer (bit): General visual representation learning. *arXiv preprint arXiv:1912.11370*, 6(2):8, 2019.

[17] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of neural network representations revisited. In *ICML*, 2019.

[18] S. Kornblith, H. Lee, T. Chen, and M. Norouzi. What's in a loss function for image classification? *arXiv preprint arXiv:2010.16402*, 2020.

[19] N. Kriegeskorte, M. Mur, and P. A. Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 2008.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[21] S. R. Kudugunta, A. Bapna, I. Caswell, N. Arivazhagan, and O. Firat. Investigating multilingual nmt representations at scale. *arXiv preprint arXiv:1909.02197*, 2019.

[22] G. W. Lindsay. Convolutional neural networks as a model of the visual system: past, present, and future. *Journal of cognitive neuroscience*, pages 1–15, 2020.

[23] W. Luo, Y. Li, R. Urtasun, and R. Zemel. Understanding the effective receptive field in deep convolutional neural networks. *arXiv preprint arXiv:1701.04128*, 2017.

[24] N. Maheswaranathan, A. H. Williams, M. D. Golub, S. Ganguli, and D. Sussillo. Universality and individuality in neural dynamics across large populations of recurrent networks. *Advances in neural information processing systems*, 2019:15629, 2019.

[25] A. Merchant, E. Rahimtoroghi, E. Pavlick, and I. Tenney. What happens to bert embeddings during fine-tuning? *arXiv preprint arXiv:2004.14448*, 2020.

[26] A. S. Morcos, M. Raghu, and S. Bengio. Insights on representational similarity in neural networks with canonical correlation. *arXiv preprint arXiv:1806.05759*, 2018.

[27] B. Mustafa, A. Loh, J. Freyberg, P. MacWilliams, M. Wilson, S. M. McKinney, M. Sieniek, J. Winkens, Y. Liu, P. Bui, et al. Supervised transfer learning at scale for medical imaging. *arXiv preprint arXiv:2101.05913*, 2021.

[28] M. Naseer, K. Ranasinghe, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang. Intriguing properties of vision transformers, 2021.

[29] T. Nguyen, M. Raghu, and S. Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. *arXiv preprint arXiv:2010.15327*, 2020.

[30] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018.

[31] S. Paul and P.-Y. Chen. Vision transformers are robust learners. *arXiv preprint arXiv:2105.07581*, 2021.

[32] M. E. Peters, M. Neumann, L. Zettlemoyer, and W.-t. Yih. Dissecting contextual word embeddings: Architecture and representation. In *EMNLP*, 2018.

[33] A. Raghu, M. Raghu, S. Bengio, and O. Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*, 2019.

[34] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *arXiv preprint arXiv:1706.05806*, 2017.

[35] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. Transfusion: Understanding transfer learning for medical imaging. *arXiv preprint arXiv:1902.07208*, 2019.

[36] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019.

[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[38] J. Shi, E. Shea-Brown, and M. Buice. Comparison against task driven artificial neural networks reveals functional properties in mouse visual cortex. *Advances in Neural Information Processing Systems*, 32: 5764–5774, 2019.

[39] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(5), 2012.

[40] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[41] Y. Tay, M. Dehghani, J. Gupta, D. Bahri, V. Aribandi, Z. Qin, and D. Metzler. Are pre-trained convolutions better than pre-trained transformers? *arXiv preprint arXiv:2105.03322*, 2021.

[42] I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, D. Keysers, J. Uszkoreit, M. Lucic, et al. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021.

[43] H. Touvron, P. Bojanowski, M. Caron, M. Cord, A. El-Nouby, E. Grave, A. Joulin, G. Synnaeve, J. Verbeek, and H. Jégou. Resmlp: Feedforward networks for image classification with data-efficient training. *arXiv preprint arXiv:2105.03404*, 2021.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[45] E. Voita, R. Sennrich, and I. Titov. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. In *EMNLP*, 2019.

[46] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020.

[47] J. M. Wu, Y. Belinkov, H. Sajjad, N. Durrani, F. Dalvi, and J. Glass. Similarity analysis of contextual word representation models. *arXiv preprint arXiv:2005.01172*, 2020.

[48] S. Wu, A. Conneau, H. Li, L. Zettlemoyer, and V. Stoyanov. Emerging cross-lingual structure in pretrained language models. *arXiv preprint arXiv:1911.01464*, 2019.

[49] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. Tay, J. Feng, and S. Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021.

[50] X. Zhai, J. Puigcerver, A. Kolesnikov, P. Ruyssen, C. Riquelme, M. Lucic, J. Djolonga, A. S. Pinto, M. Neumann, A. Dosovitskiy, et al. The visual task adaptation benchmark. 2019.

Appendix

Additional details and results from the different sections are included below.

A Additional details on Methods and the Experimental Setup

To understand systematic differences between ViT and CNNs, we use a representative set of different models of each type, guided by the performance results in [14]. Specifically, for ViTs, we look at ViT-B/32, ViT-L/16 and ViT-H/14, where the smallest model (ViT-B/32) shows limited improvements when pretraining on JFT-300M [40] vs. the ImageNet Large Scale Visual Recognition Challenge 2012 dataset [12, 37], while the largest, ViT-H/14, achieves state of the art when pretrained on JFT-300M [40]. ViT-L/16 is close to the performance ViT-H/14 [14]. For CNNs, we look at ResNet50x1 which also shows saturating performance when pretraining on JFT-300M, and also ResNet152x2, which in contrast shows large performance gains with increased pretraining dataset size. As in Dosovitskiy et al. [14], these ResNets follow some of the implementation changes first proposed in BiT [16].

In addition to the standard Vision Transformers trained with a classification token (CLS), we also trained ViTs with global average pooling (GAP). In these, there is no classification token – instead, the representations of tokens in the last layer of the transformer are averaged and directly used to predict the logits. The GAP model is trained with the same hyperparameters as the CLS one, except for the initial learning rate that is set to a lower value of 0.0003.

For analyses of internal model representations, we observed no meaningful difference between representations of images drawn from ImageNet and images drawn from JFT-300M. Figures 1, 2, 9, and 10 use images from the JFT-300M dataset that were not seen during training, while Figures 6, 3, 4, 5 7, 8, and 12 use images from the ImageNet 2012 validation set. Figures 11 and 13 involve 10-shot probes trained on the ImageNet 2012 training set, tuned hyperparameters on a heldout portion of the training set, and evaluated on the validation set.

Additional details on CKA implementation To compute CKA similarity scores, we use minibatch CKA, introduced in [29]. Specifically, we use batch sizes of 1024 and we sample a total of 10240 examples without replacement. We repeat this 20 times and take the average. Experiments varying the exact batch size (down to 128), and total number of examples used for CKA (down to 2560 total examples, repeated 10 times), had no noticeable effect on the results.

Additional details on linear probes. We train linear probes as regularized least-squares regression, following Dosovitskiy et al. [14]. We map the representations training images to $\{-1, 1\}^N$ target vectors, where N is the number of classes. The solution can be recovered efficiently in closed form.

For vision transformers, we train linear probes on representations from individual tokens or on the representation averaged over all tokens, at the output of different transformer layers (each layer meaning a full transformer block including self-attention and MLP). For ResNets, we take representation at the output of each residual block (including 3 convolutional layers). The resolution of the feature maps changes throughout the model, so we perform an additional pooling step bringing the feature map to the same spatial size as in the last stage. Moreover, ResNets differ from ViTs in that the number of channels changes throughout the model, with fewer channels in the earlier layers. This smaller channel count in the earlier layers could potentially lead to worse performance of the linear probes. To compensate for this, before pooling we split the feature map into patches and flattened each patch, so as to arrive at the channel count close to the channel count in the final block. All results presented in the paper include this additional patching step; however, we have found that it brings only a very minor improvement on top of simple pooling.

B Additional Representation Structure Results

Here we include some more CKA heatmaps, which provide insights on model representation structures (compare to Figure 1, Figure 2 in the main text.) We observe similar conclusions: ViT representation structure has a more uniform similarity structure across layers, and comparing ResNet to ViT representations show a large fraction of early ResNet layers similar to a smaller number of ViT layers.

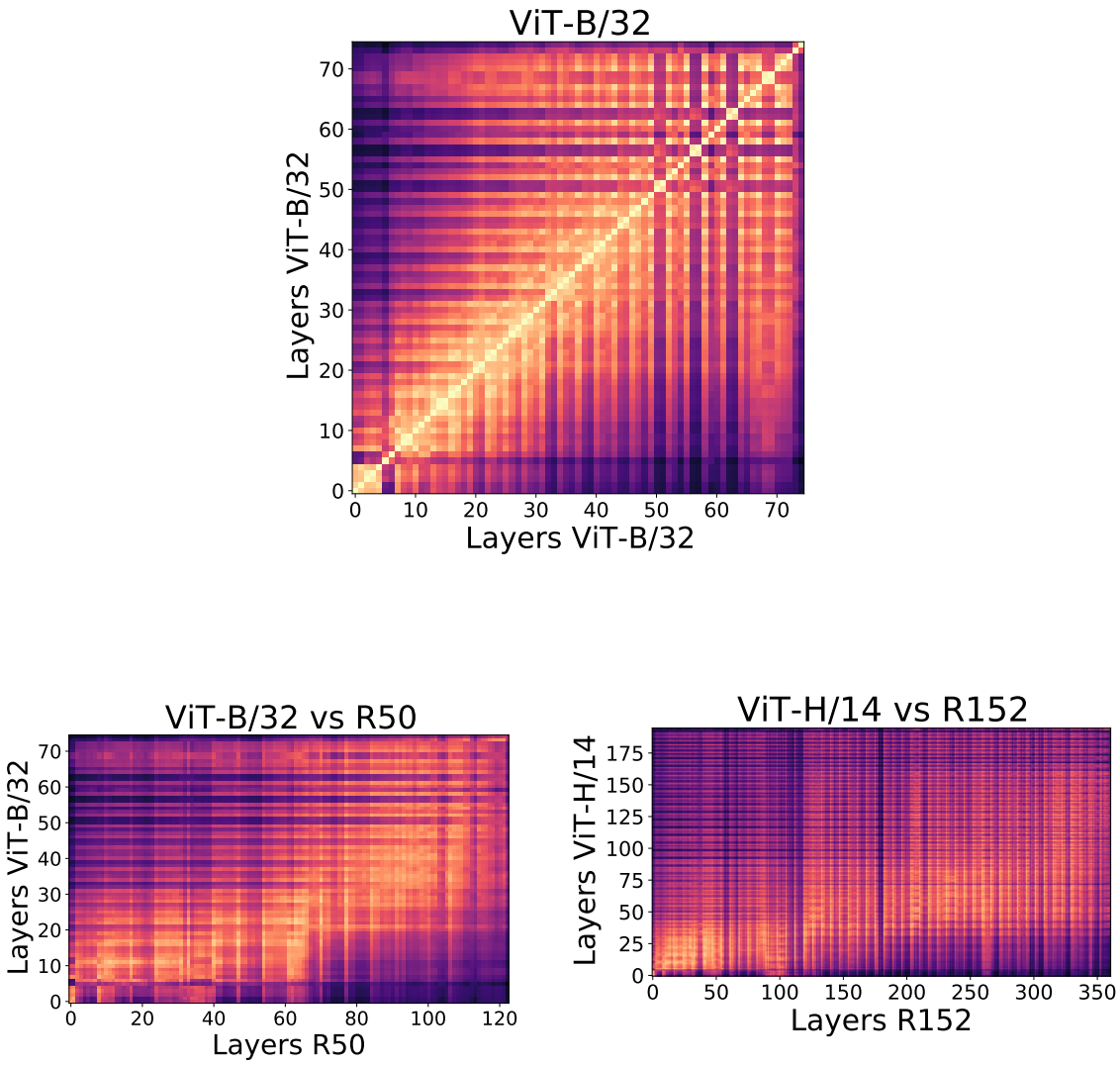


Figure B.1: Additional CKA heatmap results. Top shows CKA heatmap for ViT-B/32, where we can also observe strong similarity between lower and higher layers and the grid like, uniform representation structure. Bottom shows (i) ViT-B/32 compared to R50, where we again see that 60 of the lowest R50 layers are similar to about 25 of the lowest ViT-B/32 layers, with the remaining layers most similar to each other (ii) ViT-H/14 compared to R152, where we see the lowest 100 layers of R152 are most similar to lowest 30 layers of ViT-H/14.

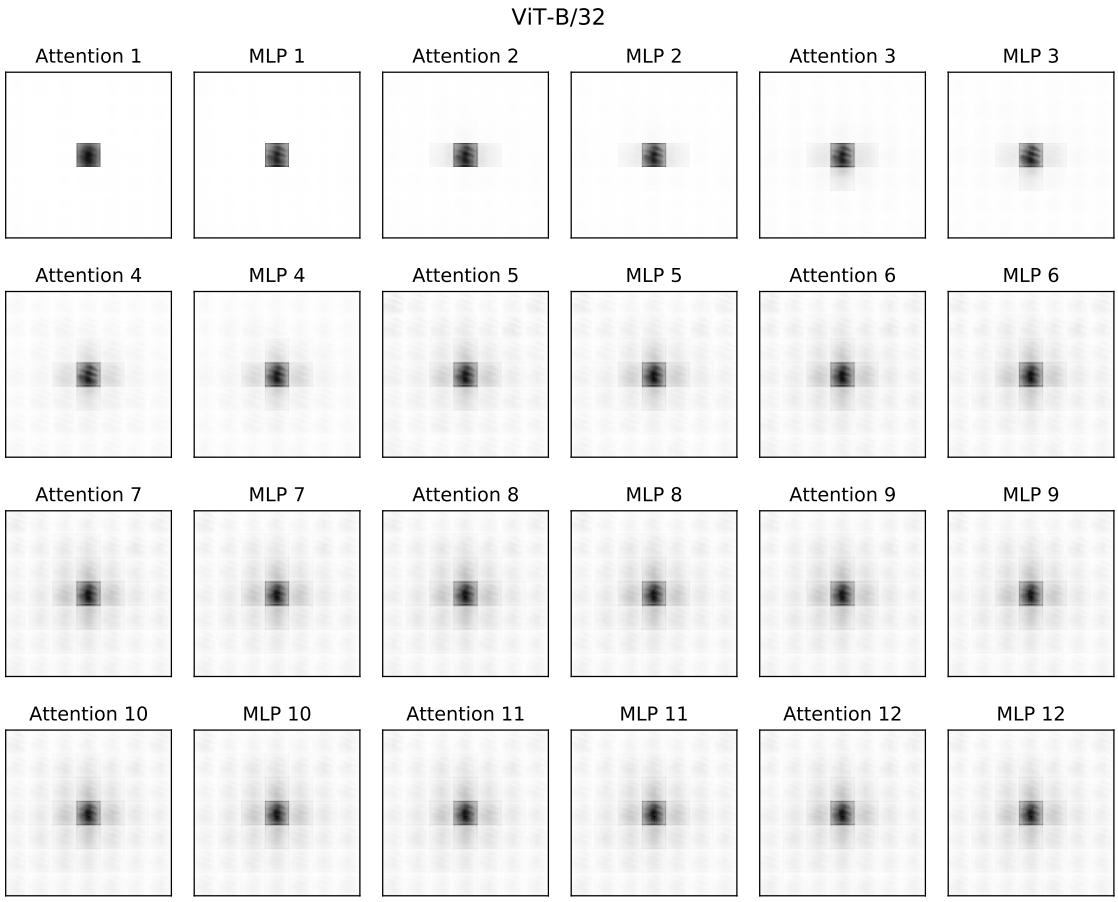


Figure C.1: Post-residual receptive fields of all ViT-B/32 sublayers.

C Additional Local/Global Information Results

In Figures C.1, C.2, and C.3, we provide full plots of effective receptive fields of all layers of ViT-B/32, ResNet-50, and ViT-L/16, taken after the residual connections as in Figure 6 in the text. In Figure C.4, we show receptive fields of ViT-B/32 and ResNet-50 taken *before* the residual connections. Although the pre-residual receptive fields of ViT MLP sublayers resemble the post-residual receptive fields in Figure C.1, the pre-residual receptive fields of attention sublayers have a smaller relative contribution from the corresponding input patch. These results support our findings in Section 5 regarding the global nature of attention heads, but suggest that network representations remain tied to input patch locations because of the strong contributions from skip connections, studied in Section 6. ResNet-50 pre-residual receptive fields look similar to the post-residual receptive fields.

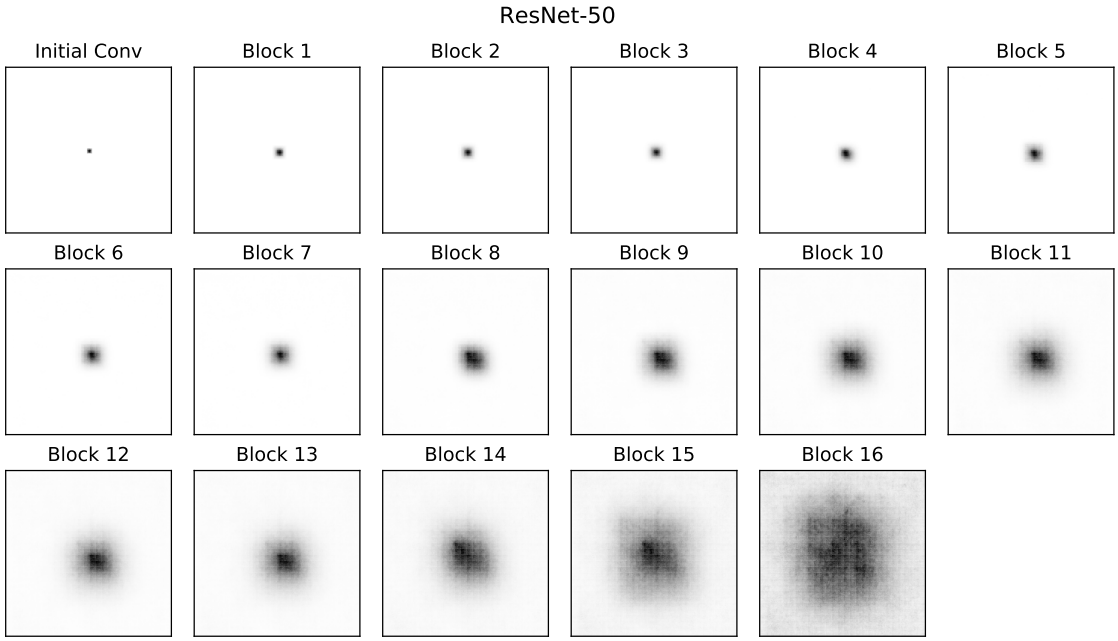


Figure C.2: Post-residual receptive fields of all ResNet-50 blocks.

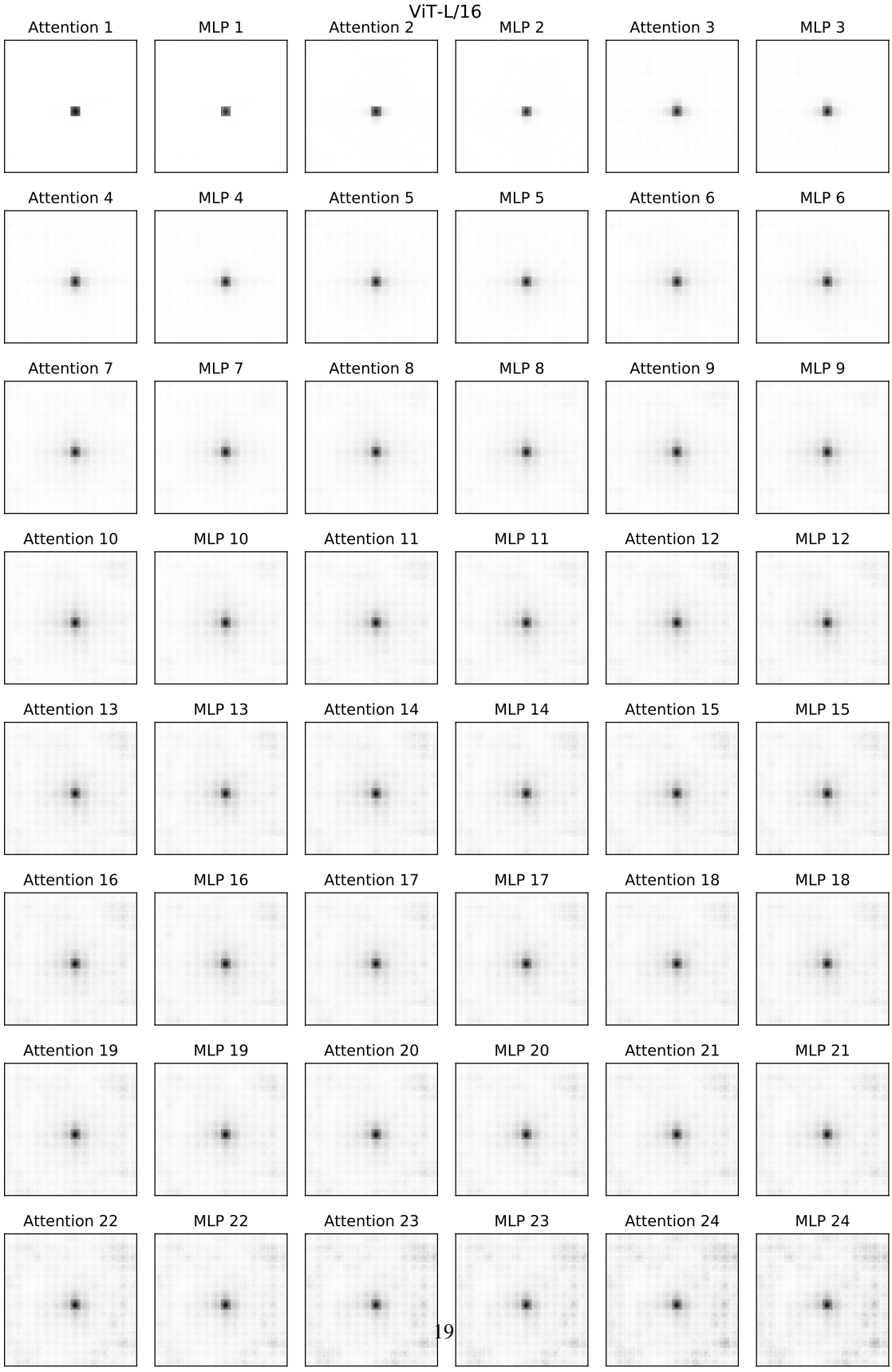


Figure C.3: Post-residual receptive fields of all ViT-L/16 sublayers.

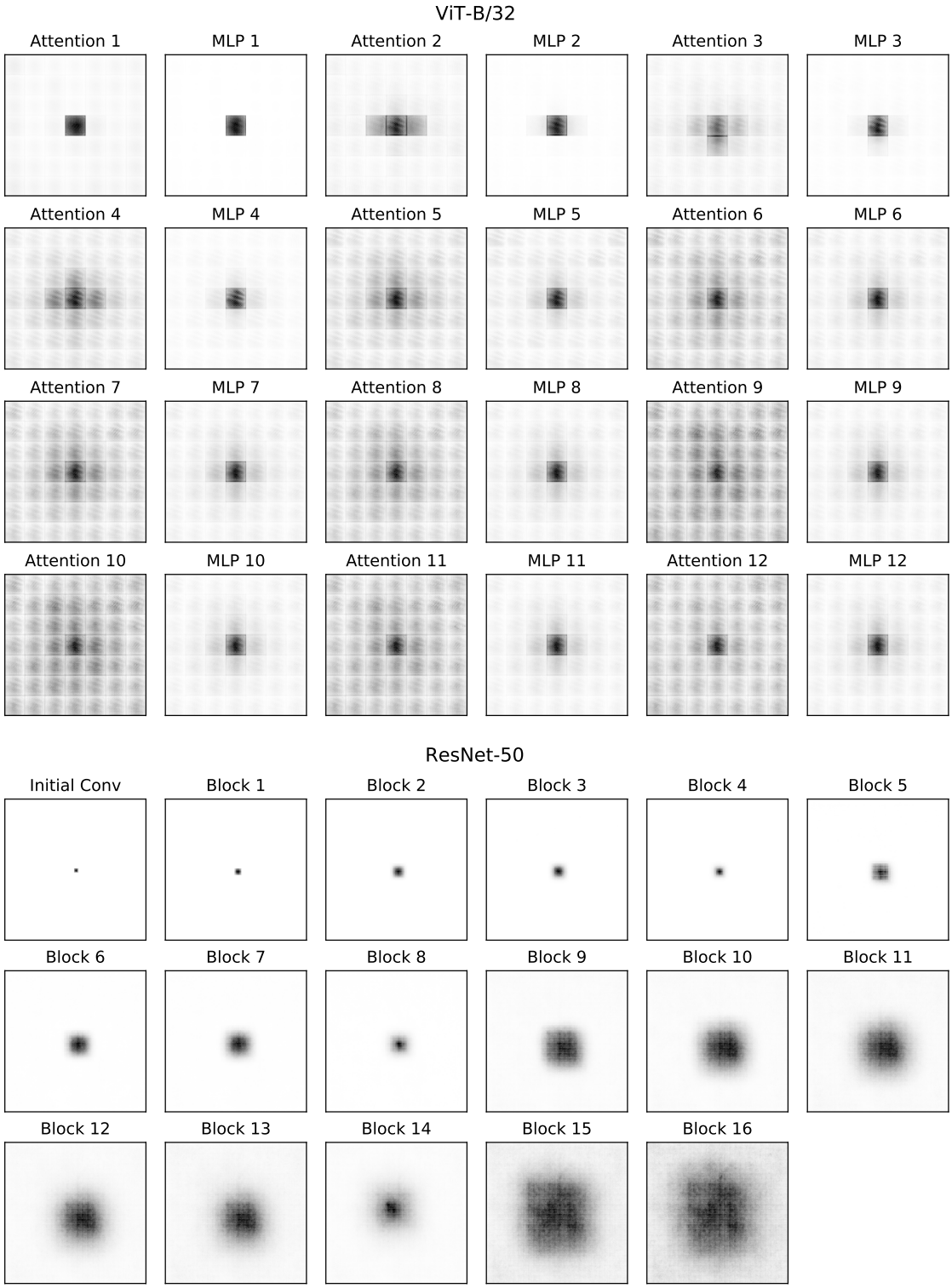


Figure C.4: Pre-residual receptive fields of all ViT-B/32 sublayers and ResNet-50 blocks. In ViT-B, we see that the pre-residual receptive fields of later attention sublayers are not dominated by the center patch, in contrast to the post-residual receptive fields shown in Figure C.1. Thus, although later attention sublayers integrate information across the entire input image, network representations remain localized due to the strong skip connections. ResNet-50 pre-residual receptive fields generally resemble the post-residual receptive fields shown in Figure C.2. The receptive field appears to "shrink" at blocks 4, 8, and 14, which are each the first in a stage, and for which we plot only the longer branch and not the shortcut. The receptive field does not shrink when computed after the summation of these branches, as shown in Figure C.2.

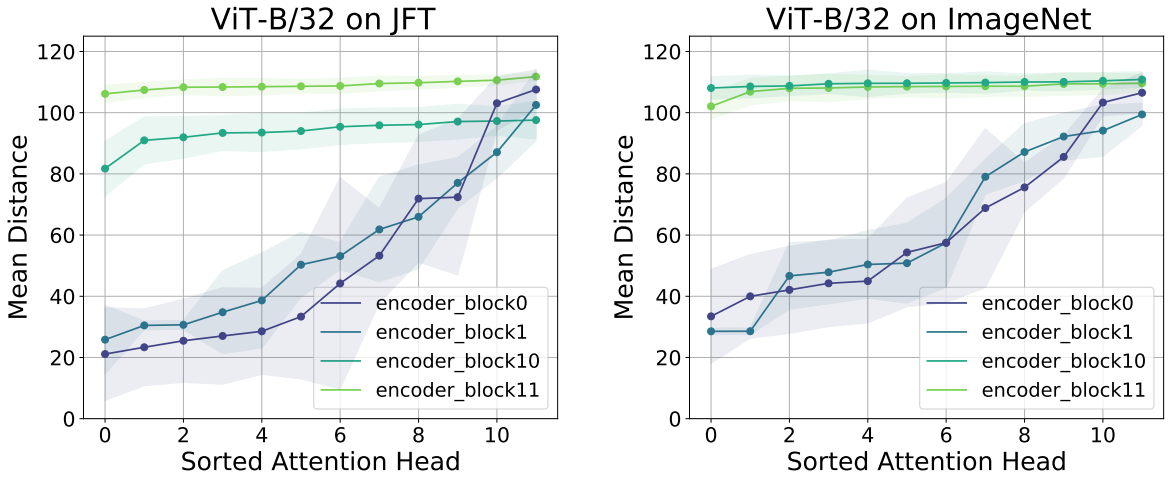


Figure C.5: Plot of attention head distances for ViT-B/32 when trained on JFT-300M and on only ImageNet shows that ViT-B/32 learns to attend locally even on a smaller dataset. Compare to Figure 3 in the main text. While ViT-L and ViT-H show large performance improvements when (i) finetuned on ImageNet having been pretrained on JFT compared to (ii) being trained on only ImageNet, ViT-B/32 has similar performance in both settings. We also observe that ViT-H, ViT-L don't learn to attend locally in the lowest layers when only trained on ImageNet (Figure 3), whereas here we see ViT-B/32 still learns to attend locally — suggesting connections between performance and heads learning to attend locally.

D Localization

Below we include additional localization results: computing CKA between different input patches and tokens in the higher layers of the models. We show results for ViT-H/14, additional higher layers for ViT-L/16, ViT-B/32 and addtional layers for ResNet.

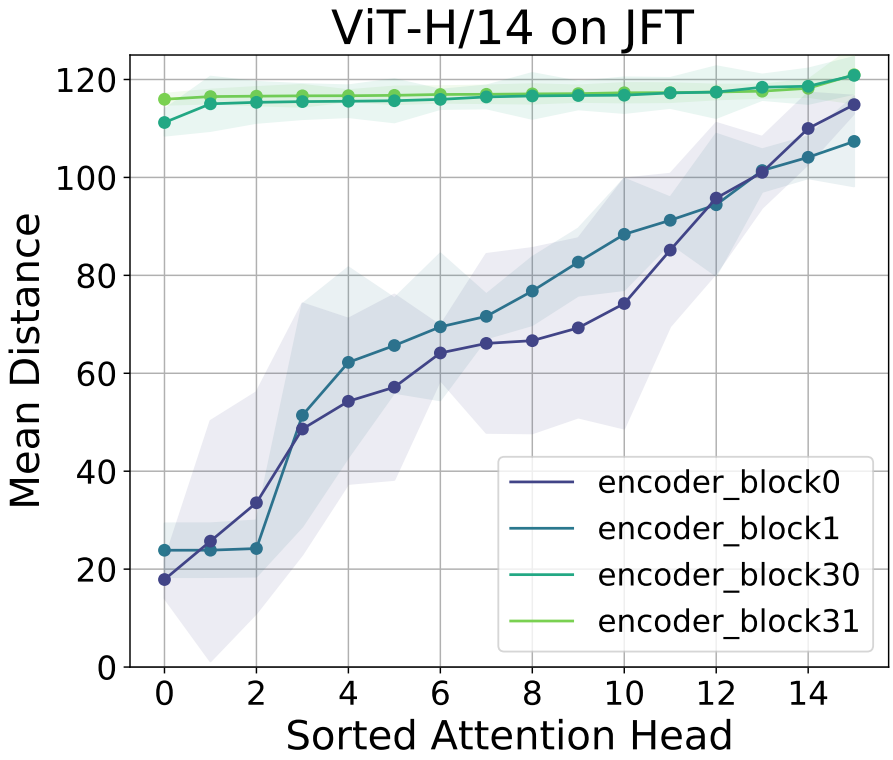


Figure C.6: Additional plot of attention head distances for ViT-H/14 on JFT-300M. Compare to Figure 3 in the main text. For each attention head, we compute the pixel distance it attends to, weighted by the attention weights, and then average over 5000 datapoints to get an average attention head distance. We plot the heads sorted by their average attention distance for the two lowest and two highest layers in the ViT, observing that the lower layers attend both locally and globally, while the higher layers attend entirely globally.

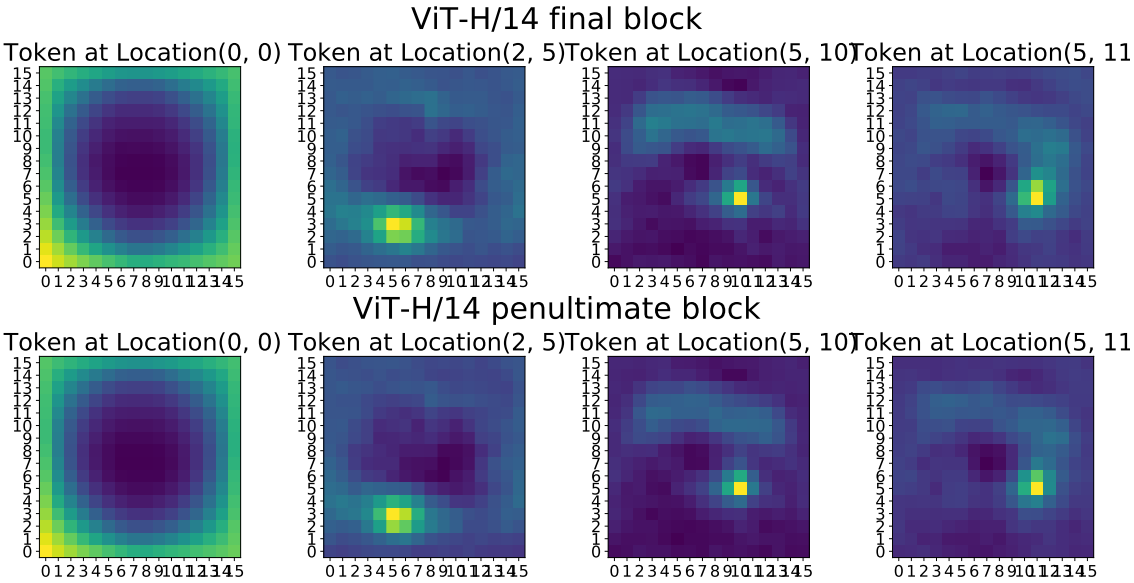


Figure D.1: Localization heatmaps for ViT-H/14. We see that ViT-H/14 is also well localized, both in the final block and penultimate block, with tokens with corresponding locations in the interior of the image most similar to the image patches at those locations, while tokens on the edge are similar to many edge positions.

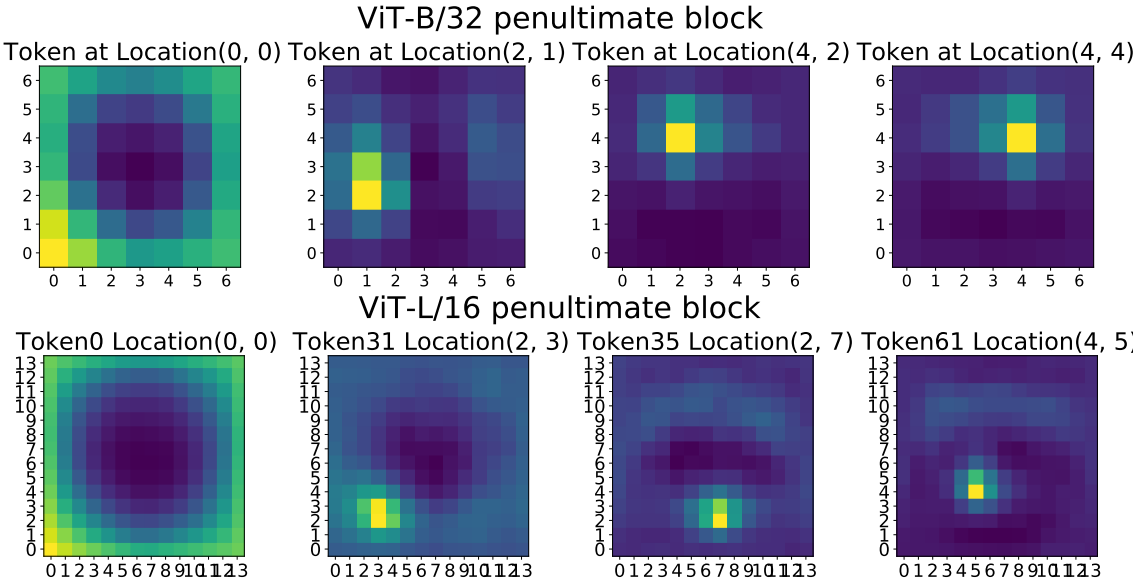


Figure D.2: Additional localization heatmaps for other higher layers of ViT-L/16 and ViT-B/32. We see that models (as expected) remain well localized in higher layers other than the final block.

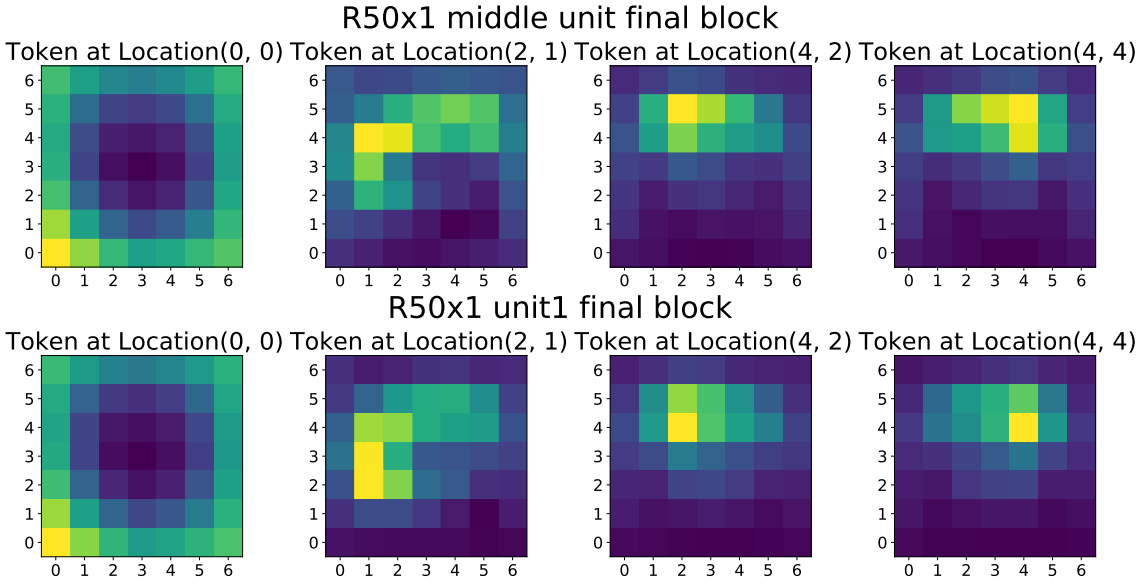


Figure D.3: Localization heatmaps for layers of ResNet below final layer. Comparing to Figure 9 in the main text, we see that layers in the ResNet lower than the final layer display better localization, but still not as clear as the CLS trained ViT models.

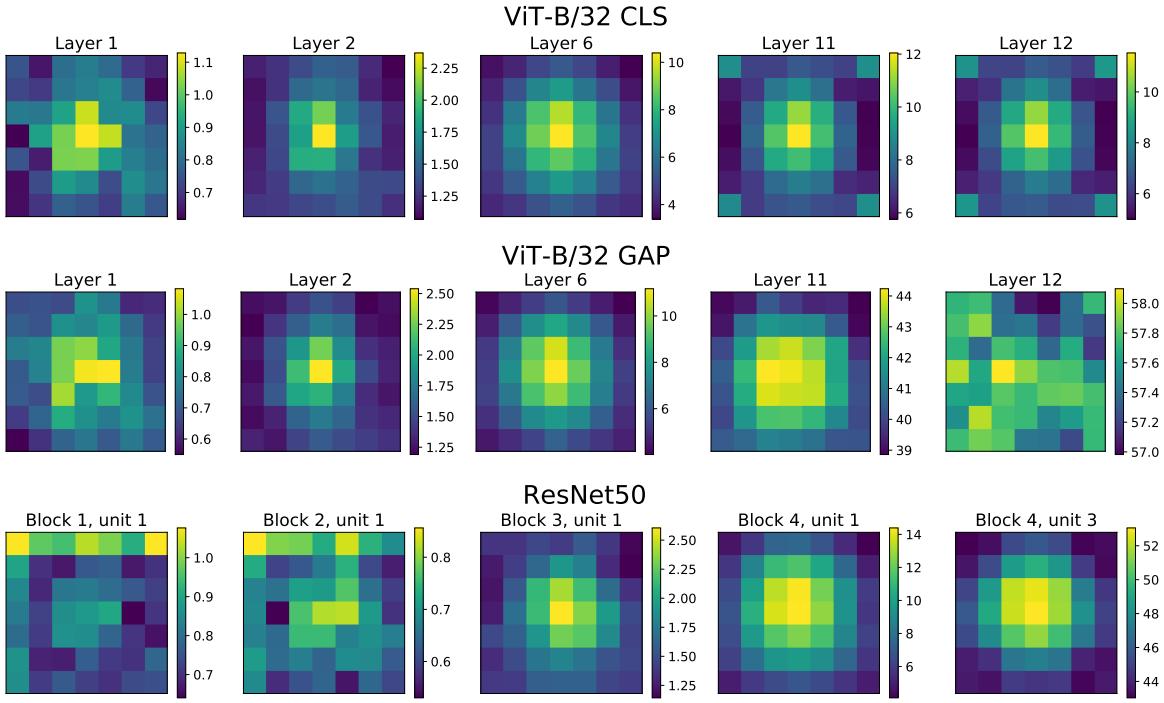


Figure D.4: Linear probes spatial localization. We train a linear probe on each individual token and plot the average accuracy over the test set, in percent. Here we plot the results for each token a subset of layers in 3 models: ViT-B/32 trained with a classification token (CLS) or global average pooling (GAP), as well as a ResNet50. Note the different scales of values in different sub-plots.

E Additional Representation Propagation Results

Figure E.1 shows the ratio of representation norms between skip connections and MLP and Self-Attention Blocks. In both cases, we observe that the CLS token representation is mostly unchanged in the first few layers, while later layers change it rapidly, mostly via MLP blocks. The reverse is true for the spatial tokens representing image patches, whose representation is mostly changed in earlier layers and does not change much during later layers. Looking at the cosine similarity of representations between output in Figure E.2 confirms these findings: while spatial token representations change more in early layers, the output of later blocks is very similar to the representation present on the skip connections, while the inverse is true for the CLS token.

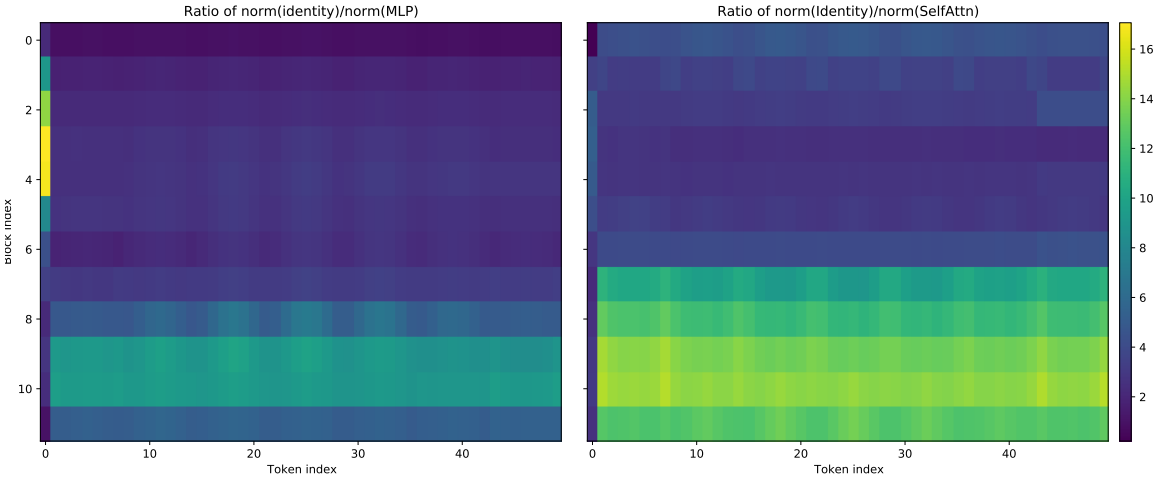


Figure E.1: Additional Heatmaps of Representation Norms Ratio Representation Norms $||z_i||/||f(z_i)||$ between skip connection and the MLP or Self-Attention block for the hidden representation of each block on ViT-B/32, separately for each Token (Token 0 is CLS).

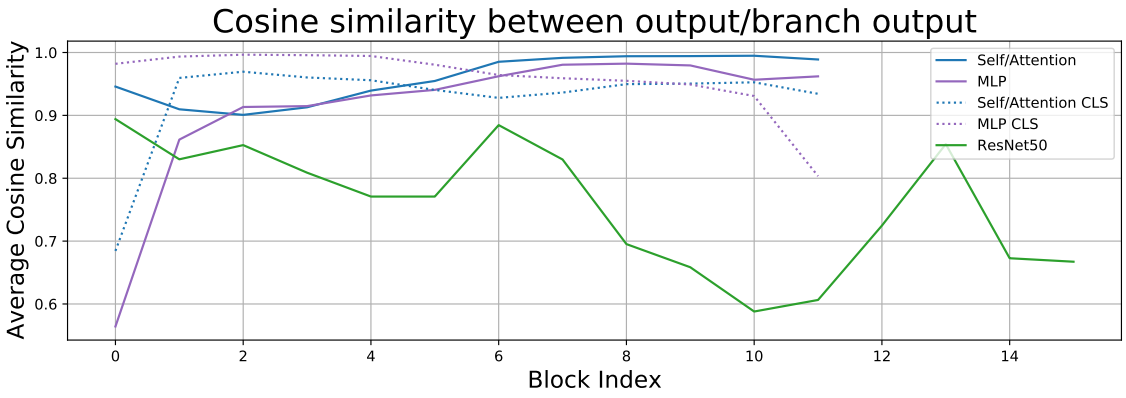


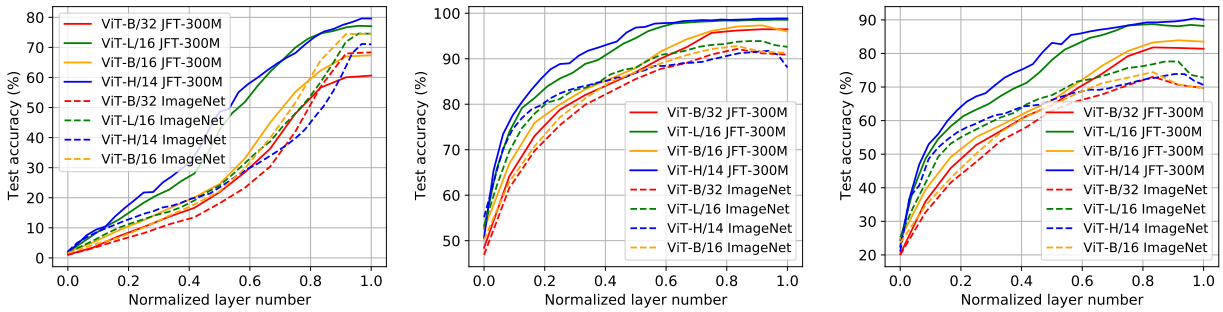
Figure E.2: Most information in ViT passes through Identity Connections. Cosine Similarity of representations between the skip-connection (identity) and the longer branch for ViT-B/16 trained on ImageNet and a ResNet v1. For ViT, we show the CLS token separately from the rest of the representation.

F Additional results on linear probes

Here we provide additional results on linear probes complementing Figures 11 and 13 of the main paper. In particular, we repeat the linear probes on the CIFAR-10 and CIFAR-100 datasets and, in some cases, add more models to comparisons. For CIFAR-10 and CIFAR-100, we use the first 45000 images of the training set for training and the last 5000 images from the training set for validation. Additional results are shown in Figures F.1, F.2, F.3.

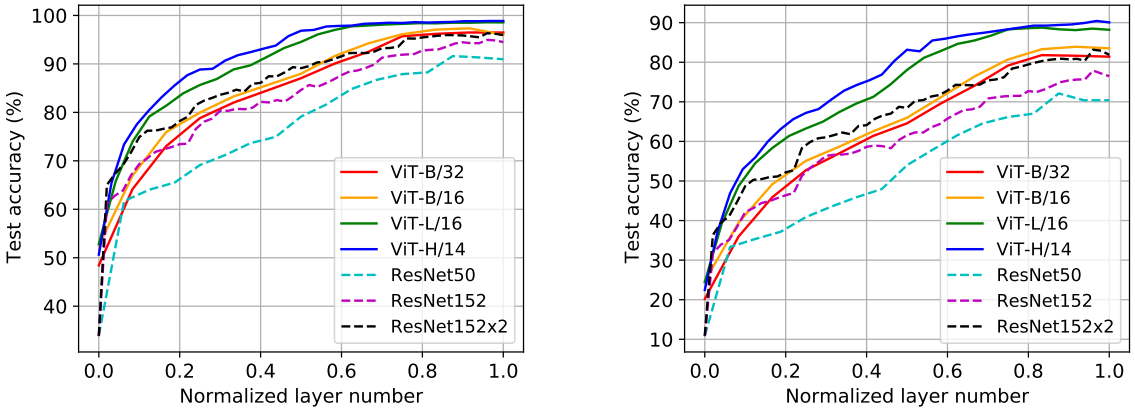
Moreover, we discuss the results in the main paper in more detail. In Figure 11 (left) we experiment with different ways of evaluating a ViT-B/32 model. We vary two aspects: 1) the classifier with which the model was trained, classification token (CLS) or global average pooling (GAP), 2) The way the representation is aggregated: by just taking the first token (which for the CLS models is the CLS token), averaging all tokens, or averaging all tokens except for the first one.

There are three interesting observations to be made. First, CLS and GAP models evaluated with their "native" representation aggregation approach – first token for CLS and GAP for GAP – perform very similarly. Second,



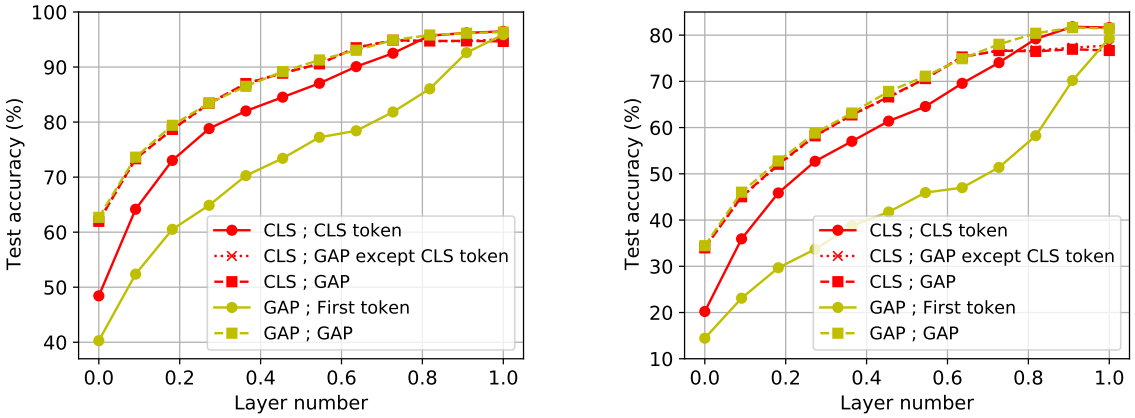
(a) ImageNet (b) CIFAR-10 (c) CIFAR-100

Figure F.1: Experiments with linear probes. Additional results on models pre-trained on JFT-300M and ImageNet (Fig. 13 left) – with the addition of ViT-B models and CIFAR-10/100 datasets.



(a) CIFAR-10 (b) CIFAR-100

Figure F.2: Experiments with linear probes. Additional results on comparison of ViT and ResNet models (Fig. 13 right) on CIFAR-10/100 datasets.



(a) CIFAR-10 (b) CIFAR-100

Figure F.3: Experiments with linear probes. Additional results on comparison of ViT and ResNet models (Fig. 11 right) on CIFAR-10/100 datasets.

the CLS model evaluated with the pooled representation performs on par with the first token evaluation up to last several layers, at which point the performance plateaus. This suggests that the CLS token is crucially contributing to information aggregation in the latter layers. Third, linear probes trained on the first token of a model trained with a GAP classifier perform very poorly for the earlier layers, but substantially improve in the latter layers and almost match the performance of the standard GAP evaluation in the last layer. This suggests all tokens are largely interchangeable in the latter layers of the GAP model.

To better understand the information contained in individual tokens, we trained linear probes on all individual tokens of three models: ViT-B/32 trained with CLS or GAP, as well as ResNet50. Figure 11 (right) plots average performance of these per-token classifiers. There are two main observations to be made. First, in the ViT-CLS model probes trained on individual tokens perform very poorly, confirming that the CLS token plays a crucial role in aggregating global class-relevant information. Second, in ResNet the probes perform poorly in the early layers, but get much better towards the end of the model. This behavior is qualitatively similar to the ViT-GAP model, which is perhaps to be expected, since ResNet is also trained with a GAP classifier.

G Effects of Scale on Transfer Learning

Finally, we study how much representations change through the finetuning process for a model pretrained on JFT-300M, finding significant variation depending on the dataset. For tasks like ImageNet or Cifar100 which are very similar to the natural images setting of JFT300M, the representation does not change too much. For medical data (Diabethic Retinopathy detection) or satellite data (RESISC45), the changes are more pronounced. In all cases, it seems like the first four to five layers remain very well preserved, even accross model sizes. This indicates that the features learned there are likely to be fairly task agnostic, as seen in Figure G.1 and Figure G.2.

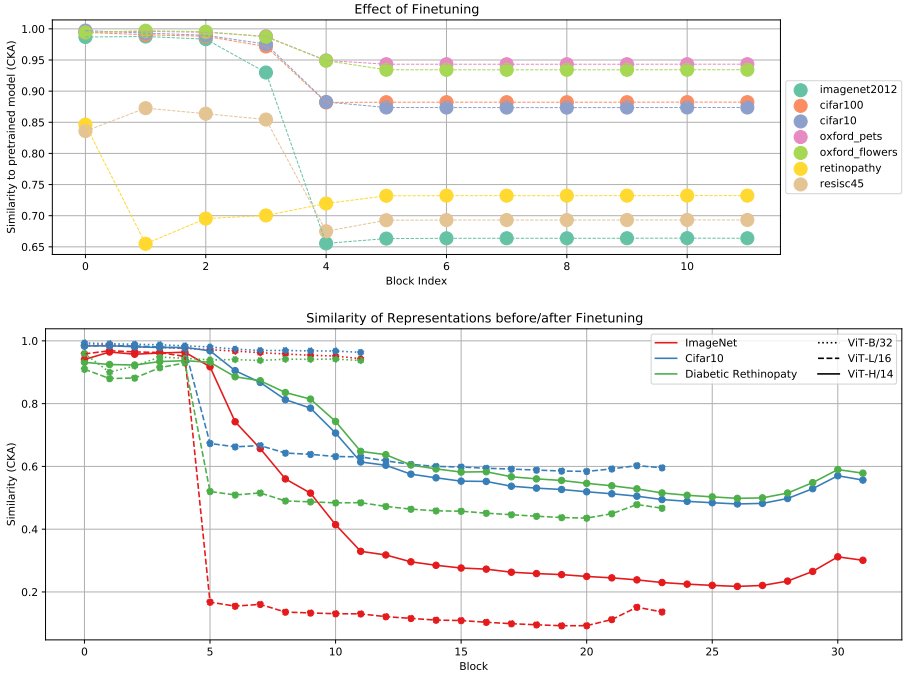


Figure G.1: (top) Similarity of representations at each block for ViT-B/16 models compared to before finetuning. (bottom) Similarity of representations at each block for different ViT model sizes.

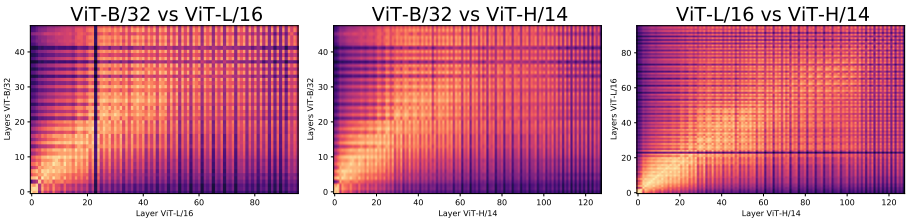


Figure G.2: Similarity of representations of different ViT model sizes.

H Preliminary Results on MLP-Mixer

Figure H.1 shows the representations from various MLP-Mixer models. The representations seem to also fall very clearly into distinct, uncorrelated blocks, with a smaller block in the beginning and a larger block afterwards. This is independent of model size. Comparing these models with ViT or ResNet as in Figure H.2 models makes it clear that overall, the models behave more similar to ViT than ResNets (c.f. Fig. 1 and 2).

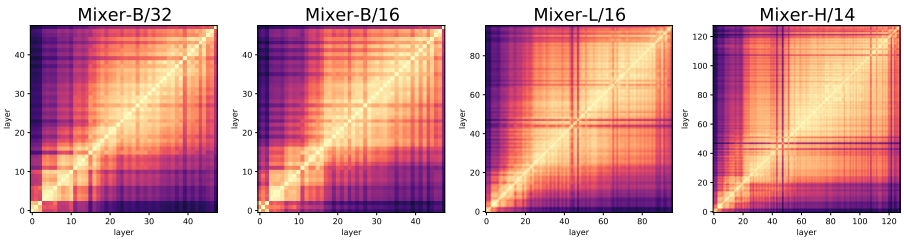


Figure H.1: Similarity of representations of different ViT model sizes.

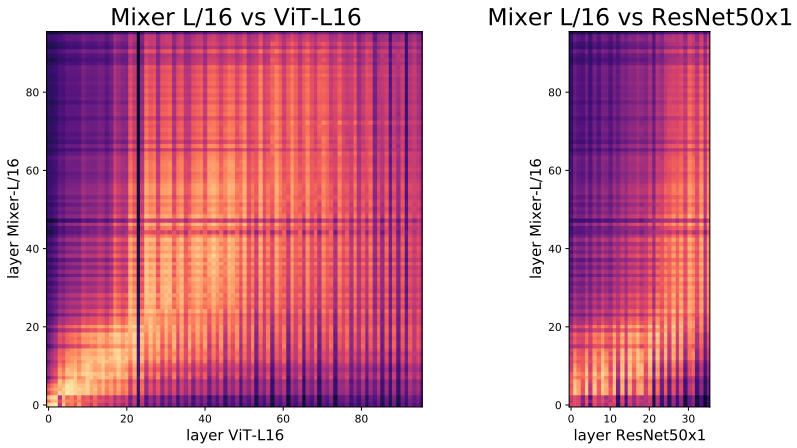


Figure H.2: Similarity of representations of different ViT model sizes.