Understanding Robustness of Transformers for Image Classification

Srinadh Bhojanapalli*, Ayan Chakrabarti*, Daniel Glasner*, Daliang Li*, Thomas Unterthiner*, Andreas Veit*
Google Research
{bsrinadh, ayanchakrab, dglasner, daliangli, unterthiner, aveit}@google.com

Abstract

Deep Convolutional Neural Networks (CNNs) have long been the architecture of choice for computer vision tasks. Recently, Transformer-based architectures like Vision Transformer (ViT) have matched or even surpassed ResNets for image classification. However, details of the Transformer architecture –such as the use of non-overlapping patches– lead one to wonder whether these networks are as robust. In this paper, we perform an extensive study of a variety of different measures of robustness of ViT models and compare the findings to ResNet baselines. We investigate robustness to input perturbations as well as robustness to model perturbations. We find that when pre-trained with a sufficient amount of data, ViT models are at least as robust as the ResNet counterparts on a broad range of perturbations. We also find that Transformers are robust to the removal of almost any single layer, and that while activations from later layers are highly correlated with each other, they nevertheless play an important role in classification.

1. Introduction

Convolutions have served as the building blocks of computer vision algorithms in nearly every application domain—with their property of spatial locality and translation invariance mapping naturally to the characteristics of visual information. Neural networks for vision tasks adopted the use of convolutional layers quite early on [11, 31], and since their resurgence with Krizhevsky et al.'s work [29], all modern networks for vision have been convolutional [41, 44, 17, 22, 25, 24, 47]—with innovations such as residual [17] connections being applied to a backbone of convolutional layers. Given their extensive use, convolutional networks have been the subject of significant analysis—both empirical [45] and analytical [13, 1].

Recently, after seeing tremendous success in language tasks [49, 7, 3], researchers have been exploring a variety of avenues for deploying attention-based *Transformer* networks [4, 9, 48, 27]—and other attention-based architectures [53, 52, 40, 32, 54]—in computer vision. Transformers are also gaining popularity in vision and language tasks [42, 33, 46, 5, 34, 39].

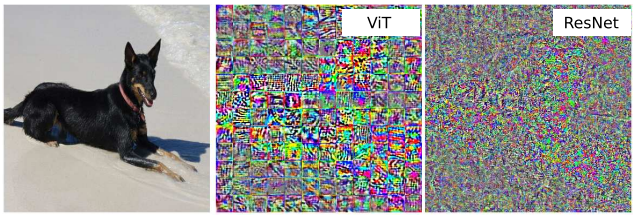

Figure 1. **Transformers vs. ResNets**. While they achieve similar performance for image classification, Transformer and ResNet architectures process their inputs very differently. Shown here are adversarial perturbations computed for a Transformer and a ResNet model, which are qualitatively quite different.

In this paper, we focus on one particular Transformer architecture—the Visual Transformer (ViT) introduced by Dosovitskiy et al. [9]—because it was shown to achieve better performance than state-of-the-art residual networks (ResNets) [17] of similar capacity, when both are pre-trained on sufficiently large datasets [28], such as JFT-300M [43]. We also focus on ViT because, unlike other Transformer models for vision, their architecture consists solely of Transformer layers.

Dosovitskiy et al.'s results [9] tell us that such an architecture is preferable in terms of performance, given enough training data to overcome the lack of an inductive bias through convolutions. But, the pure attention-based mechanism by which ViT models process their inputs is a significant departure from the familiar decade-long ubiquitous use of convolution networks for computer vision. In this paper, we seek to gain a better understanding of how these architectures behave—in terms of their robustness against perturbations to inputs and to the model parameters themselves—and build up an understanding of these models that parallels our knowledge about convolution.

We begin with an exhaustive set of experiments comparing the performance of various ViT model variants, under different perturbations to their image inputs, to similarly

*All authors contributed equally.

sized and trained ResNet architectures [17, 28]. Perturbations range from natural variations [21, 19, 18] to adversarial perturbations [45, 14, 21, 23] and spatial transformations [10]. We also evaluate texture and shape bias [12].

We then turn our attention to the action of the ViT models themselves, analyzing the evolution of information through the cascade of Transformer layers and the redundancies in their internal representations via correlation analysis and lesion studies—as has been done in the past for Transformers for language tasks [51, 38, 36, 8] and for ResNets for vision tasks [50, 15]. Moreover, since it is known that self-attention can learn to mimic convolutions [6], we also investigate the effect of enforcing spatial locality in the attention mechanism of the Transformer layers of ViT models.

Our investigations provide researchers and practitioners with a deeper understanding of how this new class of deep network architectures work, the range of applications to which they may be deployed, and provide potential avenues for how they may be improved in terms of performance or efficiency. Our contributions are as follows:

- We measure robustness of ViT models of different sizes that are pre-trained on different datasets and compare them to corresponding ResNet baselines.
- We measure robustness with respect to input perturbations, and find that ViTs pre-trained on sufficiently large datasets tend to be generally at least as robust as their ResNet counterparts.
- We measure robustness with respect to model perturbations, and find that ViTs are robust to the removal of almost any single layer, and that later layers provide only limited updates to the representations of individual patches, but focus on consolidating information in the CLS token.

2. Preliminaries

2.1. Transformers

Self-attention based Transformer architectures were introduced in [49], where they showed superior performance on machine translation. They have since been applied successfully to many tasks in NLP. Notably [7, 2] have shown that combined with pre-training, these models achieve nearly human performance on a wide range of NLP tasks.

The input to Transformer models is a sequence of vectors—typically embeddings of input tokens—that are processed by a stack of Transformer blocks. Each block consists of 1) a multi-head self-attention layer that aggregates information across tokens using dot-product attention; and 2) a tokenwise feed-forward (MLP) layer. Both use layer normalization and residual connections.

Vision Transformer ViT [9] uses the same Transformer architecture discussed above. They key difference comes in the image pre-processing layer. This layer partitions the image into a sequence of non-overlapping patches followed by a learned linear projection. For example, a 384×384 image can be broken into 16×16 patches resulting in a sequence length of 16^2. This is accomplished using a 2D convolution, where the number of filters determines the hidden size of the sequence input to the Transformer. Following [7] ViT also appends a special CLS token to the input, whose representation is used for final classification .

2.2. Model Variants

In order to better understand and contrast ViTs and ResNets, we evaluate a range of models from each architecture family. We follow [9] and use models which vary in the number of parameters, their input patch-size, and in the datasets on which they were pre-trained. Table 1 summarizes the sizes of the models used in our experiments. We append "/x" to model-names to denote models that take patches of size x as input, and use model variants that were pre-trained either on ILSVRC-2012, with $\sim$ 1.3 million images, on ImageNet-21k, with $\sim$ 12.8 million images, or on JFT-300M [43] which contains around 375M labels for 300M images. All models are finetuned on ILSVRC-2012. We obtained saved parameter checkpoints for the ViT models from the authors of [9], and those for the ResNet models from the authors of [28].

3. Robustness to Input Perturbations

In this section we compare the robustness to input perturbations of ViT models to ResNets. We do this by measuring performance of a range of representatives from each architecture family, as described in Sec. 2.2. To capture different aspects of robustness we rely on different, specialized benchmarks ImageNet-C, ImageNet-R and ImageNet-A. We also pit our models against different types of adversarial attacks. Finally, we explore the texture bias of ViTs.

3.1. Natural Corruptions

So called "natural" or "common" perturbation benchmarks provide an important yardstick for estimating real-world performance in the presence of naturally occurring image corruptions [19, 16, 30]. Robustness to such perturbations can be important for example in safety-critical applications. We use ImageNet-C, a benchmark introduced in [19] to evaluate ViT's robustness to natural corruptions. ImageNet-C includes 15 types of algorithmically generated corruptions, grouped into 4 categories: 'noise', 'blur', 'weather', and 'digital'. Each corruption type has five levels of severity, resulting in 75 distinct corruptions.

Our results, averaged over all corruptions and all severities, are shown in the second column of Fig. 2. More granular results can be found in Appendix C. We find that the size of the pre-training dataset has a fundamental effect on the

Model	ViT-B	ViT-L	ViT-H	ResNet-50x1	ResNet-101x1	ResNet-50x3	ResNet-101x3	ResNet-152x4
# Params	86M	307M	632M	23M	45M	207M	401M	965M

Table 1. **Architectures.** Model architectures used in our experiments along with the number of learnable parameters for each.

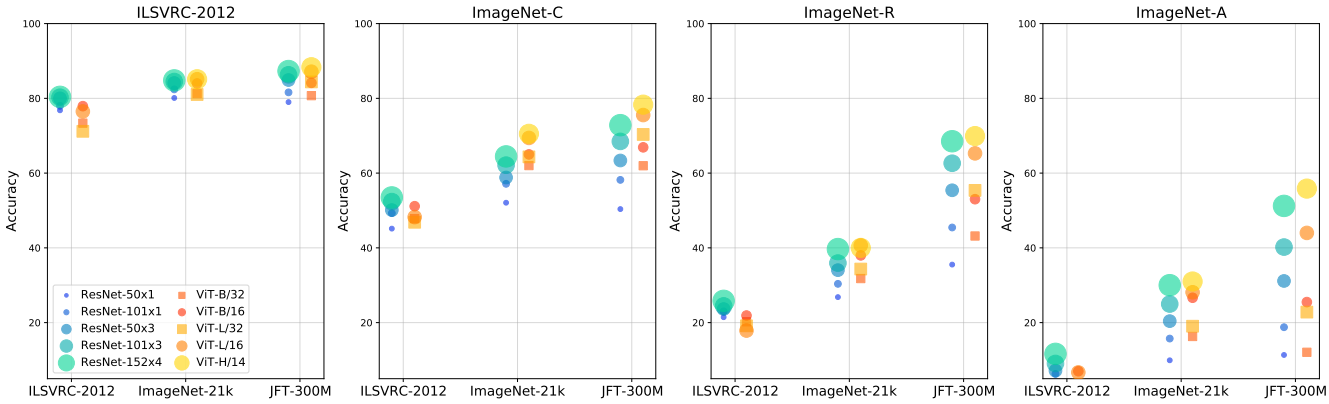

Figure 2. **Robustness Benchmarks**. Accuracy of ViT and ResNet models on ILSVRC-2012 (clean), ImageNet-C, ImageNet-R and ImageNet-A. For ImageNet-C the accuracy is averaged across all corruption types and severity levels. We observe that (i) relative accuracy on ILSVRC-2012 is generally predictive of relative accuracy on the perturbed datasets, and that when trained on sufficient data, the accuracy of ViT models (ii) outperforms ResNets, and (iii) scales better with model size. Marker size related to model size. Detailed results for ImageNet-C can be found in Appendix C.

robustness of ViTs. When the training set is small, the ViTs are less robust compared to ResNets of comparable sizes, and increasing the size of the ViTs does not lead to better robustness. This is consistent with performance on the clean set, and with the observations of [9] about the inductive bias of convolutions being useful when pre-training data is limited. However, when the training data is ImageNet-21k, we observe stronger robustness for most ViT models. This effect becomes even more pronounced when the models are pre-trained on JFT-300M, where ViTs show better robustness against most corruptions compared to ResNets. Moreover, in the larger pre-training data regime, performance gains can be achieved for ViT models by increasing the model size or by decreasing the patch size (and thus increasing the amount of computation).

3.2. Real-World Distribution Shifts

Robustness to distribution shift, can be measured in different ways. Here, we evaluate ViT models on ImageNet-R [18], a dataset with different "renditions" of ILSVRC-2012 classes. An advantage of ImageNet-R is that the renditions are real-world, naturally occurring variations, such as paintings or embroidery, with textures and local image statistics which differ from those of ImageNet images.

Despite the fundamental difference in the nature of the perturbations in ImageNet-R and ImageNet-C, the models' behavior on ImageNet-R is similar, as shown in Fig. 2. Again, ViTs underperform ResNets when the pre-training data is small and starts to outperform them when pre-trained on larger datasets. The benefit of larger model sizes is also more clear on larger datasets, especially for ViTs.

The behavior of our baseline ResNet models is in line with those observed in Appendix G of [28], where they are evaluated on objects out-of-context. The authors of [28] create a dataset of foreground objects corresponding to ILSVRC-2012 classes pasted onto miscellaneous backgrounds. They find that when using more pre-training data, better performance of the larger model on ILSVRC-2012 translates to better out-of-context performance.

Our finding that more pre-training data improves performance on out-of-distribution data is also in line with the findings in NLP. Hendrycks et al. [20] show that pre-trained Transformers improve robustness on a variety of out-of-distribution NLP benchmarks. One of their interesting findings is that for NLP, larger models are not always better. We observe a similar phenomenon for ViTs pre-trained on ILSVRC-2012, but not for ViTs pre-trained on ImageNet-21k or on JFT-300M.

3.3. Natural Adversarial Examples

Adversarial robustness is usually measured by considering the worst-case perturbation within a small radius in image space. We explore ViTs performance on such perturbations in Sec. 3.5. In contrast, the so called "natural adversarial" examples of Hendrycks et al. [21] are unmodified real-world images which were found by filtering with a trained ResNet-50 model, and have been shown to transfer to other models. In contrast to ImageNet-C and ImageNet-R, the local statistics of these images is similar to ImageNet images.

Our results on ImageNet-A are shown in the right col-

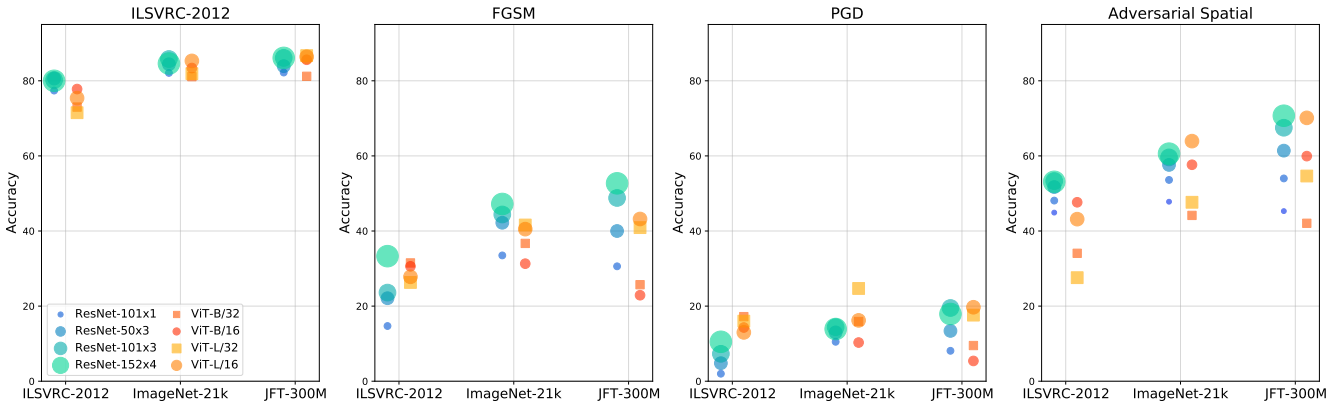

Figure 3. **Adversarial Perturbations**. Accuracy on a subset of 1000 images in ILSVRC-2012 validation of ViT and ResNet models, on clean images *(left)* vs. those subject to model-specific adversarial attacks: FGSM and PGD-based perturbations *(middle)*, and spatial (rotation and translation) transformations *(right)*. (We omit ViT-H/14 here, since it expects a different input image resolution than the other models.) ResNet models are more robust to the simpler FGSM attack than their ViT counterparts, but this advantage disappears for the more successful PGD attacks. For spatial attacks, the 16×16 ViT models exhibit equivalent robustness to ResNets of comparable size, but ViT models with the larger patch-size of 32×32 fare worse.

umn of Fig. 2. We find that ViTs, despite having a dramatically different architecture compared to ResNet-50, are susceptible to the same natural adversarial images. Again we find that larger pre-training datasets are beneficial to ViT models, which start to outperform ResNets when both are pre-trained on JFT-300M. This finding should be taken with a grain of salt, since the adversarial selection process is based on a ResNet-50, so the examples might be harder for ResNets by design.

3.4. Robustness and Model Size

On sufficiently large datasets, it is well known that for a fixed architecture, larger models lead to better quality. Kaplan et al. [26] demonstrated that such improvements on Transformers trained on large NLP datasets follow clear and predictable power laws. In previous subsections, we found that in addition to clean performance, the robustness of ViTs and ResNets against various input perturbations also improves with model size. The gap between large and small models grows as the dataset becomes bigger. It is therefore interesting to evaluate the relation between a models' robustness and its size, when pre-trained on the largest dataset, JFT-300M. The results are summarized in Fig. 4.

We find that the error rates follow consistent trends when scaling up the model size, across up to two orders of magnitude. This holds true on different robustness benchmarks, as well as the clean ILSVRC-2012 validation set. We also note that ViTs exhibit more favorable scaling compared to ResNet. This suggests that given a sufficiently large pre-training dataset, such as JFT-300M, the gap in robustness between ViTs and ResNets will further increase as the models become bigger and bigger. Note that this advantage of ViTs is only realized when the pre-training dataset is sufficiently large. In Appendix D we show that when pre-trained on ImageNet-21k, ViTs' robustness does not scale better than ResNets'.

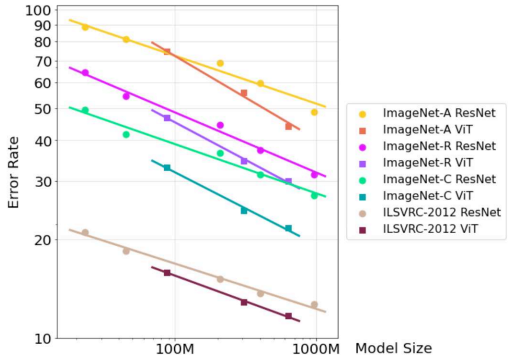

Figure 4. **Scaling**. Performance of ViT and ResNet models as a function of the number of model parameters. All models are pre-trained on JFT-300M and fine-tuned on ILSVRC-2012. We see consistent trends across different input perturbations: scaling up ViTs provides better robustness gains than scaling up ResNets.

We also find that for the same model family, the slope in the error rate vs. model size relation remains relatively consistent across different datasets, despite their drastically different characteristics. This suggests that the scaling trends we discovered might generalize to a broader set of evaluation datasets and tasks.

3.5. Adversarial Perturbations

Most deep neural network models are vulnerable to *adversarial perturbations* [45]—extremely small, but carefully crafted, perturbations to the input, that cause a model to produce incorrect predictions. In NLP, Hsieh et al. [23]

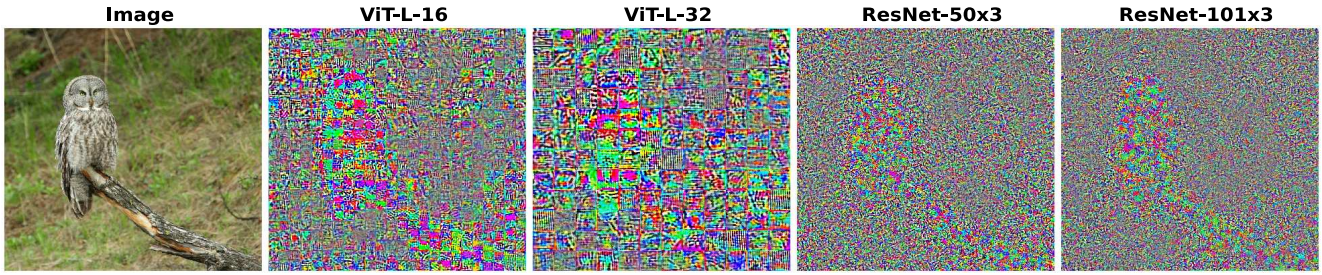

Figure 5. **Example Perturbations**. For example images from the ILSVRC 2012 validation set, we illustrate the perturbations computed using PGD for two ViT models and two ResNet models (we use models pre-trained on JFT-300M). The perturbations are visualized as images by linearly transforming their intensity from the original range of $[-1, 1]$ to $[0, 255]$.

have shown that attention-based models tend to be more robust to such perturbations than other architectures (such as recurrent networks). In this section, we evaluate the robustness of various ViT and ResNet models for image classification to adversarial perturbations.

We consider perturbations with an L_∞ norm of one gray level, computed with knowledge of the model architecture and weights (i.e., white-box attacks). We use two standard approaches to compute these perturbations: the Fast Gradient Sign Method (FGSM) [14] and Projected Gradient Descent (PGD) [35], using eight iterations with a step size of $1/8$ gray levels for the latter. Figure 3 reports the accuracies over a subset of 1000 images from the ILSVRC-2012 validation set, on the original images and after adding perturbations computed using both methods.

We see that the performance of all models degrades with these perturbations, and as expected, PGD is more successful than FGSM. Also, we find that larger models tend to be more robust than smaller ones, and that pre-training on larger datasets improves robustness to adversarial perturbations. Interestingly, among models that are trained only on ILSVRC-2012, the Transformer models appear to be more robust than ResNet models of equivalent size—quite a bit more so with perturbations computed using PGD. Among models trained with a medium amount of training data (pre-trained on ImageNet-21k), we find that ResNet models are more robust to the simpler FGSM attack than their Transformer counterparts, but the opposite is true for PGD attacks. Finally, among models trained with the most training data, robustness to FGSM appears largely to be monotonic with model size. PGD attacks are again more successful, but here, there appear to be diminishing returns with model size once one crosses 300 million parameters.

An interesting observation is that the *relative* robustness of ViT models to their ResNet counterparts appears to be lower for attacks with FGSM than PGD. This is likely due to the presence of the single large linear patch-embedding layer at the beginning of all the ViT models, which causes the single-iteration gradients used by FGSM to better correspond to a pattern coordinated across larger spatial regions.

	ViT→RN	**RN→ViT**
ViT-B/16 vs. RN-101x1	79.7% (-2.5)	85.2% (-0.4)
ViT-B/32 vs. RN-50x3	82.2% (-1.7)	80.9% (-0.3)
ViT-L/16 vs. RN-101x3	84.3% (-1.8)	85.8% (-0.6)
ViT-L/32 vs. RN-152x4	85.4% (-0.7)	86.5% (-0.2)

Table 2. **Transferability.** Accuracy when evaluating adversarial perturbations computed (with PGD) using ViT on ResNet models, and vice-versa. All models are pre-trained on JFT-300M. Numbers in parenthesis indicate difference from accuracy on unperturbed images. The results indicate that adversarial perturbations do not transfer well between ViT and ResNet models. Additional details can be found in Appendix E.

This disadvantage disappears with multiple PGD iterations.

We visualize example patterns computed (using PGD) for Transformer and ResNet models in Fig. 5, and find them to be qualitatively quite different. For all models the perturbations have the highest magnitudes around the foreground objects. For ViT, there is a clear alignment of the patterns to the patch partition boundaries. In contrast, the patterns for ResNet models are more spatially incoherent.

Finally, we find that adversarial patterns do not transfer over between ViT and ResNet architectures—i.e., patterns computed using ViT models rarely degrade the performance of ResNet models and vice-versa (see Table 2 and details in Appendix E). This stands in contrast to our observations with *natural* adversarial images described in Sec. 3.3.

3.6. Adversarial Spatial Perturbations

We now measure the spatial robustness of these models, following the approach of Engstrom et al. [10] who explore the landscape of spatial robustness using *adversarial examples*. In this setting, the adversary's attack is chosen from a given range of translations and rotations. The attack succeeds if any rotated and translated version of the image is misclassified. These attacks are chosen to particularly test the differences in input processing of these models. For example, ViTs' use of large non-overlapping patches could increase their sensitivity to subpatch-sized shifts

We test the performance of ViT and ResNet models under grid attacks, (grid search over a discrete set of rotations and translations), as these were found to be significantly more powerful than any of the other attacks considered in [10]. We consider 9 equally spaced values each, for horizontal and vertical translations in the range $[-16, 16]$ pixels, and 31 equally spaced values for rotations in the range $[-30°, 30°]$. Following [10], when rotating and translating the images, we fill the empty space with zeros (black pixels). We chose the translation ranges to span the largest patch size (32×32) used by any of the ViT models.

We present the results averaged over 1000 images from the validation set of ILSVRC-2012 in the right column of Fig. 3, and find both ViT and ResNet models to be susceptible to spatial attacks. Surprisingly, ViT models with a patch size of 16×16 mostly maintain their positions relative to the ResNet models, indicating they are no more susceptible to spatial adversarial attacks. In contrast, the performance of ViT models that use a larger patch size of 32×32, degrades much more than the comparable ResNet models. We conclude that ViT models with smaller patch size, seem to be as robust to translations and rotations as comparable ResNets. However, ViT models with larger patch sizes tend to be more susceptible to spatial attacks.

3.7. Texture Bias

Geirhos et al. [12] observe that (unlike humans) ImageNet-trained CNNs tend to rely on texture more than on shape for image classification. They further report that reducing the texture bias leads to improved robustness to previously unseen image distortions. We evaluate the texture bias of ViT models and compare it to ResNets using the Conflict Stimuli benchmark of [12]. This dataset is generated by combining 160 images of objects with white background and 48 texture images using style transfer, resulting in 1280 test images with different (possibly conflicting) shape and texture combinations. The fraction of examples in this dataset that are classified correctly by their shape determines the shape accuracy of a model.

The results are shown in Fig. 6. An interesting observation is that the larger patch-sized (32×32) ViT models perform better than the smaller patch-sized (16×16) variants. This trend is different from what we see for clean accuracy as well as for ImageNet-C ImageNet-R and ImageNet-A. This may be due to larger patch inputs preserving object shape more than the smaller patches. We also observe that unlike in all other experiments, the performance of ResNets trained on JFT-300M is not ordered by model size.

4. Robustness to Model Perturbations

In this section we present our experiments on understanding the information flow in ViT models, by computing layer correlations, lesion studies and restricting attention.

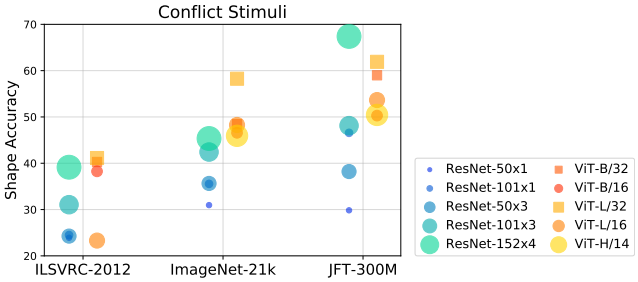

Figure 6. **Texture and Shape**. Shape accuracy of ViT and ResNet models on Conflict Stimuli [12]. In contrast to other robustness results, shape accuracy is more a function of patch size, than model size. 32×32 patch-sized ViT models do better than 16×16 ones.

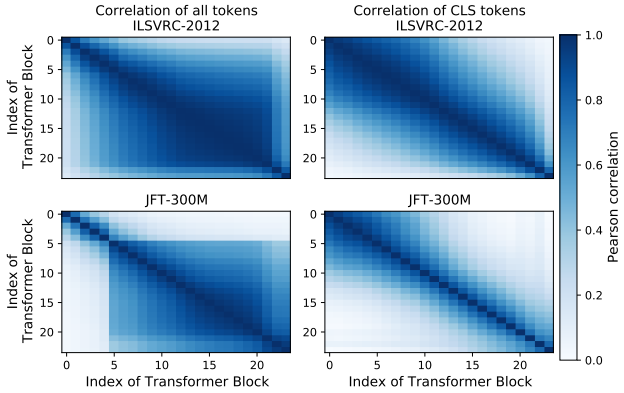

Figure 7. **Representation Correlation Study**. We compare the representations (hidden features) after each Transformer block to those of all other blocks. When taking all tokens into account *(left)*, we observe that representations are increasingly correlated towards the end of the network. In contrast, when only looking at the CLS token *(right)*, the representations become less correlated throughout the network. A potential explanation could be that early layers focus on interactions among spatial tokens whereas later layers focus on the interactions between spatial tokens and the CLS token. Additional results can be found in Appendix F.

We first study how representation of input patches evolves in the ViTs by computing their block level correlations.

Layer correlations We compute correlations between the representations of each Transformer block with the rest. In the left plots of Fig. 7 we present the correlations between representations of all blocks on ViT-L/16 for 2 datasets. Additional results for different models/datasets can be found in Appendix F. We first notice that representation from many blocks towards later layers appear to be highly correlated, indicating a large amount of redundancy. Specifically, we observe that the layers organize into larger groups. In fact a similar pattern can be observed in ResNets, where downsampling layers separate the model into groups with different spatial resolutions. Surprisingly, despite lacking such inductive bias, ViT models also appear to organize layers

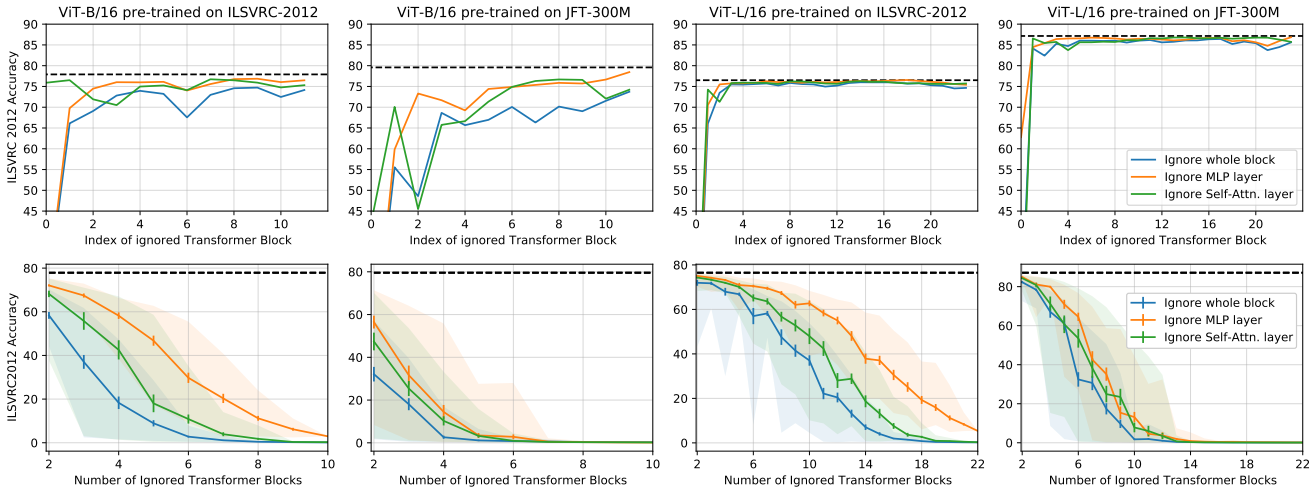

Figure 8. **Lesion Study**. *(top)* Evaluation of ViT models when individual blocks are removed from the model after training. We notice that besides the first block, one can remove any single block, self-attention or MLP from the trained model without substantially degrading performance. The larger models and the models trained only on ILSVRC-2012 are less impacted by the removal of individual layers. *(bottom)* Evaluation of ViT models when $n > 1$ random blocks are removed from the model after training, while always keeping the first block. For each n, results are from 25 independent samples of n blocks and we show the average accuracy (line), standard error (error bars) and min/max (shaded area) across samples. We observe that there is less redundancy in smaller models and their accuracy drops quickly with removal of few blocks. Interestingly, we observe that removing self-attention layers hurts more than removing MLP layers. Additional results can be found in Appendix G.

into stages—the most striking example being a very large, highly correlated group formed by the later layers, where representations appear to change only slightly.

Recall that ViT models append a special CLS token into the input sequence, whose representation is used to make the final classification. We next look at the correlation of the CLS token representations. Looking at this token in isolation, we see a different pattern (see right side of Fig. 7): the representation of CLS tokens only changes slowly at the beginning of the network, but changes rapidly during the later layers. This indicates that the later layers of the network only provide limited updates to the representations of the individual patches, but focus on consolidating the information needed for the classification in the CLS token.

Lesion study The presence of highly correlated representations across blocks raises the question whether the respective blocks are redundant. Previous works [50, 15] have shown that layers in residual networks exhibit a large amount of redundancy, and that almost any individual layer can be removed after training without hurting performance. Following that line of work, we perform a lesion study on ViTs where we remove single blocks from an already trained network during inference, such that information has to flow through the skip connection. Each block contains two skip connections, and we separately investigate the effects of deleting MLP, self-attention layers, or the whole block. This approach is similar to [37], but for whole layers and including the MLP block. As shown in the top row of Fig. 8, it indeed appears that besides the first block one can remove any single block from the model without substantially degrading performance. This is in line with results reported for ResNets.

We next investigate the effect of removing several layers, while always keeping the first block. We observe that as more layers get removed, the performance gradually deteriorates (bottom row of Fig 8), with larger models being more robust to layer removal. We also notice that the amount of training data also influences robustness: models pre-trained on large datasets are less robust to layer removal, indicating perhaps higher model utilization. The results further show that removing individual layers reduces accuracy less than removing full blocks, indicating that there is limited co-adaptation among the components within each Transformer block. Lastly, we notice that removing MLP layers hurts the model less than removing the same number of self-attention layers, indicating the relative importance of self-attention. This behavior seems to be different to transformer models in NLP, which as alluded to by [37] might behave in the opposite way. We have also observed this phenomenon in our own experiments. Additional result from our lesion study can be found in Appendix G.

Restricted attention Finally, we study the extent to which ViT models rely on long-range attention. We evaluate this by spatially *restricting* the attention between

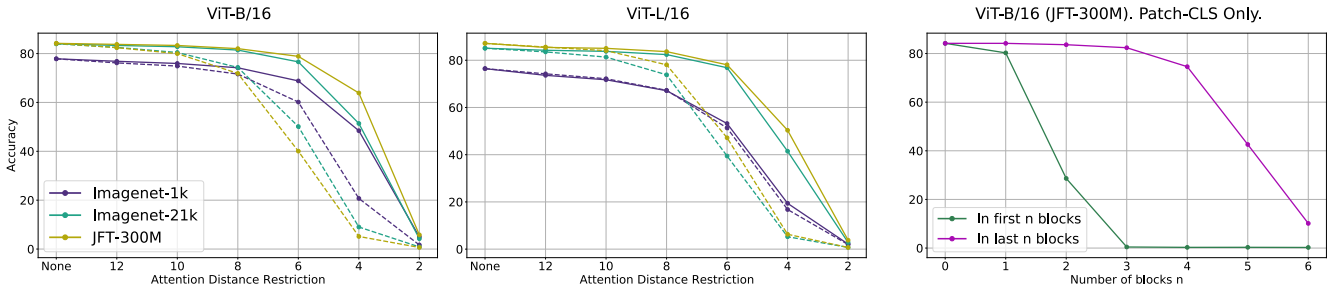

Figure 9. **Restricted Attention**. *(Left, Center)* We evaluate two ViT models, pre-trained on different datasets, in terms of ILSVRC-2012 validation set accuracy when restricting attention among patches to only pairs that lie within a certain maximum horizontal or vertical distance. Dashed lines show the results with an equivalent amount of masking where pairs of patches to mask are chosen randomly (as a random permutation of the mask matrix used for spatially restricted attention). *(Right)* For the ViT-B/16 model pre-trained with JFT-300M, we consider restricting attention to only between patches and the CLS token (without any attention between patches). We report accuracy applying this restriction to a subset of Transformer blocks—both at the beginning and at the end—of the model.

patches to those that lie within a certain distance.We apply this restriction during inference only, by passing a spatial distance-based mask for attention between patches. Note that the masks always allow attention between the CLS token and all patches.

Figure 9 shows that even though these models were trained assuming unrestricted attention, they degrade gracefully when inter-patch attention is restricted to be local. We also compare to a baseline of *randomly* restricting attention by the same amount—achieved by using a random (but fixed across experiments) permutation of the mask matrix. We see that in this case, the degradation in performance is significantly higher in most cases—with a notable exception being the large ViT model that was trained only on ILSVRC-2012 data. Our last evaluation in Fig. 9 considers the extreme version of this case, when attention is only allowed between patches and the CLS token, but not among patches, for a subset of Transformer blocks in the network—applying it either only to blocks at the beginning or at the end. Interestingly, we find that removing inter-patch attention completely at the end of the network has relatively little effect on accuracy—although this is consistent with our earlier observation that in the final few blocks of the network, it is the CLS token that is primarily being updated. In contrast, disrupting inter-patch attention in the initial blocks causes a significant degradation in accuracy.

To summarize, we find that ViT models contain a surprising amount of redundancy, which indicates that the model could be heavily pruned during inference.

5. Takeaways

In this paper, we studied different aspects of robustness in ViT models, making a number of observations. Some of these confirmed existing intuitions about neural networks for vision, while others were perhaps surprising. We summarize they key takeaways from our analysis below:

- Consistent with [9], we find that ViT models generally outperform ResNets and scale better with model size, *when trained on sufficient data*. Crucially, the above is true also of robustness. We found that relative accuracy on the standard ILSVRC-2012 validation set is predictive of performance under a diverse array of perturbations.
- We discovered that FGSM attacks fare better against ViT models than against ResNets. However, ResNet models are not fundamentally more robust since both kinds of models are equally vulnerable to perturbations computed using PGD (which is more successful than the simpler FGSM). However, the optimal perturbations for the two kinds of models are very different and do not transfer.
- We found that choice of patch size in ViT models plays a role in their robustness. Smaller patch sizes make ViT models more robust to adversarial spatial transformations, but also incresae their texture bias.
- Through correlation analysis, we discovered that ViT models organize themselves into correlated groups much like ResNets, despite having no explicit downsampling-based groups like ResNets. This analysis also showed that most updates in the later layers are to the representation of the CLS token, rather than to those of individual patches. Moreover, preventing attention between patches in later layers led to a relatively lower drop in accuracy.
- We also found that despite their ability to allow patches to communicate globally, restricting attention to be local has a relatively lower effect on accuracy.
- Finally, our lesion studies showed that ViT models are fairly robust to removing individual layers. But contrary to observations on language tasks, we found that ViT models are more robust to the removal of MLP layers than self-attention ones.

Acknowledgements. We thank the authors of [9] and of [28] for kindly sharing checkpoints of their pre-trained ViT and ResNet models, respectively.

References

[1] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. 1

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. 2

[3] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 1

[4] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020. 1

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019. 1

[6] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. In *International Conference on Learning Representations*, 2020. 2

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019. 1, 2

[8] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth, 2021. 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2, 3, 8, i

[10] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning*, pages 1802–1811. PMLR, 2019. 2, 5, 6, i

[11] Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982. 1

[12] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 2, 6, ii

[13] Ross Girshick, Forrest Iandola, Trevor Darrell, and Jitendra Malik. Deformable part models are convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 437–446, 2015. 1

[14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2, 5

[15] Klaus Greff, Rupesh K. Srivastava, and Jürgen Schmidhuber. Highway and residual networks learn unrolled iterative estimation. *ICLR*, 2017. 2, 7

[16] Sadaf Gulshad, Jan Hendrik Metzen, and Arnold Smeulders. Adversarial and natural perturbations for general robustness. *arXiv preprint arXiv:2010.01401*, 2020. 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2

[18] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020. 2, 3, ii

[19] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 2, i, ii

[20] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020. 3

[21] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019. 2, 3, ii

[22] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 1

[23] Yu-Lun Hsieh, Minhao Cheng, Da-Cheng Juan, Wei Wei, Wen-Lian Hsu, and Cho-Jui Hsieh. On the robustness of self-attentive models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1520–1529, 2019. 2, 4

[24] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 1

[25] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional net-

works. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1

[26] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. 4

[27] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021. 1

[28] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. *arXiv preprint arXiv:1912.11370*, 6(2):8, 2019. 1, 2, 3, 8, i

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 1

[30] Alfred Laugros, Alice Caplier, and Matthieu Ospici. Are adversarial robustness and common perturbation robustness independant attributes? In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[31] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[32] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11525–11538. Curran Associates, Inc., 2020. 1

[33] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 1

[34] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020. 1

[35] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 5

[36] Swetha Mandava, Szymon Migacz, and Alex Fit Florea. Pay attention when required, 2020. 2

[37] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in Neural Information Processing Systems*, 2019. 7

[38] Ofir Press, Noah A. Smith, and Omer Levy. Improving transformer models by reordering their sublayers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2996–3005, Online, July 2020. Association for Computational Linguistics. 2

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1

[40] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 1

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[42] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019. 1

[43] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 1, 2

[44] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 1

[45] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1, 2, 4

[46] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 1

[47] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 1

[48] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 1

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, 2017. 1, 2

[50] Andreas Veit, Michael Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. *arXiv preprint arXiv:1605.06431*, 2016. 2, 7

[51] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association*

for Computational Linguistics, pages 5797–5808, Florence, Italy, July 2019. Association for Computational Linguistics. 2

[52] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[53] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057. PMLR, 07–09 Jul 2015. 1

[54] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1