Comparing the Decision-Making Mechanisms by Transformers and CNNs via Explanation Methods

Mingqi Jiang, Saeed Khorram, Li Fuxin
Collaborative Robotics and Intelligent Systems (CoRIS) Institute
Oregon State University
{jiangmi, khorrams, lif}@oregonstate.edu

Abstract

In order to gain insights about the decision-making of different visual recognition backbones, we propose two methodologies, sub-explanation counting and cross-testing, that systematically applies deep explanation algorithms on a dataset-wide basis, and compares the statistics generated from the amount and nature of the explanations. These methodologies reveal the difference among networks in terms of two properties called compositionality and disjunctivism. Transformers and ConvNeXt are found to be more compositional, in the sense that they jointly consider multiple parts of the image in building their decisions, whereas traditional CNNs and distilled transformers are less compositional and more disjunctive, which means that they use multiple diverse but smaller set of parts to achieve a confident prediction. Through further experiments, we pinpointed the choice of normalization to be especially important in the compositionality of a model, in that batch normalization leads to less compositionality while group and layer normalization lead to more. Finally, we also analyze the features shared by different backbones and plot a landscape of different models based on their feature-use similarity.

1. Introduction

As attention-based Transformer networks show remarkable performance in image recognition tasks [5, 7, 9, 14, 38], understanding and comparing Transformer and convolution networks (CNNs) at a deeper level become important. Prior work [3, 17] has illustrated interesting differences between CNNs and transformers, but many questions have not been answered. Do transformers have different inner working mechanisms? Why are some transformers seemingly more robust than CNNs? Recent work, such as ConvNeXt [15], utilized design principles in transformer approaches to design a network based on depthwise convolutions and obtained excellent results. Does that indicate that the important contributing factor is not the attention itself but those design principles? If so, which specific design principles particularly affect the decision-making of networks? Better answers to those questions would help us to gain more insights into those deep and complicated black-box networks.

In this paper, we propose a novel methodology to examine these questions through *deep explanation algorithms*. Explanation algorithms have improved significantly in recent years and can generate accurate explanations that can be verified through *intervention experiments* on images [21, 26]. Recent search-based explanation algorithms can find a comprehensive set of *minimally sufficient explanations* (MSEs) [30], defined as the minimal set of patches that, if shown to the network, lead to predictions that are almost as confident as predictions from the full image. The comprehensiveness of the set of MSEs produced by the search algorithm significantly surpasses traditional saliency maps that can only produce one explanation per image.

While per-image explanation methods have greatly improved, they still do not provide a way to obtain a global understanding of the behavior of different network architectures. In this paper, we address this by extracting summary statistics from the explanations for each image and then combining them to obtain dataset-wide statistics. With this approach, we hope to obtain insights that are no longer merely anecdotal, but statistically significant and verifiable.

We propose two approaches in this paper. The first is *sub-explanation counting*, which investigates how networks perform on partial evidence by deleting patches from MSEs and examining the *likelihood ratio* between the predicted conditional probability on those subsets of patches and the full image (Fig. 1 Top-Right). The number of patch conjunctions that have high likelihood ratios indicates a type of behavior we call **compositional**, which means that the classification decision is built jointly on multiple local patches, and removing some of the patches merely lowers the confidence but may not change the classification decision.

We have observed significant differences across architectures regarding compositionality: ConvNeXt and trans-

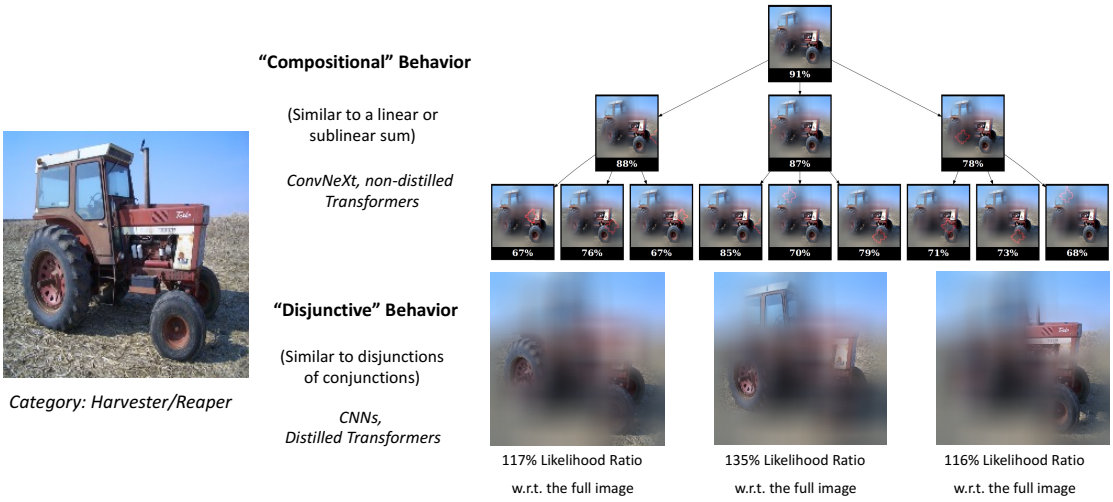


Figure 1. Different behaviors exhibited by different classes of models. Likelihood ratio refers to the ratio between the predicted class-conditional probability of the target category from the masked image and the full image. With the *compositional* behavior, a confident classification is built up jointly from multiple parts, removing some parts may only slightly reduce the likelihood ratio (shown below each node in the tree in the top-right part of the figure). With the *disjunctive* behavior, the network requires very few parts to obtain a highly confident prediction (sometimes more confident than the full image), but it can rely on any of multiple diverse combinations to obtain a confident prediction, similar to a logical OR among the different conjunctions (Best viewed in color)

former models without distillation are much more compositional, with **significantly more** subexplanations than regular CNN models. Further investigation showed that the most important factor in this difference, to our surprise, is not the choice between convolution and attention, but the **normalization** mechanisms used in the networks. Specifically, we found that the **batch normalization** commonly adopted in CNNs leads to a significantly less compositional network, compared to the layer normalization commonly used in transformers. Receptive field size of the model also impacts compositionality to a lesser extent.

The behavior of CNNs can be characterized as more *disjunctive*, which means that the network can predict confidently from a smaller number of patches, although it can recognize any of several diverse patch combinations. Fig. 1 Bottom-Right showed a few examples where a set of several revealed local patches lead to even more confident predictions than the full image, which reflects a distinctly different occlusion handling mechanism than compositional networks. We also found that commonly used **distillation** mechanisms that teach transformers with a CNN lead the transformers to become less compositional and more disjunctive, more similar to the CNNs.

To address the question whether different networks are using the same kind of visual features for classification, we developed a second methodology called *cross-testing*. In cross testing, we compute an explanation (image mask) for an image based on one network, and then submit the masked regions as input to the second network. This helps us to understand whether regions that contribute significantly to the first network are relevant to the second one. If two models rely on similar visual features, then they should score highly in cross-testing. On the other hand, if one model does not respond to the visual features that are deemed important to another model, this implies that they are relying on different features. With this approach, we are able to plot the feature-use landscape of different convolutional networks and transformers, which demonstrates that different networks indeed use different features – the cluster of CNNs, transformers and ConvNeXt are distinct from each other, although distillation can bring transformers closer to the CNNs.

In summary, our contributions are as follows:

- We propose two methodologies, subexplanation counting and cross-testing, which systematically apply model explanation approaches to examine the decision-making mechanisms of image recognition networks.
- With sub-explanation counting, we revealed that the normalization layer significantly impacts model behavior – batch normalization leads to disjunctive behavior (more combinations with fewer patches), while layer/group normalization leads to more compositional behavior (fewer combinations with more patches). Receptive field size also affects compositionality to a lesser extent.
- With cross-testing, we are able to plot the feature-use landscape of different networks and show that CNNs, transformers and ConvNeXt do not use the same visual features for classification, whereas within each group the models are more similar to each other.

2. Related Work

Multiple Explanations. [25] suggested multiple explanations might exist for the decisions made by the deep neural networks. [4] proposed sufficient input subsets so that the

observed values are sufficient to obtain output similar to the original input. They used instance-wise backward selection method to obtain such subsets. [30] proposed *Structured Attention Graphs* (SAG), which employs beam search to generate multiple sufficient patch combinations.

Explanation Using Attribution (Heat) Maps. Attribution maps (heatmaps) are some of the earliest and most widely-studied explanation tools for deep networks. They assign an *attribution* score to each input feature that contributes to the desired output of the network. A majority of the early work, known as *gradient-based* methods, generate attribution maps using the (modified) gradient of the output with respect to the input or intermediate features [2, 29, 32–34, 39]. Later, sanity check procedures showed that most gradient-based explanation methods are independent of the model predictions and mainly work as edge detectors, greatly compromising their credibility [1, 18]. There are also concerns as to whether they are indeed interpretable by humans [44]. *Perturbation-based* approaches directly perturb the image regions (E.g., in [24] which works on superpixels). Most of such approaches optimize for a real-valued mask over the input features to find the regions that significantly decrease the output probability[8, 21, 45]. However, optimization for a mask is highly non-convex and can be easily stuck in a bad local optimum. In addition, it is possible to generate *adversarial* masks [1, 12] that rely break the input features to reduce output confidence. These are easy to locate but do not necessarily explain the decision-making of visual recognition models. Recently, I-GOS [22] alleviated such issues by using the integrated-gradient as the descent direction rather than the gradient, which achieves faster convergence and locates better optima. They also proposed several tricks such as adding noise in the optimization process to avoid adversarial masks and retaining the masked image on the natural image manifold. iGOS++ [13] improved over [22] by additionally optimizing for minimal regions that improve the output confidence as well as enforcing a smoothness term inspired by bilateral filtering. This allowed faithful, non-adversarial and high-resolution masks to be discovered.

Understanding Transformers. Several works have explored the robustness of ViTs against CNNs under common perturbations [3, 17, 20]. [17] observed that ViTs are significantly more robust to occlusions than ResNet50, with DeiT-S maintaining 70% accuracy while ResNet50 drops to 0.1% accuracy on ImageNet when 50% of image regions are randomly removed. [16] examined the adversarial robustness of ViTs. [23] studied the differences in the visual representations learned from ViTs and CNNs, particularly the utilization of global and local information across different layers. [19, 42] further explored the role of self-attention in enhancing the robustness of vision transformers. [19] also revealed contrasting behaviors between attention and convolutional layers, where attention act as low-pass filters while convolutions function as high-pass filters. Different from previous work, our paper seeks to further analyze the underlying decision-making mechanisms transformers and CNNs use with explanation methods.

3. Methods

3.1. Minimal Sufficient Explanations and Structural Explanations

[30] showed that deep networks often have multiple ways to make classifications, and that a single explanation provided by heatmaps does not provide a complete understanding of the decision-making of the network. [30] proposed a more comprehensive way to find explanations by using beam search at low resolutions to systematically find different combinations of image regions that lead to high classification confidence for each image.

Given a classifier f that can predict $f_c(I) = \hat{p}(c|I)$ for an image I, we define the target class $\hat{c} = \arg\max_c \hat{p}(c|I)$ as the class of the image predicted by the classifier. For simplicity, we also call $\hat{p}(\hat{c}|I)$ the *classification confidence* of f on I. The goal is to examine whether the classification stays the same if the input is just a few patches of image I. For this goal, we can divide I into non-overlapping patches p_i, usually at a low-resolution (e.g. 7×7) to avoid adversarial perturbations to the image. Denote a union of those patches as a set N, we could predict the class-conditional probability $f_c(N) = p(c|I \cap N)$, denoting that only pixels in N are retained while the rest are replaced with a baseline image of either 0 or a blurred version of the original image [21].

A Minimal Sufficient Explanation (MSE) is defined as a minimal set of patches that achieves a sufficiently high likelihood ratio w.r.t. the full image, i.e., $f_c(N) > P_h f_c(I)$, where $P_h = 0.9$ in their and our experiments. In layman terms, MSEs are the smallest region that, when shown to the deep network, can generate a prediction almost as confident as the whole image (Fig. 2). For simplicity, we will also call them **explanations** in the rest of the paper. MSEs are not unique and a *beam search* can be used to efficiently find them. The search objective is to find all Ns that achieve a likelihood ratio higher than a threshold P_h, where no subregions in $n_j \subset N$ exceed that threshold,

$$f_c(N) \geq P_h f_c(I), \max_{n_j \subset N} f_c(n_j) < P_h f_c(I). \quad (1)$$

3.2. Sub-Explanation Counting

We propose to gain insights about different types of networks by counting *sub-explanations*, defined as a subset of patches within an MSE: $N_s \subset N$ where N is an MSE of an image I. By definition, $f_c(N_s) < P_h f_c(I)$, but there still can be two different types of behaviors: If the relationship among all patches is more similar to a logical conjunction (logical AND), then $f_c(N_s)$ could be quite low. How-

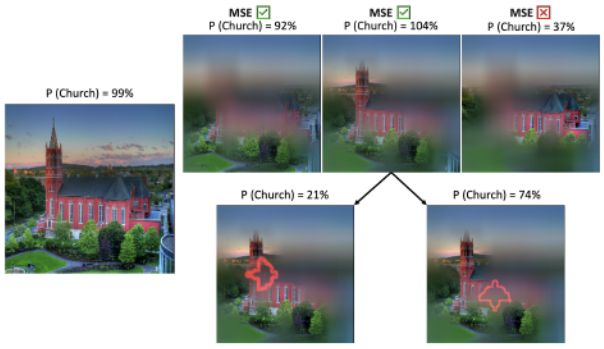


Figure 2. Illustration of Minimal Sufficient Explanations (MSEs) and sub-explanations. MSEs are minimally masked images that the deep network would recognize as the same category as the full image, with its predicted class-conditional probability at least 90% w.r.t. the one from the full image. Sub-explanations are defined as a subset of the patches of an MSE (Best Viewed in Color)

ever, deep networks are not necessarily logical, and there could be another type of behavior in that $f_c(N_s)$ still remains fairly high after occlusions of patches, which we define as a *compositional* relationship (Fig. 1 Top). Counting and comparing the number of sub-explanations across different models can help us understand which type of behavior each model is exhibiting.

Concretely, we construct a tree for each MSE by deleting one patch at a time from a parent node to generate child nodes. Every MSE for a given image is the root of a (sub)tree. In the meantime, we keep evaluating the confidence of current nodes (proper subsets of MSE) using the network f and the image I that is used to generate the MSE. When the nodes are with a likelihood ratio less than 50% compared to the full image, we stop the expansion. Afterwards, we count the number of nodes that have classification confidence above several different thresholds.

Note that being compositional is not the only way to be robust to occlusions. Instead, one could also have multiple MSEs in an OR-relationship to cover all possible occlusions. In that way, the classifier still outputs high classification confidence even with heavy occlusions as long as one of the MSEs corresponds to the occlusion pattern. In the experiments we contrast different models in this regard.

3.3. Metrics with Intervention Experiments

Next, we turn our attention to attribution maps (saliency maps), which is a very popular line of research in explanation but also controversial in that many of the algorithms have been shown to be unreliable [1], mainly because earlier evaluation methodologies based only on localization were not necessarily correlated to the network classification. A better approach to evaluate the attribution map is via perturbing the input according to the map and evaluating the change in network prediction. [26] introduced *MoRF* and *LeRF* metrics in which the patches of image pixels are first ordered based on the attribution map values. Then, the most relevant features (MoRF) and least relevant features (LeRF) are gradually replaced by random noise sampled from a uniform distribution. Finally, the perturbed images are passed through the model and their classification confidences are obtained. Similarly, [21] proposed the *deletion* and *insertion* metrics with the main difference being that during the perturbation, the substitute patches of pixels are sampled from a baseline image, e.g., a highly-blurred version of the image, rather than random noise. This way, sharp edges/boundaries are not introduced in the evaluation images, keeping them closer to the natural image distribution that the network is trained on.

One can use the area under the curve (AUC) from the MoRF/deletion and LeRF/insertion curves as metrics reflecting the effectiveness of the explanation method in finding salient regions (Fig. 3). In this paper we focus on the insertion metric, where a high insertion score indicates a sharp increase in the output confidence after the insertion of the most salient regions into the baseline image. Note that these evaluation schemes can be done automatically and do not require human-defined labels/bounding boxes [40], which makes large-scale quantitative evaluations easier. Formally, given an input image I, a baseline image $\tilde{I}$, classifier f, a target explanation class c, and an attribution map M with elements in $[0, 1]$, we can define the insertion metric as,

$$\frac{1}{2T}\left\langle \sum_{t=0}^{T-1} f_c\left(\phi^{(t)}(\tilde{I}, I, M)\right) + f_c\left(\phi^{(t+1)}(\tilde{I}, I, M)\right)\right\rangle_{p_{\text{data}}}$$
$$\phi^{(t)}(\tilde{I}, I, M) = I \odot M^{(t)} + \tilde{I} \odot (1 - M^{(t)}) \quad (2)$$

where T is the total number of perturbation steps, and $\phi^{(t)}$ generates the perturbed image after t steps, i.e., $M^{(t)}$ only keeps the top $\frac{t}{T}$ of the pixels (Perturbation Ratio) in the attribution map and the rest of the pixels, if any, are set to zero. Of course, $\phi^{(0)} = I$ and $\phi^{(T)} = \tilde{I}$.

3.4. Cross-Testing

Inspired by the intervention-based metrics, we propose to evaluate the similarity between models by using one deep model to generate an attribution map, and successively mask the images based on the attribution map to assess the insertion/deletion metrics on the second deep model. For fair comparison across different models which may have different average classification confidences, we normalize the scores based on the average top-1 classification confidence on the original image t and on fully blurred images b for each model by $\bar{s} = (s - b)/(t - b)$ [28]. This method assesses the similarity of different models under occlusion. With a pairwise similarity matrix between two cross-tested models, we then utilize kernel-based dimensionality reduction approaches [27] to visualize them in a 2D space. This gives insights about whether the features used in the first network is salient to the second one.

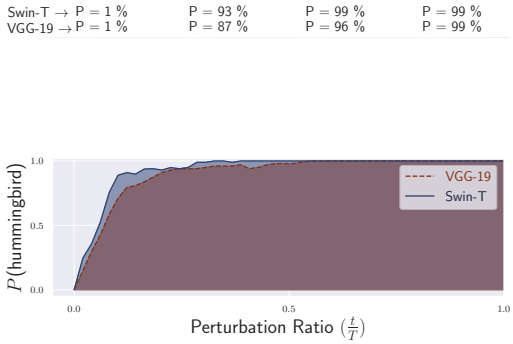


Figure 3. Cross-testing the Insertion metric between VGG-19 and Swin-T for "hummingbird". (Top) Insertion images are obtained by successively revealing pixels that are deemed salient by the heatmap; (Bottom) The Area Under the Curves (AUC) are used to compute the insertion metric for each classifier, when heatmaps are generated from only one of them (Best Viewed in Color)

4. Experiments

We compare ResNet50 [10], ResNet50-C1, ResNet50-C2, ResNet50-D [37], VGG19 [31], ConvNeXt-T [15], Swin-T [14], Nest-T [41], DeiT-S [35], DeiT-S-distilled [35], PiT-S [11], PiT-S-distilled [11] and LeViT-256 [9] in our experiments. Of these, ResNet50 and VGG19 are older CNN models trained with less data augmentation. ResNet50-C1, ResNet50-C2 and ResNet50-D are ResNet50 variants trained with modern data augmentation strategies. ConvNeXt-T is a hybrid model based on large-kernel depthwise convolutions. Swin-T, Nest-T, DeiT-S, and PiT-S are transformers with different architectural structures. DeiT-S-distilled, PiT-S-distilled and LeViT-256 were trained by distilling from a teacher CNN while DeiT-S and PiT-S were trained without distillation. The chosen models have similar sizes and similar accuracy on ImageNet (see Supplementary). To obtain standard deviations, we trained a few models with the same procedure but multiple random seeds, those results are provided in the Supplementary.

In all experiments, we use the first 5,000 images from the ImageNet validation dataset [6] due to the slow speed of running all the experiments.

4.1. Explanations and Sub-Explanations

We follow [30] to perform a beam search with width 5 on different patch combinations, with the image divided into a 7×7 grid with 49 patches. The baseline image was set to a blurred version of the original image (see Supplementary for results with a zero-image baseline). In Table 1, we count the number of MSEs and subexplanations among different networks.

Disjunctivism and Compositionality. Table 1 shows distinct differences among the different models. Most CNNs, ConvNeXts and distilled transformers have **higher** MSE counts and **smaller** MSE sizes. In contrast, Swin Transformers and other undistilled transformers have smaller MSE counts and larger MSE sizes. The differences are statistically significant (tests shown in the Supplementary).

Model Type	Model Name	MSEs Count	MSEs Size	Number of Subexplanations $\geq 80\%$	$\geq 70\%$	$\geq 60\%$	$\geq 50\%$
older CNNs	VGG19	6.93	7.17	27.45	109.99	191.15	329.97
	ResNet50	6.76	7.28	53.68	108.55	180.44	296.92
newer CNNs	ResNet50-C1	9.52	6.37	194.16	320.53	430.73	591.69
	ResNet50-C2	11.01	5.94	88.82	202.27	369.35	568.91
	ResNet50-D	9.88	6.02	146.78	216.31	272.459	332.22
ConvNeXt	ConvNeXt-T	10.28	6.14	**980.16**	**2001.67**	**3610.37**	5360.43
Transformers	Swin-T	8.90	8.01	221.58	882.72	2933.03	**7268.20**
	Nest-T	7.18	**8.77**	432.37	1093.08	2725.06	6006.22
	DeiT-S	8.95	7.72	72.09	333.84	1097.58	2408.30
	PiT-S	7.89	7.49	131.32	607.97	1803.04	3862.10
Transformers with Distillation	DeiT-S-dis	10.22	5.77	57.52	114.21	227.41	467.72
	PiT-S-dis	10.06	5.86	48.54	91.45	182.67	334.45
	LeViT-256	**12.59**	5.50	54.96	103.24	177.33	253.66

Table 1. Results of beam search to locate MSEs. The numbers on the top right are thresholds on the likelihood ratio between a subexplanation and the full image

Recalling the definition of MSEs, a higher count and smaller size means that the network needs the conjunction of **fewer** patches to form a confident classification. However, the network can be robust to occlusions or missing visual features, since it can use a **different** conjunction if a certain important feature cannot be seen. This is what we define as disjunctivism.

In contrast, for transformer models which exhibit larger MSE sizes, we note that the number of subexplanations is also significantly higher. This suggests the **compositional** mechanism for handling occlusions: in each conjunction of patches, removing some of the patches only slightly lowers the classification confidence, whereas in CNNs and distilled transformers removing some of the patches in an MSE greatly lowers the classification confidence (leading to fewer subexplanations). It is clear that disjunctivism and compositionality are different mechanisms that can both deal with occlusion and missing features.

A separate result is the effect of data augmentation: newer CNNs with better data augmentation has significantly more MSEs than older ones, showing higher robustness.

A disclaimer is that these are overall trends that can only be observed by systematically evaluating the explanations on a large dataset. We can find any network using any of these inference strategies for a specific image, and in many images all networks use a similar set of features, leading to smaller differences overall among them in Table 1 (see Fig. 4 and Supplementary for visual examples). This shows the importance of the statistical approach we are taking w.r.t. explanation methods, as it uncovers the trend from the noisy signals of individual images.

What drives the high number of sub-explanations in ConvNeXt and Swin Transformers? One specifically interesting aspect is the **significantly higher** number of subexplanations in ConvNeXt and transformers without distillation. For example, the subexplanations with $\geq 50\%$

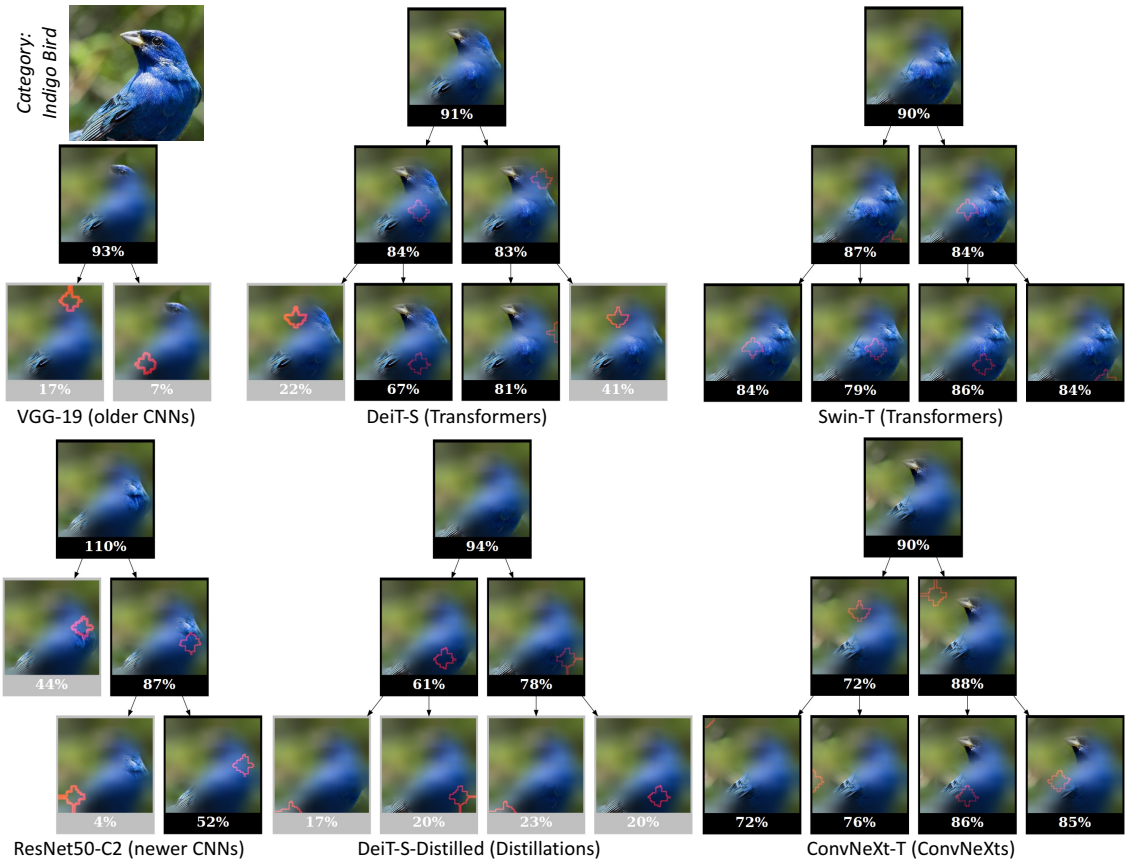


Figure 4. MSEs and some sub-explanations of different models on an image of the Indigo Bird class. Due to the space limit we only subsampled a few subexplanations. The removed patch from the parent node is indicated with a red outline. (Best viewed in Color)

confidence ratio are usually in the thousands in those networks, versus hundreds in the CNNs and transformers with distillations. ConvNeXt especially, shows up as an outlier in our analysis as it looks **both** having more MSE counts and smaller sizes similar to CNNs as well as being compositional with more subexplanations. Hence, we set out to examine which design aspect specifically drove the high number of sub-explanations.

We attempted to strip out the design elements of ConvNexT one by one, and mirrored the experiment with Swin Transformers as well. Specifically, we trained ConvNeXt-T using a 3×3 kernel size for all the ConvNeXt blocks, and Swin-T using a 4x4 window size in the first two stages, and name the resulting models *ConvNeXt-T-3* and *Swin-T-4*. Noting that the results did not fully explain the differences in subexplanations, we replaced the original layer normalization (LN) with batch normalization (BN) and group normalization (GN) and further trained models with different normalizations and smaller receptive fields: *ConvNeXt-T-3-BN*, *ConvNeXt-T-3-GN*, *Swin-T-4-BN*, and *Swin-T-4-GN*. These changes did not reduce performance on ImageNet. More information is provided in the Supplementary.

The results in Table 2 are quite **surprising** for us, as we did not expect batch normalization to play such a significant role: reducing the size of the receptive field reduced the subexplanations by about 40%, but changing layer normalization to batch normalization **very significantly** reduced the number of subexplanations by about **80%**, driving ConvNeXt and Swin Transformer back to levels similar with CNNs. This shows that although receptive field size and normalization both played a significant role in compositionality, the choice of normalization is a much stronger factor. GN exhibited compositional behaviors similar to LN, and both are distinctly different from BN.

Model		MSEs		Number of Subexplanations			
Type	Name	Count	Size	≥ 80%	≥ 70%	≥ 60%	≥ 50%
ConvNeXts	ConvNeXt-T	**10.28**	6.14	**980.16**	**2001.67**	3610.37	5360.43
	ConvNeXt-T-3	9.31	6.56	526.40	1136.17	2012.84	3089.83
	ConvNeXt-T-3-GN	6.60	**7.96**	471.87	1468.28	**3742.13**	**7476.92**
	ConvNeXt-T-3-BN	9.31	6.50	64.39	157.46	326.03	672.92
Transformers	Swin-T	8.90	8.01	**221.58**	**882.72**	**2933.03**	**7268.20**
	Swin-T-4	8.11	7.46	139.29	588.75	1885.98	4276.08
	Swin-T-4-GN	6.57	**9.42**	207.58	821.59	2641.81	7039.34
	Swin-T-4-BN	**9.40**	7.18	42.92	127.05	387.41	943.29

Table 2. Results of beam search to locate MSEs on ConvNeXt and Swin variants. The numbers on the top right are thresholds on the likelihood ratio between a subexplanation and the full image

Put in other words, using **BN strongly** leads the network to be **less compositional**, in the sense that missing features in a conjunction drops the prediction confidence more quickly. This makes the relationship among features more

like logical AND/ORs, rather than a linear sum. We attempt to explain this by examining the normalization dimension of these approaches. Batch normalization only normalizes within a single channel and **does not** normalize across different network channels, whereas GN and LN normalize across different channels. This could lead to an effect that a few large activations dominate the prediction when a network is using BN. Fig. 5 shows the activation map values in different networks and indeed the trend is clear that the top feature channels when using BN are much more dominant than with GN and LN. This effect is a preference during the network training process – it does not mean that BN and LN lead to fundamentally different network architectures. Table 1 showed that distillation from a CNN can reduce the number of subexplanations of an LN-normalized transformer and thus reduce its compositionality.

Having discovered the effect, the open question is whether it is good or bad? We do not have enough concrete evidence to support an outright answer, but some intuitive argument for compositionality can be made – over-reliance on the existence of a few local features might reduce robustness to adversarial examples. Although disjunctivism compensates by introducing more conjunctions, the increase is only about 2 MSEs per image on average. Calibration under uncertainty might also be easier with compositionality, because it is easier to generate "semi-confident" predictions. ConvNeXt and Swin Transformers usually dominate distilled ViTs on downstream visual tasks such as detection and segmentation, where one can argue it might be important to look at more features, not only the ones that are most discriminative. Potentially, one could also take a harder look at combinations of batch and group normalizations such as [43], which indeed showed better adversarial accuracy and domain adaptation capabilities but was hardly ever used in recent architectures. A potential argument for disjunctivism might be to support more consistent predictive confidence under occlusions, which could be useful if a firm decision needs to be made regardless of occlusions.

4.2. Cross-Testing with Attribution Maps

The other question we seek to answer is whether different types of networks are using similar features to classify via cross-testing their attribution maps. We use a state-of-the-art attribution map method iGOS++ [13] to generate heatmaps for each image at 28×28 resolution. This resolution is chosen because it is the highest resolution for which iGOS++ has consistently good performance across different networks. We then calculate the insertion and deletion scores based on the obtained heatmap values (full results in Supplementary). In order to better visualize these similarities, we applied Kernel PCA to project them to 2 dimensions [27], based on the similarities of the insertion scores. Figure 6 shows the projection results.

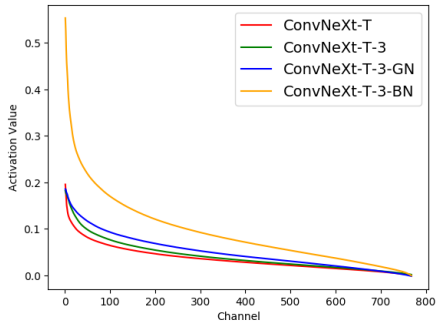


Figure 5. Sorted average values of the maximal activation in each image for each channel in the last block for ConvNeXt-T variants

It can be found that the same type of models use similar features for their predictions. We can roughly delineate clusters of older CNNs (VGG19, ResNet50), newer CNNs (ResNet50-C1, ResNet50-C2, ResNet50-D), ConvNeXt variants (ConvNeXt-T, ConvNeXt-T-3, ConvNeXt-T-3-BN and ConvNeXt-T-3-GN), non-distilled transformers (Swin-T, Swin-T-4, Swin-T-4-BN, Swin-T-4-GN, Nest-T, Deit-S, PiT-S) and distilled transformers (PiT-S-distilled, Deit-S-distilled and LeViT-256). Results using iGOS++ with setting perturbed pixels to zero and another attribution map approach, Score-CAM [36], are shown in the supplementary which show similar trends.

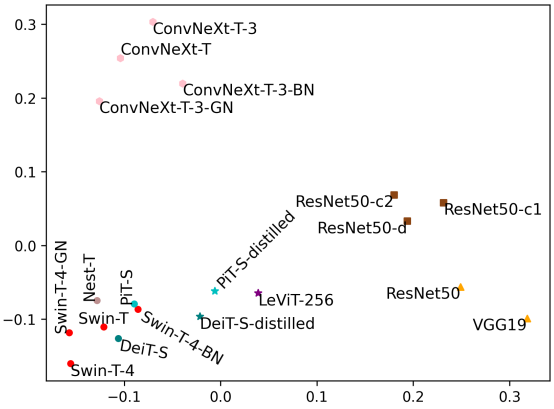
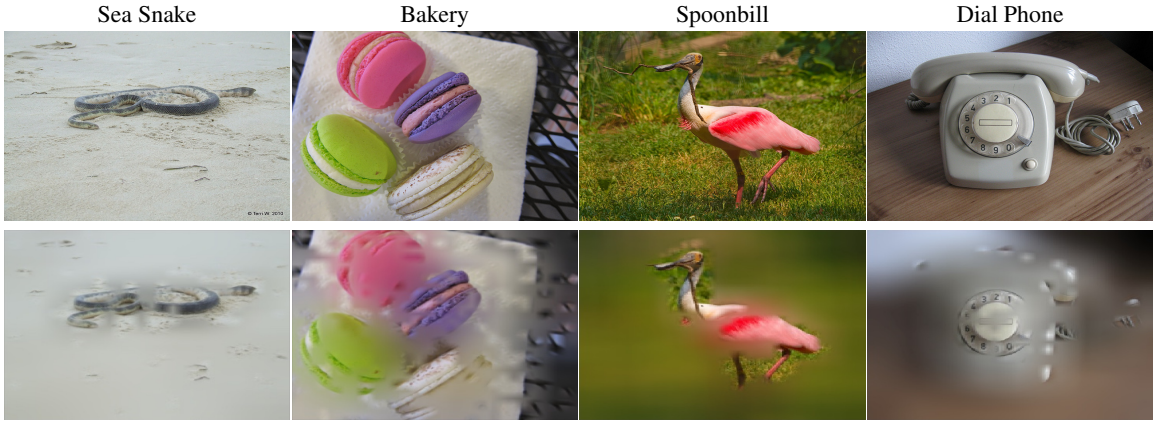


Figure 6. Kernel PCA projections of different models using the insertion metrics. We can see that models that are the same type are more similar to each other in this plot, and that distillation brings transformers closer to CNNs

We also show the average similarity among the different clusters in the confusion matrix in Fig. 8. It can be seen that the confusion matrix is not symmetric – if we generate the heatmap from old CNNs, the insertion scores among all types of networks are consistently high. However if we generate the heatmap with ConvNeXts, then the insertion score into old CNNs are significantly lower. This shows that older CNNs are more singularly minded and utilized fewer features that are more likely a subset of what is used in newer networks. Newer networks have relatively similar cross-testing insertion scores around 0.8. However, distilled trans-



Prediction Confidence on the Partially Occluded Image

VGG19	ResNet50-c2	ConvNeXt-T	VGG19	ResNet50-c2	ConvNeXt-T	VGG19	ResNet50-c2	ConvNeXt-T	VGG19	ResNet50-c2	ConvNeXt-T
0.0494	0.2988	0.5609	0.0775	0.4752	0.1447	**0.9825**	0.8261	0.9027	0.0043	0.0607	0.0595
DeiT-S	DeiT-S-dis	Swin-T	DeiT-S	DeiT-S-dis	Swin-T	DeiT-S	DeiT-S-dis	Swin-T	DeiT-S	DeiT-S-dis	Swin-T
0.3048	0.7156	**0.8593**	0.5952	0.8857	**0.8345**	0.8138	0.9983	0.8030	0.8638	**0.9896**	0.6541

Figure 7. Qualitative Cross-Testing Results. Partially occluded images were generated with iGOS++ from the model with bolded number, then tested on multiple networks and we show the prediction confidences on the ground truth class on each network. (Best viewed in color)

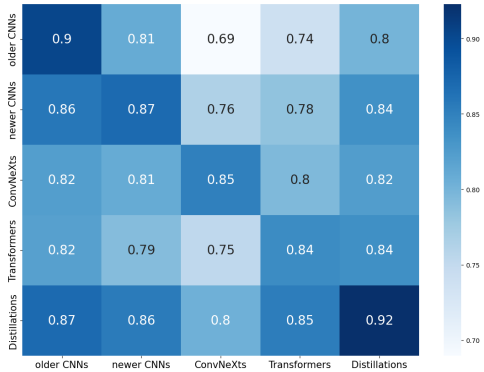


Figure 8. Confusion matrix among model groups. The rows are the models used to generate the attribution maps and the columns are the models that the attribution map is cross-tested on. The diagonal values reflect intra-group differences

formers have more similarity with both newer CNNs and other transformers. The conclusion is that these model families still sometimes use different features to classify, which points to potential benefits from ensembles with a model in each cluster. A quick test showed that a simple average of the prediction of ConvNeXt-T (82.1% accuracy on ImageNet), Swin-T (81.2%) and ResNet50-c2 (80.0%) would achieve 82.9% accuracy, surpassing all individual models.

Figure 7 shows qualitative results from cross-testing. One can again see that distilled transformer models sometimes obtain high confidence with a few regions shown. In the `Bakery` image in the second column, the heatmap is generated with Swin-T, but DeiT-S-distilled have higher confidence than Swin-T, showing that they required less information to obtain more confident predictions. On the other hand, ConvNeXt-T and the ResNets have lower confidence, showing that they may be using different features not shown in this occluded image. In the `Spoonbill` image in the 3rd column generated by VGG, most other networks were also able to obtain a confident classification. Yet, VGG fares quite poorly on masked images generated by other networks, showing their overreliance on specific features that may not be present under those masks.

5. Conclusion

In this paper, we proposed two novel methodologies, sub-explanation counting and cross-testing, that utilize deep explanation algorithms to collect dataset-wide statistics for understanding the decision-making behaviors of different visual recognition backbones. Our analysis indicates that different types of visual recognition models exhibit quite different behaviors along the concept axes of disjunctivism and compositionality. Among other findings, one finding of note is that the choice of normalization strongly affect the compositionality of the model. Receptive field size and data augmentation were shown to also affect model behavior. With cross-testing we characterized the feature-use landscape of model families. We hope the insights from our studies could help people better understand decision-making mechanisms of deep visual models and inspire thoughts about future model designs.

Acknowledgements

This work is partially supported by NSF-1751402 and ONR/NAVSEA contract N00024-10-D-6318. We also like to thank Dr. Tom Dietterich for his valuable suggestions and Lianghui Wang for assisting with HPC resources.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9505–9515, 2018. 3, 4

[2] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*, 10, 2015. 3

[3] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 1, 3

[4] Brandon Carter, Jonas Mueller, Siddhartha Jain, and David Gifford. What made you do this? Understanding black-box decisions with sufficient input subsets. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 567–576. PMLR, 2019. 2

[5] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021. 1

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 5

[7] 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. In *International Conference on Learning Representations*, 2021. 1

[8] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2950–2958, 2019. 3

[9] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. LeViT: a Vision Transformer in ConvNet's Clothing for Faster Inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12259–12269, 2021. 1, 5

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[11] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *International Conference on Computer Vision (ICCV)*, 2021. 5

[12] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019. 3

[13] Saeed Khorram, Tyler Lawson, and Fuxin Li. iGOS++: Integrated Gradient Optimized Saliency by Bilateral Perturbations. In *Proceedings of the Conference on Health, Inference, and Learning*, 2021. 3, 7

[14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1, 5

[15] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 5

[16] Kaleel Mahmood, Rigel Mahmood, and Marten van Dijk. On the robustness of vision transformers to adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7838–7847, 2021. 3

[17] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing Properties of Vision Transformers. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 3

[18] W. Nie, Y. Zhang, and A. Patel. A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizations. *ArXiv e-prints*, 2018. 3

[19] Namuk Park and Songkuk Kim. How Do Vision Transformers Work? In *International Conference on Learning Representations*, 2022. 3

[20] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 2071–2081. AAAI Press, 2022. 3

[21] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 1, 3, 4

[22] Zhongang Qi, Saeed Khorram, and Li Fuxin. Visualizing deep networks by optimizing with integrated gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 3

[23] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 12116–12128, 2021. 3

[24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016. 3

[25] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2

[26] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016. 1, 4

[27] Bernhard Scholkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *ADVANCES IN KERNEL METHODS - SUPPORT VECTOR LEARNING*, pages 327–352. MIT Press, 1999. 4, 7

[28] Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. *arXiv preprint arXiv:2001.00396*, 2020. 4

[29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 3

[30] Vivswan Shitole, Fuxin Li, Minsuk Kahng, Prasad Tadepalli, and Alan Fern. One Explanation is Not Enough: Structured Attention Graphs for Image Classification. In *Advances in Neural Information Processing Systems*, 2021. 1, 3, 5

[31] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations*, 2015. 5

[32] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 3

[33] J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for Simplicity: The All Convolutional Net. In *ICLR Workshop*, 2015.

[34] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017. 3

[35] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357, 2021. 5

[36] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020. 7

[37] Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. *CoRR*, abs/2110.00476, 2021. 5

[38] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[39] Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing. 3

[40] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *European Conference on Computer Vision*, pages 543–559. Springer, 2016. 4

[41] Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, , Sercan Ö. Arık, and Tomas Pfister. Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2022. 5

[42] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M. Alvarez. Understanding the robustness in vision transformers. In *Proceedings of the 39th International Conference on Machine Learning*, pages 27378–27394. PMLR, 2022. 3

[43] Xiao-Yun Zhou, Jiacheng Sun, Nanyang Ye, Xu Lan, Qijun Luo, Bo-Lin Lai, Pedro Esperanca, Guang-Zhong Yang, and Zhenguo Li. Batch group normalization. *arXiv preprint arXiv:2012.02782*, 2020. 7

[44] Roland S Zimmermann, Judy Borowski, Robert Geirhos, Matthias Bethge, Thomas Wallis, and Wieland Brendel. How well do feature visualizations support causal understanding of cnn activations? *Advances in Neural Information Processing Systems*, 34:11730–11744, 2021. 3

[45] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*, 2017. 3