EXAMINING THE DIFFERENCE AMONG TRANSFORMERS AND CNNS WITH EXPLANATION METHODS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a methodology that systematically applies deep explanation algorithms on a dataset-wide basis, to compare different types of visual recognition backbones, such as convolutional networks (CNNs), global attention networks, and local attention networks. We examine both qualitative visualizations and quantitative statistics across the dataset, in order to generate intuitions that are not just anecdotal, but are supported by the statistics computed on the whole dataset. Specifically, we propose two methods. The first one, *sub-explanation counting*, systematically searches for minimally-sufficient explanations of all images and count the amount of sub-explanations for each network. The second one, called *cross-testing*, computes salient regions using one network and then evaluates the performance by only showing these regions as an image to other networks. Through a combination of qualitative insights and quantitative statistics, we illustrate that 1) there are significant differences between the salient features of CNNs and attention models; 2) the occlusion-robustness in local attention models and global attention models may come from different decision-making mechanisms.

1 INTRODUCTION

As attention-based Transformer networks demonstrate outstanding performance in image recognition tasks (Dosovitskiy et al., 2021; Liu et al., 2021; Graham et al., 2021; Dai et al., 2021; Xie et al., 2021), comparisons between Transformer and CNNs attract more interest. Prior work Bhojanapalli et al. (2021); Naseer et al. (2021) have illustrated interesting differences between CNNs and transformers. However, there are still interesting questions that have not been answered, such as, are transformers finding fundamentally different salient features than CNNs in their classifications? Do they have fundamentally different inner working mechanisms? Why are some transformers seemingly more robust than CNNs? Are there differences in the transformers that utilize local attention versus the transformers that utilize global attention? Better answers to those questions would help us to gain more insights into those deep and complicated black-box networks, design better architectures, and understand whether they are suitable to be adopted in certain scenarios.

In this paper, we propose a novel methodology to examine those questions through the usage of *deep explanation algorithms*. Explanation algorithms have significantly improved in the past few years and can generate accurate explanations that can be verified through *intervention experiments* on the images (Samek et al., 2016; Petsiuk et al., 2018). The verification approach usually involves masking the images on certain areas deemed as important or salient by the explanation algorithm and checking the predicted class-conditional probability (hereafter referred to as the *classification confidence*) on the masked images. Evaluation metrics generated from these approaches put explanations on more solid footings. More recently, Shitole et al. (2021) proposed to search for explanations on each image using a beam search algorithm at a low resolution, hence reducing the chance of missing out on good explanations due to local optima in the optimization and allowing to discover different explanations on each image.

Our proposed approach is to systematically apply those explanation algorithms on a dataset-wide basis and examine both the qualitative visualizations and quantitative differences on *dataset-wide statistics* among various deep visual recognition backbones. With this approach, we can obtain insights that are both intuitive, coming from visualizations on individual images, and at the same time no longer merely anecdotal, but statistically relevant and verifiable.

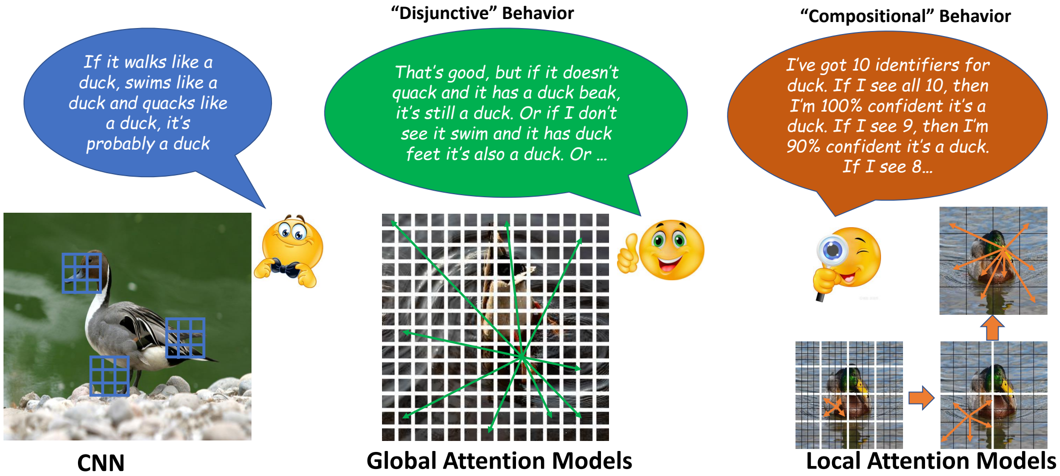

"Disjunctive" Behavior
"Compositional" Behavior
If it walks like a duck, swims like a duck and quacks like a duck, it's probably a duck
That's good, but if it doesn't quack and it has a duck beak, it's still a duck. Or if I don't see it swim and it has duck feet it's also a duck. Or ...
I've got 10 identifiers for duck. If I see all 10, then I'm 100% confident it's a duck. If I see 9, then I'm 90% confident it's a duck. If I see 8...
CNN
Global Attention Models
Local Attention Models

Figure 1: Intuitive mental models of different behaviors exhibited by different classes of models. CNN classifications are more often based on observing fixed combinations of features. Global attention models, where each patch can attend to any other patch, more often exhibit *disjunctive* behavior in that their high-confidence decisions are more likely to be based on an OR relationship among combinations of a few patches. Local attention models, where each patch attend only to a local window, tend to exhibit more often what we call *compositional* behaviors where high prediction confidence is built up by simultaneously seeing many patches. Dropping some of these patches will reduce the confidence, but the model would still be able to correctly classify the image.

Specifically, we propose two methods. The first is *sub-explanation counting*. For given explanations, defined as masks so that the masked images would generate similar predictions as the full image, we systematically remove areas from those explanations and count the number of explanations that, after the removal of these areas, can still lead to classification confidences that are higher than certain thresholds. Higher numbers tend to mean a type of behavior we name as *compositional*, in that the predictions are built on jointly considering multiple local areas (Fig. 1 Right).

The second methodology is *cross-testing*, where we first generate an attribution map (heatmap) on one recognition model, and then input masked images based on that heatmap to another model. This helps us to understand whether specific local areas (hence the visual features in those areas) that contribute significantly to one type of model would be still relevant to another one. If two models rely on similar visual features, then they should score highly in this cross-testing regime. On the other hand, if one model does not respond to the visual features that are deemed important to another model, that would mean that they are relying on potentially different features.

With these methodologies, we performed experiments on several types of deep visual recognition backbones: CNNs (Simonyan & Zisserman, 2015; He et al., 2016; Wightman et al., 2021; Liu et al., 2022), global attention models (Touvron et al., 2021; Graham et al., 2021), and local attention models (Liu et al., 2021; Zhang et al., 2022). By sub-explanation counting, we show that local attention models, such as Swin-Transformer and Nested-Hierarchical-Transformer are more *compositional* than other models. On the other hand, global attention models such as DeiT and LeViT exhibits more *disjunctive* behavior, in that they allow multiple different combinations of a smaller amount of parts to exhibit high predictive confidence. CNNs are more compositional rather than disjunctive, yet it is less robust to partial occlusions as local attention models. Besides, through cross-testing, we showed that those models do not always rely on the same type of features, as some of the salient features of one model would fail to activate other ones. Different models within the same model class are more similar to each other and the contrast between different model classes are higher. Some intuitions of these mental models are shown in Fig. 1. Those insights are going to be explained in more detail in the experiments (see Sec. 4).

2 Related Work

Multiple Explanations for Each Decision Ribeiro et al. (2018) suggested multiple explanations might exist for the decisions made by the deep neural networks. Carter et al. (2019) proposed sufficient input subsets that their observed values alone is sufficient in order to obtain similar output as the original input. They used instance-wise backward selection method in order to obtain such subsets. Shitole et al. (2021) proposed *Structured Attention Graphs* (SAG) which can generate multiple explanations by using beam search to find patch combinations. Among their findings, they showed 30% of the images from ImageNet can be explained with more than one patch combination, helping human users to obtain a better mental model than the single explanations provided by a attribution maps. Multiple explanations generated from SAG can be shown in a visualization tree where each node represents an explanation leading to certain classification confidence.

Explanation Using Attribution (Heat) Maps Attribution maps (heat maps) are some of the earliest and most widely-studied explanation tools for deep networks. They assign an *attribution* score to each input feature contributing 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 (Zeiler & Fergus, 2014; Selvaraju et al., 2017; Springenberg et al., 2015; Sundararajan et al., 2017; Smilkov et al., 2017; Bach et al., 2015). Such methods consider infinitesimal changes to the input which are not necessarily changing the output. Later on, sanity check procedures have shown that most of the gradient-based explanation methods are independent of the model predictions and mainly work as edge detectors which greatly undermined their credibility (Adebayo et al., 2018; Nie et al., 2018). There are also concerns about whether they are indeed interpretable by humans (Zimmermann et al., 2021). (Hooker et al., 2019) re-trains the model based on the heatmaps which is different from post-hoc heatmaps.
Further, *perturbation-based* approaches directly perturb the image regions such as in Ribeiro et al. (2016) which works on superpixels. Most of such approaches optimize for a real-valued mask over the input features in order to find the salient regions that significantly decrease the output (Petsiuk et al., 2018; Fong et al., 2019). However, due to the iterative process of such methods, they are relatively slow to generate. In addition, optimization for the mask is a highly non-convex problem and can easily be stuck in a bad local optimum. I-GOS (Qi et al., 2020) alleviated such issues by using the integrated-gradient as the descent direction rather than the gradient, which allows for faster convergence and finding better solutions to the optimization problem. iGOS++ (Khorram et al., 2021) improved over Qi et al. (2020) by incorporating bilateral minimal evidence, i.e., finding minimal regions over the input that influence the output the most when kept alone or removed. This largely avoided generating *adversarial* masks that solely rely on breaking the input features in order to reduce the output confidence – those are easier to locate, but they do not necessarily explain the decision-making of visual recognition models.

Understanding Transformers Recently, several works have explored the robustness of ViTs against CNNs under common perturbations (Bhojanapalli et al., 2021; Paul & Chen, 2022; Naseer et al., 2021). Bhojanapalli et al. (2021) found that ViT models, pre-trained with a large enough data, are at least as robust as the ResNet for natural and adversarial perturbations. Also, Naseer et al. (2021) found that ViTs are much more robust to occlusions than the ResNet50, and that DeiT-S keeps 70% accuracy whereas ResNet50 just has 0.1% accuracy for ImageNet when 50% of images regions are randomly removed. Mahmood et al. (2021) studies the adversarial robustness of ViTs. Raghu et al. (2021) studies the differences in the visual representations learned from ViTs and CNNs and how global and local information are utilized between lower and higher layers. Zhou et al. (2022) further studies the role of self-attention in improved robustness of vision transformers. Different from previous work, our paper seeks to further understand the *mechanism* of occlusion handling in transformers using explanation methods and points out that local and global attention models might utilize different mechanisms.

3 Methods

3.1 Minimal Sufficient Explanations and Structural Explanations

Shitole et al. (2021) showed that a single explanation provided by heatmaps does not provide a complete understanding of the decision-making of the deep network. It proposed a more comprehensive way to find explanations by using beam search at low resolutions in order to systematically find various combination of image regions that lead to high classification confidence given each image. Given an image I divided into non-overlapping patches p_i and a target class c, Shitole et al. (2021) called each image patch p_i a *literal*. A conjunction N of a set of literals establishes an image region composed of the union of them and the confidence of the classifier f w.r.t. target class c for the conjunction N can be evaluated by $f_c(N)$, the predicted class-conditional probability for class c on the region. A Minimal Sufficient Explanation (MSE) is defined as a minimal conjunction/region that achieves a certain high classification confidence P_h, i.e., $f_c(N_i) > P_h f_c(I)$, 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. We will also call them "explanations" in the rest of the paper. MSEs are not unique and *beam search* can be used to efficiently find them. The search objective is to to find all N_i such that they achieve a confidence higher than a threshold P_h where no sub-regions in $n_j \subset N_i$ exceed that threshold,

$$f_c(N_i) \geq P_h f_c(I), \max_{n_j \subset N_i} f_c(n_j) < P_h f_c(I) \tag{1}$$

In order to visualize such explanations, Shitole et al. (2021) proposed using Structured Attention Graph (SAG) which are directed acyclic graphs over attention maps of different image regions. This provides a logical structure in the form of disjunctions of conjunctions of features presented in the image. Examples of SAGs can be seen in Fig. 4.

3.2 Intervention Experiments as Verification of Explanation Algorithms

Evaluation of explanation algorithms is quite challenging as there is no ground truth. A recent approach is to *causally* evaluate a given local explanation by perturbing the input features (according to a provided attribution map) and observing its influence over the output. Samek et al. (2016) 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 patches (in MoRF) and least relevant patches (in LeRF) are gradually replaced by random noise sampled from a uniform distribution. Then, the perturbed images are passed through the model and their classification confidences are obtained. Similarly, Petsiuk et al. (2018) proposed the *deletion* and *insertion* metrics with the main difference in 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 given model 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. 2). A low deletion score conveys a sharp drop in the output confidence after successive perturbation of the most salient regions. Oppositely, 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, these evaluation schemes can be done automatically and do not require human-defined labels/bounding boxes (Zhang et al., 2016), hence easing quantitative evaluation at large scales. We use deletion and insertion scores in our evaluations. 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 deletion metric as,

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

$$\phi^{(t)}(\tilde{I}, I, M) = \tilde{I} \odot M^{(t)} + I \odot (\mathbf{1} - M^{(t)})$$

where $f_c(.)$ is the output class-conditional probability of the classifier for class c, 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. This simply implies that the $\phi^{(0)} = I$ and $\phi^{(T)} = \tilde{I}$. The Insertion score is calculated similarly where the input image I and the baseline image $\tilde{I}$ are swapped. In the experiments, the target class is chosen as the predicted class: $c = \arg\max_c f(I)$.

3.3 Sub-Explanation Counting

Firstly, we use the SAG (Shitole et al., 2021) methodology that provides a more comprehensive explanation over the predictions of the classifiers compared to attribution map methods. We use their beam search algorithm to find a comprehensive set of Minimal Sufficient Explanations (MSEs). This gives us a holistic insight into how CNNs and Transformers make predictions.

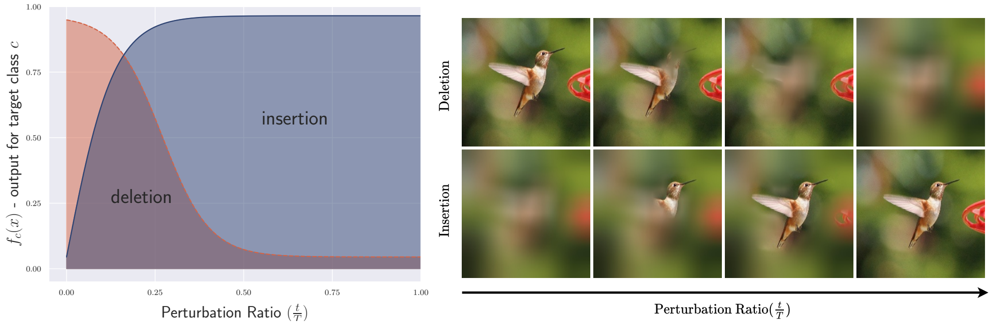


Figure 2: Illustration of the Insertion/Deletion metrics. Left: The Area under the curves (AUCs) are used to compute the deletion/insertion metrics; Right: Deletion and insertion images are obtained by masking: successive removing/revealing of pixels that are deemed most salient by the heat map.

In order to further explore which models are more robust to occlusions, 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 the tree. In the meantime, we keep evaluating the confidence of current nodes (proper subsets of MSE) on the model that is used to generate the MSE. When the nodes are with a confidence less than 50% compared to the classification confidence on the original image, we stop expansion. Afterward, we calculate the number of nodes that have classification confidence above several thresholds. This method can evaluate *compositionality*, in that it assesses whether partial explanations still retain high classification confidence.

3.4 Cross-Testing

The second method we propose is to utilize one specific deep model to generate an attribution map, and use another deep model to assess the insertion/deletion metrics by successively masking the images based on the attribution map. For fair comparison across different models which may have different average classification confidences on the dataset, we normalize the scores based on the average top-1 classification confidence and the average confidence on the fully blurred image for each model by $\bar{s} = (s - b)/(t - b)$ where s is the score and t/b are the top-1/fully-blurred confidences (Schulz et al., 2020), respectively. This method assesses the similarity of different models under occlusion. With a pairwise similarity matrix generated, we can then utilize dimensionality reduction approaches such as kernel principle component analysis (Scholkopf et al., 1999) to visualize them in a 2D space. This gives insights about which approaches are classifying more similarly.

4 Experiments

We use ResNet50 (He et al., 2016), ResNet50-C1 (Wightman et al., 2021), ResNet50-C2 (Wightman et al., 2021), ResNet50-D (Wightman et al., 2021), VGG19 (Simonyan & Zisserman, 2015), ConvNeXt-T (Liu et al., 2022), Swin-T (Liu et al., 2021), Nest-T (Zhang et al., 2022), DeiT-S (Touvron et al., 2021), DeiT-S-distilled (Touvron et al., 2021) and LeViT-256 (Graham et al., 2021) in our experiments. Of these, ResNet50 and VGG19 are famous CNN models. ResNet50-C1, ResNet50-C2 and ResNet50-D are ResNet50 variants with high performance using difference methods to train. ConvNeXt-T is hybrid model based on large-kernel depthwise convolutions. DeiT-S and LeViT-256 are global transformer models where each patch can attend to all patches in the image, Swin-T and Nest-T are local attention models where each patch can only attend to its local window (Fig. 1). DeiT-S-distilled and LeViT-256 were trained by distilling from a teacher CNN while DeiT-S was without this distillation. We chose them because they have similar sizes and similar performances among the transformer models. Table 3 shows their size and top-1 accuracy on ImageNet. In addition, we also train ResNet50-A2, Swin-T, DeiT-S, DeiT-S-distilled and LeViT-256 with the same procedure but distinct seeds to do test. Table 4 shows their mean with standard deviation of top-1 accuracy on ImageNet. For the most experiments, we use pre-trained models on ImageNet-1K, only the seed experiments we use the models trained by ourselves. In these seed experiments, we select the checkpoints in the last epoch not the epoch with maximum accuracy. In most experiments, we use the first 5,000 images from ImageNet validation dataset (Deng et al., 2009) due to the slow speed of running all the experiments.

4.1 Counting the Explanations and Sub-Explanations

We follow Shitole et al. (2021) to perform a beam search on different patch combinations when the image is divided into a 7×7 grid with 49 patches. The beam search width was set to 5 to balance speed and performance.

First, we count the number of MSEs among different networks (Table 1). It can be seen that most networks have a similar median number of MSEs. However, DeiT-S, DeiT-S-distilled and LeViT-256 have a higher mean, showing that in some images there are significantly more MSEs from these approaches, which means that the networks can look at multiple different combinations of regions to arrive at similarly confident predictions, and indicative of a more *disjunctive* behavior. If using a logical analogy, one can think about the logic exhibited by these models as the result of the **OR** relationship among a large number of different conjunctions.

On the other hand, the local attention models Swin-T and Nest-T exhibit significantly different behavior. As Table 1 shows, they have significantly more sub-explanations than all the other models. From Fig. 3 and Fig. 4, one can see that Swin-T usually has significantly larger MSEs than the other models as well. These pieces of evidence suggest that Swin-T may be more *compositional*, in that its predictions are built by simultaneously taking into account the contributions of many different parts (in a sub-additive manner). All these parts contribute to the decision in a way that if some of the parts are eliminated, the prediction confidence lowers, but not enough to misclassify the example.

Fig.3 shows the distribution of MSE sizes in a few random images. As one can see, in some of the images, all models have a similar amount of MSEs. However, in some images, the global attention models have a significantly higher amount of MSEs. For example, LeViT-256 has 26 MSEs in the `hotdog` image, much more than other types of methods which usually have $12 - 15$ MSEs. DeiT-S-distilled has 46 MSEs in the `Spoonbill` image, much higher than local attention models and CNNs that have $11 - 17$ MSEs. Although different images show different trends, the sizes of the MSEs are often smaller in the global attention models and larger in the local attention model Swin-T. The larger sizes of Swin-T MSEs also partially explain the significantly higher number of confident subexplanations shown in Table 1, and further point to its compositional behavior. More results and visualizations can be found in the supplementary material.

We would like to obtain the standard error of all the approaches to determine statistical significance of these differences, however we have very limited GPU resources. Hence, we choose to train several representative models with different random seeds to see if the differences we observe are significant. The results shown in Table 2 confirms that the two main results are statistically significant: 1) global attention models DeiT-S, DeiT-S-distilled and LeViT-256 have a higher mean of MSEs; 2) local attention model Swin-T have more sub-explanations than other models.

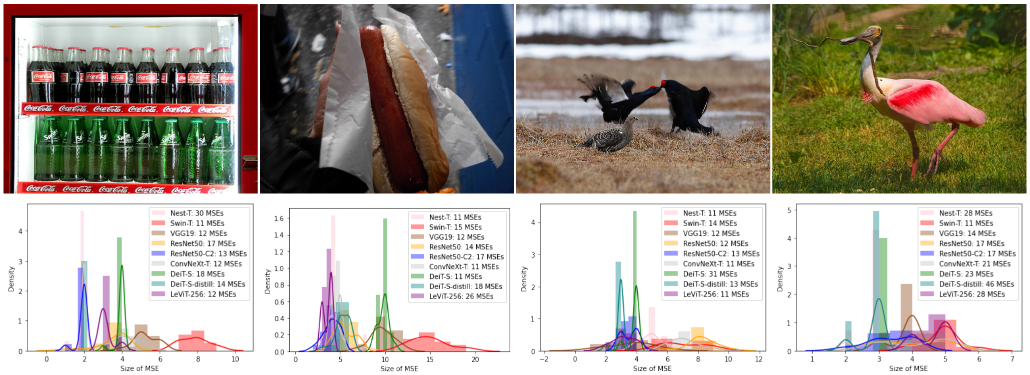
4.2 Cross-Testing with Heatmap Visualizations

We use the state-of-the-art attribution map method iGOS++ (Khorram et al., 2021) to generate heatmaps for each image at 28×28 resolution. This resolution is chosen because this is the highest resolution that iGOS++ has consistently good performance among different networks. We then calculate the insertion and deletion scores based on the obtained heatmap values. The normalized deletion and insertion scores obtained from different classifiers are presented in the Appendix. In order to better visualize these similarities, we used Kernel PCA to project them to 2 dimensions (Scholkopf et al., 1999), based on the similarities of the insertion scores. Figure 5 shows the projection results. It can be found that the same type of model is more similar, i.e. they use similar features for their predictions. CNNs (VGG19 and ResNet50) are closer to each other, 3 ResNet50 variants (ResNet50-C1, ResNet50-C2 and ResNet50-D) are closer to each other, global attention models (DeiT-S,DeiT-S-distilled and LeViT-256) are also closer to each other. One can see three clear lines from center delineating different type of methods. ConvNeXt-T is an interesting outlier that seems to be in a league of its own, having some similarities with local attention transformers and ResNets, but still very different. Another interesting thing is that the distillation (from a CNN) did bring DeiT-S closer to the center, and the additional data augmentation strategies in ResNet50-C1/C2/D also bring them closer to the center.

Figure 9 shows qualitative results from cross-testing, where partially occluded images were generated using the top-ranked patches from the iGOS++ heatmap on one model, and then the masked

Table 1: Results of beam search to locate MSEs. Confidence represents average amount of nodes with a classification confidence higher than that threshold w.r.t. the confidence of the whole image

	Number of MSEs			Confidence				
Model	**Mean**	**Std**	**Median**	**$\geq$ 90%**	**$\geq$ 80%**	**$\geq$ 70%**	**$\geq$ 60%**	**$\geq$ 50%**
VGG19	14.10	4.97	12.00	3.19	13.86	57.59	158.42	439.52
ResNet50	13.52	4.42	12.00	3.21	11.17	40.70	145.19	373.76
ResNet50-C1	13.40	3.79	12.00	**3.39**	12.62	52.45	163.72	390.93
ResNet50-C2	13.72	4.13	12.00	3.36	12.34	45.38	152.36	373.73
ResNet50-D	13.81	4.43	12.00	3.27	14.07	42.43	104.91	294.52
ConvNeXt-T	14.54	4.85	13.00	3.23	12.52	49.84	181.41	433.55
Swin-T	14.66	5.12	13.00	3.16	**21.01**	85.19	255.47	601.72
Nest-T	15.39	6.05	13.00	3.28	17.41	**101.45**	**284.26**	**673.97**
DeiT-S	15.85	6.23	**14.00**	3.31	8.93	28.37	102.15	306.71
DeiT-S-dis	**16.53**	**7.12**	**14.00**	3.19	13.39	39.35	94.12	236.84
LeViT-256	15.56	6.03	13.00	3.18	10.05	32.27	81.80	205.42



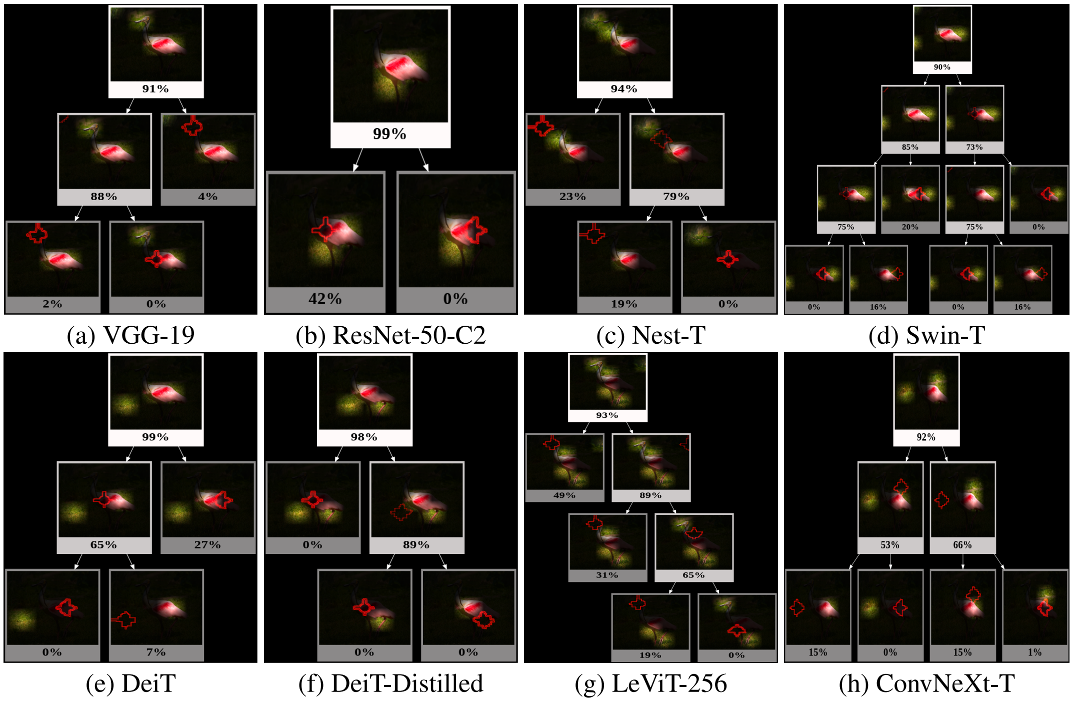
(a) Pop Bottle (b) Hotdog (c) Black Grouse (d) Spoonbill

Figure 3: A few example distributions of MSE sizes for different networks on random images. It can be seen that global attention networks such as DeiT and LeViT tends to have smaller MSEs but in some images they have significantly more MSEs than other networks, whereas Swin Transformers more often have larger MSEs than other networks

image is inputted to all the six models to evaluate their prediction confidences on the predicted class by the model used to generate the heatmap. From those images one can see that global attention models sometimes obtain high confidence with a small number of regions shown. An example is the Alp image in the last column, with a heavily occluded Alp where Swin-T only predicted with a confidence of 23.51%, LeViT-256 already obtained a confidence of 80.41% and DeiT-S-distilled has a higher confidence of 86.76%. In the Bakery image in the second column, the heatmap is generated on Swin-T, however, both LeViT-256 and DeiT-S-distilled have higher confidence than Swin-T, showing that they required less information to obtain more confident predictions. The CNNs fared poorly in some of these partially occluded cases, indicating that maybe their predictions rely too heavily on the existence of certain features that were occluded.

Table 2: Beam search results to locate MSEs of seed experiments. Confidence represents average amount of nodes with a classification confidence higher than that threshold w.r.t. the confidence of the whole image. The left side of the symbol ± is mean value and the right side is standard deviation during the same model with different seeds

		Confidence				
Model	**Number of MSEs**	**$\geq$ 90%**	**$\geq$ 80%**	**$\geq$ 70%**	**$\geq$ 60%**	**$\geq$ 50%**
ResNet50-A2	14.03 ± 0.12	3.32 ± 0.01	10.69 ± 2.59	37.85 ± 11.13	101.20 ± 25.77	244.51 ± 55.30
Swin-T	14.40 ± 0.17	3.11 ± 0.04	**17.70** ± 1.96	**85.24** ± 13.06	**267.39** ± 44.46	**757.04** ± 82.58
DeiT-S	**16.75** ± 0.13	3.33 ± 0.01	9.75 ± 0.55	34.95 ± 4.62	113.38 ± 21.64	298.51 ± 66.14
DeiT-S-dis	**16.51** ± 0.52	3.19 ± 0.00	11.51 ± 1.02	34.69 ± 1.02	98.94 ± 9.86	262.06 ± 31.53
LeViT-256	15.68 ± 0.25	3.33 ± 0.00	8.67 ± 1.25	23.31 ± 4.56	51.36 ± 9.84	125.78 ± 18.24



(a) VGG-19 (b) ResNet-50-C2 (c) Nest-T (d) Swin-T

(e) DeiT (f) DeiT-Distilled (g) LeViT-256 (h) ConvNeXt-T

Figure 4: MSEs and some sub-explanations of different models on an image of the Spoonbill class. Note that due to the space limit we only subsampled a few subexplanations. The removed patch from the higher level of the tree is indicated with a red outline. (Best viewed in Color)

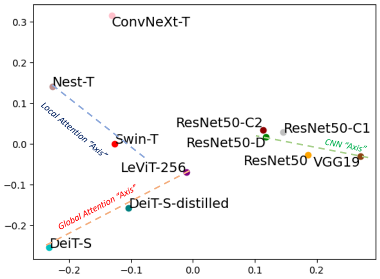
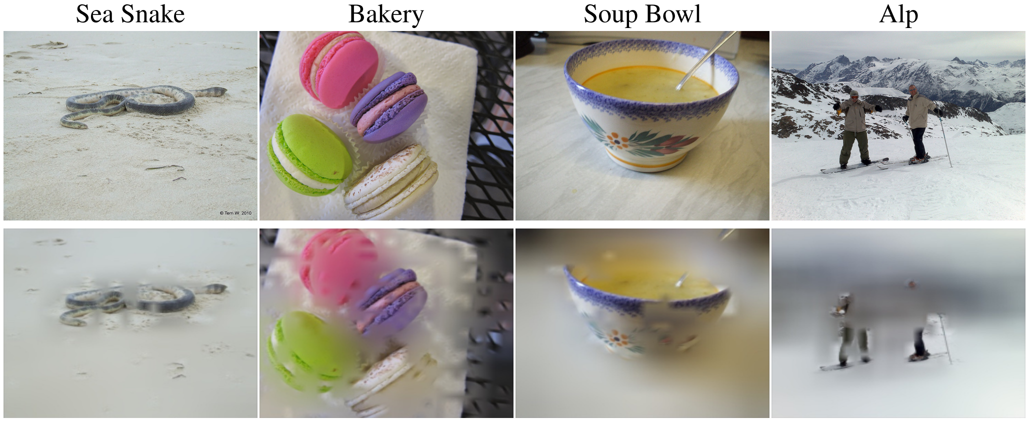


Figure 5: Kernel PCA projections of different models using the insertion metrics. We have drawn hypothetical dashed "axes" indicating different types of methods. One can see that models that are the same type are more similar to each other in this plot, and that distillation brings DeiT closer to CNNs

4.3 Discussion and Limitations

One can see that global attention methods and local attention methods both handle occlusions, but they potentially utilize different mechanisms. Global attention methods such as DeiT-S/LeViT are robust to occlusions because the models only need to see a few of the parts to obtain a confident prediction, hence, they are not dependent on any specific part to be observable. This may be linked to the global attention mechanism which can easily attend to all parts of the image (Fig. 1) and quickly locates correlated parts even if they are far away. Interestingly, distillation seems to often *reduce* the size of the MSEs in DeiT, which leads it away from building decisions using more parts. It is interesting to ponder why distillation tends to lead these models to prefer outputting higher confidence with fewer parts.

On the other hand, in local attention models such as Swin-T and Nest-T, missed parts indeed compromise the predictions, but because they base their decisions on many parts simultaneously, the reduction of confidence from a few missing parts is not very large, hence the model can still be robust to occlusions. This could potentially be traced back to the local attention mechanism, which



Prediction Confidence on the Partially Occluded Image

VGG19	ResNet50	Swin-T	VGG19	ResNet50	Swin-T	VGG19	ResNet50	Swin-T	VGG19	ResNet50	Swin-T
0.0494	0.0731	**0.8593**	0.0775	0.0794	**0.8345**	0.0135	**0.8630**	0.6054	0.0431	0.1940	0.2351
ResNet50-C1	ResNet50-C2	ResNet50-D	ResNet50-C1	ResNet50-C2	ResNet50-D	ResNet50-C1	ResNet50-C2	ResNet50-D	ResNet50-C1	ResNet50-C2	ResNet50-D
0.0088	0.2988	0.8058	0.4471	0.4752	0.9164	0.2572	0.4448	0.1875	0.0506	0.7287	0.2566
DeiT-S	DeiT-S-dis	LeViT-256	DeiT-S	DeiT-S-dis	LeViT-256	DeiT-S	DeiT-S-dis	LeViT-256	DeiT-S	DeiT-S-dis	LeViT-256
0.3048	0.7156	0.6570	0.5952	0.8857	0.9445	0.7128	0.7912	0.7985	0.0303	0.8676	**0.8041**
ConvNeXt-T			ConvNeXt-T			ConvNeXt-T			ConvNeXt-T		
0.5609			0.1447			0.6013			0.0960		

Figure 6: Qualitative Cross-Testing Results. The partially occluded images were generated using iGOS++ heatmaps on the algorithm with bolded number (not necessarily the highest). Then the same image is tested on multiple networks and we show predicted class-conditional probabilities on the ground truth class (written above). It can be seen that global attention models sometimes obtain high confidence when shown a small combination of regions, and CNN confidences vary greatly depending on whether certain features were observed. (Best viewed in color)

in lower layers cannot easily find correlated parts that are far away (Fig. 1), but they would gradually build part compositions up and at the find stages find all the information for decision-making.

CNN behaves more similarly to local attention methods than global attention methods, but it tends to be less compositional than Swin-T and may rely on more specific parts to build their decisions (Fig. 3). The inability of CNNs to utilize more relevant parts in their decisions may partially explain their lower robustness to occlusion than transformers.

Limitations There are several limitations of this work. First, due to time constraints, we have only generated the heatmaps using one heatmap algorithm. Ideally, more heatmap algorithms can be included. But iGOS++ has obtained very high insertion scores that are close to 100% (Table 7), hence we have reasons to believe that the qualitative result will hold up. Besides, the MSE beam search approach we utilized in Sec. 4.1 is independent of any heatmap algorithm. A second limitation is that we could not fully present the tendencies of each individual image. It is still important to note that what we present are general trends that do not necessarily hold on every single image. To remedy this, we present some qualitative examples that do not follow our claimed trends in the appendix. A third limitation is that although we discussed the potential reasons that could link the observed trends to the model architectures, we have not fully proven that these trends are indeed caused by the architecture. This is a difficult question and we would leave it to future work to pursue.

5 Conclusion

In this paper, we proposed a novel methodology of utilizing deep explanation algorithms to collect dataset-wide statistics and combine it with the insights from qualitative visualizations. Specifically, we propose the methodologies of sub-explanation counting and cross-testing to assess the decision-making behaviors of different classes of models. Our quantitative and qualitative analysis indicate that different types of visual recognition models exhibit quite different behaviors in their decision-making, where local attention models are more likely to exhibit a compositional behavior and global attention models are more likely to exhibit a disjunctive behavior, which could be potentially linked to their model architectures. It may provide insights and inspire future model designs.

REFERENCES

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*, pp. 9505–9515, 2018.

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, 7 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0130140.

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.

Brandon Carter, Jonas Mueller, Siddhartha Jain, and David Gifford. What made you do this? Understanding black-box decisions with sufficient input subsets. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 567–576. PMLR, 16–18 Apr 2019. URL `https://proceedings.mlr.press/v89/carter19a.html`.

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.

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.

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.

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*, pp. 2950–2958, 2019.

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*, pp. 12259–12269, 2021.

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)*, June 2016.

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.

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.

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*, pp. 10012–10022, 2021.

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.

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)*, pp. 7838–7847, October 2021.

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.

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

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*, pp. 2071–2081. AAAI Press, 2022. URL https://ojs.aaai.org/index.php/AAAI/article/view/20103.

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.

Zhongang Qi, Saeed Khorram, and Li Fuxin. Visualizing deep networks by optimizing with integrated gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2020.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 12116–12128, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/652cf38361a209088302ba2b8b7f51e0-Abstract.html.

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*, pp. 1135–1144. ACM, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

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.

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

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

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)*, pp. 618–626, 2017.

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*, volume 34, 2021.

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

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

J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for Simplicity: The All Convolutional Net. In *ICLR Workshop*, 2015. URL `http://lmb.informatik.uni-freiburg.de/Publications/2015/DB15a`.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.

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*, pp. 10347–10357, 2021.

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

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.

Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 818–833, Cham, 2014. Springer International Publishing.

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

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.

Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M. Alvarez. Understanding the robustness in vision transformers. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27378–27394. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/zhou22m.html`.

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.

A Appendix

A.1 Information of the models used in our experiments

Table 3: Information of the models used in our experiments along with the number of learnable parameters and the Top-1 accuracy on ImageNet-1K.

Model	**VGG19**	**ResNet50**	**ResNet50-C1**	**ResNet50-C2**	**ResNet50-D**	**ConvNeXt-T**	**Swin-T**	**Nest-T**	**DeiT-S**	**DeiT-S-distilled**	**LeViT-256**
Params	144M	25M	25M	25M	25M	28M	28M	17	22M	22M	19M
Top-1 acc	74.5	76.1	79.8	80.0	79.8	82.1	81.2	81.5	79.9	81.2	81.6

Table 4: The number of learnable parameters and the Top-1 accuracy on ImageNet-1K of the models we trained. The left side of the symbol ± is mean value and the right side is standard deviation

Model	**ResNet50-A2**	**Swin-T**	**DeiT-S**	**DeiT-S-distilled**	**LeViT-256**
Params	25M	28M	22M	22M	19M
Top-1 acc	79.73 ± 0.15	81.09 ± 0.05	79.72 ± 0.07	80.94 ± 0.16	78.76 ± 0.04

A.2 More Results on Minimal Sufficient Explanations and Their Sub-Explanations

In Sec. 4, we only showed the number of MSEs obtained by different models. Table 5 shows that the beam search statistics on the number of diverse MSEs by allowing for different degrees of overlap. These results are obtained from 50,000 images of ImageNet validation dataset. We see that in the case of either no overlap is allowed, or 1 overlap is allowed, LeViT-256 and DeiT-S-distilled tend to have more diverse MSEs for classification.

Fig 7 show the distributions of MSE sizes for different algorithms over the first 5,000 images from ImageNet validation dataset. We find that global attention algorithms such as DeiT and LeViT tends to have smaller MSEs on more images. The local attention algorithm, Swin-T, tends to have larger MSEs. But CNNs seem to have an even larger proportion of larger MSEs.

We also follow Shitole et al. (2021) to plot the percentage of images that can be explained with a small amount of patches. For each number of patches n, we plot the total proportion of images that contain at least one MSE with size $\leq n$. For Fig. 8, we use 50,000 images from ImageNet validation dataset to be consistent with Shitole et al. (2021), we found that the global attention models can explain the most amount of images given a small number of patches, compared with other models. This difference is significant when the number of patches is larger than 3. LeViT and DeiT-S-distilled has at least one MSE less than or equal to 10 patches (about 20% of the area of the full image) on more than 95% of the images. This again shows that these global attention models can be confident with only a few patches.

As we mentioned, what were presented in the paper were general trends rather than holding true for all the images. Fig 10 showed 2 additional examples distributions of MSE sizes for different algorithms. It can be seen that, for `hotdog`, LeViT-256 has larger MSEs than all the other algorithms. In `Silky Terrier`, ResNet50 has the most MSEs and they are relatively small. These plots show a more complete picture of what can happen in each individual image.

Finally, we show more visual examples of Structural Attention Graph (SAG) trees on several images with all the tested approaches (Fig. 11 - Fig. 37). Even if we only limited to showing 3 children per parent node, the size of the tree in Swin-T is huge, showing strong evidences that the predictions of Swin-T are built by simultaneously taking into account the contributions of many different parts. On the other hand, LeViT and DeiT have much smaller trees. What is not shown in this figure is that there are significantly more trees in those global attention models. We believe this further illustrates the stark difference between the decision-making mechanisms of these global attention models versus Swin Transformer.

A.3 More Results on Cross-Testing

In Section 4.2, we provide the the visualized results of different classifiers using Kernel PCA, which show the similarities between different models. Here we show the normalized deletion and inser-

Table 5: Number of diverse MSEs obtained by allowing for different degrees of overlap

Model	Overlap=0			Overlap=1		
	Mean	Std	Median	Mean	Std	Median
VGG19	1.27	0.69	1.00	2.08	2.12	1.00
ResNet50	1.25	0.66	1.00	1.98	1.96	1.00
ConvNeXt-T	1.20	0.52	1.00	1.89	1.75	1.00
Swin-T	1.26	0.67	1.00	2.12	2.04	1.00
Nest-T	**1.57**	**1.00**	1.00	**3.10**	**2.91**	**2.00**
DeiT-S	1.18	0.45	1.00	1.91	1.55	1.00
DeiT-S-distilled	1.41	0.88	1.00	2.57	2.60	1.00
LeViT-256	1.46	0.94	1.00	2.67	2.65	**2.00**

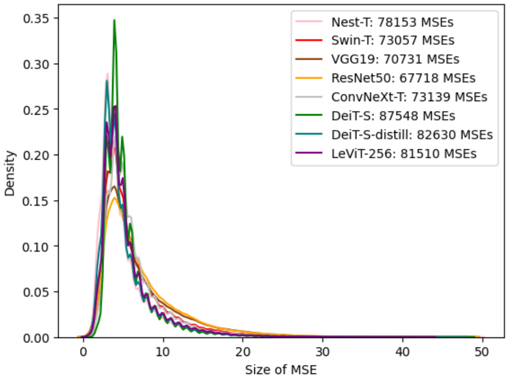


Figure 7: Distribution of MSEs over 5000 images.

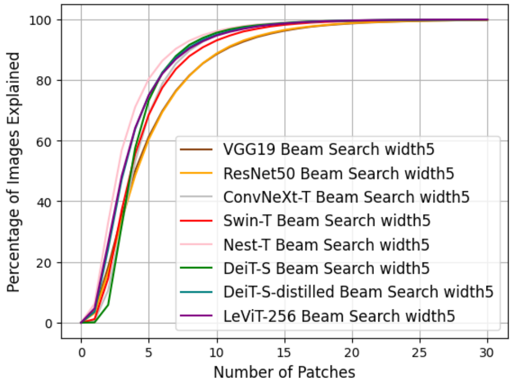


Figure 8: Percentage of images explained by different number of patches.

tion scores obtained from different classifiers.We can see that most of the algorithms have similar deletion scores and high relative insertion scores. This indicates that the heatmaps found by the algorithm explain the decisions consistently and the model is able to obtain similar confidence as the full image by only using a few top-ranked patches from the heatmap, which proves that the heatmap algorithm we use is sound as a basis for the cross-testing experiments. Note that the global attention models DeiT-S and LeViT-256 have slightly higher insertion scores which indicate they need fewer patches to achieve the same confidence as the full image than the other algorithms. DeiT-S even has a relative insertion score slightly higher than 1, indicating that in many cases, partially occluded images have been more confidently predicted than the full image.

From the results shown in Table 7 one can see the significant differences between different models. Cross-testing insertion metric is usually around 80%, which indicates that the models agree on less than 90% of the images. The similarity between models of the same category are usually higher, e.g. between VGG19 and ResNet50, and among DeiT-S, DeiT-S-distilled and LeViT-256, also during ResNet50-c1, ResNet50-c2 and ResNet50-d. The lower insertion metric in cross-testing between Swin-T and the global attention models indicate that they are quite different models. Still, the similarities between Swin-T and other global transformer models are higher than between Swin-T and the CNN models.

In Section 4.2, we stated that global attention models sometimes obtain high confidence while only showing a small number of regions. Here we provide more qualitative results from cross-testing, Fif.9. The `Cougar` image in the second column of the first row, the heatmap is generated on Swin-T, however, both DeiT-S (84.68%) and DeiT-S-distilled(96.96%) have higher confidence than Swin-T (77.39%) on this partially occluded image.

Also, we can see that in many cases even if the figures were generated by ResNet, the transformer models (especially the global attention transformers and less so the local attention transformers) have sometimes much higher confidences on occluded images. Especially, DeiT-S-distilled and LeViT more often have confidence than ResNet and VGG.

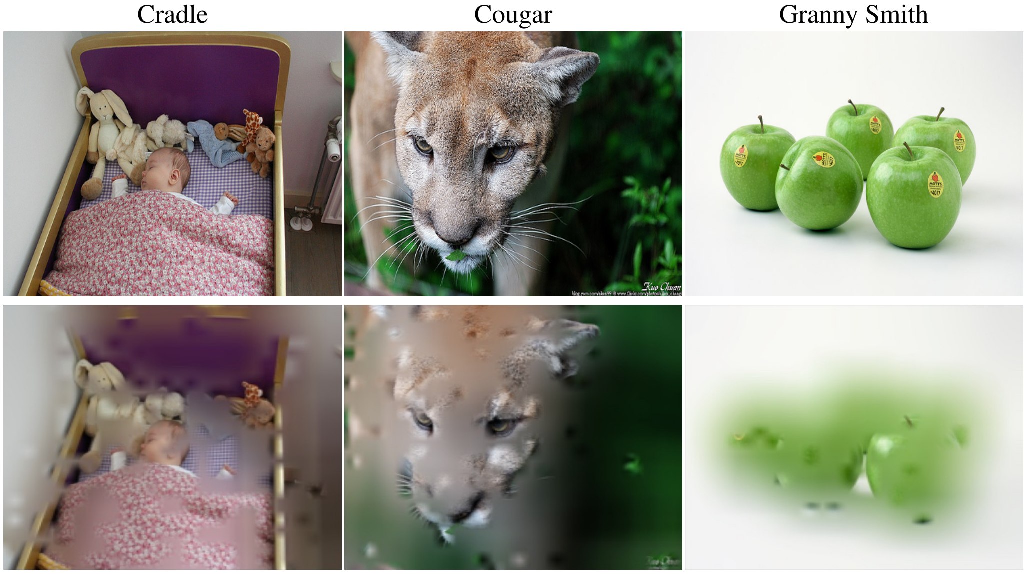
We also do the seed experiments for Cross-Testing. Table 6 show the quantitative results of Cross-Testing during the same model with different seeds. We can see, for the same model, the insertion score of Cross-Testing results with different seeds are not completely consistent, the standard deviation is not small, especially for Swin-T (0.052), DeiT-S (0.047) and LeViT-256 (0.041). However, the distillation makes the difference in DeiT-S smaller, the Ins standard deviation of DeiT-S-distilled is 0.019, lower than 0.047.

Table 6: Cross-Testing Results.

Model	**ResNet50-A2**	**Swin-T**	**DeiT-S**	**DeiT-S-distilled**	**LeViT-256**
Del	0.185 ± 0.012	0.161 ± 0.021	0.168 ± 0.028	0.122 ± 0.020	0.134 ± 0.028
Ins	0.885 ± 0.010	0.869 ± 0.052	0.926 ± 0.047	0.960 ± 0.019	0.931 ± 0.041

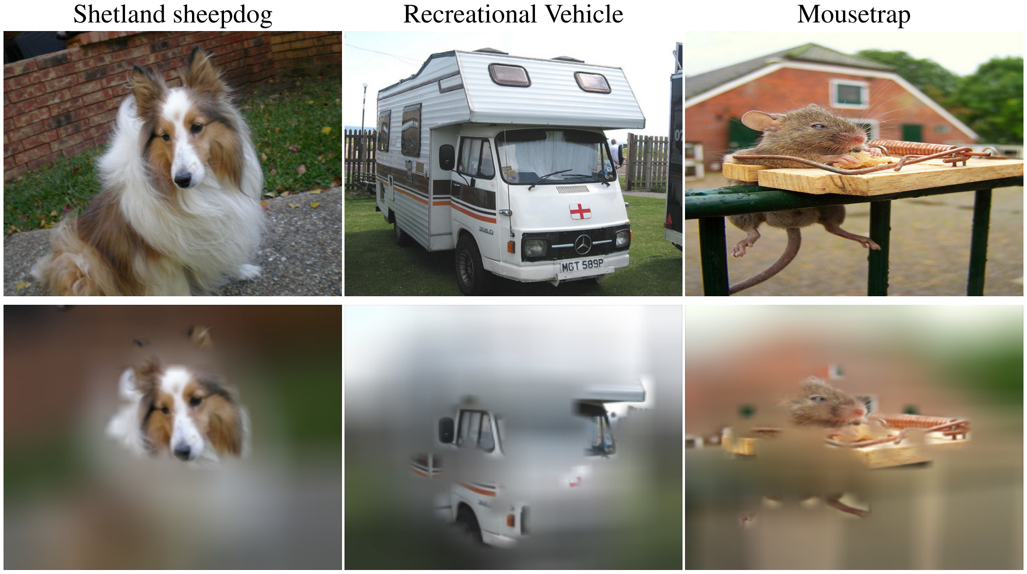
Table 7: Cross-Testing. Deletion/Insertion metrics when generating heatmaps using the model on the first row and evaluating the heatmaps by using the model on first column.

Generation →	VGG19		ResNet50		ResNet50-C1		ResNet50-C2		ResNet50-D		ConvNeXt-T		Swin-T		Nest-T		DeiT-S		DeiT-S-distill		LeViT-256	
Evaluation ↓	Del.	Ins.	Del.	Ins.	Del.	Ins.	Del.	Ins.	Del.	Ins.	Del.	Ins.	Del.	Ins.	Del.	Ins.	Del.	Ins.	Del.	Ins.	Del.	Ins.
VGG19	0.110	0.943	0.159	0.863	0.174	0.785	0.177	0.777	0.171	0.806	0.190	0.650	0.173	0.715	0.197	0.647	0.127	0.699	0.144	0.788	0.149	0.796
ResNet50	0.174	0.880	0.125	0.925	0.180	0.824	0.179	0.821	0.172	0.838	0.195	0.719	0.177	0.767	0.208	0.705	0.262	0.757	0.149	0.843	0.150	0.851
ResNet50-C1	0.203	0.854	0.193	0.850	0.169	0.910	0.195	0.838	0.192	0.851	0.214	0.745	0.198	0.770	0.232	0.717	0.223	0.763	0.169	0.832	0.167	0.839
ResNet50-C2	0.210	0.863	0.198	0.867	0.202	0.857	0.178	0.923	0.198	0.871	0.227	0.771	0.210	0.799	0.247	0.750	0.239	0.789	0.179	0.859	0.176	0.866
ResNet50-D	0.197	0.857	0.184	0.859	0.191	0.840	0.190	0.840	0.158	0.902	0.211	0.752	0.194	0.783	0.232	0.735	0.220	0.775	0.163	0.847	0.160	0.853
ConvNeXt-T	0.237	0.823	0.224	0.821	0.232	0.802	0.229	0.802	0.223	0.814	0.151	0.947	0.208	0.796	0.246	0.758	0.238	0.783	0.187	0.823	0.187	0.821
Swin-T	0.244	0.823	0.232	0.817	0.243	0.789	0.240	0.790	0.235	0.809	0.227	0.748	0.129	0.943	0.231	0.750	0.221	0.787	0.175	0.834	0.180	0.833
Nest-T	0.240	0.812	0.230	0.800	0.236	0.769	0.238	0.771	0.232	0.787	0.229	0.728	0.196	0.780	0.143	0.932	0.221	0.766	0.173	0.818	0.181	0.810
DeiT-S	0.244	0.852	0.231	0.854	0.248	0.830	0.248	0.827	0.238	0.843	0.246	0.788	0.204	0.848	0.240	0.798	0.126	1.008	0.143	0.945	0.178	0.896
DeiT-S-distill	0.244	0.852	0.231	0.872	0.246	0.850	0.246	0.850	0.236	0.866	0.245	0.797	0.204	0.849	0.244	0.801	0.203	0.890	0.095	0.994	0.173	0.911
LeViT-256	0.257	0.878	0.240	0.878	0.252	0.852	0.251	0.856	0.245	0.870	0.254	0.796	0.224	0.843	0.262	0.795	0.249	0.847	0.185	0.903	0.117	0.973



Prediction Confidence on the Partially Occluded Image

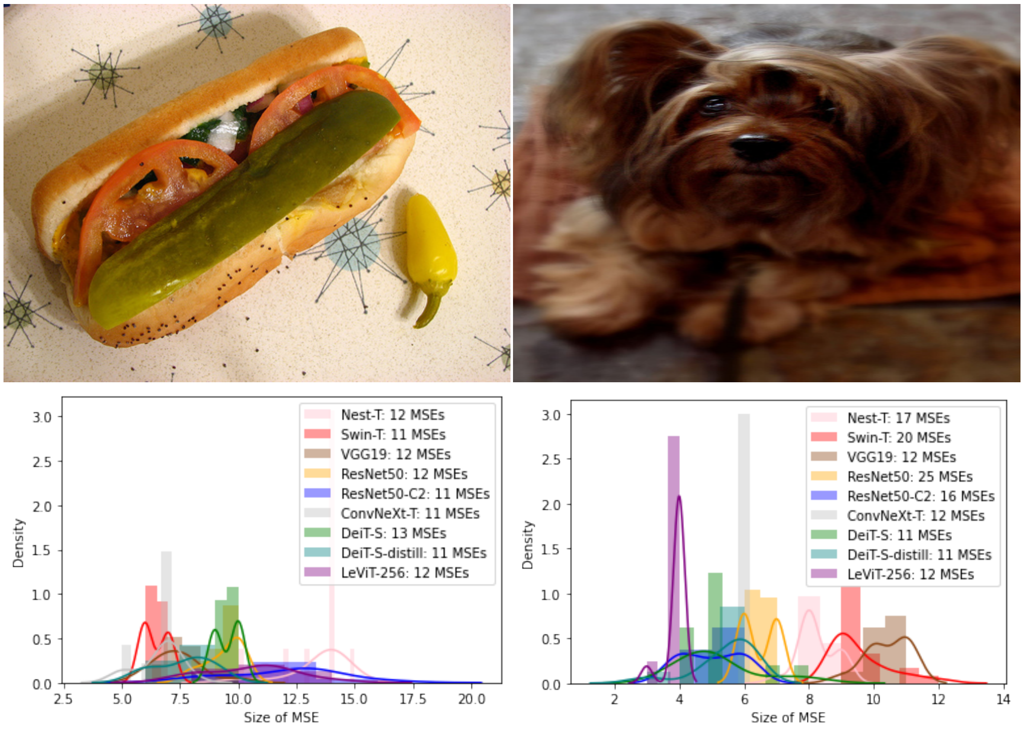
VGG19	ResNet50	Swin-T	VGG19	ResNet50	Swin-T	VGG19	ResNet50	Swin-T
0.1646	0.0914	**0.9573**	0.1609	0.1841	**0.7739**	0.3474	0.1023	**0.6497**
ResNet50-C1	ResNet50-C2	ResNet50-D	ResNet50-C1	ResNet50-C2	ResNet50-D	ResNet50-C1	ResNet50-C2	ResNet50-D
0.1252	0.1526	0.0139	0.0671	0.4226	0.2561	0.1227	0.0063	0.5517
DeiT-S	DeiT-S-dis	LeViT-256	DeiT-S	DeiT-S-dis	LeViT-256	DeiT-S	DeiT-S-dis	LeViT-256
0.1831	0.9081	0.7240	0.8468	0.9696	0.7047	0.1787	0.4536	0.4109
ConvNeXt-T			ConvNeXt-T			ConvNeXt-T		
0.3726			0.2244			0.8044		



Prediction Confidence on the Partially Occluded Image

VGG19	ResNet50	Swin-T	VGG19	ResNet50	Swin-T	VGG19	ResNet50	Swin-T
0.4301	**0.3113**	0.8467	0.2722	**0.2889**	0.4911	0.0095	0.4338	0.9279
ResNet50-C1	ResNet50-C2	ResNet50-D	ResNet50-C1	ResNet50-C2	ResNet50-D	ResNet50-C1	ResNet50-C2	ResNet50-D
0.6198	0.6249	0.7813	0.4842	0.5742	0.8244	0.6086	0.6513	0.8909
DeiT-S	DeiT-S-dis	LeViT-256	DeiT-S	DeiT-S-dis	LeViT-256	DeiT-S	DeiT-S-dis	LeViT-256
0.7590	0.9764	0.6127	0.3563	0.8125	0.8740	0.8063	0.9221	**0.9236**
ConvNeXt-T			ConvNeXt-T			ConvNeXt-T		
0.5232			0.1500			0.8196		

Figure 9: Qualitative Cross-Testing Results. The partially occluded images were generated using iGOS++ heatmaps on the algorithm with bolded number (not necessarily the highest). Then the same image is tested on multiple algorithms and we show predicted class-conditional probabilities on the ground truth class (written above).



(a) Hotdog

(b) Silky Terrier

Figure 10: A few example distributions of MSE sizes for different algorithms on random images.

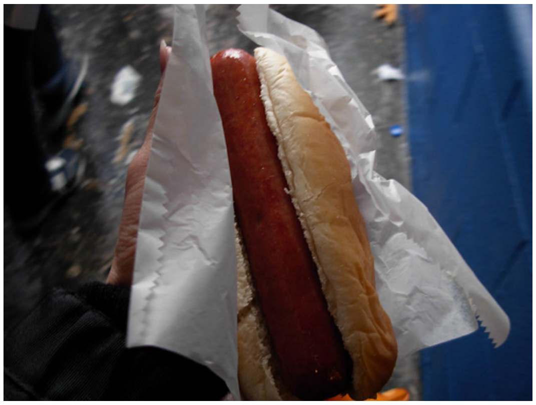


Figure 11: An image of a hotdog

Figure 12: An example SAG tree explaining Swin Transformers on Fig. 11. This tree is too big to be visualized efficiently, but the sheer size of it shows the robustness of Swin Transformers to different types of occlusions. It also justifies our approach of looking at statistics rather than the visualization themselves

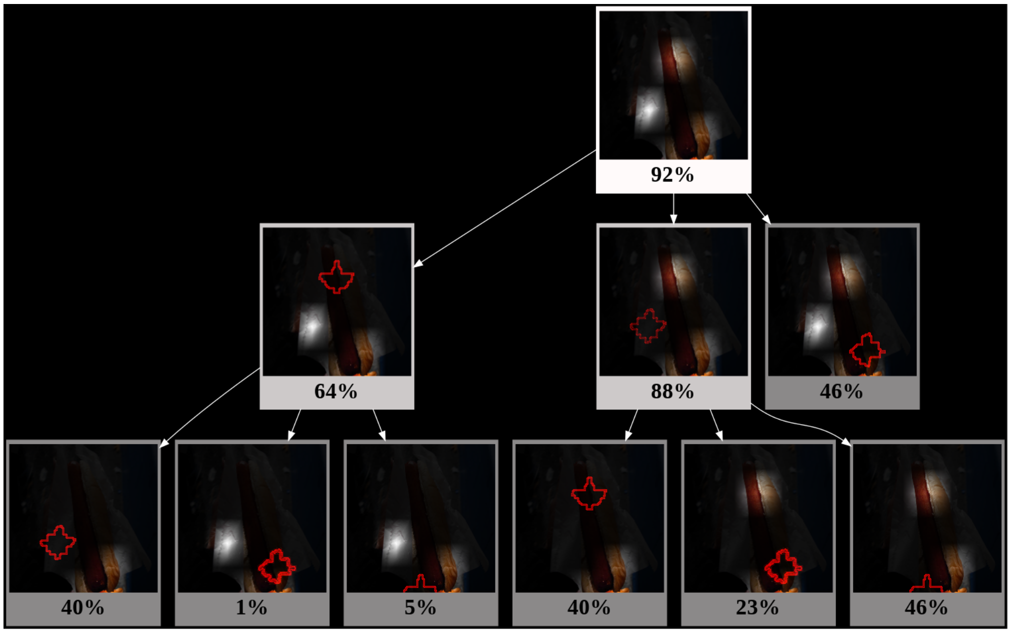


Figure 13: An example SAG tree explaining Nested Hierarchical Transformer on Fig. 11.

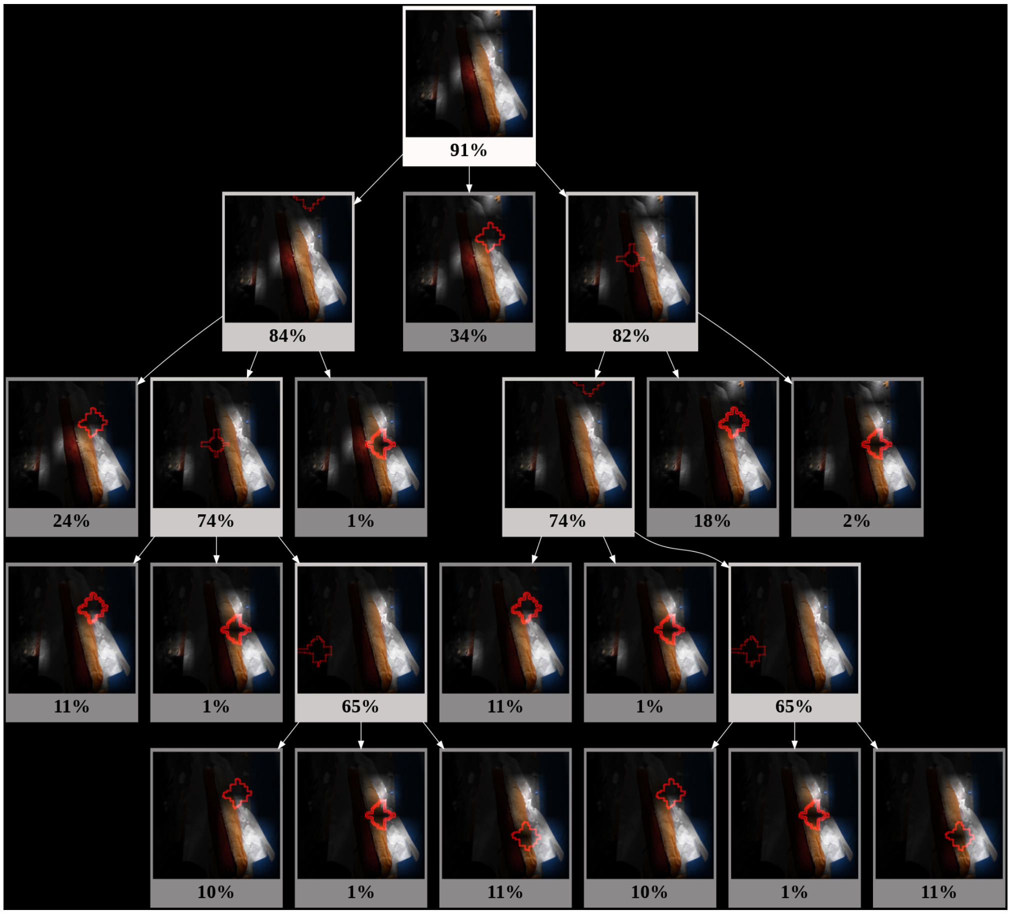


Figure 14: An example SAG tree explaining Fig. 11 for VGG. It can be seen that in many cases removal of a few parts lead to low-confidence predictions

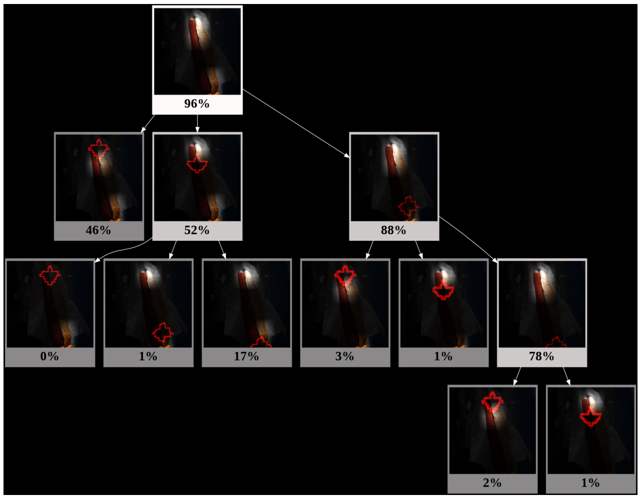


Figure 15: An example SAG tree explaining Fig. 11 for ResNet-50-C2. It can be seen that the SAG is small and focused on a very specific combination of patches of the sausage

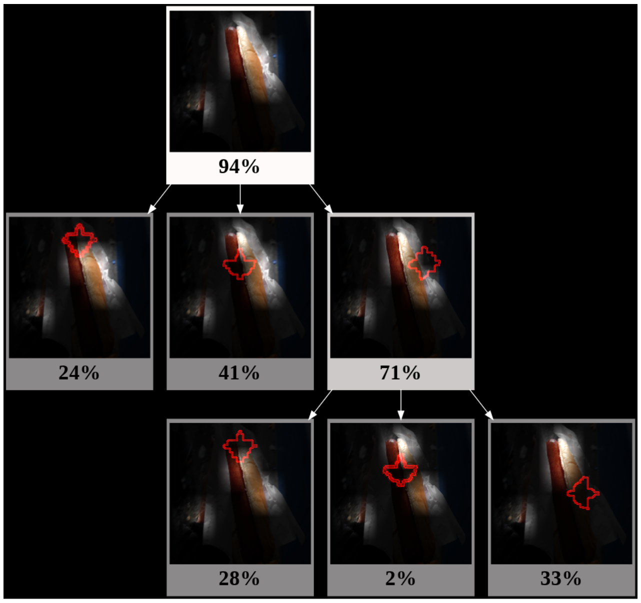


Figure 16: An example SAG tree explaining Fig. 11 for ConvNeXt-T.

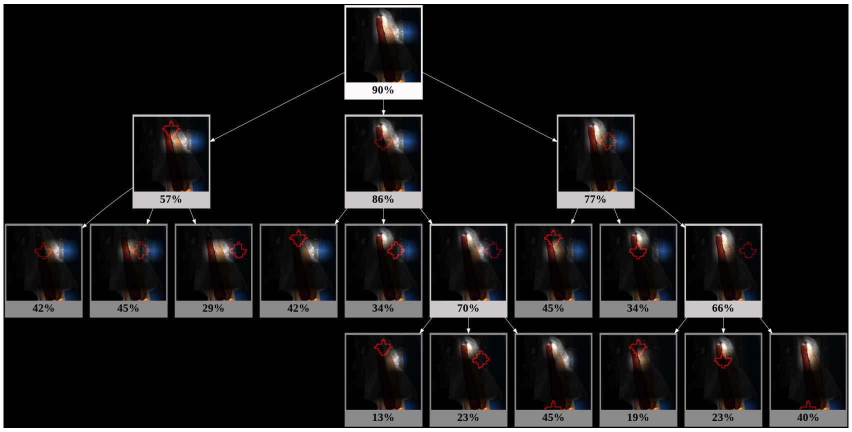


Figure 17: An example SAG tree explaining Fig. 11 for DeiT-S

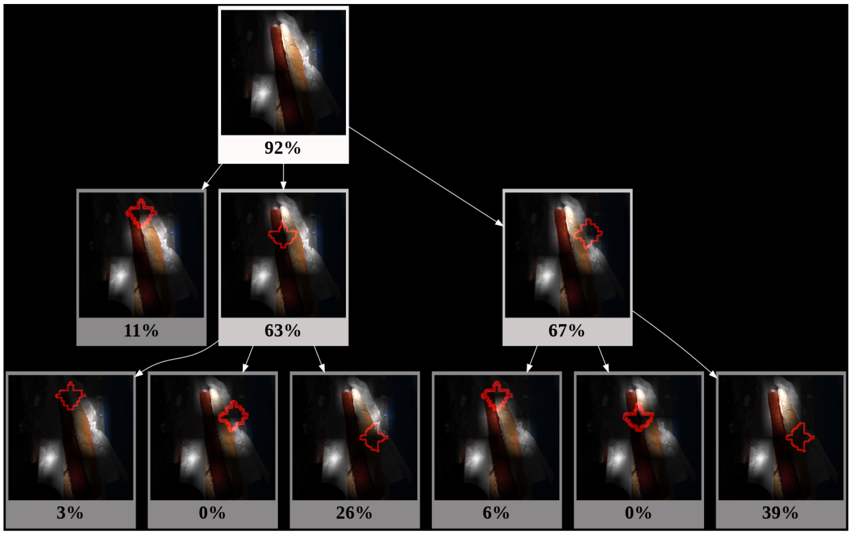


Figure 18: An example SAG tree explaining Fig. 11 for DeiT-S Distilled

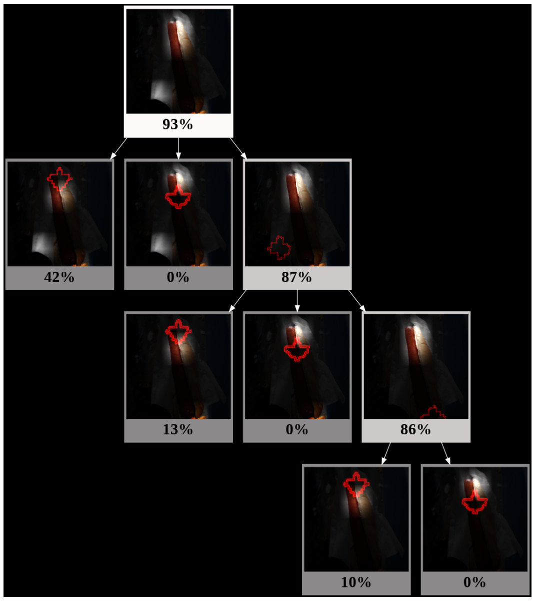


Figure 19: An example SAG tree explaining Fig. 11 for LeViT-256

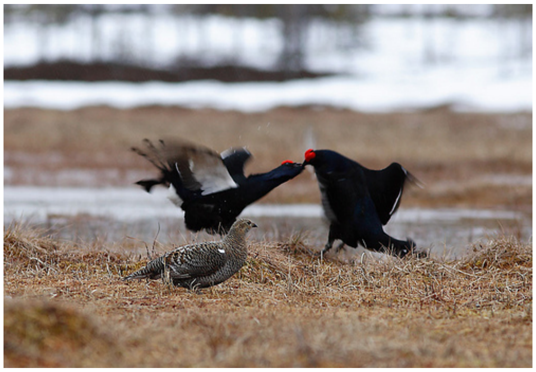


Figure 20: An image of the Black grouse

Figure 21: An example SAG tree explaining Swin Transformers on Fig. 20. Again, the tree size is too large to be visualized properly

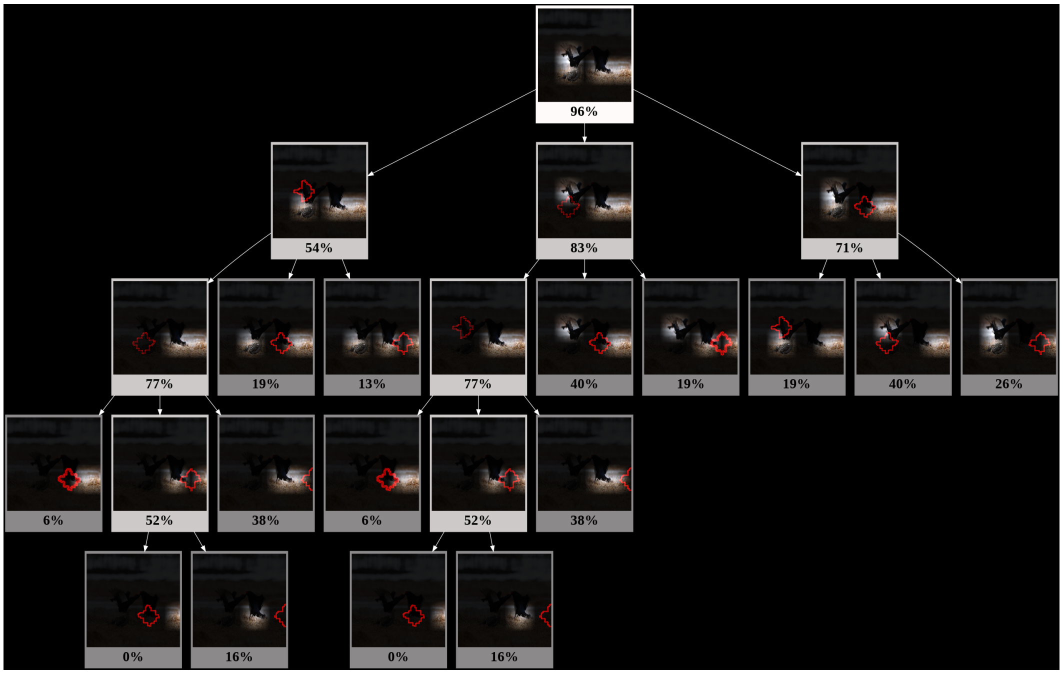


Figure 22: An example SAG tree explaining Nest-T on Fig. 20

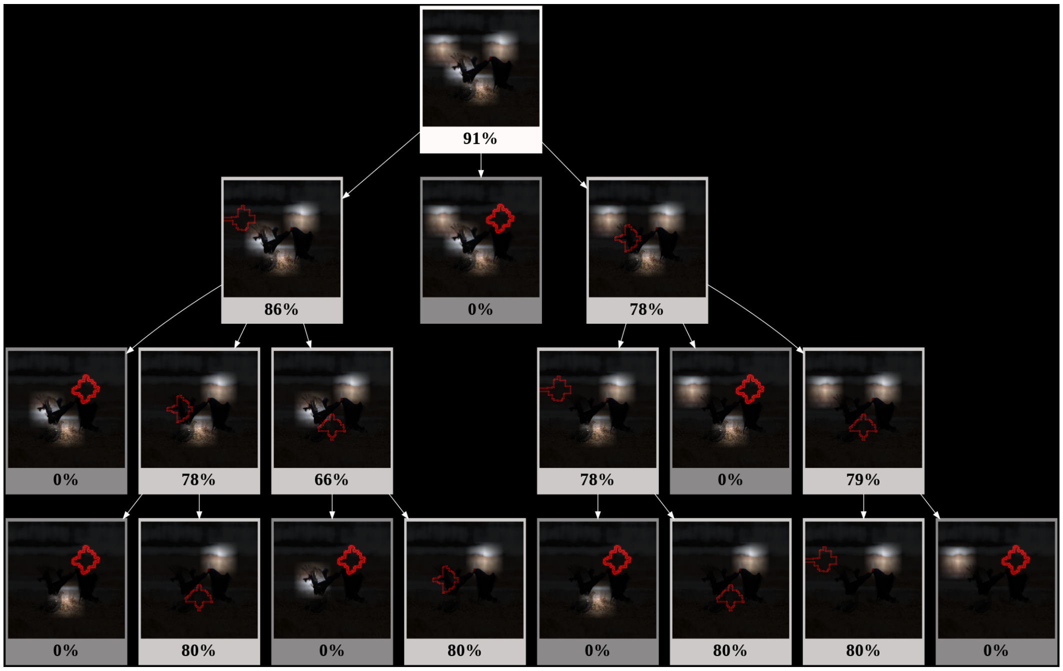


Figure 23: An example SAG tree explaining Fig. 20 for VGG.

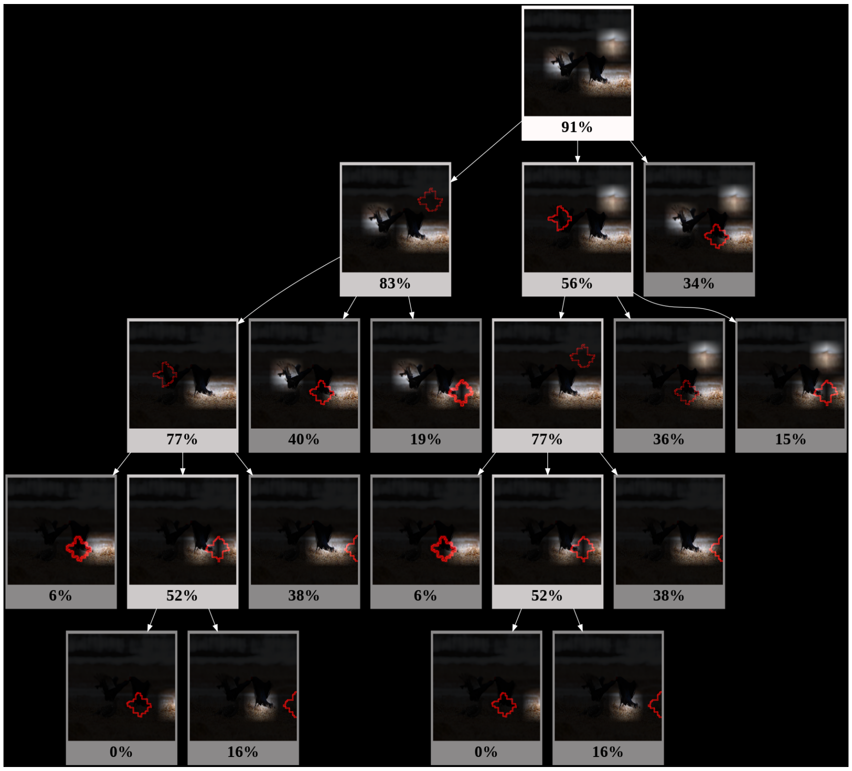


Figure 24: An example SAG tree explaining Fig. 20 for ResNet-50-C2.

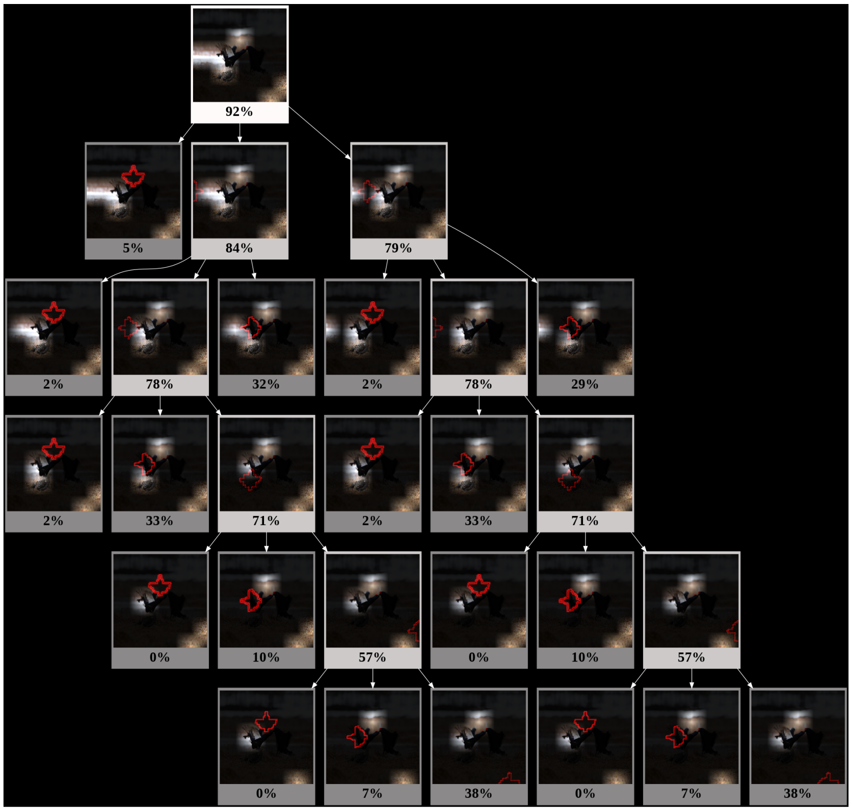


Figure 25: An example SAG tree explaining Fig. 20 for ConvNeXt-T.

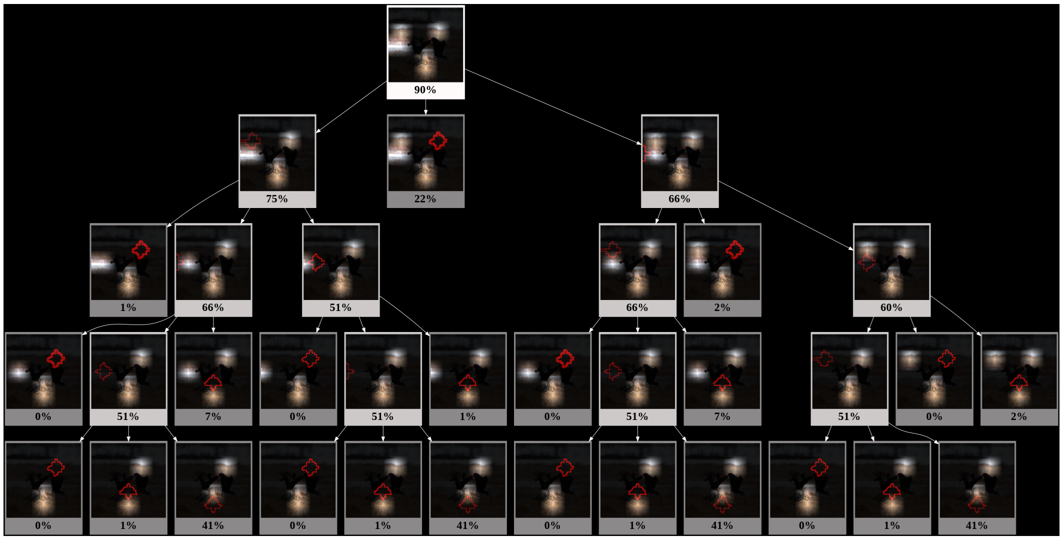


Figure 26: An example SAG tree explaining Fig. 20 for DeiT-S

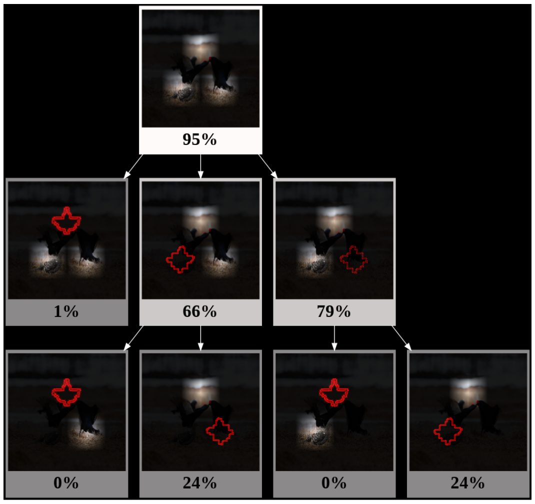


Figure 27: An example SAG tree explaining Fig. 20 for DeiT-S Distilled

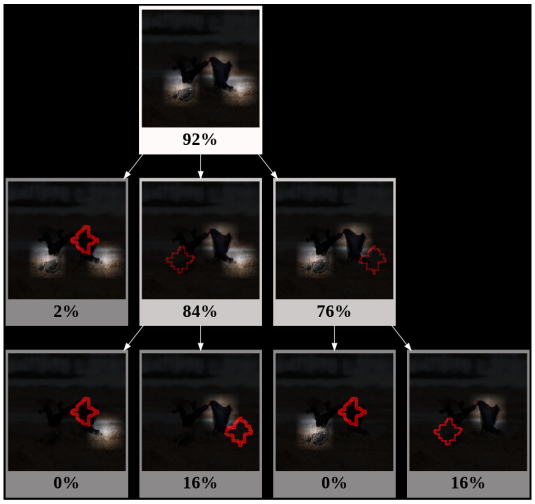


Figure 28: An example SAG tree explaining Fig. 20 for LeViT-256

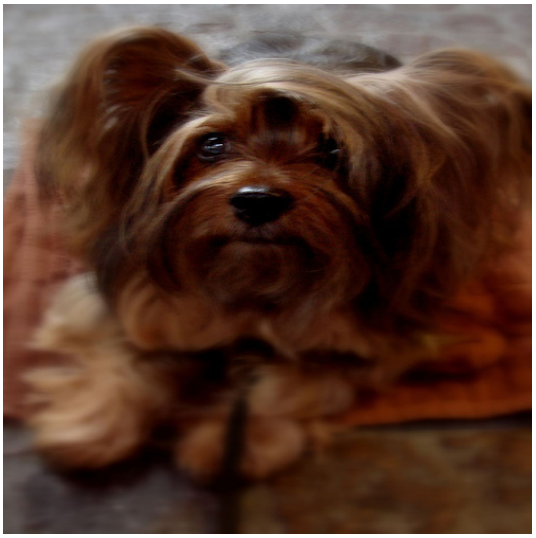


Figure 29: An images of a silky terrier

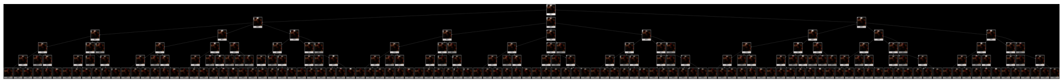


Figure 30: An example SAG tree explaining Fig. 29 for Swin-T. Again, the tree size is too large to be visualized properly

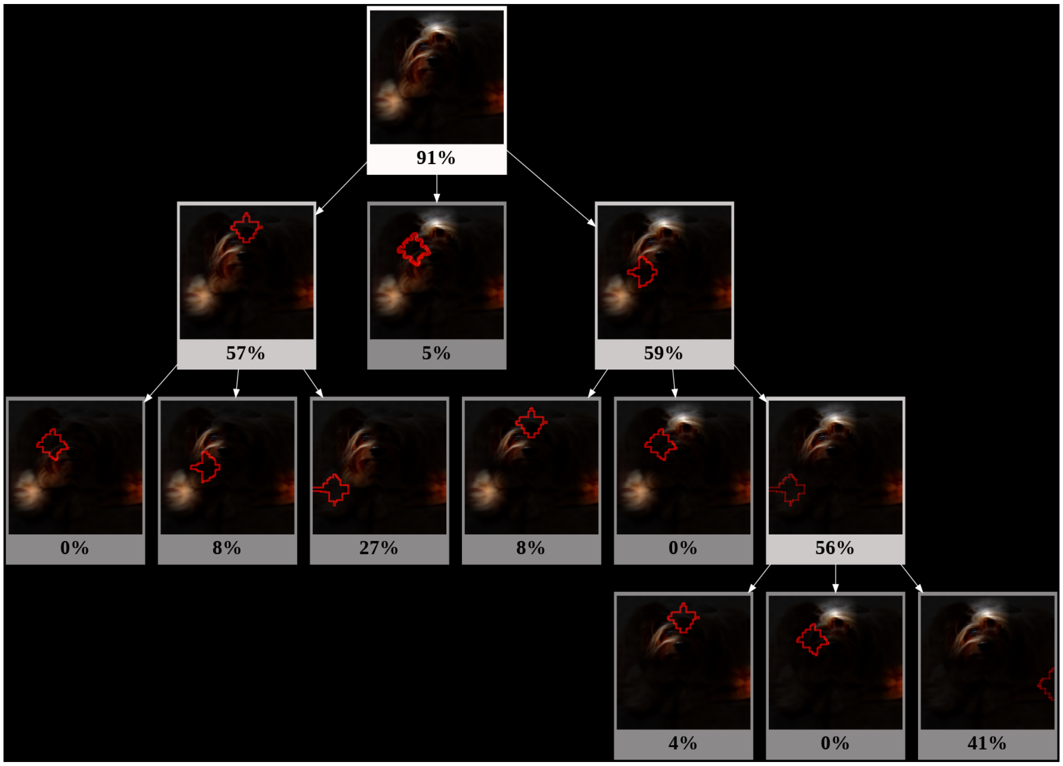


Figure 31: An example SAG tree explaining Fig. 29 for Nest-T

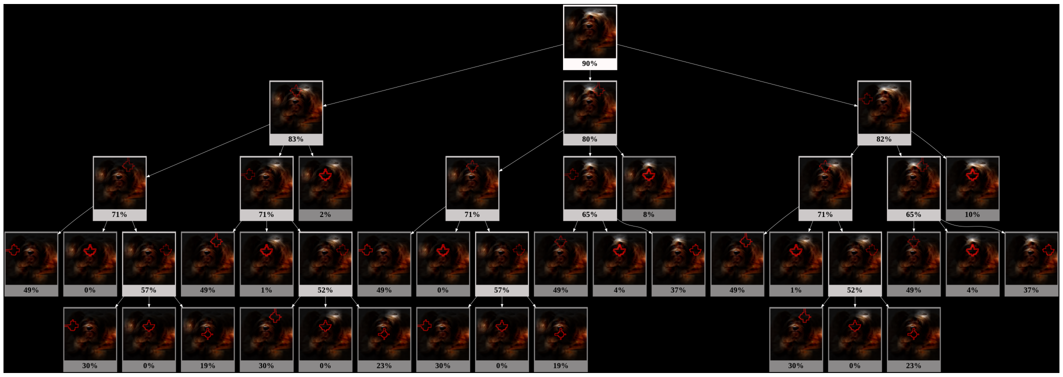


Figure 32: An example SAG tree explaining Fig. 29 for VGG. In this case the tree size is also too large to be visualized properly

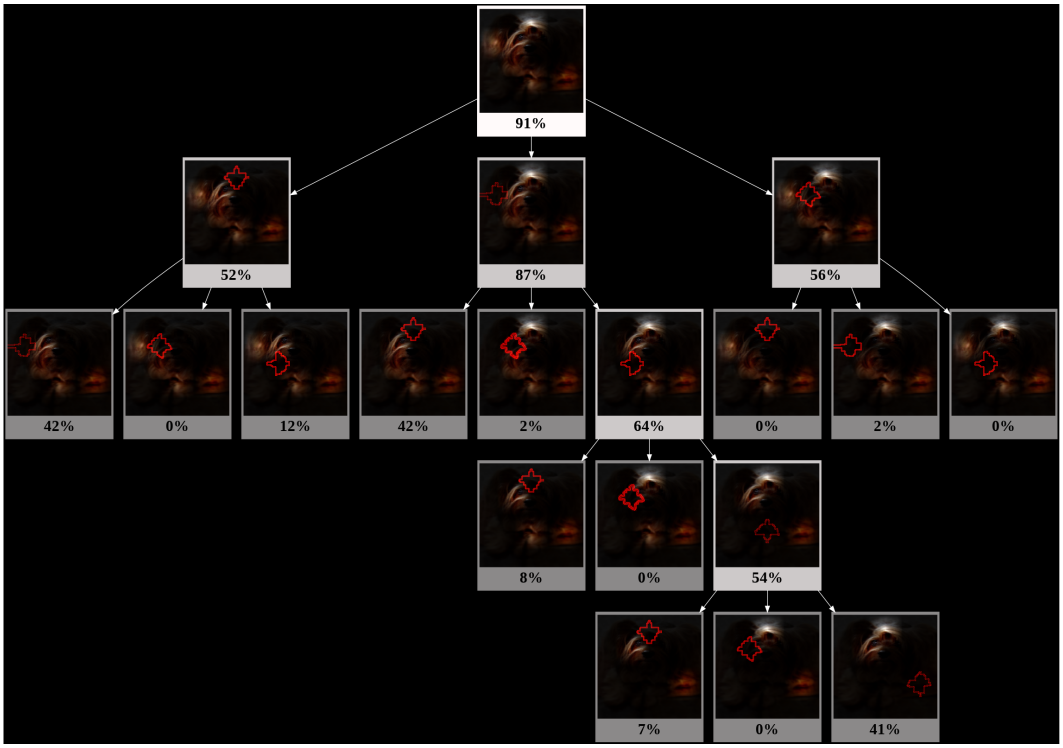


Figure 33: An example SAG tree explaining Fig. 29 for ResNet-50-C2.

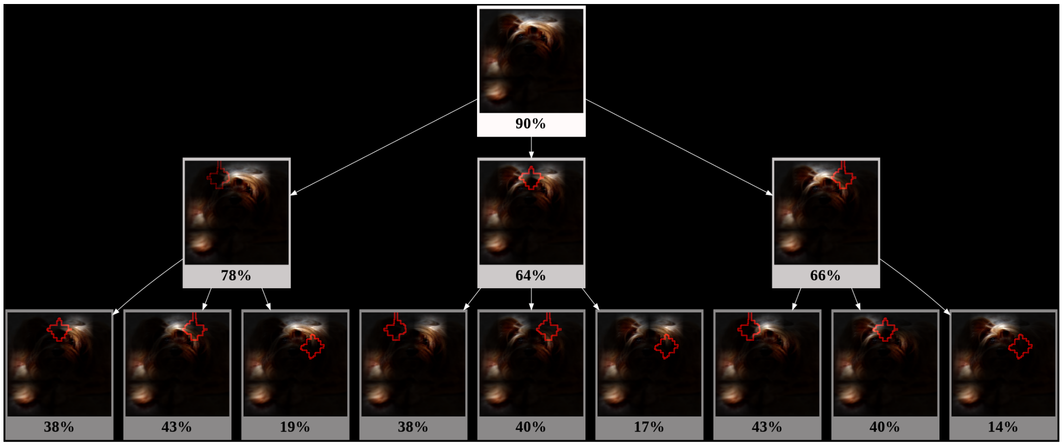


Figure 34: An example SAG tree explaining Fig. 29 for ConvNeXt-T.

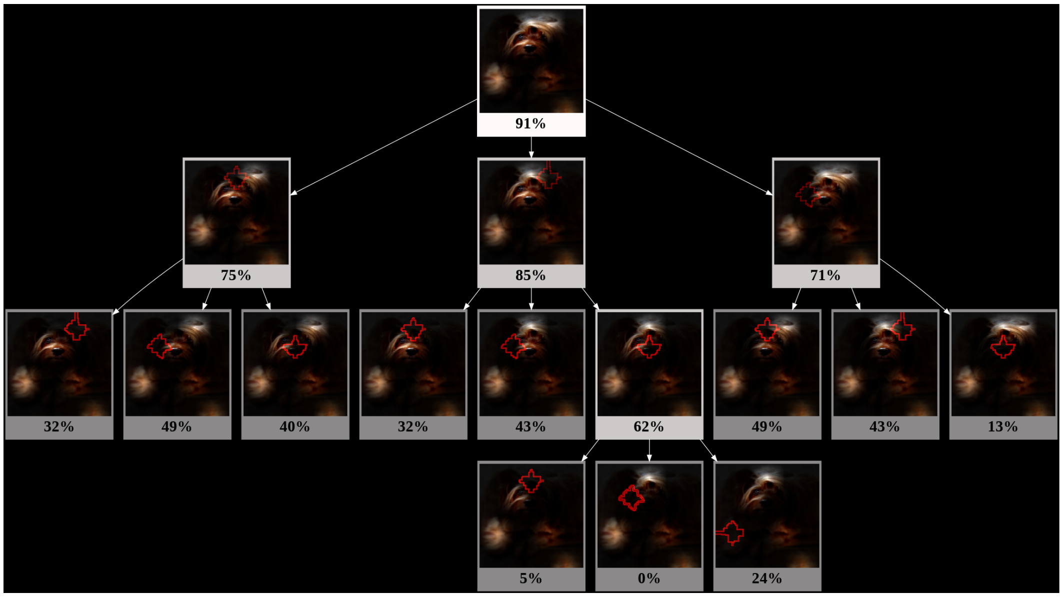


Figure 35: An example SAG tree explaining Fig. 29 for DeiT-S

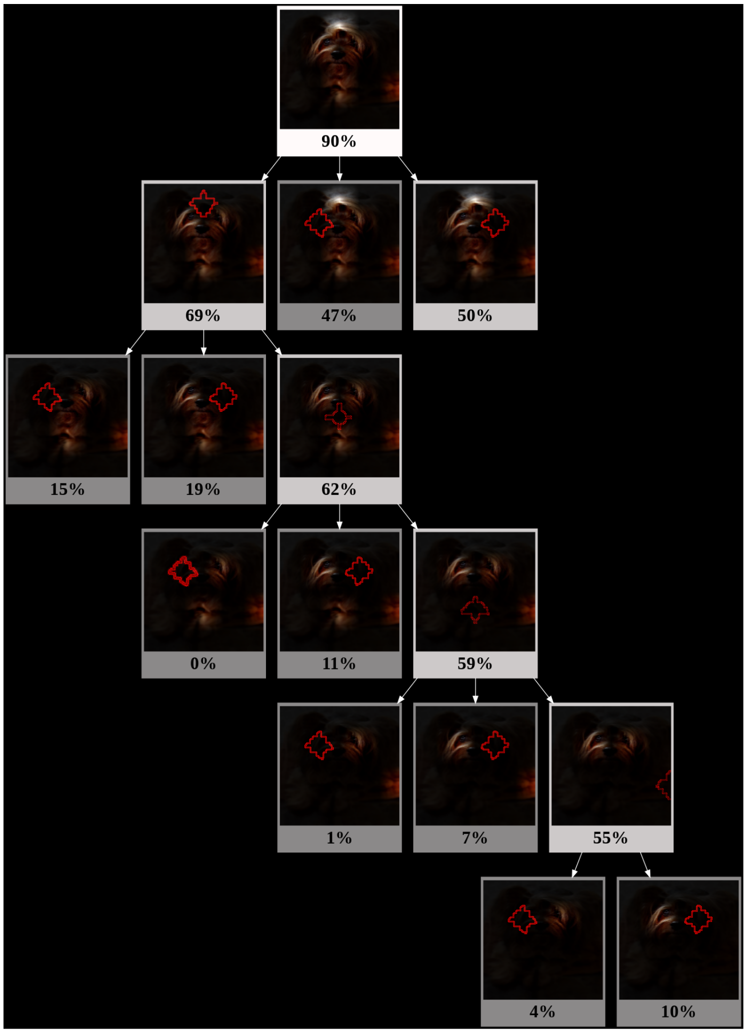


Figure 36: An example SAG tree explaining Fig. 29 for DeiT-S Distilled

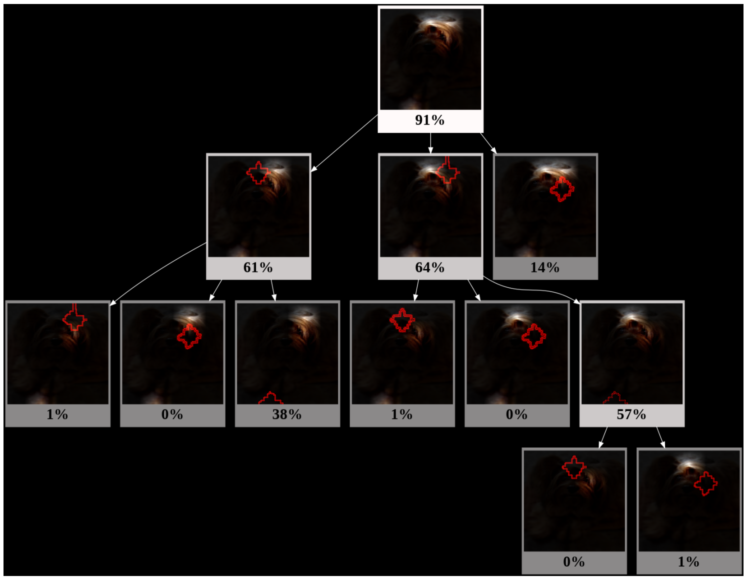


Figure 37: An example SAG tree explaining Fig. 29 for LeViT-256