I Am Big, You Are Little; I Am Right, You Are Wrong

David A Kelly
King's College London, UK
david.a.kelly@kcl.ac.uk

Akchunya Chanchal
King's College London, UK
akchunya.chanchal@kcl.ac.uk

Nathan Blake
King's College London, UK
nathan.blake@kcl.ac.uk

Abstract

Machine learning for image classification is an active and rapidly developing field. With the proliferation of classifiers of different sizes and different architectures, the problem of choosing the right model becomes more and more important. While we can assess a model's classification accuracy statistically, our understanding of the way these models work is unfortunately limited. In order to gain insight into the decision-making process of different vision models, we propose using minimal sufficient pixels sets *to gauge a model's 'concentration': the pixels that capture the essence of an image through the lens of the model. By comparing position, overlap, and size of sets of pixels, we identify that different architectures have statistically different concentration, in both size and position. In particular, ConvNext and EVA models differ markedly from the others. We also identify that images which are misclassified are associated with larger pixels sets than correct classifications.*

1. Introduction

Neural networks are now a primary component of most AI systems, especially in computer vision. There are a plethora of studies comparing image classifiers, either within or across different architectures [24, 25, 28]. Such comparisons tend to focus on accuracy, precision, and robustness [36]. These studies leave unanswered the question of what features or concepts models use to make their classifications.

Jiang et al. [18] investigate "compositionality", which they define as a conjunction of *patches* that have high likelihood ratios for a particular classification. A patch is a fixed square in a grid constructed over the image. Combinations of patches are called *minimally sufficient explanations* (MSEs), and are calculated using the explainable AI (XAI) tool, SAG [33]. As Chockler et al. [8] point out though, SAG uses a definition of explanation which is highly unusual. A combination of patches is considered sufficient if its classification likelihood is above a user-provided threshold. This threshold is a scalar of the model's confidence on an image and can be arbitrarily low. In effect, with a suitable threshold, *any* combination of patches is sufficient for the desired classification. This sufficiency is unclear: it is certainly not always sufficient to guarantee that a patch combination provides enough information to (re)produce the original classification. Indeed, given an image with low model confidence, even a relatively high threshold is likely to see many accepted patch combinations which tell us nothing about the classification and very little about the model.

We start instead with the problem of 'concentration': the smallest number of pixels required to get the original classification. We use ReX[1] [7] to find *minimal sufficient pixel sets*, or MPSs (see Figure 1). These have two important advantages over the MSEs of SAG. Firstly, they are not bound to a particular patch size. SAG breaks the image up into a grid of patches, so its MSE is minimal only in its combination of *patches* and not in its combination of *pixels*. ReX is not patch based, so an MPS is likely to contain fewer unnecessary pixels than an MSE[2]. Secondly, a ReX MPS is sufficient to recreate the original (top) classification of the image. Instead of having a confidence threshold, ReX provides an approximately minimal, non patch-based, set of pixels which really are sufficient to get the desired classification from the model.

By putting concentration onto a sound footing, other properties of models, such as compositionality — a measure of diversity in model concentration — can be studied more rigorously. We investigate concentration in terms of size, position and overlap of MPSs on 15 different image classification models. The models cover 5 different architectures and range in size from small Inception models to state-of-the-art transformers models with over a billion parameters. All models were fine-tuned on ImageNet. The goal of this comparison is to answer the following research questions:

[1] https://github.com/ReX-XAI/ReX

[2] It is still only an approximation of true minimality however. See Chockler et al. [7] for a full discussion.

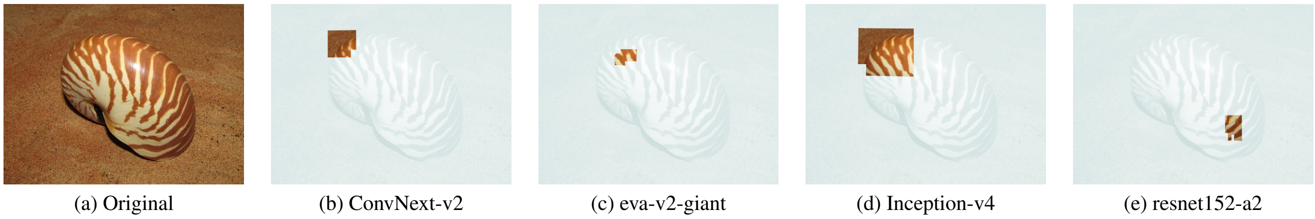

(a) Original (b) ConvNext-v2 (c) eva-v2-giant (d) Inception-v4 (e) resnet152-a2

Figure 1. Minimal, sufficient pixels for an image classified as a seashell by different models. The varying size of the pixel sets is visually evident. ResNet152 uses a completely different region of the image, and very few pixels, for the classification.

RQ1 Do different models have MPSs of different sizes on the same image?

RQ2 Do different models produce an MPS in different locations on the same input image? That is, are different models looking at different areas of the image for the classification?

RQ3 Do wrong classifications have larger or smaller MPSs than correct classifications?

We find that larger models tend to produce smaller, more spatially distinct, minimal pixel sets. In particular, EVA and ConvNext are significantly different from the other studied models. This may indicate overfitting in these large models.

Although ReX is an XAI tool, in general we avoid using the word "explanation" for the rest of the paper. Our goal is not to demonstrate the utility of causal reasoning to human-centered explainability, but rather to capture something less subjective about how models process inputs and what features and concepts they rely upon. Indeed, given the surprisingly small size of many MPSs (see Section 4) it is not always obvious how well they would act as explanations to a human.

We also do not claim that a given MPS is the *only* information relevant to the model when it makes its classification. Indeed, both [33] and [8] show that multiple explanations may exist in an image. The MPSs we consider are always the global maximum MPS (see Section 3): they are the pixels most causally responsible and show where the model was giving the majority of its concentration.

Due to the lack of space, full tables of results, models, benchmarks, and the experimental setup are submitted as a part of the supplementary material.

2. Related Work

Image classifiers have been developed since the 1980s, though it is only within the last two decades that computing power and data quantity have been sufficient to see them widely adopted [29]. Many studies compare models and architectures for accuracy and precision [23, 32, 39]. More unusually, Su *et al.* [36] compare 18 different ImageNet models for robustness in the face of adversarial attacks. We investigate, however, the minimal pixels required for a model to reproduce its original classification. We conduct this investigation using a causal XAI tool, ReX [7]. We give an overview of ReX in Section 3.

We use ReX because the causal definition of an explanation (Definition 1) is suitable to our needs. It has the advantage over other XAI tools in that the "explanation" is tested modulo the model: a causal explanation is a minimal sufficient subset of pixels in the image such that they have the same classification as the overall image. As the model acts as its own oracle, a causal explanation is less dependent on human interpretation than, for example, Shapley values. ReX is a purely black-box tool. Broadly, XAI tools split into white and black-box methods [1]. The white-box method GradCAM [31], for example, has spawned extensions [4] and forms just a small part of the wide range of layer attribution methods available[3]. All white-box methods require access to the internals of the model and need to be tailored to the specifics of the model architecture.

Black-box methods form a smaller family, but still exhibit a diversity of approaches. Among the more popular tools for image classifier explanations are LIME [30], SHAP [22], and RISE [27]. LIME builds a locally interpretable model by using perturbations of the image. SHAP uses game-theoretic Shapley values to provide a heatmap of pixel contributions to the classification. RISE utilizes random occlusions of the image to discover pixel contributions to the classification. All of these tools provide some form of pixel ranking, but do not directly isolate those pixels sufficiently for a given classification.

Jiang et al. [18] use the XAI tool SAG [33] to compare the decision making mechanisms of transformers and CNNs. SAG generates multiple explanations for a given image. These explanations are given as "patches", similar in intent to our MPSs. However, the patch itself is of a fixed size and explanations are always combinations of these fixed size patches. ReX has no such limitation. Moreover, what SAG accepts as a minimal sufficient explanation is much more generous than what ReX accepts. ReX accepts the MPS if, and only if, the top class of the MPS is the same as the top class of the overall image: SAG instead uses a

[3] See the captum library [20] for a large selection of algorithms.

confidence threshold. This means that a SAG MSE might not be the top 1 classification, but in fact might be in almost any position in the output tensor [8].

3. Minimal Pixels Sets

We use ReX [7] to construct sufficient pixel sets in image classification. The reader is referred to Halpern [14] for a detailed formal overview and more information on actual causality and responsibility. Broadly speaking, ReX views a model as a black-box causal model in the Halpern-Pearl [15] sense of the word, with its inputs being the individual pixels of an image. The variables are defined as Boolean, with the values being the original color and a baseline value. The relevant general definitions are given in [14]. Here we only present the definition of explanation for image classification.

Definition 1 (CKS explanation for image classification) *An explanation in image classification is a minimal subset of pixels of a given input image that is sufficient for the model to classify the image, where "sufficient" is defined as containing only this subset of pixels from the original image, with the other pixels set to a baseline value.*

A recent paper by Chockler and Halpern [6] proves that under the same simplifying assumptions that ReX uses, Definition 1 is equivalent to the definition of explanation in actual causality [14]. Chockler and Halpern [6] observe that the precise computation of an explanation in our setting is intractable, as the problem is equivalent to an earlier definition of explanations in binary causal models, which is DP-complete [12] (DP is the class of languages that are an intersection of a language in NP and a language in co-NP and contains, in particular, the languages of unique solutions to NP-complete problems [26]).

Algorithmic Overview We present a simplified description of the ReX algorithm. See Chockler et al. [7] for a complete description. ReX starts by dividing an image into 4 parts. We call each part a superpixel. ReX creates *mutants* of the original image by covering all combinations of superpixels with a baseline value. By default, this baseline value is 0. These combinations are tested against the model and sorted between those which satisfy the required classification and those which do not. Note that some superpixels may appear in both passing and failing mutants, depending on the exact combination. Causal responsibility is distributed over the different superpixels, where responsibility is a quantitative measure of causality and, broadly speaking, measures the amount of causal influence on the classification [5]. Passing superpixel combinations are further refined into smaller superpixel combinations and tested using the same procedure.

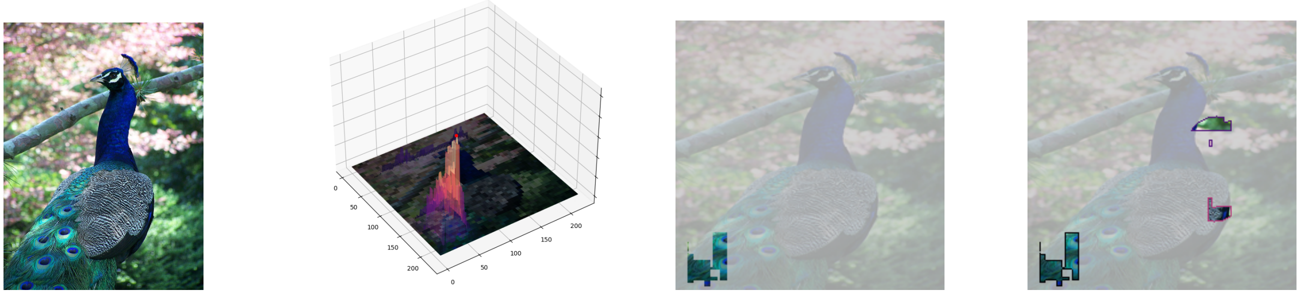
In this way, mutant generation is iteratively guided by the model, and responsibility tends to concentrate on those pixels (not superpixels) which occur most frequently in "good" combinations. This procedure is repeated many times, starting from a different random partition of the image. After a number of iterations of the algorithm, by default 20, ReX produces a responsibility landscape not dissimilar to Figure 2b. One can immediately see that causal responsibility is peaked over one small part of the image. This is not true in general, especially if multiple explanations are present ([8]).

This landscape is not itself an explanation, or MPS, according to Definition 1. ReX still needs to identify, from the landscape, those pixels which are minimal and sufficient for the classification "peacock". The pixels are ranked in the order of their *responsibility* for the classification, We note, however, that the construction of the ranked list is intractable as well (NP-complete), even in the special case of image classification, rather than the general definition of responsibility by [5]. Hence, ReX's ranking is based on the approximate degree of responsibility. As exact MPS computation is intractable, ReX's output is (approximately) minimal, but not necessarily the minimum. As ReX adds pixels according to the responsibility ranking, the first discovered MPS is the one with highest responsibility MPS (Figure 2c). It is guaranteed to be sufficient to obtain the original classification, against the provided baseline. These sufficient sets may sometimes have very low model confidence. We do not address here the issue of whether there is a threshold of model confidence below which we can no longer give credence to the results. Pixel sets based on Definition 1 have the clear advantage of being amenable to a comparison using standard approaches, such as the Sørensen–Dice coefficient (DC), which is used to gauge the similarity of any two samples [9, 35], and the Hausdorff distance [17].

Out of distribution (ood) mutants We use ReX with its default baseline masking value of 0. This masking value is applied after processing the image, and therefore does not necessarily correspond to black. It does mean that virtually every mutant created by ReX results is an *ood* image. The fact that the models are still able to classify the image in a "sane" fashion is a good indicator of their ability to cope with this type of *ood* image. As [2, 7] show, ReX MPSs are almost always inside, or overlapping, a human-provided segmentation, indicating that the MPS is in a reasonable position in the image. As we shall see, the size of a pixel set seems to be a good indicator of how robust a model is to *ood* images, at least of the sort produced by ReX.

3.1. Models and architectures

We consider the following architectures in our study, with each architecture represented by three models.



(a) A peacock

(b) Responsibility map

(c) A peacock MPS

(d) Other peacock MPSs

Figure 2. The various stages of calculating an MPS. We start with an image, Figure 2a. ReX produces a responsibility landscape, (Figure 2b), which shows an approximation of the causal responsibility of each pixel towards the classification "peacock". This map is used to rank pixels, which are then introduced over the baseline value until the pixels reproduce the classification. The result is an MPS (Figure 2c). ReX can also find multiple different MPSs (Figure 2d). In this paper, we investigate the best MPS, the one ranked most highly in Figure 2b.

Inception [37] models are a family of networks based on CNN classifiers. Rather than simply stacking convolutional layers to achieve better performance, Inception models are more complex and utilize a large body of different optimizations and heuristics. Fundamentally, they work by having multiple filters of different sizes at each level of the network. This makes the network wider rather than deeper.

ResNet [16] (*Residual Network*) architecture was introduced to solve the vanishing/exploding gradient problem common to large CNNs. ResNet uses *residual blocks*, where we skip certain intermediate layers when connecting two nodes. A ResNet is a stack of residual blocks. ConvNext [21] is a modernization of the standard ResNet architecture to bring it closer to the design of vision transformer models. The authors report that ConvNext models are competitive with vision transformers on both accuracy and scalability.

ViT [11] (*Vision Transformer*) is an image classification model which uses the encoder-only transformer architecture over patches of an image. An image is turned into a vector suitable for a standard transformer model by dividing it up into fixed-sized patches, which are then linearly embedded along with partition information. EVA [13] is a ViT pretrained to allow the model to scale to a very large number of parameters, potentially over 1 billion.

4. Evaluation

We compare 15 different models, drawn from 5 different architectural classes. All experiments were run on Ubuntu 20.04 LTS on a single A-100 GPU. ReX was run with the same hyperparameters and same random seed for all experiments. For the models, we used the latest available pretrained IN-1k weights, dated February 22nd 2024, via the PyTorch Image Models (Timm) package. All the models were then converted to the ONNX format, using ONNX v1.15 and opset 20. The model was then run under ONNX runtime 1.17 for inference when performing all the experiments. This was done due to the prevelance of deploying models using ONNX in production environments and hence, allowed for replicating real-world conditions as closely as possible.

4.1. Experimental Design

The complete list of all the different models is presented in Table 1. Two of these architectures are transformer-based (ViT and EVA), and the rest are convolutional. Each model has a different input size and a different number of internal parameters. They also differ in their pre-training data set. All models are fine-tuned on ImageNet-1k. To further validate our results, we have also fine-tuned the models on the Caltech-256 dataset to compare against the results obtained on ImageNet-1k. Due to the cost of running the experiments, we selected 500 images uniformly at random from the validation set of ImageNet-1k. We similarly selected 500 from the test set to produce a total of 1000 images.

4.2. Statistical analysis

The subsequent analyses are based on the ratio of the MPS size to the size of the whole image (as different models accept different, fixed, input sizes). We calculate this ratio on different partitions of ImageNet-1k, the validation and test set, the former of which has ground truth labels. Our first null hypothesis is that there is no statistical difference in MPS size between architectures on the test set data. Our second null hypothesis is that incorrect classifications have the same size MPS as correct classifications (see Table 1), tested on the validation set only. By having two random samples we mitigate against type 1 errors, rejecting of the null hypothesis when it is in fact true. Additionally, when we apply these post hoc tests on our data, we use the *Bonferroni correction* [3] to counteract the problem of applying multiple hypothesis tests to the same data.

We use the Kruskal-Wallis H test [10] and Friedman

	Average					
Model	Area	Correct	Incorrect	Test	Mean	Accuracy
ConvNext-V2 Large	0.081	0.075	0.122	0.078	0.089	0.880
ConvNext-V2 Huge v2	0.065	0.063	0.082	0.061	0.068	0.894
ConvNext-V2 Huge v1	0.05	0.048	0.061	0.05	0.052	0.890
EVA-02 Large V1	0.066	0.064	0.082	0.068	0.07	0.890
EVA-02 Large V2	0.07	0.068	0.09	0.065	0.073	0.894
EVA Giant	0.054	0.052	0.065	0.052	0.0558	0.894
Inception-ResNet V2	0.25	0.246	0.265	0.253	0.254	0.814
Inception V3	0.239	0.231	0.271	0.245	0.247	0.800
Inception V4	0.23	0.224	0.261	0.243	0.239	0.840
ResNet 152-B A1	0.142	0.137	0.166	0.139	0.146	0.828
ResNet 152-B A2	0.144	0.136	0.187	0.149	0.154	0.842
ResNet152-D	0.134	0.13	0.155	0.127	0.137	0.828
ViT Large	0.099	0.098	0.111	0.089	0.099	0.900
ViT Huge V1	0.156	0.154	0.17	0.152	0.158	0.882
ViT Huge V2	0.103	0.102	0.113	0.095	0.103	0.872

Table 1. Average size of MPS as percentage of entire image. *Correct* is average area for correct classifications and *Incorrect* for incorrect classifications. We do not have ground truth labels for *Test* so present average area without differentiation. The difference between the Inception models and ConvNext in particular is remarkable, being $3.6\times$ larger on average. *Accuracy* is the accuracy of the model upon the 500 images for which we have ground truth labels.

test [34], both non-parametric tests for non-normal data. The former is used for cross-architecture analysis. The latter is used to detect differences between different treatments of the same data. Because our data is matched (*i.e.* MPSs were extracted from the same images), we can use this test to detect intra-architecture differences in MPS size. We treat the threshold of statistical significance as $p < 0.01$.

We additionally calculate two sets of DC and Hausdorff values, for comparing the outputs of models across architectures and of the same architecture. For the cross-architecture analysis, we select the best-performing model from each architecture class based on accuracy scores on the randomly selected samples from ImageNet-1k and Caltech-256 (Table 4). We then find the intersections of both correct and incorrect classifications (separately) for our best performing models. We calculate our measures on these subsets. As we do not have labels for the test set of ImageNet-1k, we use the majority vote between models as a proxy for correctness. We use a similar approach for intra-architecture measures, comparing all models of a given architecture.

4.3. Results

In this section we present the analysis of our experiments. All of our analyses are over sets of MPSs, obtained from different models on the same images. Figure 1 is an illustration of such a set, for a seashell classification. We can see that the MPS in Figure 1e is in a completely different location from other explanations, having a DC of 0 and a Hausdorff distance of approximately 156 from Figure 1b. All other MPSs in this example cluster around the upper part of the shell. ReX was used with the same random seed for all runs, meaning that all *initial* superpixels were the same. Refinement of superpixels is guided by the model under test. The fact that ResNet152 has an MPS in a significantly different area indicates that causal responsibility was distributed completely differently from the other models.

We focus on the results for ImageNet in this section. We verified the ImageNet-1K results using a set of 50 randomly sampled images from CalTech-256 as a sanity check on our main results. This returns similar findings, with a statistically significant difference between architectures ($p < 0.01$), though evidence of inter-architecture differences are not found for the EVA, Inception or ResNet models at $p < 0.01$. The sample size of CalTech-256 was much smaller however.

RQ1: different MPS size

There are statistically significant differences in MPS size both across and inside different architectures. The Kruskal-Wallis H test indicates the probability of the null hypothesis being correct at $p < 0.01$.

Both the DC and Hausdorff scores for our top models demonstrate that MPSs of models with different architectures are not based on the same set of pixels. There is considerable variety of MPS size and location depending on the model. Table 1 presents the ratios of the size of the

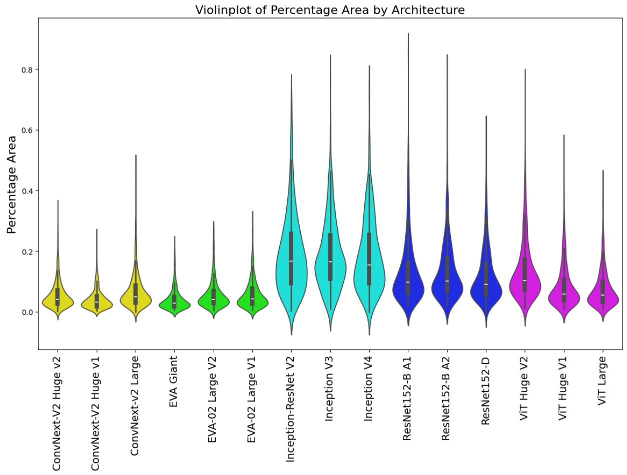


Figure 3. Violin plot of MPS size as a ratio to overall image size. Color coordinated by architecture with yellow: ConvNnet, green: EVA-02, cyan: Inception, dark blue: ResNet and purple: ViT. A Kruskal–Wallis H test indicates a significant difference amongst the architectures: $H(4) = 1176.134, p < 0.001$. To determine whether there is a difference between the models within each architecture we apply the Friedmans test results and find that there is a difference, except for the Inception models (p=0.36).

MPS with respect to the size of the input image, for all pixel sets and separately for the subsets of correct and of incorrect classifications. The Kruskal-Wallis H test indicates that, based on the ImageNet-1k dataset, the size of pixel sets found by ReX does differ by model architecture ($p < 0.01$). The Inception models, in particular, return larger pixel sets (see Figure 1d for a typical example). The Friedman test further demonstrated a significant difference in MPS size even within architectures, except for the Inception models (ConvNext, $p < 0.01$; EVA $p < 0.01$; Inception, $p > 0.01$; ResNet $p < 0.01$; ViT, $p < 0.01$). Figure 3 shows a violin plot of all models and their MPS size, with architecture coordinated by color. It is immediately clear the the ConvNext and EVA models are quite different from the other architectures.

RQ2: MPS location

Both the DC and Hausdorff scores for our top models demonstrate that MPSs do not group around the same set of pixels. There is a considerable variety of MPS location.

Model	EVA Giant	Conv Next	ViT Large	ResNet 152	Inception
EVA Giant	1.0	0.287	0.253	0.165	0.141
ConvNext	0.287	1.0	0.304	0.162	0.163
ViT Large	0.253	0.304	1.0	0.232	0.225
ResNet152	0.165	0.162	0.232	1.0	0.282
Inception	0.141	0.163	0.225	0.282	1.0

Table 2. Average DC values for pixel sets of best performing models across architectures on ImageNet-1k validation subset of correctly classified images.

Model	EVA Giant	Conv Next	ViT Large	ResNet 152	Inception
EVA Giant	0.0	99.5	98.6	85.5	139.2
ConvNext	99.5	0.0	95.6	89.9	139.4
ViT Large	98.6	95.6	0.0	78.9	121.5
ResNet152	85.5	89.9	78.96	0.0	88.4
Inception	139.2	139.4	121.4	88.4	0.0

Table 3. Average Hausdorff coefficient values for explanations of best performing models across architectures on ImageNet-1k validation subset collectively classified correctly.

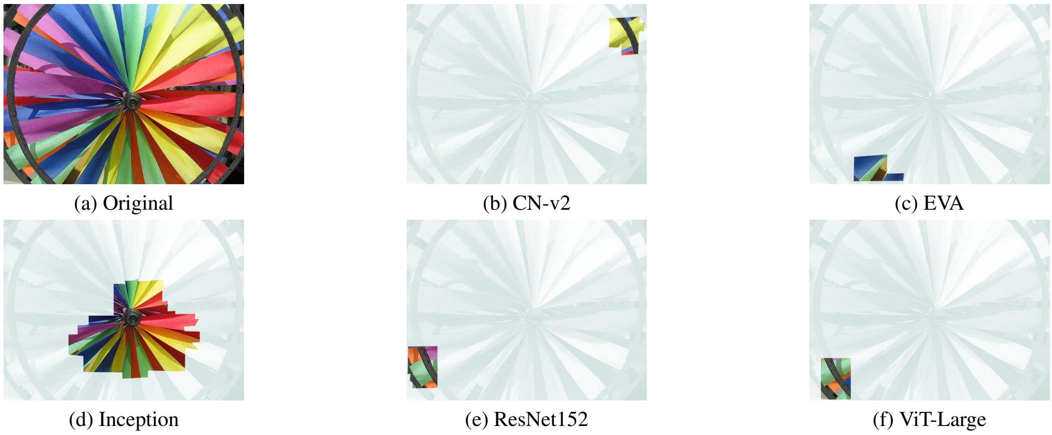
Table 2 shows the average DC for the top performing models. The average overlap is in general quite low, indicating that MPSs across models do not share a large number of pixels. Table 3 shows the average Hausdorff distance across our best performing models. Inception V4 consistently finds pixel sets in different locations from these for the other models. The closest to these of Inception V4 are produced by ResNet152.

RQ3: incorrect MPS size

There is a statistically significant difference between MPS sizes for correct and for incorrect classifications. The effect size is small, however, with only an average increase in area of 2.6%.

To assess whether the correctness of a model's output with respect to the ground truth label, or *fidelity*, has an effect on the MPS size, we partitioned the classifications of the ImageNet validation set into correct and incorrect subsets. As evident from table Table 4, different models differ in their accuracy by as much as 10%, indicating a notable confounding effect. To control for this, we applied a mixed linear model, which demonstrated that having the incorrect classification increased the explanation size by 2.6% (standard error $0.4\%, p < 0.01$). A similar effect is seen in the CalTech-256 data, with an estimated increase in size of 3.3% (standard error $0.01\%, p < 0.01$). While statistically significant, practical implications may be limited; that is something that needs to be determined qualitatively, and may change by domain area.

Figure 5 shows an example of model *infidelity* to the ground truth label. Its label in ImageNet is "grey fox". The best performing models all label the image as "hyena". The MPSs cluster around the center and rear of the body. Of note is that one of the salient distinguishing features of the



(a) Original

(b) CN-v2

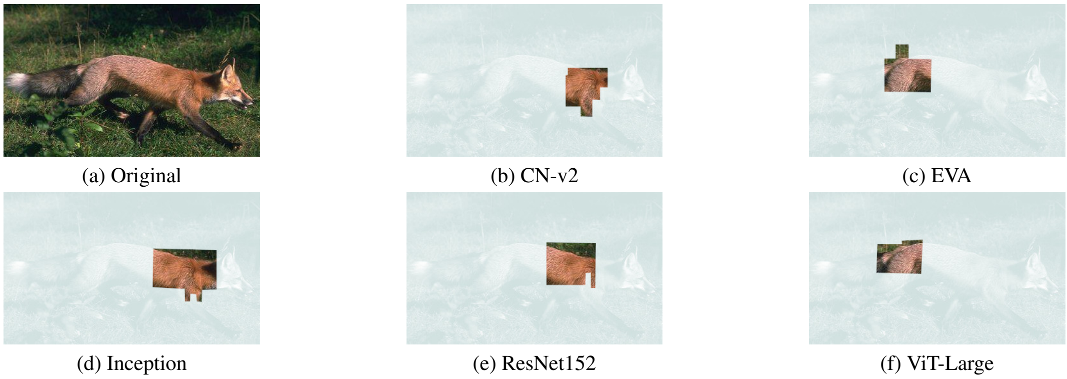
(c) EVA

(d) Inception

(e) ResNet152

(f) ViT-Large

Figure 4. A pinwheel. This image has the highest Hausdorff distance for any set of MPSs over our best performing models. While a high Hausdorff distance does not necessarily imply *no* pixel overlap, in this case only Figure 4e and Figure 4f have any pixels in common.



(a) Original

(b) CN-v2

(c) EVA

(d) Inception

(e) ResNet152

(f) ViT-Large

Figure 5. The label for this image is "grey fox". All of the best performing models classify it as "hyena". One of the salient features for a human classifier, the tail, is completely missing from the MPSs.

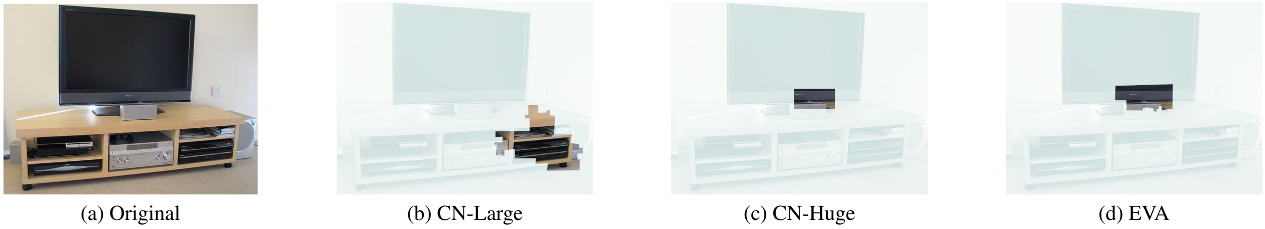
grey fox to humans, the black stripe down its tail which is also tipped with black, is completely absent from the MPSs. Conversely, Figure 6 is an example of mixed fidelity, where some models are faithful to the ground truth label and some are not. The disagreement in this case is logical, as there is potential ambiguity between the "home theater" and "television". It is interesting to note the inconsistency of the disagreement, with both Figure 6c and Figure 6d having MPSs of similar size and location, but different classifications.

4.4. Discussion

ReX mutants are produced using a baseline masking value, as opposed to the blur that SAG uses. For an MPS to be very small, the model must accept a relatively large number of images which are fundamentally out of distribution. As ReX uses a process of iterative refinement for mutant creation and usually quits only when the mutants are no longer classified appropriately, it would seem that some architectures are quite happy to make classifications based on very little "information". These very small MPSs (see Figure 4b) may indicate overfitting. Conversely, models which produce large MPSs, such as Inception, are less able to understand these *ood* images. It may be the case that different models are more sensitive to different types of *ood* images (*i.e.* the use of different baselines, or different types of blurring).

Our results highlight the importance of taking MPS characteristics into account when selecting an appropriate model for a given task, rather than relying purely on other performance metrics. The results demonstrate that large models, which are pre-trained on larger corpora of data, on average, tend to utilize very little of the available input (5.4% in the case of the 1 billion parameter EVA-Giant model). They are also highly confident in their predictions, highlighting the "myopic" tendencies of such models. This raises questions regarding the safety of these models, especially in high-stakes environments, such as healthcare, autonomous navigation or quality assurance.

MPSs provide a possible solution for mitigating some of these concerns. As there is a significant statistical difference



(a) Original

(b) CN-Large

(c) CN-Huge

(d) EVA

Figure 6. The ground truth label for this image is 598 "home theater". Only the EVA model (Figure 6d) produced the same classification. The two ConvNext models, Figure 6b and Figure 6c both classify the image as 851, "television". It is interesting to note that the MPSs for ConvNext-Huge and EVA are very similar, though the classifications are different.

Model	Imagenet-1K	Caltech-256
EVA-02 Large V1	0.890	1.00
EVA-02 Large V2	0.890	0.98
EVA Giant	0.894	0.94
ConvNext-V2 Huge v1	0.890	0.98
ConvNext-V2 Huge v2	0.894	0.98
ConvNext-V2 Large	0.880	1.00
ViT Huge V1	0.882	0.98
ViT Huge V2	0.872	0.98
ViT Large	0.900	1.00
ResNet152-B A1	0.828	0.92
ResNet152-B A2	0.842	0.94
ResNet152-D	0.828	0.94
Inception V3	0.800	0.92
Inception V4	0.840	0.82
Inception-ResNet V2	0.814	0.88

Table 4. Accuracy of different models on the randomly selected ImageNet-1k (validation set) and Caltech-256 samples.

between the size of the MPSs of a given model when it classifies an example correctly or incorrectly, this can be used as an additional check. Such a check would take place post-classification, to determine if the model's decision about the example is in the range of the MPSs of previously encountered correctly or incorrectly classified examples.

As ImageNet-1K labels are slightly "noisy" [19, 38, 40], more work is needed to discover the impact of incorrect human labeling on the size of MPSs. While the problem of ImageNet-1K labels does not affect the size or positioning of MPSs, incorrect human labeling may have an effect on our analysis of MPS size for correct and incorrect classification. With that said, our results indicate that MPSs are too myopic in general, whether they are correctly classified or not.

5. Conclusions and Future Work

In this paper, we present a large scale study on the comparison of MPSs on ImageNet-1k images. We demonstrated comprehensive experimental results on 15 different models across 5 different model architectures. We used a state-of-the-art XAI tool, ReX, based on actual causality, to generate MPSs across the ImageNet-1k dataset. Using formal statistical methods, we demonstrated that there are statistically significant differences in MPS size and location between different architectures and within the same architecture. We additionally showed that there is a small, but statistically significant, increase in explanation size for classifications which do not agree with the ground truth label.

While we are not focused on explainability, the question remains as to whether an MPS is a good explanation for a human. The answer to this question is ultimately qualitative: the datasets examined do not have human-annotated explanations, so do not allow a quantitative comparison between an MPS and a human explanation. Even if we had such information, models often base their classification on different features than humans do, as demonstrated in Figure 5.

Our results find that different models have different degrees of confidence in their MPSs; some model types remain very confident in their classifications, even when the MPS itself is small. We will extend this investigation with the effect of model confidence on the size of MPSs. It is interesting to observe the presence of at least 2 entirely disjoint explanations in Figure 1. This multiplicity of explanations is precisely the object of study of Jiang et al. [18]. Applying a similar methodology as used in this study to look at multiple explanations across multiple architectures may reveal new and interesting insights into model behavior.

Acknowledgements The authors were supported in part by CHAI—the EPSRC Hub for Causality in Healthcare AI with Real Data (EP/Y028856/1).

References

[1] Meghna P. Ayyar, Jenny Benois-Pineau, and Akka Zemmari. Review of white box methods for explanations of convolutional neural networks in image classification tasks. *Journal of Electronic Imaging*, 30(5): 050901, 2021. 2

[2] Nathan Blake, Hana Chockler, David A Kelly, Santiago Calderon Pena, and Akchunya Chanchal. Mrxai: Black-box explainability for image classifiers in a medical setting. *arXiv preprint arXiv:2311.14471*, 2023. 3

[3] C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, pages 13–60, 1936. 4

[4] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. 2

[5] Hana Chockler and Joseph Y. Halpern. Responsibility and blame: A structural-model approach. *J. Artif. Intell. Res.*, 22:93–115, 2004. 3

[6] Hana Chockler and Joseph Y. Halpern. Explaining Image Classifiers. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning*, pages 264–272, 2024. 3

[7] Hana Chockler, David A. Kelly, Daniel Kroening, and Youcheng Sun. Causal explanations for image classifiers, 2024. 1, 2, 3

[8] Hana Chockler, David A Kelly, and Daniel Kroening. Multiple different explanations for image classifiers. In *ECAI European Conference on Artificial Intelligence*, 2025. 1, 2, 3

[9] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26:297—-302, 1945. 3

[10] Yadolah Dodge. *Kruskal-Wallis Test*, pages 288–290. Springer New York, New York, NY, 2008. 4

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. 4

[12] Thomas Eiter and Thomas Lukasiewicz. Complexity results for explanations in the structural-model approach. *Artif. Intell.*, 154(1-2):145–198, 2004. 3

[13] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19358–19369, 2023. 4

[14] Joseph Y. Halpern. *Actual Causality*. The MIT Press, 2019. 3

[15] Joseph Y. Halpern and Judea Pearl. Causes and explanations: A structural-model approach. Part I: Causes. *British Journal for the Philosophy of Science*, 56(4), 2005. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[17] D.P. Huttenlocher, G.A. Klanderman, and W.J. Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993. 3

[18] Mingqi Jiang, Saeed Khorram, and Li Fuxin. Comparing the decision-making mechanisms by transformers and cnns via explanation methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9546–9555, 2024. 1, 2, 8

[19] Nikita Kisel, Illia Volkov, Kateřina Hanzelková, Klara Janouskova, and Jiri Matas. Flaws of imagenet, computer vision's favourite dataset. In *The Fourth Blogpost Track at ICLR 2025*, 2021. 8

[20] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020. 2

[21] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022. 4

[22] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. 2

[23] Sheldon Mascarenhas and Mukul Agarwal. A comparison between vgg16, vgg19 and resnet50 architecture frameworks for image classification. In *2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON)*, pages 96–99, 2021. 2

[24] MR Mustapha, HS Lim, and MZ Mat Jafri. Comparison of neural network and maximum likelihood ap-

proaches in image classification. *Journal of Applied Sciences(Faisalabad)*, 10(22):2847–2854, 2010. 1

[25] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *Advances in Neural Information Processing Systems*, pages 23296–23308. Curran Associates, Inc., 2021. 1

[26] C.H. Papadimitriou. The complexity of unique solutions. *Journal of ACM*, 31:492–500, 1984. 3

[27] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *ArXiv*, abs/1806.07421, 2018. 2

[28] 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*, pages 12116–12128. Curran Associates, Inc., 2021. 1

[29] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, 29(9): 2352–2449, 2017. 2

[30] 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*, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. 2

[31] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *International Conference on Computer Vision (ICCV)*, pages 618–626. IEEE, 2017. 2

[32] Neha Sharma, Vibhor Jain, and Anju Mishra. An analysis of convolutional neural networks for image classification. *Procedia Computer Science*, 132:377–384, 2018. International Conference on Computational Intelligence and Data Science. 2

[33] Vivswan Shitole, Fuxin Li, Minsuk Kahng, Prasad Tadepalli, and Alan Fern. One explanation is not enough: Structured attention graphs for image classification. In *Neural Information Processing Systems (NeurIPS)*, pages 11352–11363, 2021. 1, 2

[34] S. Siegel and N.J. Castellan. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, 1988. 5

[35] T. Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab.*, 5:1—-34, 1948. 3

[36] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy? – a comprehensive study on the robustness of 18 deep image classification models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[37] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 4

[38] Vijay Vasudevan, Benjamin Caine, Raphael Gontijo-Lopes, Sara Fridovich-Keil, and Rebecca Roelofs. When does dough become a bagel? analyzing the remaining mistakes on imagenet. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2022. Curran Associates Inc. 8

[39] Pin Wang, En Fan, and Peng Wang. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recognition Letters*, 141:61–67, 2021. 2

[40] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-labeling imagenet: from single to multi-labels, from global to localized labels. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2340–2350, 2021. 8