Overinterpretation reveals image classification model pathologies

Brandon Carter* MIT CSAIL
Siddhartha Jain* MIT CSAIL
Jonas Mueller* Amazon Web Services
David Gifford MIT CSAIL

Abstract

Image classifiers are typically scored on their test set accuracy, but high accuracy can mask a subtle type of model failure. We find that high scoring convolutional neural networks (CNN) exhibit troubling pathologies that allow them to display high accuracy even in the absence of semantically salient features. When a model provides a high-confidence decision without salient supporting input features we say that the classifier has overinterpreted its input, finding too much class-evidence in patterns that appear nonsensical to humans. Here, we demonstrate that state of the art neural networks for CIFAR-10 and ImageNet suffer from overinterpretation, and find CIFAR-10 trained models make confident predictions even when 95% of an input image has been masked and humans are unable to discern salient features in the remaining pixel subset. Although these patterns portend potential model fragility in real-world deployment, they are in fact valid statistical patterns of the image classification benchmark that alone suffice to attain high test accuracy. We find that ensembling strategies can help mitigate model overinterpretation, and classifiers which rely on more semantically meaningful features can improve accuracy over both the test set and out-of-distribution images from a different source than the training data.

1. Introduction

Well-founded decisions by machine learning (ML) systems are critical for high-stakes applications such as autonomous vehicles and medical diagnosis. Pathologies in models and their respective training datasets can result in unintended behavior during deployment if the systems are confronted with novel situations. For example, a recent medical image classifier for cancer detection attained high accuracy in benchmark test data, but was found to base its decision upon the presence of dermatologists' rulers in an image (present when dermatologists already suspected cancer) [24]. We define model *overinterpretation* to occur when a classifier finds strong class-evidence in regions of an image that contain no semantically salient features. Overinterpretation is related to overfitting, but overfitting can be diagnosed via reduced test accuracy. Overinterpretation can stem from true statistical signals in the underlying dataset distribution that happen to arise from particular properties of the data source (such as the dermatologists' rulers). Thus, overinterpretation can be harder to diagnose as it admits decisions that are made by statistically valid criteria, and models that use such criteria can excel at benchmarks.

It is important to understand how hidden statistical signals of benchmark datasets can result in models that overinterpret or do not generalize to examples that stem from a different distribution. Computer vision (CV) research relies upon datasets like CIFAR-10 [18] and ImageNet [28] to provide standardized performance benchmarks. Here, we analyze the overinterpretation of popular CNN architectures derived from these benchmarks to characterize pathologies.

Revealing overinterpretation requires a systematic way to identify which features are used by a model to reach its decision. Feature attribution is addressed by a large number of interpretability methods, although they propose differing explanations for the decisions of a model. One natural explanation for image classification lies in the set of pixels that is *sufficient* for the model to make a confident prediction, even in the absence of information regarding what is contained in the rest of the image. In our example of the medical image classifier for cancer detection, one might identify the pathological behavior by realizing the pixels depicting the ruler alone suffice for the model to confidently output the same classifications. This idea of Sufficient Input Subsets (SIS) has been proposed to help humans interpret the decisions of black-box models [4]. An SIS subset consists of the smallest subset of features (e.g., pixels) that suffices to yield a class probability above a certain threshold after all other features have been masked.

Here we demonstrate that models trained on CIFAR-10 and ImageNet can base their classification decisions on sufficient input subsets that only contain few pixels and lack human understandable semantic content. Nevertheless, these sufficient input subsets contain statistical signals that generalize across the benchmark data distribution, and we

*BC, SJ, JM contributed equally to conceiving the project and experimental design/analysis. In addition BC led the execution. Correspondence to BC <bcarter@csail.mit.edu> and DG <gifford@mit.edu>.

are able to train equally performing classifiers on CIFAR-10 images that have lost 95% of their pixels. Thus, there exist inherent statistical shortcuts in this benchmark that a classifier solely optimized for accuracy can learn to exploit, instead of having to learn all of the complex semantic relationships between the image pixels and the assigned class label. While recent work suggests adversarially robust classifiers rely on more semantically meaningful features [13], we find these models suffer from severe overinterpretation as well. As we subsequently show, overinterpretation is not only a conceptual issue, but can actually harm overall classifier performance in practice. We find that single ensembling of multiple networks can mitigate overinterpretation, increasing the semantic content of the resulting SIS subsets. Intriguingly, the number of pixels in the SIS rationale behind a particular classification is often indicative of whether this image will be classified correctly or not.

It may seem unnatural to use an interpretability method that produces feature attributions which look uninterpretable. However, we do not want to bias extracted rationales towards human visual priors when analyzing a model for its pathologies, but rather want to faithfully report exactly those features used by a model. To our knowledge, this is the first analysis which shows that one can extract nonsensical features from CIFAR-10 that intuitively should be insufficient or irrelevant for a confident prediction, yet these features alone are sufficient to train a classifier with a minimal loss of performance.

2. Related Work

There has been substantial research on understanding dataset bias in CV [35, 36] and the fragility of image classifiers when applied outside of the benchmark setting [26]. CNNs for image classification in particular have been conjectured to pick up on localized features like texture instead of more global features like object shape [3, 6]. Other research on deep image classifiers has also argued they heavily rely on nonsensical patterns [14, 20], and investigated this issue with artificially-generated patterns that are not in the original benchmark dataset. In contrast, we demonstrate the pathology of overinterpretation with unmodified subsets of actual training images, indicating the patterns are already present in the original dataset. Like us, [12] also recently found that sparse pixel subsets suffice to attain high classification accuracy on popular image classification datasets. In natural language processing (NLP) applications, there has been a recent effort to explore model pathologies using a similar technique [5], but this work does not analyze whether the semantically spurious patterns the models rely on are a statistical property of the dataset. Other research has demonstrated the presence spurious statistical shortcuts present in major NLP benchmarks, showing this problem is not unique to CV [21].

3. Methods

3.1. Data

CIFAR-10 [18] and ImageNet [29] have become two of the most popular image classification benchmarks. Nowadays, most classifiers are evaluated by the CV community based on their accuracy in one of these benchmarks.

We employ two additional datasets to evaluate the extent to which our CIFAR-10 models can generalize to out-of-distribution (OOD) images that stem from a different source than the training data. First, we use the CIFAR-10.1 v6 dataset [25], which contains 2000 class-balanced images drawn from the Tiny Images repository [37] in a similar fashion to that of CIFAR-10, though the authors of [25] found a large drop in classification accuracy on these images. Additionally, we use the CIFAR-10-C dataset [11], which contains variants of CIFAR-10 test images altered by various corruptions (such as Gaussian noise, motion blur, and snow). Where computing sufficient input subsets on CIFAR-10-C images, we use a uniform random sample of 2000 images from the CIFAR-10-C set.

3.2. Models

For CIFAR-10, we explore three common CNN architectures: a deep residual network with depth 20 (ResNet20) [9], a v2 deep residual network with depth 18 (ResNet18) [10], and VGG16 [31]. We train these classifiers using cross-entropy loss optimized via SGD with Nesterov momentum [33] and employ standard data augmentation consisting of random crops and horizontal flips (additional details in Section S1). After training many CIFAR-10 networks individually, we construct four different ensemble classifiers by grouping various networks together. Each ensemble outputs the average prediction over its member networks (specifically, the arithmetic mean of their logits). For each of three architectures, we create a corresponding homogeneous ensemble by individually training five copies of networks that share the same architecture. Each network has a different random initialization, which suffices to produce substantially-different models despite the fact these replicate architectures are all trained on the same data [22]. Our fourth ensemble is heterogeneous, containing all 15 networks (5 replicates of each of 3 distinct CNN architectures).

For ImageNet, we use a pre-trained Inception-v3 model [34] available in PyTorch [23]. This network achieves 22.55% and 6.44% top-1 and top-5 error on ImageNet, respectively [23].

3.3. Interpreting Learned Features

We interpret the feature patterns learned by our models using the sufficient input subsets (SIS) procedure [4], which produces rationales of a pre-trained model's decision-

making by applying backward selection locally on individual examples. These rationales are comprised of sparse subsets of input features (pixels) on which the model makes the same decision as on the original input (with the rest of pixels masked), up to a specified confidence threshold.

More formally, let $0 \leq \tau \leq 1$ be a threshold for prediction confidence. Let f predict that an image x belongs to class c with probability f_c. Let U be the total set of pixels. Then an SIS subset $S \subseteq U$ is a minimal subset of pixels such that $f_c(x_S) \geq \tau$ where the information about the pixels $R = U \setminus S$ is considered to be missing. We mask pixels in R by replacement with the mean pixel value over the entire image dataset (equal to zero when the image data has been normalized), which is presumably least informative to a trained classifier [4]. We apply SIS to the function giving the confidence toward the predicted (most likely) class. We also develop an approximation of the backward selection procedure to efficiently scale the SIS-finding procedure to higher-resolution images from ImageNet (details in Section S5).

We produce sparse variants of CIFAR-10 images where we retain the values of 5% of pixels in the image, while masking the remainder. Our goal is to identify sparse pixel subsets that contain feature patterns the model identifies as strong class-evidence as it classifies an image. We identify such pixel-subsets by local backward selection on each image as in the `BackSelect` procedure of SIS [4]. We apply backward selection to f_c, which iteratively removes pixels that lead to the smallest decrease in f_c. Our 5% pixel-subset images contain the final 5% of pixels as ordered by backward selection (with their same RGB values as in the original image) while all other pixels' values are replaced with zero.

3.4. Human Classification Benchmark

To evaluate whether sparse pixel-subsets of images can be accurately classified by humans, we asked four participants to classify images containing various degrees of masking. We randomly sampled 100 images from the CIFAR-10 test set (10 images per class) that were correctly and confidently ($\geq 99\%$ confidence) classified by our models, and for each image, kept only 5%, 30%, or 50% of pixels as ranked by backward selection (all other pixels masked). Backward selection image subsets are sampled across our three models. Since larger subsets of pixels are by construction supersets of smaller subsets identified by the same model, we presented each batch of 100 images in order of increasing subset size and shuffled the order of images within each batch. Users were asked to classify each of the 300 images as one of the 10 classes in CIFAR-10 and were not provided training images. The same task was given to each user (and is provided in Section S4).

4. Results

4.1. CNNs Classify Images Using Spurious Features

We train five replicate models of each of our three architectures (ResNet20, ResNet18, VGG16) on the CIFAR-10 training set (see Section 3.2). Table 1 shows the final model accuracies on the CIFAR-10 test set and CIFAR-10.1 and CIFAR-10-C (out-of-distribution) test sets.

To interpret the behavior of these models, we apply the sufficient input subset (SIS) interpretability procedure [4] to identify minimal subsets of features in each image that suffice for the model to make the same prediction as on the full image (see Section 3.3). For SIS, we use a confidence threshold of 0.99 and mask pixels by replacement with zeros. Figure 1 shows examples of sufficient input subsets from a randomly chosen set of CIFAR-10 test images, which are confidently and correctly classified by each model (additional examples in Section S2). Each SIS shown is classified by its corresponding model with $\geq 99\%$ confidence toward the predicted class. This result suggests that our CNNs confidently predict on images that appear nonsensical to humans (see Section 4.3), which leads to concern about their robustness and generalizability.

We observe that these sufficient input subsets are highly sparse and that the average SIS size at this threshold is $< 5\%$ of each image, so we create a sparsified variant of all CIFAR-10 images (both train and test). As in SIS, we apply backward selection locally on each image to rank pixels by their contribution to the predicted class (as described in Section 3.3). We retain 5% of pixels as ordered by backward selection on each image and mask the remaining 95% with zeros. Note that because backward selection is applied locally on each image, the specific pixels retained differ across images.

We first verify that the original models are able to classify these sparsified images just as accurately as their full image counterparts (Table 1). Moreover, the predictions on the pixel-subsets are just as confident: the mean drop in confidence for the predicted class between original images and these 5% subsets is -0.035 (std dev. $= 0.107$), -0.016 (0.094), and -0.012 (0.074) computed over all CIFAR-10 test images for our ResNet20, ResNet18, and VGG16 models, respectively, which suggests severe overinterpretation by each model (negative values imply greater confidence on the 5% subsets). We also find that these pixel subsets chosen through backward selection are more predictive than equally large pixel-subsets chosen uniformly at random from each image (Table 1), on which the models are unable to predict as accurately as on the original images or on the pixel-subsets found through backward selection. Figure 2 shows the frequency of each pixel location in the 5% backward selection pixel-subsets derived from each model across all CIFAR-10 test images.

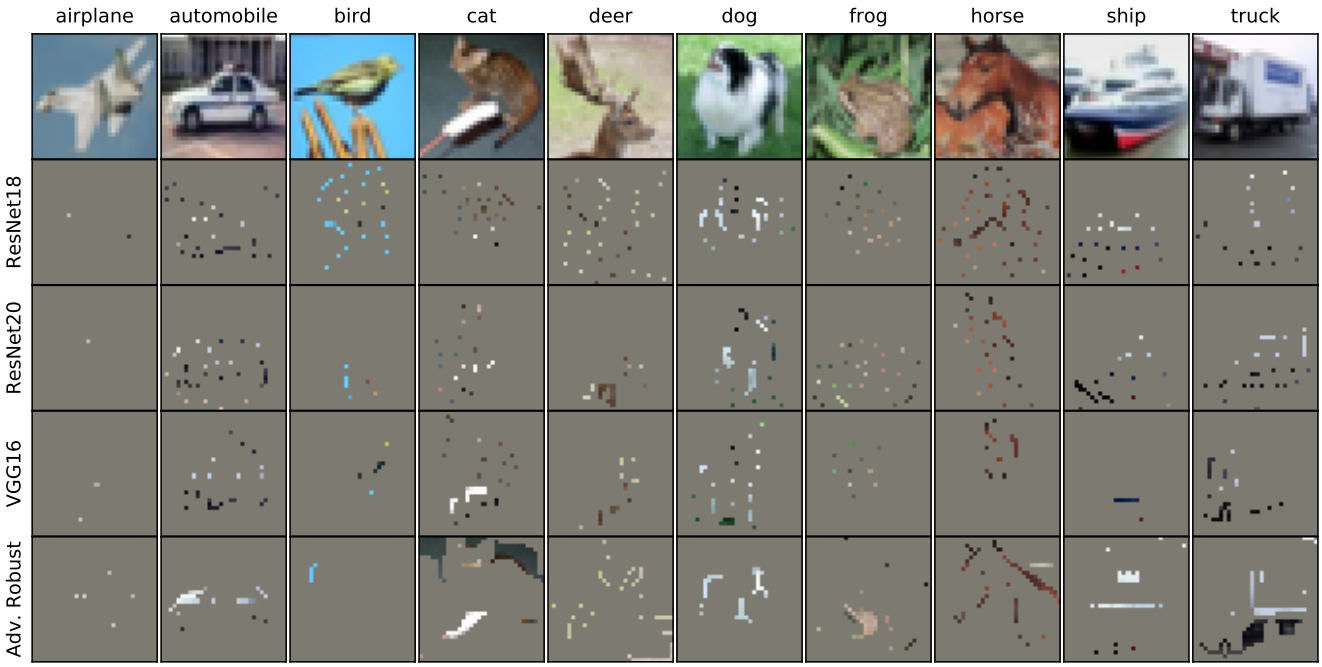


Figure 1: Sufficient input subsets (SIS) for a sample of CIFAR-10 test images (top). Each SIS image shown below is classified by the respective model with $\geq 99\%$ confidence. The "Adv. Robust" (pre-trained adversarially robust) model we use is from [19] and robust to l_∞ perturbations.

We additionally find that the SIS subsets for one model do not transfer to other models. That is, a sparse pixel subset which one model confidently classified is typically not confidently identified by the other models. For instance, 5% pixel-subsets derived from CIFAR-10 test images using one ResNet18 model (which classifies them with 94.8% accuracy) are only classified with 27.6%, 29.2%, and 27.5% accuracy by another ResNet18 replicate, ResNet20, and VGG16 models, respectively. This result suggests there exist many different statistical patterns that a flexible model might learn to rely on, and thus CIFAR-10 image classification remains a highly under-determined problem. Producing high-capacity classifiers that make the right predictions for the right reasons may require clever regularization strategies and architecture design to ensure the model favors salient features over such sparse pixel subsets.

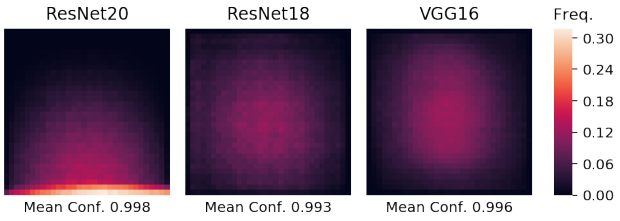


Figure 2: Heatmaps of pixel locations comprising 5% pixel-subsets across CIFAR-10 test set for each model. Frequency indicates fraction of subsets containing each pixel. Mean confidence indicates confidence on 5% pixel-subsets.

4.1.1 Analysis on ImageNet

We also find that models trained on the higher-resolution images from ImageNet suffer from severe overinterpretation. As it is computationally infeasible to scale the original backward selection procedure of SIS [4] to ImageNet, we introduce a more efficient gradient-based approximation to the original SIS procedure that enables us to find sufficient input subsets on ImageNet images (details in Section S5). Figure 3 shows examples of images confidently classified by Inception-v3, along with the corresponding SIS subsets that identify which pixels alone suffice to for the network to reach a similarly confident prediction (additional examples are provided in Figure S6). These sufficient input subsets appear visually nonsensical, yet the network nevertheless classifies them with $\geq 90\%$ confidence. Of great concern is the fact that nearly none of the SIS pixels are located within the actual object that determines the class label. For example, in the "pizza" image, the SIS is concentrated on the shape of the plate and the background table, rather than the pizza itself, which indicates that the model could generalize poorly when the image contains a different circular item on the table. In the "giant panda" image, the SIS

Model	Train On	Evaluate On	CIFAR-10 Test Acc.	CIFAR-10.1 Acc.	CIFAR-10-C Acc.
ResNet20	Full Images	Full Images	92.23 ± 0.35	83.85 ± 0.75	69.99 ± 0.67
		5% BS Subsets	92.48	82.80	70.65
		5% Random	9.99 ± 0.07	10.01 ± 0.02	10.03 ± 0.03
	5% BS Subsets	5% BS Subsets	92.50 ± 0.02	82.74 ± 0.02	70.57 ± 0.08
	5% Random	5% Random	50.05 ± 0.18	39.53 ± 0.24	43.89 ± 0.15
ResNet18	Full Images	Full Images	95.28 ± 0.17	89.07 ± 0.70	75.28 ± 0.51
		5% BS Subsets	94.76	89.35	75.15
		5% Random	10.01 ± 0.15	10.08 ± 0.12	10.02 ± 0.08
	5% BS Subsets	5% BS Subsets	94.97 ± 0.04	89.57 ± 0.08	75.27 ± 0.08
	5% Random	5% Random	51.20 ± 0.78	39.79 ± 1.18	44.77 ± 0.56
VGG16	Full Images	Full Images	93.62 ± 0.10	86.07 ± 0.59	73.96 ± 0.59
		5% BS Subsets	93.27	86.45	73.95
		5% Random	9.97 ± 0.17	10.02 ± 0.24	10.08 ± 0.12
	5% BS Subsets	5% BS Subsets	92.56 ± 0.05	85.65 ± 0.16	73.26 ± 0.22
	5% Random	5% Random	53.80 ± 1.31	41.32 ± 1.30	47.19 ± 1.02
Ensemble (5x ResNet18)	Full Images	Full Images	96.15	90.50	77.21
		5% Random	9.98	10.00	10.00

Table 1: Accuracy of various models on CIFAR-10 images trained and evaluated on full images, 5% backward selection (BS) pixel-subsets, and 5% randomly chosen pixel-subsets. Where possible, we report accuracy given as mean $\pm$ standard deviation (%) over five runs. For training/evaluation on BS pixel-subsets, we only run backward selection on all CIFAR-10 images for a single model of each type, but average over five models trained on these subsets.

contains bamboo, which likely appeared in the collection of ImageNet photos for this class. In the "traffic light" and "street sign" images, the SIS is focused on the sky, suggesting that autonomous vehicle systems that may depend on these models should be carefully evaluated for overinterpretation pathologies.

We randomly sample 1000 images from the ImageNet validation set that are classified with $\geq 90\%$ confidence and generate a heatmap of sufficient input subset pixel locations (Figure 4). Here, we use SIS subsets to generate the heatmap rather than 5% pixel-subsets. The SIS tend to be strongly concentrated along the image borders rather than near the center, suggesting the model relies too heavily on image backgrounds in its decision-making. This is a serious problem because objects corresponding to ImageNet classes are often located near the center of images, and thus this network fails to focus on salient features. The fact that the model confidently classifies the majority of images by seeing only their border pixels suggests it suffers from severe overinterpretation.

4.2. Sparse Subsets are Real Statistical Patterns

CNNs are known to be overconfident for image classification [8]. Thus one might reasonably wonder whether the overconfidence on the semantically meaningless SIS subsets is an artifact of CNN overconfidence rather than a true statistical signal in the dataset. To probe this question, we evaluate whether the CIFAR-10 sparse 5% image subsets contain sufficient information to train a new classifier to solve the same task. We run our backward selection procedure on all train and test images in CIFAR-10 using one of our three model architectures (chosen at random). We then train a new model of the same type on these 5% pixel-subset variants of the CIFAR-10 training images. We use the same training setup and hyperparameters as with the original models (see Section 3.2) without data augmentation of training images (results with data augmentation in Section S3). Note that we apply backward selection to the function giving the confidence of the *predicted* class from the original model, which prevents leaking information about the true class for misclassified images, and we use the true labels for training new models on pixel-subsets. As a baseline to the 5% pixel-subsets identified by backward selection, we create variants of all CIFAR-10 images where the 5% pixel-subsets are selected at random from each image (rather than by backward selection). We use the same random pixel-subsets for training each new model.

As shown in Table 1, models trained solely on these 5% backward selection image subsets can classify corresponding 5% test image subsets nearly as accurately as models trained and evaluated on full images. Models trained on random 5% pixel-subsets of images have significantly lower accuracy on test images (Table 1) compared to models trained on 5% pixel-subsets found through backward

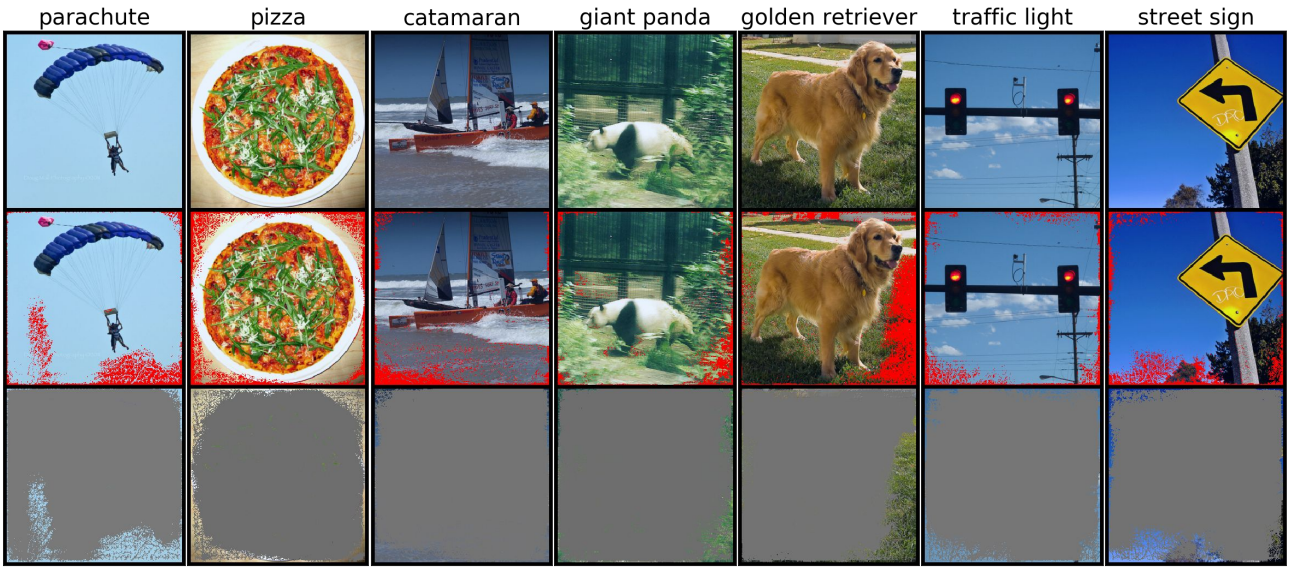


Figure 3: Sufficient input subsets for images from the ImageNet validation set (top). The middle row shows the location of the SIS pixels (red) and the bottom row shows the image with all pixels outside of the SIS masked, which is still classified by the Inception-v3 model with $\geq 90\%$ confidence.

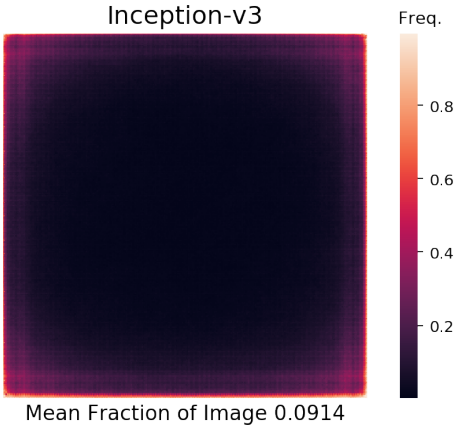


Figure 4: Heatmap of pixel locations comprising sufficient input subsets (threshold 0.9) across ImageNet validation images from Inception-v3. Frequency indicates fraction of SIS containing each pixel.

selection of existing models. This result suggests that the highly sparse subsets found through backward selection offer a valid predictive signal in the CIFAR-10 benchmark that can be exploited by models to attain high test accuracy.

4.3. Humans Struggle to Classify Sparse Subsets

Table 2 shows the accuracy achieved by humans asked to classify our sparse pixel subsets (Section 3.4). Unsurprisingly, there is strong correlation between the fraction of unmasked pixels in each image and human classification accuracy. Human classification accuracy on pixel subsets of CIFAR-10 is significantly lower than accuracy when presented original, unmasked images (estimated around 94% in previous work [16]). Moreover, human accuracy on 5% pixel-subsets is very poor, though greater than purely random guessing. Presumably this effect is due to correlations between features such as color in images (for example, blue pixels near the top of an image may indicate a sky, and hence increase likelihood for certain CIFAR-10 classes such as airplane, ship, and bird).

However, CNNs (even when trained on full images to achieve accuracy on par with human accuracy on full images) can classify these sparse image subsets with very high accuracy (Table 1, Section 4.2). This indicates the benchmark images contain statistical signals that are unknown to humans. Models solely trained to minimize prediction error may thus latch onto these signals while still accurately generalizing to the test set, but such models may behave counterintuitively when fed images from a different source which does not share these exact statistics. The strong correlation ($R^2 = 0.94$, Figure S5) between the size of pixel subsets found through backward selection and the corresponding human classification accuracy clearly suggests that larger subsets contain greater semantic content and more salient features. Thus, a model whose confident classifications have corresponding sufficient input subsets that are larger in size is presumably better than a model with smaller SIS subsets, as the former model exhibits less over-interpretation. We investigate this further in Section 4.4.

Fraction of Images	Human Classification Acc. (%)
5%	19.2 ± 4.8
30%	40.0 ± 2.5
50%	68.2 ± 3.6

Table 2: Human classification accuracy on a sample of CIFAR-10 test image pixel-subsets of varying sparsity (see Section 3.4). Accuracies given as mean $\pm$ standard deviation.

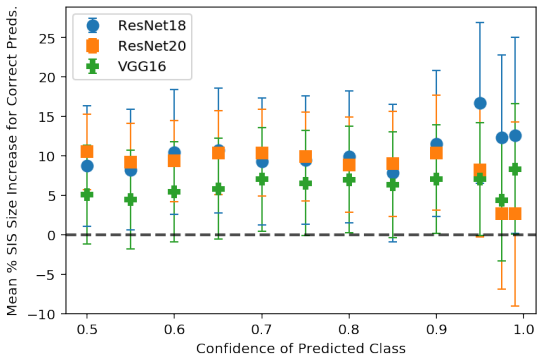
4.4. SIS Size is Predictive of Model Accuracy

Given that smaller SIS contain fewer salient features according to human classifiers, models that justify their classifications based on these sparse SIS may be limited in terms of attainable accuracy, particularly in out-of-distribution settings. Here, we investigate the relationship between a model's predictive accuracy and the size of the SIS subsets in which it identifies class-evidence. For each of our three classifiers, we compute the average SIS size increase for correctly classified images as compared to incorrectly classified images (expressed as a percentage) for both the CIFAR-10 test set and out-of-distribution CIFAR-10-C test set. Figure 5 (A for CIFAR-10 test set, B for CIFAR-10-C test set) shows that for varying SIS confidence thresholds, SIS subsets of correctly classified images are consistently significantly larger than those of misclassified images. This is especially striking in light of the fact that model confidence is uniformly lower on the misclassified inputs, as one would hope (Figure S3). Lower confidence would normally imply a larger SIS subset at a given confidence level, as one expects that fewer pixels can be masked before the model's confidence drops below the SIS confidence threshold. Thus, we can rule out overall model confidence as an explanation of the smaller SIS in misclassified images. This result suggests that the sparse SIS subsets highlighted in this paper are not just a curiosity, but may be leading to bad generalizations on real images.

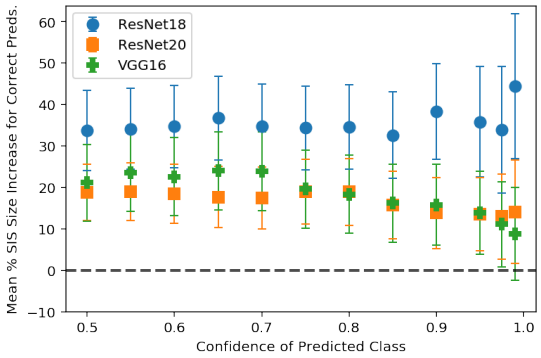
We notice similar behavior by comparing SIS subset size and model accuracy at varying confidence thresholds (Figure 6). Models with superior accuracy have higher SIS size and thus tend to suffer less from model overinterpretation.

4.5. Pathologies in Adversarially Robust Models

Recent work has suggested semantics can be better captured via models that are robust to adversarial inputs, which fool standard neural networks via human-imperceptible modifications to images [19, 30]. Here, we find that models trained to be robust to adversarial attacks classify the highly sparse sufficient input subsets as confidently as the models in Section 4.1. We use a pre-trained wide residual network provided by [19] that is adversarially robust for



(a) CIFAR-10 test set



(b) CIFAR-10-C test set

Figure 5: Percentage increase in mean SIS size of correctly classified images compared to misclassified images across (a) the CIFAR-10 test set and (b) a random sample of CIFAR-10-C test set. Positive values indicate larger mean SIS size for correctly classified images. Error bars indicate 95% confidence interval for the difference in means.

CIFAR-10 classification (trained against an iterative adversary that can perturb each pixel by at most $\varepsilon = 8$). Figure 1 ("Adv. Robust") shows examples of sufficient input subsets identified for a sample of CIFAR-10 test images. The adversarially robust model classifies each SIS image shown with $\geq 99\%$ confidence. We find that the property of adversarial robustness alone is insufficient to prevent models from overinterpreting sparse feature patterns in CIFAR-10, and these models confidently classify images that are indiscernible to humans.

4.6. Ensembling Mitigates Overinterpretation

Model ensembling is a well-known technique to improve classification performance [7, 15]. Here we test whether ensembling alleviates the overinterpretation problem as well. We explore both homogeneous and heterogeneous ensembles of our individual models (see Section 3.2). We show that SIS subset size is strongly correlated with human accuracy on image classification (Section 4.3). Thus our metric

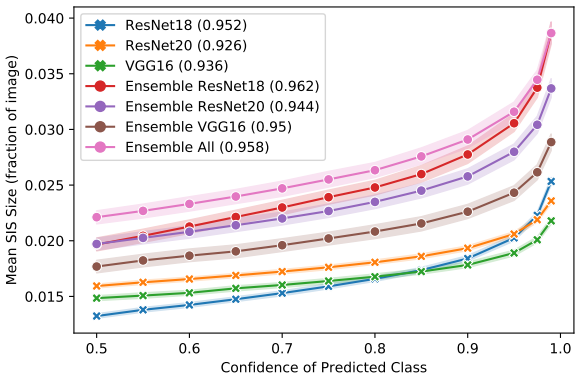


Figure 6: Mean SIS size on CIFAR-10 test images as SIS threshold varies. Corresponding model accuracies are shown in the legend. SIS size indicates fraction of pixels needed for model to make the same prediction at each confidence. Shaded region indicates 95% confidence interval around each mean.

for measuring how much ensembling can alleviate the problem is the increase in SIS subset size. Figure 6 shows that ensembling uniformly increases the model accuracy which is expected but also increases the SIS size (and given results from Section 3.4 on humans), mitigating the overinterpretation problem.

We conjecture that the cause of both the increase in the accuracy and SIS size for ensembles is the same. In our experiments we observe that SIS pixel-subsets are generally not transferable from one model to another — i.e., an SIS for one model is rarely an SIS for another (see Section 4.1). Thus, different models often consider independent pieces of evidence to arrive at the same prediction. Ensembling forces the consideration of the independent sources of evidence together for its prediction, increasing the accuracy of the prediction and forcing the SIS size to be larger by requiring simultaneous activation of multiple independently trained feature detectors. We find that the ensemble's SIS are larger than the SIS of its individual members (examples in Figure S2).

5. Discussion

We find that state of the art image classifiers overinterpret small nonsensical patterns present in popular benchmark datasets, identifying strong class evidence in the pixel subsets that constitute these patterns. Despite their lack of salient features, these sparse pixel subsets are underlying statistical signals that suffice to accurately generalize from the benchmark training data to the benchmark test data. We found that different models rationalize their predictions based on different sufficient input subsets, suggesting that optimal image classification rules remain highly underdetermined by the training data. Models with superior accuracy tend to suffer less from model overinterpretation, which suggests that reducing overinterpretation can lead to more accurate models. In high-stakes image classification applications, we recommend using ensembles of diverse networks rather than relying on just a single model.

Our results call into question model interpretability methods whose outputs are encouraged to align with prior human beliefs regarding proper classifier operating behavior [1]. Given the existence of non-salient pixel subsets which alone suffice for correct classification, a model might solely rely on those patterns in its predictions. In this case, an interpretability method that faithfully describes the model should output these nonsensical rationales, whereas interpretability methods that bias rationales toward human priors may produce results that mislead users to think their models are behaving as intended.

Mitigating model overinterpretation and the broader task of ensuring classifiers are accurate for the right reasons remain significant challenges for ML. While we discovered ensembling tends to help, pathologies remain even for heterogeneous ensembles of classifiers. One alternative is to regularize CNNs by constraining the pixel attributions generated via a saliency map [27, 32, 38]. Unfortunately, such methods require a human image annotator that highlights the correct pixels as an auxiliary supervision signal. Furthermore, saliency maps have been shown to provide unreliable insights into the operating behavior of a classifier and must be interpreted as approximations [17]. In contrast, our SIS subsets constitute actual pathological examples that have been misconstrued by the model.

Future work should investigate regularization strategies and architectures to identify how to better learn semantically-aligned features without explicit supervision. Imposing the right inductive bias is critical given the issue of underdetermination from multiple sets of non-salient patterns that serve as valid statistical signals in benchmarks. Before deploying current image classifiers in critical situations, it is imperative to assemble benchmarks composed of a greater diversity of image sources in order to reduce the likelihood of spurious statistical patterns [2].

Acknowledgements

This work was supported by the National Institutes of Health [R01CA218094] and Schmidt Futures.

References

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

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David

Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[3] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019.

[4] Brandon Carter, Jonas Mueller, Siddhartha Jain, and David Gifford. What made you do this? Understanding black-box decisions with sufficient input subsets. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 567–576, 2019.

[5] Shi Feng, Eric Wallace, II Grissom, Mohit Iyyer, Pedro Rodriguez, Jordan Boyd-Graber, et al. Pathologies of neural models make interpretations difficult. *arXiv preprint arXiv:1804.07781*, 2018.

[6] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Texture and art with deep neural networks. *Current Opinion in Neurobiology*, 46:178–186, 2017.

[7] King-Shy Goh, Edward Chang, and Kwang-Ting Cheng. Svm binary classifier ensembles for image classification. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 395–402. ACM, 2001.

[8] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org, 2017.

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

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

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

[12] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 9734–9745, 2019.

[13] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019.

[14] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019.

[15] Cheng Ju, Aurélien Bibaut, and Mark van der Laan. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of Applied Statistics*, 45(15):2800–2818, 2018.

[16] Andrej Karpathy. Lessons learned from manually classifying cifar-10. *Published online at http://karpathy. github. io/2011/04/27/manually-classifying-cifar10*, 2011.

[17] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280. Springer, 2019.

[18] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

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

[20] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, 2015.

[21] Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. *ACL*, 2019.

[22] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. In *Advances in neural information processing systems*, pages 4026–4034, 2016.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[24] Neel V. Patel. *Why Doctors Arent Afraid of Better, More Efficient AI Diagnosing Cancer*, Dec 22, 2017 (accessed Nov 11, 2019).

[25] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

[26] Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

[27] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017.

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[30] Shibani Santurkar, Dimitris Tsipras, Brandon Tran, Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Computer vision with a single (robust) classifier. *arXiv preprint arXiv:1906.09453*, 2019.

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

[32] Becks Simpson, Francis Dutil, Yoshua Bengio, and Joseph Paul Cohen. Gradmask: Reduce overfitting by regularizing saliency. *arXiv preprint arXiv:1904.07478*, 2019.

[33] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum

in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

[34] 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.

[35] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. A deeper look at dataset bias. In *Domain adaptation in computer vision applications*, pages 37–55. Springer, 2017.

[36] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[37] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

[38] Joseph D Viviano, Becks Simpson, Francis Dutil, Yoshua Bengio, and Joseph Paul Cohen. Underwhelming generalization improvements from controlling feature attribution. *arXiv preprint arXiv:1910.00199*, 2019.

Supplementary Information for: Overinterpretation reveals image classification model pathologies

S1. Details of Models and Training

Here we provide implementation and training details for the models used in this paper (Section 3.2). The ResNet20 architecture [9] has 16 initial filters and a total of 0.27M parameters. ResNet18 [10] has 64 initial filters and contains 11.2M parameters. Our VGG16 architecture [31] uses batch normalization and contains 14.7M parameters.

All models are trained for 200 epochs with a batch size of 128. We minimize cross-entropy via SGD with Nesterov momentum [33] using momentum of 0.9 and weight decay of 5e-4. The learning rate is initialized as 0.1 and is reduced by a factor of 5 after epochs 60, 120, and 160. Datasets are normalized using per-channel mean and standard deviation, and we use standard data augmentation training strategies [10].

The adversarially robust model we evaluated is the `adv_trained` model of [19], available on GitHub[1].

To apply the SIS procedure to CIFAR-10 images, we use an implementation available on GitHub[2]. For confidently classified images on which we run SIS, we find one sufficient input subset per image using the `FindSIS` procedure. When masking pixels, we mask all channels of each pixel as a single feature.

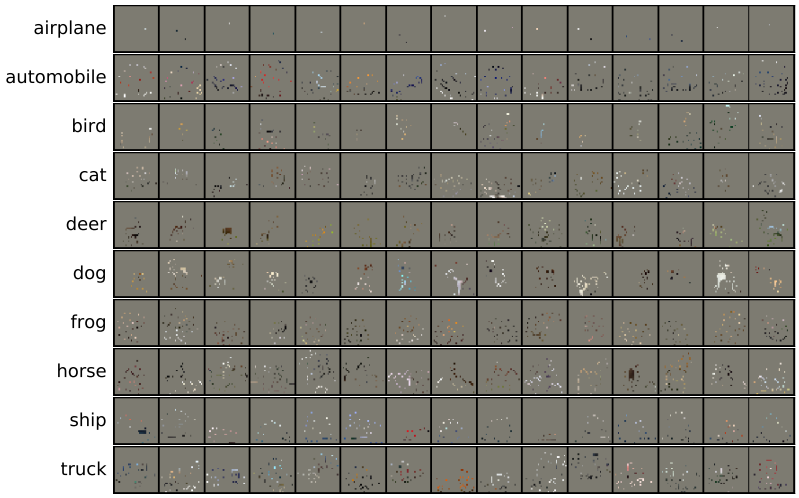
S2. Additional Examples of CIFAR-10 Sufficient Input Subsets

SIS of Individual Networks

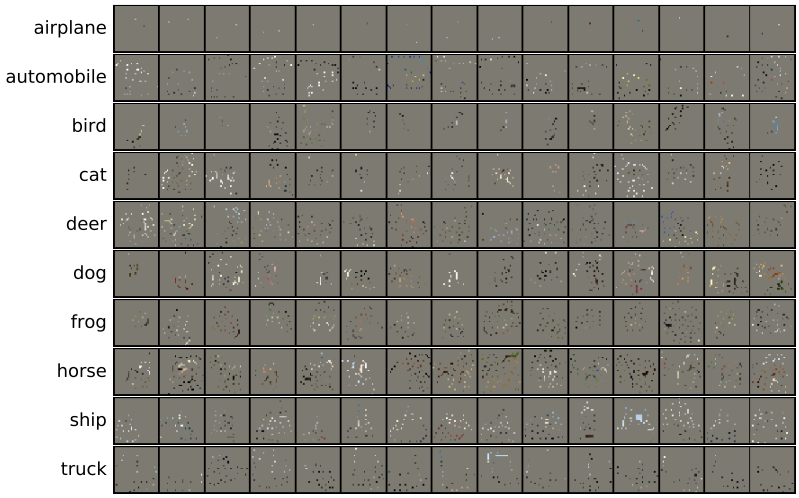
Figure S1 shows a sample of SIS for each of our three architectures. These images were randomly sampled among all CIFAR-10 test images confidently (≥ 0.99) predicted to belong to the class written on the left. SIS are computed under a threshold of 0.99, so all images shown in this figure are classified with probability $\geq 99\%$ confidence as belonging to the listed class.

[1] https://github.com/MadryLab/cifar10_challenge

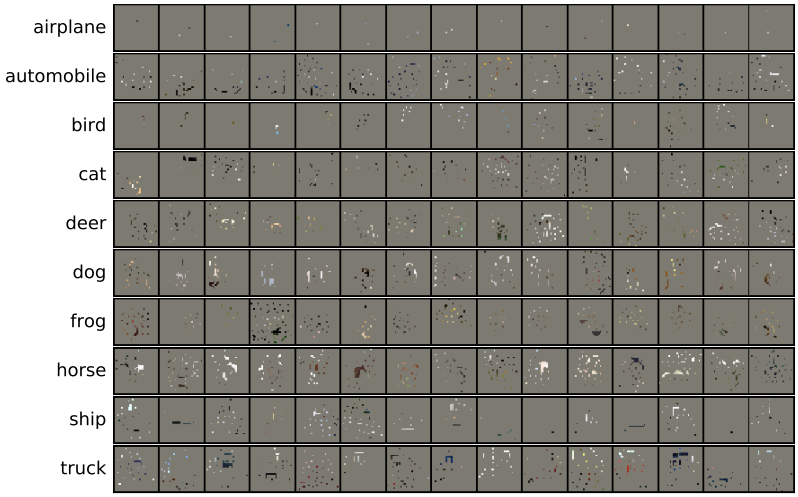
[2] https://github.com/google-research/google-research/blob/master/sufficient_input_subsets/sis.py



(a) ResNet20



(b) ResNet18



(c) VGG16

Figure S1: Examples of SIS (threshold $= 0.99$) on random sample of CIFAR-10 test images (15 per class, different random sample for each architecture). All images shown here are predicted to belong to the listed class with $\geq 99\%$ confidence.

SIS of Ensemble

Figure S2 shows examples of SIS from one of our model ensembles (a homogeneous ensemble of ResNet18 networks, see Section 3.2), along with corresponding SIS for the same image from each of the five member networks in the ensemble. We use a SIS threshold of 0.99, so all images are classified with confidence $\geq 99\%$. These examples highlight how the ensemble SIS are larger and draw class-evidence from the individual members' SIS.

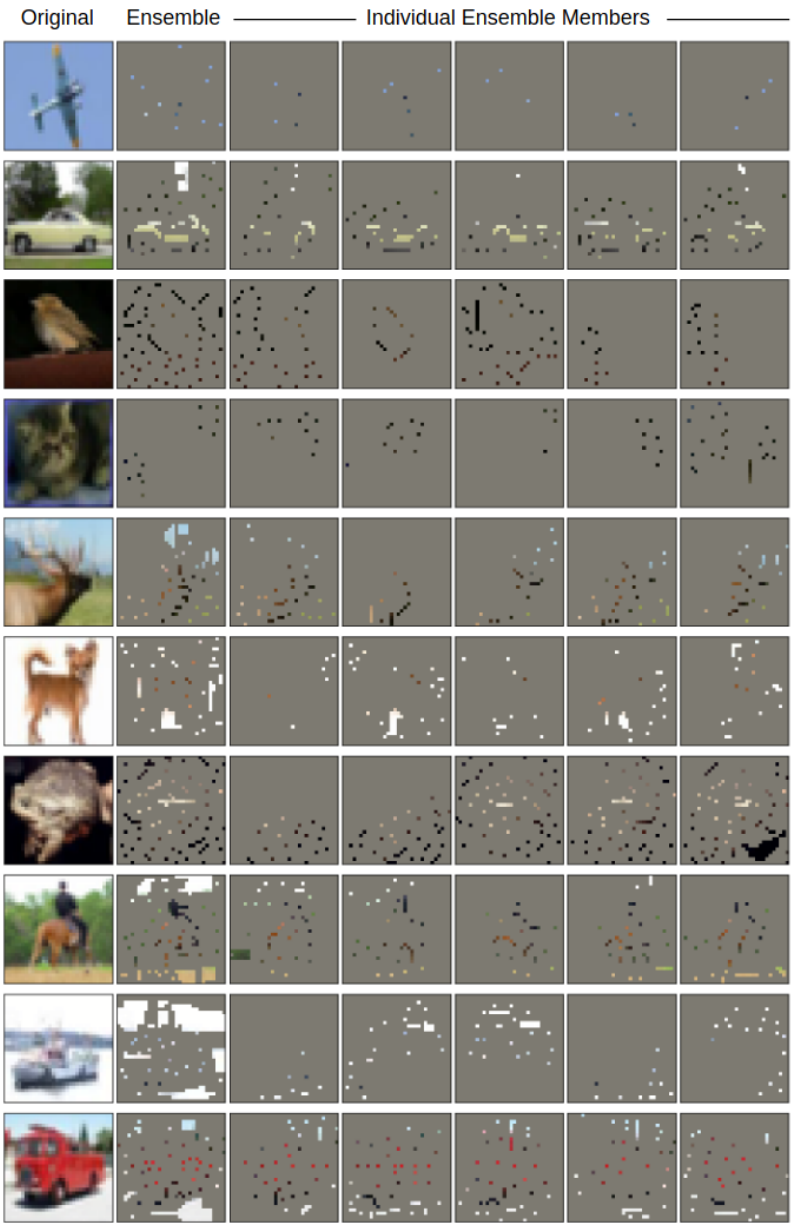


Figure S2: Examples of SIS from ResNet18 homogeneous ensemble (see Section 3.2) and its member models. Each row shows original CIFAR-10 image (left), followed by SIS from the ensemble (second column) and the SIS from each of its 5 member networks (remaining columns). Each image shown is classified with $\geq 99\%$ confidence by its respective network.

S3. Additional Model Performance Results

Training on Pixel-Subsets Without Data Augmentation

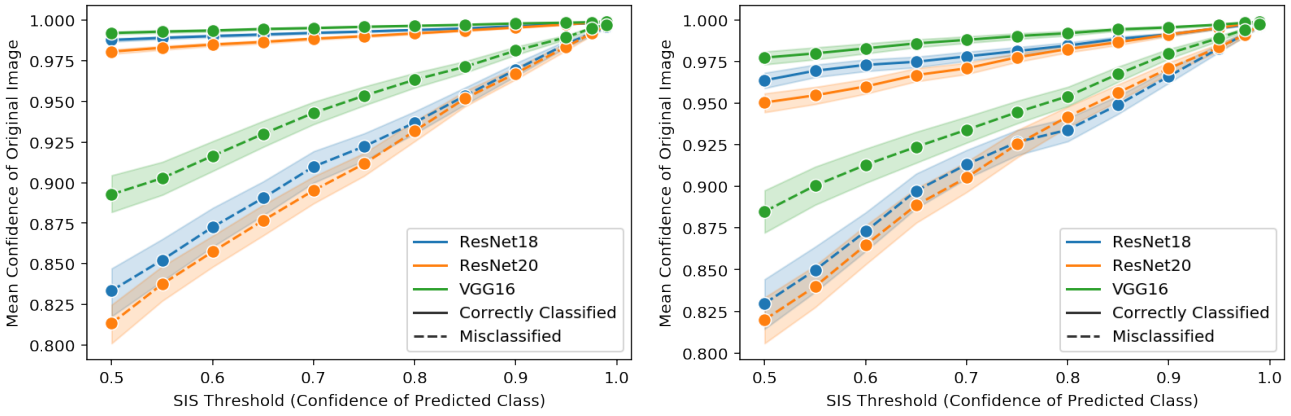
In Table S1, we present results akin to those in Section 4.2 and Table 1, but where the models here are trained on 5% pixel-subsets are trained with data augmentation. We find that training without data augmentation slightly improves accuracy when training models on 5% pixel-subsets.

Model	Train On	Evaluate On	CIFAR-10 Test Acc.	CIFAR-10.1 Acc.	CIFAR-10-C Acc.
ResNet20	5% BS Subsets (+)	5% BS Subsets	92.23 ± 0.03	82.42 ± 0.12	70.33 ± 0.14
	5% Random (+)	5% Random	48.85 ± 0.17	37.52 ± 0.30	42.58 ± 0.13
ResNet18	5% BS Subsets (+)	5% BS Subsets	94.67 ± 0.02	89.11 ± 0.13	75.00 ± 0.06
	5% Random (+)	5% Random	48.69 ± 0.92	37.74 ± 0.97	42.77 ± 0.52
VGG16	5% BS Subsets (+)	5% BS Subsets	91.13 ± 0.12	84.07 ± 0.24	72.16 ± 0.19
	5% Random (+)	5% Random	51.55 ± 1.14	39.96 ± 2.68	44.93 ± 1.05

Table S1: Performance of various models on CIFAR-10 images trained and evaluated on 5% backward selection (BS) image subsets and 5% randomly chosen image subsets *with* data augmentation (+). Accuracy given as mean $\pm$ standard deviation (%) over five runs. For results without data augmentation, see Table 1 in the main text.

Additional Analysis for SIS Size and Model Accuracy

Figure S3 shows the mean confidence of each group of correctly and incorrectly classified images that we consider at each confidence threshold (at each confidence threshold along the x-axis, we evaluate SIS size in Figure 5 on the set of images that originally were classified with at least that level of confidence). We find that as one would hope, model confidence is uniformly lower on the misclassified inputs.



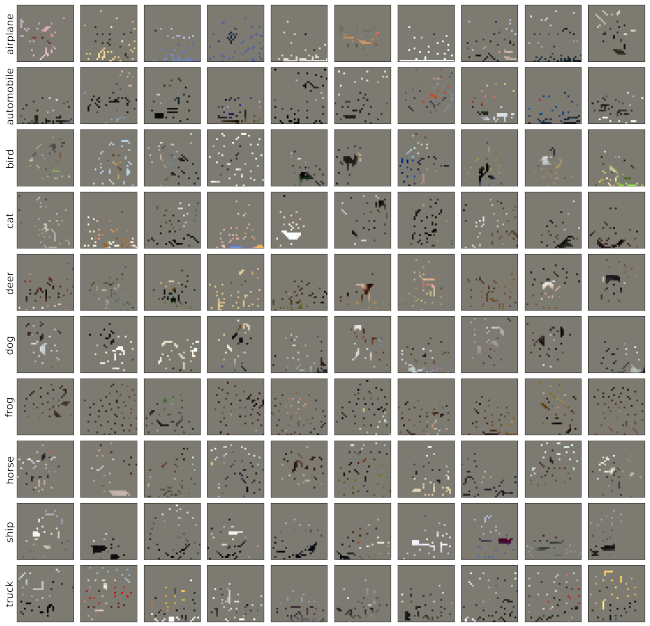
(a) CIFAR-10 test set

(b) CIFAR-10-C test set

Figure S3: Mean confidence of correctly vs. incorrectly classified images for each corresponding SIS threshold we evaluate in Figure 5 across the (a) CIFAR-10 test set and (b) our random sample of the CIFAR-10-C test set. Shaded region indicates 95% confidence interval.

S4. Details of Human Classification Benchmark

Here we include additional details on our benchmark of human classification accuracy of sparse pixel subsets (Section 3.4). Figure S4 shows all images shown to users (100 images each for 5%, 30% and 50% pixel-subsets of CIFAR-10 test images). Each set of 100 images has pixel-subsets stemming from each of the three architectures roughly equally (35 ResNet20, 35 ResNet18, 30 VGG16). Figure S5 depicts the correlation between human classification accuracy and pixel-subset size.



(a) 5% Pixel-Subsets

(b) 30% Pixel-Subsets

(c) 50% Pixel-Subsets

Figure S4: Pixel-subsets of CIFAR-10 test images shown to participants in our human classification benchmark (Section 3.4).

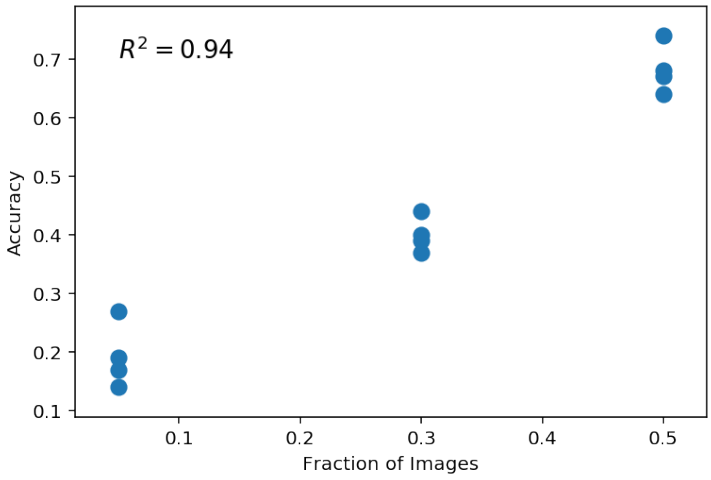


Figure S5: Human classification accuracy on a sample of CIFAR-10 test image pixel-subsets (see Section 3.4).

S5. Scaling SIS to ImageNet

It is computationally infeasible to scale the original backward selection procedure of SIS [4] to ImageNet. As each ImageNet image contains $299 \times 299 = 89401$ pixels, running backward selection to find one SIS for an image would require ~ 4 billion forward passes through the network. Here we introduce a more efficient gradient-based approximation to the original SIS procedure (via **Batched Gradient SIScollection**, **Batched Gradient BackSelect**, and **Batched Gradient FindSIS**) that allows us to find SIS on larger ImageNet images in a reasonable time. The **Batched Gradient SIScollection** procedure described below identifies a complete collection of disjoint masks for an input $\mathbf{x}$, where each mask M specifies a pixel-subset of the input $\mathbf{x}_S = \mathbf{x} \odot (1 - M)$ such that $f(\mathbf{x}_S \geq \tau)$. Here f outputs the probability assigned by the network to its predicted class (i.e., its confidence).

The idea behind our approximation algorithm is two-fold: (1) Instead of separately masking every remaining pixel to find the least critical pixel (whose masking least reduces the confidence in the network's prediction), we use the *gradient* with respect to the mask as a means of ordering. (2) Instead of masking just 1 pixel at every iteration, we mask larger subsets of pixels in each iteration. More formally, let $\mathbf{x}$ be an image of dimensions $H \times W \times C$ where H is the height, W the width, and C the channel. Let $f(\mathbf{x})$ be the network's confidence on image $\mathbf{x}$ and τ the target SIS confidence threshold. Recall that we only compute SIS for images where $f(\mathbf{x}) \geq \tau$. Let M be the mask with dimensions $H \times W$ with 0 indicating an unmasked feature (pixel) and 1 indicating a masked feature. We initialize M as all 0s (all features unmasked). At iteration i, we compute the gradient of f with respect to the input pixels and mask $\nabla M = \nabla_M f(\mathbf{x} \odot (1 - M))$. Here M is the current mask updated after each iteration. In each iteration, we find the block of k features to mask, G^*, chosen in descending order by value of entries in ∇M. The mask is updated after each iteration by masking this block of k features until all features have been masked. Given p input features, our **Batched Gradient SIScollection** procedure returns j sufficient input subsets in $\mathcal{O}(\frac{p}{k} \cdot j)$ evaluations of ∇f (as opposed to $\mathcal{O}(p^2 j)$ evaluations of f in the original SIS procedure [4]).

We use $k = 100$ in this paper, which allows us to find one SIS for each of 32 ImageNet images (i.e., a mini-batch) in $\sim$1-2 minutes using **Batched Gradient FindSIS**. Note that while our algorithm is an approximate procedure, the pixel-subsets produced are real sufficient input subsets, that is they always satisfy $f(\mathbf{x}_S \geq \tau)$. For CIFAR-10 images (which are smaller in size), we use the original SIS procedure from [4]. For both datasets, we treat all channels of each pixel as a single feature.

Batched Gradient SIScollection(f, $\mathbf{x}$, τ, k)

$M = \mathbf{0}$
for $j = 1, 2, \ldots$ **do**
  $R =$ **Batched Gradient BackSelect**$(f, \mathbf{x}, M, k)$
  $M_j =$ **Batched Gradient FindSIS**$(f, \mathbf{x}, \tau, R)$
  $M \leftarrow M + M_j$
  if $f(\mathbf{x} \odot (1 - M)) < \tau$: **return** M_1,...,M_{j-1}
end

Batched Gradient BackSelect(f, $\mathbf{x}$, M, k)

$R =$ empty stack
while $M \neq \mathbf{1}$ **do**
  $G^* = \text{Top}_k\left(\nabla_M f(\mathbf{x} \odot (1 - M))\right)$
  Update $M \leftarrow M + G^*$
  Push G^* onto top of R
end
return R

Batched Gradient FindSIS(f, $\mathbf{x}$, τ, R)

$M = \mathbf{1}$
while $f(\mathbf{x} \odot (1 - M)) < \tau$ **do**
    Pop G from top of R
    Update $M \leftarrow M - G$
end
if $f(\mathbf{x} \odot (1 - M)) \geq \tau$: **return** M
else: **return** *None*

Additional Examples of SIS on ImageNet

Figure S6 shows additional examples of SIS (threshold $= 0.9$) on ImageNet images (see Section 4.1.1).

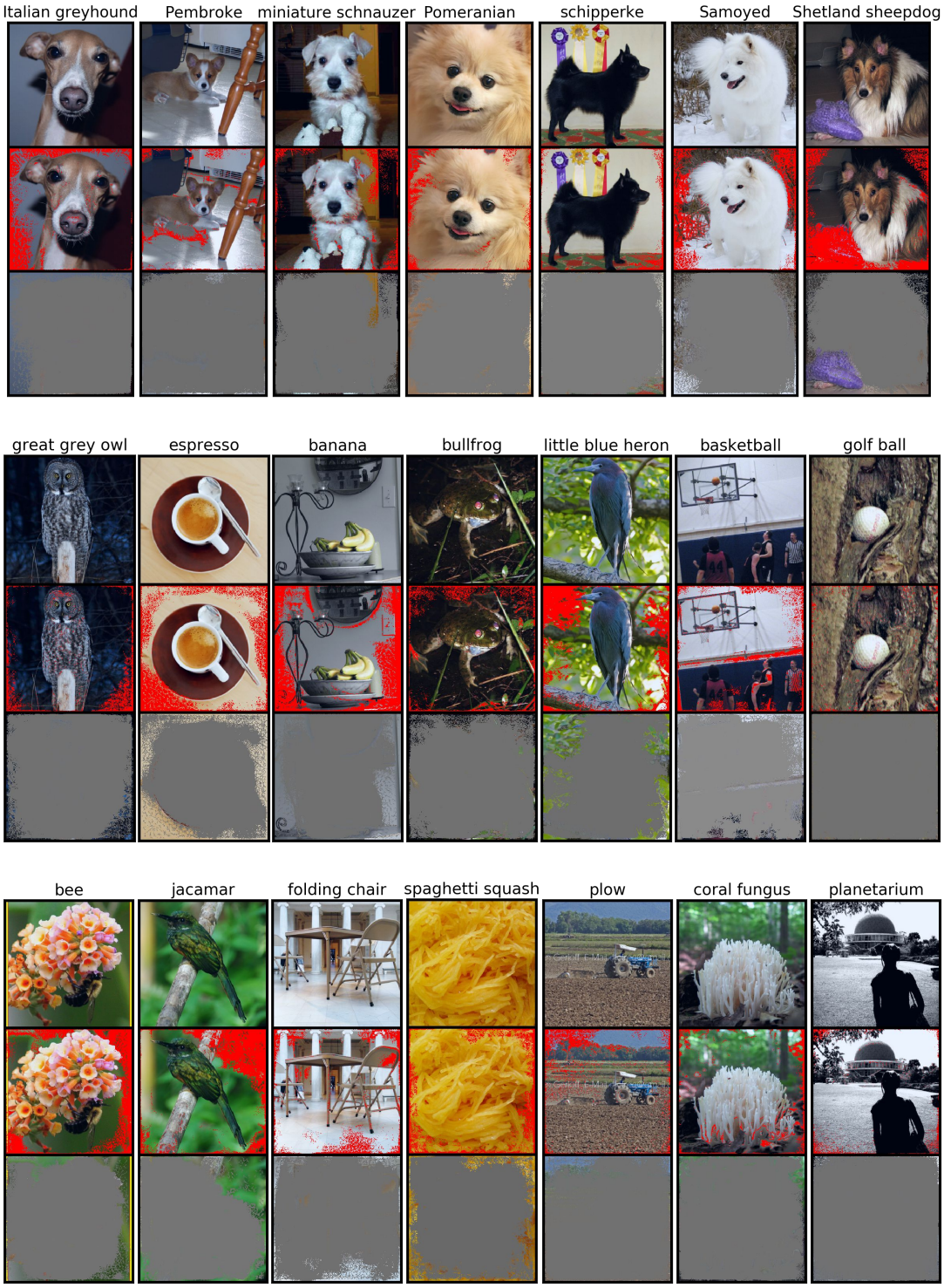


Figure S6: Examples of SIS (threshold $= 0.9$) from the ImageNet validation set (top row of each block). The middle rows show the location of the SIS pixels (red) and the bottom rows show the image with all pixels outside of the SIS masked, which is still classified by the Inception-v3 model with $\geq 90\%$ confidence.