Towards Global Explanations of Convolutional Neural Networks with Concept Attribution

Weibin Wu[1], Yuxin Su[1*], Xixian Chen[2], Shenglin Zhao[2], Irwin King[1], Michael R. Lyu[1], Yu-Wing Tai[2]
[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, [2]Tencent
{wbwu, yxsu, king, lyu}@cse.cuhk.edu.hk, {xixianchen, henryslzhao, yuwingtai}@tencent.com

Abstract

With the growing prevalence of convolutional neural networks (CNNs), there is an urgent demand to explain their behaviors. Global explanations contribute to understanding model predictions on a whole category of samples, and thus have attracted increasing interest recently. However, existing methods overwhelmingly conduct separate input attribution or rely on local approximations of models, making them fail to offer faithful global explanations of CNNs. To overcome such drawbacks, we propose a novel two-stage framework, Attacking for Interpretability (AfI), which explains model decisions in terms of the importance of user-defined concepts. AfI first conducts a feature occlusion analysis, which resembles a process of attacking models to derive the category-wide importance of different features. We then map the feature importance to concept importance through ad-hoc semantic tasks. Experimental results confirm the effectiveness of AfI and its superiority in providing more accurate estimations of concept importance than existing proposals.

1. Introduction

Convolutional neural networks (CNNs) have emerged as a cutting-edge solution to a broad spectrum of real-world applications, such as object recognition [21], audio processing [17], and natural language analysis [49]. Despite the startling advance of these powerful computational architectures, their inner workings remain a mystery. Interpreting and understanding the behaviors of CNNs have become an increasingly crucial topic of research. It can not only justify decisions of CNNs to promote model trustworthiness, but also spot their latent defects to inspire the development of better models [12, 16, 9, 45].

Among diverse explanation techniques, attribution endeavors to succinctly summarize how CNNs arrive at their final decisions [28, 10]. Under the context of image classification, the convention is to measure the importance of human-understandable units to model predictions, such as pixels (*i.e.,* input attribution) and concepts (*i.e.,* concept attribution) [19]. Concept attribution can overcome the ambiguity of input attribution and thus have attracted growing attention recently [19, 10, 52].

There are two explanation interfaces of concept attribution studied in the literature: local explanations [52] and global ones [19], and we focus on the latter in this work, which is imperative but under-explored. Local explanations investigate the rationale of model predictions on individual data points, which are helpful when we only care about a specific instance. In contrast, global explanations center on mining generic decision modes that apply to an entire class of examples. For instance, global explanations can answer to what extent the banded texture is related to a zebra class in model cognition. Therefore, such global explanations are conducive to summarize the model knowledge succinctly and understand the model as a whole [19].

In general, existing concept attribution methods implicitly follow a two-stage procedure [19, 10, 52]. First, since the model decisions are built upon a cornucopia of feature detectors, they conduct *feature attribution* to quantify the importance of individual feature detectors to model predictions[1]. In this step, current attempts simply employ back-propagated gradients as the estimation of feature importance. Second, they achieve *concept attribution* by translating feature importance into concept importance. Most turn to first settle the embedding of a concept in the model feature space (*i.e.,* the concept vector), and then measure the alignment between this concept vector and the vector of feature importance. As for works that focus on global explanations, they just analyze individual predictions in isolation with the above procedure and then return summary statistics [19, 10].

It is doubtful whether such a strategy to obtain global explanations indeed sees "globally". The deficiency primar-

*Corresponding author.

[1]To avoid confusion, we consistently use the term "feature" to refer to the visual patterns detected by feature filters of CNNs (*e.g.,* the banded texture), rather than the input pixels.

2575-7075/20/$31.00 ©2020 IEEE
DOI 10.1109/CVPR42600.2020.00868

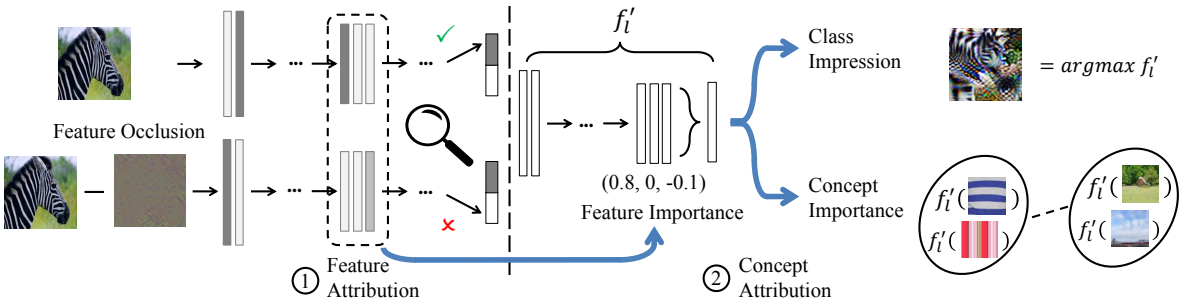


Figure 1: The workflow of our framework: Attacking for Interpretability (AfI).

ily originates from the process of feature attribution with backpropagated gradients, which implicitly builds upon a local linear approximation of CNNs. Unfortunately, such an approximation holds merely when we deal with the proximity of individual instances or the last linear layer of CNNs. Worse still, inspecting individual predictions separately with respective gradients ignores the connections among examples of the same class, and may not be able to capture the generic properties of the class embedded in model knowledge.

To surmount the pitfalls of existing proposals, we propose a novel concept attribution framework for global explanations of CNNs. It explicitly builds upon the two-stage prototype of prior efforts. As such, we systematize the process to model explanations in that we make each step grounded and propose to evaluate intermediate results. More crucially, we thoughtfully extend the methodology of input occlusion to feature occlusion, which enables learning a global explanation and delving into model internals for layer-wise inspections (Section 4.4).

Figure 1 outlines the workflow of our framework: Attacking for Interpretability (AfI). In the first stage, we conduct feature attribution through a thoughtful feature occlusion analysis. Based on an opposite view of attribution, and the fact that feature detectors in CNNs can be depressed by structured patterns [33], we proceed by learning such a feature occluder in the input space for an entire category of images. The feature occluder is applied to undermine critical feature filters of the class so that models will deviate from their original predictions. Such a feature occlusion procedure coincides with that of attacking CNNs to fool their decisions (attacking). We then record the resultant activation alterations of feature detectors, and score the importance of different features accordingly.

In the second stage, we accomplish concept attribution via directly anchoring feature importance to concept importance (interpretability). We first directly combine feature detectors as per their importance scores to obtain a class-specific meta-detector, and then run semantic tests for a concept of interest. As such, higher performance of the meta-detector in the semantic test implies greater importance of the investigated concept to the class.

In summary, the main contributions of this work are:

- We propose a novel concept attribution framework for global explanations of CNNs. Our framework explicitly builds upon a two-stage procedure and employs a novel feature occlusion methodology to learn a global interpretation. As such, we systematize the process to model explanations.
- We overcome the deficiencies of most existing global explanation techniques that bank on a local approximation of CNNs. Experimental results validate the effectiveness of our approach and showcase its superiority to previous efforts.
- With the global explanations our framework affords, we demonstrate its use cases in providing insights into CNNs, like grounding model decisions and revealing biases in model cognition.

2. Related Work

2.1. Attribution

Under the context of image classification, attribution aims to quantify the importance of human-readable units to model decisions [28, 10]. Based on the unit to which it attributes model predictions, there are two attribution techniques: input and concept attribution.

Input attribution explains model behaviors in terms of the importance of different input pixels. The outcome of input attribution, coined saliency maps, can highlight the most responsible parts of input images for model decisions. There is a vast body of work under this track, such as the gradient-based [36, 39, 35, 41, 38, 34, 28], structure-based [1, 24, 46, 51], proxy model-based [30, 23], and decision-based approach [8, 5, 53, 31, 6, 48, 46].

Unfortunately, despite being intuitive, input attribution also suffers from confining itself to input space. The primary culprit is that the semantic meanings of image pixels

are highly dependent on others and diverse. Consequently, the saliency maps returned by input attribution are subject to human perceptions before they become a human-readable interpretation. Unfortunately, human judgments are error-prone and can lead to contradicting conclusions [19].

Concept attribution attempts to address this issue by directly measuring the importance of human-understandable concepts to model decisions. It affords two interpretation interfaces: local explanations that work for individual predictions [52, 28] and global ones that apply for a whole category of examples [19, 10].

Both lines of concept attribution overwhelmingly follow an implicit two-stage procedure. They first conduct feature attribution to derive feature importance, and then translate it to concept importance to accomplish concept attribution. In the feature attribution step, prior schemes coincidentally employ backpropagated gradients as the estimation of the importance of individual features to a class (the feature importance vector). In the concept attribution phase, they usually exploit concept classification to derive the embedding of a concept in hidden layers of CNNs (the concept vector). Such a concept vector denotes the combinations of feature filters that can best detect the concept. They then project the feature importance vector along the direction of the concept vector to gauge the importance of the corresponding concept to model decisions [52, 19, 10]. For a global explanation, they simply run the above routine for individual samples in isolation and report the average concept importance [19, 10].

Our concept attribution framework defeats the pitfalls of both local approximations of models and separate investigations of samples in existing global explanation approaches. During feature attribution, we devise a novel feature occlusion analysis. It abandons local model approximations, and learns a global interpretation that considers the extensive connections among samples of the same class in model cognition. Motivated by the prior art [26, 19, 7, 3, 52], our concept attribution scheme directly combines feature filters as per their importance and estimates their representation capacity of a concept of interest to measure concept importance. Consequently, compared with current attempts, our concept attribution procedure is more general, which also offers the opportunity to integrate prior model visualization techniques [26, 50] into concept attribution.

Like us, a few efforts also aim to overcome the above shortcomings in existing global explanation methods [47, 13]. However, they possess less general applicability than us. [13] proposes to perform a direct concept occlusion analysis, whereas they assume access to the generation process of natural images for given concepts. [47] counts on an inherently more interpretable model, where each feature filter independently and exclusively responds to one concept. In contrast, our technology widely applies to post-training CNN image classifiers, without the need for the data generation mechanism or model modification.

2.2. Adversarial Susceptibility

The functional units of CNNs are surprisingly sensitive to adversarial patterns, namely, the so-called adversarial perturbations. [43] first uncovers that despite imperceptible to humans, they can deviate CNNs from correct decisions when attached to clean images. [33] further reveals that such purposeful distortions can mislead the feature filters of CNNs. Therefore, they can manipulate the hidden representations of legitimate images. Successor studies, such as [25, 4], discover that adversarial noise can be extremely effective and universal for image groups. Based on these findings, we propose a novel feature attribution scheme, where we conduct feature occlusion from the image space, and employ it to learn a global interpretation.

3. Method

In this section, we will detail the design of our framework. As illustrated in Figure 1, our two-stage approach proceeds by tackling the following tasks sequentially: (a) how to learn a feature occluder to perform feature occlusion (Section 3.1.1), (b) how to complete feature attribution with feature occluders (Section 3.1.2), and (c) how to achieve concept attribution via aligning feature importance with concept importance (Section 3.2).

We first set up some notations. We regard an input image as a vector $\mathbf{x} \in \mathbb{R}^n$ with label prediction $y \in \mathbb{Y}$, where $\mathbb{Y} := \{1, ..., K\}$ is a categorical set of interest. By convention, images will be normalized such that $\mathbf{x}$ stays within the range of $[-1, 1]^n$ with zero mean before feeding into models. In a CNN classifier with L layers, the l^{th} layer with m neurons learns a mapping from inputs to hidden representations $f_l : \mathbb{R}^n \to \mathbb{R}^m$. In particular, the final layer computes a logit vector $Z(\mathbf{x}) \in \mathbb{R}^K$ and then yields a probability vector $f_L(\mathbf{x})$ after softmax normalization. The y^{th} entry $f_L(\mathbf{x})[y]$ corresponds to the probability of $\mathbf{x}$ belonging to class y. A CNN classifier will output label predictions in the end, and thus its decision function is $f : \mathbb{R}^n \to \mathbb{Y}$.

3.1. Feature Attribution

For feature attribution, we propose to extend the methodology of input occlusion to feature occlusion. The general procedure of input occlusion is to occlude some input pixels and regard the resultant alterations of model output as their importance score [46]. Unfortunately, a straightforward adaptation scarcely applies to feature occlusion. In modern CNN architectures, there are innumerous neurons that work in close collaboration [7]. Therefore, separately occluding individual neurons ignores their intensive interconnections, while exhausting all possible combinations is prohibitively expensive.

We circumvent this difficulty via an opposite view of attribution via occlusion. Given an image $\mathbf{x}$ and its prediction y, the fundamental problem in attribution is to explain how a model discriminates class y from all the others. Furthermore, in the form of feature attribution, we can summarize the reasoning process of a model in this binary classification task as:

$$\textit{the features of class } y \textit{ in image } \mathbf{x} \textit{ are more prominent} \iff \textit{the label prediction for image } \mathbf{x} \textit{ is } y. \quad (1)$$

Consequently, it reduces to spot supporting features for model decisions. To this end, we first transform the forward reasoning of (1) into its logic equivalence:

$$\textit{the label prediction for image } \mathbf{x} \textit{ is } \textbf{\textit{not}}\ y \longrightarrow \textit{the features of class } y \textit{ in image } \mathbf{x} \textit{ are } \textbf{\textit{less}} \textit{ prominent}. \quad (2)$$

Then by combining with the backward reasoning of (1):

$$\textit{the label prediction for image } \mathbf{x} \textbf{\textit{ is}}\ y \longrightarrow \textit{the features of class } y \textit{ in image } \mathbf{x} \textit{ are } \textbf{\textit{more}} \textit{ prominent}, \quad (3)$$

it leads us to an opposite procedure for attribution with occlusion. Specifically, we can conservatively undermine feature filters of a model until it is forced to abandon its original decisions. As such, the resultant variations of neuron activations represent their importance to model predictions.

Moreover, since feature filters of CNNs are susceptible to structured noise [33], such an opposite view empowers us to perform feature occlusion from the input space. Specifically, we can first learn such a malicious perturbation to "subtract" minimal image features, which suffice to flip model predictions. We coin such perturbations *feature occluders*, which effectively work by disturbing responsible feature detectors [33, 2]. Therefore, it means that feature occluders need not destroy images in a human-recognizable manner, or align with actual regions where filters extract features. Then we examine the change of neuron outputs to rate their importance.

3.1.1 Global Feature Occluder

As we seek a global explanation of samples under the same category, we start by crafting a *global feature occluder* for them. Formally, given image collection $\{\mathbf{x}_i : i = 1, \ldots, N\}$ with identical classification y, we define their global feature occluder δ^* as:

$$\begin{aligned} \delta^* = \ & \text{argmin } D(\delta) \\ \text{such that } \ & f(\mathbf{x}_i - \delta) \neq y \\ & f(t(\mathbf{x}_i - \delta)) \neq y \qquad i = 1, \ \ldots, N \\ & f(t(\mathbf{x}_i)) = f(\mathbf{x}_i) = y \\ & \mathbf{x}_i - \delta \in [-1, 1]^n. \end{aligned} \quad (4)$$

We elucidate the definition as follows.

In the object function of (4), distance function D measures the magnitude of δ. As such, we aim to search for minimal perturbations, which reflects the appeal of disturbing minimal feature filters so that we can identify the most critical features of the class. In light of the sliding-window scheme in CNNs [11], we implement D via l_1 distance.

The first condition of (4) further requires that a global feature occluder is the minimal noise needed to flip the model predictions on all the given instances simultaneously. Therefore, it will prefer to impede decisive feature detectors common to images of the same class, which takes into account the relations among samples embedded in model memory. Therefore, our approach conducts a sort of reverse engineering of the model training process, which is conducive to expose a more global picture of model logic.

The second condition of (4) conducts regularization, where t denotes image transformations, like random noising. We suppose that purely learning deceptive distortion may end up spoiling some fragile filters less relevant to essential image features. To eliminate such artifacts, we additionally require that a global feature occluder should remain effective when applied to the transformed versions of original images. We expect that the outputs of supporting feature filters can maintain relatively unchanged compared to the others when inputting transformed images. Consequently, such a requirement can make feature occluders focus on dimming critical features rather than arrive at the cheapest structure.

Besides, to constitute an effective regularization, we ensure that t will not harm the judgment of the model on clean images (the third condition). The last condition of (4) guarantees that occluded images are still valid inputs for models.

As CNNs are involved, directly solving (4) is intractable. We instead obtain an approximation by employing Adam optimizer [20] to minimize the following object function iteratively:

$$\frac{1}{N}\Sigma_{i=1}^{N}(Z(\mathbf{x}_i - \delta)[y] + Z(t(\mathbf{x}_i - \delta))[y]) + \lambda \cdot D(\delta). \quad (5)$$

Our algorithm terminates once the occluder satisfies all the constraints in (4), or when we exceed preset maximum iterations.

3.1.2 Feature Importance Score

Now we can calculate feature importance scores with the obtained global feature occluder for class y. Specifically, the importance score of the feature that the j^{th} neuron in the l^{th} layer detects is:

$$s_l^j = \frac{1}{N}\Sigma_{i=1}^{N}(f_l(\mathbf{x}_i)[j] - f_l(\mathbf{x}_i - \delta^*)[j]). \quad (6)$$

The sign of the importance score differentiates two sorts of features related to model decisions. Neurons with positive scores account for supporting features, while the ones with negative scores vote for antagonistic counterparts [50, 34]. Similar to conventional practice, we focus on features and concepts that have positive contributions to model decisions [34, 52]. Therefore, we zero out negative importance scores in s_l^j to obtain the final feature importance score (FIS) we adopt:

$$s_l^{'j} = \max(s_l^j, 0). \tag{7}$$

3.2. Concept Attribution

This step communicates feature importance in terms of the importance of semantic notions readily accessible to humans. Some prior proposals first examine CNN units separately to work out their concept labels. They then read the importance scores of these concepts from the feature importance scores of corresponding units [28, 47]. However, such strategies overlook concepts with entangled encodings in CNNs [3, 7]. To surmount this defect, we propose a two-step procedure. We first combine CNN units as per their importance scores, which leads to a class-specific meta-detector. Then we estimate the representation capacity of the meta-detector for a concept of interest through carefully designed semantic tasks, where higher representation power signifies greater importance of the concept to the investigated class.

Specifically, in the first step, to acquire a class-specific meta-detector, we also regard feature maps as basis CNN units like the prior art [7, 3]. We denote the c^{th} feature map in layer l as A_l^c. Therefore, for class y, we normalize the total importance scores of neurons within A_l^c as its channel importance score (CIS):

$$w_l^c = \frac{1}{B}\Sigma_{j \in P_l^c} s_l^{'j}. \tag{8}$$

Here P_l^c is the index set of neurons in A_l^c, and B is a normalizing constant such that $w_l^c \in [0, 1]$. We view the fully connected layers with C neurons as C feature maps with a spatial resolution of 1×1. Subsequently, we combine feature maps in layer l with CIS to get the meta-detector:

$$f_l^{'} = \Sigma_c w_l^c \cdot A_l^c. \tag{9}$$

It encodes the relevance of various concepts to class y in model cognition.

In the second step, inspired by the work [26, 19, 7, 3, 52], we propose two kinds of semantic tasks to evaluate the representation power of the meta-detector. They are tailored for qualitative and quantitative concept attribution, respectively.

For qualitative concept attribution, we devise a generation task. Specifically, we adapt the technology of model visualization [26] to synthesize images, which can maximize the total activation of the meta-detector for class y. The crafted image corresponds to a class impression. It qualitatively depicts the most distinct characteristics of the class concept y in the memory of the model.

For quantitative concept attribution, we reify it as a concept classification task, where we gauge the capability of the meta-detector to distinguish different concepts, and rank the importance of these concepts accordingly. Specifically, we resort to probe datasets with concept labels as in [19, 10]. For each probe image, we first obtain the outputs from the meta-detector as its new representation. Then for a concept of interest, we compute the discrepancy of its samples to the benchmark ones with irrelevant concept tags. The discrepancy quantifies the discriminative power of the meta-detector regarding this concept. We adopt the Maximum Mean Discrepancy (MMD) as the discrepancy metric [14]. Therefore, we sum the MMD values calculated in all the middle layers, and view the normalized results as the importance score of the corresponding concept.

4. Experiments

We first report the intermediate attacking results in Section 4.1. Then we evaluate our feature and concept attribution results in Section 4.2 and Section 4.3, respectively. Finally, we present some qualitative and quantitative explanations we obtain in Section 4.4 and Section 4.5, respectively, which showcase the use cases of our framework.

We demonstrate the effectiveness of our framework with three CNNs trained for ImageNet (ILSVRC2012) classification: ResNet-50, GoogLeNet, and VGG-16 [15, 42, 37, 32]. These models cover representative sorts of models for image classification and have wide application in practice [29]. Therefore, such a model choice can confirm the general applicability of our approach. Besides, we focus on the ImageNet dataset since it is a widely recognized dataset for evaluating explanation techniques [36, 10] and diverse pre-trained models for ImageNet classification are publicly available. Accordingly, such a dataset choice facilitates fair comparisons with the existing efforts [19, 10].

We adopt the training set of ImageNet to learn global feature occluders so that we can work on the same page as models. Parameters are settled experimentally. The transformation function t is a composition of: (1) applying uniform random noise within $[-0.04, 0.04]^n$ and (2) random rotation within $[-5°, 5°]$. λ is set to balance the contribution of each term in (5).

4.1. Attacking Results

As experimental demonstrations, we first randomly select 100 classes from all the 1000 classes in the ImageNet dataset [32], and fix these classes for our experiments. We then learn one global feature occluder for each class. To examine the attack success rates of the resultant global feature

Model	Clean	Perturbed
ResNet-50	0.8771	0.0973
GoogLeNet	0.8115	0.0907
VGG-16	0.8095	0.1001

Table 1: Average top-1 accuracy of different models on clean images and the counterparts perturbed with corresponding global feature occluders.

Teacher Model	Gradient-based	AfI (Without t)	AfI
ResNet-50	0.8899	0.8918	**0.9592**
GoogLeNet	0.8383	0.8896	**0.9826**
VGG-16	0.8531	0.8679	**0.9468**

Table 2: The average accuracy of student models derived from different approaches.

occluders, we perturb the images with the corresponding global feature occluders and calculate the average top-1 accuracy of the model on these samples. Table 1 reports the results. We can see that our global feature occluders can severely undermine the model performance on perturbed images. Therefore, it is feasible to learn global feature occluders with our approach.

Besides, based on our preliminary experiments, we note that we can obtain fairly accurate global attribution results as long as the attack success rates are high enough (not necessarily 100%). It may be because that global concept attribution should spot concepts that are frequently important for a class in model cognition (*e.g.,* leaves for trees though some trees may not have leaves at present), and have to pay less attention to unrepresentative samples. On the other hand, if occluders fail to achieve high success rates, the performance of our global explanation approach will degenerate. Consequently, we mitigate it by class-specific fine-tuning in our experiments.

4.2. Evaluation of the Feature Attribution Results

To examine our feature attribution results—feature importance scores, we propose a distillation test similar to that in [44, 22]. We regard a model we aim to explain as a teacher model. If for class y, the teacher model owns outstanding accuracy, and our feature importance scores are correct, the derived meta-detector should also possess high discriminative competence for the class concept y. In other words, given the activation of the meta-detector as inputs, a compact student model can differentiate class y from the others. Higher performance of the student model indicates that the feature attribution results are more precise.

Therefore, we implement the distillation test as binary classification tasks in ImageNet. For each class, we first randomly sample a balanced dataset, which consists of the same number of instances from the class and complement ones. We also make sure that the teacher model can correctly recognize all the included images. Then for each sample, we compute the outputs from the meta-detectors of the teacher model, which are flattened as the representation of the image. Finally, we train student models to conduct binary classification on the resultant data as per the original training-validation partition of ImageNet.

For comparison, we also conduct the same distillation test based on the feature attribution results from the state-of-the-art baseline - TCAV [19, 10]. Specifically, TCAV proposes to perform feature attribution for individual samples with backpropagated gradients. Since TCAV does not acquire a global feature importance score (FIS) for a class, we average its feature attribution results over the whole class of examples as the FIS to test.

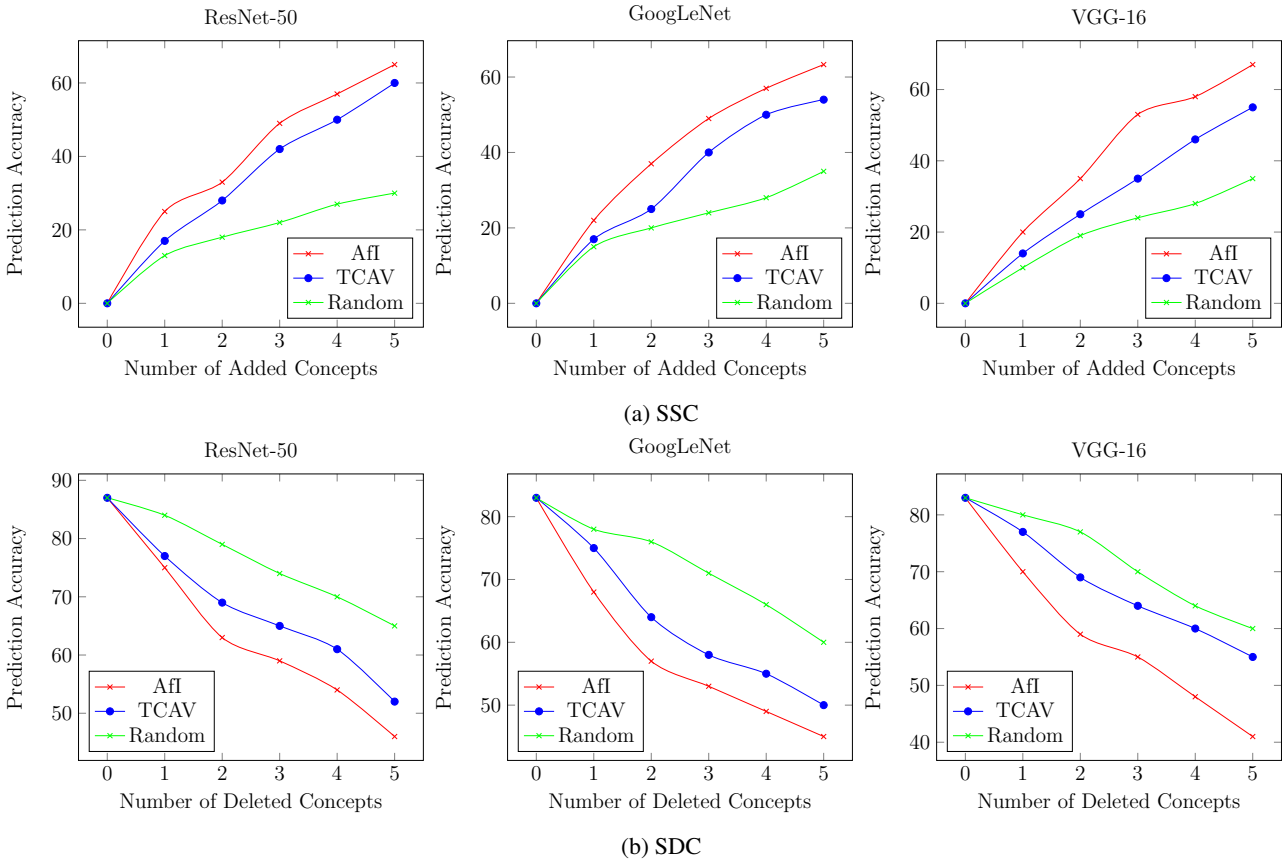
Table 2 reports the average accuracy of student models over 100 classes. All the student models we exploit are neural networks with three fully connected layers, where there are 32, 16, and 2 neurons, respectively. Student models derived from our method (AfI) can obtain remarkable accuracy, exceeding the gradient-based baseline (TCAV) by a significant margin. It validates the effectiveness of our feature attribution mechanism and its superiority to the state-of-the-art benchmark. Besides, we run an ablation study to verify the contribution of the transformation function t, where we remove it from (5) when learning feature occluders. The performance degradation of the resultant student models confirms the regularization efficacy of t.

Moreover, under our method, student models of GoogLeNet manifest the best performance compared to the other teacher models. Since we obtain student models via global explanations of model decisions, it may indicate that GoogLeNet relies on more consistent combinations of features to identify samples from the same class, and thus adopts more category-generic decision modes than the other models.

4.3. Evaluation of the Concept Attribution Results

We follow [10] to evaluate our concept attribution results—concept importance scores, since [10] can conduct extensive quantitative assessments with high efficiency. Specifically, [10] regards semantic image segments as concept data. It leads to two metrics: the smallest sufficient concepts (SSCs) and the smallest destroying concepts (SDCs). SSCs are the smallest set of concepts sufficing for models to predict the target class, while SDCs are the smallest concept collections whose absence will incur wrong predictions. More accurate concept importance scores can lead to a more precise estimation of SSCs and SDCs.

Therefore, given a class, we first segment images of the class and cluster similar segments. Each cluster represents



(a) SSC

(b) SDC

Figure 2: Model accuracy variation when we start editing the most important SSCs/SDCs estimated by different approaches. For our method (AfI), the top-5 SSCs are enough to recover over 74% of the original accuracy across all models, while removing the top-5 SDCs can result in a degradation of over 45% of the original accuracy across all models. We also plot the effect of editing concepts in random order for comparison. The concept importance scores derived by our method (AfI) are consistently more accurate than the benchmark (TCAV), since the change of model accuracy is more drastic for our approach.

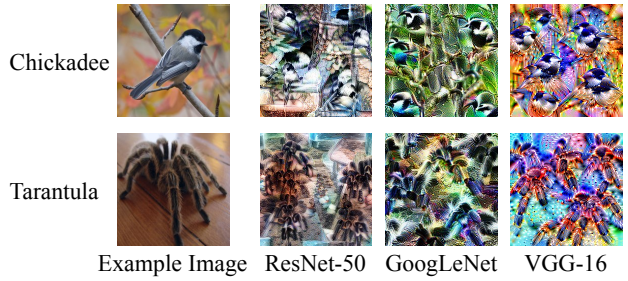


Figure 3: Class concepts captured by different models. Example images of the corresponding class are exhibited for better comparison.

examples for one concept. With these concept data, we then calculate the importance score of each concept, and curate the most important concepts as SSCs and SDCs. Finally, we sequentially add SSCs to a blank image or remove SDCs from the source image as per their importance order. We record the change of model accuracy to examine the concept importance scores we derive. We also test the state-of-the-art baseline (TCAV) under the same setup for comparison [10, 19].

Figure 2 exhibits the average result over 100 classes. It confirms that our estimation of SSCs and SDCs is remarkably more accurate than TCAV, as the change of model accuracy during concept adding/removing is more drastic. Therefore, our estimated concept importance scores are more precise than the state-of-the-art benchmark.

4.4. Class Concept Visualization

With our qualitative concept attribution strategy, we visualize class concepts captured by a model. Specifically, for a random class, we first separately generate images that can highly activate the meta-detector in each middle layer.

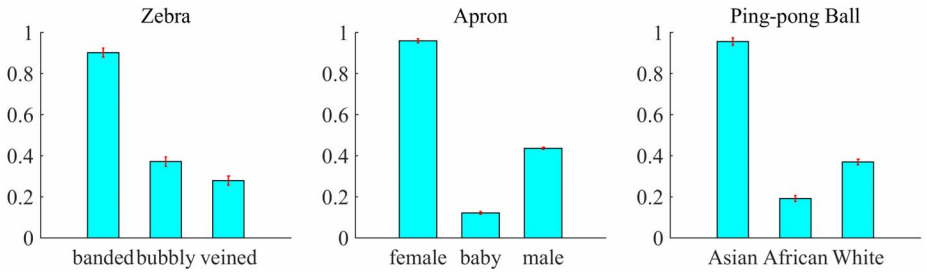


Figure 4: Importance scores of different concepts to classification results. Error bars indicate the standard deviation.

Model	Layer Name	Original Output Shape
ResNet-50	ResBlock_4c	7 x 7 x 2048
GoogLeNet	Mixed_5b	7 x 7 x 832
VGG-16	Fc_6	1 x 1 x 4096

Table 3: The layer selected to craft class impressions and its original output shape (spatial resolution $\times$ channel number).

We then spot the first layer where class concepts emerge through visual investigation. The visualization of class concepts from this layer is regarded as class impressions. During the generation of class impressions, except for the total variation penalization, we do not resort to any other natural image priors, such as a generative network [27]. Accordingly, it ensures that the class impressions are only born of the knowledge of the model under inspection.

Figure 3 displays some class impressions we obtain, along with example images of the corresponding classes for better comparison. It illustrates that CNNs can capture the most prominent characteristics of image classes, for example, the texture for the tarantula class. Additionally, ResNet-50 appears to better capture and exploit the color property of images than the other models, because the class impressions of ResNet-50 are more similar to raw images of the corresponding classes in terms of their color.

Table 3 reports the layer we choose to craft class impressions for each model. We note that in the middle layers, it is non-trivial to infer the links of copious neurons to image categories. Because unlike the last logit layer, their mappings are not specified during training. Consequently, the competence to uncover class concept embeddings in the middle layers of CNNs further verifies the effectiveness of our framework.

4.5. User-defined Concept Attribution

With our quantitative concept attribution scheme, we measure the importance of user-defined concepts to classification. We center on explaining widely-used ResNet-50, which has been less covered in the literature. As experimental examples, we gauge the importance of concepts from three representative groups (*i.e.,* texture, gender, and race) to three classes, respectively. We follow [19] to curate probe concept data [32, 3, 18]. Concretely, for each pair of the concept type and image class, we first randomly select the same number of images as the concept data for each concept. Then we fix a random benchmark set of the same size. We finally compute concept importance scores with the probe data.

Figure 4 reports the average result over 100 runs. It validates that CNNs can extract rational grounds for their decisions, like the banded texture for the zebra. However, consistent with the findings of [40], we discover that they also sometimes learn undesirable stereotypes about some classes, such as the relatively stronger positive connections of women to the apron and Asians to the ping-pong ball. Therefore, it demonstrates the use case of our framework in model confirmation and bias revelation.

5. Conclusion

We propose a novel two-step framework for global explanations of CNNs. It first derives feature importance via a novel feature occlusion analysis, and then communicates such information in terms of the importance of human-comprehensible concepts. Empirical results corroborate the effectiveness and superiority of our technique in explaining model behaviors. More crucially, we demonstrate that we can achieve concept attribution via two semantic tasks. It showcases the exciting opportunity to integrate prior feature visualization efforts into our framework, which is a promising direction for future work.

Acknowledgment

We thank anonymous reviewers for their valuable comments. The work described in this paper was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (CUHK 14210717 of the General Research Fund and CUHK 2300174 of the Collaborative Research Fund, No. C5026-18GF).

References

[1] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015. 2

[2] Randall Balestriero et al. A spline theory of deep networks. In *International Conference on Machine Learning (ICML)*, pages 383–392, 2018. 4

[3] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network Dissection: Quantifying interpretability of deep visual representations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6541–6549, 2017. 3, 5, 8

[4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017. 3

[5] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *Advances in Neural Information Processing Systems (NIPS)*, pages 6967–6976, 2017. 2

[6] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems (NIPS)*, pages 592–603, 2018. 2

[7] Ruth Fong and Andrea Vedaldi. Net2Vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8730–8738, 2018. 3, 5

[8] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *International Conference on Computer Vision (ICCV)*, pages 3449–3457. IEEE, 2017. 2

[9] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019. 1

[10] Amirata Gohorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems (NIPS)*, 2019. 1, 2, 3, 5, 6, 7

[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org. 4

[12] Bryce Goodman and Seth Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". *ICML Workshop on Human Interpretability in Machine Learning*, 2016. 1

[13] Yash Goyal, Uri Shalit, and Been Kim. Explaining classifiers with Causal Concept Effect (CaCE). *arXiv preprint arXiv:1907.07165*, 2019. 3

[14] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012. 5

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[16] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *The European Conference on Computer Vision (ECCV)*, pages 793–811. Springer, 2018. 1

[17] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. CNN architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135. IEEE, 2017. 1

[18] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 8

[19] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *International Conference on Machine Learning (ICML)*, pages 2673–2682. PMLR, 2018. 1, 3, 5, 6, 7, 8

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 4

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105. Curran Associates, Inc., 2012. 1

[22] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *International Conference on Computer Vision (ICCV)*, pages 365–372. IEEE, 2009. 6

[23] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4765–4774, 2017. 2

[24] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017. 2

[25] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1765–1773, 2017. 3

[26] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks. *Google Research Blog*, 2015. Retrieved: October 2018. 3, 5

[27] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for

neurons in neural networks via deep generator networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3387–3395. Curran Associates, Inc., 2016. 8

[28] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 2018. https://distill.pub/2018/building-blocks. 1, 2, 3, 5

[29] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, 29(9):2352–2449, 2017. 5

[30] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1135–1144. ACM, 2016. 2

[31] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*, 2018. 2

[32] 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. 5, 8

[33] Sara Sabour, Yanshuai Cao, Fartash Faghri, and David J Fleet. Adversarial manipulation of deep representations. In *International Conference on Learning Representations (ICLR)*, 2016. 2, 3, 4

[34] 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, 5

[35] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning (ICML)*, pages 3145–3153. PMLR, 2017. 2

[36] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations (ICLR)*, 2014. 2, 5

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 5

[38] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: Removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 2

[39] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR Workshop*, 2015. 2

[40] Pierre Stock and Moustapha Cisse. ConvNets and ImageNet beyond accuracy: Understanding mistakes and uncovering biases. In *The European Conference on Computer Vision (ECCV)*, pages 498–512, 2018. 8

[41] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*, pages 3319–3328. PMLR, 2017. 2

[42] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. 5

[43] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. 3

[44] Guanhong Tao, Shiqing Ma, Yingqi Liu, and Xiangyu Zhang. Attacks meet interpretability: Attribute-steered detection of adversarial samples. In *Advances in Neural Information Processing Systems (NIPS)*, pages 7717–7728, 2018. 6

[45] Weibin Wu, Hui Xu, Sanqiang Zhong, Michael R. Lyu, and Irwin King. Deep Validation: Toward detecting real-world corner cases for deep neural networks. In *International Conference on Dependable Systems and Networks (DSN)*, pages 125–137. IEEE, 2019. 1

[46] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *The European Conference on Computer Vision (ECCV)*, pages 818–833. Springer, 2014. 2, 3

[47] Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. Interpreting CNNs via decision trees. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6261–6270, 2019. 3, 5

[48] Xin Zhang, Armando Solar-Lezama, and Rishabh Singh. Interpreting neural network judgments via minimal, stable, and symbolic corrections. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4874–4885, 2018. 2

[49] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems (NIPS)*, pages 649–657, 2015. 1

[50] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene CNNs. In *International Conference on Learning Representations (ICLR)*, 2015. 3, 5

[51] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. 2

[52] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *The European Conference on Computer Vision (ECCV)*, pages 119–134, 2018. 1, 3, 5

[53] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. In *International Conference on Learning Representations (ICLR)*, 2017. 2