Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

Ramprasaath R. Selvaraju[1*] Michael Cogswell[1] Abhishek Das[1] Ramakrishna Vedantam[1*]
Devi Parikh[1,2] Dhruv Batra[1,2]
[1]Georgia Institute of Technology [2]Facebook AI Research
{ramprs, cogswell, abhshkdz, vrama, parikh, dbatra}@gatech.edu

Abstract

*We propose a technique for producing 'visual explanations' for decisions from a large class of Convolutional Neural Network (CNN)-based models, making them more transparent and explainable. Our approach – Gradient-weighted Class Activation Mapping (Grad-CAM), uses the gradients of any target concept (say 'dog' in a classification network or a sequence of words in captioning network) flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept. Unlike previous approaches, Grad-CAM is applicable to a wide variety of CNN model-families: (1) CNNs with fully-connected layers (*e.g.* VGG), (2) CNNs used for structured outputs (*e.g.* captioning), (3) CNNs used in tasks with multi-modal inputs (*e.g.* visual question answering) or reinforcement learning, all* without architectural changes or re-training. *We combine Grad-CAM with existing fine-grained visualizations to create a high-resolution class-discriminative visualization, Guided Grad-CAM, and apply it to image classification, image captioning, and visual question answering (VQA) models, including ResNet-based architectures. In the context of image classification models, our visualizations (a) lend insights into failure modes of these models (showing that seemingly unreasonable predictions have reasonable explanations), (b) outperform previous methods on the ILSVRC-15 weakly-supervised localization task, (c) are more faithful to the underlying model, and (d) help achieve model generalization by identifying dataset bias. For image captioning and VQA, our visualizations show even non-attention based models can localize inputs. Finally, we design and conduct human studies to measure if Grad-CAM explanations help users establish appropriate trust in predictions from deep networks and show that Grad-CAM helps untrained users successfully discern a 'stronger' deep network from a 'weaker' one even when both make identical predictions. Our code is available at* https://github.com/ramprs/grad-cam/ *along with a demo on CloudCV [2]*[1] *and video at* youtu.be/COjUB9Izk6E.

*Work done at Virginia Tech.
[1]http://gradcam.cloudcv.org

1. Introduction

Convolutional Neural Networks (CNNs) and other deep networks have enabled unprecedented breakthroughs in a variety of computer vision tasks, from image classification [24, 16] to object detection [15], semantic segmentation [27], image captioning [43, 6, 12, 21], and more recently, visual question answering [3, 14, 32, 36]. While these deep neural networks enable superior performance, their lack of decomposability into *intuitive and understandable* components makes them hard to interpret [26]. Consequently, when today's intelligent systems fail, they often fail spectacularly disgracefully, without warning or explanation, leaving a user staring at an incoherent output, wondering why.

Interpretability and Explainability matters. In order to build trust in intelligent systems and move towards their meaningful integration into our everyday lives, it is clear that we must build 'transparent' models that explain *why they predict what they predict*. Broadly speaking, this transparency and ability to explain is useful at three different stages of Artificial Intelligence (AI) evolution. First, when AI is significantly weaker than humans and not yet reliably 'deployable' (*e.g.* visual question answering [3]), the goal of transparency and explanations is to identify the failure modes [1, 17], thereby helping researchers focus their efforts on the most fruitful research directions. Second, when AI is on par with humans and reliably 'deployable' (*e.g.*, image classification [22] on a set of categories trained on sufficient data), the goal is to establish appropriate trust and confidence in users. Third, when AI is significantly stronger than humans (*e.g.* chess or Go [39]), the goal of explanations is in machine teaching [20] – *i.e.*, a machine teaching a human about how to make better decisions.

There typically exists a trade-off between accuracy and simplicity or interpretability. Classical rule-based or expert systems [18] are highly interpretable but not very accurate (or robust). Decomposable pipelines where each stage is hand-designed are thought to be more interpretable as each individual component assumes a natural intuitive explanation. By using deep models, we sacrifice interpretable

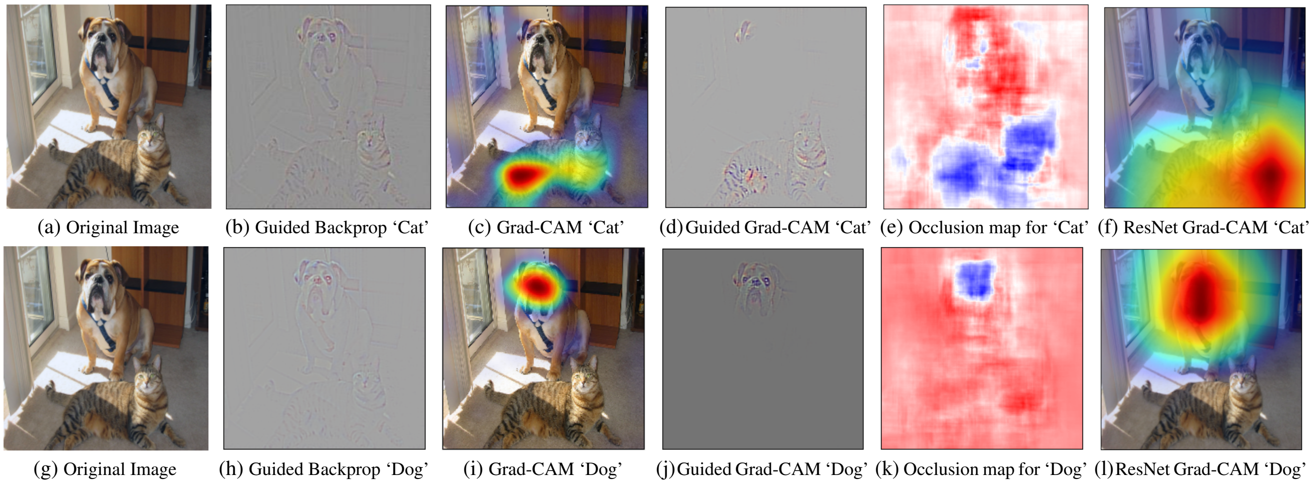


Figure 1: (a) Original image with a cat and a dog. (b-f) Support for the cat category according to various visualizations for VGG-16 and ResNet. (b) Guided Backpropagation [42]: highlights all contributing features. (c, f) Grad-CAM (Ours): localizes class-discriminative regions, (d) Combining (b) and (c) gives Guided Grad-CAM, which gives high-resolution class-discriminative visualizations. Interestingly, the localizations achieved by our Grad-CAM technique, (c) are very similar to results from occlusion sensitivity (e), while being orders of magnitude cheaper to compute. (f, l) are Grad-CAM visualizations for ResNet-18 layer. Note that in (c, f, i, l), red regions corresponds to high score for class, while in (e, k), blue corresponds to evidence for the class. Figure best viewed in color.

modules for uninterpretable ones that achieve greater performance through greater abstraction (more layers) and tighter integration (end-to-end training). Recently introduced deep residual networks (ResNets) [16] are over 200-layers deep and have shown state-of-the-art performance in several challenging tasks. Such complexity makes these models hard to interpret. As such, deep models are beginning to explore the spectrum between interpretability and accuracy.

Zhou *et al*. [47] recently proposed a technique called Class Activation Mapping (CAM) for identifying discriminative regions used by a restricted class of image classification CNNs which do not contain any fully-connected layers. In essence, this work trades off model complexity and performance for more transparency into the working of the model. In contrast, we make existing state-of-the-art deep models interpretable without altering their architecture, thus avoiding the interpretability *vs*. accuracy trade-off. Our approach is a generalization of CAM [47] and is applicable to a significantly broader range of CNN model families: (1) CNNs with fully-connected layers (*e.g*. VGG), (2) CNNs used for structured outputs (*e.g*. captioning), (3) CNNs used in tasks with multi-modal inputs (*e.g*. VQA) or reinforcement learning, without requiring architectural changes or re-training or any secondary learning component.

What makes a good visual explanation? Consider image classification [9] – a 'good' visual explanation from the model for justifying any target category should be (a) class-discriminative (*i.e*. localize the category in the image) and (b) high-resolution (*i.e*. capture fine-grained detail).

Fig. 1 shows outputs from a number of visualizations for the 'tiger cat' class (top) and 'boxer' (dog) class (bottom). Pixel-space gradient visualizations such as Guided Backpropagation [42] and Deconvolution [45] are high-resolution and highlight fine-grained details in the image, but are not class-discriminative (Fig. 1b and Fig. 1h are very similar).

In contrast, localization approaches like CAM or our proposed method Gradient-weighted Class Activation Mapping (Grad-CAM), are highly class-discriminative (the 'cat' explanation exclusively highlights the 'cat' regions but not 'dog' regions in Fig. 1c, and *vice versa* in Fig. 1i).

In order to combine the best of both worlds, we show that it is possible to fuse existing pixel-space gradient visualizations with Grad-CAM to create Guided Grad-CAM visualizations that are both high-resolution and class-discriminative. As a result, important regions of the image which correspond to any decision of interest are visualized in high-resolution detail even if the image contains evidence for multiple possible concepts, as shown in Figures 1d and 1j. When visualized for 'tiger cat', Guided Grad-CAM not only highlights the cat regions, but also highlights the stripes on the cat, which is important for predicting that particular variety of cat.

To summarize, our contributions are as follows:

(1) We propose Grad-CAM, a class-discriminative localization technique that can generate visual explanations from *any* CNN-based network without requiring architectural changes or re-training. We evaluate Grad-CAM for localization (Section 4.1), and faithfulness to model (Section 5.3), where it outperforms baselines.

(2) We apply Grad-CAM to existing top-performing classification, captioning (Section 7.1), and VQA (Section 7.2) models. For image classification, our visualizations help identify dataset bias (Section 6.2) and lend insight into failures of current CNNs (Section 6.1), showing that seemingly unreasonable predictions have reasonable explanations. For captioning and VQA, our visualizations expose the somewhat surprising insight that common CNN + LSTM models are often good at localizing discriminative image regions despite not being trained on grounded image-text pairs.

(3) We visualize ResNets [16] applied to image classification and VQA (Section 7.2). Going from deep to shallow layers, the discriminative ability of Grad-CAM significantly reduces as we encounter layers with different output dimensionality. **(4)** We conduct human studies (Section 5) that show Guided Grad-CAM explanations are class-discriminative and not only help humans establish trust, but also help untrained users successfully discern a 'stronger' network from a 'weaker' one, *even when both make identical predictions.*

2. Related Work

Our work draws on recent work in CNN visualizations, model trust assessment, and weakly-supervised localization.
Visualizing CNNs. A number of previous works [40, 42, 45, 13] have visualized CNN predictions by highlighting 'important' pixels (*i.e*. change in intensities of these pixels have the most impact on the prediction's score). Specifically, Simonyan *et al*. [40] visualize partial derivatives of predicted class scores w.r.t. pixel intensities, while Guided Backpropagation [42] and Deconvolution [45] make modifications to 'raw' gradients that result in qualitative improvements. These approaches are compared in [30]. Despite producing fine-grained visualizations, these methods are not class-discriminative. Visualizations with respect to different classes are nearly identical (see Figures 1b and 1h).

Other visualization methods synthesize images to maximally activate a network unit [40, 11] or invert a latent representation [31, 10]. Although these can be high-resolution and class-discriminative, they visualize a model overall and not predictions for specific input images.
Assessing Model Trust. Motivated by notions of interpretability [26] and assessing trust in models [37], we evaluate Grad-CAM visualizations in a manner similar to [37] via human studies to show that they can be important tools for users to evaluate and place trust in automated systems.
Weakly supervised localization. Another relevant line of work is weakly supervised localization in the context of CNNs, where the task is to localize objects in images using only whole image class labels [7, 33, 34, 47].

Most relevant to our approach is the Class Activation Mapping (CAM) approach to localization [47]. This approach modifies image classification CNN architectures replacing fully-connected layers with convolutional layers and global average pooling [25], thus achieving class-specific feature maps. Others have investigated similar methods using global max pooling [34] and log-sum-exp pooling [35].

A drawback of CAM is that it requires feature maps to directly precede softmax layers, so it is only applicable to a particular kind of CNN architectures performing global average pooling over convolutional maps immediately prior to prediction (*i.e*. conv feature maps $\rightarrow$ global average pooling $\rightarrow$ softmax layer). Such architectures may achieve inferior accuracies compared to general networks on some tasks (*e.g*. image classification) or may simply be inapplicable to any other tasks (*e.g*. image captioning or VQA). We introduce a new way of combining feature maps using the gradient signal that does not require *any* modification in the network architecture. This allows our approach to be applied to any CNN-based architecture, including those for image captioning and visual question answering. For a fully-convolutional architecture, Grad-CAM reduces to CAM. Thus, Grad-CAM is a generalization to CAM.

Other methods approach localization by classifying perturbations of the input image. Zeiler and Fergus [45] perturb inputs by occluding patches and classifying the occluded image, typically resulting in lower classification scores for relevant objects when those objects are occluded. This principle is applied for localization in [4]. Oquab *et al*. [33] classify many patches containing a pixel then average these patch class-wise scores to provide the pixel's class-wise score. Unlike these, our approach achieves localization in one shot; it only requires a single forward and a partial backward pass per image and thus is typically an order of magnitude more efficient. In recent work Zhang *et al*. [46] introduce contrastive Marginal Winning Probability (c-MWP), a probabilistic Winner-Take-All formulation for modelling the top-down attention for neural classification models which can highlight discriminative regions. This is slower than Grad-CAM and like CAM, it only works for Image Classification CNNs. Moreover, quantitative and qualitative results are worse than for Grad-CAM (see Sec. 4.1 and [38]).

3. Approach

A number of previous works have asserted that deeper representations in a CNN capture higher-level visual constructs [5, 31]. Furthermore, convolutional features naturally retain spatial information which is lost in fully-connected layers, so we can expect the last convolutional layers to have the best compromise between high-level semantics and detailed spatial information. The neurons in these layers look for semantic class-specific information in the image (say object parts). Grad-CAM uses the gradient information flowing into the last convolutional layer of the CNN to understand the importance of each neuron for a decision of interest. *Although our technique is very generic and can be used to visualize any activation in a deep network, in this work we focus on explaining decisions the network can possibly make.*

As shown in Fig. 2, in order to obtain the class-discriminative localization map Grad-CAM $L^c_{\text{Grad-CAM}} \in \mathbb{R}^{u \times v}$ of width u and height v for any class c , we first compute the gradient of the score for class c, y^c (before the softmax), with respect to feature maps A^k of a convolutional layer, *i.e*. $\frac{\partial y^c}{\partial A^k}$. These gradients flowing back are global-

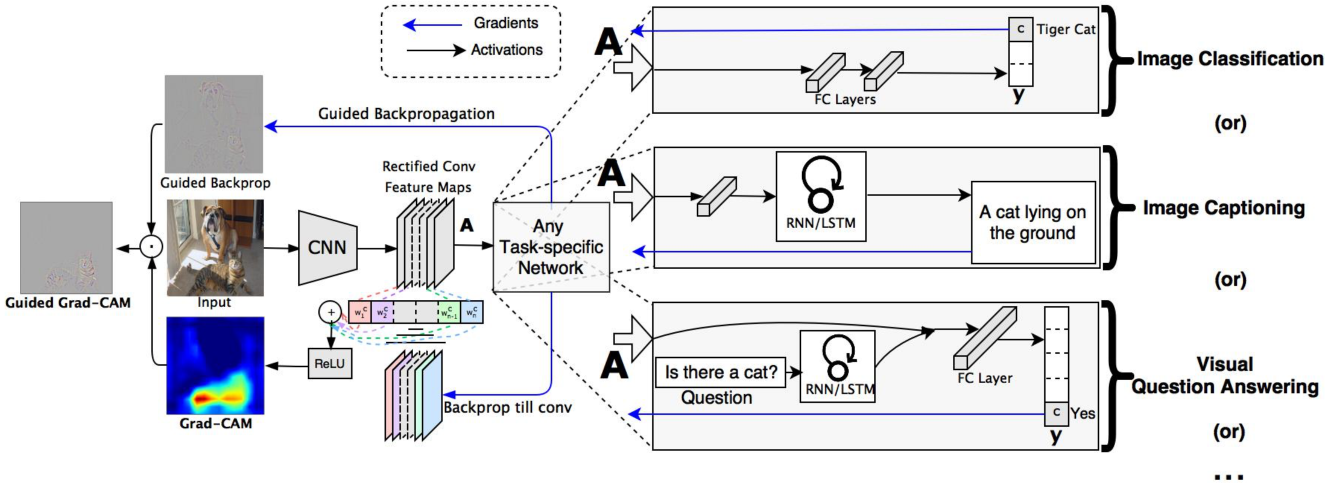


Figure 2: Grad-CAM overview: Given an image and a class of interest (*e.g.*, 'tiger cat' or any other type of differentiable output) as input, we forward propagate the image through the CNN part of the model and then through task-specific computations to obtain a raw score for the category. The gradients are set to zero for all classes except the desired class (tiger cat), which is set to 1. This signal is then backpropagated to the rectified convolutional feature maps of interest, which we combine to compute the coarse Grad-CAM localization (blue heatmap) which represents where the model has to look to make the particular decision. Finally, we pointwise multiply the heatmap with guided backpropagation to get Guided Grad-CAM visualizations which are both high-resolution and concept-specific.

average-pooled to obtain the neuron importance weights α_k^c:

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}} \tag{1}$$

This weight α_k^c represents a *partial linearization* of the deep network downstream from A, and captures the 'importance' of feature map k for a target class c.

We perform a weighted combination of forward activation maps, and follow it by a ReLU to obtain,

$$L_{\text{Grad-CAM}}^c = ReLU \underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}} \tag{2}$$

Notice that this results in a coarse heat-map of the same size as the convolutional feature maps (14×14 in the case of last convolutional layers of VGG [41] and AlexNet [24] networks). We apply a ReLU to the linear combination of maps because we are only interested in the features that have a *positive* influence on the class of interest, *i.e.* pixels whose intensity should be *increased* in order to increase y^c. Negative pixels are likely to belong to other categories in the image. As expected, without this ReLU, localization maps sometimes highlight more than just the desired class and achieve lower localization performance. Figures 1c, 1f and 1i, 1l show Grad-CAM visualizations for 'tiger cat' and 'boxer (dog)' respectively. Ablation studies and more Grad-CAM visualizations can be found in [38]. *In general, y^c need not be the class score produced by an image classification CNN. It could be any differentiable activation including words from a caption or the answer to a question.*

Grad-CAM as a generalization to CAM. Recall that CAM [47] produces a localization map for an image classification CNN with a specific kind of architecture where global average pooled convolutional feature maps are fed directly into softmax. Specifically, let the penultimate layer produce K feature maps, $A^k \in \mathbb{R}^{u \times v}$. These feature maps are then spatially pooled using Global Average Pooling (GAP) and linearly transformed to produce a score S^c for each class c,

$$S^c = \sum_k \underbrace{w_k^c}_{\text{class feature weights}} \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{A_{ij}^k}_{\text{feature map}} \tag{3}$$

To produce the localization map for modified image classification architectures, such as above, the order of summations can be interchanged to obtain L_{CAM}^c,

$$S^c = \frac{1}{Z}\sum_i \sum_j \underbrace{\sum_k w_k^c A_{ij}^k}_{L_{\text{CAM}}^c} \tag{4}$$

Note that this modification of architecture necessitates retraining because not all architectures have weights w_k^c connecting features maps to outputs. When Grad-CAM is applied to these architectures $\alpha_k^c = w_k^c$ — making Grad-CAM a strict generalization of CAM (see appendix A of [38] for details).

The above generalization also allows us to generate visual explanations from CNN-based models that cascade convolutional layers with much more complex interactions. Indeed, we apply Grad-CAM to 'beyond classification' tasks including models that utilize CNNs for image captioning and Visual Question Answering (VQA) (Sec. 7.2).

Guided Grad-CAM. While Grad-CAM visualizations are class-discriminative and localize relevant image regions well, they lack the ability to show fine-grained importance like pixel-space gradient visualization methods (Guided Backpropagation and Deconvolution). For example in Figure 1c, Grad-CAM can easily localize the cat region; however, it is unclear from the low-resolutions of the heat-map why the network predicts this particular instance as 'tiger cat'. In order to combine the best aspects of both, we fuse Guided Backpropagation and Grad-CAM visualizations via pointwise multiplication ($L^c_{\text{Grad-CAM}}$ is first up-sampled to the input image resolution using bi-linear interpolation). Fig. 2 bottom-left illustrates this fusion. This visualization is both high-resolution (when the class of interest is 'tiger cat', it identifies important 'tiger cat' features like stripes, pointy ears and eyes) and class-discriminative (it shows the 'tiger cat' but not the 'boxer (dog)'). Replacing Guided Backpropagation with Deconvolution in the above gives similar results, but we found Deconvolution to have artifacts (and Guided Backpropagation visualizations were generally less noisy), so we chose Guided Backpropagation over Deconvolution.

4. Evaluating Localization

4.1. Weakly-supervised Localization

In this section, we evaluate the localization capability of Grad-CAM in the context of image classification. The ImageNet localization challenge [9] requires competing approaches to provide bounding boxes in addition to classification labels. Similar to classification, evaluation is performed for both the top-1 and top-5 predicted categories. Given an image, we first obtain class predictions from our network and then generate Grad-CAM maps for each of the predicted classes and binarize with threshold of 15% of the max intensity. This results in connected segments of pixels and we draw our bounding box around the single largest segment.

We evaluate the pretrained off-the-shelf VGG-16 [41] model from the Caffe [19] Model Zoo. Following ILSVRC-15 evaluation, we report both top-1 and top-5 localization error on the val set in Table. 1. Grad-CAM localization errors are significantly lower than those achieved by c-MWP [46] and Simonyan *et al*. [40] for the VGG-16 model, which uses grabcut to post-process image space gradients into heat maps. Grad-CAM also achieves better top-1 localization error than CAM [47], which requires a change in the model architecture, necessitates re-training and thereby achieves worse classification errors (2.98% increase in top-1), whereas Grad-CAM makes no compromise on classification performance.

Method	Top-1 loc error	Top-5 loc error	Top-1 cls error	Top-5 cls error
Backprop on VGG-16 [40]	61.12	51.46	30.38	10.89
c-MWP on VGG-16 [46]	70.92	63.04	30.38	10.89
Grad-CAM on VGG-16 (ours)	56.51	46.41	30.38	10.89
VGG-16-GAP (CAM) [47]	57.20	45.14	33.40	12.20

Table 1: Classification and Localization on ILSVRC-15 val (lower is better).

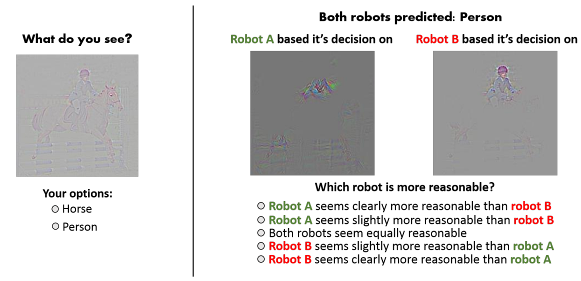


Figure 3: AMT interfaces for evaluating different visualizations for class discrimination (left) and trustworthiness (right). Guided Grad-CAM outperforms baseline approaches (Guided-backprop and Deconvolution) showing that our visualizations are more class-discriminative and help humans place trust in a more accurate classifier.

5. Evaluating Visualizations

Our first human study evaluates the main premise of our approach: are Grad-CAM visualizations more class-discriminative than previous techniques? Having established that, we turn to understanding whether it can lead an end user to trust the visualized models appropriately. For these experiments, we compare VGG-16 and AlexNet CNNs fine-tuned on PASCAL VOC 2007 train set and use the val set to generate visualizations.

5.1. Evaluating Class Discrimination

In order to measure whether Grad-CAM helps distinguish between classes we select images from VOC 2007 val set that contain exactly two annotated categories and create visualizations for each one of them. For both VGG-16 and AlexNet CNNs, we obtain category-specific visualizations using four techniques: Deconvolution, Guided Backpropagation, and Grad-CAM versions of each these methods (Deconvolution Grad-CAM and Guided Grad-CAM). We show visualizations to 43 workers on Amazon Mechanical Turk (AMT) and ask them "Which of the two object categories is depicted in the image?" as shown in Fig. 3.

Intuitively, a good prediction explanation is one that produces discriminative visualizations for the class of interest. The experiment was conducted using all 4 visualizations for 90 image-category pairs (*i.e*. 360 visualizations); 9 ratings were collected for each image, evaluated against the ground truth and averaged to obtain the accuracy. When viewing Guided Grad-CAM, human subjects can correctly identify the category being visualized in 61.23% of cases (compared to 44.44% for Guided Backpropagation; thus, Grad-CAM improves human performance by *16.79%*). Similarly, we also find that Grad-CAM helps make Deconvolution more class-discriminative (from 53.33% to 61.23%). Guided Grad-CAM performs the best among all the methods. Interestingly, our results seem to indicate that Deconvolution is more class discriminative than Guided Backpropagation, although Guided Backpropagation is more aesthetically pleasing than Deconvolution. To the best of our knowledge, our evaluations are the first to quantify this subtle difference.

5.2. Evaluating Trust

Given two prediction explanations, we evaluate which seems more trustworthy. We use AlexNet and VGG-16 to compare Guided Backpropagation and Guided Grad-CAM visualizations, noting that VGG-16 is known to be more reliable than AlexNet with an accuracy of 79.09 mAP (*vs*. 69.20 mAP) on PASCAL classification. In order to tease apart the efficacy of the visualization from the accuracy of the model being visualized, we consider only those instances where *both* models made the same prediction as ground truth. Given a visualization from AlexNet and one from VGG-16, and the predicted object category, 54 AMT workers were instructed to rate the reliability of the models relative to each other on a scale of clearly more/less reliable (+/-2), slightly more/less reliable (+/-1), and equally reliable (0). This interface is shown in Fig. 3. To eliminate any biases, VGG and AlexNet were assigned to be *model1* with approximately equal probability. Remarkably, we find that human subjects are able to identify the more accurate classifier (VGG over AlexNet) *simply from the different explanations, despite identical predictions.* With Guided Backpropagation, humans assign VGG an average score of 1.00 which means it is slightly more reliable than AlexNet, while Guided Grad-CAM achieves a higher score of 1.27 which is closer to saying that VGG is clearly more reliable. Thus our visualization can help users place trust in a model that can generalize better, just based on individual prediction explanations.

5.3. Faithfulness *vs*. Interpretability

Faithfulness of a visualization to a model is its ability to accurately explain the function learned by the model. Naturally, there exists a trade-off between the interpretability and faithfulness of a visualization: a more faithful visualization is typically less interpretable and *vice versa*. In fact, one could argue that a fully faithful explanation is the entire description of the model, which in the case of deep models is not interpretable/easy to visualize. We have verified in previous sections that our visualizations are reasonably interpretable. We now evaluate how faithful they are to the underlying model. One expectation is that our explanations should be locally accurate, *i.e*. in the vicinity of the input data point, our explanation should be faithful to the model [37].

For comparison, we need a reference explanation with high local-faithfulness. One obvious choice for such a visualization is image occlusion [45], where we measure the difference in CNN scores when patches of the input image are masked. Interestingly, patches which change the CNN score are also patches to which Grad-CAM and Guided Grad-CAM assign high intensity, achieving rank correlation 0.254 and 0.261 (*vs*. 0.168, 0.220 and 0.208 achieved by Guided Backpropagation, c-MWP and CAM, respectively) averaged over 2510 images in PASCAL 2007 val set. This shows that Grad-CAM visualizations are more faithful to the original model compared to all existing methods. Through localization experiment and human studies, we see that Grad-CAM visualizations are *more interpretable*, and through correlation with occlusion maps we see that Grad-CAM is *more faithful* to the model.

6. Diagnosing image classification CNNs

6.1. Analyzing Failure Modes for VGG-16

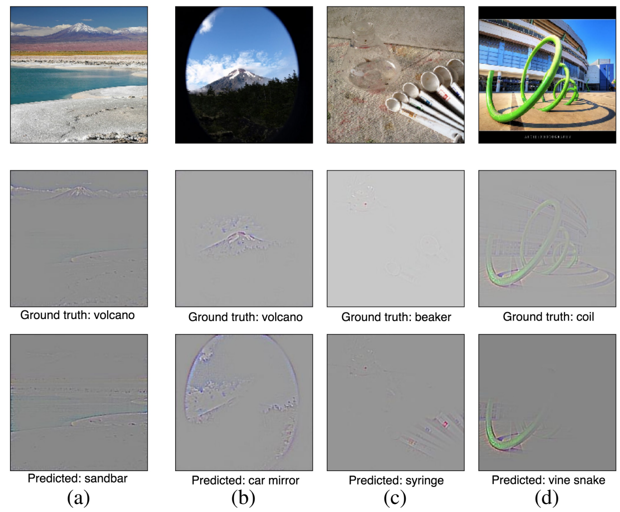


Figure 4: In these cases the model (VGG-16) failed to predict the correct class in its top 1 (a and d) and top 5 (b and c) predictions. Humans would find it hard to explain some of these predictions without looking at the visualization for the predicted class. But with Grad-CAM, these mistakes seem justifiable.

We use Guided Grad-CAM to analyze failure modes of the VGG-16 CNN on ImageNet classification [9]. In order to see what mistakes a network is making we first get a list of examples that the network (VGG-16) fails to classify correctly. For the misclassified examples, we use Guided Grad-CAM to visualize both the correct and the predicted class. A major advantage of Guided Grad-CAM visualization over other methods that allows for this analysis is its high-resolution and its ability to be highly class-discriminative. As seen in Fig. 4, some failures are due to ambiguities inherent in ImageNet classification. We can also see that *seemingly unreasonable predictions have reasonable explanations*, an observation also made in HOGgles [44].

6.2. Identifying bias in dataset

In this section we demonstrate another use of Grad-CAM: identifying and thus reducing bias in training datasets. Models trained on biased datasets may not generalize to real-world scenarios, or worse, may perpetuate biases and stereotypes (w.r.t. gender, race, age, *etc*.). We finetune an ImageNet trained VGG-16 model for the task of classifying “doctor” *vs*. “nurse”. We built our training dataset using the top 250 relevant images (for each class) from a popular image search engine. Although the trained model achieved a good validation accuracy, it did not generalize as well (82%).

Grad-CAM visualizations of the model predictions revealed that the model had learned to look at the person's face / hairstyle to distinguish nurses from doctors, thus learning a gender stereotype. Indeed, the model was misclassifying several female doctors to be a nurse and male nurses to be a doctor. Clearly, this is problematic. Turns out the image search results were gender-biased (78% of images for doctors were men, and 93% images for nurses were women).

Through this intuition gained from our visualization, we reduced the bias from the training set by adding in male nurses and female doctors to the training set, while maintaining the same number of images per class as before. The re-trained model now generalizes better to a more balanced test set (90%). Additional analysis along with Grad-CAM visualizations from both models can be found in [38]. This experiment demonstrates that Grad-CAM can help detect and remove biases in datasets, which is important not just for generalization, but also for fair and ethical outcomes as more algorithmic decisions are made in society.

7. Image Captioning and VQA

Finally, we apply our Grad-CAM technique to the image captioning [6, 21, 43] and Visual Question Answering (VQA) [3, 14, 32, 36] tasks. We find that Grad-CAM leads to interpretable visual explanations for these tasks as compared to baseline visualizations which do not change noticeably across different predictions. Note that existing visualization techniques are either not class-discriminative (Guided Backpropagation, Deconvolution), simply cannot be used for these tasks or architectures, or both (CAM or c-MWP).

7.1. Image Captioning

In this section, we visualize spatial support for an image captioning model using Grad-CAM. We build Grad-CAM on top of the publicly available neuraltalk2[2] implementation [23] that uses a finetuned VGG-16 CNN for images and an LSTM-based language model. Note that this model does not have an explicit attention mechanism. Given a caption, we compute the gradient of its log probability w.r.t. units in the last convolutional layer of the CNN ($conv5\_3$ for VGG-16) and generate Grad-CAM visualizations as described in Section 3. See Fig. 5a. In the first example, the Grad-CAM maps for the generated caption localize every occurrence of both the kites and people in spite of their relatively small size. In the next example, notice how Grad-CAM correctly highlights the pizza and the man, but ignores the woman nearby, since 'woman' is not mentioned in the caption. More qualitative examples can be found in [38].

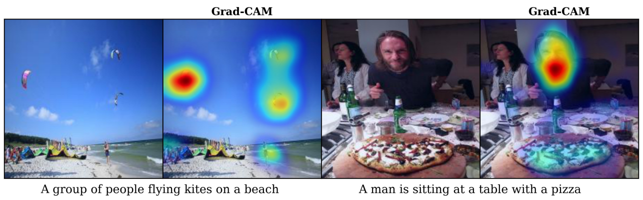
[2] https://github.com/karpathy/neuraltalk2

Comparison to dense captioning. Johnson *et al*. [21] recently introduced the Dense Captioning (DenseCap) task that requires a system to jointly localize and caption salient regions in a given image. Their model consists of a Fully Convolutional Localization Network (FCLN) and an LSTM-based language model that produces both bounding boxes for regions of interest and associated captions in a single forward pass. Using DenseCap, we generate 5 region-specific captions per image with associated ground truth bounding boxes. A whole-image captioning model (neuraltalk2) should localize a caption inside the box it was generated for, which is shown in Fig. 5b. We measure this by computing the ratio of average activation inside *vs*. outside the box. A higher ratio is better because it indicates stronger attention on the region that generated the caption. Uniformly highlighting the whole image results in a baseline ratio of 1.0 whereas Grad-CAM achieves 3.27 ± 0.18. Adding high-resolution detail gives an improved baseline of 2.32 ± 0.08 (Guided Backpropagation) and the best localization at 6.38 ± 0.99 (Guided Grad-CAM). This means Grad-CAM localizations correspond to regions in the image that the DenseCap model described, even though the holistic captioning model was not trained with any region or bounding-box level annotations.

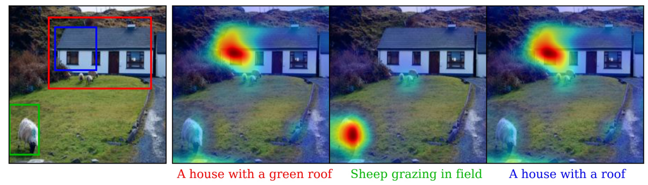
7.2. Visual Question Answering

Typical VQA pipelines [3, 14, 32, 36] consist of a CNN to model images and an RNN language model for questions. The image and the question representations are fused to predict the answer, typically with a 1000-way classification. Since this is a classification problem, we pick an answer (the score y^c in (3)) and use its score to compute Grad-CAM to show image evidence that supports the answer. Despite the complexity of the task, involving both visual and language components, the explanations (of the VQA model from [28]) described in Fig. 6 are surprisingly intuitive and informative. We quantify the performance of Grad-CAM via correlation with occlusion maps, as in Section 5.3. Grad-CAM achieves a rank correlation (with occlusion map) of 0.60 ± 0.038 whereas Guided Backpropagation achieves 0.42 ± 0.038, indicating higher faithfulness of our Grad-CAM visualization.

Comparison to Human Attention. Das *et al*. [8] collected human attention maps for a subset of the VQA dataset [3]. These maps have high intensity where humans looked in the image in order to answer a visual question. Human attention maps are compared to Grad-CAM visualizations for the VQA model from [28] on 1374 val question-image (QI) pairs from [3] using the rank correlation evaluation protocol developed in [8]. Grad-CAM and human attention maps have a correlation of 0.136, which is statistically higher than chance or random attention maps (zero correlation). This shows that despite not being trained on grounded image-text pairs, even non-attention based CNN + LSTM based VQA models are surprisingly good at localizing discriminative

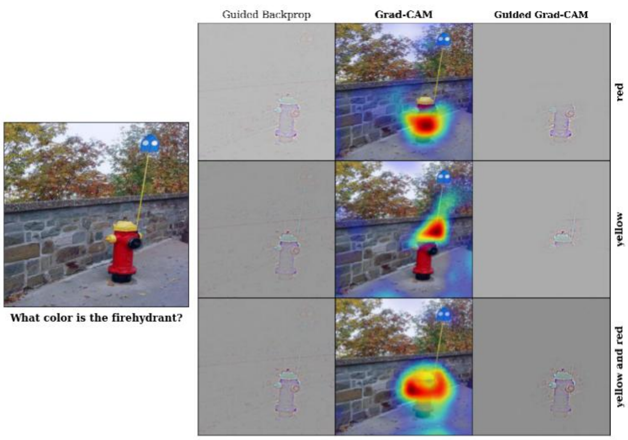


(a) Image captioning explanations

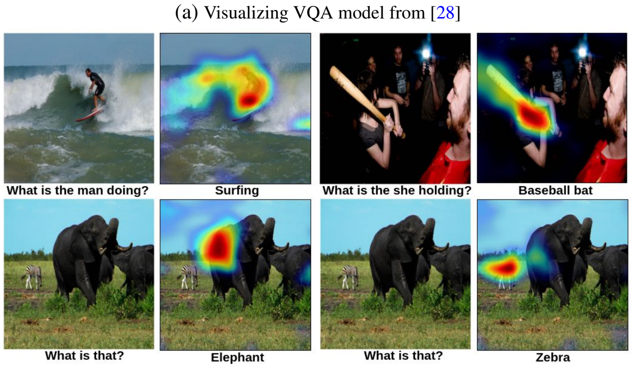


(b) Comparison to DenseCap

Figure 5: Interpreting image captioning models: We use our class-discriminative localization technique, Grad-CAM to find spatial support regions for captions in images. Fig. 5a Visual explanations from image captioning model [23] highlighting image regions considered to be important for producing the captions. Fig. 5b Grad-CAM localizations of a *global* or *holistic* captioning model for captions generated by a dense captioning model [21] for the three bounding box proposals marked on the left. We can see that we get back Grad-CAM localizations (right) that agree with those bounding boxes – even though the captioning model and Grad-CAM techniques do not use any bounding box annotations.



(a) Visualizing VQA model from [28]



(b) Visualizing ResNet based Hierarchical co-attention VQA model from [29]

Figure 6: Qualitative Results for our VQA experiments: (a) Given the image on the left and the question "What color is the firehydrant?", we visualize Grad-CAMs and Guided Grad-CAMs for the answers "red", "yellow" and "yellow and red". Grad-CAM visualizations are highly interpretable and help explain any target prediction – for "red", the model focuses on the bottom red part of the firehydrant; when forced to answer "yellow", the model concentrates on it's top yellow cap, and when forced to answer "yellow and red", it looks at the whole firehydrant! (b) Our approach is capable of providing interpretable explanations even for complex models.

regions required to output a particular answer.

Visualizing ResNet-based VQA model with attention. Lu *et al*. [29] use a 200 layer ResNet [16] to encode the image, and jointly learn a hierarchical attention mechanism on the question and image. Fig. 6b shows Grad-CAM visualization for this network. As we visualize deeper layers of the ResNet we see small changes in Grad-CAM for most adjacent layers and larger changes between layers that involve dimensionality reduction. Visualizations for various layers in ResNet can be found in [38]. To the best of our knowledge, we are the first to visualize decisions made by ResNet-based architectures.

8. Conclusion

In this work, we proposed a novel class-discriminative localization technique—Gradient-weighted Class Activation Mapping (Grad-CAM)—for making *any* CNN-based models more transparent by producing visual explanations. Further, we combined our Grad-CAM localizations with existing high-resolution visualizations to obtain high-resolution class-discriminative Guided Grad-CAM visualizations. Our visualizations outperform all existing approaches on both aspects: interpretability and faithfulness to original model. Extensive human studies reveal that our visualizations can discriminate between classes more accurately, better reveal the trustworthiness of a classifier, and help identify biases in datasets. Finally, we showed the broad applicability of Grad-CAM to various off-the-shelf available architectures for tasks including image classification, image captioning and VQA providing faithful visual explanations for possible model decisions. We believe that a true AI system should not only be intelligent, but also be able to reason about its beliefs and actions for humans to trust it. Future work includes explaining the decisions made by deep networks in domains such as reinforcement learning, natural language processing and video applications.

9. Acknowledgements

This work was funded in part by NSF CAREER awards to DB and DP, ONR YIP awards to DP and DB, ONR Grant N00014-14-1-0679 to DB, a Sloan Fellowship to DP, ARO YIP awards to DB and DP, an Allen Distinguished Investigator award to DP from the Paul G. Allen Family Foundation, ICTAS Junior Faculty awards to DB and DP, Google Faculty Research Awards to DP and DB, Amazon Academic Research Awards to DP and DB, AWS in Education Research grant to DB, and NVIDIA GPU donations to DB. SK was supported by ONR Grant N00014-12-1-0903. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government, or any sponsor.

References

[1] A. Agrawal, D. Batra, and D. Parikh. Analyzing the Behavior of Visual Question Answering Models. In *EMNLP*, 2016. 1

[2] H. Agrawal, C. S. Mathialagan, Y. Goyal, N. Chavali, P. Banik, A. Mohapatra, A. Osman, and D. Batra. CloudCV: Large Scale Distributed Computer Vision as a Cloud Service. In *Mobile Cloud Visual Media Computing*, pages 265–290. Springer, 2015. 1

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 1, 7

[4] L. Bazzani, A. Bergamo, D. Anguelov, and L. Torresani. Self-taught object localization with deep networks. In *WACV*, 2016. 3

[5] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. 3

[6] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft COCO captions: Data Collection and Evaluation Server. *arXiv preprint arXiv:1504.00325*, 2015. 1, 7

[7] R. G. Cinbis, J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 2016. 3

[8] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? In *EMNLP*, 2016. 7

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 2, 5, 6

[10] A. Dosovitskiy and T. Brox. Inverting Convolutional Networks with Convolutional Networks. In *CVPR*, 2015. 3

[11] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing Higher-layer Features of a Deep Network. *University of Montreal*, 1341, 2009. 3

[12] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From Captions to Visual Concepts and Back. In *CVPR*, 2015. 1

[13] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann. Devnet: A deep event network for multimedia event detection and evidence recounting. In *CVPR*, 2015. 3

[14] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering. In *NIPS*, 2015. 1, 7

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *CVPR*, 2014. 1

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 2, 3, 8

[17] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing Error in Object Detectors. In *ECCV*, 2012. 1

[18] P. Jackson. *Introduction to Expert Systems*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 3rd edition, 1998. 1

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. In *ACM MM*, 2014. 5

[20] E. Johns, O. Mac Aodha, and G. J. Brostow. Becoming the Expert - Interactive Multi-Class Machine Teaching. In *CVPR*, 2015. 1

[21] J. Johnson, A. Karpathy, and L. Fei-Fei. DenseCap: Fully Convolutional Localization Networks for Dense Captioning. In *CVPR*, 2016. 1, 7, 8

[22] A. Karpathy. What I learned from competing against a ConvNet on ImageNet. http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/, 2014. 1

[23] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 7, 8

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1, 4

[25] M. Lin, Q. Chen, and S. Yan. Network in network. In *ICLR*, 2014. 3

[26] Z. C. Lipton. The Mythos of Model Interpretability. *ArXiv e-prints*, June 2016. 1, 3

[27] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[28] J. Lu, X. Lin, D. Batra, and D. Parikh. Deeper LSTM and normalized CNN Visual Question Answering model. `https://github.com/VT-vision-lab/VQA_LSTM_CNN`, 2015. 7, 8

[29] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 2016. 8

[30] A. Mahendran and A. Vedaldi. Salient deconvolutional networks. In *European Conference on Computer Vision*, 2016. 3

[31] A. Mahendran and A. Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, pages 1–23, 2016. 3

[32] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015. 1, 7

[33] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014. 3

[34] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free? – weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015. 3

[35] P. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015. 3

[36] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *NIPS*, 2015. 1, 7

[37] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *SIGKDD*, 2016. 3, 6

[38] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization. *CoRR*, abs/1610.02391, 2016. 3, 4, 7, 8

[39] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016. 1

[40] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013. 3, 5

[41] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, 2015. 4, 5

[42] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller. Striving for Simplicity: The All Convolutional Net. *CoRR*, abs/1412.6806, 2014. 2, 3

[43] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 1, 7

[44] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. HOGgles: Visualizing Object Detection Features. *ICCV*, 2013. 6

[45] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 2, 3, 6

[46] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down Neural Attention by Excitation Backprop. In *ECCV*, 2016. 3, 5

[47] B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. In *CVPR*, 2016. 2, 3, 4, 5