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

Abstract

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 a captioning network) flowing into t
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 multimodal 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 ques

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 robust to adversarial perturbations,
(d) are more faithful to the underlying model
(e) help achieve model generalization by identifying dataset bias.

For image captioning and VQA, our visualizations show that even non-attention based models learn to localize discriminative regions of the input image.
We devise a way to identify important neurons through Grad-CAM and combine it with neuron names [4] to provide textual explanations for model decisions.

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 success

Introduction

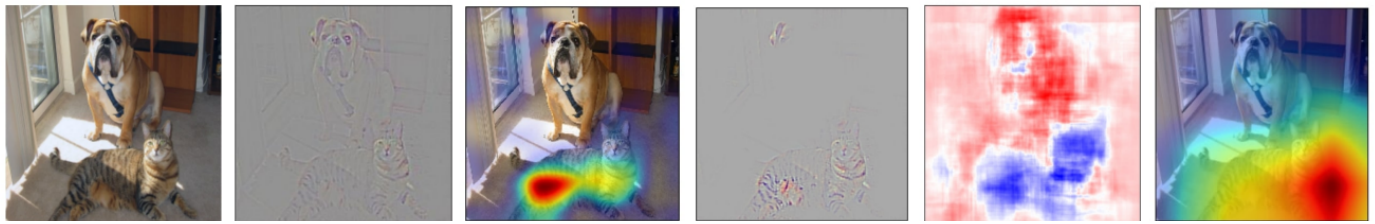
Broadly speaking, this transparency and ability to explain is useful at three different stages of Artificial Intelligence (AI) evolution.
First, when AI is significantly weaker than humans and not yet reliably deployable (e.g. visual question answering [3]), the goal of transparency and explanations is to identify the failure modes [1,25], thereby helping
Second, when AI is on par with humans and reliably deployable (e.g., image classification [30] 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 [50]), the goal of explanations is in machine teaching [28] – i.e., a machine teaching a human about how to make better decisions.

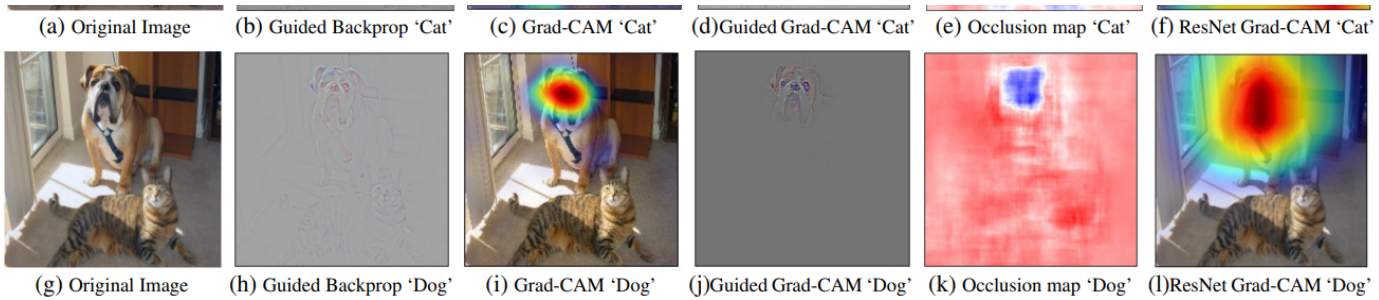
There typically exists a trade-off between accuracy and simplicity or interpretability.
As such, deep models are beginning to explore the spectrum between interpretability and accuracy.

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 [59] and is applicable to a significantly broader range of CNN model families:

What makes a good visual explanation?





Consider image classification [14] – 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)
(b) high-resolution (i.e. capture fine-grained detail).

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.

To summarize, our contributions are as follows:
(1) We introduce Grad-CAM, a class-discriminative localization technique that generates visual explanations for any CNN-based network without requiring architectural changes or re-training.
(2) We apply Grad-CAM to existing top-performing classification, captioning (Sec. 8.1), and VQA (Sec. 8.2) models.
(3) We show a proof-of-concept of how interpretable Grad-CAM visualizations help in diagnosing failure modes by uncovering biases in datasets.
(4) We present Grad-CAM visualizations for ResNets [24] applied to image classification and VQA (Sec. 8.2).
(5) We use neuron importance from Grad-CAM and neuron names from [4] and obtain textual explanations for model decisions (Sec. 7).
(6) We conduct human studies (Sec. 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' net

Related Work

Visualizing CNNs
Although these can be high-resolution and class-discriminative, they are not specific to a single input image and visualize a model overall.

Assessing Model Trust

Aligning Gradient-based Importances.

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 holistic image class labels only [8,43,44,59].
Most relevant to our approach is the Class Activation Mapping (CAM) approach to localization [59].
This approach modifies image classification CNN architectures, replacing fully-connected layers with convolutional layers and global average pooling [34], thus achieving class-specific feature maps.

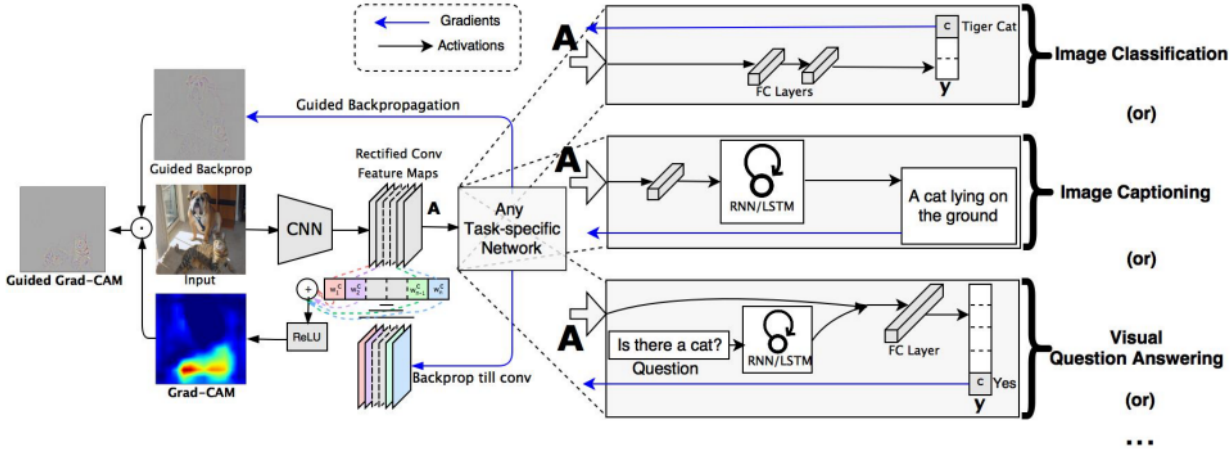
A drawback of CAM is that it requires feature maps to directly precede softmax layers, so it is only applicable to a particular kind of CNN architectures performing global average pooling over convolutional maps imm
Such architectures may achieve inferior accuracies compared to general networks on some 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 off-the-shelf CNN-based architectures, including those for image captioning and visual question answering.

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.

Grad-CAM

Grad-CAM uses the gradient information flowing into the last convolutional layer of the CNN to assign importance values to each neuron for a particular decision of interest.
Although our technique is fairly general in that it can be used to explain activations in any layer of a deep network, in this work, we focus on explaining output layer decisions only.



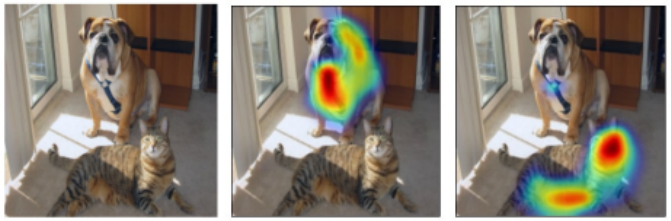
We apply a ReLU to the linear combination of maps because we are only interested in the features that have a positive influence on the class of interest, i.e. pixels whose intensity should be increased in order to inc

Grad-CAM generalizes CAM

Recall that CAM 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.
Grad-CAM is a strict generalization of CAM.
This generalization allows us to generate visual explanations from CNN-based models that cascade convolutional layers with much more complex interactions, such as those for image captioning and VQA (Sec. 8.2

Counterfactual Explanations

Using a slight modification to Grad-CAM, we can obtain explanations that highlight support for regions that would make the network change its prediction.
As a consequence, removing concepts occurring in those regions would make the model more confident about its prediction.



(a) Original Image (b) Cat Counterfactual exp (c) Dog Counterfactual exp

Evaluating Localization Ability of Grad-CAM

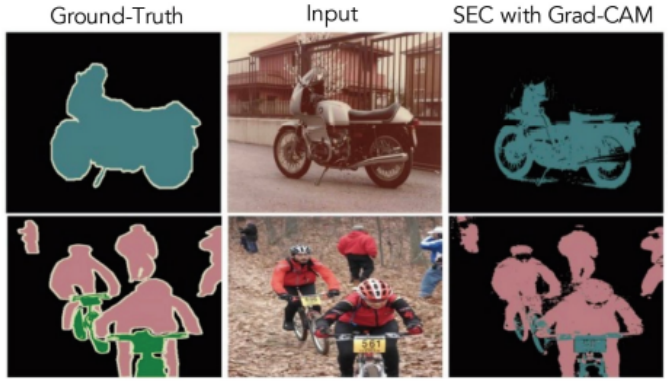
Weakly-supervised Localization

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 them with a threshold of 15% of the max intensity.
This results in connected segments of pixels and we draw a bounding box around the single largest segment.
Note that this is weakly-supervised localization – the models were never exposed to bounding box annotations during training.

		Classification		Localization	
		Top-1	**Top-5**	**Top-1**	**Top-5**
VGG-16	Backprop [51]	30.38	10.89	61.12	51.46
	c-MWP [58]	30.38	10.89	70.92	63.04
	Grad-CAM (ours)	30.38	10.89	**56.51**	46.41
	CAM [59]	33.40	12.20	57.20	**45.14**
AlexNet	c-MWP [58]	44.2	20.8	92.6	89.2
	Grad-CAM (ours)	44.2	20.8	68.3	56.6
GoogleNet	Grad-CAM (ours)	31.9	11.3	60.09	49.34
	CAM [59]	31.9	11.3	60.09	49.34

Weakly-supervised Segmentation

The task of weakly-supervised segmentation involves segmenting objects with just image-level annotation, which can be obtained relatively cheaply from image classification datasets.
In recent work, Kolesnikov et al. [32] introduced a new loss function for training weakly-supervised image segmentation models.
Their loss function is based on three principles
1) to seed with weak localization cues, encouraging segmentation network to match these cues,
2) to expand object seeds to regions of reasonable size based on information about which classes can occur in an image,
3) to constrain segmentations to object boundaries that alleviates the problem of imprecise boundaries already at training time.



Pointing Game

Their evaluation protocol first cues each visualization technique with the ground-truth object label and extracts the maximally activated point on the generated heatmap.
It then evaluates if the point lies within one of the annotated instances of the target object category, thereby counting it as a hit or a miss.

However, this evaluation only measures precision of the visualization technique.

Evaluating Visualizations

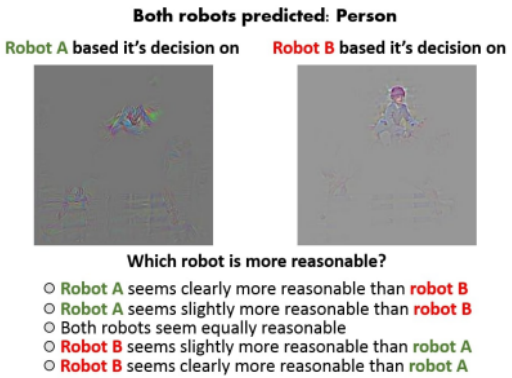
In this section, we describe the human studies and experiments we conducted to understand the interpretability vs. faithfulness tradeoff of our approach to model predictions.

Evaluating Class Discrimination

We show these visualizations to 43 workers on Amazon Mechanical Turk (AMT) and ask them “Which of the two object categories is depicted in the image?” (shown in Fig. 5).

(a) Raw input image. Note that this is not a part of the tasks (b) and (c)

(b) AMT interface for evaluating the class-discriminative property



(c) AMT interface for evaluating if our visualizations instill trust in an end user

Method	Human Classification Accuracy	Relative Reliability	Rank Correlation w/ Occlusion
Guided Backpropagation	44.44	+1.00	0.168
Guided Grad-CAM	61.23	+1.27	0.261

Evaluating Trust

We use AlexNet and VGG-16 to compare Guided Backpropagation and Guided Grad-CAM visualizations
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.
Thus, our visualizations can help users place trust in a model that generalizes better, just based on individual prediction explanations.

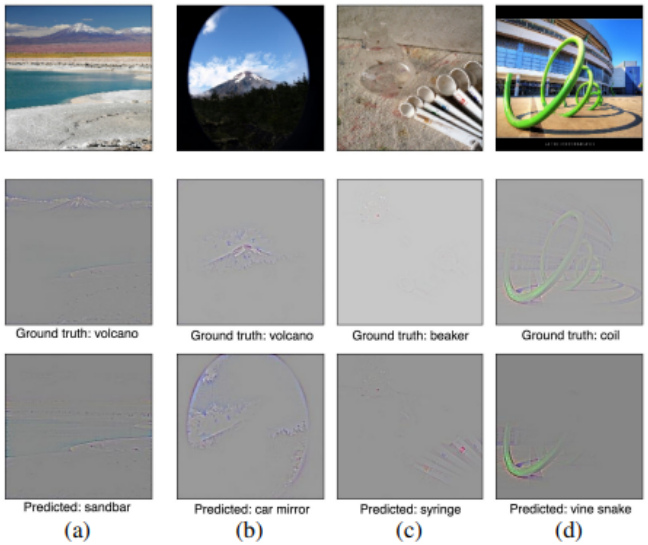
Faithfulness vs. Interpretability

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.

This shows that Grad-CAM is more faithful to the original model compared to prior methods.
Through localization experiments 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.

Diagnosing image classification CNNs with Grad-CAM

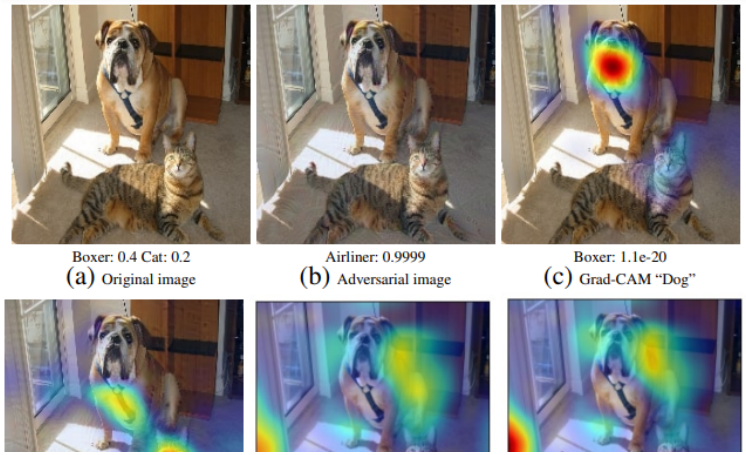
Analyzing failure modes for VGG-16

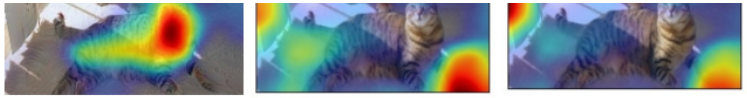


We can also see that seemingly unreasonable predictions have reasonable explanations, an observation also made in HOGgles [56].
A major advantage of Guided Grad-CAM visualizations over other methods is that due to its high-resolution and ability to be class-discriminative, it readily enables these analyses.

Effect of adversarial noise on VGG-16

This shows that Grad-CAM is fairly robust to adversarial noise.





Tiger Cat: 6.5e-17 | Airliner: 0.9999 | Space shuttle: 1e-5

(d) Grad-CAM "Cat" (e) Grad-CAM "Airliner" (f) Grad-CAM "Space Shuttle"

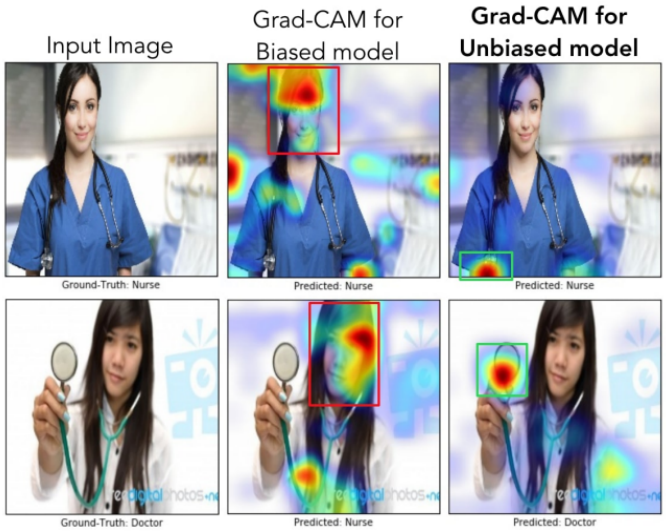
Identifying bias in dataset

We finetune an ImageNet-pretrained VGG-16 model for a "doctor" vs. "nurse" binary classification task.

Although the trained model achieves good validation accuracy, it does not generalize well (82% test accuracy).

Through these intuitions gained from Grad-CAM visualizations, we reduced bias in the training set by adding images of male nurses and female doctors, while maintaining the same number of images per class as b

This experiment demonstrates a proof-of-concept that Grad-CAM can help detect and remove biases in datasets, which is important not just for better generalization, but also for fair and ethical outcomes as more al

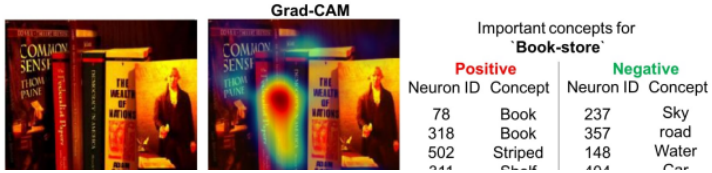


Textual Explanations with Grad-CAM

Using their approach. we first obtain neuron names for the last convolutional layer.

Next, we sort and obtain the top-5 and bottom-5 neurons based on their class-specific importance scores, αk.

The names for these neurons can be used as text explanations.



(a)

Grad-CAM for Image Captioning and VQA

Finally, we apply Grad-CAM to vision & language tasks such as image captioning [7, 29, 55] and Visual Question Answering (VQA) [3, 20, 42, 46].

Note that existing visualization techniques either are not class-discriminative (Guided Backpropagation, Deconvolution), or simply cannot be used for these tasks/architectures, or both (CAM, c-MWP).

Image Captioning

Note that this model does not have an explicit attention mechanism.

(a) Image captioning explanations

Comparison to dense captioning

Using DenseCap, we generate 5 region-specific captions per image with associated ground truth bounding boxes.

Higher ratios are better because they indicate stronger attention to the region the caption was generated for.

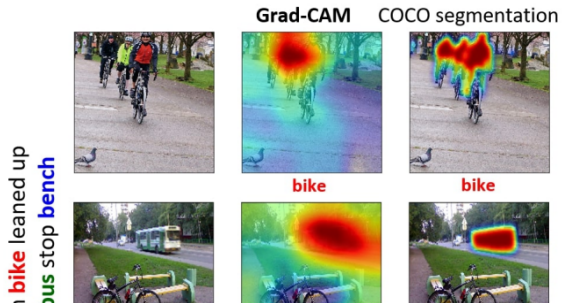
Uniformly highlighting the whole image results in a baseline ratio of 1.0 whereas Grad-CAM achieves 3.27 ± 0.18.

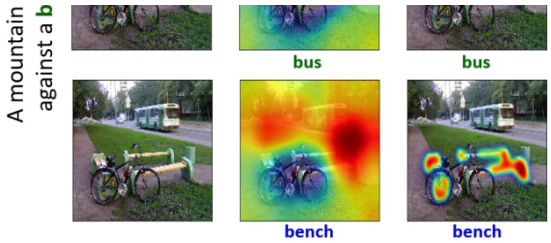
Adding high-resolution detail gives an improved baseline of 2.32 ± 0.08 (Guided Backpropagation) and the best localization at 6.38 ± 0.99 (Guided Grad-CAM).

Thus, Grad-CAM is able to localize regions in the image that the DenseCap model describes, even though the holistic captioning model was never trained with bounding-box annotations.

Grad-CAM for individual words of caption

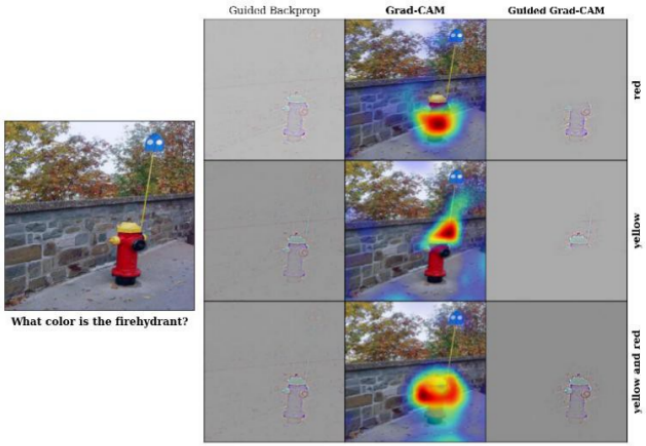
Comparison to Human Attention





Visual Question Answering

Despite the complexity of the task, involving both visual and textual components, the explanations (of the VQA model from Lu et al. [38]) described in Fig. 12 are surprisingly intuitive and informative.



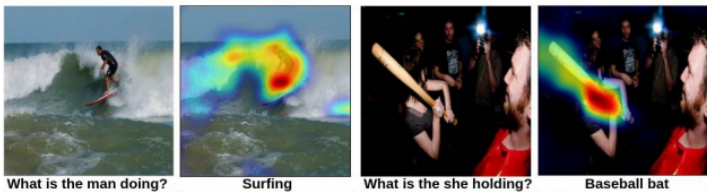
Comparison to Human Attention

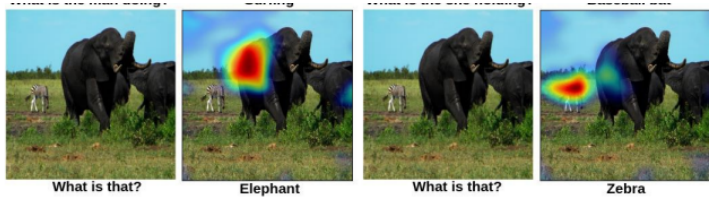
These maps have high intensity where humans looked in the image in order to answer a visual question.

Human attention maps are compared to Grad-CAM visualizations for the VQA model from [38] on 1374 val question-image (QI) pairs from [3] using the rank correlation evaluation protocol as in [9]. Grad-CAM and human attention maps have a correlation of 0.136, which is higher than chance or random attention maps (zero correlation).

Visualizing ResNet-based VQA model with co-attention.

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.





(b) Visualizing ResNet based Hierarchical co-attention VQA model from [39]

Conclusion

In this work, we proposed a novel class-discriminative localization technique – Gradient-weighted Class Activation Mapping (Grad-CAM) – for making any CNN-based model more transparent by producing visual ex
Further, we combined Grad-CAM localizations with existing high-resolution visualization techniques to obtain the best of both worlds – high-resolution and class-discriminative Guided Grad-CAM visualizations.

Finally, we show the broad applicability of Grad-CAM to various off-the-shelf architectures for tasks such as image classification, image captioning, and visual question answering.
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 and use it.
Future work includes explaining decisions made by deep networks in domains such as reinforcement learning, natural language processing, and video applications.