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

Paper Authors: Ramprasaath R. Selvaraju · Michael Cogswell · Abhishek Das · Ramakrishna Vedantam · Devi Parikh · Dhruv Batra

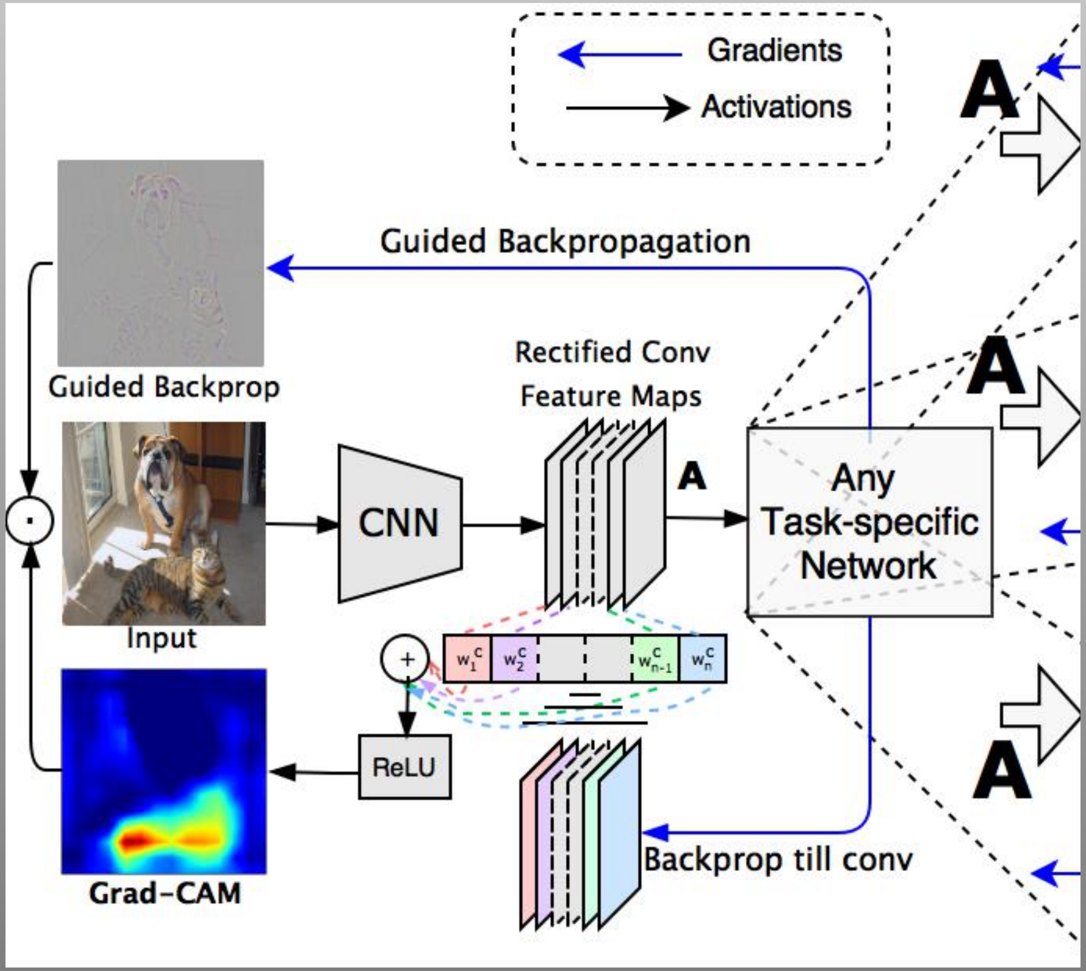


Image reproduced under fair use from https://arxiv.org/pdf/1610.02391.pdf

Overview

- Goal: Produce ‘visual explanations’ for decisions
    - CNNs
    - Transparent, Explainable AI
- Approach: Grad-CAM
    - Gradient-weighted Class Activation Mapping (Grad-CAM)
    - uses gradients
    - identifies important regions in an image
- Wider applicability without architectural changes or re-training:
    - CNNs with fully-connected layers
    - CNNs for structured outputs such as captioning
    - CNNs for multi-modal inputs such as visual question answering
    - reinforcement learning

Overview - II

- Grad-CAM visualizations
    - Explain failure modes
    - Outperform others on the ImageNet localization task
    - Robust against adversarial perturbations (?)
    - Assist with model generalization by identifying dataset bias.
- Identify influential neurons
- Explain decisions via text using neuron names.
- User studies

Introduction

- DNN’s interpretability challenged by lack of decomposability
    - decomposability into intuitive components
- AI needs to fail more gracefully
    - Explain cause of failure
- Transparent models
    - Why they predict what they predict
- Accuracy vs. Explainability
    - Expert rule-based systems more explainable

Three phases of AI

- AI < Human
  - Identify failure models
- AI ≅ Human
  - Trust and confidence
- AI > Human
  - Machine teaching
  - Enable better decision making in human beings

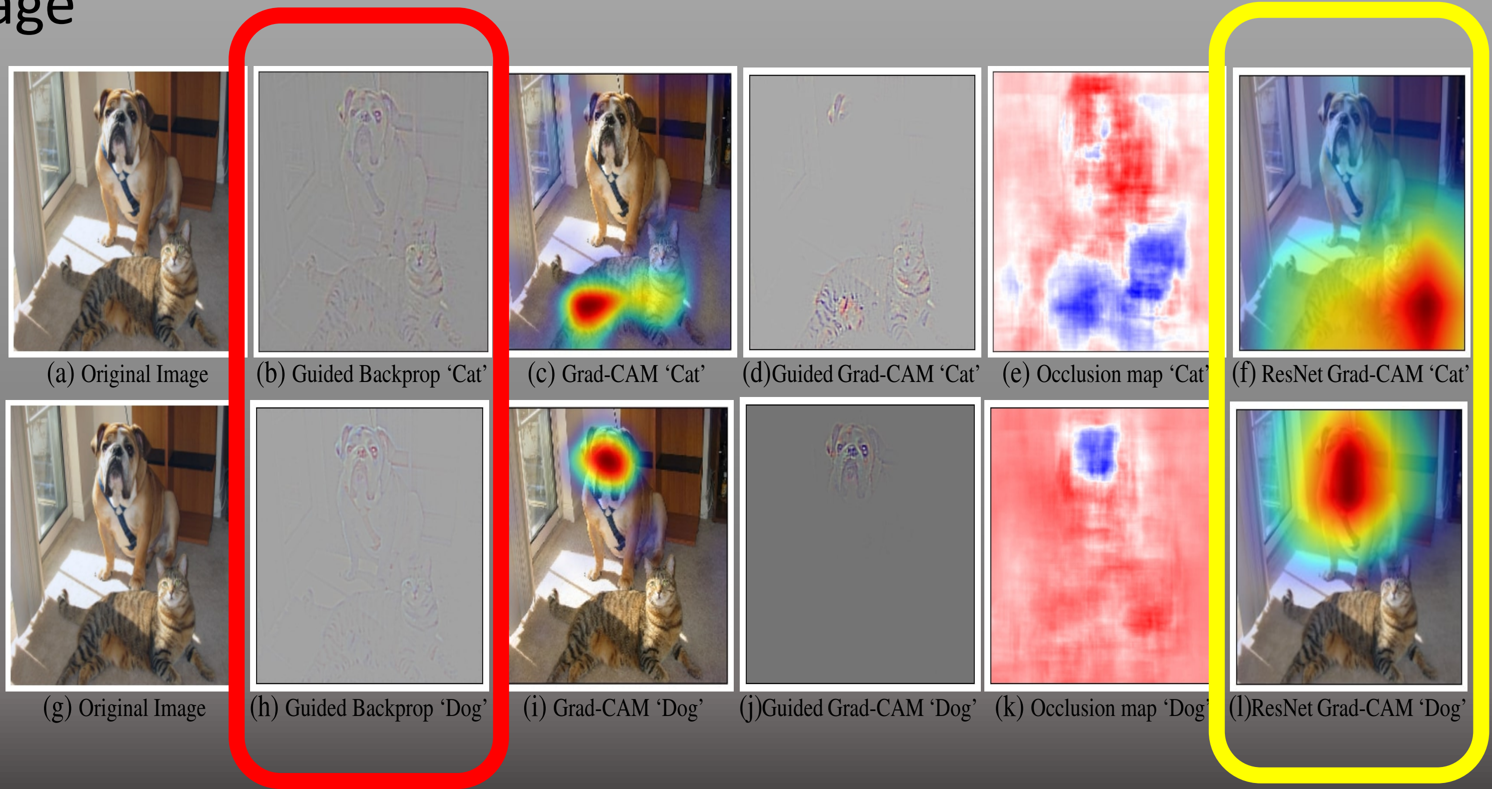
Prior Work: CAM

- Class Activation Mapping (CAM)
  - Attribution analysis for images
  - Subset of CNNs with no fully-connected layers.
- Grad-CAM focusses on SOTA DNNs such as ResNet
  - Fully connected layers
  - Structured outputs
  - Multi-modal inputs
  - RL

B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba.
Learning Deep Features for Discriminative Localization.
In *CVPR*, 2016.

Good visual explanation

- class-discriminative
  - localize the object in the image
- high-resolution
  - capture fine-grained detail
- Attributions in (b) and (h)



(a) Original Image (b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat' (d)Guided Grad-CAM 'Cat' (e) Occlusion map 'Cat' (f) ResNet Grad-CAM 'Cat'

(g) Original Image (h) Guided Backprop 'Dog' (i) Grad-CAM 'Dog' (j)Guided Grad-CAM 'Dog' (k) Occlusion map 'Dog' (l)ResNet Grad-CAM 'Dog'

Image reproduced under fair use from
https://arxiv.org/pdf/1610.02391.pdf

Prior Work

- **Visualizing CNNs**
  - Identify influential pixels or synthesize images for maximal activation
  - Simonyan *et al*. visualize partial derivatives

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

  - Modify these partial derivatives
    - Guided Backpropagation
    - Deconvolutions

M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

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

Prior Work - II

- **Assessing Model Trust**
  - Human subject studies to understand trust in AI.

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

- **Aligning Gradient-based Attributions to Human Attention Maps**
  - Map gradient-based attributions to class-specific human knowledge
  - Align gradient-based attributions to human attention maps

R.R.Selvaraju, S.Lee, Y.Shen, H.Jin, S.Ghosh, L.Heck, D.Batra, and D. Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. In ICCV 2019

Prior Work - III

- **Weakly-supervised localization**
  - Localize objects using image class labels

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

- **Class Activation Mapping (CAM)**
  - Modifies CNNs
    - feature maps must precede softmax
  - fully-connected layers replaced by
    - convolutional layers and
    - global average pooling
  - Related ideas: Gobal max pooling; also, log-sum-exp pooling

B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. In *CVPR*, 2016.

Prior Work - IV

- **Perturbing the input**
  - Classifying images with occluding patches

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

  - Use average score of multiple patches containing a pixel

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

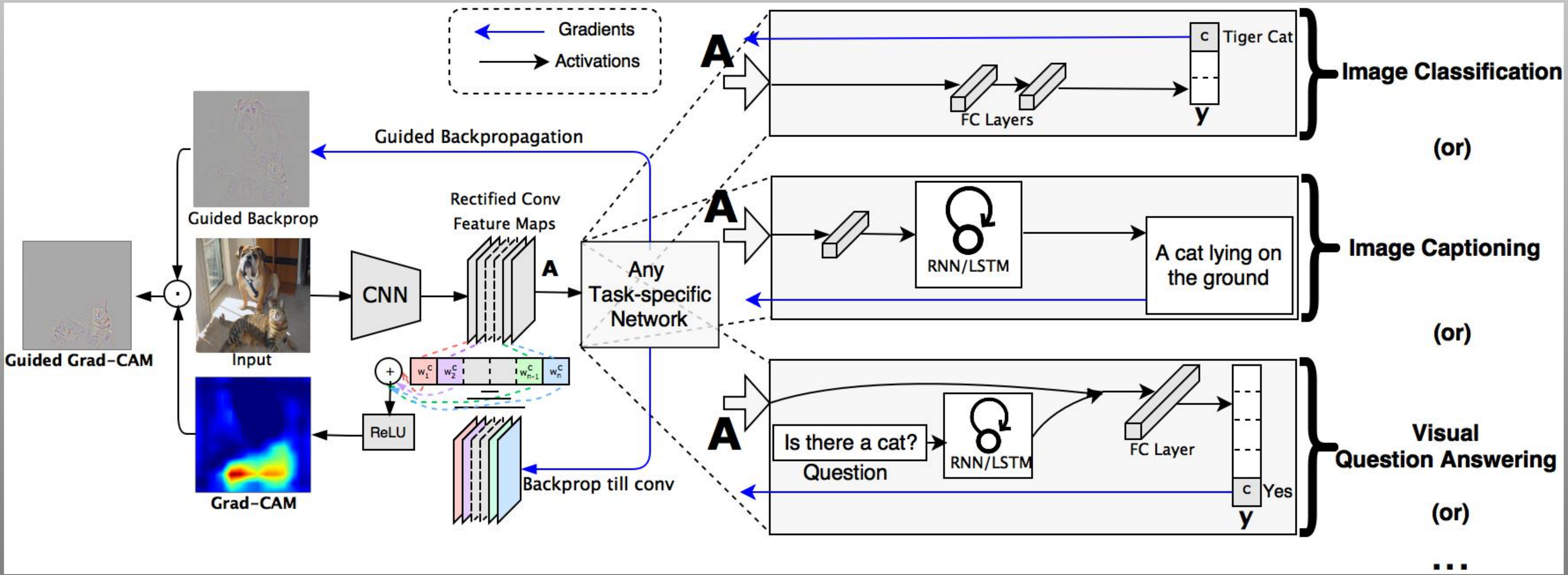


Fig. 2: Grad-CAM overview: Given an image and a class of interest (*e.g.*, 'tiger cat' or any other type of differentiable output) as input, we forward propagate the image through the CNN part of the model and then through task-specific computations to obtain a raw score for the category. The gradients 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.

Image reproduced under fair use from
https://arxiv.org/pdf/1610.02391.pdf

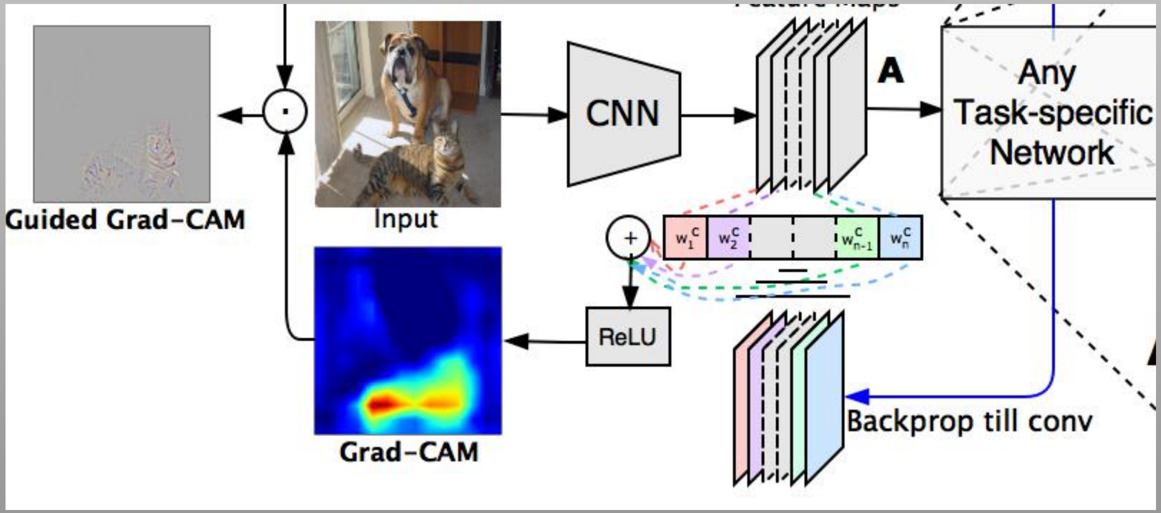
Motivation

- Deeper CNN representations describe higher-level visual information.

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.

- Convolutional layers naturally contain spatial information
- Lost in fully connected layers
- Intuitively, anticipate last convolutional layers to be most informative.
  - Spatial information
  - Discriminative semantic value
- High-level Idea of Grad-CAM: Employ the gradient of the last CNN convolutional layer for attribution analysis.

Technical Details - I



over the width and height dimension indexed by i and j

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

Neural importance weight i.e. importance of feature map k for class c

y^c: score of label c before softmax

A^k: Feature activation map of kth convolutional layer

Global-average-pooling empirically better than global-max-pooling

Technical Details - II

over the width and height dimension indexed by i and j

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

y^c: score of label c before softmax

A^k: Feature activation map of kth convolutional layer

Neural importance weight i.e. importance of feature map k for class c

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

Localization map for class c

Positive component of a linear combination of activation maps weighted by α_k^c

GradCAM + Guided Backpropagation

- Grad-CAM identifies image regions and can discriminate among classes.
- Does not perform detailed attribution analysis in the pixel space
  - Guided Backpropagation
    - Visualizes gradients in the image space
      - suppressing negative gradients while backpropagating through ReLU
  - Deconvolution
- $L^{c}_{\text{Grad−CAM}}$ upsampled to image resolution via bilinear interpolation
- Fuse Guided Backpropagation and Grad-CAM visualizations
  - element-wise multiplication

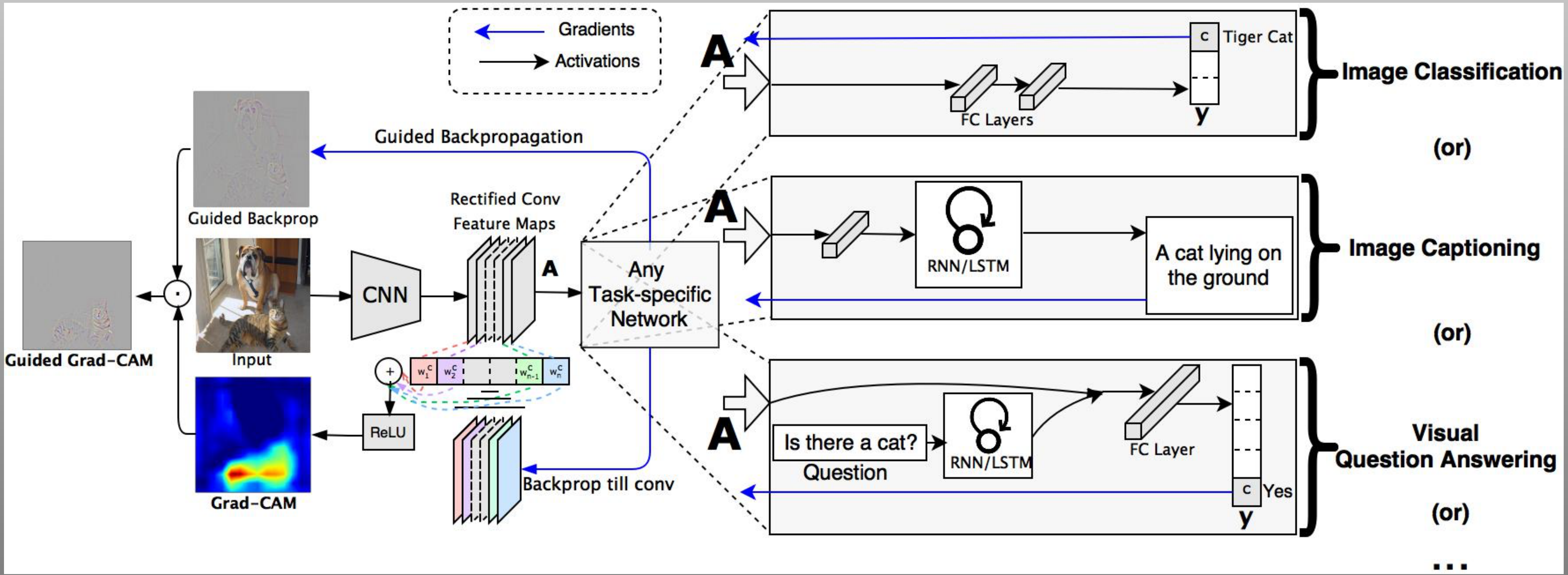


Fig. 2: Grad-CAM overview: Given an image and a class of interest (*e.g.*, 'tiger cat' or any other type of differentiable output) as input, we forward propagate the image through the CNN part of the model and then through task-specific computations to obtain a raw score for the category. The gradients 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.

Image reproduced under fair use from
https://arxiv.org/pdf/1610.02391.pdf

Grad-CAM generalizes CAM

- CAM (Class Activation Maps)

$$\underset{\text{Model score for class c}}{Y^c} = \sum_k \underbrace{w_k^c}_{\text{class feature weights}} \boxed{\overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{A_{ij}^k}_{\text{feature map}}} \longrightarrow \underset{\text{Global average pooled output}}{F^k}$$

Grad-CAM as generalization of CAM

Model score for class c

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

Global average pooled output

Grad-CAM as generalization of CAM

Model score for class c

$$Y^c = \sum_k \underbrace{w_k^c}_{\text{class feature weights}} \overbrace{\boxed{\frac{1}{Z}\sum_i \sum_j \underbrace{A_{ij}^k}_{\text{feature map}}}}^{\text{global average pooling}} \longrightarrow \underset{\text{Global average pooled output}}{F^k} \longrightarrow \frac{\partial F^k}{\partial A_{ij}^k} = \frac{1}{Z}$$

Grad-CAM as generalization of CAM

Model score for class c

$$Y^c = \sum_k \underbrace{w_k^c}_{\text{class feature weights}} \overbrace{\boxed{\frac{1}{Z}\sum_i \sum_j \underbrace{A_{ij}^k}_{\text{feature map}}}}^{\text{global average pooling}} \longrightarrow F^k \longrightarrow \frac{\partial F^k}{\partial A_{ij}^k} = \frac{1}{Z}$$

Global average pooled output

$$\frac{\partial Y^c}{\partial F^k} = w_k^c$$

$$\frac{\partial Y^c}{\partial F^k} = \frac{\frac{\partial Y^c}{\partial A_{ij}^k}}{\frac{\partial F^k}{\partial A_{ij}^k}}$$

Grad-CAM as generalization of CAM

Model score for class c

$$\begin{tikzcd}
{} & Y^c = \sum_k \underbrace{w_k^c}_{\text{class feature weights}} \boxed{\overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{A_{ij}^k}_{\text{feature map}}} \arrow[dl] \arrow[d] \arrow[r] & \underset{\text{Global average pooled output}}{F^k} \arrow[r] & \frac{\partial F^k}{\partial A_{ij}^k} = \frac{1}{Z} \arrow[d] \\
\frac{\partial Y^c}{\partial F^k} = w_k^c & \frac{\partial Y^c}{\partial F^k} = \frac{\frac{\partial Y^c}{\partial A_{ij}^k}}{\frac{\partial F^k}{\partial A_{ij}^k}} \arrow[rr] & & \frac{\partial Y^c}{\partial F^k} = \frac{\partial Y^c}{\partial A_{ij}^k} \cdot Z
\end{tikzcd}$$

Grad-CAM as generalization of CAM

Model score for class c

$$Y^c = \sum_k \underbrace{w_k^c}_{\text{class feature weights}} \overbrace{\boxed{\frac{1}{Z}\sum_i\sum_j \underbrace{A_{ij}^k}_{\text{feature map}}}}^{\text{global average pooling}} \longrightarrow F^k \longrightarrow \frac{\partial F^k}{\partial A_{ij}^k} = \frac{1}{Z}$$

Global average pooled output

$$\frac{\partial Y^c}{\partial F^k} = w_k^c$$

$$\frac{\partial Y^c}{\partial F^k} = \frac{\frac{\partial Y^c}{\partial A_{ij}^k}}{\frac{\partial F^k}{\partial A_{ij}^k}} \longrightarrow \frac{\partial Y^c}{\partial F^k} = \frac{\partial Y^c}{\partial A_{ij}^k} \cdot Z$$

$$w_k^c = Z \cdot \frac{\partial Y^c}{\partial A_{ij}^k}$$

Grad-CAM as generalization of CAM

Model score for class c

$$Y^c = \sum_k \underbrace{w_k^c}_{\text{class feature weights}} \boxed{\overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{A_{ij}^k}_{\text{feature map}}} \longrightarrow F^k \longrightarrow \frac{\partial F^k}{\partial A_{ij}^k} = \frac{1}{Z}$$

Global average pooled output

$$\frac{\partial Y^c}{\partial F^k} = w_k^c$$

$$\frac{\partial Y^c}{\partial F^k} = \frac{\frac{\partial Y^c}{\partial A_{ij}^k}}{\frac{\partial F^k}{\partial A_{ij}^k}} \longrightarrow \frac{\partial Y^c}{\partial F^k} = \frac{\partial Y^c}{\partial A_{ij}^k} \cdot Z$$

$$w_k^c = Z \cdot \frac{\partial Y^c}{\partial A_{ij}^k}$$

$$\sum_i \sum_j w_k^c = \sum_i \sum_j Z \cdot \frac{\partial Y^c}{\partial A_{ij}^k} \longrightarrow w_k^c = \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k}$$

CAM Grad-CAM

Counterfactual Explanations

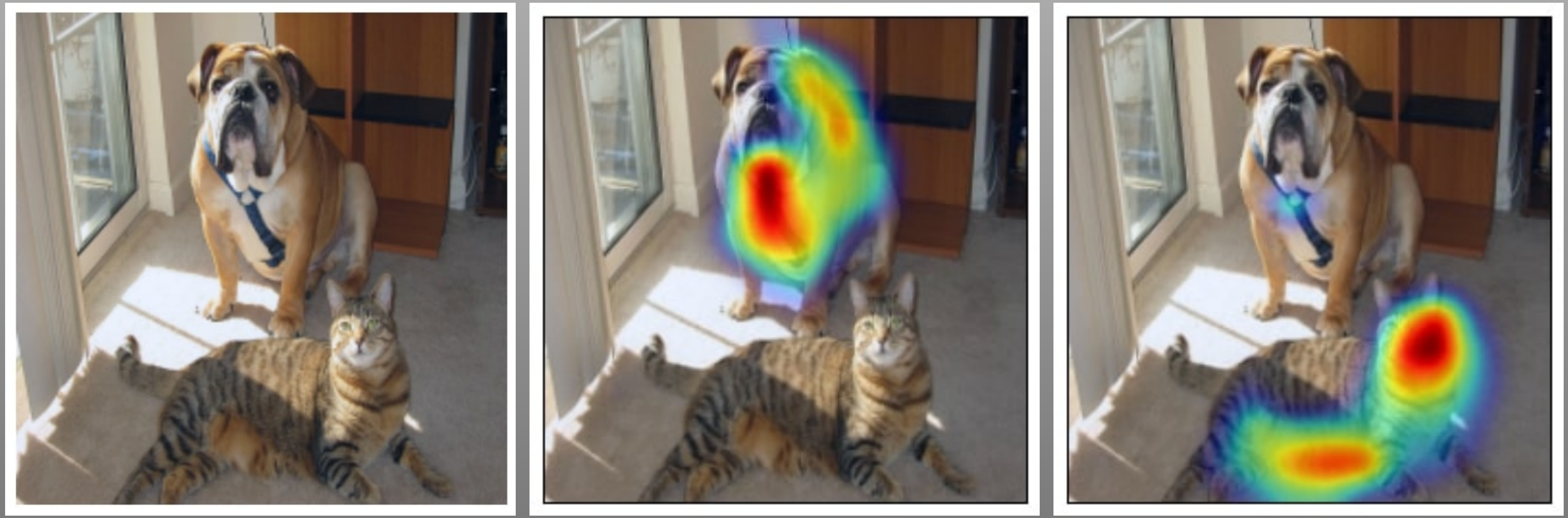
- Counterfactuals: Area that causes change in classification.
  - Removing these should enhance model confidence in prediction.
- How?
  - Negate the gradient in computing the neural importance weight

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

- Weighted sum of activation maps, A, with weights α_k^c, + RELU

Counterfactual Explanations - II

- Counterfactuals: Area that causes change in classification.
  - Removing these should enhance model confidence in prediction.



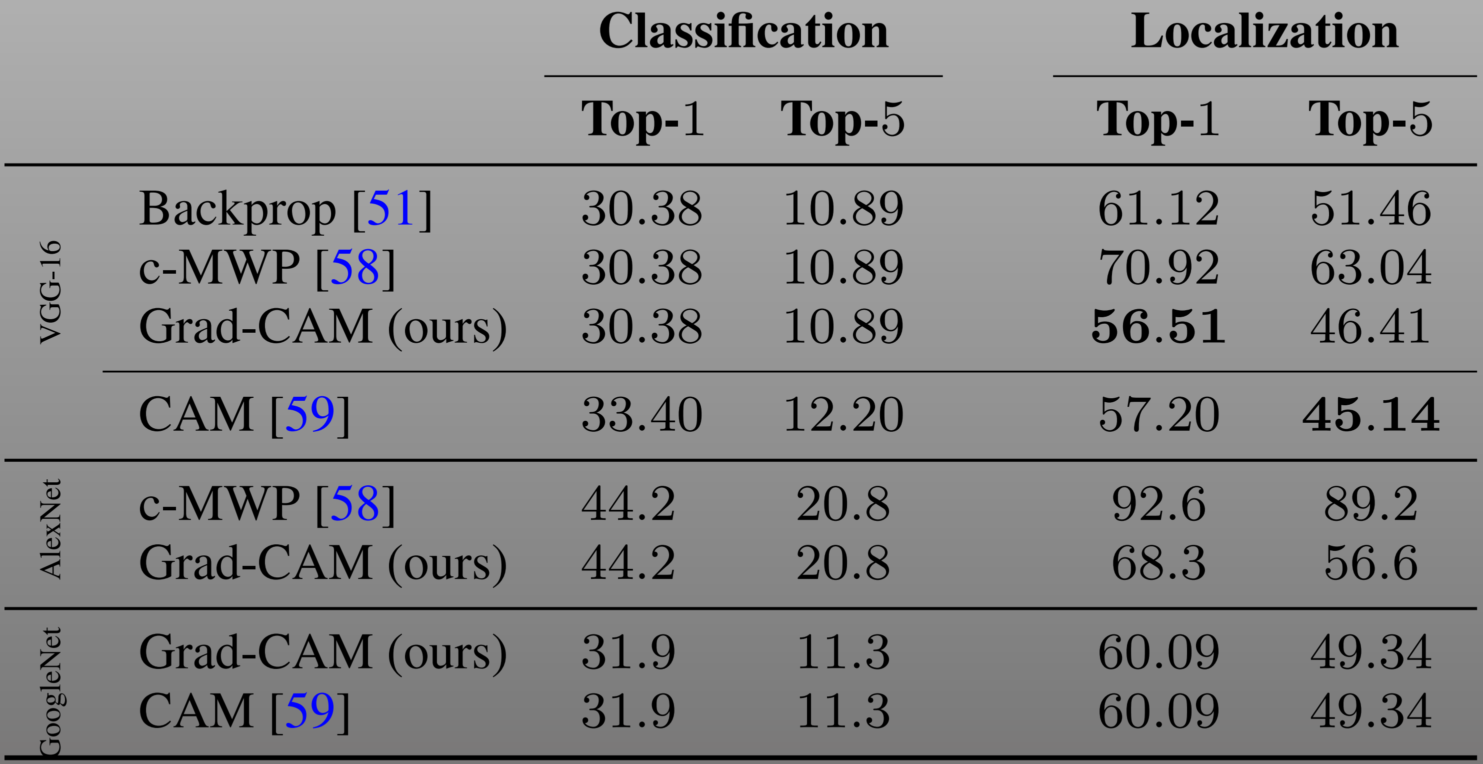
(a) Original Image (b) Cat Counterfactual exp (c) Dog Counterfactual exp

Image reproduced under fair use from
https://arxiv.org/pdf/1610.02391.pdf

Evaluations - I

Weakly-supervised Localization on ImageNet

- Given an image,
  - obtain class predictions
  - generate Grad-CAM
  - binarize them with a threshold of 15%
  - results in connected components of pixels
  - draw a bounding box for the largest component/segment.



		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

Image reproduced under fair use from https://arxiv.org/pdf/1610.02391.pdf

Evaluations - II

- **Weakly-supervised Segmentation**
  - Assign each pixel an object label/class.

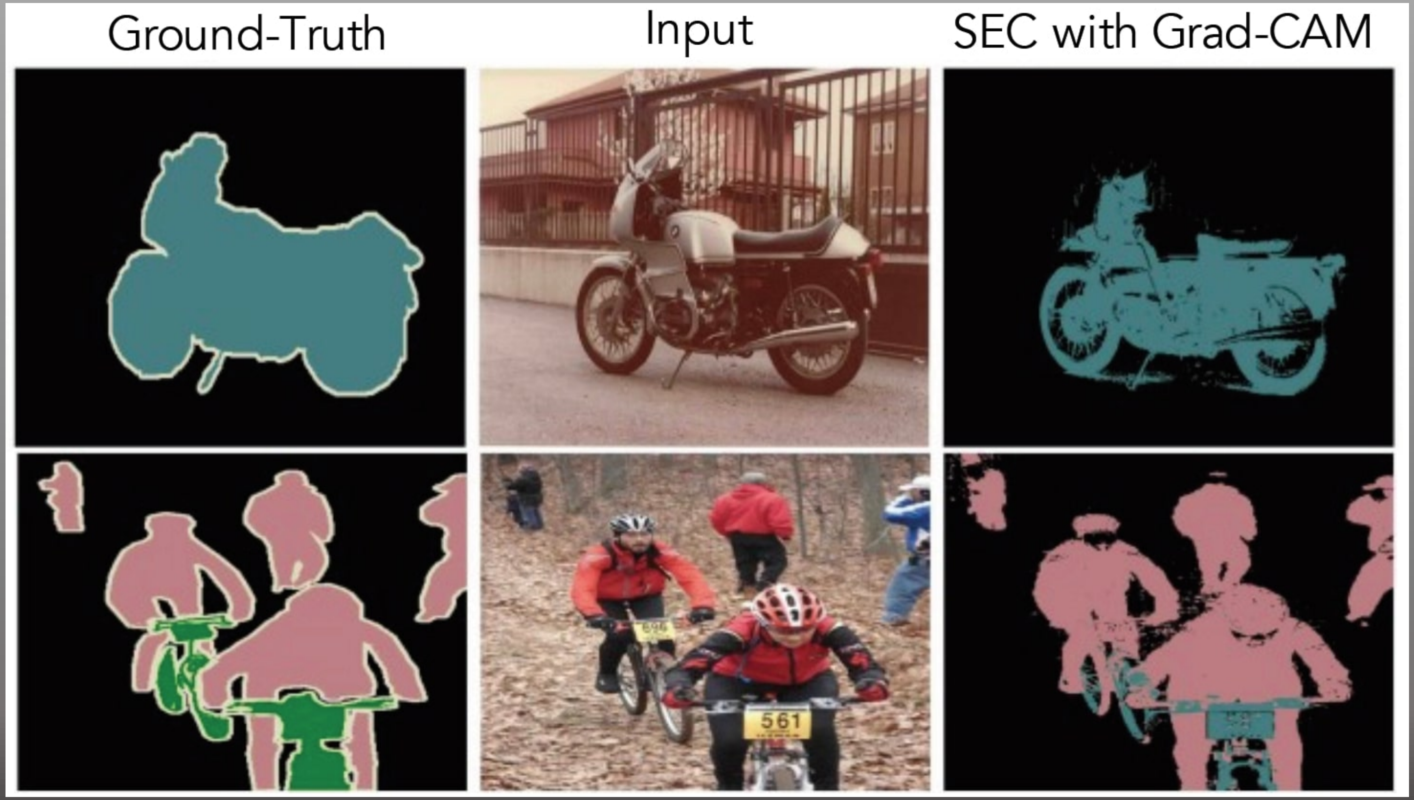


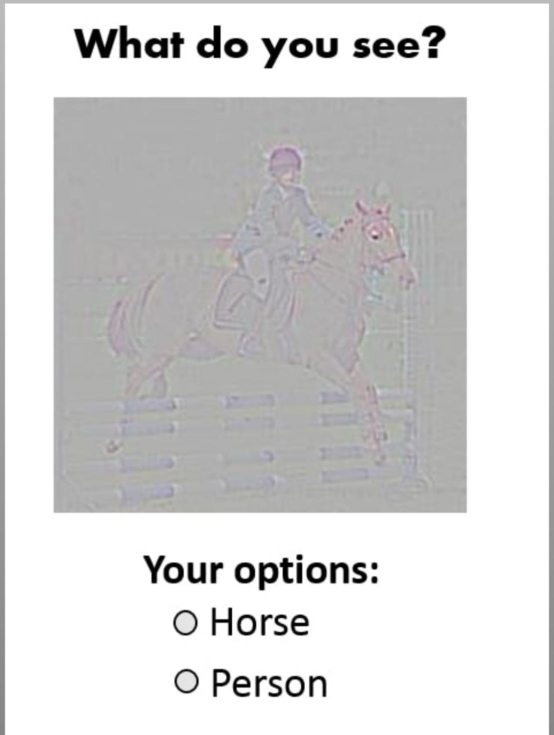
Image reproduced under fair use from https://arxiv.org/pdf/1610.02391.pdf

Evaluation - III

- VGG-16 and AlexNet on PASCAL VOC 2007 data
- Human study
- Is Grad-CAM more class discriminative
  - Compared to earlier approaches
- Does grad-CAM lead a user to trust models
  - *Appropriately*

Evaluating Class Discrimination

- Select images with 2 annotated categories
- Create visualizations for each one of them.
  - Deconvolution
  - Guided Backpropagation
  - Grad-CAM
- Query 43 humans on Amazon Mechanical Turk (AMT)
  - 4 visualizations for 90 image-category pairs
  - 9 ratings
- Which of the two object categories is depicted in the image?



What do you see?

Your options:

○ Horse

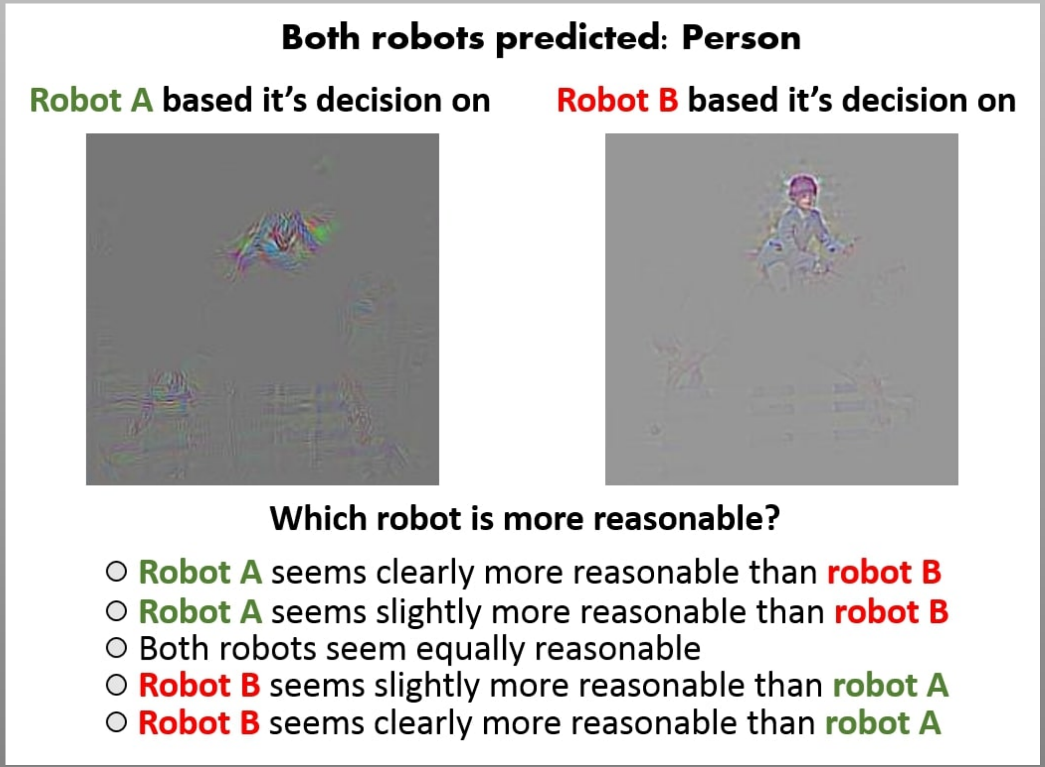
○ Person

Method	Human Classification Accuracy
Guided Backpropagation	44.44
Guided Grad-CAM	61.23

Image reproduced under fair use from https://arxiv.org/pdf/1610.02391.pdf

Evaluating Trust

- Given two attributions, which one is more trustworthy?
- AlexNet and VGG-16
  - VGG-16 mean Average Precision 79.09
  - AlexNet mAP 69.20 on PASCAL.
- Focus only on images where both models were correct
- 54 AMT workers rate reliability
  - clearly more/less reliable (+/-2),
  - slightly more/less reliable (+/-1),
  - equally reliable (0).



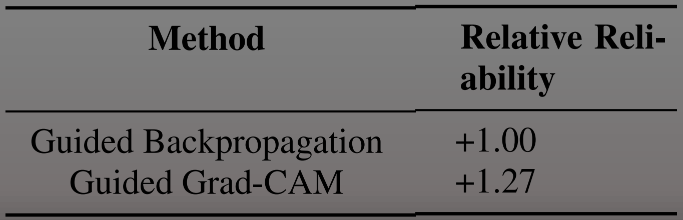
Both robots predicted: Person

Robot A based it's decision on **Robot B based it's decision on**

Which robot is more reasonable?

○ **Robot A** seems clearly more reasonable than **robot B**
○ **Robot A** seems slightly more reasonable than **robot B**
○ Both robots seem equally reasonable
○ **Robot B** seems slightly more reasonable than **robot A**
○ **Robot B** seems clearly more reasonable than **robot A**

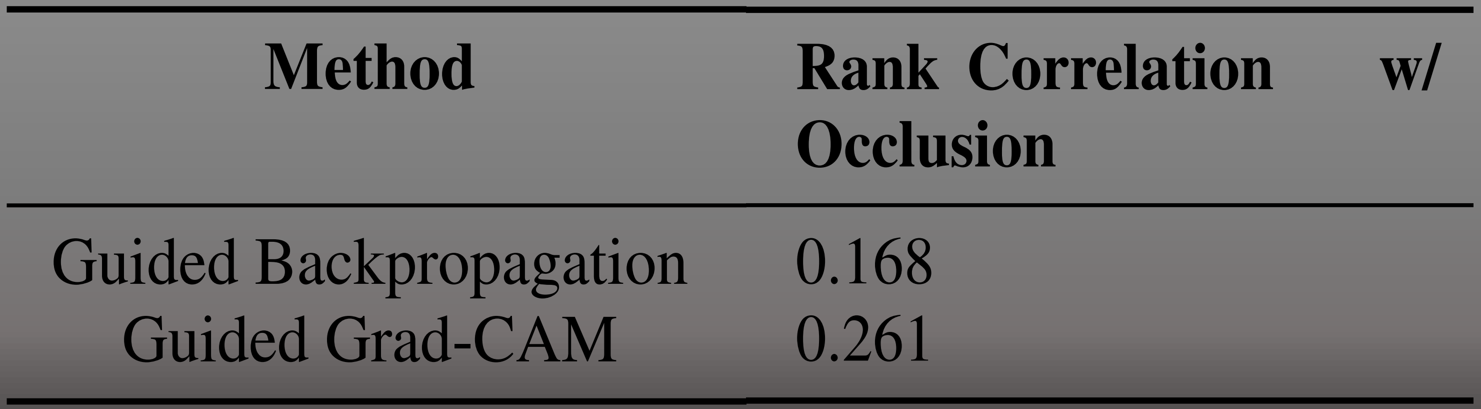
Image reproduced under fair use from https://arxiv.org/pdf/1610.02391.pdf



Method	Relative Reliability
Guided Backpropagation	+1.00
Guided Grad-CAM	+1.27

Faithfulness

- Select a reference explanation with high "faithfulness" near the input
  - image occlusion
- X = Patches that affect CNN score
- Y = Patches that have high Grad-CAM and Guided Grad-CAM
- X and Y are correlated: 0.261
- 2510 images from the PASCAL 2007 validation set.



Method	Rank Correlation w/ Occlusion
Guided Backpropagation	0.168
Guided Grad-CAM	0.261

Image reproduced under fair use from https://arxiv.org/pdf/1610.02391.pdf

Conclusions

- New class-discriminative localization technique for any CNN
  - Gradient-weighted Class Activation Mapping (Grad-CAM)
- Grad-CAM combined with high- resolution visualization
- Outperform for interpretability and faithfulness
- Human studies
  - discriminate more accurately,
  - better expose trustworthiness
- Future work
  - reinforcement learning
  - natural language processing
  - video applications