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

Ramprasaath R. Selvaraju et al.

Published at ICCV 2017

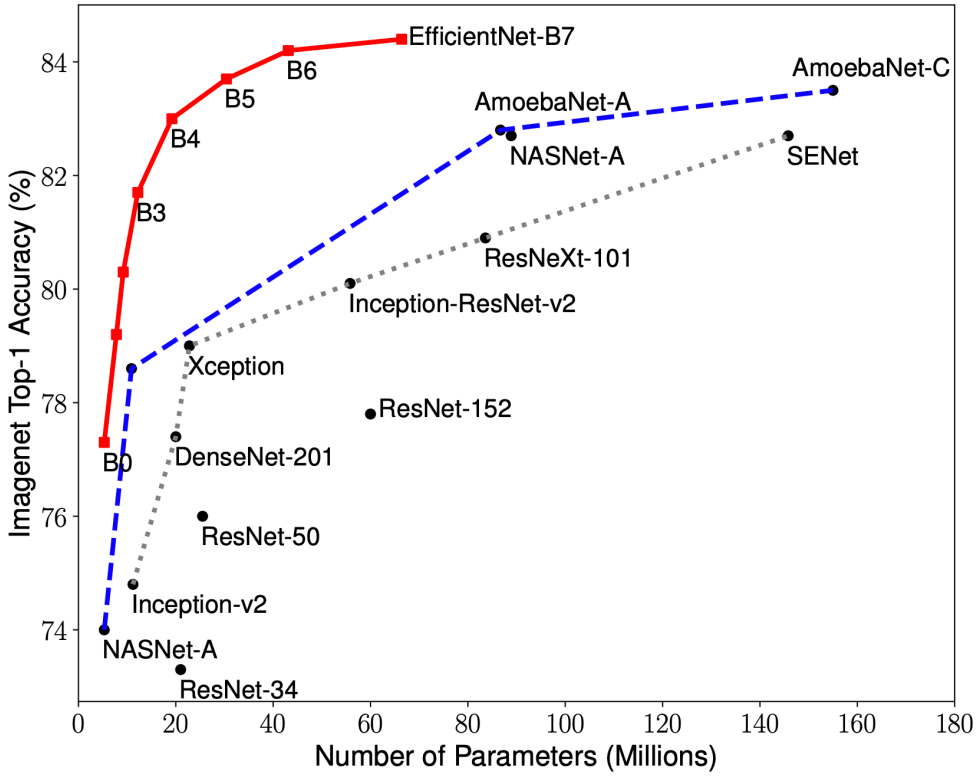
CUAI 5th

Hayun Lee

22.03.22

Background

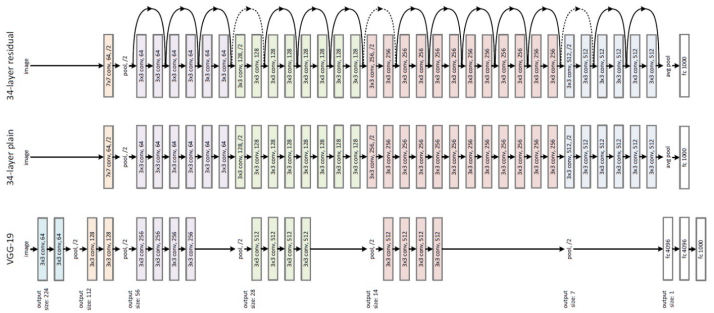
- Black Box



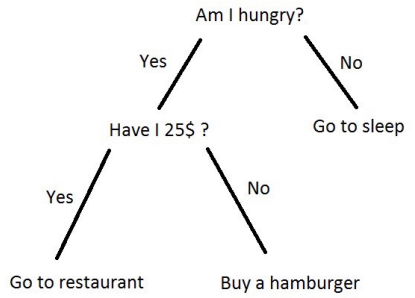
Background

- We must ‘transparent’ model that have the ability to explain why they predict what they predict.
  - AI < Human : To identify the failure modes
  - AI == Human : To establish appropriate trust and confidence in users
  - AI > Human : Machine teaching a human about now to make better decisions

- Accuracy VS Interpretability



Deep & Accurate



Shallow & Interpretable

Background

- What makes a good visual explanation?
  - Class Discriminative
  - High-Resolution

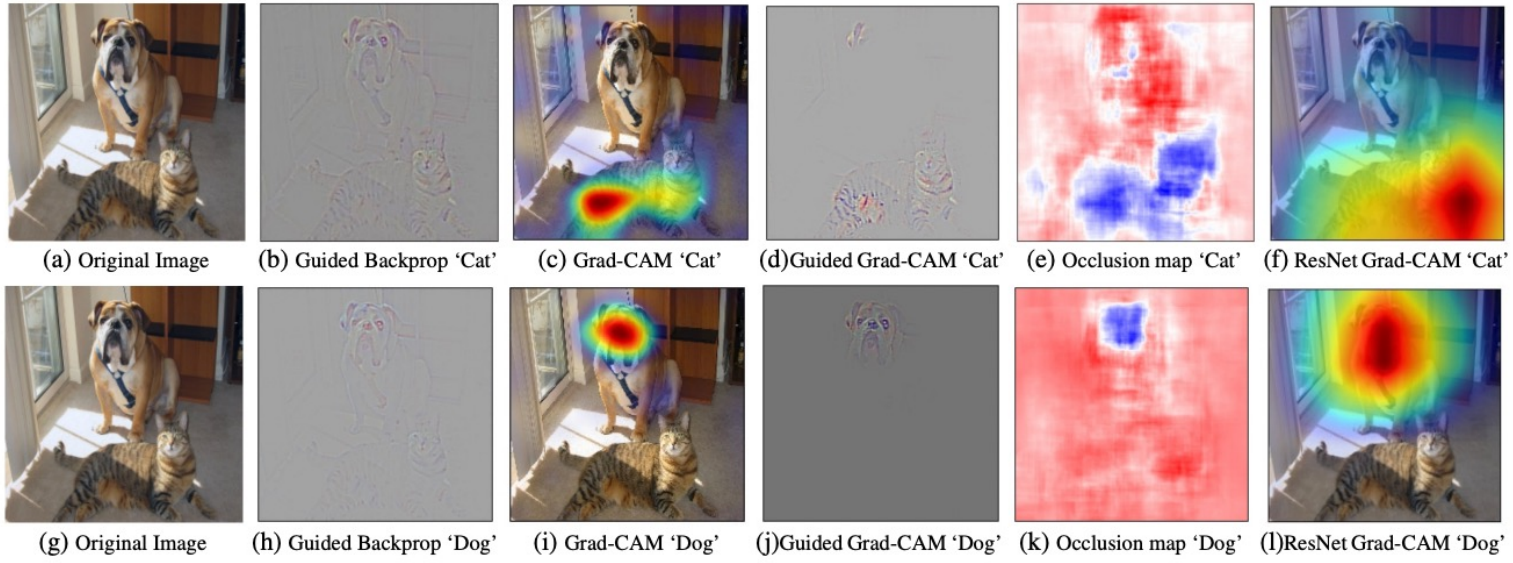


Fig. 1: (a) Original image with a cat and a dog. (b-f) Support for the cat category according to various visualizations for VGG-16 and ResNet. (b) Guided Backpropagation [53]: highlights all contributing features. (c, f) Grad-CAM (Ours): localizes class-discriminative regions, (d) Combining (b) and (c) 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.

Related Work

- Visualizing CNNs

(b) Guided Backprop 'Cat'

(h) Guided Backprop 'Dog'

Related Work

- Assessing Model Trust
  - ["Why Should I Trust You?": Explaining the Predictions of Any Classifier, 2016]
    - About trust in models
    - Grad-CAM was inspired by this paper

- Aligning Gradient-based Importances
  - Uses the gradient-based neuron importances and maps it to class-specific domain knowledge from humans in order to learn classifiers for novel classes.

Related Work

- Weakly-supervised Localization
  - CAM

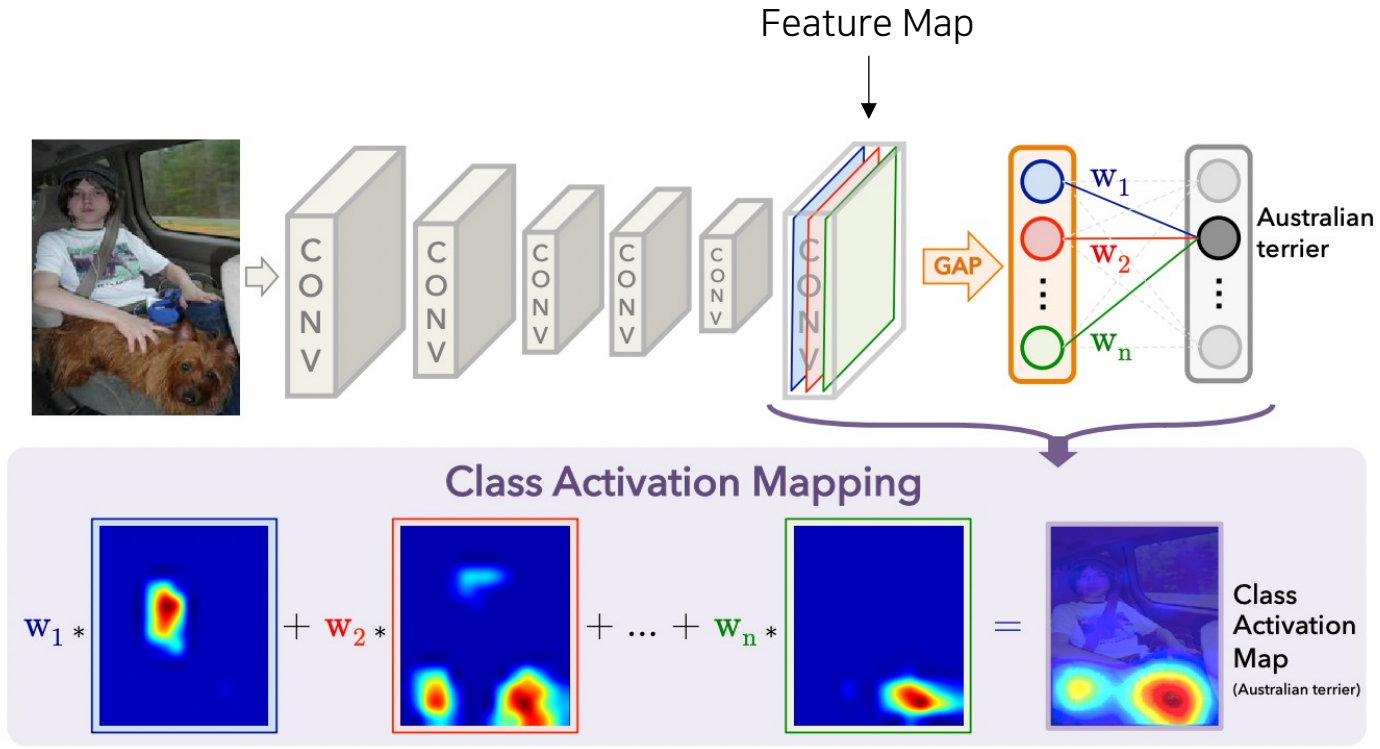
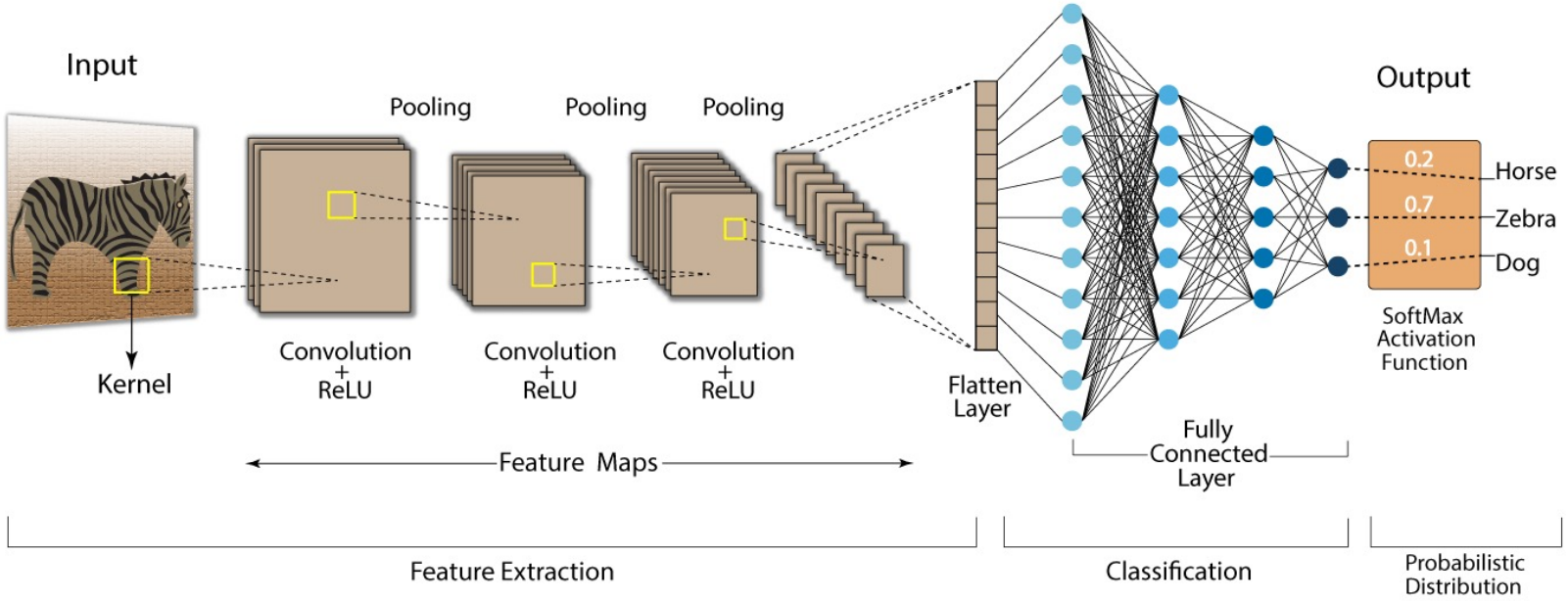


Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

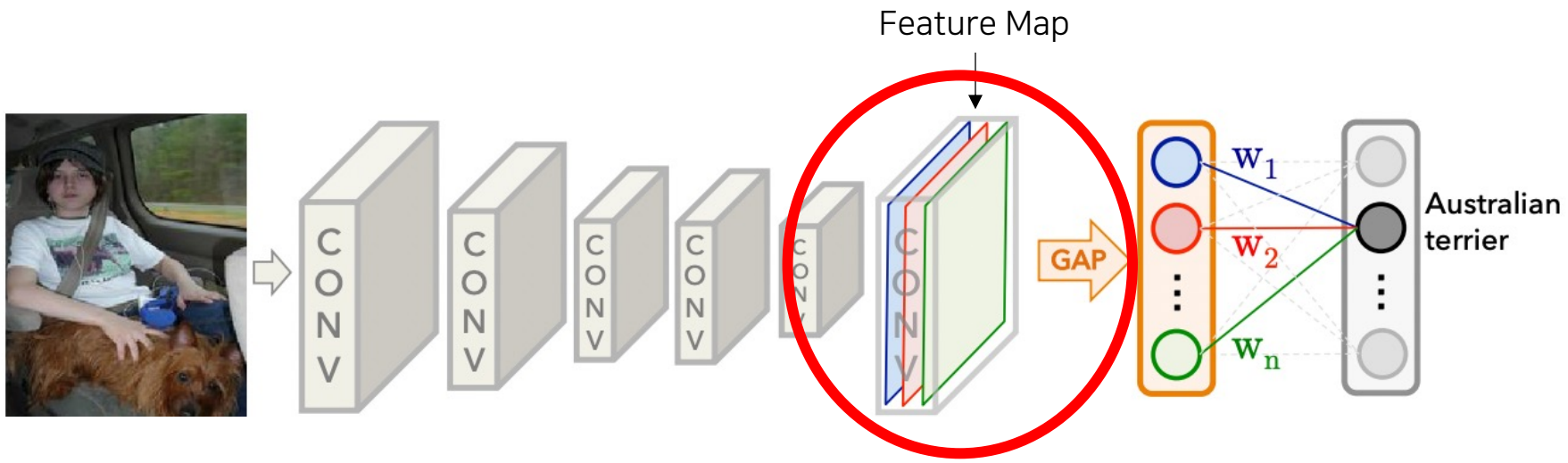
Related Work: CAM

- Convolutional Neural Network(CNN)
  - Classification Model
  - conv & pooling -> image feature extraction
  - Fully connected layer



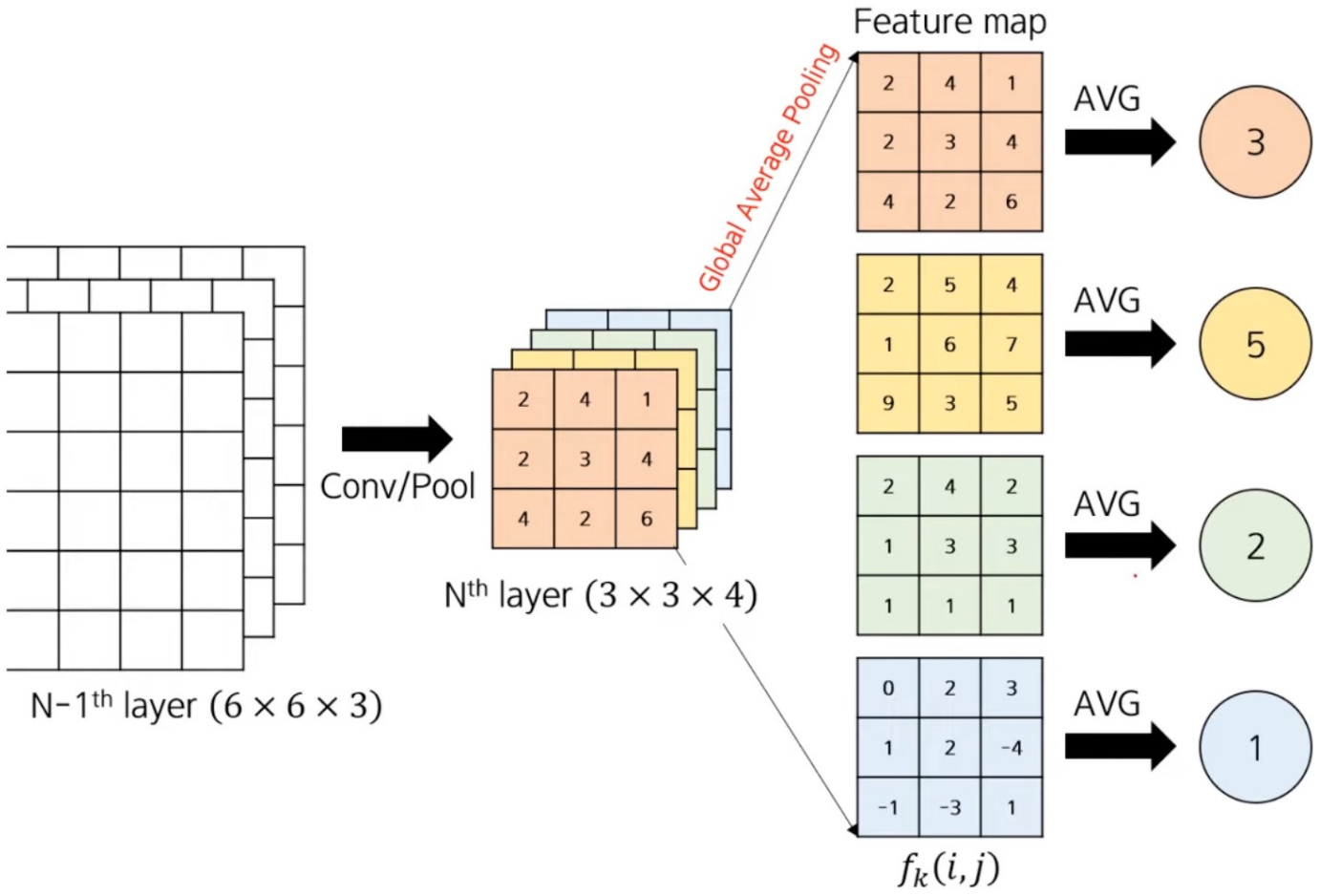
Related Work: CAM

- CNN + CAM Architecture
  - Convolutional layer -> global average pooling
  - GAP?



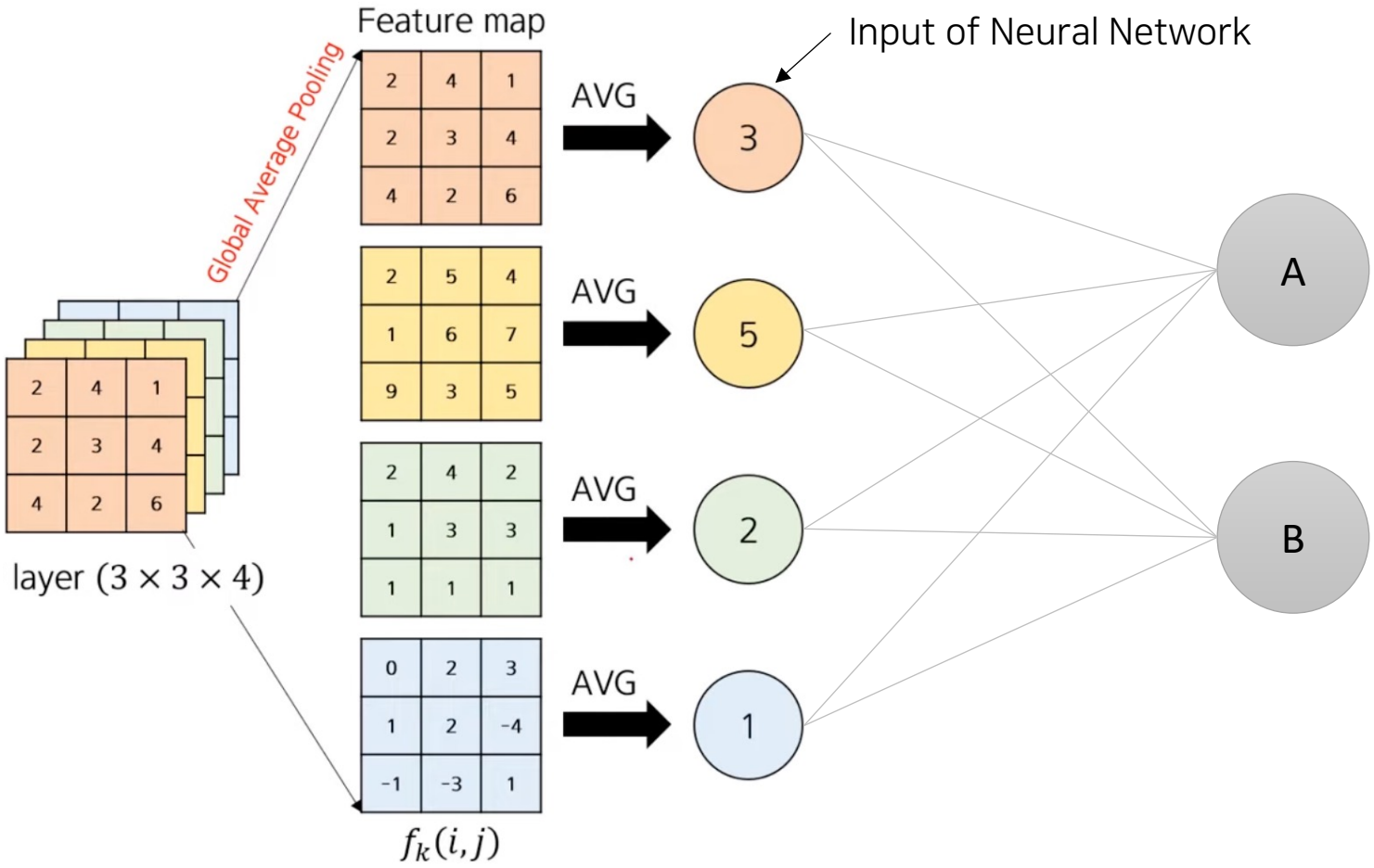
Related Work: CAM

- Global Average Pooling : average of feature map's weight



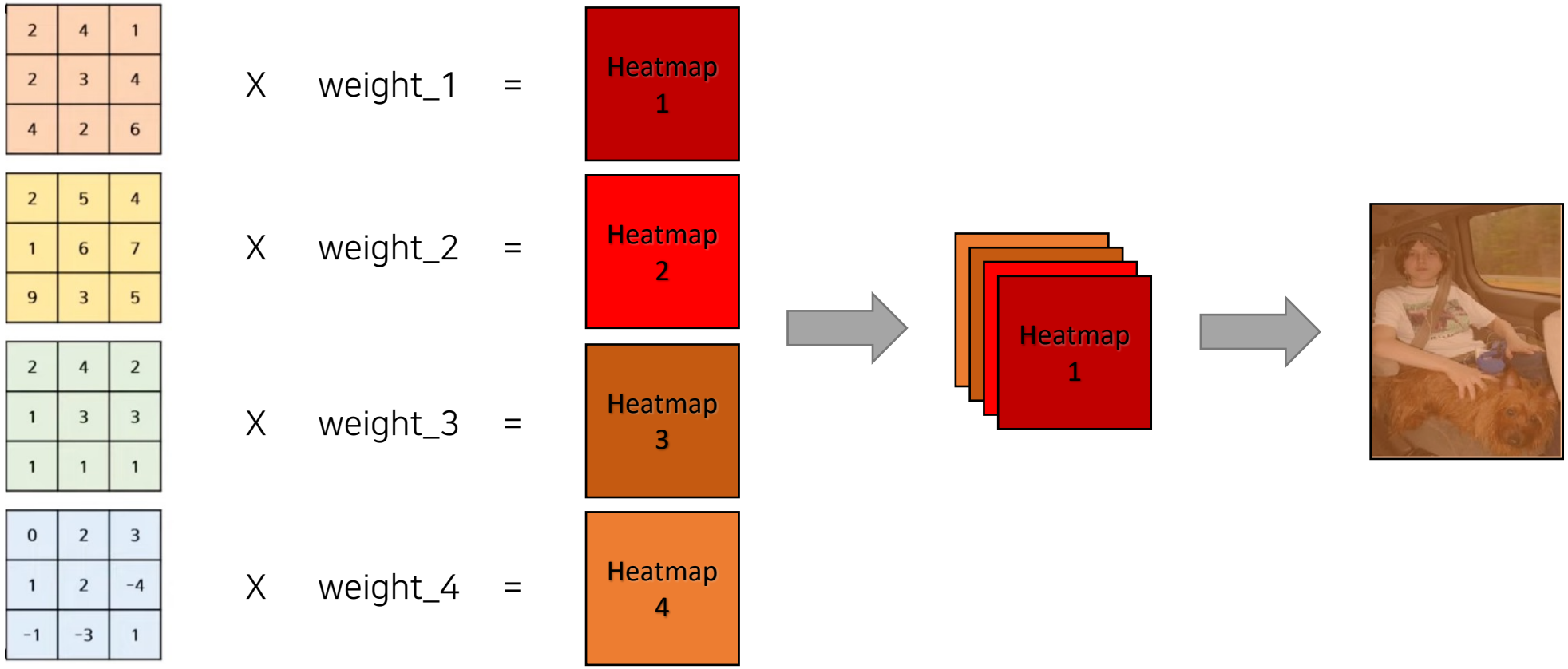
Related Work: CAM

- Global Average Pooling : average of feature map's weight



Related Work: CAM

- Heatmap for each feature map



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.

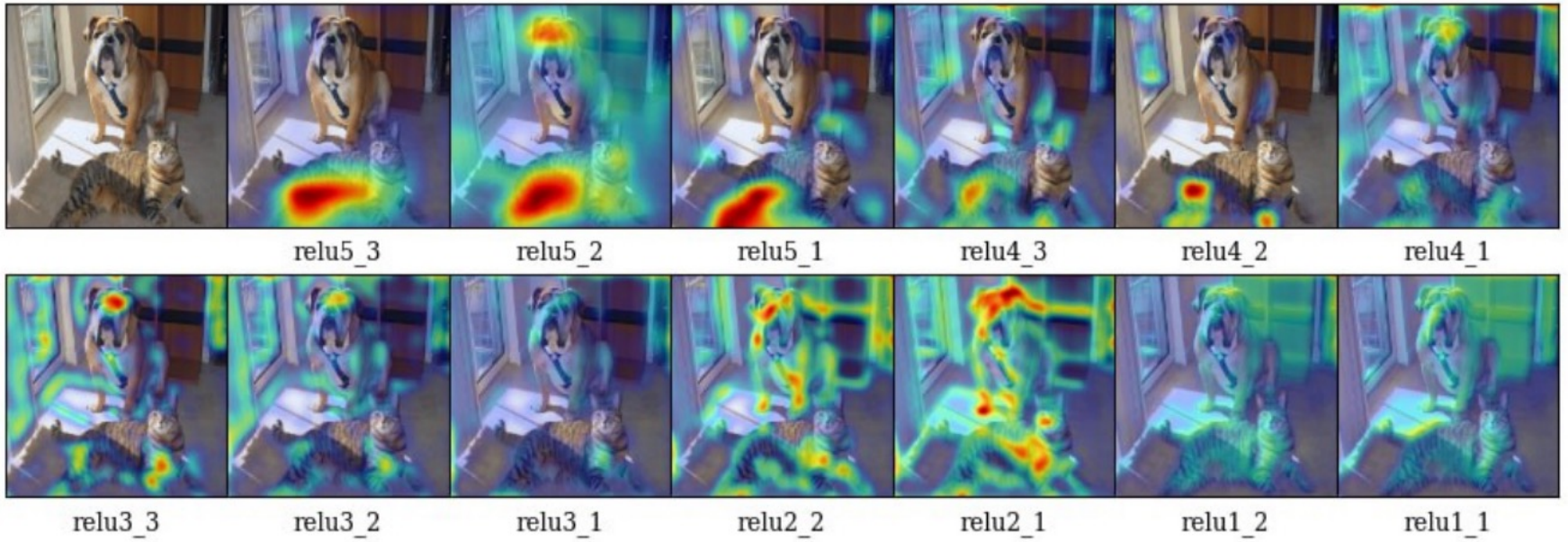


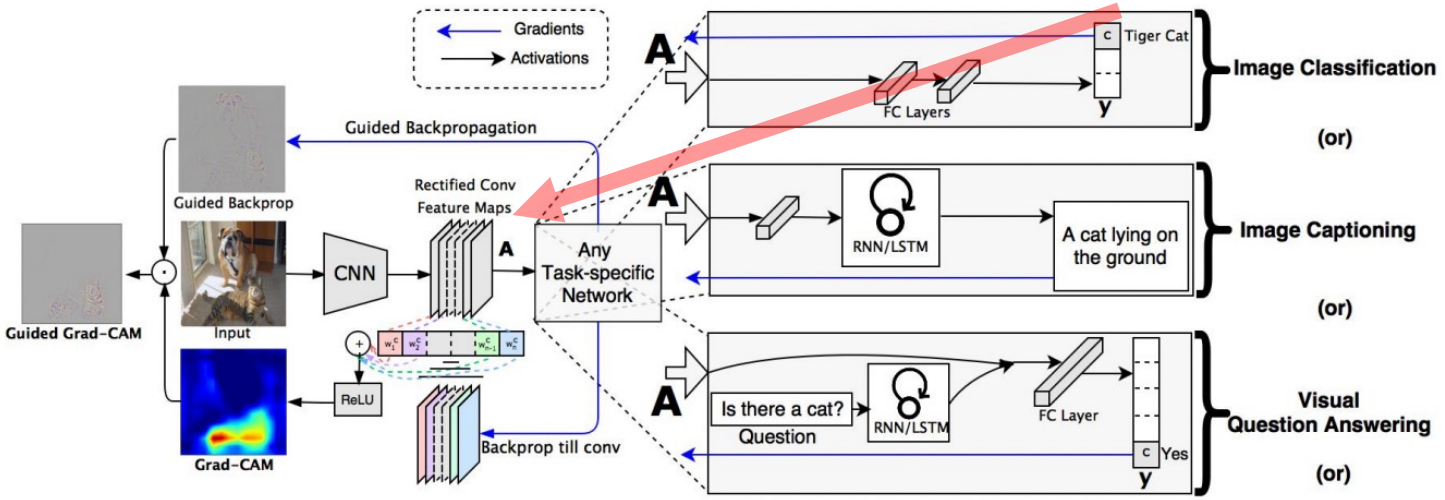
Fig. 13: Grad-CAM at different convolutional layers for the 'tiger cat' class. This figure analyzes how localizations change qualitatively as we perform Grad-CAM with respect to different feature maps in a CNN (VGG16 [52]). We find that the best looking visualizations are often obtained after the deepest convolutional layer in the network, and localizations get progressively worse at shallower layers. This is consistent with our intuition described in Section 3 of main paper, that deeper convolutional layer capture more semantic concepts.

Grad-CAM: Architecture

① Gradient of y^c (score for class c) w.r.t feature map activation A^k $\frac{\partial yc}{\partial A^k}$

- Which feature map influenced the model's prediction and how much?

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

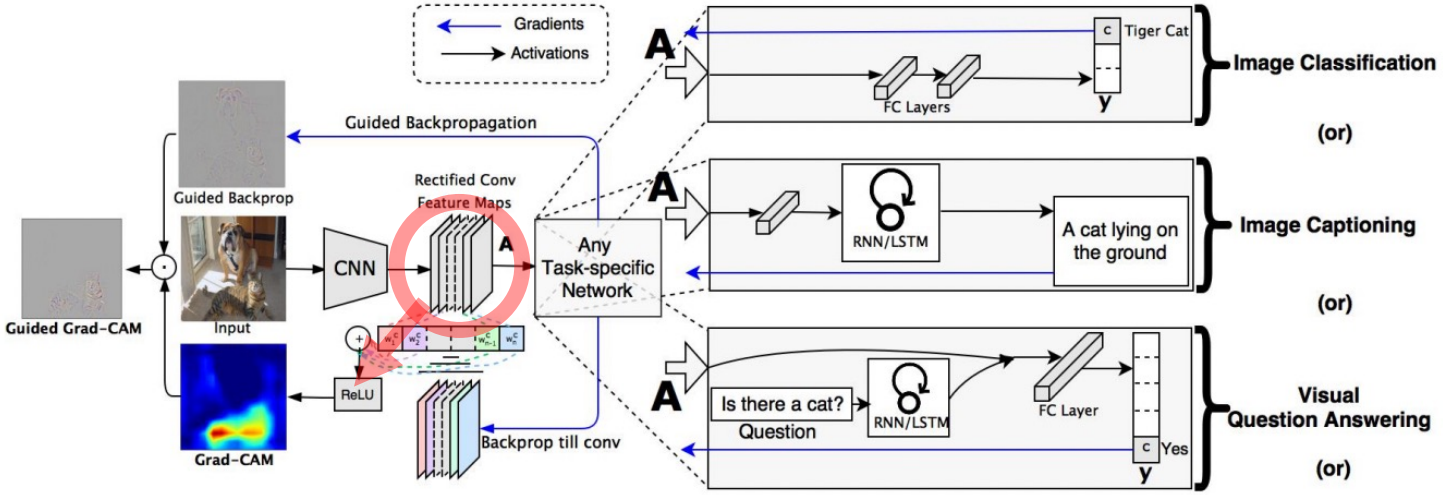


Grad-CAM: Architecture

② Weighted combination

- α : feature map's importance
- ReLU : Only positive effect

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



Grad-CAM generalizes CAM

CAM

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

$$F^k = \frac{1}{Z}\sum_i\sum_j A_{ij}^k \quad (4)$$

$$Y^c = \sum_k w_k^c \cdot F^k \quad (5)$$

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

$$\frac{\partial F^k}{\partial A_{ij}^k} = \frac{1}{Z}$$

$$\frac{\partial Y^c}{\partial F^k} = \frac{\partial Y^c}{\partial A_{ij}^k} \cdot Z \quad (7)$$

$$w_k^c = Z \cdot \frac{\partial Y^c}{\partial A_{ij}^k} \quad (8)$$

Each pixel

$$\sum_i\sum_j w_k^c = \sum_i\sum_j Z \cdot \frac{\partial Y^c}{\partial A_{ij}^k} \quad (9)$$

$$Z w_k^c = Z\sum_i\sum_j \frac{\partial Y^c}{\partial A_{ij}^k} \quad (10)$$

$$w_k^c = \sum_i\sum_j \frac{\partial Y^c}{\partial A_{ij}^k} \quad (11)$$

Grad-CAM generalizes CAM

- Grad-CAM
  - weights are calculated based on gradients, **visualization is possible without GAP**

Grad-CAM

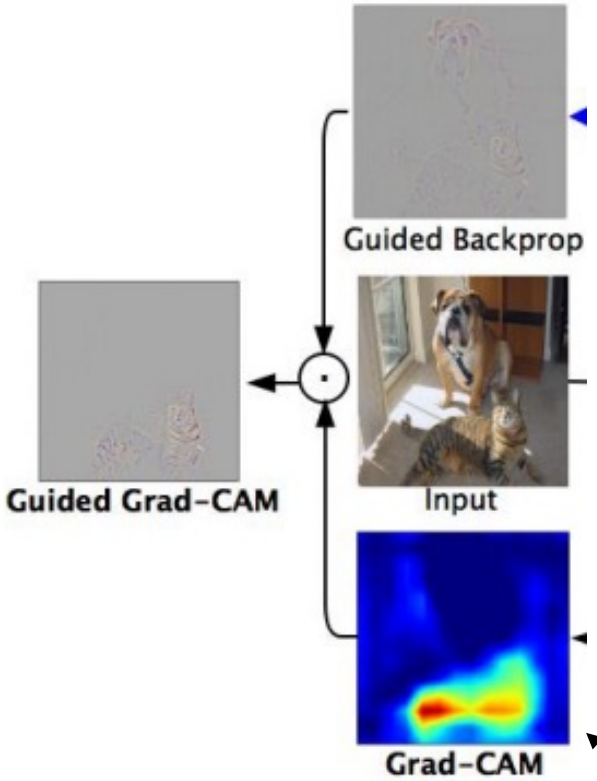
$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

CAM

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

Guided Grad-CAM

- element-wise product



(d)Guided Grad-CAM 'Cat'

(j)Guided Grad-CAM 'Dog'

Evaluating Localization Ability of Grad-CAM

- Weakly-supervised Localization
  - Class predictions from 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. *(weakly-supervised with no labeled bbox)*

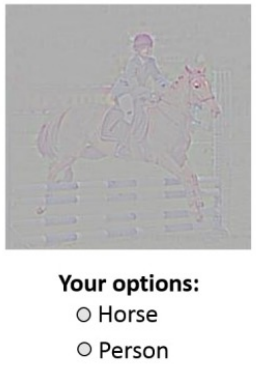
		Classification		Localization	
		Top-1	**Top-5**	**Top-1**	**Top-5**
VGG-16	Backprop [51]	30.38	10.89	61.12	51.46
	c-MWP [58]	30.38	10.89	70.92	63.04
	Grad-CAM (ours)	30.38	10.89	**56.51**	46.41
	CAM [59]	33.40	12.20	57.20	**45.14**
AlexNet	c-MWP [58]	44.2	20.8	92.6	89.2
	Grad-CAM (ours)	44.2	20.8	68.3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

Table 1: Classification and localization error % on ILSVRC-15 val (lower is better) for VGG-16, AlexNet and GoogleNet. We see that Grad-CAM achieves superior localization errors without compromising on classification performance.

Evaluating Visualizations

- Class Discrimination
  - 43 workers on Amazon Mechanical Turk(AMT)
  - “Which of the two object categories is depicted in the image?”
  - 90 images

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



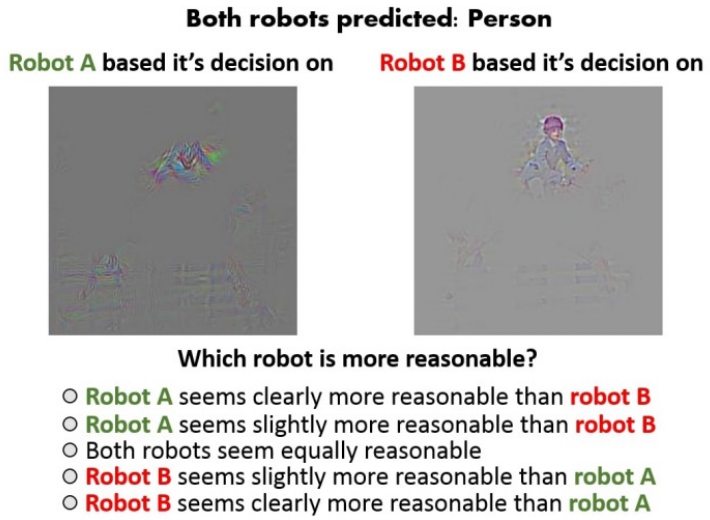
(b) AMT interface for evaluating the class-discriminative property

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

Table 2: Quantitative Visualization Evaluation. Guided Grad-CAM enables humans to differentiate between visualizations of different classes (Human Classification Accuracy) and pick more reliable models (Relative Reliability). It also accurately reflects the behavior of the model (Rank Correlation w/ Occlusion).

Evaluating Visualizations

- Evaluating Trust
  - 54 workers on AMT
  - VGG-16, AlexNet
  - -2,-1, 0, 1, 2



(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

Table 2: Quantitative Visualization Evaluation. Guided Grad-CAM enables humans to differentiate between visualizations of different classes (Human Classification Accuracy) and pick more reliable models (Relative Reliability). It also accurately reflects the behavior of the model (Rank Correlation w/ Occlusion).

Diagnosing Image Classification

- Analyzing failure models for VGG-16

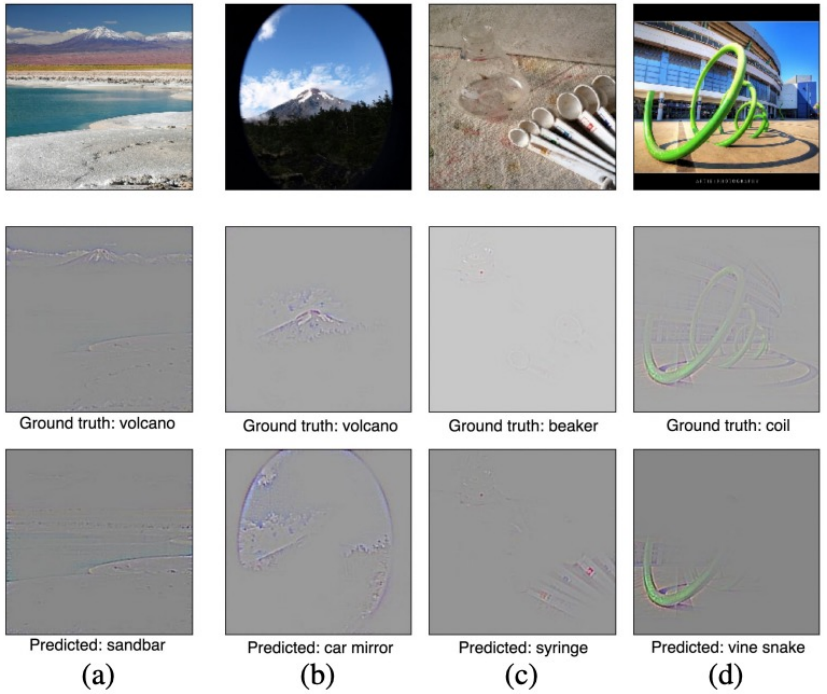


Fig. 6: In these cases the model (VGG-16) failed to predict the correct class in its top 1 (a and d) and top 5 (b and c) predictions. Humans would find it hard to explain some of these predictions without looking at the visualization for the predicted class. But with Grad-CAM, these mistakes seem justifiable.

Diagnosing Image Classification

- Effects of adversarial noise on VGG-16

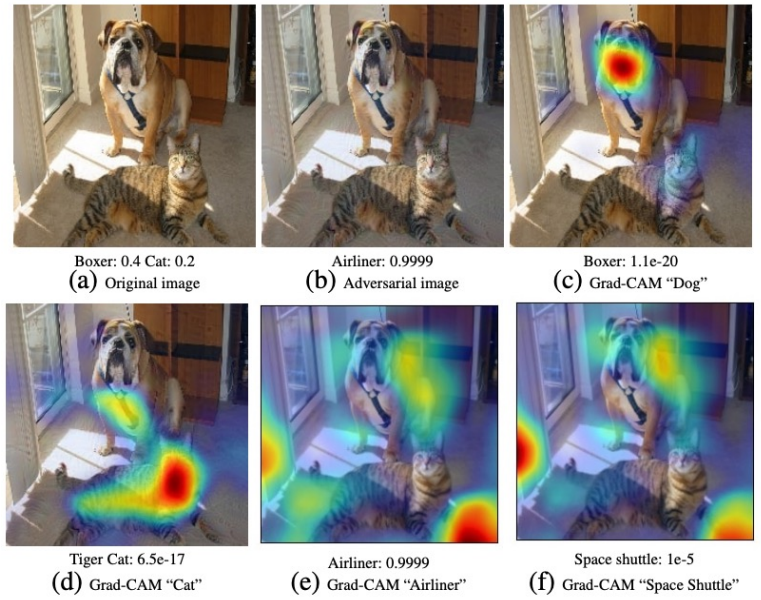


Fig. 7: (a-b) Original image and the generated adversarial image for category "airliner". (c-d) Grad-CAM visualizations for the original categories "tiger cat" and "boxer (dog)" along with their confidence. Despite the network being completely fooled into predicting the dominant category label of "airliner" with high confidence (>0.9999), Grad-CAM can localize the original categories accurately. (e-f) Grad-CAM for the top-2 predicted classes "airliner" and "space shuttle" seems to highlight the background.

Diagnosing Image Classification

- Identifying bias in dataset

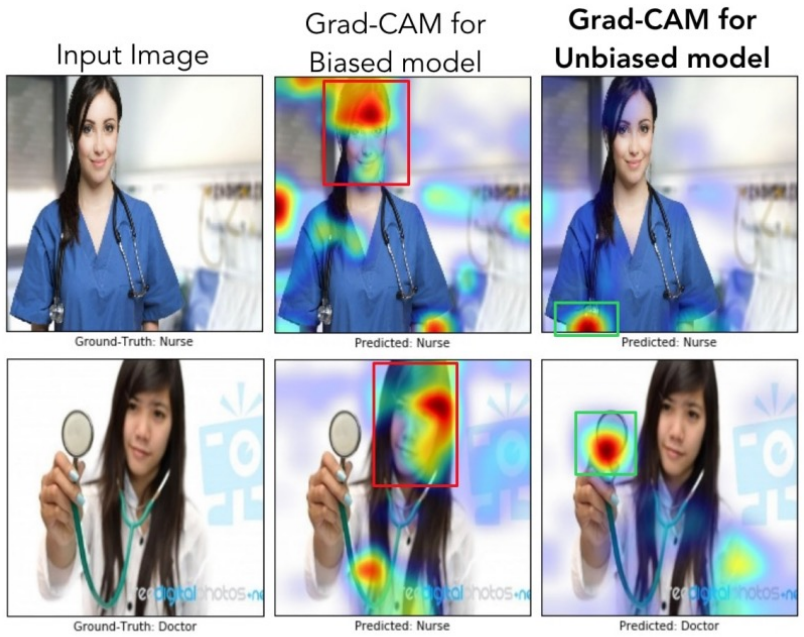


Fig. 8: In the first row, we can see that even though both models made the right decision, the biased model (model1) was looking at the face of the person to decide if the person was a nurse, whereas the unbiased model was looking at the short sleeves to make the decision. For the example image in the second row, the biased model made the wrong prediction (misclassifying a doctor as a nurse) by looking at the face and the hairstyle, whereas the unbiased model made the right prediction looking at the white coat, and the stethoscope.

Grad-CAM for Image Captioning and VQA

- Image captioning
  - Build the Grad-CAM on top of the publicly available Neuraltalk2

(a) Image captioning explanations

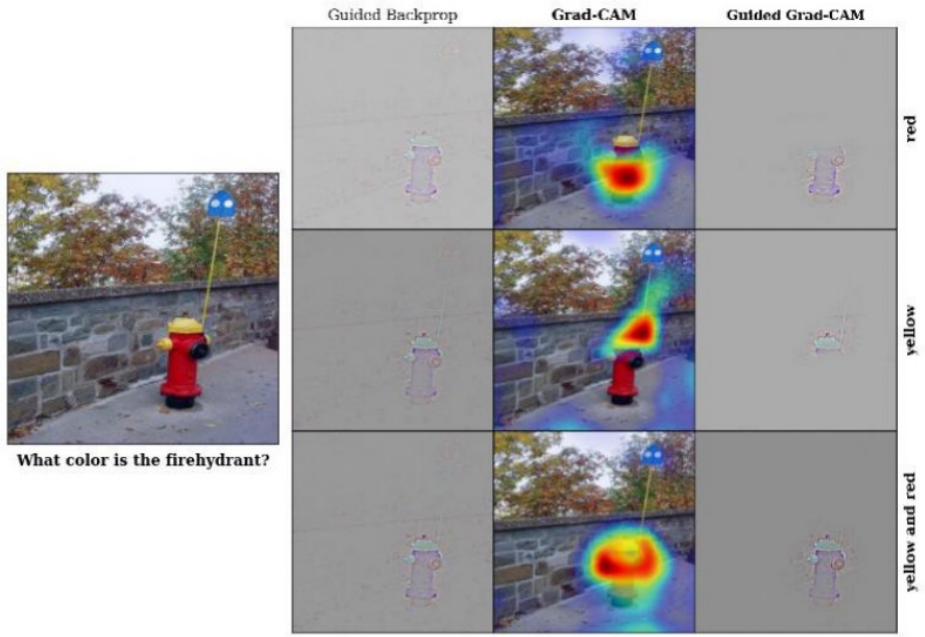
Grad-CAM for Image Captioning and VQA

- Dense captioning
  - Fully convolutional localization network + LSTM-based language model

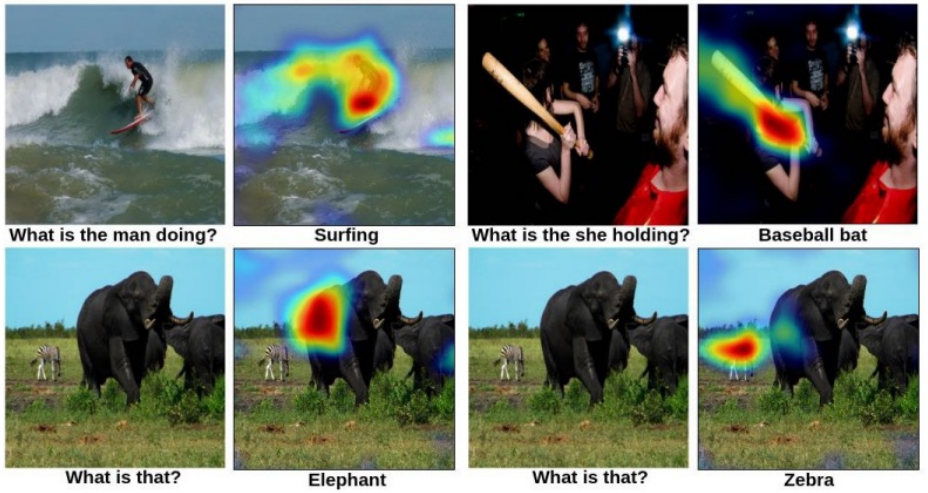
(b) Comparison to DenseCap

Grad-CAM for Image Captioning and VQA

- VQA(Visual Question Answering) explained by Grad-CAM
  - Image representation + question representation -> 1000-way(answer space, classification)



(a) Visualizing VQA model from [38]



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

Conclusion

- 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 explanations.
- Our visualizations outperform existing approaches on both axes – interpretability and faithfulness to original model.
- 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.
- 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.