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

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra

Presented by M.Rodin

April 30, 2020

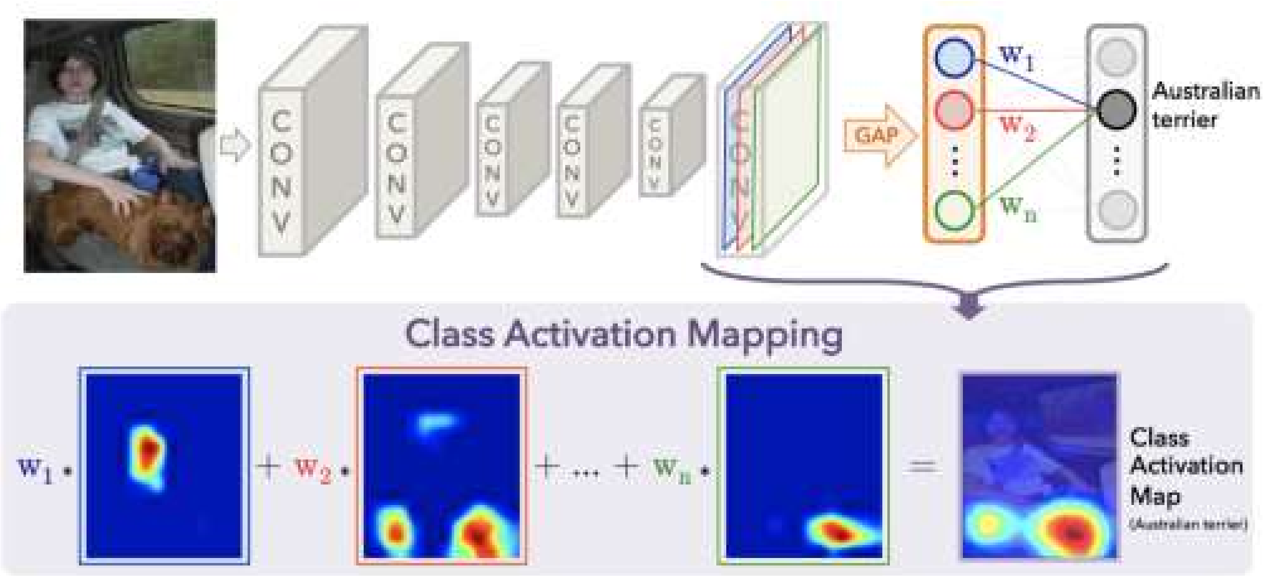
Outline

1. Introduction
    - Problem formulation
    - CAM: Class Activation Mapping
2. GradCAM architecture
3. Evaluations
    - Evaluating Localization
    - Evaluating Visualizations
4. Cases
5. Conclusion

Problem formulation

- We have a not very big dataset and there are 2 models giving the same predictions. Which model to choose?
- We have a classifier model and there is a picture on which it is mistaken. How to find out why this happens?

CAM: Class Activation Mapping

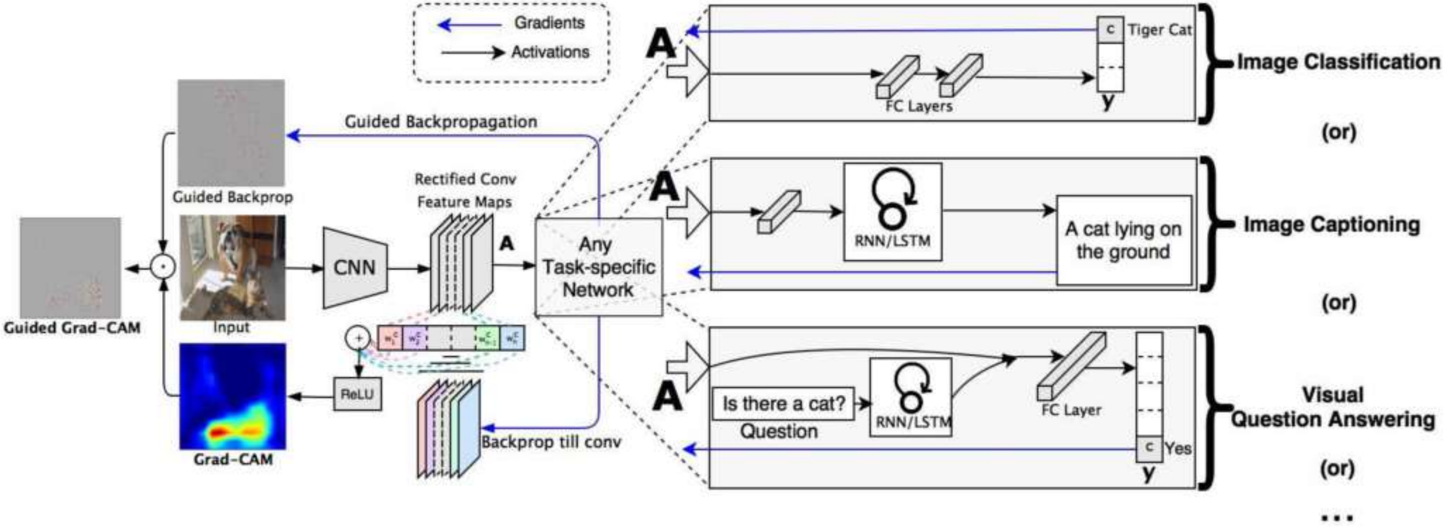


- Learning deep features for discriminative localization
- Class Activation Mapping is applicable to only GAP layers
- Make CAM to applicable to a wide variety of CNN models

Contributions

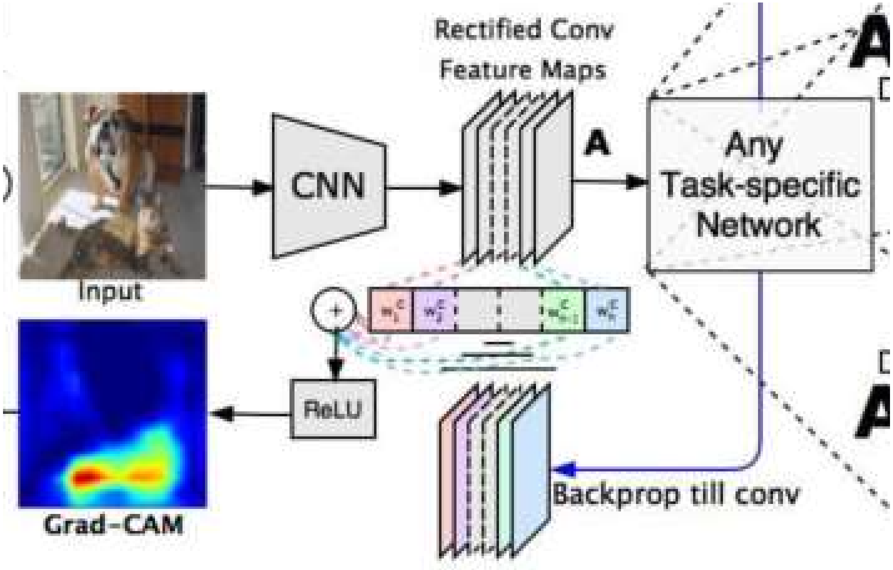
- Apply Grad-CAM to any CNN-based network without requiring architectural changes or re-training
- Apply Grad-CAM to existing top-performing classification, captioning, and VQA.
- Conduct human studies if it helps establish human trust and untrained user can discern a stronger network.

GradCAM



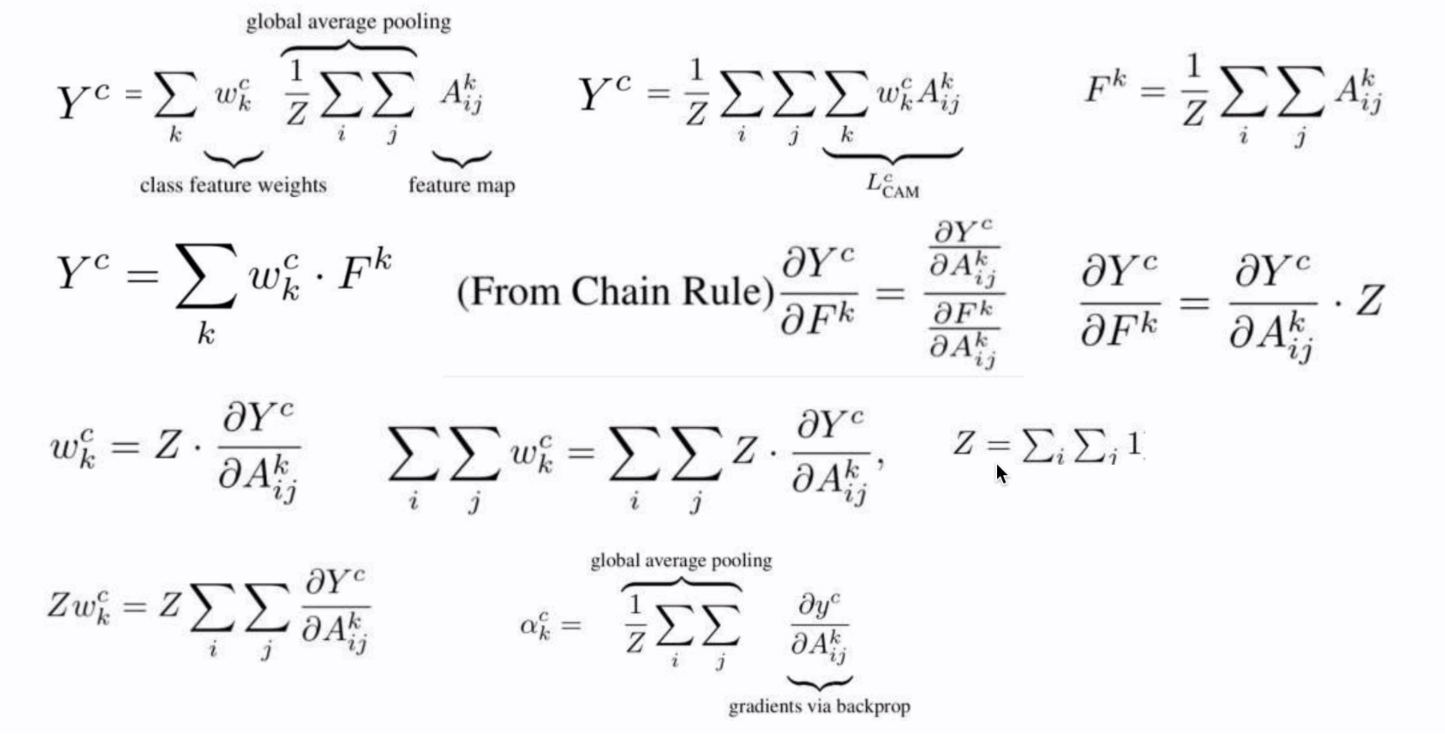
Gradients
Activations
Guided Backpropagation
Guided Backprop
Guided Grad-CAM
Input
CNN
Rectified Conv
Feature Maps
A
Any
Task-specific
Network
ReLU
Grad-CAM
Backprop till conv
A
c
Tiger Cat
FC Layers
y
Image Classification
(or)
A
RNN/LSTM
A cat lying on
the ground
Image Captioning
(or)
A
Is there a cat?
Question
RNN/LSTM
FC Layer
c
Yes
y
Visual
Question Answering
(or)
...

GradCAM



$$L^c_{GradCAM} = ReLU \underbrace{\left(\sum_k a^c_k A^k\right)}_{\text{linear combination}} \qquad a^c_k = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{GAP}} \underbrace{\frac{\partial y^c}{\partial A^k_{ij}}}_{\text{gradients}} \qquad (1)$$

Grad-CAM as a generalization of CAM



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

$$Y^c = \frac{1}{Z}\sum_i\sum_j \underbrace{\sum_k w_k^c A_{ij}^k}_{L^c_{\text{CAM}}}$$

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

$$Y^c = \sum_k w_k^c \cdot F^k$$

(From Chain Rule) $$\frac{\partial Y^c}{\partial F^k} = \frac{\frac{\partial Y^c}{\partial A_{ij}^k}}{\frac{\partial F^k}{\partial A_{ij}^k}}$$

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

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

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

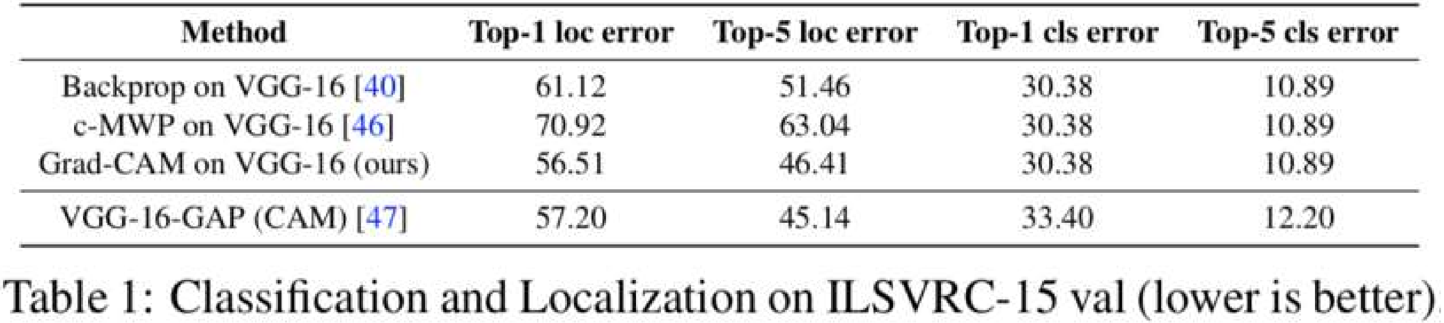
$$Z = \textstyle\sum_i \sum_j 1$$

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

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

Evaluating Localization

1. Weakly supervised localization
   - Use off-the-shelf VGG-16 from Caffe Model Zoo
   - Binarize Grad-CAM with 15
   - Draw bounding box around the single largest segment
2. Weakly supervised segmentation
   - Replace CAM with Grad-CAM in Seed, Expand, Constrain (SEC) algorithm

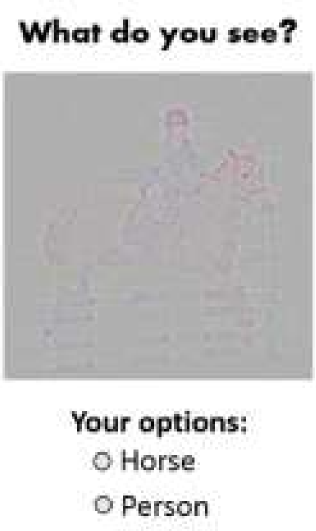


Method	Top-1 loc error	Top-5 loc error	Top-1 cls error	Top-5 cls error
Backprop on VGG-16 [40]	61.12	51.46	30.38	10.89
c-MWP on VGG-16 [46]	70.92	63.04	30.38	10.89
Grad-CAM on VGG-16 (ours)	56.51	46.41	30.38	10.89
VGG-16-GAP (CAM) [47]	57.20	45.14	33.40	12.20

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

Evaluating Visualizations

1. Class Discrimination
   - 43 AMT workers, 4 visualizations, 90 image category pairs, 9 ratings each
   - Deconv vs. Guided backprop vs. Guided Grad-CAM vs. Deconv Grad-CAM
   - 53.33% vs. 44.44% vs. 61.23% vs. 61.23%



What do you see?

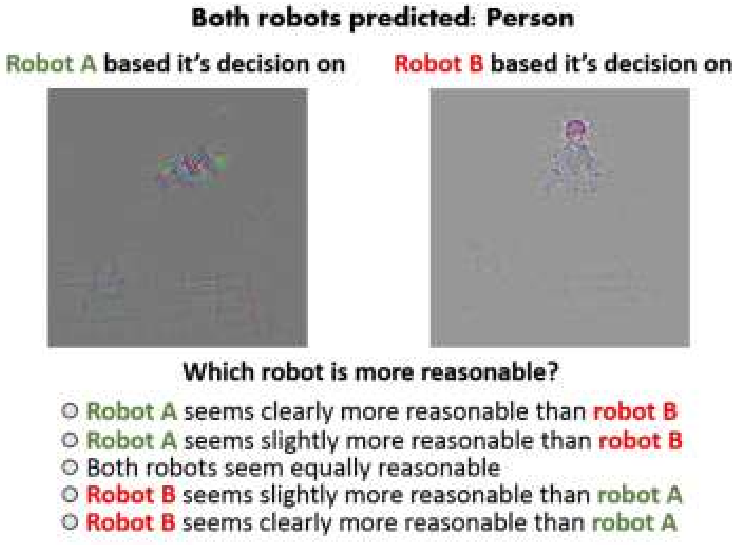
Your options:

○ Horse

○ Person

Evaluating Visualizations

1. Trust worthiness
   - 54 AMT workers, 2 classifiers (AlexNet, VGG-16), 2 visualizations
   - Show same prediction with similar output score
   - Human can identify VGG-16 is better
   - Guided Grad-CAM shows higher difference
   - 1.27 (vs. 1.0 with Guided Backprop)



Both robots predicted: Person

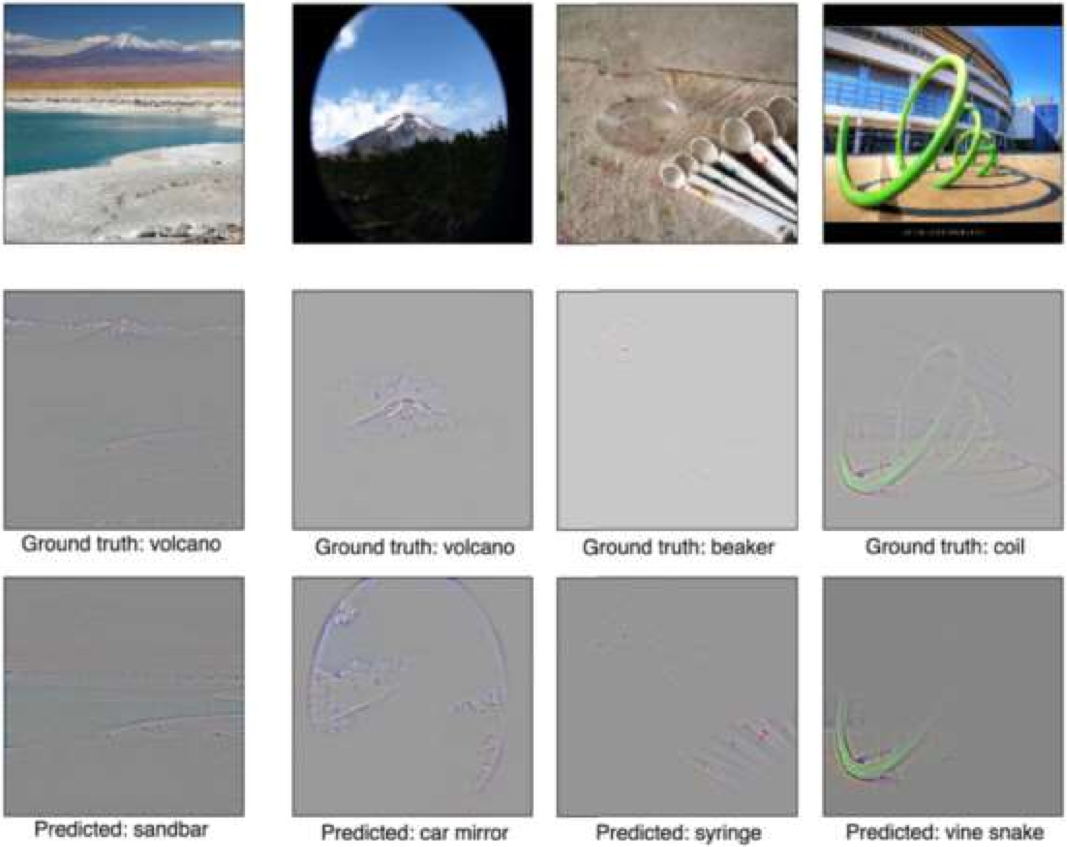
Robot A based it's decision on

Robot B based it's decision on

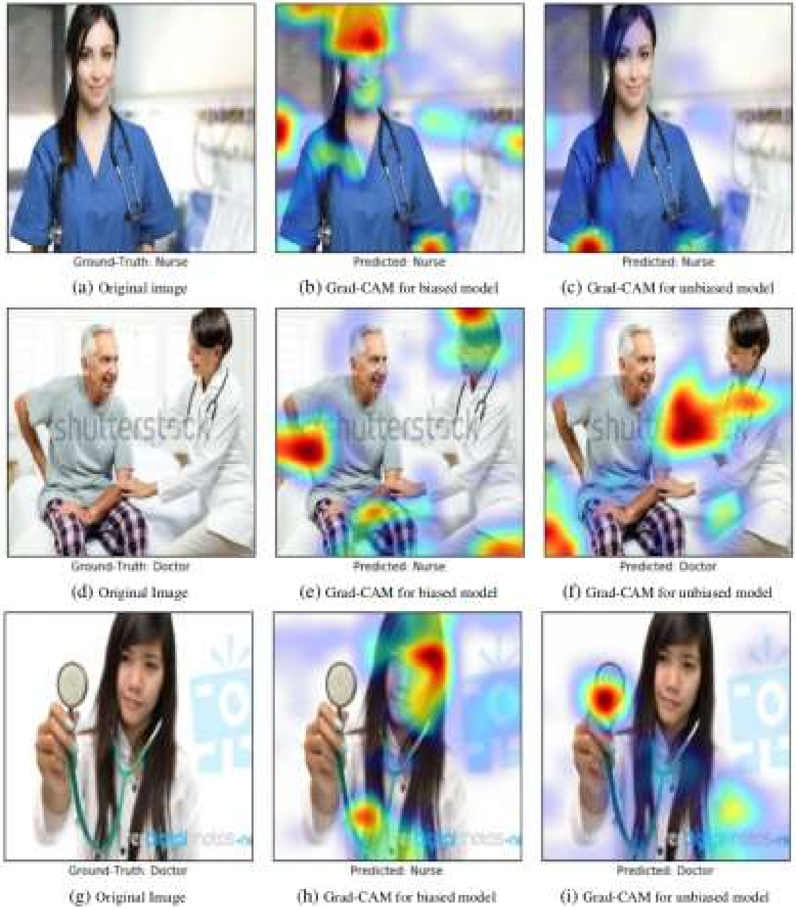
Which robot is more reasonable?

○ 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

Cases: Analyzing failure modes



Cases: Identifying bias in dataset



Ground-Truth: Nurse
(a) Original image

Predicted: Nurse
(b) Grad-CAM for biased model

Predicted: Nurse
(c) Grad-CAM for unbiased model

Ground-Truth: Doctor
(d) Original Image

Predicted: Nurse
(e) Grad-CAM for biased model

Predicted: Doctor
(f) Grad-CAM for unbiased model

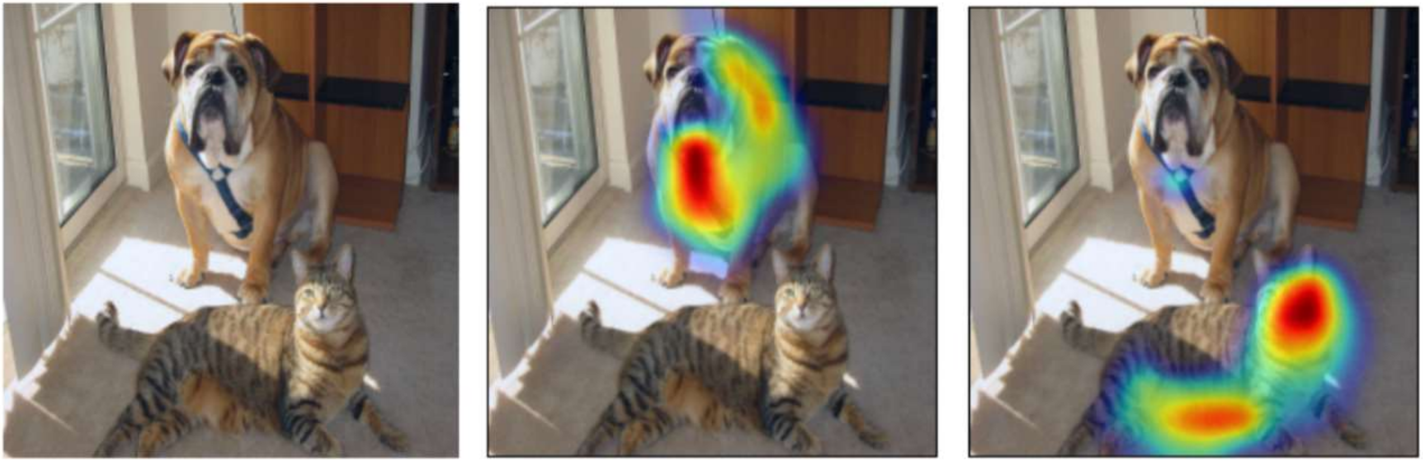
Ground-Truth: Doctor
(g) Original Image

Predicted: Nurse
(h) Grad-CAM for biased model

Predicted: Doctor
(i) Grad-CAM for unbiased model

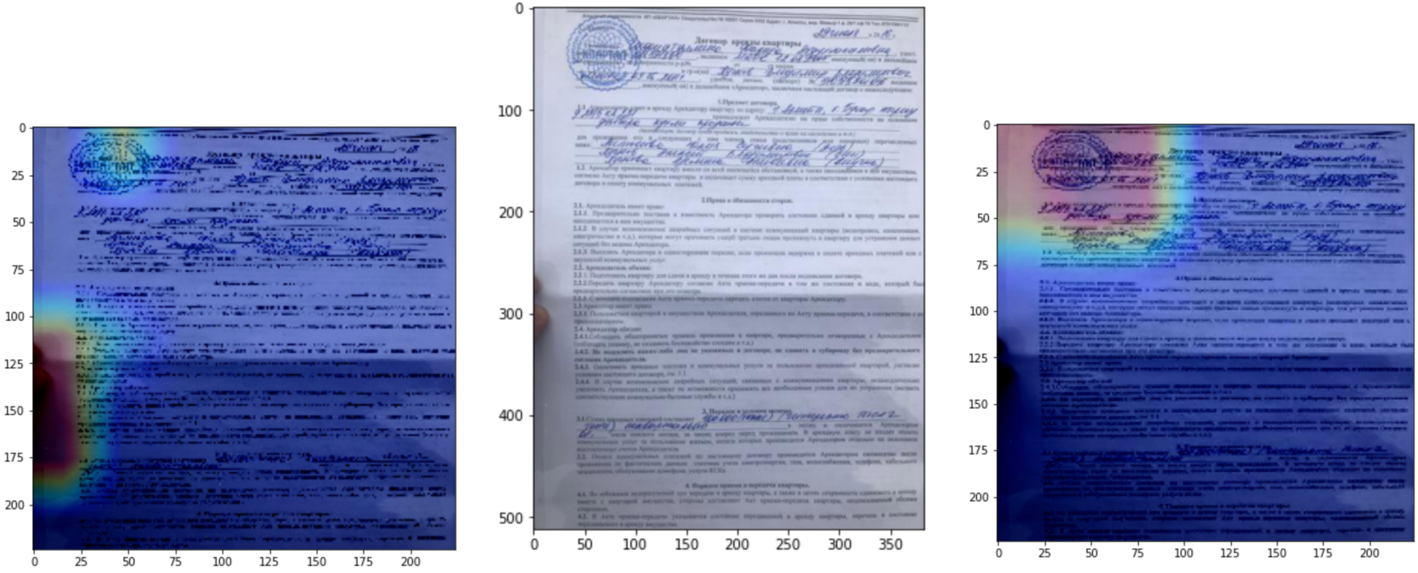
Cases: Counterfactual explanations

$$a_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{GAP}} \underbrace{-\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{negative grads}} \tag{2}$$



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

Cases: My experience with GradCAM



Conclusion

- The paper proposed Gradient-weighted Class Activation Mapping as a generalization of CAM
- Combined Grad-CAM with existing high-resolution visualizations (Guided Grad-CAM)
- Human studies reveal the trustworthiness of a classifier, and help identify biases in datasets
- AI system should not only be intelligent, but also be able to reason about its beliefs and actions for human to trust it