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

3 Feb 2020

Presenter: Sanchit Sinha

https://qdata.github.io/deep2Read/

Motivation

- Answer these questions:
  - Why do neural networks predict the way they do?
  - Why do NNs make predictions which seem to be totally irrelevant
  - What parts of an image are the most useful in predictions
  - Does adversarial perturbation of image change where the NN “look”?
- Generalize an explainability method which works across all types and varieties of CNNs
- Should also work on different domains - classification, segmentation, VQA, etc.

Background

- Explainability and performance are often a tradeoff
  - Simple rule based classifiers with very high explainability do not perform well on complex tasks
  - Complex DNNs are often considered “black boxes” but are very good at complex tasks (sometimes better than humans)
- GradCAM builts with inspiration from Class Activation Mapping which was proposed to find the “active” regions in pure CNNs
- Guided Backprop was the first such technique to venture into explainability - it gives high quality pixel-space gradient visualization methods.
- Deconvolution is also similar to Guided Backprop

Related Work

- Guided Backprop
- Deconvolutions
- CAM
- VQA
- Localization/Segmentation

Claim / Target Task

- Class-discriminative localization technique that generates visual explanations for any CNN-based network
- Apply Grad-CAM to existing top-performing classification, captioning and VQA models.
- Proof-of-concept of how interpretable GradCAM visualizations help in diagnosing failure modes
- Present Grad-CAM visualizations for ResNets
- Neuron importance from Grad-CAM and neuron names

Proposed Solution

- First taking derivatives with respect to the output (before the softmax) of a particular class.
- Global average pooling over the width*height of the desired map. Obtaining map level importance score.

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

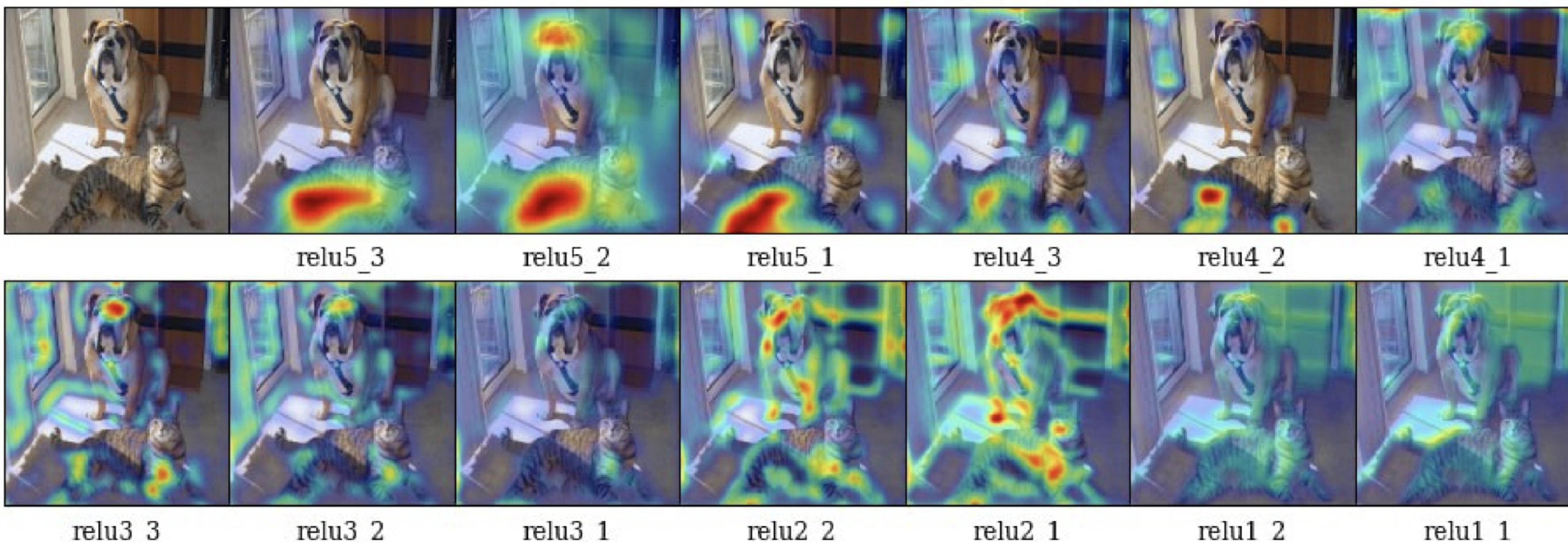
- As we are only interested in the activations which give “positive” influence on the scores, we have to remove the negative gradients. So we apply the ReLU

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

Implementation

- Only used the Convolution layer output from the last convolution layer before the fully connected layers.
    - Why? - The last layer will have the largest receptive field and will give the best spatial information.
    - Why only Conv layers? - If we use it in FC layers, the spatial information is lost

- Guided GradCAM - Hadamard product (element-wise) of the heatmaps from the GradCAM and Guided Backpropagation.
    - Why? - Guided Backprop gives much higher quality output. Taking element-wise product with the GradCAM output will definitely only highlight the most important and high quality areas of the maps


relu5_3
relu5_2
relu5_1
relu4_3
relu4_2
relu4_1
relu3_3
relu3_2
relu3_1
relu2_2
relu2_1
relu1_2
relu1_1

Data Summary

Too many different to summarize.
Imagenet
Pascal VOC
COCO

Experimental Results - Localization

		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

Experimental Results - Segmentation

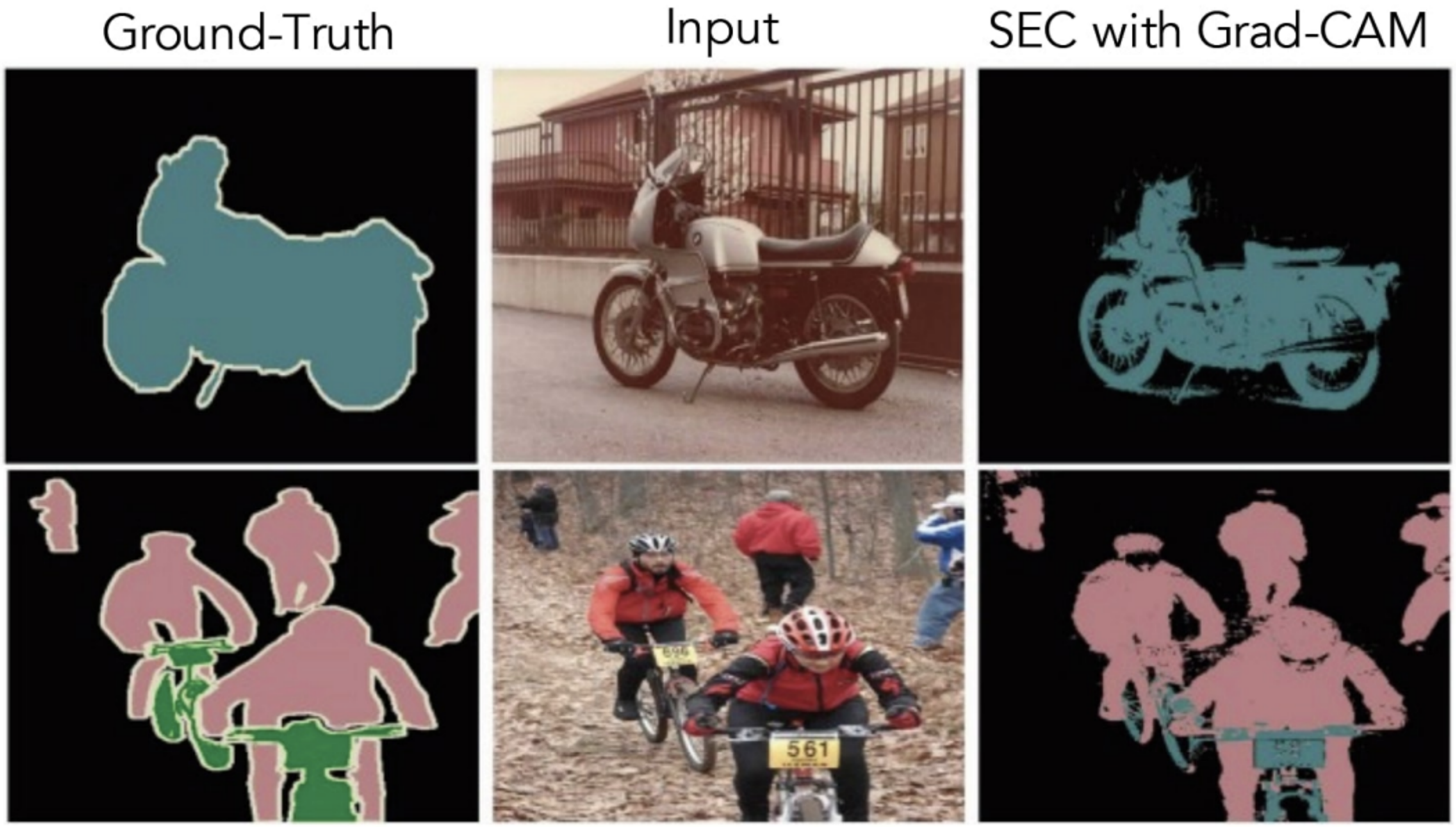
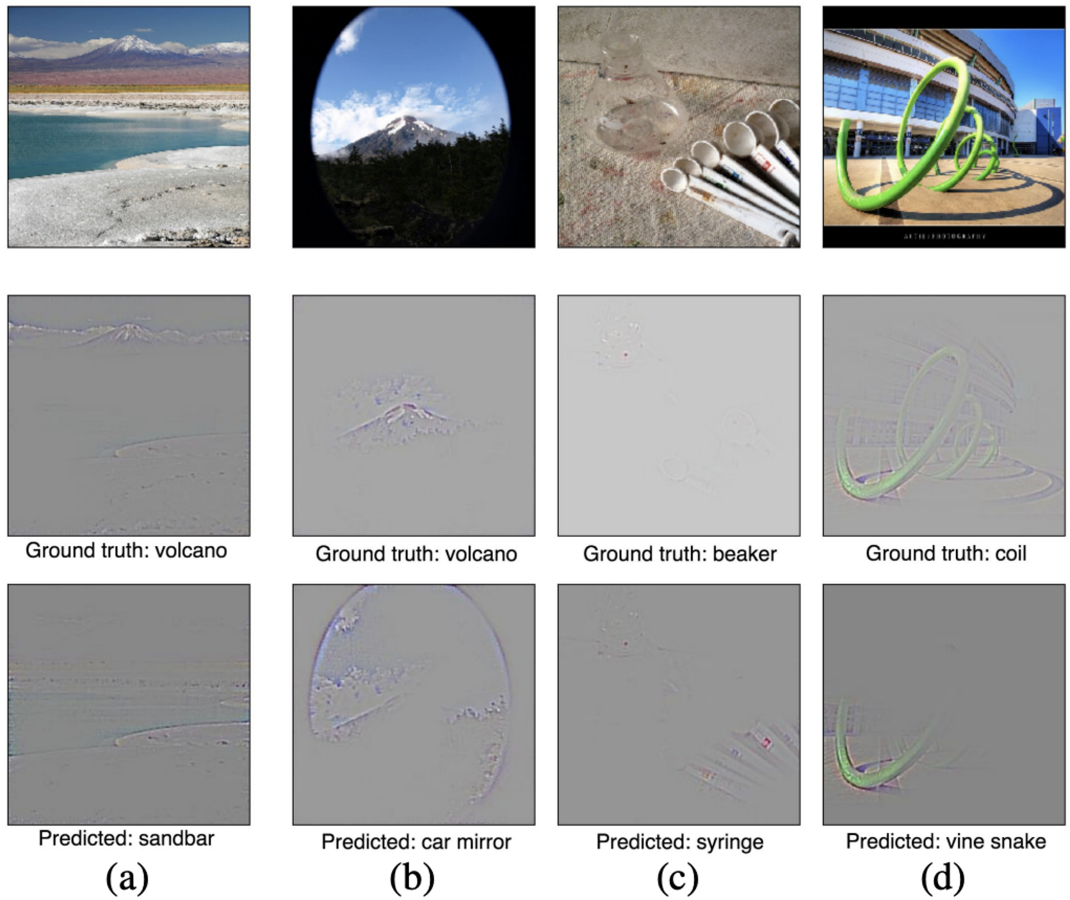


Fig. 4: PASCAL VOC 2012 Segmentation results with Grad-CAM as seed for SEC [32].

Experimental Results - Diagnosis

6.1 Analyzing failure modes for VGG-16



Experimental Results - Adversarial

Boxer: 0.4 Cat: 0.2
(a) Original image

Airliner: 0.9999
(b) Adversarial image

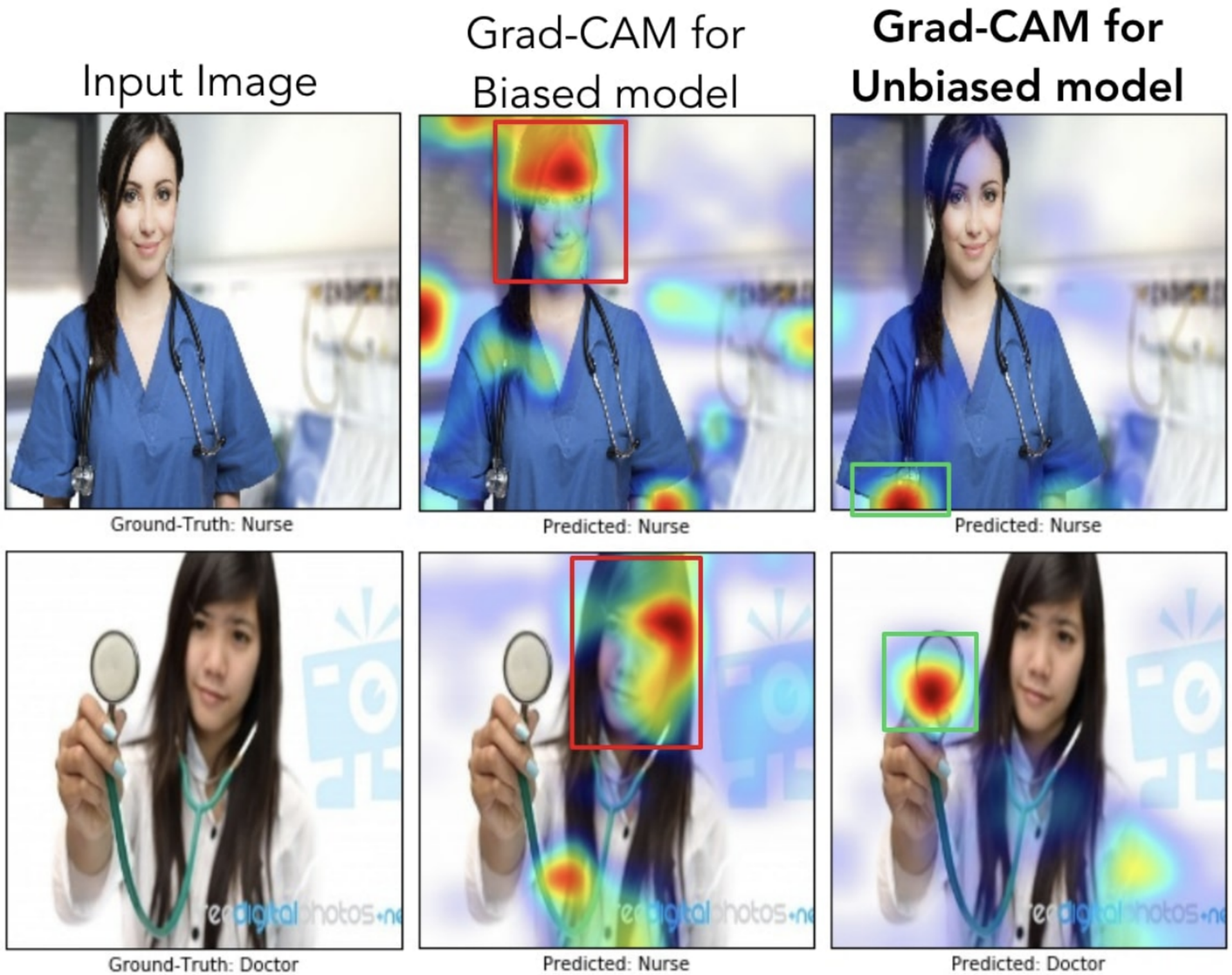
Boxer: 1.1e-20
(c) Grad-CAM “Dog”

Tiger Cat: 6.5e-17
(d) Grad-CAM “Cat”

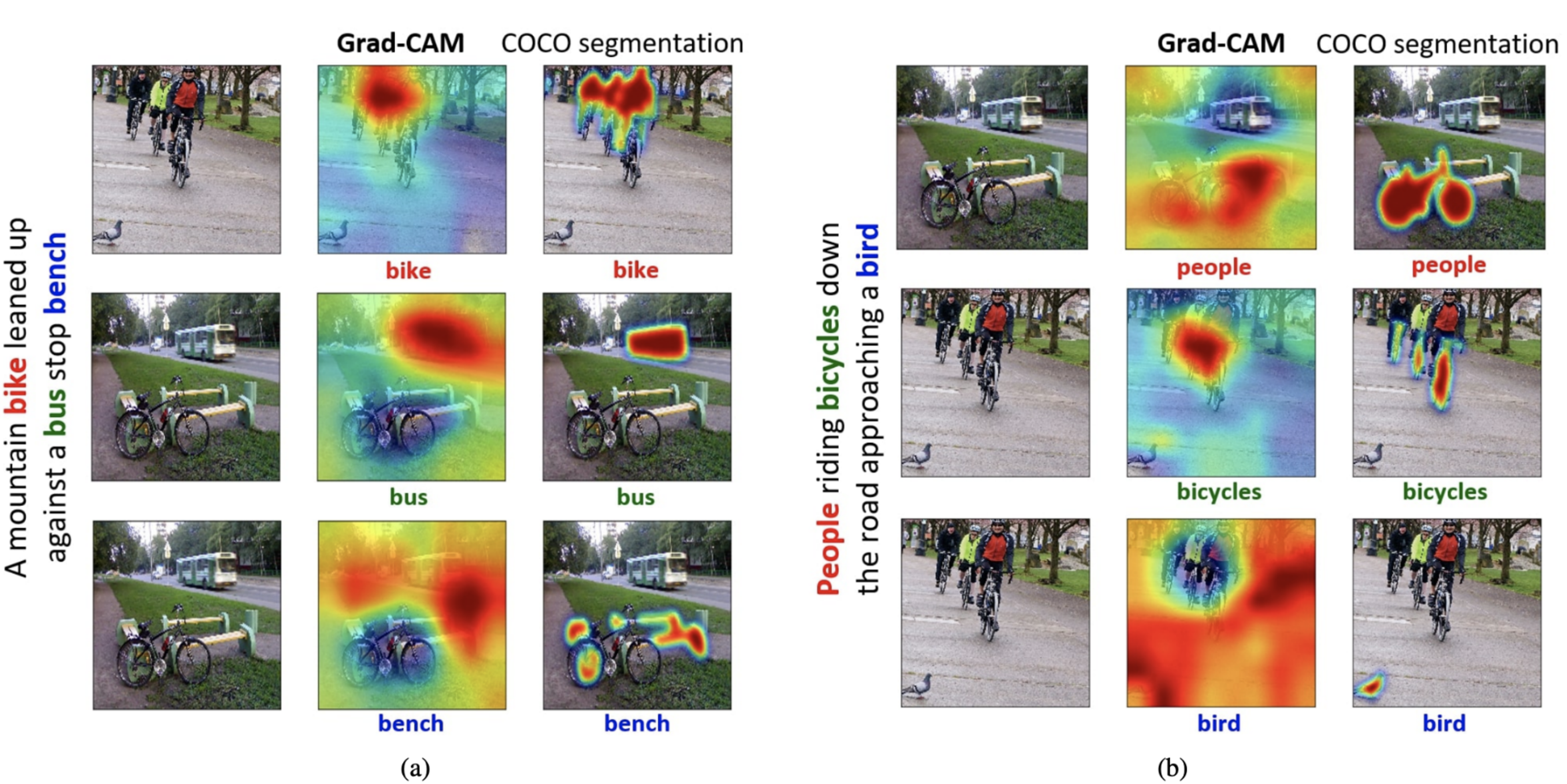
Airliner: 0.9999
(e) Grad-CAM “Airliner”

Space shuttle: 1e-5
(f) Grad-CAM “Space Shuttle”

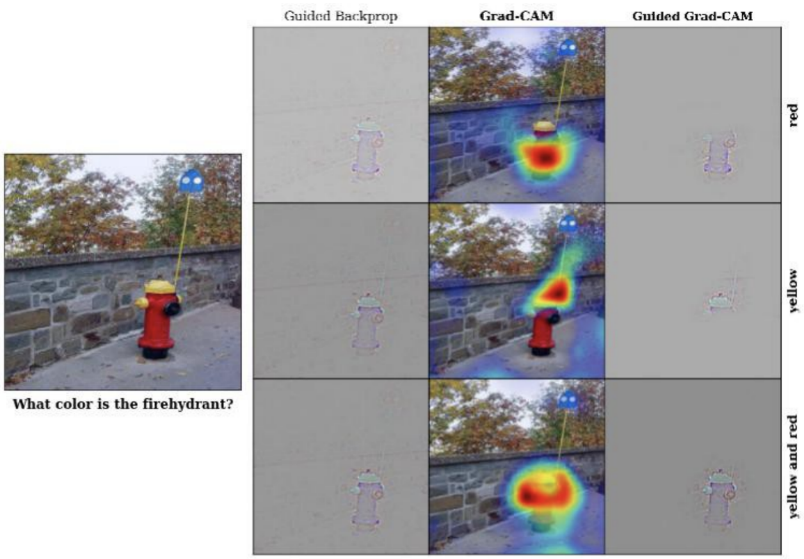
Experimental Results - Bias

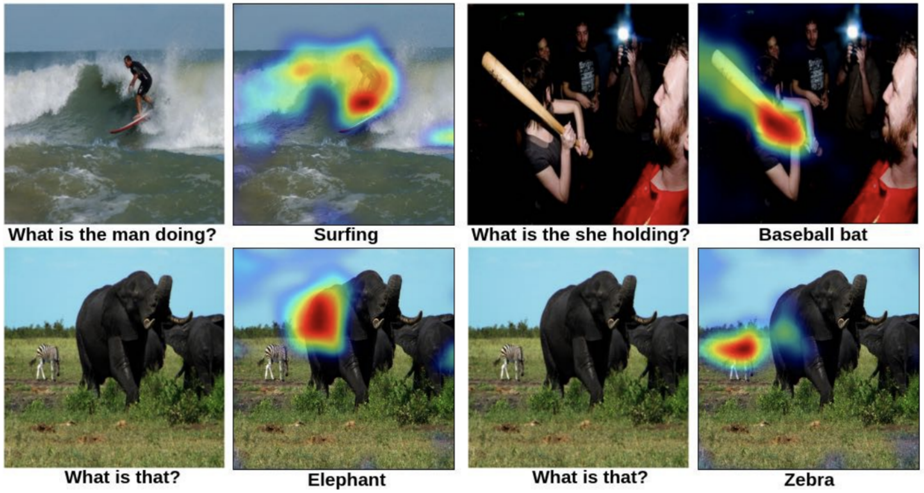


Experimental Results - Captioning



Experimental Results - VQA





References

- D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing Higher-layer Features of a Deep Network. University of Montreal, 1341, 2009. 3 17.
- M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.
- M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free? – weakly-supervised learning with convolutional neural networks. In CVPR, 2015.