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

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

Supervisors

Dr. Amro NAJJAR

Dr. Sana NOUZRI

University of Luxembourg

Presented by

Saddam Hossain

(09027635C)

Agenda

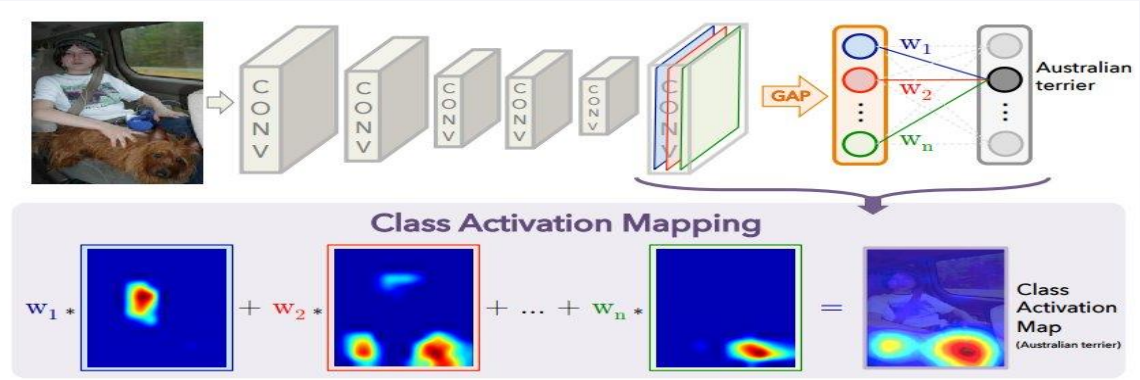
- Why interpretability matters?
- Motivation
- Contributions
- Approach
- Evaluating Localization
- Evaluating Visualizations
- Diagnosing image classification CNNs
- Image Captioning and VQA
- Related Work
- Demo
- Conclusion

Why interpretability matters?

- The lack of decomposability of deep network into intuitive and understandable components makes them hard to interpret
- Transparent model is necessary
  - To build trust in intelligent systems and move towards into our everyday life
- When AI is weaker
  - To identify failure modes
- When AI is on par with humans and reliably deployable
  - The goal is to establish trust and confidence in users
- When AI is significantly stronger than humans
  - Machine teaching a human about how to make better decisions

Motivation

- CAM: Learning deep features for discriminative localization



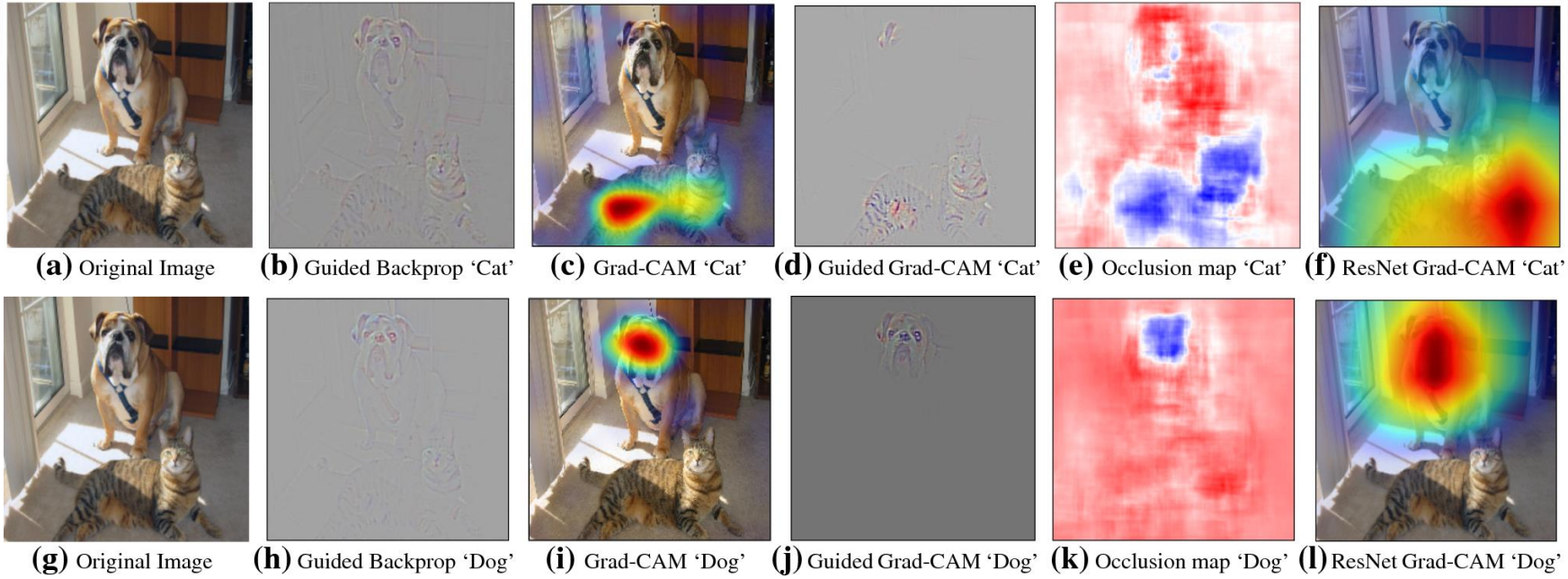
- Class Activation Mapping is applicable to only GAP layers
- Make CAM to applicable to a wide variety of CNN models
  - CNNs with fully-connected layers (e.g. VGG)
  - CNNs for structured outputs (e.g. captioning)
  - CNNs used in tasks with multi-modal inputs (e.g. VQA)

Contributions

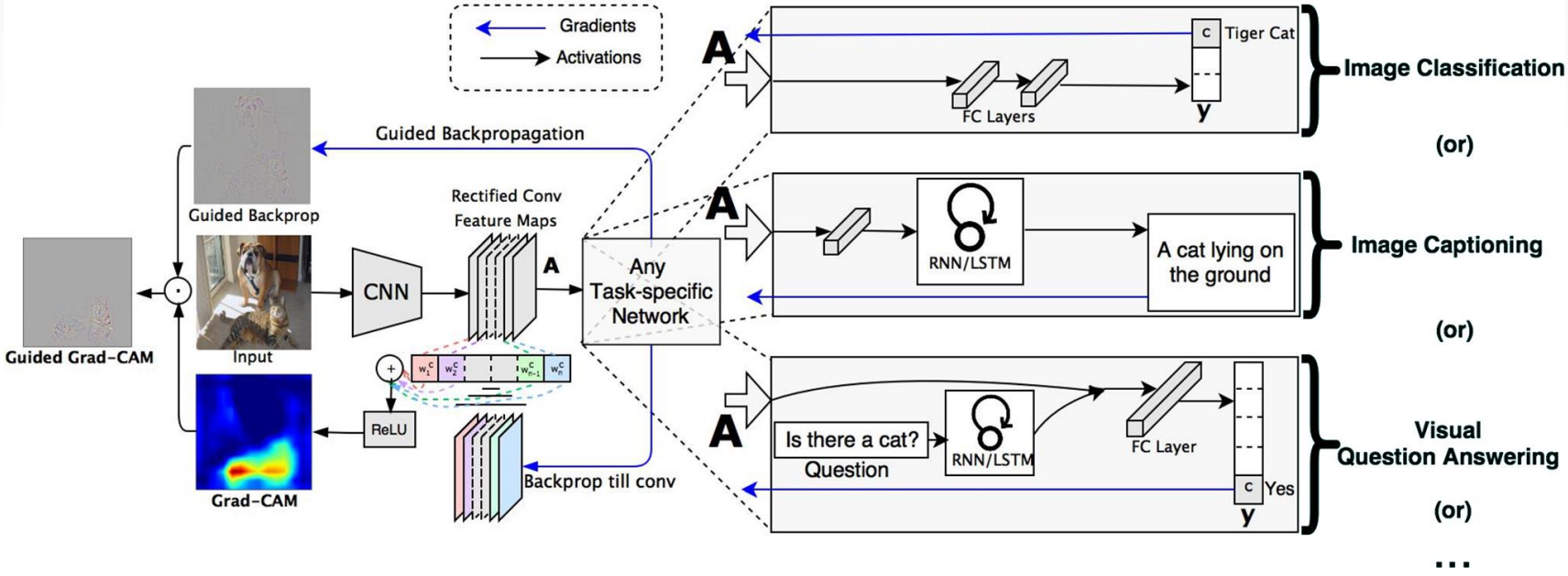
- Apply Grad-CAM to any CNN-based network without requiring architectural changes or re-training
- Authors show a proof-of-concept of how interpretable Grad-CAM visualizations.
- Apply Grad-CAM to existing top-performing classification, captioning, and VQA
- Authors present Grad-CAM visualizations for ResNets
- Authors use neuron importance from Grad-CAM
- Conduct human studies if it helps establish human trust and untrained user can discern a stronger network

What makes a good visual explanation?

a) class-discriminative (b) high-resolution



Approach


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
FC Layers
c
Tiger Cat
y
Image Classification
(or)
RNN/LSTM
A cat lying on the ground
Image Captioning
(or)
Is there a cat?
Question
RNN/LSTM
FC Layer
c
Yes
y
Visual Question Answering
(or)
...

Grad-CAM as a generalization of CAM

- Formally prove that Grad-CAM generalizes CAM for a wide variety of CNN-based architectures
- This approach modifies image classification CNN architectures replacing fully-connected layers with convolutional layers and global average pooling , thus achieving class-specific feature maps
- Authors introduce a new way of combining feature maps using the gradient signal that does not require any modification in the network architecture
- For a fully-convolutional architecture, Grad-CAM reduces to CAM. Thus, Grad-CAM is a generalization to CAM

Evaluating Localization Ability of Grad-CAM

Weakly-Supervised Localization

Weakly-Supervised Segmentation

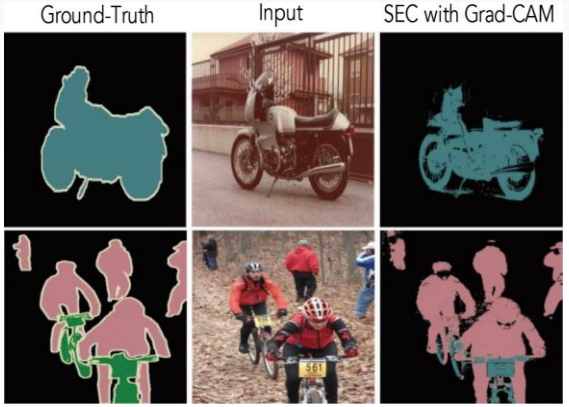
Pointing Game

Weakly-Supervised Localization

	Classification		Localization	
	Top-1	**Top-5**	**Top-1**	**Top-5**
VGG-16				
Backprop (Simonyan et al. 2013)	30.38	10.89	61.12	51.46
c-MWP (Zhang et al. 2016)	30.38	10.89	70.92	63.04
Grad-CAM (ours)	30.38	10.89	**56.51**	46.41
CAM (Zhou et al. 2016)	33.40	12.20	57.20	**45.14**
AlexNet				
c-MWP (Zhang et al. 2016)	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 (Zhou et al. 2016)	31.9	11.3	60.09	49.34

Weakly-Supervised Segmentation

- To seed with weak localization cues, encouraging segmentation network to match these cues
- To expand object seeds to regions of reasonable size based on information about which classes can occur in an image
- To constrain segmentations to object boundaries that alleviates the problem of imprecise boundaries already at training time



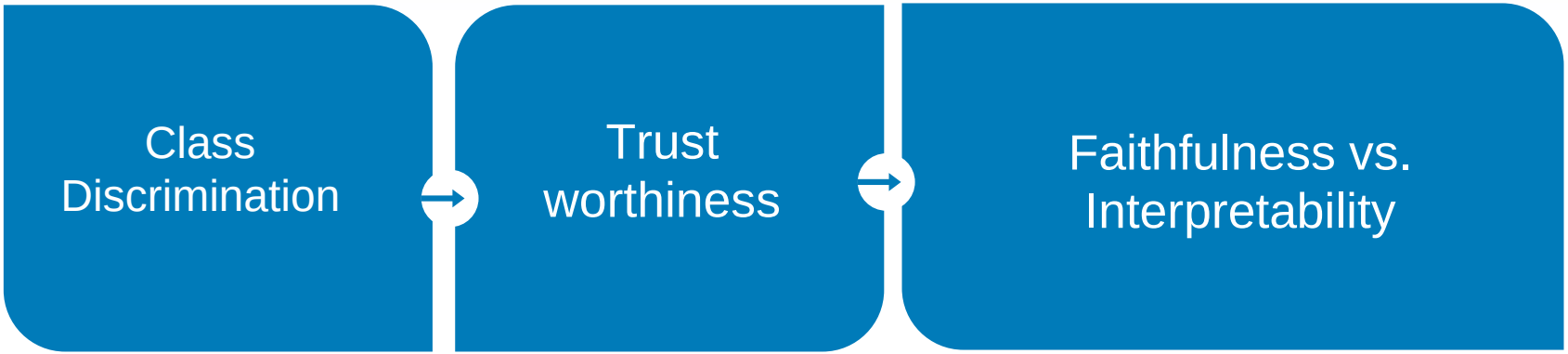
Pointing Game

- Zhang et al. (2016) introduced the Pointing Game experiment to evaluate the discriminativeness of different visualization methods for localizing target objects in scenes

$$Acc = \frac{\#Hits}{\#Hits + \#Misses}$$

- Grad-CAM outperforms c-MWP (Zhang et al. 2016) by a significant margin (70.58% *vs*.. 60.30%)

Evaluating Visualizations

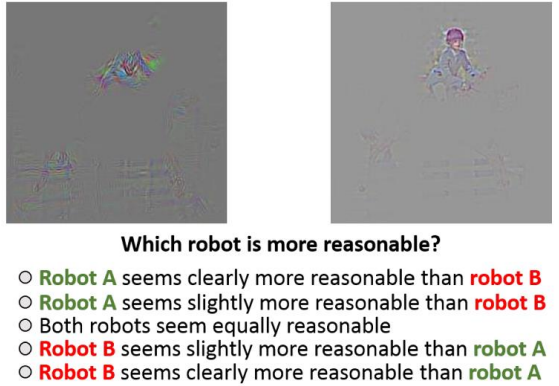


Class Discrimination

- 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%

(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

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)

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

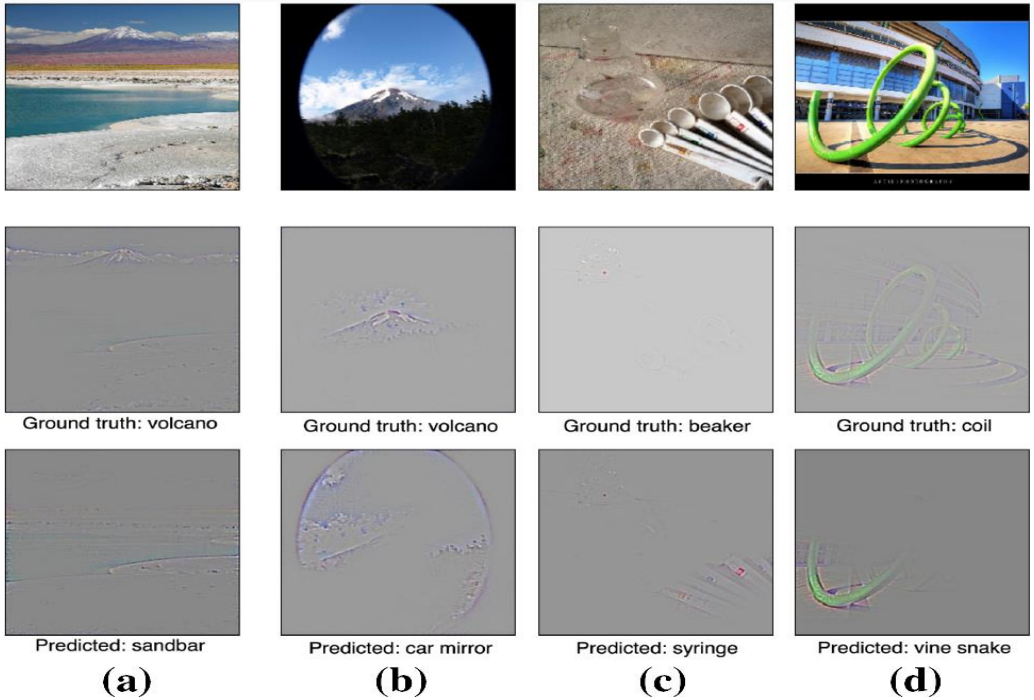
Faithfulness vs. Interpretability

- CNN score after occlude image patches
- Guided Grad-CAM assign high intensity
- Grad-CAM visualizations are *more interpretable*
- Score correlates highly with Grad-CAM
- Grad-CAM is *more faithful* to the model

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

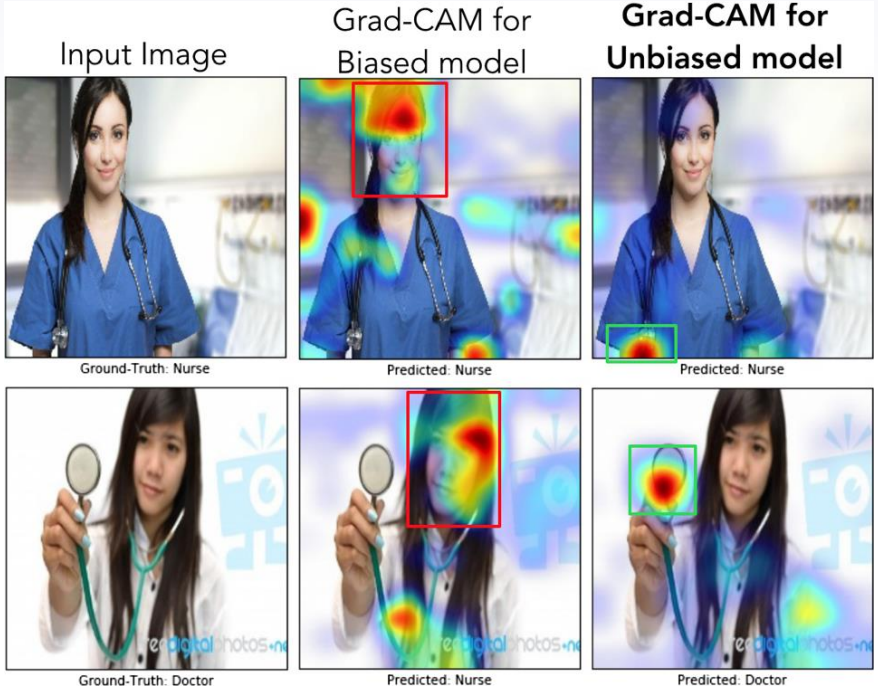
Diagnosing image classification CNNs

Analyzing failure modes for VGG-16



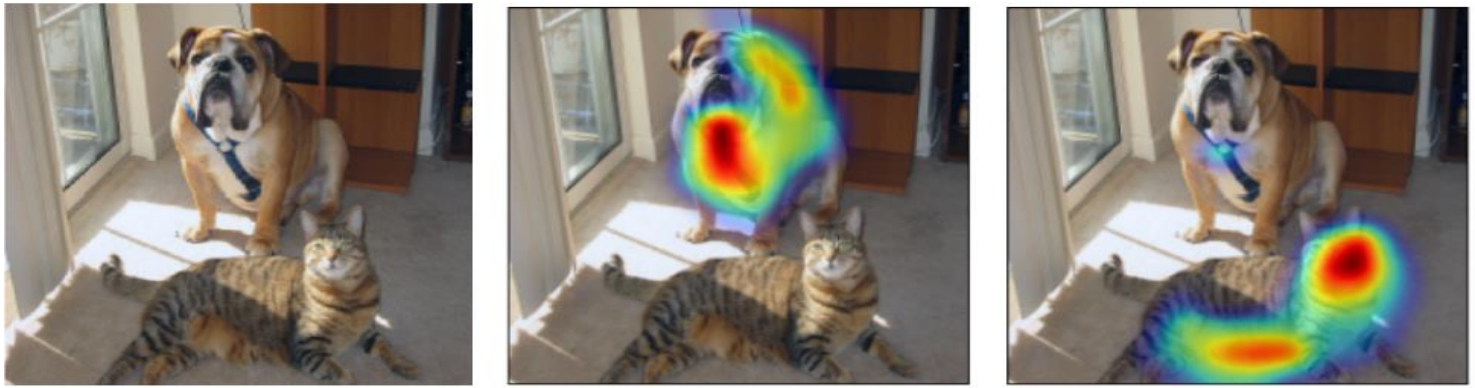
Diagnosing image classification CNNs

- Identifying bias in dataset



Counterfactual explanations

- Using a slight modification to Grad-CAM
- Use negative values to find regions that decreases output score



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

Image captioning

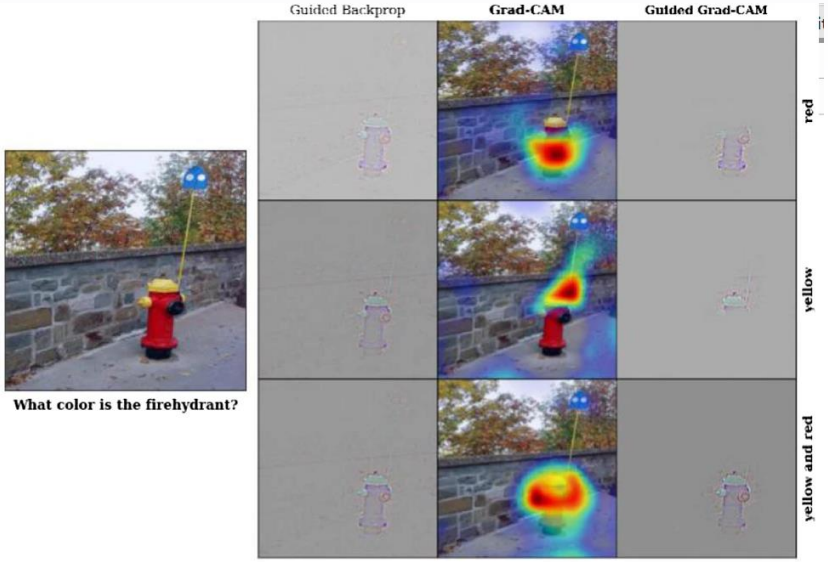
- Use neuraltalk2: VGG-16 for image and LSTM language model
- No explicit attention
- Compare with DenseCap
- Consist of Fully Convolutional Localization Network and LSTM

(a) Image captioning explanations

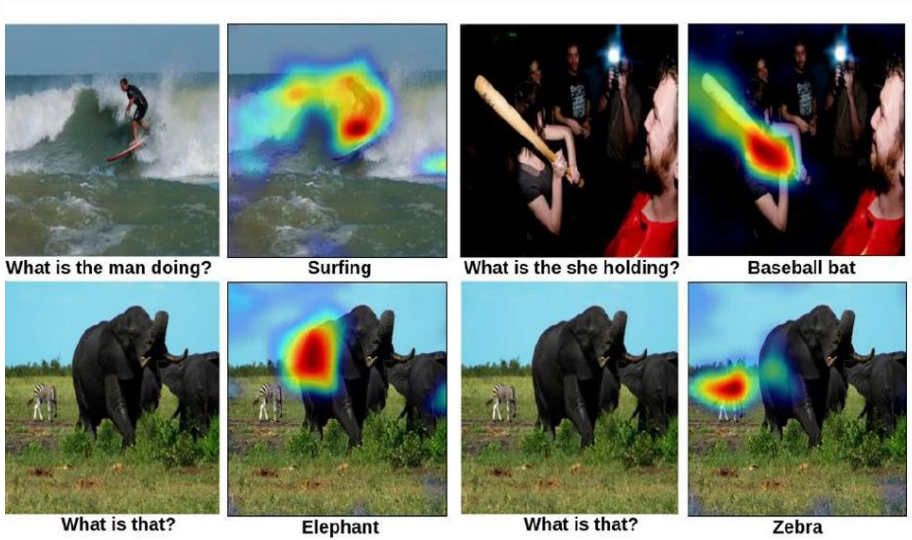
(b) Comparison to DenseCap

Visual Question Answering

- Grad-CAM correlation (with occlusion maps) of 0.60±0.038



(a) Visualizing VQA model from [38]



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

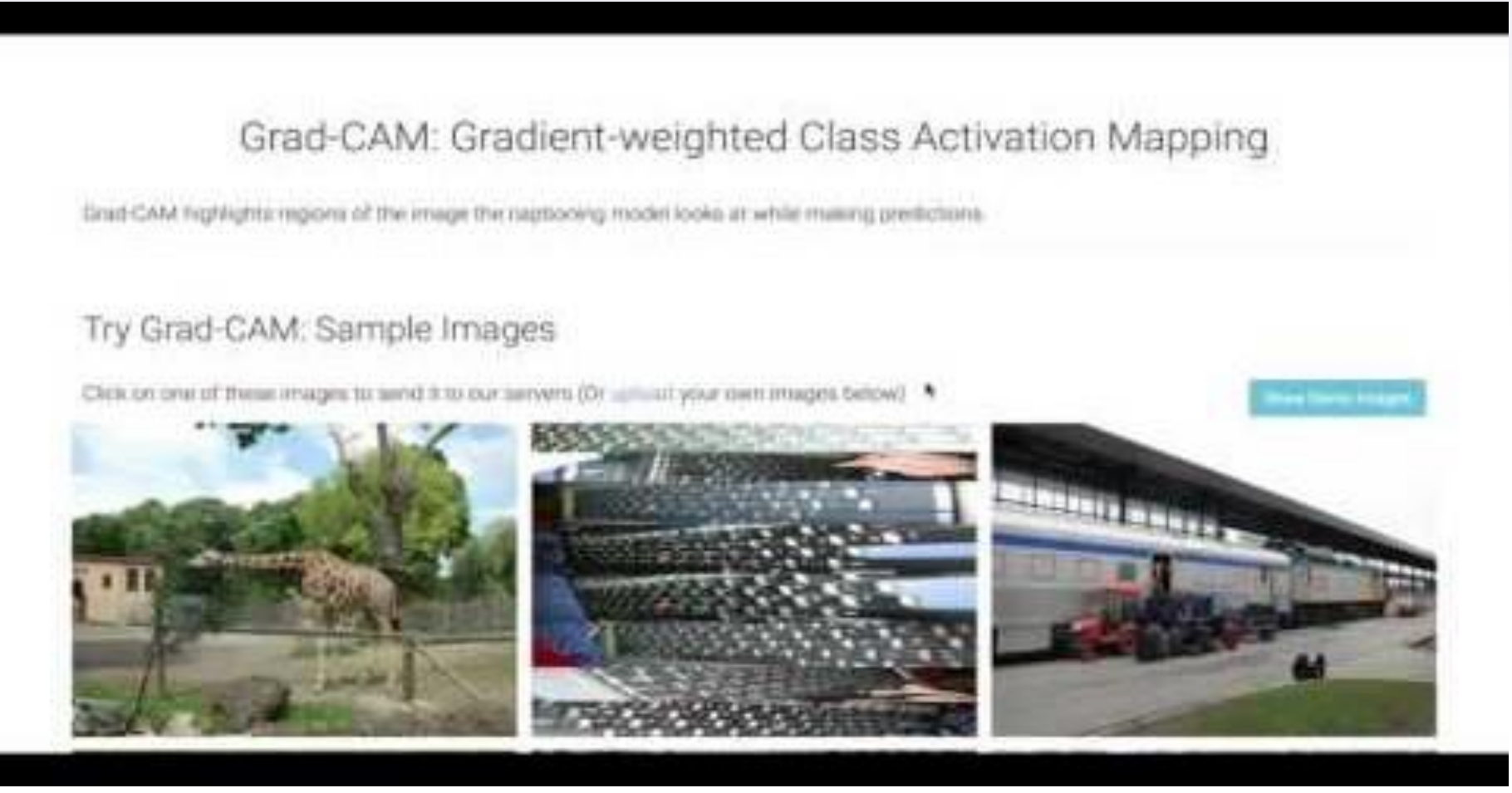
Visual Question Answering

- Comparison to Human Attention
  - Collected human attention maps for a subset of the VQA dataset
  - Grad-CAM and human attention maps have a correlation of 0.136, which is higher than chance or random attention maps
- Visualizing ResNet-Based VQA Model with Co-Attention
  - Use a 200 layer ResNet to encode the image

Related Work

- Visualizing CNNs
  - Highlight important pixels: non discriminative
  - Synthesize images to maximally activate a network unit or invert a latent representation: not for specific input images
- Assessing Model Trust
  - Motivated by notions of interpretability
  - There are some methods to assess trust in models
- Aligning Gradient-Based Importances
- Weakly-Supervised Localization
  - – Perturbing inputs by occlusion

Demo


Grad-CAM: Gradient-weighted Class Activation Mapping
Grad-CAM highlights regions of the image the captioning model looks at while making predictions
Try Grad-CAM: Sample Images
Click on one of these images to send it to our servers (Or upload your own images below)

THANKS!

Any questions?