

# **Grad-CAM**

## **Visual Explanations from Deep Networks via Gradient-based Localization**

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

**Presenter: Maulik Shah**

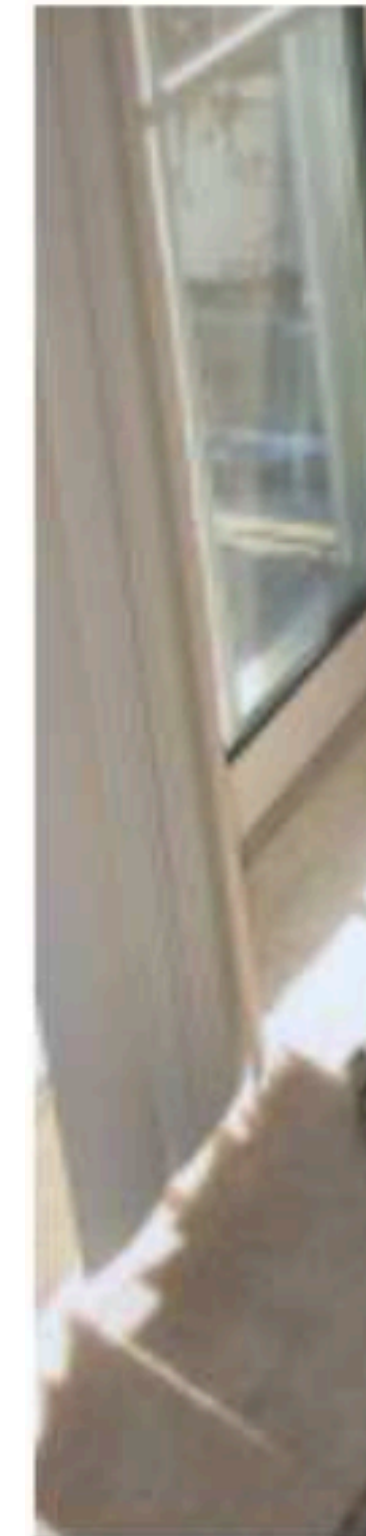
**Scribe: Yunjia Zhang**

# **Explaining Deep Networks is Hard!**

**What's a good visual explanation?**

# Good visual explanation

- Class discriminative - localize the category in the image
- High resolution - capture fine-grained detail



(b) Guided Backprop 'Cat'



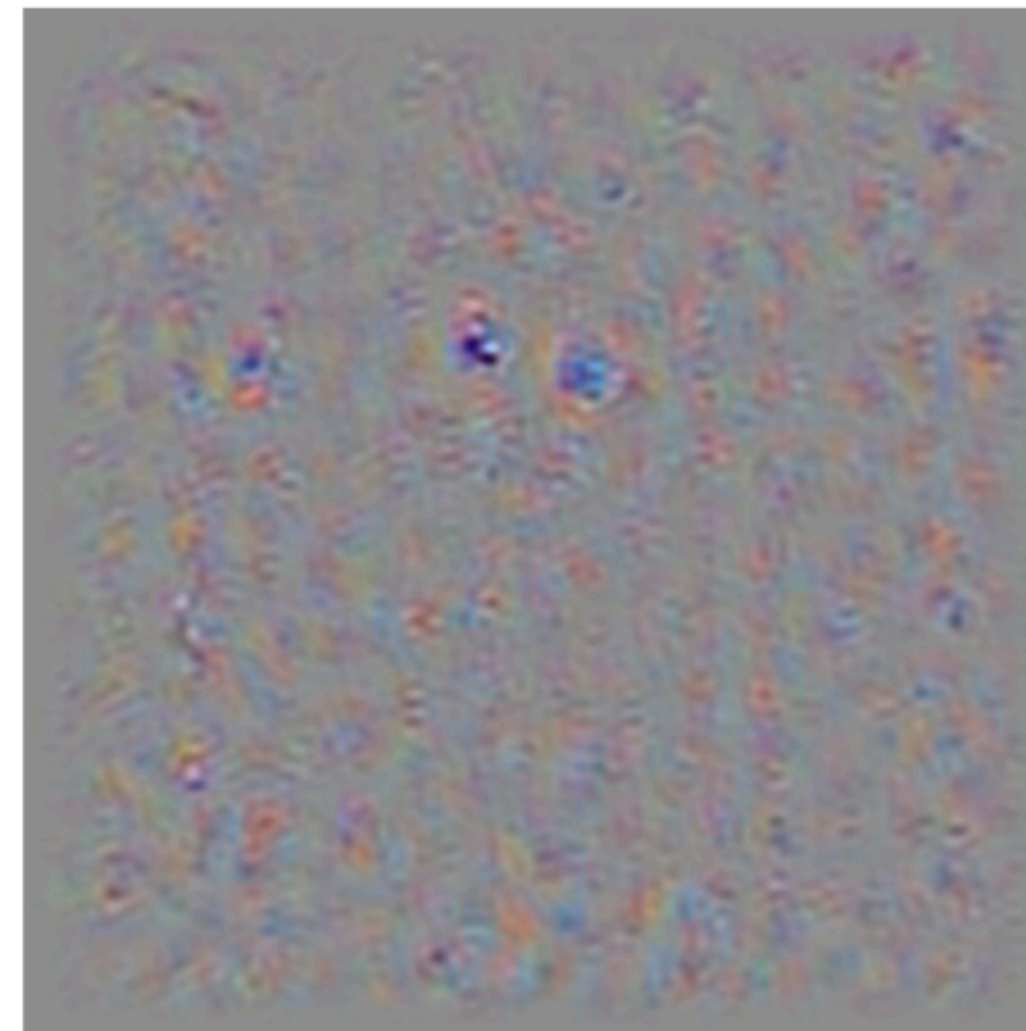
(h) Guided Backprop 'Dog'



# Work done in explaining Deep Networks

- CNN visualization
  - Guided Backpropagation
  - Deconvolution
- Assessing Model Trust
- Weakly supervised localization
- Class Activation Mapping (CAM)

Deconvolution



Guided Backprop



# Class Activation Mapping

## What is it?

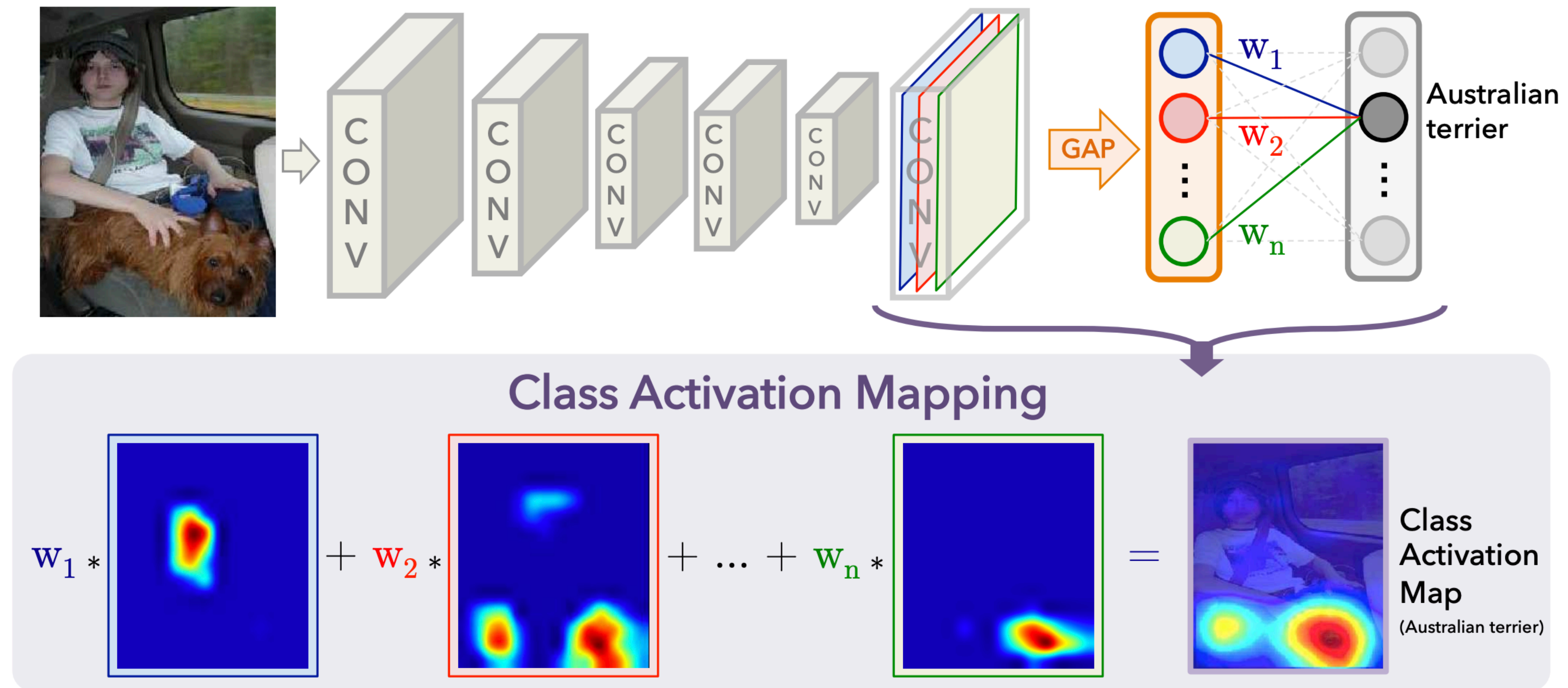
- Enables Classification CNNs to learn to perform localization
- CAM indicates the discriminative regions used to identify that category
- No explicit bounding box annotations required
- However, it needs to change the model architecture:
  - Just before the final output layer, they perform global average pooling on the convolutional feature maps
  - Use these features for a fully-connected layer that produces the desired output

# Class Activation Mapping

## How does it work?

- $f_k(x, y)$  : Activation of unit  $k$  in spatial location  $(x, y)$
- $F^k = \sum_{x,y} f_k(x, y)$  : Result of global average pooling
- $S_c = \sum_k w_k^c F_k$  : input to Softmax layer for class  $c$
- $M_c(x, y) = \sum_k w_k^c f_k(x, y)$  : CAM for class  $c$

# Class Activation Mapping



# Class Activation Mapping

## Drawbacks

- Requires feature maps to directly precede softmax layers
- Such architectures may achieve inferior accuracies compared to general networks on other tasks
- Inapplicable to other tasks like VQA, Image Captioning
- Need a method that doesn't need any modification to existing architecture
- Enter Grad-CAM!

# Gradient weighted Class Activated Mappings

## Overview

- A class discriminative localization technique that can work on *any* CNN based network, without requiring architectural changes or re-training
- Applied to existing top-performing classification, VQA, and captioning models
- Tested on ResNet to evaluate effect of going from deep to shallow layers
- Conducted human studies on Guided Grad-CAM to show that these explanations help establish trust, and identify a ‘stronger’ model from a ‘weaker’ one though the outputs are the same

# Grad-CAM

## Motivation

- Deeper representations in a CNN capture higher-level visual constructs
- Convolutional layers retain spatial information, which is lost in fully connected layers
- Grad-CAM uses gradient information flowing from the last layer to understand the importance of each neuron for a decision of interest



# Grad-CAM

## How it works

- Compute  $\frac{\partial y^c}{\partial A^k}$ : gradient of score  $y^c$  for class  $c$  wrt feature maps  $A^k$
- Global average pool these gradients to obtain neuron importance weights
$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$
- Perform weighted combination of forward activations maps and follow it by ReLU to obtain

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



# Grad-CAM

## How it works

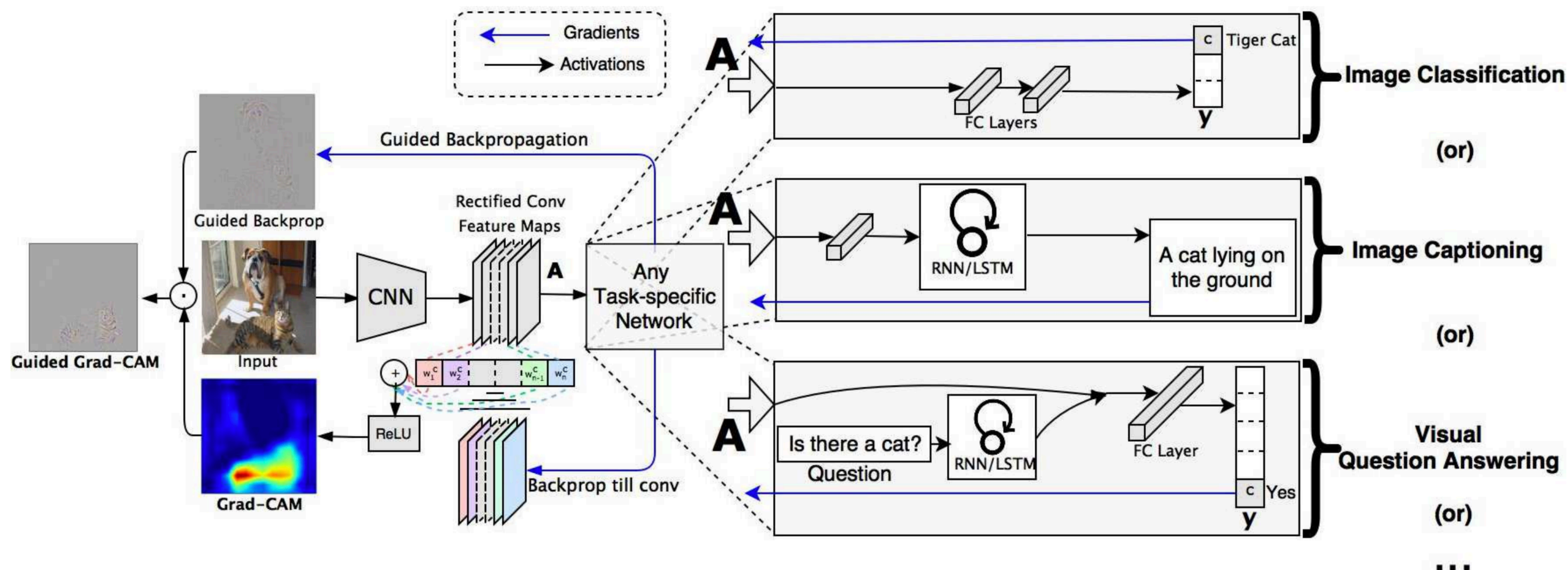
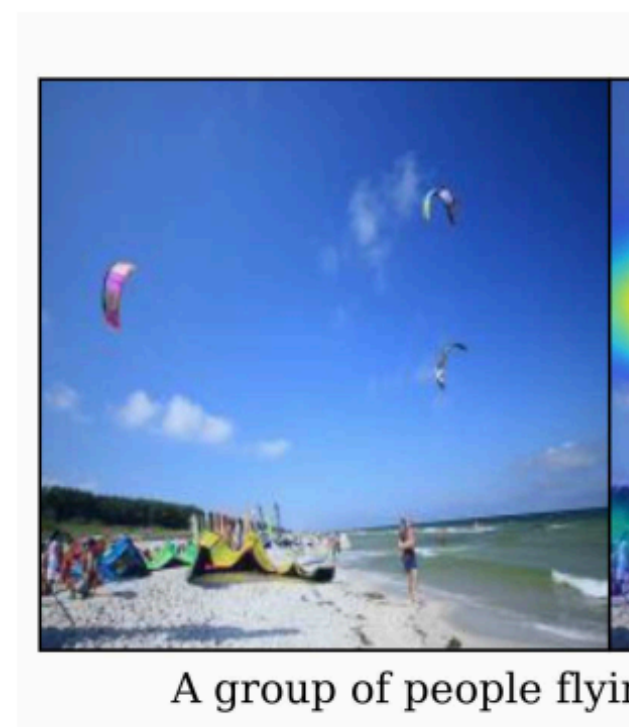


Figure 2: Grad-CAM overview: Given an image and a class of interest (e.g., 'tiger cat' or any other type of differentiable output) as input, we forward propagate the image through the CNN part of the model and then through task-specific computations to obtain a raw score for the category. The gradients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

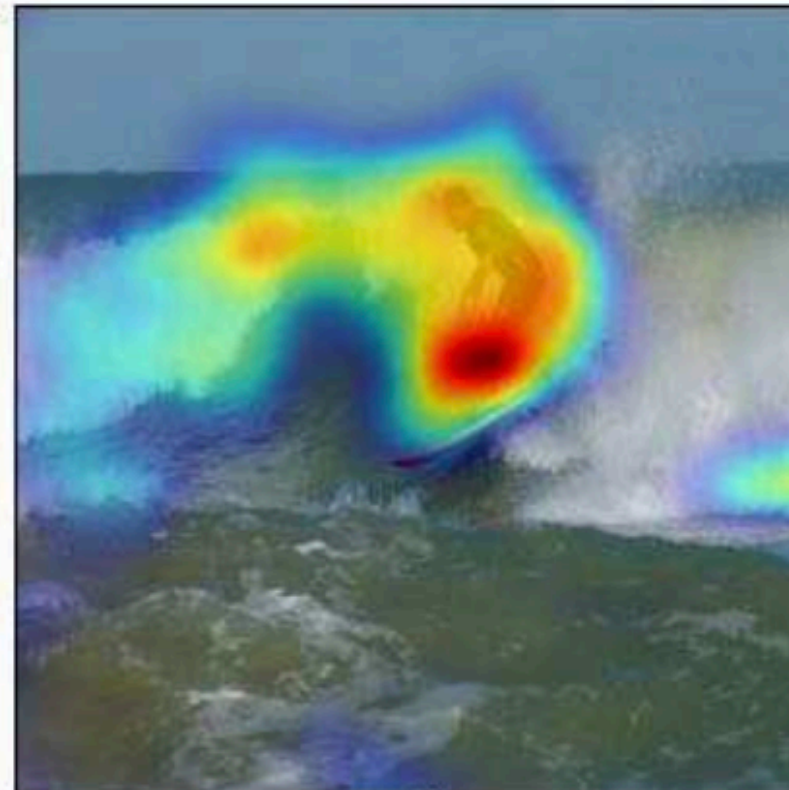


# Grad-CAM

## Results



What is the man doing?



Surfing



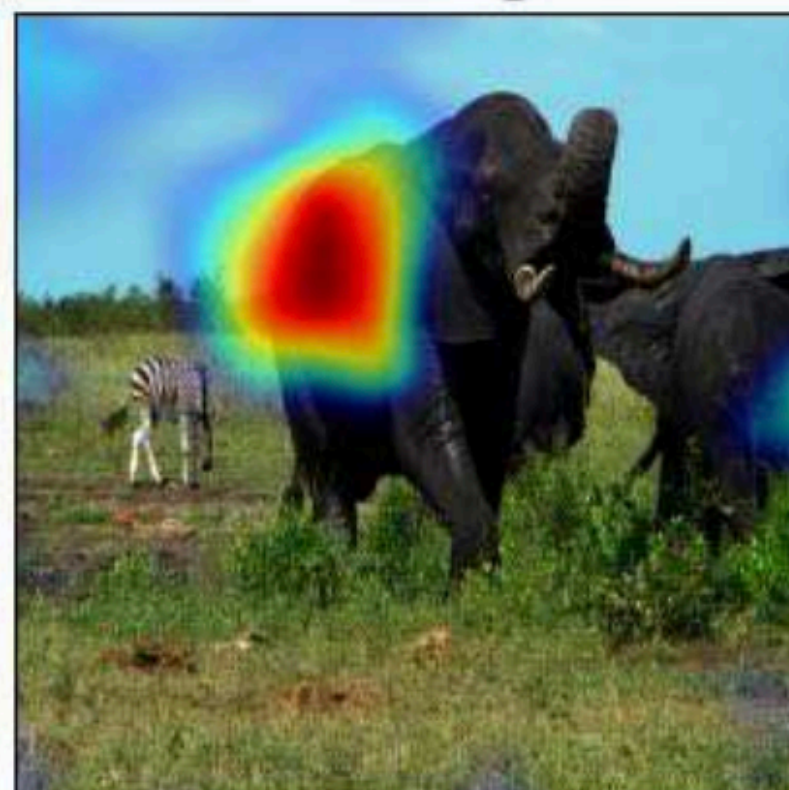
What is the she holding?



Baseball bat



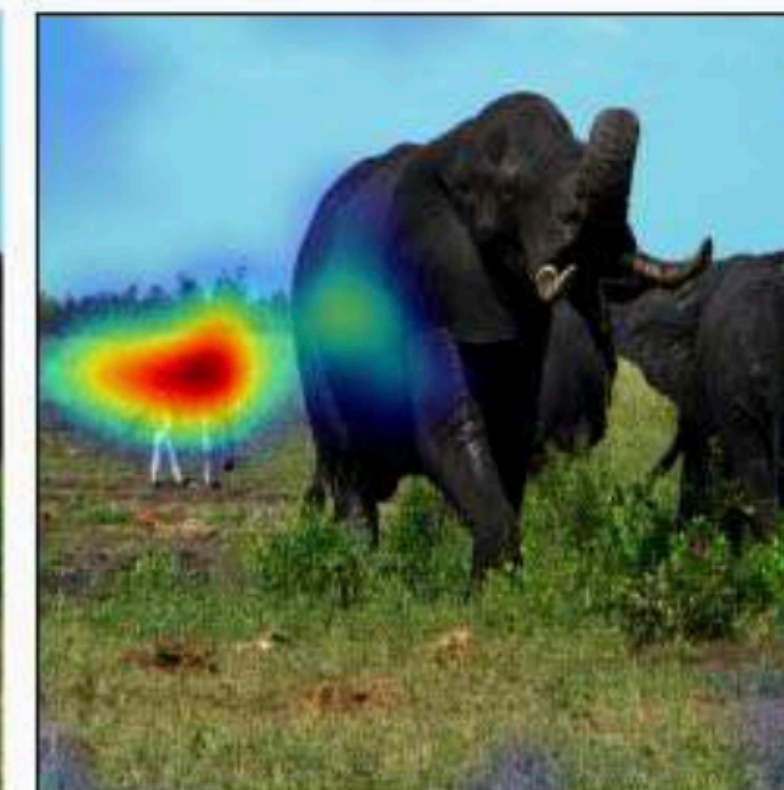
What is that?



Elephant



What is that?



Zebra



A house with a roof

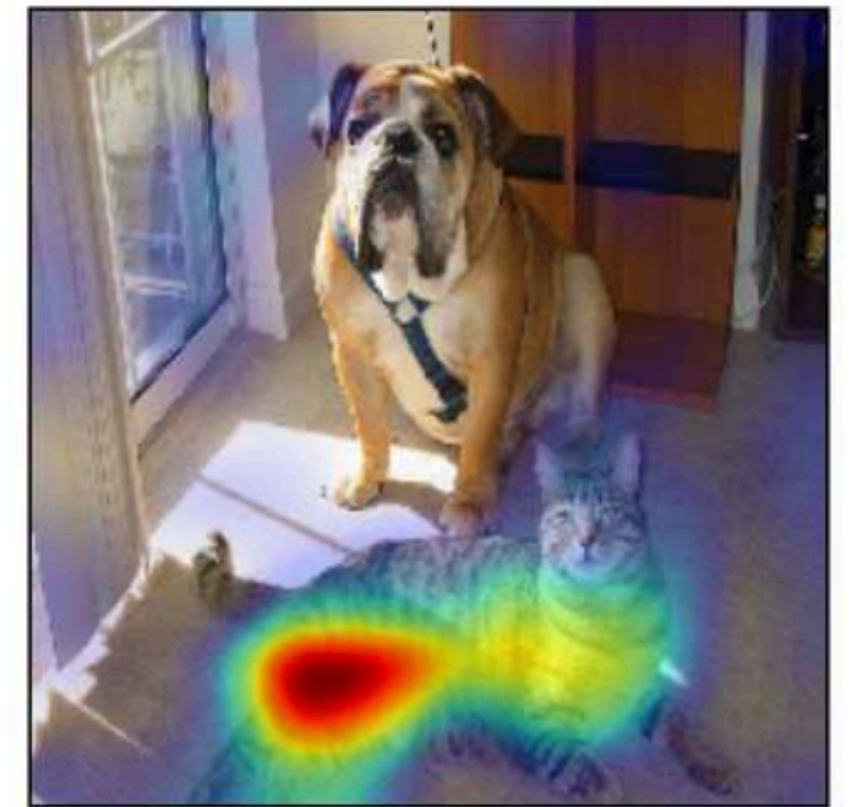
(b) Visualizing ResNet based Hierarchical co-attention VQA model from [29]



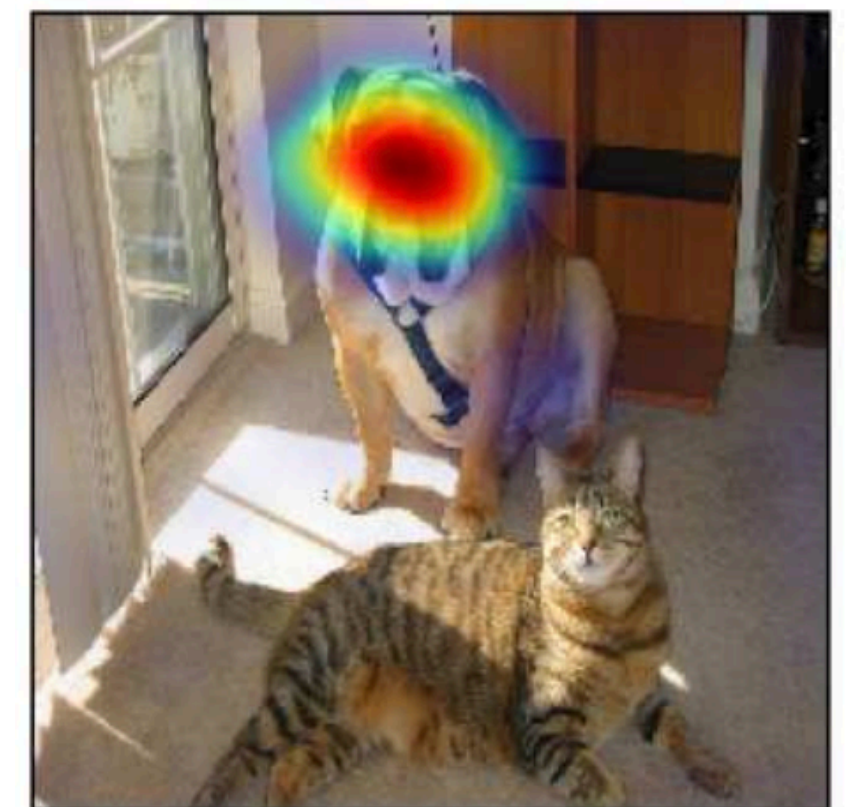
# Guided Grad-CAM

## Motivation

- Grad-CAM provides good localization, but it lacks fine-grained detail
- In this example, it can easily localize cat
- However, it doesn't explain why the cat is labeled as 'tiger cat'
- Point-wise multiplying guided backpropagation and Grad-CAM visualizations solves the issue



(c) Grad-CAM 'Cat'



(i) Grad-CAM 'Dog'



# Guided Grad-CAM

## How it works

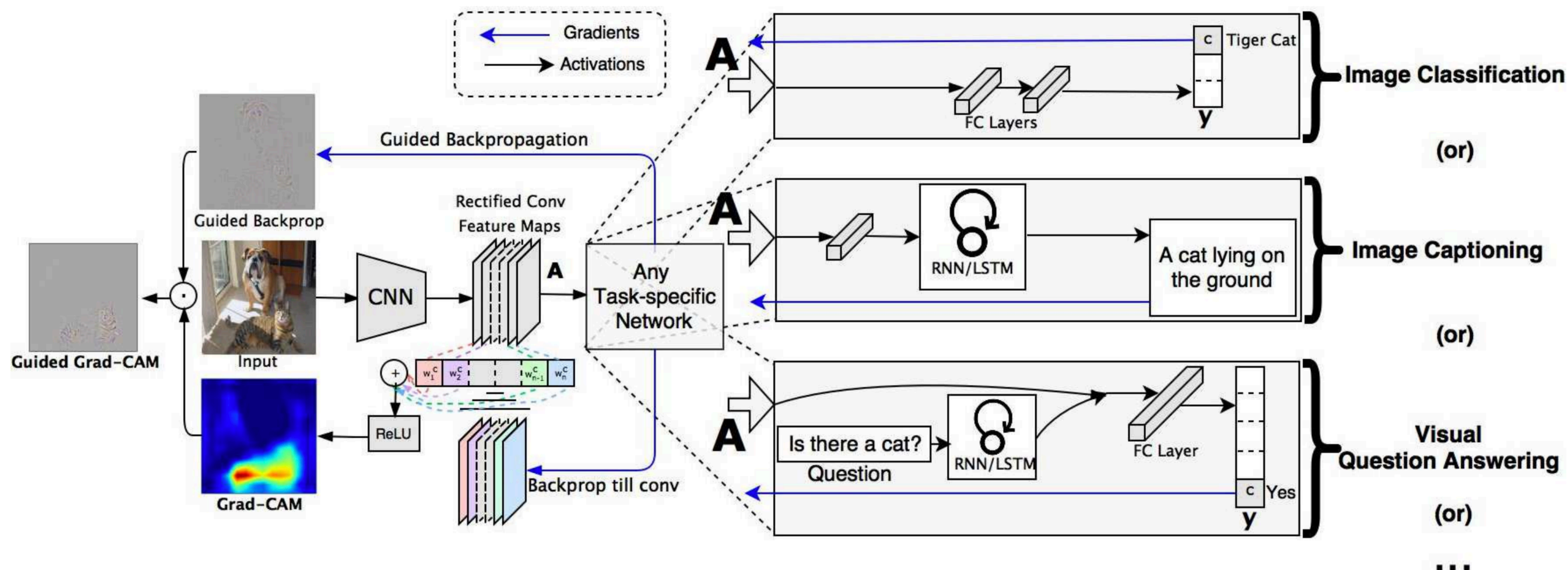
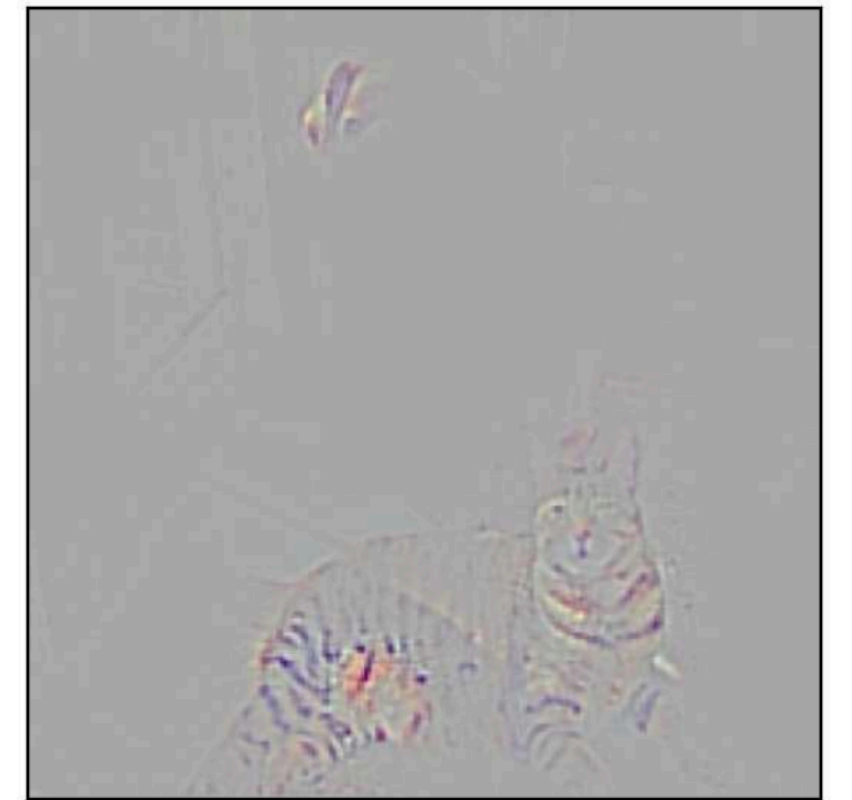


Figure 2: Grad-CAM overview: Given an image and a class of interest (e.g., 'tiger cat' or any other type of differentiable output) as input, we forward propagate the image through the CNN part of the model and then through task-specific computations to obtain a raw score for the category. The gradients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

# Guided Grad-CAM

## Results

- With Guided Grad-CAM, it becomes easier to see which details went into decision making
- For example, we can now see the stripes and pointed ears by using the model predicted it as ‘tiger cat’



(d) Guided Grad-CAM ‘Cat’



(j) Guided Grad-CAM ‘Dog’

# Evaluations

## Localization

- Given an image, first obtain class predictions from the network
- Generate Grad-CAM maps for each of the predicted classes
- Binarize with threshold of 15% of max intensity
- Draw bounding box around single largest connected segment of pixels



# Evaluations

## Localization

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).

# Evaluations

## Class Discrimination

- Evaluated over images from VOC 2007 val set that contain 2 annotated categories, and create visualizations for each of them
- For both VGG-16 and AlexNet CNNs, category-specific visualizations are obtained using four techniques:
  - Deconvolution
  - Guided Backpropagation
  - Deconvolution with Grad-CAM
  - Guided Backpropagation with Grad-CAM



# Evaluations

## Class Discrimination

- 43 workers on AMT were asked “Which of the two object categories is depicted in the image?”
- The experiment was conducted for all 4 visualizations, for 90 image-category pairs
- A good prediction explanation should produce distinctive visualizations for each class of interest

**What do you see?**



**Your options:**

- ☐ Horse
- ☐ Person

# Evaluations

## Class Discrimination

Model	Accuracy(%)
Deconvolution	53.33
Deconvolution + Grad-CAM	61.23
Guided Backpropagation	44.44
Guided Backpropagation + Grad-CAM	61.23

# Evaluations

## Trust - Why is it needed?

- Given two models with the same predictions, which model is more trustworthy?
- Visualize the results to see which parts of the image are being used to make the decision!

# Evaluations

## Trust - Experimental Setup

- Use AlexNet and VGG-16 to compare Guided Backprop and Guided Grad-CAM visualizations
- Note that VGG-16 is more accurate (79.09mAP vs 69.20)
- Only those instances considered where both models make same prediction as ground truth

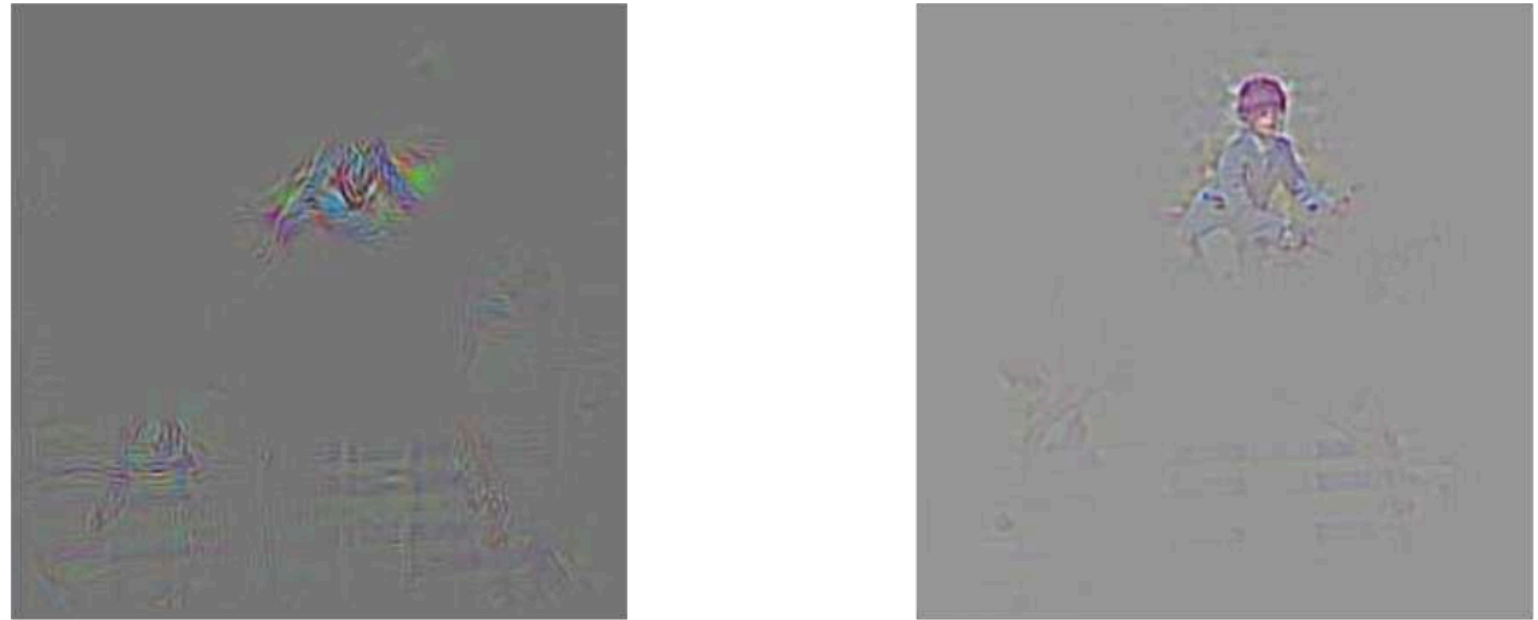
# Evaluations

## Trust - Experimental Setup

- Given visualizations from both models, 54 AMT workers were asked to rate reliability of the two models as follows
  - 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**

# Evaluations

## Trust - Result

- Humans are able to identify the more accurate classifier, despite identical class predictions
- With Guided Backpropagation, VGG was assigned a score of 1.0
- With Guided Grad-CAM, it achieved a higher score of 1.27
- Thus, the visualization can help place trust in a model which will generalize better, just based on individual predictions

# Evaluations

## Faithfulness vs Interpretability

- Faithfulness of a visualization to a model is defined as its ability to explain the function learned by the model
- There exists a trade-off between faithfulness and interpretability
- A fully faithful explanation is the entire description of the model, which would make it not interpretable/easy to visualize
- In previous sections, we saw that Grad-CAM is easily interpretable

# Evaluations

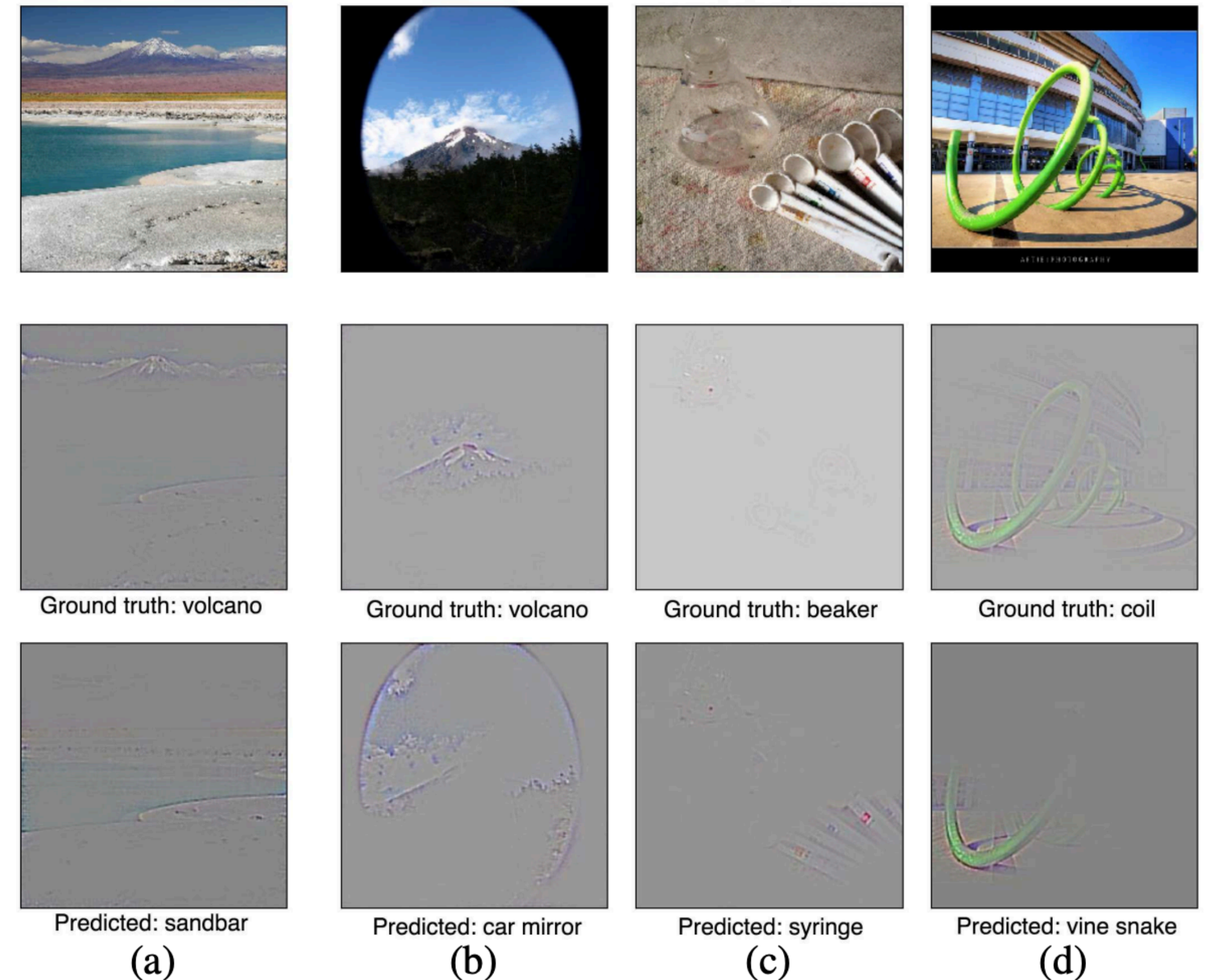
## Faithfulness vs Interpretability

- Explanations should be locally accurate
- For reference explanation, one choice is image occlusion
- CNN scores are measured when patches of the input image are masked
- Patches which change CNN scores are also patches which are assigned high intensity by Grad-CAM and Guided Grad-CAM
- Rank correlation of 0.261 achieved over 2510 images in PASCAL 2007 val set



# Analyzing Failure Modes for VGG-16

- In order to see what mistakes a network is making, first collect the misclassified examples
- Visualize both the ground truth class as well as the predicted class
- Some failures are due to ambiguities inherent in the dataset
- Seemingly unreasonable predictions have reasonable explanations



# Identifying Bias in Dataset

- Fine-tuned an ImageNet trained VGG-16 model for the task of classifying “Doctors” vs “Nurses”
- Used top 250 relevant images from a popular image search engine
- Trained model achieved good validation accuracy, but didn’t generalize well(82%)
- Visualizations helped to see that the model had learnt to look at the person’s face/hairstyle to make the predictions, thus learning gender stereotypes

# Identifying Bias in Dataset

- Image search results were 78% male doctors, and 93% female nurses
- Through this intuition, we can reduce bias by adding more examples of female doctors, as well as male nurses
- Retrained model generalizes well (90% test accuracy)
- This experiment helps demonstrate that Grad-CAM can help detect and remove biases from the dataset, thus making fair and ethical decisions

# Image Captioning

- Build Grad-CAM over a public available neuraltalk2 implementation, which uses VGG-16 CNN for images and an LSTM-based language model
- Given a caption, compute gradient of its log-probability wrt units in the last convolutional layer of the CNN



# Image-Captioning

## How it works

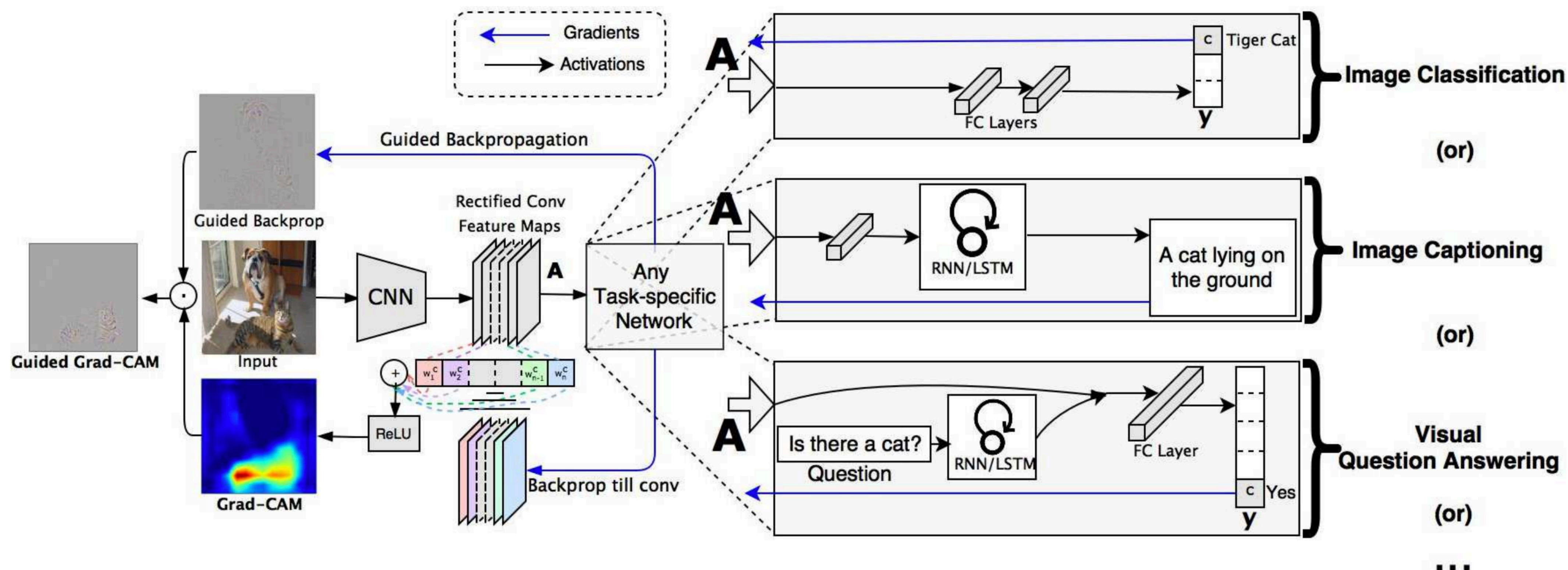
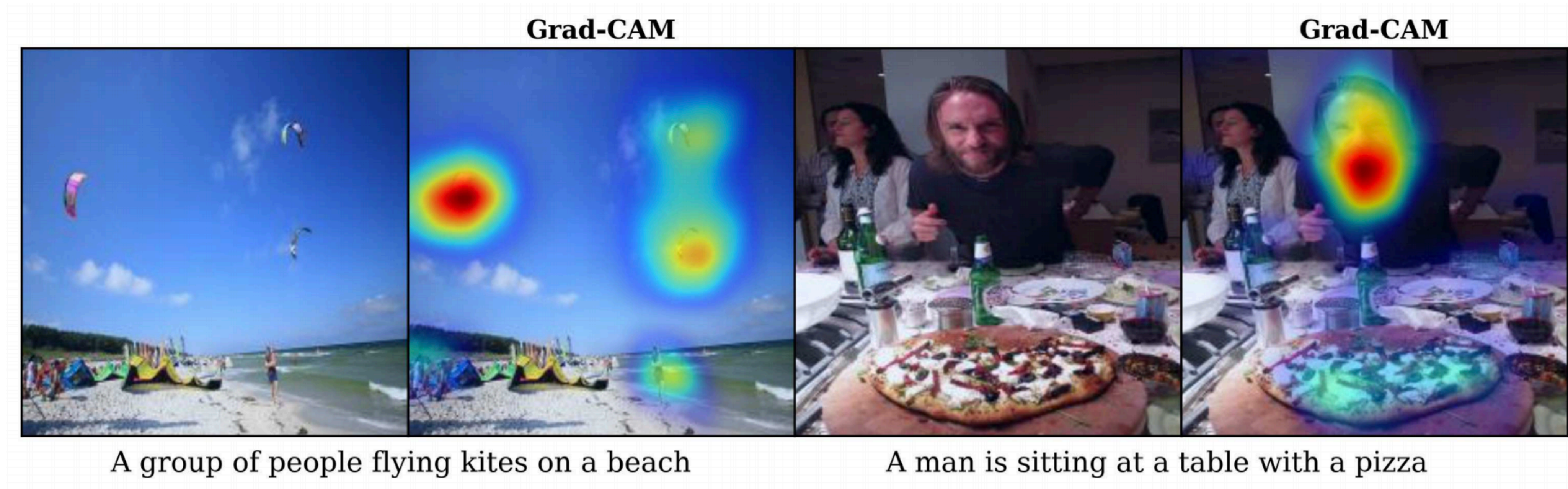


Figure 2: Grad-CAM overview: Given an image and a class of interest (e.g., ‘tiger cat’ or any other type of differentiable output) as input, we forward propagate the image through the CNN part of the model and then through task-specific computations to obtain a raw score for the category. The gradients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



(a) Image captioning explanations

# Image Captioning

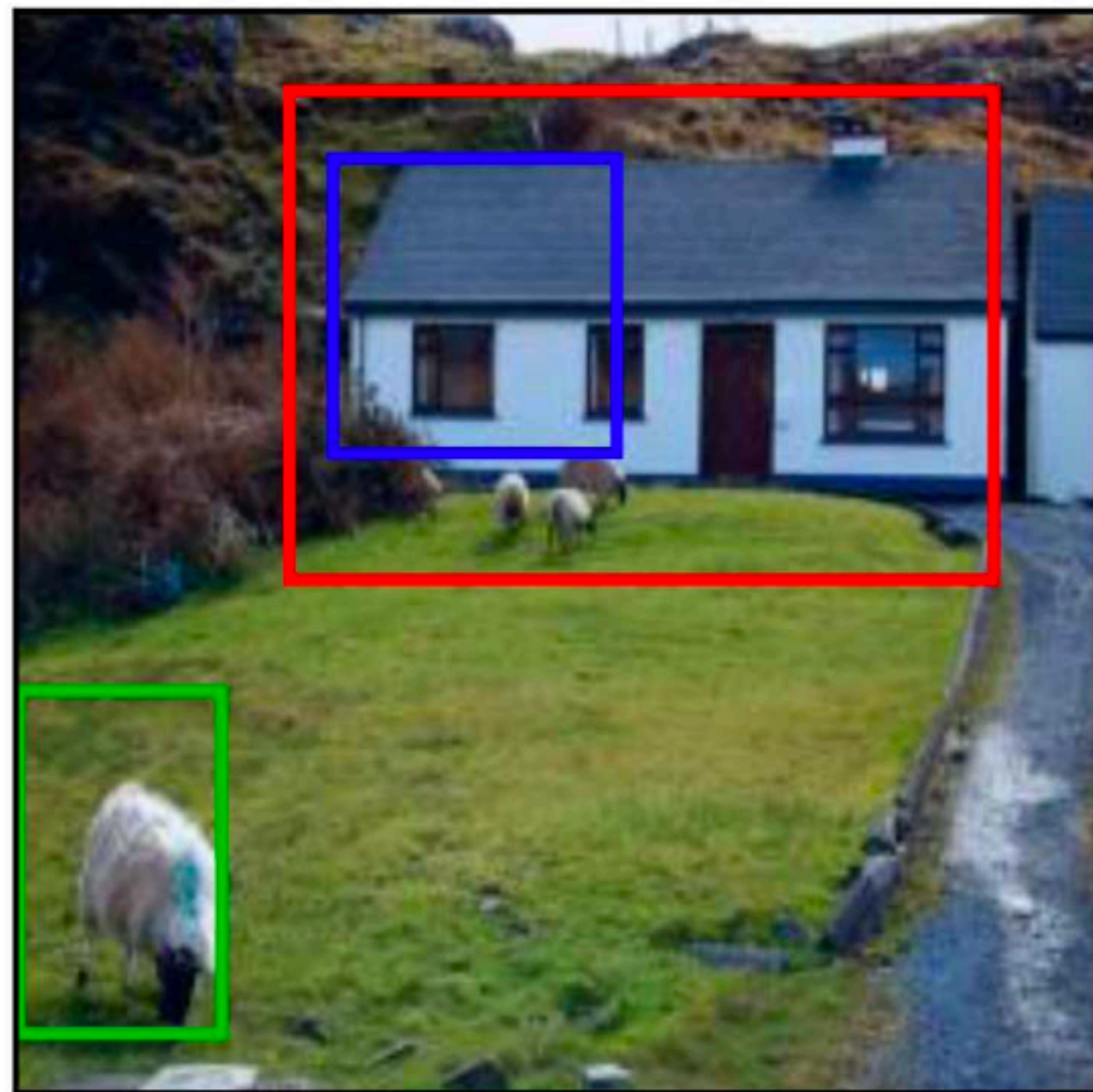
## Comparison to Dense Cap

- Dense Captioning task requires a system to jointly localize and caption salient regions of the image
- Johnson et. al.'s model consists of a Fully Connected Localization Network (FCLN) and an LSTM based language model
- It produces bounding boxes and associated captions in a single forward pass
- Using DenseCap, generate 5 region-specific captions with associated bounding boxes
- A whole-image captioning model should localize the caption inside the bounding box it was generated for



# Image Captioning

## Comparison to Dense Cap



A house with a green roof

Sheep grazing in field

A house with a roof



# Image Captioning

## Comparison to Dense Cap

- Measured by computing the ratio of average activation inside vs outside the box
- Uniformly highlighting the whole image gives a baseline of 1.0
- Grad-CAM achieves  $3.27 \pm 0.18$
- Guided Backpropagation(adding high resolution detail) gives  $2.32 \pm 0.08$
- Best localization seen for Guided Grad-CAM at  $6.38 \pm 0.99$

# Visual Question Answering

- Typical VQA pipelines consist of a CNN to model images and an RNN language model for questions
- Image and question representations are fused to predict the answer as a 1000 way classification problem
- Thus, we can take the scores  $y_c$  for for the answer and use that to compute Grad-CAM to show image evidence that supports the answer
- Despite the complexity, the results are surprisingly intuitive

# Visual Question Answering

## How it works

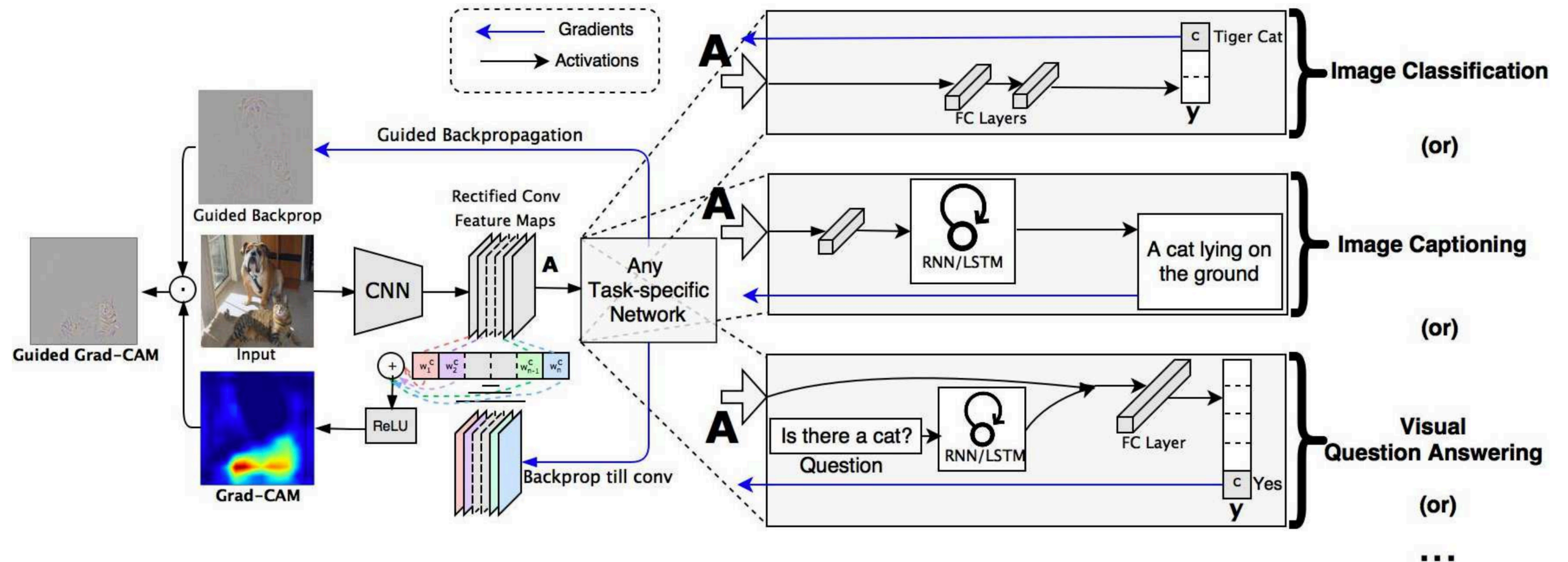


Figure 2: Grad-CAM overview: Given an image and a class of interest (e.g., 'tiger cat' or any other type of differentiable output) as input, we forward propagate the image through the CNN part of the model and then through task-specific computations to obtain a raw score for the category. The gradients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

# Visual Question Answering

## Comparison to Human Attention Maps

- Das et. al collected human attention maps for a subset of VQA dataset
- These maps have high intensity where humans looked in the image in order to answer a visual question
- Human attention maps are compared to Grad-CAM visualizations on 1374 val QI pairs using the rank correlation evaluation protocol
- They have a correlation of 0.136, which is statistically higher than chance or random attention maps (zero correlation)
- This shows that even non-attention based VQA models are surprisingly good at localizing regions required to output a particular answer

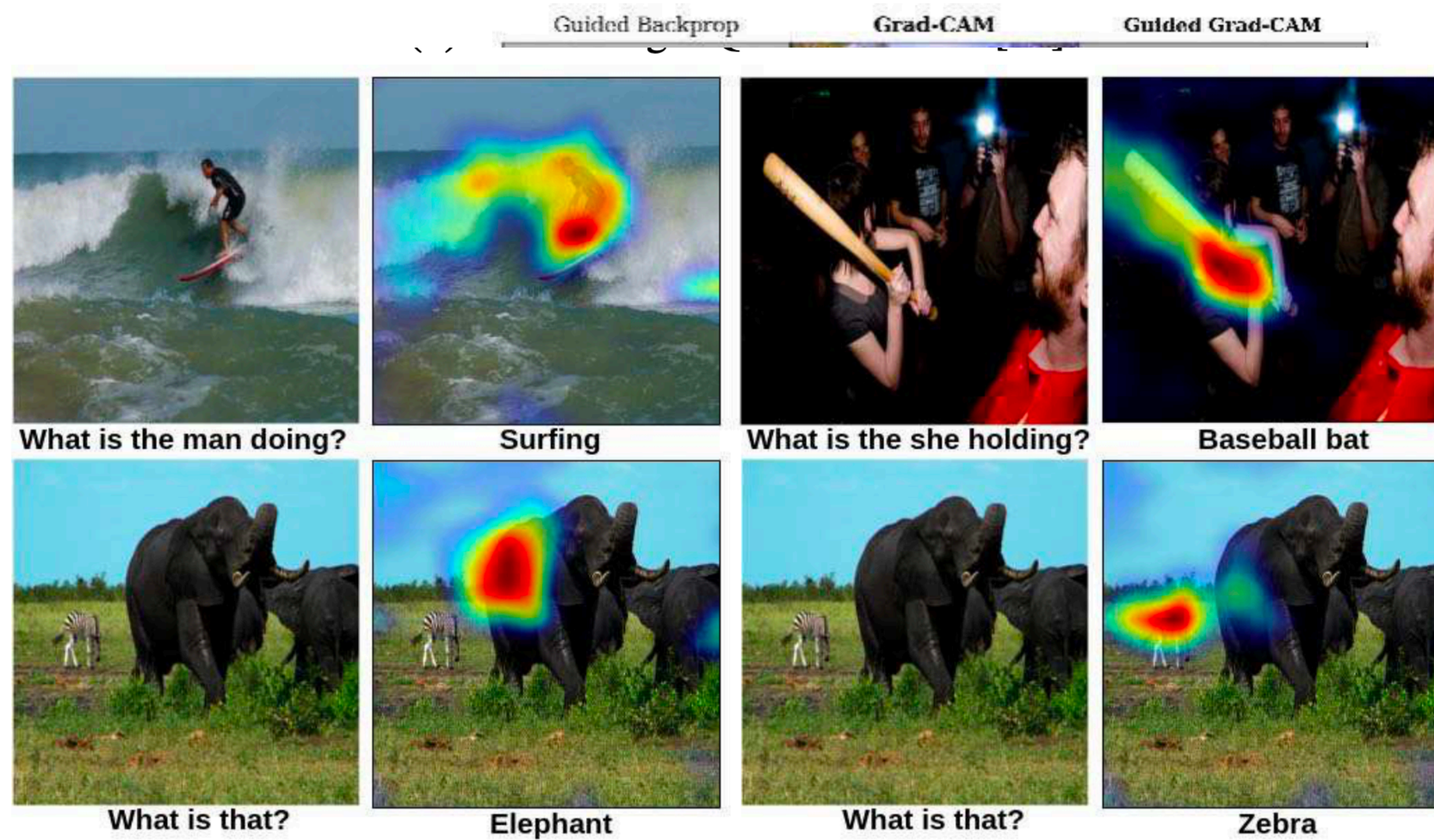
# Visual Question Answering

## Visualizing ResNet-based VQA model with attention

- Lu et. al use a 200 layer ResNet to encode the image and jointly learn a hierarchical attention mechanism on the question and the image
- As we visualize deeper layers, we find small changes for most adjacent layers, but larger changes for layers which involve dimensionality reduction
- This shows that the same approach works for even complicated models



# Visual Question Answering



(b) Visualizing ResNet based Hierarchical co-attention VQA model from [29]

(a) Visualizing VQA model from [28]



# Conclusion

- Proposed a novel class-discriminative localization technique - Grad-CAM
- Works for any CNN based architecture, without having to modify the network
- Combined Grad-CAM localizations with existing high-resolution visualizations
- Outperforms all existing approaches on both interpretability and faithfulness
- Extensive human studies reveal that visualizations can discriminate between classes more accurately, better reveal trustworthiness, and help identify biases
- Showed the broad applicability to off-the-shelf architectures

# Questions?



**Thank You!**