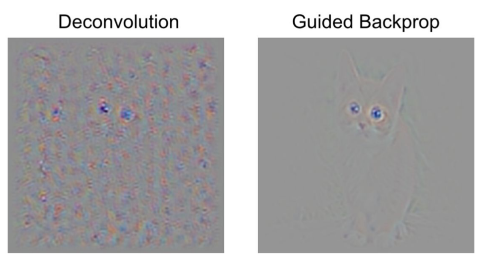
Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

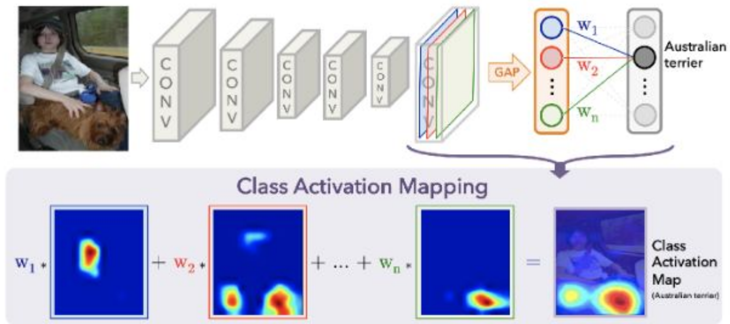
Maulik Shah, Yunjia Zhang

1. Introduction
    a. Explaining deep networks is hard!
    b. What makes a good interpretation?
        - Class discriminative - localize the category in the image
        - High resolution - capture fine-grained detail in the image
2. Related works
    a. CNN visualization
        i. Guided backpropagation
        ii. Deconvolution
        Here are some plots shown for deconvolution and guided backprop. We can see that the results from deconvolution if not clear enough to identify the object, while guided backprop can provide better resolution.



    b. Assessing model trust
    c. Weakly supervised localization
    d. CAM: Class activation mapping, our Grad-CAM is a generalization of CAM
3. CAM and Grad-CAM approach
    a. What is Grad-CAM:
        i. Enables classification CNNs to learn to perform localization
        ii. CAM indicates the discriminative regions used to identify that category
        iii. No explicit bounding box annotations required
        iv. However, it needs to change the model architecture
        Just before the final output layer, they perform global average pooling on the convolutional feature maps. And Use these features for a fully-connected layer that produces the desired output.
    b. How does CAM work:
        i. $f_k(x, y)$: Calculate activation of unit k in spatial location (x, y)
        ii. $F^k = \sum_{x,y} f_k(x, y)$: Result of global average pooling
        iii. $S_c = \sum_{k} w_k^c F_k$: input to Softmax layer for class c

iv. $M_c(x,y) = \sum_k w_k^c f_k(x,y)$: calculate CAM for class c



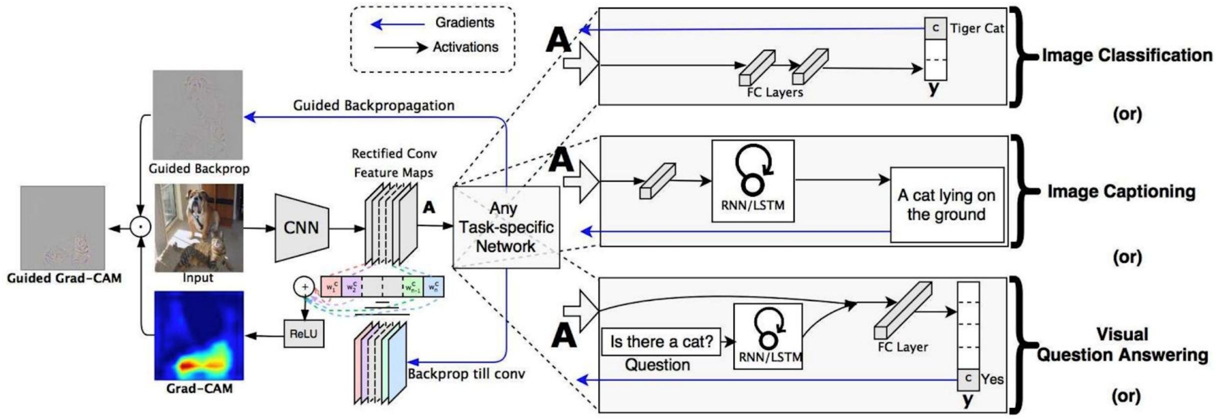
c. Drawback of CAM:
  - Requires features maps to directly precede softmax layers
  - Such architectures may achieve inferior accuracies compared to general networks on other tasks
  - Need a method that doesn't need any modification to existing architecture

d. How does Grad-CAM work:
  i. 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

  ii. 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

  iii. Approach
    - Compute gradient $\frac{\partial y^c}{\partial A^k}$: gradient of score y^c for class c wrt feature maps A^k
    - Global average pool these gradients to obtain neuron importance weights: $\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{i,j}^k}$

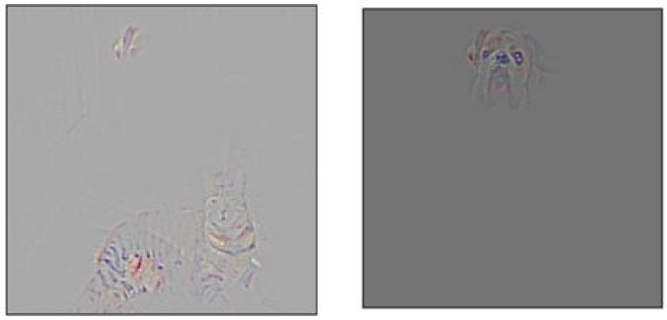
- Perform weighted combination of forward activations maps and follow it by ReLU to obtain: $$L^c_{Grad-CAM} = ReLU\left(\sum_k \alpha^c_k A^k\right)$$



4. Guided-Grad-CAM
   a. 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'

   b. How it works
      - It produces backward propagation on the neural network to get guided backprop.
      - Then pointwise multiplication with Grad-CAM to generate Guided Grad-CAM
   c. Some 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'

5. Experimental evaluation
    a. 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

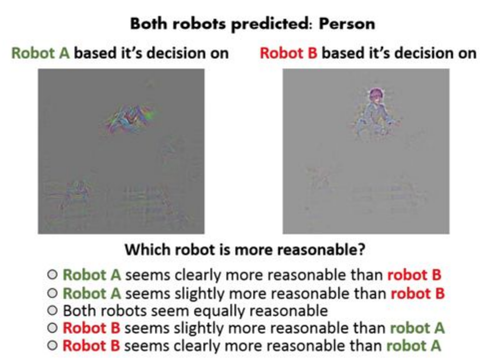
Method	Top-1 loc error	Top-5 loc error	Top-1 cls error	Top-5 cls error
Backprop on VGG-16 [40]	61.12	51.46	30.38	10.89
c-MWP on VGG-16 [46]	70.92	63.04	30.38	10.89
Grad-CAM on VGG-16 (ours)	56.51	46.41	30.38	10.89
VGG-16-GAP (CAM) [47]	57.20	45.14	33.40	12.20

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

    b. 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
        - 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

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

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

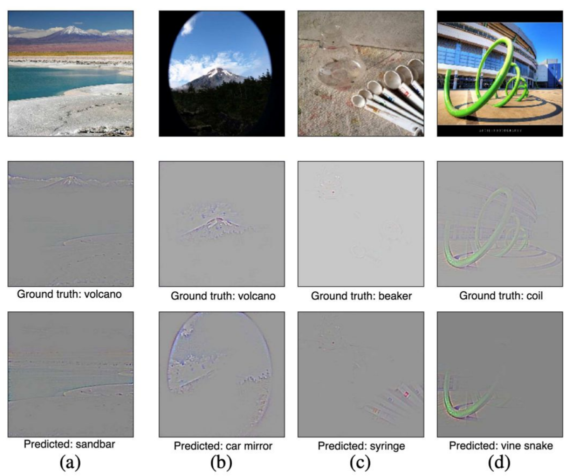


   - Results:
     - Humans are able to identify the more accurate classifier, despite identical class predictions
     - With Guided Backpropagation, VGG was assigned a score of 1.0
     - With Guided Grad-CAM, it achieved a higher score of 1.27
     - Thus, the visualization can help place trust in a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

d. 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
- 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

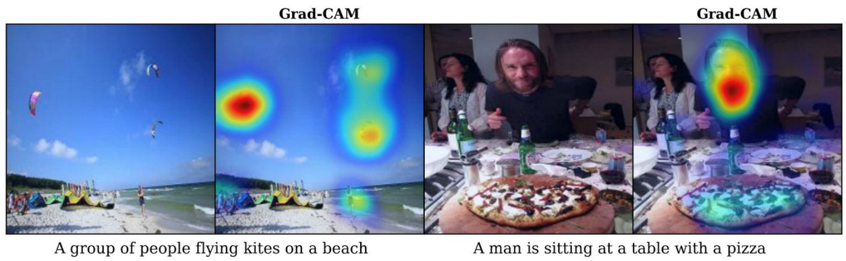
e. Identifying failure modes:
- In order to see what mistakes a network is making, first collect the misclassified examples
- Visualize both the ground truth class as well as the predicted class
- Some failures are due to ambiguities inherent in the dataset
- Seemingly unreasonable predictions have reasonable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



f. 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
- 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

g. 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



(a) Image captioning explanations

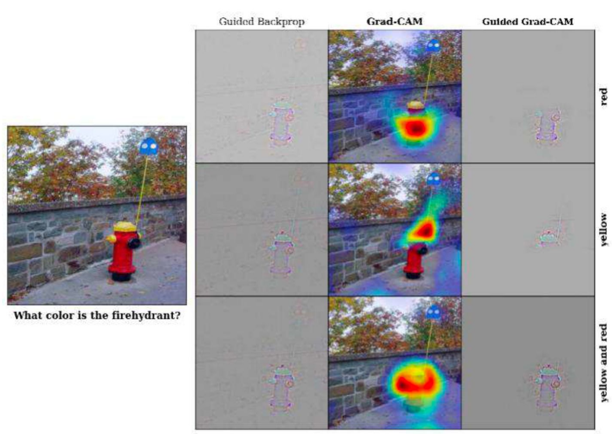
- Compared with Dense Cap:
    - Dense Captioning task requires a system to jointly localize and caption salient regions of the image
    - Johnson et. al.'s model consists of a Fully Connected Localization Network (FCLN) and an LSTM based language model
    - It produces bounding boxes and associated 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

    - 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 ± 0.18
    - Guided Backpropagation(adding high resolution detail) gives 2.32 ± 0.08
    - Best localization seen for Guided Grad-CAM at 6.38 ± 0.99

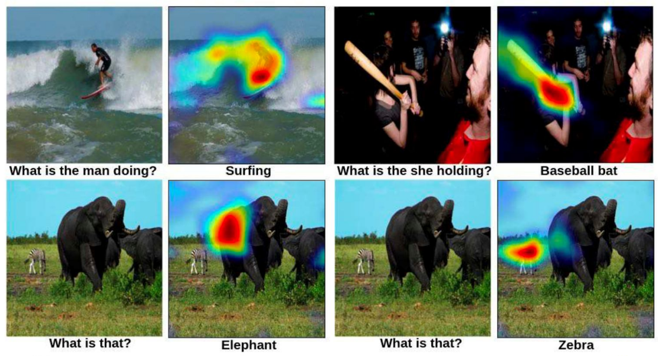
h. Visual Question Answering (VQA)

- 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
- Compared with 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
- 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



(a) Visualizing VQA model from [28]



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

6. 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

Discussion

1. Question: Why is CAM restricted to image classification problems?
   Answer: In the architecture of CAM, it used a Softmax layer as the last layer to do the prediction. The corresponding output values of Softmax layer is the probability of each of the final classes. So CAM is restricted to image classification problems. However, Grad-CAM can attach any task-specific neural network at the end of CNN based subnetwork, so it can be applied to any CNN based off-the-shelf models.

2. Question: While you are using the test accuracy to evaluate the faithfulness of Grad-CAM, how can we use something else to understand whether it is performing unfaithfully?
   Answer: They provide the distribution of the test set, i.e. the doctors example, having more male doctors and female nurses. My guess is that in that they found it (the bias of the data) out by looking at the area that CNN is looking at. This visualizaiton can help with telling whether the model can perform faithfully.

3. Question: In the figure for VQA (Visual Question Answering), given the same picture, how do they induce different answers?
   Answer: The model has a softmax layer at the top, so we have different confidence for each of the classes. Based on each of the scores, we can have the heatmap for different classes.