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

CONTENTS

目录

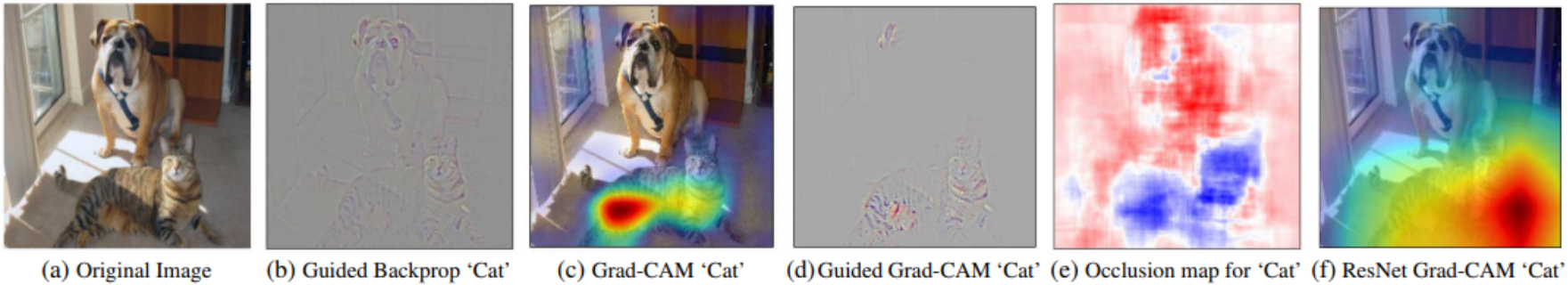
Contribution

Contribution

- **high**-resolution class-**discriminative** visualization
- applicable to a wide variety of CNN model-families :

  ResNet 、CNN+LSTM 、FCLN
- without architectural changes or re-training

Contribution



(a) Original Image (b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat' (d) Guided Grad-CAM 'Cat' (e) Occlusion map for 'Cat' (f) ResNet Grad-CAM 'Cat'

(g) Original Image (h) Guided Backprop 'Dog' (i) Grad-CAM 'Dog' (j) Guided Grad-CAM 'Dog' (k) Occlusion map for 'Dog' (l) ResNet Grad-CAM 'Dog'

Figure 1: (a) Original image with a cat and a dog. (b-f) Support for the cat category according to various visualizations for VGG-16 and ResNet. (b) Guided Backpropagation [42]: highlights all contributing features. (c, f) Grad-CAM (Ours): localizes class-discriminative regions. (d) Combining (b) and (c) gives Guided Grad-CAM, which gives high-resolution class-discriminative visualizations. Interestingly, the localizations achieved by our Grad-CAM technique, (c) are very similar to results from occlusion sensitivity (e), while being orders of magnitude cheaper to compute. (f, l) are Grad-CAM visualizations for ResNet-18 layer. Note that in (c, f, i, l), red regions corresponds to high score for class, while in (e, k), blue corresponds to evidence for the class. Figure best viewed in color.

2 Guided-BackProp

Saliency Map（BP）

train or vis

$$S_c(I) \approx w^T I + b,$$

一阶泰勒近似

$$w = \left.\frac{\partial S_c}{\partial I}\right|_{I_0}$$

pixel importance

$$M_{ij} = |w_{h(i,j)}|$$

$$M \in \mathcal{R}^{m \times n}$$

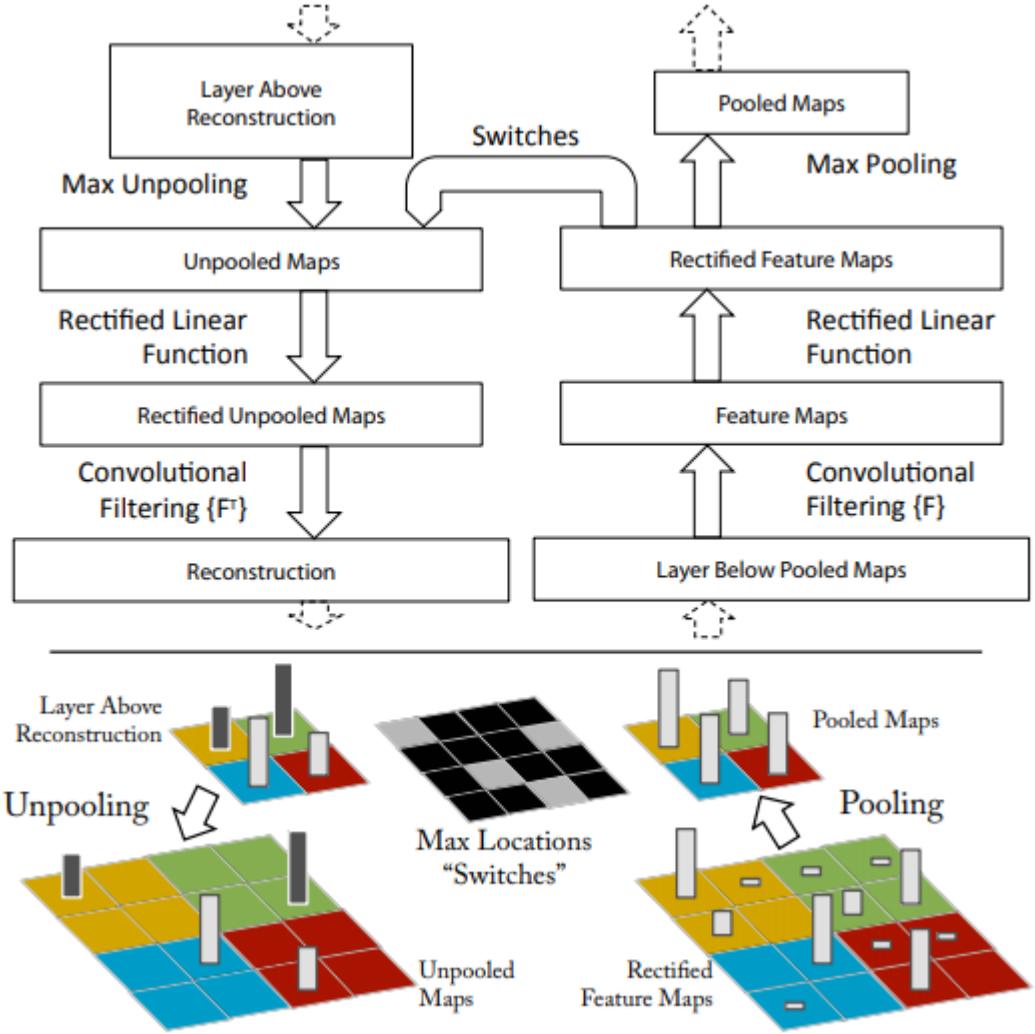
$$M_{ij} = \max_c |w_{h(i,j,c)}|$$

channel

Deconvolution

- **Unpooling**
- **RELU**
- **Deconv**

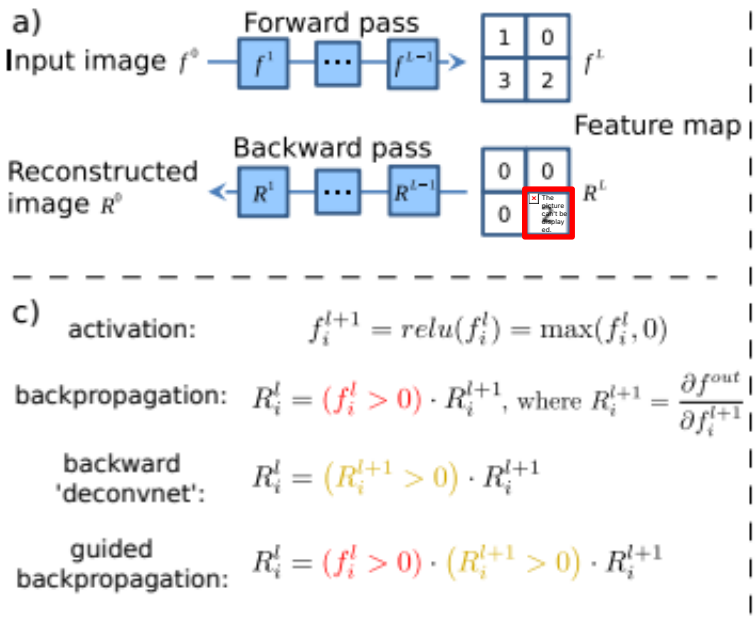
- img→conv→relu→maxpool →unpool→relu→deconv
- switches中存储着max pool中的每一个池化块的最大值的坐标

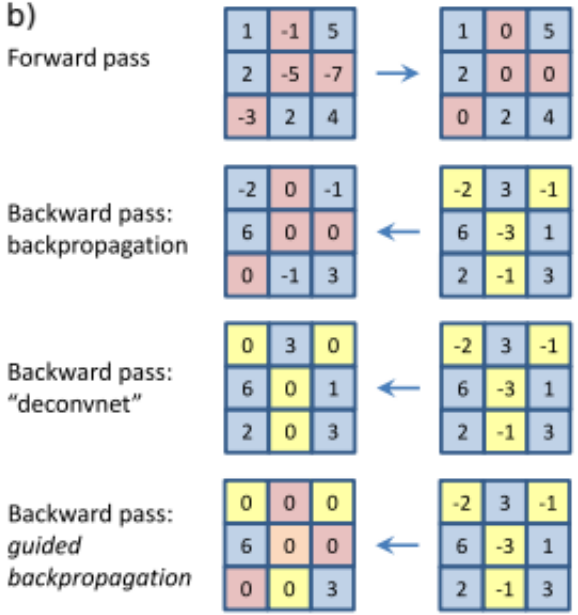


Guided-BackProp

bottom-up signal

RELU





3 Grad-CAM

CAM

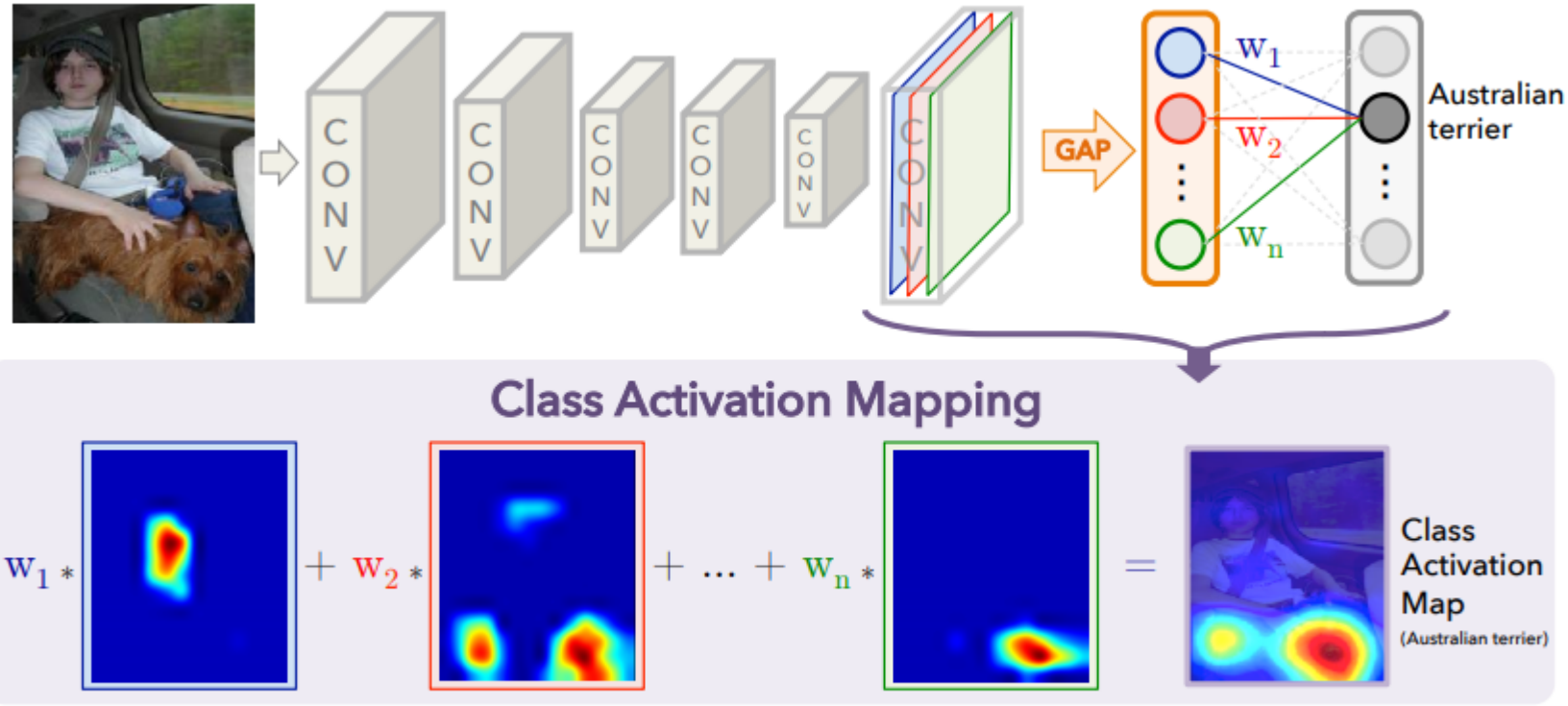


Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

CAM

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

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

Grad-CAM

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

第k个channel

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

positive effect

4 Guided Grad-CAM

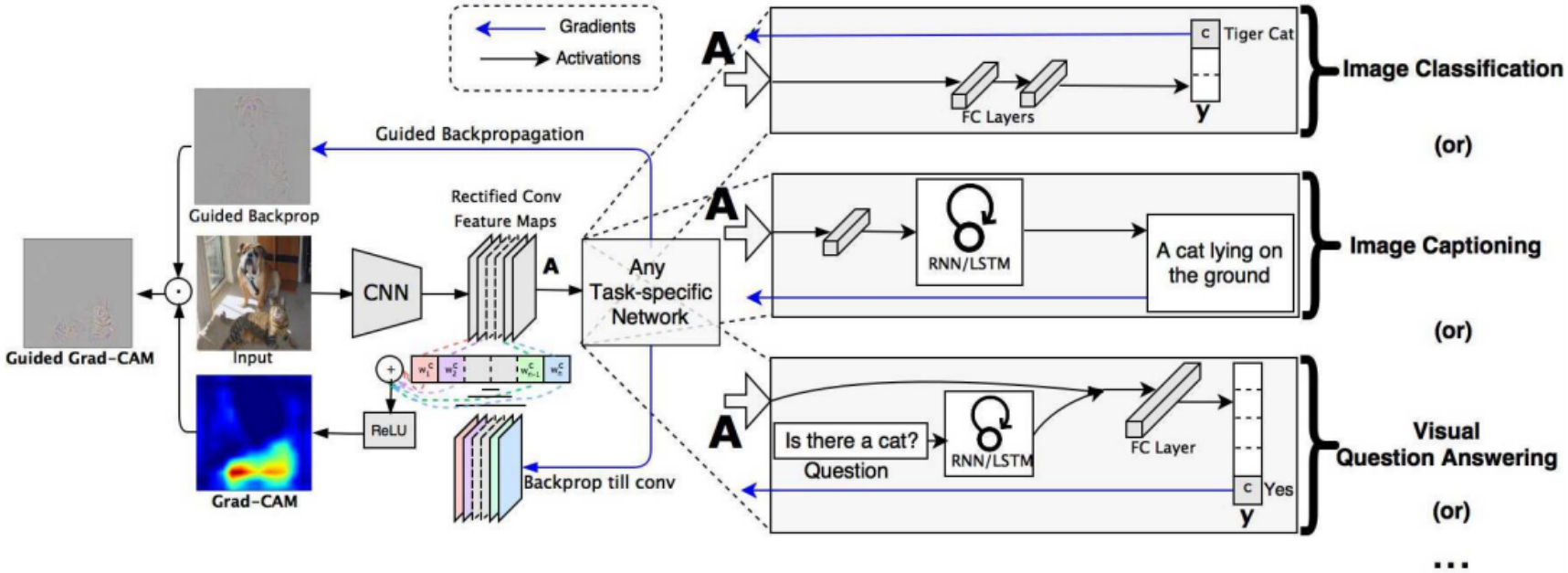


Figure 2: Grad-CAM overview: Given an image and a class of interest (*e.g.*, 'tiger cat' or any other type of differentiable output) as input, we forward propagate the image through the CNN part of the model and then through task-specific computations to obtain a raw score for the category. The gradients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

THANKS!

LIVE AND LEARN