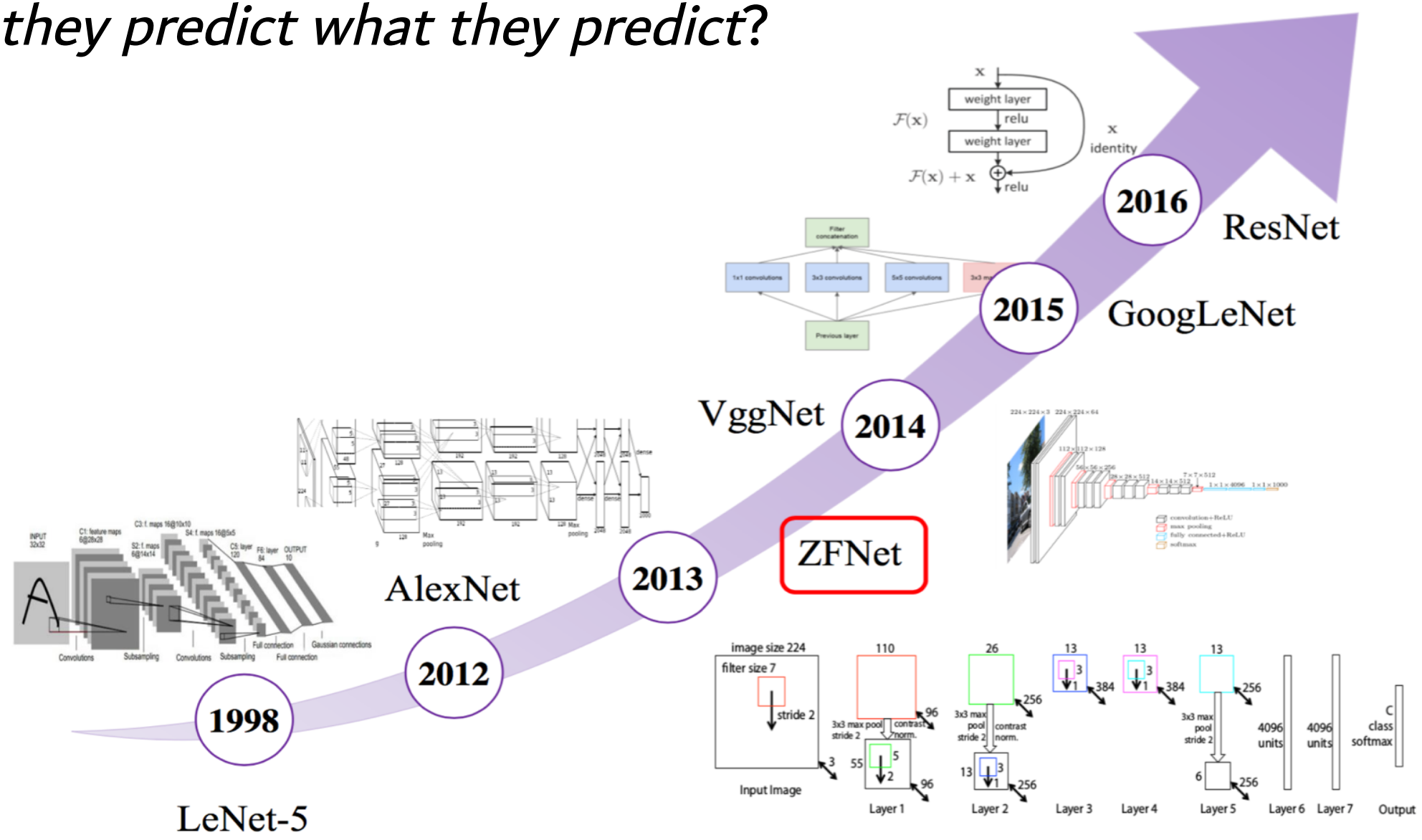
Grad-CAM:
Visual Explanations from Deep Networks via Gradient-based Localization

Ramprasaath R. Selvaraju[1*] Michael Cogswell[1] Abhishek Das[1] Ramakrishna Vedantam[1*]
Devi Parikh[1,2] Dhruv Batra[1,2]
[1]Georgia Institute of Technology [2]Facebook AI Research
{ramprs, cogswell, abhshkdz, vrama, parikh, dbatra}@gatech.edu

ICCV 2017

报告人：李帆
2020.01.02

why they predict what they predict?



Outline

Approch

- Grad-CAM
- Guided Grad-CAM

Experiments

- Evaluating Localization
- Evaluating Visualizations
- Diagnosing image classification CNNs
- Image Captioning and VQA

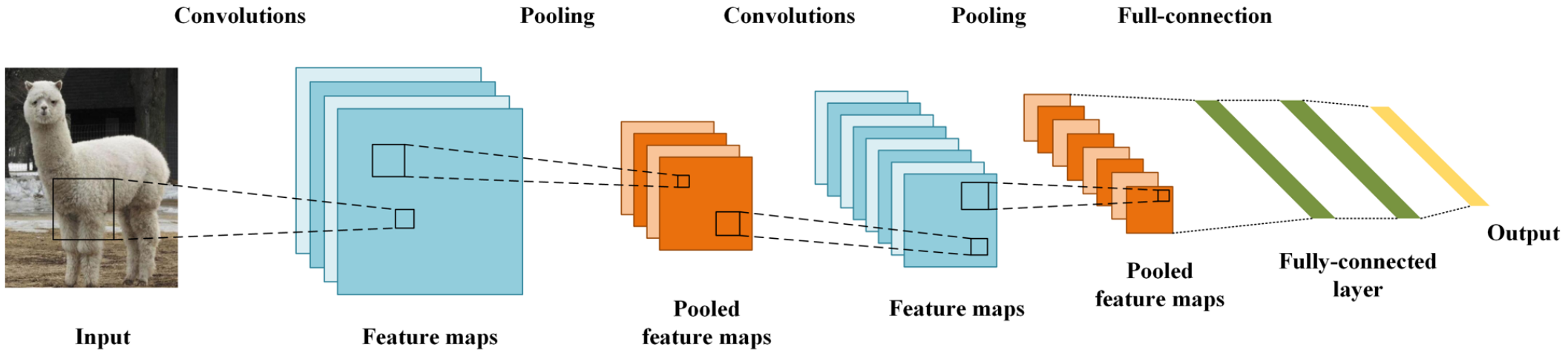
Outline

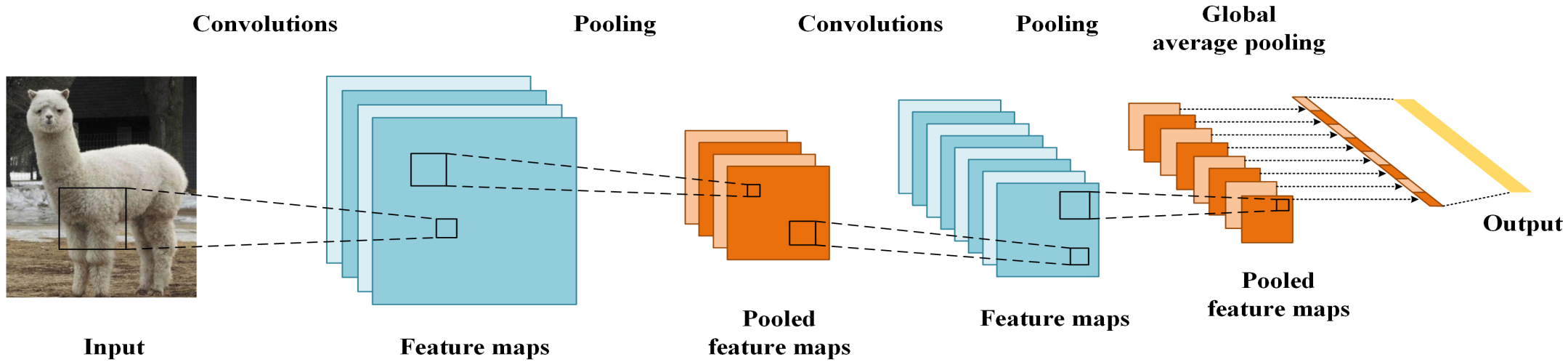
Approch

- Grad-CAM
  - a generalization to CAM
  - applicable to a wide variety of CNN model-families without architectural changes or re-training
- Guided Grad-CAM

Experiments

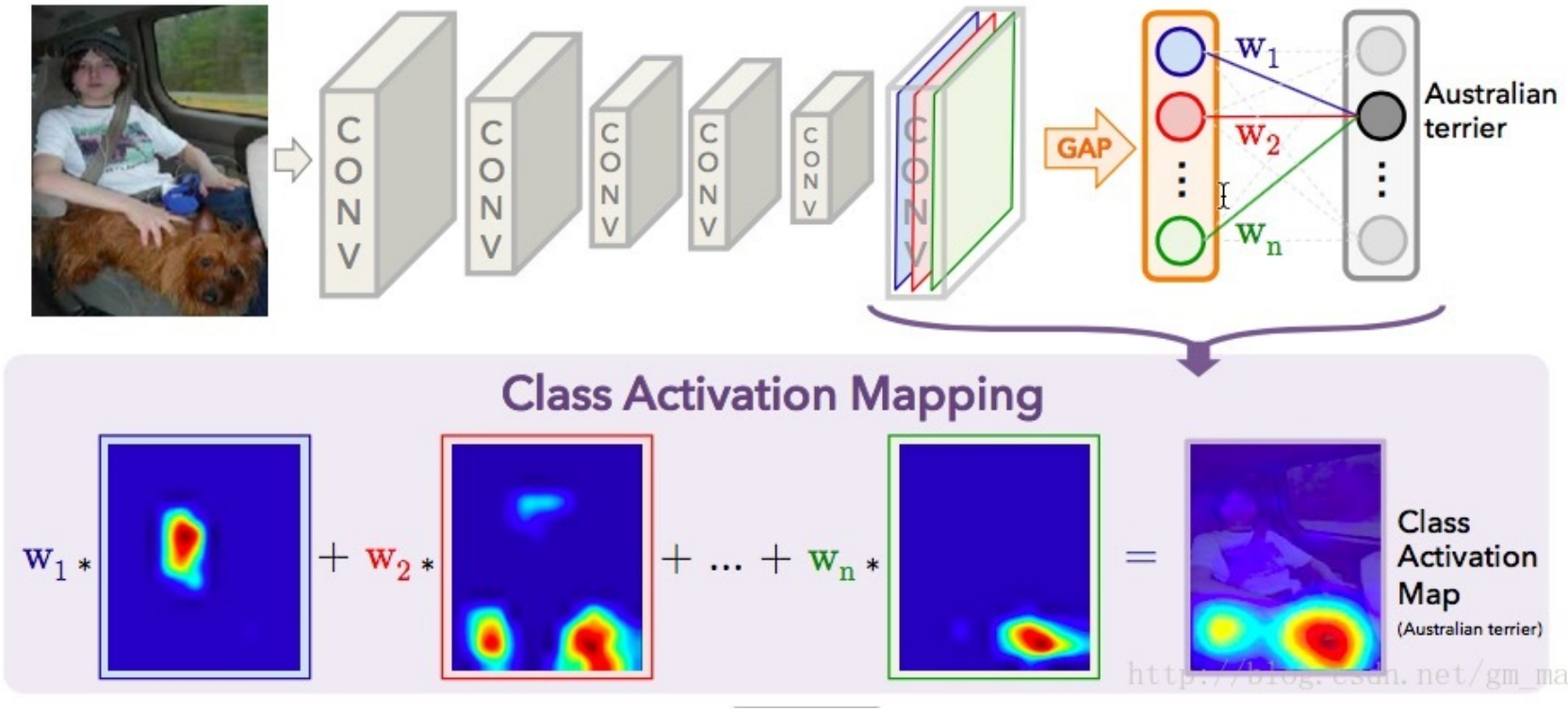
- Evaluating Localization
- Evaluating Visualizations
- Diagnosing image classification CNNs
- Image Captioning and VQA





Global Average Pooling(GAP) : Lin, Min, Qiang Chen, and Shuicheng Yan. "Network in network." *arXiv preprint arXiv:1312.4400* (2013).

Class Activation Map(CAM)

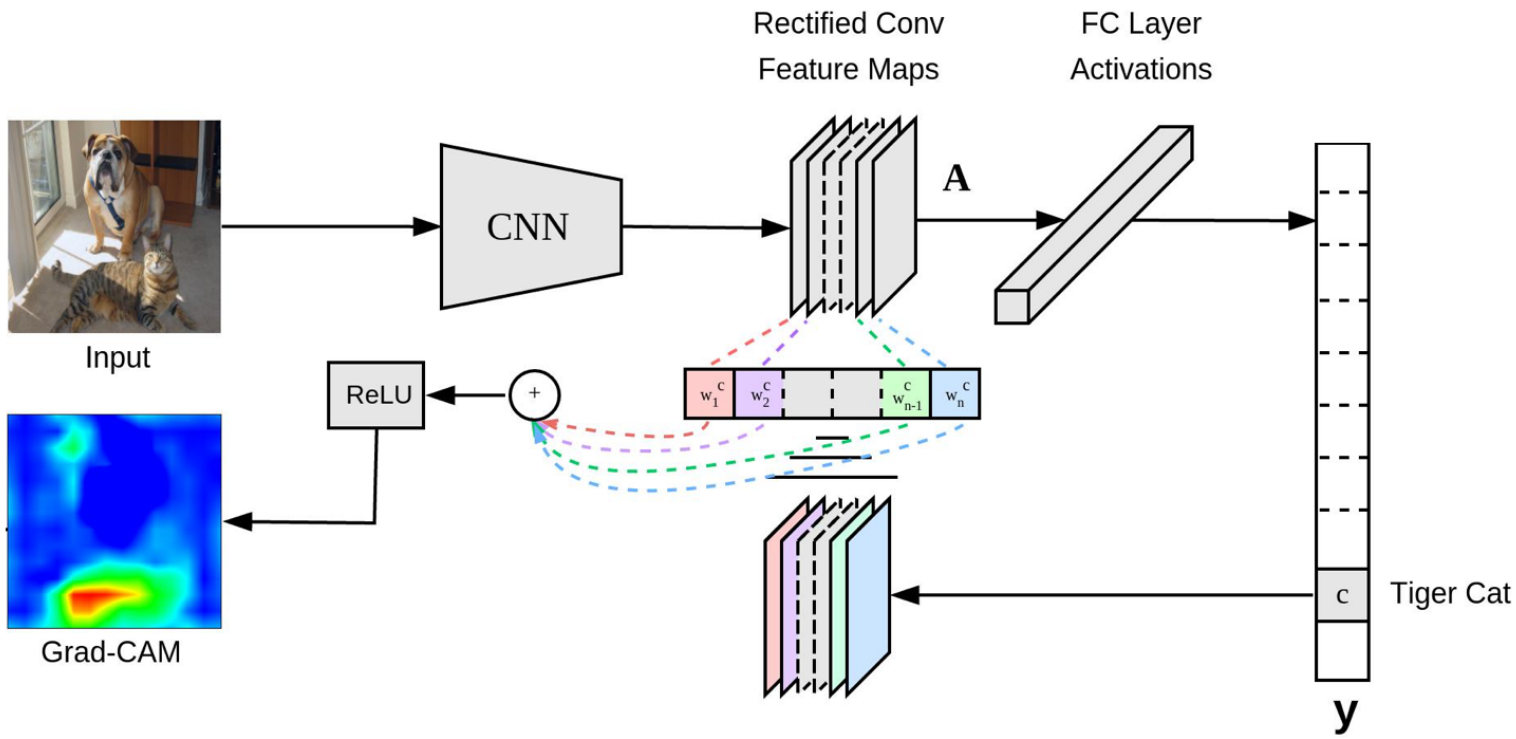


Zhou, Bolei, et al. "Learning deep features for discriminative localization." *CVPR*. 2016.

Grad-CAM

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

$$L^c_{\text{Grad-CAM}} = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}} \quad (2)$$



We apply a ReLU to the linear combination of maps because we are only interested in the features that have a *positive* influence on the class of interest, *i.e.* pixels whose intensity should be *increased* in order to increase yc. Negative pixels are likely to belong to other categories in the image.

Outline

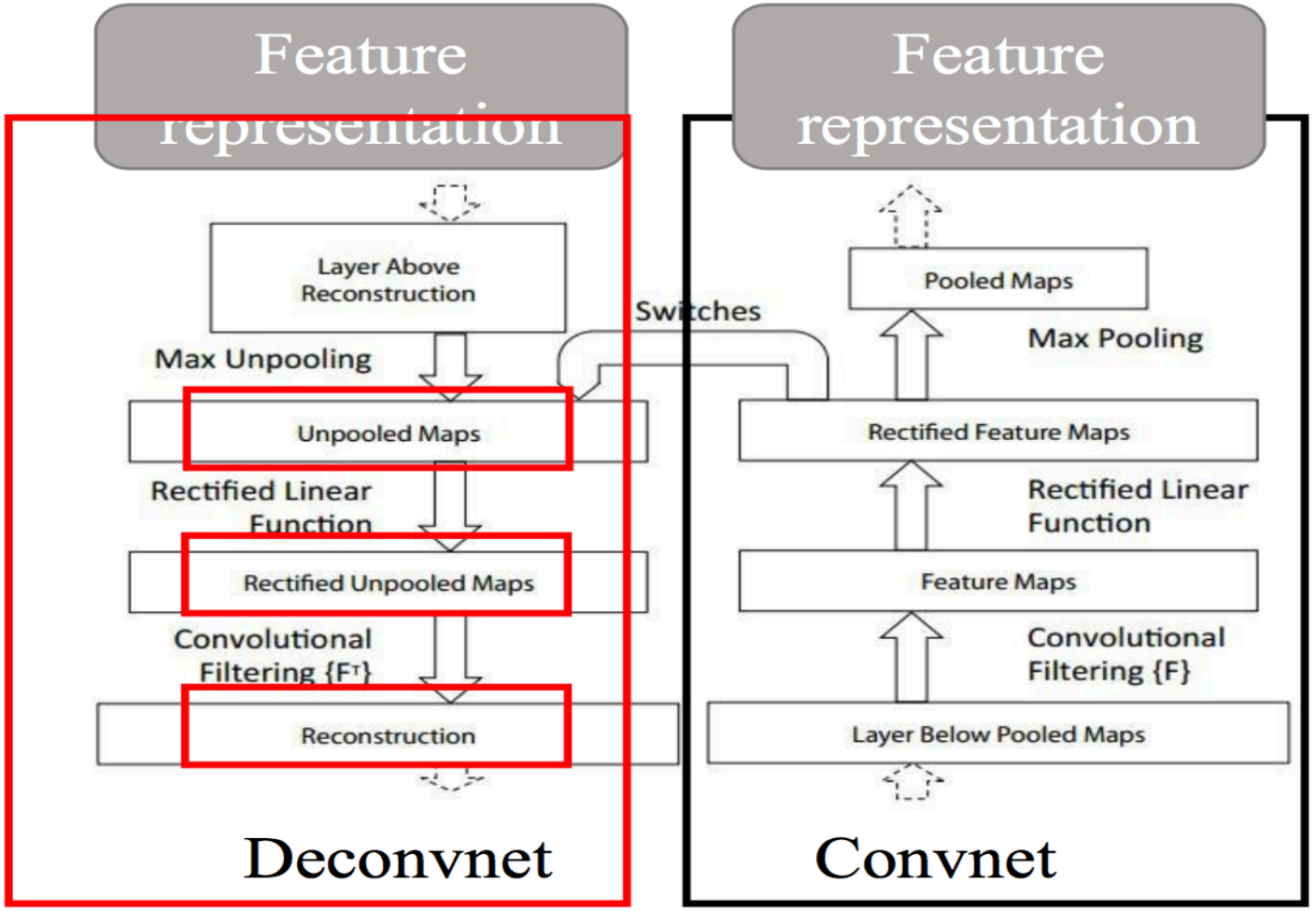
Approch

- Grad-CAM
- Guided Grad-CAM
  - combine Grad-CAM(coarse localization) with existing fine-grained visualizations(Guided Backpropagation to create a high-resolution class-discriminative visualization
  - Guided Backpropagation: Backpropagation+Deconvolution

Experiments

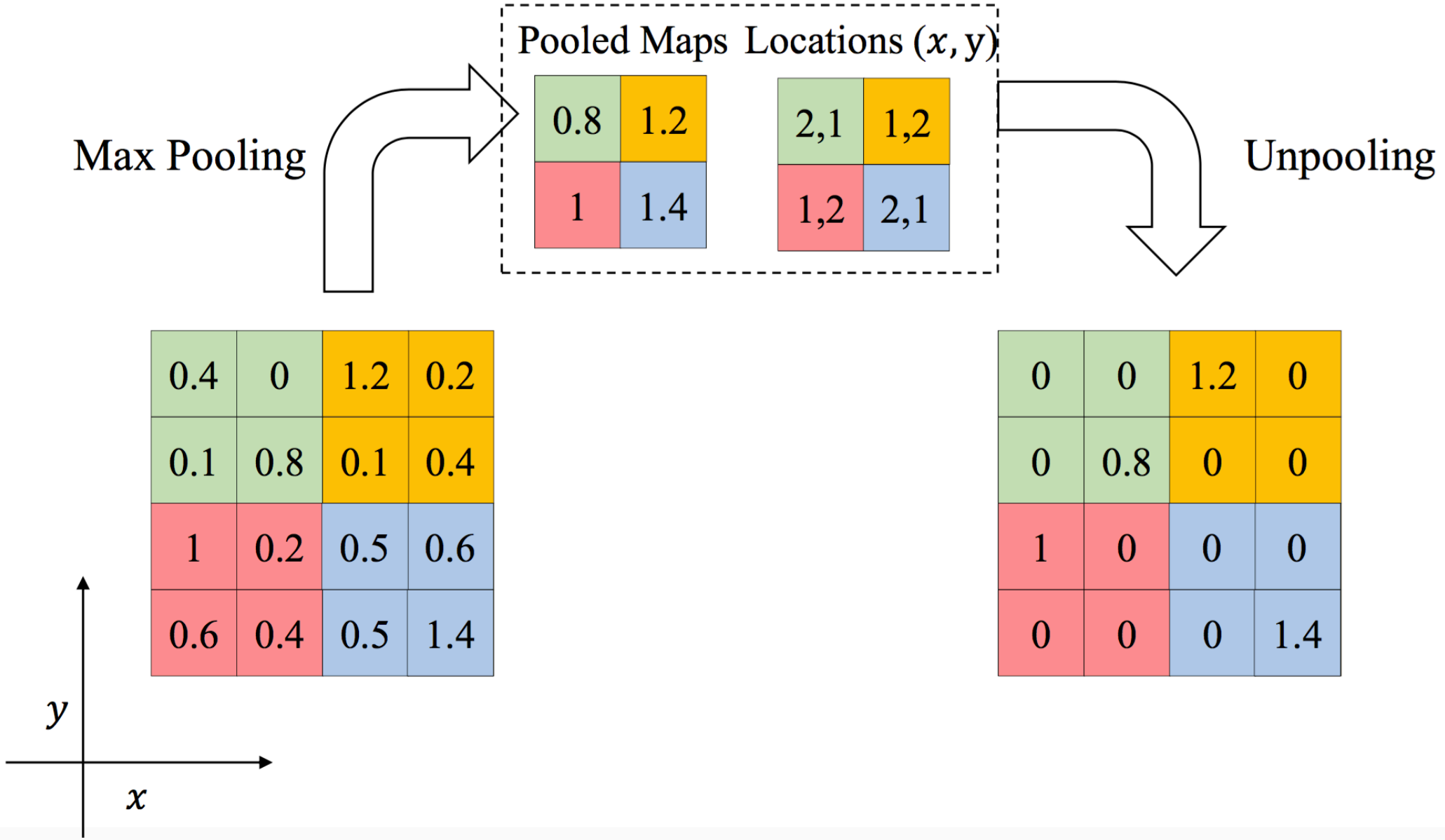
- Evaluating Localization
- Evaluating Visualizations
- Diagnosing image classification CNNs
- Image Captioning and VQA

Deconvolution visualization

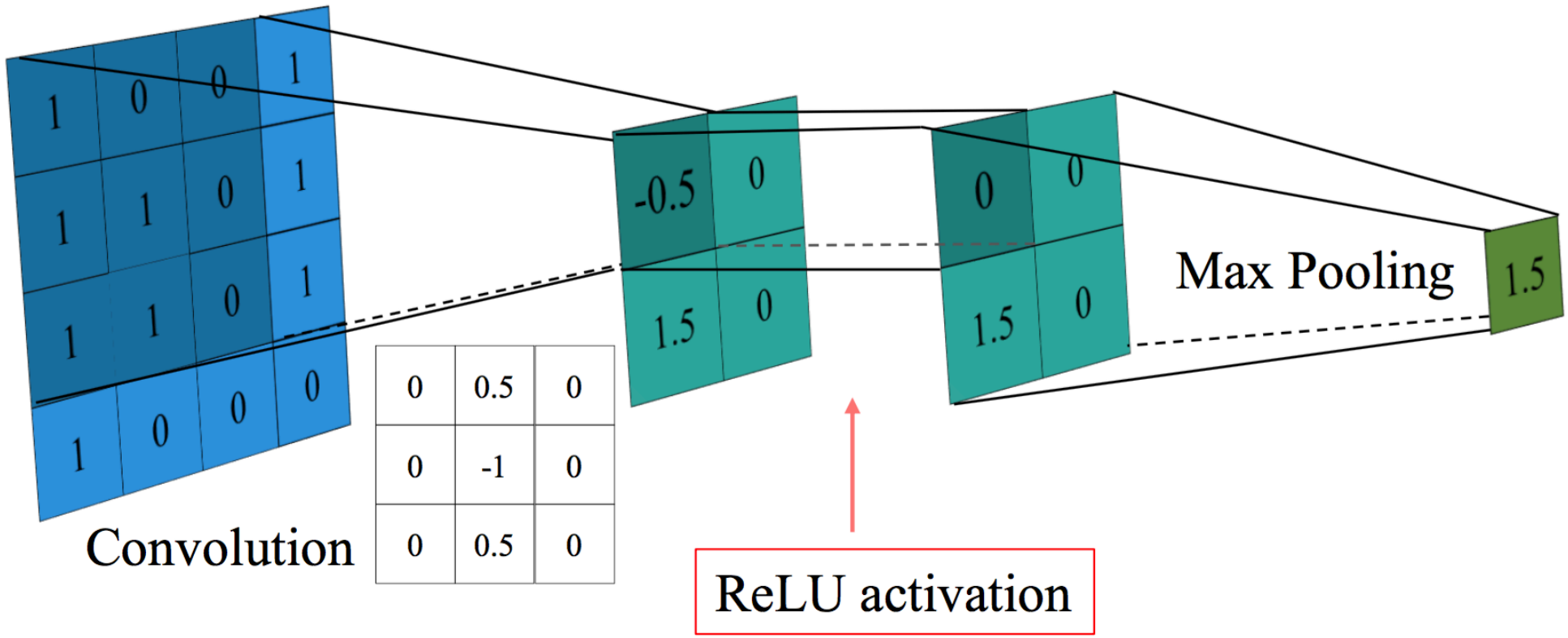


Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]. ECCV (2014)

Unpooling

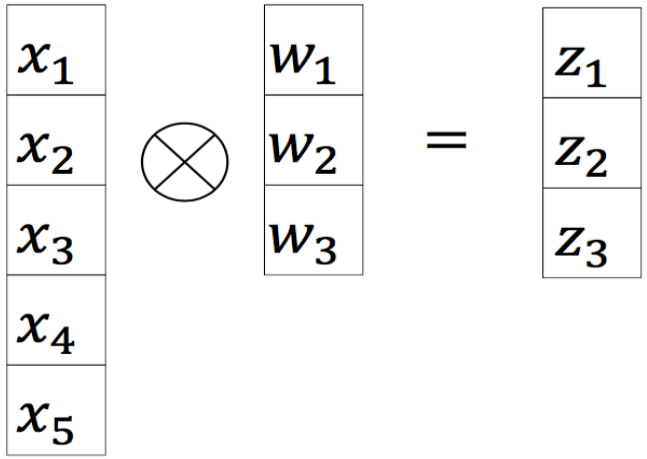

Pooled Maps Locations (x, y)
Max Pooling
0.8
1.2
1
1.4
2,1
1,2
1,2
2,1
Unpooling
0.4
0
1.2
0.2
0.1
0.8
0.1
0.4
1
0.2
0.5
0.6
0.6
0.4
0.5
1.4
0
0
1.2
0
0
0.8
0
0
1
0
0
0
0
0
0
1.4
y
x

Rectification



Convolution

1d CNNs feedforward: no padding, no strides

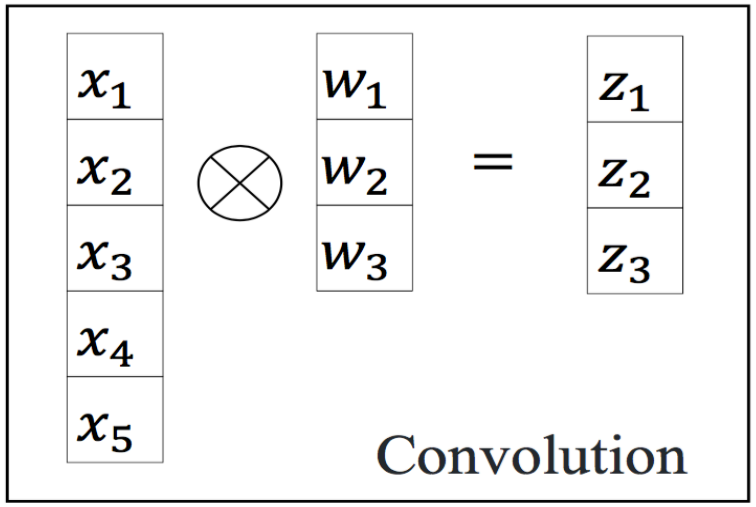


$$z_1 = w_1x_1 + w_2x_2 + w_3x_3$$
$$z_2 = w_1x_2 + w_2x_3 + w_3x_4$$
$$z_3 = w_1x_3 + w_2x_4 + w_3x_5$$

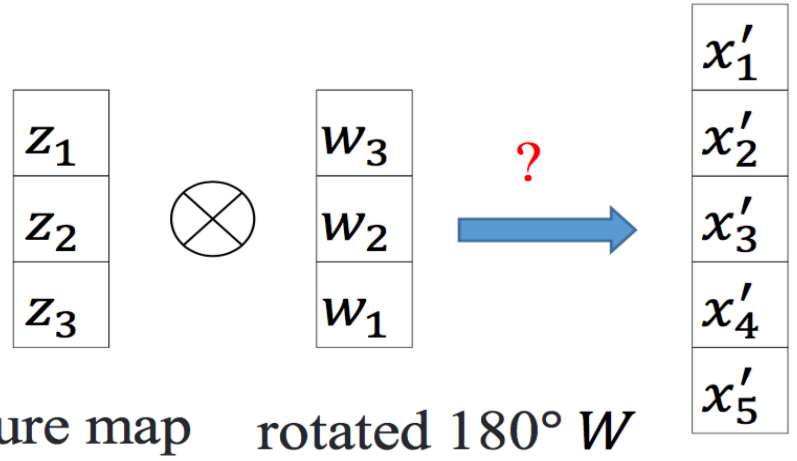
Matrix formulation:

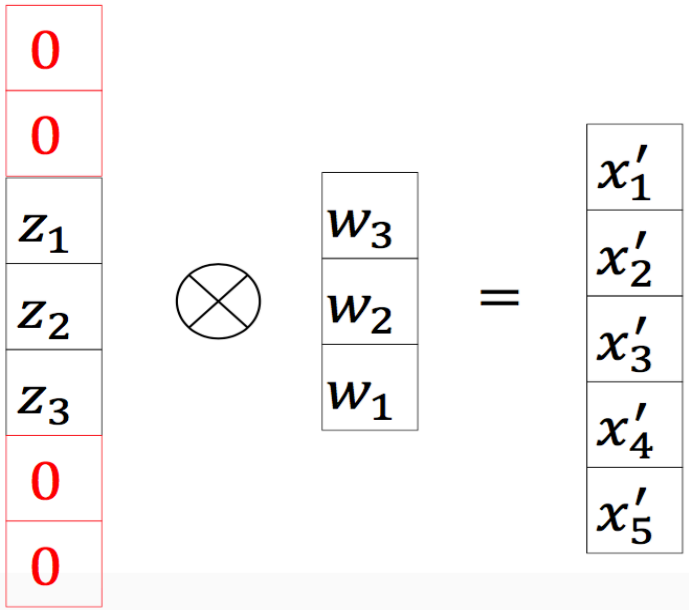
$$\begin{bmatrix} w_1 & w_2 & w_3 & 0 & 0 \\ 0 & w_1 & w_2 & w_3 & 0 \\ 0 & 0 & w_1 & w_2 & w_3 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

Deconvolution



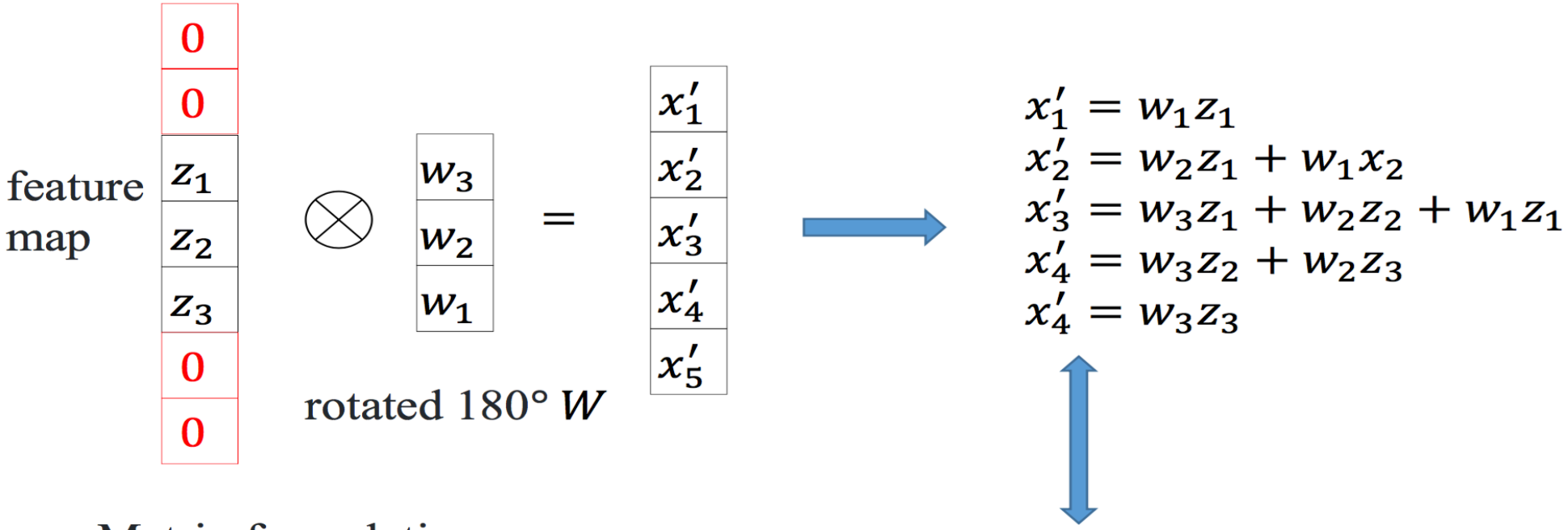
Deconvolution:





$$x'_1 = w_1 z_1$$
$$x'_2 = w_2 z_1 + w_1 x_2$$
$$x'_3 = w_3 z_1 + w_2 z_2 + w_1 z_1$$
$$x'_4 = w_3 z_2 + w_2 z_3$$
$$x'_4 = w_3 z_3$$

Deconvolution



Matrix formulation:

$$\begin{bmatrix} w_1 & 0 & 0 \\ w_2 & w_1 & 0 \\ w_3 & w_2 & w_1 \\ 0 & w_3 & w_2 \\ 0 & 0 & w_3 \end{bmatrix} \times \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} x'_1 \\ x'_2 \\ x'_3 \\ x'_4 \\ x'_5 \end{bmatrix}$$

Deconvolution

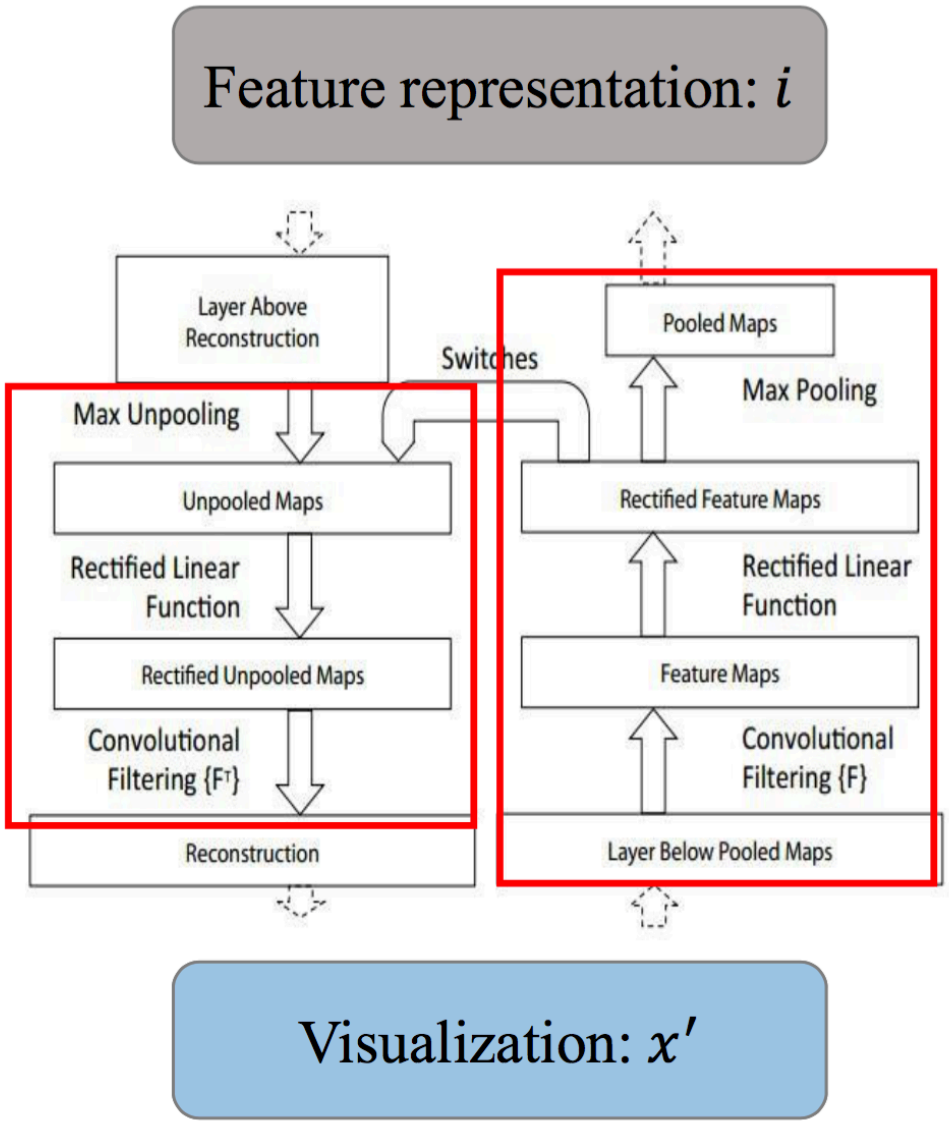
Matrix formulation 1:

$$\begin{bmatrix} w_1 & w_2 & w_3 & 0 & 0 \\ 0 & w_1 & w_2 & w_3 & 0 \\ 0 & 0 & w_1 & w_2 & w_3 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

Matrix formulation 2:

$$\begin{bmatrix} w_1 & 0 & 0 \\ w_2 & w_1 & 0 \\ w_3 & w_2 & w_1 \\ 0 & w_3 & w_2 \\ 0 & 0 & w_3 \end{bmatrix} \times \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} x'_1 \\ x'_2 \\ x'_3 \\ x'_4 \\ x'_5 \end{bmatrix}$$

Deconvolution visualization



For a given image x and a fully trained model, to visualize the i-th feature map in the layer l:

1, for a given x, feedforward computing feature activities in the layer l

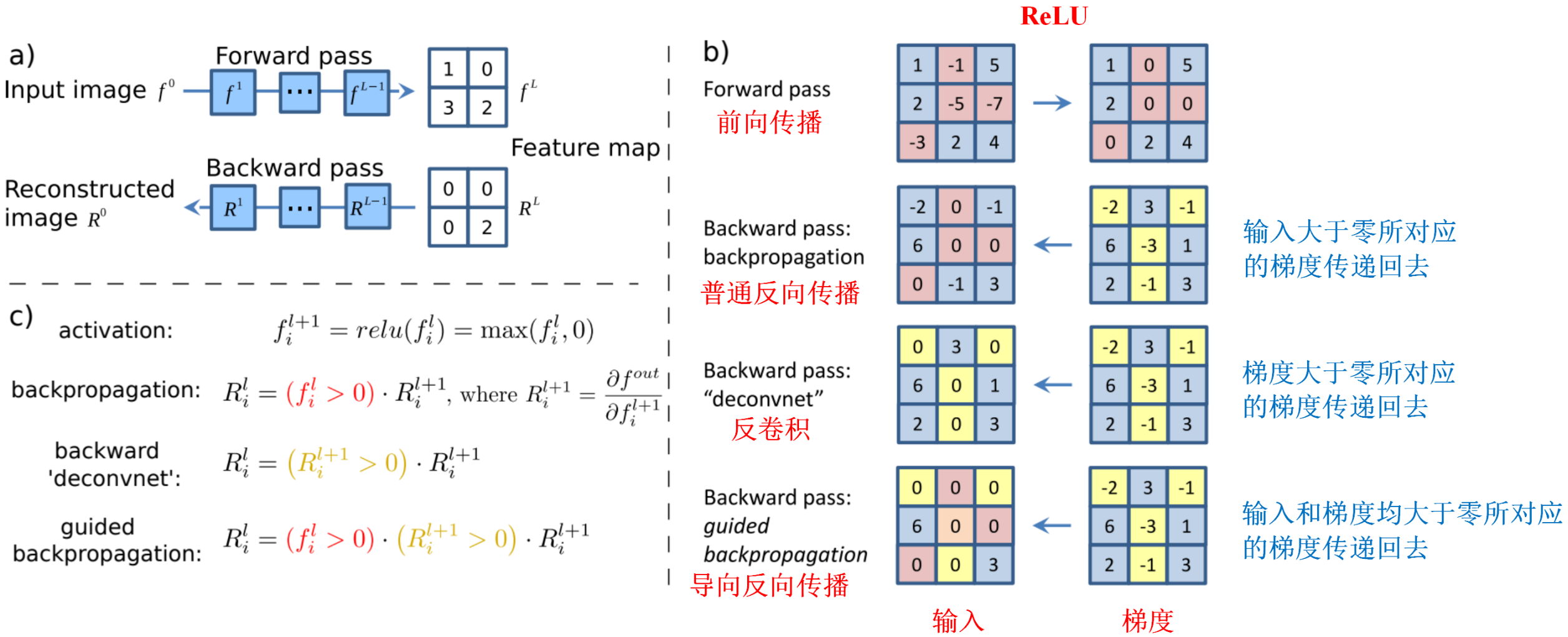
2, set all other activations (not the i-th feature map) in the layer to zero

3, for layer l to the first layer:

if the layer beneath is pooling layer:
use **Unpooling** operation

else if the layer beneath is convolutional layer:
use **Rectify** first **and Transposed filters** operation

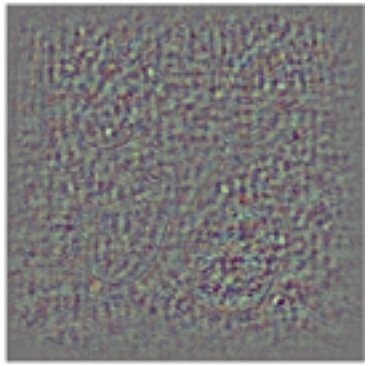
Guided BP



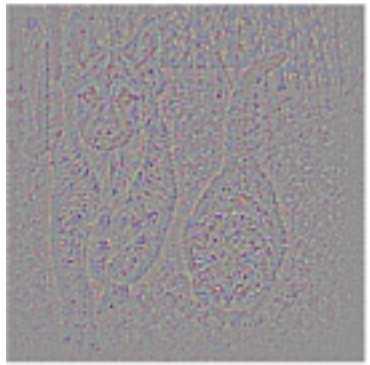
Springenberg, Jost Tobias, et al. "Striving for simplicity: The all convolutional net." *arXiv preprint arXiv:1412.6806* (2014)

Visualization comparison

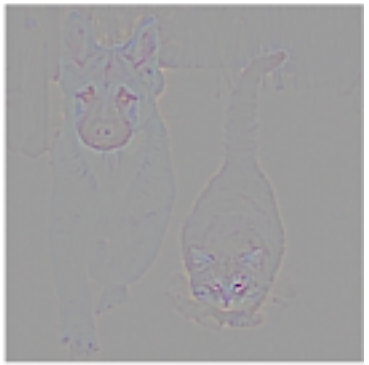
原图



普通反向传播



反卷积



导向反向传播

- 普通反向传播得到的图像噪声较多。
- 反卷积可以大概看清楚猫和狗的轮廓，但是有大量噪声在物体以外的位置上。
- 导向反向传播基本上没有噪声，特征很明显的集中猫和狗的身体部位上。
- 反卷积和导向反向传播不能拿来解释分类的结果，因为它们对类别并不敏感，直接把所有能提取的特征都展示出来了。

Application of Guided BP

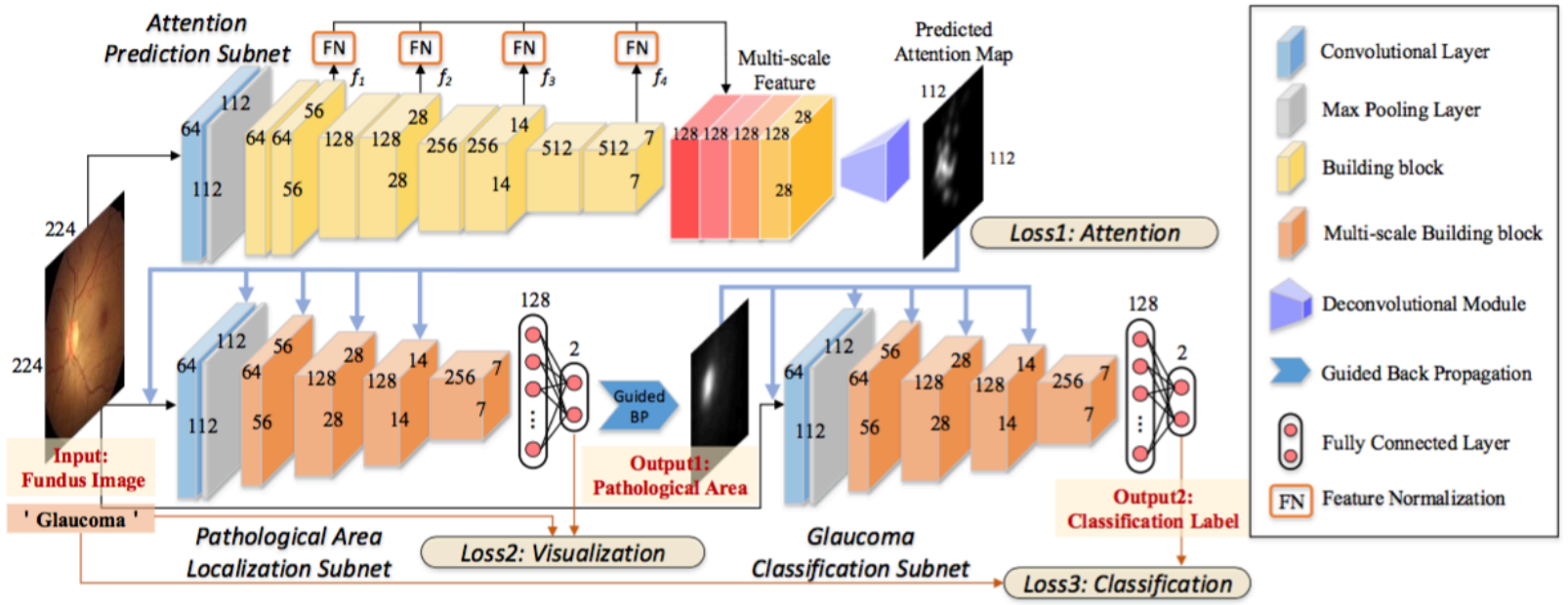
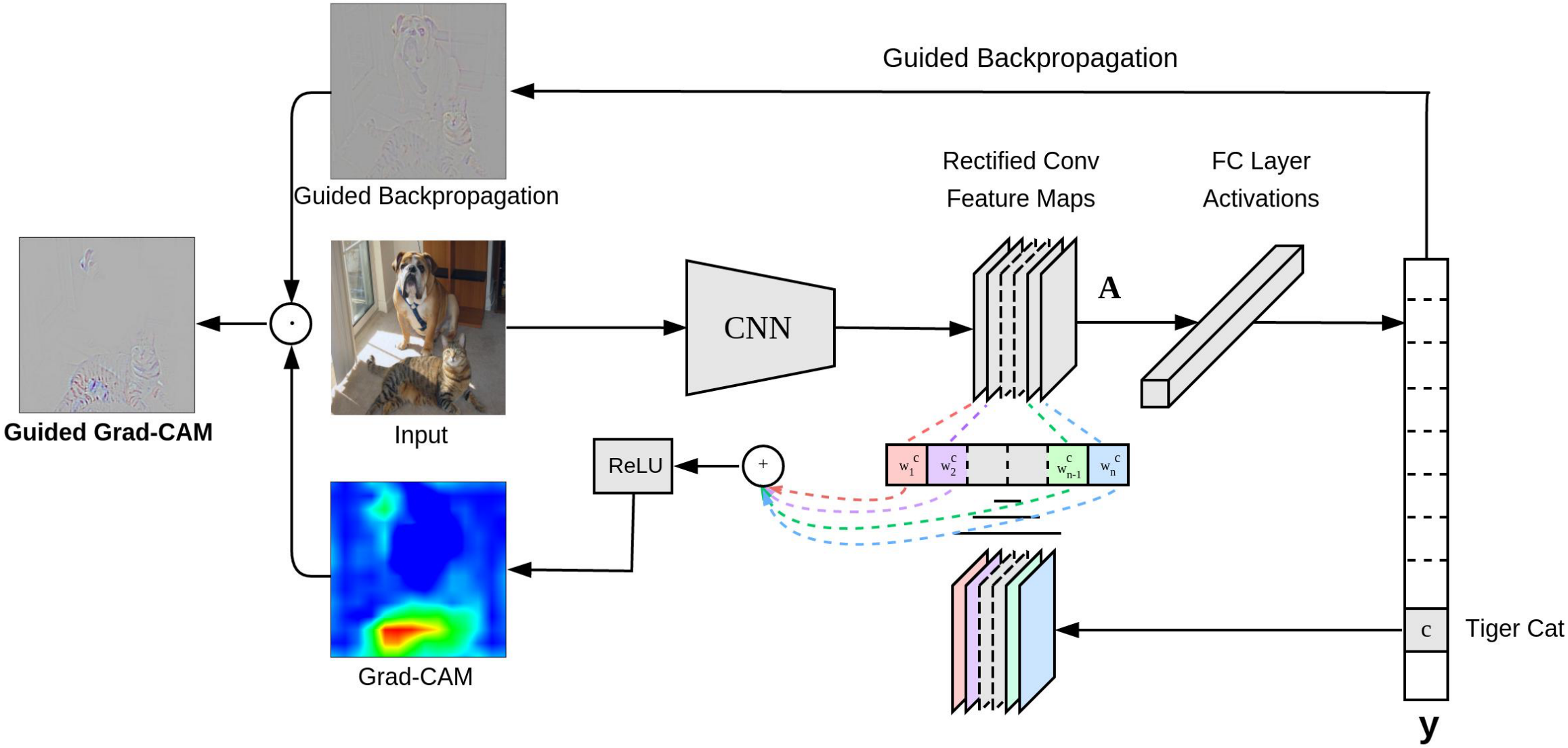


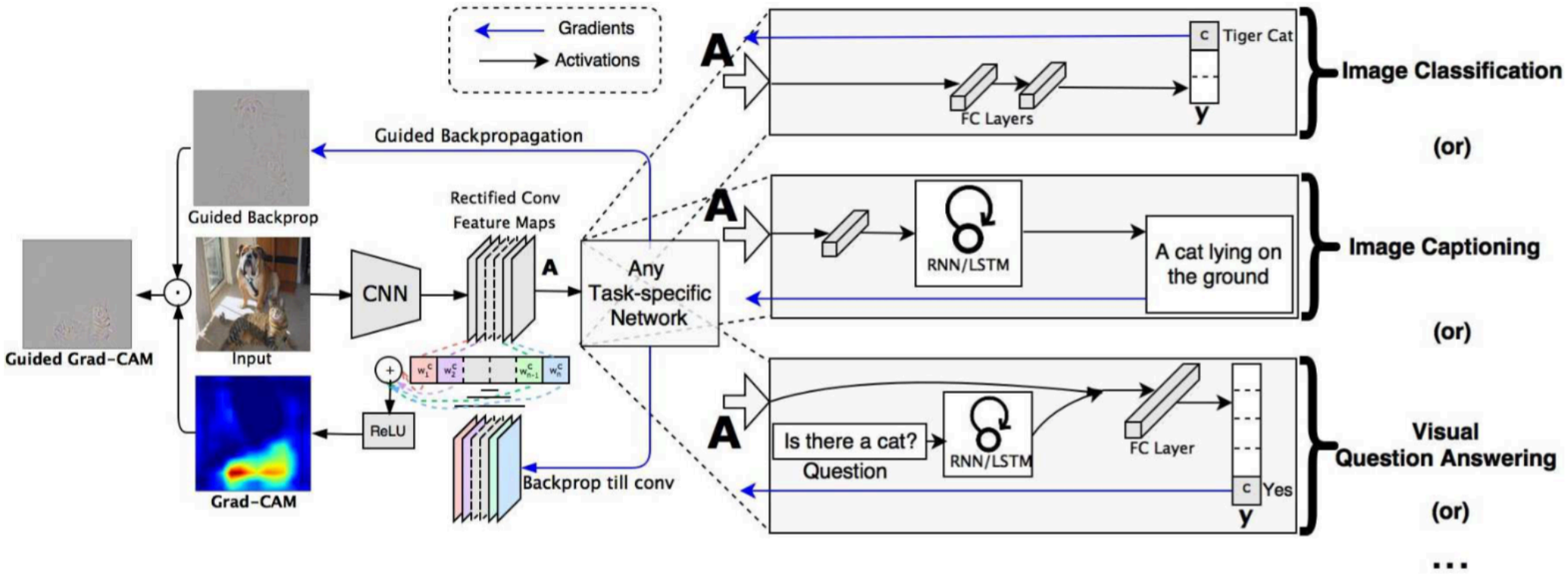
Figure 6. Architecture of our AG-CNN network for glaucoma detection. The sizes of the feature maps and convolutional kernels are shown in this figure.

Li, Liu, et al. "Attention Based Glaucoma Detection: A Large-scale Database and CNN Model." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

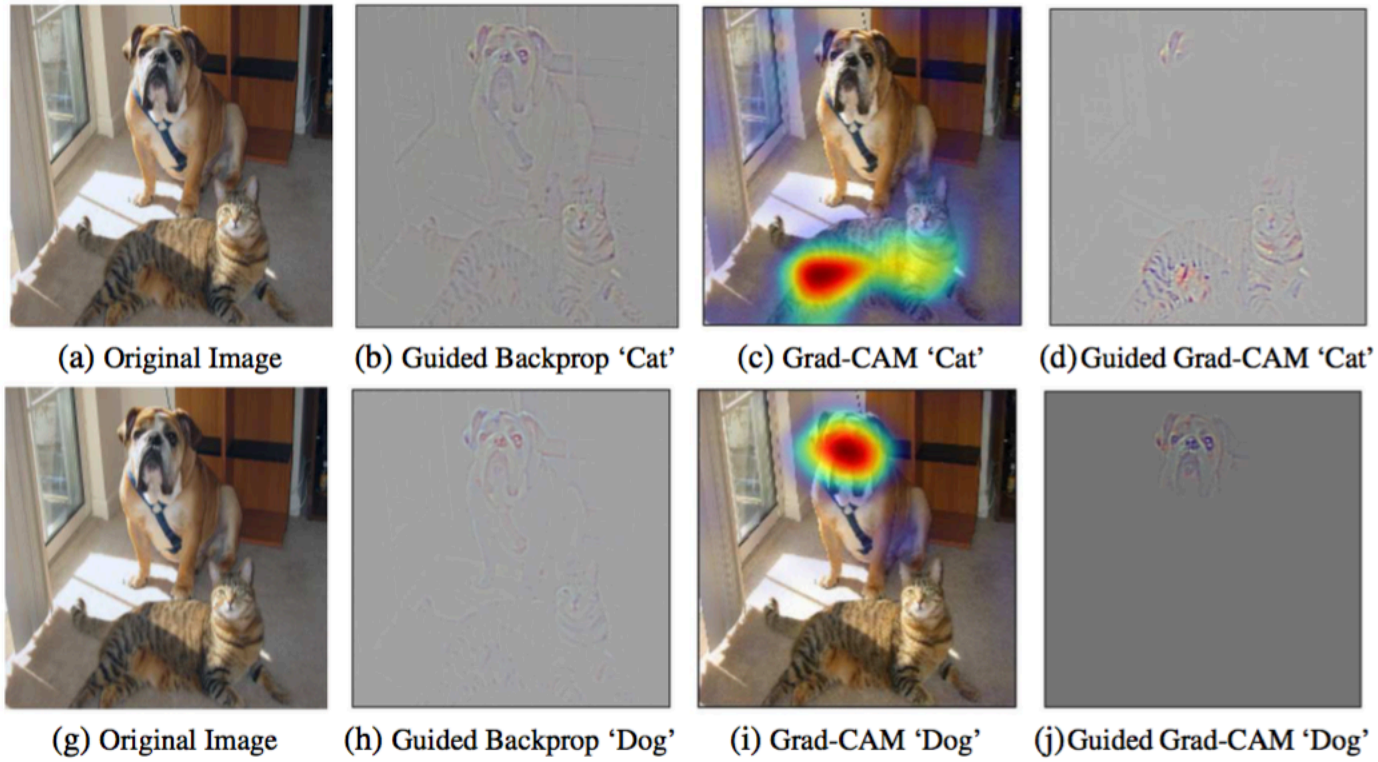
Guided Grad-CAM



Guided Grad-CAM



Visualization comaration



(a,g) Original image with a cat and a dog.
(b,h) Guided Backpropagation:highlights all contributing features.
(c,i) Grad-CAM localizes class-discriminative regions.
(d,j) Guided Grad-CAM gives high-resolution class-discriminative visualizations.

Outline

Approch

- Grad-CAM
- Guided Grad-CAM

Experiments

- Evaluating Localization
  - Weakly-supervised Localization
- Evaluating Visualizations
- Diagnosing image classification CNNs
- Image Captioning and VQA

Weakly-supervised Localization

(1)Given an image, first obtain class predictions from network.
(2)generate Grad-CAM maps for each of the predicted classes.
(3)binarize with threshold of 15% of the max intensity which results in connected segments of pixels.
(4)draw bounding box around the single largest segment.

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).

Outline

Approch

- Grad-CAM
- Guided Grad-CAM

Experiments

- Evaluating Localization
- Evaluating Visualizations
  - Evaluating Class Discrimination
  - Evaluating Trust
  - Faithfulness *vs*. Interpretability
- Diagnosing image classification CNNs
- Image Captioning and VQA

Evaluating Class Discrimination

Purpose:measure whether Grad-CAM helps distinguish between classes

Steps:

(1)train VGG-16 and AlexNet CNNs fine-tuned on PASCAL VOC 2007 train set.

(2) select images from VOC 2007 val set that contain exactly two annotated categories and create visualizations for each one of them. For both VGG-16 and AlexNet CNNs, we obtain category-specific visualizations using four techniques: **Deconvolution**, **Guided Backpropagation**, **Deconvolution Grad-CAM** and **Guided Grad-CAM**.

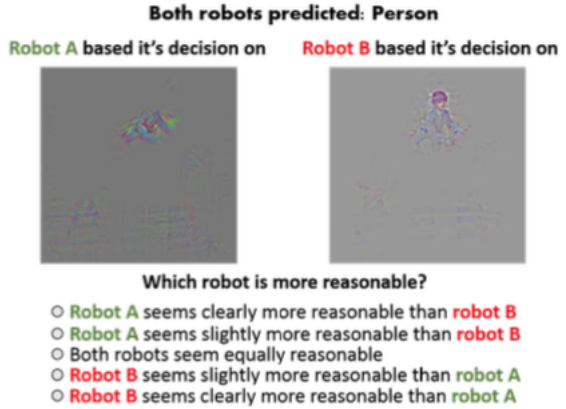
(3) show visualizations to 43 workers on Amazon Mechanical Turk (AMT) and ask them "Which of the two object categories is depicted in the image?"

Evaluation:4 visualizations for 90 image-category pairs (*i.e.* 360 visualizations); 9 ratings were collected for each image, evaluated against the ground truth and averaged to obtain the accuracy.

Conclusion: Guided Grad-CAM performs the best among all the methods.

	Guided Backpropagation	Deconvolution	Guided Grad-CAM
accuaracy	44.44%	53.33%	61.23%

Evaluating Trust



Purpose:Given two prediction explanations, we evaluate which seems more trustworthy.

Method:use AlexNet and VGG-16 to compare Guided Backpropagation and Guided Grad-CAM visualizations.

Steps:noting that VGG-16 is known to be more reliable than AlexNet with an accuracy of 79.09 mAP (*vs.* 69.20 mAP) on PASCAL classification. In order to tease apart the efficacy of the visualization from the accuracy of the model being visualized, we consider only those instances where both models made the same prediction as ground truth. Given a visualization from AlexNet and one from VGG-16, and the predicted object category, 54 AMT workers were instructed to rate the reliability of the models relative to each other on a scale of clearly more/less reliable (+/-2), slightly more/less reliable (+/-1), and equally reliable (0).

Conclusion: human subjects are able to identify the more accurate classifier (VGG over AlexNet) simply from the different explanations, despite identical predictions.

Faithfulness *vs*. Interpretability

Purpose:evaluate how faithful they are to the underlying model.

Method:image occlusion (measure the difference in CNN scores when patches of the input image are masked)

Result:patches which change the CNN score are also patches to which Grad-CAM and Guided Grad-CAM assign high intensity, achieving rank correlation 0.254 and 0.261 (*vs*. 0.168, 0.220 and 0.208 achieved by Guided Backpropagation, c-MWP and CAM, respectively) averaged over 2510 images in PASCAL 2007 val set.

Conclusion:This shows that Grad-CAM visualizations are more faithful to the original model compared to all existing methods.

Outline

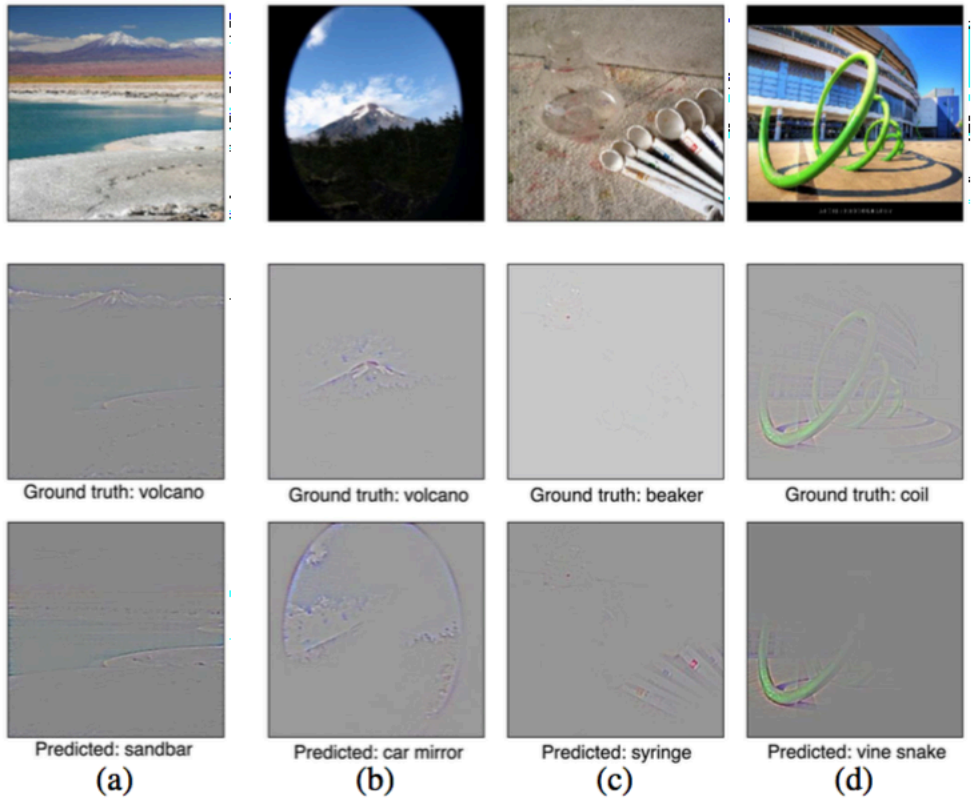
Approch

- Grad-CAM
- Guided Grad-CAM

Experiments

- Evaluating Localization
- Evaluating Visualizations
- Diagnosing image classification CNNs
  - Analyzing Failure Modes for VGG-16
  - Identifying bias in dataset
- Image Captioning and VQA

Analyzing Failure Modes for VGG-16



In these cases the model (VGG-16) failed to predict the correct class in its top 1 (a and d) and top 5 (b and c) predictions. Humans would find it hard to explain some of these predictions without looking at the visualization for the predicted class. But with Grad-CAM, these mistakes seem justifiable.

Identifying bias in dataset

- 有偏差的数据集训练的模型可能不会推广到现实世界的场景。
- 将微调的ImageNet训练的VGG-16模型用于“医生”与“护士”的分类任务。模型预测的Grad-CAM可视化表明，该模型学会了看人的面部/发型以区分护士和医生，从而学习了性别的刻板观念。这个模型把几名女医生错误地归类为护士，男护士分类成医生。结果发现图像搜索结果存在性别偏差（78%的医生图像是男性，93%的护士图像是女性）。
- 通过从可视化中获得的直观结果，将男护士和女医生添加到训练集中，减少了训练集中的偏差。重新训练的模型现在更好地推广到更平衡的测试集。

Outline

Approch

- Grad-CAM
- Guided Grad-CAM

Experiments

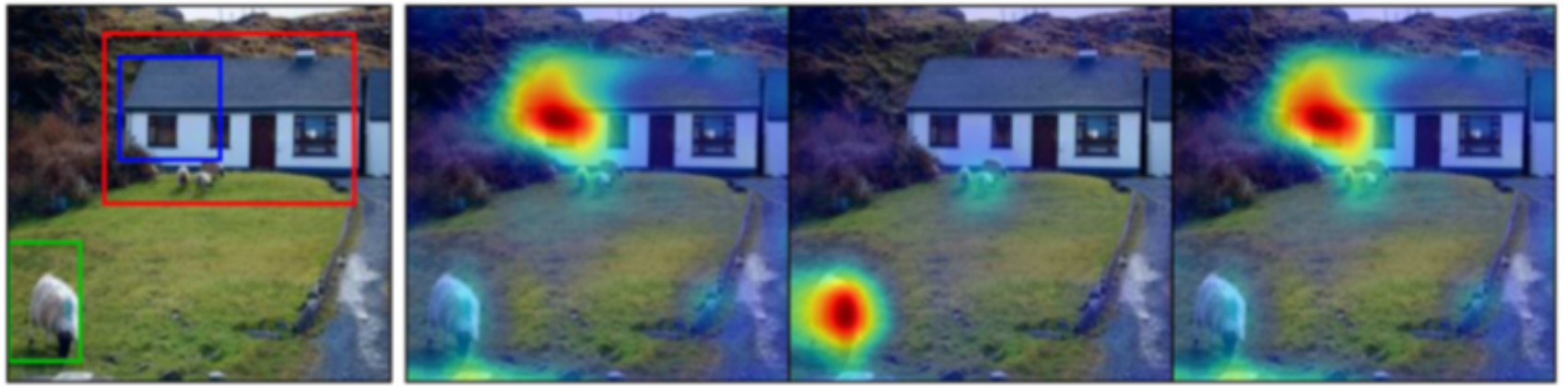
- Evaluating Localization
- Evaluating Visualizations
- Diagnosing image classification CNNs
- Image Captioning and VQA
  - Image Captioning
  - Visual Question Answering

Image Captioning

(a) Image captioning explanations

(a) Visual explanations from image captioning model highlighting image regions considered to be important for producing the captions.

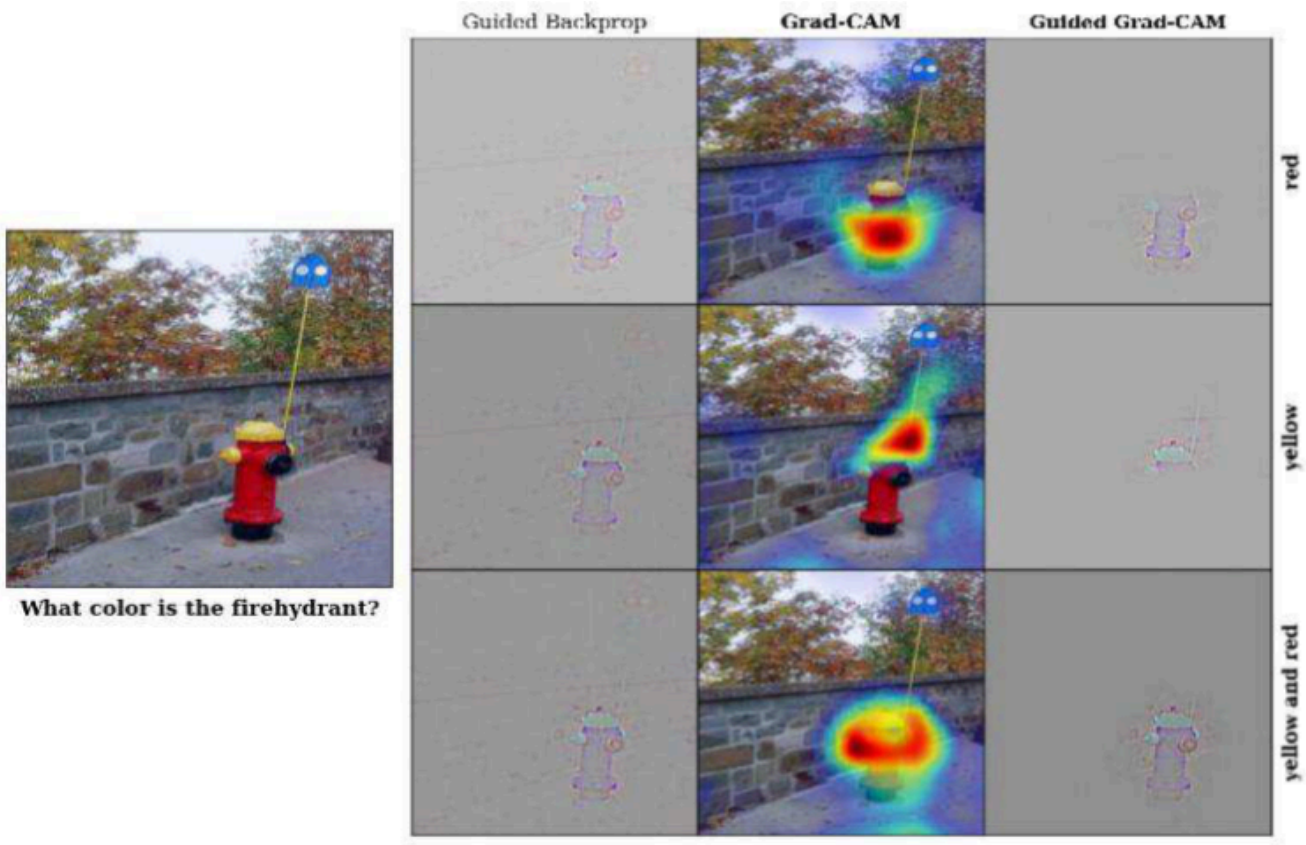
Image Captioning



(b) Comparison to DenseCap

(b) Grad-CAM localizations of a global or holistic captioning model for captions generated by a dense captioning model for the three bounding box proposals marked on the left. Grad-CAM localizations (right) agree with those bounding boxes.

Visual Question Answering



Grad-CAM visualizations are highly interpretable and help explain any target prediction (“red” , “yellow” , “yellow and red”)