Class Activation Map

20184448
MinKi Jo (조민기)

1

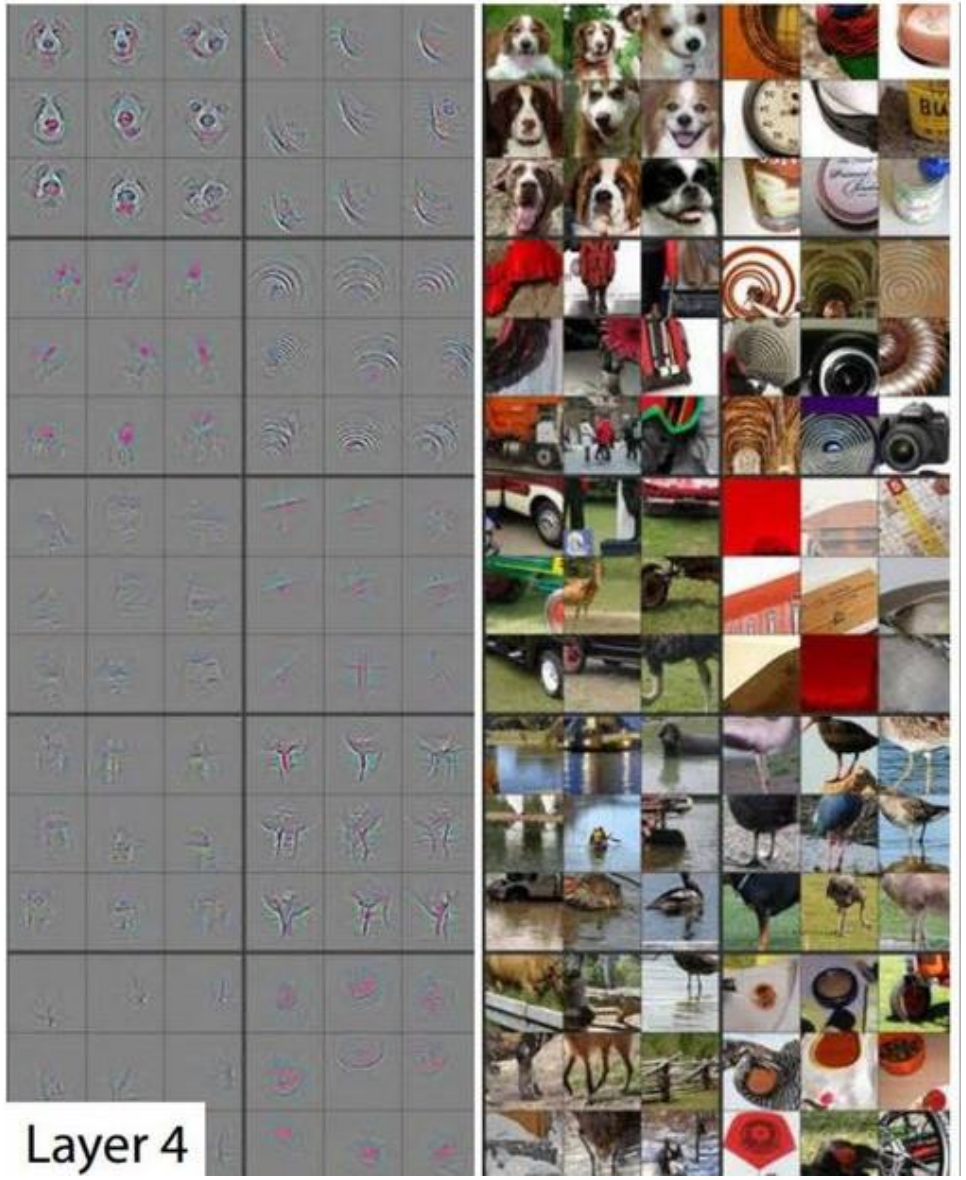
BRIEF SUMMARY

PROBLEM

FUNCTION

Objective

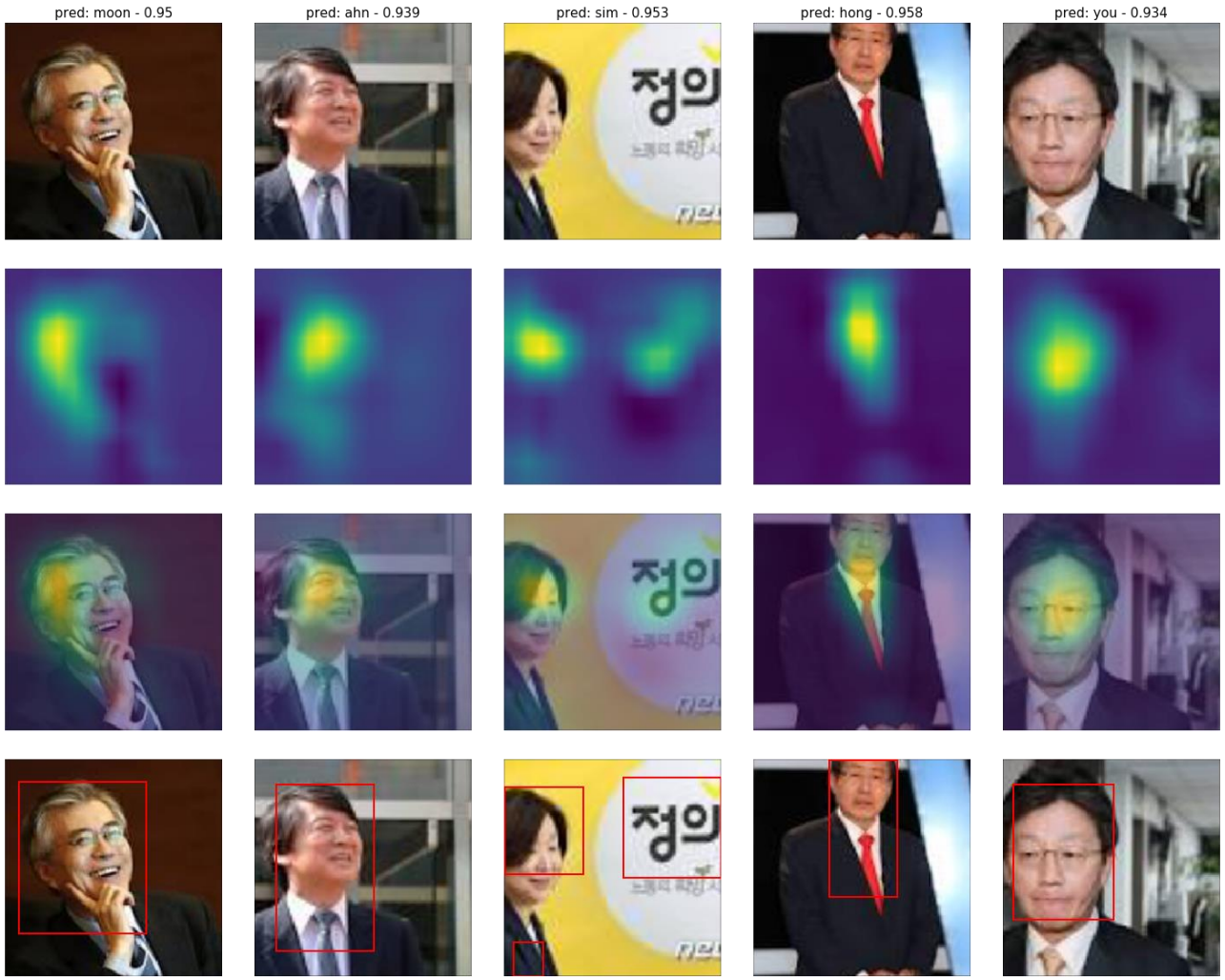
- Convolutional neural network achieved great work for the computer vision task, but people could not figure out how it works.
- This work is one of the attempt to visualize the logic of the network.



M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. Proc. ECCV, 2014

Function

- CAM draws the heatmap of the network that shows the activated region.
- This Could be used for the localization of the object from the image without the annotated data.



Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba, Learning Deep Features for Discriminative Localization, CVPR 2015

2

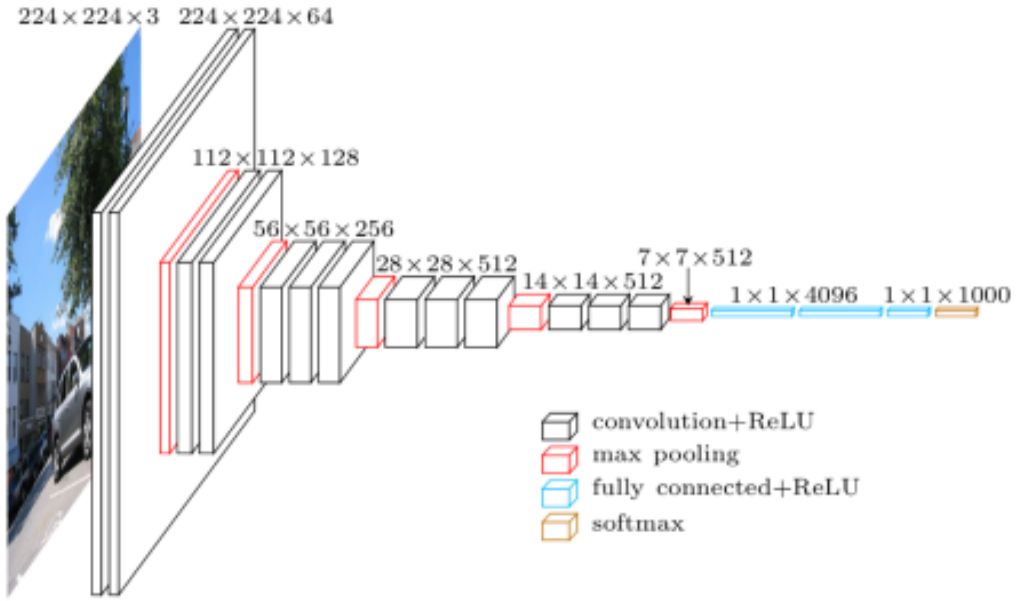
MODEL

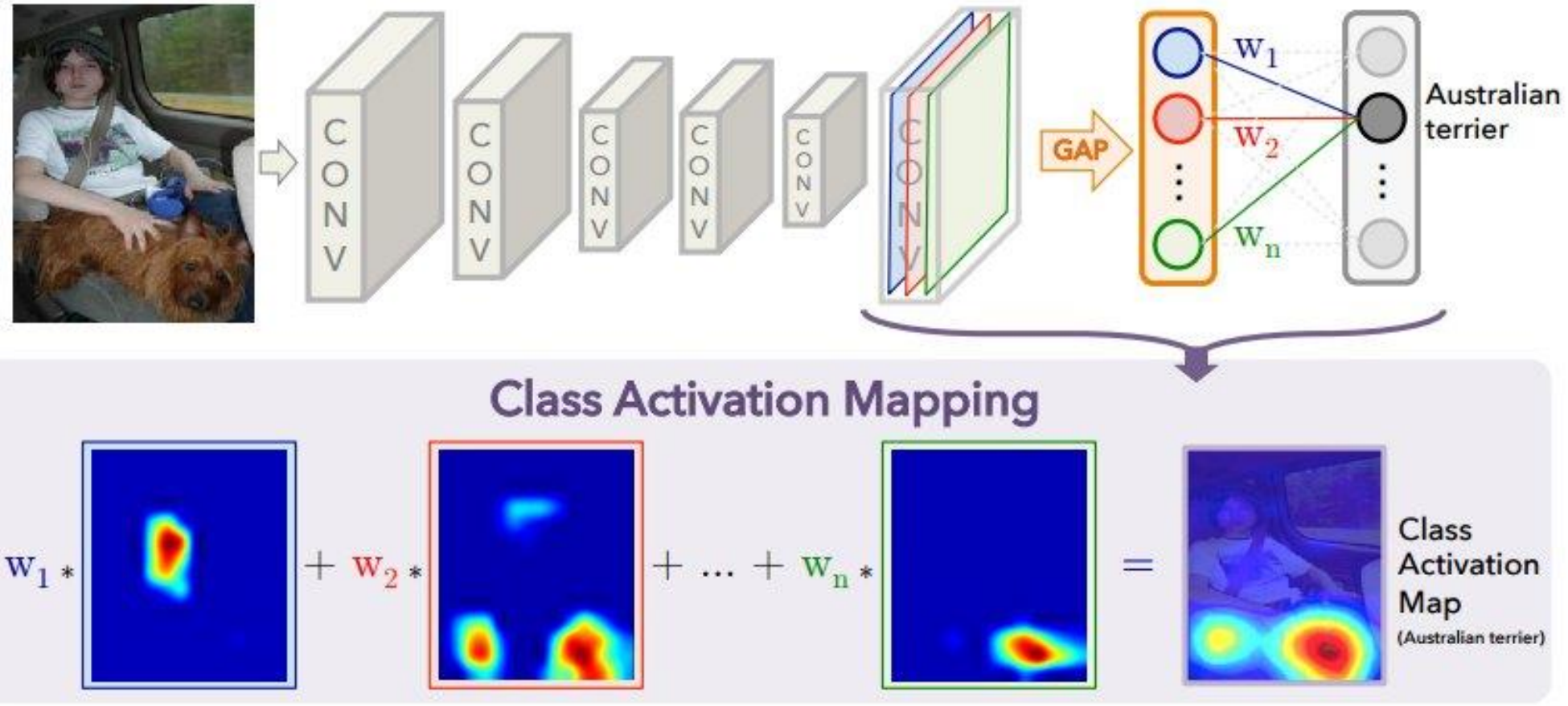
THEORY

ARCHITECTURE

Theoretical assumption

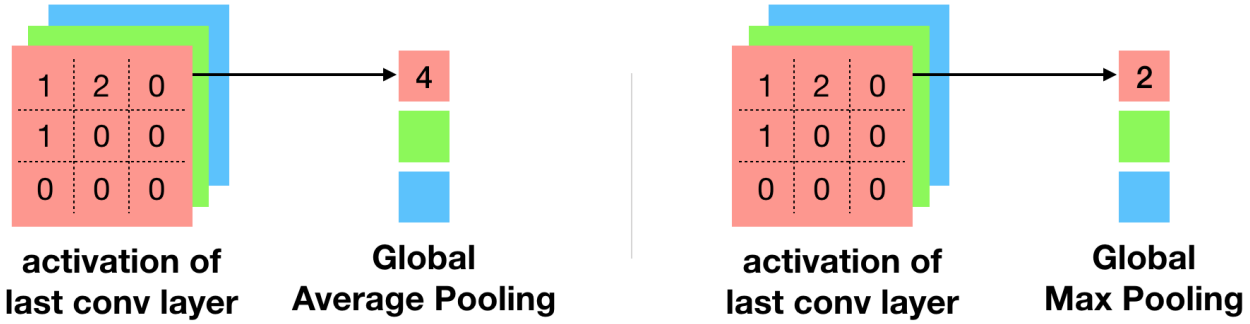
- The assumption is that the features from some convolutional layer must contain the location information.
- Thus, if we can calculate the activation of the feature that contains the location information, we can see the vision of the network.




CONV
CONV
CONV
CONV
CONV
CONV
GAP
w_1
w_2
w_n
Australian terrier
Class Activation Mapping
w_1 *
+ w_2 *
+ ... + w_n *
=
Class Activation Map
(Australian terrier)

Global Average Pooling

- The Network in Network (NIN) employed the GAP first for replacement of the FC layer.
- Without the FC layer, It shows the fine performance for the classification task.



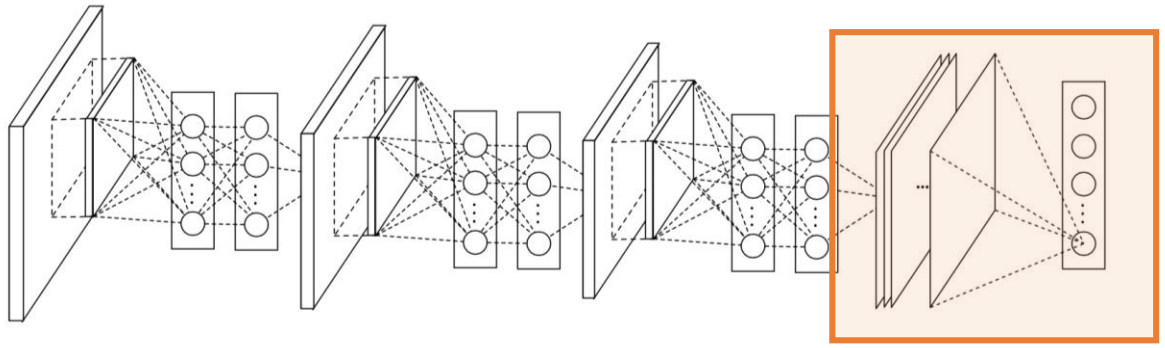


Figure 2: The overall structure of Network In Network. In this paper the NINs include the stacking of three mlpconv layers and one global average pooling layer.

M. Lin, Q. Chen, and S. Yan. Network in network. International Conference on Learning Representations, 2014

Global Average Pooling

- The Network in Network (NIN) employed the GAP first for replacement of the FC layer.
- Without the FC layer, It shows the fine performance for the classification task.

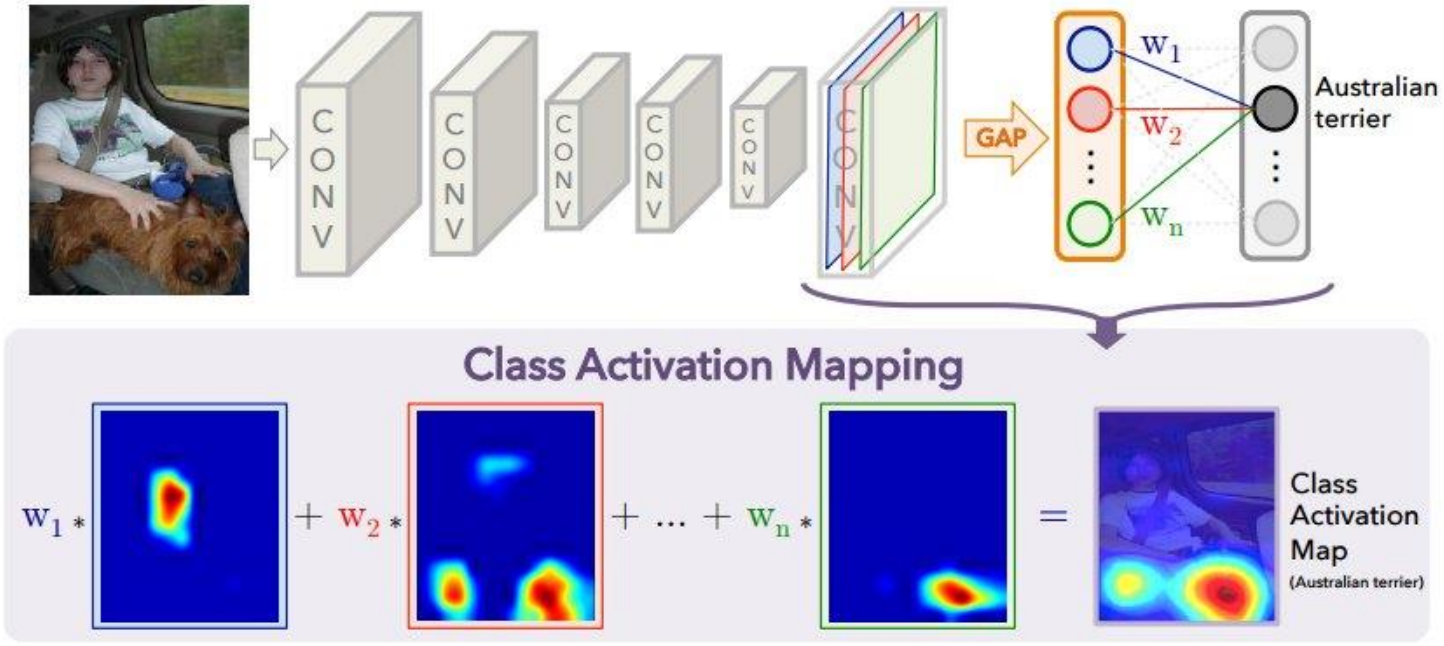
Table 1. Classification error on the ILSVRC validation set.

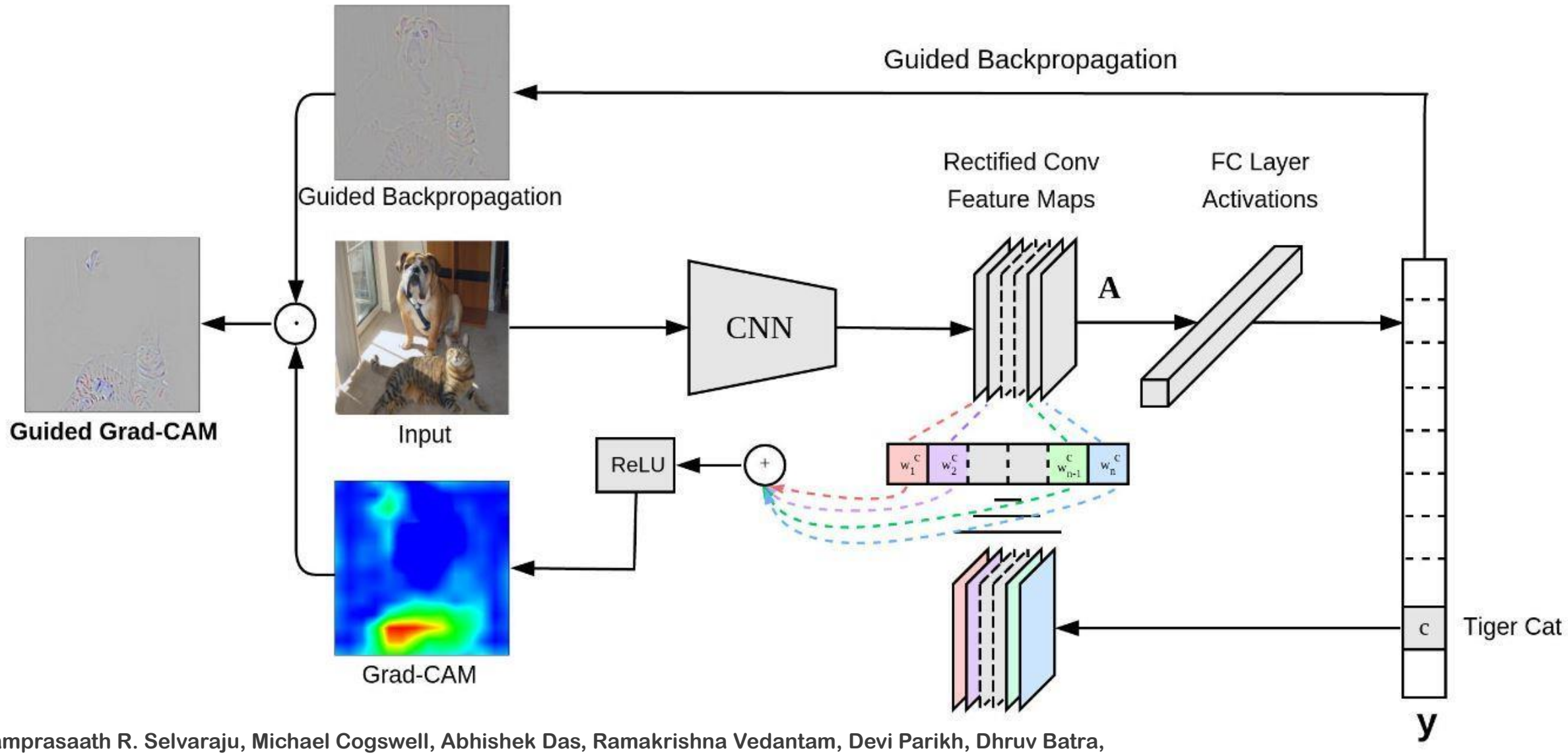
Networks	top-1 val. error	top-5 val. error
VGGnet-GAP	33.4	12.2
GoogLeNet-GAP	35.0	13.2
AlexNet*-GAP	44.9	20.9
AlexNet-GAP	51.1	26.3
GoogLeNet	31.9	11.3
VGGnet	31.2	11.4
AlexNet	42.6	19.5
NIN	41.9	19.6
GoogLeNet-GMP	35.6	13.9

Except the Alex Net, the network shows almost similar performance when it lost the FC layer

Method

- Rid off the FC layer and attach the GAP layer right after the conv-layer & retrain.
- Calculate the weight of each image feature by Global Average Pooling (GAP).
- Conduct the weighted sum for every feature and normalize it.

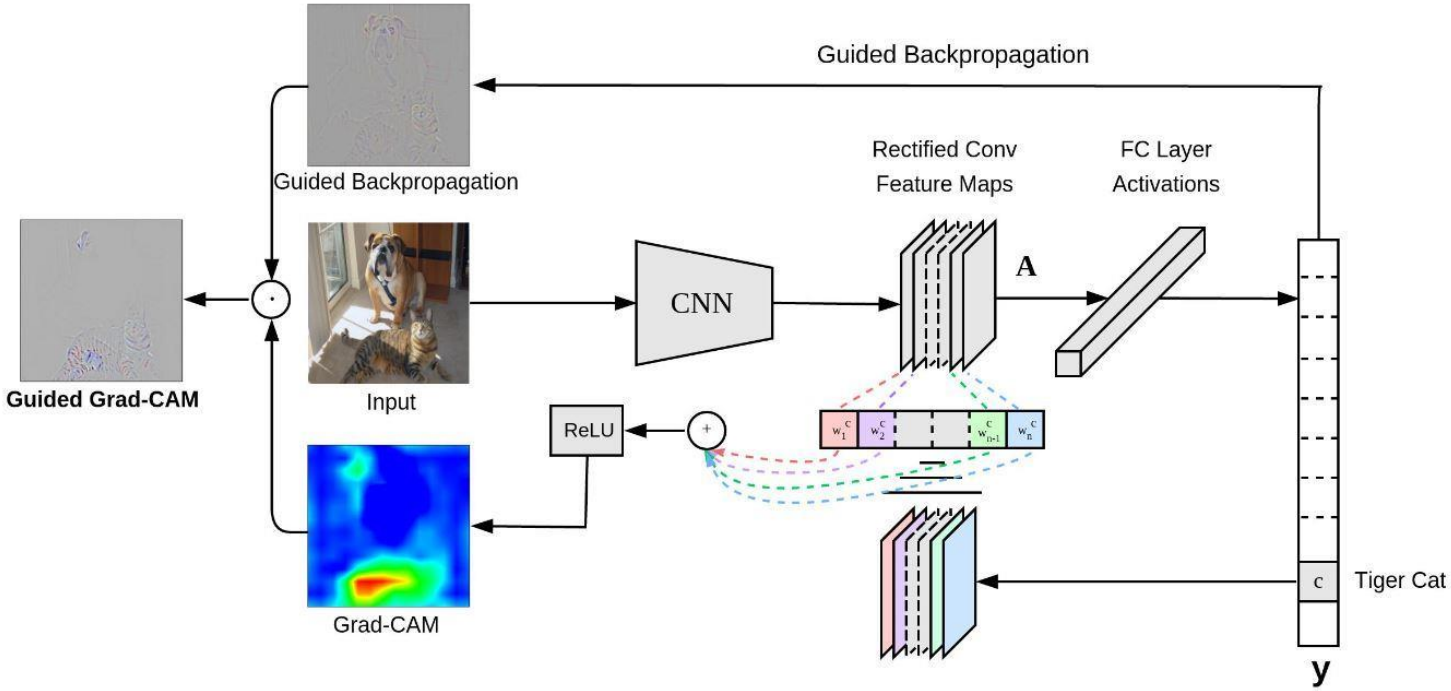




Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, ICCV 2017

Grad-CAM

- Instead of the GAP, Grad CAM get the gradient value for the weight.
- Grad-CAM does not need to rid off the FC layer. Furthermore, it can be applied for every task where CAM could be only applied on the Classification.



3

EXPERIMENTS

EXPERIMENTS

PRACTICE

Global Average Pooling

- The Network in Network (NIN) employed the GAP first for replacement of the FC layer.
- Without the FC layer, It shows the fine performance for the classification task.

Table 1. Classification error on the ILSVRC validation set.

Networks	top-1 val. error	top-5 val. error
VGGnet-GAP	33.4	12.2
GoogLeNet-GAP	35.0	13.2
AlexNet*-GAP	44.9	20.9
AlexNet-GAP	51.1	26.3
GoogLeNet	31.9	11.3
VGGnet	31.2	11.4
AlexNet	42.6	19.5
NIN	41.9	19.6
GoogLeNet-GMP	35.6	13.9

Except the Alex Net, the network shows almost similar performance when it lost the FC layer

- Activation map for each class
- We can interpret the machine's logic with the CAM.

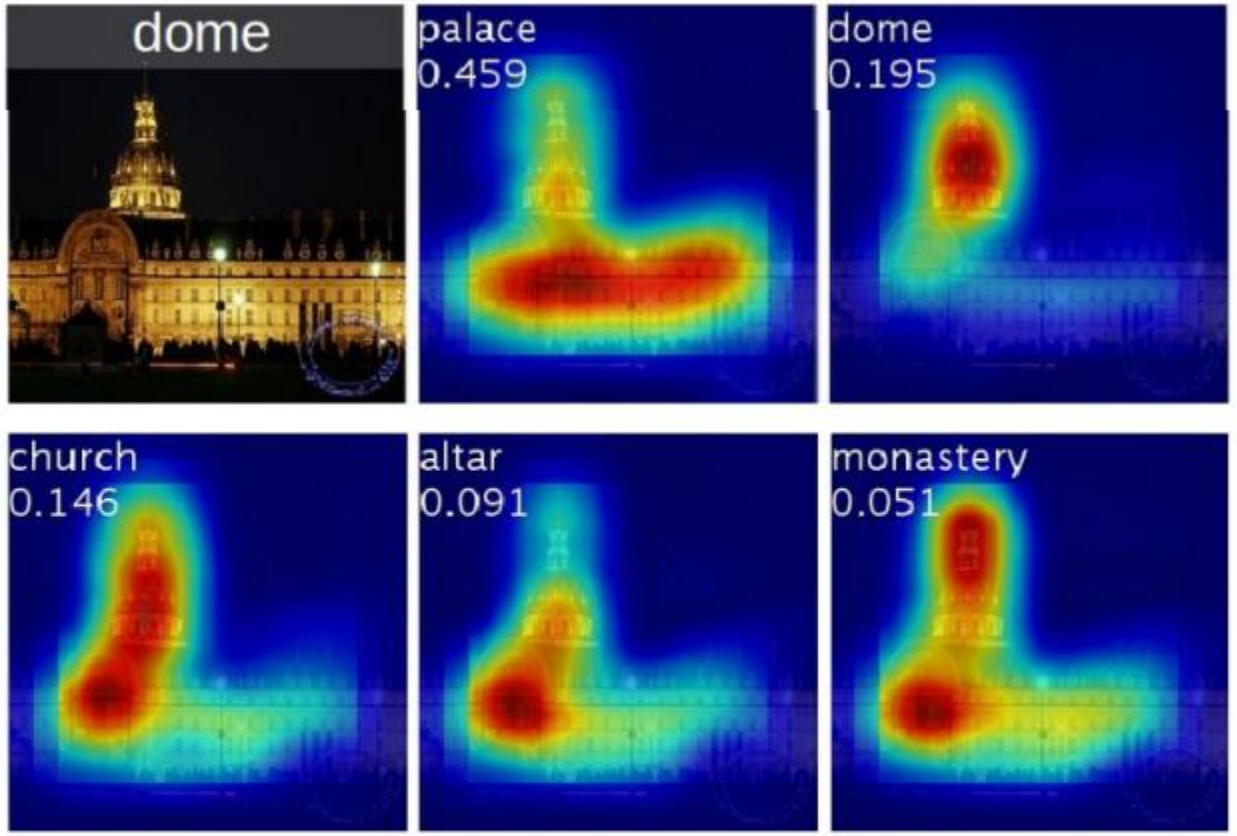
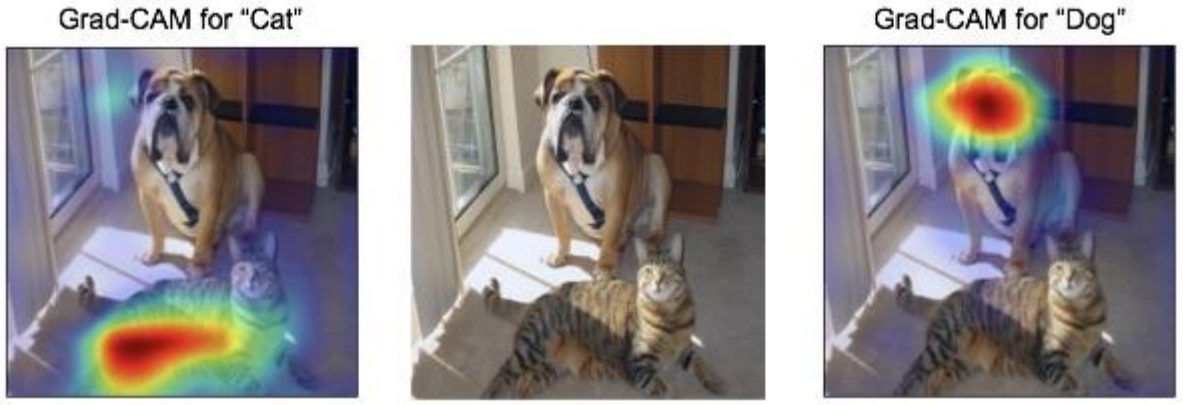


Figure 4. Examples of the CAMs generated from the top 5 predicted categories for the given image with ground-truth as dome. The predicted class and its score are shown above each class activation map. We observe that the highlighted regions vary across predicted classes e.g., *dome* activates the upper round part while *palace* activates the lower flat part of the compound.



Unsupervised (weakly supervised) Localization

Red box is the ground truth and green box is the prediction.

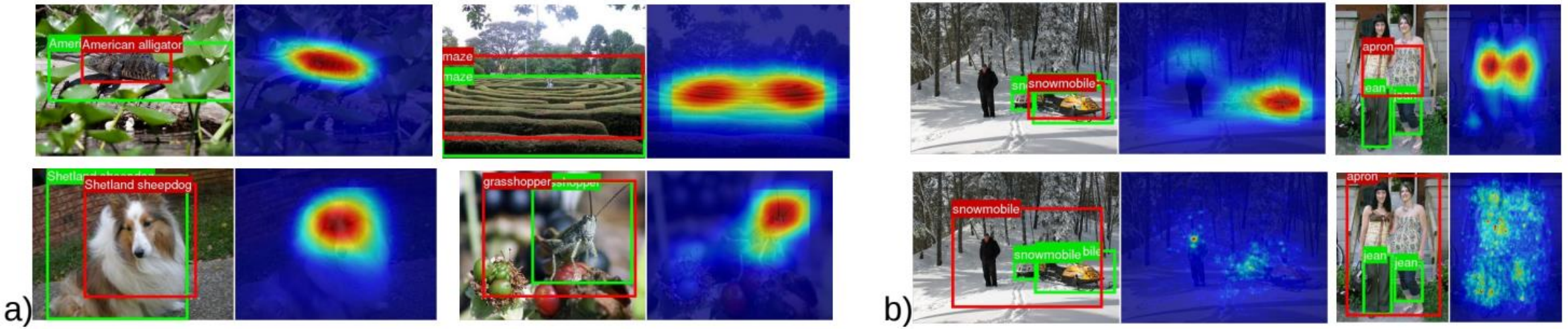


Figure 6. a) Examples of localization from GoogleNet-GAP. b) Comparison of the localization from GooleNet-GAP (upper two) and the backpropagation using AlexNet (lower two). The ground-truth boxes are in green and the predicted bounding boxes from the class activation map are in red.

Localization

- GAP models showed the noticeable performance drop.
- However, come to think about the setting (no location information at the training), this is an amazing performance.

Table 2. Localization error on the ILSVRC validation set. *Back-prop* refers to using [22] for localization instead of CAM.

Method	top-1 val.error	top-5 val. error
GoogLeNet-GAP	**56.40**	**43.00**
VGGnet-GAP	57.20	45.14
GoogLeNet	60.09	49.34
AlexNet*-GAP	63.75	49.53
AlexNet-GAP	67.19	52.16
NIN	65.47	54.19
Backprop on GoogLeNet	61.31	50.55
Backprop on VGGnet	61.12	51.46
Backprop on AlexNet	65.17	52.64
GoogLeNet-GMP	57.78	45.26

Table 3. Localization error on the ILSVRC test set for various weakly- and fully- supervised methods.

Method	supervision	top-5 test error
GoogLeNet-GAP (heuristics)	weakly	**37.1**
GoogLeNet-GAP	weakly	42.9
Backprop [22]	weakly	46.4
GoogLeNet [24]	full	26.7
OverFeat [21]	full	29.9
AlexNet [24]	full	34.2

Practical Characteristic

1. CAM does not work well for the high level (deeper) feature.

2. Sometimes the network does not see the object but catch the sufficient evidence.

3. CAM also work for the GMP, but it works better with the GAP.

4

DISCUSSION

ANALISYS

CONTRIBUTION

PROS & CONS

FURTHER WORK

Practical Characteristic

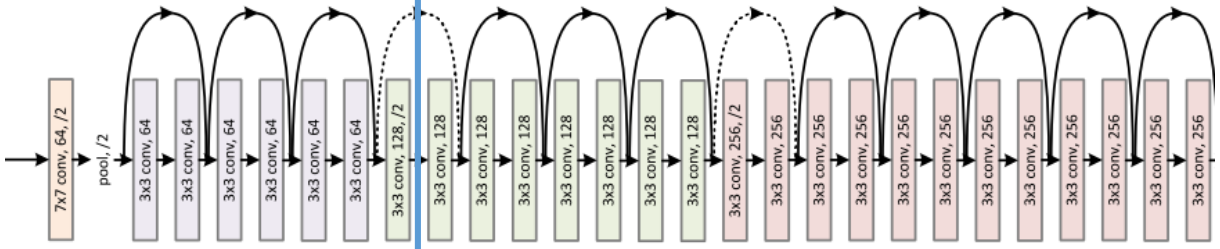
1. CAM does not work well for the high level (deeper) feature.

2. Sometimes the network does not see the object but catch the sufficient evidence.

3. CAM also work for the GMP, but it works better with the GAP.

Analysis

1. CAM does not work well for the high level (deeper) feature.
2. Sometimes the network does not see the object but catch the sufficient evidence.
3. CAM also work for the GMP, but it works better with the GAP.



Localization feature

High Level feature

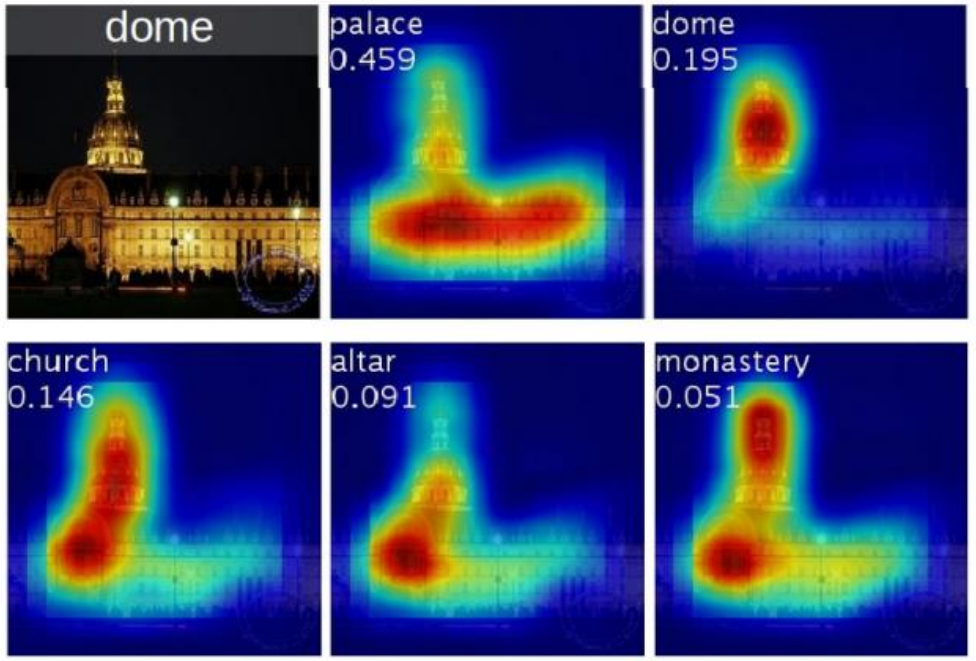
High level feature might not have the location feature but only have the semantic feature.

Analysis

1. CAM does not work well for the high level (deeper) feature.

2. Sometimes the network does not see the object but catch the sufficient evidence.

3. CAM also work for the GMP, but it works better with the GAP.



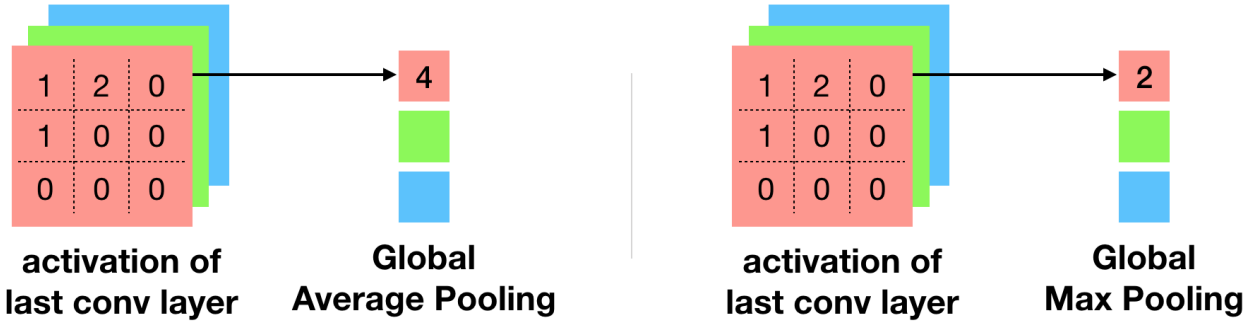
Since the network is not a human, computer only see the distribution of the image. In other word, if there some statistical evidence which is not the object, the network would catch that.

Analysis

1. CAM does not work well for the high level (deeper) feature.

2. Sometimes the network does not see the object but catch the sufficient evidence.

3. CAM also work for the GMP, but it works better with the GAP.

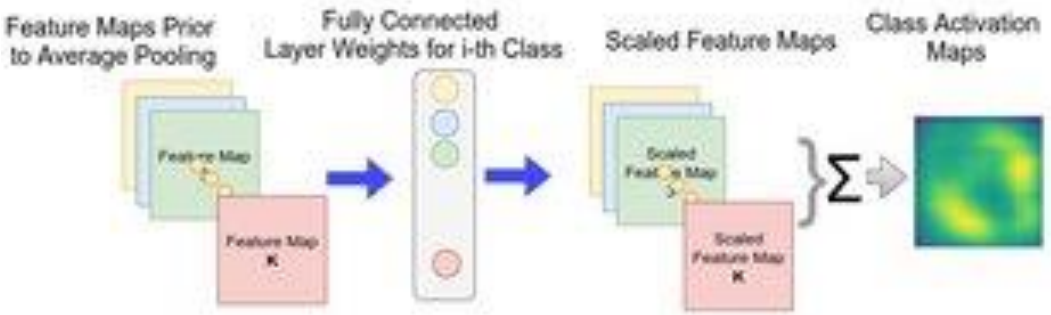


GMP only extract the one max value and lost the of the information about rest of the pixel.

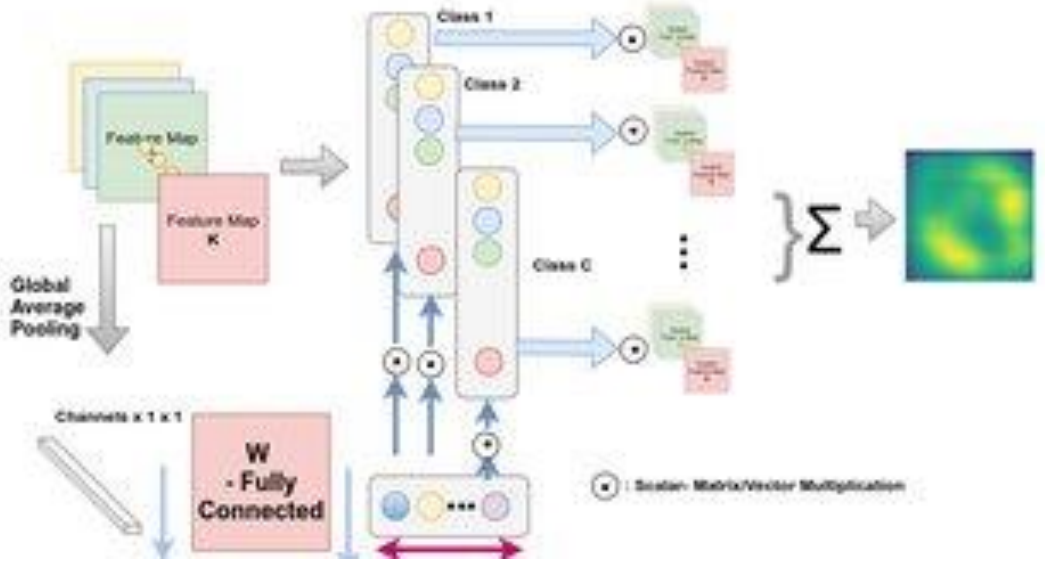
In contrast, GAP keep the information about the every pixel.

Contribution & Pros

- Current work was only available to extract the activated edge map from the CNN (Matthew et al., 2013)
- Object localization is available without additional annotation. (weakly supervised)
- This could be applied for the attention mechanism with the classifier unit.



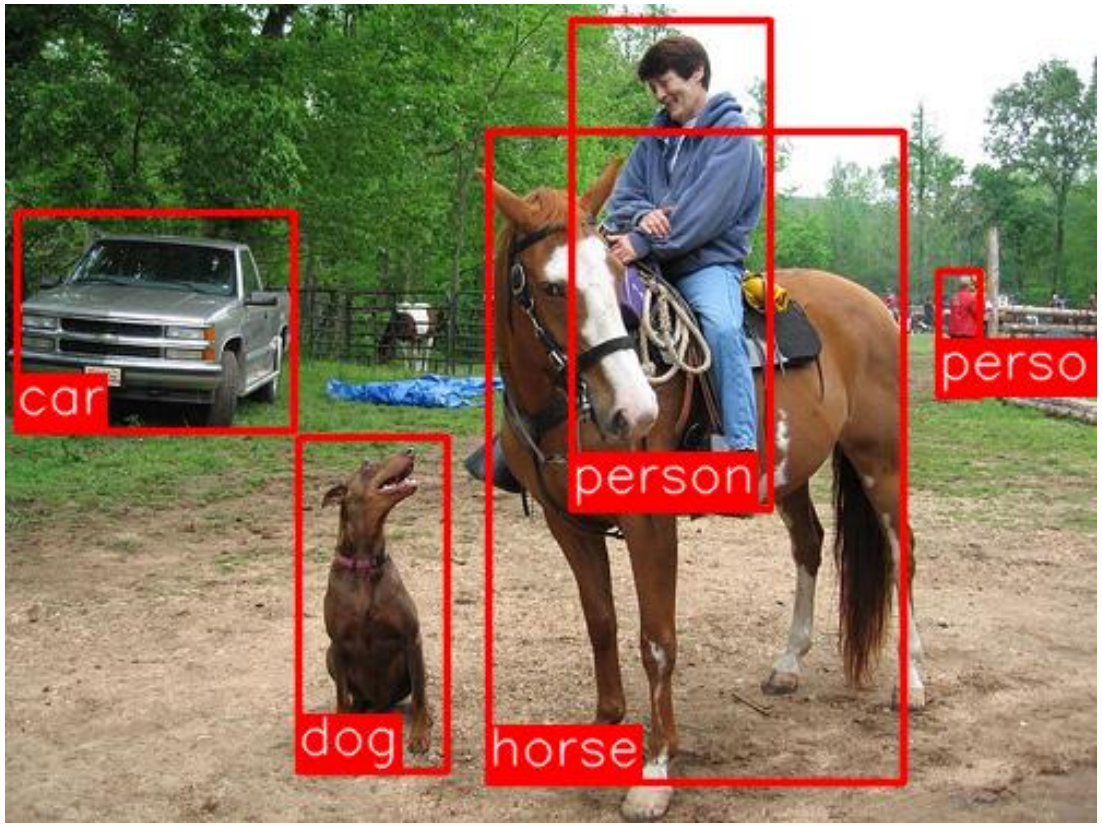
Proposed Method- Soft CAM



GAN + CAM

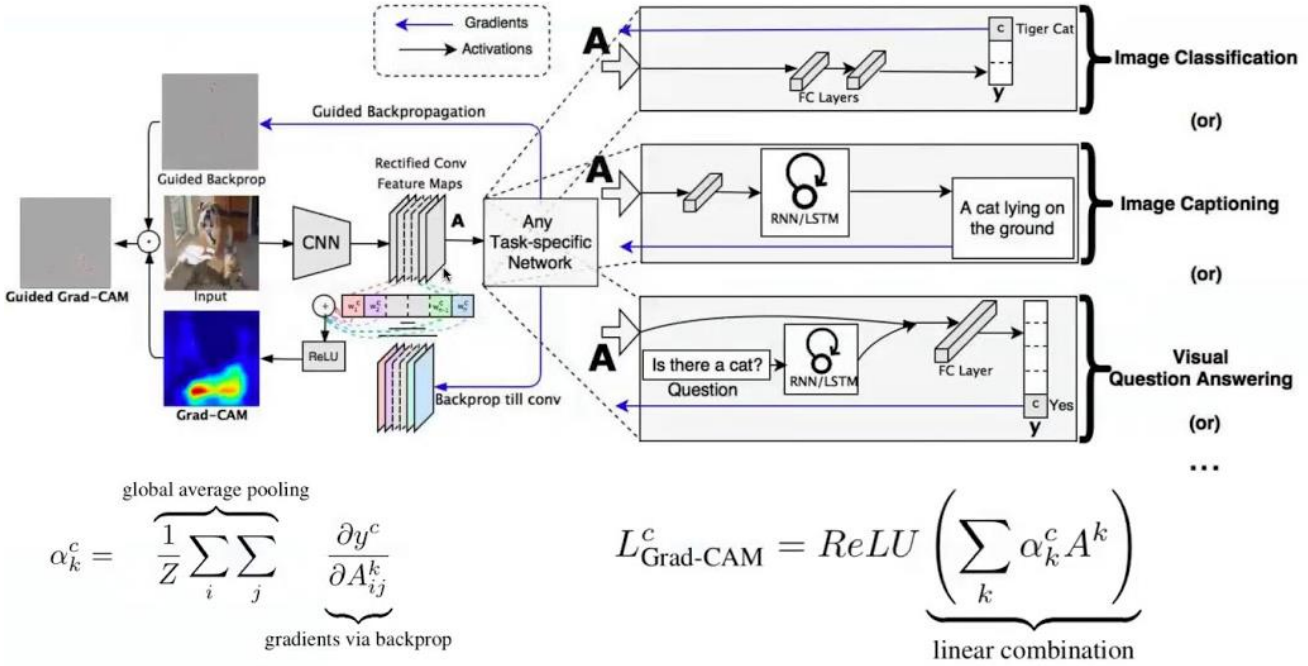
Cons

- CAM has lower performance compare to the supervised methods such as RCNN.
- Because this is not exactly the localization, sometimes CAM see the wrong position based on the machine's logic.



Cons for neat CAM
(Solved by the Grad CAM)

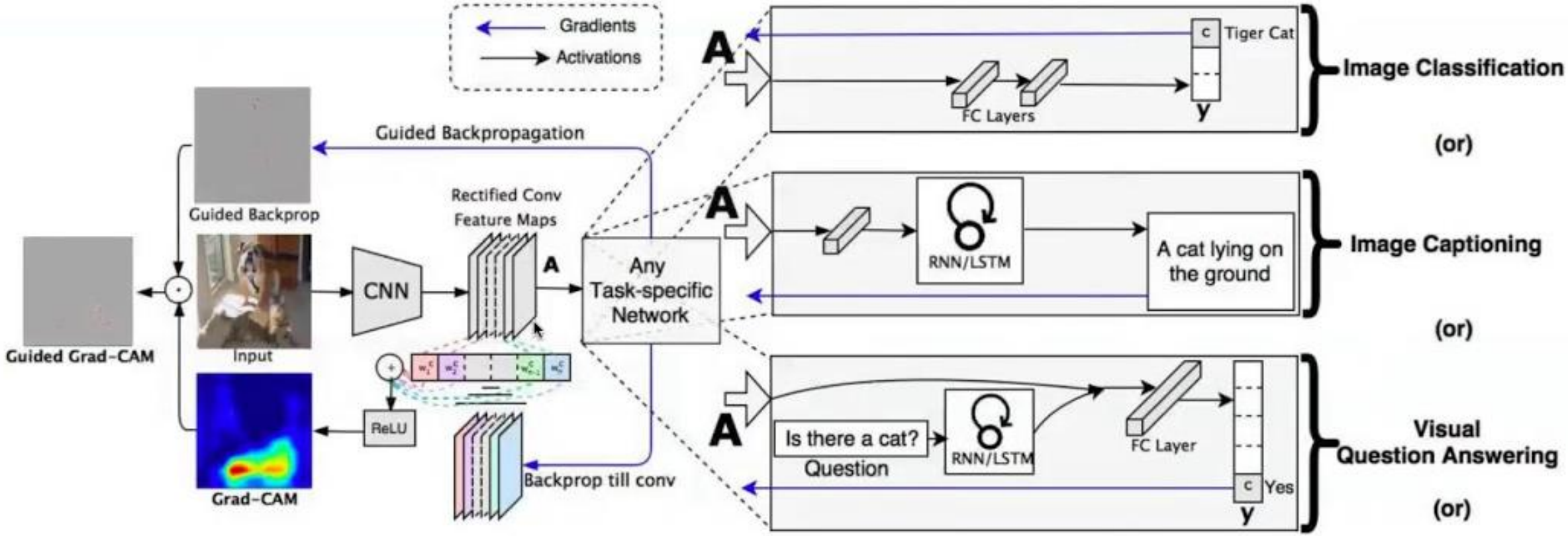
- This method is only for the classification task. Therefore, it couldn't be applied on the other tasks such as segmentation.
- FC removal is required which brings the performance drop & GAP network has different structure from the original network.



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

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

Grad cam can be applied for the Classification, Captionaing, Question answering…



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

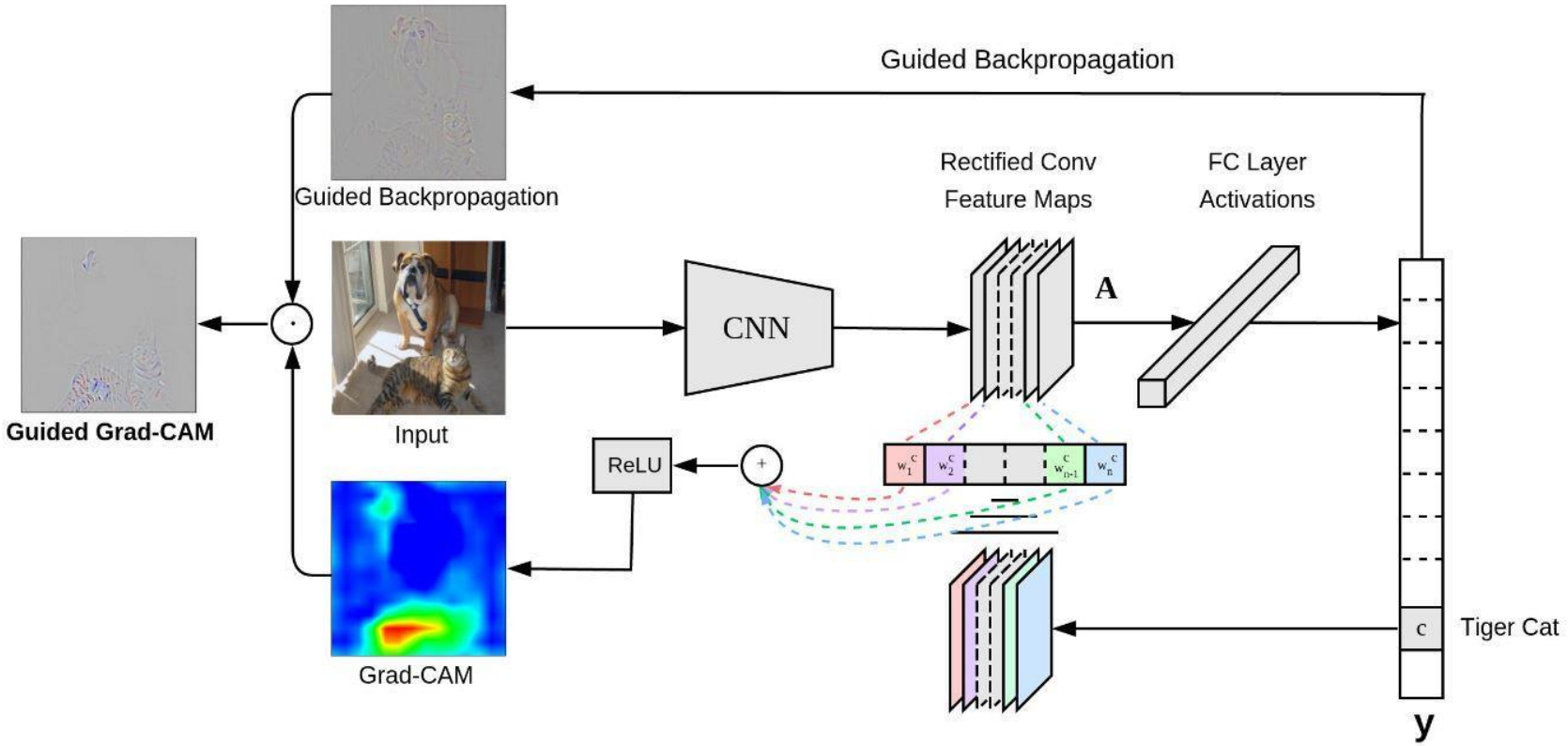
$$L^c_{\text{Grad-CAM}} = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

Further work

Grad CAM

Attention module

GAN discriminator



Thank You!

Reference

- M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. Proc. ECCV, 2014
- Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba, Learning Deep Features for Discriminative Localization, CVPR 2015
- M. Lin, Q. Chen, and S. Yan. Network in network. International Conference on Learning Representations, 2014
- Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, ICCV 2017

CAM / NIN / Grad-CAM / VGG / Visualizing CNN