DLLAB WEEK 7 Pre-Report : About Class Activation Maps

Hyunsoo Cha[1], Kwanghoon Sohn[2]
[1]Yonsei University. [2]Yonsei University.

1 Learning Deep Features for Discriminative Localization

1.1 Introduction

In recent studies, convolutional units of many layers of convolutional neural networks play a role in finding objects[5]. Although it is a fairly well-performed network, there is a disadvantage that using a fully-connected layer can cause local information to disappear. Therefore, it is difficult to know where an object is located when classifying it. They propose a structural regularizer to create high performance without using full-connected layers. The network implemented structural regularizer by using global average pooling. The role of this pooling is to prevent overfitting[1]. The advantage of this layer is its scalability.

Figure 1: As shown in the picture above, the researchers can see that CNN has successfully performed object categorization. they can see that they have successfully classified things that interact with humans rather than humans themselves. As shown in the photo above, it shows that brushing one's teeth is well caught, and that it also shows cutting wood.

The CAM approach is simple in common sense. Nevertheless, ILSVRC showed a top-5 test error of 37.1% in the wakely supervised object localization. This is a figure similar to AlexNet.

1.2 Class Activation Mapping

This is a description of how class activation maps can be created through global average pooling. A class activation map refers to a discretionary image region used in CNNs that recognize categories.

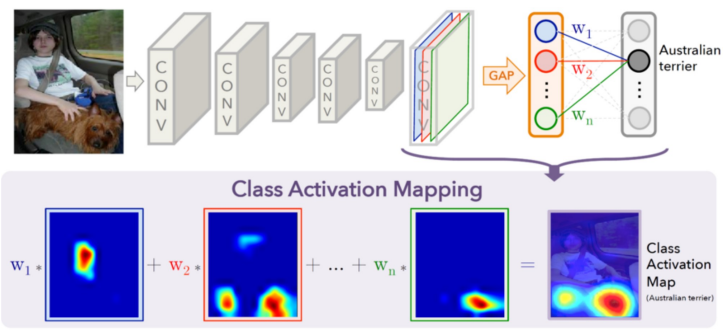

Figure 2: The figure above is a schematic illustration of class activation mapping. The classifying score returns to the convolutional layer before creating class activation maps. CAM is effective in identifying class-specific identification areas.

This network structure is similar to Network in Network or GoogLeNet. These networks are composed of convolutional layers, but the last layer is softmax for categorization. However, as it goes through the flatten, it loses the spatial information of adjacent pixels in the feature map. In addition, the number of fully connected layers input increases, resulting in a surge in the number of parameters.

To overcome this, researchers implemented global average pooling. As shown in the figure above, only one value of each color is taken from each channel. At the end, fully connected is applied. The weight for each layer is called $w_1, w_2, w_3, \cdots$. As shown in Figure, multiplying the feature map by weight becomes a class activation map. This allows visualization of CNNs.

Let $f_k(x,y)$ denote the activation of unit k in the last convolutional layer of space coordinates (x,y). Global average pooling can be represented as follows.

$$F_k = \sum_{x,y} f_k(x,y) \tag{1}$$

For class c, the input S_c of softmax can be expressed as follows.

$$S_c = \sum_k \omega_k^c \sum_{x,y} f_k(x,y) = \sum_{x,y} \sum_k \omega_k^c f_k(x,y) \tag{2}$$

And ω_k^c refers to the importance of F_k of class c. Finally, the softmax function can be represented as follows.

$$\textbf{Softmax for class } c, P_c = \frac{\exp(S_c)}{\sum_c \exp(S_c)} \tag{3}$$

The class activation map for class c can be defined as follows. This information includes spatial information.

$$M_c(x,y) = \sum_k \omega_k^c f_k(x,y), \quad S_c = \sum_{x,y} M_c(x,y) \tag{4}$$

In weakly supervised object localization, the global average pooling loss further identifies the range of objects compared to global max pooling. In the case of average, if the weight decreases due to low activation, the discriminative part of the object is maximized[2]. On the other hand, for max, the score is not affected because the low score in all image regions goes to the maximum.

1.3 Weakly-supervised Object Localisation

The researchers measure the localization possibility of the class average map through the ILSVRC 2014 benchmark. Setup is simply as follows. Thus, they let the resulting layer be 14×14. For GoogleLeNet, they delete the inception4e layer. Equally, they let 14×14 layer be the output size. In addition, they add 3×3, stride 1 convolutional layer.

For classification, comparisons are made using existing networks. The results are as follows. They can see that AlexNet is affected when the layer is removed the most significantly. Overall, models using GAP are competitive, but errors are larger than existing models.

1.4 Conclusion

In conclusion, global average pooling was used over class activation mapping for CNN. It allows CNN to classify learning that performs object localization. It have the advantage of visualizing a predicated class score for any input image. I was able to detect objects that I wanted to find relatively accurately. CAM technology can use in the field, saying fine-grained recognition, pattern discovery.

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

2.1 Introduction

From now on, I describe Grad-CAM compared to CAM. CAM is used only in a few CNNs. This is because global average pooling was used directly in front of the softmax layer just before the prediction via the convolution map. Grad-CAM generalizes CAM, giving it a wider range of applications for CNN architectures. For example, it can be used in the following models. Examples include CNN with full-connected layer, CNN with structured output such as captioned, CNN with multi-model input or reinforcement learning networks.

What would be a good visualization? A target category must be class-discriminative in order to make a good visual expansion. It also needs to be high resolution for good visual expansion.

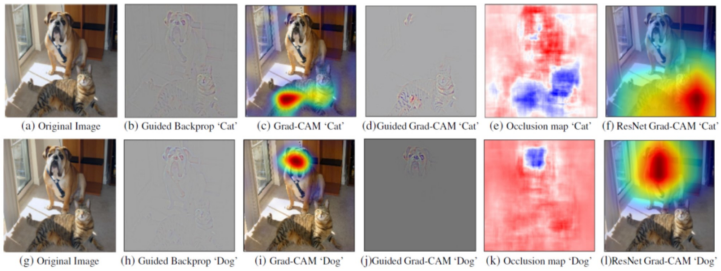

Figure 3: The figure above outputs the number of visualizations of the classes Tiger cat and Boxer dog. Guided Backpropagation[3] and Deconvolution[4] do not have good performance in distinguishing high resolution or class. In contrast, CAM and Grad-CAM show higher class-discriminative. In conclusion, the researchers achieved the following results.

For any CNN-based network, Grad-CAM is enabled without re-training or structural changes. Grad-CAM is also applicable as the highest performance classifier. They also provide a tool to interpret Grad-CAM visualization to diagnose failures in the model due to unknown biases in the dataset. They provide visualizations of ResNet applicable to VQA and image classification via Grad-CAM. Researchers then use neuron importance to name the Grad-CAM and use texture descriptions to determine models. Finally, the Guided Gramd-CAM description is class-discriminative and not only allows humans to be trusted, but also allows untrained users to be recognized from the weaker network as a stunner network.

2.2 Grad-CAM

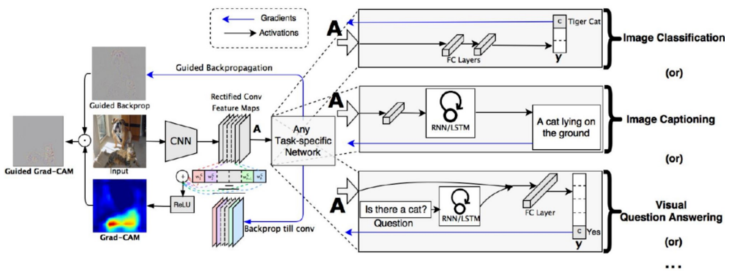

Figure 4: The figure above shows the overall schematic of Grad-CAM. In order to know the starting score of a category, the researchers first input the given image and the class that they want to classify. Gradient is set to zero for all classes, except for classes that you want to find. This signal is backpropagated with a corrected conv feature map. This is somewhat rough because it is the first result of creating a heat map for Grad-CAM. To obtain high-resolution and concept-specific guided grad-CAM visualizations, they multiply the heat map in the direction of the dots.

By using **class** c, **height** v, **width** u, they can get $L^c_{\textbf{Grad-CAM}} \in \Re^{u \times v}$ Grad-CAM class-discriminative localisation map. To do this, they can get neuron importance weight. $\alpha^c_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}}$.

The preceding summing means global average pooling and the posterior partial differentiation means gradients via backpropagation. This weight is called partial linearization. Starting from A, it is a linearization of deep network downstream. By passing through ReLU through a linear combination of α obtained in this way, we can obtain the following. $L^c_{\textbf{Grad-CAM}} = \textbf{ReLU}(\sum_k \alpha^c_k A^k)$.

The value in the ReLU function is called linear combination. They calculated linear combination. y^c is a differentiable activation function. This is activation that answers questions or includes captions. When they define global average pooled output as follows. $F^k = \frac{1}{Z}\sum_i \sum_j A^k_{ij}$. CAM calculates the final score as follows. $Y^c = \sum_k \omega^c_k \cdot F^k$.

ω^c_k stands for the weight of the k^{extth} feature map. They can differentiate the score with a feature map to obtain gradients.

$$\frac{\partial Y^c}{\partial F^k} = \frac{\frac{\partial Y^c}{\partial A^k_{ij}}}{\frac{\partial F^k}{\partial A^k_{ij}}}, \quad \omega^c_k = Z \cdot \frac{\partial Y^c}{\partial A^k_{ij}} \tag{5}$$

Consequently, an entity may write:

$$\omega^c_k = \sum_i \sum_j \frac{\partial Y^c}{\partial A^k_{ij}} \tag{6}$$

2.3 Results

Classificaiton and Localisation They measure the accuracy of Grad-CAM with VGG-16, AlexNet, and GoogLeNet. Calculate the classification, localization of top1 and top5. Grad-CAM for VGG-16 recorded the lowest error rate. This is a better result than CAM[6]. More detail results on Figure 5.

		Classification		Localization	
		Top-1	**Top-5**	**Top-1**	**Top-5**
VGG-16	Backprop [51]	30.38	10.89	61.12	51.46
	c-MWP [58]	30.38	10.89	70.92	63.04
	Grad-CAM (ours)	30.38	10.89	**56.51**	46.41
	CAM [59]	33.40	12.20	57.20	**45.14**
AlexNet	c-MWP [58]	44.2	20.8	92.6	89.2
	Grad-CAM (ours)	44.2	20.8	68.3	56.6
GoogleNet	Grad-CAM (ours)	31.9	11.3	60.09	49.34
	CAM [59]	31.9	11.3	60.09	49.34

Figure 5: The brief result table of classification and localisation with three networks.

Segmentation Grad-CAM is also doing well in semantic segmentation. In detail, it shows better results than ground-truth. The IoU value is 49.6, which is better than 44.6 in CAM.

Visualizations The result was that Deconvolution was more class discriminative than Guided Backpropagation. This reliability starts on the grounds that VGG-16 is more reliable than AlexNeet. VGG-16 is 79.09 mAP. Subjective human evaluation determined that the classifier of VGG-16 beyond AlexNet is more accurate. With Guided Backpropagation, humans set VGG-16 to an average of 1.00. This is slightly more reliable than AlexNet. Grad-CAM earned 1.27.

2.4 Conclusion

Grad-CAM allowed explicit visual expansion to be generated for any CNN-based model. Existing high-resolution visual technologies can be integrated into Grad-CAM localization. The visualizations produced by the researchers were approached with interpretability and faithfulness. With the expansion of human engineering, they find that the human visual system can classify classes more accurately with assistance from bias in datasets and trustworthiness in classifiers. Finally, the researchers built Grad-CAM to perform tasks.

3 Reference

[1] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[2] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 685–694, 2015.

[3] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[4] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[5] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014.

[6] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.