Robust Discriminative Localization Maps

Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, James Storer
Brandeis University
{aprakash,nemtiax,solomongarber,dilant,storer}@brandeis.edu

Abstract

Activation maps obtained from CNN filter responses have been used to visualize and improve the performance of deep learning models. However, as CNNs are susceptible to adversarial attack, so are the activation maps. While recovering the predictions of the classifier is a difficult task and often requires complex transformations, we show that recovering activation maps is trivial and does not require any changes either to the classifier or the input image.

Code: github.com/iamaaditya/robust-activation-maps

1. Activation Maps

Convolution filters learned from natural images are sensitive to the object classes in the training data [7], [5]. In other words, regions of an image which contain relevant objects tend to produce high activations when convolved with learned filters, while regions that do not contain any object class have lower activations, and this is particularly true for the deeper levels of the CNN. Taking the spatial average of the feature map (global average pooling) at the last convolution layer reveals possible locations of objects within the image [7], [10]. These kind of activations maps are popularly known as Class Activation Maps (CAM).

Consider a convolutional network with k output channels on the final convolution layer (f) with spatial dimensions of x and y, and let $\boldsymbol{w}$ be a vector of size k which is the result of applying a global max pool on each channel. This reduces channel to a single value, $\boldsymbol{w_k}$. The class activation map, M_c for a class c is given by:

$$M_c(x, y) = \sum_k \boldsymbol{w}_k^c \; f_k(x, y) \tag{1}$$

Class activation maps are a valuable tool for approximate semantic object localization. CAMs have been used to improve image classification [10], visual question answering [11], semantic image compression [8] and image retrieval [2] and as attention networks [4] for various computer vision related tasks.

2. Impact of adversary

It has been established that most image classification models can easily be fooled [9, 1]. Several techniques have been proposed which can generate an image that is perceptually indistinguishable from another image but is classified differently. This can be done robustly when model parameters are known, a paradigm called *white-box attacks* [1, 3, 6].

Generally, one is interested in the activation map for the class for which the model assigns the highest probability. However, in the presence of adversarial perturbations to the input, the highest-probability class is likely to be incorrect. This implies that Class Activation Maps will also be incorrect for these adversarial images. The left column of 2 depicts CAM for clean images. The middle column shows CAM for the same image after it has been altered by the adversary. In all cases, it is evident that CAM under an adversary is an incorrect depiction of the object and its associated activations

Fortunately, our experiments show that an adversary which successfully changes the most likely class tends to leave the rest of the top-k predicted classes unchanged. Our experiments show that 38% of the time the predicted class of adversarial images is the second highest class of the model for the clean image.

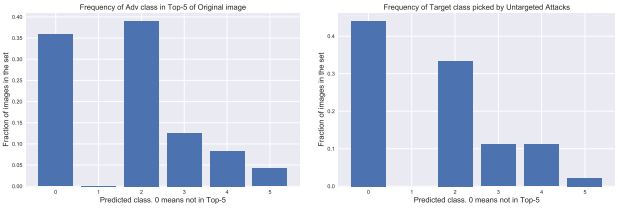

Figure 1. Left: Rank of adversarial class within the top-5 predictions for original images. Right: Rank of original class within the top-5 predictions for adversarial images. In both cases, 0 means the class was not in the top-5.

In 64% of the cases the mis-classified class was still within Top-5 picks barring the most likely class. As seen in Figure 1, the predicted class of the perturbed image is very frequently among the classifier's top-5 predictions for

the original image. In fact, nearly 40% of the time, the adversarial class was the second most-probable class of the original image.

ImageNet has one thousand classes, many of which are fine-grained. Frequently, the second most likely class is a synonym or close relative of the main class (*e.g.* "Indian Elephant" and "African Elephant"). To obtain a map which is robust to fluctuations of the most likely class, we take an exponentially weighted average of the maps of the top-k classes.

$$\widehat{M}(x,y) = \sum_{i}^{k} \frac{M_{c_i}(x,y)}{2^i} \tag{2}$$

We will refer to this as robust activation maps ($\widehat{M}(x,y)$). We normalize the map by diving it by its max so that values are in the range of $[0,1]$. Even if the top-1 class is incorrect, this averaging reduces the impact of mis-localization of the object in the image.

The appropriate number of classes k to average over depends on the total number of classes. For ImageNet-1000, we used a fixed $k = 5$. While each possible class has its own class activation map (CAM), only a single robust activation map is generated for a particular image, combining information about all classes. ImageNet covers wide variety of object classes and most structures found in other datasets are represented in ImageNet even if class names are not bijectional. Therefore, Robust Activation Maps is trained once on ImageNet but can also localize objects from Pascal-VOC or Traffic Signs.

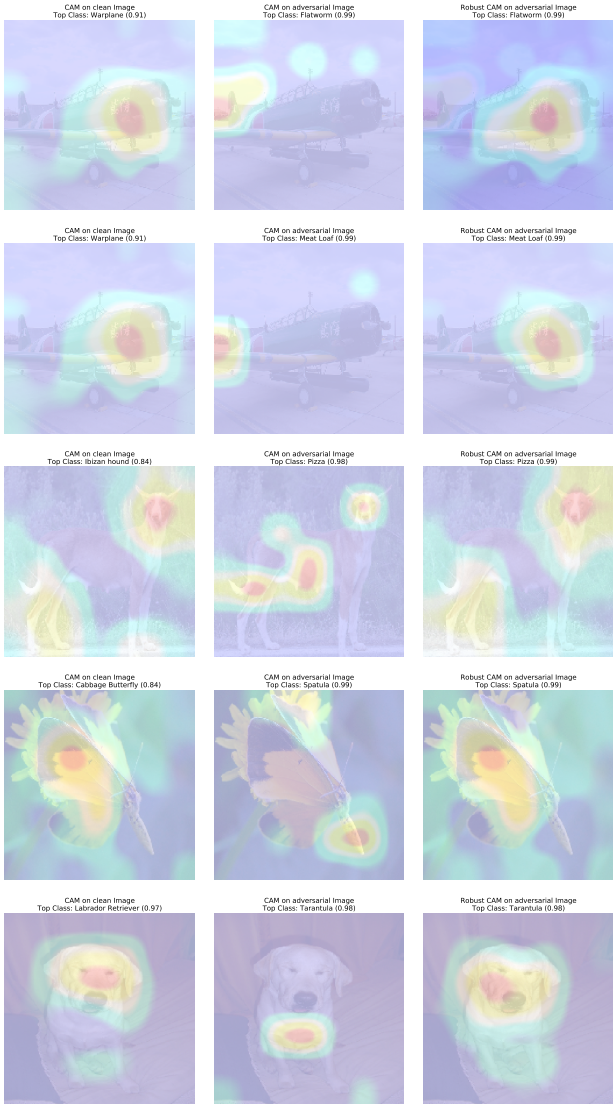

Figure 2. Difference between class activation maps and robust activation maps under the presence of an adversary.

References

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014.

[2] A. Jimenez, J. M. Alvarez, and X. Giró. Class-weighted convolutional features for visual instance search. *CoRR*, abs/1707.02581, 2017.

[3] A. Kurakin, I. J. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *CoRR*, abs/1607.02533, 2016.

[4] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu. Tell me where to look: Guided attention inference network. *CoRR*, abs/1802.10171, 2018.

[5] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[6] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *CoRR*, abs/1706.06083, 2017.

[7] M. Oquab, L. Bottou, I. Laptev, and J. 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*, 2015.

[8] A. Prakash, N. Moran, S. Garber, A. DiLillo, and J. Storer. Semantic perceptual image compression using deep convolution networks. *2017 Data Compression Conference (DCC)*, 2017.

[9] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013.

[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[11] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus. Simple baseline for visual question answering. *CoRR*, abs/1512.02167, 2015.