Course Project: CS689
Github Repo

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

Abhinav Kumar 180050003

Advisor: Prof Harish Guruprasad Ramaswamy

1 Introduction

Deep Neural Networks have largely surpassed previous techniques such as Decision Trees, Kernel SVMs etc in terms of accuracy. However, this superior performance comes at the cost of lack of simple, interpretable building blocks in the model. Consequently, when today's intelligent systems fail, they often fail spectacularly disgracefully without warning or explanation, leaving a user staring at an incoherent output, wondering why the system did what it did. In order to build trust in intelligent systems and move towards their meaningful integration into our everyday lives, it is clear that we must build 'transparent' models that have the ability to explain why they predict what they predict.

Zhou et al. recently proposed a technique called Class Activation Mapping (CAM) for identifying discriminative regions used by a restricted class of image classification CNNs which do not contain any fully-connected layers. This paper discuses a more general method to visualise the decisions for the model . Broadly, the authors of the paper introduce Grad-CAM, a class-discriminative localization technique that generates visual explanations for any CNN-based network without requiring architectural changes or re-training and show the effect on GradCam to adverserial inputs.

2 Related Work

Visualising CNNs A number of previous works [1] [2] [3] have visualized CNN predictions by highlighting 'important' pixels (i.e. change in intensities of these pixels have the most impact on the prediction score). Specifically, Simonyan et al. [1] visualize partial derivatives of predicted class scores w.r.t. pixel intensities, while Guided Backpropagation [2] and Deconvolution [3] make modifications to 'raw' gradients that result in qualitative improvements. Despite producing fine-grained visualizations, these methods are not class discriminative.

Most relevant to our approach is the Class Activation Mapping (CAM) approach to localization [4]. This approach modifies image classification CNN architectures replacing fully-connected layers with convolutional layers and global average pooling, thus achieving class-specific feature maps.

3 Method

In the experiments, we have implemented both GradCAMs as well as Guided Backpropagation [2] based Guided GradCam. In this section, we discuss both the methods.

3.1 GradCam

A number of previous works have asserted that deeper representations in a CNN capture higher-level visual constructs. Furthermore, convolutional layers naturally retain spatial information which is lost in fully-connected layers, so we can expect the last convolutional layers to have the best compromise between high-level semantics and detailed spatial information. The neurons in these layers look for semantic class-specific information in the image (say object parts). In order to obtain the class discriminative localization map $L^c_{GradCam} \in R^{uxv}$, we first compute the gradient of score of class c i.e y^c w.r.t class activation A^k of a convolutional layer. These gradients then are global-averaged-pooled over width and height to obtain neuron importance weight α^c_k for each channel in the layer.

$$\alpha^c_k = (1/Z) \sum_i \sum_j \delta y^c / \delta A^k \tag{1}$$

To obtain the gradCam Map, we take a weighted combination of α^c_k with the class activation maps following by a ReLU operation.

$$L^c_{GradCam} = ReLU(\sum_k \alpha^c_k A^k) \tag{2}$$

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 y^c Negative pixels are likely to belong to other categories in the image.

3.2 Guided GradCam

While Grad-CAM is class-discriminative and localizes relevant image regions, it lacks the ability to highlight finegrained details like pixel-space gradient visualization methods. Guided Backpropagation visualizes gradients with respect to the image where negative gradients are suppressed when backpropagating through ReLU layers. Intuitively, this aims to capture pixels detected by neurons, not the ones that suppress neurons.

In order to combine the best aspects of both, we fuse Guided Backpropagation and Grad-CAM visualizations via element-wise multiplication ($L^c_{GradCam}$ is first upsampled to the input image resolution using bilinear interpolation).

4 Experiments and Results

In this section, we summarise the results obtained on sample images from the ImageNet dataset using Vgg16 model. We also present the results of GradCam on adverserially perturbed images

4.1 GradCam visualization

The GradCam output has been superimposed on the original image after resizing it to original image size using bilinear interpolation so that meaningful observations can be

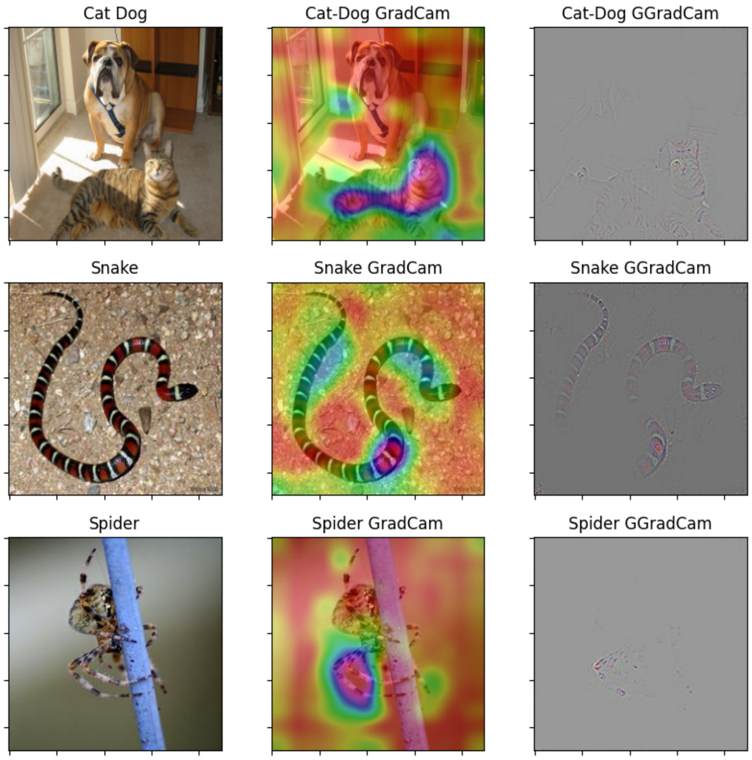

Figure 1: GradCam and Guided GradCam output

made in the image such as in the first image, for the class label "cat" the gradCam assigns much importance to the head and front body. Alongside, the plots for guided gradCam also have been provided which are more fine-grained as compared to the heat maps.

4.2 Effect of adverserial noise on VGG-16

For generating the adverserial examples, we employ the Fast Gradient Sign Method introduced by GoodFellow et al. [5] and then produce the GradCam output for the originally predicted class and the predicted class on adverserially perturbed image.

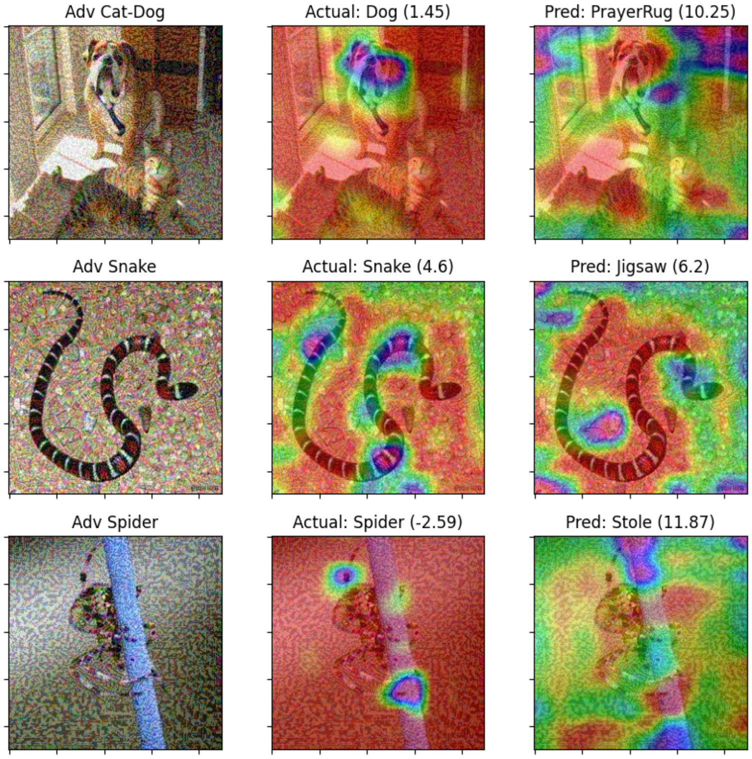

Figure 2: Robustness of GradCam to Adverserial Perturbation

As evident from Fig.2, for the correct class in the adverserial image, the GradCam output indeed matches the intuition while the output for the predicted class is not meaningful. For e.g in the second image, the highest value in the map is on regions of the floor which is unrelated to the correct decision making.

Another observation which can be made is that, by using GradCam, we can detect why the model is failing and thus make changes in the dataset or discover some hidden biases in it.

5 Conclusion & Future Work

In this project, we surveyed various class activation mapping based visualisation techniques along with their downsides. We also discussed the importance on GradCam in discovering why model may be making irrelevant prediction on some images and the robustness of GradCam to adverserial pertubations. As a future work, we plan to experiment with the use of GradCam in weakly supervised semantic segmentations as discussed in some recent works [6]

References

[1] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

[2] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net, 2015.

[3] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks, 2013.

[4] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization, 2015.

[5] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

[6] Alexander Kolesnikov and Christoph H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation, 2016.