class in case there are more than one, where the particular class can be positioned in a scattered or attached manner in the image. Thus, the heatmap generated better explains the model's behaviour when multiple instances of a single class are present in an image. This is shown with appropriate examples and comparisons with Grad-CAM results in section 3.

- Grad-CAM++ can localize the predicted class more accurately than Grad-CAM, which increases faithfulness to the model. We generated heatmaps for both the techniques (Grad-CAM and Grad-CAM++) and fused it with the actual image via point-wise multiplication. The behaviour of the confidence score (of the deep network) for that particular class, when presented with the original image and heatmap-fused image, can be analyzed to conclude which method generates a more class-specific heatmap. A lower or no drop in class score would indicate a higher localization of class-discriminative regions for a particular class. We provide results for this experiment in Section 4.1 which show a better localization capacity of Grad-CAM++ over Grad-CAM.
- We generated heatmaps with both Grad-CAM++ and Grad-CAM for a considerable number of images and

straight-forward technique is to visualize the layer activations, where they generally show sparse and localized patterns. Visualizing the Convolution filters is also useful as as they reveal the nature of content extracted by the network in a given layer. Another method is to visualize the images which maximally activate a certain neuron in a trained network. This involves a large dataset of images to be feed forwarded through the network to understand what that neuron is looking for.

Zeiler & Fergus [18] proposed the deconvolution approach to better understand what the higher layers in a given network has learnt. "Deconvnet" makes data flow from a neuron activation in the higher layers, down to the image. In this process, parts of the image that strongly activate that neuron gets highlighted. Later, Springenberg *et al*. [16] extended this work to a new method called *guided backpropagation* which helped understand the impact of each neuron in a deep network w.r.t the input image. These visualization techniques were compared in [9]. Yosinski *et al*. [17] proposed a method to synthesize the input image, that causes a specific unit in a neural network to have a high activation, for visualizing the functionality of the unit. A more guided approach to synthesizing input images that maximally activate a neuron was proposed by Simonyan *et al*. [14]. In this work, they generated class-specific saliency maps, by per-

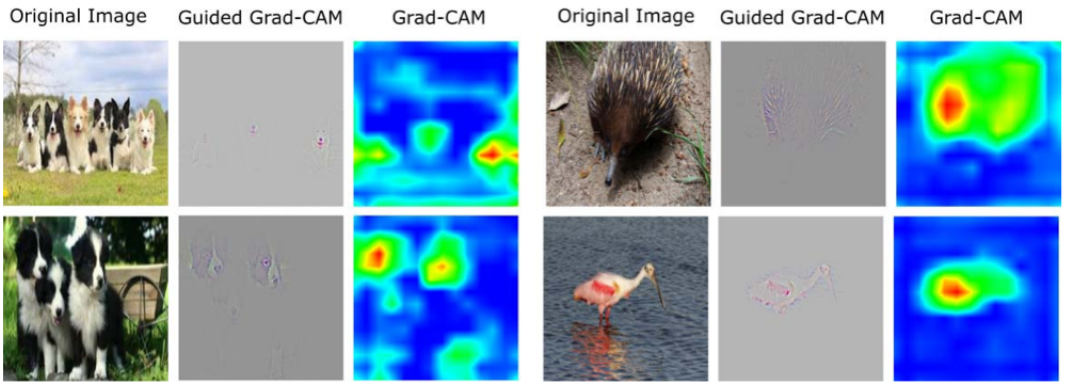

Figure 1. Weaknesses of Grad-CAM (a) multiple occurrences of the same class (Columns 1-3), and (b) localization capability of an object in an image (Columns 4-6). Note: All dogs are not visible for the first image under column 1. Full portion of the dogs are not visible for the second image under column 1. Both the classes are not visible in entirety under column 4 (the hedgehog's nose and the bird's legs are missing in the generated saliency maps).

forming a gradient ascent in pixel space to reach a maxima. This synthesized image serves as a class-specific visualization and enables us to delve deeper inside a CNN.

More recently, Ribeiro *et al*. [12] introduced a method

tions of the results of deep models used across domains, which can then be investigated upon failures to make them more robust. Another important reason is to build trust in these systems for their proper integration into our daily lives