# Research Questions

Amartya Dutta

## 1 Abstracts of Relevant Papers

The first paper that is related to my area is Learning Deep Features for Discriminative Localization. The abstract of the paper is as follows:
In this work, we revisit the global average pooling layer proposed in [13], and shed light on how it explicitly enables the convolutional neural network to have remarkable localization ability despite being trained on image-level labels. While this technique was previously proposed as a means for regularizing training, we find that it actually builds a generic localizable deep representation that can be applied to a variety of tasks. Despite the apparent simplicity of global average pooling, we are able to achieve 37.1% top-5 error for object localization on ILSVRC 2014, which is remarkably close to the 34.2% top-5 error achieved by a fully supervised CNN approach. We demonstrate that our network is able to localize the discriminative image regions on a variety of tasks despite not being trained for them.
While the abstract of this paper does not specifically mention the problem that exists and what it tries to solve. Similarly, there's no mention of why the problem even counts as a problem in the first place. However, they do mention the method they employ. They use the Global Average pooling while training deep neural networks to not only classify objects but also produce a heatmap indicating the location of the object in the image. The proposed method is called the "Convolutional Neural Network (CNN) with a Localization (Loc) layer," and it allows for discriminative localization of objects in images using a single forward pass of the network. The results improved object localization. It led to a better understanding of deep neural networks. Furthermore, the approach proposed in the paper could have been extended to tasks such as semantic segmentation, instance segmentation, and image captioning. Thus making the work important to the field of computer vision.
A second paper relevant to my area is Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. The abstract for this is as follows:
We propose a technique for producing "visual explanations" for decisions from a large class of CNN-based models, making them more transparent. Our approach - Gradient-weighted Class Activation Mapping (Grad-CAM), uses the gradients of any target concept, flowing into the final convolutional layer to produce a coarse localization map highlighting important regions in the image for predicting the concept. Grad-CAM is applicable to a wide variety of CNN model-families: (1) CNNs with fully-connected layers, (2) CNNs used for structured outputs, (3) CNNs used in tasks with multimodal inputs or reinforcement learning, without any architectural changes or re-training. We combine Grad-CAM with fine-grained visualizations to create a high-resolution class-discriminative visualization and apply it to off-the-shelf image classification, captioning, and visual question answering (VQA) models, including ResNet-based architectures. In the context of image classification models, our visualizations (a) lend insights into their failure modes, (b) are robust to adversarial images, (c) outperform previous methods on localization, (d) are more faithful to the underlying model and (e) help achieve generalization by identifying dataset bias. For captioning and VQA,

we show that even non-attention based models can localize inputs. We devise a way to identify important neurons through Grad-CAM and combine it with neuron names to provide textual explanations for model decisions. Finally, we design and conduct human studies to measure if Grad-CAM helps users establish appropriate trust in predictions from models and show that Grad-CAM helps untrained users successfully discern a 'stronger' model from a 'weaker' one even when both make identical predictions.
Though the abstract of the paper does not explicitly mention the underlying problem however with contextual knowledge it can be understood from one of the claims of the paper. They mention that Grad-CAM is applicable to a wide variety of CNN model families which is a problem that existed prior to that. And since the previous method wasn't generic enough, it was a problem that required solving. Hence, they propose a method for generating visual explanations of the predictions made by deep neural networks. The Grad-CAM method generates a heatmap that highlights the regions of an image that are most important for a particular prediction made by a deep neural network. However, unlike previous methods for generating visual explanations from deep neural networks, Grad-CAM does not require any modifications to the original network architecture or training procedure. Finally, their methods lend insights into the robustness of adversarial images, outperform previous methods on localization, are more faithful to the underlying model, and help achieve generalization by identifying dataset bias, thus making it a very important contribution.
In the third paper titled, the Threshold Matters in WSSS: Manipulating the Activation for the Robust and Accurate Segmentation Model Against Thresholds the abstract is as follows:
Weakly-supervised semantic segmentation (WSSS) has recently gained much attention for its promise to train segmentation models only with image-level labels. Existing WSSS methods commonly argue that the sparse coverage of CAM incurs the performance bottleneck of WSSS. This paper provides analytical and empirical evidence that the actual bottleneck may not be sparse coverage but a global thresholding scheme applied after CAM. Then, we show that this issue can be mitigated by satisfying two conditions; 1) reducing the imbalance in the foreground activation and 2) increasing the gap between the foreground and the background activation. Based on these findings, we propose a novel activation manipulation network with a per-pixel classification loss and a label conditioning module. Per-pixel classification naturally induces two-level activation in activation maps, which can penalize the most discriminative parts, promote the less discriminative parts, and deactivate the background regions. Label conditioning imposes that the output label of pseudo-masks should be any of true image-level labels; it penalizes the wrong activation assigned to non-target classes. Based on extensive analysis and evaluations, we demonstrate that each component helps produce accurate pseudo-masks, achieving the robustness against the choice of the global threshold. Finally, our model achieves state-of-the-art records on both PASCAL VOC 2012 and MS COCO 2014 datasets.
This abstract states that the existing WSSS methods show that the sparse coverage of CAM incurs the performance bottleneck of WSSS. This is a concerning issue because if the problem gets bottlenecked, further improvements stop, and achieving a higher segmentation score becomes difficult. Thus, they propose a novel activation manipulation network with a per-pixel classification loss and a label conditioning module to mitigate the bottleneck condition. Since it achieves State-of-the-art performance on two benchmark datasets, it's of high importance to the community.

## 2 Research Goal

The existing state-of-the-art methods in WSSS employ a Class Activation Map (CAM) based method. The reason is that Saliency Maps are relatively noisier. However, CAMs are only able

to detect the most discriminative regions of an object. In order to detect the rest of the image, a lot of post-processing is done which requires a huge amount of computational time. Furthermore, the results rely also on the post-processing rather than simply the initial method. Therefore, we are suggesting a return to Saliency Maps which does not require such post-processing. We propose an accumulation-based Saliency Map approach that effectively handles the noisy aspect of Saliency Maps. Our method breaks the structural correlation between the components of the image. This allows the classifier to also look at regions of the object it doesn't otherwise, thus resulting in the Saliency Maps detecting even the non-discriminative regions of the object. This method cannot be employed on CAMs because they cannot be backpropagated upon. Thus, using the accumulation-based method which breaks the structural correlation of the image, we aim to achieve state-of-the-art performance in WSSS.