Literature Review

Amartya Dutta

1 Weakly supervised semantic segmentation

Weakly supervised semantic segmentation (WS3) aims to train segmentation models using coarse-scale annotations that are not as precise as pixel-level locations of objects. In recent years, several WS3 methods have been proposed that use class labels to generate pseudo-ground truths for training segmentation models. Specifically, these methods employ localization maps, such as Class Activation Maps (CAMs) [7, 5], which are generated from a pre-trained classifier, to guide the segmentation process.

Recently a large body of work has focused on providing explanations of the model outputs which can potentially satisfy regulatory experiments, help practitioners debug their model and identify unintended bias in the model. This area of model explainability techniques can be categorized into two broad categories: (1) activation map based methods [7, 5], (2) attribution map based methods [6, 1].

Activation maps are feature maps from the last layer of CNNs which show regions in an image that are most responsible for the final prediction of the network. The activations are aggregated to generate a heatmap, highlighting the regions in the image that have the highest activations. One of the most popular activation map based methods is Class Activation Maps (CAMs) [7], where the activation map of the last CNN layer is combined with weights of the Global Average Pooling (GAP) layer to generate the visual explanations. Gradient-weighted Class Activation Mapping (Grad-CAM) [5] is an extension of CAM that uses gradient information to improve the localization accuracy of the generated heatmaps. **Attribution maps** on the other hand are a form of interpretability technique that attempts to assign a score to each pixel in an image, representing its contribution to the final prediction of the network. There are various methods to compute attribution maps such as gradient-based methods. Gradient-based saliency map methods [6, 1] is the most common example of such a category, where the attribution map is generated by computing the gradient of the target class score w.r.t. the input image. The gradient quantifies how much change in the input pixel values corresponds to the change in target class scores.

While Class Activation Maps (CAM) are good at highlighting discriminative regions (DR) of an image (i.e., regions that contribute significantly to the classifier's decision), CAMs are known to disregard regions of the target object class that do not contribute to the classifier's prediction, termed non-discriminative regions (NDR). This is because Class Activation Maps (CAMs) are essentially **activation maps** generated by the last convolutional neural network (CNN) layer, which are integrated with the weights of the final fully-connected layer. It has been shown that the final layer feature maps only contain information that is relevant to classification, a phenomenon called *information bottleneck* [3]. The RIB [3] demonstrates that an information bottleneck occurs in later layers as only the task-relevant information is passed to the output. As a result, CAMs that are computed at the last layer have sparse coverage of the target object. Thus, CAMs are biased towards mostly finding DR while missing the NDR of the target object, which is equally important for the purpose of segmentation. A number of WS3 solutions thus require further processing of the

CAM outputs to recover NDR for high segmentation accuracy [4, 2].
In contrast to activation maps, **attribution maps** provide an alternative approach for assigning a score to every pixel based on its contribution to the final neural network prediction. The most commonly used attribution map is the gradient-based Saliency Maps [6]. The basic idea of saliency is to calculate the gradient of the target class score with respect to every pixel in the input image. Attribution maps are fundamentally distinct from the activation maps obtained from the last layer of CNN models. However, despite the frequent use of attribution maps for interpretability purposes, their utility in WS3 has not yet been fully explored.

References

[1] David Baehrens et al. "How to explain individual classification decisions". In: *The Journal of Machine Learning Research* 11 (2010), pp. 1803–1831.

[2] Qibin Hou et al. "Self-erasing network for integral object attention". In: *Advances in Neural Information Processing Systems* 31 (2018).

[3] Jungbeom Lee et al. "Reducing information bottleneck for weakly supervised semantic segmentation". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 27408–27421.

[4] Kunpeng Li et al. "Tell me where to look: Guided attention inference network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 9215–9223.

[5] Ramprasaath R Selvaraju et al. "Grad-CAM: Why did you say that?" In: *arXiv preprint arXiv:1611.07450* (2016).

[6] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034* (2013).

[7] Bolei Zhou et al. "Learning deep features for discriminative localization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2921–2929.