Visual Explanations from Deep Networks via Gradient-based Localization

Summary and review by - Purvanshi Mehta

September 23, 2019

1 Summary

The paper proposes an architecture independent high resolution localization technique for highlighting the concepts used in the prediction of a particular class. Grad - CAM is an extension to the Class Activation Mapping (CAM) method for localization and Guided backpropagation.

The CAM approach modifies the image classification CNN architechture by replacing the fully connected layers with convolution layers and global pooling which helps in achieving class specific feature maps. The approach balances the accuracy and interpretability trade off. Grad-CAM uses gradient information flowing into the last layer of CNN. To obtain a class specific localization gradient, the partial derivative of the class score is computed with respect to the feature maps of the convolutional layer. These gradients are then global average pooled to obtain the neuron importance weights. The heat map is obtained by the weighted combination of these weights with the feature maps followed by a ReLU. One of the strong points of the paper is that the output with respect to which we take the gradients can be *any differential function* and not compulsorily a class score. This makes the approach generalizable to tasks such as image captioning.

To increase the resolution of the localizations produced by Grad-CAM, visualizations were combined with the guided backpropagation output via a pointwise product.

Experimental results are formulated in various domains - weakly supervised localization, pointing and faithfulness to original model. The method is shown to be ineffective towards adversarial inputs as it still highlights the original categories despite the network being completely uncertain of their presence. Biases are detected using doctor vs nurse classification task by fine tuning a VGG16 - the network was instead classifying male as doctors and female as nurses just by looking at the facial features. The method is also formulated as a concept identifying algorithm. Results are shown on image captioning, visual question answering.

2 Review

For making the heatmaps more interpretable to a human eye, an average of N heatmaps could be taken. This would also avoid a randomly activated region

(random to a human eye) which could be the cause of the network actually using that area for prediction or just the effect of gradients. To mark the activated region more precisely, constraints could be put on the nearby pixels like avoiding large differences in partial derivatives.

Another application of Grad-CAM could be - reasoning in question answering. Giving the heatmaps as input so that the NLG model knows why certain prediction has been made and then it generates an explanation looking to the activated regions.