Workshop Week 2: Feature attribution in vision

Payanshi Jain: 14230143

June 16, 2023

Lab 1: Integrated Gradients

Integrated Gradients are commonly associated with superior performance compared to alternative saliency methods due to their comprehensive approach. By considering the entire path from a baseline to the input and capturing gradient information at each step, Integrated Gradients provide a holistic understanding of the input-output relationship. The utilization of the baseline as a representation of feature absence and the comparison of gradients between the baseline and the input contributes to the effective highlighting of feature contributions. The role of the baseline is crucial in saliency methods, as an improper selection can detrimentally affect the results. Black or random baselines, for instance, are inadequate in capturing genuine feature absence or natural variations, leading to inaccurate saliency maps and misleading attributions. Conversely, alternative baselines such as Maximum distance, blurred, uniform, and Gaussian baselines offer enhanced efficiency by establishing clearer absence-presence differentiations, capturing overall context, providing a neutral reference point, and accounting for data noise or uncertainties. These alternative baselines produce more meaningful reference points, thereby improving the accuracy of saliency results. For example, the Maximum distance baseline establishes the baseline as the farthest point in the input space, guaranteeing a conspicuous differentiation between the absence and presence of features. The blurred baseline, on the other hand, entails blurring the input image, facilitating the capture of the overall structure and context rather than fixating on intricate particulars. In contrast, the uniform baseline assigns a uniform value to all pixels, serving as an impartial reference point that circumvents partiality towards specific features. Lastly, the Gaussian baseline employs a Gaussian noise pattern to the input, accommodating the inherent noise or uncertainties present in the data.

Lab 2: LIME

The accuracy of the method on the test set is 36.36%. By applying the LIME algorithm and examining the classifier's performance on the images from the LIME_test_files folder, it becomes apparent that the model $f()$ faces challenges in accurately categorizing husky and wolf images. The primary concerns with this classifier may originate from biases present in the training data, resulting in flawed generalizations and erroneous classifications. The classifier's subpar accuracy highlights the necessity for enhancing its ability to discern the distinctive features that differentiate huskies from wolves.

The distinction between the distribution of real samples and the perturbations can be demonstrated by examining the classification of huskies and wolves. When considering real samples, the classifier f demonstrates the ability to accurately discern significant features such as the distinctive fur patterns of huskies and the characteristic attributes of wolves. However, in the context of adversarial perturbations created to deceive post hoc explanation techniques, the focus shifts towards manipulating less crucial regions of the input. These perturbations may introduce subtle modifications to irrelevant areas or the background of the image while preserving the essential features. As a consequence, post hoc explanation methods like LIME can assign importance to these manipulated regions, leading to a misguided understanding of feature contributions and potentially obscuring the biases inherent in f. By highlighting this contrast between real samples and perturbations, we underscore the objective of adversarial attacks to undermine the dependability of post hoc explanations and conceal the underlying biases of the classifier.