SMOOTHLRP: SMOOTHING EXPLANATIONS OF NEURAL NETWORK DECISIONS BY AVERAGING OVER STOCHASTIC INPUT VARIATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

With the excessive use of neural networks in safety critical domains the need for understandable explanations of their predictions is rising. Several methods were developed which identify the most relevant inputs, such as sensitivity analysis and most prominently layerwise relevance propagation (LRP). It has been shown that the noise in the explanations from the sensitivity analysis can be heavily reduced by averaging over noisy versions of the input image, a method referred to as SmoothGrad. We investigate the application of the same principle to LRP and find that it smooths the resulting relevance function leading to improved explanations for state-of-the-art LRP rules. The method, that we refer to as SmoothLRP, even produces good explanations on poorly trained neural networks, where former methods show unsatisfactory results. Interestingly, we observed, that SmoothLRP can also be applied to the identification of adversarial examples.

8 pages of text in total - unlimited citations and many pages in appendices - use citet and citep