Explainable AI Evaluation Metrics Overview

April 19, 2021

0.1 Metric: MoRF

Most **R**elevant **F**irst (abbreviated as *MoRF*), or simply *Deletion* is an intuitive evaluation metric which is, with slight variations, proposed in multiple papers [3, 4].

0.2 Metric: ROAR

Hooker *et al.* propose the metric **R**em**o**ve **a**nd **R**etrain (abbreviated as *ROAR*) [2].

0.3 Metric: Cropping

Dabkowsky and Gal have proposed a metric based on another intuition compared to the others we have seen so far [1].

0.4 Metric: Max-Sensitivity

The *Max-Sensitivity* metric is proposed by Yeh *et al.*, originally for evaluating saliency explanations for black-box models [5].

0.5 Metric: Infidelity

The last metric is the *Infidelity* metric, which is proposed by in the same paper as the *Max-Sensitivity* metric [5].

References

[1] Piotr Dabkowski and Yarin Gal. “Real Time Image Saliency for Black Box Classifiers”. In: (May 2017).

[2] Sara Hooker et al. “A Benchmark for Interpretability Methods in Deep Neural Networks”. In: Jan. 2020.

[3] Vitali Petsiuk, Abir Das, and Kate Saenko. “RISE: Randomized Input Sampling for Explanation of Black-box Models”. In: *BMVC*. 2018.

[4] Wojciech Samek et al. “Evaluating the visualization of what a deep neural network has learned”. In: *IEEE transactions on neural networks and learning systems* 28.11 (2016), pp. 2660–2673.

[5] Chih-Kuan Yeh et al. “On the (In)fidelity and Sensitivity of Explanations”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019.