Summary of “Can Post-hoc Explanations Effectively Detect Out-of-Distribution Samples?”

Sumedha Chugh
PhD21123

In order to assess risk factors and the reliability of AI systems, the European Commission in 2018 introduced the General Data Protection Regulation law, which gives individuals the right to get logical explanations of the decision made by the AI system. This research work checks if the explanations generated for individual samples (local) after the model has been deployed (Post-hoc), can also identify if an instance being fed to it has not been seen during training i.e. it is Out of Distribution (OOD). A method LEO (Local Explainations based OOD detector) is proposed. It samples N data points of all labels from training data (in-distribution) and specifies a heatmap for each in-distribution label, by assigning a distance to every pair of explanations and clustering similar explanations. For the sample to be detected, it generates a local explanation and calculates f_{ood} by computing the distance of explanation from the cluster corresponding to the given output. If f_{ood} > threshold, it is considered OOD, otherwise in-distribution.

Experiments using LEO and other baselines are performed on datasets Fashion MNIST, MNIST, CIFAR10 & SVHN to evaluate AUROC (Area Under the Receiver Operation Characteristic curve), AUPR (Area Under the Precision-Recall curve) & FPR95 (False Positive Rate at 95% True Positive Rate) matrices. MNIST contains images of digits from 0 to 9, Fashion MNIST consists of clothes images across ten classes, CIFAR10 has images of 10 classes of animals & automobiles, and the SVHN dataset is made of images of house plates with numbers. Experiments are divided into three case studies Case A, where MNIST is in distribution and FMNIST is OOD; Case B with FMNIST being in distribution & MNIST being OOD; case C where CIFAR10-C is in distribution & SVHN is OOD. Since data is not augmented, rotated images are also considered OOD.

In case A and case B LEO has high FPR95 scores implying sensitivity to noise for the training dataset and has AUPR & AUROC scores comparable to baselines. The proposed model performs very poorly for case C. This exists because of the absence of positional bias in case C. On comparing LEO to already existing OOD methods, similar results can be seen i.e. for the first two cases its performance is comparable to State of the Art methods but performs worse than a random classifier for case C. The results show that LEO’s performance is affected by bias in the data and has a lower detection score for methods specifically designed for detecting OOD data. But it can be used to evaluate the trustworthiness of OOD samples in real-world applications by detecting bias.

Reference:

1. A. Martinez-Seras, J. D. Ser and P. Garcia-Bringas, "Can Post-hoc Explanations Effectively Detect Out-of-Distribution Samples?," 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Padua, Italy, 2022, pp. 1-9, doi: 10.1109/FUZZ-IEEE55066.2022.9882726.