Advanced Machine Learning, 2023

Home Assignment 3 - Explainable AI

Abstract

In this home assignment, the task was to implement LIME (Local Interpretable Model-Agnostic Explanations) to explain classifications generated by the ResNet18 model. LIME is a popular technique for explaining the predictions of black-box machine learning models by creating interpretable and locally faithful explanations.

The assignment involved implementing LIME to generate explanations for the predictions made by the ResNet18 model on various image segments. The ResNet18 model, a famous deep-learning architecture, was trained on a dataset containing these segments. LIME works by perturbing the input images and observing how these perturbations affect the model's predictions. It then constructs simplified, interpretable models around these perturbed instances to explain the local behavior of the predictions.

Experiment Flow

We created interpretable representations of the images by splitting them into `k` super-pixels (AKA segments). The process worked as follows:

1. Given an image and an estimator (Resnet), we generated the top 3 classes predicted by the model.
2. We segmented the given image, whose predictions we would like to explain, and generated ~ 50 segments per sample.
3. We generated 100 new samples by perturbing the original image and randomly removing 20% of the original segments.
4. We assigned each perturbation a label according to the model prediction on that image.
5. We trained a Lasso classifier using the interpretable representation of the generated perturbations (X) and their respective labels (y). We weighted each sample by its inverse L2 distance from the original image (the similar the generated sample is to the original image, the higher weight it receives).
6. We then chose the top k coefficients (by magnitude) of the trained Lasso model and plotted them (these are the top k most essential features for predicting the image class by the Resnet model)

Findings

We saw that k=7 produced the most informative features for the images we used in our experiment, as seen when plotted. Segmenting looks like a good representation of the image's semantic properties, but we would also like to continue research for an even more interpretable representation.