HW3 - Advanced Machine Learning Report

In this assignment, we implemented and utilized the LIME (Local Interpretable Model-agnostic Explanations) algorithm to generate interpretable explanations for image classification models. The LIME algorithm can provide insights into the behavior of image classification models and identifies important features (super-pixels) for each class prediction.

Data Generation and Visualization

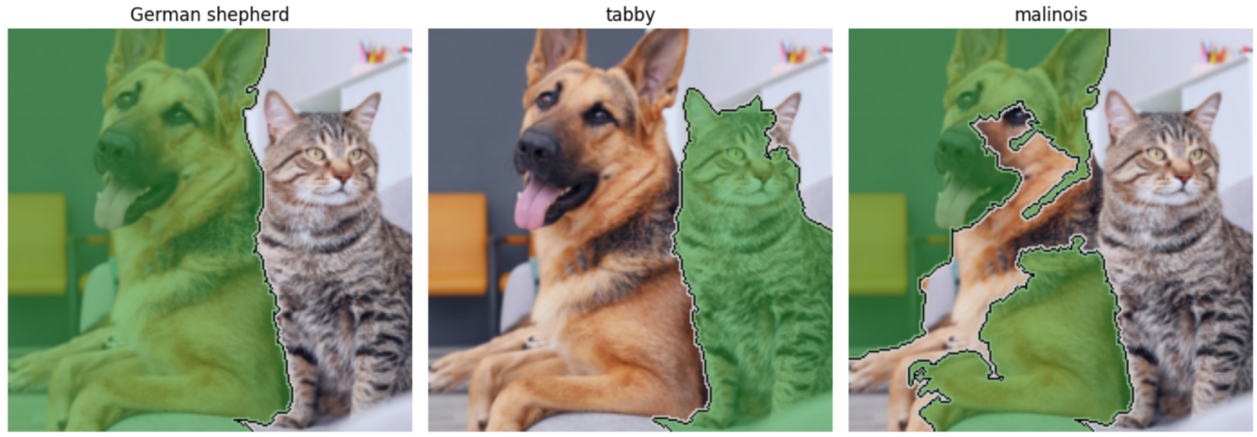
A pre-trained RESNET50 classification model was chosen as the base model for prediction. Four different images representing a dog, dog and a cat, pandas, and a zebroid (hybridization of a horse and a zebra) were used for each image, the top three labels were predicted using the RESNET50 model.

Methodology

1. Image Interpretation and Binary Vector Representation:
    a. The selected images were interpreted using the scikit-image package, which split the pictures into super-pixels.
    b. Interpretable instances were represented as binary vectors, where the vector entries indicated the inclusion or exclusion of super-pixels.
2. Local Dataset Generation and Perturbation:
    a. For each class, a local dataset was generated by applying random perturbations to the interpretable instances.
    b. Perturbations were created by uniformly choosing which parts (super-pixels) to include in the modified instances.
3. Local Surrogate Model Fitting and Explanations:
    a. A local surrogate Lasso Model, with locally weighted loss and L2 regularization, was fitted using the generated dataset.
    b. Feature selection was performed using K-Lasso to identify the set of important features (super-pixels) for each prediction.

The LIME algorithm successfully generated interpretable instances and explanations for the image classification model using the selected four images although some images were still not perfectly interpreted (e.g. the 2nd top class of the pandas).

The explanations provided insights into the model's behavior and identified important super-pixels for each class prediction. Boundary plots were utilized to visualize the separation of classes and highlight the significant super-pixels in the image's top classes in an elegant visualization technique (as demonstrated below).



German shepherd

tabby

malinois

In conclusion, the LIME algorithm demonstrated effectiveness in generating interpretable explanations for image classification models. By applying perturbations and fitting local surrogate models, the algorithm could provide insights into the model's predictions and highlight important features for each class.

The LIME algorithm proved valuable in providing insights and understanding the behavior of the model, making it a useful tool also for interpreting and explaining image classification predictions.