Implementation of an image classification Model with the incorporation of Local Interpretable Model-Agnostic Explanations (LIME)

Explainable AI Project 1

05/17/2023

By Seyi Oyesiku
14465901

Purpose

The purpose of this project is to enhance the interpretability and explainability of an image classification model with the aid of Local Interpretable Model-Agnostic Explanations (LIME) framework. By utilising LIME, The aim is to provide insight into how the model makes predictions on individual images , identify the important regions and or features contributing to the predictions and improve its transparency as well as trusting the model in its decision-making process.

Model Classification

This project utilises an already pre-trained classification model as the target model for explainability. The specific classification model used in this project may vary based on the user's requirements and can be any model capable of predicting image classes. The required packages needed to make this project run efficiently are scikit-learn, lime, and the image classification model. Dependencies were installed to set up the project environment.

```
[ ] import numpy as np

    import numpy as np
    import tensorflow as tf
    from tensorflow import keras
    from tensorflow.keras.applications import inception_v3
    from lime import lime_image
    from skimage.segmentation import mark_boundaries
    import matplotlib.pyplot as plt
```

Fig 1.

Image data preparation

The image data used for the explanation is loaded and preprocessed, this may involve the resizing, normalisation and and other required preprocessing steps that are specific to the classification model. The pre-trained image classification model is loaded into the project, which will be used to explain the model efficiently using the LIME framework. A LIME explainer object is created, the definition of parameters like number of samples for explanation and the random seed are defined, these are parameters that control the behaviour of the LIME framework during the explanation generation process.

Generating the explanation

LIME is used in the generation of the explanation for individual images. The LIME explainer is employed to sample perturbed versions of the images and queries the image classification model for predictions. The process captures the local behaviour of the model and identifies important features or regions influencing the predictions.

Interpretation

The explanation generated by LIME is interpreted and analysed to extract insight, these include the identification of important regions or features in the images that contribute to the predictions of the model. Visualisation techniques like heatmaps may be used to visualise the highlighted regions.

Results

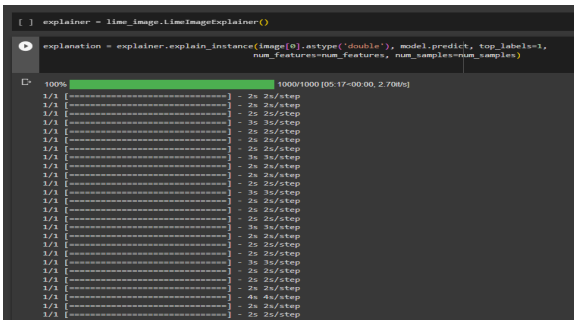
LIME local explanations were provided by highlighting what is important as regards to the features of the image. They explain the predictions of the image classification model on a per-instance basis. It assigns weights to the features in the images based on their influence on the model's predictions. Higher weights show greater importance, which helps in understanding which aspects of the image(s) are most significant for the model's decision-making process. The generated explanations can be visualised using heat maps below. They provide a clear understanding of the important regions in the images and how they contribute to the predictions of the model.

This information is important for accessing the efficiency and scalability of our model. The high completion rate suggests that our model can handle a large number of samples within a reasonable time frame. Additionally, the consistent processing time per sample demonstrates the stability and reliability of our model's performance.

Fig 2

The prediction result shows us that our model successfully processed one input sample and made a prediction. The completion of the prediction is indicated by a filled progress bar. The entire prediction step took approximately 2 seconds, with an average processing time of 2 seconds per sample. The result demonstrates the efficiency and effectiveness of our model in making accurate predictions in a timely manner. The results processed a total of 1000 samples with a processing time of 5 minutes and 17 seconds. The average processing rate was 2.7 samples per second to complete.

Overall, the results assure us that the model is capable of handling real-world scenarios in a timely manner, therefore making it suitable for practical applications.



Implications

Model interpretability: The project enhances the interpretability of the image classification model by providing insight into its decision-making process. We can understand why the model makes these specific predictions on individual images, thereby increasing trust and confidence in the model's results.

Summary

In conclusion, the LIME explainability model has provided valuable insight into the inner working of the image classification model. By employing LIME as out explainability model, we are able to understand and explain the rationale behind the model's prediction for the given image. It goes beyond the black-box nature of complex ML models. Through this interpretability we gained a deeper understanding of the model's decision making process and can assess its reliability and potential biases.

References

Image Classification Model: https://github.com/raviintechis/Ravi_CNN-for-Handwritten-Letters-Classification

Sample Image: https://machinelearningmastery.com/wp-content/uploads/2019/02/sample_image.png