Galaxy morphology classification and captioning

Overview

The goal of the project was to build reproducible machine learning models using the Galaxy Zoo dataset, to 1) predict the label of a galaxy from these categories: spiral, elliptical, edge-on, merger, ambiguous; and 2) generate a caption for an image that would describe morphological features solely off that image.

Data

The dataset (Galaxy Zoo 2) was downloaded from Kaggle. A sample of the images was saved to data/labels and the rest was kept locally. The initial encoding of the file was obscure, and I had to resort to AI to transform it into a regular csv.

Classifying

Preprocessing

Details on preprocessing can be found in the first and second notebooks. In a nutshell, I deleted all rows with missing values (a really small amount). Then I built a pipeline to split the data into training, validation and test sets which can be used by the model directly.

Training

The training loop consisted of a convoluted neural network (CNN) with a pretrained ResNet50 base. I used tensorflow's functional syntax. The third notebook explains it in more detail. The model trained for 50 epochs on the full training data.

Evaluation

The model's performance on the test set was **0.66** accuracy and **0.62** weighted F1 accuracy. An inference module and a grad-cam visualizer were added to explain the guesses of the model. The confusion matrix reveals that the most difficult classes to discern were merger and ambiguous. Due to class imbalance, the model poorly performed on underrepresented classes (merger, ambiguous) and had higher accuracy on overrepresented ones (spiral, elliptical). The results might be improved by fixing class imbalance.

Captioning

The captioning model was built using a Long Short Term Memory (LSTM) architecture with teacher forcing. The captions the model trained on were originally inferred from the voted columns in the dataset by templating. The results were evaluated on the test set using BLEU score: BLEU-1: **0.47**, BLEU-2: **0.38**, BLEU-3: **0.28**, BLEU-4: **0.19**. The captions were fine grammatically, but fundamentally lacked variety. Additionally, there were frequent mistakes occurring due to the inability to understand fine features (the model always went for the 'safe' option), the main issue being the greedy search algorithm in inference. The result was the same 2-3 phrases used 80% of the time.

Conclusion

The project may not have yielded perfect results, but it was my first experience in building something of this scale. There is definitely room for improvement, but a strong baseline is set.