Name : Muhammad Eri Mushthofa

PSU email: mmm7769@psu.edu

CLASSIFICATION TASK ON MOVIE DATASET

Dataset Source: https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset

1. Using RNN (Recurrent Neural Network)

This project involves the construction of a Recurrent Neural Network (RNN) using Keras, aimed at classifying text data. The data, sourced from 'train_data.txt,' is loaded into pandas DataFrame with columns like "ID," "Title," "Genre," and "Description." The pre-processing stage involves normalizing the text in the 'Description' column by converting it to lowercase and removing non-alphanumeric characters, a common practice in natural language processing. The 'Genre' column, the target variable, is transformed into numerical values using LabelEncoder.

To prepare the data for the RNN, a tokenizer is implemented, restricting its focus to the top 5000 words and converting the descriptions into sequences of integers. These sequences are then padded to a uniform length of 150 for consistency in training. The dataset is partially divided into training, validation, and test sets, with proportions of 60%, 20%, and 20%.

The RNN model architecture includes an Embedding layer, an LSTM (Long Short-Term Memory) layer, and a Dense layer. The LSTM and Dense layers' parameters, such as the number of units and the learning rate, are not preset but are subject to optimization through hyperparameter tuning using kerastuner. The tuning process, limited to 10 trials, seeks the best combination of LSTM units and learning rates based on validation accuracy.

The training employs early stopping, ceasing if the validation accuracy fails to improve for two consecutive epochs within a maximum of 5 epochs, thereby preventing overfitting. The optimal model is retrained and evaluated on the test set after hyperparameter tuning. The evaluation includes a classification report and an accuracy score, offering insights into the model's performance across different genres.

2. Using CNN (Convolutional Neural Network)

The project begins with loading data from 'train_data.txt' into a pandas DataFrame, structured with columns "ID," "Title," "Genre," and "Description." In preprocessing, the text in the 'Description' column undergoes standard normalization - conversion to lowercase and removal of non-alphanumeric characters. Meanwhile, the 'Genre' column is numerically encoded using LabelEncoder.

The next step involves tokenization, where a tokenizer is developed to convert the text descriptions into integer sequences, focusing on a vocabulary of the top 5000 words. These sequences are then padded to a standardized length of 150. The dataset is strategically divided into training, validation, and test sets, with a distribution of 60%, 20%, and 20%, respectively.

A CNN architecture is defined through a function for model building, incorporating hyperparameters like embedding dimension, filters, kernel size, and learning rate. The model's structure includes an Embedding layer, a Conv1D layer, a GlobalMaxPooling1D layer, and a Dense output layer. The hyperparameter tuning employs a Keras Tuner RandomSearch tuner, aiming for the highest validation accuracy. It experiments with up to 10 different configurations.

During the tuning phase, early stopping is utilized, halting training if there is no improvement in validation accuracy for two epochs. Following this, the best-performing model is identified, trained on the training data, and further validated over 5 epochs.

In the final stage, this model is assessed on the test set, with the primary metric being test accuracy. This comprehensive approach, involving careful preprocessing, hyperparameter tuning, and systematic evaluation, aims to optimize CNN's performance in classifying text data.

3. Conclusion

During my research, I encountered several challenges, particularly regarding the running time of the experiments. I implemented a cap of 10 maximum trials for each test run to address this issue. Despite this limitation, each trial required at least an hour to complete. This extensive time requirement was a significant factor in the project's overall progress. Regarding results, the Recurrent Neural Network (RNN) model achieved a peak accuracy of 53%. In comparison, the Convolutional Neural Network (CNN) model showed a slightly better performance, reaching a top accuracy of 54%. These outcomes highlight the nuanced differences in the effectiveness of these models for the tasks they were assigned.