Topic: Multi-Label Image Classification on Movies Poster using CNN

Project Objective:

The objective of our project is to predict the genre of a movie using its poster image.

A movie usually belongs to multiple genre that means more than one genre and not necessarily belong to one category like drama or thriller. Most of the times a movie is a combination of more than one genre. Hence, we are doing multi-label image classification here. The CNN model we will make, will be classifying a poster among 25 different genres based on its processing of the image and classify it to the appropriate genre of that movie. In the end, our CNN model should predict the genre of a movie judging from its poster.

Information About the Dataset:

Dataset Link: https://www.kaggle.com/raman77768/movie-classifier#

This dataset was collected from the IMDB website. One poster image was collected from one (mostly) Hollywood movies released from 1980 to 2015. Each poster image is associated with a movie as well as some metadata like ID, genres. The ID of each image is set as its file name.
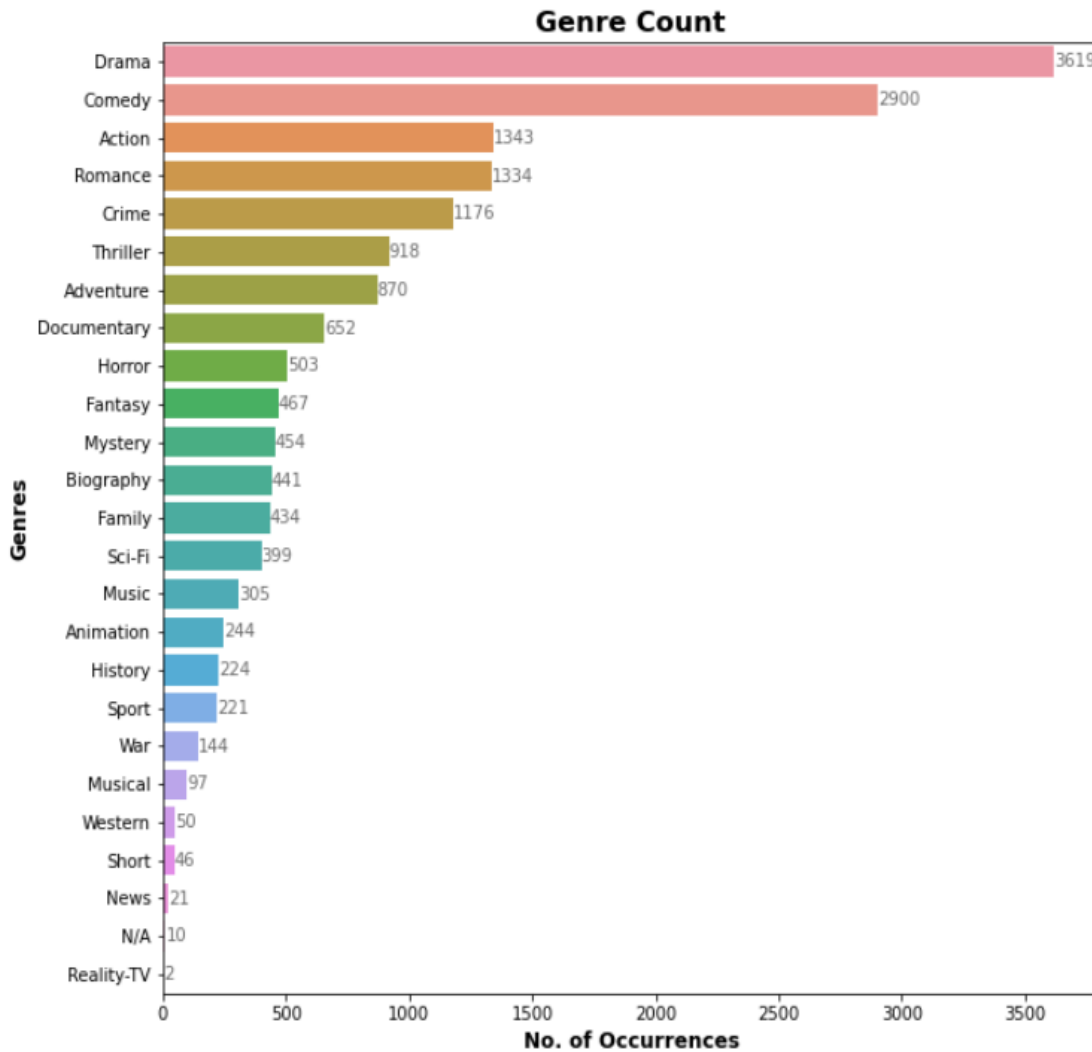
The dataset contains 7254 rows and 27 columns along with the movie's posters.

Acknowledgement of BDA Course Curriculum:

Subjects Like Machine Learning and Predictive & Prescriptive Analysis which are part of our Course curriculum has helped us greatly in understanding advanced machine learning and deep learning tools and algorithms required in building a model. We also got the opportunity to work on Project related to machine learning and neural networks which gave us hands-on experience and taught us how to approach a problem and select the right model. We would like to express our heartfelt appreciation to the entire PGDM-BDA fraternity for their continuing support and insightful feedback. The course had greatly enabled us to explore new domains as well use technical skills to build insightful machine learning models.

Details of the Work and Insights:

Visualization of the Counts of Genres Occurrence

Genre Count

Genre	No. of Occurrences
Drama	3619
Comedy	2900
Action	1343
Romance	1334
Crime	1176
Thriller	918
Adventure	870
Documentary	652
Horror	503
Fantasy	467
Mystery	454
Biography	441
Family	434
Sci-Fi	399
Music	305
Animation	244
History	224
Sport	221
War	144
Musical	97
Western	50
Short	46
News	21
N/A	10
Reality-TV	2

We can see that the genre distributions are uneven, meaning the dataset is biased.

Visualization of Model Accuracy and Model Loss
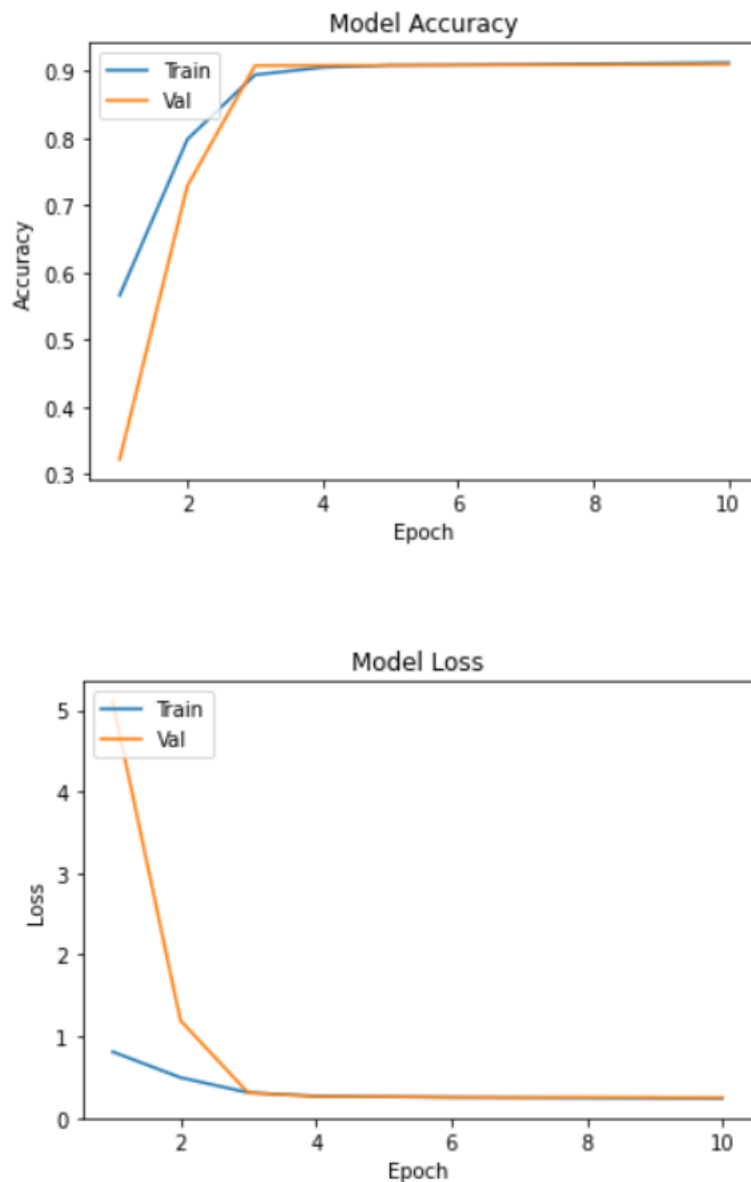
The validation accuracy (around 90%) is almost same with the training accuracy and the same with the validation loss. Hence, the model is working fine but may have overfitting problem due to biasedness.

If a certain genre is repeating in most of the training images, the model may overfit on that genre. For every new image, the model may predict the same genre. To overcome the overfitting problem, the dataset should have an equal distribution of genre categories.

Testing of the Model

Testing the model on some post 2015 movie posters to check how it's performing.

The predicted top 3 genres for the following movies below:

1. Black Panther - Adventure, Comedy, Action
2. Coolie No 1 - Drama, Comedy, Romance
3. Joker - Drama, Thriller, Crime
4. La La Land - Drama, Comedy, Romance

The predictions are not perfect but almost accurate with the actual genre of the movies.
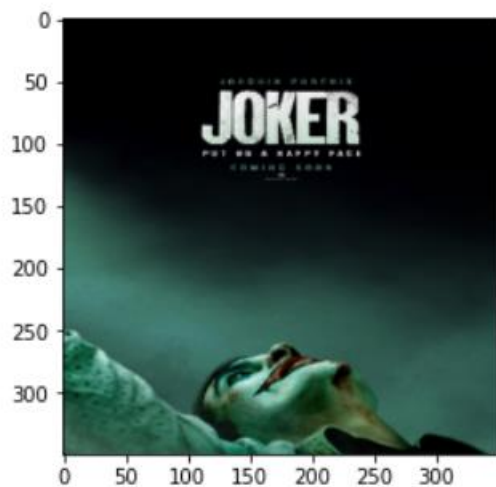

Adventure
Comedy
Action
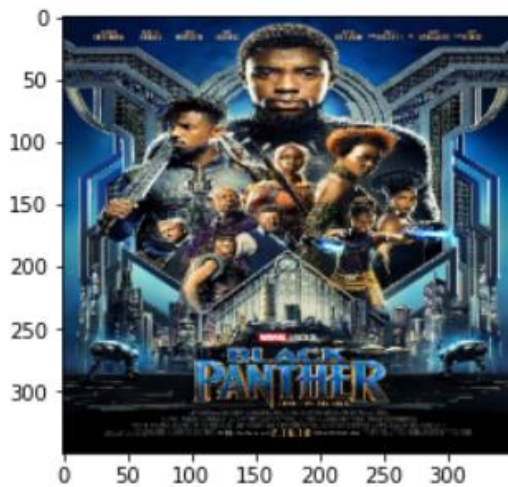
Drama
Comedy
Romance

Drama
Thriller
Crime

Drama
Comedy
Romance