Can you Judge a Movie by Its Poster?

Matthew Afsahi, Skylar Furey, Jimmy Kruse, Sivani Pillutla

The task our project will perform is to project the revenue of a movie by the image of its poster. The goal of the project is to determine if you can determine how monetarily successful a movie will be by looking at its poster. This model could be used by movie producers to determine if their poster mockups will draw crowds to their product or not. Similarly, movie theaters can determine which posters to highlight as they are the most likely to draw crowds into the theater. We will likely separate the movies into their different genres as each genre has different key characteristics, for example Horror movies tend to have darker posters. We will use image regression to linearly predict revenue with training and test datasets. One attempt at a similar task, by Jingles[1], created a model that would correctly predict if a movie would potentially be profitable approximately 68% of the time. We will be using a different data set but will aim to improve on their results.

The base of our dataset will come from Kaggle[2], which contains over 700,000 movies from the past 100 plus years, consisting of different genres and having a variety of budgets. We will then take the names of these movies and scrape OMDb API[3] to find the movie posters and a couple other additional fields for each of the movies in our dataset. To ensure data consistency, we will preprocess the posters by standardizing the image resolution and normalizing color channels across the dataset before feeding them into the model.

The metric we will use for our dataset is accuracy using the following calculation:

abs((predicted_revenue – actual_revenue) / actual_revenue)

This metric will allow us to fine tune our model such that it most accurately predicts the revenue of a given movie based on its poster. In addition, we will evaluate the model using metrics like mean squared error (MSE) to better understand its performance on outlier cases.

In addition to our main task we will classify the movies into their genres using softmax. We will use accuracy as the metric to determine the performance of our classifier. We plan to use the same CNN to provide both outputs.

Challenges with this task will be determining which features of a movie poster most impact the value of the movie it is designed for. This challenge will be dealt with by using standards developed for other projects that can help determine key parts of an image that leads to a better movie which sells more.

We plan to implement our project in Python using TensorFlow to create a CNN with optimized layers. We will also likely use packages like Pillow to preprocess the image data. Finally, we will use packages like scikit-image to detect features in the image that may help the CNN.

Links:

1. https://dev.to/jinglescode/predict-movie-earnings-with-posters-k72
2. https://www.kaggle.com/datasets/akshaypawar7/millions-of-movies
3. https://www.omdbapi.com/