Movie Genre Classification of Movie Posters

By Pratik Patel & Harsh Patel

**Introduction:**

We all have heard the idiom *“Don’t judge a book by its cover”*, typically in a metaphorical sense, however when taken literally while we may not be able to learn everything, a person can learn alot just from a cover. Just by looking at the cover of book people can learn what the book is mostly about. In the world of movies, movie posters are an important part of aesthetically appealing to people as well as allowing people to understand a few things about what the movie is about such as the genre of the film. While a human can easily distinguish a movie poster between a horror film from a comedy film, we want a computer to complete the task of classifying movie posters by genres. We first drew inspiration of this project due to how I(Pratik) am an advent moviegoer and love watching movies to the point where I keep track of the movies I have seen on an account on IMDB. With movies being the central focus of looking for a topic we found a couple of research papers that did the same project but only discussed their findings, errors and overall build. We wanted to make our own version with the general outline of their procedure while expanding on some things they did not do like how they only used black and white movie posters to expanding to color ones.

**Proposed Work Section:**

The idea we had was to first create a comprehensive collection greater than what was done by other individuals so about 10,000 movies with it’s metadata, the movie poster, and the genre of the movie. This would be done using the both TMDb and IMDb databases which both are movie databases. Specifically we would use TMDb, as they have an API to grab all details about movies and the actual image of movie poster. IMDb will be used to extract genres for a movie poster we grabbed if TMDb does not provide for some movies done through the module IMDbPY for Python. We will then take a look at the genres and do some pre-processing depending on how the data looks as their may be multiple subgenres that fall into one general genre such as movies being classified as melodrama or legal drama but to make training better it would make sense to classify them as drama, we could also have to give each movie poster a distinct genre if multiple genres are given for one movie. Some other preprocessing will be done if there are too many genres thus some movies will be removed from actual data if the genre is not really popular in dataset. We will convert the genres so that each genre will be a unique numerical value , for example Adventure would be 1, Drama would be 2 and so on to make it viable to train. We will most likely create a CSV file which will hold the all the movie metadata with a path to the picture and then include the genres (numerical values) as we believe that is how the TMDb API will save them. We then plan on creating a convolutional neural network similar to our lab assignments, where the inputs(X values) are the movie posters and outputs(Y value) will be a numerical value representing each genre. Like always the data will be split into training and test data before and we also have to pre-process the images so that the movie posters have all same size and the colors are split into correct RGB layers. By the looks of it we may end up using AWS to speed up process of actually training neural network if the image processing takes a long time.

**Timeline:**

By the first progress report we plan on gathering all the data and pre-process the genre portion in which changing the subgenres to one overall genre, removing uncommon genres, and converting genres to numerical values. For the second progress report we plan to have pre-processed the images and run a simple convolutional neural network to train data to acquire some results we can see so that we can began fine-tuning. For the final report, we would have most likely have to run our project on a AWS server to speed up process for a deeper neural network.

Data Collection

Pre-Processing Genre

April 17-25

April 25- May 1

Pre-Processing Images

Simple CNN

May 1- May 7

Fine Tuned CNN (AWS?)

Final Report

**References:**

<https://www.cs.ccu.edu.tw/~wtchu/papers/2017MUSA-chu.pdf>

<https://pdfs.semanticscholar.org/a47f/d31a18b3b1b5d02815a9220ab806ceb58edd.pdf>