Predicting movie genres from movie posters

Nirman Dave
Hampshire College

npd15@hampshire.edu

Abstract

Movies are an essential part of our lives and today when we come across a movie poster we can quickly grasp elements like color, facial expression, objects and much more to get an accurate intuitive understanding of what kind of a movie it would be. Inspired by this human behavior, in this work, I propose to construct a 7-layer convolutional neural network that can describe visual appearance of a movie poster to classify it into one or more genres. Since a movie can be classified into multiple genres, this is a multi-label classification problem. To facilitate this, I have collected a large-scale movie poster dataset which documents over 40,000 movies from 1997-2017 across 28 different genres. Based on this dataset, the network is fine-tuned on two hyper-parameters: learning rate and input image size. In the evaluation, I show that the proposed method yields a good performance on versatile movie posters and other various images.

1. Introduction

Humans have a very unique ability to understand the world around them in an instance. When it comes to movies, we can look at a poster, analyze its colors, textures, faces, expressions, objects, etc. and quickly determine what the movie could be about. For example, we can quickly tell Notebook *(Fig 1, a)* is a romantic movie and Interstellar *(Fig 1, b)* is a Sci-Fi movie. But for movie posters like Up *(Fig 1, c)* and Lés Mirablés *(Fig 1, d)* we need much more context. Looking at the Lés Mirablés poster, stand alone, we cannot say if it is a war movie, a documentary or drama? Similarly, the poster in for Up just shows an animated boy but tells us nothing about what the movie could be, yet

Figure 1: Movie posters. (a) The Notebook – Romance; (b) Interstellar – Sci-Fi; (c) Up – Comedy, Animation; (d) Lés Mirablés – Drama, History, Musical.

somehow, we have a sense of what to expect – something comic and delighting. This is truly a fascinating process. The idea with this project is to capture some of this process through construction of a Convolutional Neural Network (CNN) that trains on thousands of different movie posters with multiple genres as their labels and eventually predict the genre(s) of an unseen movie poster.

There are various ways to approach this – 1. using a pre-trained net, 2. using an established architecture like VGG-16 or 3. building from scratch. This project takes the third approach to construct a 7-layer CNN for this multi-label classification problem. The general idea is to mimic a VGG net with customized kernel size and output layer. This is an interesting approach because VGG-16 has proven to be accurate in extracting small details from an image on the ImageNet dataset. Therefore, the idea is to see if this customized, mini VGG styled CNN can extract granular features from movie posters and eventually predict its genre.

This study is important because it can have major implications in various areas – building a movie recommender system, discovering poster design cues for amateur designers and if extended, this work could be a basis for predicting box office sales.

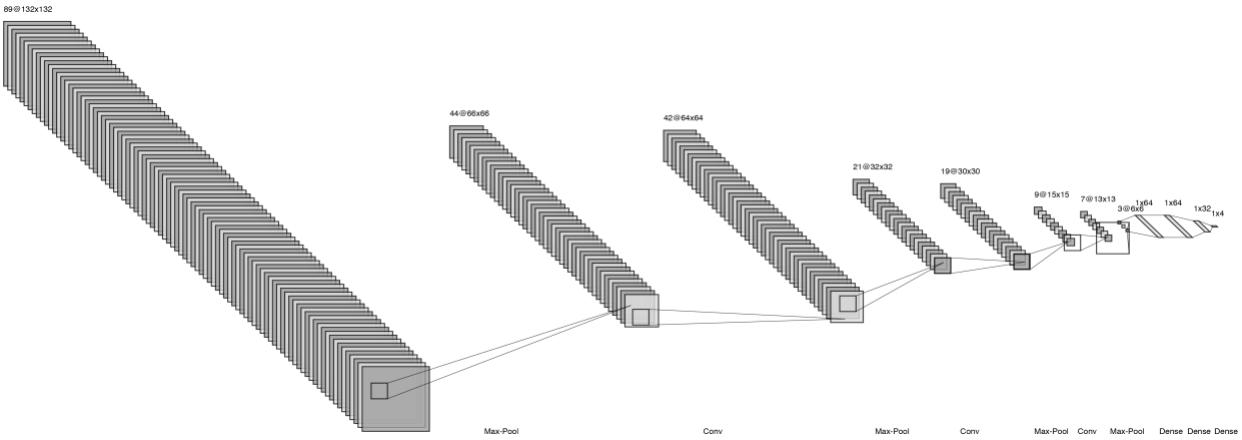


Figure 2: Seven layer CNN for genre classification (architecture for proposed solution)

2. Related literature

One of the earliest works in genre classification came from Rasheed and Shah in 2002 [1] where they used a movie trailer to analyze features like average shot length and motion content to classify the trailer into "Action" or "Not Action" labels. Since then machine learning technologies have significantly evolved and we can now see more advanced architectures being used to solve this problem.

One such work is from Kos, Pobar and Mikec in 2014 [2] where they used distance ranking, Naïve Bayes and RAKEL for multi-label genre classification. They were able to achieve fairly good validation accuracy with a small dataset of only 1500 movie posters across 6 genres.

Following the same, Kos, Pobar and Ipsic came back in 2015 [3] with another different approach to the same problem. This time they used a larger dataset of 6000 posters across 18 genres. Then they converted the multi-labels to single-labels and ran predictions using k-NN, Naïve Bayes and RAKEL to classify posters into genres. The key breakthrough in this paper was the transformation of multi-labels where they created a continuous embedded space with all labels where similar genres were clubbed together.

However, a Neural Network based approach was first taken by Kjartansson and Ashavsky in 2016 [4] where they used an ensemble of FC Net, Conv Net, VGGNet, Naïve Bayes, SqueezeNet and ResNet to predict a book's genre from its cover. They were able to achieve incredible validation accuracy of 80% using an ensemble of models. The problem they are tackling is very similar to the movie poster genre classification problem at hand and that is why their work can easily be applied to movie posters as well.

Similarly, in 2017, Chu and Guo [5] designed a new approach for this multi-label genre classification problem. They proposed to use a deep neural network combined with YOLO object detection to identify objects in movie posters and eventually classify these posters into multiple genres. They did not achieve high accuracy, however, were able to model the relationship between objects found in a poster and its genre.

Beyond academia, there are many individuals that attempt to solve this problem on independent code sharing platforms like Kaggle and GitHub. One interesting approach came from Jianda Zhou, Data Scientist at Facebook, who published a movie genre classification code on GitHub [6] that used a seven layer CNN on movie posters data coming from TMDB. Zhou obtained a validation accuracy of 73.6%.

3. Approach

To solve the multi label classification problem for genre prediction, a 7-layer CNN is constructed as shown above in Fig 2.

3.1 Architecture

The general idea is to mimic a mini-VGG styled net with 4 convolution layers and 3 fully connected layer. The CNN was implemented in Python3.6 with the Keras package as follows:

Conv Layer 1 (size=3x3, filters=32) → Max-Pool (size=2x2) →

Conv Layer 2 (size=3x3, filters=32) → Max-Pool (size=2x2) →

Conv Layer 3 (size=3x3, filters=32) → Max-Pool (size=2x2) →

Conv Layer 4 (size=3x3, filters=64) → Max-Pool (size=2x2) →

Fully Connected Layer (size=64) → Dropout (probability=0.5) →

Fully Connected Layer (size=32) → Dropout (probability=0.5) →

Fully Connected Layer (num_classes)

Each layer in the network has a ReLU activation function to introduce non-linearity in the network. The ReLU function is defined by:

$$f(x) = \begin{cases} 0 \; for \; x < 0 \\ x \; for \; x \geq 0 \end{cases}$$

Only the last fully connected layer in the network uses a Softmax activation function that squashes the output vector between interval (0, 1) to denote a probability of a poster image being classified into a set of possible genres.

3.2 Optimization

Optimizers are used to help the model "learn" by tweaking weights in the network. For this project, we use a well-known method for stochastic optimization called Adam.

Adam was chosen because inherits properties from RMSProp and AdaGrad allowing it to work well on noisy gradients (like the ones we will face with poster image data). Additionally, Adam requires little hyper parameter tuning and the step size is approximately bounded by the learning rate hyper parameter. Adam is defined by:

$$m = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot dx$$
$$v = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot dx^2$$
$$loss = loss_{t-1} - \alpha \cdot \frac{m}{\sqrt{v + \varepsilon}}$$

where β_1 is used to decay running average of gradient, β_2 is used to decay running average of gradient square, α is learning rate step size and ε is a small number used to prevent division by 0.

3.3 Loss

Loss is a summation of errors for each data point made by training and validation set. The idea is to tune the model to decrease the loss as much as possible. For this project, we use a loss function called Categorical Crossentropy (CC Loss) [7].

The reason for using this loss is because the targets for genres are encoded in multi-hot encoding since one movie poster can be associated with multiple genres. The CC Loss measures the average number of bits needed to identify an event drawn from a set. In this case, the event is correct set of genres for a given movie poster. CC Loss is defined by:

$$H(p, q) = -\sum_x p(x) \cdot \log(q(x))$$

where $p(x)$ is a 2-d tensor with each row representing a distribution and $q(x)$ is 2-d tensor where each element represents the position of 1 in multi-hot encoding.

3.4 Regularization

Regularization is a process of reducing overfitting in network. For this network, we use a regularization technique called inverted Dropout. This technique drops-out randomly chosen visible/hidden units in a neural network at a probability of p. This reduces interdependent learning amongst the neurons and reduces overfitting. For this project,

Dropout probability is fixed at p=0.5 and Dropout is not used on test set.

3.5 Tweaking hyper parameters

The CNN architecture described in Fig 2 can be tweaked across various different hyper-parameters. For the purpose of this project, two hyper parameters will be dynamically tweaked and tested: 1. Learning rate, and 2. Input image size

3.5.1 Learning rate

For the Adam optimizer discussed in 3.2, the learning rate α will be tweaked across different values in the set $[10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}]$. While keeping the betas fixed at β_1=0.9, β_2=0.999.

3.5.2 Input image size

Original poster images are shrunken by 30, 40, 50, 60 and 70% respectively. The idea is to see if the CNN can pick up on granular details on images that are less shrunk. Original image size is 182x268 pixels. Here are shrunken sizes:

Table 1: Shrunken image sizes

Shrink ratio (%)	Size (pixels)
30	55 x 80
40	73 x 107
50	91 x 134
60	109 x 161
70	127 x 188

4. Experiment

The network architecture proposed in 3.1 was trained over a dataset found on data sharing platform – Kaggle [8].

4.1 Dataset

Training data consisted of 44,000 movie posters from year 1997 to 2017. Each movie poster was accompanied with a list of genres that the movie fits into. A total of 28 unique genres were listed throughout the dataset. All data was obtained from the Internet Movie Database (IMDB). The genre distribution plot below shows bias in the dataset.

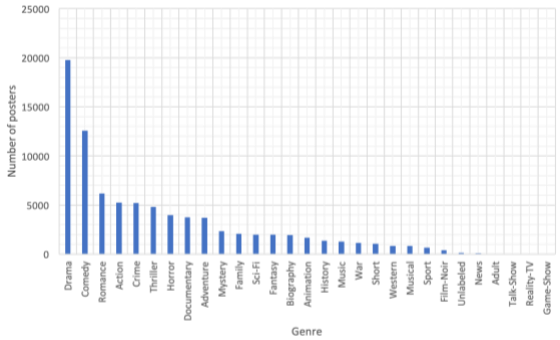


Figure 3: Poster genre distribution

In the above graph, we can see that the number of posters per genre are not equally distributed – Drama, Comedy and Adventure combined contribute for half of the dataset. To fix this, we look at the genres that co-occur the most and pick ones that are representational of the larger population of the data.

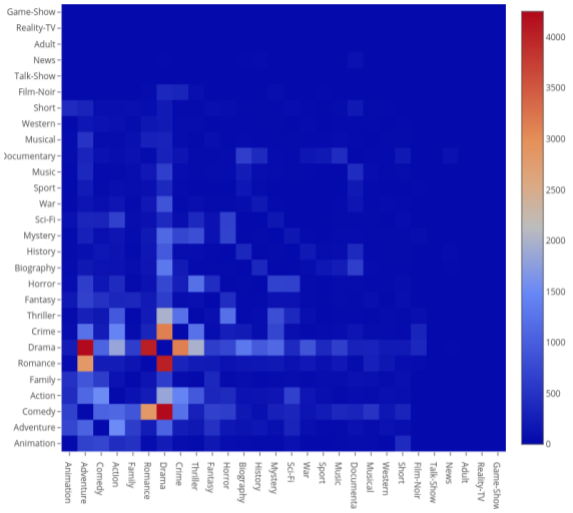


Figure 4: Genre co-occurrence heat map

Based on the genre co-occurrence heat map in Fig 4, we will pick four key genres from different co-occurrence levels that can be indicative of the larger population size. These genres are: Romance, Action, Horror and Documentary.

The network was trained on movie posters from 1997-2017 in genres: Romance, Action, Horror and Documentary.

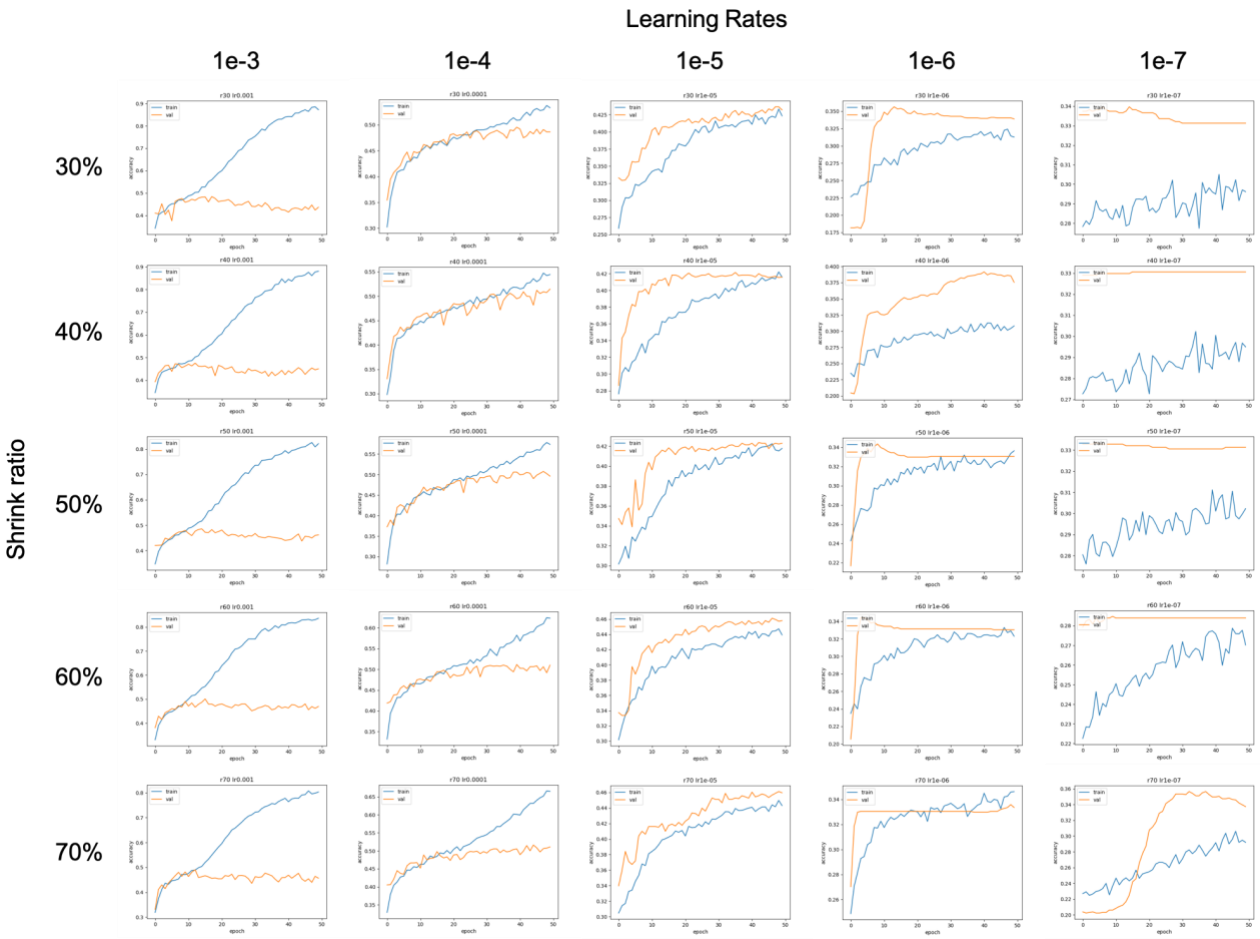


Figure 5: Training vs. Validation accuracy for all combination of hyper parameters

4.2 Training Results

Using the architecture and approach mentioned in 3, the training was conducted on an AWS G3 GPU server with 4 NVIDIA Tesla M60 GPUs. The dataset was divided into training, validation and testing folds. For training the model, only training and validation folds were used. Training the model for all combinations of hyper parameters took in total 3 days of training. Fig 5 above shows the results of training the model.

As we can observe see some models produce a validation accuracy (orange) that is higher than the training accuracy (blue). This is primarily because we have used Dropout regularization while training. When training, 50% of features are set to 0 using Dropout. When testing all features are used and scaled appropriately. Generating a model that is more robust during test time – which leads to higher accuracies than training data.

Another significant observation we make is that when learning rates are higher (1e-3), the training accuracy is much higher than validation accuracy. Here, we see that the model is remarkably overfitting. Similarly, when learning rates are lower (1e-7) the validation accuracy falls flat around 30% which goes to show that the posters are being classified almost randomly. Therefore, the learning rate within range 1e-7 and 1e-3 end up showing the most promising results. Additionally, larger image sizes only tend to give minor, but negligible, improvements on training and validation accuracies.

4.3 Testing results

The models were tested on a brand-new set of movie posters from the dataset that were not used for training or testing. The table below outlines the results from testing the models.

Table 2: Testing set accuracies (%)

Shrink ratio (%) \ Learning rates	1e-3	1e-4	1e-5	1e-6	1e-7
30	42.7	48.2	43.3	32.6	31.5
40	43.2	50.5	42.6	39.0	31.6
50	44.5	49.9	44.6	32.1	31.6
60	46.2	49.8	45.7	32.4	29.1
70	45.2	49.1	45.3	32.8	32.2

Consistent with the output in validation accuracies, testing accuracies also yield similar results. Additionally, we also see that during testing, the best model results in an accuracy of 50.5% with learning rate of 1e-4 and input image shrink ratio of 40%. Below are some of the correctly classified movie posters from each genre.

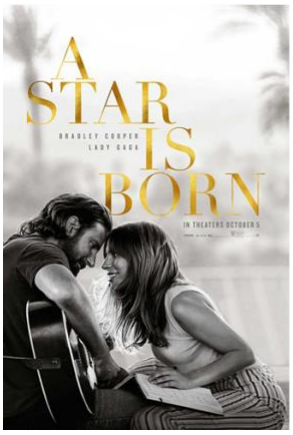
When we look closely at these correctly classified images, we see that there are certain elements that stand out for the image to fit that genre. E.g. for 'A Star is Born' in Fig 7, we see two human figures close to each other smiling which can classify as 'Romance' and IT poster in Fig 6 is completely dark and blood red colored which can classify it as 'Horror'.

Prediction:

Horror 76%
Action 22%
Documentary 2%
Romance 0%

Correct genre: Horror

Figure 6: IT movie poster - correctly classified



Prediction:

Romance 98%
Documentary 2%
Action 0%
Horror 0%

Correct genre: Romance

Figure 7: A Star is Born movie poster - correctly classified

Prediction:

Documentary 87%
Romance 13%
Action 0%
Horror 0%

Correct genre: Documentary

Figure 8: Truth to Power movie poster - correctly classified

Prediction:

Action 85%
Horror 15%
Documentary 0%
Romance 0%

Correct genre: Action

Figure 9: Star Wars movie poster - correctly classified

Similarly, below are some images that are incorrectly classified. This happens when certain elements that frequently occur in a particular genre show up in another genre. E.g. Dark background (occurring in horror movie posters) shows up in a romantic movie poster.

Prediction:

Romance 25%
Documentary 25%
Action 25%
Horror 25%

Correct genre:
Romance

Figure 10: La La Land movie poster - incorrectly classified

Prediction:

Romance 73%
Documentary 19%
Horror 5%
Action 3%

Correct genre:
Horror

Figure 11: The Nun movie poster - incorrectly classified

Prediction:

Romance 48%
Action 26%
Documentary 16%
Horror 10%

Correct genre:
Action

Figure 12: Wonder Woman movie poster - incorrectly classified

As we see, we can deduce explanations for some misclassifications such as La La Land in Fig 10 or Stories we tell in Fig 13 where the images are too tiny. However, there is not much explanation to deduce when it comes to misclassification of a movie like The Nun in Fig 11, which clearly to a human eye is a horror movie poster.

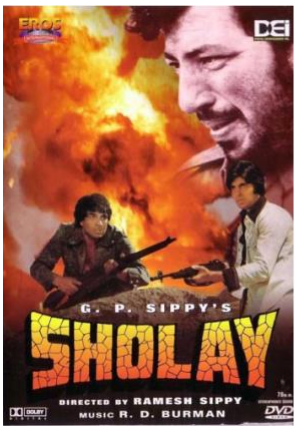
Prediction:

Romance 49%
Documentary 19%
Action 17%
Horror 15%

Correct genre:
Documentary

Figure 13: Stories We Tell movie poster - incorrectly classified

Since the model was only trained on Hollywood movies, I decided to run the model predictions on Bollywood movies to check how the model functions on posters from a different culture. The following are Bollywood movie posters that were classified correctly by the model.

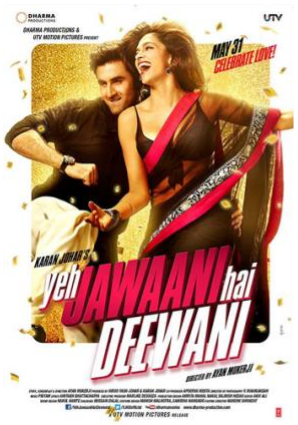


Prediction:

Action 94%
Romance 6%
Documentary 0%
Horror 0%

Correct genre:
Action

Figure 14: Sholay movie poster - correctly classified



Prediction:

Romance 90%
Action 5%
Documentary 4%
Horror 1%

Correct genre:
Romance

Figure 15: Yeh Jawani Hai Deewani movie poster – correctly classified

And the Bollywood movie poster that was classified incorrectly by the model.

Figure 16: Om Shanti Om movie poster - incorrectly classified

Moving one step further, I decided to run the model through everyday images that are not movie posters to better understand if the model could pick up on certain visual cues.

Figure 17: Obama handshake - correctly classified

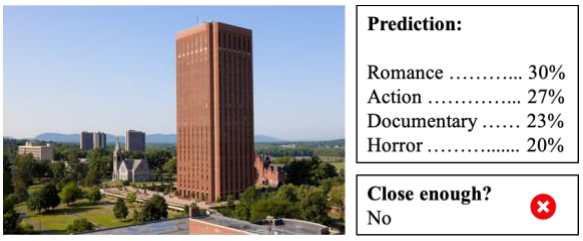


Figure 18: UMASS Library Building - incorrectly classified

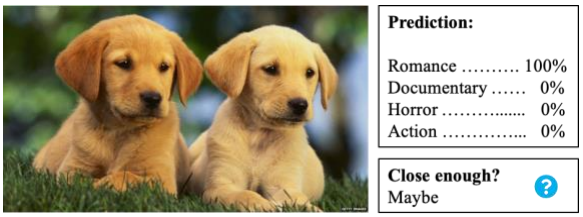


Figure 19: Puppies - not sure

Therefore, we observe that the model may pick up on certain visual cues. E.g. a photo of U.S. Presidents in Fig 17 is more likely to appear on a documentary cover. Similarly, two individuals next to each other often appears in Romance posters which is mapped to two dogs sitting next to each other.

5. Conclusion

We note that the model proposed in this paper does well on certain movie posters where the elements in the poster are clearly visible and large enough. Movie posters in Fig 6, 7 and 8 attest to this fact. A learning rate of 1e-4 yields the highest accuracy of 50.5% while input image sizes only have a minor effect on the model's overall accuracy.

This work can be extended further by training the model on all 28 genres or training the model on movie posters across different cultures – e.g. Chinese, French, Indian movies. Another interesting approach would be to train a Recurrent Neural Network (RNN) with Long Short Term Memory (LSTM) to predict the plot of a movie from its poster. Similarly, we could also use a Generative Adversarial Network (GAN) to automatically create a unique movie poster based on its genre.

At the end, as simple as this idea may be, it truly has the potential to expand in different areas of life. This project by itself is the first baby step to help a machine understand storytelling through identifying movie genres.

6. References

[1] Rasheed, Zeeshan & Shah, Mubarak. (2002). Movie Genre Classification By Exploiting Audio-Visual Features Of Previews. 2. 10.1109/ICPR.2002.1048494.

[2] Ivašić-Kos, Marina & Pobar, Miran & Mikec, Luka. (2014). Movie posters classification into genres based on low-level features. 1198-1203. 10.1109/MIPRO.2014.6859750.

[3] Ivašić-Kos, Marina & Pobar, Miran & Ipsic, Ivo. (2015). Automatic Movie Posters Classification into Genres. Advances in Intelligent Systems and Computing. 311. 319-328. 10.1007/978-3-319-09879-1_32.

[4] Kjartansson, Sigtryggur & Ashavsky, Alexander. (2017). Can you Judge a Book by its Cover? Retrieved from http://cs231n.stanford.edu/reports/2017/pdfs/814.pdf
[5] Chu, Wei-Ta & Guo, Hung-Jui. (2017). Movie Genre Classification based on Poster Images with Deep Neural Networks. 39-45. 10.1145/3132515.3132516.
[6] Zhou, Jianda. (2017). Genre classification on movie posters. Retrieved from https://www.linkedin.com/pulse/genre-classification-movie-posters-jianda-zhou/
[7] Deeplearning.net. Ops for neural networks. Retrieved from http://deeplearning.net/software/theano/library/tensor/nnet/nnet.html#theano.tensor.nnet.nnet.categorical_crossentropy
[8] Kaggle.com. (2018) Movie genre from its poster. Retrieved from https://www.kaggle.com/neha1703/movie-genre-from-its-poster