Power in the Pixels: Predicting Film Genres and Box Office Revenue

Jacob Baehr*, Minnie Kim†, John K. Lewis‡, and Nhan Phan§
Department of Computer Science
Florida Southern College
Lakeland, Florida
Email: *baehrjacob@gmail.com, †minniekim00@hotmail.com, ‡jklewis99@gmail.com, §phuocnhan18062001@gmail.com

***Abstract*—A considerable amount of financial risk is associated with the production of films today. Producers are looking for ways to identify trends in marketing and to predict whether a movie will be successful at the box office. We propose that machine learning models can be useful to help solve these problems, specifically for genre prediction and box office revenue prediction. Accurately predicting the genre of a film is useful for recognizing trends in poster design, which is an important part of marketing for the film. Popularity of a movie and the awards it receives can also be important factors for production. Using only an image of a movie's poster, we predict the movie's genres using Convolutional Neural Networks, achieving overall genre accuracy of 95.9%. We also build Linear Regression, Support Vector Regression, and Random Forest Regression models to predict box office revenue given a movie's popularity, awards and genres, achieving an R^2 value of 0.66.**

***Index Terms*—CNN, multi-label classification, regression, movies**

I. Introduction

The fears that Artificial Intelligence will take over the world are prevalent in many of the click-bait headlines from fear-mongering media outlets and desperately relevant plots in cinema and television. These fears, though, seem a little extreme. A.I. lacks the emotional conscience to really encapsulate the character-driven circumstance of human suffering or to color a canvas with the pigments of nostalgia. But what if A.I. could help find the features of the art that inflict such an inclination of thought or expenditure to see more? Using only the digital image of a movie poster, we seek to answer questions like "Can we accurately classify the genre of a movie?" or "What is present in the movie posters that lead to a film's success?" Though more marketing funds for 21st century movies are directed toward digital media, the poster remains encased in theaters and streaming service icons. Evidently, it is a core element in the marketing strategy and can help summarize a film in just a few colors and images.

A movie is more than just a poster, though. Its success is often driven by its cultural popularity, awards of which it is a recipient, and its genre(s). Therefore, we suggest that these features can be good predictors of a movie's box office revenue. Indeed, as the movie industry continues to grow, online resources can offer a considerable amount of information that can be used for data analysis. For example, the Internet Movie Database (IMDb) [1] provides information on movies, such as their production features and awards, as well as an outlet for users to write reviews and rate movies, effectively capturing the popularity of a film.

We seek to use these posters and movie information for machine learning tasks. Specifically, the goal of this research is to build machine learning models to both predict the genres of a film based on the movie poster and predict box office revenue given movie metadata associated with popularity, awards and genres. Using data generated from [2], derived from IMDb, we utilize pre-built Convolutional Neural Network (CNN) architectures–XceptionNet [3], InceptionResNetV2 [4], NasNetLarge [5]–for multi-label classification of movie genres. Using these posters' corresponding movie information [6], we develop several regression models–Linear Regression, Support Vector Regression (SVR), and Random Forest Regression (RFR)–for box office prediction. We compare the success of the CNNs on overall genre prediction accuracy and fully correct prediction accuracy, with NasNetLarge achieving 95.9% and 41.9% in each category, respectively. We compare the success of our linear regression models in terms of the coefficient of determination, R^2, and the best regression model, RVR, achieves an R^2 value of 0.66. Our code has been made publicly available at [7].

II. Methods

A. Data Acquisition

This work is a composition of our approaches to two tasks: genre classification and box office prediction. For genre classification, we used data from the Movie Posters Dataset [2]. This dataset consists of a comma-separated values (CSV) file with columns "Id" and "Genre", and a column for each of the 25 genres. Additionally, this dataset includes a JPEG image for each movie in the CSV file, identified by the "Id". In total, the Movie Posters dataset consists of 7,254 movies. However, due to some movies containing duplicate entries in the CSV, we deleted the duplicates, which reduced the total to 7,242 movies. For box office prediction, we went to the source of the Movie Posters Dataset for movie metadata [6]. Additional features about the movie were gathered, and these data were distributed among .txt files, so we wrote a script to translate the metadata into a CSV file. This dataset consists of 8,873 movies with features "_id", "Language", "Country", "Rated", "imdbID", "Awards", "Poster", "Director", "Released", "Writer",

TABLE I: Value Count and Average Number of Genres

Number of Movies	Number of Genres
1321	1
2239	2
3682	3
Average	2.33

“imdbVotes”, “Runtime”, “Response”, “imdbRating”, “Title”, “Genre”, “Actors”, “Type”, “Metascore” and “Box_office”. The values for these features were scraped from the IMDb website by the creators of this dataset. The movies in this dataset span from 1980 to 2015, inclusively.

No data cleaning or augmentation was performed in regards to the Movie Posters Dataset. As shown in Fig 1, most movies in the dataset have at least one of their genres classified as “Comedy” or “Drama”. Notice that these totals span beyond the size of the dataset of 7,242 movies; this is because most movies have more than one genre, as shown in Table I. It is therefore evident that the imbalance in the labels of the data is likely to influence our results. This dataset was split into training data and testing data for multi-label classification with pre-built CNN architectures. 80% of the complete dataset was used for training, and the remaining 20% was used for testing.

In contrast, the movie metadata was required to be cleaned due to most features having a large number of missing values, as shown in Table II. To clean this data, we first removed any duplicates and eliminated all movies without a poster, with the goal of aligning the Movie Posters Dataset with this source dataset (as we wanted to potentially include the poster as a feature for box prediction). After this step, 9 movies did not have a value for “Director” so, we removed them as well. Next, only films with Motion Picture Association of America (MPAA) ratings of “G”, “PG”, “PG-13”, “R”, “Not Rated” and “Unrated” were considered, which simultaneously eliminated films that were released only on TV. Missing values for “Language”, “Country” and “Runtime” were filled with their respective most common values, which were “English”, “USA” and “90 min”. Missing values for “imdbVotes” and “Metascore” were filled with their respective averages. Of the films without a value for “Released”, if a value was found on Wikipedia for a specific release month, day and year, it was replaced with that value, but if it was not found, it was removed. Missing values for “Actor” and “Writer” were filled with the string, “Unknown”. In addition, columns such as “Oscar_noms”, “Oscar_wins”, “Golden_globe_noms”, “Golden_globe_wins”, “BAFTA_noms”, “BAFTA_wins”, “Other_noms” and “Other_wins” were created to translate the string values for “Awards”, where each column was named based on the presence of those awards in the “Awards” column. Following this additional cleaning, the only feature with missing values was “Box Office”, with 1,680 movies missing a value for this feature; thus, we eliminated these 1,680 films to ensure the dependent variable was accurate and unassumed. It made most sense to eliminate movies with rarely present features such as those with genres “Short”, “Reality-TV” and “N/A”, as these values had totals less than 2. Furthermore, due to the fact that the number of categories for languages, countries, release dates, writers and actors were so large, it was unreasonable to one-hot encode or to include these features by taking significant steps to parse the values. After all of the cleaning, 4,992 movies were left. We split the data into training and testing with an 80%, 20% split, respectively.

TABLE II: Features and their Count of Missing Values

Feature	Number of Missing Values
Language	70
Country	22
Rated	1507
imdbID	0
Awards	2755
Poster	639
Director	118
Released	241
Writer	634
imdbVotes	346
Runtime	236
Response	0
imdbRating	345
Title	0
Genre	31
Actors	158
Type	0
Metascore	3997
Box Office	2729

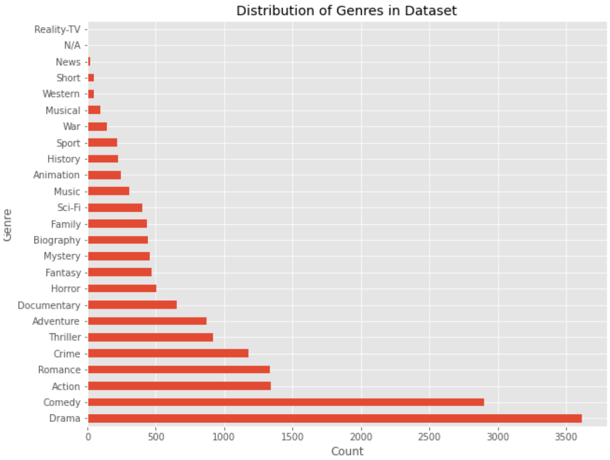


Fig. 1: An overview of the number of films in each genre

B. *Genre Classification*

For genre classification, we tested three convolutional neural networks (CNN): XceptionNet [3], InceptionResNet [4], and NasNetLarge [5]. Each was trained with the following hyperparameters:

- batch size of 4

- 100 epochs
- Adam optimizer with learning rate of 0.001
- binary cross entropy loss function
- patience of 10 epochs

We chose a smaller batch size due to limitations in our hardware. Additionally, our models had converged without improvement before 100 epochs, which is the reason behind setting and maintaining this *epochs* parameter at 100 and the *patience* parameter to 10. Binary cross entropy is used because of the definition of this genre classification problem: multi-label classification. This means that the output layer can have 1 or more activated output neurons. For each output neuron, we apply a minimum threshold, or confidence, requirement to be considered activated. We use the Adam optimization as is traditional in much of the research on CNNs.

For each of these models, the final classification layer was transformed to fit the number of genres possible for our dataset. Specifically, we used a multilayer perceptron in our final layer with input size equal to the output size of the image classification model, with 25 output neurons. This sigmoid activation function is calculated on this final output layer to reduce the values of the neurons between values of 0 and 1. The NasNetLarge model consisted of approximately 85 million parameters, the InceptionResNetV2 model consisted of approximately 59 million parameters, and the XceptionNet model consisted of approximately 22 million parameters. The input into each of these models differed: for NasNetLarge, the input size is an image of shape (331, 331, 3); for InceptionResNet, the input shape is (300, 300, 3); for XceptionNet, the input shape is (299, 299, 3).

C. Box Office Prediction

In order to predict the box office revenue, we needed to determine the relationship between the dependent and the independent variables. Therefore, we used three regression models: Linear Regression, Support Vector Regression (SVR), and Random Forest Regression (RFR). To compare the performance of our models regarding their accuracy in predicting the actual box-office revenue, we use the coefficient of determination, R^2. For a perfect model, the best R^2 value is 1, which means all variance in the data is accounted for. However, values can be less than 0, implying that the model performs worse than a constant prediction of the mean of the data. We try to improve our models by making changes that push the R^2 closer to 1.

Hyperparameter tuning for the regression models was brief. No parameters were changed for our linear regression model, meaning the regressor does not assume centered data. For the RFR model, however, we adjusted parameters for the number of trees in the forest and the max depth of each tree. We test a variable number of trees used in the forest within the range 20 to 100, with an increment size of 5. We save the forest with the best R^2 value within the range based on the set hyperparameters. For our SVR model, we implemented 4 kernels: linear, polynomial, radial basis function (rbf) and sigmoid. Data is scaled for the SVR model to prevent any one feature from trumping others. No other hyperparameter tuning was performed.

III. Results

A. Genre Classification

To compare the success of the CNNs, we use *overall* genre accuracy and *fully correct* prediction accuracy. We define overall genre accuracy as the sum of all correctly predicted outputs divided by the product of the number of samples times and number of output neurons (25). Fully correct prediction accuracy is defined as the number of samples in which all of the associated labels were correctly classified, meaning all output neurons (genres) that were a part the true label of the movie were activated and no other output neurons were activated.

Of our CNNs, the NasNetLarge model outperforms XceptionNet and InceptionResNetV2 in terms of both overall genre accuracy and fully correct prediction accuracy. The NasNetLarge model achieves overall genre accuracy of 95.9% and fully correct prediction accuracy of 41.9% on the test set, both with a threshold of 0.4. The InceptionResNetV2 model achieves overall genre accuracy of 93.0% and fully correct prediction accuracy of 16.9% on the test data, both using a threshold of 0.4. The XceptionNet model achieves overall genre accuracy of 92.5% and a fully correct prediction accuracy of 10.7%, both with a threshold of 0.4. Based on these results, the best threshold, or minimum confidence, to use on the output nodes is 0.4 for each model, meaning that, for a genre to be predicted by the model, the output node requires a confidence value greater-than or equal to 0.4. We show the comparison of the model accuracies across varying thresholds in Fig. 2. A table of the results is also shown in Table III.

These results suggest that, generally, most genres are accurately predicted, but when it comes to predicting all genres perfectly, the model can still be improved. Still, the model significantly outperforms a single random guess (without considering prior probabilities), which, for a maximum number of 3 genres, has an expected accuracy of 0.038%, as defined by

$$100 \times \frac{1}{\binom{25}{1} + \binom{25}{2} + \binom{25}{3}}.$$

This is an increase in accuracy by more than 100,000% for the NasNetLarge model. We show in Fig. 3 examples of *fully correct* prediction accuracy, *partially correct* predictions, and *fully incorrect* prediction accuracy.

B. Box Office Prediction

When predicting box office revenue, R^2 values were relatively low for most regression models. For the Linear Regression model, we obtained an R^2 value of 0.58. The SVR models, with use of the four kernels–linear, polynomial, RBF, and sigmoid–achieved R^2 values of 0.56, 0.05, 0.52 and -146.05, respectively. Of the RFR models, we obtained a best R^2 value of 0.66 using 30 trees and no max depth parameter set.

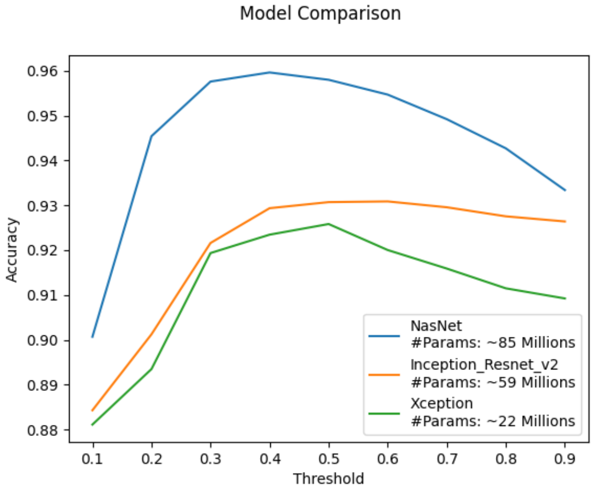


Fig. 2: Comparison of different thresholds on overall genre prediction accuracy

TABLE III: Comparison of the Different Metrics for each CNN

Model	**Overall Genre Accuracy**	**Fully Correct Prediction Accuracy**
NasNetLarge	95.9%	41.9%
InceptionResNetV2	93.0%	16.9%
XceptionResNet	92.5%	10.7%

By comparing these different R^2 values, we can confirm that the RFR model outperforms other regression models in predicting box office revenue, accounting for 66% of the variance in the box office revenue. The Linear Regression model and SVR models with linear and RBF kernels had R^2 values similar to, but not better than, the RFR model. Referring to Fig 4, a generally linear trend between the model's predictions and the actual box office revenue values is evident. Still, as suggested by the R^2 value, a large amount of variance is still present. For the best performing models, most of the predicted revenue values range between \$10 and \$200 million, which is within the same range of the actual values. In comparison, the models with lower R^2 values have a less defined relationship between the predictions and actual values and do not fit most of the predictions in this aforementioned range. Hence, it is not surprising that these models have lower R^2 values.

Regarding regression models, to evaluate which features are most influential in predicting revenue, we calculated the pearson-r correlation coefficient between some features, specifically for those features without a significant number of of values equal to 0. These included “imdbVotes”, “imdbRating”, and “Metascore”. We show these pearson-r values in Table IV.

TABLE IV: Pearson-R Calculations

Feature	**Pearson-R**
imdbVotes	0.68
imdbRating	0.20
Metascore	0.13

The feature “IMDb votes” has the highest pearson-r correlation with the box office revenue of the features we tested, suggesting, expectantly, that the more reviews a movie has, the more tickets it has sold. Ratings on both IMDb and Metascore are shown to have less correlation than “imdbVotes”, which is also expected, suggesting that movies that are rated highly do not always sell the most tickets.

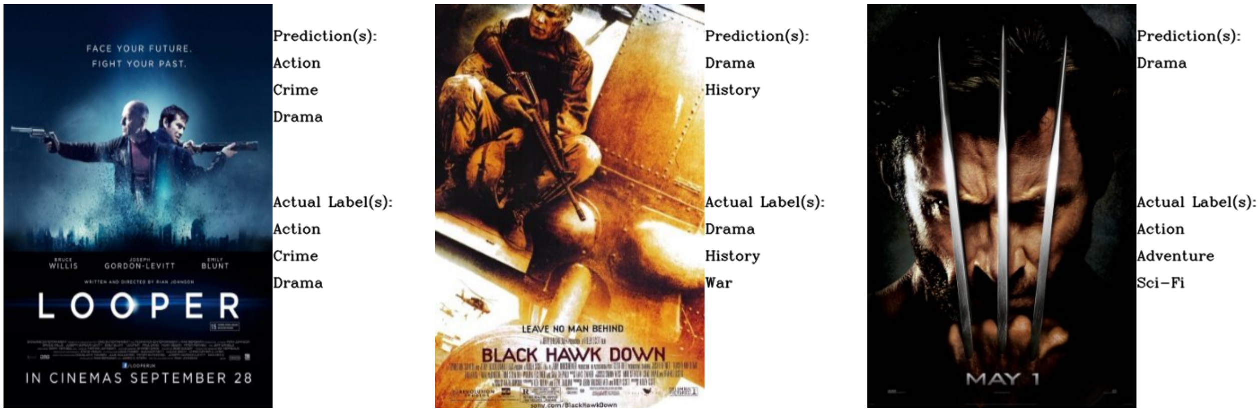
IV. Conclusion

In this work, we have shown success in identifying genres of a movie using only an image of the movie poster, achieving an accuracy of 95.9% and a fully correct prediction accuracy of 41.9%. Additionally, we have shown that RFR models outperform other regression models with an R^2 value of 0.66.

Though NasNet outperformed all other models, its large number of parameters makes it difficult to implement in practice. Because it is so large, predictions for single films without an NVIDIA-enabled GPU take approximately 60 seconds to run. In contrast, the XceptionNet model, which still consists of approximately 20 million parameters, can generate a single prediction in only a few seconds. This evaluation of speed is important to consider in application, but for our purposes, the model which has the best genre accuracy is defined as the best performing model.

Some methods could be altered to improve the results. For example, because of our own hardware limitations, the batch size could be increased to improve both training speed and overall accuracy. Additionally, fully correct prediction accuracy is still relatively low at 41.9%, so changes to the activation function in the output neurons could potentially improve results, specifically when the maximum number of possible genres is specified. For example, when 3 genres is the known maximum number of genres, a softmax activation function in tandem with a threshold of 0.25 plus a small epsilon value could limit the number of predictions to three. This could only work for confident models though, and in practice, will likely be difficult to refine. The softmax activation function is generally most appropriate for multi-class classification, but it could potentially be effective in multi-label classification, at the very least for confident models.

The data, itself, also has revealed some opportunities for change. As mentioned, “Drama” and “Comedy” dominate the other genres, causing a strong imbalance in the data. Data augmentation could be used to reduce the influence of certain genres. Approaches such as random sampling, transformations and reflections on the image to increase the number of infrequent genres with less samples could improve results. Moreover, because of each model's dependency on its training data–IMDb's classifications of genre–approaches to verify genres like cross-validating with other sources, such



(a) Fully Correct Predictions (b) Partially Correct Predictions (c) Fully Incorrect Predictions

Fig. 3: NasNetLarge Predictions using a confidence threshold of 0.4

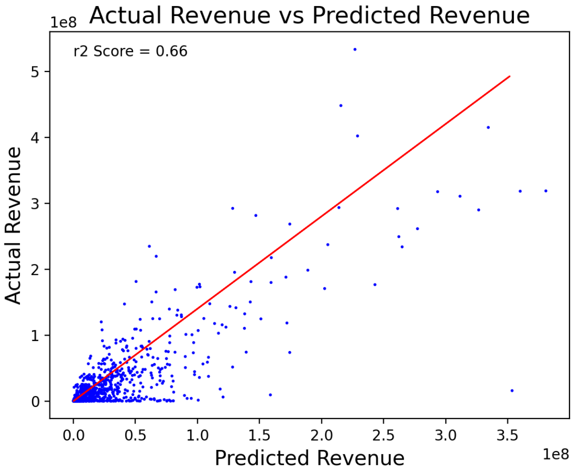


Fig. 4: Actual Box Office Values vs Predictions of SVR Model

as The Movie Database (TMDb), could potentially improve accuracy by reducing mis-labeled movies.

Regarding box office prediction, our features suggest that popularity, award count, and genres are not the best features for predicting box office revenue. Additionally, these features are not the most relevant for developing a prediction model in industry, as the features used for this research are mostly only generated after large sums of money have been invested into a production and marketing. We suggest that it would be most useful to collect more data on pre-released features which would allow our models to make box office predictions before a movie is released. This would help both prediction accuracy and provide a real-world application for this research. With the addition of new features, we argue the model can be better understood with evaluation of multicollinearity between features as well.

Acknowledgment

We want to express our gratitude to our professor, Dr. Eicholtz, for his endless encouragement and commitment to providing resourceful machine learning education.

References

[1] "IMDb." https://www.imdb.com.

[2] Movie Posters. *Kaggle/raman*. Available: https://kaggle.com/raman77768/movie-classifier. Accessed: Oct. 30, 2020.

[3] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

[4] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," 2016.

[5] B. Zoph, V. Vasudevan, J. Shlens, and Q. Le, "Learning transferable architectures for scalable image recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8697–8710, 2018.

[6] W.-T. Chu and H.-J. Guo, "Movie Genre Classification based on Poster Images with Deep Neural Networks," in *Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes - MUSA2 '17*, (Mountain View, California, USA), pp. 39–45, ACM Press, 2017.

[7] J. K. Lewis, N. Phan, M. Kim, and J. Baehr, "jklewis99/magical-movie-poster-processing." https://github.com/jklewis99/magical-movie-poster-processing, 2020.