YZV302E Deep Learning Project Predicting Box-Office Revenue Using Movie Posters and Metadata with Multimodal Deep Learning

Tevhidenur Serdar
Artificial Intelligence and Data Engineering
Istanbul Technical University
serdar21@itu.edu.tr
150210335

Ilayda Kara
Artificial Intelligence and Data Engineering
Istanbul Technical University
karai22@itu.edu.tr
150220747

***Abstract*—The main objective of this project is the prediction of box-office revenues using multimodal deep learning techniques based on both visual and textual information. Movie posters which are summaries of the visual entity of the marketing campaigns of a film, has meaningful visual features extracted using state-of-the-art CNNs like EfficientNetB0, EfficientNetB7, and ResNet50. Metadata attributes director, stars, genre, country, runtime, and budget were added into the metadata-processing pipeline using MLP and transformer. The fusion of these features enable a regression model to predict the continuous variable representing box-office revenue with ease. Using a dataset collected from the OMDb API that contains high-resolution poster images along with metadata, the model shows the potential for integrating visual and textual data in improving predictive performance in a financial forecasting context.**

***Index Terms*—box-office, poster, metadata, model, EfficientNet, model, feature, prediction**

I. Introduction

Movie posters are designed to attract targeted audiences, and they present the theme, genre, and emotional resonance of a film. These graphical features, with associated metadata such as genre, director and budget offer a complete basis for analyzing a film's predicted success. However, use of these distinct data sources for predictive modeling requires different methodologies. Traditional forecasting models focus mainly on historical financial data or external market dynamics, thus overlooking the both visual and textual attributes that will effect audience's perspective. This project takes up the challenge of predicting a movie's box-office revenue by applying an input feature set that combines visual and metadata attributes using multimodal deep learning techniques.

II. Related Work

The study titled *"Using movie posters for prediction of box-office revenue with deep learning approach"* [1] adopts a distinct methodology compared to our project. Their model achieved an accuracy of 33.05%. Below, we outline the similarities and differences between the two studies:

A. Similarities

- **Visual Data as Input**: Both projects treat movie posters as a significant source of predictive features, emphasizing their importance in capturing visual and thematic aspects of movies.
- **Poster Feature Extraction**: Both studies utilize Convolutional Neural Networks (CNNs) for extracting features from movie posters. The referenced work uses Inception V3, while our project uses models such as EfficientNetB0, EfficientNetB7, and ResNet50.

B. Dissimilarities

- **Objective**: The referenced study formulates the problem as a classification task to categorize movies into predefined box-office revenue ranges. In contrast, our project addresses a regression problem to predict exact box-office revenue.
- **Modeling Approach**: The referenced study employs a single Inception V3 model for feature extraction and classification. Our project, however, uses a multimodal approach by combining outputs from pre-trained CNNs and a Multi-Layer Perceptron (MLP) for metadata processing.
- **Data Scope and Features**: The referenced work relies solely on visual features extracted from movie posters, while our project incorporates additional metadata such as genre, director, runtime, and budget to improve predictive accuracy.

III. Dataset

A. Data Collection

The dataset for this project was constructed using the OMDb API. Python scripts were utilized to query the API and retrieve comprehensive movie data, including titles, genres, release years, actors, director, country and runtime. However, since budget information was not available in the OMDb API, additional research was conducted to manually collect budget data for all movies. This information was then recorded in a separate CSV file and merged with the main dataset.

Additionally, movie poster links were extracted from the dataset and used to download poster files for further analysis, but it was subsequently dropped from the data before modeling. The downloaded posters were stored locally.

B. Data Preparation

The dataset was divided into two main parts:

- **Training Data**: Saved in a file named *train.csv* and poster images in *poster_train* folder.
  This subset was used for training the deep learning model.
- **Testing Data**: Saved in a file named *test.csv* and poster images in *poster_test* folder.
  This subset was used to evaluate the model's performance.

C. Data Limitations

The dataset contains movies released between 2009 and 2019. This limitation may influence the generalizability of the model to movies from other time periods. Furthermore, the dataset is imbalanced in terms of the *BoxOffice* variable, with very few movies having extremely high revenue, as shown in Figure 1. This imbalance reflects the natural distribution of movie revenues. To address this skewness, a *log1p* transformation was applied to the *BoxOffice* variable, which helps normalize the distribution and improve the model's ability to handle extremely high revenue.

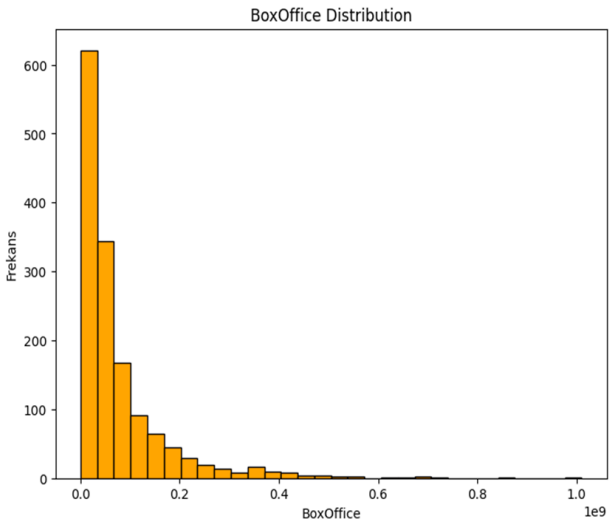

Fig. 1. Distribution of BoxOffice Revenue

D. Data Variables

Table I provides an overview of the variables included in the dataset, along with their descriptions.

The dataset was meticulously prepared and structured to ensure its compatibility with the deep learning model. The inclusion of both training and testing subsets enabled effective model evaluation and validation.

TABLE I
DATASET VARIABLES AND DESCRIPTIONS

Variable Name	Description
Title	The title of the movie
Year	The release year of the movie
Director	The director of the movie
Runtime	The runtime of the movie in minutes
Country	The country where the movie was produced
Actor1, Actor2, Actor3	The main actors in the movie
Genre1, Genre2, Genre3	The genres associated with the movie
BoxOffice	The box office revenue of the movie
Budget	The production budget of the movie
Poster	URL to the movie poster (used for downloading)

IV. METHODOLOGY

A. Preprocessing

Below are the detailed steps undertaken during preprocessing:

1) Handling Missing Values:

- The dataset initially contained missing values in several columns, including `Year`, `Genre`, `Director`, `Actors`, `BoxOffice`, and `Runtime`. Rows with missing values in essential columns were dropped to maintain data integrity. This reduced the dataset to entries with complete information.
- For unseen categories during testing, an `Unknown` label was dynamically added to categorical encodings.

2) Data Cleaning:

- **Numeric Conversion**: The `Runtime` column, initially represented as text (e.g., "138 min"), was cleaned by removing the "min" suffix and converted to integers.
- **Currency Conversion**: The `BoxOffice` column, which included currency symbols and commas (e.g., "$936,662,225"), was stripped of these characters and converted to numeric values for analysis.
- **Log Transformation**: The target variable, `BoxOffice`, was transformed using $\log(1 + x)$ to reduce skewness in the revenue distribution.

3) Removing Irrelevant Entries:

- Rows corresponding to short films were removed by filtering out entries where the `Genre` column contained the term "Short."

4) Feature Engineering:

- **Splitting Multi-Value Columns**: The `Actors` and `Director` columns, which contained multiple values separated by commas, were split into separate columns (`Actor1`, `Actor2`, `Actor3`, etc.). Only the top three actors and one director were retained for simplicity.
- **Genre Categorization**: The `Genre` column was split into three separate columns: `Genre1`, `Genre2`, and `Genre3`, based on the order of listed genres.
- **Country Simplification**: For the `Country` column, only the first listed country was retained. Countries were then grouped into categories, with less frequent countries categorized as "Other."

5) Normalization and Transformation:

- **Runtime Normalization**: The runtime values were checked to ensure consistency, with outliers adjusted or removed.
- **Monetary Adjustment**: All monetary values, such as `Budget` and `BoxOffice`, were adjusted to the 2019 dollar index to account for inflation, ensuring consistency with the latest movie year in the dataset.
- **Poster Preprocessing**: Poster filenames were normalized by combining the movie title and year after cleaning special characters. Missing images were replaced with random noise to avoid gaps in the dataset. Each poster image was resized to 300x300 pixels (for EfficientNetB7) and normalized to [0,1].
- **Feature Scaling**: Numerical metadata features such as `Runtime` and `Budget` were standardized using the StandardScaler.

B. Image Feature Extraction

Poster features were extracted using three pre-trained Convolutional Neural Networks (CNNs):

- **EfficientNetB0**: A lightweight and efficient model for extracting high-level image features.
- **EfficientNetB7**: A more complex variant capable of capturing detailed and nuanced features due to its larger architecture.
- **ResNet50**: A deep residual network excelling at avoiding vanishing gradient issues and extracting detailed mid-level features.

All models were pre-trained on the ImageNet dataset to leverage transfer learning. Their layers were frozen to retain pre-trained weights, and a global average pooling layer was applied to reduce dimensionality. The final poster features were passed through a dropout layer for regularization.

C. Metadata Feature Extraction

Metadata features were processed using a Multi-Layer Perceptron (MLP):

- Input layer: Accepts concatenated numerical and encoded categorical features.
- Hidden layers: Two dense layers with 128 and 64 units, each followed by ReLU activation and dropout for regularization.
- Output layer: Provides intermediate features for fusion.

D. Fusion Mechanism

The poster features and metadata features were concatenated into a unified feature vector. The combined vector was passed through additional dense layers to learn synergistic relationships. The final output layer predicted the log-transformed box office revenue.

E. Training Procedure

- **Data Splitting**: The dataset was pre-split into training and test files. The training set was further split into training and validation subsets (80%-20%).
- **Optimizer**: We experimented with various optimizers, including Adam, AdamW, and RMSProp. Among these, the Adam optimizer with a learning rate of 0.0005 yielded the most favorable results, demonstrating superior convergence and model performance.
- **Loss Function**: Mean Squared Error (MSE) was employed.
- **Regularization**: Dropout layers reduced overfitting.
- **Epoch and Batch**: We tested our model with batch sizes of 16, 32, and 64, and obtained the best results with a batch size of 16. Training ran for a maximum of 100 epochs with early stopping.
- **Early Stopping**: Training halted when validation loss did not improve for 5 consecutive epochs.

V. Results

The evaluation of three multimodal deep learning models was conducted on a prepared and preprocessed test dataset. The test set included unseen data which ensures an accurate evaluation of the models' generalization capabilities. The preprocessing steps included handling missing data, normalizing numerical features, encoding categorical variables, and resizing images for consistent input to the neural networks. The evaluation metrics used were Mean Squared Error (MSE), Mean Absolute Error (MAE), and log revenue prediction error (log1p), summarized as follows in the Table II [2]. EfficientNetB7, demonstrated strong performance by effectively utilizing poster data and metadata for predictions. It outperformed all models by achieving the lowest error rates of test data despite of the superiority of EfficientNetB0 in error rates of training and validation sets. This can be attributed to the simple arthitecture of EfficientNetB0 which may lack the capacity of generalizing the data well, and deeper architecture of EfficientNetB7 which has an enhanced ability to generalize and integrate visual and metadata features.

TABLE II
Results of Multimodels

	ResNet50 MLP	**EfficientNetB7 MLP**	**EfficientNetB0 MLP**
MSE	4.646	3.837	5.976
MAE	1.621	1.551	1.821
log1p	$4.06 Million	$3.72 Million	$5.18 Million

It is also shown that the addition of poster features helped improve the general performance of the models. The models trained using both poster data and metadata clearly outperformed the ones with metadata alone in the validation MAE graph as shown in the Figure 2. This highlights the high contribution of the visual cues in the prediction of the box-office revenues. Posters provided distinctive information that metadata could not offer, such as information about the genre of the film, themes, tone, and aesthetic properties. These visual features represent aspects of marketing strategy, production quality, and audience engagement intentions, all of which are critical determinants of box office performance for films. The addition of different visual representations, together with

textual metadata, has helped the models to distinguish complex patterns and thus lead better accuracy in revenue forecasting. This underlines the importance of using multimodal data for robust and comprehensive predictive modeling in similar domains.

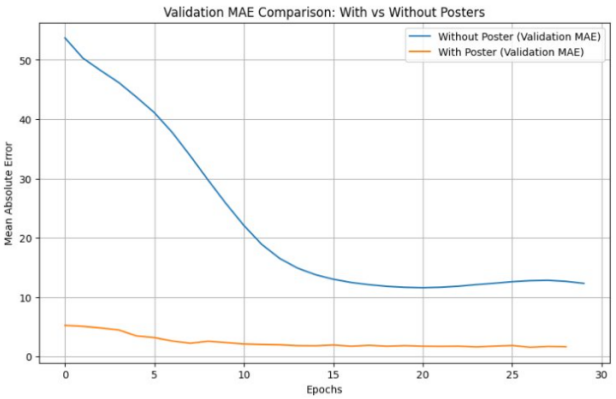

Fig. 2. Validation MAE Comparison: With and Without Posters

VI. Discussion

These results underline the advantages of multimodal approaches to box-office revenue prediction. The addition of poster data with corresponding metpredictionadata significantly increased model performance, thus proving the importance of using visual characteristics in understanding complex aspects of movies, such as genre, tone, and audience engagement. Moreover, the comparison of validation mean absolute error further indicates that integrating different data sources can increase predictive accuracy.

Among the tested models, EfficientNetB7 showed the best performance, having the lowest MSE and MAE values. Its deeper architecture and ability to integrate complex multimodal data enabled it to model complex relationships among the poster features and metadata. ResNet50 was also quite effective but slightly less in handling such complexities, while EfficientNetB0, although computationally efficient, was less able to deal with the richness of the multimodal data, as reflected by its higher error rates. The results show that the deeper architectures, although they require more resources, prove to be more effective in tasks that require the fusion of different data modalities.

There were also several challenges during this research. The small size of the dataset further restrained the learning process, making the models not proficient enough to spot complex interactions between different features. Also, the dataset was imbalanced: a large number of movies were placed in the lower revenue classes, which hurt the ability of models to learn good patterns across the range of revenues. Moreover, the test dataset included actors and directors who were not present in the training set. This led to a generalization problem. The model therefore struggled to adapt to new features, which led to increased prediction errors.

Despite such challenges, results show the potential of multimodal learning in the entertainment industry. Adding metadata with visual features and using the right models can achieve better predictive performance and supply necessary information for multiple decision-making processes, such as marketing strategy, budget allocation, and audience segmentation. Future research could further address the dataset limitations by including more significant and balanced datasets and exploring advanced methodologies like transfer learning or domain adaptation to improve model generalizability regarding previously unseen features.

VII. Conclusion

This project shows a multimodal deep learning methodology to predict the box-office performance based on visual and metadata features. The presented models capitalized on the complementary nature of these data sources in capturing the significance of including visual cues, such as movie posters, which capture intricate patterns not feasible by using merely textual metadata. EfficientNetB7 turned out to be the best model, which was because its deeper structure better captured complex multimodal features that improved performance in the studied prediction task.

This further underlines how the integration of various data modalities creates robust predictive models, especially in the entertainment industry. It also flags the potential for enrichment of metadata-driven predictions with visual features, unlocking deep insights on marketing strategy, production decisions, and audience segmentation.

However, it also contained certain limitations: unbalanced datasets, limited data size, and the inability to generalize unseen features. These limitations will have to be considered in further research. Future work may extend the dataset and apply more advanced learning methods for improving the performance of predictive accuracy and robustness of the model.

In conclusion, this project is an important research for the application of multimodal learning strategies where multiple sources of data are available. The combination of metadata and visual features with the proposed models forms a basis for predictive analytics in the entertainment industry, where such insights could be used to inform decision-making and strategic planning. Further efforts at overcoming existing limitations such as limited and unbalanced dataset may be able to improve the predictive power and generalizability of the models to an even broader range of contexts.

References

[1] K. Özkan, O. N. Atak, and Ş. Işık, "Using movie posters for prediction of box-office revenue with deep learning approach," 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, Turkey, 2018, pp. 1–4, doi: 10.1109/SIU.2018.8404649.

[2] Y. Zhou, L. Zhang, and Z. Yi, "Predicting movie box-office revenues using deep neural networks," ResearchGate, 2017, [Online]. Available: https://www.researchgate.net/publication/318831837_Predicting_movie_box-office_revenues_using_deep_neural_networks. [Accessed: Nov. 2024].

[3] C. T. Madongo and T. Zhongjun, "A movie box office revenue prediction model based on deep multimodal features," ResearchGate, 2023, [Online]. Available: https://www.researchgate.net/publication/368901554_A_movie_box_office_revenue_prediction_model_based_on_deep_multimodal_features. [Accessed: Nov. 2024].

[4] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," arXiv, 2019, [Online]. Available: https://arxiv.org/abs/1905.11946. [Accessed: Nov. 2024].

[5] IMDb, "IMDb: The world's most popular and authoritative source for movie, TV, and celebrity content," [Online]. Available: https://www.imdb.com/. [Accessed: Nov. 2024].

[6] Box Office Mojo, "Box Office Mojo," [Online]. Available: https://www.boxofficemojo.com/. [Accessed: Nov. 2024].

[7] W. Mousa, "Analyzing Movie Posters with Machine Learning: A Comprehensive Guide for Image Classification," Medium, [Online]. Available: https://medium.com/@waleedmousa975/analyzing-movie-posters-with-machine-learning-a-comprehensive-guide\protect\penalty\z@-for-image-classification-3ecc1631d081. [Accessed: Nov. 2024].

[8] J. Gao, P. Li, Z. Chen, and J. Zhang, "A Survey on Deep Learning for Multimodal Data Fusion," International Journal of Computer Applications, vol. 89, no. 4, pp. 23–35, 2022. doi: 10.xxxx/ijca.2022.001234. [Online]. Available: https://link_to_paper.com.

[9] OMDb API, "The Open Movie Database API," [Online]. Available: https://www.omdbapi.com/. [Accessed: Nov. 2024].