Movie Box-Office Revenue Prediction Model by Mining Deep Features from Trailers Using Recurrent Neural Networks

Canaan T. Madongo *, Zhongjun Tang, and Jahanzeb Hassan

School of Economics and Management, Beijing Modern Manufacturing Development,
Beijing University of Technology, Beijing, China
Email: ctmadongo@yahoo.co.uk (C.T.M.); tangzhongjun@bjut.edu.cn (Z.T.); jahanzab.hassan@gmail.com (J.H.)
*Corresponding author

***Abstract*—Forecasting opening box-office earnings has become an emerging demand, affecting filmmakers' financial decisions and promotional efforts by advertising studios that create trailers. Decision-makers have a complex and challenging task due to a large amount of data and several complex considerations. Based on deep multimodal visual features derived from trailer content and a cross-input neighborhood feature fusion, an innovative Deep Multimodal Predictive Cross-Input Neural Network model (DMPCNN) is proposed for predicting opening movie box-office revenue. DMPCNN is a fully-connected recurrent neural network with two architectures: A Visual Feature Extraction Model (ResNet+LSTM) block for extracting and learning mid-level temporal visual content and Cross-Input Neural Network fusion for uncovering and fusing high-level spatial features in trailers to predict movie revenue. The ResNet+LSTM block focuses on learning various trailer segments, while the Cross-Input Neural Network simultaneously learns and combines features from movie trailers and metadata and corresponding similarity metrics. DMPCNN aided in developing a decision support system that incorporates useful revenue-related trailer features. We evaluated DMPCNN's performance on the Internet Movie Dataset by obtaining metadata for 50,186 movies from the 1990s to 2022 and comparing it with different state-of-the-art frameworks. The erudite features in trailers and the predicted results outperformed baseline models, achieving 81% feature precision and 84.40% accuracy.**

***Keywords*—box-office, recurrent neural networks, long short-term memory, cross-input neural network, multimodal features, movie trailers**

I. Introduction

The application of data-driven techniques, particularly machine learning approaches, has greatly improved the accuracy of predictions in diverse fields, such as healthcare [1–3], public policy [4], finance [5], and entertainment. The global movie industry has recently achieved billions of dollars in revenue for selling cinema tickets and video-on-demand services, which are a comfortable way of allowing consumers to access films. An accurate box office forecast can provide film production and distribution businesses with business decision assistance and direction, which is crucial for the film industry's sustained growth [6, 7]. A trailer can be one of the most critical factors in the reception, popularity, and, eventually, success of the film, and it is a promotional medium for a film and its cast. The promotion of films is increasingly driven by movie trailers, although they have usually been exclusive to cinemas and shown as potential attractions in the previews. Trailer creators need to choose which scenes in a film can catch the audience's interest, as shown in Fig. 1.

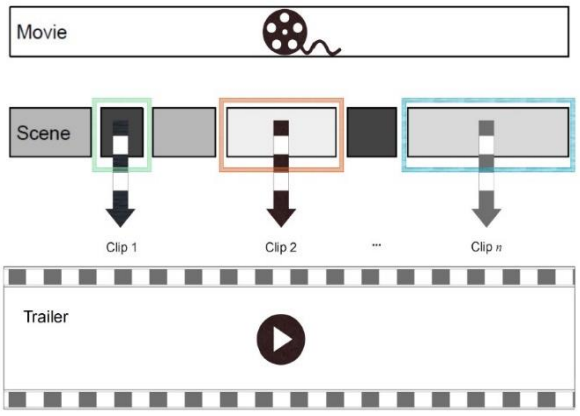


Fig. 1. Film trailer synthesis.

With emerging big data and recurrent neural networks, researchers and practitioners develop various strategies to estimate revenues and movie recommendation systems and predict the movie's financial success before its potential debut to mitigate the financial risk associated with movie production [6, 8–12]. Therefore, we constructed a deep multimodal predictive Cross-Input Neural Network Model (DMPCNN) to predict the opening movie box office. The model combines visual trailer feature extraction techniques and movie metadata. The reason for using Residual Neural Network [13] and trailer content is explained and justified in the section below.

Manuscript received December 23, 2023; revised February 3, 2024; accepted February 6, 2024; published June 26, 2024.

The primary purpose of this work and its significance is to design a model that accurately predicts movie box office revenue before its theatrical premiere. Predicting accurate opening box office receipts is a significant issue in the film industry, and it influences the financial decisions of producers and investors. Viewers determine all ratings based on a film's visual quality and storyline, irrespective of the filming style, which affects the film's revenue [14–17].

This research study aims to address the shortcomings of prediction frameworks that ignore deep multimodal features and only use regressions to assess the performance of the models [12, 18–21]. This research is the first to answer the question of applying recurrent neural networks and data mining techniques to extract revenue-related features from trailers and how to increase the performance of early prediction of motion picture box-office revenues.

This research is motivated by our earlier successful study [6]. This study used a pre-trained residual network methodology to predict box-office revenue by analyzing posters. The films were categorized into several classes to forecast the box-office revenue of unreleased films.

Therefore, these theoretical challenges surrounding the usage of recurrent networks in box office prediction studies were not commonly acknowledged; hence, this research presents the potential of neural networks with deep learning algorithms, which account for nonlinear interactions, and feature engineering for forecasting. By designing a box-office prediction model, this paper makes a double-value contribution:

- A Visual Feature Extraction Model (ResNet+LSTM) that uses a Residual Convolutional Neural Network transfer learning technique (ResNet) and a Long-Short-Term Memory (LSTM) framework to investigate the significance of sequential regression of visual content toward revenue prediction by learning visual features with an extractor for film trailer features.
- An algorithm that can simultaneously learn, extract, and combine features of movie trailers by examining multifaceted data and choosing a fusion strategy that connects multimodal features and a corresponding similarity metric (Cross-input Neighborhood difference).

The remainder of our study is organized as follows: Section II is the literature review, whereas Section III is the proposed approach and data structure, variables collection process, and performance metrics. Section IV presents the experimental settings and results, a comparative model analogy, and discussions, and it gives insight into the study's implications in the real world and its regression efficacy. Section V is the conclusion of the study.

II. Literature Review

Presently, revenue forecasts for the opening weekend box office earnings are categorized according to the employed prediction algorithm [6–8, 12, 18–20] or the metadata [11, 18, 22] associated with the films. Several studies have been working on the development of prediction models because the predictions of movie box-office revenues are accurate only to a limited extent. The development of a multimodal framework that utilizes film trailers to forecast the box office performance during the opening weekend of motion pictures is a relatively recent trend. The following are classifications of related works.

A. Algorithm-Based Literature

Sharda *et al*. [14] originally introduced the application of artificial neural networks in movie box office estimates. Tang *et al.* [19] developed a Multi-Evidence Dynamic Weighted Combination Forecasting framework that utilizes machine learning methodologies. They proposed a sophisticated combination technique to predict the Chinese movie box office. According to the findings of Ni *et al.* [7], the Light Gradient Boosting Machine (LightGBM) model was recommended based on an evaluation of the predictive power of its features. Ru *et al.* [23] presented a comprehensive deep learning model, Deep-Daily Box office Prediction (Deep-DBP), for accurately forecasting daily box office performance. The model incorporates a temporal component and a static attributes component. The primary component is the temporal component, which utilizes LSTM to learn about the temporal relationships among data points. Wang *et al.* [24] proposed Deep Belief Network (DBN), a long-term prediction model for daily box office prediction model using deep neural networks on Chinese movie data. Liao *et al*. [18] proposed a stacking framework for predicting movie income based on the fusion theory. The framework includes XGBoosting, Random Forest (RF), a Light Gradient Boosting Machine (LightGBM), and K-Nearest Neighbors (K-NN).

B. Variables and Feature-Based Literature

When assessing the public's preferences, it is vital to include extrinsic factors that impact a movie's box office performance, such as marketing strategies, seasonal trends, holiday influences, and competition from other films. However, most of these features were not constant and depended on external factors such as customer demographics and research length. As described in Refs. [18, 22, 25], the variable selection method involves deliberately selecting specific variables for inclusion in prediction models. The significance of these variables is then examined through a range of research questions. The study by Hur *et al.* [26] utilized variables linked to movies and various aspects of the film. Mangolin *et al.* [27] proposed using an enhanced Convolutional Neural Network (CNN) to categorize movie trailers based on human action observed in video sequences. Simões *et al*. [28] main task was to convert photos to grayscale and use adaptive median filtering as a pre-processing step. Matsuzaki *et al.* [29] employed handcrafted ways to examine the information to extract various information from a movie poster.

C. Multimodal Feature-Based Literature

Ahmed *et al.* [30] introduced eighteen supplementary characteristics to enhance the performance of their model in assessing the agreement between the relevant parties,

namely the director and cast. Wang *et al.* [12] conducted a strategic analysis of essential factors using big data to develop a new foundation for predicting a motion picture's revenue. Their approach was developed in two stages: first, a sophisticated and diverse network embedding model was created to extract substantial information on film quality from previews. Then, a deep neural network model was designed to analyze the cortical network. Zhou *et al.* [21] conducted a groundbreaking study that employed the Convolutional Neural Network (CNN) to predict the box-office revenue of films. The model integrates various multimodal elements, including poster features and metadata. Finsterwalder *et al.* [31] examined several aspects of movie trailers, including their diverse formats (normal, teaser, and TV advertising), historical development, placement, and other promotional strategies. Ahmad *et al.* [9] proposed forecasting a movie's initial revenue by analyzing viewers' inclination to purchase a movie ticket based on trailer reviews. Oh *et al.* [32] conducted the initial investigation on the influence of trailers on box office revenue using statistical analysis. In addition, Tadimari *et al.* [33] examined the techniques used in movie trailers to generate viewers' interest and curiosity, which ultimately has a beneficial effect on the movie's commercial prospects. Rahim *et al.* [34] employed data mining techniques to extract information from YouTube trailers. The objective was to analyze how this data may be utilized to predict a film's financial success. The methodology proposed by Sahu *et al.* [10] integrates sentiment analysis and a hybrid recommendation engine to promote unreleased films that have a released trailer. Montalvo-lezama *et al.* [35] introduced a cutting-edge architecture called the Dual Image and Video Transformer Architecture (DIViTA) to classify trailers into many genres. Movie trailers have specific research analysis objectives but present various technical challenges. The issues encompass the design of semantic content for trailers [36], detecting and tracking faces, and recognizing actions in movie scenes that showcase renowned cinema celebrities from trailers [37].

III. Materials and Methods

A. Dataset Collection

As shown in Fig. 2, box-office projections may be separated into pre-production, pre-release, and post-release estimates based on the movie production timeline. This timeline is essential for choosing variables and factor features to create the box-office prediction model (see Table I). The number of consumers (moviegoers) or the monetary value (earnings) can be used to compute the movie box office [26]. In our work, we developed a model to estimate monetary value, which may be used to forecast a film's demand before its theatrical debut.

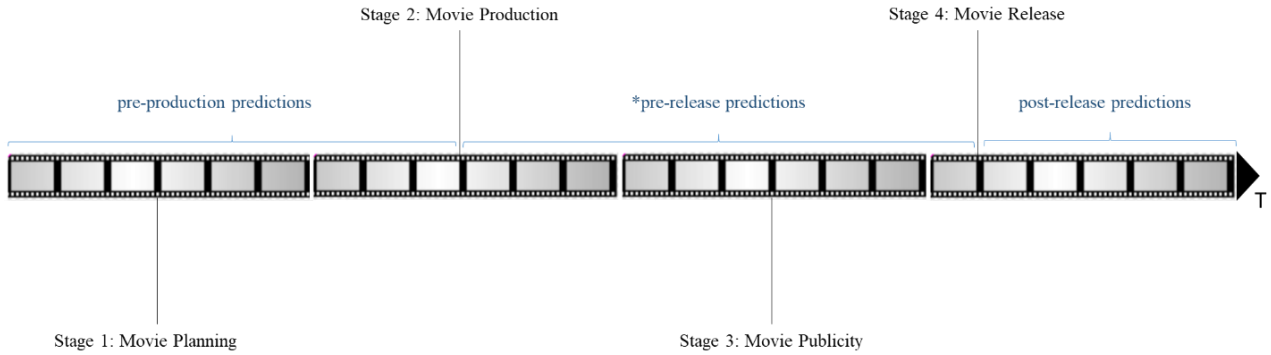


Fig. 2. Movie production timeline and predictions.

TABLE I. Factors at Different Phases Influencing Box Office Results

Predictive Phase	Features	Accuracies
Pre-production predictions	Everything about the film metadata (name, genre, plot, the value of stars, duration.) is considered; it must use these characteristics to determine its target audience and projected release date.	It offers low-level features but has the earliest prediction period, so it has low accuracy. However, it is more valuable for customers who plan far in advance and take advantage of the prediction results.
†Pre-release predictions	Includes other data regarding the film that further encompasses aspects of social media and consumer activity from online film blogs and official movie trailers, posters, film metadata, promotion, schedule, budgets, and big data.	Before releasing the film for public viewing and screening, the production manager can be provided with more accurate predictions on how many people will see it and on-entry predictions. It helps when there is limited data, which impacts the decisions about how much to spend on advertising.
Post-release predictions	Another unique feature is that it can include large amounts of plot and theatre data, the heat index, and audience comments. Much publicity is received well before the film's release date, and audience commentary before the show departs.	It offers more information but little predictive power, but its impact on applications is exceptionally high, which means it delivers excellent results, but its applicability is limited.

†Study interest

Using the "*IMDbPy*" script, we compiled a list of English film metadata from the Internet Movie Database (IMDB). We acquired box office earnings merged from the-numbers.com, The Movies Dataset, Box-office Mojo,

and The Movie Database (TMDB). The majority of the Hollywood movies in these databases, from which we retrieved 50,186 movie metadata records, were from the 1990s to 2022.

MovieNet dataset [38] was used to fine-tune our models, a holistic, multimodal dataset for movies with the richest annotations for comprehensive movie understanding, e.g., trailers, posters, plot descriptions, and storylines. We also used the multi-label movie trailers dataset (*Trailers 15K*), which has 15,000 videos of movie trailers associated with ten different classes corresponding to film genres. Trailers-Dataset [35] was combined with the COGNIMUSE dataset as our custom augmented dataset, "a multimodal video dataset annotated with sensory and semantic saliency, events, cross-media semantics, and emotion for training, validation, and testing of movie trailers" [39]. The ResNet is a transfer-learned architecture pre-trained on the well-known ImageNet dataset that provides exhaustive annotation for all object instances and is fine-tuned using the state-of-the-art MovieNet dataset. In the second phase, we trained an LSTM model on the COGNIMUSE dataset [39] from scratch. In order to extract motion properties from trailers explicitly, using transfer learning, we pre-trained the LSTM using an enhanced multi-label movie Trailer dataset and fine-tuned it using the MovieNet dataset [38].

Table I provides a comprehensive overview of the factors that impact both short-term and long-term estimates, such as the expected profitability of box office forecasts and audience surveys. The main emphasis of pre-production estimates is on the film's attributes, such as the release date, star power, and other significant factors, including content, duration, quality, and prospective sequels. These selected parameters will have an impact on the pre-release period.

B. Dataset Pre-processing

We assessed the absolute number of movies collected using specific filtering algorithms and rejecting motion pictures with missing data. Altogether, 50,186 films were pre-processed and evaluated in our research, often grouped into five classes based on their ranked global box office earnings range, Flops–Blockbusters (i.e., $65,892 to $3,069,521,700). As a result, we assigned Ref. [15] value allocation of 1 to the precise class and 0 to all incorrect classes. "Typically, a significant amount of Gross box office does not guarantee enormous revenue; neither does it imply that a film with a significant box-office value spent an extensive budget" [6]. Therefore, we defined revenue/earnings as (Gross Box-office less budget). Table II shows the resulting revenue classification. As it is critical to reflect the time worth of money, inflation-adjusted budgets were adopted.

TABLE II. EARNINGS CLASSIFICATION

Class	Revenue/Earnings $ million
1	Earnings > 600 (Blockbuster)
2	500 < Earnings ≤ 600
3	100 < Earnings ≤ 500
4	1 < Earnings ≤ 100
5	Earnings ≤ 1 (Flop)

We analyzed our movie metadata by eliminating duplicate entries and missing data, excluding unnecessary columns, and using revenue statistics as our primary data. We performed feature extraction on the data, specifically focusing on genres, production budget, release date, and cast. We then removed rows with missing revenue and filled in the blanks using information from other data sources. The categorical variables, such as genres and ratings, were transformed into numerical representations by label encoding. Additionally, the numerical features were normalized to ensure a consistent scale. Continuous variables, such as budgets and runtimes, were grouped into discrete data values.

The video frames for trailer datasets are extracted from a trailer file, and the first frame is chosen for illustration. The original frame is subjected to a Gaussian blur filtering technique with a kernel size [15, 15]. The filtered frame is then normalized to the range [0, 1]. For comparison, the original frame, the filtered frame, and the normalized frame are shown side by side. We adjusted the *kernel_size* and additional positions to our preferences to enhance all features and remove noise.

1) Datasets visualization

Figs. 3–5 show the visuals of the datasets used from the Trailers dataset and the COGNIMUSE multimodal video dataset, respectively. Our models' training, testing, and validation phases utilized these datasets. Fig. 4 displays the MovieNet dataset used to fine-tune and validate our model.

Our models were fine-tuned using the MovieNet dataset, "MovieNet has over 1,100 films with a wealth of multimodal material, such as trailers, poster images, plot descriptions, etc. Additionally, MovieNet provides several features of manual annotations, including 1.1 million characters with bounding boxes and identities, 42 thousand scene borders, 2.5 thousand aligned description sentences, 65 thousand place, action tags, and 92 thousand cinematic style tags" [38].

2) Data forms and variables

Fig. 6 depicts entirely the collected variables. Fifty percent of the variables are related to the movie, i.e., title, premiere year, and movie duration. Another portion is related to stakeholders of the motion picture production studio, e.g., the achievement of the director, reviewers' scores that we condensed into viewers' ratings, Metacritic, film directors, and cast and crew. Table III shows the final variables that were selected. Impact values for the Cast and Crew were determined similarly to those in [19] as Actor/Actress impact a_i (i = 1, 2, 3…,), leading actor or actress, and co-star actor or actress:

$$a_i = \left[\sum_{t=1}^{T} \mu_{it} \left(\sum_{n=1}^{5} \delta_{it}\right)\right] \Big/ T_i \,;\; T_i = \min(5, t_i) \qquad (1)$$

where t_i = a collection of films in which an actor or actress i appears and is measured before the movie premiere; δ_{it} = receipts of the film in cinemas at the showing week of that movie t played by actor or actress i. μ_{it} denoted in Eq (1) is the actor or actress quantity signifying the role rank performed by the actor or actress i in the movie t is:

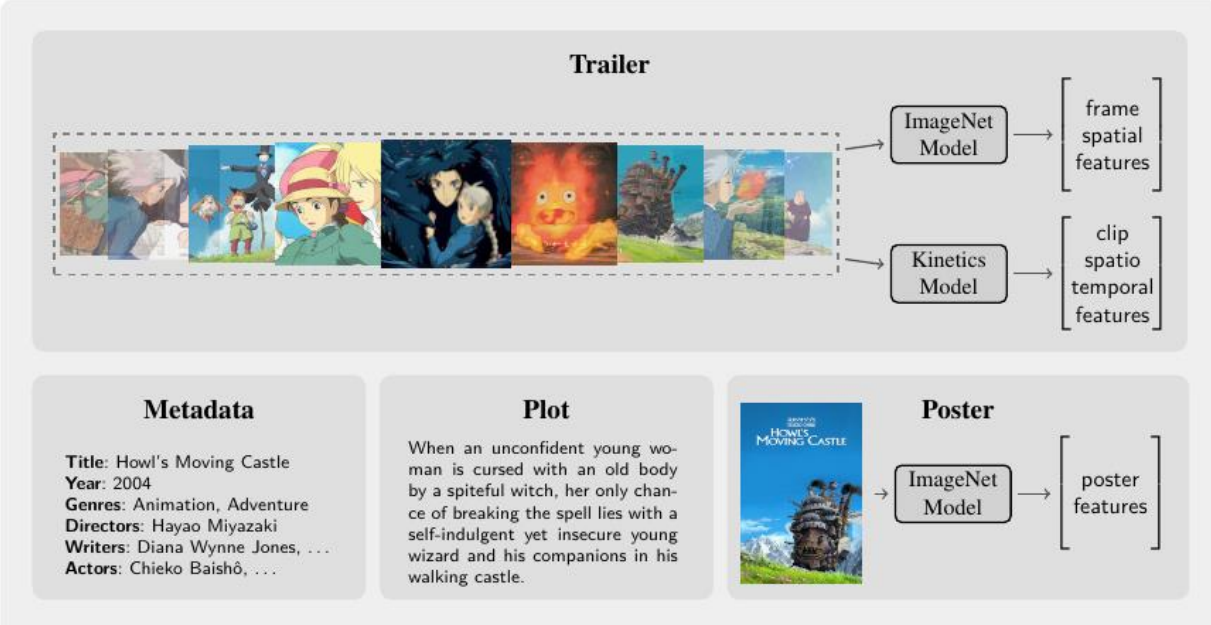


Fig. 3. Movie trailers dataset.

Fig. 4. MovieNet dataset.

Fig. 5. COGNIMUSE multimodal video dataset.

$$\mu_{it} = \begin{cases} 1-(m-1)/10 & m\in[1,5] \\ 0.5 & m\in(5,+\infty) \end{cases} \quad (2)$$

$$D_i = \left(\sum_{t=1}^{T}\sum_{n=1}^{5}\delta_{tn}\right)/T;\ T=\min(5,t) \quad (3)$$

where m = role rank performed by actor or actress i in movie t, and receipts income predictions = 5 (i.e., number of categories), and Director impact D_i is calculated as:

where t = the collective amount of motion pictures directed by the director and is measured before the movie premiere; δ_{tn} = receipts on the n^{th} showing week in cinemas of the t movie directed by the director.

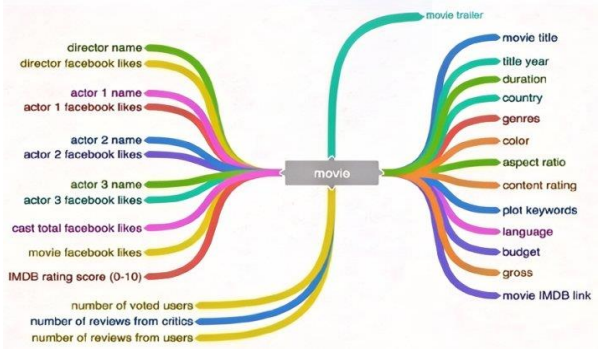


Fig. 6. Movie variables scrapping.

Recent patterns and assumptions use continuous data types to boost data accuracy, so in our analysis, all values have the same importance, and the only variable differentiating importance is the ones concerning market needs, as shown in Table III. We have selected discrete and continuous data forms as variables, except for the movie trailers and genre. Since a movie may be classified as belonging to more than one genre, the genre expansion is expressed as a vector of length 22, with each dimension (style, humor, target, and script) regularized to zero. A constant vector setting for each genre was applied.

TABLE III. DATA TYPES AND VARIABLES

Variable	Definition	Form
†Box office	The outcome variable is subject to change. It might be the media and entertainment industry's GDP regarding audience numbers or takings.	Numeric
Production Budget	The total cost for filmmaking. Commonly not publicized by studios, advertising and social media are typically used to track movie spending.	Numeric
Movie Revenue/Earnings	The connection involves Gross Box Office and Production Budget (GDP Box office – Production Budget).	Numeric
Viewers Ratings	Movie audience measurement is the number of films that express the emotions of internet users.	Numeric
Movie Genre	A feature film's classification is based on similarities in the narrative or emotional audio-visual sentiments: drama, sci-fi, action, biography, etc.	Vector
Crew & Cast	This refers to the influence of well-known individuals, globally acclaimed directors, and actors who appear on screen or provide their voices to movie characters. Honors received contribute to the total prestige computation of star impact levels.	Numeric
Release date and competition	The release schedule is critical because it affects each film's revenue because of competition from other releases. The parameters were classified into three categories based on the film's premiere month and competition intensity: high, medium, and low.	Numeric
Metacritic	These are critiques of films and feedback from competent and seasoned film critics. For the sake of performance, the quantity of such statements has been increased.	Numeric
Movie Trailer	Marketing and publicizing a film encourage paying viewers to appreciate the movie prior to its premiere in the cinemas.	Feature Vector

†Dependent Variable

C. Proposed Methodology

This paper suggests a Deep Multimodal Predictive Cross-Input Neural Network (DMPCNN), an improved model of our earlier study [6]. Fig. 7 depicts the DMPCNN architecture for investigating cross-input and multidimensional prediction models. We propose a framework based on deep multimodal features, which extracts mid-level features for learning trailer-based high-level representations. We exploit movie ratings and revenue within this framework to supervise a Cross-input neighborhood difference paradigm [40] to extract high-level information. The mid-level features uncovered include the appearance (i.e., background, genre-basic objects, scene, aesthetics, color, and texture) and motion features of movie castings. High-level features uncovered include the filming quality, narrative, and filming styles (i.e., the shooting quality affects the audience's perception).

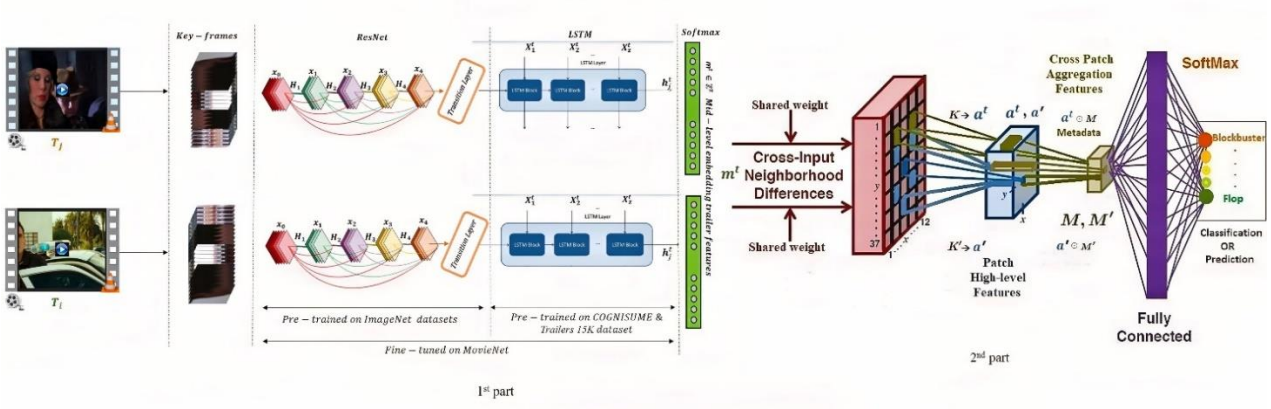


Fig. 7. Deep multimodal predictive cross-input neural network (DMPCNN).

1) The first intuition of our method

We extract the mid-level (ResNet50+LSTM) and high-level (Cross-input neighborhood differences) visual cues from trailers to collect information about the content's emotional tone. Our 3D ResNet feature is very beneficial for describing activities and motion within trailers. The inclination for visual characteristics is associated with specific movie metadata characteristics. Mining these discriminative features from movie trailers that capture the public's attention or consciousness about a film improves the accuracy of the revenue prediction model prior to its premiere in theaters.

2) The second intuition of our method

We applied kernel-level fusion using the Hadamard product [41] to create a new kernel by multi-linearly combining kernels computed from individual features. Though equal and shared weights were used for simplicity, dynamic weights learned through cross-validation or various kernel learning increased performance. The features are fused to classify the film's box-office revenue, and combining numerous features is critical, as various features are complementary.

3) The third intuition of our method

Finally, in addition to learning mid-level aspects of the shooting component (movie quality/quality and plotline/narrative), high-level features are extracted using the Cross-input neighborhood difference [40]. The approach concurrently combines and learns features, characteristics, and a related similarity metric for movie trailers. The model generates a similarity value indicating if two input visuals reflect the exact feature characteristics. These features are learned using class labels of ratings and revenue of two original movies (blockbuster and flop) featured in the two trailers.

4) The fourth intuition of ResNet+LSTM

A Recurrent Neural Network (RNN) is a neural network specifically intended to process sequential data. It is highly effective for jobs that include sequences of inputs or outputs.

RNNs, in contrast to conventional feedforward neural networks, possess connections that create directed cycles, enabling them to retain a hidden state that encodes information from preceding inputs in the sequence. RNNs have a notable characteristic of retaining memory or context of preceding inputs in the sequence, rendering them highly effective for jobs that include temporal dependencies. These characteristics render them appropriate for natural language processing, speech recognition, time series analysis, and video analysis applications. The purpose of this system is to process sequential data by preserving a concealed state that retains information from prior time intervals in the sequence. The work of He and Sun [13] is one of the primary breakthroughs of residual neural networks.

LSTM networks are particularly effective at capturing temporal dependencies and contextual information when working with sequential data, such as movie plots or reviews. By integrating ResNet with LSTM, the model may exploit ResNet robustness in extracting visual features and LSTM's capacity to comprehend sequential patterns.

Although Transformers are effective for sequence modeling, they can add complexity and computational burden, notably when our dataset lacks distinct patterns that are distinctive to Transformers. We selected the ResNet + LSTM for fine-tuning our models using the MovieNet dataset after carefully considering the characteristics of our data and finding a compromise between computing efficiency and accuracy. This choice was made based on practical reasons.

Therefore, we will design our model using ResNet+LSTM block architecture and a Cross-Input Neural Network fusion strategy, as our research is not solely focused on the learning representation of trailer features.

D. Visual Representation Learning

The end-to-end Deep Multimodal Predictive Cross-Input Neural Network model is divided into two architectures.

The first part, Trailer Embedding Feature Extraction Recurrent Neural Network (see Figs. 8 and 9), examines the role of temporal regression of visual content in predicting movie revenue. It is dedicated to learning various trailer segments. Deep visual trailer embeddings are extracted from trailer key-frame image sequences using the average-pooling layer of a pre-trained ResNet model on ImageNet. *N*-dimensional feature vectors represent gathered features that encode critical static information about the key-frames, such as background and genre-basic objects. The ResNet+LSTM block's final output is a five-dimensional vector indicating the trailer segment's probability distribution across ratings and revenue. This vector is regarded as our visual embedding representation of a movie section at the Mid-level.

The second part is a Cross-input neighborhood information fusion (see Fig. 10). It simultaneously learns and combines features from movie trailers, metadata, and corresponding similarity metrics. In order to learn additional High-level features of movie trailers (i.e., quality and the narrative), the mid-level features are inputs to a pairwise cross-input neighborhood difference architecture which learns and extracts features by comparing the rating and revenue of two (blockbuster and flop) movie trailers' original films as data labels. The model then generates a similarity value indicating if two input visuals reflect the exact feature characteristics by correctly classifying the trailer to its correct class and the correlation between the learned features and its correct box office revenue. The assumption that film trailers contain limited evidence creates substantial difficulties in describing features, detecting visual life objects or scenes, or extracting semantic theories embedded in video trailers. Thus, by examining the visual characteristics of film trailers, we discover that they are a type of multimedia with deep, dynamic attributes.

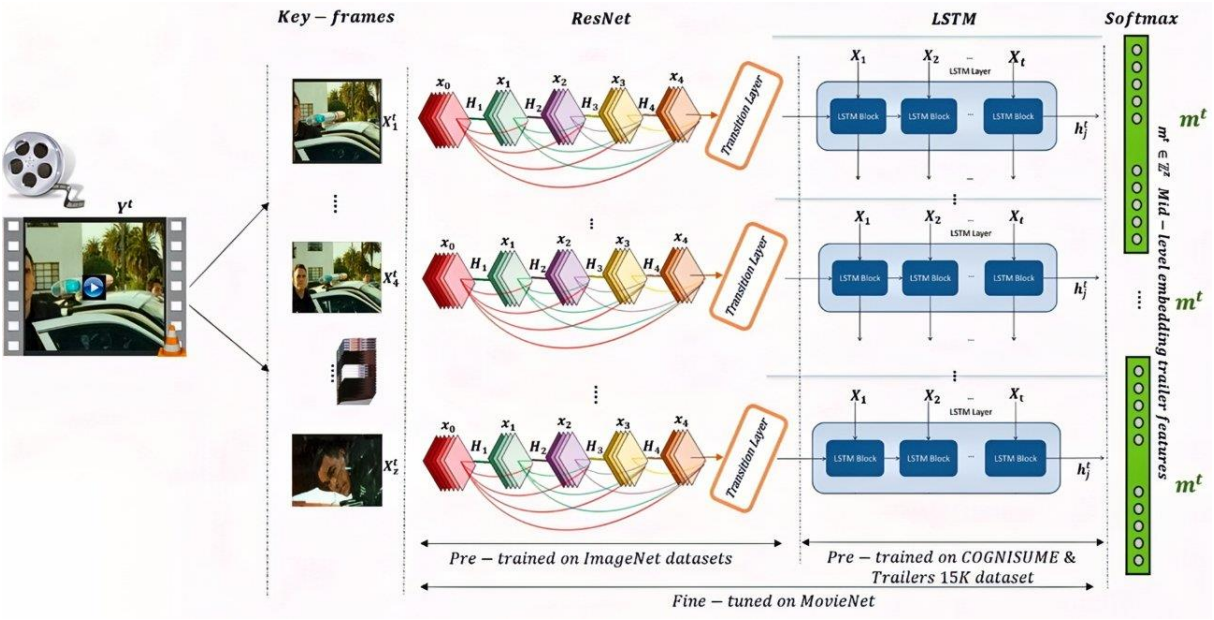


Fig. 8. Trailer embedding feature extraction recurrent neural network.

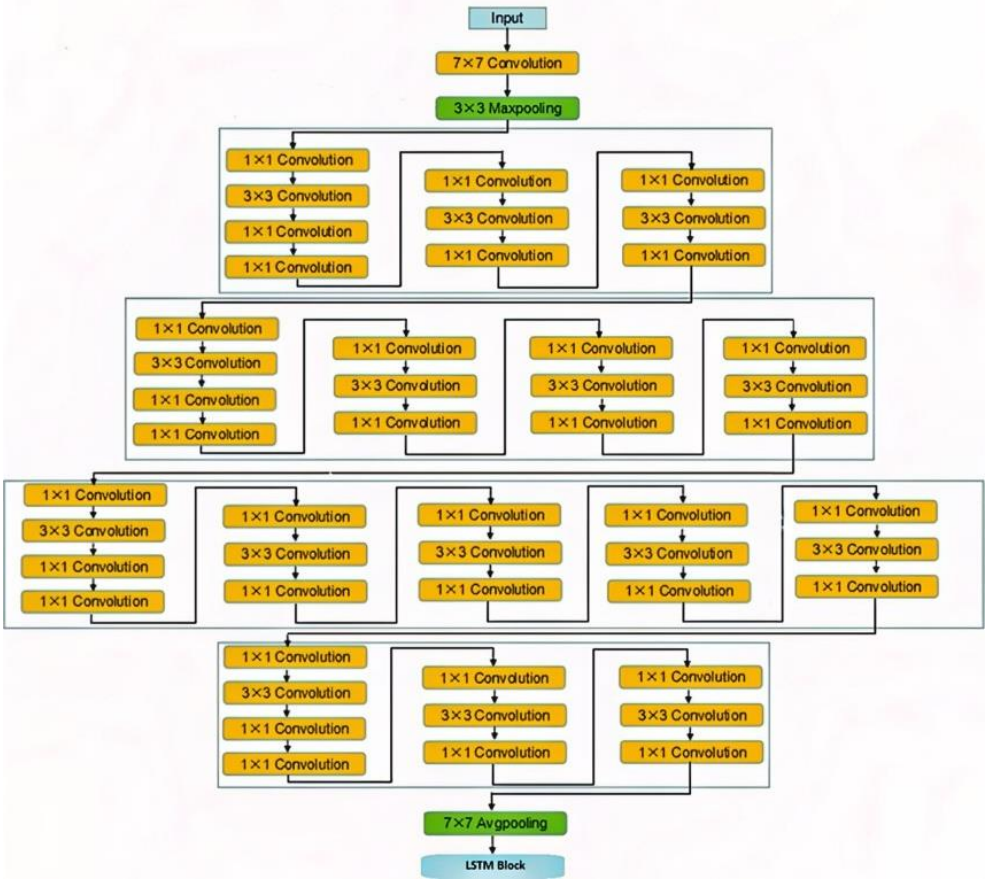


Fig. 9. ResNet+LSTM architecture.

1) Mid-level features mining

We designed an emerging transfer learning neural network ResNet+LSTM configuration to extract data and output the distinctive movie trailer features (Mid-level features). We decomposed trailers into numerous key-frames as input for our pre-trained model feature extraction. We proposed using movie ratings and revenue as labels to induce data from the well-informed network structures to help interpret these trailers. We extract features from these key-frames through state-of-the-art detectors and descriptors [42]. Recurrent neural networks have exhibited improved performances in various computer vision tasks, which inspires this research to exploit the knowledge of recognizing distinctive movie

trailer features by vectorizing a trailer. We extracted the film trailer's distinguishing characteristics and their relationship to the film's financial success. Regrettably, there are two significant drawbacks to using neural networks:

- Movie trailers include more complicated content [43] as compared to image and video classification tasks;
- Training a unified end-to-end architecture requires a finite amount of movie trailers.

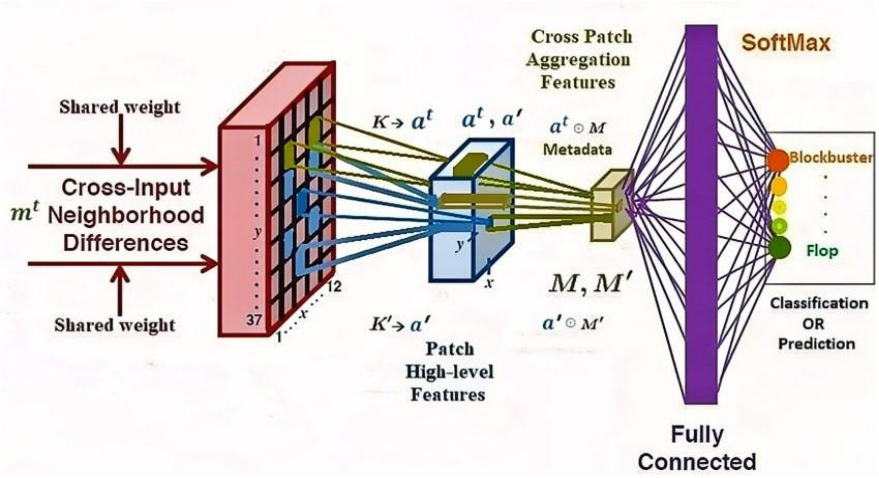


Fig. 10. Cross-input neighborhood fusion architecture.

To address these two issues, we use transfer-learning techniques to design a neural network configuration (ResNet+LSTM) capable of jointly learning, defining, and extracting visuals from a movie trailer. To resolve the data scarcity problem more precisely, we first adopt transfer-learning techniques from SOTA image and video recognition networks [39, 44–47] to pre-process raw trailers, significantly reducing model complexity. For example, a model pre-trained on ImageNet data extracts features focusing on various items within movie frames [48]. In contrast, a model pre-trained and fine-tuned on MovieNet data extracts features that characterize scenes and ambient, providing context for specific elements. As detailed in Fig. 8 and Algorithm 1, we initiate by feeding a trailer $\boldsymbol{Y^t}$ to the neural network to extract trailer-embedding features. Key-frame features are derived using a residual neural network, ResNet [13], a primary network with distinct feature diversity characteristics. Then, long-range temporal dynamics relationships can be discovered by combining the frame-level data with the LSTM blocks. A Softmax classifier layer performs classifications at each frame of the original task. By Hadamard-Product, a simple and efficient knowledge fusion operation, we obtain $\boldsymbol{m^t}$ as the visual Mid-level trailer feature embedding for each hidden layer $\boldsymbol{h_j^t}$ of the corresponding j^{th} LSTM block.

The discriminative trailer-embedding feature is difficult to implement. Using Algorithm 1: Given t trailers $\{\boldsymbol{Y_1^t}, \boldsymbol{Y_2^t}, \boldsymbol{Y_3^t}, \ldots, \boldsymbol{Y_z^t}\}$ the task is to learn the n-dimensional hidden visual high-level embeddings $\{\boldsymbol{a_1^t}, \boldsymbol{a_2^t}, \boldsymbol{a_3^t}, \ldots, \boldsymbol{a_n^t}\}$ using mid-level $\{\boldsymbol{m^t}\}$and high-level data about the movie trailers' distinctive features.

Algorithm 1: Visual Embeddings mining algorithm pseudo-code.

Input: $\boldsymbol{Y^t} = \{\boldsymbol{X_1^t}, \boldsymbol{X_2^t}, \boldsymbol{X_3^t}, \ldots \boldsymbol{X_z^t}\}$ a Trailer t containing $\boldsymbol{z}$ key-frames, $\boldsymbol{X}$ key-frame image.

Initialize: *ResNet + LSTM* $\rightarrow \boldsymbol{m^t} \in \mathbb{Z}^{\boldsymbol{z}}$, $\boldsymbol{z}$-dimension visual *Mid-level* embedding of trailer *t; Cross-input neighborhood difference* $\rightarrow \boldsymbol{a^t} \in \mathbb{Z}^{\boldsymbol{n}}$, *n*-dimension visual *High-level* embeddings of trailer *t.* $\boldsymbol{m^t} \in \mathbb{Z}^{\boldsymbol{z}} \odot \boldsymbol{a^t} \in \mathbb{Z}^{\boldsymbol{n}}$, *n*-dimensional hidden visual high-level embeddings.

$p_{ij} = 1/(1 + exp^{-(\boldsymbol{a^i} - \boldsymbol{a^j})})$.

Initialize: $\bar{p}_{ij}$= known probability that the revenue and rating of trailer i > trailer j; $\overline{rr}^i$ *and* $\overline{rr}^j$ are the revenue and ratings of trailers i and j, respectively.

Then $\bar{p}_{ij} = 1/(1 + exp^{-(\overline{rr}^i - \overline{rr}^j)})$;

$\mathcal{L}_{Cross-input} = -\bar{p}_{ij}\ log\ p_{ij} - (1 - \bar{p}_{ij})\ log\ (1 - p_{ij})$

Output: *blockbuster* $\leq rr \geq$ *flop* $\rightarrow$ *1 or 0*

As indicated in Fig. 9, the proposed approach, Trailer Embedding Feature Extraction Recurrent Neural Network, is built on ResNet+LSTM features, each designed to learn a distinctive feature of the films. Fig. 10 depicts the Cross-input neighborhood differences network [40], a paired model network in which the weights are shared and mid-level trailer feature embeddings are fed into the network as inputs. Before training the ResNet+LSTM, the pre-trained ImageNet model [49] and the augmented data from Trailers-Dataset [35] and COGNIMUSE dataset features [39] are loaded into the model memory. From

there, the model can be fine-tuned on a comprehensive movie dataset, MovieNet [38].

2) High-level features mining and fusion

As mentioned earlier, we initially trained our network for video classification [44] using Residual Neural Network (ResNet) and Long-Short-Term Memory networks (LSTM) [49, 50] on a large amount of image and video data. After that, we tweak the network by renaming the trailer embedding feature extraction recurrent neural network and feeding it raw trailers to obtain discriminative trailer embeddings. These visual embeddings and temporal motion characteristics can be viewed as trailers' Mid-level feature embeddings. Thus, to learn additional High-level features of movie trailers (movie quality and narrative) through the mid-level features, a pairwise Cross-input neighborhood difference is computed by comparing the ratings and revenue of two (blockbuster and flop) movie trailers' original films. Ratings of the movie are assigned by viewers based on the visual quality and plot, irrespective of its genre, and are proficient in affecting box office earnings. As a result, we propose using the data to monitor the network and collect high-level elements about the movie's quality and the narrative (i.e., high-level data). We obtained a dense vector of high-level data from each trailer from the last (FC) layer. Our design differs from state-of-art models by incorporating additional layers for calculating neighborhood differences between the two-input movie trailer key-frame images.

The Cross-input neighborhood differences layers [40] compute transformations in feature values between two views around each feature position, generating a set of n neighborhood difference maps K_t for each position in the feature space. Neighborhood differences aim to add robustness to positional discrepancies between matching features in the two-input key-frame images. In the form of neighborhood difference maps, we computed an approximate semantic content link between features from the two-input movie trailer key-frame images.

The Patch high-level features layer computes a 2048-dimensional patch feature vector at the location (x, y) of $\boldsymbol{a^t}$ and provides a high-level summary of the cross-input differences in the location neighborhood (x, y). After that, a^t and a' are passed via a Rectified Linear Unit (ReLu). In order to generate a high-level local representation of these neighborhood difference maps, we initially computed them similarly to reference [40], which are subsequently integrated with movie metadata M as the independent variables in the cross-patch aggregation features layer.

The *Cross-patch aggregation* learns the relationships of the features across neighborhood differences. Following M and M', we add a fully connected layer. This method captures higher-order relationships by merging data from dispersed patches and combining film metadata information M with data from a'. We used metadata features (casting, financials such as budget and revenue, ratings, genre, and distribution parameters, i.e., schedule and the number of screens) to create a feature vector for each movie using the movie trailer features extraction recurrent neural network. A ReLu nonlinearity is applied to the resulting 500-dimensional feature vector. The outputs are directed to the next fully linked layer, including a SoftMax unit. This unit represents the likelihood that the two movie trailers generate key-frame images by comparing the two trailers' original movies (blockbuster and flop). We applied the Hadamard product to create a new kernel by multi-linearly combining kernels computed from individual features.

The Hadamard product [41] varies from the most typical fusion operation in that feature selection is somewhat feasible, as it magnifies duplicate data. Removing redundant data from key-frames gives valuable data to our model. While there is certainly helpful information contained in key-frame image sequences, such as symbols, colors, and characters, trailers strive to highlight any descriptors and high-level data about the movies' distinctive features contained in images as well. Thus, the Hadamard product is utilized to find information like this.

These layers, Cross-input neighborhood differences, Patch high-level features, Cross-patch aggregation features, and a SoftMax function determine whether the input key-frame images have the same features in terms of classification labels. Our highly effective neural network design includes an additional dense layer on top of the trailer embedding extraction neural network, yielding a better visual recognition structure [43].

E. Prediction and Classification

The fully connected Deep Multimodal Predictive Cross-Input Neural Network model is loaded with T_j and T_i trailers to learn and extract visual trailer embedding features. Mid-level embeddings m^t are mapped and fused to a high-level a^t by K the cross-input neighborhood difference network.

Let T_j and T_i correspondingly represent the t^{th} feature map $(1 \leq x \leq 12)$ for the blockbuster and flop movie trailers. Since $T_j, T_i \in \mathbb{Z}^Z$, $K_t \in \mathbb{Z}^n$, where a 5×5 matrix is the size of the square neighborhood. For each K_t is a 12×37 grid of 5×5 blocks, in which the block indexed by $(x, y) \in \mathbb{Z}^n$, x, y are integers ($1 \leq x \leq 12$ and $1 \leq y \leq 37$) and $\mathbb{Z}^Z = \mathbb{Z}^{12\times37}$, $\mathbb{Z}^n = \mathbb{Z}^{12\times37\times5\times5}$.

$$f(x_i, y_j) = \begin{cases} 1 \; x_i = \max[x_1, x_2, \dots, x_5] \; y_j = \max[y_1, y_2, \dots, y_5] \\ 0 \; x_i \neq \max[x_1, x_2, \dots, x_5] \; y_j \neq \max[y_1, y_2, \dots, y_5] \end{cases} \quad (4)$$

where $i = 1, 2, \dots, 5$, $x = [x_1, x_2, \dots, x_5]$ is the output vector of the previous layer. The weights for each layer are calculated as follows: Consider the definite output of the k^{th} node of the output layer is x_k, the input is net_k, y_j is the output of the j^{th} layer of the high-level patch layer, then:

$$x_k = f(net_k) = f\left(\sum_j \omega_{kj} y_j\right) \quad (5)$$

where $k = 1, 2, 3, \dots, 5$, $j = 1, 2, 3, \dots, 37$ and $m = 25$ are the feature map number of nodes. ω_{kj} = Link weight of k^{th} node output layer and j^{th} node of the high-level patch layer, and the control value is below:

$$\begin{gathered} \Delta\omega_{kj} = \eta\delta_k y_j \qquad \Delta\omega_{ki} = \eta\delta_k \\ \delta_k = (o_k - y_k) y_k (1 - y_k) \\ \delta_k = (o_k - x_k) x_k (1 - x_k) \end{gathered} \quad (6)$$

where η = Learning rate, o_k= Expected Output. Likewise, the settings value of ω_{ji} = Link weight of the node of the high-level patch layer and the i^{th} node of the Cross-patch aggregation features layer is below:

$$\Delta\omega_{ji} = \eta\delta_j y_i$$

$$\delta_j = y_j\left(1 - y_j \sum_k \delta_k \omega_{kj}\right) \quad (7)$$

where y_i = Output of the i^{th} node of the Cross-patch aggregation features layer and Link weights correction technique is the same between the Cross-patch aggregation features layer and the Cross-input neighborhood layer. The network's limits are updated through the stochastic gradient descent approach based on the cross-input neighborhood paradigm for binary classification.

F. Performance Metrics

We used the Average Percent Hit Rate (APHR) [14], Root Mean Square Error (RMSE), and Accuracy/Precision (ACC) to assess and compute the accuracy of our network architectures. The following describes two APHR kinds that are utilized for various class performance indicators:

- *Absolute Accuracy*: Computes the exact amount of outcome (B*ingo*). Accounts for correctly predicted classes.
- R*elative Accuracy*: Based on Bingo, the predictions (*1-away*) account for the predictions that are just 1 class apart.

Average Percent Hit Rate (APHR) is expressed as:

$$APHR = \frac{Accurately\ predicted\ number\ of\ samples}{Total\ class\ sample\ number} \quad (8)$$

$$APHR_{Bingo} = \frac{1}{n}\sum_{i=1}^{c} M_i \quad (9)$$

$$APHR_{1-away} = \frac{1}{n}\sum_{i=1}^{c}(M_{i-1} + M_i + M_{i+1}) \quad (10)$$

C = total classes (=5), n = entire samples from class i, and M_i = total samples predicted as class i, and *if* $i \leq 1$ or $i \geq 5$, $M_i = 0$. Precision or **Acc**uracy denotes the classification performance:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (12)$$

Represented as true positives, true negatives, false positives, and false negatives, respectively. Root Mean Square Error (RMSE) is the average difference between values predicted by a model and the actual values for image and video classification:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (13)$$

$\hat{y}_i = i^{th}$ predicted value of the model, $y_i = i^{th}$ actual value, $\bar{y}$ = the average of all classifications, and n = the data size. The better the model effect, the lower the RMSE number.

IV. Result and Discussion

We validated the efficiency of our Deep Multimodal Predictive Cross-Input Neural Network model against existing baseline models in box-office revenue prediction and image and video classification tasks. The current box-office revenue prediction techniques include a Content-based model [11], ensemble methods [7, 51], Hybrid Model [30], Stacking Fusion Model [18], PRBO [12], evolving DNN (best) [20], Deep Neural Network (DNN) [21]. We also compared our model with state-of-the-art models on image and video classification benchmarks, including DIViTA [35], ViT [52], temporal convolutions [42, 43, 52, 53], and ConvNets [54], on benchmark datasets.

A. Experimental Settings

We executed our models in "TensorFlow," utilizing "Intel Xeon Processor E5-2680 v3 (30M Cache, 2.50 GHz) GPU–NVidia TitanX Pascal (12 GB VRAM) RAM–128 GB DDR4 2133 MHz" [6] on the Linux operating system. During the image and video classification tasks, we used the ViT [52] experimental setup for pre-training and fine-tuning to obtain comparable results, and we only focused on fine-tuning performance (see Table IV).

Initially, we commenced with a modest look-back value of 10 frames and closely monitored our model's performance. This gave us a starting point to evaluate whether the model accurately captures relevant temporal information. Subsequently, we conducted experiments with a range of values, ranging from 20 to 50, and meticulously monitored the alterations in our model's performance. This experimentation was conducted as a part of a hyperparameter tuning process. Our model had difficulties capturing relevant patterns at a frame rate of 20. At 50 frames, the model generalizes to new trailers and captures long-term dependencies, improving performance without causing computational issues. We used this as the optimal value for our specific COGNIMUSE and Trailers dataset. We also used this for our fine-tuning MovieNet dataset as the value that balances capturing relevant temporal dependencies and computational efficiency. We performed this process iteratively, using cross-validation and hyperparameter tuning to find the best look-back value for our specific datasets and prediction task.

Training and testing data split: Unless otherwise noted, we used settings derived from experiments of numerous alternative neural network setups and training settings (51%, 14%, and 34 %) for training, validation, and testing on datasets. The distribution of videos by genre classes for training, validation, and testing for the LSTM is shown in Fig. 11. The rationale behind our utilization of a 51% training set is that it was of utmost importance due to the complex structure of our combined models and the substantial size of the datasets. The 14% validation set is utilized to fine-tune hyperparameters and monitor the model's performance throughout the training process. This smaller portion was allocated for validation to guarantee adequate evaluation examples without compromising the training set's size. The testing set, which accounts for 34% of the data, is kept separate until the completion of model

training. This collection aimed to assess the model's performance on data it had not been trained on, thereby objectively estimating its generalization ability. The larger size compared to the validation helped ensure a robust assessment.

TABLE IV. HYPERPARAMETERS OF MODELS

Category	Hyperparameter	Value(s)
Model Architecture	ResNet Depth	ResNet-50
	Input Image size (ResNet)	224×224×3
	LSTM Units	256
	LSTM Layers	3
	LSTM Activation Function	tanh
	LSTM Look back value	50 frames
	Cross-input neighborhood Layers	3
Data Pre-processing	Input Image Size (ResNet)	224×224
	Sequence Length (LSTM)	16
	Data Augmentation Parameters	Rotation, Flip, Zoom
	Normalization Strategy	Mean-std normalization
Training Hyperparameters	Learning Rate	0.0001
	Learning Rate Schedule	Step decay:0.01 every ten epochs
	Optimizer	Stochastic Gradient Descent (SGD)
	Batch Size	64
	Loss Function	Categorical cross-entropy (5kfold)
	Dropout Rate	0.5
	Weight Decay (L2 Regularization)	0.00001
	Gradient Clipping	5.0
	Early Stopping Patience	5
Training Loop Control	Maximum Training Steps	40,000
	Evaluation & Checkpointing Frequency	Every 1000 steps
	Logging Frequency	Every 100 steps
Validation and Testing	Metrics for Evaluation	APHR, Accuracy, RMSE
	Test Batch Size	54
	Test Time Augmentation	Enabled
Input Pipeline	Parallelization	12 (# of CPU processes)
	Data Prefetching	6(# of batches to prefetch)
Miscellaneous	Random Seed	42
	Model Initialization	He normal initialization

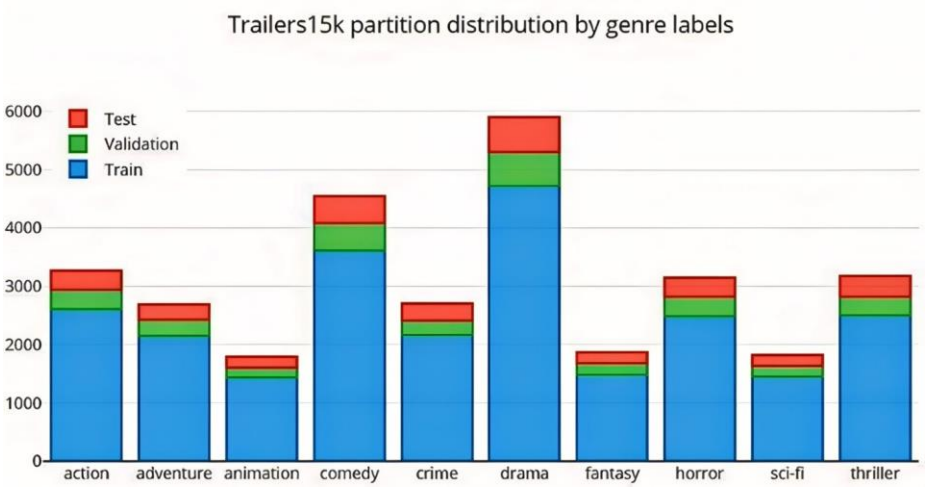


Fig. 11. Distribution of videos by genre classes for training, validation, and testing.

Based on our previous work [6], we proposed a reliable and systematic method for performing *k*-fold = (5-times) cross-validation, where the folds are generated based on collective conceptions rather than random selection. The entire dataset (M) is partitioned into k distinct classes M_1, M_2, …, M_K of equal size. The classification model is then trained and tested using *k*-fold cross-validation. Cross-validation *k*-fold was used to obtain more robust performance estimates, especially on datasets of limited size. Ultimately, we used this strategy, considering the trade-offs between having enough data for training, effective model tuning, and obtaining reliable performance metrics.

The ResNet testing, training, and validation sets are illustrated in Table V and Fig. 11. We pre-trained our model with the following datasets: Trailer 15k [35], ImageNet dataset [49], and the MovieNet dataset [38], fine-tuned our models, which provide exhaustive box annotation for all instances. Table VI shows the YouTube-8M dataset [53] partition, which comprises frame-level features for over 1.9 billion video frames and 8 million videos used for video classification tasks.

TABLE V. MULTI-LABELED TRAILER DATASET PARTITION

Revenue & Rating classes	Training	Validation	Testing	Total
1	7140	1960	4760	14000
2	7140	1960	4760	14000
3	7140	1960	4760	14000
4	510	140	340	1000
5	510	140	340	1000
Frames	9,856,600	1,582,000	2,850,000	14,288,600

The uneven distribution of movie data among genres might significantly impact the training outcomes since the model may be biased toward the dominant class. In order to tackle this issue, we employed data augmentation, resampling, and weighted loss function approaches. To address the issue of an imbalanced dataset on genres, we employed data augmentation by generating diverse versions of the minority class samples (see Table VI).

TABLE VI. DATASET PARTITION

Dataset	Training	Validation	Testing	Total
YouTube-8M	5,786,881	1,652,167	825,602	8,264,650

Subsequently, we applied a resampling technique that involved oversampling the minority class and under sampling the majority class, achieving a more equitable distribution. During the process of fine-tuning, we implemented a weighted loss function to provide greater significance to the minority class during the training phase. Ensuring a balance between addressing the class imbalance and avoiding overfitting is crucial. Our selection of approaches was contingent upon the dataset's unique characteristics and our model's effectiveness on validation sets.

B. *Box-Office Revenue Prediction Results*

Table VII and Fig. 12 compare proposed and existing state-of-the-art visual features for box office prediction. The results show that our visual feature vectors have an average precision of 81%, showing the dominance of the fusion strategy, the detailed learned features, and the predictive power of *Mid-level* + *High-level* multimodal features. The proposed fusion technique is effective because it incorporates embeddings with robust discriminative visual features. These types of embedding are strongly connected to the prediction task. The proposed model integrates mid-level features of movie frames/scenes from a trailer (motion content, scene recognition, high-level elements about the movie's quality and the narrative from the trailer styles, and object detection) to achieve maximum efficiency in classification and prediction.

TABLE VII. PRECISION COMPARISONS BETWEEN EXISTING AND PROPOSED TRAILER FEATURES

Visual Feature Vectors	1	2	3	4	5	Average
Mid-level [33]	47%	47%	48%	46%	47%	47%
Mid-level learned (proposed)	49%	49%	50%	50%	50%	51%
High-level [32]	46%	45%	45%	44%	43%	45%
High-level learned (proposed)	68%	66%	67%	67%	65%	66%
PRBO (Mid + High) [12]	57%	58%	59%	57%	57%	58%
Mid-level + High-level (proposed)	80%	81%	81%	80%	81%	81%

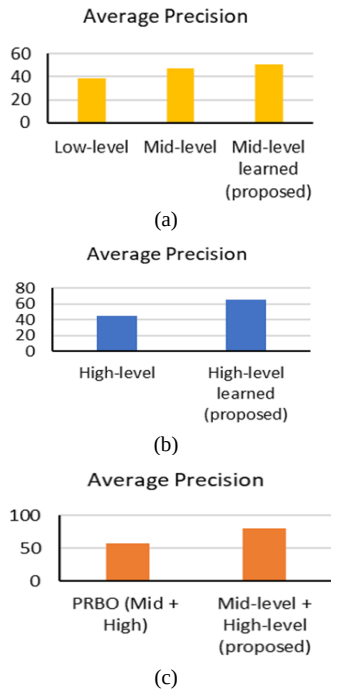


Fig. 12. Comparison between existing and proposed feature vectors' Accuracy, (a) Low-level & Mid-level Features, (b) High-level Features, (c) Combined Features.

We also compared our model to multi-model ensemble algorithms using Root Mean Square Error (RMSE). DMPCNN feature outperformed all the ensemble models as shown in Table VIII below:

TABLE VIII. RMSE COMPARISONS BETWEEN EXISTING AND PROPOSED TRAILER FEATURES

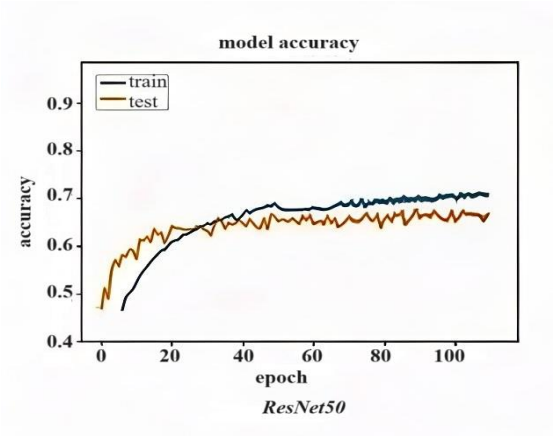
Models	RMSE
XGBoosting [7]	19,137.7581
LightGBM [7]	16,105.0243
Stacking Fusion Model [18]	17,974.5699
DMPCNN	13,845.2351

Furthermore, the multimodal feature vector's effectiveness is due to its comprehensive, erudite features revealed from movie trailers, demonstrating the efficiency of our feature extraction and fusion model algorithm.

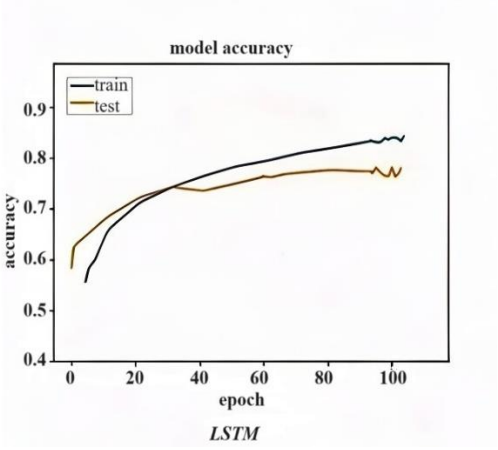
Fig. 13 demonstrates the training and testing accuracies of each model. The ResNet model alone (top left) had an accuracy of 69.76% during testing when using background and essential object features alone. The LSTM model alone (top right) had an accuracy of 74.02% during testing when we used temporal regression visual content features alone. We combined the two models to improve the performance of extracting and learning mid-level features. These deep visual trailer embeddings are extracted from

trailer key-frame images used in the ResNet+LSTM model (*bottom center*) and had an accuracy of 75.83% during testing. The performance is slightly improved for the combined ResNet+LSTM model for mid-level features without the Cross-input neighborhood difference. The effectiveness of introducing the Cross-input neighborhood difference is shown in Table IX (last row) at 79.44%.

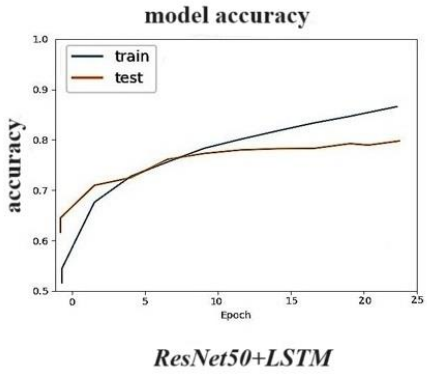
Table IX shows the ablation study; performance improves by introducing the Cross-input neighborhood difference technique to the ResNet+LSTM model, achieving an accuracy of 79.44%. The cross-input neighborhood difference jointly extracts high-level and mid-level features to learn and classify movie box office revenue. Accuracy continually improves, demonstrating the usefulness of our learning architecture and the process of optimizing the parameters during experiments.



(a)



(b)



(c)

Fig. 13. (a) ResNet50, (b) LSTM, (C) ResNet50 + LSTM, Training and Testing Model accuracies.

TABLE IX. ABLATION PERFORMANCE EVALUATION OF THE MODEL DESIGN

Model Design		1(%)	2(%)	3(%)	4(%)	5(%)
ResNet	Bingo	63.17	69.98	70.64	67.78	69.76
	1-Away	53.49	70.79	80.10	87.47	97.84
LSTM	Bingo	65.78	71.33	72.17	73.62	74.02
	1-Away	94.13	95.52	96.18	96.53	96.91
ResNet + LSTM (w/o Cross-input)	Bingo	72.25	73.65	75.23	76.96	75.83
	1-Away	97.87	97.94	98.28	98.17	98.33
with Cross-input (before fine-tuned)	Bingo	78.36	76.66	78.96	77.19	79.44
	1-Away	98.15	98.33	98.25	98.23	98.33

As illustrated in Fig. 14, the fine-tuned model testing is also outstanding, reaching 83.40% when the complete architecture's hyper-parameters are integrated within our fully connected Deep Multimodal Predictive Cross-input Neural Network (DMPCNN) model. We fine-tuned our model parameters and hyperparameters using the MovieNet dataset. The results show that the proposed model's effectiveness in learning and extracting trailer features strongly correlated with box office revenue before the movie's theatrical release. Table X compares existing box-office baseline models and the proposed model at each testing set.

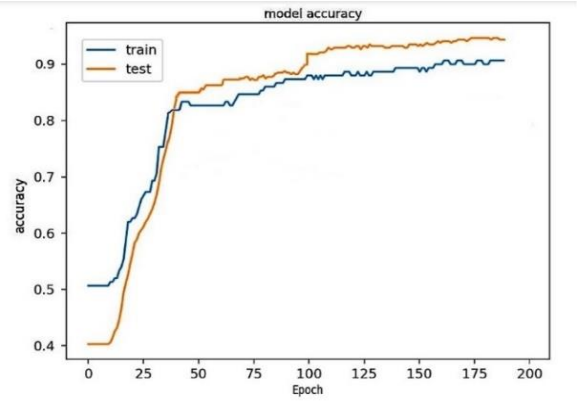


Fig. 14. Deep multimodal predictive cross-input neural network model training and testing.

From the comparisons in Fig. 15 and Table X, we can determine that our proposed framework significantly outperformed the existing box-office baseline models: multi-model ensemble, PRBO, Stacking Model, Hybrid, DNN, Evolved DNN (best and mean), Hybrid, and Content-based methods using APHR Bingo.

APHR results show that the proposed model outperforms existing methodologies due to the predictive power of combining movie trailer features and metadata, improving the performance of forecasting models. Our upgraded model, Deep Multimodal Predictive Cross-input Neural Network (DMPCNN), had an average of 98.61% and 84.40%, respectively.

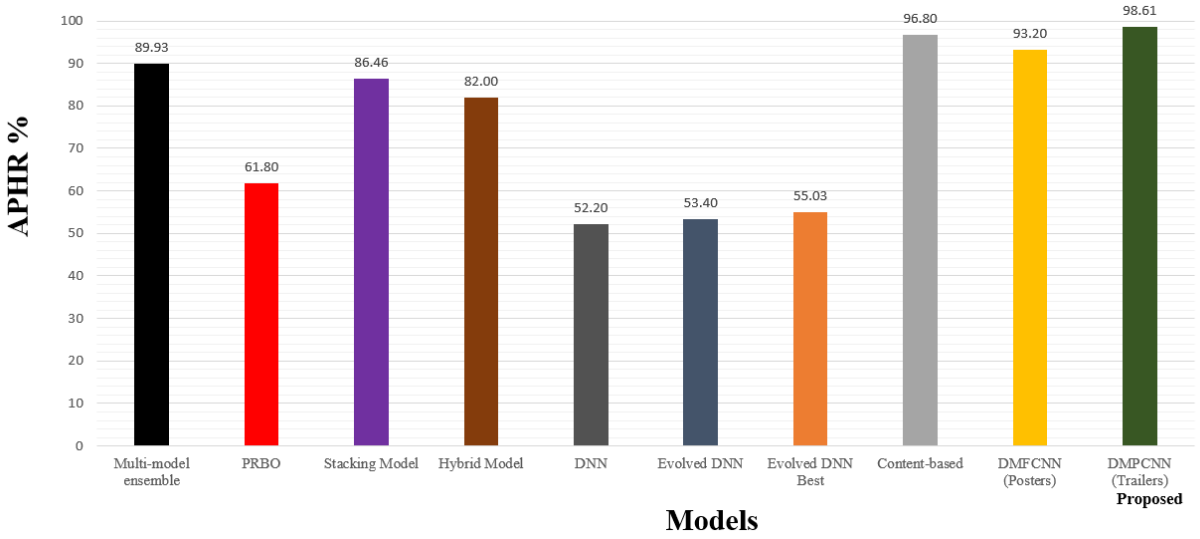


Fig. 15. Performance comparison of existing and proposed approaches.

TABLE X. APHR ANALYSIS OF EXISTING BOX-OFFICE BASELINE AND PROPOSED MODEL

Model		1(%)	2(%)	3(%)	4(%)	5(%)	Average (%)
†Multi-model ensemble [7]		-	-	-	-	-	89.93
†Content-based [11]		-	-	-	-	-	96.80
Random Forest [55]		-	-	-	-	-	96.70
KHDEM [56]		-	-	-	-	-	96.04
SHAP [57]		-	-	-	-	-	79.00
Gradient Boosting [58]		-	-	-	-	-	92.40
Hybrid Features [22]		-	-	-	-	-	86.30
Stacking Fusion Model [18]	Bingo	-	-	-	-	-	69.16
	1-Away	-	-	-	-	-	86.46
Hybrid Model [30]	Bingo	-	-	-	-	-	82.00
	1-Away	-	-	-	-	-	95.00
DNN [21]	Bingo	53.22	50.00	54.56	50.44	52.78	52.20
Evolved DNN(best) [20]	Bingo	55.60	54.32	56.39	52.53	55.13	55.03
	1-Away	92.12	91.14	89.49	91.54	90.80	91.33
PRBO [12]	Bingo	45.00	47.00	54.00	61.00	63.00	61.80
DMFCNN [6]	Bingo	57.90	55.38	55.19	52.90	55.13	59.30
	1-Away	93.40	92.89	91.68	91.25	91.80	93.20
Proposed DMPCNN (Trailers)	Bingo	83.25	82.13	82.36	84.21	85.00	84.40
	1-Away	90.36	93.25	93.53	93.53	95.61	98.61

† Target Audience Prediction, not revenue prediction

Using 5-fold cross-validation on our dataset, we tested the significance of our model against other classifiers. From Table XI, our *p*-value of 0.004 is relatively small, indicating that the model result is unlikely to have occurred by random chance compared to other classifiers. Our T-statistic of 2.98 indicates a relatively significant difference between the other classifiers and is statistically significant. *P*-value (0.004) and the large T-statistic (2.98) suggest strong evidence against the other classifiers. The conclusion shows that the observed difference in model skill is likely due to a difference in the models' construction and predictive power.

TABLE XI. SIGNIFICANCE T-TEST EVALUATION BETWEEN CLASSIFIERS

Classifier	SVM	Naïve Bayes	Decision Tree	Random Forest	k-NN	LR	Proposed DMPCNN
T-statistics	2.75	–3.20	–0.86	–2.40	–4.25	–1.43	2.98
p-value	0.234	0.264	0.398	0.125	0.421	0.983	0.004*

*Indicate statistical significance (*p* value<0.05)

C. Image and Video Classification Results

Though our primary task was movie box-office revenue prediction using trailers and metadata, we tested our model performance on image and video classification tasks. For image classification in Table XII, the DMPCNN was evaluated against state-of-the-art image classification models when pre-trained on vast quantities of data and applied to numerous small to medium-sized image classification benchmarks (ImageNet, ImageNet-21k, CIFAR-100, CIFAR-10) [52]. For video classification in Fig. 16, we used the YouTube-8M dataset [53].

TABLE XII. PERFORMANCE EVALUATION ON IMAGE CLASSIFICATION BENCHMARKS

Models	ImageNet	ImageNet21k	CIFAR-100	CIFAR-10	Augmented Dataset
GRU [59]	76.22 ± 2.88	78.95 ± 0.66	–	–	–
ConvNets [54]	77.03 ± 2.78	79.30 ± 1.03	–	–	–
Vision Transformers (ViT) [52]	85.30 ± 0.02	88.62 ± 0.05	93.25 ± 0.05	99.15 ± 0.03	99.74 ± 0.00
Transformers [35]	77.49 ± 1.18	80.02 ± 0.47	–	–	–
DMPCNN	87.54 ± 0.02	90.54	93.51 ± 0.08	99.37 ± 0.06	99.63 ± 0.03

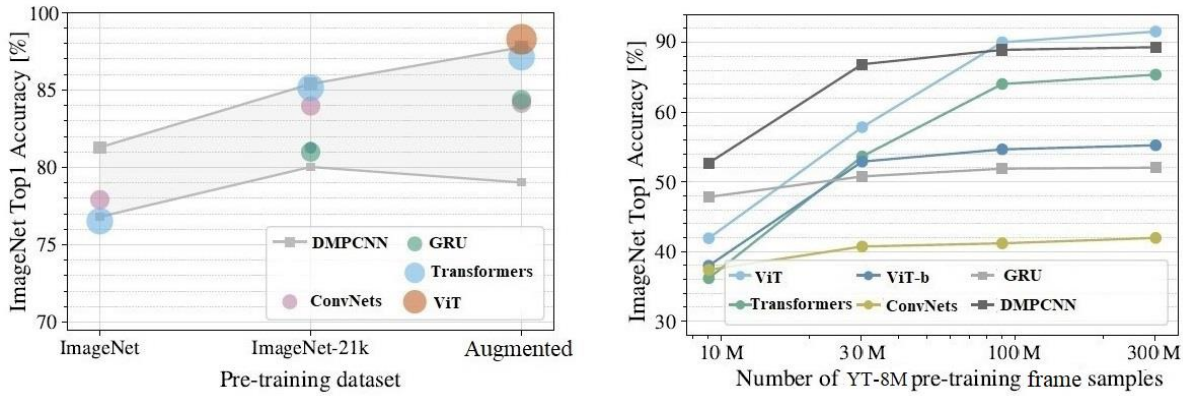


Fig. 16. Performance comparison of pre-training on benchmark datasets. (a) Pre-training datasets, (b) Pre-training frames.

Table XII shows that the pre-trained Vision Transformer (ViT) model on the Augmented dataset (ImageNet + CIFAR) outperforms our model DMPCNN (ResNet+LSTM) while requiring significantly fewer CPU resources to train. ViT had (99.74 ± 0.00) as compared to DMPCNN (99.63 ± 0.03), though the difference of (0.11) is not that big, Vision Transformers performs better when pre-trained on enormous public datasets; however, DMPCNN outperforms baseline models when pre-trained on smaller datasets achieving (87.54 ± 0.02) on ImageNet, (90.54) on ImageNet21k, (93.51 ± 0.08) on CIFAR-100 and (99.37 ± 0.06) on CIFAR-10 respectively. ViT slightly exceeds our models, achieving (99.74 ± 0.00) when pre-trained on larger augmented datasets.

Fig. 16, based on Ref. [52], shows that DMPCNN outperforms ViT with fewer pre-training datasets (shaded region) but reaches a plateau faster. Vision Transformers over-fit more than DMPCNN on smaller datasets with equivalent computational cost. This finding supports the assumption that, while the convolutional inductive bias is adequate for smaller datasets, learning the necessary structures from the data is adequate, if not advantageous, for bigger ones. When the datasets grow more extensive, ViT versions outperform all models. To pre-train, DMPCNN requires considerably less computation than the previous state-of-the-art approaches. Nevertheless, we should highlight that the architectural choice and other parameters such as training schedule, optimizer, and weight decay might impact pre-training efficiency.

D. Practical Implications of the Study

We analyzed each attribute and the regression effectiveness of our proposed model using actual box office data. The correlation between proposed learned feature knowledge and box office revenue is shown in Fig. 17, which depicts popular films' characteristics and model Regression effectiveness (i.e., flop, breakeven, and blockbuster) regarding revenue.

Experiments on the Internet Movie Database (IMDB) market demonstrate the practical application of our suggested deep multimodal feature vectors and likelihood outcomes. Experimental findings demonstrate the DMPCNN model's relevance and regression efficacy in the real world, as the framework provides studios, investors, and production teams with distinct twofold types of conclusions (i.e., prediction accuracy and learned knowledge to be incorporated into trailers). Ultimately, the hypothetical relevance motivates us to consider implementing our model as a commercial facility to support investment decisions and business intelligence in the film industry. Our model's success demonstrates the efficacy of our feature extraction framework, as it confirmed high distinguishability due to incorporating semantic and structural information.

Movies with a comparable box office gross tend to cluster. Proposed learned feature knowledge and box office revenue are highly correlated. Fig. 17 shows popular movies' features and model Regression efficacy (i.e., flop, breakeven, and blockbuster).

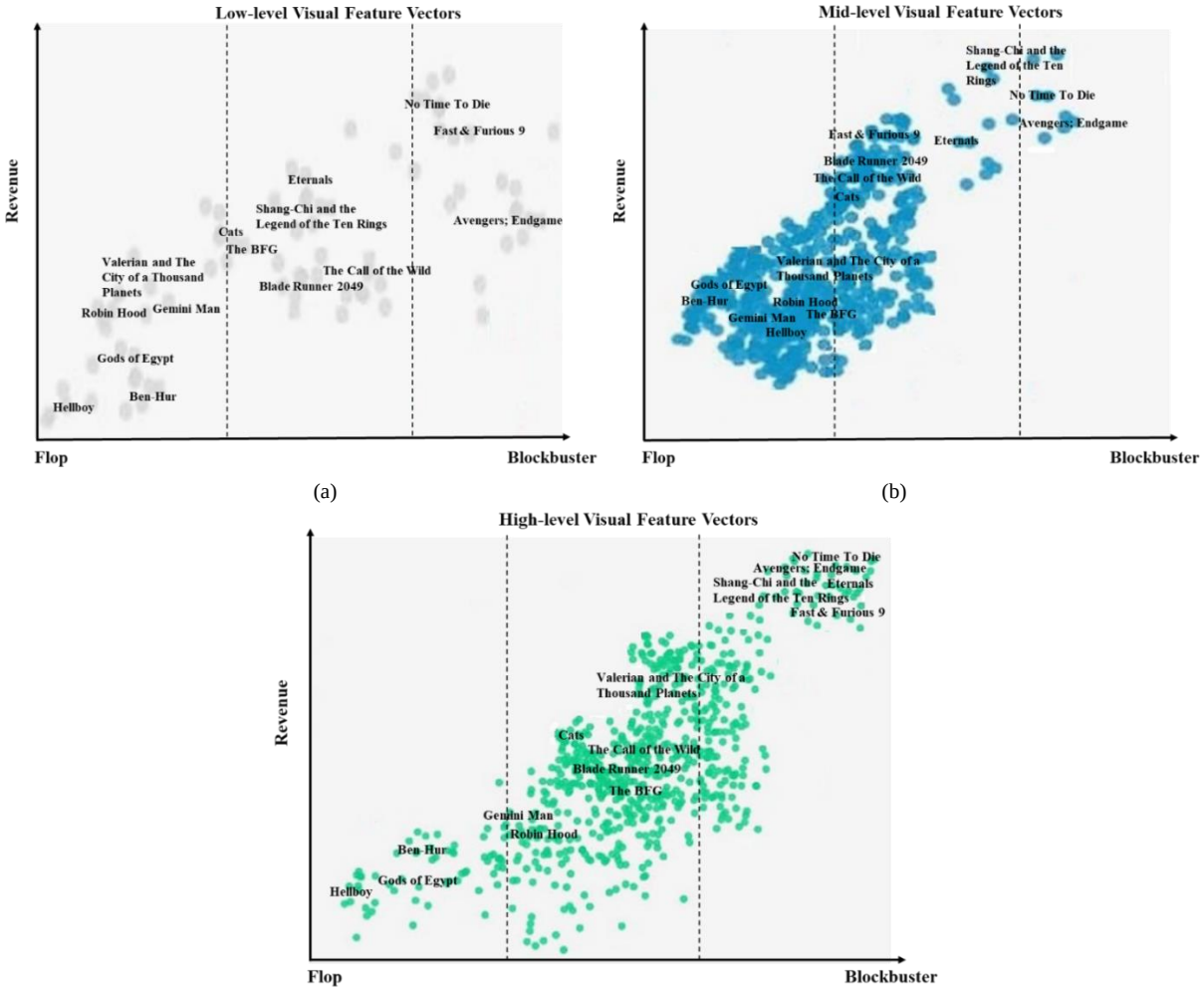


Fig. 17. Features and Model Regression Efficacy in Practical Box office prediction. (a) Low-level, (b) Mid-level, (c) High-level.

We analyzed each feature and our proposed model regression efficacy using real-world box office data. For instance, action/adventure movies with a similar genre are distributed close, assisting in the final prediction. Indeed, blockbuster movies such as No Time to Die, Avengers Endgame, Shang-Chi and the Legends of the Ten Rings, F9, and Eternals clustered together with the learned high-level features. High-level feature vectors assist in the decisive prediction objective, as shown below. Movies with similar features and box office revenue tend to cluster together, as shown in Fig. 17(c) of High-level visual feature vectors.

Limitations: Most current studies are primarily concerned with film metadata [8, 10, 11] and sentimental analysis of trailer reviews, and no investigation has been directed at the benefit of movie trailers and the use of recurrent networks in forecasting box office revenue. Therefore, some of the present study's drawbacks are described below:

- Researchers predominantly use trailer features for video recommendation models and textual content to recommend interesting videos to users.
- Movie trailer content is more difficult to correlate to its box office income than genre classification tasks.
- Neural networks require many computational resources, are challenging to train with limited movie trailers, and are hard to grasp forecasts.

Significant limitations to our model are that we are yet to test our suggested model against cross-domain datasets and biologically inspired neural networks unrelated to the motion picture box office. The other drawback of our model is that it did not perform very well in image and video classification and recognition when applied to large datasets compared to other state-of-the-art Transformers for image classification tasks. Vision Transformers (ViT) [52] recently demonstrated competitive performance in benchmarks for various computer vision tasks, including semantic segmentation of images, object detection, and image and video classification. Training them on smaller datasets requires more computational resources and results in a weaker inductive bias, necessitating a greater dependence on data Augmentation (AugReg) or model regularization. This necessitates training the Vision Transformer model on a massive dataset before fine-tuning and validation. It is challenging to implement ViT models effectively if deployed on devices with limited resources. Consequently, computer vision research indicates that ViT frameworks are as

robust as ResNet models when pre-trained with adequate data.

Therefore, we designed our model using ResNet+LSTM block architecture and a Cross-Input Neural Network fusion strategy, as our research is not solely focused on the learning representation of trailer features.

V. Conclusion

Our research paper suggested a two-stage deep multimodal learning model for predicting a motion picture's box-office revenue before its theatrical premiere. We propose using a mixture of ResNet+LSTM features and a Cross-input neighborhood difference fusion, each of which aims to learn distinct parts of the movie trailer by applying several feature-learning representations to extract different features from the movie's multimodal data. We propose an innovative Deep Multimodal Predictive Cross-Input Neural Network embedding framework for collectively learning a trailer's mid-level and high-level movie frames/scenes features. The model can effectively map a sequence of frames into intangible movie box office ratings and revenue predictions. We develop a methodology based on recurrent neural networks to extract mid-level and high-level depictions of motion picture quality from trailers. Then, a feature vector for the film is constructed using these vectors and used as an input of a fully-connected prediction model to generate the movie box office prediction. The mid-level features identified include the appearance (i.e., background, genre-basic objects, scene, aesthetics, color, and texture) and motion features of movie castings. Essential high-level features were uncovered, including the filming quality, narrative, and filming styles (i.e., the shooting quality affects the audience's perception). The proposed learned features accurately predicted the opening motion picture box-office revenue prior to its premiere in theaters with correctly predicted classes accuracy (Bingo) of 84.40% and with just one class apart category (1-Away) of 98.61% accuracy, and visual feature vectors of 81% precision. Our model outperformed all the existing baseline approaches, confirming our hypothesis that trailers affect audience perception and ultimately affect box office revenue. We confirmed our hypothesis that recurrent neural networks solve complex tasks involved in feature engineering by extracting and learning deep multimodal visual features related to movie revenue in movie trailers. These features have a substantial influence on the audience's perception of the movie, and they are the ones that capture the audience's attention and have a persuasive effect. They create a mood, enigma, interest, and anticipation of what the film offers.

Our future work will improve prediction performance by combining rich information in movie trailers and movie posters and their effect on the short-term life cycle of movie box-office revenues (before, during, and after) when they launch in movie theaters. Including audio-visual and textual features in movie abstracts and scripts using a Transformer encoder as an NLP technique for movie review, sentimental analysis is worth investigating.

Conflict of Interest

The authors declare no conflicts of interest.

Author Contributions

Canaan Tinotenda Madongo: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing, Review and Editing, Visualization; Zhongjun Tang: Supervision, Validation, Investigation, Resources, Writing, Review and Editing, Project administration, Funding acquisition; Jahanzeb Hassan: Data Curation, Writing, Review and Editing, Visualization; all authors had approved the final version.

Funding

This work is supported by the National Nature Science Foundation of China under Grant No. 71672004.

Acknowledgment

We gratefully acknowledge the support of the National Nature Science Foundation of China and the Beijing University of Technology for providing all the necessary funds and support during this research work.

References

[1] B. Abdualgalil, S. Abraham, and W. M. Ismael, "COVID-19 infection prediction using efficient machine learning techniques based on clinical data," *Journal of Advances in Information Technology*, vol. 13, no. 5, pp. 530–538, 2022. doi: 10.12720/jait.13.5.530-538

[2] J. C. T. Arroyo and A. J. P. Delima, "An optimized neural network using genetic algorithm for cardiovascular disease prediction," *Journal of Advances in Information Technology*, vol. 13, no. 1, pp. 95–99, 2022. doi: 10.12720/jait.13.1.95-99

[3] M. Ashrafuzzaman, S. Saha, and K. Nur, "Prediction of stroke disease using deep CNN based approach," *Journal of Advances in Information Technology*, vol. 13, no. 6, pp. 604–613, 2022. doi: 10.12720/jait.13.6.604-613

[4] W. Kusonkhum, K. Srinavin, N. Leungbootnak, P. Aksorn, and T. Chaitongrat, "Government construction project budget prediction using machine learning," *Journal of Advances in Information Technology*, vol. 13, no. 1, pp. 29–35, 2022. doi: 10.12720/jait.13.1.29-35

[5] J. Wang, J. Shi, D. Han, and X. Zhao, "Internet financial news and prediction for stock market: An empirical analysis of tourism plate based on LDA and SVM," *Journal of Advances in Information Technology*, vol. 10, no. 3, pp. 95–99, 2019. doi: 10.12720/jait.10.3.95-99

[6] C. T. Madongo and T. Zhongjun, "A movie box office revenue prediction model based on deep multimodal features," *Multimedia Tools and Applications*, no. 100, 2023. doi: 10.1007/s11042-023-14456-4.

[7] Y. Ni, F. Dong, M. Zou, and W. Li, "Movie box office prediction based on multi-model ensembles," *Information (Switzerland)*, vol. 13, no. 6, 2022. doi: 10.3390/info13060299

[8] Y. An, J. An, and S. Cho, "Artificial intelligence-based predictions of movie audiences on opening saturday," *International Journal of Forecasting*, vol. 37, no. 1, pp. 274–288, 2021. doi: 10.1016/j.ijforecast.2020.05.005

[9] I. S. Ahmad, A. A. Bakar, and M. R. Yaakub, "Movie revenue prediction based on purchase intention mining using YouTube trailer reviews," *Information Processing and Management*, vol. 57, no. 5, Sep. 2020. doi: 10.1016/j.ipm.2020.102278

[10] S. Sahu, R. Kumar, P. Mohdshafi, J. Shafi, S. Kim, and M. F. Ijaz, "A hybrid recommendation system of upcoming movies using

sentiment analysis of YouTube trailer reviews," *Mathematics*, vol. 10, no. 9, pp. 1–22, 2022. doi: 10.3390/math10091568

[11] S. Sahu, R. Kumar, M. S. Pathan, J. Shafi, Y. Kumar, and M. F. Ijaz, "Movie popularity and target audience prediction using the content-based recommender system," *IEEE Access*, vol. 10, pp. 42030–42046, 2022. doi: 10.1109/ACCESS.2022.3168161

[12] Z. Wang, J. Zhang, S. Ji, C. Meng, T. Li, and Y. Zheng, "Predicting and ranking box office revenue of movies based on big data," *Information Fusion*, vol. 60, no. June 2019, pp. 25–40, 2020. doi: 10.1016/j.inffus.2020.02.002

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. the IEEE Computer Society* Conference *on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90

[14] R. Sharda and D. Delen, "Predicting box-office success of motion pictures with neural networks," *Expert Systems with Applications*, vol. 30, no. 2, pp. 243–254, 2006. doi: 10.1016/j.eswa.2005.07.018

[15] L. Zhang, J. Luo, and S. Yang, "Forecasting box office revenue of movies with BP neural network," *Expert Systems with* Applications, vol. 36, no. 3, pp. 6580–6587, 2009. doi: 10.1016/j.eswa.2008.07.064

[16] N. Quader, M. O. Gani, D. Chaki and M. H. Ali, "A machine learning approach to predict movie box-office success," in *Proc. 2017 20th International Conference of Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh, 2017, pp. 1–7. doi: 10.1109/ICCITECHN.2017.8281839.

[17] R. Parimi and D. Caragea, "Pre-release box-office success prediction for motion pictures," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7988, pp. 571–585, 2013. doi: 10.1007/978-3-642-39712-7_44

[18] Y. Liao, Y. Peng, S. Shi, V. Shi, and X. Yu, "Early box office prediction in China's film market based on a stacking fusion model," *Annals of Operations Research*, 2020. doi: 10.1007/s10479-020-03804-4

[19] Z. Tang and S. Dong, "A total sales forecasting method for a new short life-cycle product in the pre-market period based on an improved evidence theory: Application to the film industry," *International Journal of Production Research*, pp. 1–15, 2020. doi: 10.1080/00207543.2020.1825861

[20] Y. Zhou and G. G. Yen, "Evolving deep neural networks for movie box-office revenues prediction," in *Proc. 2018 IEEE Congress on Evolutionary Computation*, 2018. doi: 10.1109/CEC.2018.8477691

[21] Y. Zhou, L. Zhang, and Z. Yi, "Predicting movie box-office revenues using deep neural networks," *Neural Computing and Applications*, vol. 31, no. 6, pp. 1855–1865, 2019. doi: 10.1007/s00521-017-3162-x

[22] M. T. Lash and K. Zhao, "Early predictions of movie success: The who, what, and when of profitability," *Journal of Management Information Systems*, vol. 33, no. 3, pp. 874–903, 2016. doi: 10.1080/07421222.2016.1243969

[23] Y. Ru, B. Li, J. Liu, and J. Chai, "An effective daily box office prediction model based on deep neural networks," *Cognitive Systems Research*, vol. 52, pp. 182–191, 2018. doi: 10.1016/j.cogsys.2018.06.018

[24] W. Wang, J. Xiu, Z. Yang, and C. Liu, "A deep learning model for predicting movie box office based on deep belief network," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 1, 2018. doi: 10.1007/978-3-319-93818-9_51

[25] M. Mestyán, T. Yasseri, and J. Kertész, "Early prediction of movie box office success based on wikipedia activity big data," *PLoS ONE*, vol. 8, no. 8, 2013. doi: 10.1371/journal.pone.0071226

[26] M. Hur, P. Kang, and S. Cho, "Box-office forecasting based on sentiments of movie reviews and Independent subspace method," *Information Sciences*, vol. 372, pp. 608–624, 2016. doi: 10.1016/j.ins.2016.08.027

[27] R. B. Mangolin *et al.*, "A multimodal approach for multi-label movie genre classification," *Multimedia Tools and Applications*, 2020. doi: 10.1007/s11042-020-10086-2

[28] G. S. Simões, J. Wehrmann, R. C. Barros, and D. D. Ruiz, "Movie genre classification with convolutional neural networks," in *Proc. the International Joint Conference on Neural Networks*, 2016, pp. 259–266. doi: 10.1109/IJCNN.2016.7727207

[29] Y. Matsuzaki *et al.*, "Could you guess an interesting movie from the posters? An evaluation of vision-based features on movie poster database," in *Proc. the 15th IAPR International Conference on Machine Vision Applications, MVA 2017*, 2017, pp. 538–541. doi: 10.23919/MVA.2017.7986919

[30] U. Ahmed, H. Waqas, and M. T. Afzal, "Pre-production box-office success quotient forecasting," *Soft Computing*, vol. 24, no. 9, pp. 6635–6653, May 2020. doi: 10.1007/s00500-019-04303-w

[31] J. Finsterwalder, V. G. Kuppelwieser, and M. de Villiers, "The effects of film trailers on shaping consumer expectations in the entertainment industry—A qualitative analysis," *Journal of Retailing and Consumer Services*, vol. 19, no. 6, pp. 589–595, 2012. https://doi.org/10.1016/j.jretconser.2012.07.004

[32] S. Oh, J. Ahn, and H. Baek, "Viewer engagement in movie trailers and box office revenue," in *Proc. Annual Hawaii* International *Conference on System Sciences*, 2015, pp. 1724–1732. doi: 10.1109/HICSS.2015.207

[33] A. Tadimari, N. Kumar, T. Guha, and S. S. Narayanan, "Opening big in box office? Trailer content can help," in *Proc. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 2777–2781. doi: 10.1109/ICASSP.2016.7472183

[34] M. S. Rahim, A. Z. M. E. Chowdhury, M. R. M. A. M. R. Islam, and M. R. M. A. M. R. Islam, "Mining trailers data from youtube for predicting gross income of movies," in *Proc. 5th IEEE Region 10 Humanitarian Technology Conference 2017, R10-HTC 2017*, 2018, pp. 551–554. doi: 10.1109/R10-HTC.2017.8289020

[35] R. Montalvo-lezama, B. Montalvo-lezama, and G. Fuentes-pineda, "Improving transfer learning for movie trailer genre classification using a dual image and video transformer," *Information Processing and Management*, vol. 60, no. 3, 2023. https://doi.org/10.1016/j.ipm.2023.103343

[36] T. V. Wenzlawowicz and O. Herzog, "Semantic video abstracting: Automatic generation of movie trailers based on video patterns," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7297, pp. 345–352, 2012. doi: 10.1007/978-3-642-30448-4_44

[37] I. U. Haq *et al.*, "Movie scene segmentation using object detection and set theory," *International Journal of Distributed Sensor* Networks, vol. 15, no. 6, pp. 1–7, 2019. doi: 10.1177/1550147719845277

[38] Q. Huang, Y. Xiong, A. Rao, J. Wang, and D. Lin, "MovieNet: A holistic dataset for movie understanding," in *Proc. 2020 European Conference on Computer Vision, ECCV 2020*, 2020, pp. 709–727. doi: 10.1007/978-3-030-58548-8_41

[39] A. Zlatintsi *et al.*, "COGNIMUSE: A multimodal video database annotated with saliency, events, semantics and emotion with application to summarization," *Eurasip Journal on Image* and *Video Processing*, vol. 2017, no. 1, pp. 1–24, 2017. doi: 10.1186/s13640-017-0194-1

[40] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3908–3916. doi: 10.1109/CVPR.2015.7299016

[41] B. Duke and G. W. Taylor, "Generalized hadamard-product fusion operators for visual question answering," in *Proc. the 2018 15th Conference on Computer and Robot Vision, CRV 2018*, 2018, pp. 39–46. doi: 10.1109/CRV.2018.00016

[42] M. Y. Yang, X. Yong, and B. Rosenhahn, "Feature regression for multimodal image analysis," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 770–777. doi: 10.1109/CVPRW.2014.118.

[43] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, 2017. doi: 10.1109/TPAMI.2016.2599174

[44] J. Y. H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, vol. 3. doi: 10.1109/CVPR.2015.7299101

[45] X. Du, Y. Li, Y. Cui, R. Qian, J. Li, and I. Bello, "Revisiting 3D ResNets for video recognition," arXiv preprint, arXiv:2109.01696, 2021.

[46] I. C. Duta, L. Liu, F. Zhu, and L. Shao, "Improved residual networks for image and video recognition," in *Proc. International Conference on Pattern Recognition*, 2020, pp. 9415–9422. doi: 10.1109/ICPR48806.2021.9412193

[47] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. F. Li, "Large-scale video classification with convolutional neural networks," in *Proc. the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, vol. 223. doi: 10.1109/CVPR.2014.223

[48] J. Wehrmann, R. C. Barros, G. S. Simoes, T. S. Paula, and D. D. Ruiz, "(Deep) learning from frames," in *Proc. the 2016 5th Brazilian Conference on Intelligent Systems, BRACIS 2016*, Feb. 2017, pp. 1–6. doi: 10.1109/BRACIS.2016.012

[49] G. E. Krizhevsky *et al*., "ImageNet classification with deep convolutional neural networks," in *Proc. Advances in Neural Information Processing Systems*, 2012. doi: 10.1201/9781420010749

[50] Q. Huang, Y. Xiong, Y. Xiong, Y. Zhang, and D. Lin, "From trailers to storylines: An efficient way to learn from movies," arXiv preprint, arXiv.1806.05341, 2018.

[51] S. Lee, K. C. Bikash, and J. Y. Choeh, "Comparing performance of ensemble methods in predicting movie box office revenue," *Heliyon*, vol. 6, no. 6, e04260, 2020. doi: 10.1016/j.heliyon.2020.e04260

[52] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," arXiv preprint. arXiv:2010.11929, 2021.

[53] S. Abu-El-Haija *et al.*, "YouTube-8M: A large-scale video classification benchmark," arXiv preprint, arXiv:1609.08675, 2016.

[54] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR* 2022*)*, pp. 11976–11986. https://doi.org/10.48550/arXiv.2201.03545

[55] D. Li and Z. P. Liu, "Predicting box-office markets with machine learning methods.," *Entropy (Basel, Switzerland)*, vol. 24, no. 5, May 2022. doi: 10.3390/e24050711

[56] S. Sahu, R. Kumar, H. V. Long, and P. M. Shafi, "Early-production stage prediction of movies success using K-fold hybrid deep ensemble learning model," *Multimedia Tools and Applications*, vol. 82, no. 3. 2023. doi: 10.1007/s11042-022-13448-0

[57] S. B. Kumar and S. D. Pande, "Explainable neural network analysis on movie success prediction," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 8, 2024. doi: 10.4108/eetsis.4435

[58] M. H. Shahid and M. A. Islam, "Investigation of time series-based genre popularity features for box office success prediction," *PeerJ Computer Science*, vol. 9, e1603, 2023. doi: 10.7717/peerj-cs.1603

[59] Z. Niu *et al.*, "Recurrent attention unit: A new gated recurrent unit for long-term memory of important parts in sequential data," *Neurocomputing*, vol. 517, pp. 1–9, 2023. doi: 10.1016/j.neucom.2022.10.050