A Cross-Modal Transformer Based Model for Box-office Revenue Prediction

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.J.T.); jahanzab.hassan@gmail.com (J.H.)
*Corresponding author

***Abstract*—In the dynamic entertainment industry, predicting a movie's opening box office revenue remains critical for filmmakers and studios. To address this challenge, we present a novel Cross-modal transformer and a Hierarchical Fusion Neural Network (CHFNN) model tailored to predict movie box office earnings based on multimodal features extracted from movie trailers, posters, and reviews. The Cross-modal Transformer component of the CHFNN model captures intricate inter-modal relationships by performing a cross-modal fusion of the extracted features. It employs self-attention mechanisms to dynamically weigh the importance of each modality's information. This allows the model to learn to focus on the most relevant information from trailers, posters, and reviews, adapting to the unique characteristics of each movie. The Hierarchical Fusion Neural Network within CHFNN further refines the fused features, enabling a deeper understanding of the inherent hierarchical structure of multimodal data. By hierarchically combining the cross-modal features, our model learns to capture both global and local interactions, enhancing its predictive capacity. We evaluate the performance of the CHFNN model on a comprehensive Internet Movie Dataset by obtaining metadata for 50,186 movies from the 1990s to 2022, which includes movie trailers, posters, and review data. Our results demonstrate that the CHFNN model outperforms existing models in prediction accuracy, achieving 95.80% prediction accuracy. The CHFNN model provides state-of-the-art predictive power and offers interpretability through attention mechanisms, allowing insights into the factors contributing to a movie's box office success.**

***Keywords*—box-office, movie posters, movie trailers, movie reviews, cross-modal transformers, predictions**

I. Introduction

Transformers and pre-trained models have demonstrated significant accomplishments in various domains of artificial intelligence, including but not limited to natural language processing, computer vision, and audio processing [1, 2]. Data-driven methodologies, specifically machine learning approaches, have significantly enhanced prediction accuracy in various domains, including entertainment. Movie income forecasting is critical given the risks inherent in film production, notwithstanding its high cost [3]. 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 [4–7]. Reliable forecasts reduce the risk of loss and allow filmmakers to make operational decisions, such as budget, promotion, and advertising expenditures, throughout the filmmaking process apportionments [7]. An accurate box office prediction can provide film production and distribution businesses with business decision assistance and direction, which is crucial for the film industry's sustained growth [6, 8]. The cast, genre, or marketing budget of a movie are just a few examples of the solitary aspects that are often the focus of traditional methods for predicting revenue.

The proposed model includes three primary modalities: movie posters, movie trailers, and reviews. A film's essence and visual appeal are captured in a single image by movie posters, which provide static visual cues. Trailers, however, include dynamic visual elements, such as shot changes, aural features, and pacing, essential for generating audience interest. The sentiments, opinions, and comments surrounding a movie are normally best understood through textual reviews.

To effectively combine these diverse modalities, we utilize a Cross-modal Transformer and a Hierarchical Fusion Neural Network model (CHFNN), a novel architecture that captures cross-modal interactions and dependencies. The model extracts modality-specific features from posters, trailers, and reviews using modality-specific branches. These features are then fused and interacted with through cross-modal attention mechanisms, allowing the model to learn the relationships and correlations between different modalities. The fused representation is then fed into multiple layers of Transformers, enabling the model to capture complex interactions and dependencies within the combined feature space. Finally, a prediction head maps the transformed representations to the predicted box-office revenue. This enables the model to generate revenue predictions based on the integrated information from posters, trailers, and reviews.

Manuscript received January 13, 2024; revised February 24, 2024; accepted March 13, 2024; published July 8, 2024.

Our solution to the prediction difficulty was to consider it a classification problem. Combining static and dynamic features from posters, trailers, and reviews, our model provides a holistic view of a movie's potential box-office performance. It captures the visual static appeal, visual dynamics, and audience sentiments, enabling a more nuanced and accurate prediction of box-office revenue. The proposed model can potentially assist movie industry professionals in making informed decisions regarding marketing strategies, resource allocation, and release planning.

Through this cross-modal approach, we aim to leverage the visual appeal captured by posters, the dynamic elements depicted in trailers, and the textual sentiment and context conveyed by reviews, as shown in Fig. 1. By combining these modalities, our model can potentially uncover valuable insights that can enhance the accuracy of box-office revenue predictions, empowering industry professionals to make more informed decisions. Viewers determine all reviews based on a film's visual quality and storyline, irrespective of the filming style, which affects the film's revenue [9–12].

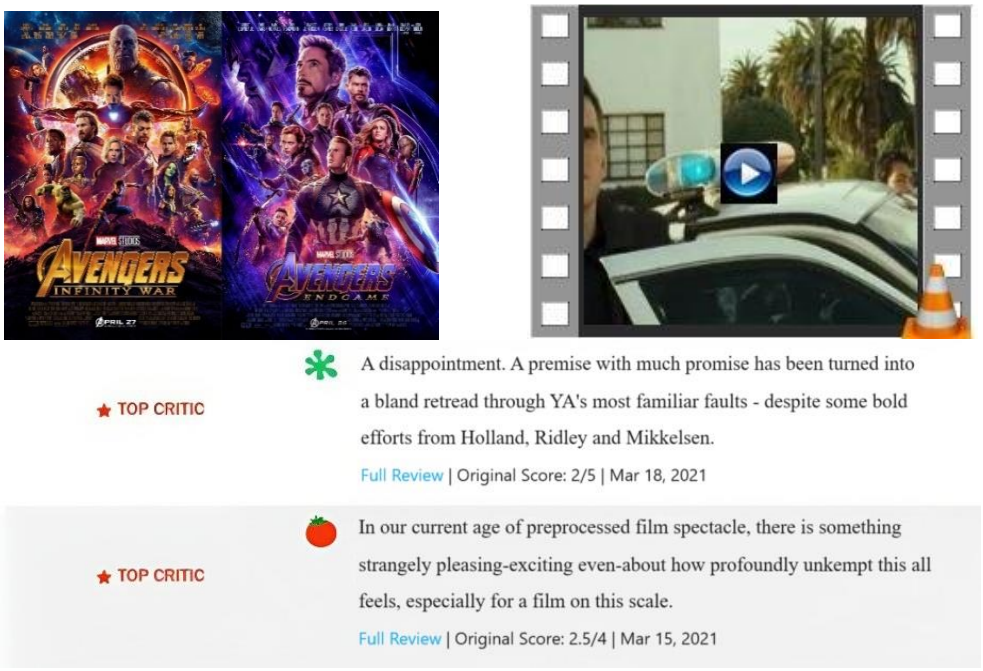

Fig. 1. Movie modalities: Poster (top left), Trailer (top right), and Reviews (bottom).

This research aims to address the shortcomings of prediction frameworks [13–17] that ignore deep multimodal features and only use regressions to assess the performance of the models.

The significance of our cross-modal, compared to prediction frameworks [13–17], lies in its ability to capture and analyze the dynamic visual elements of a movie's marketing campaign, which have become increasingly important in the digital age. Movie posters and trailers are essential components of a movie's marketing strategy, and they play a crucial role in generating audience interest and engagement, hence movie reviews. By using Cross-modal Transformers to analyze and extract features from these visual and textual elements, our model can provide valuable insights into the audience's emotional response to the movie, which can significantly predict its box-office success, which is overlooked by prediction frameworks [13–17].

Therefore, the theoretical perspectives underlying the proposed model's design that combines static and dynamic features from posters, trailers, and reviews using cross-modal transformers are rooted in machine learning and marketing research.

From a machine-learning perspective, Vision Transformers believes that deep-learning models can learn to extract and analyze complex visual features from images, such as movie posters and trailers. Vision Transformers are particularly effective in capturing the spatial relationships between different visual elements in an image or image sequence, which is essential for movie revenue prediction [7]. This allows our model to capture the dynamic visual elements of a movie's marketing campaign, which can significantly predict its box-office success compared to prediction frameworks [13–17]. For reviews, the utilization of Natural Language Processing (NLP) techniques, for instance, Pre-Trained Language Models (BERT) [18], to obtain textual features as reviews are the backbone of any conversation about film as a key to helping moviegoers understand why they might spend on that film, or why they should not, which is overlooked by prediction frameworks [13–17].

Finally, from a marketing research perspective, the cross-modal's ability to combine static and dynamic features is based on the empirical finding that both are essential for predicting a movie's success. Studies have shown that while static features such as cast, director, and genre are significant predictors of a movie's success, dynamic features such as marketing campaigns, critical reception, and audience engagement also play a crucial role in driving box-office revenue. By combining these

perspectives and findings, the cross-modal can provide a more accurate and comprehensive understanding of a movie's box-office potential, leading to more informed decision-making and better movies.

The novelty of our model lies in its integration of innovative technologies from machine learning and computer vision with insights from market research. This approach can potentially revolutionize how the film industry approaches marketing and audience engagement, leading to better-informed decision-making and better movies that resonate with audiences and generate higher revenues.

By designing a cross-modal transformer for box-office prediction, this paper makes a double-value contribution:

- It combines static, dynamic, and textual features from posters, trailers, and reviews using Cross-modal Transformers. This paper is the first work that combines cross-modal data to improve the accuracy and reliability of movie box-office revenue predictions.
- By considering both the statistical features used in movie revenue prediction models, such as cast, director, and genre, as well as static and dynamic features extracted from movie posters and trailers using Vision Transformers and textual features from reviews, the cross-modal model provides a more significant and comprehensive understanding of a movie's potential success.

The remainder of our study is organized as follows: Section II is the literature review, whereas Section III is our 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 gives an 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

Several studies are still pushing to develop prediction models due to low accuracy with the current predictions of movie box-office revenues. Most current revenue forecast models for opening weekend box office are categorized into prediction algorithms [6–8, 13–16] or the metadata [5, 14, 19] associated with the films. Developing a cross-modal box-office revenue prediction model that combines static and dynamic features from posters, trailers, and reviews for movie sentiments using Cross-modal Transformers has emerged as a promising solution. Below are categories of related works.

A. Algorithm-Based

Ni *et al*. [8] suggested the LightGBM model by evaluating the predictive value of model features. Wang *et al.* [20] offered a box office theoretical framework for motion pictures that utilized an extensive network of values, a novel idea for Barry [13]. Sharda *et al.* [9] were the first to apply artificial neural networks in movie box office forecasts. Using the fusion theory, Liao *et al.* [14] suggested a stacking framework for movie revenue prediction that included XGBoosting, RF, light gradient boosting machine, and k-NN. Tang *et al.* [15] created a Multi-Evidence Dynamic Weighted Combination Forecasting framework based on machine learning approaches, suggesting a complex version of the combined technique at the Chinese movie box office.

B. Variable and Feature-Based

Extrinsic variables related to the box office performance of a motion picture rather than the film itself, such as marketing methods, seasons, holiday effects, and other competitor movie screens, are crucial when determining the public's preferences. Nonetheless, most of these characteristics were inconsistent and were contingent on external factors such as customer demography and the duration of the investigation. Liao *et al*., Lash *et al*., and Mestyán *et al.* [14, 19, 21] involve variable selection, and various research investigations have used certain variables in predictive models and examined their significance. Hur *et al.* [22] used movie-related variables and several parameters from the film. Shambharkar *et al.* [23] demonstrated classifying movie trailers using human action based on an improved CNN in video sequences. They primarily transformed images to grayscale and pre-processed them through adaptive median filtering. Matsuzaki *et al*. [24] used handcraft techniques to investigate the information a machine-learning-based model can extract from still images.

C. Multimodal Feature-Based

Zhou *et al.* [17] are the first researchers to employ ANN "*Artificial Neural Networks*" to predict movie box-office revenue using multimodal features extracted from movie poster and movie metadata. Using big data, Wang *et al*. [13] offered a strategic analysis of key elements and analyzed their features to establish a novel basis for forecasting a motion picture's income. Their model was constructed in two phases: a complex heterogeneous network embedding model capable of eliciting a high informational level of film quality from previews and a deep neural network model focused on the cortical network. Ahmed *et al.* [25] presented eighteen additional features to improve their model performance on the agreement between the related parties (director and cast). Sahu *et al.* [4] developed a methodology incorporating sentiment analysis and a hybrid recommendation system for promoting unreleased movies for which a trailer has been released. Montalvo-Lezama *et al.* [26] proposed a state-of-the-art Dual Image and Video Transformer Architecture (DIViTA) for multi-label genre classification of trailers. Movie trailers have specific research analysis objectives but present various technical challenges. The challenges include trailer semantic content design [27], face detection and tracking, and action recognition in movie scenes featuring well-known film superstars from trailers [28]. These were previously overlooked, but due to the rising achievement of artificial neural networks in other study areas, most notably computer vision, researchers in other domains are investigating their effects. Oh *et al.* [29] were

the first to investigate trailers' impact on box office revenue using statistical analysis. Tadimari *et al.* [30] also studied how movie trailers invoke viewers' interest and curiosity, positively impacting the movie's financial future. Rahim *et al.* [31] used data mining algorithms to mine data from YouTube trailers and how it can be used to forecast a film's gross revenue. Finsterwalder *et al.* [32] investigated movie trailers and their different forms (regular, teaser, and TV ads), history, position, and other promotional techniques. Ahmad *et al.* [3] suggested predicting a movie's opening box office income by mining viewers' intent to buy a movie ticket from trailer reviews.

In this context, a cross-modal prediction model that combines static and dynamic features from posters, trailers, and reviews using Cross-modal Transformers has emerged as a promising approach. Transformers are a deep learning model with remarkable performance in image classification tasks. Combining static and dynamic features extracted from posters and trailers and textual features from reviews using Transformers, our model can better predict a movie's box-office revenue. This innovative approach to box-office revenue prediction has the potential to revolutionize the film industry, giving studios and analysts a more reliable tool to forecast a movie's success.

Multimodal features significantly improve the predictive abilities of a cross-modal transformer-based model for box-office revenue prediction compared to unimodal prediction frameworks [13–17]. Utilizing several modalities like textual reviews, visual posters, visual trailers, and metadata offers a more detailed and complete understanding of the patterns affecting box-office performance. Traditional statistical features are helpful within statistical-based models but may lack the comprehensive perspective for compelling predictions. Utilizing multimodal features enables the model to comprehensively analyze many elements of audience preferences, attitudes, and contextual information simultaneously, resulting in a more intricate comprehension of the complex components influencing box-office performance. A cross-modal transformer can enhance its adaptability to changing trends and improve predictive accuracy in box-office revenue prediction by utilizing information from different sources to exploit synergies between modalities and reduce the influence of noisy data.

III. Materials and Methods

A. Dataset Collection

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, Box-office Mojo, and The Movie Database (TMDB). Most of the Hollywood movies in these databases, from which we retrieved 50186 movie metadata records, were from the 1990s to 2022. MovieNet dataset [33] 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. The movie trailers dataset contains multi-labels with ten classes corresponding to film genres. The ResNet is a transfer-learned architecture pre-trained on the well-known ImageNet dataset and Open Image dataset V6 [34] that provides exhaustive annotation for all object instances and is fine-tuned using the state-of-the-art MovieNet dataset. We trained an LSTM model on the Trailers-Dataset [26] from scratch in the second phase, and we used the YouTube-8M dataset [35] for video classification. 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 [33]. The CMU-MOSEI dataset [36], which contains 23,500 sentence expressions, was used to fine-tune our textual model.

B. Dataset Pre-processing

We assessed the absolute number of movies collected using specific filtering algorithms and rejecting motion pictures with missing data. Altogether, 50186 films were pre-processed and evaluated in our research, often grouped into five classes based on their opening box office earnings (Blockbusters $\Longrightarrow$ Flops). As a result, we assigned the reference [10] 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…" [7]. Therefore, the number of consumers (moviegoers) or the monetary value (earnings) is used to compute the movie box office [22]. In our work, we developed a model to estimate monetary value, forecasting a film's demand before its theatrical debut. Table I shows the resulting revenue classification. As it is critical to reflect the time worth of money, inflation-adjusted budgets were adopted.

TABLE I. Earnings Classification

Class	Opening Revenue/Earnings $ million
1	Earnings > 100 (Blockbuster)
2	95 < Earnings ≤ 100
3	85 < Earnings ≤ 95
4	70 < Earnings ≤ 70
5	Earnings ≤ 70 (Flop)

Using revenue statistics as our primary data, we analyzed our movie metadata by eliminating duplicate entries and missing data and excluding unnecessary columns. 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 of (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

These datasets test, train, and validate our frameworks. Fig. 2 shows the visuals of the datasets used from the Trailers dataset [26]. Fig. 3 displays the MovieNet dataset used to fine-tune and validate our model.

Fig. 2. Movie trailers dataset.

Fig. 3. MovieNet dataset.

The MovieNet dataset was used to fine-tune our models due to its comprehensiveness, "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" [33].

2) Data forms and variables

Fig. 4 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 II shows the final variables which were selected. Impact values for the Cast and Crew were determined similarly to those in the reference [15] as Actor/Actress impact a_i (i = 1,2,3,…), leading actor or actress, and co-star actor or actress:

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

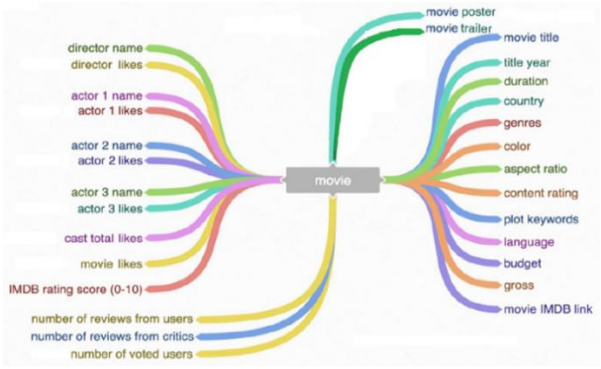

Fig. 4. Movie variables scrapping.

where t_i = a collection of films in which an actor or actress i appears, 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 (2), is the actor or actress quantity signifying the role rank performed by the actor or actress i in the movie t is:

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

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 calculated as:

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

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} week in cinemas of the t movie directed by the director.

In our study, we utilize continuous data types to enhance data accuracy. All values are considered equally important, with the only variable distinguishing importance being the ones related to market needs, as indicated in Table II. We have used discrete and continuous data formats as variables except for the movie trailers and genre. The genre expansion of a movie is represented as a vector, where each dimension (style, comedy, target, and narrative) is normalized to zero, classifying a movie into many genres. A fixed vector configuration was implemented for each genre.

TABLE II. DATA TYPES AND VARIABLES

Variable	Definition	Form
Box office [a]	It could refer to the media and entertainment industry's GDP regarding audience size or revenue.	Numeric
Movie Genre [b]	A feature film's classification is based on similarities in the narrative or emotional audio-visual sentiments: drama, sci-fi, action, and biography.	Vector
Crew & Cast	The impact of famous personalities and internationally renowned filmmakers. Honors received contribute to the total prestige computation of star impact levels.	Numeric
Release date & 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
Movie Poster	To promote and advertise a film, attract potential consumers, and cultivate their appreciation for it before its debut in theaters.	Feature Vector
Reviews	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 before its premiere in the cinemas.	Feature Vector

[a] Dependent Variable. [b] Control variable.

C. Proposed Methodology

This research paper suggests a Cross-modal Transformer and a Hierarchical Fusion Neural Network Model (CHFNN), an improved model of our earlier study [7]. Fig. 5 depicts the CHFNN architecture that combines static and dynamic features from posters, trailers, and reviews, utilizing Cross-modal Transformers and a Hierarchical Fusion Neural Network. We propose a framework based on deep multimodal features, which extracts and learns poster features, trailer-based high-level representations, and textual features from movie reviews. The poster features uncovered include the appearance (i.e., background, genre-basic objects, scene, aesthetics, color, and texture). Trailer features uncovered include the filming quality, narrative, filming styles (i.e., the shooting quality affects the audience's perception), and motion features of movie casting. Review features offer significant insights into the audience's sentiment, critical reception, and overall discourse about a film. Utilizing our 3D ResNet function is advantageous in effectively capturing the latent visual static and motion features in posters and trailers. Extracting these distinctive characteristics from movie posters and trailers, which effectively capture the public's interest and awareness of a film, enhances the precision of the revenue forecasting model.

The end-to-end Cross-modal Transformer with a Hierarchical Fusion Neural Network Model (CHFNN) is divided into four architectures: 1. feature extraction from movie posters, 2. text encoding of movie reviews, 3. feature extraction from movie trailers, 4. cross-modal feature fusion and prediction.

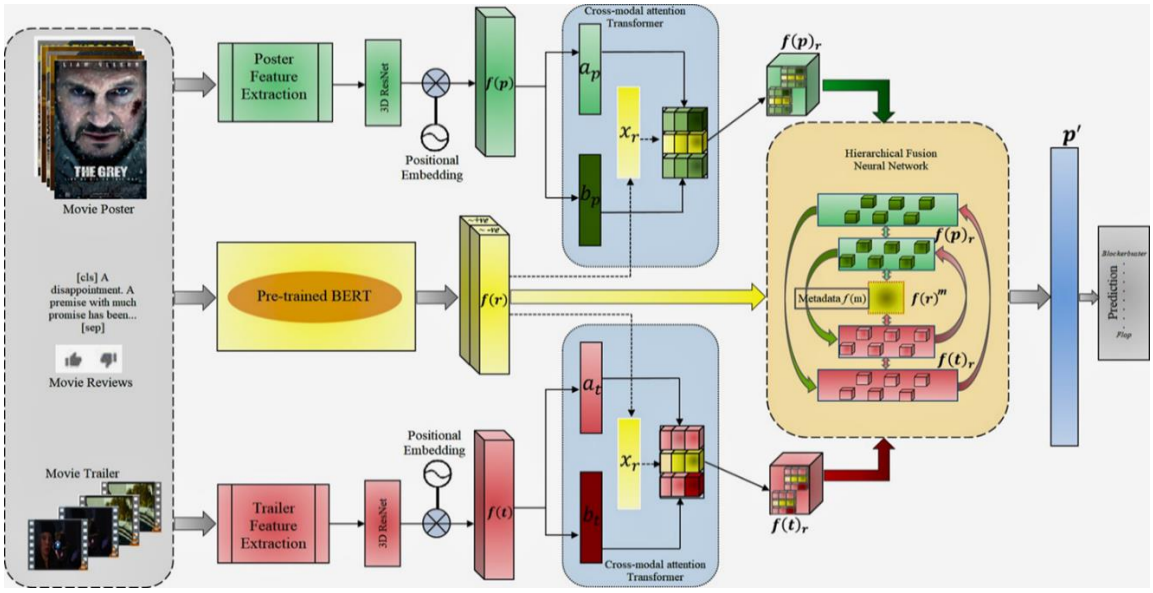

Fig. 5. Cross-modal transformer with hierarchical Fusion Neural Network (CHFNN).

1) Poster feature extraction

Using the novel ResNet50+ViT model (Fig. 6), we extract visual cues from posters to extract information about the aesthetic attributes and hidden revenue-related visual features such as dominant color, edges, blobs, and textual information on the poster. As standardized dimensions, we scaled our image posters to a uniform size of 224×224×3. This study employed a pre-trained ResNet model to mine visual static features from movie posters. Specifically, we obtained feature vectors from the last fully connected layer of ResNet, which effectively captured the visual attributes of each movie poster. The model is pre-trained on ImageNet and Open Image databases V6, according to Madongo *et al.* [7]. The utilization of Vision Transformers (ViT) [37], which has undergone fine-tuning on MovieNet [33], enables the extraction of revenue-related visual semantics *AP* and other object representations b_p from movie posters. These extracted visual features, *f(p)*, are the final poster discriminative feature vector extracted.

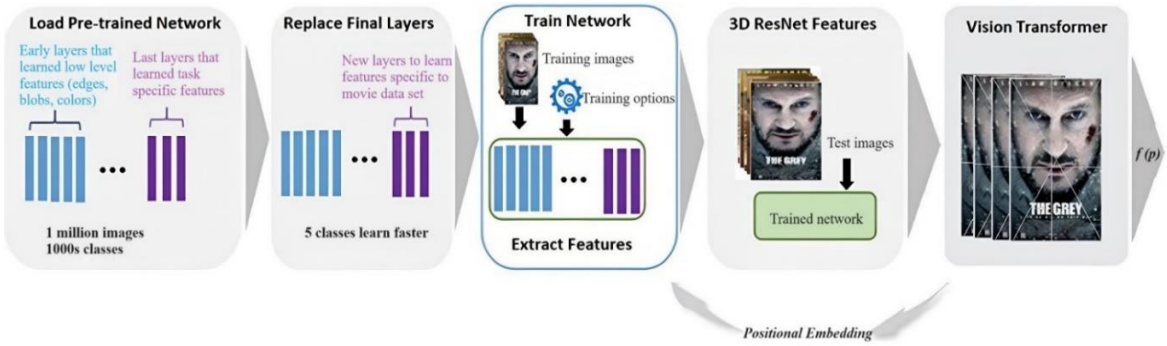

Fig. 6. Movie poster feature extraction (ResNet+ViT).

2) Trailer feature extraction

We designed an emerging transfer learning neural network ResNet+LSTM configuration (Fig. 7) to extract data and output the distinctive movie trailer features. It examines the role of temporal regression of visual content in predicting movie revenue. We decomposed trailers into numerous keyframes as input for our pre-trained model feature extraction. We extract features from these keyframes through state-of-the-art detectors and descriptors [38]. 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 used 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 [39–43] 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 [44].

In contrast, a model pre-trained and fine-tuned on MovieNet data extracts features that characterize scenes and ambient, providing context for specific elements and learning mid-level aspects of the shooting component (movie quality/quality and plotline/narrative) (see Fig. 7).

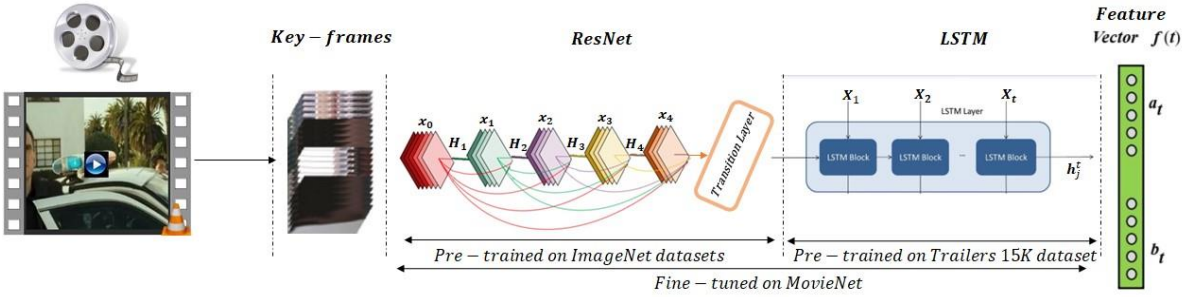

Fig. 7. Movie trailer feature extraction (ResNet+LSTM).

Keyframe features are derived using a PTM ResNet [45], a primary network with distinct feature diversity characteristics. Then, long-range temporal dynamics relationships are discovered by combining the frame-level data with the LSTM blocks. Trailer feature embeddings for each hidden layer h^t_j of the corresponding j^{th} LSTM block as output *f(t)* As indicated in Fig. 7, the proposed approach is built on ResNet+LSTM features, each designed to learn a distinctive feature of the films. Before training the ResNet+LSTM, the pre-trained ImageNet model [46] and the data from Trailers-Dataset [26] are loaded into the model memory. From there, the model is fine-tuned on a comprehensive movie dataset, MovieNet [33]. The model is dedicated to learning various trailer segments.

Deep visual trailer embeddings are extracted from trailer keyframe 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 keyframes, such as background and genre-basic objects. The ResNet+LSTM block's final output is a five-dimensional vector indicating the trailer segments' feature.

3) Text encoding using transformers

Process the textual reviews by tokenizing them, utilizing pre-trained language models to obtain review embeddings. The movie reviews are pre-processed using natural language processing methods, including tokenization, stop word removal, and stemming through embedding by appending two unique elements, [*cls*] and [*sep*] (see Fig. 8).

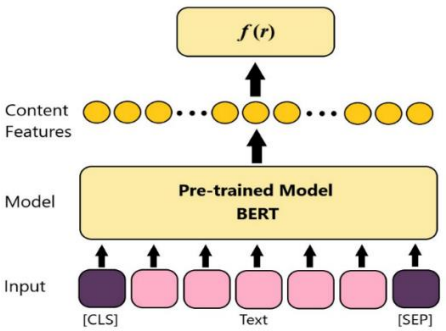

Fig. 8. Movies reviews encoding (Pre-trained BERT).

$$r_i = \{[\boldsymbol{cls}], w_1, w_2, w_3, \ldots w_n, [\boldsymbol{sep}]\} \quad (4)$$

We employed a pre-trained transformer model, BERT [18], to encode the movie review text into dense representations. Adjusted the pre-trained transformer models using the movie review dataset to tailor them to predict box office revenue. We fine-tuned the 12-layered transformer BERT model pre-trained to extract text features ***f(r)***, where r = initial review text sequence and θ_r^{bert} = hyperparameters of the pre-trained BERT.

$$f_r = BERT(r_i, \theta_r^{bert}) \in R^{d_r}, i \in [1, n] \quad (5)$$

4) Cross-modal feature fusion and prediction

Our Cross-modal Transformer architecture effectively combines the different modalities (posters, trailers, and reviews). The architecture has separate branches for each modality, allowing the model to learn the representations independently. It incorporates cross-modal attention mechanisms to capture the interdependencies between the different modalities. This way, information from posters, trailers, and reviews can be fused and combined to make predictions. Once modality-specific features are identified, the architecture incorporates cross-modal attention mechanisms to capture the interdependencies between different modalities. This allows the model to learn and fuse the relationships between visual elements (posters and trailers) and textual information (reviews). To manage multimodal data effectively, we employ a cross-modal transformer. The Cross-modal attention transformer ($C_{attention}$) simultaneously learns additional hidden features of the movie (i.e., quality and the narrative) and combines features from movie posters, trailers, and reviews. It takes the multimodal features as input and learns to address the most informative elements across different modalities. It effectively integrates the information from trailers, posters, and reviews, allowing the model to identify complementary patterns and cues that impact box office performance.

We extract the visual features from the movie poster head and trailer head and the textual embeddings from the movie review head to create a comprehensive representation of each movie, where a_p, b_t, and x_r are key vectors and values for all the modalities (poster, trailer, and review) and T represents transpose and d_n = dimensionality of the modality as follows:

$$C_{-attention_{p-r}(a,b,x)} = softmax \odot \left(\frac{a_p b_r^T}{\sqrt{d_n}}\right) x_r \quad (6)$$

$$C_{-attention_{t-r}(a,b,x)} = softmax \odot \left(\frac{a_t b_r^T}{\sqrt{d_n}}\right) x_t \quad (7)$$

5) Cross-modal feature fusion layer

We propose a Hierarchical Fusion architecture that aggregates information at distinct levels of granularity.

The model first processes individual modalities (posters, trailers, and reviews) separately, and then it further fuses the modality-specific representations to capture higher-level relationships and interactions of visual and textual features into a unified feature representation. We connect all the heads of the modalities as follows:

$$head_y^i == attention_y(x_r W_i^{x_r}, a_p W_i^{a_p}, b_p W_i^{b_p}) \quad (8)$$

$$P_y^* = MultiHead(x_r a_p b_p) \quad (9)$$

$$P_y^* = \odot (head_y^1, head_y^2, head_y^3, \dots, head_y^n) w_y^0 \quad (10)$$

$$f(p)_r = MultiHead(P_y^*; \theta^{c-att}) \quad (11)$$

$$f(t)_r = MultiHead(P_y^*; \theta^{c-att}) \quad (12)$$

$$P_y = \odot \left(f(p)_r, f(t)_r, f(r)^m \right) \quad (13)$$

where w = weight matrix of each modality and all connected modality heads = P^*_y and w^0_y = weight matrix applied after combining the head of y modalities and Θ = fusion operator, w^0_y is the weight matrix multiplied after the splicing the head of y modalities and n = the number of self-attention heads.

We employ the fusion framework described in [7] to combine the visual and textual features into a unified feature representation. Therefore, the cross-modal feature representations $f(p)_r$ and $f(t)_r$ are input to the recurrent fusion network, and the movie metadata $f(r)^m$ is calculated similarly in reference [7], i.e., cast and crew influence, director influence, and release date influence. The recurrent fusion neural network calculates the fusion weights of the respective modality ω_i similar to [7] to ensure that each modality is given a significant weight, resulting in a combined feature:

$$P_y \equiv P = \sum_{i=1}^{N} \widehat{\omega}_i P_i$$

By doing this, the latent data among the diverse modalities is entirely recognized, and the cyclic mechanism is used to eliminate redundant and irrelevant information. Especially reviews because some are ambiguous, and it is hard to interpret the audience sentiment, eventually impacting the opening box office performance.

6) Prediction layer

The prediction procedure has been simplified to a classification problem of classifying a movie into the correct class using their opening box office shown in Table I and calculated similarly to [7]. The recurrent fusion neural network P^*_y output is fed into the fully connected layer $\boldsymbol{P'}$ to predict box-office revenue based on the fused feature representation—the final representation of the classification task of P^*_y.

$$P_y^* = ReLu\left(w_{l1}^{y^T} \otimes P_y + \mathcal{P}_{l1}^y\right) \quad (14)$$

$$\boldsymbol{P'} = ReLu\left(w_{l1}^{y^T} \otimes P_y^* + \mathcal{P}_{l5}^y\right) \quad (15)$$

The features are fused to predict the film's box-office revenue, and combining numerous features is critical, as various features are complementary. An iterative method for optimizing the learning rate 0.1 for smoother convergence is applied to adjust the objective and ***relu*** activation functions.

IV. Result and Discussion

A. Performance Metrics

We employed the Average Percent Hit Rate (APHR) [9], Root Mean Square Error (RMSE), and Accuracy/Precision (ACC) as metrics to evaluate and calculate the accuracy of our network configurations. Two types of APHR are used to measure different class performance metrics:

- *Absolute Accuracy*: Calculates the precise result (B*ingo*). Quantifies the number of accurately predicted classes.
- R*elative Accuracy*: The concept of 1-away predictions refers to predictions that are only 1 class different from the actual outcome.

Average Percent Hit Rate (APHR) is expressed as:

$$APHR = \frac{Accurately\ predicted\ number\ of\ samples}{Total\ class\ sample\ number} \quad (16)$$

$$APHR_{Bingo} = \frac{1}{n} \sum_{i=1}^{c} M_i \quad (17)$$

$$APHR_{1-away} = \frac{1}{n} \sum_{i=1}^{c} (M_{i-1} + M_i + M_{i+1}) \quad (18)$$

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:

$$\textbf{Acc}\text{uracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (19)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (20)$$

Accordingly, these values are denoted as true positives, true negatives, false positives, and false negatives. The Root Mean Square Error (RMSE) is a statistical measure that quantifies the average deviation between the predicted values of a model and the actual values in the context of image and video classification:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (21)$$

$\hat{y}_i = i^{th}$ 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.

B. 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" [7] on the Linux operating system. We used ViT [37] experimental setup during the image and video classification tasks. The experimental setup aimed to provide comparable results between pre-training and fine-tuning, focusing on the performance during fine-tuning. During the training process, the weights were improved using the Stochastic Gradient Descent (SGD) technique, with a small batch size of 64.

Upon thoroughly testing the deep learning parameters, we have ascertained that the optimal values for the degree and quantity of maximums are 0.0001 and 64, respectively. These parameters were selected to encompass all pertinent data for optimal training. The loss function is determined by binary cross-entropy. Table III shows our experimental setup and the hyperparameters used in the models.

TABLE III. 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
	Hierarchical Fusion Network	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 10 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

Initially, we started with a small value of 10 frames for retrospective analysis and closely observed the performance of our model. This gave us an initial reference point to assess if the model effectively captures pertinent temporal information.

Afterward, we performed trials using various numbers from 20 to 50 and carefully observed our model's performance changes. This experiment was conducted as a component of a hyperparameter optimization process. Our model encountered challenges in accurately detecting significant patterns at a frame rate of 20. When the model reaches 50 frames, it can apply its knowledge to new trailers and understand the relationships between events that occur over a lengthy period. This leads to better performance without producing any problems with computational resources. We used this as the optimal value for our specific datasets: Trailer 15k [26], ImageNet dataset [46], Open Image Dataset V6 [34], YouTube-8M dataset [35], and the MovieNet dataset [33]. We utilized this approach for fine-tuning the MovieNet dataset, selecting a value that optimizes the trade-off between capturing significant temporal relationships and maintaining computational efficiency. We employed an iterative approach, utilizing cross-validation and hyperparameter tuning to determine our dataset's optimal look-back value and prediction objective.

C. Training and Testing Data Split

We pre-trained our visual features model with the following datasets: Trailer 15k [26], ImageNet dataset [46], Open Image Dataset V6 [34], and fine-tuned our models with the MovieNet dataset [33], which provide exhaustive box annotation for all instances. Tables IV–VI show the Trailer 15k dataset [26], YouTube-8M dataset [35], which comprises frame-level features for over 1.9 billion video frames and 8 million videos used for video classification tasks and Open Image Dataset V6 [34] partition.

Unless specified otherwise, we employed configurations obtained from experiments, including various alternative neural network structures and training parameters (51%, 14%, and 34%) for training, validating, and testing datasets.

The reason for using a 51% training set is that it is crucial for our integrated models, which have a complex structure and large datasets. The 14% validation set is

employed to fine-tune hyperparameters and check the model's performance throughout training. This smaller fraction was designated for validation to ensure enough examples for evaluation without sacrificing the size of the training set.

The testing set, comprising 34% of the data, is segregated until the conclusion of model training. This collection aimed to evaluate the model's performance on unseen data, thereby offering an unbiased evaluation of its generalization capacity. The increased dimensions of the validation process contributed to a more comprehensive and reliable evaluation.

Following our prior research [7], we have introduced a dependable and methodical approach for doing *k*-fold = (5-times) cross-validation. In this approach, the folds are created based on collective concepts rather than random selection. The complete dataset (*D*) is partitioned into *k* distinct classes D_1, D_2, …, *and* D_K, each having the same size. The classification model is subsequently trained and evaluated using *k*-fold cross-validation. The *k*-fold cross-validation technique was employed to obtain more reliable performance estimates, mainly when dealing with restricted-sized datasets. Ultimately, we employed this approach, considering the compromises between acquiring sufficient training data, optimizing the model, and collecting dependable performance measurements.

TABLE IV. MULTI-LABELED TRAILER DATASET PARTITION

Revenue 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

TABLE V. VIDEO DATASET PARTITION

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

TABLE VI. OPEN IMAGE DATASET V6 PARTITION

Dataset	Training	Validation	Testing	Total
Images	1,743,042	41,620	125,436	~
Boxes	14,610,229	303,980	937,327	600

D. Experimental Results

The current state-of-the-art box-office revenue prediction techniques include a Content-based model [5], DMFCNN [7], multi-model ensembles [8], PRBO [13], Stacking Fusion Model [14], evolving DNN (mean and best) [16], Deep Neural Network (DNN) [17], Hybrid Features [19], Hybrid Model [25], ensemble method [47], Self-MM [48], MGHF [49], SPECTRA [50], Custom SVM-based methods [51], KHDEM [52] and SHAP [53]. We compared our model and features against these SOTA models and existing state-of-the-art textual and visual features for box office prediction.

1) Textual features box-office prediction results

Table VII shows the performance comparing state-of-the-art textual features for box office prediction with ours. Textual features also performed better than existing methods, with a precision rate of 88.25%, owing to our comprehensive and erudite textual feature-extracting method. The CMU-MOSEI dataset [36], which contains 23,500 sentence expressions, was used to fine-tune our textual model.

TABLE VII. TEXTUAL FEATURES PERFORMANCE ON CMU-MOSEI DATASET

Textual Feature Vectors	Acc/Precision (%)
Self-MM [48]	53.32
MGHF [49]	53.70
SPECTRA [50]	87.34
CHFNN	**88.25**

2) Visual features box-office prediction results

In addition, we conducted a comparative analysis of our model with state-of-the-art multi-model ensemble techniques, employing the Root Mean Square Error metric for evaluation. The CHFNN feature demonstrated superior performance to all ensemble models, as seen in Table VIII.

TABLE VIII. VISUAL FEATURE ROOT MEAN SQUARE ERROR PERFORMANCE

Visual Feature Vectors	RMSE
XGBoosting [8]	19,137.7581
LightGBM [8]	16,105.0243
Stacking [8]	17,974.5699
CHFNN	**11,984.6481**

More importantly, the effectiveness of the multimodal feature vector has been attributed to its extensive and discriminative hidden features derived from movie cinematic assets. This demonstrates our method's efficiency for feature extraction and fusion. The findings indicate a significant correlation between the efficacy of the suggested model in acquiring and discerning movie cinematic assets and the box office revenue generated before the theatrical premiere of the film.

3) Model ablation performance results

Table IX shows the ablation study; combining each model at every stage improves performance. Accuracy continually improves from 62.16% to 85% for AHPR and from 17,001.232 to 12,659.563 for RMSE, demonstrating the usefulness of our learning architecture and the process of optimizing the parameters during experiments.

TABLE IX. ABLATION PERFORMANCE EVALUATION OF THE MODEL DESIGN

Model Design	RMSE	AHPR (Bingo) %
ResNet + ViT (Poster Features)	17,001.232	63.17
ResNet + LSTM (Trailer Features)	16,845.235	53.49
Cross-modal Transformer (Visual + Textual Features)	15,568.235	65.78
Hierarchical Fusion Neural Network (before fine-tuned)	**12,659.563**	**94.13**

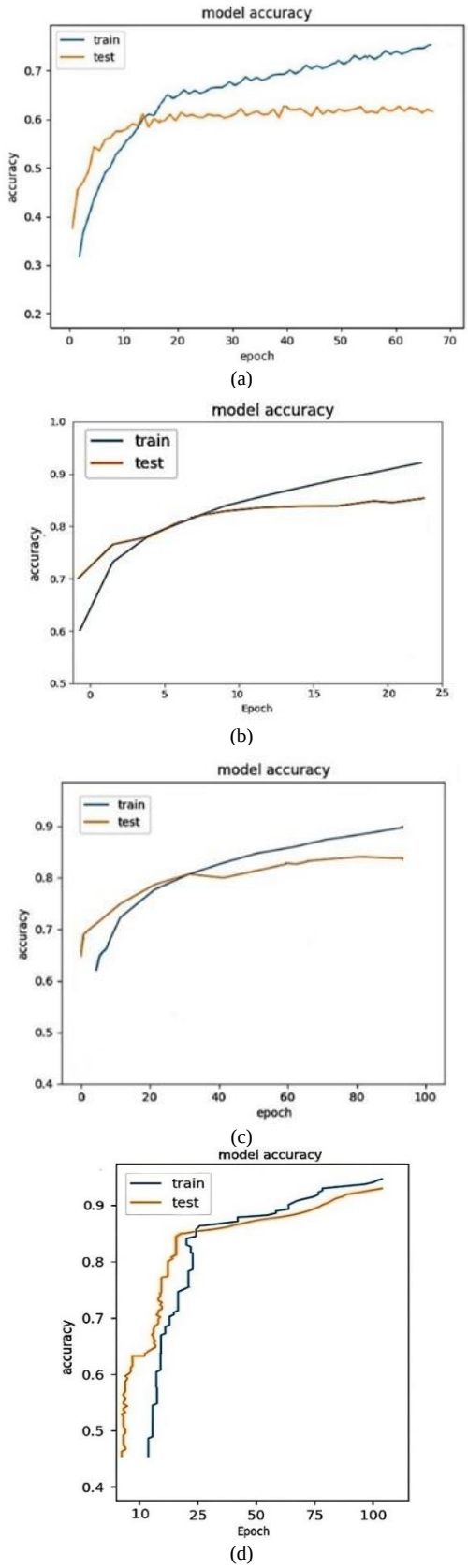

Fig. 9. Training and testing model accuracies; (a). ResNet+ViT, (b) ResNet+LSTM, (c) Cross-modal transformer, (d) Hierarchical fusion network.

Figs. 9 and 10 demonstrate each model's training and testing accuracies. The ResNet + ViT model (top left) extracted high-level and discriminative revenue-related visual semantics and object representation features from movie posters to learn and classify movie box office revenue and achieved an accuracy of 62.76%. These deep visual trailer embeddings are extracted from the trailer keyframe used in the ResNet+LSTM model (top right) and had an accuracy of 80.23% during testing. The performance is slightly improved to 82.10% for the Cross-modal attention Transformer (bottom left) for dynamic visual and textual features. The effectiveness of introducing the Hierarchical Fusion Neural Network (bottom right) improved to 85%, which further refines the fused features, enabling a deeper understanding of the inherent hierarchical structure of multimodal data.

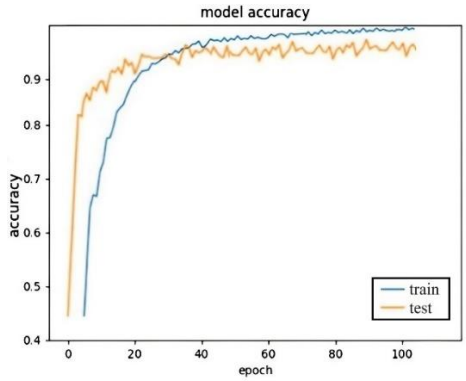

Fig. 10. Training and testing CHFNN model accuracy.

The fine-tuned CHFNN model testing performance is exceptional, as shown in Fig. 10, achieving a high accuracy of 95.80%. This result is obtained by including the hyper-parameters of the complete architecture into our fully linked Cross-modal Transformer and a Hierarchical Fusion Neural Network model (CHFNN). The model parameters and hyperparameters were fine-tuned using the MovieNet dataset. This was accomplished by modifying the adapters' weights while maintaining the model's unchanged weights.

Our cross-modal transformer model addressed causation complexity and overfitting issues using robust causal inference techniques and regularization methods. By incorporating causal models, actual relationships between different modalities are identified, which helps to prevent false correlations and improve the interpretability of the model. The approach used sentiment analysis and attention techniques to capture essential opinions, sentiments, and context in textual reviews and handle subjectivity. This enhanced comprehension of moviegoers' preferences and reduced the influence of subjective language on predictions.

An effective Hierarchical Fusion Neural Network was used to address cross-modal issues, enabling the smooth integration of static and dynamic visual features and textual features from posters, trailers, and reviews. We utilized multimodal pre-training and joint embeddings to facilitate the model's acquisition of strong cross-modal representations, allowing it to synergistically utilize information from textual reviews, metadata, trailers, and

posters. Data quality and bias issues were handled by thorough pre-processing, quality assessments, and a fairness-aware algorithm to reduce biases in the training data and predictions. Furthermore, including an extensive and inclusive MovieNet dataset enhances the precision and impartiality of the model during the fine-tuning phase.

The model incorporated recurrent neural networks and attention mechanisms to capture changing patterns. Consistent model updates and retraining using real-world data ensured our model remained responsive to evolving trends and dynamics. External aspects, such as world events or marketing plans, were considered by integrating external metadata sources and collaborating with professional movie reviewers. This allowed our model to consider its predictions in the larger external context and improve its ability to handle unexpected factors. The predictive accuracy for box-office revenue predictions was enhanced using a cross-modal transformer-based model that addressed these difficulties thoroughly.

Box-office revenue prediction results Table X compares existing box-office baseline models and the proposed model at each testing set. 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. Our suggested framework exhibits superior performance compared to many existing box-office baseline models.

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

Model		1(%)	2(%)	3(%)	4(%)	5(%)	Average (%)
†Multi-model ensemble [8]		–	–	–	–	–	89.93
†Content-based [5]		–	–	–	–	–	96.80
Custom SVM-based [51]		–	–	–	–	–	96.70
KHDEM [52]		–	–	–	–	–	96.04
SHAP [53]		–	–	–	–	–	79.00
Gradient Boosting [54]		–	–	–	–	–	92.40
Hybrid Features [19]		–	–	–	–	–	86.30
Stacking Fusion Model [14]	Bingo	–	–	–	–	–	69.16
	1-Away	–	–	–	–	–	86.46
Hybrid Model [25]	Bingo	–	–	–	–	–	82.00
	1-Away	–	–	–	–	–	95.00
DNN [17]	Bingo	53.22	50.00	54.56	50.44	52.78	52.20
Evolved DNN(best) [16]	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 [13]	Bingo	45.00	47.00	54.00	61.00	63.00	61.80
DMFCNN [7]	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 CHFNN	Bingo	**85.05**	**86.27**	**86.73**	**88.61**	**93.30**	**95.80**
	1-Away	93.02	**94.58**	**95.94**	**97.82**	**98.97**	**99.15**

† Target Audience Prediction, not Revenue prediction

The experimental results validate the CHFNN model's practical applicability as it is supposed to benefits studios, investors, and production teams. 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 novel model, Cross-modal Transformer with Hierarchical Fusion Neural Network (CHFNN), had an average percent hit rate of 95.80% (bingo) and 99.15% (1-away), respectively.

4) Image and video classification results

We conducted a 5-fold cross-validation on our dataset to assess the importance of our model compared to other classifiers. Based on the data presented in Table XI, the *p*-value of 0.005 suggests that the model result is implausible to have occurred randomly compared to other classifiers. Additionally, the T-statistic of 2.95 indicates a significant difference between the model and the other classifiers. Combining a *p*-value of 0.005 and a T-statistic of 2.95 provides compelling evidence against the other classifiers. The conclusion suggests that the disparity in model proficiency is attributable to model design and predictive capability variations.

Though our primary task was movie box-office revenue prediction using trailers and metadata, we assessed our model performance on image and video classification tasks. For image classification in Table XII, the CHFNN 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) [37]. For video classification, we used the YouTube-8M dataset [35].

TABLE XI. SIGNIFICANCE T-TEST EVALUATION BETWEEN CLASSIFIERS

Classifier	SVM	Naive Bayes	Decision Tree	Random Forest	k-NN	LR	Proposed CHFNN
T-statistics	2.75	–3.20	–0.86	–2.40	–4.25	–1.43	2.95
p-value	0.234	0.264	0.398	0.125	0.421	0.983	0.005*

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

TABLE XII. PERFORMANCE EVALUATION ON IMAGE CLASSIFICATION BENCHMARKS

Models	ImageNet	ImageNet21k	CIFAR-100	CIFAR-10	Augmented Dataset
GRU [55]	76.22 ± 2.88	78.95 ± 0.66	–	–	–
ConvNets [56]	77.03 ± 2.78	79.30 ± 1.03	–	–	–
Vision Transformers (ViT) [37]	85.30 ± 0.02	88.62 ± 0.05	93.25 ± 0.05	99.15 ± 0.03	99.74 ± 0.00
Transformers [26]	77.49 ± 1.18	80.02 ± 0.47	–	–	–
CHFNN	**88.54 ± 0.02**	**95.54**	**94.81 ± 0.02**	**99.89 ± 0.01**	**99.93 ± 0.02**

Table XII shows that our Cross-modal Transformer (CHFNN) model on the YouTube-8M outperforms the pre-trained Vision Transformer (ViT) model while requiring significantly fewer CPU resources to train. ViT had (99.74 ± 0.00) as compared to CHFNN (99.93 ± 0.03), though the difference of (0.19) is not that big, Vision Transformers generally performs better when pre-trained on enormous public datasets; however, CHFNN outperforms baseline models when pre-trained on smaller datasets achieving (88.54 ± 0.02) on ImageNet, (95.54) on ImageNet21k, (94.81 ± 0.02) on CIFAR-100 and (99.37 ± 0.06) on CIFAR-10, respectively. CHFNN slightly exceeds baseline models when pre-trained on larger augmented datasets like YouTube-8M (see Fig. 11).

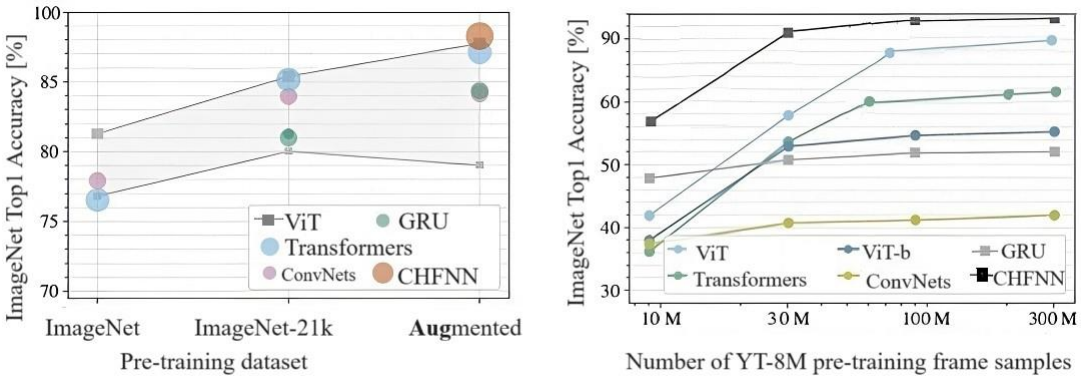

Fig. 11. Performance comparison of pre-training on benchmark datasets.

Fig. 11 shows that CHFNN outperforms ViT with fewer pre-training datasets (shaded region) but reaches a plateau faster. Vision Transformers overfit more than CHFNN on smaller datasets with equivalent computational costs. 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, CHFNN requires 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.

V. CONCLUSION

Our research paper suggested a Cross-modal Transformer and a Hierarchical Fusion Neural Network (CHFNN) model tailored for the opening box office predictions based on multimodal features extracted from movie trailers, posters, and reviews. We propose using a Cross-modal Transformer and a Hierarchical Fusion Neural Network, each of which aims to learn distinct parts of the movie's cinematic assets by applying several feature-learning representations to extract visual and textual features from the movie's multimodal data. We developed a methodology based on recurrent neural networks and vision transformers to extract visual features from posters and trailers and a pre-trained bidirectional encoder representations transformer to obtain textual features from review embeddings. Then, a feature vector for the film is constructed using these features and used as an input of a cross-modal attention transformer prediction model to predict the opening movie box office revenue. The retrieved visual elements encompass several aspects, such as the background, genre-basic objects, scene, aesthetics, color, and texture. The learned features effectively predicted the opening box-office earnings before its release in cinemas, achieving a correct class accuracy of 95.80% (bingo) and an accuracy of 99.15% for one class deviation (1-Away). The visual features showed a precision of 81%, while the textual features exhibited a precision of 88.25%. Through the integration of multiple modalities, our model has the potential to reveal significant insights that can improve the precision of box-office revenue projections, enabling industry experts to make more knowledgeable decisions. The visual features evoke a mood, mystery, intrigue, and expectation. The features profoundly influenced the audience's perception of the movie, as seen by their reviews, which captivated their attention and exerted a persuasive force. Our next objective is to enhance the accuracy of predictions by integrating diverse data, such as audio-visual and textual characteristics from movie abstracts and screenplays, utilizing Transformer encoders.

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 & Editing, Visualization. Zhongjun Tang: Supervision, Validation, Investigation, Resources, Writing—Review & Editing, Project administration, Funding acquisition. Jahanzeb Hassan: Data Curation, Writing—Review & 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] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, vol. 3, no. October, pp. 111–132, 2022. doi: 10.1016/j.aiopen.2022.10.001

[2] X. Han *et al.*, "Pre-trained models: Past, present and future," *AI Open*, vol. 2, no. June 2021, pp. 225–250, 2021. doi: 10.1016/j.aiopen.2021.08.002

[3] 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

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] D. Delen and R. Sharda, "Predicting the financial success of Hollywood movies using an information fusion approach," *Endüstri Mühendisligi Dergisi*, vol. 21, no. 1, pp. 30–37, 2010.

[12] 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

[13] 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

[14] 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

[15] 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

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

[17] 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

[18] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conference the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2019*, 2019, pp. 4171–4186.

[19] 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

[20] W. Wang, J. Xiu, Z. Yang, and C. Liu, "A deep learning model for predicting movie box office based on deep belief network," *Lecture Notes in Computer Science*, vol. 2, 2018. doi: 10.1007/978-3-319-93818-9_51

[21] 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

[22] 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

[23] P. G. Shambharkar and M. N. Doja, "Movie trailer classification using deer hunting optimization based deep convolutional neural network in video sequences," *Multimedia Tools* and *Applications*, vol. 79, no. 29–30, pp. 21197–21222, 2020. doi: 10.1007/s11042-020-08922-6

[24] 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. 15th IAPR International Conference on Machine Vision Applications, MVA 2017*, pp. 538–541, 2017. doi: 10.23919/MVA.2017.7986919

[25] 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

[26] R. M. Lezama, B. M. Lezama, and G. F. Pineda, "Improving transfer learning for movie trailer genre classification using a dual image and video transformer," *Information Processing & Management*, vol. 60, no. 3, 2023. https://doi.org/10.1016/j.ipm.2023.103343

[27] 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

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

[29] 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*, vol. 2015-March, pp. 1724–1732, 2015. doi: 10.1109/HICSS.2015.207

[30] A. Tadimari, N. Kumar, T. Guha, and S. S. Narayanan, "Opening big in box office? Trailer content can help," in *Proc. IEEE*

International Conference on Acoustics, Speech and Signal Processing,. 2016, pp. 2777–2781. doi: 10.1109/ICASSP.2016.7472183
[31] M. S. Rahim, A. Z. M. E. Chowdhury, 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
[32] J. Finsterwalder, V. G. Kuppelwieser, and M. 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
[33] Q. Huang, Y. Xiong, A. Rao, J. Wang, and D. Lin, "MovieNet: A holistic dataset for movie understanding," in *Proc. Computer Vision, ECCV 2020*, 2020, pp. 709–727. doi: 10.1007/978-3-030-58548-8
[34] A. Kuznetsova *et al.*, "The open images dataset V4: Unified image classification, object detection, and visual relationship detection at scale," *International Journal of Computer Vision*, vol. 128, no. 7, pp. 1956–1981, 2020. doi: 10.1007/s11263-020-01316-z
[35] S. A. E. Haija *et al.*, "YouTube-8M: A large-scale video classification benchmark," arXiv preprint, arXiv:1609.08675, 2016.
[36] A. Zadeh *et al.*, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, 2018, pp. 2236–2246. doi: 10.18653/v1/p18-1208
[37] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," arXiv preprint, arXiv:2010.11929, 2021.
[38] M. Y. Yang, X. Yong, and B. Rosenhahn, "Feature regression for multimodal image analysis," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 770–777, 2014. doi: 10.1109/CVPRW.2014.118
[39] 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. doi: 10.1109/CVPR.2015.7299101
[40] 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.
[41] 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
[42] 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
[43] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. F. Li, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, vol. 3. doi: 10.1109/CVPR.2014.223
[44] J. Wehrmann, R. C. Barros, G. S. Simoes, T. S. Paula, and D. D. Ruiz, "(Deep) Learning from frames," in *Proc. 2016 5th Brazilian Conference on Intelligent Systems, BRACIS 2016*, 2017. doi: 10.1109/BRACIS.2016.012
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90
[46] 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
[47] 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, 2020. doi: 10.1016/j.heliyon.2020.e04260
[48] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. 35th AAAI Conference on Artificial Intelligence, AAAI 2021*, 2021, vol. 12B, pp. 10790–10797. doi: 10.1609/aaai.v35i12.17289
[49] Z. Quan, T. Sun, M. Su, and J. Wei, "Multimodal sentiment analysis based on cross-modal attention and GATED cyclic hierarchical fusion networks," *Computational Intelligence and Neuroscience*, 2022. doi: 10.1155/2022/4767437
[50] T. Yu *et al.*, "Speech-Text Pre-training for spoken dialog understanding with explicit cross-modal alignment," in *Proc. 61st Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 7900–7913. doi: 10.18653/v1/2023.acl-long.438
[51] 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
[52] S. Sahu, R. Kumar, H. V. Long, and P. M. Shafi, "Early-Production stage prediction of movies success using K-fold cross deep ensemble learning model," *Multimedia Tools and Applications*, vol. 82, no. 3, 2023. doi: 10.1007/s11042-022-13448-0
[53] S. B. Kumar and S. D. Pande, "Explainable neural network analysis on movie success prediction," *EAI Endorsed Transactions on Scalable Information Systems*, 4435, 2024.
[54] 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
[55] 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
[56] 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.