Journal of Cloud Computing:
Advances, Systems and Applications

RESEARCH

Open Access

Enhancing multimedia management: cloud-based movie type recognition with hybrid deep learning architecture

Fangru Lin[1,2], Jie Yuan[1,2], Zhiwei Chen[2] and Maryam Abiri[3*]

Abstract

Film and movie genres play a pivotal role in captivating relevant audiences across interactive multimedia platforms. With a focus on entertainment, streaming providers are increasingly prioritizing the automatic generation of movie genres within cloud-based media services. In service management, the integration of a hybrid convolutional network proves to be instrumental in effectively distinguishing between a diverse array of video genres. This classification process not only facilitates more refined recommendations and content filtering but also enables targeted advertising. Furthermore, given the frequent amalgamation of components from various genres in cinema, there arises a need for social media networks to incorporate real-time video classification mechanisms for accurate genre identification. In this study, we propose a novel architecture leveraging deep learning techniques for the detection and classification of genres in video films. Our approach entails the utilization of a bidirectional long- and short-term memory (BiLSTM) network, augmented with video descriptors extracted from EfficientNet-B7, an ImageNet pre-trained convolutional neural network (CNN) model. By employing BiLSTM, the network acquires robust video representations and proficiently categorizes movies into multiple genres. Evaluation on the LMTD dataset demonstrates the substantial improvement in the performance of the movie genre classifier system achieved by our proposed architecture. Notably, our approach achieves both computational efficiency and precision, outperforming even the most sophisticated models. Experimental results reveal that EfficientNet-BiLSTM achieves a precision rate of 93.5%. Furthermore, our proposed architecture attains state-of-the-art performance, as evidenced by its F1 score of 0.9012.

Keywords Video classification, Deep learning, Service management, Cloud computing, Movie genres, Bidirectional LSTM, EfficientNet

Introduction

The ease of creating and distributing visual media, including photographs and videos, has led to their increasing prevalence as information carriers [1]. Major multimedia platforms like Netflix, attracting millions of viewers in the entertainment industry, have propelled image and video processing to new heights in recent years. Additionally, various methodologies have been employed to analyze image content [2], including picture classification [3, 4] and multivariate learning [5–7]. Furthermore, the attention mechanism in deep learning is crucial for video analysis due to the higher complexity and greater diversity inherent in videos [8]. For instance, a video may contain various scenes, different lighting conditions, and diverse camera angles, each of which may have an impact on the recognition of the video's genre. This complexity in video perception stems from the integration of temporal elements into the spatial components of the medium,

*Correspondence:
Maryam Abiri
M_abiri@sbu.ac.ir
[1] Weifang University of Science and Technology, Weifang 262700, China
[2] Dongseo University, Busan 47011, South Korea
[3] Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran 146899-5513, Iran

making video processing challenging to conduct solely through image-based techniques. While creating films from individual frames is a straightforward process, it often yields less precise outcomes.

One subfield of study concerning video comprehension is the classification of films based on preexisting concepts. Genres can be classified into specific instances [9, 10], and duties can be executed [11]. Several computer vision applications, such as video retrieval and recommendation systems, rely on video categorization [12, 13]. The implementation of convolutional neural networks (CNNs) in computer vision has led to significant progress [14]. Despite the considerable research dedicated to this subject, the categorization of films remains complex and requires further examination [15].

Classification of video content becomes challenging in identifying and classifying genres of trailers. Themes-based genre classification applies to the categorization of trailers, such as "dramas" or "comedies." Critics and consumers of the Internet Movie Database (IMDB) frequently classify films manually, ultimately determining the classification in the database. Trailers serve multiple purposes for major movie streaming platforms such as Hulu and Netflix, in addition to file categorization and film recommendation. Consequently, there is a growing need for algorithms capable of autonomously determining the genre of a given piece of content. Unlike alternative computer vision disciplines like activity detection and object tracking, this field encompasses a wide range of genres that intersect, giving rise to two fundamental challenges [16]. One primary concern pertains to the incapability of media sources to visually depict their disciplines. To identify the film's genre, it is imperative to view the entirety of the film, as opposed to focusing on a select few frames or moments. Another significant challenge in genre classification is the application of multiple labels [17]. A multitude of genres are depicted within a single film.

The implementation of personalized recommendations has the potential to significantly augment user engagement through the customization of content to their particular requirements and inclinations. Enhancing user contentment and personalization while simultaneously augmenting the probability that they will discover content that they appreciate. As a result, users are inclined to exhibit higher frequency of interaction, prolong their stay on the platform, and participate in intended activities like sharing content, remarking, or viewing. For instance, in the case where a video is classified as comedic, personalized suggestions might comprise additional comedic videos that the viewer might find intriguing. By enabling users to discover fresh content that corresponds to their personal interests and preferences, this feature enhances the overall viewing experience. In service administration, content filtering assists in limiting or permitting access to particular categories of videos according to predetermined criteria. By implementing this policy, organizations can effectively regulate employee access to pertinent and suitable materials, thereby promoting efficiency and upholding security protocols. Moreover, content filtration can facilitate targeted advertising by permitting the customization of marketing campaigns to particular video genres, thereby enhancing the efficacy of promotional endeavors. The term "Digital Media Management" encompasses a broad range of activities and processes involved in the organization, storage, retrieval, and manipulation of digital media content. In the context of our article, it serves as the overarching framework within which video genre identification, automatic detection, and edge computing are situated. Moreover, the concept of "Digital Media Management" serves as the foundational framework that underpins video genre identification, automatic detection, and edge computing in our article. By elucidating the connections between these concepts, we aim to highlight the importance of an integrated approach to managing and analyzing digital media content effectively.

Neural networks that integrate spatial and temporal characteristics exhibit optimal performance when representing the sequential sequence of frames in a movie. Prediction [18], data generation [19], and classification [20] are domains in which sophisticated deep learning models have been developed. In order to address this matter, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTMs) are robust neural network architectures that integrate historical and current inputs. Our research on the development of a sophisticated 1D convolutional network for the classification of movie genres is the subject of this article. Our research employed two widely acknowledged recurrent neural network architectures: CNN and LSTM. These models are employed to analyze spatial and sequential data streams in order to extract pertinent characteristics. Additionally, the efficacy of genre recognition using various movie modes is investigated [21–24].

To offer spectators a personalized experience, cloud-based media services [25] and recommendation systems [26, 27] may propose comparable films based on automatically generated movie genres. This can assist users in discovering film genres and techniques that they might not have otherwise encountered. Furthermore, genre-driven recommendations are simpler to implement and analyze compared to more intricate personalized recommendations, thereby optimizing and streamlining the recommendation process.

The development of an LSTM strategy for multimodal video genre identification is a component of our work. The hybrid network incorporates the EfficientNet-B7 architecture as a means to alleviate data imbalances. Subsequently, the proposed model will be assessed in comparison to two sets of benchmarks: the most sophisticated spatio-temporal networks presently in existence and other extensively employed networks. It functions with greater precision and requires fewer computational resources than its predecessors. This article addresses the issue of multi-class video classification through the LMTD-9 dataset analysis. It does so by identifying the most effective video representations for genre identification and classification. This will be achieved by employing a CNN structure with EfficientNet-B7 architecture and LSTM models. Three more elements are incorporated into this study:

- Our proposal entails the utilization of a cutting-edge deep learning framework, namely EfficientNet-B7 and a biLSTM, to effectively identify and categorize genres and video content through the implementation of a CNN.
- We provide a model capable of analyzing video information for many applications in the video industry. We meticulously searched LMTD for all of these videos by querying well-known movie titles.
- We evaluated the CNN-BiLSTM architecture we proposed. Our multiclass video classifier achieved 93.5% validation accuracy. Furthermore, many state-of-the-art machine learning and deep learning structures are assessed and compared to classify genre video material.
- The use of an optimized hybrid convolutional network to identify video genres in service management applications can greatly benefit organizations in several ways. Firstly, it can enhance content categorization, making it easier for service providers to manage and organize videos based on different genres. This can improve user experience and streamline content delivery. Additionally, the network can provide valuable insights into customer preferences and trends, helping businesses tailor their services and offerings accordingly.
- Ultimately, our work can be used on any platform that facilitates video sharing. Additionally, it might prove beneficial for developing browser extensions or plugins that parents can use to ensure their children's internet safety. This is the platform where users receive automated recommendations for movies and videos.

The structure of the remainder of the article is as follows. Section "Related work" provides a comprehensive assessment of the latest and most sophisticated methods for classifying movie genres. Section "Proposed model" provides a detailed explanation of the recommended method. The dataset and experiments may be found in Section "Results". The article concludes with Section "Conclusion", which discusses the further actions to be taken.

Related work

A growing number of scholarly articles have been published on the subject of automated movie genre categorization, which is gaining increasing attention and importance. Using low-level features such as the average duration of shots and color variation, Rasheed et al. [28] suggest identifying genres in video films. Using just one label, a neural network classifier is suggested by [29] for genre categorization. A visual and an aural input are used to achieve this classifier's objective. Huang and Wang used a support vector machine (SVM) classifier to classify both visual and auditory data [30].

Using picture descriptors, Zhou et al. [23] propose extracting high-level visual features. Several image descriptors exist, including Gist, CENTRIST, and w-CENTRIST [31, 32]. It is also possible to predict genres using a K-nearest neighbor classifier. An image description is produced by using a ConvNet [33]. There are several advantages to using a multimodal approach in video classification. Using it, we can capture and analyze the visual, audio, and textual information present in videos. This improves classification accuracy and provides a comprehensive understanding of the content. Furthermore, the model can be made more robust to changes in one modality by integrating multiple modalities. The methods used by Ogawa et al. [34] were a combination of different techniques. In the first place, they employed multimodal learning, which means they processed data from various sources, including visual, audio, and textual. Furthermore, they used a bidirectional Long Short-Term Memory (LSTM) network to capture both forward and backward dependencies in the video data. To label favorite and non-favorite videos, they used a classification approach.

The brain structure of some concepts in video genre classification enables progressive trait integration [18]. As a result, some networks acquire spatial and temporal properties simultaneously. To represent auditory and visual information, Ben et al. [35] used ResNet and SoundNet [36]. It is utilized to better evaluate the temporal dimension of visual input using the LSTM network.

According to Alvarez et al. [37] focusing solely on fundamental movie attributes like shot duration and black

and white usage may improve results. As a result, this method significantly hinders the effectiveness of genre categorization applications, since it only considers movies with a single genre designation.

Yu et al. [38] have developed a bipartite paradigm that emphasizes attention, place, time, and sequence. Utilize a profound Convolutional Neural Network (CNN) to extract the most advanced information from movie frames. In the final stage, a bi-LSTM attention model is used to identify genres.

A probabilistic methodology is presented in [39] that considers the importance of each background scene in each video category. The proposed approach involves two phases: training a support vector machine to categorize scenes and then analyzing the results. A key-frame moment is used as the basis for categorizing videos in part II. In [40], Choros investigates whether it is possible to evaluate shot length. According to Choros, different genres require different durations for shots.

A framework for genre classification of movie trailers using deep networks was presented by Yadav and Vishwakarma [41]. Deep neural networks are used to classify movie trailers into different genre categories based on features extracted from them. Additionally, the paper provides promising results for genre classification accuracy and discusses the advantages and limitations of the methodology. To predict the category of unseen trailers, they trained a deep neural network using a large dataset of labeled movie trailers. In genre classification, the network was trained using a variety of deep network architectures.

In the study [42], convolutional neural networks were used to classify video games into different genres. To predict the genre of new games, a model is trained on labeled game data. It also explores various methods for improving the model's accuracy, such as augmentation of data and fine-tuning of hyperparameters.

Study [43] combines visual, textual, and audio features to categorize movie genres using a multimodal approach. A combination of visual features extracted from movie trailers, textual features extracted from movie descriptions, and auditory features extracted from movie trailer soundtracks is used by the authors. Combining these different modes of information improves genre classification accuracy and reliability.

Using multimodal data, Behrouzi et al. [44] combined audiovisual and textual features to classify movie genres. A recurrent neural network (RNN) was also used to process sequential data, capturing temporal dependencies between frames and sentences. After combining the information from multiple modalities, the RNN is used to classify movies into different genre categories. Multimodal features, such as audio, visual, and textual data, improved movie genre classification accuracy when integrated. Based on the results, the proposed approach accurately classified different genres of movies with an accuracy rate of over 90%.

Since current filmmaking sometimes combines multiple genres inside a single film, it requires multi-label classification. An uneven distribution of samples among different genres can also lead to unfair genre recognition. This study aimed to address these issues by developing a multi-modal classification method for categorizing movie genres. The process begins with the classification of genres solely based on their superficial visual characteristics. Frames are used in the proposed CNN model with BiLSTM structure in order to grasp the conceptual importance of genres. To improve genre recognition, the last stage combines visual and auditory elements. With few data samples, the technique we suggest for classifying movies by genre outperforms the current state-of-the-art algorithms.

Proposed model

We propose an approach to address the issue of disturbing movies in a large movie database with cloud services. Our deep learning architecture has proven effective in resolving challenges related to video classification across a variety of scenarios. As shown in Fig. 1, the recommended model includes three main elements: video preprocessing, deep feature extraction, and video representation and classification.

The dataset movies are initially preprocessed to remove any redundant or superfluous material. Moreover, the frames acquired from each video clip are resized to a consistent size of 224 × 224. In order to obtain characteristics from each video clip, preprocessed frames are input into an ImageNet model known as EfficientNetB7. Thus, to obtain efficient video representations, the extracted features are explored within the BiLSTM architecture prior to being input into the softmax and fully connected layers. In addition, each step is comprehensively described in the subsequent subsections.

In order to identify video genres, a three-part framework architecture is proposed. In the initial step, movies are preprocessed to eliminate unnecessary frames and changed the other frames into photos with 224 × 224 pixels and three-color channels. Following that, we use EfficientNet-B7 to derive feature vectors from the frames. A two-layer BiLSTM stack is used to depict the video after all feature vectors are transformed. In this process, a fully connected layer is used to calculate the likelihood of a video clip belonging to a particular movie genre, followed by an output layer that uses softmax activation.

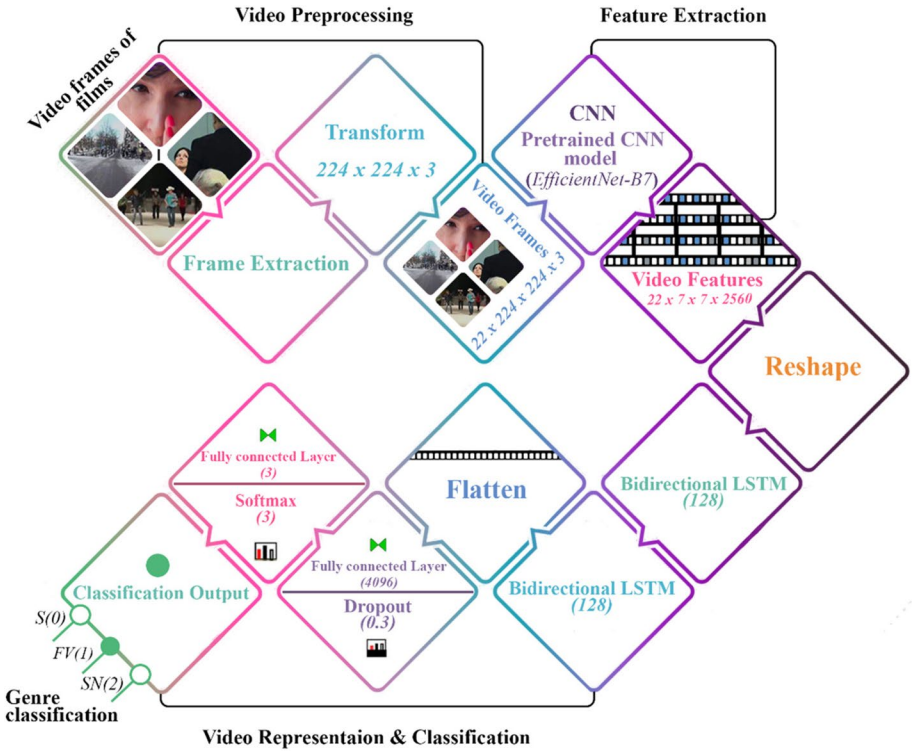

Fig. 1 In this figure, the proposed method for the genre classification of movies is shown.

Preprocessing video frames

We present our overall system architecture diagram, which visually illustrates the model/system layout, thereby facilitating easier comprehension, as depicted in Fig 2. In this figure, the structure illustrates stages such as pre-processing, as well as testing and training the model, all conducted within the cloud platform.

The discrepancy adjacent to the Edge computing platform represents a transition or interface point between the cloud platform and the Edge computing infrastructure. This transition point signifies the handoff of computational tasks or processing between the cloud and Edge environments. It may encompass various aspects such as data transmission, synchronization, or task allocation, depending on the specific requirements of the system architecture.

The collection of necessary information for deep learning techniques for video classification relies on video preprocessing. Therefore, N one-second video clips are used to represent a video (V_i) in this work, abbreviated as $c^i_1, c^i_2, \ldots, c^i_N$. All video clips are manually annotated, but those without complete context or information are rejected. As a result of dividing and labeling the video segments, it was found that each clip contained data from the previous clip. The j^{th} movie frame (c^i_j) is sampled by excluding some initial frames. The last frame of a movie frame with a frame rate below the standard video frame rate will be duplicated. As a general rule, the frames in movie c^i_j are labeled $f^{i,j}_1, f^{i,j}_2, \ldots, f^{i,j}_{22}$, while $f^{i,j}_k$ represents the k^{th} frame. The final step involves resizing the chosen frames from each video clip to ensure they conform to the input dimensions of the pre-trained CNN structure, which is 224×224.

EfficientNet and extraction of deep features

Preprocessed movie clips are used in this part to extract features using cutting-edge deep learning architecture. To extract visual representations from video frames, this study used the pre-trained CNN architecture called EfficientNet rather than constructing a new CNN model from scratch.

The EfficientNet is a CNN structure that uses a scaling technique. Compound coefficients ensure equal scaling of depth, breadth, and resolution of the networks. 1.3 million photos representing 1000 different item classes were used as training data for the model [45]. According to Tan and Le [46], EfficientNet achieved state-of-the-art accuracy on the ImageNet data while having a much smaller size and a speedy inference time than leading CNN strategies. EfficientNet's baseline structure is B0, while its scaling networks are B1 to B7. The computational speed of floating-point operations per second

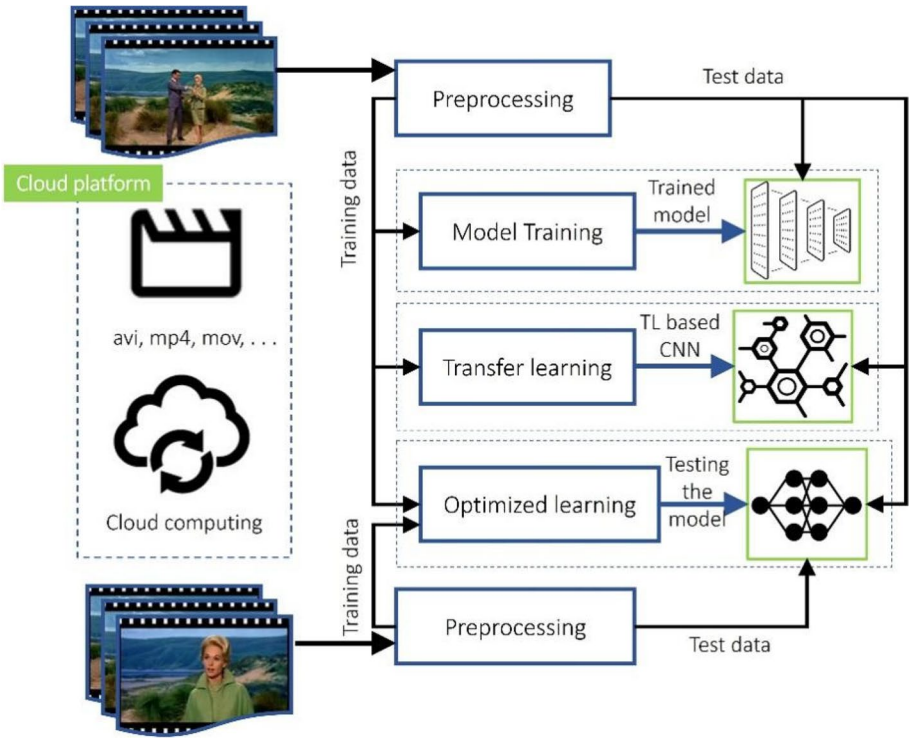

Fig. 2 The general diagram of the system architecture visually illustrates the layout of the model/system, incorporating the implementation of steps within the cloud platform

(FLOPS) is generally sacrificed in favor of enhanced precision in all scaling networks. Additionally, $f^{i,j}{}_1, f^{i,j}{}_2, \ldots, f^{i,j}{}_{22}$ preprocessed extracted frames from each video clip c^i_j were used as input to the EfficientNet-B7. Utilizing the transfer learning technique, the EfficientNet module utilized a stack of 813 layers to extract features. For each input frame, the width, height, and RGB channel were 224×224×3. Moreover, $X^{ij}{}_k$ is produced by EfficientNet-B7 by eliminating the three layers preceding it, including the fully linked layer that generates 1000 ImageNet labels per frame. Using these feature descriptors as input, BiLSTM depicts and categorizes films based on their features.

Representing and classifying

Training a bidirectional LSTM network through supervised learning is the third stage of the pipeline. Using feature descriptors of video clips, this stage allows effective learning of video representations. To the proposed system for genre classification of videos, two interconnected layers are added.

Bidirectional LSTM

Within a standard LSTM cell unit, the cell state c^t functions as an internal memory and governs data flow. By using (1), it becomes evident that the state c^{t-1} is the result of the forget gate $\boldsymbol{f^t}$ acting on the current cell state. In addition, the input gate it acting on a candidate cell state.

$$i^t = sigmoid(W^i_x x^t + W^i_h h^{t-1}) \tag{1}$$

$$\tilde{c}^t = tanh(W^c_x x^t + W^c_h h^{t-1}) \tag{2}$$

$$f^t = sigmoid(W^f_x x^t + W^f_h h^{t-1}) \tag{3}$$

The variables in question are as follows: H represents the total number of hidden nodes, h^{t-1} denotes the first hidden state produced by c^{t-1}, $\{W_x \in \mathbb{R}^{H\times D}, W_h \in \mathbb{R}^{H\times H}\}$ refers to the first set of network parameters, etc. The number of characteristics or measurements recorded at each event is denoted as D, and H represents the quantity of hidden nodes. Approaches to attaining an altered cellular condition include:

$$c^t = i^t \otimes \tilde{c}^t + f^t \otimes c^{t-1} \tag{4}$$

where, $\otimes$ denotes the aggregate of the constituent elements. Ultimately, we construct the concealed state by directing the updated cell state through an output gate layer called $\boldsymbol{o^t}$. Additionally, we employ a sigmoid

function with parameter $\boldsymbol{U}$ to estimate the likelihood of septic shock occurring at time *t*.

$$o^t = sigmoid(W_x^o x^t + W_h^o h^{t-1}) \tag{5}$$

$$h^t = o^t \otimes tanh(c^t) \tag{6}$$

$$p^t = sigmoid(Uh^t) \tag{7}$$

An inherent constraint of LSTM, when applied to video classification, focuses on preceding context. When comprehending a film, it is advantageous to analyze it from two perspectives: the historical context and the potential implications. Therefore, BiLSTM become visible to be a favorable option for movie categorization due to its ability to retain data in both directions (Fig. 3).

BiLSTM consists of two separate hidden layers: the forward hidden layer (h^f_t) and the backward hidden layer (h^b_t). In the forward hidden layer h^f_t, the input vector x_t is sequentially processed in ascending order, starting at 1 and moving up to *T*. In the backward hidden layer h^b_t, the input vector x_t is processed in descending order, starting at *T* and moving down to *T-1*, *T-2*, and so on. Finally, we merge the results of h^f_t and h^b_t to obtain the output y_t.

In order to implement the BiLSTM model, we utilize the following equations:

$$h^f_t = tanh(W^f_{xh} x^t + W^f_{hh} h^f_{t-1} + b^b_h) \tag{8}$$

$$h^b_t = tanh(W^b_{xh} x_t + W^b_{hh} h^f_{t+1} + b^f_h) \tag{9}$$

$$y_t = W^f_{hy} h^f_t + W^b_{hh} h^b_t + b_y \tag{10}$$

By increasing the number of BiLSTM layers excessively, training durations and network complexity increase. Therefore, this study used two layers of bidirectional LSTM to understand visual representations. In our study, the optimized method encapsulates a tailored set of techniques and strategies meticulously designed to elevate the performance and efficiency of the BiLSTM-based video genre identification system. This optimization endeavor encompasses several pivotal steps, foremost among them being the refinement of network architecture through the integration of attention mechanisms and the meticulous fine-tuning of hyperparameters. The overarching objective of this

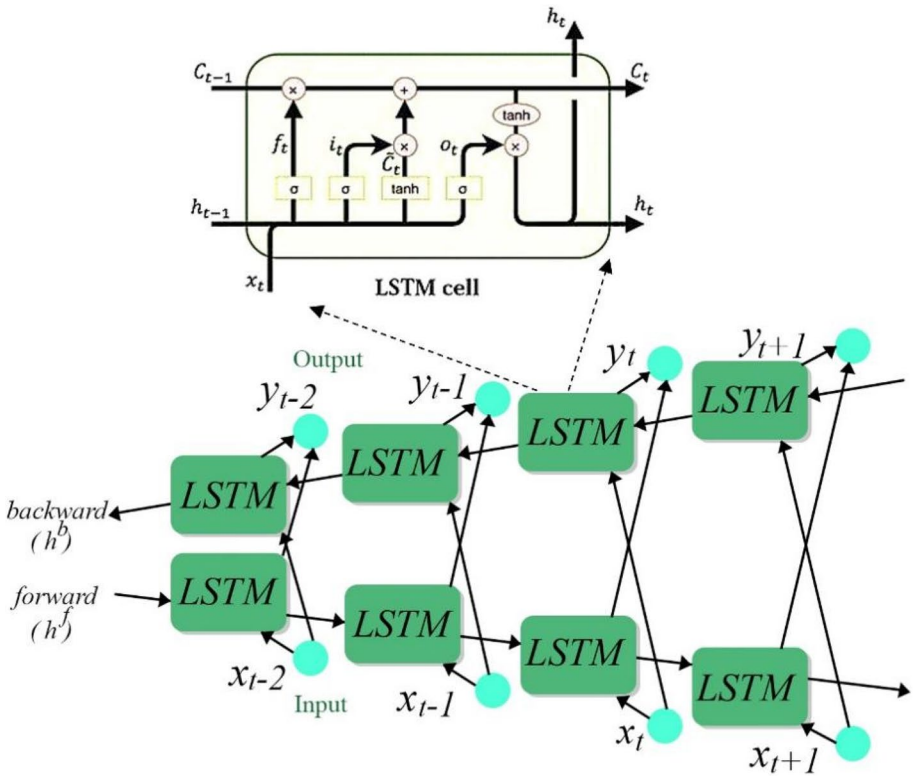

Fig. 3 The BiLSTM architecture is depicted in this image

optimization initiative is to bolster the model's accuracy, resilience, and adaptability across diverse video genres and datasets. By incorporating attention mechanisms and fine-tuning parameters, we endeavor to address multifaceted objectives: mitigating the risk of overfitting, curtailing computational overheads, and optimizing the efficacy of the BiLSTM-based approach for real-world applications. This holistic optimization process is envisioned to imbue the BiLSTM model with the requisite robustness and versatility to navigate the intricate landscape of video genre identification tasks effectively.

Attention mechanism in BiLSTM

An attention system in a neural network model finds the optimal instances to examine data, such as segments of films, by giving more priority to feature vectors that contain more valuable information. Figure 4 depicts the integration of an attention mechanism into the BiLSTM architecture.

Considering the final hidden state of the i^{th} bidirectional LSTM as "h_{it}", which is calculated as:

$$h_{it} = [h_t^f, h_t^b] \tag{11}$$

The below equations are employed to ascertain the attention mechanism:

$$e_{it} = tanh(b_a + W_a h_{it}) \tag{12}$$

moreover,

$$a_{it} = exp(e_{it}) \times \left(\sum_{j=1}^{T} exp(e_{jt}) \right)^{-1} \tag{13}$$

$$v_t = \sum_{t=1}^{T} a_{it}.h_{it} \tag{14}$$

Equation 13 determines the attention weight a_{it} assigned to the i^{th} BiLSTM output vector at time t by the attention mechanism.

The attention layer weight and bias are denoted as W_a and b_a, respectively. Notably, the output of the attention layers yields an attention vector v_t. This is obtained by summing the i^{th} BiLSTM output vector at time t with the attention weight a_{it}.

Softmax

In the output layer of a deep learning structure, an activation function (AF) called a softmax classifier is utilized. This function makes use of the softmax algorithm. To classify films into various genres, the suggested method utilized a softmax AF in the last fully connected layer to specify the relative probabilities of numerous output parts. The softmax AF (σ) is used to compute it.

$$\sigma(z_i) = exp(z_i) \times \left(\sum_{c=0}^{N-1} exp(e_c) \right)^{-1} \tag{15}$$

Final structure

The structure for multiclass movie categorization is shown in Fig. 1. The input comprises a series of 22

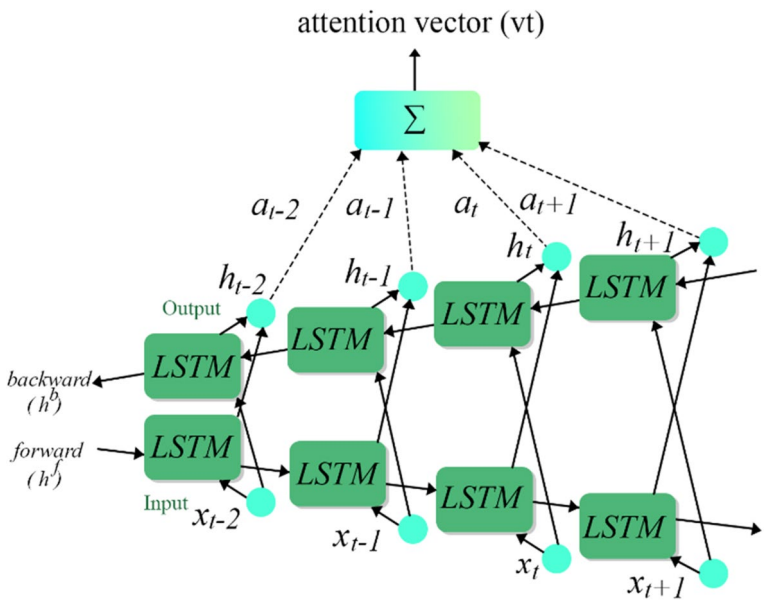

Fig. 4 The structure of the model including the attention mechanism and BiLSTM may be observed here

frames, each with a consistent resolution of 224×224×3 pixels. In order to get a feature vector of dimensions 22×7×7×2560, the frames are subjected to feature extraction using EfficientNet-B7. The modified high-level characteristics are fed into the BiLSTM network's two-layer stack. The flattening part (layer) is employed to convert the feature representations into a 1D vector. Following that, we will incorporate a densely linked layer consisting of 4096 neurons, utilizing the rectified linear unit (ReLU) AF. To address overfitting, a dropout rate of 0.3 is employed. This is done because a completely linked layer has the ability to produce a diverse set of probabilities. This is done by establishing connections between all inputs of one layer and all activation units of the succeeding layer. The final classification scores are generated by the softmax output layer consisting of three neurons.

Results

In this section, we provide an overview of the dataset utilized, followed by the adjustments made to the hyperparameters and decision bounds in order to get optimal outcomes. The equations employed to assess our system are further elaborated upon thereafter. Ultimately, the recommended method's effectiveness is assessed by comparing it to cutting-edge alternatives using the AUC score. The tables highlight the best performers in each metric by displaying their results in bold.

Dataset and setting

LMTD-9 multi-label trailer database [47] with 4021 movies was used for this purpose. LMTD-9 has a greater variety of trailer types than any other dataset currently available. A total of 603 sets are consisted in the database, containing 2815 sets for testing and 603 sets for validation. Trailers can be categorized into nine different categories. There are 693 thrillers, 313 science fiction, 651 romances, 436 horrors, 2032 dramas, 659 crime films, 1562 comedies, 593 adventure films, and 856 action films on this list.

We sample the video clips at a rate of 22 frames per second by excluding a portion of the first frame to avoid frame overlap. To compensate for video clips with frame rates below the standard range of 23-24 frames per second, the last frame is duplicated and added as padding. A frame sampling rate of 22 frames per second is used for all experiments conducted to train, verify, and test neural network models.

When using a pre-trained CNN model from ImageNet for feature extraction from video frames, it is necessary to adjust the hyperparameters of the BiLSTM and fully connected (dense) layer to optimize performance. Two concurrent layers of bidirectional LSTMs can significantly improve the performance of video classification, surpassing the effectiveness of single or multiple layers of bidirectional LSTMs. A thorough evaluation of several design choices for bidirectional LSTM layers led to this conclusion. In addition, each bidirectional LSTM layer can have 64, 128, 256, or 512 embedded hidden units. Consistency is ensured by keeping the number of hidden unit's constant in both bidirectional layers. It has been demonstrated that a BiLSTM network with two layers and 128 hidden units has the highest validation accuracy. In identifying the most appropriate and relevant labels for the categorization layer, a fully connected (FC) layer with 4096 units and a ReLU AF outperforms a FC layer with 2048 units and the same AF. A dropout layer with a dropout rate of 0.3 is added before the final dense layer, which serves as a fully connected output layer, in order to mitigate overfitting. In a three-unit output layer, the softmax classifier is used to obtain the ultimate probability scores.

There is a distinct mix of output size and learnable parameters in each of the nine layers of the model. The recommended model's 152 million parameters, which represent the number of neurons, are adjusted to optimize their performance during backpropagation. Due to its non-trainable parameters, the pre-trained EfficientNet-B7 model does not have any parameters that can be optimized. The cost function can be used to quantify a mismatch between actual predicted values. Moreover, the optimizer function aims to minimize the total loss or errors in the neural network model, thus enhancing its accuracy. The classification cross-entropy loss function is employed to classify multiclass videos.

$$L_{CrossEntropy}(\widehat{y}, y) = -\sum_{c=0}^{N-1} y_{i,c} * log(\widehat{y}_{i,c}) \tag{16}$$

This calculation is performed as follows: For each instance i, $y_{i,c}$ is the forecasted probability of class c for instance i, where c denotes one of N classes. In order to increase the precision of the cost function, we employ the Adam optimizer, an enhanced variant of stochastic gradient descent (SGD) that operates at a learning rate of 1e-5. As a result of memory and processing constraints, small subsets of the training dataset can be used to train models. During each repetition of one epoch, a subset of 1000 instances from the training datasets is analyzed concurrently. The EfficientNet-BiLSTM structure is trained using a mini-batch gradient descent optimization approach, which divides the training dataset into n subsets. Trial results using different mini-batch sizes (8, 16, and 32) showed that a batch size of 16 achieves the fastest convergence and generates the most accurate models. During a 90-iteration period, the proposed EfficientNet-BiLSTM model modifies its weights once every epoch. As

part of each iteration, the training dataset chunk was processed in 63 batches with a mini-batch size of 16.

Evaluation

Table 1 demonstrates a comparison of the test outcomes obtained through our approaches. Our suggested movie genre categorization model offers an alternative to basic LSTM, LSTM+EfficientNet, BiLSTM, and other hybrid deep learning approaches

The study's dataset contains preference samples of several video genres, ranging in magnitude from 1 to 9. Based on our research, this strategy has proven to be far more effective and applicable in a wider range of situations than previously believed. Our approach achieved a remarkable accuracy rate of 93.5% when tested on 9 distinct genres. To accommodate diverse individual preferences, this proposal challenge utilized

Table 1 Our proposed approach is shown in the table below, along with experimental results from three similar 5-fold CV experiments that used other similar models. Different videos were collected from a variety of films for the analysis

Genre	Method	5-fold (1)				5-fold (2)				5-fold (3)			
		Acc	Sen	Spec	F1	Acc	Sen	Spec	F1	Acc	Sen	Spec	F1
Thrillers	LSTM	88.7	85.6	96.2	91.6	88.6	86.9	97.2	90.1	90.7	89.6	96.3	89.4
	biLSTM	90.3	90.1	97.6	92.3	90.0	88.7	98.6	92.3	92.6	90.3	96.8	91.1
	LSTM+EfficientNet	93.2	92.1	1.00	94.5	91.8	90.2	99.6	96.0	94.2	94.0	98.5	93.2
	Proposed	**93.6**	**1.00**	**1.00**	**1.00**	**92.5**	**91.9**	**1.00**	**97.9**	**95.4**	**95.4**	**1.00**	**94.7**
Science fiction	LSTM	88.7	70.3	96.3	82.5	88.6	78.6	95.3	85.6	87.5	65.8	97.8	80.3
	biLSTM	90.2	75.6	98.5	85.6	90.2	**82.7**	97.9	**89.9**	88.6	67.9	98.1	81.6
	LSTM+EfficientNet	91.6	68.2	99.5	86.6	90.0	80.3	97.2	89.6	90.6	**70.2**	98.6	**85.9**
	Proposed	**93.2**	**83.0**	**99.8**	**89.9**	**92.5**	79.3	**1.00**	88.5	**93.1**	68.2	**1.00**	81.1
Romances	LSTM	88.1	84.3	95.6	80.3	87.3	70.1	95.3	80.6	89.3	75.3	94.6	80.1
	biLSTM	89.7	82.3	96.5	82.9	89.8	74.3	99.1	82.5	91.0	78.4	96.3	82.3
	LSTM+EfficientNet	91.6	**84.5**	97.3	84.2	91.3	76.2	98.6	84.0	92.1	**80.3**	97.5	**85.3**
	Proposed	**93.3**	82.7	**99.5**	**86.6**	**92.6**	**77.8**	**99.7**	**86.1**	**93.0**	79.2	**99.3**	84.7
Horrors	LSTM	87.4	80.2	96.5	84.1	87.4	68.6	95.9	81.2	86.5	79.6	96.0	83.7
	biLSTM	89.3	82.0	97.4	85.3	88.6	70.3	96.5	82.2	88.6	80.3	96.5	84.6
	LSTM+EfficientNet	91.7	**84.0**	99.1	88.0	90.3	72.3	97.2	84.7	90.8	**82.6**	97.4	85.3
	Proposed	**93.4**	82.7	**99.7**	**88.3**	**92.1**	**75.8**	**99.8**	**85.1**	**93.0**	80.6	**99.5**	**86.0**
Dramas	LSTM	90.6	94.8	90.2	92.4	91.6	96.0	88.6	91.8	86.3	85.1	95.4	91.4
	biLSTM	92.6	95.7	92.3	92.3	93.1	97.6	89.6	92.3	88.9	88.4	96.5	92.4
	LSTM+EfficientNet	95.5	96.5	95.1	94.9	94.3	98.6	92.3	93.8	90.6	89.3	97.3	94.7
	Proposed	**96.9**	**99.7**	**96.8**	**95.6**	**96.8**	**99.7**	**96.7**	**95.4**	**93.4**	**91.2**	**98.0**	**96.9**
Crime	LSTM	88.9	78.2	92.7	88.1	88.9	78.5	97.5	82.9	88.6	86.9	98.0	87.4
	biLSTM	90.1	80.2	94.6	89.2	88.9	78.4	97.4	82.6	90.3	88.3	98.5	89.6
	LSTM+EfficientNet	91.3	82.0	95.3	90.1	90.7	80.6	98.7	86.3	92.3	89.3	99.0	91.0
	Proposed	**93.2**	**85.6**	**96.2**	**93.8**	**92.5**	**82.5**	**99.4**	**87.5**	**93.0**	**90.9**	**99.2**	**91.2**
Comedies	LSTM	92.5	96.8	93.1	91.5	90.0	96.5	89.9	88.2	90.2	89.2	92.5	86.1
	biLSTM	92.3	96.5	92.3	91.0	89.9	96.5	89.6	88.1	93.2	94.6	94.1	86.5
	LSTM+EfficientNet	94.1	97.0	93.8	92.2	91.6	97.2	90.9	88.6	94.6	95.6	95.0	88.7
	Proposed	**94.3**	**97.7**	**94.2**	**93.8**	**92.5**	**98.7**	**91.0**	**90.3**	**96.2**	**97.1**	**95.5**	**90.4**
Adventure	LSTM	88.3	87.4	93.6	90.2	88.2	84.1	98.6	88.6	89.3	79.6	98.4	85.4
	biLSTM	89.2	89.1	94.7	91.0	88.6	84.2	99.1	88.6	90.6	81.1	98.5	86.3
	LSTM+EfficientNet	91.3	**90.2**	96.5	92.7	90.3	85.3	99.0	90.0	92.1	82.6	99.1	88.8
	Proposed	**93.5**	89.1	**99.9**	**93.8**	**92.5**	**86.5**	**99.8**	**91.9**	**93.0**	**83.1**	**99.7**	**89.5**
Action	LSTM	91.5	90.6	99.2	94.6	91.9	91.5	96.8	90.3	90.2	89.8	95.6	89.8
	biLSTM	92.3	91.0	99.1	95.4	92.3	93.6	97.1	92.0	92.6	92.3	96.3	92.4
	LSTM+EfficientNet	94.3	92.3	1.00	97.4	94.3	94.2	98.5	93.1	94.6	93.4	97.1	94.6
	Proposed	**97.4**	**97.1**	**1.00**	**98.5**	**96.6**	**95.5**	**1.00**	**95.2**	**97.9**	**97.6**	**1.00**	**98.2**

video frames. The previously listed regions all meet the desired accuracy level.

The results demonstrate that the 5-fold cross validation (CV) technique generated reliable and accurate results. The technique not only enhanced the experiment's precision but also substantially mitigated bias potential. Some items in the video frame collection may exhibit low classification accuracy.

In order to assess the distribution of these elements using video frames, three enhanced models were employed, as indicated in Table 1: LSTM, BiLSTM, LSTM+EfficientNet, and the recommended model. Our research indicates that these models exhibit high computational accuracy, sensitivity, and specificity.

Therefore, our architecture outperforms the most advanced methods in terms of accuracy and computational cost. Based on experimental results, EfficientNet-BiLSTM outperforms other frameworks with an accuracy of 93.52%. Moreover, EfficientNet with BiLSTM delivers performance with an F1 score of 0.9012.

The categorizing method yielded productive results, with 93% accuracy. The recommended approach demonstrated 93.52% accuracy when compared to similar models in a nine-class classification task, utilizing all three 5-fold cross-validation iterations. In addition, we performed multi-class categorization for each model scenario. Figure 5 displays confusion matrixes, which represent classification outcomes and allow model comparison. The results demonstrate a classification accuracy percentage of 93.52% for various movie genres. The data exhibit no statistically significant variation or deviation. The film is classified as a single genre because classification is subjective. Moreover, variance estimation is employed to assess the classification output and determine the method's efficacy in different movie genre classification scenarios.

Uncertainty becomes more noticeable when there is a significant disparity in outcomes. Figure 6 shows classification results obtained from several tests, calculated based on the loss function. They are achieved through convergence. Upon examining the proposed model's architecture in comparison to LSTM, BILSTM, and LSTM+EfficientNet for genre classification, it is evident that BiLSTM+EfficientNet outperforms the other structures consistently throughout the classification process. BiLSTM+EfficientNet demonstrates higher classification accuracy and robustness in uncertainty.

Discussion

Table 2 demonstrates that the inclusion of the multimodal component in the network resulted in a significant increase in the scores of most genres. This demonstrates that movie soundtracks may significantly improve genre prediction accuracy.

The decrease in the science fiction score with audio inclusion can be attributed to the symmetry between action and science fiction movie scores. The confusion matrixes and loss convergence reveal a significant disparity in the number of action movie trailers compared to sci-fi trailers. This discrepancy has the potential to cause confusion within the network, resulting in misclassification of these two genres. While the area under the curve (AUC) scores of the Gated recurrent unit (GRU)+SVM [44] and 1D-Conv+SVM [44] models are quite close, the GRUs+SVM model performs better than the 1D-Conv+SVM model for five out of nine genres. Furthermore, the GRU+SVM approach demonstrates superior performance compared to the most advanced models in all categories, with the exception of drama, as measured by the AUC metric. Particularly, GRU+SVM has significantly improved performance in genres with relatively few data, such as thriller and horror videos.

The findings from previous sections indicate that the EfficientNet-B7 model, pre-trained on ImageNet, outperforms other feature extractors. Furthermore, when this architecture is paired with any deep learning method, it demonstrates a high level of effectiveness in accurately identifying and categorizing hazardous films. Due to the suitability of BiLSTM, a deep learning strategy, we have selected it to build a deep convolutional learning-based method for our movie classification challenge. Our studies involve a BiLSTM structure, followed by FC layers (with 4096 units and ReLU activation), drop-out (with a value of 0.3), and softmax (with three output parts) layers.

Additionally, we examine an attention mechanism-based BiLSTM model that integrates an attention unit after each bidirectional layer. This model also includes layers for fully linked operations (with 4096 units and ReLU activation), dropout (with a value of 0.3), and softmax (with 3 output units). The LMTD dataset is divided into 5 parts for training and assessment in all investigations, using a 5-fold cross-validation split. The models are trained and tested for 100 Epochs. Ultimately, video classification scores are acquired by evaluating the trained model at the latest epoch (epoch = 100). The findings demonstrate that including attention mechanisms into EfficientNet-BiLSTM models leads to superior performance than models without attention mechanisms. This was shown by experimenting with different numbers of hidden units in each BiLSTM layer of the suggested structure.

The imbalance in the number of action and sci-fi trailers can cause confusion within the network, resulting in misclassification. Additionally, the GRU+SVM

Fig. 5 This output presents confusion matrices, which depict the classification results and enable comparison of models

model outperforms the 1D-Conv+SVM model for five out of nine genres, demonstrating superior performance across most categories except for drama. Furthermore, findings suggest that the EfficientNet-B7 model pre-trained on ImageNet excels as a feature extractor, particularly when paired with deep learning methods, showcasing high effectiveness in accurately categorizing films. The chosen approach involves a BiLSTM structure with FC layers, dropout, and softmax layers. Additionally, an attention mechanism-based BiLSTM model is examined, demonstrating enhanced performance compared to models without attention mechanisms. The inclusion of attention mechanisms in EfficientNet-BiLSTM models yields superior performance, as evidenced by experiments with different numbers of hidden units in each BiLSTM layer. Overall, the findings highlight the effectiveness of the proposed hybrid deep learning architecture in cloud-based video genre recognition, especially when incorporating attention mechanisms into EfficientNet-BiLSTM models,

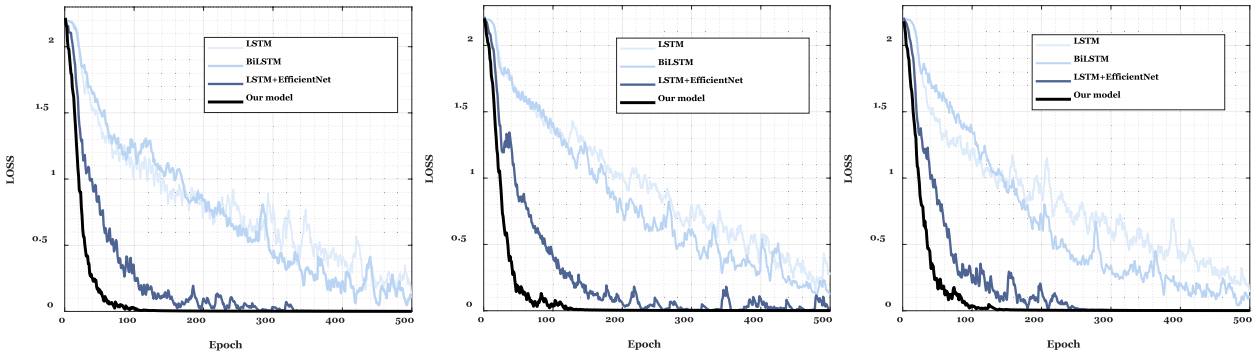

Fig. 6 A comparison of the proposed model with other comparable approaches has been conducted to demonstrate the convergence of the Loss criteria in various schemes. To determine the movie genre, the model was tested three times in completely random situations

Table 2 This table lists present technological advancements as well as the specific domains in which each film proposes to apply its methods

Model	Thriller	Sci-Fi	Romance	Horror	Drama	Crime	Comedy	Adventure	Action
CTT-MMC-TN [18]	0.522	0.401	0.456	0.667	0.841	0.547	0.870	0.672	0.835
LSTM [18]	0.437	0.237	0.313	0.478	0.740	0.421	0.792	0.573	0.687
LLFM [37]	0.520	0.192	0.468	0.424	0.641	0.628	0.871	0.752	0.852
1D-Conv-V [44]	0.763	0.787	0.655	0.810	0.733	0.659	0.803	0.807	0.777
GRU+SVM [44]	0.825	0.628	0.725	0.865	0.829	0.846	0.910	0.929	0.909
Our model	**0.856**	**0.893**	**0.896**	**0.929**	**0.901**	**0.853**	**0.923**	**0.924**	**0.912**

leading to improved classification accuracy across various genres.

Our experimental results indicate that while noise can potentially affect the performance of the model to some extent, our architecture demonstrates a degree of robustness and adaptability in handling noisy input data. By leveraging advanced techniques in deep learning and data preprocessing, including denoising algorithms and feature extraction methods, we have taken proactive measures to mitigate the adverse effects of noise on the model's performance. Furthermore, we have incorporated strategies such as data augmentation and regularization to enhance the model's generalization capabilities and improve its robustness to noisy input conditions.

An extensive evaluation of EfficientNet's performance with and without the attention-based mechanism. These are all the courses belonging to each category. Confusion matrices are utilized to visualize BiLSTM models. The diagonal numbers in each class represent the count of correctly identified occurrences, whereas cases that are not on the diagonal indicate incorrect classification.

By offering personalized movie recommendations [48–50] based on the user's preferences and viewing history, users can discover upcoming films that align with their interests. This not only adds to the enjoyment of the service but also enriches the user's content consumption experience. This leads to increased customer satisfaction and loyalty. Furthermore, integrating recommendation systems for film genres can help with discoverability and content filtering. This makes it easier for users to find the specific types of movies they want.

Using a hybrid convolutional network in service management allows for more accurate and efficient categorization of videos based on their content. This, in turn, enables targeted advertising, content filtering, and personalized recommendations, leading to improved user experience and increased customer satisfaction. Additionally, the hybrid convolutional network's ability to recognize different video genres can help in detecting inappropriate or copyrighted content, enabling better content moderation and compliance with regulations. Targeted advertising in service management allows for personalized and relevant advertisements to be presented to the viewers. This means that businesses can reach their specific target audience with higher efficiency and accuracy, increasing the likelihood of conversion and revenue generation. Additionally, targeted advertising helps to minimize wasted resources and ensures that advertisements are aligned with the preferences of the viewers, leading to a better user experience.

The proposed model harnesses the prowess of deep CNNs, heralding a new era in the realm of visual pattern recognition. Beyond merely discerning video genres, this model delves deeper into understanding the context within videos, thereby elevating its predictive accuracy. Notably, its cost-effectiveness renders it conducive for integration into cloud-based services.

Edge computing involves performing lightweight computations and inference tasks on edge devices, such as smartphones, edge servers, or IoT devices. These computations are typically less resource-intensive compared to centralized cloud servers, as they involve processing data locally at the network edge. Therefore, the computational complexity of edge computing in our method is relatively low, making it suitable for real-time video analysis and classification on resource-constrained edge devices.

Secondly, regarding the EfficientNet-BiLSTM model, it is essential to consider the computational demands associated with both components:

1. EfficientNet: EfficientNet is a state-of-the-art CNN architecture that achieves high accuracy with significantly fewer parameters compared to traditional CNNs. While EfficientNet is computationally efficient compared to larger CNN architectures, it still requires substantial computational resources during training and inference, particularly for large-scale datasets and high-resolution video inputs.
2. BiLSTM: Our optimized network captures temporal dependencies in sequential data, such as video frames. The computational complexity of BiLSTM primarily depends on the sequence length and the number of hidden units in the network. While BiLSTM is more computationally intensive compared to feedforward neural networks, it offers superior performance in capturing long-range dependencies and contextual information in video data.

Envisioned as a catalyst for revolutionizing video content analysis, the model exhibits unparalleled precision compared to prevailing methodologies. Its potential to enhance streaming services through refined content recommendations holds promise for enriching user experiences. Moreover, its capacity to precisely classify videos empowers streaming platforms to tailor content more effectively to their audience's preferences. Indeed, the proposed model marks a milestone in multimedia management, particularly in the domain of video genre recognition. By leveraging deep convolutional learning, it outshines conventional approaches, offering a robust framework for genre classification in cloud services.

At the heart of the model's efficacy lies its ability to automatically extract intricate features from video data using deep convolutional layers. Unlike traditional techniques reliant on manual feature engineering, deep learning models decipher complex patterns directly from raw input, capturing the subtleties inherent in multimedia content. Furthermore, the model showcases remarkable adaptability and generalizability across diverse datasets and video genres. Leveraging deep learning architectures enables it to glean insights from vast datasets, thereby enhancing performance and scalability in real-world scenarios—a critical trait in the ever-evolving landscape of multimedia content. Despite its strides, avenues for further refinement persist. Enhancing the model's robustness to variations in video content, such as scene changes or lighting conditions, represents a paramount challenge. Additionally, addressing scalability concerns with larger datasets and optimizing computational efficiency remain imperative areas for future exploration. The proposed model not only heralds a paradigm shift in video genre recognition but also serves as a springboard for future advancements, poised to tackle the evolving challenges of multimedia content analysis with aplomb.

By leveraging deep convolutional learning techniques, our approach offers novel insights into enhancing multimedia management within cloud services. The utilization of cloud infrastructure provides scalability and flexibility, enabling efficient processing and analysis of multimedia content, which is particularly crucial in today's data-intensive environments. Therefore, our work not only addresses the challenges of multimedia management but also underscores the pivotal role of cloud computing in facilitating advanced computational tasks. Our research explores the practical implications of deploying deep learning models for video genre recognition within cloud environments. By harnessing the computational power of the cloud, we demonstrate the feasibility of real-time genre classification, thereby enhancing the user experience and accessibility of multimedia content. This aspect underscores the direct relevance of our work to cloud computing, as it showcases the potential of cloud-based solutions in addressing multimedia-related challenges.

Additionally, we plan to incorporate deep transformer-based models with attention mechanism [51, 52] in our future endeavors. These models play a crucial role in video genre classification, boasting advanced architectures adept at extracting intricate features from video data. Given the diverse temporal and spatial characteristics present in videos, transformer models excel at recognizing complex patterns, thereby enhancing the accuracy of genre classification. Moreover, their capability to handle long-form data commonly found in videos promises a significant performance boost. The proposed deep convolutional learning approach to video genre recognition in cloud services represents a notable advancement

in multimedia management. Its ability to automatically learn discriminative features from raw video data, coupled with its adaptability and superior performance metrics, distinguishes it from traditional methods. While there is room for improvement, particularly in enhancing robustness and scalability, the proposed model serves as a promising foundation for future research and innovation in the field.

Conclusion

In this article, an effective deep learning-based architecture for genre categorization in movies is proposed. The EfficientNet-B7 structure is used to extract video features through transfer learning. With the extracted movie features, the BiLSTM structure learns effectual movie representations and conduct multi-class film classification. Experimental evaluations are conducted with a large dataset that includes videos from multiple genres. EfficientNet-BiLSTM with hidden units 128 exhibits higher accuracy (93.5%) than similar structures including LSTM, BiLSTM, and LSTM+EfficientNet, according to the evaluation results. Furthermore, our BiLSTM-driven framework outperformed existing state-of-the-art models with the highest recall score of 93.5% when compared to existing state-of-the-art methods. The suggested deep learning-inspired genre classification for videos has the advantage that it considers real-time conditions using deep learning frameworks such as EfficientNet-B7 and BiLSTM, which determine movie genres based on real-time conditions. Hybrid convolutional networks can also identify different video genres in service management. By categorizing videos based on their content, targeted advertising, content filtering, and personalized recommendations can be generated. The LMTD dataset has also been used to develop a methodology for classifying genres. This commendation underscores the transformative impact of employing hybrid convolutional networks within the realm of multimedia management, particularly in the nuanced domain of video genre recognition in cloud services. By leveraging the intricate capabilities of deep learning, our proposed model not only enhances the accuracy and efficiency of content classification but also opens up new avenues for refining content management systems across various industries. This achievement reflects not only the culmination of rigorous research and development but also highlights the potential of cutting-edge technologies to revolutionize how we interact with multimedia content. Moreover, the seamless integration of cloud services into our model further amplifies its practical utility, enabling scalable and efficient deployment in real-world scenarios. Beyond the realm of academia, the implications of this advancement extend to commercial applications such as targeted advertising and personalized service recommendations, where precise understanding and categorization of multimedia content are paramount. In essence, our work signifies a significant step forward in the ongoing quest to harness the power of deep convolutional learning for transformative advancements in multimedia management and beyond. In the future, we aim to enhance the model's performance by gaining a deeper understanding of video global representations. Additionally, we will focus on improving the categorization labels to better target film genres.

Authors' contributions

F. L. wrote and revised the main manuscript; J. Y. prepared Tables 1 and 2 & Figs. 1, 2, 3, 4 and 5; M. A. guided and reviewed the manuscript;

Funding

No funding was received.

Declarations

Availability of data and materials

All the data and codes are available through the corresponding author.

Competing interests

The authors declare no competing interests.

Received: 19 March 2024 Accepted: 7 May 2024
Published online: 17 May 2024

References

1. Chen Z, Ye S, Chu X, Xia H, Zhang H, Qu H, Wu Y (2021) Augmenting sports videos with viscommentator. IEEE Trans Visual Comput Graph 28(1):824–34
2. Ma J, Jiang X, Fan A, Jiang J, Yan J (2021) Image matching from handcrafted to deep features: a survey. Int J Comput Vision 129:23–79
3. Wang W, Yang Y, Wang X, Wang W, Li J (2019) Development of convolutional neural network and its application in image classification: a survey. Opt Eng 58(4):040901
4. Saini P, Kumar K, Kashid S, Saini A, Negi A (2023) Video summarization using deep learning techniques: a detailed analysis and investigation. Artif Intell Rev 56(11):12347–12385
5. Singh AS, Bevilacqua A, Nguyen TL, Hu F, McGuinness K, O'Reilly M, Ifrim G (2023) Fast and robust video-based exercise classification via body pose tracking and scalable multivariate time series classifiers. Data Min Knowl Discov 37(2):873–912
6. Yang Y, Qi Y, Qi S (2024) Relation-consistency graph convolutional network for image super-resolution. Vis Comput 40(2):619–635
7. Kumar S, Kumar N, Dev A, Naorem S (2023) Movie genre classification using binary relevance, label powerset, and machine learning classifiers. Multimed Tools Appl 82(1):945–968
8. Dastbaravardeh, E., et al., (2024). Channel Attention-Based Approach with Autoencoder Network for Human Action Recognition in Low-Resolution Frames. Int J Intell Syst. 2024
9. Motamedi E, Kholgh DK, Saghari S, Elahi M, Barile F, Tkalcic M (2024) Predicting movies' eudaimonic and hedonic scores: a machine learning approach using metadata, audio and visual features. Inf Process Manag 61(2):103610
10. Yousaf K, Nawaz T (2022) A deep learning-based approach for inappropriate content detection and classification of youtube videos. IEEE Access 28(10):16283–98
11. Yi Y, Li A, Zhou X (2020) Human action recognition based on action relevance weighted encoding. Signal Process 1(80):115640

12. Almeida A, de Villiers JP, De Freitas A, Velayudan M (2022) The complementarity of a diverse range of deep learning features extracted from video content for video recommendation. Expert Syst Appl 15(192):116335
13. Mahadevkar SV, Khemani B, Patil S, Kotecha K, Vora DR, Abraham A, Gabralla LA (2022) A review on machine learning styles in computer vision—Techniques and future directions. IEEE Access 26(10):107293–329
14. Tulbure AA, Tulbure AA, Dulf EH (2022) A review on modern defect detection models using DCNNs–Deep convolutional neural networks. J Adv Res 1(35):33–48
15. Montalvo-Lezama R, Montalvo-Lezama B, Fuentes-Pineda G (2023) Improving transfer learning for movie trailer genre classification using a dual image and video transformer. Inf Process Manag 60(3):103343
16. Bi T, Jarnikov D, Lukkien J. (2022) Shot-Based Hybrid Fusion for Movie Genre Classification. InInternational Conference on Image Analysis and Processing. pp. 257-269. Cham: Springer International Publishing
17. Pant P, Sai Sabitha A, Choudhury T, Dhingra P (2018) Multi-label classification trending challenges and approaches. Emerg Trends Expert Appl Secur 2019:433–44
18. Wehrmann J, Barros RC (2017) Movie genre classification: a multi-label approach based on convolutions through time. Appl Soft Comput 1(61):973–82
19. Zhang X, Yang Q (2019) Transfer hierarchical attention network for generative dialog system. Int J Autom Comput 16:720–36
20. Rezaee K et al (2024) A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance. Personal and Ubiquitous Computing 28(1):135–151
21. Badamdorj T, Rochan M, Wang Y, Cheng L. (2021) Joint visual and audio learning for video highlight detection. InProceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8127-8137
22. Tian Y, Xu C. (2021) Can audio-visual integration strengthen robustness under multimodal attacks?. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5601-5611
23. Zhou H, Hermans T, Karandikar AV, Rehg JM. (2010) Movie genre classification via scene categorization. InProceedings of the 18th ACM international conference on Multimedia. pp. 747-750
24. Cai Z, Ding H, Wu J, Xi Y, Wu X, Cui X (2023) Multi-label movie genre classification based on multimodal fusion. Multimed Tools Appl 15:1–8
25. Yang X, Esquivel JA. (2023) LSTM network-based Adaptation Approach for Dynamic Integration in Intelligent End-edge-cloud Systems. Tsinghua Sci Technol
26. Li D, Esquivel JA (2024) Accuracy-enhanced E-commerce recommendation based on deep learning and locality-sensitive hashing. Wireless Networks 2:1–6
27. Li D, Esquivel JA. Trust-aware Hybrid Collaborative Recommendation with Locality-Sensitive Hashing. Tsinghua Science and Technology. 2023.
28. Rasheed Z, Sheikh Y, Shah M (2005) On the use of computable features for film classification. IEEE Trans Circuits Syst Video Technol 15(1):52–64
29. Jain SK, Jadon RS. (2009) Movies genres classifier using neural network. In2009 24th International Symposium on Computer and Information Sciences. pp. 575-580.
30. Huang YF, Wang SH. (2012) Movie genre classification using svm with audio and video features. InActive Media Technology: 8th International Conference, AMT 2012, Macau, China, December 4-7, 2012. Proceedings 8 pp. 1-10. Springer Berlin Heidelberg
31. Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vision 42:145–75
32. Wu J, Rehg JM. (2008) Where am I: Place instance and category recognition using spatial PACT. In2008 IEEE Conference on Computer Vision and Pattern Recognition pp. 1-8
33. Simoes GS, Wehrmann J, Barros RC, Ruiz DD. (2016) Movie genre classification with convolutional neural networks. In2016 International Joint Conference on Neural Networks (IJCNN) pp. 259-266
34. Ogawa T, Sasaka Y, Maeda K, Haseyama M (2018) Favorite video classification based on multimodal bidirectional LSTM. IEEE Access 18(6):61401–9
35. Ben-Ahmed O, Huet B. (2018) Deep multimodal features for movie genre and interestingness prediction. In2018 international conference on content-based multimedia indexing (CBMI) pp. 1-6. IEEE
36. Aytar Y, Vondrick C, Torralba A. (2016) Soundnet: Learning sound representations from unlabeled video. Adv Neural Inf Process Syst ;29
37. Álvarez F, Sánchez F, Hernández-Peñaloza G, Jiménez D, Menéndez JM, Cisneros G (2019) On the influence of low-level visual features in film classification. PloS One 14(2):e0211406
38. Yu Y, Lu Z, Li Y, Liu D (2021) ASTS: attention based spatio-temporal sequential framework for movie trailer genre classification. Multimed Tools Appl 80:9749–64
39. Varghese J, Ramachandran Nair KN. (2019) A novel video genre classification algorithm by keyframe relevance. InInformation and Communication Technology for Intelligent Systems: Proceedings of ICTIS 2018, Volume 1 pp. 685-696. Springer Singapore
40. Choroś K (2019) Fast method of video genre categorization for temporally aggregated broadcast videos. J Intell Fuzzy Syst 37(6):7657–67
41. Yadav A, Vishwakarma DK (2020) A unified framework of deep networks for genre classification using movie trailer. Appl Soft Comput 1(96):106624
42. Jiang Y, Zheng L (2023) Deep learning for video game genre classification. Multimed Tools Appl 17:1–5
43. Mangolin RB, Pereira RM, Britto AS Jr, Silla CN Jr, Feltrim VD, Bertolini D, Costa YM (2022) A multimodal approach for multi-label movie genre classification. Multimed Tools Appl 81(14):19071–96
44. Behrouzi T, Toosi R, Akhaee MA (2023) Multimodal movie genre classification using recurrent neural network. Multimed Tools Appl 82(4):5763–84
45. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. (2009) Imagenet: A large-scale hierarchical image database. In2009 IEEE conference on computer vision and pattern recognition pp. 248-255
46. Tan M, Le Q. (2019) Efficientnet: Rethinking model scaling for convolutional neural networks. InInternational conference on machine learning. pp. 6105-6114. PMLR
47. Wehrmann J, Barros RC. (2017) Convolutions through time for multi-label movie genre classification. InProceedings of the Symposium on Applied Computing. pp. 114-119
48. Yang X, Esquivel JA (2023) Time-aware LSTM neural networks for dynamic personalized recommendation on business intelligence. Tsinghua Sci Technol 29(1):185–96
49. Mu Y, Wu Y (2023) Multimodal movie recommendation system using deep learning. Mathematics 11(4):895
50. Zhang Z, Gu Y, Plummer BA, Miao X, Liu J, Wang H. (2024) Movie genre classification by language augmentation and shot sampling. InProceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7275-7285
51. Tabatabaei S et al (2023) Attention transformer mechanism and fusion-based deep learning architecture for MRI brain tumor classification system. Biomed Signal Process Control 1(86):105119
52. Ullah W, Hussain T, Ullah FU, Lee MY, Baik SW (2023) TransCNN: Hybrid CNN and transformer mechanism for surveillance anomaly detection. Eng Appl Artif Intell 1(123):106173

Publisher's Note