




Multimodal movie genre classification using recurrent neural network

Tina Behrouzi¹ · Ramin Toosi² · Mohammad Ali Akhaee² 

Received: 28 December 2020 / Revised: 28 March 2022 / Accepted: 2 July 2022 /
Published online: 30 July 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

Genre is one of the features of a movie that defines its structure and type of audience. The number of streaming companies interested in automatically deriving movies' genres is rapidly increasing. Genre categorization of trailers is a challenging problem because of the conceptual nature of the genre, which is not presented physically within a frame and can only be perceived by the whole trailer. Moreover, several genres may appear in the movie at the same time. The multi-label learning algorithms have not been improved as significantly as the single-label classification models, which causes the genre categorization problem to be highly complicated. In this paper, we propose a novel multi-modal deep recurrent model for movie genre classification. A new structure based on Gated Recurrent Unit (GRU) is designed to derive spatial-temporal features of movie frames. The video features are then concatenated with the audio features to predict the final genres of the movie. The proposed design outperforms the state-of-art models based on accuracy and computational cost and substantially improves the movie genre classifier system's performance.

Keywords Movie genre detection · Multi-label classification · Gated recurrent unit (GRU) · Long short-term memory (LSTM) · 1D Convolutional neural network (1D_Conv)

1 Introduction

Due to the ease of preparation and distribution, increasing the production of images and videos has made them one of the most important sources of information. Thus, image and

✉ Mohammad Ali Akhaee
akhaee@ut.ac.ir

Tina Behrouzi
tina.behrouzi@mail.utoronto.ca

Ramin Toosi
r.toosi@ut.ac.ir

¹ Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada

² School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran

video processing is getting more attention currently. Understanding the content of images is widely studied in recent years, such as image classification [37], and manifold learning [9, 10, 20]. However, video achievements are less successful. Video understanding is a more challenging problem because of the temporal dimension of video alongside its spatial ones. Solutions designed for images still cannot be generalized to videos. For example, creating a video as individual frames is simple but usually results in less accurate systems.

One of the subsets of video understanding is video classification, in which we try to classify the video based on a concept. Action recognition [42] and genre detection [5] are two examples of this kind. Video classification has numerous applications in the computer vision field, such as video retrieval and recommendation systems. The success of convolutional neural networks (CNNs) in computer vision has also spread to this field and has made many improvements. While there are many research activities in this field, video classification is still a challenging task that requires more attention.

Trailer genre classification (or detection) is an example of a video classification problem. Genre classification aims to categorize movies based on their genres, such as drama and comedy, using the content provided by the movie trailer. Movie genres are still tagged manually by users, and the Internet Movie Database (IMDB) decides the final genre of a movie based on the suggestions of the users and critics. Moreover, big movie stream service providers such as Netflix and Hulu employ the trailer genre for movie recommendation and document categorization, among other applications. Therefore, the ability to predict genre automatically is getting more attention. Compared to other computer vision fields, such as action recognition or object tracking, two main challenges of this problem arise from the holistic and overlapping genres' nature. The first challenge is that the media does not physically express its genres. The movie genre is a concept that must be perceived from the entire video content, not just from a frame or a few shots. The second significant challenge is the multilabel [23] nature of genre classification. A movie usually has more than one genre at the same time.

The movie has a time-sequential dependency between its frames, best represented by neural networks that consider both spatial and temporal features. A great number of deep learning models have been developed to handle complex sequence problems for applications such as classification, prediction [45], and data generation [29]. Recurrent Neural Networks (RNN) are a powerful family of neural networks which try to address the same problem by considering past and current data information [19]. In this paper, we investigated and designed a deep 1D convolutional network and two of the most popular RNN structures, i.e., LSTM and GRU, for movie genre classification. These models try to capture meaningful features of sequential and spatial data streams. The effectiveness of considering different movies' modalities [3, 34] for genre detection is also investigated in this work.

The contributions of this paper are twofold: first, we investigate and propose a novel GRU model for multi-modal movie genre classification based on movie acoustic and visual features, which is not examined in previous works. The GRU model is added to the SVM network to lessen the influence of unbalance data problem. Second, the proposed model is compared to other famous spatio-temporal networks and also state-of-the-art models. Higher accuracy with less computational cost is achieved compared to other models. The method is implemented in Python, and the code is freely available online.¹

The rest of the paper is organized as follows. In Section 3, the details of the proposed method are demonstrated. Section 2 summarizes state-of-the-art models for movie genre

¹<https://github.com/Tinbeh97/MovieGenre>

classification. The dataset and experiments are included in Section 4. Finally, Section 5 concludes the paper and discusses the future direction.

2 Related work

The attention and importance of automatic movie genre classification are increasing, and several works have addressed this problem. Rasheed et al. [25] proposes a genre detection method by applying the mean shift classifier and low-level features such as average shot length and color variance. A neural network classifier with both visual and audio features to solve a single label genre detection problem is considered in [17]. Huang and Wang also used both visual and audio features alongside with SVM classifier in their method [15].

Taking advantage of image descriptors and extracting high-level visual features is suggested by Zhou et al. [44]. The image descriptors include Gist [22], CENTRIST [41], and w-CENTRIST. On the other hand, a simple K-nearest neighbor classifier is exploited to predict the genre. The ConvNet is used as an image descriptor in [30]. In this work, the extracted features are employed to form semantic histograms of the scene. These histograms with MFCC features are fed to an SVM classifier for prediction. Similarly, an SVM classifier with a set of different networks as descriptors is used in [39]. This idea is extended by adding a deep neural architecture to learn features through time [38]. Thus, their network learns spatial-temporal features simultaneously. Ben et al. [4] applied the SoundNet [1] and ResNet-125 [11] as audio and visual descriptors. The temporal aspect of visual features is further evaluated using the LSTM network.

Alvarez et al. [2] claim that better results can be achieved when considering only videos' low-level features, such as shot length and black and white rates. However, this model only considers movies with one genre label, which limits genre classification applications considerably. On the other hand, the importance of various modalities, which are movie's image, audio, synapses, poster, and subtitle, is examined in [21]. The best result is achieved considering the synopses and frames' features derived from the LSTM and CNN network. The final genres are determined using probability production of the multimodal features. Yu et al. [43] proposes an Attention-based Spatio-temporal Sequential Framework which consists of two main parts. First, movie frames' high-level features are extracted employing a deep CNN network. Then, a bi-LSTM attention model decides the final genres.

A probabilistic approach based on the significance of each background scene for each video category is introduced in [36]. Their proposed method consists of two steps: I) training an SVM classifier for scene classification II) video classification by considering the relevance of the key-frames scenes. The usefulness of shot length analysis is investigated by Choros [6]. Choros determined that different genres have distinguishable shot lengths.

Due to the intricateness of current film production, usually, each movie has several genres, which makes the problem a multi-label classification. Moreover, the unequal number of samples of each genre can lead to unfairness in genre detection. In this paper, we defined a multi-modal movie genre classification to address the above concerns. First, we classify genres only based on high-level visual features. The gated recurrent unit (GRU) captures the conceptual meaning of genres through frames. Subsequently, the final visual features are fused with acoustic properties to distinguish between different genres more robustly. Our proposed method considerably improves the performance of classes with less sample size and exceeds other state-of-the-art models' performance for the movie's genre classification.

3 Proposed method

In this section, we explain the details of four proposed methods, 1D_Conv_V, GRU_V, 1D_Conv+SVM_M, and GRU+SVM_M, for classifying movies’ genres. The block diagram of these four methods is shown in Fig. 1. First, both visual and acoustic features derived from the trailers are introduced. Then, GRU and 1D_Conv methods for extracting high-level visual characteristics are presented. Additionally, these two methods are employed to predict a movie’s genres. Finally, visual and audio features are fused, and a multi-label SVM model is applied to classify the movie’s ultimate genre.

3.1 Feature extraction

For both audio and visual features, the first five seconds of a trailer are trimmed because these sections of the video do not customarily contain useful information.

3.1.1 Visual features

Scenes are detected based on the color histogram of the successive images, in which two frames have a minimum color similarity. From the middle frame of the two consecutive shots, key-frames are obtained [25]. Eventually, 240 key-frames are derived from each movie. For some trailers, the percentage of key-frames is increased to meet the criterion of 240 key-frames.

Using raw frames of the video without deriving its high-level features does not perform well in the network [26, 32]. For extracting visual features of the video, VGG16 [27] and Resnet_152 [11] backbones are used. However, the performance of the VGG16 model is not as significant as the Resnet_152. Thus, the results of the VGG16 backbone are not reported. The last layer of Resnet_152 (softmax layer) is popped out, and the keyframes are fed into the Resnet_152 model. The final dimension of each frame’s features, V , is 2048. In other words, $V \in \mathbb{R}^{2048}$.

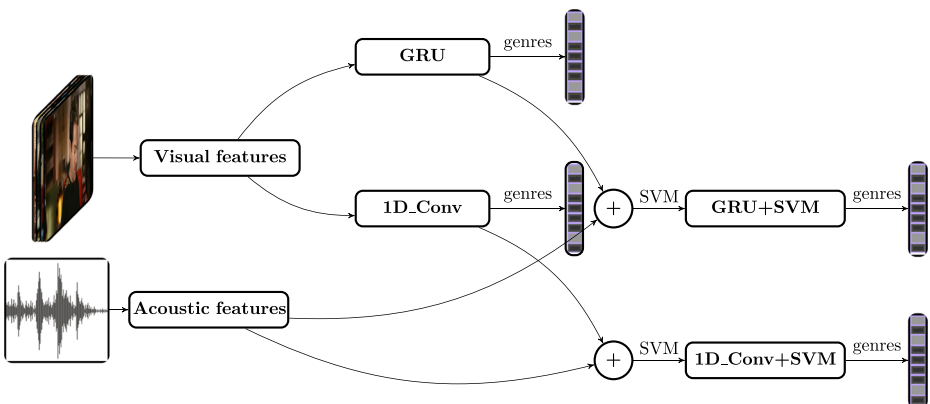


Fig. 1 The block diagram of the proposed methods

3.1.2 Acoustic features

The movie's background music and its moving images have different modalities. Sound's features are extracted from the trailers to bring visual and auditory modalities into the common space. Consequently, visual and acoustic features can be fused.

Usually, different genres of movies have specific music timbres. Timber is the quality of tone that helps the listener distinguish between sound productions that have different pitch and loudness [40]. For example, action movies' timber is usually bright, and its texture consists of rising melodies [33]. To represent a sound's timber, Mel-frequency cepstral coefficient (MFCC) and linear predictive coding (LPC) features are derived from the audio of the trailer [28].

MFCC with 13 Mel-frequency coefficients is computed for each movie and displayed in Fig. 2. MFCC coefficients are shown with $C_{i,j}$, where i and j represent the mel coefficient of each frame and time, respectively. Instead of considering the whole spectrum, the mean of MFCC and its delta in five equal time sections is considered. The delta of MFCC reflects the variation in time and is defined as (1). The audio of the movie consists of speech and music, which changes dynamically with time; therefore, the delta is an appropriate representation for our database, as shown in Fig. 2. Moreover, the average of MFCC and its delta is computed to reduce the complexity of the system, leading to less simulation time.

$$MFCC_{\Delta_{i,j}} = \frac{\sum_{n=1}^N n \times (C_{i,j+n} - C_{i,j-n})}{2 \times \sum_{n=1}^N n^2} \quad (1)$$

Another acoustic feature that we use is 9th-order LPC. LPC is widely applied in speech recognition because of its prediction performance [14]. From each trailer, we collect 10 LPC and 130 MFCC features. They are concatenated to create a vector of 140 nodes for each trailer; to rephrase it, $A \in \mathbb{R}^{140}$.

By combining sound A and visual features V the final network input length is 2188.

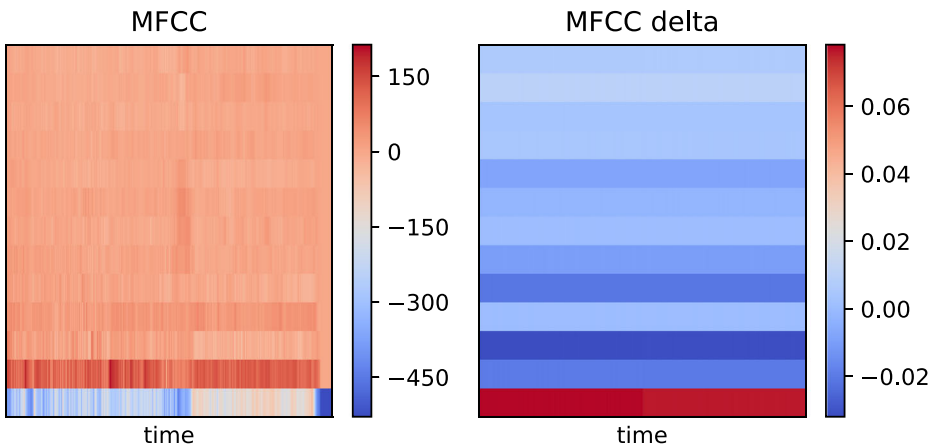


Fig. 2 MFCC and MFCC-delta are based on a random trailer. The time is between 5 seconds to 2:20 minutes of the movie

3.2 Visual-based classification

Two models are designed to predict the genres of a trailer based on its visual features. At the final layer of both models, the sigmoid activation function is applied because of the multi-label nature of the genre classification problem. The sigmoid function represents each class as a binary classification.

3.2.1 GRU model

RNNs can capture the time dependencies of the time-series data which make them highly compatible for deriving spatio-temporal features of consecutive movie's frames. To derive visual embedding of a movie, we developed and optimized two of the most known RNN networks, GRU and Long Short Term Memory (LSTM) [7, 13]. Compared with LSTM, GRU has two fewer gates and thus less number of parameters. For movie genre classification our novel GRU based model results in less computational cost and more accuracy. The full comparison of the designed GRU and LSTM models is illustrated in Appendix A

The GRU network overcomes the vanishing gradient problem of a simple RNN network when employing the back-propagation algorithm. GRU consists of two gates, reset and update gates, which address the vanishing problem. A reset gate is responsible to decide what past or current information to keep, (2). An update gate determines the usefulness of the past information, (3). W , U , b are the weight matrices and bias vector of i th layer. x_t is the input vector at time unit t .

$$r_t^i = \sigma(W_r^i x_t^i + U_r^i h_{t-1}^i + b_r^i) \quad (2)$$

$$z_t^i = \sigma(W_z^i x_t^i + U_z^i h_{t-1}^i + b_z^i) \quad (3)$$

The output vector h_t^i and candidate activation vector \hat{h}_t^i are defined by:

$$h_t^i = (1 - z_t^i) \odot h_{t-1}^i + z_t^i \odot \hat{h}_t^i \quad (4)$$

$$\hat{h}_t^i = \tanh(W_h^i x_t^i + U_h^i (r_t^i \odot h_{t-1}^i) + b_h^i) \quad (5)$$

Where \odot denotes the Hadamard product.

The description of the proposed GRU model is shown in Table 1. First, the visual features of 240 trailer key-frames are extracted, as described in Section 3.1.1. The features are fed into a GRU layer with an output vector sequence corresponding to each time frame. The comparison of feeding all frames to GRU layer and a parallel structure (separating frames

Table 1 Proposed GRU model layer description. The None value corresponds to the batch size and #param is the number of parameters

Layer	Output shape	#param
InputLayer	(None, 240, 2048)	0
GRU	(None, 240, 120)	781200
Max pooling	(None, 80, 120)	0
GRU	(None, 64)	35712
Dense	(None, L)	585
Total params: 817,497		
Trainable params: 817,497		

into 12 consecutive inputs of nine frames) is demonstrated in Appendix B. The max-pooling is applied after the GRU layer to collect important information of the layer and to avoid overfitting. The pooling size and stride number for the max pooling layer is 3 and 0, respectively. The GRU layer is employed to predict final features of the last frame. The final layer of the network has L dimensions to represent the movie’s genre.

3.2.2 1D convolution model

Considering variations in time is crucial for detecting the movie’s genre. The 1D convolutional neural network regards the whole frame together and computes the contrast between different frames. However, in convolutional networks, the differences are measured based on the kernel size. The proposed 1D.Conv system is shown in Fig. 3.

The input of the system is visual features of all 240 trailers’ keyframes. The batch normalization is employed to reduce the covariance shift and decrease the simulation time [16]. Moreover, the max-pooling and dropout layers help to avoid overfitting and lower the complexity of the system. The probability of each genre is obtained using the final FC layer with a sigmoid activation function.

For calculating the error in backpropagation, binary cross-entropy, which is highly compatible with multi-label classification, is utilized [18].

3.3 SVM

Since the derived visual and audio features have different ranges, the acoustic/visual data scale is linearly transformed by the (6) between zero and one to bring both modalities into the same range. In this equation, A_N is the normalized audio/visual features. The normalization helps the network to be unbiased to the modality variance of data.

$$A_N = \frac{A - \min(A)}{\max(A) - \min(A)} \tag{6}$$

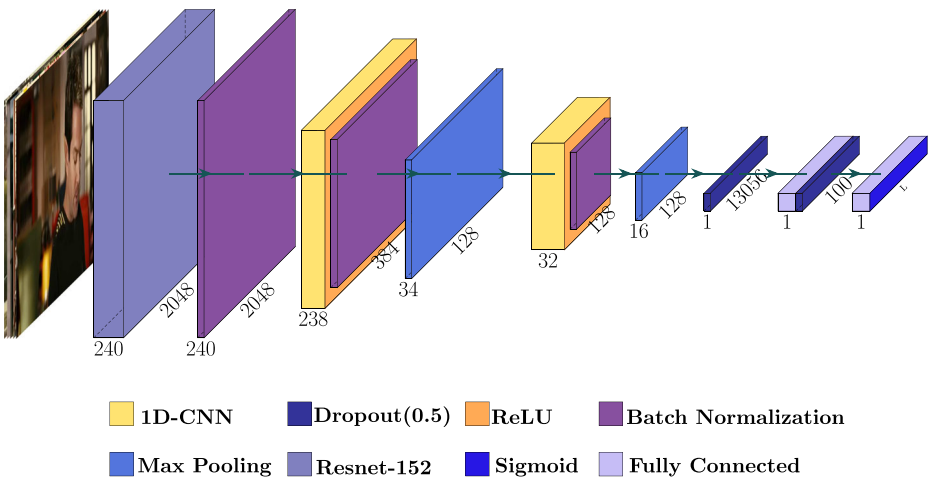


Fig. 3 1D convolutional architecture for visual-based movie’s genre classification. The input of the system is the whole 240 keyframes’ features, $Input \in \mathbb{R}^{240 \times 2048}$

SVM [8] is a well-known method to classify audio signals using MFCC and LPC features. SVM can reduce the effect of class imbalance problem. Combining neural network and SVM helps to improve the accuracy of minority classes, which is more illustrated in the results section. Therefore, the final genre is decided by applying the fusion of two properties to the SVM network. Our classes are multi-label, thus, we treat each genre as a different problem and fed them into One-vs-One SVM. The radial basis function (RBF) is employed as a kernel because the RBF hyperplane decision boundary leads to better results in our experiments than the linear kernel.

Recently, CNN is getting very popular for classifying the spectrogram of audio represented as 2D input [12]. However, for our problem, the SVM results are better than CNN, based on accuracy, by almost two percent.

The results of our proposed method are mentioned in the following. These four methods are 1D_Conv_V, GRU_V, 1D-Conv+SVM_M, and GRU+SVM_M. The GRU_V and 1D-Conv_V methods, which are introduced in Sections 3.2.1 and 3.2.2, respectively, classify movie genres only based on the 2048 visual features of each frame. The V in these two models stands for visual based classification. For the two other methods, the final layer of GRU and 1D-Conv models before applying the thresholds are concatenated with acoustic features leading to a total of $L + 140$ features. The resulting image and sound's properties are passed through SVM, defined in Section 3.3, in order to create GRU+SVM_M and 1D-Conv+SVM_M models; where M represents the multimodal classification.

4 Experiment

In this section, first, we describe the applied dataset; we also discuss the hyperparameters and decision boundary applied to optimize the results. Later, the equations employed to analyze our system are explained. Finally, the performance of the proposed method is compared with state-of-the-art methods in terms of the AUC score and the Hamming loss. In all tables, the best performance in each metric is marked as bold.

4.1 Dataset

We used the multi-label trailer database, LMTD [30], which consists of 3500 movies. As far as we know, LMTD is the largest dataset available up to this day with a great variation of trailers. The database is split into 2874 training, 773 testing, and 374 validation sets. Each trailer is classified into nine genres. The classes considered here are action, adventure, comedy, crime, drama, horror, romance, sci-fi, and thriller. The distribution of genres is shown in Fig. 4. We also extract the four most popular genres for further analysis, since many applications only require these classes. The four main labels are action, drama, horror, and romance.

4.2 Implementation details

In both GRU and 1D_Conv models proposed in Section 3.2, the batch-size of 32 is considered. Moreover, the simulation stops when the validation error increases in five consecutive epochs. The maximum epoch of 100 is applied for both models. The final number of epochs for each model is shown in Fig. 5. The GRU merge and LSTM merge are explained in Appendices B and A, respectively.

The proposed method is implemented in TensorFlow. Adam optimizer [18] with initial learning rate of $1e^{-4}$ is set to train the model. For the network parameters, 1 to 4 layers (not

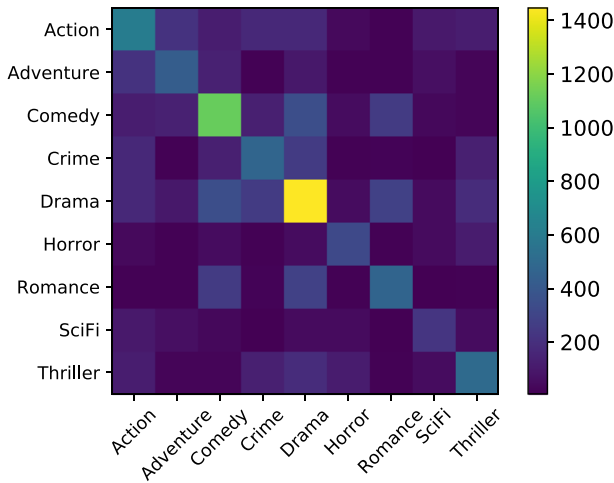


Fig. 4 Training data labels’ distribution

considering max pooling and dropout) with layer size 240, 120, 64, 32, and 16 is considered for each model.

The proper hyper-parameters of SVM are computed using the grid-search; where the range of C is between 10^{-2} and 10^5 with 14 points in the logarithmic axis. Moreover, the gamma range is determined from 14 logarithmically spaced points between 10^{-5} and 1. In order to determine the deciding decision-boundary, the ROC-curve of the training results is considered [24]. The curve indicates a true positive rate (TPR) and a false negative rate (FNR) for different thresholds. The optimal boundary, \hat{t} , is estimated for each genre exclusively as (7).

$$\hat{t} = \underset{t_i}{\operatorname{argmax}}(TPR_i - FNR_i) \tag{7}$$

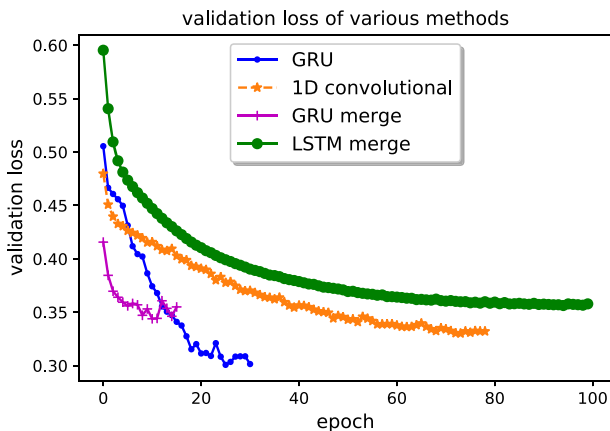


Fig. 5 Validation error of each epoch

4.3 Evaluation metrics

For evaluating the results, we analyzed the hamming loss, area under the curve (AUC) score, F1 score, and confusion matrix. We applied these metrics because they are highly employed in multi-label classifications and they can represent a fair estimation for multi-label data. These quantitative metrics are used in the previous works and thus it allows us to compare our results with the state-of-the-art methods. The hamming loss gives us the overall accuracy of the network and is defined as below:

$$HL = \frac{\sum_{q=1}^N \sum_{j=1}^L y_{pred_{qj}} \oplus y_{true_{qj}}}{NL} \quad (8)$$

Where HL is the hamming loss, $y_{true_{qj}}$ and $y_{pred_{qj}}$ represent the q th test data's label and output's prediction of the j th genre, respectively.

The well-known AUC score is determined for each genre, which allows us to compare the accuracy of each class separately. Moreover, to globally evaluate the performance of different methods, the weighted, micro, and macro averages of the AUC metric is estimated. The macro and weighted averages compute the AUC rate of each class independently and take their unweighted and weighted means, respectively. While micro-average calculates the AUC mean without considering labels. The F1-score determines the harmonic mean of the precision and recall metrics. We reported the weighted average of the genres' F1 score for each model.

By multi-label confusion matrix, we want to see the effectiveness of different genres on each other. In other words, to see which class helps to reduce the uncertainty or increase the confidence of another class prediction. Therefore, the confusion matrix is defined in (9). In this formula, $\sim y_{true_{iq}} \oplus y_{true_{jq}}$ stands for outputs where both genres i and j have the same label. As a result, the defined confusion matrix demonstrates the number of times the genre i is incorrectly predicted when a movie does or does not include both classes i and j . This proposed equation distinguishes the effect of different genres on each other.

$$S_{ij} = \frac{\sum_{q=1}^N (\sim y_{true_{iq}} \oplus y_{true_{jq}})(y_{true_{iq}} \oplus y_{pred_{iq}})}{\sum_{q=1}^N (\sim y_{true_{iq}} \oplus y_{true_{iq}})} \quad (9)$$

4.4 Comparison

In this section, we evaluate our visual-based models, GRU_V and 1D_Conv_V, and multimodal models, GRU+SVM_M, and 1D_Conv+SVM_M, considering the Hamming loss, AUC score, F1 score, and confusion matrix. The state-of-the-art methods to which we compare our results are [2, 30, 31, 38], because they consider the same database for evaluation. Moreover, the computational cost and accuracy of different classifiers for the multimodal classification are examined. Finally, the experiments are divided into two parts, nine and four genres-based classification.

4.4.1 Nine-genre results

First, we compare the methods based on the Hamming loss and F1 score. As shown in Table 2, adding sound to the network has improved both visual-based proposed systems, i.e. 1D_Conv_V and GRU_V losses improved by 10 and 12 percent, respectively. Furthermore, the proposed model, the GRU+SVM_M network, has the lowest Hamming loss and F score.

Table 2 Comparison of the different methods based on Hamming loss

Model	Hamming loss	F-score
ID_Conv_V	0.242	0.61
ID_Conv+SVM_M	0.142	0.63
GRU_V	0.253	0.65
GRU+SVM_M (Proposed method)	0.133	0.66
Low Level Feature Model (LLFM) (uni-label) [2]	0.3168	-

The LLFM's loss is calculated based on its released network

We believe a GRU_V network is better able to derive semantic features of genres compared to the ID_Conv_V model.

Based on Table 3, most of the genres' scores have increased when the acoustic feature is included in the network. It indicates that the movies' audio can improve genre prediction significantly. However, the Scifi score has reduced after adding audio, which we believe it is due to the resemblance of action and sci-fi movies' music. The number of action movies is much higher than Scifi trailers based on Fig. 4, which can make the network tends to get confused between Scifi and action movies.

The GRU+SVM_M and ID_Conv+SVM_M have relatively close AUC score; however, the score of GRU+SVM_M model is higher for 5 out of 9 genres. Moreover, the AUC score of the GRU+SVM_M method in all classes, except for drama, is higher than the state-of-the-art models. Notably, GRU+SVM_M has improved the score of genres with fewer data in comparison with other genres, such as thriller and horror videos.

Table 4 indicates that the GRU+SVM_M outperforms states-of-art methods considering the micro, macro, and weighted AUC scores; 24, 8, and 8 percent higher than the previous methods, respectively. GRU+SVM_M and ID_Conv+SVM_M differ only by 1% but still GRU+SVM_M performs better than ID_Conv+SVM_M model.

Figure 6 shows that the movie's sound can reduce the genre confusion. For example, horror and comedy confusion matrices have similarities, but their sounds are very different. When the audio is included in the network, these two genres' confusion has decreased significantly; this improvement is around 50 percent.

Another exciting interpretation is that the error of labels with fewer data has altered more considerably than other tags. Therefore, the music of the scene makes it possible to reduce the bias of the network due to the class imbalance. For instance, the comedy error has changed from 0.193 to 0.169, while the sci-fi error has been significantly decreased from 0.23 to 0.072. As shown in Fig. 6 and Table 3, the confusion matrices and the individual genre's AUC score correspond to each other.

Table 5 indicates two neural networks model time and space complexity based on CPU @2.5 GHz Intel Core i7 processor. The GRU_V model has less number of parameters which means in each epoch less gradient requires to be computed which results in less training time per epoch. Moreover, GRU_V converges to its optimal point 48 epochs sooner than ID_Conv_V network. The storage required to store GRU_V model and weights is 71% less than ID_Conv_V model. This finding indicates that GRU_V model outperforms the ID_Conv_V model considering both computational cost and all scoring metrics.

Table 3 The state of the arts and the proposed methods' AUC score for each movie's genre

Model	Action	Adventure	Comedy	Crime	Drama	Horror	Romance	SciFi	Thriller
ID_Conv_V	0.777	0.807	0.803	0.659	0.733	0.81	0.655	0.787	0.763
ID_Conv+SVM_M	0.887	0.876	0.908	0.813	0.837	0.854	0.741	0.691	0.813
GRU_V	0.824	0.844	0.83	0.77	0.732	0.819	0.726	0.721	0.795
GRU+SVM_M	0.909	0.929	0.91	0.846	0.829	0.865	0.725	0.628	0.825
LLFM(uni-i-label) [2]	0.852	0.752	0.871	0.628	0.641	0.424	0.468	0.192	0.520
CTT-MMC-TN [38]	0.835	0.672	0.870	0.547	0.841	0.667	0.456	0.401	0.522
LSTM [38]	0.687	0.573	0.792	0.421	0.740	0.478	0.313	0.237	0.437

The input of all models is LMTD dataset [30]

Table 4 General genre AUC score. Micro, macro, and weighted represent the aggregation methods to derive AUC scores for all genres

Model	Micro	Macro	Weighted
ID_Conv_V	0.756	0.755	0.754
ID_Conv+SVM_M	0.870	0.825	0.843
GRU_V	0.794	0.785	0.783
GRU+SVM_M	0.88	0.824	0.847
LLFM (uni-label) [2]	0.506	0.594	0.762
CTT-MMC-TN [38]	0.646	0.742	0.724
LSTM [38]	0.520	0.640	0.590

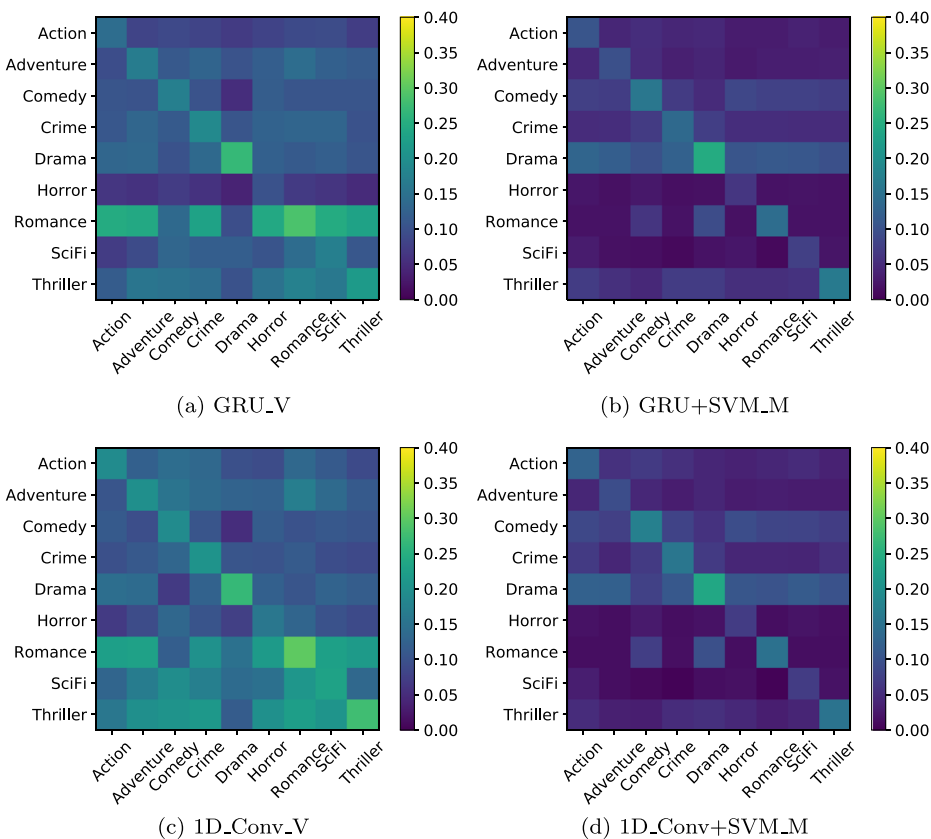


Fig. 6 Confusion matrix of the data fault prediction

Table 5 Space and time complexity of the 1D_Conv_V and GRU_V model

Model	time/epochs(s)	#epochs	#trainable parameters	storage
1D_Conv_V	59	79	2,714,097	11.5MB
GRU_V	41	31	817,497	3.3MB

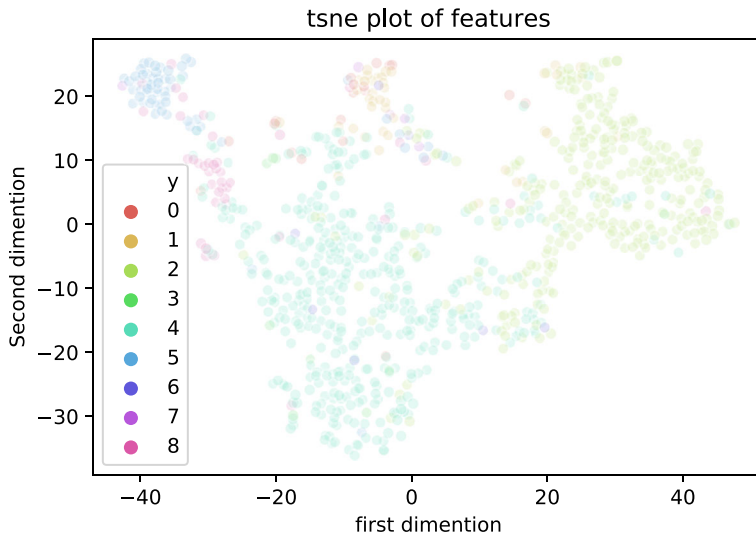


Fig. 7 T-SNE visualization of audio and video features for training data with only one genre. The legend labels from 0 to 8 are, respectively, as follows: Action ,Adventure ,Comedy ,Crime ,Drama ,Horror ,Romance ,Scifi , and Thriller

Table 6 Comparison of the accuracy and computational cost of the popular classification methods

Classifier	Hamming loss	t_{train} (s)	t_{test} (s)
SVM	0.133	21.766	0.878
KNN	0.136	0.07	1.771
Multinomial-NB	0.148	0.016	0.006

t_{train} and t_{test} correspond to training and testing time for whole training and testing data, respectively

Table 7 Comparison of different methods based on the Hamming loss

Model	Hamming loss	F-score
1D_Conv_V	0.21	0.73
1D_Conv+SVM_M	0.15	0.75
GRU_V	0.17	0.75
GRU+SVM_M	0.13	0.76
LSTM (uni-label) [31]	0.15	-
CNN-MoTion- P_{AHP} [30]	0.26	-

Table 8 Comparison of the 4 proposed methods’ AUC scores for the individual genre

Model	Action	Drama	Horror	Romance
ID_Conv_V	0.830	0.883	0.645	0.520
ID_Conv+SVM_M	0.930	0.804	0.904	0.784
GRU_V	0.866	0.798	0.766	0.694
GRU+SVM_M	0.925	0.842	0.916	0.821

Table 9 General genre AUC score

Model	Micro	Macro	Weighted
ID_Conv_V	0.834	0.745	0.809
ID_Conv+SVM_M	0.900	0.856	0.838
GRU_V	0.889	0.819	0.808
GRU+SVM_M	0.914	0.876	0.864

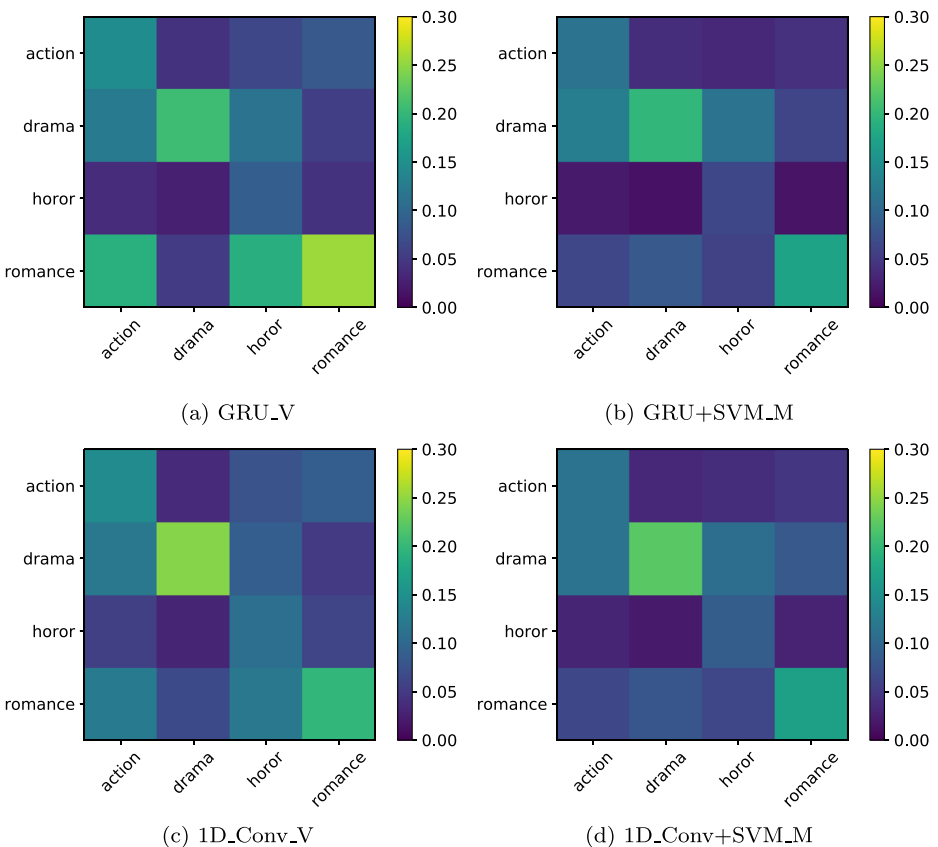


Fig. 8 Data fault prediction confusion matrix

The T-SNE method [35] is employed to visualize high-dimensional audio and video features in a two-dimensional space. Figure 7 illustrates that comedy and drama videos with more data are better distinguished from other labels. Because the problem is multi-label, we considered trailers with only one label for better virtualization.

In the final experiment, for the nine genres' problem, we employed three of the most famous machine learning supervised models in Table 6. These models are computationally optimal compared to deep neural networks and help to reduce the effect of unbalance data. The Multinomial Naive Bayes (Multinomial-NB) has least training and testing time because it only requires to calculate simple arithmetic calculation in both stages. However, this model considers features that are independent which results in more loss compared to other models. The k-nearest neighbors algorithm (KNN) with $k=50$ only stores features and its corresponding labels in the training phase, which cause very low training time. Moreover, KNN is a non-parametric model which requires low testing time but still higher than SVM model. Although SVM's training time is higher than the other two models, it has the least hamming loss. Therefore, we considered SVM model for the proposed classification method.

4.4.2 Four-genre results

Table 7 shows that similar to the result of nine-genre classification problem, adding sound to the system deducts the Hamming loss. Also, GRU+SVM_M method has the least error and the most F1 score compared to other methods.

Moreover, according to Table 8, multi-modal detection significantly increased the AUC score of genres with less data. Therefore, the bias of the system due to the database has been reduced. However, the score of drama movies with the most extensive data size has decreased after adding music's features to the 1D_Conv_V network.

Eventually, based on Table 9, the overall performance of the GRU+SVM_M method has improved compared to the other methods. All micro, macro, and weighted's scores of the GRU+SVM_M method are 1.4, 2, and 2.6 percent higher than the second-best results.

Figure 8 indicates that, like nine genre results, considering music for classification reduces the confusion of data, especially those with fewer data sizes like romance and horror movies. Moreover, the GRU+SVM_M method has the lowest confusion loss for all genres compared to the other methods.

4.5 Ablation experiment

We conducted both quantitative and qualitative experiments to indicate the importance of our multi-modal movie genre classification which considers both movie's sounds and video. Table 10 illustrates the qualitative performance of the proposed models for randomly selected trailers of the test set. Based on these examples, the false positive labels happen more commonly in the case of not considering movies' sound. In these cases, GRU+SVM_M model is the most accurate and has the least false positive and no true negative genres compared to other models.

5 Conclusion

In this paper, we proposed four methods for genre classification employing deep neural networks. Of these four, the GRU+SVM_M as a multi-modal method outperforms the other

Table 10 Examples of predictions in the test set


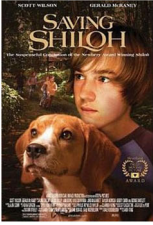



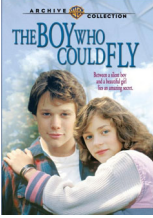
Movie	Genres
	Ground Truth 1D_Conv_V 1D_Conv+SVM_M GRU_V GRU+SVM_M Action Crime Drama Action Crime Drama Horror Scifi Thriller Action Crime Drama Thriller Action, Crime Drama Thriller Action Crime Drama Thriller
	Ground Truth 1D_Conv_V 1D_Conv+SVM_M GRU_V GRU+SVM_M Drama Drama Comedy Romance Drama Comedy Romance Drama Comedy Romance Drama
	Ground Truth 1D_Conv_V 1D_Conv+SVM_M GRU_V GRU+SVM_M Comedy Drama Romance Comedy Drama Romance Comedy Drama Romance Comedy Drama Romance Comedy Drama Romance
	Ground Truth 1D_Conv_V 1D_Conv+SVM_M GRU_V GRU+SVM_M Action Adventure Scifi Action Adventure Scifi Horror Thriller Action Adventure Scifi Action Adventure Scifi Thriller Action Adventure Scifi
	Ground Truth 1D_Conv_V 1D_Conv+SVM_M GRU_V GRU+SVM_M Comedy Horror Comedy Horror Romance Scifi Thriller Comedy Horror Romance Scifi Thriller Comedy Horror Comedy Horror

Table 10 (continued)

Movie	Genres	
	Ground Truth	Drama
	1D_Conv_V	Drama Comedy
	1D_Conv+SVM_M	Drama
	GRU_V	Drama Comedy Horror Thriller
	GRU+SVM_M	Drama

Blue and red colors represent predicted false positive and true negative genres, respectively

ones. In this method, first, high-level features of the sequence of frames are passed through the GRU model to address the variation importance of the visual properties in time. Subsequently, the fusion of the GRU output and audio features is fed into the SVM, creating GRU+SVM_M. The SVM method outperforms other machine learning and neural network methods.

We have shown that considering both the sound and image of a movie improves the performance of movie genre classification. The Hamming loss has been improved between 10 to 12 percent after including trailers' sound for classification. Notably, acoustic features enhance the accuracy of genres with low performance due to their rarity in the database. This result indicates that we have reduced the effect of our problem's main limitation, which is the unfairness of the number of genres in the dataset. As experiments show, the average, macro, and micro AUC scores of the GRU+SVM_M method for the nine genres are, respectively, 6, 8, and 24 percent higher than the best performance of the state-of-the-art methods on the same dataset.

One of the challenges of movie genre classification is that different movie scenes have different genres, and there is no comprehensive database that contains the genres of each scene. In the future, we intend to extend the current method for the whole movie based on semi-supervised learning techniques. Furthermore, we aim to include movie subtitles to determine their effect on classification.

Appendix A: LSTM merge

LSTM network is a well-known system for classifying data that changes through time and can learn long-term dependencies inside a sequence [13]. The designed LSTM model is presented in Fig. 9, where L represents the number of classes. Tables 11 and 12 indicate that LSTM+SVM network has less score and more hamming loss than both GRU+SVM and 1D_Conv+SVM models.

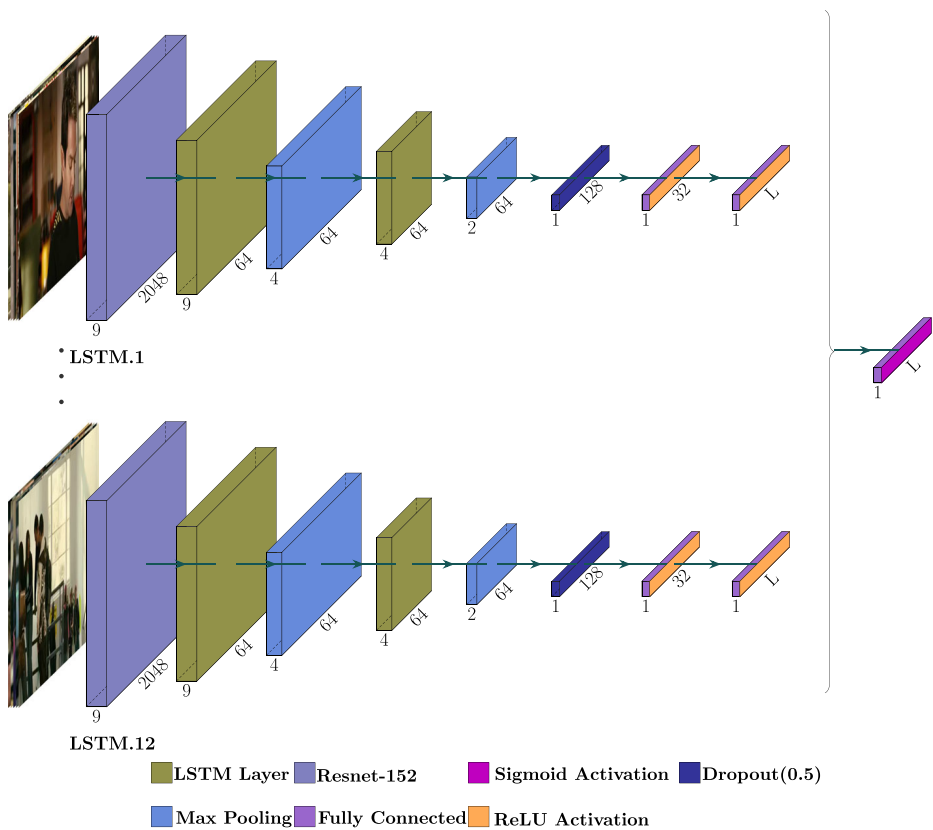


Fig. 9 LSTM model. LSTM.*i* $i = 1, \dots, 12$ indicates *i*th parallel LSTM network. The outputs of these 12 networks are merged and fed into the final sigmoid layer

Table 11 LSTM network hamming loss and AUC micro, macro, and weighted scores

Model	Hamming loss	Micro	Macro	Weighted
LSTM	0.2530	0.754	0.741	0.724
LSTM+SVM	0.1593	0.876	0.824	0.814

Table 12 LSTM network AUC score for 9 genres data

genre precision	Action	Adventure	Comedy	Crime	Drama	Horror	Romance	SciFi	Thriller
LSTM	0.771	0.815	0.776	0.698	0.669	0.759	0.678	0.670	0.739
LSTM+SVM	0.820	0.893	0.870	0.742	0.740	0.844	0.625	0.512	0.794

Appendix B: GRU merge

Figure 10 indicates the network structure of the GRU merge model. Separating frames into 12 consecutive inputs of nine frames reduces the computational complexity Fig. 5 but increases the Hamming loss and decreases all AUC and F1 scores, Tables 13 and 14. For GRU merge, the training time required to train each epoch reduces to half compared to the proposed GRU model.

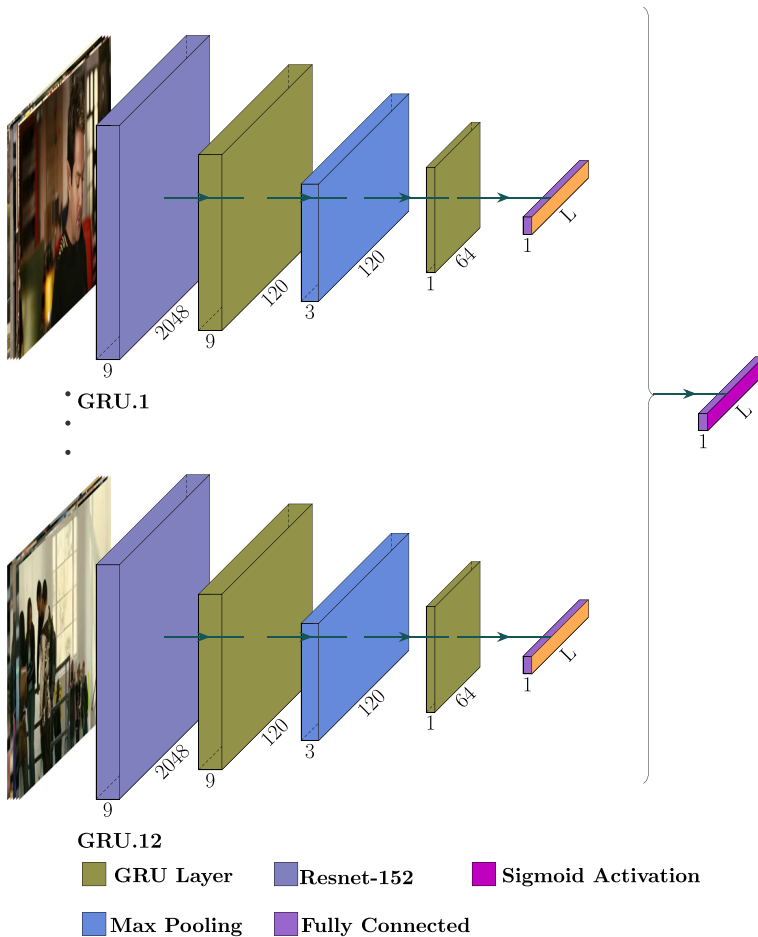


Fig. 10 GRU merge model description. GRU.i $i = 1, \dots, 12$ indicates i th parallel GRU network. The outputs of these 12 networks are merged and fed into the final sigmoid layer

Table 13 GRU merge network hamming loss and AUC micro, macro, and weighted scores

Model	Hamming loss	Micro	Macro	Weighted	F1 score
GRU	0.253	0.838	0.781	0.795	0.611
GRU+SVM	0.16	0.869	0.836	0.831	0.618

Table 14 GRU merge network AUC score for 9 genres data

genre	precision	Action	Adventure	90Comedy	Crime	Drama	Horror	Romance	SciFi	Thriller
GRU		0.831	0.855	0.877	0.762	0.78	0.847	0.677	0.699	0.704
GRU+SVM		0.866	0.892	0.885	0.795	0.775	0.873	0.798	0.815	0.824

Declarations

Competing interests The authors declare they have no competing interests.

References

- Aytar Y, Vondrick C, Torralba A (2016) Soundnet: Learning sound representations from unlabeled video. In: Advances in neural information processing systems, pp 892–900
- 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
- Badamdorj T, Rochan M, Wang Y, Cheng L (2021) Joint visual and audio learning for video highlight detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 8127–8137
- Ben-Ahmed O, Huet B (2018) Deep multimodal features for movie genre and interestingness prediction. In: 2018 International Conference on Content-Based Multimedia Indexing (CBMI), IEEE, pp 1–6
- Bhoraniya DM, Ratanpara TV (2017) A survey on video genre classification techniques. In: 2017 International conference on intelligent computing and control (I2C2), IEEE, pp 1–5
- Choroś K (2019) Fast method of video genre categorization for temporally aggregated broadcast videos. *Journal of intelligent & fuzzy systems*. Preprint, pp 1–11
- Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555
- Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20(3):273–297
- Fu S, Liu W, Tao D, Zhou Y, Nie L (2020) hesGCN: Hessian graph convolutional networks for semi-supervised classification. *Inf Sci* 514:484–498
- Fu S, Liu W, Zhang K, Zhou Y (2021) Example-feature graph convolutional networks for semi-supervised classification. *Neurocomputing* 461:63–76
- He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition, pp 770–778
- Hershey S, Chaudhuri S, Ellis DP, Gemmeke JF, Jansen A, Moore RC, Plakal M, Platt D, Saurous RA, Seybold B et al (2017) CNN architectures for large-scale audio classification. In: 2017 IEEE international conference on acoustics, speech and signal processing (icassp), IEEE, pp 131–135
- Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9(8):1735–1780
- Huang X, Acero A, Hon H-W, Foreword By-Reddy R (2001) Spoken language processing: A guide to theory, algorithm, and system development Prentice hall PTR
- Huang Y-F, Wang S-H (2012) Movie genre classification using svm with audio and video features. In: International conference on active media technology, Springer, pp 1–10
- Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167
- Jain SK, Jadon R (2009) Movies genres classifier using neural network. In: 2009 24th International Symposium on Computer and Information Sciences, IEEE, pp 575–580
- Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv:1412.6980
- Li Y, Tarlow D, Brockschmidt M, Zemel R (2015) Gated graph sequence neural networks. arXiv:1511.05493
- Liu W, Ma X, Zhou Y, Tao D, Cheng J (2018) p -laplacian regularization for scene recognition. *IEEE Trans Cybern* 49(8):2927–2940
- Mangolin RB, Pereira RM, Britto AS, Silla CN, Feltrim VD, Bertolini D, Costa Y M (2020) A multimodal approach for multi-label movie genre classification. *Multimedia Tools and Applications*, pp 1–26

22. Oliva A, Torralba A (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int J Comput Vis* 42(3):145–175
23. Pant P, Sabitha AS, Choudhury T, Dhingra P (2019) Multi-label classification trending challenges and approaches. In: *Emerging trends in expert applications and security*. Springer, pp 433–444
24. Pillai I, Fumera G, Roli F (2013) Threshold optimisation for multi-label classifiers. *Pattern Recogn* 46(7):2055–2065
25. 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
26. Rawat W, Wang Z (2017) Deep convolutional neural networks for image classification: A comprehensive review. *Neural Comput* 29(9):2352–2449
27. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. [arXiv:1409.1556](https://arxiv.org/abs/1409.1556)
28. Schwarz D, O’Leary S (2015) Smooth granular sound texture synthesis by control of timbral similarity. In *Sound and Music Computing (SMC)*, p. 6
29. Serban IV, Sordoni A, Bengio Y, Courville A, Pineau J (2015) Hierarchical neural network generative models for movie dialogues. [arXiv:1507.04808](https://arxiv.org/abs/1507.04808). 7(8), pp 434–441
30. Simões GS, Wehrmann J, Barros RC, Ruiz DD (2016) Movie genre classification with convolutional neural networks. In: *2016 International joint conference on neural networks (IJCNN)*, IEEE, pp 259–266
31. Sivaraman K, Somappa G (2016) *Moviescope: Movie trailer classification using deep neural networks* University of Virginia
32. Srinivas S, Sarvadevabhatla RK, Mopuri KR, Prabhu N, Kruthiventi SS, Babu RV (2016) A taxonomy of deep convolutional neural nets for computer vision. *Front Robot AI* 2:36
33. Thompson K, Smith J (2008) *Film art: An introduction* McGraw-Hill Higher Education
34. Tian Y, Xu C (2021) Can audio-visual integration strengthen robustness under multimodal attacks? In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 5601–5611
35. Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9:11
36. Varghese J, Nair KR (2019) A novel video genre classification algorithm by keyframe relevance. In: *Information and communication technology for intelligent systems*. Springer, pp 685–696
37. 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
38. Wehrmann J, Barros RC (2017) Movie genre classification: A multi-label approach based on convolutions through time. *Appl Soft Comput* 61:973–982
39. Wehrmann J, Barros RC, Simões GS, Paula TS, Ruiz DD (2016) (Deep) learning from frames. In: *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, IEEE, pp 1–6
40. Wessel DL (1979) Timbre space as a musical control structure. *Computer music journal*, pp 45–52
41. Wu J, Rehg JM (2008) Where am I: Place instance and category recognition using spatial PACT. In: *2008 IEEE Conference on computer vision and pattern recognition*, IEEE, pp 1–8
42. Yi Y, Li A, Zhou X (2020) Human action recognition based on action relevance weighted encoding. *Signal Process Image Commun* 80:115640
43. 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(7):9749–9764
44. Zhou H, Hermans T, Karandikar AV, Rehg JM (2010) Movie genre classification via scene categorization. In: *Proceedings of the 18th ACM international conference on multimedia*, pp 747–750
45. Zhou Y, Zhang L, Yi Z (2019) Predicting movie box-office revenues using deep neural networks. *Neural Comput & Applic* 31(6):1855–1865

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.