Abstract This work analyzes a machine committee for content-based automatic movie genre classification. The committee is consisted of diverse deep-learning architectures, each aimed at one different information type (text, poster, video or audio). This approach was evaluated in the LMTD9 dataset which is composed by movie trailers. Results show that the multi-modal proposal leads to prediction results that are comparable to the state of the art, with reduced inter-class accuracy variance. Information type influences classification in a different way for each genre under study. Additionally, classification accuracy show that text information is highly relevant for movie genre classification. This indicates that multi-modal classification is greatly relevant for the problem of automatic movie genre classification.

Keywords Multimedia Information Analysis, Movie genre classification, Deep learning architectures, Multi-label classification, Natural Language Processing

The relevance of text, image, video, audio for automatic classification of trailers movie genres

Paulo Renato de Faria · Tiago Fernandes Tavares

Received: date / Accepted: date

1 Introduction

Recent advances on massive video distribution such as YouTube, Netflix, Metacafe, Crackle, and Vimeo have fostered a large growth in the generation of media collections. These collections are often organized using text labels. They can contain metadata that highlights some of their relevant characteristics, e.g., the genres corresponding to a particular movie.

Labeling videos can be time-consuming due to the high media generation rate in social networks. However, the organization of video collections is related to a consumption demand and to marketing deadlines, thus it must be quickly accomplished. Also, consumers and cinephiles can benefit from this organization method to improve/organize their own personal collections. This demand brought researchers to explore Automatic Movie Genre Classification as a tool to facilitate the organization, navigation, and labeling of movie catalogues Rasheed et al. [2005].

Movies are often related to multi-modal information, which comprises text (title and plot descriptions), image (posters), audio (trailer audio or soundtracks) and video. Up to the authors knowledge, this combination of information modes has not been used before for movie labelling.

In principle, all modes can provide useful information for an automatic movie genre labeling process. However, using less modes can lead to faster labeling processes, which is desirable in the organization of personal collections or large datasets.

In this paper, we investigate the trade-off between the labeling accuracy and the use of information mode diversity. For such, isolated state-of-the-art classifiers for each of the information modes were implemented and then combined using both early-fusion and late-fusion strategies. The proposed system was evaluated in the large movie trailer dataset (LMTD9) Wehrmann and Barros [2017] dataset, which has been used in other recent automatic movie labelling proposals Chu and Guo [2017], Minami et al. [2018], Alvarez et al. [2019] and Choroś [2018].

The results show that the multi-modal system yields accuracy that is close to the state-of-the-art. However, the inter-class accuracy variance is sensibly lower, which indicates a greater generalization capability of the proposed system. Also, the analysis of the label-wise accuracy shows that some information modes are more effective regarding particular genres.

This paper presents important contributions for the multimedia information retrieval, as follows:

1. We propose a multi-modal movie labelling system based on both text, image, audio, and video, which is a combination that has not been experimented before;
2. We perform a label-wise evaluation of our proposal, demonstrating that the addition of multiple modes can favour the system generalization capability in unbalanced datasets;
3. We evaluate the impact of each information mode to the overall accuracy, can be useful for the development of smaller-scale retrieval systems;

Paulo Renato de Faria
School of Electrical and Computer Engineering,Department of Computer Engineering and Industrial Automation Av. Albert Einstein, 400, Office 311, Cidade Universitária Zeferino Vaz zipcode 13083-852 Campinas-SP-Brazil
E-mail: prfaria@dca.fee.unicamp.br

Tiago Fernandes Tavares
E-mail: tavares@dca.fee.unicamp.br

These contributions can be useful for the development of specialized multimodal retrieval systems. Also, the methodological insights presented in this paper can be applied in future developments in this field.

This paper is organized as follows: section 2 describes related work in the field of movie genre classification. Section 3 presents a detailed description of our proposed method, whereas Section 4 discusses the results. Conclusions and final remarks are stated in Section 5.

2 Related Work

Movie Genre Classification is the problem of assigning categories to movies. Previous work in this field has tackled two problem variations: *multi-class*, in which a single category is assigned to each movie, and *multi-label*, in which a movie can be assigned to more than one category.

In the multi-class context, Huang et al. Huang et al. [2007] worked with shallow neural networks using only visual (Video) information. More than one information source, that is, a multi-modal approach, was first employed by Jain et al. Jain and Jadon [2009], who included audio information in the classification process. In their work, audio processing was used to yield volume, pitch, Mel Frequency Cepstral Coefficients (MFCC)s, and sub-band energies features. As a result, classification accuracy was improved from 80.2% to 87.5%.

The influence of multi-modal (audio and video) features in each genre was later explored by Huang and Wang Huang and Wang [2012]. They proposed Self-adaptive harmony search (SAHS), a technique to select the best features for each of the genres in the dataset. They found that Color Histogram, Motion Vector, MFCC and Linear Prediction Coefficients (LPC) are four critical feature types appearing independently of genres. As a result, multi-class classification accuracy was improved from 87.5% to 91.9%.

Although the multi-class problem is interesting, many movies can be simultaneously associated to different categories. This scenario leads to a multi-label movie classification problem, which was first approached by Rasheed et al. Rasheed et al. [2005]. Their approach consisted of automatic clustering movies based on low-level features, and then assigning label of dominating genres in each cluster , they found that 17 out of the 101 movies in the dataset could be interpreted as outliers in their clusters.

Later, Zhou et al. Zhou et al. [2010] developed two descriptors, namely *Centrist* and *Shot Clustering*. Centrist is a descriptor that aims at recognizing semantic categories of natural scenes and indoor environments. Shot Clustering is a technique based on K-means that allows grouping scenes into a visual word codebook (called bag of visual words (bovw)).

As it was the case for the multi-class problem, different information types started being used in multi-label problems. One of the first single-mode works of this type, by Ivasic-Kos et al. Ivasic-Kos et al. [2015], employed posters as input. Their dataset was much larger than previous works. This work studied different clustering techniques: kNN (k-nearest neighbors), and RAKEL (random k-label sets) which is an ensemble method. Naive Bayes was used to the final classification with an accuracy reported of 70%. An approach for posters using deep learning was used in Chu and Guo [2017] with maximum accuracy of 18.73% in a dataset of 8,191 poster images.

Also, text information was used by Helmer and Qinghui Helmer and Ji [2012]. This work analyzed movie genre classifiers using either subtitles (in Spanish) together with video features. They found a classification accuracy of 33.95% for text features and 40.32% for video features.

Recent work by Simoes et al. Simões et al. [2016] proposed using Convolutional Neural Networks (Convutional neural networks (CNN)s) to extract relevant video features from movie trailers. The features were yielded to a Kmeans algorithm that aimed at clustering scenes. Their proposal employs audio features, such MFCC, together with video features, with a late fusion strategy. After the feature fusion, they use a feature selection algorithm and the final classification step is performed using Support-Vector Machines.

Later, Wehrmann and Barros Wehrmann and Barros [2017] extended Simoes el al. Simões et al. [2016] multi-modal work by proposing a novel classification module called convolution through time (CTT) multi-label movie classification (MMC) two-nets (TN) (CTT-MMC-TN). They employs transfer learning for video based on ImageNet and Place365 architectures. Video is processed by a deep neural network which learns spatio-temporal relationship by stacking vectors of 2048 dimensions for each frame. Audio is processed in 3 second frames, also using a deep neural network to generate intermediate weights. Both networks are merged using a late fusion approach and a weighted average defined by researchers through empirical results.

Wehrmann and Barros Wehrmann and Barros [2017] provide a dataset called LMTD9, which comprises 4020 movies divided into 9 genres. They found that adventure, crime, romance, and thriller genres are the most difficult genres to classify.

Cascante-Bonilla et al. [2019] improved LMTD9 and created a new IMDB based list of 5.043 movie records split into 13 genres.

All these works, which are summarized in Table 1, indicate that multi-modal information can improve classification results. Also, it is important to highlight that previous work has been tested in different datasets, which is harmful to the direct comparison of results. Last, we can see that deep machine learning techniques have consistently yielded better classification accuracies than approaches based on feature engineering.

Our proposal is to perform multi-label movie genre classification using multiple information modes. These modes comprise text (movie title and plot summary), image (movie posters), audio (extracted from movie trailers) and video (also extracted from movie trailers). Using the LMTD9 dataset, we evaluated the impact of each information source to the classification accuracy in each genre. Also, we investigated both early and late fusion techniques Snoek [2005] for multi-modal classification. Our results show that multi-modal information reduces the classification accuracy variation between genres, regardless of the dataset unbalance.

3 Proposed Method

In this paper, we propose to use text and image features, in addition to the modes used by Wehrmann et al. Wehrmann and Barros [2017]. As shown in Figures 1 and 2, each information mode was processed in a different branch of the network: the text branch uses title and plot information, the image branch uses plot, and the audio and video branches use information extracted from the movie trailers.[1]

The network was trained using a one-hot encoding for genres, using Adam optimizer with a learning rate of 10^{-4}. Training used 50 epochs, dropout rate of 0.5 and batch size 32.

The hidden layers of the neural network use Rectified Linear Unit (ReLU) activations, which avoid vanishing and exploding gradients problems. Vanishing gradients may cause slower convergence of the loss function slow. Exploding gradients may lead to oscillation around loss function minima.

Dense layers with ReLU activation were initialized using He's uniform technique He et al. [2015], which is a common heuristic to avoid exploding/vanishing gradients. Instead of drawing from standard normal distribution, this technique draws initial weights W from normal distribution with variance k/n, where k depends on the activation function.

For dealing with binary classification, the last layer uses a sigmoid activation, which allows predicting more than one genre per movie. As the activation function is Sigmoid, the most common heuristic to deal with gradient issues is Glorot uniform density function Glorot and Bengio [2010].

We evaluated two strategies for feature fusion. The first is early fusion, in which features derived from each network branch are merged before any classification steps are taken. The second is late fusion, in which each branch of the network performs an independent classification step and then the classification results are used in a further classification step.

Each of the network branches is further described in the next sections.

3.1 Text features

Text information was obtained from plot and title information downloaded from IMDb.

A tokenizer transforms text phrases into individual words. Then, English stopwords are removed. The remaining words are then mapped into a vector space using Glove embeddings Pennington et al. [2014].

The embeddings are yielded to fruther classification steps. It is worth noticing that these embeddings are fine-tuned during the training stages, which makes them more specialized to the movie trailer texts on use.

3.1.1 Long-short term memory (LSTM)

LSTM cells are able to extract patterns from signals using gates that store previous inputs/outputs and an internal cell state. Figure 3 shows operation performed over signal being processed X, where t is and index related to the word sequence. Each LSTM cell memorizes a cell-state C_t and outputs H_t, these values are calculated based on previous values C_{t-1} and H_{t-1}.

In this work, we used Bidirectional LSTMs. They are capable of capturing capture patterns from sentences in forward order and backward (reverse) order. This is illustrated in Figure 4.

3.1.2 Soft Attention mechanism

As a way to extract context from sentences in words nearby, a Soft Attention mechanism Bahdanau et al. [2014] was used. This mechanism follows an implementation adapted from S. Mani et al. Mani et al. [2018]. Differently from LSTM which uses previous cell state to

[1] Source code with implemented algorithms, techniques and results are available on github at `https://github.com/paulorfbr/music_video_information_retrieval`

Table 1: Movie genre classification works

Author(s)	Year	Modes	Type	Features	Techniques	Dataset (movies)	# Genres	Accuracy
Z. Rasheed et al. Rasheed et al. [2005]	2005	Visual	Multi-label	4	Mean Shift Classification	101	4	83.00%
H.Y. Huang et al. Huang et al. [2007]	2007	Visual	Multi-class	4	2-Layer Neural Network (NN)	44	3	80.20%
S.K. Jain et al. Jain and Jadon [2009]	2009	Visual-Audio	Multi-class	21	NN	300	5	87.50%
H. Zhou, et al. Zhou et al. [2010]	2010	Visual	Multi-label	6200	CENTRIST +Shot Clustering	1239	4	74.70%
Edmund Helmer, Qinghui Ji Helmer and Ji [2012]	2012	Visual-Text	Multi-class	4	Random Forest	312 (100 subtitles)	7	33.95% (subtitles)
Y. F. Huang, S. H. Wang Huang and Wang [2012]	2012	Visual-Audio	Multi-class	277	Support Vector Machines (SVM)s + SAHS	223	7	91.90%
M. Ivasic-Kos et al. Ivasic-Kos et al. [2015]	2015	Poster	Multi-label	728	RAKEL, ML-kNN, Naive Bayes	6739 posters	18	70.00%
Simões et al. Simões et al. [2016]	2016	Visual-Audio	Multi-label	2048	CNN Kmeans SVM	1067 (LMTD4)	4	73.45%
Wehrmann, J. Barros, R. C. Wehrmann and Barros [2017]	2017	Visual-Audio	Multi-label	2048	CTT-MMC-TN	4007 (LMTD9)	9	Not reported
Cascante-Bonilla et al. Cascante-Bonilla et al. [2019]	2019	Visual-Audio-Text-Poster-Metadata	Multi-label	multiple	VGG16, Glove, fastVideo, fastText	5.043 (Moviescope)	13	Not reported

keep short and long memory, Soft Attention maintain context weights for deciding output of cell state.

Figure 4 shows relationship among input X, bidirectional LSTM and how soft attention is inserted on top of it to create a context vector that will be part of overall neural network state S, and afterwards be part of output classification Y.

Soft Attention estimation starts by calculating:

$$e_{ij} = a(s_{i-1}, h_j), \tag{1}$$

where a is called Alignment model. It calculates e as a measure on how well each encoded input h matches the current output of the decoder s. After this, a Softmax function is used to align scores:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=0}^{n}\exp(e_{ik})} \tag{2}$$

The context vector is a weighted sum of the annotations (h_j) that scale alignment scores as follows:

$$C_{ij} = \sum_{j=0}^{n} \alpha_{ij} * h_j \tag{3}$$

3.2 Poster (Image) features

To work with images, we downloaded each poster image from the IMDB URLs pointed by LTMD9 dataset.

In this work, we used the ImageNet architecture as basis. The ImageNet contains weights learned from 14 million images and labels. Image objects have a broad range of high level categories such: people, musical instruments, animal, devices, plant, food. ImageNet labels comes from WordNet Press [1995] Soergel [1998]

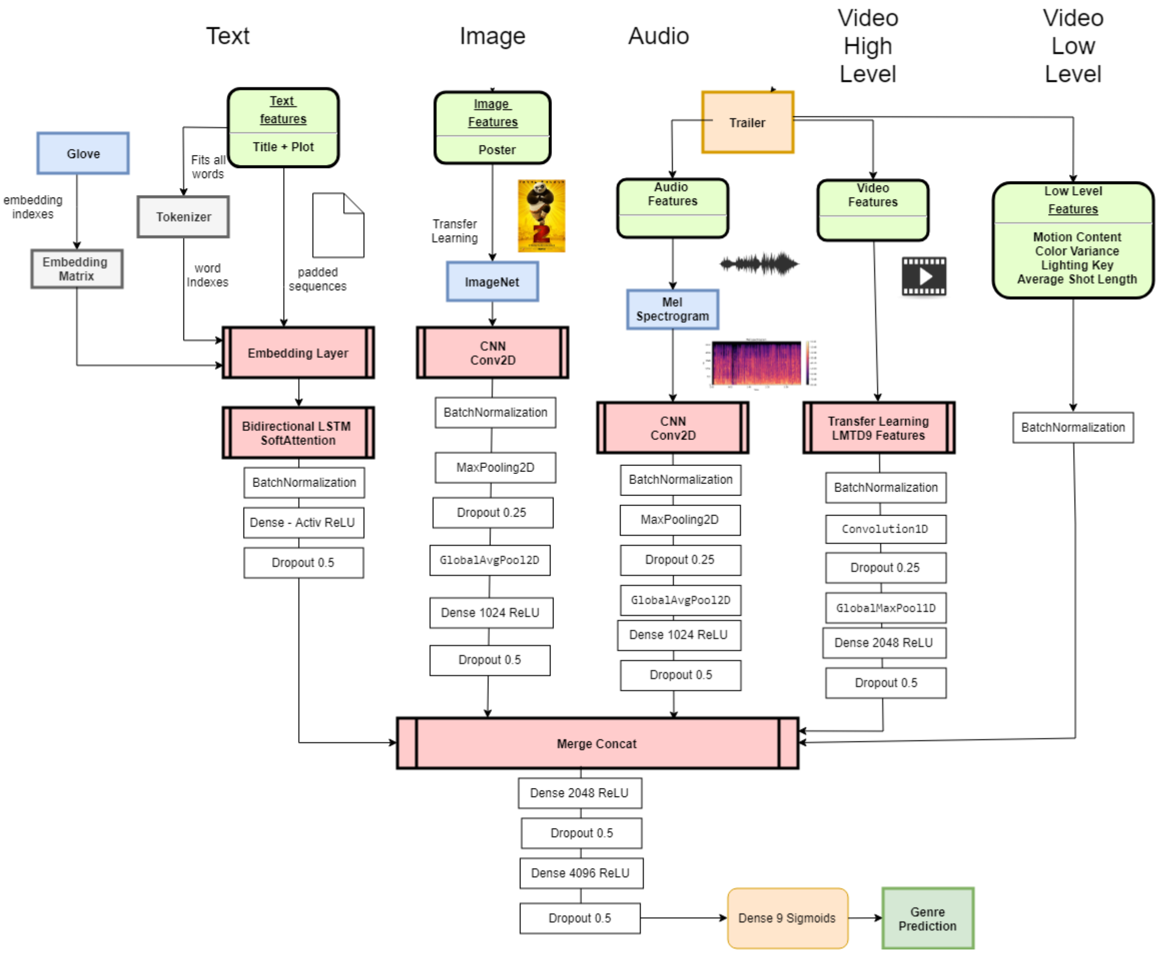

Fig. 1 TIV-MMC Proposed (Text Image Video) Movie Genre Classification Early Fusion flow

sets of synonyms (synsets). We applied Transfer Learning to ImageNet, Deng et al. [2009], that is, we retrained the neural network reusing images and labels related to movie posters.

The Deep Learning pipeline following the ImageNet feature extractor is as follows. A CNN was applied with 2 Conv2D(32), 2 Conv2D(64), both with 3 x 3 kernel, maximum and global 2D pooling. The convolution operation on posters aims at finding patterns on small figure segments. Pattern recognition occurs as different genres may have different objects recognized through ImageNet. Maximum and global pooling functions tend to extract patterns from greater objects found at poster.

3.3 Audio features

Audio information signal was extracted using librosa McFee et al. [2015]. Over deep neural network, it was inserted a melspectogram Choi et al. [2017] layer. This layer parameters were: number of Discrete Fourier Transform (DFT) 256, hops 128, number of mels 64. Melspectogram layer transforms audio signal into image. With the output image, we applied a CNN using 2 Conv2D(16), 2 Conv2D(32), both with 3 x 3 kernel, maximum and global 2D pooling.

Convolution on audio targets at identifying patterns on frequency spectrum on small periods of trailer. Maximum and global pooling functions tend to extract frequency patterns from greater duration of trailer audio for different genres.

3.4 Video features

As proposed by Rasheed et al. on Rasheed et al. [2005], scene detection is used to preprocess video frames.

3.4.1 Scene Detection

Scene detection is the first preprocessing to be accomplished on video. As images varies few from frame to frame, scene detectors serve to diminish the number of frames to process, as to find the best transition frame that are relevant to further analysis.

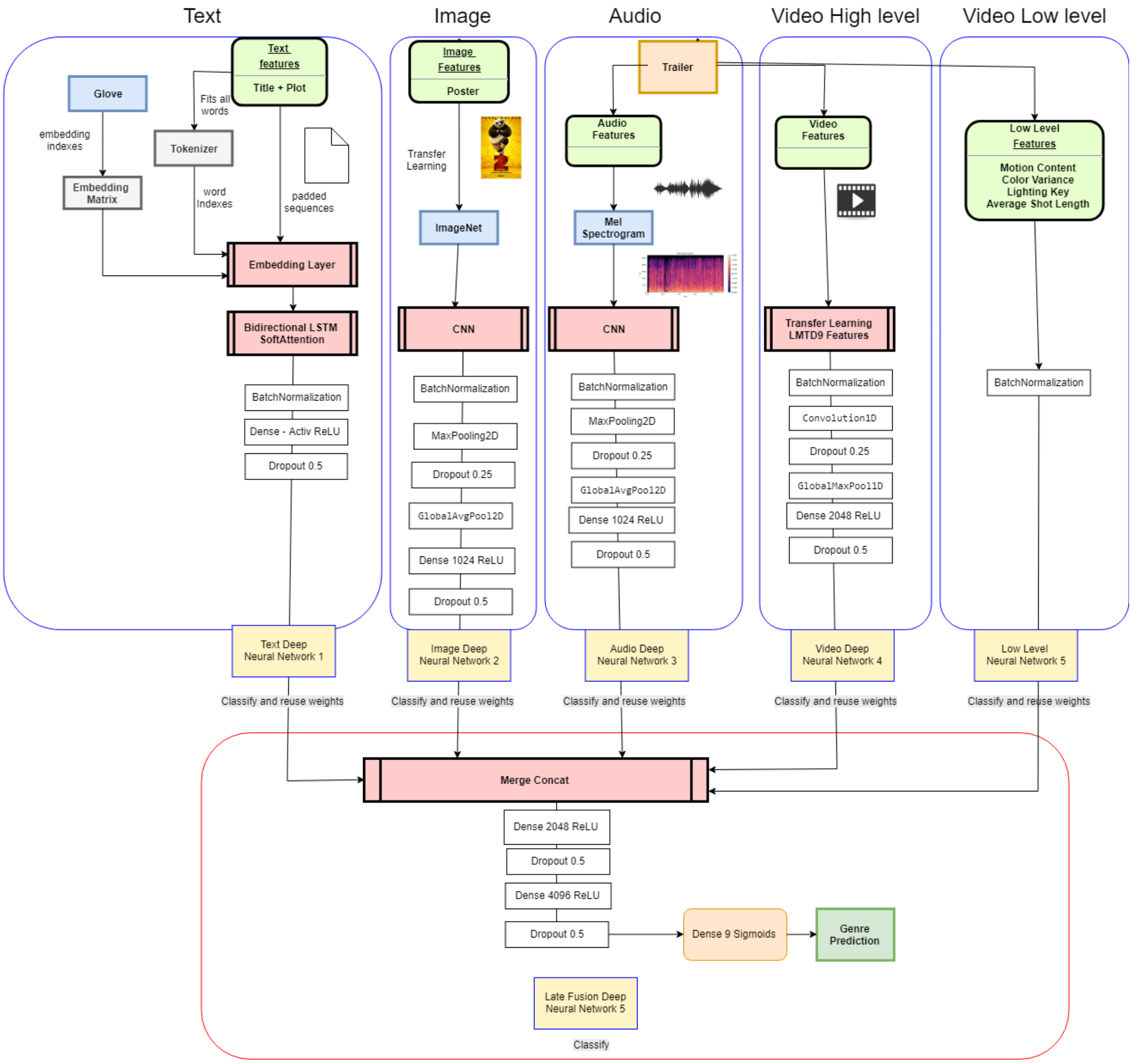

Fig. 2 TIV-MMC Proposed (Text Image Video) Movie Genre Classification Late Fusion flow

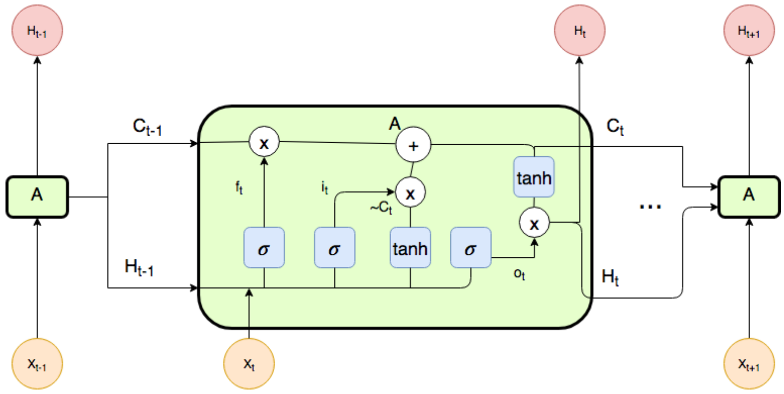

Fig. 3 LSTM cell operations

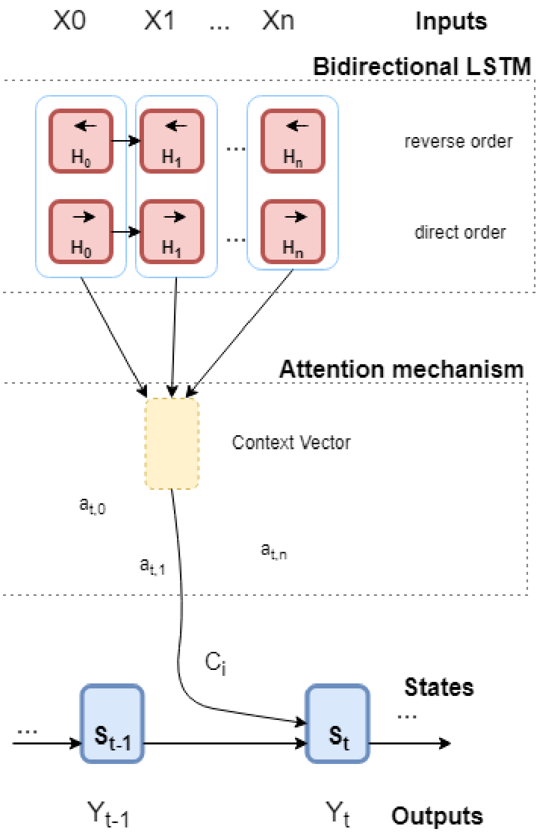

Fig. 4 Relationship among text input, bidirectional LSTM, Soft Attention mechanism, neural network state S and classification output Y

A signal smoothing can be done by applying a low-level filter (mean, Gaussian). In this work, we employed a Gaussian filter, as shown in Equation (4).

$$\exp\left(-\frac{x}{k}\right)^2 \tag{4}$$

After filtering, signal smoothing is accomplished through comparing the frame delta (gradient) with a threshold given by the Gaussian-filtered signal.

The gradients are computed as follows (5) - (9):

$$gE_i = frame_{after} - frame_{current} \tag{5}$$

$$gW_i = frame_{before} - frame_{current} \tag{6}$$

$$CE_t = \exp\left(-\frac{|gE_i|}{k}\right)^2 \tag{7}$$

$$CW_t = \exp\left(-\frac{|gW_i|}{k}\right)^2 \tag{8}$$

$$st_{next} = \lambda * (CE_t * gE_i + CW_t * gW_i) \tag{9}$$

Where k = 0.1 and λ is 0.1

3.4.2 HSV Histogram

We represented each image using a HSV Histogram. For such, each image was transformed into the HSV (Hue, Saturation and Value, i.e. Brightness/Illumination) representation. Then, a histogram was created for each of these dimensions. The Saturation and Illumination histograms have 4 bins, while the Hue one uses 8 bins.

3.4.3 Scene Detection through Histogram intersection

With the intersection of histograms it is possible to use smoothing technique from (9) to find the points where there is major variations of HSV values over the time. According to Rasheed et al. on Rasheed et al. [2005], a value of 0.7 is defined as threshold on signal values of (9). Detected frames are minimum critical points where there is a strong probability of having a scene transition in a movie/trailer.

3.4.4 Video feature through transfer learning

After detecting the scenes to remove overhead of very small changes in frames, each resulting transition image is given as input to a neural network where transfer learning is accomplished over already learnt weights from CTT-MMC-TN technique (proposed by Wehrmann et al. on Wehrmann and Barros [2017]). These extracted video features are processed along next layers to finish classification using a convolution 1D, kernel with size 3 and global max pooling.

3.5 Low level features

As stated by Rasheed et al. Rasheed et al. [2005], the following low level features are extracted from trailers:

3.5.1 Average frames per trailer

Action movies tend to have more transitions than drama or suspense ones.

From scene detection, it is possible to define a metric (10) to get average number of frames until a transition is found.

$$avgFrames = \frac{totalFrames}{totalSceneTransitionsDetected} \tag{10}$$

3.5.2 Motion content

This feature helps separating movies with action scenes from other movies.

To calculate motion content, a Sobel filter is applied to approximate derivative of two consecutive frames, as shown in equations (11) and (12).

$$H_x = \text{Sobel}(\text{frame}_x) \tag{11}$$

where subscript x, means space in image.

$$H_{xx} = H_x * H_x \tag{12}$$

Temporal derivatives (subscript t) are calculated by subtracting frames pixel by pixel, that is:

$$H_t = \text{frame}_{\text{current}} - \text{frame}_{\text{before}} \tag{13}$$

$$H_{tt} = H_t * H_t \tag{14}$$

We can calculate the partial derivative:

$$H_{xt} = H_x * H_t \tag{15}$$

Finally, a global measure for movement of pixels is calculated as a sum of the differences in each frame, where T is the total time and X are the space:

$$J_{xx} = \sum_{x=1}^{X} \sum_{x=1}^{X} H_{xx} \tag{16}$$

$$J_{xt} = \sum_{x=1}^{X} \sum_{t=1}^{T} H_{xt} \tag{17}$$

$$J_{tt} = \sum_{t=1}^{T} \sum_{t=1}^{T} H_{tt} \tag{18}$$

The ratio of movement r is calculated by:

$$r = \frac{2 * J_{xt}}{(J_{xx} - J_{xt})} \tag{19}$$

The rotation angle theta is estimated using ration movement r:

$$theta = \frac{1}{2} * \arctan(r) \tag{20}$$

Following Rasheed et al. Rasheed et al. [2005], pixels in which there is more than 10 degrees of rotation among consecutive frames are considered “active”. This allows to define *motion content* as:

$$motionContent = \frac{\sum_{a \in \text{activePixels}} a}{\sum_{p \in pixels} p} \tag{21}$$

3.5.3 Color Variance

Another low level feature is Color Variance. It is calculated converting keyframes - central frame for a scene - into CIE Luv space. This feature helps as a feature separator because horror often portrays lower variance than comedies.

$$covLUV = cov(L, U, V) \tag{22}$$

Color variance is found through matrix determinant (23).

$$colorVariance = \det(covLUV) \tag{23}$$

3.5.4 Lighting Key

Lighting Key is extracted by HSV color space calculating mean (μ) and standard deviation (σ) of the pixel values (24). Low-key lighting indicates dark or more dramatic scenes, probably indicating horror, while comedies and action movies show high-key lighting, i.e. brighter or less dramatic scenes.

$$lightingKey = \mu_{pixels \in HSV} * \sigma_{pixels \in HSV} \tag{24}$$

3.6 Evaluation measures

The outputs of TIV-MMC for each class are probability values, and the same is true for the baseline algorithms. Following Wehrmann Wehrmann and Barros [2017], precision-recall curves (PR-curves) were employed as the evaluation criterion for comparing the different approaches. Using this approach, 3 derived measures are obtained: weighted average, macro average, micro average. Each of these measures point to different aspects regarding each method's performance.

Precision definition is (25):

$$Pr = \frac{TP}{(TP + FP)} \tag{25}$$

where TP are true-positive, FP are false-positive.

Recall is given by (26):

$$Re = \frac{TP}{(TP + FN)} \tag{26}$$

where FN are false-negative.

3.7 PR Curve

To obtain Precision-Recall curve, it is used several thresholds in interval [0,1] to apply the classifiers under evaluation. After that, it can get a Receiver-Operation curve (ROC) with y-axis as precision and x-axis as Recall. The area under this curve is called Area Under curve (AU).

3.7.1 AUPR Macro

For each class c, it is calculated a ROC and the areas for each class. Macro area under curve precision-recall (AUPR) is given by the average of these areas (27).

$$AUPR_{macro} = \frac{\sum AUPR_{c \in classes}}{\sum_{c \in classes}} \tag{27}$$

For instance, by averaging the areas of all classes, macro measure causes less frequent classes to have more influence in the results.

3.8 AUPR Micro

It calculated one ROC curve including all classes together, as equations below (28) - (30):

$$P_{micro} = \frac{\sum_c TP_c}{\sum_c TP_c + \sum_c FP_c} \tag{28}$$

$$R_{micro} = \frac{\sum_c TP_c}{\sum_c TP_c + \sum_c FN_c} \tag{29}$$

$$AUPR_{micro} = areaCurve(P_{micro}, R_{micro}) \tag{30}$$

By using all labels globally, micro measure provides information regarding the entire dataset, making high-frequency classes to have greater influence in the results.

3.9 AUPR weighted

Weighted measure was calculated by averaging the area under PR curve per genre, weighting instances according to the class frequencies (31).

$$AUPR_{weighted} = \frac{(\sum w_{c \in classes} * AUPR_{c \in classes})}{\sum w_{c \in classes}} \tag{31}$$

4 Results and discussion

Table 3 shows results for multi-label classification and compares them to our results[2]. It can be noticed that the proposed method outperforms others techniques reported in literature regarding accuracy. In the macro measure, results were very close to CTT-MMC-TN. This fact is explained by the better results on low-frequency items such as adventure, crime, romance, SciFi. Conversely, the micro and weighted measures are better on CTT-MMC-TN due to the worst results on high-frequency items.

From a genre perspective, presented in Table 4 and summarized in Figure 7, TIV-MMC outperformed CTT-MMC-TN in adventure, crime, romance, and scifi.

Figure 5 presents a detailed view on individual multimedia features (text, poster, video or audio) and the influence on each genre. Action, Adventure, Comedy, and Scifi have shown to be more sensible to video than other modes. Posters have shown a greater predicting power in Crime, Horror and Thriller genres.Text features have a larger impact in Drama, Romance and Horror.Audio does not have a great impact in any of the studied genres.

It is important to notice that interclass variance was reduced with the introduction of text features, as shown

[2] Source code with implemented algorithms, techniques and results are available on github at `https://github.com/paulorfbr/music_video_information_retrieval`

in Figure 6. As a consequence, TIV methods tend to be a more balanced prediction among classes even for low frequency items on test set.

Late fusion interclass variance is 44.71% lower than CTT-MMC-TN. Compared to Early fusion, Late fusion technique presented better results in all genres, except Adventure and Romance genres.

Regardless of early or late fusion approaches, text features plays an important role. Results are comparable or surpasses video results for some genres such as crime, drama, horror, romance and thriller. By far, the greatest advantage is shown on table 2 where it can be noticed that Text features are stored in orders of megabytes whilst Video features are in gigabytes of magnitude. Depending on storage restrictions to classify or train the model, Text features can achieve an accurate and balanced result with much less storage space.

From the difficult genres for audio-video the unique item that has not been improved was Thriller. TIV-MMC also lost in comparison to CTT-MMC-TN in action, comedy, drama, and horror genres.

5 Conclusion

In this paper, the task of automatic classification of movie previews was tackled. We explored text and image features (plot, title, posters) to apply on multi-label movie genre classification problem. This work proposed TIV-MMC method as a deep neural network flow to deal with difficult genres on current state-of-art for video-audio features: adventure, crime, romance and thriller. As hypothesized in literature review, text and image (marketing poster) features helped machine to improve accuracy, precision, and recall results specially for Adventure, Crime and Romance. The unique non-explained item was Thriller, which, despite as individual feature showed a greater result than video feature, in the end of classification it did not improve results from CTT-MMC-TN, so it could be further investigated in future works. An improvement on LMTD9 genres, for example Western, Animation, Documentary, History, War and Western can also be considered to further explore movie genre diversity. The main contribution, although it does not surpasses CTT-MMC-TN in some genres is that text and image features can now be considered as an important aid to classify difficult genres automatically.

6 Acknowledgments

The authors would like to show their gratitude to Rodrigo Barros, Associate Professor and Jônatas Wehrmann, PhD, both from Pontifícia Universidade Católica do Rio Grande do Sul - PUCRS for sharing movie dataset LTMD9 to use as baseline and deep learning tasks.

We also thank Carlos A. S. Junior, MS for comments that greatly improved the manuscript.

References

Z. Rasheed, Y. Sheikh, and M. Shah. On the use of computable features for film classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(1):52–64, January 2005. ISSN 1051-8215. doi: 10.1109/TCSVT.2004.839993. URL `http://ieeexplore.ieee.org/document/1377360/`.

Jônatas Wehrmann and Rodrigo C. Barros. Movie genre classification: A multi-label approach based on convolutions through time. *Applied Soft Computing*, 61:973–982, December 2017. ISSN 1568-4946. doi: 10.1016/j.asoc.2017.08.029. URL `http://www.sciencedirect.com/science/article/pii/S1568494617305112`.

Wei-Ta Chu and Hung-Jui Guo. Movie genre classification based on poster images with deep neural networks. In *Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes*, MUSA2 '17, pages 39–45, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5509-4. doi: 10.1145/3132515.3132516. URL `http://doi.acm.org/10.1145/3132515.3132516`.

Daichi Minami, Mio Ushijima, and Taketoshi Ushiama. How do viewers react to drama?: Extraction of scene features of dramas from live commentary tweets. In *Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication*, IMCOM '18, pages 87:1–87:4, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-6385-3. doi: 10.1145/3164541.3164616. URL `http://doi.acm.org/10.1145/3164541.3164616`.

Federico Alvarez, Faustino Sánchez, Gustavo Hernández-Peñaloza, David Jiménez, José Manuel Menéndez, and Guillermo Cisneros. On the influence of low-level visual features in film classification. *PLoS ONE*, 14(2), February 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0211406. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6386315/`.

Kazimierz Choroś. Video Genre Classification Based on Length Analysis of Temporally Aggregated Video Shots. In Ngoc Thanh Nguyen, Elias Pimenidis, Zaheer Khan, and Bogdan Trawiński, editors, *Computa-*

Table 2: Storage size of extracted features

Feature	Train size	Validation size	Test Size	Sum (kB)	Sum Size (%)
Low level	92 kB	12 kB	25 kB	129	0.00052%
Text	1.1 MB	150 kB	310 kB	1586	0.0065%
Audio	2.0 GB	264 MB	545 MB	2925568	11.90%
Poster	3.5 GB	450 MB	930 MB	5083136	20.68%
Video	11.3 GB	1.5 GB	3 GB	16567500.8	67.41%

Table 3: Area Under Precision-Recall (AUPR) curve for Multi-label movie genre classification on Test Set

Author(s)	Year	Machine Learning	Accuracy	AU PR Macro	AU PR Micro	AU PR Weighted
Z. Rasheed et al. Rasheed et al. [2005]	2005	Mean Shift Classification	83.00%	Not reported	Not reported	Not reported
H. Zhou, et al. Zhou et al. [2010]	2010	CENTRIST + Shot Clustering	74.70%	Not reported	Not reported	Not reported
M. Ivasic-Kos et al. Ivasic-Kos et al. [2015]	2015	RAKEL, ML-kNN, Naive Bayes	70.00%	Not reported	Not reported	Not reported
Simões et al. Simões et al. [2016]	2016	CNN + Kmeans + SVM	73.45%	Not reported	Not reported	Not reported
Wehrmann, J. Barros, Rodrigo C. Wehrmann and Barros [2017]	2017	CTT-MMC-TN	Not reported	0.646	**0.742**	**0.724**
TIV-MMC Early Fusion (Ours)	2018	CNN, LSTM/SoftAttention Text, Image, Video features	84.31%	0.6488	0.7065	0.7009
TIV-MMC Late Fusion (Ours)	2018	CNN, LSTM/SoftAttention Text, Image, Video features	**85.75%**	**0.6538**	0.7211	0.7070

tional Collective Intelligence, Lecture Notes in Computer Science, pages 509–518. Springer International Publishing, 2018. ISBN 978-3-319-98446-9.

H. Y. Huang, W. S. Shih, and W. H. Hsu. Movie classification using visual effect features. In *2007 IEEE Workshop on Signal Processing Systems*, pages 295–300, Oct 2007. doi: 10.1109/SIPS.2007.4387561.

S. K. Jain and R. S. Jadon. Movies genres classifier using neural network. In *2009 24th International Symposium on Computer and Information Sciences*, pages 575–580, September 2009. doi: 10.1109/ISCIS.2009.5291884.

Yin-Fu Huang and Shih-Hao Wang. Movie Genre Classification Using SVM with Audio and Video Features. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Runhe Huang, Ali A. Ghorbani, Gabriella Pasi, Takahira Yamaguchi, Neil Y. Yen, and Beijing Jin, editors, *Active Media Technology*, volume 7669, pages 1–10. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-35235-5 978-3-642-35236-2. doi: 10.1007/978-3-642-35236-2_1. URL `http://link.springer.com/10.1007/978-3-642-35236-2_1`.

Howard Zhou, Tucker Hermans, Asmita V. Karandikar, and James M. Rehg. Movie genre classification via scene categorization. page 747. ACM Press, 2010. ISBN 978-1-60558-933-6. doi: 10.1145/1873951.1874068. URL `http://dl.acm.org/citation.cfm?doid=1873951.1874068`.

Marina Ivasic-Kos, Miran Pobar, and Ivo Ipsic. Automatic Movie Posters Classification into Genres. In Ana Madevska Bogdanova and Dejan Gjorgjevikj, editors, *ICT Innovations 2014*, volume 311, pages 319–328. Springer International

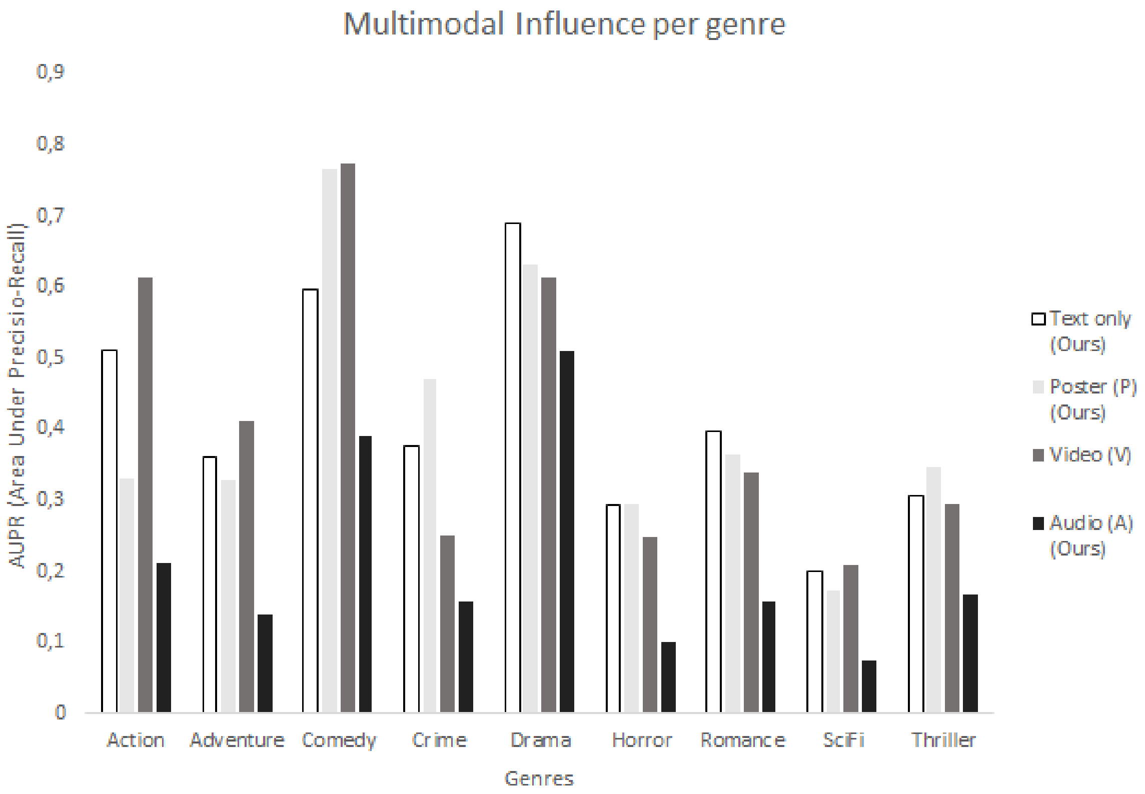

Fig. 5 TIV-MMC results for different information modes.

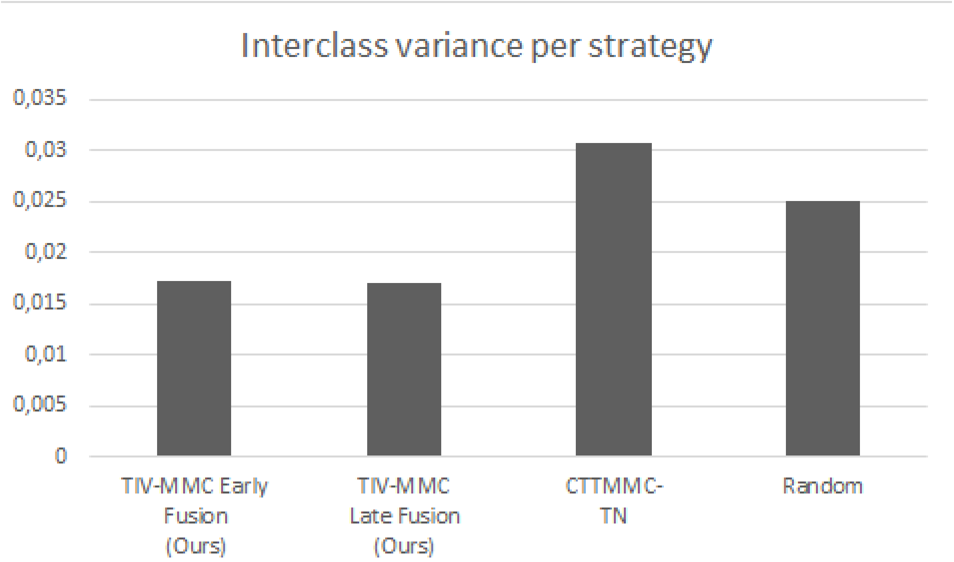

Fig. 6 Interclass variance comparison per strategy

Publishing, Cham, 2015. ISBN 978-3-319-09878-4 978-3-319-09879-1. doi: 10.1007/978-3-319-09879-1_32. URL `http://link.springer.com/10.1007/978-3-319-09879-1_32`.

Edmund Helmer and Qinghui Ji. Film Classification by Trailer Features. page 5, 2012.

G. S. Simões, J. Wehrmann, R. C. Barros, and D. D. Ruiz. Movie genre classification with Convolutional Neural Networks. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 259–266, July 2016. doi: 10.1109/IJCNN.2016.7727207.

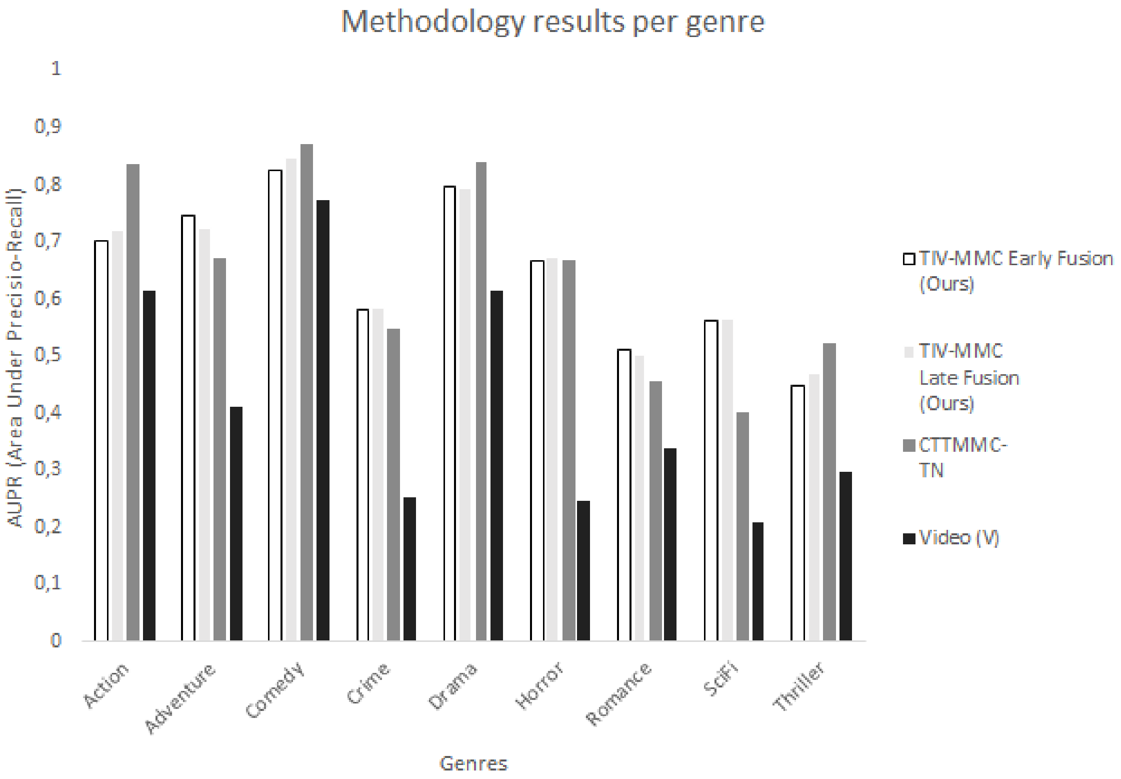

Fig. 7 Per-genre results for different classification methods.

Table 4: Per-genre AUPR for Multi-label movie genre classification on Test Set

Genre	Text only (Ours)	Poster (P) (Ours)	Video (V) (Ours)	TIV-MMC Early Fusion (Ours)	TIV-MMC Late Fusion (Ours)	CTT-MMC-TN Wehrmann and Barros [2017]	Random	Test set (%)
Action	0.5120	0.3312	0.6141	0.7006	0.722	**0.835**	0.158	21.24%
Adventure	0.3600	0.3296	0.4110	**0.7461**	0.725	0.672	0.131	13.99%
Comedy	0.5979	0.7663	0.7744	0.8268	0.8467	**0.87**	0.512	38.99%
Crime	0.3753	0.471	0.2515	0.5796	**0.5853**	0.547	0.140	15.67%
Drama	0.6889	0.6310	0.6138	0.7973	0.7944	**0.841**	0.435	51.03%
Horror	0.2925	0.2949	0.2480	0.6655	**0.6719**	0.667	0.088	10.10%
Romance	0.3982	0.3651	0.3384	**0.5126**	0.5022	0.456	0.129	15.80%
SciFi	0.2010	0.1721	0.2092	0.5612	**0.5668**	0.401	0.063	07.38%
Thriller	0.3075	0.3468	0.2957	0.4491	0.4696	**0.522**	0.196	16.71%
Variance	**0.01602**	0.03303	0.03746	**0.01713**	**0.01702**	0.03079	0.02512	N/A

Paola Cascante-Bonilla, Kalpathy Sitaraman, Mengjia Luo, and Vicente Ordonez. Moviescope: Large-scale analysis of movies using multiple modalities. *ArXiv*, abs/1908.03180, 2019.

Cees G. M. Snoek. Early versus late fusion in semantic video analysis. In *In ACM Multimedia*, pages 399–402, 2005.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. pages 1026–1034. IEEE, December 2015. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.123. URL `http://ieeexplore.ieee.org/document/7410480/`.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and*

Statistics, pages 249–256, March 2010. URL http://proceedings.mlr.press/v9/glorot10a.html.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. pages 1532–1543. Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1162. URL http://aclweb.org/anthology/D14-1162.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473 [cs, stat]*, September 2014. URL http://arxiv.org/abs/1409.0473. arXiv: 1409.0473.

Senthil Mani, Anush Sankaran, and Rahul Aralikatte. DeepTriage: Exploring the Effectiveness of Deep Learning for Bug Triaging. *arXiv:1801.01275 [cs]*, January 2018. URL http://arxiv.org/abs/1801.01275. arXiv: 1801.01275.

The MIT Press. WordNet, 1995. URL http://mitpress.mit.edu/books/wordnet.

Dagobert Soergel. WordNet. An Electronic Lexical Database. October 1998.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. ImageNet: a Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P W Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and Music Signal Analysis in Python. page 8, 2015.

Keunwoo Choi, Deokjin Joo, and Juho Kim. Kapre: On-GPU Audio Preprocessing Layers for a Quick Implementation of Deep Neural Network Models with Keras. *arXiv:1706.05781 [cs]*, June 2017. URL http://arxiv.org/abs/1706.05781. arXiv: 1706.05781.