Movie Genre Classification Using Text, Image and Video features

024826

***Abstract*—Movie genre classification challenges Computer Vision (CV) field as classes to assign can not be pinpointed in direct ways within any region of movie frames. Movies may belong to multiple genres - called multi-label problem - making works in this area more defying. This work presents a deep learning architecture called Text Image Video multi-label movie classification (TIV-MMC), which comprises usual audio-visual features and newly incorporated features for text and image. Text processing – title and plot – employs bidirectional Long-short term memory (LSTM) and soft attention mechanisms, while image – i.e marketing poster – uses Convutional neural network (CNN). Audio is transformed into melspectrogram and then treated as an image within a CNN. Video features are captured by transfer learning over LMTD9 using state-of-art convolution through time (CTT) multi-label movie classification (MMC) two-nets (TN) (CTT-MMC-TN) technique. Results showed that our method outperforms state-of-art for area under curve precision-recall (AUPR) in some difficult genres to classify using only video-audio: adventure 11%, crime 5%, romance 12%. As a conclusion, despite the challenge of no direct information from audio-video frames available for machines, movie genre classification can be improved by using text and image poster features. Similar to humans who can infer genre classes from other media sources besides trailers, this research paves new ways for future movie academic/commercial classifiers.**

***Index Terms*—Movie genre classification, Convolutional neural networks, Multi-label classification, Natural Language Processing, Deep learning paradigm.**

June 17, 2018

I. Introduction

VIDEO content analysis has potential of helping human beings to solve time-consuming and expensive problems such as automatic movie genre classification. This classification is challenging as a Computer Vision (CV) task because classes to predict are not presented in obvious ways within any region of the movie frames.

In this work, it was investigated the use of text, image, audio and video features using a Deep Convolutional Neural Network (ConvNet) architecture to solve multi-label – i.e, each movie may be labeled as belonging to more than one genre at the same time – classification of 9 movie genres based on trailers. To help identifying genres, besides the most common used features: trailer – video features – and audio – converted to image as melspectrogram –, we propose using plot and title – text features – and poster – image features – as a way to improve classification metrics: accuracy, precision and recall, respectively.

This paper is organized as follows: section II describes related work in the field of movie genre classification. Section III presents a detailed description of our proposed method, whereas Sections IV and V show the experimental setting and obtained results. Paper ends with conclusion Section at VI.

II. Related Work

According to the literature review, there are two major areas on movie classification. In initial field works, researchers dealt with multi-class problems, i.e. there are several possible classes but just one is selected for each movie. More recent works deal with multi-label, i.e. assigning more than one class to a movie each time.

Table I summarizes works on multi-class classification including feature types, number of genres used and dataset size.

The unique work related with text features was from Helmer, E. and Qinghui, J. [1] applying Random Forest using subtitles as text features, but accuracy was very low 33.95%.

Table II presents works on multi-label classification papers. M. Ivasic-Kos et al. [2] tried to use posters to classify movies with accuracy of 70%, however it did not use video features. Simoes et al. [3] created LTMD4 dataset and proposed CNN on audio-video features to classify genres and used Neural Network (NN) representation as an input for Kmeans to cluster scenes and Support Vector Machines (SVM) to make the final classification.

Wehrmann, J. and Barros, Rodrigo [4] created an improved version called large movie trailer dataset (LMTD9) comprising 9 genres (training set 2873, validation 374, and test 773 movies) and proposed a novel method called CTT-MMC-TN that is the current state-of-art using audio-video features in a large movie dataset. From this last paper, the most difficult genres to classify were adventure, crime, romance, and thriller.

As humans have the ability to classify genres based not only on trailer, but also reading text extracts about the movie, such as title and plot, and also looking into marketing movie posters, this led us to hypothesize that: could we improve movie genre classification using features such as image and text features besides commonly used audio-video features?

III. Proposed Method

In this paper, the use of new text and image (poster) features is proposed to extend work of Wehrmann et al. [4] to improve classification for difficult genres (adventure, crime, romance, and thriller) called TIV-MMC. Figure 1 details the overall process showing how each feature was extracted and how all features were merged to enable a multi-label genre prediction using sigmoid function in the end of classification.

024826 is with the Department of Electrical and Computer Engineering, State University of Campinas, SP, BR e-mail: p024826@unicamp.br.

TABLE I: Multi-class movie genre classification works

Author(s)	Year	Feature Types	Feature Selection	Feature Dim.	Machine Learning	Dataset (movies)	Nbr. Genres
H.Y. Huang et al. [5]	2007	Visual	N	4	2-Layer NN	44	3
S.K. Jain et al. [6]	2009	Visual-Audio	N	21	NN	300	5
Y. F. Huang, S. H. Wang [7]	2012	Visual-Audio	Y (SAHS)	277	SVMs	223	7

TABLE II: Multi-label movie genre classification works

Author(s)	Year	Feature Types	Feature Selection	Feature Dim.	Machine Learning	Dataset (movies)	Nbr. Genres
Z. Rasheed et al. [8]	2003	Visual	N	4	Mean Shift Classification	101	4
H. Zhou, et al. [9]	2010	Visual	N	6200	CENTRIST +Shot Clustering	1239	4
M. Ivasic-Kos et al. [2]	2015	Poster	N	728	RAKEL, ML-kNN, Naive Bayes	6739 posters	18
Simões et al. [3]	2016	Visual-Audio	Y (SAHS)	2048	CNN Kmeans SVM	1067 (LMTD4)	4
Wehrmann, J. Barros, R. C. [4]	2017	Visual-Audio	N	2048	CTT-MMC-TN	4007 (LMTD9)	9

All activations used in the neural network were Rectified Linear Unit (ReLU) except 9 sigmoids, one for each genre, in the last dense layer (this last layer used Glorot uniform technique [10] as initializer). For Gradient descent, Adam optimizer was used with learning rate of $1*10-4$. The number of epochs was 50, dropout of 0.5 and batch size was 32.

A. Text features

Following Natural Language Processing (NLP) guidelines, basic preprocessing was applied as removing stopwords from title and plot. We tried to use Lancaster stemming as well, but best results were found using no stemming on texts. Wordnet [11] [12] was employed to fit each title and plot and extract an embedding of 100 items (using maximum number of words in dictionary as 50,000). GloVe [13] word-word co-occurrence was employed for measuring distance among words. Then, each sequence was padded to be 100 size to be input on a stacked 2-bidirectional LSTM with 128 size in hidden layer. After that, a SoftAttention mechanism [14] was employed, i.e. a way to select words near to extract context, following an implementation adapted from [15].

B. Poster (Image) features

We downloaded each poster image from IMDB URLs pointed by LTMD9 database. Transfer learning was applied over ImageNet [16] and then applied a 2 Conv2D(32), 2 Conv2D(64), both with 3 x 3 kernel, maximum and global 2D pooling. Kernel was initialized using He uniform technique [17].

C. Audio features

Audio was extracted using librosa [18] and Kapre [19] layer was applied on Keras [20] to transform audio into melspectrogram image. Parameters used were: number of DFT 256, hops 128, number of mels 64. Then we applied a CNN using 2 Conv2D(16), 2 Conv2D(32), both with 3 x 3 kernel, maximum and global 2D pooling.

D. Video features

As proposed by Wehrmann on [4], we employed transfer learning over features from LTMD9 to extract video features based on CTT-MMC-TN technique. The next layers, were convolution 1D, kernel 3 and global max pooling.

E. Low level features

As proposed by Rasheed et al. on [8], features were included to capture motion content that help separating movies with action scenes from other movies, Color Variance is calculated converting keyframes - central frame for a scene - into CIE Luv space, it helps as a feature separator because horror often portrays lower variance than comedies. Lighting Key is extracted by HSV color space calculating mean (μ) and standard deviation (σ) of the pixel values, 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.

IV. Experimental Setup

All experiments were accomplished on core i7 3.40 GHz, video memory 4 Gb (GTX1050TI), 16 Gb RAM. IPython notebook was prepared to run Keras and Tensorflow.

A. Evaluation measures

The outputs of TIV-MMC for each class are probability values, and the same is true for the baseline algorithms. Following Wehrmann [4], precision-recall curves (PR-curves) were employed as the evaluation criterion for comparing the

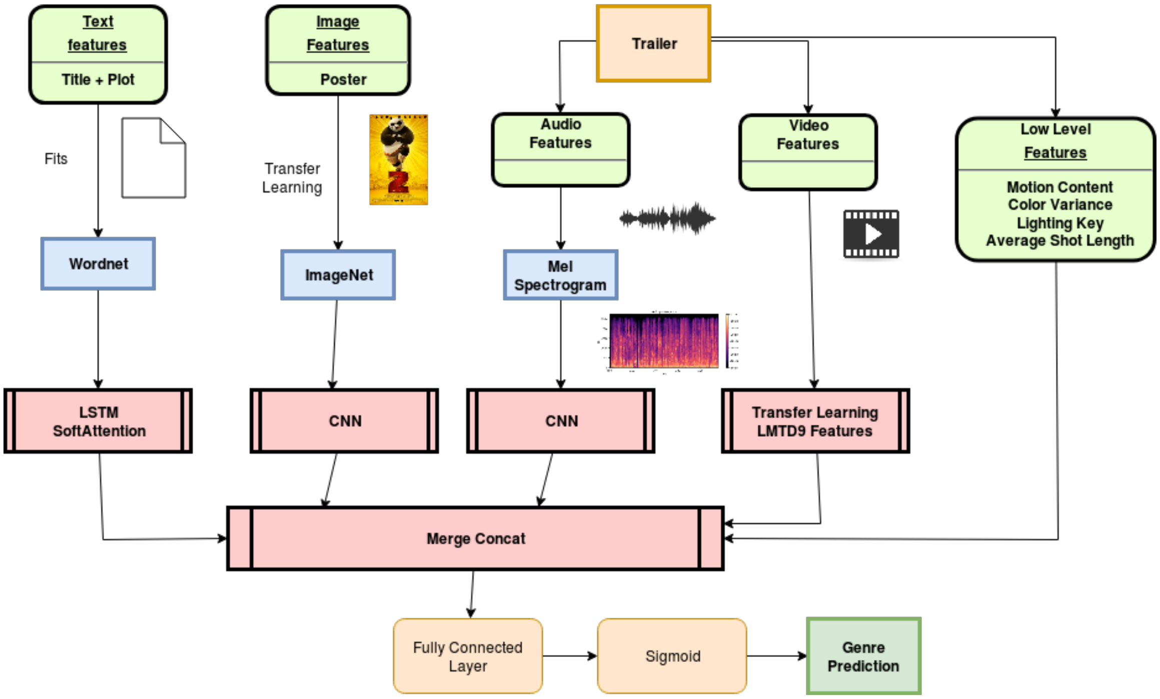

Fig. 1. TIV-MMC Movie Genre Classification flow

different approaches. Using this approach, we have 3 derived measures: weighted average, macro average, micro average. Each of these measures point to different aspects regarding each method's performance. For instance, by averaging the areas of all classes, macro measure is calculated, which causes less frequent classes to have more influence in the results. By using all labels globally, micro measure was obtained, providing information regarding the entire dataset, making high-frequency classes to have greater influence in the results. Weighted measure was calculated by averaging the area under PR curve per genre, weighting instances according to the class frequencies.

V. Results and discussion

Table III presents results for multi-class movie genre classification. Huang and Wang [7] have the best results in accuracy of 91.9% specially by applying a feature selection technique called Self-adaptive harmony search (SAHS) with applies combinations of SVMs.

Table IV shows results for multi-label classification and compares it to our results. It can be noticed that the proposed method outperforms others techniques reported in literature regarding accuracy. For 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. Micro and weighted are still better on CTT-MMC-TN due to the worst results on high-frequency items.

From a genre perspective, presented on table V TIV-MMC outperformed CTT-MMC-TN in adventure 11%, crime 5%, romance 12% and scifi 39% genres for AUPR metrics. Scifi seems the one with the greatest influence on specific text or poster features. From individual feature perspective text is advantageous on crime and romance, and fairly comparable for Scifi and Thriller. Poster features compares with video on comedy, winning on crime, drama, horror, romance and thriller. Both features combined explain the better results on adventure, crime, romance and Scifi.

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.

VI. Conclusion

In this paper, the challenging task of automatic classification of video content was tackled. We explored new text – plot and title – and image – poster – features 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.

TABLE III: Results Multi-class movie genre classification

Author(s)	Year	Machine Learning	Accuracy
H.Y. Huang et al. [5]	2007	2-Layer Neural Network	80.20%
S.K. Jain et al. [6]	2009	Neural Network	87.50%
Edmund Helmer, Qinghui Ji [1]	2012	Random Forest	33.95%
Y. F. Huang, S. H. Wang [7]	2012	SVMs	91.90%

TABLE IV: Results Multi-label movie genre classification

Author(s)	Year	Machine Learning	Accuracy	AU PR Macro	AU PR Micro	AU PR Weighted
Z. Rasheed et al. [8]	2003	Mean Shift Classification	83.00%	Not reported	Not reported	Not reported
H. Zhou, et al. [9]	2010	CENTRIST + Shot Clustering	74.70%	Not reported	Not reported	Not reported
M. Ivasic-Kos et al. [2]	2015	RAKEL, ML-kNN, Naive Bayes	70.00%	Not reported	Not reported	Not reported
Simões et al. [3]	2016	CNN + Kmeans + SVM	73.45%	Not reported	Not reported	Not reported
Wehrmann, J. Barros, Rodrigo C. [4]	2017	CTT-MMC-TN	Not reported	0.646	**0.742**	**0.724**
Ours	2018	CNN, LSTM/SoftAttention Text, Image, Video features	**84.31**%	**0.6488**	0.7065	0.7009

TABLE V: Results per genre AUPR for Multi-label movie genre classification

Genre	Text only (Ours)	Poster (P) (Ours)	Video (V)	TIV-MMC (Ours)	CTT-MMC-TN	Random	Test set (%)
Action	0.5120	0.3312	0.6141	0.7006	**0.835**	0.158	21.24%
Adventure	0.3600	0.3296	0.4110	**0.7461**	0.672	0.131	13.99%
Comedy	0.5979	0.7663	0.7744	0.8268	**0.87**	0.512	38.99%
Crime	0.3753	0.471	0.2515	**0.5796**	0.547	0.140	15.67%
Drama	0.6889	0.6310	0.6138	0.7973	**0.841**	0.435	51.03%
Horror	0.2925	0.2949	0.2480	0.6655	**0.667**	0.088	10.10%
Romance	0.3982	0.3651	0.3384	**0.5126**	0.456	0.129	15.80%
SciFi	0.2010	0.1721	0.2092	**0.5612**	0.401	0.063	07.38%
Thriller	0.3075	0.3468	0.2957	0.4491	**0.522**	0.196	16.71%

REFERENCES

[1] Edmund Helmer, Qinghui Ji,*Film Classification by Trailer Features*, 2012

[2] M. Ivasic-Kos et al.,*Automatic Movie Posters Classification into Genres*, 2015

[3] Simões et al.,*Movie Genre Classification with Convolutional Neural Networks*, 2016

[4] Wehrmann, J. Barros, Rodrigo C.,*Movie genre classification : A multi-label approach based on convolutions through time*, 2017

[5] H.Y. Huang et al., *Movie Classification using visual effects features*, 2007

[6] S.K. Jain et al., *Movies Genres Classifier using Neural Network*, 2009

[7] Y. F. Huang, S. H. Wang,*Movie Genre Classification Using SVM with Audio and Video features*, 2012

[8] Z. Rasheed et al.,*On Use of Computable Features for Film Classification*, 2003.

[9] H. Zhou, et al.,*Movie Genre Classification via Scene Categorization*, 2010.

[10] Xavier Glorot and Yoshua Bengio, *Understanding the difficulty of training deep feedforward neural networks*, AISTATS, http://jmlr.org/proceedings/papers/v9/glorot10a/glorot10a.pdf, 2010

[11] George A. Miller, *WordNet: A Lexical Database for English.*, Communications of the ACM Vol. 38, No. 11: 39-41, 1995.

[12] Christiane Fellbaum, *WordNet: An Electronic Lexical Database.*, Cambridge, MA: MIT Press.,1998.

[13] Jeffrey Pennington and Richard Socher and Christopher D. Manning, *GloVe: Global Vectors for Word Representation*, Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, http://www.aclweb.org/anthology/D14-1162, 2014

[14] Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio,*Neural Machine Translation by Jointly Learning to Align and Translate*, 2014

[15] Mani, S., Sankaran, A. and Aralikatte, R., *DeepTriage: Exploring the Effectiveness of Deep Learning for Bug Triaging*, 2018

[16] Deng, J. and Dong, W. and Socher, R. and Li, L.-J. and Li, K. and Fei-Fei, L., *ImageNet: A Large-Scale Hierarchical Image Database*,2009

[17] Kaiming He and Xiangyu Zhang and Shaoqing Ren and Jian Sun, *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*, http://arxiv.org/abs/1502.01852, 2015

[18] McFee, Brian, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. *”librosa: Audio and music signal analysis in python.”*, In Proceedings of the 14th python in science conference, pp. 18-25. 2015.

[19] Choi, Keunwoo and Joo, Deokjin and Kim, Juho, *Kapre: On-GPU Audio Preprocessing Layers for a Quick Implementation of Deep Neural Network Models with Keras*, Machine Learning for Music Discovery Workshop at 34th International Conference on Machine Learning, 2017

[20] Chollet, François and others, *Keras*, https://keras.io, 2015