Convolutions through Time for Multi-Label Movie Genre Classification

Jônatas Wehrmann
Pontifícia Universidade Católica do RS
Av. Ipiranga, 6681, Porto Alegre-RS, Brazil
jonatas.wehrmann@acad.pucrs.br

Rodrigo C. Barros
Pontifícia Universidade Católica do RS
Av. Ipiranga, 6681, Porto Alegre-RS, Brazil
rodrigo.barros@pucrs.br

ABSTRACT

In this paper, we explore the suitability of employing Convolutional Neural Networks (ConvNets) for multi-label movie trailer genre classification. Assigning genres to movies is a particularly challenging task because genre is an immaterial feature that is not physically present in a movie frame, so off-the-shelf image detection models cannot be easily adapted to this context. Moreover, multi-label classification is more challenging than single-label classification considering that one instance can be assigned to multiple classes at once. We propose a novel classification method that encapsulates an ultra-deep ConvNet with residual connections. Our approach extracts temporal information from image-based features prior to performing the mapping of trailers to genres. We compare our novel approach with the current state-of-the-art techniques for movie classification, which make use of well-known image descriptors and low-level handcrafted features. Results show that our method significantly outperforms the state-of-the-art in this task, improving the classification accuracy for all genres.

CCS Concepts

•**Computing methodologies → Supervised learning by classification; Neural networks;**

Keywords

multi-label classification, movie genre classification, deep neural networks, convolution through time

1. INTRODUCTION

Video analysis is an important Computer Vision (CV) task that is capable of helping human beings to solve a variety of problems that are currently either too tedious or too expensive for them to solve on their own. Most current effort on machine learning approaches for CV tasks focus on classifying images as belonging to one within a thousand of labels [3, 16], whereas video-based applications have shown to be much more challenging. This task is very complex and many traditional and well-established machine learning algorithms have difficulties in properly handling it.

Recent work on video analysis [5, 6, 17] have approached several problems under the perspective of Deep Convolutional Neural Networks (ConvNets) [10], showing exciting first results and paving the way for many novel applications to be fully explored. Indeed, ConvNets are the state-of-the-art method for many CV tasks (e.g., supervised image classification), borrowing concepts from neuroscience to create a mathematical structure capable of assigning meaning to what is seen. They consist of multiple layers of small sets of neurons that process portions of the input data (receptive fields), tiling the outputs so that their input regions overlap. The hierarchy of concepts that are sequentially learnt allows complex mappings from input to desired output. The well-known backpropagation algorithm is employed for training the multiple layers of neurons by backpropagating the gradients of a loss function with respect to the network's weights.

In this paper, we investigate the use of ConvNets for automatically classifying movies according to their genre (e.g., action, horror, drama, comedy). Movie genre classification is a much more challenging task than object detection or scene recognition because of the following problems. First, the classes to be predicted are not physically present within any region of the movie frames. Genres are intangible immaterial features that cannot be pinpointed in a frame or even in a sequence of movie frames like an object can. Furthermore, movie trailers have a much more diverse content than other video-based tasks (e.g., analysis of surveillance cameras), and movie trailers are much larger than typical videos found in action recognition datasets such as UCF [15].

For properly addressing the previously-mentioned issues, we propose a novel method that encapsulates an ultra-deep residual ConvNet that contains a convolution-through-time (CTT) module in the top of the architecture. The CTT module allows the mapping of a sequence of frames into intangible genres in an end-to-end fashion.

This paper is organised as follows. Section 2 discusses related work in the area of movie genre classification.

SAC 2017,April 03-07, 2017, Marrakech, Morocco
Copyright 2017 ACM 978-1-4503-4486-9/17/04. . . $15.00
http://dx.doi.org/10.1145/3019612.3019641

Section 3 describes in detail our proposed approach, whereas Sections 4 and 5 present the experimental analysis that was conducted for validating our novel method. Finally, we end this paper with our conclusions and suggestions for future work in Section 6.

2. RELATED WORK

In this section we describe the related work for the movie genre classification task. To the best of our knowledge, the best methods for genre classification are designed only for the single-label classification scheme. Nevertheless, in real world scenarios, a movie rarely belongs exclusively to a single genre.

Rasheed et al. [13] propose the extraction of low-level features to detect movie genres through the application of the mean-shift classification algorithm [1]. Such features are responsible for describing raw video elements, such as the average shot length, color variance, lighting key, and motion presence. One important elements for low-level feature extraction from movies is the *shot detection* algorithm. A scene boundary is found when the inter-frame similarity is low. Frame similarity is computed via histogram intersection in the HSV color space. A scene boundary is set at the local minima of the inter-frame similarity smoothed function. Each scene is then represented by a single static frame known as the *keyframe*, which is the central frame from the scene.

Another approach for movie genre classification makes use of well-known image descriptors to compute high-level features for each keyframe. The work of Zhou et al. [21] employ the image descriptors Gist [12], CENTRIST [19], and w-CENTRIST to extract high-level features from frames and then perform movie genre classification via the k-NN algorithm. The Gist descriptor tries to encode semantic information like naturalness, openness, roughness, expansion, and ruggedness that represent the dominant spatial structure of a scene [12]. CENTRIST [19] is an image descriptor that applies a spatial pyramid at different levels, breaking the image into smaller patches. This process enables the detection of both local and global information. Finally, w-CENTRIST [21] modifies CENTRIST by taking into account colour information, neither present in Gist nor in CENTRIST. A global multi-dimensional histogram is then built for each trailer using a bag-of-features, where each dimension encodes a part of the trailer. In its final step, each trailer in the test set is processed by the k-NN algorithm that computes its neighbours according to the χ^2 histogram similarity measure.

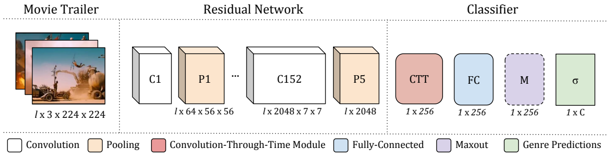

Figure 1: CTT-MMC architecture.

Huang and Wang [4] propose a hybrid approach that combines both low-level visual features and audio information, resulting in a total of 277 features. They make use of the well-known *jAudio* tool [11] to extract audio features such as audio intensity (measured in terms of the the RMS amplitude), timbre (based on different structures of amplitude spectrum), and rhythm. They extract more than 200 audio features via *jAudio*, including the well-known Mel-Frequency Cepstral Coefficients (MFCCs). During the classification process, they make use of the self-adaptive harmony search (SAHS) algorithm in order to search for the optimal subset of features for each of the one-vs-one SVMs that are designed to classify 223 movie trailers from the Apple website.

Simões et al. [14] propose to make use of a ConvNet for extracting visual high-level features from movie frames. The ConvNet is trained at frame level in the LMTD-4 dataset proposed by the authors themselves. They use the extracted features to find scene-clusters in order to build semantic histograms of the trailer scenes. Such histograms are concatenated with a pool of predictions and MFCC information, generating a novel representation for the entire movie trailer. Finally, they use an SVM to predict movie genres for each movie trailer.

Similarly, Wehrmann et al. [18] propose the use of five neural networks (4 ConvNets and 1 MLP) to learn different aspects from the movie trailers. In their approach, 3 GoogleNets [16], 1 C3D [17], and 1 MLP form an ensemble of networks. The generated predictions from the networks are employed in different voting schemes at scene and movie-levels. An SVM is employed to perform the final genre classification by using several class predictions as features.

Both [14, 18] discussed in this section train or fine-tune ConvNets by using the single-label softmax loss function. Adapting both methods for a multi-label scenario is not trivial. For instance, using a multi-label loss function for training a whole ConNet may prevent the convergence of the approaches due to the vanishing gradient problem, or may affect negatively the performance of the proposed methods.

3. CTT-MMC

In this paper, we propose **C**onvolution-**T**hrough-**T**ime for **M**ulti-label **M**ovie genre **C**lassification (CTT-MMC), which is a deep neural network architecture that is designed to take the advantage of the movie trailer frames across time. More specifically, a CTT module is employed so convolutions can be used to learn spatio-temporal feature relationships within the whole trailer. The CTT module draws inspiration from the work of Kim [7] in Natural Language Processing (NLP).

The movie trailer features are learnt by an ultra-deep 152-layer ConvNet with residual connections [3]. It is pre-trained in both ImageNet [9] and Places365 [20] datasets. ImageNet is an image dataset that comprises 14 million images divided in 21,000 classes, being widely used for CV tasks such as object classification and detection. The residual network is pre-trained over 1.2 million images from the ILSVRC 2012 subset of ImageNet. In addition, it was also pre-trained over Places365, which is a scene-centric dataset that contains roughly 1.8 million images from 365 classes. Hence, our ultra-deep ConvNet is suited to learn features regarding both objects and environmental aspects.

Formally, a movie trailer $\mathcal{T} \in \{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3..., \mathcal{T}_l\}$ consists of l keyframes of dimension $w \times h \times c$, where w is the width, h is the height and c is the number of channels. Let $\phi(\mathcal{T}_i) = \mathbf{k_i}$ denote the forward pass of the i^{th} keyframe $\mathcal{T}_i$ through the pre-trained ultra-deep ConvNet, generating feature vector $\mathbf{k_i} \in \mathbb{R}^m$ with m features that represent the respective scene. CTT-MMC generates a temporal representation of the entire movie trailer, X, by stacking all $\mathbf{k_i}$ vertically, i.e., $X \in \mathbb{R}^{l \times m}$. This novel temporal representation is then fed to the CTT module so it can extract temporal relationships among trailer scenes.

In a nutshell, the CTT module in CTT-MMC convolves the frame-based features across time, allowing for temporal relationships to be naturally learnt by the network. After convolving these features, a max-pooling-over-time operation is employed to extract the most representative features from the trailer. Such an operation results in a feature vector that encodes information from the whole movie trailer, enabling the proper mapping from trailer to genre. Figure 1 shows the architecture of CTT-MMC. Note that the architecture is totally connected and allows for end-to-end learning, i.e., the gradients can be backpropagated through the entire scheme, though in this paper we do not modify the weights of the pre-trained ultra-deep residual ConvNet.

For training the models, CTT-MMC optimizes a multi-label loss function $\mathcal{L}(\hat{y}, y)$ – the binary cross-entropy for multiple classes, given by $-\sum_{i=1}^{C} [y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)]$, where, C is the number of classes (i.e, number of movie genres), y_i is the actual class, and $\hat{y}_i$ is the probability predicted by CTT-MMC that the movie trailer belongs to the i^{th} class. Note that y is a binary vector of length C and $\hat{y} \in \mathbb{R}^C$.

3.1 CTT Module

In order to learn temporal and spatial representations of a given movie trailer, CTT-MMC comprises a CTT module. Such a module is composed of a two-dimensional convolution operator that is applied over the features extracted from the frames by the ultra-deep residual ConvNet. The convolution operation comprises n filters of size $f \times m$ that are convolved over X. This operation generates a resulting volume of $l \times 1 \times n$, which is max-pooled over l with a filter of size $l \times 1$, generating $\mathbf{x} \in \mathbb{R}^n$. Let $\psi(X)$ be the computation performed by the convolutional and pooling layers in the CTT module, then $\psi(X) = \mathbf{x}$. Figure 2 shows how the CTT module works.

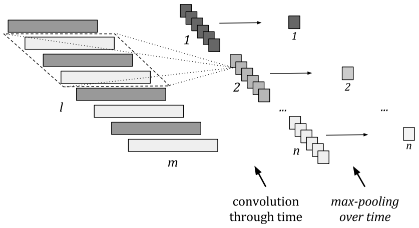

Figure 2: CTT module. In CTT-MMC, $m = 2048$ features and $n = 256$ is the number of convolutional filters.

The post-convolutional movie trailer representation is given by $\mathbf{x}$. Although such representation can be mapped directly to the classes, we also investigate the use of an additional fully-connected (FC) layer and a Maxout [2] activation layer. Hence, the prediction vector for all genres is calculated by $\zeta(\mathbf{x})$, where ζ is a (non-) linear operation over $\mathbf{x}$. The resulting prediction scores undergo the logistic sigmoid activation, i.e., $\mathcal{P} = \sigma(\zeta(\mathbf{x}))$, so they are turned into a probability vector $\mathcal{P} \in \mathbb{R}^C$, where C is the number of genres.

We experiment over three main CTT-MMC's architectural variations, namely: i) using a single FC layer mapping to the classes (hereby called CTT-MMC-A); ii) using two FC layers at the end (hereby called CTT-MMC-B); and iii) using a Maxout layer before the class prediction (hereby called CTT-MMC-C). A Maxout layer [2] performs linear classification by using two distinct weight matrices: Θ_1 and Θ_2. The activation of the Maxout layer is given by the larger of the inner products: $max(\Theta_1^T \mathbf{x}, \Theta_2^T \mathbf{x})$.

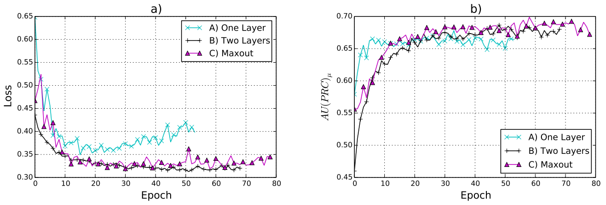

Figure 3: Training behaviour of CTT-MMC's variations. a) training loss across epochs; and b) validation performance (average area under the precision-recall curve) across epochs.

Figure 3 shows the training behaviour of all CTT-MMC's variations using default hyper-parameters. CTT-MMC-A show signs of underfitting, whereas CTT-MMC-B provides a more stable training. We can see that CTT-MMC-C presents the best overall results in validation data.

4. EXPERIMENTAL SETUP

In this section, we detail the dataset that is used in our experiments (Section 4.1), as well as the baseline algorithms (Section 4.2) that we use to compare with CTT-MMC, the evaluation criteria that we employ to measure predictive performance (Section 4.3) and the hyper-parameter settings for CTT-MMC (Section 4.4).

4.1 Dataset

For validating the performance of CTT-MMC, we performed several experiments over the recently-introduced movie trailer dataset LMTD (Labelled Movie Trailer Dataset) [14,18]. As far as we know, LMTD is the larger movie trailer dataset publicly made-available. Its full extend comprises about 10,000 movie trailers from 22 genres, which were assigned according to the IMDB meta-data. In addition, LMTD comprises roughly 400 hours of video and ≈ 30 million frames. Note that the action recognition dataset UCF-101 [15] comprises ≈ 27 hours of video (15× smaller than LMTD).

We make use of LMTD's novel multi-label subset, henceforth called LMTD-9 (Table 1). As the name suggests, it comprises movie trailers from 9 movie genres, namely action, adventure, comedy, crime, drama, horror, romance, sci-fi and thriller. These genres were selected as a consequence of limiting each class to have at least 10% of the total number of instances in LMTD. In LMTD-9, each instance is assigned to at least 1 and at the most 3 genres. In addition, to mitigate problems with outliers, LMTD-9 discards movie trailers with more than 6500 frames and with a release year older than 1980. It is randomly divided in training, validation and test sets as presented in Table 1. LMTD-9 is the largest dataset for movie genre classification to date. It comprises 4007 movie trailers, which is roughly half of all movie trailers in LMTD. The dataset used in [4], in turn, was composed of only 223 trailers: ≈ 18× smaller than LMTD-9.

Table 1: LMTD-9 dataset.

Genre	Training	Validation	Test
Action	611	78	164
Adventure	432	51	108
Comedy	1109	148	301
Crime	477	59	121
Drama	1437	192	394
Horror	324	33	78
Romance	468	59	122
SciFi	229	26	57
Thriller	502	61	129
1 genre	884 (30.90%)	124 (33.15%)	251 (32.47%)
2 genres	1226 (42.85%)	167 (44.65%)	340 (43.98%)
3 genres	751 (26.25%)	83 (22.20%)	181 (23.46%)
Total Movies	2861	374	772

4.2 Baseline Algorithms

To the best of our knowledge, this is the first study that analyzes the movie genre classification problem in a multi-label approach. Hence, we compare CTT-MMC to the state-of-the-art of single-label classification methods, namely video low-level features (VLLF) [13] and audio-visual features (AV) [4]. Note that such methods are not naturally suited for multi-label classification. Thus, for the sake of fairness, we employ the features described in each baseline in a one-vs-all SVM scheme. To generate probabilities for each genre, we use the normalized distance from the margin of the SVM decision boundary. The hyper-parameters for the baselines are set to default values ($c = 1$ and Gaussian kernel with $\gamma = 0.5$). We also provide as baseline the results that would be expected from a random classifier. Such an evaluation is important due to the existence of unbalanced classes. To generate that baseline, we sample random probabilities for the test set from an uniform distribution. These random probabilities are then used to compute all evaluation measures described in Section 4.3.

4.3 Evaluation Measures

The outputs of CTT-MMC for each class are probability values, and the same is true for the baseline algorithms. We follow the trend of multi-label classification research in which we avoid choosing thresholds by employing precision-recall curves (PR-curves) as the evaluation criterion for comparing the different approaches. For generating a PR-curve for a given classification method, one must select a predefined number of thresholds within [0, 1] to be applied over the outputs of each method, finally generating several precision and recall points in the PR plane. The interpolation of these points constitute a PR-curve, and the quantitative criterion one analyzes is the area under such a curve ($AU(\overline{PRC})$).

We employ the following derived measures: $AU(\overline{PRC})$ (micro average), $\overline{AU(PRC)}$ (macro average) and $AU(\overline{PRC})_w$ (weighted average). Each of which allows understanding different aspects regarding each method's capabilities. For instance, $\overline{AU(PRC)}$ measure is calculated by averaging the areas of all classes, which causes less-frequent classes to have more influence in the results. $AU(\overline{PRC})$ is calculated by using all labels globally, providing information regarding the whole dataset, which makes high-frequency classes to affect more the results. Finally, $AU(\overline{PRC})_w$ is calculated by averaging the area under precision-recall curve per genre weighting instances according to the class frequencies. We average the three measures generating a fourth measure called $AU(PRC)_\mu$.

4.4 Hyper-Parameters Settings

For training CTT-MMC, we employ the Stochastic Gradient Descent (SGD) with mini-batches of 64 instances, 256 neurons in the FC layers, learning rate (α) of 1×10^{-3}, weight decay regularization (γ) of 1×10^{-4}, dropout of 0.5 and Adam [8] rule for parameter update. In the test phase, we use the 10-crop strategy described in [16]. In order to provide a fair comparison with the baselines, we have not optimized the hyper-parameters nor performed any kind of feature selection. We train the CTT module for a maximum of 200 epochs, early-stopping when the predictive performance in the validation does not improve for 10 consecutive epochs. In average, our models reach convergence within 70 epochs.

5. RESULTS AND DISCUSSION

In this section we present the experimental results comparing the predictive performance of the following algorithms: CTT-MMC-(A/B/C), VLLF [13], and AV [4].

In Table 2 we report the predictive performance of all methods. Note that all variations of CTT-MMC outperform all baselines by a large margin (values in bold). CTT-MMC's performance regarding $AU(PRC)_\mu$ in the test set is ≈ 38% greater than AV, which is the strongest baseline. For comparison, AV is only ≈ 34% superior than VLLF, showing that the impact of using deep high-level features is far superior than using audio and handcrafted motion information as in [4]. By analyzing the values of $AU(\overline{PRC})$ we can see that the improvement provided by our method over AV is almost twice larger than the improvement provided by AV when compared to VLLF.

The values of $\overline{AU(PRC)}$ show that CTT-MMC presents a solid performance even for trailers of rare classes, such as sci-fi and horror. In this particular scenario, the VLLF method presents virtually random results (0.262 vs 0.206) and AV's performance decreases in ≈ 32% when compared to the $AU(\overline{PRC})$ values. CTT-MMC 's decrease

Table 2: Results comparing CTT-MMC with the state-of-the-art methods and with random classification.

	CTT-MMC-A	CTT-MMC-B	CTT-MMC-C	Random	VLLF [13]	AV [4]
$AU(\overline{PRC})$	**0.712**	**0.704**	<u>**0.722**</u>	0.204	0.457	0.537
$\overline{AU(PRC)}$	**0.618**	**0.599**	<u>**0.624**</u>	0.206	0.262	0.408
$AU(\overline{PRC})_w$	**0.683**	**0.661**	<u>**0.697**</u>	0.294	0.387	0.535
$AU(PRC)_\mu$	**0.671**	**0.676**	<u>**0.680**</u>	0.235	0.368	0.493

in performance is of $\approx$ 16%, which shows it is much more consistent when predicting rare genres. We believe that CTT-MMC's capability of learning semantic features helps it to better discriminate those genres.

Recall that $AU(\overline{PRC})_w$ values gives greater weight for high-frequency classes (e.g, drama and comedy). This fact explains the improvement in performance of the random classifier (0.294). In this scenario, CTT-MMC outperforms VLLF by $\approx$ 80% and AV by $\approx$ 30%. Also, we observe that the difference in $AU(\overline{PRC})_w$ between CTT-MMC and VLLF is proportional to the difference between AV and random classification.

We present in Figure 4 the precision-recall curves generated for all genres regarding the trailers from the test set. We compare the curves among all methods in order to provide some intuition about each method's learning capabilities. We notice that the features extracted by our deep learning architecture provide large improvement in predictive performance for all movie genres, though we highlight the more *subjective* ones. For instance, the *crime* genre is considerably more subjective than *action* or *sci-fi*, and neither visual low-level features nor audio features were helpful for classifying such a genre. In this case, both VLLF and AV presented nearly-random results. CTT-MMC, on the other hand, was capable of correctly predicting most of the *crime*, *sci-fi* and *romance* trailers. We believe that CTT-MMC's ability on finding objects within the trailer scenes is greatly responsible for achieving such good results in these more subjective genres. Furthermore, the good results when classifying *adventure* and *action* genres are probably due to the Places-based learnt features.

Horror is the only movie genre for which CTT-MMC does not present a large improvement. Our hypothesis is that the horror training trailers are overly-heterogeneous, probably requiring more trailers during training for allowing learning to take place. We do believe that extracting audio features could help when classifying horror movies.

Figure 5 depicts predictions generated by CTT-MMC. CTT-MMC correctly predicted all genres for several movies, including *Shakespeare in Love* (1998) and *Fast Five* (2011) (it also predicted the genres in the correct order!). These predictions include more subjective and complex genres such as *crime*, *comedy*, *thriller*, and *romance.* However, CTT-MMC also makes some understandable mistakes, as seen in the predictions for the following movies: (1) *Star Wars: The Force Awakens*, where it predicts sci-fi as the third most probable genre; and (2) *Lord of the Rings: The Return of the King*, whose movie trailer contains several horror-like scenes, such as monsters and dark scenes.

6. CONCLUSION

This paper proposed CTT-MMC, a novel method for multi-label movie genre classification. It comprises an ultra-deep neural network with residual connections that was pre-trained in two large datasets, and also a convolutional-through-time module that makes use of a 2D-convolutional layer and of a max-pooling-over-time layer for encoding temporal information. We have shown that CTT-MMC comfortably outperforms the current approaches for all genres in a large movie trailer dataset, establishing itself as the novel state-of-the-art for multi-label movie genre classification. For future work, we plan to integrate audio information in CTT-MMC but guaranteeing that it remains and end-to-end learning architecture.

7. ACKNOWLEDGMENTS

The authors would like to acknowledge research agencies CAPES (23038.006924/2014-00), CNPq (442231/2014-8), and FAPERGS, as well as Google for funding this research.

8. REFERENCES

[1] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, May 2002.

[2] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. C. Courville, and Y. Bengio. Maxout networks. *Journal of Machine Learning Research*, 2013.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[4] Y.-F. Huang and S.-H. Wang. Movie genre classification using svm with audio and video features. In R. Huang, A. A. Ghorbani, G. Pasi, T. Yamaguchi, N. Y. Yen, and B. Jin, editors, *AMT*, volume 7669 of *Lecture Notes in Computer Science*, pages 1–10. Springer, 2012.

[5] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):221–231, 2013.

[6] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732. IEEE, 2014.

[7] Y. Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[8] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

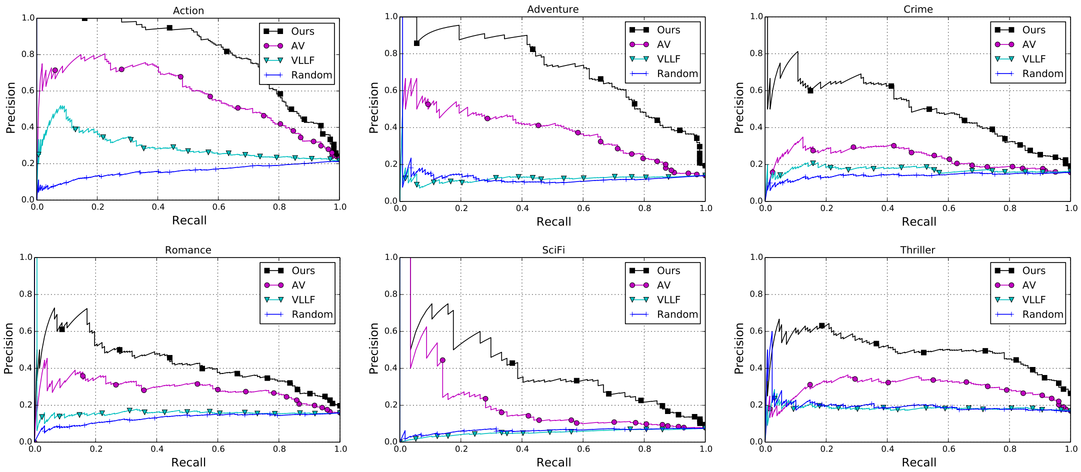

Figure 4: Precision-recall curves for some movie genres.

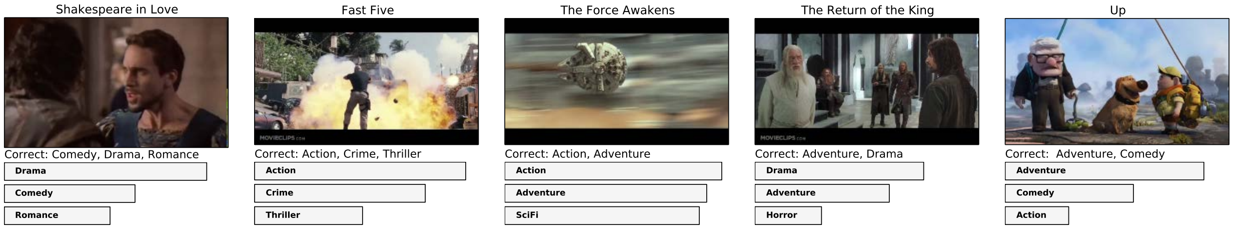

Figure 5: Example of genre predictions generated by CTT-MMC.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[11] D. McEnnis, C. McKay, I. Fujinaga, and P. Depalle. jaudio: An feature extraction library. In *ISMIR*, pages 600–603, 2005.

[12] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[13] Z. Rasheed, Y. Sheikh, and M. Shah. On the use of computable features for film classification. *Circuits and Systems for Video Technology, IEEE Transactions on*, 15(1):52–64, 2005.

[14] G. Simões, J. Wehrmann, R. C. Barros, and D. D. Ruiz. Movie genre classification with convolutional neural networks. In *International Joint Conference on Neural Networks*. IEEE, 2016.

[15] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497. IEEE, 2015.

[18] J. Wehrmann, R. C. Barros, G. Simões, T. S. Paula, and D. D. Ruiz. (deep) learning from frames. In *Brazilian Conference on Intelligent Systems*, 2016.

[19] J. Wu and J. M. Rehg. Where am i: Place instance and category recognition using spatial pact. In *CVPR*, pages 1–8, 2008.

[20] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.

[21] H. Zhou, T. Hermans, A. V. Karandikar, and J. M. Rehg. Movie genre classification via scene categorization. In *Proceedings of the international conference on Multimedia*, pages 747–750. ACM, 2010.