(Deep) Learning from Frames

Jônatas Wehrmann, Rodrigo C. Barros, Gabriel S. Simões, Thomas S. Paula, Duncan D. Ruiz
Pontifícia Universidade Católica do Rio Grande do Sul
Faculdade de Informática
Porto Alegre, RS, Brazil
Email: rodrigo.barros@pucrs.br

Abstract—Learning content from videos is not an easy task and traditional machine learning approaches for computer vision have difficulties in doing it satisfactorily. However, in the past couple of years the machine learning community has seen the rise of deep learning methods that significantly improve the accuracy of several computer vision applications, e.g., Convolutional Neural Networks (ConvNets). In this paper, we explore the suitability of ConvNets for the movie trailers genre classification problem. Assigning genres to movies is particularly challenging because genre is an immaterial feature that is not physically present in a movie frame, so off-the-shelf image detection models cannot be directly applied to this context. Hence, we propose a novel classification method that encapsulates multiple distinct ConvNets to perform genre classification, namely CoNNECT, where each ConvNet learns features that capture distinct aspects from the movie frames. 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 CoNNECT significantly outperforms the state-of-the-art approaches in this task, moving towards effectively solving the genre classification problem.

1 Introduction

Most of the modern computer-based systems and applications make use of Machine Learning (ML) at some extent. ML algorithms aim to automatically learn from experience, outperforming human beings in several tasks from a variety of application domains. Successful applications of ML algorithms include handwritten digit recognition [1], autonomous driving [2], gene expression classification [3], [4], protein function prediction [5], [6], software metrics estimation [7], [8], and real-time stream sensor analysis [9].

Automatically analyzing videos and learning from their content is an important Computer Vision (CV) application that could help humans to solve a plethora of problems that are currently either too tedious or expensive for them to solve on their own. Whereas the number of efficient ML approaches for classifying images as belonging to one within a thousand of labels grows almost exponentially (e.g.,[10], [11]), video-based applications have shown to be much more challenging. Such a task has a high complexity level, and most traditional and well-established ML algorithms have difficulties in handling it effectively.

Learning from videos is a broad concept and offers many research possibilities, such as action recognition, categorization, element recognition, context analysis, and many other tasks. Recent work [12], [13], [14] address video analysis with Deep ConvNets [15], showing exciting first results and possibly paving the way for many applications to be further explored. ConvNets are the state-of-the-art method for supervised image classification, borrowing concepts from image processing to ensure some degree of scale, position, and distortion invariance. They consist of multiple layers of small neuron sets that process portions of the input data, tiling the outputs so that their input regions overlap, thus generating a better representation of the original input.

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 two main problems. First, the classes to be predicted by the ML algorithm are not present within any region of the movie frames. Genres are intangible, immaterial features that cannot be pinpointed in a frame or sequence of movie frames like an object can. Second, since classification is performed over movie frames (or sequences of frames), the training dataset is intrinsically weakly-annotated, i.e., each frame is labeled according to the genre of its respective movie. Note that this weak annotation is problematic given that movies from distinct genres present similar content in most of their frames (e.g., images of people talking, landscapes, roads with cars, etc.). Hence, the ML algorithm will have difficulties in understanding why frames with dialogues are sometimes classified as *drama* and sometimes classified as *comedy*, for example. For properly addressing these issues, our hypothesis is that multiple ConvNets that are trained to learn different aspects of the movie frames/scenes (e.g., motion content, scene recognition, object detection) can actually perform the mapping of a sequence of frames into intangible genres. Our proposed approach is named *CoNNeCT* (Convolutional Neural Networks for Classifying Trailers).

We highlight two important contributions in our work. First, we make publicly available a novel movie trailers dataset, which comprises more than 3500 trailers from 22 genres. To the best of our knowledge, this is the most complete dataset that was publicly provided to date. Second, we present CoNNeCT in detail, and we empirically demonstrate that it significantly outperforms the current

state-of-art movie trailer classification techniques, which employ traditional image descriptors such as Gist [16], CENTRIST [17], and w-CENTRIST [18], or that perform low-level feature classification [19], [20].

This paper is organized as follows. Section 2 presents related work in the area of movie genre classification. Section 3 describes in detail our novel approach, whereas Sections 4 and 5 present the experimental analysis that was conducted for validating our research hypotheses. Finally, we end this paper with our conclusions and suggestions for future work in Section 6.

2 Related Work

Rasheed et al. [19] propose the extraction of low-level features to detect movie genres through the application of the mean-shift classification algorithm [21]. Such features are responsible for describing raw video elements, such as the average shot length, color variance, lighting key, and motion.

A second 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. [18] employs the image descriptors Gist [16], CENTRIST [17], and w-CENTRIST to extract high-level features from frames and then perform movie genre classification via k-NN. The Gist descriptor tries to encode semantic information like naturalness, openness, roughness, expansion, and ruggedness that represent the dominant spatial structure of a scene [16]. CENTRIST [17] 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. Each patch is processed through the Census Transform which compares the pixels with its neighbors. This step produces an 8-bit vector replacing the current pixel. Afterwards, it is appended to the final vector containing all values from the patches. Finally, w-CENTRIST [18] modifies CENTRIST by taking into account color information, neither present in Gist nor in CENTRIST.

It is often the case that the image descriptors output is employed to build a bag-of-visual-words (BOVW) via the well-known k-means clustering algorithm [18], [16], [17]. The final centroids generated by k-means are known as codewords, and each keyframe is assigned to one cluster represented by a codeword. Finally, a global multi-dimensional histogram is built for each trailer, 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 neighbors according to the χ^2 histogram similarity measure.

Huang and Wang [20] propose a hybrid approach that combines both low-level visual features and audio information, reaching a total of 277 features. They make use of the well-known *jAudio* tool [22] 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 with the aid of *jAudio*, including the well-known Mel-Frequency Cepstral Coefficients (MFCCs). Next, 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 used to classify 223 movie trailers from the Apple website.

3 CoNNeCT

In this section we present our approach for movie genre classification, namely *CoNNeCT* (Convolutional Neural Networks for Classifying Trailers). Considering that movie genre classification is a much more complex task than simple object classification/detection, we claim that a single off-the-shelf ConvNet model is not enough for solving the problem (this claim is supported by the experiments performed in Section 5). Hence, our approach makes use of a combination of features extracted by multiple ConvNet models, each of them designed to learn different aspects from the videos, as follows.

The first model is an implementation of the GoogLeNet [10] architecture, which is pre-trained on the well-known ImageNet dataset [24] and fine-tuned with our own movie trailer dataset, namely LMTD (Labeled Movie Trailer Dataset, details in Section 4). The second model is also resulting from a fine-tuning procedure over a GoogLeNet architecture, but pre-trained on the Places dataset [25]. In the third model, instead of fine-tuning a pre-trained model, we trained a GoogLeNet architecture on LMTD from scratch. In order to explicitly extract motion features from the trailers, the fourth model is a 3D ConvNet pre-trained on the Sports 1M dataset [13] and fine-tuned over LMTD as well. Finally, the fifth model is a simple Multi-Layer Perceptron (MLP) whose input are features extracted from the audio (MFCCs) of the videos.

With the five above-mentioned models, we believe we can cover different aspects from the movies, allowing for an easier mapping to a pre-defined movie genre. For instance, the model pre-trained with ImageNet data extracts features that focus on particular elements of the movie frames, whereas the model pre-trained on the Places dataset extracts features which characterize scenes and ambients, providing the context in which particular elements are positioned over. The 3D ConvNet is particularly helpful in characterizing actions and motion in general within the trailers. Finally, the MLP performs the direct mapping of audio (in the form of Mel-Frequent Cepstral Coefficients) into genres.

For illustrating how each fine-tuned ConvNet learns different aspects from the movie trailer, we show positive-class sensitivity heat maps in Figure 1. The frame in the first column was extracted from the movie "Battlefield", and it depicts an aircraft carrier in the middle of the ocean. The following frames show the sensitivity to the positive class (the action genre) as indicated by each model. For instance, the model pre-trained on the Places dataset is very sensitive to the environment, associating the positive class to global aspects from the scene (note the sensitivity to the ocean and sky). Conversely, the model pre-trained on ImageNet is particularly focused on relevant objects, being mostly invariant to the background. Finally, one can see that the GoogleNet trained from scratch on LMTD considers both local and global features, being sensitive to the spatial disposition of elements within each frame. We found that when the convolution layers are locked and the fine-tuning is performed only in the fully-connected layers, the learning

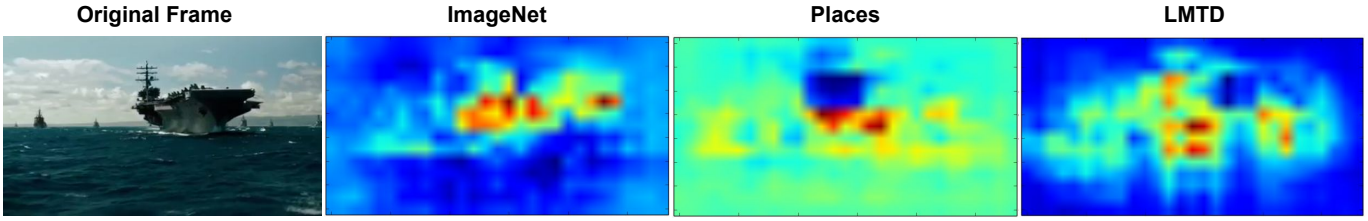

Fig. 1. Positive-class sensitivity analysis following the general procedure described in [23]. The original frame is presented in the left, followed by three heat maps: GoogLeNet pre-trained on ImageNet and fine-tuned on LMTD, GoogLeNet pre-trained on Places and fine-tuned on LMTD, and GoogLeNet trained on LMTD from scratch. A strong sensitivity to the positive class is indicated by warm colors, whereas cold colors represent low sensitivity.

is suboptimal. Considering that LMTD is a large dataset, we fine-tuned all layers of the original architecture with confidence that it would not lead to overfitting. Notwithstanding, we did use much smaller learning rates so we would not distort the pre-trained weights too much (or too quickly).

Considering that each network model captures distinct (albeit complementary) aspects from the frames/scenes of a movie trailer, *CoNNeCT* employs a post-processing learning step which uses the predictions from each model as input to an SVM classifier to generate the final genre predictions. Instead of going for an ensemble strategy of majority voting among the models, we show in Section 5 that this post-processing learning step is much more accurate in predicting genres. We detail the post-processing learning step in Section 3.3.

3.1 Pre-Processing Step

We performed multiple procedures to collect, clean, and augment the data from LMTD prior to the training/tuning of each model. The three models based on the GoogLeNet architecture make use of the same set of frames extracted from individual scenes from each trailer. First, we employ the shot detection algorithm described in [19] for detecting the set of scenes that a trailer contains. Next, we identify the keyframe (central frame) of each scene, and we collect it along with 20% of the frames in that same scene, in order to have diversity within the collected data without resorting to all frames in a trailer. Regarding the 3D ConvNet, we collected a sequence of 16 frames from each detected scene. If a scene has 32 frames or more, we split it into two instances, each of which containing 16 frames. If a scene has less than 16 frames, we discard it and continue to process the next scenes. Each instance in the 3D ConvNet is thus a sequence of 16 frames, and multiple instances may refer to the same movie scene whereas short scenes are discarded.

All frames are downsized to 256×256, and then randomly cropped during training to 224×224 and eventually mirrored (uniform probability). Considering that we use colored images, the 2D ConvNets have inputs of size $224 \times 224 \times 3$ (height, width, color channels), whereas the 3D ConvNet inputs are of size $112 \times 112 \times 3 \times 16$ (height, width, color channels, frames). Height and width were reduced 2-fold in the 3D model due to the available computational resources.

For processing the audio of the videos, we extracted the Mel Frequency Cepstral Coefficients (MFCCs) from each trailer scene. Since MFCCs extraction generates 13 variable-size feature vectors, we computed four different statistics from each one, namely the minimum and maximum values, standard deviation, and average. The combination of the statistics obtained from each of the 13 vectors and of the two corresponding deltas resulted in a single 156-long vector ($13 \times 4 \times 3$), which is then used as input to the Multi-Layer Perceptron.

3.2 *CoNNeCT* Models

With the goal of identifying particular elements within the movies frames, *CoNNeCT* makes use of a GoogLeNet model [10] that is pre-trained on the ImageNet dataset [24]. ImageNet (ILSVRC12) is the well-known image dataset that comprises 1.2 million images divided in 1.000 classes, being widely used for computer vision tasks such as object classification. In the ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC14), a GoogLeNet based model was responsible for establishing the state-of-the-art in object classification [10].

CoNNeCT employs the GoogLeNet pre-trained on the ImageNet dataset and fine-tuned on LMTD. The fine-tuning is executed in batches of 128 images for 10 epochs. The initial learning rate is set to 1×10^{-4} and it decreases 10-fold whenever the validation loss plateaus.

The second model is an implementation of the GoogLeNet architecture focused on extracting features based on scenes and environments. For such, this model is pre-trained on the Places dataset [25] and fine-tuned on LMTD. Places is a scene-centric dataset that contains over 2.4 million labeled images of scenes divided in 205 classes. The size of batches, epochs and learning rate strategy are the same as previously described.

Rather than using a pre-trained model, the third *CoNNeCT* model is a GoogLeNet trained on LMTD from scratch. The main goal of this model is to capture information that relates the movie frames with genres regardless of any previous knowledge other than the movie itself. We run it in batches of 128 images for 20 epochs, with the initial learning rate set to 1×10^{-3}, with the same decreasing policy as the previous networks.

In order to extract motion features from the trailers, *CoNNeCT* comprises a 3D ConvNet based on the C3D architecture [26], [27], which is more suitable for spatiotemporal feature learning than conventional 2D ConvNets. This model is pre-trained on the Sports-1M dataset [13] and fine-tuned on LMTD with batches of 32 frame-sequences (16

frames per sequence) executed for 20 epochs, initial learning rate of 1×10^{-4}, with the same decreasing policy as before.

The last *CoNNeCT* model is an MLP that receives as input the MFCCs audio features. The MLP architecture is $156 \times 312 \times 312 \times 4$ with hyperbolic tangent neurons, trained with Nesterov's Accelerated Gradient with an initial learning rate of 0.01 and dropout of 50%.

3.3 Post-Processing Learning Step

The final step of *CoNNeCT* is to perform the concatenation of the predictions from each of the five models. Since the models generate predictions in different granularities (per-frame or per-scene), we need to put them into the same granularity (per-scene). The three GoogLeNet-based models are the only ones to provide per-frame predictions, so we average the predictions for all frames in the same scene.

Once all predictions are scene-based, we noticed that averaging over scenes in order to generate per-trailer predictions would lead to severe information loss. Hence, we divided the trailer in p parts with uniform frequency of scenes, averaging the predictions from those scenes located in a same part. After this procedure, the final feature set contains a total of $c \times p \times 5$ features (number of classes $\times$ number of parts $\times$ number of models), which serve as input to an SVM classifier. Since each instance is now a set of genre probabilities from a single movie that vary in time, the classifier can learn the function that maps such features to the desired genre, properly addressing the issues of learning intangible information from frames.

4 Experimental Methodology

To validate the hypothesis that movie trailer genres can be properly identified by *CoNNeCT* , we need a labeled movie trailer dataset. Zhou et. al. [18] describe their own movie trailer data, though it is not made publicly available for the research community. Moreover, 54% of the trailers in their dataset belong at the same time to three out of the four genres, the same genres evaluated by our research. Their reported accuracy values consider a correct classification whenever their approach classifies the movie trailer as belonging to any of the labeled genres, which means movies with 3 genres has a 75% probability of being correctly classified simply by chance.

We have developed a novel movie trailers dataset called LMTD (Labeled Movie Trailer Data), which comprises more than 3500 trailers whose genres are known, and we make it publicly available for the interested reader[1]. The ≈ 3500 movie trailers are distributed over 22 different genres. To avoid the problems identified in the work of Zhou et. al. [18], we have selected a subset of 999 movie trailers from LMTD, as presented in Table 1, where each trailer belongs to one of 4 disjoint genres (action, comedy, drama, or horror). Note that this subset is a consequence of i) restricting to 4 genres among the 22 existing ones; and ii) selecting all disjoint movie trailers from the 4 selected genres. This subset is called LMTD-4. The training, validation, and test sets were chosen randomly among the available trailers.

1. http://www.inf.pucrs.br/gpin/

TABLE 1
LMTD-4 dataset.

Genre	Training	Validation	Test	Total
Action	160	15	90	265
Comedy	160	15	95	270
Drama	160	15	90	285
Horror	114	10	55	179
Frames	1,425,600	132,000	792,000	2,349,600

To validate our results we compare *CoNNeCT* with the state-of-the-art methods in movie genre classification, namely Gist [16], CENTRIST [17], w-CENTRIST [18], and two approaches based on low-level features extraction [19], [20]. For Gist, CENTRIST, and w-CENTRIST we set the same parameters as defined in [18], namely: BOVW of 200 codewords and 100 bin histogram with $t = 3$. We replaced the k-NN classification performed by the authors by an SVM classification (RBF kernel, $\gamma = 0.1$ and $C = 1$), considering the vast improvement achieved in validation data.

The low-level features extraction approach presented by Rasheed et al. [19] is not directly comparable to other methods since its main goal is not to classify genres but to understand the relationship between features and genres. Therefore, we employed the same strategy than for the previous methods, which is performing SVM classification with RBF kernel, $\gamma = 0.1$, and $C = 1$. For the second low-level features based approach, proposed by Huang and Wang [20], we set the parameters as suggested by the authors in their experimental analysis: SAHS with HMS set to 50 and HMCR set to 0.99; SVMs with RBF kernel and parameters γ and C tuned in the validation set considering a grid of 6x6 combinations between $[2^{-4}, ..., 2^{1}]$ and $[2^{-2}, ..., 2^{3}]$.

We also set as baseline approaches each individual network that is part of *CoNNeCT* , with the goal of verifying the hypothesis that multiple models capable of learning distinct features would outperform any single model being used individually. Our last baseline is a modification of *CoNNeCT* that performs ensemble-like classification by aggregating the predictions of the multiple models instead of performing the post-processing step with SVMs. Our goal here is to verify whether making use of the predictions from each model to feed an SVM in a post-processing learning step is a more robust approach than aggregating predictions in a weighted vote scheme. We refer to each individual model in *CoNNeCT* as follows: G-ImageNet, G-Places, G-LMTD, 3D ConvNet, Audio MLP, and we refer to the *CoNNeCT* approach that performs ensemble classification instead of the post-processing step as *E-CoNNeCT* .

5 Results and Discussion

We first analyze the performance of each network comprised by *CoNNeCT* when used individually to predict genres. In Table 2, we show the performance of each network in the validation set, averaging their predictions from frames/scenes to the entire movie trailer. Note that we also present the performance of *E-CoNNeCT* , which is the modified version of *CoNNeCT* that does not perform the

post-processing learning step. In *E-CoNNeCT* , we average the frame-based predictions generated by G-Places, G-ImageNet, and G-LMTD, to the scene granularity, and then we average the aggregated scene predictions with those generated by the 3D ConvNet and Audio MLP. Finally, we average the predictions from scenes to trailers.

Table 2 shows that the overall accuracy of each individual network is quite low, with G-Places outperforming the other networks. It also shows that combining the multiple models into a single scheme and weight-averaging the results does not provide the best overall accuracy (*E-CoNNeCT* reaches 42% versus 44% for G-Places). We argue that by aggregating the predictions from the distinct models we lose important information for defining genre. Our claim is that genre should be defined based on the relationship among these distinct predictions, and not based on their aggregation.

TABLE 2
Per-genre and overall accuracy in the validation set. Predictions are averaged from frames/scenes to trailers.

Network	Action	Comedy	Drama	Horror	Overall accuracy
G-Places	**0.60**	0.47	**0.40**	0.2	**0.44**
G-ImageNet	0.47	0.33	0.27	0.1	0.31
G-LMTD	0.47	0.4	0.27	0.1	0.33
3D ConvNet	0.47	**0.53**	**0.40**	0.2	0.42
Audio MLP	0.33	**0.53**	0.13	0.2	0.31
E-CoNNeCT	0.4	**0.53**	**0.40**	**0.30**	0.42

For backing up that claim, we show in Table 3 the performance of each individual network followed by the post-processing learning step with SVMs. For the frame-based networks, we average the predictions from frames to scenes, and then we divide each trailer in 12 parts[2] with uniform frequency of scenes. Next, we average the predictions from scenes to parts, resulting in 48 features (4 predictions $\times$ 12 parts) per trailer, which are used as input to a SVM classifier with RBF kernel, $\gamma = 1$ and $C = 1$ (default values). The same rationale is applied to the scene-based networks, averaging predictions from scenes to parts and then performing SVM classification. The final number of features that are used as input to the SVM classifier is 240 (48×5 models). We also show in Table 3 the *CoNNeCT* performance, so we can evaluate whether predictions from multiple models improve over the individual ones.

TABLE 3
Per-genre and overall accuracy in the validation set. Predictions are averaged from frames/scenes to parts followed by the post-processing learning step with SVMs.

Network	Action	Comedy	Drama	Horror	Overall Accuracy
G-Places	0.66	**0.80**	0.73	**0.70**	0.73
G-ImageNet	**0.73**	**0.80**	0.73	**0.70**	0.75
G-LMTD	0.47	0.6	0.67	0.6	0.58
3D ConvNet	**0.73**	0.73	0.6	0.70	0.70
Audio MLP	0.53	**0.80**	0.47	0.20	0.53
CoNNeCT	**0.87**	**0.80**	**0.87**	**0.80**	**0.84**

Table 3 shows that the post-processing learning step substantially improves the accuracy of all models (from $\approx 40\%$ to $\approx 70\%$). Moreover, note that *CoNNeCT* substantially outperforms the most accurate individual networks (G-Places and G-ImageNet), showing an improvement of $\approx 10\%$ by combining predictions from multiple models. Figure 2 shows the effect of sequentially adding the predictions of each model to the set that initially contains the G-LMTD predictions. A given position in the x axis indicates the model whose predictions are being added to the set of predictions from the models on its left.

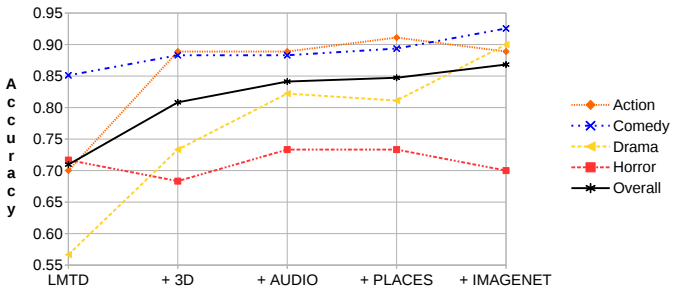

Fig. 2. Per-genre test accuracy computed by sequentially adding predictions from the baseline models into the post-processing learning step. From left to right: 1) 48 predictions from the G-LMTD model; 2) 48 predictions from the 3D ConvNet plus the previous 48 predictions from G-LMTD; 3) 48 predictions from Audio MLP plus the previous 96 predictions; 4) 48 predictions from G-Places plus the previous 144; 5) 48 predictions from G-ImageNet plus the previous 192.

Note how the overall accuracy substantially increases when sequentially adding more information from the network models, confirming our hypothesis regarding the benefits of extracting features of different aspects from the frames/scenes. The *drama* genre, in particular, greatly benefits from a multiple-model approach, going from $\approx 55\%$ to $\approx 90\%$ of accuracy. Other interesting finding is the gain of accuracy for the *action* genre when using a 3D ConvNet, which makes sense considering that *action* movies can be more naturally described by motion-based features. The only genre that does not seem to benefit from knowledge extracted by multiple models is *horror*, which only gains in accuracy when adding audio features and scene information, and has its accuracy decreasing when making use of object-oriented features.

In our last analysis, we present the performance of all baseline algorithms along with *CoNNeCT* in Table 4. Note that *CoNNeCT* outperforms the current state-of-the-art in 13% of accuracy, showing once again the power of ConvNets in Computer Vision applications. Whilst the baseline approaches struggle when predicting the *drama* genre, observe that *CoNNeCT* comfortably reaches 90% of accuracy, an improvement of 23% over the second-best approach! Yet, the downside is the *horror* genre, which is outperformed by the work of Huang and Wang [20]. We are still not certain of the reasons for *CoNNeCT* 's lack of performance when classifying *horror* trailers, but we believe that performing feature selection over the set of 240 features may increase its performance, albeit not by a great margin.

Our final remark is regarding scalability and adaptation to multi-label classification: *CoNNeCT* can naturally perform multi-label classification, which is not true for neither of the baseline approaches. Moreover, adding extra classes to the problem does not impact severely on *CoNNeCT* 's

2. This value was selected based on the accuracy in validation data. We found that values of $4 \leq p \leq 19$ differ marginally in terms of accuracy.

TABLE 4
Per-genre and overall test accuracy of all baseline algorithms and *CoNNeCT* .

Method	Action	Comedy	Drama	Horror	Overall
Low-Level + SVM [19]	0.54	0.35	0.50	0.10	0.41
GIST + SVM [18]	0.57	0.61	0.31	0.40	0.48
CENTRIST + SVM [18]	0.58	0.55	0.47	0.37	0.51
w-CENTRIST + SVM [18]	0.55	0.54	0.44	0.35	0.49
One-vs-One SVM [20]	0.74	0.83	0.67	**0.72**	0.74
CoNNeCT [Ours]	**0.89**	**0.92**	**0.90**	0.70	**0.87**

computational cost. The work of Huang and Wang [20], on the other side, would require 231 SVM classifiers to recognize the full extent of genres in LMTD (22).

6 Conclusions

We presented a novel approach to learn genre from movie trailers based on Convolutional Neural Networks (ConvNets), namely *CoNNeCT* . It recognizes disjoint movie genres with 87% of accuracy, substantially surpassing the current state-of-the-art approaches. *CoNNeCT* innovates by combining predictions from multiple models in order to address a semantic gap between the frame/scene granularity and the movie granularity. This work has shown that it is possible for multiple ConvNets to learn an intangible feature such as movie genre even resorting to a weak labeled dataset, in which frames were labeled according to the overall movie genre. As future work, we intend to investigate the movie genre domain under the perspective of multi-label classification, eventually making use of the entire set of 22 movie genres. Another interesting venue is to investigate automatic approaches for labeling scenes in order to avoid the weak-labeling issue previously described.

Acknowledgements

The authors would like to thank CAPES, CNPq, and FAPERGS for funding this research. Also, they would like to gratefully acknowledge the support of NVIDIA Corporation with the donation of two Tesla K40 GPUs.

References

[1] B. B. Le Cun, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in neural information processing systems*, 1990.

[2] U. Dogan, J. Edelbrunner, and I. Iossifidis, "Autonomous driving: A comparison of machine learning techniques by means of the prediction of lane change behavior," in *2011 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Dec 2011, pp. 1837–1843.

[3] R. C. Barros, M. P. Basgalupp, A. C. P. L. F. de Carvalho, and A. A. Freitas, "Automatic design of decision-tree algorithms with evolutionary algorithms," *Evol. Comput.*, vol. 21, no. 4, pp. 659–684, Nov. 2013. [Online]. Available: http://dx.doi.org/10.1162/EVCO_a_00101

[4] R. C. Barros, M. P. Basgalupp, A. A. Freitas, and A. C. P. L. F. de Carvalho, "Evolutionary design of decision-tree algorithms tailored to microarray gene expression data sets," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 6, pp. 873–892, Dec 2014.

[5] R. Cerri, R. C. Barros, and A. C. P. L. F. de Carvalho, "A genetic algorithm for hierarchical multi-label classification," in *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, ser. SAC '12. New York, NY, USA: ACM, 2012, pp. 250–255.

[6] "Hierarchical multi-label classification using local neural networks," *Journal of Computer and System Sciences*, vol. 80, no. 1, pp. 39 – 56, 2014.

[7] R. C. Barros, D. D. Ruiz, N. N. Tenorio, M. R. Basgalupp, and K. Becker, "Issues on estimating software metrics in a large software operation," in *32nd Annual IEEE Software Engineering Workshop (SEW '08)*, Oct 2008, pp. 152–160.

[8] M. P. Basgalupp, R. C. Barros, T. S. da Silva, and A. C. P. L. F. de Carvalho, "Software effort prediction: A hyper-heuristic decision-tree based approach," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, ser. SAC '13. New York, NY, USA: ACM, 2013, pp. 1109–1116.

[9] J. A. Silva, E. R. Faria, R. C. Barros, E. R. Hruschka, A. C. P. L. F. d. Carvalho, and J. a. Gama, "Data stream clustering: A survey," *ACM Comput. Surv.*, vol. 46, no. 1, pp. 13:1–13:31, Jul. 2013.

[10] 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.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[12] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[13] 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 (CVPR)*, 2014, pp. 1725–1732.

[14] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *Human Behavior Understanding*. Springer, 2011, pp. 29–39.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[16] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[17] J. Wu and J. M. Rehg, "Where am i: Place instance and category recognition using spatial pact," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[18] 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*. ACM, 2010, pp. 747–750.

[19] 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*, vol. 15, no. 1, pp. 52–64, 2005.

[20] Y.-F. Huang and S.-H. Wang, "Movie Genre Classification Using SVM with Audio and Video Features," in *AMT*, ser. Lecture Notes in Computer Science, R. Huang, A. A. Ghorbani, G. Pasi, T. Yamaguchi, N. Y. Yen, and B. Jin, Eds., vol. 7669. Springer, 2012, pp. 1–10. [Online]. Available: http://dblp.uni-trier.de/db/conf/amt/amt2012.html#HuangW12

[21] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, May 2002.

[22] D. McEnnis, C. McKay, I. Fujinaga, and P. Depalle, "jaudio: An feature extraction library." in *ISMIR*, 2005, pp. 600–603. [Online]. Available: http://dblp.uni-trier.de/db/conf/ismir/ismir2005.html#McEnnisMFD05

[23] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 818–833.

[24] J. Deng, W. Dong, R. Socher, L. jia Li, K. Li, and L. Fei-fei, "Imagenet: A large-scale hierarchical image database," in *In CVPR*, 2009.

[25] 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*, 2014, pp. 487–495.

[26] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "C3D: generic features for video analysis," *CoRR*, vol. abs/1412.0767, 2014. [Online]. Available: http://arxiv.org/abs/1412.0767

[27] ——, "Learning Spatiotemporal Features with 3D Convolutional Networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4489–4497.