Multimodality and Deep Learning when predicting Media Interestingness

Eloïse Berson, Claire-Hélène Demarty, Ngoc Q. K. Duong
Technicolor, France
eloise.berson@gmail.com,{claire-helene.demarty,quang-khanh-ngoc.duong}@technicolor.com

ABSTRACT

This paper summarizes the computational models that Technicolor proposes to predict interestingness of images and videos within the MediaEval 2017 Predicting Media Interestingness Task. Our systems are based on deep learning architectures and exploit the use of both semantic and multimodal features. Based on the obtained results, we discuss our findings and obtain some scientific perspectives for the task.

1 INTRODUCTION

Understanding interestingness of media content such as images and videos, has gained a significant attention from the research community recently as it offers numerous practical applications in *e.g.,* content selection or recommendation [1, 2, 5]. Following the success of the 2016 edition [4], the MediaEval 2017 Predicting Media Interestingness Task extends the benchmark to larger datasets, also annotated with a greater human annotation effort. A complete description of the task can be found in [3].

For both subtasks, Technicolor's motivation was to build incrementally from last year's systems [11], i.e., re-use similar features and DNN architectures, while adding some contextual information to the content. To this end, two new features were added, so as to capture additional semantic information related to the content, following a similar idea as in [8]. These new features (section 2), were expected to bring contextual information related to the content. In a second step (section 3), and for the video subtask only, several embeddings of this semantic information at different network levels were experimented. The aim was to investigate how this was influencing the temporal modeling of this new information.

2 MULTIMODALITY AND CONTEXTUAL FEATURES

As in 2016, CNN coming from the fc7 layer of the pre-trained caffeNet model (image modality, both subtasks) and MFCCs concatenated with their first and second derivatives (audio modality, video subtask) were extracted following the protocol described in [11]. Dimensions for these features were 4096 and 180, respectively.

To capture some additional semantic information, Image-Captioning Based (ICB) features [7] were computed for each image or frame, depending on the subtask. These features correspond to the projection of an image in a visual-semantic embedding space [7], obtained from a jointly-trained model for images and captions dedicated to automatic captioning. In this embedded space, where semantic distances between projected image and captioning features are minimized, the resulting representation features are more likely to contain semantic information than the CNN features alone. Dimension of the ICB feature is 1024.

To go further in this vein of adding semantic and contextual information, textual metadata was directly extracted from ImDB [1], exploiting the fact that the MediaEval 2017 dataset was built from Hollywood-like movie extracts. Except for 3 movies (for 2 of them, a short summary was built from descriptions found on the internet; for the last one, description was left empty), ImDB information was available: each movie description and/or storyline was proposed at the input of the RAKE algorithm [10], for keyword extraction. Thus, several keywords were extracted per movie, from which we derived a textual feature of dimension 300, classically using the Word2Vec [9] representations (pretrained on GoogleNews dataset) of this batch of keywords and averaging them.

3 DNN ARCHITECTURES

Global workflows for all submitted runs and for both subtasks are shown in Figure 1. As stated in the introduction, most components used to build the systems' architectures for both subtasks were the same as in [11]. Thus, to cope with the unbalance of the dataset, some resampling of the data was applied during training. Several parameter configurations were investigated by splitting the dataset in 80% for training and 20% for validation. A final retraining of the best model was then applied on the complete development set.

For the image subtask, different concatenations of the features were investigated to understand the contribution of each modality and to conclude on the input of contextual information to the task. Thus each submitted run differs from the others by the input features, and the adaptation of the layer sizes, while the DNN architecture remains the same: a single MLP layer, with rectified linear unit (ReLU) activation and a dropout of 0.5. All submitted runs are summarized in Figure 1a, with different colors depending on the feature concatenation; Run#1, corresponding to 2016 best system, will serve as a baseline.

For the video subtask, three levels of embedding for the W2V features were investigated (see Figure 1b), except for Run#1 which re-uses one of last year's systems (Run#3 in [11], see Figure 1 for the used layers, each of them with ReLU activation function, followed by a dropout equal to 0.5). Run#1's architecture is kept for the other runs, with some adaptation of the multimodal block depending on the input feature sizes (one or two LSTM layers, with a residual block). In Run#1, our baseline, audio and video modalities only are used. For the image channel, a first modal-specific learning step was implemented with a MLP layer followed by a LSTM layer. For the audio, a single LSTM layer is used. After merging, both channels serve as input to two LSTM layers, with a residual part (ResNet [6]).

MediaEval'17, 13-15 September 2017, Dublin, Ireland

[1] see http://www.imdb.com

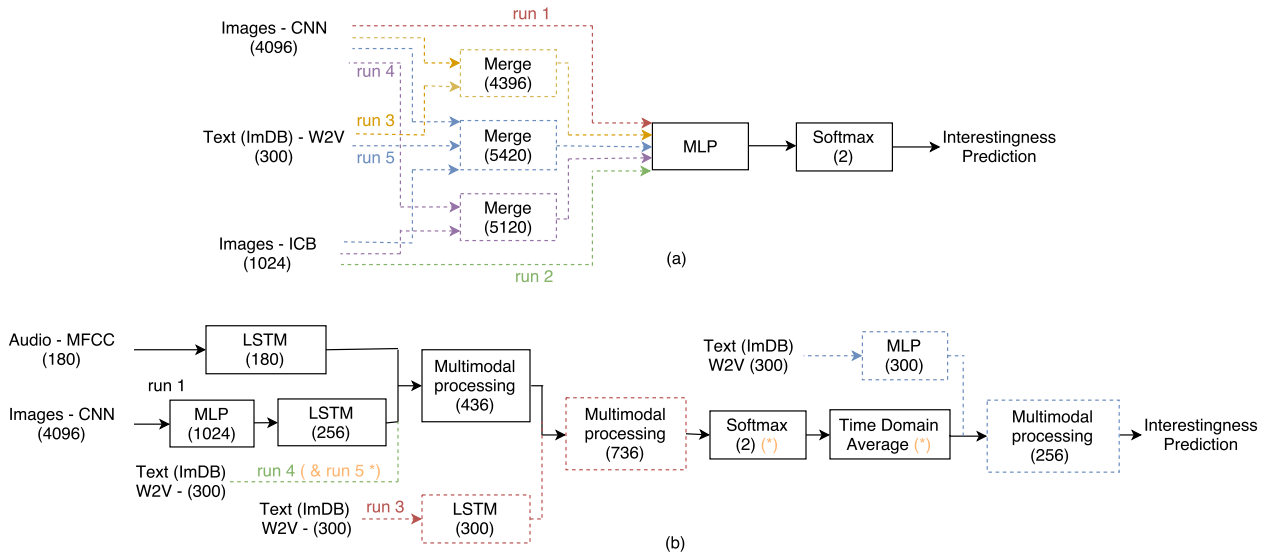

Figure 1: Workflows for all run submissions: (a) image subtask. (b) video subtask.*: block swapping.

In Run#2, W2V features were simply merged to the result of the temporal modeling for the other modalities, whereas for Run#3 and Run#4, they were duplicated for each frame and merged into the workflow either in parallel to the audio and video channels (Run#4) or after a first merging of these two modalities (Run#3) (See Figure 1). For each run, some potential processing steps were added to realize the merge with the other modalities thanks to either additional LSTM-ResNet layers when temporal modeling was possible (Runs#3 and 4), or simple MLP layers otherwise (Run#2). These steps were followed by a simple concatenation of the features from the different modalities. Run#5 is similar to Run#4 except that the *Time Domain Average* and *Softmax* steps were swapped. The motivation for this last run was to test whether the location of the decision step (softmax) had an influence on the performance.

4 RESULTS AND DISCUSSION

Results are summarized in Table 1. First runs for both subtasks show slightly improved MAP values compared to last year's results. As the systems remain the same for these runs over the two years, it tends to show that the dataset size increase and/or the refinement of the annotations had an effect on the modeling performance. Unexpectedly, the MAP@10 values are very low (lower than during the validation process, when MAP@10 values were of the same range or slightly lower than MAP for both tasks). Another unexpected result is that contextual features, either ICB or W2V features, did not bring any improvement to the image subtask, although we had an opposite conclusion during validation on the development set with MAP values of resp. 0.36 and 0.38 for those features (for comparison, we obtained 0.31 with CNN features). This suggests that using more features might have led to over-fitting, probably because of the small size of the dataset during training. This over-fitting might have also been reinforced as, because of a lack of computation resources, cross-validation was done with one fold only. In the future, some cross-validation process with more folds might lead to a better system. However, once the test set is released,

Runs	Image Subtask		Video Subtask	
	MAP	MAP@10	MAP	MAP10
Run#1	**0.2615**	0.1028	0.1856	0.0589
Run#2	0.2525	**0.1054**	0.1768	0.0465
Run#3	0.2244	0.0693	0.1825	0.0563
Run#4	0.2382	0.0875	0.1878	**0.0641**
Run#5	0.2347	0.0861	**0.1918**	0.0609

Table 1: Results on both subtasks (Official metric: MAP@10).

further analysis of the differences between the development and test sets should be done to better understand this observation.

For the video subtask, as expected, W2V features slightly improved MAP and MAP@10 when considered as a frame-based feature. Although they are simply repeated for each frame, i.e., each frame of a given video shares the same textual feature, the concatenation of this new information did bring some useful information for the video subtask. This difference between the two subtasks reinforces the difference between image and video interestingnesses which was already stated last year. Run#5 of the video subtask suggests that keeping the classification for the final step of the system is maximizing the performance, which is understandable as it allows to keep continuous values as much as possible before switching to a binary classification. It also corresponds better to the annotation protocol where the annotation is done for each video segment as a whole; thus the softmax prediction should also be done for the whole segment and not for every single frame.

As a conclusion, a lot of our findings on the evaluation step were different from those of the test set. We definitely need to understand what differs from these two sets that is responsible for the differences in performance. E.g., some new, significantly longer and thus more meaningful segments (243 out of 2435) were added to **the test set only**, representing a duration of 46min over a total duration of 87min, i.e., more than half of the test set.

REFERENCES

[1] Sharon Lynn Chu, Elena A Fedorovskaya, Francis KH Quek, and Jeffrey Snyder. 2013. The effect of familiarity on perceived interestingness of images.. In *Human Vision and Electronic Imaging*.

[2] Claire-Helène Demarty, Mats Sjöberg, Gabriel Constantin, Ngoc Q. K. Duong, Bogdan Ionescu, Thanh-Toan Do, and Hanli Wang. 2017. Predicting Interestingness of Visual Content.

[3] Claire-Helène Demarty, Mats Sjöberg, Bogdan Ionescu, Thanh-Toan Do, Michael Gygli, and Ngoc Q. K. Duong. 2017. MediaEval 2017 Predicting Media Interestingness Task. *MediaEval 2017 Workshop* (September 2017).

[4] Claire-Helène Demarty, Mats Sjöberg, Bogdan Ionescu, Thanh-Toan Do, Hanli Wang, Ngoc Q. K. Duong, and Frederic Lefebvre. 2016. MediaEval 2016 Predicting Media Interestingness Task. *MediaEval 2016 Workshop* (October 2016).

[5] Helmut Grabner, Fabian Nater, Michel Druey, and Luc Van Gool. 2013. Visual Interestingness in Image Sequences. In *Proceedings of the 21st ACM International Conference on Multimedia (MM '13)*. ACM, New York, NY, USA, 1017–1026. https://doi.org/10.1145/2502081.2502109

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. In *arXiv prepring arXiv:1506.01497*.

[7] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* (2014).

[8] Brian A. Plummer, Matthew Brown, and Svetlana Lazebnik. 2017. Enhancing Video Summarization via Vision-Language Embedding. In *Computer Vision and Pattern Recognition, 2017. CVPR 2017. Proceedings of the International Conference on*. IEEE.

[9] Xin Rong. 2014. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738* (2014).

[10] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory* (2010), 1–20.

[11] Yuesong Shen, Claire-Hélène Demarty, and Ngoc Q. K. Duong. 2016. Technicolor@ MediaEval 2016 Predicting Media Interestingness Task.. In *MediaEval*.