Cross-modal Interaction for Video Memorability Prediction

Youwei Lu[1], Xiaoyu Wu[1]
[1]Communication University of China, China
wowanglenageta@sina.com

ABSTRACT

It is important to select memorable videos from the huge amount of videos, which can serve other fields, such as video summary, movie production, *etc.* The Predicting Media Memorability task in MediaEval2021 focuses on predicting how well a video is remembered. In this paper, we use a text-guided visual cross-modal guidance approach for the video memorability prediction task. Based on this, we use a late fusion approach to fuse features from multiple modalities and predict the final video memorability scores.

1 INTRODUCTION

The image memorability task is already a relatively mature field, and much work has been proposed to study it [1, 8–10]. However, the video memorability prediction task is a brand new task from an artificial intelligence perspective. For images, people may memorize a certain region in the image, which leads to high memorability scores. For videos, people may memorize certain frames in a video and video memorability prediction is a more complex and difficult task. The Predicting Media Memorability task in the MediaEval 2021 workshop [11] is designed for this purpose, with the aim of investigating how to assess better the degree to which a video gives a moment of memory. Video memorability scores are used to measure this metric. Over the past two years, work has been done on the video memorability prediction task in the 2019 [14, 19] and 2020 [12, 13] editions of the task, where we looked at and considered the advantages and disadvantages of other methods, and finally we proposed our own method for predicting video memorability scores.

2 RELATED WORK

There are multiple attributes in videos, such as vision and audio, which play important roles in video memorability prediction. Researchers have used different methods to extract the features of multiple modalities to obtain a good feature representation. For example, the authors in [19] tried to extract features of video frames using 2D convolution method, Inception-V3 and used them to compose features of the whole video. Authors in [20] tried to extrat textual features with Glove model [17], which is a common model used in the NLP field. Researchers in [12] used a VGGish model [12] to extract audio features.

Cross-modal interaction approaches are widely used in the field of computer vision. For example, in [15], textual features are used to enhance the representation of visual features in the image captioning task and this is achieved with good results. We therefore try to introduce the approach of cross-modal interaction methods to the field of video memorability prediction.

3 APPROACH

As we have previously described, visual, textual, and audio information play an important role in the video memorability prediction task. We therefore carefully considered the feature extraction steps for each modality. At the same time, we argued that since the text was manually annotated based on the video content, there was semantic consistency between the textual and visual content, and since previous studies have shown that textual information played a role in memorability prediction tasks [3, 18], textual features were used to guide the representation of visual features, and the two modal features were interacted. After obtaining the features from each of the above three modalities, they were passed through several MLP structures and predicted their respective video memorability scores. Finally, we used an adaptive score fusion strategy to fuse the scores of the three modalities.

3.1 Visual Feature

The 3D and 2D convolutional neural networks each have their own advantages when dealing with video contents. The 3D convolutional neural network takes into account the temporal features of the video, while the 2D convolutional neural network has a smaller number of parameters. We use a 3D convolutional neural network, SlowFast [5], to extract features from the video as Global-level features. We also use a ResNet-101 network [6] to extract features from the video frames. For each input video, we sample 8 frames evenly. These video frame features are fed into the GRU network [2] to solve the timing-independent problem, and the GRU network outputs the features as Temporal-aware level features. Afterwards, these features are fed into a 1D convolutional neural network with different convolutional kernel sizes 2,3,4,5 to sense visual features of different local sizes, and the output of the 1D convolutional neural network is used as the Local level features. We splice the Global, Temporal-aware, and Local level features as visual features.

3.2 Textual Feature

We used the Bert model [4] to extract the textual features of the video. For each text, we first perform a word separation operation and prefix each text with a [CLS] token. The features corresponding to the last layer of [CLS] token in Bert was used as features for the whole text. For videos with multiple texts, we average the features of multiple texts as the textual features corresponding to the video because of the similarity of these texts.

3.3 Audio Feature

We used the VGGish model [7] to extract audio features. First, we cut each video into segments without overlapping in 0.96s, and each

MediaEval'21, December 13-15 2021, Online

segment was fed into the VGGish network and a 128-D vector was generated. We fed this vector into an MLP structure and predicted the video memorability score of the segment. We take the median score of these segments as the audio stream video memorability prediction score for the video.

3.4 Cross-modal Interaction

With the visual and textual features already extracted above, we used the textual features to interact with the visual features. For the visual features extracted above, we first cut them into M=8 segments and mapped the visual features and textual features into the same semantic space. Afterwards, the mapped textual features and each segment of visual features were integrated to calculate the weight of each segment of visual features. We used this weight to weight and sum the cut M-segment visual features to obtain the interacted features. Through this interaction, the visual features exploited the semantic consistency with the textual features to enhance the expressiveness of their own features.

3.5 Score Fusion

We trained simple MLP networks using visual, textual, and audio features separately as regressors for predicting the video memorability scores of the respective modalities. MLP network is composed of several fully connected layers and non-linear activation functions. Afterwards, an adaptive weight assignment strategy was used to fuse the three scores. We varied the weights of each modality score in steps of 0.05, but ensured that the total weight sums to 1. In this way, we fused the three scores and predicted the final video memorability score.

4 RESULTS AND ANALYSIS

In this section, we describe specifically how we used the TRECVid and Memento10k dataset in our experiments and present the results in Table 1 and Table 2 below. And this is followed by a brief analysis of the results of the experiments.

Table 1 shows the experimental results of our method on TRECVid 2021. w/(dev) in the table means that the development set was used, while w/o(dev) means that the development set was not used. This is because the development set was not officially released at the beginning of the competition, so we only used the training set to train the model. When the development set was not used, we divided the training set of 590 videos into 479 as the training set and 111 as the validation set to train our model. When the development set was used, we considered it unreasonable to use only 590 videos as the training set and more than 1000 videos as the validation set, considering that the development set contains nearly 1000 videos. So we mixed the training set and development set together and divided the data set into training and validation sets at a ratio of 0.8/0.2. We believe that more data would be beneficial to the model. We were surprised to find that when training a short-term video memorability prediction model, the model without the development set achieved better performance, both in terms of raw and normalized scores, while when training a long-term video memorability prediction model, using the development set improved the performance significantly. Now we do not know the reason for this phonomenon. Additionally, in score fusion stage, visual feature occupies the greatest weight and textual feature is scondary to it.

Table 1: Results of our method on TRECVid2021 Dataset validation set and test set

Run	test set (RC)	validation set (RC)
short-term w/(dev)	0.113	0.330
short-term w/o (dev)	0.123	0.432
normalized short-term w/(dev)	0.106	0.296
normalized short-term w/o(dev)	0.132	0.462
long-term w/(dev)	0.11	0.331
long-term w/o(dev)	0.037	0.298

Table 2: Results of our method on Memento10k Dataset validation set and test set

Run	test set (RC)	validation set (RC)
short-term	0.628	0.642
normalized short-term	0.649	0.655

Table 2 shows the results of our method on the Memento10k dataset. When training with the Memento10k dataset, we trained our model using the officially published training/validation set partitioning method. We should also explain that we did not use audio features when training the Memento10k dataset, partly because some of the videos lack audio, and partly because in [16] the authors did not use audio features, so we did not use audio features either. Our model achieves better performance on the Memento10k dataset, and we speculate that the reason for this is that more data allows for better training of the model and mitigates the effects of overfitting.

5 DISCUSSION AND OUTLOOK

In this competition, we first extracted features from multiple modalities, then we used cross-modal interaction to enhance the representation of visual features, and finally we used late fusion to fuse the video memorability scores predicted by multiple modalities to obtain the final video memorability scores. In addition to this, we observed that optical flow was used to predict video memorability scores in multiple methods, which is one of our future research directions. However, as optical flow is time-consuming and labour-intensive, we did not use optical flow features in this experiment for the time being.

ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China (No. 61801441, No. 61701277,No. 61771288), National Key R & D plan of the 13th Five-Year plan (No. 2017YFC0821601), cross media intelligence special fund of Beijing National Research Center for Information Science and Technology (No. BNR2019TD01022), discipline construction project of "Beijing top-notch" discipline (Internet information of Communication University of China) and in part by the State Key Laboratory of Media Convergence and Communication, Communication University of China.

REFERENCES

[1] Erdem Akagunduz, Adrian G Bors, and Karla K Evans. 2019. Defining image memorability using the visual memory schema. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 9 (2019), 2165–2178.

[2] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*.

[3] Romain Cohendet, Karthik Yadati, Ngoc QK Duong, and Claire-Hélène Demarty. 2018. Annotating, understanding, and predicƒting long-term video memorability. In *Proc. 2018 ACM on International Conference on Multimedia Retrieval*. 178–186.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of Association for Computational Linguitics: Human Language Technologies*, Vol. 1. 4171–4186.

[5] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proc. IEEE/CVF international conference on computer vision*. 6202–6211.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. IEEE conference on computer vision and pattern recognition*. 770–778.

[7] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, and others. 2017. CNN architectures for large-scale audio classification. In *Proc. 2017 IEEE international conference on acoustics, speech and signal processing (icassp)*. IEEE, 131–135.

[8] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2011. What makes an image memorable?. In *Proc. IEEE CVPR 2011*. 145–152.

[9] Peiguang Jing, Yuting Su, Liqiang Nie, Huimin Gu, Jing Liu, and Meng Wang. 2018. A framework of joint low-rank and sparse regression for image memorability prediction. *IEEE Trans. Circuits Syst. Video Technol.* 29, 5 (2018), 1296–1309.

[10] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. 2015. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE international conference on computer vision*. 2390–2398.

[11] Rukiye Savran Kiziltepe, Mihai Gabriel Constantin, Claire-Hélène Demarty, Graham Healy, Camilo Fosco, Alba García Seco de Herrera, Sebastian Halder, Bogdan Ionescu, Ana Matran-Fernandez, Alan F. Smeaton, and Lorin Sweeney. 2021. Overview of The MediaEval 2021 Predicting Media Memorability Task. In *Working Notes Proceedings of the MediaEval 2021 Workshop*.

[12] Ricardo Kleinlein, Cristina Luna-Jiménez, Zoraida Callejas, and Fernando Fernández-Martínez. 2020. Predicting Media Memorability from a Multimodal Late Fusion of Self-Attention and LSTM Models. In *Working Notes Proceedings of the MediaEval 2020 Workshop (CEUR Workshop Proceedings)*.

[13] Phuc H Le-Khac, Ayush K Rai, Graham Healy, Alan F Smeaton, and Noel E O'Connor. 2020. Investigating Memorability of Dynamic Media. In *Working Notes Proceedings of the MediaEval 2020 Workshop (CEUR Workshop Proceedings)*.

[14] Roberto Leyva, Faiyaz Doctor, AG Seco de Herrera, and Sohail Sahab. 2019. Multimodal deep features fusion for video memorability prediction. In *Working Notes Proceedings of the MediaEval 2019 Workshop (CEUR Workshop Proceedings)*.

[15] Jonghwan Mun, Minsu Cho, and Bohyung Han. 2017. Text-guided attention model for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[16] Anelise Newman, Camilo Fosco, Vincent Casser, Allen Lee, Barry McNamara, and Aude Oliva. 2020. Multimodal memorability: Modeling effects of semantics and decay on video memorability. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*. Springer, 223–240.

[17] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proc. 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[18] Sumit Shekhar, Dhruv Singal, Harvineet Singh, Manav Kedia, and Akhil Shetty. 2017. Show and recall: Learning what makes videos memorable. In *Proc. IEEE International Conference on Computer Vision Workshops*. 2730–2739.

[19] Le-Vu Tran, Vinh-Loc Huynh, and Minh-Triet Tran. 2019. Predicting Media Memorability Using Deep Features with Attention and Recurrent Network.. In *Working Notes Proceedings of the MediaEval 2019 Workshop (CEUR Workshop Proceedings)*.

[20] Shuai Wang, Linli Yao, Jieting Chen, and Qin Jin. 2019. RUC at MediaEval 2019: Video Memorability Prediction Based on Visual Textual and Concept Related Features.. In *Working Notes Proceedings of the MediaEval 2019 Workshop (CEUR Workshop Proceedings)*.