HCMUS at MediaEval2021: Attention-based Hierarchical Fusion Network for Predicting Media Memorability

E-Ro Nguyen[1,3], Hai-Dang Huynh-Lam[1,3], Hai-Dang Nguyen[1,3], Minh-Triet Tran[1,2,3]
[1]University of Science, VNU-HCM, [2]John von Neumann Institute, VNU-HCM
[3]Vietnam National University, Ho Chi Minh city, Vietnam
{nero,nhdang}@selab.hcmus.edu.vn
hlhdang19@apcs.fitus.edu.vn,tmtriet@fit.hcmus.edu.vn

ABSTRACT

Predicting Media Memorability is a task offered by The Benchmarking Initiative for Multimedia Evaluation in the set of challenges for the MediaEval 2021 Workshop. This task aims at predicting the memorability of visual media to explore the possibility of automated supporting systems in multiple areas of application such as advertisement, recommendations, education, and more. To approximate the memorability score of media, we employ an attention-based fusion network with a hierarchical structure that resembles binary computation trees with the embedding of root nodes used to compute the final memorability score.

1 INTRODUCTION

The task of Predicting Media Memorability at MediaEval 2021 [8] requires participants to automatically predict the probability that a human may remember a specific visual media of type video after a specified time period. This task offered us two datasets for the training and evaluation of our methods, namely the Memento10k dataset [10] with short term memorability and the TRECVid dataset [2] with both short term and long term scores.

To aid readers in understanding our approach, we organize our paper as follow: Section 2 visits some prior works with concepts related to our approach that might help readers gain preliminary knowledge; Section 3 introduces the proposed architecture as well as elaborating details about our network; Section 4 provides detailed results of our runs together with multiple insights that guided us through our experiments; Section 5 discuss about the conclusion of our research and possible future approaches based on our method.

2 RELATED WORK

In Predicting Media Memorability task, participants need to approximate the probability of each video sample being memorized by human and hence, this task may be categorized as video regression with input being 4D features sampled from each video [5, 7]. Regression and classification on video has long been studied in academic literature [1, 13, 14] with many achievements recently when Transformer-based architecture of neural networks [12] being applied on this category [9, 11].

MediaEval'21, December 13-15 2021, Online

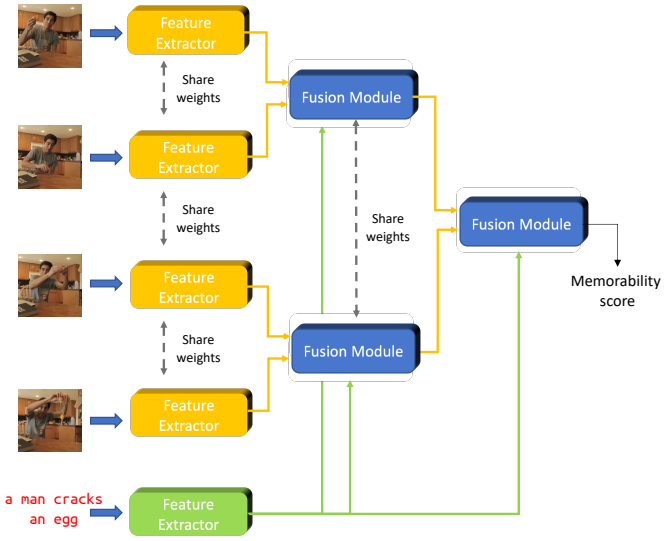

Figure 1: Overview of our proposed method AHFNet.

3 APPROACH

Figure 1 show an overview of our proposed method. The core of our method is the attention-based hierarchical fusion network (AHFNet). Its mechanism is to repeat the process of pairing and fusing two consecutive visual features into a high-level semantic feature by a fusion module according to the hierarchical structure as a binary tree from the leaf to the root node.

3.1 Fusion Module

We propose a *Fusion Module* to fuse two visual features into one based on the attention mechanism, given those two features are computed using similar method on different inputs. As illustrated in figure 2, for any two visual features $V_1, V_2 \in \mathbf{R}^{H \times W \times C}$, we first add a positional encoding $PE \in \mathbf{R}^{H \times W}$ to each of two features and employ a multi-headed self-attention [12] which is responsible for learning the association or correlation among the targets within each current frame:

$$SA(V) = MHA(V + PE, V + PE, V + PE) \quad (1)$$

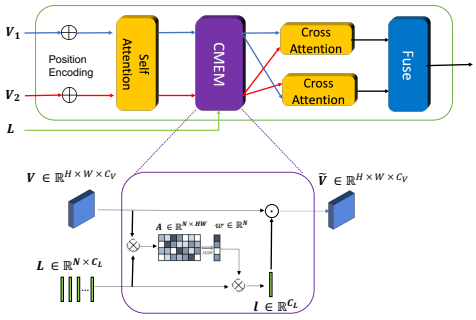

Figure 2: Illustration of our Fusion module.

where $V \in \{V_1, V_2\}$.

For consistency, $V_1 = f(\hat{V}_1)$ and $Y = f(\hat{V}_2)$ where $\hat{V}_1, \hat{V}_2 \in \mathbf{R}^{T_f \times H \times W \times C}$ are inputs while f is the computation acts on subset of V used to compute V_1 and V_2, respectively.

The cross-attention between two visual features are then taken before fusing both of them into a single feature, helping the network learn the relationship of those two:

$$CA(X, Y) = MHA(SA(X), SA(Y), SA(Y)) \tag{2}$$

where $X, Y \in \{V_1, V_2\}$.

The fusion operator is then applied to merge two visual features into single one:

$$V' = Fuse(CA(V_1, V_2), CA(V_2, V_1)) \tag{3}$$

where *Fuse* may be any reduction operator.

In our approach, we adopt the summation as Fusion operator:

$$V' = CA(V_1, V_2) + CA(V_2, V_1) \tag{4}$$

3.2 Hierarchical Fusion Network

We extracted 8 frames from each video, which were used for our image-based feature extraction. The ResNet-50 [6] (pre-trained on ImageNet) was used to extract a 2048-dimensional feature vector for each frame. And then, we make pair of the features of the frame (1, 2), (3, 4), (5, 6), (7, 8) and then fuse each pair with a fusion module to achieve a higher semantic feature from each pair. Then, the number of features is reduced by half. We continue doing the same process until the final feature is fused. The final feature has the video's high-level information that can now be used to predict the memorability score.

In the figure 1 we show only the short version with only 4 frames with 2-levels of fusion module. However, our work uses 8 frames with 3-levels of fusion module.

3.3 Cross-Modal With Text Captions

We use a pre-trained BERT [3] to extract the linguistic features of each video's text caption. These features are inserted into Fusion Module to highlight the visual features that are matched with corresponding linguistic clues by the CMEM module (Cross-Modal Excitation Modulation) [4]. The CMEM module is illustrated as a violet component in Figure 2.

Metrics	Caption	Short term Normalised	Short term Raw	Long term Raw
Spearman	No	0.06	0.066	0.013
(higher better)	Yes	**0.069**	**0.101**	**0.059**
Pearson	No	0.085	0.1	-0.023
(higher better)	Yes	**0.101**	**0.11**	**0.067**
MSE	No	**0.02**	**0.01**	**0.04**
(lower better)	Yes	**0.02**	**0.01**	0.06

Table 1: Results of our *Subtask 01* on TRECVid test set for Short Normalised, Short Raw and Long Raw memorability

Metrics	Caption	Short term Normalised	Short term Raw
Spearman	No	**0.508**	**0.516**
(higher better)	Yes	0.473	0.456
Pearson	No	**0.531**	**0.534**
(higher better)	Yes	0.476	0.461
MSE	No	**0.01**	**0.01**
(lower better)	Yes	**0.01**	**0.01**

Table 2: Results of our *Subtask 01* on Memento10k test set for Short Normalised, Short Raw memorability

4 RESULTS AND ANALYSIS

We have 2 different runs for each dataset (TRECVid, Memento10k) with each type of score (short raw, short normalised, long raw) in Subtask 01. The first run of each is the AHFNet without the text captions (AHFNetWTC), and the second is the full version of AHFNet. Table 1 and 2 show our results on the TRECVid and Memento10k, respectively.

With our experiments, we observe that the raw short term is almost better than the normalised one for both datasets. Our AHFNetWTC is better on the Memento10k test set. On the TRECVid test set, however, our AHFNet achieves higher results in all metrics and score types. These better results can be explained that our network extracts the TRECVid's text captions better than the Memento10k.

5 CONCLUSION

This paper describes a hierarchical fusion network with the attention-based proposed for the 2021 Predicting Media Memorability task of MediaEval. The main contributions of this paper are to propose a fusion module to capture the high-level semantics of two consecutive frames, leverage the binary hierarchical structure to fuse the video's features and highlight the visual features by the corresponding text caption.

In the future, we plan to conduct this task with additional features like audio in videos and a more robust feature extractor. So that can extract high-level features from dynamic videos effectively.

ACKNOWLEDGMENTS

This work was funded by Gia Lam Urban Development and Investment Company Limited, Vingroup and supported by Vingroup Innovation Foundation (VINIF) under project code VINIF.2019.DA19.

REFERENCES

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Apostol (Paul) Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. YouTube-8M: A Large-Scale Video Classification Benchmark. In *arXiv:1609.08675*. https://arxiv.org/pdf/1609.08675v1.pdf

[2] George Awad, Asad A. Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Jesse Zhang, Eliot Godard, Lukas Diduch, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, and Georges Quenot. 2020. TRECVID 2019: An Evaluation Campaign to Benchmark Video Activity Detection, Video Captioning and Matching, and Video Search Retrieval. (2020). arXiv:cs.CV/2009.09984

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[4] Zihan Ding, Tianrui Hui, Shaofei Huang, Si Liu, Xuan Luo, Junshi Huang, and Xiaoming Wei. 2021. Progressive Multimodal Interaction Network for Referring Video Object Segmentation. The 3rd Large-scale Video Object Segmentation Challenge, Workshop in conjunction with CVPR 2021 (virtual). (June 2021).

[5] Christoph Feichtenhofer. 2020. X3D: Expanding Architectures for Efficient Video Recognition. (2020). arXiv:cs.CV/2004.04730

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385* (2015).

[7] Hirokatsu Kataoka, Tenga Wakamiya, Kensho Hara, and Yutaka Satoh. 2020. Would Mega-scale Datasets Further Enhance Spatiotemporal 3D CNNs? *CoRR* abs/2004.04968 (2020). arXiv:2004.04968 https://arxiv.org/abs/2004.04968

[8] 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*.

[9] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. HERO: Hierarchical Encoder for Video+ Language Omni-representation Pre-training. In *EMNLP*.

[10] 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. (2020). arXiv:cs.CV/2009.02568

[11] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. VideoBERT: A Joint Model for Video and Language Representation Learning. (2019). arXiv:cs.CV/1904.01766

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[13] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2019. Temporal Segment Networks for Action Recognition in Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 11 (Nov 2019), 2740–2755. https://doi.org/10.1109/TPAMI.2018.2868668

[14] Chao-Yuan Wu, Ross B. Girshick, Kaiming He, Christoph Feichtenhofer, and Philipp Krahenbuhl. 2020. A Multigrid Method for Efficiently Training Video Models. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 150–159.