Modelling of Video Memorability using Ensemble Learning and Transformers

Muhammad Mustafa Ali Usmani[1,*,†], Sumaiyah Zahid[1,†] and Muhammad Atif Tahir[1,†]

[1]National University of Computer Emerging Sciences, (NUCES-FAST), Karachi Campus, Pakistan

Abstract

The modeling of video memorability is still an open challenge for researchers in machine learning. This paper presents our methodology for the MediaEval 2022: Predicting Media Memorability Challenge. The proposed approach investigated ensemble learning methods using the pre-extracted image features: Alexnet, Resnet, and Densenet. In addition to that Transformers and TF-IDF modeling were done on text features. Further, image and text features were ensembled using late fusion that helped in predicting the video memorability on the Memento10k dataset with an accuracy of 0.661.

1. Introduction

The expansion in social media demands for tools in social platforms that provide users' attention towards content creation. The content creators require their digital content to get remembered by the user interacting with a stream of other content. The modeling of human behavior towards predicting human-centric behavior involves mathematical and logical formulation of several approaches [1]. Human information perseverance toward media memorability is mostly dependent on an individual's experience and acceptance of degree patterns. This degree of pattern leads to the problem of video media memorability.

Video media memorability is the proportion of people that were able to remember watching a video on a second viewing while their memory is tested. This objective was provided by MediaEval 2022 in its consecutive 5th episode under the 'Predicting Media Memorability' challenge. The participants were provided with a Memento10k dataset and requested to provide an approach for modeling and predicting the degree of memorability for each individual video. The details of the challenge and the provided dataset are available at [2].

The paper further discusses the related advancements in Section 2 that provides the media memorability predictive methods. Section 3 explains the applied approach and methodology to predict media memorability and Section 4 discusses the results obtained, followed by the conclusion in Section 5.

2. Related Work

The study [3] investigated the effect and dependency of video image color, brightness, and hue on predicting media memorability, in comparison to complex data-driven representations such

MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2023, Bergen, Norway and Online

*Corresponding author.

†These authors contributed equally.

mustafaausmani@gmail.com (M. M. A. Usmani); sumaiyah@nu.edu.pk (S. Zahid); atif.tahir@nu.edu.pk (M. A. Tahir)

as image classification, composition, and recognition. It is observed that high-level representations (image composition, recognition, and classification) are best suited to predict media memorability.

The studies [4, 5] provide knowledge of the most commonly remembered media are war-like scenes, nature, and open spaces [6]. The brain's capability to remember any piece of media is highly dependent on the abstraction of the same level of scene and object representations. This helps in understanding the remembered media would include the cinematic and high-level object attributes. The best-suited models are the Transformers models providing better results [7, 8, 9] due in-depth representation of input features. These models are proposed as an alternative to the other neural architectures.

3. Approach

The models were applied to the textual features provided as the video description in the Memento10k dataset. Several machine learning techniques are applied based on text inputs as well as visual information. The video frames were modeled and only the best-performed models were used to predict the media memorability. The multi-modal approach is applied to encode textual and visual features. These embeddings are provided to linear regression models to predict the best media memorability.

The pre-extracted features are provided as image features from video frames of the beginning, middle and last of each video dataset. The machine learning model is applied where the top-scoring pre-extracted features are selected. The top resulting features include ALEX-NET, RES-NET, DENSE-NET, EFFICIENT-NET, and VGG. Further analysis is carried out by applying XG Boost, ADA Boost, Random Forest, KNN, and MLP (Multi-layer Perceptron) where ADA Boost outperformed and provide the best results. Figure 1 shows the complete understanding of the approach taken to accurately predict the media memorability.

3.1. Textual Features

The video caption or information helps in understanding the context of the video frame using semantic techniques. These features help in providing a baseline understanding of the video/image frame. TF-IDF[10] is a statistical formulation to understand the similarity of words in a document. This depends on the number of words and frequency of each word appearing in the same document. The TF-IDF is used to predict memorability as it provides the best among other techniques using textual information.

3.2. Text Transformers

Text Transformers provides another approach that helps in understanding the semantics of a video. There are different text transformers that are applied over the provided Memento10k dataset. The applied transformers help in getting the similarities and dissimilarities among the sentences which helped in identifying topics in text data. We got the best result using DistilBERT[11], so for further analysis and ensemble, we consider results from only DistilBERT Transformer. Since the text transformers are able to generate synthetic texts, therefore, this helped in automatic text generation. The extracted textual features are used to build language representation. The text transformers help in the classification of images when the image features and textual features are combined whereas the text encoders help to encode sentences to understand the context of the video.

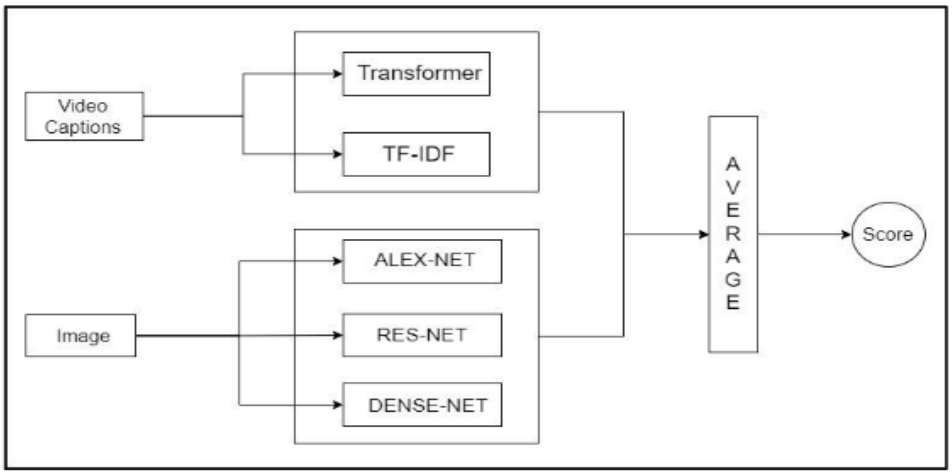

Figure 1: Architectural diagram of the proposed method.

3.3. Image Features

Videos are nothing but a continuous transition of images. For image-related features, 3 frames of every video were considered i.e. first, last and middle. Our approach consisted of learning the best pre-extracted feature that is already provided in the Memento10k dataset. After training on several machine learning models, only the top 3 image features were selected which were giving best accuracies on validation data. Namely, Alexnet[12], Resnet[13], and Densenet [14] outperformed the remaining image features. We ran multiple machine-learning models on these 3 image features, and Adaboost outperform all of them. Therefore for further analysis, we consider only the result of Adaboost.

3.4. Ensemble Learning

The ensemble combines more than one machine learning model that outperforms in predicting more accurately. This technique also provides a concept of a multi-classification model where the resulting factor to predict is based on combining multiple models. After running the base model on different features, late fusion was applied. Linear regression and average-out techniques were used in the ensemble step.

3.4.1. Runs Details

Run 1: Only image features were considered and then an average was taken of all 3 of them wto visualize the effect of only image features in predicting media memorability. Run 2: Average of Transformers and TF-IDF was taken to see if there was any effect of the image features as captions were the depiction of the same video. Run 3: Image features and TF-IDF scoring were considered and averag was taken. Run 4: All the best models were considered, from the image/video baseline Ada Boost was picked for all the features, and for text, Transformers and TF-IDF both were considerd. And then late fusion was done by taking the average of them, this gave us an improvement on the baseline methods. Run 5: Results from image features together with TF-IDF were passed to a Linear Regression Model.

4. Results and Analysis

The results elaborate on Spearman's correlation coefficient, Pearson's correlation coefficient, and Mean Square Error (MSE) values over the validation and testing set of data. The results in Table 1 show the obtained correlation against the validation and the results in Table 2 shows the obtained correlation on testing of the dataset of Memento10k for each submitted execution.

It is observed that Spearman's correlation coefficient started predicting with 0.421 accuracies and suddenly improved in the prediction accuracy in its second execution and till the fifth execution. The maximum accuracy gained from Spearmen's correlation coefficient is 0.661. On the other side, Pearson's correlation coefficient provides a little improved accuracy when compared with Spearman's correlation coefficient. Pearson's correlation coefficient obtained an accuracy of 0.439 on the test dataset improved quickly in next following executions. The maximum obtained accuracy from Pearson's correlation coefficient is 0.667. Furthermore, the MSE also reduced from 0.021 to 0.006 during several executions on the testing dataset.

Table 1
Media Memorability results on Validation dataset

Correlations	Run 1	Run 2	Run 3	Run 4	Run 5
Spearman	0.583	0.623	0.592	0.701	0.654
Pearson	0.599	0.624	0.601	0.718	0.663
MSE	0.018	0.006	0.005	0.005	0.004

Table 2
Media Memorability results on Test dataset

Correlations	Run 1	Run 2	Run 3	Run 4	Run 5
Spearman	0.421	0.611	0.547	0.661	0.619
Pearson	0.439	0.618	0.559	0.667	0.618
MSE	0.021	0.006	0.008	0.006	0.006

5. Conclusion

The approach of multi-classification through ensemble provides a better approach to predicting media memorability. The textual features and visual features predict well when modeled through ensemble learning. The textual features provide the semantic context while the visual features provide the understanding of each image through pre-extracted features from the Memento10k dataset. The accuracy to predict the media memorability increased during the testing dataset in its second execution by changing the hyperparameters of the machine learning models. TF-IDF is already a better approach to predict media memorability but our approach showed that TF-IDF performed way better while modeling through Linear Regression.

6. Acknowledgements

This work was supported in part by the Higher Education Commission (HEC) Pakistan, and in part by the Ministry of Planning Development and Reforms under the National Center in Big Data and Cloud Computing.

References

[1] R. Arnheim, Art and Visual Perception: A Psychology of the Creative Eye, University of California Press, 1974. URL: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0520243838.

[2] L. Sweeney, M. G. Constantin, C.-H. Demarty, C. Fosco, A. García Seco de Herrera, S. Halder, G. Healy, B. Ionescu, A. Matran-Fernandez, A. F. Smeaton, M. Sultana, Overview of the MediaEval 2022 predicting video memorability task, in: MediaEval Multimedia Benchmark Workshop Working Notes, 2023.

[3] P. Isola, J. Xiao, D. Parikh, A. Torralba, A. Oliva, What makes a photograph memorable?, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (2014) 1469–1482. doi:10.1109/TPAMI.2013.200.

[4] A. Jaegle, V. Mehrpour, Y. Mohsenzadeh, T. Meyer, A. Oliva, N. Rust, Population response magnitude variation in inferotemporal cortex predicts image memorability, eLife 8 (2019) e47596. URL: https://doi.org/10.7554/eLife.47596. doi:10.7554/eLife.47596.

[5] T. Konkle, T. F. Brady, G. A. Alvarez, A. Oliva, Conceptual distinctiveness supports detailed visual long-term memory for real-world objects, J Exp Psychol Gen 139 (2010) 558–578.

[6] T. Konkle, T. F. Brady, G. A. Alvarez, A. Oliva, Scene memory is more detailed than you think: the role of categories in visual long-term memory, Psychol Sci 21 (2010) 1551–1556.

[7] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, CoRR abs/2010.11929 (2020). URL: https://arxiv.org/abs/2010.11929. arXiv:2010.11929.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: http://proceedings.mlr.press/v139/radford21a.html.

[10] C. Sammut, G. I. Webb (Eds.), TF–IDF, Springer US, Boston, MA, 2010, pp. 986–987. URL: https://doi.org/10.1007/978-0-387-30164-8_832. doi:10.1007/978-0-387-30164-8_832.

[11] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter, CoRR abs/1910.01108 (2019). URL: http://arxiv.org/abs/1910.01108. arXiv:1910.01108.

[12] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: F. Pereira, C. Burges, L. Bottou, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, volume 25, Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[13] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, CoRR abs/1512.03385 (2015). URL: http://arxiv.org/abs/1512.03385. arXiv:1512.03385.

[14] G. Huang, Z. Liu, K. Q. Weinberger, Densely connected convolutional networks, CoRR abs/1608.06993 (2016). URL: http://arxiv.org/abs/1608.06993. arXiv:1608.06993.