# The Impact of Transformers Ensemble on Model Memorability and Generalizability

Muhammad Mustafa Ali Usmani[1], Humna Faisal[1] and Muhammad Atif Tahir[1]

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

Abstract

Video memorability, the extent to which a video is retained in human memory, is a crucial aspect in various multimedia applications such as advertising, education, and entertainment. This research explores the utilization of video features and text captions to predict and understand video memorability. Transformer models are employed to predict memorability scores, training is performed on the Memento10k dataset and the approach is tested on the VideoMem dataset. The highest Spearman co-efficient obtained is 0.337 on features from AlexNet with ensemble models also performing well.

## 1. Introduction

In the contemporary landscape of multimedia content creation and consumption, the concept of video memorability is a pivotal metric. It dictates the lasting impact of visual narratives on human cognition. The inherent ability of videos to resonate within our memory is a fundamental element not only in entertainment but also in influential sectors such as advertising and education. Understanding and predicting video memorability has emerged as critical research area, offering insights that can revolutionize content creation strategies and user engagement across diverse platforms.

For this Proc. of the MediaEval 2023 Workshop, Amsterdam, The Netherlands, 2024 challenge [1] we research the interplay between video features, textual cues, and the concept of video memorability. This study explores the fusion of visual and textual elements within videos. The primary focus lies in the integration of Transformer models, leveraging their power in handling sequential data and capturing intricate relationships within multimodal inputs. Through the utilization of Transformer architectures, this research aims to predict video memorability scores by encoding both video features and accompanying text captions.

The training phase used the Memento10k dataset [2], a rich repository spanning diverse video content. Validation and assessment are conducted on the VideoMem dataset [3], serving as a test for the efficacy and generalization capabilities of the proposed approach.

## 2. Approach

There has been a lot of ongoing research on the topic of video memorability. There are multiple focal points within this area. These include using raw and unprocessed videos for feature extraction and then using those features to predict memorability. In previous works Transformers

*MediaEval'23: Multimedia Evaluation Workshop, February 1–2, 2024, Amsterdam, The Netherlands and Online*

*Corresponding author.

†These authors contributed equally.

mustafaausmani@gmail.com (M. M. A. Usmani); humnaf29@gmail.com (H. Faisal); atif.tahir@nu.edu.pk (M. A. Tahir)

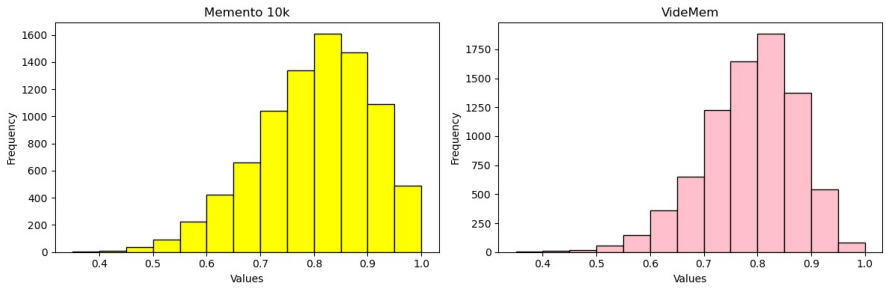


**Figure 1:** A comparison of memorability score distributions across the two data sets

have been used as a method of extracting features from videos that are then used to predict memorability [4] [5].

There is also a lot of work being done on using existing feature to develop models. A previous study [6] has used different video features with different models to predict memorability. We extend this work and use the video features that had the best results to create new models.

Using captions to predict video memorability is also an ongoing research topic. Ensemble models that combine text and video features are becoming increasingly common and give good results. Transformers and TF-IDF have been previously used on text features as a part of an ensemble model [7].

Transformers are a top choice for video memorability tasks owing to their proficiency in processing sequential data and integrating diverse types of information. Videos consist of sequential frames, and Transformers excel in handling this sequential nature, effectively capturing temporal relationships within the video content. Their attention mechanisms enable focused processing on specific aspects of the video, facilitating the identification of crucial visual cues or textual information influencing memorability.

## 2.1. Video Features - VidFormer

Primarily features extracted from AlexNet [8] , DenseNet [9] and ResNet [10] were directly used to train the Transformers. These three features were the top choice as they got the best results in previous such works [6]. The model architecture has two sequential instances of the PositionalEmbedding layer. This layer generates positional embeddings for the input sequences, crucial for the Transformer-based architecture to understand the sequence's order. The TransformerEncoder layer is then applied, utilizing the Transformer architecture to process the sequence data, capturing intricate patterns and dependencies. After this, the GlobalAveragePooling1D layer aggregates the Transformer's output across the temporal dimension, summarizing the learned features. To prevent overfitting, a Dropout layer is incorporated, serving as a regularization technique. Finally, there is an output layer composed of a Dense layer utilizing a linear activation function. Captions and annotations associated with each video were used for this approach. DistilBERT [11] architecture was used to train a model using these text features. Furthermore, as an extension to this approach the text features were also augmented using BERT [12], seven thousand captions from the Memento 10k dataset were augmented. The same Transformer model was again trained on the augmented features. This was done to achieve better generalizability with the text features.

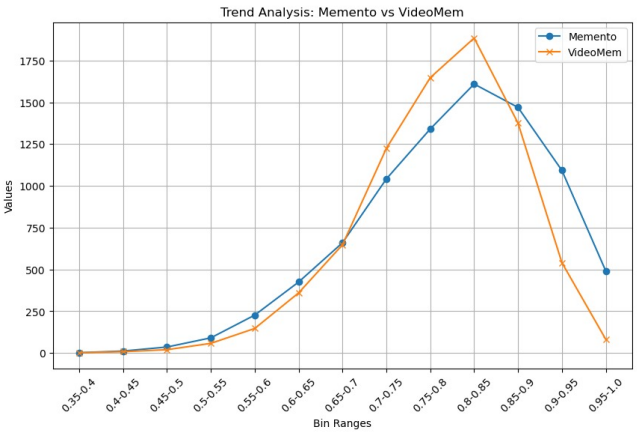


**Figure 2:** A comparison of memorability score trends across the two data sets

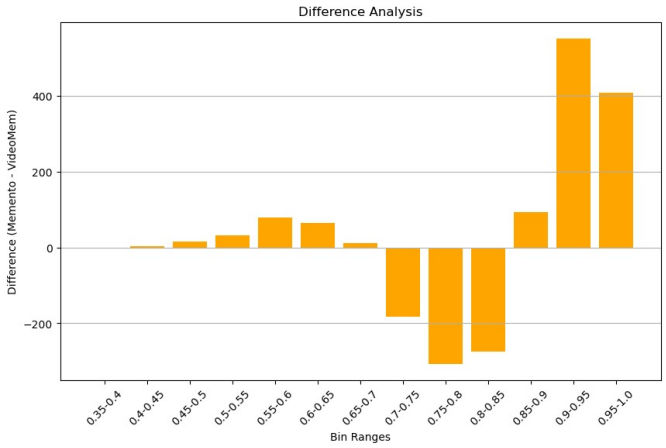


**Figure 3:** Difference of number of data points between the two data sets

### 2.2. Ensemble Approach - MultiFormer

In addition to both the above described models an ensemble approach was also used. A late fusion technique was used in which the predictions from the video and text models were combined using weighted average. This gave us an edge over just using one feature and increased generalizability of the text models.

## 3. Results and Analysis

Results show that the VidFormer with the AlexNet video features performed the best on the VideoMem test set with Spearman co-efficient of 0.377 even though it wasn't the best performing model on Memento 10k. The MultiFormer ensemble approach that was a mix of both the text and video features reached 0.68 while training on Memento 10k dataset but came in second for the VideoMem dataset. The TexFormer model performed the best on the Memento 10k dataset achieving 0.7 but was not able to perform at all on the VideoMem dataset. These results are summarized in Table 1.

The models were trained on the Memento 10k data sets and it was observed that they were not able to generalize as well on the VideoMem data set. According to our analysis a major

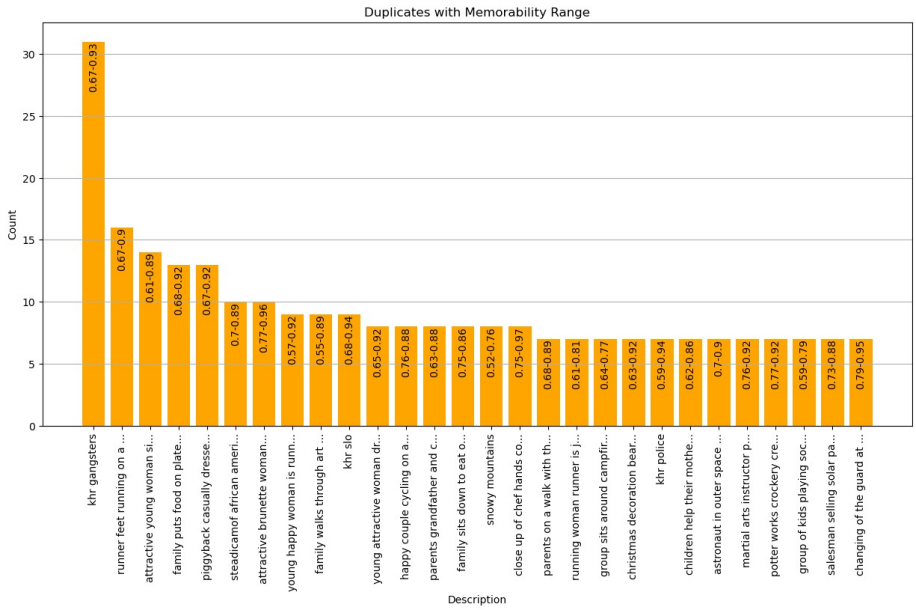


**Figure 4:** Duplicates in the VideoMem Dataset

**Table 1**
Summary of Spearman's Co-efficient across different models

| Model | Memento 10k Train | Memento 10k Test | VideoMem Test |
|---|---|---|---|
| VidFormer | 0.55 | 0.52 | 0.337 |
| MultiFormer | 0.68 | 0.66 | 0.234 |
| TexFormer | 0.7 | 0.68 | -0.008 |

reason for this is the difference in distributions in the two data sets. These are further illustrated in Figure 1, Figure 2 and Figure 3. Each bin represents different ranges for the memorability scores. The difference between Memento 10k and VideoMem fluctuates significantly across the bin ranges. Initially, the differences are smaller, but they increase notably as the bin ranges widen towards the higher end (0.7-1.0). There's a substantial divergence between Memento 10k and VideoMem values, especially in the latter bin ranges.

It was observed in the VideoMem data set there were some videos that had the same caption but different memorability scores and different videos ids, there were also instances with no captions, illustrated in Figure 4. Whereas, in Memento 10k there were only three such occurrences. In total there were 1860 such data points, this greatly skewed our accuracy measures. From this we infer that solely using text features for prediction is not sufficient and will not give the desired results.

## 4. Discussion and Outlook

From our results we can conclude that the simply using text data doesn't generalize well for memorability tasks. Using ensemble models is a much a stable and robust approach. However, a model that performs well on one dataset may not perform as well on another dataset this is owed to the difference in data distribution across both these datasets.

## References


[1] M. G. Constantin, C.-H. Demarty, C. Fosco, A. G. S. de Herrera, S. Halder, G. Healy, B. Ionescu, A. Matran-Fernandez, R. S. Kiziltepe, A. F. Smeaton, L. Sweeney, Overview of the mediaeval 2023 predicting video memorability task, in: MediaEval Multimedia Benchmark Workshop Working Notes, 2024.

[2] A. Newman, C. Fosco, V. Casser, A. Lee, B. McNamara, A. Oliva, Multimodal memorability: Modeling effects of semantics and decay on video memorability, 2020. arXiv:2009.02568.

[3] R. Cohendet, C.-H. Demarty, N. Q. Duong, M. Engilberge, Videomem: Constructing, analyzing, predicting short-term and long-term video memorability, in: Proceedings of the International Conference on Computer Vision, Seoul, 2019.

[4] M. G. Constantin, B. Ionescu, Using vision transformers and memorable moments for the prediction of video memorability, in: Working Notes Proceedings of the MediaEval 2021 Workshop, 2021.

[5] R. Kleinlein, C. Luna-Jiménez, F. Fernández-Martínez, Thau-upm at mediaeval 2021: From video semantics to memorability using pretrained transformers, in: Working Notes Proceedings of the MediaEval 2021 Workshop, 2021.

[6] M. M. A. Usmani, S. Zahid, M. A. Tahir, Quest for insight: Predicting memorability based on frequency of n-grams, in: Working Notes Proceedings of the MediaEval 2022 Workshop, 2023.

[7] M. M. A. Usmani, S. Zahid, M. A. Tahir, Modelling of video memorability using ensemble learning and transformers, in: Working Notes Proceedings of the MediaEval 2022 Workshop, 2023.

[8] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems 25 (2012) 1097–1105.

[9] G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger, Densely connected convolutional networks, 2018. arXiv:1608.06993.

[10] 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.

[11] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. arXiv:1910.01108.

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.