Memorability: Predicting movie and commercial memorability using visual and audio features

Prabavathy Balasundaram[1], Dhanush Parthasarathy[1], Tuhina Shaw[1], Nunna Jahnavi[1] and Praveen M[1]

[1]*Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, Rajiv Gandhi Salai, Chennai, Tamil Nadu, India*

Abstract

Predicting video memorability is crucial for creating engaging media and understanding human cognition. Using datasets like VIDEM and movie memorability benchmarks, this study uses a **Multimodal Attention Fusion Network** that integrates visual embeddings (EfficientNetB3, ResNet50, ViT, R3D), color and texture descriptors (LBP, HSV, RGB), audio features, and ASR text embeddings. An attention module learns modality-specific importance, enhancing generalization across diverse content. Evaluated using RMSE and R^2, the model shows that combining visual, audio, and textual cues improves long-term memorability prediction. This multimodal approach offers insights for cinematic and commercial applications, optimizing content creation and audience engagement.

1. Introduction

Memorability is a fundamental aspect of human cognition that determines how well visual and auditory experiences are retained in memory. In today's visual media landscape, predicting the memorability of the movies and commercials is increasingly important. With audiences inundated with content, understanding memorability contributes to the creation of media that exhibits greater effectiveness and audience engagement. Beyond marketing, it benefits education, information delivery, and human-computer interaction by enhancing learning, engagement, and retention. Datasets like VIDEM and movie memorability benchmarks enable study of short term and long term recall, bridging human cognition and computational modeling. By incorporating multimodal cues and brand-level challenges, this study deepens our understanding of how sensory and semantic factors shape memory.

In this paper, Multimodal Attention Fusion Network is used to predict the long-term memorability of videos, combining visual embeddings using EfficientNetB3, ResNet50, ViT, R3D, color histograms and texture descriptors such as LBP, HSV, RGB histograms, audio features, and Automatic Speech Recognition text embeddings. The attention module learns modality-specific importance, enhancing generalization across varied video content.

Performance was evaluated using RMSE and R^2, showing that textual and metadata information complement visual cues for memorability estimation.

2. Related Work

Video memorability has been an active area of research since the introduction of the image memorability concept by Isola et al. [1]. The first large-scale video memorability benchmark, VideoMem, was introduced by Cohendet et al. [2], proposing the construction of both short-term and long-term memorability datasets. Subsequent work within the *MediaEval Predicting Media Memorability Tasks* [3, 4] extended this concept to diverse video domains, emphasizing multimodal prediction based on visual, auditory, and semantic cues.

MediaEval'25: Multimedia Evaluation Workshop, October 2025, Dublin, Ireland

prabavathyb@ssn.edu.in (P. Balasundaram); dhanush2560001@ssn.edu.in (D. Parthasarathy); tuhina2520051@ssn.edu.in (T. Shaw); nunna2520055@ssn.edu.in (N. Jahnavi); praveen2520032@ssn.edu.in (P. M)

Recent research has highlighted the potential of transformer-based architectures and attention mechanisms for spatio-temporal modeling of memorability [5]. These approaches show that contextual and cross-modal relationships strongly influence how viewers retain visual content. The newly introduced *Video Effectiveness and Memorability (VIDEM)* dataset [6] further expands this research to include brand memorability and commercial video effectiveness.

Our work builds on these foundations and innovates by integrating multimodal attention fusion with audio and automatic speech recognition (ASR) embeddings for **Challenge 1.1**, while extending to Ridge and Random Forest regressors for **Challenges 2.1** and **2.2** [6]. Unlike prior methods that concatenate features or rely on unimodal representations, our system dynamically weighs each modality, offering improved interpretability and generalization across cinematic and commercial video domains.

3. Methodology

This section outlines the methodological framework adopted for predicting both video and brand memorability across Challenge 1.1, 2.1, and 2.2. The approach employs a multimodal deep learning pipeline integrating visual, auditory, and textual modalities. Three model architectures were explored: a Multimodal Attention Fusion Network (MAFN), Ridge Regression, and Random Forest Regressor. The goal was to combine interpretability, scalability, and multimodal representational learning within a unified system.

3.1. Multimodal Attention Fusion Network

The Multimodal Attention Fusion Network (MAFN) is designed to capture interdependencies among heterogeneous modalities by learning attention-based fusion weights. Each modality — visual, audio, and Automatic Speech Recognition (ASR) — is first encoded as a fixed-length feature vector.

Feature Encoding. Visual features are extracted using deep convolutional backbones such as EfficientNetB3, ResNet50, ViT, DenseNet121, and R3D [2]. Audio features are derived using MFCC, chroma, and spectral contrast representations from Librosa, while ASR textual features are embedded using Sentence-BERT [7] applied on transcripts obtained via Whisper ASR [8].

Attention-based Fusion. Each feature vector is linearly projected into a shared embedding space of dimension $d = 512$ using fully connected layers. The projected embeddings are then concatenated and passed through a multi-head self-attention layer [9] with $h = 8$ heads. This layer learns modality-specific attention weights that capture contextual relevance between modalities, allowing the network to emphasize modalities most predictive of memorability. The attention outputs are aggregated via a residual connection followed by layer normalization and fed to a two-layer feed-forward regressor.

Optimization. The model is trained using Mean Squared Error (MSE) loss with the Adam optimizer [10]. Early stopping is employed to prevent overfitting, and dropout (rate = 0.3) is applied between layers for regularization. A train-validation split (80/20) is used, and final evaluation metrics include MSE, R^2, and Spearman's rank correlation, consistent with prior MediaEval benchmarks [11, 12].

3.2. Ridge Regression

Ridge Regression [13] serves as a strong linear baseline, particularly suitable for high-dimensional multimodal features. It minimizes the following objective:

$$L(\beta) = \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

where λ is the regularization strength controlling the penalty on large coefficients. Ridge Regression provides interpretability by quantifying each feature's contribution to memorability, while mitigating overfitting in correlated feature spaces. It is implemented using Scikit-learn with hyperparameters tuned via grid search.

3.3. Random Forest Regressor

The Random Forest Regressor [14] is employed as a non-linear ensemble learning method that aggregates predictions from multiple decision trees trained on bootstrap samples. Each tree considers a random subset of features at each split, ensuring decorrelation among trees. The final prediction is the average of individual tree outputs. Random Forests have shown competitive performance in memorability prediction tasks [2, 12], owing to their robustness to noisy multimodal inputs and lack of need for normalization. Key hyperparameters include 500 estimators, a maximum depth of 15, and minimum samples per leaf of 2.

3.4. Multimodal Fusion Strategy

The overall fusion strategy follows a hierarchical late-fusion paradigm. Each modality is first processed independently and projected into a uniform latent dimension. Their embeddings are then fused either through attention-weighted averaging (in MAFN) or concatenation (in baseline models). This structure ensures that modality-specific information is preserved while still enabling joint reasoning across modalities.

3.5. Evaluation Protocol

The dataset is split into training, validation, and test sets following the MediaEval VIDEM 2025 structure. The model is first trained and validated on the dev set, followed by retraining on the full training data before inference on the hidden test set. Metrics include Mean Squared Error (MSE), Coefficient of Determination (R^2), and Spearman's rank correlation, in line with previous MediaEval memorability evaluation practices [11, 2, 12].

4. Result and Analysis

Table 1 summarizes the quantitative results obtained across the three subtasks: video memorability (Challenge 1.1), commercial video memorability (Challenge 2.1), and brand memorability (Challenge 2.2). All models were evaluated using *Spearman's rank correlation*, *Pearson's correlation*, and *Mean Squared Error (MSE)* between predicted and ground-truth memorability scores.

Challenge	Spearman	Pearson	MSE
Movie Memorability	0.257	0.261	0.068
Commercial Memorability	0.130	0.188	0.027
Brand Memorability	0.092	0.048	0.040

Table 1. Performance metrics for all submitted runs.

4.1. Quantitative Analysis

The multimodal attention fusion model for **Challenge 1.1** achieved the best overall performance, with a Spearman correlation of 0.257 and an MSE of 0.068. These results indicate that integrating visual, auditory, and ASR features improves the model's ability to capture human memorability judgments. The relatively high Pearson correlation further supports the linear consistency between predicted and true memorability scores.

For **Challenge 2.1**, which focuses on long-term memorability of commercial videos, the Ridge Regression model yielded a moderate Spearman correlation of 0.13 with a low MSE of 0.027. Although the correlation was weaker than in Challenge 1.1, the reduced error suggests the model could effectively predict general memorability trends within the structured VIDEM dataset.

In **Challenge 2.2**, the Random Forest model obtained a Spearman correlation of 0.092 and MSE of 0.040. The results show that brand-level memorability is harder to model due to the abstract and conceptual nature of branding, which depends not only on visual appeal but also on symbolic and emotional associations.

4.2. Qualitative and Failure Analysis

Qualitatively, the attention-based fusion model captured strong correlations between emotionally salient audio-visual cues and memorability. Videos containing human faces, bright color palettes, and consistent rhythm in background music were often predicted as more memorable, aligning with known cognitive patterns.

Failure analysis revealed three main challenges: (1) **Music-only clips** often lacked linguistic or semantic cues, limiting ASR and text-based representations. (2) **Visually subtle scenes** with low motion or muted tones tended to produce under-confident predictions, suggesting the model's sensitivity to visual dynamics. (3) **Brand memorability** errors frequently arose from advertisements emphasizing narrative or humor over explicit brand exposure, which the current multimodal representation struggled to quantify.

5. Discussion and Outlook

Participation in the MediaEval 2025 memorability challenges offered key insights into multimodal memorability prediction. The attention-based fusion model showed that integrating visual, audio, and ASR features enhances recall estimation, yet certain modalities contributed unevenly—ASR was useful in dialogue-rich scenes but ineffective for music-only clips, and handcrafted features added little beyond deep embeddings. These findings underline the need for adaptive modality selection.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] P. Isola, J. Xiao, A. Torralba, A. Oliva, What makes an image memorable?, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 145–152. URL: https://web.mit.edu/phillipi/www/publications/WhatMakesAnImageMemorable.pdf.

[2] R. Cohendet, C.-H. Demarty, N. L. Duong, A. Goude, Videomem: Constructing, analyzing, predicting short-term and long-term video memorability, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2019, pp. 2531–2540. URL: https://openaccess.thecvf.com/content_ICCV_2019/papers/Cohendet_VideoMem_Constructing_Analyzing_Predicting_Short-Term_and_Long-Term_Video_Memorability_ICCV_2019_paper.pdf.

[3] A. G. De Herrera, R. Cohendet, C.-H. Demarty, A. Goude, R. Lienhart, Overview of the mediaeval 2020 predicting media memorability task, in: Working Notes Proceedings of the MediaEval Workshop, 2020. URL: https://arxiv.org/abs/2012.15650.

[4] C. Sweeney, R. Cohendet, C.-H. Demarty, A. Goude, Overview of the mediaeval 2022 predicting video memorability task, in: Working Notes Proceedings of the MediaEval Workshop, 2022. URL: https://ceur-ws.org/Vol-3583/paper17.pdf.

[5] S. Kiziltepe, S. Aksoy, C.-H. Demarty, An annotated video dataset for computing video memorability, arXiv preprint arXiv:2112.02303 (2021). URL: https://arxiv.org/abs/2112.02303.

[6] I. Martín-Fernández, M. G. Constantin, C.-H. Demarty, M. Gil-Martín, S. Halder, B. Ionescu, A. Matran-Fernandez, R. Savran Kiziltepe, A. García Seco de Herrera, Overview of the mediaeval 2025 predicting movie and commercial memorability task, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025.

[7] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019). URL: https://arxiv.org/abs/1908.10084.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, arXiv preprint arXiv:2212.04356 (2023). URL: https://arxiv.org/abs/2212.04356.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems (NeurIPS) 30 (2017). URL: https://arxiv.org/abs/1706.03762.

[10] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014). URL: https://arxiv.org/abs/1412.6980.

[11] A. G. De Herrera, C.-H. Demarty, et al., Overview of the mediaeval 2020 predicting media memorability task, arXiv preprint arXiv:2012.15650 (2020). URL: https://arxiv.org/abs/2012.15650.

[12] C. Sweeney, C.-H. Demarty, et al., Overview of the mediaeval 2022 predicting video memorability task, in: MediaEval Workshop Proceedings, 2022. URL: https://ceur-ws.org/Vol-3583/paper17.pdf.

[13] A. E. Hoerl, R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics 12 (1970) 55–67.

[14] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.