


Article

Adaptive Multi-Modal Ensemble Network for Video Memorability Prediction

Jing Li ¹, Xin Guo ², Fumei Yue ^{3,*}, Fanfu Xue ³ and Jiande Sun ^{3,*} ¹ School of Journalism and Communication, Shandong Normal University, Jinan 250061, China² Shandong Haiyi Digital Technology Co., Ltd., Zibo 256410, China³ School of Information Science and Engineering, Shandong Normal University, Jinan 250061, China

* Correspondence: fumei.yue@hotmail.com (F.Y.); jiandesun@hotmail.com (J.S.)

Abstract: Video memorability prediction aims to quantify the credibility of being remembered according to the video content, which provides significant value in advertising design, social media recommendation, and other applications. However, the main attributes that affect the memorability prediction have not been determined so that making the design of the prediction model more challenging. Therefore, in this study, we analyze and experimentally verify how to select the most impact factors to predict video memorability. Furthermore, we design a new framework, Adaptive Multi-modal Ensemble Network, based on the chosen vital impact factors to predict video memorability efficiently. Specifically, we first conduct three main impact factors that affect video memorability, i.e., temporal 3D information, spatial information and semantics derived from video, image and caption, respectively. Then, the Adaptive Multi-modal Ensemble Network integrates the three individual base learners (i.e., ResNet3D, Deep Random Forest and Multi-Layer Perception) into a weighted ensemble framework to score the video memorability. In addition, we also design an adaptive learning strategy to update the weights based on the importance of memorability, which is predicted by the base learners rather than assigning weights manually. Finally, the experiments on the public VideoMem dataset demonstrate that the proposed method provides competitive results and high efficiency for video memorability prediction.



Citation: Li, J.; Guo, X.; Yue, F.; Xue, F.; Sun, J. Adaptive Multi-Modal Ensemble Network for Video Memorability Prediction. *Appl. Sci.* **2022**, *12*, 8599. <https://doi.org/10.3390/app12178599>

Academic Editor: Andrea Prati

Received: 26 June 2022

Accepted: 20 August 2022

Published: 27 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: multi-modal; video memorability; ensemble learning

1. Introduction

Human memorability can be regarded as the criterion for judging whether the content of multi-media can be remembered or not [1]. In current social media, humans are exposed to continual and vast multi-media information, such as images, videos, audio and text. Human memorability has different responses to different multi-media contents, some of which are stuck in our memory while others are forgotten easily. For example, if the content appeals to the observers, such as their favorite celebrities or shocking natural scenes, the media information can be remembered easily. Therefore, research on media information memorability can contribute to advertising, intelligent recommendation, and so many other applications.

Media memorability aims to score the probability of being remembered according to the media content. Intuitively, media memorability is affected by subjective factors, but some studies have proved that memorability is the inherent attribute of media information [2]. For example, Isola et al. [2] demonstrate that image memorability is a stable intrinsic property. The image memorability (IM) score is defined as the percentage of correct detection by different observers. Based on observation, many works focus on exploiting the factors affecting image memorability and designing models for predicting image memorability. More recently, the amount of videos on social media, such as Tik-Tok, Youtube, etc., has been growing exponentially. Therefore, video memorability is nascent as a new research field. Compared with images, videos are dynamic and contain temporal information.

Therefore, the study of video memory is different from that of images. Kar et al. [3] began to wonder what makes a video memorable and propose that the video frames can be used to extract features for predicting video memorability. Similarly, the video memorability (VM) score is defined as the percentage of correct detection by different observers.

However, the drawback of current research on video memorability is that the main attributes affecting memorability prediction have not been determined, making the prediction model's design more challenging. This paper analyzes and experimentally verifies the critical impact factors in video memorability and proposes a framework for efficiently predicting video memorability. Firstly, visual and semantic factors play a decisive role in video memorability prediction. The combination of vision and semantics has produced a powerful effect on human memory. Secondly, multi-source feature fusion usually yields better results than isolated modeling generally. However, merging multiple features of the same information source adds complexity and causes heterogeneous errors [4]. Thirdly, it has been studied that the memory interval affects the video memorability [5]. Thus, we consider the memory interval by taking the memory after memorizing a few minutes as the short-term memory and using the long-term memory to define the memory performance after 24–72 h. Choosing different media information sources can predict the change in video memorability from different angles. Finally, we choose temporal 3D information, spatial information, and semantics derived from the three media information, including videos, images and captions, respectively.

Considering the redundancy effect of feature fusion, we use different base individual learners for each media information source to predict memorability scores. Concretely, we train the Resnet3D network [6] model to predict video memorability scores. For the selection of the spatial information, based on the previous experimental results [4], Local Binary Patterns (LBP) [7] features are more favorable for predicting video memorability scores than RGB [8] and High Osmolarity Glycerol (HOG) [9]. As for the text features, the semantics are phrases or sentences sufficient to describe the video scene, and we extract features from semantics as input to the semantic modal model. Then, we design a new framework, the Adaptive Multi-modal Ensemble Network (AMEN), to predict video memorability efficiently. Specifically, we integrate the memorability scores of different models using the weighted method and then use the stochastic gradient descent algorithm to obtain the best prediction results. In addition, the framework updates the weights of each learner adaptively instead of assigning weights manually. In general, the ensemble learning [10] appliance has a more vital generalization ability than the base learning appliance. Compared with feature fusion, the weighted ensemble learning method can fully guarantee the independent information of each mode and reduce the error and redundancy caused by feature fusion. As mentioned above, people's memory for a certain video decreases with time, so we predict the video memorability scores for both short-term and long-term memory, that is, how well humans remember a video a few hours after watching it and how well humans remember the same video a few days later. Our experiments are conducted in both short-term and long-term memory.

The main contributions of our method are summarized as follows:

- We analyze and experimentally verify how to select impact factors to predict video memorability and conduct three main factors that affect video memorability, i.e., temporal 3D information, spatial information and semantics derived from video, image and caption, respectively.
- We propose a new adaptive multi-modal ensemble network (AMEN) for video memorability according to the selected impact factors. It eliminates the error caused by the heterogeneous gap via integrating the optimal base learners corresponding to each media source instead of fusing the heterogeneous modalities directly.
- We design an adaptive learning strategy to update the weights based on the importance of memorability which is predicted by the base learners and contributes to obtaining the best performance without any manual tuning.

The rest of our paper is organized as follows. Related works about media memorability are summarized in Section 2. Our method is described in Section 3. Experimental results are presented in Section 4. The conclusion is made in Section 5.

2. Related Works

In this section, we summarize the previous works on media memorability and related concepts of media memorability. The work of Isola et al. [2] led to a pioneering study of media memorability. To measure image memorability, Isola et al. first proposed the memorability score of each image by the memory game. Based on the visual memory game, Isola et al. [2] proved that image memorability is the intrinsic property of each image and showed that the image memorability score has sufficient consistency of each image across various viewers by the consistency analysis. With the development of research on image memorability, more researchers have devoted themselves to studying what makes the image memorable and how to utilize machine learning algorithms to predict image memorability scores. Isola et al. [11] further explored the fact that image features, attributes and labels have a positive effect on the memorability of an image and use the support vector regression algorithm to predict image memorability scores. Intrinsic and extrinsic properties [12] that make media information memorable have been studied in recent years. Moreover, objects, emotions, saliency, and aesthetics contribute to making an image memorable [13–16]. It has to be mentioned that Dubey et al. [17] specifically discussed what makes an object memorable. To better understand what makes an image forgettable or memorable, Basavaraju et al. [14] were committed to studying the role of depth and motion and showed that depth and motion are helpful. Although many properties have been mined related to the memorability of the image, researchers only use traditional and simple image features, such as Pixels, GIST [2,11,18], to predict image memorability. In other words, much of the image information has not been mined and utilized.

Contrast to that, scholars attempted to extract deep features from deep learning algorithms to predict image memorability with the deep learning algorithm becoming extremely popular in the 2010s [19,20]. Khosla et al. [21] first used fine-tuned Hybrid Convolutional Neural Networks (CNNs) [22] to extract deep features, and the performance outperformed all other features at that time. Then, Zarezadeh et al. [23] used three common convolutional networks types to derive deep features from predicting image memorability and they drew a conclusion that deep features outperformed traditional features which are universally used, such as the SIFT, SSIM, and HOG_{2×2}. Due to the application of deep learning, Squalli-Houssaini et al. [24] presented their computation model, which is based on a deep learning frame to predict memorability scores while support vector regression is used in previous studies. When predicting image memorability, several different datasets and ground truth have been constructed by various scholars. Isola et al. [2] established the Scene UNderstanding (SUN) memorability dataset and the Large-scale Memorability (LaMem) dataset was built by Khosla et al. [21]. SUN memorability and the LaMem dataset are also the most used in further research. The point of adaptive semi-supervised feature selection [25] is to attempt to apply memorability. Akagunduz et al. [18] proposed the concept of Visual Memory Schema (VMS) and built the VISHEMA image set. Based on these datasets, more research on image memorability will be carried out in the future.

As we have mentioned before, different media leave different impressions on people, some of which can be remembered, while others are ignored. Getting inspired by the memorability of the image, video memorability is studied gradually, and video memorability scores are defined as the percentage of correct detection of each video by different participants [3]. The research on image memorability can be said to be the cornerstone of other media memorability research. In the paper by Goswami et al. [26], they contributed to face memorability. WuLin Wang et al. [27] raised a video hashing method based on memorability features. A system for memorability estimation was proposed by Han et al. [28], which predicts the memorability scores of a video clip by learning from brain functional magnetic resonance imaging (fMRI). Influenced by these studies, Kar et al. [3] began to

wonder what makes a video memorable. Kar et al. [3] proposed that the video frames can be used to extract features for predicting video memorability. Their studies have shown that saliency, color, scene complexity, background simplicity, object occurrence, and object attributes are related to the memorability scores. In the same year, Shekhar et al. [29] focused on analyzing which fusing features could more accurately predict memorability scores. However, in previous experiments measuring video memorability scores, the data volume of the annotated dataset is insufficient. Cohendet et al. [5] constructed a large-scale dataset named VideoMem, and the composition of VideoMem is 10,000 videos with corresponding video memorability scores. It is worth noting that VideoMem includes short-term human annotations and long-term annotations since the memory of people changed over time. Cohendet et al. proved that video semantics information is beneficial for predicting video memorability scores.

At the same time, Awad et al. [30] attempted to enrich the memorability annotations of the dataset TRECVID 2019. They provided partial data with memorability annotations to carry out the memorability prediction task in MediaEval 2020 [31]. A great many novel ideas have been conceived at previous MediaEval conferences [32–36]. Such as MediaEval 2018, many researchers have put forward their ideas and carried out experiments to verify them. Lryva et al. [37] used CNNs to extract video, text, and image features. Then, features were fused to obtain a vector as global features to predict video memorability. Lryva et al. [38] proposed a video memorability prediction framework based on late fusion of text, visual and motion features. Kleinlein et al. [39] proved that text features are effective in the representation of visual semantics required for the video memorability prediction model. Ali et al. [40] propose a novel framework to fuse the text, visual and motion features to predict video memorability. In recent years, with the development of ensemble learning and cross-modal [41–45], studies on ensemble learning and multi-modal are applied in various fields. Chen et al. [46] presented a Group Ensemble Network (GENet) and a survey on ensemble learning [47] was expanded. Zhou et al. [10] proposed the method of domain adaptive ensemble learning. These research studies also encourage scholars to use ensemble learning methods to predict video memorability. Zhao et al. [48] proposed the ensemble methods with text, image, audio, and video features that are extracted. However, the method of the multi-modal approach to predict memorability also has some problems. Simply speaking, the ensemble weight is not updated, and the size of the dataset they used is small. In the paper by Azcona et al. [49], in MediaEval 2019, they used ensemble transfer learning methods with semantics and their extract features to predict media memorability scores. Specifically, each feature gets a relatively good result through the base learner, then each base learner is integrated into a strong learner by weight or others. Zhou et al. [50] introduce the knowledge about ensemble learning in detail and provide the latest inspiration for the current research. In the latest study, Newman et al. [7] analyze the influence of semantics and time decay for video memorability and construct a multi-modal memorability dataset named Memento10k. Unlike VideoMem, Memento10k includes the memorability annotations that occur delays ranging from several seconds to ten minutes. They propose a SemanticMemNet that can predict video memorability at an arbitrary delay.

With the enlightenment of the above academic research, we propose a new adaptive multi-modal ensemble network (AMEN) for video memorability according to the selected impact factors. It eliminates the error caused by the heterogeneous gap via integrating the optimal base learners corresponding to each media source instead of fusing the heterogeneous modalities directly. In addition, we design an adaptive learning strategy to update the weights based on the importance of memorability which is predicted by the base learners and contributes to obtaining the best performance without any manual tuning.

3. Proposed Methods

Many studies [5,7,31,46,48] have proved that both vision and semantics play a key role in video memorability prediction. However, there is no definite content about the selection

of impact factors for video memorability. We take the video itself, the video key frame and the corresponding captions of the video as video, image and text, respectively, and mine the feature information contained in these three media information first. Next, the features of each media information source and the individual learner required are analyzed and selected to be put into the ensemble network. In Sections 3.1–3.3, we specifically elaborate on the reason for selecting features and individual learners needed in the proposed Adaptive Multi-model Ensemble Network (AMEN). Now, we briefly explain the symbols needed for feature selection and video memorability scores calculation. Given a video instance $V_i = (v_i, i_i, t_i)$, where v_i is video feature, i_i is image feature and t_i is text feature. The memorability score obtained from the video information source of video i is called video-output, termed as Vo_i , and the score obtained from the image source is addressed as image-output and defined as Io_i . Similarly, the score obtained from the text source is called text-output, which is denoted as To_i . The final video memorability score for video i obtained by our model is named video-memorability-output, which is referred to Vmo_i and n represents the number of samples.

3.1. Video Representation

ResNet [51] made a name for itself and influenced the direction of deep learning in academia and industry. Research shows that the depth of the network is an essential factor in achieving good results. However, the gradient dispersion/explosion becomes an obstacle to the training of the deep network, leading to the failure of convergence. The ResNet model provides better performance of the network to make up for this disadvantage, but also becomes a relatively advanced network model. Considering the feature redundancy and other problems brought by the direct use of the network to extract video features, such as C3D features, we adopt the end-to-end Resnet3D model by feedback process to directly predict the video memory score, which can achieve better prediction effect and make full use of video information. The experimental results also prove our guess. Based on this, we fine-tune the ResNet3D model [6] to use an end-to-end approach to predict video memorability scores rather than extracting features directly for video features.

Based on previous studies, we chose the ResNet3D model with 34 residual blocks and replaced the classification layer of the last layer with the full connection layers to obtain the memorability scores. Compared with the ResNet3D model fine-tuned by Cohendet et al. [5], we added two hidden layers and one output layer in the full connection layer, with 100 neurons in the first hidden layer, ten neurons in the second hidden layer and output in the output layer. In the training stage, the loss function was adjusted from the $L1$ loss function to the MSE loss function, and the optimization algorithm was adapted from the Adam algorithm to the SGD algorithm. Given a video instance $V_i = (v_i, i_i, t_i)$, output values of the trained model were selected to obtain with the fine-tuned ResNet3D model instead of extracting the feature v_i directly. For example, let us define γ_v as the parameter of the video branch, the learned video-output Vo_i of video i can be represented as follows:

$$Vo_i = f_v(v_i, \gamma_v) \quad (1)$$

where f_v represents the function of video features v_i , which is used to obtain the memorability score of the video, and $Vo \in \mathbb{R}^{n \times 1} = \{Vo_i\}_{i=1}^n$.

3.2. Text Representation

Firstly, for text data, stopwords were used to process some unnecessary words. Then, we used the count-vectorizer method to extract text content as semantic features. For each training text, it only considers the frequency of each word. Countvectorizer converts the words in the text into a word frequency matrix, which uses the fit-transform function to calculate the number of times each word appears. After converting words into vectors, we used principal component analysis (PCA) to reduce dimension. PCA is a standard data analysis method that is often used for dimensionality reduction of high-dimensional data and can extract the main feature components of data. After PCA dimensionality reduction,

the dimension of text features corresponding to each video is 500-dimensional. Considering that the Random Forest (RF) algorithm has strong model generalization ability and fast training speed and the multi-layer perceptron (MLP) can quickly regress high-dimensional features into a memorability score, we compared and analyzed these two methods through experiments to more accurately predict the memorability score of the base learner, and then the final video memorability score is predicted more accurately. Finally, we used Spearman's correlation coefficient to judge the correlation. The experimental results show that the multi-layer perceptron (MLP) can predict the memorability scores more efficiently as an individual learner. The following results are described in Section 4.2.1.

Given a video instance $V_i = (v_i, i_i, t_i)$, transforms semantic information of video i into 500-dimensional vector features t_i . Let us define γ_t as the parameter of the text branch, the learned image-output To_i of video i can be represented as follows:

$$To_i = f_t(t_i, \gamma_t) \quad (2)$$

where f_t defines the function of text features t_i , which is used to obtain the memorability score of the semantic derived from the video i , and $To \in \mathbb{R}^{n \times 1} = \{To_i\}_{i=1}^n$.

3.3. Image Representation

For the selection of image features, we referred to some valuable image memorability literature, such as HOG, LBP, and RGB, which can be considered. Considering that the image information source may cause a lot of feature redundancy after fusing multiple features, or the size of the constituent feature vector is too large compared with the size of the dataset [1]. After our previous experimental comparison [4], we finally chose the LBP feature as the image feature of modeling. LBP is the abbreviation of Local Binary Pattern (Local Binary Pattern) and an effective texture description operator, which measures and extracts the local texture information of the image, which has obvious advantages such as gray invariance and rotation invariance. To a certain extent, the problem of illumination change is eliminated. In addition, it has the advantages of rotation invariance, low texture feature dimension, and fast calculation speed.

After determining the image features, based on our previous method of predicting video memorability [4], Random Forest (RF), Support Vector Regression (SVR), and fully connected layer (MLP) were chosen as regression models to predict memorability scores. Random Forest (RF) shows higher efficiency in predicting memorability scores using LBP features than the other two methods. Given a video instance $V_i = (v_i, i_i, t_i)$, we used the obtained LBP features i_i . Let us define γ_t as the parameter of the text branch, the learned image—output Io_i of video i can be represented as follows:

$$Io_i = f_i(i_i, \gamma_t) \quad (3)$$

where f_i denotes the function of text features i_i , which is used to obtain the memorability score of the image derived from the video i , and $Io \in \mathbb{R}^{n \times 1} = \{Io_i\}_{i=1}^n$.

3.4. Weighted Ensemble

Ensemble learning is an algorithm that builds and combines multiple primary learners to achieve a more robust learning capability. Most integration methods use the same basic learning algorithm to produce homogeneous basic learners, that is, the same kind of learners produce homogeneous integrations, but in this paper, we tried to use a variety of learning algorithms to train different kinds of learners to produce heterogeneous integrations, and these learners were called individual learners. The ensemble learning method improves the generalization ability by combining a group of individual learners rather than choosing the best one among them, so the combination method used is very important.

After consideration, we used the weighted method to combine the individual learners of video, image, and text. The method is the most popular and primary combination method, assuming a set of T individual learner h_1, \dots, h_T , where the output of the learner

h_i on example x is $h_i \in R$. The task is to combine the output result of h_i to obtain the final prediction result on the actual value variable. The weighted method brings the combined results by assigning different importance weights to the output results of each learner. The weighted method obtains mixed results by giving different importance weights for the output results of each learner. Specifically, it obtains the combined output $H(x)$ in such a way that

$$H(x) = \sum_{i=1}^T \omega_i h_i(x) \quad (4)$$

where ω_i represents the weight of h_i , usually with the constraint $\omega_i \geq 0$ and $\sum_{i=1}^T \omega_i = 1$.

After we obtained Vo_i , To_i and Io_i , as mentioned in Sections 3.1–3.3, we utilized the weighted method to obtain the Vmo_i . We assigned different initial weights to the outputs of respective models such as ω_{1i} , ω_{2i} , ω_{3i} for video i . The initial learned $Vmo_{i'}$ of video i can be expressed as follows:

$$Vmo_{i'} = \omega_{1i} Vo_i + \omega_{2i} Io_i + \omega_{3i} To_i \quad (5)$$

As described in Section 1, video memorability is the probability of being remembered by people. That is to say, the final predicted video-memorability-output (Vmo_i) is in the range of $[0, 1]$. Therefore, the sigmoid function was used to control the prediction result within range $[0, 1]$. The final video-memorability-output (Vmo_i) is

$$Vmo_i = \frac{1}{1 + e^{-Vmo_{i'}}} \quad (6)$$

Based on the video-memorability-output (Vmo_i) for each video i , we used short-term memory and long-term memory as groundtruth. We optimized our model by minimizing the mean square error (MSE) between the predicted value and the true value through the stochastic gradient descent (SGD) algorithm, in which the loss function L can be calculated as follows:

$$\min_{\omega_{1i}, \omega_{2i}, \omega_{3i}} L = \frac{1}{n} \sum_{i=1}^n (Y_{groundtruth_i} - Vmo_i)^2 \quad (7)$$

where $Y_{groundtruth_i}$ represents the groundtruth of short-term or long-term memorability scores, the number of the sample is n , and Vmo_i denotes the predicted video memorability scores, which are video-memorability-outputs.

To prevent overfitting, L1 regularization and L2 regularization were added, and the final loss function L was

$$\min_{\omega_{1i}, \omega_{2i}, \omega_{3i}} L = \frac{1}{n} \sum_{i=1}^n (Y_{groundtruth_i} - Vmo_i)^2 + \lambda_1 \sum_{j=1}^3 \|\omega_{ji}\| + \lambda_2 \sum_{j=1}^3 \|\omega_{ji}\|_2 \quad (8)$$

where λ is the regularization coefficient, $\sum_{j=1}^3 \|\omega_{ji}\|$ is L1-norm and $\sum_{j=1}^3 \|\omega_{ji}\|_2$ is L2-norm.

For n samples, the video-memorability-outputs (Vmo) is calculated represented as follows:

$$Vmo = \omega_1 Vo + \omega_2 Io + \omega_3 To \quad (9)$$

where $\omega_1 = (\omega_{11}, \omega_{12}, \dots, \omega_{1n})$, $\omega_2 = (\omega_{21}, \omega_{22}, \dots, \omega_{2n})$, $\omega_3 = (\omega_{31}, \omega_{32}, \dots, \omega_{3n})$, $Vmo = (Vmo_1, Vmo_2, \dots, Vmo_n)^T$, $Io = (Io_1, Io_2, \dots, Io_n)^T$ and $To = (To_1, To_2, \dots, To_n)^T$.

3.5. Evaluation Indicator: Spearman's Rank Correlation Coefficient

In this paper, Spearman's rank correlation coefficient is used to calculate the correlation between the video memorability scores predicted by our model and the real video memorability scores. The closer the Spearman coefficient is to 1, the higher the correlation between the predicted value and groundtruth. On the contrary, the closer the coefficient is to -1 , the lower the correlation. Indicators commonly used in statistics to measure the correlation between two variables include Person correlation coefficient [52], Spearman's

rank correlation coefficient [53] and Kendall correlation coefficient [54]. Pearson's correlation coefficient applies to continuous data where there is a linear relationship between two variables and the population of the two variables is normally distributed or near-normal unimodal distribution. However, as can be seen from Figure 1, our data are not normally distributed, whether short-term or long-term memory scores. Although Spearman's rank correlation coefficient and Kendall's can be used for data with non-uniform distribution, Kendall's coefficient is more suitable for the multi-column rank correlation degree method. Furthermore, Spearman's rank correlation coefficient has less strict requirements on data conditions than Pearson's correlation coefficient. As long as the observed values of the two variables are paired with rank assessment data, Spearman's rank correlation coefficient can be used to analyze the overall distribution patterns and sample sizes of the two variables.

Spearman's correlation coefficient [53] is defined as the Pearson correlation coefficient between rank variables. For sample size n , n original data are converted into rank data, and the correlation coefficient ρ is

$$\rho = \frac{\sum_i^n (Vmo_i - \overline{Vmo})(Y_{groudtruth_i} - \bar{Y})}{\sqrt{\sum_i^n (Vmo_i - \overline{Vmo})^2 \sum_i^n (Y_{groudtruth_i} - \bar{Y})^2}} \quad (10)$$

where Vmo_i and $Y_{groudtruth_i}$ are the predicted value and the groundtruth of video i , \overline{Vmo} and \bar{Y} are the mean of Vmo_i and Y . Original data is assigned a rank based on its average descending position in the overall data. In practice, the links between variables are irrelevant, so the ρ can be calculated in a simple step. The difference between the ranks of the two variables observed, ρ is

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (11)$$

$$d_i = rank(Vmo_i) - rank(Y_{groudtruth_i}) \quad (12)$$

where $rank(Vmo_i)$ and $rank(Y_{groudtruth_i})$ are the order in the list after rearranging the original data Vmo_i and $Y_{groudtruth_i}$ in ascending order.

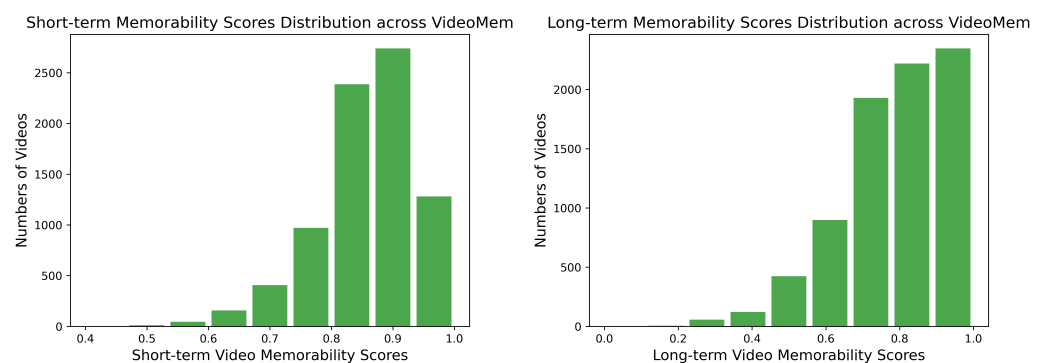


Figure 1. The Distribution of Short-term and Long-term Video Memorability Scores.

3.6. Framework Overview

As is shown in Figure 2, the AMEN framework adaptively integrates three individual learners to predict the final video memorability score based on the selection of the key features of the three media information sources and the corresponding individual learners. Specifically, the fine-tuned ResNet3D model, RF, and MLP that have been fine-tuned for the three information sources of video, image, and text are trained to obtain their respective output values and then use the weighted method to update the weight of each output value to obtain the final video memorability score.

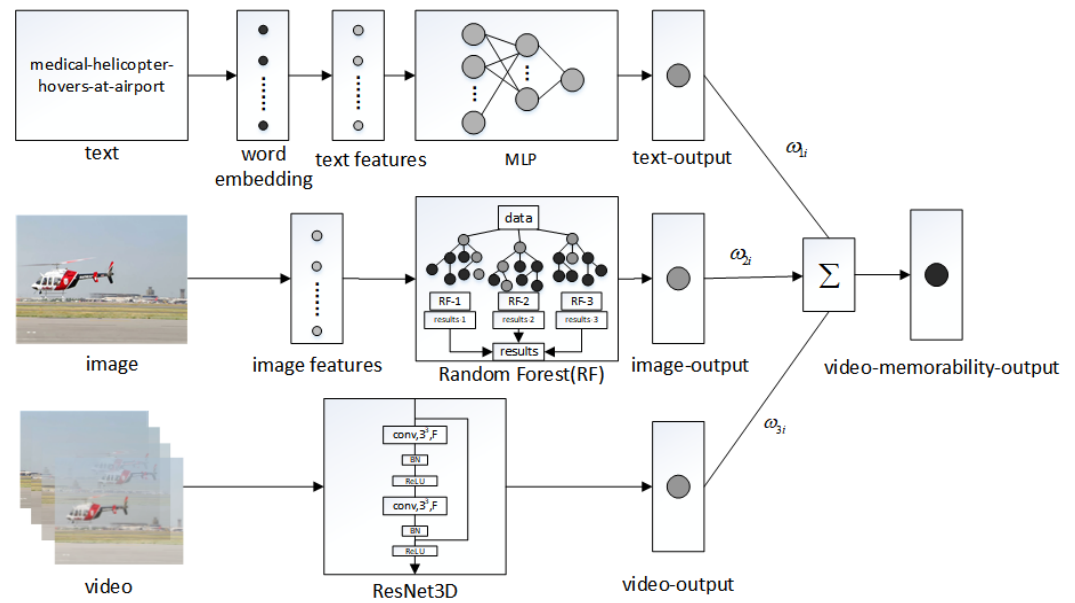


Figure 2. The overall framework of the proposed Adaptive Multi-model Ensemble Network (AMEN), see Algorithm 1.

Algorithm 1: AMEN training process.

```

1 Data:  $Vo, Io, To \leftarrow$  Video features, image features and text features
   Result:  $Vmo \leftarrow$  The video-memorability-outputs
2 while not converged do
3    $Vo_i = f_v(v_i, \gamma_v) \leftarrow$  Learning video features via Equation (1)
4    $To_i = f_t(t_i, \gamma_t) \leftarrow$  Learning text features via Equation (2)
5    $Io_i = f_i(i_i, \gamma_i) \leftarrow$  Learning image features via Equation (3)
6    $Vmo_{i'} = \omega_{1i}Vo_i + \omega_{2i}Io_i + \omega_{3i}To_i$ 
7    $Vmo_i = \frac{1}{1+e^{-Vmo_{i'}}}$ 
8    $\min_{\omega_{1i}, \omega_{2i}, \omega_{3i}} L = \frac{1}{n} \sum_{i=1}^n (Y_{groudtruth\_i} - Vmo_i)^2 + \lambda_1 \sum_{j=1}^3 \|\omega_{ji}\| + \lambda_2 \sum_{j=1}^3 \|\omega_{ji}\|_2$ 
9 end

```

4. Experiments

4.1. Dataset

In 2019, Romain Cohendet et al. [5] introduced a new protocol to measure video memorability scores and constructed a VideoMem dataset with short-term and long-term memorability scores. In this paper, we adapt 8000 silent videos from the VideoMem dataset, where each video contains a semantic shot for seven seconds. The types of videos are colorful and include different scenes such as nature, people, animals, etc. Each video corresponds to its own short-term and long-term memory scores, and the distribution of short-term and long-term memorability scores are shown in Figure 1.

In the experiment to measure the video memorability score, participants were given a series of videos, including target and non-target videos, which we call filler videos. The role of the filler video is to provide the influence of the time interval and other memory points. When the target video appeared, the participant clicked the space bar according to whether he remembered and then asked the participant to measure whether he remembered the target video again after 24–72 h. The experimental process is the same as shown in the Figure 3.



Figure 3. Experimental process of measuring video memory score.

In this paper, we divide the 8000 videos in a 3:1:1 ratio into the training set (4800 videos), the validation set (1600 videos), and the test set (1600 videos) for training, evaluating, and testing three models with text, image and video as input, respectively.

4.2. Prediction Results Analysis

4.2.1. Selection of Individual Learners from Different Media Information Sources

Firstly, 4800 videos were used to pre-train the fine-tuned ResNet3D model to obtain the memorability scores after the optimal model verification by using 1600 validation videos. At the same time, we compared the results of directly extracting C3D features and fine-tuning ResNet3D for short-term and long-term memorability scores. The comparative effects of the experiment are shown in Table 1. It can be seen that compared with the directly extracted C3D features, the fine-tuned ResNet3D model makes fuller use of the video information source, which means that it can predict the memorability scores more effectively. Moreover, the memorability scores obtained from the video information source are used as the video inputs of the weighted model.

Table 1. Results in terms of Spearman's rank correlation of **video features**.

Feature	ResNet3D		C3D	
	Short-Term	Long-Term	Short-Term	Long-Term
Video	0.331	0.147	0.291	0.132

Secondly, as mentioned in Section 3.2, after converting text information into vectors to obtain text features, the text feature was used to predict video memorability scores. Next, based on previous experiments and through the research of this paper, we determined using the Random Forest (RF) algorithm and the full connection layer (MLP) method for comparison. We took short-term memory and long-term memory as groundtruth successively. That is to say, short-term memory scores and long-term memory scores were, respectively, taken as the target scores. For the full-connection layer method, the input is a 500-dimensional text feature, a hidden layer with ten neurons. The activation function is set as tanh, and the optimization algorithm is set as the L-BFGS algorithm.

Table 2 demonstrates that whether it is short-term or long-term memory, the MLP method is more useful for us to make predictions for text information. Then, we used the video memorability scores obtained by using the semantic information as the semantic input of the weighted model.

Table 2. Results in terms of Spearman's rank correlation of **semantic feature**.

Feature	RF		MLP	
	Short-Term	Long-Term	Short-Term	Long-Term
Semantics	0.297	0.110	0.384	0.136

Following by selecting LBP features, we considered which individual learner can use LBP features to obtain the best prediction results. We chose Random Forest (RF), Support Vector Regression (SVR), and fully connected layer (MLP) as regression models to predict

memorability and used the Spearman coefficient to compare the results of the three models. Table 3 proves that the random forest algorithm as a regression model is more suitable for predicting the memorability scores of the image source.

Table 3. Results in terms of Spearman’s rank correlation of **LBP feature** across different learners.

Feature	RF		MLP		SVR	
	Short-Term	Long-Term	Short-Term	Long-Term	Short-Term	Long-Term
LBP	0.242	0.068	0.179	0.066	0.115	0.040

4.2.2. Comparison of Experimental Results

In this section, Table 4, respectively, shows the prediction results obtained using only video, image, or text sources. As mentioned in Section 3.4, video-memorability-outputs (Vmo) is weighted by video-outputs (Vo), image-outputs (Io) and text-outputs (To). So we attempt to weigh only the video source and the text source, and simultaneously weigh the video source and the text source, and compare the memorability scores predicted by the model after the weighting of the video, image, and text. (1) ResNet3D: Only the video features are used for prediction. (2) LBP: Only the image features are used for prediction. (3) Semantics: Make predictions just from text features. (4) ResNet3D+Semantics: The fusion features of video and text are used for prediction. (5) ResNet3D+LBP: The fusion features of video and image are used for prediction. (6) Semantic embedding model: The state-of-the-art method for comparison. (7) AMEN: The proposed method makes predictions by adaptively weighting text, image and video features. The experiments have proved that weighting the three media information sources has a positive effect on the prediction of the memory of the video. That is to say, the addition of image information and text information has a positive effect on the prediction of short-term or long-term memorability. Compared with the experimental results of the existing methods provided by [5], the results in Table 5 show that our method improves the video memorability score to a certain extent.

Meanwhile, to verify the effectiveness of the model, we draw the final Loss diagram. As shown in Figure 4, the model we proposed converges both for short-term and long-term memorability scores during training and validation.

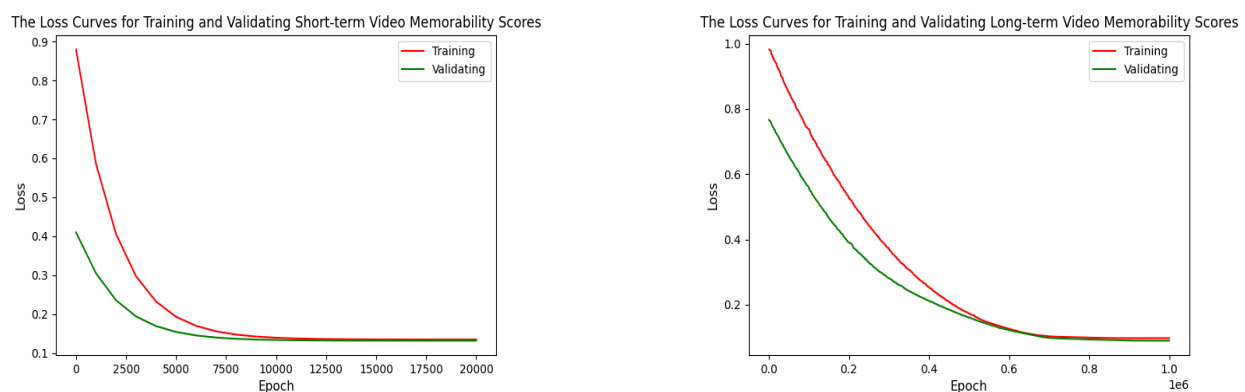


Figure 4. The loss curves for training and validating short-term and long-term video memorability scores.

Table 4. Results in terms of Spearman’s rank correlation of our model.

Model	Short-Term		Long-Term	
	Validation	Test	Validation	Test
ResNet3D	0.331	0.557	0.147	0.130
LBP	0.247	0.242	0.085	0.068
Semantics	0.383	0.384	0.159	0.136
ResNet3D+Semantics	0.718	0.573	0.251	0.126
ResNet3D+LBP	0.559	0.567	0.214	0.148
AMEN	0.829	0.604	0.923	0.259

Table 5. Comparisons with different methods in terms of Spearman’s rank correlation.

Model	Short-Term		Long-Term	
	Validation	Test	Validation	Test
MemNet [21]	0.397	0.385	0.195	0.168
Squalli [24] et al.	0.401	0.398	0.201	0.182
C3D [55]	0.319	0.322	0.175	0.154
HMP [56]	0.469	0.314	0.222	0.129
Semantic embedding model [5]	0.503	0.494	0.260	0.256
AMEN	0.829	0.604	0.923	0.259

5. Conclusions

In this work, we provided a new framework, the Adaptive Multi-modal Ensemble Network (AMEN), to predict the video memorability scores. We identified three impact factors that affect video memorability prediction, including temporal 3D information, spatial information, and semantics information. AMEN integrated three individual learners using the weighted method rather than feature fusion based on these three factors. In addition, we updated the weight based on the importance of memorability, which is predicted by each individual learner automatically rather than assigning weight manually. Through training and testing, experimental results on the VideoMem dataset proved that our method could better predict the video memorability scores. We understand that the current research on multi-modal memorability is only the fusion of features. A large number of features fusion will cause feature redundancy. Even though different researchers have experimented on various factors, there is still no definite feature to guide the study of video memorability. Therefore, the research on video memorability prediction will be more challenging in the future.

Author Contributions: Data curation, X.G.; Methodology, J.S.; Software, F.Y. and F.X.; Writing—original draft, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Scientific Research Leader Studio of Jinan (No. 2021GXRC081), the Natural Science Foundation of Shandong Province (No. ZR2021LZH010), and Joint Project for Smart Computing of Shandong Natural Science Foundation (No. ZR2020LZH015).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cohendet, R.; Yadati, K.; Duong, N.Q.; Demarty, C.H. Annotating, understanding, and predicting long-term video memorability. In Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, Yokohama, Japan, 11–14 June 2018; pp. 178–186.

2. Xiao, J.; Hays, J.; Ehinger, K.A.; Oliva, A.; Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3485–3492.
3. Kar, A.; Mavin, P.; Ghaturlu, Y.; Vani, M. What makes a video memorable? In Proceedings of the 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Tokyo, Japan, 19–21 October 2017; pp. 373–381.
4. Yue, F.; Li, J.; Sun, J. Insights of Feature Fusion for Video Memorability Prediction. In *International Forum on Digital TV and Wireless Multimedia Communications*; Springer: Singapore, 2020; pp. 239–248.
5. Cohendet, R.; Demarty, C.-H.; Duong, N.Q.K.; Engilberge, M. VideoMem: Constructing, analyzing, predicting short-term and long-term video memorability. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 2531–2540.
6. Hara, K.; Kataoka, H.; Satoh, Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6546–6555.
7. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987.
8. Hunt, R.W.G. *The Reproduction of Colour*; Fountain Press: London, UK, 1995.
9. Schüller, C.; Brewster, J.L.; Alexander M.R.; Gustin, M.C.; Ruis, H. The HOG pathway controls osmotic regulation of transcription via the stress response element (STRE) of the *Saccharomyces cerevisiae* CTT1 gene. *EMBO J.* **1994**, *13*, 4382–4389. [[PubMed](#)]
10. Zhou, K.; Yang, Y.; Qiao, Y.; Xiang, T. Domain adaptive ensemble learning. *IEEE Trans. Image Process.* **2021**, *30*, 8008–8018. [[PubMed](#)]
11. Isola, P.; Xiao, J.; Parikh, D.; Torralba, A.; Oliva, A. What makes a photograph memorable? *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 1469–1482. [[CrossRef](#)] [[PubMed](#)]
12. Bylinskii, Z.; Isola, P.; Bainbridge, C.; Torralba, A.; Oliva, A. Intrinsic and extrinsic effects on image memorability. *Vis. Res.* **2015**, *116*, 165–178. [[CrossRef](#)] [[PubMed](#)]
13. Yoon, S.; Kim, J. Object-centric scene understanding for image memorability prediction. In Proceedings of the 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Miami, FL, USA, 10–12 April 2018; pp. 305–308.
14. Basavaraju, S.; Mittal, P.; Sur, A. Image memorability: The role of depth and motion. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 699–703.
15. Bainbridge, W.A. Memorability: How what we see influences what we remember. In *Psychology of Learning and Motivation*; Academic Press: Cambridge, MA, USA, 2019; Volume 70, pp. 1–27.
16. Constantin, M.G.; Redi, M.; Zen, G.; Ionescu, B. Computational understanding of visual interestingness beyond semantics: literature survey and analysis of covariates. *ACM Comput. Surv. (Csur)* **2019**, *52*, 1–37. [[CrossRef](#)]
17. Dubey, R.; Peterson, J.; Khosla, A.; Yang, M.H.; Ghanem, B. What makes an object memorable? In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1089–1097.
18. Akagunduz, E.; Bors, A.G.; Evans, K.K. Defining image memorability using the visual memory schema. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2165–2178. [[CrossRef](#)] [[PubMed](#)]
19. Le, N.Q.K.; Ho, Q.T. Deep transformers and convolutional neural network in identifying DNA N6-methyladenine sites in cross-species genomes. *Methods* **2022**, *204*, 199–206. [[CrossRef](#)] [[PubMed](#)]
20. Tng, S.S.; Le N.Q.; Yeh, H.Y.; Chua, M.C. Improved prediction model of protein lysine Crotonylation sites using bidirectional recurrent neural networks. *J. Proteome Res.* **2021**, *21*, 265–273. [[CrossRef](#)] [[PubMed](#)]
21. Khosla, A.; Raju, A.S.; Torralba, A.; Oliva, A. Understanding and predicting image memorability at a large scale. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2390–2398.
22. Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; Oliva, A. Learning deep features for scene recognition using places database. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 27.
23. Zarezadeh, S.; Rezaeian, M.; Sadeghi, M.T. Image memorability prediction using deep features. In Proceedings of the 2017 Iranian Conference on Electrical Engineering (ICEE), Tehran, Iran, 2–4 May 2017; pp. 2176–2181.
24. Squalli-Houssaini, H.; Duong, N.Q.; Gwenaëlle, M.; Demarty, C.H. Deep learning for predicting image memorability. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2371–2375.
25. Yu, E.; Sun, J.; Li, J.; Chang, X.; Han, X.H.; Hauptmann, A.G. Adaptive semi-supervised feature selection for cross-modal retrieval. *IEEE Trans. Multimed.* **2018**, *21*, 1276–1288. [[CrossRef](#)]
26. Goswami, G.; Bhardwaj, R.; Singh, R.; Vatsa, M. MDLFace: Memorability augmented deep learning for video face recognition. In Proceedings of the IEEE International Joint Conference on Biometrics, Clearwater, FL, USA, 29 September–2 October 2014.
27. Wang, W.; Sun, J.; Liu, J. A memorability based method for video hashing. In Proceedings of the 2015 IEEE 16th International Conference on Communication Technology (ICCT), Hangzhou, China, 18–20 October 2015; pp. 309–313.
28. Han, J.; Chen, C.; Shao, L.; Hu, X.; Han, J.; Liu, T. Learning computational models of video memorability from fMRI brain imaging. *IEEE Trans. Cybern.* **2014**, *45*, 1692–1703. [[CrossRef](#)] [[PubMed](#)]
29. Shekhar, S.; Singal, D.; Singh, H.; Kedia, M.; Shetty, A. Show and recall: Learning what makes videos memorable. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 2730–2739.

30. Awad, G.; Butt, A.A.; Curtis, K.; Lee, Y.; Fiscus, J.; Godil, A.; Delgado, A.; Zhang, J.; Godard, E.; Diduch, L.; et al. Trecvid 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval. *arXiv* **2020**, arXiv:2009.09984.
31. De Herrera, A.G.; Kiziltepe, R.S.; Chamberlain, J.; Constantin, M.G.; Demarty, C.H.; Doctor, F.; Ionescu, B.; Smeaton, A.F. Overview of MediaEval 2020 predicting media memorability task: What makes a video memorable? *arXiv* **2020**, arXiv:2012.15650.
32. Smeaton, A.F.; Corrigan, O.; Dockree, P.; Gurrin, C.; Healy, G.; Hu, F.; McGuinness, K.; Mohedano, E.; Ward, T.E. Dublin's Participation in the Predicting Media Memorability Task at MediaEval 2018. MediaEval. 2018. Available online: <https://www.youtube.com/watch?v=yEOtjq6Qu3s&t=11s> (accessed on 25 June 2022).
33. Chaudhry, R.; Kilaru, M.; Shekhar, S. Show and Recall@ MediaEval 2018 ViMemNet: Predicting Video Memorability. *Group* **2018**, 1, G1.
34. Tran-Van, D.T.; Tran, L.V.; Tran, M.T. Predicting Media Memorability Using Deep Features and Recurrent Network. MediaEval. 2018. Available online: <https://www.semanticscholar.org/paper/Predicting-Media-Memorability-Using-Deep-Features-Tran-Van-Tran/44cfbfca6008248f4a9cd75d182cbeca15c1ab9e> (accessed on 25 June 2022).
35. Gupta, R.; Motwani, K. Linear Models for Video Memorability Prediction Using Visual and Semantic Features. MediaEval. 2018. Available online: <https://www.semanticscholar.org/paper/Linear-Models-for-Video-Memorability-Prediction-and-Gupta-Motwani/147ee939c1bffe633b646d729b8edac98edc7093> (accessed on 25 June 2022).
36. Cohendet, R.; Demarty, C.H.; Duong, N.Q.K. Transfer Learning for Video Memorability Prediction. MediaEval. 2018. Available online: https://www.youtube.com/watch?v=kFyw3vwl_e4 (accessed on 25 June 2022).
37. Leyva, R.; Doctor, F.; Seco De Herrera, A.G.; Sahab, S. Multimodal Deep Features Fusion for Video Memorability Prediction. MediaEval. 2019. Available online: <http://repository.essex.ac.uk/id/eprint/26580> (accessed on 25 June 2022).
38. Leyva, R.; Sanchez, V. Video memorability prediction via late fusion of deep multi-modal features. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 2488–2492.
39. Kleinlein, R.; Luna-Jiménez, C.; Arias-Cuadrado, D.; Ferreira, J.; Fernández-Martínez, F. Topic-Oriented Text Features Can Match Visual Deep Models of Video Memorability. *Appl. Sci.* **2021**, *11*, 7406. [CrossRef]
40. Ali, H.; Gilani, S.O.; Khan, M.J.; Waris, A.; Khattak, M.K.; Jamil, M. Predicting Episodic Video Memorability Using Deep Features Fusion Strategy. In Proceedings of the 2022 IEEE/ACIS 20th International Conference on Software Engineering Research, Management and Applications (SERA), Las Vegas, NV, USA, 20–22 June 2022; pp. 39–46.
41. Ghosal, D.; Akhtar, M.S.; Chauhan, D.; Poria, S.; Ekbal, A.; Bhattacharyya, P. Contextual inter-modal attention for multi-modal sentiment analysis. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 3454–3466.
42. Wang, L.; Zhu, L.; Yu, E.; Sun, J.; Zhang, H. Fusion-supervised deep cross-modal hashing. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; pp. 37–42.
43. Patrick, M.; Asano, Y.M.; Kuznetsova, P.; Fong, R.; Henriques, J.F.; Zweig, G.; Vedaldi, A. Multi-modal self-supervision from generalized data transformations. *arXiv* **2020**, arXiv:2003.04298.
44. Gabeur, V.; Sun, C.; Alahari, K.; Schmid, C. Multi-modal transformer for video retrieval. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 214–229.
45. Liu, J.; Inkawhich, N.; Nina, O.; Timofte, R. NTIRE 2021 multi-modal aerial view object classification challenge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 588–595.
46. Pintelas, P.; Livieris, I.E. Special issue on ensemble learning and applications. *Algorithms* **2020**, *13*, 140. [CrossRef]
47. Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; Ma, Q. A survey on ensemble learning. *Front. Comput. Sci.* **2020**, *14*, 241–258. [CrossRef]
48. Zhao, T.; Fang, I.; Kim, J.; Friedl, G. Multi-modal ensemble models for predicting video memorability. *arXiv* **2021**, arXiv:2102.01173.
49. Azcona, D.; Moreu, E.; Hu, F.; Ward, T.E.; Smeaton, A.F. Predicting Media Memorability Using Ensemble Models. MediaEval. 2019. Available online: <https://www.semanticscholar.org/paper/Predicting-Media-Memorability-Using-Ensemble-Models-Azcona-Moreu/09cd29b6082a127a49bab414862a0b7a6fa3f8b1> (accessed on 25 June 2022).
50. Zhou, Z.H. Ensemble learning. In *Machine Learning*; Springer: Singapore, 2021; pp. 181–210.
51. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
52. Benesty, J.; Chen, J.; Huang, Y.; Cohen, I. Pearson correlation coefficient. In *Noise Reduction in Speech Processing*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1–4.
53. Myers, J.L.; Well, A.D.; Lorch, R.F., Jr. *Research Design and Statistical Analysis*; Routledge: London, UK, 2013.
54. Abdi, H. The Kendall rank correlation coefficient. In *Encyclopedia of Measurement and Statistics*; Sage: Thousand Oaks, CA, USA, 2007; pp. 508–510.
55. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
56. Ciptadi, A.; Goodwin, M.S.; Reh, J.M. Movement pattern histogram for action recognition and retrieval. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 695–710.