Fine-grained Video Attractiveness Prediction Using Multimodal Deep Learning on a Large Real-world Dataset

Xinpeng Chen[†*], Jingyuan Chen[♭*], Lin Ma[‡♮], Jian Yao[†], Wei Liu[‡♮], Jiebo Luo[§], Tong Zhang[‡]

[†]Wuhan University, [‡]Tencent AI Lab, [♭]National University of Singapore, [§]University of Rochester

ABSTRACT

Nowadays, billions of videos are online ready to be viewed and shared. Among an enormous volume of videos, some popular ones are widely viewed by online users while the majority attract little attention. Furthermore, within each video, different segments may attract significantly different numbers of views. This phenomenon leads to a challenging yet important problem, namely fine-grained video attractiveness prediction, which only relies on video contents to forecast video attractiveness at fine-grained levels, specifically video segments of several second length in this paper. However, one major obstacle for such a challenging problem is that no suitable benchmark dataset currently exists. To this end, we construct the first fine-grained video attractiveness dataset (FVAD), which is collected from one of the most popular video websites in the world. In total, the constructed FVAD consists of 1,019 drama episodes with 780.6 hours covering different categories and a wide variety of video contents. Apart from the large amount of videos, hundreds of millions of user behaviors during watching videos are also included, such as "view counts", "fast-forward", "fast-rewind', and so on, where "view counts" reflects the video attractiveness while other engagements capture the interactions between the viewers and videos. First, we demonstrate that video attractiveness and different engagements present different relationships. Second, FVAD provides us an opportunity to study the fine-grained video attractiveness prediction problem. We design different sequential models to perform video attractiveness prediction by relying solely on video contents. The sequential models exploit the multimodal relationships between visual and audio components of the video contents at different levels. Experimental results demonstrate the effectiveness of our proposed sequential models with different visual and audio representations, the necessity of incorporating the two modalities, and the complementary behaviors of the sequential prediction models at different levels.

CCS CONCEPTS

• **Computing methodologies** → *Scene understanding*;

KEYWORDS

Video Attractiveness, Fine-grained, Multimodal Fusion, Long Short-Term Memory (LSTM)

ACM Reference Format:

Xinpeng Chen[†*], Jingyuan Chen[♭*], Lin Ma[‡♮], Jian Yao[†], Wei Liu[‡♮], Jiebo Luo[§], Tong Zhang[‡] [†]Wuhan University, [‡]Tencent AI Lab, [♭]National University of Singapore, [§]University of Rochester. 2018. Fine-grained Video Attractiveness Prediction Using Multimodal Deep Learning on a Large Real-world Dataset. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3184558.3186584

1 INTRODUCTION

Today, digital videos are booming on the Internet. It is stated that traffic from online videos will constitute over 80% of all consumer Internet traffic by 2020[1]. Meanwhile, due to the advance of mobile devices, millions of new videos are streaming into the Web everyday. Interestingly, among an enormous volume of videos, only a small number of them are attractive to draw a great number of viewers, while the majority receive little attention. Even within the same video, different segments present different attractiveness to the audiences with a large variance. A video or segment is considered attractive if it gains a high view count based on the statistics gathered on a large number of users. The larger the view count is, the more attractive the corresponding video or segment is. The view count directly reflects general viewers' preferences, which are thus regarded as the sole indicator of the video attractiveness within the scope of this paper. Considering one episode from a hot TV series, as an example in Fig 1 (a), the orange line indicates the view counts (attractiveness) for the short video segments, which are crawled from one of the most popular video websites. As can be seen, video attractiveness varies greatly over different video segments, where the maximum view count is more than twice of the minimum value.

Predicting the attractiveness of video segments in advance can benefit many applications, such as online marketing [3] and video recommendation [4]. Regarding online marketing, accurate early attractiveness prediction of video segments can facilitate optimal planning of advertising campaigns and thus maximize the revenues. For video recommender systems, the proposed method provides an opportunity to recommend video segments based on their attractiveness scores.

However, predicting the video attractiveness is a very challenging task. First, the attractiveness of a video can be influenced by many external factors, such as the time that the video is posted

[*] Work done while Xinpeng Chen and Jingyuan Chen were Research Interns with Tencent AI Lab.

[♮]Correspondence: forest.linma@gmail.com; wliu@ee.columbia.edu.

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186584

[1]https://goo.gl/DrrKcn.

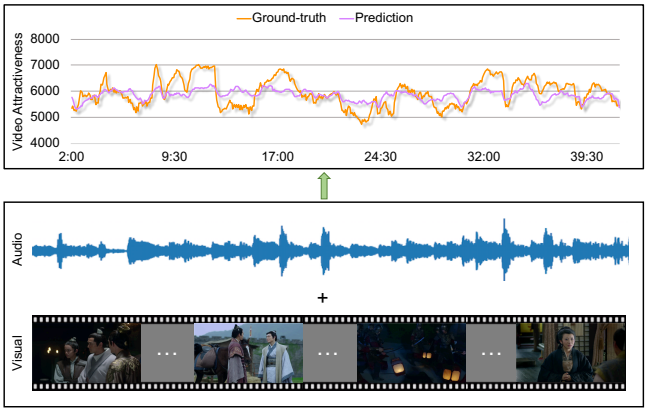

Figure 1: Definition of fine-grained video attractiveness prediction. The view counts (video attractiveness) for fine-grained video segments are shown, where the orange line denotes the ground-truth view counts based on hundreds of millions of active users. It can be observed that video attractiveness varies significantly over time. The main reason is that the contents of different video segments are of great diversity, with different visual information (*e.g.*, peaceful scenery vs. fierce battle) and different audio information (*e.g.*, soft background music vs. meaningful conversations). Such visual and audio contents together greatly influence the video attractiveness. Note that the purple line predicted based on both the visual and audio data using our proposed model in Sec. 4 can well track the trends of the ground-truth video attractiveness.

online, the advertisement intensity in the video, and so on. For the same category of videos, the more timely a video is delivered, the more views it will receive. Second, video attractiveness is also content-sensitive as shown in Fig. 1. Therefore, in order to make reliable predictions of video attractiveness, both visual and audio contents need to be analyzed. Several existing works [13, 26, 27, 38] have explored video interestingness or popularity. [13, 38] aimed at comparing the interestingness of two videos, while [26] relied on the historical information given by early popularity measurements. One problem is that the existing models only work on the video-level attractiveness prediction, while the fine-grained segment-level attractiveness prediction remains an open question without any attention. Another challenging problem is the lacking of large-scale real-world data. Recently released video datasets mostly focus on video content understanding, such as classification and captioning, specifically Sports-1M [15], YouTube-8M [1], ActivityNet [9], UCF-101 [30], FCVID [14], and TGIF [19]. These datasets do not incorporate any labels related to video attractiveness. In order to build reliable video attractiveness prediction systems, accurately labeled datasets are required. However, the existing video datasets for interestingness prediction [13, 28] are annotated by crowdsourcing. Such annotations only reflect the subjective opinions of a small number of viewers. Thus it cannot indicate the true attractiveness of the video sequence or segment.

In order to tackle the fine-grained video attractiveness prediction problem, we construct the **F**ine-grained **V**ideo **A**ttractiveness **D**ataset (FVAD), a new large-scale video benchmark for video attractiveness prediction. We collect the popular videos from one of the most popular video websites, which possesses thousands of millions of registered users. To date, FVAD contains 1,019 video episodes of 780.6 hours long in total, covering different categories and a wide variety of video contents. Moreover, the user engagements associated with each video are also included. Besides the view counts (attractiveness), there are other 9 types of engagement indicators associated with a video sequence to record the interactions between the viewers and videos, as illustrated in Fig. 3. We summarize our contributions as follows:

- We build the largest real-world dataset FVAD for dealing with the task of fine-grained video attractiveness prediction. The video sequences and their associated "labels" in the form of view count, as well as the viewers' engagements with videos are provided. The relationships between video attractiveness and engagements are examined and studied.
- Several sequential models for exploiting the relationships between visual and audio components for fine-grained video attractiveness prediction are proposed. Experimental results demonstrate the effectiveness of our proposed models and the necessity of jointly considering both the visual and audio modalities.

2 RELATED WORK

2.1 Video Datasets

Video datasets have played a critical role in advancing computer vision algorithms for video understanding. Several well labeled small-scale datasets, such as XM2VTS [23], KTH [18], Hollywood-2 [22], Weizmann [2], UCF101 [30], THUMOS'15 [12], HMDB [17], and ActivityNet [9], provide benchmarks for face recognition [21], human action recognition [11] and activity understanding. There are also other video datasets focusing on visual content recognition, video captioning, and so on, such as FCVID [14] and TGIF [19]. In order to make a full exploitation on the video content understanding, super large video datasets have been recently constructed. Sports-1M [15] is a dataset for sports video classification with 1 million videos. YFCC'14 [35] is a large multimedia dataset including about 0.8 million videos. The recent YouTube-8M [1] is so far the largest dataset for multi-label video classification, consisting of about 8 million videos. However, it is prohibitively expensive and time consuming to obtain a massive amount of well-labeled data. Therefore, these datasets inevitably introduce label noise when the labels are produced automatically. The most important thing is that all these datasets focus on understanding video contents, without touching on the video attractiveness task. MediaEval [28] is the only known public dataset, which is closely related to our work. It is used for predicting the interesting frames in movie trailers. However, MediaEval is a small dataset that only consists of 52 trailers for training and 26 trailers for testing. In addition, the interesting frames in MediaEval are labeled by a small number of subjects, which is not consistent with the real-life situation of massive diverse audiences.

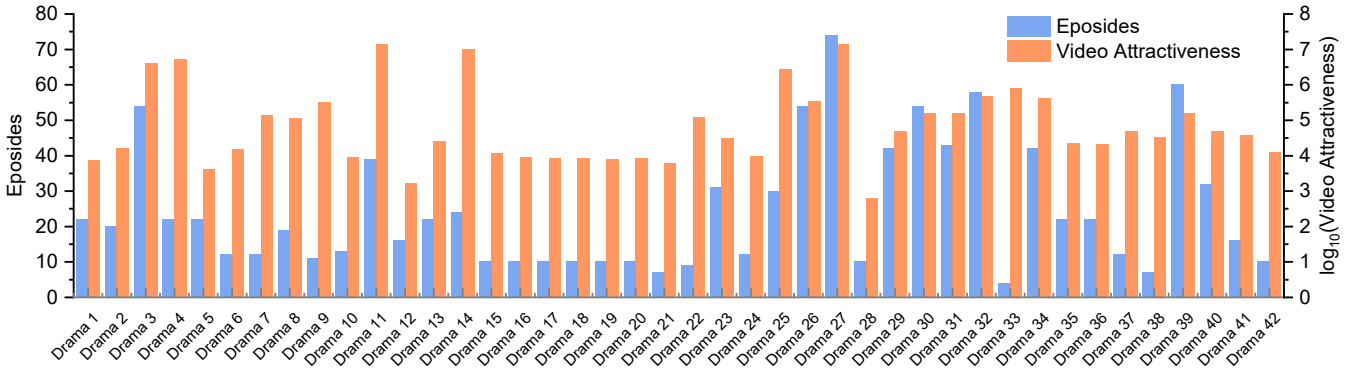

Figure 2: The statistics of 42 kinds of dramas in our constructed FVAD. The blue bars indicate the number of videos per TV series, while the orange ones indicate the video attractiveness (view counts) in the Log_{10} domain.

2.2 Video Attractiveness Prediction

A thread of work for predicting video interestingness or popularity is related to our proposed video attractiveness prediction. In [20], where Flickr images are used to measure the interestingness of video frames. Flickr images were assumed to be mostly interesting compared with many video frames since the former are generally well-composed and selected for sharing. A video frame is considered interesting if it matches (using image local features) with a large number of Flickr images. In [38], after extracting and combining the static and temporal features using kernel tricks, a relative score is predicted to determine which video is more interesting than the other using ranking SVM given a pair of videos. In [13], two datasets are collected based on interestingness ranking from Flickr and YouTube, and the interestingness of a video is predicted in the same way as [38]. In [26], the historical information given by early popularity measurements is used for video popularity prediction. A Hawkes intensity process is proposed to explain the complex popularity history of each video according to its type of content, network of diffusion, and sensitive to promotion [27]. Different from [13, 20, 38], video contents are not explicitly used for video popularity prediction [26, 27].

Our work is fundamentally different from the previous works. First, large-scale real-world user behaviors on one of the most popular video websites, are crawled to construct the proposed FVAD. Second,we aim to predict the fine-grained actual video attractiveness (view counts), compared with the video-level interestingness [13, 20, 38] and popularity [26, 27]. Third, we develop different sequential multimodal models to jointly learn the relationships between visual and audio components for the video attractiveness prediction. To the best of our knowledge, there is no existing work to handle and study the fine-grained video attractiveness prediction problem.

3 FVAD CONSTRUCTION

This section elaborates on the FVAD dataset construction, covering the video collecting strategy, the video attractiveness and engagements, and the analysis of their relationships.

3.1 Video Collection

To construct a representative dataset which contains video segments with diverse attractiveness degrees, video contents should cover different categories and present a broad range of diversities. We manually select a set of popular TV serials from the website. For different episodes and fragments within each episode, as the story develops, it is obvious that the attractiveness degree goes upward and downward. As shown in Fig. 1, the video contents, including the visual and audio components, significantly affect the video attractiveness presenting diverse view counts. For our FVAD dataset, we collected 1,019 episodes with a total duration of 780.6 hours long. The number of episodes with respect to each TV series is illustrated by the blue bars in Fig. 2. The average duration of all the episodes in FVAD is 45 minutes. Moreover, all the episodes were downloaded in high quality with the resolution of 640×480.

3.2 Video Attractiveness

In this paper, we focus on the fine-grained video attractiveness. Therefore, we need to collect the attractiveness indicators of the fine-grained video fragments. As aforementioned, the attractiveness degree for each video fragment is quantified by the total number of views. As shown in [31], visual media tends to receive views over some period of time. To normalize this effect, we divide the number of views by the duration from the upload date of the given episode to the collection date, which is 30th November, 2017. The orange bar in Fig. 2 illustrates the total video attractiveness of the TV series by summing all the view counts from all the episodes in each season. In order to make a better visualization, the attractiveness value is displayed in the $\textbf{Log}_{10}$ domain. It can be observed that the video attractiveness varies significantly among different TV series. Even for the same TV series, different seasons present different attractiveness.

3.3 Video Engagements

In addition to video views, we also collected 9 user engagement-related indicators regarding each video fragment, namely Exit, Start of Fast-Forward, End of Fast-Forward, Start of Fast-Rewind, End of Fast-Rewind, Fast-Forward Skips, Fast-Rewind Skips, Bullet Screens, and Bullet Screen Likes. The first 7 engagements are the natural

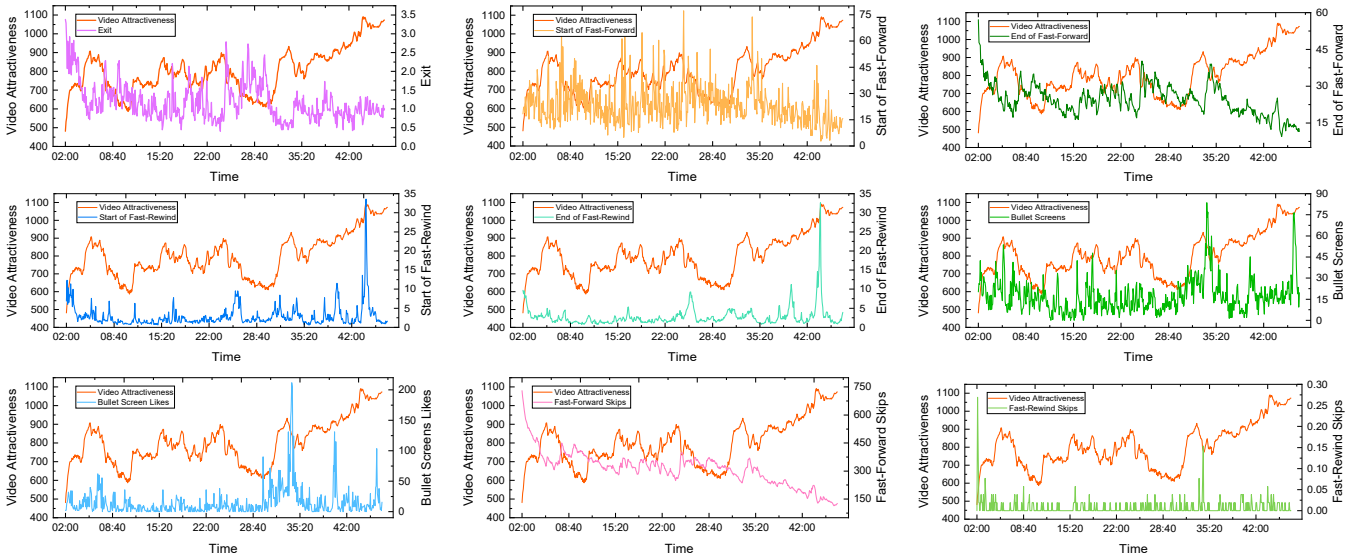

Figure 3: The other 9 types of viewers' engagements with the video sequences while watching. The view counts are also presented together with each engagement. It can be observed that the view counts present different correlations to these 9 types of viewers' engagements. From top-left to bottom-right: 1) Exit: the number of viewers exiting the show, 2) Start of Fast-Forward (FF): the number of viewers beginning FF, 3) End of Fast-Forward: the number of viewers stopping FF, 4) Start of Fast-Rewind (FR): the number of viewers beginning FR, 5) End of Fast-Rewind: the number of viewers stopping FR, 6) Bullet Screens: the number of bullet screens sent by viewers, 7) Bullet Screen Likes: the number of bullet screen likes of the viewers, 8) Fast-Forward Skips: the number of skip times during FF, and 9) Fast-Rewind Skips: the number of skip times during FR.

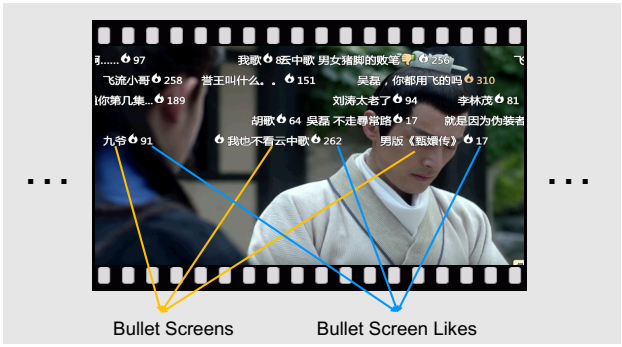

Figure 4: A simple example of bullet screens. Different users may express real-time opinions directly upon their interested frames.

user behaviors during the watching process, while the last two engagements, namely the Bullet Screens and Bullet Screen likes, involve deep interactions between viewers and videos.

Bullet Screens, also named as time-synchronized comments and first introduced in [37], allow users to express opinions directly on the frames of interest in a real-time manner. Intuitively, the user behaviors of commenting on a frame can be regarded as implicit feedback reflecting the frame-level preference, while the image features of the reviewed frame and the textual features in the posted comments can further help model the fine-grained preference from different perspectives. Fig. 4 shows a simple example of bullet screen. As can be seen, different users may express real-time opinions directly upon their interested frames. The number after each bullet screen in Fig. 4 indicates the total number of likes received by the corresponding bullet screen from the audience. The comment words from Bullet Screens can more accurately express the viewers' preferences and opinions. However, for this paper, we only collect the numbers of the Bullet Screens as well as their associated number of likes.

Fig. 3 illustrates the 9 different engagement indicators as well as the video attractiveness of one episode. It is noticed that the distributions of these different engagements are different. Each of them measures one aspect of users' engagement behaviors. These engagement characters intuitively correlate with the video attractiveness indicator (view counts). For example, high Fast-Forward Skips values always correspond to low attractiveness, while high Start of Fast-Rewind values correspond to high attractiveness.

3.4 Relationships between Video Attractiveness and Engagements

To evaluate the above correlations quantitatively, three kinds of coefficients, including Pearson correlation coefficient (PCC), cosine similarity (CS), and Spearman rank-order correlation coefficient (SRCC), are used to measure the strength and direction of the association between each engagement indicator and attractiveness. The correlations are provided in Table 1. It is demonstrated that different engagement indicators present different correlations with the attractiveness, where some present positive correlations while others present negative correlations. It is not surprising that the Start of Fast-Forward and Fast-Forward Skips present the largest

Table 1: The correlations between video attractiveness and different engagement indicators, in terms of Pearson correlation coefficient (PCC), cosine similarity (CS), and spearman's rank correlation coefficient (SRCC).

Indicator Name	PCC	CS	SRCC
Exit	-0.149	-0.148	-0.210
Start of Fast-Forward	-0.117	-0.117	-0.200
End of Fast-Forward	-0.537	-0.536	-0.522
Start of Fast-Rewind	0.327	0.327	0.368
End of Fast-Rewind	0.227	0.227	0.256
Bullet Screens	-0.139	-0.139	-0.191
Bullet Screen Likes	0.027	0.027	-0.020
Fast-Forward Skips	-0.351	-0.350	-0.315
Fast-Rewind Skips	0.022	0.022	0.013

positive and negative correlations, respectively. However, the indicator Bullet Screens shows negative correlations with video views. One possible reason is that the actual commented frame should be the one corresponding to the time when the user began to type the bullet screen, rather than the frame when the bullet screen was posted out. Therefore, the main reason is that the data about bullet screens is not well aligned. Another possible reason is that most bullet screens are complaints about the stories, therefore being not able to represent the attractiveness of the video. It is noted that both Bullet Screen Likes and Fast-Rewind Skips show less correlations with video views. One possible reason is that the value of each indicator is relatively small, which thereby cannot reflect statistical regularities.

4 VIDEO ATTRACTIVENESS PREDICTION USING DEEP LEARNING ON LARGE DATASETS

Video attractiveness prediction is a very challenging task, which may involve many external factors. For example, social influence is an important external factor, which makes a great impact on the number of views. In the Western world, the drama series such as *The Big Bang Theory* have a huge number of fans, which are of high attractiveness. However, for Chinese viewers, *The Big Bang Theory* are less attractiveness than some reality shows, such as *The Singer*. In the constructed FVAD, since user profile data is not available, we cannot track users' culture backgrounds or consider other social-related factors. Another important external factor is the director and starring list of the corresponding TV series. Specifically, a strong cast always boosts the base attractiveness of the whole series. For example, some dramas such as *Empresses in the Palace* with many famous stars attract billions of views.

Besides different external factors, video contents play the most important role in the task of video attractiveness prediction. In this paper, we aim at discovering the relationships between video contents and video attractiveness. Even further, we would like to make the prediction on the video attractiveness based solely on the video contents. Therefore, we need to first eliminate the effects of external factors. We use one simple method, namely the standardization, on the attractiveness as well as the other 9 engagement indicators. With such normalization, we can obtain the video relative attractiveness, which is regarded to be determined by the video contents only, specifically the visual and audio components. In the following, we will employ the normalized video attractiveness to perform the video attractiveness prediction.

4.1 Video Representation

To comprehensively understand video contents, we extract both visual and audio representations.

Visual representation. Recently developed convolutional neural networks (CNNs), such as VGG [29], Inception-X [10, 32–34] and ResNet [8], are usually utilized to generate global representations of images. Relying on these CNNs, we decode each video with FFmpeg, select 1 frame per second, feed each visual frame into a CNN model, and fetch the hidden states before the classification layer as the visual feature. Specifically, to exploit the capacity of different kinds of CNN models, we experiment a variety of CNNs, namely VGG-16, VGG-19, ResNet-152, Inception-X, and the recently developed model NasNet [39].

Audio representation. For the acoustic modality, mel-frequency cepstral coefficient (MFCC) [5] is widely used in many audio-related tasks [7, 36]. In this paper, MFCC feature is also used for audio representation. Specifically, for a given audio file, the length of the sampling window is set to 25 milliseconds and meanwhile the step between successive windows is set to 10 milliseconds. In this way, there will be 100 MFCC features per second. To reduce the feature dimension, we take the average of the MFCC feature every second. Since there are two channels in the audio file, we first extract the MFCC features for each channel and then concatenate them together. As a result, the dimension of the MFCC feature for a given audio signal is $T \times 26$, where T is the length of a audio signal. In addition to MFCC feature, we also use NSynth [6] to encode the audio signals. NSynth is a recently developed WaveNet-style [25] auto-encoder model. Concretely, we take audio fragment every 5 seconds as input into NSynth and get the output of the encoder as the audio representation.

4.2 Proposed Multimodal Deep Learning Models

Our proposed multimodal deep learning model for video attractiveness prediction consists of three layers, namely the context gating layer, the multimodal fusion layer, and the sequential prediction layer.

Context gating layer. In order to further enrich the representative properties of the visual and audio features, context gating is used, which is shown to be beneficial to video representation learning [24]. Context gating is formulated as:

$$\hat{X} = \sigma(WX + b) \odot X,$$

where X is the input feature vector, which can either be visual or audio representation. σ is the element-wise sigmoid activation function. $\odot$ denotes the element-wise multiplication. $\hat{X}$ is the gated representation. It can be observed that context gating acts like a sentinel, which can adaptively decide which part of the input feature is useful. Moreover, with multiplication, the original representation X and the transformed representation $\sigma(Wx + b)$ are nonlinearly fused together, thus enhancing and enriching their representative abilities.

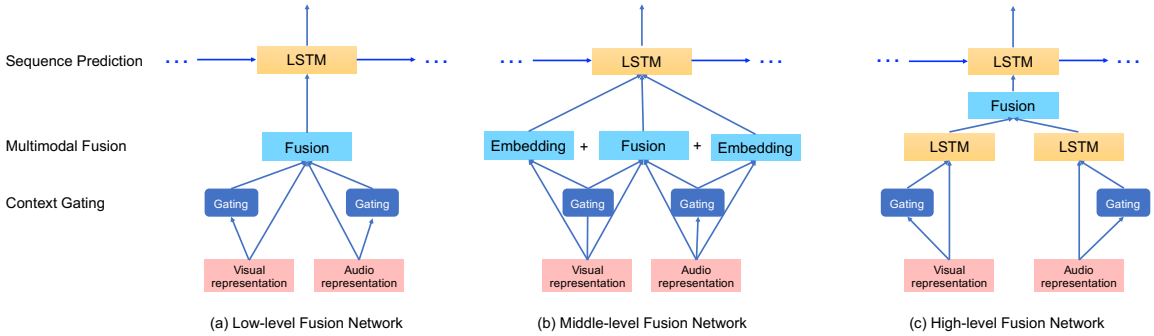

Figure 5: An overview of our video attractiveness prediction framework. Context gating is first applied to the visual and audio representations to enrich their corresponding representations. Based on the gated representations, different multimodal fusion strategies performed at different levels are used to exploit the relationships between visual and audio components. Finally, LSTM acts as the prediction layer to make the attractiveness prediction.

Multimodal fusion layer. Video contents consist of both visual and audio information, which are complementary to each other for the video representation learning [1]. Therefore, in this paper, we propose several multimodal fusion models to exploit the relationships between the gated visual and audio features to yield the final video representation. As illustrated in Fig. 5, three different multimodal fusion layers performed at different levels are proposed to yield the video representation for the final attractiveness prediction.

Low-level fusion. Fig. 5 (a) illustrates the low-level fusion layer. Specifically, we directly concatenate the visual and audio features after the aforementioned context gating layer and project them into a common space with a single embedding layer. As such, the low-level fusion strategy allows the visual and audio features to be fused at low levels. However, the contributions of visual and audio modalities are not equal. Normally, visual components will present more semantic information than audio. Simply concatenating them together may make the audio information be concealed by the visual part.

Middle-level fusion. To tackle the information concealment problem, we propose a middle-level fusion layer to learn the comprehensive representations from the two modalities. The architecture is shown in Fig. 5 (b). Specifically, we transform the gated visual and audio features with non-linear operations into three independent embeddings: the visual embedding, the audio embedding, and the joint embedding. The joint embedding captures the common semantic meanings between visual and audio modalities, while the visual and audio embeddings capture the corresponding independent semantic meanings.

High-level fusion. Furthermore, to fully exploit the temporal relations among the representations at every time step, we propose a more effective fusion method which is termed as the high-level fusion layer. As illustrated in Fig. 5 (c), we take two individual long short-term memory (LSTM) networks to encode the features of visual and audio data into the higher-order representations, which are further fused together as the video representation for the attractiveness prediction. With two different yet dependent LSTMs employed to learn the complicated behaviors within each individual modality, the semantic meanings carried by visual and audio components are extensively discovered, which is expected to benefit the final video attractiveness prediction.

Sequential prediction layer. After we obtain the multimodal embedding with both visual and audio components considered, we use a sequential prediction network to estimate the video attractiveness. More specifically, we take the output of the multimodal fusion layer x_t at t-th time step as input of another LSTM for prediction. We formulate the prediction process as follows:

$$h_t = \text{LSTM}(x_t, h_{t-1}). \tag{1}$$

The LSTM transition process is formulated as follows:

$$\begin{aligned} \begin{pmatrix} i_t \\ f_t \\ o_t \\ g_t \end{pmatrix} &= \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \mathbf{T} \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix}, \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t, \\ h_t &= o_t \odot \tanh(c_t), \\ y' &= W_o(h_t), \end{aligned} \tag{2}$$

where i_t, f_t, o_t, c_t, h_t, and σ are input gate, forget gate, output gate, memory cell, hidden state, and sigmoid function, respectively. $\mathbf{T}$ is a linear transformation matrix. $\odot$ represents an element-wise product operator. The hidden state h_t is used to predict a value y' as video attractiveness at fine-grained levels through a linear transformation layer W_o.

4.3 Training

Mean squared error (MSE) is a widely used as the objective function in sequence prediction tasks, which can be formulated as follows:

$$\mathcal{L}_{\text{MSE}} = \sum_{i=1}^{T} (y'_i - y_i)^2. \tag{3}$$

y'_i is the attractiveness value predicted by our model. y_i is the ground truth attractiveness (view counts). T is the fragment length of the video clip. Then we can use gradient descent methods to train the whole model in an end-to-end fashion.

5 EXPERIMENTS

In this section, we first introduce the experiment settings, including the data processing, evaluation metrics, baselines, as well as our

Table 2: Performance comparisons of our proposed multimodal deep learning models with different visual and audio representations, as well as their combinations. The best performance (except LSTM-EGG) for each metric entry is highlighted in boldface.

Model Name	SRCC (ρ)	MAE	RMSE	RMSLE
LSTM-EGG	0.795	0.381	0.499	0.039
LSTM-AUD-MFCC [5]	0.210	0.600	0.775	0.076
LSTM-AUD-NSynth [6, 25]	0.213	0.606	0.802	0.082
LSTM-VIS-VGG-16 [29]	0.323	0.572	0.726	0.069
LSTM-VIS-VGG-19 [29]	0.322	0.569	0.725	0.067
LSTM-VIS-ResNet-152 [8]	0.241	0.602	0.773	0.075
LSTM-VIS-NasNet-large [39]	0.359	0.570	0.724	0.069
LSTM-VIS-Inception-V1 [33]	0.336	0.570	0.719	0.066
LSTM-VIS-Inception-V2 [10]	0.337	0.569	0.724	0.067
LSTM-VIS-Inception-V3 [34]	0.335	0.571	0.725	0.068
LSTM-VIS-Inception-V4 [32]	0.365	0.567	0.713	0.067
Low-level fusion (Inception-V4+MFCC)	0.313	0.580	0.740	0.070
Low-level fusion (Inception-V4+NSynth)	0.243	0.601	0.793	0.079
Middle-level fusion (Inception-V4+MFCC)	0.330	0.575	0.731	0.069
Middle-level fusion (Inception-V4+NSynth)	0.318	0.573	0.733	0.070
High-level fusion (Inception-V4+MFCC)	0.387	0.562	0.708	0.066
High-level fusion (Inception-V4+NSynth)	0.371	0.551	0.698	0.063
Ensemble of high, middle and low level fusion (Inception-V4+MFCC)	**0.401**	0.554	0.699	0.065
Ensemble of high, middle and low level fusion (Inception-V4+NSynth)	0.393	**0.544**	**0.690**	**0.062**

implementation details. Afterward, we will illustrate and discuss the experimental results.

5.1 Experimental Settings

Data processing. To keep the diversity of training samples, for episodes in each category[2], we use 70% for training, 20% for testing and 10% for validation. Recall that the average duration of videos in FVAD is 45 minutes, which is difficult for LSTM to model such long video sequence due to the capacity limitations of LSTM. Therefore, we divide each video in the training set into a series of non-overlapping video clips with the length of 5 minutes. However, during the testing phase of our model, we take the video as a whole into the prediction model without any partitioning.

Evaluation metrics. To evaluate the performance of fine-grained video attractiveness prediction, we adopted mean absolute error (MAE), root mean square error (RMSE) and root mean squared logarithmic error (RMSLE). Besides, as in [16], we adopt Spearman rank-order correlation coefficient (SRCC) to evaluate the correlation between the video attractiveness predicted by our model and the true values. According to the definitions, larger SRCC value and smaller MAE, RMSE, and RMSLE values indicate more accurate predictions, demonstrating a better performance.

Baselines. The framework of our baseline models is similar to the model illustrated in Fig. 5 (a). The only difference is that the baseline model only takes one kind of feature as input. More specifically, given any types of representation X, we first transform X into an embedding vector with dimension size of 512. Then the embedding vector is input into the sequence prediction layer to estimate the video attractiveness. In our experiments, *LSTM-EGG* represents the model which predicts the attractiveness with 9 engagement indicators. *LSTM-AUD-*$*$ and *LSTM-VIS-*$*$ are the baseline models which only take the audio and the visual representations as input, respectively.

Implementation details. In this paper, the hidden unit size of LSTM are all set to 512. We train the model with the adam optimizer by a fixed learning rate 5×10^{-4}. The batch size is set as 16. And the training procedure is terminated with early stopping strategy when value of (3 × SRCC − MAE − RMSE − RMSLE) reaches the maximum value on the validation set.

5.2 Results and Discussions

The experimental results are illustrated in Table 2. Different video and audio representations, as well as their variant combinations, are used to perform the visual attractiveness prediction.

Recall that in Section 3 we verified that there indeed exists correlations between video attractiveness and other user engagement indicators. To investigate the combined effect of all engagement indicators, we show the performance of LSTM-EGG. We observed that LSTM-EGG obtains the best result which indicates that users' engagement behaviors as a whole shows a strong correlation with video attractiveness (view counts). This also validates that the features developed from engagement domain are much discriminative, even though they are of low-dimension. However, such features are not available for practical applications. That is also the main reason why we resort to the content features, specifically the visual and audio contents, for video attractiveness prediction.

Through the comparison among LSTM-AUD-$*$, LSTM-VIS-$*$ and different fusion methods, it is observed that visual features are more useful than audio features for video attractiveness prediction. Moreover, by incorporating more modalities, better performances can be obtained. This implies the complementary relationships rather than mutual conflicting relationships between the visual and audio modalities. To further examine the discriminative properties of the audio and visual features, we conduct experiments over

[2]A season of TV series can be seen as a category in this scenario.

different kinds of features using the proposed model. The general trend is that the more powerful the visual or audio features, the better performance it obtained. Specifically, the visual features in the form of NasNet and Inception-X are more powerful than those of VGG.

It is obvious that high-level fusion performs much better than low-level fusion methods. Regarding the low-level fusion, features extracted from various sources may not fall into the same common space. Simply concatenating all features actually brings in a certain amount of noise and ambiguity. Besides, low-level fusion may lead to the curse of dimensionality since the final feature vector would be of very high dimension. High-level fusion methods introduce two separate LSTMs to well capture the semantic meanings of the visual and audio content, respectively, which thus make a more comprehensive understanding of video contents. Additionally, the ensemble results among all levels of fusion achieve the best performance, which demonstrating that ensembling different level fusion models can comprehensively exploit the video content for attractiveness prediction.

6 CONCLUSIONS

In this paper, we built to date the largest benchmark dataset, dubbed FVAD, for tackling the emerging fine-grained video attractiveness prediction problem. The dataset was collected from a real-world video website. Based on FVAD, we first investigated the correlations between video attractiveness and nine user engagement behaviors. In addition, we extracted a rich set of attractiveness oriented features to characterize the videos from both visual and audio perspectives. Moreover, three multimodal deep learning models were proposed to predict the fine-grained fragment-level attractiveness relying solely on the video contents. Different levels of multimodal fusion strategies were explored to model the interactions between visual and audio modalities. Experimental results demonstrate the effectiveness of the proposed models and the necessity of incorporating both the visual and audio modalities.

REFERENCES

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

[2] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. 2005. Actions as Space-Time Shapes. In *ICCV*.

[3] Jingyuan Chen, Xuemeng Song, Liqiang Nie, Xiang Wang, Hanwang Zhang, and Tat-Seng Chua. 2016. Micro Tells Macro: Predicting the Popularity of Micro-Videos via a Transductive Model. In *ACM Multimedia*.

[4] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive Collaborative Filtering: Multimedia Recommendation with Item and Component-Level Attention. In *SIGIR*.

[5] Steven Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing* (1980).

[6] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. 2017. Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. *arXiv preprint arXiv:1704.01279* (2017).

[7] John N. Gowdy and Zekeriya Tufekci. 2000. Mel-scaled discrete wavelet coefficients for speech recognition. In *ICASSP*.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. In *CVPR*.

[9] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *CVPR*.

[10] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *ICML*.

[11] Yugang Jiang, Qi Dai, Xiangyang Xue, Wei Liu, and Chong-Wah Ngo. 2012. Trajectory-based modeling of human actions with motion reference points. In *ECCV*.

[12] Yugang Jiang, Jingen Liu, Amir Roshan Zamir, Ivan Laptev, Massimo Piccardi, Mubarak Shah, and Rahul Sukthankar. 2013. THUMOS Challenge: Action Recognition with a Large Number of Classes. http://crcv.ucf.edu/ICCV13-Action-Workshop/. (2013).

[13] Yugang Jiang, Yanran Wang, Rui Feng, Xiangyang Xue, Yingbin Zheng, and Hanfang Yang. 2013. Understanding and Predicting Interestingness of Videos. In *AAAI*.

[14] Yugang Jiang, Zuxuan Wu, Jun Wang, Xiangyang Xue, and Shih-Fu Chang. 2015. Exploiting Feature and Class Relationships in Video Categorization with Regularized Deep Neural Networks. *arXiv preprint arXiv:1502.07209* (2015).

[15] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale Video Classification with Convolutional Neural Networks. In *CVPR*.

[16] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. 2014. What makes an image popular?. In *WWW*.

[17] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. 2011. HMDB: a large video database for human motion recognition. In *ICCV*.

[18] Laptev and Ivan. 2005. On space-time interest points. *International journal of computer vision* 64, 2-3 (2005), 107–123.

[19] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. 2016. TGIF: A new dataset and benchmark on animated GIF description. In *CVPR*.

[20] Feng Liu, Yuzhen Niu, and Michael Gleicher. 2009. Using Web Photos for Measuring Video Frame Interestingness. In *IJCAI*.

[21] Wei Liu, Zhifeng Li, and Xiaoou Tang. 2006. Spatio-temporal embedding for statistical face recognition from video. In *ECCV*.

[22] Marcin Marszałek, Ivan Laptev, and Cordelia Schmid. 2009. Actions in Context. In *CVPR*.

[23] K Messer, J Matas, J Kittler, J Luettin, and G Maitre. 1999. XM2VTSDB: The Extended M2VTS Database. In *Second International Conference on Audio and Video-based Biometric Person Authentication*.

[24] Antoine Miech, Ivan Laptev, and Josef Sivic. 2017. Learnable pooling with Context Gating for video classification. *arXiv preprint arXiv:1706.06905* (2017).

[25] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[26] Henrique Pinto, Jussara M. Almeida, and Marcos A. Gonçalves. 2013. Using Early View Patterns to Predict the Popularity of Youtube Videos. In *WSDM*.

[27] Marian-Andrei Rizoiu, Lexing Xie, Scott Sanner, Manuel Cebrian, Honglin Yu, and Pascal Van Hentenryck. 2017. Expecting to Be HIP: Hawkes Intensity Processes for Social Media Popularity. In *WWW*.

[28] Yuesong Shen, Claire-Hélène Demarty, and Ngoc Q. K. Duong. 2016. Technicolor@MediaEval 2016 Predicting Media Interestingness Task. In *MediaEval*.

[29] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556* (2014).

[30] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. In *arXiv:1212.0402*.

[31] Gábor Szabó and Bernardo A. Huberman. 2010. Predicting the popularity of online content. *Commun. ACM* 53, 8 (2010), 80–88.

[32] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex A. Alemi. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *ICLR Workshop*.

[33] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper with Convolutions. In *CVPR*.

[34] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *CVPR*.

[35] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The New Data in Multimedia Research. *Commun. ACM* 59, 2 (Jan. 2016), 64–73.

[36] George Tzanetakis and Perry Cook. 2002. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing* 10, 5 (2002), 293–302.

[37] Bin Wu, Erheng Zhong, Ben Tan, Andrew Horner, and Qiang Yang. 2014. Crowdsourced time-sync video tagging using temporal and personalized topic modeling. In *SIGKDD*.

[38] Sejong Yoon and Vladimir Pavlovic. 2014. Sentiment Flow for Video Interestingness Prediction. In *Proceedings of the 1st ACM International Workshop on Human Centered Event Understanding from Multimedia (HuEvent '14)*.

[39] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. 2017. Learning transferable architectures for scalable image recognition. *arXiv preprint arXiv:1707.07012* (2017).