Prediction of instantaneous likeability of advertisements using deep learning

Dipayan Saha¹ | S.M. Mahbubur Rahman¹ | Mohammad Tariqul Islam² | M. Omair Ahmad³ | M.N.S. Swamy³

¹Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
²Department of Electrical Engineering, Princeton University, Princeton, New Jersey, USA
³Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada

Correspondence
S.M. Mahbubur Rahman, Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh.
Email: mahbubur@eece.buet.ac.bd

Funding Information
NVIDIA Corporation

Abstract
The degree to which advertisements are successful is of prime concern for vendors in highly competitive global markets. Given the astounding growth of multimedia content on the internet, online marketing has become another form of advertising. Researchers consider advertisement likeability a major predictor of effective market penetration. An algorithm is presented to predict how much an advertisement clip will be liked with the aid of an end-to-end audiovisual feature extraction process using cognitive computing technology. Specifically, the usefulness of different spatial and time-domain deep-learning architectures such as convolutional neural and long short-term memory networks is investigated to predict the frame-by-frame instantaneous and root mean square likeability of advertisement clips. A data set named the ‘BUET Advertisement Likeness Data Set’, containing annotations of frame-wise likeability scores for various categories of advertisements, is also introduced. Experiments with the developed database show that the proposed algorithm performs better than existing methods in terms of commonly used performance indices at the expense of slightly increased computational complexity.

1 INTRODUCTION

Over the last decade, there has been a tremendous increase in multimedia-based advertisements on the internet. The Interactive Advertising Bureau reports that in the United States, internet advertising generated revenue of $17.9 billion in 2017, which was more than that of TV advertising [1]. Total online digital advertisement revenue witnessed a 33% increase in 2017 from the previous year [1]. The incremental advancement of communication technology, the rapid growth of media content in social media, and innovative advertising ideas have contributed significantly to shaping the trend of digital marketing. Billions of dollars are being spent by advertising agencies to reach their desired customers. Advertising research has a history of more than 50 years. Over this time span, researchers from related fields have reached a consensus that the success and likeability of an advertisement are highly correlated [2, 3].

1.1 Likeability of advertisement

Marketing researchers analyse the effectiveness of advertisements from various perspectives. For example, Wells et al. [4] define the effectiveness of advertising as a measurement of sales volume. A commercial is said to be effective if the money spent on it is returned by customers in increased sales. Several researchers have investigated the four factors that influence advertising success [2, 5, 6]: cognition (conscious thoughts and attention) [7–9], effect (liking and preference) [10], memory (two retrieval measures—recall and recognition) [11, 12], and desirability (intent to purchase) [10].

In the past, customer perceptions about the product and brand were the prime concern of an advertisement [13]. Nowadays, however, a commercial is both for selling the product and a source of entertainment to gain customers’ attention. Factors such as a unique presentation concept, a novel way of storytelling, advanced cinematography, ease of
understanding, and soothing audiovisual stimuli have emerged as driving forces to make an effective advertisement that ultimately leads to product purchases. Digital marketing strategies are evolving in the methods they use to grab customers' attention because viewers usually do not want to waste their valuable time watching commercials [14, 15]. Sometimes an advertisement is regarded as an unwelcome intrusion and irritation [16]. Hence, advertisement likeability is believed to be an effective strategy for eliminating such a situation [17]. Studies describe advertisement likeability as a significant predictor of effectiveness (see, e.g. [2]). Biel et al. [7] demonstrate that likeable advertisements possess more than twofold the effectiveness of traditional commercials. Apart from such an association with advertisement effectiveness, another positive aspect of advertisement likeability is that it is measurable [18]. Because of such significance, researchers have enquired about what drives advertisement likeability. It is understood that creativity [19], meaningfulness and relevance [7], and special elements [6] like characters, action, storyline, music, and visual elements are linked to advertisement likeability. Product category [7], culture [20], and celebrity endorsement [21] can also have an impact on advertisement likeability.

In marketing research, the purchase intent is also an important yardstick that refers to the willingness of consumers to purchase a product. Purchase intent not only measures the desirability of a customer but also explains market behaviour. Psychological studies propose four models, namely, the affective transfer, dual mediation, reciprocal mediation, and independent influences hypotheses, to explain how advertisement likeability impacts purchase intent [21]. Because advertisement likeability is related to both effectiveness and purchase intent, its accurate prediction can be a common solution to generate an effective and purchase-driven advertisement prior to market launch.

Prediction of advertisement likeability is a challenging task, as the decision-making process involves non-linear elements. Whether viewers will like an advertisement video or not is a question without an obvious answer. But it is clear that many uncontrollable factors are involved in the process of understanding viewer perception. Because media contents such as video, audio, image, text, graphics are closely related to the digital advertisement, a non-linear relationship about advertisement likeability might be hidden between these contents. The best possible way to predict the likeability from media contents is to adopt an algorithm that mimics the human brain's reasoning process, similar to cognitive computing, to identify the pattern hidden in the vast amount of data. Thus, cognitive computation that includes machine learning, deep learning, and sentiment analysis can simulate critical thought processes by understanding the contents and predicting the likeability of advertisement videos.

1.2 Related works

There are three approaches to measuring advertisement likeability, namely, self-reported, response-based, and content-based. In self-reported estimation, marketers collect the overall opinions of customers through a questionnaire. This estimation approach is time-consuming, and the precision of such estimation is highly dependent on the sampling of customers. Response-based estimation of advertisement likeability employs two types of reaction measures, namely, physiological response and facial response. The physiological reactions include those obtained from eye engagement [22], heart rate [23], respiration rate, skin conductance, and neurophysiological monitoring (e.g. electroencephalogram [10, 24] and functional magnetic resonance imaging [10]). To understand consumer behaviour and advertising phenomena, a significant number of research studies have been conducted by combining different types of physiological responses [25–28]. For example, Venkatraman et al. [10] conducted an experiment that shows advertisements to participants and measures the overall effectiveness not only from their self-reports but also from their neurophysiological responses. It is noted that the measurements of likeability from neurophysiological response are expensive and require the full cooperation of customers and thus can only be performed in copy testing or pretesting.

Investigations are carried out to find the relationship between advertisement likeability and the facial responses of viewers. For example, in [3], McDuff et al. demonstrated that facial response to online advertisements can predict the consumer's liking and desire for further viewing by analysing only a few. Later, McDuff et al. [29] performed a large-scale analysis of facial responses and advertisements to model the relationship between the viewer's purchase intent and facial responses. Teixeira [30] et al. assessed the concentration of attention through eye-tracking and retention of viewers by recording their zapping behaviour. This study reports that surprise and joy effectively concentrate the attention and retain the viewers. Recently, Okada et al. [31] have collected a number of facial reaction videos watching advertisements, integrated the facial expression and physiological responses such as the heart rate, and then estimated the effectiveness of an advertisement. The accuracy of estimation of advertisement likeability using facial response may be compromised because of the introverted nature of humans, in which case, changes in facial appearance may be negligible.

Forbes, a leading business magazine, describes the visual and auditory properties of the advertisement as the main strengths of marketing [32]. Intuitively, the visual properties like the colour, intensity, arrangement of pixels, and image textures and acoustic properties like the pitch, rhythm, loudness, speech, beat arrangement, sonic textures should be investigated to find how these factors can aid to determine advertisement likeability. In this context, it can be noted that a number of research studies are conducted on the prediction of personality traits and popularity using the contextual information of the media. For example, Biel et al. [33] used multimodal cues (audio and video) for the automatic prediction of personality traits of vloggers. In [34], the popularity of music was predicted in terms of eight popularity
metrics using mel-frequency cepstrum coefficient into a trained support vector machine classifier. Tomasz et al. [35] proposed a support vector regression (SVR) method for predicting the popularity of online media using convolution neural network (CNN)-based visual features. All these works have adopted content-based approaches for social behaviour analysis and popularity prediction. Thus, it can be inferred that content-based estimation can be a better approach to predict the likeability of advertisement. To the best of our knowledge, the content-based estimation of advertisement likeability is rarely attempted. This paper presents how the visual and audio features can determine the likeability of an advertisement.

Cognitive computation includes technologies like machine learning, natural language processing, neural networks, and deep learning. With the advent of deep learning, end-to-end feature extraction through CNN is being practised widely in miscellaneous applications. Zhou et al. adopted deep CNN architecture with improved softmax loss for face and expression recognition [36]. In [37], Basnet et al. estimated instantaneous emotional states from facial video using audiovisual features extracted from CNNs. Later, they introduced a two-stream architecture to learn the audiovisual features simultaneously [38]. Long short-term memory (LSTM) [39], a special type of recurrent neural network (RNN), is also being practised by researchers to address different problems of sequences, especially for prediction of time series data. Gharibshah et al. [40] proposed an LSTM-based framework to predict the probability of a user clicking on a website of advertisement either independently or in-campaign from the chronological sequence of historically visited websites. The architecture of CNN and LSTM are also combined together, in which the former is used for extraction of audiovisual features and the latter is for supporting sequence prediction [41, 42]. For example, to estimate the affective states, Brady et al. [43] extracted features from the end-to-end CNN and then fed them to the RNN. In [44], the same problem is addressed in a similar manner but with different objective functions and structures of the CNN and LSTM. Gao et al. [45] proposed an attention-based LSTM model with semantic consistency for video captioning.

Thus, there remains a scope to investigate the performance of the CNN-LSTM-based network for predicting the frame-by-frame likeability of advertisement clips from audiovisual contents. Such an algorithm will facilitate the advertiser not only to evaluate the overall likeability of a video clip using a root mean square (RMS) value but also to analyse the clip at frame level during copy test [46] to increase the possibility of the success of the advertisement. Event level likeability score can also be estimated from such frame-level prediction of likeability of advertisement. In addition, to the best of our knowledge, there is no publicly available data set that deals with frame-level likeability scores of advertisement clips. Hence, a comprehensive advertisement data set with annotated likeability scores can be developed for evaluating the performance of the proposed network.

1.3 | Specific contributions

The specific contributions of this paper are threefold:

- Developing a CNN-LSTM-based architecture for predicting frame-by-frame as well as RMS likeability of advertisement clips using audiovisual content.
- Introducing a publicly released new data set of advertisement clips that has frame-level annotation of likeability scores.
- Experiments of the CNN-based architectures on the database developed to evaluate the performance of estimation of instantaneous likeability of advertisement.

The remainder of the paper is organized as follows. Section 2 presents the proposed deep-learning architectures. In Section 3, the details of the data set developed are given. The training scheme, evaluation process, and the results of the experiments are given in Section 4. Finally, Section 5 provides the conclusion.

2 | PROPOSED METHOD

In the proposed method, a feed-forward CNN is trained to extract features from the audiovisual clips of an advertisement. The proposed CNN has two streams—one stream for the extraction of visual features and another for the audio features. Because of the unique nature of the audio and visual contents, the two streams are designed with different structures. Estimation of the frame-level likeability of advertisement clips is formulated as a time series prediction. Hence, a two-layer LSTM, which uses the CNN-based features, is proposed with the view that the LSTM is capable of looking at the recent past data to identify a pattern. In this section, the proposed CNN and LSTM networks are described first. Next, the details of the regression layer, loss function, and optimization process are described.

2.1 | CNN architecture

The function of two-stream CNN is to extract features from the audiovisual contents in an end-to-end manner and feed them to the following LSTM network. Figure 1 provides a stick diagram of the CNN-based feature extraction layers. The visual and audio streams of the CNN are clearly marked in this figure.

2.1.1 | Visual stream

The visual stream has four stages. The first stage starts with a convolution layer followed by a ReLU layer, a batch normalization layer, and a max-pooling layer. Initially, the convolution operation extracts the high-level features. Afterwards, ReLU is placed to add non-linearity to the model. Batch normalization speeds up the learning and reduces the
overfitting. The Max-pooling layer selects the dominant features and decreases the computational cost through dimensionality reduction. Let $V_0$ represent an input frame of a video. The output of the first stage, denoted by $V_1$, can be written as

$$V_1 = MP(BN(max(0, V_0 \ast W_{1v} + b_{1v})))$$ (1)

where $\ast$, $max(0, \cdot)$, $MP(\cdot)$, and $BN(\cdot)$ denote the convolution operation, ReLU operation, max-pooling operation, and batch normalization, respectively. The terms $W_{kv}$ and $b_{kv}$ imply the weight matrix and bias vector at the $k$-th stage, respectively, where the subscripts $k$ and $v$ refer to the stage number and visual stream, respectively. The second stage is identical to the first stage in terms of functionality, and thus, the output of this stage $V_2$ can be written as

$$V_2 = MP(BN(max(0, V_1 \ast W_{2v} + b_{2v})))$$ (2)

In the third and fourth stages, there exist two back-to-back fully connected layers that are different in terms of dimension. The ReLU acts as an activation function for these two layers. The output of these stages is given by

$$V_3 = max(0, W_{3v}^T V_2 + b_{3v})$$ (3)

$$V_4 = max(0, W_{4v}^T V_3 + b_{4v})$$ (4)

where the superscript $T$ represents the transpose of a matrix, and $V_4$ represents the CNN-based extracted feature vector for the visual content.

2.1.2 | Audio stream

In the audio stream, the first stage commences with a convolution layer followed by a max-pooling layer. Let $A_0$ represent an input audio vector corresponding to a single frame. The output of this stage $A_1$ can be written as

$$A_1 = MP(A_0 \ast W_{1a} + b_{1a})$$ (5)

where $W_{ka}$ and $b_{ka}$ denote the weight matrix and bias term at the $k$-th stage, and the subscript $a$ refers to audio stream. Similarly, the output of the second stage $A_2$ can be written as

$$A_2 = MP(0, A_1 \ast W_{2a} + b_{2a})$$ (6)

The third and fourth stages of the audio stream are identical to that of the visual stream. Thus, the output of these stages is given by

$$A_3 = max(0, W_{3a}^T A_2 + b_{3a})$$ (7)

$$A_4 = max(0, W_{4a}^T A_3 + b_{4a})$$ (8)

where $A_4$ represents the CNN-based extracted feature vector for the audio content.

2.1.3 | Feature construction

A merging process is required to combine the video and audio features. This fused feature vector is obtained as

```plaintext
```

FIGURE 1 A stick diagram of the proposed two-stream convolutional neural network architecture. The operations from left to right are considered the forward path. Each operation or layer in the network is represented by its corresponding geometric shape. The convolution, rectified linear unit (ReLU), batch normalization, max-pooling, fully connected, and concatenation layers are represented by rectangle, solid rectangle, multistriped rectangle, converging trapezoid, emerald shape, and single-striped rectangle, respectively.
\[ s = V_4 || A_4 \]  

where \( \| \) denotes the concatenation operation. To predict the frame-wise likeability score of a video, the fused features are fed into the LSTM network.

### 2.2 | LSTM network

Let \( x^n \) be audiovisual feature vector obtained from CNN at the frame number \( n \), and \( N \) be the number of feature vectors obtained from the video frames between the frame number \( n - (N - 1) \) and \( n \) that are fed into the proposed two-layer LSTM network. Let \( x^n \) represent the input to the first layer of the LSTM network given by

\[ x^n = [z^{n-(N-1)} x^{n-N+2} ... x^{n-2} x^{n-1} x^n] \]  

(10)

Each of the two layers of the proposed LSTM network has three gates and two cell states. The output of the \( j \)-th layer is given by

\[ b_j^n = o_j^n \circ \tanh(c_j^n) \]  

(11)

**Output Gate:** \( o_j^n = \sigma(W_j^n b_j^{n-1} + U_j^n b_j^{n-1} + R_j^n c_j^{n-1} + b_j^n) \)  

(12)

**Cell State:** \( c_j^n = (o_j^n \circ c_j^n) + (f_j^n \circ c_j^{n-1}) \)  

(13)

**Input Gate:** \( f_j^n = \sigma(W_j^n b_j^{n-1} + U_j^n b_j^{n-1} + R_j^n c_j^{n-1} + b_j^n) \)  

(14)

**Forget Gate:** \( c_j^n = \sigma(W_j^n b_j^{n-1} + U_j^n b_j^{n-1} + R_j^n c_j^{n-1} + b_j^n) \)  

(15)

**Intermediate Cell State:** \( c_j^n = \tanh(W_j^n b_j^{n-1} + U_j^n b_j^{n-1} + b_j^n) \)  

(16)

where \( j \in (1, 2) \), \( b_0^n = x^n \), and \( \sigma \), \( \tanh \) denote element-wise multiplication, sigmoid, and tangent activation functions, respectively. \( W, U, \) and \( R \) represent weight vectors, and \( b \) indicates bias vector. Figure 2 shows an overview of the proposed LSTM network, where it can be seen that the output of the first layer is fed into the input to the second layer.

### 2.3 | Regression layer

Regression layer that consists of a fully connected layer and a smoothing layer operates on the output of the LSTM network.

The fully connected layer uses a linear activation function, and the output is given by

\[ y_o^n = W_r^T b_2^n + b_r \]  

(17)

where \( W_r \) and \( b_r \) represent weight matrix and bias terms, respectively.

In practice, a frame of advertisement clip spans around 30-50 milliseconds. Intuitively the likeability of an advertisement does not change instantly, rather to take a decision about the degree of likeability, viewers require near about 1.5–2.5 s. This is why a smoother is required that can minimize the rapid fluctuation of the output of the fully connected layer and provide fairly smooth patterns of likeability score for approximately two seconds. Thus, the proposed architecture includes a smoothing layer that discards maximum 10% and minimum 10% values for every 50 frames, finds the RMS value from the remaining scores, and sets the RMS value as the final output for those 50 frames. Thus, the estimated likeability score is given by

\[ \hat{y}^n = \Phi(y_o^n) \]  

(18)

where \( \Phi \) denotes the proposed smoothing operation. Figure 3 shows the overall block diagram of the proposed method, wherein the CNN, LSTM network, and regression layers are clearly outlined.

### 2.4 | Loss function

The proposed CNN and LSTM networks are trained separately with the aid of a fully connected layer at the output. The fully connected layer with the LSTM network is kept in the regression layer. In training, a loss function is required to be defined. Because the purpose is not only to minimize the distance between estimated score and ground truth but also to track the pattern of the variation in likeability score, in the proposed method, the loss function is defined based on the combination of concordance correlation coefficient (CCC) \([47]\) and mean squared error (MSE) \([44]\) given by

\[ L = (1 - \text{CCC}) + \text{MSE} \]  

(19)

\[ \text{CCC} = \frac{2 \rho \sigma \sigma_{\hat{y}}}{\sigma^2 + \sigma_{\hat{y}}^2 + (\mu - \mu_{\hat{y}})^2} \]  

(20)

\[ \text{MSE} = \frac{1}{N} \sum_{n=1}^{N} (y^n - \hat{y}^n)^2 \]  

(21)
where $y^n$ and $\hat{y}^n$ represent the ground truth and estimated scores of likeability, respectively, $\mu^c$ and $\mu^\hat{c}$ are the mean values, $\sigma^c$ and $\sigma^\hat{c}$ are the variances and $\rho^c$ is the correlation coefficient between $y^n$ and $\hat{y}^n$, and $N$ denotes the number of samples.

### 2.5 Optimization

In the training phase, the well-known Adam stochastic optimization algorithm [48] is employed. In this algorithm, the trainable parameters are updated based on the first moment $m$ and second moment $v$ of the gradient of the loss function $L$ with respect to weight $w$. For a given iteration $\eta$, the update process is given by

$$m(\eta) = \beta_1 m(\eta - 1) + (1 - \beta_1) \frac{dL(\eta)}{dw(\eta)} \frac{1}{1 - \beta_1^m} \quad (22)$$

$$v(\eta) = \beta_2 v(\eta - 1) + (1 - \beta_2) \left( \frac{dL(\eta)}{dw(\eta)} \right)^2 \frac{1}{1 - \beta_2^v} \quad (23)$$

$$w(\eta) = w(\eta - 1) - \alpha \frac{m(\eta)}{\sqrt{v(\eta)} + \epsilon} \quad (24)$$

where $\alpha (\alpha > 0), \beta_1, \beta_2 (0 < \beta_1, \beta_2 < 1)$ denote the learning rate and decay values for the first and second moments, respectively, and $\epsilon (\epsilon > 0)$ is a numerical stability factor.

### 3 DATA SET

Due to the absence of a publicly available data set containing annotation of frame-wise likeability scores for advertisement clips, a new data set is compiled. The data set is named as
‘BUET Advertisement Likeability Data set’, which can be accessed from https://urlz.com/uMm9S with the password ‘BUETSMMR’. The details of the advertisement clips, their annotation process, and the characterization of annotation are presented in the following sections.

3.1 Advertisement clips

The data set includes 25 Bangladeshi advertisement clips annotated by raters. These clips are publicly available on YouTube and televised on more than 20 TV channels in Bangladesh. Because all the raters are native speakers, only Bangladeshi advertisement clips are chosen to ensure their full understanding and to connect to their natural flow of expression. The advertisement clips are classified into five different categories—entertaining, creative, emotional, humorous, and miscellaneous, with each having unique characteristics. The categorization of a video is performed based on the overall impression obtained from the raters. Each category has a number of videos ranging from 4 to 6. The frame rate of the videos varies between 25 fps and 30 fps. The total number of frames in each category ranges from 5850 to 9550 with a duration of 224 to 370 s, respectively. The resolution of the frame of these videos is 1280 × 720. Summary statistics of the data set, including the total duration of videos, number of frames, and characteristics for each category of advertisement clip are given in Table 1.

3.2 Annotation process

There were 30 raters who participated in the annotation process. Each rater watched each advertisement clip with sound in a quiet environment. With permission, their reaction video to each advertisement was recorded using a web camera. Then, each rater was asked to watch the video again and to give any real number as a score of likeability in the range between 1 and 5 for every 2 s of the video scene. Based on the significant change in their facial expression in response videos, the best 20 raters were chosen. The final likeability score of a particular 2 s shot was obtained by averaging the annotations of these 20 raters. Such annotation process was inspired by Title-based Video Summarization (TVSum) data set that is a benchmark data set for video understanding and to connect to their natural flow of expression.

3.3 Interrater reliability

Interrater reliability is investigated to determine how much different raters agree in their annotations. Figure 4 displays the change in annotations among the videos for different raters. The figure shows the diversity of colour from light to dark, which indicates the diversity of the annotations. Along horizontal rows, certain cohesion is found among the raters for a single advertisement clip. This is a clear indication of homogeneity among raters. For objective measurement of internal consistency of raters, two evaluation metrics, Cronbach’s alpha [50] and Intraclass Correlation Coefficient (ICC) [51] are calculated. ICC is estimated through two-way random model and average measurements. The IBM SPSS software package [52] has been used to calculate these metrics. In [53], it is mentioned that 0.70 and 0.60 are minimum acceptable values of Cronbach’s alpha ICC, respectively. For our data set, these values are found to be 0.901 and 0.852, respectively. Hence, the developed data set shows the high consistency and agreement among raters that are effective in related research areas.

4 EXPERIMENTS AND RESULTS

To evaluate the performance of the proposed method for predicting of likeability of advertisement clips, the database is split into training and testing sets. Among the 25 videos, 5 are chosen randomly for testing purposes in a way such that a single video from each category is included in the testing set. During training, the networks are trained on 15 videos and while the remaining five videos are employed for validation purposes. To eliminate the randomness in the results obtained, fivefold cross-validation is performed. To elaborate, the evaluation process is conducted five times, and in each round, different training and validation videos are selected. This section describes the parameter settings of the CNN and LSTM network as well as that of the training process. Next, the comparison methods and the performance metrics employed to compare the methods are described. Finally, a comparison of results on the data set is presented.

4.1 Model setup

The values of the parameters of the proposed CNN and LSTM networks are chosen in accord with the dimensions of the samples in the data sets as well as those recommended in practice. As mentioned in Sections 2.1 and 2.2, the proposed method includes both a CNN and an LSTM network.

The input of the visual stream of the proposed CNN is considered to be RGB frames of 128 × 128 pixels. The number of filters in the convolution filters $W_{1v}$ and $W_{2v}$ and fully connected weights $W_{3v}$ and $W_{4v}$ of the video stream of the CNN and the corresponding bias terms $b_{1v}, b_{2v}, b_{3v},$ and $b_{4v}$ are set to 256, 512, 512, and 128, respectively. The kernel size of the convolution filters and the max-pooling operation is set to $3 \times 3$. The value of the stride in each max-pooling operation is 3. In effect, the total number of features extracted from the visual stream is 128.

The input audio vector for the audio stream is formed using the rectangular windowing method. The length of each audio input vector is equivalent to the duration of a video frame. Because the frame rate of the video files is 25 or 30 fps, the equivalent time duration of each audio vector varies according to the frame rate of the video, but the length of each
Table 1: Overall statistics of 'BUET advertisement likeability data set'

| Category     | Total Duration (s) | No. of Frames | Characteristics                                                      |
|--------------|--------------------|---------------|---------------------------------------------------------------------|
| Entertaining | 370                | 9550          | Have festive mode, celebration of life with music                   |
| Creative     | 328                | 8640          | Represent product in a unique way with outstanding cinematography    |
| Emotional    | 336                | 8400          | Create intense feeling                                              |
| Humorous     | 245                | 6400          | Create laughter with humorous dialogue or concept                    |
| Miscellaneous| 224                | 5850          | Nostalgic, awareness-spreading, celebrity endorsement                |

Figure 4: Process of annotation of personal likeability score for an advertisement clip

Figure 5: Characteristics of annotation

Audio vector is kept the same. For a video file with a frame rate of 25 fps, the input audio vector is formed using a rectangular window of 60 ms where 20 ms comes from the previous time step. For a video file with a frame rate of 30 fps, the duration of the window is 44 ms with an overlap of 11 ms from the previous time step. Just like a video stream, the number of
filters in the first two convolution filters $W_{1d}$ and $W_{2d}$ and the corresponding bias terms $b_{1d}$ and $b_{2d}$ are chosen to be 256 and 512, respectively. The kernel size for both convolution layers is chosen to be 20, and the kernel size in the max-pooling layer is chosen to be 10. The stride of each max-pooling operation is set to 10. The number of activations in two fully connected layers has been set to 128 and 64, respectively. In effect, the total number of features extracted from the audio stream is 64. The concatenation layer given by Equation (9) merges 128 visual and 64 audio features that results in the total number of features extracted from CNN being 192.

For LSTM network, a collection of 192 audiovisual features from current and previous time steps acts as input for the first layer of LSTM network. The value of $N$ in Equation (10) is chosen to be 10. The number of hidden units in the activation vectors of input gate, forget gate, intermediate cell state, and output gate is set to 192. The output of each LSTM layer given by Equation (11) contains 192 features. Each of the weight matrices $W_{i}^{f}, W_{i}^{o}, W_{i}^{u}, U_{i}^{f}, U_{i}^{o}, U_{i}^{u}, R_{i}^{f}, R_{i}^{o}$ and $R_{i}^{u}$ mentioned in Equations (11)–(16) has 192 $\times$ 192 parameters. Each of the corresponding bias terms $b_{i}^{f}, b_{i}^{o}, b_{i}^{u}, b_{i}^{f}$ and $b_{i}^{u}$ have 192 parameters.

The CNN and the LSTM networks have been trained sequentially. First, CNN is trained on the training data set by adding a dummy regression layer. This makes the CNN learn the features from the training data set. The dummy regression layer is discarded after training the CNN. Then, the features learnt by the CNN are used to train the LSTM network as well as the final regression layer. For CNN and LSTM networks, the batch size is chosen to be 10 and 128, respectively. For both training steps, an Adam optimizer is employed with parameters $\alpha, \beta_{1}, \beta_{2}$, and $\epsilon$ in Equations (22)–(24) set to $10^{-5}$, 0.9, 0.999, $10^{-8}$, respectively. After the training, the CNN, LSTM, and regression layers constitute a deployable neural network.

\subsection*{4.2 Comparing methods}

The proposed method is compared with different deep-learning approaches, in which the CNN is employed alone. In addition, the efficacy of audiovisual features is investigated by training a CNN and CNN-LSTM structure on visual information alone. It has been observed that audio features alone in a CNN or CNN-LSTM structure do not provide competitive results. As such, this experiment is not included in the paper. To compare the different deep-learning approaches with a traditional technique, a singular value decomposition (SVD)-based feature extraction method is employed. In this method, by performing SVD on a frame, a feature vector is generated that contains the singular values [54]. SVR [55] is applied to the extracted features to predict the score of likeability of advertisement. This approach is referred to as ‘SVD-SVR’. The following methods are used to evaluate the score of advertisement likeability:

- SVD-SVR-based method
- CNN-based methods (visual and audiovisual)
- CNN-LSTM-based methods (visual and audiovisual)

Previously, the concepts of CNN visual and CNN audiovisual have been employed in [38, 47], respectively.

\subsection*{4.3 Performance metrics}

The performance of different methods in predicting the likeability score of advertisement clips is evaluated in terms of CCC [56], Pearson correlation coefficient (PCC) [57] and mean absolute percentage error (MAPE) [58]. While PCC estimates a linear correlation between the ground truth and predicted score, CCC additionally measures the agreement between the scores in value labels. MAPE gives a measure of how these values deviate from each other in percentage.

\subsection*{4.4 Results}

Table 2 shows a comparison among all the experimental methods in terms of CCC, MAPE, and PCC, considering all the test scenarios. The evaluation metrics shown in Table 2 are calculated by taking the mean value and the standard deviation of obtained performance parameters for all five rounds of testing procedures. It is seen from Table 2 that the CNN-visual method performs much better than the traditional SVD-SVR method in all performance metrics. Introduction of LSTM layers to CNN-visual architecture in CNN-LSTM visual method boosts the mean value of CCC, MAPE, and PCC by 5.92%, 5.71%, and 3.33%, respectively. Combining visual and audio features in CNN audiovisual method generates 6.76% higher CCC than the CNN-visual method. A similar trend is found in other evaluation metrics. Although CNN-LSTM visual and CNN audiovisual methods are close to each other in terms of evaluation metrics, there is a clear indication that performance is enhanced in the combination of CNN and LSTM layers and in audiovisual feature combinations. In comparison with the CNN audiovisual method, the CNN-LSTM audiovisual method performs 11.61%, 1.64%, and 14.81% better in terms of mean values of CCC, MAPE, and PCC, respectively. In addition to these significant improvements, the CNN-LSTM audiovisual method shows the lowest standard deviation, which indicates the robustness of this method. In brief, CNN-LSTM audiovisual method outperforms the other compared methods in all performance parameters.

In Table 2, the total number of trainable and non-trainable parameters for different methods are mentioned. It is evident from Table 2 that CNN-based deep-learning methods outperform the traditional SVR-based methods in terms of three performance metrics, namely, CCC, MAPE, and PCC. Among the CNN-based methods, the proposed method requires 2.38%, 7.58%, and 8.90% higher number of parameters than that required for the CNN audiovisual, CNN-LSTM visual, CNN-visual methods, respectively, but that results in 11.61%, 12.5%, and 19.15% better performance in terms of CCC. Similar improvements are also observed in terms of other metrics. Thus, in the case of the proposed CNN-LSTM
| Methods               | No. of parameters | CCC Mean ± std. | MAPE (%) Mean ± std. | PCC Mean ± std. |
|-----------------------|-------------------|-----------------|----------------------|-----------------|
| SVD-SVR [54]         | 2,218,155         | 0.116 ± 0.147   | 8.19 ± 1.33          | 0.206 ± 0.214   |
| CNN-visual [47]      | 23,372,441        | 0.355 ± 0.098   | 7.88 ± 0.77          | 0.3985 ± 0.078  |
| CNN-LSTM visual      | 23,635,738        | 0.376 ± 0.122   | 7.43 ± 0.65          | 0.418 ± 0.127   |
| CNN audio-visual [38] | 24,837,105        | 0.379 ± 0.089   | 7.28 ± 0.77          | 0.412 ± 0.082   |
| CNN-LSTM audiovisual | 25,428,658        | 0.423 ± 0.087   | 7.16 ± 0.64          | 0.473 ± 0.067   |

Abbreviations: CCC, concordance correlation coefficient; CNN, convolutional neural network; LSTM, long short-term memory; MAPE, mean absolute percentage error; PCC, Pearson correlation coefficient; SVD, singular value decomposition; SVR, support vector regression.

| TABLE 2 Performance for different methods in terms of evaluation metrics and computational complexity |
|-----------------------------------------------------|

FIGURE 6 Performance analysis for test videos for (a) CNN-visual method, (b) CNN-LSTM visual method, (c) CNN audiovisual method, and (d) CNN-LSTM audiovisual method. CNN, convolutional neural network; LSTM, long short-term memory.

Audiovisual method, the performance has improved to a higher extend at the cost of comparatively low computational complexity.

Improvement of performance for different methods is shown in Figure 6. Figures 6a, 6b, 6c, and 6d compare the predicted frame-wise score of likeability and ground truth of five test videos of round 1 for CNN-visual, CNN-LSTM visual, CNN audiovisual, and CNN-LSTM audiovisual methods, respectively. In Figure 6a, for frame number 8000 to 8500, the gap between the ground truth and predicted score is high, whereas the CNN-LSTM visual method shown in Figure 6b minimizes such gaps. In Figure 6c CNN audiovisual method also has been found successful in minimizing gaps between predicted likeability score and ground truth, but sudden jumps can be seen in many frames, such as the frames 1501 to 2000. The CNN-LSTM audiovisual method in Figure 6d restricts such jumps and predicts the ascend and descend nature of the advertisement likeability score more accurately than other methods.

The RMS likeability of the advertisement video clip is measured by calculating the RMS value of the likeability score of frames of a video clip. The predicted values of RMS likeability of video clips by the proposed CNN-LSTM audiovisual method are compared with the corresponding ground truth values and are
shown in a bar chart in Figure 7. The figure shows that the difference between the ground truth and predicted RMS values ranges from −0.281 to 0.174. Such marginal differences ensure that apart from frame-by-frame likeability, the proposed method can also predict the overall likeability of advertisement video clips with a high level of accuracy.

Table 3 compares the results of fivefold cross-validation of the two most competitive methods, namely, CNN audiovisual and CNN-LSTM audiovisual. From this table, it can be seen that CNN-LSTM audiovisual outperforms CNN audiovisual in all folds in terms of CCC. The CNN audiovisual method shows a better performance in a few cases with marginal differences in terms of MAPE and PCC. It can also be seen that at a given fold, two metrics out of three are the best for the proposed method as compared with the other. Thus, the superiority of the proposed CNN-LSTM audiovisual method in all folds ascertains its efficiency in general.

Table 4 shows the category-wise performance for the CNN-LSTM audiovisual method. The performance metrics
shown in the table are estimated for all test videos of a certain category for all five rounds of the testing procedure. It is reported in Table 4 that the highest CCC, lowest MAPE, and highest PCC can be observed for test videos in the ‘emotional’ category. Videos in the ‘emotional’ category have a gradual build-up of intense feelings from start to finish. In Figure 6d for video No. 12, the proposed CNN-LSTM audiovisual method successfully predicts this trend in the score of advertisement likeability. Table 4 shows high performance for test videos of ‘entertaining’ and ‘creative’. Usually, videos in the ‘entertaining’ category include music. Due to the presence of such audio stimuli, CNN audiovisual and CNN-LSTM audiovisual show improvement compared with CNN-visual and CNN-LSTM visual for video No. 18 in Figure 6. Intuitively, the likeability of advertisement videos of ‘Creative’ depends on the concept and cinematography of the advertisement, and there is a relatively small number of sudden changes in the score of likeability. That is the reason that CNN-LSTM visual and CNN-LSTM audiovisual methods perform relatively better than other methods for video No. 24 in Figure 6. Although overall performance for videos of the ‘miscellaneous’ category is found to be not so high, in Figure 6d for video No. 25, the predicted scores can track the ground truth. Frame-wise likeability of a ‘humorous’ video depends on the ‘punch line’ of the narrative of the video. Such a punch line is largely linked to an understanding of language and a sense of humour rather than visual and audio stimuli. Some humour is a product of its time and does not hold up as time moves forward. It is suspected that for this reason, the proposed method performs relatively poorly for this category.

5 CONCLUSION AND FUTURE WORKS

With the astounding growth of video advertising content, predicting advertisement likeability has become important for advertisers in refining their strategies and making appropriate content to attract the attention of consumers. Although advertising research has a history going back more than 5 decades, very few attempts have been made to predict advertisement likeability. Recently, deep learning has ushered in a new dimension of research in many fields. This paper presents different supervised convolutional and LSTM models using audio and visual content to predict the frame-by-frame instantaneous as well as RMS likeability of advertisement clips. The work also presents a new complete data set that contains frame-wise preference score annotation and facial responses of annotators.

Intuitively, facial expression can be linked to the likeability that leads to the success of an advertisement. In the future, the facial response section of the data set can be explored to ensure enhanced advertisement effectiveness, which is the prime target of digital marketing strategy markers. In addition, facial characteristics for each category of video can be analysed, and the difference in video likeability of users for each category can be investigated. As a continuation of this work, future efforts can also be made for a better understanding of humorous advertisements.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of NVIDIA Corporation for the donation of a Titan Xp GPU that was used in this research. Also, we would like to thank the anonymous reviewers for their valuable comments, which have been useful in improving the quality of the paper.

REFERENCES

1. IAB, PwC: IAB Internet Advertising Revenue Report 2017 Full-Year Results, Interactive Advertising Bureau (IAB) and PricewaterhouseCoopers (PwC). Tech. Rep. (2018). Available: https://www.iab.com/wp-content/uploads/2018/05/IAB-2017-Full-Year-Internet-Advertising-Revenue-ReportREV_.pdf
2. Haley, R.L., Baldinger, A.L.: The ARF copy research validity project. J. Advert. Res. 31(2), 11–32 (1991)
3. McDuff, D., et al.: Automatic measurement of ad preferences from facial responses gathered over the internet. Image Vis. Comput. 32(10), 630–640 (2014)
4. Wells, W.D.: Measuring Advertising Effectiveness. Lawrence Erlbaum Associates Inc., USA (1997)
5. Morwitz, V.G., Steekel, J.H., Gupta, A.: When do purchase intentions predict sales? Int. J. Forecast. 23(3), 347–364 (2007)
6. Walker, D., Dubitsky, T.M.: Why liking matters. J. Advert. Res. 34(3), 9–18 (1994)
7. Bier, A.L., Bridgewater, C.A.: Attributes of likeable television commercials. J. Advert. Res. 30(5), 38–44 (1990)
8. Sonit, E.G., Van Meurs, L., Neijens, P.C.: Effects of advertising likeability: a 10-year perspective. J. 46(1), 73–83 (2006)
9. Lavidge, R.J., Steiner, G.A.: A model for predictive measurements of advertising effectiveness. J. Market. 25, 59–62 (1961)
10. Venkatraman, V., et al.: Predicting advertising success beyond traditional measures: new insights from neurophysiological methods and market response modelling. J. Market. Res. 52(4), 436–452 (2015)
11. Du Plessis, E.: Recognition versus recall. J. Advert. Res. 34(3), 75–91 (1994)
12. Mandler, G.: Recognising: the judgement of previous occurrence. Psychol. Rev. 87(3), 252–271 (1980)
13. Fam, K.S., Waller, D.S.: Identifying likeable attributes: a qualitative study of television advertisements in Asia. Q. J. Mark. Res: An Int. J. 9(1), 38–50 (2006)
14. Collett, P.: Youth: how do they view your commercials? In: Transcript Proceedings on ARF Youth Research Workshop, pp. 38–44. Advertising Research Foundation, New York (1994)
15. Hollis, N.S.: Like it or not, liking is not enough. J. Advert. Res. 35(5), 7–17 (1995)
16. Agee, T.: Recall the brand. Market Mag. 16, 63 (1997)
17. Sacharin, K.: Attention!: How to Interrupt, Yell, Whisper, and Touch consumers. John Wiley & Sons (2004)
18. Spaeth, J., Hess, M., Tang, S.: The anatomy of liking. In: Annual ARF Copy Research Workshop. Advertising Research Foundation, New York (1990)
19. Stone, G., Besser, D., Lewis, L.E.: Recall, liking, and creativity in TV commercials: a new approach. Jat. 40(3), 7–18 (2000)
20. Fam, K.-S.: What drives ad likeability/dislikeability in Hong Kong and Thailand? Int. J. Bus. Soc. 7(2), 3–19 (2006)
21. Rimoldi, O.: The Impact of ‘likeability’ on Advertising Effectiveness: To What Extent Does Liking an Advert Have a Persuasive Influence on Consumer Behaviour? University of Auckland (2008)
22. Venkatraman, V., Payne, J.W., Huettel, S.A.: An overall probability of winning heuristic for complex risky decisions: choice and eye fixation evidence. Organ. Behav. Hum. Decis. Process. 125(2), 73–87 (2014)
23. Lang, A., et al.: The effects of production pacing and arousing content on the information processing of television messages. J. Broadcast Electron. Media. 43(4), 451–475 (1999)

24. Wang, Y.J., Minor, M.S.: Validity, reliability, and applicability of psychophysiological techniques in marketing research. Psychol. Mark. 25(2), 197–232 (2008)

25. Lee, N., Broderick, A.J., Chamberlain, L.: What is ‘neuromarketing’? A discussion and agenda for future research. Int. J. Psychophysiol. 63(2), 199–204 (2007)

26. Ariely, D., Berns, G.S.: Neuromarketing: the hope and hype of neuroscience in business. Nat. Rev. Neurosci. 11(4), 284–292 (2010)

27. Smidts, A., et al.: Advancing consumer neuroscience. Mark. Lett. 25(3), 257–267 (2014)

28. Venkatraman, V., et al.: New scanner data for brand marketers: how neuroscience can help better understand differences in brand preferences. J. Consum. Psychol. 22(1), 143–153 (2012)

29. McDuff, D., et al.: Predicting ad liking and purchase intent: large-scale analysis of facial responses to ads. IEEE Trans. Affective Comput. 6(3), 223–235 (2015)

30. Teixeira, T., Wedel, M., Pieters, R.: Emotion-induced engagement in internet video advertisements. J. Market Res. 49(2), 144–159 (2012)

31. Okada, G., Masui, K., Tsumura, N.: Advertisement effectiveness estimation based on crowdsourced multimodal affective responses. In: Proceedings on IEEE International Conference Computer Vision and Pattern Recognition Workshops, pp. 1263–1271. Salt Lake City (2018)

32. Bowman, M.: Video Marketing: The Future of Content Marketing Forbes. https://www.forbes.com/sites/forbesagencycouncil/2017/02/05/video-marketing-the-future-of-content-marketing/ (2017). Accessed 22nd December 2020 [Online]

33. Biel, J., Gatica-Perez, D.: The YouTube lens: crowdsourced personality impressions and audiovisual analysis of vlogs. IEEE Trans. Multimed. 15(1), 41–55 (2012)

34. Lee, J., Lee, J.: Music popularity: metrics, characteristics, and audio-based prediction. IEEE Trans. Multimedia. 20(11), 3173–3182 (2018)

35. Trzcinski, T., Rolita, P.: Predicting popularity of online videos using support vector regression. IEEE Trans. Multimedia. 19(11), 2561–2570 (2017)

36. Zhou, J., et al.: Improved softmax loss for deep learning-based face and expression recognition. Cogn. Comput. and Syst. 1(4), 97–102 (2019)

37. Basnet, R., et al.: Statistical selection of CNN-based audiovisual features for instantaneous estimation of human emotional states. In: Proceedings on IEEE International Conference New Trends in Computing Sciences, pp. 50–54. IEEE, Amman (2017)

38. ———: Estimation of affective dimensions using CNN-based features of audiovisual data. Pattern Recogn. Lett. 128, 290–297 (2019)

39. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9(8), 1735–1780 (1997)

40. Gharibshah, Z., et al.: Deep learning for user interest and response prediction in online display advertising. Data Sci. Eng. 5(1), 12–26 (2020)

41. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings on IEEE International Conference Computer Vision and Pattern Recognition, pp. 2625–2634. Boston (2015)

42. Xingjian, S., et al.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Advances in Neural Information Processing Systems, pp. 802–810. Montreal (2015)

43. Brady, K., et al.: Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. In: Proceedings on ACM International Workshop. Audio/visual Emotion Challenge, pp. 97–104. ACM, Amsterdam (2016)

44. Tzirakis, P., et al.: End-to-end multimodal emotion recognition using deep neural networks. IEEE J. Sel. Top. Signal Process. 11(8), 1301–1309 (2017)

45. Gao, L., et al.: Video captioning with attention-based LSTM and semantic consistency. IEEE Trans. Multimedia. 19(9), 2045–2055 (2017)

46. Pechmann, C., Andrews, C.: Copy test methods to pretest advertisements. In: Wiley International Encyclopaedia of Marketing, vol. 4, pp. 54 (2010)

47. Al Nahian, M., et al.: CNN-based prediction of frame-level shot importance for video summarization. In: Proceedings on IEEE International Conference on New Trends in Computing Sciences, pp. 24–29. IEEE, Amman (2017)

48. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Proceedings on International Conference on Learning Representations, Banff (2014)

49. Song, Y., et al.: TVSum: Summarising web videos using titles. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 5179–5187. Boston (2015)

50. Cronbach, L.J.: Coefficient alpha and the internal structure of tests. Psychometrika. 16(3), 297–334 (1951)

51. Bartko, J.J.: The intraclass correlation as a measure of reliability. Psychol. Rep. 19(1), 3–11 (1966)

52. IBM Corp.: IBM SPSS Statistics for Windows, Version 25.0 Armonk, IBM Corp., Released (2017)

53. Koo, T.K., Li, M.Y.: A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropractic Med. 15(2), 155–163 (2016)

54. Ntalianis, K.S., Kollias, S.D.: An optimized key-frames extraction scheme based on SVD and correlation minimization. In: Proceedings on IEEE International Conference Multimedia and Expo, pp. 792–795. IEEE, Amsterdam (2005)

55. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Stat. Comput. 14(3), 199–222 (2004)

56. Lin, T.L.-K.: A concordance correlation coefficient to evaluate reproducibility. Biometrics. 45(1), 255–268 (1989)

57. Pearson, K.: Notes on regression and inheritance in the case of two parents. Proc. Royal Society of London. 58, 240–242 (1895)

58. Spyros, M.: Accuracy measures: theoretical and practical concerns. Int. J. Forecast. 9(4), 527–529 (1993)

How to cite this article: Saha D., Rahman SM, Mahbubur, Islam MT, Ahmad MO, Swamy MNS. Prediction of instantaneous likeability of advertisements using deep learning. Cogn. Comput. Syst. 2021;1–13. https://doi.org/10.1049/ccs2.12022