Multimodal Deep Learning Framework for Image Popularity Prediction on Social Media

Fatma S. Abousaleh, Wen-Huang Cheng, Neng-Hao Yu, and Yu Tsao, *Member, IEEE,*

***Abstract*—Billions of photos are uploaded to the web every day through various types of social networks. A few of these images receive millions of views and become popular, whereas others remain completely unnoticed. This raised the problem of predicting image popularity on social media. The popularity of an image can be affected by several factors, such as visual content, aesthetic quality, user, post metadata, and time. Thus, considering all these factors is essential for accurately predicting image popularity. In addition, the efficiency of the predictive model also plays a crucial role. In this study, motivated by multimodal learning, which uses information from various modalities, and the current success of convolutional neural networks (CNNs) in various fields, we propose a deep learning model, called visual-social convolutional neural network (VSCNN), that predicts the popularity of a posted image by incorporating various types of visual and social features into a unified network model. VSCNN first learns to extract high-level representations from the input visual and social features by utilizing two individual CNNs. The outputs of these two networks are then fused into a joint network to estimate the popularity score in the output layer. We assess the performance of the proposed method by conducting extensive experiments on a dataset of approximately 432K images posted on Flickr. The simulation results demonstrate that the proposed VSCNN model significantly outperforms state-of-the-art models, with a relative improvement of more than 2.33%, 7.59%, and 14.16% in terms of Spearman's Rho, mean absolute error, and mean squared error, respectively.**

***Index Terms*—Popularity prediction, social media, convolutional neural networks, multimodal learning.**

I. Introduction

SOCIAL media websites (e.g., Flickr, Twitter, and Facebook) allow users to create and share content (e.g., by liking, commenting, or viewing). Consequently, social media platforms have become an inseparable part of our daily lives, and massive amounts of social content are generated on these platforms. The explosive growth of social media content (i.e., texts, images, audios, and videos) and the interactive behavior between web users result in that only a small amount of online social content attracts significant attention and becomes popular, whereas its vast majority either receives little attention or is totally overlooked. Therefore, extensive effort has been expended in the past few years to predict social media content popularity, understand its variation, and evaluate its growth [1]–[7]. This popularity reflects people's interests and provides opportunities to understand user interaction with online content as well as the information diffusion through social media websites. Hence, an accurate popularity prediction of online content may improve user experience and service effectiveness. Moreover, it can greatly influence several important applications, such as online advertising [8], [9], information retrieval [10], online product marketing [11], and content recommendation [12].

Popularity prediction on social media is usually defined as the problem of estimating the rating scores, view counts, or click-through of a post [13]. The present study is concerned with image popularity prediction on social media websites, so that insight may be gained into the popularity factors for a particular image. Although this problem has recently received significant attention [14]–[17], it remains a challenging task. For example, image popularity prediction could be greatly influenced by various factors (and features), such as visual content, aesthetic quality, user, post metadata, and time. Thus, considering all this multimodal information is crucial for efficiently predicting image popularity. Moreover, it is nontrivial to select an appropriate model that can make better use of the various features contributing to image popularity and accurately predict it. For example, simple machine learning schemes (e.g., support vector and decision tree regression) learn to predict by feeding them with highly-structured data, and thus time as well as skill are required to fine-tune their hyperparameters. However, to obtain accurate prediction results, it is critical to construct a prediction model capable of learning through a more abstractive data representation and the optimization of the extracted features.

Accordingly, we address the image popularity prediction problem by analyzing a large-scale dataset collected from Flickr to investigate two essential components that may contribute to the popularity of an image, namely visual content and social context. Particularly, we examine the effect of the visual content of an image on its popularity by adopting different types of features that describe various visual aspects of the image, including high-level, low-level, and deep learning features. These are extracted by applying several techniques from machine learning and computer vision. Additionally, we explore the significant role of social context information associated with images and their owners by analyzing three types of social features: user, post metadata, and time. To demonstrate the efficacy of the proposed features, we propose a computational deep learning model, called visual-social convolutional neural network (VSCNN), that uses two individual CNNs to learn high-level representations of the visual and social features independently. The outputs of the

Fatma S. Abousaleh is with Social Networks and Human-Centered Computing Program, Taiwan International Graduate Program, Institute of Information Science and Research Center for Information Technology Innovation, Academia Sinica, and also with Department of Computer Science, National Chengchi University, Taipei 11529, Taiwan. (e-mail: fatma@iis.sinica.edu.tw).

Wen-Huang Cheng is with the Institute of Electronics, Department of Electronics Engineering, National Chiao Tung University, Taipei, Taiwan. (e-mail: whcheng@nctu.edu.tw).

Neng-Hao Yu is with the College of Design, Department of Design, National Taiwan University of Science and Technology, Taipei, Taiwan. (e-mail: jonesyu@mail.ntust.edu.tw).

Yu Tsao is with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 11529, Taiwan. (e-mail: yu.tsao@citi.sinica.edu.tw).

two networks are then merged into a shared network to learn joint multimodal features and compute the popularity score in the output layer. End-to-end learning fashion is employed to train the entire model, and the weights of its parameters are learned through back-propagation. In summary, the main contributions of this study are as follows:

- We demonstrate a comprehensive exploration of independent benefits and predictive power of various types of visual and social context features towards the popularity of an image. We further demonstrate that these multimodal features can be combined effectively to enhance prediction performance.
- We propose a deep learning VSCNN model for predicting image popularity on social media. VSCNN uses dedicated CNNs to learn structural and discriminative representations from the input visual and social features, achieving considerable performance in predicting image popularity compared with several traditional machine learning schemes.
- We demonstrate that processing visual and social features using the late fusion scheme is significantly better than using the early fusion scheme.
- We use a large-scale dataset of approximately 432K images posted on Flickr to evaluate the performance of the proposed VSCNN model. The simulation results demonstrate that VSCNN achieves competitive performance and outperforms six baseline models and other state-of-the-art methods.

The rest of this paper is organized as follows. Section II reviews related work. The extraction of visual and social features is discussed in Section III. Section IV presents the details of the proposed framework for image popularity prediction and discusses the six baseline models used for comparison. The experimental setup and results are described in Section V. Section VI concludes the paper and provides some directions for future work.

II. RELATED WORK

In recent years, predicting the popularity of social media content has received substantial attention [17]–[20]. Regarding image popularity prediction, the related studies differ in terms of the definition of the popularity metric (e.g., view, reshare, and comment counts), but they all share the same basic pipeline consisting of extracting and testing several types of features that influence popularity, and then applying a classification or regression model for prediction. Therefore, we review these studies by categorizing them according to the features used and the prediction model.

Regarding the features used, existing research has primarily focused on investigating the relative effectiveness of different types of features in predicting image popularity, including social context, visual content, aesthetic, and time features. For instance, Khosla *et al.* [7] demonstrated that image content (e.g., gist, color histogram, texture, color patches, gradient, and deep learning features) and social cues (e.g., number of followers or number of posted images) have a great effect on image popularity. Gelli *et al.* [15] employed visual sentiment features along with context and user features to predict a succinct popularity score of social media images. They demonstrated that sentiment features are correlated with popularity and have considerable prediction power if they are used together with context features. Cappallo *et al.* [14] demonstrated that latent image features can be used to predict image popularity. They explored the visual cues that determine popularity by identifying themes from both popular and unpopular images. McParlane *et al.* [16] performed image classification using a combination of four broad feature types, that is, image content, image context, user context, and tags, to predict whether an image will obtain a high or low number of views and comments in the future.

Compared with the above-mentioned approaches, relatively few studies have been conducted to demonstrate the effect of time and aesthetic features on image popularity. For instance, Wu *et al.* [13] developed a new framework, called multi-scale temporal decomposition, to predict image popularity based on popularity matrix factorization. They explored the mechanism of dynamic popularity by factoring popularity into user-item and time-sensitive context. Furthermore, Almgren *et al.* [21] employed social context, image semantics, and early popularity features to predict the future popularity of an image. Specifically, they considered the popularity changes over time by collecting information about the image within an hour of uploading and keeping track of its popularity for a month. Totti *et al.* [22] analyzed the effect of visual content on image popularity and its propagation on online social networks. Along with social features, they proposed using aesthetical properties and semantic content to predict the popularity of images on Pinterest.

We observe that most of the aforementioned studies rely only on a part of the useful features for image popularity prediction, and they do not consider the interactions between other pertinent types.

In regard to the models used for prediction, previous studies have introduced several types of machine learning schemes. Both [7] and [15] regarded image popularity prediction as a regression problem in which support vector regression (SVR) [23] was used to predict the number of views that an image receives on Flickr. Totti *et al.* [22] reduced the problem to a binary classification task and utilized a random forest classifier [24] to predict whether an image would be extremely popular or unpopular based on the number of reshares on Pinterest. Moreover, the authors in [25] predicted the number of views of an image on Flickr using a gradient boosting regression tree [26]. Even though most of these prediction models perform satisfactorily, they tend to generate smoothed results, and thus the popularity of images with overly high or low scores will be difficult to predict accurately. In addition, it may be time-consuming to fine-tune hyperparameters that significantly influence the performance of these models.

Recently, deep learning techniques have attracted widespread attention and achieved outstanding performance in various tasks [27]–[30]. This due to the capability of deep neural networks to learn complex representations from data at each layer, where they imitate learning in the human brain by abstraction. Nevertheless, little effort has been expended for predicting image popularity using these techniques. In

this regard, Wu *et al.* [31] proposed a new deep learning framework to investigate the sequential prediction of image popularity by integrating temporal context and attention at different time-scales. Moreover, Meghawat *et al.* [32] developed an approach that integrates multiple multimodal information into a CNN model for predicting the popularity of images on Flickr. Even though these studies have achieved satisfactory performance, they are not sufficiently powerful to capture and model the characteristics of image popularity. For instance, the authors in [32] investigated the effect of the visual content of an image on its popularity by utilizing only one feature obtained by the pre-trained InceptionResNetV2 model, whereas they ignored other important visual cues, such as low-level computer vision, aesthetic, and semantic features. Moreover, although it has been demonstrated that time features have a crucial effect on image popularity [13], [31], they were not considered in the proposed model. They also adopted the early fusion scheme for processing the proposed multi-modal features, even though several studies have demonstrated that this scheme is outperformed by the late fusion scheme in processing heterogeneous information [7], [33].

To address the above issues, we propose a multimodal deep learning prediction model that uses numerous types of features associated with image popularity, including multi-level visual, deep learning, social context, and time features. The proposed model uses dedicated CNNs for separately learning high-level representations from the input features and then efficiently merges them into a unified network for popularity prediction.

III. FEATURES

In this section, we analyze various types of features that can influence image popularity. First, in Section III-A, we investigate visual features that could be used to describe different facets of images based on their content, including low-level, high-level, and deep learning features. Then, in Section III-B, we explore several social features based on context information of images and their owners.

A. Visual Content Features

1) Low-level Features

There are several types of low-level computer vision features (e.g., texture, color, shape, gist, and gradient) that are likely to be used for visual processing. In this study, we adopt three of such features: color, texture, and gist.

Color: A perfect color distribution in an image attracts viewers' attention and aids in determining object properties and understanding scenes. In this study, we use the color histogram descriptor, resulting in a vector of 32 dimensions that characterizes the color feature [34].

Texture: Texture can describe the homogeneity of the colors or intensities in an image. It can also be utilized to identify the most interesting objects or regions [35]. To investigate its effect on image popularity, we employ one of the most widely used features for texture description, namely, local binary patterns (LBP) [36]. More precisely, we use the uniform LBP [37] descriptor, resulting in a 59-dimensional feature vector.

Gist: The GIST descriptor provides a rough description of a scene by epitomizing the gradient information (scales and orientations) for various parts of a photo. To extract the GIST feature of an image, we adopt the widely used GIST descriptor proposed in [38], resulting in a feature vector with 512 dimensions.

2) High-level Features

The quality and aesthetic appearance of an image are important for its popularity. Based on the various photographic techniques and the aesthetic standards used by professional photographers, we adopt certain aesthetic features for popularity prediction. These features are developed to evaluate the visual quality of a photograph by separating the subject area from the background using the blur detection technique [39]. Then, based on the result of this separation process, six types of aesthetic features are computed as described below.

- **Clarity contrast:** To attract the viewer's attention to the key point of a photograph and to isolate the subject region from the background, professional photographers normally adjust the lens to keep the subject in focus and bring the background out of focus. Accordingly, a clear photograph will have comparatively more high-frequency components than a blurred photograph. To characterize this property, we define a clarity contrast feature based on the method presented in [39].
- **Hue count:** The hue count of an image is a metric of its simplicity. It can also be used to assess image quality. Although professional photographs appear bright and vivid, their hue number is normally less than that of amateur photographs. We thus compute the hue count feature of an image using a 20-bin histogram H_c quantized on the good hue values. This can be formulated as follows [40]:

$$f_l = 20 - |N_c|, \tag{1}$$

$$N_c = \big\{ i \mid H_c(i) > \beta m \big\}. \tag{2}$$

  where N_c denotes the set of bins with values larger than βm, m is the maximum histogram value, and β is used to control the noise sensitivity of the hue count. We select $\beta = 0.05$ in our experiments.
- **Brightness contrast:** In high-quality photographs, the subject area's brightness significantly differs from that of the background because professional photographers frequently use different subject and background lightings. Nevertheless, most amateurs use natural lighting and allow the camera to adjust the brightness of a picture automatically; this usually reduces the difference in brightness between the subject area and the background. To discern the difference between these two types of photographs, we calculate a brightness contrast feature based on the method described in [41].
- **Color entropy:** Owing to the distinct interrelationship between the color planes of drawings and natural images, the entropy of RGB and Lab color space components is computed to differentiate natural images from drawings [42].
- **Composition geometry:** The proper geometrical composition is a fundamental demand to obtain high-quality photographs. The *Rule of Thirds* is one of the most

important photographic composition principles utilized by professional photographers to bring more balance and high quality to their photos. To formulate this criterion, we define a composition feature based on the method introduced in [41].

- **Background simplicity:** Professional photographers normally maintain simplicity to enhance the composition. This is because photographs that are clean and free from distracting backgrounds are considerably more appealing and naturally draw the attention of the viewer to the subject. It is known that the color distribution in a simple background tends to be less dispersed. For that reason, we compute a feature that represents the background simplicity of an image using the method proposed in [39], which is based on the color distribution of the background.

In this study, we combine the six aesthetic features described above, resulting in an 11-dimensional feature vector.

3) Deep Learning Features

Recently, deep learning methods have been widely used for image representation owing to their effectiveness [27], [30]. In this study, the CNN architecture of the VGG19 model is employed to learn the deep features of photographs [27]. The VGG19 model was trained on 1.2 million images from the ImageNet database to classify these images into 1000 categories [30]. We use the Keras framework of the VGG19 pre-trained CNN model [43] for feature extraction from the layer situated immediately before the final classification layer, (i.e., the last fully connected layer (fc7)). The output of this layer is a 4,096-dimensional feature vector. Some images selected from the dataset and the plots of their respective deep feature vectors values are shown in Fig. 1.

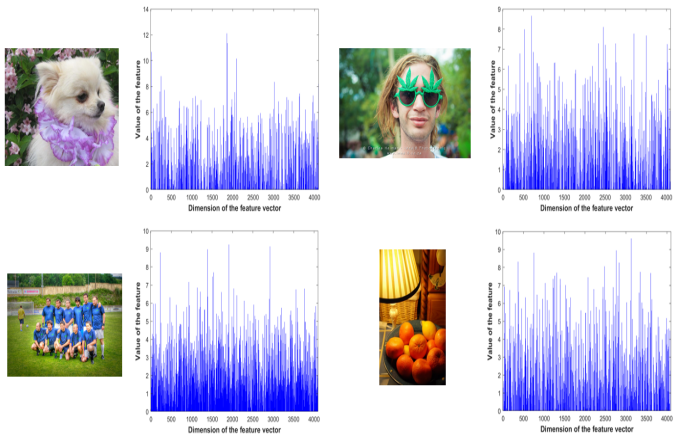


Fig. 1. The plots of deep learning feature vector values of different images from the dataset.

B. Social Context Features

Previous studies demonstrated that the popularity of an image depends not only on its content but also on the social information regarding the user who uploaded it and the textual information associated with it [7], [17]. In this section, we attempt to determine to what extent social features can influence image popularity. We analyze three types of such features: user, post metadata, and time.

TABLE I: Spearman's Rho values for the correlation of user features with popularity score.

Feature	Spearman's Rho
User id	**0.74**
Average views	**0.74**
Group count	0.067
Member count	0.19
Image count	- 0.35

1) User

There is no doubt that the popularity of a user is highly correlated with the popularity of his/her posted images. For this purpose, we adopt several user-centered features. These features are listed below so that they will have the same value for all photographs posted by the same user.

User id: This is defined as a unique integer number (from 1 to 135) according to the average view count of all images of a user (i.e., a greater average number of views implies a larger value). Thus, this number is a unique identification of each user and can be used directly in the prediction model.

Average views: The average view count of all user-uploaded photos.

Group count: The number of groups to which a user subscribes.

Member count: The mean number of members in the groups to which a user subscribes.

Image count: The total number of photographs posted by a user.

To characterize the effects of these features on predicting image popularity, we compute their rank correlation with the image popularity score using Spearman's rank correlation coefficient (Spearman's Rho) [44]. The value of the correlation coefficient ranges in [-1, 1], where a score of 1 (resp. -1) indicates an ideal positive (resp. negative) association, and a score of zero indicates no correlation. The results are shown in Table I. It can be seen that both user id and image views have a strong positive correlation with image popularity (Spearman's Rho = 0.74).

2) Post Metadata

The contextual information associated with an uploaded image (e.g., tags, comments, or title) can also influence its popularity. For instance, an image with a large number of tags would be expected to appear more frequently in search results. Therefore, we consider certain image context features for popularity prediction. These features refer to image-related metadata, and most of them are entered by the user. The image-context features that we adopted are listed below.

Tag count: The number of tags annotated by a user on a posted photograph.

Title length: The number of characters in the title of a photograph.

Description length: The character count in the image description.

Tagged people: A binary number (1 or 0) indicating whether a given photograph has tagged people or not.

Comment count: The number of comments an image has obtained from other users.

We calculate the relationship between each of these features and the popularity of an image as in the case of user features.

TABLE II: Spearman's Rho values for the correlation of post metadata features with popularity score.

Feature	Spearman's Rho
Tag count	0.49
Title length	0.22
Description length	0.51
Tagged people	0.0043

The results of Spearman's Rho are listed in Table II. We notice that most of the post features have a large positive correlation with popularity, except the tagged people feature, which has a slightly positive correlation of 0.0043.

From the results shown in Table I and Table II, we can note that the user id and image views features have the highest Spearman's Rho scores, which implies that user-centric features are the most effective in predicting image popularity. This also agrees with what we have expected because popular users usually have a large number of followers, meaning their images are more likely to receive a larger amount of views after uploading on social media and thus becoming popular.

3) Time

Along with user and post features, there is a strong dependence on time features in predicting image popularity. For instance, users tend to become more active on social websites during the weekend. Thus, images that are posted at these time slots would naturally be expected to receive more views and therefore more ratings. Hence, we consider the following time features: post day, post month, post time, and the post duration. The definitions of these features are as follows:

Post day: It represents the day of the week on which a photograph is posted. We encoded the day number using one-hot encoding of a 7-dimensional vector.

Post month: It represents the month in which a photograph is posted. We encoded the month number using one-hot encoding of a 12-dimensional vector.

Post time: It represents the period of the day during which a photograph is posted. We divide the day into four segments (six hours in each segment) and assume that the photograph is posted either in the morning (06:00 to 11:59), afternoon (12:00 to 17:59), evening (18:00 to 23:59), or night (00:00 to 05:59). Then, we encoded the post time using one-hot encoding of a 4-dimensional vector.

Post duration: The amount of time in days during which the photograph remained posted on Flickr.

IV. METHODOLOGY

In this section, we explain the details of the proposed framework for predicting social media image popularity; moreover, we present a brief description of the baseline models.

A. Overview of Proposed Framework

The overall diagram of the proposed framework is shown in Fig. 2. It consists of two phases: feature extraction and VSCNN regression. In the feature extraction phase, for each post in the dataset, we extract visual features from the image and social context features from its corresponding post context information, as shown in Fig. 2 (a). We integrate the extracted visual features to obtain a final feature vector of 4,710 dimensions that describes different visual facets of the image. Then, principal component analysis (PCA) [45] is performed to decrease the dimensionality of this vector from 4,710 to 20 and to select only the prevalent features. This results in a visual-PCA descriptor of 20 dimensions, which is denoted by X. Finally, the values of the features of X are normalized so that all of them belong to the same scale. Similarly, the same procedure is applied to the corresponding extracted social features to obtain a normalized social-PCA descriptor of 14 dimensions, which is denoted by Z. The obtained X and Z will be used as input to the proposed SVCNN model to predict the popularity of the corresponding post.

As shown in Fig. 2 (b), the proposed SVCNN model consists of two individual CNNs that are used to derive the structural and discriminative representations from the visual (X) and social (Z) features, namely, the visual and the social network. Each of these networks is a one-dimensional CNN consisting of three convolutional layers. The rectified linear unit (ReLU) is employed as the activation function for each convolutional layer to avoid the vanishing gradient problem in the training phase. A fusion network is used to combine the outputs of these networks into a unified network. It consists of one merged and two fully-connected layers. The merged layer is used to concatenate the outputs of the last convolutional layer of the visual network and that of the social network and generate the inputs of the first fully-connected layer. Finally, the outputs of the second fully-connected layer are summed at the final node, generating the predicted popularity score. In the diagram, the convolutional and fully-connected layers are denoted by $\mathrm{Conv1D_v}1$, $\mathrm{Conv1D_v}2$, $\mathrm{Conv1D_v}3$, $\mathrm{Conv1D_s}1$, $\mathrm{Conv1D_s}2$, $\mathrm{Conv1D_s}3$, FC1, and FC2, where the subscripts "v" and "s" indicate visual and social features, respectively.

B. Training the VSCNN Model

As in other CNNs, we first prepare a set of training samples (N) to train the VSCNN model. Each sample comprises the posted image, the post context information, and their corresponding popularity score (Y). We then calculate the visual feature descriptor (X) and the corresponding social feature descriptor (Z) for each sample, as described above. For each iteration, we obtain the output of the visual network as

$$V_i = Conv1D_v3\Big(Conv1D_v2\big(Conv1D_v1(X_i)\big)\Big),\ i = 1...N. \tag{3}$$

Similarly, the output of the social network is as follows:

$$S_i = Conv1D_s3\Big(Conv1D_s2\big(Conv1D_s1(Z_i)\big)\Big),\ i = 1...N. \tag{4}$$

Then, we flatten V_i and S_i, and concatenate the two feature vectors using a merge layer, and then we use the concatenated feature vector as the input of the fusion network, $F_i = [V_i^{'}\ S_i^{'}]^{'}$. Thus, a fully-connected cascade feed-forward network can be calculated as

$$\hat{Y_i} = FC2\Big(FC1\big(F_i\big)\Big),\ i = 1...N. \tag{5}$$

Let θ denote the parameters of the VSCNN model. First, they are initialized using random values between -1 and 1, and then

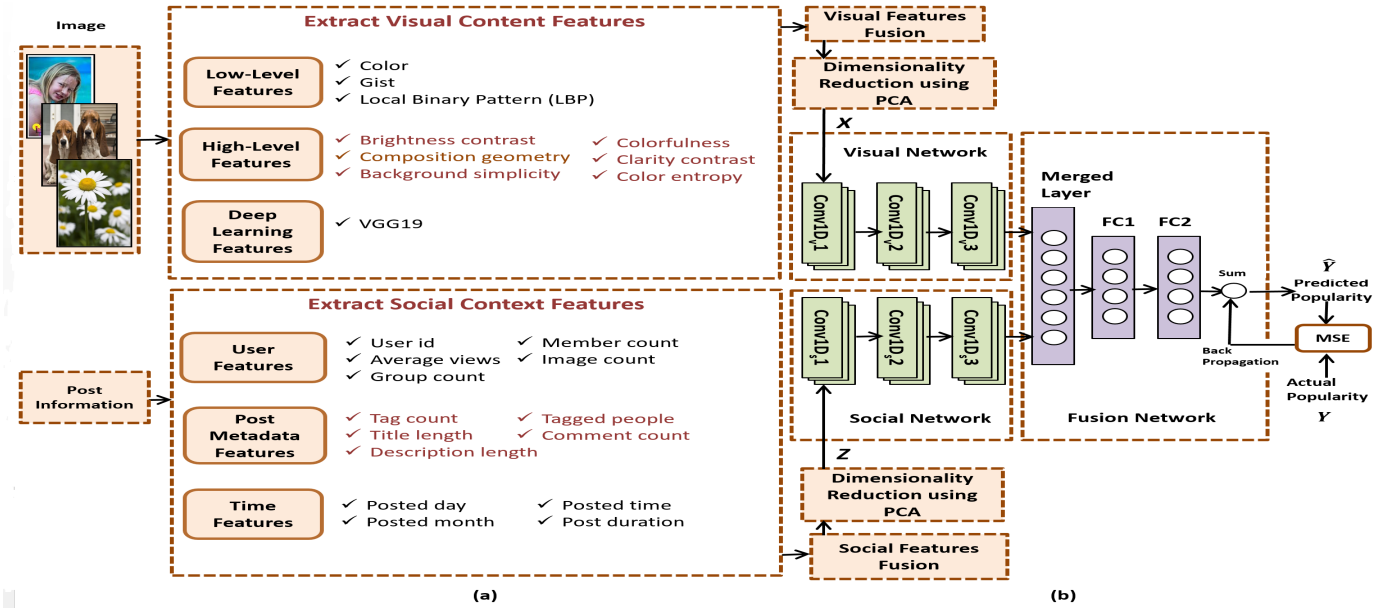


Fig. 2. Diagram of the proposed framework for image popularity prediction. (a) Feature extraction, and (b) Proposed VSCNN regression model.

TABLE III: Configuration of the VSCNN Model.

Layer	Kernel	Activation Function	Number of Neurons
$\text{Conv1D}_\text{V}1$	3	ReLU	32
$\text{Conv1D}_\text{V}2$	3	ReLU	64
$\text{Conv1D}_\text{V}3$	3	ReLU	128
$\text{Conv1D}_\text{S}1$	2	ReLU	32
$\text{Conv1D}_\text{S}2$	2	ReLU	64
$\text{Conv1D}_\text{S}3$	2	ReLU	128
Merged Layer			4736
FC1		ReLU	1024
FC2		ReLU	500

they are trained by optimizing the following mean squared error (MSE) cost function using back-propagation:

$$MSE(\theta) = \min_{\theta} \left(\frac{1}{N} \sum_{i=1}^{N} \left\| \hat{Y}_i - Y_i \right\|_2^2 \right). \tag{6}$$

We adopt a stride of size 1 in the networks of the VSCNN model. To avoid overfitting, a dropout of 0.1 is adopted after each layer, except the last fully-connected layer, in which a dropout of 0.2 is used. Additionally, we apply batch normalization in each convolutional layer to increase the stability of the CNNs. Other details about the configuration of the VSCNN model are presented in Table III.

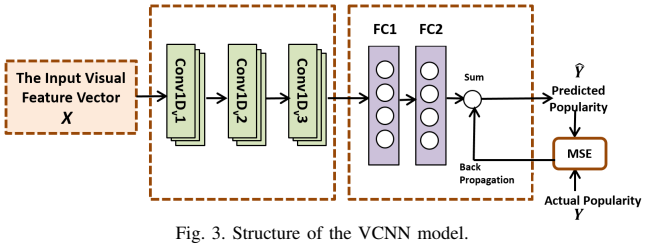


Fig. 3. Structure of the VCNN model.

C. Baseline Models

As popularity prediction for social media images is a regression problem, only a small number of current machine learning models can be directly used. Therefore, we first compare the proposed VSCNN model with two CNN-based models, namely the visual-only CNN (VCNN) and the social-only CNN (SCNN) model. Then, we compare the proposed VSCNN with four conventional regression models: linear regression (LR), SVR, decision tree regression (DTR), and gradient boosting decision tree (GBDT). We provide a short description of each of these models in the following subsections.

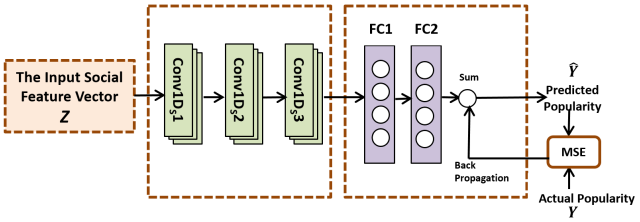


Fig. 4. Structure of the SCNN model.

1) Visual-only Convolutional Neural Network

As shown in Fig. 3, the VCNN model disjoints all social-related parts in the proposed VSCNN (cf. Fig. 2 (b)) and retains the rest. The VCNN is trained using the same procedure as in the case of the VSCNN model, which was described in Section IV-B.

2) Social-only Convolutional Neural Network

The structure of the SCNN model is shown in Fig. 4. It detaches all visual-related components in the proposed VSCNN model (cf. Fig. 2 (b)).

3) Linear Regression

LR is a statistical model designed for modeling the relationship between a single dependent variable (output) and a set of independent variables (inputs) by finding a linear regression function that best describes the input variables. To predict the popularity score of an image using this model, we assume a linear relationship between the features of an input image and

the popularity score, as follows:

$$y = w_0 + w_1 x_1 + ... + w_n x_n + \varepsilon = \mathbf{x}^{\mathsf{T}}\mathbf{w} + \varepsilon, \quad (7)$$

where y indicates the predicted popularity score of the input image, $\mathbf{x}$ denotes the feature vector, $\mathbf{w}$ is the model's weight vector, and ε is the error term. The gradient descent algorithm [46] is employed to learn the weight coefficients during the training phase.

4) Support Vector Regression

SVR [23] is a regression version of a support vector machine [47]. It can construct advanced optimal approximation functions using the training data. Given M training samples of popularity feature vectors $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M\}$, where $\mathbf{x}_i \in \mathbb{R}^d$, and their corresponding popularity score values $\{y_1, y_2, ..., y_M\}$, where $y_i \in \mathbb{R}$, the regression is performed by finding a continuous mapping $f : \mathbb{R}^d \rightarrow \mathbb{R}$ that best predicts the set of training samples with approximation function $y = f(\mathbf{x})$. This is defined as

$$y = f(\mathbf{x}) = \sum_{i=1}^{M} (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) + b, \quad (8)$$

where α_i and α_i^* are the Lagrange multipliers associated with each training sample $\mathbf{x}_i$, K denotes the kernel function, and b is the bias term. In this study, the Gaussian radial basis function (RBF) [48] is utilized as the kernel function.

5) Decision Tree Regression

A decision tree can be used to predict the value of a continuous dependent variable from a set of continuous predictors by constructing a predictive model with a tree-like structure. In this study, we use the classification and regression tree (CART) algorithm [49] to construct a decision tree. Using this algorithm, we construct a model that can predict the popularity score by learning simple decision rules derived from the features of the data. For each feature, the CART algorithm splits the data at different points. It then selects the part that minimizes the sum of squared errors (SSE) and generates more homogeneous subsets. The splitting process results in a fully grown tree such that the value (popularity score) obtained at each terminal node (leaf node) is the mean of all label values at the node.

6) Gradient Boosting Decision Trees

GBDT [26] is a machine learning algorithm that recursively constructs an ensemble of weak decision tree models using boosting. It has been proven highly efficient in various data mining competitions [50], [51]. The general principle of the GBDT algorithm is the sequential training of a series of simple decision tree estimators [49], where each successive tree attempts to minimize a certain loss function formed by the preceding trees. That is, in each stage, a new regression tree is sequentially added and trained based on the residual error of the previous ensemble model. The GBDT algorithm then updates all the predicted values by adding the predicted values of the new tree. This process is recursively continued until a maximum number of trees have been generated. Thus, the final prediction value of a single instance would be the sum of the predictions of all the regression trees.

V. Experiments and Results

In this section, we present the experimental setting and discuss the results.

A. Experimental Setup

1) Popularity Measurement

Social media websites allow the user to interact with the posted content in various ways, and this results in different social signals that can be utilized to measure the popularity of social content (e.g., images, texts, and videos) on these websites. For instance, on Twitter, popularity can be gauged by the number of re-tweets, whereas the number of likes or comments can be used to measure popularity on Facebook. In this study, we use Flickr as the major image sharing platform to predict the popularity of social media images. Previous studies have used various metrics to measure image popularity on Flickr. For example, Khosla *et al.* [7] determined the popularity of an image based on the number of views it received. McParlane *et al.* [16] adopted both view and comment count as the principal metrics.

The dataset used in our experiments complies with Khosla *et al.* [7], and the number of views is adopted as a popularity metric. To handle the large variation in the number of views of various photos from the dataset, the log function is applied. Moreover, the images receive views during the time they are online. Thus, a log-normalization approach is used to normalize the effect of the time factor. The score proposed in [7] can be defined as

$$\text{Score}_i = \log_2\left(\frac{p_i}{d_i}\right) + 1, \quad (9)$$

where p_i is the popularity metric (the original number of views) of image i, and d_i is the number of days since the image first appeared on Flickr.

2) Parameter Setting of Baseline Models

We implement all the baseline models using the scikit-learn machine learning library [52], [53]. In the experiments, we observe that the performance of the baseline models is significantly influenced by several hyper-parameters. Therefore, we identified the values of some important parameters of SVR as follows: C = 3, epsilon = 0.1, gamma = auto, and kernel = RBF. Regarding the DTR model, we note that when the max_depth parameter is set to 10, the best performance is achieved. Moreover, we identified several parameters of GBDT: n_estimators = 2000, max_depth = 10, and learning_rate = 0.01. Finally, the remaining parameters are set to their default values in all the models.

3) Dataset

We now present details about the real-world dataset that is used to evaluate the performance of the proposed approach. We use the Social Media Prediction (SMP-T1) dataset presented by ACM Multimedia Grand Challenge in 2017 [31], [54]. The dataset consists of approximately 432K posts collected from the personal albums of 135 different users on Flickr. Every post in the dataset has a unique picture id along with the associated user id that signifies the user who posted the picture. Additionally, some image metadata are as follows: post date (postdate), number of comments (commentcount),

number of tags in the post, whether the photo is tagged by some users or not (haspeople), and character length of the title and image caption (titlelen or deslen). Furthermore, some user-centric information, namely, average view count, group count, and average member count, is also given in the dataset. Each image has a label representing its popularity score (log-normalized views of the image). Some images selected from the dataset are shown in Fig. 5. In our experiments, 60% of the images were used for training, 20% for validation, and 20% for testing.

Fig. 5. Sample images from the dataset. The popularity of the images is sorted from more popular (left) to less popular (right).

4) Evaluation Metrics

In this study, we use the same metrics as those in the ACM Multimedia Grand Challenge [31], [54] to assess prediction accuracy: Spearman's Rho [44], MSE, and mean absolute error (MAE).

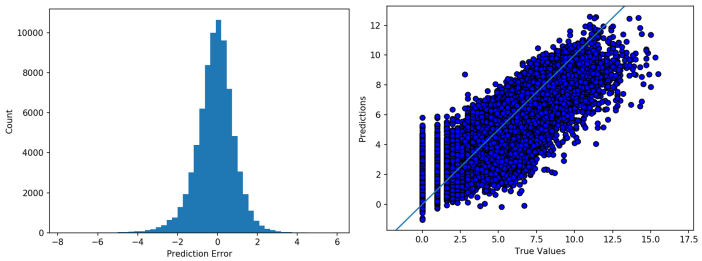
- **Spearman's Rho:** Spearman's Rho is used to calculate the correlation between the predicted popularity scores and the actual scores for the set of tested images.
- **MSE:** It is usually used to measure the average of the sum of squared prediction errors. Each prediction error represents the difference between the actual value of the data point and the predicted value obtained by the regression model. MSE has simple mathematical properties, which make its gradient easier to be calculated. Besides, it is often presented as a default metric for most of the predictive models because it is smoothly differentiable, computationally simple, and hence can be better optimized. One important limitation of MSE is that it heavily penalizes large prediction errors by squaring them. Because each error in MSE grows quadratically, the outliers in the data will greatly contribute to the total error. This means that MSE is sensitive towards outliers and puts too much weight on the effects of them, which leads to under-estimate the model performance. The drawback of MSE only becomes evident when we have outliers in our data, so using MAE is a good alternative in that case.
- **MAE:** It is a simple measure usually used to evaluate the accuracy of a regression model. It measures the average of the absolute values of individual prediction errors of the model over all samples in the test set. In MAE metric, each prediction error contributes proportionally to the total amount of error, meaning that larger errors contribute linearly to the overall error. Because we use the absolute value of the prediction error, the MAE does not indicate underperformance or overperformance of the model, i.e., whether the regression model is over-predicting the input samples or under-predicting those samples. Thus, it offers a relatively impartial comprehension of how the model performs. By taking the absolute value of the prediction error and not squaring it, the MAE becomes more robust than MSE in dealing with the outliers, as it will not heavily penalize the large errors as in MSE. Hence, MAE has its advantages and disadvantages where, on the one hand, it assists in handling outliers but, on the other hand, it fails to penalize the large prediction errors.

B. Results

Using the features extracted for model learning, we train the proposed VSCNN model to predict the popularity score. In the training stage, we use Adam [55] and stochastic gradient descent as the learning optimizer and obtain the initialized parameters for VSCNN. We set the initial learning rate to 0.001. In the experiments, we run the model for 50 training epochs over the entire training set. In each epoch, the model is iterated over batches of the training set, where the size of each batch is 20 samples. Furthermore, we add the following features to the training process: 1) The learning rate is reduced by 0.1 every 10 epochs using the learning rate scheduler function. This facilitates learning. 2) The best validation accuracy is saved using the model checkpoint function, and this assists in saving the best learning model. The cost function generally converges during the training phase. In the testing stage, the trained VSCNN model is applied to the test samples for evaluation. The evaluation results demonstrate that VSCNN can achieve a Spearman's Rho of 0.9014, an MAE of 0.73, and an MSE of 0.97. These are listed in Tables IV and V, and they will be used for comparison with the baseline models.

We add some essential visual analytics for model quality evaluation by computing the error distribution histogram, which shows the distribution of the errors made by the model when predicting the popularity score for each test sample, as shown in Fig. 6 (a). It is known that a larger number of errors close to zero on the histogram indicates a higher prediction accuracy. Moreover, Fig. 6 (b) shows a scatterplot of the actual values on the x-axis versus the predicted values obtained by the model on the y-axis. This scatterplot shows the correlation between the actual and predicted values. If the data appear to be around a straight diagonal line, then this indicates a strong correlation. Thus, a perfect regression model would yield a straight diagonal line from the data. From the results shown in Fig. 6, we observe that there are certain outliers that are not correctly predicted by the VSCNN model. Hence, we provide a discussion on these outliers below and explain in detail why our model fails to predict them.

In certain regression problems, the distribution of the target variable may have outliers (e.g., large or small values far from the mean value), which can affect the performance of the predictive model. As shown in Fig. 7, the distribution of the target variable (view counts) of the training samples is highly non-uniform in our dataset so that the proposed model attempts to minimize the prediction errors of the largest cluster of view counts of training samples. However, as the number of training samples with extremely high view counts is relatively low, it is more likely that the proposed model cannot correctly predict such high view counts, and then they will be seen as outliers in the predictive results.



(a) (b)

Fig. 6. Quality evaluation of the VSCNN model. (a) Error distribution histogram of the model, and (b) Scatterplot of true values (x-axis) versus predicted values (y-axis).

In Fig. 8, we show some good and bad predictions made using our proposed model on some images from the test set. The examples correctly predicted are shown in Fig. 8 (a). From these examples, it is noted that our model achieves superior performance with just 0.001 - 0.009 errors relative to actual scores. For example, the popularity score of the first four images in Fig. 8 (a) is correctly predicted with errors of 0.001, 0.009, 0.002, and 0.001, respectively. In addition, the popularity score of the last two images in Fig. 8 (a) is perfectly predicted with zero prediction error. On the other hand, some wrongly predicted examples are shown in Fig. 8 (b). For example, the actual popularity score of the first image in this figure is 3 while the score obtained by our model is 7.472, resulting in a substantial error of 4.472 in prediction. This disparity is due to the strong indications of some user's features of this image, such as average views and member count, which are 993.42 and 10,672, respectively, which contribute significantly to the model prediction when integrating all features. Likewise, the last two images in Fig. 8 (b) are other badly predicted examples of our proposed model. It is observed that the actual popularity score of these two images is too high. Therefore, it is suggested that our model cannot correctly predict the popularity of these images because the number of training samples with high popularity score is extremely limited in our dataset, as shown in Fig. 7.

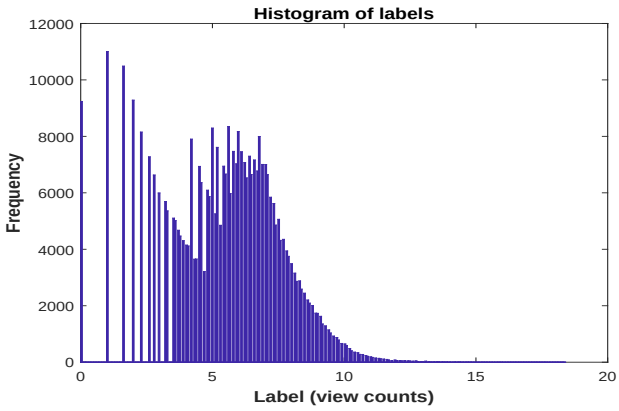
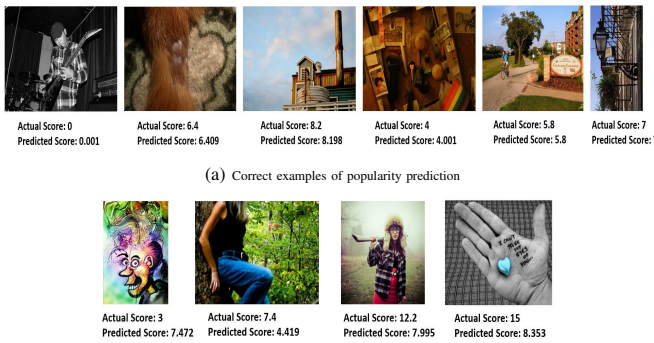


Fig. 7. A distribution of the view counts of the training samples.

1) Comparison with Baseline Models

First, we train the SCNN model using three different types of social features to explore the influence of each type of them on predicting popularity. Subsequently, we train the SCNN model using all the social features as inputs. The prediction results of the SCNN model with the different types of input



(a) Correct examples of popularity prediction

(b) Wrong examples of popularity prediction

Fig. 8. Examples of correct and wrong predictions of some images from our dataset using the VSCNN model. The actual popularity score and its corresponding predicted score are displayed below each image.

features are summarized in Table IV. In the table, we note that the user features perform exceptionally well in predicting the popularity of an image relative to the other two types of social features (i.e., post metadata and time), with a Spearman's Rho of 0.7537, MAE of 1.13, and MSE of 2.17. This indicates that the popularity of an image is closely related to the popularity of the user who uploads it. The main explanation for this is that images shared on social media by popular users have higher chances of obtaining more views. However, not all images posted by popular users are popular. To justify this, we use the popularity score and average view count as popularity metrics for images and users, respectively. We then follow the previous works [16], [56] and use the Pareto Principle (or 80-20 rule) to select a threshold to split between images with high (20%) and low (80%) popularity scores. Likewise, we set a threshold to split between users with high (20%) and low (80%) average view count. Based on these splits, the top 20% of the images and users are considered as high popular (or popular), while the remaining 80% are considered as low popular (or common). Accordingly, we find that, on average, 69.19% and 16.17% of the images posted by popular and common users, respectively, are popular. Thus, we conclude that not all images posted by popular users are always popular.

Other noteworthy features are the post metadata. The SCNN model using these features achieves 0.6590, 1.35, and 2.98 in terms of Spearman's Rho, MAE, and MSE, respectively. This shows that image-specific social features such as tag count, title length, description length, and comment count also play an important role in predicting popularity. This is to be expected as the image with more tags or a longer description/title tends to be more popular, as it has a greater opportunity to show up in the search results when people use keywords to search for images. Similarly, having more comments on the image suggests more users interact with the image, which might lead to more number of views and thus to greater popularity. By looking at the results, we also find that the time features make a significant contribution to popularity prediction. This means that the time when the image is posted may influence its popularity. For example, users tend to browse social networking sites at a particular time of the day, such as weekend leisure time, which means that images posted during that time are more likely to receive a large number of views and become popular.

TABLE IV: Performance comparison of SCNN, VCNN, and VSCNN models.

Models	Features	Spearman's Rho	MAE	MSE
SCNN	User	0.7537	1.13	2.17
	Post_Metadata	0.6590	1.35	2.98
	Time	0.5317	1.42	3.43
	All_Social	0.8809	0.79	1.13
VCNN	Color	0.3278	1.66	4.46
	Gist	0.2612	1.72	4.67
	LBP	0.3287	1.66	4.45
	Aesthetic	0.2000	1.77	4.91
	Deep	0.4101	1.61	4.13
	All_Visual	0.4168	1.58	4.08
VSCNN	Visual+Social	**0.9014**	**0.73**	**0.97**

Furthermore, we note that while each type of social feature performs well, the SCNN model achieves the best predictive performance when all the social features are combined, as shown in the fourth row of Table IV. This suggests that all the social features proposed are strongly correlated and provide complementary information to each other. Fig. 9 (a) shows a scatterplot of the predicted values obtained by SCNN and the corresponding actual values.

Similarly, we train the VCNN model using each of the individual visual features to study their effect on predicting image popularity. We also integrate all the visual features and then use them as input to the model. The evaluation results are listed in Table IV. We observe that deep learning features outperform other visual features. However, it is important to note that the VCNN model achieves the best performance in terms of all the evaluation metrics when all the visual features are combined. In addition, it is seen from the results of the VCNN model that visual features are less effective than social features in terms of image popularity prediction. This finding is consistent with the previous research studies [7], [25], [57], [58]. Nevertheless, the visual features are useful when there is no post metadata existed, or to address scenarios such as the case where no social interactions were recorded before publishing the image (e.g., because the user has just joined the social network). This indicates that image content also plays a critical role in popularity prediction, and it may complement the social features.

A scatterplot of the predicted values obtained using VCNN and the corresponding actual values is shown in Fig. 9 (b). Finally, we compare the performance of the proposed model with the best performance of both VCNN and SCNN in terms of all the evaluation metrics, and the results are listed in Table IV. It can be seen that VSCNN outperforms VCNN and SCNN, with a relative improvement of 2.33% (in SCNN) and 116.27% (in VCNN) in terms of Spearman's Rho, and with a decrease of 7.59% (in SCNN) and 53.80% (in VCNN), and of 14.16% (in SCNN) and 76.23% (in VCNN) in terms of MAE and MSE, respectively.

Subsequently, we train the other four baseline models (i.e., LR, SVR, DTR, and GBDT) using each single feature and various combinations thereof to demonstrate the effectiveness of the proposed features in predicting image popularity. The predictions are shown in Table V, where it can be seen that the user feature yields the best results. This indicates that the characteristics of the person who posts a photo determine its popularity to a large extent. Further, we notice that post metadata and time features are also good predictors.

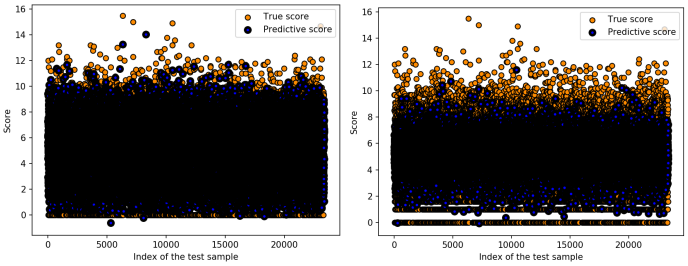


Fig. 9. Scatterplots of the predicted values obtained using the CNN-based baseline models and their corresponding ground truth values. (a) SCNN, and (b) VCNN.

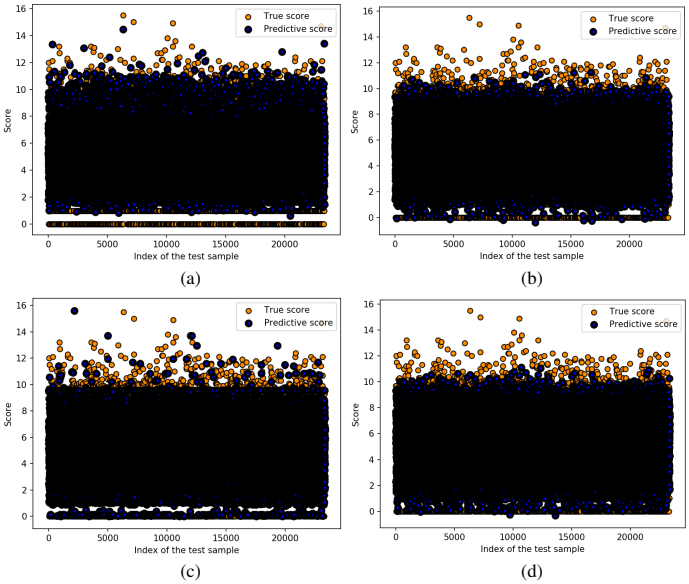


Fig. 10. Scatterplots of the predicted values obtained using the four machine learning baseline models and their corresponding ground truth values. (a) LR, (b) SVR, (c) DTR, and (d) GBDT.

Additionally, when all social context features are combined and used as inputs, the performance improves significantly for all the models, and GBDT achieves the best performance in terms of all the evaluation metrics.

The deep learning feature is quite important and outperforms the other visual features, namely, color, gist, LBP, and aesthetic, even though these features perform fairly well in all models. Nevertheless, the performance of all models is improved when all visual features are combined. Moreover, it is notable that combining visual and social features lead to significant improvement in the performance of all the models compared with that exhibited using either set of these features independently.

Fig. 10 shows scatterplots of the predicted values obtained using the four machine learning baseline models and their corresponding ground truth values. As presented in Table V and shown in Fig. 10, GBDT outperforms all other machine learning models, with a relative improvement from 5.16% (in SVR) to 19.57% (in LR) in terms of Spearman's Rho, and with decreases from 13.98% to 34.43% and from 25.32% to 52.87% in terms of MAE and MSE, respectively. Finally, we compare the performance of the proposed VSCNN model with the best performance obtained by each of the four baseline models (LR, SVR, DTR, and GBDT), and the results are shown in Table V. Compared with GBDT, VSCNN improves the prediction performance by approximately 2.69%, 8.75%,

and 15.65% in terms of Spearman's Rho, MAE, and MSE, respectively.

Fig. 11 shows the best prediction performance for all the models in terms of the three evaluation metrics. From this figure, we observe that VSCNN outperforms all the six baseline models in predicting the popularity of an image. Overall, VSCNN achieves the best prediction performance, with the highest Spearman's Rho (0.9014) and lowest MAE and MSE (0.73 and 0.97, respectively). This suggests that CNNs are more powerful than other machine learning methods in processing heterogeneous information for popularity prediction. Another significant finding is that both the social and image content features are essential and complement each other in predicting image popularity on photo-sharing websites.

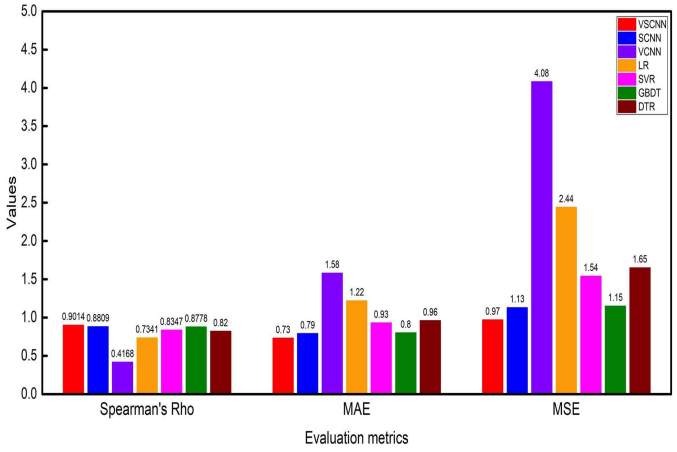


Fig. 11. Best prediction performances for all the models in terms of Spearman's Rho, MAE, and MSE metrics.

2) Comparison with state-of-the-art methods on SMP-T1 dataset

The ACM Multimedia Grand Challenge 2017 presented a Social Media Prediction Task (SMP-T1) in the shape of challenge [31], [54]. This challenge's task is to predict the popularity of images posted by users on social media. Several teams participated in this challenge, and they proposed different models based on the provided SMP-T1 dataset. We compare the performance of the proposed VSCNN model with that of these models, and the evaluation results are listed in Table VI. It can be seen that VSCNN outperforms all the other models. Compared with the best team's model (i.e., TaiwanNo.1 SMP-T1), VSCNN improves the prediction performance by approximately 9.02%, 31.62%, and 52.75% in terms of Spearman's Rho, MAE, and MSE, respectively.

In addition to the aforementioned compared models, the multimodal approach presented in [32] integrates several multimodal information extracted from the same SMPT1 dataset into a CNN model for predicting the popularity of images. Although this approach adopts multimodal features (e.g., image, textual, contextual, and social features) for popularity prediction, it ignores some other important features, such as low-level computer vision, aesthetic, and time features. It also adopts the early fusion scheme to merge the features extracted from various modalities into a single large feature vector before feeding them to the CNN regression model. Even though early fusion can create a joint representation of the input features from multiple modalities, it requires the features to be extremely engineered and preprocessed so that they align well before the fusion process. It also suffers from the difficulty of representing the time synchronization between the multimodal features. Moreover, the increase in the number of modalities makes it hard to learn the cross-correlation amongst the overly heterogeneous features. Eventually, a single model is used to make predictions by assuming that the model is well suited for all modalities. However, the architecture of the CNN model used in [32] is not sufficiently powerful to process features from different modalities and then accurately predict the popularity of an image.

Unlike [32], our model employs the late fusion scheme in which the features of each modality (i.e., visual and social features) are examined and trained independently using two CNNs with highly designed architecture. The obtained results are then fused using a merged layer into another network for further processing and obtaining the final prediction. The fusion process in our model becomes easy to execute, and does not suffer from the data representation problem which early fusion scheme has. This is because the semantic vectors resulting from the two CNNs models usually have the same form of data. Also, late fusion allows us to use the most suitable model for analyzing each modality and learning its features, allowing more flexibility. Furthermore, the robust interpretation of incomplete and inconsistent multimodal input becomes more reliable at later stages, because of more semantic knowledge becomes available from the various sources. Due to these advantages, the late fusion scheme is extensively used in multimodal systems [7], [33], [57]. To confirm the efficiency of our model, we also compare it with the multimodal approach proposed in [32], and the prediction results are summarized in Table VI. It is clearly seen that VSCNN outperforms the multimodal approach, with a relative improvement of 20.19% in terms of Spearman's Rho, and a decrease of 34.82% and 59.41% in terms of MAE, and MSE, respectively.

3) Late and Early Fusion Schemes for the VSCNN model

The framework proposed in this study adopts the late fusion scheme, that is, we first employ two convolutional neural networks to process visual features and the corresponding social context information individually. Then, the outputs of these two networks are merged into another network that fully connected all the information into a final layer of the deep architecture. We also test the early fusion scheme by integrating visual and social features at the input of the convolutional layers. The early fusion scheme, denoted by VSCNN-EF, replaces the visual and social networks in Fig. 2 with a unified CNN whose inputs comprise fused visual-social features obtained by concatenating visual and social features of a given image into a final feature vector of 4,744 dimensions (4,710 visual and 34 social). Then, PCA [45] is applied to reduce the dimensionality of this vector from 4,744 to 20 and to select only the most prevalent features. The numbers of parameters of VSCNN and VSCNN-EF are of the same order. These schemes are optimized and tested, and then their prediction performance is compared in terms of the three performance metrics. The results are listed in Table VII. It

TABLE V: Performance comparison of LR, SVR, DTR, GBDT, and VSCNN models.

Features	LR			SVR			DTR			GBDT			VSCNN		
	Spearman's Rho	MAE	MSE	Spearman's Rho	MAE	MSE	Spearman's Rho	MAE	MSE	Spearman's Rho	MAE	MSE	Spearman's Rho	MAE	MSE
Color	0.0856	1.81	5.10	0.2569	1.72	4.74	0.1915	1.76	4.93	0.3381	1.66	4.41			
Gist	0.1337	1.79	5.04	0.3209	1.67	4.55	0.1436	1.80	5.12	0.3176	1.69	4.51			
LBP	0.1546	1.79	5.04	0.3028	1.69	4.63	0.1640	1.78	5.01	0.3126	1.68	4.52			
Aesthetic	0.1221	1.8	5.06	0.1866	1.77	4.97	0.1661	1.79	5.00	0.2040	1.77	4.88			
Deep	0.3701	1.66	4.43	0.4754	1.53	3.88	0.2330	1.76	4.96	0.4403	1.59	4.05			
All_Visual	0.3837	1.65	4.35	0.5018	1.50	3.73	0.2384	1.76	4.95	0.4890	1.53	3.82			
user	0.6449	1.41	3.17	0.7548	1.12	2.18	0.7579	1.12	2.15	0.7580	1.12	2.15			
Post_Metadata	0.5266	1.68	4.41	0.6126	1.42	3.28	0.6682	1.32	2.91	0.6962	1.27	2.72			
Time	0.1337	1.80	5.00	0.2681	1.70	4.65	0.3485	1.61	4.19	0.6285	1.29	2.84			
All_Social	0.7114	1.28	2.72	0.8292	0.94	1.58	0.8049	1.01	1.74	0.8611	0.86	1.30			
Visual+Social	0.7341	1.22	2.44	0.8347	0.93	1.54	0.8200	0.96	1.65	0.8778	0.80	1.15	**0.9014**	**0.73**	**0.97**

TABLE VI: Comparison with the state-of-the-art methods on SMP-T1 dataset.

Methods		Spearman's Rho	MAE	MSE
SMP Challenge Teams [31, 54]	TaiwanNo.1 SMP-T1	0.8268	1.0676	2.0528
	heihei SMP-T1	0.8093	1.1059	2.1767
	NLPR_MMC_Passerby SMP-T1	0.7927	1.1783	2.4973
	BUPTMM SMP-T1	0.7723	1.1733	2.4482
	bluesky SMP-T1	0.7406	1.2475	2.7293
	WePREdictIt SMP-T1	0.5631	1.6278	4.2022
	FirstBlood SMP-T1	0.6456	1.6761	6.3815
	ride_snail_to_race SMP-T1	-0.0405	2.4274	9.2715
	CERTH-ITI-MKLAB SMP-T1	0.3554	3.8178	19.3593
Multimodal approach [32]		0.75	1.12	2.39
VSCNN		**0.9014**	**0.73**	**0.97**

TABLE VII: Performance comparison of VSCNN and VSCNN-EF models.

Models	Spearman's Rho	MAE	MSE
VSCNN	**0.9014**	**0.73**	**0.97**
VSCNN-EF	0.8898	0.76	1.07

is obvious that VSCNN consistently outperforms VSCNN-EF, suggesting that the proposed late fusion scheme, which initially processes visual and social information independently and merges them later, is better than the early fusion scheme, which incorporates the heterogeneous data at the beginning.

VI. Conclusion

Recently, deriving an effective computational model to characterize human behavior or predict decision making has become an emergent topic. In this study, we developed a multimodal deep learning framework for predicting the popularity of images on social media. First, we analyzed and extracted different types of image visual content features and social context information that greatly affect image popularity. Then, we proposed a novel CNN-based visual-social computational model for image popularity prediction, called VSCNN. This model uses individual networks to process input data with different modalities (i.e., visual and social features), and the outputs from these networks are then integrated into a fusion network to learn joint multimodal features and estimate the popularity score. We trained the proposed model in an end-to-end manner. The experimental results on the provided dataset demonstrated the effectiveness of the proposed model in predicting image popularity. Further experiments demonstrated that VSCNN achieved a notably superior prediction performance. Specifically, it outperformed four traditional machine learning schemes, two CNN-based models, and other state-of-the-art methods in terms of three standard evaluation metrics (i.e., Spearman's Rho, MAE, and MSE). This emphasizes the effectiveness of the proposed model in combining visual and social information to predict the popularity of an image.

In the future, we will extend our work by considering not only internal but also external factors that may affect images popularity, such as real-world events. Meanwhile, we will investigate the influence of various aspects on image popularity based on geographical location and cultural background. Additionally, we plan to use a generative model as suggested in [59] to automatically generate natural sentences describing the content and title for each image in the SMP-T1 dataset, and using also an image annotation model as proposed in [60] to create a set of keywords (hashtags) that are related to the content of the image. We then incorporate the obtained textual information into our model to explore its effect on image popularity. Finally, we aim to optimize the parameters and overall structures of the CNNs used in the proposed model so that prediction performance may be improved.

References

[1] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Communications of the ACM*, vol. 53, no. 8, pp. 80–88, 2010.

[2] X. Niu, L. Li, T. Mei, J. Shen, and K. Xu, "Predicting image popularity in an incomplete social media community by a weighted bi-partite graph," in *2012 IEEE International Conference on Multimedia and Expo*. IEEE, 2012, pp. 735–740.

[3] S. Petrovic, M. Osborne, and V. Lavrenko, "Rt to win! predicting message propagation in twitter." *ICWSM*, vol. 11, pp. 586–589, 2011.

[4] A. O. Nwana, S. Avestimehr, and T. Chen, "A latent social approach to youtube popularity prediction," in *Global Communications Conference (GLOBECOM), 2013 IEEE*. IEEE, 2013, pp. 3138–3144.

[5] H. Pinto, J. M. Almeida, and M. A. Gonçalves, "Using early view patterns to predict the popularity of youtube videos," in *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 2013, pp. 365–374.

[6] D. A. Shamma, J. Yew, L. Kennedy, and E. F. Churchill, "Viral actions: Predicting video view counts using synchronous sharing behaviors." in *ICWSM*, 2011.

[7] A. Khosla, A. Das Sarma, and R. Hamid, "What makes an image popular?" in *Proceedings of the 23rd international conference on World wide web*. ACM, 2014, pp. 867–876.

[8] C. Li, Y. Lu, Q. Mei, D. Wang, and S. Pandey, "Click-through prediction for advertising in twitter timeline," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 1959–1968.

[9] F. Figueiredo, J. M. Almeida, M. A. Gonçalves, and F. Benevenuto, "On the dynamics of social media popularity: A youtube case study," *ACM Transactions on Internet Technology (TOIT)*, vol. 14, no. 4, p. 24, 2014.

[10] C.-C. Wu, T. Mei, W. H. Hsu, and Y. Rui, "Learning to personalize trending image search suggestion," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 2014, pp. 727–736.

[11] N. Aggrawal, A. Ahluwalia, P. Khurana, and A. Arora, "Brand analysis framework for online marketing: ranking web pages and analyzing popularity of brands on social media," *Social Network Analysis and Mining*, vol. 7, no. 1, p. 21, 2017.

[12] M. A. Gonçalves, J. M. Almeida, L. G. dos Santos, A. H. Laender, and V. Almeida, "On popularity in the blogosphere," *IEEE Internet Computing*, vol. 14, no. 3, pp. 42–49, 2010.

[13] B. Wu, T. Mei, W.-H. Cheng, Y. Zhang *et al.*, "Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition." in *AAAI*, 2016, pp. 272–278.

[14] S. Cappallo, T. Mensink, and C. G. Snoek, "Latent factors of visual popularity prediction," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 2015, pp. 195–202.

[15] F. Gelli, T. Uricchio, M. Bertini, A. Del Bimbo, and S.-F. Chang, "Image popularity prediction in social media using sentiment and context features," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 907–910.

[16] P. J. McParlane, Y. Moshfeghi, and J. M. Jose, "Nobody comes here anymore, it's too crowded; predicting image popularity on flickr," in *Proceedings of International Conference on Multimedia Retrieval*. ACM, 2014, p. 385.

[17] S. Aloufi, S. Zhu, and A. El Saddik, "On the prediction of flickr image popularity by analyzing heterogeneous social sensory data," *Sensors*, vol. 17, no. 3, p. 631, 2017.

[18] L. Hong, O. Dan, and B. D. Davison, "Predicting popular messages in twitter," in *Proceedings of the 20th international conference companion on World wide web*. ACM, 2011, pp. 57–58.

[19] E. F. Can, H. Oktay, and R. Manmatha, "Predicting retweet count using visual cues," in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 2013, pp. 1481–1484.

[20] R. Bandari, S. Asur, and B. A. Huberman, "The pulse of news in social media: Forecasting popularity." *ICWSM*, vol. 12, pp. 26–33, 2012.

[21] K. Almgren, J. Lee *et al.*, "Predicting the future popularity of images on social networks," in *Proceedings of the The 3rd Multidisciplinary International Social Networks Conference on SocialInformatics 2016, Data Science 2016*. ACM, 2016, p. 15.

[22] L. C. Totti, F. A. Costa, S. Avila, E. Valle, W. Meira Jr, and V. Almeida, "The impact of visual attributes on online image diffusion," in *Proceedings of the 2014 ACM conference on Web science*. ACM, 2014, pp. 42–51.

[23] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," in *Advances in neural information processing systems*, 1997, pp. 155–161.

[24] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[25] W. Wang and W. Zhang, "Combining multiple features for image popularity prediction in social media," in *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 2017, pp. 1901–1905.

[26] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[29] F. S. Abousaleh, T. Lim, W.-H. Cheng, N.-H. Yu, M. A. Hossain, and M. F. Alhamid, "A novel comparative deep learning framework for facial age estimation," *EURASIP Journal on Image and Video Processing*, vol. 2016, no. 1, p. 47, 2016.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[31] B. Wu, W.-H. Cheng, Y. Zhang, Q. Huang, J. Li, and T. Mei, "Sequential prediction of social media popularity with deep temporal context networks," *arXiv preprint arXiv:1712.04443*, 2017.

[32] M. Meghawat, S. Yadav, D. Mahata, Y. Yin, R. R. Shah, and R. Zimmermann, "A multimodal approach to predict social media popularity," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2018, pp. 190–195.

[33] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[34] A.-E. Hassanien and A. Abraham, *Computational Intelligence in Multimedia Processing: Recent Advances*. Springer, 2008, vol. 96.

[35] M. Heikkila and M. Pietikainen, "A texture-based method for modeling the background and detecting moving objects," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, pp. 657–662, 2006.

[36] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[37] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, "Local binary patterns and its application to facial image analysis: a survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 6, pp. 765–781, 2011.

[38] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[39] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *European Conference on Computer Vision*. Springer, 2008, pp. 386–399.

[40] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1. IEEE, 2006, pp. 419–426.

[41] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1930–1943, 2013.

[42] X. Chen, Q. Zhang, M. Lin, G. Yang, and C. He, "No-reference color image quality assessment: from entropy to perceptual quality," *EURASIP Journal on Image and Video Processing*, vol. 2019, no. 1, p. 77, 2019.

[43] https://keras.io/.

[44] C. Spearman, "The proof and measurement of association between two things," *The American journal of psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[45] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.

[46] J. Lakshmi, "Stochastic gradient descent using linear regression with python," *International Journal of Advanced Engineering Research and Applications*, vol. 2, no. 8, pp. 519–525, 2016.

[47] N. Cristianini, J. Shawe-Taylor *et al.*, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[48] M. Hofmann, "Support vector machines—kernels and the kernel trick," *Notes*, vol. 26, 2006.

[49] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[50] S. B. Taieb and R. J. Hyndman, "A gradient boosting approach to the kaggle load forecasting competition," *International journal of forecasting*, vol. 30, no. 2, pp. 382–394, 2014.

[51] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016, pp. 785–794.

[52] http://scikit-learn.org/stable/.

[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[54] https://social-media-prediction.github.io/MM17PredictionChallenge/index.html.

[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[56] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, 2007, pp. 1–14.

[57] J. Lv, W. Liu, M. Zhang, H. Gong, B. Wu, and H. Ma, "Multi-feature fusion for predicting social media popularity," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1883–1888.

[58] X. Huang, Y. Gao, Q. Fang, J. Sang, and C. Xu, "Towards SMP challenge: stacking of diverse models for social image popularity prediction," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1895–1900.

[59] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 652–663, 2016.

[60] B. Wu, W. Chen, P. Sun, W. Liu, B. Ghanem, and S. Lyu, "Tagging like humans: Diverse and distinct image annotation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7967–7975.

Fatma S. Abousaleh received the B.S. degree in mathematics and computer science from Zagazig University, Zagazig, Egypt, in 2003, and the M.S. degree in computer science from Ain Shams University, Cairo, Egypt, in 2010. She is currently pursuing the Ph.D. degree with the Taiwan International Graduate Program in Social Networks and Human-Centered Computing, Institute of Information Science, Academia Sinica, Taipei, Taiwan, and the Department of Computer Science, Faculty of Science, National Chengchi University, Taipei, Taiwan. Her research interests include image processing and social network analysis.

Wen-Huang Cheng received the B.S. and M.S. degrees in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 2002 and 2004, respectively, where he received the Ph.D. (Hons.) degree from the Graduate Institute of Networking and Multimedia in 2008. He is a Professor with the Institute of Electronics, National Chiao Tung University (NCTU), Hsinchu, Taiwan, where he is the Founding Director with the Artificial Intelligence and Multimedia Laboratory (AIMMLab). Before joining NCTU, he led the Multimedia Computing Research Group at the Research Center for Information Technology Innovation (CITI), Academia Sinica, Taipei, Taiwan, from 2010 to 2018. His current research interests include multimedia, artificial intelligence, computer vision, machine learning, social media, and financial technology. He has received numerous research and service awards, including the 2018 MSRA Collaborative Research Award, the Outstanding Reviewer Award of 2018 IEEE ICME, the 2017 Ta-Yu Wu Memorial Award from Taiwan's Ministry of Science and Technology (MOST), the 2017 Significant Research Achievements of Academia Sinica, the 2016 Y. Z. Hsu Scientific Paper Award, the Outstanding Youth Electrical Engineer Award from the Chinese Institute of Electrical Engineering in 2015, the Top 10% Paper Award from the 2015 IEEE MMSP, the K. T. Li Young Researcher Award from the ACM Taipei/Taiwan Chapter in 2014. He is APSIPA Distinguished Lecturer.

Neng-Hao Yu received Ph.D. degree in the Graduate Institute of Networking and Multimedia from National Taiwan University, Taipei, Taiwan, in 2011. He is an Assistant Professor with the Department of Design, National Taiwan University of Science and Technology (NTUST), Taipei, Taiwan. Before joining NTUST, he created and directed the Innovative User Interface Lab (IUI Lab) at the Department of Computer Science, National Chengchi University (NCCU), Taipei, Taiwan, as an Assistant Professor from 2011 to 2018. His current research interests include human-computer interaction, user experience design, virtual reality, and multimedia technology. He has been a reviewer and chair for a wide range of international conferences in the field such as CHI, UIST, MobileHCI, Siggraph Asia and ChineseCHI.

Yu Tsao (M'09) received the B.S. and M.S. degrees in electrical engineering from National Taiwan University in 1999 and 2001, respectively, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology in 2008. From 2009 to 2011, He was a Researcher with the National Institute of Information and Communications Technology, Japan, where he was involved in research and product development in automatic speech recognition for multilingual speech-to-speech translation. He is currently an Associate Research Fellow with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. His research interests include speech and speaker recognition, acoustic and language model-ing, audio-coding, and bio-signal processing. He received the Academia Sinica Career Development Award in 2017.