Social Media Popularity Prediction Based on Multi-Modal Self-Attention Mechanisms

HUNG-HSIANG LIN\textsuperscript{1}, JIUN-DA LIN\textsuperscript{1}, JOSE JAENA MARI OPLE\textsuperscript{1}, JUN-CHENG CHEN\textsuperscript{2}, (Member, IEEE), AND KAI-LUNG HUA\textsuperscript{1}\textsuperscript{1}
\textsuperscript{1}Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan
\textsuperscript{2}Research Center for Information Technology Innovation, Academia Sinica, Taipei 115, Taiwan
Corresponding author: Kai-Lung Hua (hua@mail.ntust.edu.tw)

This work was supported in part by the Center for Cyber-Physical System Innovation and the Center of Intelligent Robots from The Featured Areas Research Center Program within the Framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan, in part by the Ministry of Science and Technology of Taiwan under Grant MOST109-2221-E-011-125-MY3 and Grant MOST110-2923-E-011-004, and in part by the Wang Jhan-Yang Charitable Trust Fund under Contract WJY 2020-HR-01.

\textbf{ABSTRACT} Popularity prediction using social media is an important task because of its wide range of real-world applications such as advertisements, recommendation systems, and trend analysis. However, this task is challenging because social media is affected by multiple factors that cannot be easily modeled (e.g. quality of content, relevance to viewers, real-life events). Usually, other methods adopt the greedy approach to include as many modalities and factors as possible into their model but treat these features equally.

To solve this phenomenon, our proposed method leverages the self-attention mechanism to effectively and automatically fuse different features to achieve better performance for the popularity prediction of a post, where the features used in our model can be mainly categorized into two modalities, semantic (text) and numeric features. With extensive experiments and ablation studies on the training and testing data of the challenging ACM Multimedia SMPD 2020 Challenge dataset, the evaluation results demonstrate the effectiveness of the proposed approach as compared with other methods.

\textbf{INDEX TERMS} Social media popularity prediction, ensemble learning, multi-modality, self-attention, image caption.

\textbf{I. INTRODUCTION}

Social media provides a public platform to easily exchange information with each other, and nowadays people spend a lot of time every day on various social media platforms. Since social media occupies a large part of the daily lives of modern people, many people are interested in researching how to extract data from social media. An example of information that could be gained from social media is the popularity score. Specifically, this score tells how many people viewed a post, and a larger number of views means more influence. Social media popularity prediction (SMP) is the task of estimating the popularity score using the available data of a given social media post.

Estimating the popularity score is hard because of the many and complex factors that affect popularity. Quality of content and relevance to viewers are some of the factors, and these are difficult to measure. Other factors such as real-life events are tough to include in a prediction model. Recent SMP methods attempt to tackle these complex factors by adding more modalities \cite{4}, \cite{5}, \cite{7}, \cite{12}, \cite{17}, such as images \cite{14}, \cite{37}, relationship networks \cite{25}, temporal context \cite{13}, tags, and categories.

Although increasing the number of modalities is a good approach to the works, it also increases the complexity of the model, in terms of architecture, memory consumption, number of modules, etc. Alternatively, the paper \cite{7}, \cite{26}--\cite{30} is also a multi-modal approach but in its pipeline, it represented images as captions (i.e. texts). Different modalities could be converted to another modality using existing technologies. Image captioning converts images to texts. There exist speech-to-text methods already. From the social graph of a post, we could extract different numeric values, such as the number of the neighbors for each node.

Moreover, the popularity of posts may be affected by user information. Many studies have shown that there is a high correlation between image popularity and users \cite{20}, \cite{31}, \cite{32}. One of the reasons is that the users have their own followers,
different users may have different numbers of followers. Generally, posts written by the user with more followers have a higher chance to receive more views and likes. And the temporal and spatial information may affect the popularity as well, the earlier post should get more people’s attention, and if the user uploads the post in a special location, it will attract more attention too.

In this paper, we proposed a network that exploits semantic (text) and numerical (number) modalities to estimate the popularity of a social media post based on the self-attention mechanism. Due to the data type discrepancy, we divided the data into semantic and numerical branches. In the semantic branch, the image contents are transferred to caption texts and tags, all of the textual features are converted into tokens, each token has an associated with word embedding [23], since the attention mechanism [9] is shown effective to extract contextual information, to better aggregate the sequence of embedding, we also develop a feature attention mechanism for the purpose, which can deal with dispensing recurrence, and convolutions entirely. Using only the semantic features modality is not sufficient for some types of social media posts, so we used the numerical features as well which can be easily converted into scalars, such as timestamps, geolocation. After preprocessing, we extracted and fused the features in both modalities respectively, and assemble two models to calculate the popularity score. The contributions of this work are 3 fold:

- We designed a network that adopts an attention mechanism and exploits multiple features in two modalities to perform model ensemble, the network can be easily extended to include more different modalities furthermore, which is able to solve problems with heavy categories.
- We analyzed the influence of semantic features on the model performance. Moreover, we generated additional numerical features, the result indicates the derived features are beneficial to improve our network performance.
- We demonstrated that our method outperforms the other state-of-the-art methods in Social Media Popularity Dataset.

II. RELATED WORKS

Recently, social media popularity prediction receives much attention, there is a lot of research on this topic in both academia and industry. These studies cover a wide range of applications such as recommendations, image and video annotation, personality detection, human behavior prediction, and media popularity prediction. They share a common way to figure out the final popularity score which involves feature extraction and using regression models [33], [29].

Khosla et al. [1] used the image content and the user context to predict the image popularity based on millions of images. They methodically analyzed the impact of low-level, middle-level, and high-level features on prediction accuracy.

Wu et al. [2] merged multiple time-scale dynamics into a sequential prediction of popularity. In [3], Van Zwol studied the characteristics of users’ social behavior on Flickr. He revealed that photos received the majority of their views within the first two days of being uploaded. Moreover, the popularity of images was influenced by the owners’ contacts and social groups to which he or she belonged. There are also several works studied on other platforms. Hessel et al. [4] analyzed that the combination of visual and textual modalities generally leads to the best accuracies for predicting relative popularity on Reddit. Mazloom et al. [5] proposed that there are several important features, called engagement parameters, such as sentiment, vividness, and entertainment. They used these parameters for predicting the popularity of brand-related posts on Instagram.

Many researchers predicted social media popularity based on ACM Multimedia Challenge 2019 or earlier [19], [29], [30], [29]. For example, Hsu et al. [7] employed word-to-vector models to encode the text information and image semantic features extracted by image caption. Ding et al. [15] fused textual and numerical data with deep neural network techniques to predict the popularity score. Li et al. [19] presented a Doc2Vec model and an effective text-based feature fusion engineering, but these works only concatenated the different types of features then fed them to the regression model, they did not consider the correlation between different features. Hsu et al. [21] proposed an iterative refinement method to compensate for prediction error and [22] computed the view count of a post by residual learning. However, this works only adopted limited types of social media data, there are still a lot of useful data that can improve the performance of prediction.

With the rapid development of machine learning or deep learning, many works present vision-based applications, for example, Lin et al. [35] employed multiple residual dense blocks to perform pattern removal. Yeh et al. [36] proposed a visual attention module to enhance image classification capability. Ortis et al. [38] considered visual and textual information to perform sentiment analysis through the SVM...
classifier, and Katsurai et al. [39] exploited the SentiWordNet to retrieve sentiment information and fused the visual and textual views to classify the post belongs positive or negative via SVM as well, however, the SVM model cannot afford the large-scale dataset, and it is hard to apply to high dimensional data.

In 2016, He et al. [10] proposed a novel deep learning architecture, Residual Network (ResNet), generally, the deeper network will get better performance, however, there exists a degradation problem: when the number of layers increases, the accuracy will decrease. ResNet has an identity mapping mechanism to solve problems of gradient vanishing and explosion. In this case, we used a ResNet-50 [10] based on pre-trained ImageNet weight with average pooling over $K \times M$ grids in the image, yielding $N = KM$ output vectors of 2048 dimensions each. For every image, input images are resized, center-cropped at 224 x 224, and normalized, which is the standard for ResNet.

Transfer learning from supervised ImageNet features is the well-known approach [13] in computer vision and self-supervised word and sentence embedding [12] has become ubiquitous in natural language processing. However, fine-tuning self-supervised language modeling system [34] revolutionized the field recently, language modeling enables systems to learn embedding in a contextualized method, and it yielded even better results on a variety of tasks. We built our self-attention model based on the multi-modal BiTransformer method [14], which enhances the strength of text-only self-supervised representations with the power of state-of-the-art CNN architectures.

### III. METHODOLOGY

Our method addresses the tasks of Social Media Popularity Prediction (SMP). In SMP, we try to estimate a popularity score from a social media post, specifically, a Flickr post (Figure 1). We proposed a multi-modal approach that applied semantic and numerical features. Because of the datatype disparity, the features are divided into two partitions, and different models are used to extract the features. The system overview of the proposed method is shown in Figure 2. We will explain the components in more detail subsequently.

#### A. SEMANTIC FEATURE EXTRACTION ($Z_{text}$)

A post contains a lot of information, such as image, description, or attribute data, they don’t have a fixed length, some of them are sentences and some of them are individual words. The proposed method fused the following features:

- **Caption Features** — Social media posts could have images or videos attached. To simplify the pipeline of our method, these attached images and videos are converted to text using a pre-trained captioning model [7, 12] and are treated similarly as textual features.

- **User-Related Features** — User-related information is directly related to the user who created the social media post. For simplicity, we used two features in this type: Unique User ID and Pro-member Flag (i.e. Flickr paid membership). The user ID can provide our model with useful information to distinguish unique users, for instance, celebrities usually have higher popularity than ordinary people because of their inherent reputation. And according to the analysis, the user who is pro-member has a higher popularity score on average.

- **Categorical Features** — A social media post could be categorized using different systems. In this paper, a Flickr post has different levels of categorization, which are: (main) category, subcategory, concept descriptions. There are 11 classes for categories, 77 classes for subcategories, and 668 classes of concept descriptions.

- **Tag Features** — Tag features are composed of several keywords given by the user when they are creating a
In this portion, we define another network \( G \), which is based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Furthermore, this architecture allows flexibly scanning the input sequence, hence reliably extracting the needed features.

The attention mechanism considers all of the encoder hidden states when computing the token information \( T_c \), the weight vector \( A_t \) is computed by comparing the current hidden state with each source hidden state, then the weighted word vector is obtained by dot product and softmax operation, shown in formula 2. We adopted this self-attention module in the feature aggregator \( G \), in order to derive a weighted context vector \( T_a \) that captures relevant source-side information to help determine the important ratio in each position of the category information from \( X_{\text{text}} \). Finally, in order to leverage attentional context information and keep original information, we applied a concatenation layer to fuse vector \( T_a \) and source data \( T_r \), namely \( C_r \).

\[
T_a = \text{softmax} \left( T_c^T T_r \right) T_c
\]  

(2)

**B. NUMERIC FEATURE EXTRACTION (**\( Z_{\text{num}} \)**)

Numerical features are features that could be expressed quantitatively (i.e. ordinal, interval, ratio). Examples of these features in terms of a social media post are timestamps, longitude, and latitude. Other numeric values could be computed using the existing features such as tag count and posting frequency.

1) PREPROCESSING

The semantic features are just words and cannot be converted yet to a feature vector \( Z_{\text{text}} \). Technically, the raw features are a bunch of text. We lowercase the raw text and convert the raw features to word embeddings using a tokenizer and a word embedding model [23]. The final output is a sequence of word embeddings, which we define as a \( X_{\text{text}} \).

2) FEATURE EXTRACTION

In this portion, we define another network \( G_{\phi} \) that outputs \( Z_{\text{text}} \) as the feature representation given a \( X_{\text{text}} \). We define:

\[
Z_{\text{text}} = G_{\phi}(X_{\text{text}})
\]  

(1)

where \( \phi \) are the parameters of network \( G \). To handle sequential data, we used the proposed self-attention-based feature aggregator \( G_{\phi} \) for feature aggregation, which is based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Furthermore, this architecture allows flexibly scanning the input sequence, hence reliably extracting the needed features.

The attention mechanism considers all of the encoder hidden states when computing the token information \( T_c \), the weight vector \( A_t \) is computed by comparing the current hidden state with each source hidden state, then the weighted word vector is obtained by dot product and softmax operation, shown in formula 2. We adopted this self-attention module in the feature aggregator \( G_{\phi} \), in order to derive a weighted context vector \( T_a \) that captures relevant source-side information to help determine the important ratio in each position of the category information from \( X_{\text{text}} \). Finally, in order to leverage attentional context information and keep original information, we applied a concatenation layer to fuse vector \( T_a \) and source data \( T_r \), namely \( C_r \).

**C. ENSEMBLE REGRESSOR MODEL**

The architecture of the network follows similarly to [14], stacks multiple regression models in multiple layers, which enables the different models to complement and reinforce each other. We define an ensemble regressor [7] network \( F_{\theta} \) such that:

\[
p = F_{\theta}(Z)
\]  

(5)

where \( p \) is the computed popularity score, \( Z \) is the extracted features from the social media post, and \( \theta \) is the parameters of \( F \). The features \( Z \) is computed using two intermediate values as seen in formula 6:

\[
Z = \text{Concat}(Z_{\text{text}}, Z_{\text{num}})
\]  

(6)

where \( \text{Concat}(\cdot) \) is a feature-wise concatenation function, \( Z_{\text{text}} \) is the intermediate value generated from the nominal or textual features of the input post, while \( Z_{\text{num}} \) is computed using the numeric (i.e. ordinal, interval, ratio) features from the social media post.

**D. LEARNING OBJECTIVE**

We train our model, end-to-end, using Mean Absolute Error (MAE). MAE is computed as follows:

\[
\text{MAE} = \frac{1}{n-1} \sum_{k=1}^{n} |\hat{P}_k - P_k|
\]  

(7)
IV. EXPERIMENTS

In this section, we describe the implementation details of the proposed method, evaluation metric, and evaluation results. We also conduct an ablation study to analyze the importance of each feature used in our proposed algorithm.

A. DATASET

SMPD 2020 dataset [24] contains 305K posts with over 70K users from Flickr. It has rich information including user profiles, images, times, location, tags, and other metadata. To further improve our model, we have also generated some features from existing data such as posting frequency and tag count. We calculated posting frequency using the duration between the first and last posts of the user then divided by the number of posts. The ground-truth population score is computed using the log-scaled number of views of each post.

B. IMPLEMENTATION DETAILS

We implemented our algorithm using the TensorFlow framework with an Intel Core i7-9750H CPU and an NVIDIA RTX-2070 GPU. We use Adam optimizer with $\beta_1$ set as 0.9 and $\beta_2$ as 0.999, The learning rate of 0.001, and 9 epochs of training.

C. EVALUATION METRICS

We evaluate our method using the metrics Spearman’s Rho ranking correlation (SRC) and Mean Absolute Error (MAE). SRC measures the correlation between the actual popularity set $P$ and the predicted popularity set $\hat{P}$. For $n$ test samples, SRC could be expressed as:

$$SRC = \frac{1}{n-1} \sum_{k=1}^{n} \frac{(P_k - \bar{P})(\hat{P}_k - \bar{\hat{P}})}{\sigma_P \sigma_{\hat{P}}}$$

where $\bar{P}$ and $\sigma_P$ are the mean and variance of their corresponding popularity set. For MAE, it is computed using Eq. 7.

For some experiments, we used a ranking system to show the overall performance of the model. Specifically, we independently ranked the performance in terms of SRC and MAE, where rank 1 has the best performance. Using the preliminary rank of both metrics, we get the average rank and use this value as a measure of the overall ranking.

D. ABLATION STUDY

We performed some experiments to test the components of our method. First, in Figure 4 we tested which backbone architecture for $G_\phi$ is the best. For a quick recall, $G_\phi$ is a network that receives sequential inputs. We tested the following architectures: LSTM, CNN, RoBERTa [11], and our proposed self-attention-based feature aggregation network.

We choose the long short-term memory (LSTM) as our baseline, which is a classical Recurrent Neural Network (RNN) architecture. In our experiment, the score of LSTM and CNN are almost the same because there are too many custom tags (roughly 250k unique words) and LSTM cannot afford the long-term dependency for a sequence. Furthermore, the category features are not sentences, which has high word correlations. We try to see if we can learn how to extract features by themselves using the attention mechanism, and the result shows that attention is the best way to solve giant categories problems.

In the next ablation studies, we tested the effectiveness of the different features we used on our method. We separated this into two experiments: one for the semantic features and the other for the numerical features.

In Table 1, Textural Features are not part of the columns since they are used by default. Based on the table, we could see that using almost all of the features provides a good performance boost to the overall model.

In Table 2, we tested whether the additionally computed numerical features are useful or not. Based on the results, under the benefit of “category data length”, we can get higher SRC scores since the features become more diverse, however,
H.-H. Lin et al.: Social Media Popularity Prediction Based on Multi-Modal Self-Attention Mechanisms

### TABLE 1. Ablation study on different features.

| User Info | Category Feature | Tags Feature | Image Feature | Misc. / Ensemble | SRC / MAE | Rank |
|-----------|------------------|--------------|---------------|------------------|-----------|------|
| X         | X                | √            | X             | X                | 0.46(5) / 1.87(5) | 4.5 |
| X         | X                | √            |              | X                | 0.47(5) / 2.01(5) | 4.5 |
| X         | √                | √            |              | X                | 0.52(3) / 1.75(3) | 3   |
| √         | √                | √            |              | X                | 0.57(2) / 1.73(2) | 2   |
| √         | √                | √            |               | √                | 0.58(4) / 1.61(4) | 1   |

Ablation studies of fusing different semantic features for our self-attention-based feature aggregation network. In the experiment, we found that we cannot reach good performance if we use the image feature directly. The reason may be the higher quality images cannot represent higher popularity.

### FIGURE 6. The sample information of each post. The samples have popularity score of 10.68, 11.22 and 3.81 from left to right, which are given by SMP dataset. And predicted score of these posts by our method are 8.44, 11.47 and 3.32, respectively. (The higher score, the more popular).

### TABLE 2. Ablation studies of Model performance using the additionally computed numerical features.

| Category Data Length | Post Frequency | SRC / MAE | Rank |
|----------------------|----------------|-----------|------|
| X                    | X              | 0.560(3) / 1.614(1) | 2    |
| √                    | √              | 0.568(3) / 1.729(3) | 2.5  |
| √                    | ×              | 0.569(4) / 1.714(1) | 1.5  |

We performed an ablation study to determine the importance of numerical features, namely, "Category Data Length" and "Post Frequency".

MAE also rise a little bit, but the ranking result shows that if we combine both of “category data length” and “post frequency”, we can lower the MAE a little than just using “category data length” feature.

### E. ENSEMBLE COMPARISON

In order to understand whether we can fuse the different attributes through ensemble two modality models effectively, we did an experiment with different combinations of the models. Figure 5 shows the training losses with three experiments, we collected the “tag” and “concept” features to train the model in experiment 1, which get the MAE of 0.48; experiment 2 adopted not only the features used in experiment 1 but also add the categorical features (“subcategory”, “category”, “path alias”, “uid”, “ispro”), increasing the performance by 27%; In experiment 3, we combined the features which adopted in last two experiments as model-1 and designed another model which combines numerical data and hand-crafted attributes, like “photo number”, “length of tag”, based on the result of ensemble two models, the MAE score is even made progress by 3% to 0.34. All training processes converge in 10 epochs, the comparison shows that in experiment 3 which ensemble can get the best score in terms of MAE, which also means that these data can complement each other based on ensemble mechanics.

### F. PERFORMANCE COMPARISON

We compare the performance of our proposed method with five state-of-the-art methods. The first method is a multi-modal method proposed in [22], the second method is an iterative refinement method proposed in [21], the third method is a word encoding method proposed in [19], the fourth method is the fifth place of SMP challenge 2019 [15], and the fifth method is the third place of SMP challenge 2019 [24]. However, the SMP datasets (SMPD) are a little bit different between 2019 and 2020 at the user information which is only available in SMP challenge 2020. Other than that, other parameters remain the same. For fair comparisons, the evaluation results as below did not use user information at all. As seen in Table 3, we have the best average ranking in terms of SRC and MAE.

### TABLE 3. Performance comparison with other SMP methods.

| Model                  | SRC / MAE         | Rank |
|------------------------|-------------------|------|
| multi-modal [22]       | 0.528(6) / 1.942(6) | 6    |
| Iterative Refinement [21] | 0.537(4.5) / 1.939(5) | 4.75 |
| Effective Word Encoding [19] | 0.548(4) / 1.938(5) | 4.5  |
| SMPCityU [15]          | 0.537(5) / 1.748(3) | 4    |
| CobotVisionLab [24]    | 0.651(1) / 1.760(4) | 2.5  |
| Proposed method (w/o user information) | 0.559(3) / 1.677(2) | 2.5  |
| Proposed method (w/ user information) | 0.58(2) / 1.61(1) | 1.5  |

Our model had the lowest MAE and relatively higher SRC than state-of-the-art methods.
G. PREDICTION EXAMPLES

In this subsection, we evaluated our social media popularity prediction model with some posts, the demonstrations are shown in Figure 6, the evaluated standard metric is the popularity score given from SMP-Challenge 2020 dataset. We compared the scores from the dataset with the popularity score predicted by the proposed method. Please note that Figure 6 is the example of given information in a post corresponding to the SMP dataset, however, the dataset does not provide the view number of each post.

V. CONCLUSION

In this paper, we proposed a social media popularity prediction method with multi-modal input and attention-based mechanisms. Specifically, our method uses semantic and numerical features to compute the popularity score. Semantic features are text-based and sequential hence attention-based networks (i.e. Transformer) have good synergy with this task. We also converted images to semantic features using existing image captioning algorithms. Furthermore, we augmented the existing numerical features to increase the performance of our model. We showcased that our method performs reasonably well against other state-of-the-art methods.

REFERENCES

[1] A. Khosla, A. D. Sarma, and R. Hamid, “What makes an image popular,” in Proc. Int. Conf. World Wide Web, 2014, pp. 867–876.
[2] B. Wu, W. H. Cheng, Y. Zhang, and T. Mei, “Time matters: Multi-scale temporalization of social media popularity,” in Proc. 24th ACM Int. Conf. Multimedia, Oct. 2016, pp. 1336–1344.
[3] R. van Zool, “Flickr: Who is looking?” in Proc. IEE/WIC/ACM Int. Conf. Web Intell. (WI), Nov. 2007, pp. 184–190.
[4] J. Hessel, L. Lee, and D. Mimno, “Cats and captions vs. Creators and the clock: Comparing multimodal content to context in predicting relative popularity,” in Proc. 26th Int. Conf. Multimedia, Apr. 2017, pp. 927–936.
[5] M. Mazlooom, R. Rietveld, S. Rudinac, M. Worrning, and W. van Dolen, “Multimodal popularity prediction of brand-related social media posts,” in Proc. 24th ACM Int. Conf. Multimedia, Oct. 2016, pp. 179–201.
[6] SMP Challenge Organization. (2020). Social Media Prediction Challenge. Available: http://smp-challenge.com
[7] C.-C. Hsu, L.-W. Kang, C.-Y. Lee, J.-Y. Lee, Z.-X. Zhang, and S.-M. Wu, “Popularity prediction of social media based on multi-modal feature mining,” in Proc. 27th ACM Int. Conf. Multimedia, Oct. 2019, pp. 2687–2691.
[8] F. Gelli, T. Uriochio, M. Bertini, A. Del Bimbo, and S.-F. Chang, “Image popularity prediction in social media using sentiment and context features,” in Proc. 23rd ACM Int. Conf. Multimedia, Oct. 2015, pp. 907–910.
[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, N. Adam Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2017, pp. 6000–6010.
[10] K. He, X. Zhang, 26th Conf. World Wide Web, Apr. 2017, pp. 927–936.
[11] M. Mazlooom, R. Rietveld, S. Rudinac, M. Worrning, and W. van Dolen, “Multimodal popularity prediction of brand-related social media posts,” in Proc. 24th ACM Int. Conf. Multimedia, Oct. 2016, pp. 179–201.
[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, N. Adam Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2017, pp. 6000–6010.
[13] M. He, X. Zhang, 26th Conf. World Wide Web, Apr. 2017, pp. 927–936.
[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “RoBERTa: A robustly optimized BERT pretraining approach,” 2019, arXiv:1907.11692.
[15] D. Kita, S. Bhosshan, H. Firooz, E. Perez, and D. Testuggine, “Supervised multimodal bitransformers for classifying images and text,” 2019, arXiv:1909.02950.
[16] 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,” in Proc. 27th Int. Joint Conf. Artif. Intell., Aug. 2017, pp. 3062–3068.
[17] L. Liu, F. Huang, Y. Li, Z. Yang, and W. Liu, “A layer-wise deep stacking model for social image popularity prediction,” in Proc. Int. Conf. World Wide Web, 2018, pp. 1639–1655.
[38] A. Ortis, G. M. Farinella, G. Torrisi, and S. Battiato, “Exploiting objective text description of images for visual sentiment analysis,” Multimedia Tools Appl., vol. 80, pp. 22323–22346, Jan. 2020.

[39] M. Katsurai and S. Satoh, “Image sentiment analysis using latent correlations among visual, textual, and sentiment views,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2016, pp. 2837–2841.

HUNG-HSIANG LIN received the B.S. degree in computer science and information engineering from the National Taiwan University of Science of Technology (NTUST), Taipei, Taiwan, in 2020, where he is currently pursuing the M.S. degree with the Department of Computer Science and Information Engineering. His research interests include multimedia networking and deep learning applied to computer vision.

JIUN-DA LIN received the B.S. degree in computer science and information engineering from the National Taiwan University of Science of Technology (NTUST), Taipei, Taiwan, in 2020, where he is currently pursuing the M.S. degree with the Department of Computer Science and Information Engineering. His research interests include image processing and deep learning applied to computer vision.

JOSE JAENA MARI OPLE received the B.S. degree in computer science from De La Salle University, Philippines, in 2018, and the M.S. degree from the National Taiwan University of Science and Technology (NTUST), in 2020, where he is currently pursuing the Ph.D. degree with the Department of Computer Science and Information Engineering. His research interests include digital image processing and deep learning applied to computer vision.

JUN-CHENG CHEN (Member, IEEE) received the B.S. and M.S. degrees in computer science and information engineering from the National Taiwan University, Taiwan, in 2004 and 2006, respectively, advised by Prof. Ja-Ling Wu, and the Ph.D. degree in computer science from the University of Maryland, College Park, MD, USA, in 2016, advised by Prof. Rama Chellappa. From 2017 to 2019, he was a Postdoctoral Research Fellow at the Institute for Advanced Computer Studies, University of Maryland. He is currently an Assistant Research Fellow at the Research Center for Information Technology Innovation, Academia Sinica. His research interests include computer vision, machine learning, deep learning, and their applications to biometrics, such as face recognition/facial analytics, and activity recognition/detection in the visual surveillance domain. He was a recipient of the ACM Multimedia Best Technical Full Paper Award in 2006.

KAI-LUNG HUA received the B.S. degree in electrical engineering from the National Tsing Hua University, Hsinchu, Taiwan, in 2000, the M.S. degree in communication engineering from the National Chiao Tung University, Hsinchu, in 2002, and the Ph.D. degree from the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA, in 2010. Since 2010, he has been with the National Taiwan University of Science and Technology, where he is currently a Professor with the Department of Computer Science and Information Engineering. Since 2021, he has also been the Dean of the Office of Industry-Academia Collaboration. He is also the Director of the Artificial Intelligence Research Center. His current research interests include digital image and video processing, computer vision, and machine learning. He is a member of Eta Kappa Nu and Phi Tau Phi, as well as a recipient of MediaTek Doctoral Fellowship. He has received several research awards, including the 2019 Excellent Research Award of Taiwan Tech, the 2019 Outstanding Research Award of Taiwan Tech, the 2018 Young Scholar Award of Taiwan Tech, the Top Performance Award of 2017 ACM Multimedia Grand Challenges, the Top 10% Paper Award of 2015 IEEE International Workshop on Multimedia Signal Processing, the Second Award of the 2014 ACM Multimedia Grand Challenge, the Best Paper Award of the 2013 IEEE International Symposium on Consumer Electronics, and the Best Poster Paper Award of the 2012 International Conference on 3D Systems and Applications.