Cross-Media Keyphrase Prediction: A Unified Framework with Multi-Modality Multi-Head Attention and Image Wordings

Yue Wang\textsuperscript{1}, Jing Li\textsuperscript{2}, Michael R. Lyu\textsuperscript{1}, and Irwin King\textsuperscript{1}
\textsuperscript{1}Department of Computer Science and Engineering
The Chinese University of Hong Kong, HKSAR, China
\textsuperscript{2}Department of Computing, The Hong Kong Polytechnic University, HKSAR, China
\textsuperscript{1}\{yuewang,lyu,king\}@cse.cuhk.edu.hk
\textsuperscript{2}jing-amelia.li@polyu.edu.hk

Abstract
Social media produces large amounts of contents every day. To help users quickly capture what they need, keyphrase prediction is receiving a growing attention. Nevertheless, most prior efforts focus on text modeling, largely ignoring the rich features embedded in the matching images. In this work, we explore the joint effects of texts and images in predicting the keyphrases for a multimedia post. To better align social media style texts and images, we propose: (1) a novel \textit{Multi-Modality Multi-Head Attention} (M\textsuperscript{3}H-Att) to capture the intricate cross-media interactions; (2) \textit{image wordings}, in forms of optical characters and image attributes, to bridge the two modalities. Moreover, we design a novel \textit{unified} framework to leverage the outputs of keyphrase classification and generation and couple their advantages. Extensive experiments on a large-scale dataset\textsuperscript{1} newly collected from Twitter show that our model significantly outperforms the previous state of the art based on traditional co-attention networks. Further analyses show that our multi-head attention is able to attend information from various aspects and boost classification or generation in diverse scenarios.

1 Introduction
The prominent use of social media platforms (such as Twitter) exposes individuals with an abundance of fresh information in a wide variety of forms such as texts, images, videos, etc. Meanwhile, the explosive growth of multimedia data has far outpaced individuals’ capability to understand them. This presents a concrete challenge to digest the massive amount of data, distill the salient contents therein, and provide users with a quicker access to the information they need when navigating a large multitude of noisy social media data.

\begin{itemize}
\item \textbf{Post (a):} Contemplating the mysteries of life from inside my egg carton...\textsuperscript{2} \\
\textsuperscript{2}We consider a hashtag as a post’s keyphrase annotation following the common practice (Zhang et al., 2016, 2018).
\end{itemize}

\begin{flushright}
\texttt{#cat #cats #CatsOfTwitter #NBAFinals}
\end{flushright}

\begin{itemize}
\item \textbf{Post (b):} The \texttt{<mention>} have the slight lead at halftime!
\end{itemize}

\begin{flushright}
\texttt{#2019 NBA FINALS}
\end{flushright}

\begin{figure}[ht]
\centering
\includegraphics[width=\textwidth]{t1.png}
\caption{Two multimedia posts from Twitter, where texts offer limited help in identifying their keyphrases while images provide essential clues.}
\end{figure}

To that end, extensive efforts have been made to \textbf{social media keyphrase prediction}\textsuperscript{2} — aiming to produce a sequence of words that reflect a post’s key concern. Nevertheless, previous work mostly focuses on the use of textual signals (Zhang et al., 2018; Wang et al., 2019a,b), which sometimes provide limited features as social media language is essentially informal and fragmented. To enrich the contexts, here we resort to exploiting the matching images, which are widely used in social media posts to deliver auxiliary information from authors (e.g., opinions, feelings, topics, etc.), primarily due to the flourish of mobile Internet.

To illustrate our motivation, Figure 1 shows the texts and images of two Twitter posts (tweets). The left is tagged with a keyphrase “cat”, which can be clearly signaled with its image while the paired text is an anthropomorphic description and hardly unveils its real semantics. For the right, the image depicts a basketball game scene with optical characters “2019 NBA FINALS”, directly indicating its keyphrase, which is difficult to identify from the...
texts. In both examples, images play a more vital role than texts in reflecting the key information. These points motivate our cross-media keyphrase prediction study that examines how the salient contents can be indicated by the coupled effects of post texts and their matching images.

Previous work (Zhang et al., 2017, 2019) employs co-attention networks (Lu et al., 2016; Xu and Saenko, 2016) to encode multimedia posts, where a single attention function is concurrently performed to infer either visual or textual distributions. We argue that they might be suboptimal to model intricate text-image associations, as a recent finding (Vempala and Preotiuc-Pietro, 2019) points out there can be four diverse semantic relations held by images and texts on Twitter. To allow for better modeling, in this work, we take advantage of the recent advances in multi-head attention (Vaswani et al., 2017), which is capable of learning representation from different subspaces. We extend it to capture diverse cross-media interactions, named as Multi-Modality Multi-Head Attention (M$^3$H-Att). Moreover, to well align the images’ semantics to texts’, we adopt image wordings and define two forms for that — explicit optical characters (such as “NBA Finals” in post (b)) detected from the optical character reader (OCR) and implicit image attributes (Wu et al., 2006), high-level text labels predicted to summarize the image’s semantic concepts (such as a “cat” label for post (a)).

Furthermore, unlike prior work employing either classification (Gong and Zhang, 2016) or generation models (Wang et al., 2019a), we propose a unified framework to couple the advantages of keyphrase classification and generation. Specifically, in addition to the joint training of both modules, we further extend the copy mechanism (See et al., 2017) to adaptive aggregate classification outputs together with source input tokens. Empirical results show that our proposed unified model not only keeps classification’s superiority to predict common keyphrases (Figure 5 (c)) while enabling keyphrase creation beyond a predefined candidate list, but also largely benefits the keyphrase prediction especially for absent keyphrases (Figure 5 (b)).

For experiments, we collect a large-scale tweet dataset with texts and images, which is presented as part of our work. Extensive results show that our model significantly outperforms the state-of-the-art (SOTA) methods using traditional attention mechanisms. For example, we obtain 47.06% F1@1 compared with 43.17% by Wang et al. (2019a) (keyphrase generation from texts only) and 42.12% by Zhang et al. (2017) (multi-modal keyphrase classification). We then examine how we perform to handle absent and present keyphrases, and varying keyphrase frequency and post length. The results indicate the consistent performance boost brought by our M$^3$H-Att design and unified framework in diverse scenarios (§5.3). We further quantify the effects of different settings of multi-head attention and image wordings to see when and how they work the best (§5.4). Lastly, we provide qualitative analysis to interpret why our model results in superior multimedia understanding (§5.5).

2 Related Work

Social Media Keyphrase Prediction. Traditional keyphrase prediction studies focus on using two-step pipeline methods: candidates are first extracted with handcrafted features (e.g. part-of-speech tags (Witten et al., 1999)) and then ranked by unsupervised (Wan and Xiao, 2008) or supervised algorithms (Medelyan et al., 2009). These methods undergo labor-intensive feature engineering and hence lead to the growing popularity of adopting data-driven neural networks. Specifically for social media keyphrase prediction, most efforts are based on sequence tagging style extraction (Zhang et al., 2016, 2018) or classification from a predefined candidate list (Gong and Zhang, 2016; Zhang et al., 2017), which cannot produce keyphrases absent in the post or the fixed list.

Inspired by the recent success of keyphrase generation (Meng et al., 2017; Chan et al., 2019) for scientific articles, Wang et al. (2019a,b) employ sequence-to-sequence (seq2seq) models to allow unseen keyphrases to be flexibly created for social media posts. Unlike them, we propose a novel unified framework to combine the benefits of keyphrase classification and generation. Similar to this, Chen et al. (2019) also exploit the power of classification for keyphrase generation but in a separated retrieval manner, where we elegantly integrate them with a tailored copy mechanism and allow for the end-to-end joint training. While most of prior work focuses on the modeling of texts, we additionally exploit the matching images and study their coupled effects for indicating keyphrases.

Cross-media Research. We are also related to cross-media research, where texts and images are jointly exploited for a variety of applications,
such as personalized image captioning (Park et al., 2019), event extraction (Li et al., 2020), sarcasm detection (Cai et al., 2019), and text-image relation classification (Vempala and Preotiuc-Pietro, 2019). Some of them have pointed out the usefulness of OCR texts (Chen et al., 2016) and image attributes (Wu et al., 2016) to endow images with higher-level semantics beyond visual features, where we are the first to study how OCR texts and image attributes work together to indicate keyphrases. Closest to our work, Zhang et al. (2017, 2019) study multimedia hashtag classification and employ co-attention networks (Lu et al., 2016; Xu et al., 2019) study multimedia hashtag classification and employ co-attention networks (Lu et al., 2016; Xu and Saenko, 2016) to model the text-image associations, while we extend the multi-head attention (Vaswani et al., 2017) to better capture such diverse styles of cross-media interactions.

While multi-head attention has been widely exploited in many vision-language (VL) tasks, such as image captioning (Zhou et al., 2020), visual question answering (Tan and Bansal, 2019; Lu et al., 2019), and visual dialog (Kang et al., 2019; Wang et al., 2020), its potential benefit to model flexible cross-media posts has been previously ignored. Due to the informal style in social media, cross-media keyphrase prediction brings unique difficulties mainly in two aspects: first, its text-image relationship is rather complicated (Vempala and Preotiuc-Pietro, 2019) while in conventional VL tasks the two modalities have most semantics shared; second, social media images usually exhibit a more diverse distribution and a much higher probability of containing OCR tokens (§4), thereby posing a hurdle for effectively processing.

3 Our Unified Cross-Media Keyphrase Prediction Framework

Given a collection $C$ with $|C|$ text-image post pairs $\{(x^n, I^n)\}_{n=1}^{|C|}$ as input, we aim to predict a keyphrase set $\hat{Y} = \{\hat{y}^i\}_{i=1}^{|Y|}$ for each of them. Following Meng et al. (2017), we copy the source input pair multiple times to allow each paired to have one keyphrase. We represent each input as a triplet $(x, I, y)$, where $x$ and $y$ are formulated as word sequences $x = \langle x_1, ..., x_{l_x} \rangle$ and $y = \langle y_1, ..., y_{l_y} \rangle$ ($l_x$ and $l_y$ denote the number of words).

We show the overview of our proposed cross-media keyphrase prediction model in Figure 2. We first encode a text-image tweet into three modalities: text, attribute, and vision (§3.1), and propose a Multi-Modality Multi-Head Attention (M$^3$H-Att) to capture their intricate interactions (§3.2). Then, we feed the learned multi-modality representations for either keyphrase classification or generation, followed with a tailored aggregator to combine their outputs (§3.3). Lastly, the entire framework can be jointly trained via multi-task learning (§3.4).

### 3.1 Multi-modality Encoder

**Learning Text Representation.** We first embed each token $x_i$ from the input sequence into a high-dimensional vector via a pre-trained lookup table, and then employ bidirectional gated recurrent unit (Bi-GRU) (Cho et al., 2014) to encode the embedded input token $e(x_i)$:

$$\hat{h}^{i} = GRU(e(x_i), \hat{h}_{i-1}), $$

(1)

$$\hat{h}^{l} = GRU(e(x_i), \hat{h}_{i+1}). $$

(2)

Forward hidden state $\hat{h}^{i}$ and backward one $\hat{h}^{l}$ are later concatenated into $h_i = [\hat{h}^{i}; \hat{h}^{l}]$. We employ it as the context-aware representation of $x_i$ and pack all of them in the input sequence into a textual memory bank $M_{text} = \{h_1, ..., h_{l_x}\} \in \mathbb{R}^{l_x \times d}$, where $d$ denotes the hidden state dimension.

**Encoding OCR Text.** To detect optical characters from images, we use an open-source toolkit (Smith, 2007) to extract OCR texts in form of a word sequence. It is then appended into the post text with a delimited token (sep) to notify the change of text genres, which is shown to be a simple yet effective design to combine OCR features.

**Learning Image Representation.** We consider two types of image representations: grid-level or object-level visual features. For the former, we

![Figure 2: The overview of our unified cross-media keyphrase prediction model.](image-url)
apply a pre-trained VGG-16 Net (Simonyan and Zisserman, 2015) to extract $7 \times 7$ convolutional feature maps for each image $I$. For the latter, inspired by bottom-up attention (Anderson et al., 2018), we use the Faster-RCNN (Ren et al., 2015) pre-trained on Visual Genome (Krishna et al., 2017) to detect the objects and extract their features. Each feature map is further transformed into a new vector $v_i$ through a linear projection layer. As such, we construct a visual memory bank as $M_{vis} = \{v_1, ..., v_{l_v}\} \in \mathbb{R}^{l_v \times d}$, where $l_v$ denotes the number of image regions or objects.

**Encoding Image Attribute.** Following Cai et al. (2019), we first train an attribute predictor based on the Resnet-152 (He et al., 2016) features on MS-COCO 2014 caption dataset (Lin et al., 2014). Specifically, we extract noun and adjective tokens from the image captions as the attribute labels. After that, the top five predicted attributes of each image are transformed with another linear layer to an attribute memory bank $M_{attr} = \{a_1, ..., a_5\} \in \mathbb{R}^{5 \times d}$, which aims to capture images’ high-level semantic concepts.

### 3.2 Multi-modality Multi-Head Attention

Our design of multi-head attention is inspired by its prototype in Transformer (Vaswani et al., 2017). We extend it to capture multiple forms of cross-modality interactions for a multimedia post, which is therefore named as $M^3$H-Att, short for Multi-Modality Multi-Head Attention. Compared to its prototype in Transformer, it is therefore named as $M^3$H-Att, instead operate on three modalities (text, attribute, and vision) in a pairwise co-attention manner.

For each co-attention, we perform scaled dot attention $\mathcal{A}$ on a set of $\{Query, Key, Value\}$:

$$\mathcal{A}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V,$$ (3)

$$\mathcal{A}^H(Q, K, V) = [head_1; ..., head_H]W^Q,$$ (4)

where $head_h = \mathcal{A}(QW^Q_h, KW^K_h, VW^V_h)$, (5)

where $W^Q_h, W^K_h, W^V_h \in \mathbb{R}^{d \times d_H}$ are learnable weights to project the query, key, and value from dimension $d$ to a lower space of $d_H$-dimension and $H$ is the head number. Outputs from all the heads are concatenated (in $\mathcal{A}^H$) and passed to a feedforward network with residual connections (He et al., 2016) and layer normalization (Ba et al., 2016).

Specifically, we employ the text features as a query to attend to the vision/attribute modality and vice versa. Here max/average-pooling is adopted to obtain one holistic query vector for each modality instead of token-level queries considering the noisy nature of social media data. Moreover, we stack multiple co-attention layers to empower its modeling capability, where $L_{text}, L_{attr}, L_{vis}$ denote the number of stacked layers for text, attribute, and vision queries, respectively. After that, the outputs from all co-attention layers are summed up with a linear multi-modal fusion layer to produce a context vector $c_{fuse} \in \mathbb{R}^d$. It will be fed into a keyphrase classifier and generator for the unified prediction. Notably, this indicates that our $M^3$H-Att’s great potential to serve as a generic module for benefiting other cross-media applications.

### 3.3 Unified Keyphrase Prediction

We describe how we combine the keyphrase classification and generation into a unified prediction for coupling their advantages below.

**Keyphrase Classification.** As each keyphrase $y$ usually consists of only several tokens, it can be considered as a discrete integral label and predicted it with a keyphrase classifier. Here we directly pass the multi-modal context vector $c_{fuse}$ into a two-layer of multi-layer perceptron (MLP) and map it to the distribution over the label vocabulary $V_{cls}$:

$$P_{cls}(y) = \text{softmax}(\text{MLP}_{cls}(c_{fuse})).$$ (6)

**Keyphrase Generation with Pointer.** For keyphrase generation, we base on a sequence-to-sequence framework to predict the keyphrase word sequence $y = \langle y_1, ..., y_{l}\rangle$, where the generation probability is defined as $\prod_{t=1}^{l} P(y_t | y < t)$. Concretely, we use an unidirectional GRU decoder to model the generation process, which emits the hidden state $s_t = \text{GRU}(s_{t-1}, u_t) \in \mathbb{R}^d$ based on the previous hidden state $s_{t-1}$ and the embedded decoder input $u_t$. The decoder state is initialized

---

**Figure 3:** Overview of $M^3$H-Att to fuse multi-modal features from text, attribute, and vision modalities.

---

1We also try other combinations, e.g., $M^3$H-Att between the vision and attribute, but the improvements are negligible.

---

**Figure 3:** Overview of $M^3$H-Att to fuse multi-modal features from text, attribute, and vision modalities.
by the last hidden state \( h_t \) of the text encoder. Here attention mechanism (Bahdanau et al., 2015) is adopted to obtain a textual context \( c_{text} \):

\[
c_{text} = \sum_{i=1}^{l_w} \alpha_{t,i} h_i,
\]

\[
\alpha_{t,i} = \text{softmax}(S(s_t, h_i)),
\]

\[
S(s_t, h_i) = v_\alpha^T \tanh(W_\alpha s_t + b_\alpha),
\]

where \( S(s_t, h_i) \) is a score function to measure the compatibility between the \( t \)-th word to be decoded and the \( i \)-th word from the text encoder. \( W_\alpha \in \mathbb{R}^{d \times 2d}, b_\alpha, v \in \mathbb{R}^d \) are trainable parameters.

Next, we incorporate the static multi-modal vector \( c_{fuse} \) (produced by M^3H-Att and independent of the decoding step \( t \)) to construct a context-rich representation \( c_t = [u_t; s_t; c_{text} + c_{fuse}] \). Based on it, we apply another MLP with softmax to produce a word distribution over vocabulary \( V_{gen} \):

\[
P_{gen}(y_t) = \text{softmax}(\text{MLP}_{gen}(c_t)).
\]

To further allow the decoder to explicitly extract words from the source post, we apply the copy mechanism (See et al., 2017) by calculating a soft switch \( \lambda_t \in [0, 1] \) with a sigmoid-activated MLP on \( c_t \). It indicates whether to generate the word from the vocabulary \( V_{gen} \) or copy it from the input sequence, where the extractive distribution is decided by the text attention weights \( \alpha_{t,i} \) in Eq. (8).

**Classification Output Aggregation.** We further extend the copy mechanism to aggregate the classification’s outputs to benefit keyphrase generation. First, we retrieve the top-K predictions from the classifier and convert each into the word sequence \( \mathbf{w} = \langle w_1, ..., w_{l_w} \rangle \), where \( l_w \) is the sequence length of the combined predictions. Then, we normalize their classification logits using softmax into a word-level distribution \( \beta \in \mathbb{R}^{l_w} \), which represents the extractive probability from the classification output. Finally, we obtain the unified prediction via:

\[
P_{unf}(y_t) = \lambda_t \cdot P_{gen}(y_t) + (1 - \lambda_t) \cdot \left( a \cdot \sum_{i : x_i = y_t} \alpha_{t,i} + b \cdot \sum_{j : y_j = y_t} \beta_j \right),
\]

where \( a, b \) (\( a + b = 1 \)) are hyper-parameters to decide whether to copy from the input sequence or the classification outputs. To stabilize the aggregation of classification outputs, we warm up the classifier for several epochs first by setting \( a \) to 1 and \( b \) to 0 and then both to 0.5 for further training.

### 3.4 Joint Training Objective

We employ the standard negative log-likelihood loss and define the entire framework’s training objective with the linear combination of the label classification loss and the token-level sequence generation loss for multitask learning:

\[
\mathcal{L}(\theta) = -\sum_{n=1}^{N} \left[ \log P_{cls}(y_n^a) + \gamma \cdot \sum_{t=1}^{l_y} \log P_{unf}(y_t^n) \right],
\]

where \( N \) is the size of the training text-image pairs and \( \gamma \) is a hyper-parameter to balance the two losses (empirically set to 1) and \( \theta \) denotes the trainable parameters shared for the whole framework. Intuitively, jointly training keyphrase classification would benefit the unified prediction by not only implicitly better parameter learning, but also explicitly providing more precise outputs to be copied to the keyphrase generator by the aggregation module.

### 4 Multi-modal Tweet Dataset

**Data Collection and Statistics.** Since there are no publicly available datasets for multi-modal keyphrase annotation, we contribute a new dataset with social media posts from Twitter. Specifically, we employ the Twitter advanced search API\(^4\) to query English tweets that contain both color image in JPG form; 2) we remove tweets with less than 4 tokens or more than 5 hashtags to filter out noise data; 3) rare hashtags (occurring less than 10 times) and their corresponding tweets are removed to alleviate sparsity issue; 4) we remove the duplicate tweets (e.g., retweets) and images and obtain 53,701 tweets with each containing a distinct tweet text-image pair.

**Preprocessing.** We employ an open-source Twitter preprocessing tool (Baziotis et al., 2017) to tokenize the tweets, segment the hashtags, and apply common spelling corrections. To reduce the errors introduced by the automatic hashtag segmentation, we manually check them and construct a complete

\[^4\text{https://twitter.com/search-advanced}\]
Table 1: Data split statistics. KP: keyphrase; |KP|: the size of unique keyphrase; % of occ. KP: percentage of keyphrases occurring in the source post.

| Split | #Post | Post Len | |KP| /Post | |KP| | Len | % of occ. KP | Vocab |
|-------|-------|----------|---|-----|---|-----|-----|--------|--------|
| Train | 42,959 | 27.26 | 1.33 | 4,261 | 1.85 | 37.14 | 48,019 |
| Val   | 5,370  | 26.81 | 1.34 | 2,544 | 1.85 | 36.01 | 16,892 |
| Test  | 5,372  | 27.05 | 1.32 | 2,534 | 1.86 | 37.45 | 17,021 |

Figure 4: Image type distribution of 200 sampled text-image tweets in our collected dataset.

Tweet Image Analysis. To further analyze the Twitter image characteristics, we sample 200 text-image tweets and analyze their distributions over varying types in Figure 4. We observe a diverse set of categories and only around half of the images (54%) are natural photos, which is rather different from other standard image data such as MS-COCO. Moreover, we conduct a pilot study to categorize the text-image relations following Vempala and Preotiuc-Pietro (2019) and find 52% of them have either texts or images useless to represent semantics (see Figure 9 for some examples). Such diverse category and complex text-image relationship pose unique challenges compared to traditional vision-language tasks like image captioning and visual question answering, where they focus on more natural images, and more importantly, their two modalities have most semantics shared. To deal with this, we propose M³H-Att and image wordings to better capture essential information from noisy cross-media data.

Image Wording Analysis. Here we shed light on some interesting statistics on image wordings. We first analyze the top 5 attributes predicted from the images in our dataset: \{man, shirt, woman, sign, white\}, which shows that most of the images on Twitter are about people and daily life. For OCR texts, we employ a widely used OCR engine Tesseract\(^5\) to extract optical characters. From all matching images, there are around 35% of them contain characters, significantly larger than the corresponding number in COCO images (4%), indicating social media users’ preference to post images containing optical characters. To mitigate the effects of OCR errors, we only consider tokens present in the vocabulary of tweet texts and find about 17% left with a median length of 16 tokens. Besides, 32% of the remaining data have words appearing in their corresponding keyphrases and 13% contain the entire keyphrases, suggesting its potential help in keyphrase prediction.

5 Experiments and Analyses

5.1 Experimental Setup

Evaluation Metrics. We mainly evaluate our model with popular information retrieval metrics macro-average F1@K, where K is 1 or 3 as there are 1.33 keyphrases on average per tweet (Table 1). To further measure the keyphrase orders (as we can generate a keyphrase ranking list with beam search), we employ mean average precision (MAP) for the top five predictions following Chen et al. (2019). The higher scores from all the metrics indicate better performance. For word matchings in evaluation, we consider the results after processed with Porter Stemmer following Meng et al. (2017).

Comparison Models. We first consider the upper-bound performance of extractive methods, denoted as EXT-ORACLE. Then, the following baselines are compared. (1) Image-only models: we apply max/average pooling on the grid-level VGG features or object-level BUTD (Anderson et al., 2018) and aggregate them for classification. (2) Text-only models: we consider classification-based (CLS) or sequence generation-based (GEN) methods. For CLS models, we consider simple max/average pooling on the text features learned from Bi-GRU encoder and the Topic Memory Network (TMN) (Zeng et al., 2018) (a SOTA short text classification model). For GEN models, we employ the seq2seq with attention (Bahdanau et al., 2015), copy mechanism (See et al., 2017), and latent topics (Wang et al., 2019a) (the SOTA topic-
We first report the main comparison results in Table 2, and we employ beam search with beam size set to 10 with a learning rate as 0.01. For our models, we first adopt the basic variants with $M^3$H-Att separately applying to either CLS or GEN. Then we additionally combine image wordings and the joint training strategy (Eq. (12)). Our full model is obtained by further aggregating the CLS and GEN outputs (Eq. (11)).

**Parameter Settings.** We maintain a generation vocabulary $V_{gen}$ of 45K tokens and the keyphrase classification vocabulary $V_{cls}$ with 4,262 labels. We apply 200-d Twitter GloVe embedding (Pennington et al., 2014) for encoding inputs. We employ two layers of Bi-GRU for the encoder and a single layer GRU for the decoder with hidden size set to 300. For visual signals, we extract either 49 grid-level VGG 512-d features or 36 object-level BUTD 2048-d features. For the $M^3$H-Att, we employ 4 heads with 64-d subspace, where 4 layers are stacked for attention to text modality, and 1 layer for vision or attribute modality. In training, we set the loss coefficient $\gamma = 1$ and employ Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.001. We decay it by 0.5 if validation loss does not drop and apply gradient clipping with the max gradient norm as 5. Early stop (Caruana et al., 2000) is adopted via monitoring the change of validation loss. For inference, we employ beam search with beam size set to 10 to generate a ranking list of keyphrases. For the baselines, we re-implement CLS-IMG-ATT and CLS-CO-ATT, and employ the released codes to produce results for CLS-TMN⁶, GEN-TOPIC⁷, and CLS-BAN⁸.

### 5.2 Main Comparison Results

We first report the main comparison results in Table 2 and draw the following observations:

- **Textual features are more important than visual signals.** It is seen from the text-only models’ better performance compared with their counterparts relying solely on images. For image-only models, we find that object-level BUTD outperforms grid-level VGG, while for pooling methods, average pooling works better for visual signals while max pooling is more suitable for texts.

- **Vision modality can provide complementary information to the text.** Most models considering cross-media signals perform better than text-only and image-only baselines. An exception is observed on the classification models with traditional attention, where the best F1@1 score 42.12 from CLS-CO-ATT is still less than the text-only model GEN-TOPIC’s 43.17. This indicates the limitation of traditional co-attention to well exploit multi-modal features from social media.

- **Both $M^3$H-Att and image wordings are helpful to encode social media features.** We find that both $M^3$H-Att and image wordings contribute to the performance boost of keyphrase classification or generation or their joint training results, which showcase their ability to handle multi-modality data from social media. We will discuss more in §5.4.

- **Our output aggregation strategy is effective.** Seq2seq-based keyphrase generation models (especially armed with the copy mechanism to enable better extraction capability) perform better than

---

⁶https://github.com/zengjichuan/TMN
⁷https://github.com/yuewang-cuhk/TAKG
⁸https://github.com/jnhwkim/ban-vqa

| Models | F1@1 | F1@3 | MAP@5 |
|--------|------|------|-------|
| EXT-ORACLE | 39.30 | 23.20 | 39.26 |
| **Image-only** | | | |
| CLS-VGG-MAX | 14.20 | 12.02 | 17.68 |
| CLS-VGG-AVG | 15.69 | 13.67 | 19.70 |
| CLS-BUTD-MAX | 17.65 | 15.00 | 21.77 |
| CLS-BUTD-AVG | 20.02 | 16.97 | 24.73 |
| **Text-only** | | | |
| CLS-AVG | 35.96 | 27.59 | 41.84 |
| CLS-MAX | 38.33 | 28.84 | 44.15 |
| CLS-TMN | 40.33 | 30.07 | 46.82 |
| **CLS-ATT** | | | |
| GEN-ATT | 45.69 | 37.90 | 55.85 |
| GEN-COPY | 42.10 | 33.91 | 50.35 |
| GEN-TOPIC | 43.17 | 30.73 | 48.07 |
| **Text-image** | | | |
| CLS-BAN | 38.73 | 29.68 | 45.03 |
| CLS-IMG-ATT | 41.48 | 31.22 | 47.93 |
| CLS-CO-ATT | 42.12 | 31.55 | 48.39 |
| **CLS-M$^3$H-ATT (ours)** | | | |
| + image wording | 44.46 | 32.82 | 50.39 |
| + joint-train | 45.16 | 33.17 | 51.48 |
| **CLS-M$^3$H-ATT (ours)** | | | |
| + image wording | 44.56 | 32.72 | 49.95 |
| + joint-train | 45.69 | 32.78 | 51.37 |
| **GEN-CLS-M$^3$H-ATT (ours)** | | | |
| + image wording | 47.06 | 33.10 | 52.07 |

Table 2: Comparison results (in %) displayed with average scores from 5 random seeds. Our GEN-CLS-$M^3$H-ATT significantly outperforms all the comparison models (paired t-test $p < 0.05$). Subscripts denote the standard deviations (e.g., 47.06±0.04).
most classification models and even upper bound results of extraction models. It is probably because of the high absent keyphrase rate and the large size of keyphrases (Table 1) exhibited in the noisy social media data. Nevertheless, GEN-CLS-M^3H-ATT, coupling advantages of classification and generation, obtains the best results (47.06 F1@1), dramatically outperforms the SOTA text-only model (43.17) and text-image one (42.12).

5.3 Quantitative Analyses

We examine how our models perform in diverse scenarios: present vs. absent keyphrases and varying keyphrase frequency and post length in Figure 5.

Present vs. Absent Keyphrases. We report the F1@1 for evaluating present keyphrases and recall@5 for absent keyphrases. As shown in Figure 5 (a-b), generation models with copy mechanism consistently outperform classification models for present keyphrases, while the latter works better for absent keyphrases. Nonetheless, our output aggregation strategy is able to cover generation models’ inferiority for absent keyphrases and exhibits much better results (41.19 vs. 35.83 recall@5 score) from GEN-CLS-M^3H-ATT than GEN-M^3H-ATT. Besides, visual signals are helpful for generation and classification models in predicting both present and absent keyphrases, where a larger boost is observed for the latter, probably owing to the inadequate clues available from texts.

Keyphrase Frequency. From Figure 5 (c), we observe better F1@1 from all models to produce more frequent keyphrases, because common keyphrases allow better representation learning from more training instances. For extremely rare keyphrases (occur < 10 times in training), generation models with copy mechanisms exhibit better capability to handle them than classification ones.

Post Length. From Figure 5 (d), we observe that longer post length does not guarantee better performance and the best results are obtained for posts with 15 ~ 35 tokens. It might be attributed to the noisy nature of social media data — longer posts provide both richer contents and more noise. For the posts with < 15 tokens, all multi-modal methods perform better than the text-only ones, indicating that the image modality plays a more important role when texts contain limited features.

5.4 Analyses of M^3H-Att and Image Wording

We proceed to quantify the effects of different settings in M^3H-Att and image wording.

M^3H-Att Analysis. We investigate how various configurations \(L_{vis} \in \{1, 2, 3, 4\}, H \in \{2, 4, 8, 12\}, d_H \in \{64, 128, 256\}\) of our M^3H-Att affect the prediction results in Table 3. Here we only show the classification results (and similar trends are observed from generation). We notice that more complex models do not always present better results and even render performance deterioration in some cases due to the overfitting issue. The best performance is attained by 4 stacked layers of 4 heads with a 64-d subspace.

Image Wording Analysis. To examine image wording effects, we compare four models in three settings: no image wording, OCR (only), and image attributes (only) in Table 4. The results are shown in three test sets: the entire test set (Full), the 889 subset instances with OCR tokens (OCR), and the 266 ones containing keyphrases from ImageNet labels (Attr) (Russakovsky et al., 2015). For the CLS-MAX and GEN-COPY, we add at-
tributes by using its max-pooled features to attend the text memory, which is later used for prediction.

We observe that either OCR texts or image attributes contribute to better F1@1 on the entire test set for all chosen models, while much more performance gain can be observed on their subsets with OCR texts or ImageNet keyphrases, indicating that images with optical characters and natural styles can benefit more from image wordings.\(^\text{10}\)

5.5 Qualitative Analysis

To explore whether M\(^3\)H-Att is able to attend different aspects from the image, we probe into its attention weights via heatmap visualization in Figure 6. Here CLS-M\(^3\)H-ATT is employed with a single layer of 12 heads, whose image-to-text and text-to-image attention are examined. The top figure shows that all its heads attend to the text based on the visual cues, where some attend to “turtle” while others attend to “world” and “globe” with various emphasis. Interestingly, Head 11 highlights the “happy” token, which also appears in the image. For the text-to-image attentions (bottom), we find some heads tend to highlight the specific local objects, such as the two players by Head 0 and 5 and the textual regions by Head 9, while some capture a more global view of the image like Head 11. More examples are shown in Figure 8.

We further illustrate how images (visual signals, image attributes, and OCR tokens) help cross-media keyphrase prediction by analyzing their predictions in Figure 7. In post (a), visual features help both CLS-CO-ATT and our model correctly predict its keyphrase, where our model precisely attends the cat’s face (key region reflecting the image’s semantics). Without such context, GEN-COPY wrongly predicts “star wars”, which might be caused by the misleading token “mysterious” in the texts. Besides, the cat keyphrase is also revealed in the top predicted attribute. In post (b-c), only our model with image wordings makes correct predictions, where we observe that the ground-truth keyphrases directly appear in the attributes or OCR texts. See Figure 10 and 11 for more examples.

6 Conclusion

This paper studies cross-media keyphrase prediction on social media and presents a unified framework to couple the advantages of generation and classification models for this task. Moreover, we propose a novel Multi-Modality Multi-Head Attention to capture the dense interactions between texts and images, where image wordings explicit in optical characters and implicit in image attributes are further exploited to bridge their semantic gap. Experimental results on a large-scale newly-collected Twitter corpus show that our model significantly outperforms SOTA generation and classification models with traditional attention mechanisms.

Acknowledgements

We thank Hou Pong Chan, Jian Li, and our anonymous reviewers for their insightful feedback on our paper. This work was partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (No. CUHK 14210717, General Research Fund; No. CUHK 2410021, Research Impact Fund, R5034-18). Jing Li is supported by the Hong Kong Polytechnic University internal funds (1-BE2W and 1-ZVRH) and NSFC Young Scientists Fund 62006203.
Post (e): So excited to hear her new song never really over every hour all day

Figure 8: More attention weight visualizations for both image-to-text attention and text-to-image attention.
**Post (a):** Sharing is caring. Good girl Kit, cause I know how much you love your bed. #Dogs #Kindness

**Post (b):** Waves crash against the North Pier this evening at Tynemouth, River Tyne in the UK @david1hirst #StormHour

**Post (c):** “I am declaring an emergency that only I can fix” #BoycottTrumpPrimeTime

**Post (d):** The whole of the UK when armadillo and danny say anything #LoveIsland

---

**Figure 9:** Example tweets of four different types of text-image relationship in our dataset. Post (a): text is represented and image adds to. Post (b): text is represented and image does not add to. Post (c): text is not represented and image adds to. Post (d): text is not represented and image does not add to.

---

**Post (a):** I thought Older Hanzo died after D’Vorah killed him? @NetherRealm #MortalKombat11

**Post (b):** Congrats producer of the year, non-classical winner - Williams #Grammys

**Post (c):** Last year’s highest rated animated movie spider-man into the Spider-Verse is now streaming on Netflix! #SpiderMan

**Post (d):** We need to make sure the ratings are high #SaveShadowhunters

---

**Figure 10:** More qualitative examples showing the effectiveness of encoding OCR texts. Among various models, only our model that considers OCR tokens correctly predicts the keyphrases for all these cases (in bold). Purple tokens are the detected OCR tokens, where we observe that the keyphrases directly appear in them.

---

**Post (a):** Good night, everyone. I hope that you have had a delightful day and a restful weekend. #hoorayfordogs

**Post (b):** Head up, chest out! A handsome purple finch poses for a shot. #birds #wildlife #photography

**Post (c):** I was watching all the bees Honeybee collecting pollen on the flowers Bouquet #SaveTheBees #CatsOfTwitter

**Post (d):** For 1970, Plymouth intended to make its GTX model a street powerhouse. #MuscleCar #ClassicCar

---

**Figure 11:** More qualitative examples showing the effectiveness of encoding image attributes. Our model that considers image attributes correctly predicts the keyphrases for all these cases (in bold). Blue tokens are the top five predicted attributes, which reveal the main image contents and thus help to indicate keyphrases.