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

Yue Wang[1], **Jing Li**[2], **Michael R. Lyu**[1], **and Irwin King**[1]
[1]Department of Computer Science and Engineering
The Chinese University of Hong Kong, HKSAR, China
[2]Department of Computing, The Hong Kong Polytechnic University, HKSAR, China
[1]{yuewang,lyu,king}@cse.cuhk.edu.hk
[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 *Multi-Modality Multi-Head Attention* (M^{3}H-Att) to capture the intricate cross-media interactions; (2) *image wordings*, in forms of optical characters and image attributes, to bridge the two modalities. Moreover, we design a novel *unified* framework to leverage the outputs of keyphrase classification and generation and couple their advantages. Extensive experiments on a large-scale dataset[1] newly collected from Twitter show that our model significantly outperforms the previous state of the art based on traditional co-attentions. 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, presenting a concrete challenge to digest the massive amount of data, distill the salient contents therein, and provide users with a quick access to the information they need when navigating noisy online data.

Post (a): Contemplating the mysteries of life from inside my egg carton...☺
#cat #cats #CatsOfTwitter

Post (b): The <*mention*> have the slight lead at halftime!
#NBAFinals

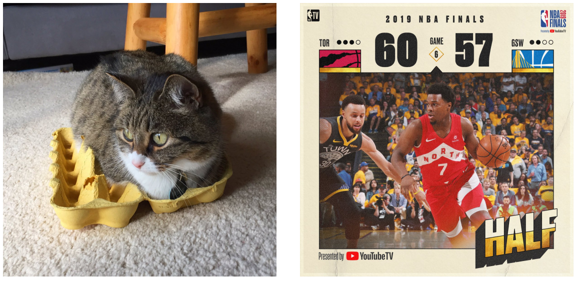


Figure 1: Two multimedia posts from Twitter, where texts offer limited help in identifying their keyphrases while images provide essential clues.

To that end, extensive efforts have been made to **social media keyphrase prediction**[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

[1] Our code and dataset are released at https://github.com/yuewang-cuhk/CMKP.

[2] We consider a hashtag as a post's keyphrase annotation following the common practice (Zhang et al., 2016, 2018).

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 3311–3324,
November 16–20, 2020. ©2020 Association for Computational Linguistics

the texts. In both examples, images play a more vital role 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 advance of multi-head attention (Vaswani et al., 2017) capable of learning from different representation subspaces and extend it to capture diverse cross-media interactions, named as *Multi-Modality Multi-Head Attention* (M^3H-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 explicitly aggregate classification outputs together with tokens from the source input. Empirical results show that integrating classification outputs not only keeps classification's superiority to predict common keyphrases (Figure 5(c)) while enables keyphrase creation beyond a predefined candidate list, but also largely benefits the keyphrase generation with better absent keyphrase prediction (Figure 5(b)).

For experiments, we collect a large-scale tweet dataset with texts and images, which is presented as part of our work. The empirical results show that our model significantly outperforms the state-of-the-art (SOTA) methods using traditional attention. 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^3H-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 are however unable to produce keyphrases absent in the post or the fixed list. Inspired by the recent success of keyphrase generation for scientific articles (Meng et al., 2017; Chan et al., 2019), 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 exploits 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 their matching images and study the 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 (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 diverse styles of cross-modal interactions in social media.

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 $\mathcal{C}$ with $|C|$ text-image post pairs $\{(\mathbf{x}^n, I^n)\}_{n=1}^{|C|}$ as input, we aim to predict a keyphrase set $\mathcal{Y} = \{\mathbf{y}^i\}_{i=1}^{|\mathcal{Y}|}$ 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 $(\mathbf{x}, I, \mathbf{y})$, where $\mathbf{x}$ and $\mathbf{y}$ are formulated as word sequences $\mathbf{x} = \langle x_1, ..., x_{l_x} \rangle$ and $\mathbf{y} = \langle y_1, ..., y_{l_y} \rangle$ (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).

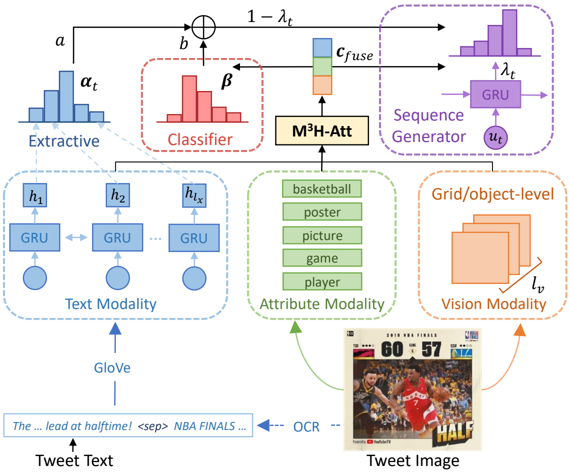


Figure 2: The overview of our unified cross-media keyphrase prediction model.

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 $e(x_i)$:

$$\overrightarrow{\mathbf{h}_i} = GRU(e(x_i), \ \overrightarrow{\mathbf{h}_{i-1}}), \quad (1)$$

$$\overleftarrow{\mathbf{h}_i} = GRU(e(x_i), \ \overleftarrow{\mathbf{h}_{i+1}}). \quad (2)$$

Forward hidden state $\overrightarrow{\mathbf{h}_i}$ and backward one $\overleftarrow{\mathbf{h}_i}$ are later concatenated into $\mathbf{h}_i = [\overrightarrow{\mathbf{h}_i}; \overleftarrow{\mathbf{h}_i}]$. 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 $\mathbf{M}_{text} = \{\mathbf{h}_i, ..., \mathbf{h}_{l_x}\} \in \mathbb{R}^{l_x \times d}$, 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 $\langle sep \rangle$ 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 apply a pre-trained VGG-16 Net (Simonyan and Zisserman, 2015) to extract 7×7 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 $\mathbf{v}_i$ through a linear projection layer. As such, we construct a visual memory bank as $\mathbf{M}_{vis} = \{\mathbf{v}_1, ..., \mathbf{v}_{l_v}\} \in \mathbb{R}^{l_v \times d}$, 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. Afterward, the top five predicted attributes of each image are transformed with another linear layer to an attribute memory bank $\mathbf{M}_{attr} = \{\mathbf{a}_1, ..., \mathbf{a}_5\} \in \mathbb{R}^{5 \times d}$, 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 original use as a self-attention over texts only, we 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 $\{Query, Key, Value\}$:

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

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

$$\text{where} \quad head_h = \mathcal{A}(QW_h^Q, KW_h^K, VW_h^V), \quad (5)$$

where $W_h^Q, W_h^K, W_h^V \in \mathbb{R}^{d \times d_H}$ 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 $\mathcal{A}^M$) and passed to a feedforward network with residual connections (He et al., 2016) and layer normalization (Ba et al., 2016).

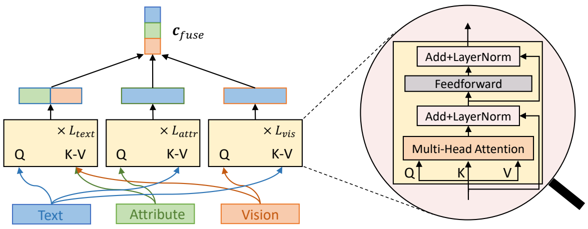


Figure 3: Overview of M^{3}H-Att to fuse multi-modal features from text, attribute, and vision modalities.

Specifically, we employ the text features as a query to attend to the vision/attribute modality and vice versa.[3] 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 $\mathbf{c}_{fuse} \in \mathbb{R}^d$. 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 $\mathbf{y}$ 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 $\mathbf{c}_{fuse}$ into a two-layer of multi-layer perceptron (MLP) and map it to the distribution over the label vocabulary V_{cls}:

$$P_{cls}(\mathbf{y}) = \mathtt{softmax}(\mathtt{MLP}_{cls}(\mathbf{c}_{fuse})). \quad (6)$$

Keyphrase Generation with Pointer. For keyphrase generation, we base on a sequence-to-sequence framework to predict the keyphrase word sequence $\mathbf{y} = \langle y_1, ..., y_{l_y} \rangle$, where the generation probability is defined as $\prod_{t=1}^{l_y} P(y_t \,|\, \mathbf{y}_{<t})$.

Concretely, we use an unidirectional GRU decoder to model the generation process, which emits the hidden state $\mathbf{s}_t = GRU(\mathbf{s}_{t-1}, \mathbf{u}_t) \in \mathbb{R}^d$ based

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

on the previous hidden state $\mathbf{s}_{t-1}$ and the embedded decoder input $\mathbf{u}_t$. The decoder state is initialized by the last hidden state $\mathbf{h}_{l_x}$ of the text encoder. Here an attention mechanism (Bahdanau et al., 2015) is adopted to obtain a textual context $\mathbf{c}_{text}$:

$$\mathbf{c}_{text} = \sum_{i=1}^{l_x} \alpha_{t,i} \mathbf{h}_i, \tag{7}$$

$$\alpha_{t,i} = \texttt{softmax}(S(\mathbf{s}_t, \mathbf{h}_i)), \tag{8}$$

$$S(\mathbf{s}_t, \mathbf{h}_i) = \mathbf{v}_\alpha^T \tanh(\mathbf{W}_\alpha[\mathbf{s}_t; \mathbf{h}_i] + \mathbf{b}_\alpha), \tag{9}$$

where $S(\mathbf{s}_t, \mathbf{h}_i)$ 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. $\mathbf{W}_\alpha \in \mathbb{R}^{d\times 2d}$, $\mathbf{b}_\alpha, \mathbf{v} \in \mathbb{R}^d$ are trainable weights.

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

$$P_{gen}(y_t) = \texttt{softmax}(\texttt{MLP}_{gen}(\mathbf{c}_t)). \tag{10}$$

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 $\mathbf{c}_t$. 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:

$$P_{unf}(y_t) = \lambda_t \cdot P_{gen}(y_t) \; + \tag{11}$$
$$(1 - \lambda_t) \cdot (a \cdot \sum_{i:x_i=y_t}^{l_x} \alpha_{t,i} + b \cdot \sum_{j:w_j=y_t}^{l_w} \beta_j),$$

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:

$$\mathcal{L}(\theta) = -\sum_{n=1}^{N} [\underbrace{\log P_{cls}(\mathbf{y}^n)}_{\text{Classification}} + \gamma \cdot \sum_{t=1}^{l_y^n} \underbrace{\log P_{unf}(y_t^n)}_{\text{Unified}}], \tag{12}$$

where N is size of the training text-image pairs and γ is a hyper-parameter to balance the two losses (empirically set to 1) and θ 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 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 images and hashtags from January to June 2019. For keyphrases, we consider to use user-generated hashtags following common practice (Zhang et al., 2016, 2018). We further clean the raw data in the following ways: 1) we only retain tweets with one 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. We randomly split the data into 80%, 10%, 10% corresponding to training, validation, and test set. The data split statistics of tweet texts are displayed in Table 1.

[4] https://twitter.com/search-advanced

Split	#Post	Post Len	#KP /Post	|KP|	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

Table 1: Data split statistics. KP: keyphrase; |KP|: the size of unique keyphrase; % of occ. KP: percentage of keyphrases occurring in the source post.

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 mapping list. Following Wang et al. (2019a), we retain tokens in hashtags (without # prefix) for those occurring in the middle of the posts due to their inseparable semantic roles. We further remove all the non-alphabetic tokens and replace links, mentions (@username), digits into special tokens as ⟨*url*⟩, ⟨*mention*⟩, and ⟨*number*⟩ respectively.

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 in the Appendix). 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 M^{3}H-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 Tesserocr[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.

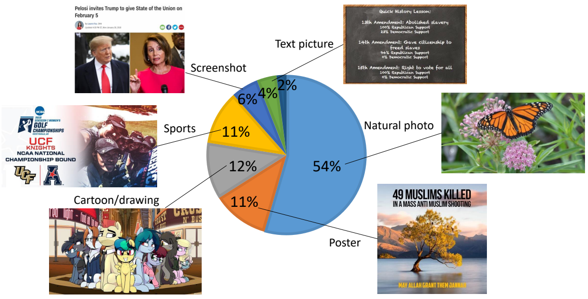


Figure 4: Image type distribution of 200 sampled text-image tweets in our collected dataset.

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)

[5] https://pypi.org/project/tesserocr/

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-aware model for social media keyphrase generation). (3) **Text-image** models: we consider the SOTA CLS model for multi-modal hashtag recommendation (Zhang et al., 2017) using co-attention and its variant with image-attention (Yang et al., 2016), as well as Bilinear Attention Networks (BAN) (Kim et al., 2018) (a SOTA variant for Visual Question Answering (Antol et al., 2015)). 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 as 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-TMN[6], GEN-TOPIC[7], and CLS-BAN[8].

	Models	F1@1	F1@3	MAP@5
	EXT-ORACLE	39.50	23.20	39.26
Image-only	CLS-VGG-MAX	14.20_{35}	12.20_{24}	17.68_{31}
	CLS-VGG-AVG	15.69_{21}	13.67_{06}	19.70_{20}
	CLS-BUTD-MAX	17.65_{32}	15.00_{21}	21.77_{29}
	CLS-BUTD-AVG	20.02_{27}	16.97_{06}	24.73_{11}
Text-only	CLS-AVG	35.96_{11}	27.59_{05}	41.84_{14}
	CLS-MAX	38.33_{47}	28.84_{09}	44.15_{34}
	CLS-TMN	40.33_{39}	30.07_{28}	46.28_{27}
	GEN-ATT	38.36_{28}	27.83_{15}	43.35_{20}
	GEN-COPY	42.10_{19}	29.91_{30}	46.94_{35}
	GEN-TOPIC	43.17_{24}	30.73_{13}	48.07_{23}
Text-Image	CLS-BAN	38.73_{18}	29.68_{23}	45.03_{15}
	CLS-IMG-ATT	41.48_{33}	31.22_{14}	47.93_{34}
	CLS-CO-ATT	42.12_{38}	31.55_{33}	48.39_{34}
	CLS-M^{3}H-ATT (ours)	44.11_{17}	31.47_{14}	49.45_{11}
	+ image wording	44.46_{12}	32.82_{24}	50.39_{15}
	+ joint-train	45.16_{09}	33.27_{10}	51.48_{11}
	GEN-M^{3}H-ATT (ours)	44.25_{05}	31.58_{13}	49.35_{10}
	+ image wording	44.56_{09}	31.77_{23}	49.95_{22}
	+ joint-train	45.69_{17}	32.78_{09}	51.37_{12}
	GEN-CLS-M^{3}H-ATT (ours)	**47.06**$_{04}$	**33.11**$_{01}$	**52.07**$_{03}$

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 deviation (e.g., $47.06_{04} \Rightarrow 47.06 \pm 0.04$).

5.2 Main Comparison Results

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

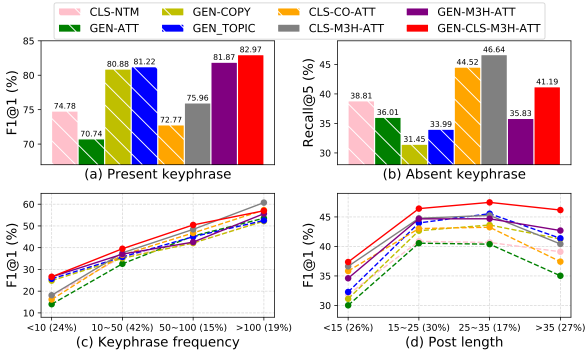
• *Textual features are more important than visual signals.* It is seen from the text-only models' better performance compared with their counterparts relying solely on images. For image-only models, we find that object-level BUTD outperforms grid-level VGG, while for pooling methods, average pooling works better for visual signals while max pooling is more suitable for texts.[9]

• *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 CLS-CO-ATT, which indicates the limitation of traditional co-attention to well exploit multi-modality representations 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.

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

[9] In experiments, we find that VGG works better than BUTD features for M^{3}H-Att in our variants. We show results with the better setting without otherwise specified.



(a) Present keyphrase

(b) Absent keyphrase

(c) Keyphrase frequency

(d) Post length

Figure 5: Model comparison over: (a) present keyphrases, (b) absent keyphrases, (c) varying keyphrase frequency, and (d) varying post length. Striped bars or dashed lines denote previous models while solid ones denote ours.

• *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 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 keyphrase tags (Table 1) exhibited in the noisy social media data. Nevertheless, GEN-CLS-M^{3}H-ATT, coupling advantages of classification and generation, obtains the best results (47.06 F1@1), drastically 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. 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 better results from GEN-CLS-M^{3}H-ATT than GEN-M^{3}H-ATT. Besides, visual signals are helpful to both generation and classification to yield either present or 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.

# Layer	2 Head			4 Head			8 Head			12 Head		
	64-d	128-d	256-d	64-d	128-d	256-d	64-d	128-d	256-d	64-d	128-d	256-d
1	42.06	43.32	43.01	43.11	43.98	43.63	43.75	44.18	43.43	43.48	43.81	43.53
2	43.22	44.36	44.26	44.27	44.38	44.27	44.58	44.59	43.12	45.05	38.16	39.97
3	43.51	44.23	43.62	44.50	44.25	43.00	44.70	43.27	36.05	44.49	35.70	31.35
4	44.38	44.42	31.72	45.29	36.03	30.47	37.17	32.73	31.69	37.85	34.99	30.91

Table 3: Analysis of M^{3}H-Att with various stacked layer number, head number, and subspace dimension.

Models	No Image Wording			Add OCR				Add Attribute			
	Full	OCR	Attr	Full	Δ (%)	OCR	Δ (%)	Full	Δ (%)	Attr	Δ (%)
CLS-MAX	38.31	36.11	32.04	38.75	+1.1	40.67	+12.6	41.09	+7.3	37.87	+18.2
GEN-COPY	42.01	40.81	35.55	42.86	+2.0	43.58	+6.8	43.11	+2.6	38.10	+7.2
CLS-M^{3}H-ATT	44.19	42.93	36.93	44.27	+0.2	46.53	+8.4	44.38	+0.4	38.73	+4.9
GEN-M^{3}H-ATT	44.33	43.26	35.93	44.48	+0.3	46.31	+7.1	44.77	+1.0	39.90	+11.0

Table 4: F1@1 over three test sets with various settings: no image wording, adding either OCR or attribute. Δ: the relative improvements over no image wording.

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 $\sim$ 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, as the image modality enriches the context for short texts.

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

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

M^{3}H-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^{3}H-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 deteriorate 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-

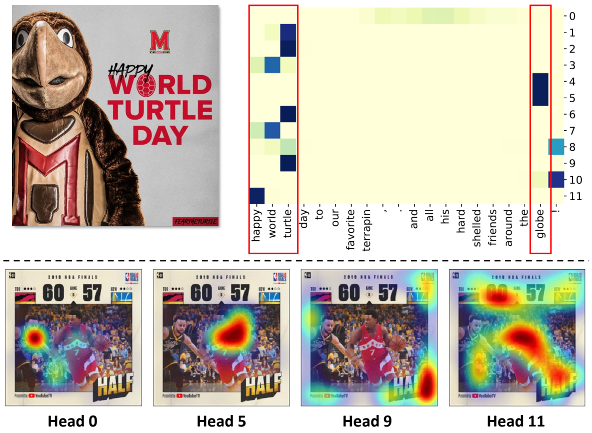
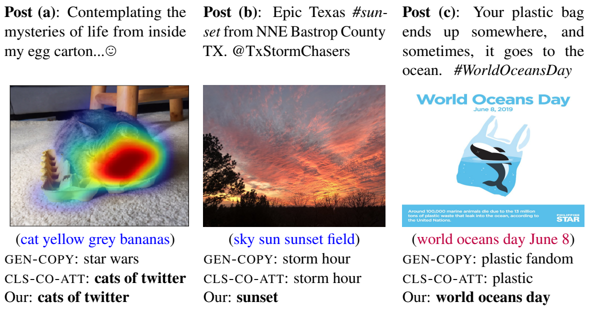


Figure 6: Attention weight visualization of M^{3}H-Att for two example posts with image-to-text (top) and text-to-image attention (bottom). Best viewed in color.



Post (a): Contemplating the mysteries of life from inside my egg carton...☺

Post (b): Epic Texas *#sunset* from NNE Bastrop County TX. @TxStormChasers

Post (c): Your plastic bag ends up somewhere, and sometimes, it goes to the ocean. *#WorldOceansDay*

(cat yellow grey bananas)
GEN-COPY: star wars
CLS-CO-ATT: **cats of twitter**
Our: **cats of twitter**

(sky sun sunset field)
GEN-COPY: storm hour
CLS-CO-ATT: storm hour
Our: **sunset**

(world oceans day June 8)
GEN-COPY: plastic fandom
CLS-CO-ATT: plastic
Our: **world oceans day**

Figure 7: Tweet image's effects for keyphrase prediction. Blue tokens are the top 4 attributes and purple ones are OCR tokens. Correct predictions are in **bold**.

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.[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 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 either generation or classification models with traditional attention.

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.

[10] Here we assume that multimedia posts with ImageNet keyphrases have a higher probability to contain natural photos.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086. IEEE Computer Society.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433.

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 747–754.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multimodal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2506–2515.

Rich Caruana, Steve Lawrence, and C. Lee Giles. 2000. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*, pages 402–408.

Hou Pong Chan, Wang Chen, Lu Wang, and Irwin King. 2019. Neural keyphrase generation via reinforcement learning with adaptive rewards. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2163–2174.

Tao Chen, Xiangnan He, and Min-Yen Kan. 2016. Context-aware image tweet modelling and recommendation. In *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*, pages 1018–1027. ACM.

Wang Chen, Hou Pong Chan, Piji Li, Lidong Bing, and Irwin King. 2019. An integrated approach for keyphrase generation via exploring the power of retrieval and extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2846–2856. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734.

Yuyun Gong and Qi Zhang. 2016. Hashtag recommendation using attention-based convolutional neural network. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 2782–2788.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pages 630–645.

Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. 2019. Dual attention networks for visual reference resolution in visual dialog. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2024–2033. Association for Computational Linguistics.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 1571–1581.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020. Cross-media structured common space for multimedia event extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2557–2568. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 13–23.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 289–297.

Olena Medelyan, Eibe Frank, and Ian H. Witten. 2009. Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1318–1327.

Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. 2017. Deep keyphrase generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 582–592.

Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2019. Towards personalized image captioning via multimodal memory networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(4):999–1012.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

R. Smith. 2007. An overview of the tesseract OCR engine. In *9th International Conference on Document Analysis and Recognition (ICDAR 2007), 23-26 September, Curitiba, Paraná, Brazil*, pages 629–633.

Hao Tan and Mohit Bansal. 2019. LXMERT: learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5099–5110. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Alakananda Vempala and Daniel Preotiuc-Pietro. 2019. Categorizing and inferring the relationship between the text and image of twitter posts. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2830–2840. Association for Computational Linguistics.

Xiaojun Wan and Jianguo Xiao. 2008. Single document keyphrase extraction using neighborhood knowledge. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008*, pages 855–860.

Yue Wang, Shafiq R. Joty, Michael R. Lyu, Irwin King, Caiming Xiong, and Steven C. H. Hoi. 2020. VD-BERT: A unified vision and dialog transformer with BERT. *CoRR*, abs/2004.13278.

Yue Wang, Jing Li, Hou Pong Chan, Irwin King, Michael R. Lyu, and Shuming Shi. 2019a. Topic-aware neural keyphrase generation for social media language. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2516–2526.

Yue Wang, Jing Li, Irwin King, Michael R. Lyu, and Shuming Shi. 2019b. Microblog hashtag generation via encoding conversation contexts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1624–1633.

Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning. 1999. KEA: practical automatic keyphrase extraction. In *Proceedings of the Fourth ACM conference on Digital Libraries, August 11-14, 1999, Berkeley, CA, USA*, pages 254–255.

Hua Wu, Haifeng Wang, and Zhanyi Liu. 2006. Boosting statistical word alignment using labeled and unlabeled data. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*.

Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony R. Dick, and Anton van den Hengel. 2016. What value do explicit high level concepts have in vision to language problems? In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 203–212.

Huijuan Xu and Kate Saenko. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*, volume 9911 of *Lecture Notes in Computer Science*, pages 451–466. Springer.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2016. Stacked attention networks for image question answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 21–29.

Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R. Lyu, and Irwin King. 2018. Topic memory networks for short text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3120–3131. Association for Computational Linguistics.

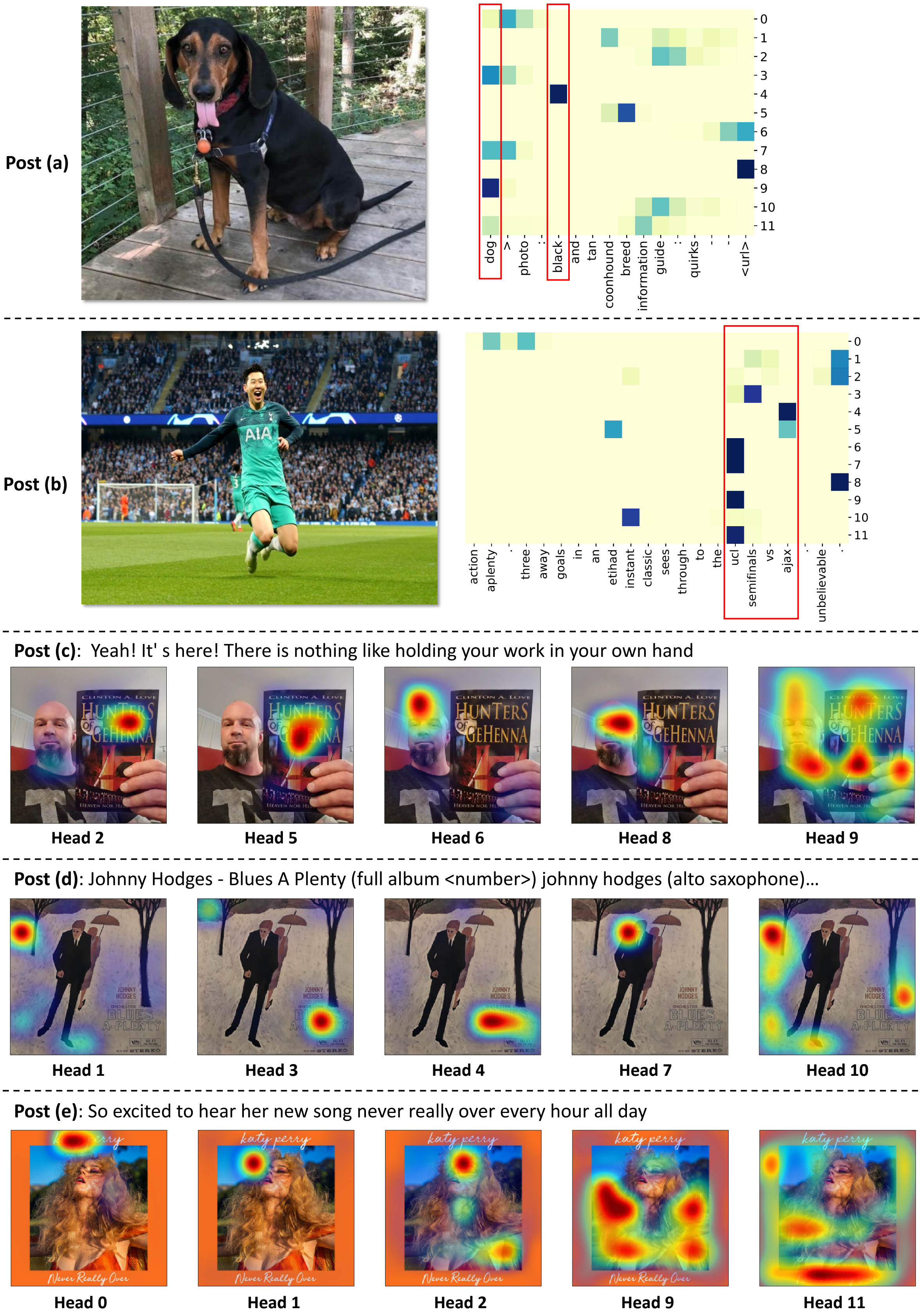
Qi Zhang, Jiawen Wang, Haoran Huang, Xuanjing Huang, and Yeyun Gong. 2017. Hashtag recommendation for multimodal microblog using co-attention network. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 3420–3426.

Qi Zhang, Yang Wang, Yeyun Gong, and Xuanjing Huang. 2016. Keyphrase extraction using deep recurrent neural networks on twitter. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 836–845.

Suwei Zhang, Yuan Yao, Feng Xu, Hanghang Tong, Xiaohui Yan, and Jian Lu. 2019. Hashtag recommendation for photo sharing services. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 5805–5812.

Yingyi Zhang, Jing Li, Yan Song, and Chengzhi Zhang. 2018. Encoding conversation context for neural keyphrase extraction from microblog posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1676–1686.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13041–13049. AAAI Press.



Post (c): Yeah! It' s here! There is nothing like holding your work in your own hand

Post (d): Johnny Hodges - Blues A Plenty (full album <number>) johnny hodges (alto saxophone)...

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*

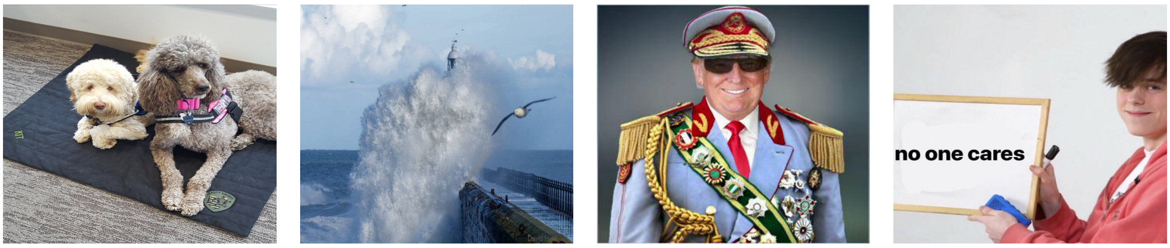


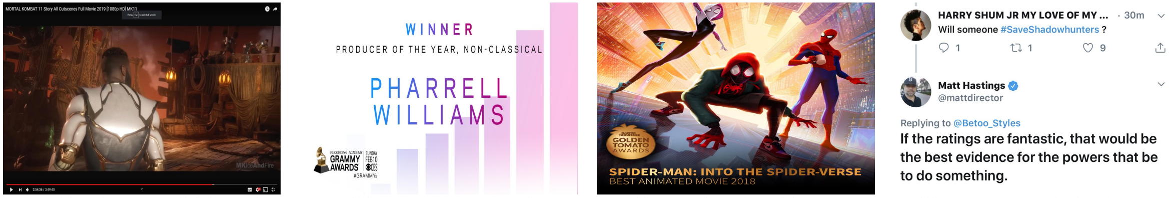
Figure 9: Example tweets of four different types of text-image relationship in our dataset. Post (a): text is represented and image adds to. Post (b): text is represented and image does not add to. Post (c): text is not represented and image adds to. Post (d): text is not represented and image does not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



HARRY SHUM JR MY LOVE OF MY ... · 30m
Will someone #SaveShadowhunters ?
1 1 9

Matt Hastings
@mattdirector

Replying to @Betoo_Styles
If the ratings are fantastic, that would be the best evidence for the powers that be to do something.

(mortal kombat story all full movie)
GEN-COPY: quote
CLS-CO-ATT: destiny 2
Our: **mortal kombat 11**

(williams at grammy awards)
GEN-COPY: live under par
CLS-CO-ATT: a star is born
Our: **grammys**

(spider man into the spider-verse)
GEN-COPY: spider verse
CLS-CO-ATT: marvel
Our: **spider man**

(will someone save shadow hunters)
GEN-COPY: teacher goals
CLS-CO-ATT: brexit
Our: **save shadowhunters**

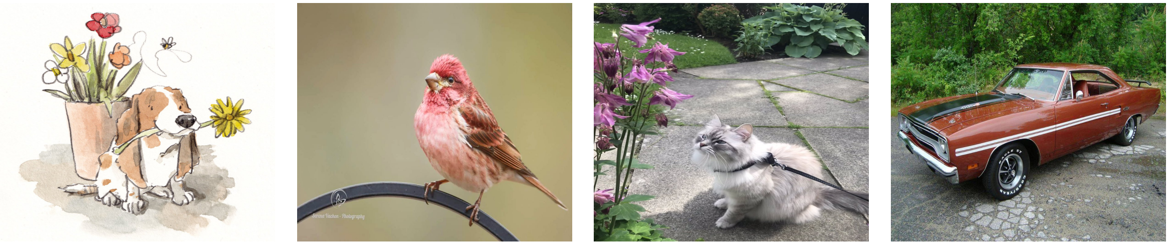
Figure 10: More qualitative examples showing the effectiveness of encoding OCR texts. Among various models, only our model that considers OCR tokens correctly predicts the keyphrases for all these cases (in bold). Purple tokens are the detected OCR tokens, where we observe that the keyphrases directly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



(dog white yellow brown plate)
GEN-COPY: friday feeling
CLS-CO-ATT: **hooray for dogs**
Our: **hooray for dogs**

(branch bird red top small)
GEN-COPY: gap ol
CLS-CO-ATT: **birding**
Our: **birds**; **wildlife**

(cat white pink grey flowers)
GEN-COPY: photography
CLS-CO-ATT: springwatch
Our: **cats of twitter**

(car roof park old meter)
GEN-COPY: plymouth
CLS-CO-ATT: mopar
Our: **classic car**

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.