Journal of Big Data

RESEARCH Open Access

Cross-modality representation learning from transformer for hashtag prediction

Mian Muhammad Yasir Khalil[1], Qingxian Wang[1], Bo Chen[1] and Weidong Wang[1*]

*Correspondence:
wdwang@uestc.edu.cn

[1] School of Information and Software Engineering, University of Electronic Science and Technology of China, Jianshe North Road, Chengdu 610054, China

Abstract

Hashtags are the keywords that describe the theme of social media content and have become very popular in influence marketing and trending topics. In recent years, hashtag prediction has become a hot topic in AI research to help users with automatic hashtag recommendations by capturing the theme of the post. Most of the previous work mainly focused only on textual information, but many microblog posts contain not only text but also the corresponding images. This work explores both image-text features of the microblog post. Inspired by the self-attention mechanism of the transformer in natural language processing, the visual-linguistics pre-train model with transfer learning also outperforms many downstream tasks that require image and text inputs. However, most of the existing models for multimodal hashtag recommendation are based on the traditional co-attention mechanism. This paper investigates the cross-modality transformer LXMERT for multimodal hashtag prediction for developing LXMERT4Hashtag, a cross-modality representation learning transformer model for hashtag prediction. It is a large-scale transformer model that consists of three encoders: a language encoder, an object encoder, and a cross-modality encoder. We evaluate the presented approach on dataset InstaNY100K. Experimental results show that our model is competitive and achieves impressive results, including precision of 50.5% vs 46.12%, recall of 44.02% vs 38.93%, and F1-score of 47.04% vs 42.22% compared to the existing state-of-the-art baseline model.

Keywords: Attention mechanism, Hashtag recommendation, Multimodal data, Transfer learning

Introduction

In recent years, social media networks such as Twitter, Instagram, and Sina Weibo have gained huge popularity. People use these platforms for communication and sharing their opinions on different daily activities. The rapid adoption of social media results in a huge volume of social media content on a daily basis. According to the latest statistics, Twitter has 330 million monthly active users, while the number for Pinterest has reached 444 million,[1] and Instagram has more than 2 billion active users.[2]

[1] https://www.socialpilot.co/blog/social-media-statistics/.

[2] https://techcrunch.com/2018/06/20/instagram-1-billion-users/.

Fig. 1 An Example of multimedia post from Instagram, where text offers limited information, without visual information, we can't recommend the correct hashtags

To avoid being overwhelmed, a good choice that improves information diffusion is through the use of hashtags. Hashtags indicate the theme of the social media post and have proven to be useful for influence, opinion analysis, forecasting, prediction, and other purposes. So, hashtag predictions have become an important research topic and have received considerable attention in recent years. Although many researchers have done work on hashtag predictions, most of the previous work has focused only on text information, which sometimes does not provide the context of user opinions. According to the data analysis, we observe that social media posts contain different sources of data, e.g., (texts, images, and videos) that express the user's opinions or daily life moments. So, it is not easy to correctly recommend hashtags for multimodal data using the model, which is designed based on text information. Figure 1 illustrates a multimodal micro-blog post from the Instagram[3] with hashtag #cat and the cat information is not given in the text content of the post. With only textual information, we may predict the hashtag about gifts and celebrations. However, the hashtag # cat is hardly to be identified.

Many approaches have been developed for hashtag recommendations for social media content. Most of these approaches have used traditional deep learning techniques, while recently, the transformer [1] base BERT [2] model has been adopted as an effective method for many classification problems. Unlike the previous research works, which either use image text simple fusion or the traditional co-attention approach, we adopt the cross-modality transfer learning model for the hashtag prediction task and convert this task into a transfer learning problem.

Inspired by the achievements of visual linguistic transformer in many downstream tasks, we adopt the cross-modality representation learning transformer mechanism to

[3] https://www.instagram.com/p/CcmHYf_rzVU/.

extract the features from the multimodal dataset with a cross-attention transformer layer to capture the interactions among images, texts, and hashtags. A multilabel classification head is added on top of the cross-attention layer(the last hidden state) to generate the probability score of each hashtag from the hashtags list. Our approach is based on the recently developed Learning Cross-Modality Encoder Representations from Transformers (LXMERT) [3] for hashtag prediction. LXMERT extends the recent language model BERT's self-attention mechanism for cross-modality vision-and-language interaction. We trained this cross-modality model on a multimodal hashtag dataset called LXMERT4Hashtag, a cross-modality transfer learning model for hashtag prediction. It is an efficient method of hashtag prediction from cross-modal representation learning. This model is based on separate streams for an object encoder and a language encoder that communicates through a cross-modality transfer encoder. In summary, in this article, we present the cross-modality representation learning transformer model for hashtag recommendation. The key contributions can be summarized as follows:

- We investigate the hashtag recommendation task for a multimodal dataset and propose a cross-attention-based transformer framework to extract the features from both image and text. And capture the correlation between hashtags and image-text features called LXMERT4Hashtag.
- We frame the hashtag recommendation task as a multi-label classification problem. To formulate this task, we employ the cross-modality representation learning transformer architecture to model this recommendation process.
- Extensive experimental results on the big dataset, crawled from the Instagram named InstaNY100K, demonstrate that our proposed model achieves better results for hashtag recommendation tasks by fully exploiting the cross-modal representation learning.

The rest of the paper is organized as follows: related Works section reviews the related research works. Approach section presents the proposed model, and Experiments section presents the experimental result and evaluates the performance of the proposed model. Conclusion section presents the conclusion and future work.

Related works

Our work relates to the hashtag recommendation for multimodal content using a cross-modality transformer model. While previous work has practised traditional deep learning. There is also increasing recognition of the importance of being able to handle multimodal social media content and cross-modal attention. We review some of this work in this section.

Hashtag recommendation

Due to the usefulness of hashtags in influence marketing, opinion mining, and many other purposes, hashtag recommendation has become an attractive research field in recent years. Researchers have proposed many approaches from different perspectives. Zangerle et al. [4] introduced an approach for highly appropriate hashtag recommendation based on TF-IDF content similarity of the user tweet and other tweets. The

recommendation aim is to encourage the user to use more appropriate hashtags and avoid synonymous hashtags. The Ref. [5] used Latent Dirichlet Allocation for hashtag recommendation for microblogs using a topic-specific translation model. Sedhai et al. [6] proposed a solution by learning-to-rank method for hyperlink tweets. First, they select the hashtags through five similarity schemes: similar documents, similar tweets, the named entities contained in the document, the domain of the link, and for hashtag recommendation, they adopt the RankSVM approach. Hashtag-LDA proposed by [7] finds meaningful latent topics and the relationships between topics and hashtags. Motivated by the success of convolutional neural networks (CNNs) in some natural language processing tasks, [8] adopted CNNs to perform the hashtag recommendation problem and proposed a novel architecture with an attention mechanism. The work [9] proposed TAB_LSTM, a hashtag recommendation model with Topical Attention-Based LSTM, which is an attention mechanism to merge local hidden representations with global topic vectors. They consider the attention process and construct a unique Topical Attention-Based LSTM model for hashtag recommendation. Li et al. [10] proposed a Long Short-Term Memory Recurrent Neural Network(LSTM-RNN) model for hashtag recommendation. They utilized tweet vector features to categorize hashtags without any feature engineering. In their work, distributed word representations are used with the skip-gram model. Then, the convolutional neural network is trained by semantic phrase vectors, the phrase vectors are then used to train an LSTM-RNN. Hashtag prediction based on multi-features of the microblogs proposed by [11] utilized a user-based method for hashtag prediction by taking advantage of similar users. Hashtag2Vec proposed by [12] explores the multiple relations of hashtag-tweet, tweet-word, word-word, and hashtag-hashtag relationships based on the hierarchical heterogeneous network. The work [13] proposed a framework DeepTagRec, a content-cum-user, based deep learning model to recommend appropriate hashtags on Stack-Overflow. The Ref. [14] proposed a Topical Co-Attention Network (TCAN), which learns the content representation based on a bidirectional LSTM and constructs the topical word matrix to represent the topic and combine them with the co-attention mechanism. Parallel Long Short-term Memory (PLSTM) proposed in [15] for hashtag recommendation task based on current post contents and the post history representation. Zhang et al. [16] proposed Semantically Enhanced Tag Recommendation (STR), a deep learning-based approach that recommends the tags through semantics learning of both tags and questions on software(e.g. Stack Overflow).

Many researchers have worked on image datasets for hashtag recommendation and have been using different approaches. Most of the researchers pay attention to the tags annotated by users through social media services such as Flickr. For example [17], introduced tag recommendation strategies to assist the user with photo annotation by recommending a set of tags to the photo. The Ref. [18] studied the problem of personalized tag recommendation tasks for the images. Liu et al. [19] proposed a tag ranking system that ranks the tags associated with given Flickr photos according to their relevance to the image content. In Li et al. [20] trains efficient ensembles of Support Vector Machines per tag, enabling fast classification. To learn the semantics of images, most recent approaches rely on CNNs as HARRISON [21] introduced a hashtag recommendation model for real-world photos in social media, which included a visual feature extractor

based on a (CNN)convolutional neural network for multi-label hashtag classifier. On the HARRISON dataset, two single feature-based models, object-based and scene-based models, as well as an integrated model of them, are evaluated using this framework. For personalized hashtag recommendation [22], proposed a deep learning model that considers the user's preferences and visual information for image tag recommendation. For sequence relationships between social media images and hashtags, an Attention-based neural image hashtag network (A-NIH) is proposed in [23]. CNN Inception V3 with LSTM-Atten is used for the sequential image feature, and GRU to generate the output. Kao et al. [24] combined image classification and semantic embedding models to design an efficient and resource-aware hashtag recommendation model based on deep neural networks. Voting Deep Neural Network with Associative Rules Mining (VDNN-ARM) framework proposed in [25], for image hashtag recommendation. For user representation learning for image hashtag prediction, Durand et al. [26] first extracted the visual representation for each image and then computed the vectorial representation of each image hashtag into pairs.

From the brief descriptions given above, we can observe that most of the previous works focused on either textual information or visual features. In this work, the proposed method incorporates both textual and visual information.

Multimodal hashtag recommendation

As information gets more diverse on social media, hashtag recommendations for multimodal data have attracted a lot of attention in recent years. Various approaches have been studied for multimodal hashtag recommendation tasks in different aspects. Previous works on hashtag recommendation for multimodal datasets usually focused on traditional co-attention for multimodal representation learning. For example [27], first extracted the image feature by 16-layer VGGNet [28], text feature by LSTM, and then fed them into a co-attention network that incorporates textual and visual information to recommend the hashtags for multimodal tweets. User habit-based hashtag recommendation proposed in [29] is based on two modules: a content model for post image and text data based on a parallel co-attention mechanism and a model for users' tagging habits. Finally, the post feature vector from the content modelling module and the habit influence vector from the user habit module are concatenated for hashtag recommendations. Attention-based Multimodal Neural Network model (AMNN) proposed by [30] for hashtag sequence prediction. To extract the image feature representations, a hybrid neural network architecture was adopted. In the first step, the preliminary feature map of a given image is captured by the CNN, and then LSTM is applied to process the intermediate features sequentially. Text feature extraction is considered by the BiLSTM model. Multimodal representation was obtained by concatenating image and text-distributed representations, and a gated recurrent unit (GRU) [31] network decoder was used for the final hashtag recommendation. Co-Attention Memory Network developed by [32] for multimodal microblog's hashtag recommendation combines the attention mechanism with memory networks for hashtag recommendation tasks. With the co-attention mechanism, it first gets text-based visual attention and image-based textual attention. It then feeds them into a cross-attention memory framework to extract the users' interests from the users' microblog history. CACNet [33] proposed a Cross-Active

Table 1 Summary of the state-of-the-art selected literature on hashtag recommendation

Refs.	Year	Name	Features	Techniques	Recommendation
[9]	2016	TAB_LSTM	Text	Topical Attention-Based LSTM model that merge local hidden representations with global topic vectors.	General
[11]	2018	HRMF	Text	Hashtag recommendation method based on multi-features(hashtag, user, text) of micro-blogs. To find similar users they proposed a new topic model User-Hashtag Topic Model Based on Short Text Expansion (UHTME).	Personalized
[12]	2018	Hashtag2Vec	Text	Explored the multiple relations of hashtag-tweet, tweet-word, word-word, and hashtag-hashtag relationships based on the hierarchical heterogeneous network.	General
[13]	2019	DeepTagRec	Text	A neural network model for tag recommendation on Stack Overflow, Quora, etc that leverages both the textual content (title and body) of the questions and the user-tag network.	Personalized
[14]	2019	TCAN	Text	Topical Co-Attention Network which learns the content representation with BiLSTM and constructs the topical word matrix and combines them with the co-attention mechanism.	General
[15]	2019	PLSTM	Text	Parallel Long Short-term Memory based on current post contents and the post history representation.	Personalized
[16]	2018	STR	Text	Semantically Enhanced Tag Recommendation approach that recommends the tags through semantics learning of both tags and questions on Stack Overflow.	General
[22]	2017	CNN-PerMLP	Image	For personalized hashtag recommendation consider the user's preferences as well as visual information for image tags recommendation.	Personalized
[23]	2018	A-NIH	Image	Attention-based Neural Image Hashtag network for sequence relationships between images and hashtags process with Inception V3 and GRU	General
[25]	2020	VDNN-ARM	Image	A voting deep neural network with associative rules mining approach.	General
[26]	2020	–	Image	A user conditional joint embedding model, it first extracted the visual representation for each image and then computed the vectorial representation of each image-hashtags into pairs.	Personalized
[27]	2017	CoA	Image, text	A deep learning framework that incorporates textual(LSTM) and visual(CNN) features of multimodal tweets with co-attention model.	General
[29]	2019	MACON	Image, text	Memory augmented co-attention network which learns the image and text fractures with user tagging habits.	Personalized
[30]	2020	AMNN	Image, text	A sequence generation Attention-based Multimodal Neural Network that extracts the features of images and texts and incorporates them into the sequence-to-sequence GRU model for hashtag recommendation.	General
[33]	2021	CACNet	Image, text	VGG and weighted average Word2Vec based Cross-Active Connection model for Image-Text Feature Fusion.	General

Table 1 (continued)

Refs.	Year	Name	Features	Techniques	Recommendation
[34]	2022	TweetEmbd_Net	Image, text	Explore the problem of jointly modeling tweet components (image, text, hashtags, user, and, location) in a common embedding space.	General

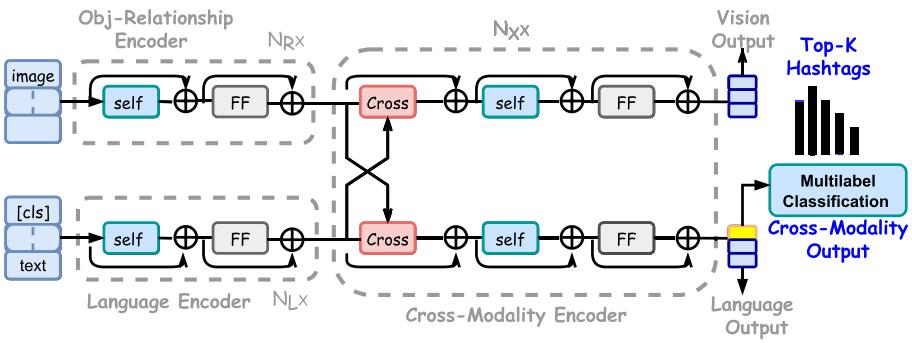

Fig. 2 The overview of LXMERT4 Hashtag. The [CLS] token on the top of the text embedding passes through the cross-attention layers and learns the joint representation. The corresponding feature vector of [CLS] token is represented as the yellow square on the top of the language output vector, where multilabel classification is used for top K hashtags recommendation

Connection Network for multimodal feature fusion. The work [34] proposed a deep neural network framework that combines the tweet components for representation learning of multimodal Twitter data for downstream applications, including hashtag recommendations tasks. An overview of the selected literature is shown in Table 1.

As mentioned before in the introduction section, these works are based on the traditional co-attention mechanism. Inspired by the recent success of the self-attention mechanism of transformer [1], BERT [2] has gained great achievement in natural language processing, and cross-modal learning has propelled great advancement in vision-and-language joint learning tasks. In this work, we adopt the cross-modal representation learning transformer mechanism for better learning cross-visual linguistic features for hashtag prediction.

Approach

We frame the hashtag recommendation task as a multi-label classification problem. To formulate this task, we adopt the cross-modality transformer architecture to model this recommendation process and use the cross-modality head of the LXMERT model similar to visual question answering. It learns the joint representations on the top of image and text embedding and assigns the probability score to each hashtag based on its relevancy to the hashtags list on the base of image-text features from cross-modality representations, where top k hashtags are selected and used for hashtag prediction. The model is proposed as follows: Given the input image with corresponding content pairs, we aim to train a cross-modality representation learning model with cross-attention following the self-attention mechanism of the transformer for the hashtag recommendation task. In this section, we describe the model architecture in detail.

Model architecture

Figure 2 presents the architecture of the proposed model. Self-attention sub-layers and cross-attention sub-layers are abbreviated as 'Self' and 'Cross,' respectively. A feed-forward sub-layer is denoted by the letter 'FF'. First, each sentence from the input text is represented as a sequence of word tokens, and the image as a sequence of objects. The model consists of two separate encoders for text and image representation that are based on the transformer self-attention mechanism and a cross-modality encoder with cross-attention layers following the transformer attention mechanism for visual linguistic interaction. The output is a vector, and each vector is represented by the probability of the hashtag, where probability is obtained from Binary Cross-Entropy loss.

Pre-processing

Both image and text input need to convert into two sequences of features. Each sentence from the input text is represented as a sequence of word tokens, and the image as a sequence of objects. Later, these embedding features will be further processed by the corresponding encoding layers.

Text embedding

Text is tokenized through Word Piece tokenizer [35] as in BERT [2]. Each sentence is split into words, and each word has an index $\{w_1,..., w_n\}$. Embeddings are generated as vectors for each word and its corresponding index through embedding layers. The final word embeddings are generated by adding both word and position indices as follows:

$$\begin{aligned} \hat{w}_i &= WordEmbed(w_i) \\ \hat{u}_i &= IdxEmbed(u_i) \\ h_i &= LayerNorm(\hat{w}_i + \hat{u}_i) \end{aligned} \tag{1}$$

Image embedding

The image embedding is based on object-level image embeddings. Faster R-CNN [36] with Bottom-up top-down attention [37] used for image embedding. Objects are detected via an object detector. Each object is represented by a position feature p_j and a region-of-interest feature f_j. The positioning feature represents the bounding box coordinates of the object, and the region-of-interest feature represents a 2048-dimensional region of each object. The final object embedding (v_j) is formed by adding both features p_j and f_j). Thus, image embedding is a position-aware object embedding. In order to obtain a balance between the two types of features, each feature embedding is normalized before addition.

$$\begin{aligned} \hat{f}_j &= LayerNorm\left(W_F f_j + b_F\right) \\ \hat{p}_j &= LayerNorm\left(W_P p_j + b_P\right) \\ v_j &= \left(\hat{p}_j + \hat{f}_j\right)/2 \end{aligned} \tag{2}$$

Single-modality encoders

The proposed model consists of two single modality encoders, one for text representation and the other for image representation. This two-stream architecture has some

advantages, we can reuse the existing pre-trained model, such as for the text encoder, and we can benefit from the pre-trained text model BERT. Another advantage of the two-stream design is that the model can be fine-tuned for retrieval tasks separately for images and text. Both encoders are based on self-attention and adopt the multi-head attention as described by Transformer [1], same as BERT for text encoder. In the following section, we describe each type in detail.

Text encoder

Index-aware word embeddings that consist of word w_i and its index $_i$ are fed to a transformer encoder. The N_L layers in the text encoder based on BERT are formed of a self-attention layer and a feed-forward layer followed by residual connection and normalization layers.

Object-relationship encoder

The object-relationship encoder includes one layer of location features in addition to the visual features to follow the order-less self-attention mechanism of the Transformer model. After obtaining the representation of 36 objects that are represented by a position feature and a region-of-interest feature, the embedding is sent to the object-relational encoder for subsequent processing. The N_R layers in the object-relationship encoder mainly include a self-attention layer and a feed-forward layer. In addition, the "+" sign in Fig. 2 represents the newly added layer structure, which is residual connection and layer normalization. The overall structure of this encoder is similar to the text encoder, where a transformer encoder takes the image embedding as input.

Cross-modality encoder

The cross-modality module is formed of (a) two self-attention layers, (b) one bi-directional cross-attention layer, and (c) two feed-forward layers. Each layer output is used as input to the next layer. The bi-directional cross-modality layer is formed of two uni-directional cross-attention layers, from language to image and from image to language. Residual connections and normalization layers are added after each layer. The output vectors represent the language features and vision features as follows:

$$\hat{h}_i^k = CrossAtt_{L\to R}\left(h_i^{k-1}, v_1^{k-1}, \ldots, v_m^{k-1}\right) \tag{3}$$

$$\hat{v}_i^k = CrossAtt_{R\to L}\left(v_j^{k-1}, h_1^{k-1}, \ldots, h_n^{k-1}\right) \tag{4}$$

The Cross-attention layer aims to exchange information between the language and vision modalities. The output of the cross-attention layer is then passed to the self-attention layers.

$$\begin{aligned} \widetilde{h}_i^k &= SelfAtt_{L\to L}\left(\hat{h}_i^k, \hat{h}_1^k, \ldots, \hat{h}_n^k\right) \\ \widetilde{v}_i^k &= SelfAtt_{R\to R}\left(\hat{v}_i^k, \hat{v}_1^k, \ldots, \hat{v}_m^k\right) \end{aligned} \tag{5}$$

Finally, the output of self-attention layers is passed to the feed-forward layers to give the final output.

Output representations

We consider the hashtag recommendation as a multi-label classification task. As shown in the right-most part of Fig. 2, the cross-modality encoder has three outputs for language, vision, and cross-modality. We adopt the cross-modality output representation for the hashtag recommendation task. For the cross-modality representation learning, a special [CLS] token on the top of the text embedding passes through the cross-attention layers from language to image and from image to language and learns the joint representation. The corresponding feature vector of [CLS] token is used as the cross-modality output which is represented by the yellow square on the top of the language output vector. Finally, we adopt the last hidden state of [CLS] token for multilabel classification to assign the probability score to each hashtag based on its relevancy on the base of image-text features, where top k hashtags are selected and used for hashtag prediction.

Algorithm 1 LXMERT for Cross-Modal Hashtag Prediction

```
procedure LXMERT4HASHTAG(captions, images, hashtags, max_sequence_length, batch_size, num_epochs)
    Initialize LXMERT model with a binary classification head
    • Load the pretrained LXMERT model, which includes the Text Encoder, Object-Relationship Encoder, and Cross-Modality Encoder.
    • Add a binary classification head.
    Tokenize captions and prepare images segments
    Split data into training and validation sets
    for epoch ← 1 to num_epochs do
        for batch ← 1 to (number_batches) do
            Perform Forward pass-through model to obtain logit
            Compute Binary Cross-Entropy Loss between logits and targets
            Backpropagate gradients and update model parameters
        end for
        Perform validation on the validation set and adjust hyperparameters.
    end for
    Initialize predicted_tags list
    for tweet in tweets do
        Tokenize caption and prepare image segments
        Perform Forward pass-through model to obtain predictions
        Threshold predictions to obtain predicted hashtags
        Append predicted hashtags to predicted_hashtags list
    end for
    return predicted_hashtags
end procedure
```

Experiments

We apply the cross-modality representation learning transformer encoder to the task of hashtag recommendation to evaluate the performance. In this section, we design experiments to answer the following research questions: (i) How much can cross-modality representation learn from transformer help in hashtag recommendation as compared to the traditional baseline methods? (ii) Does the cross-attention mechanism inspired by the self-attention approach of the transformer help for this task?

Table 2 Statistics of the InstaNY100K dataset

Posts	Images	Hashtags	Ave_h
100000	100000	164243	15.9

Ave_h is the average number of hashtags per post

Dataset

Since there are few public data sets available that contain both text and image for hashtag recommendation, we evaluate our model using the InstaNY100K[4] dataset collected by the authors of [38]. A brief introduction of the dataset is given as follows: InstaNY100K contains 100k Instagram public microblogs associated with New York City. This data is part of the InstaCities1M dataset that contains 1 M Instagram microblogs associated with one of the top 10 English-speaking cities all over the world. InstaNY100K contains 100k samples with both text and image. The detailed statistics are shown in Table 2. It can be observed that microblogs in InstaNY100K provide many more hashtags with a higher mean value in a more reasonable range. The unique hashtag in the corpus was 164243, and the average number of hashtags per microblog post was 15.9.

Considering the computer processing power for quality data selection, we keep the microblog posts that have at least five words in the text. Further, for fair training and testing, we remove the re-posted and keep the original ones. For hashtag selections, first, we cleaned all the hashtags that are not in the top five hundred most frequent ones. Furthermore, we remove all the hashtags that are specific to city names to train the model for general purposes, as the InstaNY100K dataset is associated with New York City. Finally, we remove the posts if they have less than five hashtags and the hashtags with low frequency in the list of the top hashtags. For better learning, we remove all the mentions(@name), non-alphabetic, and links(URL) tokens. To construct a mapping list, we apply the spelling correction for the hashtag and manually check it. We keep the hashtag word that appears in the middle of the text without the '#' prefix due to their semantic roles.

Experimental settings

Evaluation metrics

To compare the performance of the proposed model with baseline models for the hashtag prediction task, we select the following metrics for evaluation: We use precision, recall, and F1-score as the evaluation metrics to measure the overall performance. Precision means the percentage of "hashtags truly assigned" among "hashtags assigned by model". Recall denotes that "hashtags truly assigned" among "hashtags manually assigned". F1-score is the average of Precision and Recall.

[4] https://gombru.github.io/2018/08/01/InstaCities1M/.

$$
\begin{aligned}
Accuracy &= \frac{CorrectHashtagPredictions}{TotalNumberofPosts} \\
Precision &= \frac{TP}{TP + FP} \\
Recall &= \frac{TP}{TP + FN} \\
F1Score &= 2\frac{Precision * Recall}{Precision + Recall}
\end{aligned}
\tag{6}
$$

Where TP is True Positive, FP is False Positive, and FN is False Negative.

Baseline methods

We refer to our proposed model as LXMERT4Hashtag, a cross-modality representation learning transformer encoder for hashtag recommendation (Fig. 2). For evaluation of the proposed model, we compare the following baseline methods against our model:

- Topical Attention-based LSTM: TAB_LSTM [9] proposed a hashtag recommendation model for text-only content. It's a novel attention-based LSTM model which incorporates topic modeling into the LSTM architecture through an attention mechanism.
- LSTM-CNN Concat: We concatenate the LSTM and Vgg for the hashtag prediction task. For combining text and image features, we project the image and text representation to the same dimension and then concatenate the vectors for the hashtag prediction task.
- Co-Attention Network: CoA [27] is the state-of-the-art hashtag recommendation method for multimodal(image and text) microblog contents. This model applies the traditional co-attention mechanism to extract post features and then directly uses the features to make recommendations.
- Attention-based Multimodal Neural Network model: AMNN [30] a sequence generation attention-based multimodal neural network that first extracts text and image features separately using the neural network with attention mechanism (encoder). Then, it merges the image and text representations and feeds the output values into GRU networks to generate a sequence of the recommended hashtags (decoder).
- Tweet Embedding Network: TweetEmbd_Net [34] proposed a task-agnostic model for multimodal Twitter data. This model combines the tweet components (image, text, hashtags, user, location, and time) for representation learning for downstream applications, including hashtag recommendation tasks.

Experimental setup

We perform the training task as follows: We split the entire dataset into three parts with a ratio of 8:1:1 for training, validation, and test set, respectively. We train our model on the training dataset to learn cross-modal joint embedding. A multilabel classification head is used on the top of the cross-modal joint embedding (the last hidden state of [CLS] token) to generate the probability score of each hashtag from the hashtags list. We save the model which has the best performance on the validation dataset. For the

Table 3 Evaluation results of different models for hashtag recommendation task. LXMERT4Hashtag obtained better performance over other methods

Methods	Precision	Recall	F1
TweetEmbd_Net	0.272	0.106	0.152
AMNN	0.327	0.103	0.156
TAB_LSTM	0.469	0.177	0.257
Con(LSTM_CNN)	0.51	0.192	0.279
CoA	0.532	0.202	0.293
LXMERT4Hashtag	0.599	0.233	0.335

performance evaluation of the proposed model, we use the test dataset embedding through our trained model and then perform the multilabel classification to generate the probability score for each hashtag. For training all the baseline models and our proposed model, we truncated or padded each text sequence into 128-word tokens length. Instead of using the feature output from the convolutional neural network, we adopt the FRCNN [36] with pre-trained BUTD [37] for the proposed model to extract the sequence of 36 objects from each image by bounding boxes on the image. Each object from the sequence is represented by its position feature and its 2048-dimensional region-of-interest (RoI). For the baseline models, image features are extracted by using their visual encoder. BertAdam optimizer is used to optimize the proposed model. For all the baseline methods and our model, we use the validation data to tune the hyperparameters and report the performance of all the models on the same test dataset. All the models are trained for 75 epochs with a batch size of 18. Each model can be fit into 1 Nvidia Titan V GPU with a batch size of 18.

The model is proposed as given the input image with corresponding content pairs, where we aim to train a cross-modality modal for hashtag recommendation task with cross-attention based on the transformer attention mechanism. The whole process is illustrated in Fig. 2. We adopt the cross-modality head of the model, which learns the cross-modality representation on the top of image and text representations, and a multilabel classification head is used to assign the probability score to each hashtag based on its relevancy to the hashtags list on the basis of image-text features from cross-modality representation, where top k hashtags are selected and used for hashtag prediction.

Results and discussion

Since a large number of social media users prefer to share posts with corresponding images, therefore finding an effective way to get meaningful information, both from the textual and visual content of the post, is very important for social media analyses.

Effectiveness Comparisons Observing the comparisons of our proposed model with base models, it is clear that the cross-modality representation learning transformer encoder can significantly improve the performance of the hashtag recommendation task. We evaluated the proposed model with state-of-the-art baseline methods using the same dataset. In Table 3, we compare the result of our proposed model and the state-of-the-art baseline multiple methods. The result based on the evaluation metrics shows that our proposed model outperforms the other methods. The evaluation results presented

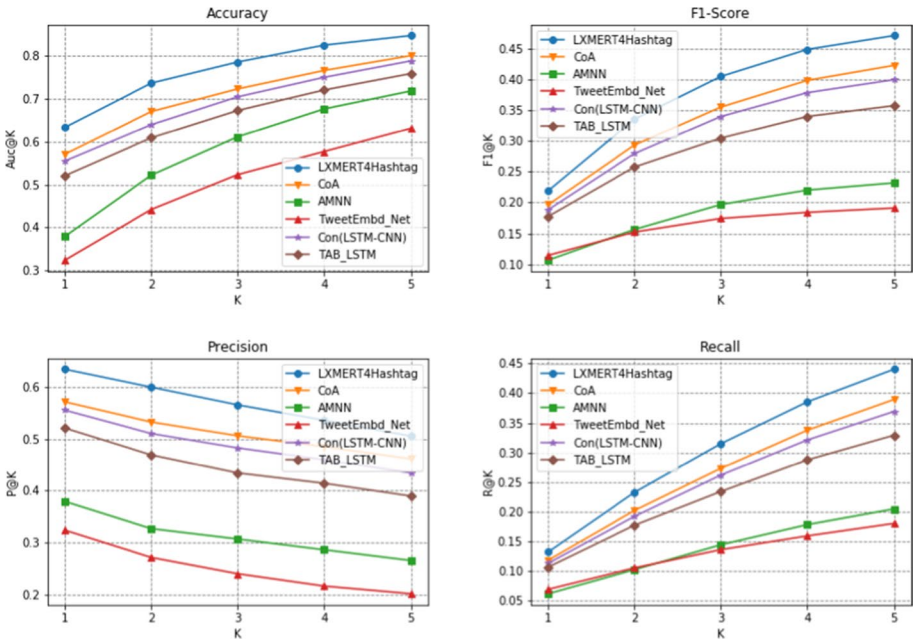

Fig. 3 Accuracy, precision, recall, and F1-score curves with different numbers of hashtags recommendation. LXMERT4Hashtag significantly outperforms the compared methods in all the evaluation metrics

in Table 3 were obtained with the top two hashtag recommendations for each post in the test dataset.

Figure 3 shows the result of the evaluation metrics with top K numbers of recommended hashtags on the test dataset. Each point on the curve represents the number of recommended hashtags ranging from 1 to 5. The y-axis represents the accuracy, precision, recall, and F1-score, respectively; the x-axis indicates the top K numbers of recommended hashtags. A model that has the highest curve on the graph indicates the best performance compared to other methods. We can observe from the curve that the performance of our proposed model is the highest compared to all the baseline methods. Figure 3 indicates that when K varies from 1 to 5, our proposed method can achieve 6.2% - 4.4%, 1.4% - 5.1%, and 2.3% - 4.8% absolute improvements in terms of the precision, recall, and F1-score, respectively compared with the best competitor CoA. The proposed model significantly outperforms all the baseline methods on all the evaluation metrics. These remarkable improvements represent the effectiveness of the cross-attention mechanism of the proposed model in the hashtag recommendation task.

In the baseline methods, CoA comparatively performs better than the other methods because of its co-attention mechanism. Although AMNN also considers both visual and linguistic contents, it performs comparatively poorly compared to CoA. The main reason is that AMNN is a sequence generation model that uses GRU networks to generate a sequence of recommended hashtags. However, in the dataset, we removed all the hashtags below the threshold, which means that the input hashtags in the dataset do not represent a good sequence of hashtags. This discrepancy in the dataset affects the performance of AMNN, leading to suboptimal results compared to CoA. The performance of the TweetEmbed_Network depends on components such as image, text, hashtags, user, location, and time. In our case, we used images, text, and

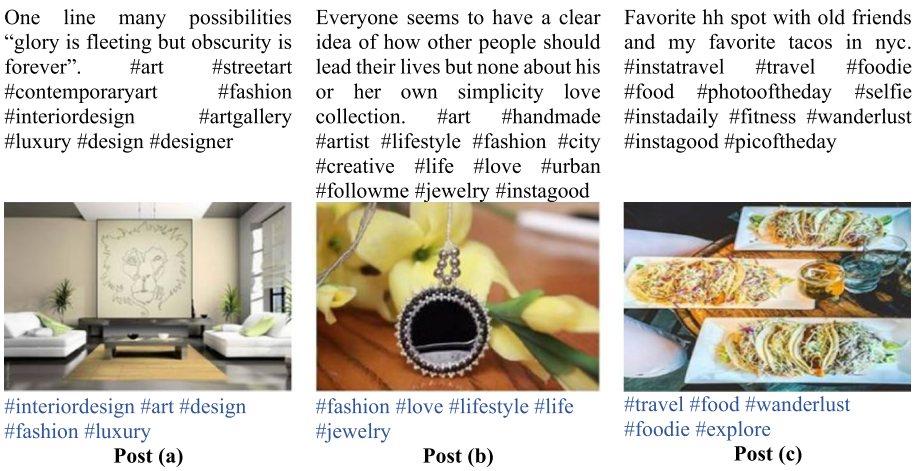

Fig. 4 Some examples of hashtag prediction from LXMERT4Hashtag

hashtags. However, we acknowledge that location can have an important impact, and another possible reason could be the loss function. TAB_LSTM is a text-only method, and we observe that TAB_LSTM performs much better. This indicates the usefulness of topic attention in hashtag recommendation tasks.

The experimental results of the proposed model demonstrate that the cross-modality representation learning with a transformer can generate better correlations between the hashtags and the visual-linguistic contents of the social media post. The competitive performance of the proposed model shows that the cross-attention of the cross-modality transformer encoder can produce the high-level representation learning of the multimodal social media content compared to other baseline methods. Based on the above discussion, it can be concluded that the proposed LXMERT4Hashtag performs better in the hashtag recommendation task compared to the state-of-the-art methods.

Qualitative analysis

In this section, we present some examples to conduct the qualitative analysis of our proposed model result. These examples cover a broad range of feature learning of the proposed model, including the objects, cross-attention mechanism, and implicit learning. Post (a), in Fig. 4, we show the rich features representation learned by LXMERT4Hashtag from the corresponding image of an Instagram post. In this example, the hashtags are correctly recommended by using the features of the post image. We observe that the hashtags #interiordesign, #art, and #design are very difficult to predict with post text, but the visual features give the clue. Post (b), we observe that the hashtags #fashion and #lifestyle do not appear in the microblog post. The hashtags #fashion and #lifestyle are both correctly predicted by the cross-media representation learned by LXMERT4Hashtag of an Instagram post. From the post (c), we observe that in addition to the explicit features, our model can also extract implicit meaning such as #travel and #explore, which demonstrates that the context ability of the cross-attention mechanism of the proposed model is very effective.

Conclusion

In this article, we presented the cross-attention mechanism inspired by the self-attention approach of the transformer for the hashtag recommendation task. The presented algorithm is based on the multimodality of social media content and hashtags. We converted the hashtag recommendation task into a multi-label classification and introduced a cross-modality representation learning transfer model for this task. As both image and text contents of the user's social media post are important in the hashtag recommendation task, the proposed model adopts the cross-attention layer to exchange the information and align the entities between image and text modalities in order to learn joint cross-modality representations.

We designed the experiments to evaluate the proposed model against several state-of-the-art models. Our model significantly outperformed and achieved state-of-the-art performance for hashtag prediction tasks over the baseline methods, including the traditional co-attention mechanism, and we found that cross-modality representation learning from the transformer encoder significantly helped in this task.

In the future, there are some options that we would like to investigate in the domain of multimodal hashtag prediction. We want to consider the different architecture of the cross-modality transformer encoder where the features from different modalities can fuse in a better way. Further, we also want to investigate the effectiveness of the visual transformer encoder for the improvement of our model based on the recent competitive visual transformer model. Furthermore, we also want to explore the training strategies to make it generalize for the settings where some hashtags have never been observed during training. We believe this work could be designed with different model architectures to achieve even better results that can handle the cross-modality features from social media posts for hashtag prediction. We left these issues to be further optimized and improved in our future work.

Acknowledgements
We would like to thank the authors of the InstaNY100K dataset and LXMERT model for the availability of code on Github. We also want to thank Adham Alkhadrawi for his support throughout this project and Ata Ur Rahman Khalid for helpful discussions.

Author Contributions
MMYK conducted the experiments and wrote the manuscript. QW assisted with experiments and researched related work. BC provided guidance in the direction of the study, and WW provided guidance and coordination in the experiments and manuscript writing. All authors read and approved the final manuscript.

Funding
Parts of this research were funded by the National Natural Science Foundation of China under grants 62072429, the Sichuan Science and Technology Program under Grant 2021YFG0305, 2021YFQ0054 and 2022YFG0175, the Intelligent Terminal Key Laboratory of Sichuan Province under Grant SCITLAB-0003 and SCITLAB-1003, and the Fundamental Research Funds for the Central Universities under grant ZYGX2020ZB021.

Availability of data and materials
The dataset InstaNY100K used in our experiments is available at GitHub https://gombru.github.io/2018/08/01/InstaCities1M/ collected by the authors of [38].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 31 May 2022 Accepted: 12 September 2023
Published online: 28 September 2023

References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A.N, Kaiser Ł, Polosukhin I. Attention is all you need. Adv Neural Inf Process Syst. 2017-Decem(Nips), 2017; 5999–6009. arXiv:1706.03762
2. Devlin J, Chang M.W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol - Proc Conf. 1(Mlm), 2019; 4171–4186. arXiv:1810.04805
3. Tan H, Bansal M. 2020 LXMert: Learning cross-modality encoder representations from transformers. EMNLP-IJCNLP 2019 - 2019 Conf Empir Methods Nat Lang Process. 9th Int Jt Conf Nat Lang Process Proc Conf. 2020; 5100–5111. https://doi.org/10.18653/v1/d19-1514. arXiv:1908.07490
4. Zangerle E, Gassler W, Specht G. Recommending #-tags in Twitter. CEUR Workshop Proc. 2011;730:67.
5. Ding Qi Zhang uanJ ing Huang Z.X. Automatic hashtag recommendation for microblogs using topic-specific translation model TITLE AND ABSTRACT IN CHINESE, 2012; 265–274.
6. Sedhai S, Sun A. Hashtag recommendation for hyperlinked tweets. SIGIR 2014 - Proc 37th Int ACM SIGIR Conf Res Dev Inf Retr., 2014; 831–834: https://doi.org/10.1145/2600428.2609452
7. Zhao F, Zhu Y, Jin H, Yang LT. A personalized hashtag recommendation approach using LDA-based topic model in microblog environment. Futur Gener Comput Syst. 2016;65:196–206. https://doi.org/10.1016/J.FUTURE.2015.10.012.
8. Yuyun G, Qi Z. Hashtag recommendation using attention-based convolutional neural network. IJCAI Int Jt Conf Artif Intell. 2016;16:2782–8.
9. Li Y, Liu T, Jiang J, Zhang L. Hashtag recommendation with topical attention-based LSTM. COLING 2016 - 26th Int Conf Comput Linguist Proc. COLING 2016 Tech Pap. 2016; 3019–3029
10. Li J, Xu H, He X, Deng J, Sun X. Tweet modeling with LSTM recurrent neural networks for hashtag recommendation. Proc Int Jt Conf. Neural Networks 2016-October, 2016; 1570–1577: https://doi.org/10.1109/IJCNN.2016.7727385
11. Kou FF, Du JP, Yang CX. Hashtag recommendation based on multi-features of microblogs. J COM-PUTER Sci Technol. 2018;33(4):711–26. https://doi.org/10.1007/s11390-018-1851-2.
12. Liu J, He Z, Huang Y. Hashtag2Vec: Learning hashtag representation with relational hierarchical embedding model. 2018.
13. Maity SK, Panigrahi A, Ghosh S, Banerjee A, Goyal P, Mukherjee A. DeepTagRec: a content-cum-user based tag recommendation framework for stack overflow. In: Azzopardi L, Stein B, Fuhr N, Mayr P, Hauff C, Hiemstra D, editors. Lecture notes computer science. Cham: Springer; 2019. p. 125–31. https://doi.org/10.1007/978-3-030-15719-7-16.
14. Li Y, Liu T, Hu J, Jiang J. Topical Co-attention networks for hashtag recommendation on microblogs. Neurocomputing. 2019;331:356–65. https://doi.org/10.1016/J.NEUCOM.2018.11.057.
15. Peng M, Bian Q, Zhang Q, Gui T, Fu J, Zeng L, Huang X. Model the Long-Term Post History for Hashtag Recommendation. Lect Notes Comput Sci. (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 11838 LNAI, 2019; 596–608. https://doi.org/10.1007/978-3-030-32233-5_46
16. Zhang J, Sun H, Tian Y, Liu X. Poster: Semantically enhanced tag recommendation for software CQAs via deep learning. https://doi.org/10.1145/3183440.3194977
17. Sigurbjörnsson B, Van Zwol R. Flickr tag recommendation based on collective knowledge. Proceeding 17th Int Conf World Wide Web 2008, WWW'08, 2008; 327–336. https://doi.org/10.1145/1367497.1367542
18. Garg N, Weber I. Personalized, interactive tag recommendation for flickr. RecSys'08 Proc. 2008 ACM Conf Recomm Syst. 2008; 67–74. https://doi.org/10.1145/1454008.1454020
19. Liu D, Hua X.S, Yang L, Wang M, Zhang H.J. Tag ranking. WWW'09 - Proc. 18th Int. World Wide Web Conf. 2009; 351–360 . https://doi.org/10.1145/1526709.1526757
20. Li X, Snoek C.G.M. Classifying tag relevance with relevant positive and negative examples. MM 2013 - Proc. 2013 ACM Multimed Conf. 2013; 485–488. https://doi.org/10.1145/2502081.2502129
21. Park M, Li H, Kim J. HARRISON: a Benchmark on HAshtag Recommendation for Real-world Images in Social Networks 2016; arXiv:1605.05054
22. Nguyen H.T.H, Wistuba M, Grabocka J, Drumond L.R, Schmidt-Thieme L. Personalized deep learning for tag recommendation. Lect Notes Comput Sci. (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 10234 LNAI, 2017; 186–197 . https://doi.org/10.1007/978-3-319-57454-7_15
23. Wu G, Li Y, Yan W, Li R, Gu X, Yang Q. Hashtag Recommendation with Attention-Based Neural Image Hashtagging Network. Lect Notes Comput Sci. (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 11302 LNCS, 2018; 52–63. https://doi.org/10.1007/978-3-030-04179-3_5
24. Kao D, Lai K.T, Chen M.S. An efficient and resource-aware hashtag recommendation using deep neural networks. Lect Notes Comput Sci. (including Subser. Lect Notes Artif Intell Lect Notes Bioinformatics) 11440 LNAI, 2019; 150–162: https://doi.org/10.1007/978-3-030-16145-3_12
25. Hachaj T, Miazga J. Image hashtag recommendations using a voting deep neural network and associative rules mining approach. Entropy. 2020;22(12):1351. https://doi.org/10.3390/E22121351.
26. Durand T. Learning user representations for open vocabulary image hashtag prediction. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2020; 9766–9775. https://doi.org/10.1109/CVPR42600.2020.00979
27. ; Zhang Q, Wang J, Huang H, Huang X, Gong Y. Hashtag recommendation for multimodal microblog using co-attention network. IJCAI Int Jt Conf Artif Intell. 0, 2017; 3420–3426: https://doi.org/10.24963/ijcai.2017/478
28. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2015; arXiv:1409.1556v6

29. Zhang S, Yao Y, Xu F, Tong H, Yan X, Lu J. Hashtag recommendation for photo sharing services. 33rd AAAI Conf. Artif. Intell. AAAI 2019, 31st Innov Appl Artif Intell Conf. IAAI 2019 9th AAAI Symp Educ Adv Artif Intell. EAAI 2019. 2019; 5805–5812 . https://doi.org/10.1609/aaai.v33i01.33015805
30. Yang Q, Wu G, Li Y, Li R, Gu X, Deng H, Wu J. AMNN: attention-based multimodal neural network model for hashtag recommendation. IEEE Trans Comput Soc Syst. 2020;7(3):768–79. https://doi.org/10.1109/TCSS.2020.2986778.
31. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. EMNLP 2014 - 2014 Conf Empir Methods Nat Lang Process Proc Conf. 2014; 1724–1734 . https://doi.org/10.3115/V1/D14-1179. arXiv:1406.1078
32. Ma R, Qiu X, Zhang Q, Hu X, Jiang YG, Huang X. Co-attention memory network for multimodal microblog's hashtag recommendation. IEEE Trans Knowl Data Eng. 2021;33(2):388–400. https://doi.org/10.1109/TKDE.2019.2932406.
33. Im J.H, Cho W, Kim D.S. Cross-active connection for image-text multimodal feature fusion. vol. 12801 LNCS, pp. 343–354. Springer. 2021; https://doi.org/10.1007/978-3-030-80599-9_30
34. Rivas R, Paul S, Hristidis V, Papalexakis EE, Roy-Chowdhury AK. Task-agnostic representation learning of multimodal twitter data for downstream applications. J Big Data. 2022. https://doi.org/10.1186/s40537-022-00570-x.
35. Wu Y, Schuster M, Chen Z, Le Q.V, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, Klingner J, Shah A, Johnson M, Liu X, Kaiser Ł, Gouws S, Kato Y, Kudo T, Kazawa H, Stevens K, Kurian G, Patil N, Wang W, Young C, Smith J, Riesa J, Rudnick A, Vinyals O, Corrado G, Hughes M, Dean J.: Google's neural machine translation system: bridging the gap between human and machine translation. arXiv:1609.08144v2
36. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2015;39(6):1137–49. https://doi.org/10.1109/TPAMI.2016.2577031.
37. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L. Bottom-up and top-down attention for image captioning and visual question answering. InProceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 6077–6086).
38. Gomez R. Learning to learn from web data through deep semantic embeddings. arXiv:1808.06368v1

Publisher's Note