



國立台灣大學管理學院資訊管理學研究所

碩士論文

Department of Information Management

College of Management

National Taiwan University

Master Thesis

基於社群媒體內容之個人化 Hashtag 推薦模型

A Personalized Hashtag Recommendation Model Based
on Social Media Contents

李欣

HSIN LEE

指導教授：李瑞庭 博士

Advisor: Anthony J. T. Lee, Ph.D.

中華民國 113 年 1 月

January 2024

謝辭

在臺大資管的碩士生生活即將結束。回顧這段時光，由衷感謝許多人的支持與鼓勵。多虧有導師的悉心指導，家人朋友的無私支持，以及同窗們的相互鼓舞，使我能夠順利地完成研究所學業。



首先，十分感謝我的指導老師李瑞庭教授，在研究過程中的細心指導與關照。從文獻探討、研究主題選擇、模型實驗設計到論文撰寫等方面，給予相當全面的專業建議，且在過程中傾注了大量時間和心力，細心地審視每一個細節。在碩士研究的歷程中，老師不僅指導我們該如何做研究，也時常在研究遇到困難時，引領我們解決問題。老師的專業知識、嚴謹治學態度和對教學的熱情，讓我受到了相當大的啟發，在加深專業知識積累的同時，還學到了正確的學術態度。除此之外，也十分感謝兩位口試委員吳怡瑾教授和向倩儀教授提供寶貴的建議，使得我的研究更加完善。

接著，我想感謝家人和朋友們的支持。你們的陪伴、鼓勵和無私付出，使我能夠毫無後顧之憂地完成碩士學位。在我面臨挫折和疲憊時，你們給予的安慰和鼓勵是我堅持下去的力量源泉。

最後，特別感謝實驗室的同儕，包括溫暖的學長姐、可愛貼心的學弟妹以及一同奮鬥的同屆夥伴，感謝有你們一起努力、成長，互相支持，才能讓我的研究順利進行，創造了許多難忘的回憶。希望未來工作中我們仍能保持聯繫，成為永遠的朋友。

感謝所有在我求學路上給予幫助和鼓勵的人們，讓我的學生生活豐富多彩、充滿挑戰。雖然即將告別碩士生生活，我將把這些寶貴經驗帶到人生的下一個階段，繼續成長與前行。

李欣 謹識

於臺灣大學資訊管理學研究所

中華民國一百一拾三年一月

論文摘要



論文題目：基於社群媒體內容之個人化主題標籤推薦模型

作者：李欣

指導教授：李瑞庭 博士

為了提升貼文能見度以及讓貼文呈現個人化或品牌形象，在本研究中，我們提出一個個人化主題標籤推薦模型，我們提出的模型包括四個階段，首先，我們萃取每篇貼文的視覺與文字特徵，接著，我們設計一個貼文與標籤的注意力機制以取得使用者標注的習慣以及貼文與標籤的關係，然後，我們開發一個六重共注意力機制產生貼文表示式，最後，我們利用關係圖卷積網絡來強化貼文表示式，並利用強化後的貼文表示式推薦主題標籤。實驗結果顯示，我們的模型在命中率、精確度、召回率和 F1 分數等指標上均優於現有模型。我們的模型可幫助使用者使用個人化的主題標籤，亦可幫助企業加速主題標籤標注的工作，進而提升貼文的能見度、社群參與度以及使用者間的互動。

關鍵字：主題標籤推薦、注意力機制、關係圖卷積網絡、記憶網絡

THESIS ABSTRACT

A Personalized Hashtag Recommendation Model Based on Social Media Contents

By Hsin, Lee

MASTER DEGREE OF BUSSINESS ADMINISTRATION

DEPARTMENT OF INFORMATION MANAGEMENT

NATIONAL TAIWAN UNIVERSITY

JANUARY 2024

ADVISOR: Anthony J. T. Lee, Ph.D.

To increase the post visibility and present personalized or brand image, in this study, we propose a novel personalized hashtag recommendation framework to recommend hashtags for user posts. The proposed framework contains four phases. First, we extract the visual and textual features from each post. Second, we devise the post-tag attention mechanism to acquire user tagging habits and post-tag relationships through similar posts. Third, we employ the sextuplet co-attention mechanism to derive the representation of each post. Finally, we adopt the relational graph convolution network (R-GCN) to construct the interaction graph between posts and update the post representations by propagating information among neighboring nodes. Then, we use the convoluted post representations to recommend hashtags for users. The results from the experiments demonstrate that the suggested framework surpasses the performance of current state-of-the-art models, achieving higher scores in hit rate, precision, recall, and F1 score. Our proposed framework may offer an effective tool for users to quickly make personalized hashtags on their posts and for businesses to accelerate the task of creating hashtags, which in turn increases post visibility, fosters community engagement, and enhances user interactions.

Keywords: Hashtag recommendations, Attention mechanism, Relational Graph Convolution Network, Memory Network



Table of Contents



Chapter 1 Introduction	1
Chapter 2 Related Work.....	4
2.1 Hashtag Recommendation.....	4
2.2 Memory Network	6
2.3 Attention Mechanism	6
Chapter 3 The Proposed Framework	8
3.1 Feature Extraction	9
3.1.1 Image Encoder.....	9
3.1.2 Text Encoder	12
3.2 Post-Tag Attention Mechanism.....	12
3.3 Sextuplet Co-attention Mechanism	14
3.4 Graph Convolution and Hashtag Recommendation.....	15
Chapter 4 Experimental Results.....	18
4.1 Dataset and Experiment Setup	18
4.2 Performance Evaluation	19
4.4 Effects of Attention Mechanisms	22
4.5 Recommendation Examples	26
Chapter 5 Conclusions and Future Work.....	29
References.....	32
Appendix.....	36

List of Figures



Fig. 1. Example Posts.....	1
Fig. 2. The Proposed Model.....	8
Fig. 3. Memory Network	10
Fig. 4. Sextuplet Co-Attention Mechanism	15
Fig. 5. Performance of Our and Comparing Models	19
Fig. 6. Heatmap of the Post-Tag Attention Mechanism	23
Fig. 7. Posts of User 11016.....	24
Fig. 8. Heatmaps of the Sextuplet Co-attention Mechanism	24
Fig. 9. Posts of Users 111, 2058 and 3378.....	25
Fig. 10. Recommendation Examples	28

List of Tables

Table 1. Statistics of the Dataset	18
Table 2. Performance of Recommending Top 9 Hashtags	20
Table 3. Performance of Variant Models	20
Table 4. Impact of Each Type of Features.....	22
Table 5. Impact of Visual and Textual Relations	22
Table 6. Hyperparameters Used in Our Model.....	36



Chapter 1 Introduction

Social media becomes indispensable for users, influencers, and companies to expand their reach and engage with their target audience nowadays. However, it is challenging for them to make their posts visible in a large number of media contents. To increase post visibility and present personalized or brand image, they often put hashtags on their posts. According to Statista, people often use three to four hashtags in a post to effectively increase the social exposure of a post on Instagram.¹ Moreover, posts with at least one hashtag get 29% more interactions than posts without any hashtag.² Also, Rauschnabel et al. (2019) showed that users tend to use personalized hashtags as a way of self-presentation on social media platforms. Personalized hashtags help brands and users

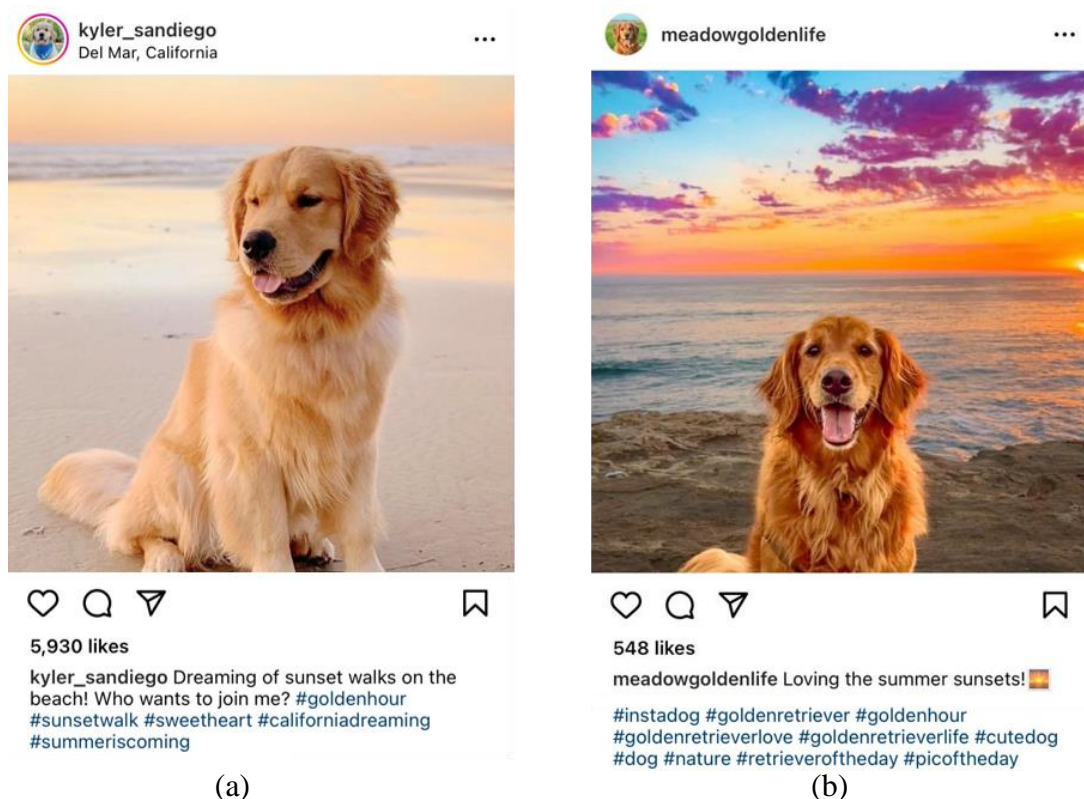


Fig. 1. Example Posts

¹ <https://www.statista.com/statistics/1305917/instagram-post-engagement-by-number-of-hashtags/>

² <https://comparecamp.com/hashtags-statistics/>


create a specific image, which has created market segmentation and achieved social media marketing effects.

Fig. 1. shows the example posts made by two users on Instagram. Both posts look quite similar, but they have very different hashtags. Therefore, it is essential and desirable to develop a personalized hashtag recommendation method for users to choose appropriate hashtags based on their tagging preferences.

Many hashtag recommendation methods (Ding et al. 2013, Sedhai and Sun 2014, Gong and Zhang 2016, Yang et al. 2016, Zhang et al. 2017, Gong et al. 2018, Wu et al. 2018, Yang et al. 2020) recommend hashtags for a post according to the post content. However, they do not consider user tagging habits and post-tag relationships that similar posts tend to have similar hashtags. Some methods (Denton et al. 2015, Huang et al. 2016, Veit et al. 2018, Ma et al. 2019, Zhang et al. 2019, Javari et al. 2020, Chen et al. 2022) consider post contents and user tagging habits but do not consider the post-tag relationships.

Therefore, in this study, we propose a novel personalized hashtag recommendation framework to integrate the post contents, user-tagging habits and post-tag relationships into the recommendation model. The proposed framework contains four phases. First, we extract the visual and textual features from each post. Second, we devise the post-tag attention mechanism to acquire user tagging habits and post-tag relationships through similar posts. Third, we develop the sextuplet co-attention mechanism to derive the representation of each post. Finally, we adopt the relational graph convolution network (R-GCN) (Schlichtkrull et al. 2017) to construct the interaction graph between posts and update the post representations by propagating information among neighboring nodes. Then, we use the convoluted post representations to recommend hashtags for users.

The contributions of this study are summarized as follows.

- 
- A novel sextuplet co-attention mechanism is developed for integrating the features of image, text, user tagging habits, and post-tag relationships to derive the post representation.
 - Based on the Differentiable Neural Computer (DNS) model (Graves et al. 2016), a memory network is proposed for aggregating visual features from multiple aspects while deriving the post representations.
 - A new graph convolution module is designed for generating node (post) representations by aggregating the structural information of neighboring nodes.

The rest of this thesis is organized as follows. In Chapter 2, we conduct a survey of the relevant literature, followed by a comprehensive exploration of our proposed framework in Chapter 3. Moving forward, Chapter 4 showcases the experimental results, while Chapter 5 encapsulates the conclusion and outlines future avenues for work.

Chapter 2 Related Work

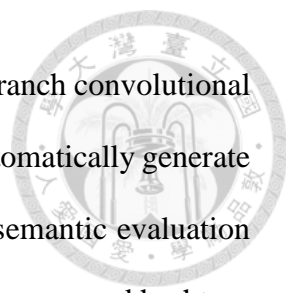


Our study aligns with three distinct research streams: hashtag recommendation, memory network, and attention mechanism. Subsequently, we will delve into a literature survey focused on these three research streams in the upcoming sections.

2.1 Hashtag Recommendation

Hashtag recommendation methods can be classified into two categories: content-based method, and personalized method (Zhang et al. 2019). The content-based methods suggest post hashtags based on content, while the personalized methods consider both post content and user tagging habits for recommending hashtags.

For the content-based methods, Ding et al. (2013) proposed a topical translation model to recommend hashtags for microblogs, where the posts may be written in different languages. Sedhai and Sun (2014) used the term frequency–inverse document frequency (TF-IDF) weighting scheme to compare the similarity between posts for recommending hashtags for a post. Gong and Zhang (2016) adopted the convolutional neural networks (CNNs) (Lecun et al. 1998) to encode sentences in social media posts. Wu et al. (2018) applied CNN to encode images and used Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) to capture the temporal relationships among hashtags. Yang et al. (2016) proposed an image question-answering model, which could also be used for hashtag recommendations by treating posting images as the question and target hashtag as the answer. Zhang et al. (2017) and Gong et al. (2018) incorporated both textual and visual information to recommend hashtags. Yang et al. (2020) extracted the features of texts and images and incorporated them into the sequence-to-sequence model (Sutskever et al. 2014) to generate the recommended hashtags. Połap (2023) combined the U-NET



model (Ronneberger et al. 2015) for image segmentation, the multi-branch convolutional neural networks and the skip-gram model (Mikolov et al. 2013), to automatically generate hashtags for social media images. Alsini et al. (2023) introduced a semantic evaluation approach for hashtag recommendation. The content-based methods recommend hashtags for a post based on the post content; however, they do not consider user tagging habits and the post-tag relationships, i.e., similar posts tend to have similar hashtags.

For the personalized methods, Denton et al. (2015) combined user metadata and image features derived by CNN to predict image hashtags. Huang et al. (2016) recommended hashtags for a post by using the features extracted from the post's text content and user tagging habits learned by the external memory network (Weston et al. 2015). Ma et al. (2019) and Zhang et al. (2019) employed an external memory network to understand user tagging habits. They leveraged textual and visual features extracted from posts, along with user tagging habits, to provide hashtag recommendations. Javari et al. (2020) proposed a personalized hashtag recommendation model by building a graph-based profile of users to learn user interest from their links towards hub nodes. Chen et al. (2022) introduced the triplet co-attention mechanism to capture the interplay among text, image, and user interactions, used the Graph Convolutional Network (GCN) (Schlichtkrull et al. 2017) to represent the relationships between posts, and then recommended hashtags based on the node features derived by the GCN.

Many previous methods (Ding et al. 2013, Sedhai and Sun 2014, Gong and Zhang 2016, Yang et al. 2016, Zhang et al. 2017, Gong et al. 2018, Wu et al. 2018, Yang et al. 2020, Alsini et al. 2023, Połap 2023) consider post contents but do not consider user tagging habits and post-tag relationships. Some methods (Denton et al. 2015, Huang et al. 2016, Ma et al. 2019, Zhang et al. 2019, Javari et al. 2020, Chen et al. 2022) consider post contents and user tagging habits but do not consider the post-tag relationships. Therefore,

we propose a novel hashtag recommendation model by integrating the post contents, user-tagging habits and post-tag relationships to derive the post representations, and adopting the relational graph convolution network (R-GCN) (Schlichtkrull et al. 2017) for recommending hashtags for user post.

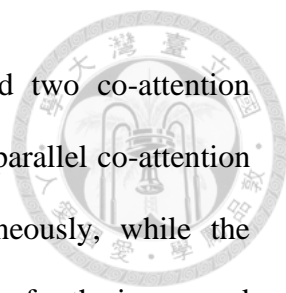
2.2 Memory Network

Graves et al. (2014) proposed the architecture of Neural Turing Machine, which used neural networks to control what values were written to which memory blocks or read from which memory blocks. Later on, Graves et al. (2016) introduced a machine learning model called a differentiable neural computer (DNC). The DNC memory is selectively writable and readable, and can be iteratively modified the memory contents. Also, Sukhbaatar et al. (2015) proposed an end-to-end memory network that required less supervised data for training. Miller et al. (2016) introduced a key-value memory network, which made reading documents more viable by utilizing different encodings in the addressing and output stages of the memory read operation.

To utilize memory network in hashtag recommendation, Zhang et al. (2020) proposed a method by using two coupled memory networks to enhance the expressiveness of post embeddings. Peng et al. (2019) used a memory network to consolidate the relationships between post contents and hashtags in historical posts. Different from the previous studies (Huang et al., Ma et al. 2019, Peng et al. 2019, Zhang et al. 2019, 2020) which use the memory network to characterize user tagging habits, our proposed framework adopts the DNC memory network to learn visual features from posts.

2.3 Attention Mechanism

Attention mechanism can help a model discriminate and put focus on important



information (Bahdanau et al. 2016). Lu et al. (2017) introduced two co-attention mechanisms: parallel co-attention and alternating co-attention. The parallel co-attention mechanism attends to image regions and the question simultaneously, while the alternating co-attention mechanism generates attentive representations for the image and question sequentially. Also, they showed that the parallel co-attention mechanism has better performance than the alternating co-attention mechanism, but was more difficult to train. For hashtag recommendation, Zhang et al. (2019) used parallel co-attention mechanism to model the correlation between image and text contents in a post.

Besides the parallel co-attention model, Zhang et al. (2017) proposed a co-attention network to incorporate tweet-guided visual attention and image-guided textual attention. Inspired by the studies (Zhang et al. 2017, Zhang et al. 2019), Chen et al. (2022) proposed a triplet co-attention mechanism to fuse multimodal information such as image, text and user interest. For each post, the triplet co-attention framework first calculated user-image, image-text, and text-user co-influential vectors, and then summed up these vectors to derive the post representation.

Motivated by these previous studies, we incorporate two attention mechanisms into our framework. First, based on the parallel co-attention mechanism, we present the post-tag attention mechanism to model the interactions among posts and frequently used hashtags. Next, inspired by the triplet co-attention mechanism (Chen et al. 2022), we develop the sextuplet co-attention mechanisms to consider the interactions among the visual features, textual features, user tagging habits, and post-tag relationships.

Chapter 3 The Proposed Framework

Given a collection of posts p_1, p_2, \dots, p_n , we propose a Personalized Hashtag Recommendation model, called PHR, to recommend the hashtags for a new post p made by user u , where each post p_i contains multiple hashtags and the owner of the post, n

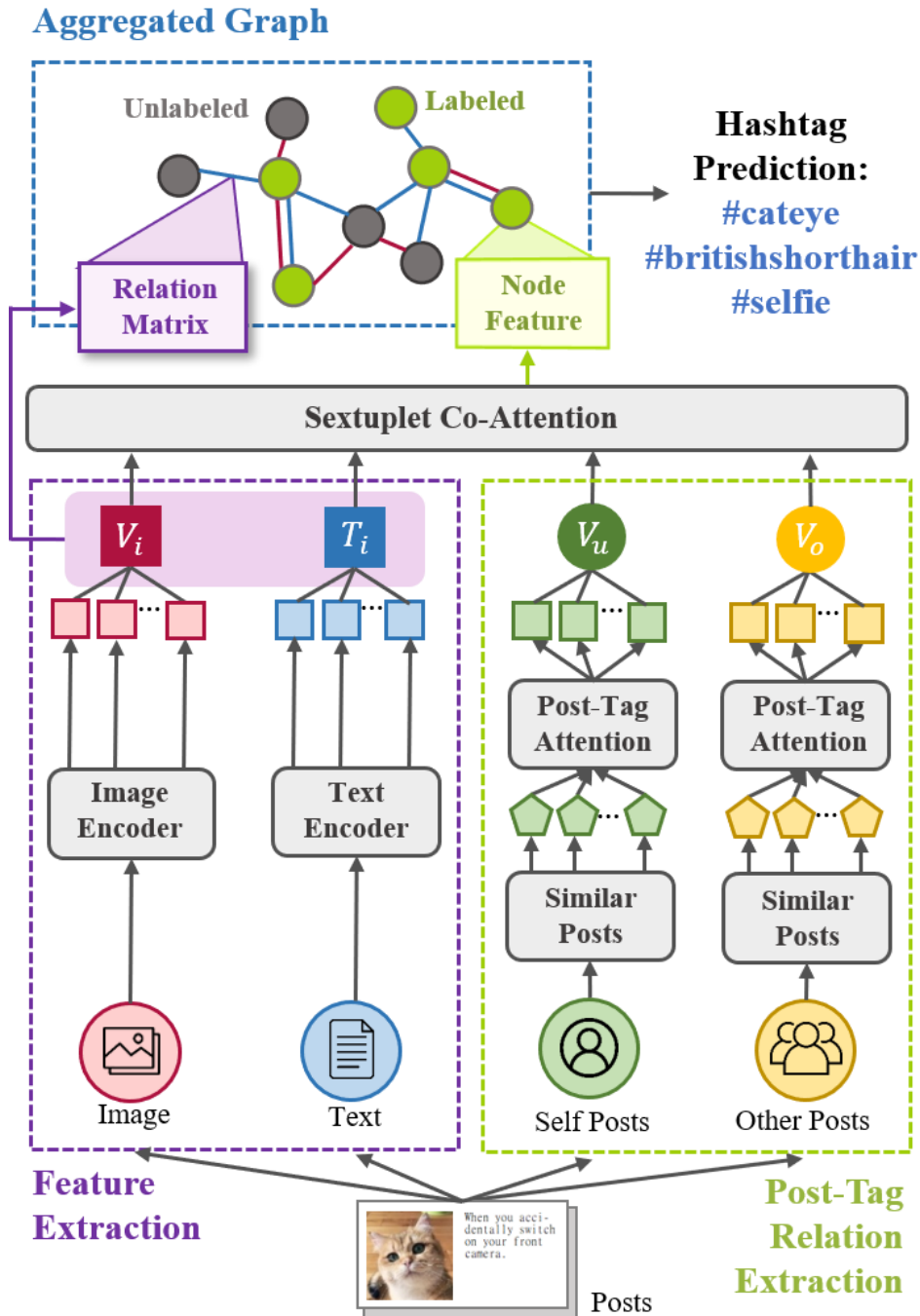


Fig. 2. The Proposed Model

denotes the number of posts in the collection, $1 \leq i \leq n$. The proposed model contains four phases as shown in Fig. 2.

First, we employ the image and text encoders to extract the visual and textual features for each post. Second, we devise the post-tag attention mechanism to acquire user tagging habits and post-tag relationships through similar posts. Third, we develop the sextuplet co-attention mechanism to derive the representation of each post. Finally, we adopt the Relational Graph Neural Network (R-GCN) to construct the interaction graph between posts and update the post representations by propagating information among neighboring nodes. Then, we use the convoluted post representation to recommend hashtags for the new post p .

3.1 Feature Extraction

In our approach, we utilize the image encoder to extract visual features and the text encoder to extract textual features for each post. The detailed descriptions of both encoders will be provided in the following subsections.

3.1.1 Image Encoder

Inspired by Graves et al. (2016), we exploit the memory network to develop the image encoder for extracting the visual features from the image in each post. Since hashtags are sometimes highly associated with some objects in the image, we first use the pre-trained ResNet-50 model (He et al. 2015) to retrieve the visual proposals from the image, each of which is compiled by a sequence of convolution and pooling operations, and extracts the features of a certain region in the image. For example, a visual proposal could represent the color, shape, or textual distribution of a certain object such as a cat. Next, we utilize the memory network to consolidate the extracted visual proposals from

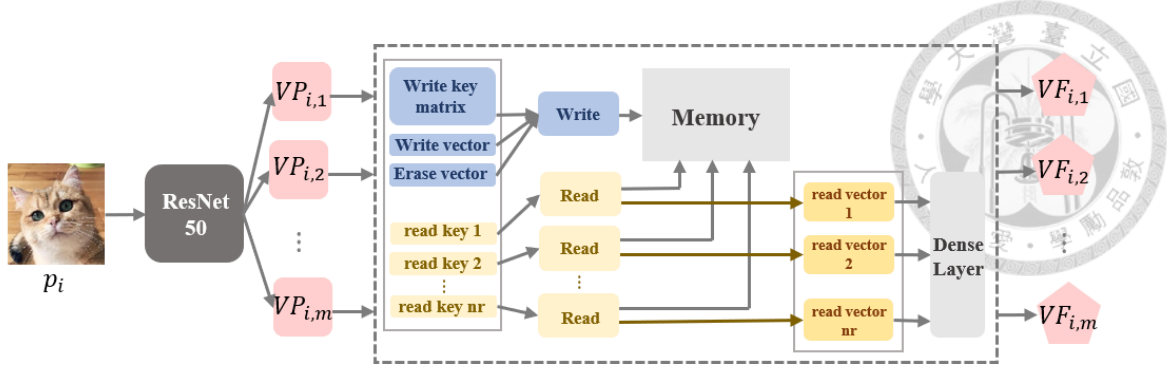


Fig. 3. Memory Network

multiple aspects such as color, shape, or textual, for better learning the visual features from a post for hashtag recommendation.

Specifically, for each post p_i , the image encoder employs the pre-trained ResNet-50 model to extract m visual proposals for the image in the post by Eq. (1), where $VP_i = \{VP_{i,1}, VP_{i,2}, \dots, VP_{i,m}\}$ contains all the visual proposals retrieved from p_i , $VP_{i,j} \in \mathbb{R}^a$ denotes the j th visual proposal retrieved from p_i , m denotes the number of the visual proposals, a denotes the dimensionality of $VP_{i,j}$, $1 \leq j \leq m$.

$$VP_i = \text{ResNet}(p_i) \quad (1)$$

Next, the image encoder exploits the memory network shown in Fig. 3 to consolidate the visual features from multiple aspects by Eq. (2), where Mem denotes the memory network, $V_{i,j} \in \mathbb{R}^f$ denotes the visual feature vector of $VP_{i,j}$, and f denotes the dimensionality of $V_{i,j}$. By consolidating these visual features, our model can better learn the visual features from each post for hashtag recommendation. The memory network has m memories, $M_1, M_2, \dots, M_m \in \mathbb{R}^{s \times f}$, where each memory contains s slots. Each slot is used to implicitly learn the feature from a certain aspect. The memory network contains two modules, read and write. For each proposal $VP_{i,j}$, the image encoder performs the write module once and then the read module once on M_j to generate its visual feature vector.

$$V_{i,j} = \text{Mem}(M_j, VP_{i,j}) \quad (2)$$

For each $VP_{i,j}$, the image encoder computes the write key matrix $K_{i,j}^w \in \mathbb{R}^{s \times f}$ by Eq. (3), write content vector $CV_{i,j}^w \in \mathbb{R}^f$ by Eq. (4), and the erase vector $EV_{i,j} \in \mathbb{R}^f$ by Eq. (5), where $\cos(\cdot)$ denotes the cosine similarity between both arguments, ϕ denotes a dense layer, and σ denotes the sigmoid function. The write key matrix stores the information learned from the j th proposal of p_i . The write content vector $CV_{i,j}^w$ contains the information that will be written into memory. The erase vector $EV_{i,j}$ decides which information would be erased.

$$\cos(K_{i,j,x}^w, VP_{i,j}) = \frac{K_{i,j,x}^w (VP_{i,j})^T}{|K_{i,j,x}^w|^2 |VP_{i,j}|^2} \quad (3)$$

$$CV_{i,j}^w = \phi(VP_{i,j}) \quad (4)$$

$$EV_{i,j} = \sigma(\phi(VP_{i,j})) \quad (5)$$

Then, the image encoder computes the write weight $W_{i,j}^w \in \mathbb{R}^s$ for the memory M_j by Eq. (6), where $W_{i,j,l}^w \in \mathbb{R}^+$ represents the write weight of the l th slot in M_j , and c represents the strength parameter. This write weight is crucial for ensuring that similar proposals have similar weights. Subsequently, we update the memory matrix M_j by Eq. (7), where $W_{i,j}^w$ represents the write weight of M_j , and $\mathbf{1}$ denotes the matrix of ones.

$$W_{i,j,l}^w = \frac{\exp(c \cdot \cos(K_{i,j,l}^w, VP_{i,j}))}{\sum_{x=1}^s \exp(c \cdot \cos(K_{i,j,x}^w, VP_{i,j}))} \quad (6)$$

$$M_j = M_j \times (\mathbf{1} - W_{i,j}^w (EV_{i,j})^\top) + W_{i,j}^w (CV_{i,j}^w)^\top \quad (7)$$

Following that, the image encoder executes the read module to extract salient visual features from memory, obtaining the visual feature vector of p_i . Initially, it utilizes a dense layer to generate nr read keys by projecting $VP_{i,j}$ nr times by Eq. (8). For each read key $K_{i,j,d}^r \in \mathbb{R}^f$, it calculates its read weight $W_{i,j,d,x}^r \in \mathbb{R}^+$ of the d th read head for the memory slot $M_{j,x}$ through Eq. (9).

$$K_{i,j,d}^r = \phi(VP_{i,j}) \quad (8)$$

$$W_{i,j,d,l}^r = \frac{\exp(\cos(K_{i,j,d}^r, M_{j,l}))}{\sum_{x=1}^{nr} \exp(\cos(K_{i,j,d}^r, M_{j,x}))} \quad (9)$$

Subsequently, the image encoder generates the readout vector $O_{i,j,d}^r \in \mathbb{R}^c$ by Eq. (10), comprising the information read from memory. Here, $O_{i,j,d}^r$ represents the readout vector of the d th read head, and $W_{i,j,d}^r \in \mathbb{R}^s$ represents the read weights of the d th read head of M_j . Afterwards, it derives the visual feature vector $V_{i,j} \in \mathbb{R}^f$ for $VP_{i,j}$ by Eq. (11). The image encoder performs Eqs. (3)-(11) repeatedly for each $VP_{i,j}$, $j = 1, 2, \dots, m$. That is, the image encoder derives the visual feature vector of each proposal for each post.

$$O_{i,j,d}^r = M_j^\top W_{i,j,d}^r \quad (10)$$

$$V_{i,j} = \phi([O_{i,j,1}^r, O_{i,j,2}^r, \dots, O_{i,j,d}^r]) \quad (11)$$

3.1.2 Text Encoder

The text encoder converts each word token in p_i into a one-hot encoding vector and uses an embedding layer transform it into a word embedding vector. Next, the text encoder employs the Bidirectional Long Short-Term Memory Network (BiLSTM) (Graves and Schmidhuber 2005) to derive the textual features from the word embedding vectors of p_i , where $T_i = \{T_{i,1}, T_{i,2}, \dots, T_{i,m}\}$.

$$T_i = BiLSTM(p_i) \quad (12)$$

3.2 Post-Tag Attention Mechanism

User habit refers to the tendency of users to employ the same tags for a post as those used in similar historical posts created by themselves (Zhang et al. 2019). Thus, we use the post-tag attention mechanism to acquire the user habits. Let C_u contain all the posts made by u . We select top k posts from C_u , where the selected posts are most similar to p_i . The similarity between posts p_i and p_j is computed by Eq. (13), where ω is a

hyperparameter used to decide the proportion of visual similarity, $0 \leq \omega \leq 1$. For each top k post, we employ the pretrained Bidirectional Encoder Representations from Transformers model (BERT) (Devlin et al. 2019) to derive its embedding vector. Let $EM_u \in \mathbb{R}^{k \times g}$ denotes the embedding vectors of the top k posts, where g denotes the dimensionality of a BERT embedding vector. Afterwards, we find the top γ most frequently used hashtags in the top k posts and use BERT to convert each hashtag found into an embedding vector. Let $EH_u \in \mathbb{R}^{\gamma \times g}$ denotes the embedding vectors of the top γ hashtags.

$$sim(p_i, p_j) = \omega \cdot cos(V_i, V_j) + (1 - \omega)cos(T_i, T_j) \quad (13)$$

Next, we use the post-tag attention mechanism to obtain the post embedding $V_u \in \mathbb{R}^g$ by considering the correlations between posts and tags by Eq. (14), where $W_H \in \mathbb{R}^{g \times g}$, $W_E \in \mathbb{R}^{g \times g}$, $W_{H'} \in \mathbb{R}^{g \times g}$, and $W_F \in \mathbb{R}^{1 \times g}$ are learnable parameter matrices, and $b_F \in \mathbb{R}^k$ is the bias. We first compute the correlation matrix $Q \in \mathbb{R}^{k \times \gamma}$ which records the correlations between posts and hashtags. Next, we define the new post features $F \in \mathbb{R}^{g \times k}$ by using the hashtag embedding vectors to guide the attention learning of posts. Finally, we compute the attention weights $\pi \in \mathbb{R}^k$ and derive the attentive user-post vector $V_u \in \mathbb{R}^g$. That is, we utilize the attentive user-post vector to learn the correlations between text contents and hashtags in the top k posts for better recommending hashtags for the posts made by u .

$$\begin{aligned} Q &= \tanh((EM_u)W_H(EH_u)^T) \\ F &= \tanh(W_E(EM_u)^T + (W_{H'}(EH_u)^T)Q^T) \\ \pi &= \text{Softmax}(W_F F + b_F) \\ V_u &= \sum_{j=1}^k \pi_j (EM_{u,j}) \end{aligned} \quad (14)$$

In addition, users tend to use the same tags for a post as those of similar posts made by the other users. Let C_o contains all the posts not made by u . We select the top μ

posts most similar to p_i from C_o , find the top ν most frequently used hashtags from the top μ posts, and use BERT to convert each hashtag found into an embedding vector. Then, we use the post-tag attention mechanism to derive attentive other-post vector $V_o \in \mathbb{R}^g$. That is, we utilize the attentive other-post vector to learn the correlations between text contents and hashtags in the top μ posts for better recommending hashtags for the posts made by u .

3.3 Sextuplet Co-attention Mechanism

Inspired by Chen et al. (2022), we develop the sextuplet co-attention mechanism to derive the post feature vector of each post as shown in Fig. 4. We use the sextuplet co-attention mechanism to learn the pairwise correlations of V_i , T_i , V_u , and V_o . Specifically, we first exploit the co-attention mechanism to fuse the information of visual and textual features (V_i and T_i) within a post by Eqs. (15)-(16). We first use a single-layer tanh function to integrate the vectors V_i and T_i , get the attention probability distribution by a softmax layer, and generate the integrated textual representation $IT_i^{T,V} \in \mathbb{R}^{f \times m}$ by computing the weighted-sum of the integrated visual and textual features of each proposal j by Eq. (16), where $R^{T,V} \in \mathbb{R}^{mf \times m}$ denotes the integrated visual and textual features, $\pi^{T,V} \in \mathbb{R}^m$ denotes attention probability of each proposal, $W_T^{T,V} \in \mathbb{R}^{mf \times f}$, $W_V^{T,V} \in \mathbb{R}^{mf \times f}$ and $W_R^{T,V} \in \mathbb{R}^{1 \times mf}$ are learnable parameter matrices, and $b^{T,V} \in \mathbb{R}^m$ is bias.

$$\begin{aligned}
 R^{T,V} &= \tanh(W_T^{T,V} T_i + W_V^{T,V} V_i) \\
 \pi^{T,V} &= \text{Softmax}(W_R^{T,V} R^{T,V} + b^{T,V}) \\
 IT_i^{T,V} &= \sum_{j=1}^m \pi_j^{T,V} R_j^{T,V}
 \end{aligned} \tag{15}$$

To consider the mutual influence between visual and textual features, we use the V_i and integrated vector $IT_i^{T,V}$ to generate the combined vector $R^{V,T} \in \mathbb{R}^{mf \times m}$, the attention probability distribution $\pi^{V,T} \in \mathbb{R}^m$, and the integrated visual representation

$IV_i^{T,V} \in \mathbb{R}^{f \times m}$ by Eq. (16), where $W_T^{V,T} \in \mathbb{R}^{mf \times f}$, $W_V^{V,T} \in \mathbb{R}^{mf \times f}$, and $W_R^{V,T} \in \mathbb{R}^{1 \times mf}$ are learnable parameter matrices, and $b^{V,T} \in \mathbb{R}^m$ is the bias. Then, we sum up $IT_i^{T,V}$ and $IV_i^{T,V}$ to get the co-influential vector $V_i^{TV} \in \mathbb{R}^{f \times m}$.

$$\begin{aligned} R^{V,T} &= \tanh(W_V^{V,T} V_i + W_T^{V,T} IT_i^{T,V}) \\ \pi^{V,T} &= \text{Softmax}(W_R^{V,T} R^{V,T} + b^{V,T}) \\ IV_i^{T,V} &= \sum_{j=1}^m \pi_j^{V,T} R_j^{V,T} \end{aligned} \quad (16)$$

For V_u and V_o , we first divide it into f segments and convert each segment into m dimensions to derive $V_u' \in \mathbb{R}^{f \times m}$ and $V_o' \in \mathbb{R}^{f \times m}$. Similarly, we apply the co-attention mechanism in Eqs. (15)-(16) to fuse V_i and V_u' to obtain co-influential vector V_i^{UV} , V_i and V_o' to obtain co-influential vector V_i^{OV} , T_i and V_u' to obtain co-influential vector V_i^{UT} , T_i and V_o' to obtain co-influential vector V_i^{OT} , and V_u' and V_o' to obtain co-influential vector V_i^{OU} . Finally, we get the post feature vector r_i of post p_i by summing up all the co-influential vectors.

3.4 Graph Convolution and Hashtag Recommendation

Posts with similar text and image contents often have similar hashtags. Therefore, we create the graph to consider the relations between posts with respect to their integrated visual and textual representations, where each node denotes a post feature vector. Two nodes are connected by a red edge if the cosine similarity between their visual feature vectors is high enough, and by a blue edge if the cosine similarity between their textual feature vectors is high enough. Specifically, the adjacency matrix A^{red} for red edges is formulated by Eq. (17), where $\text{sim}_{i,j}^V$ denotes the cosine similarity between the visual feature vectors of the i th and j th posts. Similarly, the adjacency matrix A^{blue} for blue edges is formulated by Eq. (18), where $\text{sim}_{i,j}^T$ denotes the cosine similarity between the textual feature vectors of the i th and j th posts. The thresholds θ and λ are

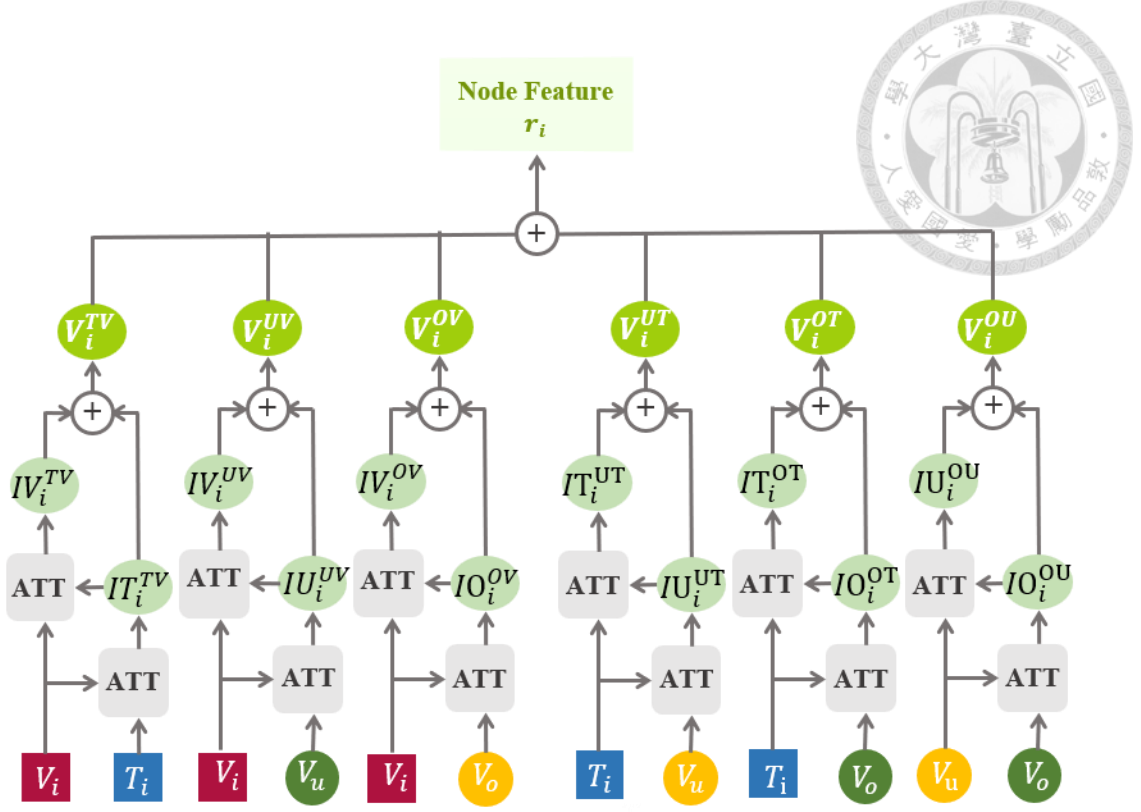


Fig. 4. Sextuplet Co-Attention Mechanism

set to filter out the edges with low similarity scores.

$$A_{i,j}^{red} = \begin{cases} sim_{i,j}^V, & \text{if } sim_{i,j}^V \geq \theta \\ 0, & \text{if } sim_{i,j}^V < \theta \end{cases} \quad (17)$$

$$A_{i,j}^{blue} = \begin{cases} sim_{i,j}^T, & \text{if } sim_{i,j}^T \geq \lambda \\ 0, & \text{if } sim_{i,j}^T < \lambda \end{cases} \quad (18)$$

We adopt the Relational Graph Neural Network (R-GCN) (Schlichtkrull et al. 2017) to construct the interaction graph between posts and update the node features by propagating information among neighboring nodes in Eq. (19), where β denotes a neighboring node of node r_i with respect to the red (blue) edges, r_β denotes the post representation of node β , W_γ and W_0 are learnable parameter matrices, $\tau_{i,\alpha,\beta}$ is a normalization constant, and N_{r_i} contains the neighboring nodes of node r_i .

$$r_i' = \sigma \left(\sum_{\alpha \in \{red, blue\}} \sum_{\beta \in N_{r_i}} \frac{1}{\tau_{i,\alpha,\beta}} W_\gamma A_{i,\beta}^\alpha r_\beta + W_0 r_i \right) \quad (19)$$

After propagating information over the graph, we derive the convoluted post

representation for each post. Next, we use the multi-layer perceptron (MLP) to predict the hashtags for each post by Eq. (20).

$$z_i = MLP(r'_i) \quad (20)$$

The loss function is defined in Eq. (21), where $y_{i,t}$ denotes whether hashtag t is in the post p_i or not, e is the number of posts, and t denotes the number of unique hashtags.

$$L_{loss} = \frac{1}{et} \sum_{i=1}^e \sum_{j=1}^t -\log(z_{i,j}) * y_{i,j} \quad (21)$$

Chapter 4 Experimental Results



In this chapter, we evaluate the performance of our proposed framework. The assessment includes the following sections: the dataset and experimental setup in Section 4.1, performance evaluation in Section 4.2, performance of each variant model in Section 4.3, effect of each attention mechanism in Section 4.4, and recommendation examples in Section 4.5.

4.1 Dataset and Experiment Setup

We sample a subset from the MaCon dataset (Zhang et al. 2019), which contains a collection of posts curated from Instagram users. The subset is strategically selected to align with our computational constraints and research objectives, comprising 20,000 posts out of the original dataset which contains 624,520 posts. This subset encapsulates a diverse range of 3,832 unique hashtags and is characterized by a lexical diversity with 58,435 distinct words. Our subset represents a microcosm of the social media landscape, featuring posts from 3,600 distinct users. Each post contains 8.722 hashtags and 59.465 words on average. Table 1 lists the statistics of the dataset.

In our experimental setup, we divide the dataset into 90% for training and 10% for testing. We set the batch size to 500. The training leverages a learning rate of 0.05, using Stochastic Gradient Descent (SGD) as the optimizer, and is conducted on the TensorFlow platform. This configuration aims to effectively evaluate and enhance the performance of our hashtag recommender system. The remaining hyperparameter configurations are detailed in Table 6 in the Appendix.

Table 1. Statistics of the Dataset

#Posts	#Users	#Hashtags	#Words	#Avg_tag	#Avg_word
20,000	3,600	3,822	58,435	8.722	59.465



4.2 Performance Evaluation

We conduct a comparison of our model with two state-of-the-art methods, MaCon (Chen et al. 2022) and TAGNet (Chen et al. 2022). Fig. 5 presents the performance of our and comparing methods. The performance metrics are evaluated in terms of hit rate, precision, recall, and F1 score for recommending the top k hashtags for users, where $k=1, 3, 5, 7$, and 9 . Table 2 shows the performance of recommending top 9 hashtags (HitRate@9, Precision@9, Recall@9, F1@9).

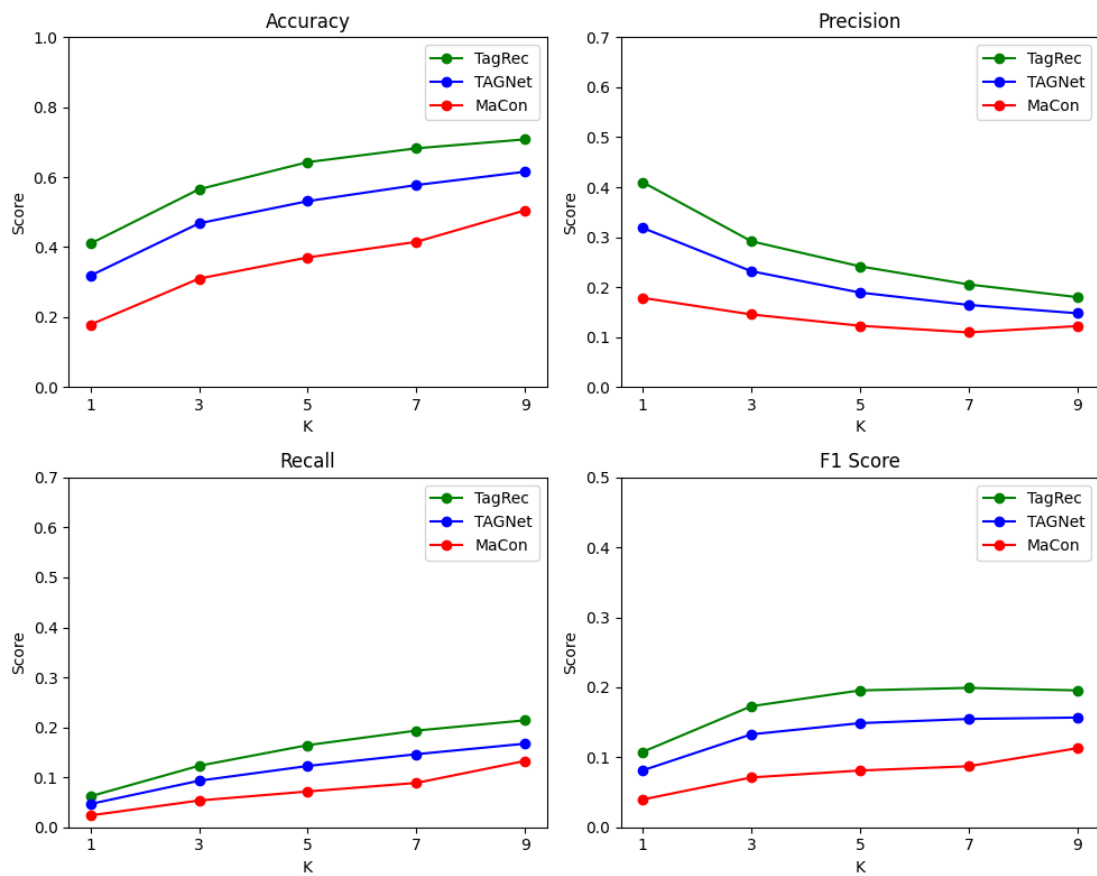


Fig. 5. Performance of Our and Comparing Models

TAGNet employs the triplet attention mechanism to learn the correlations between visual, textual, and user features for a comprehensive understanding of how different types of data interact and complement each other. Thus, it performs better than MaCon. Our proposed model uses the memory network to learn visual representation of a post,

which enables our model to learn the representations from multiple aspects. This multifaceted representation captures a more comprehensive and nuanced understanding of the content, leading to more accurate predictions. Also, we introduce the sextuplet co-attention mechanism and the post-tag attention mechanism in learning correlations between features. These mechanisms help our model learn intricate interactions between features, significantly enhancing the recommendation performance of our model. Therefore, our proposed model outperforms MaCon and TAGNet across all metrics.

4.3 Analysis of the Variant Models

Table 3 illustrates the performance the variants of our model. PHR-MN is the variant of our model without using the component, memory network. PHR-PTA is the variant of our model without using the component, post-tag attention mechanism. PHR-SCA is the variant of our model without using the component, sextuplet co-attention mechanism.

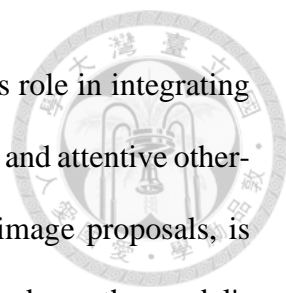
The sextuplet co-attention mechanism is shown to be most crucial, as its absence

Table 2. Performance of Recommending Top 9 Hashtags

Model	HitRate@9	Precision@9	Recall@9	F1@9
MaCon	0.5205	0.1321	0.1473	0.1393
TAGNet	0.6160	0.1474	0.1676	0.1569
PHR	0.6635	0.1719	0.1959	0.1831

Table 3. Performance of Variant Models

Model	HitRate@9	Precision@9	Recall@9	F1@9
PHR-MN	0.6345	0.1632	0.1828	0.1724
PHR-PTA	0.6390	0.1662	0.1863	0.1757
PHR-SCA	0.4350	0.1015	0.1020	0.1017
PHR	0.6635	0.1719	0.1959	0.1831



leads to a significant drop in all performance metrics, highlighting its role in integrating various types of features, including visual, textual, attentive user-post, and attentive other-post vectors. The memory network, which incorporates additional image proposals, is also vital for learning from visual contents; its absence noticeably reduces the model's performance. While the post-tag attention mechanism contributes to refining the model's performance, it is least crucial compared to other components. Removing it leads to a modest decline in performance, primarily because it mainly influences text processing. This suggests that text contents, while significant, have a lesser impact on hashtag recommendation than visual contents. Overall, each component plays a significant role in enhancing model performance, with the sextuplet co-attention mechanism being the most influential in hashtag recommendation.

Table 4 shows the impact of each type of features. PHR-V denotes our model without using visual feature vectors. PHR-T denotes our model without using textual feature vectors. PHR-U denotes our model without using attentive user-post vectors. PHR-O denotes our model without using attentive other-post vectors.

The visual features emerge as the most critical factor. Given that our dataset originates from Instagram, a platform predominantly focused on visual contents, it is evident that visual features significantly influence the model's performance. Users typically select hashtags that align closely with the image content. In contrast, among the commonly used hashtags. On the other hand, the attentive other-post vectors have a more modest effect on model performance. It includes data from posts akin to the target post and associated tagging practices. While its influence on performance is less pronounced than that of user posts, the posts made by other users are instrumental in suggesting a broader spectrum of hashtags to the user.

Table 4. Impact of Each Type of Features

Model	HitRate@9	Precision@9	Recall@9	F1@9
PHR-V	0.5575	0.1363	0.1489	0.1423
PHR-T	0.5995	0.1528	0.1673	0.1598
PHR-U	0.5745	0.1458	0.1573	0.1513
PHR-O	0.5965	0.1465	0.1638	0.1547
PHR	0.6635	0.1719	0.1959	0.1831

Table 5. Impact of Visual and Textual Relations

Model	HitRate@9	Precision@9	Recall@9	F1@9
PHR-TR	0.6355	0.1622	0.1813	0.1712
PHR-VR	0.2865	0.0443	0.0468	0.0456
PHR	0.6635	0.1719	0.1959	0.1831

Our model introduces both visual and textual relations to construct the graph. This enhances the diversity of the convolution effects between posts and validates the effectiveness of incorporating text relations in predictions. Table 5 shows the impact of visual and textual relations. PHR-TR denotes the variant of our model without using textual relations, where the edge weights are derived by the similarity between visual features. PHR-VR denotes the variant of our model without using visual relations, where the edge weights are derived by the similarity between textual features. We can observe that the addition of the textual relations offers a slight improvement to the model. However, overall, the importance of visual relations remains higher than that of text.

4.4 Effects of Attention Mechanisms

Fig. 6 shows the heatmap of the post-tag attention mechanism, where each row denotes a post made by user 11016, each column denotes a hashtag, and the color of each cell indicates the attention weight. The darker the color is, the higher the weight is.

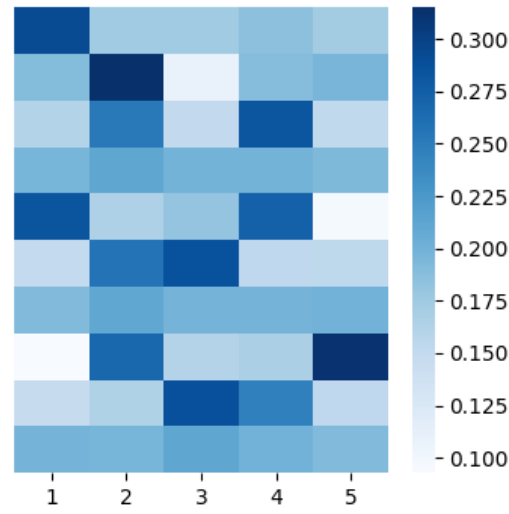


Fig. 6. Heatmap of the Post-Tag Attention Mechanism

Fig. 7 illustrates three posts made by user 11016. These three posts demonstrate that user 11016 often use dog-related words such as puppy and dogs, and dog-related hashtags such as #puppiesofinstagram and #puppylove. The dog-related hashtags will have higher weights (dark colors in Fig. 7). On the other hand, the hashtags like #powder, #repost, and #climbinglife, which are commonly-used hashtags but not related to the text contents of the posts. Thus, these hashtags will have less weights (light colors in Fig. 7). Therefore, the post-tag attention mechanism can help our model to recommend the hashtags related to the target post.

Fig. 8 shows the heatmaps of the sextuplet co-attention mechanism for visual and textual features, where each row denotes a post, and each column denotes a feature vector. The first five posts are made by user 111, the middle five by 2058, and the last five by 3378. Fig. 8 (a) presents the attention probability distribution $\pi^{T,V}$ in Eq (15), while Fig. 8 (a) presents the attention probability distribution $\pi^{V,T}$ in Eq (16). Fig. 9 displays the

posts made by each user.

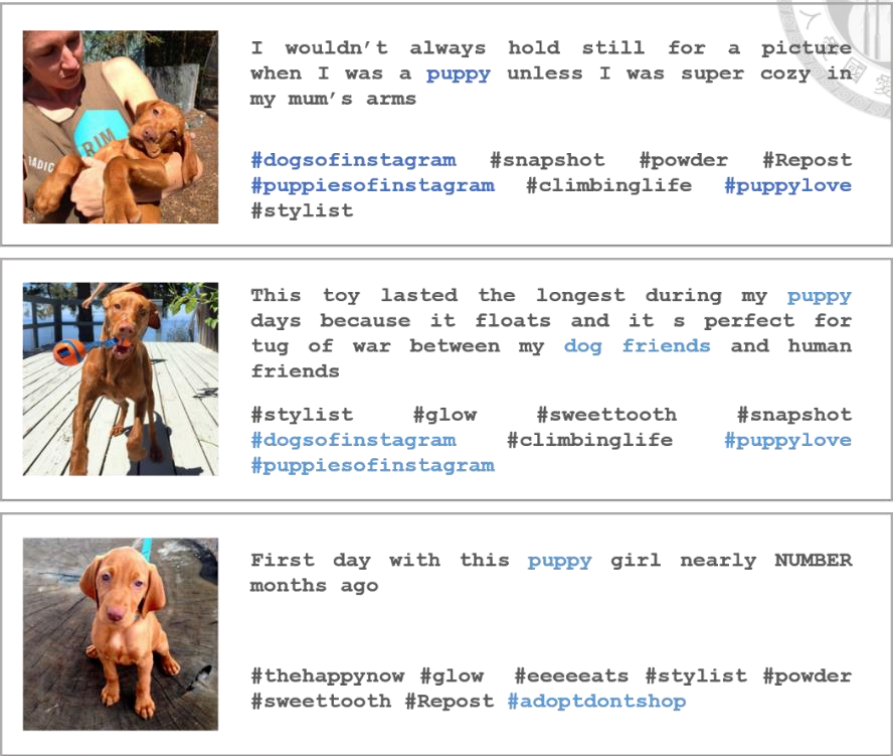


Fig. 7. Posts of User 11016

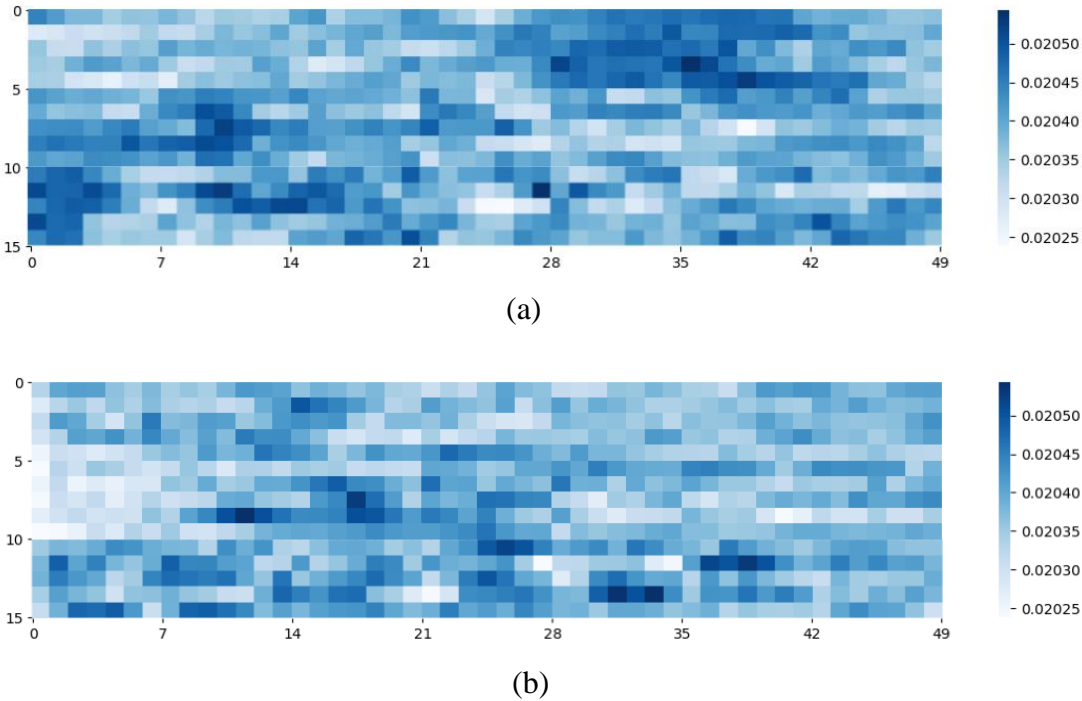
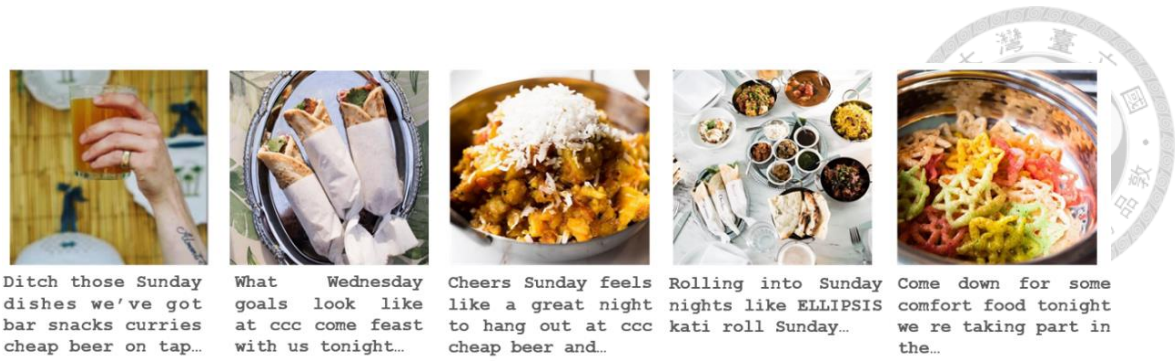


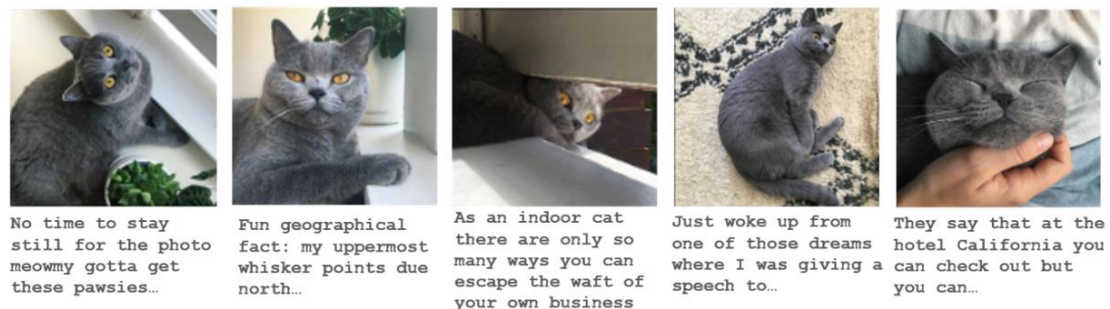
Fig. 8. Heatmaps of the Sextuplet Co-attention Mechanism



(a) User 111



(b) User 2058



(c) User 3378

Fig. 9. Posts of Users 111, 2058 and 3378

For the posts made by user 111, a noticeable concentration of darker areas appears in the 28th to 40th columns, aligning with the thematic focus of their posts in Fig. 8(a). User 2058's heatmap also displays darker areas in the 5th to 11th columns, while user 3378's heatmap reveals darker areas in the 1st to 4th columns. Thus, the sextuplet co-attention mechanism assigns higher weights to specific regions for highlighting significant content variations among users. The concentration areas in the heatmap in Fig.

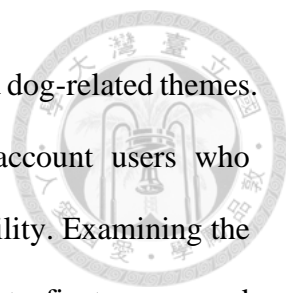
8(b) are more separated compared to those in Fig. 8(a). This separation signifies the nuanced interaction between visual and textual features, emphasizing that the model assigns importance weights to specific combinations of these features.

The sextuplet co-attention mechanism is able to adapt to diverse thematic emphases within user-generated contents as shown in the heatmap in Fig. 8(a), while the complex interplay between visual and textual features is revealed in the heatmap in Fig. 8(b). Therefore, the sextuplet co-attention mechanism can effectively capture and quantify the interactive relationship between images and text.

4.5 Recommendation Examples

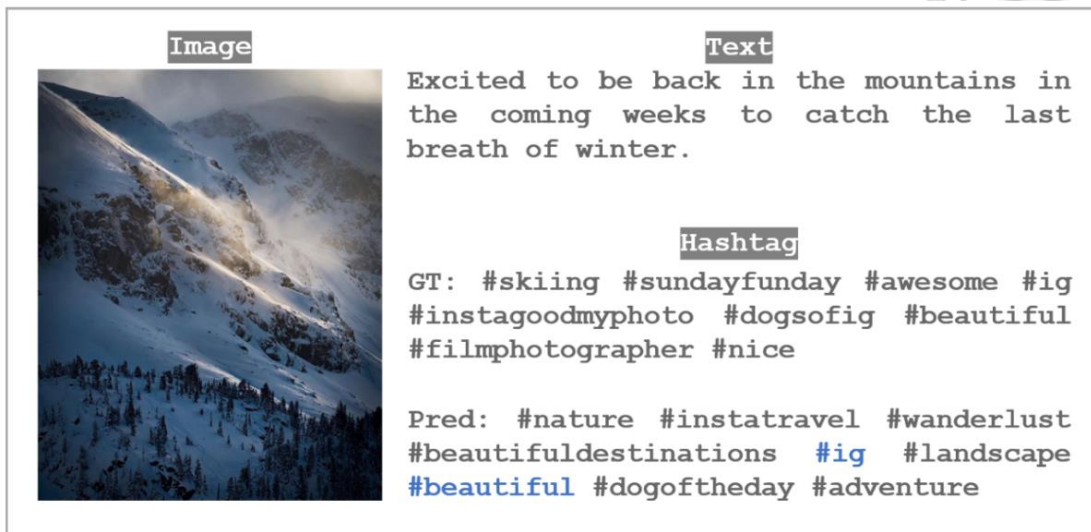
Fig. 10 illustrates two examples of the recommended hashtags for posts made by user 10953, user 11016. **GT** contains the original hashtags made by the user, while **Pred** contains the hashtags recommended by our model. In Fig. 10(a), impressively, our model accurately predicts hashtags such as #ig and #beautiful, which are consistent with the user's previous tagging habits. Furthermore, it suggests relevant tags like #nature and #landscape, derived from the image and text contents. It also proposes popular hashtags like #beautifuldestinations and #wanderlust, commonly used by nature photographers on Instagram, demonstrating our model's ability to adapt to recommend the hashtags frequently used by other users. Notably, the hashtag #dogoftheday, despite its apparent irrelevance to the post content, is included. This reflects a frequently-used tagging strategy where users tag unrelated yet popular hashtags to gain visibility. Thus, our model occasionally recommends such popular tags.

In Fig.10(b), our model makes accurate predictions, successfully identifying hashtags such as #dogphotography and #dogs based on the post content. Additionally, leveraging user tagging habits, our model correctly anticipates the hashtag #cats. Notably,



the hashtags recommended are predominantly centered around cat and dog-related themes. This observation aligns with the common practice among pet account users who strategically employ both dog and cat hashtags to enhance post visibility. Examining the predictions further, our model suggests relevant tags like #catsofinstagram and #bestmeow, capturing the essence of popular cat-related content. Expanding beyond the specific animal categories, the model also outputs #instadog and #pets based on visual content analysis. Furthermore, the inclusion of #Repost is attributed to user tagging habits, showcasing the model's proficiency in recognizing and adapting to these patterns.

The examples showcase our model's competence in not only predicting user-specific hashtags but also suggesting similar tags prevalent in similar posts, content-based hashtags, and popular tags that could enhance post visibility.



(a)



(b)

Fig. 10. Recommendation Examples

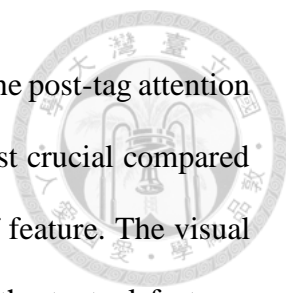
Chapter 5 Conclusions and Future Work



We propose a framework to recommend hashtags for user posts on social media platforms, based on post content and personalized hashtag preference. The proposed framework uses the memory network to memorize the visual proposal features, which helps our model to contain more variety of visual elements. Next, to obtain more information from the historical posting behavior of each user, our proposed framework uses the frequently-used hashtags in the posts similar to the target post, to increase the capability of personalized recommendation. Also, it employs the sextuplet co-attention mechanism to learn the interactions between various types of features. Last, our framework considers both visual and textual relations between posts in Relational Graph Convolution Network (RGCN), which aggregates the neighboring nodes to enhance the post representations to better recommend hashtags for user posts.

Our experimental results demonstrate that our model outperforms the state-of-the-art models and its variant models across all metrics. Our model consistently delivers superior recommendations compared to the two comparing models. This is because that the proposed model uses the memory network to learn visual representation of a post, which enables our model to learn the representations from multiple aspects. Also, our model introduces the sextuplet co-attention and the post-tag attention mechanisms in learning correlations between features. These mechanisms help our model learn intricate interactions between features, significantly enhancing the recommendation performance of our model. Therefore, our proposed model outperforms MaCon and TAGNet across all metrics.

In addition, we conduct the impact of each component on the performance of our model. The results show that the sextuplet co-attention mechanism is most crucial, as its



absence leads to a significant drop in all performance metrics. While the post-tag attention mechanism contributes to refining the model's performance, it is least crucial compared to other components. Also, we analyze the influence of each type of feature. The visual features exhibit the most impact on the model performance while the textual features exhibit the least effect. This is because Instagram is a platform predominantly focused on visual content. Users usually select hashtags that align closely with the image content. Thus, the visual features significantly influence the model performance.

The theoretical implications of our model are multifaceted. Unlike traditional models that rely on single or dual inputs, ours integrates a broad range of inputs such as images, text, user posts, user tags, and other related posts from other users. This comprehensive approach provides a deeper and more nuanced understanding of the post contents and users' tagging preference, leading to more accurate recommendations. Furthermore, we introduce a novel sextuplet co-attention mechanism coupled with a post-tag attention mechanism, facilitating complex interactions among the various features and substantially boosting the model's representational capacity. Another key advancement is the integration of a memory network, which enables the model to remember and utilize similar visual proposals, significantly enhancing the relevance of its recommendations through learning from historical data. Finally, the Relational Graph Convolution Network (RGCN) enables our model to analyze relations from both visual and textual perspectives. This enhances its capacity to capture the diverse range of relationships between posts.

The personalized hashtag recommendation model presented in this study offers significant managerial implications, particularly in the realm of social media content management and digital marketing. Primarily, it streamlines the process of personalized tagging for users, enabling them to quickly label their content. For businesses, this model simplifies and accelerates the task of creating personalized tags. These content-based tag

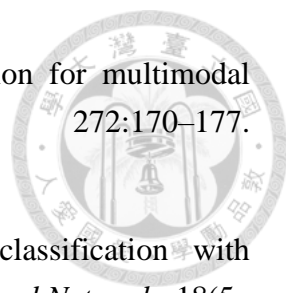
recommendations facilitate user discovery of relevant posts, aiding in effective content navigation and enhancing brand image through tailored tagging. Additionally, widespread adoption of this model on social media platforms can lead to increased tag usage. This, in turn, makes content more accessible to interested users, fostering greater community engagement, enhancing interactions between posts, and expanding the overall reach of the content.

Looking ahead, there are potential extensions for our framework in the following directions. First, incorporating engagement rates and reach into our model could enhance post popularity. Second, we can integrate time series analysis into our model to offer dynamic, time-sensitive recommendations. By extracting the timestamps associated with each post, we can perform time series analysis. This addition empowers our model to identify and respond to trends, delivering timely recommendations that align with ongoing events and discussions in the social media landscape. Third, we can enhance our model by integrating specialized models tailored for specific types of contents. For instance, we could establish a dedicated recommendation model for pet-related posts or a model designed to address situations where the image and text contents do not align. This targeted approach allows our system to provide more accurate and context-aware recommendations, catering to the diverse content categories within social media. Last, we can extend our model to deal with various types of media, including multi-photo posts and short videos. For instance, in the short video type of media, we could get visual features by frames and object detection technique, audio features by audio spectrogram and speech recognition, and append all these features in model, which will allow our model to offer relevant hashtags across different content forms, catering to the diverse and evolving landscape of social media. These advancements aim to make the model more versatile and effective in social media marketing and engagement.

References



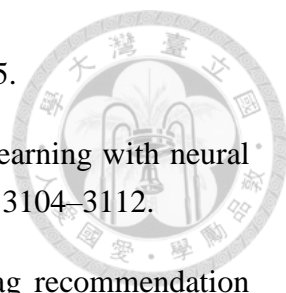
- 【1】 Alsini A, Huynh DQ, Datta A (2023) #REVAL: A semantic evaluation framework for hashtag recommendation. *arXiv:2305.18330*. <https://doi.org/10.48550/arXiv.2305.18330>.
- 【2】 Bahdanau D, Cho K, Bengio Y (2016) Neural machine translation by jointly learning to align and translate. *ArXiv:1409.0473*. <https://doi.org/10.48550/arXiv.1409.0473>.
- 【3】 Chen YC, Lai KT, Liu D, Chen MS (2022) TAGNet: Triplet-attention graph networks for hashtag recommendation. *IEEE Transaction on Circuits and Systems for Video Technology* 32(3):1148–1159. <https://doi.org/10.1109/TCSVT.2021.3074599>.
- 【4】 Cheng L, Leung ACS, Ozawa S eds. (2018) Hashtag recommendation with attention-based neural image hashtagging network. *Proceedings of International Conference on Neural Information Processing*. 52–63. https://doi.org/10.1007/978-3-030-04179-3_5.
- 【5】 Denton E, Weston J, Paluri M, Bourdev L, Fergus R (2015) User conditional hashtag prediction for images. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1731–1740. <https://dl.acm.org/doi/10.1145/2783258.2788576>.
- 【6】 Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4171–4186. <https://doi.org/10.18653/v1/N19-1423>.
- 【7】 Ding Z, Qiu X, Zhang Q, Huang X (2013) Learning topical translation model for microblog hashtag suggestion. *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*. 2078–2084. <https://dl.acm.org/doi/10.5555/2540128.2540427>.
- 【8】 Gong Y, Zhang Q (2016) Hashtag recommendation using attention-based convolutional neural network. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. 2782–2788. <https://dl.acm.org/doi/10.5555/3060832.3061010>.

- 
- 【9】 Gong Y, Zhang Q, Huang X (2018) Hashtag recommendation for multimodal microblog posts. *Neurocomputing* 272:170–177. <https://doi.org/10.1016/j.neucom.2017.06.056>.
- 【10】 Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18(5–6):602–610.
- 【11】 Graves A, Wayne G, Danihelka I (2014) Neural turing machines. *ArXiv:1410.5401*. <https://doi.org/10.48550/arXiv.1410.5401>.
- 【12】 Graves A, Wayne G, Reynolds M, Harley T, Danihelka I, Grabska-Barwińska A, Colmenarejo SG, et al. (2016) Hybrid computing using a neural network with dynamic external memory. *Nature* 538(7626):471–476. <https://doi.org/10.1038/nature20101>.
- 【13】 He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. *ArXiv:1512.03385*. <http://arxiv.org/abs/1512.03385>.
- 【14】 Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Computation* 9(8):1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>.
- 【15】 Huang H, Zhang Q, Gong Y, Huang X (2016) Hashtag recommendation using end-to-end memory networks with hierarchical attention. *Proceedings of the International Conference on Computational Linguistics*. 943–952.
- 【16】 Javari A, He Z, Huang Z, Jeetu R, Chen-Chuan Chang K (2020) Weakly supervised attention for hashtag recommendation using graph data. *Proceedings of The Web Conference*. 1038–1048. <https://doi.org/10.1145/3366423.3380182>.
- 【17】 Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proceedings of the IEEE*. 86(11):2278–2324. <https://doi.org/10.1109/5.726791>.
- 【18】 Lu J, Yang J, Batra D, Parikh D (2016) Hierarchical question-image co-attention for visual question answering. *Proceedings of the International Conference on Neural Information Processing Systems*. 289–297. <https://dl.acm.org/doi/10.5555/3157096.3157129>.
- 【19】 Ma R, Qiu X, Zhang Q, Hu X, Jiang YG, Huang X (2019) Co-attention memory

network for multimodal microblog's hashtag recommendation. *IEEE Transactions on Knowledge and Data Engineering* 33(2):388–400. <https://doi.org/10.1109/TKDE.2019.2932406>.

- 【20】 Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in Vector Space. *ArXiv:1301.3781*. <http://arxiv.org/abs/1301.3781>.
- 【21】 Miller A, Fisch A, Dodge J, Karimi AH, Bordes A, Weston J (2016) Key-value memory networks for directly reading documents. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1400–1409. <https://doi.org/10.18653/v1/D16-1147>.
- 【22】 Peng M, Lin Y, Zeng L, Gui T, Zhang Q (2019) Modeling the long-term post history for personalized hashtag recommendation. *Proceedings of the Conference on Chinese Computational Linguistics*. 495–507. https://doi.org/10.1007/978-3-030-32381-3_40.
- 【23】 Połap D (2023) Hybrid image analysis model for hashtag recommendation through the use of deep learning methods. *Expert Systems with Applications* 229:120566. <https://doi.org/10.1016/j.eswa.2023.120566>.
- 【24】 Rauschnabel PA, Sheldon P, Herzfeldt E (2019) What motivates users to hashtag on social media? *Psychology and Marketing*. 36(5):473–488. <https://doi.org/10.1002/mar.21191>.
- 【25】 Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention*. 234-241. https://doi.org/10.1007/978-3-319-24574-4_28.
- 【26】 Schlichtkrull M, Kipf TN, Bloem P, Berg R van den, Titov I, Welling M (2018) Modeling relational data with graph convolutional networks. *The Semantic Web Lecture Notes in Computer Science*. vol 18043. Springer Science and Business Media LLC, Berlin/Heidelberg, Germany.
- 【27】 Sedhai S, Sun A (2014) Hashtag recommendation for hyperlinked Tweets. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 831–834. <https://doi.org/10.1145/2600428.2609452>.
- 【28】 Sukhbaatar S, Szlam A, Weston J, Fergus R (2015) End-to-end memory networks. *Proceedings of the 28th International Conference on Neural Information Processing*

Systems. 2440–2448. <https://doi.org/10.48550/arXiv.1503.08895>.

- 
- 【29】 Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*. 3104–3112.
- 【30】 Tao H, Khan L, Thuraisingham B (2022) Personalized hashtag recommendation with user-level meta-learning. *Proceedings of International Joint Conference on Neural Networks*. 1–8.
- 【31】 Weston J, Chopra S, Bordes A (2015) Memory networks. *ArXiv:1410.3916*. <http://arxiv.org/abs/1410.3916>.
- 【32】 Yang Q, Wu G, Li Y, Li R, Gu X, Deng H, Wu J (2020) AMNN: Attention-based multimodal neural network model for hashtag recommendation. *IEEE Transactions on Computational Social Systems* 7(3):768–779. <https://doi.org/10.1109/TCSS.2020.2986778>.
- 【33】 Yang Z, He X, Gao J, Deng L, Smola A (2016) Stacked attention networks for image question answering. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 21–29. <https://doi.org/10.1109/CVPR.2016.10>.
- 【34】 Zhang Q, Wang J, Huang H, Huang X, Gong Y (2017) Hashtag recommendation for multimodal microblog using co-attention network. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. 3420–3426. <https://doi.org/10.24963/ijcai.2017/478>.
- 【35】 Zhang S, Yao Y, Xu F, Tong H, Yan X, Lu J (2019) Hashtag recommendation for photo sharing services. *Proceedings of the AAAI Conference on Artificial Intelligence* 33:5805–5812. <https://doi.org/10.1609/aaai.v33i01.33015805>.
- 【36】 Zhang Z, Bu J, Ester M, Zhang J, Yao C, Li Z, Wang C (2020) Learning temporal interaction graph embedding via coupled memory networks. *Proceedings of The Web Conference*. 3049–3055. <https://doi.org/10.1145/3366423.3380076>.

Appendix A



Table 6. Hyperparameters Used in Our Model

Hyperparameter	Value
The dimensionality of image proposals	512
The dimensionality of text proposals	300
The dimensionality of visual, textual, self post and other post features	300
The number of top k self posts	5
The number of top k other posts	10
The number of most frequently-used hashtags in self posts	10
The number of most frequently-used hashtags in other posts	10
The thresholds of visual feature similarity	0.6
The thresholds of textual feature similarity	0.95