Learning Long- and Short-Term User Literal-Preference with Multimodal Hierarchical Transformer Network for Personalized Image Caption

Wei Zhang,[1,2*] **Yue Ying,**[1] **Pan Lu,**[3] **Hongyuan Zha**[4]
[1]School of Computer Science and Technology, East China Normal University
[2]Key Laboratory of Artificial Intelligence, Ministry of Education, Shanghai
[3]Departments of Statistics and Computer Science, University of California, Los Angeles
[4]School of Computational Science and Engineering, Georgia Institute of Technology
{zhangwei.thu2011, yingyue2011, lupantech}@gmail.com, zha@cc.gatech.edu

Abstract

Personalized image caption, a natural extension of the standard image caption task, requires to generate brief image descriptions tailored for users' writing style and traits, and is more practical to meet users' real demands. Only a few recent studies shed light on this crucial task and learn static user representations to capture their long-term literal-preference. However, it is insufficient to achieve satisfactory performance due to the intrinsic existence of not only long-term user literal-preference, but also short-term literal-preference which is associated with users' recent states. To bridge this gap, we develop a novel multimodal hierarchical transformer network (MHTN) for personalized image caption in this paper. It learns short-term user literal-preference based on users' recent captions through a short-term user encoder at the low level. And at the high level, the multimodal encoder integrates target image representations with short-term literal-preference, as well as long-term literal-preference learned from user IDs. These two encoders enjoy the advantages of the powerful transformer networks. Extensive experiments on two real datasets show the effectiveness of considering two types of user literal-preference simultaneously and better performance over the state-of-the-art models.

Introduction

Inspired by the success of learning multi-modal representations in recent years, image caption (Karpathy and Li 2015; Xu et al. 2015) has become a hotspot for scientific and industrial exploration, aiming at generating natural language descriptions for target images. It finds a wide range of applications such as the reduction of heavy manual cost of writing descriptions for tens of thousands of images and the promotion of visual understanding for machines. Typically, the pipeline for this task involves the following two most fundamental components: a visual understanding module (e.g., convolutional neural network (Krizhevsky, Sutskever, and Hinton 2012)) and a language-oriented decoder (e.g., recurrent neural network (RNN) (Mikolov et al. 2010)).

Despite the remarkable progress in the traditional image caption task, there is an intrinsic limitation that the generated captions are not tailored for individual users. In other words, through the above pipeline, the generated caption of the same image keeps always the same for different users who would like to manually write the captions or mark their lives with photo annotation. Actually, each user has its own literal-preference depending on different writing styles and user states. For example, as shown in Figure 1, the two images are much the same, with strawberries in a plate. However, the captions provided by User 1 and User 2 are apparently different, for User 1 has an objective statement of his breakfast, while user 2 expresses his love to strawberries. As such, it is more practical to conduct personalized image caption to meet users' real demands.

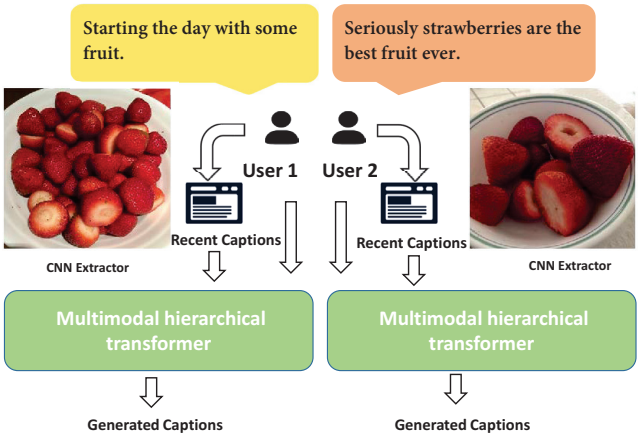

Figure 1: The sketch of personalized image caption with multimodal hierarchical transformer and users' recent captions. The two real examples are gotten from Instagram.

Regarding to this task, only a few pioneering studies investigate the impact of user literal-preference in generating effective personalized captions (Park, Kim, and Kim 2017; Wang et al. 2018; Long, Yang, and Xu 2019; Shuster et al. 2019; Park, Kim, and Kim 2019). They rely on users' active vocabularies and self-descriptions (e.g., tags), as well as their unique IDs to learn latent user representations. They are further integrated with visual representations to generate final captions. Since the user representations are associated

*Corresponding author.
Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

with each user's static characteristics, they are deemed to be able to capture long-term user literal-preference, resulting in improvements over traditional models without considering user personality.

However, we argue that only using the long-term representation is insufficient to achieve satisfactory performance due to the intrinsic existence of both long- and short-term user literal-preference (see Table 1 and Figure 2 for verification). On the one hand, long-term literal-preference commonly reflects a user's personal writing style and its active vocabulary. On the other hand, it is intuitive that a user's recent state will impact his short-term literal-preference, which in turn affects the image caption to be given. Taking a real example from our datasets for illustration. A user first posted an image with the caption "wedding time good luck" and several hours later, he delivered another image with the caption "wedding breakfast". It is obvious that the second image caption depends on the first caption due to the user's specific state. As a result, it is promising to consider the two types of literal-preference into a unified model.

To this end, we develop a novel multimodal hierarchical transformer network (MHTN) for personalized image caption (see Figure 1). It is partially inspired by the powerful transformer network (Vaswani et al. 2017) which can model the complex dependencies among different elements and acquire contextualized representations for each of them. In particular, we first learn to encode users' recent captions through a short-term user encoder at the low level of MHTN, followed by a user-guided attention to obtain the short-term representation of user literal-preference. Afterwards, at the high level, another multimodal encoder is applied to jointly model user short-term representations and target image representation, as well as long-term user representations of literal-preference encoded by user IDs. The contextualized multimodal representations are finally utilized to generate target image captions through a transformer decoder. By this way, our model augments the original transformer network with the ability to encode short-term literal-preference, as well as to capture the multimodal interactions among user ID, text, and image in the task.

We summarize the contributions of this paper as follows:

(1) To our best knowledge, we are the first to address the joint learning of both long- and short-term user literal-preference in the personalized image caption task.

(2) We devise a novel multimodal hierarchical transformer network to encode the two types of literal-preference, as well as to combine target image representation.

(3) We conduct extensive experiments on two publicly available datasets, demonstrating our MHTN achieves the best performance in image caption and the benefit of learning the two types of literal-preference.

Related Work

In this section, we briefly review the literature from the following two aspects, image caption and personalized content generation.

Image caption. Image caption has been a long-standing task which involves both textual and visual modalities, thus attracting researchers from both natural language processing and computer vision communities. Some previous studies (Jia, Salzmann, and Darrell 2011; Kuznetsova et al. 2012) formulate image caption as a retrieval task by searching similar images in the database and their corresponding captions are taken as the captions of query images. Another line of researches (Farhadi et al. 2010; Kulkarni et al. 2011; Lu et al. 2018) focuses on utilizing basic templates to fill the words relevant to the images. Recent studies have shown that deep neural networks with an encoder-decoder framework are effective and flexible in image caption task (Karpathy and Li 2015; Xu et al. 2015), which is motivated by the success in machine translation (Cho et al. 2014). In addition, to overcome the exposure bias (Ranzato et al. 2016) suffered in the decoding stage, techniques like reinforcement learning have been leveraged (Rennie et al. 2017).

Despite much progress in general image caption, personalized image caption, which is more practical to meet users' real demands, did not received attention until the recent several years. The pioneering studies (Park, Kim, and Kim 2017; 2019) address the personalized image caption task by incorporating each user's active vocabularies into memory networks to capture their writing styles. Since users might be associated with self-annotated tags, (Wang et al. 2018) regards these tags as the reflection of users' preference to captions. (Shuster et al. 2019) specifies 215 different personality traits to characterize each user and makes the caption generation dependent on them. However, descriptions of users, including user tags and personality traits, might not always exist in every scenario. An alternative is to learn user representations based on user IDs to denote user latent preference (Long, Yang, and Xu 2019). However, all of the above approaches only learn user static representations to capture the long-term literal-preference, motivateing this work to simultaneously consider both long- and short-term literal-preference which is learned based on users' recent captions and thus is dynamic over time.

Personalized content generation. In the era of user-generated content, automatically generating personalized content has incurred great interest and gotten a thriving development. The researches (Li et al. 2017; 2019) couple the two tasks of personalized rating score prediction (Zhang et al. 2016) and tip generation to benefit each other. (Zhou et al. 2017) generates reviews given user and item factors, as well as sentiment polarity. (Li et al. 2016) also learns from user IDs to incorporate personalization into dialogue generation. (Zeng et al. 2019) utilizes user descriptions such as age and gender to generate social media comments. In addition to the above text generation which commonly has a similar encoder-decoder framework (Sutskever, Vinyals, and Le 2014), (Lin et al. 2019) investigate the personalized fashion generation by generating images through a deconvolutional neural network (Zeiler, Taylor, and Fergus 2011). (Wang, Zhang, and He 2019) synthesizes continuous states and medication dosages of patients with generative adversarial networks. Although the above studies share some spirits with personalized image caption, their problem settings are not exactly the same. Moreover, the short-term user preference is overlooked to some extent by these studies as well.

Preliminaries

We now give the basic notations and formulation of the personalized image caption problem, followed by a real data analysis to verify the motivation of considering both long- and short-term user literal-preference.

Problem Formulation

Assume we have a set of image-caption-user tuples (posts), i.e., $\mathcal{D} = \{(I_i, C_i, U_i)\}_{i=1}^{M}$, where M is the total size of the set. I_i is the raw pixel input of the i-th image. C_i is the caption of the image which contains a list of one-hot encoding of words from a predefined vocabulary $\mathcal{V}$, i.e., $C_i = \{\mathbf{w}_1^i, \cdots, \mathbf{w}_{L_i}^i\}$ where L_i is the length of the caption. U_i consists of two parts, i.e., $U_i = \{\mathbf{u}_i, C_i^U\}$, where $\mathbf{u}_i$ is the one-hot encoding of the corresponding user ID and C_i^U covers one or more of the user's recently posted captions, which is utilized for modeling short-term user literal-preference.

Given the above formulations, the goal of personalized image caption is to learn a model: $\mathrm{f}(I_*, U_*) \rightarrow C_*$, which can generate a caption for any given target image (*) with a specified user. In what follows, we empirically demonstrate the existence of both long- and short-term literal-preference.

Data Verification of Long- and Short-term Literal-preference

We have two real datasets (Park, Kim, and Kim 2019) which come from Instagram and Flickr, respectively. They are named as Instagram and YFCC100M for short.

We first show the existence of user long-term literal-preference by comparing the text similarities of captions belonging to the same user and captions of different users. In particular, each caption is represented by a commonly adopted term frequency−inverse document frequency (TF-IDF) based vector. Given this, we define *intra-user caption similarity* as the average cosine similarity of the TF-IDF based vectors for a single user, and *inter-user caption similarity* as the average cosine similarity for different users. As shown in Table 1, the degrees of intra-user caption similarity are obviously greater than those of inter-user caption similarity. Since the above similarity calculation covers a long time interval, the comparison shows that each user has its own long-term literal preference.

Table 1: Caption similarity analysis

	User-intra caption similarity	User-inter caption similarity
Instagram	0.0225	0.0086
YFCC100M	0.0450	0.0055

In each user caption set, we sort the captions in a chronological order and further calculate the average caption similarity w.r.t. the number of position interval between two captions. This is an in-depth analysis of caption similarity by considering the temporal information in similarity computation. The results in Figure 1 depict an interesting phenomenon that as the position interval gets larger, the caption similarity becomes smaller, with a dramatic decline in the first several position intervals. This consistent observation on the two datasets indicates that even for the same user, the captions have an intrinsic regularity of similarity change. Consequently, we can draw a conclusion that the current caption of a user is more relevant to his recent captions, demonstrating the existence of short-term user literal-preference.

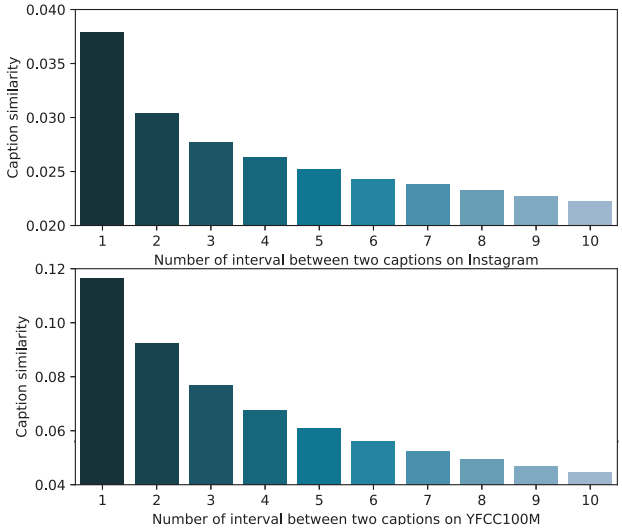

Figure 2: Caption similarity w.r.t. different number of intervals.

The above analysis motivates our study of incorporating long- and short-term literal-preference into personalized image caption. Specifically, user IDs are leveraged to learn long-term user literal-preference and users' recently generated captions are employed to encode short-term user literal-preference.

Proposed Approach

We present multimodal hierarchical transformer network to consolidate the textual and visual modalities, as well as the long- and short-term user literal-preference. The model consists of an input representation module, a hierarchical transformer encoder, and a transformer decoder. The input representation module involves the encoding of target images, words in users' recent captions, and user IDs. Transformer encoder hierarchically encodes short-term literal preference and multimodal representations. The transformer decoder is employed for generating captions as usual. In what follows, we take the image-caption-user tuple (I_i, C_i, U_i) as an example to illustrate the details of our approach.

Input Representation

Image feature extraction We adopt 101-layer ResNet (He et al. 2016) pretrained on the ImageNet dataset as our feature extractor to obtain image feature as follows:

$$\mathbf{i}_i = \mathbf{W}_I \mathrm{CNN}(I_i), \tag{1}$$

where CNN returns the pool5 feature of RestNet, following (Park, Kim, and Kim 2017). $\mathbf{W}_I$ is used to convert the output to the multimodal embedding space, with the dimension $\mathrm{K} = 512$.

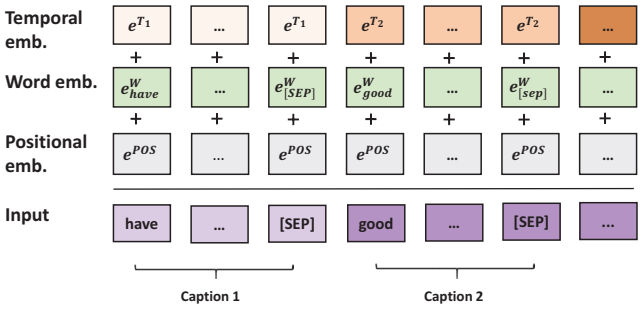

Figure 3: The input embeddings of short-term user encoder.

Hybrid Word Embedding As shown in Figure 3, we employ three types of embeddings to represent each word $\mathbf{w}_j$ in C_i^U, wherein the first two types are commonly used. To be specific, the first type is standard word embedding obtained through a look-up operation, i.e., $\mathbf{e}_j^W = \mathbf{E}^W \mathbf{w}_j$, where $\mathbf{E}^W$ is an embedding matrix. The second type is the positional encoding based on sine and cosine functions proposed in (Vaswani et al. 2017), which we denote as $\mathbf{e}_j^{POS}$ correspondingly.

Since the words might come from different captions which were posted at different time, we propose temporal embedding to characterize the time interval. The intuition behind temporal embeddings is hoping to learn to concentrate more on the captions which were posted more recently. In particular, we empirically set a time interval threshold set as $\mathcal{T} = \{10min, 30min, 2h, 6h, 1d, 3d, 6d, 10d, 1month, 3month, +\infty\}$, wherein each threshold is associated with a temporal embedding $\mathbf{e}^T$ to be learned. A specific time interval is represented by the closest threshold larger than it. We have also tried other similar settings for the threshold set and found the results are close.

Finally, for a given word, its input representation is constructed by summing the above three types of embeddings, denoted as $\hat{\mathbf{e}}^W$. And all the word representations in C_i^U compose an input embedding sequence $\hat{\mathbf{E}}_i^W \in \mathbb{R}^{\mathrm{K} \times |C_i^U|}$, which is later fed into short-term user encoder. It is worth noting that for transformer decoder, the input word representation matrix $\bar{\mathbf{E}}^W$ decoded in previous steps only involves the first two types of embeddings.

Long-term user representation Users who post their captions online are typically associated with user IDs. We aim to leverage the IDs to learn static user representations to capture their long-term literal-preference in image caption. We define a user embedding matrix $\mathbf{E}^U$. And the long-term user representation is then obtained through a look-up operation as well, i.e., $\mathbf{e}_i^{UL} = \mathbf{E}^U \mathbf{u}_i$.

Hierarchical Transformer Encoder

Hierarchical transformer encoder is composed of a low-level short-term user encoder and a high-level multimodal encoder.

Short-term user encoder We first adopt transformer encoder to model the dependencies between different words

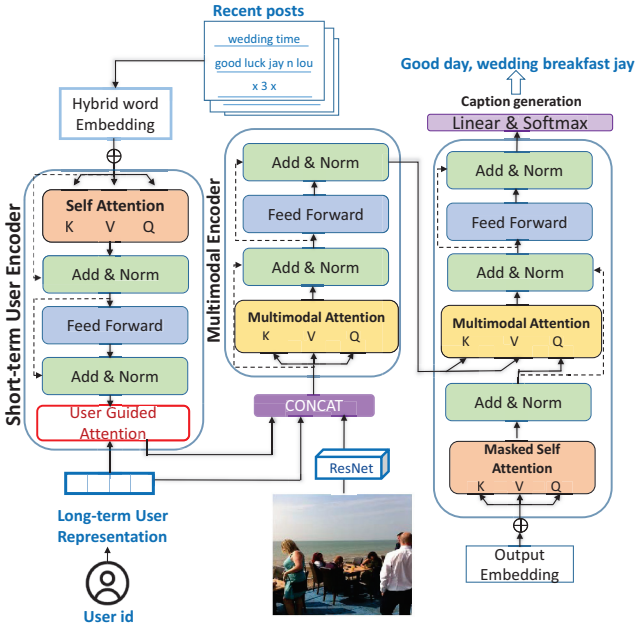

Figure 4: The architecture of multimodal hierarchical transformer network.

in the user's recent captions, which is beneficial for obtaining contextualized word representations. Specifically, multi-head attention (Vaswani et al. 2017) is used, where each header associates all positions in the word sequence with the weighted combination of input word embeddings. Formally, it is defined as follows:

$$\mathrm{H}_m(\hat{\mathbf{E}}_i^W) = \mathrm{softmax}\Big(\frac{(\mathbf{W}_m^Q \hat{\mathbf{E}}_i^W)^\top (\mathbf{W}_m^K \hat{\mathbf{E}}_i^W)}{\sqrt{\mathrm{K/M}}}\Big) \cdot (\mathbf{W}_m^V \hat{\mathbf{E}}_i^W)^\top, \tag{2}$$

where M is number of headers and $m \in \{1, \cdots, \mathrm{M}\}$. $\mathbf{W}_m^Q$, $\mathbf{W}_m^K$, and $\mathbf{W}_m^V$ ($\in \mathbb{R}^{\mathrm{K/M} \times \mathrm{K}}$) correspond to the trainable parameters of query, key, and value, respectively. The representations from each header is fused to form a multi-header based representation as follows:

$$\mathrm{MH}(\hat{\mathbf{E}}_i^W) = \mathrm{MLP}([\mathrm{H}_1(\hat{\mathbf{E}}_i^W); \cdots; \mathrm{H}_\mathrm{M}(\hat{\mathbf{E}}_i^W)]^\top), \tag{3}$$

where MLP denotes a multi-layer perceptron for linear transformation and [;] indicates a row-wise concatenation. Furthermore, residual connection, layer norm (LN), and MLP are combined together to get the contextualized word embeddings as follows:

$$\hat{\mathbf{E}}_{1i}^W = \mathrm{LN}\big(\hat{\mathbf{E}}_i^W + \mathrm{MLP}(\mathrm{LN}(\hat{\mathbf{E}}_i^W + \mathrm{MH}(\hat{\mathbf{E}}_i^W)))\big), \tag{4}$$

where $\hat{\mathbf{E}}_{1i}^W$ denotes the output of the first transformer encoder. In practice, the transformer encoder could be stacked L times and finally the output word embedding matrix is represented as $\hat{\mathbf{E}}_{Li}^W$.

To learn short-term user representation, we introduce a simple user-guided attention mechanism over $\hat{\mathbf{E}}_{Li}^W$. That is, we leverage user long-term representation as a query to attend each word embedding $\hat{\mathbf{E}}_{Lij}^W$ ($j \in \{1, \cdots, |C_i^U|\}$).

Specifically, the attention weight for each word is given as:

$$\alpha_j = \mathrm{softmax}\big(\omega^\top \tanh(\mathbf{W}_U^{ATT}\mathbf{e}_i^{UL} + \mathbf{W}_W^{ATT}\hat{\mathbf{E}}_{Lij}^W + \mathbf{b})\big), \tag{5}$$

where $\mathbf{W}_U^{ATT}$ and $\mathbf{W}_W^{ATT}$ are matrix parameters, while ω and $\mathbf{b}$ are vector parameters. After that, the short-term user representation is encoded as,

$$\mathbf{e}_i^{US} = \sum_{j=1}^{|C_i^U|} \alpha_j \hat{\mathbf{E}}_{Lij}^W. \tag{6}$$

Multimodal encoder The multimodal encoder takes user long- and short-term representations, as well as image representation as input, and adopts another transformer encoder to model their inter-modal interaction. A multimodal embedding matrix is first formed based on the column-wise concatenation $\mathbf{E}_i^M = [\mathbf{e}_i^{UL}, \mathbf{e}_i^{US}, \mathbf{i}_i]$. In a similar way as described in Equation 2, 3, and 4, we obtain a contextualized multimodal embedding matrix, i.e., $\mathbf{E}_{Li}^M$, where the encoder is also stacked L times, without loss of generality.

Caption Generation and Training

In caption generation, the transformer decoder develops a masked self-attention operation to ensure the word generation for position j to be only influenced by the generated words before this position. In multimodal attention of the decoder, the obtained multimodal embedding matrix is employed as key and value, and the output word embedding is regarded as query. This ensures that the word generation is directly affected by both long- and short-term user representations, as well as target image representations. The training target of our model is to maximize the likelihood of generating true descriptions for images in the dataset $\mathcal{D}$. We leave the incorporation of other training methods such as reinforcement learning as future work.

Experiments

Experimental Setup

Datasets As aforementioned, we have Instagram and YFCC100M based on the InstaPIC-1.1M dataset and the YFCC100M benchmark dataset respectively (Park, Kim, and Kim 2019). Since the original InstaPIC-1.1M dataset has not stored the time information of each post, we crawl the raw data of the posts from the website via each user name appearing in the dataset. For both datasets, we follow (Park, Kim, and Kim 2019) to remove duplicate posts and lengthy captions. To prevent models from peeking users' future literal preference, we sort all posts of each user in a chronological order. We split the two datasets by taking the first 85% posts as training sets, then 5% of posts as validation sets, and the last 10% of posts as test sets. The main statistics of the two datasets are summarized in Table 2.

Implementation details We tune the hyper-parameters of all adopted models by their performance on validation datasets for a fair comparison. To train the MHTN model, we use the Adam optimizer with $\alpha = 0.9, \beta = 0.999, \epsilon = 1 \times 10^{-8}$, and the batch size to be 100. We set the dropout ratio to 0.1 for intermediate layers. We also apply label smoothing (Szegedy et al. 2016) with factor of 0.1 to our training procedure. Gradient clipping is used with the range $[-0.1, 0.1]$. The hyper-parameters of transformer are $N = 6$, $\mathrm{M} = 8$, $K = 512$. The default number of recent posts considered by our model is set to 5.

Following (Park, Kim, and Kim 2017), we report the caption generation results by decoding each position with the most likely word for all approaches. Beam search with different small sizes are also conducted and similar conclusions w.r.t. performance comparison can be drawn. Due to the space limitation, we do not report these results.

Table 2: Statistics of the datasets.

Data	Post	User	Time Span	Vocab. Size
Instagram	363,656	2,888	2010-2016	40,000
YFCC100M	353,259	5,868	2004-2014	40,000

Baselines The involved baselines are as follows:

1NN-IM, 1NN-Usr, and 1NN-UsrIM (Park, Kim, and Kim 2017): They are retrieval based baselines by taking the captions of the nearest training image and nearest user as generated captions.

ShowTell (Vinyals et al. 2015): ShowTell is a pioneering encoder-decoder based model for generating captions with an RNN decoder.

ShowAttTell (Xu et al. 2015): ShowAttTell incorporates a visual attention computation to capture the importance of each image region in word decoding.

Transformer (Vaswani et al. 2017): We take the image representation by outputing of the last convolutional layer of 101-layer ResNet with the size 196×2048, as the input of transformer encoder.

Attend2u (Park, Kim, and Kim 2017): Attend2u is the first model for personalized image caption by modeling a user's active vocabulary as its memory context.

CDPIC (Long, Yang, and Xu 2019): This model utilizes user IDs and takes their frequently used words as context, and also adopts an RNN based encoder-decoder framework.

EICP (Shuster et al. 2019): The one-hot encoding of personality is used in EICP with the UPDOWN (Anderson et al. 2018) strategy for image caption. To ensure fairness, we regard user ID as user personality, and only use top-down attention since bottom-up attention involves image region box detection which is out the scope of this paper.

Experimental Results

Model comparison Table 3 mainly presents the caption generation performance on the two adopted datasets by MHTN and compared models. The evaluation metrics include language similarity metrics (BLEU, CIDEr, METEOR, and ROUGE-L) and the tailored image caption performance metric (SPICE). The retrieval based models in the first part of the table perform poorly compared with other generative models. In the second part where all models do not consider personalization, Transformer outperforms ShowTell and ShowAttTell, showing its good modeling capability in image caption. The three baselines in the third

Table 3: Evaluation results by our model and compared models on the Instagram and YFCC100M datasets.

Instagram										
Methods	BLEU1	BLEU2	BLEU3	BLEU4	CIDEr	METEOR	ROUGE-L	SPICE	Time-TR	Time-TE
1NN-Im	0.026	0.001	0.000	0.000	0.011	0.009	0.026	0.005	—	—
1NN-Usr	0.042	0.008	0.002	0.001	0.021	0.020	0.038	0.004	—	—
1NN-UsrIm	0.037	0.008	0.002	0.001	0.018	0.019	0.034	0.003	—	—
ShowTell	0.055	0.016	0.006	0.002	0.045	0.020	0.061	0.009	0.22s	0.15s
ShowAttTell	0.049	0.015	0.005	0.003	0.056	0.021	0.063	0.014	—	—
Transformer	0.060	0.019	0.008	0.004	0.079	0.026	0.070	0.019	0.89s	0.91s
Attend2u	0.065	0.020	0.008	0.004	0.076	0.026	0.069	0.013	—	—
CDPIC	0.057	0.020	0.009	0.005	0.080	0.024	0.071	0.020	—	—
EICP	0.062	0.023	0.011	0.006	0.094	0.028	0.078	0.022	0.86s	0.88s
MHTN	**0.093**	**0.036**	**0.017**	**0.010**	**0.125**	**0.042**	**0.089**	**0.025**	0.56s	0.58s
YFCC100M										
Methods	BLEU1	BLEU2	BLEU3	BLEU4	CIDEr	METEOR	ROUGE-L	SPICE	Time-TR	Time-TE
1NN-Im	0.046	0.014	0.005	0.003	0.057	0.017	0.042	0.002	—	—
1NN-Usr	0.038	0.009	0.003	0.002	0.018	0.012	0.032	0.004	—	—
1NN-UsrIm	0.042	0.010	0.003	0.001	0.016	0.012	0.032	0.004	—	—
ShowTell	0.070	0.024	0.010	0.001	0.069	0.021	0.064	0.016	0.22s	0.15s
ShowAttTell	0.079	0.032	0.017	0.011	0.101	0.024	0.078	0.026	—	—
Transformer	0.082	0.036	0.020	0.013	0.138	0.029	0.085	0.032	0.89s	0.91s
Attend2u	0.076	0.025	0.010	0.004	0.075	0.029	0.083	0.017	—	—
CDPIC	0.101	0.053	0.034	0.024	0.205	0.037	0.100	0.042	—	—
EICP	0.117	0.066	0.044	0.032	0.263	0.044	0.112	0.047	0.86s	0.89s
MHTN	**0.145**	**0.091**	**0.066**	**0.053**	**0.408**	**0.063**	**0.133**	**0.059**	0.55s	0.58s

part behave better in most of the metrics, indicating the necessity of considering personality in achieving good caption performance. By comparing our model MHTN with the other baselines, we can see consistent and significant improvements. Specifically, the improvements of MHTN over state-of-the-art approach EICP are statistically significant, from 0.094 to 0.125 by CIDEr on Instagram and from 0.047 to 0.059 by SPICE on YFCC100M. Since EICP also learns from user IDs and uses advanced attention mechanism for encoder-decoder modeling, the comparison with EICP reveals our improvements are attributed to the advantage of encoding short-term user literal-preference and the powerful hierarchical multimodal transformer architecture.

In addition, the right region of Table 3 shows the average running time of MHTN and several other baselines w.r.t. training (Time-TR) and testing (Time-TE) on each batch data. The number of training epochs needed to converge is similar for them. We find MHTN runs faster than EICP and Transformer but slower than ShowTell. In total, the training of MHTN can be completed in less than 18 hours with only 1 GPU, which is feasible for these image caption datasets.

Ablation study We conduct ablation experiments to further verify the contribution of individual component design in our model. In particular, we consider the following variants: 1) "w/o Temporal Emb." removes temporal embedding from the hybrid input embeddings of short-term user encoder; 2) "w/o Image" does not input target image; 3) "w/o Long-term User Rep." removes the long-term user representation; 4) "w/o Short-term User Rep." removes short-term user encoder; and 5) "RNN+Transformer" replaces transformer based short-term user encoder with time-aware RNN based encoder which also considers temporal embedding to get the short-term user representation.

Table 4 shows the results of MHTN and its variants. From a whole perspective, temporal embedding makes less contribution than other components, but still make a positive contribution on the two datasets. Moreover, the performance degradation of "w/o Image" reveals the visual content is an indispensable component in personalized image caption, which confirms to intuition. By further comparing MHTN with "w/o Long-term User Rep." and "w/o Short-term User Rep.", we can see the performance goes through significant improvements, showing the indeed positive effects brought by both long- and short-term user representations. Finally, we compare our model with "RNN+Transformer" and the better results show the benefit of proposing to use transformer as short-term user encoder.

Qualitative analysis Figure 5 shows some selected images and their captions in two parts. The left part, out of the dotted box, compares the captions generated by different methods. It is undeniable that image caption is a hard task because the ground-truth captions from different users involve diverse perspectives, literal-preference, and even some named entities. For example, the first image in the second column contains a location entity "washington" in its caption. However, the generated captions by the selected models are relevant to the images to some extent. More importantly, the colored words captured by our model provide some details about the images and seem to be related to users' recent state or literal preference. In addition, the captions from our model apparently have a richer vocabulary than other models, making the caption more descriptive. In the part of the dotted box, we can find: (i) different query users indeed generate captions from different perspectives for the same images; and (ii) considering users' recent captions benefit capturing more details might relevant to the images.

Table 4: Ablation study of MHTN on the two datasets.

Instagram								
Methods	BLEU1	BLEU2	BLEU3	BLEU4	CIDEr	METEOR	ROUGE-L	SPICE
MHTN	**0.093**	**0.036**	**0.017**	**0.01**	**0.125**	**0.042**	**0.089**	**0.025**
w/o Temporal Emb.	0.089	0.034	0.015	0.008	0.119	0.039	0.084	0.022
w/o Image	0.078	0.028	0.016	0.007	0.087	0.036	0.079	0.016
w/o Long-term User Rep.	0.078	0.028	0.012	0.007	0.097	0.034	0.080	0.022
w/o Short-term User Rep.	0.080	0.030	0.014	0.007	0.107	0.036	0.082	0.022
RNN+Transformer	0.089	0.032	0.015	0.008	0.117	0.038	0.088	**0.025**

YFCC100M								
Methods	BLEU1	BLEU2	BLEU3	BLEU4	CIDEr	METEOR	ROUGE-L	SPICE
MHTN	**0.145**	**0.091**	**0.066**	**0.053**	**0.408**	**0.063**	**0.133**	**0.059**
w/o Temporal Emb.	0.138	0.085	0.060	0.046	0.361	0.060	0.126	0.055
w/o Image	0.130	0.082	0.060	0.048	0.350	0.053	0.118	0.045
w/o Long-term User Rep.	0.126	0.070	0.050	0.038	0.297	0.050	0.119	0.049
w/o Short-term User Rep.	0.118	0.072	0.052	0.040	0.332	0.051	0.113	0.051
RNN+Transformer	0.141	0.081	0.058	0.046	0.364	0.058	0.130	0.058

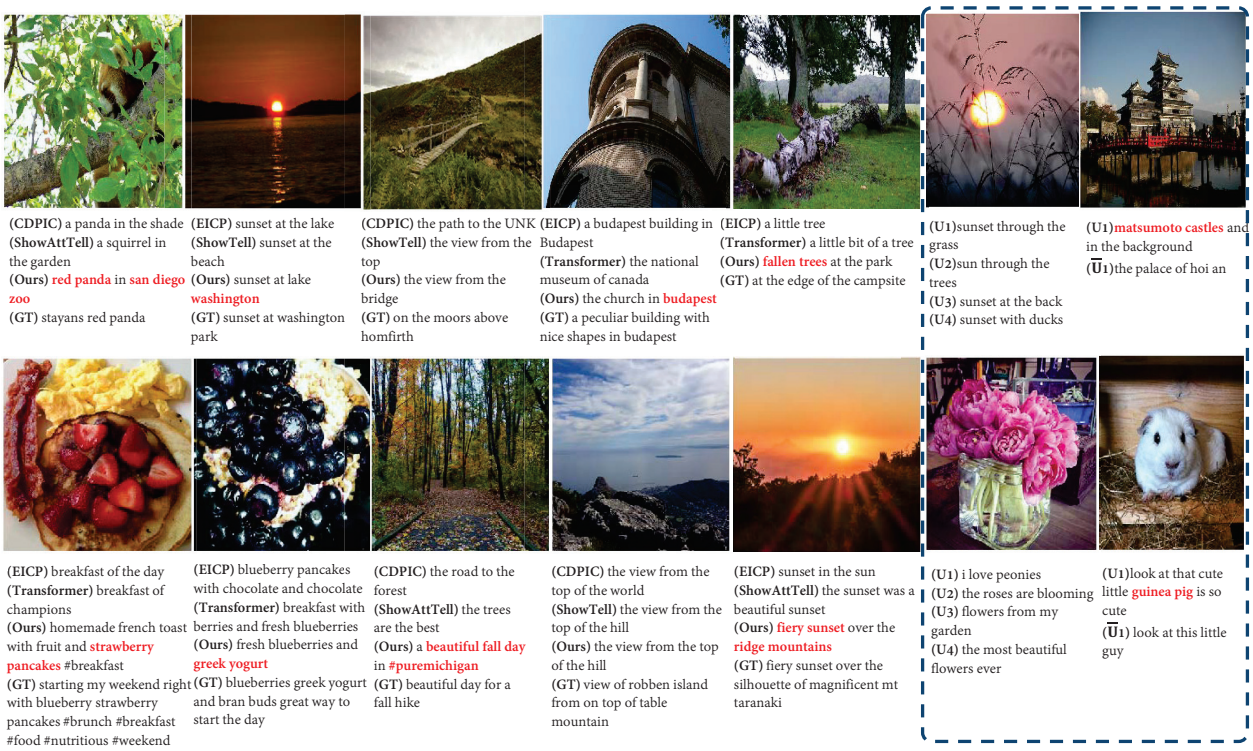

Figure 5: Examples from YFCC100M (top) and Instagram (bottom). For each image out of the dotted box, we present its grond truth (GT) caption, accompanied by the ones generated by our model MHTN and some strong baselines. And for each image in the dotted bos, we present captions by different query users, denoted by U1 for example. Ū1 corresponds to captions generated without considering the user's recent posts, in comparison with the full version of MHTN.

Effect of the number of recent posts Table 5 shows that, as the number of posts increases, "w/o Time-emb." gains better performance in terms of CIDEr at first and then the performance drops, implying the latest posts have larger contributions on caption generation. By contrast, the results of MHTN become better at first and remain stable. This is because temporal embedding could help differentiate the recent posts from other older posts.

Conclusion

In this paper, we develop a novel multimodal hierarchical transformer network to encode both long- and short-term user literal-preference for personalized image caption. The goal is achieved by the low-level user-dependent transformer encoder to learn short-term user representations from users' recent posts, and the high-level multimodal transformer encoder to integrate short-term user representations and long-

Table 5: Results of different number of recent posts.

Method	Length	Instagram	YFCC100M
MHTN	1	0.120	0.386
	2	0.120	0.394
	5	0.125	0.408
	10	0.126	0.407
w/o Temporal Emb.	1	0.117	0.356
	2	0.119	0.361
	5	0.113	0.344
	10	0.110	0.346

term user representations of user IDs, as well as image representations. We have conducted experiments on two publicly available datasets, showing the superiority of our model and validating the contributions of its main components.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable suggestions. This work is supported by National Key Research and Development Program (2019YFB2102600), NSFC (61702190), Shanghai Sailing Program (17YF1404500), NSFC-Zhejiang (U1609220), and the foundation of Key Laboratory of Artificial Intelligence, Ministry of Education, P.R. China.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and VQA. In *CVPR*, 6077–6086.

Cho, K.; van Merrienboer, B.; Gülçehre, Ç.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, 1724–1734.

Farhadi, A.; Hejrati, S. M. M.; Sadeghi, M. A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; and Forsyth, D. A. 2010. Every picture tells a story: Generating sentences from images. In *ECCV*, 15–29.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Jia, Y.; Salzmann, M.; and Darrell, T. 2011. Learning cross-modality similarity for multinomial data. In *ICCV*, 2407–2414.

Karpathy, A., and Li, F. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 3128–3137.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*, 1106–1114.

Kulkarni, G.; Premraj, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A. C.; and Berg, T. L. 2011. Baby talk: Understanding and generating simple image descriptions. In *CVPR*, 1601–1608.

Kuznetsova, P.; Ordonez, V.; Berg, A. C.; Berg, T. L.; and Choi, Y. 2012. Collective generation of natural image descriptions. In *ACL*, 359–368.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G. P.; Gao, J.; and Dolan, W. B. 2016. A persona-based neural conversation model. In *ACL*.

Li, P.; Wang, Z.; Ren, Z.; Bing, L.; and Lam, W. 2017. Neural rating regression with abstractive tips generation for recommendation. In *SIGIR*, 345–354.

Li, P.; Wang, Z.; Bing, L.; and Lam, W. 2019. Persona-aware tips generation? In *TheWebConf*, 1006–1016.

Lin, Y.; Ren, P.; Chen, Z.; Ren, Z.; Ma, J.; and de Rijke, M. 2019. Improving outfit recommendation with co-supervision of fashion generation. In *TheWebConf*, 1095–1105.

Long, C.; Yang, X.; and Xu, C. 2019. Cross-domain personalized image captioning. *Multimedia Tools and Applications*.

Lu, D.; Whitehead, S.; Huang, L.; Ji, H.; and Chang, S. 2018. Entity-aware image caption generation. In *EMNLP*, 4013–4023.

Mikolov, T.; Karafiát, M.; Burget, L.; Cernocký, J.; and Khudanpur, S. 2010. Recurrent neural network based language model. In *INTERSPEECH*, 1045–1048.

Park, C. C.; Kim, B.; and Kim, G. 2017. Attend to you: Personalized image captioning with context sequence memory networks. In *CVPR*, 6432–6440.

Park, C. C.; Kim, B.; and Kim, G. 2019. Towards personalized image captioning via multimodal memory networks. *TPAMI* 41(4):999–1012.

Ranzato, M.; Chopra, S.; Auli, M.; and Zaremba, W. 2016. Sequence level training with recurrent neural networks. In *ICLR*.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *CVPR*, 1179–1195.

Shuster, K.; Humeau, S.; Hu, H.; Bordes, A.; and Weston, J. 2019. Engaging image captioning via personality. In *CVPR*, 12516–12526.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *NIPS*, 3104–3112.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *CVPR*, 2818–2826.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*, 5998–6008.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*, 3156–3164.

Wang, L.; Chu, X.; Zhang, W.; Wei, Y.; Sun, W.; and Wu, C. 2018. Social image captioning: Exploring visual attention and user attention. *Sensors* 18(2):646.

Wang, L.; Zhang, W.; and He, X. 2019. Continuous patient-centric sequence generation via sequentially coupled adversarial learning. In *DASFAA*, 36–52.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A. C.; Salakhutdinov, R.; Zemel, R. S.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2048–2057.

Zeiler, M. D.; Taylor, G. W.; and Fergus, R. 2011. Adaptive deconvolutional networks for mid and high level feature learning. In *ICCV*, 2018–2025.

Zeng, W.; Abuduweili, A.; Li, L.; and Yang, P. 2019. Automatic generation of personalized comment based on user profile. In *ACL, Student Research Workshop*, 229–235.

Zhang, W.; Yuan, Q.; Han, J.; and Wang, J. 2016. Collaborative multi-level embedding learning from reviews for rating prediction. In *IJCAI*, 2986–2992.

Zhou, M.; Lapata, M.; Wei, F.; Dong, L.; Huang, S.; and Xu, K. 2017. Learning to generate product reviews from attributes. In *EACL*, 623–632.