Towards Personalized Aesthetic Image Caption

Kun Xiong, Liu Jiang,Xuan Dang,Guolong Wang,Wenwen Ye,Zheng Qin*
School of software, Tsinghua University, Beijing, China
{xk18,jiangl16}@mails.tsinghua.edu.cn
dangxuaner@163.com
{wanggl16, yeww14, qingzh}@mails.tsinghua.edu.cn

***Abstract*—Image captioning (IC) is a commonly-used technique for generating textual image description, which finds its applications on semantic image retrieval and multi-modal image understanding, among many others. This paper focuses on an important IC method specialized for generating aesthetic descriptions of images, i.e., aesthetic image captioning (AIC). Despite some effectiveness of initial work on AIC, their performances are inherently limited due to a lack of consideration of user preferences on aesthetics and better aesthetic feature, making it unusable for real-world applications where human users present a large variation on evaluating visual aesthetics of images. To tackle this, we propose a novel personalized aesthetic image caption (PAIC) approach for capturing and incorporating user preferences for AIC tasks. Our approach mainly contains Aesthetic feature Extraction Network(AEN), User Encoder network(UEN) and a personalized image caption model. AEN is designed to extract more expressive feature, UEN is introduced for learning the user vector from the limited information in our AVA-PCap dataset. Personalized image caption model is constructed to generate the caption when given the user id and photo pairs. The experimental results show that our methods outperform baselines by 10% , which is encouraging for a first step towards personalized aesthetic image caption.**

***Index Terms*—image caption, neural network, deep learning, personalized, aesthetic**

I. Introduction

Image Captioning (IC) is an image processing technique that generates textual descriptions for images of interest [1] [2] [3] [4]. The description usually contains key information of the given image, such as attributes of entities and relationships between entities. IC has a wide variety of applications, which include semantic image retrieval and image-based chat-bots. In this paper, we focus on a particular IC technique variant, i.e., Aesthetic Image Captioning(AIC), which focuses on the aesthetic features of a photograph and generates textual descriptions that best present visual aesthetic attributes.

In the context of AIC, the focused aspect of an image, i.e., aesthetics, is mostly based on subjective notions such as beautiful and ugly, which varies a lot among human users. Therefore, compared with general IC methods, AIC is more related to visual aesthetics perceived by each human individual, and the essence of its task is to capture and evaluate what a user likes and dislikes, other than to simply list the objects and actions present in the image.

* is corresponding author.

Very cool. I love the translucency of the balloons.

A group of people shopping at an outdoor market.

Fig. 1. The differences between IC and AIC: left picture comes from our AVAPCap dataset, right one comes from flick-30k dataset. We can see that AIC pays more attention on the feelings of the user and aesthetic features.

Fig. 2. This photo and reviews come from our proposed dataset AVAPCap. Different users give various comments to the same photo. These comments focus on different aspect. Green words show the preferences of the users. Red number is the user id.

To elaborate on this, Figure 1 intuitively reveals the differences between IC and AIC. This example implies that approaches for AIC should pay more attention to what the user feels about the image, but not only what the image presents. In the left picture, the caption doesn't illustrate the balloons, but the feeling of the user.

To the best of our knowledge, there has just been little efforts on aesthetic image caption [5], [6]. Chang et al. [5] is the first study of the AIC problem. This paper firstly contributes a dataset PCCD that contains around 4000 images and 60000 reviews for seven well-known aspects, and then proposed an Aspects-fusion approach(AF) to generate caption by fusing the fitness of the seven aspects. In Chang's model,

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

vanilla CNN-LSTM caption approach is applied in every aspects to fuse and different aspects have different LSTMs but share the same CNN network. Considering [5]'s poor use of aesthetic information and small dataset, Ghosal et al. [6] follows Chang's work and proposes a caption filtering strategy, compiling a cleaner and larger dataset AVA-Caption and proposing a strategy for training a convolutional neural network, which applies LDA topic model on the reviews and learns CNN parameters by fitting the topic distribution.

However, There exist two main defects in the aforementioned models. Firstly, the two models only use the high-level output of CNN as feature to predict the aesthetic review, but low-level visual information and local visual information contribute more to some aesthetic styles, where high-level information refers to objects and low-level feature refers to color or shape. To prove this, we list the 14 most used aesthetic styles in aesthetic photos from AVA dataset [7] and make a table to show which level is decisive for each style. In table I, we can see almost all styles relate to low-level color feature.

The second defect of those approaches is that they predict the caption only by visual information and ignore an important fact that different users have different aesthetic preferences. In figure 2, different users give different comments to the same picture and the discrepancy comes from various concerns of users . We can easily identify that the #15711 user pays more attention to low level aesthetic feature such as color and composition, however, the #97225 user thinks more about the high level scene of the photo. The preferences of users lead to very distinct comments. In real-world applications, the different preferences in aesthetics between users are very common and they play an important role in determining how a specific user feels. Therefore, we argue that modeling user preference and fully incorporate this into image caption models is essential for a more informed and flexible AIC, rendering it more possible to bridge the gap between AIC models and real applications. However, the problem of capturing user preference for aesthetic image captioning is less explored by existing works.

To resolve the first defect, this paper presents an Aesthetic feature Extraction Network(AEN) to get the more comprehensive visual feature rather than applies a vanilla CNN network. To collect a multi-level feature and reduce the aesthetic quality loss due to resizing image to fit CNN, AEN collects a Multi-level Spatial Pooled feature(MLSP [8]), which is extracted from every layer of an Inception-v3 network and maked unrelated to image size by global average pool. MLSP feature considers multi-level spatial information and improves the aesthetic quality assessment(AQA) accuracy, which can solve aforementioned defect well. A multi-columns CNNs feature is concatenated with MLSP, which is a common thought in AQA task and is proven to be beneficial to aesthetic tasks.

The second defect our model aimed to resolve is to utilize the user preference information to improve the caption model. Our model assumes that every single item contains image, review and the information of user who produces the comment. There are many large scale aesthetic datasets in AQA task such as AVA dataset [7] and aforementioned PCCD , however, there is no aesthetic dataset that contains any user information. To evaluate our model, we compiled a dataset for aesthetic image captioning called AVA-PCap[1]. Every item of AVA-PCap consists of aesthetic image from AVA dataset and pairwise data of user id and related review, and users can have various comments for the same images. To the best of our knowledge, though AVA-PCap only contain the limited user information, this is the first and only aesthetic caption dataset with user information.

Based on common encoder-decoder architecture, We propose a novel approach to utilize the user information and resolve the second main defect. As the Figure 3 shows, our approach consists of two steps. In the first step, a User Encoder Network(UEN) is presented to extract the user preference information and encode it into a user vector. In the second step, we transform the preference vector into vocabular preference vector and visual preference vector. The vocabulary preference vector helps the caption decoder model to choose proper word, and the visual preference vector helps the encoder to decide which visual levels are more important.

After addressing the above defects, we integrate the above sub-networks together with a customized encoder-decoder language model and name our model as Personalized Aesthetic Image Caption model (PAIC). Distinct from generic image captioning task, PAIC receives a user identification and an image which the user never commented before, and predicts the review that is most likely to be give by the user, which can be very helpful for prediction of the aesthetic preferences of a particular user. In our experiments we evaluate the effectiveness of our PAIC model by predicting the comment of a particular pair of user and image. Metric scores which are common used in IC task, is calculated to measure the effectiveness. In ablation study, we evaluate the impact of applying AEN and UEN; both the quantitative and qualitative show that our sub-models do improve the ability of aesthetic caption prediction.

Main contributions of this paper include:

- To the best of our knowledge, The paper is the first one to consider the task of personalized aesthetic image caption.
- We compile the first aesthetic review dataset contains personal information, and evalute our model on it.
- We propose the two defects of the prior works in AIC, and propose a novel PAIC model which fully exploits the personal information and extracts more accurate aesthetic feature to resolve the defects.
- We do exhaustive experiments and outperform sever competing alternatives.

II. Related work

This section reviews existing efforts related to our proposed personalized aesthetic image caption method. Relevant work can be mainly categorized into two lines of research: image captioning and aesthetic quality assessment .

[1]download: https://cloud.tsinghua.edu.cn/f/4fa024c5606248b185b1/?dl=1

TABLE I
AESTHETIC STYLES AND WHICH LEVEL IT DEPENDS ON

Aesthetic Feature	low level feature	high level feature
Duotones	✓(color)	
HDR	✓(color)	
Image Grain	✓(shape)	
Light On White	✓(color)	
Long Exposure		✓(object)
Macro		✓(object)
Motion Blur	✓(color)	✓(object)
Negative Image	✓(color)	✓(object)
Rule of Thirds		✓(position)
Shallow DOF	✓(color)	✓(object)
Silhouettes	✓(color)	✓(object)
Soft Focus	✓(color)	✓(object)
Vanishing Point	✓(shape)	

TABLE II
DETAILS OF TWO DATASET

dataset	images	users	reviews
Raw AVA-PCap	39180	13037	453528
AVA-PCap	8288	4779	31551

Image Captioning. In recent years, lots of works has been published on image captioning [6] [5] [1] [2] [3] [9] [10]. General image captioning task aims to generate a sentence to describe the main content of an image. [1] transferred the encoder-decoder framework from machine translation to image caption. [2] brought in attention mechanism and improved the performance of ShowTell model. There are many follow-up works after the introduction of attention mechanism. [10] exploited semantic attention to combine top-down and bottom-up strategies to extract richer information from images, and coupled it with an LSTM decoder. At the same time, a lot of researchers tried to extend general image caption task to many related areas. [11] proposed StyleNet to generate various style captioning for a given image. [9] used a memory net as decoder and tried to modify general image caption model to attend on particular user. [5] and [6] transferred the general image caption model to aesthetic dataset and set up aesthetic image caption task. Our research is based on aesthetic image caption and introduces an effective personalized model into this area.

Aesthetic Quality Assessment. Our research is also related to aesthetic quality assessment(AQA). AQA is the first task concerning aesthetic images. The goal is to evaluate aesthetic score for a given photo. Datta et.al. [12] first casted the image aesthetics assessment problem as a classification or regression problem. With the prospering of CNN network for image processing, Lu et.al. [13] used CNN to extract features from multi-patch of the aesthetic image and concatenated them togather as final visual feature. In [14], Lu et.al. also found that two identical CNNs with no shared parameters can help improve the aesthetic feature extraction progress. Based on the work of Lu, Wang et.al. [15] extended the two CNNs to Multi-CNNs and tried to prove the effectiveness by the aesthetic cognitive process of brain. [16] tried to train a personalized regression model to predict a personalized offset score try to model the bias of different user. Our work take in some great ideas and fit them to our task.

III. MODEL AND APPROACH

A. Problem Formulation

We formulate the personalized AIC problem as follows. Let U and I denotes a set of users and images, respectively. The set cardinality, $|U|$ and $|I|$, represents the number of the users and images, and $|W|$ represents the vocaburay size. Based on the above notations, the user-image interactions form a subset $S \subseteq U \times I$; each interaction is associated with a sequence of words $C_{u,i} = \{C_{u,i}^1, C_{u,i}^2 ... C_{u,i}^{|C_{u,i}|}\}$, which is sourced from the comment posted by a user to describe the visual image aesthetics perceived by him. The task of the personalized AIC problem is to predict the sequence $C_{u,i}$ given $u \in U$ and $i \in I$.

B. Overview

The proposed PAIC model contains three main modules, i.e., a Aesthetic feature Extraction Network(AEN), a User Embedding Network(UEN) and a Personalied Caption Network based on encoder-decoder framework. Figure 3 shows the overview of our model and the relation between the three modules. AEN aims to generate multi-level visual feature to better fit aesthetic task for aesthetic image; UEN aims to collect the user visual perferences and linguistic perferences from reviews by user. The Personalized caption model takes in the outputs from AEN and UEN to predicts the reviews. The rest of this section introduces the three modules in order.

C. Aesthetic feature Extraction Network(AEN)

AEN accepts a image as input and outputs a aesthetic vector as a output. We will describe its details as follows. Firstly, we present the extraction of MLSP feature. Then we introduce the basic idea of Multi-Column CNN and show how to it is appied on our AEN network.

MLSP feature: MLSP are visual features proposed for aesthetic quality assessment task. MLSP are mainly focused on the extraction of the multi-level visual attributes based on a inception network. In [8], Vlad Hosu et al. collects the MLSP feature by gethering the outputs of every convolution block layer of Inception-v3 network and then applying a global average pooling on each of them. After concatenating these $1 \times 1 \times N_{i-layer}$ features, the MLSP feature is composed of multi-perception-field visual feature and it is invariant to the image size, which indicates that MLSP feature can capture the multi-level feature and reduce the loss of aesthetic quality by avoiding resizing the input image.

Multi-Column CNNs: In aesthetic quality assessment task, Multi-Column CNNs approach is proven to be a effective way to improve the feature quality. [14] [13] The main idea of Multi-Column CNNs is to exploit additional CNNs to extract features of the same images from different aspects. The additional CNN networks can only share the lower level parts of parameters but not the high level parts. In [14], the

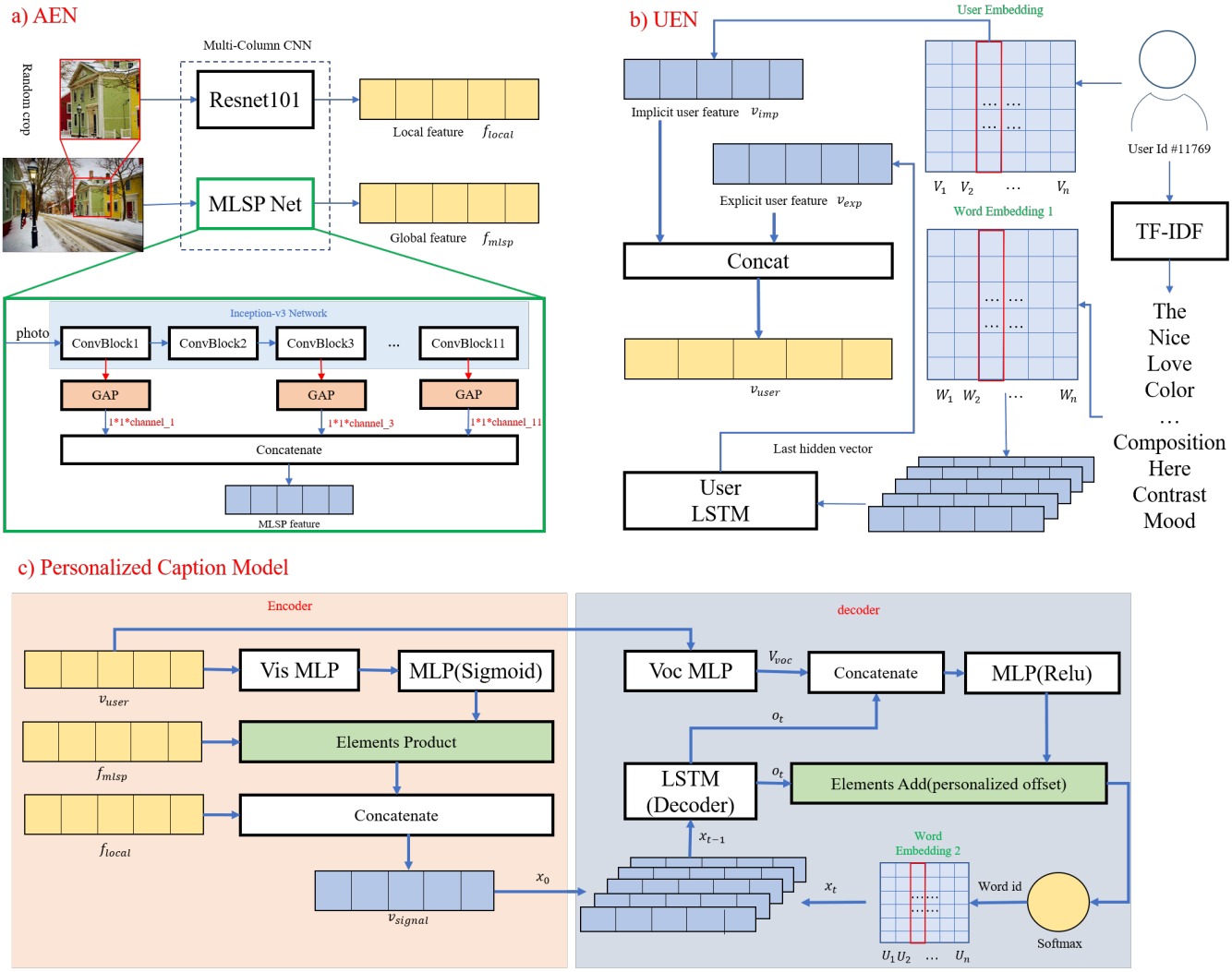

Fig. 3. Framework of our PAIC model: **a)** is AEN submodel, introduced to extract effective aesthetic feature. AEN take the idea of multi-patch to extract local feature and aggregates MLSPNet to extract size-irrelevant multi-level feature. **b)** is our UENetwork. UEN exploits the limited information from our dataset and models the perference of users in a vector. **c)** is the personalized image caption model, which take in the outputs from the AEN and UEN to predict the review of given user and image.

author use two identical CNNs to extract the local and global feature respectively. Based on this, Lu et al. [13] adds more addictional CNNs to apply on the same task.

AEN network : Noticing the aforementioned two aproaches are effective for improving aesthetic feature, our AEN feature integrates them into an organic one: using MLSP for multi-preception-field information encoding and Multi-Column CNNs thought for local information encoding. In our AEN feature, to reduce the memory usage and avoid overfitting, our MLSP feature is extracted by:

$$f_{mlsp} = [Incep_1||Incep_3||Incep_5||...||Incep_{11}] \quad (1)$$

where the $Incep_i$ means the ith-level output of the Inception-v3 network, and $||$ is the concatenate operation. And our local visual information is extracted by another Resnet-101, which receives a random small crop of images and scale it as input images.

$$f_{local} = Resnet(I_{crop}) \quad (2)$$

The above two features $\{f_{local}, f_{mlsp}\}$ are the output of our AEN network and they are encoded to one single vector together with user vectors later in the personalized caption network.

D. User Embedding Network

The user embedding network is a function f_u that maps available user information to user vector. We model the user vector from two different aspects as follows, explicit aspect and implicit aspect. explicit aspect exploit the textual information and implicit aspect help to model the text-independent perferences as a supplement.

Explicit aspect: We realize the user information used to encode perferences must be the most distinctive statistics in the dataset. In AVA-PCap dataset, The small amount user information we can get is user id and the textual reviews given by the same user. So we introduce the Term Frequency–Inverse Document Frequency from natural language processing into this area [17]. TF-IDF is a numerical statistic invented to find the importance of words to the document in a copus. TF-IDF assigns every word a importance score for each document. The higher the score is, the more distinctive the word is in this document comparing other documents. In our method, we group reviews into documents by user id and gather the documents as a copus, then we apply TF-IDF and select the K most scored words defined as $\{w_i\}_{i=1}^{K}$, For simplicity, we ignore the user label of words. we feed the embedded word vectors $\{E(w_i)\}_{i=1}^{K}$ into a user-LSTM in a increasing order of scores. The reason of the increasing order is to make the most important words have more impact. The embedding process is a lookup table which is shown in equation 3.

$$\begin{aligned} E_{w1} : \mathbb{R} &\rightarrow \mathbb{R}^{D_w} \\ x &\rightarrow W_w e_x \end{aligned} \tag{3}$$

where the D_w is dimensions of word vector, $W_w \in \mathbb{R}^{|W| \times D_w}$ is a parameter to be learned by the model and initialized randomly. e_x is an one-hot vector with element indexed by x being 1 and other elements set as 0.

$$v_{exp} = LSTM_u^K(E_{w1}(w_i)) \tag{4}$$

we name the v_{exp} as explicit perference vector which is the output of user-LSTM at time K.

Implicit aspect: v_{exp} encode the user perferences from textual reviews, however, some perferences relate to other information which is not provided in the AVA-PCap dataset. In order to increase the capacity of preference fitting, we let the network to learn hidden perference parameters by itself. we use the similar embedding function E_u to get the implicit perference vector.

$$v_{imp} = E_u(uid) \tag{5}$$

where the uid is the user identity and E_u has the similar parameters W_u with different shape $W_u \in \mathbb{R}^{|U| \times D_u}$. Our user vector is represented as $v_{user} = [v_{imp}||v_{exp}]$.

E. Personalized Caption Model

Personalized Caption Model utilizes the vectors $f_{mlsp}, f_{local}, v_{user}$ to generate reviews for every image-user pairs. The caption model in our method is based on encoder-decoder framework [18] [19]. Considering user perference can have different influence in encoder process and decoder process, we first extract two different type of vectors from v_{user}:

$$\mathbf{v_{voc}} = \sigma_l(\mathbf{W_{voc}} \mathbf{v_{user}} + b_{voc}) \tag{6}$$

$$\mathbf{v_{vis}} = \sigma_l(\mathbf{W_{vis}} \mathbf{v_{user}} + b_{vis}) \tag{7}$$

where the σ_l are leaky relu function and $\mathbf{W}_*, b_*$ is parameters of this layer. we let D_{voc} and D_{vis} be the dimisions of $\mathbf{v_{voc}}$ and $\mathbf{v_{vis}}$ which represent the vocabulary perferences and visual perferences of a user. Next, we introduce the encode and decode module and how the $\mathbf{v_{vis}}$ and $\mathbf{v_{voc}}$ be used.

Encoder module: In the encoder part of PAIC, v_{vis} and f_{mlsp}, f_{local} are used to encode the image into a visual signal which is the initial state of our decoder module. Inspired by [20], we transform v_{vis} to a mask vector which every elements between $[0, 1]$ and has the same dimisions as f_{mlsp}, then multiply the mask vector to f_{mlsp} element-wise.

$$\mathbf{v_{mlsp}} = \sigma_s(\mathbf{W_{mlsp}} \mathbf{v_{vis}} + b_{mlsp}) * \mathbf{f_{mslp}} \tag{8}$$

where the σ_s is the sigmoid activation function. Because every element of mlsp feature encode different level information, the multiplication assigns different attention score to each levels according to the user preferences. The final visual signal is:

$$\mathbf{v_{signal}} = [\mathbf{v_{mlsp}}||\mathbf{f_{local}}] \tag{9}$$

Decoder module: the object of decoder module is to predict the word at each time step. There are many decoder based on differnt network in the prior work [1] [9], and the most common one is based on LSTM which is alse the one we used in our approach. We use anthor word embedding function E_{w2} similar to Equation 3 to vectorize the words in decoder-lstm. In the following texts, we represent the output vector of decoder-LSTM at time step t as:

$$DLSTM(input)^t. \tag{10}$$

and represent the gt_i as the i-th word id in the ground-truth review sentences. The training process of the decoder is the following formular, the t is $1, 2, \ldots, n$:

$$\mathbf{x_0} = \mathbf{v_{signal}} \tag{11}$$

$$\mathbf{x_t} = E_{w2}(gt_t) \tag{12}$$

$$\mathbf{o^t} = DLSTM(\mathbf{x_{t-1}})^t \tag{13}$$

$$\mathbf{off^t} = MLP(\mathbf{o^t}||\mathbf{v_{voc}}) \tag{14}$$

$$\mathbf{p^t} = \mathbf{o^t} + \mathbf{off^t} \tag{15}$$

where the MLP is a multi-layer perception to mine the relation between the $\mathbf{o^t}$, $\mathbf{v_{voc}}$ and $\mathbf{off^t}$. The thought of general vector plus offset vector comes from other personalized network applied in aesthetic assessment task [16]. Note that, we add special EOS(end of sequence) token in the end of $\mathbf{gt}$ before the training.

The training loss is a function of the $\mathbf{p^t}$:

$$L = \sum_t -\mathbf{log}(\mathbf{softmax}(\mathbf{p^{t-1}}))[gt^t] \tag{16}$$

the $[*]$ means get the element of vector by the index.

During the test period, the $\mathbf{x_t}$ is the word id predicted in the $t-1$ step rather than the gt. Word at t time step is predicted by :

$$w^t = \mathbf{argmax}(\mathbf{softmax}(\mathbf{p^t})) \tag{17}$$

If the predicted word is EOS, then finish the process.

IV. Experiment

A. Datasets

Table III summarize the quantitative results of baselines and our PAIC model. We evaluate our proposed model on a new dataset called AVA-PCap. The Raw AVA-PCap(RAVA-PCap) dataset is a middle sized personalied aesthetic image caption dataset containing 39180 images, 13037 users and 453528 reviews, where the images and reviews is collected from DPChallenge website. DPChallenge is a great source for aesthetic information, because lots of professional photographers post their masterpieces with lots of reviews from different aesthetic aspects given by photography peers.

To assure the quality of images and save time, our crawler program reuses the AVA dataset image list and only download the reviews and user information. We get the RAVA-PCap in a week and our AVA-PCap will be extended in the future to cover more images.

Next we process a series of filtering and cleaning. Our task is related to the user activity, so we filter the inactive user by setting a threshold(=30) and only keeping the user whoes reviews is more than the threshold. After noticing many reviews is too simple and too short to contain aesthetic information, we follow the strategy of [6] to filter these valueless comments. The strategy filter the sentences by assigning every sentence a score to measure the aesthetic value. After the above process, we get a clean AVA-PCap dataset. The above filtering strategy may be a little strict, AVA-PCap only contains 8288 images, so we provide download link for both AVA-PCap and Raw AVA-PCap. Table II show the details of the two datasets.

B. Experimental Setup

Because of the differences between AVA-PCap dataset and normal image caption dataset, We use the following two principles to split our AVA-PCap into test set and training set. Firstly, we want to ensure high activation of users in test set. Secondly, we guarantee that the images in test set never appeal in training set. To achieve the two principles, we rank the users by review number and select the top-N activest users where N is the size of the test set and we let it be the 10% of tot images. For each selected users, we sort the commented images, and select the image that is with the least comments and never selected before. By this way, we can get our test set which contains N pairs of user and image and satisfy the two principles. The rest pairs unrelated with the images in test set are used for training. Finally, we get 30722 pairs for training and 820 pairs for test(because the number of images are around 8200).

To quantitatively compare our method with baselines, we compute the language similarity between predicted sentences and ground truths. We calculate BLEU [21], CIDEr [22], ROUGE-R [23] and METEOR [24] as scores. In the aforementioned metrics, higher scores indicate better performance. Different from MSCOCO [25] or Flick30K [26], in our test set we just select one reviews per images, which means our GTs just contains one review for each image.

When we do experiments, some tricks are used to improve the final results. We transform the least used words to UNK token to reduce the size of vocabulary and we also truncate long sentences. In our experiments, the threshold of length is 15.

C. Baselines

We select several image caption methods as baselines containing several nearest neighbor approches and previous image caption methods. We also select two variants of our model as ablation study.

1NN-*: This set of methods contain three variants. Both of the methods encode the images with the output of average pooling layer of Resnet101. In 1NN-Image, we randomly select a sentence of the image with smallest Euclidean distance to the given image. In 1NN-User, we randomly select a review given by the same user. 1NN-UserImage is a baseline that collects the images commented by the same user, and then find the reviews of nearest image along the collected images as output.

ShowTell and ShowAttendTell: These methods are proposed for image caption. We use the implementation of ImageCaption.pytorch codebase [27].

CWS: This method is proposed for aesthetic image caption. LDA topics are computed to supervise the training of Resnet. The decoder part of the model is the same with ShowTell caption model.

PAIC-FULL: This approach is our PAIC model with complete aforementioned submodule. To reduce the bias, we uniform the evaluation code to compute the metric score. We use the evaluation code from MSCOCO dataset.

PAIC-noUEN: This baseline come from our PAIC-FULL model. we cut off the UEN network and make our model unrelated with user id. Comparing with PAIC-noUEN, we can find the effectiveness of our UEN module.

PAIC-noAEN: This baseline also come from our PAIC-FULL model. we cut off the AEN network and the image feature used in this model is output of resnet101. Comparing with PAIC-noUEN, we can find the effectiveness of our AEN module.

In ablation study, we will introduce the last two variants of our model. Table III lists all the above baselines, and the * means the model is personalized. Bode number means the best.

D. Comparative analysis and ablation study

Table III summarize the quantitative results of aesthetic image caption. First of all, we can easily see that our PAIC-FULL model significantly outperforms other baselines in almost all metrics. Secondly, we can note that the prediction scores of 1NN-User exceeds the scores of 1NN-Image in all aspects, which means that user information is significant to the prediction of comments. Comparing results of 1NN-User and 1NN-UserImage, we can also learn that the similarity of normal Resnet101 feature can't provide better performance, but comparing PAIC-noUEN with ShowTell, the effectiveness

TABLE III
EXPERIMENTS, * MEANS PERSONALIZED MODEL

Method	B-1	B-2	B-3	B-4	METEOR	CIDEr	ROUGE-L
1NN-Image	0.073	0.018	0.004	0.001	0.035	0.042	0.062
1NN-User*	0.104	0.039	0.021	0.013	0.050	**0.137**	0.089
1NN-UserImage*	0.100	0.037	0.017	0.009	0.048	0.108	0.089
ShowTell [1]	0.111	0.053	0.028	0.007	0.038	0.111	0.035
ShowAttendTell [2]	0.125	0.059	0.031	0.010	0.045	0.065	0.129
CWS [6]	0.127	0.060	0.032	0.013	0.045	0.053	0.124
PAIC-noUEN	0.132	0.060	0.031	0.012	0.045	0.063	0.121
PAIC-noAEN*	0.126	0.054	0.029	0.008	0.041	0.120	0.128
PAIC-FULL*	**0.142**	**0.069**	**0.035**	**0.015**	**0.058**	0.094	**0.135**

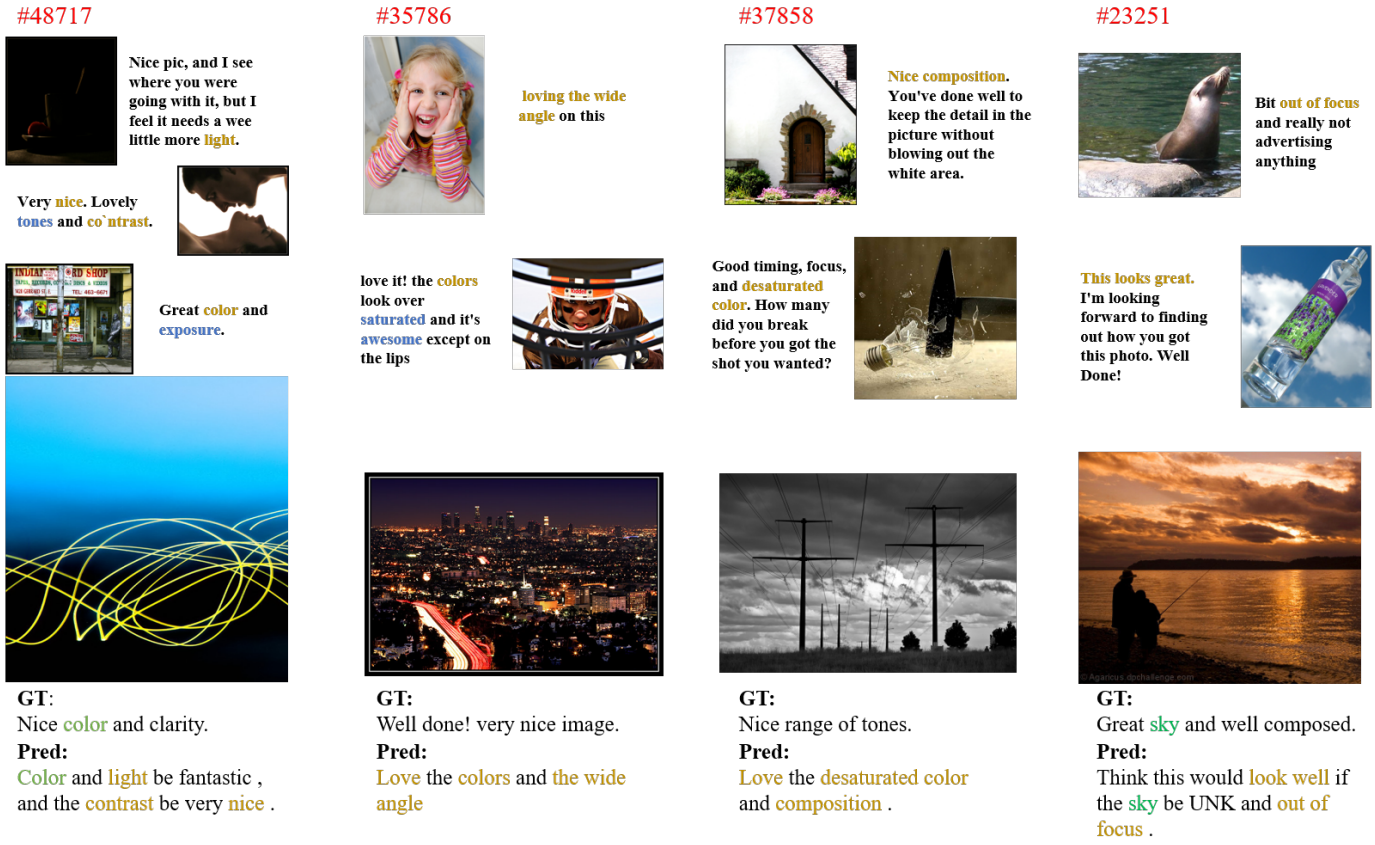

Fig. 4. Cases from our AVA-PCap dataset. In every column, the above images is commented by the same user in training set. The nether image is the predicted image. Green word is the correct word in groud truth. Yellow word means this word not appear in GT, but it is related to the perferences of users in train set.

of our AEN model is proven and our aesthetic model truly extract the better feature from our aesthetic photo.

We can easily know that, after applying attention mechanism, ShowAttendTell outperforms ShowTell, which means that attention of image region can improve the results of model in this task. This can also explain why our PAIC model works: In our UEN model, our visual part of encoder can be thought as a attention model, and it attends on the visual level of images, i.e. the color, shape or objects of the image. This can bring advantages for predicting.

As ablation study, we can know personalization is more important for comment prediction task than image feature extraction. Also, we can know our proposed multi-level feature can describe a photo in a more accurate way.

E. Case study

Cases are showed in figure 4. We can see that PAIC model does exploit the personalized information: PAIC model combines the related review segments of given user and aggregates them to a whole review. Such as in #35786, the second column of figure 4, groud truth gives meaningless sentences, but our PAIC model gives more information about this images based on the user's pervious remarks. By the cases, we realize that a better aesthetic filtering strategy is critical for better

performance because meaningless comments do weanken the effectiveness of our PAIC model.

V. CONCLUSION

After analyzing the aesthetic image caption task, we find two main defects and find way to address them. Realizing the importance of personalization, We propose our PAIC model to solve the aesthetic image caption task. We also compile a AVA-PCap dataset based on images of AVA and do experiments on the dataset. With quantitative evaluation, we show that our PAIC model outperform several baselines. Our experiments and other arguments support the proposed defects of previous methods. There are several promising future directions that go beyond this work. Firstly, we can introduce the image region attention mechanism in our PAIC model to improve it. Secondly, AVAP-Cap is a little small comparing to other dataset in aesthetic image caption task, which leads to the low scores in our experiments, so a larger dataset should be constructed or a more effecient aesthetic reviews filter approach should be proposed.

REFERENCES

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[2] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[3] S. He, H. R. Tavakoli, A. Borji, and N. Pugeault, "Human attention in image captioning: Dataset and analysis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8529–8538.

[4] L. Ke, W. Pei, R. Li, X. Shen, and Y.-W. Tai, "Reflective decoding network for image captioning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8888–8897.

[5] K.-Y. Chang, K.-H. Lu, and C.-S. Chen, "Aesthetic critiques generation for photos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3514–3523.

[6] K. Ghosal, A. Rana, and A. Smolic, "Aesthetic image captioning from weakly-labelled photographs," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[7] N. Murray, L. Marchesotti, and F. Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2408–2415.

[8] V. Hosu, B. Goldlucke, and D. Saupe, "Effective aesthetics prediction with multi-level spatially pooled features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9375–9383.

[9] C. Chunseong Park, B. Kim, and G. Kim, "Attend to you: Personalized image captioning with context sequence memory networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 895–903.

[10] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[11] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng, "Stylenet: Generating attractive visual captions with styles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3137–3146.

[12] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *European conference on computer vision*. Springer, 2006, pp. 288–301.

[13] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 990–998.

[14] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rating image aesthetics using deep learning," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2021–2034, 2015.

[15] Z. Wang, S. Chang, F. Dolcos, D. Beck, D. Liu, and T. S. Huang, "Brain-inspired deep networks for image aesthetics assessment," *arXiv preprint arXiv:1601.04155*, 2016.

[16] J. Ren, X. Shen, Z. Lin, R. Mech, and D. J. Foran, "Personalized image aesthetics," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 638–647.

[17] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242. Piscataway, NJ, 2003, pp. 133–142.

[18] I. Sutskever, O. Vinyals, and Q. Le, "Sequence to sequence learning with neural networks," *Advances in NIPS*, 2014.

[19] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1243–1252.

[20] A. Veit, S. Belongie, and T. Karaletsos, "Conditional similarity networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 830–838.

[21] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[22] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[23] C.-Y. Lin and F. J. Och, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2004, p. 605.

[24] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[26] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[27] R. Luo, "An image captioning codebase in pytorch," https://github.com/ruotianluo/ImageCaptioning.pytorch, 2017.