Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

Collaborative and Attentive Learning for Personalized Image Aesthetic Assessment

Guolong Wang[1], **Junchi Yan**[2] and **Zheng Qin**[1*]
[1] BNRist, School of Software, Tsinghua University, China
[2] Shanghai Jiao Tong University
wanggl16@mails.tsinghua.edu.cn, yanjunchi@sjtu.edu.cn, qingzh@tsinghua.edu.cn

Abstract

The ever-increasing volume of visual images has stimulated the demand for organizing such data by aesthetic quality. Automatic and especially learning based aesthetic assessment methods have shown potential by recent works. Existing image aesthetic prediction is often user-agnostic which may ignore the fact that the rating to an image can be inherently individual. We fill this gap by formulating the personalized image aesthetic assessment problem with a novel learning method. Specifically, we collect user-image textual reviews in addition with visual images from the public dataset to organize a review-augmented benchmark. Using this enriched dataset, we devise a deep neural network with a user/image relation encoding input for collaborative filtering. Meanwhile an attentive mechanism is designed to capture the user-specific taste for image semantic tags and regions of interest by fusing the image and user's review. Extensive and promising experimental results on the review-augmented benchmark corroborate the efficacy of our approach.

1 Introduction

With the continuously generated and ever-expanding volume of visual images, automatic image aesthetics assessment is increasingly important in many applications e.g. image retrieval and editing, content management and photography [Datta *et al.*, 2007; Marchesotti *et al.*, 2015; Lu *et al.*, 2015b]. Among various vision problems, there is a particular challenge to assess an image as 'good' or 'bad', due to the highly subjective and complex nature of human aesthetic preference.

Many research efforts have been made to address the computational image aesthetic quality assessment problem. A major line of research formulates the image aesthetic prediction as classification or regression problem that map the image to users' ratings [Datta *et al.*, 2006; Nishiyama *et al.*, 2011] etc. Among them, early attempts are based on the intuition of how an image is perceived, which design hand-crafted features by following the standard photographic rules of visual design such as the rule of thirds, the golden ratio, and color harmonies[Tong *et al.*, 2004; Ke *et al.*, 2006; Dhar *et al.*, 2011; Bhattacharya *et al.*, 2010; Zhang *et al.*, 2013]. Despite the success of handcrafted features, recent work show that deep convolutional neural network (CNN) feature based methods [Kang *et al.*, 2014; Lu *et al.*, 2015b; Dong *et al.*, 2015; Tian *et al.*, 2015; Lu *et al.*, 2015a] can yield leading performance on public benchmarks.

One key complication of aesthetic analysis compared with other vision tasks is that different people may have diverse conclusions on the same image based on their subjective preferences. In [Segalin *et al.*, 2016], pictures are mapped into attributed human traits defined as the Big-Five factor structure [Goldberg, 1990]. However, human traits refer to common group characteristics rather than individual ones. So far little personalization technique has been introduced to image aesthetic assessment problem, which we believe is the essence of this problem and plays as the key motivation of this paper.

Such limitation also exists at existing public image aesthetic assessment datasets, including the popular Aesthetic Visual Analysis (AVA) dataset [Murray *et al.*, 2012] and the Aesthetics and Attributes Database (AADB) dataset [Kong *et al.*, 2016]. In fact, these datasets only contain visual images without any textual reviews that can semantically reflect users' preferences. Moreover, the used ground truth for performance evaluation is often set as the average of individual users' ratings of an image, making the evaluation protocol inherently personalization-free.

We enrich the AVA benchmark by crawling its user-specific textual reviews and derive the user-specific ratings to each image. It is important to note that review comments can reflect users' personalized tastes. As such, we re-create a review-augmented version of AVA with multi-modal data (visual image and textual review), as well as derived personalized opinions. Note that [Zhang, 2016] developed a deep learning architecture for image aesthetic assessment, where image semantic tags were used in addition to image content. However, no (unstructured) text data was involved.

In this work, we propose a novel personalized multi-task

*indicates the corresponding author

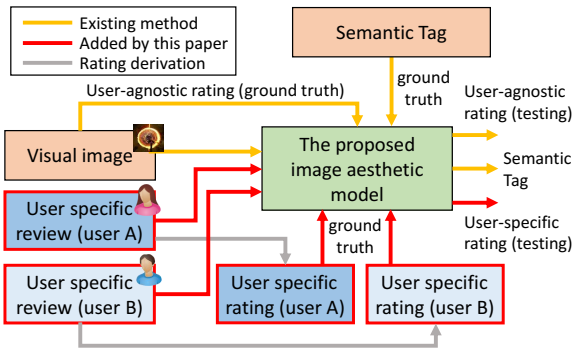

Figure 1: Working flow and difference of our model against existing pure visual image based methods. Our personalized model is multi-modal by using user review data.

image aesthetic assessment approach. User textual reviews, visual content of images and their semantic tags are simultaneously modeled based on the above motivation and intuition. We identify the contributions as follows:

i) we address the personalized user-specific image aesthetic assessment learning task. To our best knowledge, this is the first work for formally formulating this problem. We also create the review-augmented image aesthetic assessment benchmark based on the Aesthetic Visual Analysis (AVA) dataset [Murray *et al.*, 2012].

ii) We present a novel network for integrating multi-modal information including image, review, and user-image relation for (personalized) image aesthetic assessment learning. In particular, attentive mechanism is devised for capturing users' personalized taste (reflected by their reviews) to an image w.r.t. the area of interest as well as the semantic tag of the images. The attention is fulfilled by the joint modeling of images and user-specific reviews.

iii) In addition with the new capability of predicting personalized user-image ratings, we also achieve the state-of-the-art performance on traditional user-agnostic image aesthetic assessment benchmark.

2 Problem Settings

Review-augmented AVA As discussed above, we enrich the AVA dataset [Murray *et al.*, 2012] by re-collecting the user-specific reviews for an image which is tagged with its semantic tags such as *family*, *landscape* etc. There are in total 66 semantic tags in AVA and in total over 250,000 images.

For each image, the data source of the AVA dataset, i.e. the website (http://www.dpchallenge.com/) in fact contains both user-image ratings and user-image text reviews. For each review, the user id is available. However, for the raw rating score (with 10), its user information is de-identified and anonymous thus it is only possible to obtain the rating distribution but unable to link each raw rating to a corresponding user id (and its reviews). As a tentative effort, for each user's review, we try to re-generate users' ratings to the image and assign a binary label ('good' or 'bad'), if the number of positive words is larger than the negative ones in the review to certain extent. Fig. 2 shows two image examples with their

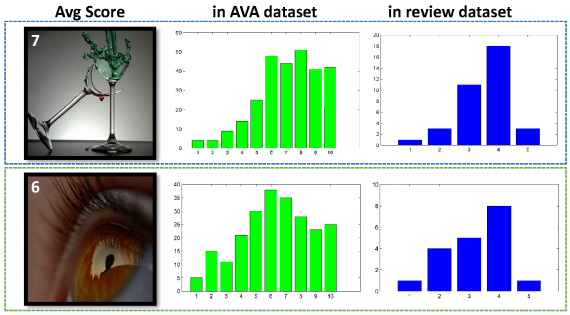

Figure 2: Two examples of the images in the AVA dataset: 1) raw image; 2) anonymous rating's distribution for the image from the raw dataset; 3) re-generated rating's distribution from the text review data. Two observations are made: 1) the raw rating's distribution is similar to our re-generated one which verifies our derived user-dependent ratings are reasonable; 2) the distributions can be either gaussian or non-gaussian which showcases the complexity and diversity of the ratings, and the need for a personalized model.

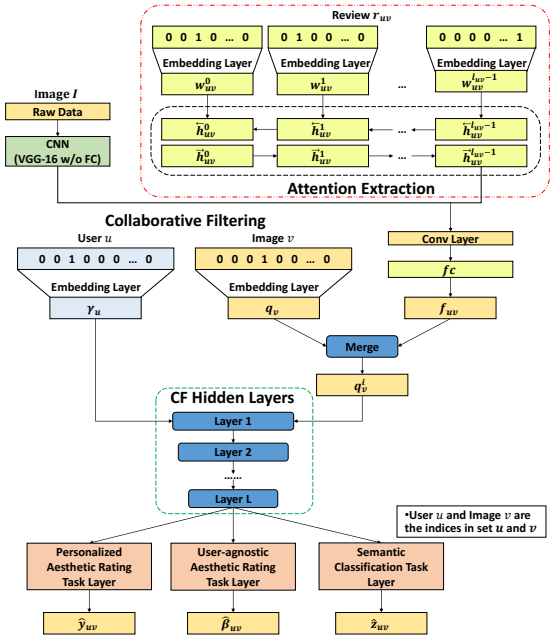

Figure 3: Attentive collaborative filtering network (CFAN) for personalized aesthetic assessment. Each word in the textual reviews $\mathbf{r}_{uv}$ is first one-hot encoded by the word dictionary, the embedding $\mathbf{w}_{uv}$ then is fed into a bidirectional GRU recurrent neural network. Its output is convolved with the image CNN feature to generate the user-specific feature $\mathbf{f}_{uv}$ (see more illustration in Fig. 6). Finally the aesthetic rating task and semantic tagging task are learned jointly to benefit each other. The rest part is the same with the CF network.

rating distributions: one is from de-identified raw score, the other is from the re-generated user-specific score from their review by our approach. In fact, we are able to reach 81% consistency between the raw ratings and the derived ratings from reviews for each image in average – see more details in the experiment section about the comparison protocol.

Metrics	Portrait	Still Life	Abstract	Landscape	Nature	Overa
Avg. review # per image	9.031	7.464	7.542	8.214	7.733	7.92
Avg. user # per image	9.011	7.455	7.525	8.201	7.718	7.91
Avg. review # per user	9.977	8.748	10.648	12.725	12.885	35.2(
Range review # per image	[1, 125]	[1, 83]	[1, 114]	[1, 110]	[1, 105]	[1, 8
Range user # per image	[1, 123]	[1, 87]	[1, 116]	[1, 111]	[1, 109]	[1, 8
Range review # per user	[1, 868]	[1, 680]	[1, 800]	[1, 1158]	[1, 1197]	[1, 396
Image #	73071	74324	97117	109623	115776	2555

Table 1: Partial statistics of AVA review dataset.

Table 1 discloses its preliminary statistics: for each imag at least there is one review text. The overview of the prc lem and working pipeline of our method is sketched in F 1, which highlights the difference from existing problem s ting: reviews are collected as extra inputs and user person ized ratings are additional outputs.

Problem & Notations Specifically, we integrate both vi- sual image and associate user-specific reviews into a deep net- work for personalized image aesthetic assessment. Attention model is developed to attend to both particular image areas and visual attributes, to capture the users' personalized taste. First we introduce notations.

Denote users as $\mathbf{u} = \{u_1, u_2, \cdots, u_m\}$, and images as $\mathbf{v} = \{v_1, v_2, \cdots, v_n\}$, the binary user-image matrix is defined where 1-entry means the image is reviewed by a user and 0 otherwise. User u's review on image v is defined as r_{uv} with binary aesthetic rating s_{uv} ('good' or 'bad'). The word list in r_{uv} is defined as $\mathbf{w}_{uv} = \{w_{uv}^0, w_{uv}^1, \cdots, w_{uv}^{L_{uv}-1}\}$, where L_{uv} is the length of the review text. The attention map which means the focus of user u on image v is denoted by $\mathbf{a}_{uv}$. The original visual features of image v is defined as $\mathbf{f}_v$ and the user specific one is $\mathbf{f}_{uv}$.

Given all user specific visual features $\mathcal{F} = \{\mathbf{f}_{uv} | u \in \mathbf{u}, v \in \mathbf{v}\}$, reviews $\mathcal{W} = \{\mathbf{w}_{uv} | u \in \mathbf{u}, v \in \mathbf{v}\}$, ratings $\mathcal{S} = \{s_{uv} | u \in \mathbf{u}, v \in \mathbf{v}\}$, and image tag (e.g. portrait, landscape) $\mathcal{T} = \{t_1, t_2, \cdots, t_l\}$, for a target user u and an unseen image $v_o \in \mathbf{v}$, the task is to find a function $g(s_{uv_o} | u, \mathcal{W}, \mathcal{S}, \mathcal{F}, \mathcal{T})$ to predict if user u would rate 'good'/'low' to v_o where s_{uv_o} is the predicted rating.

3 Proposed Models and Algorithms

In this section, we will introduce our network which involves standard building blocks like CNN (VGG-16) [Simonyan and Zisserman, 2014], multi-task loss [Caruana, 1997; Kao *et al.*, 2017]. In particular, the personalization is fulfilled by a collaborative filtering network component that infuses the user and image relation information. Using this model as a starting point, we further add users' textual review data and correspondingly develop an attentive mechanism to bridge the visual images and textual reviews. Importantly, this mechanism can capture the particular region of interest in an image as well as the latent factors of semantic tags individual for rating an image.

3.1 Personalized Collaborative Filtering Attentive Aesthetic Assessment Network

As shown in Fig. 3, for image v reviewed by user u, we first embed its index in set $\mathbf{v}$ (by one-hot encoding vector) into an embedding layer with output $\mathbf{q}_v$. Then $\mathbf{q}_v$ is combined with

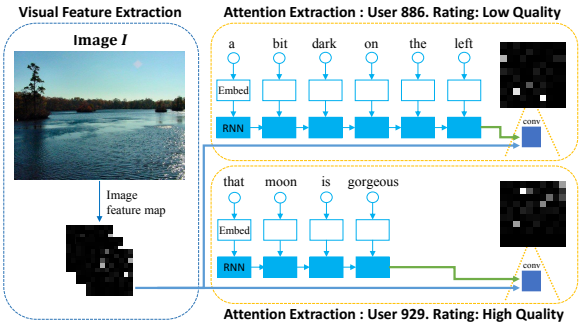

Figure 4: Different feature maps extracted with attention maps generated by two users' reviews. User ID.886 focused on the tree and its reflection on the left bottom of the image while user ID.929 focused on the moon in the right top.

the image visual content feature $\mathbf{f}_v$ (weighted by $\mathbf{W}^i$) in to a merger operation by:

$$\mathbf{q}_v^i = Merge(\mathbf{q}_v, \mathbf{W}^i \cdot \mathbf{f}_v) \tag{1}$$

where $Merge(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a function that merges two d dimension vectors into one. Element-wise multiplication is used in this paper while other mappings can be used depending on specific applications.

The latent variable $\mathbf{q}_v^i$ is combined with the user-wise index embedding $\boldsymbol{\gamma}_u$ to generate the binary prediction $\hat{y}_{uv}$:

$$\widehat{y}_{uv} = f(\boldsymbol{\gamma}_u, \mathbf{q}_v^i) \tag{2}$$

where $f(\cdot, \cdot)$ is interaction function learned through hidden layers.

For the CNN structure used for image content feature $\mathbf{f}_v$ extraction, VGG-16 [Simonyan and Zisserman, 2014] pretrained on the 1000-class ImageNet classification challenge 2012 dataset [Deng *et al.*, 2009] is adopted (by removing its fully connected layers). Then we use bidirectional GRU [Bahdanau *et al.*, 2014] model to combine the word embedding with the CNN visual features. $\overrightarrow{GRU}$ reads the review r_{uv} from w_{uv}^0 to $w_{uv}^{l_{uv}-1}$ while $\overleftarrow{GRU}$ reads reversely.

$$\begin{aligned} \overrightarrow{h}_{uv}^i &= \overrightarrow{GRU}(w_{uv}^i), i \in [0, L_{uv} - 1] \\ \overleftarrow{h}_{uv}^i &= \overleftarrow{GRU}(w_{uv}^i), i \in [L_{uv} - 1, 0] \end{aligned} \tag{3}$$

As a result the annotation for word w_{uv}^i can be concatenated of forward hidden state $\overrightarrow{h}_{uv}^i$ and backward one $\overleftarrow{h}_{uv}^i$:

$$\mathbf{h}_{uv} = [\overrightarrow{\mathbf{h}}_{uv}, \overleftarrow{\mathbf{h}}_{uv}] \tag{4}$$

The annotation $\mathbf{h}_{uv}$ is projected to visual space from contextual information as the convolutional kernel $\mathbf{K}$ by $\mathbf{K} = \sigma(W_k \mathbf{h}_{uv} + b_k)$, highlighting the visual features focused by users (reflected by their reviews). $\mathbf{K}$ has the same number of channels as the visual feature $\mathbf{f}_v$. The review-image attention map is calculated by:

$$a_{uv}^{ij} = \frac{e^{f_{uv}^{ij}}}{\sum_i \sum_j e^{f_{uv}^{ij}}} \tag{5}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$, and $\mathbf{f}_{uv} = \mathbf{K} * \mathbf{f}_v$ is the convolution of the kernel $\mathbf{K}$ with $\mathbf{f}_v$. Note a_{uv}^{ij} is the element of the attention

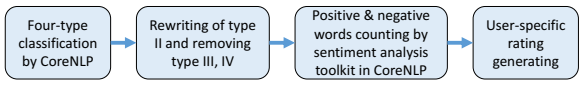

Figure 5: Data preprocessing for recovering user-specific ratings for each image from raw textual reviews. See main text for the definition of type I, II, III, IV.

map at position (i, j), and $\mathbf{a}_{uv}$ has the same size as $\mathbf{f}_v$, It characterizes the attention distribution across feature map.

We conjecture the attention framework can be useful in two aspects: 1) It can locate subjects in the photo and features in less relevant regions would be filtered out; 2) It can also focus on personalized semantic patterns (e.g. colorfulness, lighting) of a specific user. Fig. 4 shows such diversity for attention maps among different users. Note the attention map is user-specific.

Finally a multi-task learning part is added to the attentive network. The hidden layers and convolutional layers in this network can be regarded as shared layers, the added fully-connected layers in multi-task learning part are task-specific layers. The final loss function is:

$$\begin{aligned}\ell_{cfa} = &\sum y_{uv} \log \hat{y}_{uv} + (1 - y_{uv}) \log(1 - \hat{y}_{uv}) \\ &+ \sum \lambda_z \left(z_{uv} \log \hat{z}_{uv} + (1 - z_{uv}) \log(1 - \hat{z}_{uv})\right) \\ &+ \sum \lambda_\beta \left(\beta_{uv} \log \hat{\beta}_{uv} + (1 - \beta_{uv}) \log(1 - \hat{\beta}_{uv})\right) \quad (6)\end{aligned}$$

where λ_z and λ_β are the weighting hyper-parameters, $\hat{y}_{uv}$ is the prediction of user-specific rating task, $\hat{z}_{uv}$ is the prediction of semantic tag classification task, and $\hat{\beta}_{uv}$ is the prediction of user-agnostic rating task.

4 Experimental Results and Discussion

We evaluate the proposed method on one of the most large-scale and challenging datasets, i.e. AVA dataset for visual aesthetic quality assessment (augmented by users' reviews). It contains more than 255,000 images gathered from www.dpchallenge.com, with each image tagged by 0 to 2 semantic tags. There are around 200 users giving a rating score ranging from 1 to 10 for each image.

We follow the protocol by the state-of-the-art work [Kao *et al.*, 2017] to identify 29 major semantic tags (e.g. *abstract*, *fashion*, *family*, *sky*, *sports*) with 185,751 images used for our evaluation: i) the chosen tag has at least 3000 images; ii) the chosen image has at least one tag. If an image has two tags, the primary tag will be set as its tag in line with the protocol used by the state-of-the-art [Kao *et al.*, 2017].

4.1 Protocol & Data Preparation

Review Augmentation

We try to recover the user-specific ratings which are missing in the AVA dataset. The main idea is to explore the text reviews to derive the underlying ratings. There are some challenges in extracting semantic ratings from textual reviews which consist of many short sentences. Moreover, the reviews may not well follow the grammar in written form. We design the following empirical protocol to recover user-image ratings, as illustrated in Fig. 5.

- **Review classification** To identify useful reviews, we first classify reviews into four types by the tool CoreNLP [Manning *et al.*, 2014]: I) comments based on the photographic rules like "*The color is wonderful and the composition is great!*". II) users' subjective feelings like "*I like this photo!*". III) ambiguous not literally related to image aesthetic quality like "*How handsome the boy is!*" IV) only contain equivocal messages like "*Do you live in the steeple?*" The part-of-speech tagger, the parser and the open information extraction tools in CoreNLP are used to analyze the grammatical structures of sentences and extract relation tuples of words.
- **Review cleaning** The reviews in the type II are rewritten as "*The photo is good (resp. bad)*" and those in the type III and type IV are removed due to their inherent ambiguity.
- **Review sentiment counting** The high-frequency words are sorted and those with the highest relevance to photo aesthetic quality assessment (e.g. shot, composition, focus, light, angle, color, exposure) are selected. Then positive and negative words in reviews are counted by the sentiment analysis toolkit in CoreNLP, by which the reviews are divided into five levels of user's preference.
- **Rating generation** Since traditionally existing image assessment work almost use binary rating as ground truth [Murray *et al.*, 2012], we derive the binary ratings $s_{uv} \in \{0, 1\}$ for each review-to-image by setting the first two levels as 'bad', and the other three as 'good'.

Validity of Recovered User-specific Ratings

We perform a cross-check to verify the reliability of our recovered binary rating s by indirectly comparing them with the raw de-identified ratings ϕ from 1 to 10 as recorded in the AVA dataset. Note that there is no exact one-to-one matching between the review-recovered ratings and the raw ratings because sometimes a user may write a review with no rating and vice versa. Hence we try to compute the positive vs. negative ratio consistency between ϕ and s. For each image v, we first compute its overall binary rating, i.e. user-agnostic score b_v – this is also the tradition widely adopted in the AVA dataset related studies [Murray *et al.*, 2012]: let $b_v = 1$ if the average of image v's K ratings $\{\phi_v\}_{k=1}^K$ is larger than 5 otherwise $b_v = 0$. Then we compute the ratio of the 'good' samples against 'bad' ones for the recovered binary scores $\{s_v\}$. More specifically the consistency ratio score is defined as follows for each image ($\#(\cdot)$ indicates the number):

$$\mathcal{C} = \sum_{v, b_v == 1} \frac{\#(\{s_v == 1\})}{\#(\{s_v\})} + \sum_{v, b_v == 0} \frac{\#(\{s_v == 0\})}{\#(\{s_v\})} \quad (7)$$

By dividing the above number to the total number of image, we reach a consistency of 81.06% which we believe shows our recovered ratings are reasonable. As such, the review dataset contains 1,163,258 positive review derived ratings ($s_{uv} = 1$) and 271,862 negative ones ($s_{uv} = 0$).

Experiment Settings

We use Fig. 6 to illustrate the detailed setting of our model. More detailed description can be found in the caption. The input raw images are resized to 320×320, and the feature map

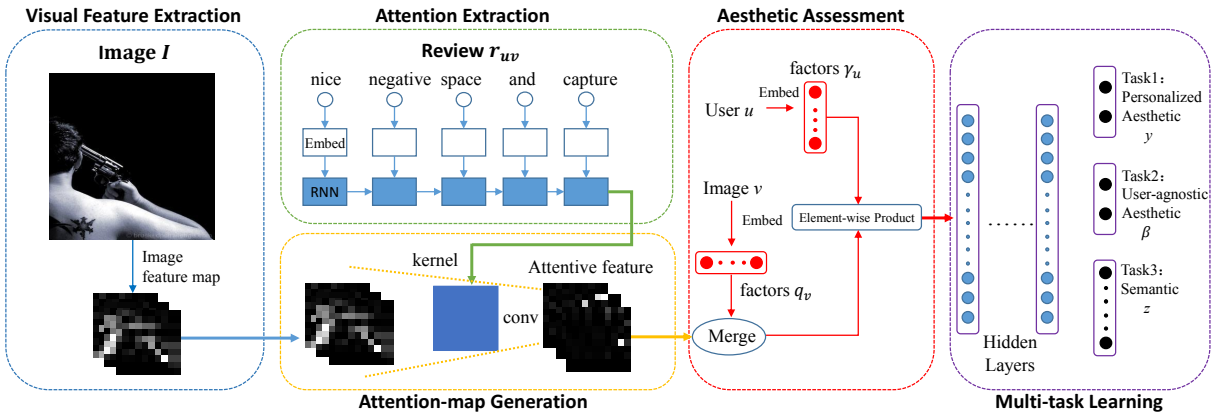

Figure 6: Detailed view of our approach. Blue box is used to extract CNN feature maps from raw images. Green box denotes the attention extraction phase. In yellow box, the output of the RNN (GRU) is transformed to a kernel convoluted with CNN feature maps to generate attention-map. Red box denotes the personalized processing phase, where the attention-map is merged with image index one-hot encodings to derive a new embedding, which is then multiplied by the user index encoding to generate both the image semantic and aesthetic labels, in the sense of multiple task learning as shown in the purple box.

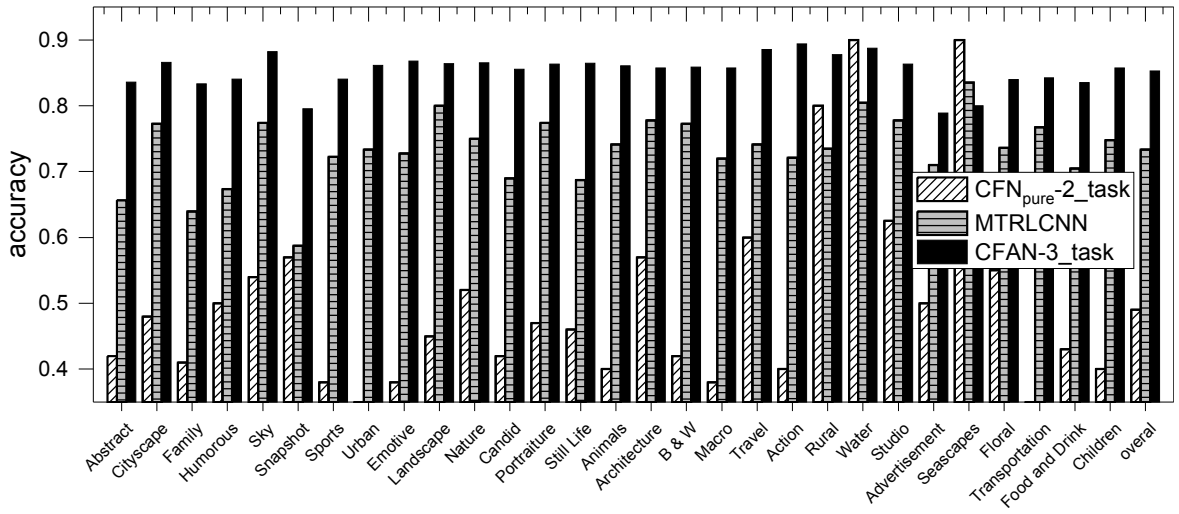

Figure 7: Evaluation for user-agnostic image aesthetic assessment on the 29 major tags. Rightmost bar is the overall accuracy.

is reduced into $10\times10\times512$ after five pooling layers in VGG-16. The dimension of the attention-map is $10\times10\times512$ which is the same as the feature map. The merging of attention-map and image feature is through transform matrix according to a normal distribution. The user encoding and image embedding have the same dimension and after element-wise product the result can be used to predict the aesthetic ratings for different users. In all of our networks, the VGG-16 part is initialized by pretraining on ImageNet, and the resting parameters are randomly initialized according to a standard normal distribution $N(0, I)$, and then updated by executing stochastic gradient descent (SGD). The hyper-parameters in our models are tuned by conducting 10-fold cross validation on the training set. We set 90% of the data as training set, and the rest is testing set. Specifically, the weighting parameter λ_β and λ_z in Eq. 6 is set to balance the contribution of semantic classification loss and aesthetic rating loss. We set the original learning rate as 0.005, the decay rate as 0.99, the decay step as 1000. k is set as 50, λ_β is set as 1, and λ_z is set as 0.015.

4.2 Results & Discussion

User-agnostic Aesthetic Assessment

In the previous study, image aesthetic assessment is performed at the aggregated level, including the state-of-the-art **MTRLCNN** (**M**ulti-**T**ask **R**elationship **L**earning **C**onvolutional **N**eural **N**etwork) [Kao *et al.*, 2017]. For a fair comparison, we use the same metric, i.e. binary classification accuracy on the rating to evaluate the performance of **MTRLCNN** and our proposed models.

The accuracy on the major 29 image semantic tags is illustrated in Fig. 7, which involves three compared methods: 1) **CFN$_{pure}$-2_task**: pure CF based network without using any image or text review as input. Two tasks, i.e. image assessment prediction and image tag classification are jointly

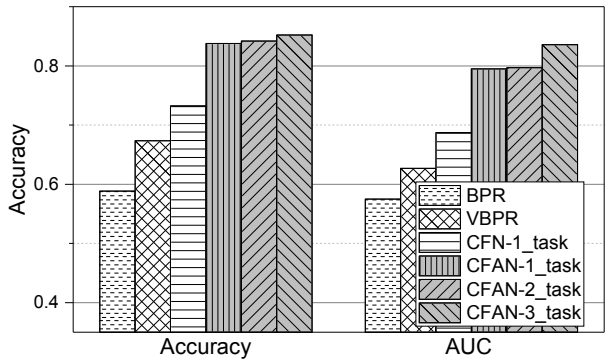

Figure 8: Evaluation of methods: **BPR**, **VBPR**, **CFN-1_task**, **CFAN-1_task**, **CFAN-2_task** and **CFAN-3_task**.

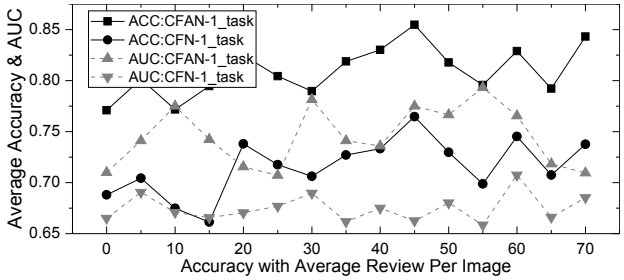

Figure 9: Classification accuracy and AUC score as the average review number per image changes.

learned; 2) **MTRLCNN**: the state-of-the-art user-agnostic assessment approach which is a multi-task learning based network using only visual image as input; 3) our approach **CFAN-3_task** as depicted in Fig. 3 which uses CF and attention mechanism to fuse both text and image data. In general, our method outperforms in most tags, which verifies the efficacy of our model to traditional user-agnostic aesthetic assessment task. Note that our model meanwhile learns the user-specific assessment task.

Personalized Aesthetic Assessment

For comparison of our approach **CFAN-3_task** for personalized prediction, AUC score for ROC [Bradley, 1997] is compared. Since there are few tailored personalized models for image aesthetic assessment, we compare two recently proposed general personalized prediction.

BPR (**B**ayesian **P**ersonalized **R**anking) [Rendle *et al.*, 2012]: is a well-known personalized recommendation algorithm to model user implicit feedback (e.g. clicks, purchases) without visual features. In the experiment, for each positive feedback, a negative sample is randomly generated.

VBPR (**V**isual **B**ayesian **P**ersonalized **R**anking) [He and Mcauley, 2016]: is an Bayesian Personalized Ranking method which models raw visual features for item recommendation. In the test, visual features are also extracted by the same deep features [Simonyan and Zisserman, 2014].

We also test degraded versions of our model as follows: 1) **CFN-1_task**: the input are user factors, image factors and image raw data. Only collaborative filter layers are included; 2) **CFAN-1_task**: the review is added as another modal of input and a RNN model is used to extract attention map; 3) **CFAN-2_task**: the network which jointly performs personalized assessment task and semantic tag classification task. For **CFN-1_task** and **CFAN-1_task**, no multi-task is involved and the loss focuses on the user-specific prediction.

Fig. 8 shows the averaged performance among various tags of images. Note our final model as depicted in Fig. 3 involves three loss, so it is termed by **CFAN-3_task** in the plot. We make the following observations and analyses:

- The basic method **BPR** only considers user and image factors. It performs worst because of the limited input information. **VBPR** and **CFN-1_task** have the same input but the latter performs better in binary classification problem. **CFAN-1_task** improves the performance by nearly 10% with the help of additional text/image information. Our main network **CFAN-3_task** is superior against other methods and the result shows that different tasks can synergetically contribute to each other.
- No surprisingly, the network with attentive mechanism, i.e. **CFAN-1_task** outperforms the **CFN-1_task** by a notable margin. It can be explained that attention map extracted from user review can reflect users' focuses. Features in less relevant regions would be filtered out. Fig. 4 shows the feature maps extracted by different users. In **CFN-1_task** network, the connection between user and image is based on the combination of the factors. In **CFAN-1_task** network, that is further based on the user focused image visual features.

As one of our main contributions is modeling the user reviews, we also test the performance of our methods changing reviews in each image. Fig. 9 shows the performance for **CFN-1_task** and **CFAN-1_task** learning user personalized assessment. The accuracy and AUC have minor fluctuations within 10%. Though reviews are not modeled in **CFN-1_task**, the result can confirm that the number of reviews makes no significant influence on the results of aesthetic rating. A special case is that when we only use reviews without image nor rating relation input, the personalized task accuracy is 0.18. In fact, the modeling of personalized information involves user ID, image ID, review, and image. The reviews contribute to but not dominate the results.

5 Conclusion

In this work, to the best of our knowledge, we are the first to predict personalized user specific rating of image aesthetics by training a neural network from images and texts. Our framework models visual features, semantic tags, and user reviews, leading to increasing accuracy of personalized image aesthetic rating. In the framework, the collaborative filtering and RNN network are combined to generate user specified visual features with attention maps. The aesthetic assessment tasks (i.e. user-specified and user agnostic) and semantic classification task are performed simultaneously. Empirical study shows state-of-the-art performance of our approach. In the future, we aim at using this research as a first step to perform explainable image aesthetic assessment.

References

[Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[Bhattacharya *et al.*, 2010] Subhabrata Bhattacharya, Rahul Sukthankar, and Mubarak Shah. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *Multimedia*, pages 271–280. ACM, 2010.

[Bradley, 1997] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[Caruana, 1997] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[Datta *et al.*, 2006] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, pages 288–301. Springer, 2006.

[Datta *et al.*, 2007] Ritendra Datta, Jia Li, and James Z Wang. Learning the consensus on visual quality for next-generation image management. In *Multimedia*, pages 533–536. ACM, 2007.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li Jia Li, Kai Li, and Fei Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[Dhar *et al.*, 2011] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. High level describable attributes for predicting aesthetics and interestingness. In *CVPR*, pages 1657–1664. IEEE, 2011.

[Dong *et al.*, 2015] Zhe Dong, Xu Shen, Houqiang Li, and Xinmei Tian. Photo quality assessment with dcnn that understands image well. In *International Conference on Multimedia Modeling*, pages 524–535. Springer, 2015.

[Goldberg, 1990] Lewis R Goldberg. An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229, 1990.

[He and Mcauley, 2016] Ruining He and Julian Mcauley. Vbpr: visual bayesian personalized ranking from implicit feedback. In *AAAI*, pages 144–150, 2016.

[Kang *et al.*, 2014] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *CVPR*, pages 1733–1740, 2014.

[Kao *et al.*, 2017] Yueying Kao, Ran He, and Kaiqi Huang. Deep aesthetic quality assessment with semantic information. *IEEE Transactions on Image Processing*, 26(3):1482–1495, 2017.

[Ke *et al.*, 2006] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. In *CVPR*, volume 1, pages 419–426. IEEE, 2006.

[Kong *et al.*, 2016] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, 2016.

[Lu *et al.*, 2015a] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang. Rating image aesthetics using deep learning. *IEEE Transactions on Multimedia*, 17(11):2021–2034, 2015.

[Lu *et al.*, 2015b] Xin Lu, Zhe Lin, Xiaohui Shen, Radomir Mech, and James Z Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *ICCV*, pages 990–998, 2015.

[Manning *et al.*, 2014] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *ACL*, pages 55–60, 2014.

[Marchesotti *et al.*, 2015] Luca Marchesotti, Naila Murray, and Florent Perronnin. Discovering beautiful attributes for aesthetic image analysis. *IJCV*, 113(3):246–266, 2015.

[Murray *et al.*, 2012] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*, pages 2408–2415. IEEE, 2012.

[Nishiyama *et al.*, 2011] Masashi Nishiyama, Takahiro Okabe, Imari Sato, and Yoichi Sato. Aesthetic quality classification of photographs based on color harmony. In *CVPR*, pages 33–40. IEEE, 2011.

[Rendle *et al.*, 2012] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. pages 452–461, 2012.

[Segalin *et al.*, 2016] Cristina Segalin, Alessandro Perina, Marco Cristani, and Alessandro Vinciarelli. The pictures we like are our image: Continuous mapping of favorite pictures into self-assessed and attributed personality traits. *IEEE Transactions on Affective Computing*, pages 1–14, 2016.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014.

[Tian *et al.*, 2015] Xinmei Tian, Zhe Dong, Kuiyuan Yang, and Tao Mei. Query-dependent aesthetic model with deep learning for photo quality assessment. *IEEE Transactions on Multimedia*, 17(11):2035–2048, 2015.

[Tong *et al.*, 2004] Hanghang Tong, Mingjing Li, HongJiang Zhang, Jingrui He, and Changshui Zhang. Classification of digital photos taken by photographers or home users. In *Pacific-Rim Conference on Multimedia*, pages 198–205. Springer, 2004.

[Zhang *et al.*, 2013] Fang-Lue Zhang, Miao Wang, and Shi-Min Hu. Aesthetic image enhancement by dependence-aware object recomposition. *IEEE Transactions on Multimedia*, 15(7):1480–1490, 2013.

[Zhang, 2016] Luming Zhang. Describing human aesthetic perception by deeply-learned attributes from flickr. *arXiv preprint arXiv:1605.07699*, 2016.