Aesthetic Critiques Generation for Photos

Kuang-Yu Chang*, Kung-Hung Lu*, and Chu-Song Chen
Institute of Information Science, Academia Sinica, Taipei, Taiwan
{kuangyu, henrylu, song} @iis.sinica.edu.tw

Abstract

It is said that a picture is worth a thousand words. Thus, there are various ways to describe an image, especially in aesthetic quality analysis. Although aesthetic quality assessment has generated a great deal of interest in the last decade, most studies focus on providing a quality rating of good or bad for an image. In this work, we extend the task to produce captions related to photo aesthetics and/or photography skills. To the best of our knowledge, this is the first study that deals with aesthetics captioning instead of AQ scoring. In contrast to common image captioning tasks that depict the objects or their relations in a picture, our approach can select a particular aesthetics aspect and generate captions with respect to the aspect chosen. Meanwhile, the proposed aspect-fusion method further uses an attention mechanism to generate more abundant aesthetics captions. We also introduce a new dataset for aesthetics captioning called the Photo Critique Captioning Dataset (PCCD), which contains pair-wise image-comment data from professional photographers. The results of experiments on PCCD demonstrate that our approaches outperform existing methods for generating aesthetic-oriented captions for images.

1. Introduction

Aesthetic computing has long been an important topic in the field of computer vision. In this paper, we consider the problem of image captioning from the aesthetic viewpoint. There are many studies on caption generation [11, 9, 12, 22, 2, 21, 8, 32, 9, 20]; however, most of them focus on producing a single caption that depicts the objects or the relative positions of the objects in a picture.

In this paper, we study a new problem, namely, **aesthetic analysis of photos**. Aesthetic quality (AQ) assessment has generated a great deal of interest in the last decade. Many studies tackled this problem with various feature representations and/or learning architectures [19, 14, 17, 16, 24].

*indicates equal contribution.

Photo Critique Captioning: *racing makes for interesting pictures because of the speed the movement the bright colors*

Image Captioning: *a man riding a motorcycle down a street*

Figure 1: Photo critique captioning versus image captioning

However, the purpose of AQ assessment is to provide a binary decision, which yields a quality rating of good or bad for a specific photo. In this paper, we address a more general problem, namely, captioning of photo aesthetics and/or photography skills. To the best of our knowledge, this is the first study that considers the problem, which covers a broader range of applications than AQ assessment only. Besides AQ, our system analyzes the reasons why photos are (or are not) appealing in some respect, so that a meaningful caption can be generated for a photo from an aesthetic perspective. Figure 1 shows the difference between photo aesthetics captioning and common image captioning.

Figure 2 provides more examples of captions produced by our system. In Figure 2(a), the space reserved on the right-hand side of the photo needs to be refined; in Figure 2(b), the vanishing point and lines are good and admired; and in Figure 2(c), the subjects gaze creates further space that enhances the AQ. With the captions provided, the topic addressed in this paper suggests a better applicability. Besides the simple assessment of AQ, it is possible to provide in-depth descriptions and comments, which are informative and can improve the photographic skills of users.

Our learning model assumes there is an input dataset of images and their sentence descriptions. Every sentence relates to a specific aspect. To evaluate our work thoroughly, we compiled a dataset for photo aesthetics captioning called PCCD. The dataset contains pairwise data of images and sentences, where an image could have multiple sentences related to different aspects of the aesthetics. To the best of our knowledge, this is the first publicly available dataset for photo aesthetics captioning.

We propose two approaches to solve the aesthetic critique problem. The first is our baseline approach called the aspect oriented (AO) approach; and the second is an improvement of AO called the aspect fusion (AF) approach. In the AO approach, the training data are divided into disjoint subsets based on the aspects of sentences, and we apply a CNN-LSTM model to create a photo captioning system for each aspect. The CNN model is also trained for regression and then used to select the most interesting aspect of the input image. Instead of enforcing a single aspect, our AF approach fuses the captions learned from the individual aspects to create a new caption. We propose a soft attention mechanism in the AF approach to produce a caption from the established LSTMs. In our evaluation, the AF approach performs better than learning a CNN-LSTM model directly from the training data of all aspects.

The contributions of this paper are as follows.

From judgement to critiques: As well as AQ assessment, our approach provides a caption for the aesthetic value of the input image.

Photo aesthetics captioning dataset: We compiled a dataset for the performance evaluation. It is the first publicly available dataset in this new area.

Multi-aspect aesthetics captioning: We propose a new captioning approach to generate aesthetic critiques for images and the generated sentences are aspect-oriented which are more diverse and favorable for human.

2. Related Work

Image Captioning: Recently, many approaches [11, 9, 12, 22, 2, 21, 8, 32, 9, 20] have achieved promising results by describing objects in images and videos with natural language. Most of them [12, 11] apply a CNN-RNN framework comprised of high-level features extracted from a CNN model trained on object recognition and the Recurrent Neural Networks (RNN) language model. Johnson et al. [11] consider the dense captioning task and use the CNN-RNN framework to generalize object detection and generate dense annotations of images. Mao et al. [22, 21, 20] propose a multimodal Recurrent Neural Networks model that embeds the recurrent language features and image features in a multimodel space. [2, 31] leverage external data so that the CNN-LSTM captioning model does not require paired image-sentence data for training.

Figure 2: Examples of captions generated by our system for photo aesthetic analysis.

Some works focus on captioning with visual attentions. You et al. [36] extract visual attributes from the image and combine them as semantic attention to guide image captioning. Xu et al. [33] enable the model to focus on a local patch of the image when generating a sequence of words. In contrast to the CNN-RNN models, some approaches exploit retrieval techniques. For example, Devlin et al. [6] proposes a nearest neighbor method to retrieve captions and outputs the top ranked one. Fang et al. [9] extract visual concepts by training visual detectors for words and use a maximum entropy model conditioned on the detected words to generate captions. Most recent works [15, 27] extend the attention mechanism with the ability to interact with language model to choose the attention areas dynamically.

Video captioning [34, 37, 25] has also generated a great deal of interest. Tapaswi et al. [29] align a movie scene with a suitable book chapter by using dialogs and character identities as cues; while Zhu et al. [38] match books and movies on the sentence/paragraph level. These two works try to provide rich descriptive explanations of visual content, which are far beyond existing captioning works in terms of semantic meaning.

AQ Assessment: AQ assessment of photos has been investigated for a long time [5, 23, 24, 7]. The first challenge is how to represent the aesthetics of an image. Traditional low level features, such as color histograms, hue and saturation, are utilized in AQ assessment. Moreover, some studies [24, 7, 35] focus on designing semantic feature representations. The inspiration might come from the photography or image processing, e.g., the rule of thirds, sky illumination and simplicity [35].

With the success of deep learning techniques, some approaches [19, 14, 17, 16] exploit deep CNN to learn powerful representations from the data in an end-to-end manner

Figure 3: Examples of the captions generated by our approach for different aspects (composition and color) of the same picture.

for AQ assessment. Kong et al. [14] propose a CNN-based method that combines different loss functions and attributes in their aesthetic dataset. Lu et al. [16] introduced a double-column CNN architecture that uses holistic images and image patches as global and local features respectively. The style attributes are aggregated to leverage the performance. Subsequently, Lu et al. [17] proposed a multiple instance learning CNN model that generates multiple patches from a single image. More precisely, the statistics aggregation layer aggregates the multiple patches and achieves a better performance than comparable approaches.

In this paper, we introduce a two-stage LSTM model that can integrate the LSTM features of different aspects for photo captioning from the aesthetic viewpoint.

3. Our Framework

Multi-aspect captioning: A worth noting issue of the photo aesthetics captioning is its *multi-aspect* nature; that is, more than one aspect of an image can be commented on. For example, a photo could be characterized in terms of the *composition*, *color-arrangement* and *subject-contrast* aspects, which relate to the aesthetics or photographic skills. Different aspects would have different captions to be synthesized, as shown by the example in Figure 3. The caption produced by our system is "i like the way you have used the rule of thirds $\cdots$" for the composition aspect. However, for the color-arrangement aspect, "the color is great and i love the bright greens" would be a more suitable alternative produced by the system.

3.1. Aspect-oriented (AO) approach – baseline

In our baseline approach, AO, the training data are separated into different aspects. Assume a dataset containing N triplets, $\mathbf{D} = (\Phi_i, C_i, a_i)$, $i = 1 \cdots N$, is available to train our photo aesthetics captioning system, where Φ_i is the i-th image and C_i is its caption. The images can be repeated (i.e., $\Phi_i = \Phi_j$ for some i and j), but the captions C_i vary with i; a_i in $\{1 \cdots L\}$ is the aspect of the caption, where L is the number of aspects. Besides the images and captions, a likelihood (namely, $p_{i,l} \in [0, 1]$) is also available as the degree of aesthetic appeal of the image Φ_i on the aspect l.

In the AO approach, the training data asociated with the triplets whose captions are for a single aspect, namely, $(\Phi_i, C_i, l), i = 1 \cdots N_l$, are used, where N_l is the amount of training data in the aspect l. We employ a CNN-LSTM architecture to train the captioning model for each single aspect. To proceed, we give a brief review of the CNN-LSTM as follows. Given a training caption (desired output) C_i comprised of the words $\{w_1, w_2, \cdots, w_T\}$, a total of $T + 2$ feature vectors $\{x_{-1}, x_0, x_1, \cdots, x_T\}$ are fed into the LSTM model, where x_{-1} is the feature vector extracted from the CNN for the input image Φ_i, x_0 is a special START token, and x_t are the feature vectors converted from w_t in the feature-embedded layer for $t = 1 \cdots T$. The LSTM model computes a sequence of hidden states h_t and outputs the word probability prediction y_t by the recurrence formula for $t = 1 \cdots T$,

$$\{h_t, y_t\} = f(h_{t-1}, x_t). \tag{1}$$

Thus, given an input image Φ_i, we can get L aspect-specific captioning models. We denote the hidden states (a.k.a. hidden annotations) of the l-th aspect LSTM model to be $\mathbf{h}_l = \{h_{l,t} | t = 1 \cdots T\}$, for $l = 1 \cdots L$.

Without loss of generality, the neuraltalk2 model [12] is adopted in our approach, despite our framework can use other models as well. The CNN-LSTM models have some variations in recent studies [33, 36]. Because this paper deals with a new problem, evaluation (or comparison) of more updated CNN-LSTM models that are favorable for conventional captioning tasks is not our main focus. On the contrary, we focus more on handling the aesthetic critiques of different aspects to develop a better photo aesthetics captioning system (Section 3.2).

The learned caption generator is then used to produce a caption associated with the photo aesthetics of the aspect focused on. For example, if the model is trained on the aspect of composition, the captions generated will target the compositional analysis of the image. In the AO approach, a single aspect l^* is selected from the L aspects, and the caption generated by the CNN-LSTM model for the l^*-th aspect serves as the output. To choose the aspect of appealing, we use the CNN model to train L predictors based on the pairs $\{(\Phi_i; p_{i,l})\}$ $(l = 1 \cdots L)$ in our dataset. The output of the CNN model has L nodes, each of which has a regression output in the range [0,1]. Then, given an input image, we select the aspect with the highest prediction value in the AO approach as l^*. Figure 4 shows the flowchart of the AO approach, which combines the aspect predictor and the

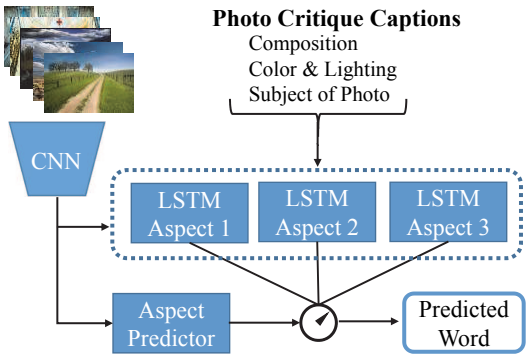

Figure 4: Overview of the aspect-oriented (AO) approach, where the number of aspects $L = 3$.

individual aspect-oriented captioning systems.

3.2. Aspect-fusion (AF) approach

However, irrespective of the type of evaluation used (automatic or human evaluation), we found that the performance of the AO approach is limited because it is trained on a restricted set of data for the chosen aspect. As the training data in a single aspect is less than those in the whole dataset, AO cannot exploit the interrelated sentences between different aspects to produce a more diverse caption. Hence, the caption generated by the AO approach tends to be monotonous.

A possible remedy to this lack-of-diversity problem is to apply the CNN-LSTM to the whole dataset of image-caption pairs, $\{(\Phi_i, C_i)|i = 1 \cdots N\}$, which contains the training captions from all aspects; we refer this approach to as the *CNN-LSTM on the whole dataset* (CNN-LSTM-WD). However, we found that this approach still suffers from the same problem on either automatic or human evaluation, possibly because of the inter-aspect difference of the words and sentence forms of the captions.

To address this issue, we develop the AF approach that also uses the entire dataset to learn a CNN-LSTM model. Unlike CNN-LSTM-WD whose inputs are the images $\{\Phi_i \in \mathbf{D}\}$ in the learning process, we propose leveraging the L aspect-specific models already trained. In the AF approach, the hidden annotations $\mathbf{h}_l$ $(l = 1 \cdots L)$ generated by the L aspect-specific captioning models are further used for learning the CNN-LSTM model; hence, both the images and hidden annotations, namely, $\{(\Phi_i, \mathbf{h}_{i;l})|i = 1 \cdots N\}$, are used as the inputs in the learning process, and the captions $\{C_i|i = 1 \cdots N\}$ remain as the desired outputs. As the output words in the LSTM models are directly dependent on the hidden states $h_{l,t}$, the hidden-layer outputs can serve as the feature representations extracted by using the CNN-LSTM models. Thus, the hidden annotations are deep features extracted from the models already trained and established for every aspect, which are better sources for training and make the AF approach potentially more effective in learning a new caption model. The recurrence formula in the new LSTM of the AF approach is established as

$$(g_\tau, y_\tau) = F(g_{\tau-1}, x_\tau, s_\tau). \tag{2}$$

In contrast to Eq. (1), we denote 'τ' as the time index and 'F' as the recursive function in Eq. (2) to avoid the confusion with the symbols 't' and 'f'. In Eq. (2), the output y_τ is conditioned on the input words x_τ and the previous hidden state $g_{\tau-1}$, as well as on s_τ that is a context vector relying on the aspect-specific hidden annotations, $\mathbf{h}_l$ $(l = 1 \cdots L)$. The formulation of the context vector s_τ will be detailed in the following.

To fuse the hidden annotations $\mathbf{h}_l = \{h_{l,t}|t = 1 \cdots T\}$ from different aspects l, a worth-of-noting issue is that the time indices t in different aspects are not aligned inherently, and thus they should not be combined directly to generate the output at time $\tau = t$. In our AF model, there are L sources, and we introduce a soft-attention layer to predict the aspect-fusion coefficients from the context information. The context vector s_τ is determined by combining the aspect-specific hidden annotations as follows:

$$s_\tau = \sum_{l=1}^{L} \sum_{t=1}^{T} \alpha_{lt}^{\tau}(\mathbf{h}, g_{\tau-1}) h_{lt}. \tag{3}$$

In Eq. (3), the fusion coefficients $\alpha_{lj}^{\tau}(\mathbf{h}, g_{\tau-1})$ of time position τ is dependent on the aspect-specific hidden annotations, $\mathbf{h}$, and hidden state of the previous time postion, $g_{\tau-1}$. The coefficients provide soft attention on the entire period $(t = [1 : T])$ of the aspect-specific hidden annotations. Note that for different aspects $l = 1 \cdots L$, different coefficients $\alpha_{l,[1:T]}$ are used, and thus an asynchronized attention mechanism is enforced. To provide the capability of non-uniform alignment of reference, recent advances of sequence to sequence models [3][18] also embed soft attention in their formulations. However, unlike their approaches where only a single source sequence is used, there are L sources in the AF model and an updated soft-attention mechanism is proposed in our study. To generate the fusion coefficients α_{lt}^{τ} in the soft-attention layer of AF, we first produce an intermediate representation e_{lt}^{τ} that is adaptive to the previous state $g_{\tau-1}$ and aspect-specific hidden annotations h_{lt} by using

$$e_{lt}^{\tau} = A(g_{\tau-1}, h_{lt}), \tag{4}$$

with $A(\cdot, \cdot)$ a feed-forward network established as

$$A(g_{\tau-1}, h_{lt}) = \mathbf{W}\gamma(\mathbf{U}g_{\tau-1} + \mathbf{V}h_{lt}), \tag{5}$$

where γ is the ReLU activation function, $\mathbf{W} \in R^{n \times n}$, $\mathbf{U} \in R^{n \times n}$ and $\mathbf{V} \in R^{n \times n}$ are learnable weighted matrices, and n is the dimension of the hidden state vector (in

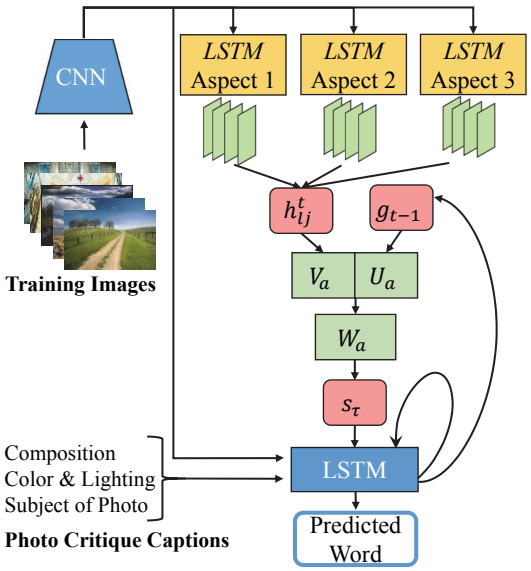

Figure 5: Overview of the aspect-fusion (AF) approach in the case where $L = 3$.

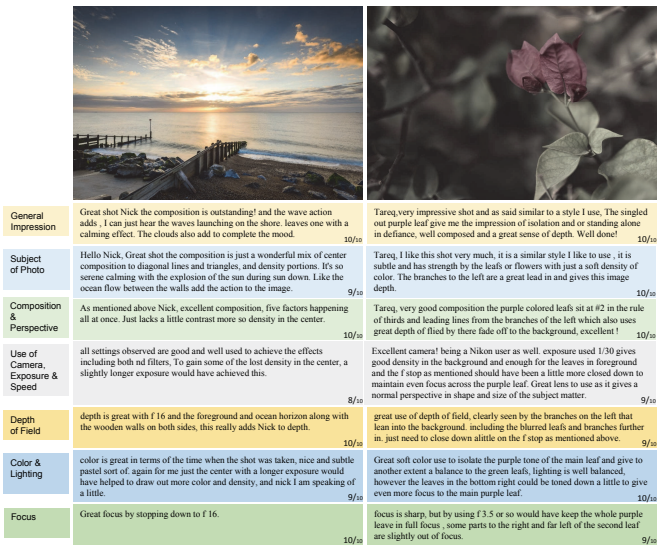

Figure 6: Samples in the Photo Critique Captioning Dataset (PCCD).

our implementation, $n = 768$). Then, α_{lt}^τ is obtained by normalizing e_{lt}^τ,

$$\alpha_{lt}^\tau = \frac{exp(e_{lt}^\tau)}{\sum_{p=1}^{L}\sum_{q=1}^{T} exp(e_{pq}^\tau))}. \quad (6)$$

The proposed AF approach is illustrated in Figure 5. By considering the context vector, the AF approach can leverage the hidden annotations of different aspects and choose the proper combination dynamically over time for caption generation. Because the aspect-oriented hidden annotations, together with the image CNN features and word sequences of captions, are fed into the LSTM model (blue part) to generate the output caption of our AF approach, the sentences in different aspects are likely to be softly merged in the learned model to enhance the formation of captions.

4. Dataset and Evaluation Criteria

In this section, we present the dataset for photo critiques learning, and the criteria for the performance evaluation.

4.1. Dataset

To validate the proposed method on the aesthetic-related photo caption generation problem, we compiled a dataset called the Photo Critique Captioning Dataset (PCCD), which is available to the public and can be used for future studies in this area. The dataset is based on a professional photo critique website[1] that provides experienced photographers reviews of photos. On the website, photos are presented and several professional comments are given about the following seven aspects: general impression, composition and perspective, color and lighting, subject of photo, depth of field, focus, and use of camera, exposure and speed. For those aspects that have comments for a given photo, a paragraph containing one or more sentences are presented. Figure 6 shows some sample examples. The photos together with their sentences in the respective aspects are used to establish the triplets $\mathbf{D} = (\Phi_i, C_i, a_i)$. Table 1 shows the statistics of PCCD. It contains 4235 images and more than sixty thousands captions. The source data used to compile our dataset also contains a rating (from 1 to 10) per aspect for a photo; the higher the rating for the aspect, the better will be the quality of the input image for the aspect. The ratings are normalized to [0,1] and serve as the likelihood of aesthetic appeal of the aspect $\{p_{i,l}\}$ described in Section 3.1.

In our experiments, because not all aspects are commented for a photo, we select $L = 3$ most frequent aspects, namely, composition and perspective, color and lighting and subject of photo, which contain 3840 images and 30254 sentences for training and 300 images for testing. We use a threshold to control the vocabulary of words. Words that appear fewer times than the threshold are collapsed into the <UNK> category. A higher threshold yields a smaller vocabulary because less frequent words are grouped. In our implementation, the threshold is 5, which provides a vocabulary of 10390 words. Though not large, PCCD would be a good start to the new step on aesthetics-critique captioning.

The dataset described above contains pairwised information of images and aesthetic captions. We use another pairwised data, MSCOCO image captioning dataset, as the outside data to enhance the performance of photo aesthetic

[1] https://gurushots.com/

Table 1: Statistics of our photo critique captioning dataset.

Aspect	# photos	# sentences	# words
General Impression	4123	12908	237337
Composition & perspective	4000	12848	262194
Color & Lighting	3769	9384	168028
Subject of photo	3812	8022	129179
Depth of field	3017	4864	86391
Focus	2994	5421	89626
Use of camera, exposure	3396	8255	156721
Total	4235(union)	61702	1129476

critiques. MSCOCO caption dataset contains over 160K images and 1 million captions about objects. In our implementation, the CNN-LSTM model is pre-trained on the MSCOCO image captioning dataset as an initial model, and then fine-tuned on PCCD. Compared to training with PCCD directly, we found that the pre-training is useful to enrich the object-recognition and sentence-formation capabilities of our photo aesthetics captioning system, and results in a better captioning performance.

As MSCOCO is large about object descriptions and PCCD is relatively small on the aesthetic critiques, when training either the AO or the CNN-LSTM-WD approach, the CNN pre-trained on MSCOCO are fixed to keep its object-description capability, and only the LSTMs are fine-tuned on PCCD. This strategy is helpful to avoiding over-fitting and provides a better subject-identification capability. Similarly, when training the AF approach, the aspect-specific LSTMs are also fixed to avoid over-fitting. The strategy also benefits the efficiency of training. A computer mounted with a single Titan X GPU is used in our implementation. It takes about one day to train a single CNN-LSTM model with the AO approach, and two days with the AF approach, respectively.

4.2. Evaluation Criteria

As we are handling a new topic, no existing studies are available for comparisons. The evaluation criteria suitable to this new topic become also an issue. Traditional criteria such as BLEU [26] and METEOR [4] use simple n-gram overlaps for evaluation, which produce inaccurate results because two sentences may share similar meanings without a high n-gram overlap. Note that there is more than one reference caption that corresponds to a single image in PCCD. As our dataset (PCCD) is not designed for object recognition, unlike common image captioning datasets, these reference captions are often not synonymous sentences. This characteristic makes the criterion computing the occurrence frequency of n-grams in the reference captions (such as CIDEr [30]) inconsequential for the evaluation either.

SPICE: A recent advance in automatic evaluation metrics [1] captures more semantics in a photo for the comparison. Though imperfect either, we suggest that the SPICE criterion presented in [1] is more suitable for the performance evaluation of photo aesthetics critiques. The SPICE method parses a sentence into a graph, and evaluates the similarity based on the parsed results between the generated and reference sentences and then reports the F-score. The criterion in [1] adopts a variant of the rule-based version of the Stanford Scene Graph Parser [28]. A Probabilistic Context-Free Grammar (PCFG) dependency parser [13] is followed by simplifying quantificational modifiers, resolving pronouns and handling plural nouns. It has been shown that SPICE performs better than traditional metrices such as BLEU, METEOR, and CIDEr in capturing human judgment over the generated captions. Hence, for an image, we compute the SPICE (F-score) between the generated caption and all of its reference captions, and then use the highest one as the evaluation score for the image of interest.

Diversity: In contrast to the other captioning problems, repetition of the captions generated is an issue for photo aesthetic critiques. For example, if the same sentence "I like the composition and perspective of this image" is repeated for different photos, people will feel tedious because the critiques generated for the test photos are not plentiful enough. However, the problem caused by the repeated or monotonous captions cannot be reflected by the traditional captioning evaluation criteria mentioned above. To address this issue, we propose a measure called *diversity*, which takes the near-duplication sentences into consideration to establish an evaluation measure. We treat two sentences duplicate if the ratio of common words between them is larger than a threshold (in our implementation 70% is used), and then call the non-duplication rate (one minus the duplication rate) of the captions generated for the test photos as *diversity*. This criterion is used to evaluate our photo critiques problem as well.

5. Experiment Results

In the experiments, we compare the AF approach with two baseline approaches, AO and CNN-LSTM-WD. First, we show the results on PCCD based on the automatic evaluation criteria in Section 5.1. Then, we compare the performance based on human evaluations and present the results in Section 5.2. Finally, we show the cross-dataset results on the AVA dataset [24] in Section 5.3.

5.1. Automatic Evaluation

As mentioned above, we use both the SPICE (F-score) and diverity for automatic evaluation of the test dataset.

Table 2 shows the SPICE evaluations of compared approaches at generated critiques, and we also report the precision and recall scores for reference. In terms of the SPICE criterion, CNN-LSTM-WD yields better performance than AO, and AF performs better than both AO and CNN-LSTM-WD. We attribute these results to that the AO ap-

Table 2: Evaluation of the proposed approaches via the SPICE criterion.

Method	SPICE	Precision	Recall
CNN-LSTM-WD	0.136	0.181	0.156
AO Approach	0.127	0.201	0.121
AF Approach	0.150	0.212	0.157

proach trains the models by using only the aspect-specific captions, which are limited and thus performs worse than the CNN-LSTM-WD approach that uses the whole dataset for training. In contrast, the AF approach further employs the hidden annotations as intermediate representation for training the captioning model, which can yield the best performance.

Then, we use the diversity criterion to evaluate these methods. The results are shown with the x-axis of Figure 7. In contrast to the SPICE criterion, the CNN-LSTM-WD approach performs worse than the AO approach on the diversity criterion. It appears that applying CNN-LSTM-WD to the whole dataset that contains captions of different aspects tends to yield more monotonous sentences. The AF approach that leverages and fuses the hidden annotations of different aspects still performs more favorable than the other approaches. Figure 7 combines both the diversity and SPICE measures in a diagram. As can be seen, the AF model that integrates the learned sentence representations can produce more diverse sentences (in terms of diversity) with higher semantic similarity (in terms of SPICE).

5.2. Human Evaluation

In human evaluation, we ask the subjects to rate the generated captions on a 3-point scale: Good, Common and Bad. We define the judge that Good means that the caption contains details presented in the picture and its suggestions are helpful; Common means that the caption is safe but not impressive. As photo critiques are subjective, there are many comments like "I like your composition" or "Nice photo, I think your photo is good" which does not describe the detail of image but expresses critics's preference. They might not be thought as wrong captions but to be honest, they are not useful advice for photographers so we classify these kind of captions into Common. Bad means the caption contains obvious error. This setting is similar to the design in [10].

Main Results: It is natural that our generated critiques should be judged by professional photographers. However, we also care about the comments from common users as the eventual goal of this task is to help people take satisfied pictures. We find three experts with more than five-year experience in photography for expert evaluation, and also establish an experiment involving five people through Amazon Mechanical Turk (AMT).

As shown in Table 3, we can find that the AF approach

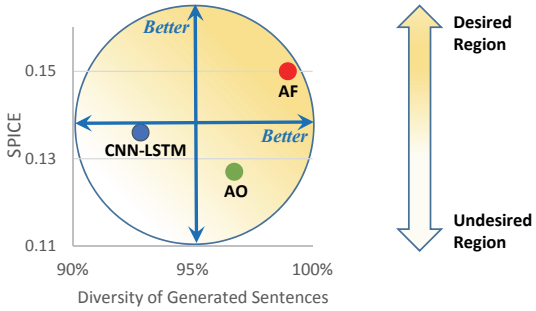

Figure 7: The automatic evaluation results of the three approaches compared, where the X and Y axes represent the diversity and SPICE, respectively.

achieves the best performance in terms of average score among the three methods. AF also generates much more "Good" score judged by both experts and common users. This demonstrates that our AF approach has more favorable user experience. We also note that the AO approach has the largest number of "Common" score as well as the least "Bad" score. However, as mentioned before, "Common" critique cannot provide useful advices to user. Hence, the AO approach is more like a "safe" method but not an ideal solution for this task in terms of human evaluation. One possible explanation is that in the AO approach, only the captions of a single aspect are used without sharing information with the other aspects.

Another noteworthy outcome is the consistency of the judgements from both experts and common users. This matches our assumption that experts have stricter standard to photos critique and thus the average scores judged by experts are lower than those by common users. However, the AF model still outperforms the baselines in both testing groups. Figure 8 shows some examples of captions generated by the three approaches.

Comparison with Groundtruth: In particular, we ask some experts to compare a computer generated critique to the ground-truth captions when presented with an image. For each image, we calculate the ratio that the generated critiques are no worse than the ground-truth captions, and the average result is shown in Table 4. We can find that the AF approach reaches the highest score, which proves that our method could generate better photo aesthetic critiques.

Novel Sentences Generated: The novel captions (not present in the training data) generated by AF, AO and CNN-LSTM are 66%, 37% and 48% respectively, revealing that AF inclines to generate diverse and integrated sentences.

Failed Cases: Figure 9 shows some failed cases generated by our approach. Although the AF approach performs better than the other approaches on photo aesthetics captioning, there is still room for improvement. Our approach can serve as a baseline for future studies in this area.

Table 3: Human evaluation by workers on Amazon Mechanical Turk and experts.

	Evaluation by users on AMT				Evaluation by experts			
Method	Good 3	Common 2	Bad 1	Average Score	Good 3	Common 2	Bad 1	Average Score
CNN-LSTM-WD	19.0%	65.7%	15.3%	2.04	7.1%	75.0%	17.9%	1.89
AO Approach	19.7%	72.1%	8.2%	2.11	10.8%	78.0%	11.2%	1.99
AF Approach	28.8%	57.2%	14.0%	2.15	16.8%	69.2%	14.0%	2.03

Figure 8: Examples of the critiques generated by the three models, AF, AO, and CNN-LSTM-WD.

Table 4: Comparison of the generated captions with the ground truths by human.

Method	Better	Worse	Total
CNN-LSTM-WD	47.4%	52.6%	100%
AO Approach	51.4%	48.6%	100%
AF Approach	58.4%	41.6%	100%

5.3. Cross-dataset Results

We apply the AF models to a large-scale aesthetic-quality-assessment dataset, Aesthetic Visual Analysis (AVA) [24]. Unlike PCCD, the AVA dataset has no ground-truth captions, and thus we cannot fine-tune the models on this dataset. Therefore, the AF model trained by using PCCD is directly applied to the AVA dataset. Some results on this cross-dataset testing are shown in the respective figure on the supplementary material. From the results,

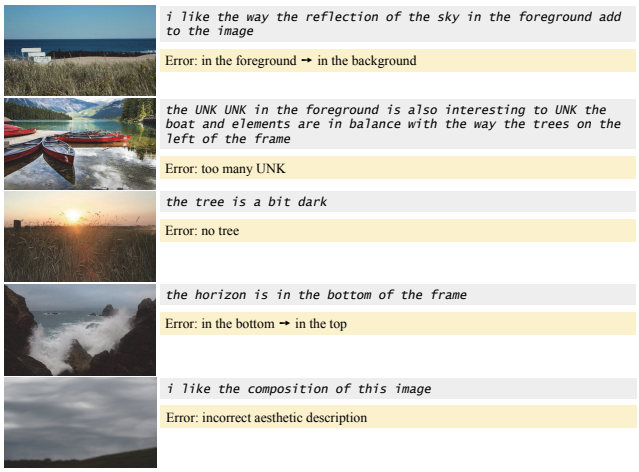

Figure 9: Some failed-case captions generated by our approach.

it can be seen that the learned model can be used for generating photo aesthetic critiques for other image datasets as well, which demonstrates the generalization ability of the proposed approach.

6. Concluding Remarks

In this paper, we study a new problem, namely, captioning different aesthetic aspects of an image. To resolve the prolem, we introduce the baseline approach, AO, which can produce aspect-specific captions by using existing CNN-LSTM methods. We then extend the AO approach to the AF approach, which can exploit the hidden annotations learned from different aspects to generate captions that are more semantically meaningful and diverse. In addition, we show the outcomes on the proposed PCCD dataset as well as the quantitative results judged by both automatic criteria and human evaluation. The results demonstrate the effectiveness and application potential of our approach.

Acknowledgment

This work was supported in part by Ministry of Science and Technology under the grants MOST 104-2221-E-001-023-MY2, MOST 105-2218-E-001-006 and MOST 106-2221-E-001-016.

References

[1] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016. 6

[2] L. Anne Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, 2016. 1, 2

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015. 4

[4] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005. 6

[5] R. Datta, D. Joshi, J. Li, and J. Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, 2006. 2

[6] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell. Language models for image captioning: The quirks and what works. In *ACL*, 2015. 2

[7] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *CVPR*, 2011. 2

[8] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 1, 2

[9] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From captions to visual concepts and back. In *CVPR*, 2015. 1, 2

[10] T.-H. K. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, et al. Visual storytelling. In *NAACL*, 2016. 7

[11] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016. 1, 2

[12] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 1, 2, 3

[13] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *ACL*, 2003. 6

[14] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, 2016. 1, 2, 3

[15] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017. 2

[16] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rapid: Rating pictorial aesthetics using deep learning. In *ACM MM*, 2014. 1, 2, 3

[17] 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 *ICCV*, 2015. 1, 2, 3

[18] M.-T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, 2015. 4

[19] L. Mai, H. Jin, and F. Liu. Composition-preserving deep photo aesthetics assessment. In *CVPR*, 2016. 1, 2

[20] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 1, 2

[21] J. Mao, X. Wei, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *ICCV*, 2015. 1, 2

[22] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *ICLR*, 2015. 1, 2

[23] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *ICCV*, 2011. 2

[24] N. Murray, L. Marchesotti, and F. Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012. 1, 2, 6, 8

[25] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. In *CVPR*, 2016. 2

[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 6

[27] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek. Areas of attention for image captioning. *arXiv preprint arXiv:1612.01033*, 2016. 2

[28] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *EMNLP 4th Workshop on Vision and Language*, 2015. 6

[29] M. Tapaswi, M. Bauml, and R. Stiefelhagen. Book2movie: Aligning video scenes with book chapters. In *CVPR*, 2015. 2

[30] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 6

[31] S. Venugopalan, L. A. Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko. Captioning images with diverse objects. *arXiv preprint arXiv:1606.07770*, 2016. 2

[32] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 1, 2

[33] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 2, 3

[34] R. Xu, C. Xiong, W. Chen, and J. J. Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *AAAI*, 2015. 2

[35] M.-C. Yeh and Y.-C. Cheng. Relative features for photo quality assessment. In *ICIP*, 2012. 2

[36] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *CVPR*, 2016. 2, 3

[37] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *CVPR*, 2016. 2

[38] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *CVPR*, 2015. 2