Aesthetic Critiques Generation for Photos: Supplementary Material

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

This supplementary material contains additional experiments and discussions.

- First, we provide more in-depth analysis to compare different automatic evaluation metrics and to judge whether they are suitable for this novel task.
- Second, we present more experimental results on PCCD and the cross-dataset results on AVA.

1. Limitation of automatic evaluation criteria

As mentioned in our paper, we study a new problem, aesthetic critiques generation, which is different from conventional image captioning. Many recent works [1, 2] started to argued that conventional evaluation criteria (BLEU, METEOR and CIDEr) borrowed from machine translation community are unsuitable for image captioning task. How to choose a suitable criterion is still a tricky problem in image captioning, and this issue is more significant in our proposed new problem. In this section, we provide the results of our approaches on conventional automatic evaluation criteria. Then we give an example to explain why these criteria are improper. Compared to them, SPICE suits better to our task, although human evaluation could be regarded as a more reliable measure.

The results of our approaches on BLEU, METEOR, and CIDEr are presented in Table 1. Compared to recent captioning works, the values shown in Table 1 are quite low. In the beginning, we are curious about why the automatic evaluation results are different from human's (as shown in Section 5.2 of our main paper). So, we start to study whether these automatic criteria are reasonable. First, let us note that the ground truths of common image captioning and our PCCD datasets are different. In the former (e.g., MSCOCO and Flickr30k), an image is described by using multiple similar sentences, and thus the ground-truth captions are near-duplicates for the same image. In PCCD, an image is described by multiple sentences that are dissimilar (or not synonymous), and thus the ground-truth captions are not near duplicates.[1] Second, note that the target of image captioning is objects, while our goal is to produce aesthetic critiques that involve high-level semantic concepts.

One of the problems with using the conventional criteria is that they heavily rely on n-gram matching which could be too picky on literal level when we want to convey the same meaning for two texts. So it is unequitable to judge image captioning with such metrics, let alone the novel photo critique captioning task emphasizing more on expressing abstract concepts. Following is a simple example with its corresponding scores on different criteria shown in Table 2.

(a) the ocean on the bottom of the image is too dark.

(b) the tree on the foreground is too close to the right.

These two sentences express different suggestions on different targets. If these sentences are compared by using any of the previously mentioned n-gram metrics, a high similarity score is obtained due to the presence of the similar structure like 'on the' or 'is too' which has less useful information though they are important components to make the sentences fluent. However, on the semantic level, sentences (a) and (b) convey totally different concepts. They contain distinct subjects (ocean and tree) and adjectives (dark and right) that are critical elements in the photo critiques, and so it is strange that they get such high scores with these metrics. To be honest, it is reasonable for natural language community to use n-gram metrics to ensure the fluency of the sentences for machine translation tasks. But it emphasizes too much on lexical matching that misleads the scoring process for our photo critique captioning task. This appears to be the reason why our AF approach performs worse than CNN-LSTM-WD on these metrics, while it is more favorable on human evaluations (no matter for the AMT, experts and ground-truth comparisons as shown in Tables 3 and 4 of our main paper). On the contrary, SPICE provides a more rational judgement based on its semantic graph matching.

*indicates equal contribution.

[1] This explains why CIDEr, which employs the occurrence frequency of n-grams in the ground-truth captions, is unsuitable for the evaluation of our task as mentioned in the main paper.

Table 1: Automatic evaluation of the proposed approaches.

Method	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	CIDEr
CNN-LSTM-WD	0.245	0.108	0.041	0.007	0.122	0.024
AO Approach	0.221	0.090	0.033	0.004	0.113	0.021
AF Approach	0.233	0.098	0.037	0.007	0.116	0.020

Table 2: Automatic evaluation of sentences (a) and (b).

BLEU1	BLEU2	BLEU3	BLEU4	METEOR	CIDEr	SPICE
0.462	0.196	0	0	0.156	0	0

And it is more intuitive and close to human evaluation so that we choose SPICE as our automatic evaluation criteria.

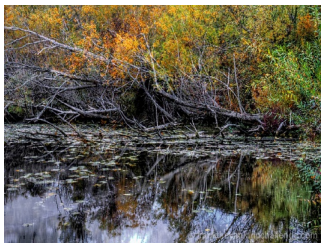
2. Example results of PCCD and AVA dataset

Some image-caption pairs generated by using our AF approach on the PCCD are shown in Figure 1. Besides, we also apply the AF model trained on PCCD to test the AVA dataset, and show the cross-dataset results in Figure 2. It can be seen that the learned model appears to be also useful for generating the photo aesthetic critiques for the AVA dataset, which demonstrates the generalization capability of the proposed approach.

References

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

[2] R. Vedantam, S. Bengio, K. Murphy, D. Parikh, and G. Chechik. Context-aware captions from context-agnostic supervision. In *CVPR*, 2017. 1



i would also like to see a little more of the river and less water

i like the composition of this image the way the bird is looking into the frame

the composition is good and is very close to the rule of thirds

the road to the left of the frame is good but the angle is not as i would have used the view to a more powerful view of the scene

the clouds have very good detail and help maintain and or help the comp

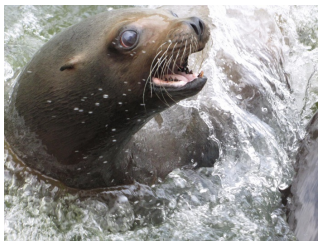
i like the way you have used the rule of thirds and this is a very good example of the colours

the horizon is perfectly placed

i like the composition and tree line

i like the way you have used the trees to frame the subject in the background as the only other element of the image

the horizon line is in the middle of the frame which i would like to see more of the house



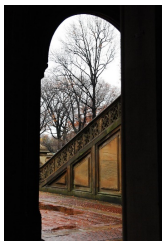
i love the seal and the decision to make it black

the trees on the right add a nice balancing component

i like your use of diagonals generated by the clouds and the structures on the horizon to give a sense of scale

i like the way you have placed the tree on the right so that the eye is drawn to the main subject and the trees on the far left of the frame

the branches in the foreground frame the image nicely though you have a very dark area in the composition that is pleasing to the eye



i like the way you have the elements set up with the eye being drawn to the circle in the upper area

Figure 1: PCCD results: more examples of the critiques generated by using our AF approach on the PCCD.

i like the way the subject is placed in the middle of the frame as it is to the left

i would have liked to have seen the subjects head further to the left of the frame

i like the way you have the buildings peak

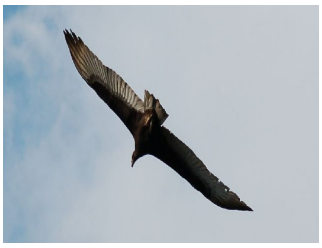
the horizon is in the bottom third of the frame

the color is great but i would have liked to see more detail on the red

the wing position presented here is really great too

i like the way you have used the rule of thirds

i would have liked to see more of the childs face



i like the simplicity of the composition

i like the way the trees on the left of the frame keep the eye in the center of the frame

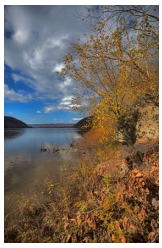
the clouds are dramatic and elegant

i like the way you have used the rule of thirds in this image

the clouds have very good detail and help maintain and or help the comp

i like the way you have used the rule of thirds to the image

the composition is good but i would like to see more of the empty space for the subject



i like the way the tree canopy comes in and curves nicely downwards directing the view towards the headland and the ocean

Figure 2: AVA results: examples of the critiques generated by using our AF model (trained by PCCD) on the AVA dataset.