
# Combining Image Caption and Aesthetic Description Using Siamese Network

Xinghui Song(✉) and Peipei Zhu

School of Artificial Intelligence and Big Data, Guangdong Business and Technology University, Zhaoqing 526000, Guangdong, China
songxinghui@tiangong.edu.cn

**Abstract.** In recent decades, the confluence of CV and NLP technologies has grown in popularity. Many researchers have focused their attention on Image caption task. In recent years, academics have been more interested in image aesthetic description because of image aesthetic indicative of the level. In this study, we present an aesthetic description technique that combines image description and aesthetic description at the same time. We use a Siamese network to acquire datasets for training from two data domains: Image caption task and Image aesthetic description task. The parameters gained from training were migrated back to the conventional Encoder-Decoder model for testing after training. On image caption task, we chose the flickr8k datasets to reduce computing cost. On aesthetic task, the PCCD datasets was used. The final findings indicate that our technique is capable of simultaneously training datasets from two data domains and producing both kinds of image descriptions.

**Keywords:** Image caption · Image quality evaluation · Aesthetic description · Siamese network · Transfer learning

## 1 Introduction

### 1.1 A Subsection Sample

Image captioning (IC) is a typical mutil-modal task transition from image visual content to natural language text, involving related areas such as computer vision (CV) and natural language processing (NLP), and is frequently used in semantic image search and multi-modal image understanding. Traditional IC tasks primarily explain important information about images [1–4], such as the link between an entity's attribute and the entity, Image aesthetic quality evaluation [5] is used to describe the aesthetic information and subjective feelings of images. Image aesthetic description is a small field of image quality assessment, which mainly describes image aesthetic information, such as color, lighting, composition. It was established by Taiwan researchers in 2017 [6]. In this paper, Image caption is referred as NIC (Natural Image Caption) and Aesthetic description is referred as AIC (Aesthetic Image Caption).

It is interesting to feed a computer a diagram that understands the content on the image and gives a description of image. On AIC, most researchers construct datasets


Y. Zhang et al. (Eds.): CENet 2023, LNEE 1125, pp. 41–51, 2024.
https://doi.org/10.1007/978-981-99-9239-3_4

of different scene kinds based on their own knowledge of aesthetic description, using more Encoder-Decoder model framework to fulfill the goal of aesthetic description in the existing study [7–10]. However, the most critical step in most researchers' methods is the construction of datasets, And because of copyright issues, can't open source. At the same time, researchers have different understandings of beauty. Therefore, in this article species. We propose five criteria for aesthetic description of images, Use standard public datasets: NIC task flickr8k datasets, which has over 8000 pictures and over 40,000 linguistic descriptions, to minimize computing costs. More than 4000 pictures and 30,000 English descriptions are included in the AIC task PCCD datasets. The topology of Siamese networks is used to simultaneously train datasets from two data domains using the integrated learning concept. The model parameters were moved to the standard NIC task model [2] after training. The final experimental results demonstrate that our technique effectively learns data from both data domains and simultaneously produces both types of picture description. In Fig. 1, we compare and analyze the picture aesthetic description provided by the current IC and AIC tasks to better demonstrate these two difficulties.

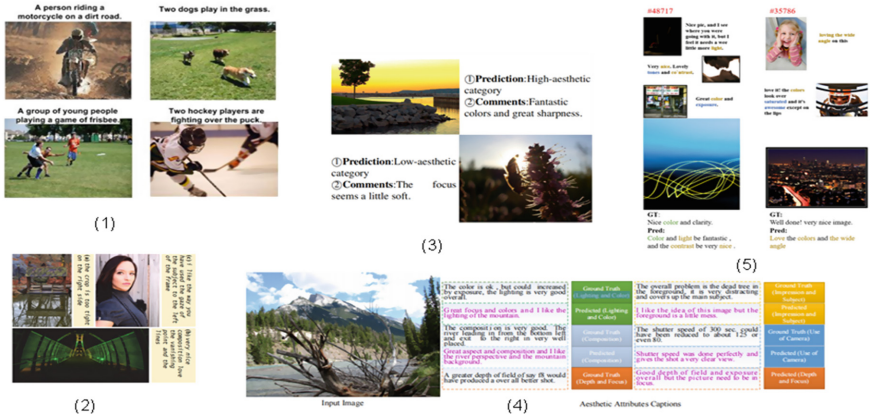


**Fig. 1.** Results of NIC task and AIC task at present

The image's important information, such as the entity's characteristics and the relationship between the entities, is depicted in Fig. 1 (1). The objective of (2) is to produce aesthetic remarks on the image, and the goal of (2–5) is to explain the evaluation of this image, sentiments, affection, and so on. (3) is tasked for perceiving visual aesthetics and generating human commentary and interpretations. The aim of (4) is to provide titles for each of the image's attributes, including color and light, composition, depth, and focus. (5) The objective is to provide the title of tailored picture aesthetic duties for various consumers.

Figure 2 The experimental results of the present image description and aesthetic description


**Fig. 2.** Part of achievement display

Our main contributions include:

1. We present aesthetic description methods combining image descriptions and, to our knowledge, we are currently the first to do this work.
2. We use Siamese networks to effectively learn datasets from two data domains, the NIC and AIC tasks, and achieve image description Combining image caption and Aesthetic description results. In our tests, we had a good rate of over 65%.
3. We implemented image descriptions exporting both styles on the model.

## 2 Related Work

Because the aesthetic description encompasses a wider range of areas, the relevant work is divided into three categories.

### 2.1 Image Aesthetics Quality Evaluation

Researchers began using deep learning for automated picture feature extraction in 2014, when deep learning algorithms matured [11–16].When compared to typical manual design characteristics, using deep learning has a far higher categorization accuracy. Google proposed the NIMA model [13] in 2018. The NIMA model produces a hierarchical distribution—from 1 to 10 for any given image, presenting better predictions of human preference, reaching 81.51% on the AVA datasets. 2019. Xin Jin et al., On the AVA aesthetic classification datasets, the suggested ILGNet obtains a classification accuracy of 85.53%. The researchers' emphasis is expanded to additional factors when they have developed a pretty sophisticated aesthetic judgment. Kong et al. [15] published a study

in 2016. A new image aesthetic datasets, AADB, comprises eight aesthetic variables, and some work has been built on AADB to assess the influence of aesthetic components on picture aesthetic quality, as well as related work on image aesthetics.

### 2.2 Image Caption

Image Caption, as the name implies, creates a descriptive text from a photograph, which is a difficult process. The areas of CV and NLP have made significant progress in producing text and comprehending pictures and video during the last two years. While both areas include techniques that are comparable to AI and machine learning, they have grown independently in the past and have had little contact in the scientific community. In Google CVPR2015 year papers [2], however, there has been a spike in problematic interest in the need to integrate linguistic and visual information in recent years. In natural language processing (NLP) machine translation, the Encoder-Decoder is used [17]. It connects better with recent breakthroughs in computer vision and machine translation by replacing the original Encoder RNN with the CNN structure utilized in the image. Deep learning techniques were used to turn images into text descriptions. In this new linguistic vision community, auto-description has become a key responsibility. Cho et al. [18] suggested an Attention method that does not employ unified semantic features and allows a Decoder to freely select the desired features in the in-put sequence. Xu and colleagues [3]. The Attention mechanism was used to improve the basic Encoder-Decoder system. Specifically, we employ CNN's spatial properties to extract a feature for each of the image's 196 positions, allowing Decoder to select these 196 position features during decoding [19]. The total number of labels for each image was first collected from the C words that appeared most frequently in all descriptions, and the training data for each image was obtained directly from its reported words. Chen and colleagues [20] Many improvements to Decoder RNN's structure, allowing the RNN network to not only translate picture features into text, but also to extract image features from text, all while boosting speed. 2019 [4] Starting with natural language, picture caption is proposed at the word level by predicting noun chunk sequences while carefully analyzing visual and linguistic variations, offering further grounding. External signals are in charge 2020 [1], Using abstract scene plots (Abstract Scene Graph, ASG), simultaneously controlling the desired expressed objects, properties, and relationships through the graph structure can not only reflect the user's fine-grained description intent, but also generate more diverse image descriptions.

### 2.3 Aesthetic Description Task

The image aesthetic task is divided into five tiers of tasks in the image aesthetic quality evaluation system, including aesthetic distribution, aesthetic score, aesthetic distribution, aesthetic components, and aesthetic description. The task was Image aesthetic/photography skills related titles for specific aesthetic aspects of color, clarity, and composition of images. The first datasets, PCCD. For the task was produced Image aesthetic/photography skills related titles for specific aesthetic aspects of color, clarity, and composition of images. Microsoft in 2018 [21] Poetically generated poetry from

images needs to satisfy both the correlation with the image and the rules of poetic language. Ghosal et al. [8] Following Chang's work, we propose a title filtering strategy, which has compiled a cleaner, larger dataset AVA-Caption, and proposed a strategy for training convolutional neural networks that applies the LDA topic model to reviews and learns CNN parameters by fitting the topic distribution. In Jin [7], Kun et al. [6] used a larger datasets of DPC-caption; with change in their analysis. The paper comparison produced a description and score for the image aesthetic of 5 criteria (light/color, composition, depth of field/focus, theme/impression, camera use). According to Kun et al. [10], the network can simultaneously output a description of many dimensional features. A new personalized aesthetic image title (PAIC) approach for gathering and combining user preferences and improving aesthetic features for AIC tasks will be proposed in the study.

## 3 Methods

### 3.1 Training Stage

We chose to adopt [3] based on the CNN + LSTM + ATTENTION framework, with the caveat that it is generic and can be used for any framework-based Image Caption task in which visual features are extracted from ResNet101 networks trained with ImageNet and passed as input to LSTM, with the traditional image captioning dataset NIC captioning C1 and image aesthetic description captioning C2 to simultaneously at For our framework, we use a twin CNN network with shared parameters on the training visual model, trained on data from both datasets, trained training. Finally, our model training phase is shown in Fig. 3.

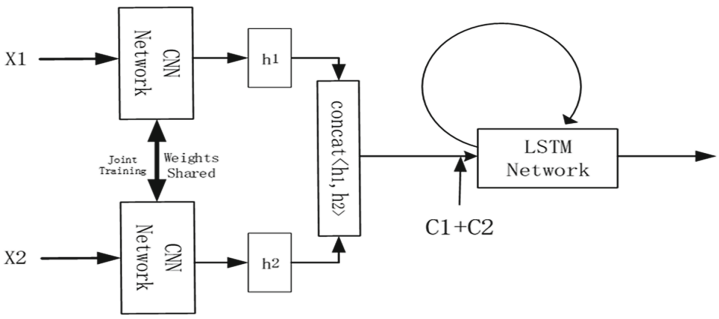


**Fig. 3.** Based on the twin network image description model

In Fig. 3, X1 represents the NIC image, X2 represents the AIC image, images from two datasets, after a shared parameter CNN feature extractor, yielding two eigenvectors h1, h2. The data splicing with the concat function yields an eigenvector h0. The h0 contains information about the two images.C1 represents the description of image X1 of NIC and C2 represents the description of AIC image X2.The C1 and C2 data were spliced and sent to LSTM training along with h0.Training yielded shared CNN parameters, and LSTM parameters.

### 3.2 Test Phase

The used testing phase [3] for CNN + LSTM + ATTENTION, is shown in Fig. 4.

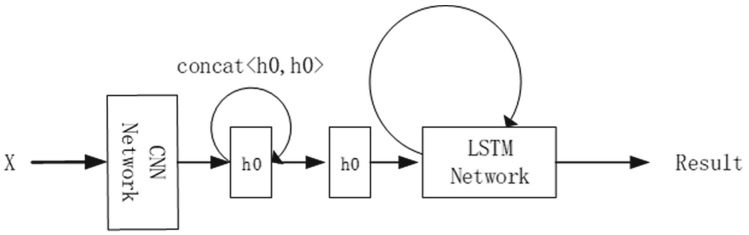


**Fig. 4.** The image caption generation model

We migrated the parameters trained by Fig. 3 model to model 4. In Fig. 4, X represents our test images, and the images passed via the CNN feature extractor to get the vector h0, In order to remain consistent with Fig. 3—1 model dimensions, we replicate h0. The h0 was data stitched and then sent to the LSTM network to obtain the predicted results.

### 3.3 Training Details

The image feature extractor CNN was partially performed using the pre-trained Resnet101 network. The text generator uses the LSTM network. Increase the prediction effect using the Attention mechanism.

There are over 8000 photos and over 40,000 linguistic descriptions in the data from two datasets called flickr8k. More than 4,000 photos and 30,000 linguistic descriptions are included in the PCCD datasets. A json dictionary is used to store the first description of the two datasets. After doing embedding, C1 (which represents picture description) and C2 (which represents aesthetic description) were connected. The first problem encountered was the pairing problem of the two datasets. We provide two viable methods. The first, the few datasets replicate themselves. Mount data into datasets with loader and reload part of list, to data balance. The second, multiple datasets remove the excess. The balance parameter is set in the code. By True (means less copy itself), False means more removed excess. The model during tested, entering an image and then experiencing encoder, requires double at concat connection so that the data dimensions can be kept consistent before being sent to the decoder.

## 4 Experiments

### 4.1 Experimental Data

The NIC image captioning datasets uses the flickr8k datasets to reduce computational overhead. The data contained 8000 images, each paired with five descriptions. These descriptions provide a content description of the objects and events in the picture. The AIC image captioning datasets used PCCD (Photo Critique Captioning Datasets): This

datasets is generated from [6] Introduced, and is based on www.gurushots.com. Professional photographers have provided seven comments on the uploaded photos: general impression, composition and perspective, color and lighting, subject of photo, depth of field, focus and use of camera, exposure and speed.

### 4.2 Experiment Results and Analysis

Training indicators:

The train_losses_step: training set changes in the loss function with increasing step length.

The train_accs_step: training set changes accuracy with step length.

The train_losses_epoch: training sets change in the loss function with increasing epoch.

The train_accs_epoch: training set changes in accuracy with increasing epoch.

The val_losses_step: validation sets change in the loss function with increasing step length.

The val_losses_epoch: validation set changes in accuracy with increasing step size.

The val_accs_step: validation sets change in the loss function with increasing epoch.

The validation set of val_accs_epoch: changes in accuracy with increasing epoch (Figs. 5 and 6).

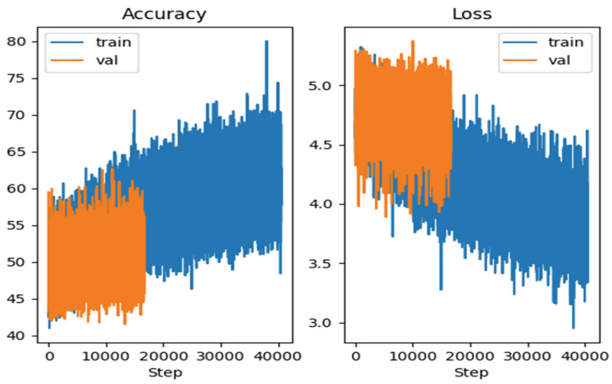


**Fig. 5.** Accuracy, loss curve with step

From the above two graph observations, the training set accuracy and loss are in oscillatory convergence, and the validation set accuracy and loss are also in oscillatory convergence. The accuracy is 62.5%, the main reason for this result is that the text we generated is first diverse and novel, and at training as a random combination of two datasets, the equivalent to each epoch is a brand new dataset in training.

Subjective evaluation results:

For the subjective evaluation, 200 images were randomly selected from the flickr8k datasets, and 100 images from the PCCD datasets were tested. The test results have been uploaded to https://github.com/SongANIC/SampleANIC.

We divided the evaluation results into three levels of perfect, good, general, poor.

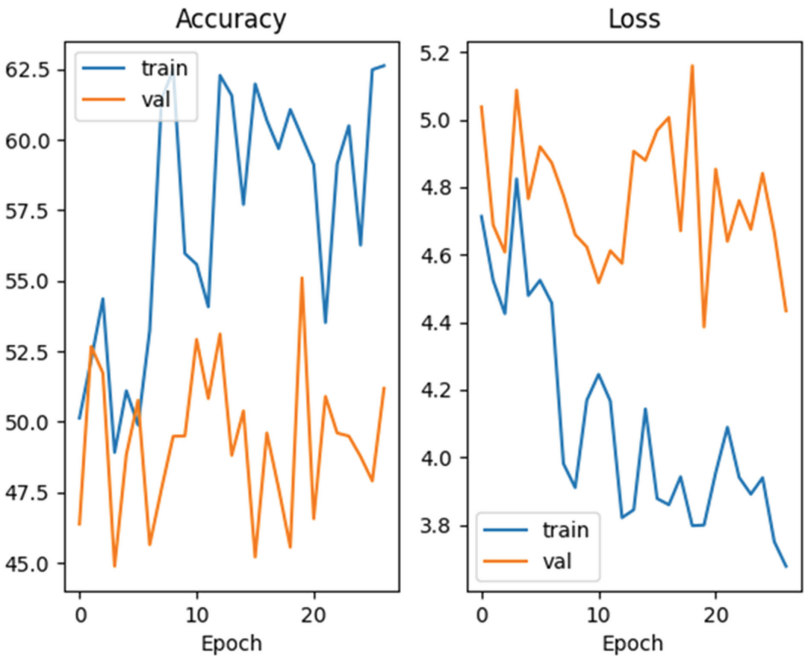


**Fig. 6.** Accuracy, loss curve with epoch

The standard of good (Good) is to accurately describe the image description and the aesthetic description.

The general (General) standard is, there are certain errors, but you can describe the image description more accurately, you can describe the aesthetic description.

Differential (Bad) criteria, poor statements, no sentence or generate only one style description.

The test result was generated with 300 images:

good images of 138, general 71, and bad 91.

Overall generated images meeting our requirements of $138 + 71 = 209$ bad 91 sheets.

The qualification rate is currently at 69.67%. There will be many flaws and unobjective elements in subjective judgment because of its subjectivity. The experimental results have been posted to https://github.com/SongANIC/SampleANIC to make future study and analysis easier.

First, there are four types of images:

Good standard: able to accurately express visual descriptions and smooth statements while also describing the aesthetic description (Fig. 7).

General (General) standard: there are certain errors, but you can more accurately describe the image description, can describe the aesthetic description (Fig. 8).

Bad standard: poor statements, no sentences or only one style description (Fig. 9).


**Fig. 7.** Generates a AIC aesthetic image describing the good classification


**Fig. 8.** Generates a AIC aesthetic image describing the general classification


**Fig. 9.** Generates a AIC aesthetic image describing the bad classification

## 5 Conclusions

For present aesthetic description informative challenges, we offer an aesthetic description strategy combining image description to create an end-to-end neural network system designed to combine the critical visual information of the image. In the picture feature extraction component of our system, we use a twin network. The concept of creating a twin CNN network with identical parameters. Allows it to receive picture features from

both datasets at the same time and learn how to extract the image's primary visual information as well as the aesthetic description's weakly supervised multi-level information. Our model is also nimble and light, equating to simultaneously extracting picture key and aesthetic information with a CNN feature extraction network, thanks to the same parameter characteristics. So when it comes to testing, all we need is a CNN network and an input image to get our results. We incorporate the language description of traditional image description from the NIC datasets and image aesthetic description from the AIC datasets into the text output model, and simultaneously input both styles of language description into the Language generation model LSTM for training, so that the computer realizes a network can output both styles of language description at the same time. If there are enough datasets and various styles of language descriptions, our model idea should be able to output more than two kinds of language descriptions at the same time. At the same time, as the datasets grows and the accuracy of the annotation data improves, our model's generalization performance and text output capability improve.

## References

1. Chen, S., Jin, Q., Wang, P., Wu, Q.: Say as you wish: fine-grained control of image caption generation with abstract scene graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9962–9971 (2020)
2. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2015)
3. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057. PMLR (2015)
4. Cornia, M., Baraldi, L., Cucchiara, R.: Show, control and tell: a framework for generating controllable and grounded captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8307–8316 (2019)
5. Jin, X., Zhou, B., Zou, D., et al.: Image aesthetic quality evaluation technology development trend. Sci. Technol. Guide **9**, 36–45 (2018)
6. Chang, K.Y., Lu, K.H., Chen, C.S.: Aesthetic critiques generation for photos. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3514–3523 (2017)
7. Jin, X., Wu, L., Zhao, G., Li, X., Zhang, X., Ge, S., Zou, D., Zhou, B., Zhou, X.: Aesthetic attributes assessment of images. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 311–319 (2019)
8. Ghosal, K., Rana, A., Smolic, A.: Aesthetic image captioning from weakly-labelled photographs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, p. 0 (2019)
9. Wang W, Yang S, Zhang W, et al.Neural aesthetic image reviewer[J].IET Computer Vision, 2019,13(8):749–758
10. Xiong, K., Jiang, L., Dang, X., Wang, G., Ye, W., Qin, Z.: Towards personalized aesthetic image caption. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2020)
11. Mai, L., Jin, H., Liu, F.: Composition-preserving deep photo aesthetics assessment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 497–506 (2016)

12. Jin, X., Wu, L., Li, X., et al.: ILGNet: inception modules with connected local and global features for efficient image aesthetic quality classification using domain adaptation. IET Comput. Vis. **13**(2), 206–212 (2019)
13. Talebi, H., Milanfar, P.: NIMA: neural image assessment. IEEE Trans. Image Process. **27**(8), 3998–4011 (2018)
14. Lee, H., Hong, K., Kang, H., et al.: Photo aesthetics analysis via DCNN feature encoding. IEEE Trans. Multimedia **19**(8), 1921–1932 (2017)
15. Kong, S., Shen, X., Lin, Z., Mech, R., Fowlkes, C.: Photo aesthetics ranking network with attributes and content adaptation. In: European Conference on Computer Vision, pp. 662–679. Springer, Cham (2016)
16. Schwarz, K., Wieschollek, P., Lensch, H.P.: Will people like your image? Learning the aesthetic space. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 2048–2057. IEEE (2018)
17. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation (2014). arXiv preprint arXiv:1406.1078
18. Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate (2014). arXiv preprint arXiv:1409.0473
19. Wu, Q., Shen, C., Liu, L., Dick, A., Van Den Hengel, A.: What value do explicit high level concepts have in vision to language problems?. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 203–212 (2016)
20. Chen, X., Zitnick, C.L.: Learning a Recurrent Visual Representation for Image Caption Generation (2014). arXiv preprint arXiv:1411.5654
21. Liu, B., Fu, J., Kato, M.P., Yoshikawa, M.: Beyond narrative description: Generating poetry from images by multi-adversarial training. In Proceedings of the 26th ACM International Conference on Multimedia, pp. 783–791 (2018)

## The Introduction of Author

22. Song Xinghui: Male, born in 1994, graduated from the School of Computer Science and Technology, TianGong University with a master's degree. Now working in Guangdong Business and Technology University: Once published the paper "CNTK communication optimization based on parameter server", "Research on gene coexpression network based on RNA-seq data", etc.
23. Zhu Peipei: Master degree, now working in Guangdong Business and Technology University.