AESTHETIC VISUAL QUESTION ANSWERING OF PHOTOGRAPHS

Xin Jin[1], *Yuchen Li*[1], *Wu Zhou*[1], *Xinghui Zhou*[2], *Hongtao Yang*[3*]

[1]Beijing Electronic Science and Technology Institute,
[2]University of Science and Technology
[3]Communication University of China
[*]Corresponding Author's Email: yhtbyr@163.com

ABSTRACT

Aesthetic assessment of images can be categorized into two main forms: numerical assessment and language assessment. In this paper, we propose a new task of aesthetic language assessment: aesthetic visual question and answering (AVQA) of images. We use images from *www.flickr.com*. The objective QA pairs are generated by the proposed aesthetic attributes analysis algorithms. Moreover, we introduce subjective QA pairs that are converted from aesthetic numerical labels and sentiment analysis from large-scale pre-train models. We build the first aesthetic visual question answering dataset, AesVQA, that contains 72,168 high-quality images and 324,756 pairs of aesthetic questions. This is the first work that both addresses the task of aesthetic VQA and introduces subjectiveness into VQA tasks. The experimental results reveal that our methods outperform other VQA models on this new task.

Index Terms— aesthetic dataset with attributes, multi-tasking, external attribute features, ECA channel attention

1. INTRODUCTION

In the researches of computer vision and natural language processing, researchers always focus on direct and obvious goals instead of image aesthetics with different attributes. In the past decade, many researchers were interested in image aesthetics. However, the popular methods were to score images and describe attributes of images. Either giving a score ranging from 0 to 10 to evaluate aesthetics of images or describing a images' attributes[1], such as color, lighting, and composition, in one sentence are limited and lack description in details.

The state-of-the-art model of visual question answering (VQA) has achieved multiple purposes on many large-scale datasets. According to the category of the datasets, the subtask of VQA can be divided into task-based, inference-based, and text-based, like TextVQA [2] dataset. Many famous models [3, 4, 5] have been developed in these visual question answering tasks. However, they lack the overall investigation and reasoning of the picture.

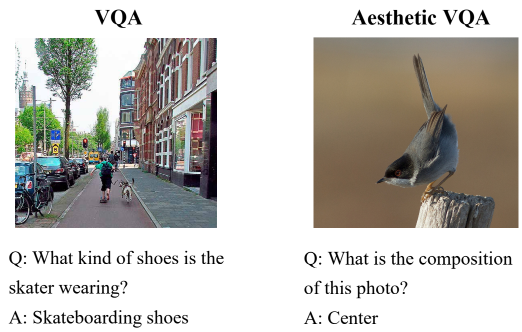

Fig. 1. VQA and Aesthetic VQA. The AVQA focuses on the overall aesthetic attributes of the image such as shot, lighting and colors.

Most of current VQA researches focus on objective QA pairs. This paper mainly addresses aesthetic visual question answering (AVQA) of images, as shown in Figure 1. This task is both aesthetic and subjective; also, it is a complementary task for the previous QA studies. This problem is important since subjective questions and answers are very prevalent among human discourses. In this paper, we present a dataset with aesthetic QA pairs, aesthetic visual question answering (AesVQA). Images in the new dataset are all marked with aesthetic labels through series of multiple computer vision subtasks. These labels cover composition, color, subject, lighting, genres, techniques and emotions of photos.

To obtain reliable labels, we develop some unsupervised and semi-supervised computer vision methods. At the same time, we propose a method and design a small amount of manual labelling image adjustment function to get a uniform and credible dataset distribution.

In particularly, the labels of lights and colors are designed based on the specific objects in the image. The labels of scene and composition are designed based on the whole picture. The other labels based on subjective evaluations of images, especially techniques and emotions, provide extra dimensional information in question and answering. The task

Table 1. VQA datasets' answer words length accounted for the proportion of all answers table. Answers with two words or more are 94.6% in AesVQA dataset, which leads a more difficult VQA task.

	The Length of The Answer Words			
Dataset	One Word	Yes/No	Two Words or More	The Number of QA Pairs
VQAv2 [6]	51.2%	45.0%	3.8%	1,105,904
Visual7W [7]	51.2%	21.2%	27.6%	139,868
Visual Genome [8]	52.5%	21.2%	26.3%	1,445,233
AesVQA	5.4%	**0%**	**94.6%**	324,756

becomes more challenging by adding some human psychological evaluations. The main contributions of this paper can be summarized as:

- **New Task**: The new task of image aesthetics question and answer, and it is the first work that introduces subjective answering to VQA problems.
- **New Labels**: Semi-automatic QA labelling of images gets more information when answering and methods of dataset distribution adjustment.
- **New Points of VQA**: The traditional visual question answering method based on objective detection is optimized to half-object-detection-based (such as lighting and color labels) and half-whole-image-based (such as subject and composition labels).

2. RELATED WORK

Aesthetics and computer vision are inextricably linked, and people are demanding higher quality of image data. How to process the images into a form that is more compatible with human preferences has become a critical issue. It is feasible to use computer technology to study the aesthetics of images[9].

Visual question answering is a difficult problem that straddles computer vision and natural language processing, and its task requires the extraction of not only image features but also textual partial features. Recent VQA datasets such as DAQUAR[10], Visual Madlibs[11], FM-IQA[12], etc. are available. Unlike look-and-talk tasks, simply fusing image and text features often does not yield the desired features, i.e., answers.

Visual question answering requires the model to process from image and language to combine and give the answer effectively. So, VQA is both interdisciplinary and challenging. Also, due to characteristics of image aesthetics, there is no professional image aesthetics Q & A(Question and Answer) dataset. Therefore, it is important to propose a common aesthetic Q & A datasets. By the visual model, we use classical Q& A datasets as reference to generate our dataset. The classic Q & A datasets are always generated by machines, such as: TextVQA[2], EST-VQA[13] and VizWiz-VQA[14].

One of the difficulties of the image aesthetics task is the disadvantage of inaccurate subjective evaluations but low number of objective ones. Borrowing a large amount of data as a foundation, it will be possible to mine a sufficient number of images and corresponding comments with a high enough standard, and then further convert the comments into the desired Q&A pairs. Proxy objectification with statistical features of subjective evaluation is a common approach in current image aesthetics tasks.

To date, there is no comprehensive aesthetic VQA dataset. Previous research work proposed the AQUA[15] dataset, which focuses more on studying the artistic aspects of paintings, while our dataset focuses more on the aesthetic aspects of photographs. In previous studies, researchers would apply attention mechanisms to images or text to obtain better results, but it is difficult to obtain the desired features from the many image features because of the different feature spaces and the presence of tensor features with ultra-high dimensionality of images.

3. AESTHETIC VISUAL QUESTION ANSWERING

To study the answers to questions related to image aesthetics and the overall content of the image, and to reduce the cost of manual annotation, we proposed AesVQA, as shown in Figure 2 and Table 1. This dataset is an image aesthetic question and answer dataset based on an unsupervised model. We will start by describing how to select the images used in AesVQA. Then, we explain our data collection pipeline to collect questions and answers.

3.1. Images of AesVQA

We use pictures from the Flickr website as our source pictures, which conforms to developing a VQA model based on image aesthetics and the overall content of the image. We are most interested in pictures in categories such as landscapes, people, still lifes, and animals. Several categories in the Explore section of the Flickr website meet this condition.

To automate this process to identify categories that tend to contain images with text, we select 100 random images for each category (if the maximum number of images for that category is less than 100, select all images). We run the state-of-the-art OCR model Rosetta [6] on these images and calculate the average number of OCR boxes in the category.

Q: What level is the composition of this picture?

A: Good composition

Q: What is the photo represent?

A: Happy

Q: How to evaluate the picture?

A: Beautiful shot

Fig. 2. Pictures, Questions and Answers in the Aesthetic VQA dataset.

Based on the 190,000 Flickr images we crawled, after artificially filtering out abstract images and unrealistic images from image photography, filtering these (and a small amount of noisy data from manual annotations) can obtain 72168 images. image. Manual screening requires that the image must be taken from the real world, not obtained through PS or other software editing tools.

3.2. Generation and filtering of AesVQA

In order to generate question and answer pairs, we designed our own model. Our model based on sequence to sequence method, including three modules: text content encoder, answer encoder and decoder. In addition, the quality of the generated AesVQA data is uneven, and there is a big gap with the common VQA dataset. Therefore, the data needs to be screened again after QA is generated. We propose an aesthetic question and answer filtering method using LDA[16].

The text content coder is used to process the input comments. At the same time, it needs to calculate the attention weight of different words and the weight of specific words in the replication mechanism. The answer encoder is used to process the selected answer words or phrases, filter the words according to the attention weight, and encode the words into a format suitable for the decoder. The decoder is responsible for using the cyclic neural network to process the input tensor data, and generating questions according to the weight data between the answer and the comment.

For the generated question and answer pairs, the manual filtering method is used to remove the inappropriate question and answer pairs, so as to ensure that at least one question and answer pair generated by each picture is reasonable. These question and answer pairs are formed into a small text set, and the correlation between the text set and multiple topic words obtained from previous LDA topic calculation is calculated, which is used as the threshold to evaluate whether a question and answer pair meets the aesthetic question and answer model.

Calculate the relevance between the remaining sentences and the topics selected by LDA, and judge whether the relevance of the selected sentences is greater than the topic threshold. When the relevance value of the selected sentence is greater than the threshold, it is determined as a question and answer pair that meets the requirements, otherwise it is determined as a question and answer pair that does not meet the requirements.

Finally, since the quality of the generated issues is not very high, the solution is to make manual corrections.

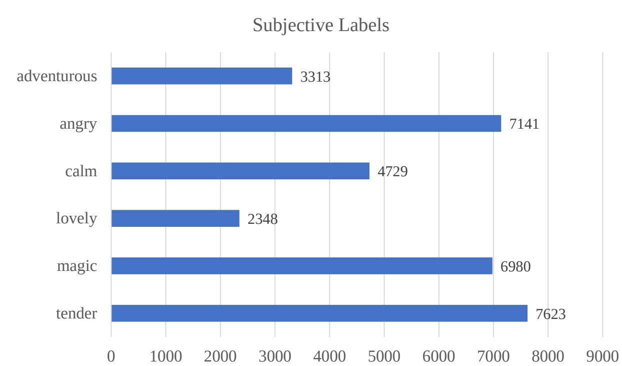

Fig. 3. The distribution of photos's subjective labels in the AesVQA dataset. Including tender, magic, lovely, calm, angry, adventurous and many other attributes.

3.3. Basic Aesthetic Labels

The first part of the AADB[17] dataset is the basic image aesthetic tags, including composition, color, and subject. Among them, in order not to reduce the VQA task to a simple classification task, and to ensure that the specific image attributes are associated with specific objects, we design light labels and color labels to be related to the content of the image, while composition labels and subjects labels are related to the overall image.

Among them, the symmetry of the composition labels is obtained by the SSIM[18] calculation of the image. When the left and right parts of the image or the upper and lower parts of the image after the mirror symmetry, the SSIM is higher than the threshold, and it is judged to be an asymmetric image. The rule of thirds and the center composition use the faster R-CNN based on the visual genome to obtain the position of the most distinctive object in the image.

Color labels depend on the type and color of objects in the image obtained by faster R-CNN. To enrich the diversity of colors, we have added warm tone and cold tone, using target detection to cut out the selected objects in the image, and classify their average colors, and judge based on the correlation between the average color and other known main colors warm tone or cold tone.

The attributes of the subject are labeled using the CLIP[19] pre-training model. According to the classification results of the subject on the AVA[20], it can be known that CLIP can deal with the category labeling of multiple subjects. At the same time, to make up for the errors that may be caused by automatic marking, some manually marked image attributes are subsequently used to correct the results.

3.4. Distribution in VQA Dataset

The distribution of the dataset is extremely important for the VQA task. Some work [21, 22, 23] have proved it. For the aesthetic question and answer task based on the whole image, three dimensions of data distribution need to be considered: the distribution of the whole picture and the specific objects in the image in the question, the distribution of pictures with obvious objects, and the overall scenery in the data set, the answer to the question is Distribution under the influence of subjective factors.In response to the impact of data distribution in Chapter Four, we add a subjective evaluation. Mark the content of the picture as emotionally related.

We show the distributions of the basic aesthetic attribute labels, the genre labels, the subjective labels and photography labels in Appendix.

3.5. Subjective Labels and Photography Component

To overcome the possible impact of data distribution, we divide the attributes of images and aesthetics into subjective feelings and objective photography techniques. The subjective description comes from people's direct perception of the image, and the label of this part comes from the pre-training label of CLIP. Including tender, magic, lovely, calm, angry, adventurous, and many other attributes. The reason for selecting these words is that the results of the sample survey show that these words can get a more balanced data distribution.

On the other hand, we provide related attributes of photography, including genres and techniques, which contain multiple subcategories. Same as subjective, we use CLIP to mark images. Some properties are shown in Appendix. Many attributes of this part have the limitation of vague definition. Therefore, each category has a confidence level. Only when the confidence level is high enough, the mark of the picture is considered effective.

3.6. Bias of Data

Because of the use of unsupervised models such as CLIP to predict the types of pictures, how to design the distribution of options in answers for pictures is the key.

Usually, we will give up pictures with too high an accuracy rate and too low an accuracy rate, that is, a picture can be judged as B category on the A attribute, or it is difficult to judge which category it is. We use the design threshold and the accuracy of a certain type of problem model to estimate how the threshold of this type of problem should be, including two parameters, the mean and variance. Only when the mean and the standard deviation of answers' distribution meet the accuracy of this type of baseline model (LXMERT for example) higher than 50 %, answers contribution is efficacious.

$$\text{accuracy}(V, Q, A, D(\text{avg}), D(\text{std})) > 0.5 \quad (1)$$

where $D(\text{avg})$ means the average of the answers' distribution, the $D(\text{std})$ means the standard deviation of the answers' distribution. 72,168 pictures satisfy the above formula.

4. EXPERIMENTS

In the experiment, we used three models to train and test each category. The ratio of the training set and validation set is 8:1, and the number of test sets is the same as the validation set. To test the influence of data distribution on the model, we constantly adjusted the proportions of each sub-category during training and obtained more balanced training results.

The data distribution without artificial interference comes from all pictures. Such a data distribution will not be conducive to the training of the neural network. The operation of adjusting the data set mainly includes: adjusting the confidence threshold of each sub-category in a large category, and the distribution of the question and answer results of the pictures in each large category. The number of pictures is limited by the different types of pictures that have different requirements in the selection process, so it is impossible to achieve close results. The results of the comparative experiment with AoC (Adjustment of Confidence) and AoDA (Adjustment of the Distribution of Answers) are shown in Appendix. It can be seen that UNITER, based on multiple datasets has achieved the best results in AesVQA.

In order to comprehensively compare the performance of the methods proposed and introduced in this paper, several baseline models are designed in the experiment, and our final

Table 2. VQA model accuracy

	5 reference ans	10 reference ans
VGG19+Single LSTM	37.1%	22.4%
VGG19(ImageNet)+LSTM	40.9%	27.3%
VGG19+LSTM(without Aesthetic Loss)	43.2%	32.5%
VGG19+LSTM+ Aesthetic Loss	**46.6%**	**35.7%**

structure is proposed. Each baseline model and final structure is described below.

- **VGG19 (ImageNet) +LSTM**: VGG19 (ImageNet) +LSTM is a model for migration learning directly on ImageNet data set without parameter freezing, which does not use AVA data set. But using the loss function of image aesthetics.
- **VGG19+LSTM without AESLoss**: VGG19 (ImageNet) +LSTM is a model of migration learning using AVA data set for parameter freezing. The loss function of this model is only cross entropy loss function.
- **VGG19+Single LSTM**: The traditional visual question answering method based on objective detection is optimized to half-object-detection-based (such as lighting and color labels) and half-whole-image-based (such as subject and composition labels).
- **VVGG19+LSTM without AESLoss**: VGG19 (ImageNet) +LSTM is a model of migration learning using AVA data set for parameter freezing. The model uses a loss function for image aesthetics..

The accuracy of each model is shown in the table 2.

4.1. Adjustment of Confidence

Taking genres and techniques as examples, we adjusted the confidence of different categories through training and test results, called it adjustment of confidence (AoC). For different genres and techniques, it is generally believed that the judgment criteria are also different, which is reflected in the labeling process with unsupervised learning as the main method, and the confidence level needs to be adjusted according to the specific class. As is shown in Appendix, The adjusted result is artificially limited to the range of 0.3 to 0.7, this range will help the training and prediction of the model.

4.2. Adjustment of Distribution of Answers

Limited to automatic labeling, our proposed dataset requires multiple models to label images. To prevent the occurrence of some over-fitting phenomena, ten answers are set for each question. For adjusting the distribution of answers (AoDA), we let questions have no controversial answers and salient answers.

The salient answers mean that it is obvious to know which is the right choice in all options. The controversial answers mean there are two or more answers with high confidence. For example, a picture may be classified in ”landscape” and ”wildlife” at the same time, such a picture needs to be removed, as shown in Appendix.

The adjustment of the distribution of answers will help eliminate images in the data set that lack diversity in labels. These images will affect the model's preference and make the model more prone to overfitting. The example in Appendix illustrates which pictures should be discarded and which should be kept during adjustment.

5. CONCLUSIONS

In this paper, we propose and solve a new task of aesthetic quality evaluation: VQA of image aesthetics. This paper constructs the VQA dataset for image aesthetics. We get basic image aesthetics labels, genres labels, and techniques labels from photography, and subjective emotional labels, and transfer these labels to question-answers pairs. We used three advanced VQA models for training and testing and obtained further performance improvement by adjusting the confidence in image classification and the distribution of answers in VQA. It can be proved by experiments that exporting the answer to aesthetic questions and answers is available by training on our datasets. As a result, display performance model and the advantage of databases, which filled in the blank in the field of aesthetics in the VQA.

6. REFERENCES

[1] Xin Jin, Le Wu, Geng Zhao, Xiaodong Li, Xiaokun Zhang, Shiming Ge, Dongqing Zou, Bin Zhou, and Xinghui Zhou, “Aesthetic attributes assessment of images,” in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 311–319.

[2] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach, “Towards vqa models that can read,” in *Proceedings of the IEEE/CVF Conference on Com-*

puter Vision and Pattern Recognition, 2019, pp. 8317–8326.

[3] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al., "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *European Conference on Computer Vision*. Springer, 2020, pp. 121–137.

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu, "Uniter: Universal image-text representation learning," in *European Conference on Computer Vision*. Springer, 2020, pp. 104–120.

[5] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[6] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh, "Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei, "Visual7w: Grounded question answering in images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4995–5004.

[8] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.

[9] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang, "Rating image aesthetics using deep learning," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2021–2034, 2015.

[10] Mengye Ren, Ryan Kiros, and Richard Zemel, "Exploring models and data for image question answering," *Advances in neural information processing systems*, vol. 28, 2015.

[11] Licheng Yu, Eunbyung Park, Alexander C Berg, and Tamara L Berg, "Visual madlibs: Fill in the blank description generation and question answering," in *Proceedings of the ieee international conference on computer vision*, 2015, pp. 2461–2469.

[12] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu, "Are you talking to a machine? dataset and methods for multilingual image question," *Advances in neural information processing systems*, vol. 28, 2015.

[13] Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang, "On the general value of evidence, and bilingual scene-text visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10126–10135.

[14] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al., "Vizwiz: nearly real-time answers to visual questions," in *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 2010, pp. 333–342.

[15] Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura, "A dataset and baselines for visual question answering on art," in *European Conference on Computer Vision*. Springer, 2020, pp. 92–108.

[16] David M Blei, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[17] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *European conference on computer vision*. Springer, 2016, pp. 662–679.

[18] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021.

[20] Naila Murray, Luca Marchesotti, and Florent Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2408–2415.

[21] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel, "Unshuffling data for improved generalization," *arXiv preprint arXiv:2002.11894*, 2020.

[22] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4971–4980.

[23] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi, "Swag: A large-scale adversarial dataset for grounded commonsense inference," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 93–104.