TOWARDS UNIFIED AESTHETICS AND EMOTION PREDICTION IN IMAGES

Jun Yu[1], *Chaoran Cui*[2,⋆], *LeiLei Geng*[2], *Yuling Ma*[1], *Yilong Yin*[1,⋆]

[1]School of Computer Science and Technology, Shandong University
[2]School of Computer Science and Technology, Shandong University of Finance and Economics
yu.jun@mail.sdu.edu.cn, crcui@sdufe.edu.cn, leileigeng_njust@163.com,
mayuling@mail.sdu.edu.cn, ylyin@sdu.edu.cn

ABSTRACT

Aesthetics assessment and emotion recognition are two fundamental problems in user perception understanding. While the two tasks are correlated and mutually beneficial, they are usually solved separately in existing studies. In this paper, we resort to multi-task learning to deal with aesthetics assessment and emotion recognition for images in a unified framework. Towards this goal, we extend a large scale emotion dataset by further manually rating the aesthetic qualities of images. To our best knowledge, the new dataset is the first collection of images that are associated with both aesthetic and emotional labels. Besides, we present a novel Aesthetics-Emotion hybrid Network (AENet) for multi-task learning on aesthetics assessment and emotion recognition. Task-specific and shared features have been explicitly separated by different network streams, and effectively fused at multiple network levels. Experiments on our new and benchmark datasets verify the effectiveness of our approach for unified aesthetics and emotion prediction.

Index Terms— Aesthetics assessment, emotion recognition, multi-task learning

1. INTRODUCTION

Inferring the high-level semantics of an image has reached many outstanding milestones with the latest achievements in computer vision and multimedia communities [1, 2, 3]. Recently, researchers have drawn ideas from the aforementioned to address yet more challenging problems such as understanding users' psychological perceptions of visual content. Typically, aesthetics assessment [4] and emotion recognition [5] are two fundamental problems in user perception understanding, which aim to predict human aesthetic and emotional reactions evoked by visual stimuli, respectively. The potential applications include image retrieval [6], album curation [7], and photo enhancement [8].

In recent years, there have emerged some large-scale image datasets with peer-rated aesthetic or emotional labels, which facilitate the development of learning-based methods [9, 10]. Generally, image aesthetics assessment is cast as a classification or regression problem to distinguish high-aesthetic images from low-aesthetic ones, while image emotion recognition is formulated to classify images into the predefined emotional categories. The key challenge is to extract discriminative visual features. Recent research advances [11, 12, 13] stem from elaborately designing handcrafted features, and evolve into automatically learning deep representations for visual aesthetics or emotion.

Despite the remarkable progress in existing studies, image aesthetics assessment and emotion recognition are usually considered as two separate and independent tasks. Intuitively, aesthetic and emotional perceptions are correlated and interact with each other at the human cognitive level. For example, if an image could provide a feeling of pleasure in aesthetics, it is much likely to arouse positive emotions for viewers. In neuroscience, it has also proven that an aesthetic experience is a cognitive process accompanied by continuously upgrading affective states, resulting in an emotion, and vice versa [14]. Therefore, we argue that aesthetics assessment and emotion recognition needs to be coupled and solved as a whole.

Motivated by this, in this paper, we resort to multi-task learning [15] to deal with image aesthetics assessment and emotion recognition in a unified framework. Nevertheless, this idea faces two major challenges:

- The use of multi-task learning requires images to be associated with both aesthetic and emotional labels, which is hardly satisfied in existing datasets.
- The architecture of most multi-task learning frameworks is inflexible, in which the private and shared information of different tasks are determined merely based on the sharing or separation of some parameters.

To address the above issues, we collect a large scale set of images that are associated with both aesthetic and emotional

⋆ Corresponding author.

This work was supported by National Natural Science Foundation of China (Grant 61701281, Grant 61573219, and Grant 61876098), Shandong Provincial Natural Science Foundation (Grant ZR2017QF009), and the Fostering Project of Dominant Discipline and Talent Team of Shandong Province Higher Education Institutions.

978-1-5386-6249-6/19/$31.00 ©2019 IEEE

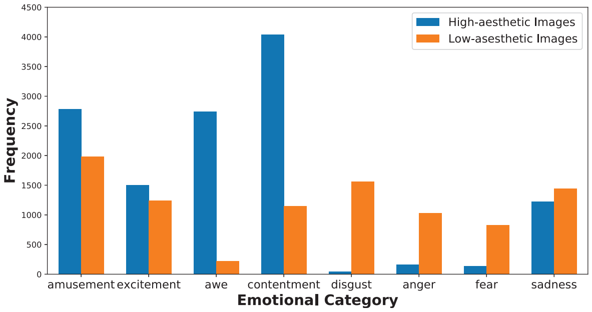

Fig. 1. The number of high-aesthetic and low-aesthetic images grouped on eight emotional categories. The first four emotions (i.e., amusement, excitement, awe, and contentment) are positive, while the last four (i.e., disgust, anger, fear, and sadness) are negative.

labels. It could laid the foundation for us to develop the collaborative modeling of aesthetic and emotional perceptions. This new dataset will be released to the research community[1]. Moreover, we present a novel deep neural network architecture for multi-task learning on aesthetics assessment and emotion recognition. The network explicitly separates the features specific to each single task and the features shared between tasks. A fusion layer is introduced at multiple network levels to effectively combine the task-specific and shared knowledge. Experiments on our new and benchmark datasets verify the effectiveness of our approach for unified aesthetics and emotion prediction in images.

2. DATASET CONSTRUCTION

In this section, we introduce a large scale dataset to facilitate multi-task learning for unified image aesthetics and emotion prediction. We refer to this dataset as the “Images with Aesthetics and Emotions”, or **IAE** for short. To our best knowledge, IAE is the first collection of images associated with both aesthetic and emotional labels. Specifically, IAE is an extension of the earlier work in [10], where 22,086 images are manually divided into eight emotion categories, i.e., amusement, anger, awe, contentment, disgust, excitement, fear, and sadness. Each category consists of more than 1,100 images.

We further rate the quality of these images from the aesthetic perspective. To ensure the quality and integrity of the rating process, ten volunteers were invited to rate each image with one of the four quality levels: Excellent (score 10), Good (score 7) , Fair (score 4) and Bad (score 1). The aesthetic quality of each image is measured by the average of the scores from individual raters, and thus takes values on the scale of 1 to 10. Similar to previous rating datasets [9], we find that the average scores are well fit by a Gaussian distribution. Then, images with average scores smaller than 5 were labeled as low quality, and the others were labeled as high quality. Finally, we identified 12,641 high-aesthetic and 9,445 low-aesthetic images, respectively.

Fig. 1 displays the number of high-aesthetic and low-aesthetic images grouped on each emotional category. As can be seen, if images arouse positive emotions, they are more likely to have high-aesthetic quality; otherwise, they are more likely to be low-aesthetic ones. This phenomenon provides empirical support for our claim that aesthetic and emotional perceptions are correlated and interact with each other.

3. METHOD

Recently, neural-based models for multi-task learning have become popular, since they provide a convenient way of combining information from multiple tasks [15]. Following this idea, we present a novel Aesthetics-Emotion hybrid Network (AENet) for unified aesthetics and emotion prediction.

3.1. Overall Architecture

Previous works on neural-based multi-task learning determine the private and shared features of different tasks merely based on the sharing or separation of certain network layers. As pointed out in [16], such a strategy may lead to the mutual interference between the private and shared features. In this paper, we attempt to explicitly separate the features specific to each single task and the features shared between tasks. Fig. 2 displays the overall architecture of the proposed AENet. AENet is composed of three streams, which are all designed based on the well-known ResNet50 model [1]:

- **Aesthetic stream** (***$\mathcal{A}$-stream***) extracts the information that is unique to the task of aesthetics assessment. We exploited a ResNet50 model pre-trained on the AVA aesthetics dataset [9], and transferred the model weights to this stream as initialization. This stream get the aesthetic task-specific features from input image.
- **Emotional stream** (***$\mathcal{E}$-stream***) extracts the emotion-related information for emotion recognition. In a similar way, we pre-trained the stream on the FI emotion dataset [10].
- **Shared stream** (***$\mathcal{S}$-stream***) extracts the features that are shared across aesthetics assessment and emotion recognition. Specially, it was initialized with the pre-trained weights on ImageNet [17]. The features output from this stream can be used for the two related high-level tasks.

Note that ResNet50 partitions network layers into multiple blocks, each of which contains similar operations of convolution and pooling. At the end of each block, we combine *$\mathcal{S}$-stream* with *$\mathcal{A}$-stream* and *$\mathcal{E}$-stream* through a fusion layer (see Section 3.2). In this way, the task-specific and shared features are jointly leveraged to improve the performance of

[1] https://github.com/junfish/IEA-dataset

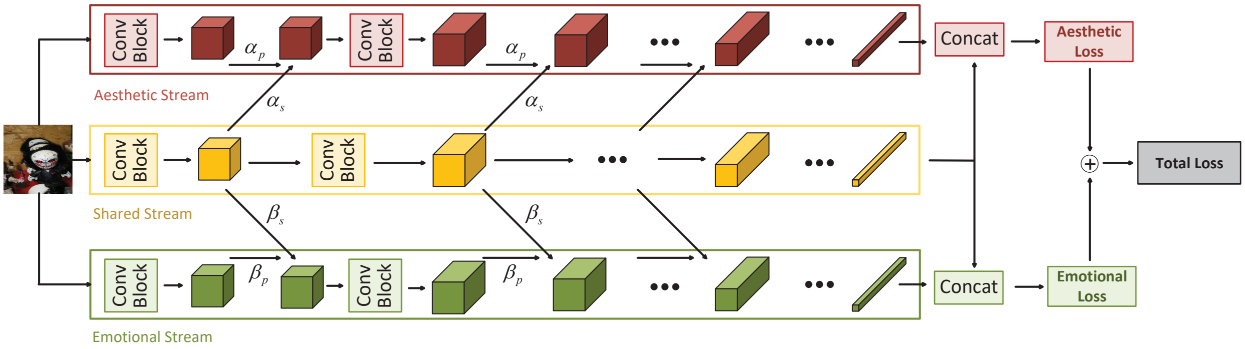

Fig. 2. Overview of the proposed AENet. The cuboids represent the activation maps, and the rectangles represent the operations such as convolution and concatenation.

each single task. Besides, we concatenate the final output vectors of $\mathcal{S}$-*stream* and $\mathcal{A}$-*stream* for aesthetics assessment, and those of $\mathcal{S}$-*stream* and $\mathcal{E}$-*stream* for emotion recognition.

Since both aesthetics assessment and emotion recognition are formulated as classification problems, we choose the binary cross-entropy loss L_a for the former and the softmax cross-entropy loss L_e for the latter. The total loss is computed as the weighted sum of the individual losses at each task, i.e.,

$$L = L_a + \lambda L_e \ , \tag{1}$$

where λ is a trade-off hyper-parameter balancing the two terms.

3.2. Fusion Layer

Fusion layers are designed to effectively combine the knowledge of $\mathcal{S}$-*stream* and that of $\mathcal{A}$-*stream* and $\mathcal{E}$-*stream*. Inspired by the work in [18], we realize the fusion by a linear combination of the activation maps of the three streams. Given the activation maps f_s, f_a, and f_e from $\mathcal{S}$-*stream*, $\mathcal{A}$-*stream*, and $\mathcal{E}$-*stream*, we learn the linear combinations $\widetilde{f}_a$ and $\widetilde{f}_e$, and feed them as input to the next layer of $\mathcal{A}$-*stream* and $\mathcal{E}$-*stream*, respectively. Specifically, at the location (x, y) in the activation maps, $\widetilde{f}_a$ and $\widetilde{f}_e$ are defined as

$$\begin{bmatrix} \widetilde{f}_a(x,y) \\ f_s(x,y) \\ \widetilde{f}_e(x,y) \end{bmatrix} = \begin{bmatrix} \alpha_p & \alpha_s & 0 \\ 0 & 1 & 0 \\ 0 & \beta_s & \beta_p \end{bmatrix} \begin{bmatrix} f_a(x,y) \\ f_s(x,y) \\ f_e(x,y) \end{bmatrix} . \tag{2}$$

Here, the scaling parameters α_s and α_p control the relative importance of the shared and task-specific features for aesthetics assessment, while β_s and β_p control that for emotion recognition.

Since the fusion layer is modeled as a linear combination by the above scaling parameters, their partial derivatives for the loss L can be easily computed as

$$\begin{aligned} \begin{bmatrix} \frac{\partial L}{\partial \alpha_p} & \frac{\partial L}{\partial \alpha_s} \end{bmatrix}^\top &= \frac{\partial L}{\partial \widetilde{f}_a(x,y)} \begin{bmatrix} f_a(x,y) & f_s(x,y) \end{bmatrix}^\top , \\ \begin{bmatrix} \frac{\partial L}{\partial \beta_p} & \frac{\partial L}{\partial \beta_s} \end{bmatrix}^\top &= \frac{\partial L}{\partial \widetilde{f}_e(x,y)} \begin{bmatrix} f_e(x,y) & f_s(x,y) \end{bmatrix}^\top . \end{aligned} \tag{3}$$

Eq. (2) and Eq. (3) formalize the forward-propagation and back-propagation calculations through a fusion layer, respectively. As can be seen, the end-to-end learning is unimpeded in this layer. The optimal values of the scaling parameters will be automatically determined in learning.

3.3. Implementation Details

The network is trained with the mini-batch stochastic gradient descent algorithm. We set the batch size to 64. The parameters in the fusion layers were initially set to $\alpha_p = \beta_p = 0.9$ and $\alpha_s = \beta_s = 0.1$, which are about two or three orders of magnitude larger than the typical values of the other layer parameters that were initialized using the Xavier method [19]. Therefore, we need to use higher learning rates for the fusion layers. In practice, we set the learning rate of the scaling parameters to 10^2, and that of the other parameters to 10^{-4}. This leads to faster convergence and the best performance. Besides, since aesthetics assessment is formulated as binary classification and emotion recognition as multi-label classification, for the best and balanced performance between two related tasks, we set the trade-off hyper-parameter to balance the variation of two losses by setting $\lambda = 1/4$.

4. EXPERIMENTS

4.1. Experimental Settings

4.1.1. *Datasets*

We evaluated our approach on the new IAE dataset for aesthetics assessment and emotion recognition, respectively. We

Methods		*Aes Acc* (%)	*Emo Acc* (%)
Single-task	ResNet50	74.85	61.09
	WRN	75.16	62.65
Multi-task	SSNet	77.79	60.23
	CSNet	77.70	63.17
	AENet-FL	76.68	61.46
	AENet (Ours)	**81.05**	**66.23**

Table 1. Classification accuracy on IAE for aesthetics assessment and emotion recognition, respectively.

randomly picked out 70% of images for training, 10% for validation, and the remaining for testing. In order to verify the generalization ability, we also tested on an aesthetics and an emotion benchmark dataset separately, i.e., AVA [9] and ArtPhoto [20]. Note that AVA and ArtPhoto were only used as test sets.

4.1.2. Baselines

Single-task Baselines: We first compared our approach against single-task learning models. The first competitor is **ResNet50**, which serves as the base network of our proposed AENet. It should be noted that AENet has more parameters (nearly three times) than ResNet50. To clarify whether performance improvement is only due to more number of parameters, we adopted the **Wide ResNet (WRN)** [21] as another baseline. WRN has a widened architecture of ResNet blocks, and was configured to have an approximate number of parameters with AENet in our case.

Multi-task Baselines: We also introduced three multi-task learning models as baselines. One is the traditional **Share-Split Network (SSNet)** [22], where all convolutional layers are shared and the split takes place after the last convolutional layer for task-specific losses. Another one is the **Cross-Stitch Network (CSNet)** [18], in which the individual streams of each task are directly fused via element-wise linear combinations of their activation maps. Besides, we implemented a variant method of AENet without the fusion layers, which is denoted by **AENet-FL**.

4.2. Experimental Results

In our experiments, classification accuracy is adopted as the evaluation metric. Table 1 summarizes the results of different methods on IAE. Compared to the single-task baselines, our AENet exhibits at least 5.9% and 3.6% performance improvement for aesthetics assessment and emotion recognition, respectively. In particular, the superiority of AENet over WRN suggests that the improvement is derived from the full use of shared knowledge between tasks, rather than the simple increase of the number of model parameters. For the multi-task baselines, we find that it is difficult for the traditional SSNet to enhance both tasks simultaneously. This may be attributed to the less flexibility of its parameter sharing mechanism. In addition, AENet substantially outperforms CSNet. A possible reason is that AENet explicitly separates the features specific to each task as well as the features shared between tasks, which is overlooked by CSNet. We also notice that AENet achieves higher performance than AENet-FL, which indicates the importance of the fusion layers in combining the task-specific and shared knowledge at multiple network levels.

Datasets		*Aes Acc* (%)	*Emo Acc* (%)
Methods		AVA	ArtPhoto
Single-task	ResNet50	71.18	24.83
	WRN	71.33	27.02
Multi-task	SSNet	68.39	23.86
	CSNet	70.48	27.30
	AENet-FL	70.29	24.22
	AENet (Ours)	**72.83**	**27.92**

Table 2. Classification accuracy of different methods on AVA and ArtPhoto, respectively.

As aforementioned, we performed a cross-set evaluation, where all models were trained only on IAE, but tested on AVA and ArtPhoto, respectively. Table 2 shows the comparison results. Compared to the results listed in Table 1, all methods experience a sharp degradation in performance, especially on ArtPhoto. We believe this is because there exists a significant difference in data distribution between training and testing sets. As expected, AENet still outperforms all its counterparts. For example, AENet enjoys about 1.7% improvement over the runner-up method on AVA. The observation highlights the better generalization ability of AENet.

5. CONCLUSION

In this paper, we have addressed the problem of unified image aesthetics and emotion prediction by introducing an end-to-end multi-task deep learning framework, i.e., AENet. Task-specific and shared features have been separately extracted by different network streams, and effectively fused at multiple network levels. Experimental results have verified the promise of our approach for aesthetics assessment and emotion recognition, respectively. We believe that the AENet architecture can be also used for other related tasks, and the research on the relationships between aesthetics and emotion is thought-provoking. In addition, we have collected the new IAE dataset, which initially associates images with both aesthetic and emotional labels. We hope that it will promote the development of the understanding on high-level visual perceptions in the future.

6. REFERENCES

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[2] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks.," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[3] Chaoran Cui, Jialie Shen, Liqiang Nie, Richang Hong, and Jun Ma, "Augmented collaborative filtering for sparseness reduction in personalized poi recommendation," *ACM Transactions on Intelligent Systems and Technology*, vol. 8, no. 5, pp. 71, 2017.

[4] Yubin Deng, Chen Change Loy, and Xiaoou Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 80–106, 2017.

[5] Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z Wang, Jia Li, and Jiebo Luo, "Aesthetics and emotions in images," *IEEE Signal Processing Magazine*, vol. 28, no. 5, pp. 94–115, 2011.

[6] Chaoran Cui, Huidi Fang, Xiang Deng, Xiushan Nie, Hongshuai Dai, and Yilong Yin, "Distribution-oriented aesthetics assessment for image search," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017, pp. 1013–1016.

[7] Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J Foran, "Personalized image aesthetics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 638–647.

[8] Sicheng Zhao, Xin Zhao, Guiguang Ding, and Kurt Keutzer, "Emotiongan: Unsupervised domain adaptation for learning discrete probability distributions of image emotions," in *Proceedings of the ACM Multimedia Conference*, 2018, pp. 1319–1327.

[9] Naila Murray, Luca Marchesotti, and Florent Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2408–2415.

[10] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016, pp. 308–314.

[11] Baohan Xu, Yanwei Fu, Yu-Gang Jiang, Boyang Li, and Leonid Sigal, "Video emotion recognition with transferred deep feature encodings," in *Proceedings of the ACM International Conference on Multimedia Retrieval*, 2016, pp. 15–22.

[12] Yueying Kao, Ran He, and Kaiqi Huang, "Deep aesthetic quality assessment with semantic information," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1482–1495, 2017.

[13] Chaoran Cui, Huihui Liu, Tao Lian, Liqiang Nie, Lei Zhu, and Yilong Yin, "Distribution-oriented aesthetics assessment with semantic-aware hybrid network," *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1209–1220, 2019.

[14] Helmut Leder, Benno Belke, Andries Oeberst, and Dorothee Augustin, "A model of aesthetic appreciation and aesthetic judgments," *British journal of psychology*, vol. 95, no. 4, pp. 489–508, 2004.

[15] Sebastian Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[16] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang, "Adversarial multi-task learning for text classification," *arXiv preprint arXiv:1704.05742*, 2017.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[18] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert, "Cross-stitch networks for multi-task learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3994–4003.

[19] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[20] Jana Machajdik and Allan Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2010, pp. 83–92.

[21] Sergey Zagoruyko and Nikos Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[22] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.