FILLING THE GAPS: REDUCING THE COMPLEXITY OF NETWORKS FOR MULTI-ATTRIBUTE IMAGE AESTHETIC PREDICTION

Magzhan Kairanbay *John See* *Lai-Kuan Wong* *Yong-Lian Hii*

Faculty of Computing and Informatics, Multimedia University, Malaysia

ABSTRACT

Computational aesthetics have seen much progress in recent years with the increasing popularity of deep learning methods. In this paper, we present two approaches that leverage on the benefits of using Global Average Pooling (GAP) to reduce the complexity of deep convolutional neural networks. The first model fine-tunes a standard CNN with a newly introduced GAP layer. The second approach extracts global and local CNN codes by reducing the dimensionality of convolution layers with individual GAP operations. We also extend these approaches to a multi-attribute network which uses a style network to regularize the aesthetic network. Experiments demonstrate the capability of attaining comparable accuracy results while reducing training complexity substantially.

Index Terms— Aesthetics, Style, Convolutional Neural Network, Global Average Pooling, Multi-attribute Network

1. INTRODUCTION

Automatic aesthetic evaluation of photographs is a challenging problem in the pattern recognition field. The difficulties lie in the fact that aesthetics is a subjective notion, where different people may evaluate aesthetics differently, or have differing ideas about it. The machine evaluation of such a task can be of great help to many. For instance, journalists and designers are able to retrieve or search for aesthetically high photos from large scale repositories. At a more personal level, people often capture a large number of photos without realizing, and they are then faced with a quandary to choose their most beautiful photos to be shared on social networks, or to get their photo gallery sorted out intelligently.

The approach widely undertaken by the research community is to regard this problem as of a binary decision; images are categorized as aesthetically *high* or *low*. Recent state-of-the-art solutions for photo aesthetic evaluation are shifting towards using deep learning techniques [1, 2, 3, 4], which has been shown to produce significant improvement over conventional hand-crafted features [5, 6, 7, 8] and generic descriptors [9]. This was spurred on by the availability of a large benchmark AVA dataset [10] that was made publicly available for this task. In addition to providing aesthetics rating, the AVA dataset also provides other meta-attributes such as style and semantic information, which have been utilized by some researchers [2, 11, 12] to train aesthetics evaluation models. However, these multi-attribute deep-learning approaches are computationally costly due to the large amount of parameters.

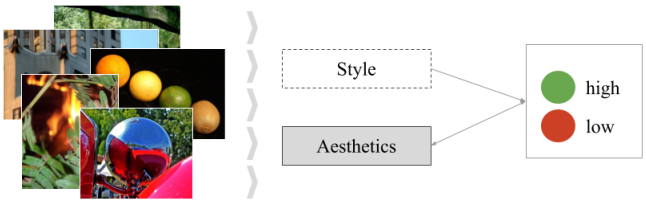

Fig. 1. Multi-attribute architecture.

In this paper, we propose two efficient deep-learning approaches; (1) **AnGAP-Finetuned** (fine-tuned GAP based AlexNet model for photo aesthetic task) and (2) **AnGAP-FeatEns** (AlexNet model with GAP based ensemble of features) that require substantially less parameters for model training, with the aim of improving the computational efficiency of current deep architectures. These two proposed approaches are first used to train the aesthetic classification models using only information from the image content. We then extend both the AnGAP-Finetuned and AnGAP-FeatEns approaches to utilize additional attribute information; the style meta-information provided by the dataset. Fig. 1 illustrates the conceptual overview of the proposed multi-attributes approaches. Our experiments show that our proposed approaches produce comparable accuracy with existing works but with significantly less parameters.

2. RELATED WORK

Among early deep learning approaches in this domain, Lu et al. [12] proposed a single-column (SCNN) and a double-column convolutional neural networks (DCNN). Different types of image transformation like center crop, random crop, warping and padding of input images were considered and tested. One of the inputs of the DCNN is responsible for the global view while the other considers the local view. The authors further regularize their proposed models at the fully connected layer with an additional column containing

pre-trained features from style or semantic attributes.

Jin et al. [13] proposed ILGNet (Inception Local Global Net), a new form of CNN which consists of 13 layers. ILGNet is based on GoogleNet which uses inception and pretreatment modules. The main idea behind this net is to connect local and global features together. The concatenation of the two intermediate inception layers of local features with one inception layer of global features forms a 1024-dimension layer, followed by a fully connected layer. To adapt to this task, the ILGNet is first trained on the ImageNet dataset and then fine-tuned on the AVA dataset. The ensemble of local visual features from earlier inception layers to the global visual features was able to describe the overall photo aesthetics and it gave promising results.

Wang et. al [11] introduced a brain-inspired deep network (BDN) which made use of style information from the AVA dataset. First, 14 fully convolutional neural networks (FCNNs) are trained for each style. Each FCNN consists of 4 convolution layers. The first two convolutional layers of each FCNN are trained on the whole AVA dataset in an unsupervised way using Stacked Convolutional Auto Encoders, whereas the subsequent layers are trained in a supervised manner. Three primitive features (hue, saturation, value) of the input images are also fused with the output of the third convolutional layer of the 14 FCNNs to form an input cube for another FCNN, which predicts the overall aesthetics ratings. While their idea is inspired by neuroscience models, it is computationally heavy.

3. DATASET

The Aesthetic Visual Analysis (AVA) [10] is the state-of-the-art benchmark dataset used for large-scale photo aesthetics analysis. The dataset contains 250K photos, where 230K photos are used for training and the rest are used for testing. On an average, each photo is evaluated by around 210 users; each of whom rated the photo with an aesthetic score between 1 and 10. In our work, we use the mean rating to indicate the aesthetic score for an image. Apart from the aesthetics ratings, the dataset provides style annotation for 14,079 photos, where 11,270 photos are designated for training while the rest for testing. Notably, the photos in the training set have only one style but the photos in the testing set may be assigned to multiple styles. The authors of AVA dataset evaluated their solution on different subsets of the dataset by changing the δ value, which is used to filter out ambiguous photos. The aesthetics score of each photo in the subset is in the range of 1 to $(5-\delta)$ or $(5+\delta)$ to 10.

4. SINGLE COLUMN ARCHITECTURE

Combining multiple single column CNNs can be advantageous as it capacitate the sharing of features that a single column CNN could not extract. In this section, we first introduce two variations of single column networks; (1) AnGAP-Finetuned and (2) AnGAP-FeatEns. AnGAP-Finetuned is adapted from the concept by Zhou et al. [14], which uses the GAP layer as the key technique to reduce the computation time and number of parameters. AnGAP-FeatEns extracts features from an ensemble of GAP layers appended to various convolutional layers of the AlexNet model. These features are then aggregated and passed to a logistic regression step for classification.

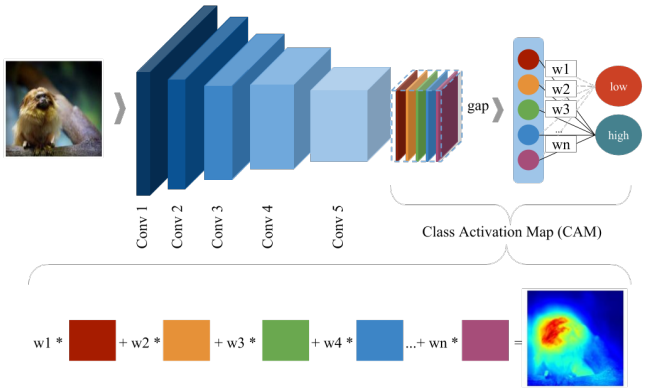

Fig. 2. AnGAP-Finetuned: The usage of GAP layer in AlexNet for aesthetics evaluation task.

4.1. AnGAP-Finetuned: GAP based CNN model

Zhou et al. [14] proposed a network that uses a GAP layer for generating Class Activation Maps (CAM). The idea behind this technique is to remove the fully connected (FC) layers and replace those layers with a additional convolutional layer, followed by a GAP layer. The GAP layer generates the feature vector by calculating the global average of each feature map in the last convolutional layer. This architecture saves a huge amount of parameters and computation time.

Inspired by [14], we proposed AnGAP-Finetuned, which employs the usage of the GAP layer in AlexNet for the aesthetics evaluation task. The architecture of AnGAP-Finetuned is illustrated in Fig. 2. The usage of the GAP layer is useful for producing the CAMs, which can single out areas that most highly activating a particular aesthetic class.

4.2. AnGAP-FeatEns: GAP based ensemble of CNN features

Fig. 3 illustrates the strategy employed in the proposed AnGAP-FeatEns approach: Features are first extracted from an ensemble of GAP layers appended to each convolutional layer of the AlexNet model; the aggregated features are used for aesthetic prediction using logistic regression. The ex-

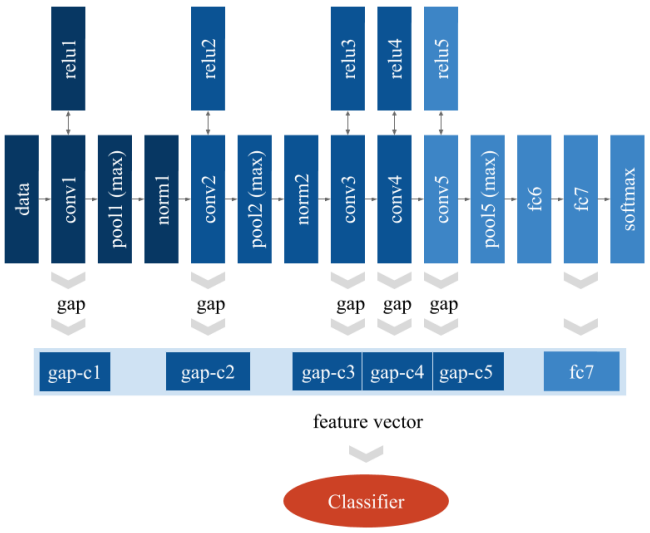

Fig. 3. AnGAP-FeatEns: Ensemble of GAP based features

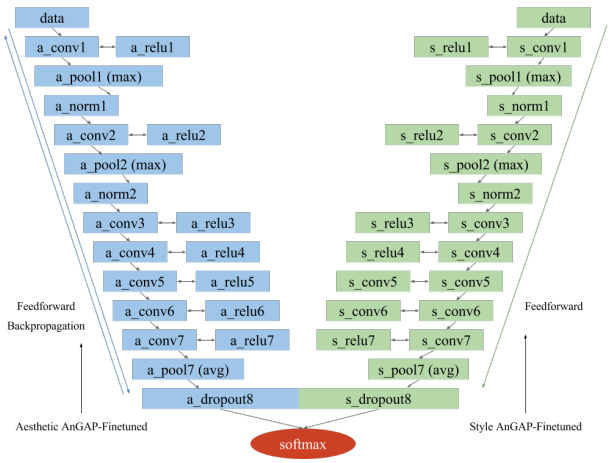

Fig. 4. The architecture of Multi-Attribute AnGAP-Finetuned.

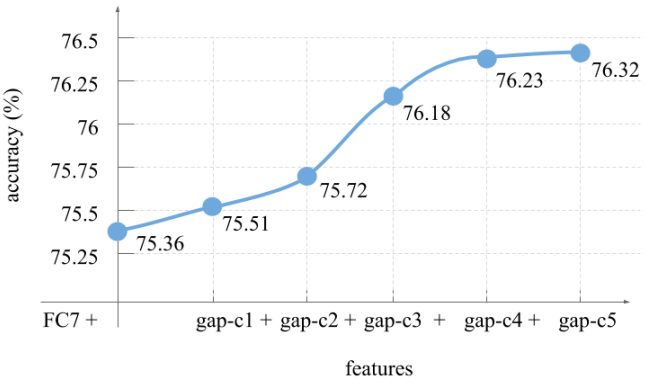

Fig. 5. Accuracy of Multi-Attribute AnGAP-FeatEns approach based on different combination of GAP feature layers

tracted GAP based features contain both low-level visual features from the earlier convolutional layers, and high-level visual features from the later layers. Since photo aesthetics is influenced directly by both local and global information [12, 15], this set of extracted features can provide a good representation of image aesthetics. Notably, since the number of feature maps in each convolutional layer is not high, the size of the aggregated feature vector is of reasonable size.

5. MULTI-ATTRIBUTE NET

Next, we design two multi-attribute networks by extending both the AnGAP-Finetuned and AnGAP-FeatEns approaches to include the style attribute. Fig. 4 illustrates the architecture of the Multi-Attribute AnGAP-Finetuned model. This multi-attribute network combines the single column aesthetics CNN and style CNN, in which the style CNN is used to regularize the main aesthetics CNN, in similar fashion as Lu et al. [12]. However, in that work, their CNN is not as deep while the usage of the fully connected layers is computationally inefficient and memory consuming. Comparatively, our network is deeper and more efficient in terms of time and memory. Both the aesthetics and style CNNs are identical but are fine-tuned for different tasks. To combine both CNNs, we remove the last softmax layer for each CNN and join the aesthetic and style CNNs by concatenating the penultimate layer of each CNN. A softmax layer follows this combined layer. We freeze the learning rates for style CNN so that the back-propagation process involves only the aesthetics CNN.

For the Multi-Attribute AnGAP-FeatEns, we concatenate the extracted GAP-based feature vectors for both the aesthetics and style CNNs into a super vector before performing classification using logistic regression. To examine the contribution of each GAP feature, we tested our approach by gradually combining the features from the FC7-GAP layers of both CNNs with the other GAP features from the convolutional layers. First, we concatenate only the GAP-features from the FC7 layers of both CNNs. Following that, we further add on the GAP features of the next convolutional layer for both the aesthetics and style tasks, until the GAP features from all convolutional layers (conv1 to conv5) have been included. Fig. 5 shows the accuracy of each test case. We can observe that the prediction accuracy is directly proportional to the the number of GAP layers used. The maximum accuracy (76.32%) is achieved when the GAP features of all convolutional layers are concatenated.

6. EXPERIMENTAL RESULTS AND DISCUSSIONS

For our experiments, we empirically tested all proposed architectures on a subset of the AVA dataset, with δ={2, 1.5, 1, 0.5, 0}. From our observation, the accuracy reaches the highest point for all tested models when $\delta = 0.5$ (approximately

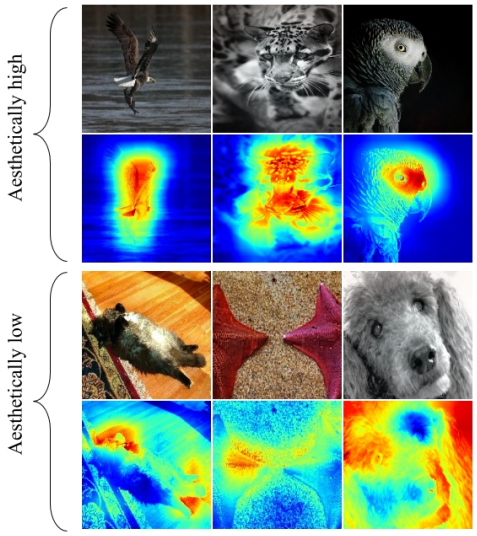

Fig. 6. Class activation maps for samples of aesthetically high and low photographs.

127K photos). This is likely be due to the fact that half of the AVA dataset contains ambiguous photos which are located in the middle of the rating range (4.5–5.5).

6.1. Overall Results

Table 1 reports the accuracy rate of different deep architectures evaluated and reported in the literature. In comparison to AlexNet, AnGAP-Finetuned uses less model parameters and is thus more computationally efficient. The accuracy of AnGAP-Finetuned is only marginally lower ($\sim 0.2\%$) than the conventionally fine-tuned AlexNet. As many other existing works do not provide model parameter counts, it is difficult to provide an accurate assessment of their models.

Comparing AnGAP-FeatEns to AnGAP-Finetuned, the AnGAP-FeatEns approach has a much larger number of parameters than AnGAP-Finetuned but it is faster in terms of the feature extraction and training speed. One of the drawbacks of AnGAP-FeatEns is that the concatenation of many GAP features require a substantial amount of memory. However, we see an improvement in performance ($\sim 1\%$) of the AnGAP-FeatEns over then AnGAP-Finetuned. Notably, the accuracy of the two multi-attribute networks that introduced an additional style CNN to the single column CNN architecture, only increases the performance slightly by $\sim 0.3\%$, when compared to their respective counterparts.

6.2. Class Activation Maps

Interestingly, the usage of GAP capacitates the extraction of class activation maps (CAM) [14], which can be used to visualize the spatial locations that contribute towards high or low aesthetics in a photograph. Fig. 6 offers some insights into images from both aesthetically high and low classes, where the hot (redder) areas indicate locations that give a sense of beauty or ugliness in photos of the respective classes. From the CAMs of aesthetically high photos, it can be observed that the hot area in each photo coincides with the main subject (parrot, eagle and tiger). In contrast, photos with low aesthetic scores show activations at locations that are uninteresting. This corresponds with how professional photographers utilize various photography rules such as simple background and low depth-of-field, to create subject dominance in their photographs. Thus, the CAMs illustrate the effectiveness of the proposed GAP-based architectures in learning the elements which influence photographic aesthetics.

Table 1. Performance comparison of the proposed schemes against various methods in literature

Method	Acc. (%)	# params
AVA Baseline [10]	68.00	-
SPP [16]	72.85	-
DCNN [12]	73.25	-
RDCNN-style [12]	74.46	-
Peng et. al [17]	74.50	-
Kao et. al [18]	74.51	-
AnGAP-Finetuned	**74.84**	**$\sim$ 4K**
AlexNet-Finetuned	**75.13**	**$\sim$ 56K**
Multi-Att AnGAP-Finetuned	**75.16**	**$\sim$ 8K**
DMA-Net [16]	75.41	-
RDCNN semantic [15]	75.42	-
AnGAP-FeatEns	**76.07**	**$\sim$ 56K**
Kao et. al [19]	76.15	-
Multi-Att AnGAP-FeatEns	**76.32**	**$\sim$ 112K**
BDN [11]	76.80	

7. CONCLUSION

In this paper, we present two approaches that leverage on the benefits of using GAP to reduce the complexity of deep convolutional neural networks. The first model fine-tunes a standard CNN with the GAP layer. The second approach extracts an ensemble of CNN features from various layers that have been reduced by individual GAP operations. We also extend these approaches to a multi-attribute network which leverages a style network to regularize the aesthetic network. Experiments show the capability of attaining comparable accuracy results while reducing training and testing complexity substantially. Such light models could be widely deployed into portable devices for various useful applications. We also show the capability of the GAP layer in producing CAMs for visualizing locations in a photo that contribute towards its aesthetic quality. For future work, we intend to explore multi-attribute networks that incorporate semantic or text information.

8. REFERENCES

[1] Weining Wang, Mingquan Zhao, Li Wang, Jiexiong Huang, Chengjia Cai, and Xiangmin Xu, "A multi-scene deep learning model for image aesthetic evaluation," *Signal Processing: Image Communication*, vol. 47, pp. 511–518, 2016.

[2] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charles Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *European Conference on Computer Vision*. Springer, 2016, pp. 662–679.

[3] Xinmei Tian, Zhe Dong, Kuiyuan Yang, and Tao Mei, "Query-dependent aesthetic model with deep learning for photo quality assessment," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2035–2048, 2015.

[4] Yubin Deng, Chen Change Loy, and Xiaoou Tang, "Image aesthetic assessment: An experimental survey," *arXiv preprint arXiv:1610.00838*, 2016.

[5] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang, "Studying aesthetics in photographic images using a computational approach," in *European Conference on Computer Vision*. Springer, 2006, pp. 288–301.

[6] Yan Ke, Xiaoou Tang, and Feng Jing, "The design of high-level features for photo quality assessment," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 1, pp. 419–426.

[7] Yiwen Luo and Xiaoou Tang, "Photo and video quality evaluation: Focusing on the subject," in *European Conference on Computer Vision*. Springer, 2008, pp. 386–399.

[8] Kuo-Yen Lo, Keng-Hao Liu, and Chu-Song Chen, "Assessment of photo aesthetics with efficiency," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 2186–2189.

[9] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1784–1791.

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

[11] Zhangyang Wang, Shiyu Chang, Florin Dolcos, Diane Beck, Ding Liu, and Thomas S Huang, "Brain-inspired deep networks for image aesthetics assessment," *arXiv preprint arXiv:1601.04155*, 2016.

[12] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang, "Rapid: Rating pictorial aesthetics using deep learning," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 457–466.

[13] Xin Jin, Jingying Chi, Siwei Peng, Yulu Tian, Chaochen Ye, and Xiaodong Li, "Deep image aesthetics classification using inception modules and fine-tuning connected layer," in *Wireless Communications & Signal Processing (WCSP), 2016 8th International Conference on*. IEEE, 2016, pp. 1–6.

[14] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

[15] 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.

[16] Xin Lu, Zhe Lin, Xiaohui Shen, Radomir Mech, and James Z Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 990–998.

[17] Kuan-Chuan Peng and Tsuhan Chen, "Toward correlating and solving abstract tasks using convolutional neural networks," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–9.

[18] Yueying Kao, Kaiqi Huang, and Steve Maybank, "Hierarchical aesthetic quality assessment using deep convolutional neural networks," *Signal Processing: Image Communication*, vol. 47, pp. 500–510, 2016.

[19] Yueying Kao, Ran He, and Kaiqi Huang, "Visual aesthetic quality assessment with multi-task deep learning," *arXiv preprint arXiv:1604.04970*, 2016.