MULTIGAP: MULTI-POOLED INCEPTION NETWORK WITH TEXT AUGMENTATION FOR AESTHETIC PREDICTION OF PHOTOGRAPHS

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

Faculty of Computing and Informatics, Multimedia University, Malaysia

ABSTRACT

With the advent of deep learning, convolutional neural networks have solved many imaging problems to a large extent. However, it remains to be seen if the image "bottleneck" can be unplugged by harnessing complementary sources of data. In this paper, we present a new approach to image aesthetic evaluation that learns both visual and textual features simultaneously. Our network extracts visual features by appending global average pooling blocks on multiple inception modules (MultiGAP), while textual features from associated user comments are learned from a recurrent neural network. Experimental results show that the proposed method is capable of achieving state-of-the-art performance on the AVA / AVA-Comments datasets. We also demonstrate the capability of our approach in visualizing aesthetic activations.

***Index Terms*—** Image aesthetics evaluation, deep neural network, CNN, textual features, aesthetic visualization

1. INTRODUCTION

In the past decade, digital photography has taken the world by storm with the ease of storing hundreds or even thousands of images in a chip that is barely larger than a fingernail. Combine that with the ubiquity of smartphones packing a capable camera, the accessibility of social networking sites, cheap cloud storage, and we are collectively taking more photos than ever. Sifting through the huge volumes of images everyday is enormously time-consuming , so the ability to automatically discern aesthetically pleasing images would therefore offer great applicability in such situations.

Recently, there has been significant interest in the area of *image aesthetic evaluation* [1, 2]. Many existing works have mostly focused on handcrafted methods, which rely on various photographic rules [3] and low-level features [4] as well as feature representations such as SIFT or color descriptors [1]. With the advent of deep neural network models, most of the attention in this field has been shifted towards the use of convolutional neural networks (CNN). Recent works [2, 5] have proved that they are effective at extracting features and are able to surpass handcrafted methods by a considerable margin of accuracy.

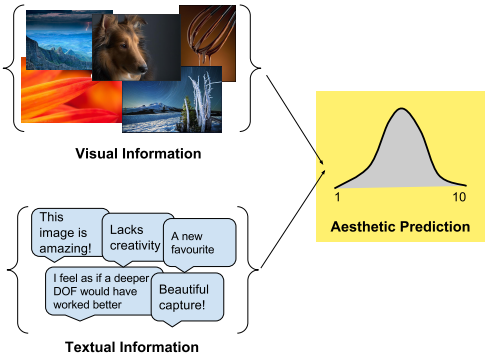


Fig. 1. Visual and textual information for aesthetic prediction

In this paper, we propose a new approach to image aesthetic evaluation. To summarize, our three main contributions are: 1) a novel deep neural network architecture called MultiGAP that exploits features from multiple inception modules pooled by global average pooling (GAP), evaluated by prediction of (10-class) categorical distribution, before performing score binarization; 2) the incorporation of textual features trained simultaneously with MultiGAP using a recurrent neural network (RNN) with gated recurrent unit (GRU) layers; 3) the ability to leverage the GAP block for visualization of activation maps based on the aesthetic category and rating. On top of that, an extensive performance comparison between competing methods in literature is also reported for the benefit of this growing field of research.

2. RELATED WORK

In this section, we will provide a brief categorization and critique of selected state-of-the-art approaches.

Works in image aesthetic evaluation can generally be divided into three distinct categories – classical handcrafted low-level features, generic features based on image descriptors, and the contemporary approach of utilizing deep learning models. In the first category, the aesthetics of images are largely determined by low level descriptors such as color, hue, saturation, light exposure, and also other heuristics driven by rule of thumbs used by professional photographers

[4, 3, 6] For the category which utilizes generic descriptors [7, 1], popular feature extraction methods such as SIFT [8] and color descriptors are adopted by combining them with a feature encoding method such as a Fisher Vector [9].

Deep learning models, particularly Convolutional Neural Networks (CNN) [10] have made incredible strides into various classification tasks in domains such as speech, text, images and videos, largely due to their capability in learning discriminative representations from raw signals and performing end-to-end classification. Lately, CNNs have found its way into aesthetics prediction. Lu et al. [2] proposed a double-column deep neural network which learns from global and local views; their subsequent work [5] continued along the same theme using multiple patches instead. Kao et al. [11] trained a regression model using CNN to estimate a continuous score. Hierarchical or parallel approaches using multiple CNNs [12, 13] have also been introduced with promise, but they are blighted by high computational overhead. More recently, the work of Zhou et al. [14] presented a multimodal network that jointly learns image and text representations, reporting state-of-the-art performance on the AVA dataset [1].

In the past, image aesthetics has mostly been directly formulated as a binary classification problem, where the task simply determines if an image is "good" or "bad", or as a regression problem to predict a score. These methods produced results that were relatively unintuitive and hard to interpret, discarding useful information regarding the rater's consensus along the way. Some recent works have started dealing with this issue by performing prediction of aesthetic rating distribution [13, 15]. However, these works attempt to match distributions using the mean and standard deviation which relies heavily on the sufficiency of sample scores. Hence, we propose a hybrid approach which uses the Kullback-Liebler loss to predict a 10-way categorical distribution, where the mean rating is binarized into the two standard aesthetic labels.

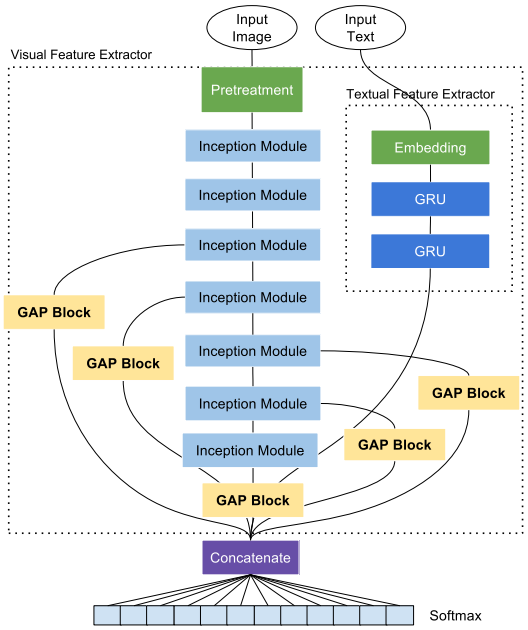


Fig. 2. The proposed RNN-augmented MultiGAP network

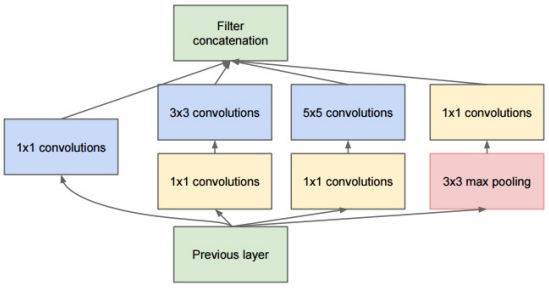


Fig. 3. Inception module [16]

3. METHODS

3.1. Visual Features

To learn visual features, we use GoogLeNet, a deep neural network architecture codenamed Inception proposed by Google [16] as our base network. Briefly, the GoogLeNet is a network consisting of *inception modules* stacked upon each other, with occasional max-pooling layers with stride 2 added to halve the grid resolution. A single inception module consists of a number of parallel filters to densely capture spatial correlations. Dimensionality reduction is incorporated through the use of 1×1 convolutions before the expensive 3×3 and 5×5 convolutions (see Fig. 3).

To construct our proposed network called MultiGAP, the final three inception modules are removed and a new "GAP block" is attached to each inception module. Each GAP block consists of an additional convolution layer followed by a global average pooling (GAP) layer, which reduces the feature maps to a 1-D vector of weights (see Fig. 4). GAP layers are useful replacements for fully connected layers, and their ability to localize strong class activations have been shown in literature [17, 18]. By adding GAP layers to the inception modules, this intuitively regularizes the "micro network" within each module, reducing the proneness to overfitting at different levels of abstraction.

We then replaced the original 1000-class softmax (of ImageNet Challenge) with a 10-class softmax, which corresponds to each of the 10 rating levels of the AVA dataset. All images are warped to fit the network receptive field of 224×224 pixels in the RGB color space with zero mean. We fine-tune the model with vanilla Stochastic Gradient Descent (SGD) over the entire training set ($\delta = 0$)[1]. Learning rate starts at 0.001, dividing by 10 every time the validation error plateaued.

[1]The parameter δ was introduced in the AVA paper[1] to discard ambiguous images with average score within $[5 - \delta, 5 + \delta]$ from the training set.

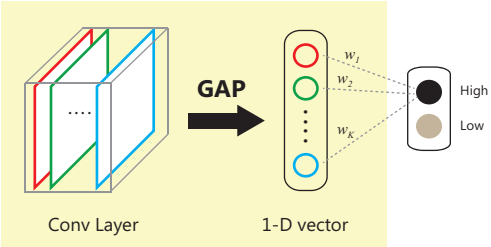


Fig. 4. GAP Block

3.2. Textual Features

For textual information, we used the AVA Comments dataset provided in [14]. The words from the comments are mapped to vector space using Global Vectors or GloVe [19] representation, which trains the word representation based on aggregated global word-word co-occurrence statistics.

Firstly, all user comments for each image are concatenated and tokenized into a sequence of word indices. The top 20,000 most commonly occurring words are considered. The length of words in all the comments on an image combined is fixed to 100 for reasons of practicality. The weights of a 300-D GloVe embedding pre-trained on the Wikipedia corpus is used to construct the word embedding vectors.

Finally, a simple 2-layer Recurrent Neural Network (RNN) model is trained using the embedding vectors as input. The constituents of the model include two layers of Gated Recurrent Units (GRU) [20]. The output of the network is then concatenated with the rest of the GAP blocks from the visual network (MultiGAP) to jointly train and predict the aesthetic categorical distribution of the input image.

3.3. Categorical Distribution Prediction

Most existing works follow the lead of key works in literature [1, 4] by directly representing the aesthetic quality of each image by a single scalar value (typically, the mean rating). These values are further assigned to binary labels, i.e. high quality and low quality. This over-simplifies the rather subjective nature of the aesthetic problem as the learning of features are supervised without the knowledge of its distribution.

To utilize more information on the distribution of ratings, each image is represented by a categorical distribution over 10 categories, which corresponds to each of the 10 ratings. We obtain the ground truth distribution by normalizing the individual ratings of each image by its sum of ratings. Aptly, we also use the Kullback-Liebler (KL) divergence loss as our choice of measuring the information gain from a predicted distribution, p_i to the actual distribution q_i:

$$KL(p_i, q_i) = \sum_i p_i \cdot log(\frac{p_i}{q_i}) \quad (1)$$

To evaluate the performance of our models based on the standard binary classification task, the expected value of the predicted categorical distribution $E[X] = \sum_i x_i p_i$ is taken as the mean rating of the test image. Following standard AVA protocol [1], the mean rating is binarized to 1 (high quality) if it is above or equal to 5, and 0 (low quality) otherwise.

4. EXPERIMENTAL RESULTS

4.1. Datasets

For evaluation, we used the large-scale AVA image aesthetic dataset [1], which contains more than 250,000 community-rated images, each has an average of 200 ratings (between 1 and 10) collected from photography community. In the original protocol, a total of 20,000 images were earmarked for testing while the remaining images were for training. Aesthetics assessment is formulated as a binary classification problem, where images with a mean rating higher than 5 are labeled as high-quality images, and vice versa.

Recently, Zhou et al. [14] crawled the user comments from DPChallenge for all images in the AVA dataset to form the AVA-Comments dataset. This supporting dataset contains more than 1.5 million user comments which were tokenized, with all quotes and extra HTML tags such as links removed. In our work, we learn textual features from the AVA-Comments data to complement the visual features trained from the AVA dataset.

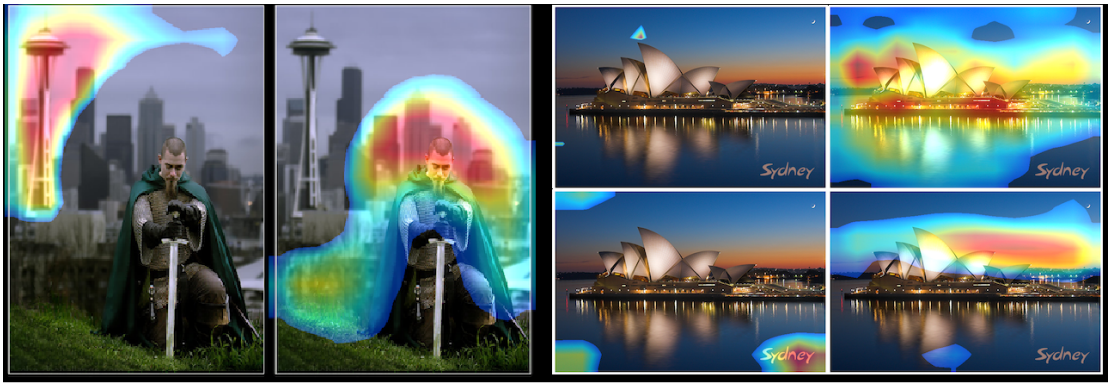
4.2. Aesthetic Prediction

Table 1 shows the results of our proposed methods benchmarked against other methods in literature, organized by type of features used. "SingleGAP" denotes the trimmed version of MultiGAP where the GAP block is applied only to the last inception module. Surprisingly, the SingleGAP network perform marginally better than MultiGAP on the visual features.

On the textual features, the use of Recurrent Neural Network (RNN) provides an improved performance over prior methods evaluated by [14]. Crucially, unlike the method in [14], there is no deterioration of performance when the joint network is trained jointly using both image and textual information. The proposed MultiGAP architecture, augmented by a RNN that captures the textual dependencies achieved a state-of-the-art accuracy of 82.27%. Fig. 6 shows a confusion table containing sample images that were correctly classified or misclassified using our best reported approach.

We also observe from other experiments we conducted that predicting based on the proposed categorical distribution consistently outperforms the standard binary classification schemes by around 1-2%[2]. Our experiments affirm the importance of preserving the distribution of ratings or consensus across voter ratings; a point well supported by a couple of recent works [13, 15].

[2]Supplementary results are included with this paper.



(a) (b)

Fig. 5. Class activation maps (CAM) of two sample images from AVA. For each pair, we show activations from high aesthetic *(right)* and low aesthetic *(left)* classes. (a) Image 1: MultiGAP method, (b) Image 2: SingleGAP *(top)*, MultiGAP *(bottom)*.

Table 1. Comparison of proposed method against state-of-the-art approaches for the AVA/AVA-Comments datasets

	Model	Accuracy
Image	DCNN [2]	73.25
	RDCNN [2]	74.46
	Kao et al. [12]	74.51
	AlexNet [10] – finetuned	75.11
	DMA [5]	75.41
	GoogLeNet [16] – finetuned	75.60
	MultiGAP	**75.76**
	SingleGAP	**76.31**
	BDN [13]	76.80
Text	word2vec [21]	78.40
	1D-CNN [22]	79.48
	Naive Bayes SVM [14]	80.90
	RNN (1-layer GRU)	**81.09**
	RNN (2-layer GRU)	**81.79**
Joint	Multimodal DBM [14]	78.88
	SingleGAP + RNN (2-layer GRU)	**80.54**
	MultiGAP + RNN (2-layer GRU)	**82.27**

5. VISUALIZING AESTHETICS

The addition of the GAP blocks in our architecture allows us to generate class activation maps (CAM), as suggested by a recent work in [18]. Briefly, these CAMs indicate the area of the image that strongly activates the model for a given class.

Based on our proposed model, we have 10 rating classes. We generate a composite CAM (of the first GAP block) by finding the mean activations for CAMs from ratings 1 to 4 (i.e. low ratings), and ratings 5 to 10 (i.e. high ratings). It may be intuitive to perceive it as a range of *what* the model likes or dislikes, and *where* the model is receptive towards. Interestingly, the categorical distribution scheme provides a transition effect where activation maps of adjacent ratings are somewhat similar and appear to gradually shift over the different ratings. Supplementary visualizations are provided.

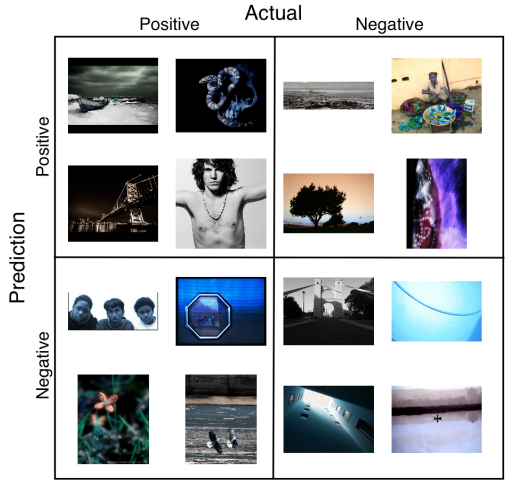


Fig. 6. Confusion table with images (our best approach)

Figure 5(a) shows the CAMs activated by a sample image classified with MultiGAP. We can see that the model "likes" the subject in focus, while it "dislikes" the background tower. In Figure 5(b), we can observe the difference in the CAMs generated from SingleGAP (top) and MultiGAP (bottom). Areas close to the building are strong activated while the word 'Sydney' interestingly, contributes to low rating.

6. CONCLUSION

In this paper, we propose a deep neural network that learns from both visual and textual features for image aesthetic prediction. Our experiments demonstrate the effectiveness of complementing both types of features, achieving state-of-the-art performance on a large-scale dataset. Also, additional meta information such as style or semantic tags could be incorporated to provide better context-sensitive judgments. The visualization of activation maps is an interesting avenue to pursue for better understanding of the *whats* and *wheres* in image aesthetics.

7. REFERENCES

[1] Naila Murray, Luca Marchesotti, and Florent Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *Proc. of CVPR*, June 2012, pp. 2408–2415.

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

[3] Pere Obrador, Ludwig Schmidt-Hackenberg, and Nuria Oliver, "The role of image composition in image aesthetics," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, 2010, pp. 3185–3188.

[4] 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.

[5] 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.

[6] Wei Luo, Xiaogang Wang, and Xiaoou Tang, "Content-based photo quality assessment," in *IEEE Int. Conf. on Computer Vision*. IEEE, 2011, pp. 2206–2213.

[7] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *IEEE Int. Conf. on Computer Vision*. IEEE, 2011, pp. 1784–1791.

[8] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[9] Gabriela Csurka and Florent Perronnin, "Fisher vectors: Beyond bag-of-visual-words image representations," in *International Conference on Computer Vision, Imaging and Computer Graphics*. Springer, 2010, pp. 28–42.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in NIPS*, 2012, pp. 1097–1105.

[11] Yueying Kao, Chong Wang, and Kaiqi Huang, "Visual aesthetic quality assessment with a regression model," in *Image Processing (ICIP), IEEE Int. Conf. on*. IEEE, 2015, pp. 1583–1587.

[12] 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.

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

[14] Ye Zhou, Xin Lu, Junping Zhang, and James Z Wang, "Joint image and text representation for aesthetics analysis," in *Proc. of the ACM Int. Conf. on Multimedia*. ACM, 2016, pp. 262–266.

[15] Bin Jin, Maria V Ortiz Segovia, and Sabine Süsstrunk, "Image aesthetic predictors based on weighted cnns," in *Image Processing (ICIP), IEEE Int. Conf. on*, 2016, pp. 2291–2295.

[16] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proc. of IEEE CVPR*, 2015, pp. 1–9.

[17] Min Lin, Qiang Chen, and Shuicheng Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[18] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *Proc. of IEEE CVPR*, 2016, pp. 2921–2929.

[19] Jeffrey Pennington, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation.," in *EMNLP*, 2014, vol. 14, pp. 1532–43.

[20] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[21] Quoc V Le and Tomas Mikolov, "Distributed representations of sentences and documents.," in *ICML*, 2014, vol. 14, pp. 1188–1196.

[22] Yoon Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.