Modeling Image Composition for Visual Aesthetic Assessment

Dong Liu, Rohit Puri, Nagendra Kamath, Subhabrata Bhattacharya
Netflix Inc.
{dongl,rpuri,nkamath,sbhattacharya}@netflix.com

Abstract

Composition information is an important cue to characterize the aesthetic property of an image. We propose to model the image composition information as the mutual dependencies of its local regions, and design an architecture to leverage such information to boost aesthetic assessment. We adopt a Fully Convolutional Network (FCN) as the feature encoder of the input image and use the encoded feature map to represent the individual local regions and their spatial layout in the image. Then we build a region composition graph in which each node denotes one region and any two nodes are connected by an edge weighted by the similarity of the region features. We perform reasoning on this graph via graph convolution, in which the activation of each node is determined by its highly correlated neighbors. Our method achieves the state-of-the-art performance on the benchmark visual aesthetic dataset [15].

1. Introduction

Automatic image aesthetic assessment has evolved from the conventional shallow machine learning models trained with hand-crafted visual features [3, 4, 16, 17, 18] to the end-to-end deep models that jointly learn visual aesthetic features and infer aesthetic ratings [8, 9, 10]. However, some unique properties related to image aesthetics are still not fully explored. Among them, the composition information of image plays a crucial role in aesthetic assessment. In visual arts, the visual elements in an image never stand alone but rather are mutually dependent on each other and collectively manifest the aesthetic property of the whole image. Therefore, it is important to design a deep neural network architecture that allows us to encode such information and leverage it to boost the performance.

Our main contributions include: (1) An end-to-end image aesthetic assessment network that leverages the composition of local regions in the image to learn aesthetics, (2) a unique feature encoding mechanism tailored to image aesthetic analysis that not only preserves visual elements and their relations but also seamlessly integrates fine grained visual details with high level semantics in the image and (3) a graph based learning framework to uncover mutual dependencies of local regions in the image. Notably, none of these contributions have been exploited in the existing works.

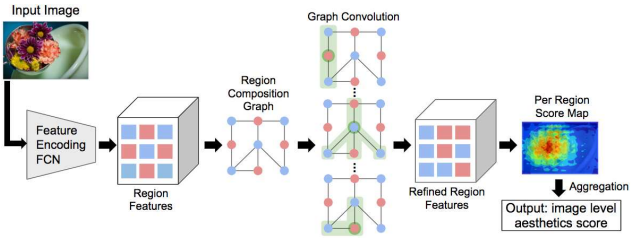

Figure 1. The proposed framework for image aesthetic assessment.

2. Our Approach

As shown in Figure 1, our model is an end-to-end feed-forward network architecture composed of three modules. The first is an Fully Convolutional Network (FCN) [22] style feature encoder that generates a 3D feature map to represent the local regions and their spatial layout in the image. The second module is a set of graph convolution [21] blocks that perform message passing across regions in the graph. The third module is a classification head that maps the feature map to the image level aesthetic score.

Feature Encoding FCN. The low level features from the shallow layers of a network describe the fine grained image details and should be fully leveraged in aesthetic assessment. Therefore, we choose DenseNet [5] as the backbone of our FCN feature encoder to preserve the fine-grained visual details in the image. DenseNet uses dense connections to feed the output of each convolution layer to all unvisited layers ahead. In this way, the low features can be maximally integrated with semantic features output at the end of the network, and serve as powerful features for learning aesthetics. Specifically, we convert the fully connected DenseNet-121 architecture [5] to an FCN. To increase the resolution of the feature map, we first remove the classification layer and the last two pooling layers in DenseNet-121, and then set the dilation rates of the convolution layers after the two removed layers to be 2 and 4 to make the pre-trained weights reusable. In this way, the dilated DensetNet-121 outputs a feature map of $1/8$ input image resolution.

Graph Convolution Over Regions. With the feature map obtained from FCN, we construct a region composition graph over the local image regions. In the graph, each node represents a local region, and we connect each pair of nodes with an edge weighed by their similarity. Mathematically, given the FCN feature map of dimensions $H \times W \times d$, we stack the feature vectors on the individual spatial locations into a matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{N \times d}$, where $N = H \times W$ denotes the total number of feature vectors, and each $\mathbf{x}_i \in \mathbb{R}^d$, $i = 1, 2, \ldots, N$ denotes the feature representation of one local region. Then the pairwise similarity function $s(\mathbf{x}_i, \mathbf{x}_j)$ between any two local regions can be defined as $s(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \theta(\mathbf{x}_j)$, where $\phi(\mathbf{x}_i) = \mathbf{A}\mathbf{x}_i$ and $\theta(\mathbf{x}_j) = \mathbf{B}\mathbf{x}_j$ are two linear transformations applied on the feature vectors [24] with $\mathbf{A} \in \mathbb{R}^{d \times d}$ and $\mathbf{B} \in \mathbb{R}^{d \times d}$ being weight matrices optimized via back propagation. After performing row-wise softmax normalization, the similarity matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$ is taken as the graph adjacent matrix representing the relations between the nodes, which characterizes the mutual dependencies of local regions in the image. Finally, we perform reasoning on the graph by applying graph convolution [21] defined as $\mathbf{Z} = \mathbf{S}\mathbf{X}\mathbf{W}$, where $\mathbf{W} \in \mathbb{R}^{d \times d}$ is the weight matrix for one layer and $\mathbf{Z} \in \mathbb{R}^{N \times d}$ is the output feature from the layer. In this work, we stack three graph convolution layers to model region dependencies.

Aesthetic Prediction Head. We use a $1 \times 1 \times d$ small kernel followed by `softmax` function to produce one score $Y_{ij}^c \in [0, 1]$ at each of the $H \times W$ spatial locations on the feature map and for each of the aesthetics classes $c \in \{0, 1\}$, where 0 denotes the class of "*low easthetics*" and 1 denotes that of "*high aesthetics*". To predict the aesthetic label at the image level, we choose a smooth convex function called *Log-Sum-Exp* (LSE) as our aggregation function [19]: $y^c = \log\left(\sum_{i,j} \exp(rY_{ij}^c)/HW\right)/r$, where r (set as 4) is a hyper-parameter controlling the smoothness of the approximation. The output $y^c \in \mathbb{R}$ denotes the image level aesthetic score aggregated over the local regions. We further convert the image-level scores into class conditional probabilities by applying a `softmax` function.

Implementation Details. The DenseNet-121 used as backbone of our feature encoding FCN is pre-trained on ImageNet [20]. We then fine-tune our proposed network with images in the domain of aesthetic assessment. The *cross entropy* loss is employed as the training objective function. During training, the input images are resized to 300×300, followed by data augmentations including randomly flipping, randomly scaling images in the range of $[1.05, 1, 25]$ the input image size and then randomly cropping 300×300 image patches. The model is trained on a 4-GPU machine with a min-batch size of 32. Adam optimizer [7] is used to train our model for 80 epochs, starting with learning rate of 10^{-4} and reducing it exponentially. Batch normalization [6] is used before each weight layer to ease the training.

Method	Acc. (%)
FC-CNN	80.45
FCN	81.43
RDCNN [10]	74.46
DMA-Net-ImgFu [11]	75.41
MT-CNN [8]	76.58
BDN [25]	76.80
AA [26]	77.00
Adaptive-Rank-CNN [9]	77.33
MNA-CNN-Scene [13]	77.40
APM [12]	80.30
NIMA [23]	81.51
A-Lamp [14]	82.50
FCN-GC (our)	82.33
ASPP FCN-GC (our)	**83.59**

Table 1. Comparison to the baselines as well as the state-of-the-art aesthetic models on the standard AVA test set.

3. Experiments

Dataset. We use the benchmark database for Aesthetic Visual Analysis (AVA) [15] as the testbed to evaluate the proposed model. It contains around $250,000$ images downloaded from *DPChallenge* [1]. To ensure fair comparison, we follow the same training/test data partition of the AVA dataset as the previous work [10, 11, 12, 13], in which there are around $230,000$ images for training and $20,000$ images for testing. From the training set, we further hold out $2,000$ images as the validation set for validating model hyper-parameters. We report the binary classification accuracy, the standard AVA evaluation metric for all the experiments.

Performance. In Table 1, we compare our model with some baselines as well as the state of the arts. The first two methods are the baseline methods, in which *FC-CNN* is the fully connected DenseNet-121, and *FCN* is the DenseNet FCN without Graph Convolutions (GC). Although these methods use the same network backbone as our method, their performances are much worse. This verifies the value of modeling local region dependencies using FCN and graph convolution. We also compare the proposed FCN-GC to the state-of-the-art aesthetic prediction models. As seen, our method beats most of them by wide margins. It is interesting to highlight that if we further introduce *Atrous Spatial Pyramid Pooling* (*ASPP*) [2], a mechanism to bring multiscale context information to the feature output of FCN, our method ASPP FCN-GC can achieve the accuracy as high as 83.59%, outperforming all existing methods.

4. Conclusion

We have presented an end-to-end network architecture for learning image aesthetics from the composition information of image. Our method builds a region graph to represent the visual elements and their spatial layout in the image, and performs reasoning on the graph to uncover the region dependencies. Our method achieves the state-of-the-art performance on the benchmark visual aesthetic dataset.

References

[1] www.dpchallenge.com. 2

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *TPAMI*, 2018. 2

[3] R. Datta, D. Joshi, J. Li and J. Wang. Studying Aesthetics in Photographic Images Using a Computational Approach. In *ECCV*, 2006. 1

[4] S. Dhar, V. Ordonez and T. Berg. High Level Describable Attributes for Predicting Aesthetics and Interestingness. In *CVPR*, 2011. 1

[5] G. Huang, Z. Liu, L. Maaten and K. Weinberger. Densely Connected Convolutional Networks. In *CVPR*, 2017. 1

[6] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *ICML*, 2015. 2

[7] D. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. 2

[8] Y. Kao, R. He and K. Huang. Deep Aesthetic Quality Assessment with Semantic Information. *TIP*, 2017. 1, 2

[9] S. Kong, X. Shen, Z. Lin, R. Mech and C. Fowlkes. Photo Aesthetics Ranking Network with Attributes and Content Adaptation. In *ECCV*, 2016. 1, 2

[10] X. Lu, Z. Lin, H. Jin, J. Yang and J. Wang. RAPID: Rating Pictorial Aesthetics using Deep Learning. In *MM*, 2014. 1, 2

[11] X. Lu, Z. Lin, X. Shen, R. Mech and J. Wang. Deep Multi-Patch Aggregation Network for Image Style, Aesthetics, and Quality Estimation. In *ICCV*, 2015. 2

[12] N. Murray and A. Gordo. A Deep Architecture for Unified Aesthetic Prediction. *arXiv:1708.04890*, 2017. 2

[13] L. Mai, H. Jin and F. Liu. Composition-preserving Deep Photo Aesthetics Assessment. In *CVPR*, 2016. 2

[14] S. Ma, J. Liu and C. Chen. A-lamp: Adaptive Layout-aware Multi-Patch Deep Convolutional Neural Network for Photo Aesthetic Assessment. In *CVPR*, 2017. 2

[15] N. Murray, L. Marchesotti and F. Perronnin. AVA: A Large-Scale Database for Aesthetic Visual Analysis. In *CVPR*, 2012. 1, 2

[16] L. Marchesotti, N. Murray, and F. Perronnin. Discovering Beautiful Attributes for Aaesthetic Image Analysis. *IJCV*, 2015. 1

[17] L. Marchesotti, F. Perronnin, D. Larlus and G. Csurka. Assessing the Aesthetic Quality of Photographs using Generic Image Descriptors. In *ICCV*, 2011. 1

[18] V. Ordonez, S. Dhar and T. Berg. High Level Describable Attributes for Predicting Aesthetics and Interestingness. In *CVPR*, 2011. 1

[19] P. Pinheiro and R. Collobert. From Image-level to Pixel-level Labeling with Convolutional Networks. In *CVPR*, 2015. 2

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 2

[21] F. Scarselli, M. Gori, A. Tsoi, M. Hagenbuchner and G. Monfardini. The Graph Neural Network Model. *TNN*, 2009. 1, 2

[22] E. Shelhamer, J. Long and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. *TPAMI*, 2016. 1

[23] H. Talebi and P. Milanfar. NIMA: Neural Image Assessment. *arXiv:1709.05424*, 2017. 2

[24] X. Wang, R. Girshick, A. Gupta and K. He. Non-local Neural Networks. In *CVPR*, 2018. 2

[25] Z. Wang, D. Liu, S. Chang, F. Dolcos, D. Beck and T. Huang. Image Aesthetics Assessment using Deep Chatterjee's Machine. In *IJCNN*, 2017. 2

[26] W. Wang, J. Shen, and H. Ling. A Deep Network Solution for Attention and Aesthetics Aware Photo Cropping. In *TPAMI*, 2018. 2