Deep Aesthetic Learning

Estimating Photo Aesthetic Rating Using Deep Convolutional Neural Network

Enhao Gong
Electrical Engineering
Stanford University
enhaog@stanford.edu

Abstract

Deep Learning has been widely applied in several computer vision applications such as item recognition and image retrieval. In addition to identifying what a image contains, it is also a interesting problem to explore about whether the image is appealing or the composition of the items are highly rated aesthetically.

There are multiple studies using hand-designed or data-driven model to quantify the metrics regarding to photo quality such as color distribution, composition, sharpness etc. Aesthetic rating is such a complicated and unknown function of human beings' psychological responses to the pixels that deep networks are potentially better tools to achieve similar modeling and predictions. Existing large aesthetic rating data-set also enable the training good deep networks.

In this work, I developed a convolution neural network dedicated for learning to model and predict photo ratings from pixels. Novel data augmentation scheme and further fine tuning regression from the learned features are the two main parts of the contributions in this work. Results demonstrate better prediction performance compared with existing solutions using either conventional engineered features or deep networks. In addition, the trained network is able to out-perform human beings in estimating the aesthetic rating of images.

The resulting performance are summarized and more details of network design and tuning are further discussed. The trained networks are further quantified which also provides insights to the understanding of human being's appreciation to art and beauty.

1. Introduction

1.1. Aesthetic Qualities

The aesthetic quality of a piece of art designs, such as a photo, results in psychological responses in human beings. There are multiple aspects that enable a high aesthetic quality of a photo.

Firstly, some technical and objective metrics are obvious related to photo qualities, such as exposure and sharpness. A photo can only be fully appreciated if it is well exposed. Over exposed and under exposed photos are usually be treated as bad shots. Also, blurring often appears if a photo is not well shot or there are motions that relatively larger compared with the shutter time. There are a lot metrics quantify theses objective technical qualities.

Secondly, in contrast to objective qualities of a photo, human beings are also tend to appreciate varies subjective quality of a photo such as whether the composition is balanced and how the color distributed. It is hard to design a metric to quantify these aesthetic qualities but various studies shows learning algorithms can be designed and tuned to predict metrics related to composition (such as whether a photo matches Rule-Of-Third or Golden Angle) and color (which genre/style a photo/picture belongs to).

Last but not least, photos are also related to people's memory. Studies shows people are good at remember precise details in image. The most beautiful and meaningful moments are distinctive when they are presented in photos. Both extrinsic (related to human behavior and personal experiences) and intrinsic image features for making an image memorable are studied.

1.2. Challenges

The largest challenge to predict photo qualities is that most algorithms are utilizing low-level image features that cannot sufficiently character the high-level perception of aesthetics. Also, how to combine and integrate various qualities (technical and emotional, subjective and objective,

etc.) is still an open problem. Furthermore, in order to train good model to estimate aesthetic rating of photos, enough amount of detailed data is a necessity.

1.3. Deep Aesthetic Learning

In this work, we designed a dedicated convolutional deep neural network [2] to tackle these problems.

The detailed goal is to estimate (with regression) the average human rating on different photos.

The resulting performance are summarized and more details of network design and tuning are further discussed. The trained networks are further quantified which also provides insights to the understanding of human being's appreciation to art and beauty.

2. Data Preparation

2.1. Datasets

In order to achieve good modeling of the complicated aesthetic qualities, large dataset of photo aesthetic ratings are of great importance. Here we are mainly use the AVA dataset [4] that contains human ratings on over **250k** photos. Each data point is human's response on a specific photo that was posted on the DPChallenge website which were usually rated by over 200 users. Each rate is from 0 to 10 so the final averaged rate is a float value in the rage of (0,10). I used 230k data for training and 20k for regression. The mean averaged rating is 5.3 ± 1.4.

I downloaded the dataset information describing the photo IDs of all the photos and I developed python script to crawl all the data from DPChallenge website. .

Figure 1. Ava dataset

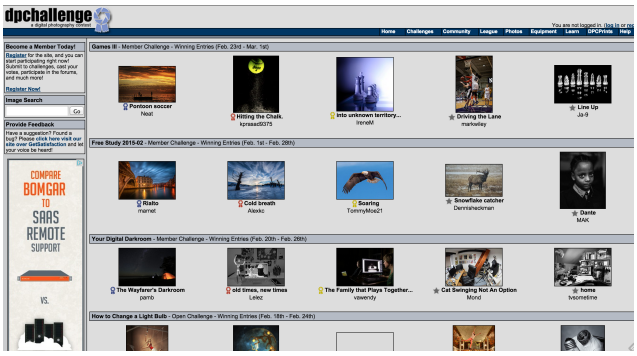

Figure 2. Example of dpChallenge website with lists of photos for people to rate

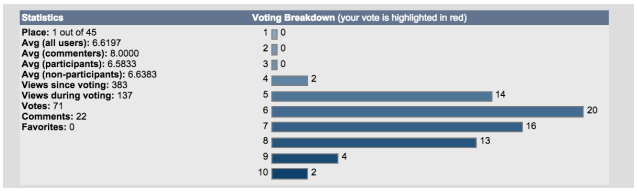

Figure 3. Example of rating distribution of images on dp challenge

3. Methods

3.1. Non-CNN methods

Here I also implemented a baseline algorithm to evaluate the performance.

3.1.1 Features and Dimension Reduction

The are a lot of studies using engineered image features to predict photo ratings. I also implemented similar algorithm with basic feature sets of a photo include:

- Color Distribution: 3D HSL Histogram.
- Composition: Blurred Image and Salient Map (Graphcut-based algorithm), GIST descriptor
- Sharpness: Gradient and Gradient entropy
- Details: SIFT BOW descriptor

The steps for prediction includes:

- Dimension Reduction with PCA
- Reduced feature based regression (Support-Vector-Regression).

The implementation of traditional algorithm provide baselines for predicting photo ratings.

3.2. CNN based methods

The main structure of the CNN is $(Conv\text{-}ReLU\text{-}Pool) \times N + Affine \times M + Regression\ Layer$. Different from the models in examples, the architecture is modified based on the new regression objective function, Euclidean Loss for regression, L_2 norm of the differences between ground-truth rating and predicted rating.

Previous studies by *Lu et.al.* [3] has developed complicated CNN to aesthetic ratings. The general idea is they combine basic single column CNN on scaled global view of image with another other CNNs integrating local and global view and boosted with style information:

- Local View CNN for local quality: a CNN trained with randomly sampled local view of images,
- Style CNN for Regularized Learning: a CNN trained with style data-sets

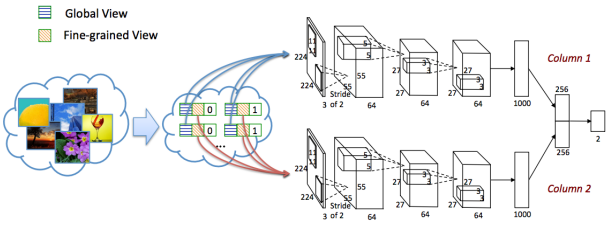

Figure 4. CNN structure from RAPID paper, integrating local and global view and boosted with style information

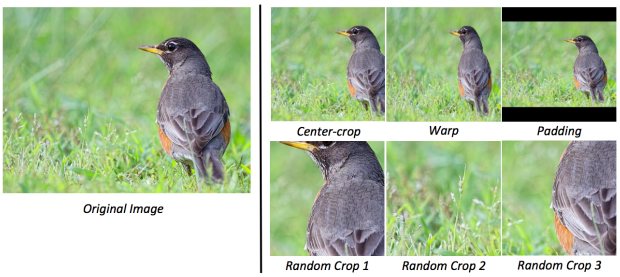

Figure 5. The sampling and data augmentation scheme used for RAPID paper which augment the data with local views and global views

New structures and schemes are developed for this work shown in figure **??** and table 1 below .

3.3. Boost CNN with Data Augmentations

In this way, they incorporate features in unity from global views and local views. In addition, they used the style attributes of images to further improve aesthetic quality quantification with style categorizations related information.

Table 3: Accuracy of Aesthetic Quality Categorization for Different Methods

δ	[24]	SCNN	AVG_SCNN	DCNN	RDCNN
0	66.7%	71.20%	69.91%	73.25%	74.46%
1	67%	68.63%	71.26%	73.05%	73.70%

Figure 6. The prediction results from RAPID paper. They built complicated regularized learning triple-column CNN model to out-perform existing machine learning algorithm with engineered features

Table 1. CNN structure and parameters

Layer	Parameters	Channels	ReLU	POOL
Data	256x256	6		
Conv1	6x5x5	32	ReLU	POOL
Conv2	32x3x3	64	ReLU	POOL
Conv3	64x3x3	64	ReLU	POOL
Conv4	64x3x3	64	ReLU	POOL
Conv5	64x3x3	64	ReLU	POOL
Conv6	64x3x3	64	ReLU	POOL
FC5	FCx1000			
FC6	FCx256			
FC7	FCx1			

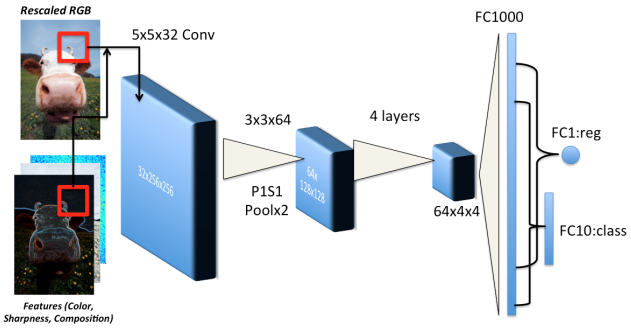

Figure 7. CNN structure for this work. Integrating global and local information with Data Augmentation. Boost the performance with fine tuned regression.

Here based on similar insight but with a different scheme, I developed deep networks with simple single column CNN structure but incorporate further information using Data Augmentation way. Since CNN works the best with local features, global information such as shapes and compositions are not among CNN's advantages. Thus, I append features with global information to the additional channels of images.

Instead of simple RGB 3 channels, the dataset fed into the CNN is a 6 channel image-liked data:

- 3 channels of RGB

- 1 channel of color distribution
- 1 channel of details(edge) distribution
- 1 channel of composition (salient map)

Based on the output table, we can see the results demonstrate the trained CNN with data augmentation but simple structure can outperform conventional machine learning algorithms with engineered features.

3.4. Further Boost CNN with Fine Tuning Regression

In order to achieve better performance, we implemented transfer learning and fine tuning. We get the CNN features learned and developed a fined tuning regression model. The regression model using here is Support Vector Regression (SVR) with input from features from layer and the ground-truth for regression is the averaged ratings.

The SVR maps features from the original finite dimensional space into a much higher-dimensional space that is able to fit what linear regression can never captures. So intuitively, this SVR based Fine Tuning Regression further boost CNN with much deeper network structures.

3.5. CNN implementation

The CNN is trained using Caffe [1] and python. I developed data preparation script in python (with image scaling, sampling and augmentation with engineered features), the Caffe CNN description script, result parsing script in python (route the console output to text file and parse the output to get the CNN loss as well as learned features). Further CNN parameter tuning is based on the parsed results. Also the Fine Tuning Regression is based on the output features.

4. Results

4.1. Goals and Metrics

The most straightforward metrics to quantify the performance is the error of rating prediction. For DPChallenge, the rating range is well defined in 0 10.

So there we explore two sets of accuracy metrics for performance estimation:

- Accuracy for binary classification for two-classes high-rated and low-rated (consistent with other studies with the same dataset). It is a percentage value.
- Root-Mean-Square-Error (RMSE) for regression about how accurate the model can predict the rating score. It is a value in unit of score.

4.2. Results

Here shows the results of both classification and regression:

Table 2. Results of Binary (2-class) Classification, CNN1:without fine tuned regression, CNN2: with fine tuned regression

Methods	AVA	CNN	RAPID	CNN1	CNN2
2-Class Acc.	67%	68.6%	73.7%	70.6%	75.2%

Table 3. Results of Regression. The RMSE of the trained model with fine tuned regression out-performance human behaviors which is quantified with the Standard Deviation of human being's rating from 0 10

Methods	CNN2	Human Performance
RMSE	0.71	1.43

5. Discussion and Future Plans

5.1. Implementation and Training

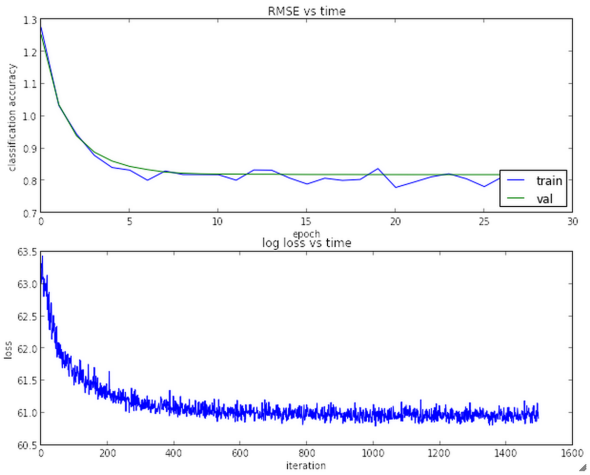

Figure 8. The loss converges shown as training of CNN goes on

5.2. Error analysis

I plotted the prediction results, sorted and outputted where the largest errors happened. Current CNN based method still perform limitedly at some conditions:

- over-estimation appears in some cases including images with dark background
- over-estimation appears in some cases including images with higher contrasts
- under-estimation appears in cases with simple structure, low contrasts but with specific content.

Some examples are shown in figure below 9.

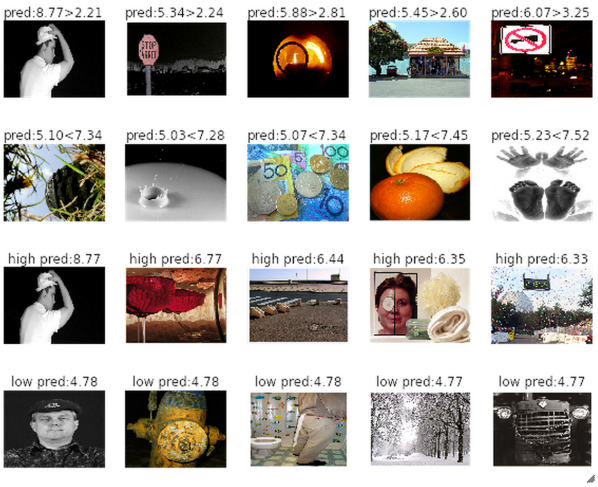

Figure 9. Example of errors of the CNN. Over-estimation appears in some cases including images with dark background and higher contrast and some under-estimation appears in cases with simple structure, low contrasts but with specific content.

5.3. More data augmentations

Also it is possible to introduce more improvement by fusing different quality metrics and augmenting data from positive examples.

In this work we used the AVA datasets to estimate the aesthetic ratings of photos directly from photos with positive and negative examples (highly and lowly rated photos).

In addition, further exploration with the learned model can be applied on the datasets about all kinds of technical qualites of photos from wikipedia database. There are example photos of optimal selection regarding to good composition, good exposure, good coloring as well as good choice of blurring/sharpness etc. Different from the rated photo from DP-Challenge, these photos are all positive examples.

Also there are labels of photos with different aethestic styles, such HDR, etc. This labels can also be used to boost the performance.

One possible solution is to use separate CNNs with regularized learning, similar to previous studies [3]. Also, we can use these data to improve the training by augmenting data from the available positive examples.

5.4. Personalized data

In this work we demonstrate how CNN can be powerful to predict human's perceptions on aesthetic quality. The CNN model even though with a very simple structure but can outperform traditional algorithms as well as human being's performance by using specific data augmentation as well as fine tuned regression to transfer the learned features.

However aesthetic preference can also be actually personalized issue and subject dependent which is shown from the large standard deviation of human's rating on DPChallenge and AVA dataset. There should not be a globally optimal predictor working for everyone.

For future exploration, the CNN framework can be used to study subject differences in photo rating, qualtify where the inter-personal effect start to appear in the deep network and possibly provide personalized predictor or training framework.

Data, personalized data instead of the global data like AVA dataset, is necessary to address the inter-personal differences of aesthetic perception. Also it can be used as a data source for real-time on-line learning and updating. For example, each personal should have different CNN weights that is updated with its own data point.

One possible solution I may further pursue is to use first-hand data on this problem from Polarr which is an online photo editing platform provide professional photo curation, editing and enhancement (`https://polarr.co`) shown in fig .

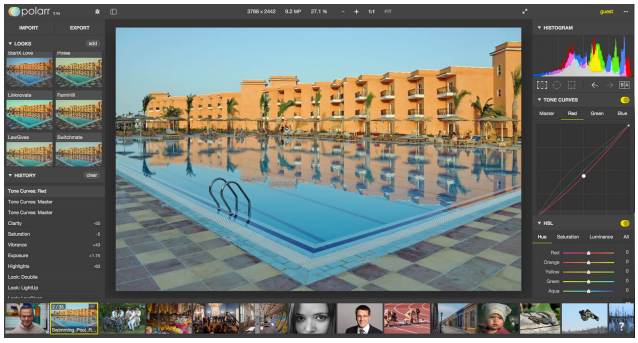

Figure 10. Personalized photo preference data is downloaded from online photo curation platform Polarr.

6. Conclusion

- Here I designed and developed a dedicated deep convolutional neural network to predict photo's aesthetic ratings.
- Data Augmentation and further Fine Tuned Regression greatly boosted the performance.
- The final results outperformed existing solutions in regression and classifications.
- To certain extent, the solution beat human performance.
- This work provided a way for automatic photo rating and possibly for recommended editing
- To further improve the model, I plan to explore different loss functions, more data augmentations (multiple

views) as well as the improvement of Transfer Learning and Regularized Learning

- Personalized data can make the model works for each individual

References

[1] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. *h ttp://caffe. berkeleyvision. org*, 2013.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[3] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the ACM International Conference on Multimedia*, pages 457–466. ACM, 2014.

[4] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2408–2415. IEEE, 2012.