Image Aesthetics Assessment based on High Level Attributes using Deep Learning

Ananya Pal, Anuja Apte, Anjali Iyer, and Saaket Joshi

Maharashtra Institute of Technology, Pune 411004, India,
www.mitwpu.edu.in

Abstract. Aesthetic assessment of images has been getting a lot of attention for the past decade in the field of computer vision. Large amounts of social media and advertising data in the form of images is continuously analysed to assign it an aesthetic quality value to improve businesses as well as for gaining more popularity across the web. Visual perception by humans cannot be fully replicated by a machine and continuously more work is being published on aesthetic classification of images. In this paper, we have presented a convolutional neural network model which automatically extracts high level features and distinguishes a set of images into pleasing and non-pleasing categories. Our dataset has been compiled from a variety of sources on the web to make it as diverse as possible. Compared to the traditional handcrafted methods and other machine learning models, our CNN model has provided a better classification accuracy of 68% on our dataset.

Keywords: machine learning, cnn, deep learning, image aesthetic classification, high level attributes

1 Introduction

Image Aesthetic Assessment refers to the evaluation of images on the basis of high and low quality of aesthetic value of the picture. It plays an important role in the field of Computer vision which deals with automatic extraction and analysis of useful information from a picture or sequence of pictures. What makes this task a little challenging is that aesthetic value of an image may be differently perceived by each individual and thus automation of this process may not be very adept at accurately judging the aesthetic quality of an image. Many visual features play a part in the perception of an image by a human. There are two basic categories of such features among others i.e. Low Level and High Level Attributes.

1.1 Low Level Attributes

Low-level features include spatial characteristics such as edge density, straight-edge density and entropy, and color characteristics such as hue, saturation, and brightness.[1]

1.2 High Level Attributes

High level attributes are built on top of low level attributes to detect larger shapes and objects in an image.

- Presence of Salient Object: The image includes a large salient object well separated from the background.
- Rule of Thirds: The main subject lies on the intersection or on the lines of a 3x3 grid.
- Depth of Field: The region of interest is in focus and background is blurred, often to emphasize on the object of interest.
- Opposing Colors: Color pairs of opposing hues are prominent in the image.[1]

2 Approach

We have considered a unique combination of High Level Features, that is, the Rule of Thirds, Depth of Field and Color Contrast to be the key factors in evaluating our model.

2.1 Handcrafted Methods

Color Contrast The perception of an image being pleasing or not pleasing depends on the colors or a combination of colors used in that image. Contrast is the difference between the color and brightness of the object and color and brightness of the background.This color difference makes the focused object distinguishable.[2]We have extracted the foreground and background of an image using the Grabcut algorithm.Further,we are calculating the RGB mean values of the extracted foreground and background images and taking there difference. The difference value of RGB mean values is compared with a given threshold and segregated into High color contrast images and Low color contrast images.

- High Contrast images are usually the images that have two contrasting colors for object and the background.These images will contain dark shadows and bright highlights.The figure below is a high contrast image.

- Low Contrast images are do not have a big difference between their shadows and highlights.These images use supplementary colors on the color wheel.The lack of brightness may result in a flat image as the object and background may not be distinctly visible.

Rule Of Thirds The rule of thirds involves dividing your image using 2 horizontal lines and 2 vertical lines. It is the positioning of the important elements in your scene along those lines, or at the points where they meet for better visual results.[3].We have extracted the foreground and background of an image using the grabcut algorithm.For Rule of Thirds,we use the foreground extracted image. On this image, we first calculated the centroid of the object and then the intersection points of the grid.Next,we checked the distance of the centroid from the intersection points. Based on a specific given threshold,the image is classified as either it follows RoT or it doesn't follow RoT.

Depth Of Field The region of interest is in sharp focus and the background is blurred in a depth of Field.It may be more effective, emphasizing the subject while de-emphasizing the background[3][4].We have extracted the foreground and background of an image using the Grabcut algorithm. Using the Laplacian filter, the noise of the extracted background image is evaluated.This noise value is eventually considered as the value of the depth of field in an image.

2.2 CNN approach

Dataset The dataset consists of 6000 images gathered from a wide range of heterogenuous sources. We have considered the following High level attributes - Rule of Thirds, Depth of Field, and Color Contrast. We have 2000 images of each category, following as well as not following each of the above rules. We have further divided the dataset into Appealing and Non-Appealing for the basis of training the model based on above attributes. The dataset is split randomly into training and testing sets. The Training set consists of 4200 images and the Testing set consists of 1800 images(0.3 of total number of images). For preprocessing, we have converted the images to grayscale, resized the images into size (128*128) and standardized the dataset by scaling.

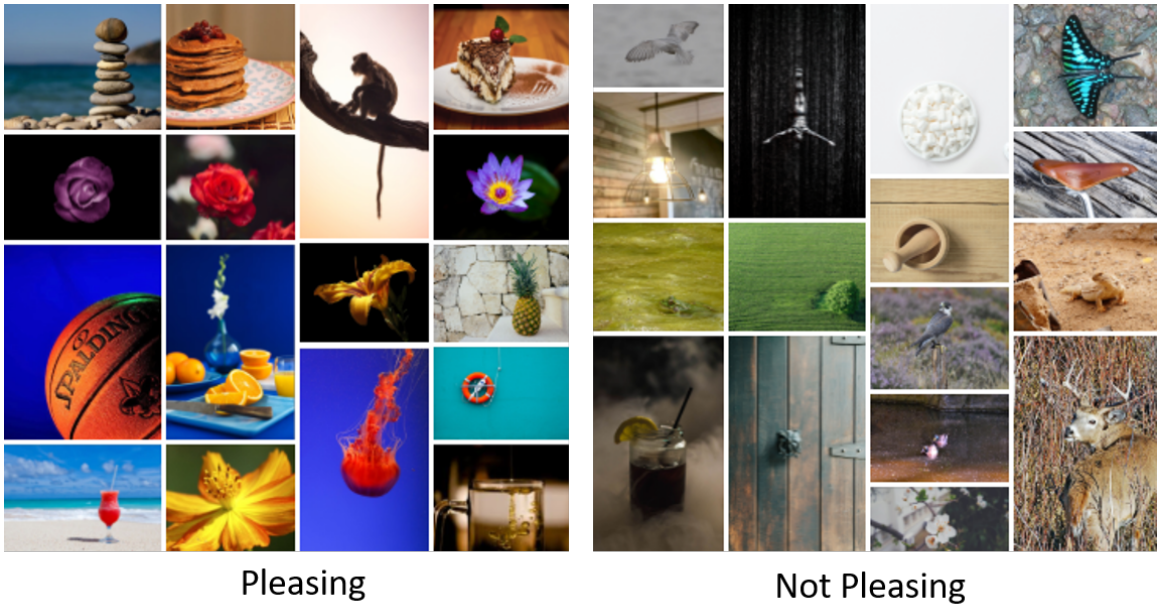

Fig. 1: The dataset consisting of Pleasing and Non-Pleasing images

CNN Design In this section, we will give the detailed explanation of the design and architecture of CNN model. The model receives black and white images of size 128*128 as input and has five convolution layers each followed by Max Pooling (of size 2*2). All these layers contribute in the automatic feature extraction for the three high level features considered. This is followed by 2 fully connected layers. All these layers use rectified linear activation function except the output layer. The output layer uses sigmoid activation for 2-class classification into Pleasing and Non-Pleasing classes. The number of filters in the convolutional layers are 64,64,128,128 and 64 respectively. The kernel size of the convolutional layers are (7*7),(3*3),(3*3),(7*7) and (5*5) respectively.

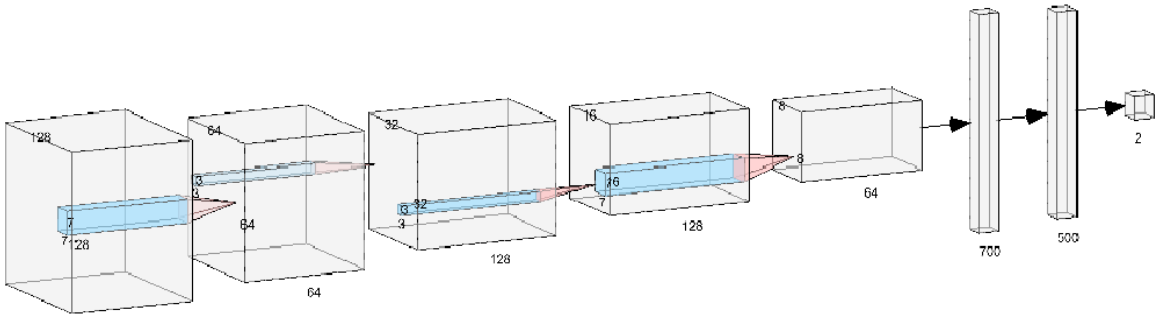

Fig. 2: CNN Architecture

2.3 Results

Our Deep Learning model provides an classification accuracy of 68% and gives a loss rate of 61%. The validation data has been tested with respect to the training dataset and graphs of accuracy and loss are plotted as shown below.

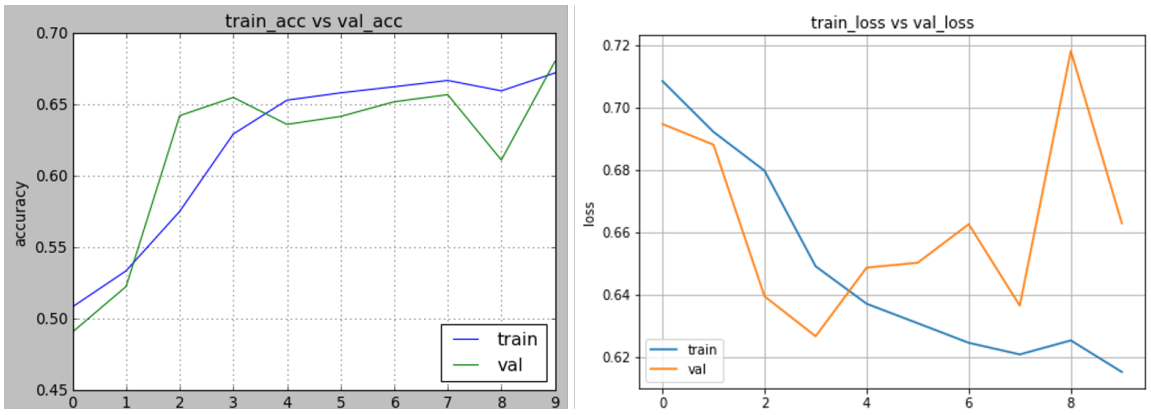

Fig. 3: Graphs showing the comparison of scores between Training and Validation sets in terms of loss rate and accuracy. The X axis represents the validation samples and Y axis gives the corresponding accuracy/loss.

2.4 Conclusion and Discussion

The classification of an image based on its aesthetic appeal can be tricky and challenging since human perception is greatly subjective and unpredictable at times.[4] However, we have presented a CNN model which takes into consideration the three high level features - Rule of Thirds, Depth of field, and Color Contrast, and gives an accurate result for most of the images based on these features. We have achieved an accuracy of 68% in our deep learning model. The outliers that are wrongly classified using the traditional handcrafted modules, are correctly classified by the deep learning module. Besides the outlier classification of the traditional handcrafted methods, machine learning algorithms like Support Vector Machine provided an accuracy of nearly 50% on our dataset. Thus, we can conclude that our Deep learning model provides better performance compared to the traditional algorithms available for aesthetic classification of images.

Outliers			
Actual Class	Appealing	Appealing	Appealing
Handcrafted	Not Appealing	Not Appealing	Not Appealing
Deep learning	Appealing	Appealing	Appealing

Fig. 4: A few examples of the Comparison between Handcrafted module and CNN module based on outliers.

Bibliography

[1] Sagnik Dhar, Vicente Ordonez, Tamara L Berg, "High Level Describable Attributes for Predicting Aesthetics and Interestingness," Colorado Springs: USA, 2011.

[2] Steven W. Smith, Ph.D.,The Scientist and Engineer's Guide to Digital Signal Processing.https://www.dspguide.com/ch23/5.htm

[3] Sagnik Dhar,Vicente Ordonez, Tamara L Berg, "High Level Describable Attributes for Predicting Aesthetics and Interestingness," Stony Brook University,Stony Brook, NY 11794, USA.

[4] Yiwen Luo and Xiaoou Tang , "Photo and Video Quality Evaluation: Focusing on the Subject," Department of Information Engineering The Chinese University of Hong Kong, Hong Kong.