CSE 700 – Machine Learning

Prof. V. V. Phoha

Theme-based Aesthetic Image Assessment

by

Pratheek Basavana Gowda

and

Amritbani Sondhi

EE&CS, Syracuse University

Spring 2018

Theme-based Aesthetic Image Assessment

** Pratheek Basavana Gowda
Syracuse University
Syracuse, New York
pgowda@syr.edu

** Amritbani Sondhi
Syracuse University
Syracuse, New York
asondhi@syr.edu

Abstract

Aesthetics in the world of art, photography and psychology refers to the principle of the nature and appreciation of the beauty. With the rise in popularity of digital camera, the visual data available on the web is growing exponentially. Some of the pictures are beautiful and aesthetically pleasing but vast images are of low quality or are not so good. This proposal aims to demonstrate a simple yet powerful way to aesthetically rate and classify the images based on the features that are present in the images that pleases the human eye.

In this work, we treat the challenge of inferring aesthetic quality of pictures using their visual content, with rated images from data sources. We will train the predictor, using already aesthetically pleasing images that will allow us to choose the best suited features. We were inspired by Deep Chatterjee's Machine (DCM) model [1] for this task. DCM first learns attributes through the parallel supervised pathway on a variety of selected feature dimensions. A high-level synthesis network is trained to associate and transform these attributes into an overall aesthetic rating.

Problem Statement

Using the computational features that are present in an image, to predict the aesthetic quality of the image. The challenge with predicting the aesthetic rating with maximum accuracy is that, aesthetics is highly subjective and sensitive to human judgement. Also, the semantic gap between low level-computable features and high- level human oriented semantics makes accurate predictions difficult. This was a problem faced by the DCM model, and the previous works too.

Applications

With exponential rise in digital images on the internet, the necessity for assessment of images has considerably increased. For example, in aesthetic image cropping where out of a large image you can derive an aesthetic image that is more pleasing. In picture editing software to produce appealing polished photography. Also, in image classification systems for image ratings, in image ranking algorithms and image retrieval systems.

Significance of our work

The DCM model achieves high accuracy as compared to the RAPID model. But, in terms of complexity, DCM's network consists of 5 parallel pathways with a high-level synthesis network, thus making the design more complex. Our approach focuses on proposing a simple model with focus on theme based classification of the images with rating them, which the DCM model didn't take into consideration, and thus suffered with deviation in ratings despite having state of the art architecture to do so.

Related Work:

Datta [4] first casted the image aesthetics assessment problem as a classification or regression problem. They extracted certain visual features based on the intuition to discriminate between aesthetically pleasing and displeasing images. Automated classifiers were built using support vector machines (SVMs) and classification trees. Linear regression on polynomial terms of the features were also applied to infer numerical aesthetics ratings.

However, the various extracted hand-crafted features including both low-level image statistics and high-level photographic rules cannot accurately and exhaustively represent the aesthetic properties. A lot of work has been done for aesthetic assessment using generic feature sets such as SIFT and Fisher Vectors, for object classification tasks. However, they are unable to attain the upper performance limits in aesthetics-related problems.

Aesthetic related features inspired from photography or psychological literature are:

- Low level image statistics: distributions of edges, color histograms
- High level photographic rules: rule of thirds

The scope of these manually designed attributes is limited. Also, the vagueness of certain photographic and psychological rules and difficulty in implementing them computationally. There is a lack of principled approach to improve the effectiveness of such features.

Rating pictorial aesthetics using deep learning (RAPID) [2] is among the firsts to apply deep convolutional neural networks to the aesthetic rating prediction. It improved the model by style annotations associated with images and even hidden activations from a generic CNN worked well for aesthetic features.

Further, improving over RAPID, the DCM introduced a specific architecture of parallel supervised pathways to learn multiple attributes on various selected feature dimensions, which are then associated and transformed into an overall aesthetics rating by a high-level synthesis network. It also predicted the distribution of human ratings, since aesthetic ratings are subjective.

DCM failed to incorporate the creative ideas and the subject of the images such as, images that do not have appealing composition or color, but that might seem high in aesthetic quality to the human eye for specific themes. This emphasizes the need to understand the context in which the image is being judged. Hence, we put forward our approach to train our predictor in a theme specific environment, so that the accuracy of prediction reaches a near-human performance to rate the images.

** people have contributed equally

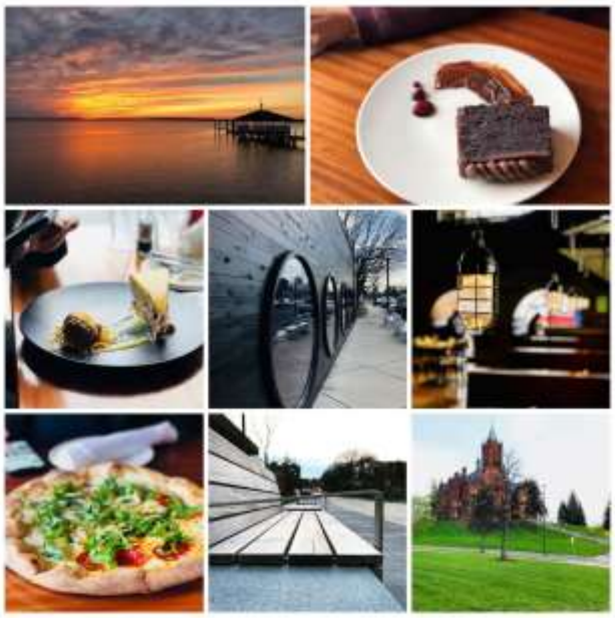

Figure 1: Aesthetically High Rated Images

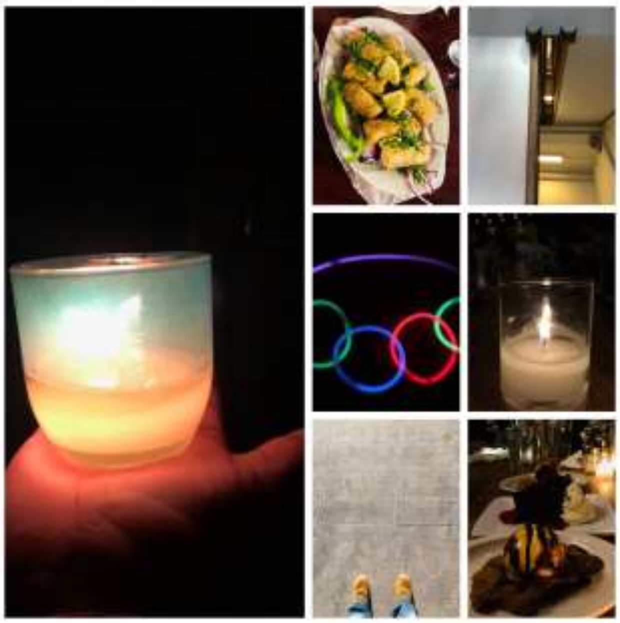

Figure 2: Aesthetically Low Rated Images

Methodology

We combined two previous approaches from the Datta et al and the RAPID model, by using simple convolutional neural networks to learn the image features and SVMs to classify it to high or a low aesthetic rating.

As suggested, by few of the authors who have worked on the same, we considered images for our dataset with varying distances. The more the distance between the high quality and low-quality images used to train the model, the better the results. To achieve this, we used a triplet loss function as suggested in Understanding Aesthetics with deep learning[7]. The objective of this loss function lets convolutional neural networks (CNNs) learn the similarity between high quality images and learn the difference between high quality and mediocre images. Therefore, by training the network, we learn feature representations. The feature distances between two high quality images will be smaller than the distance between a high-quality image and a low quality image.

The triplet loss function can be defined as,

Triplet loss = *max (0, c + Dist (ϕ(I1), ϕ(I2)) – Dist (ϕ(I1), ϕ(I3))* *[7]*

where,

ϕ is the feature representation retrieved from the CNN model,

I1, I2, I3 are the images and

c is the margin which maintains the distance between *ϕ(I1)* and *ϕ(I2)* and is always greater than *ϕ(I1)* and *ϕ(I3)*. Refer figure 3 for its diagrammatic representation.

This was done to achieve a pure classification of high and low ratings while training our model. We trained a convolutional neural network with a loss function as they perform well for image classification. A CNN basically consists of several convolutional layers on top of each other, fed to a fully connected layer in the end. Each image goes through these layers multiple times specified by the number of epochs and creates a back propagation. When the image passes through each step or epoch, the values for accuracy, recall and loss is obtained.

We divided our dataset in a 70-30 ratio. Out of the 30%, 2/3rd was used for validation and 1/3rd as the test set. Validation set is used to fine tune the trained classifier. This also helps to determine the optimal number of epochs.

In our approach, the CNN is used only for feature extraction. The benefit of using a CNN model is that it considers even the features which are hidden to the human eye. But we cannot explicitly state, which exact features are being considered by the CNN, as it keeps on training for a no. of steps to stabilize the weights. The purpose of using a CNN is only for feature learning and to determine the loss function. After the features are learnt, we needed to classify the data into 2 classes, low rated and high rated. We chose binary classification over Likert scale, as it has better and clear results and that these algorithms handle high dimensional spaces as well as large number of training examples very well. The options we considered were decision trees, random forest algorithm and SVMs. We selected SVM, as it performs better for non-linear correlations and computes losses for maximum margin classification and gives a better accuracy score. We calculate the deviation through the test dataset.

Datasets

The factors contributing to our understanding of aesthetics is the selection of datasets. The image quality and rating system reflects aesthetic value of photography. For the fulfillment of this purpose, we chose to use images from photo.net which is an online photo sharing community with over 4 lakh active photographers. Apart from photo.net, communities like eyeem.com and flicker add up to need for accurate aesthetic ratings of the images.

We fetched the images from photo.net which is an online open-source platform with high quality peer rated images from photographers around the globe. We scraped the images from the

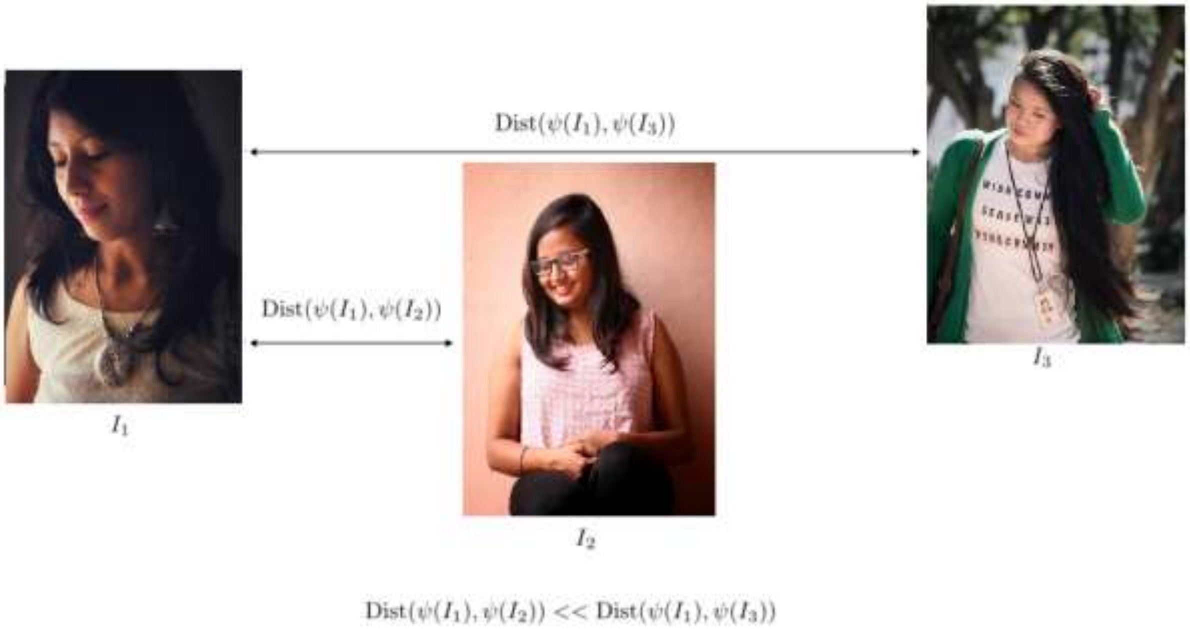
$\mathrm{Dist}(\psi(I_1), \psi(I_3))$
$\mathrm{Dist}(\psi(I_1), \psi(I_2))$
I_1
I_2
I_3
$\mathrm{Dist}(\psi(I_1), \psi(I_2)) << \mathrm{Dist}(\psi(I_1), \psi(I_3))$

Figure 3: Describing the distance between images with respect to triplet loss function

web, by using the photo ids and noting their scores. The sizes of the downloaded images were not uniform, hence we cropped and scaled the images to a dimension of 96 x 96. Through our code, we parsed the labels from the file names to retrieve the aesthetic ratings.

The downloaded images are split into train, validate and test sets. With training set containing 5652 images, validation set containing 1000 images and test set having 500 images in total.

Our design architecture

For training our convolutional network, we provide it with an 3 x 96 x 96 input. The network consists of 5 convolutional layers and 4 max pooling layers. The output of this network is a feature vector of size 1024 x 1. We connect this feature vector as an input to the support vector machine to classify it into 2 categories, that are, high quality and low-quality images.

Results and Analysis

The result from the CNN network can achieve an accuracy of 66.8%. It demonstrates the effectiveness of feature selection by using triplet loss to train the convolutional networks, but, it is still not as good as expected.

Below is the comparative analysis between our model and the other approaches, used for the same objective:

Approach	Accuracy
CNN with SVM	67%
Histogram with SVM *	64%
VGG-16 with SVM *	84%

The reasons for this might be:

- The structure of the network: The convolutional network which we used, should have been more deep and wide enough to extract more features and hence improve the score.

  Currently, our model has only 5 convolutional layers and only one fully connected layer which may not be enough to extract the best features.

- Parameter Tuning: Selecting the right parameters, takes a lot of trial and errors, resources and also time to compute the results on even a mid-sized dataset. We computed results by varying the number of epochs and learning rates.

Due to the limited resources, as well as our limited knowledge gained within the stipulated time, we successfully implemented a simplified network for applying the approach and get the results. We observe, that the minimum the loss function, the lesser the distance between the images and higher the rating of aesthetics.

Challenges faced

Extracting features from images came across as a little difficult task to achieve the results what we expected. As

* *Photography Classification in aesthetics* [8]

transforming psychological aesthetic features of images into code and formulae is difficult, there are very limited set of features which can actually be extracted and a model can be trained on it. Some of the features which we were able to extract with the help of OpenCV library in python are: R, G, B and H, S, V values, image blur, grain and edges.

As these features are limited, we decided to go for the convolutional neural network approach using tensorflow and keras. This came as a learning experience as we learned the deep learning approach from the scratch, focusing on how our own dataset is fed as an input to the connected layers and the weights are trained.

We were able to create our own dataset containing themed images and convert it to a tensor flow record to be fed to the model.

The initial approach of using a two-layer model and having a theme detector for first layer with an accuracy of 76.28% and integrating that with the second layer was a task which was not achieved at the end. The optimal value of the hyperparameters couldn't be found out for image assessment.

References

[1] Z. Wang, D. Liu, S. Chang, F. Dolcos, D. Beck, and T. Huang, "Image Aesthetics Assessment using Deep Chatterjee's Machine" in *International Joint Conference on Neural Networks (IJCNN)*, Anchorage, AK, 2017, pp. 941-948.

[2] X. Liu, 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.* ACM, 2014, pp. 457-466.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks" in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach", in *Computer Vision- EECV 2006.* Springer, 2006, pp. 288-301.

[5] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darell, "Decaf: A deep convolutional activation feature for generic visual recognition" *arXiv preprint arXiv: 1310.1531*, 2013.

[6] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-Column deep neural networks for image classification" *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

[7] Understanding Aesthetics with deep learning, *A journal from Nvidia,* *https://devblogs.nvidia.com/understanding-aesthetics-deep-learning/*

[8] Photography Classification in aesthetics, Lifei Chen, Pai Zhu, Carnegie Mellon University, 2016

[9] R. Arnheim, "Art and Visual Perception: A Psychology of the Creative Eye", University of California Press, Berkeley, 1974.