Learning Good Taste: Classifying Aesthetic Images

Prasanth Veerina
Stanford University
pveerina@stanford.edu

Abstract

When humans look at an image they not only understand the subject matter, they also make a number of subjective assessments about it such as aesthetic quality, emotional intensity, creativity, etc. In this work we attempt to classify images based on their aesthetic appeal. Much like image recognition, the features which comprise a beautiful image are difficult to describe and quantify by hand. Thus we use convolutional neural networks to automatically learn features and evaluate images. Our system is more accurate than previous systems on the same dataset.

1. Introduction

In most computer vision tasks we tend to think of images as digital dark matter, a rich but inaccessible medium to be mined for information. These typical tasks involve some sort of recognition (e.g objects, digits, poses, faces, etc.) or analysis/processing. However, when we as humans look at images we think about more than just their subject matter, we also see whether the images are visually interesting and aesthetically pleasing. Photography as a visual art is as much concerned with composition, texture, color, light and shadow as it is with the subject matter. Similar to the problem of object recognition, the features which comprise an aesthetically pleasing/displeasing image are difficult to completely describe or quantify.

A system that can automatically evaluate image aesthetics has many potential applications. For example, such a system could be used as a module in an information retrieval system to filter or curate high quality images. It could be incorporated into cameras and phones as a way to help people take better photos, or improve consumer camera features such burst mode photo selection and automatic retouching. Aesthetic quality also has a relationship to image popularity so it could be used in marketing and content creation settings (for example we could incorporate this into a system built to predict photo popularity on Instagram, 500px, or other photo sharing services).

Previous work in aesthetic evaluation of images has relied on generic image features, or hand crafted aesthetics-related features.Convolutional neural networks (ConvNet) have revolutionized the field of image classification by being able to learn complex features that can be used to create far more expressive representations of images than traditional hand-crafted features. In this work we will investigate using ConvNets to classify aesthetic images.

2. Related Work

Computer vision scientists have done a quite a bit of work in trying to quantify the aesthetic quality of images. Early approaches tried to hand-craft features based on photographic intuitions. They used low level features such as such as spatial distribution of edges, color histograms, blur [4] and high level features such as salient object detection, rule of thirds, depth of field [1] to build classifiers which performed quite strongly on their respective datasets.

In 2012 Murray et al. introduced the AVA dataset which is a large scale dataset of images with aesthetic ratings. In their original work they formulated a binary classification problem and established the experimental settings which we use in this work (described in sections 3.1 and 3.2). They trained an SVM with Fisher Vector signatures computed from SIFT descriptors which achieved a maximum of 67% accuracy [7].

Recent work done by Lu et al. showed state of the art aesthetic classification performance using convolutional neural networks [6]. Lu et al. used the AVA dataset and same experimental settings as [7] and was able to achieve a classification accuracy between 60.25% and 71.2% on this dataset using single column convolutional networks. Their most accurate architecture contained 4 convolutional layers and 2 fully connected layers. The convolutional layers each contained 64 kernels, of sizes 11, 5, 3 and 3 respectively and the fully connected layers contained 1000, and 256 neurons respectively. They achieved a maximum of 74.46% accuracy using a specialized dual column network where one column is the same as their original aesthetic classification network and the second column is used to classify image style which they then combine into a single aesthetic prediction.

3. Approach

3.1. Problem Formulation

In general, aesthetic evaluation on our dataset would be a regression problem where given an input image I we want to output an image score between 1 and 10. For our purposes, however, (and in order to be able to compare results with previous work), we will treat the problem as binary classification. Given and image I output "HIGH" or "LOW" to signify whether the image has high aesthetic quality (equivalent to $score > 5 + \delta$ in the general case where δ is some parameter we chose to reduce ambiguity) or low aesthetic quality ($score \leq 5 - \delta$ in the general case). In our experiments we have fixed $\delta = 0$

3.2. Dataset

We are using the recently developed aesthetic visual analysis (AVA) dataset [7], which contains roughly 250 thousand images of various sizes and subject matter along with aesthetic rating metadata. The images are collected from a photography competition website called digital photography challenge where users submit their photographs to various competitions and the photos are judged by other members of the site. Each photo in the dataset is rated on a scale of 1-10 based on its overall appeal for the given contest theme by on average of 200 site members. While there is certainly more to a beautiful photograph than purely visual attributes, (e.g. creative subject matter, emotional impact), we will use these user ratings as our ground truth. Since the images are collected from this contest format there is some control for subject matter variation since each contest has a theme such as boats or flowers which all submissions must relate to. We obtained most of the data (some links were broken) and separated out 200k training images (roughly 60k negative and 140k positive examples) and 20k test images (roughly 6k negative and 14k positive examples) as defined the AVA dataset specification[7]. In figure 1 we see the distribution of ratings in the dataset looks roughly Gaussian around a mean of 5.38.

3.3. Data preprocessing

As described in several of the earlier works on computational image aesthetics [4, 1, 6], there are both local and global characteristics that lead to an overall aesthetic effect. Local characteristics include noise, blur and contrast while global characteristics include composition features such as rule of thirds, foreground/background separation, depth of field etc. Since our raw input images are not of any set size or aspect ratio, we need to perform some preprocessing and data augmentation in order for our ConvNet to be able to properly pick up on these global and local characteristics. Based on the steps suggested in [6] we resize our images in two ways. The first way is by simply warp the image to be 256x256 without regard to aspect ratio simply downsampling to get the right size (see 2d). Since warping or down-sampling the image too aggressively may destroy some salient aesthetic features the second way was to take random 256x256 crops from the full size image, thus we lose no detail to down-sampling (see 2(b,c)).

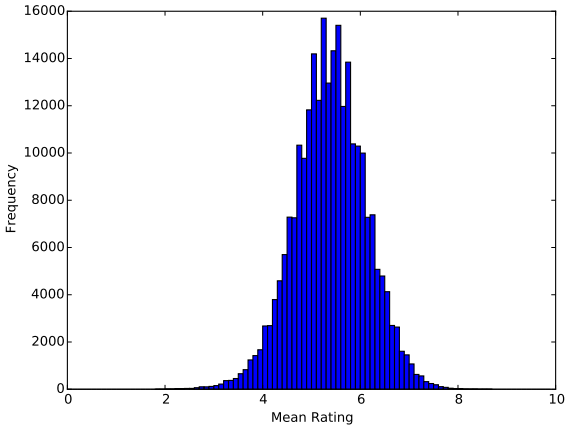


Figure 1: Distribution of Ratings

During training we augment the data further by first subtracting the mean image and taking random crops according to the input size of the network and applying horizontal mirroring.

3.4. Models

3.4.1 CaffeNet

We started with the CaffeNet model[3] which is a variation on the well known AlexNet architecture [5]. This model contains 5 convolutional layers and 3 fully connected layers and we replaced the final fully connected layer with a 2-neuron layer since our problem is binary classification. The network was initialized with ImageNet weights and all layers were finetuned (the learning rate on the initial layers was lower as not to completely disrupt the ImageNet weights). The network was trained over 100k iterations with batch size 100.

3.4.2 VGG

We then tried VGG16 the very deep network from Simonyan et al. [8]. This is a very powerful model which contains 13 Convolution, 3 Fully Connected layers. Once again we replaced the last layer with a 2-neuron layer since our problem is binary classification. The network was initialized from ImageNet and all layers were finetuned. This model proved more difficult to train than CaffeNet as the initialization was very easy to disrupt sending the train-

(a) Original (b) Random Crop 1 (c) Random Crop 2 (d) Warp

Figure 2: Example Resizing

ing loss to extremely high values after the first few iterations. We reduced the learning rate in order to prevent this. Furthermore the model is very computationally expensive to train, its memory requirements forced us to reduce batch size to 10. Due to time constraints the network was trained over 100k iterations. Because of these training challenges we believe this model can perform much better given slightly better hyperparameters and computational budget.

3.4.3 PReLU

Finally we tried replacing the the CaffeNet ReLU with a PReLU [2] which is a parameterized ReLU which learns a slope for inputs less than 0 instead of just thresholding them like ReLU. This model was trained for 60k iterations and the PReLU parameter was initialized to constant 0.1

4. Experiments & Results

4.1. Comparing Preprocessing Strategies

Warped	Random Crops
76.82%	74.10%

Table 1: CaffeNet accuracy on different resize strategies

The first experiment was to train the CaffeNet model using the squashed and random crop resize strategies (see section 3.3) and compare results. Table 1 shows the accuracies. We were slightly surprised with these results since we expected, based on the results of [6], that random image crops would outperform warped images. (Since much of the discriminatory power of ConvNets comes from local features). However we believe this may have occurred since we initialized our CaffeNet model with ImageNet weights while [6] trained from scratch. Since ImageNet input images are downsampled, not randomly cropped, there would be a difference in scale for the features that the ImageNet filters have been trained to pick out so naturally they wouldn't work as well on these non-down sampled crops.

4.2. Comparing Models

SVM [7]	ConvNet [6]	CaffeNet	PReLU	VGG16
67%	74.46%	76.82%	75.51%	77.07%

Table 2: Test Accuracies for various models

In this experiment we trained the CaffeNet, VGG16 and CaffeNet + PReLU models on the warped images. Table 2 shows the accuracies for each as well as the best reported accuracies on the same test set from [7] and [6]. We see that all three of our models beat previously reported results. There are two main differences in our approach. First the models are larger than those in [6] both in terms of number of layers and in terms of number of kernels in each layer, second our models are finetuned rather than trained from scratch.

Interestingly we see that PReLU actually performs worse than plain CaffeNet. This is likely due training issues. The weights we transfered from ImageNet were trained on a ReLU network but in PReLU network we initialized the parameter to a constant thus we had to reduce the learning rate to not destroy the weights, but this in turn made training much slower.

We see that VGG16 achieves the best performance, but that it is not very much higher than CaffeNet. This is almost certainly due to hyperparameter selection. As mentioned in the Model section, VGG16 proved quite sensitive to changes in learning rate but due to computational expense we were not able to do a thorough search.

4.3. Example Classifications

The following images were classified with the the CaffeNet model. Figure 3 shows a sample of images with the highest probability to be labeled as highly aesthetic. We see that all three of these images look like highly processed HDR photographs. They contain a lot of micro contrast, and smooth color gradients.

Figure 4 shows a sample of images with the highest probability to be labeled as not aesthetic. These images stand in stark contrast the the highly rated ones. They lack vivid color, are not as sharp and the middle image contains a lot of black pixels.

Figure 5 shows a selection of images which the classified misclassified. These three misclassification examples reveal the short comings of this ConvNet approach. In figure 5a we see that the classifier predicted that the image would be low quality while the ground truth label said it was highly aesthetic. This image is difficult because it is a very stylized still life. We clearly recognize the dramatic lighting and the sharp contrast of the fruit against the black background as creative stylistic choices, however the ConvNet has no such notion. In figure 5b we see the classifier classifies this images as low, but it is labeled as high. This image contains a lot of black pixels but it also has a mean rating of 5.005. It's a good representative of them many borderline or ambiguous images in our dataset. Human evaluators couldn't decide on this image and neither could our classifier. The third mistake is in figure 5c. We see that the classifier rated this image highly but on line it is rated as low. The ConvNet is likely influenced by the fact that the sky has a lot of contrast and color but it doesn't understand that the foreground is underexposed.

5. Conclusion

We have shown that convolutional neural networks can achieve very good results in binary classification of aesthetic images. In our experiments we have seen the effectiveness of transfer learning and the impressive generality of these models. We confirmed our intuition that the network will tend to look at local textural features and we saw the shortcomings of this since the network does not truly understand style, exposure and other photographic principles. We also found that mean rating is perhaps not the best way to measure and predict the aesthetic value of a photo since there is not much variance in our dataset. On clear next step would be to turn this from a binary classification problem into a regression problem and see how the model performs. However a more interesting future direction might be to try and predict distributions of ratings on images, such a system would capture more of the divisions in people's taste. Another interesting direction would be to incorporate more metadata so that a classifier can have context that an image is for example, a certain style. Finally it might also be interesting to add some global photographic features, such as rule of thirds, into the classifier so that it might complement the local textural features of the ConvNet.

References

[1] S. Dhar, V. Ordonez, and T. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1657–1664, June 2011.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *ArXiv e-prints*, Feb. 2015.

[3] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[4] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *in IEEE Computer Society Conference on Computer Vision and Pattern Recognition,2006,vol.1,pp.419426*, pages 419–426, 2006.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

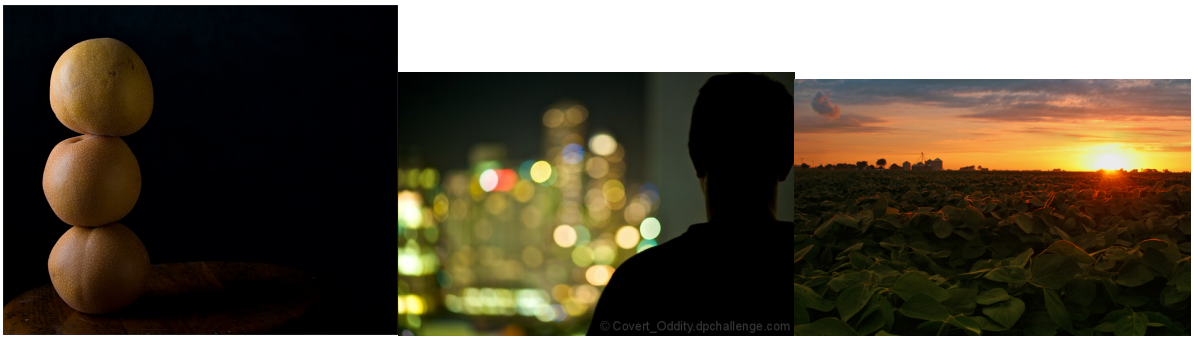
[6] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the ACM International Conference on Multimedia*, MM '14, pages 457–466, New York, NY, USA, 2014. ACM.

[7] 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, June 2012.

[8] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

Figure 3: Sample of most likely to be highly aesthetic

Figure 4: Sample of most likely to be low aesthetic



(a) Predicted=0, Label=1 (b) Predicted=0, Label=1 (c) Predicted=1, Label=0

Figure 5: Sample of misclassified images