Food Image Aesthetic Quality Measurement by Distribution Prediction

Jiayu Lou(jiayul@stanford.edu), Hang Yang(hyang63@stanford.edu)

Abstract

We built a Food Image Aesthetic Quality Classification model based on VGG16, ResNet, and Google's NIMA (Neural Image Assessment), aiming to serve as a better alternative of rating food photos' attractiveness to Yelp's published approach that utilized EXIF data in training sets. With an image of food as input, we are able to predict with 72% accuracy whether a group of people will rate it as a "high" or "low" quality food photo, with an estimated distribution of how many % of people will vote 1 through 10 if they are asked to rate the picture.

Introduction

Today we observe millions of food photos uploaded to all kinds of sites. Some are aesthetically pleasing, and some are not. Websites like Yelp rely heavily on the high quality photos to attract users to restaurants, generate demand for its usage, and ultimately profit from user activities. Users are always bombarded by food pictures of varying qualities posed by restaurants, professional photographers, bloggers, and food-loving people, which compels platforms like Yelp to locate attractive and high-quality food pictures and present the food in their best interests. Our model serves this demand: taking a food photo as input, it will output a probability distribution of ratings from 1 to 10, which will then be used to predict its quality in terms of "high" or "low."

Related Work

In the past, due to the subjective nature of metrics such as attractiveness and lack of useful training sets, there exist only a limited number of meaningful deep learning projects for the purpose of determining whether food images look attractive and delicious. Yelp's published approach [1] that utilized EXIF data in training sets and binary outcomes of "good" or "bad" seems a little outdated, and there isn't much else published that particularly addresses this interesting yet challenging problem.

There, however, exists a number of prior work that deal with generalized image aesthetic quality assessments. Murray et al. [2] is the benchmark on aesthetic assessment, showing significant improvement to previous works based on hand-crafted features. They introduce the AVA dataset and propose a technique to use manually designed features for style classification. Lu et al. [3] show that deep CNNs, such as their double-column CNN consisting of four convolutional and two fully-connected layers, are well suited to the aesthetic assessment task. Similar to Murray et al. [2], their images are also classified to low and high aesthetics based on mean human ratings. Other researchs use regression loss to learn the human ratings of the AVA dataset, such as the AlexNet inspired model in [4] and the fine-tuned VGG network in [5]. More recent papers discuss alternative advanced techniques. [6] uses an adaptive spatial pooling and a multi-net approach with each network being a pre-trained VGG. Ma et al. [7] propose using a saliency map to select patches with highest impact on predicted aesthetic score. However, none of these methods report correlation between their predictions and the ground truth ratings. Recently, Kong et al. in [8] train an AlexNet-based CNN with a rank-based loss function to learn the differences between two input images' aesthetic scores and indirectly optimize for rank correlation. From there, Google's NIMA (Neural Image Assessment) [9] further achieves higher correlation with human ratings, predicting the distribution of ratings as a histogram with normalized Earth Mover's Distance as its loss function.

Dataset

We are using two datasets for the purpose of this project. One is the AVA database [2], containing more than 250,000 photos of various topics with each scored by an average of 200 people in response to photography contests on a scale of 1 to 10. The other one is the Food-101 database (https://www.vision.ee.ethz.ch/datasets_extra/food-101/static/bossard_eccv14_food-101.pdf), containing 101,000 food photos of varying qualities, without quality scores.

Due to limitation in time and computational resources, we decide to use a normal training set of 40,000 photos in various topics with a validation set of 2,000 food-only photos, and a specific training set of 6,000 food-only photos with a validation set of 1,181 food-only photos, so we can explore whether specification in food photos during training will have a positive impact on performance.

Example of inputs:

953730.jpg

957890.jpg

ImageID	# of 1	# of 2	# of 3	# of 4	# of 5	# of 6	# of 7	# of 8	# of 9	# of 10
953730	0	0	3	4	32	48	20	12	4	3
957890	0	1	2	15	43	42	11	3	4	2

For preprocessing, we employ the standard practice of resizing the short edge to 256 pixels, random crop of 224x224 for training, center crop of 224x224 for validation, and normalization for the image input pixels so that each pixel is between 0 and 1. For the corresponding labels, we divide them by the sum to obtain a probability distribution summing up to 1 as follows:

ImageID	# of 1	# of 2	# of 3	# of 4	# of 5	# of 6	# of 7	# of 8	# of 9	# of 10
953730	0.0000	0.0000	0.0238	0.0317	0.2540	0.3810	0.1587	0.0952	0.0317	0.0238
957890	0.0000	0.0081	0.0163	0.1220	0.3496	0.3415	0.0894	0.0244	0.0325	0.0163

So that we match the prediction from soft-max activation function that we are going to use following the fully connected output layer.

Methods

We build on Google's NIMA to predict a distribution of human opinion scores. Specifically, we replace the last layer of baseline CNN with a fully connected layer of 10 neurons followed by soft-max activations. The proposed method has reached state-of-the-art performance for aesthetic qualities of images. We would like to apply the model specifically on food images with our optimizations, using VGG16 and ResNet-50 as our base network structures.

We are using the Earth Mover's loss function as our loss function, which is defined as the following:

$$\mathrm{EMD}(\mathbf{p}, \widehat{\mathbf{p}}) = \left(\frac{1}{N} \sum_{k=1}^{N} |\mathrm{CDF}_{\mathbf{p}}(k) - \mathrm{CDF}_{\widehat{\mathbf{p}}}(k)|^{r} \right)^{1/r}$$

where CDF stands for the cumulative distribution function for each score class. It has been shown that for ordered classes, the classification frameworks can outperform regression models, and training on datasets with intrinsic ordering between classes can benefit from EMD based losses. Since our predicted labels are fed into soft-max function in the last layer, we have ensured that the probabilities for each score class add up to 1, which makes calculating CDF meaningful.

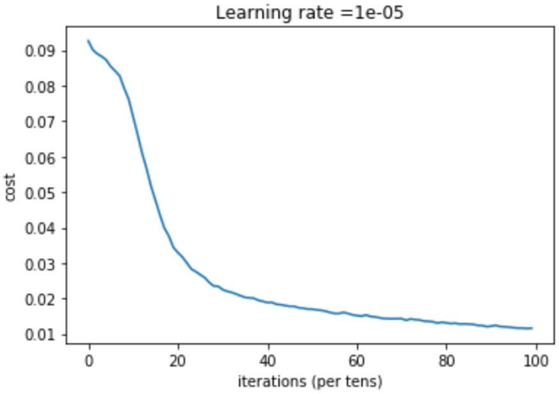
To measure the performance of our model, we plan to follow Google's binary classification. Specifically, we would attach "high quality" or "low quality" tags to each image based on if their predicted mean quality score exceeds 5, where the mean quality score is defined as:

$$\mu = \textstyle\sum_{i=1}^{N} s_i \times p_{s_i}$$

We would then calculate the metrics of the model by comparing if the predicted binary tag matches the ground truth binary tag (computed from ground truth mean quality score). In addition, we feed images from the Food-101 dataset into the trained models, to see how well the models are performing in the real-world context with no prior scoring distributions given.

Experiments & Results

We experimented with the GPU capacity of our AWS EC2 Deep Learning Environment, and found that it could accommodate a max mini-batch size of 64 without running out of memory, so that's the size we chose to maximize efficiency. To find an appropriate learning rate, we started training and validating with a smaller subset of the AVA dataset, with 2500 training examples and 250 validating examples. Our results favored a learning rate of 1e-5 for both VGG16 and ResNet-50 backed networks, with the loss decreasing as follows:



Model	# Train/Val	Train Acc	Val Acc
VGG16	2500/250	96.8%	71.6%
	40000/2000	89.0%	71.8%
ResNet-50	2500/250	90.3%	68.4%
	40000/2000	81.7%	71.5%
VGG16 Food Only	6000/1181	92.1%	70.9%
ResNet-50 Food Only	6000/1181	85.3%	70.4%

We also found that the variance, indicated by the difference between training accuracy and validation accuracy, was quite large. After initial regularization efforts such as adding Dropout layers, L2 regularization, weight decay, etc. failed, we decided that maybe training on a larger dataset would make it harder to overfit to the training set and therefore reduce variance. Our best results are listed above.

The results, while satisfying to us, could use some improvement. The variances are still large, as seen in the large gap between training and validating accuracies, despite our best efforts to train on more data and to add more regularizations. This overfit to the training set is likely due to our having insufficient time to change the model around, retrain on the large dataset, and find the best model with both low bias and low variance. With regards to confusion matrices, we have for ResNet-50 40000/2000 and ResNet-50 Food Only 6000/1181:

ResNet-50 40000/2000

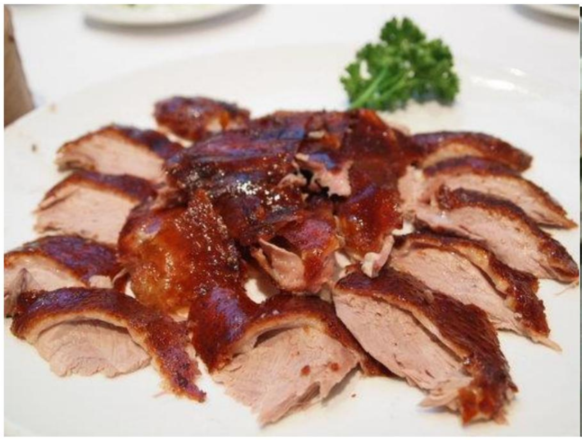
		Predicted	
		High	Low
Ground Truth	High	72.2%	27.7%
	Low	29.1%	70.9%

ResNet-50 Food Only 6000/1181

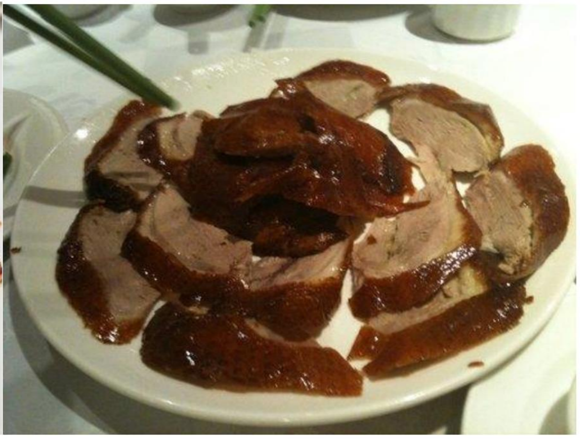
		Predicted	
		High	Low
Ground Truth	High	72.7%	27.3%
	Low	28.7%	71.3%

The VGG16 results are similar. The performances of different models in terms of overall accuracies, false positive rates, and false negative rates don't differ much, which is a slight surprise to us as food-only training specification didn't help as much as we envisioned. It would appear that photos' aesthetic qualities are more generalized than we previously thought. Nevertheless, the model predicts fairly well when applied to the Food-101 dataset, examples of the results:

High



Low



Since images in Food-101 dataset don’t have human scores associated with them, our team takes subjective judgment on whether they are correct classifications. It would seem that the false positive rate (rating High when it should be Low) is higher than the false negative rate, and it would seem to be perhaps a result of the training AVA dataset containing images all from photography competitions, possibly biasing the classifier to consider more images as high quality.

Conclusion & Future Work

We believe our classifier did a satisfactory job of predicting the food image aesthetic qualities and classifying food images to high or low quality categories with 72% accuracy compared to human ratings. Both models are performing at about equal level, with VGG16 overfitting more to the training set, and therefore potentially has more room for improvement in the bias-variance tradeoff. If we are to continue working on this project, our first direction would be to further reduce variance and improve validation accuracy by means of training on more data and adding more regularization. We would also like to experiment with other network structures and loss functions, and also employ our model to improve food image qualities based on the features it has learned.

Team Members & Contributions

Jiayu Lou(jiayul@stanford.edu): Data Import, Preprocessing, Model Building, Training, and Hyperparameter Tuning
Hang Yang(hyang63@stanford.edu): Instance & Environment Setup, Data Preprocessing, Model Building, Training, and Hyperparameter Tuning

References

[1] M., A. (2018). *Finding Beautiful Yelp Photos Using Deep Learning*. [online] Engineeringblog.yelp.com. Available at: https://engineeringblog.yelp.com/2016/11/finding-beautiful-yelp-photos-using-deep-learning.html.

[2] 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. IEEE, 2012, pp. 2408– 2415. 1, 2, 3, 7, 8, 9, 10

[3] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rating image aesthetics using deep learning," IEEE Transactions on Multimedia, vol. 17, no. 11, pp. 2021–2034, 2015. 1, 7

[4] Y. Kao, C. Wang, and K. Huang, "Visual aesthetic quality assessment with a regression model," in Image Processing (ICIP), 2015 IEEE International Conference on. IEEE, 2015, pp. 1583–1587. 1, 7

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014. 1, 3

[6] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 497–506. 1, 7

[7] S. Ma, J. Liu, and C. W. Chen, "A-lamp: Adaptive layout-aware multipatch deep convolutional neural network for photo aesthetic assessment," in Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. IEEE, 2017. 1, 6, 7

[8] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in European Conference on Computer Vision. Springer, 2016, pp. 662–679. 1, 2, 6, 7

[9] H. Talebi and P. Milanfar. 2017. NIMA: Neural Image Assessment. In arXiv:1709.05424.