Food Image Aesthetic Quality Measurement by Distribution Prediction

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

Predicting

Today we observe millions of food photos uploaded to all kinds of sites, with some aesthetically pleasing, and some not so much. 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.

We built a Food Image Aesthetic Quality Classification model based on Google's NIMA (Neural Image Assessment), aiming to serve as the better alternative 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.

Data

We are using the AVA database containing 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, and the Food-101 database containing 101,000 food photos of varying qualities, without quality scores.

Models & Results

I. NIMA Model with VGG16

Our initial model follows Google's NIMA implementation. Specifically, we replaced the last layer of VGG16 with a fully connected layer of 10 neurons followed by softmax activation in order to achieve a probability distribution over 10 classes (score of 1 to 10).

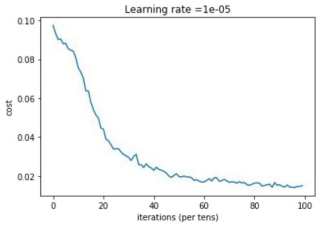
We then used the Earth Mover's Distance as our loss function as follows:

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

where CDF stands for the cumulative distribution function for each score class. Google's paper has 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.

II. Preliminary Results

With "high" or "low" tags to each image based on whether its predicted mean quality score exceeds 5, we trained and tested on a subset of 2500 images for training and 250 food images for validation, with 96.8% train and 71.6% val accuracies, loss as follows:



III. Additional Models

We then moved to experiment with more models, and since Google's Paper did not include ResNet as part of the discussion, we chose to allocate most of our time on ResNet-50, tuning hyperparameters, adding regularizations, and training and validating with larger amount of data.

We also trained on only the images with the "food & drinks" tags within AVA dataset, with the expectation that specification during training might help our objective.

IV. Final Results

Our final results are 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%

From Food-101:

High Low

Discussion

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 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.

The performances of different models in terms of validation accuracies don't differ much, which is a slight surprise to us as food-only training specification didn't help. It seems that photos' aesthetic qualities are more generalized than we previously thought.

Future

If we are to continue working on this project, our first direction would be to further reduce variance and improve val acc by means of training on more data and more regularization. We would also like to experiment with other network structures and loss functions, hopefully with more time and computing resource.

Reference

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