Yelp Restaurant Photo Classification

Pradeep Kalipatnapu
SID:26963490
University of California, Berkeley
Computer Science
prad@berkeley.edu

Sung-Li Chiang
SID:24978880
University of California, Berkeley
Mechanical Engineering
slchiang@berkeley.edu

Zhuosi Wang
SID:26967290
University of California, Berkeley
Computer Science
zhuosi.wang@berkeley.edu

Abstract

In this project we showed the ability of deep convolutional neural networks to classify a multilabel dataset. Furthermore, in our problem, the data instances themselves are not classified, but sets of data instances are. For example, a restaurant may have dozens of pictures, and is classified with binary attributes. However, the individual images themselves are not labelled with a ground truth.

Our approach involved propagating set labels to individual data instances and demonstrating neural networks ability to learn impressively on the resulting noisy dataset.

1. Introduction

Yelp has been a very popular application recently, we use it to search a restaurant for any events such as casual dinner, date, friends reunion. For different situations, we need different types of restaurants to satisfy our requirement. For example, we need a cheap and a meal which is good for dinner if we are looking for a restaurant for single person; a romantic restaurant for a wonderful dating night, and a restaurant where is able to accommodate tens of people and maybe drinks for a reunion of high school classmates.

Fortunately, there are tens of thousands of pictures and descriptions of restaurants. For each restaurant, there are photos including delicious meals, restaurants environment and menu so that you are able to pick a restaurant you like if meals look fancy and delicious and environment is classic. Photographs on Yelp are extremely popular for figuring out which dishes to order or to confirm if the restaurant looks good. On the other hand, the other handy function is 'business info , From attributes of 'business info , it provides you information of things like if they accept table reservation of table services, etc. Many times after visiting a restaurant you get a survey for Yelp asking about the attributes. This is probably the least popular thing about Yelp. Being able to resolve this automatically could be more accurate, and also lead to a better user experience.

In this project, we are more curious about that if it is possible to learn all these message from images. The challenges are making a multi-label classification and learning implicit message from given images given only businesses features. For example, giving an image with a delicious plate of food on it and restaurants features as 'good-for-lunch , 'good-for-kids , 'outdoor-seating , how could we know the corresponding features of the restaurants with similar images.

2. Data Description

2.1. Data Source

We download image data from Yelp Restaurant Photo Classification competition on Kaggle[3]. There are about 7 Gb/ 10k images for training and testing data set. Images belong to different restaurant businesses, and there are different number of images for different businesses. The size of each image is of 500 * 300 or 300 * 500. For each set of images from the same restaurant, we have the same feature attributes of the restaurant instead of the specific meaning for each image, respectively. The attributes are:

1. good-for-lunch
2. good-for-dinner

3. takes-reservations
4. outdoor-seating
5. restaurant-is-expensive
6. has-alcohol
7. has-table-service
8. ambiance-is-classy
9. good-for-kids

In this image data set, image information is actually noisy due to the duplicates uploaded accidentally by users and the same photos used for different branches of chain businesses.

For estimation of our work, we also test on images, which are from restaurants near Berkeley campus and not included in the training or testing data from Yelp.

2.2. Data Preprocessing

Since we were training on individual images, but were labelling businesses, computing a loss function and actually training the net were not straightforward. In order to make use of inbuilt caffe loss functions, we transferred labels from restaurants to images, However, as a negative, this results in the creation of an even noisier dataset. For example, some pictures of a bar may be pictures of food or people, but the picture will always have the 'has-alcolhol attribute.

While this labelling mechanism is not ideal, it is simple and as we shall show later quite effective. The lack of a manual data entry process also makes our analysis reproducible.

We also stored the data in the form of an LMDB for faster access. Our test set was pretty large, and by moving to lmdb, we reduced the run time from 6.5 hours to 2 hours.

We also random cropped our images to $227 * 227$ and used horizontal mirroring as means of data augmentation. We also applied simple transformations such as subtracting the mean for each channel.

3. Model Description

3.1. Model Training

For training our model, we modified a Caffe Pretrained Alex net model. Since we had 9 attributes to predict, we used 9 nodes in the output layer. The other layer weights were adopted from the pretrained model. The last layer used sigmoid non-linearity, and we used sigmoid cross entropy loss to train our model. We used a batch size of 128 images. The images were shuffled, and images from one restaurant were not grouped together.

The model is fine tuned over 4 epochs of the training data set. We saw a mean score of 0.8075 on the training data.

3.2. Model Testing

We set aside 20% of businesses from the training data set for validation purposes. Each of these businesses has about $50-300$ images. We used our model above to predict the scores for these businesses and achieved an F1 score of 0.78.

3.3. Evaluation Metric

Since it is a multi-label output, we used the mean F1-score[2] as an evaluation metric, which measures the accuracy by using the statistic precision and recall. The score is defined as below:

$$F1 = 2\frac{pr}{p+r} \tag{1}$$

where p is statistic precision and f is recall. Also,

$$p = \frac{tp}{tp+fp} \tag{2}$$

$$r = \frac{tp}{tp+fn} \tag{3}$$

where, tp is true possitive, $tp+fp$ is all predicted positives and $tp+fn$ is all actual positives.

3.4. Prediction Model

After the above model was trained, we were able to extract predictions for each image for every attribute. However, due to the large amount of noise in the data, the positive and negative classes were not well separated. Consider the attribute 'good-for-lunch', which is generally defined as quick, take out food such as pizza,sandwiches or burgers. The x-axis in Figure 1 is the score assigned by our model

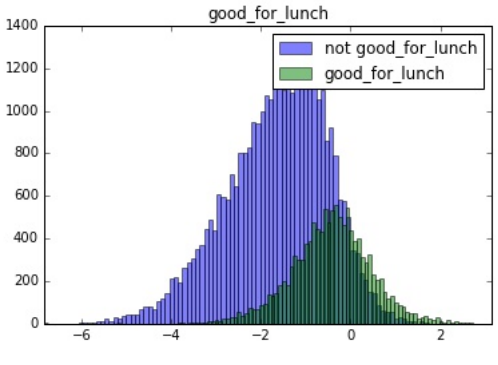


Figure 1: Good for lunch distribution

for the 'good-for-lunch' attribute. The purple bars represent images that come from restaurants that are not good

for lunch, and the green are images that from businesses that are good for lunch. However, it is clear in Figure 2 that the two distributions have different means. Based on this observation, we grouped all the images of a single restaurant and computed the mean score the for the business. The following Figure 2 shows this. Even though there is some overlap

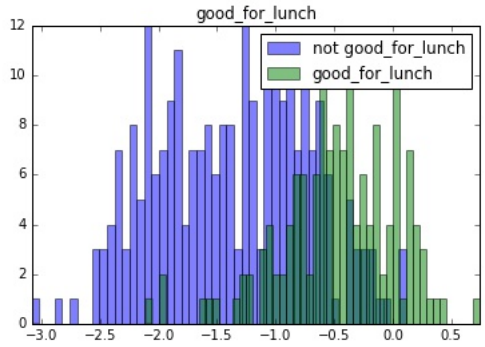


Figure 2: Good for lunch business distribution

between the two classes, we can compute the bayes' boundary here of approximately -0.6. Restaurants with a score above this value are likely to be good for lunch. We computed the bayes' optimal decision boundary on the training data for each attribute and achieved an F1 score of 0.788 on the validation data.

4. Results

4.1. Best Images for each category

The following, Figure 3, are the images from our validation data with highest scores in each category. As you can see, the images are quite representative of their categories. High correlation between some attributes such as 'is-expensive', 'takes-reservations', can also be seen by the images repeating across these categories.

4.2. Attributes

Among the attributes, 'has-alcohol, Figure 4 and 'takes-reservations', Figure 5, were the best performing attributes.

However, our model struggled with has outdoor seating, Figure 6, barely learning anything at all. Most images are taken indoors in a restaurant, and even some outdoor images are taken at night. So we feel that the model would have struggled to learn based on brightness. In fact, one of our top 5 images for outdoors, was incorrect. The top 3 images had meshed tables that you would find outdoors, which is probably a better indicator to the mode about outdoor seating.

On the contrary, Figure 7 and Figure 8 show good results for classifying images of 'restaurant-is-expensive' and

(a) Highest score image for 'have-alcohol' (b) Highest score image for 'good-for-lunch'

(c) Highest score image for 'takes-reservations' (d) Highest score image for 'outdoor-seating'

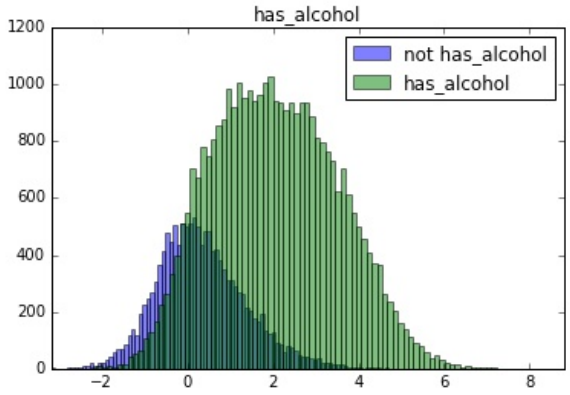
(e) Highest score image for 'restaurant-is-expensive' (f) Highest score image for 'good-for-kids'

Figure 3: Highest score images for each attribute

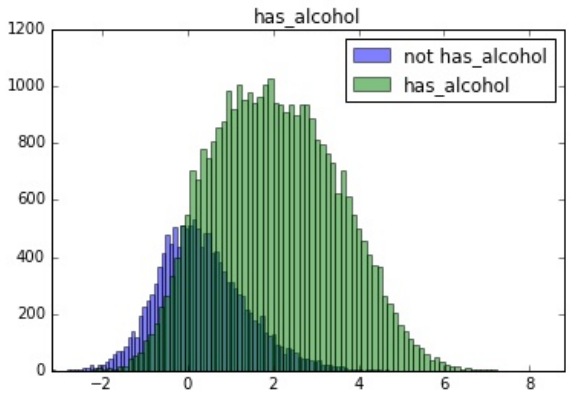
the restaurant 'takes-reservation'. For each figure, there are top five scores of images classified as the corresponding attribute. GT means the ground truth and EST means the estimation result.

4.3. Information combination of images

In Figure 9, it shows the prediction results and the ground truth of images' attributes. An interesting thing is that even though predictions are not 100% accurate, some other attributes it predicts is still making a lot of sense. For example, in Figure 9b, 'good-for-kids' is predicted but not in ground truth; however, from images, it seems really a decent and safe place for kids to be there. Also, amazingly, the classifier predicted entirely correct for Figure 9e .

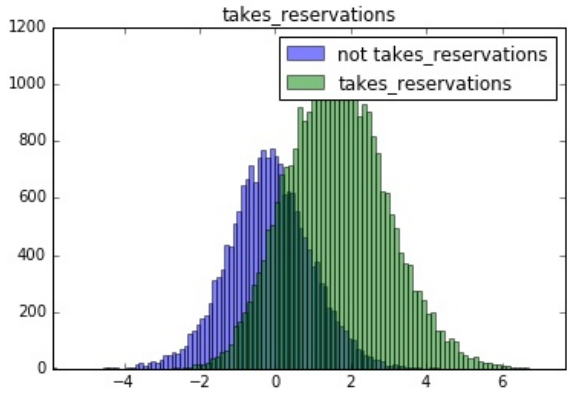


(a) has-alcohol image distribution

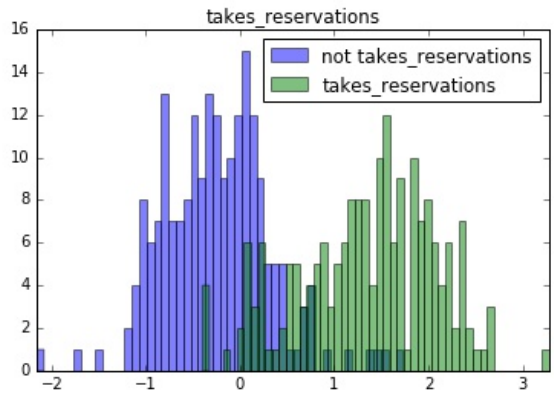


(b) has-alcohol business distribution

Figure 4: Distribution of images are 'has-alcohol' and businesses 'has-alcohol'



(a) Distribution of images are 'takes-reservations'



(b) Distribution of businesses are 'takes-reservations'

Figure 5: Distribution of images are 'takes-reservations' and businesses 'takes-reservations'

5. Conclusion

Overall, we were able to show CNN learning using set level attributes, as opposed to instance level attributes. We were also able to achieve high levels of accuracy on attributes other outdoor seating, and it is clear that computer vision could be a good solution in this field. We also managed to apply our model unchanged to restaurants specifically in Berkeley, Figure 10, with good results, especially was amazingly able to tell expensive restaurant, and thus demonstrated the portability of our model.

6. Future Work

Since most images from a restaurant contain one major object such as a dish or a cup of drink for a single image, the usage of a long-short term memory cell to deal with a batch of images from the same restaurant may be able to catch a better description of the type of a restaurant. However, the other challenge for this idea is that images chosen in a batch may have to be with enough diversity; otherwise, if all images are pizza, for example, but with different angle or ambient brightness, this may do nothing useful help with improving prediction. Moreover, the images don't form a sequence, rather they form a set. This is slightly different from current use cases of short term memory.

7. Reference

1. Github https://github.com/pkalipatnapu/YelpVisionProject
2. F1 score description https://www.kaggle.com /wiki/MeanFScore
3. Kaggle Competition https://www.kaggle.com/c/yelp-restaurant-photo-classification

(a) Highest score image for ‘outdoor-seating’

(b) Second high score image for ‘outdoor-seating’

(c) Third high score image for ‘outdoor-seating’

(d) Fourth high score image for ‘outdoor-seating’

(e) Fifth high score image for ‘outdoor-seating’

Figure 6: Top 5 high score image for ‘outdoor-seating’.

(a) Highest score image for ‘restaurant-is-expensive’

(b) Second high score image for ‘restaurant-is-expensive’

(c) Third high score image for ‘restaurant-is-expensive’

(d) Fourth high score image for ‘restaurant-is-expensive’

(e) Fifth high score image for ‘restaurant-is-expensive’

Figure 7: Top 5 high score images for ‘restaurant-is-expensive’

(a) Highest score image ‘takes-reservations’

(b) Second high score image for ‘takes-reservations’

(c) Third high score image for ‘takes-reservations’

(d) Fourth high score image for ‘takes-reservations’

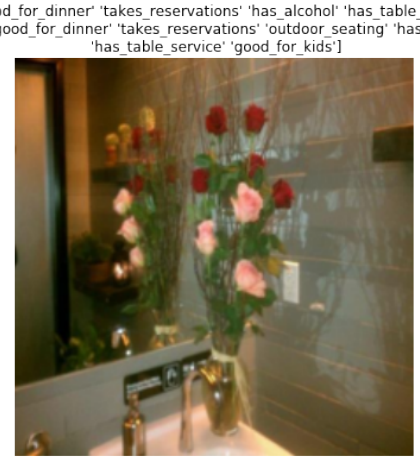
(e) Fifth high score image for ‘takes-reservations’

Figure 8: Top 5 high score images for ‘takes-reservations’.

(a) Comparison 1

(b) Comparison 2

(c) Comparison 3



(d) Comparison 4

(e) Comparison 5

Figure 9: Image results of comparison of prediction and ground truth. GT is the ground truth and EST is the estimation of attributes.

(a) Nefeli. Estimate:{'good-for-dinner', 'takes-reservations', 'outdoor-seating', 'has-alcohol', 'has-table-service', 'good-for-kids'}

(b) Chez Panisse. Estimation:{'good-for-dinner', 'takes-reservations', 'restaurant-is-expensive', 'has-alcohol', 'has-table-services', 'ambiance-is-classy' }

Figure 10: Restaurants nearby Berkeley campus.