Yelp Photo Classification by Using Convolutional Neural Networks

Xiaoman Dong, Yigao Li, Taoran Yu
Georgetown University

Abstract

Yelp becomes a popular restaurant review application in the world. There are tons of user photos uploaded every day. Currently Yelp is an user-submitted environment that users can only upload with photos and caption. However, this project aims to improve the photo classification and automatically create tags or labels for the uploaded image. To be specific, each uploaded photo is labeled as one of the five categories: food, drink, menu, restaurant interior decoration or storefront. The project tried several deep learning models to train the dataset and comes out that the multi-layer perceptron achieving at the best performance with testing accuracy 87.4%

1 Introduction

This project is about photo classification by using Yelp restaurant photos. Ideally the purpose of the projects is to create recommended labels for each uploaded photo from users. We run the baseline as Multi-Layer Perceptron. Moreover, we trained dataset with several complicated neural network models and run our models on Google Colab

2 Related Work

The first model of feature extraction dataset is building from a Multi-Layer Perceptron as a baseline and achieve accuracy about 87.4%. Then we apply other machine learning methods such as CNN-1D (Convolutional Neural Network in one dimension), combination of CNN and KNN(K Nearest Neighbor), VGG16, VGG19, RESNET and Inception V3 from Google.

3 Dataset

The dataset is downloaded directly from Yelp official website with 280,992 images. Images have attributes such as caption, photo_id, business_id and label. Also, Yelp official website provides an extracted feature vector csv dataset, by using a convolutional neural network trained for photo classification for programmers to train the machine learning model. The feature extraction dataset provides 100 features mapping to all the images from original dataset.

The figure 1 below shows a biased distribution of data with five labels. Food data contributes about 66% of whole dataset. However, the menu dataset only counts 1% of complete dataset. Figure 2 shows several sample photos.

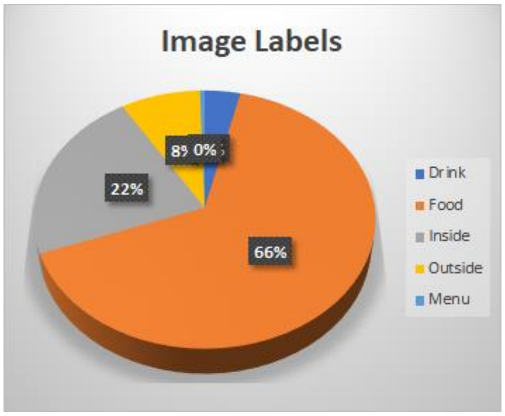

Figure 1. Photo distribution

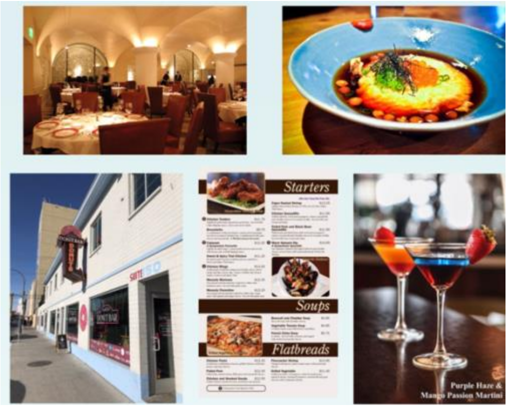

Figure 2. Sample photo

4 Methods

4.1 Train Model

First the model was trained with 48% of all images, validation dataset with 32% (since there are 100 features as the result of CNN, we want to have more validation data) of all images and 20% of all dataset as testing set.

Next, we run baseline model with 100 nodes as input layer, 512 nodes as hidden layer and 5 nodes as output layer. Then we apply “ReLU” activation function from input layer to hidden layer and use “softmax” activation function from hidden layer to output layer. Then using categorial cross-entropy softmax objective function and adadelta optimization with testing accuracy 87.4%. After the training step, we want to see how well the training model can be performed on the testing set and construct a confusion matrix heatmap as below:

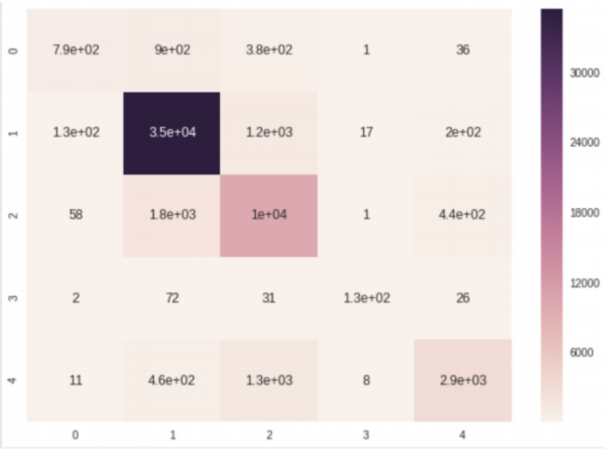

Figure 3. Confusion matrix heatmap on testing set

4.2 Trying other architectures

Overall we trained six models after baseline. They are CNN in 1 dimension, combination of CNN and KNN using CNN-1D as base, CNN using pre-trained model, for example VGG16, VGG19, RESNET, and Google’s Inception V3

The CNN was buiding with two convolutional layers, the first layer including 128 output filters, kernel size of 3, activation function as ReLU. Then applying the batch normalization, and add a maxpooling layer with pool size of 2. After we add the second convolutional layer including 256 filters, kernel size of 3, and activation function as ReLU. Then we did batch normalization again, and add a max pooling layer with pool size 2, flatten the layer from last step and put them into a dense hidden layer with 1024 nodes, then apply a drop out rate of 0.6, and finally put the result from last step into a dense layer with 5 output nodes and use softmax activation function.

Finally we train the model with 50 epochs and 2048 batches, achieving the test accuracy at 86.47%.

5 Results

Overall the performance of each trained model are acceptable. The result of testing accuracy with each model are listed in a table below:

Model	Validation Accuracy	Testing Accuracy
MLP(baseline)	87.8%	**87.4%**
CNN-1D	87.8%	86.47%
CNN+KNN	N/A	85.1%
VGG16	85.5%	84.62%
VGG19	85.66%	84.96%
RESNET	85.5%	84.78%
Inception V3	81.84%	81.61%

From above, it is clear to see that the Multi-Layer Perceptron achieved the highest accuracy rate at 87.4%, including input layer of 512 nodes, and ReLU activation function, then have a dropout rate of 0.7, and then have a dense layer of 5 output nodes, with activation function softmax.

6 Discussion

6.1 Accuracy Results

In general, the baseline model Multi-Layer Perceptron achieved the highest accuracy at 87.4%.

From the result table above, the best performance of all pre-trained model is VGG19 at test accuracy 84.96%.

Also, from the result table, the combination of KNN and CNN doesn’t increase the accuracy of model. However, the possible reason that the accuracy of those models are even worse than the baseline is they use extracted feature as input, hence it may cause information lost

6.2 Misclassification

From confusion matrix heatmap as shown in Figure 3, we can clearly see that there exists large number of misclassifications between food and drink. For testing set with label “Drink”, the baseline model would predict the drink with a food label wrongly even for the majority of the set.

For testing set with label “Outside”, it also reveals a significant misclassification. The model would misclassify the outside images as inside at a large proportion of 28%

Also, images with label “Menu” would mistake recognized as food at a large percentage

6.3 Ongoing Work

When users upload photo via Yelp, they can only enter the caption rather than selecting labels manually. Therefore it comes with an possible problem that their classification dataset may exist confliction with user’s truly selection. We also think the outcome of testing accuracy could be better if the Yelp could provide the label selection clicker to user and put all data which matching both user’s selection and current training model as the input training model.

In the future, if this project would have improvements, we would apply those pre-trained models with raw images directly. However, by trying those pre-trained models we find out VGG19 achieve at the highest accuracy rate comparing to other pre-trained models. Therefore, if this project was able to continue, we would use raw images as input for VGG19

Also, the dataset itself is very biased, since the size of menu and outside dataset are relatively small comparing with other labeled images.

7 Conclusion

Overall, this project achieved the highest accuracy of 87.4% with Multi-Layer Perceptron on classifying 5 labels.

If this project was able to continue, we would sample images of equal size of each label because we see biased image dataset lead to a significant misclassification, and use these raw images as input to train CNN with pre-trained model VGG 19 since we already see that in this project VGG19 performs best among all pre-trained models.

8 Reference

[1] Agrawal, Pulkit, and Raghav Gupta. "MIML Learning with CNNs: Yelp Restaurant Photo Classification."

[2] “Yelp restaurant photo classification,” https://www.kaggle. com/c/yelp-restaurant-photo-classification.

[3] *Shang, https://github.com/bzshang/yelp-photo-classification.*

.