Yelp Restaurant Photo Classification

Lakshmi Vaidiyanathan [*1] **Vinothini Pushparaja** [*1] **Robert Finn** [1]

Abstract

The Yelp Restaurant Photo Classification challenge is a Kaggle challenge that focuses on the problem predicting user labels of restaurants based on user review photographs. This project approached the problem with the Resnet-152 convolutional neural network architecture with transfer learning from a trained Resnet-152 followed by classification Support vector machine and Random Forest to achieve a score F1 of 0.80, that is close to the highest score of 0.82 achieved by the winner.

1. Introduction

Yelp.com and Yelp mobile App publishes crowd-sourced reviews about businesses. The Yelp Restaurant Classification Challenge (YRCC) is a challenge hosted by a data science community known as Kaggle in collaboration with Yelp. The goal of the YRCC is to build a model that automatically labels restaurants with multiple categories based on user-submitted photos. A restaurant can be associated with one or more of the following nine labels or attributes:

0 - good_for_lunch
1 - good_for_dinner
2 - takes_reservations
3 - outdoor_seating
4 - restaurant_is_expensive
5 - has_alcohol
6 - has_table_service
7 - ambiance_is_classy
8 - good_for_kids

The Data consisted of 2000 restaurants with labeled train set and around 230,000 images. The unlabeled test set consisted of 10,000 restaurants with approximately 1.2 million images , with one or more of the labels applicable.

[*]Equal contribution [1]Saint Peter's University, New Jersey, USA. Correspondence to: Lakshmi Vaidiyanathan <lvaidiyanathan@saintpeters.edu>, Vinothini Pushparaja <vpushparaja@saintpeters.edu>, Robert Finn <rfinn@saintpeters.edu>.

DS-670 Capstone: Big Data and Business Analytics, Saint Peter's University, New Jersey, USA.

There are two aspects to this problem that makes it more interesting and challenging :

1) The number of photos for each restaurant varies, thus model input also varies and is not constant.

2) Each image can be linked to more than one label, as a user can tag an image for label good for kids, has table service and take reservations at the same time.

As these lead to a multi-output binary classification problem where the restaurants tagged with 9 labels, where each label falls in to two categories: Applicable and Not_applicable.

The input for training the model was the JPEG format images submitted by Yelp users as a part of the reviews. The number of photos and the size of the photo varies. The output retrieved from the model may be (2,5,7,8) which predicts the labels 2,5,7,8 are Applicable to the restaurant and 0,1,3,4 are Not applicable.

The execution involved setting up frameworks, preparing data, training models and predicting outputs. The computation resource used for this project was a Tensor-flow enabled machine with a NVIDIA GTX1800 graphics card .

2. Relevant Work

2.1. Kaggle Winners

Yelp restaurant photo classification competition was hosted in Kaggle. In the competition Dmitrii Tsybulevskii (Tsybulevskii) won the first place with winning solution and Thuyen Ngo (Ngo) won second position. The preprocessing techniques used by the first winner was Photo-level feature extraction and Business level feature extraction. Best performing photo-level feature extraction was: Full ImageNet trained Inception-BN, Inception-V3, and ResNet. Averaging photo level features, Fishers Vectors and VLAD descriptor was used for Business level feature extraction. After preprocessing, Logistic Regression, Neural Network and XBG was used for classification. And it was proved that Neural Network performed best than LR and XGB, since multi-label problem was handled well using Neural Network.

Thuyen Ngo finished in 2nd place, with an F1 score

of 0.83168, Ngo used pre-trained convolutional networks to extract image features. Although he started using Incpetion-V3, he ended up using the pre-trained resnet-152 provided by Facebook. In order to solve the multi label and multi-aspect problem, he used multilayer perceptron as it gave the flexibility to handle both the multiple label and multiple instance simultaneously, he also used 9 sigmoid units and for multiple instance and employed a technique similar to the attention mechanism in the neural network literature, this enabled to system to learn by itself. Ngo, built a model based on business level labels and used the standard cross entropy as the loss function. The training was performed with Nesterovs accelerated SGD. Random 5-fold split was done and the results were based on the average of 5 models from 5-fold validation.

2.2. Convolutional Neural Network

Convolutional Neural Network (ConvNet) are deep artificial neural network that classifies images, cluster the images by similarity and object detection are performed within the scenes. ConvNets identify faces, individual, tumors, street signs, and many other feature of an image. CNN are flexible yet powerful deep learning models. CNN finds features in an image using Feature Detector and put them into a feature map. By having in feature map it still preserves the spatial relationship between pixels. Steps to go through images in CNN is Convolution, Max pooling, Flattening and Full Connection (Fully Connected Neural Network).

CNN performs a non-linear transformation of its input from one vector space to another. By performing non-linear transformation at each layer, we project the input vector space to new vector space, and draw decision boundary to separate classes. To form a full ConvNet layer we stack all these layers. Figure 1 shows the architecture of ConvNet.

A ConvNet classification could have the following architecture: Input [224,224,3] will hold the raw pixel value of the images, in this case an image of width 224, height 224 and with three color channels R,G,B. Conv layer is a combined integration of two functions and it shows how one function modifies the shape of the other. Where the two functions are the input and feature detector. The dot product of it gives a feature map. ReLu layer is used to increase the non-linearity in the network. Pool layer will perform downsamppling along the spatial dimensions, resulting in reduced size of [112,112,3]. Fully connected layer will compute the probability of an image belonging to a particular class. And each neuron in a layer is connected to every neuron in another layer which is why it is called fully connected layer.

In our project we will use Resnet ConvNet to extract features from images.

2.3. Deep residual learning for image recognition (ResNet)

In 2015, Microsoft Research Asia came up with Resnet, which is a 152 layer network architecture that set new records in classification, detection, and localization through one incredible architecture. ResNet won ILSVRC 2015 with an incredible error rate of 3.6% (Depending on their skill and expertise, humans generally hover around a 5-10% error rate. In (He et al., 2015), instead of few stacked layers fitting a desired underlying mapping, they explicitly let those layers fit a residual mapping which can be seen in Figure 2. They hypothesized that it is easier to optimize the residual mapping instead of optimizing the original, unreferenced mapping.

They are extremely deep residual neural networks which are easy to optimize than the other state-of-are nets whose error rate increases when the depth increases. They have experimented with 50, 101 and 152 layers. Where Resnet-152 obtained a top-5 error rate of 5.71, compared to ResNet-50 and ResNet-101 which obtained top-5 error rate of 6.71 and 6.05. Which is the reason why we chose Resnet-152 layer.

2.4. Multi-label Classification

Multi-label classification problem (Jain) is a form of classification problem where one or more target labels are assigned to each instance. There are 3 ways to solve multi-label classification problem, namely: Problem transformation which will try to transform multi-label problem into single-label problem, which is carried out in three different ways as, binary relevance, classifier chains and label powerset. Adapted Algorithm, adapting an algorithm to directly perform multi-label classification. And thirdly, ensembled approaches which always produces better results. In our project we used binary relevance approach. The advantage of using binary relevance is because of its computational simplicity and also it handles irregular labeling well, contrary to the label powerset approach where each label combination is treated as single class (Jain).

3. Research Method

3.1. Data Processing

The Data consisted of 2000 restaurants with labeled train set and around 230,000 images. The unlabeled test set consisted of 10,000 restaurants with approximately 1.2 million images, with one or more of the labels applicable. Before directly feeding the images into the model, we convert the raw data into understandable format. For this project the data preprocessing method used was Data reduction, we used one of the instance selection algorithms. We reduced the spatial dimensions of the images and also extracted the

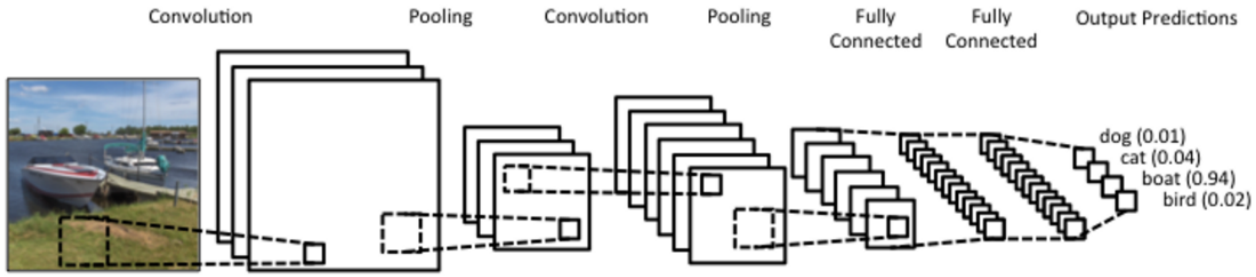

Figure 1. Convolutional Neural Network Architecture

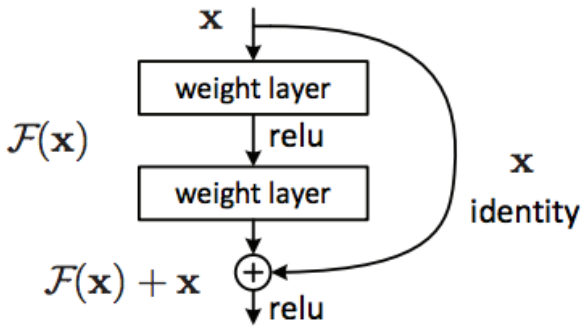

Figure 2. A Residual block

important features from images through pre-trained convolutional neural network called ResNet. Which is explained in the next section Image feature extraction. And also once we get the features from the images, we merged it together with the business data. Which will be explained in the section business feature extraction.

3.2. Image Feature Extraction

Feature extraction was done to reduce the training cost and the complexity of processing high dimensional images. Each image uploaded by the users in the Yelp had different dimensions and resolutions. In order to consider similar dimensionality to each and every image we used a spatial dimension of width 224, height of 224 and three colored channels R,G,B. Features from all images belonging to restaurants were extracted using the deep residual neural network called ResNet, which is pre-trained on Imagenet. We decided to use Resnet-152 than other state-of-art CNNs because of ResNets extremely deep neural network. Before 2014 deeper neural network were considered to be more difficult to train until ResNet came in 2015. (He et al., 2015) In ResNet they present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. The ResNet was trained on Imagenet dataset with a depth of 152 layers which is 8 times deeper than VGG nets (Deshpande) and still having reduced complexity.

3.3. Business Feature Extraction

Business feature extraction was done by taking the average of every image for each restaurant. The result of it is a vector representation for every restaurant. To combine every feature in a single vector, average pooling was the reasonable method to use. Because different image activates different features and it gave reasonable results as well.

3.4. Classification

The restaurant features were finally multi-label classified using Scikit learns one-vs-rest support vector machine (SVM) and random forest (RF). One-vs-Rest is a multi-class or multi-label strategy. When using supervised learning, on their own they are just binary classifiers, which means they cannot handle target vector with more than two classes. One-vs-rest is a clever extension to do that. In one-vs-rest SVM and RF, a separate model is trained for each class and predicted whether an observation is in that class or not, by making it as a binary classification problem as mentioned in section 2.5. And it assumes each classification problem is independent. And we also utilized 5-Fold and 10-Fold cross validation along with SVM and RF to avoid over fitting. We cross verified our results with and without the cross validation for SVM and RF, there was not much difference in their results. Therefore, we can confidently say our model does not over fit.

SVM is a supervised learning classification model, defined by a separating hyperplane. The algorithm produces an optimal hyperplane that can categorize new examples. RF is an ensemble method that contains one or more decision trees to come up with a decision, by taking the average among all the trees.

4. Results

4.1. Evaluation Metrics

The F1-score is an evaluation metric that represents the harmonic mean of precision and recall. Kaggle uses the mean F1-score (F1-samples) to evaluate the predicted labels on the test set by computing the F1-score for every testing example and computing the average over all testing examples. The F1-score per testing example is calculated by using the labels as instances in order to compute the precision and recall. Another method of computing the F1-score is by counting all true and false positives along with the true and false negatives for every testing example and calculate the F1-score (F1-micro) afterwards. Finally, the F1-score per label is used, which is important to determine a models performance for individual labels. The latter two scores can be computed only over the training set since there are no labels available for the test set.

4.2. Results on Train Set

All the following results are produced by taking average pooling to combine image features to restaurant features and fed them into the classification algorithms for decision making.

Figure 3 shows the F1-score for support vector machine and random forest. The F1-score for each labels are shown as well. It is clear from the plot that the model works significantly better on Random Forest with a F1-score of 0.83. Also the model works significantly better on some labels like has_table_service, has_alcohol, good_for_kids and worst on labels like good_for_lunch, outdoor_seating and ambiance_is_classy.

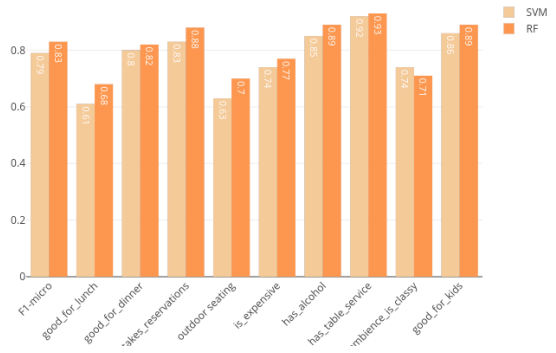

Figure 3. F1-Score on train set for SVM and Random Forest

Figure 4 shows the precision score obtained for support vector machine and random forest, and also for all the individual labels. The precision score for both SVM and RF are the same. But the model performs significantly better for has_table_service and worst for good_for_lunch.

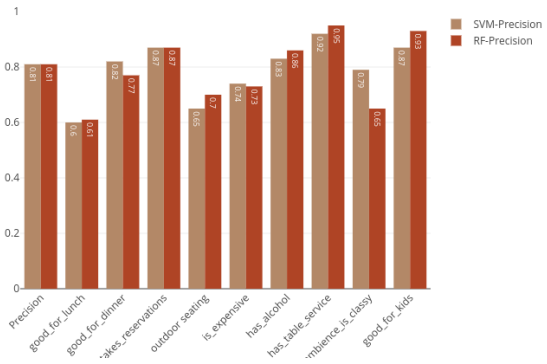

Figure 4. Precision of SVM and Random Forest

Figure 5 shows the recall score for SVM and RF and for every individual label. The recall score for SVM is 0.78 and for RF is 0.85. The RF model performs significantly better than SVM. The model performs better on has_table_service and worst on outdoor_seating.

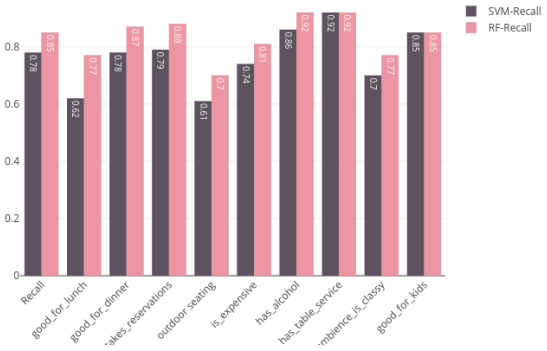

Figure 5. Recall of SVM and Random Forest

For both 5-Fold and 10-Fold cross validation with SVM gave an F1-score of 0.80 and similarly for RF an F1-score of 0.80 was obtained.

4.3. Results on Test Set

For the test set containing 10,000 restaurants and 1.2 million images, we obtained an F1-Score of 0.77 for one-vs-rest SVM model and an F1-score of 0.80 for RF model. We obtained this score by uploading the predicted labels obtained for each and every restaurant to kaggle.

5. Performance

5.1. Qualitative Analysis

Apart from the Quantitative analysis with F1 Score, we have performed Qualitative analysis on the labels predicted together and the ones that are not.

Good for lunch Restaurants that are predicted with labels good for lunch are also commonly marked as good for kids, where very few good for lunch are rarely labeled as restaurant is expensive.With the photos of lunch labels with food in economical silverware , explains that lunch serving restaurants are generally inexpensive.

Figure 6. Good for lunch

Good for dinner Good for Dinner is the most commonly used label , this label is widely predicted along with restaurant that takes reservations , that serves alcohol and has table service .Restaurants that are labeled as good for dinner are rarely good for kids. Good for lunch is the least related label to good for dinner label.

Figure 7. Good for dinner

Takes reservation Good for dinner is the label that is most frequently predicted label. followed by label: has alcohol and has table services.Surprisingly it is least related to the restaurant is expensive and ambiance is classy labels . As the images uploaded for such labels are relatively low , with more images for those labels will enable further analysis.

Outdoor seating Restaurants predicted with labels has alcohol , good for kids , has table services are the frequently predicted labels apart from takes reservation . Labels like Ambiance is classy , restaurant is classy , good for lunch , good for dinner are predicted along with Outdoor seating labels . Less than 5% of the overall prediction has outdoor seating photos.

Figure 8. Take reservations

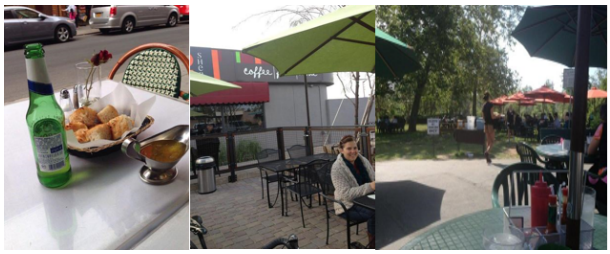

Figure 9. Outdoor seatings

Restaurant is expensive Predicted labels for good for dinner , has alcohol , takes reservations , has table service are almost same are combined with restaurant is expensive label and so the image predicted are almost the same. Outdoor seating and good for kids are the labels that are not predicted with restaurant is expensive label.

Has Alcohol Good for dinner is the label with highest number of predications with restaurants has alcohol, followed by takes reservation , outdoor seating.only 0.17% of the entire predictions has good for children and has alcohol in common.

Figure 10. Has alcohol

Has Table service This label has the most number of images while training the model.As most of the user submitted photos are in restaurants and not in fast-food.good for dinner is commonly added a label along with has table service.

Ambiance is Classy This images predicted for this label are darker and few in numbers ,hence prediction for this label cannot be accurate with the available data.

Good for kids This label is commonly used with good for

lunch , good for dinner and restaurants that takes reservations. Less common labels grouped are Has Alcohol , ambiance is classy and restaurant is expensive.

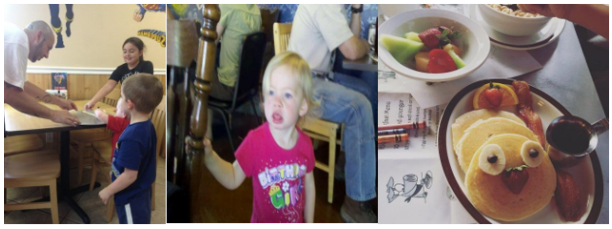

Figure 11. Good for kids

Other Photos Apart from the photos that can be tagged to their relative labels, some photos of the objects in restaurant , menu selfie of reviewers, etc.., are mentioned in user review with photos as below .These photos are used in training and testing the model , the predictions are fairly accurate as the context of these images are sometimes irrelevant to the classification performed.

Figure 12. Other photos

6. Conclusion

In this project we predicted restaurant categories based on the images uploaded by the user in the yelp application. The results obtained using one-vs rest SVM and RF on the training dataset, which were evaluated by multiple metrics and the testing set resulted on one overall F1 score. We approached the YRCC by utilizing ResNet-152 to extract features from images and combined them using average pooling and classified using one-vs-rest support vector machine. Which resulted in a F1-score of 0.77 on the test set in kaggle. The best performing combination of model was by extracting the image features from ResNet-152 and then classifying the restaurants using random forest. Which resulted in an overall F1-score of 0.80 on the test set in kaggle. The performance of ensemble approach like random forest is much better than support vector machine.

References

Adam Gibson, Chris Nicholson, Josh Patterson. A beginners guide to deep convolutional neural networks (cnns). URL https://deeplearning4j.org/convolutionalnetwork#.

Deshpande, Adit. The 9 deep learning papers you need to know about. URL https://adeshpande3.github.io/The-9-Deep-Learning-Papers-You-Need-To-Know-About.html.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. Technical report, Microsoft Research, 2015. URL https://arxiv.org/pdf/1512.03385.pdf.

Jain, Shubham. Solving multi-label classification problems. URL https://www.analyticsvidhya.com/blog/2017/08/introduction-to-multi-label-classification/.

Karpathy, Andrej. Convolutional neural networks for visual recognition. URL https://cs231n.github.io/convolutional-networks/.

Krizhevsky, Alex and nd Geoffrey E Hintonresnet, Ilya Sutskever. Imagenet classification with deep convolutional neural networks. Technical report, University of Toronto, 2012.

Ngo, Thuyen. Yelp restaurant photo classification challenge second winner. URL https://engineeringblog.yelp.com/2016/04/yelp-kaggle-photo-challenge-interview-1.html.

Tsybulevskii, Dmitrii. Yelp restaurant photo classification challenge first winner. URL https://engineeringblog.yelp.com/2016/05/yelp-kaggle-photo-challenge-interview-2.html.