COGS 260 Final Project: Multi-Instance Multi-Label Classification for Restaurant Categorization using Deep Learning

Adithya Bharadwaj Balaji
Electrical and Computer Engineering
PID: A53211421
a2balaji@ucsd.edu

Arun Kumar Chandrasekar
Electrical and Computer Engineering
PID: A53221342
a9chandr@ucsd.edu

Abstract

In this report, the authors present a model that automatically labels restaurants with multiple categories based on user-submitted photos. The model was trained and tested using four baseline AlexNets(and derivatives). Additionally, the authors also present an ensemble combination of these networks that produced superior performance. Through this technique an F1 score of 0.7781, which is close to the highest achieved score of 0.8278.

1 Introduction

Image recognition is a pivotal component in the burgeoning world of artificial intelligence. Great strides have been made with respect to image recognition and classification in the recent years, which is starting to find its way in a plethora of real life applications today and image annotation has always been one among them. Serving as an extension to this, the "Yelp Restaurant Classification Challenge" [5], hosted on the data science forum Kaggle, proposes the task of tagging restaurants with multiple labels automatically, based on the photos submitted by users.

Yelp is a popular social networking website that allows users to share reviews and information about businesses. In addition to written reviews, users can upload photos and categorize businesses. Yelp's users upload an enormous number of photos every day, but selecting the labels is optional, leaving some restaurants improperly categorized. Our goal 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: 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.

This challenge has two interesting and unconventional technical aspects to it. The first is the binary multi-label classification aspect [4], which means that every restaurant corresponds to a set of labels instead of one individual label [3]. The second aspect is the multiple-instance [4] nature of the challenge. In conventional image classification tasks, every image has a class or labels associated with it. In the challenge presented, only restaurants are labeled which means that there is no information about the images except for their correspondence to a specific restaurant. This increases the difficulty of developing predictive models based on the images. The aforementioned problems are circumvented through the proposed model by restructuring the problem based on the information provided.

2 Dataset Description and Preprocessing of images

The dataset that was used in this project was provided by Yelp through the Yelp Restaurant photo Classification competition hosted on Kaggle. The dataset comprised of 234,840 photos belonging

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

to 2000 restaurants as a part of the training set and 237152 photos as a part of the test dataset which should are mapped to their corresponding business IDs (restaurants). The number of images corresponding to each restaurant in the training set ranges from 1 to 2,974 across different restaurants. The average number of images per restaurant comes out to be 117. As mentioned earlier, each of these restaurants(business) can have nine self-explanatory labels. These labels were annotated by the Yelp community. A few samples from the training dataset can be seen in Fig. 1, Fig. 2 and Fig. 3. The Fig. 1 shows examples of restaurants that have labels like: has table service, takes reservations, is expensive and ambiance is classy. The Fig. 2 shows examples of restaurants that are good for lunch and outdoor seating. Fig. 3 shows examples of restaurants that are good for kids.

Figure 1: Photos of restaurants having images showing expensive and classy interiors

Figure 2: Photos of restaurants having images showing that they are good for lunch and have outdoor seating

Figure 3: Photos of restaurants having images showing that they are good for kids

2.1 Preprocessing of images

The images are of varying size, mostly having sizes 350 x 500 or 500x 350. These images are preprocessed by resizing them to a size of 224 x 224 suitable for feeding to the neural network.

The distribution of the photos across businesses can be seen in Fig. 4. From Fig. 4, we can see that nearly 80% of the total distribution can be explained by picking the first 100 images for each business. To deal with this unequal distribution we have subsampled the data to include only the first 100 images. Then we used a 80-20 split for training and validation.

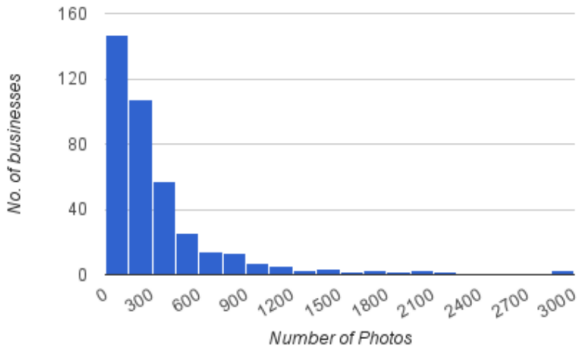

Figure 4: Distribution of photos across businesses

3 Design of the Network

The neural network architecture that we selected for solving the problem at hand was the AlexNet [1]. This was inspired from the fact that yelp actually uses AlexNet in its product deployment where photos are categorized. The feature extraction part involved four different variants of AlexNet: BVLC AlexNet, BVLC Reference CaffeNet, Places205-AlexNet and the Hybrid-AlexNet. All the models were created and pre-trained in Caffe, Deep Learning framework and on different image datasets like the Places 205 and the ILSVRC 2012.The architectures utilized are briefly explained below.

3.1 BVLC AlexNet

BVLC AlexNet is the replication of the model described in the AlexNet publication [1]. But the difference being that it was not trained with the relighting data augmentation and also intializing a non zero bias of 0.1 instead of 1. The framework of the BVLC AlexNet is shown in Fig. 5. From Fig. 5, we can see the varying dimensions of the initial image vector to the final feature vector.

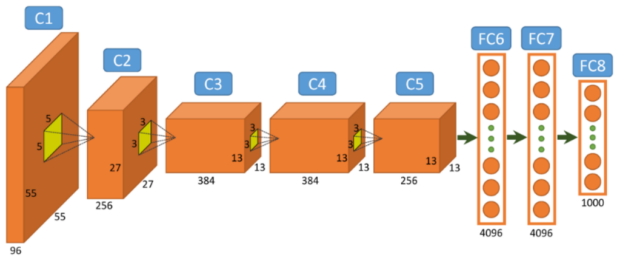

Figure 5: Architecture of the BVLC AlexNet [1]

3.2 BVLC Reference CaffeNet

BVLC Reference CaffeNet was obtained as a result of the Caffe ImageNet model training instructions. It is a replication of the model described in the AlexNet but with changes like switching up of the pooling and the normalization layers and not training with the relighting data augumentation.

3.3 Places205 AlexNet

Places 205 AlexNet is the vanilla AlexNet that was trained on a scene centric database called "Places" which was created by MIT's CS and AI Lab. It was trained on 205 scene categories and 2.5 million images with a category label.

3.4 Hybrid AlexNet

Hybrid AlexNet was developed as a fusion of applying AlexNet on Places database and the ILSVRC 2012 database. It was trained on 1183 categories with 205 scene categories from the Places database and 978 object categories from the ImageNet database.

4 Training procedure

We have carefully selected the four pre-trained models will be use to pass the training images through them and generate their corresponding feature vectors. By using pre-trained models which have been previously trained on large datasets, we can directly use the weights and architecture obtained and apply the learning on our problem statement. This is known as transfer learning. We "transfer the learning" of the pre-trained model to our specific problem statement.

On passing the images from the training dataset along with their corresponding labeling, we obtain the feature vectors for the images at the end of the neural network. We use the pre-trained models for the feature extraction purpose and remove the output layer and extract the features from the penultimate layers. We then use the entire network as a fixed feature extractor for the training dataset.

4.1 Feature Extraction

The image features are selected from the convolution layers C4, C5 and the fully connected layers FC6 and FC7 as shown in Fig.5. The features are then stacked and flattened into a single vector. the outputs from the convolutional layers C4 and C5 are given as input to a relaxed version of spatial pooling [2] which takes only 4 quadrants of a feature map. So, every image passed through one of our models would be converted into a vector containing 384*4+256*4+4096+4096 = 10752 features. The pooling performed by the spatial pyramid matching can be seen in Fig. 6.

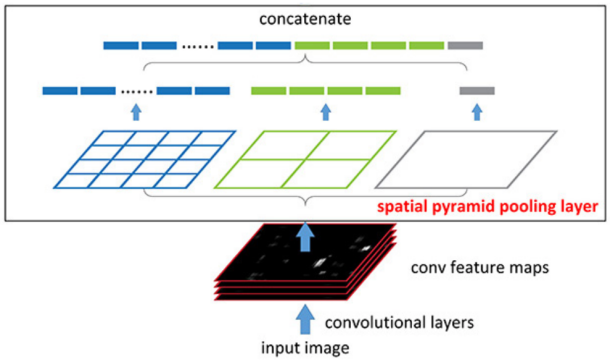

Figure 6: Relaxed Spatial Pyramid pooling [2]

Combining the image set into features of the restaurant was done by taking a mean if the constituent image features resulting in 10752 feature vector per restaurant. Also another approach was to take

the maximum of image features and populating the feature vector with it. Then an ensemble was performed along with the mean model which provided a strong boost to the evaluation metric.

4.2 Linear classifier

After the feature extraction, we obtain 8 feature sets (2 x 4 AlexNet models) and we use this to create a model for restaurant tagging. The type of classifier utilized was based on the high dimensionality of the problem (10752) and the relatively low number of instances (2000). So based on this we decided to use a linear classifier. On experimenting with linear SVM and logistic regression, we felt that logistic regression tends to perform better in this case.

4.3 Fusion of Models

Next step was deciding the fusion of the eight models that we have at hand. On experimenting with techniques like simple average, voting and blending. We found that blending via meta modeling turned out to be the best.

4.4 Label Dependencies

The dependencies between the labels were accounted for by using the the Classifier Chains method which involved feeding the logistic regression outputs, 9 label probabilities x 8 models to a meta-model of two non-linear classifiers: using the same neural network as before and XGBoost [6]. By using this technique both the model fusion and label dependencies have been accomplished in a single go.

5 Evaluation metric

The evaluation metric that was used was the Mean F1 score which is also known as the example-based F-measure in the multi-label learning literature. The F1 score is mainly used the information retrival domain to measure the accuracy using statistical parameters like precision p and recall r.

Precision is the ratio of teh true positives(tp) to all predicted positives(tp+fp). Recall is the ratio of the true positives to all actual positives(tp+fn).

The F1 score is given by:

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

where

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

The F1 metric weights recall and precision equally and a good retrival algorithm will maximize both the precision and recall simultaneously. Thus, moderately good performance on both will be favored over extremely good performance on one and poor performance on the other.

6 Testing phase and implementation

The testing was done by passing the images from the test dataset through the same neural network architecture used to train our model. Then generating labels using our linear classifier and then linking the photos to their corresponding restaurants and generating the labels for the restaurants which is performed by taking the union of the label sets of the images belonging to the same restaurant. Then a submission file was generated in the .csv format which contained two columns: restaurant Ids and their corresponding labels.

6.1 Implementation

Our implementation was done through python utilizing the Caffe frameworks and frameworks from the MIT CS AI lab. All our code was based on python 2.7 where we utilized libraries like sklearn,

Table 1: F1 scores for different Neural Network Architectures

BVLC AlexNet		BVLC Ref CaffeNet		Places205 AlexNet		Hybrid AlexNet	
Mean	Max	Mean	Max	Mean	Max	Mean	Max
0.7527	0.7402	0.7369	0.7380	0.7455	0.7204	0.7623	0.7485

Table 2: Accuracy for labels

Label No.	Label Name	Accuracy
0	Good for lunch	76%
1	Good for dinner	80%
2	Takes reservations	85%
3	Outdoor seating	66%
4	Restaurant is expensive	86%
5	Has alcohol	82%
6	Has table service	88%
7	Ambiance is classy	84%
8	Good for kids	84%

pandas, numpy and xgboost. The experiments were performed on the Amazon AWS instance with the p2.xlarge instance containing one NVIDIA GPU with a storage space of 150 GB utilized. The process of generating feature vectors took nearly 5 hours due to the large size of the training dataset. Similarly the blending of the meta models and generating the test file took another 5-7 hours.

7 Results

In this part, we report the F1 scores for the different models, both baseline and ensemble. Table 1 shows the results for the baseline. As observed, the results for both the mean and max averaging feature models are documented. For the most part, the mean combination works better than the max combination, as evidenced by the results furnished in Table 1. This seems to be different just in the case of BVLC Ref Caffenet. Since the performances of both the architectures are comparable even in this case, this particular finding can be classified as an aberration and hence can be hypothesized that the mean combination is superior. An explanation for this could be that, in the case of mean combination, features from frequently occurring images receive a heavier weight in the final restaurant feature vector, unlike max combination, which takes one value from one image per feature. Thus, the restaurant features will be similar to the features of the most occurring images for that restaurant.

Building on the results of the baseline models, the ensemble model was subsequently tested and it was found to have an accuracy of 0.7781, which is comparable to the state of the art standard of 0.82(leaderboard). It is apparent from these results that this particular ensemble works better than the individual models. Since the three models trained on different datasets produce results that are very similar, it is interesting to note that their combination works well. This could be explained by the models having a slightly different bias, which is balanced out by combining their predictions. A good amount of research has been done in this regard and the general consensus is that the use of ensembles also reduces the risk of overfitting due to a larger hypothesis space [7]. It can also be hypothesized that even though the model's overall performances are similar, they are slightly better at predicting some labels than others. Hence by combining these predictions, the overall performance on all labels improves.

The performance of the individual labels for the ensemble technique is tabulated in Table 2. It can be observed that some labels perform better than others. Noticeably, the label has table service performs the best, whereas labels like outdoor setting and good for lunch perform very poorly. The cause for this might be that, as is the case with outdoor seating, the data and labels are noisy and relevant patterns can't be learned by the CNN. Some of the properly classified images are presented in Fig. 7. From Fig. 7, we can see that our classifier performs similar to the intuitive nature of the human who tag the images on Yelp. The first image clearly shows that it is a place that serves alcohol. Similarly, the second image shows a fast food joint as good for lunch and good for kids. The third image shows us that the ambiance is classy.

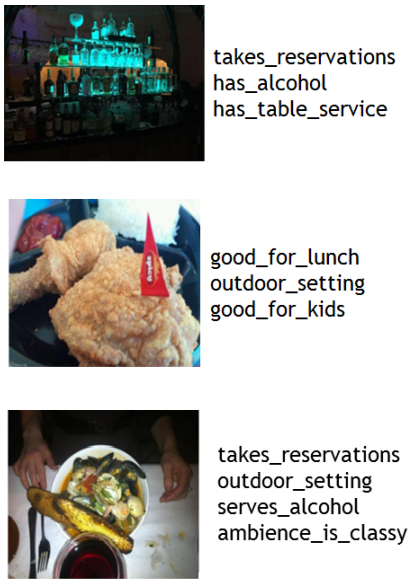

Figure 7: Sample result obtained on classifying the test images

8 Future Scope

An obvious method to improve the performance would be to use a more advanced CNN like ResNet 152, which is among the best when it comes to accuracy. But this would require more computational power, which is a trade-off. Another interesting approach would be to cluster the images into the five categories mentioned earlier, using unsupervised machine learning techniques, which would hopefully result in the five clusters. Subsequently, different clusters of images could be fed to models trained on different feature specific dataset like the Food101 CNN, for extracting feature from cluster of images depicting food, or Places CNN would extract features from pictures of the interior and exterior-seating clusters. The hypothesis is that an ensemble of such models would perform better due to the individual models being better at predicting specific labels within their domain, thus a combination of those models would perform better on all labels.

9 Conclusion

This report explored several deep learning architectures to predict restaurant categories based on user-uploaded images. Results were obtained on the test set, which resulted in an overall F1 score metric. This performance was found to be very comparable to the state of the art system. The baseline consisted of features extracted from four versions of AlexNet, combined using spatial pyramid pooling, with mean and max combinations of the resulting features being classified using a one versus rest SVM. It was found that the best performing model was the simple ensemble that combined predictions of the aforementioned four AlexNets in a mean fashion. It produced an F1 score of 0.78 on the test set, which is comparable to the top of the line performance of 0.82.

Acknowledgement

We would like to sincerely thank Professor Zhuowen Tu for providing us wondering opportunity to implement the deep learning model on the Amazon AWS Instance. It was our first time using it and it was a wholesome and a nourishing experience. We would also like to thank the TA, Saining Xie for helping us navigating through out doubts and figuring out the nitty gritty details.

References

[1] Krizhevsky, A. & Sutskever, I. & Hinton, G.E.(2012) ImageNet Classification with Deep Convolutional Nueral Networks. *Advances in Neural Information Processing Systems 7*, pp. 1097–1105. Cambridge, MA: MIT Press.

[2] Kaiming, H. & Zhang, X. & Ren, S.& Sun, J.(2014) Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *European Conference on Computer Vision*, pp. 346–361. Springer International Publishing.

[3] Jansche, M.(2007) A Maximum Expected Utility Framework for Binary Sequence Labeling *ACL*

[4] Zhou, Z.H. & Zhang, M.L. (2007) Multi-instance multi-label learning with application to scene classification. *Advances in Neural Information Processing Systems 19*, pp. 1609. Cambridge, MA: MIT Press.

[5] Yelp Restaurant Photo Classification Challenge. *Kaggle*`https://www.kaggle.com/c/yelp-restaurant-photo-classification`

[6] Chen, T. & Guestrin, C. (2016) Xgboost: A scalable tree boosting system. *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM.

[7] Dietterich, T.G. (2000) Ensemble methods in machine learning. *International workshop on multiple classifier systems.*, pp. 1–15. Springer.