Multilabel Classification of Restaurants Through User-Submitted Photos

Kaan Ertas, Neval Cam – {kertas, nevalcam}@stanford.edu

SUMMARY

In this paper we examine the Yelp Photo Classification Challenge on Kaggle, which presents a dataset of user submitted photos of restaurants and 9 possible labels for each business. The task is to predict, from several photos per business, what subset of labels apply to each business. We tackle this multi-instance, multi-label problem by utilizing convolutional neural networks with different approaches to handle data imbalance and the problem of weakly labeled data. Using these methods that can easily be transferred to other similar problems, we achieve an F1 score of 0.80 – close to the highest F1 score achieved on Kaggle which was 0.83.

DATASET

We train our model on 1000 businesses with 32 randomly selected images each. We then validate and test them on 32 businesses with 32 photos each.

The restaurants are labeled with the following tags:
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) ambience is classy 9) good for kids




Label	Frequency
good_for_lunch	0.29
good_for_dinner	0.54
takes_reservations	0.56
outdoor_seating	0.51
restaurant_is_expensive	0.31
has_alcohol	0.68
has_table_service	0.73
ambience_is_classy	0.34
good_for_kids	0.57

Frequency of Labels in Training Set

	good for lunch	good for dinner	takes reservations	outdoor seating	restaurant is expensive	has alcohol	has table service	ambience is classy	good for kids
good for lunch	1	-0.35	-0.37	0.04	-0.37	-0.33	-0.49	-0.38	0.41
good for dinner	-0.35	1	0.64	-0.08	0.51	0.53	0.5	0.53	-0.53
takes reservations	-0.37	0.64	1	-0.01	0.56	0.64	0.63	0.58	-0.52
outdoor seating	0.04	-0.08	-0.01	1	-0.03	0.04	-0.1	0	-0.05
restaurant is expensive	-0.37	0.51	0.56	-0.03	1	0.44	0.4	0.58	-0.57
has alcohol	-0.33	0.53	0.64	0.04	0.44	1	0.58	0.46	-0.49
has table service	-0.49	0.5	0.63	-0.1	0.4	0.58	1	0.42	-0.43
ambience is classy	-0.38	0.53	0.58	0	0.58	0.46	0.42	1	-0.52
good for kids	0.41	-0.53	-0.52	-0.05	-0.57	-0.49	-0.43	-0.52	1

Label Correlation Matrix for Training Set

METHODS

LOSS FUNCTION

$$L(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N} \cdot [y^{(i)} log(\hat{y}^{(i)}) + (1-y^{(i)}) log(1-\hat{y}^{(i)})]$$

WEIGHTED LOSS FUNCTION

$$L(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N} w_i \cdot [y^{(i)} log(\hat{y}^{(i)}) + (1-y^{(i)}) log(1-\hat{y}^{(i)})]$$

CUSTOM THRESHOLDS

Result: Returns array of custom threshold values for each label
Initialze empty array of thresholds;
for *each label l* **do**
  Determine the frequency f of l in the training data sets;
  Predict the training data set;
  Take predictions for l ;
  Sort the predictions for l in ascending order;
  Determine the (1-f)*100 percentile of the sorted predictions ;
  Append the percentile to the thresholds array
end

TRANSFER LEARNING WITH VGG19

We used a pre-trained state of the art model, VGG19. The weights were trained on the ImageNet dataset, which includes copious images of food and room settings. This makes the weights appropriate for our task. To use transfer learning, we removed the final softmax layer and inserted a fully connected 9-neuron sigmoid layer. We froze the training on all layers except the last 3 fully connected layers.

MULTI-INSTANCE LEARNING CONSIDERATIONS

MEAN: For each business, we take the arithmetic mean for each label across it associated photos.
MAX: For each business, we use the maximum values of sigmoidal activations across all photos of the business.

EVALUATION METRIC

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

RESULTS & DISCUSSION

	TRAIN		DEV		TEST	
	Mean	Max	Mean	Max	Mean	Max
b-CNN	0.48	**0.72**	0.40	**0.72**	0.35	0.62
VGG19	0.70	0.69	0.59	0.68	0.68	**0.65**
VGG19-CT	0.86	0.66	**0.74**	0.68	**0.80**	0.63
VGG19-CT-CL	**0.90**	0.67	0.72	0.68	**0.80**	0.63
VGG19-CL	0.75	0.70	0.58	0.70	0.63	**0.65**

Mean F1 Scores for Our Models

Applying custom thresholds (CT) makes our algorithm perform better for mean. Using CT nullifies the effect of weighted loss (CL).

We achieved the highest F1 score of 0.80— close to the highest F1 score achieved on Kaggle which was 0.83.

Comparing CT training and testing results, we find no considerable overfitting.

FUTURE WORK

1. Transfer Learning with Other Architectures (eg. ResNet)
2. Applying attention mechanisms
3. Modular approach: Some labels are object specific; recognizing bottles in images helps learn the label “has alcohol”

REFERENCES

arXiv:1802.04712

C., Wei-Hong. “Yelp.” How We Use Deep Learning to Classify Business Photos Yelp, engineeringblog.yelp.com/2015/10/how-we-use-deep-learning-to-classify-business-photos- \indent at-yelp.html.

Chollet, François. "Keras." (2015).

Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, Ghyslain Gagnon, Multiple instance learning: A survey of problem characteristics and applications, Pattern Recognition, Volume 77 (2018), Pages 329-353.

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

Tsoumakas, Grigorios & Katakis, Ioannis. (2009). Multi-Label Classification: An Overview. International Journal of Data Warehousing and Mining. 3. 1-13. 10.4018/jdwm.2007070101.

Xu, Xin, and Eibe Frank. "Logistic regression and boosting for labeled bags of instances." Pacific-Asia conference on knowledge discovery and data mining. Springer, Berlin, Heidelberg, 2004.

Yelp Restaurant Photo Classification | Kaggle, www.kaggle.com/c/yelp-restaurant- photo-classification.

Yosinski, Jason, et al. "How transferable are features in deep neural networks?." Advances in neural information processing systems. 2014.

Zhou, Zhi-Hua, et al. "Multi-instance multi-label learning." Artificial Intelligence 176.1 (2012): 2291-2320.