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) ambiance is classy
9) good for kids

METHODS

LOSS FUNCTION
\[ L_{p, q} = \frac{1}{n} \sum_{i=1}^{n} \frac{p_i^{q} - p_i^{q} \log (p_i^{p}) (1 - p_i^{q} \log (1 - p_i^{p}))}{p_i^{q} \log (p_i^{p}) (1 - p_i^{q} \log (1 - p_i^{p}))} \]

WEIGHTED LOSS FUNCTION
\[ L_{p, q} = \frac{1}{n} \sum_{i=1}^{n} \frac{p_i^{q} - p_i^{q} \log (p_i^{p}) (1 - p_i^{q} \log (1 - p_i^{p}))}{p_i^{q} \log (p_i^{p}) (1 - p_i^{q} \log (1 - p_i^{p}))} \]

CUSTOM THRESHOLDS
Result: Recall array of instance threshold values for each label
Select evenly spaced areas of thresholds
Recall is 0.80

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-return 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 its associated photos.
MAX: For each business, we use the maximum values of sigmoidal activations across all photos of the business.

EVALUATION METRIC
\[ F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \]

RESULTS & DISCUSSION

<table>
<thead>
<tr>
<th></th>
<th>TRAIN</th>
<th>DEV</th>
<th>TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>0.80</td>
<td>0.78</td>
<td>0.71</td>
</tr>
<tr>
<td>Precision</td>
<td>0.80</td>
<td>0.78</td>
<td>0.71</td>
</tr>
</tbody>
</table>

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


