locality-aware queries

Business listings, that contain representations of business categories
Introduction & Motivation

• Time-consuming and expensive
• Google Street View to automate the process
• Limited Labeled data for business category
Task

- Fine-grained storefront classification from street level imagery
- Create a large, multi-label, training dataset
- Multi-label
- Fine-grained
Outline for the paper

• Challenges in storefront classification
• Ontology based generation of training data
• Model Architecture and Training
• Evaluation
Large within-class variance

Sushi Restaurant

Bench store

Pizza place
Misleading extracted text

(a) Unexpected Language  (b) Misleading Text  (c) Stitching Errors
Business Category Distribution

- 300,000 images for Food and Drink
- 13,000 images for Laundry Services
Labeled Data Acquisition

The operators are asked to mark image areas that contain business related information from the google street view panoramas offered to them which are called biz-patches.
Ontology based generation of training data

Goal: Matching extracted biz-patches \( p \) and sets of relevant category labels

- biz-patch \( p \)
- set of text \( S \)
- location information

\( p \) is a biz-patch of \( b \) if geographical distance between them is less than approximately one city block, and enough extracted text from \( S \) matches \( T \)

- set of text \( T \)
- location information

- known business \( b \)
- Business database \( B \)

\( 3 \) millions \((p,b)\)

\( (p,s) \) where \( p \) a biz-patch and \( s \) is a matching set of labels with varying levels of granularity --> 1.3 million \( p \) and 208 unique labels

Google Map Maker's ontology
Google Map Maker

https://mapmaker.google.com/mapmaker

Add a Place  Cancel

Select category

Type to select a category

Restaurant  Cafe
University  Shopping Mall
Gym  Movie Theater
Bank  Hospital

Type to select from among 2000+ more categories.
Model and Training

- GoogLeNet
- 1.2 Million images for training and 100,000 images for testing
- Splitting is location aware
Model and Training

- Covered the globe with big and small tiles (18km and 2km)
- Tiling alternates between two types of tiles
- A boundary area of 100 meters between adjacent tiles
- Panoramas located in big tiles for training set
- Panoramas located in small tiles for testing set
Model and Training

- Pre-trained using ImageNet
- Dropout rate 70%
- Logistic Regression Top Layer
- Training
  - Each image resized to 256 * 256
  - 220 * 220 after cropping
- Testing
  - A central box of 220 * 220
Evaluation

• When building a business listing it is important to have very high accuracy.
• Top-K accuracy (a prediction is correct if $g_i \cap p_i^k \neq \emptyset$.)

(a) Accuracy at $K$

(b) First Correct Prediction
Evaluation

• recall at certain level of accuracy (90%)

Figure 7. Precision recall curves for some of the top performing categories. The precision curve of the full system is shown as a dashed line. Recall at 90% precision is shown in the legend.

Figure 8. Histogram of correct labels in the top 5 predictions for a set of 300 manually verified images. Color indicates mean prediction confidence. Note that the confidence in prediction is strongly correlated with the accuracy.
Evaluation

• Human Performance Study (agreement of labels for same business)
• 13 Top Level Category
• Full resolution images
• Study 1 two operators
• Study 2 three-four

<table>
<thead>
<tr>
<th>Operator Agreement</th>
<th>Number of images</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Study 1</td>
</tr>
<tr>
<td>100%</td>
<td>50,425</td>
</tr>
<tr>
<td>75%</td>
<td>-</td>
</tr>
<tr>
<td>66%</td>
<td>-</td>
</tr>
<tr>
<td>50%</td>
<td>-</td>
</tr>
<tr>
<td>0%</td>
<td>22,847</td>
</tr>
<tr>
<td><strong>Average Agreement</strong></td>
<td><strong>69%</strong></td>
</tr>
</tbody>
</table>

Table 1. Human Performance studies. In two large scale human studies we have found that manual labelers agree on a label for 69% and 78% of the images.
## Evaluation

- What if the text information is blurred

<table>
<thead>
<tr>
<th>automotive</th>
<th>health &amp; beauty</th>
<th>.999</th>
<th>.992</th>
</tr>
</thead>
<tbody>
<tr>
<td>gas &amp; automotive</td>
<td>health</td>
<td>.999</td>
<td>.985</td>
</tr>
<tr>
<td>shopping</td>
<td>doctor</td>
<td>.999</td>
<td>.961</td>
</tr>
<tr>
<td>store</td>
<td>emergency services</td>
<td>.999</td>
<td>.960</td>
</tr>
<tr>
<td>vehicle dealer</td>
<td>dentist</td>
<td>.998</td>
<td>.945</td>
</tr>
<tr>
<td>health &amp; beauty</td>
<td>beauty</td>
<td>.935</td>
<td>.925</td>
</tr>
<tr>
<td>cosmetics</td>
<td>beauty salon</td>
<td>.742</td>
<td>.713</td>
</tr>
<tr>
<td>hair care</td>
<td></td>
<td>.527</td>
<td>.985</td>
</tr>
<tr>
<td>automotive</td>
<td></td>
<td>.996</td>
<td>.996</td>
</tr>
<tr>
<td>gas &amp; automotive</td>
<td></td>
<td>.995</td>
<td>.995</td>
</tr>
<tr>
<td>shopping</td>
<td></td>
<td>.995</td>
<td>.995</td>
</tr>
<tr>
<td>store</td>
<td></td>
<td>.995</td>
<td>.995</td>
</tr>
<tr>
<td>transportation</td>
<td></td>
<td></td>
<td>.985</td>
</tr>
</tbody>
</table>
Conclusion

• Method for fine grained, multi-label, classification of business storefronts from street level imagery
• Using ontology of entities with geographical attributes to generate large labeled data-set
• Demonstrated the system learned to extract text information when necessary
• Achieves human level accuracy
Questions?