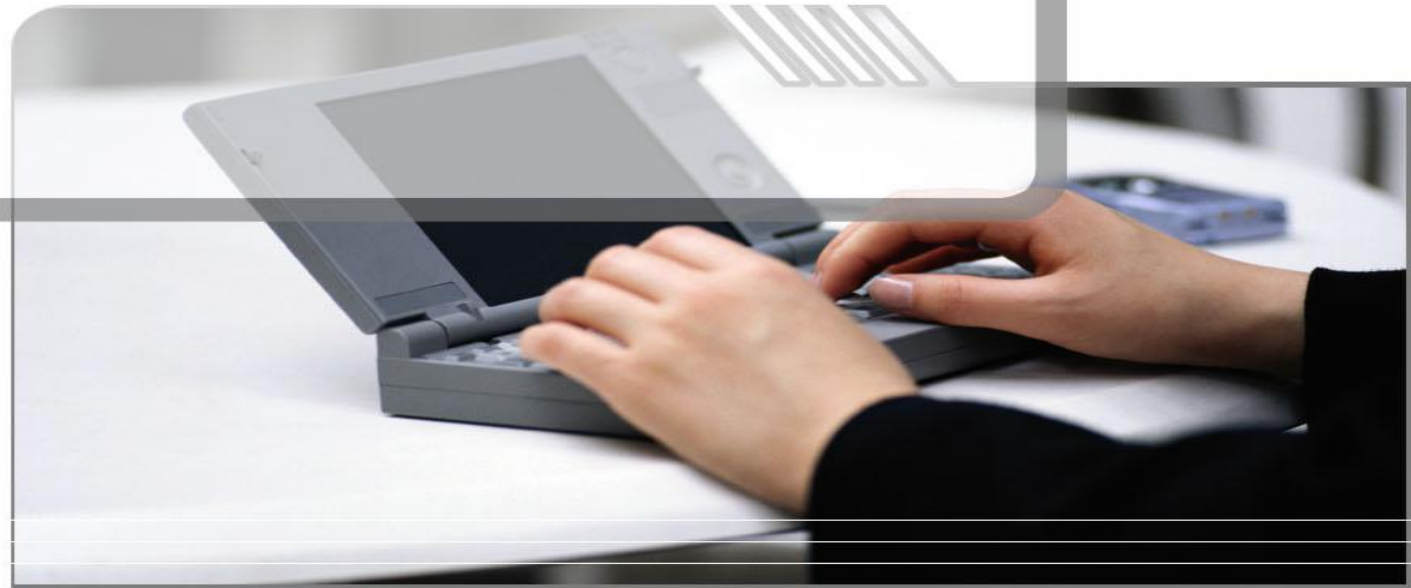


# Ontological Supervision for Fine Grained Classification of Street View Storefronts

CVPR 2015 by Yair Movshovitz-Attias, Qian Yu, Martin C. Stumpe, Vinay Shet, Sacha Arnoud, Liron Yatziv

Presenter: Minghua Jiang



# locality-aware queries



asian restaurant near me



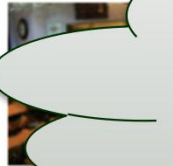
← Rating ▾ Cuisine ▾ Price ▾ Hours ▾

## Thai '99 Restaurant

4.0 ★★★★★ (49) · \$\$ · Thai

Unassuming venue with outdoor seating presents pad Thai, drunken noodles & other traditional dishes.

2210 Fontaine Ave



## Yuan Ho Carryout Restaurant

3.9 ★★★★★ (18) · \$\$ · Chinese

Modest take-out place offering Chinese noodles & stir-fries, plus some Japanese dishes.

117 Maury Ave



## Lemongrass

4.3 ★★★★★ (52) · \$\$ · Thai

Relaxed, warm, sit-down stop for Vietnamese & Thai meals served fast with some vegan options too.

104 14th St NW



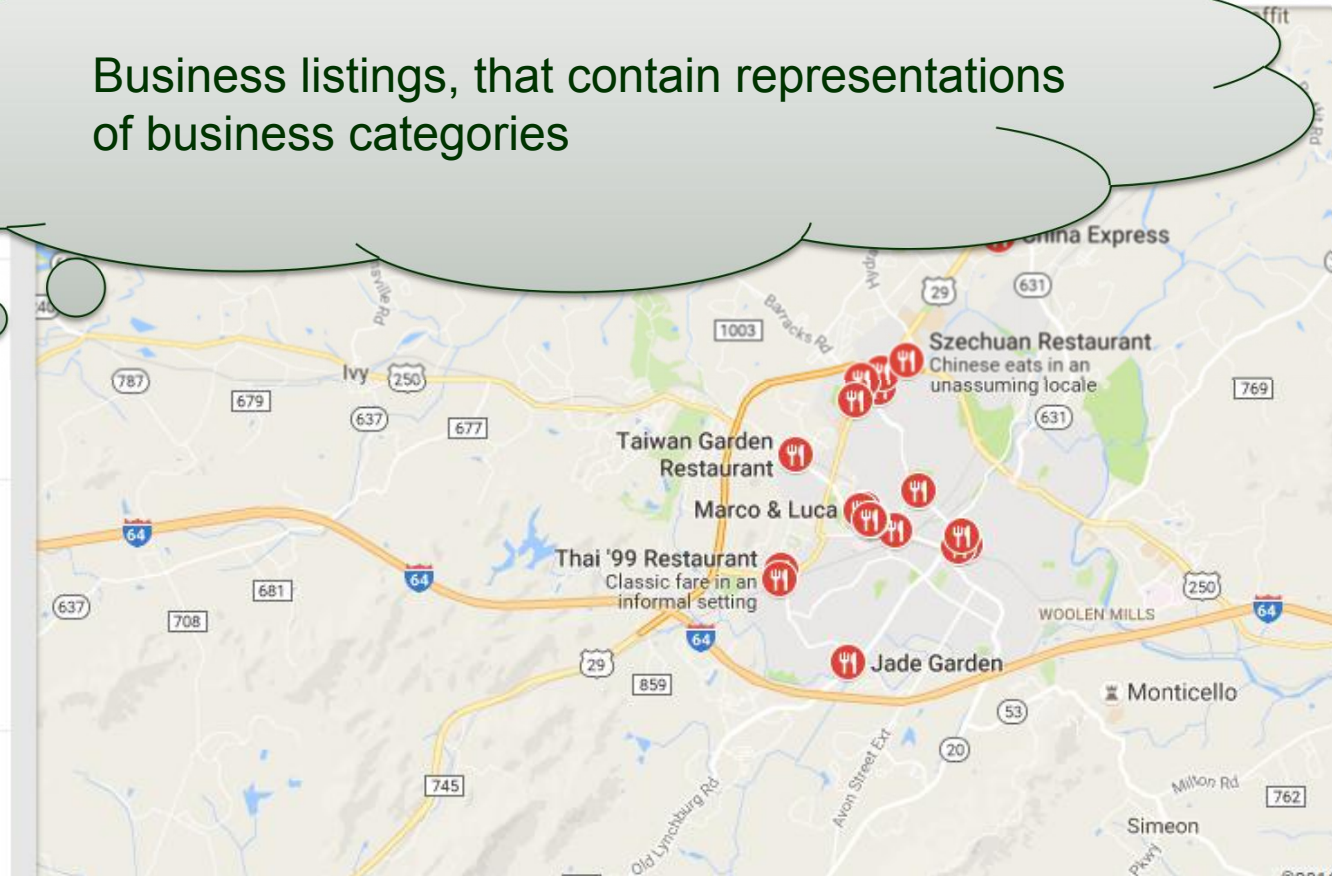
## Ginkgo Chinese Restaurant

4.1 ★★★★★ (13) · Chinese

104 14th St NW



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



Grocery



Plumbing store

# 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



(a) Area Too Small



(b) Area Too Large

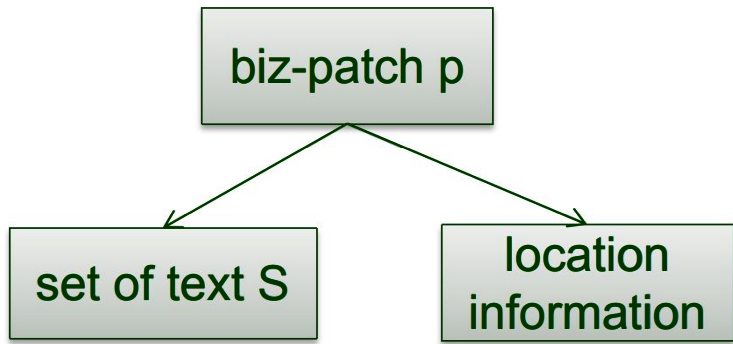


(c) Multiple Businesses

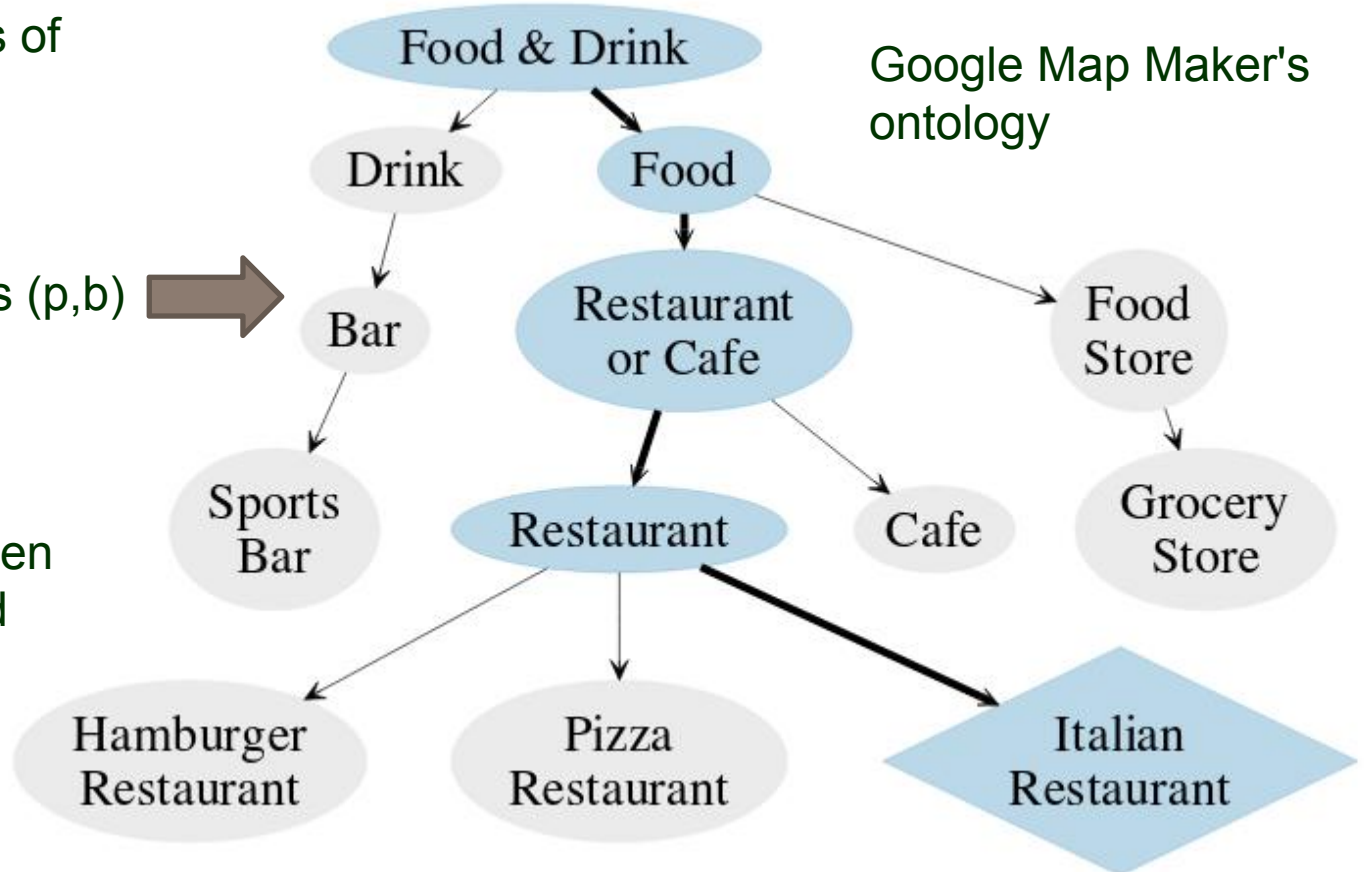
The operators are asked to mark image areas that contain business related information from the google street view panoramas offered to them which are called biz-patches

# Ontology based generation of training data

Goal: Matching extracted biz-patches  $p$  and sets of relevant category labels

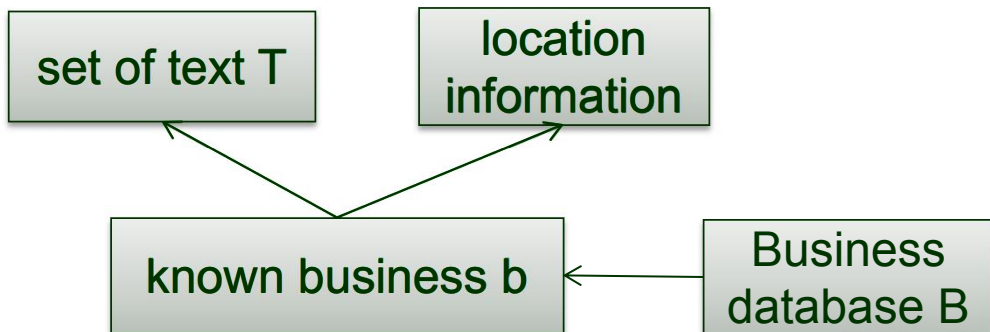


3 millions  $(p,b)$  →



Google Map Maker's ontology

$p$  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  $T$



$(p,s)$  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

The screenshot shows the Google Map Maker web interface. At the top, the browser address bar displays <https://mapmaker.google.com/mapmaker>. Below the address bar is the Google logo and a search bar containing the address "2111 Jefferson Park Avenue, Charlottesville, VA, United States".

On the left side, there is a button labeled "Add a Place" with a "Cancel" link next to it. The main map area shows a street view of "Maury Ave" with a red location pin. A "Select category" dialog box is open over the map, featuring a search input field with the placeholder text "Type to select a category".

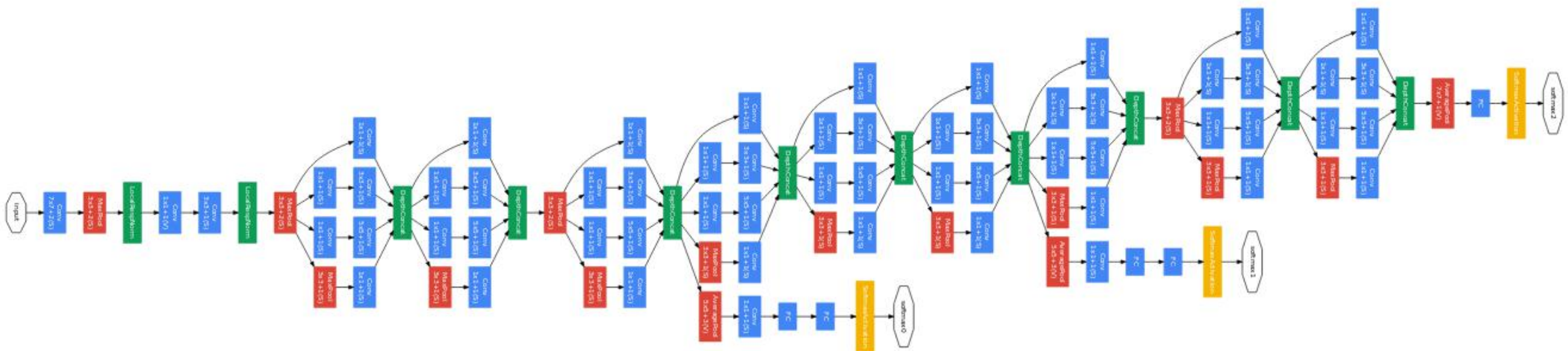
The dialog box lists the following categories:

- Restaurant
- University
- Gym
- Bank
- Cafe
- Shopping Mall
- Movie Theater
- Hospital

At the bottom of the dialog box, it says "Type to select from among 2000+ more categories." The background map shows a street labeled "Maury Ave" and a red location pin.

# Model and Training

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



# Model and Training

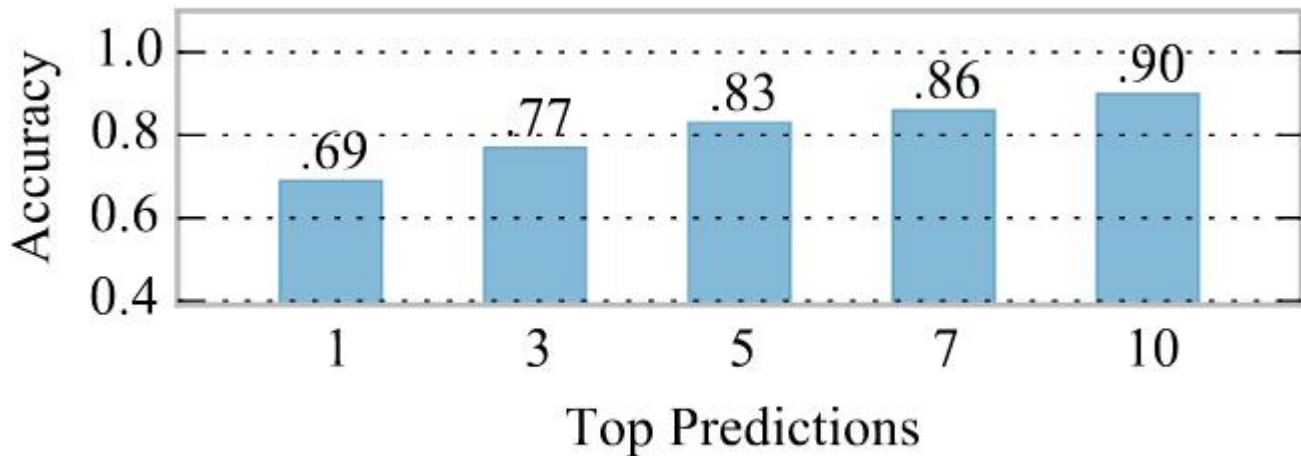
- Covered the globe with big and small tiles (18km and 2km)
- Tiling alternates between two types of tiles
- A boundary area of 100 meters between adjacent tiles
- Panoramas located in big tiles for training set
- Panoramas located in small tiles for testing set

# Model and Training

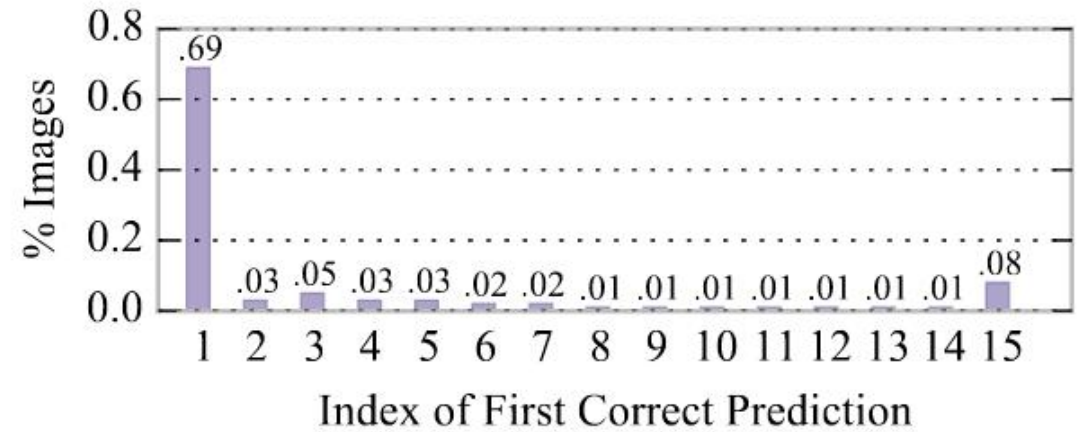
- Pre-trained using ImageNet
- Dropout rate 70%
- Logistic Regression Top Layer
- Training
  - Each image resized to  $256 * 256$
  - $220 * 220$  after cropping
- Testing
  - A central box of  $220 * 220$

# 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%)

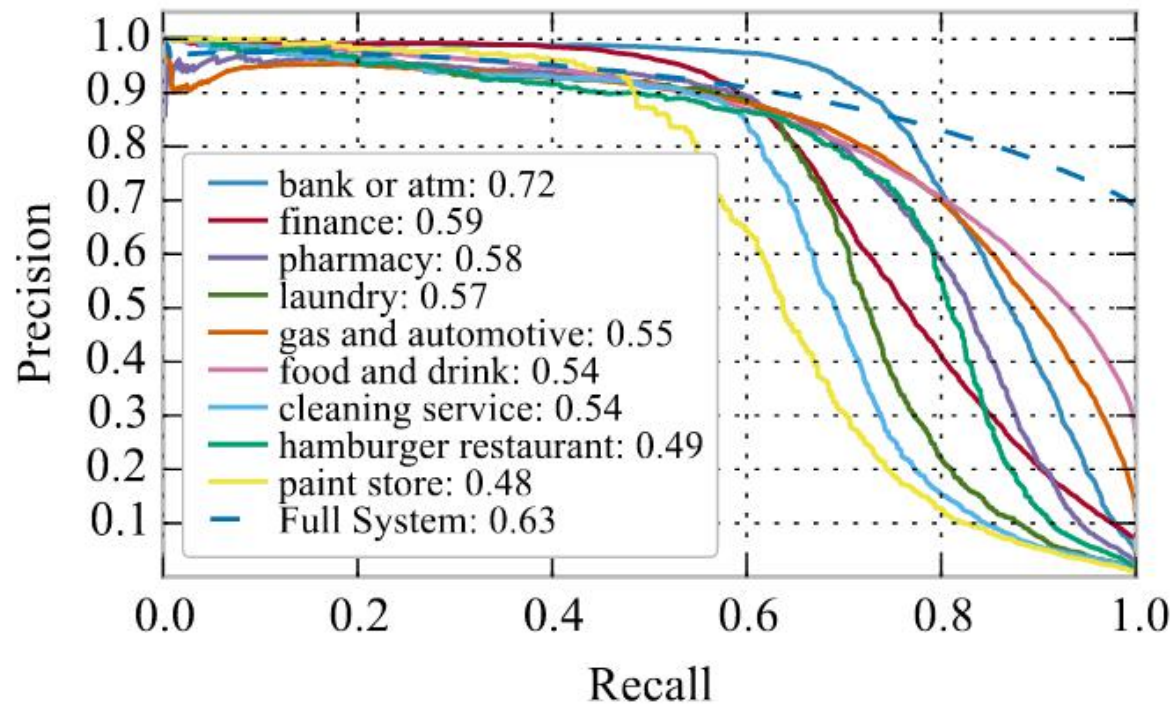


Figure 7. Precision recall curves for some of the top performing categories. The precision curve of the full system is shown as a dashed line. Recall at 90% precision is shown in the legend.

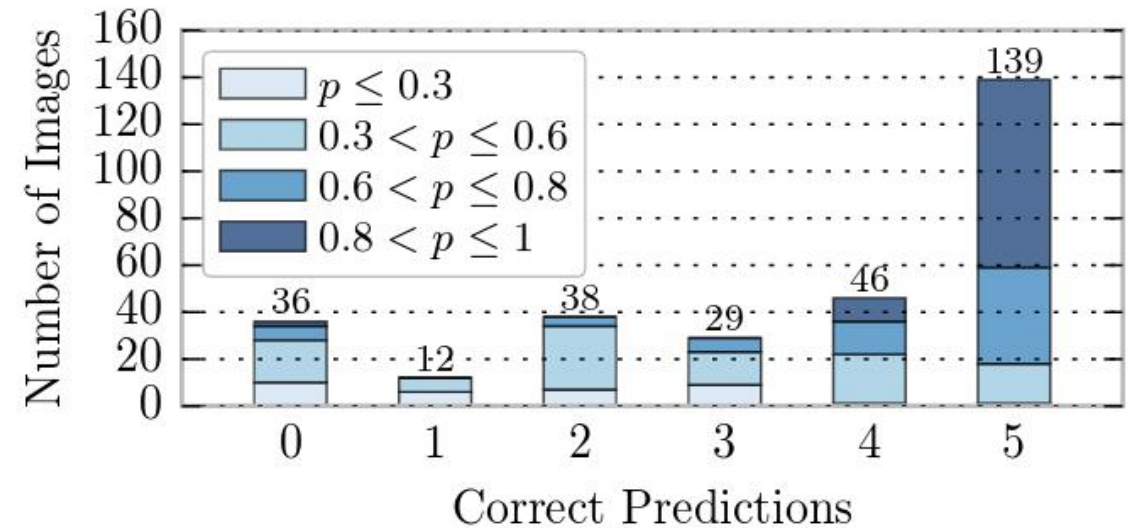


Figure 8. Histogram of correct labels in the top 5 predictions for a set of 300 manually verified images. Color indicates mean prediction confidence. Note that the confidence in prediction is strongly correlated with the accuracy.

# 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

Operator Agreement	Number of images	
	<i>Study 1</i>	<i>Study 2</i>
100%	50,425	9,938
75%	-	9
66%	-	8,535
50%	-	133
0%	22,847	1,300
<b>Average Agreement</b>	<b>69%</b>	<b>78%</b>

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



automotive	.999
gas & automotive	.999
shopping	.999
store	.999
vehicle dealer	.998



health & beauty	.992
health	.985
doctor	.961
emergency services	.960
dentist	.945

Dental



health & beauty	.935
<i>beauty</i>	<i>.925</i>
<i>cosmetics</i>	<i>.742</i>
<i>beauty salon</i>	<i>.713</i>
<i>hair care</i>	<i>.527</i>

Auto



automotive	.996
gas & automotive	.996
shopping	.995
store	.995
transportation	.985

# 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?

