Ontological Supervision for Fine Grained Classification of Street View Storefronts

Yair Movshovitz-Attias[1], Qian Yu[2], Martin C. Stumpe[2], Vinay Shet[2], Sacha Arnoud[2], Liron Yatziv[2],
[1]Computer Science Department, Carnegie Mellon University. [2]Google.

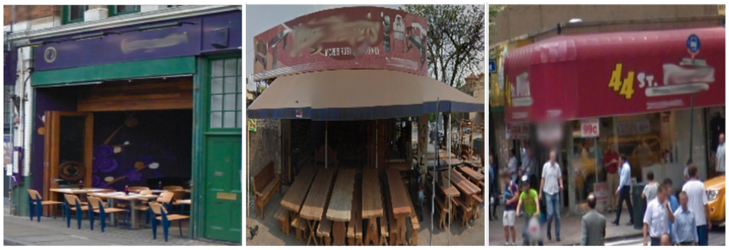


Figure 1: Examples of 3 businesses with their names blurred. Can you predict what they sell? Starting from left they are: Sushi Restaurant, Bench store, Pizza place. The intra-class variation can be bigger than the differences between classes. This example shows that the textual information in images can be important for classifying the business category. However, relying on OCR has many problems. In this paper we describe a method that classifies store fronts. We show our system implicitly learns to use text when it is available, but does not suffer from the same weaknesses.

Following the popularity of smart mobile devices, search engine users today perform a variety of locality-aware queries, such as *Japanese restaurant near me*, *Food nearby open now*, or *Asian stores in San Diego*. With the help of local business listings, these queries can be answered in a way that is tailored to the user's location.

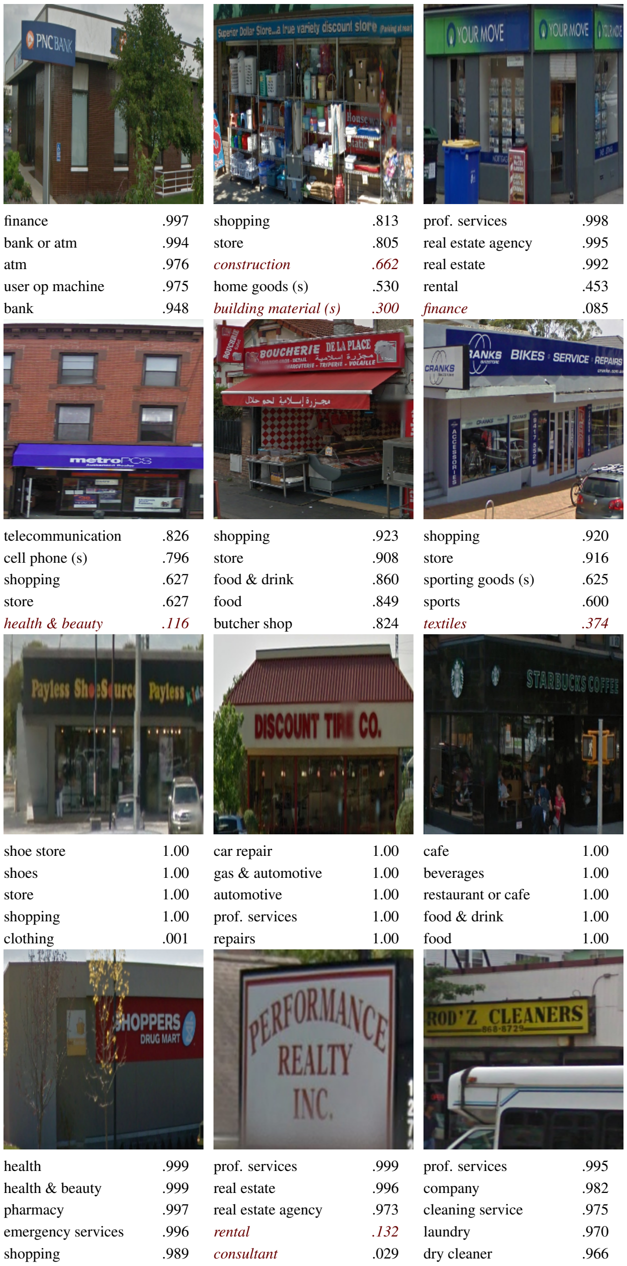
Creating accurate listings of local businesses is time consuming and expensive. To be useful for the search engine, the listing needs to be accurate, extensive, and importantly, contain a rich representation of the business category. Recognizing that a JAPANESE RESTAURANT is a type of ASIAN STORE that sells FOOD, is essential in accurately answering a large variety of queries. Listing maintenance is a never ending task as businesses often move or close down. In fact it is estimated that 10% of establishments go out of business every year, and in some segments of the market, such as the restaurant industry, the rate is as high as 30%.

The turnover rate makes a compelling case for automating the creation of business listings. For businesses with a physical presence, such as restaurants and gas stations, it is a natural choice to use data from a collection of street level imagery. Probably the most recognizable such collection is Google Street View which contains hundreds of millions of 360° panoramic images, with geolocation information.

In this work we focus on business storefront classification from street level imagery. We view this task as a form of multi-label fine grained classification. Given an image of a storefront, extracted from a Street View panorama, our system is tasked with providing the most relevant labels for that business from a large set of fine-grained labels. The problem is fine grained as business of different types can differ only slightly in their visual appearance. The discriminative information can be very subtle, and appear in varying locations and scales in the image; this, combined with the large number of categories needed to cover the space of businesses, require large amounts of training data.

The contribution of this work is two fold. First, we provide an analysis of challenges of a storefront classification system. We show that the intra-class variations can be larger than differences between classes (see Figure 1). Textual information in the image can assist the classification task, however, there are various drawbacks to text based models: Determining which text in the image belongs to the business is a hard task; Text can be in a language for which there is no trained model, or the language used can be different than what is expected based on the image location and we discuss these challenges in detail.

Finally, we propose a method for creating large scale labeled training data for fine grained storefront classification. We match street level imagery to known business information using both location and textual data extracted from images. We fuse information from an ontology of entities with geographical attributes to propagate category information such that each image is paired with multiple labels with different levels of granularity. Using this data we train a Deep Network based on the GoogLeNet architecture that achieves human level accuracy.



finance	.997	shopping	.813	prof. services	.998
bank or atm	.994	store	.805	real estate agency	.995
atm	.976	*construction*	*.662*	real estate	.992
user op machine	.975	home goods (s)	.530	rental	.453
bank	.948	*building material (s)*	*.300*	*finance*	.085

telecommunication	.826	shopping	.923	shopping	.920
cell phone (s)	.796	store	.908	store	.916
shopping	.627	food & drink	.860	sporting goods (s)	.625
store	.627	food	.849	sports	.600
health & beauty	*.116*	butcher shop	.824	*textiles*	*.374*

shoe store	1.00	car repair	1.00	cafe	1.00
shoes	1.00	gas & automotive	1.00	beverages	1.00
store	1.00	automotive	1.00	restaurant or cafe	1.00
shopping	1.00	prof. services	1.00	food & drink	1.00
clothing	.001	repairs	1.00	food	1.00

health	.999	prof. services	.999	prof. services	.995
health & beauty	.999	real estate	.996	company	.982
pharmacy	.997	real estate agency	.973	cleaning service	.975
emergency services	.996	*rental*	*.132*	laundry	.970
shopping	.989	*consultant*	.029	dry cleaner	.966

Figure 2: Top 5 predictions from our Deep Convolutional Network for sample images from the test set. Predictions in *red* disagree with ground truth labels and (s) is an abbreviation for store. The model predicts both generic and fine-grained categories.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.