Ontological Supervision for Fine Grained Classification of Street View Storefronts

Yair Movshovitz-Attias[*], Qian Yu[†], Martin C. Stumpe[†], Vinay Shet[†], Sacha Arnoud[†], Liron Yatziv[†]

[*]Carnegie Mellon University   [†]Google

yair@cs.cmu.edu   {qyu,mstumpe,vinayshet,sacha,lirony}@google.com

Abstract

Modern search engines receive large numbers of business related, local aware queries. Such queries are best answered using accurate, up-to-date, business listings, that contain representations of business categories. Creating such listings is a challenging task as businesses often change hands or close down. For businesses with street side locations one can leverage the abundance of street level imagery, such as Google Street View, to automate the process. However, while data is abundant, labeled data is not; the limiting factor is creation of large scale labeled training data. In this work, we utilize an ontology of geographical concepts to automatically propagate business category information and create a large, multi label, training dataset for fine grained storefront classification. Our learner, which is based on the GoogLeNet/Inception Deep Convolutional Network architecture and classifies 208 categories, achieves human level accuracy.

1. Introduction

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% [24].
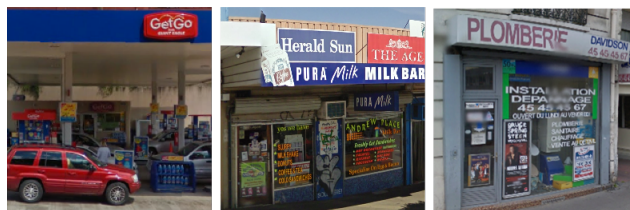

Figure 1. The multi label nature of business classification is clear in the image on the left; the main function of this establishment is to sell fuel, but it also serves as a convenience store. The remaining images show the fine grained differences one expects to find in businesses. The shop in the middle image is a grocery store, the one on the right sells plumbing supplies; visually they are similar.

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 labels. To understand the importance of associating a business with multiple labels, consider the gas station shown in Figure 1 (left). While its main purpose is fueling vehicles, it also serves as a convenience or grocery store. Any listing that does not capture this subtlety will be of limited value to its users. Similarly, stores like Target or Walmart sell a wide variety of products from fruit to home furniture, all of which should be reflected in their listings. The problem is fine grained as business of different types can differ only slightly in their visual appearance. An example of such a subtle difference is shown in Figure 1. The middle image shows the front of a grocery store, while the image on the right is of a plumbing supply store. Visually they are similar. The discriminative infor-

mation 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 2). 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 (see Figure 3). We discuss these challenges in detail in Section 3.

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 Convolutional Network that achieves human level accuracy.

2. Related Work

The general literature on object classification is vast. Object category classification and detection [9] has been driven by the Pascal VOC object detection benchmark [8] and more recently the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [26]. Here, we focus on reviewing related work on analysis of street view data, fine-grained classification and the use of Deep Convolutional Networks.

Analyzing Street View Data. Since its launch in 2007, Google Street View [28, 1] has been used by the computer vision community as both a test bed for algorithms [19, 31] and a source from which data is extracted and analyzed [12, 34, 21, 10, 6].

Early work on leveraging street level imagery focused on 3D reconstruction and city modeling. Cornelis et al. [6] focused on supplying textured 3D city models at ground level for car navigation system visualizations. Micusik et al. [21] used image segmentation cues and piecewise planar structures to build a robust 3D modeling system.

Later works have focused on extracting knowledge from Street View and leveraging it for particular tasks. In [34] the authors presented a system in which SIFT descriptors from $100,000$ Street View images were used as reference data to be queried upon for image localization. Xiao et al. [31] proposed a multi view semantic segmentation algorithm that classified image pixels into high level categories such as ground, building, person, etc. Lee et al. [19] described a

weakly supervised approach that mined midlevel visual elements and their connections in geographic data sets. Their approach discovered elements that vary smoothly over location. They evaluated their method using Street View images from the eastern coast of the United States. Their classifiers predicted location with a resolution of about 70 miles.

Most similar to our work, is that of Goodfellow et al. [12]. Both works utilize Street View as a map making source, and data mine information about real world objects. They focused on understanding street numbers, while we are concerned with local businesses. They described a method for street number transcription in Street View data. Their approach unified the localization, segmentation, and recognition steps by using a Deep Convolutional Network that operated directly on image pixels. The key idea behind their approach was to train a probabilistic model $P(S|X)$, where S is a digit sequence, and X an image patch, by maximizing $\log P(S|X)$ on a large training set. Their method, which was evaluated on tens of millions of annotated street number images from Street View, achieved above 90% accuracy and was comparable to human operators.

Fine Grained Classification. Recently there has been renewed interest in Fine Grained classification [32, 33, 14] Yao et al. [33] modeled images by densely sampling rectangular image patches, and the interactions between pairs of patches, such as the intersection of the feature vectors of two image patches. In [32] the authors proposed a codebook-free representation which samples a large number of random patches from training images. They described an image by its response maps to matching the template patches. Branson et al. [4] and Wah et al. [29] proposed hybrid human-computer systems, which they described as a visual version of the *20-question game*. At each stage of the game, the algorithm chooses a question based on the content of the image, and previous user responses.

Convolutional Networks. Convolutional Networks [11, 18] are neural networks that contain sets of nodes with tied parameters. Increases in size of available training data and availability of computational power, combined with algorithmic advances such as piecewise linear units [16, 13] and dropout training [15] have resulted in major improvements in many computer vision tasks. Krizhevsky et al. [17] showed a large improvement over state of the art in object recognition. This was later improved upon by Zeiler and Fergus [35], and Szegedy et al. [27].

On immense datasets, such as those available today for many tasks, overfitting is not a concern; increasing the size of the network provides gains in testing accuracy. Optimal use of computing resources becomes a limiting factor. To this end Dean et al. developed DistBelief [7], a distributed, scalable implementation of Deep Neural Networks. We base our system on this infrastructure.

Figure 2. 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 as discussed in Section 3.

3. Challenges in Storefront Classification

Large Within-Class Variance. Predicting the function of businesses is a hard task. The number of possible categories is large, and the similarity between different classes can be smaller than within class variability. Figure 2 shows three business storefronts. Their names have been blurred. Can you tell the type of the business without reading its name? Two of them are restaurants of some type, the third sells furniture, in particular store benches (middle image). It is clear that the text in the image can be extremely useful for the classification task in these cases.

Extracted Text is Often Misleading. The accuracy of text detection and transcription in real world images has increased significantly over the last few years [30, 22], but relying on the ability to transcribe text has drawbacks. We would like a method that can scale up to be used on images captured across many countries and languages. When using extracted text, we need to train a dedicated model per language, this requires a lot of effort in curating training data. Operators need to mark the location, language and transcription of text in images. When using the system it would fail if a business had a different language than what we expect for its location or if we are missing a model for that language (Figure 3a). Text can be absent from the image, and if present can be irrelevant to the type of the business. Relying on text can be misleading even when the language model is perfect; the text can come from a neighboring business, a billboard, or a passing bus (Figure 3b). Lastly, panorama stitching errors may distort the text in the image and confuse the transcription process (Figure 3c).

However, it is clear that the textual parts of the image do contain information that can be helpful. Ideally we would want a system that has all the advantages of using text information, without the drawbacks mentioned. In Section 6.3 we show that our system implicitly learns to use textual cues, but is more robust to these errors.

Business Category Distribution. The natural distribution of businesses in the world exhibits a "long tail". Some busi-

(a) Unexpected Language   (b) Misleading Text   (c) Stitching Errors

Figure 3. Text in the image can be informative but has a number of characteristic points of failure. (a) Explicitly transcribing the text requires separate models for different languages. This requires maintaining models for each desired language/region. If text in one language is encountered in a an area where that language was not expected, the transcription would fail. (b) The text can be misleading. In this image the available text is part of the Burger King restaurant that is behind the gas station. (c) Panorama stitching errors can corrupt text and confuse the transcription process.

(a) Area Too Small   (b) Area Too Large   (c) Multiple Businesses

Figure 4. Common mistakes made by operators: a red box shows the area marked by an operator, a green box marks the area that should have been selected. (A) Only the signage is selected. (B) An area much larger than the business is selected. (C) Multiple businesses are selected as one business.

nesses (e.g. McDonalds) are very frequent, but most of the mass of the distribution is in the large number of businesses that only have one location. The same phenomena is also true of categories. Some labels have an order of magnitude more samples than others. For example, for the FOOD AND DRINK category which contains restaurants, bars, cafes, etc, we have 300,000 images, while for LAUNDRY SERVICE our data contains only 13,000 images. We note that a large part of the distribution's mass is in smaller categories.

Labeled Data acquisition. Acquiring a large set of high quality labeled data for training is a hard task in and of itself. We provide operators with Street View panoramas captured at urban areas in many cities across Europe, Australia, and the Americas. The operators are asked to mark image areas that contain business related information. We call these areas *biz-patches*. This process is not without errors. Figure 4 shows a number of common mistakes made by operators. The operators might mark only the business signage (4a), an area that is too large and contains unneeded regions (4b), multiple businesses in the same biz- patch (4c).

4. Ontology Based Generation of Training Data

Learning algorithms require training data. Deep Learning methods in particular are known for their need of large quantities of training instances, without which they overfit. In this section we describe a process for collecting a large scale training set, coupled with ontology-based labels.

Building a training set requires matching extracted biz-patches p and sets of relevant category labels. First, we match a biz-patch with a particular business instance from a database of previously known businesses $\mathcal{B}$ that was manually verified by operators. We use the textual information and geographical location of the image to match it to a business. We detect text areas in the image, and transcribe them using an OCR software. This process suffers from the drawbacks of extracting text, but is useful for creating a set of candidate matches. This provides us with a set $\mathcal{S}$ of text strings. The biz-patch is geolocated and we combine the location information with the textual data. For each known business $b \in \mathcal{B}$, we create the same description, by combining its location and the set $\mathcal{T}$ of all the textual information that is available for it; name, phone number, operating hours, etc. We decide that p is a biz-patch of b if geographical distance between them is less than approximately one city block, and enough extracted text from $\mathcal{S}$ matches $\mathcal{T}$.

Using this technique we create a set of 3 million pairs (p, b). However, due to the factors that motivated our work, the quality and completeness of the information varies greatly between businesses. For many businesses we do not have category information. Moreover, the operators who created the database were inconsistent in the way they selected categories. For example, a McDonalds can be labeled as a HAMBURGER RESTAURANT, a FAST FOOD RESTAURANT, a TAKE AWAY RESTAURANT, etc. It is also plausible to label it simply as RESTAURANT. Labeling similar businesses with varying labels will confuse the learner.

We address this in two ways. First, by defining our task as a multi label problem we teach the classifier that many categories are plausible for a business. This, however, does not fully resolve the issue – When a label is missing from an example, the image is effectively used as a *negative* training instance for that label. It is important that training data uses a consistent set of labels for similar businesses. Here we use a key insight: the different labels used to describe a business represent different levels of specificity. For example, a hamburger restaurant *is a* restaurant. There is a containment relationship between these categories. Ontologies are a commonly used resource, holding hierarchical representations of such containment relations [3, 23]. We use an ontology that describes containment relationships between entities with a geographical presence, such as RESTAURANT, PHARMACY, and GAS STATION. Our
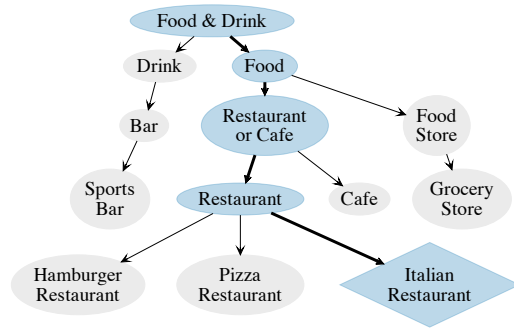

Figure 5. Using an ontology that describes relationships between geographical entities we assign labels at multiple granularities. Shown here is a snippet of the ontology. Starting from the ITALIAN RESTAURANT concept (diamond), we assign all the predecessors' categories as labels as well (shown in blue).

ontology, which is based on Google Map Maker's ontology, contains over 2,000 categories. For a pair (p, b) for which we know the category label c, we locate c in the ontology. We follow the containment relations described by the ontology, and add higher-level categories to the label set of p. The most general categories we consider are: ENTERTAINMENT & RECREATION, HEALTH & BEAUTY, LODGING, NIGHTLIFE, PROFESSIONAL SERVICES, FOOD & DRINK, SHOPPING. Figure 5 shows an illustration of this process on a snippet from the ontology. Starting from an ITALIAN RESTAURANT, we follow containment relations up predecessors in the ontology, until FOOD & DRINK is reached.

This creates a large set of pairs (p, s) where p is a biz-patch and s is a matching set of labels with varying levels of granularity. To ensure there is sufficient training data per label we omit labels whose frequency is very low and are left with 1.3 million biz-patches and 208 unique labels.

5. Model Architecture and Training

We base our model architecture on the winning submission for the ILSVRC 2014 classification and detection challenges by Szegedy et al. named GoogLeNet [27]. The model expands on the Network-in-Network idea of Lin et al. [20] while incorporating ideas from the theoretical work of Arora et al. [2]. Szegedy et al. forgo the use of fully connected layers at the top of the network and, by forcing the network to go through dimensionality reduction in middle layers, they are able to design a model that is much deeper than previous methods, while dramatically reducing the number of learned parameters. We employ the DistBelief [7] implementation of deep neural networks to train the model in a distributed fashion.

We create a train/test split for our data such that 1.2 million images are used for training the network and the remaining 100,000 images are used for testing. As a busi-
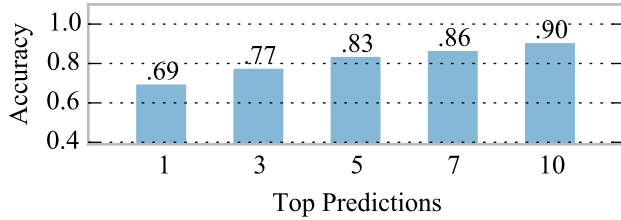
(a) Accuracy at K
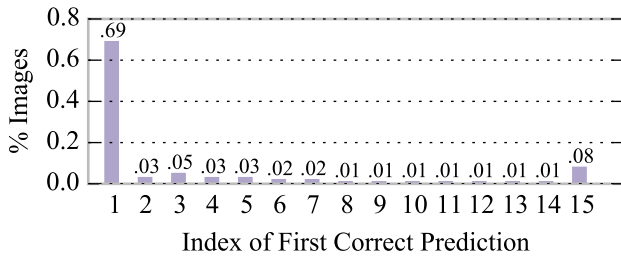
(b) First Correct Prediction

Figure 6. (a) Accuracy of classification for top K predictions. Using the top-1 prediction our system is comparable to human operators (see Table 1). When using the top 5 predictions the accuracy increases to 83%. (b) Percentage of images for which the first correct prediction was at rank K. To save space the values for $K \geq 15$ are summed and displayed at the 15th bin.

ness can be imaged multiple times from different angles, the splitting is location aware. We utilize the fact that Street View panoramas are geotagged. We cover the globe with two types of tiles. Big tiles with an area of 18 kilometers, and smaller ones with area of 2 kilometers. The tiling alternates between the two types of tiles, with a boundary area of 100 meters between adjacent tiles. Panoramas that fall inside a big tile are assigned to the training set, and those that are located in the smaller tiles are assigned to the test set. This ensures that businesses in the test set were never observed in the training set while making sure that training and test sets were sampled from the same regions. This splitting procedure is fast and stable over time. When new data is available and a new split is made, train/test contamination is not an issue as the geographical locations are fixed. This allows for incremental improvements of the system over time.

We first pre-train the network using images and ground truth labels from the ImageNet large scale visual recognition challenge with a Soft Max top layer, and once the model has converged we replace the top layer, and continue the training process with our business image data. This pre-training procedure has been shown to be a powerful initialization for image classification tasks [25, 5]. Each image is resized to 256×256 pixels. During training random crops of size 220×220 are given to the model as training images. We normalize the intensity of the images, add random photometric changes and create mirrored versions of the images

to increase the amount of training data and guide the model to generalize. During testing a central box of size 220×220 pixels is used as input to the model. We set the network to have a dropout rate of 70% (each neuron has a 70% chance of not being used) during training, and use a Logistic Regression top layer. Each image is associated with all the labels found by the method described in Section 4. This setup is designed to push the network to share features between classes that are on the same path up the ontology.

6. Evaluation

In this section we describe our experimental results. We begin by providing a quantitative analysis of the system's performance, then describe two large scale human performance studies that show our system is competitive with the accuracy of human operators and conclude with quantitative results that provide understanding as to what features the system managed to learn.

When building a business listing it is important to have very high accuracy. If a listing contains wrong information it will frustrate its users. The requirements on coverage however can be less strict. If the category for some business images can not be identified, the decision can be postponed to a later date; each street address may have been imaged many times, and it is possible that the category could be determined from a different image of the business. Similarly to the work of Goodfellow et al. [12] on street number transcription, we propose to evaluate this task based on recall at certain levels of accuracy rather than evaluating the accuracy over all predictions. For automatically building listings we are mainly concerned with recall at 90% precision or higher. This allows us to build the listing incrementally, as more data becomes available, while keeping the overall accuracy of the listing high.

6.1. Fine Grained Classification Results

As described in section 4 each image is associated with one or more labels. We first evaluate the classifier's ability to retrieve at least one of those labels. For an image i, we define the ground truth label set g_i. The predictions p_i are sorted by the classifier's confidence, and we define the top-k prediction set p_i^k as the first k elements in the sorted prediction list. A prediction for image i is considered correct if $g_i \cap p_i^k \neq \emptyset$. Figure 6a shows the prediction accuracy as a function of labels predicted. The accuracy at top-k is shown for $k \in \{1, 3, 5, 7, 10\}$. Top-1 performance is comparable to human annotators (see Section 6.2), and when the top 5 labels are used the accuracy increases to 83%. Figure 6b shows the distribution of first-correct-prediction, i.e. how far down the sorted list of predictions does one need to search before finding the first label that appears in g_i. We see that the first predicted label is by far the most likely and that the probability of having a predicted set p_i^k that does not
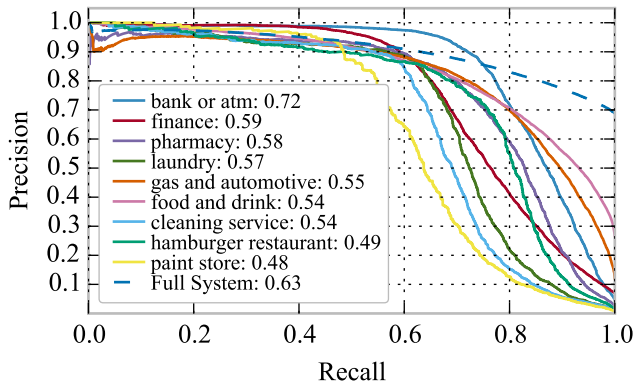
Figure 7. Precision recall curves for some of the top performing categories. The precision curve of the full system is shown as a dashed line. Recall at 90% precision is shown in the legend.

contain any of the true labels decreases with k. In order to save space we sum up all the probabilities for $k \in [15, 208]$ in one bin.

As mentioned above, one important metric for evaluation is recall at specific operating points. Figure 7 shows precision recall curves for some of the top performing categories and summary curve of the full system (dashed). The precision and recall of the full system is calculated by using the top-1 prediction. For many categories we are able to recover the majority of businesses while precision is held at 90%. For a classification system to be useful in a practical setting the classifier's returned confidence must be well correlated with the quality of its prediction. Figure 8 shows a histogram of the number of correctly predicted labels in the top 5 predictions on a set of images whose labels were manually verified. The mean prediction confidence is indicated by color intensity (darker means higher confidence). Note the strong correlation between confidence and accuracy; for confidence above 80% normally at least 4 of the top labels are correct.

The GoogLeNet network [27] incorporates a number of new design elements that make it particularly appealing: by not using fully connected upper layers, and forcing the network to go through dimensionality reduction stages, the network has far fewer parameters while being much deeper than previous methods. We evaluate the use of this architecture by comparing it to the AlexNet network proposed by Krizhevsky et al. [17] on a 13 category, *single label* classification task. We select a set of useful categories that a user might search for, and train both networks to predict one label per image. Figure 9 shows recall at 0.9 precision on all categories for which at least one method has recall ≥ 0.1. GoogLeNet outperforms the baseline for all categories, and for some, e.g. PHARMACY, recall more than doubles.

Figure 10 shows 30 sample images from the test set with their top 5 predictions. The model is able to classify these images with their high level categories and with fine
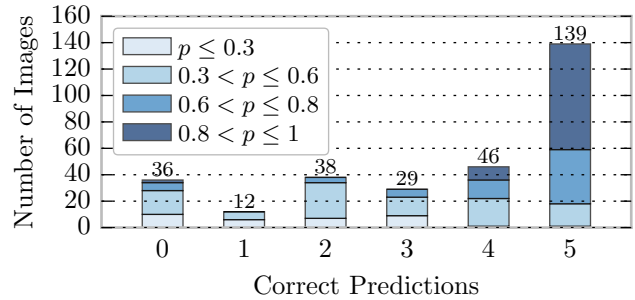
Figure 8. Histogram of correct labels in the top 5 predictions for a set of 300 manually verified images. Color indicates mean prediction confidence. Note that the confidence in prediction is strongly correlated with the accuracy.
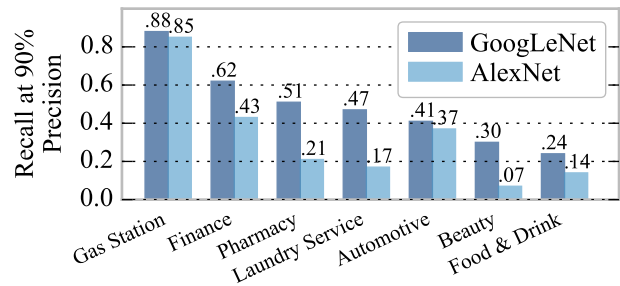

Figure 9. Recall at 90% precision for the GoogLeNet model used by our method and the AlexNet architecture [17] on a 13 category classification task. We show classes for which at least one of the models has recall ≥ 0.1. GoogLeNet performs better on all categories. On some, such as PHARMACY recall more than doubles.

grained, specialized classes. For example, the top-right image was classified by the model as: BEAUTY, BEAUTY SALON, COSMETICS, HEALTH SALON, and NAIL SALON.

6.2. Human Performance Studies

Our system needs to perform at human level accuracy. Otherwise it will require a human verification post process. To estimate human accuracy on this task we have conducted two large scale human performance studies that check the agreement of operator-provided labels for the same business. In our experiments, subjects were shown images of businesses and selected one of 13 options (12 categories, and an OTHER category that stands for "none of the above".) Categories were chosen based on their financial impact for a business listing creator. The categories are: FOOD STORE, RESTAURANT/CAFFE, FINANCE, PHARMACY, REAL ESTATE, GAS STATION, AUTOMOTIVE (other than gas stations), FASHION, LODGING, LAUNDRY, PLACE OF WORSHIP, BEAUTY. Note that the studies used full resolution images as opposed to the 256×256 images given to the algorithm. The first study had $73,272$ images, each was shown to two operators. The operators agreed on 69% of image labels. In our second

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

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.

study we had 20,000 images, but each image was shown to *three* to *four* subjects. We found that the average agreement of the human operators was 78%. Table 1 shows a detailed summary of the human study results.

6.3. Analysis: Learning to Read With Weak Labels

For some of the images in Figure 10, such as the image of the dental center (top row, second image from right) it is surprising that the model was capable of classifying it correctly. It is hard to think of a "canonical" dental center look, but even if we could, it doesn't seem likely that this image would be it. In fact, without reading the text, it seems impossible to correctly classify it. This suggests that the system has learned to use text when needed. Figure 11 shows images from Figure 10 for which we have manually blurred the discriminative text. Note especially the image of the dental center, and of the auto dealer. After blurring the test "Dental" the system is confused about the dental center; it believes it is a beauty salon of some sorts. However, for the auto dealer, it is still confident that this is a place that sells things, and is related to transportation and automotive. To take this experiment to its logical extreme, we also show a synthetic image, which contains only the word *Pharmacy*. The classifier predicts the relevant labels for it.

To us this is a compelling demonstration of the power of Deep Convolutional Networks to learn the correct representation for a task. Similar to a human in a country in which she does not know the language, it has done the best it can – learn that some words are correlated with specific types of businesses. Note that it was never provided with annotated text or language models. It was only provided with what we would consider as *weak textual labels*, images that contain text and labeled with category labels. Furthermore, when text is not available the system is able to make an accurate prediction if there is distinctive visual information.

7. Discussion

Business category classification, is an important part of location aware search. In this paper we have proposed a method for fine grained, multi label, classification of busi-

Dental		**Auto**		**Nail**	
health & beauty	.935	automotive	.996	health & beauty	.894
beauty	.925	gas & automotive	.996	beauty	.891
cosmetics	.742	shopping	.995	cosmetics	.800
beauty salon	.713	store	.995	beauty salon	.799
hair care	.527	transportation	.985	*hair*	.407

Liquor store		**McDonald's**			
food & drink	.833	food & drink	.998	health & beauty	.987
food	.745	food	.996	shopping	.985
restaurant or cafe	.717	restaurant or cafe	.992	store	.982
restaurant	.667	restaurant	.990	health	.981
beverages	.305	fast food restaurant	.862	pharmacy	.969

Figure 11. A number of images from Figure 10 with the discriminative text in the image blurred (noted above the image). For some images, without the discriminative word the algorithm is confused (left column). For example, for the dental center, without the word *dental* it predicts a beauty salon. For other images, there is enough non textual information for the algorithm to be confident of the business category even when the text is blurred, for example the car dealership. Note the image of the nail spa: when the word *nail* is blurred the classifier falls back to more generic classes that fit the visual information - beauty salon, cosmetics, etc. As a final indicator to the ability of the network to learn textual cues we show a synthetic image where the only visual information is the word *pharmacy*. The network predicts relevant labels.

ness storefronts from street level imagery. We show that our system learned to extract and associate text patterns in multiple languages to specific business categories without access to explicit text transcriptions. Moreover, our system is robust to the absence of text, and when distinctive visual information is available, it is able to make correct predictions. We show our system achieves human level accuracy.

Using an ontology of entities with geographical attributes, we propagate label information during training, and produce a large set of 1.3 million images for a fine grained, multi label task. The use of non visual information, such as an ontology, to "ground" image data to real world entities is an exciting research direction, and there is much that can be done. For example, propagating information using the ontology at test time can increase both accuracy and recall. Node similarity in the ontology can be used to guide feature sharing between classes, and improve performance for seldom viewed classes.

Figure 10. Top predictions for sample images from the test set. Predictions marked in *red* disagree with ground truth labels; in some cases the classifier is correct and the ground truth label is wrong (marked with *). For example, see asian restaurant (fourth image from left, row before last), for which the top 5 predictions are all correct, but do not agree with the ground truth. The mark (s) is an abbreviation for store.

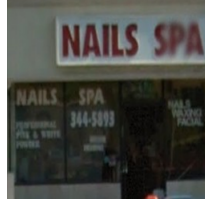|					
finance .997	shopping .813	prof. services .998	automotive .999	health & beauty .992	beauty .997
bank or atm .994	store .805	real estate agency .995	gas & automotive .999	health .985	beauty salon .997
atm .976	*construction* .662	real estate .992	shopping .999	doctor .961	cosmetics .995
user op machine .975	*building material (s)* .300	rental .453	store .999	emergency services .960	health salon .994
bank .948		*finance* .085	vehicle dealer .998	dentist .945	nail salon .953
telecommunication .826	shopping .923	shopping .920	laundromat .934	food & drink .947	automotive .999
cell phone (s) .796	store .908	store .916	cleaners .795	food .867	gas & automotive .999
shopping .627	food & drink .860	sporting goods (s) .625	prof. services .732	restaurant or cafe .722	repairs .999
store .627	food .849	sports .600	laundry .679	restaurant .621	prof. services .999
health & beauty .116	butcher shop .824	*textiles* .374	cleaning service .669	beverages .441	car repair .998
shoe store 1.00	car repair 1.00	cafe 1.00	food & drink 1.00	liquor store 1.00	food & drink .999
shoes 1.00	gas & automotive 1.00	beverages 1.00	food 1.00	beverages .999	food .998
store 1.00	automotive 1.00	restaurant or cafe 1.00	restaurant or cafe .999	shopping .998	restaurant or cafe .884
shopping 1.00	prof. services 1.00	food & drink 1.00	restaurant .999	store .998	restaurant .995
clothing .001	repairs 1.00	food 1.00	hamburger restaurant .936	food & drink .700	fast food restaurant .884
health .999	prof. services .999	prof. services .995	*food & drink** .996	food & drink .825	shopping .932
health & beauty .999	real estate .996	company .982	*food** .959	food .762	store .920
pharmacy .997	real estate agency .973	cleaning service .975	*restaurant** .931	restaurant or cafe .741	florist .896
emergency services .996	*rental* .132	laundry .970	*restaurant or cafe** .909	restaurant .672	*fashion* .077
shopping .989	*consultant* .029	dry cleaner .966	*asian** .647	beverages .361	*gift shop* .071
prof. services .594	gas station .996	shopping .489	shopping .719	beauty .999	place of worship .990
legal services .346	transportation .996	store .467	store .713	health & beauty .999	church .988
lawyer .219	gas & automotive .995	*prof. services* .289	home goods (s) .344	cosmetics .998	*education/culture** .031
insurance .129	*government* .001	*services* .246	furniture store .299	health salon .998	*assoc./organization** .029
insurance agency .103	*gastronomy* .001	*gas & automotive* .219	mattress store .240	nail salon .949	*prof. services* .027

Acknowledgments

The authors wish to thank Ian Goodfellow, Christian Szegedy, and Yaser Sheikh for helpful discussions.

References

[1] D. Anguelov, C. Dulong, D. Filip, C. Frueh, S. Lafon, R. Lyon, A. Ogale, L. Vincent, and J. Weaver. Google street view: Capturing the world at street level. *Computer*, 2010. 2

[2] S. Arora, A. Bhaskara, R. Ge, and T. Ma. Provable bounds for learning some deep representations. *arXiv:1310.6343 [cs, stat]*, oct 2013. 4

[3] J. Bhogal, A. Macfarlane, and P. Smith. A review of ontology based query expansion. *Information processing & management*, 43(4):866–886, 2007. 4

[4] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *European Conference of Computer Vision (ECCV)*. 2010. 2

[5] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *British MAchine Vision Conference (BMVC)*, 2014. 5

[6] N. Cornelis, B. Leibe, K. Cornelis, and L. Van Gool. 3d urban scene modeling integrating recognition and reconstruction. *International Journal of Computer Vision*, 2008. 2

[7] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. aurelio Ranzato, A. Senior, P. Tucker, K. Yang, Q. V. Le, and A. Y. Ng. Large scale distributed deep networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1223–1231. 2012. 2, 4

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html. 2

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2010. 2

[10] A. Flores and S. Belongie. Removing pedestrians from google street view images. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010. 2

[11] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4), 1980. 2

[12] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks. *arXiv preprint arXiv:1312.6082*, 2013. 2, 5

[13] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. *arXiv preprint arXiv:1302.4389*, 2013. 2

[14] A. B. Hillel and D. Weinshall. Subordinate class recognition using relational object models. *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, 2007. 2

[15] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012. 2

[16] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009. 2

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 2012. 2, 6

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 1998. 2

[19] Y. J. Lee, A. A. Efros, and M. Hebert. Style-aware mid-level representation for discovering visual connections in space and time. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, 2013. 2

[20] M. Lin, Q. Chen, and S. Yan. Network in network. *CoRR*, abs/1312.4400, 2013. 4

[21] B. Micusik and J. Kosecka. Piecewise planar city 3d modeling from street view panoramic sequences. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009. 2

[22] P. X. Nguyen, K. Wang, and S. Belongie. Video text detection and recognition: Dataset and benchmark. In *Winter Conference on Applications of Computer Vision (WACV)*, Steamboat Springs, CO, March 2014. 3

[23] N. F. Noy. Semantic integration: a survey of ontology-based approaches. *ACM Sigmod Record*, 33(4):65–70, 2004. 4

[24] H. Parsa, J. T. Self, D. Njite, and T. King. Why restaurants fail. *Cornell Hotel and Restaurant Administration Quarterly*, 46(3):304–322, 2005. 1

[25] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. *CoRR*, 2014. 5

[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge, 2014. 2

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv:1409.4842 [cs]*, sep 2014. 2, 4, 6

[28] L. Vincent. Taking online maps down to street level. *Computer*, 2007. 2

[29] C. Wah, S. Branson, P. Perona, and S. Belongie. Multiclass recognition and part localization with humans in the loop. In *IEEE International Conference on Computer Vision (ICCV)*, 2011. 2

[30] K. Wang and S. Belongie. Word spotting in the wild. In *European Conference on Computer Vision (ECCV)*, Heraklion, Crete, Sept. 2010. 3

[31] J. Xiao and L. Quan. Multiple view semantic segmentation for street view images. In *Computer Vision, 2009 IEEE 12th International Conference on*, 2009. 2

[32] B. Yao, G. Bradski, and L. Fei-Fei. A codebook-free and annotation-free approach for fine-grained image categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2

[33] B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 2

[34] A. R. Zamir and M. Shah. Accurate image localization based on google maps street view. In *Computer VisionECCV 2010*. Springer, 2010. 2

[35] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional neural networks. *arXiv preprint arXiv:1311.2901*, 2013. 2