Business-Aware Visual Concept Discovery from Social Media for Multimodal Business Venue Recognition

Bor-Chun Chen
University of Maryland, USA
sirius@umd.edu

Yan-Ying Chen and **Francine Chen** and **Dhiraj Joshi**
FX Palo Alto Laboratory, Palo Alto, CA, USA
{yanying, chen, dhiraj}@fxpal.com

Abstract

Image localization is important for marketing and recommendation of local business; however, the level of granularity is still a critical issue. Given a consumer photo and its rough GPS information, we are interested in extracting the fine-grained location information, i.e. business venues, of the image. To this end, we propose a novel framework for business venue recognition. The framework mainly contains three parts. First, business-aware visual concept discovery: we mine a set of concepts that are useful for business venue recognition based on three guidelines including business awareness, visually detectable, and discriminative power. We define concepts that satisfy all of these three criteria as business-aware visual concept. Second, business-aware concept detection by convolutional neural networks (BA-CNN): we propose a new network configuration that can incorporate semantic signals mined from business reviews for extracting semantic concept features from a query image. Third, multimodal business venue recognition: we extend visually detected concepts to multimodal feature representations that allow a test image to be associated with business reviews and images from social media for business venue recognition. The experiments results show the visual concepts detected by BA-CNN can achieve up to 22.5% relative improvement for business venue recognition compared to the state-of-the-art convolutional neural network features. Experiments also show that by leveraging multimodal information from social media we can further boost the performance, especially when the database images belonging to each business venue are scarce.

Introduction

Nowadays, there are a sheer amount of images being uploaded to social media sites on the web everyday. Although some of the images contain check-in information that discloses at which business venues they were taken, many of the images do not have such information available. For example, the images uploaded to Flickr or Google Photos only contain GPS information but no check-in information. Even for images which have check-in information, most check-ins are famous travel landmarks while very few of them are local business venues. There arises an interesting research problem: given image content taken in some business venue and its GPS information, we aim to infer which venue the image was taken at.

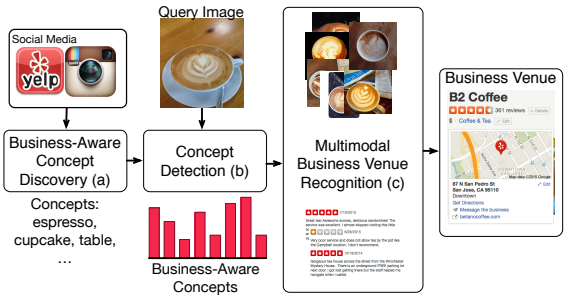

Figure 1: Given an image uploaded to social media and its rough GPS information, we want to automatically find out the business venue where it was taken. (a) We first mine a list of business-aware visual concepts from social media, (b) use the proposed BA-CNN to detect these business-aware visual concepts from the query image and (c) associate visual concepts with images and business reviews in a geo-tagged database to recognize the business venue.

Recognition of the business venue (e.g. cafe shop, local restaurant) in an image can help many applications for personalization and location-based services/marketing. For instance, it allows personalized promotion based on the business venue a user had visited, or accurate check-in suggestion in social media applications. One might think this is an easy task: since we already have the GPS information, we can just map it to the GPS information of business venue. However, GPS information is not accurate enough to achieve such fine-grained geo-localization tasks. According to experiments conducted in Maier and Kleiner (2010), modern GPS sensors can have up to 40 meter error, especially in the urban area. Hence, GPS can only help us narrow down the candidates within a nearby area, and we need a more reliable way to recognize the venue.

There are many previous works focusing on geo-localization based on matching visual content. However, most of the works only target on a coarser granularity of location (e.g., city), and they are only applicable for outdoor images while a huge portion of the images on social media websites are indoor images. The major challenge is – indoor images contain less unique visual patterns and many business venues have only a few images associated with them,

so it is hard to recognize location in such a fined-grained setting without any high-level semantic descriptions (e.g., coffee cups in the cafe). Some other previous works use text information to infer the user's location. However, these methods cannot deal with the cases when a query image is not associated with any texts and they do not utilize visual information, which can provide useful clues.

By leveraging freely available social media on the Internet, we propose a novel framework to address this challenging problem. As shown in Figure 1, our system mainly contains three parts: (1) Business-Aware Visual Concept Discovery: By mining large-scale social media text corpus, we discover a set of business-aware visual concepts that are useful for business venue recognition. (2) Business-Aware Visual Concept Detection: we detect the concepts from images using a novel convolutional neural network configuration (BA-CNN), and (3) Multimodal Business Venue Recognition: we then use Word Vector Model (Mikolov et al. 2013) to extend visually detected concepts to word representations and further combine with image content for multimodal venue recognition. Note that the extension of multimodal feature representations only relies on the visual content of a query image without being associated with any texts.

To sum up, the contributions of this paper include: (1) to the best of our knowledge, this is the first work to recognize business venues by using visual content in consumer photos; (2) we develop a systematic framework to automatically mine visually detectable business-aware concepts from reviews of local businesses; (3) we propose a novel CNN configuration to incorporate semantic signals mined from business reviews for training visual concept detector and extracting business-aware semantic features; (4) we extend a visual representation to multimodal feature representations – visual concepts and word vectors – to associate with multiple information sources on the Web for business venue recognition.

Related Work

Our work is closely related to several research directions. (1) Geo-location prediction: predicting the location information from an image or a short text description (i.e. tweets). (2) Visual concept detection: finding a semantic representation of an image. (3) Convolutional neural networks: learning visual representation based on a deep neural network. In the following section, we will discuss the related works in each area and the differences with our work.

Geo-location prediction

There are many related works for inferring the location from an image. Hays and Efros (2008) is one of the early studies that successfully infer geo-information from a single image. They use a simple data-driven approach to find geo-information based on a large-scale geo-tagged database. However, they only focus on outdoor images with coarse granularity up to city level. Schindler, Brown, and Szeliski (2007) is another early work on geo-location prediction, which focus on location recognition within a city. They developed an algorithm to select informative low-level features to improve the recognition accuracy in a large-scale setting. While their granularity is smaller, they only focus on street view images within a 20 kilometer range. In Friedland, Vinyals, and Darrell (2010), they use multimodal information to infer the geo-information of a video, but they only focus on city-scale granularity by using low-level feature such as SIFT features (Lowe 2004). In Fang, Sang, and Xu (2013), they tried to find discriminative image patches for city-level geo-location prediction. In Lin et al. (2015), they use aerial images to help geo-location prediction. While they can achieve a finer granularity, the technique can only apply to images of outdoor buildings. There are also many works that focus on landmark recognition (Zheng et al. 2009) (Li, Crandall, and Huttenlocher 2009), which is highly related to geo-location predication. However, these works relay on distinct low level visual patterns to recognize the landmarks. Note that in (Chen et al. 2011), they also use GPS information to assist the retrieval task, which is similar to our setting, but they only focus on landmark recognition.

Our work is different from the aforementioned works in many different aspects. (1) We focus on fine-grained business venue recognition, while most previous works only address city-level granularity. (2) We focus on consumer photos which contain both indoor and outdoor images, while most previous works can only deal with outdoor images. (3) We derive a semantic representation from the image content, which can be used to match the text information in the reviews of business venues available in a multimodal database.

There are also many works focusing on geo-location prediction based on texts in the social media (i.e. tweets): Chen et al. (2013) Chen et al. (2014a) Hulden, Silfverberg, and Francom (2015) DeLozier, Baldridge, and London (2015). However, text information is not always available and there might not be location-related information available in the texts. Therefore, texts and images can be viewed as complementary sources for geo-location prediction. In this work, we focus on the case where only an image is available as the query for business venue recognition.

Visual concept detection

Our work is also related to the research of visual concept detection. There are many previous works that address generic concepts discovery (Deng et al. 2009) (Li et al. 2010). However, these concepts are not mined for the purpose of business venue recognition, and therefore, as shown later in the experiments, do not perform well compared to our business-aware visual concepts.

Chen et al. (2014b) propose to mine semantic concepts from event description for event detection. Ye et al. (2015) further improve the concept definition by mining concepts from "WikiHow." Compared to these works, we have the following advantage: (1) We consider the discriminative power in terms of business categories while they define a separate set of concepts for each event. (2) We use the features learnt by CNN rather than hand crafted. The concept features in our work are further constrained by the labels of business venues, which incorporate the correlations of concepts associated with the same business venues. (3) We further represent each detected concept as a meaningful word vector that are learned by large-scale review corpus.

Convolutional neural networks

Convolutional neural networks have shown superior performance in many computer vision tasks (Razavian et al. 2014). Therefore, we adopt it for our visual concept detection. Our CNN configuration is developed based on the one in (Krizhevsky, Sutskever, and Hinton 2012), and implemented with open source framework named CAFFE (Jia et al. 2014). Different from the original network structure, our configuration is able to extract semantic concepts while maintain discriminative powers for business venue recognition.

Proposed Method

System overview

Our goal is to recognize the business venue by a single query image. This section introduces the major components of our system (cf. Figure 1): (a) Business-Aware Visual Concept Discovery: mining a list of business-aware visual concepts from a business review corpus. (b) Business-Aware Visual Concept Detection: using a novel CNN configuration to detect the semantic concepts from query images. (c) Multimodal Business Venue Recognition: extending visual concepts to multimodal representation for business venue recognition.

Business-Aware Visual Concept Discovery

We follow three guidelines to discover business-aware visual concepts: (1) *Business Awareness*: the relevance with business venues. For example, "earth" is not a good business-aware concept because it might not be commonly used in any of the business venues; on the other hand, "cat" might be a good business-aware concept because it could appears in local pet shops. (2) *Visually Detectable*: the detectability from visual content in an image. For instance, "disease" although usually appears at hospitals, is hard to be detected by image content, and thus not a good visual concept; on the other hand, "medicine" is a good visual concept because it has more consistent visual patterns for detection. (3) *Discriminability*: the discriminative power to distinguish between different business venues. For example, "person" might not have enough discriminability because it appears in general business venues, while "burger" could be a good concept as it appears more frequently in American restaurants. According to these three guidelines, we first introduce the approach of mining many candidate concepts from reviews of local businesses followed by selecting concepts with high accuracy of visual detection and low entropy across business venues. Figure 2 shows an overview of our method for business-aware visual concepts discovery.

Mining Candidate Concepts Following the guidelines mentioned above, we first mine the candidate concepts from reviews of local businesses on a social media website (i.e. Yelp) to ensure the property of business awareness. We first classify the business venues by their top-level category in the Yelp business category topology [1] (example categories include restaurants, active life, automotive, etc.) We then gather 3,000 reviews from each business category respectively. From each category, we select 500 frequent nouns based on their document frequency as our candidate concepts. Note that we use NLTK Toolkit (Bird, Klein, and Loper 2009) to tokenize the words in the reviews and find the part-of-speech tags. We only select the nouns as candidate concepts to ensure the concepts are more visually detectable. There are many overlapping concepts in each category and we find 2,143 concepts overall. In order to ensure the discriminability of the candidate concepts, we remove concepts that appears in more than ten different categories. We also remove concepts that are offensive terms that blocked by Instagram API and result in 1,723 concept candidates. Table 1 shows some candidate concepts found in each category.

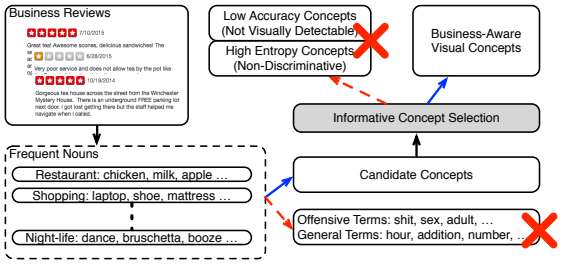

Figure 2: The overview for business-aware visual concept discovery. We first collect Yelp reviews and find frequent nouns in every business category, and then remove general terms (to every category) and offensive terms (blocked by Instagram) to construct a set of candidate concepts. Finally, we select concepts with visual consistency and low normalized entropy across locations.

Selecting Informative Concepts After finding candidate concepts, we need to select useful concepts for business venue recognition from an image. For each concept, we use it as keyword to retrieve 1,000 images from a social media website, i.e. Instagram. Since images downloaded from Instagram are quite noisy, we do two-fold cross validation by using convolutional neural networks (CNN) (Krizhevsky, Sutskever, and Hinton 2012) to select qualified images for learning accurate detectors of visual concepts.

The main idea of two-fold cross validation is – dividing the images into two sets, training a separate concept classifier for each set, and finally using each to verify images in the other set. We select top 250 images from each set based on the classification score for training the concept detectors. Figure 3 (a) shows the training data before the cross-validation selection for concept "pizza" while Figure 3 (b) shows the training data after cross-validation selection. We can see that the training data after selection are more visually consistent and therefore can achieve better accuracy for concept classification. The experiment in Table 2 shows that by cross-validation selection we can achieve up to 48.5% classification accuracy compared to 36.5% by simply using all images as training data. Finally, we remove concepts that have validation accuracy lower than 50% (using hash tag as ground-truth) to ensure the visual detectability of concepts.

We then further select the concepts with more discrim-

[1] https://www.yelp.com/developers/documentation

Table 1: Example candidate concepts in each category mined from reviews of local business.

Category	# of Concepts	Example Candidate Concepts
Restaurants	233	chicken, milk, apple, sashimi, onion, tea, chef, pasta, waiter, pizza
Pets	190	doctor, vet, furry, tail, adoption, cage, toy, cat, doggie, salon
Automotive	184	motorcycle, windshield, carpet, auto, girlfriend, stereo, wheel, gas, tank, dealership

(a) Noisy Images (b) Clean Images

Figure 3: (a) Images crawled from Instagram by the hash tag "pizza." (b) Images selected by cross-validation that are more visually consistent and correctly represent the visual concept.

Table 2: Accuracy of concept classifiers trained by all images (All), randomly selected images (Random) and the images selected by cross-validation (CRV). Note that the accuracy involves the concepts that are less visually detectable. After concept selection, CRV can reach 85% accuracy.

Training Data	All	Random	CRV
Rank-1 Accuracy	36.5%	38.7%	**48.5%**

inative power by computing the cross-location normalized entropy using the following formula:

$$\eta(X^{(c)}) = -\sum_{i=1}^{n(c)} \frac{p(x_i^{(c)}) \log_2(p(x_i^{(c)}))}{\log_2(n^{(c)})}, \quad (1)$$

where X is a random variable that denotes the venue distribution of concept c. $\eta(X^{(c)})$ is the normalized entropy for that concept. $n^{(c)}$ is the total number of business venues that have concept c and $p(x_i(c))$ is the probability of the concept appears in a business venue i. We prepared a dataset from Instagram that contains 250,000 images associated with 1,000 different business venues and computed the normalized entropy for each concept in terms of its distribution over business venues. Finally, the 490 concepts with the lowest entropy value are selected as business-aware visual concepts for business venue recognition. Figure 4 shows some example concepts and corresponding images.

Convolutional Neural Networks for Business-Aware Concepts (BA-CNN)

Convolutional Neural Networks have shown promising results in many computer vision related problems. Here we adopt the state-of-the-art visual features learned by CNN (Krizhevsky, Sutskever, and Hinton 2012) as a baseline for business venue recognition. Note that because of (1) scalability: too many business venues and (2) sparsity: only a few images for most business venues (cf. Figure 6), we cannot

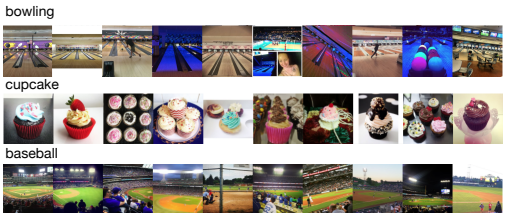

Figure 4: Example concepts and corresponding images.

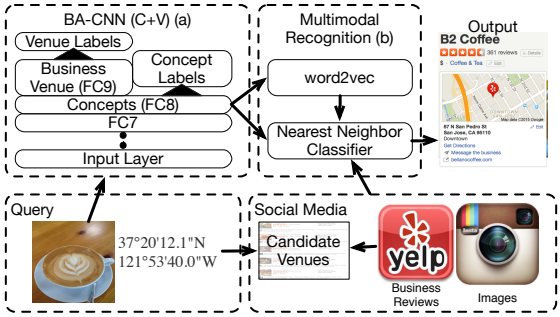

Figure 5: System framework for multimodal business venue recognition. Given an query image, we first find a list of candidate venues from social media using GPS, and detect business-aware concepts from image content using BA-CNN (C+V). We then use a Word Vector model to generate the text representation. The visual concept scores and text representation of the query image are then matched against those extracted from the reviews and images in the database. The business venue associated with the best-matched images and reviews is returned as the most likely business venue.

directly train the classifiers to distinguish different business venues. Instead, we learn the features supervised by different types of labels at the output layer of an CNN, and use the activations from the last fully-connected layer (FC7) before the output layer as the features to represent an image. The types of labels could be: general concepts used in ImageNet (ImageNet-CNN), business-aware concepts (BA-CNN (C)) and a subset of business venues (BA-CNN (V)). The comparisons of different types of labels are presented in the experiments later. Finally, we apply nearest neighbor classifier based on the CNN features of an query image and database images. The business venue associated with the most similar database image is output as the predicted business venue. Note that the GPS of the query image is used to narrow down the candidate business venues. The impact from the number of candidates is discussed in the Experiments section.

However, simply use CNN features may suffer from several problems. For ImageNet-CNN (i.e. a network trained on ImageNet labels), the concepts are predefined and not rele-

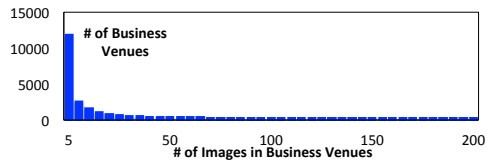

Figure 6: The number of images in each business venue sampled from social media ($> 50\%$ venues have < 5 images).

vant to local businesses; for BA-CNN (C) the discriminability only lies in separating different business-aware concepts rather than business venues; finally, BA-CNN (V) the business venues are limited to the venues comprising more training images and thus cannot cover general business venues. Furthermore, the common problem of CNN features is – they do not have semantic meaning, which is a key property to associate with other data domains.

To address these issues, we propose a new CNN configuration (BA-CNN (C+V)) to detect business-aware concepts for business venue recognition. As shown in Figure 5 (a), instead of using FC7 for recognition, we let layer (FC8) supervised by business-aware concept labels and add another layer (FC9) on top of the concept layer supervised by a subset of business venue labels. This way, we can extract features from FC8, where each dimension corresponds to a business-aware visual concept, and has the discriminative power to separate different business venues. In our experiments, BA-CNN (C+V) is demonstrated with a higher recognition accuracy compared to the other CNN features extracted from images. Moreover, it is able to associate multimodal data (e.g., text and images) for recognition since the features extracted by BA-CNN (C+V) are the responses of semantically describable concepts.

Multimodal Business Venue Recognition

Once we have the concept representation detected by BA-CNN, we can use it for business venue recognition. However, we want to further improve the recognition accuracy by extending image content to multimodal representations – visual concepts and text representation, to utilize the text information, i.e. business review, of the business venues available on the social media. Figure 5 shows our system framework for multimodal business venue recognition.

We first use review text of local businesses (e.g. Yelp reviews) to train word vector model (Mikolov et al. 2013) that can convert each word into a 500-dimensional vector representation. For each query image, we use the top-5 visual concepts detected from the query image as concept words and average the word vector representation of the top-5 concepts to represent another modality of the image. As shown in Figure 5 (b), visual concept representation and word vector representation are then fused together to form the final representation. Here we simply use early fusion (i.e. concatenate the 490 dimensional concept representation and 500 dimensional word vector representation together to form a 990 dimensional vector) to combine two modalities. Similarly, the images and reviews associated to business venues in the databases are also represented as visual concepts and word vectors, respectively. Finally, we use a nearest neighbor classifier with L2 distance based on the multimodal representation to determine the most likely business venue.

Experiments

Data Collection and Experimental Settings

For our experiments, we need images and reviews related to business venues. We use the public data, Yelp Challenge Dataset [2], which contains information and reviews of 61,184 business venues in ten different cities from Yelp for this purpose. We then map the venues to the Instagram checkin based on GPS information and venue name. 22,763 venues were found on Instagram. We collect up to 1,000 images for each venue. The distribution of images over venues is shown in Figure 6. Note that more than a half of the venues have fewer then five images. We take 250 images from each of 1,000 different venues as training data to train the BA-CNN and to compute the normalized entropy in each concept. We than take the other venues with more than eleven images as our evaluation set. In total, 7,699 venues are used for evaluation. For each venue, we randomly select one image as query image. The remaining 10 images together with 20 Yelp reviews of the venue construct a geo-tagged database, where the visual concepts (image) and the word vector (reviews) are used to represent the associated business venue. During the recognition, we use GPS information from the query image to narrow down candidate venues to two to ten neighboring venues. We use rank-1 accuracy as our evaluation metric.

Improvements by BA-CNN

We compare **BA-CNN** with several baselines and different settings: (1) **ImageNet-CNN (FC8)** (Deng et al. 2009): we use responses of general concepts (FC8) from CNN trained on ILSVRC 2012 data as a baseline feature. (2) **ImageNet-CNN (FC7)** (Razavian et al. 2014): we use CNN trained on ILSVRC 2012 to extract features (FC7) for business venues recognition. (3) **BA-CNN (C)**: we use CNN trained on Instagram images labeled with 490 business-aware visual concepts to extract features from FC7. For each of the 490 concepts, we further collect 4,000 images from Instagram and use 2,000 images with higher classification scores as training data, in total around one million images are used for training. (4) **BA-CNN (V)**: we use 250,000 images from 1,000 different business venues as training data to train CNN and extract features from FC7. (5) **BA-CNN (C+V)**: we use the configuration in Figure 5 (a) to extract the business-aware concepts for recognition.

As shown in Figure 7, for every method the accuracy drops when the number of neighborhood venues increase because the task becomes more difficult. However, BA-CNN (C+V) can achieve up to 77.5% accuracy when there are two candidates and still maintain around 45% accuracy when the candidate numbers increase to ten; overall, the performance is the best against the other baselines.

ImageNet-CNN performs much worse than BA-CNN and the relevant approaches because the concepts in ImageNet are generic concepts without considering business

[2] http://www.yelp.com/dataset_challenge

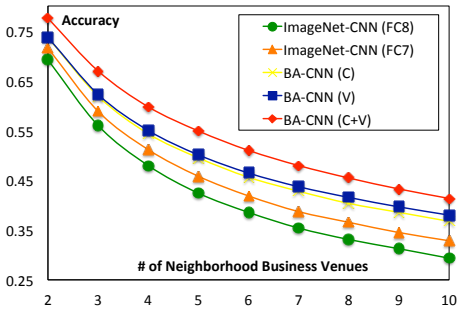

Figure 7: Recognition accuracy as different numbers of neighboring business venues are considered as candidates. When there are more business venues nearby, the performance will drop because the task becomes harder. BA-CNN (C+V) outperforms all other baseline consistently.

awareness and discriminative information between business venues. BA-CNN (C) and BA-CNN (V) have similar performance but BA-CNN (C+V) outperforms both methods because it utilizes both the concept and venue label information in a hybrid structure. Also, BA-CNN (C+V) can take advantage of the semantic representation and be used for multimodal recognition as shown in the following section.

Results of Multimodal Business Venue Recognition

We use the word vector model to convert the visual concepts detected from the query image and the reviews of each business venue in the database as a vector of text representation. Table 3 exhibits the accuracy of business venue recognition by matching the text representations only, that is, no database images are used. **WordVec (Google News)** shows the performance of the model trained with Google News dataset (about 100 billion words) and **WordVec (Business-Aware)** indicates the model trained with Yelp reviews (about 0.2 billion words). **Random Guess** is the accuracy of randomly picking one of the candidate venues. We can see both methods outperform random guessing significantly (more than 115% relative improvement), which suggests that the concepts generate from BA-CNN (C+V) indeed have semantic meaning and highly relevant to what might appear in reviews of local business. WordVec (Business-Aware) performs slightly better than WordVec (Google News) that again shows the importance of business-awareness in the application of business venue recognition.

When combining BA-CNN (C+V) with Word Vectors, we can further improve the recognition accuracy, demonstrating the complimentary nature of the image and text information. It is worth noticing that the multimodal recognition only requires an query image without any text because the proposed image representation, business-aware visual concepts, can be used directly when text representation is available.

The multimodal representation is particularly important for the image sparsity problem in the database of business venues. As shown in Figure 6, many of the business venues contains fewer than five images on Social Media website. Therefore, we also evaluate our method with different number of images (range from one to ten images) for each business venue. In Figure 8, "WordVec" indicates the accuracy of matching query image and database reviews when no database images are available. As the number of database images in the business venues decreases, the recognition accuracy by image representations drops. "ImageNet-CNN (FC7)" only outperforms "WordVec" when there are more than three images of each venues in the database. The accuracy is obviously boosted by further considering database reviews ("BA-CNN (C+V)" vs. "BA-CNN (C+V) + WordVec") when few images are available, suggesting the proposed multimodal recognition method have advantages to tackle the image sparsity issue. In social media, the associations between images and venues are mainly based on user checkins. However, because of the heavy tail and power law behavior of checkins per venue (Noulas et al. 2011), only a few famous venues feature a large number of checkin images, while general business venues have only few checkin images. In consideration of this problem, our approach poses a new opportunity to push the generality of automatic recognition to common business venues.

Table 3: The recognition accuracy with 2 and 5 candidate venues. Simply using text representation obviously outperforms random guess, suggesting the concepts extracted from BA-CNN (C+V) indeed have semantic meaning. WordVec (Business-Aware) surpasses WordVec (Google News) demonstrating the importance of business awareness. BA-CNN (C+V) + WordVec can reach the best accuracy.

Method	Acc.@2	Acc.@5
Random Guess	50.0%	20.0%
WordVec (Google News)	65.8%	39.1%
WordVec (Business-Aware)	69.1%	42.3%
BA-CNN (C+V) + WordVec	**78.5%**	**56.1%**

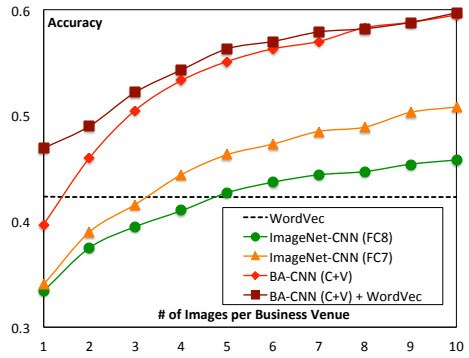

Figure 8: The accuracy with different number of images for each business venue. Image sparsity decreases the accuracy of the models using image representation, while text representation is stable, and multiple modalities (BA-CNN (C+V) + WordVec) can improve more in such cases.

Conclusion

We propose a novel framework for business venue recognition. We first mine business-aware visual concepts from reviews of local business, and then incorporate business-aware concepts with convolutional neural networks for represent-

ing images as response of visual concepts. The semantics of visual concepts can be further represented by text representation. We propose to use multimodal representation for business venue recognition and the experiments show its superiority against the single modal approaches and the state-of-the-art visual features, especially when there are insufficient images to represent the venues. In the future, we will seek the opportunity to associate more data domains, e.g., company profiles, purchase logs. Moreover, we will investigate the other metadata that can replace GPS to narrow down candidate venues, e.g., time, social network.

References

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural language processing with Python*. " O'Reilly Media, Inc.".

Chen, D. M.; Baatz, G.; Köser, K.; Tsai, S. S.; Vedantham, R.; Pylvä, T.; Roimela, K.; Chen, X.; Bach, J.; Pollefeys, M.; et al. 2011. City-scale landmark identification on mobile devices. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 737–744. IEEE.

Chen, Y.; Zhao, J.; Hu, X.; Zhang, X.; Li, Z.; and Chua, T.-S. 2013. From interest to function: Location estimation in social media. In *AAAI*. Citeseer.

Chen, F.; Joshi, D.; Miura, Y.; and Ohkuma, T. 2014a. Social media-based profiling of business locations. In *Proceedings of the 3rd ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia*, 1–6. ACM.

Chen, J.; Cui, Y.; Ye, G.; Liu, D.; and Chang, S.-F. 2014b. Event-driven semantic concept discovery by exploiting weakly tagged internet images. In *Proceedings of International Conference on Multimedia Retrieval*, 1. ACM.

DeLozier, G.; Baldridge, J.; and London, L. 2015. Gazetteer-independent toponym resolution using geographic word profiles. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 248–255. IEEE.

Fang, Q.; Sang, J.; and Xu, C. 2013. Giant: Geo-informative attributes for location recognition and exploration. In *Proceedings of the 21st ACM international conference on Multimedia*, 13–22. ACM.

Friedland, G.; Vinyals, O.; and Darrell, T. 2010. Multimodal location estimation. In *Proceedings of the international conference on Multimedia*, 1245–1252. ACM.

Hays, J., and Efros, A. 2008. Im2gps: estimating geographic information from a single image. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8. IEEE.

Hulden, M.; Silfverberg, M.; and Francom, J. 2015. Kernel density estimation for text-based geolocation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Li, L.-J.; Su, H.; Fei-Fei, L.; and Xing, E. P. 2010. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in neural information processing systems*, 1378–1386.

Li, Y.; Crandall, D. J.; and Huttenlocher, D. P. 2009. Landmark classification in large-scale image collections. In *Computer vision, 2009 IEEE 12th international conference on*, 1957–1964. IEEE.

Lin, T.-Y.; Cui, Y.; Belongie, S.; Hays, J.; and Tech, C. 2015. Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5007–5015.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60(2):91–110.

Maier, D., and Kleiner, A. 2010. Improved gps sensor model for mobile robots in urban terrain. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, 4385–4390. IEEE.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Noulas, A.; Scellato, S.; Mascolo, C.; and Pontil, M. 2011. An empirical study of geographic user activity patterns in foursquare. *ICwSM* 11:70–573.

Razavian, A. S.; Azizpour, H.; Sullivan, J.; and Carlsson, S. 2014. Cnn features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, 512–519. IEEE.

Schindler, G.; Brown, M.; and Szeliski, R. 2007. City-scale location recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 1–7. IEEE.

Ye, G.; Li, Y.; Xu, H.; Liu, D.; and Chang, S.-F. 2015. Eventnet: A large scale structured concept library for complex event detection in video. *arXiv preprint arXiv:1506.02328*.

Zheng, Y.-T.; Zhao, M.; Song, Y.; Adam, H.; Buddemeier, U.; Bissacco, A.; Brucher, F.; Chua, T.-S.; and Neven, H. 2009. Tour the world: building a web-scale landmark recognition engine. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 1085–1092. IEEE.