IMAGE-BASED USER PROFILING OF FREQUENT AND REGULAR VENUE CATEGORIES

Ryosuke Shigenaka[*], *Yan-Ying Chen*[†], *Francine Chen*[†], *Dhiraj Joshi*[‡], *Yukihiro Tsuboshita*[*]

[*]Fuji Xerox Co., Ltd., Yokohama-shi, Kanagawa, Japan
[†]FX Palo Alto Laboratory, Inc., Palo Alto, California, USA
[‡]IBM T.J. Watson Research Center, Yorktown Heights, New York, USA
{shigenaka.ryosuke|Yukihiro.Tsuboshita}@fujixerox.co.jp, {yanying|chen}@fxpal.com, djoshi@us.ibm.com

ABSTRACT

The availability of mobile access has shifted social media use. With that phenomenon, what users shared on social media and where they visited is naturally an excellent resource to learn their visiting behavior. Knowing visit behaviors would help market survey and customer relationship management, e.g., sending customers coupons of the businesses that they visit frequently. Most prior studies leverage meta-data e.g., check-in locations to profile visiting behavior but neglect important information from user-contributed content, e.g., images. This work addresses a novel use of image content for predicting the user visit behavior, i.e., the frequent and regular business venue categories that the content owner would visit. To collect training data, we propose a strategy to use geo-metadata associated with images for deriving the labels of an image owner's visit behavior. Moreover, we model a user's sequential images by using an end-to-end learning framework to reduce the optimization loss. That helps improve the prediction accuracy against the baseline as demonstrated in our experiments. The prediction is completely based on image content that is more available in social media than geo-metadata, and thus allows coverage in profiling a wider set of users.

Index Terms— user profiling, visit pattern, social media

1. INTRODUCTION

User visit behavior is essential information for marketing and customer management. Traditionally, the visit behavior could be available from customer surveys, e.g., filling out a feedback sheet after a visit. The information would be helpful for business managers to understand their customers' visit behavior and offer better services. Complementary to conventional feedback sheets, social media has become a common way to seek and reach more customers. Its scale, diversity and convenience pose great opportunities to understand user types and behavior [1] and to offer the business managers actionable insights. On the other hand, it also relies more on machine intelligence for automating analytics on the unstructured data to infer web-scale users' behavior.

Among a variety of behaviors, frequency and regularity are two of the most common patterns of interest. For example, sending relevant coupons of the businesses that the customers would visit frequently, or offering discount periodically for the users who would visit regularly. Moreover, discovering these visit patterns from user logs [2] is getting more promising with the rise of location data in social media, e.g., Foursquare, Instagram. However, relying on geo-metadata still offers limited coverage for users who either turn off the geo-sensors or prefer to use the social media platforms without location service.

Fig. 1. We propose an end-to-end system to infer the frequent and regular business venue categories such as "Japanese Restaurant" and "Stadium" that a target user would visit by using their history of image posts.

User-contributed content such as images is an integral part of users' social media posts, and the increasing convenience of mobile access has made the content more closely related to where users visited. In comparison to location metadata, image content was reported to occur more frequently in social media. A report [3] indicated that in Twitter, 36% of posts contained image content while less than 1% of posts tagged were with location [4]. Furthermore, the photo-centric social media platform Instagram only had 5% of image posts tagged with location [5]. In addition to higher availability, photos taken at a site may be an indicator of where users visited. For example, in Figure 1 the photos of ramen reveal the clues about visits at Japanese restaurants. The effectiveness of using image content for location prediction has been demonstrated in the prior works [6, 7], though most of them focus on prediction by image rather than by user.

Motivated by the high demand of visit behavior and the greater availability of image posts, we propose to predict the frequent and regular business venue categories (e.g., Japanese Restaurant, Stadium in Figure 1) that a user is likely to visit,

978-1-5090-6067-2/17/$31.00

given his/her history of images posts in social media. The consideration of both frequency and regularity is to understand a user's behavior of routinely visiting certain venue categories while avoiding the frequent but irregular behavior, e.g., taking a burst of photos at a football game but never visiting football games thereafter. The proposed model is based on a learning method that requires a user's sequential images and ground-truth labels of visit behavior. To reduce the labeling efforts for collecting the training data, we use location metadata to automatically derive the labels that approximate the ground-truth visit behavior. Furthermore, to reduce the optimization loss in offline training, we model a user's images by using an end-to-end learning framework. Finally, to profile a wider coverage of users, the online prediction only relies on given image content without any use of location metadata.

To summarize, the contributions of our work include: 1) proposing a novel system that uses image content to predict the visit behavior - frequent and regular business venue categories that the content owner is likely to visit; 2) proposing a strategy of deriving weak labels of visit behavior to reduce manual labeling efforts; 3) proposing an end-to-end framework that learns the discriminative features in users' sequential images for the visit behavior prediction. We discuss the related work in the next section. The dataset and annotation method are then introduced, following by the learning framework for predicting frequent and regular venue categories. Finally, we evaluate our system versus the baseline methods and conclude our findings.

2. RELATED WORK

The related studies are separated into the three research areas, 1) visit behavior mining, 2) content-based venue prediction and 3) learning from sequential data that either motivates or is involved in our proposed system.

The success of location-based social networks (LBSN) has brought abundant user-contributed location logs to the Internet and has motivated several prior studies to address visit behavior mining and location recommendation. Bao et al. [2] propose a location-based and preference-aware system that recommends a list of venues to a user according to his/her location history. Since users' location logs might be very sparse, considering user information such as social interaction of the user being profiled and others [8] and local preference [9] is thought to be complementary to location logs.

Visit behavior not only helps personalized recommendation but also marketing. Karamshuk et al. [10] study the problem of retail store placement by using location logs in LBSN. These prior studies have gradually confirmed the feasibility of understanding visit behavior using LBSN and its practical use in real applications. However, most of them only use metadata without addressing the analysis of content, particularly images, while the benefits from image content is the main focus of this work.

Using content analysis to infer where content was posted has been shown to be promising in some prior works. Chen et al. [7] propose to use venue-related concepts detected in image content for predicting the venue of an image. Weyand et al. [6] considers the sequential relationship between images in an album to make more accurate prediction of image location.

Venue categories such as movie theater, stadium and bookstore are also getting more attention because they reveal the information about activities at venues and might be indicative of user preference. Movshovitz-Attias et al. [11] propose to use street-view storefront images to recognize the venue category of given image content. These studies have shaped a more concrete picture of how image content is informative of a venue or a venue category where it was taken; however, all of these studies perform analysis of image-level location recognition and do not advance to a user-level visit behavior prediction that is targeted in this paper.

Pushing image-level analysis forward to user-level analysis requires effective aggregation. An intuitive approach is a two-stage approach like [12] that does image recognition and then aggregates the labels of images to infer the label(s) of the owner. However, image recognition is not perfect, especially for recognition based on a single image, because it may only comprise junk information without clear relevance to location. This phenomenon is more obvious for social media data since the content is not controlled and may be noisy.

Modeling a sequence of data items has been proposed to extract the relationship from item to item that is neglected in the prediction of single items. Kim [13] proposes an end-to-end system that learns features from sequential words in sentences for sentence-level sentiment classification and demonstrates that the end-to-end system is better than a two-stage system where the classification of a sentence is built upon the sub-classification results of phrases [14]. Learning feature representations of sequential data from end-to-end has become more compelling for improving classification and prediction, not only for text but also video [15]. As for a set of user images taken in order of time, it resembles the nature of sequential data but has not yet been addressed in this way. Motivated by the success of end-to-end systems, our work investigates end-to-end representation learning that can better aggregate a set of user's images for predicting the venue categories that the user would visit frequently and regularly.

3. DATA COLLECTION AND ANNOTATION

Our experimental data are composed of the user-contributed images associated with check-in venues. To ensure that the data can cover diverse venues and venue categories, we take the SMPBL dataset [16] as an initial repository. SMPBL is a Twitter post collection oriented to business venues and collected from the San Francisco Bay Area from June 2013 to April 2014. In this dataset, we keep around 0.39 million posts

Table 1. Examples of venue categories having lower EER in cross validation.

Example Venue Categories
Beach, Movie Theater, Stadium, Ski Area, Winery, Salon, Airport, Jewelry Store, Ice Cream Shop, Bar, Car Wash, Casino, Campground, Food Truck, Bookstore, Farm, Theme Park, Pet Service, Zoo

that are forwarded from Instagram (a photo-centric social network) and comprised of image content as well as information on venue and time stamps. The venue category of each venue is traced in a LSBN website, i.e., Foursquare [1], where the venue category information was collaboratively labeled by business owners, social media users and developers.

There are around 110,980 users who contributed these 0.39 million posts. To secure sufficient observations without loss of generality, we filter out the users with less than 5 images. Since users may change their visit behavior from time to time, we segment a user's image sequence into subsequences every 35 images. After preprocessing, our data set contains 9,534 image sequences comprising 178,739 images.

3.1. Deriving Labels

To obtain the ground-truth labels of our training data, we propose a strategy that uses the venue category and time stamp logged with each image in a user's sequence to derive weak labels of frequent and regular venue categories for that user. The automatic derivation from user logs helps reduce the subjectivity and efforts of doing manual annotation.

A user's image sequence is labeled with a list of venue categories, each is a probability P_c that indicates how likely the user is to visit the category c frequently and regularly.

$$P_c = \frac{S_c}{\sum_{i \in C} S_i}, \tag{1}$$

where i is the index of venue category and C is the whole set of venue categories. The score S_c is derived from the frequency and regularity of the user's posts as follows,

$$S_c = \frac{f_c \times (d_c + \alpha)}{\sqrt{Var(\triangle T_c)} + \alpha}. \tag{2}$$

f_c represents the frequency of image posts associated with the venue category c in the user's sequence. d_c is the time duration between the first and the last image posts associated with the venue category c in the user's sequence. $\triangle T_c$ indicates the time intervals between any two consecutive image posts associated the venue category c in the user's sequence. α is a constant to prevent extreme values of S_c if the variance was 0 or if d_c was 0. α is set to 0.3 in our experiments. A larger f_c indicates a higher frequency while a smaller $Var(\triangle T_c)$ indicates more consistent time intervals between consecutive tweets and hence higher regularity. The category that has image posts spreading over longer time (higher d_c) is assumed to be more regular and has higher S_c.

[1]The list of venue categories used in our experiments was downloaded from `https://developer.foursquare.com/categorytree`.

3.2. Evaluating Venue Category

Learning representative image features of a venue category requires a sufficient amount of training images and considerable visual consistency in image content. However, not every venue category satisfies the two characteristics and thus may have negative impacts on visit behavior prediction. To reduce the impacts, we integrate these categories as a special category "other" in the learning framework to prevent these categories from overwhelming the optimization loss but still keep certain effects like background data. We first use 2-fold cross validation (a similar procedure used in [7]) to evaluate the Equal Error Rate (EER) of the binary classification results of each venue category. Using EER as an indicator of quality, the categories are integrated to a special category "other" if their EER is higher than 30%. That roughly covers 1/3 of the total categories. After the preprocessing, there remain 164 venue categories plus one "other" category in our experiments. Table 1 presents examples of the 164 venue categories with higher classification accuracy.

4. LEARNING FROM USER IMAGES

This section describes a baseline method first and then introduces the end-to-end modeling used in our work. The methods of end-to-end modeling include non-sequential and sequential based approaches for learning a representation indicative of visit behavior from a sequence of user images. The two types of approaches are compared in the experiments to provide more insights. The inputs for training are image sequences ordered by time. The output of each sequence is a 165-dimensional vector of probability $\mathbf{P}$ formed by P_c of each venue category c that indicates how likely the user would visit that category frequently and regularly (cf. Section 3.1). Before being fed in a training model, each image in a sequence is represented by an image embedding using Convolutional Neural Networks (CNN).

4.1. Baseline Method

To the best of our knowledge, there is no existing work using images to predict frequent and regular venue categories for a user. An intuitive way to define a baseline system is 1) predicting the probabilities of venue categories for each image in a user's sequence and then 2) taking the average of the predicted probabilities as the aggregated prediction of that sequence. For the stage 1), we train a model by using AlexNet [17] for single-image prediction. A likely 2-stage model is

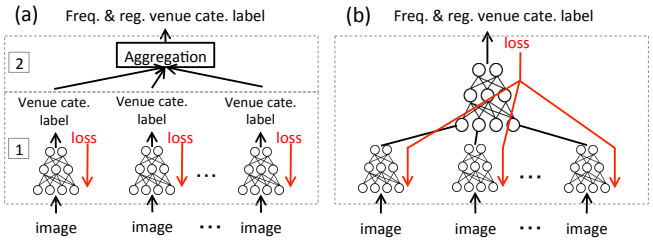

Fig. 2. The major difference of (a) the baseline 2-stage system and (b) the proposed end-to-end system.

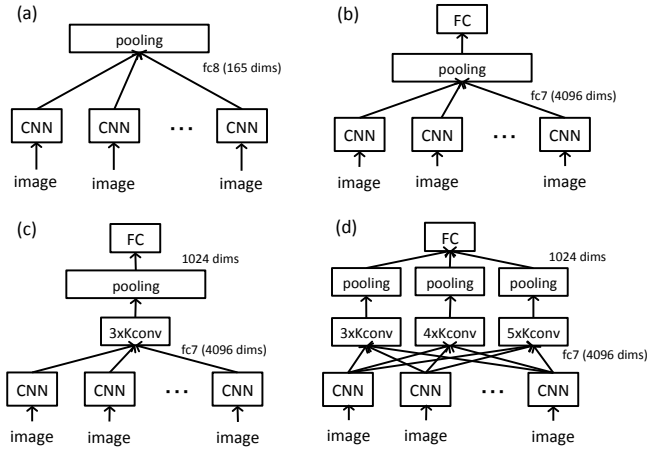

Fig. 3. The four architectures used in our end-to-end system for modeling image sequences: (a) pooling, (b) pooling+fc, (c) conv+pooling+fc and (d) multi-conv+pooling+fc.

used in another user-profiling work [12] but targets on predicting a user's gender. Figure 2 illustrates the major difference between this baseline model (referred as **2-stage**) and our model that optimizes training loss from end-to-end.

4.2. Non-sequential Modeling over Image Embeddings

A typical non-sequential modeling performs pooling over embeddings of each image in a sequence. Figure 3 (a) shows an example architecture (referred to **pooling**), where each image is represented by the output layer (fc8) of AlexNet. (b) shows another variation of pooling (referred to **pooling+fc**) that uses the hidden layer (fc7) of AlexNet as the input for pooling. On the top of pooling, there is a fully-connected layer (FC) to turn the pooling results to a 165-dimensional output. The two pooling-based models incorporate multiple images without considering the sequential order sorted by time.

4.3. Sequential Modeling over Image Embeddings

We also investigate two sequential-based architectures to understand whether time-sequence information helps for analyzing visit behavior. Figure 3 (c) shows an architecture (referred as **conv+pooling+fc**) that use $3 \times K$-dimensional kernel filters to do convolution on every sub-sequence in an image sequence where 3 is the window size of a sub-sequence and K is the dimension of an image embedding (i.e., 4,096, the hidden layer in AlexNet). This framework is motivated by the 3-gram convolution used in [13]. The convolutions over a sequence is aggregated by pooling. We use 1,024 kernel filters and thus get a 1,024-dimensional output after pooling. An FC is used to turn the pooling results to a 165-dimensional output.

To investigate how the window size of a sub-sequence affects the performance, we experiment on another sequential-based architecture (referred as **multi-conv+pooling+fc**) as presented in Figure 3 (d) that extends the window size of a sub-sequence from 3 to a set of windows of size 3, 4 and 5. The 3 window sizes form 3 types of kernel filters $3 \times K$, $4 \times K$ and $5 \times K$, and the convolution results of each type are applied with pooling and then concatenated as a 3×1024-dimensional vector. An FC layer is used to turn the vector to a 165-dimensional output.

Table 2. Experiment data

Data set	Training	Validation	Test
# of images	137,378	19,732	21,629
# of users	7,534	1,000	1,000

5. EXPERIMENTS

5.1. Dataset and Evaluation Metrics

Table 2 presents the number of images and users in the training, validation and test sets. We use two evaluation metrics, top-1 accuracy and normalized discounted cumulative gain ($nDCG_r$) in the experiments. The random guess of top-1 accuracy is about 0.6% that represents selecting one of the 165 venue categories. $nDCG_r$ is used to evaluate the ranking quality of venue categories based on their predicted probability. r indicates a particular rank position of venue categories sorted by their probabilities P_c, and is set to 1 and 3 in the experiments. For $r = 1$, only the most probable venue category, as defined in Eq. 1, is considered relevant. For $r = 3$, the three most probable venue categories are considered relevant.

5.2. Implementation Details

We use AlexNet pretrained on ImageNet [18] as a base model and fine-tune on the SMPBL training data where the weights of the first two convolution layers are fixed. In the end-to-end models, the additional layers on top of the base model are initialized with random values based on Gaussian distribution. All of the models are trained though stochastic gradient descent with momentum 0.9 and weight decay 0.0005. The mini-batch size is set to 256 before the last pooling layer and 1 after it. The starting learning rate is set to 0.001 and decreased by a factor of 10 for each 10 epochs until conver-

Table 3. The top-1 accuracy (%) and $nDCG$ (%) of the models that use different loss functions.

Models	accuracy	$nDCG_1$	$nDCG_3$
cross entropy	**33.9**	45.42	45.40
pairwise ranking	32.4	44.01	**49.58**
warp ranking	33.6	45.41	48.24

Table 4. The top-1 accuracy (%) and $nDCG$ (%) for the models using maximum pooling and average pooling.

Models	accuracy	$nDCG_1$	$nDCG_3$
pooling (max)	33.6	45.41	48.24
pooling (ave)	36.0	**48.45**	**55.29**
pooling+fc (max)	35.6	48.15	52.61
pooling+fc (ave)	**36.1**	48.31	55.07

gence. All experiments are conducted using a single NVIDIA Tesla K80 GPU and implemented with Chainer [19].

5.3. Performance of Visit Behavior Prediction

This section compares the baseline 2-stage model and the proposed end-to-end models, pooling, pooling+fc, conv+pooling+fc and multi-conv+pooling+fc (cf. Figure 3 and Section 4) in the task of predicting frequent and regular categories of an image sequence. Note that, the task is more challenging than predicting the category of a single image because the imperfect prediction of each image can impact on the prediction of a whole sequence. We also experiment with different loss functions and pooling methods to determine a better combination for the proposed end-to-end framework.

Cross-entropy loss vs. ranking loss: We compare three loss functions, cross-entropy loss [20], pairwise ranking loss [21] and warp ranking loss [22] to optimize training errors in an end-to-end framework. The experiments are conducted using the same model, pooling (Figure 3 (a)), for comparing the three loss functions. As presented in Table 3, ranking loss functions perform better in $nDCG$ while cross-entropy loss function has higher top-1 accuracy. That suggests the two types of loss functions could be used for different applications, e.g., ranking loss would be a better option for providing a list of prioritized recommendations and cross-entropy loss can help target users of a specific category. Considering a balance between top-1 accuracy and $nDCG$, warp ranking is selected as the loss function for the following experiments.

Maximum pooling vs. average pooling: We also compare two types of pooling, maximum pooling and average pooling, to identify which one works better for predicting frequent and regular venue categories. The two types are applied to the two non-sequential models, namely pooling (Figure 3 (a)) and pooling+fc (Figure 3 (b)). Table 4 reveals that average pooling is generally better than maximum pooling in both top-1 accuracy and $nDCG$, perhaps because the average pooling can better represent frequent patterns.

Table 5. The top-1 accuracy (%) and $nDCG$ (%) of the baseline 2-stage model and the end-to-end models used in our system. Random indicates a random guess from 165 categories.

Models	accuracy	$nDCG_1$	$nDCG_3$
Random	0.6	–	–
2-stage	34.6	46.16	54.41
pooling	36.0	48.45	55.29
pooling+fc	36.1	48.31	55.07
conv+pooling+fc	**37.3**	**49.42**	**55.50**
multi-conv+pooling+fc	35.7	47.41	51.70

Non-sequential vs. sequential: We compared non-sequential modeling (pooling, pooling+fc) and sequential modeling (conv+pooling+fc, multi-conv+pooling+fc). Table 5 shows that sequential modeling (conv+pooling+fc) performs better than non-sequential modeling. However, the window size of sub-sequence matters, and considering additional sub-sequences with a longer window size (multi-conv+pooling+fc) would not improve the prediction. This result suggests that user visit patterns might be more distinct in the shorter term, and therefore shorter sub-sequences could contribute more to the representation learning of whole sequences.

2-stage vs. end-to-end: We compared the baseline 2-stage model and several variations of the end-to-end optimization models. Table 5 reports that most of the end-to-end models perform better than the 2-stage model, no matter if they are modeled non-sequentially or sequentially. That might be attributed to the error propagation in the 2-stage model. The prediction error per single image (the first stage) might be propagated to the prediction per user (the second stage) because not every single image is informative enough to make an accurate prediction. The results also suggests that end-to-end systems may alleviate error propagation because it can optimize loss from end to end.

5.4. Challenges

This work addresses several aspects of modeling for predicting frequent and regular categories from an image sequence, but there remain some challenges in the defined task. First, noisy image data have considerable impacts on the prediction accuracy. Figure 4 exhibits some example results of the binary classification using 2-fold cross validation. (b) comprises the examples associated with "beach" but misclassified as negatives. The content of these false negatives varies from flowers to portraits without clear clues indicative of "beach" because image content shared by users might not be always location-aware, e.g., selfies. The same phenomenon appears to the category "farm" as displayed in Figure 4 (d), and similarly many of the false negatives comprise

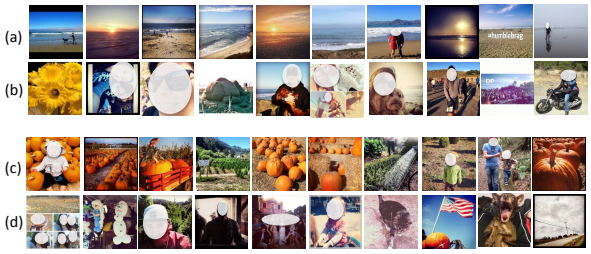

Fig. 4. Example classification results in cross validation. (a) and (b) show the true positives and false negatives, respectively, for the category "beach." (c) and (d) exhibit the true positives and false negatives, respectively, for the category "farm." Faces are blurred to avoid privacy concerns.

Table 6. The top-1 accuracy (%) and $nDCG$ (%) of the end-to-end models added with one more FC layer on the top.

Models	accuracy	$nDCG_1$	$nDCG_3$
pooling+2fc	**36.9**	**48.92**	**55.49**
conv+pooling+2fc	36.3	48.20	54.58
multi-conv+pooling+2fc	35.5	47.17	53.36

human that occludes most of the background. This also explains why the single-image prediction in the baseline 2-stage model may be imperfect. Besides of optimizing the loss from end-to-end, we expect that involving other data modalities such as text would complement the missing information.

Though we demonstrate that the idea of optimizing errors from end to end does improve the predictions, adding more units in the architectures (e.g., more FC layers) does not always bring up improvement. The results in Table 6 point out this problem – by adding one more FC layer, pooling+2fc improves the prediction results of pooling+fc (Table 5). Conversely, conv+pooling+2fc and multi-conv+pooling+2fc do not gain any improvement and are even worse than the non-sequential model pooling+2fc. This might be attributed to overfitting the increasing number of parameters, but it is supposed to be mitigated if the training data can be scaled up.

6. CONCLUSION

We propose a novel system that leverage user-contributed image content to predict the venue categories that the content owner would visit frequently and regularly. To mitigate the manual efforts in collecting ground-truth labels of visit behavior, we propose a strategy of deriving weak labels from user logs in social media. To reduce the training loss, we focus on an end-to-end learning framework for the prediction and investigate several methods including non-sequential and sequential modeling. The experiments suggests that the end-to-end modeling used in our system outperforms the conventional 2-stage model. Moreover, the investigation over data and end-to-end models poses the opportunities and challenges of image-based visit behavior prediction. In the future work, we will address the noisy data and overfitting problems and consider including text to complement image content.

7. REFERENCES

[1] Yuheng Hu, Lydia Manikonda, and Subbarao Kambhampati, "What we instagram: A first analysis of instagram photo content and user types," *ICWSM*, 2014.

[2] Jie Bao, Yu Zheng, and Mohamed F. Mokbel, "Location-based and preference-aware recommendation using sparse geo-social networking data," in *ACM SIGSPATIAL*, 2012.

[3] Sam Laird, "Twitter: A day in the life [infographic]," *Mashable*, 2012, http://mashable.com/2012/08/16/twitter-day-in-the-life-infographic/.

[4] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews, "Home location identification of twitter users," *ACM TIST*, 2014.

[5] Christopher Heine, "14 instagram data findings that every marketer needs to know," *Adweek*, 2014, http://www.adweek.com/digital/14-instagram-data-findings-every-marketer-needs-know-160969/.

[6] Tobias Weyand, Ilya Kostrikov, and James Philbin, "Planet - photo geolocation with convolutional neural networks," in *ECCV*, 2016.

[7] Bor-Chun Chen, Yan-Ying Chen, Francine Chen, and Dhiraj Joshi, "Business-aware visual concept discovery from social media for multimodal business venue recognition," in *AAAI*, 2016.

[8] Hao Wang, Manolis Terrovitis, and Nikos Mamoulis, "Location recommendation in location-based social networks using user check-in data," in *ACM SIGSPATIAL*, 2013.

[9] Hongzhi Yin, Yizhou Sun, Bin Cui, Zhiting Hu, and Ling Chen, "Lcars: A location-content-aware recommender system," in *ACM KDD*, 2013.

[10] Dmytro Karamshuk, Anastasios Noulas, Salvatore Scellato, Vincenzo Nicosia, and Cecilia Mascolo, "Geo-spotting: Mining online location-based services for optimal retail store placement," in *ACM KDD*, 2013.

[11] Yair Movshovitz-Attias, Qian Yu, Martin C. Stumpe, Vinay Shet, Sacha Arnoud, and Liron Yatziv, "Ontological supervision for fine grained classification of street view storefronts," in *CVPR*, 2015.

[12] Xiaojun Ma, Yukihiro Tsuboshita, and Noriji Kato, "Gender estimation for sns user profiling using automatic image annotation," in *ICMEW*, 2014.

[13] Yoon Kim, "Convolutional neural networks for sentence classification," in *EMNLP*, 2014.

[14] Tetsuji Nakagawa, Kentaro Inui, and Sadao Kurohashi, "Dependency tree-based sentiment classification using crfs with hidden variables," in *ACL*, 2010.

[15] Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko, "Sequence to sequence video to text," in *CVPR*, 2015.

[16] Francine Chen, Dhiraj Joshi, Yasuhide Miura, and Tomoko Ohkuma, "Social media-based profiling of business locations," in *ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia*, 2014.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[19] Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton, "Chainer: a next-generation open source framework for deep learning," in *Workshop on Machine Learning Systems at NIPS*, 2015.

[20] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid, "Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *ICCV*, 2009.

[21] Thorsten Joachims, "Optimizing search engines using clickthrough data," in *ACM KDD*, 2002.

[22] Jason Weston, Samy Bengio, and Nicolas Usunier, "Wsabie: Scaling up to large vocabulary image annotation," in *IJCAI*, 2011.