USING BUSINESS-AWARE LATENT TOPICS FOR IMAGE CAPTIONING IN SOCIAL MEDIA

Yan-Ying Chen, Francine Chen, Matthew Cooper, Dhiraj Joshi

FX Palo Alto Laboratory, Inc., Palo Alto, California, USA
{yanying,chen,cooper,dhiraj}@fxpal.com

ABSTRACT

Captions are a central component in image posts that communicate the background story behind photos. Captions can enhance the engagement with audiences and are therefore critical to campaigns or advertisement. Previous studies in image captioning either rely solely on image content or summarize multiple web documents related to image's location; both neglect users' activities. We propose business-aware latent topics as a new contextual cue for image captioning that represent user activities at business venues. The idea is to learn the typical activities of people who posted images from business venues with similar categories (*e.g.*, fast food restaurants) to provide appropriate context for similar topics (*e.g.*, burger) in new posts. User activities at businesses are modeled via a latent topic representation. In turn, the image captioning model can generate sentences that better reflect user activities at business venues. In our experiments, the business-aware latent topics are effective for adapting to captions to images captured in various businesses than the existing baselines. Moreover, they complement other contextual cues (image, time) in a multi-modal framework.

Index Terms— image caption, topic model, social media

1. INTRODUCTION

Photo-centric social media sites such as Instagram and Pinterest have shown exceptional growth and become major platforms for interaction. The trend is reflected in the amount of image posts; 70 million photos and videos were shared on Instagram per day in 2015. These posts are generally a combination of image, caption, check-in, and other metadata, each revealing certain clues about users' experiences.

Captions are a vital part of image posts in social media because they convey a richer semantic representation which can tell a story about a photo and express users' experiences including why/when/where a photo was captured. For photo sharing, the importance of captions has been supported in several empirical investigations [1, 2], presumably because a motivation for sharing photos is sharing stories. Though captions can enhance user engagement and thus help publicity campaigns and advertisement, they require much manual efforts and that motivates research of automatic captioning.

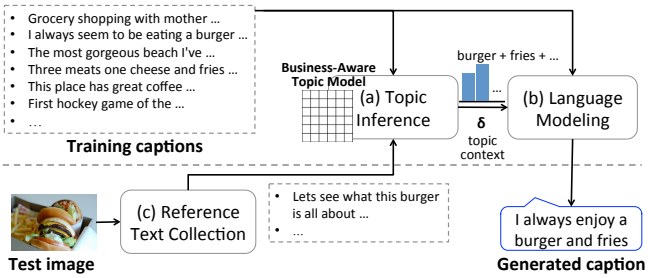

Fig. 1. Generating image captions that describe relevant user activities at businesses – (a) representing each of the given training captions as a distribution over business-aware latent topics; (b) using the topic representation as a context prior for learning language models; (c) given a test image, using the associated metadata to collect its reference texts that are then represented as the topic context to bias the generated caption.

As a summarization of a whole story, captions are influenced by multiple associated contexts. Image content is thought to provide informative context for generating visually descriptive captions [3, 4]. The metadata automatically captured with image posts, *e.g.*, locations [5, 6], are also useful clues to generate textual summaries. Captions also depend on the circumstances in which the image was captured or shared, *i.e.*, what activities the authors were doing. These complicated factors lead to noisy captions and pose great challenges to analysis and prediction.

Integrating awareness of user activities is a potential direction for improving image captioning but has not been addressed much in the literature. We propose to represent user activities at businesses by modeling the latent topics in user-contributed captions that correlate with businesses. Note that, the activities are at the local business level rather than user level. Here, we use **business venue categories** associated with location check-ins of images as a proxy for information on users' activities at local businesses to learn the latent topic model. Check-in is a concept of self-reported positioning allowing users to share physical locations on social media. Business venue categories are associated with a check-in and the category labels are contributed and verified by business owners and users in location-based social media; for example, in Foursquare [7] the category of "McDonald" is "Fast Food Restaurant." Crowdsourced knowledge ensures reason-

able credibility and makes business venue categories a practical information source to better discern users' activities.

The result is a topic model that associates business categories with words used in image posts describing relevant user activities at similar businesses, and is referred to as **business-aware latent topics** hereafter. As shown in Fig. 1, this model allows us to represent reference texts of an image in a latent space that spans a range of user activities and to bias caption generation towards descriptive words more appropriate to its business category. Our experiments demonstrate that business-aware latent topics are helpful for generating relevant captions posted at business venues and complement the other contextual data in a multi-modal captioning model.

The contributions of this work include: (1) proposing business-aware latent topics to represent user activities at different businesses; (2) using business-aware latent topics as a new contextual cue for image captioning; (3) comparing contextual cues for generating captions of images posted at specific business venue categories; (4) combining business-aware latent topics with other context (image content, time, etc.) in a multi-modal image captioning framework.

2. RELATED WORK

Research on image captioning has drawn lots of attention. Generalizing, the studies can be separated into visual-based image captioning and context-based text summarization.

Visual-based image captioning attempts to generate visually descriptive sentences for images by exploiting features extracted from visual content and the statistics of associated language data. Since this research area targets visually descriptive text, image content is most often the only available information for generating captions. Kulkarni, *et al.* [3] used the visual detection results (objects, attributes, spatial relationships, etc.) in image content with respect to a statistical prior obtained from descriptive text as constraints to generate sentences for new images. With the recent advances of deep learning in visual computing and machine translation, some studies incorporate deep learning models to generate more precise sentences for images. Kiros, *et al.* [8] proposed Multi-Modal Log-Bilinear Models for sentence generation with a prior biased by the image representation learned by a convolutional neural network. The idea can be extended to more language models such as recurrent neural networks to improve sentence generation describing images [9, 10] and image regions [4]. However, captions of image posts in social media are often motivated by storytelling rather than solely describing the visual content. These captions are thus naturally influenced by multiple contexts beyond the image content itself.

Context-based text summarization can summarize user experiences along with the image by using associated contextual sources. Fan, *et al.* [11] leveraged place names (*e.g.*, London Bridge) in Geographic Information Systems (GISs) to search for relevant documents on the web and summarize their retrieved text to generate captions for images. Li, *et al.* [6] proposed to summarize blog posts relevant to given images and locations for generating blogs. Aker, *et al.* [5] further used patterns learned from corpora of scene types (bridge, churches, etc.) to bias summarization. However, not every location has sufficient relevant captions for summarization unless they are famous travel landmarks. In photo-centric social media, reference texts for diverse user-contributed photos are not often available. Different from summarizing a set of documents associated with a specific location, we propose to train a language model of captions that is biased by the latent topics shared over venues and correlated to user activities at businesses. Latent topics have been adopted for profiling locations [12] and users [13], predicting image annotations [14], as well as incorporating memory in language models [15] because they are informative to represent text. Different from the previous studies, the proposed method targets on image captioning and can take advantage of prior knowledge from captions posted at similar business venues with minimal reference texts in the test phase.

Table 1. The experimental data of image captioning.

business venue categories	#Posts	#Venues	#Users
Food	3,998	265	2,639
Outdoor & Recreation	4,359	174	3,096
College & University	530	28	399
Art & Entertainment	4,063	118	2,858
Nightlife Spot	1,266	117	839
Professional & Others	1,870	85	1,367
Shop & Service	2,010	89	1,097
Travel & Transportation	1,456	62	1,269
Overall[1]	19,717	972	9,998

3. DATA COLLECTION

Though many websites host image posts, we focus on the posts from Instagram associated with check-in venues because they are: (1) user-contributed, (2) photo-centric, and (3) relevant to businesses. User-contributed data include the full range people's diverse expressions, visual content and activities. Examining a photo-centric platform ensures that the photos are created before the captions and thus considerably influence the captions. Check-in venues that were intentionally attached by users can associate the posts with corresponding business venue categories and provide additional context regarding users' activities at businesses.

To ensure the data cover diverse business venues, we use **SMPBL** – a data set oriented to business venues [16] as the initial seeds. SMPBL includes geo-tagged Twitter posts in the San Francisco Bay Area from June 2013 to April 2014. Around 390,200 posts in SMPBL are forwarded from Instagram, each associated with a business venue (either check-ins or location names in posts that are close to the actual business venues) and the corresponding business venue categories [7], which have been crowdsourced from users and developers of

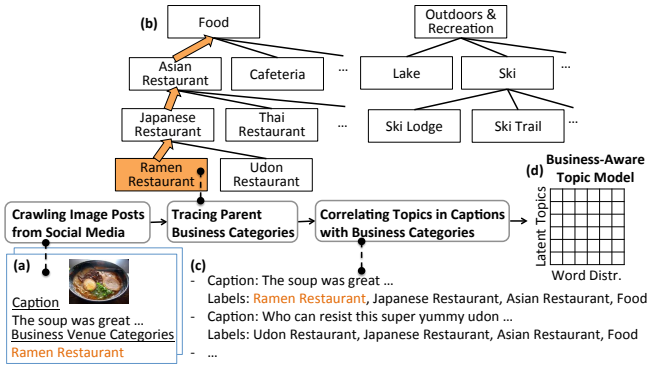

Fig. 2. Business-aware topic modeling – (a) A set of image posts are crawled via location-based social media (Sec. 3), each with the corresponding business categories that were crowdsourced by social media users. (b) Each category label is expanded to a set of categories to include its parents at each level. (c) By using Corr-LDA to correlate captions and the expanded categories, the latent topics could take the advantage of the correlation with similar business activities (Sec. 4).

the social media site Foursquare. An example post is shown in Fig. 2(a). A majority of these captions are organized by hashtags or short phrases rather than sentences, so we filter out the posts with more than 3 hashtags (the captions that are dominated by hashtags without more complete sentences) or fewer than 8 words (similar to [8]). We also remove posts with any mentions to avoid conversations between users. Based on these processes and the availability of original images, we collected 19,717 image posts of 972 business venues from 8 major business categories for experiments (Table 1 [1]).

4. BUSINESS-AWARE TOPIC MODELING

Latent variable methods [17, 18] can effectively extract topics characterizing a corpus but they are rarely used for image captioning due to the lack of relevant texts. Assuming posts from the same place are likely to reflect similar topics, an idea is to use check-in venues to align image posts and aggregate relevant captions. However, the number of business venues is huge (60 millions in Foursquare in 2014) and can keep growing along with new venues, the captions relevant to most venues are sparse [2] and not sufficient for topic modeling.

To overcome the data sparsity problem, we align image posts created at venues with similar business categories for topic modeling. This assumes people at similar business venue types are likely to discuss similar topics in captions; for example, commenting on hamburgers at fast food restaurants. Specific venues without sufficient data to highlight significant topics inherit knowledge from the data collected at the other venues with similar categories, *e.g.*, Burger King and McDonalds. We leverage Correspondence LDA (Corr-LDA) [18] to correlate captions and business venue categories.

Corr-LDA models the correspondence between two types of data, *e.g.*, images/tags [18] and images/sounds [19] for the purpose of automatic annotation given the first data type. As for the two data types captions/categories in this work, rather than predicting categories given captions, we use the joint distribution of captions and business venue categories to find latent topics in captions that are oriented by businesses. As shown in Fig. 2 (b), each block is a business venue category, and they are hierarchically organized from coarse to fine grain according to the Foursquare Category Hierarchy [7]. Given a caption associated with a check-in venue, the venue is linked to a business category, and we extend the single category to a set of categories by including its parents at each level. For example, the category "Ramen Restaurant" is expanded to a set with terms from the categories {Ramen Restaurant, Japanese Restaurant, Asian Restaurant, Food} and the set is then assigned to the caption (Fig. 2 (c)). The label expansion allows the model to take advantage of the correlation between other similar businesses under the same parent categories, *e.g.*, the information about Ramen Restaurant can be shared with Udon Restaurant.

Let $\boldsymbol{z}$ be the latent topics that generate the captions and $\boldsymbol{y}$ be discrete indexing variables. α and β are the parameters of the Dirichlet prior on the topic distributions and the prior on the word distribution per topic, respectively. For each caption $\boldsymbol{w}$ and its corresponding set of business venue categories $\boldsymbol{b}$,

- Sample $\theta \sim Dir(\theta \mid \alpha)$
- For each word w_n in a caption, $n \in 1, ..., N$:
  - Sample $z_n \sim Mult(\theta)$
  - Sample word $w_n \sim Mult(\beta_{z_n})$
- For each label b_m in the corresponding business venue categories:
  - Sample $y_m \sim Unif(1, ..., N)$
  - Sample business label $b_m \sim Mult(\beta_{y_m})$ conditioned on the z_{y_m} factor

$Mult(\cdot)$ denotes multinomial distribution and $Unif(\cdot)$ denotes discrete variables with equal probability. During inference, the posterior distribution over latent topics ϕ_n for each word w_n is updated considering the likelihood that each business venue category b_m was generated by the same latent topic as the word. For each business venue category b_m, the approximate posterior distribution over words is updated accordingly. The updates run iteratively until convergence. The word distribution over latent topics (Fig. 2 (d)) is then used to infer the topic distribution of given texts for representing the topic context feature. The number of latent topics is set to 200 empirically in the experiments.

5. CAPTION GENERATION

As shown in Fig. 1, we aim to generate sentence-level captions that reflect user activities at businesses for a given test image. (a) We first compute the topic distribution $\boldsymbol{\delta}$ of the caption in each training image post by using the business-aware

[1] The overall statistics include business categories with few instances.
[2] Less than 25% of the sampled venues have more than 20 images.

topic model. (b) $\boldsymbol{\delta}$ is then used as the context representation of the corresponding caption for learning the language models. (c) For a test image, we use its check-in venue to acquire a set of reference texts, which range from 1 to 10 captions from the other posts at the same venue (depending on the availability) in our experiments. The topic distribution of these reference texts are then used as the context representation for predicting captions of the image. There might be the other ways to collect reference texts for the test image as long as the majority of the reference texts are relevant to users' experience at businesses. Note that the labels of business venue categories are only used in learning topic models (Sec. 4) but are not required for caption generation in testing.

The topic representation $\boldsymbol{\delta}$ is incorporated into the two language models of [8]: multi-modal log bilinear model (MLBL) and its factored model MLBLF. Their performances are compared in the experiments (Sec.6.1). In MLBL, a sentence is generated word by word. The language model itself is a feed-forward neural network which makes a linear prediction of the next word representation $\boldsymbol{r}$ to compute the conditional probability of the next word w_n given previous words $w_1, ..., w_{n-1}$ and the topic context of a given image $\boldsymbol{\delta}$:

$$P(w_n = w \mid w_1, ..., w_{n-1}, \boldsymbol{\delta}) = \frac{exp(\boldsymbol{r}^\top \boldsymbol{r}_w + s_w)}{\sum_j exp(\boldsymbol{r}^\top \boldsymbol{r}_j + s_j)}, \quad (1)$$

$$\boldsymbol{r} = \left(\sum_{i=1}^{n-1} \boldsymbol{C}_i \boldsymbol{r}_{w_i} \right) + \boldsymbol{C}_q \boldsymbol{\delta}. \quad (2)$$

s_w is a word-specific bias. $\boldsymbol{C}_i$ and $\boldsymbol{C}_q$ are parameter matrices for word representation $\boldsymbol{r}_{w_i}$ and context representation $\boldsymbol{\delta}$, respectively. The language models are pre-trained using 50-dimensional scaled word representations from [20].

The MLBL model can further incorporate modality-specific word representations with respect to $\boldsymbol{\delta}$ by a factored model (MLBLF). Let $\boldsymbol{R}$ be a tensor where each matrix $\boldsymbol{R}^{(i)}$ in $\boldsymbol{R}$ is a matrix including the word representation of each word in the vocabulary specific to the i_{th} component in the context feature $\boldsymbol{\delta}$. To incorporate the modality-specific word representations into the language model, $\boldsymbol{R}$ is factorized to three sub matrices $\boldsymbol{W}_{kr}$, $\boldsymbol{W}_{k\delta}$ and $\boldsymbol{W}_{kh}$ [21]. ${\boldsymbol{W}_{kr}}^\top \boldsymbol{W}_{kh}$ is used as the word representation matrix to estimate the next word representation $\boldsymbol{r}$. The estimation is similar to Eq. 2, where $\boldsymbol{r}_{w_i}$ is replaced by $({\boldsymbol{W}_{kr}}^\top \boldsymbol{W}_{kh})_{w_i}$. Let $\boldsymbol{k}$ be $(\boldsymbol{W}_{kr}\boldsymbol{r}) \circ (\boldsymbol{W}_{k\delta}\boldsymbol{\delta})$, the conditional probability is

$$P(w_n = w \mid w_1, ..., w_{n-1}, \boldsymbol{\delta}) = \frac{exp({\boldsymbol{W}_{kh}^{(w)}}^\top \boldsymbol{k} + s_w)}{\sum_j exp({\boldsymbol{W}_{kh}^{(j)}}^\top \boldsymbol{k} + s_j)}. \quad (3)$$

The prediction of the next word is repeated until the ending criteria (cf. the evaluation metrics in Sec. 6) is fulfilled.

6. EXPERIMENTS

We assess image captioning in social media with regard to 1) the performance of the proposed business-aware topic context and the other individual contextual cues, 2) the performance impact of integrating contextual features over various business venue categories and 3) the effects of incorporating business category information for modeling users' activities. The experimental setting is reviewed first to introduce the evaluation metrics and baselines.

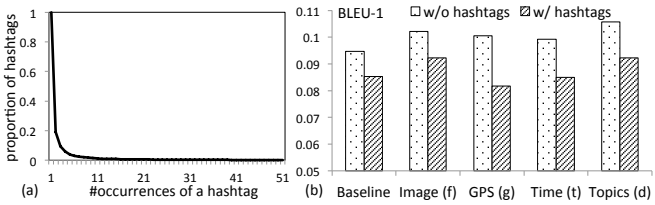

Fig. 3. (a) The statistics of hashtags in the collected data shows the sparsity issue for hashtags. (b) The BLEU-1 of caption generation shows the data with hashtags is more challenging due to data sparsity and results in lower BLEU-1.

Evaluation metrics: BLEU (Bilingual Evaluation Understudy) [22] is an approach to evaluating the relevance between any two descriptions in which a higher score is better. It is also used as the standard quantitative metrics for captioning models [8, 9, 10]. BLEU-1 indicates the BLEU for unigrams. Given a model and an image, we generate a candidate caption and then compute the BLEU score between the candidate and the ground-truth. To ensure the evaluation is independent of the number of words in a caption, the stopping criteria is to generate as many words as the ground-truth, which are the original captions created by social media users. The evaluation process is repeated 5 times and the final BLEU takes the average to account for the variability of generated texts.

Baseline: The first baseline [23] is a language model trained by the captions only without using any context (denoted as Baseline). It is a reference for the improvement gained from the additional context. The second baseline [8] uses the image content as context to generate captions (denoted as Image). To compare with more metadata, we also use GPS logs (denoted as GPS) and time stamps (denoted as Time) of the images as context to generate captions. For the representation of GPS context, we use longitude and latitude that are quantized to each venue. For the representation of time context, we use year, month, day and hour, each normalized by using zero mean and a sigmoid function. Note that, though the contextual cues are different, all four baselines and our method are evaluated on the same user generated captions in social media as introduced in Sec. 3.

6.1. Performance over Contexts

This section compares the proposed business aware topic-based context (denoted as Topic) and the alternative cues including Image, GPS, and Time. For fair comparison, the analysis first investigates the impacts of the data sets with/without hashtags and the two language models MLBL and MLBLF.

Impact of Hashtags: Hashtags pose many challenges to caption generation. In our sampled data (Fig. 3(a)), around

Table 2. The BLEU-1 of the context-augmented language models MLBL and MLBLF using the four context sources associated with the given images.

Models	Image [8]	GPS	Time	Topics
MLBL	0.0880	0.1006	0.0973	**0.1057**
MLBLF	0.1021	0.0929	0.0992	0.0975

80% of hashtags appear only once, which means the lack of training data may lead to very low predictability. Fig. 3(b) compares the BLEU-1 of the data sets with and without hashtags. Performance on the data without hashtags reaches 0.1005 on average while including hashtags results in 0.0873. The relative decrease is around 13% and the declining trend is consistent over the different context types. Since the data sparsity issue has considerable impact on the language models and the relative performance with and without hashtags is similar, the following experiments for contexts and business venue categories are conducted on the data without hashtags.

Impacts of Language Models: The effectiveness of language models varies when incorporating different types of contextual data. In Table 2 we compare two language models MLBL and MLBLF, each fed with the four context representations: image, GPS, time and the business aware topic-based context. The business aware topic context incorporated in the MLBL model outperforms the other individual contexts, followed by image content incorporated in the MLBLF model. The same context used in different language models performs differently; for example, image content with MLBL is worse than with MLBLF, GPS with MLBLF is worse than with MLBL. The obvious difference among the contextual sources is in dimensionality. The image representation has the highest dimension (4,096) over the other contexts and it thus comes with the larger word representation matrix which is further factorized in the MLBLF model. As for the GPS with two dimensional representation, the factorization might not be effective. Due to the different characteristics of each contextual cue, in the following experiments we use the more effective model specific to each for comparison: MLBLF for Image and Time, MLBL for GPS and Topic.

6.2. Performance of Contexts over Business Categories

Different contexts can have varied effectiveness for the images posted from different business venue categories. To analyze the differences, we separate the test data by business venue categories. Besides the evaluation over individual context sources, we assess the use of multiple contexts (denoted as Multiple) by concatenating the business-aware topic features, image features and time features.

The bold numbers in Table 3 highlight the two single contextual cues that obtain the highest BLEU-1 scores. Each context is combined with the language model with the higher BLEU-1 in Table 2 accordingly. For Multiple, because of the high dimension of concatenated context feature, it is incorporated with the language model MLBLF for experiments. For the four single contextual sources in Table 3, the latent business aware topics have more stable contributions over the business categories and appear as the highlighted context for most of the categories. The consistent improvement across business categories might be attributed to the business category information incorporated in the topic modeling and suggests that it can help generate sentences that better reflect user activities at various businesses.

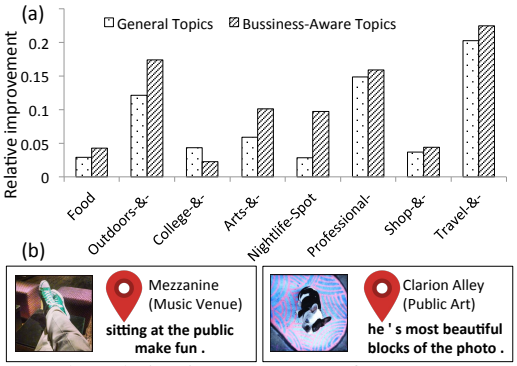

Fig. 4. (a)The relative improvement of BLEU-1 contributed by using general topics and business-aware topics compared to the Baseline. (b) Examples of generated captions for given venue-tagged images by using business-aware topics.

Image content is most effective for the "Travel" and "Outdoor" locations. The results reflect in part potentially repetitive descriptions for scenery photos. GPS improves the "Shop" and "Food" categories; this perhaps can be attributed to the relevance with the business-venue-specific information. The time context seems less informative for the categories where the primary business involves indoor activities such as shops, bars and restaurants. Overall, the effectiveness of the other single-context models is more business-dependent compared to the business aware topic-context model.

The model (Multiple) that combines business-aware topic contexts and the other single contextual cues performs the best in most of the business venue categories compared to any of the individual context features. The only exception is the "College" category, where the improvement gained from each of the single context varies more perhaps due to the smaller scale of data. The numbers in $(\cdot)$ in Table 3 indicate the relative gains of the multiple-context model compared to Baseline [23] and Image [8], respectively. The consistent improvements of the multiple-context model suggest that the topic contexts can complement the other contextual cues such as image content in a multi-modal framework. Overall, Multiple has about 9.7% relative improvement against Image.

6.3. Impacts from Business Awareness

Awareness of business category has potential to improve the representation of users' contexts. In Sec. 4, we introduce a way to incorporate business venue categories in topic mod-

Table 3. The BLEU-1 of the models learned by using different contexts performed on the posts over business venue categories. The bold numbers highlight the two single contexts that contribute the most, where the proposed topic context (Topics) is more obvious and stable than the others. The model (Multiple) that combines Topics and the other contexts performs the best in most of the business categories (underlines). Its relative improvements against Baseline [23] and Image [8] are shown in the two (·).

top-layer business	Baseline [23]	Image [8]	GPS	Time	Topics	Multiple		
Food	0.0961	0.0933	**0.1007**	0.0998	**0.1002**	0.1075	(+12%)	(+15%)
Outdoor & Recreation	0.0942	**0.1069**	0.1040	0.1036	**0.1106**	0.1185	(+26%)	(+11%)
College & University	0.0974	0.0988	**0.1109**	**0.1037**	0.0996	0.0993	(+2%)	(+1%)
Art & Entertainment	0.0947	**0.1022**	0.0995	0.1021	**0.1043**	0.1087	(+15%)	(+6%)
Nightlife Spot	0.0949	**0.1013**	0.0975	0.0948	**0.1041**	0.1075	(+13%)	(+6%)
Professional & Others	0.0951	**0.1045**	0.0983	0.1029	**0.1102**	0.1109	(+17%)	(+6%)
Shop & Service	0.0985	0.0975	**0.1023**	0.0942	**0.1028**	0.1035	(+5%)	(+6%)
Travel & Transportation	0.0895	**0.1143**	0.1080	0.1066	**0.1096**	0.1180	(+32%)	(+3%)

eling. To evaluate how much the awareness of business categories enhances the representation, we compare the general topics learned by LDA [17] without the business information and the business-aware topics modeled by Corr-LDA using the business venue categories as the labels. Fig. 4 shows the relative improvement contributed by using the general topics and the business-aware topics compared to the baseline. The results of using the general topics could reflect whether the topic distributions in the texts collected from the same check-in venue are consistent. The results of the business-aware topics emphasize the improvement as the topics are business-relevant. In most business venue categories, the caption generation shows more relative improvement of BLEU-1 from business-aware topics than general topics.

7. CONCLUSION

In light of the complex and varied factors motivating image posts in social media, we focus on the impacts of the context of user activities at businesses where the posts were created. We propose to model the latent topics driven by user activities at businesses and to incorporate the topic context in an image captioning model. Moreover, we investigate several combinations of data collections and language models to identify the influence of context over multiple business categories. The experiments demonstrate that business-aware topic context features produce captioning performance gains over different business categories that are more stable than the other single context sources. The further improvement by combining the business-aware topic contexts with the other contextual features suggests that they are complementary within a multimodal framework. In the future, we will explore more social activities beyond businesses *e.g.*, the relationships of people appearing in an image or involved behind an image, and understand how these activities influence image captioning.

8. REFERENCES

[1] K. Irby and S. Quinn, "As photos flood our screens, which ones hold our attention?," Poynter, 2015.

[2] E. Holmes, "In photo sharing, every picture tells a story, when it has the right caption," The Wall Street Journal, 2015.

[3] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C Berg, and T. L Berg, "Baby talk: Understanding and generating image descriptions," in *IEEE TPAMI*, 2013.

[4] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015.

[5] A. Aker and R. Gaizauskas, "Generating image descriptions using dependency relational patterns," in *ACL*, 2010.

[6] H. Li and X.-S. Hua, "Melog: mobile experience sharing through automatic multimedia blogging," in *ACM MM Workshop*, 2010.

[7] "Foursquare category hierarchy," `https://developer.foursquare.com/categorytree`.

[8] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *ICML*, 2014.

[9] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015.

[10] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," in *ICLR*, 2015.

[11] X. Fan, A. Aker, M. Tomko, P. Smart, M. Sanderson, and R. Gaizauskas, "Automatic image captioning from the web for gps photographs," in *MIR*, 2010.

[12] W. Nie, X. Wang, Y. Zhao, Y. Gao, Y. Su, and T. Chua, "Venue semantics: Multimedia topic modeling of social media contents," in *PCM*, 2013.

[13] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. Magnenat-Thalmann, "Who, where, when and what: discover spatio-temporal topics for twitter users," in *KDD*, 2013.

[14] T. Tran and S. Choi, "Supervised multimodal topic model for image annotation," in *ICASSP*, 2014.

[15] T. Mikolov and G. Zweig, "Context dependent recurrent neural network language model," in *SLT*, 2012.

[16] F. Chen, D. Joshi, Y. Miura, and T. Ohkuma, "Social media-based profiling of business locations," in *ACM MM GeoMM Workshop*, 2014.

[17] D. Blei and M. Jordan, "Latent dirichlet allocation," in *JMLR*, 2003.

[18] D. Blei and M. Jordan, "Modeling annotated data," in *SIGIR*, 2003.

[19] H. Xiao and T. Stibor, "Toward artificial synesthesia: Linking images and sounds via words," in *NIPS Workshop*, 2010.

[20] J. Turian, L. Ratinov, and Y. Bengio, "Word representations: A simple and general method for semi-supervised learning," in *ACL*, 2010.

[21] M. Ranzato, A. Krizhevsky, and G. E. Hinton, "Factored 3-way restricted boltzmann machines for modeling natural images," in *AISTATS*, 2010.

[22] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002.

[23] A. Mnih and G. Hinton, "Three new graphical models for statistical language modelling," in *ICML*, 2007.