Show and Tell More: Topic-Oriented Multi-Sentence Image Captioning

Yuzhao Mao, Chang Zhou, Xiaojie Wang, Ruifan Li
Center for Intelligence Science and Technology, School of Computer Science,
Beijing University of Posts and Telecommunications
{maoyuzhao,elani,xjwang,rfli}@bupt.edu.cn

Abstract

Image captioning aims to generate textual descriptions for images. Most previous work generates a single-sentence description for each image. However, a picture is worth a thousand words. Single-sentence can hardly give a complete view of an image even by humans. In this paper, we propose a novel Topic-Oriented Multi-Sentence (TOMS) captioning model, which can generate multiple topic-oriented sentences to describe an image. Different from object instances or visual attributes, topics mined by the latent Dirichlet allocation reflect hidden thematic structures in reference sentences of an image. In our model, each topic is integrated to a caption generator with a Fusion Gate Unit (FGU) to guide the generation of a sentence towards a certain topic perspective. With multiple sentences from different topics, our TOMS provides a complete description of an image. Experimental results on both sentence and paragraph datasets demonstrate the effectiveness of our TOMS in terms of topical consistency and descriptive completeness.

1 Introduction

Image captioning, usually generates a single textual description for an image, has grabbed the attention of researchers from various fields. With standard datasets available, such as Flickr8k, Flickr30k and COCO, the Single-Sentence (SS) image captioning has been progressing fast. The state-of-the-art results on COCO have been improved from 20 to 33 in BLEU [Kiros et al., 2014; Vinyals et al., 2015; Xu et al., 2015; You et al., 2016; Gan et al., 2017]. However, a picture is worth a thousand words. It is often insufficient to describe an image with SS. Not only because the same image can be described in many different ways, but also, even humans can hardly use SS to cover all the image contents describing an image. A more reasonable and complete description is using Multiple-Sentence (MS) to caption an image, especially for a semantically rich one.

A few work aimed at achieving more diverse descriptions for image captioning. By adopting a randomly sampled vector in the architecture of conditional variational autoencoder [Wang et al., 2017] or generative adversarial net-work [Dai et al., 2017], their models can capture the uncertainty about what is depicted in an image. For those methods, sentences generated are diversified, however, often describing part of the same object instances within an image. Their objective cannot guarantee the completeness of describing an image. [Johnson et al., 2016; Mao et al., 2016; Krause et al., 2017; Liang et al., 2017] reasoned about describing image regions of interest with MS or paragraph. However, to train those models require additional annotations, such as object instances and their descriptions, which is labor-intensive. Besides, one object a sentence is often monotonous as a description.

Note that, descriptions of an image are annotated by persons with different backgrounds. Those reference sentences are partially overlapped, though, for the most of time, describing an image with different emphases. Together they form a semantically complete description. In this paper, we use Latent Dirichlet Allocation (LDA) to mine topics of interest from textual descriptions, and caption an image more completely according to the mined topics. Textual topic and visual attribute are both semantic concepts. The latter, already adopted in many other captioning models [You et al., 2016; Gan et al., 2017], can be viewed as a bag of independent words showing what is depicted in an image. While the former are bags of correlational words reflecting hidden thematic structure in sentences such as the different emphases in delineating an image. It is natural to describe an image with sentences of emphases. Besides, the correlational words in a topic often cover more than one object instances which may
help enrich the description. Furthermore, topics are discovered in an unsupervised fashion, which makes it possible for MS captioning without additional annotations. As such, we propose a Topic-Oriented Multi-Sentence (TOMS) captioning model for MS captioning, one topic a sentence. In our model, each topic is represented as a topic embedding for guidance. We design a Fusion Gate Unit (FGU) to integrate the topic embeddings into Long Short-Term Memory (LSTM), so that topical consistency can be maintained between the guidance and the generated sentence. Inspired by [Liu et al., 2017], a topic classifier is added to the first step of LSTM for semantic regularization and topic prediction. In our experiments, we set up the evaluation criteria and compare our model with a set of baselines to present the advances of our TOMS.

The main contributions of this paper are as follows. 1) A novel topic-oriented captioning model, TOMS, is proposed to describe an image more completely in MS. 2) An FGU is designed to integrate topic embeddings and maintain topical consistency. 3) Extensive experiments are conducted on both sentence and paragraph datasets to demonstrate the effectiveness of our TOMS in terms of topical consistency and descriptive completeness.

2 Related Work

SS captioning. Traditionally, image captioning was rooted in the community of cognitive science. Approaches to this problem were typically template based [Kulkarni et al., 2013; Kuznetsova et al., 2014]. Extracted visual concepts were filled into predefined templates to generate image descriptions. Those models heavily relied on the extracted concepts and predefined less flexible templates. Therefore, some recent works considered captioning an image directly with an encoder-decoder framework by training recurrent neural network language models conditioned on image features [Kiros et al., 2014; Karpathy and Fei-Fei, 2015; Vinyals et al., 2015; Jia et al., 2015]. However, a single sentence is often partially descriptive to an image. Encoding the entire image to generate a single sentence often suffers from the discrepancy between encoding and decoding. Thus, some approaches to image captioning reasoned about image regions rather than the entire image with the attention mechanism [Xu et al., 2015; You et al., 2016]. To improve the interpretability of the captioning model, [Dong et al., 2017] also introduced the concept of topic for SS captioning. However, SS often describes the partial image and is regarded as an incomplete solution.

MS captioning. To completely depict an image, MS is considered as the viable description. Very recently, some methods generated MS by captioning regions of interest within an image. [Johnson et al., 2016] introduced a dense captioning task, which jointly detected and captioned regions of interest. [Mao et al., 2016] aimed at mutual inference between regions and descriptions giving each region an unambiguous text description. [Krause et al., 2017] and [Liang et al., 2017] generated a paragraph description in which each sentence was region-based. Instead, our TOMS generates different sentences from topics of interest. These textually mined topics can capture linguistic distinctions in describing an image, which is unavailable from the visual side. [Krause et al., 2017; Liang et al., 2017] also used the term of topic. However, it refers to the regions of interest in an image. This meaning of topics is different from that in our TOMS. [Dai et al., 2017] introduced a random vector for controlling the diversity of a sentence with generative adversarial nets. [Wang et al., 2017] use conditional variational auto-encoder to sample the random vector. However, the random diversity, like obtained with in beam-search, lacks clear directionality and cannot guarantee the descriptive completeness. Sentences generated by our TOMS are highly directional guaranteed by an explicit topic embeddings. Mined from all the reference sentences, topics try to cover more content in captioning an image. [Yu et al., 2016] generated MS for only video captioning by capturing strong temporal dependencies which is unavailable from image information.

3 Model

3.1 Formulation

Let I be an image, and S = {w0, ..., wT} be a sentence with T + 1 words. Traditionally, the objective of an image captioning model is to maximize the log likelihood of sentence given image, which is

\[
\log p(S|I) = \sum_{t=1}^{T} \log p(w_t|w_{t-1}, ..., 0, I)
\]

As discussed previously, reference sentences often emphasize different parts of an image. To learn those distinctions, we introduce the topic variable. Let \( z \in \{z_1, ..., z_K\} \) be the topic of a sentence. The objective function is modified as a joint distribution, \( p(S, z|I) \). The log likelihood of \( p(S, z|I) \) can be unfolded into two terms with Bayes, which is

\[
\log p(S, z|I) = \log p(S|z, I)p(z|I) = \log p(S|z, I) + \log p(z|I)
\]

(1)

where

\[
\log p(S|z, I) = \sum_{t=1}^{T} \log p(w_t|w_{t-1}, ..., 0, z, I)
\]

Figure 2 is the architecture of our TOMS. It comprises three main components, LSTM (Sec. 3.2), topic embedding (Sec. 3.3) and FGU (Sec. 3.4), and two outputs corresponding to the two terms in Eq. (1). \( p(S|z, I) \) is to formulate a topic-oriented language model, which aims to learn specific language style of a topic. \( p(z|I) \) is a topic classifier given image. For training, the topic inferred by LDA given a reference sentence, is adopted as not only the guidance for training the language model, but also the label for training the classifier. For testing, the classifier predicts topics to guide language model generating topic-oriented descriptions.

3.2 LSTM

In our model, LSTM is employed to encode image-sentence pairs into representations. Specifically, the LSTM receives the image feature in the first step, then the sentence word by word, which is formulated as

\[
h_t = \begin{cases} \text{LSTM}(x_0) & (t = 0) \\ \text{LSTM}(h_{t-1}, x_t) & (t > 0) \end{cases}
\]

(2)
where $x_0$ is the image feature and $x_t(t > 0)$ is the word embedding. The first hidden state, $h_0$, preserving only visual information is the image representation. It enforces the generated sentence pertinent to the image. $h_t(t > 0)$ is the context representation, which enforces the generated sentence fluent. Both representations are fed to the FGU for further processing.

3.3 Topic Embedding

Let $w = \{w_1, ..., w_V\}$ be the vocabulary with $V$ words, $z = \{z_1, ..., z_K\}$ be the topic set of size $K$, and $d = \{d_1, ..., d_M\}$ be the document set of size $M$ with each reference sentence as a document. LDA defines the generative process for a document $d$ as follows,

- Choose $\theta \sim \text{Dirichlet}(\alpha)$.
- For each of the $N$ words $w_n$ in $d$:
  - Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
  - Choose a word $w_n$ from $P(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic $z_n$.

where $\theta$ is the mixing proportion and is drawn from a Dirichlet prior with parameter $\alpha$. Both $\alpha$ and $\beta$ are hyperparameters for the symmetric Dirichlet distributions that the discrete distributions, per-document topic and per-topic word distribution, are drawn from. The probability of a document $d$ in a corpus is defined as,

$$P(d|\alpha, \beta) = \int P(\theta|\alpha) \prod_{n=1}^{N} P(z_k|\theta)P(w_n|z_n, \beta)d\theta$$

Learning LDA [Griffiths and Steyvers, 2004] on reference sentences provides two sets of parameters: word probabilities given topic $P(w|z)$ and topic probabilities given document $P(z|d)$. By choosing $N$ most probable words under each topic according to $P(w|z)$, we construct each topic embedding as a weighted sum of the top $N$ word embeddings, which is

$$P(w|z_k) = (\phi_{1,k}, \phi_{2,k}, ..., \phi_{N,k})$$

$$\vec{z}_k = \sum_{n=1}^{N} \phi_{n,k} \vec{w}_n$$

where $\vec{w}_n$ and $\vec{z}_k$ are the $n$-th word and $k$-th topic embedding. $\phi_{n,k}$ is the probability of the $n$-th top word in $P(w|z_k)$.

It’s worth mentioning that topic embeddings are updated along with the word embeddings while keeping $\phi_{n,k}$ unchanged. In this case, our TOMS not only maintains the relative importance of the top $N$ words in each topic, but also distributes more weight in training word embeddings for those top $N$ words. Thus, sentence generated tends to describe an image using the top $N$ words of the given topic.

3.4 FGU

We design a Fusion Gate Unit (FGU) to fuse three sources of information from image, context and topic. Specifically, the unit firstly uses Hadamard product to obtain a common representation from image and topic, then concatenate the common representation to a sequence of context representations. To emphasize, different from summation or concatenation that compute OR between two vectors, Hadamard product computes AND by remaining neurons activated in both vectors. Besides, we can balance the forces between keeping sentence fluent and being pertinent to image and topic by regulating the size ratio when concatenating.

The fusion process is mathematically depicted as follows,

$$\tilde{c}_t = \sigma(W_c h_t + b_c)(t > 0)$$

$$\tilde{v} = \sigma(W_v h_0 + b_v)$$

$$\tilde{z} = \sigma(W_z \tilde{z} + b_z)$$

$$f_t = [\tilde{z} \odot \tilde{v}, \tilde{c}_t]$$

where $\tilde{c}_t$, $\tilde{v}$ and $\tilde{z}$ are the representations of context, image and topic transformed by weight matrices $W_c$, $W_v$ and $W_z$ with biases and sigmoid activation function $\sigma$. The weight matrices regulate the vectors size to ensure the lengths of $\tilde{v}$ and $\tilde{z}$ are equal. $\odot$ denotes Hadamard product. $f_t$ is the final fused representation. $\tilde{z}$ is the supervised topic embedding. The topic for constructing the topic embedding is obtained from reference sentence when training (see Sec. 3.5), and from image when inference (see Sec. 3.6).

3.5 Training

Our TOMS outputs a topic classifier given image $p(z|x_0)$ computed by a multi-label logistic regression and a topic-oriented language model $p(x_{t+1}|x_{t},...,z)$ computed by a softmax. The formulae of the two outputs are as follows,

$$p(z_k|x_0) = \sigma(\theta_k^T \tilde{v} + b_k)$$

$$p(x_{t+1}|x_t,...,z) = \text{softmax}(W_f f_t + b_f)$$

Figure 2: The architecture of our TOMS model. It comprises three primary components which are LSTM, topic embedding and FGU. For training, a topic is inferred from a sentence with externally trained LDA. One line of the topic is used as the label to train the topic classifier, the other is represented as topic embedding to train the caption generator. For inference, topics are observed with the topic classifier, then feed to the caption generator generating MS.
where $x_{t\ldots0}$ denotes the context with $x_0$ being the image. $W_f$ and $\theta_k^T$ are the corresponding weight matrix and vector. $\hat{z}$ is the supervised topic embedding. Choosing the topic for training is a sampling process, which is

$$p(1_k = 1|d_m) = \eta_{k,m}$$  \hspace{1cm} (11)

$$\hat{z} = \sum_{k=1}^{K} 1_k \cdot \hat{z}_k$$  \hspace{1cm} (12)

where $d_m$ is the input sentence. $p(z|d_m)$, inferred by LDA, can be viewed as a categorical distribution parameterized by $\{\eta_{k,m}\}$. $1_k$ is an indicator one-hot variable that is set to 1 when the k-th topic is sampled. To avoid sampling irrelevant topics, only topn topics are considered.

The label to train $p(x_{t+1}|x_{t\ldots0}, \hat{z})$ is the ground-truth word, $y_{t+1}$. And the label to train $p(z_k|x_0)$ is the indicator variable $1_k$. Our training loss is a sum of two cross-entropy terms, the sentence loss $\ell_{sent}$ on $p(x_{t+1}|x_{t\ldots0}, \hat{z})$, and the topic loss $\ell_{topic}$ on $p(z_i|x_0)$, which is

$$\ell = \sum_{t=1} \ell_{sent}(p(x_{t+1}|x_{t\ldots0}, \hat{z}), y_{t+1}) + \sum_{i=1} \ell_{topic}(p(z_i|x_0), 1_k)$$

### 3.6 Inference

MS is generated based on topics observed from a given image, one topic a sentence. Specifically, the observed topics are obtained as follows,

1. Normalize Eq.(9) to ensure $\sum_{k=1}^{K} p(z_k|x_0) = 1$;
2. Rank the topics in descending order;
3. Sum up the probabilities of top n topics, initially n=1;
4. Repeat 3 with n+=1 until the sum larger than a threshold;
5. Choose the n topics as the observed topics.

Let $z_k$ be an observed topic. Each sentence is generated by sampling word by word using $p(x_{t+1}|x_{t\ldots0}, z_k)$ until the end of sentence token. It seems fantastic to directly observe topics from a topic model trained on both images and sentences. Note that the reference sentences are invisible during generation. It is impossible to learn sentences distinctions using image based topic model. Therefore, we use $p(z|x_0)$ to approximate $p(z|d)$ for generation.

### 4 Experiment

#### 4.1 Datasets and Metrics

We evaluate our model on two different types of datasets. First are standard datasets, including Flickr8k [Hodosh et al., 2013], Flickr30k [Young et al., 2014] and COCO [Lin et al., 2014] for sentence level MS captioning and second is a paragraph dataset collected by [Krause et al., 2017] for paragraph level MS captioning. The paragraph dataset comprises of 19,551 COCO and Visual Genome images with each annotated with a paragraph description. For making use of paragraph level data in our model, each description is split into sentences. The same preprocessing and data splits as previous works [Karpathy and Fei-Fei, 2015; Krause et al., 2017] are used in our experiments.

#### 4.2 Implementation

We build our code based on pytorch\(^1\). It provides off-the-shelf pre-trained CNN models for extracting image features. We use the resulting fully-connected 2048-dimensional activations from ResNet-152 network as image features. We implement two layers LSTM with each hidden dimension of 512. Both topic and word embedding size are set 256. In FGU, topic and image are 512-dimensional vectors, and 1024 for context representations. Dropout is adopted in both input and output layer. This is a strong implementation of image captioning model. The re-implementation of [Vinyals et al., 2015]'s NIC obtains evaluation scores of CD: 90.6, B@4: 30.2, B@3: 40.7, B@2: 54.4, B@1: 71.1, RG: 52.1, MT: 23.6 on COCO, which outperforms NeuralTalk2\(^2\) and the original implementation.

#### 4.3 MS Captioning Experiment

**Sentence level MS Captioning.** To demonstrate the difference between random and topic-oriented MS captioning, we compare TOMS with NIC and ATT-FCN [You et al., 2016]. Both comparative model randomly generate MS of n sentences with n size beam search. For a fair comparison, n is equal to the number of the observed topics (See Sec.3.6). We present the results in Table 1. Our TOMS outperforms the other models. By observing generated sentences (See Figure 3), we find that, for the most of time, Beam Search only generates outwardly different MS. Those sentences tend to tell the same conspicuous scenes in an image. However, our TOMS can mine inconspicuous scenes with topics to give more complete descriptions. The significant improvement of IC has proved that TOMS depicts more image content than the other models.

**Paragraph level MS Captioning.** For better evaluating the performance of our model, we make a comparison with existing paragraph description models, including RIT-GAN [Liang et al., 2017], Regions-Hierarchical [Krause et al. 2017], GAN [Liang et al., 2017] and RTT-FCN [You et al., 2016]. We use coco-caption\(^3\) to generally evaluate our model. The code adopts four evaluation metrics, including BELU (B@1, B@2, B@3, B@4) [Papineni et al., 2002], METEOR (MT) [Lavie and Agarwal, 2007], ROUGE_L (RG) [Lin, 2010] and CIDEr (CD) [Vedantam et al., 2014].

We also use Instance Coverage (IC) to evaluate the descriptive completeness of generated MS. With Stanford natural language parser\(^4\), we choose notional word, including ‘NN’, ‘NNP’, ‘NNPS’, ‘NNS’, ‘PRP’ as entity instances, ‘VB’, ‘VBD’, ‘VBG’, ‘VBN’, ‘VBP’, ‘VBZ’ as action instances, ‘JJ’, ‘JJR’, ‘JJS’ as attribute instances, from both reference sentences and generated MS. The same instances count only once. Let $C$ be the instances of generated MS and $R$ be instances of reference. IC is computed as,

$$IC = \frac{\text{Count}(C \cap R)}{\text{Count}(R)}$$

### References

1. https://github.com/tylin/coco-caption
Table 1: Results of sentence level MS captioning. All the models generate the same number of sentences. Our TOMS outperforms the other models, especially in term of IC metric.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Models</th>
<th>CD</th>
<th>B@1</th>
<th>B@2</th>
<th>B@3</th>
<th>B@4</th>
<th>RG</th>
<th>MT</th>
<th>IC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flickr8k</td>
<td>NIC</td>
<td>35.8</td>
<td>67.9</td>
<td>48</td>
<td>32.1</td>
<td>21</td>
<td>42.9</td>
<td>19.8</td>
<td>15.3</td>
</tr>
<tr>
<td></td>
<td>ATT-FCN</td>
<td>34.2</td>
<td>63.8</td>
<td>45</td>
<td>30.2</td>
<td>19.3</td>
<td>39.3</td>
<td>18.2</td>
<td>13.6</td>
</tr>
<tr>
<td></td>
<td>TOMS(ours)</td>
<td>37.3</td>
<td>70</td>
<td>49.8</td>
<td>33.2</td>
<td>21.4</td>
<td>44.2</td>
<td>20.1</td>
<td>25.1*</td>
</tr>
<tr>
<td>Flickr30k</td>
<td>NIC</td>
<td>35.8</td>
<td>66.5</td>
<td>45.9</td>
<td>30.3</td>
<td>20.1</td>
<td>41.8</td>
<td>17.3</td>
<td>13.1</td>
</tr>
<tr>
<td></td>
<td>ATT-FCN</td>
<td>35.6</td>
<td>62.3</td>
<td>44.6</td>
<td>30</td>
<td>21.1</td>
<td>40.2</td>
<td>17</td>
<td>12.1</td>
</tr>
<tr>
<td></td>
<td>TOMS(ours)</td>
<td>37.3</td>
<td>69.1</td>
<td>47.2</td>
<td>31.2</td>
<td>20.8</td>
<td>43.3</td>
<td>18.1</td>
<td>24.9*</td>
</tr>
<tr>
<td>COCO</td>
<td>NIC</td>
<td>88</td>
<td>68.9</td>
<td>52.6</td>
<td>38.4</td>
<td>28.3</td>
<td>50.9</td>
<td>22.5</td>
<td>16.2</td>
</tr>
<tr>
<td></td>
<td>ATT-FCN</td>
<td>88.6</td>
<td>69.3</td>
<td>53.2</td>
<td>39.4</td>
<td>28.7</td>
<td>51.4</td>
<td>23.4</td>
<td>15.3</td>
</tr>
<tr>
<td></td>
<td>TOMS(ours)</td>
<td>90.3</td>
<td>72.3</td>
<td>54.1</td>
<td>39.2</td>
<td>28.9</td>
<td>52.1</td>
<td>23</td>
<td>28.6*</td>
</tr>
</tbody>
</table>

Table 2: Results of paragraph level MS captioning on paragraph dataset. R-H and S-C denotes Regions-Hierarchical and Sentence-Concat, respectively.

<table>
<thead>
<tr>
<th>Models</th>
<th>CD</th>
<th>B@1</th>
<th>B@2</th>
<th>B@3</th>
<th>B@4</th>
<th>MT</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-C</td>
<td>6.8</td>
<td>31.1</td>
<td>15.1</td>
<td>7.6</td>
<td>4</td>
<td>12.1</td>
</tr>
<tr>
<td>R-H</td>
<td>13.5</td>
<td>41.9</td>
<td>24.1</td>
<td>14.2</td>
<td>8.7</td>
<td>16</td>
</tr>
<tr>
<td>RTT-GAN</td>
<td>20.4</td>
<td>42.1</td>
<td>25.4</td>
<td>14.9</td>
<td>9.2</td>
<td>18.4</td>
</tr>
<tr>
<td>TOMS(ours)</td>
<td>20.8</td>
<td>43.1</td>
<td>25.8</td>
<td>14.3</td>
<td>8.4</td>
<td>18.6</td>
</tr>
</tbody>
</table>

Table 3: Topical consistency results of our TOMS compared with baselines in topic groups. Higher is better.

It generates the same description in every topic groups. The poor performance of NIC is due to the lack of topic information. The same single description may obtain a high score in one topic group and low scores in the others.

NIC-MS comprises of sub-NIC models trained separately on image-sentence pairs of each topic. Compared with NIC, the outperformance of NIC-MS indicates the effectiveness of topic in discovering sentences distinctions.

4.4 Topical Consistency
Topical consistency and descriptive completeness are the two main features of our TOMS. We use IC to evaluate the descriptive completeness (See Table 1). For topical consistency, we set up experiment as follows. We choose the most probable topic in Eq.(11) to annotate each reference sentence with a topic label. Those labels cluster reference sentences of an image into groups, namely topic groups. Each description is generated based on the topic label. For instance, given an image, three reference sentences annotated with the same topic form a topic group. Our TOMS generates description of that topic and use coco-caption to compute evaluation metrics with reference sentences in that topic group. Scores in all topic groups are averaged for the final evaluation. We set up three baselines and present the results in Table 3.
NIC [Vinyals et al., 2015] is a classical SS captioning model.
4.5 Qualitative Results

**Descriptions based on different topics.** In this experiment, we example descriptions of an image generated by TOMS on different observed topics. We also list descriptions generated by NIC with beam search for comparison. The examples are depicted in Figure 3. We notice that descriptions generated by NIC are outwardly different describing the same scenes in an image. On the other hand, each description generated by our TOMS has a major theme and depict different content within the image. We highlight topic words of the supervised topic with colors. By raising the probabilities of these topic words, our TOMS caption an image with a specific topic perspective.

**Descriptions based on the same topic.** In this experiment, we example descriptions of different images generated by TOMS based on the same observed topic. Firstly, we randomly choose a topic from the topic set and list 10 most probable topic words to represent the chosen topic. Then, we randomly choose three images that have the chosen topic observed. Description of each image is generated based on the same chosen topic. Notice that, all the three images involve the same scene of using cell phone, which is also mentioned in the topic words of the chosen topic. See Figure 4, all the descriptions generated by our model correctly describe this scene, however, NIC only describes the conspicuous scenes in images.

**Topic Exploration.** One of the benefits for topic model is the interpretability. We can visualize topics with clusters of topic words. Table 4 are the topic words of several topics obtained by LDA on COCO. We notice that topics are diversified. Scenes like building, elephant and walking are described by different topics from different viewpoints.

### Table 4: Example topics of 100 topics LDA on COCO

<table>
<thead>
<tr>
<th>IDs</th>
<th>Topic Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>#7</td>
<td>front outside building house store nearby shop</td>
</tr>
<tr>
<td>#46</td>
<td>trees near surrounded buildings tall bushes around</td>
</tr>
<tr>
<td>#53</td>
<td>street city busy buildings corner crossing crowded</td>
</tr>
<tr>
<td>#58</td>
<td>light traffic signs pole city intersection buildings</td>
</tr>
<tr>
<td>#61</td>
<td>clock building tower tall brick church roof stone</td>
</tr>
<tr>
<td>#29</td>
<td>across along towards moving path elephants walks</td>
</tr>
<tr>
<td>#71</td>
<td>elephant baby nurse adult trunk mother feeding</td>
</tr>
<tr>
<td>#41</td>
<td>water elephants drinking pool swimming bushes</td>
</tr>
<tr>
<td>#8</td>
<td>umbrella walking pink rain underneath purple beach sand walking surfboards sandy chairs ocean</td>
</tr>
<tr>
<td>#16</td>
<td>people several waiting line cross walk lined wait</td>
</tr>
<tr>
<td>#36</td>
<td>side road dirt walk car crossing gravel light passing</td>
</tr>
</tbody>
</table>

5 Conclusion

This work proposes a novel TOMS captioning model. Topics, represented as topic embedding, are used to arrange the MS generation. A FGU is designed to combine information from topic, image and context. Compared with baselines, our TOMS performs better. Experimental results prove that topic-oriented MS can better capture the details of an image than SS. Evaluations demonstrate the robustness, effectiveness and interpretability of our models.

Acknowledgments

This paper is supported by NSFC (No.61273365), NSSFC(2016ZDA055), 111 Project (No. B08004), Beijing Advanced Innovation Center for Imaging Technology, Engineering Research Center of Information Networks of MOE, China.
References


