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

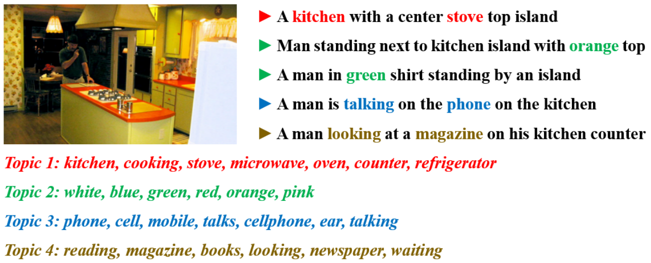
Abstract

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 auto-encoder [Wang *et al.*, 2017] or generative adversarial network [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.



▶ A kitchen with a center stove top island
▶ Man standing next to kitchen island with orange top
▶ A man in green shirt standing by an island
▶ A man is talking on the phone on the kitchen
▶ A man looking at a magazine on his kitchen counter

Topic 1: kitchen, cooking, stove, microwave, oven, counter, refrigerator
Topic 2: white, blue, green, red, orange, pink
Topic 3: phone, cell, mobile, talks, cellphone, ear, talking
Topic 4: reading, magazine, books, looking, newspaper, waiting

Figure 1: An image with human-annotated reference sentences. The colors denote LDA-inferred topics in sentences and words. We notice that sentences with different topics depict the image with different emphases.

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 (See Figure 1). 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 $S = \{w_0, ..., w_T\}$ 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

$$\begin{aligned} \log p(S, z|I) &= \log p(S|z, I)p(z|I) \\ &= \log p(S|z, I) + \log p(z|I) \end{aligned} \tag{1}$$

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

$$\boldsymbol{h}_t = \begin{cases} LSTM(\boldsymbol{x}_0) & (t = 0) \\ LSTM(\boldsymbol{h}_{t-1}, \boldsymbol{x}_t) & (t > 0) \end{cases} \tag{2}$$

where $\boldsymbol{x}_0$ is the image feature and $\boldsymbol{x}_t(t > 0)$ is the word embedding. The first hidden state, $\boldsymbol{h}_0$, preserving only visual information is the image representation. It enforces the generated sentence pertinent to the image. $\boldsymbol{h}_t(t > 0)$ 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 $\boldsymbol{w} = \{w_1, ..., w_V\}$ be the vocabulary with V words, $\boldsymbol{z} = \{z_1, ..., z_K\}$ be the topic set of size K, and $\boldsymbol{d} = \{d_1, ..., d_M\}$ 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 $\theta \sim Dirichlet(\alpha)$.
- For each of the N words w_n in d:
  - Choose a topic $z_n \sim Multinomial(\theta)$.
  - Choose a word w_n from $P(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

where θ is the mixing proportion and is drawn from a $Dirichlet$ prior with parameter α. Both α and β 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,

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

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

$$p(\boldsymbol{w}|z_k) = (\phi_{1,k}, \phi_{2,k}, ..., \phi_{N,k}) \tag{3}$$

$$\vec{\boldsymbol{z}}_k = \sum_{n=1}^{N} \phi_{n,k} \vec{\boldsymbol{w}}_n \tag{4}$$

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

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 representations 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,

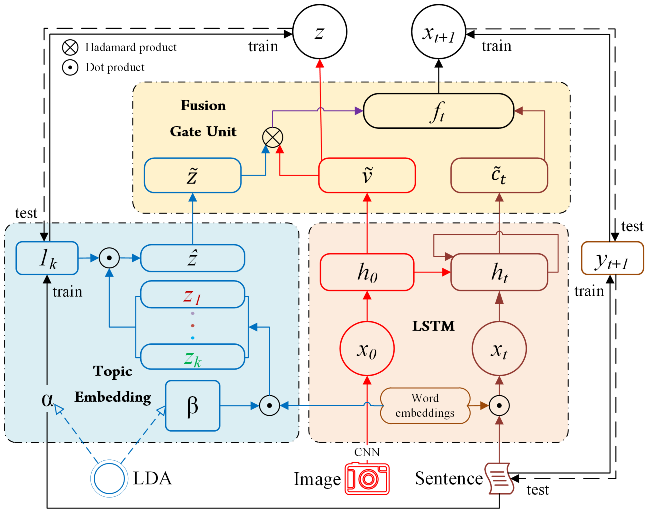


Figure 2: The architecture of our *TOMS* model. It comprises three primary components which are LSTM, topic embedding and FGU. For training, a topic is inferred from a sentence with externally trained LDA. One line of the topic is used as the label to train the topic classifier, the other is represented 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.

$$\tilde{\boldsymbol{c}}_t = \sigma(\mathcal{W}_c \boldsymbol{h}_t + \boldsymbol{b}_c)(t > 0) \tag{5}$$

$$\tilde{\boldsymbol{v}} = \sigma(\mathcal{W}_v \boldsymbol{h}_0 + \boldsymbol{b}_v) \tag{6}$$

$$\tilde{\boldsymbol{z}} = \sigma(\mathcal{W}_z \hat{\boldsymbol{z}} + \boldsymbol{b}_z) \tag{7}$$

$$\boldsymbol{f}_t = [\tilde{\boldsymbol{z}} \odot \tilde{\boldsymbol{v}}, \tilde{\boldsymbol{c}}_t] \tag{8}$$

where $\tilde{\boldsymbol{c}}_t$, $\tilde{\boldsymbol{v}}$ and $\tilde{\boldsymbol{z}}$ are the representations of context, image and topic transformed by weight matrices $\mathcal{W}_c$, $\mathcal{W}_v$ and $\mathcal{W}_z$ with biases and sigmoid activation function σ. The weight matrices regulate the vectors size to ensure the lengths of $\tilde{\boldsymbol{v}}$ and $\tilde{\boldsymbol{z}}$ are equal. $\odot$ denotes Hadamard product. $\boldsymbol{f}_t$ is the final fused representation. $\hat{\boldsymbol{z}}$ 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 $p(\boldsymbol{z}|x_0)$ computed by a multi-label logistic regression and a topic-oriented language model $p(\boldsymbol{x}_{t+1}|x_{t,..,0}, \hat{\boldsymbol{z}})$ computed by a softmax. The formulae of the two outputs are as follows,

$$p(z_k|x_0) = \sigma(\boldsymbol{\theta}_k^T \tilde{\boldsymbol{v}} + b_k) \tag{9}$$

$$p(\boldsymbol{x}_{t+1}|x_{t,..,0}, \hat{\boldsymbol{z}}) = softmax(\mathcal{W}_f \boldsymbol{f}_t + \boldsymbol{b}_f) \tag{10}$$

where $x_{t,..,0}$ denotes the context with x_0 being the image. $\mathcal{W}_f$ and $\boldsymbol{\theta}_k^T$ are the corresponding weight matrix and vector. $\hat{\boldsymbol{z}}$ is the supervised topic embedding. Choosing the topic for training is a sampling process, which is

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

$$\hat{\boldsymbol{z}} = \sum_{k=1}^{K} 1_k \cdot \vec{\boldsymbol{z}}_k \tag{12}$$

where d_m is the input sentence. $p(\boldsymbol{z}|d_m)$, 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 $p(\boldsymbol{x}_{t+1}|x_{t,..,0}, \hat{\boldsymbol{z}})$ is the ground-truth word, $\boldsymbol{y}_{t+1}$. 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 ℓ_{sent} on $p(\boldsymbol{x}_{t+1}|x_{t,..,0}, \hat{z})$, and the topic loss ℓ_{topic} on $p(z_i|x_0)$, which is

$$\ell = \sum_{t=1} \ell_{sent}(p(\boldsymbol{x}_{t+1}|x_{t,..,0}, \hat{\boldsymbol{z}}), \boldsymbol{y}_{t+1}) + \sum_{i=1} \ell_{topic}(p(z_i|x_0), 1_k)$$

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 $p(\boldsymbol{x}_{t+1}|x_{t,..,0}, z_k)$ 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 $p(\boldsymbol{z}|x_0)$ to approximate $p(\boldsymbol{z}|\boldsymbol{d})$ 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.

We use coco-caption[1] 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[2], 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,

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

4.2 Implementation

We build our code based on pytorch[3]. 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[4] 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 *RTT-GAN* [Liang *et al.*, 2017], *Regions-Hierarchical* [Krause *et*

[1] https://github.com/tylin/coco-caption

[2] https://nlp.stanford.edu/software/lex-parser.shtml

[3] http://pytorch.org/

[4] https://github.com/karpathy/neuraltalk2

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

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.

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

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

al., 2017] and its baselines, *Sentence-Concat*. To adapt our *TOMS* for paragraph level MS captioning, we also employ sentence concatenation. Different from *Sentence-Concat* that concatenates randomly generated sentences, our *TOMS* concatenates topic-oriented sentences which is proven more advanced in the previous experiment. The results are presented in Table 2. Our *TOMS* challenges the other paragraph descriptive models, even though the connection between sentences is not considered. The reason is that a paragraph description depicts more details with each sentence describing a specific theme. Our *TOMS* is good at capturing those textual details. Organized by observed topics, one topic a sentence the details are described by the generated MS. The combination of the MS can comprehensively describe the image despite the weakness in connecting sentences.

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. 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.

gLSTM [Jia *et al.*, 2015] is a modified LSTM for integrating semantic information by element-wise addition. We use *gLSTM* to replace FGU for comparison. The results prove the advance of FGU. Since a topic often contains several irrelevant words in captioning an image, FGU can better capture the common information and lead to better topical consistency.

SCN-RNN [Gan *et al.*, 2017] combines topic embedding with both hidden and input neurons by Hadamard product. We use *SCN-RNN* to take place FGU for comparison. With the same Hadamard product, FGU combines topic embedding with image feature, then concatenates to hidden neurons. The results demonstrate the advances of using image features to filter topic embedding for better topical consistency.

Datasets	Models	CD	B@4	RG	MT
Flickr8k	NIC	40	7.1	30	12.4
	NIC-MS	51	7.9	31.8	13
	gLSTM	51.6	7.5	32	13.2
	SCN-RNN	53.2	7.9	33.3	13.6
	TOMS(ours)	**55.6**	**8.5**	**34.2**	**13.9**
Flickr30k	NIC	24.6	6.3	27.7	11
	NIC-MS	35.7	7.8	28.8	11.7
	gLSTM	36	8	28.1	11.3
	SCN-RNN	38.8	8.5	29.2	11.6
	TOMS(ours)	**48.3**	**9.1**	**31.4**	**12.4**
COCO	NIC	54.6	8.9	33.5	13.6
	NIC-MS	60.2	10.3	34.8	14
	gLSTM	61.1	10.8	35.2	14.3
	SCN-RNN	63.8	11.3	36.6	15.1
	TOMS(ours)	**88.6**	**12.6**	**38.1**	**16.8**

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

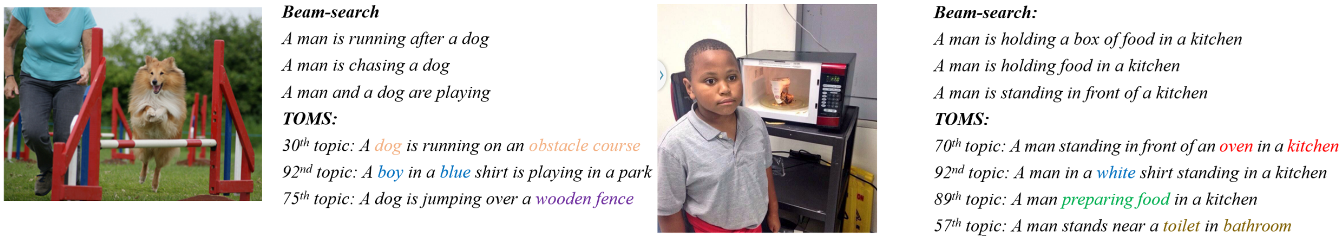

Beam-search
A man is running after a dog
A man is chasing a dog
A man and a dog are playing
TOMS:
30th topic: A dog is running on an obstacle course
92nd topic: A boy in a blue shirt is playing in a park
75th topic: A dog is jumping over a wooden fence
Beam-search:
A man is holding a box of food in a kitchen
A man is holding food in a kitchen
A man is standing in front of a kitchen
TOMS:
70th topic: A man standing in front of an oven in a kitchen
92nd topic: A man in a white shirt standing in a kitchen
89th topic: A man preparing food in a kitchen
57th topic: A man stands near a toilet in bathroom

Figure 3: Qualitative results of descriptions based on different topics. Descriptions generated by beam-search is listed for comparison. Topic words of each supervised topic are highlighted with the same color. 57th topic in the right side is an incorrect observed topic

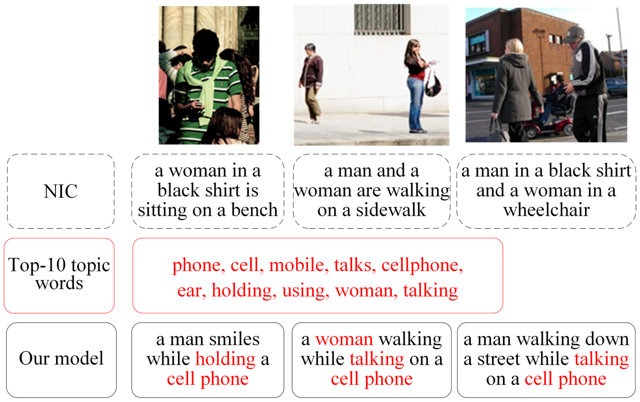

NIC
a woman in a black shirt is sitting on a bench
a man and a woman are walking on a sidewalk
a man in a black shirt and a woman in a wheelchair
Top-10 topic words
phone, cell, mobile, talks, cellphone, ear, holding, using, woman, talking
Our model
a man smiles while holding a cell phone
a woman walking while talking on a cell phone
a man walking down a street while talking on a cell phone

Figure 4: Qualitative results of descriptions based on the same topic. The top is descriptions generated by the *NIC*. The middle is the topic words of the sampled topic. And the bottom is descriptions generated by our model supervised by the sampled topic.

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 the same scene, such as building, is described from multiple perspectives. With the guidance of these topics, our model describes an image from different topic perspectives.

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

Table 4: Example topics of 100 topics LDA on COCO. We can see that topics are diversified. Scenes like *building*, *elephant* and *walking* are described by different topics from different viewpoints.

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

[Dai *et al.*, 2017] Bo Dai, Dahua Lin, Raquel Urtasun, and Sanja Fidler. Towards diverse and natural image descriptions via a conditional gan. *CVPR*, 2017.

[Dong *et al.*, 2017] Yinpeng Dong, Hang Su, Jun Zhu, and Bo Zhang. Improving interpretability of deep neural networks with semantic information. *CVPR*, 2017.

[Gan *et al.*, 2017] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *CVPR*, volume 2, 2017.

[Griffiths and Steyvers, 2004] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[Hodosh *et al.*, 2013] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[Jia *et al.*, 2015] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. Guiding long-short term memory for image caption generation. *ICCV*, 2015.

[Johnson *et al.*, 2016] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. 2016.

[Karpathy and Fei-Fei, 2015] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.

[Kiros *et al.*, 2014] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Multimodal neural language models. In *ICML*, volume 14, pages 595–603, 2014.

[Krause *et al.*, 2017] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. *CVPR*, 2017.

[Kulkarni *et al.*, 2013] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013.

[Kuznetsova *et al.*, 2014] Polina Kuznetsova, Vicente Ordonez, Tamara L Berg, and Yejin Choi. Treetalk: Composition and compression of trees for image descriptions. *TACL*, 2(10):351–362, 2014.

[Lavie and Agarwal, 2007] Alon Lavie and Abhaya Agarwal. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231. Association for Computational Linguistics, 2007.

[Liang *et al.*, 2017] Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. Recurrent topic-transition gan for visual paragraph generation. *ICCV*, 2017.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. pages 740–755, 2014.

[Lin, 2010] Lin. Rouge: A package for automatic evaluation of summaries. 2010.

[Liu *et al.*, 2017] Feng Liu, Tao Xiang, Timothy M Hospedales, Wankou Yang, and Changyin Sun. Semantic regularisation for recurrent image annotation. *CVPR*, 2017.

[Mao *et al.*, 2016] Junhua Mao, Huang Jonathan, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. *CVPR*, 2016.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. Association for Computational Linguistics, 2002.

[Vedantam *et al.*, 2014] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *Computer Science*, pages 4566–4575, 2014.

[Vinyals *et al.*, 2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015.

[Wang *et al.*, 2017] Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *NIPS*, pages 5758–5768, 2017.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

[You *et al.*, 2016] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. *CVPR*, 2016.

[Young *et al.*, 2014] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[Yu *et al.*, 2016] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *CVPR*, pages 4584–4593, 2016.