# Topic-Oriented Image Captioning Based on Order-Embedding

Niange Yu, *Student Member, IEEE*, Xiaolin Hu, *Senior Member, IEEE*, Binheng Song, Jian Yang, and Jianwei Zhang

***Abstract*— We present an image captioning framework that generates captions under a given topic. The topic candidates are extracted from the caption corpus. A given image's topics are then selected from these candidates by a CNN-based multi-label classifier. The input to the caption generation model is an image-topic pair, and the output is a caption of the image. For this purpose, a cross-modal embedding method is learned for the images, topics, and captions. In the proposed framework, the topic, caption, and image are organized in a hierarchical structure, which is preserved in the embedding space by using the order-embedding method. The caption embedding is upper bounded by the corresponding image embedding and lower bounded by the topic embedding. The lower bound pushes the images and captions about the same topic closer together in the embedding space. A bidirectional caption-image retrieval task is conducted on the learned embedding space and achieves the state-of-the-art performance on the MS-COCO and Flickr30K datasets, demonstrating the effectiveness of the embedding method. To generate a caption for an image, an embedding vector is sampled from the region bounded by the embeddings of the image and the topic, then a language model decodes it to a sentence as the output. The lower bound set by the topic shrinks the output space of the language model, which may help the model to learn to match images and captions better. Experiments on the image captioning task on the MS-COCO and Flickr30K datasets validate the usefulness of this framework by showing that the different given topics can lead to different captions describing specific aspects of the given image and that the quality of generated captions is higher than the control model without a topic as input. In addition, the proposed method is competitive with many state-of-the-art methods in terms of standard evaluation metrics.**

***Index Terms*— Image captioning, topic, order-embedding, cross-modal retrieval.**

Manuscript received February 21, 2018; revised November 4, 2018; accepted December 17, 2018. Date of publication December 27, 2018; date of current version March 21, 2019. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700904, in part by the National Natural Science Foundation of China under Grant 61332007, Grant 61621136008, and Grant 61620106010, and in part by the German Research Council (DFG) under Grant TRR-169. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Vishal Monga. *(Corresponding author: Xiaolin Hu.)*

N. Yu and X. Hu are with the State Key Laboratory of Intelligent Technology and Systems, Beijing National Research Center for Information Science and Technology, Department of Computer Science and Technology, Institute for Artificial Intelligence, Tsinghua University, Beijing 100084, China (e-mail: xlhu@tsinghua.edu.cn).

B. Song is with the Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China (e-mail: yng15@mails.tsinghua.edu.cn).

J. Yang is with the Department of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, China.

J. Zhang is with the Department of Informatics, University of Hamburg, D-22527 Hamburg, Germany.

Digital Object Identifier 10.1109/TIP.2018.2889922

## I. Introduction

NATURAL language has emerged as an elegant intermediary in visual understanding. Compared with a fixed set of visual categories, it can provide an unconstrained and far richer description of the visual scenes. Automatically generating natural language descriptions for the image, i.e. image captioning, is a challenging task. Over the past few years, the encoder-decoder framework that engages deep neural networks has largely driven the progress in this field [1]–[4]. Generally, they use a multi-modal language model that takes an image as input and generates a sentence as output.

A natural image always contains rich semantic information. It is hard to make a full description in a single caption. Typically, each caption describes the image from a "specific topic", which makes them very diverse as shown in Fig. 1. In most previous works, the language model is trained to map the image to one of those captions at each iteration. This requires the model to be complex enough to accommodate all of the captions without interfering with each other, which poses difficulties in designing and training the model.

To tackle this problem, we propose to learn a fine-grained mapping from the image-topic pairs to the captions. In our model, a topic is provided as an additional input to the language model as illustrated in Fig. 1, the corresponding caption should be closely related to the given topic. If we define the "output space" of a well-trained model as the set of all captions that the model could generate given some input, then the output space of our model is constrained such that the generated caption should be about not only the content of the image but also the given topic. With smaller output space, it would be easier for the model to learn better matching between images and captions, which may lead to more accurate captions. When doing inference or testing, given different topics, the model can generate different captions. It is therefore called *topic-oriented image captioning*.

In addition, since the content of the generated caption is controllable, this framework has great potential in applications that require human-computer interaction. In those applications, the desired caption should not only be related to the image but also reflect what the user wants to emphasize.

In this framework, during training, the topics are extracted from the ground-truth caption corpus by a topic model as in a previous work [5]. While during testing, the topics of the given image are detected by an image topic detector (ITD). The input of the language model is from a cross-modal embedding space learned for the given images, topics, and the ground-truth

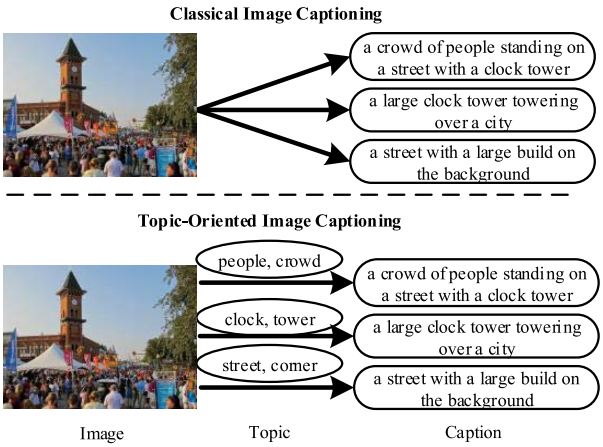



Fig. 1. Illustration of classical image captioning and topic-oriented image captioning processes. For the latter, the caption generator is guided by the given topics.

captions. This embedding method is an extension of previous work called order-embedding [6], where the two modalities are image and caption, by introducing the topic as the third modality. This additional modality reduces the size of the output space of the language model for caption prediction (Fig. 2).

To show the effectiveness of the embedding method, we conduct the bidirectional caption-image retrieval task in the embedding space. The proposed method achieves the state-of-the-art results on two benchmark datasets MS-COCO [7] and Flickr30K [8]. Based on the embedding results, the topic-oriented image captioning method can produce topic-specified captions, and the quality of the captions is comparable to some state-of-the-art image captioning methods.

To summarize, the main contributions of this work are as follows:

- We propose to encode the images, captions and the topics jointly based on the order-embedding method.
- We put the image captioning task in a new scenario, describing an image focusing on specific topics.

Experiments on two benchmark datasets MS-COCO [7] and Flickr30K [8] have validated the effectiveness of the proposed order-embedding method and topic-oriented image captioning method.

## II. Related Work

### A. Image Captioning

Generating descriptions of natural language for images has been studied for a long time. In recent years, the encoder-decoder framework [1] has become dominant in this field. It uses a CNN as the encoder to produce the image representation, which is fed into a decoder. The decoder is usually an RNN that generates a sentence of variable length word-by-word. The image representation from the encoder only captures the global information of the whole image, while every single word in a sentence predicted by the model usually relates to a specific aspect of the image. Xu *et al.* [2] introduce the attention mechanism to solve this problem. Before word prediction at each step, the attention module is used to produce a mask on the image feature map which reflects the "attention" for the current word being predicted. Many other models are proposed to improve the attention mechanism and drive the progress in this field [3], [9]–[11].

Instead of CNN features, some works use more semantic information extracted from the image as the input to the decoder [3], [12]. First, a set of attributes (a set of frequently used words in the caption corpus) are extracted from the dataset. Then a model (acting as the "encoder") is trained to predict the attributes for the given image. The "encoder" would produce a probability distribution of the attributes. This distribution is used to replace or accompany the CNN features as input to the decoder. Compared to the CNN features, the attribute distribution has a more semantic meaning.

The traditional training paradigm in image caption aims to fit the output of the model to a specific ground-truth sentence. Achieving this goal is not identical to improving the evaluation metrics usually used to judge the performance of models. Some methods are proposed to optimize the model directly based on certain evaluation metrics [13]–[15]. As the metrics are usually non-differentiable, almost all of these methods use reinforcement learning for optimization.

Gan *et al.* [16] propose to detect some semantic concepts from the image contents which are then used to aid the caption generation process. The semantic concepts are words simply picked from the caption corpus according to their frequencies. An empirical study in their work shows that the given semantic concepts have a larger probability to be seen in the generated caption than not given the semantic concepts. Different from their model, our model uses a topic model to extract semantic concepts from the caption corpus which can better capture the latent semantic information. The given semantic concept is embedded in the caption generated by the proposed approach.

### B. Cross-Modal Embedding

Cross-modal embedding, typically with image and text, has been a basis in a large amount of cross-modal research [17]–[19]. Most of the previous work uses Euclidean or cosine distance as the similarity measurement in the embedding space, which results in a symmetrical relationship between different modalities [18]. However, sometimes, an unsymmetric relationship makes more sense, e.g., caption and image. A caption is an abstraction of an image, and there exists a hypernym relationship between them. It is proposed to build this relationship as a partial order [6], and a coordinate-wise order is enforced in the embedding space. In our framework, there are three partial order relationships: topic-caption, topic-image, and caption-image. We need to preserve a three-level semantic hierarchy structure in the embedding space based on three partial orders.

### C. Topic Models

Topic models are widely used in text mining to discover the abstract "topics" in a collection of documents [20]–[23]. The documents are represented by the distributions of words.

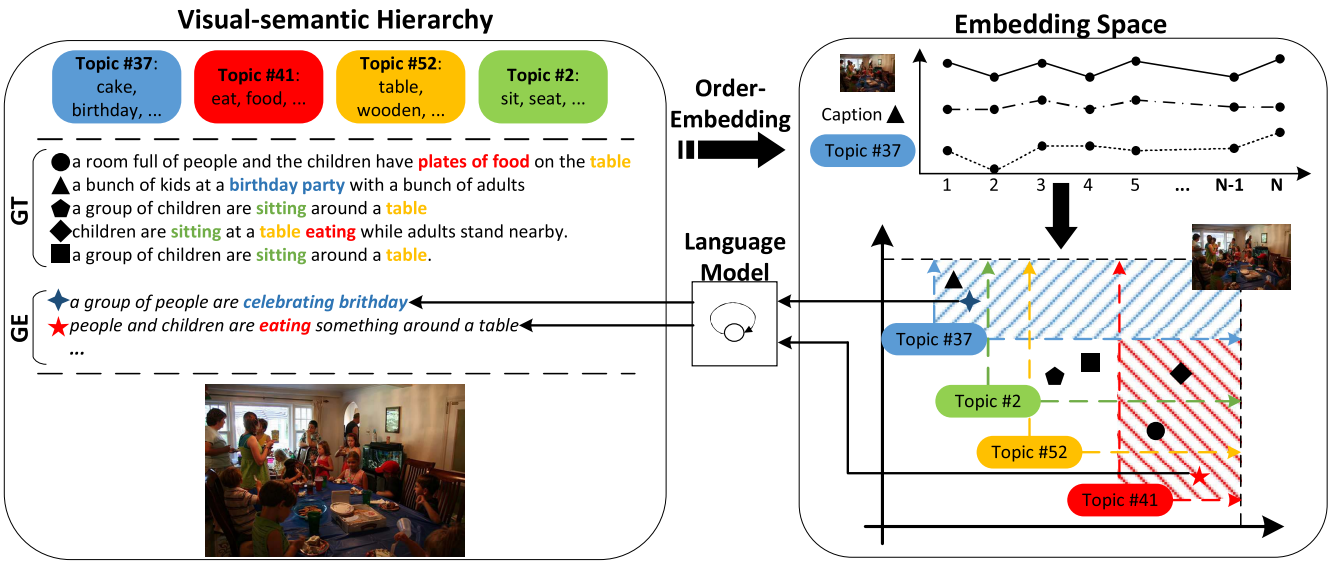



Fig. 2. The pipeline of the proposed method. There are four topics and five ground-truth captions (denoted as "GT") on the left for a single image, which are mapped to the N-dimensional embedding space by order-embedding. As shown at the top right corner, the hierarchical relationships are preserved at each coordinate. The vertical axis refers to the value of each coordinate. The relationships are further illustrated in the 2-dimensional space (bottom right). At caption generation stage, a point is sampled from a region in the embedding space bounded by the coordinates of the image embedding and a topic embedding, then decoded by the language model to generate a sentence (denoted as "GE").

Each document is viewed as a mixture of topics, and the distribution of words characterizes each topic.

One of the most popular topic models is Latent Dirichlet Allocation (LDA) [22]. It is a hierarchical Bayesian model which assumes that the distribution of topics over the documents is a Dirichlet distribution. Another widely-used topic model is non-negative matrix factorization (NMF) [23]. The input to NMF is a document-by-term matrix $V$ whose rows contain the representations for each document. NMF is used to find a decomposition of $V$ into two matrices: $H$ and $W$, where $W$ is a topic-by-term matrix whose rows are the topic representations and $H$ is a document-by-topic matrix whose rows are the topic distributions of each document. When the topic representation $W$ has been learned from the training set, it can be used to extract topics from the new documents.

## III. Methods

In the proposed framework, the topics, captions, and images are organized in a three-level visual-semantic hierarchical structure with the topic at the top and image at the bottom (Fig. 2). They are embedded in the same space by the order-embedding method, where the hierarchical structure is preserved coordinate-wise. Given an image embedding and a topic embedding, there would be a subspace bounded by them. The embeddings of target captions should be constrained in the subspace. The language model is trained to sample a point in the subspace and decode it to generate the target caption. Fig. 2 illustrates the framework.

In the following, we first present the topic models used for topic extraction in captions and use ITD to detect the topics in images. Then, the order-embedding method is introduced to model the hierarchical structure of the topic-caption-image. Finally, we describe the LSTM-based language model which is used to generate the captions.

### A. Topic Learning

For each image in the training set, we concatenate its ground-truth captions to form a document. The documents are then transformed into tf or tf-idf features as the input to the topic model. Firstly, the documents are tokenized into words and then stemmed by the poster stemming algorithm [24] to reduce semantically similar words to their root. Then, all of the words that appear there are sorted by their frequencies. We remove the "stop words" and select the top $5,000$ words to build a vocabulary. Finally, the tf or tf-idf vector is built for each document.

We use two topic models (i.e., LDA and NMF) to extract topics from documents. Tf-idf is better than tf in most cases since it can reduce the influence of the common words and concentrate more on the "distinguish words". The disadvantage of tf-idf is that it is continuous while LDA only supports discrete input. In our experiments, the tf vector is used for LDA, and the tf-idf vector is used for NMF.

During training on a topic-oriented image captioning task, since both images and ground-truth captions are given, the topics can be extracted from the captions by a topic model. However, during testing, images are given without captions, and one has to extract topics from images. We formulate this as a multi-label classification problem (an image can have multiple topics) and use a CNN-based ITD to solve it. It takes an image as input and ends up with a sigmoid transform layer as the $pred$ layer, which produces an $N_T$-dim vector as the topic prediction. The model is trained on the training split of the image-caption dataset using the cross-entropy loss.

The "ground truth" is set to the topics extracted from the images' captions.

### B. Embedding Learning

We use the order-embedding method to learn the embeddings of the image, caption, and topic. Order-embedding is originally proposed for the caption-image retrieval task where it learns the embeddings according to the partial order relationship between the image and caption. In our case, there exist three partial order relationships: topic-caption, topic-image, and caption-image. These three relationships are considered jointly in our model.

*1) Order-Embedding:* Order-embedding is used to embed $X$ into a partially ordered embedding space $(Y, \preceq_Y)$. The choice of the embedding space $Y$ and the order $\preceq_Y$ are application-dependent. The reversed coordinate-wise order on $R_+^N$ [6] is adopted, as can be seen in Fig. 2. The coordinate-wise order is defined by the conjunction of total orders on all coordinates.

Denote the mapping by $f : (X, \preceq_X) \rightarrow (Y, \preceq_Y)$. For all $u, v \in X$ and the corresponding mappings $s, t \in Y$ where $s = f(u)$ and $t = f(v)$,

$$u \preceq v \quad \text{if and only if } \forall i = 1, \ldots, N, s_i \geq t_i, \tag{1}$$

where $i$ indexes the dimension of a vector. On this basis, the penalty for an unordered pair $(s, t)$ in the embedding space $R_+^N$ is defined as

$$E(s, t) = \| \max(0, t - s) \|_2 . \tag{2}$$

This measures the degree to which pair $(s, t)$ violates the coordinate-wise order. It is used in the loss function during training (see below).

*2) Loss Function:* Like most previous works which learn embedding in cross-modal retrieval tasks, the learning target is to make the similarity between the matching pairs higher than non-matching pairs. As in [6], the similarity between a pair $(u, v)$ is defined as:

$$S(u, v) = -E(f(u), f(v)), \tag{3}$$

where $E$ is the order-violation penalty defined in (2), and $f$ is the mapping function. The loss function is the commonly used pairwise ranking loss [17], [18], [25], which encourages $S(u, v)$ for matched pairs to be greater than that for all other pairs by a margin $\alpha$:

$$L(U, V) = \sum_{(u,v)} \Big( \sum_{u'} \max\{0, \alpha - S(u, v) + S(u', v)\} + \sum_{v'} \max\{0, \alpha - S(u, v) + S(u, v')\} \Big). \tag{4}$$

In this equation, $(u, v)$ is a matched pair, $u'$ ranges over $U$ that does not match $v$, and $v'$ ranges over $V$ that does not match $u$. Using this method, three ranking losses are constructed for topic-caption, topic-image and caption-image pairs, respectively. The total loss is defined as follows

$$L = \lambda_1 L(C, I) + \lambda_2 L(T, I) + \lambda_3 L(T, C), \tag{5}$$

where $T$, $C$ and $I$ are the representations of topic, caption and image, respectively, and the hyper-parameters $\lambda_1, \lambda_2, \lambda_3$ control the relevant importance of different partial order relationships.

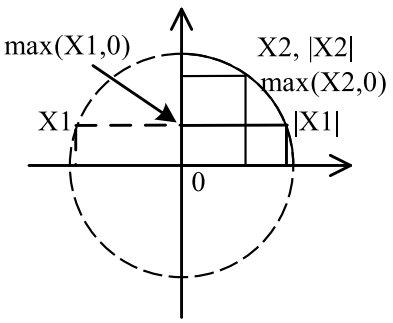



Fig. 3. Illustration of the ReLU and abs functions in 2-dim space. $X1$ and $X2$ are two cases.

*3) Embeddings of Image, Caption and Topic:* The embedding functions $f$ defined for the image and caption generally follow [6]. Specifically, as CNN is used for image representation, the embedding function for the image is formulated as

$$f_I(I) = \max\left(\frac{W_I \cdot CNN(I)}{\| W_I \cdot CNN(I) \|_2}, 0\right), \tag{6}$$

where $W_I$ is a learnable matrix and $CNN(I)$ denotes the image feature which is usually drawn from a pretrained CNN on the ImageNet classification task [26]. For caption embedding, an RNN encoder with a GRU [27] units is used

$$f_C(C) = \max\left(\frac{GRU(C)}{\| GRU(C) \|_2}, 0\right). \tag{7}$$

For topic embedding, we use the $N_T$ dimensional one-hot representation for each topic denoted by $R_T$. The embedding function is

$$f_T(T) = \max\left(\frac{W_T \cdot R_T}{\| W_T \cdot R_T \|_2}, 0\right), \tag{8}$$

where $W_T$ is a learnable matrix.

Note that the embedding functions $f$ above take the ReLU form to make the embedding space in $R_+^N$, that is, $f(x) = \max(x, 0)$. This differs from the function used orginally in [6] which takes the absolute value form, i.e., $\widetilde{f}(x) = |x|$. As in (6) (7) (8), $x$ is forced to have unit L2 norm to make the training process easier and mitigate over-fitting. But it is easy to see that if $\widetilde{f}$ is used and the unit L2 norm is kept, it is impossible for any two points to be ordered in all coordinates. In Fig. 3, take $X_1$, $X_2$ as an example. $|X_1|$ and $|X_2|$ cannot be arranged in any order. If we take the ReLU form, $\max(X_2, 0) \preceq \max(X_1, 0)$. It is possible to get the right solution, which has a big influence on both the order-embedding theory and the model's performance.

### C. Topic-Oriented Image Captioning

Our model is based on the encoder-decoder framework proposed in [1]. It can be divided into two parts: the encoder for image representation and the decoder for caption generation. The CNN features are used to represent the image and

the LSTM network to generate the sentences. The LSTM is defined as follows:

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \tag{9}$$
$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \tag{10}$$
$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \tag{11}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \tag{12}$$
$$h_t = o_t \odot c_t \tag{13}$$
$$p_{t+1} = \text{Softmax}(h_t) \tag{14}$$

where $i_t$, $f_t$ and $o_t$ are the input gate, forget gate and output gate. (12) shows the update formulation of the cell memory $c_t$. $h_t$ stands for the hidden state which is fed into the softmax to produce a probability distribution $p_{t+1}$ over all words as in (14). $W, b$ are the learnable weights and biases. $\sigma$ is the sigmoid function and $\odot$ is the element-wise product.

Denote the input image by $I$ and the ground-truth image caption by $C = (C_0, \ldots, C_N)$ where each $C_t$ denotes a word or a special character. At each timestamp ($t \in 0 \ldots N-1$), the network takes $x_t$ as input and produces $p_{t+1}$ as output. In [1], $x_t$ is defined as:

$$x_{-1} = W_I CNN(I) \tag{15}$$
$$x_t = W_e C_t, \quad t \in 0 \ldots N-1 \tag{16}$$

where $C_t$ is the one-hot representation of the input word, $C_0$ is a special start word and $C_N$ is a special stop word. The image representation is only input once at $t = -1$ as in (15). $W_I$ and $W_e$ are learnable matrices. This model is called *NeuralTalk2* [28].

Our method differs from the above method [1] only in the input to the network at $t = -1$:

$$x_{-1} = \widetilde{W}_I \odot f_I(I) + \widetilde{W}_T \odot f_T(T) \tag{17}$$
$$\widetilde{W}_T = 1 - \widetilde{W}_I \tag{18}$$

where $f_I(I)$ and $f_T(T)$ are the image and topic embeddings defined in (6) and (8). They are weighted by $\widetilde{W}_I$ and $\widetilde{W}_T$ before being summed up. We denote this model as NeuralTalk2-T-oe where T denotes the topic and oe denotes the order-embedding method. As illustrated in Fig. 2, $x_{-1}$ is the embedding point sampled in the subregion bounded by $f_I(I)$ and $f_T(T)$. The weight $\widetilde{W}_I$ is initialized as 0.5 and learnable during training. Ideally, the training captions should reside in the same subregion in the embedding space as $x_{-1}$ (Fig. 2, bottom right). However, due to the imperfect embedding learning, they may reside outside the subregion. It does not alter the fact that the lower bound at each coordinate (though some of which may be violated) set by any topic reduces the size of the output space of the language model compared with the situation without the topic. The smaller output space would enable the language model to learn accurate matching between an image and its captions more easily when they are about the same topic.

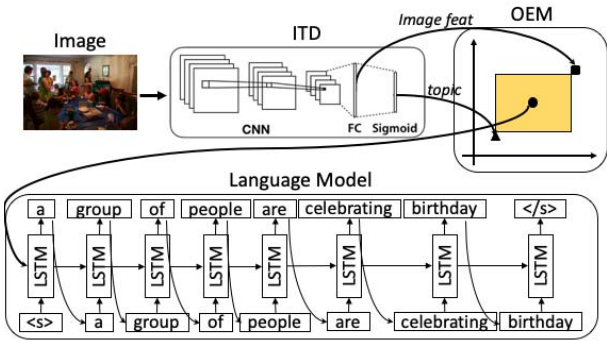



Fig. 4. Illustration of the test procedure. The input image is fed into ITD which produces two types of outputs. The first type is image features from the fully connected layer, and the second type is the topic probabilities from the prediction layer. A topic with the highest probability is selected as the topic input. The top-1 topic and image features are mapped to the order-embedding space. Initial input of the LSTM network is sampled from the region bounded by the image embedding vector and topic embedding vector, and the LSTM generates the caption word by word.

For training, the loss function is the sum of the negative log likelihoods of the words of the ground-truth caption:

$$L((I, T), C) = -\sum_{t=1}^{N} \log p_t(C_t). \tag{19}$$

To show the advantage of the proposed embedding based encoder, we design a control model which is also topic-conditioned but with a simpler encoder. Again, it differs from the model in [1] only in the input to the network at $t = -1$:

$$x_{-1} = W_{IT}[\text{CNN}(I), R_T], \tag{20}$$

where $R_T$ is the one-hot representation of the topic in (8). It simply concatenates the topic representation $R_T$, and the CNN image features CNN($I$) as a single input. $W_{IT}$ is a learnable matrix. All other settings are the same as in the model specified by (17)(18). We denote this model as NeuralTalk2-T where T denotes to the topic.

During testing, we detect the topics of the given image by ITD. The topic to be fed into the language model can be selected from these. Different topics would lead to different captions. For comparison with existing methods which output one caption for one image, we selecte the top-1 topic detected by ITD. The details can be found in the experiment section.

### *D. Summary of the Framework*

Note that two CNNs are needed in this framework, one for image topic detection and the other for extracting features as indicated in (6). For simplicity, we use the same CNN architecture for both tasks.

The entire testing procedure is illustrated in Fig. 4. Given an image, the ITD predicts a topic and extracts image features, which are mapped to the order-embedding space, respectively. A caption embedding vector is sampled from the region bounded by the topic embedding vector and image feature embedding vector according to (17) and (18). This caption embedding vector is input to an LSTM to generate the desired caption.

TABLE I
TOP-5 TOPICS SAMPLED FROM LDA AND NMF ON THE MS-COCO DATASET

| Rank | LDA |
|---|---|
| (1) | meatball wallet hip clap bloody spout demon gorilla ordinary bandage |
| (2) | shoot bullet amble graphic pane smash skylight streetcar nude multitude |
| (3) | flower vase pair pink house run purple shoe attach small |
| (4) | road state high town distance way highway direct country travel |
| (5) | penny list notepad become jail charcoal veil prison girlfriend scoreboard |
| Rank | NMF |
| (1) | tennis court racket player racquet match swing serve female male |
| (2) | skateboard concert cement step guy helmet shirt stair male ride |
| (3) | train station platform passing subway pull wait commute pass load |
| (4) | giraffe zoo enclosure tall feed neck area couple leave together |
| (5) | train track travel railroad engine rail pass come long steam |

From the above process, it is seen that there are three training stages.

1) Topic learning. A topic model (LDA or NMF) is trained to extract topics from the ground-truth captions in the dataset. These topics are then used as topic candidates for images, and another model ITD is trained to predict the most likely topic for a given image.
2) Order-embedding learning. Three mapping functions for image, caption and topic, as defined in (6) (7) and (8), are learned. However, during inference, only the image and topic mapping functions are used.
3) LSTM-based language model learning. It takes an input from the order-embedding space and generates the caption as the final result.

These stages are trained one by one, not in an end-to-end manner.

## IV. EXPERIMENTS

In this section, we first introduce the topic model used in our topic learning process and the topic detector for detecting the topics from raw images. Then we present the details of the embedding learning and its application to the caption-image retrieval task. Finally, the proposed topic-oriented image captioning framework is discussed. Both the caption-image retrieval task and the caption generation task are evaluated on the MS-COCO [7] and Flickr30K [8] datasets.[1]

The MS-COCO and Flickr30K datasets are two popular datasets that are widely used in caption-image retrieval and image caption generation tasks. They contain $123,000$ and $31,000$ images respectively, with five captions annotated for each image. We use the publicly available split which is commonly used in the caption-image retrieval and caption generation tasks [2], [3], [25], [29], where the validation set and testing set each have $5,000$ images for MS-COCO and $1,000$ for Flickr30K. The remaining $113,000$ and $29,000$ images are used for training in MS-COCO and Flickr30K respectively.

### A. Topic Learning

The NLTK toolkit [30] is used to produce the tf or tf-idf feature for documents as described in Section III-A. The implementation of the NMF and LDA model is based on the "scikit-learn" [31] software. Several topics learned by LDA and NMF on MS-COCO [7] are shown in Table I. It can be seen that NMF performs better than LDA. The full topic list of the NMF model can be found in the *Supplementary Materials*. Therefore in subsequent experiments, we only use NMF for topic learning.

For the image topic detection task, the 19-layer VGG-net [32] is used as the backbone network. The $pred$ layer in VGG is replaced with a $N_T$-dim fully connected layer followed by a sigmoid activation layer. During training, every image is rescaled to $256 \times 256$ and a $224 \times 224$ patch is randomly cropped. The network is trained on the training split with an initial learning rate of 0.002 and mini-batch size of 50. The pre-trained weights on ImageNet classification are used as initial weights. The learning rate is decreased by a factor 0.1 for every 5 epochs, and it runs for about 12 epochs in total.

During testing, following the standard procedure [6], [18], [33], we rescale the images to the shortest side length of 256 pixels and take 10 crops in $224 \times 224$ from the four corners, center, and their horizontal reflections. These crops are then fed into the network, and we get the averaged $pred$ as the topic prediction.

### B. Embedding Learning

*1) Embedding Learning:* The embedding model is trained with the hybrid loss defined in (5). The input is a batch of triplets $(t, c, i)$, where $t$ stands for a one-hot topic representation, $c$ stands for a tokenized caption representation, and $i$ stands for an image in raw pixel.

To build the triplets, at first, a batch of caption-image pairs $(c, i)$ is sampled from the training set. For each pair $(c, i)$, we randomly sample a topic $t$ which has confidence higher than 0.3 predicted by the topic model for caption $c$. Since the caption $c$ describes the image content, and the topic $t$ is extracted from $c$, it is reasonable to assume that the image $i$ and the selected topic $t$ are strongly correlated. Then the topic $t$ is added to the pair $(c, i)$ to form a triplet $(t, c, i)$. In a triplet, any of the two items can form an ordered pair.

In the experiment, the batch size is set to 128. As indicated by (4), all of the ordered pairs $(u, v)$ and the contrastive terms $(u, v')$ and $(u', v)$ are drawn from the batch. For the caption-image part, this gives us 128 ordered pairs $(c, i)$,

[1]The source code is available at https://github.com/feiyuhug/TopicOrientedImageCaption

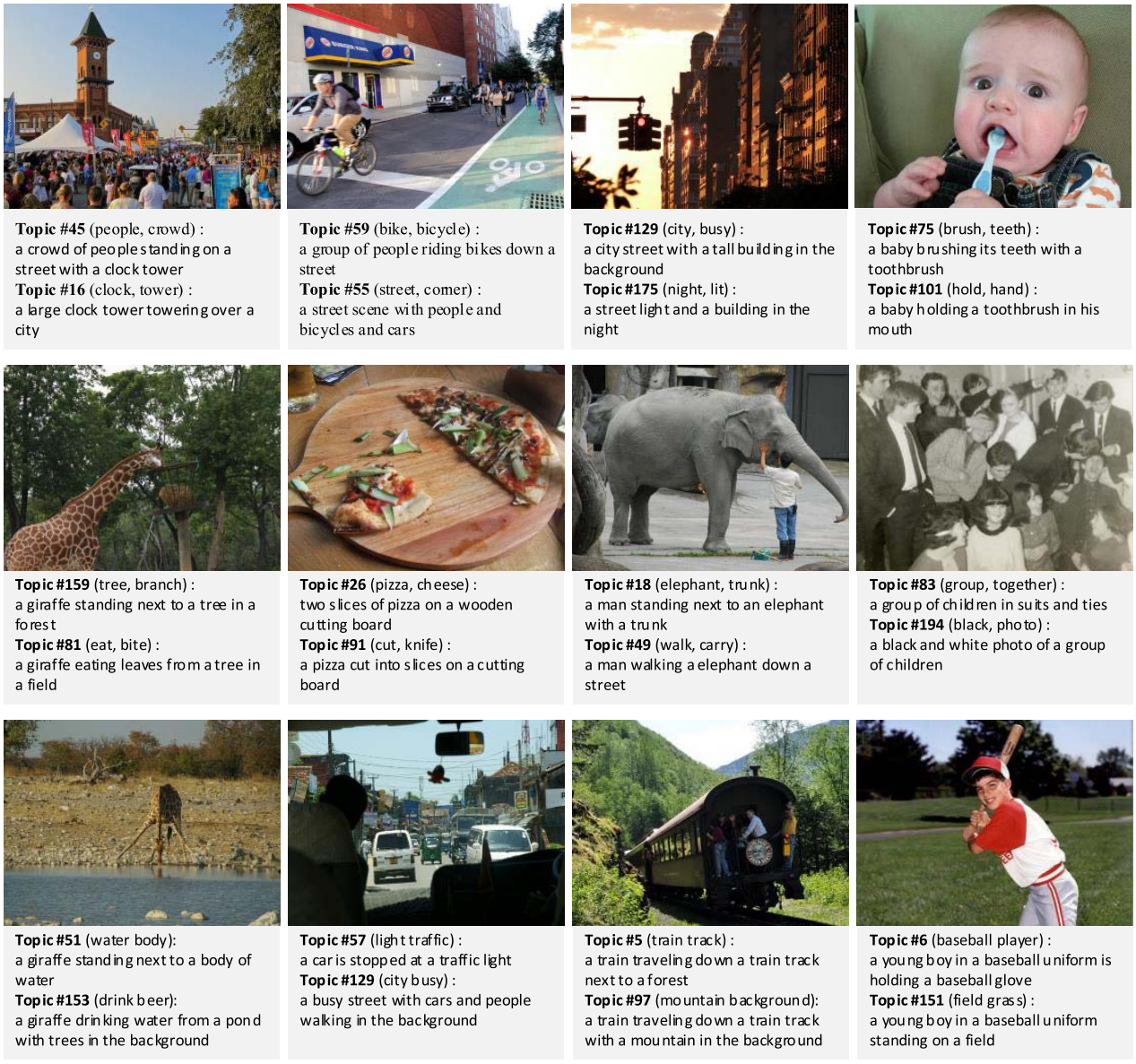



Fig. 5. Sample results of image captioning. Two captions with different topics are shown for each image. The captions are generated by running the model at beam-size=1 given the image and corresponding topic. Top-2 words of each topic are listed in brackets.

127 contrastive terms $(c, i')$ for each caption $c$ where $i'$ ranges over all of the images in the batch except $i$, and 127 contrastive terms $(c', i)$ for each image $i$ in the same way. Therefore, there are $127 \times 128 \times 2 = 32512$ contrastive terms in every single batch.

For the topic-caption part, the contrastive terms $(t', c)$ for each caption $c$ are drawn where $t'$ ranges over all of the topics in the triplets whose confidence predicted by the topic model for caption $c$ is less than 0.1. The contrastive terms $(t, c')$ for each topic $t$ are drawn in the same way. Finally, for the topic-image part, the contrastive terms $(t', i)$ for each image $i$ are drawn where $t'$ ranges over all of the topics in the triplets whose confidence predicted by ITD for the image $i$ is less than 0.1. The contrastive terms $(t, i')$ are drawn in the same way.

We compared the results of two sets of CNN features defined in (6) for image encoding. The first one is the features taken from the $fc7$ layer in VGG19-net pre-trained on ImageNet classification. The second one is the features taken from the $fc7$ layer in ITD which has the same architecture as the VGG19-net except for the output layer. The dimensions of the embedding space and the GRU hidden state in (7) are set to 1024, and the margin $\alpha$ is set to 0.05. $\lambda_1$, $\lambda_2$ and $\lambda_3$ in (5) are set to 1.0, 0.2 and 0.2 respectively. The model is trained with the full hybrid loss for about 50 epochs using the Adam optimizer with learning rate 0.001.

*2) The Caption-Image Retrieval Task:* As the embedding learning of images, captions and topics is a basic component of the proposed topic-oriented image captioning framework, we need to know the quality of the learned embeddings. This is evaluated on the bidirectional caption-image retrieval task. All of the images and captions are embedded first, then the similarity of a caption-image pair is evaluated by their distance in the embedding space. In order-embedding, the distance is

TABLE II

RESULTS OF CAPTION-IMAGE RETRIEVAL ON MS-COCO. (* INDICATES ENSEMBLE MODEL, ♯ INDICATES EXTERNAL ANNOTATION IS USED.)

| Model | Caption Retrieval | | | | Image Retrieval | | | |
|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | Med r | R@1 | R@5 | R@10 | Med r |
| STD†* [33] | 33.8 | 67.7 | 82.1 | 3 | 25.9 | 60.0 | 74.6 | 4 |
| m-RNN [34] | 41.0 | 73.0 | 83.5 | 2 | 29.0 | 42.2 | 77.0 | 3 |
| FV†* [32] | 39.4 | 67.9 | 80.9 | 2 | 25.1 | 59.8 | 76.6 | 4 |
| DVSA [25] | 38.4 | 69.9 | 80.5 | 1 | 27.4 | 60.2 | 74.8 | 3 |
| MNLM [18] | 43.4 | 75.7 | 85.8 | 2 | 31.0 | 66.7 | 79.9 | 3 |
| m-CNN* [35] | 42.8 | 73.1 | 84.1 | 2 | 32.6 | 68.6 | 82.8 | 3 |
| RNN+FV† [36] | 40.8 | 71.9 | 83.2 | 2 | 29.6 | 64.8 | 80.5 | 3 |
| DSPE+FV† [28] | 50.1 | 79.7 | 89.2 | - | 39.6 | 75.2 | 86.9 | - |
| OEM-abs [6] | 46.7 | - | 88.9 | 2 | 37.9 | - | 85.9 | **2** |
| OEM | 47.0 | 79.6 | 89.5 | 2 | 38.8 | 74.6 | 86.5 | **2** |
| OEM-FT100 | 50.4 | 82 | 90.4 | 1.4 | 41.5 | 77.1 | 88.2 | **2** |
| OEM-FT100-T | 51.3 | 83.1 | 91.7 | 1.2 | 42.7 | 78.2 | **88.9** | **2** |
| OEM-FT200 | 52.0 | 83.2 | 91.7 | **1** | 42.5 | 77.9 | 88.6 | **2** |
| OEM-FT200-T | **52.2** | **83.6** | **91.8** | **1** | **43.5** | **78.4** | **88.9** | **2** |

TABLE III

RESULTS OF CAPTION-IMAGE RETRIEVAL ON FLICKR30K. (* INDICATES ENSEMBLE MODEL, ♯ INDICATES EXTERNAL ANNOTATION IS USED.)

| Model | Caption Retrieval | | | | Image Retrieval | | | |
|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | Med r | R@1 | R@5 | R@10 | Med r |
| MNLM [18] | 23.0 | 50.7 | 62.9 | 5 | 16.8 | 42.0 | 56.5 | 8 |
| m-RNN [34] | 35.4 | 63.8 | 73.7 | 3 | 22.8 | 50.7 | 63.1 | 5 |
| FV†* [32] | 35.0 | 62.0 | 73.8 | 3 | 25.0 | 52.7 | 66.0 | 5 |
| m-CNN* [35] | 33.6 | 64.1 | 74.9 | 3 | 26.2 | 56.3 | 69.6 | 4 |
| RNN+FV† [36] | 34.7 | 62.7 | 72.6 | 3 | 26.2 | 55.1 | 69.2 | 4 |
| DSPE+FV† [28] | **40.3** | 68.9 | 79.9 | - | 29.7 | 60.1 | 72.1 | - |
| OEM-abs [6] | 34.7 | 64.3 | 77.2 | 3 | 26.9 | 56.9 | 68.3 | 4 |
| OEM | 34.4 | 67.3 | 78.3 | **2** | 27.9 | 57.8 | 69.7 | 4 |
| OEM-FT100 | 39.5 | 67.5 | 79.6 | **2** | 29.4 | 59.6 | 71.2 | 4 |
| OEM-FT100-T | 40.1 | **70.8** | **81.4** | **2** | **32.5** | **61.4** | **72.5** | **3** |

defined in (2). For a caption-image pair $(c, i)$, the distance between them is defined as $E(c, i) = \| \max(0, c-i)\|_2$, where $c$ and $i$ denote the embedding of the caption and image respectively. The bidirectional caption-image retrieval task involves two symmetrical ranking problems: ranking a set of images by similarity for a query caption (Image Retrieval) or the other way around, ranking a set of captions for a query image (Caption Retrieval).

The reason for choosing this task for evaluating the embedding method is as follows. From the bottom right of Fig. 2, it can be seen that images and captions about the same topic will be pushed closer in the embedding space by the order-embedding method, which makes sense because such images and captions ought to be related. It makes bidirectional retrieval effective, though topics are only used in training but not in testing.

The experimental results of our method on MS-COCO and Flickr30K datasets are listed in Table II and Table III respectively, together with existing results in the literature. We use the standard ranking metrics for evaluation. "R@K" means the recall rate of top-K, and "Med r" means the median rank of the first ground-truth result. To demonstrate the contributions of different components in our model, we present the models with various settings. OE is our baseline, which uses the embedding definition in (6)(7) and original VGG19 features (the $fc7$ layer). OE-abs [6] also uses the original VGG19 features; the difference is that it uses the absolute value form in its embedding definition compared with OE. OE-FT$N_T$ differs from OE in that it uses the image representation of ITD (the $fc7$ layer). $N_T$ is the topic number in the image topic detection task as defined in Section III-A. All of the models mentioned above only learn the embeddings of image and caption using the loss function in (5) with only the first term. OE-FT$N_T$-T is the full model which learns the topic embeddings jointly with the captions and images using the loss function in (5) with all of the three terms. But we find that training the full model for about 50 epochs first, then training it by dropping the last two terms for another 10 epochs works better. In other words, the last two terms in (5) help us to find a good initial solution in this setting. This model outperforms the existing models by a significant margin. This result is remarkable since unlike many existing models [29], [33], [34], [36], [37], no model ensemble or external annotations are used.

We provide three techniques to improve the performance of OE-abs: (1) changing the absolute value function to ReLU function, (2) fine-tuning the CNN for image embedding in the image topic detection task, (3) adding two topic-related terms in the loss function. From Table II, it is seen that the second technique makes the most significant contribution. It is worth mentioning that it is the introduction of the topic that leads to the multi-label task and motivates the second technique above.

### C. Image Captioning

We generate the embeddings of images as defined in (6). At the training stage, the image is resized into $256 \times 256$; then

TABLE IV

THE RESULTS OF DIFFERENT IMAGE-CAPTIONING METHODS WITH VGG-NET OR GOOGLENET AS THE IMAGE ENCODER ON THE MS-COCO DATASET. BOLDFACE INDICATES THE BEST AND UNDERSCORE INDICATES THE SECOND BEST

| Model | Encoder | Bleu-1 | Bleu-2 | Bleu-3 | Bleu-4 | METEOR | Rouge-L | CIDEr |
|---|---|---|---|---|---|---|---|---|
| Google NIC [1] | GoogleNet | 0.666 | 0.461 | 0.329 | 0.246 | - | - | - |
| LRCN [38] | VGG-net | 0.628 | 0.442 | 0.304 | 0.210 | - | - | - |
| Spatial ATT [2] | VGG-net | 0.718 | 0.504 | 0.357 | 0.250 | 0.230 | - | - |
| Semantic ATT [3] | GoogleNet | 0.709 | 0.537 | 0.402 | 0.304 | 0.243 | - | - |
| DM [13] | VGG-net | 0.713 | 0.539 | 0.403 | 0.304 | 0.251 | 0.525 | 0.937 |
| GLA [11] | VGG-net | 0.725 | 0.556 | 0.417 | 0.312 | 0.249 | 0.533 | 0.964 |
| SC [10] | VGG-net | 0.716 | 0.545 | 0.405 | 0.301 | 0.247 | - | 0.970 |
| ATT-kCC [39] | VGG-net | 0.735 | 0.566 | 0.424 | **0.316** | 0.251 | - | 0.982 |
| NeuralTalk2(stage1) | VGG-net | 0.710 | 0.502 | 0.348 | 0.270 | 0.212 | 0.402 | 0.821 |
| NeuralTalk2(stage2) | VGG-net | 0.713 | 0.525 | 0.369 | 0.298 | 0.243 | 0.522 | 0.909 |
| NeuralTalk2-T(stage1) | VGG-net | 0.713 | 0.512 | 0.346 | 0.283 | 0.232 | 0.473 | 0.877 |
| NeuralTalk2-T(stage2) | VGG-net | 0.722 | 0.523 | 0.365 | 0.302 | 0.248 | 0.528 | 0.917 |
| NeuralTalk2-T-oe(stage1) | VGG-net | 0.720 | 0.512 | 0.355 | 0.291 | 0.235 | 0.497 | 0.898 |
| NeuralTalk2-T-oe(stage2) | VGG-net | 0.737 | 0.560 | **0.434** | 0.314 | 0.253 | 0.533 | 0.980 |
| NeuralTalk2-T-oe(stage3) | VGG-net | **0.740** | **0.567** | 0.433 | 0.313 | **0.255** | **0.534** | **0.983** |

TABLE V

THE RESULTS OF DIFFERENT IMAGE-CAPTIONING METHODS WITH RESNET AS THE IMAGE ENCODER ON THE MS-COCO DATASET. ♯ INDICATES EXTERNAL TRAINING DATA IS USED. BOLDFACE INDICATES THE BEST AND UNDERSCORE INDICATES THE SECOND BEST

| Model | Encoder | Bleu-1 | Bleu-2 | Bleu-3 | Bleu-4 | METEOR | Rouge-L | CIDEr |
|---|---|---|---|---|---|---|---|---|
| SCA-CNN [40] | ResNet-152 | 0.719 | 0.548 | 0.411 | 0.311 | 0.250 | - | - |
| SCST [15] | ResNet-101 | - | - | - | 0.319 | 0.255 | 0.543 | 1.063 |
| SCN-LSTM [16] | ResNet-152 | 0.728 | 0.566 | 0.433 | 0.330 | 0.257 | - | 1.012 |
| Stack-Cap (C2F) [41] | ResNet-101 | **0.786** | **0.625** | **0.479** | **0.361** | **0.274** | **0.569** | **1.204** |
| Adaptive [42] | ResNet-152 | 0.742 | 0.580 | 0.439 | 0.332 | 0.266 | | 1.085 |
| NeuralTalk2-T-oe(stage3) | ResNet-101 | 0.739 | 0.572 | 0.432 | 0.326 | 0.261 | 0.544 | 1.038 |

TABLE VI

RESULTS OF DIFFERENT IMAGE-CAPTIONING METHODS ON THE FLICKR30K DATASET. BOLDFACE INDICATES THE BEST AND UNDERSCORE INDICATES THE SECOND BEST

| Model | Encoder | Bleu-1 | Bleu-2 | Bleu-3 | Bleu-4 | METEOR | Rouge-L | CIDEr |
|---|---|---|---|---|---|---|---|---|
| Deep VS [25] | VGG-net | 0.573 | 0.369 | 0.240 | 0.157 | - | - | - |
| Google NIC [1] | GoogleNet | 0.663 | 0.423 | 0.277 | 0.183 | - | - | - |
| m-RNN [35] | VGG-net | 0.600 | 0.410 | 0.280 | 0.190 | - | - | - |
| Soft-Attention [2] | VGG-net | 0.667 | 0.434 | 0.288 | 0.191 | 0.185 | - | - |
| Hard-Attention [2] | VGG-net | **0.669** | 0.439 | 0.296 | 0.199 | 0.185 | - | - |
| ATT [3] | GoogleNet | 0.647 | **0.460** | **0.324** | **0.230** | 0.189 | - | - |
| SCA-CNN [40] | VGG-net | 0.646 | 0.453 | 0.317 | 0.218 | 0.188 | - | - |
| NeuralTalk2(stage1) | VGG-net | 0.597 | 0.430 | 0.285 | 0.187 | 0.155 | 0.410 | 0.322 |
| NeuralTalk2(stage2) | VGG-net | 0.613 | 0.437 | 0.291 | 0.192 | 0.178 | 0.418 | 0.376 |
| NeuralTalk2-T(stage1) | VGG-net | 0.604 | 0.427 | 0.283 | 0.190 | 0.158 | 0.417 | 0.327 |
| NeuralTalk2-T(stage2) | VGG-net | 0.622 | 0.439 | 0.297 | 0.203 | 0.182 | 0.423 | 0.386 |
| NeuralTalk2-T-oe(stage1) | VGG-net | 0.618 | 0.428 | 0.291 | 0.195 | 0.177 | 0.427 | 0.365 |
| NeuralTalk2-T-oe(stage2) | VGG-net | 0.631 | 0.440 | 0.302 | 0.211 | 0.186 | 0.431 | **0.399** |
| NeuralTalk2-T-oe(stage3) | VGG-net | 0.646 | 0.438 | 0.319 | 0.224 | **0.192** | **0.438** | 0.396 |

it is cropped to $224 \times 224$ and fed to VGG-net. The VGG-net that ends up with $fc7$ is followed by a fully-connected layer initialized with $W_I$, an L2 norm layer and finally a ReLU layer as in (6). As the input of the framework is an image-topic pair $(I, T)$, for each image $I$ in a training batch, the topic $T$ is randomly selected from the top-5 topics detected by ITD. The topic embeddings are taken from the learned embedding space and fixed during training. The target caption for this pair is drawn from the five ground-truth captions that have $T$ in their top-5 topics.

We follow [25] to do the caption preprocessing, including building the dictionary, removing the low-frequency words and truncating all the captions longer than 16 tokens. The dimensions of the hidden state in LSTM and word embeddings are set to 512. We use the Adam algorithm to optimize the model, the whole process follows a curriculum learning and is divided into three stages. At the first stage, only the weights in LSTM are learned. After that, the weights of the image encoder $f_I(I)$ are freed. Finally, $\widetilde{W}_I$ is also freed. The baseline models NeuralTalk2 and NeuralTalk2-T have two stages. At the first stage, only the weights in LSTM are learned. At the second stage, all weights are learned jointly including the weights in CNN($I$) and $W_{IT}$ in (20).

*1) Qualitative Evaluation:* Some qualitative results from the MS-COCO dataset are presented in Fig. 5. It is seen that the generated captions are closely related to the given topics. For the first image, given topic #45 about *people* and *crowd*, the generated sentence is "a crowd of people standing

TABLE VII

RECALL RATIO OF THE TOPIC'S TOP WORDS IN THE GENERATED CAPTIONS

| Word-rank | 1 | 2 | 3 |
|---|---|---|---|
| Recall | 96% | 90% | 84% |

TABLE VIII

CAPTIONS GENERATED BY DIFFERENT GIVEN TOPICS (INCLUDING SOME UNRELATED TOPICS) FOR THE FIRST IMAGE IN FIG. 5

| Topic | Caption |
|---|---|
| #45 (people, crowd) | a crowd of people standing on a street with a clock tower |
| #16 (clock, tower) | a large clock tower towering over a city |
| #1 (tennis, court) | people standing on the court |
| #2 (skateboard, concrete) | a crowd of people standing |
| #3 (train, station) | a crowd of people standing in a station |
| #4 (giraffe, zoo) | a crowd of people standing |
| #5 (train track) | a crowd of people |

on a street with a clock tower" where *people* and *crowd* are presented. If the topic is #16 which is about *clock* and *tower*, the sentence is "a large clock tower towering over a city" where only *clock* and *tower* are presented and *people* is not presented.

Every topic has a list of representative words with decreasing ranks. We look into the probability that the words with high ranks in the topic can be found in the generated caption. A high probability would indicates a great influence of the topic on the generated caption. We compute the recall ratio of each topic's top-3 words in the generated caption. This is carried out on the $5,000$ images in the test set. The results of the words in different ranks are listed in Table VII. For example, the first column is the recall ratio of the first words in the topics of all $5,000$ images. As can be seen, all of the recall ratios are pretty high.

The proposed model requires the given topic to be related to the image. An interesting question is, if the given topic were unrelated to the image, what would happen? We have tested this scenario and find that the model is not very robust. Taking the first image in Fig. 5 as an example, topic #45 and #16 are the right topics and topic #1, #2, #3, #4, #5 are the wrong topics totally unrelated to the image (Table VIII). The generated captions are listed in Table VIII. It can be seen that with wrong topics, the generated captions are generally meaningful but misled by the topics to some degree. Therefore, to generate good captions in an automatic system, it is recommended to select the topics predicted by ITD. Such topics are appropriate as proved by quantitative evaluations of the generated captions (see below). In human-computer interaction applications, it is easy for users to select appropriate topics.

*2) Quantitative Evaluation:* During testing, we set the beam size to 3. The results for MS-COCO are listed in Table IV. The image encoders of all methods listed here are either VGG-net or GoogleNet which are popular in this field. The baseline model upon which our model is built is taken from publicly available code *NeuralTalk2*.[2] The performances of two baseline models NeuralTalk2 and NeuralTalk2-T and our full model NeuralTalk2-T-oe are listed. NeuralTalk2-T-oe achieves a significant improvement over the baseline NeuralTalk2 and outperforms the previous methods except ATT-kCC [39] whose BLEU-4 is higher. NeuralTalk2-T is slightly better than NeuralTalk2, which verifies the efficiency of topic-condition in image-captioning. The full model NeuralTalk2-T-oe outperforms NeuralTalk2-T by a large margin, which verifies the merit of the proposed embedding method.

Much recent work uses ResNet as the encoder. To compare with those models, we replace the VGG-net with ResNet-101 in the proposed framework. The improved results are listed in Table V. It is seen that our results are comparable to the results of some recent methods [15], [16], [40] but inferior to the results of the others [41], [42]. Note that our image caption model is a simple standard encoder-decoder model, while most existing models use the attention mechanism. Though with a simpler architecture, our model is competitive with those models.

The experiments with the same setting are also conducted on Flickr30K. The results are listed in Table VI. In the ablation study, NeuralTalk2-T-oe has the best performance and surpasses the baseline by a large margin. Compared with existing models that use VGG-net as the encoder, the proposed model ranks first on most of the metrics. ATT [3] performs better than our model as its encoder GoogleNet is stronger than VGG-net, and it adopts the attention mechanism.

## V. CONCLUSION

This paper introduces a new image captioning problem: describing an image under a given topic. To solve this problem, a cross-modal embedding of image, caption, and topic is learned. The proposed method has achieved competitive results with the state-of-the-art methods on both the caption-image retrieval task and the caption generation task on the MS-COCO and Flickr30K datasets. This new framework provides users with controllability in generating intended captions for images, which may inspire exciting applications.

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

## REFERENCES

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3156–3164.

[2] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[3] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4651–4659.

[4] A. Tariq and H. Foroosh, "A context-driven extractive framework for generating realistic image descriptions," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 619–632, Feb. 2017.

[5] Y. Dong, H. Su, J. Zhu, and B. Zhang, "Improving interpretability of deep neural networks with semantic information," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 975–983.

[6] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, "Order-embeddings of images and language," in *Proc. ICLR*, 2016, pp. 1–12.

[7] X. Chen *et al.* (2015). "Microsoft COCO captions: Data collection and evaluation server." [Online]. Available: https://arxiv.org/abs/1504.00325

[8] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, Feb. 2014.

[9] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov, "Review networks for caption generation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2361–2369.

[10] L. Zhou, C. Xu, P. Koch, and J. J. Corso. (2016). "Watch what you just said: Image captioning with text-conditional attention." [Online]. Available: https://arxiv.org/abs/1606.04621

[11] L. Li, S. Tang, L. Deng, Y. Zhang, and Q. Tian, "Image caption with global-local attention," in *Proc. AAAI*, 2017, pp. 4133–4139.

[12] H. Fang *et al.*, "From captions to visual concepts and back," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1473–1482.

[13] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1151–1159.

[14] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of SPIDEr," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 873–881.

[15] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1179–1195.

[16] Z. Gan *et al.*, "Semantic compositional networks for visual captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1141–1150.

[17] R. Socher, Q. V. L. A. Karpathy, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Trans. Assoc. Comput. Linguistics*, vol. 2, no. 1, pp. 207–218, 2014.

[18] R. Kiros, R. Salakhutdinov, and R. S. Zemel. (2014). "Unifying visual-semantic embeddings with multimodal neural language models." [Online]. Available: https://arxiv.org/abs/1411.2539

[19] F. Wu *et al.*, "Cross-modal learning to rank via latent joint representation," *IEEE Trans. Image Process.*, vol. 24, no. 5, pp. 1497–1509, May 2015.

[20] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. A. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, p. 391, Sep. 1990.

[21] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, nos. 1–2, pp. 177–196, Jan. 2001.

[22] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. 1, pp. 993–1022, Jan. 2003.

[23] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[24] P. Willett, "The Porter stemming algorithm: Then and now," *Program*, vol. 40, no. 3, pp. 219–223, 2006.

[25] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3128–3137.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[27] K. Cho *et al.*, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1724–1734.

[28] A. Karpathy and J. Johnson. (2015). *NeuralTalk2*. [Online]. Available: https://github.com/karpathy/neuraltalk2

[29] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5005–5013.

[30] S. Bird, E. Klein, and E. Loper, *Natural Language Processing With Python: Analyzing Text With the Natural Language Toolkit*. Newton, MA, USA: O'Reilly Media, Inc., 2009.

[31] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.

[33] B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Associating neural word embeddings with deep image representations using Fisher vectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4437–4446.

[34] R. Kiros *et al.*, "Skip-thought vectors," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 3294–3302.

[35] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

[36] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2623–2631.

[37] G. Lev, G. Sadeh, B. Klein, and L. Wolf, "RNN Fisher vectors for action recognition and image annotation," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 833–850.

[38] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2625–2634.

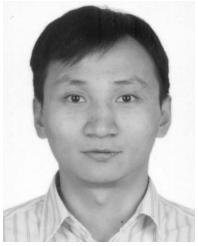
[39] J. Mun, M. Cho, and B. Han, "Text-guided attention model for image captioning," in *Proc. AAAI*, 2017, pp. 4233–4239.

[40] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6298–6306.

[41] J. Gu, J. Cai, G. Wang, and T. Chen, "Stack-captioning: Coarse-to-fine learning for image captioning," in *Proc. AAAI*, 2018, pp. 1–8.

[42] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3242–3250.

**Niange Yu** (S'15) received the B.E. degree in software engineering from Beihang University, Beijing, China, in 2015. He is currently pursuing the M.E. degree with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His current research interests include computer vision and deep learning.



**Xiaolin Hu** (S'01–M'08–SM'13) received the B.E. and M.E. degrees in automotive engineering from the Wuhan University of Technology, Wuhan, China, in 2001 and 2004, respectively, and the Ph.D. degree in automation and computer-aided engineering from The Chinese University of Hong Kong, Hong Kong in 2007. He is currently an Associate Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His current research interests include artificial neural networks, computer vision, and computational neuroscience. He is currently an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

**Binheng Song** received the B.E., M.E., and Ph.D. degrees from the Department of Applied Mathematics, Tsinghua University, Beijing, China, in 1986, 1988, and 1991, respectively. He is currently an Associate Professor with the Graduate School at Shenzhen, Tsinghua University.

**Jian Yang** received the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology (NUST) in 2002. He was a Post-Doctoral Researcher with the University of Zaragoza in 2003. From 2004 to 2006, he was a Post-Doctoral Fellow with the Biometrics Centre, Hong Kong Polytechnic University. From 2006 to 2007, he was a Post-Doctoral Fellow with the Department of Computer Science, New Jersey Institute of Technology. He is currently the Chang-Jiang Professor with the School of Computer Science and Technology, NUST. He has authored over 100 scientific papers in pattern recognition and computer vision. His journal papers have been cited over 4000 times in the ISI Web of Science and 9000 times in the Web of Google Scholar. His research interests include pattern recognition, computer vision, and machine learning. He is a Fellow of the IAPR. He is/was currently an Associate Editor of *Pattern Recognition Letters*, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and *Neurocomputing*.

**Jianwei Zhang** received the B.E. degree (Hons.) and the M.E. degree from the Department of Computer Science, Tsinghua University, Beijing, China, the Ph.D. degree from the Institute of Real-Time Computer Systems and Robotics, Department of Computer Science, University of Karlsruhe, Germany, in 1994, and the Habilitation degree from the Faculty of Technology, University of Bielefeld, Germany, in 2000. He is currently a Professor and the Head of the Department of Informatics, TAMS, University of Hamburg, Germany. In his research areas, he has published about 300 journal and conference papers (with four Best Paper Awards), technical reports, four book chapters, and five research monographs. His research interests are cognitive robotics, sensor fusion, dexterous manipulation and multimodal robot learning, and Industry 4.0. He has been coordinating numerous collaborative research projects of EU and German Research Council, including the Transregio-SFB TRR169 Crossmodal Learning. He is also a Life-Long Academician of Academy of Sciences in Hamburg. He is the General Chair of the IEEE MFI 2012 and the IEEE/RSJ IROS 2015.