Self-Guiding Multimodal LSTM - when we do not have a perfect training dataset for image captioning

Yang Xian, *Member, IEEE,* and Yingli Tian, *Fellow, IEEE*

Abstract—In this paper, a self-guiding multimodal LSTM (sgLSTM) image captioning model is proposed to handle uncontrolled imbalanced real-world image-sentence dataset. We collect FlickrNYC dataset from Flickr as our testbed with $306,165$ images and the original text descriptions uploaded by the users are utilized as the ground truth for training. Descriptions in FlickrNYC dataset vary dramatically ranging from short term-descriptions to long paragraph-descriptions and can describe any visual aspects, or even refer to objects that are not depicted. To deal with the imbalanced and noisy situation and to fully explore the dataset itself, we propose a novel guiding textual feature extracted utilizing a multimodal LSTM (mLSTM) model. Training of mLSTM is based on the portion of data in which the image content and the corresponding descriptions are strongly bonded. Afterwards, during the training of sgLSTM on the rest training data, this guiding information serves as additional input to the network along with the image representations and the ground-truth descriptions. By integrating these input components into a multimodal block, we aim to form a training scheme with the textual information tightly coupled with the image content. The experimental results demonstrate that the proposed sgLSTM model outperforms the traditional state-of-the-art multimodal RNN captioning framework in successfully describing the key components of the input images.

Index Terms—image captioning, self-guiding, real-world dataset, multimodal, recurrent neural network.

Fig. 1. Example of the description generated by the proposed sgLSTM image captioning framework compared with the result generated by the traditional multimodal RNN. Both frameworks are trained with FlickrNYC - a new dataset proposed in this paper.

I. INTRODUCTION

In the recent popularized language-vision community, image captioning has been an important task. It involves generating a textual description that describes an image by analyzing its visual content. Automatic image captioning is able to assist solving computer vision challenges including image retrieval, image understanding, object recognition, navigation for the blind, and many others.

Although image captioning is a natural task for human beings, it remains challenging from a computer vision point of view especially due to the fact that the task itself is ambiguous. There are countless ways to describe one input image, from high-level descriptions to explanations in details, while all are semantically correct. The fundamental cause is that in principle, descriptions of an image can talk about any visual aspects in it varying from object attributes to scene features, or even refer to objects that are not depicted and the hidden interaction or connection that requires common sense knowledge to analyze [1].

In general, image captioning is a data-driven task. Descriptions for query images are normally defined by the training data. Therefore, it is not uncommon to see the birth of a new dataset for a new task. Recently, Amazon Mechanical Turk (AMT) is involved in more and more dataset description generation process. Different sets of descriptions may be generated depending on the instructions provided to fit a specific captioning task. Since it is an expensive process, majority of the image captioning frameworks focus on exploring existing datasets which tend to provide a sentence description embedded with the objects, attributes, and the reactions with the scene in the image. Some other frameworks tackle the problem from a different angle, such as unambiguous descriptions [2], image stream descriptions [3], etc.

In this paper, we work with FlickrNYC - an image-sentence dataset collected directly from Flickr. The original descriptions provided by the users are utilized as the training data. Flickr data has been widely used in the dataset collection [4], [5], [6] due to its availability of billions of images. However, the descriptions provided by the users are rarely used for captioning purpose directly due to several characteristics of the Flickr

Y. Xian was with the Department of Computer Science, The Graduate Center, The City University of New York, New York, NY, USA e-mail: yxian@gradcenter.cuny.edu.

Y. Tian is with the Department of Computer Science, The Graduate Center and the Department of Electrical Engineering, The City College, The City University of New York, New York, NY, USA e-mail: ytian@ccny.cuny.edu

text data: 1) Lengths of the descriptions vary dramatically for each image. While some users talk in paragraphs about the details including the possible background that is not directly related to the image, others may just describe in a few words indicating the location or the date information. 2) Users may input descriptions for an album instead of a photo. Therefore, we may have multiple images that look visually different with the same description. 3) Unlike the labeling process performed by AMT workers, the content of the descriptions is not strictly controlled semantically or syntactically. Foreign languages exist along with personal information including copyright statement, camera information, and links to personal social media accounts. Existing natural language processing (NLP) tools provide a limited solution in preparing the training data. In the end, it becomes tedious to set filtering criteria or use regular expressions to generate the 'perfect' training dataset.

However, despite all the characteristics listed, Flickr data meets the criterion for image captioning task - it comes from millions of users who can describe anything related to the images they upload. And more importantly, it is a real-world uncontrolled valuable resource. In this paper, we use 'new york city' as our test case, i.e., 'new york city' is employed as the keyword for the query process, to build FlickrNYC dataset. As observed in FlickrNYC, descriptions in shorter lengths are more strongly correlated to the image content and are mainly related to the locations, events, or activities. They occur more repetitively compared with longer descriptions, e.g., a user may uploaded several images related to a walk in central park and they all have the description as 'central park'. On the other hand, descriptions in longer sentences or paragraphs reveal more syntactical details, but may provide concepts that are more implicitly related to the images and have a weaker or no correlation to the image content. Examples in FlickrNYC can be found in Fig. 2.

In the proposed framework, a self-guiding multimodal long short-term memory (sgLSTM) framework is presented to leverage between two portions of the data: $data_s$ (images with shorter length of descriptions) and $data_l$ (images with longer length of descriptions). We aim to make use of the part of the dataset with more reliable information to guide the training process of the caption generation. As demonstrated in Fig. 1, a direct training utilizing the state-of-the-art multimodal RNN captioning method fails to capture the core event revealed in the image due to the fact that, FlickrNYC is a noisy real-world dataset in which we have multiple images labeled as 'thanksgiving' but are visually different. Moreover, thanksgiving celebration is more frequently seen than Chinese New Year celebration within the training dataset. However, the proposed framework manages to generate accurate description that is both semantically and syntactically correct.

Contributions of the proposed framework are threefold:

- A novel image captioning framework is proposed to deal with the uncontrolled image-sentence dataset where descriptions could be strongly or weakly correlated to the image content and in arbitrary lengths. The self-guiding process looks into the learning process in a global way to balance the syntactic correctness and the semantic details

revealed in the images. This scheme can be extended to handle other tasks when we have imperfect training data.
- A new FlickrNYC dataset is introduced with $306,165$ images related to 'new york city'. Light pre-processing combining basic NLP tools and regular expression filtering are performed to remove the personal information including copyright, camera info., URLs, social network accounts, etc. Different from the majority captioning datasets, descriptions in FlickrNYC come from the original Flickr users.
- Experimental results demonstrate that the proposed guiding textual feature manages to provide additional text information which strongly correlates to the image content. Compared with the existing traditional multimodal RNN captioning framework, the self-guiding scheme is able to recover more accurate descriptions given an input image.

The rest of the paper is organized as follows: Related work is discussed in Sec. II. Sec. III provides a detailed description of the proposed image captioning framework based on self-guiding multimodal LSTM. The collected dataset including the experimental results are presented and discussed in Sec. IV. Conclusions and future work are listed in Sec. V.

II. RELATED WORK

A. Image Captioning

Based on the underlying models utilized, recent image captioning frameworks can be classified into three categories. The first group of approaches casts the problem as a retrieval problem in which description of a test image is generated by searching for similar images in a database. This group of models employs the visual space to measure the similarity during image search. Descriptions of these similar images are transferred to obtain the target description. [7] utilized the activations of the last layer of the Visual Geometry Group convolutional neural network (VGG-CNN) [8] trained on ImageNet [9] to represent the image features. The description of the query image is represented as a weighted average of the distributed representations of the retrieved descriptions. Different from [7], [10] employed the n-gram overlap F-score between the descriptions to measure the description similarity. Other than the deep learning based approaches, traditional machine learning techniques are also utilized in this task [11].

The second group of methods adopts pre-defined sentence templates to generate image descriptions. The missing components in the sentence structures are filled based on image understanding of the objects, attributes and the correlations between objects and the scene. [12] proposed a sentence generation model which parses a query image into a visual dependency representation (VDR) which then traversed to fill the missing slots in the templates. More linguistically sophisticated approaches [13], [14], [15] were proposed to tackle the sentence generation.

The third group of approaches integrates image understanding and natural language generation into a unified pipeline. In general, image content in terms of objects, actions, and attributes is represented based on a set of visual features. Later, this content information is utilized to drive a language

Fig. 2. Sample images and the corresponding descriptions in FlickrNYC dataset. (a) Examples from $data_s$ in which images are with short descriptions. (b) Examples from $data_l$ in which images are with long descriptions.

generation system, e.g., a recurrent neural network (RNN), to output the image descriptions[16], [17]. Some frameworks model image and text jointly into a multimodal space where later the joint representation space is used to perform cross-modal retrieval based on a query image. [18] presented an alignment model which uses a structured object to align the two modalities (i.e., CNN over image regions and bidirectional RNN over sentences) through a multimodal embedding. An encoder-decoder framework is presented by [19] utilizing a joint multimodal space in which the LSTM is a big success. Another represented work in this category is the mRNN model [20] in which a multimodal component is introduced to explicitly connect the language model and the vision model by a one-layer representation.

Within the encoder-decoder design, semantic concepts are employed in many frameworks. Visual detectors were used to generate conditional inputs to a maximum-entropy language model in [21]. Jia *et al.* [22] modified the LSTM model to add semantic information extracted from the image as extra input to each LSTM block. In [16], an algorithm was introduced which learns to selectively attend to semantic concept proposals and fuse them into hidden states and outputs of RNN. Liu *et al.* [23] proposed to use a semantically regularized embedding layer as the interface between CNN and RNN. Detected semantic concepts was directly utilized to generate image captions in [24]. Semantic compositional network (SCN) was proposed in [17] which detects semantic concepts from the image and uses the propability of each semantic concept to compose the parameters in a LSTM network.

With image captioning being a thriving topic, it is driven by the technical trials and improvements in both computer vision and NLP, and also importantly, the availability of relevant datasets. Other than the traditional image captioning task, efforts have been made to special captioning tasks. [25] modified mRNN to address the task of learning novel visual concepts. In [26], a hierarchical multimodal LSTM was proposed which jointly learns the embeddings of all sentences, phrases, images and image regions. Compared with mRNN, this hierarchical structure allows it to learn a dense embedding space and explicitly exploit the relations among 'parent' and 'children' image regions and phrases. [27] incorporated unpaired image data with labeling and unpaired text data to address the concept limitations in the image-sentence paired dataset. Similarly, [28] proposed the Novel Object Captioner (NOC) to describe object categories that are not present in the existing image-sentence datasets. 'Referring expression' was explored in [2], [29] to generate unambiguous descriptions. [3] presented a coherence recurrent convolutional network (CRCN) to describe an image stream in a storytelling manner utilizing blog data. Later, the authors brought up the personalized image captioning framework counting in users' vocabulary in previous documents [30]. A fill-in-the-blank image captioning task was introduced by [31].

The proposed sgLSTM captioning framework falls into the third category. A multimodal component is utilized to connect the visual and the textual spaces. Different from the existing methods, a novel guiding textual feature is proposed to emphasize the correlation between the description and the image content. The guiding text is extracted through a separate mLSTM model and serves as an additional input to the multimodal component.

B. Datasets

Due to the rising interest in image captioning task, a number of datasets have been brought up varying in sizes, formats

of the descriptions, and the collection process. One of the earliest benchmark datasets - Pascal1K [32] was proposed which consists of $1,000$ images selected from Pascal 2008 object recognition dataset [33]. Each image is associated with five sentence descriptions generated by AMT.

Later, based on Pascal 2010 action recognition dataset, [12] introduced the Visual and Linguistic Treebank (VLT2K) with $2,424$ images. AMT is again utilized with specific instructions to generate three, two-sentence descriptions for each image. Object annotation is available for a small subset of the images and VDRs are created manually for these images.

The Flickr8K [5] and Flickr30K [6] find their roots on images from Flickr. Although the images are collected based on user queries for specific objects or actions, the descriptions are generated in a manner similar to Pascal1K dataset where AMT workers provide five captions for each image. The original titles or descriptions from Flickr are not directly utilized to generate the captions in these two datasets. On the other hand, user-provided descriptions are employed in SBU1M [4] which contains approximately one million captioned images from Flickr. Strict filtering is applied that the downloaded image should contain at least one noun and one verb on predefined control lists.

The MS COCO dataset [34] is widely used recently for image captioning evaluation with $123,287$ images accompanied by five descriptions per image. Extensions of MS COCO dataset are available to meet specific needs of various tasks, e.g., question answering [35], unambiguous descriptions [2], and text detection and recognition [36]. The Déjà image captions dataset [37] makes use of the existing web data without additional human efforts. It consists 4 million images with 180K unique captions where lemmatization and stop word removal are employed to normalize the captions and create a corpus of near-identical texts. News images are explored in [38] and captions written by journalists along with auxiliary information are utilized to generate image descriptions. In [39], a large-scale dataset with over 40 million images and 300 million sentences is collected from publicly available Pins on Pinterest [40].

Although various datasets have been collected recently, expensive human label under specific instructions or strict filtering is often required especially for large datasets. However, as mentioned, image descriptions should come in different degrees of abstraction, i.e., descriptions could be abstract as in several words or a short term, or as a detailed paragraph in a storytelling way. FlickrNYC dataset is collected through the real-world user data, which meets the requirements for image description naturally and the additional human efforts for the dataset generation is minimized without NLP-based normalization.

III. DEEP CAPTIONING VIA SELF-GUIDING MULTIMODAL LSTM

The successful combination of CNN and RNN, especially LSTM, has been widely experimented in image captioning and related tasks. However, as observed by [22], the generated sentence sometimes is weakly coupled to the provided image but is strongly correlated to the high frequency sentences in the training dataset. This is due to the fact that the generated sentence is "drifted away" during the sequence prediction process. This problem exists especially for long sentences where the generation is carried out "almost blindly towards the end of the sentence". To address this issue, alternative extensions have been proposed by adding attention mechanism [41] and modifying the LSTM cell [22]. However, it is still challenging with an uncontrolled dataset with descriptions in arbitrary lengths and abstraction levels.

In this section, we first introduce the basic multimodal LSTM (mLSTM) image captioning framework which fuses the information of the input sentences and the corresponding image features in the multimodal component. It works effectively when the two input sources are strongly bonded. However, when this is not the case, it is difficult to maintain the correlation as the sentence generation goes on especially when the training dataset is not ideal for image captioning task.

As observed in FlickrNYC dataset, descriptions in shorter sentences tend to have a stronger bond with the image content compared with longer descriptions. Although they may not be syntactically sound to form a sentence, these short descriptions tend to accurately describe the locations, activities, objects, or events, as the images were taken. Some examples can be found in Fig. 2(a) where core information in these images are conveyed in the corresponding descriptions. On the other hand, long descriptions are valuable as the users may state their feelings, reasoning, personal experiences, or objects that are not depicted in the image. As shown in Fig. 2(b), these sentences are difficult to reproduce by the AMT workers even with instructions. However, some descriptions may not be strongly bonded with the visual content.

In order to generate image captions with adequate details related to the image content, we separate the data based on the different characteristics revealed. FlickNYC is divided into two subsets, $data_s$ with descriptions in short sentences or terms, and $data_l$ with descriptions in long sentences or paragraphs (the length is measured in the number of words. Details of the data separation can be found in Sec. IV-A). We start by training a mLSTM captioning model based on $data_s$. This captioning model aims to extract the key textual information provided an input image. This key information is later utilized to guide the training of sgLSTM based on $data_l$ to better link the description to the image content. This guiding information is represented through a sentence vectorizer and fed as another input to the multimodal component in sgLSTM.

A. Captioning with mLSTM

To train a caption model with $data_s$, we employ a variation of mRNN [20] due to its elegance and simplicity. The gated recurrent unit is replaced with LSTM in the proposed mLSTM model. The LSTM network [42] has been widely used to model temporal dynamics in sequences. Compared with the traditional RNN, it better addresses the issue of exploding and vanishing gradients. The basic LSTM block consists of a memory cell which stores the state over time and the gates which control how to update the state of the cell.
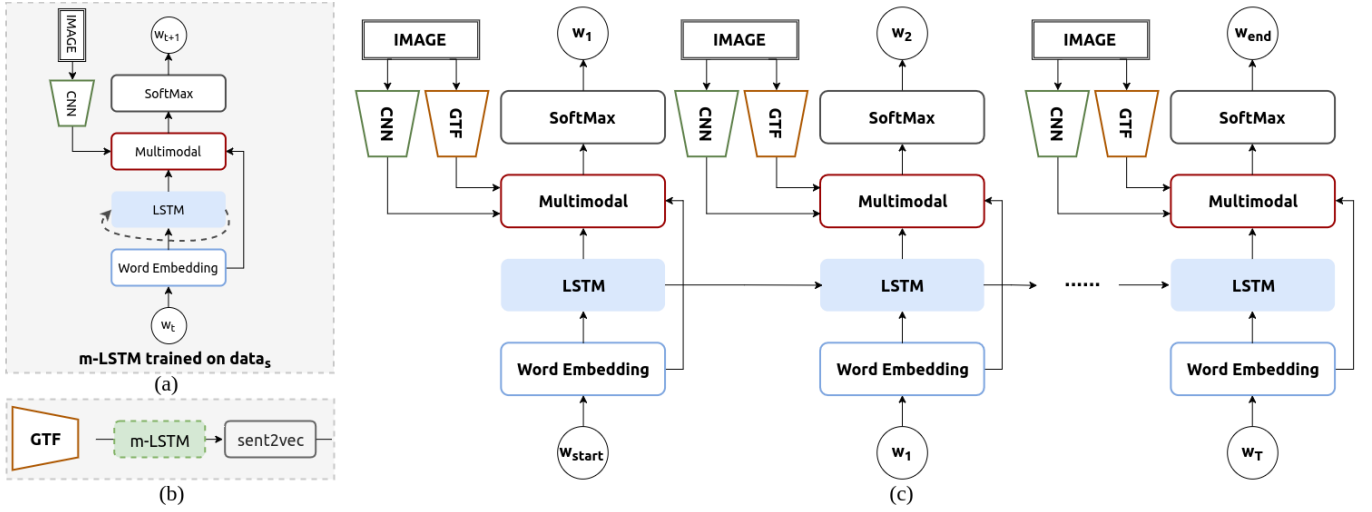
Fig. 3. Systematic flowchart of the proposed self-guiding multimodal LSTM (sgLSTM) captioning framework. (a) Basic multimodal LSTM (mLSTM) captioning framework trained on a subset of the FlickrNYC dataset with short descriptions (i.e., $data_s$). w_t denotes the t-th word in a sentence with words ranging from w_1 to w_T. A start sign w_{start} and an end sign w_{end} are added to all training instances in both $data_s$ and $data_l$. (b) Guiding text feature (GTF) extraction: to extract the text feature for self-guiding, we generate the descriptions utilizing mLSTM followed by the sentence vectorizer. (c) Illustration of the sgLSTM architecture. Compared with mLSTM, an additional textual feature is fed to the multimodal block which encodes the language information connected to the image content. The previous trained mLSTM is used to provide the additional textual feature, i.e., guiding text feature. Figure is best viewed in color.

As illustrated in Fig. 3(a), mLSTM is composed of a word embedding layer, an LSTM layer, a multimodal layer, and a softmax layer. It takes the training images and the corresponding descriptions as inputs. Each word in the sentences is encoded with one-hot representation before being fed to mLSTM training. The word embedding layer aims to map the one-hot vector to a more compact representation as shown in Eq. 1. Same as [20], we randomly initialize the embedding layer and learn W_e during training.

$$e_t = W_e \cdot w_t, \tag{1}$$

where w_t stands for the one-hot representation of word at step t. W_e is the mapping weight between the one-hot representation and the word embedding representation e_t.

There are many LSTM variants. In the proposed mLSTM model and later in sgLSTM, we adopt LSTM with peepholes [43] where the memory cell and gates within an LSTM block are defined as:

$$i_t = \sigma(W_{ic}C_{t-1} + W_{ih}h_{t-1} + W_{ie}e_t + b_i), \tag{2}$$

$$f_t = \sigma(W_{fc}C_{t-1} + W_{fh}h_{t-1} + W_{fe}e_t + b_f), \tag{3}$$

$$o_t = \sigma(W_{oc}C_t + W_{oh}h_{t-1} + W_{oe}e_t + b_o), \tag{4}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot g_1(W_{ch}h_{t-1} + W_{ce}e_t + b_c), \tag{5}$$

$$h_t = o_t \odot C_t, \tag{6}$$

in which $\odot$ denotes the element-wise product. $\sigma(\cdot)$ is the sigmoid nonlinearity-introduce function. $g_1(\cdot)$ is the basic hyperbolic tangent function. i_t, o_t, f_t, C_t, and h_t represent the state values of the input gate, output gate, forget gate, cell state, and hidden state at step t, respectively. $W_{[\cdot][\cdot]}$ and $b_{[\cdot]}$ denote the weight matrices and bias vectors for corresponding gates and states.

The word sequence is fed to the LSTM network by iterating the recurrence connection as shown in Fig. 3(a). Inception v3 [44] is used to extract the image features. They are connected with the language inputs though a multimodal component. The multimodal part fuses the language information represented as the dense word embedding and the LSTM activation with the image information represented using CNN as shown below:

$$mm_t = g_2(W_i \cdot I + W_d \cdot e_t + W_l \cdot \text{LSTM}_t), \tag{7}$$

where $g_2(\cdot)$ is the element-wise scaled hyperbolic tangent function [45] which leads to a faster training process than the basic hyperbolic tangent function. W_i, W_d, and W_l indicate the mapping weights between the image feature/ word embedding/ LSTM layer and the multimodal layer to learn during training.

The mLSTM model is learnt utilizing a log-likelihood cost function based on perplexity introduced by [20]:

$$C_{\text{mLSTM}} = \frac{1}{N_w} \sum_{i=1}^{N_s} L_i \cdot \log_2 \mathcal{P}(w_{1:L_i}^{(i)} \mid I^{(i)}) + \lambda_\theta \cdot \|\theta\|_2^2, \tag{8}$$

where $\mathcal{P}(\cdot)$ stands for the perplexity of a sentence given the image which is a standard measure for evaluating language model. N_w and N_s represent the number of words and the number of sentences in the training set. L_i is the length of the i-th sentence, and θ denotes the model parameters. λ_θ is used to leverage the two terms.

B. Captioning with sgLSTM

In this subsection, we describe in detail the training of sgLSTM with $data_l$. As mentioned, for some training instances in $data_l$, there is not a strong connection between

the textual description and the image content. In other words, additional textual features are needed during training. Therefore, as presented in Fig. 3(b), we use mLSTM (trained on $data_s$) to assist the training by connecting it to a sentence vectorizer. These two components form a guiding textual feature (GTF) extractor. This guidance feature aims to provide additional textual information for each training instance in $data_l$, which tends to emphasize the correlation between the textual and the visual domains. Compared with the basic mLSTM architecture, sgLSTM carries additional information in the multimodal component. Same as the image feature, the guiding textual features are fed into the multimodal component on each timestep as auxiliary information. This additional textual feature implicitly encodes the semantic information related to the image, such as location, activity, etc.

The sgLSTM architecture is composed of four layers in each timestep similar to mLSTM. The embedding layer encodes the one-hot word representation into a dense word representation. The weights in the embedding layer are learnt from the training data aiming at encoding the syntactic and semantic meaning of the words. The word representation after the embedding layer serves as the input to the LSTM layer. Same as mLSTM, we adopt a basic LSTM block with peepholes. After this layer, a multimodal layer is set to connect the CNN-based image feature, the dense word representation, the recurrent layer output, and the proposed guiding texture feature. The activations of these four inputs are mapped to the same multimodal feature space as the activation of the multimodal layer:

$$mm_t^{'} = g_2(W_i^{'} \cdot I + W_d^{'} \cdot e_t + W_l^{'} \cdot \text{LSTM}_t + W_t^{'} \cdot T), \quad (9)$$

where $W_i^{'}$, $W_d^{'}$, $W_l^{'}$, and $W_t^{'}$ represent the corresponding weighting matrices between the image/ word embedding/ LSTM layer/ guiding text feature and the multimodel component.

Extraction of Guiding Textual Features:

To generate the guiding textual feature for a certain image, we first utilize mLSTM trained on $data_s$ to output a short description (i.e., the raw sentence for the guiding textual feature). Beam search is adopted in the process to avoid the exhaustive search in the exponential search space. It is widely used in RNN-based captioning models [3], [20], [22] due to its efficiency and effectiveness. The top 1 ranked sentence is selected for further vectorization. Fig. 4 presents examples of the guiding texts generated by mLSTM trained on $data_s$. The images are from $data_l$ and therefore, the original descriptions are relatively long. As demonstrated, the guiding texts either provide core information that is not conveyed in the original descriptions, e.g., authorship info (jackson pollock), landmark name (radio city music hall), and season info (snowy day), or emphasize the key image content buried in long sentences, e.g., event (macy's thanksgiving day parade), and location (grand central terminal). We also observe some interesting results that reveal some underneath feelings of the images themselves, e.g., 'snow, dirt, love, and loneliness'.

A group of sentence vectorizers are investigated to vectorize the sentence or term generated by mLSTM. In general, we adopt the word2vec with fusion scheme, i.e., each word in the sentence is vectorized and then these word vectors are combined to produce the final output. Three word2vec schemes are experimented:

- word2vec-GloVe: we adopt pre-trained GloVe [46], i.e., Global Vectors for word representation, as the word vectorizer. The word vectors are trained through aggregated global word-word co-occurrence statistics from a corpus combining Wikipedia 2014 and Gigaword 5. Four different feature dimensions, 50, 100, 200, and 300 are tested.
- word2vec-NYC: compared with word2vec-GloVe, word2vec-NYC is a local word vectorizer trained on the textual data in FlickrNYC. This model is trained utilizing gensim [47] and a 128-dimensional vector is generated per word.
- word2vec-short: a word embedding mapping is learnt when training mLSTM on $data_s$. In this word vectorizer, the representation after the word embedding layer is employed directly as to map a word to a $1,024$-dimensional vector.

After representing each word in vector, two different fusing methods are investigated:

- Average: an average of all the word vectors in a sentence is calculated to obtain the final sentence vector.
- TF-IDF: the word vectors are combined using term frequency-inverse document frequency (TF-IDF) weighting scheme to generate the final representation.

The various vectorization methods look into the mapping problem from different angles, utilizing a global corpus or a local dataset, and in different dimensionality. As later shown in Table II, sgLSTM based on word2vec-GloVe with TF-IDF weighting under feature dimension 50 (denoted as sgLSTM-GloVe-tfidf-50) appears to work the best among all the 12 vectorization schemes (more details can be found in Sec. IV-C).

Training sgLSTM

Same as mLSTM, a log-likelihood cost function related to the perplexity is utilized for training sgLSTM as shown in Eq. (8). Normalization regarding the number of words corrects the bias over shorter sentences during the caption generation process, and therefore, is suitable for FlickrNYC with images in various lengths.

IV. EXPERIMENTAL RESULTS

In this section, the effectiveness of the proposed self-guiding strategy is verified experimentally on FlickrNYC. We start by a deeper introduction of FlickrNYC dataset followed by the implementation details of the proposed system. Afterwards, experimental evaluation results are presented and analyzed.

A. FlickrNYC Dataset

The FlickrNYC dataset is composed of $306,165$ images in total collected from Flickr with key word 'new york city'.

Fig. 4. Examples of the guiding text (marked in red) generated by mLSTM (trained on $data_s$) compared with the original descriptions (marked in blue) provided by the Flickr users. All 6 images come from $data_l$. The guiding text provides supplementary information that is strongly related to the image content. Figure is best viewed in color.

More specifically, Flickr search API is employed to crawl image-description data based on the key word, i.e., photos whose title, description, or tags contain 'new york city' will be fetched. After capturing the images and their corresponding metadata, each image is accompanied with 1 reference description provided by the original user. Images without valid descriptions are discarded. We only perform a light pre-processing without modifying the sentence structure utilizing NLTK Toolbox [48], textacy[1], and self-defined regular expressions, to remove unnecessary personal information (e.g., URLs, copyright declaration, camera information, personal social media accounts, advertisements, etc.) since we aim to keep the syntactic and semantic variety from the real-world user data.

After the textual pre-processing, the dataset is divided based on the number of words in the descriptions. Images with descriptions shorter than 10 form dataset $data_s$ with $165,374$ images for training and $1,000$ for testing. The rest $139,791$ images form $data_l$ in which $137,791$ is used for training, $1,000$ for validation and $1,000$ for testing. Table I provides the statistics of distributions based on the description lengths in FlickrNYC. Sample images and the corresponding descriptions can be found in Fig. 2.

Different from the traditional way to create the vocabulary which removes all words that contain non-alphanumeric characters or even non-alphabetic characters, the vocabulary

[1] http://textacy.readthedocs.io/en/latest/index.html

TABLE I
STATISTICS OF IMAGE DISTRIBUTION BASED ON THE DESCRIPTION LENGTHS IN FLICKRNYC DATASET.

sentence length	$1-5$	$6-10$	$11-15$	$16-25$	≥ 26
num of instances	$94,180$	$85,691$	$45,291$	$37,108$	$43,895$

build-up process for FlickrNYC is tricky: 1) Since the dataset is based upon New York city in which multiple landmark names contain combinations of alphanumeric characters (e.g. '5th avenue' or informally '5 ave' in some descriptions), therefore, numeric and alphanumeric words should not be eliminated in the vocabulary. Moreover, words that contain or are connected by punctuations should also be considered, e.g., 'Macy's', 'it's', 'let's', 'sight-seeing', 'African-Americans', etc. 2) Although one image is accompanied by one description, the description is not restricted to one sentence. As observed, some descriptions can be long containing multiple sentences. To better model the continuity of a paragraph of sentences, punctuations such as ' , ', ' . ', ' ! ', and ' ? ' should be considered as part of the vocabulary list. 3) FlickrNYC utilizes uncontrolled real-world text data, which indicates that the usage of words can be informal. However, we find sometimes this informality is valuable since it reveals the emotions of the users, such as Emoticons (' :-) ', ' :-P ', etc.) and exaggerated expressions ('soooo', 'superrrr', etc.). Therefore, in order to keep all the information mentioned above, after tokenization and converted to lowercase, words that appear at least 3 times

in the training set are kept to create the vocabulary[2]. The final vocabulary size is $22,230$.

B. Implementation Details

The proposed framework is built upon mRNN[3] with TensorFlow [50]. The inception v3 [44] pretrained on ImageNet [9] is used to compute CNN features as image representation. Feature dimension for this image representation is $2,048$. In both mLSTM and sgLSTM, the word embedding layer is with $1,024$ dimension. The LSTM layer and the multimodal layer are with $2,048$ dimensions. We assign 0.5 dropout rate to all three layers.

Both mLSTM and sgLSTM models are trained with RMSProp optimizer [51]. We apply the stochastic gradient descent (SGD) with mini-batches of 64. The beam search size is set to be 3. The top ranked sentence generated by mLSTM based on training data in $data_s$ is utilized for guiding textual feature extraction. As mentioned, three word2vec schemes are tested, i.e., word2vec-GloVe, word2vec-NYC, and word2vec-short. Four different sets of pre-trained word vectors are tested for word2vec-GloVe with dimensions 50, 100, 200, and 300. Dimensions for word2vec-NYC and word2vec-short based representations are 128 and $1,024$, respectively.

C. Experimental Evaluations

In order to select the best vectorization scheme for the guiding textual feature, certain objective criterion is needed to evaluate each scheme. Popular evaluation metrics for image captioning tasks include BLEU [52] (BLEU@$1,2,3,4$), METEOR [53], ROUGE-L [54], and CIDEr [55]. However, none of the criteria listed is a perfect metric for the evaluation task in our case since the ground-truth descriptions in FlickrNYC dataset are noisy. An example can be found in the bottom left image in Fig. 5 in which the original description is 'december 6th'. On the other hand, the proposed sgLSTM framework outputs description 'boaters on the lake in central park near the bow bridge' which is a much better description compared with the original one given the image content. However, this superiority will not be reflected in the numerical metrics listed above.

Despite the challenges in evaluating the vectorization schemes in the proposed framework, there still exists a large portion of data in FlickrNYC which suits 'perfectly' for captioning task. Therefore, a small validation dataset is separated from $data_l$ and utilized to evaluate the 12 different vectorization schemes. Among the various GloVe-based approaches, there is no clear correlation between the size of the feature dimensions and the numerical results. TF-IDF fusing achieves better performance for some dimensions, but not always. Based on the experimental results, sgLSTM-GloVe-tfidf-50 achieves the best performance quantitatively. Thus we adopt sgLSTM based on word2GloVe in TF-IDF weighting with dimension 50 as the final setting and all the results reported in this paper are based on this setting unless stated otherwise.

Table II presents the numerical results based on $1,000$ testing images in $data_l$. The proposed sgLSTM framework is compared with mRNN [20], mLSTM, and among different vectorization settings. The results of the top 4 performers in the previous verification step are included in this table. mRNN is trained based on all the training data in FlickrNYC (i.e., $data_s$ and $data_l$) utilizing the source code [4] provided by the authors. We use all the default settings used in the original paper. mLSTM-long represents the mLSTM captioning model trained on $data_l$ which shares the similar limitations as mRNN due to the lack of correlation between the images and descriptions in the training data. As shown in Table II, sgLSTM-GloVe-tfidf-50 gives the best performance numerically almost among all the evaluated methods, which is consistent with our observation in the verification step.

The zero numbers shown in Table II for mRNN might be better explained by looking into the results in Fig. 5. A direct training over the whole dataset tends to put a preference into high frequency sentences in the training dataset, which may be unrelated to the test image itself. Therefore, when it comes to numerical evaluations, a total miss of the core concept in the image content leads to a low score. On the other hand, by integrating the guiding textual features into the training process, the proposed sgLSTM model manages to generate accurate descriptions related to the image content, and sometimes, the generated descriptions are more meaningful than the original ones provided by the Flickr users as demonstrated in Fig. 5.

To confirm the proper sizes of the LSTM layer and the multimodal layer are selected, additional experiments are conducted by varying the sizes of these two layers. As shown in Table III, these additional networks are represented as sgLSTM-A-B where A indicates the size of the LSTM layer and B represents the multimodal layer size. All the networks are based on word2vec-GloVe with TF-IDF weighting scheme under feature dimension 50. As demonstrated, sgLSTM-2048-2048 (i.e., the final setting adopted in this paper) achieves the best performance.

Fig. 6 provides more examples by comparing the results generated by the proposed framework with the original descriptions provided by the Flickr users. Our sgLSTM model accurately generates descriptions that are closely related to the image content and successfully recovers key image features (e.g., weather, objects, activities) and the landmarks. In Fig. 7, several results are presented that could be further improved. Additional preprocessing steps could be performed before the training to remove terms such as 'more on blog' as shown in the figure. NLP techniques can be applied in avoiding a repetitive pattern shown in the 'bear' example. For certain cases, solely based on image content, it is difficult to generate accurate descriptions even for New Yorkers - as to differentiate between 'sunset' and 'sunrise', 'brooklyn' and 'bronx', or to decide the name of a certain building with little information - to that extent, it would be better to remove the ambiguous information that cannot be predicted based on the image content in the final description.

---

[2]If a word only contains alphabetic characters, we employ WordNet [49] to rule out typos and non-English words.

[3]https://github.com/mjhucla/TF-mRNN

[4]https://github.com/mjhucla/TF-mRNN

TABLE II
NUMERICAL RESULTS OF THE PROPOSED FRAMEWORK COMPARED WITH OTHER METHODS BASED ON THE TESTING IMAGES IN $data_l$.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
mRNN [20]	0.036	0.019	0.000	0.000	0.021	0.084	0.003
mLSTM-long	0.310	0.257	0.216	0.169	0.145	0.244	0.696
sgLSTM-NYC-ave	0.237	0.194	0.160	0.133	0.122	0.198	1.270
sgLSTM-GloVe-tfidf-50	0.417	**0.381**	**0.359**	**0.339**	**0.211**	**0.365**	1.010
sgLSTM-Glove-ave-200	**0.426**	0.364	0.333	0.308	0.190	0.342	**2.384**
sgLSTM-GloVe-tfidf-300	0.281	0.279	0.278	0.276	0.154	0.248	0.177

Fig. 5. Descriptions generated by the proposed framework (marked in red) compared with mRNN [20], mLSTM-long (mLSTM trained on $data_l$), mLSTM-full (mLSTM trained on all training data) and the original descriptions (marked in blue) provided by the Flickr users. The guiding texts are also provided. To help with the evaluation, the ground-truth locations are marked in each image (usage of different colors is for the best contrast). Figure is best viewed in color.

TABLE III
NUMERICAL RESULTS OF THE PROPOSED FRAMEWORK AT VARIOUS
PARAMETER SETTINGS BASED ON THE TESTING IMAGES IN $data_l$.

	BLEU-1	METEOR	ROUGE-L	CIDEr
sgLSTM-2048-4096	0.163	0.051	0.112	0.673
sgLSTM-2048-2048	**0.417**	**0.211**	**0.365**	**1.010**
sgLSTM-1024-2048	0.310	0.042	0.125	0.595
sgLSTM-1024-1024	0.023	0.004	0.016	0.047

sgLSTM with MS COCO Dataset

Although the proposed sgLSTM captioning framework is designed to tackle real-world datasets, it is worth experimenting the proposed self-guiding scheme with controlled dataset. We adopt MS COCO [34] as our testbed.

Since all the training instances in MS COCO are considered 'perfect', we randomly divided the training dataset (i.e., the combined version of Train2014 and Val2014) into two equal parts $data_a$ and $data_b$. The first training subset $data_a$ is utilized to train mLSTM. The trained mLSTM is then combined with the sentence vectorizer (GloVe-tfidf-50) to generate the guiding text feature for the second training subset $data_b$. Afterwards, an sgLSTM is trained based on $data_b$. We use the trained sgLSTM model to generate results of the MS COCO test set (i.e., Test2014) which are then evaluated by MS COCO evaluation server [56].

Due to the strong correlation between the image content and the descriptions in MS COCO, the guiding text feature actually provides duplicated information as the ground-truth descriptions when training sgLSTM. Therefore, the training of sgLSTM only sees half of the training dataset, while the compared approach observed the full dataset. As shown

Ours-sgLSTM: a rainy day in new york city. the hustle, time square.
Orig Descrip: time square happenings

Ours-sgLSTM: view of lower manhattan from dumbo section, brooklyn.
Orig Descrip: brooklyn thirteen

Ours-sgLSTM: horse and carriage ride in central park, manhattan.
Orig Descrip: new york city: rain in may.

Ours-sgLSTM: the hearst tower in nyc, located at 57th st and 8th ave, is the home of the hearst publishing company founded by william randolph hearst.
Orig Descrip: new york city trip

Ours-sgLSTM: grand central station - even without grand shafts of light it's rather photogenic
Orig Descrip: grand central terminal

Ours-sgLSTM: there are tons of beautiful flowers in the park in the summer, so of course there are also lots of bees around.
Orig Descrip: on the high line, new york city

Ours-sgLSTM: the aircraft carrier uss intrepid, as seen from the deck of the uss intrepid
Orig Descrip: uss intrepid (cv/cva/cvs-11), also known as the fighting i, is one of 24 essex-class aircraft carriers built during world war ii for the united states navy.
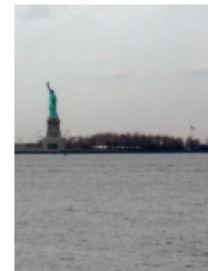
Ours-sgLSTM: taken on the walkway of the brooklin bridge, heading to brooklin.
Orig Descrip: experimental shot with a sepia effect added. not sure it works though!

Ours-sgLSTM: marble statue of a youthful hercules. the metropolitan museum of art.
Orig Descrip: there are no words.

Ours-sgLSTM: statue of liberty as seen from the staten island ferry.
Orig Descrip: i don't like boats. this is as close as we got. also, i didn't goose the color. this is how it came out of the camera.

Fig. 6. More results generated by the proposed framework compared with the original descriptions provided by the Flickr users. Blue indicates a precise description of the image content that does not even appear in the original descriptions. Green shows a successful recovery of the landmarks. Figure is best viewed in color.

TABLE IV
NUMERICAL RESULTS OF SGLSTM COMPARED WITH OTHER METHODS BASED ON THE TESTING IMAGES (TEST2014) IN MS COCO [34]. SGLSTM IS BASED ON GLOVE-TFIDF-50.

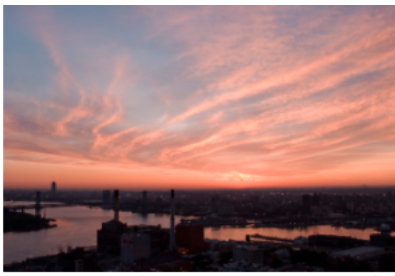|	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE-L	METEOR
mRNN-c5 [20]	0.680	0.506	0.369	0.272	0.791	0.499	0.225
mRNN-c40 [20]	0.865	0.760	0.641	0.529	0.789	0.640	0.304
sgLSTM-c5	0.679	0.499	0.365	0.269	0.821	0.499	0.235
sgLSTM-c40	0.863	0.753	0.633	0.519	0.829	0.643	0.320

in Table IV, although with only half of the training data, sgLSTM achieves comparible performance compared with mRNN measured in BLEU@1, 2, 3, 4 and ROUGE-L, even better performance measured in CIDEr and METEOR.

D. Discussions

Majority of the existing image captioning frameworks rely on the controlled training datasets, i.e., image content and the descriptions are strongly correlated and there is minimal varia-

Fig. 7. Results generated by the proposed framework that could be further improved. Purple indicates wrong or unrelated phrases. Red shows a wrong location or activity based on the original descriptions provided by the users. However, for some cases, these locations or activities cannot be recovered solely based on the image content. Figure is best viewed in color.

tion in the text format. Moreover, some semantic-based image captioning frameworks also depend on certain pre-knowledge (e.g., pretrained on other datasets such as ImageNet), which leads to degraded performance with new datasets such as FlickrNYC due to object differences.

On the other hand, the proposed self-guiding scheme can be naturally plugged into many state-of-the-art captioning algorithms when dealing with real-world datasets. Given any real-world datasets, is it always easy to separate the less noisy data from the rest? The answer is it depends. Since the data is collected from uncontrolled environment, it is important to "observe" the data which is not necessary for the mainstream captioning datasets. In the worst case, human experts can be included into the data separation process. However, different from the aforementioned description generation by humans such as AMT, this will be a simpler classification task and we still use the real-world data. As far as there is adequate useful information in the training dataset, self-guiding scheme is effective without the need of outside data.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a novel self-guiding mul-timodal LSTM captioning framework which targets at a more effective training over uncontrolled real-world dataset. A new FlickrNYC dataset is introduced as the testbed to verify the proposed self-guiding scheme. The portion of data, in which the textual description strongly correlates with the image content, is utilized to train a mLSTM model to extract the textual features. Afterwards, the additional features are utilized to guide the training process of the caption generation based on the rest of the data. Experimental results demonstrate the effectiveness of the proposed framework in generating descriptions that are syntactically correct and semantically sound.

In the future, we will demonstrate the effectiveness of self-guiding scheme by experimenting on more state-of-the-art image captioning frameworks, i.e., boosting their performance on FlickrNYC. Due to the noisy labels in real-world datasets, there are limitations in the existing evaluation tools to perform parameter selection and comparison. It is worth developing a systematic way to perform evaluation in these cases. We also plan to extend our work to use self-guiding to assist caption dataset generation and in other computer vision tasks.

REFERENCES

[1] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank, "Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures," *Journal of Artificial Intelligence Research*, vol. 55, pp. 409–442, 2016.

[2] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy, "Generation and Comprehension of Unambiguous Object Descriptions," in *CVPR*, 2016.

[3] C. C. Park and G. Kim, "Expressing an Image Stream with a Sequence of Natural Sentences," in *NIPS*, 2015.

[4] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2Text: Describing Images Using 1 Million Captioned Photographs," in *NIPS*, 2011.

[5] M. Hodosh, P. Young, and J. Hockenmaier, "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–4899, 2013.

[6] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," in *ACL*, 2014.

[7] S. Yagcioglu, E. Erdem, A. Erdem, and R. Çakıcı, "A Distributed Representation Based Query Expansion Approach for Image Captioning," in *ACL*, 2015.

[8] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *ICLR*, 2015.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "ImageNet: A Large-Scale Image Database," in *CVPR*, 2009.

[10] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell, "Language Models for Image Captioning: The Quirks and What Works," in *ACL*, 2015.

[11] Y. Verma and C. V. Jawahar, "A support vector approach for cross-modal search of images and texts," *Computer Vision and Image Understanding*, vol. 154, pp. 48–63, 2017.

[12] D. Elliott and F. Keller, "Image Description using Visual Dependency Representations," in *EMNLP*, 2013.

[13] M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, X. Han, A. Mensch, A. Berg, T. Berg, and H. D. III, "Midge: Generating Image Descriptions From Computer Vision Detections," in *EACL*, 2012.

[14] P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi, "TreeTalk: Composition and Compression of Trees for Image Descriptions," in *EMNLP*, 2014.

[15] L. G. M. Ortiz, C. Wolff, and M. Lapata, "Learning to Interpret and Describe Abstract Scenes," in *Annual Conference of the North American Chapter of the ACL*, 2015.

[16] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image Captioning with Semantic Attention," in *CVPR*, 2016.

[17] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic Compositional Networks for Visual Captioning," in *CVPR*, 2017.

[18] A. Karpathy and F.-F. Li, "Deep Visual-Semantic Alignments for Generating Image Descriptions," in *CVPR*, 2015.

[19] J. R. Kiros, R. Salakhutdinov, and R. Zemel, "Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models," in *EMNLP*, 2015.

[20] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep Cationing with Multimodal Recurrent Neural Networks (m-RNN)," in *ICLR*, 2015.

[21] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig, "From Captions to Visual Concepts and Back," in *CVPR*, 2015.

[22] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the Long-Short Term Memory model for Image Caption Generation," in *ICCV*, 2015.

[23] F. Liu, T. Xiang, T. M. Hospedales, W. Yang, and C. Sun, "Semantic Regularisation for Recurrent Image Annotation," in *CVPR*, 2017.

[24] Q. Wu, C. Shen, L. Liu, A. Dick, and A. van den Hengel, "What value do explicit high level concepts have in vision to language problems?" in *CVPR*, 2016.

[25] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille, "Learning like a Child: Fast Novel Visual Concept Learning from Sentence Descriptions of Images," in *ICCV*, 2015.

[26] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Hierarchical Multimodal LSTM for Dense Visual-Semantic Embedding," in *ICCV*, 2017.

[27] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, "Deep Compositional Captioning: Describing Novel Object Categories Without Paired Training Data," in *CVPR*, 2016.

[28] S. Venugopalan, L. A. Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko, "Captioning Images with Diverse Objects," in *CVPR*, 2017.

[29] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg, "ReferItGame: Referring to Objects in Photographs of Natural Scenes," in *EMNLP*, 2014.

[30] C. C. Park, B. Kim, and G. Kim, "Attend to You: Personalized Image Captioning with Context Sequence Memory Networks," in *CVPR*, 2017.

[31] Q. Sun, S. Lee, and D. Batra, "Bidirectional Beam Search: Forward-Backward Inference in Neural Sequence Models for Fill-in-the-Blank Image Captioning," in *CVPR*, 2017.

[32] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting Image Annotations Using Amazon's Mechanical Turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[33] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2010.

[34] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *ECCV*, 2014.

[35] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual Question Answering," in *ICCV*, 2015.

[36] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images," in *arXiv preprint arXiv:1601.07140*, 2016.

[37] J. Chen, P. Kuznetsova, D. S. Warren, and Y. Choi, "Déjà Image-Captions: A Corpus of Expressive Descriptions in Repetition," in *NAACL*, 2015.

[38] A. Tariq and H. Foroosh, "A Context-Driven Extractive Framework for Generating Realistic Image Descriptions," *IEEE Transactions on Image Processing*, vol. 26, pp. 619–632, 2017.

[39] J. Mao, J. Xu, Y. Jing, and A. Yuille, "Training and Evaluating Multimodal Word Embeddings with Large-scale Web Annotated Images," in *NIPS*, 2016.

[40] "Pinterest," https://www.pinterest.com/.

[41] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention," in *ICML*, 2015.

[42] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, pp. 1735–1780, 1997.

[43] F. Gers and J. Schmidhuber, "Recurrent nets that time and count," in *IJCNN*, 2000.

[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016.

[45] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, *Neural Networks: Tricks of the Trade: Second Edition*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[46] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *EMNLP*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[47] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010, http://is.muni.cz/publication/884893/en.

[48] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[49] P. University, "WordNet," 2010, http://wordnet.princeton.edu.

[50] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: http://tensorflow.org/

[51] G. Hinton, N. Srivastava, and K. Swersky, "Lecture 6e - RMSProp: Divide the gradient by a running average of its recent magnitude," 2014, neural Networks for Machine Learning.

[52] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *ACL*, 2002.

[53] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *ACL*, 2005.

[54] C.-Y. Lin and F. J. Och, "Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics," in *ACL*, 2004.

[55] R. Vedantam, C. L. Zitnich, and D. Parikh, "CIDEr: Consensus-based Image Description Evaluation," in *CVPR*, 2015.

[56] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, "Microsoft COCO Captions: Data Collection and Evaluation Server," in *arXiv preprint arXiv:1504.00325*, 2015.
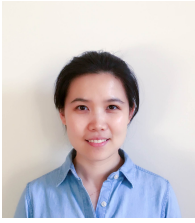
Yang Xian (S'14–M'17) received the B.E. degree from Southeast University, Nanjing, China, in 2009, the M.S. degree from New York University, New York, USA, in 2012, and the Ph.D. degree from the Graduate Center, the City University of New York, New York, USA, in 2017. Her research interests are in the synergic areas of computer vision, computational photography, and machine learning. She has been working on large-scale surveillance event detection, image captioning, and designing quality enhancement algorithms, i.e., super-resolution, completion, for both RGB images and depth maps.

YingLi Tian (M'99–SM'01–F'18) received the B.S. and M.S. degrees from Tianjin University, China, in 1987 and 1990, and the Ph.D. degree from Chinese University of Hong Kong, Hong Kong, in 1996. After holding a faculty position at National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, she joined Carnegie Mellon University in 1998, where she was a postdoctoral fellow at the Robotics Institute. She then worked as a research staff member in IBM T. J. Watson Research Center from 2001 to 2008. She is one of the inventors of the IBM Smart Surveillance Solutions. She is a professor in the Department of Electrical Engineering at the City College and the department of Computer Science at the Graduate Center, City University of New York since 2008. Her current research focuses on a wide range of computer vision problems from object recognition, scene understanding, medical imaging analysis, to human behavior analysis, facial expression recognition, and assistive technology. She is a fellow of IEEE.