KEYWORD-DRIVEN IMAGE CAPTIONING VIA CONTEXT-DEPENDENT BILATERAL LSTM

Xiaodan Zhang[1,2], Shengfeng He[3], Xinhang Song[2,4], Pengxu Wei[2], Shuqiang Jiang[2,4], Qixiang Ye[2], Jianbin Jiao[2], Rynson W.H. Lau[1]

[1]City University of Hong Kong; [2]University of Chinese Academy of Sciences;[3]South China University of Technology;
[4]Institute of Computing Technology, Chinese Academy of Sciences
zhangxiaodan10@mails.ucas.ac.cn; hesfe@scut.edu.cn; xinhang.song@vipl.ict.ac.cn; weipengxu11@mails.ucas.ac.cn;
sqjiang@ict.ac.cn; {qxye, jiaojb}@ucas.ac.cn; rynson.lau@cityu.edu.hk

ABSTRACT

Image captioning has recently received much attention. Existing approaches, however, are limited to describing images with simple contextual information, which typically generate one sentence to describe each image with only a single contextual emphasis. In this paper, we address this limitation from a user perspective with a novel approach. Given some keywords as additional inputs, the proposed method would generate various descriptions according to the provided guidance. Hence, descriptions with different focuses can be generated for the same image. Our method is based on a new Context-dependent Bilateral Long Short-Term Memory (CDB-LSTM) model to predict a keyword-driven sentence by considering the word dependence. The word dependence is explored externally with a bilateral pipeline, and internally with a unified and joint training process. Experiments on the MS COCO dataset demonstrate that the proposed approach not only significantly outperforms the baseline method but also shows good adaptation and consistency with various keywords.

Index Terms— Image Captioning, Keyword-driven, LSTM

1. INTRODUCTION

Image captioning, which aims to describe the semantic content of an image in a natural language form, has become one of the hot research problems in the artificial intelligence community. This is stimulated by practical applications including content-based image retrieval, visually-impaired assistance, and intelligent chatbots.

Existing image captioning methods can be roughly categorized into two classes: the bottom-up and the top-down approaches. The bottom-up approaches [1, 2, 3, 4] that group keywords with connection words to form a sentence are popular in early research due to their simplicity. The top-down ones [5, 6, 7, 8] share a similar pipeline that first applies convolutional neural network (CNN) to extract visual features of the image, and then fed them to a sequential model, e.g., recurrent neural network (RNN), to generate descriptions of the image content. The joint end-to-end training of CNN-RNN pipeline allows the top-down approaches achieve state-of-the-art performance.

Fig. 1. Examples of our proposed method. The values shown in the brackets are the BLEU-1 scores.

Notwithstanding the demonstrated success of the existing approaches, the existing image captioning methods are difficult to cover all the fine details of the image in the generated sentence. In fact, an image contains too much information to be precisely described in one sentence. One natural way to expand captions is to generate multiple captions through an increased beam (in beam search). However, this strategy simply increases the number of captions without introducing semantic diversity. Another way is to generate local descriptions based on image regions, while the region-based captioning works [7, 9] focus on the local region and thereby are unable to describe the image from a global perspective. On the other hand, image captioning suppose to be a highly customized task, and the user may have different emphases, which can not be fulfilled by current works.

978-1-5090-6067-2/17/$31.00 ©2017 IEEE

In this paper, we aim to inject the semantic guidance into the image captioning to enrich the image understanding with different emphases, where an additional keyword is introduced as the prior information to generate a customized sentence from a global perspective (see Fig. 1). Since the generated sentences are dependent on the given words, different given words result in different descriptions for the same image. This additionally given keyword can be obtained from different sources, such as image tags in semantic image retrieval, user input in human-computer interaction, or multiple object detection.

We propose a new language model, called Context-dependent Bilateral Long Short-Term Memory (CDB-LSTM) model, to predict the sentence driven by an additional keyword. CDB-LSTM contains two cascaded sub-models, which are jointly trained in an end-to-end pipeline. The first model generates the front part of the sentence in the reverse order (from the given word to the beginning of the sentence), while the second model generates the back part (from the given word to the end of the sentence) by taking the prediction of the first model into account. The front and the back parts of the sentence are concatenated together to generate the final sentence. These two models are unified and jointly optimized in an end-to-end training framework by considering the words dependence through a context transfer module. The proposed model can be considered as a combination of bottom-up and top-down models, as the output caption is driven by global image features and the additional keyword. Extensive experimental evaluations demonstrate that the proposed model shows good adaptation with various keywords (ground truth or detected keywords). In particular, quantitative evaluations and user study show that the sentences generated by the proposed method are more descriptive than traditional image captioning model.

2. METHOD

Given an image I with the visual feature v and an input word w, our goal is to generate a sentence $Y = \{y_1, \cdots, y_t\}$ to describe the image according to the given word w ($w \in Y$). To leverage the given word in the captioning process, we assume the sentence generation starts from the given word and thereby involves two directions: forward and backward generations. The backward prediction produces the front part of the sentence Y^b in an inverted order, and the forward prediction produces the back part of the sentence Y^f in a forward order. The final caption is defined as $Y = \{y_1, \cdots, y_t\} = \{Y^b, w, Y^f\}$. We formulate this interactive captioning process with a context-dependent bilateral LSTM model (CDB-LSTM), which combines the two directions together in an end-to-end structure and predicts the customized caption sequentially. The distribution of each time step p_t is generated by a linear transform followed by a softmax normalization:

$$p_t \propto exp\left(y_t | y_{1:t-1}, v, w; \Theta\right), \quad (1)$$

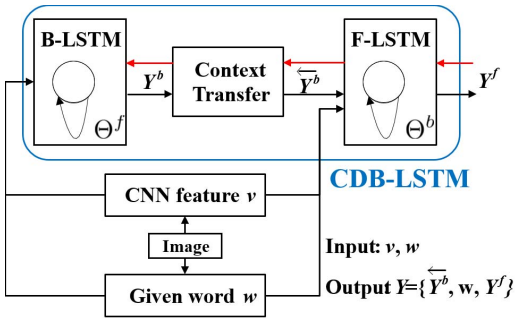

Fig. 2. Pipeline of the context-dependent bilateral LSTM model (CDB-LSTM). Our model contains two sub-models: the B-LSTM and the F-LSTM, which are connected via the context transfer module.

where Θ represents all the learning parameters of the model and exp denotes the softmax function.

2.1. Context-dependent Bilateral LSTM Model

We adopt the top-down paradigm and build upon the CNN+RNN architecture [5] in the proposed model. As shown in Fig. 2, the proposed Context-dependent Bilateral LSTM model (CDB-LSTM) contains two cascade sub-models: backward LSTM model (B-LSTM) and forward LSTM model (F-LSTM), which are connected together in a sequential way and trained in an end-to-end pipeline. The F-LSTM model takes the preliminary result of B-LSTM as input, to predict the full sentence by taking context dependence into account.

Context information between two cascade models is explored using a context transfer module. The inputs of CDB-LSTM are the CNN feature of the image v and the given word w. The final output sentence contains three parts: the backward output y_t^b, the given word w and the forward output y_t^f, which construct the sentence in a bottom-up manner.

In the traditional CNN+RNN pipeline, the CNN feature of the image v is fed to the RNN for the caption generation. As the hidden state h_t in RNN evolves over time t, the $t-$th output word y_t is drawn according to the probability vector p_t which is controlled by the current state h_t. The generated y_t will be fed back into RNN in the next time step as the input x_{t+1}. We use LSTM as the core of the RNN model, and all the LSTM models used in this work are built based on this flow path.

2.1.1. Backward Model

The backward LSTM model is the first step of the proposed CDB-LSTM model in Fig. 2, which predicts the front part of the sentence in a backward order. The technical process is similar to the traditional LSTM model, except that the start token is replaced with the given word. For an image I with the CNN features v and a given word w, assuming that the output phrase Y^b contains M words, there will be $M+1$ time steps

in this iteration. The prediction process is as follows:

$$
\begin{aligned}
x_0 &= W_{ix}\, v, && (2)\\
x_1 &= W_{wx}\, w, && (3)\\
h_t &= RNN(\, h_{t-1}\, ,\, x_t\,),\ t \in \{1, \cdots, M+1\}, && (4)\\
y_t\, , p_t &\propto exp(\, W_{hp} h_t\,),\ t \in \{1, \cdots, M+1\}, && (5)\\
x_{t+1} &= W_{ex} y_t,\ t \in \{1, \cdots, M\}, && (6)
\end{aligned}
$$

where W_{**} represents the corresponding linear encoding or decoding model: W_{ix} for the visual feature encoding, W_{wx} for the given word encoding, W_{ex} for the predicted word encoding, and W_{hp} for the output decoding. RNN is the recurrent model which uses LSTM as the core in this paper. The CNN features v is fed to the network at the first time step to provide RNN a quick overview of the image content. The given word w is fed to the network at the second time step to provide the model a strong restricted start, which is also to guarantee the following prediction is related to this word. This sub-model utilizes all the information of the last step as the input and makes the whole sentence consistent and reasonable. Eq. 5 to Eq. 6 are recursively applied during the captioning process. The loss of this sub-model is the sum of the negative log likelihood of the predicted word at each step:

$$
L(B/w) = -\sum_{t=1}^{M+1} logp\left(y_t | y_{1:t-1}, v, w; \Theta^b\right), \qquad (7)
$$

where Θ^b represents all the learning parameters of the backward model.

This B-LSTM model mostly predicts the content describes the given object. For example, when the given word is an object, the backward direction predicts the front part of the caption which is like the "adjective" phrase of the object. And the inverted output of this part $\overleftarrow{Y^b}$ will be transferred to the F-LSTM model.

2.1.2. Forward Model

The F-LSTM is the second step of our CDB-LSTM model. This sub-model predicts the back part of the sentence in a forward order. Similar to B-LSTM, the first input of the forward prediction model is the CNN feature v. The following M inputs are the inverted M outputs of the B-LSTM $\overleftarrow{Y_b}$. At $M+1$ time step, the given word is fed in and the later generated words are the new prediction:

$$
\begin{aligned}
x_0 &= W_{ix}\, v, && (8)\\
x_t &= W_{ex} \overleftarrow{Y^b}_t,\ t \in \{1, \cdots, M\}, && (9)\\
x_t &= W_{wx}\, w,\ t = M+1, && (10)\\
h_t &= RNN(\, h_{t-1}\, ,\, x_t\,),\ t \in \{1, \cdots, K\}, && (11)\\
y_t\, , p_t &\propto exp(\, W_{hp}\, h_t\,),\ t \in \{M+1, \cdots, K\}, && (12)\\
x_{t+1} &= W_{ex} y_t,\ t \in \{M+1, \cdots, K-1\}, && (13)
\end{aligned}
$$

where W_{**} and RNN have the same meaning as the backward model. The difference is that this forward model has $K = M + N + 1$ recurrent time steps. The first M time steps of this sub-model produce no output, which is just for the hidden layer to memorize the input words. The loss of this sub-model is the sum of the negative log likelihood of the later N predicted words:

$$
L(F/B, w) = -\sum_{t=M+1}^{M+N+1} logp\left(y_t | y_{M+1:t-1}, v, w; \Theta^f\right), \qquad (14)
$$

where Θ^f represents all the learning parameters of the forward model. The combination of the M output words of the B-LSTM Y^b, the N output words of the F-LSTM Y^f, and the given word w is our final prediction which focuses on the given word. As the two sub-models are cascaded-connected, the total number of the time steps of CDB-LSTM model is $2M + N + 2$.

2.2. Unified Loss Function

The training data for each image consists of input image features v, the keyword w, and the ground truth sequence Y_t. The ground truth sentence Y is divided to three parts according to the keyword: $Y_t = Y^b, w, Y^f$. Our goal is to learn the parameters of the proposed two models $\Theta^b = \{W^b_{ix}, W^b_{wx}, W^b_{ex}, W^b_{hp}, RNN^b\}$, $\Theta^f = \{W^f_{ix}, W^f_{wx}, W^f_{ex}, W^f_{hp}, RNN^f\}$. In order to train the two cascaded models in a unified fashion, the loss of one training example is defined as the total negative log-likelihood of the two sub-models:

$$
\begin{aligned}
L(B, w, F) &= L(B/w) + L(F/B, w) && (15)\\
&= -\sum_{t^b=1}^{M+1} logp\left(y^b_{t^b} | x^b_{t^b}; \Theta^b\right)\\
&\quad -\sum_{t^f=M+1}^{M+N+1} logp\left(y^f_{t^f} | x^f_{t^f}; \Theta^f\right). && (16)
\end{aligned}
$$

The combined loss $L(B, w, F)$ is able to converge faster, and achieves a lower error rate than training the two models individually (see the detailed result in the experiment section).

2.3. Context Transfer Module

The F-LSTM uses the output of the B-LSTM as part of the input, and the loss of the F-LSTM gives feedback to the B-LSTM. This inner connection is learned through the context transfer module, as shown in Fig. 2.

In the forward propagation of the proposed CDB-LSTM model, the output of the first step B-LSTM is fed into the F-LSTM through word transfer:

$$
x^f_t = W^f_{ex} \overleftarrow{Y^b_t}, \qquad (17)
$$

where x^f_t is part of the input of the F-LSTM, and Y^b_t is the output of the B-LSTM.

In the back propagation process of the CDB-LSTM model, the loss of the second step F-LSTM is sent to the B-LSTM through word transfer:

$$\frac{\partial L}{\partial y_t^b} = \frac{\partial L}{\partial L_t^b}\frac{\partial L_t^b}{\partial y_t^b} + \frac{\partial L}{\partial L_t^f}\frac{\partial L_t^f}{\partial y_t^b}. \quad (18)$$

The second part of Eq. 18 is the loss feedback through F-LSTM to B-LSTM, and this part can be a complement of the loss of B-LSTM, which allows searching the optimize result in a global view and bridge the gap of the two directions of the sentence.

3. INPUT KEYWORDS

The selection of input keywords is a key factor of our work in both training and testing. We build the keywords vocabulary from the ground truth sentences for training. While in the testing procedure, the given keywords can be arbitrary but should be image related.

3.1. Training: Ground truth

In this paper, MS COCO [10] is utilized for training, as this dataset contains both object labels and image captions, which is suitable for our keywords extracting. In the training procedure, we leverage the object information to select keywords from ground truth sentences that belong to the predefined 80 categories. We first build a dictionary D, which contains all the words that have appeared in the ground truth sentences. All the words in D with the noun tags ("NN" or "NNS") are collected via the NLTK toolbox [11]. The keywords vocabulary V is built automatically based on the similarity of the 80 object categories and the noun words in D via the word2vec tool of Gensim [12].

3.2. Testing: Arbitrary

For testing, the given keywords can be arbitrary but should be image related. In this paper, we have two different rules to select input keywords: ground truth keywords and detected object labels. To tackle the mismatching of the given words and our keywords vocabulary, we use the word2vec tool [12] to map the given word to our vocabulary V according to the word similarity.

4. EXPERIMENTS

In this section, extensive experiments are performed to evaluate the proposed method. The proposed method is implemented using Torch7 [13], and tested on a server with an i7 3.2GHz CPU, 32GB RAM, and a K40 GPU. We evaluate the proposed method on MS COCO [10] and follow the publicly available train/val/test splits of training, validating and testing sets in [6].

4.1. Implementation Details

The proposed CDB-LSTM network is fine-tuned based on the existing network [5, 6]. As a result, we share the same parameter setting, except the beam size. Theoretically, higher beam size results in better performance. Existing methods set this value to 7 [6], 10 [14] or 20 [5], which are empirically selected. In this paper, we set the beam size to 1, as we cannot observe much improvement with a higher beam size in our experiment. Moreover, lower beam size saves much computational time. The dimensions of the input layer, hidden layer and output layer of LSTM are set to 512. The learning rate is initiated to be 4e-4. The stochastic gradient descent (SGD) method of the LSTM model uses an adaptive learning rate algorithm RMSProp [15]. We use the finetuned VGGNet, which is finetuned based on the CNN-RNN pipeline of the traditional captioning task (NeuralTalk2 [1]) to extract the visual features. For the processing of ground truth sentences, we discard the words that occur less than five times in the training sentences when creating the word dictionary D. We also set the max length of a caption to 20, and captions longer than this value will be clipped. The size of D for COCO it is 9584. The keywords vocabulary V contains 537 words.

4.2. Evaluation on the Proposed Method

Evaluation metrics. To compute the similarity between the predicted sentence and the ground truth sentences, there are lots of metrics for evaluation and most of them come from machine translation. BLEU [16] is one of the most commonly used metrics so far in the image captioning literature. The main component of BLEU is n-grams precision of the generated caption with respect to the references. Each n-grams precision is computed separately and usually four grams are used to generate the final score. BLEU of high order n-grams indirectly measures the grammatical coherence. However, BLEU is criticized for favoring short sentences and it only considers precision but does not take recall into consideration. METEOR [17] is another widely used evaluation metric. It computes a score based on word level matches between the candidate sentence and references, with precision, recall, and grammaticality in consideration. The score is much lower than BLEU while METEOR has been shown more related to human judgments than any order of BLEU. CIDEr [18] and ROUGE-L [19] are also in our evaluation metrics list for comprehensive judgment.

The choice of evaluation metrics and the evaluation code are performed according to the MS COCO captioning challenges 2015 [10]. In this paper, we use the symbols B-n, M, C, R corresponding to the n-grams BLEU, METEOR, CIDEr, and ROUGE-L.

Evaluation results. To the best of our knowledge, the proposed method is the first one to introduce keyword guidance, thus we mainly compare to a bilateral LSTM model without considering context dependence, called independent-LSTM (I-LSTM). This model also consists of two LSTMs, each of which predicts the front and the back parts individually. The final sentence is the combination of these two LSTMs, which are trained without inner context dependence.

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

Table 1. Performance and comparison on the MS COCO dataset with different measures (%).

Method	B-1	B-2	B-3	B-4	M	C	R
Google NIC[5]	66.6	46.1	32.9	24.6	–	–	–
Hard-Attention[21]	71.8	50.4	35.7	25.0	23.04	–	–
gLSTM[14]	67.0	49.1	35.8	26.4	22.74	–	–
m-RNN[22]	67.-	49.-	35.-	25.-	–	–	–
ATT[8]	70.9	53.7	40.2	30.4	24.3	–	–
LSTM	69.8	52.2	38.5	28.7	23.9	53.4	42.9
I-LSTM(GR)	45.3	34.8	24.9	17.3	18.5	64.9	45.0
I-LSTM(GM)	66.1	50.6	35.0	23.4	20.9	77.2	48.0
CDB-LSTM(GR)	73.1	53.2	35.8	23.6	21.6	78.5	49.9
CDB-LSTM(GM)	78.8	58.3	40.4	27.5	23.4	83.6	51.8
CDB-LSTM(DR)	62.9	42.5	27.9	18.4	17.2	47.1	43.2
CDB-LSTM(DM)	76.3	56.1	38.9	26.5	22.5	77.3	51.4

Table 1 shows the results of the proposed method on the MS COCO dataset. In the evaluation, we have two different rules to select input words. The first one selects keywords in ground truth sentences as input, where "GR" indicates randomly selected input keyword from the ground truth sentences, and "GM" indicates the ideal maximum value obtained from all the ground truth keywords. The second rule uses an object detector [20] to generate multiple keywords via the word mapping strategy, where "DR" means to randomly select the detected keyword as input, and "DM" means the ideal maximum score. LSTM is our trained traditional language model with the finetuned VGGNet feature, and the CDB-LSTM and I-LSTM methods use the same CNN features.

The proposed context-dependent model largely outperforms independent model (I-LSTM), which also generate a sentence with the same given keyword. This is because that two models in I-LSTM are blind to each other, leading to incoherent, incomplete, or inaccurate results (see Fig. 3), which would also result in a high sentence length penalty in the evaluation. The ideal maximum result I-LSTM(GM) is significantly higher than I-LSTM(GR), which is mainly due to the weakness of length penalty. Compared to I-LSTM, the proposed CDB-LSTM shows good adaptation and consistency with various input words. The result indicates that the context dependence is important in this task. It may be intuitive that the additional input can help improve the image captioning performance, however, only using a straightforward embedding technique (I-LSTM) does not sufficient to leverage this information. In contrast, the proposed method is able to utilize the keywords to generate better sentences than the state-of-the-art methods.

It is not surprising that CDB-LSTM(GM) outperforms the others for it choose the ideal maximum value from multiple sentences driven by ground truth keywords. Randomly selecting one keyword from multiple candidates is more reasonable. One problem of our keyword-driven results is that these descriptions have different emphases for the same image, thus have a slight bias from the ground truth sentences which describe the main content of the image. Take the bottom image in Fig 1 for example, the sentences driven by the

Fig. 3. Illustration of the superiority of the proposed CDB-LSTM over I-LSTM: the independent LSTM have two directions that blind to each other, and thus more likely to predict incoherent, inaccurate, or incomplete sentences.

words "baseball" and "children" are better than the result of LSTM but the evaluation scores are lower for they are quite different from the ground truth sentences. Anyway, our CDB-LSTM(GR) outperforms the LSTM model and is comparable to the state-of-the-art methods. Compared to the traditional captioning models, our model is able to generate various sentences with different emphases from the global perspective. The following human evaluation is more reliable for comparison between CDB-LSTM and the traditional LSTM model without the evaluation bias.

The accuracy of the keywords is very important in our work. The captions driven by keywords come from object detector (DR, DM) are slightly inaccurate than using ground truth keywords (GR, GM) due to the unreliable of predicted keywords, and a good object detector and may help in the future.

4.3. Human Evaluation

Each of the automatic measures computes a score that indicates the similarity between the system output and one or more human-written reference texts. However, these automatic measures are criticized for they have weakly negative or no correlation with human judgments [1, 23, 24]. Moreover, as a task focusing on the keyword, involving the user in the system is more reasonable for evaluating the proposed model with interaction. We randomly choose 100 images from MS COCO. 50 participants are invited to give a word to describe the image first, and then they are asked to rate the descriptiveness for the two automatically generated sentences. One sentence is generated by the baseline and the other is our customized caption. The descriptiveness is rated on a four-point scale ([25, 5]), and higher is better.

Fig. 4 shows the human evaluation result. It can be observed that the proposed CDB-LSTM performs better with higher cumulative distribution. By allowing user input keywords, the generated sentences consistently to be more descriptive than the traditional LSTM.

Fig. 1 shows some example descriptions generated by the

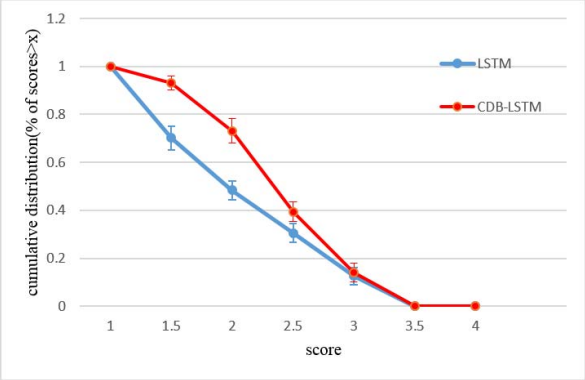

Fig. 4. Human evaluation on the random selected images.

proposed method. It can be seen that the proposed model is able to produce reasonable captions with various input keywords. Note that no matter the given word is salient or not, our model is able to generate a reasonable sentence driven by the given keyword (e.g., "gloves" in the bottom example). Our results can also focus on fine details that are not describable using the traditional image captioning pipeline (e.g., "table" in the upper example).

5. CONCLUSION

In this paper, we proposed a new problem, keyword-driven image captioning, by interactively describe the image according to user guidance. To this end, we propose a new language model, context-dependent bilateral LSTM, to generate caption based on an additional keyword. The customized caption is generated by two cascaded LSTM models, which are connected and jointly trained in an end-to-end framework. More importantly, two LSTM models heavily depend on each other to prevent generating an incoherent and inaccurate sentence. The proposed interactive captioning model is able to generate various sentences with different emphases from the global perspective and thus more suitable for practical applications. Both of the automatic metrics and human evaluation demonstrate the superiority of the proposed method.

6. ACKNOWLEDGEMENT

This work was partially supported by the NSFC under Grant 61671427 and 61532018, Beijing Municipal Science and Technology Commission under Grant Z161100001616005 and D161100001816001. This work was also partially supported by the Science and Technology Development Fund of Macao SAR (010/2017/A1).

7. REFERENCES

[1] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, Siming Li, Yejin Choi, A.C. Berg, and T.L. Berg, "Babytalk: Understanding and generating simple image descriptions," *TPAMI*, vol. 35, no. 12, pp. 2891–2903, 2013.

[2] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. & Choi, "Generalizing image captions for image-text parallel corpus," in *ACL*, 2013, pp. 790–796.

[3] P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi, "Treetalk: Composition and compression of trees for image descriptions.," *TACL*, vol. 2, no. 10, pp. 351–362, 2014.

[4] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *ECCV*, 2010, pp. 15–29.

[5] O. Vinyals, A. Toshev, S. Bengio, and D. & Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015, pp. 3156–3164.

[6] A. Karpathy and F. Li, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015, pp. 3128–3137.

[7] J. Johnson, A. Karpathy, and F. Li, "Densecap: Fully convolutional localization networks for dense captioning," in *CVPR*, 2016, pp. 4565–4574.

[8] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo, "Image captioning with semantic attention," in *CVPR*, 2016, pp. 4651–4659.

[9] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *arXiv:1602.07332*, 2016.

[10] T. Lin, M. Maire, S. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," in *ECCV*, 2014, pp. 740–755.

[11] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*, O'Reilly Media, Inc., 2009.

[12] R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," in *LREC Workshop on New Challenges for NLP Frameworks*, 2010.

[13] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," in *NIPS Workshop*, 2011.

[14] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding long-short term memory for image caption generation," in *ICCV*, 2015.

[15] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Networks for Machine Learning*, vol. 4, no. 2, 2012.

[16] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.

[17] A. Lavie and A. Agarwal, "Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments," in *Second Workshop on Statistical Machine Translation*, 2007, pp. 228–231.

[18] R. Vedantam, C Lawrence Z., and D. Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015, pp. 4566–4575.

[19] C. Lin, "Rouge: A package for automatic evaluation of summaries," in *ACL workshop*, 2004, vol. 8.

[20] S. Ren, K. He, R. Girshick, J. Sun, S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[21] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

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

[23] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, "Improving image-sentence embeddings using large weakly annotated photo collections," in *ECCV*, 2014, pp. 529–545.

[24] D. Elliott and F. Keller, "Comparing automatic evaluation measures for image description," in *ACL*, 2014, vol. 452, p. 457.

[25] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *JAIR*, vol. 47, pp. 853–899, 2013.