# A NOVEL SEMANTIC ATTRIBUTE-BASED FEATURE FOR IMAGE CAPTION GENERATION

*Wei Wang, Yuxuan Ding, Chunna Tian**

School of Electronic Engineering, Xidian University, Xian China

## ABSTRACT

Image captioning is challenging because it connects computer vision and natural language processing. It requires not only sensing objects but also the interrelations and context in an image to generate natural language descriptions. In this paper, we propose to extract a novel visual feature weighted by salient semantic attributes, which is fed to the encoder of Long Short Term Memory (LSTM). Semantic attributes are important to exploit more semantic-related information in images and describe the salient scenes to enhance the accuracy of generating image captions. Based on the Multiple Instance Learning (MIL) architecture on VGG-16 network, we design transferring rules that map high probability attributes to the feature vector in fc7 layer. It results in more semantic-related visual features. Our model can recognize richer details of images effectively and achieve the state-of-the-art performance on MSCOCO 2014 dataset under standard metrics.

***Index Terms*****—** Semantic attributes, Mapping, Captioning, Feature

## 1. INTRODUCTION

Image captioning, generating natural language descriptions for images automatically, is one of the major challenges in image understanding. It requires detecting various objects in a given image and expressesing their layout and interactions correctly. There are two kinds of approaches in image captioning task: Bottom-up and top-down based ones. Inspired by the successes of machine translation [1], the top-down based approaches usually use an encoder-decoder (Convolutional Neural Network and LSTM) architecture [2][3][4][5]. It encodes the entire image into a compact representation, then translates the CNN feature into natural language descriptions by a caption generation network. The bottom-up based approaches [6][7][8] are consist of several separate sub-tasks such as detecting objects or attributes, sorting words and phrases, generating sentences using language model and ranking for top captions [6]. Besides the classic image feature representation [9], the main visual feature extraction methods are based on deep networks at present. However, the state-of-the-art architecture still suffers from the insufficiency of image representation. Though local regions of an image contain amounts of information, most works only feed the holistic image feature to the caption generating network at the initial step. Thus, how to enhance the image representation is the key issue of generating image captions effectively.

Semantic attributes contain a set of words which can be detected by bottom-up detectors trained on labeled images. As high-level semantic cues, attributes are useful for generating image captions. Some attempts [3][10][11] discussed this task from different aspects, such as how to extract attributes and then fuse them with existing framework for better performances. Those works indicated that semantic information enhanced the description accuracy indeed.

In this paper, we propose an image captioning model which focuses on semantic salient regions of the given image through an attribute transferring mechanism. It obtains a weighted image feature, and then generates descriptions by a decoder. Figure 1 illustrates this process with an example. This network leverages the high-level semantic knowledge of image to enhance the fine visual details that may important to describe the whole image and scene context. The main contributions of this paper are summarized as follows: (1) Design a new weighting mechanism which transfers the predicted semantic attributes to the visual feature of image in order to pay more attention on semantic salient regions. (2) Re-weighting the attributes by word frequency of retrieved similar captions from training data, then obtain an attribute-based visual feature to replace the classic CNN feature for captioning, which captures more semantic contextual information. Experiments on image captioning dataset MSCOCO 2014 demonstrate that our method utilizes semantic attributes very well and achieves performance improvements over several popular methods on metrics of BLEU, CIDEr and METEOR.

## 2. RELATED WORK

Image captioning methods proposed in recent years can be divided into two main categories: Bottom-up and top-down

* The corresponding author is Chunna Tian.

This work was supported in part by the National Natural Science Foundation of China under Grants 61571354, 61671385, in part by China Postdoctoral Science Foundation under Grant 158201, in part by Natural Science Basis Research Plan in Shaanxi Province of China under No. 2017JM6001.

978-1-5386-4658-8/18/$31.00 ©2018 IEEE

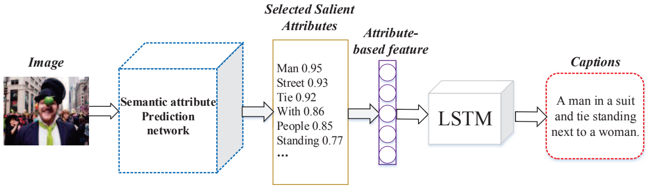


Image
Semantic attribute Prediction network
Selected Salient Attributes
Man 0.95
Street 0.93
Tie 0.92
With 0.86
People 0.85
Standing 0.77
...
Attribute-based feature
LSTM
Captions
A man in a suit and tie standing next to a woman.

**Fig. 1**. Overview of the attribute-based model.

methods. To further improve the description accuracy, some external information such as attention, attributes was leveraged. Our model combines the encode-decode captioning model with attributes.

The classic bottom-up approaches pose image captioning as a retrieval and ranking problem [12][13]. A template-based method [7] solved this problem by three sub-modules without Recurrent Neural Networks (RNNs): attributes detection, text descriptions generation by language model, sentences reranking. Similarly, Lebret et al.[6] proposed a phrase based network to achieve image captions. Aker et al.[14] extracted dependency patterns from the instance corpora of particular object types to automatically generate image descriptions.

Inspired by the substantial progress of machine translation [1], there are new top-down methods that use the encode-decode technique for image captioning. Vinyals et al.[2] and Karpathy et al.[15] used CNN features as the source language then generated captions by multi-modal RNNs (or LSTM). Mao et al.[16] proposed a multi-modal RNN (m-RNN) to predict the next word based on previously generated words and the deep CNN embedded features of an image. Xu et al.[5] proposed an attention mechanism that learned to fix gaze on salient objects and generated corresponding words in the output sequence. You et al.[4] proposed to extract semantic concepts from an image and selectively computed the attention on word candidates for captioning. Mun et al.[17] introduced an attention model which utilized text attention from similar captions to focus on related regions for details. Captioning is quite challenging because the currently image representation such as a 4096D vector extracted by CNN trained on ImageNet may not rich enough to represent the whole image context.

Attributes are more and more popular for enhancing the performance of CNN-RNN framework, some successes prove that this attempt is effective. The basic CNN-RNN approach inputs the visual vector to language generator, inspired by multi-label classification task, Wu et al.[3][10] demonstrated that high-level semantic information further improved the performance of image captioning. They used supervised learning to predict a set of attributes which were represented as image feature. Yao et al.[11] discussed the influence of semantic attributes on image captioning. They designed five different ways to combine semantic attributes and visual feature to enhance the LSTM model. The different input order of image representation and high-level attributes resulted in different results. They verified that input attributes into LSTM at each time step achieved best results. Gan et al.[18] added attributes probability to the weight matrix of LSTM in caption generation stage.

Unlike the commonly used ways of semantic attributes, which either pool the final prediction probability as a vector after softmax and then fed it to LSTM [3], or input the attributes to LSTM as additional weight information at each generation time step [4][18]. In this paper, we explore a different application of semantic attributes. Inspired by Wu et al.[3][10], we propose a novel transferring mechanism which weights the basic CNN visual feature in fc7 with high-level semantic attributes in the encode stage directly. Instead of extracting attributes from local regions then combining a probability vector [3], we predict attributes based on convolutional layers directly, so this new image embedding not only contains semantic information but also keeps more location information. In order to obtain attributes more effectively, a weakly supervised method MIL [19] is proper because it avoids the process of proposals extraction.

## 3. THE PROPOSED METHOD

In this paper, we extend the CNN-LSTM caption model by an attribute-based visual feature as Figure 2 shows. We predict high probability attributes of an image by MIL and transfer the results to weight the basic image representation, then translate the encoded attribute-based feature by LSTM model.

### 3.1. Transfer Attributes to Visual Feature

In this section, we detect attributes from an image and transfer the predicted results to image representation encoded by CNN, which pays more attention on regions that play important role in represent image context. Then, we use the new image feature that conditioned on semantic attributes to replace the classic CNN (such as GoogLeNet) feature to improve the accuracy of image captioning.

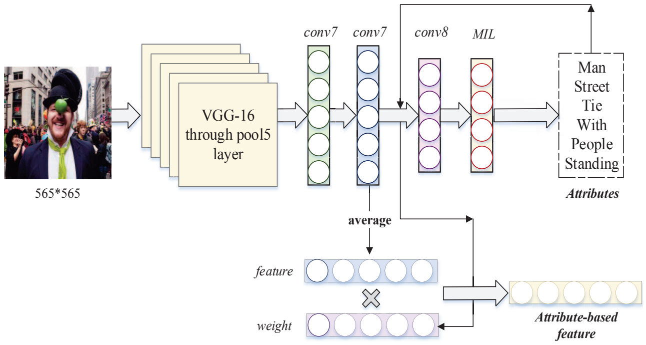


**Fig. 2**. The overview of extracting the attribute-based feature.

Before training, we collect 1000 most common words from the captions of training set to formulate an attribute vocabulary which covers about 92% words in all sentences. These 1000 words are the predefined categories of attribute detectors. However, we do not have image bounding boxes which correspond to words to train the attribute detectors. Besides, it is difficult to define clear corresponding regions in images for some words such as green, out or laying etc. So supervised learning techniques are impracticable for this task. Therefore, we follow the work of Lebret et al.[7] and use noisy-OR version of MIL [19] to train our attribute detectors.

We resize the input image into 565*565 pixels, then feed it into VGG-16 and replace the fully connected layers (fc6, fc7, fc8) by convolutions after pool5 layer. The fc8conv layer generates a spatial response map which corresponds to slide the original CNN over the whole image. The penultimate convolution layer fc7conv represents not only the image feature but also the location information of features in the input image. The response map consists of possible objects in the $j$-th regions of the $i$-th image. In order to predict final attributes, we run MIL layer on the response map to calculate a single probability $p_i^w$ from the probabilities of all regions in the given image:

$$p_i^w = 1 - \prod\nolimits_{j \in i} (1 - p_{ij}^w) \tag{1}$$

We threshold a precision value to output the top $N$ attributes $\{\mathrm{Att}^1, \mathrm{Att}^2, ..., \mathrm{Att}^N\}$ with higher probability rankings. In order to make visual attributes more reasonable, we also retrieve similar image captions [20] from the training dataset as additional semantic information. The attributes, which are frequently occur in the retrieved captions are chosen. The probability vector after thresholding is denoted as $\rho$ which is used as the importance-parameter to weight the image convolutional feature fc7conv. Consequently, the attribute-based feature is defined as Eq.(2).

$$\mathbf{I}_{att} = \rho \odot \mathbf{fc7conv} \tag{2}$$

In Eq.(2) $\odot$ represents dot multiplication. The final weighted feature $\mathbf{I}_{att}$ highlights regions with high semantic probabilities, which emphasizes the semantic context in representing the image. The feature $\mathbf{I}_{att}$ is fed to LSTM for caption generating.

### 3.2. Image Caption Generator

Recent successes in machine translation show that LSTM is more efficient than traditional RNN to decode a dimension-fixed feature into a target language. LSTM provides memory cells controlled by gates which transfer knowledge selectively at each time step according to previous results. There are three gates which determine current state whether to read input, output new value or forget current value.

The definition of gates and cell states are given as follows. All bias items have been omitted:

$$i_t = \sigma(\mathbf{W}_{ix}\mathrm{x}_t + \mathbf{W}_{ih}\mathrm{h}_{t-1}) \tag{3}$$

$$f_t = \sigma(\mathbf{W}_{fx}\mathrm{x}_t + \mathbf{W}_{fh}\mathrm{h}_{t-1}) \tag{4}$$

$$o_t = \sigma(\mathbf{W}_{ox}\mathrm{x}_t + \mathbf{W}_{oh}\mathrm{h}_{t-1}) \tag{5}$$

$$\mathrm{c}_t = f_t \odot \mathrm{c}_{t-1} + i_t \odot \tanh(\mathbf{W}_{cx}\mathrm{x}_t + \mathbf{W}_{ch}\mathrm{h}_{t-1}) \tag{6}$$

$$\mathrm{h}_t = o_t \odot \mathrm{c}_t \tag{7}$$

$$p_{t+1} = \mathrm{Softmax}(\mathrm{h}_t) \tag{8}$$

In above, $i_t, f_t, o_t$ are input gate, forget gate, and output gate, respectively. $\odot$ represents the product with a gate value, all $\mathbf{W}$ matrices are training parameters, the current cell state is fed to a softmax layer to produce a word probability distribution $p_{t+1}$.

Inspired by above methods, we follow the caption generator of Vinyals et al.[2], we input the extracted representation $\mathbf{I}_{att}$ to LSTM [21] and generate natural language captions. In detail, it generates one word at each time step and predict the next word conditioned on previous predictions. The decoder maximizing the probability of the correct description is formulated by Eq.(9)

$$\theta^* = \arg\max_{\theta} \sum\nolimits_{(\mathbf{I},\mathrm{S})} \log p(\mathrm{S}|\mathbf{I}_{att};\theta) \tag{9}$$

Where $\mathrm{S}$ is the corresponding ground-truth sentence and $\log p(\mathrm{S}|\mathbf{I}_{att})$ is the generation probability. It usually uses chain rule to model the joint probability of previously generated words as Eq.(10).

$$\log p(\mathrm{S}|\mathbf{I}_{att}) = \sum\nolimits_{t=0}^{N} \log p(\mathrm{S}_t|\mathbf{I}_{att}, \mathrm{S}_0, ..., \mathrm{S}_{t-1}) \tag{10}$$

## 4. EXPERIMENTS

### 4.1. Experimental Setup

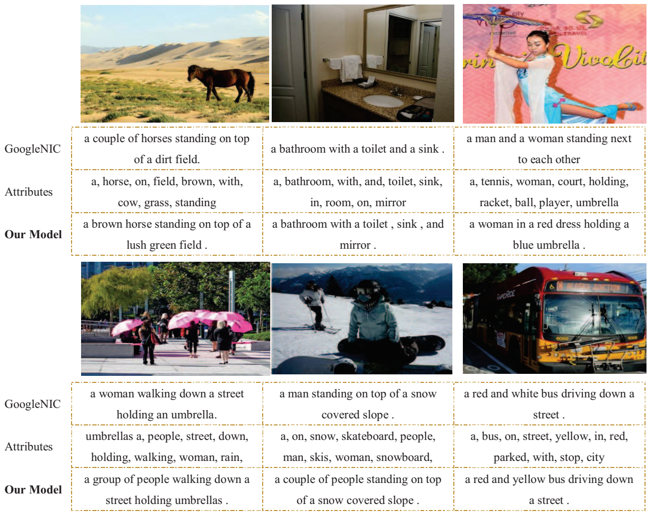
In order to validate the effectiveness of our semantic attribute-based caption model, we perform experiments on Microsoft COCO 2014 [22] dataset. This dataset contains 123,287 images in both training and validation sets. Each image has five captions annotated by Amazon Mechanical Turk. For comparison, we follow Vinyals et al.[2] to split these 123,278 images into three parts: training set, validation set and testing set. We reserve 10% of MSCOCO validation set (4000 images) randomly as the testing set.

In feature extracting part, the previous convolutional layers of VGG-16 network from conv1-1 to conv5-3 are held on, we replace fc6, fc7, and fc8 by three fully convolutional layers, and follow a MIL layer for attribute prediction. We select the attributes with probabilities higher than 70% to weight the 4096D vector of the fc7conv layer and output the attribute-based feature. We feed the extracted CNN feature to LSTM model with a 512 dimensional state vector from GoogleNIC network for caption generating. All these models trained on NVIDIA Titan Xp GPUs.

For evaluation, we adopt metrics: BLEU [23], CIDEr [24], and METEOR [25] as coco-caption [26]. BLEU is firstly used to measure the similarity between two sentences in machine translation tasks, which is defined as the geometric mean of n-gram (up to 4-gram) precision scores multiplied by a brevity penalty for short sentences. CIDEr is a specialized evaluation metric designed for image captioning. It measures the consensus between candidate image descriptions and the reference sentences. METEOR is defined as the harmonic mean of precision and recall of unigram matches between sentences.

| Model | B-1 | B-2 | B-3 | B-4 | M | C |
|---|---|---|---|---|---|---|
| NeuralTalk2 | 0.63 | 0.45 | 0.32 | 0.23 | 0.20 | 0.66 |
| LRCN | 0.63 | 0.44 | 0.30 | 0.21 | – | – |
| GoogleNIC | 0.67 | 0.45 | 0.30 | 0.20 | 0.24 | 0.86 |
| m-RNN | 0.67 | 0.49 | 0.35 | 0.25 | – | – |
| Soft-Att | 0.71 | 0.49 | 0.34 | 0.24 | 0.24 | – |
| G-LSTM | 0.67 | 0.49 | 0.36 | 0.26 | 0.23 | 0.81 |
| Hard-Att | **0.72** | 0.50 | 0.36 | 0.25 | 0.23 | – |
| **Our Model** | 0.70 | **0.53** | **0.39** | **0.30** | **0.25** | **0.90** |

**Table 1**. Attribute-based image captioning results on MSCOCO 2014 test split compared with other state-of-the-art methods. The best results are in bold and (–) indicates unknown scores.



| | | | |
|---|---|---|---|
| GoogleNIC | a couple of horses standing on top of a dirt field. | a bathroom with a toilet and a sink . | a man and a woman standing next to each other |
| Attributes | a, horse, on, field, brown, with, cow, grass, standing | a, bathroom, with, and, toilet, sink, in, room, on, mirror | a, tennis, woman, court, holding, racket, ball, player, umbrella |
| **Our Model** | a brown horse standing on top of a lush green field . | a bathroom with a toilet , sink , and mirror . | a woman in a red dress holding a blue umbrella . |

| | | | |
|---|---|---|---|
| GoogleNIC | a woman walking down a street holding an umbrella. | a man standing on top of a snow covered slope . | a red and white bus driving down a street . |
| Attributes | umbrellas a, people, street, down, holding, walking, woman, rain, | a, on, snow, skateboard, people, man, skis, woman, snowboard, | a, bus, on, street, yellow, in, red, parked, with, stop, city |
| **Our Model** | a group of people walking down a street holding umbrellas . | a couple of people standing on top of a snow covered slope . | a red and yellow bus driving down a street . |

**Fig. 3**. Example of placing a figure with experimental results.

### 4.2. Results and Analysis

Table 1 demonstrates the result comparison among our method and some state-of-the-art methods on MSCOCO 2014 dataset, where Our Model means the performance of the proposed attribute-based model. The comparing algorithms include encode-decode based architectures: NeuralTalk2 [15], LRCN [27], GoogleNIC [2], and m-RNN [16], and the attention based methods such as Guiding LSTM [28] and Soft/Hard Attention [5]. The results demonstrate that attention-based model usually outperforms the classic CNN/RNN architecture, such as NeuralTalk2 and GoogleNIC. Our model achieves almost the best performance on most metrics. It surpasses the baseline (GoogleNIC) by 4.5%, 17.8%, 30% for BLEU1, BLEU2, BLEU3. Especially for the more meaningful metrics: BLEU4 and CIDEr, our method improves at 50% and 4.7%, respectively. In order to find an appropriate threshold value for attribute prediction, we test various probabilities as the threshold. The threshold of 70% achieves the best results. Soft/Hard Attention model performances better than other models because of the attention mechanism. The attention model achieves the best BLEU1 score, because the attention focus more on salient image regions. However, our attribute model still has best results under most metrics, which shows the superiority of our visual attributes.

We show some captioning examples from the validation set in Figure 3. For better illustration, we also list the high probability attributes detected from testing images. We observe these words often obvious in images and help to describe the whole scene better. Compared with the baseline GoogleNIC, the main difference between our method and the baseline method is the CNN feature extractor. We replace the 2048D image feature obtained by GoogLeNet with our MIL-based VGG-16 network and obtain a 4096D attribute weighted image representation vector. Since the additional visual semantic information provides accurate object priors and semantic context, the captions of our attribute network always have more fine details, such as the class of objects, the number of objects, the color and the relationship of them. The results illustrate that semantic information brings much improvement on caption accuracy.

## 5. CONCLUSIONS

We propose a novel method for image captioning, which utilizes attributes to pay more attention on the regions with richer semantic information in a given image. We obtain a new feature weighted by mapping parameter between detection results and penultimate convolution layer as the encoded image information for generating sentences. Our model aims at embedding image more reasonably by highlighting the semantic details. The weighted semantic attribute feature results in more accurate captioning. It outperforms several state-of-the-art methods on MSCOCO 2014 dataset. In the short future, we plan to explore the visual attention area to bring more salient information.

## 6. REFERENCES

[1] I. Sutskever, O. Vinyals, and Q.V. Le, "Sequence to sequence learning with neural networks," in *Proc.NIPS*. MIT Press, 2003, pp. 3104–3112.

[2] O. Vinyals, A. Toshev, and S. Bengio, "Show and tell: A neural image caption generator," in *Proc.CVPR*. IEEE, 2015, pp. 3156–3164.

[3] Q. Wu, C. Shen, and L. Liu, "What value do explicit high level concepts have in vision to language problems?," in *Proc.CVPR*. IEEE, 2016, pp. 203–212.

[4] Q. You, H. Jin, and Z. Wang, "Image captioning with semantic attention," in *Proc.CVPR*. IEEE, 2016, pp. 4651–4659.

[5] K. Xu, J. Ba, and R. Kiros, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc.ICML*. ACM, 2015, pp. 2048–2057.

[6] R. Lebret, P.O. Pinheiro, and R. Collobert, "Simple image description generator via a linear phrase-based approach," in *Proc. ICLR Workshop*, 2015.

[7] H. Fang, S. Gupta, and F. Iandola, "From captions to visual concepts and back," in *Proc.CVPR*. IEEE, 2015, pp. 1473–1482.

[8] D.Elliott and F. Keller, "Image description using visual dependency representations," in *Proc.EMNLP*. ACL, 2013, pp. 1292–1302.

[9] Z. Li, J. Liu, and J. Tang, "Robust structured subspace learning for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, pp. 2085–2098, October 2015.

[10] Q. Wu, P. Wang, and C. Shen, "Ask me anything: Free-form visual question answering based on knowledge from external sources," in *Proc.CVPR*. IEEE, 2016, pp. 4622–4630.

[11] T. Yao, Y. Pan, and Y. Li, "Boosting image captioning with attributes," in *arXiv preprint*, 2016, arXiv:1611.01646.

[12] R. Socher, A. Karpathy, and Q.V. Le, "Grounded compositional semantics for finding and describing images with sentences," *Transactions of the Association of Computational Linguistic*, vol. 2, pp. 207–218, January 2014.

[13] 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. 853–899, 2013.

[14] A. Aker and R. Gaizauskas, "Generating image descriptions using dependency relational patterns," in *Proc.ACL*. ACL, 2017, pp. 5630 – 5639.

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

[16] J. Mao, W. Xu, and Y. Yang, "Deep captioning with multimodal recurrent neural networks (m-rnn)," in *Proc.ICLR*, 2015.

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

[18] Z. Gan, C. Gan, and X. He, "Semantic compositional networks for visual captioning," in *Proc.CVPR*. IEEE, 2017, pp. 5630 – 5639.

[19] C. Zhang, J.C. Platt, and P.A. Viola, "Multiple instance boosting for object detection," in *Proc.NIPS*. MIT Press, 2006, pp. 1417–1424.

[20] F. Faghri, D. J. Fleet, and J.R. Kiros, "Vse++: Improving visual-semantic embeddings with hard negatives," in *arXiv preprint*, 2017, arXiv:1707.05612v2.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, August 1997.

[22] T. Lin, M. Maire, and S. Belongie, "Microsoft coco: Common objects in context," in *Proc.ECCV*. Springer, 2014, pp. 740 – 755.

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

[24] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proc.CVPR*. IEEE, 2015, pp. 4566 – 4575.

[25] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," pp. 228–231, 2005.

[26] X. Chen, H. Fang, and T. Lin, "Microsoft coco captions: Data collection and evaluation server," in *arXiv preprint*, 2015, arXiv:1504.00325.

[27] J. Donahue, L.A. Hendricks, and S. Guadarrama, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc.CVPR*. IEEE, 2015, pp. 2625–2634.

[28] E. Gavves X. Jia and B. Fernando, "Guiding long-short term memory for image caption generation," in *arXiv preprint*, 2015, arXiv:1509.04942.