

Show, Observe and Tell: Attribute-driven Attention Model for Image Captioning*

Hui Chen^{†,ℓ}, Guiguang Ding^{†,ℓ}, Zijia Lin[‡], Sicheng Zhao^{†,ℓ}, Jungong Han[§]

[†]Beijing National Research Center for Information Science and Technology(BNRist)

^ℓSchool of Software, Tsinghua University, Beijing, China

[§]School of Computing & Communications, Lancaster University, UK

[‡]Microsoft Research, Beijing, China

{jichenhui2012,jungonghan77,schzhao}@gmail.com, dinggg@tsinghua.edu.cn, zijlin@microsoft.com

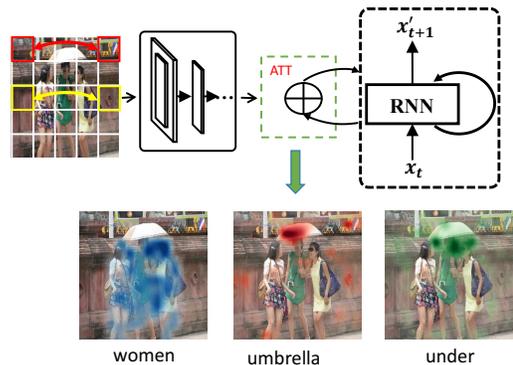
Abstract

Attribute-based approaches and attention-based approaches have been proven to be effective in image captioning. However, most attribute-based approaches simply predict attributes independently without taking the co-occurrence dependencies among attributes into account. Most attention-based captioning models directly leverage the feature map extracted from CNN, in which many features may be redundant in relation to the image content. In this paper, we propose an attribute-driven attention model for image captioning. We focus on training a good attribute-inference model via the recurrent neural network (RNN) for image captioning, where the co-occurrence dependencies among attributes can be maintained. The uniqueness of our inference model lies in the usage of a RNN with the visual attention mechanism to *observe* the image before generating captions. Additionally, it is noticed that compact and attribute-driven features will be more useful for the attention-based captioning model. Therefore, we extract the context feature for each attribute, and enable the captioning model to adaptively attend to these context features. We verify the effectiveness and superiority of the proposed approach over other captioning approaches by conducting massive experiments and comparisons on the MS COCO image captioning dataset.

1 Introduction

Image captioning enables machines to understand an image and generate a descriptive caption for it. It is challenging due to: 1) accurately recognizing all concepts attached to the given image, including objects, attributes, relationships, etc., is problematic; and 2) teaching machines to mimic how humans verbally describe an image does not seem easy. However, this

*This work was supported by the National Natural Science Foundation of China (No. 61571269), the National Natural Science Foundation of China (No. 61701273) and the Project Funded by China Postdoctoral Science Foundation (No. 2017M610897). Corresponding author: Guiguang Ding



Caption: There are women laughing under the umbrella.

Figure 1: **Top**: the proposed framework for predicting the attributes of the given image. We regard the attribute prediction as a sequential learning process and adopt the encoder-decoder framework to infer the attributes. **Bottom**: We obtain attribute information and its corresponding attention maps, which are used as the context features in the proposed approach. The attributes include object terms (*women*, *umbrella*), relational terms (*under*) and descriptive terms.

topic is of great significance to the ultimate goal of scene understanding, and thus attracts much attention from academia and industry.

So far, many pioneering works [Karpathy and Li, 2015; Mao *et al.*, 2015; Vinyals *et al.*, 2015; Jia *et al.*, 2015] have been proposed for image captioning. Early works usually adopted an encoder-decoder framework, where the information of an image is encoded into a static representation by a convolutional neural network (CNN), and a recurrent neural network (RNN), e.g., Long Short-Term Memory (LSTM), is employed as a decoder to interpret it into sentences.

In spite of great advances, these CNN-RNN captioning models exploit the word generation on the basis of the image representation directly, without explicitly taking more high-level semantic information from image into consideration [Yao *et al.*, 2017]. Later on, it has been verified that attributes with rich semantic cues about the image are effective in captioning [Yao *et al.*, 2017; Wu *et al.*, 2016; Liu *et al.*, 2017]. To integrate the attribute information into

the decoder RNN, they regarded the attribute prediction as a multiple single-label classification problem and extend CNN to predict attributes independently.

However, strong co-occurrence dependencies are common among attributes. For example, *sea* usually appears together with *wave*, but rarely co-occurs with *cars*. On the other hand, the attributes are composed of relational terms and descriptive terms, apart from the object terms. We claim that modeling the dependencies among attributes can facilitate the inference of relational terms. For example, in Fig. 1, object terms, *women* and *umbrella*, can help to recognize the relational term, *under*. Since the RNN can model the relationship among elements in a sequential manner by storing the dependencies in its internal memory, it is more natural and appropriate to adopt the RNN to infer the attributes, especially for the relational terms.

Besides, it is noticed that most attention-based captioning models directly attend to the feature map obtained from the CNN. The information of the feature map can be redundant or irrespective of the content of the image due to the uniform grid of equally sized and shaped receptive fields for feature map [Anderson *et al.*, 2017]. For example, for the image in Fig. 1, the regions of background (indicated by the red arrow) may be irrespective of the image content, and the features for regions resembling in each other (indicated by the yellow arrow), are redundant. Therefore, a more compact representation for images should be explored for attention, so that the attention results can more precisely capture the visual information related to the predicted word.

In this paper, we propose an attribute-driven attention model for image captioning. Firstly, instead of simply adopting a CNN to predict the attributes, we employ a CNN-RNN framework with the attention mechanism to predict attributes, where CNN provides feature representations and RNN acts as an inference module to predict the attributes. Secondly, we extract the context features corresponding to the attributes, as illustrated in Fig. 1, and incorporate them with the attribute information into the captioning model. In this way, the co-occurrence dependencies among attributes can be maintained in the memory of the RNN. Besides, the context features derived from the supervision of attributes are more semantically related to the image content, and thus are more compact than the feature map from CNN.

Overall, our contributions are three-fold:

- We model the co-occurrence dependencies among attributes by adopting a CNN-RNN framework and incorporating the visual attention mechanism for attribute detector.
- We filter the redundant or irrelevant features in relation to the image content by extracting the context features for attributes, and integrate them with the attribute information into the captioning model.
- We show the effectiveness and superiority of the proposed approach by conducting a massive of experiments and comparisons with other approaches on the MS COCO image captioning dataset.

2 Related Work

There have been many pioneering neural network-based works for image captioning. The early neural network-based works [Mao *et al.*, 2015; Vinyals *et al.*, 2015] aim to explore the way of aligning the visual feature and the semantic feature. They simply compressed the image information into a single and static representation, which can lead to losing information that could be useful for richer, more descriptive captions [Xu *et al.*, 2015]. Besides, the feature representations are extracted from a CNN directly and thus lack more high-level semantic information.

Attribute-based approaches. Attribute-based approaches aim at boosting the performance of a captioning model with the more high-level semantic information. You *et al.* [2016] proposed a semantic attention model to selectively attend to the most related attributes during caption generation. Liu *et al.* [2017] treated the attribute information as a semantic regularisation for the captioning model. Yao *et al.* [2017] regarded the attributes as the auxiliary information and explored different ways to incorporate them into captioning. These works achieved inspiring results by incorporating the attribute information into the captioning model. However, they simply predict the attributes with a CNN, without considering the co-occurrence dependencies among attributes. In this paper, we aim at modeling the co-occurrence among attributes via employing both the CNN and the RNN.

Visual attention models. Inspired by the presence of attention in the human visual system, Xu *et al.* [2015] firstly proposed soft attention and hard attention to make the decoder exposed to different aspects of image information at each time step. Chen *et al.* [2017a] thought that the attention should be spatial and channel-wise, and proposed SCA-CNN model to selectively attend to both salient regions and salient semantic patterns. Anderson *et al.* [2017] pre-trained an object detection model on another dataset, and used it to obtain image feature at concept level for attention. Our approach feeds the attention model with the context features corresponding to the predicted attributes, which is similar to the idea in [Anderson *et al.*, 2017]. But unlike [Anderson *et al.*, 2017], where an object detection model is pre-trained on a large dense captioning dataset with precise position information about the concepts in images, our model does not need a pre-trained object detection model and is trained on image captioning dataset without assistant datasets. Please note that we can also leverage other large datasets to boost the performance of the attribute prediction, which is out of the scope of this paper and is left as our future work.

Multi-label classification. Another related work is the multi-label classification domain. Our attribute prediction model is similar to [Wang *et al.*, 2016; Liu *et al.*, 2017]. But their works are built on the standard datasets for multi-label classification task, which means that labels of images for training are annotated by humans. While in our work, the attributes are obtained from the captions corresponding to images, which consist of not only object terms and descriptive terms but also relational terms, and thus are more diverse and more in quantity.

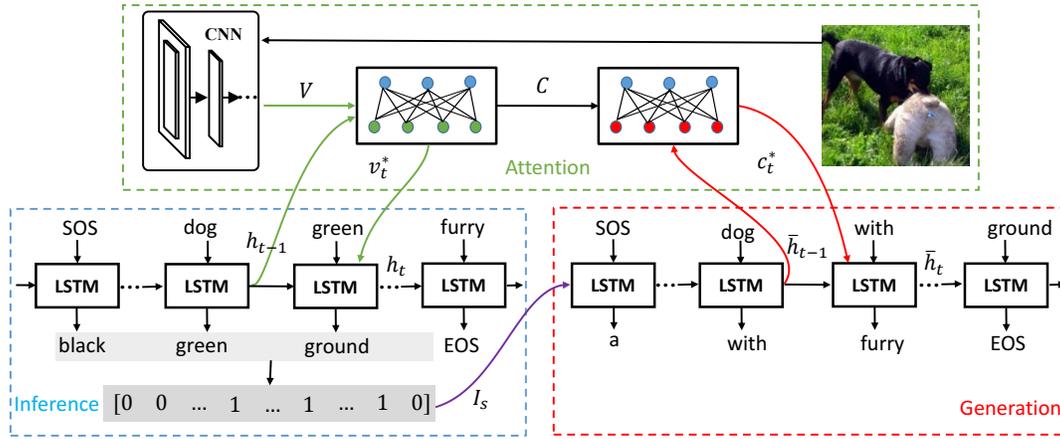


Figure 2: The framework of the proposed model, including three components: the inference module (in blue box), the attention module (in green box) and the generation module (in red box). ‘‘SOS’’ and ‘‘EOS’’ indicate the start and the end of the sequence, respectively.

3 Methodology

In this section, we introduce the whole framework of the proposed approach in detail. As illustrated in Fig. 2, our model is composed of three components: the inference module, the attention module and the generation module. The inference module aims to predict the attributes and produce their observed context features in a sequential manner. The generation module attempts to generate the sentence word by word, on the basis of the image information. And the attention module interacts with the inference module and the generation module and provides different types of features, e.g., v_t^* and c_t^* for them.

3.1 Encoder-Decoder Framework

We start by briefly introducing the encoder-decoder framework. Given an image I and a sequence $X = \{x_0, x_1, x_2, \dots, x_T\}$ which depends on the task background, e.g., attributes for multi-label classification, the encoder-decoder framework aims to predict the next element x_{t+1} conditioned on the image I and the partial sequence $\{x_0, x_1, \dots, x_t\}$ that has been generated. Generally, a CNN is employed as the encoder to extract the features of the given image, and a RNN acts as the decoder to generate the sequence conditioned on the image features. It directly maximizes the following objective:

$$\theta^* = \arg \max_{\theta} \sum_{t=0}^T \log p(x_{t+1} | I, x_0, x_1, \dots, x_t) \quad (1)$$

where the conditional probability is modeled by the decoder RNN: $\text{RNN}(z_t, x_t, h_{t-1}, m_{t-1})$, where h_{t-1} and m_{t-1} are the hidden state vector and memory cell vector of RNN at time step $t-1$, respectively. z_t is the auxiliary knowledge, like, the global image feature obtained from the last fully connected layer in CNN, or the context feature of the image from the output of the attention module.

Here we formulate the RNN function as a variant of LSTM as in [Rennie *et al.*, 2017].

3.2 Attention Module

The attention module plays a core role in the proposed approach. At different time steps, it captures the most related information of the image to the next prediction of the generated words. Our attention module can provide two types of image features, the region-based features for the inference module and the attribute-based ones for the generation module, respectively.

Region-based feature. At time step t , given the feature map from the CNN, i.e., $V = \{v_i | i = 0, 1, 2, \dots, k, v_i \in \mathbb{R}^D\}$ where each subscript i denotes the i -th image region, our attention module is able to adaptively attend to the most relevant image regions to the next prediction under the guidance of the hidden state h_{t-1} of the RNN in the inference module.

$$\alpha_t = \text{softmax}(W_a \tanh(W_{av}V + (W_{ah}h_{t-1})\mathbf{1}^T))$$

$$v_t^* = \sum_{i=1}^k \alpha_t^i v_i \quad (2)$$

where W_a , W_{av} and W_{ah} are parameters to be learned. We define v_t^* as the *context* feature corresponding to the attribute which contains the visual information of those image regions related to the attribute.

Attribute-based feature. The context feature v_t^* is more semantically related to the image content. We extract a series of context features corresponding to attributes, and when generating the next word in the generation module, the attribute-based feature is derived by adaptively attending to these context features. For the ease of explanation, we substitute v_t^* with c_i and use $C = \{c_0, \dots, c_t, \dots, c_l\}$, i.e. $c_t = v_t^*$, to represent the context feature map corresponding to the attributes, where l is the number of attributes. So the attribute-based feature is obtained as follows:

$$\beta_t = \text{softmax}(W_b \tanh(W_{bc}C + (W_{bh}\bar{h}_{t-1})\mathbf{1}^T))$$

$$c_t^* = \sum_{j=1}^l \beta_t^j c_j \quad (3)$$

where W_b , W_{bc} and W_{bh} are parameters to be learned. \bar{h}_{t-1}

is the hidden state of the RNN in the generation module at time step $t - 1$.

Our attention model can be considered as a stack attention model. The first layer of attention can be regarded as a filter, merging regions and compressing the redundant or irrelevant features, and thus the context feature corresponding to the attribute is more compact than the feature map in the CNN. The second layer of attention operates the standard attention mechanism on those compact features from the the first layer and thus can be more semantically related to the next word to be predicted.

3.3 Inference Module

The inference module aims to predict attributes for the given images, following the encoder-decoder framework. First, the information of images and attributes will be projected to the same embedding space to capture the image-text relationship. Then a RNN is adopted to take in the embedding vectors in the embedding space at each time step and maintain the label co-occurrence information in its internal memory.

Specifically, at time step t , the information of the image, v_t^* , i.e., the region-based feature described in Sec. 3.2, will be mapped into the embedding space through a non-linear mapping function:

$$\hat{v}_t = \sigma(W_v v_t^* + b_v) \quad (4)$$

where W_v and b_v are the parameters to be learned. $\sigma()$ is an activation function. \hat{v}_t is the embedding feature of v_t^* .

Then, a recurrent layer of RNN is employed to seize the co-occurrence dependencies among attributes. Its hidden state h_t is modeled as:

$$h_t, m_t = \text{RNN}(\hat{v}_t, w_t, h_{t-1}, m_{t-1}) \quad (5)$$

where m_{t-1} is the memory cell vector at time step $t - 1$ and w_t is the embedding feature of the attribute at time step t .

Finally, an inference layer is built on top of the recurrent layer by considering the current hidden state h_t and generating the distribution of the next attribute to be predicted via a softmax function:

$$\phi_t = \text{softmax}(W_{dh} h_t + b_{dh}) \quad (6)$$

where W_{dh} and b_{dh} are the parameters to be learned. b_{dh} is the bias term.

Attribute prediction. During prediction, we generate the attribute one by one under a greedy-decoding strategy according to the distribution ϕ_t at each time steps.

3.4 Generation Module

The generation module aims to generate a sentence word by word, given the features of images. To leverage the attribute information, we denote the representation of attributes obtained from the inference module with a binary (0 or 1) attribute vector I_s where 1 means that the image has the corresponding attribute, and 0 means not.

Since the attribute vector I_s is rich in semantic information, we use it to initialize the hidden state of RNN as in [Liu *et al.*, 2017]:

$$h_{\text{init}} = I_s \quad (7)$$

The idea behind this is that we force the RNN to understand the image in the beginning so that it could properly select or leave out some aspects about the image along its decoding process.

We make the generation module perceive the information of images by means of the attention mechanism. Specifically, at each time step t , a attribute-based feature vector, c_t^* , is extracted in the attention module (see Sec. 3.2). c_t^* will be further projected into the embedding space:

$$\hat{c}_t = \sigma(W_c c_t^* + b_c) \quad (8)$$

where W_c and b_c are the parameters to be learned. $\sigma()$ is an activation function. Then the RNN generates the next word by conditioning on the embedding feature \hat{c}_t and the representation of word y_t :

$$\begin{aligned} \bar{h}_t, \bar{m}_t &= \text{RNN}(\hat{c}_t, y_t, \bar{h}_{t-1}, \bar{m}_{t-1}) \\ \bar{\phi}_{t+1} &\sim \text{softmax}(\bar{h}_t) \end{aligned} \quad (9)$$

where \bar{h}_{t-1} and \bar{m}_{t-1} are the hidden state vector and the memory cell vector of the decoder RNN at time step $t - 1$, respectively. $\bar{\phi}_{t+1}$ is the distribution of the next word given the past information. y_t is the embedding feature of the current word via mapping the one-hot representation of the word into the word embedding space.

In the proposed approach, the attribute vector I_s is integrated into the decoder as a semantic regularisation, which is similar to [Liu *et al.*, 2017]. But different from [Liu *et al.*, 2017], the regularisation vector in our approach is represented as a hard-code way, where the element is either zero or one, instead of a soft-code way, where numbers ranging from zero to one depict the likelihood of the corresponding attribute to be related to the image. Besides, we also extract the context features for attributes and incorporate them into the captioning decoder by the attention mechanism.

Sentence generation. During generation, at time step t , we plug the word sampled from the distribution at last time step, $\bar{\phi}_{t-1}$, into the RNN. We employ the beam search strategy to boost the performance.

4 Experiment

4.1 Dataset and Setting

Dataset. Following previous works [Yao *et al.*, 2017; Yang *et al.*, 2016], we conduct experiments on the popular MS COCO dataset [Lin *et al.*, 2014], which consists of 82783 training images and 40504 validation images. All images are labeled with at least 5 captions by human labellers. It provides 40775 images as a test set for online evaluation as well. For offline evaluation, we follow most previous works [Chen *et al.*, 2017a; Yao *et al.*, 2017] and split the 123287 images into three parts, 5000 for validation, 5000 for test and the remains for training.

Evaluation metrics. To compare with other methods, we use the same evaluation metrics, including BLEU [Papineni *et al.*, 2002], ROUGE-L [Lin and Hovy, 2003] and CIDEr [Vedantam *et al.*, 2015]. Meanwhile, we use the MS COCO caption evaluation tool¹ to compute the scores.

¹<https://github.com/tylin/coco-caption>

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr
NIC [Vinyals <i>et al.</i> , 2015]	66.6	45.1	30.4	20.3	-	-
HA [Xu <i>et al.</i> , 2015]	71.8	50.4	35.7	25.0	-	-
SA [Xu <i>et al.</i> , 2015]	70.7	49.2	34.4	24.3	-	-
SCA-CNN [Chen <i>et al.</i> , 2017a]	71.9	54.8	41.1	31.1	53.1	95.2
ATT [You <i>et al.</i> , 2016]	70.9	53.7	40.2	30.4	-	-
SCN-LSTM [Gan <i>et al.</i> , 2017]	72.8	56.6	43.3	33.0	-	1.012
MSM [Yao <i>et al.</i> , 2017]	73.0	56.5	42.9	32.5	53.8	98.6
Ours	74.3	57.9	44.3	33.8	54.9	104.4

Table 1: Performance of the proposed approach and other approaches on the MS COCO dataset. All values are reported as percentage(%).

	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr	
	c5	c40	c5	c40										
Google NIC† [Vinyals <i>et al.</i> , 2015]	71.3	89.5	54.2	80.2	40.7	69.4	30.9	58.7	25.4	34.6	53.0	68.2	94.3	94.6
ATT† [You <i>et al.</i> , 2016]	73.1	90.0	56.5	81.5	42.4	70.9	31.6	59.9	25.0	33.5	53.5	68.2	94.3	95.8
ERD [Wu and Cohen, 2016]	72.0	90.0	55.0	81.2	41.4	70.5	31.3	59.7	25.6	34.7	53.3	68.6	96.5	96.9
SCA-CNN [Chen <i>et al.</i> , 2017a]	71.2	89.4	54.2	80.2	40.4	69.1	30.2	57.9	24.4	33.1	52.4	67.4	91.2	92.1
Adaptive Attention† [Lu <i>et al.</i> , 2017]	74.6	91.8	58.2	84.2	44.3	74.0	33.5	63.3	26.4	35.9	55.0	70.6	103.7	105.1
SCN-LSTM† [Gan <i>et al.</i> , 2017]	74.0	91.7	57.5	83.9	43.6	73.9	33.1	63.1	25.7	34.8	54.3	69.6	100.3	101.3
MSM† [Yao <i>et al.</i> , 2017]	73.9	91.9	57.5	84.2	43.6	74.0	33.0	63.2	25.6	35.0	54.2	70.0	98.4	100.3
R-LSTM [Chen <i>et al.</i> , 2017b]	75.1	91.3	58.3	83.3	43.6	72.7	32.3	61.6	25.1	33.6	54.1	68.8	96.9	98.8
Ours	78.7	93.5	61.5	85.5	46.5	74.8	34.5	63.3	25.9	34.2	55.5	69.9	106.1	108.7

Table 2: Evaluation performance of the proposed approach on the *online* MS COCO testing server. All values are reported as percentage(%). † indicates the results of ensemble models.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr
LSTM-A5 + MIL-IAC [Yao <i>et al.</i> , 2017]	73.4	56.7	43.0	32.6	54.0	100.2
LSTM-A5 + RNN(ours)	73.6	56.8	43.2	32.9	54.2	102.8
Attend to regions	73.0	56.4	42.7	32.7	54.1	101.2
Attend to attributes	73.7	57.2	43.6	33.4	54.4	102.8
Full	74.3	57.9	44.3	33.8	54.9	104.4

Table 3: Ablation study of the proposed approach on the MS COCO dataset. All values are reported as percentage(%).

Compared approaches. We compare the proposed approach with the following state-of-the-art approaches. (1) NIC [Vinyals *et al.*, 2015]: a standard neural network based approach which only injects the image into RNN at the initial time step. (2) HA and SA [Xu *et al.*, 2015]: it incorporates the spatial feature map into the decoder via attention mechanisms. HA incorporates the feature map into the decoder by adopting a *hard* way, and SA adopts a *soft* way. (3) SCA-CNN [Chen *et al.*, 2017a]: SCA-CNN adopts a spatial and channel-wise attention for captioning. (4) ATT [You *et al.*, 2016]: ATT firstly predicts the attributes for images with CNN, and then adopts the semantic attention to adaptively select the attribute features in the decoder for captioning. (5)SCN-LSTM [Gan *et al.*, 2017]: SCN-LSTM proposed a semantic compositional network by leveraging the attributes. (6)MSM [Yao *et al.*, 2017]: MSM explores different ways to integrate the attribute information into the RNN.

4.2 Implementation Details

We convert all sentences to the lower case, and filter rare words that occur less than 5 times, and we end up with a vocabulary set of 9487 tokens. We use ResNet-101 [He *et al.*, 2016] pre-trained on the ImageNet dataset to extract the image features. We do not crop or scale any image. In-

stead, we use the final convolutional layer of ResNet as image features, and apply spatially average pooling, resulting in a fixed size of $14 \times 14 \times 2048$ of the feature map. The hidden state size of LSTM, the embedding dimension of the input word and the embedding dimension of image features are all fixed to 1000. During training, the parameters are updated by ADAM optimizer with a learning rate of 5×10^{-4} and 0.9 as the learning rate decay factor. We let the learning rate decay every 2 epoches and we train the model for 30 epoches with a batch size of 16. Following [Yao *et al.*, 2017; Fang *et al.*, 2015], we use the 1000 most frequent words in the training captions as the attribute vocabulary, which cover the majority of words in the training data. To train the inference module, for each image we rank the attributes according to their frequency. For caption generation in the testing stage, we apply the beam search algorithm to boost the performance by default. The beam size is empirically set to 3.

4.3 Evaluation

Compared with other approaches. The comparisons are shown in Tab. 1. The proposed model defeats NIC model with a large margin, which shows that the attribute information is effective and can provide helpful high-level information to enhance the caption generation performance. Com-

Image	Attributes	Generated caption	Ground truth
	sitting, table, small, front, dog, food, wooden , eating, floor , piece, dish	A small dog standing on the floor with a plate of food.	1) Shaggy dog gets dinner served on a plate. 2) A small black dog standing over a plate of food. 3) A small dog eating a plate of broccoli. 4) A black dog being given broccoli to eat. 5) There is a dog staring at a plate of food.
	street, large, city, road, trees, lights , traffic , night, scene, bright	A city street at night with traffic lights .	1) It is night time and the town is quiet. 2) A nightlife scene at the park in the dark 3) A long exposure image of a street during the night. 4) A street is displayed at night with time lapse photography. 5) There is a street at night with cars passing by.
	sitting, white, cat, bathroom, laying, sink, lying	A cat laying in a bathroom sink.	1) A cute cat laying down in a sink. 2) A cat laying inside of a sink under a fixture. 3) A grey and white cat lays in a sink. 4) a cat sitting in the sink in the bathroom 5) A striped cat is laying on the sink and looking at the camera.
	man, standing, kitchen, food, sink, counter, using, preparing , lady, meal	A man standing in a kitchen preparing food.	1) A man appears to be making something in his kitchen. 2) a person in the kitchen using a mixer in a cup 3) A woman near a messy kitchen counter holds a hand blender into a green drink. 4) A picture of a person cooking some food. 5) a man using a hand blender in a kitchen

Figure 3: Examples of attributes and captions on COCO. The attributes are detected by our inference module and the captions are generated by the proposed generation model. The word in green are the terms not in the ground truth but highly related to the image.

pared with the visual attention based models, e.g., HA, SA and SCA-CNN, our model also has advantages in all metrics. Compared to SCN-LSTM, MSM and ATT models, which also adopt the attribute as high-level image features, the performance of our model is superior to theirs. We also train our model with reinforcement learning [Rennie *et al.*, 2017], and submit the result of the official test set to the test server. The compared results are shown in Tab. 2. Our model can outperform other approaches in most metrics, even though we does not utilize the ensemble approach. We left the ensemble approach as a future work. The comparisons with the state-of-the-art approaches demonstrate the effectiveness and superiority of the proposed approach.

Ablation study. Ablation study is to verify the effectiveness of each component of our proposed approach. The results are shown in Tab. 3. First, to prove the effectiveness and superiority of the proposed approach inferring attributes via RNN, we adopt the same architecture of the captioning model as the best model in [Yao *et al.*, 2017], and compare the performance with that reported in their paper. Our model can obtain a better performance than [Yao *et al.*, 2017] (see the second row), which shows that the proposed approach to predict attributes is effective. Second, to reveal the effectiveness of the proposed attention module which provides the context features corresponding to attributes for the generation module, we train a captioning model with the standard visual attention mechanism directly attending to the feature map from CNN, indicated as “Attend to regions”. And we compare it with our generation model without the semantic regularisation, indicated as “Attend to attributes” (see the third row). The result shows that the context features we propose for captioning model is effective and can potentially replace the original feature map from CNN. The performance improvement can be attributed to that the features are derived from the supervision of the semantic information and thus are more semantically powerful than features directly extracted from

CNN. By incorporating the attributes as semantic regularisation and by attending to the context features derived from the inference module, the performance of our full model can be improved further (see the fourth row).

4.4 Qualitative Analysis

Fig. 3 presents a few examples generated by our model. Our inference module can detect various kinds of attributes successfully, including object terms, relational terms and descriptive terms. Since the true attributes used to train the inference model are collected from the ground truth captions, they may not completely describe the content of the image. But our inference model can still produce attributes outside the ground truth but highly related to the content of the image. For example, for the image in the second row, the attributes, *traffic* and *lights*, predicted by our inference module, are not in the ground truth, but highly related to the image. Moreover, these attributes can assist the captioning model to generate a descriptive sentence, as illustrated in the examples.

5 Conclusion

In this paper, we proposed an attribute-driven attention model for image captioning. We trained an attribute inference module by utilizing a CNN-RNN framework to model the co-occurrence dependencies among attributes. Different from other attribute-based approaches, we incorporate the attribute information and their corresponding context features into the decoder for sentence generation. The context features corresponding to attributes contain rich semantic information and thus are more compact for representing images than the feature map in the CNN. To verify the effectiveness of the proposed approach, we conducted experiments on MS COCO, a popular dataset for image captioning. We also compared the proposed model with other state-of-the-art captioning models. The results well demonstrated the effectiveness and superiority of the proposed approach.

References

- [Anderson *et al.*, 2017] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and vqa. *arXiv preprint arXiv:1707.07998*, 2017.
- [Chen *et al.*, 2017a] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. *CVPR*, 2017.
- [Chen *et al.*, 2017b] Minghai Chen, Guiguang Ding, Sicheng Zhao, Hui Chen, Qiang Liu, and Jungong Han. Reference based lstm for image captioning. *AAAI*, 2017.
- [Fang *et al.*, 2015] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *CVPR*, pages 1473–1482, 2015.
- [Gan *et al.*, 2017] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *CVPR*, 2017.
- [He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 00:770–778, 2016.
- [Jia *et al.*, 2015] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. Guiding the long-short term memory model for image caption generation. In *IEEE International Conference on Computer Vision*, pages 2407–2415, 2015.
- [Karpathy and Li, 2015] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.
- [Lin and Hovy, 2003] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 71–78. Association for Computational Linguistics, 2003.
- [Lin *et al.*, 2014] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014.
- [Liu *et al.*, 2017] F. Liu, T. Xiang, Timothy Hospedales, W. Yang, and C. Sun. Semantic regularisation for recurrent image annotation. In *Computer Vision and Pattern Recognition (CVPR 2017)*, 2 2017.
- [Lu *et al.*, 2017] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *CVPR*, 2017.
- [Mao *et al.*, 2015] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *International Conference on Learning Representations*, 2015.
- [Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.
- [Rennie *et al.*, 2017] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. *CVPR*, 2017.
- [Vedantam *et al.*, 2015] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.
- [Vinyals *et al.*, 2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015.
- [Wang *et al.*, 2016] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2016.
- [Wu and Cohen, 2016] Zhilin Yang, Ye Yuan, Yuexin Wu, and Ruslan Salakhutdinov, William W Cohen. Encode, review, and decode: Reviewer module for caption generation. *NIPS*, 2016.
- [Wu *et al.*, 2016] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton van den Hengel. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 203–212, 2016.
- [Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.
- [Yang *et al.*, 2016] Zhilin Yang, Ye Yuan, Yuexin Wu, Ruslan Salakhutdinov, and William W. Cohen. Encode, review, and decode: Reviewer module for caption generation. *NIPS*, 2016.
- [Yao *et al.*, 2017] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *ICCV*, 2017.
- [You *et al.*, 2016] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659, 2016.