# Deliberate Attention Networks for Image Captioning

**Lianli Gao,[1] Kaixuan Fan,[1] Jingkuan Song,[1] Xianglong Liu,[2] Xing Xu,[1] Heng Tao Shen[1]***
[1]Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China. [2]Beihang University, China. {lianli.gao,201722060722,xing.xu}@uestc.edu.cn, {jingkuan.song}@gmail.com, xlliu@nlsde.buaa.edu.cn,shenhengtao@hotmail.com

## Abstract

In daily life, deliberation is a common behavior for human to improve or refine their work (e.g., writing, reading and drawing). To date, encoder-decoder framework with attention mechanisms has achieved great progress for image captioning. However, such framework is in essential an one-pass forward process while encoding to hidden states and attending to visual features, but lacks of the deliberation action. The learned hidden states and visual attention are directly used to predict the final captions without further polishing. In this paper, we present a novel Deliberate Residual Attention Network, namely DA, for image captioning. The first-pass residual-based attention layer prepares the hidden states and visual attention for generating a preliminary version of the captions, while the second-pass deliberate residual-based attention layer refines them. Since the second-pass is based on the rough global features captured by the hidden layer and visual attention in the first-pass, our DA has the potential to generate better sentences. We further equip our DA with discriminative loss and reinforcement learning to disambiguate image/caption pairs and reduce exposure bias. Our model improves the state-of-the-arts on the MSCOCO dataset and reaches 37.5% BELU-4, 28.5% METEOR and 125.6% CIDEr. It also outperforms the-state-of-the-arts from 25.1% BLEU-4, 20.4% METEOR and 53.1% CIDEr to 29.4% BLEU-4, 23.0% METEOR and 66.6% on the Flickr30K dataset.

## Introduction

Image captioning is to automatically generate a natural language sentence given an image. The sentence is required to be fluent and describe objects and scene in the image. Image captioning can facilitate lots of practical applications. For example, visually impaired people can get help from image captioning to understand the content of image. To date, encoder-decoder framework with attention mechanisms has achieved great progress for image captioning. Generally, CNN extracts image features as encoder and RNN predicts words as decoder. When each word is generated, attention mechanism (Anderson et al. 2018; Lu et al. 2017; Song et al. 2018b; Gao et al. 2018) decides which regions to attend to.

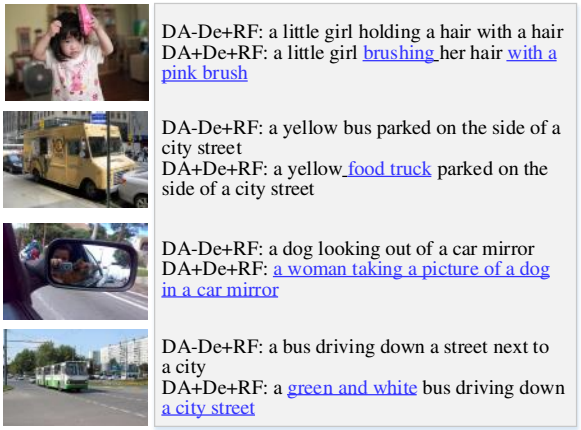


Figure 1: Examples of image captions generated by our DA. For each image, the first and second sentences are generated by the first-pass (DA-De+RF) and the second-pass respectively (DA+De+RF). This demonstrates that our deliberate attention can generate more precise concepts (e.g., "food truck" vs "bus") and more reasonable descriptions (e.g., "a women taking a picture of a dog in a car mirror" vs "a dog looking out of a car mirror") by using a deliberate process.

Besides, the encoder-decoder framework with attention mechanisms has been widely applied to solve other sequence generation tasks, e.g., machine translation (Denkowski and Lavie 2014), image annotation (Krishna et al. 2017) and video captioning (Song et al. 2018a; Zhang et al. 2019; Song et al. 2017). However, such a framework is in essential an one-pass forward process. When a model predicts the next word, it can only leverage the generated words but not the future unknown words. To humans, deliberation action is a common behavior in their daily, e.g., reading, writing or understanding an image. During the process, global information of both the past and future are leveraged. Xia *et. al.* (Xia et al. 2017) designed a deliberation network which included two levels of decoders, and it is proved to be effective for neural machine translation. The first one generates a coarse sentence and corresponding hidden states. The

*Jingkuan Song and Heng Tao Shen are corresponding authors.

second decoder refines the sentence with deliberation. In the network, the second decoder could leverage the global information of both the past and future parts.

Another major concern of encoder-decoder framework is the exposure bias. During the training process, the goal of RNN is to generate a target sequence word by word, given source sequence. When testing, RNN is used to predict words, and ground truth sequence is not available. So we have to predict the next word conditioned on generated word. But small errors on generated words may be enlarged through information flow, and result in a terrible sequence. To address this issue, Bengio *et al.* (Bengio et al. 2015) proposed a new learning process. They designed a sampling mechanism to decide whether they use a generated word or word in ground truth during training. Goyal *et al.* (Goyal et al. 2016) introduced GAN to tackle this issue, and applied on the captioning tasks. They forced hidden states distributions produced in training and sampling process to be close to each other. Both approaches aimed to bridge the gap between training and testing.

More recently, reinforcement learning has been widely used in image captioning (Rennie et al. 2017). Policy-gradient methods for reinforcement learning are proved to be suitable for training captioning models. Generally, people pretrain the captioning model using MLE, and continue training model with policy-gradient. Reward is important to improve the performance of system trained with policy-gradient. In (Rennie et al. 2017), CIDEr score is used as a reward to guide the training. Many works (Dai et al. 2017; Dognin et al. 2018) also introduce GAN to image captioning. Generally speaking, captions are required to be fluent and informative. However, discrimination of image captioning has drawn remarkable attention recently. It requires a model to generate discriminative captions for similar images. Image captioning can be seen as mapping image to sentence. Similarly, retrieval is a task to compute a map between image and sentence. Therefore, in (Luo et al. 2018; Liu et al. 2018), they introduced retrieval model into image captioning model. Retrieval model forced image captioning model to generate discriminative sentences.

To address the above concerns of encoder-decoder framework for image captioning, and inspired by the success of reinforcement learning, in this paper, we introduce a novel architecture, Deliberate Residual Attention Network for image captioning. Our model consists of three parts: two residual based attention layers and a reinforcement module. In the first attention layer, the combination of attention based LSTM and residual shortcuts produced hidden states and attended image features, which can be used to generate the raw image captions. In the second deliberate residual-based attention layer, we refine the previous hidden states and attended visual features for generating better sentences. Since the second-pass is based on the rough global features captured by the hidden layer and visual attention in the first-pass, our DA relieves the exposure bias to some extent, and has potential to generate better sentences. Some experimental results shown in Fig. 1 clearly illustrates the effect of our deliberate residual-based attention layer. To further generate discriminative sentences, we introduce reinforcement learning to guide the training process of captioning model. We evaluate our model on two dataset, COCO and Flickr30K, and compare our approach with state-of-the-art methods. Experimental results show that our model achieves the state-of-the-art performance.

# Methodology

The overall architecture of the DA net is shown in Fig.2. We begin by introducing the image encoding. Then, we describe deliberate attention mechanism and reinforcement module.

## Image Representation

Image encoder is an essential part of image captioning and it is used to extract visual content information of images. CNN designed for image classification is usually adopted to extract a global visual image feature, while R-CNN designed for object detection is usually employed to derive region visual features. Compared with a CNN-based visual feature which is obtained within a global context, a R-CNN based region visual feature contains rich information about a particular object. In this paper, we adopt a pretrained ResNet-101 (He et al. 2016) (pool5) to extract global visual feature, and use Faster R-CNN (Ren et al. 2015) to produce bounding boxes and then apply them on the pretrained ResNet-101 to extract $L$ region features. For simplicity, given an image $I$, we extract $L+1$ image features, represented as $\{v_0, V\}$, where $v_0$ is the global visual feature and $V = \{v_1, \cdots, v_L\}$ indicates the $L$ region visual features.

## Deliberate Attentions Mechanism

Our deliberate attentions mechanism consists of two residual based attention layers. The first layer aims to prepare the hidden states and visual attention for generating a preliminary version of the captions, while the second layer is applied as a proofreading process to refine them. Both layers are built upon the basic LSTM. For the $t$-th time step, where $x_t$ is the input vector of the LSTM unit, $h_t$ is the output vector of the LSTM unit, and $h_{t-1}$ is the output of the LSTM unit at the $t-1$ time step. For simplicity, we refer to the operation procedure of basic LSTM with the following notation:

$$h_t = LSTM(x_t, h_{t-1}) \tag{1}$$

**First Residual based Attention Layer** We modify the basic LSTM to generate an initial text sequence feature as depicted in Fig.2. We define the input of the LSTM unit as:

$$x_t^1 = [v_0, h_{t-1}^2, w_t] \tag{2}$$

where $v_0$ is the global image visual feature, $h_{t-1}^2$ is the output of the second LSTM layer, at the $t$-1 time step, and $w_t$ is the feature of the current word derived by embedding an one-hot word vector. It is easy to understand that the current hidden states are based on the global visual feature, the previous hidden states and the $t$-th word. We further use $h_{t-1}^2$ from a higher-level LSTM in order to utilize the more precise information to guide the learning of $h_t^1$. The we get $h_t^1 = LSTM(x_t^1, h_{t-1}^1)$.

Traditionally, the hidden state of the LSTM is directly applied to guide which region should be focused on. The

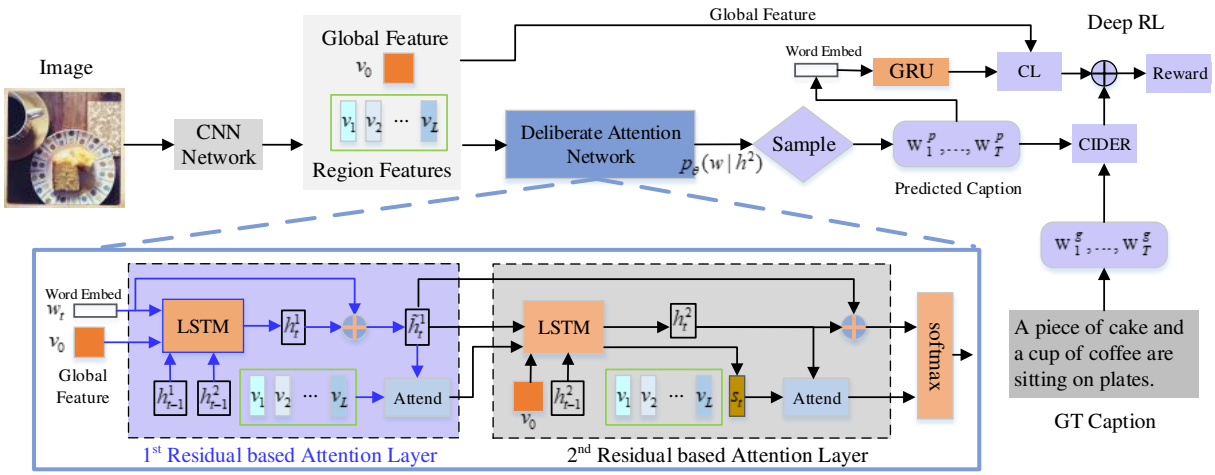


Figure 2: The proposed framework for DA, our model uses residual shortcut connection to improve information flow through two lstms. And adaptive attention is applied to calculate weights of features when predicting new word.

LSTM provides a temporal shortcut path to avoid vanishing gradients. Here, we provide an additional word shortcut from the $t$-th high frequency word for efficient training. As a result, we use a residual shortcut connection to further reduce vanishing gradients:

$$\tilde{h}_t^1 = W_{rd}[w_t; h_t^1] \tag{3}$$

where $h_t^1$ is the hidden states at the $t$-th step, $W_{rd}$ is the parameter to be learned, and [;] is a concatenation operation.

Given $L$ image region visual features $\{v_1, \cdots, v_L\}$ and the sequential context information $\tilde{h}_t^1$, we aim to selectively utilize certain region visual features by defining a visual attention mechanism below:

$$z_t^1 = w_{z1}^T \tanh(W_{v1}V, W_{h1}\tilde{h}_t^1) \tag{4}$$

$$\alpha_t^1 = soft\max(z_t^1) \tag{5}$$

where $V = [v_1, v_2..., v_L]$ indicates the $L$ image region features. And $w_{z1}^T, W_{v1}, W_{h1}$ are parameters to be learned. $\alpha_t^1 \in \mathbb{R}^L$ is the attention weight, which is then applied on region features to locate the important visual information:

$$\hat{v}_t^1 = \sum_{i=1}^{L} \alpha_{i,t}^1 v_i \tag{6}$$

where $\hat{v}_t^1$ is the attended visual information which can be used together with $\tilde{h}_t^1$ to generate the primary $t$-th word.

**Second Residual based Attention Layer** By integrating the softmax layer and the loss functions to the first residual based attention layer, we can generate a preliminary word at each step. In this subsection, we design a second residual based attention layer as a deliberate process to further purify the captioning results. Inspired by (Lu et al. 2017), we first design a visual sentinel as a complement to the visual feature, because predicting non-visual words such as "the" and "of" requires little or no visual information from the image. The revised visual sentinel is defined as:

$$g_t = \sigma(W_x x_t^2 + W_h h_{t-1}^2) \tag{7}$$

$$s_t = g_t \odot \tanh(m_t^2) \tag{8}$$

where $W_x, W_h$ are parameters to be learned, $x_t^2$ is the input of the LSTM unit. $\odot$ is an element-wise product and $\sigma$ represents the logistic sigmoid activation. More specifically,

$$x_t^2 = [v_0, \tilde{h}_t^1, \hat{v}_t^1] \tag{9}$$

After we get $h_t^2 = LSTM(x_t^2, h_{t-1}^2)$ and $s_t$ from the second LSTM, we compute an attention vector to determine when and where to look at the visual and context information. To compute the attention vector, we firstly obtain $z_t^2$ by:

$$z_t^2 = w_{z2}^T \tanh(W_{v2}V, W_{h2}h_t^2) \tag{10}$$

where $w_{z2}^T, W_{v2}, W_{h2}$ are parameters to be learned. $h_t^2$ is the LSTM output at the $t$-th time step. Next, we use the following function to calculate the attention weights $\alpha_t^2$.

$$\alpha_t^2 = soft\max([z_t^2; w_a^T \tanh(W_s s_t + W_{h3}h_t^2)]) \tag{11}$$

where $w_a^T, W_s, W_{h3}$ are parameters to be learned. $\alpha_t^2 \in \mathbb{R}^{L+1}$ contains weights for image region features and the sequential context information $s_t$. Finally, the attended results can be obtained by:

$$\hat{v}_t^2 = \sum_{i=1}^{L+1} \alpha_{i,t}^2 v_i \tag{12}$$

where $v_{L+1}$ equals to $s_t$.

To generate the $t$-th word, we combine the output of the first layer $\tilde{h}_t^1$, the output of the second LSTM $h_t^2$ and the attended visual features $v_t^2$ together by the function below:

$$\tilde{h}_t^2 = W_{sd}[\tilde{h}_t^1; h_t^2; \hat{v}_t^2] \tag{13}$$

where $W_{sd}$ is the parameter to be learned. We use softmax function to calculate the probability of the $t$-th word:

$$p_t = softma(\tilde{h}_t^2) \tag{14}$$

## Two-step Training

In essential, our training process can be divided into two stages. Firstly we pre-train the model with MLE loss, and then we fine-tune the model with reinforcement learning.

**Step 1-MLE Loss** Traditionally, parameters in image captioning model are learned by maximum likelihood estimation (MLE). The objective is to minimize the MLE loss:

$$L(\theta) = -\sum_{t=1}^{T} \log(p_\theta(w_t^*|w_1^*, ...w_{t-1}^*)) \quad (15)$$

where $\theta$ are the parameters to be learned including parameters in word embedding and two residual based attention layers. And $(w_1^*, ...w_{t-1}^*)$ represents the ground truth caption. Next, we introduce the details about reinforcement learning.

**Step 2-Reinforcement Learning** Inspired by the previous work (Luo et al. 2018; Rennie et al. 2017), we consider our DA network introduced above as "agent" to interact with external environment (i.e., words, global and region visual features), and $L(\theta)$ as the policy to conduct an action to predict a word. After the whole caption is generated, the agent observers a reward. Since CIDEr is proposed to evaluate the quality of image captioning model. We design our reward functions by combing contrastive loss (CL) with CIDEr. Next, we introduce the definition of CL.

**Contrastive Loss (CL).** Given an image $I$ and caption $c$, we obtain caption and image features by RNN and CNN, respectively. We take global image feature $v_0$ as image features. Each word in $c$ is embedded and then input into a RNN network to derive a caption feature. We define caption feature as $c_0$. Next, we map two features into a common space by $W_v^T$ and $W_c^T$, respectively:

$$f(v) = W_v^T v_0 \quad (16)$$

$$f(c) = W_c^T c_0 \quad (17)$$

Furthermore, cosine similarity is used to compute similarity between an image and caption:

$$s(I, c) = \frac{f(v) \cdot f(c)}{\|f(v)\| \, \|f(c)\|} \quad (18)$$

Parameters in the such model are learned by minimizing the contrastive loss (CL), which is a sum of two hinge losses:

$$\begin{aligned} L_{CON}(c, I) = &\max_{c'} [\alpha + \mathrm{s(I, c')} - \mathrm{s(I, c)}]_+ \\ &+ \max_{I'} [\alpha + \mathrm{s(I', c)} - \mathrm{s(I, c)}]_+ \end{aligned} \quad (19)$$

where $[x]_+ \equiv \max(x, 0)$. In Eq.19, $(c, I)$ indicates that a pair of caption and image is matched. $c$ correctly describes image $I$. Both $(I, c')$ and $(I', c)$ are mismatched. $((I, c'))$ suggests that $c'$ is the incorrect description of $I$, while $(I', c)$ suggests that $c$ is the incorrect description of $I'$. $\alpha$ is used to ensure minimum gap between scores of $(c, I)$ with $(I, c')$ and $(I', c)$.

**CIDEr+CL** In order to optimize the parameters in our model, the objective becomes to maximize the reward obtained from reward function by learning parameters. An update is implemented by computing the gradient of the expected reward:

$$\nabla_\theta E[R(\hat{c}, I)] \approx (R(\hat{c}, I) - R(c^*, I)) \nabla_\theta \log p(\hat{c}|I; \theta) \quad (20)$$

where $R(\hat{c}, I) = CIDEr(\hat{c}) - L_{CON}(\hat{c}, I)$ is reward function. $\hat{c}$ is caption generated from (14). The baseline is computed by $c^* = (BOS, w_1^*, ..., w_T^*)$, which is obtained as:

$$w_t^p = \arg\max_w (w|w_{0,...t-1}^p, I) \quad (21)$$

# Experiments

## Datasets

In this paper, we utilize two datasets including COCO (Lin et al. 2014) and Flickr30K (Young et al. 2014) to evaluate the performance of our proposed DA network.

**COCO.** It is the largest dataset for image captioning, which consists of $82,783$, $40504$ and $40,775$ images for training, validation and testing, respectively. In COCO dataset, each image has 5 captions annotated by human beings. Following previous work (You et al. 2016; Lu et al. 2017; Luo et al. 2018; Anderson et al. 2018), we use the "Karpathy" splits proposed in (Karpathy and Li 2015). In this split, $113,287$, $5,000$ and $5,000$ images are used for training, validation and testing, respectively.

**Flickr30K.** It consists of $31,783$ images collected from Flickr. In particular, each image is associated with 5 crowdscoured descriptions. Most of the images are about human beings performing activities. Following previous work (Lu et al. 2017), we use 29k images for training, 1k for validation and 1k for testing.

For both COCO and Flickr30K dataset, we conduct a preprocessing procedure by firstly truncating captions longer than 16 words. Next, all modified captions are converted to lower case. For each dataset, we build a vocabulary. For COCO, the vocabulary contains $9,487$ words, while for Flickr30K the vocabulary has 7k words.

## Evaluation Metric

Five generally used evaluation metrics are adopted to evaluate the performance of image captioning, including BLUE (Papineni et al. 2002), ROUGEL (Lin 2004), METEOR (Denkowski and Lavie 2014), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015) and SPICE (Anderson et al. 2016). More specifically, for the COCO dataset, we also report the SPICE subclass scores on 5k validation sets, including Color, Attribute, Cardinality, Object, Relation and Size. All the SPICE subclass scores are scaled up by 100.

## Implementation Details

To extract the global visual feature, we use pre-trained ResNet-101 (He et al. 2016) to process the input image and the output of "pool5" (2048-D) is used as global appearance feature. In terms of region visual features, we utilize the same region features used in (Anderson et al. 2018). As a result, for each image, we obtain 36 region features and the dimension for each region feature is also 2048-D. In addition, the hidden state size of two LSTM in our DA network is set to be 512.

To compute the contrastive loss, we use RNN as the text encoder to extract caption features and use pool5 of ResNet-101 to extract image features. The dimension of word embedding is 512, the RNN hidden state size is set as 1024 and

Table 1: Ablation study results obtained from the COCO dataset.

| model | BLUE1 | BLUE4 | METEOR | ROUGE | CIDEr | SPICE | Att | Card | Col | Obj | Rel | Size |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DA-De-RF | 74.2 | 33.7 | 26.4 | 54.6 | 104.9 | 19.4 | 9.3 | 2.1 | 11.0 | 35.5 | 5.2 | 3.8 |
| DA-De+RF | 78.4 | 35.2 | 27.3 | 56.5 | 117.3 | 21.1 | 9.2 | 13.0 | 10.5 | 39.2 | 5.6 | 2.8 |
| DA+De-RF | 75.8 | 35.7 | 27.4 | 56.2 | 111.9 | 20.5 | 10.8 | 6.1 | 14.6 | 36.8 | 5.5 | **5.6** |
| DA+De+RF | **79.9** | **37.5** | **28.5** | **58.2** | **125.6** | **22.3** | **11.2** | **14.4** | **15.2** | **40.3** | **6.4** | 3.7 |

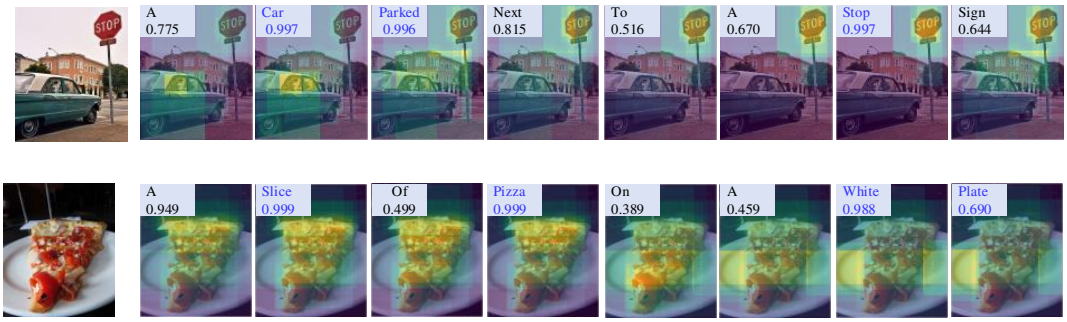


Figure 3: Visualization of first residual attention map of the DA+De+RF model. The sentence is generated by the DA+De+RF. The region with the maximum attention weight is in orange. We also show each word with the corresponding visual weight in the second residual attention block.

the dimension of embedded image feature is 1024. To train the model for calculate contrastive loss, we set epochs as 27 for both COCO and Flicker30K.

For training process, MLE is firstly used to pre-train the DA model. Next, we train it with reinforcement learning by using CIDEr and CL as reward value. For MLE training, the epoch is set as 150 for both COCO and Flickr30K, while for reinforcement learning the epoch is set as 200 for COCO and 150 for Flickr30K. All models are trained by using Adam and the batch size is set as 128. We initialize the learning rate with 5e-4 and update it by a decreasing factor 0.8 in every 15 epochs. When conduct testing, beam search is applied to predict captions, with beam size setting as 5.

## Ablation Study

In order to figure out the contribution of each component, we conduct the following ablation studies on the COCO dataset. Specifically, we remove the deliberation process (De) and reinforcement learning (RF) respectively from our DA model, and have four experiments: DA-De-RF, DA-De+RF, DA+De-RF and DA+De+RF. The experimental results are shown in Tab. 1.

**The Influence of Deliberation.** From Tab. 1, we can see that with or without reinforcement learning, our DA+De models, including DA+De+RF and DA+De-RF, perform better than DA-De models (DA-De-RF and DA-De+RF). More specifically, DA+De-RF outperforms DA-De-RF on all 12 metrics, in particular with an increase of 2% on BLUE4, 1% on METEOR, 1.6% ROUGE, 7% CIDEr and 1.1% SPICE. In addition, compared with DA-De+RF, DA+De+RF obtains higher performance on all 12 metrics. This verifies the importance of our deliberation mechanism.

In order to further demonstrate the role of deliberation mechanism, we show four visual examples in Fig.1. In Fig.1, each image contains two descriptions. The first sentence is generated by the DA-De+RF and the second sentence is generated by the DA+De+RF model. We can see that without deliberation process, the model can generate a sentence which contains error information (e.g., "a hair with a hair") or inconsistent semantic information, e.g., "a dog looking out of a car mirror". With the deliberation component, our DA model can provide more precise description, such as "a yellow food truck" instead of "a yellow bus"; and a more reasonable description: "a women taking a picture of a dog in a car mirror" instead of "a dog looking out of a car mirror". These examples also show that the first residual attention based layer can detect primary objects (e.g., girl, hair) and activities (e.g., holding), while the second residual attention based layer can refine the activities (e.g., brushing) and detect the relationship between objects (e.g., "a women taking a picture of a dog" and "a dog in a car mirror").

In addition, Fig.3 shows the attended image regions of the first residual based attention of the DA+De+RF. For each generated word, we visualize the attention weights on individual pixels, outlining the region with the maximum attention weight in orange. Moreover, for each word, we display the corresponding visual weight of second residual based attention layer. From Fig.3, we find that our first layer attention is able to locate the right objects, which enables the DA+De+RF to accurately describe objects occurred in the input image. On the other hand, the visual weights in the second layer are obviously higher when our model predicts words related to objects (e.g., car and pizza).

**The Influence of Reinforcement Learning.** From Tab. 1,

Table 2: Performance on Flicker30k test split. DA refers to the DA+De+RF mode in Tab.1.

| model | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | CIDEr |
|---|---|---|---|---|---|---|
| DeepVS (Karpathy and Li 2015) | 57.3 | 36.9 | 24.0 | 15.7 | 15.3 | 24.7 |
| Hard-Attention (Xu et al. 2015) | 66.9 | 43.9 | 29.6 | 19.9 | 18.5 | - |
| ATT-FCN (You et al. 2016) | 64.7 | 46.0 | 32.4 | 23.0 | 18.9 | - |
| Adaptive-Attention (Lu et al. 2017) | 67.7 | 49.4 | 35.4 | 25.1 | 20.4 | 53.1 |
| DA | **73.8** | **55.1** | **40.3** | **29.4** | **23.0** | **66.6** |
| Relative Improvement | 6.1 | 5.7 | 4.9 | 4.3 | 2.6 | 13.5 |

Table 3: Single-model image captioning performance on the COCO Karpathy test split. B-4, M, R, C and S are BLUE4, METEOR, ROUGE, CIDEr and SPICE scores, respectively. All methods are based on the reinforcement learning.

| model | B4 | MR | R | C | S | Att | Card | Col | Obj | Rel | Size |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SCST:Att2in (Rennie et al. 2017) | 31.3 | 26.0 | 54.3 | 101.3 | - | - | - | - | - | - | - |
| SCST:Att2all (Rennie et al. 2017) | 30.0 | 25.9 | 53.4 | 99.4 | - | - | - | - | - | - | - |
| ATTN+C+D(1) (Luo et al. 2018) | 36.3 | 27.3 | 57.1 | 114.1 | 21.1 | 9.5 | 10.5 | 9.3 | 39.0 | 5.9 | 2.6 |
| Up-Down (Anderson et al. 2018) | 36.3 | 27.7 | 56.9 | 120.1 | 21.4 | 10.0 | **18.4** | 11.4 | 39.1 | **6.5** | 3.2 |
| DA | **37.5** | **28.5** | **58.2** | **125.6** | **22.3** | **11.2** | 14.4 | **15.2** | **40.3** | 6.4 | **3.7** |

DA+De-RF : a man doing a trick on a skateboard in the air
DA+De+RF : a man doing a trick on a skateboard on a street

DA+De-RF : a man doing a trick on a skateboard in the air
DA+De+RF : a man doing a trick on a skateboard on a park

Figure 4: The role of RF. Examples of image captions generated by DA+De-RF and DA+De+RF.

we can see that the results clearly show the advantage of our reinforcement learning. From the first block of Tab. 1, we can see that DA-De+RF achieves 4.2% BLUE1, 1.5% BLUE4, 0.9% METEOR, 1.5% ROUGE, 12.4% CIDEr and 1.7% SPICE performance gain compared with DA-De-RF. Moreover, compared with DA+De-RF, DA+De+RF performs better with an increase of 4.1% BLUE1, 1.8% BLUE4, 1.1% METEOR, 2.0% ROUGE, 13.7% CIDEr and 1.8% SPICE. The increases of performance clearly demonstrate the effectiveness of the proposed reinforcement learning component.

In order to further demonstrate the discriminability, we show some qualitative results in Fig.4. By observing the two images, a human can aware that two men are performing the same activity (i.e., doing a tricks on a skateboard), but they are playing at different places (i.e., on a street and on a park). The first man plays skateboard on the street, while the second man plays skateboard on a park. From Fig.4, we can see that DA+De-RF generates the same description for both images, while DA+De+RF is able to generate precise caption to describe their difference (i.e., on a street or on a park). This indicates the reinforcement learning with contrastive loss improves the discriminability of our DA model.

## Comparing with State-of-the-Art

In this section, we use DA to represent our model DA+De+RF for convenience.

**Flickr30K.** For Flickr30K dataset, we compare our DA with DeepVS (Karpathy and Li 2015), Hard-Attention (Xu et al. 2015), ATT-FCN (You et al. 2016) and Adaptive-Attention (Lu et al. 2017) and the comparison results are shown in Tab. 2. From Tab. 2, we can see that our DA outperforms the counterparts by a large margin. Specifically, compared with Adaptive (Lu et al. 2017), DA has 6.1% BLUE-1, 5.7% BLUE-2, 4.9% BLUE-3, 4.3% BLUE-4, 2.6% METEOR and 13.5% CIDEr increases. The improvement is significant, especially for BLUE-n and CIDEr.

**COCO.** We conduct two types of evaluations on the COCO dataset. The first is conducted offline by using the "Karpathy" split that have been widely used in prior work. The second one is conducted online and the captioning results are obtained on the online MSCOCO test server.

For offline evaluation, all the image captioning models are single-model. Here, we compare DA with SCST:Att2in (Rennie et al. 2017), SCST:Att2all (Rennie et al. 2017), ATTN+C+D(1) (Luo et al. 2018) and Up-Down (Anderson et al. 2018). The offline evaluation results are shown in Tab. 3. It is clear that our DA performs the best on the widely used evaluation metrics, e.g., BLUE4, METEOR, ROUGE, CIDEr and SPICE scores. Up-Down (Anderson et al. 2018) is a strong competitor, and it performs the best for "Card" and "Rel".

For online evaluation, we compare with previous published works, including Review Net (Yang et al. 2016), Adaptive (Lu et al. 2017), PG-BCMR (Liu et al. 2017), SCST:Att2all (Rennie et al. 2017), LSTM-A3 (Yao et al. 2017), Up-Down (Anderson et al. 2018). The experimental results are shown in Tab. 4. From Tab. 4, we can see that beside METEOR, Up-down in general performs the best.

Table 4: Results on the online MSCOCO test sever. DA is a single model, both SCST:Att2all and Up-Down are an ensemble of 4 models. LSTM-A3 utilizes Resnet-152 based visual feature while DA uses RestNet101 based visual feature.

| | BLEU-1 | | BLEU-2 | | BLUE-3 | | BLUE-4 | | METEOR | | ROUGE-L | | CIDEr | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | c5 | c40 | c5 | c40 | c5 | c40 | c5 | c40 | c5 | c40 | c5 | c40 | c5 | c40 |
| Review Net | 72.0 | 90.0 | 55.0 | 81.2 | 41.4 | 70.5 | 31.1 | 59.7 | 25.6 | 34.7 | 53.5 | 68.6 | 96.5 | 96.9 |
| Adaptive | 74.8 | 92.0 | 58.4 | 84.5 | 44.4 | 74.4 | 33.6 | 63.7 | 26.4 | 35.9 | 55.5 | 70.5 | 104.2 | 105.9 |
| PG-BCMR | 75.4 | - | 59.1 | - | 44.5 | - | 33.2 | - | 25.7 | - | 55 | - | 101.3 | - |
| SCST:Att2all | 78.1 | 93.7 | 61.9 | 86.0 | 47.0 | 75.9 | 35.2 | 64.5 | 27.0 | 35.5 | 56.3 | 70.7 | 114.7 | 116.7 |
| LSTM-A3 | 78.7 | 93.7 | 62.7 | 86.7 | 47.6 | 76.5 | 35.6 | 65.2 | 27.0 | 35.4 | 56.4 | 70.5 | 116.0 | 118.0 |
| Up-Down | **80.2** | **95.2** | **64.1** | **88.8** | **49.1** | **79.4** | **36.9** | **68.5** | 27.6 | 36.7 | 57.1 | **72.4** | 117.9 | **120.5** |
| DA | 79.4 | 94.4 | 63.5 | 88.0 | 48.7 | 78.4 | 36.8 | 67.4 | **28.2** | **37.0** | **57.7** | 72.2 | **120.5** | 122.0 |


Figure 5: Examples of image captions generated by Up-Down, DA and human beings. The first column, both Up-Down and our DA provide accurate descriptions. The middle column shows that in some cases our DA is able to provide better descriptions, while the third column indicates that in complex situations both Up-Down and DA fail.

In fact, both Up-Down and SCST:Att2All are an ensemble of 4 models, while our DA uses a single-model. Although LSTM-A3 utilizes better visual features extracted from the ResNet-152, our DA with ResNet-101 visual features obtains higher performances, especially the METEOR scores reaching 28.2% on c5 and 37.0% on c40. We believe that the performance of DA could be further boosted via an ensemble of multiple DA-based models.

## Qualitative Analysis

Here, we show some qualitative results in Fig.5. From the above Tables (i.e., Tab. 2, Tab. 3 and Tab.4), we can see that the previously Adaptive (Lu et al. 2017) performs the best on the Flicker30k while Up-Down (Anderson et al. 2018) performs the best on the COCO dataset. Due to the reason that Up-Down (Anderson et al. 2018) releases the code while Adaptive (Lu et al. 2017) does not, we show some captioning examples obtained by our DA and Up-Down. The first column with two examples show that both DA and Up-Down are able to provide accurate description. For the middle column, we can see that our DA provides more accurate descriptions, especially for describing the relationships among objects. In this case, Up-Down fails to detect all objects and the relationships among them. The third column shows two images and both of them have complex background and their corresponding descriptions contain rich semantic information. For those two images, both DA and Up-Down fails. One possible reason is that a human being can conduct reasoning based on his or her background knowledge while at this stage both DA and Up-Down cannot. This proposes a potential research direction for image captioning.

## Conclusion

In this work, we propose a novel architecture, deliberate residual attention networks (DA) for image captioning. DA consists of two residual based attention layers. The first-pass residual-based attention layer learns the hidden states and visual attention which can be used for generating a raw captions, while the second-pass deliberate residual-based attention layer refines them using more information, e.g., the global visual and contextual information. Besides, we guide the process of training with reinforcement learning, by combining CIDEr with CL as the reward function. Results show that our approach outperforms the state-of-the-art methods.

## Acknowledgements

This work is supported by the Fundamental Research Funds for the Central Universities (Grant No. ZYGX2014J063, No. ZYGX2016J085), the National Natural Science Foun-

dation of China (Grant No. 61772116, No. 61872064, No. 61632007, No. 61602049).

## References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*, 382–398. Springer.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Bengio, S.; Vinyals, O.; Jaitly, N.; and Shazeer, N. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *NIPS*, 1171–1179.

Dai, B.; Fidler, S.; Urtasun, R.; and Lin, D. 2017. Towards diverse and natural image descriptions via a conditional GAN. In *ICCV*, 2989–2998.

Denkowski, M., and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, 376–380.

Dognin, P. L.; Melnyk, I.; Mroueh, Y.; Ross, J.; and Sercu, T. 2018. Improved image captioning with adversarial semantic alignment. *CoRR* abs/1805.00063.

Gao, L.; Zeng, P.; Song, J.; Liu, X.; and Shen, H. T. 2018. Examine before you answer: Multi-task learning with adaptive-attentions for multiple-choice VQA. In *ACM Multimedia*, 1742–1750.

Goyal, A.; Lamb, A.; Zhang, Y.; Zhang, S.; Courville, A. C.; and Bengio, Y. 2016. Professor forcing: A new algorithm for training recurrent networks. In *NIPS*, 4601–4609.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Karpathy, A., and Li, F. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 3128–3137.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123:32–73.

Lin, T. Y.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; and Zitnick, C. L. 2014. Microsoft common objects in context. In *ECCV*, 740–755.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Liu, S.; Zhu, Z.; Ye, N.; Guadarrama, S.; and Murphy, K. 2017. Improved image captioning via policy gradient optimization of spider. In *ICCV*, 873–881.

Liu, X.; Li, H.; Shao, J.; Chen, D.; and Wang, X. 2018. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. *CoRR* abs/1803.08314.

Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 3242–3250.

Luo, R.; Price, B. L.; Cohen, S.; and Shakhnarovich, G. 2018. Discriminability objective for training descriptive captions. volume abs/1803.04376.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 311–318.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, 91–99.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *CVPR*, 1179–1195.

Song, J.; Gao, L.; Guo, Z.; Liu, W.; Zhang, D.; and Shen, H. T. 2017. Hierarchical LSTM with adjusted temporal attention for video captioning. In *IJCAI*, 2737–2743.

Song, J.; Guo, Y.; Gao, L.; Li, X.; Hanjalic, A.; and Shen, H. T. 2018a. From deterministic to generative: Multimodal stochastic rnns for video captioning. *IEEE Transactions on Neural Networks and Learning Systems* 1–12.

Song, J.; Zeng, P.; Gao, L.; and Shen, H. T. 2018b. From pixels to objects: Cubic visual attention for visual question answering. In *IJCAI*, 906–912.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, 4566–4575.

Xia, Y.; Tian, F.; Wu, L.; Lin, J.; Qin, T.; Yu, N.; and Liu, T. 2017. Deliberation networks: Sequence generation beyond one-pass decoding. In *NIPS*, 1782–1792.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A. C.; Salakhutdinov, R.; Zemel, R. S.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2048–2057.

Yang, Z.; Yuan, Y.; Wu, Y.; Cohen, W. W.; and Salakhutdinov, R. R. 2016. Review networks for caption generation. In *NIPS*, 2361–2369.

Yao, T.; Pan, Y.; Li, Y.; Qiu, Z.; and Mei, T. 2017. Boosting image captioning with attributes. In *ICCV*, 4904–4912.

You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016. Image captioning with semantic attention. In *CVPR*, 4651–4659.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2:67–78.

Zhang, M.; Yang, Y.; Zhang, H.; Ji, Y.; Shen, H. T.; and Chua, T.-S. 2019. More is better: Precise and detailed image captioning using online positive recall and missing concepts mining. *IEEE Transactions on Image Processing* 28(1):32–44.