Multi-Level Policy and Reward Reinforcement Learning for Image Captioning

An-An Liu[1], **Ning Xu**[1], **Hanwang Zhang**[2], **Weizhi Nie**[1], **Yuting Su**[1], **Yongdong Zhang**[3]
[1] School of Electrical and Information Engineering, Tianjin University, Tianjin, China
[2] School of Computer Science and Engineering, Nanyang Technological University, Singapore
[3] University of Science and Technology of China, Hefei, China
liuanan@tju.edu.cn

Abstract

Image captioning is one of the most challenging hallmark of AI, due to its complexity in visual and natural language understanding. As it is essentially a sequential prediction task, recent advances in image captioning use Reinforcement Learning (RL) to better explore the dynamics of word-by-word generation. However, existing RL-based image captioning methods mainly rely on a single policy network and reward function that does not well fit the multi-level (word and sentence) and multi-modal (vision and language) nature of the task. To this end, we propose a novel multi-level policy and reward RL framework for image captioning. It contains two modules: 1) Multi-Level Policy Network that can adaptively fuse the word-level policy and the sentence-level policy for the word generation; and 2) Multi-Level Reward Function that collaboratively leverages both vision-language reward and language-language reward to guide the policy. Further, we propose a guidance term to bridge the policy and the reward for RL optimization. Extensive experiments and analysis on MSCOCO and Flickr30k show that the proposed framework can achieve competing performances with respect to different evaluation metrics.

1 Introduction

Image captioning is the task of describing the visual content of an image using natural language. Unlike traditional computer vision tasks, such as image classification and object detection, image captioning requires not only visual understanding the image, but also the compositions of natural language. This technique can be widely applied to semantic image retrieval [Karpathy *et al.*, 2014] and human-robot interactions [Das *et al.*, 2017].

Image captioning is a sequential word prediction task. State-of-the-art approaches [Xu *et al.*, 2015; You *et al.*, 2016; Karpathy and Fei-Fei, 2017; Li *et al.*, 2017] generally follow an encoder-decoder framework: they deploy convolutional neural networks (CNN) to encode the image into a visual embedding vector, and then use recurrent neural networks (RNN) to decode the vector into a sentence; during training and inference, they try to maximize the probability of the next word based on the current prediction context. Recently, it has been shown that Reinforcement Learning (RL) [Sutton *et al.*, 1999] can better fit in this task. The reason is that RL aims to learn a policy that decides sequential actions by maximizing the cumulative future rewards [Silver *et al.*, 2016]. Therefore, RL can help to explore more fruitful language in sentence generation, avoiding severe bias in training samples [Rennie *et al.*, 2017]. However, existing RL-based image captioning methods [Liu *et al.*, 2017b; Rennie *et al.*, 2017] mainly rely on a single policy network and reward function that does not well fit the multi-level (word and sentence) and multi-modal (vision and language) nature of the task.

In this paper, we propose a novel multi-level policy and reward reinforcement learning framework for image captioning. The multi-level policy network aims to adaptively fuse the word-level and the sentence-level policies for word generation, and the multi-level reward function aims to collaboratively leverage the vision-language and the language-language rewards to guide the policy. To further bridge the policy network and the reward function, we propose a guidance term by minimizing the distance between the sentence-level policy and the vision-language reward for optimization.

As shown in Figure 1, the multi-level policy network consists of the word-level policy and the sentence-level policy. The former is the CNN-RNN-based network, which provides the word confidence by locally predicting the next word based on the current state; the latter is a visual-semantic embedding network, which provides the sentence (context) confidence by globally evaluating the current state. The multi-level reward function consists of the vision-language reward and the language-language reward. The former is also a visual-semantic embedding network, which measures the cross-modality similarity between the visual content and the generated description, and defines a specific optimization goal for reinforcement learning; the later is the CIDEr [Vedantam *et al.*, 2015] metric on a pre-defined rule and a stable supplement to the former. Particularly, the vision-language reward and the sentence-level policy leverage the same embedding architecture while owning different parameters. We pre-train an embedding network to initialize the former while the later is directly trained in RL framework.

The contributions are summarized as follows.

- We propose a novel multi-level policy and reward reinforcement learning framework for image captioning. To the best of our knowledge, it is the first RL framework that explores the multi-level (word and sentence) and multi-modal (vision and language) nature of image captioning task.
- We design a multi-level policy network that adaptively fuses word and sentence confidences for sentence generation, and a multi-level reward function that collaboratively leverages vision-language and language-language rewards to guide the generation. A proposed guidance term further bridges the policy and the reward modules.
- We perform comprehensive evaluations on MSCOCO and Flickr30k datasets. Our framework achieves the competing performances against state-of-the-art methods. Ablative studies showcase the effect of the proposed framework.

2 Related Work

2.1 Image Captioning

Many image captioning methods have been proposed in the literature. In the early stage, the template-based methods [Farhadi *et al.*, 2010] detected objects from images to generate sentences by pre-defined grammar rules. Recently, a CNN-RNN-based framework was explored and variants were proposed [Vinyals *et al.*, 2015; Karpathy and Fei-Fei, 2017; Dai *et al.*, 2017; Liu *et al.*, 2017a; Nie *et al.*, 2013]. For examples, [Wu *et al.*, 2016] incorporated high-level semantic concepts into the CNN-RNN framework. Additionally, attention-based methods weighted each feature to exploit the spatial structure and rich intermediate description for images [Xu *et al.*, 2015; Anderson *et al.*, 2017; Zhang *et al.*, 2017; Cheng *et al.*, 2018]. For examples, [Xu *et al.*, 2015] used the attention model to learn where to focus in images during sentence generation. [Anderson *et al.*, 2017] incorporated bottom-up and top-down attention models and [Lu *et al.*, 2017] proposed the adaptive attention model that decided whether to attend to the image or to the visual sentinel.

2.2 Reinforcement Learning

RL aims to learn a policy that decides sequential actions by maximizing the cumulative future rewards [Silver *et al.*, 2016]. Recently, many challenging problems, such as the game of Go [Silver *et al.*, 2016], can be successfully solved by RL algorithms. Several RL methods have been proposed to solve the computer vision problems, such as visual tracking [Yun *et al.*, 2017] and image captioning [Ren *et al.*, 2017]. For examples, [Ren *et al.*, 2017] proposed a "policy network" and a "value network" to collaboratively generate captions, with a reward defined by visual-semantic embedding. [Rennie *et al.*, 2017] and [Liu *et al.*, 2017b] directly optimized image captioning systems by the test rewards.

3 Approach

3.1 Problem Formulation

We formulate image captioning as a RL process. In image captioning, the goal is, given an image F, to generate a sentence $\hat{S} = \{\hat{w}_1, \ldots, \hat{w}_T\}$ which correctly describes the image content, where $\hat{w}_i$ is a word in sentence $\hat{S}$ and T is the length. In RL, there is an *agent (policy)* that interacts with the *environment*, and executes a series of *actions*, so as to optimize a *goal*. Particularly, the environment is the given image F and the words predicted so far $\{\hat{w}_1, \ldots, \hat{w}_t\}$. An action is to predict the next word $\hat{w}_{t+1}$. After each action a, a state s is observed. The state s_t at time step t consists of the image F and the words predicted until t, $\{\hat{w}_1, \ldots, \hat{w}_t\}$. The action space is the dictionary $\mathcal{D}$ that the words are drawn from, i.e., $a_t \subset \mathcal{D}$. However, existing RL-based image captioning methods rely on a single policy network and reward function that does not well fit the multi-level (word and sentence) and multi-modal (vision and language) nature of the task. To this end, we propose a novel multi-level policy and reward RL framework for image captioning.

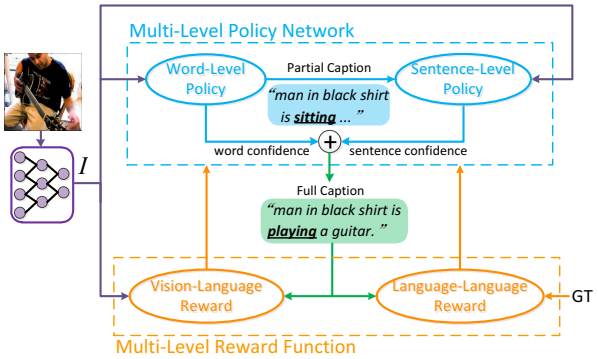

Figure 1: Illustration of the proposed multi-level policy and reward RL framework. A multi-level policy network can adaptively fuse both word-level and sentence-level policies to generate each word, and a multi-level reward function collaboratively leverages both vision-language and language-language rewards to guide the policy.

3.2 Multi-Level Policy Network

The multi-level policy network consists of the word-level policy and the sentence-level policy.

Word-level policy consists of a Convolutional Neural Network (CNN) and a Long Short Term Memory Network (LSTM). It is similar to the image captioning model [Karpathy and Fei-Fei, 2017] used in the encoder-decoder framework. We first extract the CNN feature I for the input image, and then embed it through a linear mapping. Words are represented by one hot vectors which are embedded with the same dimension as mapped image features. The beginning of each sentence is marked with a special $\langle BOS \rangle$ token, and the end with an $\langle EOS \rangle$ token. In this policy, words are generated and then fed back into LSTM, with the image feature I treated as the first word. LSTM outputs a distribution $\bar{w}_t$ over all words by updating the hidden states and cells of it. Let θ_π denote the parameters of the word-level policy. $\{\bar{w}_1, \ldots, \bar{w}_{t-1}\}$ is denoted by $\bar{S}_{1:t-1}$. The objective is to minimize the sum of the negative log likelihood of the correct word at each step:

$$\mathcal{L}(\theta_\pi) = -\sum_{t=1}^{T} log\big(p_\pi(\bar{w}_t | I, \bar{S}_{1:t-1})\big) \qquad (1)$$

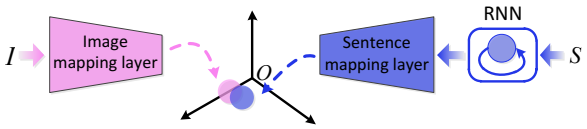

Figure 2: Illustration of the visual-semantic embedding network, which are synchronously used in the sentence-level policy and the vision-language reward. It is comprised of two mapping layers and a RNN unit. By projecting image feature I and sentence S into one common embedding space, it measures the similarity between vision and semantic.

Sentence-level policy is a visual-semantic embedding network, which has been successfully applied to image retrieval [Kiros *et al.*, 2015] and captioning [Pan *et al.*, 2016]. Inspired by [Ren *et al.*, 2017], we map the image feature I and the sentence S into one common embedding space that measures the similarity between them. As shown in Figure 2, given a sentence S, its embedding feature is represented using the last hidden state of RNN. We denote a sentence mapping layer by $h_p(\text{RNN}(S))$ and a image mapping layer by $f_p(I)$. As shown in Figure 1, the sentence-level policy is fed by the image feature I and the partially generated caption $\bar{S}_{1:t}$ from the word-level policy. The confidence between them is computed by:

$$c(I, \bar{S}_{1:t}) = \frac{f_p(I) h_p(\text{RNN}(\bar{S}_{1:t}))}{\|f_p(I)\| \|h_p(\text{RNN}(\bar{S}_{1:t}))\|} \tag{2}$$

Further, the sentence-level policy can provide the sentence confidence by globally evaluating the current state.

3.3 Multi-Level Reward Function

The multi-level reward function consists of the vision-language reward and the language-language reward.

Vision-language reward is a visual-semantic embedding network that has the same architecture with the sentence-level policy. But there are two different points between them. The first point is this reward is fed by the image feature I and the fully generated caption $\hat{S}$, instead of partially generated caption $\bar{S}_{1:t}$, from the multi-level policy network, which fuses both word-level and sentence-level policies. It can evaluate the vision-language correlation on the fully generated caption, and define a specific goal for RL optimization. As shown in Figure 2, we denote a sentence mapping layer by $h_r(\text{RNN}(\hat{S}))$ and a image mapping layer by $f_r(I)$. This reward is defined by:

$$r_{vl}(I, \hat{S}) = \frac{f_r(I) h_r(\text{RNN}(\hat{S}))}{\|f_r(I)\| \|h_r(\text{RNN}(\hat{S}))\|} \tag{3}$$

The second point is that we pre-train the embedding space for this reward while the sentence-level policy is directly trained in the RL framework. Let θ_r denote the parameters of the vision-sentence reward. Inspired by [Ren *et al.*, 2017], we use the image-sentence pairs as in the image captioning dataset, and learn the RNN weights as well as mapping layers using a bi-directional ranking loss:

$$\begin{aligned} \mathcal{L}(\theta_r) = &\sum_I \sum_{S^-} \max\big(0, \gamma - f_r(I)h_r(S) + f_r(I)h_r(S^-)\big) \\ &+ \sum_S \sum_{I^-} \max\big(0, \gamma - h_r(S)f_r(I) + h_r(S)f_r(I^-)\big) \end{aligned} \tag{4}$$

where γ is the margin cross-validated, every (I, S) are a ground truth image-sentence pair, S^- denotes a negative description for the image corresponding to I, and vice-versa with I^-.

Language-language reward is the CIDEr metric which has been successfully applied to image captioning task [Rennie *et al.*, 2017]. Because CIDEr is calculated on the pre-defined rule, it can steadily evaluate the sequential actions. We use language-language reward as the supplement to the vision-language one. It is computed by comparing the fully generated caption $\hat{S}$ with the corresponding ground truth S, which is denoted by $r_{ll}(S, \hat{S})$.

In this paper, we directly optimize a linear combination of both rewards as follows:

$$r_{total} = \begin{cases} 0 & 0 < t < T \\ \lambda r_{vl}(I, \hat{S}) + (1-\lambda) r_{ll}(S, \hat{S}) & t = T \end{cases} \tag{5}$$

where r_{total} is the linear combination of both rewards. T is the length of sentences. $0 \leq \lambda \leq 1$ is a hyperparameter.

3.4 Training Using Reinforcement Learning

The key problem of RL lies in correlating the policy and the reward parts for joint learning. Except that we co-train both parts in the traditional RL framework, we design a guidance term $\mathcal{G}$. It minimizes the distance between the vision-language reward and the sentence-level policy by calculating the mean squared loss. Since the vision-language reward is pre-trained with ground truth, it can be regarded as an expert to measure the correlation between images and sentences. However, the sentence-level policy is trained in the RL framework by leveraging all information in the environment. It can be regarded as an amateur for the similarity measure between images and sentences. Therefore, by minimizing $\mathcal{G}$, the expert reward will guide the amateur policy for optimization and further benefit joint learning both parts. Let θ_a denote the parameters of the sentence-level policy. The guidance term can be formulated as:

$$\mathcal{G}(\theta_a) = \|r_{vl}(I, \hat{S}) - c(I, \bar{S}_{1:t})\|^2 \tag{6}$$

We denote the parameters of the multi-level policy network by $\Theta = \{\theta_\pi, \theta_a\}$, and we learn Θ by minimizing the negative expected combination reward r_{total}, the guidance term $\mathcal{G}$, and the distribution of generated words $p_{\hat{w}_t}$ (will be defined later in Section 3.5). The objective function can be formulated as:

$$\mathcal{J}(\Theta) = -r_{total} \times p_{\hat{w}_t} \times \mathcal{G} \tag{7}$$

The training process consists of two steps.

- By standard supervised learning, we pre-train the word-level policy θ_π and the vision-language reward θ_r in Eq. 1 and Eq. 4, respectively.
- θ_π and θ_a are jointly trained in Eq. 7. We baseline RL with not only the sentence-level policy $c(I, \bar{S}_{1:t})$ but also the language-language reward $r_{ll}(S, \hat{S}')$ that is obtained by the current model under the inference algorithm used at test time. A sample approximation to the gradient is:

$$\nabla_{\theta_\pi} \mathcal{J} \approx \sum_{t=1}^{T} \nabla_{\theta_\pi} \, logp_\pi(\bar{w}_t | I, \bar{S}_{1:t-1}) \\ (r_{total} - \lambda c(I, \bar{S}_{1:t}) - (1-\lambda) r_{ll}(S, \hat{S}')) \quad (8)$$

$$\nabla_{\theta_a} \mathcal{J} \approx \nabla_{\theta_a} c(I, \bar{S}_{1:t}) \\ (r_{total} - \lambda c(I, \bar{S}_{1:t}) - (1-\lambda) r_{ll}(S, \hat{S}'))$$

Here, $c(I, \bar{S}_{1:t})$ and $r_{ll}(S, \hat{S}')$ serve as a combine-moving baseline by λ. The subtraction with the evaluation leads to a much lower variance estimate of the policy gradient. Scaling the gradient can be seen as an estimate of the advantage of action a_t in state s_t.

3.5 Lookahead Inference with Multi-Level Policy

For the multi-level policy network, the inference is guided by the word-level and the sentence-level policies. The former provides the word confidence that locally predicts the next word according to current state. The later provides the sentence confidence that globally evaluates the current state. These complement confidences are collaboratively used to adjust the distribution of next word towards the goal of generating captions that are similar to ground truth.

The agent executes each action by fusing both policies:

$$p_{\hat{w}_t} = \beta logp_\pi(\bar{w}_t | I, \bar{S}_{1:t-1}) + (1-\beta) c(I, \bar{S}_{1:t}) \quad (9)$$

where $p_{\hat{w}_t}$ is the adjusted distribution of the next word. $0 \leq \beta \leq 1$ is a hyperparameter.

4 Experiments

4.1 Datasets

We evaluate our framework on captioning datasets: MSCOCO and Flickr30k. For fair comparison, we adopt the splits consistent with [Karpathy and Fei-Fei, 2017], which uses 5,000 images for validation and test on MSCOCO; 1,000 images for validation and test on Flickr30k. We drop any word that has count less then five, yielding a vocabulary of size 9,487 and 7,615 words for MSCOCO and Flickr30k, respectively. All the reported results are computed using Microsoft COCO caption evaluation tool [1], including the metrics BLEU(B@N), Meteor(M), Rouge-L(R) and CIDEr(C).

4.2 Protocol

As shown in Figure 1, we take the output of the 2048-d *pool5* layer from ResNet-101 as image feature I. Both the sentence-level policy and the vision-language reward are the visual-semantic embedding networks. We adopt the same architecture for them, but train them independently. As shown in Figure 2, we use one LSTM unit with 2048-d hidden layers to construct RNN, and the dimension of both linear mapping layers is set to 2048×512.

We use two types of RNN-based captioning models to construct the word-level policy. **1) CNN-RNN model**. It is introduced in Section 3.2. **2) Attention model**. It is similar to [Xu *et al.*, 2015]. We encode the image by spatially adaptive max-pooling and then the size of output is 14×14×2048. At each time step, the attention model enables LSTM decoder to emphasize features from spatial regions depending on the current context.

In training, the LSTM hidden, image, word and attention embedding dimension are fixed to 512 for the word-level policy. We use Adam optimizer with an initial learning rate of 5×10^{-5} and minibatches of size 64. The maximum number of epochs is 30. The margin λ in Eq. 5, β in Eq. 9, and γ in Eq. 4 are set as 0.6, 0.6, and 0.2, respectively. In testing, beam search is set to 1. All experiments are implemented by PyTorch.

4.3 Comparing with State-Of-The-Art Methods

In this paper, we use CNN-RNN or Attention captioning models to construct the word-level policy, denoted by Ours-CNN-RNN or Ours-Attention, respectively. For fair comparison, we only provide the results of these two types of models and existing RL models in Table 1. Our framework consistently achieves competing performances against state-of-the-art methods across all metrics.

Particularly, we discuss it in three parts. **1) CNN-RNN models**. Ours-CNN-RNN is based on a network similar to DeepVS and Google-NIC. The significant improvement over them shows the advantages of multi-level policy and reward. ATT-CNN+LSTM and MSM@MSRA use explicit high-level attributes, and m-RNN uses external data to prove its unique transfer capacity. Even though our method performs better than them. **2) Attention models**. Ours-Attention is based on a network similar to Hard-Attention, ERD, and ATT-FCN. The results show we can further improve performances by a large margin. Comparing to Adaptive, that proposes a better attention model to decide whether to attend to the image or to the visual sentinel, Ours-Attention achieves better results, which confirms the effect of our framework. **3) RL models**. MIXER is a BLEU@4-driven RL method. It is hard to generalize to other metrics while our method performs well in all metrics. Especially, Decision-Making also uses both policies, i.e., "policy network"(CNN-RNN) and "value network", also with the visual-semantic embedding reward. However, Ours-CNN-RNN performs better, which validates the effect of the multi-level reward function. Meanwhile, Ours-Attention outperforms SCST that trains the policy network (Attention) also with the CIDEr reward. Additionally, Ours-Attention performs better than Ours-CNN-RNN. It illustrates other powerful mechanisms can be directly integrated into our word-level policy and further improve the performance.

Figure 3 shows the values of reward and loss during training. We observe that both reward curves gradually increase along with the iteration. For the language-language reward, it indicates the generated captions is becoming more and more similar to the ground truth. For the vision-language reward, it is that the distance between the semantics of the generated caption and visual content is becoming closer and closer. Meanwhile, both loss curves are gradually to converge, and stable during training. Note that the word-level policy converges faster than the sentence-level policy. It is because word confidence and sentence confidence are often asynchronous, i.e., changes in word is more sensitive than sentence.

[1] https://github.com/tylin/coco-caption

	Models	MSCOCO					Flickr30k				
		B@2	**B@3**	**B@4**	**M**	**C**	**B@2**	**B@3**	**B@4**	**M**	**C**
CNN-RNN	DeepVS [Karpathy and Fei-Fei, 2017]	0.450	0.321	0.230	0.195	0.660	0.369	0.240	0.157	0.153	0.247
	Google-NIC [Vinyals *et al.*, 2015]	0.451	0.304	0.203	-	-	0.423	0.277	0.183	-	-
	m-RNN [Mao *et al.*, 2015]	0.490	0.350	0.250	-	-	0.410	0.280	0.190	-	-
	ATT-CNN+LSTM [Wu *et al.*, 2016]	0.560	0.420	0.310	0.260	0.940	0.550	0.400	0.280	-	-
	MSM@MSRA [Yao *et al.*, 2017]	0.565	0.429	0.320	0.251	0.986	-	-	-	-	-
Attention	Hard-Attention [Xu *et al.*, 2015]	0.504	0.357	0.250	0.230	-	0.439	0.296	0.199	0.185	-
	ERD [Yang *et al.*, 2016]	-	-	0.298	0.240	0.895	-	-	-	-	-
	ATT-FCN [You *et al.*, 2016]	0.537	0.402	0.304	0.243	-	0.460	0.324	0.230	0.189	-
	Adaptive [Lu *et al.*, 2017]	0.580	0.439	0.332	**0.266**	1.085	0.494	0.354	0.251	0.204	0.531
RL	MIXER [Ranzato *et al.*, 2016]	-	-	0.290	-	-	-	-	-	-	-
	Decision-Making [Ren *et al.*, 2017]	0.539	0.403	0.304	0.251	0.937	-	-	-	-	-
	SCST* [Rennie *et al.*, 2017]	-	-	0.313	0.260	1.013	-	-	-	-	-
Ours	Ours-CNN-RNN	0.601	0.449	0.330	0.252	1.042	0.562	0.408	0.282	0.219	0.586
	Ours-Attention	**0.619**	**0.464**	**0.340**	**0.266**	**1.109**	**0.575**	**0.416**	**0.289**	**0.225**	**0.615**

Table 1: Performance of the proposed framework on MSCOCO and Flickr30k test splits. (-) indicates unknown scores. (*) indicates we only compare with the single model reported in that paper for fair comparison.

Variant ID	word-level policy	sentence-level policy	vision-language reward	language-language reward	guidance term	**B@4**	**M**	**R**	**C**
1	✓					0.295	0.239	0.528	0.921
2	✓		✓	✓		0.318	0.249	0.542	1.009
3	✓	✓		✓		0.311	0.246	0.541	1.007
4	✓	✓	✓			0.315	0.247	0.539	0.996
5	✓	✓	✓	✓		0.323	0.249	0.544	1.011
Ours	✓	✓	✓	✓	✓	**0.330**	**0.252**	**0.547**	**1.042**

Table 2: Comparison of variants for the proposed framework on MSCOCO. (✓) indicates "used". (None) indicates "removed".

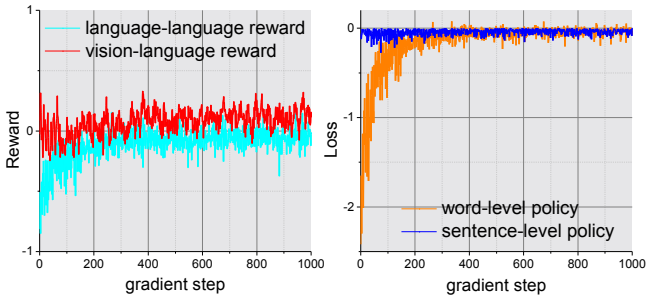

Figure 3: Performance of our framework on the validation set during the first 1,000 gradient steps.

4.4 Ablation Studies of Proposed Framework

In this section, we examine the efficacy of the proposed method by assessing several variants. We divide our framework into five components, i.e., word-level policy, sentence-level policy, vision-language reward, language-laguage reward, and guidance term. Table 2 presents five variants on Ours-CNN-RNN with respect to different IDs. The mark "✓" stands for "used" while "None" is "removed". For example, variant 3 measures the effect of vision-language reward by removing this component, where λ in Eq. 5 is set to 0. To explicitly validate the contribution of each component, we remove the guidance term in all variants.

Our method outperforms all the variants. For clarity, we discuss the results in Table 2 from Ours to variant 1. First, our method performs better than variant 5, which concludes the proposed guidance term can further bridge the policy and the reward modules. Then, we compare variant 5 with variants 2∼4 that only use single-level policy and reward. The improved results shows the merit of multi-level policy and reward in RL. Finally, variant 1 is a supervised learning baseline which results fall behind others slightly.

4.5 Qualitative Results

We show some qualitative results generated by our method and variants in Table 2. As shown in Figure 4, we can observe three points. **1) single-level vs. multi-level policies**. Variants 2∼4 use single-level policy and reward while variant 5 uses multi-level ones. The results show the later can generate competing descriptions against ground truth while the former often miss key information. For examples, variant 5 generates the phrases *wine glasses* and *talking on a cell phone* in case(b) and (f), respectively, while variant 3 and 4 miss these key words. It shows the merit of the proposed multi-level framework. **2) the effect of guidance term**. Thanks to the guidance term, our method can further improve the descriptions on variant 5, and perform better at recognizing explicit objects, e.g., the *hill* other than *field* in case(a) and the *around* other than *at* in case(b). Further, the generated captions by our method are more similar to ground truth than others, e.g., we generate the word *people* that cannot be captured by others in case(g). **3) some failure cases**. We show two failure cases in the last column, where all methods fail to understand the important visual contents, i.e., *holding an umbrella* in case(d) and *video game* in case(h). It is because our policy cannot capture the explicit objects under the noisy background. Adding more detailed visual modeling techniques such as detection can alleviate such problem in the future.

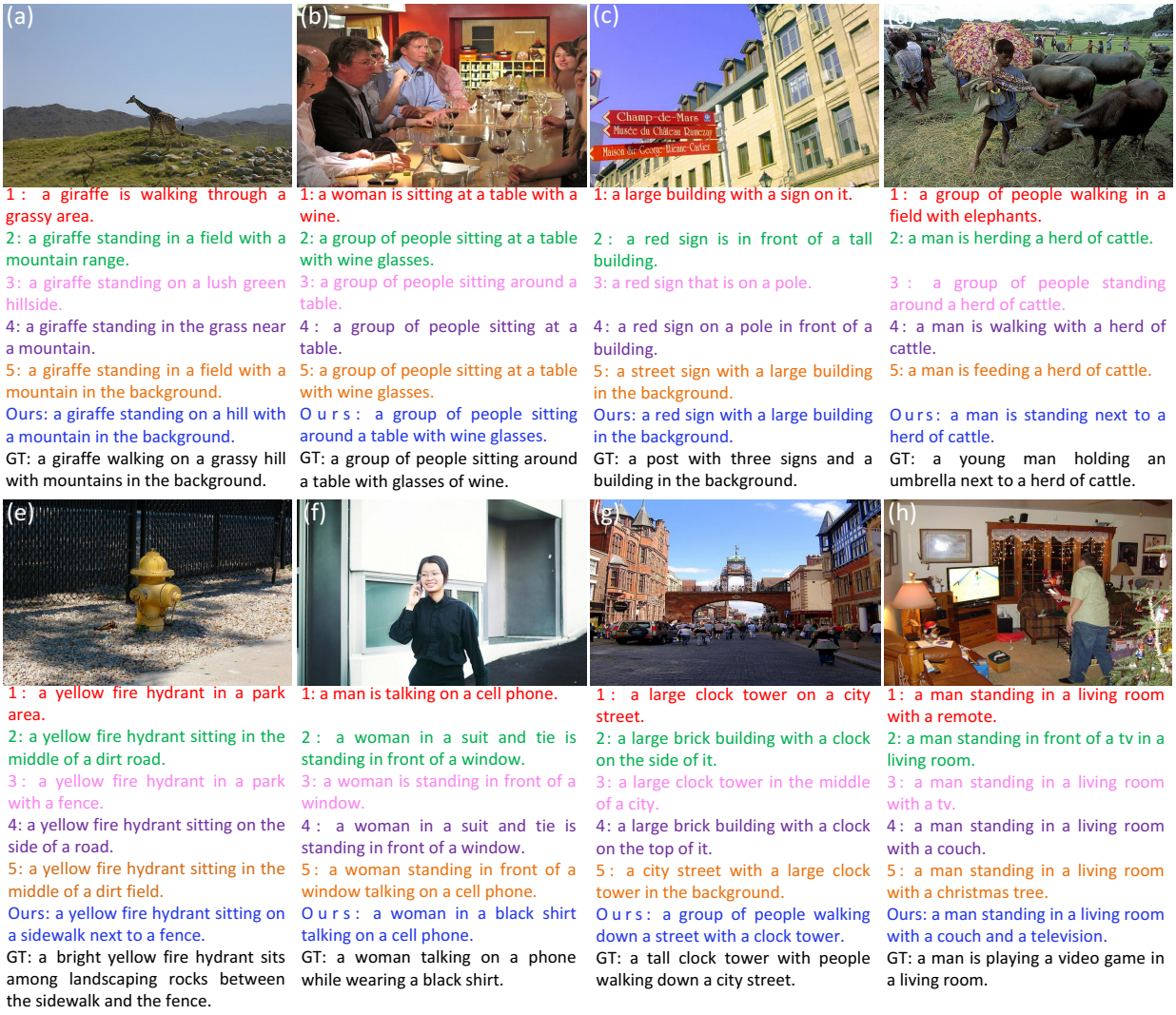

Figure 4: Qualitative results of the proposed framework on MSCOCO. The output sentences are generated by our method and the variants 1∼5 (Table 2). GT stands for the randomly selected ground truth.

5 Conclusion

In this work, we present a multi-level policy and reward reinforcement learning framework for image captioning. Different from previous methods, it explores the multi-level (word and sentence) and multi-modal (vision and language) nature of image captioning task. Particularly, the multi-level policy network can adaptively fuse the word-level and the sentence-level policies for word generation, and the multi-level reward function collaboratively leverages the vision-language and the language-language rewards to guide the policy. For the optimization, we propose a guidance term to further bridge the policy network and the reward function. Our framework achieves competing performances against state-of-the-art methods on MSCOCO and Flickr30k. Further, we explore the effect of each component by variants of the framework. In the future, we plan to investigate the multi-agent algorithm to learn the policy for image captioning.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (61772359, 61525206, 61472275, 61502337), the Beijing Advanced Innovation Center for Imaging Technology under Grant BAICIT-2016009.

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 visual question answering. In *arXiv preprint arXiv:1707.07998*, 2017.

[Cheng *et al.*, 2018] Zhiyong Cheng, Ying Ding, Lei Zhu, and Mohan S. Kankanhalli. Aspect-aware latent factor model: Rating prediction with ratings and reviews. In *WWW*, pages 639–648, 2018.

[Dai *et al.*, 2017] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descrip-

tions via a conditional GAN. In *ICCV*, pages 2989–2998, 2017.

[Das *et al.*, 2017] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, pages 1080–1089, 2017.

[Farhadi *et al.*, 2010] Ali Farhadi, Seyyed Mohammad Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David A. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, pages 15–29, 2010.

[Karpathy and Fei-Fei, 2017] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676, 2017.

[Karpathy *et al.*, 2014] Andrej Karpathy, Armand Joulin, and Fei-Fei Li. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, pages 1889–1897, 2014.

[Kiros *et al.*, 2015] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. 2015.

[Li *et al.*, 2017] Linghui Li, Sheng Tang, Lixi Deng, Yongdong Zhang, and Qi Tian. Image caption with global-local attention. In *AAAI*, pages 4133–4139, 2017.

[Liu *et al.*, 2017a] An-An Liu, Ning Xu, Yongkang Wong, Junnan Li, Yuting Su, and Mohan S. Kankanhalli. Hierarchical & multimodal video captioning: Discovering and transferring multimodal knowledge for vision to language. *CVIU*, 163:113–125, 2017.

[Liu *et al.*, 2017b] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *ICCV*, pages 873–881, 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. In *CVPR*, pages 3242–3250, 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 *ICLR*, 2015.

[Nie *et al.*, 2013] Liqiang Nie, Meng Wang, Yue Gao, Zheng-Jun Zha, and Tat-Seng Chua. Beyond text QA: multimedia answer generation by harvesting web information. *IEEE Trans. Multimedia*, 15(2):426–441, 2013.

[Pan *et al.*, 2016] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *CVPR*, pages 4594–4602, 2016.

[Ranzato *et al.*, 2016] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *ICLR*, 2016.

[Ren *et al.*, 2017] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. Deep reinforcement learning-based image captioning with embedding reward. In *CVPR*, pages 1151–1159, 2017.

[Rennie *et al.*, 2017] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, pages 1179–1195, 2017.

[Silver *et al.*, 2016] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[Sutton *et al.*, 1999] Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, pages 1057–1063, 1999.

[Vedantam *et al.*, 2015] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 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.

[Wu *et al.*, 2016] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony R. Dick, and Anton van den Hengel. What value do explicit high level concepts have in vision to language problems? In *CVPR*, pages 203–212, 2016.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. 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. In *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 *CVPR*, pages 4651–4659, 2016.

[Yun *et al.*, 2017] Sangdoo Yun, Jongwon Choi, Youngjoon Yoo, Kimin Yun, and Jin Young Choi. Action-decision networks for visual tracking with deep reinforcement learning. In *CVPR*, pages 1349–1358, 2017.

[Zhang *et al.*, 2017] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *CVPR*, pages 3107–3115, 2017.