Meta Learning for Image Captioning

Nannan Li,[1] Zhenzhong Chen,[1*] Shan Liu[2]
[1]School of Remote Sensing and Information Engineering, Wuhan University, China
[2]Tencent Media Lab, Palo Alto, CA, USA
{live, zzchen}@whu.edu.cn, shanl@tencent.com

Abstract

Reinforcement learning (RL) has shown its advantages in image captioning by optimizing the non-differentiable metric directly in the reward learning process. However, due to the reward hacking problem in RL, maximizing reward may not lead to better quality of the caption, especially from the aspects of propositional content and distinctiveness. In this work, we propose to use a new learning method, meta learning, to utilize supervision from the ground truth whilst optimizing the reward function in RL. To improve the propositional content and the distinctiveness of the generated captions, the proposed model provides the global optimal solution by taking different gradient steps towards the supervision task and the reinforcement task, simultaneously. Experimental results on MS COCO validate the effectiveness of our approach when compared with the state-of-the-art methods.

Introduction

Image captioning, which aims to automatically generate descriptions of a given image, is a prominent research problem in computer vision (Farhadi et al. 2010; Fang et al. 2015; Vinyals et al. 2015). It's a challenging task because it requires advanced techniques of object recognition and natural language processing, in order to translate an image into human-like description accurately.

Based on a CNN-LSTM structure (Vinyals et al. 2015; Mao et al. 2015; Karpathy and Fei-Fei 2015), significant progress has been made in recent years, especially by using reinforcement learning (RL) (Sutton and Barto 1998). Reinforcement learning has been exploited as a training method to deal with the out-of-context problem (Choi, Torralba, and Willsky 2008). For implicit optimization towards the evaluaition metric, (Rennie et al. 2017) train the model directly on non-differentiable metrics by using test-time reward as the baseline in the target function. For better estimation of the reward function, (Liu et al. 2017) use a linear combination of different evaluation metrics and expoit Monte Carlo rollouts instead of mixing Maximum Likelihood Estimation (MLE) training with policy gradient (Ranzato et al. 2016). (Ren et al. 2017) present a policy network and a value network using an actor-critic reinforcement learning model, where the reward is learned instead of pre-defined using the evaluation metrics .

RL: a little girl holding a cat in a of a. (1.48)
GT: girl with a yellow shirt holding a small cat. (1.39)
occupy

RL: a red fire hydrant in the snow with a. (2.57)
GT: a fire hydrant and sign covered in snow. (2.10)
occupy

RL: a street sign on a pole with a in the. (1.21)
GT: a no bicycles, skates or skateboards sign on a pole. (0.94)

RL: a bathroom with a toilet and a sink and a. (1.80)
GT: a modern looking bathroom has a toilet and a sink. (1.05)

Figure 1: Comparison of some ground truth captions (i.e., GT) and the RL generated captions using CIDEr optimization (i.e., RL). Digit next to the caption is its corresponding CIDEr score, where RL generated captions achieve significantly high CIDEr score but not necessarily better quality.

Although RL has been proven to achieve significant performance improvements on the evaluation metrics (Rennie et al. 2017; Anderson et al. 2018; Yao et al. 2017), the model may overfit to the reward function,which causes reward hacking (Irpan 2018). Specifically, given a reward as the objective function, RL can maximize the reward but the increasing reward may not come from the intended solution. Since the reward function may not accurately represent

*Corresponding Author

a caption's quality, some incorrect expressions may have higher reward than the correct ones. As shown in Figure 1, captions generated by RL have significantly higher CIDEr scores even than the ground truth, but not with necessarily better quality, especially with regard to propositional content and distinctiveness. We will elaborate on this issue in the next section.

In this paper, we propose to use a new learning method, meta learning (Finn, Abbeel, and Levine 2017), to ensure the propositional correctness and distinctiveness of the generated captions whilst optimizing the evaluation metrics. Specifically, a meta model is built to maximize the probability of the ground truth caption (i.e., supervision task) as well as maximize the reward of the generated caption (i.e., reinforcement task). In our approach, different gradient steps are taken to learn these two tasks, simultaneously, which enables the meta model to adapt to the global optimal solution of each task. The optimization of the reward function is thus guided to avoid reward hacking to some extent, and thus ensures the propositional content and distinctiveness of the generated captions.

It's worth mentioning that our method is generic and can be built upon any CNN-LSTM based captioning model. In our experiments, we adopt state-of-the-art captioning model in (Anderson et al. 2018) as a high-performing baseline. Additionally, the idea of using meta learning to utilize supervision in RL is also applicable to other RL related tasks (Ramakanth and Mohit 2017; Hongyu 2015).

Background

A large source of problem in RL stems from the difficulty in designing a proper reward function (Irpan 2018). In image captioning, caption with good quality should be correct in both content and grammar, furthermore, should be distinctive (Dai 2017). However, the evaluation metrics simply compute the generated caption's n-gram overlapping or object-relation overlapping with the ground truth, which can not ensure the propositional content and uniqueness of the generated caption. Optimizing the evaluation metrics alone can make the model overfit to achieve high score.

In fact, after empirical studies, we find RL generated captions using CIDEr optimization (Rennie et al. 2017) have abnormal patterns: some sentences tend to end with "*prep.*+a", as shown in Figure 2. This is caused by the design of the evaluation metric: CIDEr gives punishment to sentence that is too short and puts less weight on phrase that is too common. When the model generates short caption, RL enforces it to add less-weighted but commonly-used phrases to avoid punishment whilst hitting correct pairs. Therefore, the model may achieve high evaluation score, but produce captions with poor quality.

Such problem is called reward hacking in RL (Irpan 2018), meaning that out-of-the-box solution (i.e., weired sentence endings) gives more reward than the intended answer (i.e., normal sentence endings). Attempts have been made to alleviate the reward hacking problem in CIDEr optimization, by importing other evaluation metric as an additional reward term, especially SPICE (Liu et al. 2017). SPICE (Anderson et al. 2016) focuses on semantic propositional content and performs evaluation on scene graph. The abnormal endings are viewed as an unmatched object-relation pair in scene graph and thus is punished in SPICE. As shown in Table1, compared with optimizing CIDEr, maximizing CIDEr+SPICE has higher SPICE score and lower CIDEr score, indicating that optimizing SPICE helps improving grammatical and propositional content. However, SPICE has its own reward hacking problem because it does not punish repeated tuples in a scene graph. Technically, it's hard to design a perfect evaluation metric that considers every aspect of the intended goal.

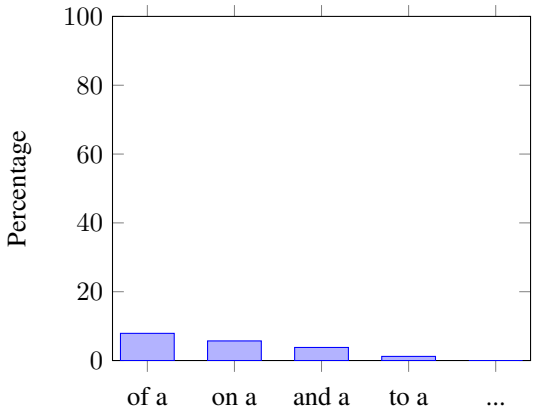


Figure 2: Examples of some abnormal sentence endings of RL generated captions on MS COCO test set using CIDEr optimization. Percentage means the ratio of sentences with the corresponding sentence ending.

Table 1: Performance on SPICE and CIDEr using different reward functions. None means MLE method without optimization of the evaluation metric.

Reward Function	SPICE	CIDEr
None	20.3	110.2
CIDEr	21.0	**120.9**
CIDEr+SPICE	**21.3**	120.4

To alleviate the reward hacking problem, supervision from the ground truth is necessary to guide the reward learning process such that the model optimizes towards the intended direction. In other words, we aim to improve the propositional content and distinctiveness of the generated captions whilst maintaining the evaluation scores.

Related Work

Based on Maximum Likelihood Estimation (MLE), methods with various model architectures have been proposed in image captioning. (Xu et al. 2015) use attention to highlight corresponding image area for each word in the caption. (Wang et al. 2017) decompose a caption to a skeleton sentence and its attributes, which are processed in different streams. In (Lu et al. 2017), a sentinel gate is designed to change attention adaptively and suppress visual attention for non-visual words. (Anderson et al. 2018) use an attention

cell and a language cell to process the visual and linguistic information separately.

To give a detailed explanation on MLE, given image I and its caption $S = w_1, w_2, ..., w_N$, MLE aims to maximize the probability of S conditioned on I:

$$p(S|I) = \sum_{t=1}^{N} \log p(w_t|I, w_{1,...}w_{t-1}) \quad (1)$$

where $p(w_t|I, w_{1,...}w_{t-1})$ is the probability of word w_t given image I and previous words $w_{1,...}w_{t-1}$.

MLE produces a description word by word in a supervised manner. However, it can not directly optimize the evaluation metric, which is non-differentiable. Regarding this issue, RL based methods are proposed in image captioning (Liu et al. 2017; Rennie et al. 2017; Ren et al. 2017). Given caption S, RL can optimize any non-differentiable reward function by maximizing its expected value $E[r(S)]$:

$$E[r(S)] = E[\sum_{t=1}^{N} r(w_t) \log p(w_t)] \quad (2)$$

As explained in the second Section, due to the reward hacking problem, RL generated captions have higher reward but not necessarily better quality, especially with regard to propositional correctness and distinctiveness. Prior to this work, some relevant ideas about this issue have also been explored. (Dai et al. 2017) construct a discriminator in conditional GAN to evaluate distinctiveness of the caption. The distinctiveness is learned as a parameterized approximation, but the approximation accuracy is not ensured in GAN. (Liu et al. 2017) adjust the coefficients of multiple evaluation metrics for a more comprehensive reward function, but the empirically set coefficients can not ensure to be the optimal solution. In our meta learning based method, aiming to improve the propositional correctness and distinctiveness of the generated captions whilst maintaining evaluation score, the evaluation metrics are directly optimized with proper supervision from the ground truth captions. And the global optimal solution is provided by taking different gradient steps towards the supervision task and the reinforcement task.

Approach

Generally, we want the model to directly optimize the evaluation metric and optimize towards the intended goal, which means supervision towards the goal is necessary in RL. However, using MLE equipped with RL directly has no benefits on either task because each task is independent and has different gradient directions. As shown in Figure 3, adding up their loss (i.e., $\lambda\nabla L_1(\theta) + \nabla L_2(\theta)$) causes the gradient to move towards a direction in between (the brown arrow), which can not ensure a optimal solution for either task.

Now we first introduce the basic concept of meta learning. Meta learning is to learn a meta model that can be transferred quickly to multiple different tasks and learns the optimal point to adapt to these tasks (Finn, Abbeel, and Levine 2017). Two gradient steps are taken to update parameter θ: the first step is to adapt parameter θ to different tasks (the black arrows in Figure 3) and to calculate their loss L_1, L_2,

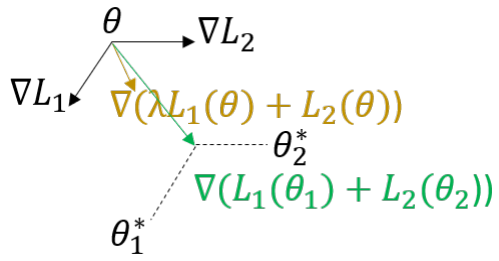


Figure 3: Illustration of the proposed meta learning model and the MLE+RL model. $(\theta_1, \nabla L_1)$, $(\theta_2, \nabla L_2)$ are (*parameter*, *gradient*) for MLE and RL task, respectively. θ is the parameter of the meta model. The meta learning model learns θ that is optimal to adapt to both tasks (marked in green) after one gradient step, whereas MLE+RL takes a gradient step in between (marked in brown).

but θ itself is not updated; the second step is to update θ itself, which is called "meta update" (the green arrow in Figure 3). In the following sections, we first introduce detailed techniques of meta learning, and then explain its formulation in our model.

Model Agnostic Meta Learning

Model agnostic meta learning trains a model's parameters to adapt to multiple tasks within a few gradient steps (Finn, Abbeel, and Levine 2017). Let θ be the model's parameter and T_i be the i_{th} task, θ adapts to T_i after one gradient step:

$$\theta_i' = \theta - \alpha\nabla_\theta L_{T_i}(\theta) \quad (3)$$

where θ_i' and L_{T_i} is the parameter and loss for task T_i, respectively, and α is the step size. The meta-objective is to minimize loss L_{T_i} using the adapted parameter θ_i' for T_i with respect to θ:

$$\min_\theta \sum_i L_{T_i}(\theta_i') = \min_\theta \sum_i L_{T_i}(\theta - \alpha\nabla_\theta L_{T_i}(\theta)) \quad (4)$$

$$\theta \leftarrow \theta - \beta\nabla_\theta \sum_i L_{T_i}(\theta_i') \quad (5)$$

where β is the step size. Generally, the meta parameter θ is adapted once in Equation (3) and updated once in Equation (5), with no extra parameters imported. Note that the gradient descent in Equation (5) is performed over θ, thus back propagation of a second derivative is required. To reduce the computational cost, we use a first-order approximation instead (Finn, Abbeel, and Levine 2017).

Meta Learning for Image Captioning

We define two tasks T_1, T_2 for the captioning model: 1) maximizing the probability of the ground truth caption (i.e., MLE), which is a supervision task; 2) maximizing the reward of the generated caption (i.e., RL), which is a reinforcement task. By achieving these two tasks, the learned meta model is sensitive to both loss and is able to optimize the evaluation metric with supervision from the ground truth caption. In the following, we use θ to denote

the parameter of the meta model, and θ'_1, θ'_2 are the adapted parameters for the corresponding tasks. w_t is the t_{th} word in sentence S and $p(w_t)$ denotes its probability.

Maximizing probability. Instead of using standard MLE method to maximize $p(S|I)$ as in Equation (1), a baseline model is imported to reduce variance of a batch and to encourage distinctiveness of the generated caption (Dai 2017). The target and the baseline model are both pretrained using standard MLE loss, which outputs $p(S|I)$ and $p_b(S|I)$, respectively. Then the baseline model is fixed to give stable output $p_b(S|I)$. $p(S|I)$ is then further maximized with $p_b(S|I)$ as its baseline. We use S^+ to represent the paired caption, and S^- to represent the unpaired caption of image I. Given positive image-caption pairs (S^+, I) and negative pairs (S^-, I), $p(S^+|I)$ is maximized whilst $p(S^-|I)$ is minimized with relative to the baseline model:

$$G((S, I)) = \sigma(\log p(S|I) - \log p_b(S|I)) \tag{6}$$

$$L_1 = -\log(G(S^+, I)) - \log(1 - G(S^-, I)) \tag{7}$$

where σ indicates sigmoid function. Same as (Dai 2017), negative samples are included in the loss function to encourage distinctiveness of the generated captions. Since $p_b(S|I)$ is fixed, the gradient of this task only involves $p(S|I)$:

$$\nabla_{\theta'_1} L_1(\theta'_1) = -\frac{1}{\nabla_{\theta'_1} G(S^+, I)} - \frac{1}{1 - \nabla_{\theta'_1} G(S^-, I)} \tag{8}$$

$$\nabla_{\theta'_1} G(S, I) = \nabla_{\theta'_1} \sigma(\log(p_{\theta'_1}(S|I))) \tag{9}$$

Maximizing reward. Following (Rennie et al. 2017), a baseline reward b is imported to Equation (2) to reduce variance of the gradient estimate without changing its expected value, which is calculated by greedy sampling:

$$E[r(S)] = E[\sum_{t=1}^{N} (r(w_t) - b) \log p(w_t)] \tag{10}$$

Using policy gradient method, the gradient of this task $\nabla_{\theta'_2} L_2(\theta'_2)$ should be:

$$\begin{aligned} \nabla_{\theta'_2} L_2(\theta'_2) &= -\nabla_{\theta'_2} E[r(S)] \\ &= -E[\sum_{t=1}^{N} (r(w_t) - b) \nabla_{\theta'_2} \log p_{\theta'_2}(w_t)] \end{aligned} \tag{11}$$

The reward function $r(S)$ is CIDEr+SPICE in our model, considering both n-gram overlapping and object-relation overlapping with the ground truth captions.

Meta learning. The meta learning steps include update to task-specific parameter (i.e., θ'_i) and update of meta parameter (i.e., θ), as in Equation (12) and Equation (13), respectively.

$$\theta'_i = \theta - \alpha \nabla_\theta L_i(\theta), i = 1, 2 \tag{12}$$

$$\theta \leftarrow \theta - \beta \nabla_\theta (\lambda L_1(\theta'_1) + L_2(\theta'_2)) \tag{13}$$

where $L_i(\theta)$ is loss for the i_{th} task with parameter θ, and α/β is the step size. λ is a constant value between 0 to 1, deciding the ratio of these two tasks in the meta model. Supervision from the ground truth gets stronger with larger λ. Note that we sample twice of the caption in the second task when taking one gradient step of the meta model: the first sampling for adapting parameter θ to θ'_2 in Equation (12), and the second sampling for meta-updating θ in Equation (13). Different from θ'_1, θ'_2 is updated using supervision from the ground truth instead of sampling.

By adapting the meta model to maximize the probability of the ground truth caption as well as maximize the reward of the generated caption in different gradient directions, its reward learning process is guided by supervision of the ground truth labels. Thus the generated caption can present both high evaluation score and good quality.

Captioning model. Since the meta learning method only modfiies the loss function, it is universal for all sorts of model architectures for captioning, including the models we referred to in Related Work. We choose the state-of-the-art architecture in (Anderson et al. 2018) for high-performing baseline, which consists of an attention LSTM and a language LSTM. Let h_t^1, h_t^2 denote the hidden state of language LSTM and attention LSTM at time t, respectively, their formulations are given as follows:

$$h_t^1 = \text{LSTM}([x_t, \bar{v}, h_{t-1}^2], h_{t-1}^1) \tag{14}$$

$$h_t^2 = \text{LSTM}([\hat{v}_t, h_t^1], h_{t-1}^2) \tag{15}$$

where $\text{LSTM}(x, h_t)$ is the LSTM function with input x and hidden state h_t. x_t be the word embedding of w_t. Given image features of N objects $v = v_1, v_2 ... v_N$, $\bar{v}$ is the average-pooled image feature and $\hat{v}_t$ is the attention-driven image feature, whose attention is calculated using soft attention (Xu et al. 2015):

$$\alpha_{i,t} = \text{softmax}(W_a^T \tanh(W_{av} v_i + W_{ah} h_t^1)) \tag{16}$$

$$\hat{v}_t = \sum_i \alpha_{i,t} v_i \tag{17}$$

where W_a^T, W_{av} and W_{ah} are learned weights. $\alpha_{i,t}$ is the attention value of object i at time t. With learned matrix W_p, the model outputs a probability distribution of the next word w_t using softmax function:

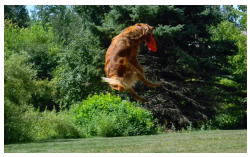
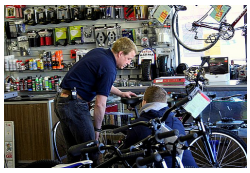
$$p(w_t) = \text{softmax}(W_p h_t^2) \tag{18}$$

When maximizing the probability of the ground truth caption (i.e., MLE), w_t is the ground truth word and $p(w_t)$ is maximized. However, when maximizing the reward of the generated caption (i.e., RL), w_t is sampled according to distribution $p(w_t)$ for higher reward, regardless of the ground truth. By learning both tasks at the same time using meta learning, the meta model has higher probability to sample the ground truth caption since it's ensured in MLE task. With supervision from the ground truth captions, RL can explore proper vocabularies in the vicinity of the ground truth, thus alleviating the reward hacking problem to some extent.

Experiments

To validate the effectiveness of our model, we conduct experiments on MSCOCO (Lin et al. 2014) dataset with

Table 2: Examples of the generated captions on MS COCO test split. The first four rows are successful cases whilst the bottom row is a typical failed one. Number in the bracket is the corresponding evaluation scores (i.e., CIDEr+SPICE).

Image	Generated Captions	Ground Truth
	MLE: a bathroom with a walk in shower and a glass shower. (0.72) RL: a bathroom with a toilet and a sink in a. (1.76) MLE+RL: a bathroom with a toilet and a sink and a. (1.87) Ours: a bathroom with a toilet and a sink and a shower. (2.46)	1. a clean, spacious bathroom with a large shower stall. 2. there are a toilet, a sink, and a shower stall in a large bathroom. 3. a bathroom with an enclosed shower next to a sink and a toilet. 4.bathroom with a shower, sink, and toilet in it. 5.a bathroom featuring a walk in shower, mirror, sink and toilet
	MLE: a horse drawn carriage on a street with a horse. (1.08) RL: a horse drawn carriage on a street.(3.25) MLE+RL: a group of horses standing next to a building. (1.47) Ours: two horses pulling a carriage in front of a building. (1.15)	1. a pair of horses carrying a carriage that is parked by a street. 2. a horse drawn carriage is in front of an old large building. 3. people walking pass a horse drawn carriage sitting at the curb. 4. a horse drawn carriage parked on the street. 5. a horse drawn carriage is parked along the curb.
	MLE: a dog is running with a frisbee in the grass. (1.97) RL: a dog jumping in the air to catch a frisbee. (1.85) MLE+RL: a dog jumping in the air to catch a frisbee. (1.85) Ours: a dog jumping to catch a frisbee in the grass. (2.27)	1. a dog in the grass catching a frisbee. 2. a tan dog leaping to catch a frisbee. 3. a dog is opening his mouth to catch a frisbee. 4. a yellow dog runs to grab a yellow frisbee in the grass. 5. a dog is on the grass playing frisbee.
	MLE: a dog jumping up catch a frisbee in its mouth. (2.73) RL: a dog jumping in the air to catch a frisbee. (3.29) MLE+RL: a dog jumping to catch a frisbee in the air. (2.72) Ours: a dog jumping in the air with a frisbee in its mouth. (4.21)	1. a dog leaping in the air to catch a frisbee. 2. a dog jumping high in the air catching a frisbee in its mouth. 3. a brown dog jumping in the air and catching a frisbee. 4. a dog jumping high in the air with a frisbee in its mouth. 5. a brown dog flying through the air with a red frisbee in his mouth.
	MLE: a man standing in a room with a bike. (1.04) RL: a man standing next to a bike in a store. (1.21) MLE+RL: a man standing with a bike in a store. (1.19) Ours: a man standing in a store with a bicycle. (1.06)	1. the bike shop employee is helping a customer. 2. a bicycle store shows two males leaning toward a bike. 3. a man and a boy are talking about a bicycle in a store. 4. a man adjust a bicycle in a bike shop with a child. 5. two people in a shop looking at a bike.

123,287 labeled images. Each image has 5 human annotated captions as reference. We use public available splits (Karpathy and Fei-Fei 2015) which have 5000 randomly selected images for validation and test. Our vocabulary size is fixed to 10,000 including special start sign <BOS>and end sign <EOS>. With Faster R-CNN (Ren et al. 2015) as image features, the number of proposals for each image is fixed to be 36 instead of chosen adaptively (Anderson et al. 2018) for shorter training schedule. For pre-processing, we convert all sentences to lower case and filter all punctuation except the period. Sentences that have more than twenty words or less than five words are discarded.

Implementation Details

The number of hidden nodes of our network is set to 512 for the LSTM cell, with word embedding size of 512. We use Adam optimizer (Kingma and Ba 2014) with learning rate decay and set initial learning rate $\alpha = 0.01, \beta = 5 \times 10^{-4}$. λ is set to be 0.1 since we find empiracally that it's a good value for maintaining the overall performance performance whilst ensuring propositional correctness and distinctiveness of the generated captions. We use 0.5 dropout before the last layer and feed back 0.05 sampled words every 4 epochs starting from the the $10th$ epoch until reaching a 0.25 feeding back rate (Bengio et al. 2015). We add a batch normal-

Table 3: Results of SPICE and breakdown of SPICE F-scores over various sub-categories on the MS COCO test split.

Method	SPICE	Objects	Relations	Attributes	Color	Count	Size
MLE	20.3	37.2	5.6	8.6	13.4	2.2	**4.9**
RL	21.3	38.5	5.8	9.2	14.3	12.7	3.7
MLE+RL	21.2	38.0	5.6	**9.7**	12.4	12.7	3.9
Ours	**21.7**	**39.0**	**7.4**	9.5	**14.7**	**12.9**	**4.9**

Table 4: Results of self-retrieval on MS COCO test split. R@k represents the top-k recall rate.

Method	R@1	R@5	R@50
MLE	55.6	78.5	89.5
RL	38.9	68.0	88.1
MLE+RL	42.6	71.8	88.9
Ours	**56.9**	**79.5**	**90.2**

ization layer (Ioffe and Szegedy 2015) in the beginning of the LSTM model to accelerate training with mini-batch size of 50.

Qualitative Analysis

Our aim is not to improve all the evaluation metrics because of the reward hacking problem in RL. Instead, we stress more on improving the propositional correctness and distinctiveness of the generated caption whilst *maintaining* the evaluation scores. To prove this point, we present the results in three aspects: 1) breakdown of SPICE F-scores for propositional correctness; 2) self-retrieval rate for distinctiveness of the caption; 3) model performance on the evaluation metrics. For fair comparison, instead of using meta learning, we train three baseline models for ablation studies: 1) train for task one alone using Equation (7), denoted as MLE; 2) train for task two alone using Equation (11), denoted as RL; 3) train for these two tasks together but simply add up their loss, i.e., $\nabla_\theta L = \nabla_\theta(\lambda L_1(\theta) + L_2(\theta)), \lambda = 0.1$, denoted as MLE+RL.

Some visualized examples are given in Table 2 for intuitive explanations about the propositional content and distinctiveness of the generated captions. The number at the end of the caption corresponds to its evaluation score (i.e., CIDEr+SPICE).

Propositional Correctness We show SPICE score and the accuracy of sub-categories in SPICE F-score in Table 3. By utilizing supervision from the ground truth in reward optimization, our model outperforms the baseline models on SPICE, and improves the recognition accuracy of most subcategories compared with RL and MLE+RL, especially on objects and relations. This proves that our model enhances the propositional content of the generated captions.

For visualized results, we give some examples in the first two rows in Table 2. In the first picture, RL generated caption improve the evaluation metric greatly but suffer from the reward hacking problem, manifested in abnormal sentence endings such as "with/in a". On the other hand, our model further optimizes the evaluation metric by essentially improving the propositional content of the captions. For example, the primary objects of the first picture: sink, toilet and shower, are included without a miss in our caption.

Note that in the second picture, the caption generated by our model has lower evaluation score because it has less overlapping with the ground truth captions, which do not contain *two horses* and *pulling*. However, it also describes the image faithfully and gives details such as *in front of a building*. It's a typical example that the generated caption is semantically correct but has low evaluation score. Since current evaluation metrics either depend on n-gram overlapping (e.g., BLEU,CIDEr) or object-relation overlapping (i.e., SPICE) with the ground truth captions, designing a new evaluation metric that relies on the semantic overlapping is regarded as our future work.

Distinctiveness As for the distinctiveness of the generated caption, we use self-retrieval rate as the evaluation criteria (Mao et al. 2015; Dai 2017). Self-retrieval (Mao et al. 2015) is a ranking problem which uses the generated caption of image I to retrieve I. Specifically, the generated caption S_j of I_j is paired with each image $I_m, m = 1, ..., N$ in the test set to calculate probability $p((S_j|I_m), m = 1, ..., N$. I_j is S_j's top-k recall if I_j is in the top-k of the ranked $p(S_j|I_m), m = 1, ..., N$. Higher recall rate indicates better uniqueness of the caption.

As summarized in Table 4, RL has a large performance drop on self-retrieval rate (38.9%) compared with MLE (55.6%). This suggests that captions generated by RL only focus on primary content of the image and loss details, thus are similar for images look alike. However, in order to be distinguishable, caption with good quality should describe unique content of the given image. By encouraging resemblance with the ground truth captions in the MLE task whilst optimizing the evaluation metric in the RL task, our model outperforms both MLE+RL and RL, improving R@1 to 56.9%.

For visualized results, we give examples in the middle two rows in Table 2. These two pictures look similar in that there is a dog jumping to catch a frisbee in both scenes. However, the difference is that the first frisbee is in the air whereas the second frisbee is in the dog's mouth. RL model ignores such difference and uses *a dog jumping in the air to catch a frisbee* to descibe both pictures, whereas our model pays attention to the details in the image and thus generates an accurate and distinctive caption for each scene.

Table 5: Results of the overall performance on MS COCO test split. B-n stands for BLEU-n metric. **-** represents unknown result. Methods marked with $*$ adopt RL for CIDEr optimization.

Method	B-1	B-2	B-3	B-4	METEOR	CIDEr	SPICE
ER (Ren et al. 2017)	71.3	53.9	40.3	30.4	25.1	93.7	-
Skel-Attr-LSTM (Wang et al. 2017)	74.2	57.7	44.0	33.6	26.8	107.3	19.6
Ada-ATT (Lu et al. 2017)	74.2	58.0	43.9	33.2	26.6	108.5	19.4
Ada-ATT+CL (Dai 2017)	75.5	59.8	46.0	35.3	27.1	114.2	-
ARNet (Chen et al. 2018)	74.0	57.6	44.0	33.5	26.1	103.4	19.0
$*$ 4 Att2in (Rennie et al. 2017)	-	-	-	34.8	26.9	115.2	-
$*$ Up-down (Anderson et al. 2018)	**79.8**	-	-	36.3	27.7	120.1	21.4
MLE	76.5	60.5	45.8	35.3	26.8	110.2	20.3
$*$ RL	78.9	63.2	48.3	36.5	27.5	120.4	21.3
$*$ MLE+RL	77.7	62.7	47.9	36.4	27.4	119.3	21.2
$*$ Ours	79.1	**63.9**	**49.4**	**37.5**	**27.8**	**121.0**	**21.7**

Table 6: Results on the online MS COCO test server. All metrics are reported using c5 and c40 references. SPICE is not included on the test server.

Method	c5							c40						
	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr
Ada-ATT	74.8	58.4	44.4	33.6	26.4	55.0	104.2	92.0	84.5	74.4	63.7	35.9	70.5	105.9
Ada-ATT+CL	74.2	57.7	43.6	32.6	26.0	54.4	101.0	91.0	83.1	72.8	61.7	35.0	69.5	102.9
4 Att2in	78.1	61.9	47.0	35.2	27.0	56.3	114.7	93.1	86.0	75.9	64.5	35.5	70.7	116.7
Up-down	**80.2**	**64.1**	49.1	36.9	**27.6**	**57.1**	**117.9**	**95.2**	**88.8**	**79.4**	68.5	**36.7**	**72.4**	**120.5**
Ours	79.3	63.7	**49.2**	**37.3**	27.4	**57.1**	117.4	94.7	88.2	79.0	**68.6**	36.2	71.5	119.0

Model Performance We report our results with frequently used evaluation metrics: BLEU-1,2,3,4 (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), CIDEr (Vedantam, Zitnick, and Parikh 2015) and SPICE (Anderson et al. 2016), as provided by MSCOCO (Chen et al. 2015). In Table 5, methods marked with $*$ involve RL for CIDEr optimization. Comparing MLE+RL with Ours, simply adding up the loss of each task results in worse performance because each task is independent and thus has different directions of gradient. However, by taking different directions of gradient corresponding to each task, our meta model is sensitive to both objective loss and learns the optimal solution to adapt to both tasks. As shown in Table 5, the proposed model outperforms the baseline models on all evaluation metrics. In Table 6, our model only outperforms Up-down (Anderson et al. 2018) on BLEU3 and BLEU4 because: 1) Up-down uses model ensembles whereas we use a single model; 2) Up-down optimizes CIDEr whereas our model optimizes CIDEr+SPICE with supervision, which stresses more on the object-relation overlapping instead of 1-gram or 2-gram overlapping, and BLEU3/BLEU4 is relatively more suitable to evaluate the object-relation overlapping.

Error Analysis Except for the successful examples, we also present a typical failed case in the bottom row in Table 2. In the bottom picture, the squatter next to the bicycle is missed out in the generated captions. In fact, we find the model sometimes misses or mis-recognizes objects under complex scene, especially when there are multiple similar objects in the foreground. This is partly due to the inaccuracy of CNN features, the model of which is pretrained on Visual Genome (Krishna et al. 2016) using Faster-RCNN, with 10.2% mAP@0.5. In future work, we may consider improving the CNN model to get better image features.

Conclusion

In this work, we first analyze the reward hacking problem in RL, and then propose to use a new learning method, meta learning, to alleviate the problem by utilizing supervision from the ground truth whilst optimizing the reward function. In our approach, aiming to improve the propositional content and distinctiveness of the generated captions whilst maintaining high performance towards the ground truth, the proposed model provides the global optimal solution by taking different gradient steps towards the supervision task and the reinforcement task, simultaneously. Experimental results on MS COCO validate the effectiveness of our approach when compared with the state-of-the-art methods.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61771348 and Tencent.

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. SPICE: semantic propositional image caption evaluation. In *ECCV*.

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*.

Banerjee, S., and Lavie, A. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop*.

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

Chen, X.; Fang, H.; Lin, T.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv:1504.00325*.

Chen, X.; Ma, L.; Jiang, W.; Yao, J.; and Liu, W. 2018. Regularizing RNNs for caption generation by reconstructing the past with the present. In *CVPR*.

Choi, M. J.; Torralba, A.; and Willsky, A. S. 2008. Context models and out-of-context objects. *Pattern Recognition Letters*.

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

Dai, Boand Lin, D. 2017. Contrastive learning for image captioning. In *NIPS*.

Fang, H.; Platt, J. C.; Zitnick, C. L.; Zweig, G.; Gupta, S.; Iandola, F.; Srivastava, R. K.; Deng, L.; Dollar, P.; and Gao, J. 2015. From captions to visual concepts and back. In *CVPR*.

Farhadi, A.; Hejrati, M.; Sadeghi, M. A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; and Forsyth, D. 2010. Every picture tells a story: Generating sentences from images. In *ECCV*.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*.

Hongyu, G. 2015. Generating text with deep reinforcement learning. In *NIPS*.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*.

Irpan, A. 2018. Deep reinforcement learning doesn't work yet. https://www.alexirpan.com/2018/02/14/rl-hard.html.

Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. In *ICLR*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; Bernstein, M.; and Fei-Fei, L. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*.

Lin, T. Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common objects in context. In *ECCV*.

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

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*.

Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Huang, Z.; and Yuille, A. 2015. Deep captioning with multimodal recurrent neural networks (m-RNN). In *ICLR*.

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

Ramakanth, P., and Mohit, B. 2017. Reinforced video captioning with entailment rewards. In *EMNLP*.

Ranzato, M.; Chopra, S.; Auli, M.; and Zaremba, W. 2016. Sequence level training with recurrent neural networks. In *ICLR*.

Ren, S.; He, K.; Ross, G.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*.

Ren, Z.; Wang, X.; Zhang, N.; Lv, X.; and Li, L.-J. 2017. Deep reinforcement learning-based image captioning with embedding reward. In *CVPR*.

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

Sutton, R. S., and Barto, A. G. 1998. *Reinforcement learning: An introduction*. MIT press Cambridge.

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

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Wang, Y.; Lin, Z.; Shen, X.; Cohen, S.; and Cottrell, G. W. 2017. Skeleton key: Image captioning by skeleton-attribute decomposition. In *CVPR*.

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

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