A Multi-task Learning Approach for Image Captioning

Anonymous NAACL submission

Abstract

In this paper, we propose a Multi-task Learning Approach for Image Captioning (*MLAIC*), motivated by the fact that humans have no difficulty performing such task because they possess capabilities of multiple domains. Specifically, MLAIC consists of three key components: (i) A multi-object classification model that learns rich category-aware image representations using a CNN image encoder; (ii) A syntax generation model that learns better syntax-aware LSTM based decoder; (iii) An image captioning model that generates image descriptions in text, sharing its CNN encoder and LSTM decoder with the object classification task and the syntax generation task, respectively. In particular, the image captioning model can benefit from the additional object categorization and syntax knowledge. To verify the effectiveness of our approach, we conduct extensive experiments on MS-COCO dataset. The experimental results demonstrate that our model achieves impressive results compared to other strong competitors.

1 Introduction

Humans possess the capability to describe an image verbally because they are naturally multi-task intelligent agents. Humans have developed those skills since childhood by not only learning to perform a single task, but rather adapting to comprehend the complex outer world via multi-channel perceptions and communications. They are trained in fact by performing multiple relevant tasks together to develop a comprehensive set of skills to understand and describe a scenario. If one desires to create a machine intelligence imitating such a comprehensive skill set of human, studying all relevant tasks that contribute its development is quite a necessary step.

In this paper, we are motivated by the fact that a cognitive AI is by nature multi-tasking and develop a computerized image captioning agent that can also perform a few other related tasks. Image captioning, as to generate a sentence describing the salient aspects of an image, is a fundamental task in computer vision and natural language processing (Bernardi et al., 2016). In recent years, it is often approached with supervised learning framework by collecting human generated examples and developing models that base on matching the generated text to those collected annotations. We suppose such learning frameworks are theoretically of limited success in two dimensions. First, the model created from the collected data can only comprehend the complexity of the problem subject to the degree of complexities of presented examples. Since the dataset is more or less a finite collection, the presented complexities should be limited. Second, the loss function used in numerically optimizing the model is often not sensitive to certain aspects of the structured output that have not been emphasized in the conventional evaluation metric of image captioning, such as object categories and syntax of generated sentences. In fact, we show that exposing the learning framework with more relevant data and objectives can be helpful in the both dimensions.

w/o syntax:	a man holding a tennis ball with a tennis ball.	a suitcase filled with lots of items and.
w/o object:	a man holding a tennis ball.	a suitcase filled with lots of items.
Our Model:	a man hitting a tennis ball with a tennis racket.	a suitcase filled with lots of items on a bed.

Table 1: Examples of captions that fail to respect to syntax structure and missing salient object "bed" and "racket".

Multi-task learning is hardly a new idea for machine learning, but often remains as a non-trivial step for building empirically successful systems. We argue that it is essentially the case for creating an effective image captioning system. In an ablation study of our models (see Table 1), a model which is unaware of the sentence syntax could generate a broken sentence for describing a picture; a model which does not recognize all presented objects could generate a descriptive sentence missing a salient object in the picture. Yet, an improved system whose components have been co-trained with multiple related tasks could generate a more satisfied sentence for accurately and correctly describing a picture. Based on this observation, we believe a multi-task learning framework helps to improve an image captioning system in dimensions that have not been quantitatively measured in conventional evaluation metrics.

Our system exploits the recent successes of the encoder-decoder framework to generate image captions (Donahue et al., 2015; Karpathy and Fei-Fei, 2015; Kiros et al., 2014; Vinyals et al., 2015). The common idea of this framework is to use a convolutional neural network (CNN) as an encoder to extract features representing the visual understandings of an input image and then feed the feature vector to a recurrent neural network (RNN) based decoder so as to generate image captions. Sharing this standard framework with other related approaches, we in this paper propose additional regularizations using multi-task learning. Firstly, our CNN encoder is regularized with the co-training to perform an additional task of multi-object classification. Secondly, our RNN decoder is also regularized with the co-training to perform another task of syntax annotation (Nadejde et al., 2017). The purpose of co-training is not to achieve the best performance on these auxiliary tasks, but rather to compensate for the missing regularization requirement of image captions in the standard framework.

We summarize our main contributions as follows:

- We propose *MLAIC*, a multi-task learning system to jointly train the task of image captioning and two other related tasks: multi-object classification and syntax generation. The auxiliary tasks help to enhance the CNN encoder and the RNN decoder in the image captioning model. Specifically,

    1. Multi-object classification co-trained with image captioning intends to learn an object-rich image encoder and improves the quality of locating contextual information of an image.
    2. The variations of style and wording of captions with respect to different object categories are explored under controlled experimental settings.
    3. The RNN decoder is capable to leverage word-level syntax to generate high-quality captions from language modeling perspective. It alleviates the issues of incomplete sentences and duplicated words.

- The experimental results show that *MLAIC* achieves outstanding performance on the widely used MSCOCO dataset according to both the offline Karpathy test spilt and the online server evaluation.

2 Related work

Generating image captions from images is a challenging problem that has been receiving much attention from the computer vision and natural language processing communities recent years. Bernardi et al. (2016) provided a detailed review of most existing approaches, the benchmark datasets, and the evaluation measures for image captioning.

Early methods for image captioning either explored template-based approaches (Elliott and Keller, 2013; Mitchell et al., 2012) or retrieval-based approaches (Gong et al., 2014; Kuznetsova et al., 2014). These models were usually heavily relied on hand-designed features or templates, and it is hard for them to generate novel sentences with new compositions.

Recent advances in deep neural networks have substantially improved the performance of image captioning task. A typical image captioning strategy is to combine CNN and RNN (Donahue et al., 2015; Karpathy and Fei-Fei, 2015; Kiros et al., 2014; Vinyals et al., 2015), where CNN is used to extract the compact representational vector of a whole image, and RNN is used to construct the language model operated on the representation vectors to generate captions. Visual attention has been proven as an effective way to improve the basic encoder-decoder framework. For example, Xu et al. (2015) introduced an attention based model that automatically learn where to attend when generating image descriptions. The attention is modeled as spatial probabilities that re-weight the feature map of the last convolutional layer in the CNN. Chen et al. (2016) dynamically modulated the sentence generation context in multi-layer feature maps, encoding where (i.e., attentive spatial locations at multiple layers) and what (i.e., attentive channels) the visual attention was.

There have been increasing interests in integrating the encoder-decoder framework and reinforcement learning paradigms for image captioning (Chen et al., 2017; Liu et al., 2016; Rennie et al., 2016). For example, Liu et al. (2016) employed policy gradient (PG) method to directly optimize a linear combination of SPICE and CIDEr metrics, where the SPICE score ensured the captions were semantically faithful to the image, and CIDEr score ensured the captions are syntactically fluent. Rennie et al. (2016) proposed a self-critical sequence training (SCST) method by employing the popular REINFORCE algorithm. Instead of estimating a "baseline" to normalize the rewards and reduce variance, SCST utilizes the output of its own test-time inference algorithm to normalize the rewards it experiences. Anderson et al. (2017) proposed a combined bottom-up and top-down attention mechanism that enables attention to be calculated at the level of objects and other salient image regions, and employed the REINFORCE algorithm to optimize CIDEr directly.

Multi-task learning is a useful learning paradigm to improve the supervision and the generalization performance of a task by jointly training it with related tasks (Caruana, 1998; Collobert and Weston, 2008). Recently, Luong et al. (2016) combined multi-task learning with sequence-to-sequence models, sharing parameters across the tasks encoders and decoders. They showed improvements in machine translation using parsing and image captioning. Venugopalan et al. (2016) explored linguistic improvements to the video caption decoder by fusing it with external language models. Pasunuru and Bansal (2017); Collobert and Weston (2008) improved video captioning by sharing knowledge with two related directed-generation tasks: a temporally-directed unsupervised video prediction task and a logically-directed language entailment generation task.

Our model differs from the above approaches in several aspects. First, We perform image captioning by multi-task learning, which shares knowledge with three related tasks: multi-label classification, image captioning, syntax generation, so that we can improve the performance of both the CNN encoder and LSTM decoder. Second, the CNN encoder is shared between object classification and image captions such that it is capable of recognizing the existence of object instances and focusing on different aspects of the object. In addition, we also project object labels onto distributed representations, and incorporate them into LSTM decoder to generate captions of different styles. Finally, we integrate word-level syntax into the LSTM decoder to generate high-quality captions, alleviating the problem of generating incomplete sentences and duplicate words.

3 Our Model

Given an image $\mathbf{x}$, for the object classification, we have $\mathbf{y}^{\text{o}} = \{y_1^o, y_2^o, ..., y_C^o\}$ denoting the object vector of each image, where $y_i^o = 1$ if object i is annotated in this image; otherwise $y_i^o = 0$, C is the number of object categories. For image captioning, we have $\mathbf{y}^w = \{y_1^w, y_2^w, ..., y_T^w\}$ denoting the image description given image $\mathbf{x}$, where T is the length of sequence. For syntax generation, we have $\mathbf{y}^{\text{s}} = \{y_1^s, y_2^s, ..., y_T^s\}$ denoting the CCG supertag sequence with respect to the corresponding caption of image $\mathbf{x}$. We use $\mathcal{W}_w$, $\mathcal{W}_s$ and $\mathcal{W}_o$ to denote the vocabularies of captions, annotated CCG supertags and object categories, respectively.

Our model *MLAIC*, whose framework is illustrated in Figure 1, jointly train the image captioning task with two related tasks: the object classification and the syntax generation. The object classifier shares its CNN encoder with the encoder of image captioning task. All object labels are encoded into low-dimensional distributed embeddings and treated as extra input to LSTM decoder, helping the LSTM decoder focus on different aspects of the images with respect to the object labels. The syntax generation task shares its LSTM decoder with the decoder of image captioning, which predicts words and syntax by training the shared LSTM decoder in a similar fashion as is done in (Nadejde et al., 2017). Next, we elaborate the three tasks in details.

3.1 Shared CNN encoder

We use the ResNet-101 (He et al., 2016) pre-trained on ImageNet (Russakovsky et al., 2015) as our CNN encoder to encode the input image $\mathbf{x}$ into L vectors, each of which is a D-dimensional vec-

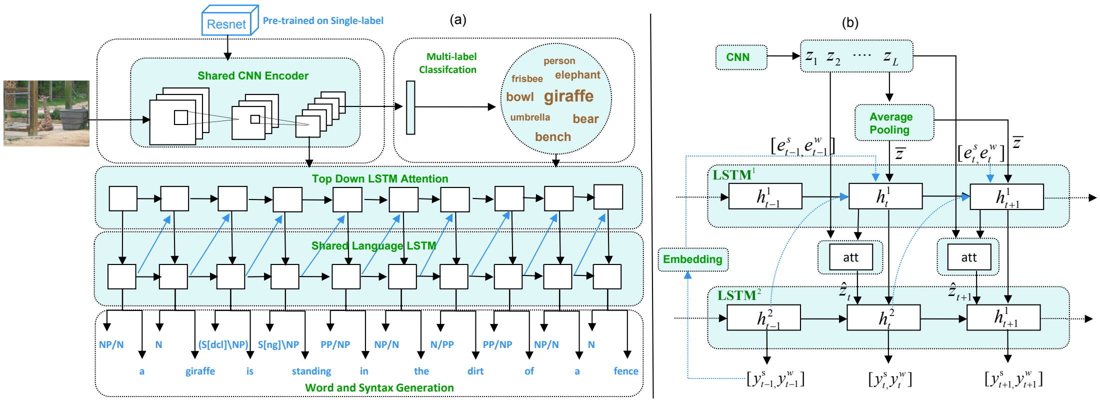


Figure 1: (a) shows the Architecture of multi-task learning on three related tasks. The font size in the ball represents the probability of object category. (b) shows the details of Top-Down LSTM attention and shared language LSTM.

tor corresponding to the features extracted at different locations of the image. These vectors are referred to as annotation vectors.

$$z = \text{CNN}(\mathbf{x}) = \{z_1, z_2, ..., z_L\} \tag{1}$$

We also introduce $\mathbf{z}^{fc}$ to represent object probability of image $\mathbf{x}$ calculated by fully connected layer: $z^{\text{fc}} = \text{FC}(\mathbf{z})$. Following the work of (Xu et al., 2015), we extract feature vectors on the lower convolutional layer rather than use the fully connected layer. In this way, the decoder can selectively attend to certain parts of an image by weighting a subset of the feature vectors. In our experiments, we use the $14 \times 14 \times 2048$ feature map in ResNet-101. That said, our LSTM decoder operates on the flattened 196×2048 (i.e. $L \times D$) representations. The shared CNN encoder are then fine-tuned with both image captioning and multi-object classification tasks.

3.2 Shared LSTM decoder

The image captioning task shares its LSTM decoder with the decoder of syntax generation task. Following (Anderson et al., 2017), the decoder consists of two stacked LSTM networks. The first LSTM layer (denoted as LSTM$^{(1)}$) is characterized as a top-down visual attention model, while the second LSTM layer (denoted as LSTM$^{(2)}$) is a language model. The attention model takes input as the concatenation of the previous output of the language LSTM ($h^{(2)}_{t-1}$), the mean-pooled image feature $\bar{z} = \frac{1}{L}\sum_i z_i$, the word embedding of the previously generated word (e^w_{t-1}), and its syntax embedding (e^s_{t-1}):

$$x^{(1)}_t = \left[h^{(2)}_{t-1}, \bar{z}, e^w_{t-1}, e^s_{t-1}\right] \tag{2}$$

where $x^{(1)}_t$ is the input of LSTM$^{(1)}$ at time step t. The hidden state of attention LSTM at time step t is then computed by:

$$h^{(1)}_t = \text{LSTM}^{(1)}\left(h^{(1)}_{t-1}, x^{(1)}_t\right) \tag{3}$$

Given the output of the attention LSTM $h^{(1)} = \left\{h^{(1)}_1, h^{(1)}_2, \ldots, h^{(1)}_T\right\}$, where T is the length of the sequence, we compute the attended image feature $\hat{z}$, which is then used as the input to the language LSTM. The attended image feature $\hat{z}$ makes sure that every time step of the decoder can get full information of the context. We calculate $\hat{z}_t$ when we decode the t-th word by

$$\hat{z}_t = \sum_{i=1}^{L} \alpha_{t,i} z_i \tag{4}$$

Here, the attention weight $\alpha_{t,i}$ for the i-th annotation of CNN encoder is computed by

$$\alpha_{t,i} = \frac{\exp(c_{t,i})}{\sum_{k=1}^{L}\exp(c_{t,k})}; \quad c_{t,i} = \sigma(h^{(1)}_{t-1}, z_i) \tag{5}$$

where σ is a feed-forward neural network, which maps a vector to a real-valued score. This attention weights $\alpha_{t,i}$ models the alignment between the image content at location i and the output word at position t.

We use the LSTM language model to produce a caption (or a CCG supertag sequence) by generating one word (or CCG supertag) at every time step, which takes as input the concatenation of the output of the attention LSTM ($h^{(1)}_t$) and the attended image feature ($\hat{z}_t$), given by:

$$x^{(2)}_t = \left[\hat{z}_t, h^{(1)}_t\right] \tag{6}$$

where $x^{(2)}_t$ is the input of LSTM$^{(2)}$.

The hidden state of the language model LSTM at time t is then computed by

$$h^{(2)}_t = \text{LSTM}^{(2)}\left(x^{(2)}_t, h^{(2)}_{t-1}\right) \tag{7}$$

Finally, the generation probabilities of the t-th word and the t-th CCG supertag are given by

$$p_t^w = p\left(y_t^w \mid y_{1:t-1}^w, y_{1:t-1}^s; \mathbf{x}\right) = \text{softmax}\left(U^w h_t^{(2)} + b^w\right),$$
$$p_t^s = p\left(y_t^s \mid y_{1:t-1}^w, y_{1:t-1}^s; \mathbf{x}\right) = \text{softmax}\left(U^s h_t^{(2)} + b^s\right) \tag{8}$$

where U^w, U^s, b^w and b^s are the parameters to be learned.

Incorporating Object Label Embeddings To explore the variations of caption styles for images of different objects, we integrate the object label embeddings in the LSTM decoder. A forgetting gate is introduced to determine when the object label information contributes, which is computed by $f_t = \text{sigmoid}\left(h_{t-1}^{(2)}, z^{\text{fc}}\right)$. Then the biased word generation probability is given as:

$$p_t^w \propto \exp\left(U^w h_t^{(2)} + b^w\right) + \exp\left(f_t \odot W z^{\text{fc}}\right) \odot I_{\mathcal{W}_o}(w) \tag{9}$$

where W is transformation matrix to project object label into word embedding space. $I_{\mathcal{W}_o}$ is the indicator vector denoting whether candidate word w is in the vocabulary $\mathcal{W}_o$ of object categories.

3.3 Multi-task Learning

Our *MLAIC* model consists of three subtasks, each has its own training objective. For the object classification subtask, the model minimizes the multi-label margin loss function following the work of (Li et al., 2017):

$$J^{\text{obj}}(\theta) = \sum_{j \notin C} \sum_{k \in C} w_j \max\left(0, b + \mathbf{z}_k^{\text{fc}} - \mathbf{z}_j^{\text{fc}}\right) \tag{10}$$

where θ is the set of parameters, $\mathbf{z}_j^{\text{fc}}$ denotes the probability that image $\mathbf{x}$ contains object j, w_j is the weight indicating the frequency of occurrence of object j in the image, b is a hyper-parameter that determines the margin, commonly set to 1.0.

For the syntax generation and image captioning subtasks, we employ the minimum negative log-likelihood estimation:

$$J_{\text{ml}}^{\text{word}}(\theta) = -\sum_t \log p_t^w, \quad J_{\text{ml}}^{\text{syn}}(\theta) \quad = -\sum_t \log p_t^s \tag{11}$$

For the purpose of improving shared CNN encoder and LSTM decoder, we train these three related tasks simultaneously. The joint multi-task objective function is minimized by:

$$J_{\text{ml}}(\theta) = \lambda_1 J^{\text{obj}}(\theta) + \lambda_2 J_{\text{ml}}^{\text{word}}(\theta) + \lambda_3 J_{\text{ml}}^{\text{syn}}(\theta). \tag{12}$$

For comparison with recent work (Rennie et al., 2016), we also optimize directly for CIDEr (Vedantam et al., 2015) using policy gradient algorithm, and minimize the negative expected rewards:

$$J_{\text{rl}}^{\text{word}}(\theta) = -E_{y_{1:t} \sim p_\theta^w}\left[r\left(y_{1:t}^w\right)\right] \tag{13}$$

where $r(.)$ is CIDEr score function. According to the policy gradient theorem (Williams, 1992), we compute the gradient of the expected reward with respect to parameters as:

$$\nabla_\theta J_{\text{rl}}^{\text{word}}(\theta) \approx -\left(r\left(y_{1:t}^w\right) - r\left(\hat{y}_{1:t}^w\right)\right) \nabla_\theta \log p_\theta\left(y_{1:t}^w\right) \tag{14}$$

After pre-training the proposed model by minimizing the negative log-likelihood with multi-task using Eq.12, we switch the model to further minimize a mixed training objective, integrating the reinforcement learning objective J_{rl} with the original multi-task loss.

$$J_{\text{mixed}}(\theta) = \beta J_{\text{ml}}(\theta) + (1-\beta) J_{\text{rl}}^{\text{word}}(\theta) \tag{15}$$

where β is a hyper-parameter, and we set $\beta = 0.005$.

4 EXPERIMENTAL SETUP

4.1 Dataset

In our experiment, we use the widely used MSCOCO 2014 image captions (Karpathy and Fei-Fei, 2015) as our dataset, which consists of totally 82,783 training, 40,504 validation, and 40,775 testing images, each of which has 5 ground truth captions. For the off-line testing, we adopt the commonly used Karpathy split (Karpathy and Fei-Fei, 2015), which uses 113,287 images for training, and 5,000 images for validation and testing, respectively. This setting has been widely adopted in previous studies. For the on-line server evaluation, our proposed model is trained on 118,287 images and validated on 5,000 images.

For multi-label objection classification, we use the MSCOCO detection dataset (Lin et al., 2014), which consists of 500,000 object instances from 80 different object categories. It contains the same images and shares the same data split with MSCOCO 2014 image captions dataset (Karpathy and Fei-Fei, 2015).

For syntax generation, following the work of (Nadejde et al., 2017), the captions of training data is annotated with CCG supertags by using EasySRL[1], where each word has a corresponding dependency label of supertags.

[1] https://github.com/uwnlp/EasySRL

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr
Google NIC (Vinyals et al., 2015)	-	-	-	27.7	-	23.7	85.5
Hard-Attention (Xu et al., 2015)	70.7	49.2	34.4	24.3	23.9	-	-
Soft-Attention (Xu et al., 2015)	71.8	50.4	35.7	25.0	23.0	-	-
VAE (Pu et al., 2016)	72.0	52.0	37.0	28.0	24.0	-	90.0
Google NICv2 (Vinyals et al., 2017)	-	-	-	32.1	25.7	-	99.8
Attributes-CNN+RNN (Wu et al., 2016)	74.0	56.0	42.0	31.0	26.0	-	94.0
CNN$_{\mathcal{L}}$+RNN (Gu et al., 2017b)	72.3	55.3	41.3	30.6	26.0	-	94.0
PG-SPIDEr-TAG (Liu et al., 2016)	75.4	59.1	44.5	33.2	25.7	55.0	101.3
Adaptive (Lu et al., 2016)	74.2	58.0	43.9	33.2	26.6	54.9	108.5
SCST:Att2in (Rennie et al., 2016)	76.9	60.2	45.1	33.3	26.3	55.3	111.4
SCST:Att2all (Rennie et al., 2016)	77.4	60.9	46.0	34.1	26.7	55.7	114.0
TopDown:ResNet (Anderson et al., 2017)	76.6	59.9	45.4	34.0	26.5	54.9	111.1
TopDown:Up-Down (Anderson et al., 2017)	79.8	63.4	48.4	36.3	27.7	56.9	120.1
StackCap (Gu et al., 2017a)	78.4	62.5	47.9	36.1	27.4	56.9	**120.4**
MLAIC (ours)	**80.7**	**63.9**	**49.0**	**36.9**	**27.7**	**57.5**	119.1

Table 2: Comparisons of our image captioning approach and existing methods on MSCOCO Karpathy test split.

4.2 Baseline Methods

In the experiments, we compare our model with state-of-the-art methods, and several recent strong competitors are described below:

Adaptive (Lu et al., 2016). This model uses adaptive attention mechanism with a visual sentinel to determine when to look at the image and which regions to attend at the decoder stage.

SCST: Att2in/Att2all (Rennie et al., 2016). This model proposes self-critical sequence training (SCST) to bypass the non-differentiable task metric issue, and uses an attention model to dynamically select and linearly combine different locations of the input image.

TopDown: ResNet/Up-Down (Anderson et al., 2017). This model combines Bottom-up features and Top-Down LSTM attention methods, which enables the attention scores to be computed at the levels of objects and salient image regions.

StackCap (Gu et al., 2017a). This model builds on a multi-stage framework of a coarse-to-fine mechanism. It consists of multiple decoders, where each decoder produces refined image descriptions based on the previous one.

4.3 Implementation Details

We first pre-train our model on the training data with cross-entropy cost, and use Adam optimizer with an initial learning rate 5×10^{-4} and a momentum parameter of 0.9 to optimize the parameters. After that, we run the proposed RL-based approach on the just trained model, which is directly optimized for the CIDEr metric. During this stage, we use Adam optimizer with learning rate $5{\times}10^{-5}$. We set λ_1=0.2, λ_2=0.7, λ_3=0.1. We set the number of hidden units in TopDown attention LSTM (LSTM$^{(1)}$) to 1,000, the number of hidden units in language model LSTM (LSTM$^{(2)}$) to 512, the size of the input word embedding to 512, and the size of the CCG supertag embedding to 100. During the decoding stage, we use a beam size of 5 to generate captions.

4.4 Evaluation Metrics

To quantitatively evaluate our image captioning method, we follow previous work to use BLEU-N (N=1,2,3,4) (Papineni et al., 2002), MENTOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015) scores for comparison. All these metrics measure the consistency between n-gram occurrences in generated captions and ground-truth captions, where this consistency is weighted by n-gram saliency and rarity.

5 EXPERIMENTAL RESULTS

In this section, we compare our model with the competitors quantitatively and qualitatively.

5.1 Quantitative Evaluation

5.1.1 Offline Evaluation

The experimental results on MSCOCO with Karpathy test split are summarized in Table 3.3. We observe that *MLAIC* substantially and consistently outperforms the existing methods by a noticeable margin on most of the evaluation metrics. In particular, our model successfully yields better scores of all evaluation metrics compared to the TopDown model which utilizes the same basic CNN-LSTM backbone as ours. Our model benefits from the fact that both the CNN encoder and

	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Google NIC	71.3	89.5	54.2	80.2	40.7	69.4	30.9	58.7	25.4	34.6	53.0	68.2	94.3	94.6
Hard-Attention	70.5	88.1	52.8	77.9	38.3	65.8	27.7	53.7	24.1	32.2	51.6	65.4	86.5	89.3
PG-SPIDEr-TAG	75.1	91.6	59.1	84.2	44.5	73.8	33.1	62.4	25.5	33.9	55.1	69.4	104.2	107.1
Adaptive (Ens.5)	74.8	92.0	58.4	84.5	44.4	74.4	33.6	63.7	26.4	35.9	55.0	70.5	104.2	105.9
Stack-Cap (Single)	77.8	93.2	61.6	86.1	46.8	76.0	34.9	64.6	27.0	35.6	56.2	70.6	114.8	118.3
SCST:Att2in (Ens.4)	78.1	93.1	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
TopDown:Up-Down (Ens.4)	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
MLAIC (Single)	80.3	94.4	63.4	87.3	48.3	77.4	36.0	66.1	27.4	36.1	57.0	71.8	113.9	116.4

Table 3: Leaderboard of the published image captioning models (as of 12/7/2017) on the online MSCOCO test server. Our single model trained with the multi-task learning yields comparable performance with the state-of-the-art approaches (include the ensemble models) on all reported metrics.

	Cross-entropy							CIDEr-optimization						
	B-1	B-2	B-3	B-4	METEOR	ROUGH	CIDEr	B-1	B-2	B-3	B-4	METEOR	ROUGH	CIDEr
MLAIC	76.2	59.7	46.1	35.7	27.0	55.6	109.8	80.7	63.9	49.0	36.9	27.7	57.5	119.1
w/o multi-label	74.6	58.3	44.8	34.5	26.8	55.3	107.7	77.2	61.1	46.1	34.5	26.9	56.5	115.7
w/o syntax	75.8	59.2	45.7	35.5	26.9	55.5	109.2	80.1	63.5	48.7	36.4	27.4	57.1	118.5

Table 4: Ablation study on MSCOCO Karpathy test split under the cross-entropy optimization and CIDEr-optimization. Here, B-n is short for BLEU-n.

the LSTM decoder are improved by jointly learning the image captioning with the object classification and the syntax generation tasks.

5.1.2 Online Evaluation

Table 4.4 reports the performance of our single model on the official MSCOCO evaluation server[2]. The previous top performers on the leaderboard (as of December 7, 2017) are also demonstrated. Our model achieves better or competitive performance compared to the state-of-the-art competitors (both single and ensemble models).

5.2 Ablation Study

For the purpose of analyzing the effectiveness of different components of our model for image captioning, in this section, we report the ablation test of our model by discarding the multi-label object classification task (w/o object) or the syntax generation task (w/o syntax), respectively. In order to investigate the performance of the policy gradient update, we also report the results of our multi-task learning model with policy gradient update (i.e., CIDEr-optimization) or with cross-entropy optimization, respectively.

The results are demonstrated in Table 6. Generally, both tasks contribute, and the multi-label object classification task contributes most. This is within our expectation since the object classification task helps the CNN encoder learn better image representations. In addition, the proposed model utilizes the classification results (label embeddings) to produce captions of different styles and wording with respect to different object categories. While the improvement of integrating syntax generation is relatively limited. This may be because that the issue of incomplete sentence has little effect on the evaluation metrics of image captions. As shown in 5, the sentences generated by w/o syntax model are incomplete but achieve high scores in terms of the evaluation metrics. It is no surprise that combining both tasks achieves the best performance for all evaluation metrics. The main advantage comes from its capability of sharing knowledge of image captioning with two related tasks and learning better encoder and decoder representations simultaneously. We can also observe that the model with policy gradient update substantially outperforms the model with cross-entropy by a noticeable margin on all the evaluation metrics. The is because that the policy gradient update is able to bypass the exposure bias and non-differentiable evaluation metrics issues, and maximize long-term reward in caption generation.

5.3 Qualitative Evaluation

To evaluate the proposed model qualitatively, we show some generated image descriptions in table 5. It is easy to see that *MLAIC* can generate rea-

[2] https://competitions.codalab.org/competitions/3221

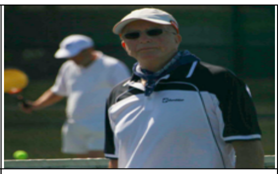
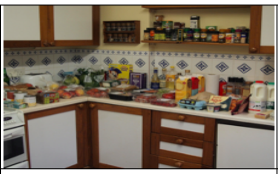
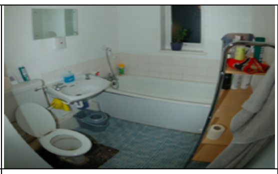
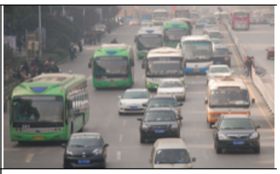
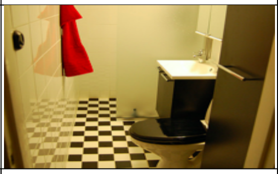
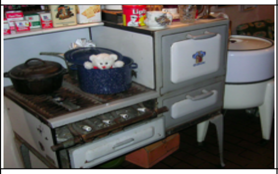
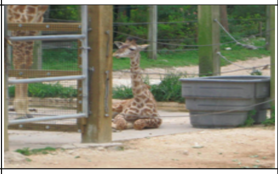
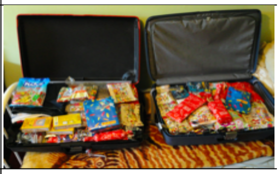
				
w/o syntax	a man holding a tennis ball with a tennis ball.	a kitchen with a stove and cabinets in the.	a bathroom with a toilet sink and a sink.	a busy city street with cars and buses on the.
w/o object	a man holding a tennis ball.	a kitchen with a stove and cabinets on the wall.	a bathroom with a toilet and a sink.	a group of cars and buses on a city street.
MLAIC	a man hitting a tennis ball with a tennis racket.	a kitchen with a stove and cabinets on the counter.	a bathroom with a toilet and a sink in it.	a bunch of cars and buses on a highway.
				
w/o syntax	a bathroom with a toilet and sink in the.	a kitchen with two pots and sitting on top of.	a giraffe is standing in the zoo of a zoo.	a suitcase filled with lots of items and.
w/o object	a bathroom with a toilet and sink in it.	a kitchen with two pots sitting on top of it.	a giraffe is standing in the dirt of a zoo.	a suitcase filled with lots of items.
MLAIC	a bathroom with a toilet and tiled floor in it.	a kitchen with two pots sitting on a stove.	a giraffe is standing in the dirt of a fence.	a suitcase filled with lots of items on a bed.

Table 5: Example captions generated by different models. The red, blue and green text indicate the in-corrected, in-appropriate and sophisticated phrases respectively.

Captions:	a	giraffe	is	standing	in	the	dirt	of	a	fence
CCG:	NP/N	N	(S[dcl]\NP)/(S[ng]\NP)	(S[ng]\NP)/PP	PP/NP	NP/N	N/PP	PP/NP	NP/N	N
Captions:	a	suitcase	filled	with	lots	of	items	on	a	bed
CCG:	NP/N	N	(S[pss]\NP)/PP	PP/NP	N/PP	PP/NP	N	(S\NP)\(S\NP)/NP	NP/N	N

Table 6: Examples of captions and corresponding CCG supertags generated by our model.

sonably relevant and plausible sentences. For example, the sentences “a man hitting a tennis ball with a tennis racket” and “a bunch of cars and buses on a highway” created by *MLAIC* are more precise in describing the content of the images. By comparing the results of w/o syntax and *MLAIC*, we reveal that the syntax generation task is able to help the LSTM decoder to alleviate the issues of generating duplicate words and incomplete sentences. On the other hand, the object classification task makes the CNN encoder be able to recognize the existence of objects, and help the LSTM decoder produce captions of different styles and generate novel words with respect to different object categories.

The performance of the syntax generation task is also improved by the multi-task learning mechanism. We provide some qualitative examples of image captions and the corresponding CCG supertags in Table 5.2 to evaluate the syntax generation model. From Table 5.2 we observe that the generated CCG supertag sequences are reasonable and consistent with the corresponding images captions. For example, the supertag (S[dcl]\NP)/(S[ng]\NP) for the verb “is” in the first caption implicates that a present participle is expected if the next word is verb. While the supertag (S\NP)\(S\NP)/NP for “on” implies that a noun phrase will be generated in the next. Our model can predict the high-quality CCG supertags, and help the image captioning alleviate the problem of generating duplicate words and incomplete sentence.

6 Conclusion

We proposed a novel multi-task learning method to improve image captioning by jointly training the image captioning with two related tasks: object classification and syntax generation. The objection classification helps learn better image representations and improve the performance of visual attention, and the syntax generation helps alleviate the problem of generating incomplete sentences and duplicate words. Our model achieves better or comparable performance with the state-of-the-art approaches on MSCOCO dataset.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and vqa. *arXiv preprint arXiv:1707.07998* .

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL*. volume 29, pages 65–72.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research* 55:409–442.

Rich Caruana. 1998. Multitask learning. In *Learning to learn*, Springer, pages 95–133.

Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, and Tat-Seng Chua. 2016. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. *arXiv preprint arXiv:1611.05594* .

Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. 2017. Show, adapt and tell: Adversarial training of cross-domain image captioner. *arXiv preprint arXiv:1705.00930* .

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*. ACM, pages 160–167.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *The IEEE conference on computer vision and pattern recognition*. pages 2625–2634.

Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *The Conference on Empirical Methods in Natural Language Processing*. volume 13, pages 1292–1302.

Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014. Improving image-sentence embeddings using large weakly annotated photo collections. In *European Conference on Computer Vision*. pages 529–545.

Jiuxiang Gu, Jianfei Cai, Gang Wang, and Tsuhan Chen. 2017a. Stack-captioning: Coarse-to-fine learning for image captioning. *arXiv preprint arXiv:1709.03376* .

Jiuxiang Gu, Gang Wang, Jianfei Cai, and Tsuhan Chen. 2017b. An empirical study of language cnn for image captioning. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pages 770–778.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition*. pages 3128–3137.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* .

Polina Kuznetsova, Vicente Ordonez, Tamara L Berg, and Yejin Choi. 2014. Treetalk: Composition and compression of trees for image descriptions. *Transactions of the Association for Computational Linguistics* 2(10):351–362.

Yuncheng Li, Yale Song, and Jiebo Luo. 2017. Improving pairwise ranking for multi-label image classification. *arXiv preprint arXiv:1704.03135* .

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL Workshop on Text Summarization Branches Out*. volume 8.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, pages 740–755.

Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2016. Optimization of image description metrics using policy gradient methods. *arXiv preprint arXiv:1612.00370* .

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2016. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *arXiv preprint arXiv:1612.01887* .

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *International Conference on Learning Representations*.

Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. pages 747–756.

Maria Nadejde, Siva Reddy, Rico Sennrich, Tomasz Dwojak, Marcin Junczys-Dowmunt, Philipp Koehn, and Alexandra Birch. 2017. Predicting target language ccg supertags improves neural machine translation. In *Proceedings of the Second Conference on Machine Translation*. pages 68–79.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *The 40th Annual Meeting on Association for Computational Linguistics*. pages 311–318.

Ramakanth Pasunuru and Mohit Bansal. 2017. Multi-task video captioning with video and entailment generation. *arXiv preprint arXiv:1704.07489* .

Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. 2016. Variational autoencoder for deep learning of images, labels and captions. In *Advances in Neural Information Processing Systems*. pages 2352–2360.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2016. Self-critical sequence training for image captioning. *arXiv preprint arXiv:1612.00563* .

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *The IEEE Conference on Computer Vision and Pattern Recognition*. pages 4566–4575.

Subhashini Venugopalan, Lisa Anne Hendricks, Raymond Mooney, and Kate Saenko. 2016. Improving lstm-based video description with linguistic knowledge mined from text. In *arXiv preprint arXiv:1604.01729*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pages 3156–3164.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2017. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence* 39(4):652–663.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning* 8:229–256.

Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton van den Hengel. 2016. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 203–212.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*. pages 2048–2057.