Show, Conceive and Tell: Image Captioning with Prospective Linguistic Information

Yiqing Huang[1][0000−0002−2143−3329] and Jiansheng Chen[1][0000−0002−2040−7938]

Department of Electronic Engineering, Tsinghua University
huang-yq17@mails.tsinghua.edu.cn, jschenthu@mail.tsinghua.edu.cn

Abstract. Attention based encoder-decoder models have achieved competitive performances in image captioning. However, these models usually follow the auto-regressive way during inference, meaning that only the previously generated words, namely the explored linguistic information, can be utilized for caption generation. Intuitively, enabling the model to conceive the prospective linguistic information contained in the words to be generated can be beneficial for further improving the captioning results. Consequently, we devise a novel Prospective information guided LSTM (Pro-LSTM) model, to exploit both prospective and explored information to boost captioning. For each image, we first draft a coarse caption which roughly describes the whole image contents. At each time step, we mine the prospective and explored information from the coarse caption. These two kinds of information are further utilized by a Prospective information guided Attention (ProA) module to guide our model to comprehensively utilize the visual feature from a semantically global perspective. We also propose an Attentive Attribute Detector (AAD) which refines the object features to predict the image attributes more precisely. This further improves the semantic quality of the generated caption. Thanks to the prospective information and more accurate attributes, the Pro-LSTM model achieves near state-of-the-art performances on the MSCOCO dataset with a 129.5 CIDEr-D.

1 Introduction

Image captioning aims at automatically generating the descriptions for images in natural language. This task can facilitate lots of practical applications such as human-machine interaction and content based image retrieval. To date, the encoder-decoder framework [1] equipped with attention modules [2–11] has become prevalent in image captioning. Generally, these models utilize CNN as the encoder to extract visual features from the image, and leverage language decoders such as RNN or the Transformer [12] to generate the captions. These decoders usually infer the caption in an auto-regressive manner. Specifically, when generating the current word, only the linguistic information contained in the previously generated words are utilized. For convenience, we call such linguistic information as the *explored information* in this work. Oppositely, the linguistic information contained in the words to be generated, namely the *prospective*

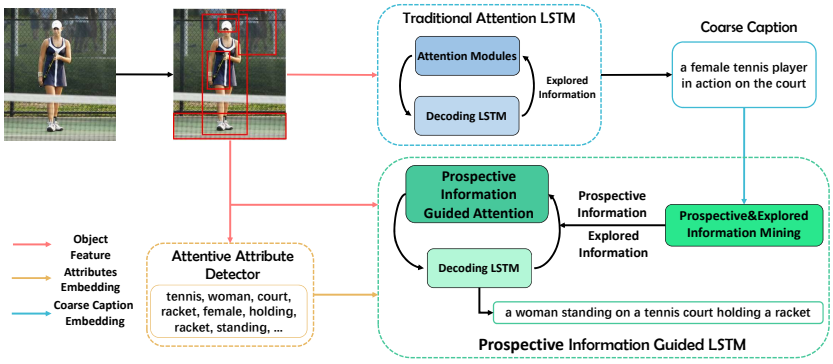

Fig. 1. While most image captioning models can only utilize the explored information, our model additionally incorporates prospective linguistic information to boost captioning. Object features extracted by the Faster-RCNN [13] are both exploited to generate a coarse caption and refined to predict the image attributes via Attentive Attribute Detector (AAD). The proposed Prospective information guided LSTM (Pro-LSTM) finally leverages the prospective information to guide the attention modules to better attend to the visual features, leading to image captions of higher quality.

information, is seldom considered. This is logical considering that in an auto-regressive process, exact prospective information keeps unknown. However, suppose there is a way to get the prospective information in advance, even roughly, it is possible for the decoder to synthesize more refined and accurate image descriptions under the guidance of a 'conceived' global linguistic context.

Actually, when human beings compose a sentence, they usually draft a coarse version first; and then polish it with both prospective and explored linguistic information to obtain a more descriptive and accurate sentence. In image captioning, such a strategy can be imitated by firstly generating a coarse caption using a pre-trained auto-regressive captioning model. Although the coarse caption may not be perfect, roughly it is still a semantically correct description of the overall content of the image, as shown in the blue circle in Fig. 1. Thus, the coarse caption can be regarded as a reasonable representation of the global linguistic information from which the prospective information can be effectively extracted. Specifically in our proposal, at each time step, the prospective&explored information mining module adaptively renews the prospective information and the explored information. It should be noticed that the prospective information may not necessarily be related to the succeeded words in the coarse caption. It denotes the information contained in words that have not yet been generated by the current time step. In our method, we extract the prospective information from coarse caption words of which the semantic information is less correlated to that contained in the previously generated words. Such words are closely related to the input image and contain additional linguistic information that is complementary to the explored information.

Both prospective information and explored information are further exploited by the Prospective information guided Attention (ProA) module in the green circle of Fig. 1. Considering that the explored information is deterministic and relatively more reliable, we utilize it to augment the visual features directly as linguistic feature. The prospective information, however, contains additional semantic information upon current linguistic context. Thus, we do not directly utilize it as a feature but combine it with the current linguistic context to guide the attention modules semantically from a global perspective. By jointly using the prospective and explored information, our model attends to more appropriate visual features based on a better grasp of the complete linguistic context.

Besides exploiting the image features and the coarse caption, we also leverage the image attributes, the most salient concepts contained in the image, in our model, as shown in the orange circle in Fig. 1. While most works [14, 15] directly adopt the image features in attribute detection, we propose to refine these features for better detection performance to further strengthen the captioning model. Our proposed Attentive Attribute Detector (AAD) leverages the Graph Convolutional Network (GCN) to model the similarity between the image features and the attribute embedding in the refinement process. We notice that with proper selections of similarity formulation and activation function, the GCN can actually be transformed to a multi-head attention [12] module.

Benefiting from both prospective information contained in the coarse caption and the explicit semantic information brought about by the image attributes, the proposed Pro-LSTM model can generate precise and detailed image captions. The main contributions of our work are as follows: **1)** We introduce Pro-LSTM which additionally utilizes prospective information to facilitate better usage of visual and language information to boost image captioning. **2)** We introduce an AAD which refines the image features in order to predict the image attributes more precisely. **3)** The Pro-LSTM model achieves state-of-the-art image captioning performance of 129.5 CIDEr-D score on the MSCOCO benchmark dataset [16]. We notice that captions generated by Pro-LSTM are sometimes even more descriptive than the human-labeled ground truth captions, indicating the effectiveness of introducing the prospective information.

2 Related Work

Neural Image Captioning. The neural network based encoder-decoder framework was first proved to be effective in image captioning in [1]. Later on, the attention mechanism has been introduced to boost the vanilla encoder-decoder framework. For example, spatial attention [5], semantic attention [6], adaptive attention [4], bottom-up and top-down attention [2] were introduced to exploit the visual information in different ways for generating better descriptions. Recently, Huang *et al.* [17] modified the attention module by adopting another attention over it to obtain more appropriate attention results. While these auto-regressive methods focus on better exploiting the explored information at each time step, we explore how to additionally utilize the prospective information to

utilize both the visual information and the language information. We will also show that our method is compatible with the state-of-the-art AoA [17] method.

Exploitation of Prospective Information. Realizing that only leveraging the explored information is not sufficient for captioning, researchers began to study the possibility of exploiting the complementarity between the explored information and the prospective information. Wang *et al.* [18] adopted the bidirectional LSTMs to generate two sentences in both forward-pass and backward-pass independently. Nevertheless, as they merely integrated the two generated sentences by selecting the one with larger probability, their method was essentially exploiting the explored and inversely-explored information but not the prospective information. Look&modify [19] was devised to modify the coarse captions with the residual information. However, they roughly integrate the word embedding of the words in the coarse caption without explicit model the relationships between the words. Recently, Ge *et al.* [20] proposed to attend to both the coarse caption and the visual features respectively to generate a better description. However, they treat the coarse caption just like the attributes and fail to model the interaction between prospective information and visual information. In our proposal, both the prospective and explored information are thoroughly exploited to guide the model towards more appropriate visual attention.

3 Preliminary

Before introducing our framework, we briefly introduce the Multi-Head Attention (MHA) [12]. The multi-head attention was first introduced in the Transformer model [12]. The *scaled dot-product attention* is the core component of multi-head attention. Given a query $\mathbf{q}_i$, a set of keys $\mathbf{K} = (\mathbf{k}_1,\ ...,\ \mathbf{k}_n)$ and a set of values $\mathbf{V} = (\mathbf{v}_1,\ ...,\ \mathbf{v}_n)$ where $\mathbf{q}_i,\ \mathbf{k}_i,\ \mathbf{v}_i \in \mathbb{R}^d$, the output of the scaled dot-product attention is the weighted sum of $\mathbf{v}_i$. The weights are determined by the dot-product of $\mathbf{q}_i$ and $\mathbf{k}_j$. Additionally, the dot-products are divided by the square root of dimension d. In practice, the queries are packed together as a matrix $\mathbf{Q} = (\mathbf{q}_1,\ ...,\ \mathbf{q}_m)$ to compute the above process in parallel as in (1).

$$\mathit{Attention}(\mathbf{Q},\ \mathbf{K},\ \mathbf{V}) = \mathit{Softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}})\mathbf{V} \tag{1}$$

Multi-head attention (MHA) is an extension of the above attention mechanism. The queries, keys and values are firstly linearly projected to h subspaces. Then the scaled dot-product attention is applied to the h heads separately in (2), where $i \in \{1,\ ...,\ h\}$. The h outputs are finally concatenated to form the output of MHA as in (3), where $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V \in \mathbb{R}^{d\times\frac{d}{h}}$ and $\mathbf{W}^{MHA} \in \mathbb{R}^{d\times d}$ are trainable parameters.

$$\mathbf{H}_i = \mathit{Attention}(\mathbf{Q}\mathbf{W}_i^Q,\ \mathbf{K}\mathbf{W}_i^K,\ \mathbf{V}\mathbf{W}_i^V) \tag{2}$$

$$\hat{\mathbf{V}} = \mathit{MHA}(\mathbf{Q},\ \mathbf{K},\ \mathbf{V}) = \mathit{Concat}(\mathbf{H}_1,\ ...,\ \mathbf{H}_h)\mathbf{W}^{MHA} \tag{3}$$

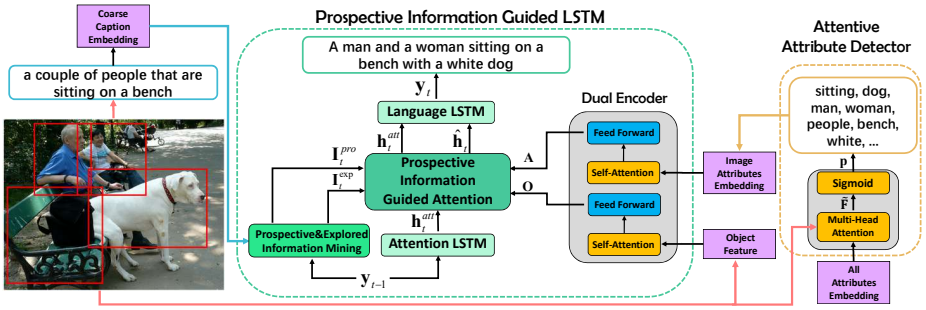

Fig. 2. The overall framework of our proposal. The framework is composed of an Attentive Attribute Detector (AAD), and a Prospective information guided LSTM (Pro-LSTM) caption generator. The object features (shown in red lines) are firstly extracted to generates a coarse caption with a pre-trained captioning model. Concurrently, AAD refines the object features with attribute embedding and predicts the probabilities $\mathbf{p}$ of the image attributes with the refined feature $\tilde{\mathbf{F}}$. The Pro-LSTM dynamically mines the prospective and explored information inside the coarse caption embedding (shown in blue line) and guide our model to more properly attend to the visual features through the Prospective information guided Attention (ProA) module.

4 Methodology

As shown in Fig. 2, the overall framework of our model follows the encoder-decoder paradigm. Firstly, we introduce the Pro-LSTM model, which is equipped with a Prospective information guided Attention (ProA) module to generate improved image captions with the guidance of prospective information. We then describe the Attentive Attribute Detector (AAD) which refines the object features with multi-head attention to better predict image attributes.

4.1 Prospective Information Guided LSTM

As shown in the middle green circle in Fig. 2, the Pro-LSTM model is composed of a *Dual Encoder* and an *LSTM decoder*. The MHA [12] based dual encoder consists of two separate components that encode the object features and the image attributes respectively. The LSTM decoder is a two-layer LSTM, where the first LSTM layer and the second LSTM layer are called the attention LSTM and language LSTM respectively. The implementations of these two LSTM layers are similar to that proposed in [2] except for the attention module and some inputs. It additionally takes in a coarse caption and leverages the ProA module to decode these features in order to obtain plausible and detailed image descriptions.

Dual Encoder Generally, the components of the dual encoder are implemented similarly using the structure defined in Eq. 4, where $\mathbf{Z}$ denotes the object feature or the attribute feature to be encoded, *FFN* and *MHA* are short for

Feed-Forward Network and Multi-Head Attention in [12]. The selected top-L attributes are ranked by the predicted confidence $\boldsymbol{p}$. We stack the attention blocks for 6 times. The outputs of the 6^{th} attention block of the object encoder and the attribute encoder are denoted as $\mathbf{O} \in \mathbb{R}^{M\times d}$ and $\mathbf{A} \in \mathbb{R}^{L\times d}$ respectively.

$$Encoder(\mathbf{Z}) = Stack(FFN(MHA(\mathbf{Z},\ \mathbf{Z},\ \mathbf{Z}))) \times 6 \tag{4}$$

Prospective&Explored Information Mining The structure of this module is shown in the left of Fig. 3. The idea of information mining is inspired by the phenomenon that the semantically similar words are close in the word embedding space [21, 22]. We notice that such a phenomenon also exists in the word embedding trained along with the image captioning network. Thus, we compute the cosine similarity between the current input word $\boldsymbol{y}_{t-1}$ and the words in the coarse caption $\hat{\mathbf{W}} = [\hat{\boldsymbol{w}}_1,\ ...,\ \hat{\boldsymbol{w}}_{T'}], \hat{\boldsymbol{w}}_i \in \mathbb{R}^{1\times d}$ at each time step to select the prospective words and the explored words.

$$cosine_t^i = \frac{\hat{\boldsymbol{w}}_i * \boldsymbol{y}_{t-1}}{\|\hat{\boldsymbol{w}}_i\| \left\|\boldsymbol{y}_{t-1}\right\|} \tag{5}$$

At the first time step, there is no word in the explored word set $\mathbf{W}_t^{exp}$ and all the words in the coarse caption are in the prospective word set $\mathbf{W}_t^{pro}$. Then, at each time step, the words in the coarse caption that are similar to ($cosine_t^i > 0.9$) current input words are recognized as previously generated words. These words are added to the explored word set and are removed from the prospective word set simultaneously. As such, the prospective word set $\mathbf{W}_t^{pro}$ always consists of words that contain unrevealed semantic information at the current time step. We then adopt self-attention to the two word sets as in Eq. 6, which transforms these sets of word embedding to a high-level representation of prospective information and explored information respectively.

$$\mathbf{I}_t^{pro} = MHA(\mathbf{W}_t^{pro},\ \mathbf{W}_t^{pro},\ \mathbf{W}_t^{pro}),\ \mathbf{I}_t^{exp} = MHA(\mathbf{W}_t^{exp},\ \mathbf{W}_t^{exp},\ \mathbf{W}_t^{exp}) \tag{6}$$

Prospective Information Guided Attention While most previous works focused on how to exploit the explored information, we additionally explore the effectiveness of using the prospective information to guide the attention mechanism in order to better utilize the visual information and the semantic information jointly. The right side of Fig. 3 shows the detailed architecture of the Prospective information guided Attention (ProA) module, which is composed of four attention sub-layers, an augmentation sub-layer, and a fusion sub-layer.

Suppose we are generating the t^{th} word $\boldsymbol{y}_t$ at the current time step. The attention LSTM takes in the concatenation of current input word $\boldsymbol{y}_{t-1}$, the averaged object feature $\bar{\mathbf{O}}$, and the output of the language LSTM in the last time step $\boldsymbol{h}_{t-1}^{lan}$ to generates the current attention query $\boldsymbol{h}_t^{att}$ as in Eq. 7.

$$\boldsymbol{h}_t^{att} = LSTM_{att}(\boldsymbol{h}_{t-1}^{att},\ [\boldsymbol{h}_{t-1}^{lan};\ \bar{\mathbf{O}};\ \boldsymbol{y}_{t-1}]) \tag{7}$$

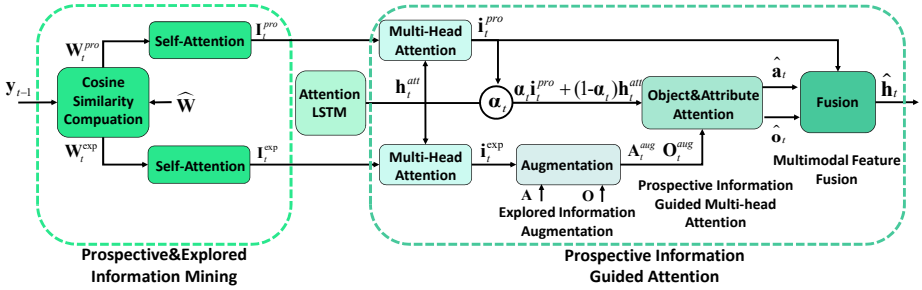

Fig. 3. The architecture of the proposed prospective&explored information mining module and the ProA module. The prospective and explored information are leveraged to guide visual attention to attend to more appropriate visual features in different ways. The object attention and attribute attention are implemented separately, they are combined in this figure since they share similar structure.

Multi-head attention is then applied to the prospective information $\mathbf{I}_t^{pro}$ and explored information $\mathbf{I}_t^{exp}$ respectively to obtain the corresponding features, namely the prospective feature $\boldsymbol{i}_t^{pro}$ and the explored feature $\boldsymbol{i}_t^{exp}$ in Eq. 8.

$$\boldsymbol{i}_t^{pro} = MHA(\boldsymbol{h}_t^{att},\ \mathbf{I}_t^{pro},\ \mathbf{I}_t^{pro}),\ \boldsymbol{i}_t^{exp} = MHA(\boldsymbol{h}_t^{att},\ \mathbf{I}_t^{exp},\ \mathbf{I}_t^{exp}) \tag{8}$$

After grasping the prospective and explored features, we leverage them for guiding the model to attend to the encoded object features $\mathbf{O}$ and image attributes $\mathbf{A}$ more properly. Considering that the explored feature is basically the integration of the information of previously generated words, it is appropriate to let the model attend to it when generating the none visual words like articles and prepositions at the current time step. Thus, we directly augment the encoded visual feature with the explored feature as in Eq. 9.

$$\mathbf{O}_t^{aug} = [\mathbf{O};\ \boldsymbol{i}_t^{exp}],\ \ \mathbf{A}_t^{aug} = [\mathbf{A};\ \boldsymbol{i}_t^{exp}] \tag{9}$$

Such an augmentation can be viewed as a modification of the Transformer [12] or the adaptive attention module [4]. While the encoded features in the Transformer are fixed in the whole captioning generation process, our model dynamically augments the explored feature to the encoded feature for facilitating the inclusion of new information at each time step. Comparing with the adaptive attention which directly attends to the visual features and the language features, we additionally leverage multi-head attention to make these features more informative.

The prospective feature $\boldsymbol{i}_t^{pro}$, however, can be viewed as the future information 'conceived' by the captioning model. Although this information may not be precise as the coarse caption may not always be satisfying, it is probably a useful supplement to the current linguistic context $\boldsymbol{h}_t^{att}$. Thus, we integrate the prospective feature $\boldsymbol{i}_t^{pro}$ with $\boldsymbol{h}_t^{att}$ to form a new guide and implement the visual attention mechanism from a global perspective. We modify the vanilla scaled

dot-product attention in Eq. 1 in the multi-head attention module to Eq. 10 and Eq. 11, so that the prospective information is leveraged as part of the query to attend to the augmented object features. In these equations, $\mathbf{W}^{attn} \in \mathbb{R}^{2d \times d}$ are trainable parameters and $\boldsymbol{\alpha}_t \in \mathbb{R}^d$ is the fusion weight vector. Similar modification is also applied to the augmented attribute features $\mathbf{A}_t^{aug}$ to generate the attribute context $\hat{\boldsymbol{a}}_t$. The corresponding equations are very similar to Eq. 10 and Eq. 11 and are omitted for conciseness.

$$Attention^{pro}(\boldsymbol{h}_t^{att}, \boldsymbol{i}_t^{pro}, \mathbf{O}_t^{aug}) = Softmax(\frac{(\boldsymbol{\alpha}_t \boldsymbol{i}_t^{pro} + (1-\boldsymbol{\alpha}_t)\boldsymbol{h}_t^{att})\mathbf{O}_t^{aug\top}}{\sqrt{d}})\mathbf{O}_t^{aug} \tag{10}$$

$$\hat{\boldsymbol{o}}_t = MHA^{pro}(\boldsymbol{h}_t^{att}, \boldsymbol{i}_t^{pro}, \mathbf{O}_t^{aug}), \;\; \boldsymbol{\alpha}_t = Sigmoid(Concat(\boldsymbol{h}_t^{att}, \; \boldsymbol{i}_t^{pro})\mathbf{W}^{attn}) \tag{11}$$

After the generation of the object context $\hat{\boldsymbol{o}}_t$ and the attribute context $\hat{\boldsymbol{a}}_t$, we integrate them with the prospective feature to form the attended feature. We concatenate the output of the attention LSTM and the prospective feature to form the fusion weights as in Eq. 12. The concatenation is then fed to a linear layer, the weight of which is $\mathbf{W}^{fuse} \in \mathbb{R}^{2d \times 3d}$. The output of the linear layer, which is of size $3d$, is reshaped to $3 \times d$ and further input to the Softmax layer to compute the fusion weights. The fused attended feature is obtained as in Eq. 13.

$$\boldsymbol{\beta}_t = Softmax(Reshape(Concat(\boldsymbol{h}_t^{att}, \; \boldsymbol{i}_t^{pro})\mathbf{W}^{fuse})) \tag{12}$$

$$\hat{\boldsymbol{h}}_t = \boldsymbol{\beta}_t^1 * \hat{\boldsymbol{o}}_t + \boldsymbol{\beta}_t^2 * \hat{\boldsymbol{a}}_t + \boldsymbol{\beta}_t^3 * \boldsymbol{i}_t^{pro} \tag{13}$$

Thanks to the global linguistic information brought about by the coarse caption, the model can exploit the visual information and language information more thoroughly by utilizing the complementarity between prospective and explored information contained in $\hat{\boldsymbol{h}}_t$. As such, image descriptions that are not only reasonable in general but also accurate in details can be possibly generated. Finally, the concatenation of the attended feature $\hat{\boldsymbol{h}}_t$ and the output of attention LSTM $\boldsymbol{h}_t^{att}$ is fed to the language LSTM to generate $\boldsymbol{h}_t^{lan}$ as in Eq. 14.

$$\boldsymbol{h}_t^{lan} = LSTM_{lan}(\boldsymbol{h}_{t-1}^{lan}, \; [\boldsymbol{h}_t^{att}; \; \hat{\boldsymbol{h}}_t]) \tag{14}$$

The output of the language LSTM if firstly sent to a fully connected layer to generate the logits and then sent to a softmax layer to generate the probability for each word in the vocabulary as is in Eq. 15, where $\mathbf{W}^p \in \mathbb{R}^{d \times k}$ and $\boldsymbol{b}^p \in \mathbb{R}^k$ are trainable parameters and k is the vocabulary size.

$$\boldsymbol{p}_t^w = Softmax(\boldsymbol{h}_t^{lan}\mathbf{W}^p + \boldsymbol{b}^p) \tag{15}$$

4.2 Attentive Attribute Detector

While previous attribute detectors [14, 15] directly utilize the image features extracted from CNN for attribute detection without further polishing, we alternatively explore the effectiveness of refining the visual features in advance.

Suppose that we need to predict the probability of P attributes with M object features, we can view the object features as M nodes in a graph and the embedding of all the attributes as P exterior nodes. We model the relationship between these two kinds of nodes and refine the object features via a special kind of one-layer Graph Convolutional Network (GCN). We follow [23] to construct the GCN layer, of which the mathematical formulation is shown in Eq. 16, where $\mathbf{W} \in \mathbb{R}^{d \times d}$ is the parameter of the linear layer and f is the activation function; function $A(\mathbf{F},\ \mathbf{E})$ models the similarity between the object features $\mathbf{F} \in \mathbb{R}^{M \times d}$ and the attribute embedding $\mathbf{E} \in \mathbb{R}^{P \times d}$.

$$\tilde{\mathbf{F}} = f(A(\mathbf{F},\ \mathbf{E})\mathbf{F}\mathbf{W}) \tag{16}$$

In our work, we take the recently proposed scaled dot-product formulation [12] to model such similarity. In this formulation, the attribute embedding $\mathbf{E}$ can be viewed as queries ($\mathbf{Q}$), while the object features $\mathbf{F}$ can be viewed as keys ($\mathbf{K}$) and values ($\mathbf{V}$) as in Eq. 17. Thus, the M object features are refined to P compound features $\tilde{\mathbf{F}} \in \mathbb{R}^{P \times d}$ that are more appropriate for predicting the probability of each image attribute.

$$\tilde{\mathbf{F}} = f(Softmax(\frac{\mathbf{E}\mathbf{F}^{\top}}{\sqrt{d}})\mathbf{F}\mathbf{W}) \tag{17}$$

In practice, the scaled dot-product formulation can be extended to the aforementioned MHA to further increase the model capacity. Actually, when the relationship is modeled by the scaled dot-product and f is set to identity mapping, GCN is transformed to MHA. Note that, when the number of interior nodes and exterior nodes is unequal ($M \neq P$), this kind of GCN cannot be stacked as usual GCN layers. That's why we call MHA a special kind of one-layer GCN with additional exterior nodes. The output of MHA is sent to a linear layer and then activated by the sigmoid function to predict the probability distribution $\boldsymbol{p} \in \mathbb{R}^{P \times 1}$ of image attributes as in Eq. 18, where $\mathbf{W}^{Attr} \in \mathbb{R}^{d \times 1}$ are trainable parameters.

$$\boldsymbol{p} = Sigmoid(MHA(\mathbf{E},\ \mathbf{F},\ \mathbf{F})\mathbf{W}^{Attr}) \tag{18}$$

The structure of AAD is shown in the right orange circle in Fig. 2. We model the task of attribute detection as P binary classification task and leverage the focal loss [24] in training. For each image, we send the embedding of top L attributes $\mathbf{G} \in \mathbb{R}^{L \times d}$ that are predicted with the highest confidences to the Pro-LSTM model to exert the attribute information.

4.3 Loss Functions

Our image captioning model is trained in two phases. In the first phase, we use the traditional cross-entropy (XE) loss. In the second phase, we modify the Self-Critical Sequence Training (SCST) [25] to directly optimize the CIDEr-D metric. The gradient of $loss^{w}_{RL}$ can be approximated as Eq. 19, where $r(w^{s}_{1:T})$, $r(\tilde{w}_{1:T})$ and $r(\hat{w}_{1:T})$ are the CIDEr-D rewards for the randomly sampled caption, greedy

Table 1. *F1* scores of two attribute detectors with top-L attributes. We reproduce the MIL based detector with Up-down feature for fair comparison.

L	5	10	15	20	25
MIL* [14]	0.286	0.420	0.435	0.422	0.395
AAD (ours)	0.352	0.442	**0.459**	0.454	0.437

decoded caption, and the input coarse caption respectively. We also impose the random caption to outperform the coarse caption in the SCST phase.

$$\nabla_\theta loss^w_{RL} = -(r(\boldsymbol{w}^s_{1:T}) - 0.5 * r(\tilde{\boldsymbol{w}}_{1:T}) - 0.5 * r(\hat{\boldsymbol{w}}_{1:T}))\nabla_\theta log(\boldsymbol{p}^w(w^s_{1:T})) \quad (19)$$

5 Experiments

5.1 Experimental Settings

We evaluate our model on the MSCOCO captioning dataset [16]. Words that appear in the training set for over 4 times are selected to form a k=10369 size vocabulary. Similar to [14, 15], the top-ranked 1000 words are selected to form the attribute vocabulary. We follow the widely adopted Karpathy's data split [26] in offline evaluation. We utilize the 36x2048 object feature released in Up-down [2] for both attribute detection and image captioning and set the hidden size d=1024. Our models are trained 15 epochs under XE loss and another 10 epochs under SCST [25] with a mini-batch size of 40. We use the following metrics in evaluation: Bleu [27], Meteor [28], Rouge-L [29], CIDEr-D [30] and SPICE [31].

5.2 Performance Evaluation and Analysis

Attribute Detection We first evaluate the attribute detection performance of our proposed AAD using the average *F1* score on the MSCOCO test split. Table 1 compares the detection performance of AAD with the MIL based [14] attribute detector. The MIL detector firstly utilizes one feature to predict the probabilities for all attributes, which may not be accurate; and then integrates the probabilities of all proposals with MIL. The proposed AAD, however, predicts the confidence of each attribute with the integrated and refined features of all the object features. It can be seen that AAD outperforms the MIL detector for different values of L, suggesting the effectiveness of our proposal.

MSCOCO Offline Evaluation Table 2 shows the single-model performance of the proposed Pro-LSTM model and recent state-of-the-art methods. The Up-Down [2] method extracts image features with bottom-up attention and generates image captions with top-down attention. The proposed two-layer LSTM is leveraged as our backbone model. Our proposed Pro-LSTM outperforms the SGAE [32] shows that incorporating coarse caption is more beneficial than utilizing the scene graph features. Although Snammani [19] and Ge *et al.* [20] also use

Table 2. Single-model image captioning performance (%) on the COCO 'Karpathy' test split, where B@N, M, R, C and S are short for Bleu@N, Meteor, Rouge-L, CIDEr-D and SPICE scores. * indicates the results obtained from the publicly available code. Top-2 scores in each column are marked in **boldface** and underline respectively.

Methods	Cross-Entropy Loss						CIDEr-D Optimization					
Metric	B@1	B@4	M	R	C	S	B@1	B@4	M	R	C	S
Up-Down [2]	77.2	36.2	27.0	56.4	113.5	20.3	79.8	36.3	27.7	56.9	120.1	21.4
SGAE [32]	-	-	-	-	-	-	80.8	38.4	28.4	58.6	127.8	22.1
Look&Modify [19]	76.9	36.1	-	56.4	112.3	20.3	-	-	-	-	-	-
MaBi-LSTM [20]	**79.3**	36.8	28.1	56.9	116.6	-	-	-	-	-	-	-
AoA [17]	77.4	**37.2**	**28.4**	**57.5**	119.8	21.3	80.2	38.9	**29.2**	**58.8**	**129.8**	22.4
AoA* [17]	77.4	37.0	28.2	57.3	118.4	21.5	80.4	39.0	28.9	58.7	128.8	22.5
Pro-LSTM (Ours)	77.8	37.1	28.2	57.3	120.2	21.5	**81.0**	**39.2**	29.0	58.8	129.5	22.6
Pro-LSTM+AoA*	77.7	37.1	28.3	57.2	**120.5**	**21.6**	80.8	39.0	29.0	**58.9**	**129.8**	**22.7**

Table 3. Performance (%) on the online MSCOCO evaluation server, where $^{\Sigma}$ denotes model ensemble. Top-2 rankings are indicated by red superscript for each metric

Methods	Bleu-1		Bleu-4		Meteor		Rouge-L		CIDEr-D	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
SCST$^{\Sigma}$ [25]	-	-	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
Up-Down$^{\Sigma}$ [2]	80.2[2]	95.2[1]	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
SGAE$^{\Sigma}$ [32]	-	-	38.5[1]	69.7[1]	28.2	37.2	58.6[1]	73.6[2]	123.8[2]	126.5[1]
AoA* [17]	79.9	94.4	38.0[2]	69.1[2]	28.6[2]	37.7[2]	58.2	73.4	123.2	125.8[2]
Pro-LSTM (ours)	80.3[1]	94.8[2]	38.5[1]	69.7[1]	28.7[1]	38.0[1]	58.4[2]	73.7[1]	124.1[1]	126.5[1]

the coarse caption, they fail to let it guide their model but treat it as semantic feature. Our method outperforms MaBi-LSTM in most metrics indicates the effectiveness of leveraging prospective information as appropriate guidance for the attention modules in the image captioning model. The AoA [17] model modifies the attention modules to achieve state-of-the-art performance. However, they fail to leverage the prospective information. We re-train the AoA* model with the released publicly available code to generate the coarse caption in our experiments. The experimental results show that our method extensively exploits the global information from the coarse caption and enhances the CIDEr-D for 1.8 and 0.7 respectively in XE and SCST. Our method is compatible with the AoA method as the combination of Pro-LSTM and AoA* yields the best performance.

MSCOCO Online Evaluation We submit the single-model captioning results of both the AoA* model and our proposed Pro-LSTM to the online testing server [1]. Table 3 shows the online performance of officially published state-of-the-art works. Our proposed Pro-LSTM achieves performance improvements over the AoA* model.The performance of Pro-LSTM is among the top-2 in all the compared methods across all the metrics. Specifically, a single Pro-LSTM model even outperforms the ensemble of SGAE models.

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

Table 4. The performance of utilizing different modules in XE training.

method	Bleu-1	Bleu-4	Meteor	Rouge-L	CIDEr-D	SPICE
baseline	76.8	36.5	28.0	56.9	116.7	21.1
baseline+AAD	77.2	36.7	28.2	57.0	118.3	21.3
baseline+ProA	77.7	36.9	28.3	57.1	119.1	21.5
Pro-LSTM	77.8	37.1	28.2	57.3	120.2	21.5

Table 5. The performance of using prospective information and explored information in XE training.

Methods	Bleu-1	Bleu-4	Meteor	Rouge-L	CIDEr-D	SPICE
Pro-LSTM - augmentation	77.8	37.0	28.2	57.2	119.7	21.5
Pro-LSTM - guidance	77.4	36.7	28.1	57.0	118.9	21.3
Pro-LSTM - fusion	77.5	36.8	28.2	57.2	119.3	21.4
Pro-LSTM	77.8	37.1	28.2	57.3	120.2	21.5

6 Ablation Study

6.1 Effectiveness of Modules

In Table 4, we assess the contribution of each module in our proposed method. We firstly form a baseline model similar to that proposed in Up-Down [2] except that we leverage multi-head attention to encode the object feature and generate the object context. We then add the attentive attribute detector to the baseline model to exploit the attribute information. Concurrently, we also replace the vanilla multi-head attention module with our proposed prospective information guided attention module as is shown in the third row. It can be noticed that leveraging the global information to guide the attention module is more beneficial than utilizing the image attributes only. Naturally, comprehensively using these two modules in our proposed Pro-LSTM leads to the most favorable performance.

6.2 Effectiveness of ProA

The effectiveness of our proposed ProA module mainly comes from the joint utilization of both prospective and explored information. More specifically, we leverage the explored information to augment the visual features and utilize the prospective information to guide the attention module. And we finally fuse the multimodal features to further enhance the performance. Thus, we test the effectiveness of each sub-module in Fig. 3 by eliminating it from the Pro-LSTM model. The corresponding results are shown in Table 5. The performance only decreases a little without using the explored information for feature augmentation. This is understandable since such information has already been embedded to the LSTM hidden states. Additionally leverage it in the attention module leads to incremental improvements only. In contrast, eliminating the guidance of prospective information results in a significant performance drop. This further verifies that grasping the global linguistic context in attention modules through

Fig. 4. Qualitative results of the generated captions and corresponding CIDEr-D scores in MSCOCO test split. One ground truth caption is shown for (d), (e), and (f) respectively. Pro-LSTM effectively utilizes the prospective information in the AoA* generated coarse captions to narrate more precise and detailed captions. In some cases, the captions generated by Pro-LSTM can more informative even when their CIDEr-D scores are lower than that of AoA* captions.

utilizing the prospective information is essentially beneficial for image captioning. Fusing the prospective information with visual information is also beneficial for captioning since this enables the model to better conceive future semantics. To conclude, we observe that effectively utilizing the prospective information in the ProA module leads to better captioning performance as we expect.

7 Qualitative Results

Fig. 4 compares the captions generated by the AoA* model and that by the Pro-LSTM model. Generally, AoA* sometimes generates grammatically correct but semantically flawed sentences. However, these sentences still contain useful descriptions that can help Pro-LSTM to correctly attend to corresponding visual features, such as *'walking'* and *'cake'* in Fig. 4(a)(b). In Fig. 4(c), AoA* only infers that the horses are *'standing'* in the field. With the help of prospective information, the Pro-LSTM model grasps the semantic that the horses are in a field via the language information, and further induces that they are *'grazing'* as it is a common behavior under this circumstance. For Fig. 4(d), Pro-LSTM describes the color scheme while AoA* fails to do so. This is probably because that *'black'* and *'white'* are successfully detected by AAD. Thus, Pro-LSTM can leverage these attributes via ProA. Interestingly, we notice that the CIDEr-D scores may even drop when the Pro-LSTM predicted captions are more accurate and detailed as in Fig. 4(e)-(f). This is due to the fact that some of the image details successfully revealed by Pro-LSTM are actually missing in the human-labeled ground truth, such as *'smoke'* in Fig. 4(e) and *'white'* in Fig. 4(f). Under

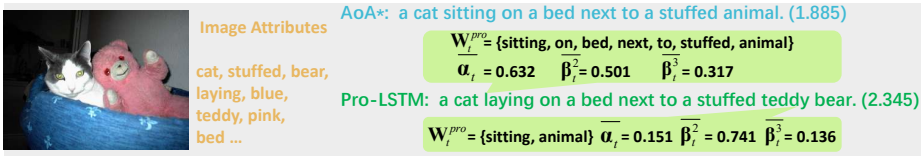

Fig. 5. Qualitative result illustrating how AAD and ProA affect the caption generation process. $\mathbf{W}_t^{pro}$ is the prospective word set, $\bar{\boldsymbol{\alpha}}_t$ is the average weight of prospective information in Eq. 10, $\bar{\boldsymbol{\beta}}_t^2$ and $\bar{\boldsymbol{\beta}}_t^3$ are average weights of attribute information and prospective information in Eq. 13 respectively.

the commonly used token-based metrics which compute exact word matching instead of considering semantic accuracy, this may lead to a performance drop.

We further demonstrate how AAD and ProA affect the captioning model in Fig. 5 by showing the weight of corresponding information when the model generates *'laying'* and *'teddy'*. Our model generates *'laying'* rather than *'sitting'* mainly due to prospective information in the word *'bed'*. The average weight $\bar{\boldsymbol{\alpha}}_t$ of the prospective information is large enough to guide our model to choose the visual features that are more correlated to *'bed'*. Moreover, as the word *'laying'* is detected as one of the image attributes, the prospective information therefore tends to assign relatively larger weights to attribute information in the fusion sub-module in Eq. 12. Consequently, *'sitting'* in the coarse caption is replaced by the more appropriate *'laying'*. When the generation of the whole sentence is about to terminate, e.g. when the Pro-LSTM generates *'teddy'*, the prospective information is not rich enough in helping to generate new words. Nevertheless, thanks to the accurate AAD, *'teddy'* and *'bear'* are successfully detected as image attributes to aid Pro-LSTM in generating a more detailed sentence by replacing *'stuffed animal'* with *'teddy bear'*.

8 Conclusions

We propose a Prospective information guided LSTM (Pro-LSTM) model which comprehensively exploits both prospective and explored linguistic information to boost image captioning. Generally, thorough utilization of the prospective information from the coarse caption makes it possible for the model to attend to proper information from a global perspective. Specifically, with the help of the proposed AAD and ProA module, the Pro-LSTM model can appropriately attend to the object features and image attributes, and adaptively decide when to utilize visual information and when to make use of the language information. As such, sentences with richer and more accurate semantics can be generated. Our method achieves state-of-the-art performances on the benchmark MSCOCO dataset. Comprehensive ablation studies further demonstrate the effectiveness of our method. For future work, we are going to streamline our model to achieve end-to-end training of Pro-LSTM and AAD. This work was supported by the National Natural Science Foundation of China under Grant 61673234.

References

1. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 3156–3164
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Volume 3. (2018) 6
3. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.S.: Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2017) 6298–6306
4. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2017) 3242–3250
5. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of International Conference on Machine Learning. (2015) 2048–2057
6. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 4651–4659
7. Yao, T., Pan, Y., Li, Y., Qiu, Z., Mei, T.: Boosting image captioning with attributes. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 22–29
8. Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring visual relationship for image captioning. In: European Conference on Computer Vision. (2018) 684–699
9. Ke, L., Pei, W., Li, R., Shen, X., Tai, Y.W.: Reflective decoding network for image captioning. arXiv preprint arXiv:1908.11824 (2019)
10. Huang, Y., Chen, J., Ouyang, W., Wan, W., Xue, Y.: Image captioning with end-to-end attribute detection and subsequent attributes prediction. IEEE Transactions on Image Processing **29** (2020) 4013–4026
11. Pan, Y., Yao, T., Li, Y., Mei, T.: X-linear attention networks for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 10971–10980
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. (2017) 5998–6008
13. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. (2015)
14. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., et al.: From captions to visual concepts and back. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 1473–1482
15. Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Carin, L., Deng, L.: Semantic compositional networks for visual captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Volume 2. (2017)

16. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
17. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on attention for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 4634–4643
18. Wang, C., Yang, H., Bartz, C., Meinel, C.: Image captioning with deep bidirectional lstms. In: Proceedings of the 24th ACM international conference on Multimedia, ACM (2016) 988–997
19. Sammani, F., Elsayed, M.: Look and modify: Modification networks for image captioning. arXiv preprint arXiv:1909.03169 (2019)
20. Ge, H., Yan, Z., Zhang, K., Zhao, M., Sun, L.: Exploring overall contextual information for image captioning in human-like cognitive style. arXiv preprint arXiv:1910.06475 (2019)
21. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
22. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proc. the 2014 Conference on Empirical Methods in Natural Language Processing. (2014) 1532–1543
23. Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 5177–5186
24. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)
25. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 1179–1195
26. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 3128–3137
27. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. (2002) 311–318
28. Denkowski, M.J., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the 9th workshop on statistical machine translation. (2014) 376–380
29. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Proc. Text Summarization Branches Out. (2004) 1–8
30. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 4566–4575
31. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: European Conference on Computer Vision, Springer (2016) 382–398
32. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-encoding scene graphs for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 10685–10694