Connecting Language to Images: A Progressive Attention-Guided Network for Simultaneous Image Captioning and Language Grounding

Lingyun Song,[1] Jun Liu,[2] Buyue Qian,[1] Yihe Chen[3]
[1]Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, 710049, China
[2]Guang Dong Xi'an Jiaotong University Academy, Shunde, China
[3]Department of Statistical Science, University of Toronto, Toronto, M5S 3G3, Canada
lingyun.a.song@gmail.com, {liukeen,qianbuyue}@xjtu.edu.cn, yiko.chen@mail.utoronto.ca

Abstract

Image captioning and visual language grounding are two important tasks for image understanding, but are seldom considered together. In this paper, we propose a **P**rogressive **A**ttention-**G**uided **Net**work (PAGNet), which simultaneously generates image captions and predicts bounding boxes for caption words. PAGNet mainly has two distinctive properties: i) It can progressively refine the predictive results of image captioning, by updating the attention map with the predicted bounding boxes. ii) It learns bounding boxes of the words using a weakly supervised strategy, which combines the frameworks of Multiple Instance Learning (MIL) and Markov Decision Process (MDP). By using the attention map generated in the captioning process, PAGNet significantly reduces the search space of the MDP. We conduct experiments on benchmark datasets to demonstrate the effectiveness of PAGNet and results show that PAGNet achieves the best performance.

1 Introduction

Deep neural networks have great advances on the tasks of image understanding, such as object detection and image captioning. Object detection aims to recognize and localize the objects that occur in images, which overlooks the relationships among objects in natural language and thus is far from the end of image understanding. Image captioning aims to compress salient visual information into descriptive language and the state-of-the-art performance is achieved by neural image captioning models (Anderson et al. 2018), which usually adopt the encoder-decoder framework (Cho et al. 2014) consisting of two components: a Convolutional Neural Network (CNN) for image feature extraction and a Recurrent Neural Network (RNN) for caption generation.

Considering the fact that each caption word is always related to partial image contents, neural image captioning models usually incorporate attention mechanisms to allow their models to attend to different parts of the input image. However, in these models the attended regions for predicting each word may be meaningless and inaccurate (Liu et al. 2017), which impedes their performance. Besides, these models cannot predict accurate locations (i.e., the bounding box) for the words of the generated captions, which is inconsistent with the ability of human vision and limits their applicability to vision tasks, such as image annotation (Song et al. 2016) and visual question answering (Anderson et al. 2018).

To address the above limitations, we propose a **P**rogressive **A**ttention-**G**uided **Net**work (PAGNet), which not only generates descriptive sentences for input images, but grounds (i.e., aligns) the words of the generated descriptions to image regions.

1.1 Contributions

The contributions of this paper can be summarized as follows: 1) We combines the tasks of image captioning and language grounding by a PAGNet, which has two appealing properties: i) PAGNet predicts caption sentences for images in a progressive manner. Specifically, the attention over the image can be progressively updated using the grounding results (i.e., bounding boxes) of the words, which enables PAGNet to refine the predictive captions. ii) By combining the frameworks of MDP and MIL, PAGNet can predict bounding boxes for caption words via only image-level annotations, without the requirement of region proposal algorithms to hypothesize target locations. 2)We conduct extensive experiments on several benchmark datasets to evaluate the performance of PAGNet and results show that PAGNet achieves state-of-the-art performance on all the datasets.

2 Related work

2.1 Image captioning

Many approaches (Anderson et al. 2018; Yao et al. 2017; Chen et al. 2016; Xu et al. 2015) based on RNN have been proposed for image captioning. Inspired by the human vision system that selectively processes salient features with attention, some approaches (Xu et al. 2015; Chen et al. 2016) incorporates different attention mechanisms, which allows their models to focus only on related image regions for predicting different words. For example, Chen et al.(2016) proposed a spatial and channel-wise attention-based model, which modulates the sentence generation context in multi-layer feature maps. Anderson et al.(2018) proposed an attention model for image captioning which combines bottom-up and top-down attention mechanisms. Lu et al.(2018) proposed an attention framework called Neural Baby Talk for image captioning, which first

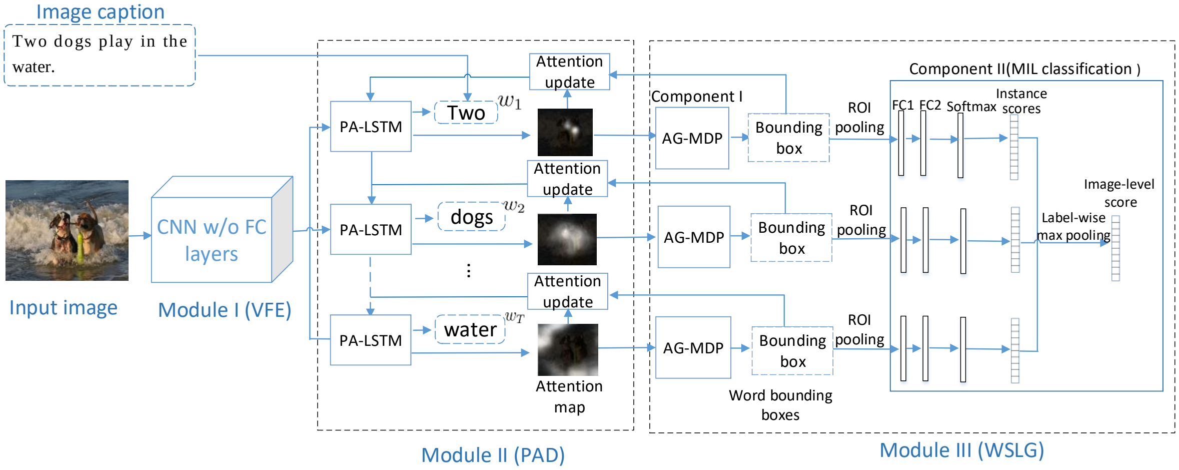

Figure 1: The framework of PAGNet, which consists of three modules: VFE, PAD and WSWA.

generates template sentences that have slot locations tied to image regions, then fills the slots with the concepts detected from the regions using object detectors. The attention used for predicting each word conveys the alignment information from language space to image space, and the correctness of the attention influences the final performance. However, the aforementioned approaches implicitly infer the attention by the hidden states of LSTM, which cannot ensure the correctness of the learned attention (Liu et al. 2017).

2.2 Visual language grounding

This task has been studied by many works (Plummer et al. 2017; 2015; Karpathy and Fei-Fei 2015; Karpathy, Joulin, and Fei-Fei 2014), which align text phrases to image regions. For example, Plummer et al.(2017) proposed a framework for localization or grounding of phrases in images. Karpathy et al. (2014) decomposed images and sentences into fragments and inferred their inter-modal alignment using a ranking objective. Different from these works that only focus on learning correspondence between text phrases and image regions, our model integrates image captioning and language grounding into one framework, which not only can generate captions for images, but can align each word of the generated captions to image regions.

2.3 Object detection

Our model is also related to weakly supervised object detection (WSD) (Durand, Thome, and Cord 2016) and Deep Reinforcement Learning (DRL) based detection (Jie et al. 2016; Caicedo and Lazebnik 2015). Many WSD approaches formulate the object localization as a MIL problem, where each image is represented as a bag of instances. Though localizing objects via only image-level annotations, most WSD approaches depend on region proposal algorithms (Zitnick and Dollár 2014) to hypothesize object locations. This limitation does not exist for DRL based approaches, which cast the object localization as a MDP. For example, Caicedo and Lazebnik; Jie et al.(2015; 2016) proposed to localize objects in images by a MDP, in which an agent is set to deform a bounding box with a sequence of predefined actions. However, during training phase these approaches need ground-truth boxes of objects, which are difficult to collect. In contrast, by combining MIL and DRL, our model localizes words in images using only image labels. Besides, as DRL is a trial-and-error process, its success relies on the agent's luck in achieving the goal by chance in the first place(Lin 1992). The MDP in previous DRL based approaches usually search the locations of objects starting from the whole image, which makes it difficult to shorten the learning time and even leads to a failure.

3 The Proposed Model

3.1 Overview

Figure 1 illustrates the framework of PAGNet, which predicts image captions by using the framework of encoder-decoder, where Module I acts as the encoder and Module II acts as the decoder. Module III predicts bounding boxes for the caption words by combining MIL and DRL. The inputs to PAGNet is an image and its text captions $\{w_t\}_{t=1}^{T}$, where T represents the number of words in the captions. Specifically, Module I is a CNN that encodes the input image into a convolutional feature map. Based on the feature map, Module II predicts captions for the image by a Progressively Attention guided LSTM (PA-LSTM), which can progressively modulate its attention over the image. Besides, PA-LSTM also outputs attention maps that comprises the attention weights over the input feature map. The attention maps are fed into Module III consisting of two components: Attention Guided MDP (AG-MDP) and MIL classification. AG-MDP aligns each caption word to an image region, whereas the MIL classification plays an auxiliary role to align the word to an region that constitutes an object.

3.2 Module I: VFE

The architecture of VFE is shown in Fig. 2. The first five blocks of convolution layers have the same design as the *VGG16* (Simonyan and Zisserman 2015). Compared with higher convolution layers (e.g., the layers in block 5), the lower convolution layers can capture more information for the tiny objects in images. Thus, we combine feature maps from lower and higher convolution layers. Specifically, we concatenate the feature maps from last convolutional layers of the Blocks $3, 4, 5$ to form a unified feature map, i.e., the *Concat* block. To apply the concatenation, the Blocks 3 and 4 are followed by max-pooling layers to synchronize their feature maps to the same size, i.e., the size of the feature maps from Blocks 5. Finally, a 1×1 convolution layer is appended to shrink the channel size of the *Concat* block. The channel size of final feature map is the same as the convolution layer of Block 5.

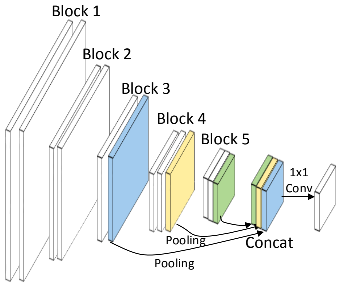

Figure 2: The architecture of Module I.

3.3 Module II: PAD

PAD consists of two components: i) PA-LSTM. ii) Attention update.

PA-LSTM Previous attention based approaches infer the attention using the hidden states of LSTM units, which cannot ensure the correctness of the inferred attention. To solve this limitation, PA-LSTM progressively modulates the attention by the results of the language grounding in Module III. In PA-LSTM, we use the LSTM units (Zaremba, Sutskever, and Vinyals 2014) consisting of an input gate $\mathbf{i}$, a forget gate $\mathbf{f}$, an output gate $\mathbf{o}$, a cell state $\mathbf{c}$ as well as the hidden state $\mathbf{h}$. At time step of t, the interaction between the gates and the hidden state is defined by

$$\mathbf{i}_t = \sigma(W_i\mathbf{y}_{t-1} + U_i\mathbf{h}_{t-1} + E_i\mathbf{x}_t + \mathbf{b}_i), \quad (1)$$
$$\mathbf{f}_t = \sigma(W_f\mathbf{y}_{t-1} + U_f\mathbf{h}_{t-1} + E_f\mathbf{x}_t + \mathbf{b}_f), \quad (2)$$
$$\mathbf{c}_t = \mathbf{f}_t\mathbf{c}_{t-1} + \mathbf{i}_t tanh(W_c\mathbf{y}_{t-1} + U_c\mathbf{h}_{t-1} + E_c\mathbf{x}_t + \mathbf{b}_c), \quad (3)$$
$$\mathbf{o}_t = \sigma(W_o\mathbf{y}_{t-1} + U_o\mathbf{h}_{t-1} + E_o\mathbf{x}_t + \mathbf{b}_o), \quad (4)$$
$$\mathbf{h}_t = \mathbf{o}_t tanh(\mathbf{c}_t), \quad (5)$$

where W, U, Z denote weight matrices, $\mathbf{b}$ denotes the biases, σ is a sigmoid function, $\mathbf{y}_t$ denotes the embedding vector of w_t, $\mathbf{x}_t$ denotes the attentive feature for generating w_t.

Specifically, $\mathbf{x}_t$ is computed by the weighted summation over the vectors of feature set $V = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_m]$, which is obtained by flattening the width W' and the height H' of the input feature map. The parameter $m = W' \times H'$ and each $\mathbf{v}_i \in \mathbb{R}^C$ $(1 \leq i \leq m)$ represents the feature of the i-th location on the feature map, where C represents the channel number of the feature map. The computation of $\mathbf{x}_t$ is formulated by

$$\mathbf{x}_t = \sum_{i=1}^{m} \alpha_{ti}\mathbf{v}_i, \quad (6)$$

where α_{ti} is the attention weight on the i-th location. At time stpe t, the attention map is denoted by $\boldsymbol{\alpha}_t = [\alpha_{t1}, \alpha_{t2}, \cdots, \alpha_{tm}]$, which is computed using a multilayer perceptron followed by a softmax function, and can be formulated by

$$\boldsymbol{\alpha}_t = softmax(M_s\mathbf{s}_t + b), \quad (7)$$
$$\mathbf{s}_t = tanh((M_vV + \mathbf{b}_s) \oplus M_h\mathbf{h}_{t-1}), \quad (8)$$

where $M_v \in \mathbb{R}^{k\times C}$, $M_h \in \mathbb{R}^{k\times d}$ and $M_s \in \mathbb{R}^k$ are transformation matrices that map V and $\mathbf{h}_{t-1}$ to a common space. The $b \in \mathbb{R}^1$ and $\mathbf{b}_s \in \mathbb{R}^k$ are biases, d denotes the dimensionality of $\mathbf{h}_t$. The symbol $\oplus$ represents the addition of a matrix and a vector, which is performed by adding each column of the matrix by the vector. After obtaining $\mathbf{x}_t$, PA-LSTM follows (Xu et al. 2015) to predict the word w_t by a deep output layer (Pascanu et al. 2013) conditioned on $\mathbf{x}_t$, $\mathbf{h}_t$, and w_{t-1}, which is formulated by

$$p(w_t|V, w_{t-1}) \propto exp(P_o(\mathbf{y}_{t-1} + P_h\mathbf{h}_t + P_z\mathbf{x}_t)), \quad (9)$$

where the parameters P_o, P_h, P_z are initialized randomly.

The $\mathbf{h}_t$ and $\mathbf{c}_t$ are initialized by inputting the average of $\mathbf{v}_i$ into two separate MLPs: $\mathbf{c}_0 = f_{init_c}(\frac{1}{m}\sum_{i=1}^{m} v_i)$, $\mathbf{h}_0 = f_{init_h}(\frac{1}{m}\sum_{i=1}^{m} v_i)$, where f_{init_c} and f_{init_h} are the functions of two MLPs.

Attention update In this part, we progressively modulate $\boldsymbol{\alpha}_t$ using the bounding box of w_t output by Module III. The motivation is that the generated bounding box straightly indicates the region highly related to w_t, which is more important than the rest regions. Thus, we update $\boldsymbol{\alpha}_t$ to $\tilde{\boldsymbol{\alpha}}_t$ by

$$\tilde{\boldsymbol{\alpha}}_t = \boldsymbol{\alpha}_t \times \mathbf{z}_t, \quad (10)$$

where $\mathbf{z}_t = [z_{t1}, z_{t2}, \cdots, z_{tm}]$ is a binary coefficient vector and each element $z_{ti} \in \{0, 1\}$. If the i-th image location is in the predicted bounding box of w_t, z_{ti} is set to 1, otherwise set to 0. In this way, the attention of PA-LSTM is progressively led to the regions that are highly related to the generated word, which prevents PA-LSTM from focusing on meaningless regions.

With $\tilde{\boldsymbol{\alpha}}_t$, PA-LSTM can refine the predictive captions. Specifically, we first substitute Eq. (10) into Eq. (6) to compute new attentive feature $\tilde{\mathbf{x}}_t$, which is then substituted into Eq.(1) - Eq. (4) to compute new $\tilde{\mathbf{h}}_t$. Finally, $\tilde{\mathbf{h}}_t$ and $\tilde{\mathbf{x}}_t$ are substituted into Eq. (9) to predict new captions. Besides, $\tilde{\mathbf{h}}_t$ is also substituted into Eq.(7) and (8) to compute new attention maps, which are fed into Module III to start new episodes. The details are described in Subsection 3.4.

3.4 Module III: WSLG

As shown in Fig. 1, WSLG is a multi-group network, where each group corresponds to one caption word and consists of two components: i) AG-MDP. ii) MIL classification.

AG-MDP MDP has been applied to localize objects in many works (Jie et al. 2016) and achieves promising performance. However, existing MDP searches targets starting from the whole image, which greatly increases the difficulty of localizing target boxes with such large state space.

To solve this problem, we propose a new AG-MDP that can be directed to first explore the most promising region containing targets, by using the attention map output by Module II. Specifically, an attention box generated from the attention map is used as the starting window for the MDP. The attention region corresponding to each word is generated on the attention map by selecting a square region which has the largest attention weights. The size of the selected region is set to 4×4, which corresponds to a 64×64 patch in the original image. We also experimentally vary the size to 8×8 and observe that the performance changes slightly.

As the attention map is progressively updated in Module II, thus the staring window of each episode in MDP will progressively approximate the correct location of targets, which is conducive to reducing the difficulty of localizing targets and improving the localization accuracy. Note that AG-MDP is performed on the last feature map of Module I, rather than the original image. AG-MDP has a set of actions A, a set of states S, and a reward function R, which are described as below.

State: The state representation is the concatenation of two components: a feature vector $\mathbf{b}$ of the current observed window and a memory vector $\mathbf{v}_h$ that captures the last ten actions selected by the agent. The vector $\mathbf{b}$ is generated based on the last feature map V of Module I and the attention weights $\boldsymbol{\alpha}$. Specifically, we first modulate the V to $\tilde{V}$ by $\tilde{V} = V \cdot \boldsymbol{\alpha} = [\tilde{v}_1, \tilde{v}_2, \cdots, \tilde{v}_m]$. Then, a ROI pooling layer (Girshick 2015) is added on top of $\tilde{V}$ to obtain a fixed-length feature vector of the current window. The history vector $\mathbf{v}_h$ is a binary vector that indicates which actions have been taken in the past. Each action in the history vector is represented by a 1-of-K vector, where only one element corresponding to the taken action is 1 and all other elements are 0. As there are 15 different actions presented in the following section, the $\mathbf{v}_h \in \mathbb{R}^{150}$.

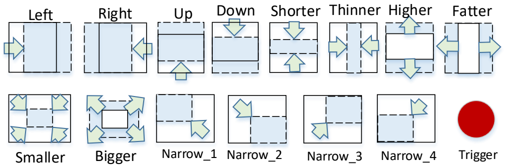

Figure 3: Illustration of the actions used in the MDP. Blue windows with dashed lines represent the windows obtained by taking one of the actions.

Actions: Figure 3 illustrates the set of actions A, which consists of fourteen *transformation* actions that can be used to deform the observed window, and one *Trigger* action used to stop the searching process. The *transformation* actions can be categorized into three groups: i) Move actions that aim to change the location of the boxes. This group contains four actions, each of which moves the window by 0.2 times of the current window size. ii) Scaling actions that aim to proportionally change the scales of the windows. This group contains six scaling actions, each of which scales the window by 0.25 times of the current window size. iii) Ratio actions that aim to modify aspect ratios of the windows. This group contains four ratio actions, each of which changes the horizontal/vertical size by 0.15 times of the current window size. The *trigger* action indicates that the agent has correctly localized an object and the sequence of the current search should be terminated.

Rewards: At time step t' of MDP, the agent receives a reward $R_{a_{t'}}$ for its action $a_{t'}$, which moves the state from $s_{t'}$ to $s_{t'+1}$. Each state $s_{t'}$ has an associated window $b_{t'}$.

As the grounding-truth bounding boxes of each word are unavailable, PAGNet measures the reward $R_{a_{t'}}$ by

$$R_{a_{t'}} = \begin{cases} 1 & \tau(\mathbf{b}_{t'+1}, \mathbf{y}_t) - \tau(\mathbf{b}_{t'}, \mathbf{y}_t) > 0.05 \\ -1 & \tau(\mathbf{b}_{t'+1}, \mathbf{y}_t) - \tau(\mathbf{b}_{t'}, \mathbf{y}_t) < 0 \\ 0 & otherwise \end{cases} \tag{11}$$

where $\mathbf{b}_{t'+1}$ and $\mathbf{b}_{t'+1}$ denote the feature vector of the window $b_{t'+1}$ and $b_{t'}$, respectively. The $\mathbf{y}_t$ denotes the embedding vector of w_t, and $\tau\left(\mathbf{b}, \mathbf{y}_t\right)$ denotes the semantic similarity between $\mathbf{b}_{t'}$ (or $\mathbf{b}_{t'+1}$) and $\mathbf{y}_t$. The semantic similarity is evaluated by the method in (Plummer et al. 2015), which first learns an embedding of the window and word features to a shared latent space using canonical correlation analysis (Hotelling 1936), and then uses cosine distance in that space to score the semantic similarity.

As seen from Eq. (11), when state $s_{t'}$ is changed to state $s_{t'+1}$, the reward function returns $+1$ if the improvement of the similarity τ is larger than 0.05, returns -1 if τ decreases, or returns 0 otherwise. As the *trigger* does not transform the box, we define its reward by

$$R_{a_{t'}} = \begin{cases} +3 & \tau(\phi(b_{t'}), \mathbf{y}_t) > \eta \\ -3 & otherwise \end{cases} \tag{12}$$

where η is a threshold that indicates the minimum semantic similarity allowed to consider the word is correctly localized. We set $\eta = 0.8$ during the training phase.

Localization policy: In MDP, the agent selects actions according to a policy. We learn the optimal policy by the deep Q-learning algorithm (Mnih et al. 2015), which estimates the value of each state-action pair using a deep Q-network. Specifically, the deep Q-network is a multi-layer neural network containing two hidden layers, which takes the state representation as input, and predicts a vector of action values $Q(s, a; \theta_q)$. The θ_q are the parameters of the neural network. The numbers of neurons in both hidden layers are 1024.

During training, the agent is set to interact with the environment in multiple episodes. Each episode starts from the initial state and ends when the action *trigger* is selected. The agent's behavior during training is ϵ-greedy (Sutton and Barto 1998). However, different from the previous MDP that immediately restarts a new episode after one episode terminate, our model returns to train Module II to refine the predictive captions after one episode is completed. Detailed training strategy is present in Subsection 3.5.

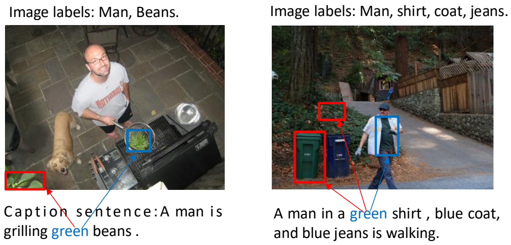
Image labels: Man, Beans.
Caption sentence:A man is grilling green beans .
Image labels: Man, shirt, coat, jeans.
A man in a green shirt , blue coat, and blue jeans is walking.

Figure 4: Examples of word-region alignments generated by AG-MDP.

MIL classification Although the reward function in AG-MDP tends to find regions that has high semantic similarity with the caption words, it still cannot ensure the obtained regions are meaningful and described by the corresponding caption words. For example, as shown in Fig. 4, AG-MDP wrongly grounds the caption word in blue to the red boxes, whereas the ground-truth is the blue boxes. This occurs because the captions usually describe the objects of images. However, the reward function used in PA-LSTM cannot guide the words to be grounded to the object regions.

To solve this problem, after AG-MDP we append a MIL classification network, which takes the bounding boxes generated in AG-MDP as inputs, to perform MIL image classification. Specifically, each grounding box is treated as an instance and one ROI pooling layer is used to extract the feature representation of each instance. The labels used for MIL classification refer to the objects (i.e., the nouns in captions) of images, and in this way the caption words are enforced to be grounded to the regions that constitute objects. As shown in Fig.1, the MIL classification network consists of two Fully Connected (FC) layers with ReLU activation, one FC layer with softmax output for instance-level classification scores. The final image-level classification scores are computed using a max-pooling on the instance-level scores.

Objective function We jointly train AG-MDP and the MIL classification network. The objective function $J(\theta)$ is

$$J(\theta) = \max_{\theta_q,\theta_c}(\frac{1}{T}\sum_{t=1}^{T} R_t(\theta_q) - \lambda H(\theta_c, \mathbf{p}^n, \tilde{\mathbf{y}}^n)), \quad (13)$$

where θ_q and θ_c are the parameters of the deep Q-network and the classification network, respectively. T represents the number of time steps in PA-LSTM. $R_t(\theta_q)$ denotes the average of the expected reward over N samples and T' time steps in AG-MDP. Specifically, $R_t(\theta_q)$ is computed by

$$R(\theta_q) = \frac{1}{NT'}\sum_{n=1}^{N}\sum_{t'=1}^{T'} \mathbb{E}[((r(a^n_{t'}) + \gamma \max a^n_{t'+1} Q(s^n_{t'+1}, \\ a^n_{t'+1}; \theta_q) - Q(s^n_{t'}, a^n_{t'}; \theta_q))^2], \quad (14)$$

where γ denotes the discount factor. The function $H(\theta_c, \mathbf{p}^n, \tilde{\mathbf{y}}^n)$ denotes the average sigmoid cross entropy loss over N training samples.

$$H(\theta_c, \mathbf{p}^n, \tilde{\mathbf{y}}^n) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}[\tilde{y}^n_k \times \log(p^n_k) + (1 - \tilde{y}^n_k) \\ \times \log(1 - p^n_k)], \quad (15)$$

where $\mathbf{p}^n = [p^n_1, p^n_2, \cdots, p^n_K]$ denotes the predictive score vector of the n-th image with respect to each label, $\tilde{\mathbf{y}}^n = [\tilde{y}^n_1, \tilde{y}^n_2, \cdots, \tilde{y}^n_K]$ denotes the corresponding ground-truth label vector and K is the total number of labels.

3.5 Training strategy

During training, we first train Module I separately to obtain visual features of the input image, then fix the layers of Module I and train module II and III alternatively. Specifically, we first train Module II to predict image captions and attention maps. Then, we fix Module II and fed the attention maps into Module III, which searches the bounding boxes for each word by a MDP. The starting window of the MDP is generated using the input attention maps. After obtaining the bounding boxes of all the caption words, we fix Module III and input the generated bounding boxes to the component *Attention update* of Module II, for modulating the attention maps and refining the results of image captioning.

The motivation behind this alternative strategy is that: i) The attention map provides important clues to find the regions related to words, and thus can act as a *teacher* to guide AG-MDP of Module III to first explore the most promising region that contains target objects. The more accurate the attention map is, the more effective it is in reducing the difficulty of the localization and improving the accuracy of localization. ii) With the progress of training, the accuracy of the bounding boxes of caption words output by Module III becomes higher and higher, which is conducive to gradually improving the accuracy of the attention maps. We repeat the alternative training several times for better performance.

4 Experiments

4.1 Datasets

For the task of image captioning, we report the experimental results obtained by different models on the the COCO (Lin et al. 2014) and Flickr30k Entities (Plummer et al. 2015) datasets. COCO contains $123,000$ images and Flickr30k Entities contains $31,783$ images. Flickr30k Entities augments the Flickr30k dataset (Young et al. 2014) with bounding boxes for each entity (noun phrases) of image captions. The entities are categorized into eight types including people (peo.), body parts (body.), animals (ani.), clothing/color (clo.), instruments (ins.), vehicles (veh.), scene (sce.), and other. In some cases where some phrases correspond to multiple boxes, we follow (Plummer et al. 2015) to treat the union of the boxes as ground truth. Each image of these two datasets has at least five ground-truth captions.

To make our results comparable to others, we use the publicly available splits[1] of training, testing and validating

[1] https://github.com/karpathy/neuraltalk

sets for both Flickr30k Entities and COCO. Specifically, for COCO dataset we use 113,287 images for training, $5,000$ images for both validation and testing. For Flickr30k Entities dataset we use 1,000 images for validation, $1,000$ images for testing and the rest for training. We convert all sentences to lowercase, filter non-alphanumeric characters and words that occur less than 5 times in the training set, which results in $7,414$, and $8,791$ words for Flickr30K Entities and COCO datasets, respectively.

4.2 Evaluation metrics

For the task of image captioning, we use BLEU (B1, B2, B3, B4)[2] (Papineni et al. 2002), METEOR (M) (Banerjee and Lavie 2005), CIDEr (C) (Vedantam, Lawrence Zitnick, and Parikh 2015), and ROUGE-L (R) (Lin 2004) as evaluation metrics. The scores of these metrics are computed with the codes[3] released by COCO Evaluation Server.

The task of language grounding is akin to object detection, we evaluate the performance using the mean Average Precision (mAP), which is computed over all the regions processed by the agent during the episodes in AG-MDP. we consider that a word is correctly grounded if the predicted bounding box has an IoU ratio of at least 50% with the corresponding ground-truth bounding box. We only report the experimental results of language grounding on Flickr30k Entities, as the ground-truth of the bounding boxes of the words in COCO is not available.

4.3 Implementation details

The size of input images is 224×224. The length of the captions longer than 18 in COCO or 22 in Flickr30k Entities are truncated. Each caption word is represented as a 300-D word2vec (Mikolov et al. 2013) feature. During training, the layers of the first five convolution blocks in Module I are initialized with the weights of the corresponding layers of the VGG16 pre-trained on ImageNet. The rest layers of Module I are randomly initialized with Gaussian distributions $G(\mu;\sigma)$, where $\mu = 0$ and $\sigma = 0.01$. We train PAGNet with SGD at a learning rate of 0.0001 and a mini-batch size of 64. The momentum and weight decay are set to 0.9 and 0.0005, respectively. During the training and testing, we align each word of the generated captions to image regions. The words other than nouns are aligned to the regions that constitute an object or a part of an object. This is because image descriptions often make frequent references to objects that occur in images. During testing, for the task of language grounding, we use the Stanford parser to identify nouns of the generated captions and evaluate the performance only on these nouns, because other words (e.g. determiner and preposition) do not have accurate location. In Flickr30k Entities dataset, the bounding boxes associated with noun-phrases is viewed as the ground-truth bounding boxes of the nouns in the phrases. As the generated captions are often different from the ground-truth captions, we evaluate the performance only on the matched nouns between the generated captions and the ground-truth captions.

4.4 Experimental results

Ablation study To reveal the contribution of each component, we test the performance of PAGNet with different configurations, including: i) PAGNet-1, which is obtained by removing the component *Attention update* from Module II. Without *Attention update*, the bounding boxes output by Module III cannot be used to update the attention and thus the predicted captions cannot be refined. ii) PAGNet-2, which is obtained by removing the component *MIL classification* from Module III.

Table 1: Results (%) of the ablation study on COCO test set.

Algorithms	B1	B2	B3	B4	M	R	C
PAGNet-1	68.6	51.8	36.7	28.5	23.2	50.8	86.1
PAGNet-2	76.3	59.7	44.9	35.8	26.9	55.2	110.3
PAGNet	**83.2**	**62.8**	**46.3**	**40.8**	**30.4**	**58.6**	**118.6**

Table 2: Results (%) of the ablation study on Flickr30k Entities test set.

Algorithms	B1	B2	B3	B4	M	R	C
PAGNet-1	65.6	43.3	30.4	21.2	19.1	53.3	49.6
PAGNet-2	70.3	51.2	38.3	27.5	22.6	57.9	54.3
PAGNet	**74.8**	**55.8**	**41.2**	**30.7**	**25.2**	**61.4**	**57.5**

The experimental results of the ablation study on image captioning are shown in Table 1 for *COCO* dataset and Table 2 for *Flickr30k Entities* dataset. From the tables, we can observe that PAGNet (w/o update) exhibits the worst performance in all metrics, which indicates that the component *Attention update* matters more than the component *MIL classification*. Without the *Attention update*, PAGNet (w/o update) cannot update attention weights for each word and thus lose the ability to refine the predictive captions, which results in a significant performance drop. With the incorporation of the *Attention update*, PAGNet (w/o MIL) outperforms PAGNet (w/o update) on both experimental datasets. However, without the MIL classification network, PAGNet (w/o MIL) cannot ensure that words are grounded to the regions of objects, which may lead the PA-LSTM to focus on meaningless background clutter and thus impedes the performance.

Table 3: AP (%) values of the ablation experiments on Flickr30k Entities test set.

Algorithms	peo.	body.	ani.	clo.	ins.	veh.	sce.	other	mAP
PAGNet-1	42.8	19.4	22.8	26.4	28.5	39.2	36.4	27.9	30.4
PAGNet-2	48.4	24.5	33.7	31.1	29.2	44.1	38.9	29.3	34.9
PAGNet	**54.6**	**28.4**	**35.8**	**39.2**	**34.3**	**52.6**	**46.2**	**34.5**	**40.7**

As the ground-truth bounding boxes of caption words are not available in COCO dataset, we only conduct an ablation study on *Flickr30k Entities* dataset for language grounding and show the results in Table 3. We can see that PAGNet (w/o update) performs worse than PAGNet (w/o MIL),

[2] Bn is the geometric average of the n-gram precision.

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

which indicates *Attention update* has a larger influence on language grounding. For example, compared with PAGNet (w/o update), PAGNet improves the mAP from 30.4% to 40.7%, with a 10.3% margin at most. This is because the introduction of the *Attention update* enables PA-LSTM to progressively update the attention map corresponding to each word, which results in a more accurate starting window for AG-MDP. Compared with PAGNet (w/o MIL), PAGNet improves the performance by 5.8%, which reveals that the *MIL Classification* does contribute to improve the performance of language grounding. The *MIL Classification* can be viewed as a discriminator to penalize the agents that align words to background clutter.

Comparison with other models To the best of our best knowledge, no previous works have conducted the experiments of grounding the words of the generated captions to image regions. Therefore, we only compare PAGNet with other models on the task of image captioning.

Table 4: Experimental results (%) of image captioning for different models on COCO dataset.

Algorithms	B1	B2	B3	B4	M	R	C
M-RNN	62.5	45.0	32.1	23.0	19.5	-	66.0
LSTM+Attn	71.8	50.4	35.7	25.0	23.0	-	-
Adaptive-Attn	74.2	58.0	43.9	33.2	26.6	-	108.5
LSTM-A	73.5	56.6	42.9	32.4	25.5	53.9	99.8
CNN+attn	71.1	53.8	39.4	28.7	24.4	52.2	91.2
up-down	77.2	-	-	36.2	27.8	56.4	113.5
PAGNet	**83.2**	**62.8**	**46.3**	**40.8**	**30.4**	**58.6**	**118.6**

Table 5: Experimental results (%) of image captioning for different models on Flickr30k entities dataset.

Algorithms	B1	B2	B3	B4	M	R	C
M-RNN	57.3	36.9	24.0	15.7	-	-	-
LSTM+Attn	66.9	43.9	29.6	19.9	18.5	-	-
Adaptive-Attn	67.7	49.4	35.4	25.1	20.4	-	53.1
PAGNet	74.8	55.8	41.2	30.7	25.2	61.4	57.5

The baseline models include non-attention models and attention-based models. The non-attention models includes *M-RNN*(Karpathy and Fei-Fei 2015) and *LSTM-A*(Yao et al. 2017). The attention-based models includes *LSTM+Attn*(Xu et al. 2015), *Adaptive-Attn*(Lu et al. 2017), *CNN+attn*(Aneja, Deshpande, and Schwing 2018), *up-down*(Anderson et al. 2018). The experimental results on *COCO* and *Flickr30k Entities* datasets are shown in Table 4 and Table 5, respectively. We also show some qualitative results and analysis in supplementary material. As shown in the tables, the results across all evaluation metrics consistently indicate that PAGNet achieves better performance than all other models. In particular, taking the results on *COCO* as an example, PAGNet makes the relative improvement over the non-attention models by at least 8.4%, 4.9%, 4.7%, 18.8% in BLUE(B-4), METEOR, ROUGR-L and CIDEr, respectively. PAGNet also outperforms all the attention-based models, improves the state-of-the-art on B-4 from 36.2% to 40.8%, METEOR from 27.8% to 30.4%, ROUGR-L from 56.4% to 58.6%, and CIDEr from 113.5% to 118.6%. Similarly, on *Flickr30k Entities*, PAGNet improves the state-of-the-art with a large margin.

The superior performance of PAGNet can be attributed to the following two reasons: i) PAGNet can progressively update the attention map for predicting each word. The correctness of the attention map is a key factor in improving the performance of image captioning. However, these baseline models cannot ensure the correctness of the attention (Liu et al. 2017). ii) By combining the framework of MIL and MDP, PAGNet grounds the word to the regions of the objects in images, which prevents the PA-LSTM from focusing on meaningless background clutter. In contrast, the baselines allow their models to attend any visual parts of images.

5 Conclusion

In this paper, we propose PAGNet that combines image captioning and language grounding, which can be mutually reinforced to improve their performance. Specifically, PAGNet generates captions for input images by a PA-LSTM, which can progressively update the attention corresponding to each word by using the outputs of language grounding. PAGNet grounds each word to image regions by a AG-MDP, which searches target bounding boxes starting from the window generated by the attention map output by PA-LSTM, rather than the whole image. Experimental results show that PAGNet outperforms all the baselines on all datasets.

6 Acknowledgments

This work was supported by National Key Research and Development Program of China (2018YFB1004500), National Natural Science Foundation of China (61532015, 61532004, 61672419, and 61672418), Innovative Research Group of the National Natural Science Foundation of China(61721002), Innovation Research Team of Ministry of Education (IRT_17R86), Project of China Knowledge Centre for Engineering Science and Technology, Science and Technology Planning Project of Guangdong Province (No. 2017A010101029), the program of China Scholarship Council (201706280198).

References

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Aneja, J.; Deshpande, A.; and Schwing, A. G. 2018. Convolutional image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Banerjee, S., and Lavie, A. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Caicedo, J. C., and Lazebnik, S. 2015. Active object localization with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 2488–2496.

Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; and Chua, T.-S. 2016. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. *arXiv preprint arXiv:1611.05594*.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Durand, T.; Thome, N.; and Cord, M. 2016. Weldon: Weakly supervised learning of deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4743–4752.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

Hotelling, H. 1936. Relations between two sets of variates. *Biometrika* 28(3/4):321–377.

Jie, Z.; Liang, X.; Feng, J.; Jin, X.; Lu, W.; and Yan, S. 2016. Tree-structured reinforcement learning for sequential object localization. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, 127–135.

Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3128–3137.

Karpathy, A.; Joulin, A.; and Fei-Fei, L. F. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, 1889–1897.

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 *Proceedings of the European Conference on Computer Vision*, 740–755.

Lin, L.-J. 1992. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine learning* 8(3-4):293–321.

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

Liu, C.; Mao, J.; Sha, F.; and Yuille, A. L. 2017. Attention correctness in neural image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4176–4182.

Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017. 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*, volume 6, 2.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2018. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7219–7228.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, 3111–3119.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540):529.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318.

Pascanu, R.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2013. How to construct deep recurrent neural networks. *arXiv preprint arXiv:1312.6026*.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, 2641–2649.

Plummer, B. A.; Mallya, A.; Cervantes, C. M.; Hockenmaier, J.; and Lazebnik, S. 2017. Phrase localization and visual relationship detection with comprehensive image-language cues. In *Proceedings of the IEEE International Conference on Computer Vision*.

Simonyan, K., and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*.

Song, L.; Luo, M.; Liu, J.; Zhang, L.; Qian, B.; Li, M. H.; and Zheng, Q. 2016. Sparse multi-modal topical coding for image annotation. *Neurocomputing* 214:162–174.

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

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4566–4575.

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 *Proceedings of the International Conference on Machine Learning*, 2048–2057.

Yao, T.; Pan, Y.; Li, Y.; Qiu, Z.; and Mei, T. 2017. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, 22–29.

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

Zaremba, W.; Sutskever, I.; and Vinyals, O. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.

Zitnick, C. L., and Dollár, P. 2014. Edge boxes: Locating object proposals from edges. In *Proceedings of the European Conference on Computer Vision*, 391–405.