More Grounded Image Captioning by Distilling Image-Text Matching Model

Yuanen Zhou\textsuperscript{1,2}, Meng Wang\textsuperscript{1,2,*}, Daqing Liu\textsuperscript{3}, Zhenzhen Hu\textsuperscript{1,2}, Hanwang Zhang\textsuperscript{4}

\textsuperscript{1}Key Laboratory of Knowledge Engineering with Big Data (Ministry of Education)
\textsuperscript{2}School of Computer Science and Information Engineering, Hefei University of Technology
\textsuperscript{3}University of Science and Technology of China
\textsuperscript{4}Nanyang Technological University

\{y.e.zhou.hb, eric.mengwang, huzhen.ice\}@gmail.com, liudq@mail.ustc.edu.cn, hanwangzhang@ntu.edu.sg

Abstract

Visual attention not only improves the performance of image captioners, but also serves as a visual interpretation to qualitatively measure the caption rationality and model transparency. Specifically, we expect that a captioner can fix its attentive gaze on the correct objects while generating the corresponding words. This ability is also known as grounded image captioning. However, the grounding accuracy of existing captioners is far from satisfactory. To improve the grounding accuracy while retaining the captioning quality, it is expensive to collect the word-region alignment as strong supervision. To this end, we propose a Part-of-Speech (POS) enhanced image-text matching model (SCAN [24]): POS-SCAN, as the effective knowledge distillation for more grounded image captioning. The benefits are two-fold: 1) given a sentence and an image, POS-SCAN can ground the objects more accurately than SCAN; 2) POS-SCAN serves as a word-region alignment regularization for the captioner’s visual attention module. By showing benchmark experimental results, we demonstrate that conventional image captioners equipped with POS-SCAN can significantly improve the grounding accuracy without strong supervision. Last but not the least, we explore the indispensable Self-Critical Sequence Training (SCST) [46] in the context of grounded image captioning and show that the image-text matching score can serve as a reward for more grounded captioning \textsuperscript{1}.

1. Introduction

Image captioning is one of the primary goals of computer vision which aims to automatically generate free-form descriptions for images [23, 53]. The caption quality has been dramatically improved in recent years, partly driven by the development of attention-based deep neural networks [56], which allow the captioning models to dynamically align image regions to caption words. Conventionally, many previous works are used to qualitatively show the attention visualizations, which aim to indicate that the learned model can fix its gaze on the correct regions while captioning. However, some quantitative analyses [28, 38] show that although
the models can achieve impressive caption quality, they still suffer from poor attention grounding. This may lead to undesired behaviors such as object hallucinations [47] and gender discrimination [14], which harm the rationality and explainability of the neural image captioning models.

There are some efforts for more grounded image captioners. Most of them supervise the learning process by the attention module [28, 65, 36]. However, they require fine-grained region-word alignment annotations, which are expensive to collect. Therefore, in this paper, we want to supervise the visual attention without region-word alignment annotations. To this end, we propose a novel knowledge distillation [15, 34, 63] approach to regularize the visual attention in captioner, by treating an image-text matching model as a weak supervision of grounding [19, 48]. By “weak”, we mean that the image-text model training only relies on the image-text alignment but not the expensive word-region alignment. The key motivation of our knowledge distillation is that compared to the caption generation task, the image-text matching task [9, 24] is a more well-posed one, because 1) the latter doesn’t have to take the sentence grammar and fluency into account, and 2) the training loss for the latter’s metric (accuracy on matched or not) is more objective and faithful to the task; while for the former’s, such as the word-level cross-entropy and sentencellevel CIDEr [52], still has a well-known gap with human judgment.

As shown in Figure 1 (a) and (b), the attention of the matching model (a) (the POS-SCAN introduced later) is more focused and reliable, e.g., it aligns shirt and restaurant to the correct regions, while the captioning model (b) doesn’t. Therefore, it is reasonable to supervise the visual attention module of a captioning model by using an image-text matching model. In this way, the image-text matching model serves as an independent “teacher” that doesn’t couple with the “student” captioning model. Note that the “independence” can avoid the model collapse of the teacher and student who are trained from the same task [38, 41].

Specifically, we use a state-of-the-art image-text matching model termed SCAN [24], which will be detailed in Section 3.1. The reason why we choose SCAN is that it can serve as a weakly-supervised visual grounding model with local region-word alignment (though it is a by-product in the original paper [24]). Note that our approach can be integrated with any matching model with a local alignment module like SCAN. Though SCAN shows good performance in image-text matching, we surprisingly find that the original SCAN model has no better grounding performance than a popular baseline: Up-Down captioning model [3]. As qualitatively shown in Figure 1, its alignment (c) is no better than the captioning model (b). We also quantitatively report their attention accuracy in Table 1: the attention accuracy of SCAN is 17.63%, while that of Up-Down is 19.83%. A plausible reason is that some non-noun words that hurt grounding are however beneficial to fit the matching model. For example, grounding non-visual function words (“a”, “the”), prepositions (“on”, “of”, “with”), and visual relationship verbs (“ride”, “jump”, “play”) are inherently challenging even with word-region strong supervision [44], not to mention for the weakly-supervised setting. Therefore, a high matching score based on all the words is possibly attributed to the bias of certain word collocations, which are widely observed in a large spectrum of vision-language tasks [58, 59, 51].

In this paper, we propose a simple but effective method to remedy the above problem. Specifically, we only keep the noun words when computing the matching score with the help of a Part-of-Speech (POS) tagger. After this, the grounding performance of the re-trained POS enhanced SCAN (POS-SCAN) model meets the requirement of the downstream task. Note that the reason why we call it POS-SCAN but not merely noun-SCAN is: we can seamlessly incorporate other POS if its visual grounding ability matures in the future. During inference, the matching model can be fully removed and there is no extra computing overhead. Without any region-word alignment annotations, our method can achieve better performance in terms of both caption quality and attention accuracy on the challenging Flickr30k Entities dataset [44].

Last but not the least, we explore the indispensable Self-Critical Sequence Training (SCST) [46] in the context of grounded image captioning. We find that although a captioning model obtains higher scores using the standard SCST metrics (e.g., CIDEr [52]), it achieves worse grounding performance. Fortunately, when we incorporate SCAN as the reward, the captioning model is encouraged to generate captions that are more faithful to the image while retaining the standard metric scores. However, when we use POS-SCAN as the reward, we empirically discover significantly worse results in terms of standard metrics, but better grounding results. By knowing that POS-SCAN is a better grounding model than SCAN, we are indeed facing a dilemma: captioning vs. grounding, whose metrics should be unified in the future. We hope that our study can offer a promising direction towards more grounded image captioning.

2. Related Work

Image Captioning. Earlier approaches for image captioning are rule-template-based [23, 40, 26]. Recently, attention-based neural encoder-decoder models prevail [53, 56, 35, 6, 60, 29, 58, 59]. Attention mechanisms have been operated on uniform spatial grids [56, 35], semantic meta-data [61, 57, 12], and object-level regions [3, 18, 60, 64]. Although attention mechanisms are generally shown to improve caption quality, some quantitative analyses [28, 38]
show that the “correctness” of the attention is far from satisfactory. This makes models less trustworthy and less interpretable. There are some efforts for more grounded image captioning. Lu et al. [36] proposed a slot-and-fill framework for image captioning that can produce natural language explicitly grounded in entities. In [28, 65], attention module is explicitly supervised. However, such methods require fine-grained region-word alignment annotations, which are expensive to collect. Although Ma et al. [38] proposed a cyclical training paradigm that requires no alignment annotations, their method has difficulty in providing sufficient attention supervision. This is because their localizer and decoder are learned jointly and coupled loosely in the attention module, easily resulting in modal collapse [41].

Image-Text Matching. The image-text matching methods can be roughly categorized into global alignment based and local alignment based. Global alignment based methods [10, 21, 54, 9, 55] map the holistic image and the full sentence into a joint semantic space. A representative global image-text matching model VSE++ [9] has been adopted in [37, 33] to improve the discriminability of generated captions. In contrast, local alignment based methods [19, 42, 24] typically infer the global image-text similarity by aligning visual objects to textual words and make image-text matching more fine-grained and interpretable. In this work, we adopt the classic local image-text matching model SCAN [24] to serve as a reinforced reward and the proposed POS-SCAN to serve as an attention supervision.

Visual Grounding. Visual grounding is the general task of locating the components of description in an image. In terms of the learning fashion, methods can be roughly divided into three categories: supervised, unsupervised and weakly supervised. Many works [39, 32, 5, 17, 62, 30] belong to the first category which requires expensive ground truth annotations. Some works [48, 4] attempt to learn by reconstruction without supervision. There are also works [19, 31, 7] which use weak supervision from image-captions pairs to perform visual grounding. Datta et al. [7] recently proposed a weakly supervised grounding model, which can also be adopted in our framework. We leave this as our future work.

Knowledge Distillation. Since Hinton et al. [15] proposed to distill the knowledge from an ensemble of models into a single model, there are a lot of follow-up works, including exploring different forms of knowledge [49, 25], cross-modality distillation [13, 1], cross-task distillation [34, 63]. Here, we only mention some representative similar works, a comprehensive survey is beyond the scope of this paper. Liu et al. [34] proposed to boost multi-label classification by distilling knowledge from a weakly-supervised detection task. Yuan et al. [63] proposed to transfer knowledge from image captioning and classification model to text-to-image synthesis model. In this work, we aim to boost the attention accuracy of the image captioning model (student with hard task) by distilling knowledge from the image-text matching model (teacher with easy task).

3. Approach

Our model comprises of two main components: a neural image caption generator and an image-text matching model, as shown in Figure 2. We will first describe the two components used in our experiments, then elaborate on how we combine the two components in a collaborative framework to generate more grounded captions. We denote the input image as $I$, which is represented by a set of regions feature $[f_1, \ldots, f_k] \in \mathbb{R}^{k \times d}$ extracted by a detector [45]. The corresponding ground truth and generated sentence $T$ with $n$ words are represented as $(y_1^*, \ldots, y_n^*)$ and $(y_1, \ldots, y_n)$, respectively.

3.1. Image-Text Matching Model

In this work, we extend the classic image-text matching model SCAN [24] to serve as a fine-grained rewarder and the POS enhanced SCAN to serve as an attention guider. SCAN is a matching model that discovers the full latent alignment using both image regions and words in a sentence as context then infers image-text similarity. Here, we only focus on the adopted text-image formulation. Specifically, given an image $I$ and a sentence $T$, it first transforms each region feature $f_i$ to appropriate dimension by:

$$v_i = W_v f_i + b_v, \quad v_i \in \mathbb{R}^{d_v},$$

and employs a bi-directional GRU [50] to embed the words:

$$x_t = W_e y_t^*, \quad \hat{h}_t = GRU(x_t), \quad \hat{h}_t = GRU(x_t),$$

where $W_e$ is an embedding matrix. The final word feature $e_t$ is the average of the forward hidden state $\hat{h}_t$ and backward hidden state $\hat{b}_t$:

$$e_t = \frac{\hat{h}_t + \hat{b}_t}{2}, \quad t \in [1, n].$$

Then the cosine similarity matrix for all possible pairs is computed as follows:

$$s_{it} = \frac{v_i^T e_t}{\|v_i\|\|e_t\|}, \quad i \in [1, k], t \in [1, n].$$

Here, $s_{it}$ denotes the similarity between the $i$-th region and the $t$-th word is normalized as $\pi_{it} = \frac{s_{it}}{\sqrt{\sum_{i=1}^k s_{it}^2}}$, where $[x]_+ = max(x, 0)$. After that, the attended image vector $a_i^t$ with respect to the $t$-th word is given by:

$$a_i^t = \sum_{i=1}^k \alpha_{it} v_i, \quad \alpha_{it} = \frac{exp(\tau \pi_{it})}{\sum_{i=1}^k exp(\tau \pi_{it})}.$$
Figure 2. The pipeline of the proposed framework. During training, the attention weights of captioning module $\beta$ are supervised with the ones of pre-trained matching model $\alpha$ via a local alignment loss (e.g., KL-div) at the visually-groundable words. Additionally, the image-text matching similarity score can serve as a fine-grained reward at the self critical sequence training stage. During testing, the matching model can be fully removed and the captioning model can generate more descriptive and grounded (regions and words are well aligned) captions. Where $h^1_t$ is the hidden state of attention LSTM.

Where $\tau$ is the inverse temperature of the softmax function and $\alpha_{it}$ is the attention weight. Finally, the global similarity score $S(I, T)$ between image $I$ and sentence $T$ is computed by summarizing the local similarity scores $R(e_i, a^v_t)$:

$$S(I, T) = \frac{1}{n} \sum_{i=1}^{n} R(e_i, a^v_t), \quad R(e_i, a^v_t) = \frac{e_i^T a^v_t}{||e_i|| ||a^v_t||}.$$  \hfill (6)

The model is optimized by a triplet loss with hard negative mining [9] in a mini-batch:

$$l_{hard}(I, T) = [m - S(I, T) + S(I, \hat{T}_h)]_+$$

$$+ [m - S(I, T) + S(\hat{I}_h, T)]_+,$$  \hfill (7)

where $m$ is the margin, $\hat{I}_h = \text{argmax}_{p \neq 1} S(p, T)$ and $\hat{T}_h = \text{argmax}_{c \neq T} S(I, c)$.

In the experiment, we find that the original SCAN model even has lower grounding performance than the adopted caption generator. The cause may be the influence of too many non-visual words. So we propose to enhance SCAN model with Part-of-Speech (POS) tags when it serves as an attention guider. We call it POS-SCAN. The Equation (6) is rewritten as:

$$S_{pos}(I, T) = \frac{1}{n} \sum_{i=1}^{n} I_{y_i^* = \text{noun}} R(e_i, a^v_t),$$  \hfill (8)

where $I_{y_i^* = \text{noun}}$ is the indicator function which equals to 1 if the POS of word $y_i^*$ is noun and 0 otherwise. The $S(I, T)$ in Equation (7) is also replaced with $S_{pos}(I, T)$. By doing so, the grounding performance of the POS-SCAN model meets the requirement of the downstream task.

3.2. Caption Generator

For the caption generator, we adopt the state-of-the-art Up-Down [3] model. It is mainly composed of two LSTM [16] layers where the first one is the attention LSTM and the second one is the language LSTM. Each layer is indicated with the corresponding subscript in the equations below. Specifically, it first transforms each region feature $f_i$ as:

$$v'_i = W'_v f_i + b'_v, \quad v'_i \in \mathbb{R}^{d_2}.$$ \hfill (9)

Then at step $t$, the attention LSTM takes previous output of the language LSTM $h_{t-1}^l$, mean-pooled image feature $\mathbf{v} = \frac{1}{k} \sum_i v'_i$ and previous word embedding $e_{t-1} = W'_e y_{t-1}$ as input and output a hidden state $h_t^l$:

$$h_t^l = LSTM_l([h_{t-1}^l; \mathbf{v}; e_{t-1}], h_{t-1}^l),$$ \hfill (10)

where $[:]$ denotes concatenation and $W'_e$ is the word embedding matrix. Given $h_t^l$, the attended image feature is calculated as:

$$\tilde{v}_t = \sum_{i=1}^{k} \beta_{i,t} v'_i, \quad \beta_t = \text{softmax}(z_t),$$ \hfill (11)

$$z_{i,t} = w'_a^T \tanh(W'_w v'_i + W'_h h_{t-1}^l).$$ \hfill (12)

Finally, the language LSTM takes the attended image feature $\tilde{v}_t$ and $h_t^l$ as input and gives the conditional distribution over possible output word as:

$$h_t^2 = LSTM_2([\tilde{v}_t; h_t^1]; h_{t-1}^2),$$ \hfill (13)

$$p(y_t | y_{1:t-1}) = \text{softmax}(W_y h_t^2 + b_y),$$ \hfill (14)

where $W_y$ and $b_y$ are learned weights and biases. $y_{1:t-1}$ refers to $(y_1, \cdots, y_{t-1})$.

3.3. Learning to Generate More Grounded Captions

The SCAN model and POS-SCAN are first pre-trained on image-caption dataset and remain fixed. They serve as
the attention guider and fine-grained reworder during the SCST [46] fine-tuning of the caption generator. The training process is divided into two stages.

In the first stage, given the target ground truth sentence $(y_1, \cdots, y_n)$, the captioning model with parameters $\theta$ is usually trained by minimizing standard cross-entropy loss. However, its attention module is not forced to correctly associate the generated words with the attended regions. To generate more grounded captions without region-word alignment annotations, we additionally regularize the attention weights $\beta_t$ of captioning model with attention weights $\alpha_t$ distilled from POS-SCAN model via KL-divergence. The combined loss function is as follows:

$$l_1(\theta) = \sum_{i=1}^{n} (-\log(p_{\theta}(y_1^i|y_1^{i-1}))$$

$$+ \lambda_1 \mathbb{1}_{y_1^i=y_{\text{noun}}} KL(\beta_t||\alpha_t).$$

(15)

If ground truth region-word alignment annotations are available, the combined loss function can be written as follows:

$$l'_1(\theta) = \sum_{i=1}^{n} (-\log(p_{\theta}(y_1^i|y_1^{i-1}))$$

$$+ \lambda'_1 \mathbb{1}_{y_1^i=y_{\text{noun}}} \sum_{i=1}^{k} -\gamma_{t_i} \log \beta_{t_i}),$$

(16)

where $\gamma_t = [\gamma_{t1}, \cdots, \gamma_{tk}]$ is the indicators of positive/negative regions and $\gamma_{t_i} = 1$ when the $i$-th region has over 0.5 IoU with the ground truth box and otherwise 0. The second term of $l'_1(\theta)$ can also be KL-divergence and negative log likelihood loss.

The second stage, the captioning model is further trained by REINFORCE algorithm. Specifically, it seeks to minimize the negative expected reward $r$:

$$l_2(\theta) = -E_{y_1^{n-1} \sim \rho_0}[r(y_1^{n-1})].$$

(17)

Following the approach described in self-critical sequence training (SCST) [46], the gradient of this loss can be approximated as:

$$\nabla_{\theta} l_2(\theta) \approx -(r(y_1^{n-1}) - r(\hat{y}_1^{n-1})) \nabla_{\theta} \log p_{\theta}(y_1^{n-1}),$$

(18)

where $y_1^{n-1}$ is a sampled caption and $r(\hat{y}_1^{n-1})$ defines the baseline reward obtained by greedily decoding the current model. Compared to [46, 37, 33], the main difference lies in the definition of the reward function $r$ and the goal. In [46], only language metric CIDEr [52] is used as the reward function. In [37, 33], a weight sum of CIDEr score and global image-text matching similarity score is used as the reward function for discriminative captions. To make full use of the local image-text matching model, we further treat the fine-grained local image-text matching score $S(I, T)$ as a reward. Our final reward function is the combination:

$$r(y_1^{n-1}) = \text{CIDEr}(y_1^{n-1}) + \lambda_2 S(I, y_1^{n-1}),$$

(19)

which has the potential to encourage captioning model to generate more grounded captions.

### 4. Experiments

#### 4.1. Datasets and Evaluation Metrics

Since the main goal of our experiments is to evaluate the effectiveness of the proposed weakly-supervised method in improving the grounding and performance of the captioning model, it’s convenient to use the Flickr30k Entities dataset [44]. The dataset contains 275k bounding boxes from 31k images associated with natural language phrases. Each image is annotated with 5 crowdsourced captions. Following [36], phrase labels for boxes are converted to a single-word object labels. We used splits from Karpathy et al. [19], which includes 29k images for training, 1k images for validation, and another 1k for test. We also reported part results on MS-COCO dataset [27].

To evaluate the caption quality, we used the standard evaluation script,

2

which reports the widely used auto-

training (SCST) [46]. Following the approach described in self-critical sequence
to minimize the negative expected reward

trained by REINFORCE algorithm. Specifically, it seeks
negative log likelihood loss.

tion weights

alignment annotations, we additionally regularize the atten-
tion weights $\beta_t$ of captioning model with attention weights $\alpha_t$ distilled from POS-SCAN model via KL-divergence.

To generate more grounded captions without region-word alignment annotations, we additionally regularize the attention weights $\beta_t$ of captioning model with attention weights $\alpha_t$ distilled from POS-SCAN model via KL-divergence. The combined loss function is as follows:

$$l_1(\theta) = \sum_{i=1}^{n} (-\log(p_{\theta}(y_1^i|y_1^{i-1}))$$

$$+ \lambda_1 \mathbb{1}_{y_1^i=y_{\text{noun}}} KL(\beta_t||\alpha_t).$$

(15)

If ground truth region-word alignment annotations are available, the combined loss function can be written as follows:

$$l'_1(\theta) = \sum_{i=1}^{n} (-\log(p_{\theta}(y_1^i|y_1^{i-1}))$$

$$+ \lambda'_1 \mathbb{1}_{y_1^i=y_{\text{noun}}} \sum_{i=1}^{k} -\gamma_{t_i} \log \beta_{t_i}),$$

(16)

where $\gamma_t = [\gamma_{t1}, \cdots, \gamma_{tk}]$ is the indicators of positive/negative regions and $\gamma_{t_i} = 1$ when the $i$-th region has over 0.5 IoU with the ground truth box and otherwise 0. The second term of $l'_1(\theta)$ can also be KL-divergence and negative log likelihood loss.

In the second stage, the captioning model is further trained by REINFORCE algorithm. Specifically, it seeks to minimize the negative expected reward $r$:

$$l_2(\theta) = -E_{y_1^{n-1} \sim \rho_0}[r(y_1^{n-1})].$$

(17)

Following the approach described in self-critical sequence training (SCST) [46], the gradient of this loss can be approximated as:

$$\nabla_{\theta} l_2(\theta) \approx -(r(y_1^{n-1}) - r(\hat{y}_1^{n-1})) \nabla_{\theta} \log p_{\theta}(y_1^{n-1}),$$

(18)

where $y_1^{n-1}$ is a sampled caption and $r(\hat{y}_1^{n-1})$ defines the baseline reward obtained by greedily decoding the current model. Compared to [46, 37, 33], the main difference lies in the definition of the reward function $r$ and the goal. In [46], only language metric CIDEr [52] is used as the reward function. In [37, 33], a weight sum of CIDEr score and global image-text matching similarity score is used as the reward function for discriminative captions. To make full use of the local image-text matching model, we further treat the fine-grained local image-text matching score $S(I, T)$ as a reward. Our final reward function is the combination:

$$r(y_1^{n-1}) = \text{CIDEr}(y_1^{n-1}) + \lambda_2 S(I, y_1^{n-1}),$$

(19)

which has the potential to encourage captioning model to generate more grounded captions.

### 4. Experiments

#### 4.1. Datasets and Evaluation Metrics

Since the main goal of our experiments is to evaluate the effectiveness of the proposed weakly-supervised method in improving the grounding and performance of the captioning model, it’s convenient to use the Flickr30k Entities dataset [44]. The dataset contains 275k bounding boxes from 31k images associated with natural language phrases. Each image is annotated with 5 crowdsourced captions. Following [36], phrase labels for boxes are converted to a single-word object labels. We used splits from Karpathy et al. [19], which includes 29k images for training, 1k images for validation, and another 1k for test. We also reported part results on MS-COCO dataset [27].

To evaluate the caption quality, we used the standard evaluation script, which reports the widely used automatic evaluation metrics, BLEU [43], METEOR [8] and CIDEr [52] and SPICE [2].

To evaluate region-word alignment quality, we followed the metrics defined in [65]. It can compute alignment quality on both ground truth and generated sentences. In the first case, we fed the ground truth sentence into the model and compared the region with the highest attention weight against the ground truth box at each annotated object word. An object word is correctly localized if the Intersection-over-Union (IoU) is over 0.5. In the second case, $F_{1\text{all}}$ and $F_{1\text{loc}}$ metrics are computed after performing standard language generation inference. In $F_{1\text{all}}$, a region prediction is considered correct if the object word is correctly predicted and also correctly localized. In $F_{1\text{loc}}$, only correctly-predicted object words are considered. For more details, please refer to the appendix in [65].

#### 4.2. Implementation Details

We mainly adopted the widely used Faster R-CNN [45] model pre-trained by Anderson et al. [3] on Visual Genomes [22] as image feature extractor. For each image, we extracted 36 regions which are represented as a sequence of feature vectors with 2,048 dimensions and bounding box coordinates with 4 dimensions. To make a fair comparison with a recent similar work [38], we additionally conducted experiments using visual features extracted by Zhou et al. [65]. If no special instruction, we used the former image features.

For the local image-text matching model, the word embedding size was set to 300, the GRU hidden state size and joint embedding size $d_j$ were set to 1,024. The margin $m$
Table 1. Attention accuracy on Flickr30k Entities val set. It is measured on annotated object words of ground truth sentences. * indicates such results are our remeasurement. +XE denotes cross entropy loss. NLL denotes negative log likelihood and KL denotes KL divergence. GT denotes grounding supervision comes from the ground truth. 0.1 is the balance weight.

<table>
<thead>
<tr>
<th>Model</th>
<th>Attention Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCAN</td>
<td>17.63%</td>
</tr>
<tr>
<td>Up-Down+XE[3]</td>
<td>19.83%</td>
</tr>
<tr>
<td>POS-SCAN</td>
<td>28.58%</td>
</tr>
<tr>
<td>Up-Down+XE+0.1NLL(GT)</td>
<td>37.17%</td>
</tr>
<tr>
<td>Up-Down+XE+0.1KL(POS-SCAN)</td>
<td>29.39%</td>
</tr>
</tbody>
</table>

Figure 3. The effect of the $\lambda_1$ on the Flickr30k entities val set. From the Figure, we can observe that both the captioning evaluation (e.g. CIDEr and SPICE) and attention evaluation (e.g. $F_{1,att}$ and $F_{1,loc}$) of the captioning model can be improved when appropriate region-word alignments supervision is enforced.

4.3. Quantitative Analysis

We will validate the effectiveness of the proposed method by answering five questions as follows.

Q1: Does the image-text matching model has higher region-word alignment accuracy than image captioning model? Our method is based on the intuition that the region-word alignments of the image-text matching model should be more reliable than the ones of the image captioning model. We validated it by feeding the ground truth sentences on validation set into the model and computing the attention accuracy, with results reported in Table 1. To our surprise, the original SCAN model even has lower attention accuracy 17.63% than the adopted caption generator Up-Down 19.83%. The cause may be the influence of too many non-visual words. We remedied this by resorting to POS to remove non-visual words when computing the matching score at the cost of image-text matching accuracy. After this, the attention accuracy of POS-SCAN model 28.58% meets the requirements of the downstream task.

Q2: Can we improve the grounding performance of the captioning model by distilling the image-text matching model? Although POS-SCAN has higher attention accuracy than Up-Down model, it is not clear to what extent can POS-SCAN transfer the grounding ability to Up-Down model. To check this, we trained four Up-Down models, which respectively corresponds to without attention supervision, with ground truth attention supervision (upper bound) and weakly supervision distilled from SCAN and POS-SCAN model in the XE Pre-Train stage. The effect of $\lambda_1$ on caption evaluation and attention supervision is shown in Figure 3. In the following experiment, we set $\lambda_1 = 0.1$ if not otherwise specified. By comparing the 1st row in each section of Table 2, we can observe that the model with POS-SCAN supervision significantly improves the attention evaluation performance without any region-word alignment annotations, while the model with original SCAN supervision can’t achieve this as expected.

Q3: Can the captioning model maintain the grounding performance after self-critical sequence training (SCST)? It is well known that SCST [46] is an effective training strategy to improve caption quality in practice. However, how the grounding performance (attention accuracy, with slightly abused) of captioning model changes remains unknown. To uncover this, captioning models were further optimized by SCST with CIDEr as reward. By comparing the 1st and 2nd row in each section of Table 2, we find that the caption quality is significantly improved while the grounding performance is degrading in most cases. The reason is that CIDEr metric encourages the n-gram consistency but not the visual semantic alignment, leading to the conflicting grounding and captioning performances.

Q4: Is it useful to incorporate the fine-grained image-text similarity score as reward? By comparing the 2nd
and 3rd row in each section of Table 2, we can find that by further incorporating the SCST as reward function, models obtain consistently improvement on the SPICE metric, which captures more semantic propositional content compared with other conventional metrics. Moreover, we find that such reward can improve the grounding performance in most cases when compared to using only CIDEr as reward. By further comparing the 3rd and 4th row in each section of Table 2, we can find that SCAN reward function is a good trade-off between the caption quality and the grounding performance when compared to POS-SCAN reward function.

Q5: How does our final model perform compared to other state-of-the-art models? We compared our final model with other state-of-the-art models on the test set, as shown in Table 3. For a fair comparison with the most similar work [38], we also run our final model using their visual feature (with $\lambda_1 = 0.2$). Our model achieves better performance on both caption evaluation and attention evaluation without any ground truth attention supervision. We also report part results on MS-COCO in Table 4.
4.4. Qualitative Result

To illustrate the advantages of our proposed method, we present some qualitative examples in Figure 4. We can observe that our proposed method can help to generate more grounded captions (e.g., it aligns the “men” to the correct region in the 2nd image). We also present some representative failure cases of the neural-based captioning model in Figure 5. Errors include pattern repetition (e.g., the 1st image), mis-recognition (e.g., the 2nd and 3rd image) and misassociation because of complex context (e.g., the 4th image).

5. Conclusions

In this work, we demonstrated that it is feasible to generate more grounded captions without grounding annotations by distilling the image-text matching model: the proposed POS-SCAN. This enhances the interpretability and transparency of existing captioning models. Additionally, by incorporating the SCAN image-text matching score as the reward, we found a practical trade-off between the caption quality and the grounding performance. In the future, it may be an interesting direction to design a learnable image-text matching metric — other than the problematic n-gram based metrics — to encourage more grounded image captioning for better model explainability.

Acknowledgments We thank all the reviewers for their constructive comments. The research is supported by National Nature Science Foundation of China under grants 61732008 and 61725203. This research was also partly supported by Major Scientific Research Project of Zhejiang Lab (No.2019DB0ZX01).
References


[54] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In CVPR, 2016.


