More Grounded Image Captioning by Distilling Image-Text Matching Model

Yuanen Zhou[1,2], Meng Wang[1,2,*], Daqing Liu[3], Zhenzhen Hu[1,2], Hanwang Zhang[4]
[1]Key Laboratory of Knowledge Engineering with Big Data (Ministry of Education)
[2]School of Computer Science and Information Engineering, Hefei University of Technology
[3] University of Science and Technology of China [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 [1].

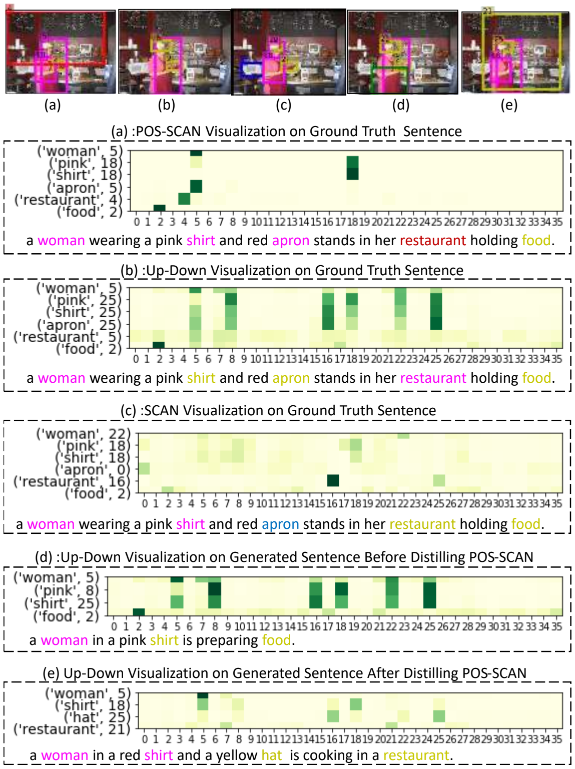

Figure 1. Visualizations of five different word-region alignment results, where all the models are trained without any word-region alignment ground-truth. Words and the corresponding attended region with maximum weight are marked with the same color. POS-SCAN (cf. Section 3.1) is a revised image-text matching model, Up-Down (cf. Section 3.2) is a state-of-the-art image captioning model. Best viewed in color.

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

[*] Corresponding Author.
[1] https://github.com/YuanEZhou/Grounded-Image-Captioning

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 sentence-level 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 metadata [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-caption 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 $[\mathbf{f}_1, \cdots, \mathbf{f}_k] \in \mathbb{R}^{k \times d}$ extracted by a detector [45]. The corresponding ground truth and generated sentence T with n words are represented as $(y_1^*, \cdots, y_n^*)$ and $(y_1, \cdots, y_n)$, 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 $\mathbf{f_i}$ to appropriate dimension by:

$$\mathbf{v}_i = \mathbf{W}_v \mathbf{f}_i + \mathbf{b}_v, \quad \mathbf{v}_i \in \mathbb{R}^{d_1}, \tag{1}$$

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

$$\mathbf{x}_t = \mathbf{W}_e y_t^*, \quad \overrightarrow{\mathbf{h}_t} = \overrightarrow{GRU}(\mathbf{x}_t), \quad \overleftarrow{\mathbf{h}_t} = \overleftarrow{GRU}(\mathbf{x}_t), \tag{2}$$

where $\mathbf{W}_e$ is an embedding matrix. The final word feature $\mathbf{e}_t$ is the average of the forward hidden state $\overrightarrow{\mathbf{h}_t}$ and backward hidden state $\overleftarrow{\mathbf{h}_t}$:

$$\mathbf{e}_t = \frac{(\overrightarrow{\mathbf{h}_t} + \overleftarrow{\mathbf{h}_t})}{2}, \quad t \in [1, n]. \tag{3}$$

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

$$s_{it} = \frac{\mathbf{v}_i^T \mathbf{e}_t}{\|\mathbf{v}_i\| \, \|\mathbf{e}_t\|}, i \in [1, k], t \in [1, n]. \tag{4}$$

Here, s_{it} denotes the similarity between the i-th region and the t-th word is normalized as $\bar{s}_{it} = [s_{it}] / \sqrt{\sum_{t=1}^{n} [s_{it}]_+^2}$, where $[x]_+ \equiv max(x, 0)$. After that, the attended image vector $\mathbf{a}_t^v$ with respect to the t-th word is given by:

$$\mathbf{a}_t^v = \sum_{i=1}^{k} \alpha_{it} \mathbf{v}_i, \quad \alpha_{it} = \frac{exp(\tau \bar{s}_{it})}{\sum_{i=1}^{k} exp(\tau \bar{s}_{it})}. \tag{5}$$

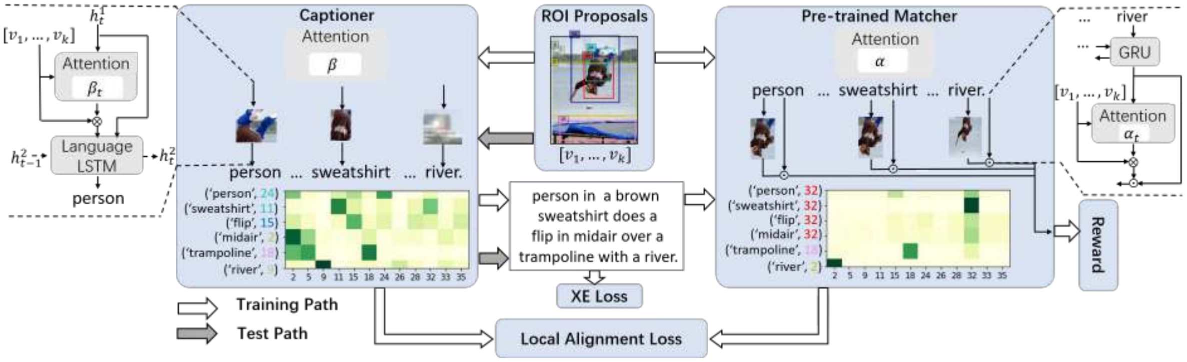

Figure 2. The pipeline of the proposed framework. During training, the attention weights of captioning module $\boldsymbol{\beta}$ are supervised with the ones of pre-trained matching model $\boldsymbol{\alpha}$ via a local alignment loss (*e.g.* KL-div) at the visually-groundable words. Additionally, 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_t^1 is the hidden state of attention LSTM.

Where τ is the inverse temperature of the softmax function and α_{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 $R(\mathbf{e}_t, \mathbf{a}_t^v)$:

$$S(I,T) = \frac{\sum_{t=1}^{n} R(\mathbf{e}_t, \mathbf{a}_t^v)}{n}, \; R(\mathbf{e}_t, \mathbf{a}_t^v) = \frac{\mathbf{e}_t^T \mathbf{a}_t^v}{\|\mathbf{e}_t\| \, \|\mathbf{a}_t^v\|}. \tag{6}$$

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

$$\begin{aligned} l_{hard}(I,T) &= [m - S(I,T) + S(I,\hat{T}_h)]_+ \\ &\quad + [m - S(I,T) + S(\hat{I}_h,T)]_+, \end{aligned} \tag{7}$$

where *m* is the margin, $\hat{I}_h = argmax_{p \neq I} S(p,T)$ and $\hat{T}_h = argmax_{c \neq T} S(I,c)$.

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:

$$S_{pos}(I,T) = \frac{\sum_{t=1}^{n} \mathbb{1}_{y_t^* = y^{noun}} R(\mathbf{e}_t, \mathbf{a}_t^v)}{\sum_{t=1}^{n} \mathbb{1}_{y_j^* = y^{noun}}}, \tag{8}$$

where $\mathbb{1}_{y_t^* = y^{noun}}$ is the indicator function which equals to 1 if the POS of word y_t^* 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 $\mathbf{f}_i$ as:

$$\mathbf{v}_i^{'} = \mathbf{W}_v^{'} \mathbf{f}_i + \mathbf{b}_v^{'}, \quad \mathbf{v}_i^{'} \in \mathbb{R}^{d_2}. \tag{9}$$

Then at time step t, the attention LSTM takes previous output of the language LSTM $\mathbf{h}_{t-1}^2$, mean-pooled image feature $\overline{\mathbf{v}} = \frac{1}{k} \sum_i \mathbf{v}_i^{'}$ and previous word embedding $\mathbf{e}_{t-1}^{'} = \mathbf{W}_e^{'} y_{t-1}$ as input and output a hidden state $\mathbf{h}_t^1$:

$$\mathbf{h}_t^1 = LSTM_1([\mathbf{h}_{t-1}^2; \overline{\mathbf{v}}; \mathbf{e}_{t-1}^{'}], \mathbf{h}_{t-1}^1), \tag{10}$$

where $[;]$ denotes concatenation and $\mathbf{W}_e^{'}$ is the word embedding matrix. Given $\mathbf{h}_t^1$, the attended image feature is calculated as:

$$\hat{\mathbf{v}}_t = \sum_{i=1}^{k} \beta_{i,t} \mathbf{v}_i^{'}, \quad \boldsymbol{\beta}_t = softmax(\mathbf{z}_t), \tag{11}$$

$$z_{i,t} = \mathbf{w}_a^T tanh(\mathbf{W}_{va} \mathbf{v}_i^{'} + \mathbf{W}_{ha} \mathbf{h}_t^1). \tag{12}$$

Finally, the language LSTM takes the attended image feature $\hat{\mathbf{v}}_t$ and $\mathbf{h}_t^1$ as input and gives the conditional distribution over possible output word as:

$$\mathbf{h}_t^2 = LSTM_2([\hat{\mathbf{v}}_t; \mathbf{h}_t^1], \mathbf{h}_{t-1}^2), \tag{13}$$

$$p(y_t | y_{1:t-1}) = softmax(\mathbf{W}_o \mathbf{h}_t^2 + \mathbf{b}_o), \tag{14}$$

where $\mathbf{W}_o$ and $\mathbf{b}_o$ 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 rewarder 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 $\boldsymbol{\theta}$ 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 $\boldsymbol{\beta}_t$ of captioning model with attention weights $\boldsymbol{\alpha}_t$ distilled from POS-SCAN model via KL-divergence. The combined loss function is as follows:

$$\begin{aligned} l_1(\boldsymbol{\theta}) = &\sum_{t=1}^{n}\{-\log(p_{\boldsymbol{\theta}}(y_t^*|y_{1:t-1}^*)) \\ &+ \lambda_1 \mathbb{1}_{y_t^*=y^{noun}} KL(\boldsymbol{\beta}_t \| \boldsymbol{\alpha}_t)\}. \end{aligned} \quad (15)$$

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

$$\begin{aligned} l_1^{'}(\boldsymbol{\theta}) = &\sum_{t=1}^{n}\{-\log(p_{\boldsymbol{\theta}}(y_t^*|y_{1:t-1}^*)) \\ &+ \lambda_1^{'} \mathbb{1}_{y_t^*=y^{noun}} \sum_{i=1}^{k} -\gamma_{ti} \log \beta_{ti}\}, \end{aligned} \quad (16)$$

where $\boldsymbol{\gamma}_t = [\gamma_{t1}, \cdots, \gamma_{tk}]$ is the indicators of positive/negative regions and $\gamma_{ti} = 1$ when the i-th region has over 0.5 IoU with the ground truth box and otherwise 0. The second term of $l_1^{'}(\boldsymbol{\theta})$ 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:

$$l_2(\boldsymbol{\theta}) = -E_{y_{1:n} \sim p_{\boldsymbol{\theta}}}[r(y_{1:n})]. \quad (17)$$

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

$$\nabla_{\boldsymbol{\theta}} l_2(\boldsymbol{\theta}) \approx -(r(y_{1:n}^s) - r(\hat{y}_{1:n}))\nabla_{\boldsymbol{\theta}} log\, p_{\boldsymbol{\theta}}(y_{1:n}^s), \quad (18)$$

where $y_{1:n}^s$ is a sampled caption and $r(\hat{y}_{1:n})$ 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:

$$r(y_{1:n}) = CIDEr(y_{1:n}) + \lambda_2 S(I, y_{1:n}), \quad (19)$$

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 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 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, $F1_{all}$ and $F1_{loc}$ metrics are computed after performing standard language generation inference. In $F1_{all}$, a region prediction is considered correct if the object word is correctly predicated and also correctly localized. In $F1_{loc}$, only correctly-predicated 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_1 were set to $1,024$. The margin m

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

Model	Attention Acc.
SCAN*[24]	17.63%
Up-Down+XE*[3]	19.83%
POS-SCAN	28.58%
Up-Down+XE+0.1NLL(GT)	37.17%
Up-Down+XE+0.1KL(POS-SCAN)	29.39%

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.

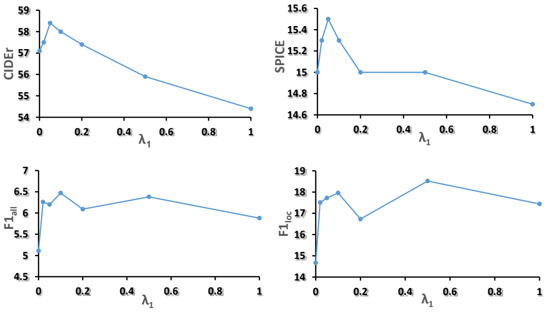

Figure 3. The effect of the λ_1 on the Flickr30k entities val set. From the Figure, we can observe that both the captioning evaluation (*e.g.* CIDEr and SPICE) and attention evaluation (*e.g.* F1$_{all}$ and F1$_{loc}$) of the captioning model can be improved when appropriate region-word alignments supervision is enforced.

and temperature τ were respectively set to 0.2 and 9. Following the training strategy in [24], we retrained both the SCAN and POS-SCAN model.

For the captioning model, we conducted experiments based on the widely used open-source codebase[3]. The word embedding size was set to 512. The image feature embedding size d_2 and LSTM hidden state size were all set to 512 (1,024 for MS-COCO). We built a dictionary by dropping the words that occur less than 5 times and end up with a vocabulary of 7,000 (9,487 for MS-COCO). We truncated captions longer than 16 words. We optimized our model with Adam [20] for 30 epochs in the first training stage. The learning rate was initialized to be $5e$-4 and decayed by a factor 0.8 every three epochs. In the second stage, we continued to train the model for another 80 epochs with an initial learning rate of $5e$-5. During inference, we disabled the beam search for the convenience of region-word alignment evaluation on Flickr30k Entities and set it to 3 on MS-COCO.

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 λ_1 on caption evaluation and attention evaluation 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

[3] https://github.com/ruotianluo/self-critical.pytorch

XE Pre-Train			SCST Fine-Tune			Caption Eval.					Attention Eval.	
GT	SCAN	POS-SCAN	CIDEr	SCAN	POS-SCAN	B@1	B@4	M	C	S	F1$_{all}$	F1$_{loc}$
Using Ground Truth Attention Supervision												
✓	✗	✗	✗	✗	✗	70.1	27.4	21.8	58.9	15.4	8.33	**23.09**
✓	✗	✗	✓	✗	✗	**73.4**	**29.6**	22.4	**67.5**	16.0	7.53	18.40
✓	✗	✗	✓	✓	✗	72.3	28.5	**22.6**	67.0	**16.5**	**8.35**	20.75
✓	✗	✗	✓	✗	✓	72.3	27.6	22.4	64.4	16.1	8.01	19.48
No Attention Supervision												
✗	✗	✗	✗	✗	✗	69.6	26.9	21.6	57.1	15.0	5.11	**14.67**
✗	✗	✗	✓	✗	✗	**73.1**	**29.1**	22.2	67.1	15.9	4.19	10.71
✗	✗	✗	✓	✓	✗	**73.1**	28.8	22.3	**67.5**	16.1	4.59	12.81
✗	✗	✗	✓	✗	✓	72.1	27.7	**22.5**	64.9	**16.3**	**5.37**	13.88
Attention Supervision Distilled from SCAN												
✗	✓	✗	✗	✗	✗	70.0	27.7	22.0	58.8	15.5	4.49	13.49
✗	✓	✗	✓	✗	✗	73.2	**29.3**	**22.5**	67.4	16.0	4.72	13.47
✗	✓	✗	✓	✓	✗	73.2	28.6	22.4	**67.8**	**16.3**	4.77	12.25
✗	✓	✗	✓	✗	✓	**73.3**	28.4	**22.5**	67.5	16.1	**5.34**	**14.79**
Attention Supervision Distilled from POS-SCAN												
✗	✗	✓	✗	✗	✗	70.4	27.5	21.8	58.0	15.3	6.47	17.96
✗	✗	✓	✓	✗	✗	73.7	**29.9**	22.3	67.5	16.0	6.62	16.97
✗	✗	✓	✓	✓	✗	**73.9**	29.4	**22.8**	**68.2**	**16.7**	7.30	**18.44**
✗	✗	✓	✓	✗	✓	72.6	28.0	22.6	64.3	16.0	**7.63**	18.33

Table 2. Ablation studies on the Flickr30k Entities val set. The baseline captioning model is Up-Down [3]. XE denotes cross entropy. In the XE Pre-Train stage: GT denotes using ground truth attention supervision; SCAN (POS-SCAN) denotes attention supervision distilled from SCAN (POS-SCAN). In the SCST [46] fine-tune stage: CIDEr denotes using CIDEr as reward function; SCAN (POS-SCAN) denotes using the image-text matching score of SCAN (POS-SCAN) model as reward.

	Caption Evaluation					Att. Eval.	
	B@1	B@4	M	C	S	F1$_{all}$	F1$_{loc}$
SR-PL[33]	72.9	29.3	21.8	65.0	15.8	-	-
Gu *et al.* [11]	**73.8**	**30.7**	21.6	61.8	15.0	-	-
NBT[36]	69.0	27.1	21.7	57.5	15.6	-	-
Unsup.† [65]	69.2	26.9	22.1	60.1	16.1	3.88	11.7
GVD(Sup.)†[65]	69.9	27.3	22.5	62.3	16.5	7.55	22.2
Cyclical† [38]	68.9	26.6	22.3	60.9	16.3	4.85	13.4
Ours†	71.4	28.0	**22.6**	66.2	**17.0**	6.53	15.79
Ours‡	73.4	30.1	**22.6**	**69.3**	16.8	**7.17**	**17.49**

Table 3. Performance comparison with the state-of-the-art methods on the Flickr30k Entities test set. † denotes using visual feature from [65] and ‡ denotes using the widely adopted bottom-up visual feature from [3]. Sup. denotes model trained with ground truth grounding annotations. The supervised method is used as upper bound and its numbers are not bolded.

Up-Down	B@1	B@4	M	C	S
XE Pre-Train*[3]	77.2	36.2	27.0	113.5	20.3
+SCST(CIDEr)*[3]	79.8	36.3	27.7	120.1	21.4
XE Pre-Train	76.2	36.4	27.7	113.1	20.5
+SCST(CIDEr)	80.0	37.8	28.1	125.2	21.6
XE Pre-Train+POS-SCAN	76.6	36.5	27.9	114.9	20.8
+SCST(CIDEr)	80.1	37.8	28.3	125.9	22.0
+SCST(SCAN)	**80.2**	**38.0**	**28.5**	**126.1**	**22.2**

Table 4. Performance on the MS-COCO Karpathy test set. $*$ denotes results reported in the original paper. Omitted balance weights equal to 1. SCST(x) means using x as reward function in SCST [46] fine-tune stage.

and 3rd row in each section of Table 2, we can find that by further incorporating the SCAN 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.

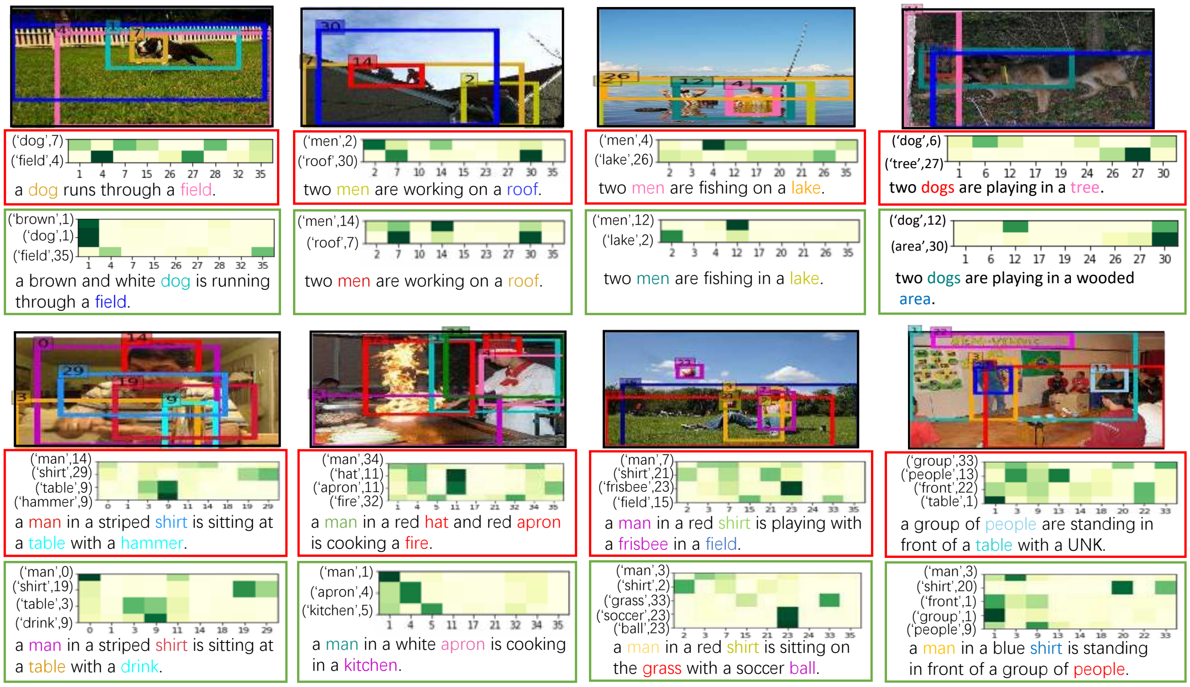

Figure 4. Generated captions and internal region-word alignments of models without and with POS-SCAN attention supervision in the XE Pre-Train stage. In each unit, caption surrounded by red box is from the former and green one is from the latter. Word and corresponding attended region with maximum weight are marked with the same color. We also visualize the attention weight distributions of some visually-groundable words on top of captions. Darker color indicates bigger weights. For space reasons, we only show a part of regions.

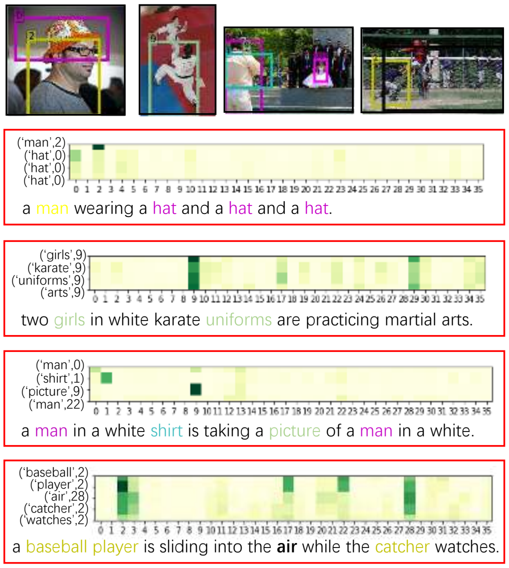

Figure 5. Some representative failure cases generated by the captioning model.

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 mis-association 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

[1] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Emotion recognition in speech using crossmodal transfer in the wild. In *ACM MM*, 2018. 3

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016. 5

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 2, 4, 5, 6, 7

[4] Kan Chen, Jiyang Gao, and Ram Nevatia. Knowledge aided consistency for weakly supervised phrase grounding. In *CVPR*, 2018. 3

[5] Kan Chen, Rama Kovvuri, and Ram Nevatia. Query-guided regression network with context policy for phrase grounding. In *ICCV*, 2017. 3

[6] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, 2017. 2

[7] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. *arXiv preprint arXiv:1903.11649*, 2019. 3

[8] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 5

[9] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improved visual-semantic embeddings. In *BMVC*, 2018. 2, 3, 4

[10] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013. 3

[11] Jiuxiang Gu, Gang Wang, Jianfei Cai, and Tsuhan Chen. An empirical study of language cnn for image captioning. In *ICCV*, 2017. 7

[12] Longteng Guo, Jing Liu, Jinhui Tang, Jiangwei Li, Wei Luo, and Hanqing Lu. Aligning linguistic words and visual semantic units for image captioning. In *ACM MM*, 2019. 2

[13] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *CVPR*, 2016. 3

[14] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018. 2

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 3

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4

[17] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *CVPR*, 2017. 3

[18] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *ICCV*, 2019. 2

[19] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 2, 3, 5

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[21] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 3

[22] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 5

[23] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013. 1, 2

[24] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, 2018. 1, 2, 3, 6

[25] Seung Hyun Lee, Dae Ha Kim, and Byung Cheol Song. Self-supervised knowledge distillation using singular value decomposition. In *ECCV*, 2018. 3

[26] Siming Li, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228. Association for Computational Linguistics, 2011. 2

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[28] Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. Attention correctness in neural image captioning. In *AAAI*, 2017. 1, 2, 3

[29] Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yongdong Zhang, and Feng Wu. Context-aware visual policy network for sequence-level image captioning. In *ACM MM*, 2018. 2

[30] Daqing Liu, Hanwang Zhang, Zheng-Jun Zha, and Wu Feng. Learning to assemble neural module tree networks for visual grounding. In *ICCV*, 2019. 3

[31] Daqing Liu, Hanwang Zhang, Zheng-Jun Zha, and Fanglin Wang. Referring expression grounding by marginalizing scene graph likelihood. *arXiv preprint arXiv:1906.03561*, 2019. 3

[32] Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. Referring expression generation and comprehension via attributes. In *ICCV*, 2017. 3

[33] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *ECCV*, 2018. 3, 5, 7

[34] Yongcheng Liu, Lu Sheng, Jing Shao, Junjie Yan, Shiming Xiang, and Chunhong Pan. Multi-label image classification via knowledge distillation from weakly-supervised detection. In *ACM MM*, 2018. 2, 3

[35] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017. 2

[36] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *CVPR*, 2018. 2, 3, 5, 7

[37] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. In *CVPR*, 2018. 3, 5

[38] Chih-Yao Ma, Yannis Kalantidis, Ghassan AlRegib, Peter Vajda, Marcus Rohrbach, and Zsolt Kira. Learning to generate grounded image captions without localization supervision. *arXiv preprint arXiv:1906.00283*, 2019. 1, 2, 3, 5, 7

[39] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 3

[40] Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. Midge: Generating image descriptions from computer vision detections. In *EACL*, 2012. 2

[41] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *NeurIPS*, 2019. 2, 3

[42] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Hierarchical multimodal lstm for dense visual-semantic embedding. In *ICCV*, 2017. 3

[43] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 5

[44] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 2, 5

[45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 3, 5

[46] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, July 2017. 1, 2, 5, 6, 7

[47] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *EMNLP*, 2018. 2

[48] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016. 2, 3

[49] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 3

[50] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997. 3

[51] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. *arXiv preprint arXiv:2002.11949*, 2020. 2

[52] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 2, 5

[53] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 1, 2

[54] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016. 3

[55] Tan Wang, Xing Xu, Yang Yang, Alan Hanjalic, Heng Tao Shen, and Jingkuan Song. Matching images and text with multi-modal tensor fusion and re-ranking. In *ACM MM*, 2019. 3

[56] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 1, 2

[57] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, 2019. 2

[58] Xu Yang, Hanwang Zhang, and Jianfei Cai. Learning to collocate neural modules for image captioning. In *ICCV*, 2019. 2

[59] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *arXiv preprint arXiv:2003.03923*, 2020. 2

[60] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018. 2

[61] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, 2016. 2

[62] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, 2018. 3

[63] Mingkuan Yuan and Yuxin Peng. Ckd: Cross-task knowledge distillation for text-to-image synthesis. *IEEE TMM*, 2019. 2, 3

[64] Zheng-Jun Zha, Daqing Liu, Hanwang Zhang, Yongdong Zhang, and Feng Wu. Context-aware visual policy network for fine-grained image captioning. *IEEE TPAMI*, 2019. 2

[65] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *CVPR*, 2019. 2, 3, 5, 7