Self-Annotated Training for Controllable Image Captioning

Zhangzi Zhu, Tianlei Wang, and Hong Qu
University of Electronic Science and Technology of China
202021080414@std.uestc.edu.cn

Abstract
The Controllable Image Captioning (CIC) task aims to generate captions conditioned on designated control signals. Several structure-related control signals are proposed to control the semantic structure of sentences, such as sentence length and Part-of-Speech tag sequences. However, due to the fact that the accuracy-based reward focuses mainly on contents rather than semantic structures, existing reinforcement training methods are not applicable to structure-related CIC models. The lack of reinforcement training leads to exposure bias and the inconsistency between the optimizing function and evaluation metrics. In this paper, we propose a novel reinforcement training method for structure-related control signals: Self-Annotated Training (SAT), to improve both the accuracy and controllability of CIC models. In SAT, a recursive annotation mechanism (RAM) is designed to force the input control signal to match the actual output sentence. Moreover, we propose an extra alignment reward to finetune the CIC model trained after SAT method, which further enhances the controllability of models. On the MSCOCO benchmark, we conduct extensive experiments on different structure-related control signals and on different baseline models, the results of which demonstrate the effectiveness and generalizability of our methods.

1 Introduction
Image captioning, which belongs to the intersection of computer vision and natural language processing, is an important part of applying artificial intelligence to many life scenes. The generated texts can be used for the image search task and visually impaired people assistance. Captioning generation is a particularly challenging task which requires models not only to recognize salient objects, attributes and relationships in an image, but also to describe various information through fluent natural language. Thanks to the proposal of encoder-decoder framework (Vinyals et al. 2015) and attention mechanism (Xu et al. 2015), current captioning models (Pan et al. 2020; Luo et al. 2021; Ji et al. 2020; Song et al. 2020) have already outperformed humans in several accuracy-based evaluation metrics.

In recent years, many efforts (Chen et al. 2021; Cornia, Baraldi, and Cucchiara 2020; Chen et al. 2020; Deng et al. 2020) have been made to endow captioning models with human-like controllability, called Controllable Image Captioning (CIC). They introduce various control signals into CIC. As classified in (Chen et al. 2021), control signals can be roughly divided into two categories: 1) Content-related: the control signal is related to the contents of images, such as guiding objects (Zheng, Li, and Wang 2019), image regions (Cornia, Baraldi, and Cucchiara 2020), abstract scene graphs (Chen et al. 2020) and verb-specific semantic roles (Chen et al. 2021). 2) Structure-related: the control signal is related to the semantic structures of sentences, including sentence length (Deng et al. 2020) and Part-of-Speech (POS) tags (Deshpande et al. 2019). Trained with cross-entropy loss, existing CIC models have achieved satisfactory performance in terms of controllability.

Nevertheless, current structure-related CIC works are unable to combine with reinforcement training methods (Renjie et al. 2016; Gao et al. 2019), which prevents models from generating more accurate sentences. It has been proved that models trained only with cross-entropy loss suffer from
exposure bias (Ranzato et al. 2015), since in the training stage, the word at each time step is generated conditioned on ground truth words while in the testing stage, the word is generated based on previously predicted words of the model. Besides, the inconsistency between the cross entropy loss in the training stage and non-differentiable evaluation metrics in the testing stage also leads to unsatisfactory results. Therefore, reinforcement training, which can solve the above problems, is crucial for structure-related CIC models.

For content-related CIC tasks, only the models in (Cornia, Baraldi, and Cucchiara 2020; Chen et al. 2021) are trained with REINFORCE algorithm, where control signals aligned with ground truth captions are used as inputs. Under the guidance of these content-related control signals, the content of the generated sentence is trained to approach the ground truth caption, so that CIC models are able to learn content-related controllability. However, this method can only be applied to content-related control signals since the reward in reinforcement learning is designed to measure the content similarity between two sentences. For structure-related control signals, the reward makes the generated sentence approach the ground truth sentence in contents rather than semantic structures. Therefore, CIC models fail to learn structure-related controllability with ground truth annotations during conventional reinforcement training (Fig. [1]).

In this paper, we propose a novel reinforcement training method, Self-Annotated Training (SAT), for structure-related control signals. The main difference between our SAT and the method in (Cornia, Baraldi, and Cucchiara 2020) is the source of input control signals. During reinforcement training, control signals in (Cornia, Baraldi, and Cucchiara 2020) come from ground truth captions, while ours are from generated sentences of CIC models. For this purpose, we design a recursive annotation mechanism (RAM) which forces the input control signal to match the actual output sentence. Moreover, we propose an extra alignment reward to finetune the CIC model trained after SAT method. Under the extra supervision of the alignment reward, the controllability of the CIC model is further enhanced. Experiments on MSCOCO dataset (Lin et al. 2014) demonstrate that SAT can effectively improve the accuracy and controllability of captioning models controlled by structure-related signals. Besides, we conduct quantitative experiments on different structure-related control signals and on different baseline models, which show the generalizability of our SAT method.

In summary, we mainly make the following contributions in this paper:

- We propose a novel reinforcement training method for structure-related CIC: Self-Annotated Training (SAT), which can be easily incorporated into existing captioning models to make them generate more accurate and controllable sentences. To the best of our knowledge, SAT is the first reinforcement training method for structure-related CIC models.

- We propose an extra alignment reward to finetune the CIC model trained after SAT method. Under the extra supervision of the alignment reward, the controllability of the CIC model is further improved.

- We perform extensive experiments on different structure-related control signals and on different baseline models, which demonstrate the effectiveness and generalizability of our methods.

2 Related Work

2.1 Image Captioning

Existing captioning models follow an encoder-decoder framework which is first introduced to image captioning tasks by (Vinyals et al. 2015). In (Xu et al. 2015; Huang et al. 2019), attention mechanism is used to select the target area of interest that needs special attention at each time step. To enhance the diversity, GAN-based methods (Dognin et al. 2019; Dai et al. 2017; Chen et al. 2018) are introduced in image captioning. Models proposed in (Zheng, Li, and Wang 2019; Ge et al. 2019) change the order of the sentence generation, starting from the middle or the end of sentences. Two-step networks are designed in (Song et al. 2020; Gao et al. 2019) to generate refined captions from raw information. In (Yang et al. 2018; Chen et al. 2020; Shi et al. 2020), scene graphs are employed to further explore the objects, attributes and relationships in the image, which improve the overall performance of captioning models. In order to solve the long-term dependency problem in the previous LSTM architectures, Transformer-based models (Luo et al. 2021; Cornia et al. 2020; Guo et al. 2020; Li et al. 2019) using multi-head self- and encoder-decoder attention mechanisms are explored. Regarding the training strategy, reinforcement learning methods (Rennie et al. 2016; Gao et al. 2019) are proposed to optimize non-differentiable metrics, solving the problem of exposure bias. In this paper, we expand the scope of application of reinforcement training from conventional captioning tasks to CIC tasks. With our self-annotated training, the accuracy and controllability of CIC models are further improved.

2.2 Controllable Image Captioning

In addition to the conventional captioning task, another related route is to generate controllable captions, which is called controllable image captioning (CIC). CIC models aim to generate captions conditioned on designed control signals. As classified in (Chen et al. 2021), control signals can be roughly divided into two categories: 1) Content-related: the control signal is related to the contents of images. Models in (Zheng, Li, and Wang 2019) generate sentences starting from a guiding object in order to contain the given word. Models in (Cornia, Baraldi, and Cucchiara 2020) describe images conditioned on a given sequence or set of image regions to control which objects are described and their orders. In (Chen et al. 2020), Abstract Scene Graphs (ASG) are taken as the control signal to control sentences at a more fine-grained level. (Chen et al. 2021) proposes a new control signal, Verb-specific Semantic Roles (VSR), which meets both event-compatible and sample-suitable requirements. 2) Structure-related: the control signal is related to the semantic structures of sentences. To explore length-aware image captioning models, sentence length is studied in (Deng et al. 2014).
as the signal of “length level”. Models in (Deshpande et al. 2019) employ signals of Part-of-Speech (POS) tag sequences to make generated sentences diverse. However, existing reinforcement training methods are not applicable to structure-related CIC models above. In this work, we propose the SAT method, the first reinforcement training method for structure-related signals, to achieve high accuracy and controllability of models.

3 Preliminaries

3.1 Embedding Method

| $\beta_{\text{len}}$ | length | $\beta_{\text{ten}}$ | tense |
|------------------|--------|--------------------|-------|
| 0                | $\leq 8$ | 5                  | no v  |
| 1                | 9      | 6                  | be + v|
| 2                | 10     | 7                  | v-ing |
| 3                | 11     | 8                  | v     |
| 4                | $\geq 12$ | 9                  | v-ed  |

Table 2: Specific settings of control levels for sentence length and tense.

In this work, we conduct experiments on structure-related control signals of sentence attributes, including sentence length and tense. For each caption, we divide it into a length level $\beta_{\text{len}}$ and a tense level $\beta_{\text{ten}}$ according to its attribute. The specific settings of control levels for sentence length and tense are shown in Table 2. We try to make the number of samples at each level evenly distributed.

Take the task of controlling the sentence length and tense simultaneously as an example. Given an input caption $Y = \{y_1, y_2, \cdots, y_T\}$, we first obtain the control signal $\beta = \{\beta_{\text{len}}, \beta_{\text{ten}}\}$ according to the attribute of $Y$. For each element in $\beta$, an embedding matrix $E \in \mathbb{R}^{k \times d}$ (k is the number of levels in Table 2, and d is the embedding dimension) is employed to embed it to a d-dimensional vector space:

$$e_{\text{len}} = W^T \Pi_{\text{len}},$$

$$e_{\text{ten}} = W^T \Pi_{\text{ten}},$$

where $\Pi_{\text{len}}$ and $\Pi_{\text{ten}}$ are the one-hot representations of $\beta_{\text{len}}$ and $\beta_{\text{ten}}$, respectively. After obtaining the length level embedding $e_{\text{len}}$ and the tense level embedding $e_{\text{ten}}$, we calculate the control level embedding $e_\beta$ as (remove another part when only one element is desired to be controlled):

$$e_\beta = e_{\text{len}} + e_{\text{ten}}.$$  

Then, each word $y_i$ in caption $Y$ is represented by adding the control level embedding $e_\beta$ with word embedding $e_{y_i}$ and, optionally (for Transformer-based decoder), positional embedding $e_{p_i}$:

$$x_i = e_\beta + e_{y_i} + e_{p_i}.$$  

Finally, $x_i$ replaces original word embedding as the input of the decoder of captioning models. By integrating the control signal information into the word embedding, CIC models naturally associate the input control signal with the output sentence, thus learning the meaning of the control signal.

3.2 Training Strategy

Cross Entropy Training (XE) Given an image $I$, a target ground truth sequence $y^s_{1:T}$, the paired control signal $\beta$ and the captioning model with parameters $\theta$, we minimize the following cross-entropy loss:

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(y^*_t|y^s_{1:t-1}, I, \beta)),$$

where $\beta$ depends on the attribute of the ground truth sequence.

CIDEr Score Optimization (RL) After pretrained with cross-entropy loss, CIC models are further trained by REINFORCE algorithm. The training process is to minimize the negative expected reward:

$$L_{RL}(\theta) = -E_{y_{1:T} \sim p_\theta}(r(y_{1:T})), $$

where the reward $r(\cdot)$ is calculated according to the score of the evaluation metric (e.g. CIDEr (Vedantam, Zitnick, and Parikh 2015)). As in (Rennie et al. 2016), the gradient can be approximated as:

$$\nabla_\theta L_{RL}(\theta) \approx -\frac{1}{k} \sum_{i=1}^{k} (r(Y^s_i) - b) \nabla_\theta \log(p_\theta(Y^s_i | I, \beta_i)),$$

where $Y^s_i$ represents the $i$-th sampled caption and $b = \langle \sum_i r(Y^s_i) \rangle/k$ is the baseline, computed as the mean of the rewards obtained by the sampled captions. In this stage, the input control signal $\beta_i$ is calculated from the attribute of the $i$-th ground truth caption aligned with image $I$.

4 Method

In this section, we first elaborate on the specific process of Self-Annotated Training (SAT) in Section 4.1. Then, we introduce how to finetune the CIC model after SAT in Section 4.2.

4.1 Self-Annotated Training (SAT)

By reviewing the cross-entropy training method, we can find that the main reason why models learn controllability well is the consistency of inputs and outputs, that is, the input control signal can match the output sentence of the model (during cross-entropy training output sentences of models are ground truth captions). In this case, CIC models are able to naturally associate the input control signal with the output sentence, thus learning the meaning of the control signal. Therefore, during reinforcement training, control signals should be aligned with actual output sentences which are generated by sample methods rather than ground truth captions. However, output sentences generated by sample methods require control signals as the input. The control signals, in turn, are calculated based on output sentences. They are prerequisites to each other, which brings great obstacles to the actual operation. Therefore, how to design an algorithm to solve the above problem is crucial.

Figure 1 (a) gives an overview of our proposed SAT method. Its central idea is that the control signal is from
actual sampled sentences rather than ground truth captions. Given an image $I$ and its $k$ paired ground truth captions $G = \{G_1, G_2, \cdots, G_k\}$, for each ground truth $G_i$, we first calculate the control signal $\beta_i$ for caption $G_i$ according to its attribute. Then, a single Monte-Carlo sample is used to generate a sampled sentence $Y_i^s$ and its corresponding distribution $p(Y_i^s | I, \beta_i)$ conditioned on both the input image $I$ and the control signal $\beta_i$. It is worth noting that the attribute of $Y_i^s$ is not necessarily consistent with the control signal $\beta_i$ due to the limited capability of the CIC model. However, the extra reward is so tailored for the specific task that it is difficult to imitate the construction of it on other tasks.

In order to force the input control signal to match the output sampled sentence, a recursive annotation mechanism (RAM) is designed. After obtaining the sampled sentence $Y_i^s$, we compute the control signal $\beta_i^s$ for $Y_i^s$. Then, the model regards $Y_i^s$ as the target caption and predicts output distributions $p(Y_i^s | I, \beta_i^s)$ conditioned on $I$ and $\beta_i^s$, the process of which is the same as that in cross-entropy training. In this case, the input control signal $\beta_i^s$ is forced to match the output sampled sentence $Y_i^s$. Therefore, when the sampled sentence $Y_i^s$ which returns higher reward than baseline is encouraged to generate, the controllability of the CIC model is also enhanced. It should be noted that due to the existence of dropout, $p(Y_i^s | I, \beta_i)$ is probably different from $p(Y_i^s | I, \beta_i^s)$ even when $\beta_i$ and $\beta_i^s$ are the same. In this situation, we tend to retain the actual sampled output distribution $p(Y_i^s | I, \beta_i)$ instead of $p(Y_i^s | I, \beta_i^s)$, since considering the negative impact of exposure bias. The final output distribution $p'(Y_i^s | I, \beta_i^s)$ is updated as:

$$
p'(Y_i^s | I, \beta_i^s) = \begin{cases} p(Y_i^s | I, \beta_i), & \beta_i = \beta_i^s \\ p(Y_i^s | I, \beta_i^s), & \beta_i \neq \beta_i^s. \end{cases} \quad (7)
$$

After that, the CIDEr reward is calculated between $Y_i^s$ and all ground truth captions paired with image $I$, and finally sent into Eq. (6) with $p'(Y_i^s | I, \beta_i^s)$.

In fact, we find that samples resulting in lower rewards than baseline make the training process unstable. The reason for this phenomenon is that suppressing the output probability of sentences which are not sampled by the model itself leads to instability in the training process. Therefore, we discard the optimization of these sentences and change Eq. (6) to:

$$
\nabla_{\theta} L_{RL}(\theta) = -\frac{1}{k} \sum_{i=1}^{k} [r_i(Y_i^s) - b]_+ \nabla_{\theta} \log(p_Y(Y_i^s | I, \beta_i^s)), \quad (8)
$$

where $[x]_+ = \max(x, 0)$. Algorithm 1 summarizes the entire process.

### 4.2 Finetuning

After the CIC model converges with self-annotated training, we propose a finetuning method to further improve the controllability of the model. Our work is inspired by the extra reward proposed in (Cornia, Baraldi, and Cucchiara 2020). It is designed to evaluate the alignment with respect to the input control signal, thus enhancing the controllability of the model. However, the extra reward is so tailored for the specific task that it is difficult to imitate the construction of it on other tasks.

In this paper, we propose a general construction of reward for structure-related controllability (Figure 1(b)). Following
the mathematics notation in Section 4.1, $\beta_i$ and $\beta^s$ are given from steps 2 to 5 of Algorithm 1. We first compute their embedding vectors $e_i = \{e_{i \text{ten}}, e_{i \text{ten}}\}$ and $e^s_i = \{e_{i \text{ten}}, e_{i \text{ten}}\}$ as in Eq. (1). Then, the extra reward, which evaluates the alignment between the input control signal $\beta_i$ and the attribute $\beta^s_i$ of the output sentence, is formulated as the form of Euclidean distance:

$$r_{\text{align}} = -\frac{\|e_{i \text{ten}} - e^s_{i \text{ten}}\|^2 + \|e_{i \text{ten}} - e^s_{i \text{ten}}\|^2}{2\sqrt{d}},$$

where $d$ is the embedding dimension. The final reward $r(Y^*_i)$ is a weighted sum of CIDEr score and the alignment score:

$$r(Y^*_i) = r_{\text{cider}} + \lambda r_{\text{align}},$$

where $\lambda$ is a trade-off parameter to balance the contributions between accuracy and controllability. In the finetuning stage, the CIC model is optimized by conventional reinforcement training with Eq. (5) instead of SAT method. The generalizability of our finetuning method comes from that as long as control signals can be encoded into vectors, our method is able to evaluate the alignment reward. Under the extra supervision of the alignment reward, the controllability of CIC models is improved.

5 Experiments

5.1 Datasets and Evaluation Metrics

We use the MSCOCO 2014 captions dataset (Lin et al. 2014) to evaluate our proposed methods. MSCOCO dataset includes 164,062 images labeled with 5 captions each. Following the Karpathy data split (Karpathy and Fei-Fei 2016) which has been widely used in prior work, we choose 113,287 for training, 5,000 images for validation and 5,000 images for test. We measure the caption quality by using five evaluation metrics, including BLEU (Papineni et al. 2002), ROUGE-L (Lin 2004), METEOR (Denkowski and Lavie 2014), CIDEr (Vedantam, Zitnick, and Parikh 2015) and SPICE (Anderson et al. 2016). Following the approach in (Chen et al. 2021), we adopt a new metric Control Precision (CP) to measure the alignment with the input control signal.

5.2 Implementation Details

We choose AoANet (Huang et al. 2019) as our baseline model and follow (Huang et al. 2019) to set its hyperparameters. Specifically, we extract image features by employing Faster-RCNN (Ren et al. 2015) pretrained on Visual Genome (Krishna et al. 2017), thus obtaining a 2048-dimensional feature vector for each region. The input word embedding size and the hidden state size are all set to 1024. We adopt Adam optimizer to minimize the cross-entropy loss for 30 epochs, and then use self-annotated training with a fixed learning rate of $5 \times 10^{-6}$ for another 20 epochs. After that, AoANet is further trained with the finetuning method for 5 epochs. The batch size is set to 10 and the beam size is set to 2. The trade-off coefficient $\lambda$ in Eq. (10) is set to 1.

5.3 Quantitative Analysis

By reviewing the previous works for CIC tasks, we find that content-related works (Cornia, Baraldi, and Cucchiara 2020; Chen et al. 2020, 2021) prefer to use “1 caption to 1 ground truth” to test the results of CIC models. “1 caption to 1 ground truth” means the accuracy-based metrics are calculated between the generated sentence and the single ground truth caption which provides the control signal. By contrast, structure-related works (Deshpande et al. 2019, Deng et al. 2020) prefer to adopt “1 caption to 5 ground truth”, which means the accuracy-based metrics are calculated between the generated sentence and all ground truth captions aligned with the image. On the surface, “1 caption to 1 ground truth” pays more attention to controllability while “1 caption to 5 ground truth” pays more attention to accuracy. In order to fully demonstrate the performance of our method, we show the results in both situations. Note that in the following sections except Section 5.3 the symbol “SAT” represents the whole process of SAT + Finetuning.

Ablative Analysis. To examine the impact of our proposed SAT method and finetuning method, we choose
The introduction of SAT or finetuning method brings an improvement both in accuracy and controllability. Specifically, the control precision rises with the increase of \( \lambda \). The experiments above validate the effectiveness of our SAT and finetuning method.

**Model Selection with \( \lambda \).** In our proposed finetuning method, we combine two rewards together with a trade-off coefficient \( \lambda \) in Eq. (10). Figure 2 shows the results of our finetuning method with different \( \lambda \). It is obvious that the control precision rises with the increase of \( \lambda \) both achieving more than 99% control precision when the \( \lambda \) is set to 5, which verifies the effectiveness of the alignment reward in controllability. However, focusing too much on controllability reduces the accuracy of CIC models. As in Figure 2 (a), when \( \lambda \) is too large the CIDEr-D score drops by almost 2. Since both models achieve their best CIDEr-D performances at \( \lambda = 1 \), we eventually select \( \lambda = 1 \) for our finetuning method. By comparing the results between Finetuning (w/ SAT) and Finetuning (w/o SAT), we find that the model trained after SAT method is 1.5 higher in CIDEr-D on average than that without using SAT method, which demonstrates the necessity of our SAT method.

**Generalizability on Different tasks.** Table 4 reports the evaluation of AoANet w/ or w/o our SAT method in different tasks. As it can be observed, our SAT method significantly outperforms the XE method in all accuracy-based evaluation metrics, especially in the single task. Meanwhile, our SAT method maintains the high controllability, improving the control precision compared with the XE method. The performance above prove the effectiveness and generalizability of our SAT method.

| Task | Training strategy | 1 caption to 1 ground truth | 1 caption to 5 ground truth | con. |
|------|-------------------|-----------------------------|-----------------------------|------|
| Length Tense | XE/RL SAT Finetune | B-1 B-4 M R C S | B-1 B-4 M R C S CP | CP |
| XE × × | 45.3 15.9 19.7 41.8 147.6 28.2 | 74.5 33.8 27.9 56.1 111.3 21.3 | 97.6 |
| RL × × | 39.6 14.2 18.6 42.0 140.7 28.0 | 74.0 36.1 27.0 57.0 116.4 20.6 | 28.7 |
| RL × ✓ | 45.8 15.2 19.8 42.1 148.0 28.4 | 76.8 34.7 28.2 56.9 120.3 21.8 | 98.3 |
| RL ✓ × | 45.9 15.6 19.9 42.2 150.4 28.6 | 77.0 35.0 28.3 57.0 120.5 21.8 | 97.7 |
| RL ✓ ✓ | 45.9 15.4 19.9 42.2 149.7 28.7 | 77.4 35.0 28.3 57.1 121.9 22.0 | 98.3 |

Table 4: Evaluation of AoANet w/ or w/o our SAT method in different tasks.
Table 5: Sample results of controllability with different control signals. In order to fully present the role of each control signal, we separately train two models with different control signals to control the sentence length and tense respectively.

| Models            | B-1 | B-4 | M   | R   | C   | S   |
|-------------------|-----|-----|-----|-----|-----|-----|
| UpDown (XE)       | 40.8| 12.0| 17.6| 37.5| 115.4| 25.2|
| UpDown (SAT)      | 42.5| 12.2| 18.2| 39.1| 120.9| 25.8|
| AoANet (XE)       | 42.3| 12.6| 18.5| 38.8| 122.1| 26.3|
| AoANet (SAT)      | 43.4| 12.9| 18.9| 39.8| 128.5| 26.9|
| Transformer (XE)  | 41.0| 12.2| 17.9| 37.6| 117.8| 25.6|
| Transformer (SAT) | 43.3| 12.6| 18.7| 39.6| 128.6| 26.5|

Table 6: Evaluation of different baseline models w/ or w/o our SAT method in the task of sentence length.

Generalizability on Different Baseline Models. Table 6 shows the generalizability of our SAT method in different baseline models, including UpDown (Anderson et al. 2018), AoANet (Huang et al. 2019) and Transformer (Vaswani et al. 2017). Due to the limited space, we only show the results of “1 caption to 1 ground truth” and leave the data of “1 caption to 5 ground truth” in the supplementary material. As reported in Table 6, all baseline models with our SAT method outperform that with XE training in terms of all evaluation metrics. Concretely, boosting the CIDEr-D score from 117.8 to 128.6 on the Transformer baseline, verifies the advantage and generalizability of our SAT method.

Comparison with Previous Works. In the field of controllable image captioning, only (Deng et al. 2020) and (Deshpande et al. 2019) focus on structure-related control signals. Since the core contribution of (Deshpande et al. 2019) is high diversity and fast speed, we mainly compare our SAT method with the performance of models in (Deng et al. 2020). In this part, we follow the settings in (Deng et al. 2020) and divide the sentence length into several levels: [1, 9], [10, 14] and [15, 19]. In the test stage, the input control signal is artificially fixed at a certain level. Table 7 shows the performance comparisons between the previous works and our proposed approach in the task of sentence length. As it can be observed, AoANet trained by our SAT method achieves the best performance according to all metrics at all three levels. Boosting all evaluation metrics on the AoANet (XE) baseline, validates that our SAT method is able to greatly improve the accuracy of CIC models. Figure 3 reports the control precision of the above methods. As we can see, AoANet equipped with our SAT method reaches almost the same controllability as the previously best model LaBERT (Deng et al. 2020). They both achieve more than 99 % control precision at all three levels, and fully satisfy the needs of CIC tasks. The results above prove that our SAT method significantly improves the accuracy-based performance of CIC models while maintaining high controllability.

5.4 Qualitative Analysis. Table 5 shows some examples controlled by different requirements of sentence attributes. Due to the limited space, we only show part of the tense level and leave the whole table in the supplementary material. When faced with different control signals, the CIC model trained with our SAT method
Table 7: Performance comparisons with the models in (Deng et al. 2020) in the task of sentence length. † denotes the results of our trained model. The other data are from (Deng et al. 2020).

Figure 3: The control precision of our SAT version and three other versions in (Deng et al. 2020).

is able to generate various captions for the same image according to the demand. Take the performance in the length task as an example, with the increase of the control level, the length of the generated sentence also increases word by word. With our self-annotated training, the generated captions achieve both high accuracy and controllability, which illustrates the effectiveness of our methods.

6 Conclusion

In this paper, we focus on the reinforcement training method for controllable image captioning. For structure-related control signals, we propose a novel reinforcement training method called SAT, which adopts a recursive annotation mechanism to force the input control signal to match the output sentence. Moreover, we propose an extra alignment reward to finetune the CIC model trained after SAT method. Extensive experiments validate the effectiveness of our methods in terms of accuracy and controllability.

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. SPICE: Semantic Propositional Image Caption Evaluation. *Adaptive Behavior*, 11(4): 382–398.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; and Lei, Z. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, C.; Mu, S.; Xiao, W.; Ye, Z.; Wu, L.; and Ju, Q. 2018. Improving Image Captioning with Conditional Generative Adversarial Nets.

Chen, L.; Jiang, Z.; Xiao, J.; and Liu, W. 2021. Human-like Controllable Image Captioning with Verb-specific Semantic Roles.

Chen, S.; Jin, Q.; Wang, P.; and Wu, Q. 2020. Say As You Wish: Fine-grained Control of Image Caption Generation with Abstract Scene Graphs.

Cornia, M.; Baraldi, L.; and Cucchiara, R. 2020. Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cornia, M.; Stefanini, M.; Baraldi, L.; and Cucchiara, R. 2020. Meshed-Memory Transformer for Image Captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dai, B.; Fidler, S.; Urtasun, R.; and Lin, D. 2017. Towards Diverse and Natural Image Descriptions via a Conditional GAN.

Deng, C.; Ding, N.; Tan, M.; and Wu, Q. 2020. Length-Controllable Image Captioning.

Denkowski, M.; and Lavie, A. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*.

Deshpande, A.; Aneja, J.; Wang, L.; Schwing, A. G.; and Forsyth, D. 2019. Fast, Diverse and Accurate Image Captioning Guided by Part-Of-Speech. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dognin, P.; Melnyk, I.; Mroueh, Y.; Ross, J.; and Sercu, T. 2019. Adversarial Semantic Alignment for Improved Image Captions. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gao, J.; Wang, S.; Wang, S.; Ma, S.; and Gao, W. 2019. Self-critical n-step Training for Image Captioning.

Ge, H.; Yan, Z.; Zhang, K.; Zhao, M.; and Sun, L. 2019. Exploring Overall Contextual Information for Image Captioning in Human-Like Cognitive Style.

Guo, L.; Liu, J.; Lu, S.; and Lu, H. 2019. Show, Tell and Polish: Ruminant Decoding for Image Captioning. *IEEE Transactions on Multimedia*, PP(99): 1–1.
