Adaptive Text Denoising Network for Image Caption Editing

MENGQI YUAN, Nanjing University of Posts and Telecommunications, China
BING-KUN BAO*, Nanjing University of Posts and Telecommunications, China
ZHIYI TAN, Nanjing University of Posts and Telecommunications, China
CHANGSHENG XU, Peng Cheng Laboratory; University of Chinese Academy of Sciences; NLPR, Institute of Automation, CAS, China

Image caption editing, which aims at editing the inaccurate descriptions of the images, is an interdisciplinary task of computer vision and natural language processing. As the task requires encoding the image and its corresponding inaccurate caption simultaneously and decoding to generate an accurate image caption, the encoder-decoder framework is widely adopted for image caption editing. However, existing methods mostly focus on the decoder, yet ignore a big challenge on the encoder: the semantic inconsistency between image and caption. To this end, we propose a novel **A**daptive **T**ext **D**enoising **Net**work (ATD-Net) to filter out noises at word level and improve the model's robustness at sentence level. Specifically, at the word level, we design a cross-attention mechanism called Textual Attention Mechanism (TAM), to differentiate the misdescriptive words. The TAM is designed to encode the inaccurate caption word by word based on the content of both image and caption. At the sentence level, in order to minimize the influence of misdescriptive words on the semantic of an entire caption, we introduce Bidirectional Encoder to extract the correct semantic representation from the raw caption. The Bidirectional Encoder is able to model the global semantics of the raw caption, which enhances the robustness of the framework. We extensively evaluate our proposals on the MS-COCO image captioning dataset and prove the effectiveness of our method when compared with the state-of-the-arts.

Additional Key Words and Phrases: Image caption editing, Sequence editing, Cross-modal semantic matching.

ACM Reference Format:
Mengqi Yuan, Bing-Kun Bao, Zhiyi Tan, and Changsheng Xu. 2022. Adaptive Text Denoising Network for Image Caption Editing. *J. ACM* 37, 4, Article 111 (August 2022), 18 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Image caption, which aims to bridge the gap between visual and language modalities, is an interdisciplinary task of computer vision (CV) and natural language processing (NLP). Beyond traditional image caption task, image caption editing aims to correct the inaccurate descriptions of the images, as illustrated in Fig. 1. This task plays a crucial role in many complex applications, such as removing misdescriptive words in image captioning datasets and content-based image retrieval [7, 9, 53, 56]. However, it also imposes higher requirements for the precise alignment between the modalities of visual and language.

*Corresponding author
†The code is publicly available at https://github.com/NJUPT-MCC/ATD-Net.

Authors' addresses: Mengqi Yuan, 2020010306@njupt.edu.cn, Nanjing University of Posts and Telecommunications, Nanjing, China; Bing-Kun Bao, bingkunbao@njupt.edu.cn, Nanjing University of Posts and Telecommunications, Nanjing, China; Zhiyi Tan, Nanjing University of Posts and Telecommunications, Nanjing, China; Changsheng Xu, Peng Cheng Laboratory; University of Chinese Academy of Sciences; NLPR, Institute of Automation, CAS, Beijing, China.

0004-5411/2022/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

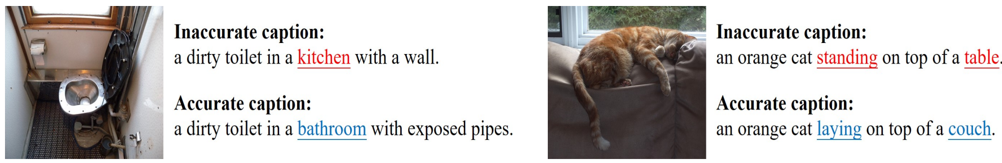

Fig. 1. Two examples of inaccurate captions versus accurate captions on corresponding images. The words in red are misdescriptive words and the words in blue are correct ones. In the first example, the task of image caption editing needs to edit the word "chair" to "table". In the second example, the task of image caption editing requires editing the word "standing" to "laying" and "table" to "couch".

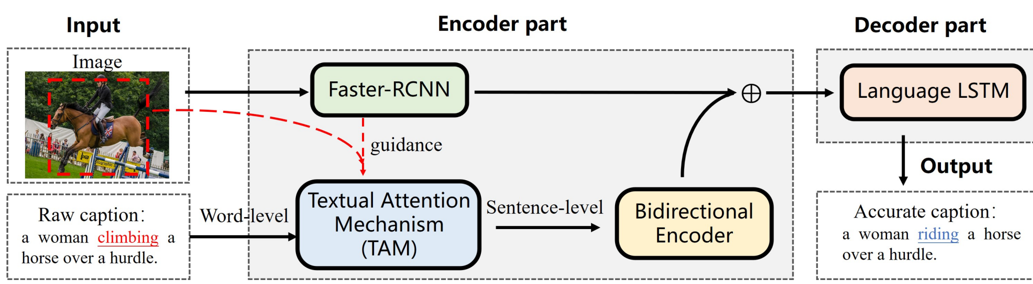

Fig. 2. For a given image and a raw caption, the encoder of ATD-Net first uses Textual Attention Mechanism (TAM) and Bidirectional Encoder to process the raw caption under the guidance of the image features extracted by Faster-RCNN. Subsequently, the decoder generates an accurate caption with the help of feature vectors from both the image and the raw caption.

In recent years, methods based on encoder-decoder architecture have achieved initial success in image caption editing [6, 7]. At one end, Convolutional Neural Network (CNN) [1, 2] and Long Short-Term Memory (LSTM) [7] are commonly used to encode the input images and the corresponding inaccurate captions into fixed-size feature vectors. While at the other end, LSTM or Transformer [13] are applied to decode the encoded feature vectors into a sequence of words. Based on this framework, Fawaz and Mahmoud [6] separately encode raw caption and image with the proposed Deep Averaging Network (DAN) and CNN, then utilize the gate function of LSTM to simultaneously decode two-mode representations to generate a new caption. Fawaz and Mahmoud [7] propose an editing network in the decoder, which generates each word by selecting the LSTM cell state corresponding to the most relevant words.

However, the text noises caused by semantic inconsistency between image and caption will affect both encoder and decoder. If the raw caption and misdescriptive words are encoded into sentence features and word features, it will induce error accumulation in the training process. For example, the misdescriptive words "chair" "standing" and "table" in Fig. 1 will cause inconsistent semantic representations of the entire sentence encoding. Moreover, existing models treat all words in the raw caption equally which leads to two major problems: First, since the entire sentence is usually encoded as a whole, it is difficult to locate and edit at word level. Second, because the misdescriptive words may distort the semantics of an entire sentence, it will cause greater misalignment of the semantics of images and text. Existing models usually adopt forward sequential encoding methods, which introduce the hidden states of misdescriptive words into all the forward positions in sentence encoding, leading to the semantic errors of the entire sentence.

Based on this observation, we propose a novel encoder-decoder architecture, called **A**daptive **T**ext **D**enoising **Net**work (ATD-Net), which adds denoising process into the encoder to reduce semantic inconsistency between image and caption (shown in Fig. 2). The major contribution of this proposal is the two-step encoding of the raw caption at word and sentence levels: (1) Differentiating the possible noises at the word level through a newly designed cross-modal attention module called Textual Attention Mechanism (TAM). The misdescriptive words can be considered as white noises in principle, because their semantics are quite different from those of correct ones, and their locations are unknown. Using the correct image content as a guide is a very effective way to differentiate them in the raw caption. Inspired by the self-attention mechanism in BERT models [28], we design TAM as a cross-attention structure. It assigns low weight to noisy words through the guidance of image content and recodes each word embedding of the raw caption at word level by attention weights. (2) Encoding weighted word embedding sequence into a text feature vector by Bidirectional Encoder at sentence level. Since misdescriptive words easily lead to deviations of the semantics at sentence level, the encoder should consider adjacent words in both forward and backward directions to improve its robustness. Through bi-directional coding of the sequence of word embeddings, ATD-Net is able to comprehensively consider the semantic information of both past and future words, which further enhances the model's robustness. Finally, the decoder generates accurate captions word by word with both text features and image features extracted from the encoder.

In the experiments, to further test the denoising ability of our model, we artificially add variable proportions of misdescriptive words into different parts of speech in the raw captions. Experimental results on the MS-COCO dataset demonstrate that the proposed ATD-Net model outperforms or is competitive with the state-of-the-art methods and shows pretty good ability on misdescriptive word editing.

Our contributions are summarized as follows:

(1) We design a Textual Attention Mechanism (TAM) in ATD-Net to differentiate the possible noises from the raw captions, which achieves the cross-modal semantic matching of image and caption at word level.

(2) We propose a Bidirectional Encoder at sentence level to make the model focus on the global information contained in the raw caption when generating each word of a new caption.

(3) We add misdescriptive words into different parts of speech with variable proportions to the widely used MS-COCO dataset. Through conducting extensive experiments on the MS-COCO, we demonstrate that the proposed ATD-Net can effectively edit inaccurate captions and has great robustness.

The rest of the paper is organized as follows: Section 2 reviews the related work. Section 3 presents our method in details. Section 4 reports the experiment results. Section 5 concludes the paper with future work.

2 RELATED WORK

In this section, we review the recent studies related to image caption editing. To conclude, this task can be divided into two sub-categories according to their inputs, that is, image captioning and sequence-to-sequence editing.

2.1 Image Captioning

Image captioning aims to find cross-modal associations between image and text. The neural encoder-decoder model has achieved initial success in the field of image caption in past few years. In the basic encoder-decoder framework, the convolutional neural network (CNN) is usually adopted to encode the image to feature vectors, while the recurrent neural network (RNN) is used to decode the feature vectors to a sentence word by word [1, 2]. Later on, inspired by the excellent performance of self-attention mechanism in the BERT [11], a number of approaches [10, 24] use

Transformer to decode visual features into captions effectively. Therefore, encoder-decoder models based on CNN and Transformer become common recently. Later, some researchers try to model the relationship between image and text by embedding new neural network modules [8, 20] into the encoder-decoder structure. For example, some methods [22, 25, 55, 57] use GCN and scene graphs to model the correlation between objects in the image. Moreover, recent advances in image captioning use deep reinforcement learning (RL) [4, 21, 26] to alleviate the "exposure bias" during cross-entropy training. In this paper, we also choose to train the basic encoder-decoder framework with cross-entropy loss and reinforcement learning respectively.

In the cross-modal image caption task, the generated sentences are susceptible to incorrect cross-modal semantic matching. To this end, the attention mechanism [19, 23, 54, 58] has been widely used in recent years to align image and text across modalities. Xu et. al. [3] integrate soft and hard attention mechanisms into LSTM based decoder, selecting the most relevant image regions for word prediction at each decoding stage. On the other hand, Anderson et al. [5] utilize bottom-up and top-down attention mechanisms, which can calculate attention weight at the level of objects and other salient image regions. After that, Guo et. al. [9] propose a ruminant image captioning framework, which attempts to introduce the polishing process into image caption generation procedure. In addition, the structure with transformer as decoder [14] has also been well applied in the field of image caption in recent years. Self-attention mechanism is used to find more fine-grained features in images [10, 28]. However, the existing work still cannot achieve effective semantic alignment of image and text, which leads to the existence of description bias. In this paper, we propose an Adaptive Text Denoising Network to further realize the semantic matching of image and caption by accurately locating and editing noisy words from raw captions.

2.2 Sequence-to-Sequence Editing

Sequence-to-sequence editing can be roughly divided into two categories according to the task modalities: one is single-modal text sequence editing, and the other is cross-modal image caption editing.

Text sequence editing is a classic single-modal natural language processing (NLP) task. In the past few years, inspired by the performance of the encoder-decoder structure in machine translation [17, 18], a number of approaches [40–43] attempt to use a deep neural network to deal with the task of sequence-to-sequence editing and have achieved initial success. Kyunghyun et. al. [40] first propose the Seq2seq encoder-decoder structure based on RNN, then a variety of attention mechanisms [41–43] are introduced into the basic model of [40] and achieve better performance. LaserTagger [44] achieves finer editing of text sequences by combining Bert encoder and autoregressive transformer decoder. Meanwhile, Recurrence [45] improves the performance of model by narrowing down the editing sentence length through a reasoning algorithm based on recursive iteration. Later, some approaches [46, 47] begin to focus on more complex text sequence editing tasks. Quantifiable Sequence Editing (QuaSE) [46] uses content and result similarities to model pseudo-parallel sentences, which makes the generated sequence closer to the pre-defined goals. Pre-training of Denoising Autoencoders (PoDA) [47] considers the influence of text noise on sequence editing, which first pre-trains the noisy data and then fine-tunes the transformer model to enhance generalization performance.

Unlike traditional text sequence editing tasks, image caption editing is a cross-modal sequence editing task, which requires not only text information in the raw caption but also visual information in the image. Therefore, the encoder needs to encode both image and text modal information at the same time [7]. Moreover, since the semantic gap between the image and caption significantly influences the quality of generated captions, the task has high requirements on cross-modal semantic alignment. Existing mainstream methods for image caption editing is based on encoder-decoder framework. The encoder encodes the image and text into feature vectors respectively, and the decoder generates accurate image descriptions on the basis of these feature vectors. For example, Fawaz and Mahmoud [6] introduce a novel framework that learns what to be kept, removed or added to the existing caption from a given framework at each timestep. This study uses Deep Averaging Network (DAN) to encode the existing captions. However, DAN ignores the word-level noises when encoding sentences into feature vectors, which may lead to serious error propagation in the training of the model. After that, Fawaz and Mahmoud [7] propose an edit network to image captioning based on iterative adaptive refinement of an existing caption. In this study, each word selected from the raw caption has its corresponding memory state and is copied into the internal structure of the LSTM at each decoding step. Although this method has shown its effectiveness in the experimental performance, it wastes much semantic information as it discards all the information after the most relevant word in the raw caption. In this paper, we introduce a new framework for image caption editing, in which we employ Textual Attention Mechanism (TAM) and Bidirectional Encoder to recode the raw caption from the word level and sentence level.

3 OUR METHOD

In this section, we introduce the whole framework of our Adaptive Text Denoising Network (ATD-Net), which is depicted in Fig. 3. Section III-A gives an overview of the framework. Section III-B and Section III-C introduce the Image Encoder and the Caption Encoder respectively. Section III-D presents the framework of Caption Decoder. Finally, we introduce the training objectives of ATD-Net in Section III-E.

3.1 Overview of the Framework

The goal of image caption editing is to generate a sentence $\widehat{S} = \{\widehat{w}_1, ..., \widehat{w}_T\}$ that accuratly describes the image content, given an image I and a raw caption C with misdescriptive words. The objective is to maximize the sum of log-likelihood of the corresponding words:

$$\theta^* = \arg\max_{\theta} \sum_{t=1}^{T} \log p(\widehat{w}_t | I, C, \widehat{w}_0, ..., \widehat{w}_{t-1}, \theta) \tag{1}$$

where $\widehat{w}_t$ is the t-th word in a sentence $\widehat{S}$, T is the sentence length and θ represents the parameters to be learned.

The framework of our ATD-Net is an encoder-decoder structure. The encoder is used to transform both image and the raw caption into fixed-size feature vectors respectively, while the decoder is used to generate accurate description by image and textual feature vectors. Specifically, the encoder consists of two parts: Image Encoder and Caption Encoder. In Caption Encoder, instead of encoding the entire sentence through DAN and LSTM as most of the existing methods do, we combine the newly designed Textural Attention Mechanism (TAM) with the Bidirectional Encoder to extract more accurate semantic representations from the raw caption. Note that, the input of Caption Encoder requires not only the raw caption but also the output of the current state in the decoder that contains the image content.

3.2 Image Encoder

Given an image I, we encode it into the spatial image features with CNN encoder followed by previous work [5]:

$$V = CNN(I) \tag{2}$$

where $V = \{v_1, v_2, ..., v_k\}, v_i \in R^{2048}$, and each image feature v_i encodes a salient region of the image. k is the number of regions, which is set to 36 in this paper. Specifically, we first use Faster

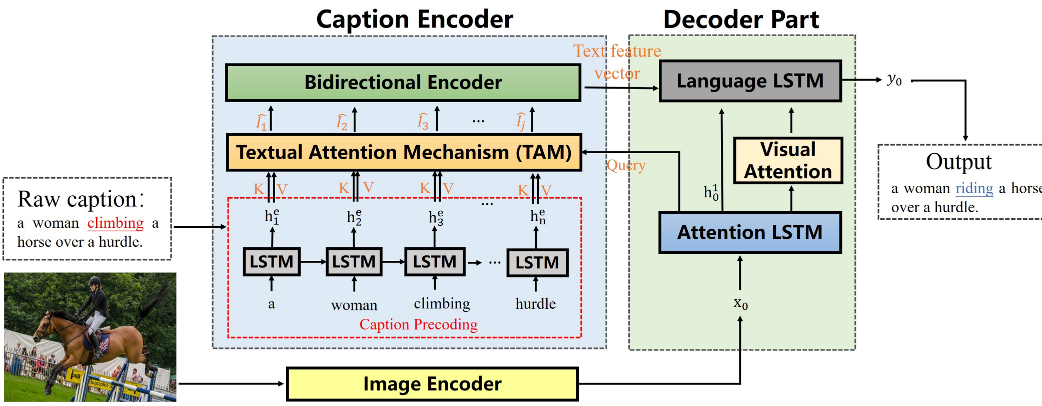

Fig. 3. The overview of our Adaptive Text Denoising Network (ATD-Net), which is an encoder-decoder based model. Image Encoder and Caption Encoder transform the input image and the raw caption into feature vectors respectively. Particularly, the Caption Encoder uses a Textual Attention Mechanism (TAM) to differentiate the possible noises in the raw caption at word level firstly. Then Bidirectional Encoder is applied to encode the extracted word feature to generate the text feature vector at sentence level. For the decoder, Attention LSTM and Language LSTM are used to fuse the text feature and image feature to generate a sentence that can describe the image accurately.

R-CNN with ResNet-101 [30] to divide every image into k sub-regions and encode them into feature vectors v_i. Subsequently, we can get the final feature vector of each image through mean pooling processing:

$$\overline{V} = \frac{1}{k} \sum\nolimits_{i=1}^{k} v_i \tag{3}$$

3.3 Caption Encoder

For the raw caption, ATD-Net encodes it in three steps: (1) Caption Precoding; (2) Textual Attention Mechanism; (3) Bidirectional Encoder. Firstly, Caption Precoding part pre-codes each word of the raw caption into a word feature vector. Subsequently, Textual Attention Mechanism designs a cross-modal attention module to differentiate the possible noises at the word level. Finally, Bidirectional Encoder encodes weighted word sequence into a text feature vector at sentence level.

3.3.1 ***Caption Precoding***. For a given caption C, we first use a one-layer LSTM to achieve the feature representation of each word in the raw caption like [7]:

$$\overline{H_c} = LSTM(C) \tag{4}$$

where $\overline{H_c} = [h^c_1, h^c_2, ..., h^c_n]$, and n is the number of words in caption. Each word in raw caption is precoded into a feature vector h^c_j by this method.

3.3.2 ***Textual Attention Mechanism***. The raw caption may have some misdescriptive words that do not match the semantic of image. Therefore, the word embedding $\overline{H_c}$ generated by Caption Precoding is noisy, and the location of these noisy word embeddings is unknown. To overcome this problem, we propose a Textual Attention Mechanism (TAM) (shown in Fig. 4(a)) inspired by Transformer to differentiate the possible noises at the word level. The input of TAM consists of two parts, one is $\overline{H_c}$ generated by Caption Precoding, and the other is hidden layer state h^1_t of Attention LSTM in the decoder. The TAM first assigns low weight to noisy word embedding in

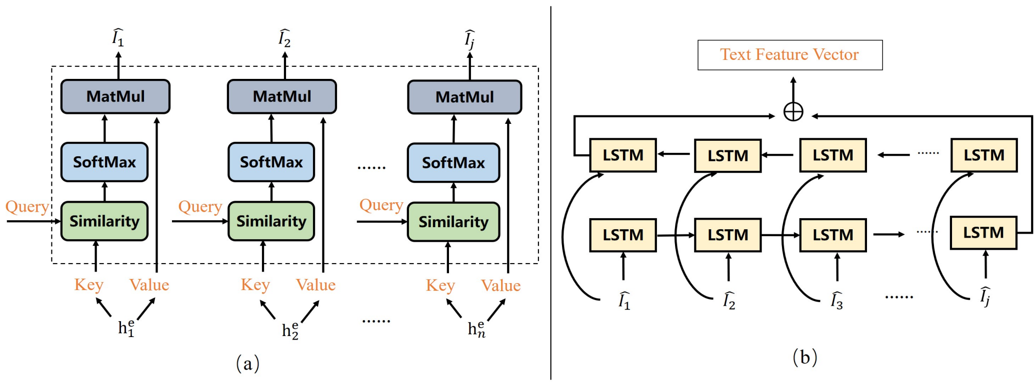

Fig. 4. (a): TAM first calculates similarity weight α_j between *Query* and *Key*. Subsequently, TAM recodes the word embedding *Value* to generate some weighted average $\widehat{I}$. (b): Bidirectional encoder module consists of two layers of LSTM units, which encodes the weighted words into a text feature vector by bidirectional coding.

$\overline{H_c}$, then multiplies the weighted word embedding with the hidden layer state h^1_t which contains the image content. Due to the semantic gap between visual and textual modalities, we first use Attention LSTM in the decoder part to map visual features into a shared dimensional space. Then, we use the feature vector h^1_t as 'Query' to distinguish noise words through TAM.

In our proposal, TAM bases on three sets of vectors, namely a set of *queries* Q, *keys* K and *values* V. *Queries* are the hidden layer state h^1_t at the current time of Attention LSTM in the decoder. The *keys* and *values* are both h^c_j generated by Caption Precoding. Firstly, we calculate the correlation weight α_j through a set of *queries* and *keys*. Then each word embedding is recoded with α_j to generate the attention weighted word embedding I_j as follows:

$$\alpha_j = softmax(W^T_\alpha \tanh(W_{c\alpha} h^c_j + W_{s\alpha} h^1_t)) \tag{5}$$

$$I_j = h^c_j \cdot \alpha_j \tag{6}$$

where $W_{c\alpha} \in \mathbb{R}^{H\times C}$, $W_{s\alpha} \in \mathbb{R}^{H\times S}$ and $W^T_\alpha \in \mathbb{R}^H$ are learned parameters.

In Eq 5, the correlation weight α_j is represented by the correlation coefficient between *query* h^1_t and *key* h^c_j. As the semantic of the misdescriptive word is white noise which is quite different from the correct words, α_j is able to differentiate the noisy words by calculating the similarity scores between the correct semantics guided by the image and the word semantics in the raw caption. In particular, word embedding that is more related to the image-guided semantic in the raw caption will obtain a higher attention score α_j, correspondingly, the correlation coefficient α_j of the noise position j will be smaller. The attention-weighted word embedding I_j then contains little semantic representation of noisy word embeddings.

In addition, unlike the traditional self-attention mechanism in Transformer, TAM is a cross-modal attention mechanism. Every time the decoder part inputs a *query* h^1_t, TAM will calculate n weighted word embedding based on $\overline{H_c}$. Specifically, in the self-attention mechanism, *key*, *value*, and *query* are the same elements in text modality, but the *key*, *value*, and *query* used by TAM in this paper are the different elements in both text and visual modalities.

3.3.3 ***Bidirectional Encoder***. Bidirectional Encoder generates a text feature vector containing the correct semantics from the sentence level of the raw caption. Through TAM, we obtain the weighted encoding I_j of each word according to the image semantic in the raw caption. However, due to the denoising process, some weighted word embeddings I_j are vacant. For the coding of these vacant word embeddings, all available input information in the past and future of a time frame needs to be considered. Therefore, in order to fully utilize the input information, we adopt Bidirectional Encoder to encode the weighted word embedding sequence. In this way, we integrate these weighted sequences and output the final text feature vector needed by each time frame of the decoder.

As shown in Fig. 4(b), The Bidirectional Encoder is designed as a Bi-LSTM based structure, which consists of two layers of LSTM with different direction encoding. The first layer is forward encoding LSTM: The input of each LSTM unit is the previous extracted word embedding I_j and the hidden layer state h^+_{j-1} of the previous time frame in forward encoding LSTM. The second layer is backward encoding LSTM: the input of each LSTM unit is also the extracted word embedding I_j and the hidden layer state h^-_{j+1} of the latter time frame in backward encoding:

$$\begin{aligned} a^+ &= BiLSTM^+(I_j) \\ a^- &= BiLSTM^-(I_j) \end{aligned} \tag{7}$$

where $BiLSTM^+$ and $BiLSTM^-$ are the forward coding layer and the backward coding layer of the Bidirectional Encoder respectively.

Finally, the text feature vector α_{text} extracted from the raw caption is generated by combining the output of the two layers as follows:

$$a_{text} = \frac{1}{2}(a^+ + a^-) \tag{8}$$

3.4 Caption Decoder

Following prior work [5], the Caption Decoder of our ATD-Net framework consists of two layers of LSTM: Attention LSTM and Language LSTM.

3.4.1 ***Attention LSTM****.* Attention LSTM generates the hidden layer state h^1_t which contains the image content at the current position and passes it to the Caption Encoder to calculate the attention weights. The structure of Attention LSTM is the same as that of traditional LSTM, and its input consists of three parts as follows:

$$x^1_t = [h^2_{t-1}, \overline{V}, W_e\Pi_t] \tag{9}$$

where h^2_{t-1} is the hidden layer state corresponding to the previous iteration step of Language LSTM, $\overline{V}$ is the mean-pooled image features, Π_t is one-hot encoding, and W_e is a word embedding matrix. Through the guidance of image content and ground-truth caption, Attention LSTM can generate noiseless hidden layer coding h^1_t at the current iteration position.

Since the training process of Attention LSTM and Caption Encoder are independent of each other, noisy words in the raw caption will not directly interfere with the training of the Attention LSTM.

3.4.2 ***Language LSTM****.* In the decoder, Language LSTM generates accurate captions word by word. The input of Language LSTM consists of four parts:

$$x^2_t = [h^1_t, h^2_{t-1}, \alpha_{text}, \widehat{V}_t] \tag{10}$$

where h^1_t is the hidden layer state of the current time frame of Attention LSTM, h^2_{t-1} is the hidden layer state of the previous time frame of Language LSTM, α_{text} is the text feature vector extracted from the raw caption by TAM and Bidirectional Encoder in Caption Encoder, and $\widehat{V}_t$ is the attended image feature used to focus on the most matching image area when generating each word with "soft" attention mechanism like [5, 7]. We use a normalized attention weight $\alpha_{i,t}$ for image features v_i of each region as follows:

$$\alpha_{i,t} = softmax(W_V^T \tanh(W_{v\alpha} v_i + W_{h\alpha} h_t^1)) \tag{11}$$

$$\widehat{V}_t = \sum_{i=1}^{K} \alpha_{i,t} \cdot v_i \tag{12}$$

where $W_{v\alpha} \in \mathbb{R}^{L \times V}$, $W_{h\alpha} \in \mathbb{R}^{L \times S}$ and $W_V^T \in \mathbb{R}^L$ are learned parameters.

Subsequently, to get an accurate sentence $y_{1:T} = (y_1, ..., y_T)$, the hidden state h_t^2 of Attention LSTM can be used to generate words at each time step t iteration position with maximum probability distribution:

$$p(y_t | y_{1:t-1}) = softmax(W_p h_t^2 + b_p) \tag{13}$$

where $W_p \in \mathbb{R}^{|\Sigma| \times S}$ and $b_p \in \mathbb{R}^{|\Sigma|}$ are learned weights and biases. The distribution is calculated as the product of the conditional distributions at all time steps:

$$p(y_{1:T}) = \prod_{t=1}^{T} p(y_t | y_{1:t-1}) \tag{14}$$

3.5 Objectives

Following previous studies on image captioning [1, 5], we train our model with a word-level cross-entropy loss (XE). Given the target ground truth sentence $y^*_{1:T} = (y^*_1, ..., y^*_T)$, we minimize the following cross entropy loss:

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(y^*_t | y^*_{1:t-1})) \tag{15}$$

where θ is the parameters of the captioning model.

Following prior work of Self-Critical Sequence Training [4], we further employ a reinforcement learning algorithm to directly optimize the metric of CiDEr-D. Specifically, the optimization objective is to minimize the negative expected reward as follows:

$$L_{RL}(\theta) = -E_{y_{1:T} \sim P_\theta}[r(y_{1:T})] \tag{16}$$

where $r(y_{1:T})$ is the cider score of the generated sentence.

The final policy gradient is calculated as follows:

$$\nabla_\theta L_{RL}(\theta) \approx -(r(y^s_{1:T}) - r(\widehat{y}_{1:T})) \nabla_\theta \log P_\theta(y^s_{1:T}) \tag{17}$$

where $r(y^s_{1:T})$ defines the cider score of a sampled caption and $r(\widehat{y}_{1:T})$ defines the baseline cider score obtained by greedily decoding the current model.

4 EXPERIMENTS

4.1 Dataset

We conduct our experiments on the most popular image caption dataset MS-COCO [48]. The whole MS-COCO dataset contains 123,287 images, in which there are 82,783 training images, 40,504 validation images, and 40,775 testing images with five human-annotated sentences. In this paper, we employ standard "Karpathy" data split [32] for model evaluation, where 113,287 images are used for training, 5,000 are used for validation, and 5,000 are used for testing.

Fig. 5. Some examples of the raw captions constructed in our experiment.

4.2 Implementation Details

For experimental data, we divide each image into 36 sub-regions through Faster R-CNN, which is pre-trained on ImageNet [31] and Visual Genome [36]. Each region is represented as a 2,048-dimensional feature vector. Similar to [7], we extract the words that appear over 5 times to form a vocabulary dictionary, and each word is represented as a "one-hot" vector.

We construct the raw inaccurate captions by artificially adding word noises into the captions generated by the previous model [7, 14]. As shown in Figure 5, we first use "AoANet" [14] to generate caption of the MS-COCO dataset, then randomly replace some common words with other words of the same part of speech, thereby obtaining the raw captions that need to be edited in our experiment.

For the proposed model, the hidden size of the LSTM in both encoder and decoder is set to 1024. We set the initial learning rate to 5×10^{-4}, and let it decay by 20% every three epochs. For the training stage, the whole architecture is firstly optimized by cross-entropy loss with 25 epochs. Then, we further optimize the metric of CIDEr scores with "Self-Critical Sequence Training" [4] for another 10 epochs. The whole experiment is trained and tested on NVIDIA Tesla V100 GPU.

Following the standard evaluation protocol, we utilize the metrics of BLEU@N [33], ROUGE-L [34], and CIDEr-D [35] to evaluate our model.

4.3 Baselines

We divide baselines into three categories according to whether the model introduces scene graphs and external pre-trained Transformers.

Group I: The first group does not utilize scene graph and pre-trained Transformer where our ATD-Net also belongs to. This group of models includes NIC [1], which uses CNN as encoder, and LSTM as decoder; SCST [4], which uses reinforcement learning to further optimize the CiDEr-D metric of the model; Adaptive [16], which uses an adaptive attention mechanism to dynamically focus on each image region in time sequence; Up-Down [5], which uses bottom-up and top-down attention mechanism to weigh the image features extracted by Faster R-CNN; RFNet [38], which fuses the visual information extracted from multi-layer CNN; MN [6], which encodes raw caption with DAN and utilizes decoder LSTM to simultaneously decode visual and textual representations to generate a new caption; AAT [49], which utilizes a novel Adaptive Attention Time module to align image and text adaptively; LBPF [50], which pays attention to both visual feature of the past and the predictive word of the future; SG-RWS [51], which adds a text retrieval module in decoder part to generate word by extracting the prior knowledge of other captions; and ETN [7], which

Table 1. Performance of our model and other state-of-the-art models on the MS-COCO "Karpathy" test split under cross-entropy training and CIDER-D score optimization. † indicates that scene graph is introduced in these models. * indicates that these models use a pre-trained transformer to perform self-attention on visual features.

Metric / Method	Cross-Entropy Loss						CIDEr-D Score Optimization			
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDEr	BLEU-1	BLEU-4	ROUGE	CIDEr
NIC[1]	-	-	-	29.6	52.6	94.0	-	31.9	54.3	106.3
SCST(Att2in)[4]	-	-	-	31.3	54.3	101.3	-	33.3	55.3	111.4
Adaptive[16]	74.2	58.0	43.9	33.2	54.9	108.5	-	-	-	-
Up-Down[5]	77.2	-	-	36.2	56.4	113.5	79.8	36.3	56.9	120.1
RFNet[38]	76.4	60.4	46.6	35.8	56.8	112.5	79.1	36.5	57.3	121.9
MN[6]	76.9	61.2	47.3	36.1	56.4	112.3	-	-	-	-
AAT[49]	-	-	-	37.0	57.3	117.2	80.1	38.5	58.2	126.7
LBPF[50]	77.8	-	-	37.4	57.5	116.4	80.5	38.3	58.4	127.6
SG-RWS[51]	77.1	-	-	36.6	56.9	116.9	80.3	38.5	58.4	129.1
ETN[7]	77.9	62.5	48.9	38.0	57.7	**120.0**	80.6	39.2	58.9	128.9
† GCN-LSTM[12]	77.3	-	-	36.8	57.0	116.3	80.5	38.2	58.3	127.9
† SGAE[39]	-	-	-	36.9	56.4	113.5	80.8	38.4	58.6	127.8
*AOA[14]	77.3	61.6	47.9	36.9	57.3	118.4	80.5	39.1	58.9	128.9
*X-Linear[13]	77.3	61.5	47.8	37.0	57.5	**120.0**	80.9	39.7	59.1	132.8
*DLCT[52]	-	-	-	-	-	-	**81.4**	**39.8**	**59.1**	**133.8**
ATD-Net(Ours)	**78.2**	**63.1**	**49.5**	**38.5**	**58.1**	118.5	80.8	39.3	59.0	128.6

generates each word by selecting the LSTM cell state corresponding to the most relevant words in the raw caption.

Group II: The second group introduces an additional Scene Graph on the basis of the first group. This group of models includes GCN-LSTM [12], which uses GCN to integrate the spatial information and semantic information extracted from the image; and SGAE [39], which constructs the visual relationship graph guided by the caption to improve the performance.

Group III: The third group introduces a pre-trained Transformer to perform self-attention on visual features, which introduces extra information from external datasets. This extra information helps the model further understand the object and visual relation in the image. This group of models includes AOA [14], which extends the conventional attention mechanism to reweight the image features; X-Linear [13], which utilizes X-Linear attention module to extract fine-grained features of image; and DLCT [52], which fuses image grid features and image area features through cross attention module.

4.4 Comparison With Baselines

Table 1 shows the comparison between our model and the baseline models on Cross-Entropy Loss and CIDEr-D Score Optimization. For a fair comparison, all the models are firstly trained under cross-entropy loss and then optimized by the CIDEr-D score. It can be seen from Table 1 that our model consistently outperforms the baseline models in Cross-Entropy Loss training stage and is competitive with the state-of-the-art models in CIDEr-D Score Optimization. In particular, compared with the first and the second groups of methods that do not apply pre-trained Transformer, our ATD-Net performs the best in both cross-entropy loss training and CIDEr-D score optimization. At the cross-entropy loss training stage, ATD-Net increases the BLEU-4 and ROUGE scores to 38.5 and 58.1 respectively. At the CIDEr-D score optimization stage, ATD-Net still achieves the highest performance in BLEU@1-4 and ROUGE-L scores and has good competitiveness in CIDEr-D score with SG-RWS [51]. Compared with the third group of methods which employ pre-trained Transformer, our ATD-Net is still competitive in performance as it can fully extract the correct semantics in the raw caption. Especially, the results in Table 1 show that our ATD-Net still achieves the best results in BLEU@1-4 and ROUGE-L score at cross-entropy loss training stage and is competitive at CIDEr-D optimization stage. Moreover, our ATD-Net also exceeds the performance of some models that use visual self-attention performed by pre-trained Transformers, such as AOA [14].

In particular, ETN [7] is the previous state-of-the-art method of image caption editing and the raw caption it uses does not contain word noises artificially introduced. If we use the same raw caption with misdescriptive words in this article on ETN [7] model, all of these metrics of ETN will be lower than the values in Table 1, especially CIDEr will drop to around 116-117 under the cross-entropy loss training, slightly lower than our ATD-Net. Since our ATD-Net can filter out some noisy words that are not related to image semantics by TAM and Bidirectional Encoder, the generated captions will be more similar to the ground-truth sentences. Therefore, our method will perform better on the metrics that compare the similarity between generated captions and ground-truth captions such as BLEU and ROUGE-L.

4.5 Qualitative Analysis

To further test the denoising ability of our ATD-Net, we add different forms of noisy words to the raw caption and judge the accuracy of the model's restoration. We test the noise reduction performance of our model from the following two aspects: (1) The denoising capability on words with different parts of speech; (2) The denoising capability under different noise ratios.

Firstly, we classify the noisy words in the caption of the MS-COCO dataset into six categories: 1. Verb; 2. Noun (human or animal); 3. Noun (food); 4. Noun (location or place); 5. Noun (common objects); 6. Noun (color). We choose about eight words from each category as the noisy words. Then, to add noise into the caption, we randomly replace each word with different noisy words of the same category. In this paper, the noise ratio is set from 5% to 30%. For example, when adding 10% noise to the word "dog", we randomly take out 10% of the images that contain "dog" in the dataset, and randomly replace the corresponding word "dog" in the raw caption with other noisy words such as "cat" and "tiger". Finally, we record whether the captions generated by the model contain the exact words "dog". The accuracy is expressed as the percentage of pictures that our model can generate the correct words.

Table 2 demonstrates the accuracy of our ATD-Net on editing noisy words, which is composed of two parts: Average precision and Variance. The average precision indicates the probability of the target word that can be accurately generated under the current noise ratio. The Variance indicates the probability where the generated sentence does not contain the target word, but it can still correctly describe image semantics. For example, we replace the correct word "boy" in the raw caption with the misdescriptive word "girl", while the word generated by our ATD-Net is "person", which can still correctly describe the image content. From Table 2, we can observe that the proposed ATD-Net model has a good ability to edit captions with noisy words.

In addition, Fig. 6 shows the average precision of our ATD-Net for words with different noise ratios and different parts of speech through the line chart. Specifically, our model can correct the noisy words with a 100% accuracy under the noise ratio of 5%, and the accuracy declines as the noise

Table 2. The noise detection accuracy on the MS-COCO dataset. The noise words are divided into six categories, each category randomly selected about 8 words, and then tested the noise reduction ability of the model when the noise ratio is 5%, 15%, 30%.

(a)

Verb	Noise ratio			Human or Animal	Noise ratio		
	5%	15%	30%		5%	15%	30%
Stand	99.37±0.27	97.90±0.77	96.67±1.35	Boy	99.16±0.41	97.10±1.25	94.91±1.66
Hold	99.21±0.42	98.08±0.99	95.38±2.27	Girl	98.00±0	97.00±1.00	96.00±2.00
Rid	99.21±0.16	98.26±0.79	96.32±1.42	Person	100±0	97.86±0.31	96.63±0.92
Sit	98.93±0.17	98.17±0.68	95.65±2.06	Dog	100±0	98.26±0	97.68±0
Lay	98.36±0	97.39±0.29	96.77±0.83	Cat	99.49±0	98.98±0	97.96±0
Walk	99.30±0	96.15±1.05	92.30±2.10	Giraffe	100±0	100±0	98.92±0
Play	98.64±0	97.96±0	96.26±0.34	Elephant	100±0	97.62±0	96.43±0
Fly	99.29±0	97.51±0.35	96.45±0.71	Horse	100±0	99.12±0	98.24±0

(b)

Objects	Noise ratio			Place	Noise ratio		
	5%	15%	30%		5%	15%	30%
Computer	99.27±0	98.54±0	98.17±0.36	Bathroom	98.50±0	97.00±0	94.38±0.38
Table	99.11±0.08	95.72±1.21	90.63±2.42	Kitchen	100±0	97.50±0.35	95.38±0.36
Bed	100±0	98.37±1.04	95.60±1.73	Street	98.63±0.23	96.34±0.92	92.55±2.06
Chair	98.42±0	97.31±1.58	93.47±3.17	Building	97.90±0.70	96.15±1.05	91.60±2.10
Umbrella	100±0	100±0	96.56±1.15	Field	98.68±0.83	96.30±2.02	92.85±3.81
Airplane	100±0	96.82±1.58	92.84±2.37	Ocean	97.60±0	93.40±1.80	90.20±4.80
Bus	99.21±0	98.02±0.39	96.45±0.39	Grass	96.12±0	95.63±0.48	92.24±0.97
Train	98.98±0	97.96±0	96.94±0.51	Beach	98.77±0	96.31±0.61	94.79±0.92

(c)

Food	Noise ratio			Color	Noise ratio		
	5%	15%	30%		5%	15%	30%
Banana	100±0	99.26±0.24	99.02±0.49	Red	98.76±0.54	97.72±2.59	94.09±5.22
Pizza	100±0	100±0	99.17±0.83	Yellow	98.75±1.25	96.25±2.50	90.00±5.82
Vegetable	98.80±1.21	96.39±3.62	92.19±5.41	White	99.18±0.41	96.71±2.05	91.99±5.13
Cake	100±0	100±0	98.71±0	Brown	100±0	97.50±2.50	91.00±5.83
Water	99.48±0.52	98.96±0.52	95.84±2.07	Green	99.28±0.71	95.43±3.57	89.57±7.14
Food	100±0	97.22±0	96.10±0	Black	98.60±0.85	95.21±3.10	90.42±6.21
Sandwich	100±0	96.83±0	92.88±0.79	Blue	99.35±0.64	96.77±3.23	92.24±6.46

ratio increases. However, even with the noise ratio of 30%, our model still has over 90% probability to generate caption with accurate words.

4.6 Ablation Studies

To prove the effectiveness of the Textual Attention Mechanism (TAM) and Bidirectional Encoder in our ATD-Net and to provide more detailed parameter analysis, we conduct extensive ablation studies on the MS-COCO dataset. The results of the component analysis are reported in Table 3.

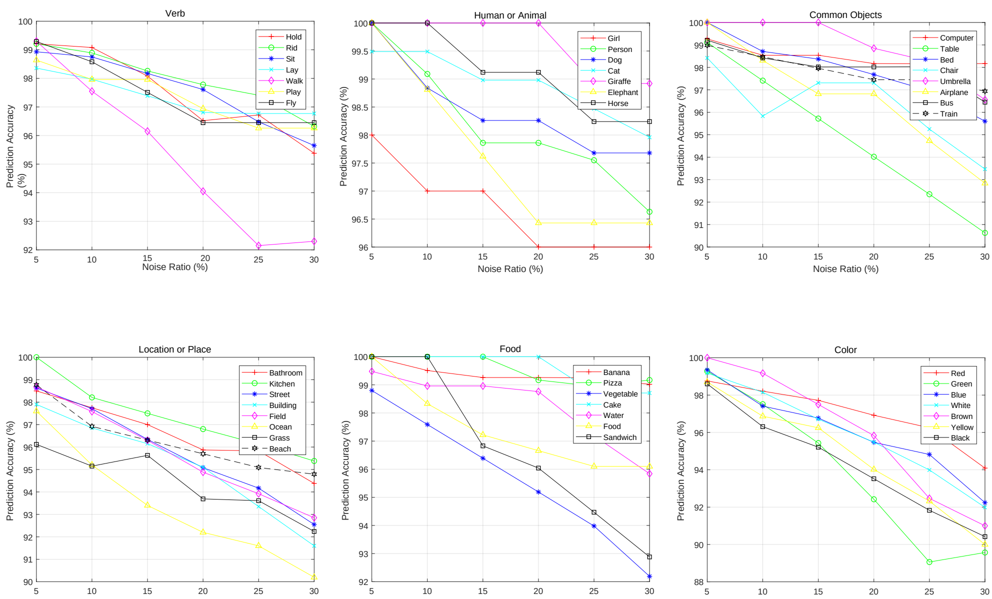

Fig. 6. The average precision of our ATD-Net for words with different noise rates and different parts of speech.

Table 3. Ablation studies on the MS-COCO dataset about essential components of ATD-Net

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDEr
Baseline	77.2	61.5	47.3	36.2	56.4	113.5
+TAM	77.6	62.8	48.8	38.0	57.5	116.7
+LSTM	78.0	63.0	49.4	38.3	58.0	117.9
+BiLSTM	**78.2**	**63.1**	**49.5**	**38.5**	**58.1**	**118.5**

In the ablation experiments, we remove the whole Caption Encoder of ATD-Net and use the decoder only with visual attention as the baseline in the ablation experiment. We first add Caption Encoder with Caption Precoding and Textual Attention Mechanism to the baseline, then each word embedding produced by Textual Attention Mechanism is directly summed and averaged as the input of language LSTM in the decoder. From Table 3, it shows that after the introduction of Textual Attention Mechanism, our ATD-Net can achieve higher performance.

TAM can achieve rich text semantics from the raw caption. To fully extract this semantic information at the sentence level, we have made two different attempts. Specifically, we try to use forward direction coding based on LSTM and bidirectional direction coding based on Bi-LSTM respectively. The experimental results show that both methods can improve the performance of the model, but the Bidirectional Encoder can better call the context semantic information to make the performance of the model reach the best. Moreover, it can be seen that ATD-Net can achieve

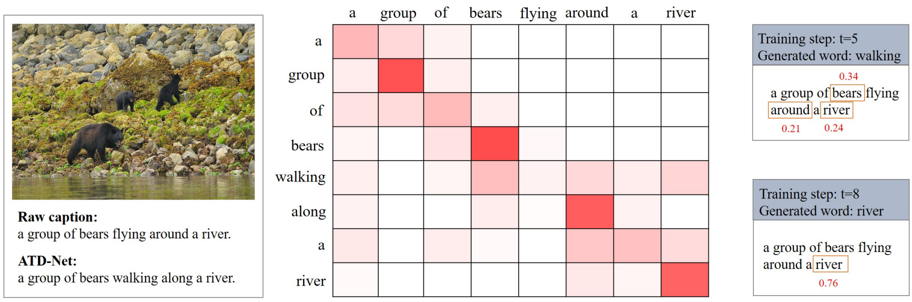

Fig. 7. Weight parameter visualization of Textual Attention Mechanism (TAM) in our ATD-Net. The x-axis and y-axis correspond to the words in the raw caption and the generated caption, respectively. Darker color illustrates higher attentional weight. The red number indicates the weight value assigned by TAM to each word in the raw caption.

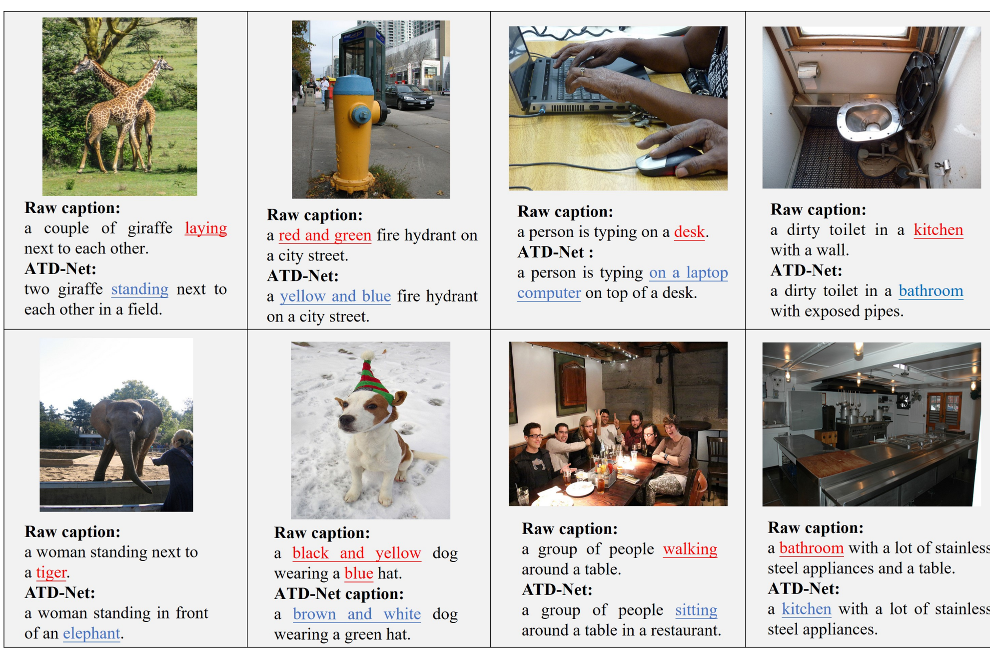

Fig. 8. Examples of captions generated by our ATD-Net and raw captions with noisy words. The words in red are misdescriptive words and the words in blue are corrected words generated by ATD-Net.

certain stability and higher performance after the introduction of Bidirectional Encoder, which reflects the robustness of ATD-Net.

Fig. 7 shows an example of Textual Attention Mechanism (TAM) weight parameter visualization. At each training step, TAM will attach different weights to each word of the raw caption. In this example, the corresponding matrix is the correlation weights of the textual attention between words in the generated caption and words in the raw caption. Moreover, we mark the words with the highest weight values at each iteration and the weight value of the word that is not marked in this example is relatively low. It can be found that, when the words of the raw caption have a high similarity with the image semantics, TAM will attach a higher weight to the corresponding word in the raw caption, such as generating the word “bears” and “river”. Meanwhile, if the semantics of the raw caption corresponding to the word to be generated is incorrect, TAM will attach a low weight to the incorrect word of the raw caption. For example, when generating the word “walking", TAM will pay more attention to the words “bear”, “around” and “river”, and attach a lower weight to the misdescriptive word “flying”.

In Fig. 8, we show some qualitative examples of the captions generated by our ATD-Net on the dataset of MS-COCO. It shows that our model is able to edit the misdescriptive words in the raw caption well and finally, generate an accurate image caption.

5 CONCLUSION

In this paper, we propose ATD-Net, a novel encoder-decoder based architecture for image caption editing. This framework solves the problem of the inconsistency between the semantics of image and caption from word level and sentence level respectively. Specifically, (1) Textual Attention Mechanism (TAM) is designed to locate and minimize noises at word level. (2) Bidirectional Encoder is designed for robust caption encoding at sentence level. Experiments on the MS-COCO dataset show that ATD-Net achieves better performance in the metric of BLEU and ROUGE. Moreover, ATD-Net can reach an accuracy rate of more than 90% in the editing accuracy of various parts of speech word noises. In the future, our framework may be extended to related tasks such as video caption editing. Specifically, we can use the video frame as a guide to correct the text semantic noises in the raw caption through a cross-modal attention mechanism similar to TAM in ATD-Net.

ACKNOWLEDGMENTS

This work was supported by National Key Research and Development Project (No.2020AAA0106200), the National Nature Science Foundation of China under Grants (No.61936005, 61872424), and the Natural Science Foundation of Jiangsu Province (Grants No.BK20200037 and BK20210595).

REFERENCES

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3156–3164.
[2] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” *IEEE Trans. Pattern Anal.Mach. Intell.*, 2017, vol. 39, no. 4, pp. 664–676.
[3] K. Xu et al., “Show, attend and tell: Neural image caption generation with visual attention,” *in Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
[4] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, “Self-critical sequence training for image captioning,” in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1179–1195.
[5] P. Anderson et al., “Bottom-up and top-down attention for image captioning and visual question answering,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.
[6] S. Fawaz and E. Mahmoud, “Look and Modify: Modification Networks for Image Captioning,” in *British Machine Vision Conference.*, 2019, pp. 75.
[7] S. Fawaz and E. Mahmoud, “Show, Edit and Tell: A Framework for Editing Image Captions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4807-4815.
[8] L. Wu, M. Xu, J. Wang and S. Perry, “Recall What You See Continually Using GridLSTM in Image Captioning,” *IEEE Transactions on Multimedia*, 2020, pp. 808-818.
[9] L. Guo, J. Liu, S. Lu and H. Lu, “Show, Tell and Polish: Ruminant Decoding for Image Captioning,” *IEEE Transactions on Multimedia*, 2019, pp. 2149-2162.
[10] L. Guo, J. Liu, X. Zhu, P. Yao, S. Lu and H. Lu, “Normalized and Geometry-Aware Self-Attention Network for Image Captioning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10324-10333.
[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, N. A. Gomez, L. Kaiser and I. Polosukhin, “Attention Is All You Need,” in *roc. Neural Inf. Process. Syst. Conf.*, 2017, pp. 5998-6008.
[12] T. Yao, Y. Pan, Y. Li and T. Mei, “Exploring Visual Relationship for Image Captioning,” in *Proc. Eur. Conf. Comput. Vis.*, 2019, pp. 711-727.

[13] Y. Pan, T. Yao, Y. Li and T. Mei, “X-Linear Attention Networks for Image Captioning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10968-0977.
[14] L. Huang, W. Wang, J. Chen and X. Wei, “Attention on Attention for Image Captioning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4633-4642.
[15] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao and T. S. Chua, “SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6298-6306.
[16] J. Lu, C. Xiong, P. Devi and S. Richard. Knowing, “When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3242-3250.
[17] I. Sutskever, O. Vinyals and V. Q. Le, “Sequence to Sequence Learning with Neural Networks,” in *Proc. Neural Inf. Process. Syst. Conf.*, 2014, pp. 3104-3112.
[18] D. Bahdanau, K. Cho and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” in *Proc. 3rd Int. Conf. Learn. Representations,* San Diego, CA, USA, 2015.
[19] J. Ji, C. Xu, X. Zhang, B. Wang and X. Song, “Spatio-temporal Memory Attention for Image Captioning,” *IEEE Transactions on Image Processing*, 2020, pp. 7615-7628.
[20] Y. Huang, J. Chen, W. Ouyang, W. Wan and Y. Xue, “Image Captioning with End-to-end Attribute Detection and Subsequent Attributes Prediction,” *IEEE Transactions on Image Processing*, 2020, pp. 4013-4026.
[21] N. Xu, H. Zhang, A. Liu, W. Nie, Y. Su, J. Nie and Y. Zhang, “Multi-Level Policy and Reward-Based Deep Reinforcement Learning Framework for Image Captioning,” *IEEE Transactions on Multimedia*, 2020, pp. 1372-1383.
[22] S. Chen, Q. Jin, P. Wang and Q. Wu, “Say As You Wish: Fine-grained Control of Image Caption Generation with bstract Scene Graphs,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9959-9968.
[23] L. Zhou, Y. Zhang, Y. Jiang, T. Zhang and W. Fan, “Re-Caption: Saliency-Enhanced Image Captioning through Two-Phase Learning,” *IEEE Transactions on Image Processing.*, 2019, pp. 694-709.
[24] M. Cornia, M. Stefanini, L. Baraldi and R. Cucchiara, “Meshed-Memory Transformer for Image Captioning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10575-10584.
[25] X. Yang, K. Tang, H. Zhang and J. Cai, “Auto-Encoding Scene Graphs for Image Captioning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10685-10694.
[26] Z. Zha, D. Liu, H. Zhang, Y. Zhang and F. Wu, “Context-Aware Visual Policy Network for Fine-Grained Image Captioning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
[27] L. Guo, J. Liu, P. Yao, J. Li and H. Lu, “MSCap - Multi-Style Image Captioning With Unpaired Stylized Text,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4204-4213.
[28] J. Devlin, M. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *The North American Chapter of the Association for Computational Linguistics*, 2018, pp. 4171-4186.
[29] R. Soricut, N. Ding, P. Sharma and S. Goodman, “Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning,” in *Proc. Assoc. Comput. Linguistics.*, 2018, pp. 2556-2565.
[30] K. He, X. Zhang, S. Ren and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770-778.
[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, S. M. Bernstein, C. A. Berg and F. Li, “ImageNet Large Scale Visual Recognition Challenge,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 221-252.
[32] K. Andrej and F. Li, “Deep visual-semantic alignments for generating image descriptions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 664-676.
[33] P. Kishore, R. Salim, W. Todd, and W. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proc. Assoc. Comput. Linguistics.*, 2002, pp. 311-318.
[34] C. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Proc. Text Summarization Branches Out*, 2004, pp. 1-8.
[35] R.Vedantam, C. L. Zitnick, and D. Parikh, “Cider:Consensus-based image description evaluation,” in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4566–4575.
[36] R. Krishna and Y. Zhu et al, “Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations,” in *International Journal of Computer Vision*, 2017, pp. 32-73.
[37] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. 3rd Int. Conf. Learn. Representations.*, San Diego, CA, USA, May 7–9, 2015.
[38] W. Jiang, L. Ma, Y. Jiang, W. Liu and T. Zhang, “Recurrent Fusion Network for Image Captioning,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 510-526.
[39] X. Yang, K. Tang, H. Zhang and J. Cai, “Auto-encoding scene graphs for image captioning,” in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10685-10694.
[40] C. Kyunghyun, V. M. Bart, G. Çaglar, B. Fethi, S. Holger and B. Yoshua, “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014,

pp. 1724-1734.

[41] B. Dzmitry, C. Kyunghyun and B. Yoshua, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Representations.*, San Diego, CA, USA, 2015.

[42] J. Sébastien, C. KyungHyun, M. Roland and B. Yoshua, "On using very large target vocabulary for neural machine translation," in *Proc. Assoc. Comput. Linguistics.*, 2015, pp. 1-10.

[43] L. Thang, P. Hieu and D. M. Christopher, "Effective Approaches to Attention-based Neural Machine Translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1412-1421.

[44] M. Eric, K. Sebastian, R. Sascha, M. Daniil and S. Aliaksei, "Encode, Tag, Realize: High-Precision Text Editing," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2019, pp. 5053-5064.

[45] N. Shi, Z. Zeng, H. Zhang and Y. Gong, "Recurrent Inference in Text Editing," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 1758-1759.

[46] Y. Liao, L. Bing, P. Li, S. Shi, W. Lam and T. Zhang, "QuaSE: Sequence Editing under Quantifiable Guidance," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 3855-3864.

[47] L. Wang, W. Zhao, R. Jia, S. Li and J. Liu, "Denoising based Sequence-to-Sequence Pre-training for Text Generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2019, pp. 4001-4013.

[48] T. Lin, M. Michael, J. B. Serge, D. B. Lubomir, B. G. Ross, H. James, P. Pietro, R. Deva, D. Piotr, and Z. C. Lawrence, "Microsoft COCO: Common objects in context," in *Proc.Eur. Conf. Comput. Vis,* 2014, pp. 740–755.

[49] L. Huang, W. Wang, Y. Xia and J. Chen, "Adaptively Aligned Image Captioning via Adaptive Attention Time," in *n Proc. Adv. Neural Inf. Process.Syst.*, 2019, pp. 8940-8949.

[50] Y. Qin, J. Du, Y. Zhang and H. Lu, "Look Back and Predict Forward in Image Captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8367-8375.

[51] L. Wang, Z. Bai, Y. Zhang and H. Lu, "Show, Recall, and Tell: Image Captioning with Recall Mechanism," in *Proc. 45nd AAAI Conf*, 2020, pp. 12176-12183.

[52] Y. Luo, J. Ji, X. Sun, L. Cao, Y. Wu, F. Huang, C. Lin and R. Ji, "Dual-Level Collaborative Transformer for Image Captioning," in *Proc. 35nd AAAI Conf*, 2021, pp. 2286-2293.

[53] A. Liu, Y. Wang, N. Xu, W. Nie, J. Nie and Y. Zhang, "Adaptively Clustering-Driven Learning for Visual Relationship Detection," *IEEE Transactions on Multimedia*, 2020

[54] M. Tao, H. Tang, F. Wu, X. Jing, B. Bao and C. Xu, "DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis." in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2022.

[55] J. Wang, B. Bao and C. Xu, "DualVGR: A Dual-Visual Graph Reasoning Unit for Video Question Answering." *IEEE Transactions on Multimedia.*, 2021, vol. 14, no. 8.

[56] L. Lo, H. Xie, H. Shuai and W. Cheng, "Facial Chirality: Using Self-face Reflection to Learn Discriminative Features for Facial Expression Recognition." in *IEEE International Conference on Multimedia Expo.*, Shenzhen, China, 5-9 July, 2021.

[57] H. Xie, L. Lo, H. Shuai and W. Cheng, "AU-assisted Graph Attention Convolutional Network for Micro-Expression Recognition." in *ACM International Conference on Multimedia.*, Seattle, USA, 12-16 October, 2020.

[58] C. Chen, L. Lo, P. Huang, H. Shuai and W. Cheng, "FashionMirror: Co-attention Feature-remapping Virtual Try-on with Sequential Template Poses," in *IEEE International Conference on Computer Vision.*, Montreal, Canada, 10-17 October, 2021.