applied sciences

MDPI

Article

Separate Syntax and Semantics: Part-of-Speech-Guided Transformer for Image Captioning

Dong Wang [1,2], Bing Liu [1,2,*], Yong Zhou [1,2], Mingming Liu [1,2], Peng Liu [3] and Rui Yao [1,2]

[1] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China
[2] Engineering Research Center of Mine Digitization, Ministry of Education of the People's Republic of China, Xuzhou 221116, China
[3] National Joint Engineering Laboratory of Internet Applied Technology of Mines, Xuzhou 221008, China
* Correspondence: liubing@cumt.edu.cn

Abstract: Transformer-based image captioning models have recently achieved remarkable performance by using new fully attentive paradigms. However, existing models generally follow the conventional language model of predicting the next word conditioned on the visual features and partially generated words. They treat the predictions of visual and nonvisual words equally and usually tend to produce generic captions. To address these issues, we propose a novel part-of-speech-guided transformer (PoS-Transformer) framework for image captioning. Specifically, a self-attention part-of-speech prediction network is first presented to model the part-of-speech tag sequences for the corresponding image captions. Then, different attention mechanisms are constructed for the decoder to guide the caption generation by using the part-of-speech information. Benefiting from the part-of-speech guiding mechanisms, the proposed framework not only adaptively adjusts the weights between visual features and language signals for the word prediction, but also facilitates the generation of more fine-grained and grounded captions. Finally, a multitask learning is introduced to train the whole PoS-Transformer network in an end-to-end manner. Our model was trained and tested on the MSCOCO and Flickr30k datasets with the experimental evaluation standard CIDEr scores of 1.299 and 0.612, respectively. The qualitative experimental results indicated that the captions generated by our method conformed to the grammatical rules better.

Keywords: image captioning; transformer; part of speech; multitask learning

Citation: Wang, D.; Liu, B.; Zhou, Y.; Liu, M.; Liu, P.; Yao, R. Separate Syntax and Semantics: Part-of-Speech-Guided Transformer for Image Captioning. *Appl. Sci.* **2022**, *12*, 11875. https://doi.org/10.3390/app122311875

Academic Editor: Silvia Liberata Ullo

Received: 17 October 2022
Accepted: 16 November 2022
Published: 22 November 2022

1. Introduction

Image captioning is the task of generating the grammatically correct description of an image, which has been attracting much attention in the field of image understanding [1–8]. With the success of deep learning, image captioning models have recently achieved great progress. A typical deep neural network for an image captioning model generally follows an encoder–decoder paradigm, where a deep convolutional neural network (CNN) is introduced as the encoder to learn visual representations from the input image, while a recurrent neural network (RNN) serves as the decoder to recursively predict each word. Recently, the transformer-based image captioning models have shown superior performance to the conventional CNN-RNN models by using fully attentive paradigms. Despite great advances made in the model architectures, existing models still have two limitations: (i) they treat the predictions of visual and nonvisual words equally at each time step, leading to ambiguous inference; (ii) they have the tendency to generate minimal sentences, which is common in datasets. Consequently, how to organize phrases and words to accurately express the semantics of an image remains a challenging task.

The neuroscience research on language processing has demonstrated that the brain contains partially separate systems for processing syntax and semantics [9,10], which

provides us a new prospective to overcome the limitations of existing image captioning models. Naturally, the traditional encoder–decoder framework can be improved by imposing an analogous separation. Considering that in English the part-of-speech (PoS) tag sequences contain rich grammatical rules available to infer the corresponding words (We use the Stanford constituency parser to obtain the PoS tags of captions. URL: https://www.nltk.org/book/ch05.html (accessed on 15 November 2022), in this paper, we intend to improve the grounding performance of image captioning by using the PoS information. Figure 1a illustrates an example of an image caption with its corresponding PoS tags. From Figure 1a, we can observe that the different parts of speech of the words play specific grammatical roles in the caption. For example, the determiners (DET) and adjectives (ADJ) are generally used to modify the nouns (NN). The adpositions (ADP), such as *in* and *on*, play the role of connecting two noun phrases so as to establish their semantic relationship. All the PoS tags play an important role in generating the caption since they correspond to words one by one. Consequently, it is essential to master the PoS of each word for generating grammatically correct sentences. Besides, some PoS tags, such as ADJ and NOUN are closely related to the visual features of the image while some PoS tags, such as the second ADP (corresponding to the word *on*) in the PoS tag sequence, are irrelevant to any visual features. As a result, there is a need to find more ways to highlight the PoS information contained in sentences so that they can provide additional guidance for one captioner to distinguish between visual and nonvisual words.

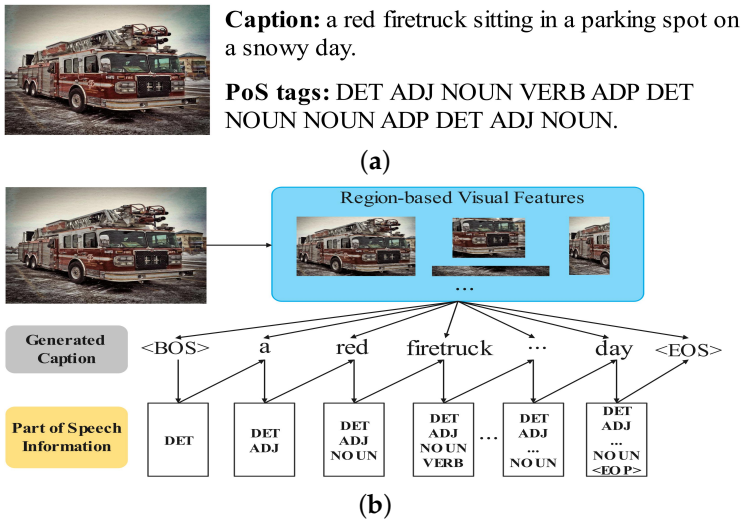

Figure 1. An example of PoS-guided caption generation. The PoS and word information are maintained in separate streams. (**a**) The PoS tags and the corresponding words in an image description sentence. (**b**) Our model first predicts the most appropriate subsequent PoS by the previous words at each time step. Then, the obtained PoS information is used to guide the visual and linguistic attention for the word prediction. "<BOS>" and "<EOS>" denote the beginning and end of all the sentences, respectively. "<EOP>" is short for "<End of PoS>", which is the end of all the PoS sequences.

Aiming to obtain the syntactic information contained in the sequence of PoS tags, we first introduce a PoS predictor to predict the PoS tag of the next word, which can be integrated with the image captioning model seamlessly. As shown in Figure 1b, the PoS tag of the next word is predicted based on the previous words while the PoS information provided by the PoS predictor is utilized to guide the generation of the next word. For instance, after the words *a* and *red* as well as their PoS tags are generated, the PoS predictor uses the word embeddings of *a* and *red* as inputs to predict the PoS tag NOUN. Meanwhile, the PoS information of DET, ADJ, and NOUN are utilized by the image caption model to predict the next word *firetruck*. Unlike the existing transformer-based captioners that treat all word predictions equally, the sequence of partially generated tags can help evaluate the effect of visual features and language signals on the word prediction. As illustrated in Figure 1a,

when the word *on* is to be generated, the visual features are actually not very helpful at the current time step. However, the conventional transformer-based image captioners take no effective measures but simply concatenate attended visual features and language signals in each decoder layer, i.e., the irrelevant visual features are also used to predict the next word. As a result, the captioners are easily distracted by irrelevant visual concepts, leading to the generation of incorrect words. In contrast, after the partial PoS tag ADP of the next word *on* is available, the captioners can exploit the information of partially generated PoS tags to balance the effect of visual features and language context, e.g., the language cues would be paid more attention to at the current time step, which facilitates the generation of the correct word *on*.

In order to make a transformer-based image captioning model effectively align the generated words with the visual or nonvisual features of an image and further generate the grammatically correct captions with the help of the PoS information, we propose a PoS-Transformer framework based on a new learning paradigm. Specifically, the process of generating captions is divided into two stages: PoS prediction and caption generation. The PoS tag of the next word is predicted in the first stage, which is much easier than predicting the next word directly, since the number of PoS tags is far less than that of words. In the second stage, two different PoS-guided attention modules are proposed on top of the PoS guiding information, visual features, and linguistic context, which enables the decoder to adaptively attend to visual features and language signals. As a result, the PoS predictor, the PoS-guided attention modules, and the encoder–decoder captioning network closely collaborate to enhance the performance of image captioning. The main contributions of our work can be summarized as follows:

- We propose two kinds of PoS-guided attention mechanisms based on the PoS information, adaptively adjusting the effect of visual features and language signals on the word prediction, to encourage the generation of more grounded captions.
- We incorporate the PoS prediction model and the PoS-guided attention modules into the transformer-based captioning architecture to build a unified end-to-end image captioning framework, boosting the performance of image captioning by separating syntax and semantics for the prediction of each word.
- We optimize the proposed PoS-Transformer network by a multitask learning method on the Flickr30k and MSCOCO benchmark datasets, respectively. Extensive experiments demonstrate the effectiveness of our method.

The remainder of this paper is organized as follows. Section 2 introduces the related work, especially the prevailing deep-learning-based methods. Our proposed framework and its multitask learning for image captioning are detailed in Section 3. The experimental results are reported in Section 4. Finally, Section 5 concludes the paper.

2. Related Work

Image captioning. The mainstream image captioning methods generally follow the encoder–decoder paradigm, where image features extracted by a CNN are fed into an RNN to generate the corresponding sentence. For example, Xu et al. [11] first utilized soft and hard attention mechanisms to attend to the different CNN grid features of an image when generating each word. Lu et al. [12] presented an adaptive attention mechanism to determine where to attend to visual features for the word prediction. After that, Anderson et al. [13] further introduced an attention mechanism over the region-based features extracted by an object detector. Despite progress made on the basis of visual attention mechanism over object features, these approaches suffer from catastrophic forgetting in long-term memory, leading to limited performance improvement. To overcome the limitations of RNN-based image captioning models, plenty of transformer-based models [14–22], following fully attentive paradigms, have recently been presented and have improved the performance remarkably. For example, Herdade et al. [15] developed an object relation transformer (ORT) captioning model, which explicitly incorporated spatial relationships between region features through geometric attention. Li et al. [23] introduced entangled

attention into a transformer-based sequence modeling framework that performs attention over visual features and semantic attributes simultaneously. Recently, a large amount of methods have been explored to improve image understanding with the help of a scene graph, as it contains rich semantic information. For example, Yang et al. [24] proposed a method that first used the sentence's scene graph to learn a dictionary, and then incorporated it with the image's scene graph for the description generation. Yao et al. [25] presented a model that integrated both the semantic and spatial object relationships as image representation. Since the scene graph constructed a series of semantic relationship information, the model achieved comparable results. Zhao et al. [26] proposed a multilevel cross-modal alignment (MCA) module to align the image scene graph with the sentence's scene graph at a different level. Although the existing captioning approaches have achieved impressive results, they still follow the conventional way of modeling language and suffer from the limitations mentioned above.

PoS-based image captioning. Recently, some works have also introduced the PoS information into image captioning models [27–29]. However, these methods are all based on long short-term memory (LSTM) networks, while our model exploits the transformer-based captioning architecture and fully attentive paradigm, which is essentially different from them. The model proposed by Zhang et al. [27] is the most related to ours; they integrated the PoS information with two popular image captioning models. However, their models suffered from dependencies between distant positions since the hidden states of LSTM were used to predict the PoS sequences. He et al. [28] utilized PoS tags as switches to guide the generation of the visual words. However, they required an external PoS tagger in both the training and test stages, which was limited in practice. In our PoS-Transformer, a PoS prediction network, as a part of the framework, is seamlessly integrated with other parts of PoS-Transformer. Consequently, the captions can be generated word by word at the inference time without any extra PoS taggers. Deshpande et al. [29] used the part-of-speech information to generate diverse captions. They first predicted a PoS sequence for an image and then employed the PoS sequence as the guiding information to generate image captions. However, they quantized the space of POS tag sequences by using a classification model, which harmed the generation of fine-grained captions. Unlike existing PoS-based image captioning models, our proposed PoS-Transformer framework is able to process both word sequences and PoS sequences in parallel during training. On one hand, by means of cross-attention, PoS-Transformer establishes the relationship between the visual features and PoS information as well as the relationship between the partially generated words and PoS information. On the other hand, PoS-Transformer also captures the self-attention within the PoS information, which is helpful to adaptively adjust the weights between visual features and language signals for the word prediction.

3. Approach

The proposed PoS-Transformer model aims to guide the process of caption generation with the part-of-speech information on top of the Transformer architecture. Notably, our method follows a novel learning paradigm, which maintains the PoS and word information in separate streams for image captioning. Specifically, PoS-Transformer is composed of four parts: (1) a visual subencoder that exploits the deep visual representation on the basis of a self-attention mechanism; (2) a language subencoder that represents language signals; (3) a self-attention PoS predictor (SAPP) which is used to predict the category of PoS and obtain the PoS information for generating the next word in the captioning process; (4) a PoS-guided multimodal decoder which provides two alternative attention mechanisms, i.e., single attention (SAT) and dual attention (DAT), to integrate and decode visual features, language signals, and PoS information. Figure 2 illustrates the overall architecture of the proposed PoS-Transformer model.

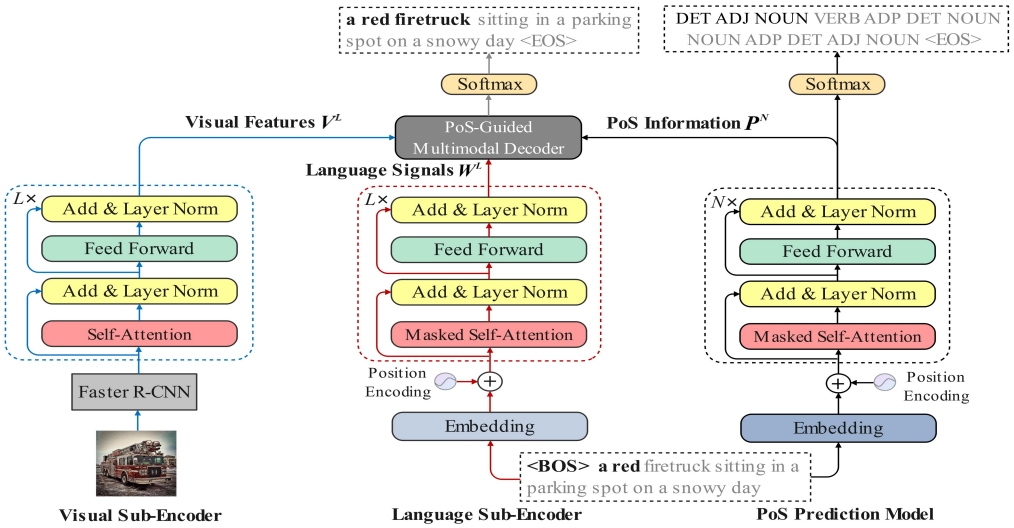

Figure 2. The overall architecture of PoS-Transformer for image captioning. Our framework consists of four parts: the visual subencoder, the language subencoder, the PoS predictor, and the PoS-guided multimodal decoder. The captioning generation includes three steps: (1) encoding the visual features and language signals separately; (2) obtaining the PoS information by predicting the subsequent PoS; and (3) generating the final caption with PoS guidance to the usage of visual and language cues. Our PoS-Transformer can dynamically adjust the weights between visual features and language signals to produce more fine-grained sentences.

3.1. Dual-Way Encoder

Different from the local operator essence of convolution [3,30], the full transformer captioning networks, effectively accessing information globally via self-attention mechanism, have recently been proposed and achieved promising performance. However, the existing transformer-based captioning architectures are still based on the conventional language model, which generates the captions word by word regardless of the grammatical structures, leading to the limitations mentioned above. Consequently, it is essential to construct a novel image captioning architecture, which not only separates syntactic structure and word semantics, but has the ability to guide the usage of visual and language information. To reach this goal, inspired by the ETA model [23], we first propose a dual-way encoder that contains a visual subencoder and a language subencoder to obtain the visual features and language signals attended to, respectively.

(1) Visual subencoder: In Figure 3, the region-based visual features of an image extracted by a pretrained Faster-RCNN model are utilized as the input of visual subencoder. Given a set of region-based visual features $V = \{v_1, v_2, \ldots, v_N\}$ extracted from an input image, where N is the number of visual regions in an image, the visual features V are first projected to a d-dimensional space via a fully connected layer to adapt to the visual subencoder's dimensionality. Then, the projected features $V^0 = \{v_1^0, v_2^0, \ldots, v_N^0\} \in \mathbb{R}^{N \times d}$ are input into the visual subencoder with L attention blocks. To be specific, the output of the lth ($0 \leq l < L$) layer is input into a multihead module (MH) [31] in the $(l+1)$th layer, which is then followed by an AddNorm operation:

$$\hat{V}^{l+1} = AddNorm(MH(V^l, V^l, V^l)), \tag{1}$$

and a positionwise feed-forward network (FFN) [31] is adopted to further transform the outputs, which is also encapsulated within the AddNorm operation:

$$V^{l+1} = AddNorm(FFN(\hat{V}^{l+1})). \tag{2}$$

Eventually, we can obtain V^L, i.e., the output of our visual subencoder, which represents the considered visual features, on basis of the self-attention mechanism.

(2) Language subencoder: Given a caption $Y = \{Y_1, Y_2, \ldots, Y_M\}$, where Y_i denotes the ith word in the sentence and M is the number of words. To adapt the language subencoder's dimensionality, all tokens are first embedded to d-dimensional vectors through an embedding matrix and then fed into the positional encoding module for the relative and absolute position information. Finally, we obtain the initial input features $W^0 = \{w_1^0, w_2^0, \ldots, w_M^0\} \in \mathbb{R}^{M \times d}$, which are input to the language subencoder with L attention blocks. Different from the visual subencoder, the output of the lth ($0 \leq l < L$) layer is passed into the masked multihead (MMH) module [31] to ensure that the prediction for the tth word w_t depends only on the previous words $w_{1:t-1}$, and the output of the $(l+1)$th layer is denoted as follows:

$$\begin{aligned} \hat{W}^{l+1} &= AddNorm(MMH(W^l, W^l, W^l)), \\ W^{l+1} &= AddNorm(FFN(\hat{W}^{l+1})). \end{aligned} \tag{3}$$

Recursively, the output of the Lth layer, denoted as W^L, can be obtained and used as the language signals to be fed into the following decoder.

3.2. Self-Attention PoS Predictor

In the self-attention PoS predictor, we also use a randomly initialized word-embedding matrix and positional encoding to project the input tokens $Y = \{Y_1, Y_2, \ldots, Y_M\}$ to d-dimensional vectors $P^0 = \{p_1^0, p_2^0, \ldots, p_M^0\} \in \mathbb{R}^{M \times d}$. The PoS prediction model takes the projected features P^0 as the initial input to N, the first attention block. Similar to the language subencoder, the output of the $(n+1)$th layer can be represented as:

$$\begin{aligned} \hat{P}^{n+1} &= AddNorm(MMH(P^n, P^n, P^n)), \\ P^{n+1} &= AddNorm(FFN(\hat{P}^{n+1})). \end{aligned} \tag{4}$$

Finally, the output of the Nth decoder stack is used as the PoS information to predict the probability distribution of the next word's PoS as follows:

$$p(s_t | Y_{t-1}) = Softmax(W_{PoS} \cdot P_{t-1}^N + b_{PoS}), \tag{5}$$

where P_{t-1}^N denotes the hidden state corresponding to the $(t-1)$th PoS, the embedded matrix $W_{PoS} \in \mathbb{R}^{d \times \mathcal{C}}$, the bias vector $b_{PoS} \in \mathbb{R}^{\mathcal{C}}$, Y_{t-1} denotes the previously generated words, and $\mathcal{C}$ is the class number of PoS. Meanwhile, as shown in Figure 3, the PoS information P_{t-1}^N is then passed to the PoS-guided multimodal decoder to guide the caption generation.

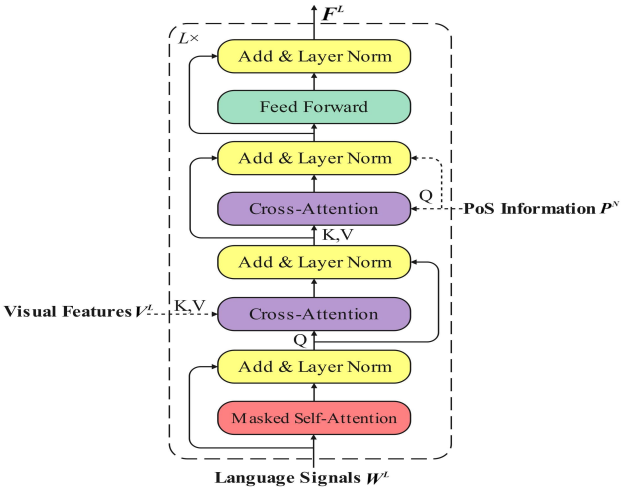

Figure 3. The structure of the single-attention-based multimodal decoder model with part-of-speech guidance.

3.3. PoS-Guided Multimodal Decoder

(1) *PoS-guided single attention:* Different from the traditional transformer decoder, we introduce a single cross-attention over the fused features of visual features and language signals by virtue of the PoS information.

As shown in Figure 3, for the $(l+1)$th layer, the input F^l is fed into an MMH module, followed by the AddNorm operation:

$$\hat{F}^{l+1} = AddNorm(MMH(F^l, F^l, F^l)). \tag{6}$$

Note that $F^0 = W^L$. Subsequently, the output $\hat{F}^{l+1}$ is fed into one multihead cross-attention module to perform the attention task over visual features V^L as follows:

$$\begin{aligned} \bar{F}_V^{l+1} &= MH(\hat{F}^{l+1}, V^L, V^L), \\ \bar{F}^{l+1} &= AddNorm(\bar{F}_V^{l+1}). \end{aligned} \tag{7}$$

Since the PoS information is beneficial for both visual words and nonvisual words, it is used to attend to the fused features of visual features and language signals during training. Meanwhile, it is also added to the considered fused features, to provide the decoder with the PoS information. To be specific, we utilize the PoS information P^N as the query vectors to perform the cross-attentions over $\bar{F}^{l+1}$ as follows:

$$\tilde{F}^{l+1} = AddNorm(MH(P^N, \bar{F}^{l+1}, \bar{F}^{l+1}), P^N). \tag{8}$$

Finally, the output of the multimodal decoder can be obtained as follows:

$$F^{l+1} = AddNorm(FFN(\tilde{F}^{l+1})). \tag{9}$$

(2) *PoS-guided dual attention:* Although the single attention mechanism utilizes the POS information to facilitate the generation of grounded captions, it cannot adaptively adjust the weights between visual features and language signals at each decoding time step. Inspired by the ETA model [23], we first introduce the dual attention mechanism into the multimodal decoder, which employs the PoS information to attend to the visual features and language signals, respectively. In addition, a gated controller module is inserted after the dual attention module, which enables the decoder to dynamically adjust the weights between the visual features and language signals.

As depicted in Figure 4, the dual attention module is inserted between the MMH and FFN modules, which allows the decoder block to apply attention over the output visual features V^L and language signals W^L of the dual-way encoder simultaneously. Similar to the single attention, we have:

$$\hat{F}^{l+1} = AddNorm(MMH(F^l, F^l, F^l)). \tag{10}$$

where $F^0 = P^N$. Then, the output $\hat{F}^{l+1}$ is passed into two multihead cross-attention modules to perform the attention task over language signals W^L and visual features V^L:

$$\begin{aligned} V^{l+1} &= MH(\hat{F}^{l+1}, V^L, V^L), \\ S^{l+1} &= MH(\hat{F}^{l+1}, W^L, W^L). \end{aligned} \tag{11}$$

Next, as shown in Figure 4, the gated controller module is introduced into the decoder to dynamically specify the weights of S^{l+1} and V^{l+1} on the word prediction. Concretely, the context gate C^{l+1} of the gated controller is determined by the visual features V^{l+1}, the language signals S^{l+1}, and the current self-attention output $\hat{F}^{l+1}$:

$$C^{l+1} = \sigma([V^{l+1}, \hat{F}^{l+1}, S^{l+1}] \cdot W_C), \tag{12}$$

where $C^{l+1} \in \mathbb{R}^{M\times 1}$, $W_C \in \mathbb{R}^{3d\times 1}$, $[\cdot]$ and $\sigma(\cdot)$ denote the vector concatenation and sigmoid function, respectively. The gate value C^{l+1} and its complement part $(1 - C^{l+1})$ control the flow of visual features V^{l+1} and language signals S^{l+1}, respectively, we have:

$$\begin{aligned} E^{l+1} &= V^{l+1} \odot C^l + S^{l+1} \odot (1 - C^l), \\ F^{l+1} &= AddNorm(FFN(AddNorm(\hat{F}^{l+1}, E^{l+1}))), \end{aligned} \tag{13}$$

where $\odot$ represents the Hadamard product and $E^{l+1} \in \mathbb{R}^{M\times d}$ denotes the output of the gated controller module.

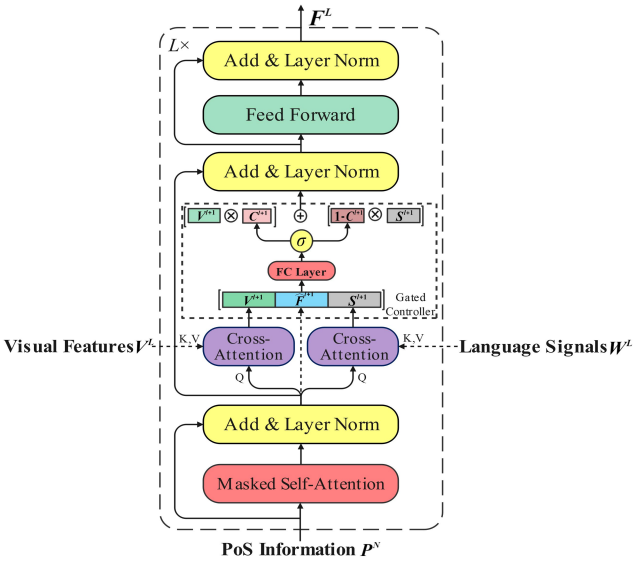

Figure 4. The structure of the dual attention based multimodal decoder model with part-of-speech guidance. The multimodal representations are first learned based on the dual attention with PoS guidance. Then, the gated controller is introduced to adaptively measure the contribution of visual and language cues for predicting words.

Finally, the output F^L of the PoS-guided SAT or DAT module is input into the word classifier to predict the next possible word as follows:

$$p(y_t|Y_{t-1}, V^L) = Softmax(W_{word} \cdot F^L_{t-1} + b_{word}), \tag{14}$$

where F^N_{t-1} is the hidden state corresponding to the $(t-1)$th word, the embedded matrix $W_{word} \in \mathbb{R}^{d\times\mathcal{D}}$, the bias vector $b_{word} \in \mathbb{R}^{\mathcal{D}}$, and $\mathcal{D}$ is the size of the vocabulary.

3.4. Training Details

As shown in Figures 3 and 4, the SAT-based and DAT-based multimodal decoder have the same input visual features, language signals, and PoS information as well as the same output vectors. The two outputs of our models are utilized to predict the next word and its PoS tag, which, respectively, correspond to two different objective functions. Thus, in practice, the network weights of these two models can be trained concurrently by a supervised multitask learning.

For an input image, assume its region-based visual feature vector as V, the corresponding ground-truth caption $Y^* = \{y^*_0, \cdots, y^*_T\}$ and the ground-truth PoS tags $S^* = \{s^*_0, \cdots, s^*_T\}$. For the self-attention PoS predictor, the cross-entropy (XE) loss for the PoS prediction is:

$$L_{PoS} = -\sum_{t=0}^{T} \log(p_\varphi(s^*_t|Y^*_{t-1})), \tag{15}$$

where φ represents the parameters of the SAPP network.

The parameters θ of our image captioning model (including dual-way encoder and PoS-guided multimodal decoder) is optimized via minimizing the following cross-entropy loss L_{word} between the generated captions and the ground truths:

$$L_{word} = -\sum_{t=0}^{T} \log(p_\theta(y_t^* | Y_{t-1}^*, V). \tag{16}$$

Combining the word prediction loss L_{word} with the PoS prediction loss L_{PoS}, the total loss function for our proposed PoS-Transformer framework can be defined as:

$$L = L_{word} + \lambda \cdot L_{PoS}, \tag{17}$$

where λ is a trade-off factor between the PoS loss and the word loss. Thus, all the parameters of the PoS-Transformer network can be optimized by minimizing the total loss function.

As can be seen from Figure 2, when minimizing the XE loss L_{word}, the parameter φ of the SAPP network will also be optimized, which indicates that the word prediction can be considered as the leading task of the whole model. At the same time, when the XE loss L_{PoS} is minimized, only the PoS predictor in the whole framework will be updated. Thus, the training of SAPP plays a role of auxiliary task for the main task. By means of the ground-truth PoS tags, the PoS prediction model can be well optimized, which provides the main task with the auxiliary optimization direction of the parameter φ. Consequently, with the guidance of SAPP, the image captioning part of our whole framework can be encouraged to generate more grounded and fine-grained captions.

At inference time, PoS-Transformer needs not employ any PoS tagger to tag each word in the generated sentences since it actually utilizes the current hidden state of the SAPP network as the PoS information to guide the caption generation.

4. Experiments

4.1. Datasets

MSCOCO [32]: This popular benchmark dataset contains 123k images and each of them is equipped with five manually annotated sentences. We adopted the offline Karpathy splits [33], which assigns 113k images for training, 5k images for validation, and 5k images for testing. Following the same settings in prior studies, we converted all sentences to lowercase, deleted the punctuation characters, tokenized each caption, and constructed a vocabulary including 9487 words by selecting the words which appeared more than five times.

Flickr30k [34]: Flickr30k consists of 31k images with five text descriptions each. Following prior studies [11,35], we used the publicly available split which divides Flickr30k into 29k/1k/1k for training/validation/test, respectively.

4.2. Evaluation Metrics

To evaluate the performance of different captioning methods, we used the full set of the standard evaluation metrics, including BLEU [36], METEOR [37], ROUGE-L [38], CIDEr [39], and SPICE [40]. All these metrics were calculated directly by using the MSCOCO caption evaluation tool (https://github.com/tylin/coco-caption (accessed on 15 November 2022)). BLEU is an n-gram precision-based metric, METEOR performs unigram matching, and SPICE computes an F1-score over caption scene-graph tuples, i.e., the balance between the precision and the recall. Notably, CIDEr is specially designed to evaluate the image captioning model. It obtains the similarity between the captions to be evaluated and the reference captions by calculating the TF-IDF weights of each n-tuple to evaluate the effectiveness of the image captioning. The number of times an n-gram w_k

occurs in a reference sentence s_{ij} is denoted by $h_k(s_{ij})$ or $h_k(c_i)$ for the candidate sentence c_i. The TF-IDF weighting $g_k(s_{ij})$ for each n-gram w_k can be formulated as:

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{w_l \in \Omega} h_l(s_{ij})} \log\left(\frac{|I|}{\sum_{I_{p \in I}} \min(1, \sum_q h_k(s_{pq}))}\right), \tag{18}$$

where ω is the vocabulary of all n-grams and I is the set of all images in the dataset. The CIDEr score for n-grams of length n is computed by using the average cosine similarity between the candidate sentence and the reference sentences, which accounts for both precision and recall:

$$\begin{aligned} CIDEr_n(c_i, S_i) &= \frac{1}{m} \sum_j \frac{g^n(c_i) \cdot g^n(s_{ij})}{||g^n(c_i)||||g^n(s_{ij})||}, \\ CIDEr(c_i, S_i) &= \sum_{n=1}^{N} w_n CIDEr_n(c_i, S_i), \end{aligned} \tag{19}$$

Empirically, the uniform weights $w_n = 1/N$ work the best and $N = 4$. The higher the CIDEr score, the better the resulting discourse quality.

4.3. Experimental Settings

(1) Data preprocessing: To gain the PoS tags of the reference captions in the training set, we employed the tagger provided by the Stanford University Natural Language Processing Research Group (https://Nlp.stanford.edu/software/tagger.shtml#Download (accessed on 15 November 2022)). Specifically, the PoS set included 12 universal PoS tags, such as verb (VERB), noun (NOUN), adjective (ADJ), etc.

(2) Implementation details: For the self-attention PoS predictor and language subencoder, we utilized randomly initialized word embeddings W^0, whose dimensionality was equal to d, and then summed the input vectors and their sinusoidal positional encodings [8]. For the visual subencoder, we used the pretrained Up-Down model [13] to extract the 2048-dimensional bottom-up features of the detected objects and linearly projected them to the 512-dimensional input visual vectors. Following the same settings as in [31], the latent dimensionality in each head was set to $d_h = d/h = 64$, where the latent dimensionality d was 512. The number of attention blocks L in the visual subencoder, language subencoder, and PoS-guided multimodal decoder ranged in $\{1, 2, 4, 6\}$ and that of the POS prediction model N was set to 3. During the training stage, we used the Adam optimizer [41] with 20,000 warm-up steps and a batch size of 10. Our models were first trained for 30 epochs with the cross-entropy loss and then further optimized with the CIDEr reward [42] for additional 30 epochs with a fixed learning rate of 5×10^{-6}. In the inference stage, the beam search strategy was adopted [8] with a beam size of three.

4.4. Ablation Studies

To validate the impacts of different modules and settings in our models on the captioning performance, we conducted extensive ablations including different numbers of encoding and decoding layers L, different values of the hyperparameter λ, and different PoS-guided attention mechanisms.

(1) Effect of encoding and decoding layers: To investigate the impact of the number of encoding and decoding layers, we applied the single-attention-based PoS-Transformer (SAT-PoS-Transformer) model with different numbers of stacked blocks $L \in \{1, 2, 4, 6\}$ on Flickr30k, as well as the dual-attention-based PoS-Transformer (DAT-PoS-Transformer) model on MSCOCO and Flickr30k, respectively. For simplicity, the numbers of stacked blocks in the encoder and decoder were set to the same value. Table 1 shows the performance of SAT-PoS-Transformer and DAT-PoS-Transformer with different L's on Flickr30k. We can observe that these two models achieved the best performance when using four encoding and four decoding layers. This was due to the fact that deeper layers enabled

the encoder of the captioner to represent more complicated relationships between objects and the decoder to provide more discriminative latent vectors for the prediction of words. However, if the number of layers becomes large, the risk of overfitting also increases. Table 2 reports the performance of DAT-PoS-Transformer with different L's on the MSCOCO dataset. Similarly, we can see that the generated image captions by our proposed models reached the highest scores on all metrics when $L = 4$. Thus, all subsequent experiments used four layers.

(2) Effect of the hyperparameter λ: To analyze the impact of the hyperparameter λ on the captioning performance, we applied our PoS-Transformer models with different values of λ on MSCOCO and Flickr30k, respectively. The experimental results of SAT-PoS-Transformer and DAT-PoS-Transformer on Flickr30k are illustrated in Table 3. It can be seen that DAT-PoS-Transformer with $\lambda = 0.75$ had the highest scores on most metrics and a pretty high BLEU-4 and ROUGE-L scores (only slightly lower than the highest 0.287 and 0.492, respectively). For SAT-PoS-Transformer, it reached relatively optimal performance when $\lambda = 1.00$. As can be seen from Table 4, when the coefficient λ of the PoS loss function increased to 0.50, DAT-PoS-Transformer obtained the highest scores in terms of all metrics on MSCOCO.

(3) Effect of single attention and dual attention: As shown in Tables 1 and 3, the image captions generated by SAT-PoS-Transformer with $L = 4$ and $\lambda = 1.00$ reached the highest scores on most metrics. It can be also observed from Table 3 that DAT-PoS-Transformer significantly outperformed SAT-PoS-Transformer in terms of all metrics. Based on the dual attention mechanism, the best CIDEr score increased from 0.601 to 0.612 on the Flickr30k dataset, which validated the superiority of dual attention over single attention.

Table 1. The performance of PoS-Transformer with different numbers of encoding and decoding Layers on Flickr30k dataset. B@1, B@2, B@3, B@4, M, R, C, and S are short for BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-L, CIDEr, and SPICE, respectively.

L	SAT-PoS-Transformer								DAT-PoS-Transformer							
	B@1	**B@2**	**B@3**	**B@4**	**M**	**R**	**C**	**S**	**B@1**	**B@2**	**B@3**	**B@4**	**M**	**R**	**C**	**S**
1	0.687	0.501	0.365	0.265	0.215	0.480	0.590	0.162	0.695	0.521	0.383	0.280	0.215	0.488	0.595	0.160
2	0.685	0.505	0.368	0.269	0.219	0.482	0.593	0.164	0.698	0.523	0.384	0.282	0.219	0.489	0.598	0.164
4	0.697	0.522	0.386	0.285	0.220	0.490	0.601	0.164	0.699	0.524	0.387	0.286	0.221	0.489	0.607	0.162
6	0.690	0.512	0.374	0.271	0.222	0.484	0.597	0.161	0.696	0.518	0.380	0.278	0.219	0.488	0.586	0.158

Table 2. The performance of DAT-PoS-Transformer with different numbers of encoding and decoding layers on MSCOCO.

L	**B@1**	**B@2**	**B@3**	**B@4**	**M**	**R**	**C**	**S**
1	0.752	0.592	0.455	0.351	0.274	0.560	1.132	0.205
2	0.755	0.594	0.459	0.356	0.278	0.563	1.139	0.209
4	0.762	0.601	0.465	0.359	0.282	0.567	1.155	0.211
6	0.756	0.595	0.460	0.355	0.280	0.564	1.146	0.210

Table 3. The performance of PoS-Transformer with different values of hyperparameter λ on Flickr30k dataset.

λ	SAT-PoS-Transformer								DAT-PoS-Transformer							
	B@1	**B@2**	**B@3**	**B@4**	**M**	**R**	**C**	**S**	**B@1**	**B@2**	**B@3**	**B@4**	**M**	**R**	**C**	**S**
1.00	0.697	0.522	0.386	0.285	0.220	0.490	0.601	0.162	0.699	0.524	0.387	0.286	0.221	0.489	0.607	0.162
0.75	0.698	0.519	0.379	0.276	0.219	0.484	0.593	0.160	0.703	0.527	0.388	0.284	0.221	0.489	0.612	0.166
0.50	0.693	0.518	0.381	0.279	0.222	0.487	0.591	0.162	0.697	0.524	0.388	0.287	0.220	0.492	0.599	0.159
0.25	0.691	0.518	0.376	0.274	0.220	0.484	0.591	0.159	0.695	0.522	0.387	0.284	0.196	0.488	0.583	0.154

Table 4. The performance of DAT-PoS-Transformer with different values of hyperparameter λ on MSCOCO "Karpathy" test split.

λ	B@1	B@2	B@3	B@4	M	R	C	S
1.00	0.762	0.601	0.465	0.359	0.282	0.567	1.155	0.211
0.75	0.764	0.605	0.469	0.360	0.279	0.565	1.150	0.210
0.50	0.766	0.606	0.469	0.363	0.282	0.569	1.161	0.211
0.25	0.760	0.602	0.464	0.357	0.280	0.566	1.145	0.209

4.5. Quantitative Analysis

According to the ablation studies, we compared our best DAT-PoS-Transformer model with the competitive methods on Flickr30k and MSCOCO datasets.

(1) Results on the MSCOCO Karpathy test splits: In Table 5, we compared DAT-PoS-Transformer with LSTM [43], SCST [42], ADP-ATT [12], LSTM-A [44], Up-Down [13], RFNet [45], GCN-LSTM [25], SGAE [24], AVSG [26], and ORT [15] on the offline COCO Karpathy test split. In addition, we also compared DAT-PoS-Transformer with part-of-speech-based image captioning methods such as PoS-Guiding [28], Inject+PoS [27], PoS-SCAN [46], and CNM [47]. LSTM introduced a deep model with two attention mechanisms to distill information in images down to the most salient objects. LSTM-A improved LSTM by emphasizing semantic attributes at the decoding stage. ADP-ATT introduced a visual sentinel and sentinel gate to adaptively determine whether to attend to the visual regions for the word prediction. Up-Down and RFNet improved the attention mechanism by having it learn to identify selective spatial regions, which further boosted the performance of the captioning generation. ORT developed an object relation transformer captioning model which explicitly incorporated spatial relationships between region features through geometric attention. GCN-LSTM, SGAE, and AVSG used a scene graph which contained rich semantic information to improve the image understanding. As can be seen from Table 5, compared with the existing PoS-based methods, our method had better performance on most metrics when optimized with the self-critical loss [42]. Remarkably, the CIDEr score and BLEU-4 score of our model could reach 129.9% and 39.3%, which were 2% and 4% better than the best comparison model CNM [47], respectively. In addition, other than [28] which exploited PoS tags as switches to decide whether or not to utilize visual features at each time step, our method did not need any PoS tagger in the test stage. Compared to [27], which also introduced a PoS prediction model to image captioning, our PoS-Transformer model not only overcame the limitation of dependencies between distant positions in language modeling, but also incorporated the novel PoS-guided attention module to more flexibly adapt to the variation of PoS for each word. Furthermore, compared with the strong baseline (Transformer), which followed the traditional language model, the proposed PoS-Transformer model achieved better performance on all metrics, which demonstrated the effectiveness of our model with the PoS guidance and dual attention mechanism.

(2) Results on the Flickr30k dataset: We also compared DAT-PoS-Transformer to other methods trained by cross-entropy loss on the Flickr30k dataset. As can be seen in Table 6, our method surpassed all other approaches in terms of BLEU-1∼BLEU-4 and CIDEr. The METEOR and ROUGE-L scores of our method were worse than those of Inject+PoS [27]. Remarkably, it improved on the performance of the Inject+PoS model on CIDEr by 0.143 points (from 0.469 to 0.612). Thus, our method achieved better performance in comparison with the existing PoS-based models. Notably, our model had superior performance over the strong baseline (the original Transformer model) on all metrics, which further validated that it was effective at generating the captions with PoS guidance.

Table 5. Comparison of image captioning performance with state-of-the-art methods on MSCOCO "Karpathy" test split.

Model	PoS	Cross-Entropy Loss					Self-Critical Loss				
		B@1	B@4	M	R	C	B@1	B@4	M	R	C
LSTM [43]	✗	-	0.296	0.252	0.526	0.940	-	0.319	0.255	0.543	1.063
SCST [42]	✗	-	0.300	0.259	0.534	0.994	-	0.342	0.267	0.557	1.140
ADP-ATT [12]	✗	0.742	0.332	0.266	-	1.085	-	-	-	-	-
Up-Down [13]	✗	0.772	0.362	0.270	0.564	1.135	0.798	0.363	0.277	0.569	1.201
RFNet [45]	✗	0.764	0.358	0.274	0.565	1.125	0.791	0.365	0.277	0.573	1.219
GCN-LSTM [25]	✗	0.773	0.368	0.279	0.570	1.163	0.805	0.382	0.285	0.583	1.276
SGAE [24]	✗	-	-	-	-	-	0.808	0.384	0.284	0.586	1.278
ORT [15]	✗	0.766	0.355	0.280	0.566	1.154	0.805	0.386	0.287	0.584	1.283
AVSG [26]	✗	-	-	-	-	-	0.807	0.387	0.285	0.586	1.289
PoS-Guiding [28]	✓	0.711	0.279	0.239	-	0.882	-	-	-	-	-
Inject+PoS [27]	✓	0.761	0.335	0.301	0.605	0.951	-	-	-	-	-
PoS-SCAN [46]	✓	-	-	-	-	-	0.802	0.380	0.285	-	1.259
CNM [47]	✓	0.776	0.371	0.279	0.573	1.166	0.808	0.389	0.284	0.588	1.279
DAT-PoS-Transformer	✓	0.766	0.363	0.282	0.569	1.161	0.808	0.393	0.290	0.589	1.299

Table 6. Comparison of image captioning performance with state-of-the-art methods on Flickr30k caption dataset under cross-entropy loss.

Methods	B@1	B@2	B@3	B@4	M	R	C
LSTM [43]	0.663	0.423	0.277	0.183	-	-	-
Soft-Att [11]	0.667	0.434	0.288	0.191	0.185	-	-
Hard-Att [11]	0.669	0.439	0.296	0.199	0.185	-	-
ATT-FCN [48]	0.647	0.460	0.324	0.230	0.189	-	-
ADP-ATT [12]	0.677	0.494	0.354	0.251	0.204	0.467	0.531
SCA-CNN [49]	0.662	0.468	0.325	0.223	0.195	-	-
Transformer (Base)	0.664	0.483	0.345	0.243	0.212	0.466	0.551
BCAN [50]	0.698	0.519	0.378	0.274	0.212	0.488	0.583
PoS-Guiding [28]	0.638	0.446	0.307	0.211	-	-	-
Inject+POS [27]	0.694	0.498	0.355	0.254	0.251	0.538	0.469
DAT-PoS-Transformer	0.703	0.527	0.388	0.284	0.221	0.489	0.612

4.6. Qualitative Analysis

Figure 5 shows some test images and their corresponding captions and PoS sequences generated by PoS-Transformer and the Transformer baseline, respectively.

Intuitively, the descriptions generated by PoS-Transformer were more precise and distinguishable compared to the Transformer baseline. The reason was that by introducing the PoS information guidance, our model was encouraged to align the visual words with the grounding visual features, while the generated captions conformed to the grammatical rules better. More specifically, our model could generate more fine-grained and grounded captions than the original Transformer model. Taking the fifth image as an example, the Transformer baseline only generated a simple sentence *a baseball player holding a bat*. Instead, our model generated the caption *a baseball game in progress with the batter up at the plate*, which was more fine-grained and had the same semantic meaning as the ground truth. In addition, in the last image, our model generated the feasible sentence *a large bird with a long beak walking on a beach*, while the Transformer baseline inferred the simple but wrong sentence *a bird that flying in the air*. Notably, the PoS tags generated by our model included two more ADJ (*large* and *long*) and one NOUN (*beak*), which made the description more vivid and detailed. Additionally, it can be seen from Figure 5 that in most cases, the self-attention PoS predictor was able to precisely predict the PoS tags. It is

worth noting that the corresponding word could also be inferred correctly even if its PoS tag was incorrect, which implied that the PoS predictor actually played a role of auxiliary task, and by means of the beam search strategy [8], the proposed model had the capability to correct errors on the PoS tags to some extent.

GT: A man with glasses and his eyes closed dressed in a black shirt and a necktie.
Transformer: A man in a suit poses for a picture.
PoS-Transformer: A man wearing a suit and tie with glasses.
PoS: DET NOUN VERB DET NOUN VERB NOUN VERB NOUN.

GT: A motorcycle in the foreground parked in a dirt parking lot.
Transformer: A motorcycle that is parked in the dirt.
PoS-Transformer: A motorcycle parked on a dirt field next to a fence.
PoS: DET NOUN VERB ADP DET NOUN NOUN ADP PRT DET NOUN.

GT: A boat that is decorated with flags on the water.
Transformer: A boat is sitting in the water.
PoS-Transformer: A small boat with a flag on it in the water.
PoS: DET NOUN NOUN ADP DET NOUN ADP DET ADP DET NOUN.

GT: Two people are snowboarding down a hill fast.
Transformer: A couple of men riding down a snow covered slope.
PoS-Transformer: Two men are snowboarding down a snowy hill.
PoS: DET NOUN VERB VERB ADP DET NOUN NOUN.

GT: A batter up at the plate in a baseball game.
Transformer: A baseball player standing next to home plate.
PoS-Transformer:A baseball game in progress with the batter up to plate.
PoS: DET NOUN NOUN ADP NOUN ADP DET NOUN ADV PRT NOUN.

GT: A red plane flying through a blue sky.
Transformer: A red plane is flying in the sky.
PoS-Transformer: A red fighter jet flying through a blue sky.
PoS: DET ADJ NOUN NOUN VERB ADP DET NOUN NOUN.

GT: The man and the little girl are walking past the statue.
Transformer: A large building with a statue in front of it.
PoS-Transformer: People walking past a statue in a town square.
PoS: NOUN VERB ADP DET NOUN ADP DET NOUN NOUN.

GT: A bird standing on top of a beach next to water.
Transformer: A bird that is flying in the air.
PoS-Transformer: A large bird with a long beak walking on a beach.
PoS: DET ADJ NOUN ADP DET NOUN NOUN VERB ADP DET NOUN.

Figure 5. Examples of captions generated by standard Transformer and our proposed model as well as ground truths. Moreover, the PoS sequences generated by our self-attention PoS predictor are also presented. The correct and incorrect PoS tags are colored in green and red, respectively. Generally, our method can generate more accurate and fine-grained captions.

We further visualized the image regions attended to and the variations of gate values in the gate controller during the caption generation in Figure 6. For each word, we mainly analyzed its gate value of the gate controller in the last decoding block since it was directly used to infer the next word. From Figure 6, we can observe that the proposed model was able to correctly attend to the corresponding image regions when predicting the visual words, e.g., *baseball*, *game*, and *batter*, while preventing itself from attending to any image region if a nonvisual word was being generated, such as *a*, *process*, *the*, etc. To be specific, our model assigned a pretty large gate value (over 0.9) for visual words. Note

that some nonvisual words following NOUN, such as *in* and *up*, may also be assigned gate values larger than 0.5, which was reasonable since these words actually represented the relationships between objects, i.e., they were closely related to the visual words. The visualization experiment could further demonstrate that our PoS-Transformer model effectively took advantage of the PoS information to adaptively adjust the effect of visual features and language signals on the word prediction.

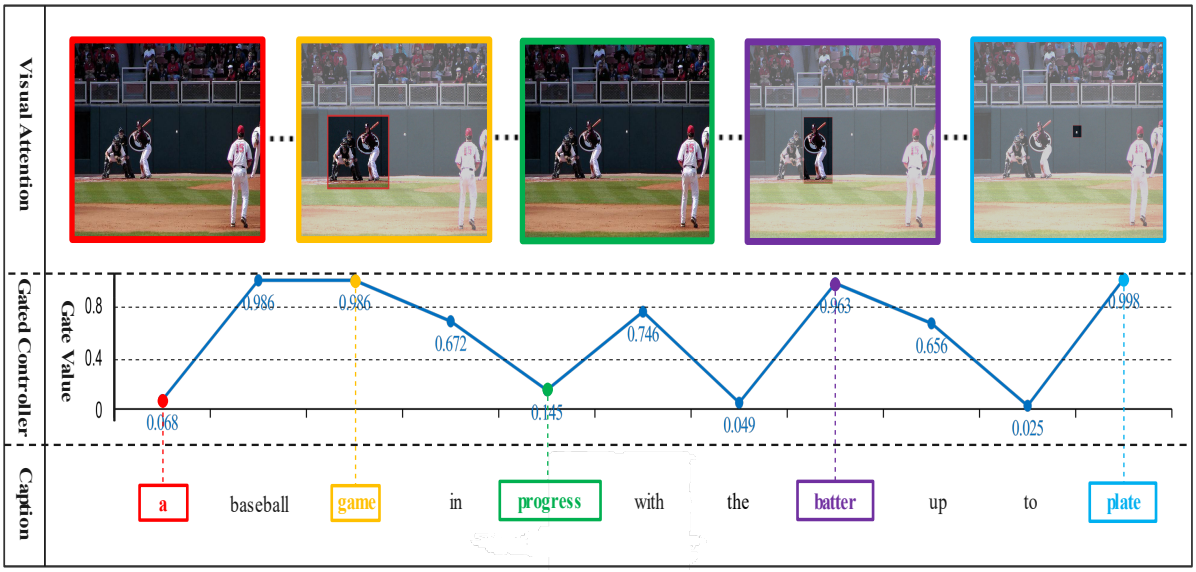

Figure 6. Visualization of attention regions in the caption generation process for PoS-Transformer and gate value of each generated word for controlling the flow of visual features and language signals attended to. By virtue of the PoS information and gated controller, PoS-Transformer is able to adaptively adjust the effect of visual features and language signals on the predicted visual or nonvisual word.

5. Conclusions

In this paper, we presented PoS-Transformer, a novel transformer-based framework for image captioning, to separate the grammatical structures and word semantics of captions and incorporate the PoS guiding information into the modeling. PoS-Transformer seamlessly integrated the PoS prediction module with the transformer-based captioner for a more grounded and fine-grained image captioning. By virtue of two proposed attention mechanisms, the PoS-Transformer decoder effectively exploited the PoS information to guide the caption generation, which not only adaptively adjusted the weights between visual and language signals for more grounded captioning, but leveraged the PoS information to generate more fine-grained sentences. Extensive experiments as well as ablation studies demonstrated that our method could significantly boost the performance of image captioning on top of the transformer-based architecture and substantially outperform other PoS-based image captioning models on the Flickr30k and MSCOCO datasets.

The current PoS-Transformer model focuses on introducing syntactic structures into the conventional language model in image captioning, which can play a better role in robot interaction, preschool education, and other application fields. Additional visual and semantic encoding approaches, such as exploiting the image attributes and the relative geometry relations between the objects, are not integrated with PoS-Transformer. However, it has been validated that these approaches can provide much richer visual and semantic information to facilitate a high-quality caption generation. In our future work, we will further enrich the representations of visual and semantic concepts to boost the performance of PoS-Transformer.

Author Contributions: Conceptualization, D.W. and B.L.; methodology, D.W. and B.L.; software, D.W. and B.L.; validation, D.W. and B.L.; formal analysis, Y.Z. and R.Y.; investigation, D.W. and M.L.; resources, D.W.; data curation, D.W. and B.L.; writing—original draft preparation, D.W.; writing—review and editing, D.W., B.L., Y.Z. and P.L.; visualization, D.W. and B.L.; supervision, M.L.; project administration, D.W.; funding acquisition, B.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (grant number no. 62276266, no. 61801198, no. 62272461), the Graduate Innovation Program of China University of Mining and Technology (grant number 2022WLJCRCZL270), and by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (grant number SJCX22_1134).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. These datasets can be found here: https://cocodataset.org/ (accessed on 15 November 2022) and http://shannon.cs.illinois.edu/DenotationGraph/ (accessed on 15 November 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xu, N.; Liu, A.; Wong, Y.; Zhang, Y.; Nie, W.; Su, Y.; Kankanhalli, M.S. Dual-Stream Recurrent Neural Network for Video Captioning. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 2482–2493. [CrossRef]
2. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; NIPS: Lake Tahoe, NV, USA, 2012; pp. 1106–1114.
3. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
4. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; NIPS: Montreal, QC, Canada, 2015; pp. 91–99.
5. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
6. Cho, K.; van Merrienboer, B.; Gülçehre, Ç.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1724–1734.
7. Kalchbrenner, N.; Blunsom, P. Recurrent continuous translation models. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 1700–1709.
8. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*; NIPS: Montreal, QC, Canada, 2014; pp. 3104–3112.
9. Miller, E.K.; Cohen, J.D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **2001**, *24*, 167–202. [CrossRef] [PubMed]
10. Thompsonschill, S. Dissecting the language organ: A new look at the role of Broca's area in language processing. In *Twenty-First Century Psycholinguistics: Four Cornerstones*; Routledge: New York, NY, USA, 2005 ; pp. 313–330.
11. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
12. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 375–383.
13. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6077–6086.
14. Yu, J.; Li, J.; Yu, Z.; Huang, Q. Multimodal Transformer with Multi-View Visual Representation for Image Captioning. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 4467–4480. [CrossRef]
15. Herdade, S.; Kappeler, A.; Boakye, K.; Soares, J. Image captioning: Transforming objects into words. *arXiv* **2019**, arXiv:1906.05963.
16. Guo, L.; Liu, J.; Zhu, X.; Yao, P.; Lu, S.; Lu, H. Normalized and geometry-aware self-attention network for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10327–10336.
17. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-memory transformer for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10578–10587.

18. Pan, Y.; Yao, T.; Li, Y.; Mei, T. X-linear attention networks for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10971–10980.
19. Ji, J.; Luo, Y.; Sun, X.; Chen, F.; Luo, G.; Wu, Y.; Gao, Y.; Ji, R. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; Volume 35, pp. 1655–1663.
20. Luo, Y.; Ji, J.; Sun, X.; Cao, L.; Wu, Y.; Huang, F.; Lin, C.; Ji, R. Dual-level Collaborative Transformer for Image Captioning. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; pp. 2286–2293.
21. Zhang, X.; Sun, X.; Luo, Y.; Ji, J.; Zhou, Y.; Wu, Y.; Huang, F.; Ji, R. RSTNet: Captioning with Adaptive Attention on Visual and Non-Visual Words. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 15465–15474.
22. Liu, W.; Chen, S.; Guo, L.; Zhu, X.; Liu, J. CPTR: Full Transformer Network for Image Captioning. *arXiv* **2021**, arXiv:2101.10804.
23. Li, G.; Zhu, L.; Liu, P.; Yang, Y. Entangled transformer for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 8928–8937.
24. Yang, X.; Tang, K.; Zhang, H.; Cai, J. Auto-encoding scene graphs for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 10685–10694.
25. Yao, T.; Pan, Y.; Li, Y.; Mei, T. Exploring Visual Relationship for Image Captioning. In *Proceedings of the European Conference on Computer Vision*; Springer: Munich, Germany, 2018; pp. 711–727.
26. Zhao, S.; Li, L.; Peng, H. Aligned visual semantic scene graph for image captioning. *Displays* **2022**, *74*, 102210. [CrossRef]
27. Zhang, J.; Mei, K.; Zheng, Y.; Fan, J. Integrating Part of Speech Guidance for Image Captioning. *IEEE Trans. Multimed.* **2021**, *23*, 92–104. [CrossRef]
28. He, X.; Shi, B.; Bai, X.; Xia, G.; Zhang, Z.; Dong, W. Image Caption Generation with Part of Speech Guidance. *Pattern Recognit. Lett.* **2019**, *119*, 229–237. [CrossRef]
29. Deshpande, A.; Aneja, J.; Wang, L.; Schwing, A.G.; Forsyth, D. Fast, diverse and accurate image captioning guided by part-of-speech. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 10695–10704.
30. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; NIPS: Long Beach, CA, USA, 2017; pp. 5998–6008.
32. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
33. Karpathy, A.; Joulin, A.; Fei-Fei, L. Deep fragment embeddings for bidirectional image sentence mapping. *arXiv* **2014**, arXiv:1406.5679.
34. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78. [CrossRef]
35. Karpathy, A.; Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3128–3137.
36. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
37. Satanjeev, B. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, MI, USA, 25–30 June 2005; pp. 228–231.
38. Szpakowicz, S. Text summarization branches out. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, 21–26 July 2004.
39. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
40. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. Spice: Semantic propositional image caption evaluation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 382–398.
41. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
42. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-critical sequence training for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1179–1195.
43. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
44. Yao, T.; Pan, Y.; Li, Y.; Qiu, Z.; Mei, T. Boosting Image Captioning with Attributes. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4904–4912.
45. Jiang, W.; Ma, L.; Jiang, Y.G.; Liu, W.; Zhang, T. Recurrent fusion network for image captioning. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 510–526.
46. Zhou, Y.; Wang, M.; Liu, D.; Hu, Z.; Zhang, H. More Grounded Image Captioning by Distilling Image-Text Matching Model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4776–4785.

47. Yang, X.; Zhang, H.; Cai, J. Learning to Collocate Neural Modules for Image Captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, South Korea, 27 October–2 November 2019; pp. 4249–4259.
48. You, Q.; Jin, H.; Wang, Z.; Fang, C.; Luo, J. Image captioning with semantic attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4651–4659.
49. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.S. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5659–5667.
50. Jiang, W.; Wang, W.; Hu, H. Bi-Directional Co-Attention Network for Image Captioning. *ACM Trans. Multim. Comput. Commun. Appl.* **2021**, *17*, 1–20. [CrossRef]