

Received March 27, 2022, accepted April 15, 2022, date of publication April 22, 2022, date of current version May 3, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3169781

# Image Captioning Model Using Part-of-Speech Guidance Module for Description With Diverse Vocabulary

JU-WON BAE<sup>1</sup>, SOO-HWAN LEE<sup>1</sup>, WON-YEOL KIM<sup>2</sup>, JU-HYEON SEONG<sup>3</sup>,  
AND DONG-HOAN SEO<sup>4</sup>

<sup>1</sup>Department of Electronics & Electrical Engineering, Interdisciplinary Major of Maritime AI Convergence, Korea Maritime and Ocean University, Busan 49112, South Korea

<sup>2</sup>Artificial Intelligence Convergence Research Center for Regional Innovation, Korea Maritime & Ocean University, Busan 49112, South Korea

<sup>3</sup>Department of Liberal Education, Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, Busan 49112, South Korea

<sup>4</sup>Division of Electronics and Electrical Information Engineering, Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, Busan 49112, South Korea

Corresponding author: Dong-Hoan Seo (dhseo@kmou.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government [Ministry of Science and ICT (MSIT)] under Grant NRF-2021R1A2C1014024.

**ABSTRACT** Image captions aim to generate human-like sentences that describe the image's content. Recent developments in deep learning (DL) have made it possible to caption images for accurate descriptions and detailed expressions. However, since DL learns the relationship between images and captions, it constructs sentences based on commonly frequented words in the dataset. Although these generated sentences are highly accurate, they have low lexical diversity, unlike humans due to limited vocabulary. Therefore, in this paper, we propose a Part-Of-Speech (POS) guidance module and a multimodal-based image captioning model that determines the intensity of images and word sequences and generates sentences through POS to enhance the lexical diversity of DL. The proposed POS guidance module enables rich expression by controlling the information of images and sequences based on the predicted POS guidance to predict words. Then, the POS multimodal layer adds POS and output vector of Bi-LSTM using the multimodal layer to predict the next caption, considering the grammatical structure. We trained and tested the proposed model on the Flickr 30K and MS COCO datasets and compared them with current state-of-the-art studies. Also, we analyzed the lexical diversity of the caption model through the Type-Token Ratio (TTR) and confirmed that the proposed model generates sentences using several words.

**INDEX TERMS** Deep learning, image captioning, multimodal layer, part of speech.

## I. INTRODUCTION

Artificial intelligence has been targeting several segments of human life in various domains. Among them, image captioning is a technique for modeling human actions performed to understand a scene. Image captioning aims to generate human-like captions that enable machines to understand images and describe context through natural language. Image captioning is challenging in computer vision and natural language processing fields because they analyze two disparate data: images and natural language [1]–[5]. Furthermore, since the machine is described based on a

probabilistic approach, this caption lacks a human-like systematic presentation focusing on semantic significance. Therefore, most research focused on generating sentences with correct grammar while including all the image contexts.

Most studies in the image caption area apply the encoder-decoder framework, which consists of an encoder that extracts features from an image and a decoder that generates sentences due to the development of deep learning. Unlike conventional methods [6]–[9], this structure can create various captions from scenes without using a fixed sentence template [11]–[14]. This description has a more unconstrained structure than before, detailing the context. This strength made this approach the mainstream of image captioning research. Recent research has significantly improved word

The associate editor coordinating the review of this manuscript and approving it for publication was Li He<sup>1</sup>.

prediction accuracy through attention mechanisms that focus on important features of images and sentences [15]–[18]. The attention mechanism maps importance based on the frequency of appearance between images and words according to a sequence. Therefore, these methods have a lexical bias that mainly uses vocabulary that appears a lot in the dataset.

Some studies have improved the structure and accuracy of sentences by adding metadata. He *et al.* [19] and Zhang *et al.* [11] improved the image captioning model by using the part-of-speech (POS) of a sentence as metadata. These studies introduced the method to insert POS as an attribute in the decoder of the language model. However, this structure does not consider the inherent characteristics of POS, so it has no advantages such as lexical diversity. POS classifies each element in a sentence and contains information on the role of words according to sequences. Also, the POS represented by nouns, adjectives, verbs, and adverbs does not simply indicate the position of the vocabulary in the sentence but also affects the selection of objects in the image because it is the order of description. In detail, stopwords such as adverbs and conjunctions, which are POSs that affect structure rather than meaning in a sentence, are attributes of the language model. Conversely, the rest of the POS containing the meaning of the text is defined as attributes for the image. Therefore, the language model can select visual and linguistic information through POS and be used directly for word prediction. According to the POS characteristics, image captioning can generate accurate captions and obtain sentences with enhanced lexical diversity if both methods are applied.

In this paper, we propose the POS Guidance Module, which determines the intensity of image features and word sequences according to POS, and the Multimodal-based image captioning model that combines POS and caption information with a multimodal layer. And we used Bi-directional Long Short-Term Memory (Bi-LSTM) [20] as a decoder to create captions reflecting bi-directional context. We tested our model on the Flickr 30K [21] and MS COCO [22] datasets. And we evaluate the model's performance using standard metrics such as BLEU [23], METEOR [24], ROUGE-L [25], CIDEr [26], and Word Mover's Distance (WMD) [27] compared with the state-of-art model. In addition, to check the lexical diversity in the sentences, we use TTR [28] metrics by comparing the variety of vocabularies of the sentences generated through the test result caption of models. The milestones which we proposed in this paper are:

- We propose a POS guidance module that improves lexical diversity by focusing on the image vector and caption embedding vector considering direct POS guidance.
- We propose a multimodal layer that combines the time-series network's output with POS information, applying to the decoder's predicted captions that conform to the grammatical structure.

The rest of this paper is organized as follows: Section II briefly reviews image captioning research and other studies

like the proposed model. Section III describes the details of the proposed model. Section IV describes experimental results using evaluation metrics and examples of generated captions. Section V concludes the paper and describes the expected direction.

## II. THEORETICAL BACKGROUND

Studies of image captions using deep learning have achieved excellent performance in computer vision and natural language processing [29]–[33]. These select the correct words corresponding to the image rather than conventional methods and take a structurally appropriate form.

Vinyals' model, Google Neural Image Caption (NIC) [1], transforms the visual features of an image into intermediate representations and uses them to generate sentences. This structure achieved high-accuracy performance by applying CNN and RNN, individually suitable for vision and natural language. In addition, models based on this structure can design networks end-to-end from input to output. This structure is called the encoder-decoder framework as the basic framework of recent image captioning research.

Mao *et al.* [34] proposed a model in which LSTM [35] is applied to the decoder, and multimodal structure is used for the intermediate representation. This model minimizes the loss of LSTM-encoded information and increases the accuracy of the generated sentences by directly combining the visual presentation. Yao *et al.* [36] proposed inserting attributes as metadata and providing visual information as input to the decoder. This model can focus on the scene and create more accurate captions by providing additional image information as an attribute. Guan *et al.* [37] improved the memory loss problem by repeatedly inputting feature vectors into all cells. Also, this research analyzed various inputting features to the decoder methods.

The attention mechanism, which is widely applied due to its high performance in machine translation, has also been used to image captions. Xu *et al.* [2] proposed a model based on the soft/hard attention mechanism, which finds notable areas of interest in an image and generates captions by focusing on those areas. You *et al.* [3] proposed a semantic attention method that generates sentences focusing on the attributes closely associated with the image content based on the visual attributes detected in an image. Johnson *et al.* [14] proposed a method of generating multiple sentences for a large volume of information, which can be obtained from an image simultaneously.

Vaswani *et al.* [38] proposed the Transformer network by significantly improving the attention mechanism widely applied in language models. Since then, based on this network, research has been conducted to use the Transformer network in the image captioning field. Cornia *et al.* [39] proposed a Meshed Memory Transformer to improve the accuracy of sentences according to the characteristics of image captions using many types of data. Guo *et al.* [40] proposed a Normalized Self-Attention Network (NSA) to

effectively learn new data distributions. They also proposed a Geometry-Aware Self-Attention Network (GSA) to model the geometric relationships of objects in an image.

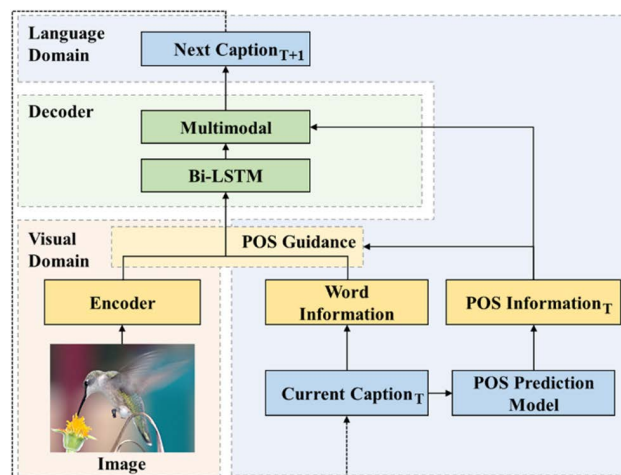
Also, there are other image captioning studies using various information that can be obtained from sentences. Ge et al. [41] proposed a model using Mutual-aid Bi-direction LSTM that is consider human cognitive styles that describe the image to be explained and the sentence to be generated while recognizing the overall image. Ke et al. [42] proposed the model correctly generates captions for long sequences, using Reflective Decoding Networks (RDNs) that can recognize the relative position of words. Yang et al. [43] proposed Collocate Neural Modules (CNMs) for learning dynamic structures biased to language collocate. This model is designed in a form similar to VQA, extracts information about objects, features, and relationships, and creates sentences with a decoder. Yi et al. [44] proposed caption-based Visual Relationship Graphs (VRGs) to strengthen the connection between the predicate that describes an object in a sentence and the target region in which the object is located. And using VRG as attention, they generated captions in which objects and predicates are semantically connected.

In addition, there are image captioning studies using POS. He et al. [19] proposed a model for generating sentences by inputting the POS of the predicted word and image feature information at each time step into the decoder's LSTM. While this method was the first attempt toward integrating the POS to emphasize the images, the words with multiple POSs were not considered because a pre-made POS dictionary was used. Zhang et al. [11] proposed the POS Guidance module, a method to use POS as a guide in image captioning. They proposed two models using POS as a guide for the inject-based method and the merge-based method, which are the most used among the image captioning methods analyzed by Tanti et al. [45]. These image captioning models benefited from making the sentence structure precise and diversified by using the POS as an auxiliary indicator. While in our model, we use POS guide to word embedding vector and image feature vector in the inject-based method. Also, we consider that POS is regarded as the main attribute corresponding to image or language information. So, we approach it from a multimodal perspective and apply the multimodal layer for optimization.

### III. PROPOSED IMAGE CAPTIONING MODEL

We propose the image captioning model that uses POSs more directly to reinforce lexical diversity and keep grammatical rules. Figure 1 shows the architecture of the proposed image captioning model. The proposed model follows the inject-based encoder-decoder framework and is structurally divided into Visual Domain, Decoder, and Language Domain. Visual Domain extracts the features of the image input by using the encoder. In the Language Domain, POS Prediction Model predicts POS from the current caption, in timestep T. Also, POS Guidance Module uses the POS information as Guidance to process the information to generate captions by

focusing on image feature information and caption information. In other words, the model semantically connects images and sentence information using POS guidance directly generates rich expressions. The decoder uses Bi-LSTM to generate sentence information considering the bi-directional context. And the POS multimodal layer combines the sequential information from the Bi-LSTM and the sequential POS information from the POS prediction model in the Language Domain to predict the next timestep  $T + 1$  caption considering the grammatical rules. That is, we use POS as a guide for data application before the decoder and as the main attribute that can be directly involved in word prediction after the decoder. And since we approach the output of the POS and Bi-LSTM in a multimodal approach, we can get optimized data. Therefore, the proposed model can obtain lexical diversity and well-described sentences. The following section describes the proposed POS guidance module and multimodal-based image captioning model.

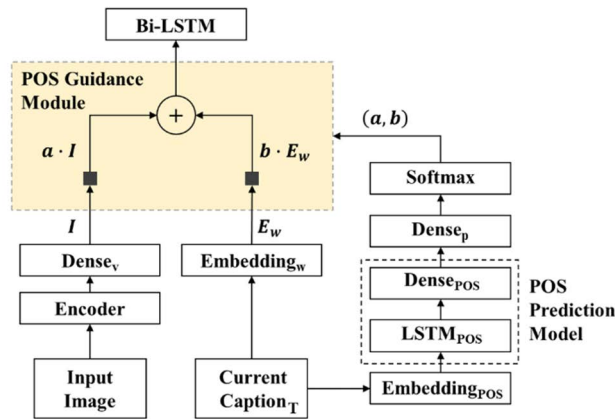


**FIGURE 1. Overview of the proposed image-captioning model. The proposed model use image and caption. And use POS Guidance to consider what features to consider, image and word information. And the model uses POS multimodal layer to combine Bi-LSTM's output and POS information, predict next caption can be considered POS.**

#### A. POS GUIDANCE MODULE

This study aims to generate sentences using the POS with various word expressions and accurate descriptions. Therefore, it is necessary to focus on the image feature information and sentence information that POS has. To use this point, we propose the POS Guidance Module in Figure 2. The POS Guidance Module connects image features and caption embedding vectors according to predicted POS information from POS Prediction Model. This module links the two sets of data by assigning probability according to how much focus should be placed on the image vectors and caption embedding vectors and transmitting it to the decoder. Therefore, information on the POS of each word is required. However, the POS of the word can be obtained only when the entire sentence is constructed. We use the POS Prediction model

proposed by Zhang *et al.* [11] for obtaining POS data. The POS prediction model uses LSTM to predict the POS of the word displayed at the next time step according to the word at the current time step. In Figure 2, POS Prediction Model receives  $Embedding_{POS}$  as input, which has same form as  $Embedding_w$  and name differently for differentiation. Then, the POS of the next time step for the current input word is predicted through the  $LSTM_{POS}$  layer. Finally, data is processed according to the dimension of POS categories through the  $Dense_{POS}$  layer. This data can be seen as a POS for a word predicted at the next time step. And because the information obtained from this POS prediction model is trained through a time series network, it also contains the sequential POS information. The output of the POS Prediction Model is processed through the fully connected layer,  $Dense_p$ .



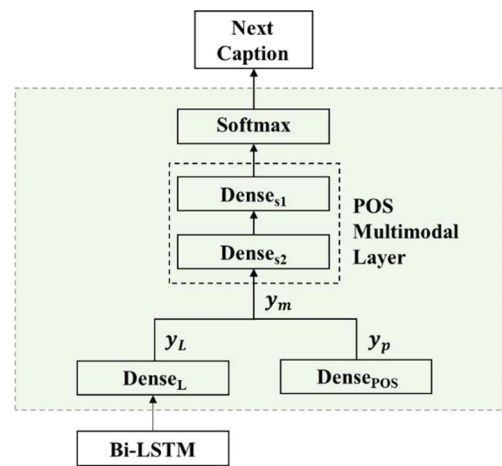
**FIGURE 2.** Structure of the proposed captioning encoder along with the POS guidance module. The POS guidance module can focus on the image feature and caption feature, using predicted POS information which is the output of the POS prediction model.

POS Guidance Module calculates the POS guide vector. POS guide vector controls the degree of focus on image information and caption information, using the sequential POS information obtained from the POS prediction model’s reprocessed data, the output of  $Dense_p$ . We can get richly expressive sentences with lexical diversity by considering image information and language information using POS. The POS guide vector  $(a, b)$ , serve as a probability of focusing more on the image features and caption information according to the sequential POS information. This vector can adaptively focus on the image and caption information according to the POS and predict the semantically appropriate word for the next step based on the image content and syntax. Image features obtained from the encoder are aligned in the same dimension as the caption embedding vector through  $Dense_v$ , and the caption information is vectorized through the embedding layer. The processed image vector and caption embedding vector are multiplied by the POS guide vector to focus on each piece of data according to the direct POS guidance. The vector  $x_L$  output from the POS guidance module

is calculated as

$$x_L = a \cdot I + b \cdot Emb_w. \tag{1}$$

Here,  $I$  and  $Emb_w$  denote the encoded image feature and vectorized the caption embedding information, respectively. After the multiplication of  $I$  with  $a$  and  $Emb_w$  with  $b$ , the two values are added to calculate  $x_L$ . Here,  $x_L$  is focused differently depending on the POS at each time step in the caption prediction. It can be used to generate rich expressions, and various lexical diversity, providing an accurate description by using the POS as a guide. And  $x_L$  is input to Bi-LSTM and used for learning the sequence of the generated caption.



**FIGURE 3.** Structure of proposed captioning decoder with POS multimodal layer. POS multimodal layer combines Bi-LSTM’s output and POS’s sequential vector from POS prediction model in Figure 2.

### B. CAPTION GENERATION USING POS MULTIMODAL LAYER

According to the attribute of the POSs mentioned in Section I, POS contains information about the word’s sequential information and the grammatical role of the words. We combine the sequential information of the Bi-LSTM-generated output and the sequential POS predict information to take advantage of this. However, POS is a different kind of data from predicted caption information, which is the output of Bi-LSTM. So, if we just add each data together, the dimensional complexity of the data may increase and performance is decreased. Therefore, a method to optimize dimensional complexity is required, and the proposed POS Multimodal Layer will be a solution to this problem. The POS Multimodal Layer combines and optimizes the POS prediction result and the Bi-LSTM’s output, predicted caption information. As a result, the model through this process is excellent in choosing words for the next level by considering the grammatical rules.

Figure 3 shows the structure of the POS multimodal layer. The information obtained from the POS prediction model,  $y_p$ , consists of the POS predict information, such as sequential POS information. To generate the entire caption according to the sequence of the POS, the POS multimodal layer combines



the two types of information to correct the information on the generated sentence considering the POS. As the combined information,  $y_m$ , includes information on the word predicted for the next time step as well as the POS information predicted for the current time step, the word predicted for that timestep can be corrected following the POS. Equation 2 shows the calculation of  $y_m$ ,

$$y_m = y_L + y_P. \quad (2)$$

where,  $y_L$  is a vector calculated through  $Dense_L$  to combine the output of Bi-LSTM with POS information.  $y_m$  is the word predicted for the next time step, corrected according to the POS prediction information and predicted captions. Here, injecting the POS information directly into the sentence information is equivalent to combining two types of data in different dimensions. However, this results in the problem that learning is not properly performed with a simple linear layer owing to increased complexity. To solve this complexity, we used a simple fully-connected layer ( $Dense_s$ ) via the nonlinear activation function, ReLU, for a multimodal layer. The input and output equation for the  $Dense_s$  layer is defined as in

$$y_s = ReLU(W_s \cdot y_m + b_s). \quad (3)$$

where,  $W_s$  and  $b_s$  are the weight and bias of  $Dense_s$ , respectively. The  $y_s$  output is a vector combining the sequential information of the generated sentence, combined with the sequential POS information. To efficiently process different types of high-dimensional data such as the sequential information of the generated caption and the sequential POS information, two layers of  $Dense_s$  are stacked for optimization. Finally, probability  $P(w_{t+1} | w_{1:t})$  of the word to be output at the next time step is obtained through the output layer, using Softmax activation function. The value  $P(w_{t+1} | w_{1:t})$  output through the proposed POS multimodal layer is expressed as in equation 4,

$$P(w_{t+1} | w_{1:t}) = softmax\{W_{st} \cdot y_s + b_{st}\}. \quad (4)$$

Here,  $W_{st}$  and  $b_{st}$  are the weight and bias of the Softmax layer, respectively, and  $w_{1:t}$  is the set of words output from the first output word to the current time step, in other words, the sentence obtained up to that step. Because these sentences, in which the predicted words are generated, are based on the collected sequential POS information. Thus, we can follow certain grammatical rules and generate rich vocabulary.

## IV. EXPERIMENTS AND RESULTS

### A. EXPERIMENT SETTINGS

To validate the performance of the proposed model, we used the Flickr 30K and MS COCO datasets for training and testing. Flickr 30K dataset contains a total of 31,783 images, which were divided into sets of 29,381 for training, 1,000 for verification, and 1,000 for testing. Each image had five human-made descriptive sentences, and there is a

total of 158,915 sentences. And in the MS COCO dataset, we used Karpathy's split [46]. This split had 123,287 images, of which 113,287 were used for training, 5000 for validation, and 5000 for testing. This split is widely used for validating various models. Each image had five man-made descriptive sentences, so, there are 616,435 sentences in this dataset.

For preprocessing the optimization of the dataset, vocabularies with a frequency of less than five were not used for training, and they were converted into the "unk" tokens instead. Thus, the total number of vocabularies in all datasets was 7415, including the added "unk," "startseq," "endseq," and "0" tokens, which were used as the dimensions of the embedding layer and the final output layer. In addition, the start token("startseq") was added at the beginning of each sentence and the end-of-sentence token("endseq") was added at the end of each sentence to train the start and end of the sentences, respectively. For training, all sentence data must be of the same length, so we adjust the length to 50 by zero-padding before sentences with a length less than 50 to fit the dimension of caption data. To get POS data for each word, we used the "DefaultTagger" of NLTK's POS tagger. DefaultTagger follows the Penn Treebank POS Tagset [47] defined by the University of Pennsylvania.

The proposed model was an inject-based one that simultaneously inputs image information and caption information into the model. We used Inception-ResNetV2 [48] model as the encoder to extract the image features from the images. The dimension of the output image feature vector was 1,536 and the output dimension of the embedding layer for vectorizing sentence information was set to 300, which was the same as that of all fully connected layers and Bi-LSTM layers. The dimensions of the LSTM and the fully connected layer used in the POS prediction model were set to 40, including additional tokens in the number of POSs to consider all possible POSs. The Dense layer's dimension in the POS multimodal layer is set to 300, to maintain and optimize data volume. The batch size for training was set to 128, and Adam [49] was used for model optimization. Categorical-Cross-entropy was used as the loss function for each of the sentences and POS outputs, and the loss function of the entire learning was the sum of these two loss functions. Equation 5 depicts the loss function,

$$Loss_{all} = Loss_{POS} + Loss_{Word} \times \gamma. \quad (5)$$

where,  $Loss_{all}$  is the loss function for the entire learning,  $Loss_{POS}$  is the loss function for the POS prediction output, and  $Loss_{Word}$  is the loss function for the sentence prediction output.  $\gamma$  is the variable that adjusts the ratio of the loss function obtained from the POS prediction model. Inspired by Zhang et al. [11], we aimed to transform the sentences in consideration of each time step's POS conditionally. Therefore, in this study,  $\gamma$  in (5) was set to 0.5. As for the caption generation method, we used greedy search which collected the Top 1 candidate word to output sentences to make sure using a variety of words.

**TABLE 1.** Performance comparison between the proposed model and parallel-inject model on Flickr 30K dataset.

Method	B-1	B-4	M	R	C
Parallel-Inject Model [45]	54.4	15.5	15.2	38.8	31.8
Proposed Model (LSTM)	59.0	18.5	23.5	49.9	30.8
Proposed Model (Bi-LSTM)	<b>64.8</b>	<b>22.1</b>	<b>26.3</b>	<b>54.4</b>	50.4

**TABLE 2.** Performance comparison between the proposed model and parallel-inject model on MS COCO dataset.

Method	B-1	B-4	M	R	C
Parallel-Inject Model [45]	66.7	26.5	21.9	49.3	81.8
Proposed Model (LSTM)	73.0	29.4	28.9	59.4	82.5
Proposed Model (Bi-LSTM)	<b>75.7</b>	<b>33.2</b>	<b>30.1</b>	<b>60.3</b>	<b>85.1</b>

## B. ANALYSIS OF EXPERIMENTAL RESULTS

In this study, we aimed to generate sentences of various expressions with accurate contents by considering POS. We used BLEU-1, 4, METEOR, ROUGE-L, CIDEr, and WMD as evaluation metrics for comparison. Also, we calculate the TTR metric for lexical diversity. Moreover, we checked and analyzed the generated sentences to evaluate whether richly expressive sentences were generated.

### 1) COMPARISON OF EVALUATION METRICS FOR STRUCTURAL VALIDITY

The proposed model has a structure with a POS guidance module to the encoder-decoder framework, and Bi-LSTM and multimodal layer applied as the decoder.

First, we compared the encoder-decoder framework-based formal Parallel Injection model [45] without using POS information and using the POS information; the proposed model (LSTM) to evaluate the achievement using POS. In addition, we compared LSTM and Bi-LSTM model's output sentences to compare the decoder's performance. Table 1 shows the results of the performance comparison using the Flickr 30K dataset. A large difference in evaluation metric scores was found between the parallel-inject model not using POS and the proposed model using POS. All evaluation metrics of the proposed model (LSTM) were increased from a minimum of 3.0 to a maximum of 11.1, when compared to those of the parallel-inject model. In particular, the ROUGE-L score showed the greatest increase by 11.1.

And the MS COCO dataset's test illustrated in Table 2, all evaluation metrics of the proposed model (LSTM) were increased when compared to those of the Parallel-Inject model. Especially, ROUGE-L showed the greatest increase by 10.1. It can be concluded that METEOR and ROUGE-L are significantly higher than the Parallel-Inject model. Notably, this increase shows that even long, expressive sentences are highly accurate. Moreover, more accurate sentences are generated using POS. Through the comparisons in Table 1 and Table 2, it was verified in this study that more accurate sentences can be generated using POS, and that high-dimensional data (including POS, sentences, and images) can be processed using Bi-LSTM.

### 2) COMPARISON OF EVALUATION METRICS WITH OTHER STATE-OF-ART MODELS

Here we compared the standard evaluation metrics compared to other state-of-art image captioning models in Tables 3, 4.

In Table 3, we compared evaluation metrics trained by Flickr 30K. The BLEU-1 score of Hard Attention was the highest at 66.9. Meanwhile, the METEOR and BLEU-4 scores of the Hard Attention were the lowest among those of the compared models. Hard Attention appears to focus only on the image feature, but other metrics are lower than Inject-Tag and Proposed Model, using POS. It can be seen that when using POS, we can create accurate captions while considering the sentence structure.

Table 4 compares evaluation metrics with the state-of-art models using MS COCO dataset. It can comprehensively utilize image and sentence features and use information more accurately through attention and better utilize the information when generating sentences, such as M<sup>2</sup> Transformer, X-LAN, and X-Transformer. Therefore, the BLEU metrics, which perform a simple linguistic manners comparison, have a particularly high. It can be seen that over 80 in BLEU-1, over 39 in BLEU-4. Also, CIDEr is based on TF-IDF against the n-gram captions. M<sup>2</sup> transformer is 131.2, the highest CIDEr score. Recently, there are many studies on optimization using CIDEr, so it seems to be less meaningful to compare the proposed model or Inject-Tag. Moreover, transformer-based models are computationally heavy. The proposed model has advantages in speed and calculation. Although the proposed model is lower than the latest models in BLEUs and CIDEr, it achieved the highest score in the evaluation metrics, METEOR and ROUGE-L. It can be seen that over 29.5 in METEOR and over 59 in ROUGE-L. A high score on these metrics means accurate descriptions are produced, even in long sentences because of the variety of vocabularies usages. It can be seen that POS can create captions with different word sequences for the same meaning.

Additionally, we try to compare semantic similarity by comparing the WMD metric. WMD uses Word2Vec to define the semantic distance distribution of the meaning of a word and compares the optimal distance between two words in another sentence. That is, by using the WMD, we can check

**TABLE 3. Comparison of evaluation metrics with state-of-art models on Flickr 30K dataset.**

Method	B-1	B-4	M	R	C
Google NIC [1]	63.6	20.4	24.7	52.2	40.4
m-RNN [34]	60.0	19.0	-	-	-
Hard Attention [2]	<b>66.9</b>	19.9	18.5	-	-
Inject-Tag [11]	62.6	21.3	25.7	53.3	46.3
<b>Proposed Model</b>	64.8	<b>22.1</b>	<b>26.3</b>	<b>54.4</b>	<b>50.4</b>

**TABLE 4. Comparison of evaluation metrics with state-of-art models on MS COCO dataset.**

Method	B-1	B-4	M	R	C	W
Google NIC [1]	-	29.6	25.2	52.6	85.5	0.349
LSTM-A [36]	73.9	33	25.6	54.2	98.4	-
Bottom-up [50]	74.5	33.4	26.1	54.4	105.4	-
GCN-LSTM [51]	80.5	38	28.5	58.3	127.6	-
RFNet [52]	79.1	36.5	27.7	57.3	121.9	-
SGAE [53]	80.8	38.4	28.4	58.6	127.8	-
AoA Net [15]	80.2	38.9	29.2	58.8	<b>129.8</b>	-
LBPF [54]	77.8	37.4	28.1	59.5	116.4	-
M <sup>2</sup> Transformer [39]	80.8	39.1	29.2	58.8	<b>131.2</b>	-
X-LAN [55]	<b>81.4</b>	<b>40</b>	29.4	59.2	128	-
Inject-Tag [11]	75.2	32.4	<b>29.8</b>	<b>59.9</b>	95.1	0.334
<b>Proposed Model</b>	75.7	33.2	<b>30.1</b>	<b>60.3</b>	97.6	<b>0.33</b>

the semantic similarity between captions. To get a WMD score, we measure word travel cost, denoted as

$$c_{ij} = \|z_i - z_j\|_2 \tag{6}$$

where,  $c_{ij}$  is word travel cost and meant as Euclidean distance between word  $i$  and  $j$  in the Word2Vec embedding space,  $z_i$  and  $z_j$ . This refers to the cost that one-word moves to another word. And we get the word distribution between the caption which word  $i$  is belonging and the other caption which word  $j$  is belonging. WMD uses each word's normalized Bag-Of-Words (nBOW) to the word distribution, denoted as

$$d_i = x_i / \sum_{k=1}^n x_k \tag{7}$$

where,  $d$  in equation (7) means the nBOW representation. It can be calculated using the count of the word  $i$   $x_i$ , and all amount of  $n$  word's count,  $\sum_{k=1}^n x_k$ . Let  $T \in R^{n \times n}$  is a flow matrix, where  $T_{ij}$  means the distance between word  $i$ 's distribution and word  $j$ 's distribution. So,  $T_{ij}$  and  $T_{ji}$  can be defined as

$$T_{ij} = d_i, \quad T_{ji} = d'_j \tag{8}$$

Finally, WMD metric is calculated in

$$\min \sum_{i,j=1}^n T_{ij} c_{ij} \tag{9}$$

where, WMD aims to optimizing to minimizing the equation (9) and consider  $\sum_{i,j=1}^n T_{ij}$  to  $d_i$ , WMD can follow the linear

**TABLE 5. Type-token ratio comparison in MS COCO dataset.**

Method	Whole word counts	vocabs	TTR
Test dataset	253006	9297	0.0367
Google NIC[1]	4454	254	0.0057
Inject-Tag [12]	47056	810	0.0172
Proposed Model	46665	930	0.0199

program and solve two caption's distance minimum cumulative cost. It can be seen that the smaller this value is, the closer it is semantically. NIC model, without the POS, is 0.349, the inject-tag is 0.334, and the proposed model is 0.33. It can be seen that the caption generated by the proposed model generated an expression including semantically similar words.

Through a quantitative comparison of the evaluation metrics between recent image captioning models, it was confirmed from the METEOR and ROUGE-L scores that the proposed model generates rich expressions of captions and word sequences when compared with other models. Also, by comparing WMD, our model generates captions semantically similar.

### C. COMPARISON AND ANALYSIS OF GENERATED SENTENCES

The goal of the proposed model is to generate correct sentences with rich expression and lexical diversity. For a

qualitative comparison of this, we compared and analyzed the example of captions generated by each model to confirm that they are constructed with lexical diversity. And we use the evaluation metrics called TTR. TTR is the total number of unique words divided by the total number of words in each sentence. Thus, we can check lexical diversity through TTR numerically.

In Table 5, we compare the total number of words in generated sentences, types of words used, and Type-Token Ratio was compared in MS COCO Karpathy’s split [40] Test Dataset. Here, Whole words are the total number of words in output captions and calculated by adding up all the words in test outputs. Words can check whether various vocabulary is used by categorizing words and POS. According to the results, it can be confirmed that more words are used when the POS is used. Also, the TTR and the words are the highest among other methods. It can be demonstrated by comparison in Table 5 that the proposed model can improve the lexical diversity of output captions.

Furthermore, to confirm the use of POS contains semantic information even when using various vocabulary, we try to prove this by comparing POS used with the captions made. For this comparison, we compared 8 POS categories by semantically connecting them from 36 POS of Penn Treebank POS Tag sets. We divided POS tags into Noun, Verb, ADJ, Connective words, ADV, DT, and the rest including exclamations were classified as ETC, such as Symbols, and Cardinal numbers. The graph in Figure 4, Figure 5 demonstrate the classification of POS that words used in each caption.

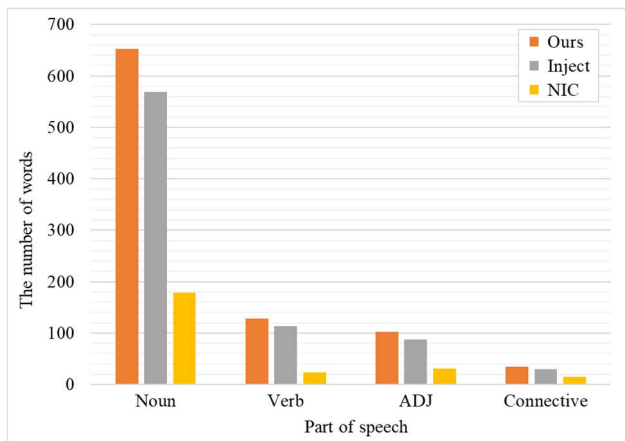


FIGURE 4. Comparison of the number of words between Noun, Verb, ADJ, and Connective words in the output of each model.

In Figure 4, the use of nouns has risen significantly using POS models. And Verbs, ADJs are also increased when using POS information. It can be seen that various words are used with the same meaning.

In Figure 5, the use of ADV, DT, ETC words is also risen using POS models. Specifically, when NIC captions are not using ADV, but using POS models are generated ADV words. ADV modifies verbs and ADJs in sentences, reinforcing the meaning of words. Therefore, it can be confirmed that by

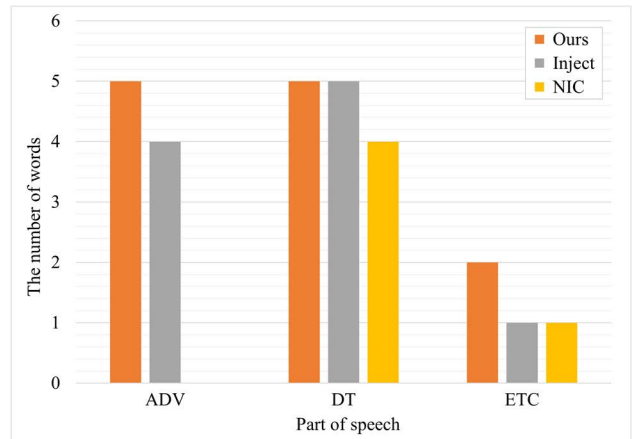


FIGURE 5. Comparison of the number of words between ADV, DT, ETC in the output of each model.

using the POS, sentences with various contents were created using words of rich expression.

Figure. 6 shows the sentences generated by each model for the given images to compare whether the sentences generated by the implemented models provide an accurate description. The sentences generated by the implemented Google NIC, the Inject-Tag, and the proposed model was compared with the ground truth sentences.

In the sentences generated by each model for each image, all captions have detected the object used for the subject well. In Figure. 6(a), the proposed model’s caption has more accurate like GT sentences by explaining the man’s state, ‘in the air over the snow.’ Also, in Figure. 6(b), the proposed model’s captions are more detailed and reinforced statements about the object, such as ‘a bag of clothes.’ These modifiers provide a more detailed description of the type and state of an object. In Figure 6(b), GT caption’s ‘luggage’ is the bag where the dog sitting, and the proposed model caption describes the luggage as ‘a bag of clothes.’ This word is semantically consistent with bag and is an element that can confirm that it is lexically rich. And Figure 6(c), It can be seen that considering the POS more directly, information around the subject that other models could not describe was expressed in more detail, like ‘on a blue surfboard.’ In Figure 6(d) can be viewed as an example of the above. By predicting the word ‘tooth brush,’ It made detailed captions that more express the surrounding information of the subject.

It was also confirmed that the proposed model’s captions use more diverse connective words to generate richly expressive sentences. In Figure. 6(e), the proposed model uses the expression “leaning against” for an image of a bicycle parked on an iron bar on the side of the road to generate a description of various contents. Figure. 6(f) shows a scene in which people watch a person playing a video game. Other models described only the person playing a game, but the proposed model accurately described that others were watching with a structure using “while.” It is clear that the proposed model can generate rich explanations using more diverse connective






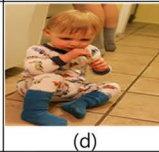



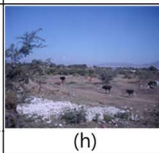
 <p>(a)</p>	<p>GT: A man flying through the air while riding a snowboard.                  NIC: A man riding a snowboard down a snow covered slope.                  Inject-tag: A man is snowboarding down a snowy hill.                  Proposed: A snowboarder is <b>in the air over the snow</b>.</p>	 <p>(b)</p>	<p>GT: There is a big dog that is sitting inside of a luggage.                  NIC: A dog laying on top of a bed.                  Inject-tag: A dog laying on a bed with a blanket.                  Proposed: A dog is sitting in a suitcase with <b>a bag of clothes</b>.</p>
 <p>(c)</p>	<p>GT: The surfer rides a wave on a blue surfboard.                  NIC: A man in a blue shirt is riding a wave.                  Inject-tag: Man in a wetsuit is surfing in the ocean.                  Proposed: A man in a wetsuit is surfing on a <b>blue surfboard</b></p>	 <p>(d)</p>	<p>GT: A toddler in pajamas and socks cleaning his teeth.                  NIC: A little girl sitting on a couch with a teddy bear.                  Inject-tag: A baby is holding a remote control in his hand.                  Proposed: A baby is brushing her teeth with a <b>tooth brush</b>.</p>
 <p>(e)</p>	<p>GT: A bicycle that has been locked onto metal bars.                  NIC: A red fire hydrant sitting on the side of a street.                  Inject-tag: A bicycle is chained to a pole on a sidewalk.                  Proposed: A bike leaning <b>against</b> a post on a side walk.</p>	 <p>(f)</p>	<p>GT: A man and woman watching a man play a video game.                  NIC: A group of people playing a video game.                  Inject-tag: A man is playing a video game with a wii controller.                  Proposed: A man is playing a video game <b>while</b> another man watches.</p>
 <p>(g)</p>	<p>GT: People carrying surfboards walking across a beach next to the ocean.                  NIC: A man riding a wave on top of a surfboard.                  Inject-tag: A man riding a surfboard on top of a wave.                  Proposed: A man is walking <b>along</b> the beach with a surfboard</p>	 <p>(h)</p>	<p>GT: Steers are walking on a dirt path through a barren terrain                  NIC: A group of people standing on top of a beach.                  Inject-tag: A herd of cattle grazing on a lush green hillside.                  Proposed: A group of cows walking <b>along</b> a dirt road.</p>

FIGURE 6. Comparison of examples of image captions by using NIC, Inject-tag and the proposed model.

words. In addition, in Figure. 6(g) and Figure. 6(h), using the preposition of “along” in the description of each image, an accurate description of the subject’s surrounding environment and an accurate sentence were described in terms of content. Moreover, the generated sentences are written in a structure that is easy to understand by people who follow certain grammatical rules. In this way, if the POS is directly considered in sentence generation, it is possible to obtain a sentence with accurate content, word order, and rich expression. The sentences generated by the proposed model were compared with those of the Parallel-Inject model to prove that the direct use of POS for sentence generation would result in sentences with more accurate grammatical structures and with more abundant contents.

As reported in Section IV-B, the sentence of the proposed model using POS can generate much more precise captions. Also, by using Bi-directional structures, we can process POS and caption information simultaneously. And the accuracy of sentence generation and abundance of expressions were confirmed by comparing the proposed model with the latest models, demonstrating its potential to compete with state-of-art models. Specifically, the proposed model showed a high increase in the METEOR and ROUGE-L scores, indicating that the model generated sentences containing many word sequences using rich expressions of captions and word sequences. And in WMD metrics, we found that using POS generates sentences with more semantic similarity. In Section IV-C, we compare real generated captions and use TTR, which compares the diversity of vocabulary with the content of actually generated sentences. By comparing TTR metrics, it can be seen that the proposed model can improve the lexical diversity of output captions. Also, example

captions are demonstrated that these sentences with certain grammatical rules and can enhance lexical diversity using various modifiers and conjunctions as well as accurate content.

## V. CONCLUSION

In this paper, we proposed a POS Guidance Module and a multimodal-based image captioning model to generate sentences with rich expressions for enhancing lexical diversity. The proposed image captioning model focuses more on specific pieces of information among the image and sentence information according to direct POS guidance, using the POS Guidance Module. Then, the predicted output from Bi-LSTM is combined with POS information through a POS multimodal layer to generate captions with various expressions and certain grammatical rules. We can get more lexically diverse captions by directly injecting the POS.

The proposed model was compared with other models that learned Flickr 30K and MS COCO datasets based on the objective image captioning evaluation metrics such as BLEU, METEOR, ROUGE-L, and CIDEr. Especially, the METEOR, ROUGE-L, and the WMD metrics of the proposed model showed a high increase respectively for each dataset, suggesting that accurate descriptions containing similar word sequences were generated. In addition, we compare lexical diversity and semantic similarities using real-generated caption and caption’s TTR. It can be seen that the generated captions have lexical diversity and use various vocabularies with high semantic similarity.

This image captioning model presented in this study is expected to enable the generation of sentences in an accurate structure and rich representation with an abundance

of expressions for wide commercialization in the fields requiring analysis of given images, such as medical care, image summary and surveillance.

## ACKNOWLEDGMENT

The work presented in this article is based on J. W. Bae's M.S. Dissertation ("A Study on POS Guidance Module and Multimodal-Based Image Captioning Model," Graduate School of Korea Maritime and Ocean University).

## REFERENCES

- [1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 652–663, Apr. 2017, doi: 10.1109/TPAMI.2016.2587640.
- [2] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2015, pp. 2048–2057.
- [3] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4651–4659.
- [4] X. Li and S. Jiang, "Know more say less: Image captioning based on scene graphs," *IEEE Trans. Multimedia*, vol. 21, no. 8, pp. 2117–2130, Aug. 2019.
- [5] A. U. Haque, S. Ghani, and M. Saeed, "Image captioning with positional and geometrical semantics," *IEEE Access*, vol. 9, pp. 160917–160925, 2021.
- [6] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2010, pp. 15–29.
- [7] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *J. Artif. Intell. Res.*, vol. 47, no. 1, pp. 853–899, 2013.
- [8] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2Text: Describing images using 1 million captioned photographs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011, pp. 1143–1151.
- [9] Y. Yang, C. Teo, H. Daumé III, and Y. Aloimonos, "Corpus-guided sentence generation of natural images," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2011, pp. 444–454.
- [10] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "BabyTalk: Understanding and generating simple image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891–2903, Dec. 2013.
- [11] J. Zhang, K. Mei, Y. Zheng, and J. Fan, "Integrating part of speech guidance for image captioning," *IEEE Trans. Multimedia*, vol. 23, pp. 92–104, 2021.
- [12] J. H. Lee, S. H. Lee, S. H. Tae, and D. H. Seo, "Parallel injection method for improving descriptive performance of Bi-GRU image captions," *J. Korea Multimedia Soc.*, vol. 22, no. 11, pp. 1223–1232, 2019.
- [13] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–36, Nov. 2019.
- [14] J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully convolutional localization networks for dense captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4565–4574.
- [15] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4634–4643.
- [16] W. Wang, Z. Chen, and H. Hu, "Hierarchical attention network for image captioning," in *Proc. 32nd AAAI Conf. Artif. Intell.*, Jul. 2019, pp. 8957–8964.
- [17] S. Wang, L. Lan, X. Zhang, G. Dong, and Z. Luo, "Cascade semantic fusion for image captioning," *IEEE Access*, vol. 7, pp. 66680–66688, 2019.
- [18] W. Jiang, X. Li, H. Hu, Q. Lu, and B. Liu, "Multi-gate attention network for image captioning," *IEEE Access*, vol. 9, pp. 69700–69709, 2021.
- [19] X. He, B. Shi, X. Bai, G.-S. Xia, Z. Zhang, and W. Dong, "Image caption generation with part of speech guidance," *Pattern Recognit. Lett.*, vol. 119, pp. 229–237, Mar. 2019.
- [20] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
- [21] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2641–2649.
- [22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. Zürich, Switzerland: Springer*, 2014, pp. 740–755.
- [23] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2001, pp. 311–318.
- [24] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, 2005, pp. 65–72.
- [25] C. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*, 2004, pp. 74–81.
- [26] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2016, pp. 382–398.
- [27] M. Kusner, Y. Sun, N. Kolkun, and K. Weinberger, "From word embeddings to document distances," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 957–966.
- [28] M. C. Templin, "Certain language skills in children: Their development and interrelationships," *Inst. Child Welfare, Univ. Minnesota, Minneapolis, MN, USA, Tech. Rep.*, vol. 10, 1957.
- [29] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.
- [30] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
- [31] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 3104–3112.
- [32] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. Interspeech*, Sep. 2010, vol. 2, no. 3, pp. 1045–1048.
- [33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
- [34] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (M-RNN)," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–17.
- [35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [36] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4894–4902.
- [37] J. Guan and E. Wang, "Repeated review based image captioning for image evidence review," *Signal Process., Image Commun.*, vol. 63, pp. 141–148, Apr. 2018.
- [38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
- [39] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10578–10587.
- [40] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10327–10336.
- [41] H. Ge, Z. Yan, K. Zhang, M. Zhao, and L. Sun, "Exploring overall contextual information for image captioning in human-like cognitive style," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1754–1763.
- [42] L. Ke, W. Pei, R. Li, X. Shen, and Y. W. Tai, "Reflective decoding network for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8888–8897.
- [43] X. Yang, H. Zhang, and J. Cai, "Learning to collocate neural modules for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4250–4260.

[44] Y. Yi, H. Deng, and J. Hu, "Improving image captioning evaluation by considering inter references variance," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Jul. 2020, pp. 985–994.

[45] M. Tanti, A. Gatt, and K. P. Camilleri, "Where to put the image in an image caption generator," *Natural Lang. Eng.*, vol. 24, no. 3, pp. 467–489, May 2018.

[46] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3128–3137.

[47] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of English: The Penn treebank," *Comput. Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[48] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-V4, inception-ResNet and the impact of residual connections on learning," in *Proc. 31st. AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[50] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.

[51] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 684–699.

[52] W. Jiang, L. Ma, Y. G. Jiang, W. Liu, and T. Zhang, "Recurrent fusion network for image captioning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 499–515.

[53] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10685–10694.

[54] Y. Qin, J. Du, Y. Zhang, and H. Lu, "Look back and predict forward in image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8367–8375.

[55] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10971–10980.



**SOO-HWAN LEE** received the B.S. and M.S. degrees in electrical and electronics engineering from Korea Maritime and Ocean University, South Korea, in 2016 and 2018, respectively, where he is currently pursuing the Ph.D. degree in electrical and electronics engineering. His research interests include positioning systems, sense networks, and embedded signal processing.



**WON-YEOL KIM** received the B.S., M.S., and Ph.D. degrees in electrical and electronics engineering from Korea Maritime and Ocean University, South Korea, in 2014, 2016, and 2021, respectively. He worked with the Artificial Intelligence Convergence Research Center for Regional Innovation, Korea Maritime and Ocean University. His research interests include positioning systems, sense networks, and embedded signal processing.



**JU-HYEON SEONG** received the B.S., M.S., and Ph.D. degrees in electrical and electronics engineering from Korea Maritime and Ocean University, South Korea, in 2014 and 2016, respectively. He is currently an Associate Professor with the Department of Liberal Education, Korea Maritime and Ocean University. His research interests include positioning systems, sense networks, and embedded signal processing.



**JU-WON BAE** received the B.S. degree in ocean engineering and the M.S. degree in electrical and electronics engineering from Korea Maritime and Ocean University, South Korea, in 2019 and 2021, respectively, where he is currently pursuing the Ph.D. degree in electrical and electronics engineering. His research interests include image captioning, computer vision, and anomaly detection.



**DONG-HOAN SEO** received the B.S., M.S., and Ph.D. degrees in electronic engineering from Kyungpook National University, South Korea, in 1996, 1999, and 2003, respectively. Since 2004, he has been with Korea Maritime and Ocean University, where he is currently a Professor with the Division of Electronics and Electrical Information Engineering. His research interests include sense networks, signal processing, and computer vision.

...