IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 21, 2019, accepted April 3, 2019, date of publication April 16, 2019, date of current version May 3, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2911656

Multimodal Neural Machine Translation With Weakly Labeled Images

YOONSEOK HEO[1], SANGWOO KANG[2], AND DONGHYUN YOO[3]

[1]Department of Computer Science and Engineering, Sogang University, Seoul 04107, South Korea
[2]Department of Software, Gachon University, Gyeonggi-do 13120, South Korea
[3]NCSOFT Corporation, Seongnam-si 13494, South Korea

Corresponding author: Sangwoo Kang (swkang@gachon.ac.kr)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Science and ICT, under Grant NRF-2019R1C1C1006299.

ABSTRACT Machine translation refers to a fully automated process that translates a user's input text into a target language. To improve the accuracy of machine translation, studies usually exploit not only the input text itself but also various background knowledge related to the text, such as visual information or prior knowledge. Herein, in this paper, we propose a multimodal neural machine translation system that uses both texts and their related images to translate Korean image captions into English. The data in the experiment is a set of unlabeled images only containing bilingual captions. To train the system with a supervised learning approach, we propose a weak-labeling method that selects a keyword from an image caption using feature selection methods. The keywords are used to roughly determine an image label. We also introduce an improved feature selection method using sentence clustering to select keywords that reflect the characteristics of the image captions more accurately. We found that our multimodal system achieves an improved performance compared to a text-only neural machine translation system (baseline). Furthermore, the additional images have positive impacts on addressing the issue of under-translation, where some words in a source sentence are falsely translated or not translated at all.

INDEX TERMS Human–computer interaction, multi-layer neural network, natural language processing, image classification, multimodal neural machine translation, weak label.

I. INTRODUCTION

Recent advances in deep learning have made it possible to handle a number of artificial intelligence-related tasks such as natural language processing (NLP), computer vision, and signal processing. Machine translation, commonly known as MT, is one of the most challenging NLP tasks and is best addressed with a deep-learning approach. MT usually refers to automatic translation for texts from a source (original) language into a target language without human intervention. Traditional approaches to MT are generally divided into two categories: rule-based approaches and statistical approaches.

The rule-based approach was the very first solution developed in the field. A rule-based MT system [1], [2] strongly depends on human-generated linguistic knowledge and therefore can precisely translate the input sentence if the predefined rules are exactly applied to it. The key limitations of the approach are the necessity of the expertise needed to create the rules and the vast number of rules and linguistic resources required for the system.

On the other hand, the statistical approach [3]–[6] is data-driven, requiring only a large bilingual corpus. This means that linguists are not required to develop the translation rules. In addition, the statistical model can generate various types of translations for one sentence. However, certain core issues still need to be addressed: data sparseness and an inability to capture the overall semantic information contained in a sentence.

However, several neural approaches [7]–[11] to MT are good at learning semantic representations and modeling a wide context without severe data sparseness. At first, the translation task can be formulated in a sequence-to-sequence framework [12], [13] consisting of two neural networks called an encoder and a decoder. An input sentence is represented by a single vector using the encoder. This vector, called the "context vector," is considered to imply the

The associate editor coordinating the review of this manuscript and approving it for publication was Dong Wang.

overall meaning of the sentence. Then, for every decoding step, the decoder determines a word based on the context vector and the words that were generated in the previous step, ultimately generating one translated sentence. Furthermore, since the neural-based approach considers the whole context of the sentence, it can solve the problem of word reordering that occurs frequently in translation of sentences with complicated structures. The performance of this approach with regards to long sentences has proven to be robust if the encoder and the decoder consist of recurrent neural networks (RNNs) using long short term memory (LSTM) [14] or a gated recurrent unit (GRU) [15].

However, serious problems remain in neural-based MT (NMT) systems, such as over-translation and under-translation [16]. Over-translation refers to when some words are duplicated in the translation of the source sentence. In contrast, under-translation occurs when words in the source sentence are mistakenly or falsely translated. Error analysis of the translations generated from conventional attention-based NMT models [17], [18] has shown that the occurrences of under-translation are comparatively higher than those of over-translation [16]. Therefore, we have studied a multimodal MT system that utilizes supplementary resources in the form of images, going beyond the existing text-only MT studies to overcome under-translation issues and to improve the quality of the translations. Recent studies have applied multimodality to MT and can be found in the Workshop on Statistical Machine Translation 2016 (WMT16) shared task [19]. The task uses data that extends the Flickr30K Entities dataset [20], where the entities in the image are labeled. Each image has its own caption, each of which is a source sentence (e.g., in Korean) and a target sentence (e.g., in English) that a person translated.

However, the WMT shared task is constrained by the fact that each image must have its own label concerning the object that appears in the image. Labels are necessary to train the model to generate feature representations for the image with a supervised learning approach. For resource-poor languages like Korean, bilingual caption data with object-labeled images are scarce. Therefore, in this paper, we propose a multimodal NMT system that translates Korean captions into English using unlabeled images. The main contributions are as follows: To the best of our knowledge, we are the first to propose a multimodal MT from Korean to English. In addition, we propose a weak-labeling method to roughly determine a label on the image as a keyword chosen from its caption using a feature selection methodology to train the convolution neural network (CNN) used to obtain information from images with a supervised learning approach. Moreover, we also propose an improved feature selection method using a K-means algorithm to select keywords that can distinguish the semantic differences between captions as much as possible.

The remainder of this paper is as follows. Section II briefly discusses the multimodal NMT and the two feature selection methods. Section III describes the methodology for determining weak labels. Section IV introduces our proposed multimodal NMT system. Section V discusses our experimental setup and analyzes our results. Finally, we draw conclusions in Section VI.

II. RELATED WORK

This section consists of two parts. The first briefly introduces the multimodal NMT and the second describes the two feature selection methods, bag-of-words (BoW) and term frequency and inverse document frequency(TF-IDF).

A. MULTIMODAL NEURAL MACHINE TRANSLATION

Motivation to combine multimodality with translation tasks arises from a considerable amount of work on visual recognition and image caption generation. Visual recognition detects an object in an image, and it has been actively studied with research efforts like the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [21]. In this challenge, deep CNNs such as VGG19 [22] and ResNet [23] showed considerable performances for obtaining effective visual features from images.

Image caption generation refers to research on how to generate a description of an image from the image itself [24], [25]. The task is generally approached with a sequence-to-sequence framework. The encoder mainly uses VGG19, which demonstrated a state-of-the-art performance in ILSVRC for extracting information from input images. The decoder usually consists of an RNN with an attention mechanism. These two types of studies provide a sufficient foundation for combining multimodality with MT.

Based on the results of the aforementioned studies, there have been several research efforts that incorporated text-based MT with visual information. First, there is a WMT shared task [19], which is a competition focused on multimodal MT that analyzes how image functions improve translation quality. Huang et al. [26] proposed an attention-based multimodal MT model that can incorporate two types of visual features extracted from VGG19 and region-based CNN [27]. Calixto et al. [28] proposed an attentive encoder-decoder variant with a conditional GRU https://github.com/nyu-dl/ dl4mt-tutorial/blob/master/docs/cgru.pdf. The system in this study adds an attention mechanism to the image, whereas the standard NMT model only includes an attention structure for text. In previous WMT tasks, purely neural-based multimodal systems did not show improvement compared with text-only NMT or statistical MT models. Notably, visual information could not fully contribute to the translation quality. However, this study showed reasonable performance improvement for multimodal NMT for the first time.

Nonetheless, the method in this study is still constrained by its strong reliance on a large-scale parallel corpus with labeled images. CNNs are widely used to learn meaningful representation of the images with object labels to obtain meaningful information [22], [29]. However, according to a study by [30], not all images have labels that can be determined to be explicitly true, even if the images used in the benchmark set are human-annotated. The image labels tend to be determined by objects within the image with high importance. The images in this work have no proper labels. Therefore, we define a "weak label" on the image as a keyword chosen from its caption and propose a weak-labeling method using two feature selection methods, BoW and TF-IDF, which are described in the following subsection.

B. FEATURE SELECTION TECHNIQUES IN NATURAL LANGUAGE PROCESSING

The quality of features for data representation has a critical impact on the performance of machine-learning-based models. Even for different kinds of tasks that use the same data, the types of feature must be designed to suit the specific task's purpose. Also, a large number of features that represent the data does not necessarily guarantee a high model performance, and it can sometimes even cause performance degradation. Therefore, selecting which features will represent the data and which features are suitable for solving a given task has always been an important research topic in machine learning. In this section, we introduce the BoW and TF-IDF techniques, which are frequently used in the NLP domain.

The BoW method is a way of representing a text (sentence or document) based on frequency. A text can be defined as a set of frequencies of words within the text, or a so-called "bag-of-words." Then, the degree of occurrences of each word is used as a feature. This technique assumes that the more frequently a word appears in a text, the more information it represents. Therefore, the high-frequency words are selected as features that characterize the document. With its ease of understanding, this method has seen great success in NLP tasks such as document classification [31] and email filtering [32].

Although term frequency has its strengths, its premise can also be regarded as a significant weakness. For example, common words like articles or pronouns are almost always terms with the highest frequencies in the text. Consequently, the method comes to an inappropriate conclusion that such words are the features that principally characterize the text. To address this issue, a new approach called TF-IDF [33] has been proposed that integrates term frequency (TF) with an additional factor called the inverse-document frequency (IDF).

III. WEAKLY LABELED IMAGES FOR TRAINING A CONVOLUTIONAL NEURAL NETWORK

The proposed model consists of two parts, as shown in Fig. 1: a CNN and an attention-based sequence-to-sequence (Seq2Seq) network [17]. The CNN is used to obtain visual features from the input images. Specifically, we use a 19-layer VGG network (VGG19) [22], which displayed a considerable performance in the ImageNet challenge in 2014. The attention-based Seq2Seq model incorporates a source sentence with visual information extracted from the VGG19 and generates a target sentence. In this section, we only concentrate on the VGG19.

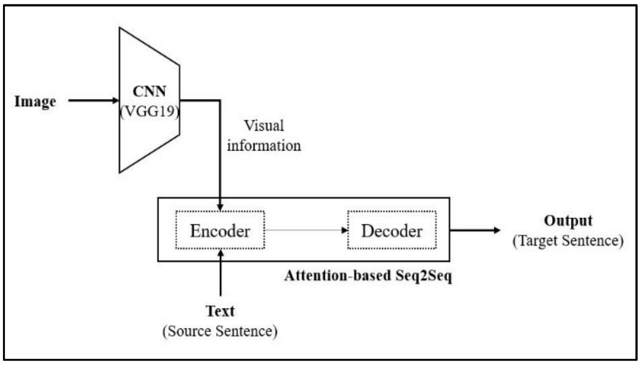


FIGURE 1. Abstract description of the proposed model.

The primary purpose of the VGG19 is to classify an object appearing in the image. However, the overall training procedure for the model is to calculate the cross-entropy loss function between the output from the decoder and its corresponding human-translated sentence. If the CNN is trained using the calculated loss from the translated result, the information of the propagated loss can be considered ambiguous from the viewpoint of the CNN, which must generate the features fully implying the input image. This is because the errors both in the translation process and in the image feature generation are mixed. In order to generate better features from the image, the VGG19 is pre-trained with a supervised learning approach for object classification.

However, the images in the dataset have no proper object labels. Instead, the data is composed of a set of images, each of which has only two pairs of captions, one in Korean and one in English. Therefore, we propose a weak-labeling method, wherein the label on an image is roughly determined by a keyword from its caption. The keyword is defined as the best descriptive word in the current caption compared to other captions. The keyword is chosen by the feature selection techniques in the NLP. The following sections describe this weak-labeling method, which uses the two feature selection techniques described previously, BoW and TF-IDF. In addition, we further propose an improved feature selection technique using a K-means clustering algorithm to select keywords by maximizing the semantic differences between different image captions.

A. WEAK LABELING METHOD WITH FEATURE SELECTION

A label for an image is generally determined by the object within the image that contains the largest amount of information. Taking Fig.2 as an example, the objects in the image can include "giraffe," "tree," and, to some extent, "forest" when bound to a set of trees. Between these three objects, the representative object in the image is "giraffe," since it represents the largest amount of information in the image. Therefore, the label for the image in Fig.2 can be determined as "giraffe." The image annotation becomes clearer upon considering the English captions. The key subject that the

Caption 1

Source1(Korean)	나무들로 된 숲 한가운데 서 있는 기린
Target1(English)	A giraffe standing in the middle of a forest with trees

Caption 2

Source(Korean)	많은 종류의 나무들 옆에 서 있는 기린
Target(English)	A giraffe standing next to many types of trees

FIGURE 2. An example of image annotations.

two English captions share is "giraffe," which is the most important keyword in each caption.

However, for all images, it is not possible to select one representative object from the image as the label, like it was possible to do in Fig.2. According to [30], some of the human-annotated labels used in the benchmark set have considerable ambiguity depending on the subjectivity of the person. Therefore, when determining a label from any of the objects in an image, the global visual feature obtained from CNN can still be considered as representing the entire image if it can guarantee the validity to some degree.

In this paper, the keyword from the image caption is determined to be the label of the image. Since captions typically describe the most represented object in an image, it can be assumed that the keyword of the caption is generally a representative object of the image. The method for selecting the keyword from all the words in the caption is based on BoW and TF-IDF, the two kinds of feature-selection methodologies used in NLP. Likewise, a label that is heuristically generated and not human-annotated is called a "weak label," and an image that has a weak label is called a "weakly labeled image." Next, to understand the weak-labeling method, we must understand how to select a keyword among all the terms in a caption using feature selection.

Prior to the introduction of the methodology, we must first define a feature set that expresses a caption to apply the feature selection method. One of the critical indicators for deciding the quality of caption translation is whether the objects in the image described in the caption are completely represented or not. Thus, the keyword of the caption is very likely to be one of the objects in the image. In this case, objects are mostly nouns. Therefore, in this paper, we define a feature set representing a caption as all the nouns in all the captions in the training data. In addition, the label on the image is defined as the keyword of the target sentence of the caption, not of the source sentence. The reason for this is to learn very direct information related to the translation itself, whose anticipated result is the target language caption.

The following sections describe the weak labeling method using BoW and TF-IDF for feature selection.

1) KEYWORD SELECTION USING BAG-OF-WORDS

BoW is a feature selection method based on the frequency of each feature. Thus, the weight of each feature is determined by the frequency of the feature, where the feature with the highest frequency is evaluated as a crucial factor. Using BoW to label an object on an image proceeds as follows.

First, among the entire feature sets, the frequency of noun words(features) appearing in the target description is calculated. Then, the term with the highest frequency is set as a keyword representing the caption, and finally, the keyword is set as a weak label on the image. It is very rare in this field to find a word appearing more than once in a caption, except in compound sentences containing various modifiers. In other words, most features exist only once. Therefore, in this paper, we concatenate two target captions for one image and then calculate the feature vector using BoW.

Suppose that a feature set is defined as a 5-dimensional vector consisting of "giraffe," "forest," "trees," "sky," and "water." Then, if the feature vector has the following elements: "giraffe": 2, "forest": 1, "trees": 2, and "sky" and "water": 0, then either "giraffe" or "trees" can be used as the label on the image under a uniform distribution.

2) KEYWORD SELECTION USING TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY

TF-IDF is a feature selection method widely used in the information retrieval domain, such as for document classification [34], [35], and is generally used to calculate the weights of the words constituting a document. The TF-IDF of a word is defined as the product of term frequency and inverse document frequency of a word, which takes into account both the frequency of its occurrence in the document and the amount of information that the word contains within the entire document set. In this paper, TF-IDF is used as a method for selecting keywords from captions.

In this work, images have two different captions, each of which is represented bilingually. Like the document classification task, an image is defined as a category and is assumed to be a document composed of two captions. Then, an image can be expressed as a set of TF-IDF values for the entire feature set, which is defined as all the nouns of all the captions in the training data. The TF for each feature is defined as the frequency of its occurrence in the image. The IDF for each feature is inversely proportional to the number of other images containing the feature. Thus, the TF-IDF for each value is defined by the below equations. The feature with the highest TF-IDF is selected as a representative keyword, and finally, it is defined as the weak label of the image.

$$\mathrm{TF}-\mathrm{IDF}\,(\mathrm{f},\mathrm{m}) = TF\,(f,m) * IDF(f,M) \tag{1}$$

$$\mathrm{TF}\,(\mathrm{f},\mathrm{m}) = \log(1+\mathrm{freq\ \ of\ \ feat(f)\ \ in\ \ img(m)}) \tag{2}$$

$$\mathrm{IDF}\,(\mathrm{f},\mathrm{M}) = \log(\frac{|M|}{1+\{m \in M|f \in m\}|}) \tag{3}$$

where f denotes the features representing the image, m denotes an image, and $|M|$ represents the number of all the images in the training data.

The difference between BoW and TF-IDF comes from the IDF. The IDF is defined by (3) and refers to the amount of information that a feature contains within the whole image set. If a feature occurs frequently over the whole image set, the feature cannot be regarded as a key feature that characterizes the current image. That is, the amount of information that the feature has is small over the entire set of images, and its IDF value becomes small. Therefore, its TF-IDF value is small if its TF value is similar to that of other features. Likewise, a feature with a higher TF-IDF value can be interpreted as a distinctive feature representing the current image.

B. AN IMPROVED FEATURE SELECTION METHOD WITH SENTENCE CLUSTERING BY K-MEANS ALGORITHM

In general, when using TF-IDF for tasks such as document classification, the number of categories is only about a dozen. However, in this paper, since one image is defined as one category, the total number of categories is 7,500, which is extraordinarily large compared with other tasks. Therefore, when calculating the IDF using (3) in section III.A.2, the possibility that a feature exists in other images is considerably high, which causes the denominator of the IDF value to increase. As a result, most of the features of the image have IDF values of almost zero. Thus, the TF-IDF values become extremely zero. That is, such features become meaningless, resulting in a failure to select a differentiated keyword. Therefore, in this paper, we propose an improved TF-IDF calculation method by grouping semantically similar captions into one cluster using a K-means algorithm [36] and redefining one cluster as one category.

First, the TF value of each feature is defined as the frequency of each feature in an image, as in the previous method. However, the IDF value is defined to be inversely proportional to the number of clusters containing the feature, not to the number of images. The TF-IDF value of all the features for an image can be calculated as shown in (4) below. Then, the feature with the largest TF-IDF value is selected as the representative keyword and defined as the weak label of the image. The improved TF-IDF calculation method proposed in this paper is as follows:

$$\mathrm{TF-IDF\,(f, m, C)} = TF\,(f, m) * IDF(f, C) \tag{4}$$

$$\mathrm{TF\,(f, m) = log(1{+}freq\ \ of\ \ feat(f)\ \ in\ \ img(m))} \tag{5}$$

$$\mathrm{IDF\,(f, C)} = \log\Big(\frac{|M|}{1 + \{m \in C | f \in c\}|}\Big) \tag{6}$$

where f denotes the features representing the image, m denotes an image, c denotes a cluster, and $|M|$ is the number of images in the entire clusters.

To apply the improved TF-IDF method, clusters between semantically similar captions should be processed in advance. First, a sentence embedding model is required to express a caption as a single vector. The sentence embedding of a caption is defined as the average word embeddings of each word constituting the caption. Assuming a caption to be a set of words with length T, a sentence-embedding vector ($\vec{S}$) for the caption can be represented as follows:

$$\vec{S} = \frac{1}{T}\sum_{i=0}^{T-1}\vec{w}_i = \frac{1}{T}(\vec{w_0} + \vec{w_1} + \ldots + \vec{w_{T-1}}) \tag{7}$$

where $\vec{w}_i$ denotes a word-embedding vector for the i^{th} word in a caption and $\vec{S}$ denotes the sentence-embedding vector for the caption. The pre-trained Glove [37] is exploited for the embedding of words.

Next, sentence-embedding vectors for each caption are grouped into clusters with semantic similarity using a K-means algorithm. We set the number of clusters to 20 since the performance was robust. The experimental results are shown in Section V.C. Thus, captions with semantic similarity are grouped into one cluster, and eventually new categories can be redefined in cluster units.

The improved IDF calculation using (6) preserves the amount of information that each feature has over the whole image set as in the conventional IDF calculation. In addition, the number of categories to be classified is adjusted to the number of clusters grouped together by semantically similar captions to compensate for the problem of extremely small IDF values, resulting in reasonable IDF values for features. As a result, the improved TF-IDF method can select keywords that better distinguish the characteristics of captions with different meanings.

C. PRE-TRAINING

We obtained three supervised VGG models that were trained independently with weakly labeled images generated by three feature selection methods. The total numbers of weak labels on the training images generated by each method were 163 (BoW), 757 (TF-IDF), and 1020 (improved). All the layers in the VGG19 models, except the softmax layers, which are dependent on the number of weak labels in the image, were initialized with parameters trained by ImageNet. The softmax layers were designed to match the number of weak labels in the training data. Then, each of the VGG19 models can be trained with a supervised learning approach to predict weak labels for the input images.

IV. NETWORK ARCHITECTURE FOR MULTIMODAL MACHINE TRANSLATION

In this section, we propose a multimodal NMT system that exploits the information from the image as an additional resource when translating the source language caption into the target language caption. As shown in Fig.3, the proposed model consists of two parts. One is a 19-layer VGG network (VGG19) pre-trained with the weakly labeled images mentioned in the previous section, which contributes to the creation of additional resources from the image needed for translation. The other is an attention-based sequence-to-sequence model that includes an encoder with two additional cells added to the conventional attention-based models [17] to combine additional information generated from the VGG19 with the text. One of the additional cells, called the image encoding cell, integrates the global visual features of the image with the encoder. The other, called the object encoding cell, feeds into the model a set of words that are highly anticipated to be generated when translating the source-language caption. As a result, the system

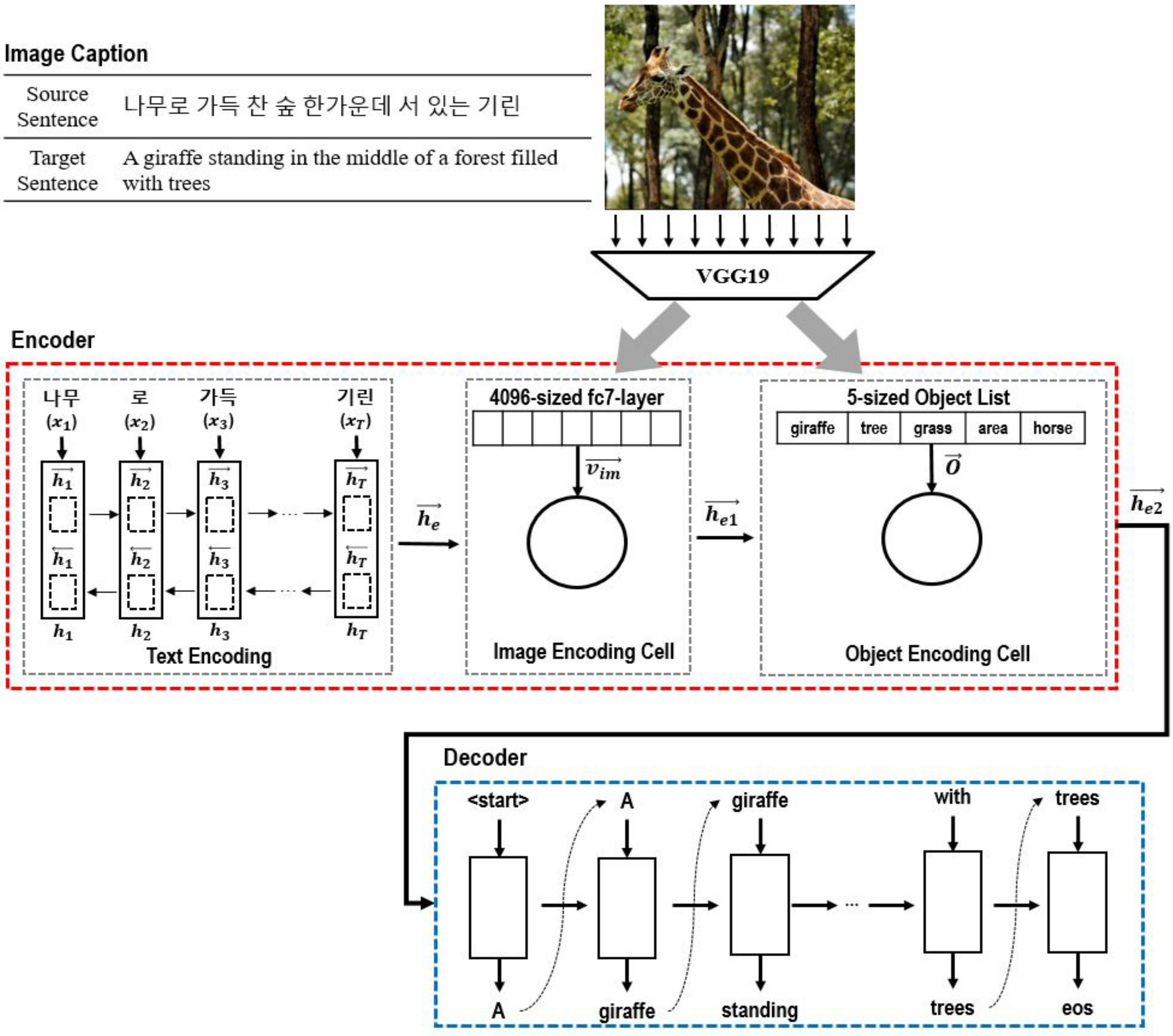


FIGURE 3. Network architecture for multimodal neural machine translation. Attention mechanism in the decoder is omitted for clarity.

is more likely to capture the words that the existing NMT model [17], [18] fails to generate, thereby improving the quality of the translation. The following sections describe the VGG19, the three encoding cells, and the attention-based decoder.

A. CONVOLUTIONAL NEURAL NETWORK FOR EXTRACTING INFORMATION FROM IMAGES

The VGG19 in Fig.3 has two roles: one is to generate a feature vector that represents the image, and the other is to generate the label of the image to be directly used as the keyword for the sentence to be translated.

The VGG19 is a deep CNN that includes 16 convolutional layers, 5 max-pooling layers, and 2 fully connected layers. It can effectively learn rich feature representations from images [22]. In this work, the global visual features for each image are represented by a 4096-sized feature vector extracted from the so-called fc7 VGG19 layer. Then, this vector is fed into the encoder, especially the image encoding cell explained later.

Next, the label that the VGG19 generates for the image is considered as the keyword of the target caption because the VGG19 is trained with the weak labels, as described in Section III. Therefore, the output of the VGG19 is the most important word information for translation, which is expected to improve the translation performance. Then, these additional translation hints are fed into the encoder.

B. TEXT ENCODING

Korean sentences are generally embedded in the encoder as one of three units: eojeol (Korean spacing unit), morpheme, or syllable. Fig.3 includes a text encoding cell that encodes the Korean caption tokenized into morpheme units. Given an input X as a source sentence of length T, where each token consists of a word-embedding vector with dimension E, a bi-directional RNN produces the two final hidden states

$(\overleftarrow{h}_1, \overrightarrow{h_T})$ by iteratively reading the input sequence X in a forward and backward fashion. Then, the final context state $(\vec{h_e})$ in the text encoding is computed by the concatenation of the two, followed by one projection layer to be compatible with the hidden size (D) of the encoder.

$$\vec{h_e} = W_e \left[\overrightarrow{h}_T : \overleftarrow{h}_1 \right] + b_e \tag{8}$$

where W_e and b_e are the weights and bias for the projection layer, respectively.

C. IMAGE ENCODING

The final text encoding state $(\vec{\boldsymbol{h}}_{\boldsymbol{e}})$ can be regarded as the initial state of the image encoding cell, as shown in Fig.3. As mentioned before, the global image features($\vec{\boldsymbol{v}}_{\boldsymbol{im}}$) are defined as the 4096-size vector in the fc7 layer of VGG19. Then, the cell reads the global visual features corresponding to the current caption and incorporates them with the previously encoded source caption per (9).

$$\vec{h}_{e1} = f(W_{im}\vec{v}_{im}, \vec{h}_e) \tag{9}$$

where f is the RNN (e.g., LSTM or GRU) and W_im is a transformation matrix to project image features into the same dimensionality as the word-embedding vector.

D. OBJECT ENCODING CELL

In the previous two encoding steps, the encoder obtains information about the text and the image as a whole. However, the global visual information has a weakness. Since the global visual features from the fc7 layer fully contain the information from the image, it is reasonable to think that the features could provide a better basis for solving under-translation than solely relying on the text information. However, much of the image information is implicitly used for translation, since the image feature is expressed as a 4096-dimensional vector and fed into the RNN cell at one time. This is considerable compared with the fact that the information for each word in the target caption is compactly stored into the RNN cell as a 300-dimensional vector representation. As a result, the effect of the image information, which is necessary for the translation, is weakened, and in terms of translation, it becomes noise, as can be shown in the experimental results.

However, let us consider the label that the VGG19 generates for the image. The label is reasonably related to the target caption describing the image, as given in (10), as shown at the bottom of this page. The label that the VGG19 generates is the keyword of the target caption. This is because each image has a weak label, which is defined as the keyword of its target caption, and the VGG19 is trained with the weakly labeled images. We assume in Section III that the feature set is defined as all the nouns of the target captions in the training data to select the keywords in the target captions. The reason is that the feature that is mainly selected as a keyword in the caption is an object in the image mentioned in the source caption, and the object is usually a noun. Thus, the predicted label from the VGG19 is a noun keyword in the target caption, which in turn is expected to refer to one of the objects appearing in the image. Thus, it can be concluded that direct encoding of the image's label increases the likelihood of generating the keyword of the translated caption at the decoding stage.

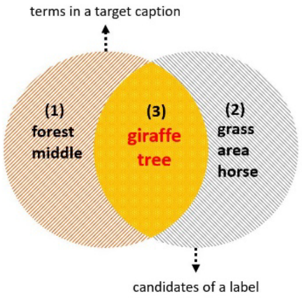


FIGURE 4. Relation between terms in a target caption and candidates of a label on an image (Figure 3).

Therefore, the encoder in the proposed model includes an object encoding cell, as shown in Fig.3, which directly encodes a word that refers to an object in the image, which is expected to be used for translation. The input of the cell is composed of K-labels having the upper Kth probability from the softmax layer of the VGG19. This is not only the predicted label of the image with the highest probability, but also candidate information for K-1 labels. This is to compensate for the following two possible problems. First, there can be more than one object in the image, and one or more may be described in the image caption. Let us revisit the structural meaning of the softmax layer in VGG19. The softmax layer consists of the probability that each candidate can be an image label. Among these candidates, some candidates above a certain level of probability may be the objects included in the image, and they may actually be described in the target caption.

Take Fig.4 as an example. It is based on the example in Fig.3. The left and middle parts of the Venn diagram, (1) and (3), contain the nouns in the target caption of the image in Fig.3. The middle and right parts, (2) and (3), contain the candidates for the label corresponding to the fifth highest probability from the softmax layer of the VGG19 for the image. This is the object list that is used as the input to the object encoding cell in Fig.3. The leftmost value in the object list has the highest probability of occurrence in the softmax layer, and its value is predicted as the label (giraffe) of the image.

In Fig.4, the words in (3) correspond to the label candidates with high probability in the softmax layer of the VGG19 among the actual words comprising the translated

$$\{\, \textit{terms denoting objects in an image} \} \ni \textit{label} := \textit{keywordinatargetcaption} \in \{\, \textit{all noun terms in a target caption} \} \tag{10}$$

caption. At the same time, these are objects that represent the image and eventually are likely to be the words in the translated caption. Indeed, "giraffe" is the image label predicted with the highest probability, and "tree" is the label candidate with the next highest probability. These two labels are not only included in the image but also in the actual caption. Thus, this cell can contain explicit information needed for translation, as opposed to the previous image encoding cell, by encoding one or more objects described in the target caption with the label candidates generated by VGG19.

The second reason for using K-labels as the input of the cell is that, even if the VGG19 predicts a label as an object that is not included in the image (e.g., the words in (2) of the Fig.4), one of the label candidates with the next highest probability in the softmax layer will possibly be the correct label. If the label predicted by the VGG19 is solely encoded in the cell, the effect of addition of the cell is highly dependent on VGG19 performance. However, if the cell contains label candidates with a certain probability or more, it is highly expected to improve the recall of the translation quality.

Each of the K label candidates is represented as a vector by sharing the target embedding used in the decoder. Each embedding vector is concatenated and finally used as the input to the cell. The final context vector of the encoder is shown in (12).

$$\vec{O} = [\vec{o_1}; \vec{o_2}; \ldots; \vec{o_K}] \tag{11}$$

$$\vec{h_{e2}} = f\left(W_O\, \vec{O},\, \vec{h}_{e1}\right) \tag{12}$$

where $\vec{O}$ is a concatenation of K label candidates consisting of target-embedding vectors, f is the RNN (e.g., LSTM or GRU), and W_o is a transformation matrix to project $\vec{O}$ into the same dimensionality as the source word-embedding vector.

E. DECODER WITH ATTENTION MECHANISM

The proposed model is based on the standard attention mechanism [17] where, at each decoding time step, every word in the target caption is generated considering both hidden state information at the previous decoding step and the current context vector that denotes the relevant information of the source sentence. Compared with the existing mechanism, the output of all the RNN cells in the encoder is not used to derive the context vector generated in the attention layer. Instead, the decoder is designed only to attend to the output of the text-encoding part in the encoder, excluding the output of the two cells associated with the image. This is because when the text encoding is performed, the information may be distorted or omitted due to the long-term dependency such that the text information can be given a larger weight by focusing only on the text information. According to a study by [38], the influence of the attention vector is greater than the previous hidden state of the RNN when generating words for each decoding step. Therefore, to effectively use the text information, which is the largest basis for the translation, the decoder is designed to focus only on the text.

TABLE 1. Summary of dataset.

Dataset	Language	# of Sentences	# of Tokens	Average Lengths (Words)
Our Dataset	Korean	15,000	14,130	8.68
	English	15,000	7,291	11.3
Multi30K Dataset	English	31,014	11,420	11.9

The decoder is a uni-directional RNN that generates the target sentence Y. The conditional probability of choosing the *j*th word (y_j) is:

$$P(y_j|y_{<j}, x; \theta) = softmax\ (y_{j-1}, s_j, c_j) \tag{13}$$

where y_{j-1} is the previously generated word, s_j is the current RNN hidden state, and c_j is a source-side context vector. The attention mechanism generates the context vector by means of a weighted sum of the final annotation vectors in the encoder:

$$c_j = \sum_{i=1}^{T} (\alpha_{i,j} * h_i) \tag{14}$$

where $\alpha_{i,j}$ denotes the attention weight, which is defined as

$$\alpha_{i,j} = \frac{exp(e_{i,j})}{\sum_{k=1}^{T} exp(e_{k,j})} \tag{15}$$

$$e_{i,j} = v_a^T tanh\ (W_a s_j + U_a h_i) \tag{16}$$

where $e_{i,j}$ is an alignment model that calculates the degree of the relevance between s_j and h_i. In addition, v_a, W_a, and U_a are the weights for the matrix transformation.

V. EXPERIMENT

This section consists of three parts. Section V.A introduces our new dataset and Section V.B describes the training details. Lastly, we show the performance of our proposed model and investigate the results at Section V.C.

A. DATASET

We developed a new dataset similar to the Multi30K dataset [39] adopted as the primary data in the WMT16 shared task [19]. Our dataset consists of 7,500 images, each with two Korean descriptions and two corresponding human-translated English descriptions. The specification of the image captions is described in TABLE 1. Compared with the pre-existing data, the main difference is that the images in our dataset have no object labels, whereas all the images in the Multi30K dataset have appropriate entity labels.

The training, validation, and test sets consist of 6000, 1000, and 500 images, respectively, each containing two pairs of sentences (the original Korean captions and their translations into English). In addition, we use the entire English captions in the Multi30K dataset to produce high-quality sentence embedding. Furthermore, we employed the pre-trained VGG19 to better extract image features [22].

The Moses Statistical MT Toolkit [40] is used to normalize and tokenize the English captions. We also use the komoran

TABLE 2. BLEU-4 score of the baseline model (NMT) on test set.

Model \ Korean Token Unit	Eojeol	Morpheme	Syllable
NMT (Baseline)	31.3	34.1	**34.3**

class in the KoNLPy package [41] to split Korean descriptions into morphemes, which are the basic units of Korean words. Moreover, we discard sentences longer than 50 words.

B. TRAINING DETAILS

We trained two Seq2Seq models: one for the text-based NMT system as a baseline, and the other for the multimodal NMT as a proposed model. Both encoders are bidirectional RNNs with GRU cells and hidden sizes of 512, and both decoders are unidirectional RNNs with GRU cells and hidden sizes of 512. The basic attention mechanism comes from [17], but its implementation is based on [18] due to its effectiveness. The size of the source word embedding depends on its input unit: eojeol(100), morpheme(200), syllable(50). The size of target word embedding is 300. All the weights in the models are initialized using a Xavier scheme [42], and the biases are set to zero. Image features are extracted from the fc7 layer of VGG19 pre-trained on the ImageNet and fine-tuned on our dataset. The number of elements in the object list is 5.

Both models are trained using a stochastic gradient descent with Adam [43]. The learning rate is set to 0.001. The mini-batch size is set to 80 for the NMT or 32 for the multimodal NMT. We use early stopping by choosing the model where the BLEU-4 [44] over the validation set does not improve for 20 epochs. The translation quality is automatically evaluated by BLEU-4.

C. RESULTS

The performances of the baseline and multimodal NMT systems are presented in Table 2 and TABLE 3 using BLEU-4 as an automatic evaluation metric. The baseline is a conventional attention-based Seq2Seq model (NMT), and two types of multimodal NMT models are tested: one containing only the image encoding cell and the other having both the image encoding cell and the object encoding cell. Experiments were carried out by learning three input units of Korean for each model: eojeol (Korean spacing unit), morpheme, and syllable.

Multimodality, as in previous studies, has a notably positive impact on the MT results, even though the data in this work is a set of bilingual unlabeled images compared with the previous studies [26], [28] that used human-annotated images. As shown in Table 2 and TABLE 3, the model using only image features ($\text{MNMT}_{\text{image}}$) showed a maximum improvement of +0.8 compared to the baseline model (NMT). In addition, the image and object model ($\text{MNMT}_{\text{image+object}}$), including the labels generated from the image, showed up to a +1.0 BLEU performance improvement when comparing the performance between models for the same Korean input unit (morpheme).

The results can be analyzed considering three aspects: the performance change corresponding to Korean input unit, the effect of image features, and the effect of label candidates. First, of the three input units, the morpheme is the Korean input unit most suitable for multimodal NMT. When image information is added, the model with the morpheme unit tends to have a higher performance. On the other hand, both the eojeol and syllable unit input models in the baseline demonstrated inadequate results in terms of multimodality. The reason comes from the size of the image features. Unlike, the word embedding which delivers compact word information in each time step, the image features broadly reflect the overall characteristics of the image. Therefore, the information contains somewhat redundant or unnecessary elements in terms of translation, and thus it eventually acts as noise.

Next, we discuss the effectiveness of additional image features. Considering the token of source sentences as the morpheme, the $\text{MNMT}_{\text{image}}$ system showed the best performance (34.9) by adopting the weak-labeling method based on TF-IDF with sentence clustering, which showed up to a +1.0 BLEU performance improvement compared with the BoW-based weak-labeling method. Therefore, it follows that changes in translation performance due to the addition of image features are dependent on the weak-labeling methods.

TABLE 3. BLEU-4 Scores of the proposed models (MNMT) on the test set. $\text{MNMT}_{\text{image}}$ indicates the MNMT model containing only the image encoding cell. $\text{MNMT}_{\text{image+object}}$ denotes the MNMT model with both the image encoding cell and the object encoding cell.

	Image Weak Labeling Method								
	Bag-of-Word			TF-IDF			TF-IDF with sentence clustering		
Model \ Korean Token Unit	Eojeol	Morpheme	Syllable	Eojeol	Morpheme	Syllable	Eojeol	Morpheme	Syllable
$\text{MNMT}_{\text{image}}$	30.7	33.9	32.9	30.8	34.5	34.4	30.2	34.9	33.6
$\text{MNMT}_{\text{image+object}}$	29.6	34.3	33	29.8	34.9	32.4	30.9	35.1	34.1

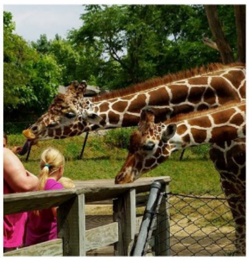
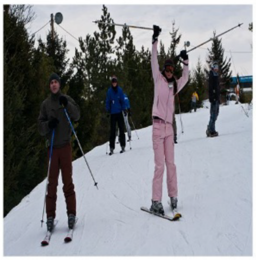
		
BoW	People	People
TF-IDF	Giraffe	Ski

FIGURE 5. Examples of weak labels.

For example, the number of image labels determined by the BoW method is 153 in total, and on average, one label represents approximately 40 among all images in the training data. On the other hand, the number of image labels determined by the TF-IDF is 757 in total, and one label includes approximately 10 images. As a result, the model trained with TF-IDF-based weakly labeled images may create more discriminative image features compared with the one trained with BoW-based weakly labeled images when it comes to images with different characteristics, as long as the capacity of the models is all the same.

Take Fig.5 as an example. In the case of the model learned with BoW-based weakly labeled images, the labels of both images in Fig.5 were created as "people." However, the weights of the "people" in each picture are quite different. In contrast, the model trained with TF-IDF based weakly labeled images generates the labels for Fig.5 as "giraffe" and "ski." Therefore, it is reasonable to judge that the image features generated by the model trained with TF-IDF based weakly labeled images are more sophisticated than those with BoW-based weakly labeled images. This inference can be proven based on the results in TABLE 3. As shown in TABLE 3, the image features extracted from the model trained with BoW-based weakly labeled images caused noise in translation. On the other hand, when using TF-IDF based weakly labeled images, the quality of translation improved, especially for the model using morpheme units.

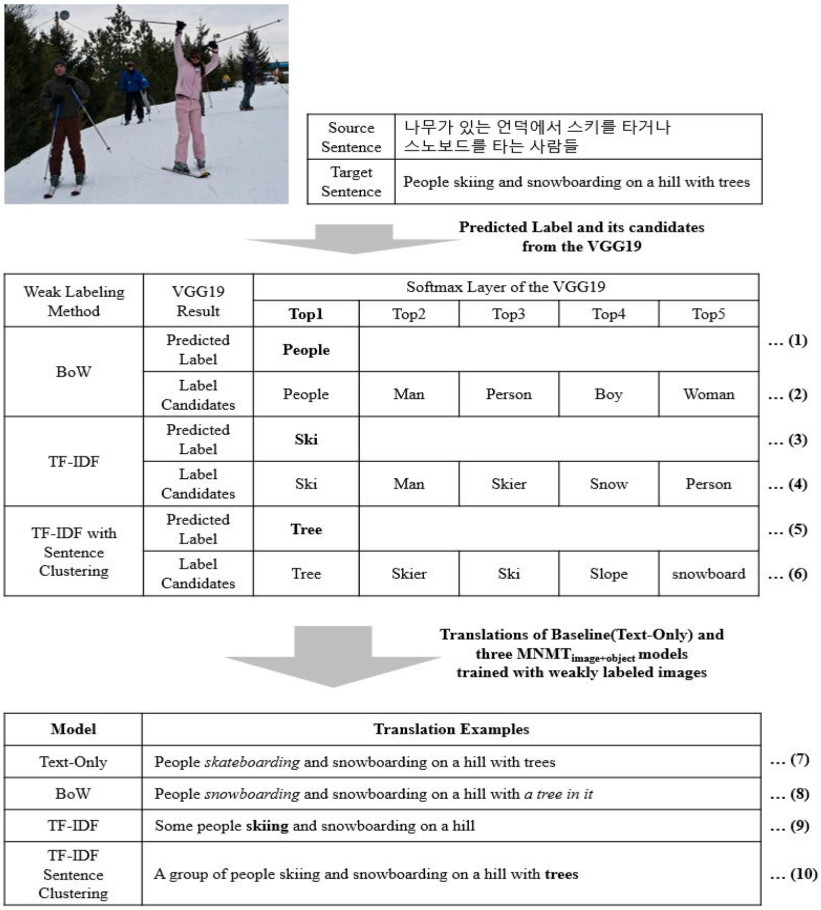
Let us now consider the effect of label candidates used in the object encoding cell, which is the last cell in the encoder. First, it was found that the labels generated by the VGG19 using weakly labeled images with TF-IDF are more suitable for resolving under-translation problems than the labels using the BoW-based approach. For example, Fig.6 shows the result of the predicted labels (e.g., 1, 3, 5) and 5-sized object lists (e.g., 2, 4, 6) generated by the VGG19 independently trained with weakly labeled images using three different weak-labeling methods. Also, the sentences from (7) to (10) in Fig. 6 are the translation results of the baseline model and each of three multimodal MT models trained with weakly labeled images using three different weak-labeling methods.

For the image in Fig.6, the model trained with BoW-based weakly labeled data generated mostly the words in (2) that are related to "people" of the target caption. This corresponds to the word "사람들 [salamdeul]" in the source caption. Therefore, it seems to be advantageous to translate the source caption with the additional information about label candidates. However, syntactically explicit words in a sentence, such as the subject, are generally translated sufficiently by text alone, and there is no case where they are confirmed to be useful in our actual experimental results. Therefore, label candidates like (2) are not useful as additional resources in translation, but rather serve as superfluous features.

However, the case of TF-IDF is different. Label candidates such as "ski" (predicted label), "skier," and "snow" as in (4) of Fig.6 have no corresponding word in the target sentence but are significantly semantically similar to "skiing" or "snowboarding." As such related resources are encoded, it is found that the term "skiing," which both the multimodal NMT model trained with weakly labeled images using BoW and the baseline model cannot translate, is normally generated as in (9) of Fig.6.

The final matter to address is the effect of weakly labeled images using sentence clustering using the K-means algorithm. For morpheme-based models that have a positive effect on combining with image information, MNMT$_{\text{image+object}}$ showed an improvement of +1.0 BLEU compared with the baseline model. Specifically, the MNMT$_{\text{image}}$ model trained with weakly labeled images using improved TF-IDF showed a +0.4 BLEU performance improvement compared with that using the existing TF-IDF. The number of labels that weakly labeled images with improved TF-IDF is 1020, which, on average, means that one label contains only approximately 8 images. This is less than the number of labels of weakly labeled images using the existi1ng TF-IDF. Therefore, it follows that the VGG19 trained with weakly labeled images using the improved TF-IDF can produce more accurate image features than that with the existing TF-IDF.

In addition, when exploiting label candidates as an additional source from the VGG19, the MNMT$_{\text{image+object}}$model showed a +0.2 higher performance improvement, which is the best score in this work. The translation result of (9) in Fig.6 fails to generate the word "tree" corresponding to the word "나무[namu]" in the source caption. However, if the model is trained with the data using improved TF-IDF, then the model can generate an accurate sentence like (10) in Fig.6, which reflects all the object information appearing in the source caption. In fact, the VGG19 trained with weakly labeled images using improved TF-IDF generated "tree" as a label in (5). Additionally, the words "ski" and "skier," which were also generated by the model trained with the data using the existing TF-IDF method, are also included in the object list. As a result, the MNMT$_{\text{image+object}}$model correctly generates the word "tree" that the model trained with weakly labeled images using the existing TF-IDF failed to generate and includes the word "skiing" that the baseline model could



Source Sentence	나무가 있는 언덕에서 스키를 타거나 스노보드를 타는 사람들
Target Sentence	People skiing and snowboarding on a hill with trees

Predicted Label and its candidates from the VGG19

Weak Labeling Method	VGG19 Result	Softmax Layer of the VGG19					
		Top1	Top2	Top3	Top4	Top5	
BoW	Predicted Label	**People**					**... (1)**
	Label Candidates	People	Man	Person	Boy	Woman	**... (2)**
TF-IDF	Predicted Label	**Ski**					**... (3)**
	Label Candidates	Ski	Man	Skier	Snow	Person	**... (4)**
TF-IDF with Sentence Clustering	Predicted Label	**Tree**					**... (5)**
	Label Candidates	Tree	Skier	Ski	Slope	snowboard	**... (6)**

Translations of Baseline(Text-Only) and three MNMT$_{image+object}$ models trained with weakly labeled images

Model	**Translation Examples**	
Text-Only	People *skateboarding* and snowboarding on a hill with trees	**... (7)**
BoW	People *snowboarding* and snowboarding on a hill with *a tree in it*	**... (8)**
TF-IDF	Some people **skiing** and snowboarding on a hill	**... (9)**
TF-IDF Sentence Clustering	A group of people skiing and snowboarding on a hill with **trees**	**... (10)**

FIGURE 6. Translation examples with weak labeling by BoW, TF-IDF, and TF-IDF with sentence clustering.

not generate. Thus, the translation quality can be improved by addressing the under-translation issue via directly encoding the word information likely to be generated in the translation process.

VI. CONCLUSION

In this paper, we have introduced a multimodal MT system that incorporates image information with the conventional text-based NMT. To exploit visual information from images without any object labels, we proposed a novel approach that creates weak labels by selecting keywords from image descriptions by means of feature selection methods in NLP. Moreover, the quality of the translation proves to be valid when the label candidates generated from the VGG19 are directly encoded. As a result, our proposed model improved the performance by +1.0 BLEU compared to the text-based NMT model. This demonstrates that visual information from the images can be effectively used to translate captions for unlabeled images.

In the future, we will extend the architecture by incorporating both visual and keyword components with an attention mechanism and explore more accurate ways to extract features from images and descriptions.

REFERENCES

[1] A. Barreiro, B. Scott, W. Kasper, and B. Kiefer, "OpenLogos machine translation: Philosophy, model, resources and customization," *Mach. Transl.*, vol. 25, no. 2, pp. 107–126, 2011.

[2] M. Costa-Jussa, M. Farrús, J. B. Marino, and J. Fonollosa, "Study and comparison of rule-based and statistical catalan-spanish machine translation systems," *Comput. Informat.*, vol. 31, no. 2, pp. 245–270, 2012.

[3] U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada, "Fast and optimal decoding for machine translation," *Artif. Intell.*, vol. 154, nos. 1–2, pp. 127–143, 2004.

[4] A. Kazemi, A. Toral, A. Way, A. Monadjemi, and M. Nematbakhsh, "Syntax- and semantic-based reordering in hierarchical phrase-based statistical machine translation," *Expert Syst. Appl.*, vol. 84, pp. 186–199, Oct. 2017.

[5] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol.*, vol. 1, 2003, pp. 48–54.

[6] R. Wang, H. Zhao, B. Lu, M. Utiyama, and E. Sumita, "Bilingual continuous-space language model growing for statistical machine translation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 7, pp. 1209–1220, Jul. 2015.

[7] B. Zhang, D. Xiong, and J. Su, "Neural machine translation with deep attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[8] K. Chen *et al.*, "A neural approach to source dependence based context model for statistical machine translation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 2, pp. 266–280, Feb. 2018.

[9] Q.-P. Nguyen, A.-D. Vo, J. C. Shin, and C. Y. Ock, "Effect of word sense disambiguation on neural machine translation: A case study in Korean," *IEEE Access*, vol. 6, pp. 38512–38523, 2018.

[10] D. Banik, A. Ekbal, and P. Bhattacharyya, "Machine learning based optimized pruning approach for decoding in statistical machine translation," *IEEE Access*, vol. 7, pp. 1736–1751, 2019.

[11] B. Zhang, D. Xiong, J. Su, and Y. Qin, "Alignment-supervised bidimensional attention-based recursive autoencoders for bilingual phrase representation," *IEEE Trans. Cybern.*, to be published.

[12] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 4, Sep. 2014, pp. 3104–3112.

[13] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in *Proc. Conf. Empirical Methods Natural Language Process.*, vol. 3, Oct. 2013, pp. 1700–1709.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734.

[16] Z. Tu, Z. Lu, Y. Liu, X. Liu, and H. Li, "Coverage-based neural machine translation," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 76–85.

[17] D. Bahdanau, K. Cho, and Y. Bengio. (2014). "Neural machine translation by jointly learning to align and translate." [Online]. Available: https://arxiv.org/abs/1409.0473

[18] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1412–1421.

[19] L. Specia, S. Frank, and K. Sima'an, and D. Elliott, "A shared task on multimodal machine translation and crosslingual image description," in *Proc. 1st Conf. Mach. Transl.*, vol. 2, 2016, pp. 543–553.

[20] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, Feb. 2014.

[21] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 1–42, 2015.

[22] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 770–778.

[24] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3156–3164.

[25] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[26] P.-Y. Huang, F. Liu, S. Shiang, J. Oh, and C. Dyer, "Attention-based multimodal neural machine translation," in *Proc. 1st Conf. Mach. Transl.*, vol. 2, 2016, pp. 639–645.

[27] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[28] I. Calixto, Q. Liu, and N. Campbell, "Doubly-attentive decoder for multi-modal neural machine translation," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 1913–1924.

[29] Y. Wei *et al.*, "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Feb. 2017.

[30] F. Tian and X. Shen, "Image annotation with weak labels," in *Proc. 14th Int. Conf. Web-Age Inf. Manage.*, 2013, pp. 375–380.

[31] M. J. Blosseville, G. Hébrail, M. G. Monteil, and N. Pénot, "Automatic document classification: Natural language processing, statistical analysis, and expert system techniques used together," in *Proc. 15th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 1992, pp. 51–58.

[32] E. Blanzieri and A. Bryl, "A survey of learning-based techniques of email spam filtering," *Artif. Intell. Rev.*, vol. 29, no. 1, pp. 63–92, 2008.

[33] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, 1988.

[34] Y.-J. Ko and J.-Y. Seo, "Issues and empirical results for improving text classification," *J. Comput. Sci. Eng.*, vol. 5, no. 2, pp. 150–160, 2011.

[35] T.-H. Jo, "Representation of texts into string vectors for text categorization," *J. Comput. Sci. Eng.*, vol. 4, no. 2, pp. 110–127, 2010.

[36] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient K-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2002.

[37] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for Word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[38] D. Britz, A. Goldie, M.-T. Luong, and Q. Le, "Massive exploration of neural machine translation architectures," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1442–1451.

[39] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30K: multilingual English-German image descriptions," in *Proc. 5th Workshop Vis. Lang.*, 2016, pp. 70–74.

[40] P. Koehn *et al.*, "Moses: Open source toolkit for statistical machine translation," in *Proc. 45th Annu. Meeting Assoc. Comput. Linguistics*, Jun. 2007, pp. 177–180.

[41] E. L. Park and S. Cho, "KoNLPy: Korean natural language processing in Python," in *Proc. 26th Annu. Conf. Hum., Cognit. Lang. Technol.*, 2014, pp. 4–7.

[42] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *J. Mach. Learn. Res.*, vol. 9, pp. 249–256, May 2010.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, Dec. 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[44] K. Papineni, S. Roukos, T. Ward, and W. W. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. ACL*, Apr. 2002, pp. 311–318.

YOONSEOK HEO received the B.S. and M.S. degrees in computer science (majoring in natural language generation) from Sogang University, where he is currently pursuing the Ph.D. degree with the Department of Computer Science. He has worked as a Researcher with Gachon University, in 2018. He is interested in spoken dialogue systems, machine translation, question answering, machine reading comprehension, and named entity recognition. His current research focuses on the way of exploiting multimodal resources for machine translation and addressing large-scale open domain texts for machine reading comprehension.

SANGWOO KANG received the Ph.D. degree in computer science from Sogang University, where he was also a Research Fellow Professor. He has been an Assistant Professor with the Department of Software, Gachon University, since 2016, where he is currently leading the Intelligent Software and Natural Language Processing Laboratory. His specialty is Natural Language Processing and is interested in spoken dialogue interface, information retrieval, text mining, opinion mining, big data, and UI/UX. His recent focus has been in applying deep learning techniques to his research.

DONGHYUN YOO received the master's degree by writing a thesis on spoken dialogue interface in computer science from Sogang University. He is currently a Researcher with the Department of Natural Language Processing, Artificial Intelligence Center, NCSoft Corporation. His specialty is natural language processing and is interested in spoken dialogue interface, information extraction, question and answer, and deep learning techniques.

• • •