Multimodal Neural Machine Translation Using CNN and Transformer Encoder

Hiroki Takushima[†], Akihiro Tamura[†], Takashi Ninomiya[†] and Hideki Nakayama[‡]

[†] Graduate School of Science and Engineering, Ehime University
[‡] Graduate School of Information Science and Technology, The University of Tokyo
{takushima@ai., tamura@, ninomiya@}cs.ehime-u.ac.jp
nakayama@nlab.ci.i.u-tokyo.ac.jp

Abstract. Multimodal machine translation uses images related to source language sentences as inputs to improve translation quality. Pre-existing multimodal neural machine translation (NMT) models that incorporate the visual features of each image region into an encoder for source language sentences or an attention mechanism between an encoder and a decoder cannot catch the relationship between the visual features from each image region. This paper proposes a new multimodal NMT model that encodes an input image using a convolutional neural network (CNN) and a Transformer encoder. In particular, our proposed image encoder extracts visual features from each image region using a CNN then encodes an input image based on the extracted visual features using a Transformer encoder, where the relationship between the visual features from each image region is captured by a self-attention mechanism of the Transformer encoder. Our experiments with English-German translation tasks using the Multi30K data set showed our proposed model improves 0.96 BLEU points against a baseline Transformer NMT model without image inputs and improves 0.47 BLEU points against a baseline multimodal Transformer NMT model without a Transformer encoder for images.

Key words: multimodal neural machine translation, machine translation, multimodal learning, neural network, transformer, CNN

1 Introduction

Various neural network (NN)-based methods have been actively studied in the area of neural machine translation (NMT). Recently, a Transformer NMT model [1] attracted attention for performing state-of-the-art translation above other NMT models. Like recurrent neural network (RNN)-based NMT models [2][3] and convolution neural network (CNN)-based NMT models [4], a Transformer NMT model consists of an encoder that generates intermediate expressions from source language sentences and a decoder that predicts target language sentences from intermediate representations. Each of the Transformer encoder and decoder has a self-attention mechanism that catches the relationship between the words

in a sentence. In particular, the self-attention mechanism in the Transformer encoder catches the relationship between input words in a source sentence, and the mechanism in the Transformer decoder catches the relationship between output words in a target sentence.

Multimodal NMT [5] uses images related to source sentences and source language sentences as inputs to improve translation quality. Multimodal NMT models assume that additional input images could help improve translation performance as clues that decrease translation ambiguity. For example, the English word "bank" has multiple meanings and can be translated into two Japanese words, "銀行 (a business that holds and lends money and provides other financial services)" and "土手 (land along the side of a river or lake)"; therefore, "bank" has translation ambiguity. However, the relevant term in English for "bank" can be translated into Japanese by using an image in addition to the word. "Bank" can be translated into "銀行" if a financial image is input, and "bank" can be translated into "土手" if a river image is input.

Some of the pre-existing multimodal NMT models incorporate visual features extracted from each region of the input image using a CNN in an attention mechanism between an encoder and decoder of the NMT model; therefore, target language sentences could be generated by input images being reflected through the cross-lingual attention mechanism [6]. In addition, multimodal NMT models that incorporate the visual features of each region into an encoder for source sentences have been proposed [7, 8]. However, pre-existing multimodal NMT models cannot catch the relationship between the visual features from each image region.

Therefore, we propose a new multimodal NMT model that encodes an input image using a CNN and a Transformer encoder. Specifically, the transformer decoder in our model generates a target language sentence from the concatenation of the intermediate expression of an input image, which is the output of an encoder for an image (hereinafter referred to as "image encoder"), and that of a source language sentence, which is the output of an encoder for a source language sentence (hereinafter referred to as "source language encoder"). The image encoder first extracts visual features from each image region using a CNN, then encodes an image based on the extracted visual features using a Transformer encoder. Our model could catch the relationship between visual features from each image region by a self-attention mechanism of the transformer encoder in the image encoder.

The experiments with the English-German translation tasks using the Multi30K data set [9] showed that our model improved 0.96 BLEU points against a baseline Transformer NMT model that does not use image inputs and improved 0.47 BLEU points against a baseline multimodal Transformer NMT model that does not use a Transformer encoder in the image encoder.

2 Related work

In this section, we overview the Transformer NMT model [1] we based our model on and describe previous multimodal NMT models.

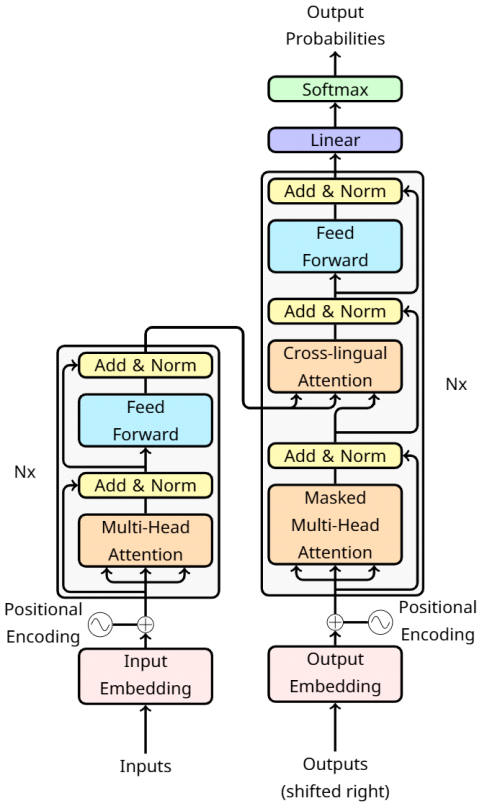

Fig. 1. Transformer NMT Model

2.1 Transformer

Figure 1 overviews the Transformer NMT model [1]. The Transformer model accepts a source language sentence $X = (x_1, x_2, ..., x_M)$ as an input and outputs a target language sentence $Y = (y_1, y_2..., y_L)$. In training, objective function $p(Y|X)$ is learned from a parallel corpus. The Transformer model is an encoder-decoder model in which the Transformer encoder generates the intermediate representation h_t $(t = 1, ..., M)$ from the source language sentence X and the Transformer decoder generates a target language sentence Y from the intermediate representation h_t:

$$h_t = TransformerEncoder(X), \tag{1}$$

$$Y = TransformerDecoder(h_t). \tag{2}$$

N layers are stacked in the Transformer encoder and the Transformer decoder. Each layer of the encoder is composed of two sublayers, self-attention and

point-wise, fully connected layers (hereinafter referred to as "Feed Forward"). Each layer of the decoder is composed of three sublayers, an attention mechanism between the source language and the target language (hereinafter referred to as "cross-lingual attention") and the two sublayers of the encoder. Layer normalization [1] and a residual connection [1] are used between the sublayers in the encoder and decoder.

Each attention mechanism $Attention(\cdot)$ (i.e., a self-attention mechanism and a cross-lingual attention mechanism) is computed as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_{model}}})V, \tag{3}$$

where Q, K, and V represent an internal representation of the encoder/decoder and d_{model} is the dimension of the internal representation. The inner product of Q and K represents the similarity between each element of Q and K and is converted to a probability by using the softmax function, which can be treated as weights of attention of Q to K. Finally, the attention mechanism computes a weighted sum of V with the attention weights. In this way, the attention mechanism generates an expression that reflects the degree of association between a word in Q and that in K. The self-attention mechanism computes the degree of association between words in the same sentence by using the same input source for Q, K, and V. In particular, the self-attention in the encoder catches the degree of association between words in the input sentence by using the internal expressions in the encoder as Q, K, and V. Meanwhile, the self-attention in the decoder catches the degree of association between words in the output sentence by using the internal expressions in the decoder as Q, K, and V. The cross-lingual attention mechanism computes the degree of association between a word in the source language sentence and that in the target language sentence by using the internal representation of the decoder as Q and the output of the last layer of the encoder as K and V.

Vaswani et al. [1] found that the multi-head attention mechanism that uses multiple attention functions $Attention(\cdot)$ is more beneficial than a single attention mechanism. In the multi-head attention with h heads, Q, K, and V are linearly projected to h subspaces, and then the attention function is performed in parallel on each subspace. Finally, these are concatenated, and the concatenation is projected to a space with the original dimension. The multi-head attention mechanism $MultiHead(\cdot)$ is represented as follows:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^o, \tag{4}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V), \tag{5}$$

where $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}, W_i^K \in \mathbb{R}^{d_{model} \times d_k}, W_i^V \in \mathbb{R}^{d_{model} \times d_k}$ are weight matrices for linear transformations of Q, K, V from d_{model} dimension to d_k ($= d_{model}/h$) dimension, respectively, h is the number of head, and $Concat$ is a function that concatenates two matrices. Multi-head attention enables the model to aggregate information from different representation spaces at different positions.

The Transformer uses positional encoding (hereinafter referred to as "PE") to encode the positional information of each word in a sentence because the Transformer does not have any recurrent or convolutional structure. PE is calculated as follows:

$$PE(pos, 2i) = sin(pos/10000^{2i/d_{model}}), \tag{6}$$

$$PE(pos, 2i+1) = cos(pos/10000^{2i/d_{model}}), \tag{7}$$

where pos is the absolute position of the word and i is the dimension. At the bottom of the encoder and the decoder in Fig. 1, PE is added to the input embeddings.

2.2 Previous Multimodal NMT

Since NMT appeared, multimodal NMT has been actively researched. For example, many multimodal NMT models have been proposed in the multimodal machine translation shared task [5] at the WMT conference. Calixto et al. [6] improve translation performance by introducing attentions between images and target language sentences into the RNN-based sequence-to-sequence NMT model [3]. Calixto et al. [8] propose a method for incorporating visual features extracted from input images by using a CNN in the initial hidden state of the RNN encoder and a method for injecting visual features into the input of a decoder. Huang et al. [7] incorporate both features extracted from each image region by using a CNN and visual features caught by object detection using a region-based convolutional network (R-CNN) [10] in the initial hidden state of the LSTM encoder.

Recently, some Transformer-based multimodal NMT models have been proposed. Grönroos et al. [11] added a gating layer to each output of the Transformer encoder and decoder, and their model uses visual features in the gate. They showed that the proposed gating layer in the encoder decreases ambiguity in encoding source language sentences and that in the decoder suppresses the outputs of unnecessary words.

Note that such previous multimodal NMT models could not establish the relationship between the visual features from each image region.

3 Proposed Model

In this section, we propose a multimodal NMT model that encodes an input image using a CNN and a Transformer encoder to catch the relationship between the features of each image region. Figure 2 overviews our model. Our model has two encoders: an image encoder and a source language encoder. In our model, an input image is encoded by the image encoder and a source language sentence is encoded by the source language encoder. Then, the concatenation of intermediate expressions generated by the two encoders is fed into the conventional Transformer decoder and the Transformer decoder generates a target language sentence from the intermediate expression based on image and source language sentence information.

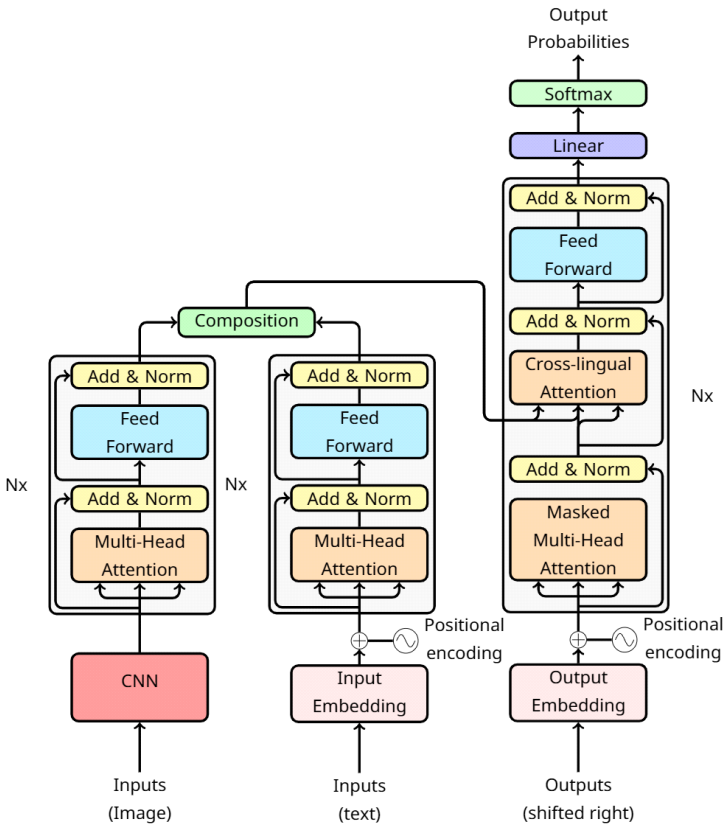

Fig. 2. Proposed Multimodal NMT Model

In the rest of this section, we describe the proposed image encoder in Section 3.1. We describe the composition layer that concatenates the intermediate expression of the image encoder and that of source language encoder in Section 3.2. The details of the source language encoder and the decoder of our model are the same as the encoder and decoder of the conventional Transformer NMT model described in Section 2.1, respectively.

3.1 Image Encoder

Our image encoder first uses a CNN to extract visual features from each image region. CNN is a neural network that includes multiple convolution layers and pooling layers. In our experiments, we used VGG16 [12], which has 16 layers of 13 convolution layers and 3 fully connected layers.

Then we fed the visual features output by a CNN into the conventional Transformer encoder:

$$feature = CNN(image), \tag{8}$$

$$h_v = TransformerEncoder(W \cdot feature), \tag{9}$$

where $image$ is an input image, $feature$ is the visual feature extracted by a CNN, h_v is the intermediate representation generated by the image encoder, and $W \in \mathbb{R}^{d_{feature} \times d_{model}}$ is a weight matrix for linear transformation (from $d_{feature}$ dimension to d_{model} dimension). Note that positional encoding PE is not used in the Transformer encoder of the image encoder.

3.2 Composition Layer

After the intermediate representations of an input image and a source language sentence are generated by the image encoder and the source language encoder, respectively (see Eq. (1) and Eq. (9)), the two representations are concatenated in the composition layer:

$$h = Concat(h_v, h_t). \tag{10}$$

The concatenation h is the output of the proposed encoder and is fed to the decoder of our model.

4 Experiments and Results

This section describes the experiments for evaluating the multimodal NMT.

4.1 Experiment Setting

We performed the English-German translation tasks using the Multi30K dataset [9]. We prepared 29,000 sentences as a training dataset, 1,014 sentences as a variation dataset, and 1,000 sentences as a test dataset. We used byte pair encoding (BPE) [13] for sub-wording (i.e., words with a low number of occurrences were decomposed into subword units). The vocabulary set was shared between encoder and decoder, and there were 6,150 words in the vocabulary set.

We resized an input image to 256x256, and then used the center cropped 224x224 image as the input of the image encoder. We used the output of the final convolution layer of VGG16 [12] for image regions as the input of the Transformer encoder of the image encoder. Following the original paper on Transformers [1], the Transformer models had six layers stacked for each encoder and decoder where each layer had 8 heads and the embedded vectors had 512 dimensions ($d_{model} = 512$). We used Adam for optimization with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$. The learning rate was set following the [1]. The mini batch size was set to 80, and 50 epochs were repeated. We did not perform CNN fine-tuning while training the NMT. Greedy decoding was used to generate target language sentences at inference.

BLEU [14] evaluated translation quality. We used the model with the best BLEU score for the validation dataset to evaluate the test dataset.

Table 1. Results

Method	BLEU
$Transformer$ without image	34.30
$CNN + Trans_{Dec}$	34.79
$CNN + Trans_{Enc} + Trans_{Dec}$	35.26

Table 2. Translation Example

Input	man scaling a wall with fire in his hand
Reference	mann erklettert mauer mit feuer in der hand
$Transformer$ without image	ein mann , der eine wand in der hand fordert , geht eine wand in der hand（a man who demands a wall in his hand, goes a wall in his hand）
$CNN + Trans_{Dec}$	ein mann , der ein feuer in der hand hlt （a man holding a fire in his hand）
$CNN + Trans_{Enc} + Trans_{Dec}$	ein mann geht mit feuer in der hand eine wand entlang（a man is walking along a wall with a fire in his hand）

4.2 Evaluation

Table 1 shows the experimental results. In the table, "$Transformer$ without image" means a baseline Transformer model without image inputs. "$CNN + Trans_{Dec}$" indicates a baseline multimodal MNT model, where a CNN encoder is used for image encoding and a Transformer is used for translation. Note that "$CNN + Trans_{Dec}$" does not use a Transformer encoder for image encoding. "$CNN + Trans_{Enc} + Trans_{Dec}$" indicates our model, which is comprised of a CNN encoder and a Transformer encoder for image encoding and a Transformer for translation. As shown in the table, our model improved 0.96 BLEU points against the baseline Transformer ($Transformer$ without image), and improved 0.47 BLEU points against the baseline multimodal NMT ($CNN + Trans_{Dec}$).

5 Discussion

Table 2 and Fig. 3 show translation examples of each model and the input image for multimodal NMT models (i.e., the proposed model ($CNN + Trans_{Enc} + Trans_{Dec}$) and the baseline model ($CNN + Trans_{Dec}$)). As Table 2 shows, while a $Transformer$ without an image (which does not use the input image) misses the information of "feuer (fire)," our model and $CNN+Trans_{Dec}$ (which utilize the input image) successfully include the information of "fire." Moreover, while $CNN + Trans_{Dec}$ misses the information of "wand/mauer (wall)", our model successfully includes the information of "wall."

To confirm whether additional input images are effective in our model, we evaluated the last layer's cross-lingual attentions from the word "feuer (fire)" to the input image and those from the word "wand (wall)" to the input image.

Fig. 3. Original image

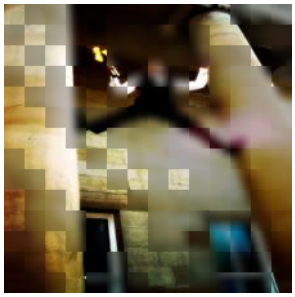

Fig. 4. Cross-lingual Attention from "fire"

Fig. 5. Cross-lingual Attention from "wall"

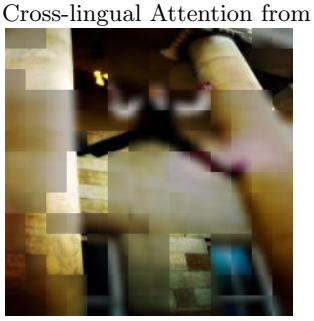

Fig. 6. Self-attention from a region of a wall

Figures 4 and 5 visualize the cross-lingual attentions. In Figs. 4 and 5, the clearer region represents a higher attention score. Figures 4 and 5 show that "feuer (fire)" and "wand (wall)" pay attention to the regions representing the words. This indicates that the input images could help prevent missing translations.

To confirm the effectiveness of our model, we evaluated self-attention from a region of a wall to each region of the input image. Figure 6 illustrates the self-attentions. As shown in the figure, a region of a wall is associated with other regions of a wall through self-attentions in the Transformer encoder of the image encoder. The proposed model might avoid missing the word "wand (wall)" by associating multiple regions in the image, while $CNN + Trans_{Dec}$ might not capture a "wall" from visual features of individual regions. This indicates that our model can determine the associations between regions of an input image, and the associated features can contribute to machine translation.

6 Conclusion

We proposed a new multimodal NMT model that uses a CNN and a Transformer encoder as an image encoder. Our experiments with English-German translation tasks using the Multi30K data set showed that our model outperforms the baseline Transformer NMT model without image inputs and the baseline multimodal NMT model without the Transformer encoder for images. Through the discussions, we showed that our model can determine the relationship between the

visual features from each image region and improve machine translation performance.

In the future, we would like to improve our model by incorporating object-detection technology into our image encoder.

Acknowledgement

The research results have been achieved by "Research and Development of Deep Learning Technology for Advanced Multilingual Speech Translation", the Commissioned Research of National Institute of Information and Communications Technology (NICT) , JAPAN. This work was partially supported by JSPS KAKENHI Grant Number JP18K18110, JP25280084.

References

1. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.
2. Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014.
3. Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421. Association for Computational Linguistics, 2015.
4. Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann Dauphin. Convolutional sequence to sequence learning. In *Proceedings of Thirty-fourth International Conference on Machine Learning (ICML2017)*, pages 1243–1252, 2017.
5. Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323. Association for Computational Linguistics, 2018.
6. Iacer Calixto, Qun Liu, and Nick Campbell. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924. Association for Computational Linguistics, 2017.
7. Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 639–645. Association for Computational Linguistics, 2016.
8. Iacer Calixto and Qun Liu. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003. Association for Computational Linguistics, 2017.

9. Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics, 2016.
10. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015.
11. Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. The MeMAD submission to the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 603–611. Association for Computational Linguistics, 2018.
12. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Proceedings of 3rd International Conference on Learning Representations (ICLR2015)*, abs/1409.1556, 2014.
13. Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics, 2016.
14. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.