Supervised Visual Attention for Multimodal Neural Machine Translation

Tetsuro Nishihara
Ehime University
t_nishihara@ai.cs.ehime-u.ac.jp

Akihiro Tamura
Doshisha University
aktamura@mail.doshisha.ac.jp

Takashi Ninomiya
Ehime University
ninomiya@cs.ehime-u.ac.jp

Yutaro Omote
Ehime University
omote@ai.cs.ehime-u.ac.jp

Hideki Nakayama
The University of Tokyo
nakayama@ci.i.u-tokyo.ac.jp

Abstract

This paper proposed a supervised visual attention mechanism for multimodal neural machine translation (MNMT), trained with constraints based on manual alignments between words in a sentence and their corresponding regions of an image. The proposed visual attention mechanism captures the relationship between a word and an image region more precisely than a conventional visual attention mechanism trained through MNMT in an unsupervised manner. Our experiments on English-German and German-English translation tasks using the Multi30k dataset and on English-Japanese and Japanese-English translation tasks using the Flickr30k Entities JP dataset show that a Transformer-based MNMT model can be improved by incorporating our proposed supervised visual attention mechanism and that further improvements can be achieved by combining it with a supervised cross-lingual attention mechanism (up to +1.61 BLEU, +1.7 METEOR).

1 Introduction

As mainstream machine translation, Neural Machine Translation (NMT) model, widely used since the early days, is the Recurrent Neural Network (RNN)-based NMT with attention mechanism (Luong et al., 2015). This model achieves higher translation accuracy than conventional RNN-based NMT by using a cross-lingual attention mechanism that captures the relationship between words in source and target language sentences. In recent years, the Transformer model (Vaswani et al., 2017) has been attracting much attention because it achieves higher accuracy than methods using RNN and Convolutional Neural Network (CNN). In addition to the conventional cross-lingual attention mechanism, the Transformer model introduces a self-attention mechanism that captures relationships between words in a sentence.

Various studies have been conducted on methods to improve NMT's performance and on one method constraining the cross-lingual attention mechanism (Liu et al., 2016; Mi et al., 2016; Garg et al., 2019). These cited researchers have improved translation performance by using a tool to obtain alignments between words in source and target language sentences in advance and then providing alignments as supervisions to train the cross-lingual attention mechanism.

Multimodal NMT (MNMT) (Barrault et al., 2018) is a machine translation task that aims to improve translation performance by using images in addition to source language sentences because input images are considered useful for disambiguation and omission completion. Helcl et al. (2018) have proposed an MNMT model that introduces a visual attention mechanism into the Transformer-based MNMT model's decoder to capture alignment between words in a sentence and image regions to utilize image features from CNNs. Additionally, Delbrouck and Dupont (2017) have proposed a model that introduces a visual attention mechanism into an RNN-based MNMT model's encoder. However, these visual attention mechanisms are automatically trained through MNMT in an unsupervised manner, and, hence, they do not always capture the relationship between words in a sentence and image regions that they should.

This paper proposes a supervised visual attention mechanism trained with constraints based on manual alignments between words in a sentence and their corresponding image regions to improve MNMT

Proceedings of the 28th International Conference on Computational Linguistics, pages 4304–4314
Barcelona, Spain (Online), December 8-13, 2020

performance. In the MNMT's conventional visual attention mechanism, the image's region where a word should pay attention is not given as a supervision but is learned automatically through MNMT in an unsupervised manner. In our method, we prepare data that shows alignments between words in a source sentence and objects in an image as supervisions and then directly train the visual attention mechanism in the Transformer-based MNMT model's encoder from the supervisions. Note that when only alignments between target language words and image objects can be obtained, our method converts them to alignments between source language words and image objects by using word alignments between source and target sentences.

We experimented with English-German and German-English translation using the Multi30k dataset (Elliott et al., 2016) and with English-Japanese and Japanese-English translation using the Flickr30k Entities JP dataset (Nakayama et al., 2020). These experiments show that the proposed supervised visual attention mechanism improves a Transformer-based MNMT model's performance (i.e., METEOR and BLEU scores).

2 Background

In this section, we overview the Transformer NMT model (Vaswani et al., 2017) and then explain supervised cross-lingual attention for the Transformer model (Garg et al., 2019).

2.1 Transformer NMT

The Transformer NMT model consists of an encoder, which converts the source language sentence into an intermediate representation, and a decoder, which generates a target language sentence from the intermediate representation. The encoder and decoder are stacked with multiple encoder and decoder layers. Each encoder layer has two sub-layers—a self-attention and position-wise fully connected feed-forward network. Each decoder layer consists of three sub-layers—the same two sub-layers used for the encoder and the cross-lingual attention layer. Residual connection (He et al., 2016) and layer normalization (Ba et al., 2016) are used between sub-layers.

The self-attention and the cross-lingual attention mechanisms are computed as follows:

$$\mathrm{Att}(Q, K, V) = AV, \tag{1}$$

$$A = \mathrm{softmax}(\frac{QK^\top}{\sqrt{d_k}}), \tag{2}$$

where A is called *attention matrix*, Q, K, and V are hidden states of the encoder/decoder, and d_k is the dimension of Q, K, and V. In the self-attention mechanism, Q, K, and V are given as the previous sub-layer's output. In the cross-lingual attention mechanism, Q is given as output of the decoder's previous sub-layer, and K and V are given as the encoder's output. The self-attention mechanism calculates relationships between words in the same sentence, and the cross-lingual attention mechanism calculates relationships between words in the source language sentence and words in the target language sentence.

The Transformer is also characterized by a multi-head attention mechanism, in which hidden states are split into sub-spaces to represent various types of information in each sub-space. The multi-head attention mechanism consisting of h heads is expressed as follows:

$$\mathrm{MHA}(Q, K, V) = [head_1; \ldots; head_h]W^O, \tag{3}$$

$$head_i = \mathrm{Att}(QW_i^Q, KW_i^K, VW_i^V), \tag{4}$$

where $[\,]$ denotes the concatenation operation, and $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d_{model} \times d_k}$ and $W^O \in \mathbb{R}^{hd_k \times d_{model}}$ are parameter matrices. d_{model} is the embedding dimensions and $d_k = d_{model}/h$.

The position-wise fully connected feed-forward network is represented as follows:

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2, \tag{5}$$

where $W_1 \in \mathbb{R}^{d_{model} \times d_{ff}}$, $W_2 \in \mathbb{R}^{d_{ff} \times d_{model}}$, $b_1 \in \mathbb{R}^{d_{ff}}$, and $b_2 \in \mathbb{R}^{d_{model}}$ are parameter matrices. d_{model} is the number of input and output dimensions, and d_{ff} is the inner-layers' number of dimensions.

The Transformer introduces positional encoding to incorporate word order information into hidden states. By adding positional encoding to a word's embedded representation, word order information is added.

2.2 Supervised cross-lingual attention

Garg et al. (2019) have proposed a supervised learning method for the Transformer NMT's cross-lingual attention mechanism by providing word alignment between source and target languages for supervision of cross-lingual attentions. Word alignment between languages is obtained by using an alignment tool. The cross-lingual attention mechanism is learned by minimizing the difference between an attention matrix computed by one attention head of the multi-head cross-lingual attention mechanism and word alignment.

The difference is calculated as cross entropy loss as follows:

$$L_a(A) = -\frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N} G_{m,n} \times \log(A_{m,n}), \tag{6}$$

where M and N are the length of a target sentence and that of a source sentence, respectively, A is an attention matrix computed by Equation (2), and G is a word-alignment matrix where $G_{m,n} = 1$ if the n-th source word is aligned to the m-th target word (otherwise $G_{m,n} = 0$). The objective function L for the NMT model using the supervised cross-lingual attention mechanism is given as a loss function that adds $L_a(A)$ to the translation loss L_t as follows:

$$L = L_t + \lambda L_a(A), \tag{7}$$

where λ is a hyperparameter.

3 Proposed method

This section explains the proposed method for supervised training of visual attention mechanisms to improve MNMT translation performance. First, we describe our Transformer-based MNMT model and then propose a supervised training for the visual attention mechanism.

3.1 Architecture of Transformer-based MNMT model

Figure 1 shows our Transformer-based MNMT model's overall image. The model has the image encoder in addition to the source sentence encoder and the decoder. Given an input image, the image encoder first applies CNN to the input image to obtain its image features. Then, self-attention is applied over the CNN output; this enables the image encoder to learn relationships between image regions. Finally, the self-attention layer's output is applied to position-wise fully-connected feed-forward neural networks to generate the image encoder's output.

Next, visual attention (Libovický et al., 2018) is calculated in the source sentence encoder by using the self-attention's output for the source sentence and the image encoder's output. The visual attention mechanism calculates relationships between image regions and words. In Equation (1), Q is the self-attention layer's output in the source sentence encoder, and K and V are the image encoder's output in the case of visual attention.

Figures 5 (c) and 6 (c) show examples of visual attention in which darker cells represent higher attention for the words "man" and "lamp," respectively. The CNN maps the input image to image features that are high-dimensional coarse-grained bitmaps. Each pixel of the coarse-grained bitmaps corresponds to a region in the original bitmaps and has high-dimensional features. For example, image features for Figures 5 (c) and 6 (c) have 7×7 regions, and each region has 2,048 features. In the visual attention mechanism, each word attends to regions in the image features. Therefore, visual attention can be visualized as a region heatmap.

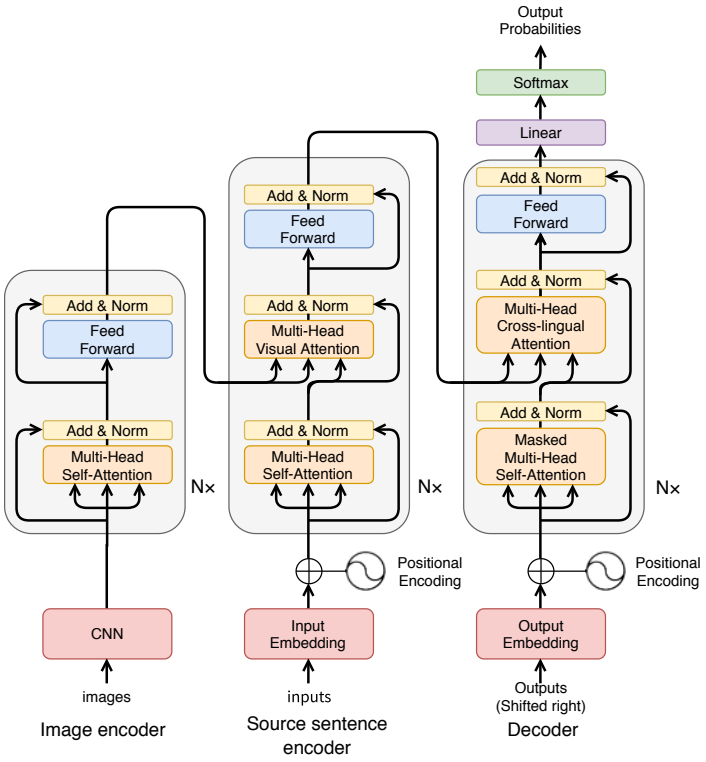


Figure 1: Multimodal Transformer NMT model

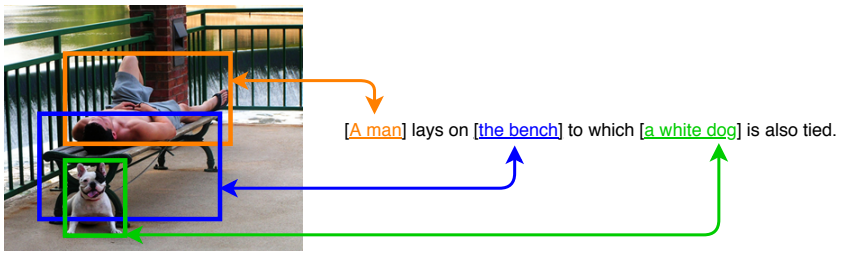


Figure 2: Flickr30k entities sample

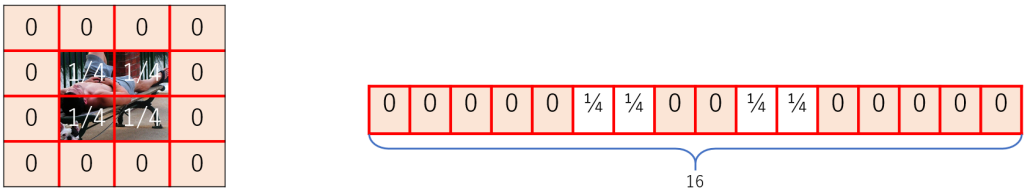
3.2 Supervised training for the visual attention mechanism

The proposed method learns relationships between source language words and image objects by providing constraints based on manual alignments between source language words and their corresponding image regions as supervisions for the visual attention mechanism. Specifically, constraints are applied to attentions between the image encoder's output and the self-attention mechanism's output in the source sentence encoder.

Constraints on the visual attention mechanism are applied to minimize the difference between a *supervision matrix* and the attention matrix in Equation (2) in the visual attention mechanism, where we suppose that the supervision matrix is provided as manually annotated alignment between words in the source language sentence and corresponding image regions. The difference is calculated as cross entropy loss as follows:

$$L_{img\_src}(A) = -\frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N} G_{m,n} \times \log(A_{m,n}), \tag{8}$$

where M and N are the length of a source language sentence and the number of image regions, respectively, A is an attention matrix computed by Equation (2), and G is a supervision matrix, where $G_{m,n} = 1$ if the m-th source word is aligned to the n-th image region (otherwise $G_{m,n} = 0$).



(a) Annotation for the word "man" (b) Linearization

Figure 3: Example of making a supervision matrix for the visual attention mechanism

In this study, we create supervision matrices for the visual attention mechanism by using the Flickr30k entities dataset (Plummer et al., 2017) made from the Flickr30k dataset (Young et al., 2014), in which five caption sentences are attached to one image, and if a word in each captioned sentence is related to an object in the image, the dataset shows to which region in the image the word is related (Figure 2). From this dataset, We extract relationships between words in the source language sentence and objects in the image.

First, we re-scale relationships between words and objects, that is, bounding boxes in the Flickr30k entities dataset, to CNN output regions. For example, if the image is mapped to 4×4 by the CNN used in the image encoder, we compute relationships between a word and its object in the image's 16 regions. When multiple regions are related, values are averaged so that each region is equally related. This means that the value is set to $1/n_{regions}$, where $n_{regions}$ is the number of related regions (Figure 3(a)). Then, two-dimensional regions are linearized to one dimension (Figure 3(b)). This process is performed for all words in the source language sentence. As for words that do not correspond to objects in the image, special tokens are used to process them, inspired by Liu et al. (2016) and Mi et al. (2016). Note that they introduced special tokens for supervised cross-lingual attention. In our method, a special token is attached to the beginning of the matrix in Figure 3(b), and words that do not relate to objects in the image are associated to the special token.

The objective function L for the MNMT with supervised visual attention become as follows:

$$L = L_T + \lambda_1 L_{img\_src}, \tag{9}$$

where L_T is the loss function of the multimodal Transformer NMT model, L_{img_src} is the loss function between the attention matrix of the visual attention mechanism and the supervision matrix, and λ_1 is the hyperparameter to control weights between translation loss and supervised visual attention loss.

3.3 Supervised training of visual attention and cross-lingual attention

In this study, we also introduce the supervised cross-lingual attention explained in Section 2.2 to our MNMT model to improve translation performance. To supervise the cross-lingual attention mechanism, alignment must be acquired between words in the source language sentence and the target language sentence. In this study, we use an alignment tool following Liu et al. (2016) and Garg et al. (2019). As an alignment tool, we use MGIZA (Gao and Vogel, 2008), an implementation of GIZA++ (Och and Ney, 2003), to run multi-threaded on multi-core machines.

When a word in the target language sentence is related to multiple words in the source language sentence, values are averaged so that they are equally related. This means the value is set to $1/n_{words}$, where n_{words} is the number of related words in the source language sentence. Similar to Section 3.2, for words not related to the word in the source language sentence, special tokens are used for processing. In this method, as shown in Figure 4, a special token is set at the beginning of the source language sentence, and words that do not relate to the word in the source language sentence are associated with the special token.

The objective function L for the MNMT with supervised visual attention and cross-lingual attention becomes as follows:

$$L = L_T + \lambda_1 L_{img\_src} + \lambda_2 L_{src\_tgt}, \tag{10}$$

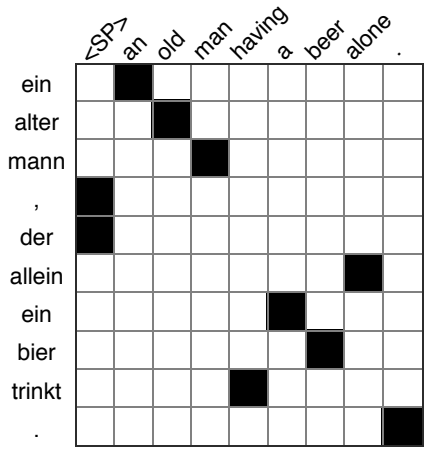


Figure 4: Example of word alignment

	en → de		de → en		en → ja		ja → en	
	B	M	B	M	B	M	B	M
NMT	38.76	57.59	42.58	39.19	43.69	59.27	44.21	40.03
MNMT	38.89	57.35	42.29	39.13	44.09	59.59	44.42	40.03
MNMT+SVA	39.91	58.11	42.52	38.86	44.51	60.03	44.76	40.40
MNMT+SVA+SCA	**40.50**	**59.05**	**43.76**	**39.71**	**44.79**	**60.23**	**45.36**	**40.65**

Table 1: Experiment results. B and M denote BLEU and METEOR, respectively.

where L_{src_tgt} is the supervised cross-lingual attention loss, and λ_1 and λ_2 are hyperparameters to control weights among the translation loss, the supervised visual attention loss, and the supervised cross-lingual attention loss.

4 Experiments

We performed four translation experiments: English-German, German-English, English-Japanese, and Japanese-English. For English-German and German-English translation experiments, we used the Multi30k dataset (Elliott et al., 2016), which consists of a pair of images and their captions. 29,000 pairs of training data, and 1,014 pairs of development data. We used the '2016 test set' as test data; 1,000 test pairs.

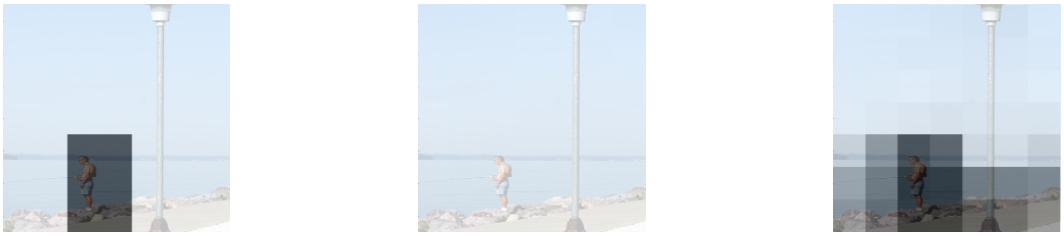
For English-Japanese and Japanese-English translation experiments, we used the Flickr30k Entities JP dataset (Nakayama et al., 2020)[1]. Following text pre-processing in the Multi30k dataset, we applied lowercase, punctuation normalization, and the Moses tokenizer[2] for English sentences. We used Kytea[3] for word segmentation of Japanese sentences. Only pairs in which both English and Japanese sentences were less than 100 words were used in training data. We used 59,516 pairs for training, 2,017 pairs for development, and 2,000 pairs for testing.

As pre-processing, the image was resized to 256×256, and then the image's central part was cropped to 224×224. ResNet50 (He et al., 2016) was used for CNN in the image encoder. Output of the final convolutional layer in ResNet50 was used for image features, which consisted of $7 \times 7 \times 2048$ features. CNN fine-tuning was not performed during training. The image encoder, source sentence encoder, and decoder were stacked with 6 layers. The number of heads was 8, and the number of embedding dimensions was 512. The size of the inner feed-forward network layer was set to 2048. We

[1] https://github.com/nlab-mpg/Flickr30kEnt-JP

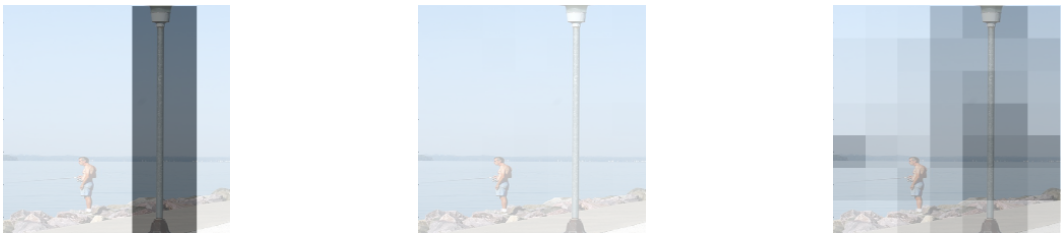
[2] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

[3] http://www.phontron.com/kytea/index-ja.html



(a) Reference (b) MNMT (w/o constraint) (c) MNMT+SVA

Figure 5: Visual attentions for the word “man”



(a) Reference (b) MNMT (w/o constraint) (c) MNMT+SVA

Figure 6: Visual attentions for the word “lamp”

used the Adam optimizer (Kingma and Ba, 2014), and we trained a model for 40 epochs with a mini-batch size of 128. We applied Byte Pair Encoding (BPE) (Sennrich et al., 2016) for English-German and German-English experiments, in which we used 3,492 words for English vocabulary and 4,536 words for German vocabulary. During inference, target sentences were generated by the greedy method.

Translation performance was evaluated by BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014). We selected the epoch model with the highest BLEU value for development data and evaluated performance for the test data. In experiments, we compared: (i) the text-only Transformer NMT model (NMT), (ii) the unconstrained multimodal Transformer NMT model (MNMT), (iii) the model with only the supervised visual attention mechanism (MNMT+SVA), and (iv) the model with supervised both visual and cross-lingual attention mechanisms (MNMT+SVA+SCA). The supervised cross-lingual attention mechanism is the same as Garg et al. (2019)’s supervised attention mechanism described in Section 2.2. Constraints to the attention mechanism for supervision were applied to the sixth layer (i.e., the final layer) of the source language sentence encoder and the decoder. Constraints were placed on one head of the visual attention mechanism and the cross-lingual attention mechanism. In Equation (10), hyperparameters were $\lambda_1 = 0.05$ and $\lambda_2 = 0.05$, following Garg et al. (2019). In the model with only the supervised visual attention mechanism, hyperparameters were set to $\lambda = 0.05$. Note that alignments between words and image objects are tagged only in English sentences for both experimental datasets. As for English-German and English-Japanese tasks, supervision matrices for the visual attention mechanism were created directly from manually annotated alignments. As for German-English and Japanese-English tasks, manually annotated alignments in English sentences were first converted to alignments between German/Japanese words and image objects by using word alignments obtained by MGIZA, and then supervision matrices for the visual attention mechanism were created from converted alignments.

Table 1 displays experimental results, showing that in all tasks, BLEU and METEOR were the highest when both visual and cross-lingual attention mechanisms were supervised. These results demonstrate the proposed supervised visual attention mechanisms’ effectiveness.

Source: 赤いシャツを着た男の子が、黄色いシャベルで砂を掘っている。

MNMT: a boy in a red shirt is shoveling sand .

MNMT+SVA: a boy in a red shirt is digging in the sand with a yellow shovel .

Reference: a boy wearing a red shirt digs into the sand with a yellow shovel .

(a) Examples of translations in the Japanese-English translation task

Source: ein rothaariger mann mit dreadlocks sitzt und spielt auf einer akustischen gitarre .

MNMT: a man with red-hair sits on an acoustic guitar .

MNMT+SVA: a red-haired man with dreadlocks is sitting and playing an acoustic guitar .

Reference: a red-haired man with dreadlocks is sitting playing and acoustic guitar .

(b) Examples of translations in the German-English translation task

Figure 7: Translation examples

5 Analysis

5.1 Examples of visual attentions

Figures 5 and 6 show visual attentions from the word "man" and "lamp", respectively, in the English-Japanese translation task's test data. In the figures, the darker region represents a higher attention score, and when the constraint on visual attentions is not used, each of the two words pays equal attention to the entire image (Figures 5(b) and 6(b)). In contrast, when supervised visual constraint is used, image regions related with each of the two words can be identified more precisely through visual attentions (Figures 5(c) and 6(c)). These results demonstrate that our supervised visual attention encourages visual attentions to attend to each word's related regions.

5.2 Examples of translations

From examples of translation results on test data of Japanese-English and German-English tasks, Figure 7 shows that sentences translated by the MNMT model without proposed supervised visual attentions do not include some information from the source language sentence. For example, the "MNMT" translation loses the information "a yellow shovel" in Figure 7(a) and the information on the man's characteristic "dreadlocks" in Figure 7(b). In contrast, the MNMT model with the proposed mechanism translates the source language sentences correctly. This might be because the source language sentences could be encoded more properly by linking each source language word with the image's related regions through the proposed supervised visual attention mechanism, indicating that the proposed supervised visual attention mechanism avoids under-generation errors.

5.3 Experiments with manual word alignments

In the Flickr30k Entities JP dataset, word alignments between source language words and target language words are manually tagged. We evaluate the translation performance of our proposed models, "MNMT+SVA" and "MNMT+SVA+SCA", when using manual word alignments rather than automatic word alignments. Table 2 shows that in the experimental results, both "MNMT+SVA" and "MNMT+SVA+SCA" outperform "MNMT", and "MNMT+SVA+SCA" is better than "MNMT+SVA" on both English-Japanese and Japanese-English tasks. In other words, our proposed supervised attention mechanism using manual word alignments is also effective, and combination with the supervised cross-lingual attention mechanism achieves further improvement.

	en → ja		ja → en	
	B	M	B	M
NMT	43.69	59.27	44.21	40.03
MNMT	44.09	57.35	44.42	40.03
MNMT+SVA	44.51	60.03	44.76	40.40
MNMT+SVA+SCA	44.85	60.35	44.90	40.41

Table 2: Translation performance when using manual word alignments

6 Related Work

NMT has been improved by introducing a supervised cross-lingual attention mechanism trained with constraints based on automatic or manual word alignments between source and target language words. Mi et al. (2016) and Liu et al. (2016) have proposed supervised cross-lingual attention mechanisms for RNN-based NMT models, and Garg et al. (2019) have proposed one for a Transformer-based NMT model. Note that, as far as we know, supervised cross-lingual attention mechanisms have not been applied to MNMT.

Recently, some studies have proved the effectiveness of images for machine translation (Elliott, 2018; Caglayan et al., 2019). Various models have been proposed for MNMT. In the early days, RNN-based MNMT was a dominant architecture (e.g., (Calixto et al., 2017; Caglayan et al., 2017; Delbrouck and Dupont, 2017)) that extended RNN-based NMT (Bahdanau et al., 2015). In recent years, Transformer-based MNMT models have been actively studied (e.g., (Helcl et al., 2018; Libovický et al., 2018; Grönroos et al., 2018; Ive et al., 2019; Zhang et al., 2020)), based on Transformer NMT (Vaswani et al., 2017). Most MNMT models have incorporated an input image's features with a visual attention mechanism. Some studies have introduced a visual attention mechanism that captures relationships between source language words and image regions (Delbrouck and Dupont, 2017; Zhang et al., 2020), while others have used a visual attention mechanism that captures relationships between target language words and image regions (Calixto et al., 2017; Helcl et al., 2018; Libovický et al., 2018; Ive et al., 2019; Takushima et al., 2019). Note that these visual attention mechanisms were trained in an unsupervised manner, and, as far as we know, a supervised visual attention mechanism has not yet been proposed.

7 Conclusion

We have proposed an MNMT supervised visual attention mechanism that generates supervisions for visual attentions from manual alignments between words in a sentence and their corresponding regions of an image and trains visual attentions of an MNMT model's encoder from the supervisions. Experiments showed that our proposed supervised attention mechanism improved a Transformer-based MNMT model on English-German and German-English translation tasks using the Multi30k dataset and on English-Japanese and Japanese-English translation tasks using the Flickr30k Entities JP dataset.

In future work, we will explore our proposed supervised visual attention mechanism's effectiveness for other Transformer-based MNMT models.

Acknowledgements

These research results were 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 JP20K19864.

References

Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *ArXiv*, abs/1607.06450.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. 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.

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. 2017. LIUM-CVC submissions for WMT17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*, pages 432–439, Copenhagen, Denmark, September. Association for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. 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, Vancouver, Canada, July. Association for Computational Linguistics.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017. Modulating and attending the source image during encoding improves multimodal translation. *CoRR*, abs/1712.03449.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium, October-November. Association for Computational Linguistics.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio, June. Association for Computational Linguistics.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4452–4461, Hong Kong, China, November. Association for Computational Linguistics.

Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. 2018. The MeMAD submission to the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 603–611, Belgium, Brussels, October. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Jindřich Helcl, Jindřich Libovický, and Dušan Variš. 2018. CUNI system for the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 616–623.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6538, Florence, Italy, July. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Jindřich Libovický, Jindřich Helcl, and David Mareček. 2018. Input combination strategies for multi-source transformer decoder. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 253–260, Brussels, Belgium, October. Association for Computational Linguistics.

Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Neural machine translation with supervised attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3093–3102.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Haitao Mi, Zhiguo Wang, and Abe Ittycheriah. 2016. Supervised attentions for neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2283–2288, Austin, Texas, November. Association for Computational Linguistics.

Hideki Nakayama, Akihiro Tamura, and Takashi Ninomiya. 2020. A visually-grounded parallel corpus with phrase-to-region linking. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4204–4210, Marseille, France, May. European Language Resources Association.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123(1):74–93.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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, Berlin, Germany, August. Association for Computational Linguistics.

Hiroki Takushima, Akihiro Tamura, Takashi Ninomiya, and Hideki Nakayama. 2019. Multimodal neural machine translation using cnn and transformer encoder. In *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2019)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural machine translation with universal visual representation. In *8th International Conference on Learning Representations, ICLR 2020, April 26th through May 1st*.