

# Double Attention-based Multimodal Neural Machine Translation with Semantic Image Regions

Yuting Zhao<sup>1</sup>, Mamoru Komachi<sup>1</sup>, Tomoyuki Kajiwara<sup>2</sup>, Chenhui Chu<sup>2</sup>

<sup>1</sup>Tokyo Metropolitan University, 6-6 Asahigaoka, Hino, Tokyo 191-0065, Japan

<sup>2</sup>Osaka University, 2-8 Yamadaoka, Suita, Osaka 565-0871, Japan

zhao-yuting@ed.tmu.ac.jp

komachi@tmu.ac.jp

{kajiwara, chu}@ids.osaka-u.ac.jp

## Abstract

Existing studies on multimodal neural machine translation (MNMT) have mainly focused on the effect of combining visual and textual modalities to improve translations. However, it has been suggested that the visual modality is only marginally beneficial. Conventional visual attention mechanisms have been used to select the visual features from equally-sized grids generated by convolutional neural networks (CNNs), and may have had modest effects on aligning the visual concepts associated with textual objects, because the grid visual features do not capture semantic information. In contrast, we propose the application of semantic image regions for MNMT by integrating visual and textual features using two individual attention mechanisms (double attention). We conducted experiments on the Multi30k dataset and achieved an improvement of 0.5 and 0.9 BLEU points for English→German and English→French translation tasks, compared with the MNMT with grid visual features. We also demonstrated concrete improvements on translation performance benefited from semantic image regions.

## 1 Introduction

Neural machine translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015) has achieved state-of-the-art translation performance. Recently,

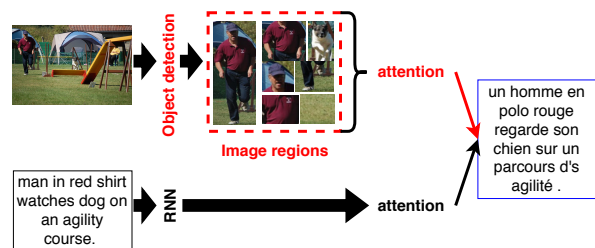


Figure 1: Overview of our MNMT model.

many studies (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018) have been increasingly focusing on incorporating multimodal contents, particularly images, to improve translations. Hence, researchers in this field have established a shared task called multimodal machine translation (MMT), which consists of translating a target sentence from a source language description into another language using information from the image described by the source sentence.

The first MMT study by (Elliott et al., 2015) demonstrated the potential of improving the translation quality by using image. To effectively use an image, several subsequent studies (Gao et al., 2015; Huang et al., 2016; Calixto and Liu, 2017) incorporated global visual features extracted from the entire image by convolutional neural networks (CNNs) into a source word sequence or hidden states of a recurrent neural network (RNN). Furthermore, other studies started using local visual features in the context of an attention-based NMT. These features were extracted from equally-sized grids in an image by a CNN. For instance, multimodal attention (Caglayan et al., 2016b) has been designed for a mix of text and local visual features. Additionally, double attention mechanisms (Calixto et al., 2017) have been proposed for text

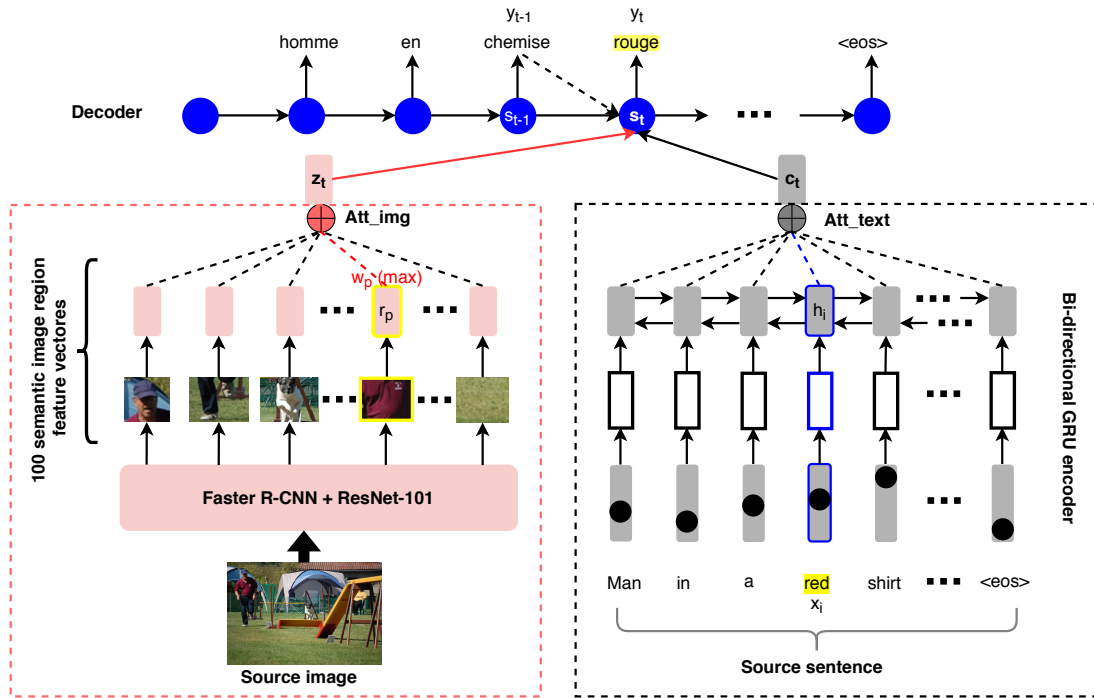


Figure 2: Our model of double attention-based MNMT with semantic image regions.

and local visual features, respectively. Although previous studies improved the use of local visual features and the text modality, these improvements were minor. As discussed in (Delbrouck and Dupont, 2017), these local visual features may not be suitable to attention-based NMT, because the attention mechanism cannot understand complex relationships between textual objects and visual concepts.

Other studies utilized richer local visual features to MNMT such as dense captioning features (Delbrouck et al., 2017). However, their efforts have not convincingly demonstrated that visual features can improve the translation quality. Caglayan et al. (2019) demonstrated that, when the textual context is limited, visual features can assist in generating better translations. MMT models disregard visual features because the quality of the image features or the way in which they are integrated into the model are not satisfactory. Therefore, which types of visual features are suitable to MNMT, and how these features should be integrated into MNMT, still remain open questions.

This paper proposes the integration of semantic image region features into a double attention-based NMT architecture. In particular, we combine object detection with a double attention mechanism to fully exploit visual features for MNMT. As shown in Figure 1, we use the semantic im-

age region features extracted by an object detection model, namely, Faster R-CNN (Ren et al., 2015). Compared with the local visual features extracted from equally-sized grids, we believe that our semantic image region features contain object attributes and relationships that are important to the source description. Moreover, we expect that the model would be capable of making selective use of the extracted semantic image regions when generating a target word. To this end, we integrate semantic image region features using two attention mechanisms: one for the semantic image regions and the other one for text. Code and pre-trained models are publicly available at: <https://github.com/Zhao-Yuting/MNMT-with-semantic-regions>.

The main contributions of this study are as follows:

- We verified that the translation quality can significantly improve by leveraging semantic image regions.
- We integrated semantic image regions into a double attention-based MNMT, which resulted in the improvement of translation performance above the baselines.
- We carried out a detailed analysis to identify the advantages and shortcomings of the proposed model.

## 2 MNMT with Semantic Image Regions

In Figure 2, our model comprises three parts: the source-sentence side, source-image side, and decoder. Inspired by (Calixto et al., 2017), we integrated the visual features using an independent attention mechanism. From the source sentence  $X = (x_1, x_2, x_3, \dots, x_n)$  to the target sentence  $Y = (y_1, y_2, y_3, \dots, y_m)$ , the image-attention mechanism focuses on all semantic image regions to calculate the image context vector  $z_t$ , while the text-attention mechanism computes the text context vector  $c_t$ . The decoder uses a conditional gated recurrent unit (cGRU)<sup>1</sup> with attention mechanisms to generate the current hidden state  $s_t$  and target word  $y_t$ .

At time step  $t$ , first, a hidden state proposal  $\hat{s}_t$  is computed in cGRU, as presented below, and then used to calculate the image context vector  $z_t$  and text context vector  $c_t$ .

$$\begin{aligned}\hat{\xi}_t &= \sigma(W_\xi E_Y[y_{t-1}] + U_\xi s_{t-1}) \\ \hat{\gamma}_t &= \sigma(W_\gamma E_Y[y_{t-1}] + U_\gamma s_{t-1}) \\ \hat{s}_t &= \tanh(W E_Y[y_{t-1}] + \hat{\gamma}_t \odot (U s_{t-1})) \\ \hat{s}_t &= (1 - \hat{\xi}_t) \odot \hat{s}_t + \hat{\xi}_t \odot s_{t-1}\end{aligned}\quad (1)$$

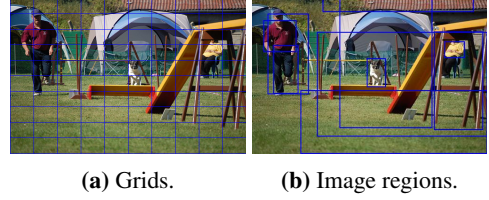
where  $W_\xi$ ,  $U_\xi$ ,  $W_\gamma$ ,  $U_\gamma$ ,  $W$ , and  $U$  are training parameters;  $E_Y$  is the target word vector.

### 2.1 Source-sentence side

The source sentence side comprises a bi-directional GRU encoder and ‘‘soft’’ attention mechanism (Xu et al., 2015). Given a source sentence  $X = (x_1, x_2, x_3, \dots, x_n)$ , the encoder updates the forward GRU hidden states by reading  $x$  from left to right, generates the forward annotation vectors  $(\vec{h}_1, \vec{h}_2, \vec{h}_3, \dots, \vec{h}_n)$ , and finally updates the backward GRU with the annotation vectors  $(\overleftarrow{h}_1, \overleftarrow{h}_2, \overleftarrow{h}_3, \dots, \overleftarrow{h}_n)$ . By concatenating the forward and backward vectors  $h_i = [\vec{h}_i; \overleftarrow{h}_i]$ , every  $h_i$  encodes the entire sentence while focusing on the  $x_i$  word, and all words in a sentence are denoted as  $C = (h_1, h_2, \dots, h_n)$ . At each time step  $t$ , the text context vector  $c_t$  is generated as follows:

$$\begin{aligned}e_{t,i}^{\text{text}} &= (V^{\text{text}})^T \tanh(U^{\text{text}} \hat{s}_t + W^{\text{text}} h_i) \\ \alpha_{t,i}^{\text{text}} &= \text{softmax}(e_{t,i}^{\text{text}}) \\ c_t &= \sum_{i=1}^n \alpha_{t,i}^{\text{text}} h_i\end{aligned}\quad (2)$$

<sup>1</sup><https://github.com/nyu-dl/dl4mt-tutorial/blob/master/docs/cgru.pdf>



**Figure 3:** Comparing between (a) coarse grids and (b) semantic image regions.

where  $V^{\text{text}}$ ,  $U^{\text{text}}$ , and  $W^{\text{text}}$  are training parameters;  $e_{t,i}^{\text{text}}$  is the attention energy;  $\alpha_{t,i}^{\text{text}}$  is the attention weight matrix of the source sentence.

### 2.2 Source-image side

In this part, we discuss the integration of semantic image regions into MNMT using an image attention mechanism.

**Semantic image region feature extraction.** As shown in Figure 3, instead of extracting equally-sized grid features using CNNs, we extract semantic image region features using object detection. This study applied the Faster R-CNN in conjunction with the ResNet-101 (He et al., 2016) CNN pre-trained on Visual Genome (Krishna et al., 2017) to extract 100 semantic image region features from each image. Each semantic image region feature is a vector  $r$  with a dimension of 2048, and all of these features in an image are denoted as  $R = (r_1, r_2, r_3, \dots, r_{100})$ .

**Image-attention mechanism.** The image-attention mechanism is also a type of ‘‘soft’’ attention. This mechanism focuses on 100 semantic image region feature vectors at every time step and computes the image context vector  $z_t$ .

First, we calculate the attention energy  $e_{t,p}^{\text{img}}$ , which is an attention model that scores the degree of output matching between the inputs around position  $p$  and the output at position  $t$ , as follows:

$$e_{t,p}^{\text{img}} = (V^{\text{img}})^T \tanh(U^{\text{img}} \hat{s}_t + W^{\text{img}} r_p) \quad (3)$$

where  $V^{\text{img}}$ ,  $U^{\text{img}}$ , and  $W^{\text{img}}$  are training parameters. Then the weight matrix  $\alpha_{t,p}^{\text{img}}$  of each  $r_p$  is computed as follows:

$$\alpha_{t,p}^{\text{img}} = \text{softmax}(e_{t,p}^{\text{img}}) \quad (4)$$

At each time step, the image-attention mechanism dynamically focuses on the semantic image region features and computes the image context vector  $z_t$ ,

as follows:

$$z_t = \beta_t \sum_{p=1}^{100} \alpha_{t,p}^{\text{img}} r_p \quad (5)$$

For  $z_t$ , at each decoding time step  $t$ , a gating scalar  $\beta_t \in [0, 1]$  (Xu et al., 2015) is used to adjust the proportion of the image context vector according to the previous hidden state of the decoder  $s_{t-1}$ .

$$\beta_t = \sigma(W_\beta s_{t-1} + b_\beta) \quad (6)$$

where  $W_\beta$  and  $b_\beta$  are training parameters.

### 2.3 Decoder

At each time step  $t$  of the decoder, the new hidden state  $s_t$  is computed in cGRU, as follows:

$$\begin{aligned} \xi_t &= \sigma(W_\xi^{\text{text}} c_t + W_\xi^{\text{img}} z_t + \bar{U}_\xi \hat{s}_t) \\ \gamma_t &= \sigma(W_\gamma^{\text{text}} c_t + W_\gamma^{\text{img}} z_t + \bar{U}_\gamma \hat{s}_t) \\ \bar{s}_t &= \tanh(W^{\text{text}} c_t + W^{\text{img}} z_t + \gamma_t \odot (\bar{U} \hat{s}_t)) \\ s_t &= (1 - \xi_t) \odot \bar{s}_t + \xi_t \odot \hat{s}_t \end{aligned} \quad (7)$$

where  $W_\xi^{\text{text}}$ ,  $W_\xi^{\text{img}}$ ,  $\bar{U}_\xi$ ,  $W_\gamma^{\text{text}}$ ,  $W_\gamma^{\text{img}}$ ,  $\bar{U}_\gamma$ ,  $W^{\text{text}}$ ,  $W^{\text{img}}$ , and  $\bar{U}$  are model parameters;  $\xi_t$  and  $\gamma_t$  are the output of the update/reset gates;  $\bar{s}_t$  is the proposed updated hidden state.

Finally, the conditional probability of generating a target word  $p(y_t | y_{t-1}, s_t, C, R)$  is computed by a nonlinear, potentially multi-layered function, as follows:

$$\text{softmax}(L_o \tanh(L_s s_t + L_c c_t + L_z z_t + L_w E_Y[y_{t-1}])) \quad (8)$$

where  $L_o$ ,  $L_s$ ,  $L_c$ ,  $L_z$ , and  $L_w$  are training parameters.

## 3 Experiments

### 3.1 Dataset

We conducted experiments for the English→German (En→De) and English→French (En→Fr) tasks using the Multi30k dataset (Elliott et al., 2016). The dataset contains 29k training and 1,014 validation images. For testing, we used the 2016 testset, which contains 1,000 images. Each image was paired with image descriptions expressed by both the original English sentences and the sentences translated into multiple languages.

For preprocessing, we lowercased and tokenized the English, German, and French descriptions with

the scripts in the Moses SMT Toolkit.<sup>2</sup> Subsequently, we converted the space-separated tokens into subword units using the byte pair encoding (BPE) model.<sup>3</sup> Finally, the number of subwords in a description was limited to a maximum of 80.

### 3.2 Settings

**Ours.** We integrated the semantic image regions by modifying the double attention model of (Calixto et al., 2017). In the source-sentence, we reused the original implementation. In the source-image, we modified the image attention mechanism to focus on 100 semantic image region features with a dimension of 2048 at each time step. The parameter settings were consistent with the baseline doubly-attentive MNMT model, wherein we set the hidden state dimension of the 2-layer GRU encoder and 2-layer cGRU decoder to 500, source word embedding dimension to 500, batch size to 40, beam size to 5, text dropout to 0.3, and image region dropout to 0.5. We trained the model using stochastic gradient descent with ADADELTA (Zeiler, 2012) and a learning rate of 0.002, for 25 epochs. Finally, after both the validation perplexity and accuracy converged, we selected the converged model for testing.

**Baseline Doubly-attentive MNMT.** We trained a doubly-attentive MNMT model<sup>4</sup> as a baseline. For the text side, the implementation was based on OpenNMT model.<sup>5</sup> For the image side, attention was applied to the visual features extracted from  $7 \times 7$  image grids by CNNs. For the image feature extraction, we compared three pre-trained CNN methods: VGG-19, ResNet-50, and ResNet-101.

**Baseline OpenNMT.** We trained a text-only attentive NMT model using OpenNMT as the other baseline. The model was trained on En→De and En→Fr, wherein only the textual part of Multi30k was used. The model comprised a 2-layer bidirectional GRU encoder and 2-layer cGRU decoder with attention.

For baselines, we used the original implementations and ensured the parameters were consistent with our model.

<sup>2</sup><https://github.com/moses-smt/mosesdecoder>

<sup>3</sup><https://github.com/rsennrich/subword-nmt>

<sup>4</sup><https://github.com/iacercalixto/MultimodalNMT>

<sup>5</sup><https://github.com/OpenNMT/OpenNMT-py>

Model	En→De		En→Fr	
	BLEU	METEOR	BLEU	METEOR
OpenNMT (text-only)	34.7±0.3	53.2±0.4	56.6±0.1	72.1±0.1
Doubly-attentive MNMT (VGG-19)	36.4±0.2	55.0±0.1	57.4±0.4	72.4±0.4
Doubly-attentive MNMT (ResNet-50)	36.5±0.2	54.9±0.4	57.5±0.4	72.6±0.4
Doubly-attentive MNMT (ResNet-101)	36.5±0.3	54.9±0.3	57.3±0.2	72.4±0.2
<b>Ours (Faster R-CNN + ResNet-101)</b>	<b>37.0±0.1<sup>†</sup></b>	<b>55.3±0.2</b>	<b>58.2±0.5<sup>†‡</sup></b>	<b>73.2±0.2</b>
vs. OpenNMT (text-only)	(↑ 2.3)	(↑ 2.1)	(↑ 1.6)	(↑ 1.1)
vs. Doubly-attentive MNMT (ResNet-101)	(↑ 0.5)	(↑ 0.4)	(↑ 0.8)	(↑ 0.9)
Caglayan et al. (2017) (text-only)	38.1±0.8	57.3±0.5	52.5±0.3	69.6±0.1
Caglayan et al. (2017) (grid)	37.0±0.8	57.0±0.3	53.5±0.8	70.4±0.6
Caglayan et al. (2017) (global)	38.8±0.5	57.5±0.2	54.5±0.8	71.2±0.4

**Table 1:** BLEU and METEOR scores for different models on the En→De and En→Fr 2016 testset of Multi30k. All scores are averages of three runs. We present the results using the mean and the standard deviation. † and ‡ indicate that the result is significantly better than OpenNMT and double-attentive MNMT at p-value < 0.01, respectively. Additionally, we report the best results of using grid and global visual features on Multi30k dataset according to (Caglayan et al., 2017), which is the state-of-the-art system for En→De translation on this dataset.

### 3.3 Evaluation

We evaluated the quality of the translation according to the token level BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014) metrics. We trained all models (baselines and proposed model) three times and calculated the BLEU and METEOR scores, respectively. Based on the calculation results, we report the mean and standard deviation over three runs.

Moreover, we report the statistical significance with bootstrap resampling (Koehn, 2004) using the merger of three test translation results. We defined the threshold for the statistical significance test as 0.01, and report only if the p-value was less than the threshold.

## 4 Results

In Table 1, we present the results for the OpenNMT, doubly-attentive MNMT and our model on Multi30k dataset. Additionally, we also compared with Caglayan et al. (2017), which achieved the best performance under the same condition with our experiments.

Comparing the baselines, the doubly-attentive MNMT outperformed OpenNMT. Because there did not exist a big difference amongst the three image feature extraction methods for the doubly-attentive MNMT model, we only used ResNet-101 in our model.

Compared with the OpenNMT baseline, the pro-

posed model improved both BLEU scores and METEOR scores for En→De and En→Fr tasks. Additionally, the results of our proposed model are significantly better than the results obtained by the baseline with a p-value < 0.01 for both tasks.

Compared with the doubly-attentive MNMT (ResNet-101) baseline, the proposed model also improved the BLEU scores and METEOR scores for both tasks. Moreover, the results are significantly better than the baseline results with a p-value < 0.01 for En→Fr task.

For comparison with Caglayan et al. (2017), we report their results for the text-only NMT baseline, grid and global visual features for MNMT method. With the grid visual features, their results surpassed the text-only NMT baseline for En→Fr, but failed to surpass the text-only NMT baseline for En→De with regard to both metrics. With the global visual features, their results surpassed the text-only NMT baseline.

For En→De, though Caglayan et al. (2017) (global) achieved higher scores than our model, the improvements were minor. In terms of relative improvement compared with the text-only NMT baseline, their results improved the BLEU score by 1.8% and METEOR score by 0.3%. In contrast, our model improved the BLEU score by 6.6% and METEOR score by 3.9%.

For En→Fr, our results outperform Caglayan et al. (2017) (global) with regard to both metrics.

In terms of relative improvement compared with the text-only NMT baseline, their results improved the BLEU score by 1.9% and METEOR score by 1.1% with the grid visual features and improved the BLEU score by 3.8% and METEOR score by 2.3% with the global visual features. Our model improved the BLEU score by 2.8% and METEOR score by 1.5%.

## 5 Analysis

### 5.1 Pairwise evaluation of translations

We randomly investigated 50 examples from the En→Fr task to evaluate our model in detail. We compared the translations of our model with the baselines to identify improvement or deterioration in the translation. Then we categorized all examples into five types: 1) those whose translation performance were better than both baselines; 2) those whose translation performance were better than the doubly-attentive MNMT (ResNet-101) baseline; 3) those whose translation performance were better than the OpenNMT baseline; 4) those whose translation performance did not change; 5) those whose translation performance deteriorated. We counted the number and proportion of all types.

In Table 2, we can see that in nearly half of the examples, the translation performance is better than at least one baseline. Moreover, amongst a total of 50 examples, 14 examples are better than the doubly-attentive MNMT (ResNet-101) baseline and just two examples of local deterioration were found compared with the baselines.

### 5.2 Qualitative analysis

In Figure 4, we chose four examples to analyze our model in detail. The first two rows explain the advantages of our model, while the last two rows explain the shortcomings.

At each time step, the semantic image region is shown with deep or shallow transparency in the image, according to its assigned attention weight. As the weight increases, the image region becomes more transparent. Considering the number of 100 bounding boxes in one image and the overlapping areas, we visualized the top five weighted semantic image regions. The most weighted image region is indicated by the blue lines, and the target word generated at that time step is indicated by the red text along with the bounding box. Then, we analyzed whether the semantic image regions had a positive or negative effect at the time step when

Better than both baselines	8	(16%)
Better than MNMT baseline	6	(12%)
Better than NMT baseline	10	(20%)
No change	24	(48%)
Deteriorated	2	(4%)

**Table 2:** The amount and proportion of each type of examples in all investigated examples.




the target word was generated.

**Advantages.** In the first row, we can see that our model is better at translating the verb “grabbing” compared with both baselines. For the text-only OpenNMT, the translation of the word “grabbing” is incorrect. In English it is translated as “strolling with.” The doubly-attentive MNMT (ResNet-101) translated “grab” into “agrippe,” which failed to transform the verb into the present participle form. In contrast, although the reference is “saisissant” and our model generated “agrippant,” the two words are synonyms. Our approach improved the translation performance both in terms of meaning and verb deformation, owing to the semantic image regions. We visualized the consecutive time steps of generating the word “agrippant” in context. Along with the generation of “agrippant,” the attention focused on the image region where the action was being performed, and thus captured the state of the action at that moment.

In the second row, the noun “terrier” could not be translated by the baselines. This word means “a lively little dog” in English. As we can see, when the target word “terrier” was generated in our model, the attended semantic image region at that time step provided the exact object-level visual feature to the translation.

**Shortcomings.** The example in the third row reflects improvement and deficiency. Both baselines lack the sentence components of the adverbial “happily.” In contrast, our model translated “happily” into “joyusement,” which is a better translation than both baselines. However, according to the image, the semantic image region with the largest attention weight did not carry the facial expression of a boy.

Although the maximum weight of the semantic image region was not accurately assigned, other heavily weighted semantic image regions, which contain the object attributes, may assist the translation. There may be two reasons for this: the func-

			Source (En) a man in a blue coat <b>grabbing</b> a young boy's shoulder . Reference (Fr) un homme en manteau bleu <b>saisissant</b> l's épaule d's un jeune garçon . NMT un homme en manteau bleu <b>se baladant avec (strolling with)</b> l's épaule d's un jeune garçon . MNMT un homme en manteau bleu <b>agrippe (grab)</b> l's épaule d's un jeune garçon . Ours un homme en manteau bleu <b>agrippant (grabbing)</b> l's épaule d's un jeune garçon .
			Source (En) a boston <b>terrier</b> is running on lush green grass in front of a white fence . Reference (Fr) un <b>terrier</b> de boston court sur l's herbe verdoyante devant une clôture blanche . NMT un <b>garde (guard)</b> de boston court sur l's herbe souple devant une clôture blanche . MNMT un <b>croreur (croror)</b> court sur l's herbe verte devant une clôture blanche . Ours un <b>terrier (terrier)</b> de boston terrier court sur l's herbe verte devant une clôture blanche .
			Source (En) a small child wearing a blue and white t-shirt <b>happily holding</b> a yellow plastic <b>alligator</b> . Reference (Fr) un petit enfant avec un t-shirt bleu et blanc <b>tenant joyeusement</b> un <b>alligator</b> en plastique jaune . NMT un petit enfant vêtu d's un t-shirt bleu et blanc <b>brandissant (brandishing)</b> une <b>bouteille (bottle)</b> en plastique jaune . MNMT un petit enfant vêtu d's un t-shirt bleu et blanc <b>tenant (holding)</b> un <b>fusil (rifle)</b> en plastique jaune . Ours un petit enfant vêtu d's un t-shirt bleu et blanc <b>met (put) joyeusement (happily) une forme (shape)</b> en plastique jaune .
			Source (En) men playing volleyball , with one player missing the ball but hands still <b>in the air</b> . Reference (Fr) des hommes jouant au volleyball , avec un joueur ratant le ballon mais avec les mains toujours <b>en l's air</b> . NMT des hommes jouant au volleyball , un joueur à l's attraper , mais les autres mains ayant toujours <b>dans les airs</b> . MNMT des hommes jouant au volley-ball , avec un joueur qui le regarde <b>dans les airs (in the air)</b> . Ours des hommes jouant au volleyball , avec un joueur qui passer le ballon mais les mains <b>du vol (of the flight)</b> .

**Figure 4:** Translations from the baselines and our model for comparison. We highlight the words that distinguish the results. Blue words are marked for better translation and red words are marked for worse translation. We also visualize the semantic image regions that the words attend to.

tion of the attention mechanism is not sufficiently effective, or there exists an excessive amount of semantic image regions.

On the other hand, for the generation of the word “holding” and “alligator,” the most weighted semantic image regions were not closely attended to. There was a slight deviation between the image regions and semantics. Owing to the inaccuracy of the image region that was drawn upon the object, the semantic feature was not adequately extracted. This indicates that the lack of specificity in the visual feature quality can diminish the detail of the information being conveyed.

In the last row, we presented one of the two examples with local deterioration. The “air” is correctly translated by baselines. However, our model translated “in the air” into “du vol (of the flight).” We observed that the transparent semantic image regions with the five top weights in the image were very scattered and unconnected. Amongst them, none of the semantic image regions matched the feature of “air.” We speculate that the word “air” is difficult to interpret depending on visual features. On the other hand, our model translated it into “vol (flight),” which is close to another meaning of the polysemous “air,” not something else.

**Summary.** In our model, the improvement of translation performance benefits from semantic image regions. The semantic image region visual features include the object, object attributes, and scene understanding, may assist the model in performing a better translation on the verb, noun, adverb and so on.

On the other hand, there are some problems:

- In some cases, although the translation performance improved, the image attention mechanism did not assign the maximum weight to the most appropriate semantic image region.
- When the object attributes cannot be specifically represented by image regions, incorrect visual features conveyed by the semantic image regions may interfere with the translation performance.
- If the image attention mechanism leads to the wrong focused semantic image region, it will bring negative effects on translation performance.

In our investigation, we did not identify any clear examples of successful disambiguation. In

contrast, there is one example of detrimental results upon disambiguation. If the semantic image regions did not have good coverage of the semantic features or the image attention mechanism worked poorly, the disambiguation of polysemous words would not only fail, but ambiguous translation would also take place.

## 6 Related Work

From the first shared task at WMT 2016,<sup>6</sup> many MMT studies have been conducted. Existing studies have fused either global or local visual image features into MMT.

### 6.1 Global visual feature

Calixto and Liu (2017) incorporated global visual features into source sentence vectors and encoder/decoder hidden states. Elliott and Kádár (2017) utilized global visual features to learn both machine translation and visually grounding task simultaneously. As for the best system in WMT 2017,<sup>7</sup> Caglayan et al. (2017) proposed different methods to incorporate global visual features based on attention-based NMT model such as initial encoder/decoder hidden states using element-wise multiplication. Delbrouck and Dupont (2018) proposed a variation of the conditional gated recurrent unit decoder, which receives the global visual features as input. Calixto et al. (2019) incorporated global visual features through latent variables. Although their results surpassed the performance of the NMT baseline, the visual features of an entire image are complex and non-specific, so that the effect of the image is not fully exerted.

### 6.2 Local visual features

**Grid visual features.** Fukui et al. (2016) applied multimodal compact bilinear pooling to combine the grid visual features and text vectors, but their model does not convincingly surpass an attention-based NMT baseline. Caglayan et al. (2016a) integrated local visual features extracted by ResNet-50 and source text vectors into an NMT decoder using shared transformation. They reported that the results obtained by their method did not surpass the results obtained by NMT systems. Caglayan, Barrault, and Bougares (2016b) proposed a multimodal attention mechanism based on (Caglayan et al., 2016a). They integrated two modalities by

computing the multimodal context vector, wherein the local visual features were extracted by the ResNet-50 CNN. Similarly, Calixto et al. (2016) incorporated multiple multimodal attention mechanisms into decoder using grid visual features by VGG-19 CNN. Because the grid regions do not contain semantic visual features, the multimodal attention mechanism can not capture useful information with grid visual features.

Therefore, instead of multimodal attention, Calixto, Liu, and Campbell (2017) proposed two individual attention mechanisms focusing on two modalities. Similarly, Libovický and Helcl (2017) proposed two attention strategies that can be applied to all hidden layers or context vectors of each modality. But they still used grid visual features extracted by a CNN pre-trained on ImageNet. Caglayan et al. (2017) integrated a text context vector and visual context vectors by grid visual features to generate a multimodal context vector. Their results did not surpass those of the baseline NMT for the English–German task.

Helcl, Libovický, and Variš (2018) set an additional attention sub-layer after the self-attention based on the Transformer architecture, and integrated grid visual features extracted by a pre-trained CNN. Caglayan et al. (2018) enhanced the multimodal attention into the filtered attention, which filters out grid regions irrelevant to translation and focuses on the most important part of the grid visual features. They made efforts to integrate a stronger attention function, but the considered regions were still grid visual features.

**Image region visual features.** Huang et al. (2016) extracted global visual features from entire images using a CNN and four regional bounding boxes from an image by a R-CNN.<sup>8</sup> They integrated the features into the beginning or end of the encoder hidden states. Because the global visual features were unable to provide extra supplementary information, they achieved slight improvement above the attention-based NMT. Notably, detailed regional visual features lead to better NMT translation performance.

Toyama et al. (2017) proposed a transformation to mix global visual feature vectors and object-level visual feature vectors extracted by a Fast R-CNN.<sup>9</sup> They incorporated multiple image features into the encoder and the head of the source se-

<sup>6</sup><http://www.statmt.org/wmt16/multimodal-task.html>

<sup>7</sup><http://www.statmt.org/wmt17/multimodal-task.html>

<sup>8</sup><https://github.com/rbgirshick/rcnn>

<sup>9</sup><https://github.com/rbgirshick/fast-rcnn>



quence and target sequence. Their model does not benefit from the object-level regions because the integration method cannot adequately handle visual feature sequences. Delbrouck, Dupont, and Seddati (2017) used two types of visual features, which had been extracted by ResNet-50 pretrained on ImageNet, and DenseCap<sup>10</sup> pretrained on Visual Genome, respectively. They integrated the features into their multimodal embeddings and found that the regional visual features (extracted by DenseCap) resulted in improved translations. However, they did not clarify whether the improvement in the regional visual features was brought by the multimodal embeddings or the attention model.

For the best system in WMT 2018,<sup>11</sup> Grönroos et al. (2018) used different types of visual features, such as the scene type, action type, and object type. They integrated these features into the transformer architecture using multimodal settings. However, they found that the visual features only exerted a minor effect in their system. Anderson et al. (2018) proposed a bottom-up and top-down model, which calculates attention at the level of objects. This model was used in visual question answering and image captioning tasks.

## 7 Conclusion

This paper proposed a model that integrates semantic image regions with two individual attention mechanisms. We achieved significantly improved translation performance above two baselines, and verified that this improvement mainly benefited from the semantic image regions. Additionally, we analyzed the advantages and shortcomings of our model by comparing examples and visualization of semantic image regions. In the future, we plan to use much finer visual information such as instance semantic segmentation to improve the quality of visual features. In addition, as English entity and image region alignment has been manually annotated to the Multi30k dataset, we plan to use it as supervision to improve accuracy of the attention mechanism.

## Acknowledgments

This work was supported by Microsoft Research Asia Collaborative Research Grant, Grant-in-Aid for Young Scientists #19K20343 and Grant-in-Aid for Research Activity Start-up #18H06465, JSPS.

<sup>10</sup><https://github.com/jcjohnson/densecap>

<sup>11</sup><http://www.statmt.org/wmt18/multimodal-task.html>

## References

- Anderson, Peter, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086.
- Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *ICLR*, abs/1409.0473.
- Barrault, Loïc, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *WMT*, pages 304–323.
- Caglayan, Ozan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, and Joost van de Weijer. 2016a. Does multimodality help human and machine for translation and image captioning? In *WMT*, pages 627–633.
- Caglayan, Ozan, Loïc Barrault, and Fethi Bougares. 2016b. Multimodal attention for neural machine translation. *CoRR*.
- Caglayan, Ozan, 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 *WMT*, pages 432–439.
- Caglayan, Ozan, Adrien Bardet, Fethi Bougares, Loïc Barrault, Kai Wang, Marc Masana, Luis Herranz, and Joost van de Weijer. 2018. LIUM-CVC submissions for WMT18 multimodal translation task. In *WMT*, pages 597–602.
- Calixto, Iacer and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *EMNLP*, pages 992–1003.
- Calixto, Iacer, Desmond Elliott, and Stella Frank. 2016. DCU-UvA multimodal MT system report. In *WMT*, pages 634–638.
- Calixto, Iacer, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *ACL*, pages 1913–1924.
- Calixto, Iacer, Miguel Rios, and Wilker Aziz. 2019. Latent variable model for multi-modal translation. In *ACL*, pages 6392–6405.
- Delbrouck, Jean-Benoit and Stéphane Dupont. 2017. An empirical study on the effectiveness of images in multimodal neural machine translation. In *EMNLP*, pages 910–919.
- Delbrouck, Jean-Benoit and Stéphane Dupont. 2018. UMONS submission for WMT18 multimodal translation task. In *WMT*, pages 643–647.

- Delbrouck, Jean-Benoit, Stéphane Dupont, and Omar Seddati. 2017. Visually grounded word embeddings and richer visual features for improving multimodal neural machine translation. In *GLU*, pages 62–67.
- Denkowski, Michael and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *WMT*, pages 376–380.
- Elliott, Desmond and Ákos Kádár. 2017. Imagination improves multimodal translation. In *IJCNLP*, pages 130–141.
- Elliott, Desmond, Stella Frank, and Eva Hasler. 2015. Multi-language image description with neural sequence models. *CoRR*.
- Elliott, Desmond, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual English-German image descriptions. In *VL*, pages 70–74.
- Elliott, Desmond, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *WMT*, pages 215–233.
- Fukui, Akira, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *WMT*, pages 457–468.
- Gao, Haoyuan, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question answering. In *NIPS*, pages 2296–2304.
- Grönroos, Stig-Arne, 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 *WMT*, pages 603–611.
- He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*, pages 770–778.
- Helcl, Jindřich, Jindřich Libovický, and Dušan Variš. 2018. CUNI system for the WMT18 multimodal translation task. In *WMT*, pages 616–623.
- Huang, Po-Yao, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *WMT*, pages 639–645.
- Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP*, pages 388–395.
- Krishna, Ranjay, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and F. Li. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73.
- Libovický, Jindřich and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *ACL*, pages 196–202.
- Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.
- Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *ICCV*, pages 91–99.
- Specia, Lucia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *WMT*, pages 543–553.
- Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112.
- Toyama, Joji, Masanori Misono, Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. 2017. Neural machine translation with latent semantic of image and text. *ArXiv*, abs/1611.08459.
- Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057.
- Zeiler, Matthew D. 2012. ADADELTA: an adaptive learning rate method. *CoRR*.