Master's Thesis

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

Yuting Zhao

January 21, 2020

Tokyo Metropolitan University
Graduate School of Systems Design
Department of Computer Science

A Master's Thesis
submitted to Graduate School of Systems Design,
Tokyo Metropolitan University
in partial fulfillment of the requirements for the degree of
MASTER of ENGINEERING

Yuting Zhao

Thesis Committee:

KOMACHI Mamoru	(Supervisor)
YAMAGUCHI Toru	(Co-supervisor)
TAKAMA Yasufumi	(Co-supervisor)

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

Yuting Zhao

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.

Keywords:

Multimodal neural machine translation, Semantic image regions

*Master's Thesis, Department of Computer Science, Graduate School of Systems Design, Tokyo Metropolitan University, January 21, 2020.

Contents

List of Figures

1 Introduction

Neural machine translation (NMT) [1, 2] has achieved state-of-the-art translation performance. Recently, many studies [3–5] 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 [6] demonstrated the potential of improving the translation quality by using images. To effectively use an image, several subsequent studies [7–9] 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 [10] has been designed for a mix of text and local visual features. Additionally, double attention mechanisms [11] have been proposed for text and local visual features, respectively. Although previous studies improved the use of local visual features and the text modality, these improvements were moderate. As discussed in [12], 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 [13]. However, their efforts have not convincingly demonstrated that visual features can improve the translation quality. Caglayan et al. (2019) [14] 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.

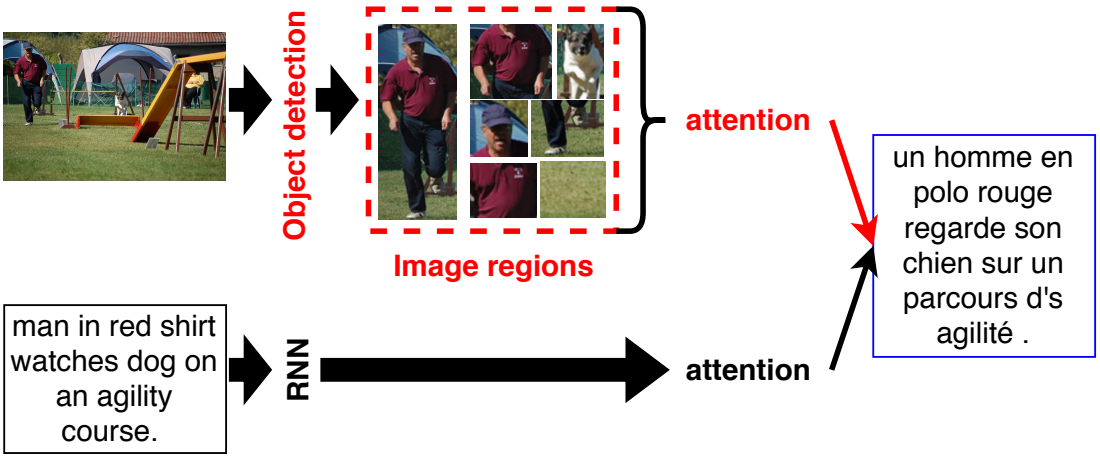


Figure 1.1: Overview of our MNMT model.

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 1.1, we use the semantic image region features extracted by an object detection model, namely, Faster R-CNN [15]. 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.

The main contributions of this study are as follows:

- We verified that the translation quality can significantly be improved 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 Related Work

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

2.1 Global visual features

Calixto and Liu [9] incorporated global visual features into source sentence vectors and encoder/decoder hidden states. As for the best system in WMT 2017,† Caglayan et al. [16] proposed different methods to incorporate global visual features based on attention-based NMT such as initial encoder/decoder hidden states using element-wise multiplication. Delbrouck and Dupont [17] proposed a variation of the conditional gated recurrent unit decoder, which receives the global visual features as input. 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.

2.2 Local visual features

2.2.1 Grid visual features

Fukui et al. [18] applied multimodal compact bilinear pooling to combine the grid visual features and text vectors, but their model does not convincingly surpass a text-only NMT baseline. Caglayan et al. [19] integrated local visual features extracted by ResNet-50 [20] and source text vectors into an NMT decoder using shared transformation. They reported that the results obtained by their method did not surpass the

*http://www.statmt.org/wmt16/multimodal-task.html
†http://www.statmt.org/wmt17/multimodal-task.html

results obtained by text-only NMT. Caglayan, Barrault, and Bougares [10] proposed a multimodal attention mechanism based on [19]. They integrated two modalities by computing the multimodal context vector, wherein the local visual features were extracted by the ResNet-50. 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 [11] proposed two individual attention mechanisms focusing on two modalities. Similarly, Libovický and Helcl [21] 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. Elliott et al. [22] considered the quality of learning visually grounded representations, but they only selected three methods to extract the grid features, and there is not much improvement between the three methods. Caglayan et al. [16] integrated a text context vector and visual context vectors extracted 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 et al. [23] set an additional attention sub-layer after the self-attention based on the Transformer architecture, and integrated grid visual features extracted by a pretrained CNN. Caglayan et al. [24] 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.

2.2.2 Image region visual features

Huang et al. [8] extracted global visual features from entire images using a CNN and four regional bounding boxes from an image by an R-CNN.[‡] 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.

Toyama et al. [25] proposed a transformation to mix global visual feature vectors and object-level visual feature vectors extracted by a Fast R-CNN.[§] They incorporated

[‡]https://github.com/rbgirshick/rcnn

[§]https://github.com/rbgirshick/fast-rcnn

multiple image features into the encoder as the head of the source sequence 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 [13] used two types of visual features, which had been extracted by ResNet-50, and DenseCap[¶]. 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,[‖] different types of visual features have been used in [26], 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. [27] proposed a bottom-up model, which calculates attention at the level of image regions. This model was used in visual question answering and image captioning tasks.

[¶]https://github.com/jcjohnson/densecap

[‖]http://www.statmt.org/wmt18/multimodal-task.html

3 Faster R-CNN

Faster region-based convolutional neural network (Faster R-CNN [15]) is a model that performs object detection as shown in Figure 3.1. The following steps are adopted in a Faster R-CNN model:

- The model processes an image from a dataset with a convolutional neural network (CNN) to generate a feature map on the last convolutional layer.
- The generated feature map is processed by a separately trained network, called the region proposal network (RPN), that outputs regions of interest (RoIs).
- The RoIs from RPN are processed by a region of interest pooling (RoI pooling) layer and several fully connected (FC) layers to output object classes and refined bounding box coordinates.

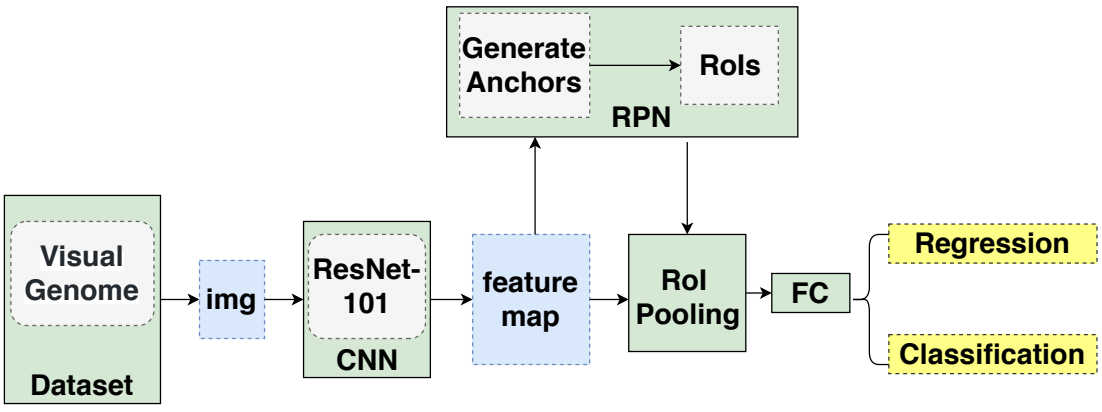


Figure 3.1: The architecture of Faster R-CNN.

3.1 Convolutional neural network (CNN)

We take an image as input and pass it to a CNN which generates a feature map for that image. The CNN is generally composed of convolutional layers, pooling layers and the last layer which is a FC layer that will be used for an specific task like classification or detection. Many pre-trained models are developed to directly use them without training, like VGG19, ResNet50 and ResNet101 [20].

3.2 Region proposal network (RPN)

The RPN is a small neural network that generates proposals for objects. It slides on the generated feature map from the last convolutional layer and finds all possible bounding boxes where objects can be located. In other words, the RPN ranks region boxes (called anchors) and proposes the ones most likely containing objects. The output of this network is a list of bounding boxes including the likely positions of objects. These are called RoIs.

3.3 Region of interest pooling (RoI pooling)

A RoI pooling layer is used to rescale all the RoIs into the same size. The FC layer always expects the same input size, but input RoIs to the FC layer may have different sizes. The function of the RoI pooling is to perform max pooling over inputs of variable sizes into a fixed length output. In this stage, the RoI pooling extracts object feature vectors which correspond to the RoIs. The vectors are used as semantic image region features with the dimension of 2048.

3.4 Prediction

At the end of the model, for every RoI, the model uses another FC layer to decide whether it belongs to one of the target classes and to refine the coordinates of bounding boxes.

4 MNMT with Semantic Image Regions

In Figure 4.1, our model comprises three parts: the source-sentence side, source-image side, and decoder. Inspired by [11], we integrated the visual features using an independent attention mechanism. From the source sentence $X = (x_1, x_2, x_3, \cdots, x_n)$ to the target sentence $Y = (y_1, y_2, y_3, \cdots, y_m)$, 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)* 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.

$$\begin{aligned}
\hat{\xi}_t &= \sigma(W_\xi E_Y[y_{t-1}] + U_\xi s_{t-1}) \\
\hat{\gamma}_t &= \sigma(W_\gamma E_Y[y_{t-1}] + U_\gamma s_{t-1}) \\
\ddot{s}_t &= \tanh\left(W E_Y[y_{t-1}] + \hat{\gamma}_t \odot (U s_{t-1})\right) \\
\hat{s}_t &= (1 - \hat{\xi}_t) \odot \ddot{s}_t + \hat{\xi}_t \odot s_{t-1}
\end{aligned} \tag{4.1}$$

where W_ξ, U_ξ, W_γ, U_γ, W, and U are training parameters; E_Y is the target word vector.

4.1 Source-sentence side

The source sentence side comprises a bi-directional GRU encoder and “soft" attention mechanism [28]. Given a source sentence $X = (x_1, x_2, x_3, \cdots, x_n)$, the encoder updates

*https://github.com/nyu-dl/dl4mt-tutorial/blob/master/docs/cgru.pdf

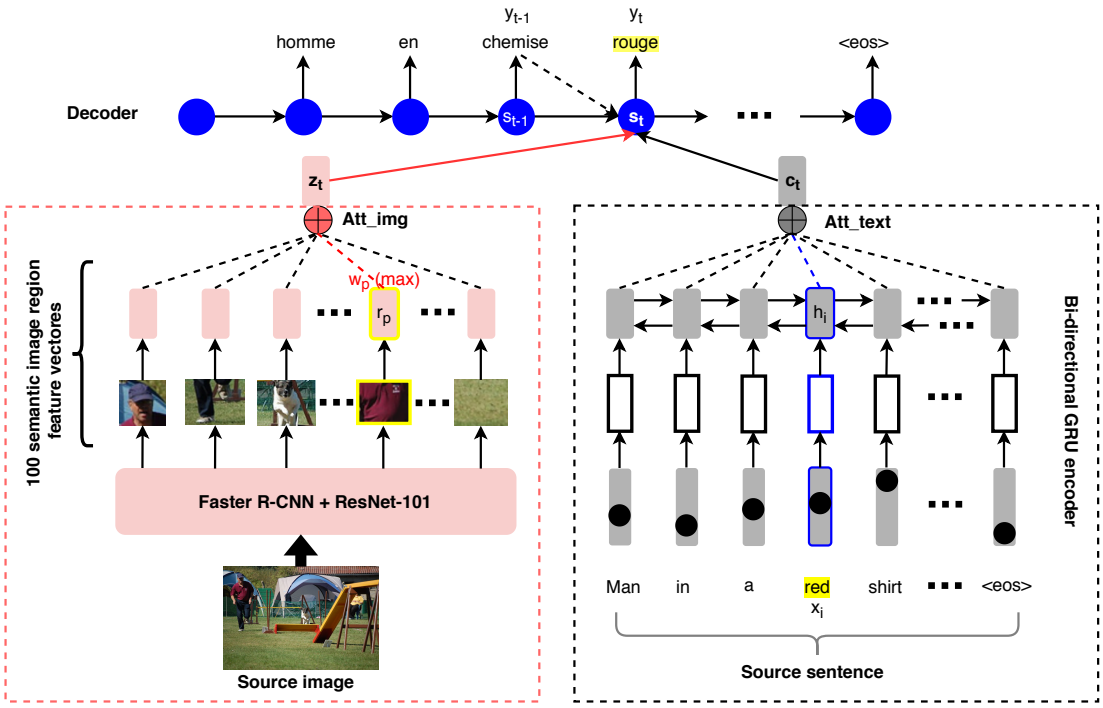
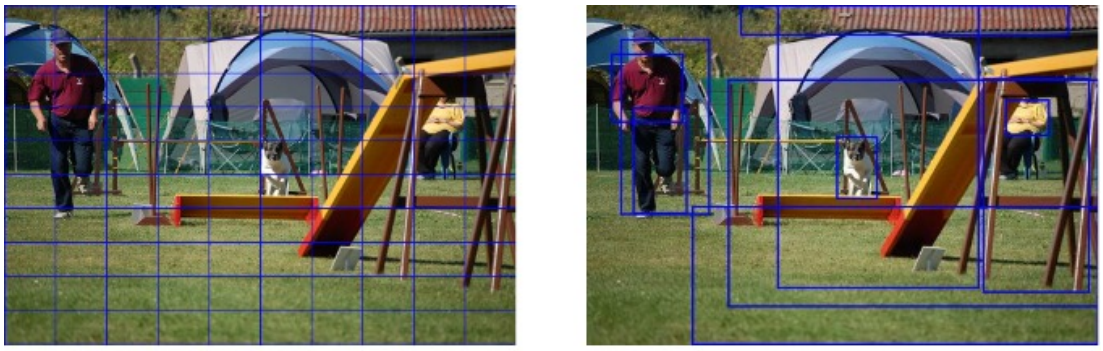


Figure 4.1: Our model of double attention-based MNMT with semantic image regions.

the forward GRU hidden states by reading x from left to right, generates the forward annotation vectors $(\overrightarrow{h_1}, \overrightarrow{h_2}, \overrightarrow{h_3}, \cdots, \overrightarrow{h_n})$, and finally updates the backward GRU with the annotation vectors $(\overleftarrow{h_1}, \overleftarrow{h_2}, \overleftarrow{h_3}, \cdots, \overleftarrow{h_n})$. By concatenating the forward and backward vectors $h_i = [\overrightarrow{h_i}; \overleftarrow{h_i}]$, every h_i encodes the entire sentence while focusing on the x_i word, and all words in a sentence are denoted as $C = (h_1, h_2, \cdots, h_n)$. At each time step t, the text context vector c_t is generated as follows:

$$\begin{aligned} e_{t,i}^{\text{text}} &= (V^{\text{text}})^{\text{T}} \tanh(U^{\text{text}} \hat{s}_t + W^{\text{text}} h_i) \\ \alpha_{t,i}^{\text{text}} &= \text{softmax}(e_{t,i}^{\text{text}}) \\ c_t &= \sum_{i=1}^{n} \alpha_{t,i}^{\text{text}} h_i \end{aligned} \tag{4.2}$$

where V^{text}, U^{text}, and W^{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.



(a) coarse grids. (b) semantic image regions.

Figure 4.2: Comparing between (a) coarse grids and (b) semantic image regions.

4.2 Source-image side

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

4.2.1 Semantic image region feature extraction

As shown in Figure 4.2, 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 CNN pre-trained on Visual Genome [29] 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 $R = (r_1, r_2, r_3, \cdots, r_{100})$.

4.2.2 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 $e^{\text{img}}_{t,p}$, 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:

$$e_{t,p}^{\text{img}} = (V^{\text{img}})^{\text{T}} \tanh(U^{\text{img}} \hat{s}_t + W^{\text{img}} r_p) \tag{4.3}$$

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

$$\alpha_{t,p}^{\text{img}} = \text{softmax}(e_{t,p}^{\text{img}}) \tag{4.4}$$

$$\tag{4.5}$$

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:

$$z_t = \beta_t \sum_{p=1}^{100} \alpha_{t,p}^{\text{img}} r_p \tag{4.6}$$

For z_t, at each decoding time step t, a gating scalar $\beta_t \in [0, 1]$ [28] is used to adjust the proportion of the image context vector according to the previous hidden state of the decoder s_{t-1}.

$$\beta_t = \sigma(W_\beta s_{t-1} + b_\beta) \tag{4.7}$$

where W_β and b_β are training parameters.

4.3 Decoder

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

$$\begin{aligned}
\xi_t &= \sigma(W_\xi^{\text{text}} c_t + W_\xi^{\text{img}} z_t + \bar{U}_\xi \hat{s}_t) \\
\gamma_t &= \sigma(W_\gamma^{\text{text}} c_t + W_\gamma^{\text{img}} z_t + \bar{U}_\gamma \hat{s}_t) \\
\bar{s_t} &= \tanh\left(W^{\text{text}} c_t + W^{\text{img}} z_t + \gamma_t \odot (\bar{U} \hat{s}_t)\right) \\
s_t &= (1 - \xi_t) \odot \bar{s_t} + \xi_t \odot \hat{s}_t
\end{aligned} \tag{4.8}$$

where W_ξ^{text}, W_ξ^{img}, $\bar{U}_\xi$, W_γ^{text}, W_γ^{img}, $\bar{U}_\gamma$, W^{text}, W^{img}, and $\bar{U}$ are model parameters; ξ_t and γ_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:

$$\mathrm{softmax}(L_o \tanh(L_s s_t + L_c c_t + L_z z_t + L_w E_Y[y_{t-1}])) \tag{4.9}$$

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

5 Experiments

5.1 Dataset

We conducted experiments for the English→German (En→De) and English→French (En→Fr) tasks using the Multi30k dataset [30]. 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.* Subsequently, we converted the space-separated tokens into subword units using the byte pair encoding (BPE) model.†

5.2 Settings

5.2.1 Ours

We integrated the semantic image regions by modifying the double attention model of [11]. 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

*https://github.com/moses-smt/mosesdecoder

†https://github.com/rsennrich/subword-nmt

descent with ADADELTA [31] 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.

5.2.2 Baseline doubly-attentive MNM

We trained a doubly-attentive MNMT model[‡] as a baseline. For the text side, the implementation was based on OpenNMT model.[§] For the image side, attention was applied to the visual features extracted from 14×14 image grids by CNNs. For the image feature extraction, we compared three pre-trained CNN methods: VGG-19, ResNet-50, and ResNet-101.

5.2.3 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.

We used the original implementations and ensured the parameters were consistent with our model.

5.3 Evaluation

We evaluated the quality of the translation according to the BLEU [32] and METEOR [33] metrics. We trained all models (baselines and proposed model) three times and calculated the BLEU and METEOR scores. Based on the calculation results, we report the mean and standard deviation over three runs.

Moreover, we report the statistical significance with bootstrap resampling [34] 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.

[‡]https://github.com/iacercalixto/MultimodalNMT
[§]https://github.com/OpenNMT/OpenNMT-py

6 Results

In Table 6.1, we present the results for the OpenNMT, doubly-attentive MNMT [11], our model and Caglayan et al. [16] on Multi30k dataset.

Comparing the baselines, the doubly-attentive MNMT outperformed OpenNMT by 1.8 BLEU points and 1.7 METEOR points for En→De, and by 0.7 BLEU points and 0.3 METEOR points for En→Fr. 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 proposed model improved the BLEU scores by 2.3 points and METEOR scores by 2.1 points for En→De. Additionally, it improved the BLEU scores by 1.6 points and the METEOR scores by 1.1 points for En→Fr. The results obtained by this study are significantly better than the results obtained by the baseline for both tasks with a p-value < 0.01.

Compared with the doubly-attentive MNMT (ResNet-101) baseline, the proposed model improved the BLEU scores by 0.5 points and the METEOR scores by 0.4 points for En→De. Additionally, it improved the BLEU scores by 0.9 points and the METEOR scores by 0.8 points for En→Fr. Moreover, the results are significantly better than the baseline results with a p-value < 0.01.

For comparison with [16], we report their results for the text-only NMT baseline, grid-based MNMT method and global-based MNMT method. With the grid-based method, their results failed to surpass the text-only NMT baseline for En→De with regard to both metrics, and surpassed the text-only NMT baseline by 1.0 BLEU points and 0.8 METEOR points for En→Fr. With the global-based method, their results surpassed the text-only NMT baseline by 0.7 BLEU points and 0.2 METEOR points for En→De, and by 2 BLEU points and 1.6 METEOR points for En→Fr.

For En→De, their global-based method achieved higher scores than our model. In terms of relative improvement compared with the text-only NMT baseline, their re-

sults 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 surpassed their results with regard to both methods. In terms of 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-based method and improved the BLEU score by 3.8% and METEOR score by 2.3% with the global-based method. Our model improved the BLEU score by 2.8% and METEOR score by 1.5%.

Hence, when using local visual features, our model achieves the best improvement compared with previous methods. However, the improvement achieved by our model does not surpass the improvement achieved by their global-based method.

	En→De		En→Fr	
Model	BLEU	METEOR	BLEU	METEOR
OpenNMT (text-only)	34.7±0.3	53.2±0.4	56.6±0.1	72.1±0.1
Calixto et al. [11] (VGG-19)	36.4±0.2	55.0±0.1	57.4±0.4	72.4±0.4
Calixto et al. [11] (ResNet-50)	36.5±0.2	54.9±0.4	57.5±0.4	72.6±0.4
Calixto et al. [11] (ResNet-101)	36.5±0.3	54.9±0.3	57.3±0.2	72.4±0.2
Ours	**37.0±0.1**†	**55.3±0.2**	**58.2±0.5**†‡	**73.2±0.2**
vs. OpenNMT	(↑ 2.3)	(↑ 2.1)	(↑ 1.6)	(↑ 1.1)
vs. Calixto et al. [11]	(↑ 0.5)	(↑ 0.4)	(↑ 0.8)	(↑ 0.9)
Caglayan et al. [16] (text-only)	38.1±0.8	57.3±0.5	52.5±0.3	69.6±0.1
Caglayan et al. [16] (grid)	37.0±0.8	57.0±0.3	53.5±0.8	70.4±0.6
Caglayan et al. [16] (global)	38.8±0.5	57.5±0.2	54.5±0.8	71.2±0.4

Table 6.1: 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 [11] (ResNet-101) at p-value < 0.01, respectively. Additionally, we report the best results of using grid and global visual features on Multi30k dataset according to [16], which is the state-of-the-art system for En→De translation on this dataset.

7 Analysis

7.1 Pairwise evaluation of translations

We 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 7.1, 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.

7.2 Qualitative analysis

In Figure 7.1, 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

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

Table 7.1: The amount and proportion of each type of examples in all investigated examples.

along with the bounding box. Then, we analyzed whether the semantic image regions had a positive or negative effect at the time step when the target word was generated.

7.2.1 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 attented semantic image region at that time step provided the exact object-level visual feature to the translation.

Figure 7.1: 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.

7.2.2 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 "joyeusement," 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 function 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.

7.2.3 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, our model also has 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.

8 Conclusion

The thesis 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.

Acknowledgements

I would like to express my deep gratitude to Associate Professor Mamoru Komachi for providing me with a comfortable environment for studying as well as giving me guidance in conducting research activities. Through my research life, I owe it all to my supervisor that I was able to have a great deal of valuable experiences such as conference presentations and collaborative research. I also sincerely thank Dr. Chenhui Chu and Dr. Kajiwara Tomoyuki, in their dedication to help and careful guidance, I can successfully complete this study. In addition, thanks to senior Longtu Zhang for kindly guiding and helping me start my first research. And thank you to all the students in the laboratory who spent my research life together. I would like to thank Prof. Yamaguchi Toru and Prof. Takama Yasufumi for the advice of the co-supervisor.

References

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112, 2014.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, abs/1409.0473, 2015.

[3] Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. A shared task on multimodal machine translation and crosslingual image description. In *WMT*, pages 543–553, 2016.

[4] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. In *WMT*, pages 215–233, 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 *WMT*, pages 304–323, 2018.

[6] Desmond Elliott, Stella Frank, and Eva Hasler. Multi-language image description with neural sequence models. *CoRR*, 2015.

[7] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. In *NIPS*, pages 2296–2304, 2015.

[8] Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. Attention-based multimodal neural machine translation. In *WMT*, pages 639–645, 2016.

[9] Iacer Calixto and Qun Liu. Incorporating global visual features into attention-based neural machine translation. In *EMNLP*, pages 992–1003, 2017.

[10] Ozan Caglayan, Loïc Barrault, and Fethi Bougares. Multimodal attention for neural machine translation. *CoRR*, 2016.

[11] Iacer Calixto, Qun Liu, and Nick Campbell. Doubly-attentive decoder for multi-modal neural machine translation. In *ACL*, pages 1913–1924, 2017.

[12] Jean-Benoit Delbrouck and Stéphane Dupont. An empirical study on the effectiveness of images in multimodal neural machine translation. In *EMNLP*, pages 910–919, 2017.

[13] Jean-Benoit Delbrouck, Stéphane Dupont, and Omar Seddati. Visually grounded word embeddings and richer visual features for improving multimodal neural machine translation. In *GLU*, pages 62–67, 2017.

[14] Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. Probing the need for visual context in multimodal machine translation. In *NAACL*, pages 4159–4170, 2019.

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *ICCV*, pages 91–99, 2015.

[16] Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. LIUM-CVC submissions for WMT17 multimodal translation task. In *WMT*, pages 432–439, 2017.

[17] Jean-Benoit Delbrouck and Stéphane Dupont. UMONS submission for WMT18 multimodal translation task. In *WMT*, pages 643–647, 2018.

[18] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *WMT*, pages 457–468, 2016.

[19] Ozan Caglayan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, and Joost van de Weijer. Does multimodality help human and machine for translation and image captioning? In *WMT*, pages 627–633, 2016.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[21] Jindřich Libovický and Jindřich Helcl. Attention strategies for multi-source sequence-to-sequence learning. In *ACL*, pages 196–202, 2017.

[22] Desmond Elliott and Ákos Kádár. Imagination improves multimodal translation. In *IJCNLP*, pages 130–141, 2017.

[23] Jindřich Helcl, Jindřich Libovický, and Dušan Variš. CUNI system for the WMT18 multimodal translation task. In *WMT*, pages 616–623, 2018.

[24] Ozan Caglayan, Adrien Bardet, Fethi Bougares, Loïc Barrault, Kai Wang, Marc Masana, Luis Herranz, and Joost van de Weijer. LIUM-CVC submissions for WMT18 multimodal translation task. In *WMT*, pages 597–602, 2018.

[25] Joji Toyama, Masanori Misono, Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Neural machine translation with latent semantic of image and text. *ArXiv*, abs/1611.08459, 2017.

[26] 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 *WMT*, pages 603–611, 2018.

[27] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018.

[28] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.

[29] Ranjay Krishna, 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. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017.

[30] 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, 2016.

[31] Matthew D. Zeiler. ADADELTA: an adaptive learning rate method. *CoRR*, 2012.

[32] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002.

[33] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *WMT*, pages 376–380, 2014.

[34] Philipp Koehn. Statistical significance tests for machine translation evaluation. In *EMNLP*, pages 388–395, 2004.

Publication List

International Conferences

[1] Longtu Zhang and Yuting Zhao and Mamoru Komachi. TMU Japanese-Chinese Unsupervised NMT System for WAT 2018 Translation Task. In The 5th Workshop on Asian Translation, HongKong, China. December 3, 2018.

Domestic Conferences

[1] Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, Chenhui Chu. Double Attention-based Multimodal Neural Machine Translation with Semantic Image Regions. In The 14th Symposium of Young Researcher Association for NLP Studies, P02, Hokkaido, Japan. August 26, 2019.

[2] Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, Chenhui Chu. Double Attention-based Multimodal Neural Machine Translation with Semantic Image Region. In The 241st Meeting of Special Interest Group of IPSJ Natural Language Processing, Hokkaido, Japan. August 30, 2019.

[3] Yuting Zhao, Longtu Zhang and Mamoru Komachi. Application of Unsupervised NMT Technique to Japanese-Chinese Machine Translation. In The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, Niigata, Japan. June 6, 2019.