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

Yuting Zhao[1,a)] Mamoru Komachi[1,b)] Tomoyuki Kajiwara[2,c)] Chenhui Chu[2,d)]

Abstract: Current work on multimodal neural machine translation (MNMT) has mostly paid attention to the effect of combining visual and textual modalities in improving translation performance. However, it has been suggested that the visual modality is only marginally beneficial. As conventional visual attention mechanisms are used to select visual features from grids of equal size in an image generated by convolutional neural net, the feature of a grid that is not related to image content may arise limited effects in aligning visual concepts associated with the textual object. In contrast, we propose to apply semantic image regions for MNMT with integrating visual and textual features by means of two separate attention mechanisms (double attention) in order to improve predictive token generation. Our approach on the Multi30k dataset achieves 0.5 and 0.9 BLEU point improvement on English–German and English–French translation tasks compared with the baseline double attention-based MNMT.

Keywords: Multimodal neural machine translation, semantic image regions, double attention, Multi30k

1. Introduction

In recent years, significant amount of research has paid more attention to the joint modeling of natural language processing (NLP) and computer vision. There appeared a lot of shared task such as image caption generation [1–5], visual question answering [6–8] as well as Multimodal Machine Translation (MMT) [9–12] that takes image into account as the auxiliary inputs. In the first MMT published in [13], a full image is used to support a translation task. It became popular to heuristically integrating two modalities such as text and image into machine translation. Consequently, more research has been conducted to show that visual attributes would assist the neural network to generate better and more accurate translations. In [12], they put a feature vector extracted from a whole image into an encoder hidden state followed by text sequence, and use attention focusing on each part in the encoding phase. In addition, Calixto and Liu [9] propose to represnet a whole image as a word in the head or tail of source sentence or to use the whole image to initialize the encoder or decoder hidden state as an input to improve attention-based Neural Machine Translation (NMT) performance. Furthermore, the idea of combining image preprocessing with attention mechanisms is also emerging. As in [11], image is first preprocessed into grid by CNN, and then visual features and text annotation are used together to compute multimodal context vector with a multimodal attention mechanism. Subsequently, an improved model is proposed in [10], in which they publish the concept of two separated attention mechanisms. Similarly, they also preprocessed the image into feature vectors in grid units. Different from their previous work, their two attention mechanisms work on image features and text annotation, respectively. Their research has made steady improvement by using image features in machine translation, while the full potential of the image feature contributing to machine translation still needs to be more deeply explored.

In this paper, we combine attention mechanism and object detection to take full advantage of image features in NMT. Firstly, we refine the image feature extraction compared with previous work. We take advantage of the bottom-up attention model [14] which is updated from Faster R-CNN [15] as an object detection model. Thanks to this model, the instances of objects can be identified and be localized with bounding boxes. We use it to enhance the image preprocessing; instead of obtaining a grid feature, we acquire the object-level semantic image region feature. Secondly, in order to make full use of the extracted semantic image region, we utilize an additional image-attention mechanism to the decoder to assist generating the target word. In this way, the prediction can be generated from not only source text but also from the semantic image region.

Our main contributions are:

- We propose semantic image regions, which are attended from one attentional mechanism in double attention-based NMT to maximize the use of visual feature, so as to make an improvement of adequacy on translation effect.
- We demonstrate the impact of introducing the semantic image region to the attention-based NMT and double attention-based Multimodal NMT (MNMT).
- We discuss the effect of the semantic image region acting on the actual translation examples.

[1] Tokyo Metropolitan University, 6-6 Asahigaoka, Hino, Tokyo 191-0065, Japan
[2] Osaka University, 2-8 Yamadaoka, Suita, Osaka 565-0871, Japan
[a)] zhao-yuting@ed.tmu.ac.jp
[b)] komachi@tmu.ac.jp
[c)] kajiwara@ids.osaka-u.ac.jp
[d)] chu@ids.osaka-u.ac.jp

2. Related Work

2.1 Attention mechanism

The concept of attention was first put forward in [16]. Their motivation comes from the mechanism of human attention. When people look at an image, they do not actually look at every pixel of the whole image at one time but focus their attention on specific parts according to their needs. Based on this idea, they propose a framework for attention-based task-driven visual processing with neural networks.

In the field of machine translation, after the emergence of traditional encoder decoder NMT, the first attention-based mechanism comes from [17]. Their motivation is that learning to align and translate the source word and the current predicted word jointly according to calculated attention weight via attention mechanism.

On the application of image captioning, Xu et al. [2] introduce the attention mechanism for the first time. The attention mechanism is extended into soft attention and hard attention in this paper. For soft attention, it is the initial attention method similar to [17], but it applied the attention over the convolutional image features extracted from VGG. For hard attention, it does not generate weights for all features but only focused on one position at one time.

2.2 Multimodal NMT

In [11], they improve Multimodal NMT (MNMT) with a multimodal attention mechanism. It is designed to integrate two modality-specific context vectors from the textual annotation vector and visual annotation vectors by the multimodal attention mechanism. They extract visual feature from ResNet-50 CNN, and compute a multimodal context vector by a non-linearity function acting on two modalities. At last, they calculate the current decoder hidden state considering the multimodal context vector and previous hidden state.

On the other hand, in [10], they improve MNMT with two separated attention mechanisms. They compute two modality-specific context vectors independently by two separated attention mechanisms. One is used for source text and the other one is used for visual vectors extracted by ResNet-50 CNN. Then, they generate the current decoder hidden state considering both modality-specific context vectors.

Furthermore, in [18], they propose to re-implement transformer NMT architecture with a multimodal settings. Then they extract image features by a couple of different ways, including CNN-based, scene-type, action-type and boject-type. They make an improvement from incorporating multimodal information, while showing that the effect of the visual features in their system is small.

2.3 Image region

In [12], they propose to integrate multimodal information (text and regional features) into the attention-based NMT. For extracting regional features, they use the region-based convolutional neural networks (R-CNN) [19]. They take only top 4 regional objects arranging in ascending order of size, and then combine them together with global image feature to obtain a new visual sequence. They put the new visual sequence followed by the text sequence. Avoiding dimension mismatch and the inherent difference in content between the visual and textual vectors, they use a transformation matrix to learn the mapping. Then, they calculate the attention weights of all the possible hidden states in the final sequence. Considering the length of the encoding phrase (visual sequence plus text sequence), they just make the visual sequence from 4 regional features.

In [14], they propose a bottom-up and top-down attention model for image captioning and visual question answering. The task of the bottom-up attention model is to acquire the image region of interest to extract image features, and the top-down attention model is used to learn to adjust the feature weights, realizing the time attention of the image content, and generating the description word by word. It is worth noting that the bottom-up attention model enables attention to be calculated at the level of objects and other salient image regions.

3. Methodology

As illustrated in Figure 1, the overall double attention-based MNMT model consists of three parts: source-text side, source-image side and decoder side. We address the problem of learning to generate target words by means of concentrating on semantic image regions instead of the coarse grids in terms of image-attention mechanism. From a source sequence $\chi = (\chi_1, \chi_2, \chi_3, \cdots, \chi_n)$ to its target sequence $\tau = (\tau_1, \tau_2, \tau_3, \cdots, \tau_m)$, the image-attention mechanism concentrates on all the semantic image region features so as to calculate image context vector z_t, which is passed into the decoder layer together with the text context vector c_t to assist prediction. We consider that the more clearly the image region can express the semantic features of the source word, the more valuable information can be provided for the translation process from the source word to the target word, and the more accurate the word generation can be.

3.1 Source-text side

This is a common part both in the attentive NMT and MNMT. There are a bi-directional RNN encoder and a soft attention mechanism. Given a source sequence $\chi = (\chi_1, \chi_2, \chi_3, \cdots, \chi_n)$, the model updates forward RNN hidden states via reading χ from left to right, and generates forward annotation vectors $(\overrightarrow{\upsilon_1}, \overrightarrow{\upsilon_2}, \overrightarrow{\upsilon_3}, \cdots, \overrightarrow{\upsilon_n})$, as well as reversely updating backward RNN with annotation vectors $(\overleftarrow{\upsilon_1}, \overleftarrow{\upsilon_2}, \overleftarrow{\upsilon_3}, \cdots, \overleftarrow{\upsilon_n})$. Concatenating forward and backward vectors $\upsilon_i = [\overrightarrow{\upsilon_i}\,;\,\overleftarrow{\upsilon_i}]^{\mathrm{T}}$, every υ_i encodes the whole sentence with a focus on the i-th word. At each time step t, text context vector c_t is computed based on annotation vectors $V = (\upsilon_1, \upsilon_2, \ldots, \upsilon_n)$, and the decoder previous hidden state s_{t-1}:

$$e_{t,i}^{\text{text}} = A(s_{t-1}, V_i) \tag{1}$$

$$\alpha_{t,i} = \frac{\exp(e_{t,i}^{\text{text}})}{\sum_{k=1}^{n} \exp(e_{t,k}^{\text{text}})} \tag{2}$$

$$c_t = \sum_{i=1}^{n} \alpha_{t,i} * V_i \tag{3}$$

where A is presented an attention model.

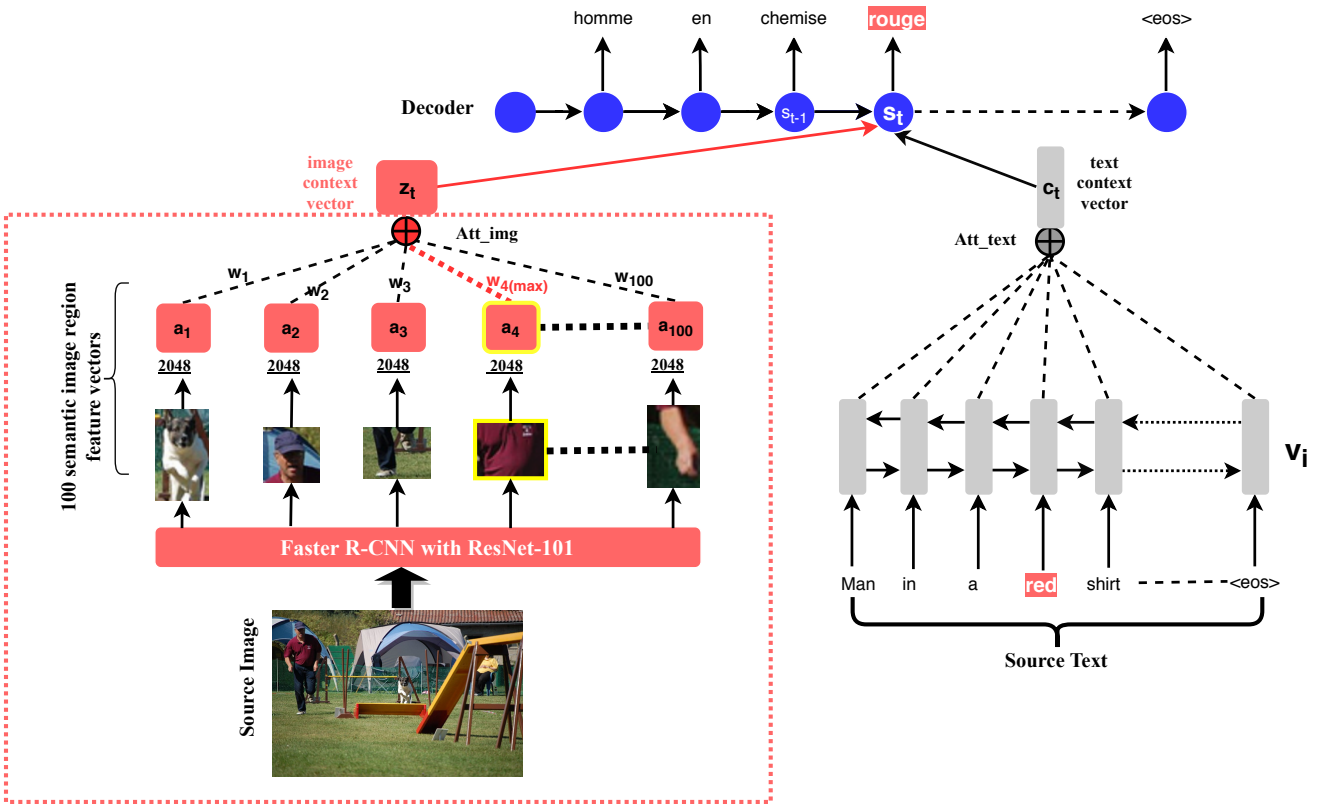
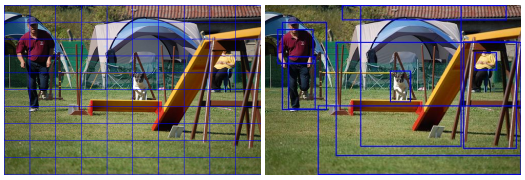


Fig. 1: The model of double attention-based MNMT with semantic image regions.



(a) Grid. (b) Image regions.

Fig. 2: Comparing between the coarse grid (a) and semantic image regions (b).

3.2 Source-image side

In this part, through the image-attention mechanism, the semantic image regions are integrated into the MNMT model.

3.2.1 Semantic image region extraction

As illustrated in Figure 2, instead of extracting grid feature in a uniform size, we consider extracting a feature which is related to an object-level meaning by means of a bounding box as shown in the picture. Therefore, we put forward the concept of integrating semantic image region, extracted 100 bounding box with object-level information from the image to make them to be paid attention to by the image-attention mechanism.

In this work, we employed the Faster R-CNN [15] in conjunction with the ResNet-101 [20] CNN and trained by Visual Genome [21] data. We use it to extract 100 bounding box features from each image in the Multi30k dataset. Each feature is generated by a vector with size of 2048×1, and marked as $\Lambda = (a_1, a_2, a_3, \cdots, a_{100})$.

3.2.2 Image-attention mechanism

This part is an extension of the text-attention mechanism, which concentrates on 100 semantic image region feature vectors at each time step. The attention distribution of input image regions corresponding to each token generated by the target sentence can be understood as the probability of alignment between the input image region and the target token.

We apply the "soft" attention [2] considering image region feature a_i in $\Lambda = (a_1, a_2, a_3, \cdots, a_{100})$, the previously generated target word τ_{t-1} and previous hidden state s_{t-1}. A hidden state proposal $\tilde{s}_t$ is computed as follows:

$$\tilde{s}_t = f(s_{t-1}, \tau_{t-1}) \tag{4}$$

We then project it to get the attention energy $e_{t,i}$, which is an attention model that scores the degree of output matching between the inputs around position i and the output at position t:

$$e_{t,i} = V_a^{\mathrm{T}} * \tanh\left(U_a \tilde{s}_t + W_a a_i\right) \tag{5}$$

where V_a, U_a and W_a are model parameters.

Then the weight matrix $w_{t,i}$ of each a_i is computed:

$$\begin{aligned} w_{t,i} &= \mathrm{softmax}(e_{t,i}) \\ &= \frac{\exp(e_{t,i})}{\sum_{k=1}^{100} \exp(e_{t,k})} \end{aligned} \tag{6}$$

At each time step, the image-attention mechanism dynamically selects the image region feature that is most relevant for predicting the current target word. The result is an image context vector z_t:

$$z_t = \beta_t * \sum_{i=1}^{100} w_{t,i} * a_i \tag{7}$$

Among z_t, the $\beta_t \in [0, 1]$ following [2] is used to adjust the proportion of the image context vector in GRU (relative to the proportion of s_{t-1}, τ_{t-1}):

$$\beta_t = \sigma(f_\beta(s_{t-1})) \tag{8}$$

3.3 Decoder side

The decoder is a conditional GRU. Its initial hidden state results from a projection of the encoder final vector. The new hidden state s_t employing GRU gated hidden units is computed by the hidden state proposal $\tilde{s}_t$, the time-dependent text context vector c_t and image context vector z_t:

$$s_t = (1 - \xi_t) \odot \bar{s}_t + \xi_t \odot \tilde{s}_t \tag{9}$$

in which $\bar{s}_t$ is the proposed updated hidden state:

$$\bar{s}_t = \tanh\left(C^{\text{text}} * c_t + C^{\text{img}} * z_t + \gamma_t \odot (U * \tilde{s}_t)\right) \tag{10}$$

where $\odot$ is an element-wise multiplication, C^{text}, C^{img} and U are model parameters, and ξ_t and γ_t is the output of update / reset gates:

$$\xi_t = \sigma(C_\xi^{\text{text}} c_t + C_\xi^{\text{img}} z_t + U_\xi \tilde{s}_t) \tag{11}$$

$$\gamma_t = \sigma(C_\gamma^{\text{text}} c_t + C_\gamma^{\text{img}} z_t + U_\gamma \tilde{s}_t) \tag{12}$$

Finally, each conditional probability of generating target token τ_t is computed by g which is a nonlinear, potentially multi-layered, function:

$$p(\tau_t|\tau_1, \ldots, \tau_{t-1}, V, \Lambda) = g(s_t, \tau_{t-1}, c_t, z_t) \tag{13}$$

4. Experiments

4.1 Data set

In this work , we conduct experiments on the English→German (EN→DE) and English→French (EN→FR) data from Multi30k dataset [22], which is an extension of the Flickr30K dataset [23]. It is a dataset that contains 30k training, 1014 validation and 1k test images respectively, and each image paired with image descriptions contains both original English and translated sentences in multiple languages.

For preprocessing, we lowercase and tokenize English, German and French descriptions with the scripts in Moses SMT Toolkit [24]. We convert space-separated tokens into subword units via byte pair encoding (BPE) model [25]. We limit the number of tokens in a description to a maximum of 80. We train model to translate from English to German and English to French tasks. We use entire training set for training , its validation set for model selection and its 2016's test set to report evaluation of cased, tokenized sentences with punctuation.

4.2 Baseline Approaches for comparisons

4.2.1 Attentive NMT

We train a text-only attentive NMT model using OpenNMT [26] as our baseline. We train it on EN→DE and EN→FR in which only the textual part of Multi30k is used for training. It is consisted of a 2-layer bidirectional GRU encoder and a 2-layer conditional GRU decoder. The function of attention mechanism is to allow the decoder to attend to different parts of the source sentence at each time step of the output prediction.

4.2.2 Doubly-attentive MNMT

In addition, we train a doubly-attentive MNMT model [10] as another baseline, in which text and image have been integrated into the model by means of two attention mechanisms, respectively. On the source-text side, it has the same attention mechanism as the one used in the OpenNMT model. On the source-image side, however, attention mechanism acts on image features which are encoded by pre-trained convolutional neural net (CNN). For image feature extraction, we encode images with three pre-trained CNN methods: VGG-19, ResNet-50, and ResNet-101, respectively. We use the authors' implementation on the GitHub [*1] to keep the settings consistent.

4.3 Implementation Details

4.3.1 Feature

We extract image regions from bottom-up attention model [14], which is based on the training of Faster R-CNN [15] with ResNet-101, using Visual Genome [21] data. We obtain output features corresponding to salient image regions, which contain bounding box feature vectors and horizontal and vertical coordinates of each bounding box on an image. We set 100 bounding box on a image and a bounding box feature vector with a size of 2048×1.

4.3.2 Learning

We change doubly-attentive MNMT model to make it applicable to act on our image regions. On the source-text side, we keep the settings consistent. On the image-source side, we adjust image attention mechanism to attend to image regions of size 1×100×2048 at each time step. Then, 2-layer conditional GRU decoder is conditioned on the previous hidden state from the first layer of decoder, the previously emitted token, the source context vector and image region context vector. The settings of other parts and parameters are consistent with the basic doubly-attentive MNMT model, in which setting we use batch size of 40, text dropout with a probability of 0.3 and image region dropout with 0.5, as well as we train the model using stochastic gradient descent with ADADELTA [27] with a learning rate of 0.002. We select the model by analyzing the learning curve via calculating validation perplexity and validation accuracy. We stop learning when the values of both metrics are stable; in this work, it is set to run till 25 iterations.

4.3.3 Evaluation

We evaluate translation quality in terms of BLEU [28] and METEOR [29]. We train all the models (baseline and our proposal) three times and calculate BLEU score and METEOR score, based on which we report the mean and standard deviation over three runs. Furthermore, we report statistical significance with bootstrap resampling by [30] using the merger of three test translation results. We define the threshold for statistical significance test as 0.05 and report if the p-value is less than the threshold or not.

[*1] https://github.com/iacercalixto/MultimodalNMT

	EN→DE		EN→FR	
Model	BLEU	METEOR	BLEU	METEOR
Attentive NMT + (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
Doubly-attentive MNMT + (Semantic Image Regions)	**37.0±0.1** (↑ 0.5)	**55.3±0.2** (↑ 0.4)	**58.2±0.5** (↑ 0.8) **(p-value < 0.05)**	**73.2±0.2** (↑ 0.9)

Table 1: BLEU and METEOR scores for different models on the 2016's EN→DE and EN→FR test set of Multi30k. All scores are averages of three runs. We present the results using the mean and the standard deviation.

Source (En)	a man in a blue coat grabbing a young boy's shoulder .
Reference (Fr)	un homme en manteau bleu saisissant l'paule d'un jeune garon .
Attentive NMT (Fr)	un homme en manteau bleu se baladant avec l'paule d'un jeune garon .
Attentive MNMT (Fr)	un homme en manteau bleu agrippe l'paule d'un jeune garon .
Our proposal (Fr)	un homme en manteau bleu agrippant l'paule d'un jeune garon .

Table 2: Comparison of translation performance generated by our proposal and baselines.

5. Results

In Table 1, we show results for attentive NMT and MNMT baselines as well as our proposal on translating from EN→DE and EN→FR. We note that our proposal performs consistently better than the strong attentive NMT and MNMT baselines in both directions, moreover in both evaluation methods.

Comparing with attentive NMT baseline, our proposal improves 2.3 BLEU scores and 2.1 METEOR scores in EN→DE, as well as an improvement of 1.6 BLEU scores and 1.1 METEOR scores in EN→FR.

Comparing with attentive MNMT with ResNet-101 baseline, our proposal improves 0.5 BLEU scores and 0.4 METEOR scores in EN→DE. On the other hand, there is an improvement of 0.9 BLEU scores and 0.8 METEOR scores in EN→FR, moreover, results are significantly better than the baseline with p-value < 0.05.

6. Qualitative Analysis

In order to intuitively evaluate the influence of semantic image region acting on doubly-attentive MNMT, we choose examples from EN→FR task randomly and conduct qualitative analysis from three aspects:

- **Comparative observation of translations.** As shown in Table 2, we compare the translation results generated by attentive NMT, doubly-attentive MNMT with ResNet-101 and doubly-attentive MNMT with semantic image regions. The names of these three models are abbreviated as "Attentive NMT," "Attentive MNMT" and "Our proposal" in the following tables.
- **Visualization of the semantic image region and target word attention.** As shown in Figure 4, at each time step, the semantic image region is shown in deep or shallow transparency on the image according to the attention weight assigned to it. The larger the weight, the more clearly it appears on the image. Considering a large number of 100 bounding boxes and overlapping areas, we visualize the five most weighted bounding boxes. The most weighted image region are marked with blue lines, and the target token generated at that time step are marked with red text along with the bounding box.
- **Visualization of the source and target word attention.** Figure 3 is the visualization of the attention weights from the original English tokens to the translation tokens (French), respectively. Each pixel shows the weight of the matching relation of source token and corresponded target token. The larger the weight, the brighter the corresponding pixel (0: black, 1: white).

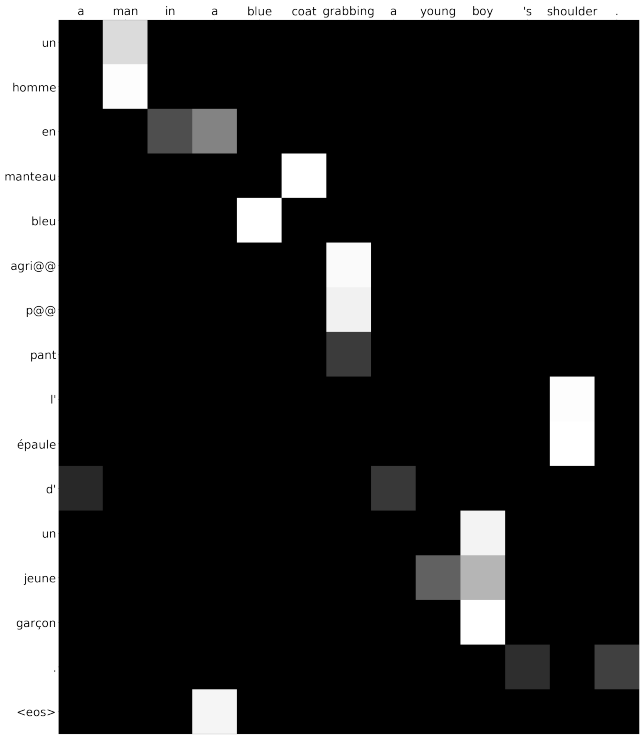


Fig. 3: Source–Target (EN→FR) word attention visualization corresponding to Table 2.

(a) "agrippant" → "*grabbing*".

(b) " l' " → "*the*/*a*". (c) "paule" → "*shoulder*". (d) Source image.

Fig. 4: Image–Target attention visualization corresponding to Table 2.

6.1 Contribution of semantic image region

As in the example of Table 2, we list source English sentence, French reference and French translations from three methods. In this example, translation is broken down into two subtasks: translate "grab" into French, and then transform it into the present participle. Verbs in French have many surface forms in the context according to different tenses, such as present simple, future, past, and others. Moreover, polysemy is ubiquitous. It is difficult to translate the verb into the exact meaning and form from the context only. Concentrating on the semantic image region during translation improved this deficiency.

By analyzing source and target token attention shown in Figure 3, it can be seen that the bright spots exactly represent the highest match degree between the source and generated tokens.

By comparing in Table 2, it can be obviously found that in the translation of the verb, our proposal gives better translation than both baselines. In the text-only attentive NMT, the word "grabbing" is not translated correctly. Its translation into English is "a man in a blue coat strolling with the shoulder of a young boy," which is different from the source sentence. On the other hand, in the doubly-attentive MNMT with grid image information from ResNet-101 CNN, only one of the two points was broken. It successfully translates "grab" into "agrippe," but fails to transform it into the present participle form. It does not catch the state of the verb. By contrast, our approach improves the translation performance benefited from the advantage of semantic image regions. We improve the translation of "grabbing" from "se baladant avec" to "agrippant," both in meaning and verb deformation.

In this part, in accordance with Figure 4, we evaluate what semantic image regions actually contribute. We visualize the time step of generating tokens of "agrippant" as well as its neighbor time step in context. Along with the generation of "agrippant," the semantic image regions lock onto the area of the image where the action is being performed, capturing the state of the action at the moment. Coordinately, the image-attention mechanism concentrates on the most weighted image region which is loaded with description feature, so as to affect the output token of the decoder.

To analyze this improvement quantitatively, we specifically extracted 20 source sentences which have present participle as accompanying adverbial. The results show that: the accuracy of translation into correct verb form in attentive NMT is 60% and that of doubly attentive MNMT is 65%, in contrast, our accuracy reaches 90%.

7. Conclusion

Image feature plays a positive role on machine translation and improves the accuracy of translation. Our main idea is to maximize the influence of semantic image features in NMT. Our proposal is a model that coalesces semantic image region and double attention to generate more vivid translation. By comparing with baselines, we achieve improvement benefited from semantic image regions. In the future, we will continue to explore what image features are more conducive to the translation process, and how to better integrate them.

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.

References

[1] Hodosh, M., Young, P. and Hockenmaier, J.: Framing Image Description As a Ranking Task: Data, Models and Evaluation Metrics, *J. Artif. Int. Res.*, Vol. 47, No. 1, pp. 853–899 (2013).
[2] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S. and Bengio, Y.: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, *ICML* (2015).
[3] Kiros, R., Salakhutdinov, R. and Zemel, R. S.: Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, *CoRR* (2014).
[4] Mao, J., Xu, W., Yang, Y., Wang, J. and Yuille, A. L.: Explain Images with Multimodal Recurrent Neural Networks, *CoRR* (2014).
[5] Vinyals, O., Toshev, A., Bengio, S. and Erhan, D.: Show and Tell: A Neural Image Caption Generator, *ICML* (2014).
[6] Malinowski, M., Rohrbach, M. and Fritz, M.: Ask Your Neurons: A Neural-based Approach to Answering Questions about Images, *ICCV* (2015).
[7] Shih, K. J., Singh, S. and Hoiem, D.: Where To Look: Focus Regions for Visual Question Answering, *CVPR* (2015).
[8] Xiong, C., Merity, S. and Socher, R.: Dynamic Memory Networks for Visual and Textual Question Answering, *ICML* (2016).
[9] Calixto, I. and Liu, Q.: Incorporating Global Visual Features into Attention-based Neural Machine Translation., *EMNLP*, pp. 992–1003 (2017).
[10] Calixto, I., Liu, Q. and Campbell, N.: Doubly-Attentive Decoder for Multi-modal Neural Machine Translation, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, Association for Computational Linguistics, pp. 1913–1924 (2017).
[11] Caglayan, O., Barrault, L. and Bougares, F.: Multimodal Attention for Neural Machine Translation, *CoRR* (2016).
[12] Huang, P.-Y., Liu, F., Shiang, S.-R., Oh, J. and Dyer, C.: Attention-based Multimodal Neural Machine Translation, *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, Berlin, Germany, Association for Computational Linguistics, pp. 639–645 (2016).
[13] Specia, L., Frank, S., Sima'an, K. and Elliott, D.: A Shared Task on Multimodal Machine Translation and Crosslingual Image Description, *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, Berlin, Germany, Association for Computational Linguistics, pp. 543–553 (2016).
[14] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S. and Zhang, L.: Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, *CVPR* (2018).
[15] Ren, S., He, K., Girshick, R. B. and Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, *NIPS* (2015).
[16] Mnih, V., Heess, N., Graves, A. and Kavukcuoglu, K.: Recurrent Models of Visual Attention, *NIPS* (2014).
[17] Bahdanau, D., Cho, K. and Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate, *ICLR* (2015).
[18] Grönroos, S., Huet, B., Kurimo, M., Laaksonen, J., Mérialdo, B., Pham, P., Sjöberg, M., Sulubacak, U., Tiedemann, J., Troncy, R. and Vázquez, R.: The MeMAD Submission to the WMT18 Multimodal Translation Task, *WMT* (2018).
[19] Girshick, R. B., Donahue, J., Darrell, T. and Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation, *CVPR* (2013).
[20] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *CVPR* (2015).
[21] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S. and Fei-Fei, L.: Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations, *International Journal of Computer Vision*, Vol. 123, pp. 32–73 (2016).
[22] Elliott, D., Frank, S., Sima'an, K. and Specia, L.: Multi30K: Multilingual English-German Image Descriptions, *CoRR* (2016).
[23] Young, P., Lai, A., Hodosh, M. and Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, *Transactions of the Association for Computational Linguistics*, Vol. 2 (2014).
[24] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation, *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 177–180 (2007).
[25] Sennrich, R., Haddow, B. and Birch, A.: Neural Machine Translation of Rare Words with Subword Units, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, Association for Computational Linguistics, pp. 1715–1725 (2016).
[26] Klein, G., Kim, Y., Deng, Y., Senellart, J. and Rush, A. M.: OpenNMT: Open-Source Toolkit for Neural Machine Translation, *ACL* (2017).
[27] Zeiler, M. D.: ADADELTA: An Adaptive Learning Rate Method, *CoRR* (2012).
[28] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: Bleu: a Method for Automatic Evaluation of Machine Translation, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, Association for Computational Linguistics, pp. 311–318 (2002).
[29] Denkowski, M. and Lavie, A.: Meteor Universal: Language Specific Translation Evaluation for Any Target Language, *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA, Association for Computational Linguistics, pp. 376–380 (2014).
[30] Koehn, P.: Statistical Significance Tests for Machine Translation Evaluation, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, Association for Computational Linguistics, pp. 388–395 (2004).