Region-attentive multimodal neural machine translation

Zhao, Yuting
Tokyo Metropolitan University

Komachi, Mamoru
Tokyo Metropolitan University
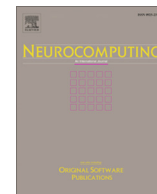
Kajiwara, Tomoyuki
Ehime University

Chu, Chenhui
Kyoto University

https://hdl.handle.net/2324/7173551

Region-attentive multimodal neural machine translation

Yuting Zhao [a,*], Mamoru Komachi [a], Tomoyuki Kajiwara [b], Chenhui Chu [c]

[a] *Tokyo Metropolitan University, 6-6 Asahigaoka, Hino, Tokyo 191-0065, Japan*
[b] *Ehime University, 3 Bunkyo-cho, Matsuyama, Ehime 790-8577, Japan*
[c] *Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan*

ARTICLE INFO

ABSTRACT

We propose a multimodal neural machine translation (MNMT) method with semantic image regions called region-attentive multimodal neural machine translation (RA-NMT). Existing studies on MNMT have mainly focused on employing global visual features or equally sized grid local visual features extracted by convolutional neural networks (CNNs) to improve translation performance. However, they neglect the effect of semantic information captured inside the visual features. This study utilizes semantic image regions extracted by object detection for MNMT and integrates visual and textual features using two modality-dependent attention mechanisms. The proposed method was implemented and verified on two neural architectures of neural machine translation (NMT): recurrent neural network (RNN) and self-attention network (SAN). Experimental results on different language pairs of Multi30k dataset show that our proposed method improves over baselines and outperforms most of the state-of-the-art MNMT methods. Further analysis demonstrates that the proposed method can achieve better translation performance because of its better visual feature use.

1. Introduction

Neural machine translation (NMT) has achieved state-of-the-art translation performance [43,16,3,45]. The strength of NMT lies in its ability to learn directly, in an end-to-end fashion, mapping from the input text to the associated output text. In the context of recurrent neural network (RNN), it was proposed to use internal state (memory) to process variable-length sequences of inputs, that is much better at capturing long-term dependencies [3]. In the context of self-attention network (SAN), a special attention mechanism was proposed for selecting specific parts of an input sequence by relating its elements at different positions, dispensing with recurrence entirely [45].

Multimodal NMT (MNMT) is a novel take on NMT that translates sentences in the presence of multimodal contents, aiming to tackle specific situations wherein only the textual contexts, such as ambiguous words and grammatical gender, are not sufficient for translation. Hence, many studies [42,21,4] have increasingly been focusing on incorporating visual input, particularly images, to improve translation.

The potential for improving translation quality using images has been pioneered by [22]. Subsequent studies [26,30,12] have started using a global visual feature extracted from an entire image to initialize encoder/decoder RNN hidden states to contextualize language representations. However, the effect of images cannot be fully exerted, as the visual features of an entire image are complex and non-specific.

To effectively use images, some studies [46,8,13] used spatial convolutional features extracted through convolutional neural networks (CNNs) pre-trained on ImageNet [40]. As these equally sized grid local visual features do not convey specific semantics, the role of visual modality only provides dispensable help for translation.

Other studies utilize richer local features for MNMT, such as DenseCap[1] in [18]. However, their efforts have not convincingly demonstrated that visual features can improve the translation quality. According to [10], when the textual context is limited, visual features can help generate better translations. MNMT disregards visual features because the quality of the image features or the way in which they are integrated is not satisfactory. Therefore, the types of visual features that are suitable and how these features should be integrated remain open questions.

In this study, as shown in Fig. 1, we attempt to combine object detection with an additional region-dependent attention mecha-

---

* Corresponding author.
 E-mail addresses: zhao-yuting@ed.tmu.ac.jp (Y. Zhao), komachi@tmu.ac.jp (M. Komachi), kajiwara@cs.ehime-u.ac.jp (T. Kajiwara), chu@i.kyoto-u.ac.jp (C. Chu).

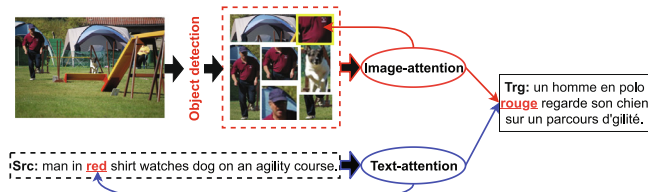[1] https://github.com/jcjohnson/densecap.

Fig. 1. Overview of region-attentive multimodal neural machine translation (RA-NMT).

nism for fully exploiting semantic image region features upon NMT architectures, which is called region-attentive multimodal neural machine translation (RA-NMT). In RA-NMT, it is possible to focus on different parts of the source sentence and different object-level regions of the image at the same time. The main motivation behind this is that we expect the proposed method to take advantage of useful visual information by attending to specific regions of the image to assist in translating source words. As suggested in [20], attending to specific regions of the image is crucial for improving the translation.

Technically, rather than equally sized grid local visual features, we present that semantic image region features containing object attributes and relationships are essential to MNMT. Furthermore, inspired by previous studies [8,13,37,17] on the investigation of the attention mechanism for multi-source learning, we introduce that a region-dependent attention mechanism is a promising way to make MNMT attend to the salient regions of an image. Therefore, instead of utilizing regional features to initialize/contextualize language representations [30,31], we propose integrating semantic image region features into MNMT with two modality-dependent attention mechanisms, one for text and the other for the semantic image regions, which is significantly different from the previous studies.

Although we have implemented and verified the proposed method on the RNN-based NMT architecture in our previous study [51], how to implement the proposed method and whether it is compatible and effective on the SAN-based NMT architecture remain challenging. In this study, we implemented and verified the proposed method on not only the RNN-based architecture but also SAN-based architecture, which are called region-attentive multimodal RNN (RA-RNN) method and region-attentive multimodal SAN (RA-SAN) method, respectively.

Additionally, the effect of the number of semantic image regions on translation performance is explicitly clarified in this study. We conducted experiments on different numbers of semantic image regions based on both the RA-RNN and RA-SAN methods, and a detailed analysis was performed. Furthermore, we also carried out a pairwise evaluation and qualitative analysis within/between RNN-based and SAN-based architectures to demonstrate the translation performance of our proposed method.

In contrast to our previous study [51], the main contributions of this study are fourfold:

- We propose multimodal method that combines object detection with an additional region-dependent attention mechanism to fully exploit semantic image region features on NMT architectures, which is called RA-NMT. This proposal is implemented and verified on two types of NMT architectures: RNN and SAN.
- Extensive experimental results show that our proposed method improves over baselines on both RNN and SAN architectures. A further experimental comparison shows that our proposed method outperforms most existing MNMT methods.

- To investigate the effect of the number of semantic image region features on our proposal, we conducted experiments on both the RA-RNN and RA-SAN methods and performed a detailed analysis.
- Further analysis demonstrates that the proposed method can make better use of visual information by attending to specific semantic image regions with an additional region-dependent attention mechanism.

2. Related work

Some MNMT models integrate visual information using a single global visual feature vector extracted by CNNs. For example, some models use the global visual feature in the following ways: initializing the encoder/decoder hidden states [22,30,12]; performing element-wise multiplication with target word embeddings [5]; impacting the text encoder by learning an image representation jointly [24,29]. In addition, some models use the global visual feature to interact between the sources through a latent variable [14], a shared space [52], or a universal representation [50]. Although they aim to combine text and image sources to generate a good translation, it is difficult to summarize all the semantic information of an entire image into a single feature vector.

Other studies represent visual information with a sequence of equally sized grid local visual feature vectors extracted by CNNs. These grid features are used to preserve the spatial correspondence with the input image. For example, a joint representation is generated by combining visual and textual representations [25], compute a multimodal context vector using a multimodal or filtered attention mechanism [8,6,7], and focus on textual and visual annotations independently by different strategies on attention mechanisms [11,13,37,17]. Although they aim to use part of the image related to the text's semantics, it is difficult to distinguish the equally sized grid local features in the image.

To overcome the above difficulties, current studies attempt to represent an image using multiple object-level regional features. [30], for example, integrated regional features followed by the text sequence. [44] proposed a transformation to mix global visual features and regional features. [27] and [31] generated a single representation of regional features to initialize the encoder or target word embeddings. Furthermore, [47] proposed a multi-head co-attention upon regional features. [49] used a unified multimodal graph to capture semantic relationships between words and objects. So far, how to fully exploit visual information in MNMT remains an open question.

3. Proposed methods

3.1. RA-RNN: Region-Attentive Multimodal RNN

As shown in Fig. 2, the proposed RA-RNN, based on [13], comprises three parts: sentence encoder, image encoder, and decoder.

We integrate the visual features using an additional attention mechanism. From the source sentence $X = (x_1, x_2, x_3, \cdots, x_n)$ to the target sentence $Y = (y_1, y_2, y_3, \cdots, y_g)$, the image attention mechanism focuses on all semantic image region features to calculate the image context vector z_t, whereas the text-attention mechanism computes the text context vector c_t. The decoder is an RNN with conditional gated recurrent unit (cGRU)[2] to generate the current hidden state s_t and target word y_t on two attention mechanisms.

At time step t, a hidden state proposal $\hat{s}_t$ is initially computed in cGRU, and then the image context vector z_t and text context vector c_t are calculated.
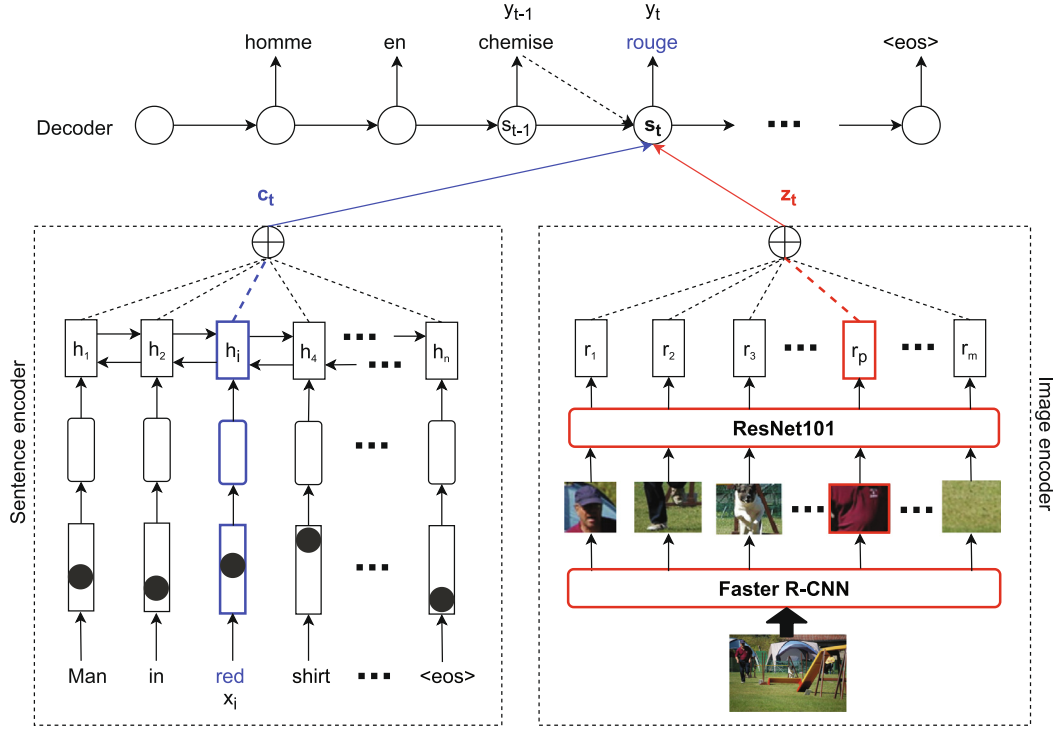
---

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

Fig. 2. RA-RNN: Region-Attentive Multimodal RNN.

$$\hat{\xi}_t = \sigma(W_\xi E_Y[y_{t-1}] + U_\xi s_{t-1})$$
$$\hat{\gamma}_t = \sigma(W_\gamma E_Y[y_{t-1}] + U_\gamma s_{t-1})$$
$$\dot{s}_t = \tanh(W E_Y[y_{t-1}] + \hat{\gamma}_t \odot (U s_{t-1}))$$
$$\hat{s}_t = \left(1 - \hat{\xi}_t\right) \odot \dot{s}_t + \hat{\xi}_t \odot s_{t-1}$$

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

3.1.1. Sentence encoder

The sentence encoder is a bi-directional RNN with GRU [15]. Given a sentence $X = (x_1, x_2, x_3, \cdots, x_n)$, the encoder updates the forward hidden states with annotation vectors $(\overrightarrow{h_1}, \overrightarrow{h_2}, \overrightarrow{h_3}, \cdots, \overrightarrow{h_n})$, and updates the backward with annotation vectors $(\overleftarrow{h_1}, \overleftarrow{h_2}, \overleftarrow{h_3}, \cdots, \overleftarrow{h_n})$. By concatenating the forward and backward vectors $h_i = \left[\overrightarrow{h_i}; \overleftarrow{h_i}\right]$, each h_i encodes the entire sentence while focusing on the x_i word, and all words in a sentence are denoted as $C = (h_1, h_2, \cdots, h_n)$.

3.1.2. Image encoder

The image encoder is an object-detection-based approach following [1], acting as a feature extractor in the object-level image region.

As shown in Fig. 2, when given an input image, the image encoder first employs an object detection method, which is Faster R-CNN [39] pre-trained on Visual Genome [36], to propose m object-level image regions from each image. Then, based on the detected object-level image regions, a ResNet101 [28] pre-trained on ImageNet [40] is utilized to extract semantic image region features. Finally, each semantic image region feature is represented as a vector r with dimensions d_r, and all of these features in each image are denoted as $R = (r_1, r_2, r_3, \cdots, r_m)$.

3.1.3. Decoder

The decoder comprises three parts: the text-attention mechanism, image-attention mechanism, and generation.

Text-attention mechanism. At time step t, the text context vector c_t is generated as follows:

$$e_{t,i}^{\text{text}} = \left(V^{\text{text}}\right)^{\text{T}} \tanh(U^{\text{text}} \hat{s}_t + W^{\text{text}} h_i)$$
$$\alpha_{t,i}^{\text{text}} = \text{softmax}\left(e_{t,i}^{\text{text}}\right)$$
$$c_t = \sum_{i=1}^{n} \alpha_{t,i}^{\text{text}} h_i$$

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

Image-attention mechanism. At time step t, the image-attention mechanism focuses on the m semantic image region features and computes the image context vector z_t.

We initially calculate the attention energy $e_{t,p}^{\text{img}}$, which scores the degree of output matching between the inputs around position p and the output at position t, as follows:

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

where $V^{\text{img}}, U^{\text{img}}$, and W^{img} are trainable parameters.

Then, the weight matrix $\alpha_{t,p}^{\text{img}}$ of each r_p is computed as follows:

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

At time step t, the image-attention mechanism dynamically focuses on the m semantic image region feature vectors and computes the image context vector z_t, as follows:

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

For z_t, at each decoding time step t, a gating scalar $\beta_t \in [0, 1]$ [46] was used to adjust the proportion of the image context vector according to the previous hidden state s_{t-1}.

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

where W_β and b_β are trainable parameters.

Generation. At time step t of the decoder, the new hidden state s_t is generated in the cGRU, as follows:

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

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 $\overline{U}$ 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 output probability is computed as follows:

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

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

3.2. RA-SAN: Region-Attentive Multimodal SAN

As shown in Fig. 3, RA-SAN comprises three parts: encoder, decoder, and image encoder. We propose RA-SAN based on transformer architecture [45]. In the decoder, we implement two modality-dependent cross-attention mechanisms over the multi-source (image, text). The image encoder follows the method described in Section 3.1.2.

3.2.1. Encoder

To represent source sentences, an input embedding layer acts as a lookup table to map each word to a vector representation. Because the encoder in the transformer has no recurrence like that in RNN, it is necessary to inject positional information into the input embeddings, which is done using positional encoding.

The encoder comprises a stack of N identical layers. Each layer has self-attention and feed-forward sublayers. The self-attention sub-layer is a multi-head attention mechanism that allows the model to jointly attend to information from different representation subspaces. The feed-forward sub-layer is a basic, position-wise, fully connected feed-forward network, which is applied to each position separately and identically.

In addition to the two sub-layers described above, the residual connection [28] and layer normalization [2] are also key components of the transformer. There is a residual connection around every one of the two sublayers and a layer normalization inside the residual connection in our model. Therefore, the output of each sublayer is defined as $(x + \text{Sublayer}(\text{LayerNorm}(x)))$, where Sublayer() is the function implemented by the sublayer itself. To encourage these residual connections, all sublayers and embedding layers produce outputs of dimension d_{model}.

3.2.2. Decoder

The decoder comprises a stack of N identical layers. In addition to the two sub-layers similar to the encoder, the decoder inserts two cross-attention mechanisms between them. One is text cross-attention, which performs multi-head attention on encoder output features. The other is image cross-attention, which performs multi-head attention over semantic image region features. There is also a residual connection around every sublayer and a layer normalization inside the residual connection, similar to the encoder.
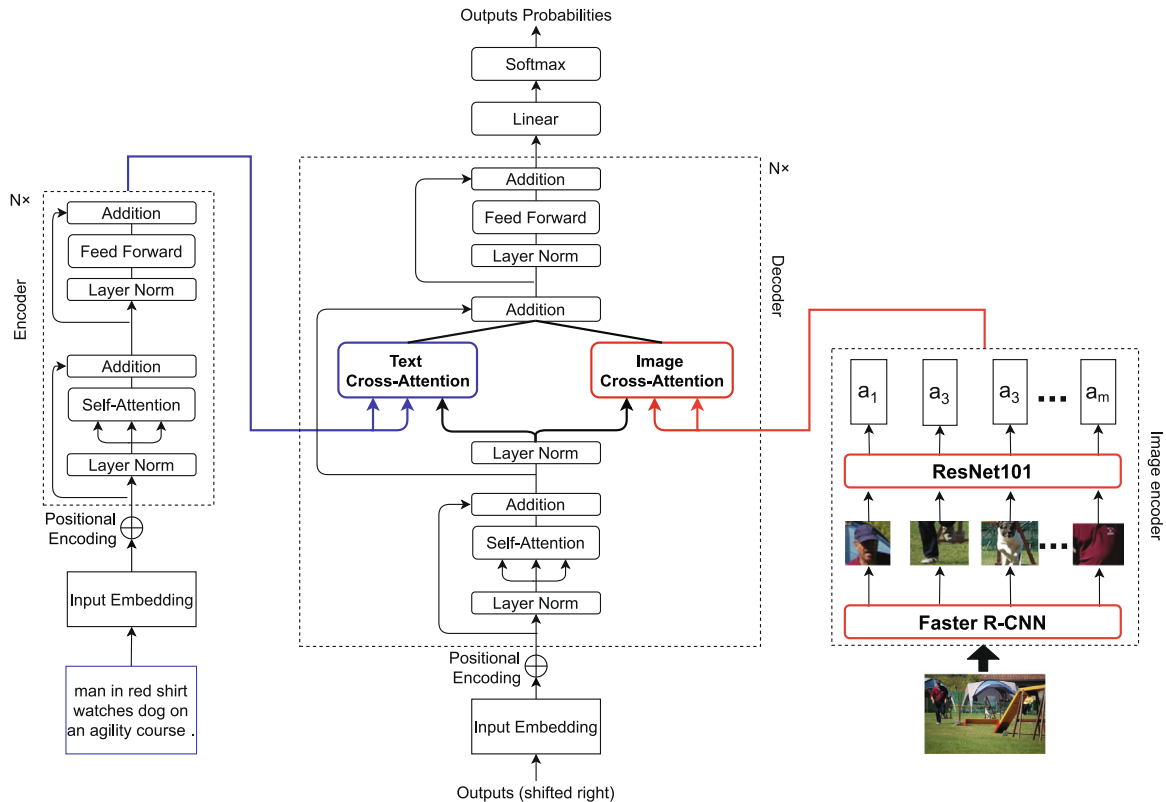

Fig. 3. RA-SAN: Region-Attentive Multimodal SAN.

When generating a target word at a time step t, the attention from one of the sources may be strong or weak from the other, and thus, summing two cross-attention outputs would help learn the better translation. Therefore, the summarized output from two cross-attentions is fed into the feed-forward network sublayer, which consists of two linear transformations with a ReLU activation in between:

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

While the linear transformations are the same across different positions, they use different parameters from layer to layer, where W_1, W_2, b_1, and b_2 are trainable parameters. In this equation, the dimensions of the input and output are d_{model}, and the inner feed-forward neural network layer has dimensions d_{ff}. Finally, the decoder is capped off with a linear layer that acts as a classifier and a softmax layer to obtain the target word probabilities.

3.2.3. Double cross-attentions

As illustrated on the left of Figs. 4 and 5, conventional cross-attention in the transformer acts as a query mapping of key-value sets to an output, which is multi-head attention that performs the attention function on the encoder output features using H heads in parallel. Each scaled dot-product attention process is called one head. Each head produces an output vector that is concatenated into a single vector before passing through the final linear layer.

The input involves queries and keys of dimension d_k and values of dimension d_v. Each query is multiplied with all keys by dot product multiplication and scaled by $\sqrt{d_k}$; then, there is an src_padding on padding source text input into the maximum length. Finally, the softmax function is applied to obtain the weights of the values. The final output of the scaled dot-product attention is computed as the weighted sum of the values. The weight assigned to each value is calculated using the compatibility function of the query with the corresponding key.

The cross-attention is simultaneously calculated on a set of queries, keys, and values and packed together into a matrix Q, K_t, V_t. The output matrix is computed as follows:

$$Attention(Q, K_t, V_t) = \text{softmax}\left(\frac{QK_t^T}{\sqrt{d_k}}\right)V_t$$

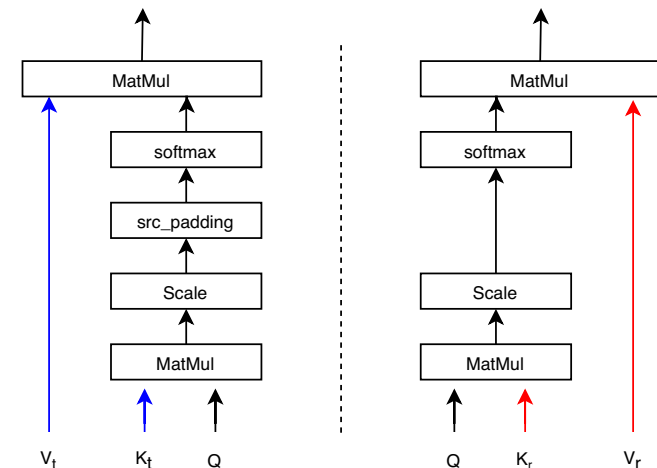$$MultiHead(Q, K_t, V_t) = \text{Concat}\left(head_t^1, \ldots, head_t^H\right)W^O$$

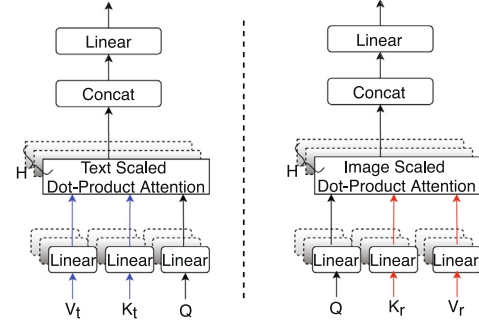**Fig. 4.** Left: Text Scaled Dot-Product Attention; Right: Image Scaled Dot-Product Attention.

Fig. 5. Left: Text Multi-Head Attention; Right: Image Multi-Head Attention.

where $head_t^{i\in[1,H]} = Attention\left(QW_i^Q, K_t W_i^K, V_t W_i^V\right)$.

The projections are parameter matrices:

$$W_i^Q \in \mathbb{R}^{d_{model}\times d_k}$$
$$W_i^K \in \mathbb{R}^{d_{model}\times d_k}$$
$$W_i^V \in \mathbb{R}^{d_{model}\times d_v}$$
$$W^O \in \mathbb{R}^{Hd_v\times d_{model}}$$

In the proposed RA-SAN, we use the conventional cross-attention in the transformer as the text cross-attention mechanism. We implemented an additional image cross-attention mechanism, which is multi-head attention, that performs the attention function on m semantic image region features using H heads in parallel.

The image cross-attention mechanism is illustrated on the right side of Figs. 4 and 5. Unlike text-scaled dot-product attention, the image-scaled dot-product attention has no source input padding because the number of semantic image regions is fixed. The image cross-attention mechanism is defined as:

$$Attention(Q, K_r, V_r) = \text{softmax}\left(\frac{QK_r^T}{\sqrt{d_k}}\right)V_r$$

$$MultiHead(Q, K_r, V_r) = \text{Concat}\left(head_r^1, \ldots, head_r^H\right)W^o$$

where $head_r^{i\in[1,H]} = Attention\left(QW_i^q, K_r W_i^k, V_r W_i^v\right)$.

The projections are parameter matrices:

$$W_i^q \in \mathbb{R}^{d_{model}\times d_k}$$
$$W_i^k \in \mathbb{R}^{d_{model}\times d_k}$$
$$W_i^v \in \mathbb{R}^{d_{model}\times d_v}$$
$$W^o \in \mathbb{R}^{Hd_v\times d_{model}}$$

4. Experiments

4.1. Dataset

We experimented on English→German (En→De) and English→French (En→Fr) tasks of the Multi30k dataset [23]. The dataset contained 29 k training images and 1,014 validation images. For testing, we used the 2016 test set, which included 1,000 images. Each image was paired with its English descriptions as well as human translations of German and French. We used Moses [35] toolkit[3] to normalize and tokenize all sentences. Then, we converted the space-separated tokens into sub-word units using the byte pair encoding (BPE) model [41].[4] With 10k merge operations, the result-

[3] https://github.com/moses-smt/mosesdecoder.
[4] https://github.com/rsennrich/subword-nmt.

ing vocabulary sizes of each language pair were 5,202→7,065 tokens for En→De and 5,833→6,575 tokens for En→Fr. The number of tokens in the sentence was limited to a maximum of 100. We trained models to translate from English to German/French and report the evaluation of cased, tokenized sentences with punctuation.

4.2. Evaluation metrics

We evaluated the quality of translation according to the token-level BLEU [38] and METEOR [19] metrics.

We trained all models three times and calculated the BLEU and METEOR scores. Finally, we reported the average over three runs. Moreover, we reported the statistical significance of BLEU using bootstrap resampling [34] over a merger of three test translation results. We defined the statistical significance test threshold as 0.05, and reported only when the p-value was less than the threshold.

4.3. Baselines

4.3.1. RNN-based baselines

RNN. We trained a text-only RNN model using the OpenNMT [33] toolkit[5] as a baseline. The RNN was trained on En→De and En→Fr, wherein only the textual part of Multi30k was used. This architecture comprises a 2-layer bidirectional GRU encoder and a 2-layer cGRU decoder with an attention mechanism.

Grid-attentive multimodal RNN (GA-RNN). We trained a GA-RNN [13] model[6] as another baseline, which was extended from OpenNMT. This architecture comprises a 2-layer bidirectional GRU encoder and a 2-layer cGRU decoder with two attention mechanisms. We trained this model with 7×7 equally sized gird local visual features from each image extracted by a ResNet101 pre-trained on ImageNet. Each grid-based local visual feature was represented as a 2,048-dimension vector.

4.3.2. SAN-based baselines

SAN. We trained a text-only SAN model using the transformer's settings in the OpenNMT toolkit as a baseline. The SAN was also trained on only the textual part of Multi30k on En→De and En→Fr tasks.

Grid-attentive multimodal SAN (GA-SAN). We trained a GA-SAN model based on the GA-RNN model by modifying the transformer's settings in the OpenNMT toolkit as another baseline. An image cross-attention mechanism was implemented on the grid-based local visual features in the transformer's settings. This architecture was also trained with 7×7 gird-based local visual features from each image extracted by a ResNet101 pre-trained on ImageNet, and each feature was represented as a 2,048-dimension vector.

4.4. Setup

We implemented our proposed RA-RNN and RA-SAN based on GA-RNN and GA-SAN baselines, respectively, by modifying the image attention mechanism to focus on m semantic image region feature vectors generated from the image encoder. For the image encoder in both the RA-RNN and RA-SAN methods, the number of semantic image region features was set to $m = 100$ and the dimension of regional feature vectors was set to $d_r = 2,048$.

4.4.1. Settings of RNN-based architectures

We set the hidden state dimension of the bi-directional GRU encoder and cGRU decoder to 500, source word embedding dimen-

sion to 500, sentence-minibatches 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 ADAM [32] and a learning rate of 0.002 for 25 epochs. Finally, after both the validation perplexity and accuracy converged, the model with the highest BLEU score of the validation set was selected to evaluate the test set.

4.4.2. Settings of SAN-based architectures

We set $N = 6$ layers for the encoder and decoder. The number of dimensions of all the input and output layers was set to $d_{model} = 512$. The inner feed-forward neural network layer was set to $d_{ff} = 2,048$. The heads of all the multi-head modules were set to $H = 8$ in both the encoder and decoder layers. We applied linear projection on visual features to reduce the dimensions from 2,048 to 512 to have the same size as word embeddings. We applied a dropout of 0.3 on linear projection. During training, the sentence-minibatches were set to 40, the value of label smoothing was set to 0.1, and the attention dropout and residual dropout were 0.3. An Adam optimizer was used to tune the model parameters. The learning rate was set to two with a warm-up step of 8,000. We trained the model up to 100 epochs, and the model with the highest BLEU score of the validation set was selected to evaluate the test set.

4.5. Results

Table 1 presents the experimental results of RNN-based architectures, showing that the proposed RA-RNN achieves better performance than both the text-only RNN baseline and the GA-RNN baseline in all translation tasks. In particular, the results of the RA-RNN are significantly better than those of the text-only RNN baseline with a p-value of < 0.05 on both language pairs. This illustrates that integrating semantic image region visual features is capable of promoting translation performance, and our proposed method can make better use of visual information.

Table 2 presents the experimental results of the SAN-based architectures, showing that the proposed RA-SAN outperforms the baselines on both the En→De and En→Fr tasks. It is worth noting that our RA-SAN results are significantly better than not only the text-only SAN baseline but also the GA-SAN baseline with a p-value of < 0.05 on both tasks. This demonstrates that the proposed method is universal, which can result in consistent improvements in performance on different NMT architectures. Thus, we confirm the effectiveness and generality of the proposed method.

4.6. Comparison with existing methods

To further verify the merit of the proposed method, we also implemented the proposed method on the state-of-the-art text-only NMT baseline mentioned in [14] and the state-of-the-art transformer baseline mentioned in [49], respectively. Furthermore, we compared the experimental results of our proposed method with the following state-of-the-art MNMT methods:

Parallel RCNNs [30]: The encoder of RNN is composed of multiple encoding threads. In each thread, a regional visual feature is followed by a text sequence.

$NMT_{SRC+IMG}$ [13]: Integrates two separate attention mechanisms over the source words and conventional grid local visual features in a cGRU decoder.

IMG_D [12]: Integrates global visual features as additional data to initialize the decoder hidden state.

Imagination [24]: Jointly learns a translation model and visually grounded representations.

{*Soft, Stochastic*} *Attention + Grounded Image (GI)* [17]: Employs two kinds of attention mechanisms, which are superimposed by

---

[5] https://github.com/OpenNMT/OpenNMT-py.
[6] https://github.com/iacercalixto/MultimodalNMT.

Table 1
The experimental results of RNN-based architectures. The best performance is highlighted in bold. † indicates that the result is significantly better than the text-only RNN baseline at a p-value of < 0.05.

Methods	En→De		En→Fr	
	BLEU	METEOR	BLEU	METEOR
RNN	34.8	53.4	56.5	71.9
GA-RNN	36.5	54.8	57.8	72.8
RA-RNN	**36.9**†	**55.5**	**58.1**†	**73.2**
v.s. RNN	(↑ 2.1)	(↑ 2.1)	(↑ 1.6)	(↑ 1.3)
v.s. GA-RNN	(↑ 0.4)	(↑ 0.7)	(↑ 0.3)	(↑ 0.4)

Table 2
The experimental results of SAN-based architectures. The best performance is highlighted in bold. † and ‡ indicate that the result is significantly better than the text-only SAN and GA-SAN baselines at p-value < 0.05, respectively.

Methods	En→De		En→Fr	
	BLEU	METEOR	BLEU	METEOR
SAN	35.4	52.8	57.4	72.2
GA-SAN	37.5	55.6	59.5	74.4
RA-SAN	**38.0**†‡	**56.0**	**60.1**†‡	**74.8**
v.s. SAN	(↑ 2.6)	(↑ 3.2)	(↑ 2.7)	(↑ 2.6)
v.s. GA-SAN	(↑ 0.5)	(↑ 0.4)	(↑ 0.6)	(↑ 0.4)

an additional grounding attention method, for considering visual annotations of image feature maps to generate context vectors.

VMMT$_F$ [14]: An MNMT model that incorporates image context through a latent variable model.

Del+Obj [31]: A transformer-based deliberation model enriched with object-level features.

MTF [48]: A transformer-based NMT model with multimodal self-attention to integrate text and image features.

GMFE-NMT [49]: A transformer-based NMT model integrated with a multimodal graph neural network (GNN) encoder on the grounding-based correspondences between phrase-level words and regions.

As shown in Table 3, all the existing methods are divided into two groups: RNN-based methods and SAN-based methods. Then, we display the experimental results of the proposed method and the state-of-the-art methods' results for the respective group. Note that previous methods mainly report the results on the En→De language pair of the Multi30k 2016 test set, and hence, the existing results on the En→Fr task are fewer than those of the En→De task.

By comparing the performance of the proposed method with the state-of-the-art methods, we draw two conclusions as follows:

First, the proposed method outperforms the state-of-the-art text-only baselines on different basic neural architectures. For instance, the proposed method in the respective group outperforms the text-only NMT baseline and the transformer baseline by 1.6 and 0.6 BLEU scores, respectively, on the En→Fr task. Therefore, we can confirm the effectiveness and generality of the proposed method.

Second, the evaluation results of the proposed method outperform most of the existing MNMT methods. Among the SAN-based methods, our proposed method achieves the best performance evaluated by the METEOR score on both language pairs; furthermore, the results of the proposed method surpass all the METEOR scores in the RNN-based methods on different language pairs as well. This demonstrates that our proposed method is competitive among all the state-of-the-art MNMT methods.

On the other hand, we notice that our method underperforms [17,49]. We conjecture that the reason for this is inseparable from the superimposed use of the grounding methods in [17,49]. In contrast, our method does not take advantage of additional methods to eliminate interference information on the image side but still achieves effective visual information usage. Although our results do not outperform theirs, our method is simpler and less computationally expensive than theirs.

5. Analyses

5.1. Effect of the number of semantic image region features m

In order to analyze the influence of the number of semantic image region features on our proposed method, we show the experimental results on different numbers of semantic image region features in Fig. 6 and report the computational cost in Table 4.

Specifically, for the setup of the image encoder in the proposed method mentioned in Section 4.4, instead of setting $m = 100$, the number of semantic image region features was set to $m = [10, 20, 40, 60, 80, 100, 120, 140, 160, 180, 200]$, respectively. All the variants are attempted in both the RA-RNN and RA-SAN methods on the En→De translation task. In Fig. 6, all the experimental results evaluated by BLEU and METEOR are the average score over three runs.

Holistically, either on RA-RNN or on RA-SAN, there is no big gap in the performance of the proposed method on different numbers of semantic image region features. As m increases, the BLEU/METEOR scores fluctuate up and down within a limited range. This proves the stability of our proposed method and indicates that the performance of the proposed method is not sensitive to the number of semantic image region features, which further illustrates that it is the effective use of semantic image region features rather than their numbers that determine the translation performance.

Additionally, we also show the computational cost of different variants in Table 4, which includes training speed, decoding speed, and elapsed time per training epoch. All the variants with different m in the RA-RNN and RA-SAN methods have experimented on the En→De translation task. In order to make a clear illustration, we report the computational cost of the variants, including the maximum value of $m = 200$, the minimum value of $m = 10$, and $m = 100$.

In general, with the increase of m, there is a drop of the training and decoding speed simultaneously, and the elapsed time per training epoch has become longer correspondingly. Overall training and decoding speed are reasonable. In more detail, in the case of the same m, the training and decoding speed of the RA-SAN are slower than the RA-RNN. We conjecture that as the calculation of the image-attention in the RA-RNN is additive, but the calculation of the image cross-attention in the RA-SAN is scaled dot-product, the computational cost of the RA-SAN is slightly larger than that of the RA-RNN.

5.2. Pairwise evaluation

To further analyze the translation performance of our proposed method, we performed a pairwise evaluation and statistical analysis. The results of the pairwise evaluation of the En→Fr language pair are summarized in Table 5.

Based on two kinds of NMT architectures, we conducted three groups of comparisons. Specifically, we compared the proposed RA-RNN/RA-SAN translations with their corresponding baselines' translations to identify improvement or deterioration of translation performance, and we compared the translations of RA-RNN

Table 3
Comparison with existing methods. Among all the results, we highlight the best performance in bold. All the experimental results of our proposal are the average scores over three runs.

Methods	En→De		En→Fr	
	BLEU	METEOR	BLEU	METEOR
RNN-Based Methods				
Text-only NMT [33,14]	35.0	54.9	56.5	71.9
Parallel RCNNs [30]	36.5	54.1	N/A	N/A
NMT$_{SRC+IMG}$ [13]	36.5	55.0	57.8	72.8
IMG$_D$ [12]	37.3	55.1	N/A	N/A
Imagination [24]	36.8	55.8	N/A	N/A
Soft Attention + GI [17]	37.6	55.3	N/A	N/A
Stochastic Attention + GI [17]	38.2	55.4	N/A	N/A
VMMT$_F$ [14]	37.7	56.0	N/A	N/A
Our Proposal$_{(OpenNMT)}$[11]	36.9	55.5	58.1	73.2
SAN-Based Methods				
Text-only transformer [45,49]	38.4	56.5	59.5	73.7
Del+Obj [31]	38.0	55.6	59.8	74.4
MTF [48]	38.7	55.7	N/A	N/A
GMFE-NMT [49]	**39.8**	57.6	**60.9**	74.9
Our proposal$_{(Nmtpytorch)}$[12]	38.6	**57.7**	60.1	**75.0**

[11] The results of our proposal reported here are implemented on the state-of-the-art text-only NMT baseline mentioned in [14] using the OpenNMT toolkit. The experimental settings are consistent with the setup described in Section 4.4.1.
[12] The results of our proposal reported here are implemented on the state-of-the-art transformer baseline mentioned in [49] using the Nmtpytorch toolkit [9]. The experimental settings are consistent with the setup in Section 4.4.2, except that the learning rate was tuned to 0.03 and the model was trained up to 300 epochs.
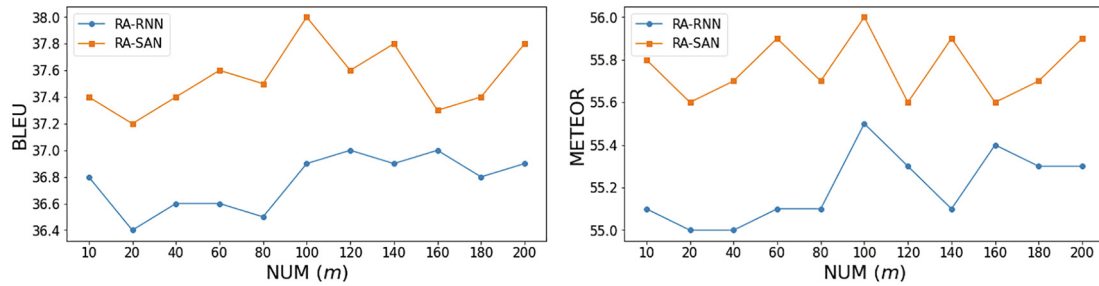

Fig. 6. Investigation of the effect of the number of semantic image region features m on our proposed method. The BLEU and METEOR results in the RA-RNN and RA-SAN methods are the average score over three runs on the En→De task..

Table 4
Computational cost: training speed (source tokens per second), decoding speed (target tokens per second), and elapsed time in seconds (s) per epoch. All the variants on both the RA-RNN and RA-SAN methods are attempted with $m = [10, 100, 200]$ on the En→De task, respectively.

Variants		Training	Decoding	Time (s)
RA-RNN	$m = 10$	9,140	7,973	52
	$m = 100$	5,486	5,296	78
	$m = 200$	3,847	3,767	110
RA-SAN	$m = 10$	5,575	5,788	72
	$m = 100$	4,267	4,549	91
	$m = 200$	2,868	3,073	135

and RA-SAN to identify which architecture can achieve better translation performance. For each group, we randomly selected 50 examples for evaluation and categorized 50 investigated examples into various categories by counting the number and proportion.

After statistical analysis, we find that almost half of the investigated examples show that our RA-RNN performs better than at least one baseline model. Similarly, half of the investigated examples show that our RA-SAN outperforms at least one baseline model. It is further verified the effectiveness and generality of our proposed method. Moreover, the number of examples in which our RA-SAN is better than both baselines is slightly improved

compared with a similar case of the RA-RNN. By comparing the translation performance of our RA-SAN and RA-RNN, we find that the number of examples where RA-SAN is better than RA-RNN is four times larger than the opposite cases. This illustrates that our RA-SAN can achieve a better translation performance compared with RA-RNN.

5.3. Qualitative analysis

For qualitative analysis, we analyze translation performance by comparing the translation results of the proposed method and its baselines, along with visualizing the semantic image regions that are attended by the image-attention mechanism at every time step.

According to the attention weight assigned to each region, the semantic image regions are shown with deep or shallow transparency in the image at every time step. 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. In the image, a blue bounding box indicates the most weighted image region, and the red text along with the bounding box shows the target word generated at that time step. Then, we analyze whether the semantic image regions have a positive or negative effect at the time step when a target word is generated.

Table 5
Pairwise evaluation. We counted the number and proportion of various categories among 50 random examples.

RA-RNN v.s. RNN-based baselines		
Better than both baselines	8	(16%)
Better than GA-RNN baseline	6	(12%)
Better than RNN baseline	10	(20%)
No change	24	(48%)
Deteriorated	2	(4%)
RA-SAN v.s. SAN-based baselines		
Better than both baselines	10	(20%)
Better than GA-SAN baseline	4	(8%)
Better than SAN baseline	11	(22%)
No change	24	(48%)
Deteriorated	1	(2%)
RA-SAN v.s. RA-RNN		
RA-SAN is better than RA-RNN	8	(16%)
RA-RNN is better than RA-SAN	2	(4%)
No change	40	(80%)

To distinguish the translation quality, we highlight the better translation with blue and worse translation with red.

5.3.1. Analysis within RNN-based models

In Fig. 7, we present two examples to analyze the effect of semantic image regions on translation quality within RNN-based models. The first is an example of a positive effect, whereas the second is the opposite.

For the first example, it illustrates that the semantic image regions of the proposed method can play a positive role in providing object attributes.

In detail, by comparing the translation result of our RA-RNN and its baselines, we find that the RA-RNN translates "striped beach chairs" better, which is a phrase made up of an adjective and a noun. From the visualization of the most weighted semantic image region, we can identify the semantic of "chairs" and "striped," respectively.

For the second example, it presents that attending to the semantic image regions that are not related to the text's semantics is not helpful for translation performance.

As shown in the example, "air" is correctly translated by baselines. However, the RA-RNN translates "in the air" into "du vol (of the flight)." We observe that the transparent semantic image regions with the top five weights in the image are scattered and unconnected. We can not understand any semantic information in the visualized image regions. We speculate that the word "air" is challenging to interpret depending on visual features. Furthermore, the proposed method translates it into "vol (flight)," which is close to another meaning of the polysemous "air," not completely different from the original meaning.

5.3.2. Analysis within SAN-based models

In Fig. 8, we present two examples to analyze the effect of semantic image regions on translation quality within SAN-based models. The first is an example of a positive effect, whereas the second is the opposite.

For the first example, it shows that the semantic image regions of the proposed method can play a positive role in providing verb attributes.

In this example, compared with baselines' translations, the RA-SAN translates "operating" better, which is a verb. By visualizing the most weighted semantic image region, we can identify the semantic of "operate.".

For the second example, we find that the semantic image regions of the proposed method have no effect on distinguishing synonyms.

As illustrated in the example, "wearing" is correctly translated by baselines. However, the RA-RNN translates the verb into "in," which is a preposition. Although we can identify the semantic of "wearing" from the most weighted semantic image region, it can also be understood as "in.".

5.3.3. Analysis between RA-RNN and RA-SAN

In Fig. 9, we present two examples to analyze the effect of semantic image regions on translation quality between RA-RNN

	chaises	rayées			
			Src (En)	two people are sitting fishing on striped beach chairs in a body of water .	
			Ref (Fr)	deux personnes sont assises dans des fauteuils de plage rayés , pêchant dans une étendue d's eau .	
			RNN	deux personnes sont assises sur une structure de plage rayée (a striped beach structure) dans un plan d's eau .	
			GA-RNN	deux personnes sont assises , pêchent sur une plage de sable (a sandy beach) dans un plan d's eau .	
			RA-RNN	deux personnes sont assises à pêcher sur des chaises rayées (striped chairs) dans un plan d's eau .	

	du	vol			
			Src (En)	men playing volleyball , with one player missing the ball but hands still in the air .	
			Ref (Fr)	des hommes jouant au volleyball , avec un joueur ratant le ballon mais avec les mains toujours en l's air .	
			RNN	des hommes jouant au volleyball , un joueur à l's attraper , mais les autres mains ayant toujours dans les airs (in the air) .	
			GA-RNN	des hommes jouant au volley-ball , avec un joueur qui le regarde dans les airs (in the air) .	
			RA-RNN	des hommes jouant au volleyball , avec un joueur qui passer le ballon mais les mains du vol (of the flight).	

Fig. 7. Examples for recurrent neural network (RNN), grid-attentive multimodal RNN (GA-RNN), and region-attentive multimodal RNN (RA-RNN). Red and blue words indicate incorrect and correct, respectively.

Fig. 8. Examples for self-attention network (SAN), grid-attentive multimodal SAN (GA-SAN), and region-attentive multimodal SAN (RA-SAN). Red and blue words indicate incorrect and correct, respectively.

Fig. 9. Examples for region-attentive multimodal RNN (RA-RNN) and region-attentive multimodal SAN (RA-SAN). Red and blue words indicate incorrect and correct, respectively.

and RA-SAN architectures. The first is an example of the case where RA-SAN is better than RA-RNN, whereas the second is the opposite case.

For the first example, it reflects that the performance of the image attention mechanism is also crucial to the translation quality of the proposed method. In another word, the semantic image region features and the effectiveness of the image attention mechanism are indispensable.

As shown in this example, the RA-SAN translates "equipment" better than RA-RNN. From the top five weighted semantic image regions, we can identify the semantic of "equipment," either in RA-RNN or RA-SAN. However, as the most weighted semantic image region by the image attention mechanism in RA-RNN does not provide any relevant semantic information to the text's semantics, it eventually leads to worse translation.

For the second example, it demonstrates that the improvement in translation performance benefits from attending to the specific semantic image region features.

As shown in this example, the RA-RNN translates "yard work" better than RA-SAN. We find that the RA-RNN focuses on a potted plant in a small garden from the most weighted semantic image region, however, the RA-SAN focuses on a boy's work activities. Moreover, we observe that the top five weighted semantic image regions on which the two architectures focus are quite different. The RA-RNN mainly focuses on the garden, whereas the RA-SAN focuses on the action.

6. Conclusion

This study proposed a multimodal NMT method, namely, RA-NMT, with semantic image regions. The proposed method was implemented on two types of NMT architectures. Experimental results showed that the proposed method achieved a significant improvement above its baselines on either of the two neural architectures. Furthermore, the proposed method implemented on the state-of-the-art NMT baselines can not only achieve better performance than the baselines but can also outperform most of the existing MNMT methods, which verifies its effectiveness and competitiveness. In addition, we investigated the effect of the number of semantic image region features and proved that the performance of the proposed method is not sensitive to the number of semantic image region features, which reflects the stability of the proposed method. Further analysis demonstrated that the proposed method effectively improves translation performance, and the improvement benefits from attending to specific semantic image region features, leading to better use of visual information.

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 the English entity and image region alignment have been manually annotated to the Multi30k dataset, we plan to use it as supervision to improve the performance of the attention mechanism.

CRediT authorship contribution statement

Yuting Zhao: Investigation, Conceptualization, Methodology, Data Curation, Software, Writing - Original Draft, Writing - Review & Editing. **Mamoru Komachi:** Methodology, Writing - Review \& Editing, Funding acquisition. **Tomoyuki Kajiwara:** Methodology, Writing - Review \& Editing. **Chenhui Chu:** Conceptualization, Methodology, Writing - Review \& Editing, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

References

[1] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L., 2018. Bottom-up and top-down attention for image captioning and visual question answering, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pp. 6077–6086. URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Bottom-Up_and_Top-Down_CVPR_2018_paper.html..

[2] Ba, L.J., Kiros, J.R., Hinton, G.E., 2016. Layer normalization. CoRR abs/1607.06450. URL: http://arxiv.org/abs/1607.06450, arXiv:1607.06450..

[3] Bahdanau, D., Cho, K., Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate. CoRR abs/1409.0473. URL: https://arxiv.org/abs/1409.0473v7..

[4] Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., Frank, S., 2018. Findings of the third shared task on multimodal machine translation, in: Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Jimeno-Yepes, A., Koehn, P., Monz, C., Negri, A., Neves, M.L., Post, M., Specia, L., Turchi, M., Verspoor, K. (Eds.), Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018, pp. 304–323. doi:10.18653/v1/w18-6402..

[5] Caglayan, O., Aransa, W., Bardet, A., García-Martínez, M., Bougares, F., Barrault, L., Masana, M., Herranz, L., van de Weijer, J., 2017a. LIUM-CVC submissions for WMT17 multimodal translation task, in: Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno-Yepes, A., Koehn, P., Kreutzer, J. (Eds.), Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7–8, 2017, pp. 432–439. doi:10.18653/v1/w17-4746..

[6] Caglayan, O., Aransa, W., Wang, Y., Masana, M., García-Martínez, M., Bougares, F., Barrault, L., van de Weijer, J., 2016a. Does multimodality help human and machine for translation and image captioning?, in: Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, Berlin, Germany, pp. 627–633. doi:10.18653/v1/w16-2358..

[7] Caglayan, O., Bardet, A., Bougares, F., Barrault, L., Wang, K., Masana, M., Herranz, L., van de Weijer, J., 2018. LIUM-CVC submissions for WMT18 multimodal translation task, in: Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Jimeno-Yepes, A., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M.L., Post, M., Specia, L., Turchi, M., Verspoor, K. (Eds.), Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018, pp. 597–602. doi:10.18653/v1/w18-6438..

[8] Caglayan, O., Barrault, L., Bougares, F., 2016b. Multimodal attention for neural machine translation. CoRR abs/1609.03976. URL: http://arxiv.org/abs/1609.03976, arXiv:1609.03976..

[9] Caglayan, O., García-Martínez, M., Bardet, A., Aransa, W., Bougares, F., Barrault, L., 2017b. NMTPY: A flexible toolkit for advanced neural machine translation systems. Prague Bull. Math. Linguistics 109, 15–28. URL: http://ufal.mff.cuni.cz/pbml/109/art-caglayan-et-al.pdf..

[10] Caglayan, O., Madhyastha, P., Specia, L., Barrault, L., 2019. Probing the need for visual context in multimodal machine translation, in: Burstein, J., Doran, C., Solorio, T. (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers), pp. 4159–4170. doi:10.18653/v1/n19-1422..

[11] Calixto, I., Elliott, D., Frank, S., 2016. DCU-UvA multimodal MT system report, in: Proceedings of the First Conference on Machine Translation, WMT 2016,

collocated with ACL 2016, August 11-12, Berlin, Germany, pp. 634–638. doi:10.18653/v1/w16-2359..

[12] Calixto, I., Liu, Q., 2017. Incorporating global visual features into attention-based neural machine translation, in: Palmer, M., Hwa, R., Riedel, S. (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9–11, 2017, pp. 992–1003. doi:10.18653/v1/d17-1105..

[13] Calixto, I., Liu, Q., Campbell, N., 2017. Doubly-attentive decoder for multi-modal neural machine translation, in: Barzilay, R., Kan, M. (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, pp. 1913–1924. doi:10.18653/v1/P17-1175..

[14] Calixto, I., Rios, M., Aziz, W., 2019. Latent variable model for multi-modal translation, in: Korhonen, A., Traum, D.R., Màrquez, L. (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pp. 6392–6405. doi:10.18653/v1/p19-1642..

[15] Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y., 2014a. On the properties of neural machine translation: Encoder-decoder approaches, in: Wu, D., Carpuat, M., Carreras, X., Vecchi, E.M. (Eds.), Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014, pp. 103–111. URL: https://www.aclweb.org/anthology/W14-4012/..

[16] Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014b. Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: Moschitti, A., Pang, B., Daelemans, W. (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 1724–1734. doi:10.3115/v1/d14-1179..

[17] Delbrouck, J., Dupont, S., 2017. An empirical study on the effectiveness of images in multimodal neural machine translation, in: Palmer, M., Hwa, R., Riedel, S. (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9–11, 2017, pp. 910–919. doi:10.18653/v1/d17-1095..

[18] Delbrouck, J., Dupont, S., Seddati, O., 2017. Visually grounded word embeddings and richer visual features for improving multimodal neural machine translation. CoRR abs/1707.01009. URL: http://arxiv.org/abs/1707.01009, arXiv:1707.01009..

[19] Denkowski, M.J., Lavie, A., 2014. Meteor universal: Language specific translation evaluation for any target language, in: Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26–27, 2014, Baltimore, Maryland, USA, pp. 376–380. doi:10.3115/v1/w14-3348..

[20] D. Elliott, Adversarial evaluation of multimodal machine translation, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, October 31 - November 4, 2018, Brussels, Belgium, 2018, pp. 2974–2978, https://doi.org/10.18653/v1/d18-1329.

[21] Elliott, D., Frank, S., Barrault, L., Bougares, F., Specia, L., 2017. Findings of the second shared task on multimodal machine translation and multilingual image description, in: Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno-Yepes, A., Koehn, P., Kreutzer, J. (Eds.), Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7–8, 2017, pp. 215–233. doi:10.18653/v1/w17-4718..

[22] Elliott, D., Frank, S., Hasler, E., 2015. Multi-language image description with neural sequence models. CoRR abs/1510.04709. URL: http://arxiv.org/abs/1510.04709, arXiv:1510.04709..

[23] Elliott, D., Frank, S., Sima'an, K., Specia, L., 2016. Multi30k: Multilingual english-german image descriptions, in: Proceedings of the 5th Workshop on Vision and Language, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, VL@ACL 2016, August 12, Berlin, Germany, pp. 70–74. doi:10.18653/v1/w16-3210..

[24] Elliott, D., Kádár, Á., 2017. Imagination improves multimodal translation, in: Kondrak, G., Watanabe, T. (Eds.), Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers, pp. 130–141. URL: https://www.aclweb.org/anthology/l17-1014/..

[25] A. Fukui, D.H. Park, D. Yang, A. Rohrbach, T. Darrell, M. Rohrbach, Multimodal compact bilinear pooling for visual question answering and visual grounding, in: J. Su, X. Carreras, K. Duh (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, November 1–4, 2016, Austin, Texas, USA, 2016, pp. 457–468, https://doi.org/10.18653/v1/d16-1044.

[26] Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., Xu, W., 2015. Are you talking to a machine? dataset and methods for multilingual image question, in: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, pp. 2296–2304. URL: https://proceedings.neurips.cc/paper/2015/file/fb508ef074ee78a0e58c68be06d8a2eb-Paper.pdf..

[27] Grönroos, S., Huet, B., Kurimo, M., Laaksonen, J., Mérialdo, B., Pham, P., Sjöberg, M., Sulubacak, U., Tiedemann, J., Troncy, R., Vázquez, R., 2018. The MeMAD submission to the WMT18 multimodal translation task, in: Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Jimeno-Yepes, A., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M.L., Post, M., Specia, L., Turchi, M., Verspoor, K. (Eds.), Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018,

Belgium, Brussels, October 31 - November 1, 2018, pp. 603–611. doi:10.18653/v1/w18-6439..

[28] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90..

[29] Helcl, J., Libovický, J., Varis, D., 2018. CUNI system for the WMT18 multimodal translation task, in: Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Jimeno-Yepes, A., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M.L., Post, M., Specia, L., Turchi, M., Verspoor, K. (Eds.), Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018, pp. 616–623. doi:10.18653/v1/w18-6441..

[30] Huang, P., Liu, F., Shiang, S., Oh, J., Dyer, C., 2016. Attention-based multimodal neural machine translation, in: Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany, pp. 639–645. doi:10.18653/v1/w16-2360..

[31] Ive, J., Madhyastha, P., Specia, L., 2019. Distilling translations with visual awareness, in: Korhonen, A., Traum, D.R., Màrquez, L. (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pp. 6525–6538. doi:10.18653/v1/p19-1653..

[32] Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization, in: Bengio, Y., LeCun, Y. (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings, pp. 1–15. URL: http://arxiv.org/abs/1412.6980..

[33] Klein, G., Kim, Y., Deng, Y., Senellart, J., Rush, A.M., 2017. OpenNMT: Open-source toolkit for neural machine translation, in: Bansal, M., Ji, H. (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, System Demonstrations, pp. 67–72. doi:10.18653/v1/P17-4012..

[34] Koehn, P., 2004. Statistical significance tests for machine translation evaluation, in: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25–26 July 2004, Barcelona, Spain, pp. 388–395. URL: https://www.aclweb.org/anthology/W04-3250/..

[35] 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., Herbst, E., 2007. Moses: Open source toolkit for statistical machine translation, in: Carroll, J.A., van den Bosch, A., Zaenen, A. (Eds.), ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23–30, 2007, Prague, Czech Republic, pp. 177–180. URL: https://www.aclweb.org/anthology/P07-2045/..

[36] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D.A. Shamma, M.S. Bernstein, L. Fei-Fei, Visual genome: Connecting language and vision using crowdsourced dense image annotations, Int. J. Comput. Vis. 123 (2017) 32–73, https://doi.org/10.1007/s11263-016-0981-7.

[37] Libovický, J., Helcl, J., 2017. Attention strategies for multi-source sequence-to-sequence learning, in: Barzilay, R., Kan, M. (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers, pp. 196–202. doi:10.18653/v1/P17-2031..

[38] Papineni, K., Roukos, S., Ward, T., Zhu, W., 2002. Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA, pp. 311–318. URL: https://www.aclweb.org/anthology/P02-1040/..

[39] Ren, S., He, K., Girshick, R.B., Sun, J., 2015. Faster R-CNN: towards real-time object detection with region proposal networks, in: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7–12, 2015, Montreal, Quebec, Canada, pp. 91–99. URL: https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf..

[40] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M.S. Bernstein, A.C. Berg, F. Li, Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (2015) 211–252, https://doi.org/10.1007/s11263-015-0816-y.

[41] Sennrich, R., Haddow, B., Birch, A., 2016. Neural machine translation of rare words with subword units, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7–12, 2016, Berlin, Germany, Volume 1: Long Papers, pp. 1715–1725. doi:10.18653/v1/p16-1162..

[42] Specia, L., Frank, S., Sima'an, K., Elliott, D., 2016. A shared task on multimodal machine translation and crosslingual image description, in: Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany, pp. 543–553. doi:10.18653/v1/w16-2346..

[43] Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks, in: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8–13 2014, Montreal, Quebec, Canada, pp. 3104–3112. URL: http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks..

[44] Toyama, J., Misono, M., Suzuki, M., Nakayama, K., Matsuo, Y., 2016. Neural machine translation with latent semantic of image and text. CoRR abs/1611.08459. URL: http://arxiv.org/abs/1611.08459, arXiv:1611.08459..

[45] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need, in: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pp. 5998–6008. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need..

[46] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention, in: Bach, F.R., Blei, D.M. (Eds.), Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, pp. 2048–2057. URL: http://proceedings.mlr.press/v37/xuc15.html..

[47] Yang, P., Chen, B., Zhang, P., Sun, X., 2020. Visual agreement regularized training for multi-modal machine translation, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7–12, 2020, pp. 9418–9425. URL: https://aaai.org/ojs/index.php/AAAI/article/view/6484..

[48] Yao, S., Wan, X., 2020. Multimodal transformer for multimodal machine translation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020, pp. 4346–4350. doi:10.18653/v1/2020.acl-main.400..

[49] Y. Yin, F. Meng, J. Su, C. Zhou, Z. Yang, J. Zhou, J. Luo, A novel graph-based multi-modal fusion encoder for neural machine translation, in, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, July 5–10, 2020, Online, 2020, pp. 3025–3035, https://doi.org/10.18653/v1/2020.acl-main.273.

[50] Zhang, Z., Chen, K., Wang, R., Utiyama, M., Sumita, E., Li, Z., Zhao, H., 2020. Neural machine translation with universal visual representation, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, pp. 1–11. URL: https://openreview.net/forum?id=Byl8hhNYPS..

[51] Zhao, Y., Komachi, M., Kajiwara, T., Chu, C., 2020. Double attention-based multimodal neural machine translation with semantic image regions, in: Forcada, M.L., Martins, A., Moniz, H., Turchi, M., Bisazza, A., Moorkens, J., Arenas, A.G., Nurminen, M., Marg, L., Fumega, S., Martins, B., Batista, F., Coheur, L., Escartín, C.P., Trancoso, I. (Eds.), Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020, Lisboa, Portugal, November 3-5, 2020, pp. 105–114. URL: https://www.aclweb.org/anthology/2020.eamt-1.12/..

[52] M. Zhou, R. Cheng, Y.J. Lee, Z. Yu, A visual attention grounding neural model for multimodal machine translation, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, October 31 - November 4, 2018, Brussels, Belgium, 2018, pp. 3643–3653, https://doi.org/10.18653/v1/d18-1400.

Yuting Zhao received her B.Eng. degree from Liaoning Technical University in 2014, and her M.Eng. degree from Tokyo Metropolitan University (TMU) in 2020. She is currently a Ph.D. Candidate at TMU. Her research interests include natural language processing, unsupervised learning and multimodal machine learning.

Mamoru Komachi is an associate professor at Tokyo Metropolitan University (TMU). He received his M.Eng. and Ph.D. degrees from the National Institute of Science and Technology (NAIST) in 2007 and 2010, respectively. He was an assistant professor at NAIST before joining the TMU. His research interests include semantics, information extraction, and the educational applications of natural language processing.
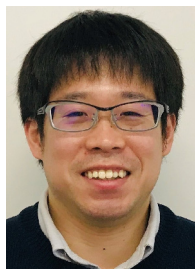
Tomoyuki Kajiwara is an assistant professor at Ehime University. He received his B.S. and M.S. degrees in engineering from Nagaoka University of Technology in 2013 and 2015, respectively, and his Ph.D. degree in engineering from Tokyo Metropolitan University in 2018. He was a specially-appointed assistant professor at Osaka University before joining Ehime University. His research interests include natural language processing, particularly paraphrasing and quality estimation.
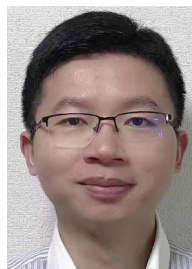
Chenhui Chu received his B.S. in software engineering from Chongqing University in 2008, and his M.S. and Ph. D. in Informatics from Kyoto University in 2012 and 2015, respectively. He is currently a program-specific associate professor at Kyoto University. His research interests include natural language processing, particularly machine translation and multimodal machine learning.