Information Sciences 554 (2021) 47–60

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Multi-modal neural machine translation with deep semantic interactions

Jinsong Su [a,b], Jinchang Chen [a], Hui Jiang [a], Chulun Zhou [a], Huan Lin [a], Yubin Ge [c], Qingqiang Wu [a,*], Yongxuan Lai [a,*]

[a] Xiamen University, Xiamen, China
[b] Peng Cheng Laboratory, Shenzhen, China
[c] University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

ARTICLE INFO

Article history:
Received 21 July 2018
Received in revised form 2 November 2020
Accepted 9 November 2020
Available online 28 November 2020

Keywords:
Multi-modal neural machine translation
Semantic interaction
Bi-directional attention
Co-attention

ABSTRACT

Based on the conventional attentional encoder-decoder framework, multi-modal neural machine translation (NMT) further incorporates spatial visual features through a separate visual attention mechanism. In this aspect, most current multi-modal NMT models first separately learn the semantic representations of text and image and then independently produce two modalities of context vectors for word predictions, neglecting their semantic interactions. In this paper, we argue that learning text-image semantic interactions is more reasonable in the sense of jointly modeling two modalities for multi-modal NMT and propose a novel multi-modal NMT model with deep semantic interactions. Specifically, our model extends the conventional multi-modal NMT by introducing the following two attention neural networks: (1) a bi-directional attention network for modeling text and image representations, where the semantic representations of text are learned by referring to the image representations, and vice versa; (2) a co-attention network for refining text and image context vectors, which first summarizes the text into a context vector, then attends it to the image for obtaining the text-aware visual context vector. The final context vector is calculated by re-attending the visual context vector to the text. Results on the Multi30k dataset for different language pairs show that our model significantly improves on the state-of-the-art baselines. We have released our code at https://github.com/DeepLearnXMU/MNMT.

1. Introduction

As an extension of the conventional text-based machine translation, multi-modal NMT exploits both visual and textual semantic information for translation. The studies of multi-modal machine translation have the following significances: First, to promote the integration of computer vision into natural language processing; Second, to push multi-modal language processing towards multi-lingual multi-modal language processing; Third, to investigate the effectiveness of image information on machine translation. Hence, multi-modal NMT has attracted extensive attention from the industry and become a research hotspot in academia as well [37,18,4].

* Corresponding authors.
E-mail addresses: jssu@xmu.edu.cn (J. Su), chenjinchang@stu.xmu.edu.cn (J. Chen), hjiang@stu.xmu.edu.cn (H. Jiang), clzhou@stu.xmu.edu.cn (C. Zhou), huanlin@stu.xmu.edu.cn (H. Lin), yubinge2@illinois.edu (Y. Ge), wuqq@xmu.edu.cn (Q. Wu), laiyx@xmu.edu.cn (Y. Lai).

https://doi.org/10.1016/j.ins.2020.11.024
0020-0255/

In this aspect, many models have been proposed to incorporate visual semantic information into NMT. For instance, Calixto et al. [8] proposed a doubly-attentive decoder which naturally incorporates spatial visual features via an independent attention mechanism. Moreover, Calixto et al. [9] investigated the effectiveness of incorporating visual features into different components of multi-modal NMT model. Additionally, Delbrouck and Dupont [15] employed Compact Bilinear Pooling to combine the attention features of the two modalities by computing the outer product of their corresponding context vectors. Besides, Delbrouck and Dupont [14] compared several attention mechanisms on the basis of the existing multi-modal NMT model, surpassing the state-of-the-art scores on the Multi30k data set. However, among the aforementioned works, there are two drawbacks:

(1) In terms of learning semantic representations, the previous models only utilized text information to refine the visual semantic representations, ignoring the strong semantic association between text and image. Intuitively, it is more reasonable to generate better representations via the semantic interactions between text and image representations.
(2) With respect to the generation of context vectors, the above-mentioned models adopted two separate attention mechanisms to generate text and image contexts, respectively. However, the semantic interactions between text and image contexts are apparently under-utilized in these models, although intuitively beneficial for multi-modal NMT.

It should be noted that designing an effective mechanism capturing multi-modal semantic interactions can significantly refine the representation learned in distinct modalities, which has been demonstrated effective in many tasks, such as VQA [32,44,22,42,31,41,52], NER [51], QA [43] etc. Therefore, we believe that exploiting multi-modal semantic interactions has potential to further improve multi-modal NMT.

In this paper, we propose a novel multi-modal NMT model with semantic interactions between text and image. Our proposed model is based on the conventional multi-modal NMT and enhanced by introducing the following two additional attention mechanisms:

(1) A bi-directional attention network for representation modeling, which learns enhanced text and image representations via modeling semantic interactions between them. Concretely, we introduce an attention matrix to capture fine-grained semantic alignments between text and image to generate improved representations, which incorporates more interaction information between modalities.
(2) A co-attention mechanism for context vector modeling, which summarizes multi-modal representations in a way that exploits the complementarity and redundancy between modalities. More specifically, the co-attention mechanism is implemented in the following manner: It first derives a context vector merely from previously generated text representations. Then, this context vector is used to guide the generation of image context vector. Finally, the image context vector is exploited in the subsequent re-attention process to obtain the updated text context vector that encodes more interaction information.

To summarize, the contributions of our work are as follows: (1) From the perspective of modeling, we investigate how to exploit semantic interactions between text and image to improve multi-modal NMT. (2) We propose a bi-directional attention network to obtain visual-aware text representations and text-aware image representations by modeling alignments between text and image. Meanwhile, we introduce a co-attention mechanism to refine the generation of context vectors for multi-modal NMT via jointly performing text-guided visual attention and image-guided text attention. (3) Experimental results and in-depth analyses on public datasets show that our model achieves significant improvements over the conventional multi-modal NMT models. (4) We release our code at https://github.com/DeepLearnXMU/MNMT.

2. Related work

Recently, multimodal deep learning has attracted increasing attention due to its wide applications, such as image clustering Li and Lu [29], disease detection Li et al. [28], image segmentation Fang et al. [21], and cross-modal retrieval Li et al. [27]. Particularly, as a significant extension of the conventional text-only NMT, multi-modal NMT has also become a hot research topic in the community of machine translation, and has been listed as a new shard WMT task at the intersection of natural language processing and computer vision [37,18,4]. Currently, the dominant models can be classified into the following categories:

As for the first kind, only textual attention mechanism is employed, where image information plays a complementary role in refining text representations. For example, Huang et al. [25] extended the traditional attentional NMT framework via integrating the semantic representation of image as an additional input into encoder. Calixto et al. [9] further explored how to initialize the decoder hidden states with the semantic representation of image. In a multitask learning framework, Elliott and Kádár [20] decomposed multimodal translation into learning a translation model and visually grounded representations. In this way, the multi-modal model can be trained over external datasets of parallel text or described images, making it possible to take advantage of existing resources. Qian et al. (2018) [34] put forward a novel algorithm on the basis of the Advantage Actor-Critic algorithm [2] to investigate the effectiveness of reinforcement learning in multi-modal NMT.

Unlike the first kind, approaches of the second category hold that both text and image information are vital in multi-modal NMT. Therefore, two attention mechanisms are simultaneously used to capture text and image contexts for translation, respectively. In this aspect, Caglayan et al. [6,7] first proposed an end-to-end attentional multi-modal NMT model effectively incorporating two attention mechanisms that share the same parameters for text and image. Furthermore, Calixto et al. [8] investigated the effectiveness of applying two independent attention mechanisms respectively on incorporating text and image information. Delbrouck et al. [16] made an empirical investigation on improving multi-modal NMT by using enhanced visual and word representations. Among these work, however, the text and image were assumed to be mutually independent. To utilize semantic interactions to refine the learned image semantics, Delbrouck and Dupont [15] applied a multi-modal Compact Bilinear Pooling operation to remove the noisy information of image representations according to the text representations. Recently, Yin et al. [46] proposed a graph-based multi-modal fusion encoder for NMT, which is based on a unified graph representing various semantic relationships between multi-modal semantic units. Lin et al. [30] introduced capsule network to better dynamically extract image features for translation. Yang et al. [45] jointly trained the source-to-target and target-to-source translation models, which are encouraged to share the same focus on the visual information when generating semantically equivalent visual words.

In this work, we mainly explore image-text semantic interactions for multi-modal NMT, which have been proven effective in many tasks, such as VQA [32,44,22,42,31,41,52], NER [51], QA [43] etc. Overall, our work differs from the above-mentioned works in the following two aspects:

First, our work is the first one to introduce image-text mutual interactions to refine their semantic representations, which is significantly different from previous works that only focus on image-to-text one-way operation.

Second, we explore not only the semantic interactions between text and image representations but also that of their context vectors, which is also very crucial for multi-modal NMT. More precisely, the context vector, generated from text representations, is attended to the image representations to build a second layer of context vector, which is used to implement attention on the text representations again.

3. Background

In this section, we give a detailed description of the multi-modal NMT model [8], which is an extension of the attention-based NMT [3] with the addition of a separate visual attention mechanism to incorporate image features, as shown in Fig. 1. Given an image I, a source sentence $X = (x_1, x_2, \cdots, x_N)$ that describes the image and its corresponding translation $Y = (y_1, y_2, \cdots, y_M)$, the multi-modal NMT aims to construct an end-to-end neural network to model $P = (Y|X, I)$. Formally, the multi-modal NMT model is composed of one text encoder, one visual encoder and one decoder with two attention mechanisms. The text encoder and visual encoder learn a text representation C and a visual representation A for the source sentence and image, respectively, and then the doubly-attentive decoder sequentially generates target words by decomposing the following conditional probability:

$$logp(Y|X, I) = \sum_{i=1}^{M} logp(y_t|y_{<t}, C, A). \tag{1}$$

As for the specific implementation of encoder, a bi-directional Recurrent Neural Network (RNN) with Gated Recurrent Unit (GRU) [11] is constructed, where a forward RNN $\vec{\Phi}_{enc}$ reads the source sentence word by word, from left to right, and produces a sequence of forward hidden states $\left(\vec{h}_1, \vec{h}_2, \cdots, \vec{h}_N\right)$. Similarly, a backward RNN $\overleftarrow{\Phi}_{enc}$ scans the source sentence reversely and generates a sequence of backward hidden states $\left(\overleftarrow{h}_1, \overleftarrow{h}_2, \cdots, \overleftarrow{h}_N\right)$, as shown in Eq. (2) and Eq. (3):

$$\vec{h}_i = \vec{\Phi}_{enc}\left(E_x[x_i], \vec{h}_{i-1}\right), \tag{2}$$

$$\overleftarrow{h}_i = \overleftarrow{\Phi}_{enc}\left(E_x[x_i], \overleftarrow{h}_{i-1}\right), \tag{3}$$

where $\vec{\Phi}_{enc}$ and $\overleftarrow{\Phi}_{enc}$ are the GRU activation functions in two directions, $E_x[x_i]$ denotes the embedding of source word x_i. The final hidden state at a given timestep is the concatenation of forward and backward hidden states $h_i = \left[\vec{h}_i; \overleftarrow{h}_i\right]$. By doing so, we can use the set of hidden states $C = (h_1, h_2, \cdots, h_N)$ to represent the input sentence. The visual encoder is a pre-trained convolutional neural network (CNN), of which parameters are fixed during training. Specifically, we employ a 50-layer Residual Network (ResNet-50) [23] to represent the visual semantic information as a matrix $A = (a_1, a_2, \ldots, a_{196}$, where $a_i \in \mathbb{R}^{1024})$ and each row is composed of a 1024D feature vector encoding the information of a specific image region. The decoder is a conditional GRU[1] (cGRU) with two separate attention mechanisms: one is text attention mechanism and the other is visual

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

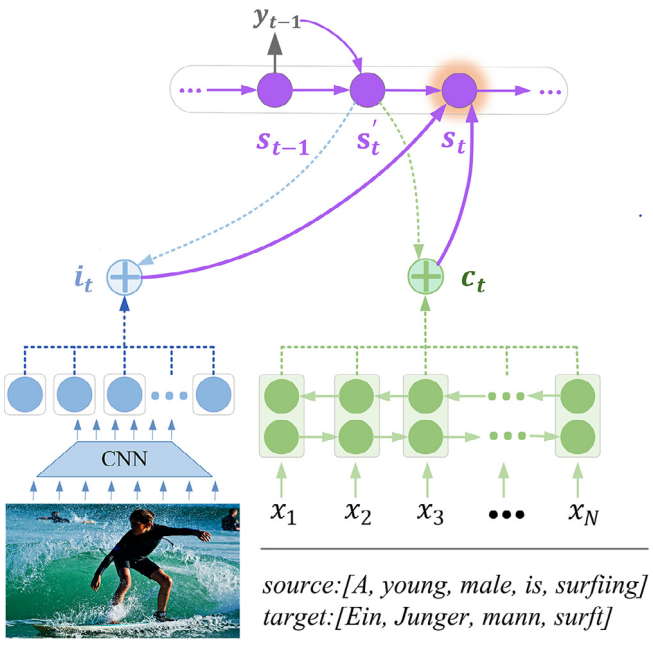

Fig. 1. Conventional multi-modal NMT model.

attention mechanism. In details, the cGRU consists of two stacked GRU activations called REC_1 and REC_2. To generate the target word y_t at the time step t, REC_1 firstly computes a hidden state proposal s'_t given the previous hidden state s_{t-1} and the previously generated target word y_{t-1} as follows:

$$s'_t = (1 - z'_t) \odot \underline{s}'_t + z'_t \odot s_{t-1}, \tag{4}$$
$$\underline{s}'_t = \tanh(W' E_y[y_{t-1}] + r'_t \odot (U' s_{t-1})), \tag{5}$$
$$r'_t = \sigma(W'_r E_y[y_{t-1}] + U'_r s_{t-1}), \tag{6}$$
$$z'_t = \sigma(W'_z E_y[y_{t-1}] + U'_z s_{t-1}), \tag{7}$$

where $E_y[y_i]$ is the embedding vector of the target word y_i. In this process, the text attention generates a time-dependent context vector c_t based on the hidden state proposal s'_t and a sequence of hidden states C in the following way:

$$c_t = f_{att\_text}(C, s'_t). \tag{8}$$

Meanwhile, the visual attention computes a time-dependent context vector i_t based on the hidden state proposal s'_t and the visual feature maps A as follows:

$$i_t = f_{att\_img}(A, s'_t). \tag{9}$$

Then, the second GRU activation REC_2 generates a new hidden state s_t using the temporary hidden state proposal s'_t, c_t and i_t:

$$s_t = (1 - z_t) \odot \underline{s}_t + z_t \odot s'_t, \tag{10}$$
$$\underline{s}_t = \tanh(W c_t + W i_t + r_t \odot (U s'_t)), \tag{11}$$
$$r_t = \sigma(W_r c_t + W_r i_t + U_r s'_t), \tag{12}$$
$$z_t = \sigma(W_z c_t + W_z i_t + U_z s'_t). \tag{13}$$

Finally, the probabilities for the next target word are calculated on the basis of the previous hidden state s_t, the previously decoded symbol y_{t-1}, c_t and i_t:

$$p(y_t | y_{<t}, C, A) \propto \exp(L_o \tanh(L_s s_t + L_w E_y[y_{t-1}] + L_{cs} c_t + L_{ci} i_t)), \tag{14}$$

where $L_o, L_s, L_w, L_{cs}, L_{ci}$ are the related model parameters.

4. The proposed model

As shown in Fig. 2, our model is an extension of the conventional multi-modal NMT model, which is equipped with a bi-directional attention network and a co-attention network to model underlying semantic interactions between text and image.

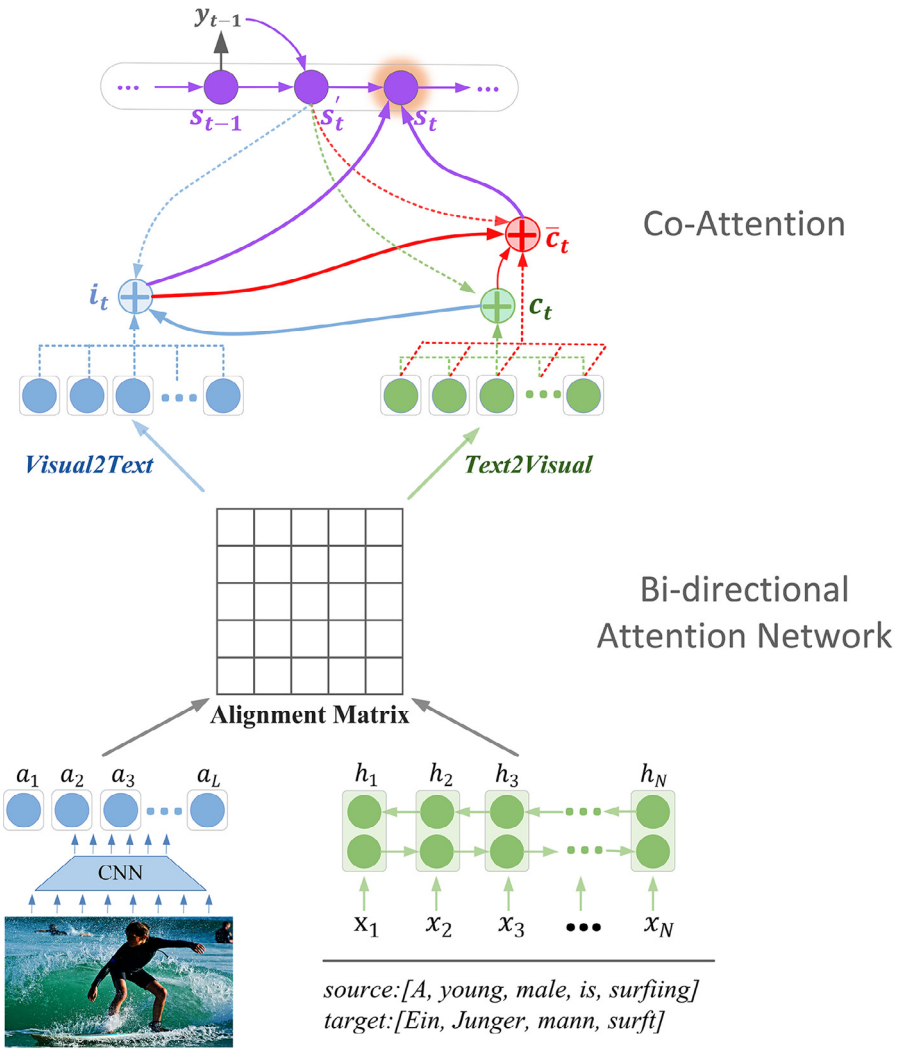

Fig. 2. The architecture illustration of our model.

4.1. Encoder with bi-directional attention network

As illustrated in the lower part of Fig. 2, we propose a bi-directional attention network to enhance text and image representations by capturing their semantic interactions. To do this, we first implement attentions in two directions: from text to image as well as from image to text, where a shared alignment matrix $S \in \mathbb{R}^{N \times L}$ between the text hidden states $C = (h_1, h_2, \ldots, h_N)$ and the image feature maps $A = (a_1, a_2, \cdots, a_L)$ are derived for bi-directional attentions. In this way, each source word is able to semantically interact with all image regions and vice versa. Formally, the alignment matrix is computed as follows: [2]

$$S_{i,l} = g(h_i \cdot a_l), \tag{15}$$

where $g(*)$ is a scalar function and $S_{i,l} \in \mathbb{R}$ measures how well the i-th row vector in C semantically matches the l-th row vector in A.

[2] In fact, there are many ways can be used to model the bi-directional attention network. In this work, we follow common practices [48,13,43] to directly apply a scalar function to calculate the shared similarity matrix S. Please note that the dimension of S is $N \times L$, and thus it is not a symmetric matrix. More importantly, although S is shared between text-visual and visual-text representations, it can induce two different text-to-visual and visual-to-text attention weight matrices, which are normalized by rows and columns, respectively (See Eqs. (16) and (18)). In this way, we not only obtain text-visual and visual-text attention weight matrices quickly, but also avoid the over-fitting caused by additional matrix parameters. Our experimental results demonstrate the effectiveness of our such approach for text-visual and visual-text similarity calculation.

Text-to-visual Attention. It signifies which image regions are most relevant to each source word. We use the vector $w_i^{t2v} \in \mathbb{R}^L$ to denote attention weights on the image regions, where $\sum w_{il}^{t2v} = 1$ for all i and w_i^{t2v} is computed in the following way:

$$w_i^{t2v} = softmax(S_{i:}). \tag{16}$$

Subsequently, each attended text vector $\bar{h}_i$ is calculated as

$$\bar{h}_i = h_i + \sum_l w_{il}^{t2v} a_l. \tag{17}$$

Thus, we can obtain a matrix $\overline{H} = (\bar{h}_1, \bar{h}_2, \ldots, \bar{h}_N)$ containing the vectors for the whole source sentence. .

Visual-to-text Attention. This attention denotes which source words semantically match each visual region mostly. Let $w_l^{v2t} \in \mathbb{R}^N$ represents the attention weight of the l-th visual region acting on the source words, where $\sum w_{li}^{v2t} = 1$ for all l and we obtain w_l^{v2t} by

$$w_l^{v2t} = softmax(S_{:l}). \tag{18}$$

Thus, every attended visual vector $\bar{a}_l$ is induced as follows:

$$\bar{a}_l = a_l + \sum_i w_{li}^{v2t} h_i, \tag{19}$$

constituting a matrix $\overline{A} = (\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_L)$ that includes the attended text vectors for all visual regions.

4.2. Decoder with co-attention mechanism

The upper half of Fig. 2 shows the architecture of decoder with co-attention network. Different from doubly-attentive decoder described in Section 3, which learns to attend to visual feature maps and source sentence separately, our decoder is enhanced by a co-attention mechanism implemented in the following three phases: (1) It first generates a text context vector c_t by attending to the text hidden states; (2) It applies c_t to attend to image feature maps for producing a text-aware visual context vector i_t; (3) It forms the final context vector through attending to text hidden states on the basis of previously derived context vector c_t and i_t.

Specifically, we employ a single-layer feed-forward network to compute an expected alignment $e_{t,i}$ between text hidden state $\bar{h}_i$ and target word y_t:

$$e_{t,i} = (v_a)^T \tanh\left(U_a s'_t + W_a \bar{h}_i\right), \tag{20}$$

where $e_{t,i}$ quantifies the importance of the i-th source word in generating y_t, and v_a, U_a and W_a are the related model parameters. We then normalize the alignment scores and define the text context vector c_t as the weighted sum over text hidden states:

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{i'=1}^{N} \exp(e_{t,i'})}, \tag{21}$$

$$c_t = \sum_{i=1}^{N} \alpha_{t,i} \bar{h}_i, \tag{22}$$

where $\alpha_{t,i}$ is the normalized alignment weight.

At the second phase, we obtain the time-dependent visual context vector i_t based on the visual feature maps $\overline{A}$, the context vector c_t and the hidden state proposal s'_t in the following way:

$$i_t = \sum_{l=1}^{L} \alpha_{t,l}^{img} \bar{a}_l, \tag{23}$$

$$\alpha_{t,l}^{img} = \frac{\exp\left(e_{t,l}^{img}\right)}{\sum_{l'=1}^{L} \exp\left(e_{t,l'}^{img}\right)}, \tag{24}$$

$$e_{t,l}^{img} = \left(v_a^{img}\right)^T \tanh\left(U_a^{img} s'_t + V_a^{img} \bar{a}_l + W_a^{img} c_t\right), \tag{25}$$

where $e_{t,l}^{img}$ evaluates how much attention should be put on the l-th visual region at the t-th time step, $\alpha_{t,l}^{img}$ is the normalized attention weight, and $v_a^{img}, U_a^{img}, V_a^{img}$ and W_a^{img} are the related model parameters. Note that unlike conventional multi-modal NMT, we take the text context vector c_t as an additional input when computing the visual context vector i_t.

Next, in a similar way, we use both c_t and i_t as additional inputs to re-attend to text hidden states, resulting in another text context vector $\bar{c}_t$:

$$\bar{c}_t = \sum_{i=1}^{N} \bar{\alpha}_{t,i} \bar{h}_i, \tag{26}$$

$$\bar{\alpha}_{t,i} = \frac{\exp(\bar{e}_{t,i})}{\sum_{i'=1}^{N} \exp(\bar{e}_{t,i'})}, \tag{27}$$

$$\bar{e}_{t,i} = (\bar{v}_a)_T \tanh\left(\overline{U}_a s'_t + \overline{V}_a \bar{h}_i + \overline{W}_a c_t + Q_a i_t\right), \tag{28}$$

where $\bar{e}_{t,i}$ is an unnormalized attention weight, $\bar{\alpha}_{t,i}$ is the normalized weight, and $\bar{v}_a, \overline{U}_a, \overline{V}_a, \overline{W}_a$ and Q_a are the related parameters.

Finally, our decoder employs a modified version of REC_2 to generate the final hidden state on the basis of the hidden state proposal s'_t and all previously derived context vectors c_t, i_t and $\bar{c}_t$:

$$s_t = (1 - z_t) \odot \bar{s}_t + z_t \odot s'_t, \tag{29}$$

$$\bar{s}_t = \tanh\left(Wc_t + Wi_t + W\bar{c}_t + r_t \odot (Us'_t)\right), \tag{30}$$

$$r_t = \sigma\left(W_r c_t + W_r i_t + W_r \bar{c}_t + U_r s'_t\right), \tag{31}$$

$$z_t = \sigma\left(W_z c_t + W_z i_t + W_z \bar{c}_t + U_z s'_t\right). \tag{32}$$

We then calculate the probability distribution of the next target word as follows:

$$p(y_t | y_{<t}, C, A) \propto \exp\left(L_o \tanh\left(L_s s_t E_y[y_{t-1}] + L_{cs} c_t + L_{ci} i_t + L_{c\bar{s}} \bar{c}_t\right)\right), \tag{33}$$

where $L_{c\bar{s}}$ is a new vector parameter for $\bar{c}_t$.

5. Experiment

5.1. Datasets

To investigate the effectiveness of the proposed model, we carried out two groups of experiments on the translated and the comparable Multi30k dataset [3] [19] respectively. We referred to these two datasets as M30K_T and M30K_C, both of which are expansions of the Flick30K [47] and have been widely used in the community of multi-modal NMT. Each instance in M30K_T is a triple consisting of an image, an English description and a German description manually translated by a professional translator. As for M30K_C, it consists of a set of triples and each of them includes an image, five descriptions in English and five descriptions in German collected independently. In two groups of experiments, training, validation and test sets contain 29 K, 1,014 and 1 K triples, respectively.

The entire training set in M30k_T was used for training while its validation set served for model selection based on BLEU score [33]. After that, we used M30K_T test set to evaluate the models. Following Calixto et al. [8], we trained the DL4NMT [3] on the M30K_T dataset without images to build a back-translation model. Then, we employed the model to back-translate all English(German) descriptions in M30K_C into German(English). Finally, we used the triples (synthetic English description, German description, image) to pre-train the model when translating English into German, and the triples (synthetic German description, English description, image) when translating German into English.

We utilized the *MOSES* scripts[4] to normalize, tokenize and lowercase both English and German descriptions. Then, following Calixto et al. [8], we preprocessed our data by converting words to subword tokens using a bilingual *Byte Pair Encoding* (BPE) [35] model with 10 K merge operations. In this way, the English and German vocabularies consist of a total of 5,201 and 7,066 subwords respectively.

5.2. Setup

We implemented our proposed model and its variants on the top of $\text{MNMT}_{\text{SRC+IMG}}$ [8] [5], which was developed based on OpenNMT [26]. The encoder is a bi-directional RNN with GRU (one 256D single-layer forward RNN and one 256D single-layer backward RNN). We set both embedding sizes of source and target words to 128 and randomly initialized the word

[3] https://github.com/multi30k/dataset
[4] http://www.statmt.org/moses/
[5] https://github.com/iacercalixto/MultimodalNMT

embeddings and other related model parameters according to a uniform distribution ($min = -0.1$ and $max = 0.1$). We used a modified cGRU as our decoder, which is a neural language model [5] conditioned on the previously generated target word, the previous hidden state, the source sentence and the image via a co-attention mechanism. To accurately represent the visual semantic information, we extracted visual features from the `res4f` layer of pre-trained ResNet-50 [23] and transformed them from 1024D to 256D to make them consistent with the text. Please note that the visual features are fixed during training. Besides, we also adopted dropout strategy [38] with rate 0.3 to further enhance our models.

As implemented in [7], we used *Adam* optimizer with mini-batches size of 32 to train all models. Particularly, we applied *early stopping* to determine the optimal model parameters based on the system performance on the validation set, where the training procedure was stopped if the system performance remains unchanged for more than 20 epochs.

For the evaluation of the generated translations, we adopted three commonly used metrics: BLEU4 [33], METEOR [17] and TER [36]. All these metrics were computed using scripts in MultEval [12].

Baseline models. To empirically verify the merit of our proposed model, we referred to our model as MNMT and compared it with the following state-of-the-art methods, namely:

- *Parallel RCNN*$_s$ [25]: It employs an encoder consisting of multiple encoding threads where all the long short-term memory [24] parameters are shared. In each thread, a regional visual feature is followed by the text sequence.
- *MNMT*$_{SRC+IMG}$ [8]: It uses a doubly-attentive decoder to exploit visual feature maps and text hidden states independently when generating translations. Please note that we developed our proposed model based on its released code.
- *IMG*$_D$ [9]: This model utilizes image features as additional input to initialize the first hidden state of the decoder.
- *IMG*$_{E+D}$ [9]: The first hidden states of both encoder and decoder are initialized with visual features extracted from the images.
- *Soft-Attention* [14]: An additional spatial attention on visual feature maps is incorporated into the attention-based NMT framework by taking all representations into account when generating context vectors.
- *Hard-Attention* [14]: This model employs two separate attention mechanisms, one attending to all text representations and the other considering only one image feature at each time step, to generate image and context vectors.
- *MNMT*$_{\_CO\_ATT\_BI\_ATT}$: As a variant of our model, we removed both the bi-directional neural network and the co-attention mechanism from MNMT. In this way, this variant degenerates to MNMT$_{\text{SRC+IMG}}$ [8], which uses a doubly-attentive decoder to exploit visual feature maps and text hidden states independently when generating translations. Please note that it is our most important baseline.
- *MNMT-Big*$_{\_CO\_ATT\_BI\_ATT}$: It is a variant of MNMT$_{_\text{CO_ATT_BI_ATT}}$, where the size of RNN hidden state is increased from 256 to 280. Compared with MNMT$_{_\text{CO_ATT_BI_ATT}}$, it increases the parameter number by 9%, which is slightly more than that of our model. Through this comparative experiment, we aim to fully analyze the impact of the additional parameters on our model.

To further verify the effectiveness of different components in our model, we also provided the performance of the following ablated versions of our model:

- *MNMT*$_{\_CO\_ATT}$: The model only introduces the bi-directional neural network to refine both image and text representations.
- *MNMT*$_{\_BI\_ATT}$: This model only exploits the co-attention mechanism to improve the modeling of both image and text context vectors.

5.3. Experimental results

As shown in Table 1, for English-to-German multi-modal translation task, the lower and upper halves show the performance of the models with and without pre-training, respectively. Using almost the same hyper-parameters, MNMT$_{\text{CO_ATT_Bi_ATT}}$ exhibits comparable performance to MNMT$_{\text{SRC+IMG}}$, demonstrating that our re-implemented baseline is competitive in performance. In the upper part of Table 1, the results across three evaluation metrics consistently show that our proposed model and its variants MNMT$_{_\text{CO_ATT}}$ and MNMT$_{_\text{BI_ATT}}$ improve translation quality by large margins comparing to other models. Specifically, MNMT$_{_\text{CO_ATT}}$ obtains the relative improvements over the baseline models in all metrics. The result suggests the effectiveness of employing the bi-directional attention network to refine both text and image representations. Meanwhile, MNMT$_{_\text{BI_ATT}}$ also outperforms the baseline models according to all metrics, demonstrating that the semantic interactions between text and image context vectors also contribute to the translation improvement. Finally, our MNMT achieves the highest performance on all three metrics, 39.2, 56.6 and 40.5, respectively. This result indicates that text and image are semantically complementary and thus they are able to act as mutual reinforcement for multi-modal NMT. When pre-trained with additional back-translated data, our proposed model and its variants still outperform their corresponding baselines. The results indicate that proposed model and its variants could effectively exploit the back-translated data.

The upper part of Table 2 provides the experimental results of the models without pre-training on the German to English multi-modal translation task and the lower part gives the results of the models that are pre-trained on back-translated M30k$_C$. Similar to the previous experimental results, our model also achieve better performance compared to all other base-

Table 1
Experimental results on English to German multi-modal translation task. † indicates previously reported scores. Here we reported the mean and standard deviation over 4 independent runs.

Model	BLEU↑	METEOR↑	TER↓
Text-only NMT [8,9] †	33.70	52.3	46.7
Text-only NMT [14] †	34.11	52.4	46.2
Parallel RCNNs†	36.50	54.1	-
$MNMT^{\dagger}_{SRC+IMG}$	36.50	55.0	43.7
$IMG^{\dagger}_{D}$	37.30	55.1	42.8
Soft-Attention†	37.62	55.3	41.8
Hard-Attention†	38.17	55.4	41.5
$MNMT_{-CO\_ATT-BI\_ATT}$	37.0(0.3)	54.9(0.5)	42.7(0.2)
$MNMT\text{-}Big_{-CO\_ATT-BI\_ATT}$	37.7(0.2)	55.3(0.3)	42.4(0.5)
$MNMT_{-CO\_ATT}$	38.5(0.5)	56.0(0.5)	41.1(0.4)
$MNMT_{-BI\_ATT}$	38.3(0.2)	55.8(0.3)	41.3(0.5)
MNMT	**39.2(0.3)**	**56.6(0.3)**	**40.5(0.3)**
Pre-training dataset: back-translated $M30k_C$			
Text-only NMT [8,9]†	35.50	53.4	43.3
$IMG^{\dagger}_{D}$	38.50	55.9	41.6
$MNMT^{\dagger}_{SRC+IMG}$	37.10	54.5	42.8
$MNMT_{-CO\_ATT-BI\_ATT}$	37.4(0.4)	55.3(0.2)	42.0(0.2)
$MNMT\text{-}Big_{-CO\_ATT-BI\_ATT}$	38.2(0.5)	55.7(0.4)	41.0(0.1)
$MNMT_{-CO\_ATT}$	39.6(0.4)	57.0(0.3)	40.6(0.1)
$MNMT_{-BI\_ATT}$	40.1(0.4)	57.4(0.3)	**39.8(0.4)**
MNMT	**40.3(0.3)**	**57.5(0.2)**	39.9(0.3)

Table 2
Experimental results on German to English multi-modal translation task. † indicates previously reported scores. Here we reported the mean and standard deviation over 4 independent runs.

Model	BLEU↑	METEOR↑	TER↓
Text-only NMT [8,9]†	38.20	35.8	40.2
$MNMT^{\dagger}_{SRC+IMG}$	40.60	37.5	37.7
$IMG^{\dagger}_{E+D}$	41.90	37.9	37.1
$MNMT_{-CO\_ATT-BI\_ATT}$	40.6(0.5)	37.4(0.2)	37.9(0.4)
$MNMT\text{-}Big_{-CO\_ATT-BI\_ATT}$	40.7(0.4)	37.5(0.2)	37.9(0.2)
$MNMT_{-CO\_ATT}$	41.2(0.3)	38.1(0.3)	37.4(0.3)
$MNMT_{-BI\_ATT}$	42.2(0.5)	38.4(0.2)	36.5(0.4)
MNMT	**42.3(0.3)**	**38.6(0.2)**	**36.4(0.4)**
Pre-training dataset: back-translated $M30k_C$			
Text-only NMT [8,9]†	42.6	38.9	36.1
$MNMT^{\dagger}_{SRC+IMG}$	43.2	39.0	35.5
$MNMT_{-CO\_ATT-BI\_ATT}$	43.1(0.3)	38.9(0.3)	35.5(0.2)
$MNMT\text{-}Big_{-CO\_ATT-BI\_ATT}$	43.2(0.2)	39.0(0.1)	35.4(0.2)
$MNMT_{-CO\_ATT}$	43.2(0.3)	39.0(0.2)	35.4(0.4)
$MNMT_{-BI\_ATT}$	43.5(0.2)	39.2(0.3)	**35.1(0.4)**
MNMT	**43.7(0.4)**	**39.3(0.3)**	35.3(0.2)

line models. Furthermore, we note that $MNMT_{-BI\_ATT}$ exhibits better performance than $MNMT_{-CO\_ATT}$ and is comparable to MNMT. This indicates that high-level text-image semantic interactions can achieve a comparable or even better performance than the model with fine-grained text-image semantic interactions, when translating source language into an "easier" target language, i.e. the language with less morphology. When pre-trained on back-translated $M30k_C$, both $MNMT_{-BI\_ATT}$ and MNMT significantly improve over the baseline models while $MNMT_{-CO\_ATT}$ is only comparable to the baselines. The results suggest that although multi-modal model could benefit from additional back-translated data, introducing text-image semantic interactions can further improve translation quality.

5.4. Effect of pre-trained shared bilingual word embeddings

In this group of experiments, we investigated the impact of a pre-trained shared bilingual word embeddings. The large-scale training corpus we used is the commonly-used WMT 2015 English-German parallel corpus, which consists of about 4.46 M sentence pairs with 116.1 M English words and 108.9 M German words. According to the BPE vocabularies extracted from our Multi30k dataset, we first converted the parallel sentences of the WMT corpus into subword sequences, and then followed Artetxe et al. [1] to obtain the shared bilingual subword embeddings. Afterwards, we used these subword embeddings to initialize our model, which was finally fine-tuned using $M30K_T$.

Table 3 reports the experimental results. We can observe that using pre-trained WMT bilingual BPE embeddings leads to performance declines on our model and its two variants, respectively. The underlying reason is that the domain (*News* topic) of the WMT corpus is significantly different from that of our Multi30k dataset. The only exception is MNMT-Big$_{-\text{CO_ATT}-\text{BI_ATT}}$, of which performance is slightly improved. For this result, we speculate that its larger parameter dimension enables the model to make better use of the pre-trained ebmeddings.

5.5. Effect of pre-training our decoder

In this group of experiments, we investigated the effect of another pre-training strategy on our model. Specifically, we first pre-trained the decoder of our model using the German sentences of commonly-used WMT English-German parallel corpus, and then fine-tuned our whole model using M30K$_{\text{T}}$. It should be noted that both MNMT$_{-\text{CO_ATT}}$ and MNMT are not suitable for pre-training using large parallel text-only corpora, since their bi-attention mechanisms involve the utilizations of images. Thus, we conducted experiments using MNMT$_{-\text{BI_ATT}}$.

From Table 4, we can observe that when directly using the pre-trained decoder parameters, the performance of MNMT$_{-\text{BI_ATT}}$ significantly drops. Even if we fine-tune the decoder parameters on our Multi30k dataset, the performance of MNMT$_{-\text{BI_ATT}}$ is still slightly reduced due to the domain conflict between the pre-trained WMT corpus and the training corpus.

5.6. Case study

In order to know how our model improves the performance of the multi-modal NMT, we compared the translations produced by different models. Table 5 shows an English-German multi-modal example from the M30k$_{\text{T}}$ test set. In this case, we notice that MNMT$_{-\text{CO_ATT}-\text{BI_ATT}}$ misses the key word "*claps*" in the source sentence. Meanwhile, both Soft-Attention and Hard-Attention suffer from the same problem even if they exploit two separate attention mechanisms to capture related text and image contexts. However, it is remarkable that our proposed model and its variants produce correct translations. This result proves the effectiveness of introducing text-image semantic interactions and reveals the weakness of traditional multi-modal NMT model, where the processes of modeling text and image semantics are independent of each other.

Table 6 shows a German-English multi-modal translation example from the M30k$_{\text{T}}$ test set. We notice that multi-modal NMT models (IMG$_{\text{E+D}}$, MNMT$_{-\text{CO_ATT}-\text{BI_ATT}}$ and MNMT$_{-\text{CO_ATT}}$) are unable to correctly translate the German words "*trinkt shots*" (drinking shots) into English. In contrast, the translations generated by MNMT$_{-\text{BI_ATT}}$ and MNMT, although not novel, are still correct. The underlying reason may be that: co-attention mechanism aims at dealing with high level text-image semantic interactions that have incorporated global information while bi-attention mechanism is used to cope with local fine-grained text-image semantic interactions in a relatively lower level. Considering that more global information is needed during the translation of ambiguous words, co-attention mechanism performs better at discriminating ambiguous semantics of words than bi-attention does.

To further verity the merit of our model, in Fig. 3, we visualized the visual and textual attention weights when generating the target word "*klatscht*". We observe that during the first attention operation, our model fails to focus on the corresponding source word "*claps*" and instead places greater weights on the "*riding*", "*on*" and ".". Then, our model assigns greater weights on regions corresponding to the action "*claps*" when generating the visual context vector. Finally, our model performs text-image semantic interactions and then reconsiders the source sentence. It is worth noticing that the source word "*claps*" obtains the highest weight at this time. This result demonstrates that introducing text-image semantic interactions is able to enhance the capability of attention mechanism and further improve the translation quality.

5.7. Experimental results on english to French, English to Czech Multi-modal Translation Tasks

To further investigate the effectiveness and generality of our model, we carry out experiments on English to French, and English to Czech multi-modal translation tasks. We also chose Multi30k datasets as our experimental datasets. Besides, we used the same settings as our previous English-German experiments. In this way, the sizes of our French and Czech vocabularies are 9,426 and 6,574, respectively. However, since the Multi30k datasets do not contain the extra corpus as English-German M30k$_{\text{C}}$, we are unable to investigate the performance of various models pre-trained using back-translated pseudo parallel sentences.

Tables 7,8 provide the experimental results on the two translation tasks. Similar to previous experiment results, our model and its variants significantly outperform both MNMT$_{-\text{CO_ATT}-\text{BI_ATT}}$ and MNMT-Big$_{-\text{CO_ATT}-\text{BI_ATT}}$ on almost all test sets. Therefore, we confirm the validity and generality of our model in different language pairs.

6. Conclusion and future work

In this paper, we have successfully extended the conventional multi-modal NMT to a novel model with cross-modal semantic interaction modeling, which introduces a bi-directional attention for refining text and image semantic representations and a co-attention for enhancing context vectors. Experimental results on English-German and German-English trans-

Table 3
Experimental results on English to German multi-modal translation task with large-scale pre-trained bilingual word embeddings. Similarly, we listed the mean and standard deviation over 4 independent runs.

Model	BLEU↑	METEOR↑	TER↓
MNMT$_{-\text{CO_ATT}-\text{BI_ATT}}$	37.0(0.3)	54.9(0.5)	42.7(0.2)
MNMT-Big$_{-\text{CO_ATT}-\text{BI_ATT}}$	37.7(0.2)	55.3(0.3)	42.4(0.5)
MNMT$_{-\text{CO_ATT}}$	38.5(0.5)	56.0(0.5)	41.1(0.4)
MNMT$_{-\text{BI_ATT}}$	38.3(0.2)	55.8(0.3)	41.3(0.5)
MNMT	39.2(0.3)	56.6(0.3)	40.5(0.3)
Using pre-trained WMT 2015 shared bilingual word embeddings			
MNMT$_{-\text{CO_ATT}-\text{BI_ATT}}$	36.3(0.2)	54.7(0.2)	43.4(0.1)
MNMT-Big$_{-\text{CO_ATT}-\text{BI_ATT}}$	37.9(0.1)	55.5(0.2)	42.2(0.1)
MNMT$_{-\text{CO_ATT}}$	37.6(0.3)	55.4(0.1)	42.3(0.2)
MNMT$_{-\text{BI_ATT}}$	38.1(0.3)	55.8(0.1)	42.0(0.1)
MNMT	38.8(0.1)	56.3(0.1)	41.3(0.3)

Table 4
Experimental results on English to German multi-modal translation task with a pre-trained decoder. Likewise, we reported the mean and standard deviation over 4 independent runs.

Model	BLEU↑	METEOR↑	TER↓
MNMT$_{-\text{CO_ATT}-\text{BI_ATT}}$	37.0(0.3)	54.9(0.5)	42.7(0.2)
MNMT$_{-\text{BI_ATT}}$	38.3(0.2)	55.8(0.3)	41.3(0.5)
pre-trained decoder w/o fine-tune	22.5(0.5)	24.4(0.9)	56.4(0.7)
pre-trained decoder w/ fine-tune	37.9(0.3)	55.3(0.3)	41.7(0.2)

Table 5
An English-German multi-modal translation example generated by different models. Note that the word "*klatscht*" in blue is the German translation of source word "*claps*". Here we directly cited the translation results of Soft-attention and Hard-Attention from [14].

Source	a child claps while riding on a woman 's shoulders.
Reference	ein kind sitzt auf den schultern einer frau und klatscht.
Soft-attention	ein kind, das sich auf der schultern eines frau reitet, fahrt auf den schultern.
Hard-Attention	ein kind in der haltung, wahrend er auf den schultern einer frau fahrt.
MNMT$_{-\text{CO_ATT}-\text{BI_ATT}}$	ein kind kniet auf der schultern einer frau.
MNMT$_{-\text{CO_ATT}}$	ein kind klatscht und reitet auf den schultern einer frau.
MNMT$_{-\text{BI_ATT}}$	ein kind klatscht in den schultern einer frau.
MNMT	ein kind klatscht auf den schultern einer frau.

Table 6
A German-English multi-modal translation example generated by different models. Here we directly cited the translation result of IMG$_{\text{E+D}}$ from [9].

Source	eine gruppe junger menschen trinkt shots in einem mexikanischen setting.
Reference	a group of young people take shots in a mexican setting.
IMG$_{\text{E+D}}$	a group of young people drinking dishes in a Mexican restaurant.
MNMT$_{-\text{CO_ATT}-\text{BI_ATT}}$	a group of young people are drinking pants in a mexican.
MNMT$_{-\text{CO_ATT}}$	a group of young people are drinking shorts in a mexican club.
MNMT$_{-\text{BI_ATT}}$	a group of young people are having a drink in a mexican setting.
MNMT	a group of young people are drinking in a mexican setting.

lation verified that incorporating text-image semantic interactions remarkably improves translation quality compared to the conventional multi-modal NMT model.

Our future works include the following aspects. First, since previous approaches just use image features that does not explicitly represent high-level semantic concept (e.g. attribute, object), we want to extract fine-grained semantic facts from images and investigate how to incorporate it into the multi-modal NMT. Second, we will adapt the proposed model into the multi-modal NMT models with other encoders, such as context-aware based encoder [49], hierarchical RNN-based encoder [40]. Third, inspired by the success of variational NMT [50,39,10], variational multi-modal NMT with semantic interactions will also be one focus of our future work. Besides, we will explore a more effective multi-modal NMT framework to better utilize additional image description datasets, such as the Microsoft COCO Dataset.

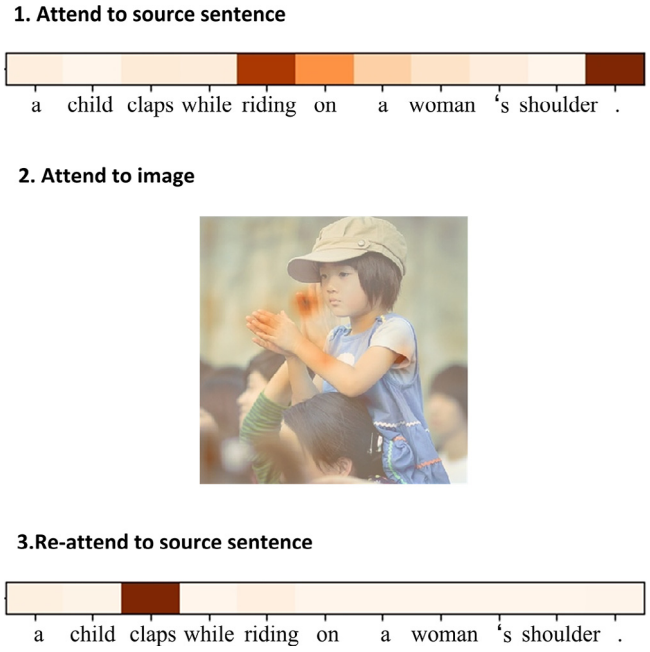

Fig. 3. Visualization of the co-attention process for the M30k$_T$ test set.

Table 7
Experimental results on English to French multi-modal translation task. Here we reported the mean and standard deviation over 4 independent runs.

Model	BLEU↑	METEOR↑	TER↓
MNMT$_{-CO_ATT-BI_ATT}$	57.3(0.5)	71.4(0.5)	27.5(0.4)
MNMT-Big$_{-CO_ATT-BI_ATT}$	57.8(0.5)	71.6(0.5)	27.2(0.2)
MNMT$_{-CO_ATT}$	58.6(0.5)	72.5(0.4)	26.8(0.5)
MNMT$_{-BI_ATT}$	58.7(0.5)	72.8(0.1)	26.6(0.3)
MNMT	59.2(0.3)	73.4(0.2)	26.0(0.4)

Table 8
Experimental results on English to Czech multi-modal translation task. We also provided the mean and standard deviation over 4 independent runs.

Model	BLEU↑	METEOR↑	TER↓
MNMT$_{-CO_ATT-BI_ATT}$	30.6(0.4)	28.8(0.3)	46.2(0.4)
MNMT-Big$_{-CO_ATT-BI_ATT}$	30.6(0.5)	28.9(0.2)	46.4(0.3)
MNMT$_{-CO_ATT}$	31.1(0.3)	29.2(0.2)	46.1(0.3)
MNMT$_{-BI_ATT}$	31.2(0.4)	29.3(0.4)	45.9(0.2)
MNMT	31.6(0.5)	29.7(0.2)	45.4(0.5)

CRediT authorship contribution statement

Jinsong Su: Conceptualization, Methodology, Writing - original draft, Writing - review & editing. **Jinchang Chen:** Software, Validation, Visualization, Investigation. **Hui Jiang:** Software, Validation, Visualization, Investigation. **Chulun Zhou:** Software, Writing - review & editing. **Huan Lin:** Writing - review & editing. **Yubin Ge:** Visualization, Investigation, Writing - review & editing. **Qingqiang Wu:** Writing - review & editing. **Yongxuan Lai:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing.

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.

Acknowledgement

The project was supported by National Key Research and Development Program of China(2019QY1803), National Natural Science Foundation of China (No. 61672440, No. 61672441), and Natural Science Foundation of Fujian Province of China (No.2020J06001).

References

[1] A. Artetxe, G. Labaka, E. Agirre, A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings, in: Proceedings of ACL 2018, 2018. pp. 789–798..
[2] D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, A.C. Courville, Y. Bengio, An actor-critic algorithm for sequence prediction, in: Proceedings of ICLR 2017, 2017..
[3] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: Proceedings of ICLR 2015, 2015..
[4] L. Barrault, F. Bougares, L. Specia, C. Lala, D. Elliott, S. Frank, Findings of the third shared task on multimodal machine translation, in, in: Proceedings of the Third Conference on Machine Translation: Shared Task Papers, 2018, pp. 304–323.
[5] Y. Bengio, R. Ducharme, P. Vincent, C. Janvin, A neural probabilistic language model, J. Mach. Learn. Res. 3 (2003) 1137–1155.
[6] O. Caglayan, W. Aransa, Y. Wang, M. Masana, M. García-Martínez, F. Bougares, L. Barrault, J. van de Weijer, Does multimodality help human and machine for translation and image captioning?, in: Proceedings of WMT 2016, 2016, pp 627–633.
[7] O. Caglayan, L. Barrault, F. Bougares, Multimodal attention for neural machine translation. arXiv preprint arXiv:1609.03976, 2016b..
[8] I. Calixto, Q. Liu, N. Campbell, Doubly-attentive decoder for multi-modal neural machine translation, in: Proceedings of ACL 2017, 2017a. pp. 1913–1924..
[9] I. Calixto, Q. Liu, N. Campbell, Incorporating global visual features into attention-based neural machine translation, in: Proceedings of EMNLP 2017, 2017b. pp. 992–1003..
[10] I. Calixto, M. Rios, W. Aziz, Latent variable model for multi-modal translation, in: Korhonen, A., Traum, D.R., Màrquez, L. (Eds.), Proceedings of ACL 2019, 2019. pp. 6392–6405..
[11] K. Cho, B. van Merrienboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, in: Proceedings of SSST@EMNLP 2014, 2014. pp. 103–111..
[12] J.H. Clark, C. Dyer, A. Lavie, N.A. Smith, Better hypothesis testing for statistical machine translation: Controlling for optimizer instability, in: Proceedings of ACL 2011, 2011, pp. 176–181.
[13] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, G. Hu, Attention-over-attention neural networks for reading comprehension, in: Proceedings of ACL 2017, 2017, pp. 593–602.
[14] J. Delbrouck, S. Dupont, An empirical study on the effectiveness of images in multimodal neural machine translation, in: Proceedings of EMNLP 2017, 2017, pp. 910–919.
[15] J.B. Delbrouck, S. Dupont, Multimodal compact bilinear pooling for multimodal neural machine translation. arXiv preprint arXiv:1703.08084, 2017b..
[16] J.B. Delbrouck, S. Dupont, O. Seddati, Visually grounded word embeddings and richer visual features for improving multimodal neural machine translation. arXiv preprint arXiv:1707.01009, 2017..
[17] M. Denkowski, A. Lavie, Meteor universal: Language specific translation evaluation for any target language, in: Proceedings of WMT 2014, 2014, pp. 376–380.
[18] D. Elliott, S. Frank, L. Barrault, F. Bougares, L. Specia, 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 WMT 2017, 2017. pp. 215–233..
[19] D. Elliott, S. Frank, K. Sima'an, L. Specia, Multi30k: Multilingual english-german image descriptions, in, in: Proceedings of the 5th Workshop on Vision and Language, 2016, pp. 70–74.
[20] D. Elliott, Á. Kádár, Imagination improves multimodal translation, in: Kondrak, G., Watanabe, T. (Eds.), Proceedings of IJCNLP 2017, 2017. pp. 130–141..
[21] L. Fang, X. Wang, L. Wang, Multi-modal medical image segmentation based on vector-valued active contour models, Inf. Sci. 513 (2020) 504–518.
[22] 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: Proceedings of EMNLP 2016, 2016, pp. 457–468.
[23] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of CVPR 2016, 2016, pp. 770–778.
[24] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780.
[25] P.Y. Huang, F. Liu, S.R. Shiang, J. Oh, C. Dyer, Attention-based multimodal neural machine translation, in: Proceedings of WMT 2016, 2016, pp. 639–645.
[26] G. Klein, Y. Kim, Y. Deng, J. Senellart, A.M. Rush, Opennmt: Open-source toolkit for neural machine translation, in: Proceedings of ACL 2017, System Demonstrations, 2017. pp. 67–72..
[27] J. Li, M. Li, G. Lu, B. Zhang, H. Yin, D. Zhang, Similarity and diversity induced paired projection for cross-modal retrieval, Inf. Sci. 539 (2020) 215–228.
[28] J. Li, B. Zhang, G. Lu, J. You, D. Zhang, Body surface feature-based multi-modal learning for diabetes mellitus detection, Inf. Sci. 472 (2019) 1–14.
[29] Y. Li, H. Lu, On multi-modal fusion learning in constraint propagation, Inf. Sci. 462 (2018) 204–217.
[30] H. Lin, F. Meng, J. Su, Y. Yin, Z., Yang, Y. Ge, J. Zhou, J., Luo, Dynamic context-guided capsule network for multimodal machine translation, in: Proceedings of ACM MM 2020, 2020..
[31] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (Eds.), Proceedings of NIPS 2019, 2019. pp. 13–23..
[32] J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, in: Proceedings of NIPS 2016, 2016, pp. 289–297.
[33] K. Papineni, S. Roukos, T. Ward, W. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of ACL 2002, 2002, pp. 311–318.
[34] X. Qian, Z. Zhong, J. Zhou, Multimodal machine translation with reinforcement learning. arXiv preprint arXiv:1805.02356, 2018..
[35] R. Sennrich, B. Haddow, A. Birch, Neural machine translation of rare words with subword units, in: Proceedings of ACL 2016, 2016, pp. 1715–1725.
[36] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, J. Makhoul, A study of translation edit rate with targeted human annotation, in: Proceedings of AMTA 2006, 2006..
[37] L. Specia, S. Frank, K. Sima'an, D. Elliott, A shared task on multimodal machine translation and crosslingual image description, in: Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, 2016. pp. 543–553..
[38] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.
[39] J. Su, S. Wu, D. Xiong, Y. Lu, X. Han, B. Zhang, Variational recurrent neural machine translation, in: Proceedings of AAAI 2018, 2018, pp. 5488–5495.
[40] J. Su, J. Zeng, D. Xiong, Y. Liu, M. Wang, J. Xie, A hierarchy-to-sequence attentional neural machine translation model, IEEE/ACM Trans. Audio, Speech Language Processing 26 (2018) 623–632.
[41] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, J. Dai, VL-BERT: pre-training of generic visual-linguistic representations, in: Proceedings of ICLR 2020, 2020..
[42] P. Wang, Q. Wu, C. Shen, A. van den Hengel, The vqa-machine: Learning how to use existing vision algorithms to answer new questions, in: Proceedings of CVPR 2017, 2017, pp. 3909–3918.

[43] C. Xiong, V. Zhong, R. Socher, Dynamic coattention networks for question answering, in: Proceedings of ICLR 2017, 2017..
[44] H. Xu, K. Saenko, Ask, attend and answer: Exploring question-guided spatial attention for visual question answering, in: Proceedings of ECCV 2016, 2016, pp. 451–466.
[45] P. Yang, B. Chen, P. Zhang, X. Sun, Visual agreement regularized training for multi-modal machine translation, in: Proceedings of AAAI 2020, 2020, pp. 9418–9425.
[46] 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: Proceedings of ACL 2020, 2020, pp. 3025–3035.
[47] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, Trans. Assoc. Comput. Linguistics 2 (2014) 67–78.
[48] B. Zhang, D. Xiong, J. Su, Battrae: Bidimensional attention-based recursive autoencoders for learning bilingual phrase embeddings, in: Proceedings of AAAI 2017, 2017, pp. 3372–3378.
[49] B. Zhang, D. Xiong, J. Su, H. Duan, A context-aware recurrent encoder for neural machine translation, IEEE/ACM Trans Audio, Speech Language Process. 25 (2017) 2424–2432.
[50] B. Zhang, D. Xiong, J. Su, H. Duan, M. Zhang, Variational neural machine translation, in: Proceedings of EMNLP 2016, 2016, pp. 521–530.
[51] Q. Zhang, J. Fu, X. Liu, X. Huang, Adaptive co-attention network for named entity recognition in tweets, in: Proceedings of AAAI 2018, 2018, pp. 5674–5681.
[52] L. Zhou, H. Palangi, L. Zhang, H. Hu, J.J. Corso, J. Gao, Unified vision-language pre-training for image captioning and VQA, in: Proceedings of AAAI 2020, 2020, pp. 13041–13049.