Applied Intelligence (2022) 52:14194–14203
https://doi.org/10.1007/s10489-022-03331-8

Dual-level interactive multimodal-mixup encoder for multi-modal neural machine translation

Junjie Ye[1,2] · **Junjun Guo**[1,2]

Accepted: 31 January 2022 / Published online: 3 March 2022

Abstract
Multi-modal neural machine translation (MNMT), which mainly focuses on the use of image information to guide text translation. Recent MNMT approaches have been shown that incorporating visual features into textual translation framework is helpful to improve machine translation. However, visual features always contain textual unrelated information, but the noisy visual feature fusion problem is rarely considered for traditional MNMT methods. How to extract the useful visual features to enhance textual machine translation is the key point need to be considered for MNMT. In this paper, we propose a novel Dual-level Interactive Multimodal-Mixup Encoder (DLMulMix) based on multimodal-mixup for MNMT, which can extract the useful visual features to enhance textual-level machine translation. We first employ the Textual-visual Gating to extract text related visual features, which we believe that regional features are crucial for MNMT. Then visual grid features are employed in order to establish the image context of the effective regional features. Moreover, an effective visual-textual multimodal-mixup is adopted to align textual features and visual features into multi-modal common space to improve textual-level machine translation. We evaluate our proposed method on the Multi30K dataset. The experimental results show that the proposed approach outperforms the previous efforts for both EN-DE and EN-FR tasks regarding BLEU and METEOR scores.

Keywords Multi-modal neural machine translation · Dual-level interactive multimodal-mixup encoder · Transformer · Feature fusion

1 Introduction

Multi-modal neural machine translation (MNMT) has recently attracted widespread attention, which is an important direction of neural machine translation [1–3]. Different from conventional text-based neural machine translation [4], MNMT aims to use images to guide textual machine translation, which only requires a small amount of data to achieve superior translation performance. Most of the works are focused on the Multi30K dataset for training the model [5, 6].[1]

How to explore and extract the related visual features to enhance textual machine translation is the key point for MNMT. To achieve this goal, many studies have been made recently, roughly consisting of: (1) Apply attention mechanism to extract useful visual-context information [2, 7–10]; (2) Visual features are used as an additional source language input [1, 2] to expand the amount of data, or use the concatenation method [11] to connect to the source language sentence. (3) Employ visual context to enhance text through the gating mechanism [12, 13]; Moreover, [14] constrain the consistency of the distributions of visual and textual outputs for MNMT.

Despite their successes, these methods still have various shortcoming. Grid features and regional features are two main common visual features extracted from image. Grid features are used in earlier methods [2, 15–17]. However, grid features contain a lot of textual-unrelated information

✉ Junjun Guo
guojjgb@163.com

Junjie Ye
junjieye.cdx@qq.com; junjieye@stu.kust.edu.cn

[1] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, Yunnan Province, China

[2] Yunnan Key Laboratory of Artificial Intelligence, Kunming, 650500, Yunnan Province, China

[1] The source code is available at: https://github.com/DLMulMix/DLMulMix

(noise), as shown in the left of Fig. 1, only girl and tennis racket are textual-related information (useful). Among all the grid features for a given image, a large part of them are not related to the corresponding text, directly fusion grid features and textual features may introduce unrelated visual information, e.g., background visual feautes.

Regional features can provide object-level information, some salient regions are generally useful for textual-level machine translation. However, regional features are still criticized for the lack of global visual features and the contextual information. As shown in the right of Figure 1. *'a young lady'* and *'tennis racket'* in the source language are closely related to the red area, separately, which are helpful for the machine translation, however, there are still a lot of unrelated object information for machine translation, as shown in the yellow areas. Moreover, regional features lack of global scene information. Most of earlier MNMT approaches only use grid features [11] or regional features [12], which are unable to offer sufficient visual guidance.

To address the visual features fusion problem, this paper proposes dual-level interactive multimodal-mixup encoder (DLMulMix) for MNMT. Both grid features and regional features are leveraged to enrich textual-level representations. We first employ the standard transformer embedding layer to initial the textual features and the pretrained Resnet-101 network to initial both the grid features and the regional features. Then we leverage the textual-guided visual encoder to extract the textual related regional features through textual-visual gating mechanism and region-grid self-attention. Finally, the textual-visual multimodal mixup module is proposed to align the textual related visual features into visual-textual common space via multimodal mixup strategy. And we employ a traditional transformer decoder for MNMT.

Compared with previous models, DLMulMix is able to effectively combine two types of visual features with the textual features, and learns better multi-modal joint representations for MNMT. In summary, the major contributions of our work are listed as follows:

- We introduce a novel dual-level interactive multimodal-mixup encoder to capture effectively image features extraction and multi-modal fusion. To the best of our knowledge, this paper is the first attempt to fuse grid features and regional features for MNMT.
- A textual-guided visual extraction approach is proposed to obtain the text closely related visual features with textual-visual gating mechanism and visual-level self-attention mechanism based on both grid features and regional features.
- A multimodal mixup strategy is adopted for visual-textual feature fusion. As far as we know, our work is the first attempt to employ multimodal mixup for multi-modal fusion.
- Experiments on the Multi30k [5] dataset are conducted and the results show that our model significantly improves machine translations performance, and achieves competitive BLEU and METEOR scores.

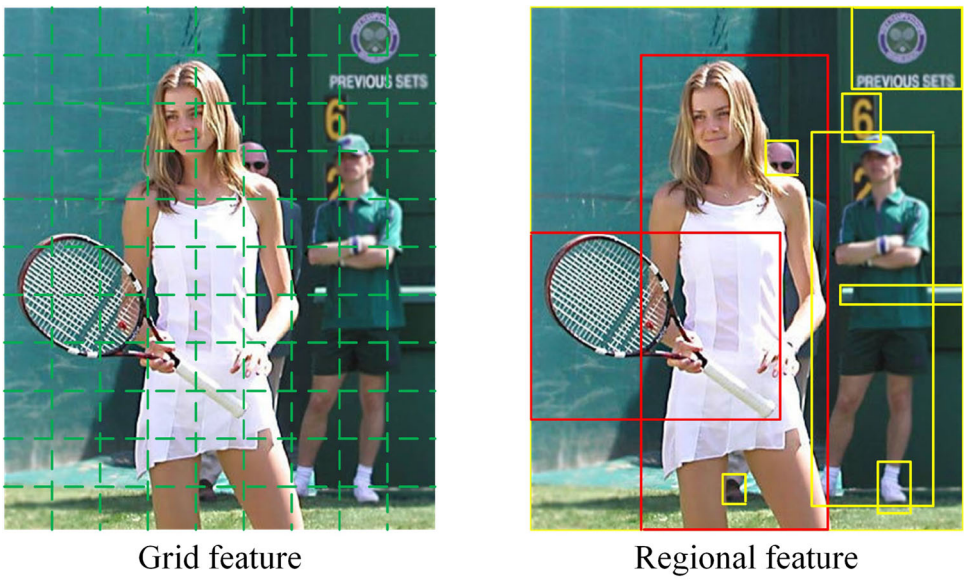

Fig. 1 Grid features and regional features. (a) an example of grid features (left), regional features (right). (b) The green grids are global visual features. (c) The red areas represent the textual related regions, and the yellow areas are textual unrelated regions

2 Related work

Multi-modal Neural Machine Translation (MNMT) It has been proven that the employ of image information can improve the performance of machine translation [18], and the research of machine translation by integrating multiple modal information has become a hot spot in machine translation research. How to extract and fuse image features has attracted great attention over the past several years. Multi-modal feature extraction and fusion approaches for MNMT can be summarized as:

The grid features are widely used in the field of natural language processing (NLP), as shown in Fig. 1 (left). [19] adopt multimodal compact bilinear pooling to fuse the grid features and textual features; [15] employ grid visual features with multi-modal attention mechanism on the decoder; [2] adopt two independent attention mechanisms to process source language words and grid visual features separately; [11] concatenate text and grid features and then use multimodal self-attention to fuse the two modalities on the encoder;

The Regional features also is an object-level features that contains all semantic visual features, as shown in Fig. 1 (right). [1] extract region visual features and fuse them to the input source language word sequence; [20] adopt object-relationship encoder and language encoder to learn the representations of text and regional features, separately, and then employ a cross-modality encoder to fuse the two representations. [12] propose a graph-based MNMT architecture by using regional features to enhance textual features; [14] employ supervised region visual features attention for MNMT.

Multimodal Mixup Mixup [21], a fast and effective learning principle, which has been widely used in computer vision (CV) and natural language processing (NLP) domains. In CV field, a manifold mixup approach [22] is presented with interpolated hidden states to improve the representations learned by deep networks, and a AdaMixUp method [23] is developed with out-of-manifold regularization for image classification, moreover, a cutmix augmentation strategy [24] is adopted to enhance image classification performance in robustness, and furthermore, a Attentive CutMix approach [25] is proposed to improve Cutmix with attention mechanism to outperform other methods by a significant margin; another mixup strategies are explored in [26–28]. In addition, mixup technique is also attempted to applied in many NLP tasks [29, 30]. Specifically, [31] adopts two adaptive mixup strategies for sentence classification, and a nonlinear mixup method [32] is presented for text classification to address the nonconvex combination problems.

3 Dual-level interactive multimodal-mixup encoder

Our proposed DLMulMix model is stacked six Transformer layers [33] in both encoder and decoder. Our work mainly focuses on building an effective encoder to incorporate the visual features into MNMT framework. As shown in Fig. 2, The DLMulMix comprises of several parts: source text embedding and visual embedding, source sentence self-attention, textual-guided regional features extraction, Multi-modal Visual-Text attention, visual-textual multimodal mixup.

3.1 Source text embedding and visual embedding

Source sentence is embedded via traditional embedding layer with position embedding, image is extracted as grid features and regional features via the pretrained Resnet-101 [34] and Faster R-CNN [35] respectively. Denote by $x_k = \{x_1^k, \cdots, x_n^k\}$ and z_k the k-th data-pair of source sentences and the corresponding image, respectively, where n is the source length of x_k. Formally, source sentence embedding and visual embedding can be expressed as follows:

$$\mathbf{C}_k^x = emb_x(x_k) \tag{1}$$

$$\mathbf{C}_{k,g}^z = emb_{z,g}(z_k) \tag{2}$$

$$\mathbf{C}_{k,r}^z = emb_{z,r}(z_k) \tag{3}$$

Where emb_x is the textual embedding layer with both word embedding and position embedding, $emb_{z,g}$ is the grid visual feature extraction layer based on Resnet-101, $emb_{z,r}$ is the regional visual feature extraction layer based on Faster R-CNN.

3.2 Source sentence self-attention

We leverage the multi-head self-attention module to generate the contextual representation of the source sentence by collecting information about nearby words by text self-attention. Formally, we have that

$$H_{k,x}^l = \text{MultiHead}(C_k^x, C_k^x, C_k^x) \tag{4}$$

where $l = \{0, 1, \cdots, 5\}$ is the transformer layer index, Multihead$(*)$ denotes the multi-head self-attention and textual feature is set as the query/key/value matrix.

3.3 Textual-guided regional features extraction

Text-based textual-visual gating mechanism Inspired by [11, 12], we employ a text-based textual-visual gating mechanism to filter out unrelated regional features for the

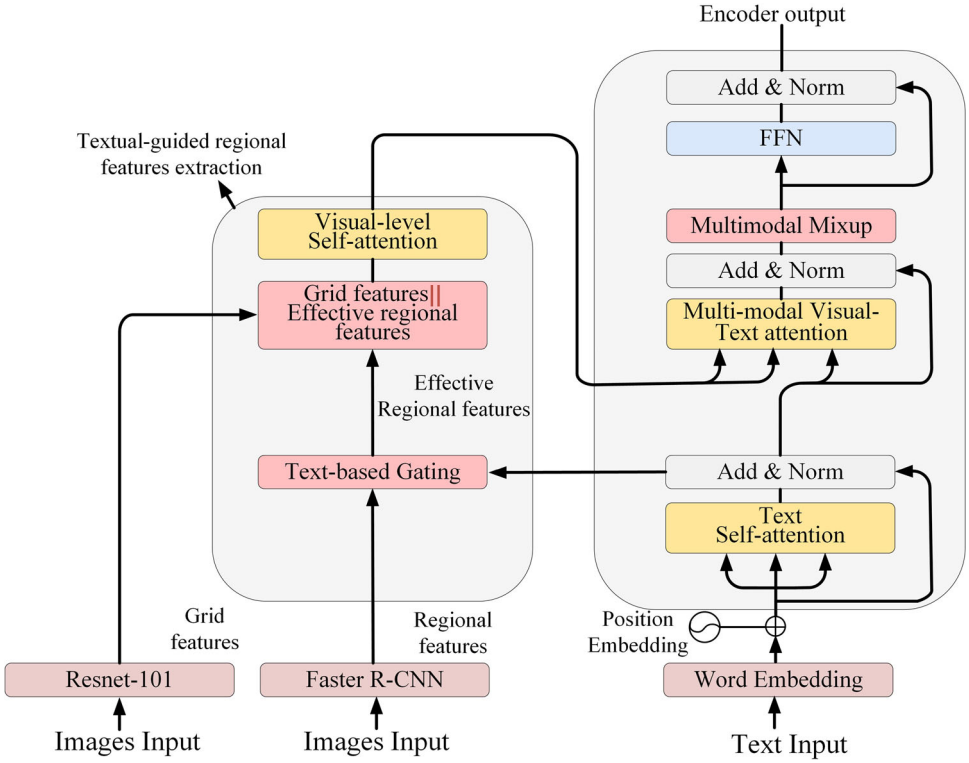

Fig. 2 The Dual-level interactive multimodal-mixup encoder for MNMT

source sentence. Our proposed multi-modal gating module is given as follows:

$$D^l_{k,r} = \alpha \odot C^z_{k,r} \tag{5}$$

$$\alpha = \text{Sigmoid}(W^l_x C^x_k + W^l_z C^z_{k,r}) \tag{6}$$

where $D^l_{k,r}$ is the textual related regional feature, α is the similarity weight between the regional feature and the textual feature. W^l_x and W^l_z are parameter matrices.

Visual-level self-attention Although we have obtained regional features consistent with the text, these regional features have not established a good connection with each other. Thus unlike most of the works we know before, we conduct a visual-level self-attention module is built between the regional features and the grid features to obtain the global vision for the textual related regional features. Specifically, we first concatenate grid features and regional features as follows:

$$E^l_{k,r,g} = C^z_{k,g} \,||\, D^l_{k,r} \tag{7}$$

Where $||$ denotes the concatenation operation. Then the visual-level self-attention is used to generate an effective regional features network that is interconnected and consistent with the text. Similar to text self-attention, visual-level self-attention can be represented as follows:

$$H^l_{k,r,g} = \text{MultiHead}(E^l_{k,r,g}, E^l_{k,r,g}, E^l_{k,r,g}) \tag{8}$$

3.4 Multi-modal visual-text attention

Inspired by multi-head visual attention fusion of different modalities proposed by [14], we use the textual features $H^l_{k,x}$ as the query matrix and the visual features $H^l_{k,r,g}$ as the key/value matrix, visual features and textual features are fused with the multi-modal visual-text attention, thus we have that

$$H^l_{k,r,g,x} = \text{MultiHead}(H^l_{k,x}, H^l_{k,r,g}, H^l_{k,r,g}) \tag{9}$$

We fuse visual information and textual information by multi-modal visual-text attention. Compared with previous work, our proposed approach can extract text closely related visual features better in MNMT. For the convenience of description, we omit the description of residual connection and layer normalization. The specific structure is shown in Fig. 2.

3.5 Visual-textual multimodal mixup

In this part, we introduce a new multimodal mixup model, the detailed multimodal mixup fusion architecture is given

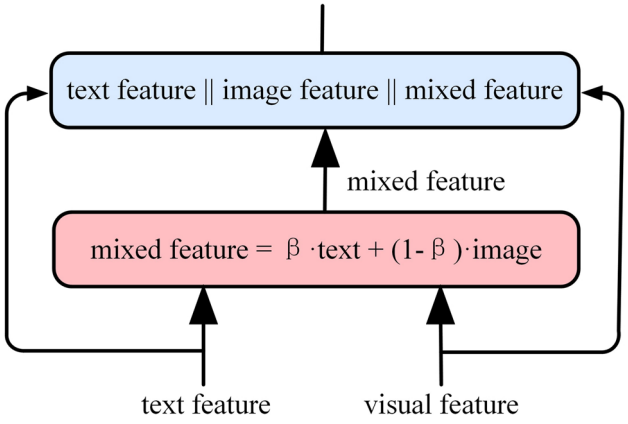

Fig. 3 Visual-textual multimodal mixup architecture

in the following Fig. 3. To the best of our knowledge, this article is the first attempt to use multimodal mixup to fuse two different text and image modalities. Thus we have that

$$E_k^l = \beta \cdot H_{k,r,g,x}^l + (1-\beta) \cdot g(C_{k,g}^z) \tag{10}$$

$$E_{k,d}^l = H_{k,r,g,x}^l \,||\, C_{k,g}^z \,||\, E_k^l \tag{11}$$

Where β is a scalar of balancing textual features and visual features, sampled from a Beta (α, α) distribution [32] with a hyper-parameter α. Similar to [24, 31, 32], we directly set α to 1 when training the model. $g(\cdot)$ is the linear transformation function. E_k^l is the mixed multi-modal features for the k-th source sentence and its corresponding image. It is worth noting that we can conduct this method on different layers.

Ultimately, Similar to conventional transformer-model architecture, we employ position-wise Feed-Forward network (FFN), forming the representation S^l of the source sentence as:

$$S^l = \text{FFN}(E_{k,d}^l) \tag{12}$$

3.6 Decoder

We leverage the traditional stacked Transformer Layer in the decoder [33], and each decoder layer is composed of three sub-layers: 1) target language self-attention layer; 2) the cross-lingual attention layer; 3) position-wise Feed-Forward network layer. Finally, the output of the last layer of the decoder is used as the softmax input, and the probability distribution of the target sentence is predicted by the softmax layer. We can express it as

$$Q = \text{Self-MultiHead}(C_{k,t}^l, C_{k,t}^l, C_{k,t}^l) \tag{13}$$

$$Y = \text{Cross-MultiHead}(Q, S^l, S^l) \tag{14}$$

$$L = \text{FFN}(Y) \tag{15}$$

$$P = \text{Softmax}\{W_s L + b\} \tag{16}$$

Where $C_{k,t}^l$ is the target sentence embedding, b and W_s are the parameters, S^l is the output of the encoder.

4 Experiment

We conduct our model on English→German and English→French translation tasks.

4.1 Datasets

Text: We use the Multi30K[2] dataset [5] like most previous works (see Table 1). The training, validation and testing sets contain 29k, 1014 and 1000 text-image pairs, respectively. We also employ the WMT17 test set containing 1000 text-image pairs and the ambiguous MSCOCO test set with 461 text-image pairs to evaluate our model. Then directly use the preprocessed sentence pairs [12] via byte pair encoding (BPE[3]) [39] segmentation with 10000 merge operations.

Image: Here we extract the visual features as grid features and regional features separately. The grid features are extracted by the pre-trained Resnet-101 [34], and the spatial features are $7\times7\times2048$-dimensional vectors with 49 local spatial regional features of the image. The regional features are extracted by the pre-trained Faster-RCNN based on Resnet-101 [35], and the spatial features are $J\times2048$-dimensional vectors with objects features of the image. Here J represents the number of objects in the image.

4.2 Settings

Our model architecture is similar to transformer. We train our model on a single GTX 3090 GPU with fp16. Specifically, we set the word embedding dimension size to 512, and the max tokens is set to 1536, and the warmup is set to 4000. In addition, the attention head is 4 and the dropout probability is 0.3. we set the label smoothing value to 0.2, and the max-update is set to 4700. Then if the BLUE score does not improve on the validation data of the 10 validation steps, we will apply early stopping. Finally, we average the checkpoints of the last 11 models parameter as our model parameter, and employ the metrics BLEU [40] and METEOR[4] [41] to evaluate the performance of our model.

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

[3] https://github.com/rsennrich/subword-nmt

[4] http://www.cs.cmu.edu/~alavie/METEOR/

Table 1 An introduction to dataset size

	Language	Sentences	Tokens	Avg-length
train	en	29000	377534	13.0
	de	29000	360706	12.4
	fr	29000	409845	14.1
valid	en	1014	13308	13.1
	de	1014	12828	12.7
	fr	1014	14381	14.2
Test2016	en	1000	12968	13.0
	de	1000	12103	12.1
	fr	1000	13988	14.0
Test2017	en	1000	11376	11.4
	de	1000	10758	10.8
	fr	1000	12596	12.6
mscoco	en	461	5239	11.4
	de	461	5158	11.2
	fr	461	5710	12.4

Avg-length represents the average length of sentences, in words/sentence

4.3 Results on the EN→DE translation task

As shown in Table 2, we report our best experimental results on the English→German translation task. It can be clearly seen that our model performance has been greatly improved compared with previous works. In particular, the BLEU scores of our model in the three test sets obtained the competitive results, and as far as we know we achieve the

Table 2 Comparison results on the EN→DE translation task on the Multi30k dataset. Best results are highlighted in bold

Model	EN→DE					
	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Traditional MNMT Systems Based on RNN						
Doubly-att (RNN) [2]	36.5	55.0	–	–	–	–
Soft-att (RNN) [7]	37.6	55.3	–	–	–	–
Stochastic-att (RNN) [7]	38.2	55.4	–	–	–	–
Fusion-conv (RNN) [36]	37.0	57.0	29.8	51.2	25.1	46.0
Trg-mul (RNN) [36]	37.8	57.7	30.7	**52.2**	26.4	47.4
VMMT (RNN) [37]	37.7	56.0	30.1	49.9	25.5	44.8
Existing MNMT Systems Based on Transformer						
DCCN [13]	39.7	56.8	31.0	49.9	26.7	45.7
Supervised Visual Attention [14]	40.50	**59.05**	–	–	–	–
Multimodal Transformer [11]	39.5	56.9	–	–	–	–
Graph-based MMT [12]	39.8	57.6	32.2	51.9	28.7	47.6
Our MNMT Systems Based on Transformer (Fairseq Framework)						
Transformer [33]	38.80	56.94	31.25	50.23	28.42	46.97
Doubly-ATT [38] †	39.30	57.48	31.96	50.96	28.37	47.31
Multimodal Transformer [11] †	39.60	57.53	32.14	51.36	28.46	48.15
Graph-based MMT [12] †	39.99	57.82	32.42	51.24	29.20	48.43
Our model	**41.77**	58.93	**33.07**	51.85	**29.90**	**49.09**

† means to reproduce previous work based on our framework

Table 3 Ablation study on multimodal fusion method

Ours model	EN→DE					
	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Full Model	41.77	58.93	32.89	51.58	29.78	48.72
Remove mixup	40.56	57.90	31.84	50.73	29.13	48.06
Remove gating	40.43	57.95	32.25	51.05	28.94	48.47
Remove mixup and gating	40.28	57.85	31.89	50.78	28.08	47.35
Transformer	38.80	56.94	31.25	50.23	28.42	46.97

Influence of Multimodal Mixup and gating in our model

new SOTA in MNMT. Compared with previous efforts, we draw the following conclusions:

First, our proposed multimodal attention module can pay more attention on text than picture. We establish dual-level interactive multimodal encoder to extract and fuse useful visual features to improve NMT. Our proposed textual-guided regional features extraction approach can selectively extract textual related visual information to enhance machine translation.

Second, different from the previous fusion methods, we propose multimodal mixup to fuse textual features and visual features. This is a first attempt to adopt mixup strategy for MNMT.

4.4 Ablation study

Ablation study on multimodal fusion To explore the influence of multimodal mixup and the visual-textual gating mechanism on the performance of our model, we conduct the experiments to remove the multimodal mixup and the visual-textual gating mechanism. The experimental results are reported in Table 3:

(1) *Effectiveness of mixup.* When we remove multimodal mixup. It is obvious that both BLEU and METEOR scores on all three test sets significantly decreased. This implies that fusion visual features are helpful to improve the performance of machine translation through multimodal mixup.

(2) *Effectiveness of gating.* When we remove the gating, the scoring criteria of the three test sets that examine the performance of the model all drop. The results suggest that visual-textual gating mechanism is helpful to enhance the performance of machine translation. In addition, when we remove both multimodal mixup and gating subnetworks, the experimental results are lower than removing mixup or gating. This experimental

Table 4 Ablation study on image features

Ours model	EN→DE					
	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Full Model	41.77	58.93	32.89	51.58	29.78	48.72
Remove grid features	41.38	58.89	32.92	51.41	29.62	48.69
Remove regional features	37.17	55.44	29.77	49.09	26.00	45.77
Regional features replace with random features	39.23	57.22	32.56	50.77	28.62	47.72
Grid features replace with random features	40.92	58.51	32.04	51.25	28.77	47.64
Transformer	38.80	56.94	31.25	50.23	28.42	46.97

Influence of grid features and regional features in our model

Source:	a little girl with blond-hair is playing and splashing in a mud puddle .
Reference:	ein kleines mädchen mit blonden haaren spielt und spritzt in einer schlammpfütze .
Transformer:	ein kleines blondes mädchen spielt und planscht in einer pfütze .
Remove mixup+gating:	ein kleines mädchen mit blondem haar spielt und planscht in einer pfütze .
Remove mixup:	ein kleines mädchen mit blondem haar spielt und planscht in einer schlammpfütze .
Full model:	ein kleines mädchen mit blonden haaren spielt und planscht in einer schlammpfütze .

Fig. 4 Translation example on EN-DE task. Example comparison of ablation study on multimodal fusion on the Test2016 test set

result further demonstrates the effectiveness of our proposed method.

(3) *Transformer.* Compare with the Transformer baseline method and our proposed approach, experimental results show that not adopt our DLMulMix model caused a significant performance degradation. This proves the effectiveness of our proposed model.

Ablation study on image features Our model aims to construct a visual region closely related to the text to guide textual machine translation, and obtain the related visual-textual common representations. In this section, we explore the effectiveness of grid features and regional features for our proposed MNMT approach. We conduct the experiment to remove grid features and regional features. The results are reported in Table 4:

(1) *The effectiveness of the grid features.* Removing grid features resulted in a slight drop in the experimental results, especially in the Test2016 test set. In addition, when grid features are replaced with random features, both BLEU and METEOR scores drop significantly in all three test datasets, which implies that visual grid features are beneficial to improve the performance of machine translation.

(2) *The effectiveness of the regional features.* Removing regional features, only grid features are used to guide machine translation. It can be seen from the results in the third row that this change causes a significant decrease in the performance of the model. This suggests that regional features are very important in multimodal neural machine translation. Grid features contain a lot of noise that is not related to the text, which leads to a sharp drop in the experimental results.

4.5 Case analysis

To better demonstrate the effect of the model in this paper, we conduct a case analysis of the ablation study of multimodal fusion, and the results are reported in Fig. 4.

Table 5 Experimental results on the EN→FR translation task. Best results are highlighted in bold

EN→FR Model	Test2016		Text2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Existing MNMT Systems						
Fusion-conv(RNN) [36]	53.5	70.4	51.6	68.6	43.2	63.1
Trg-mul(RNN)[36]	54.7	71.3	52.7	69.5	43.5	63.2
Deliberation Network(TF) [42]	59.8	74.4	-	-	-	-
DCCN [13]	61.2	76.4	54.3	70.3	**45.4**	65.0
Graph-based MMT [12]	60.9	74.9	53.9	69.3	-	-
Our MNMT Systems						
Transformer [33]	60.38	75.23	52.17	70.33	42.24	64.08
Doubly-ATT [38] †	61.22	76.42	53.99	71.67	44.26	65.36
Multimodal Transformer [11] †	61.57	76.47	54.21	71.85	42.87	64.40
Graph-based MMT [12] †	61.64	75.78	54.06	71.91	43.88	65.39
Our Model	**62.23**	**76.85**	**55.18**	**73.37**	44.42	**66.41**

† means to reproduce previous work based on our framework

Colors highlight improvement. *'mädchen mit blonden haaren'* and *'einer schlammpfütze'* are the two effective regional features on the left of Fig. 4. Both the full model and the remove mixup model are translated correctly, thus illustrating the effectiveness of text-based gating mechanism. In addition, compared with other ablation experimental models, the full model accurately translates the two words *'blonden'* and *'haaren'*. This reveals that our Dual-level interactive multimodal encoder learns more accurate representations.

4.6 Results on the EN→FR translation task

To verify the robustness of our model, we also conduct experiments on English-French translation task. The experimental results are reported in Table 5. Compared with previous works, our model has better performance, and realize the SOTA BLEU and METEOR. In Table 5, we provide strong evidence that our model performs well in different language pairs. In particular, our model achieve +3.1 improvement in METEOR on the test2017 test set. Therefore, this experiment once strongly prove the effectiveness and versatility of our model in MNMT.

5 Conclusion

In order to bridge the gap between the textual features and visual features for MNMT, in this paper, we construct a dual-level interactive multimodal-mixup encoder based on text and image. We present a text-guided visual extraction module and visual-level self-attention mechanism to filter and extract the useful visual features for the given source sentence. Moreover, we adopt a multimodal mixup strategy to incorporate different modalities into common feature space, and then leverage the aligned features for MNMT. Experimental results are given to show the effectiveness of our proposed method. In future works, we plan to adopt the multimodal mixup strategy presented in this article to other multimodal tasks. Such as voice and text multi-modal tasks, video and text multi-modal tasks.

Acknowledgments This paper is supported by National Key Research and Development Program of China (2020AAA0107900, 2020AAA0107904). National Natural Science Foundation of China (61866020,61762056,61672271), Natural Science Foundation project of Yunnan Science and Technology Department (2019FB082,2019QY1801). Yunnan provincial major science and technology special plan projects (202002AD080001).

References

1. Huang P-Y, Liu F, Shiang S-R, Oh J, Dyer C (2016) Attention-based multimodal neural machine translation. In: Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pp 639–645
2. Calixto I, Liu Q, Campbell N (2017) Doubly-attentive decoder for multi-modal neural machine translation. arXiv:1702.01287
3. Pota M, Ventura M, Fujita H, Esposito M (2021) Multilingual evaluation of pre-processing for bert-based sentiment analysis of tweets. Expert Syst Appl 181:115119
4. Liu Y, Gu J, Goyal N, Li X, Edunov S, Ghazvininejad M, Lewis M, Zettlemoyer L (2020) Multilingual denoising pre-training for neural machine translation. Trans Assoc Comput Linguist 8:726–742
5. Elliott D, Frank S, Sima'an K, Specia L (2016) Multi30k: Multilingual english-german image descriptions. arXiv:1605.00459
6. Su J, Chen J, Jiang H, Zhou C, Lin H, Ge Y, Wu Q, Lai Y (2021) Multi-modal neural machine translation with deep semantic interactions. Info Sci 554:47–60
7. Delbrouck J-B, Dupont S (2017) Multimodal compact bilinear pooling for multimodal neural machine translation. arXiv:1703.08084
8. Helcl J, Libovickỳ J, Variš D (2018) Cuni system for the wmt18 multimodal translation task. arXiv:1811.04697
9. Zhou M, Cheng R, Lee YJ, Yu Z (2018) A visual attention grounding neural model for multimodal machine translation. arXiv:1808.08266
10. Guarasci R, Silvestri S, De Pietro G, Fujita H, Esposito M (2021) Assessing bert's ability to learn italian syntax: a study on null-subject and agreement phenomena. J Ambient Intell Human Comput:1–15
11. Yao S, Wan X (2020) Multimodal transformer for multimodal machine translation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp 4346–4350
12. Yin Y, Meng F, Su J, Zhou C, Yang Z, Zhou J, Luo J (2020) A novel graph-based multi-modal fusion encoder for neural machine translation. arXiv:2007.08742
13. Lin H, Meng F, Su J, Yin Y, Yang Z, Ge Y, Zhou J, Luo J (2020) Dynamic context-guided capsule network for multimodal machine translation. In: Proceedings of the 28th ACM International Conference on Multimedia, pp 1320–1329
14. Nishihara T, Tamura A, Ninomiya T, Omote Y, Nakayama H (2020) Supervised visual attention for multimodal neural machine translation. In: Proceedings of the 28th International Conference on Computational Linguistics, pp 4304–4314
15. Calixto I, Elliott D, Frank S (2016) Dcu-uva multimodal mt system report. In: Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pp 634–638
16. Catelli R, Gargiulo F, Casola V, De Pietro G, Fujita H, Esposito M (2020) Crosslingual named entity recognition for clinical de-identification applied to a covid-19 italian data set. Appl Soft Comput 97:106779
17. Guarasci R, Silvestri S, De Pietro G, Fujita H, Esposito M (2022) Bert syntactic transfer: A computational experiment on italian, french and english languages. Comput Speech Lang 71:101261
18. Zhang Z, Chen K, Wang R, Utiyama M, Sumita E, Li Z, Zhao H (2019) Neural machine translation with universal visual representation. In: International Conference on Learning Representations
19. Fukui A, Park DH, Yang D, Rohrbach A, Darrell T, Rohrbach M (2016) Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv:1606.01847
20. Tan H, Bansal M (2019) Lxmert: Learning cross-modality encoder representations from transformers. arXiv:1908.07490
21. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D (2017) mixup: Beyond empirical risk minimization. arXiv:1710.09412
22. Verma V, Lamb A, Beckham C, Najafi A, Courville A, Mitliagkas I, Bengio Y (2018) Manifold mixup: Learning better representations by interpolating hidden states. stat 1050:4

23. Guo H, Mao Y, Zhang R (2019) Mixup as locally linear out-of-manifold regularization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 33, pp 3714–3722
24. Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y (2019) Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 6023–6032
25. Walawalkar D, Shen Z, Liu Z, Savvides M (2020) Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification. arXiv:2003.13048
26. Zhang Z, He T, Zhang H, Zhang Z, Xie J, Li M (2019) Bag of freebies for training object detection neural networks. arXiv:1902.04103
27. He T, Zhang Z, Zhang H, Zhang Z, Xie J, Li M (2019) Bag of tricks for image classification with convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 558–567
28. Berthelot D, Carlini N, Goodfellow I, Papernot N, Oliver A, Raffel C (2019) Mixmatch: A holistic approach to semi-supervised learning. arXiv:1905.02249
29. Sun L, Xia C, Yin W, Liang T, Yu PS, He L (2020) Mixup-transformer: Dynamic data augmentation for nlp tasks. arXiv:2010.02394
30. Wu Y, Inkpen D, El-Roby A (2021) Mixup regularized adversarial networks for multi-domain text classification. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 7733–7737
31. Guo H, Mao Y, Zhang R (2019) Augmenting data with mixup for sentence classification: An empirical study. arXiv:1905.08941
32. Guo H (2020) Nonlinear mixup: Out-of-manifold data augmentation for text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 34, pp 4044–4051
33. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
34. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
35. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst 28:91–99
36. Caglayan O, Aransa W, Bardet A, García-Martínez M, Bougares F, Barrault L, Masana M, Herranz L, Van de Weijer J (2017) Lium-cvc submissions for wmt17 multimodal translation task. arXiv:1707.04481
37. Calixto I, Rios M, Aziz W (2018) Latent variable model for multi-modal translation. arXiv:1811.00357
38. Arslan HS, Fishel M, Anbarjafari G (2018) Doubly attentive transformer machine translation. arXiv:1807.11605
39. Sennrich R, Haddow B, Birch A (2015) Neural machine translation of rare words with subword units. arXiv:1508.07909
40. Papineni K, Roukos S, Ward T, Zhu W-J (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp 311–318
41. Denkowski M, Lavie A (2014) Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the ninth workshop on statistical machine translation, pp 376–380
42. Ive J, Madhyastha P, Specia L (2019) Distilling translations with visual awareness. arXiv:1906.07701

Junjie Ye is now studying for an MS degree at Kunming University of Science and Technology, China. He received his BS degree in Kunming University of Science and Technology, China. His current research interests include multi-modal fusion and machine translation.

Junjun Guo is currently an associate professor at Kunming University of Science and Technology, China. He graduated from China University of Petroleum, China, in 2010. He received the Doc. degree from Xi'an Jiaotong University, China, in 2017. His research interests include pattern recognition, multi-modal fusion and machine translation.