Noise-robust Cross-modal Interactive Learning with Text2image Mask for Multi-modal Neural Machine Translation

Junjie Ye[1,2], Junjun Guo[1,2]*, Yan Xiang[1,2], Kaiwen Tan[1,2], Zhengtao Yu[1,2]
[1] Faculty of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming, China
[2] Yunnan Key Laboratory of Artificial Intelligence, Kunming, China
junjieye.cdx@qq.com, guojjgb@163.com,
sharonxiang@126.com, kwtan0909@qq.com, ztyu@hotmail.com

Abstract

Multi-modal neural machine translation (MNMT) aims to improve textual level machine translation performance in the presence of text-related images. Most of the previous works on MNMT focus on multi-modal fusion methods with full visual features. However, text and its corresponding image may not match exactly, visual noise is generally inevitable. The irrelevant image regions may mislead or distract the textual attention and cause model performance degradation. This paper proposes a noise-robust multi-modal interactive fusion approach with cross-modal relation-aware mask mechanism for MNMT. A text-image relation-aware attention module is constructed through the cross-modal interaction mask mechanism, and visual features are extracted based on the text-image interaction mask knowledge. Then a noise-robust multi-modal adaptive fusion approach is presented by fusion the relevant visual and textual features for machine translation. We validate our method on the Multi30K dataset. The experimental results show the superiority of our proposed model, and achieve the state-of-the-art scores in all En-De, En-Fr and En-Cs translation tasks [1].

1 Introduction

Multi-modal Neural Machine Translation (MNMT) aims to optimize the conventional text-only machine translation systems by using multi-modal information (eg., image, video, sound), which has received growing research attentions in the fields of CV and NLP, recently. A reasonable assumption is that visual information is helpful to improve textual-level machine translation (Elliott et al., 2017; Barrault et al., 2018; Ye and Guo, 2022), and many studies have been carried out to conduct the benefits of image for NMT (Caglayan et al., 2019; Yin et al., 2020; Li et al., 2021a). As expected, the fusion of visual information actually improves the performance of machine translation (Caglayan et al., 2019).

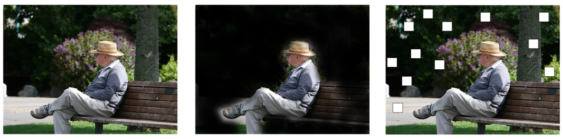
Most existing MNMT methods mainly focus on how to design a excellent multi-modal fusion framework to bridge the semantic gap between image and text, while visual noise is often ignored. Unfortunately, it is often difficult to obtain the image that exactly match the textual information. What is worse, image information and textual information may even be weakly correlated with each other. Visual noise is generally unavoidable(Li et al., 2022; Yao and Wan, 2020). As shown in Figure 1 (left), objects such as old man, brown hat and bench included in the image, these visual objects correspond to the '*old man*', '*brown hat*' and '*bench*' in the source sentence, which is the useful visual information for machine translation. However, the image also contains some irrelevant visual information (e.g., tree, flower, grasse) for the source sentence, the mismatched visual-textual information may distract the multi-modal fusion and then lead to machine translation performance decay. Therefore, it is necessary to consider noise-robust text-image fusion problem for MNMT.

How to effectively and efficiently extract useful visual information is one of the core issues of MNMT, there are three main multi-modal fusion methods: 1) Multi-modal attention mechanism, such as cross-modal interactive attention mechanism (Kwon et al., 2020; Song et al., 2021; Zhao et al., 2021) and adaptive feature selection mechanism (Wang and Xiong, 2021; Zhao et al., 2022; Li et al., 2022) between visual features and textual features. 2) Multi-modal Transformer fusion methods, which utilizes Transformer to encode textual features and visual features separately (Takushima et al., 2019; Nishihara et al., 2020), and then a multi-head cross-modal attention mechanism (Yao and Wan, 2020; Gain et al., 2021; Li et al., 2021a) is

[*] Corresponding author.

[1] https://github.com/nlp-mmt/Noise-robust-Text2image-Mask

Proceedings of the 29th International Conference on Computational Linguistics, pages 5098–5108
October 12–17, 2022.



Source *(En)*: the old man in the brown hat is sitting on the bench .
Target *(Fr)*: le vieil homme au chapeau brun est assis sur le banc .

Figure 1: An example of an En→Fr translation that illustrates the need to consider for image noise in the translation model.

adopted to integrate them. 3) Gating fusion methods (Yin et al., 2020; Lin et al., 2020; Li et al., 2021b), which are leveraged to ensure both textual semantic representations and visual semantic representations are consistent with each other. Above existing methods mainly focus on designing multimodal feature fusion architectures by leveraging visual information to enhance traditional machine translation, however, visual noise problem is ignored.

This paper endeavors to address visual noise-robust multi-modal fusion problem for MNMT, we attempt to explore robust multi-modal interactive fusion strategy with cross-modal relation-aware mask mechanism for MNMT in Transformer framework. Concretely, a text-image relation-aware attention module is constructed in the visual transformer encoder by cross-modal interactive mask mechanism, and the visual features are extracted based on text-to-image interactive mask knowledge. Then a noise-robust multi-modal fusion approach is adopted to integrate visual features into seq2seq framework more efficiently and effectively. Comparing with previous works, the major contributions of our paper are three-fold.

- A noise-robust multi-modal fusion approach is proposed with cross-modal relation-aware mask for MNMT. To the best of our knowledge, it is the first attempt to explore mask-based multi-modal representation for MNMT.
- A text-image relation-aware module is constructed with cross-modal interaction masking mechanism to obtain text-image interaction mask knowledge for noise-robust multi-modal representation and fusion in noisy scenes.
- The extensive experimental results show that our proposed model outperforms other state-of-the-art MNMT approaches and significantly improves machine translation performance on En-De, En-Fr and En-Cs translation tasks. Furthermore, we emphasize the interpretability of the model, the in-depth analysis of the experimental results show the effectiveness of our proposed method.

2 Related Work

MNMT Early attempts mainly focused on RNN-based encoder-decoder architecture with attention (Huang et al., 2016; Calixto et al., 2017; Delbrouck and Dupont, 2017). Recently, Transformer-based seq2seq framework has achieved significant improvement for MNMT. Zhao et al. (2021) utilized object detection features with an additional region-dependent attention mechanism to fusion visual regional features and textual features; Nishihara et al. (2020) presented a supervised cross-modal attention module to align textual features and visual features; Song et al. (2021) employed a co-attention graph updating module at each Transformer encoder layer to align multi-modal features. Yao and Wan (2020) used multi-modal Tranformer to align both visual features and textual features; Yin et al. (2020) proposed a graph-based MNMT approach to extract multi-model features through text-image gating attention mechanism; Lin et al. (2020) adopted a gating mechanism to fuse visual features extracted by a dynamic context-guided capsule network;

All the above methods focus on multimodal feature fusion methods, and they assume that visual information is closely related to textual information, which heavily restricts their robustness. However, text and its corresponding image may not match exactly, visual noise is generally inevitable. In this work, we systematically investigate whether masking visual noise helps machine translation.

Mask Strategy Mask strategy is one of the most effective ways of representation learning, which has been widely used in vision and textual pre-training models. We summarize the existing mask strategies in the three aspects as follows: 1) Vision mask-based pre-trained models (Li et al., 2021c; Peng and Harwath, 2022; Xie et al., 2021), the main purpose is to mask image patch-level for better visual robust-representation learning. 2) Textual mask-based pre-trained models (Joshi et al., 2020; Fu et al., 2022; Devlin et al., 2019), the tokens of the input sentences are randomly masked, and then are predicted in decoder, which aims to generate more fine-grained textual representations. 3) Cross-

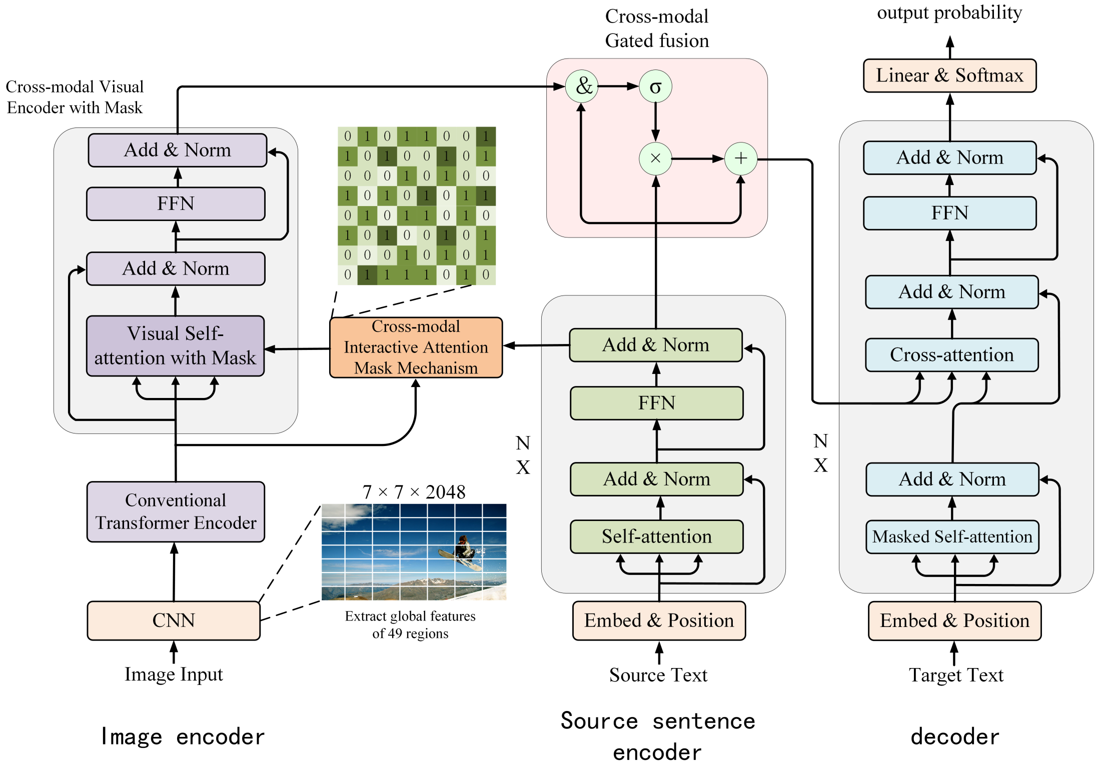


Figure 2: The overview of our proposed model, which consists of four components: (a) the image encoder to encode visual information with cross-modal interactive attention mask mechanism based on Transformer encoder; (b) the source sentence encoder to encode textual information; (c) the cross-modal gated fusion module to fuse helpful visual features and textual features; (d) the decoder to generate target translation conditioned on encoded textual features with helpful visual information;

modal mask-based pre-trained models (Li et al., 2020; Zhou et al., 2021; Shin et al., 2022), both text tokens and vision tokens are randomly masked, which aims to learn multimodal representations between vision and language in a pre-training manner. The mask strategy has been shown effective in many pre-training representation learning tasks. Inspired by Li et al. (2021c), in this work, we try to exploit the mask strategy to address the noise-robust multi-modal fusion problem for MNMT.

3 Methodology

In this section, we introduce our proposed noise-robust multi-modal neural machine translation approach, as illustrated in Figure 2. Our proposed model is based on the structure of Transformer, which contains four subnetworks, 1) source sentence encoder, 2) image encoder with robust masking matrix, 3) cross-modal gated fusion module and 4) target sentence decoder.

Without loss of generality, input words are embedded via traditional embedding layer with position embedding. As an example, denote by $x_j = \{x^j_1, \cdots, x^j_n\}$ and v_j as the j-th data-pair of source sentence input and its corresponding image, respectively, where n is the source length of x_j. Formally, the source sentence representation E^x_j and visual representation E^v_j are calculated as $\mathrm{E}^x_j = \mathrm{Emb}_x(x_j)$ and $\mathrm{E}^v_j = \mathrm{Emb}_v(v_j)$, where, Emb_x is the textual embedding layer with both word embedding and position embedding, Emb_v is the visual feature extraction layer with Resnet-101, $\mathrm{E}^x_j \in R^{n \times d_1}$ and $\mathrm{E}^v_j \in R^{m \times m \times d_2}$.

3.1 Source Sentence Encoder

As shown in the middle part of Figure 2, our encoder is employed the same as the conventional multi-head Transformer encoder, and each encoder layer is composed of two sublayers: 1) self-attention layer and 2) position-wise feedforward network (FFN) layer. Concretely, we first employ the multi-head self-attention module is used here by taking the sourece textual representation as a query/key/value matrix to establish word-to-word interconnections, which can be expressed as,

$$\mathrm{H}^l_{x_j} = \mathrm{Multihead}(\mathrm{E}^x_j, \mathrm{E}^x_j, \mathrm{E}^x_j) \quad (1)$$

$$= \mathrm{Concat}(head^1_j, \cdots, head^M_j) \quad (2)$$

where, M denotes the number of heads, Multihead$(\cdot)$ is a multi-head attention layer, $l = \{0, \cdots, 3\}$ is the Transformer layer index. Formally, the output of Multi-head attention is computed as follows:

$$head_j^{c\in[1,M]} = \sum_{k=1}^{n} \alpha_{ik}(\mathrm{E}_{j_k}^{x}\mathbf{W}_{j,c}^{V}) \tag{3}$$

where n is the source length of x_j, the weight coefficient of α_{ik} is calculated by the softmax function:

$$\alpha_{ik} = \mathrm{softmax}\left(\frac{(\mathrm{E}_{j_i}^{x}\mathbf{W}_{j,c}^{Q})(\mathrm{E}_{j_k}^{x}\mathbf{W}_{j,c}^{K})^{\mathrm{T}}}{\sqrt{d}}\right) \tag{4}$$

where α_{ik} is the dot-product attention matrix of the textual features and multi-modal features, $\mathbf{W}_{j,c}^{V}$, $\mathbf{W}_{j,c}^{Q}$, $\mathbf{W}_{j,c}^{K}$ are parameter matrices.

Then the position-wise Feed-Forward neural network is used to update the state of each position of the sequence for produce $\mathrm{F}_{x_j}^{l}$ as follows:

$$\mathrm{F}_{x_j}^{l} = \mathrm{FFN}(\mathrm{H}_{x_j}^{l}) \tag{5}$$

3.2 Image Encoder with Robust Masking Matrix

As shown in the left part of Figure 2, our image encoder layer is composed of two sublayers: 1) conventional Transformer encoder and 2) cross-modal visual encoder with mask. To reduce the number of parameters of the proposed model, we only use a single Transformer layer in image encoder.

3.2.1 Conventional Transformer Encoder for Visual

The image feature is extracted by the pretrained Resnet-101 models, and the image spatial feature is 7×7×2048-dimensional vector with 49 local spatial region features of each image. And we then transfer them into a 49× d feature matrix by linear transformation, where d denote the word-embedding-dimensional. Then, an internal relationship is established between the 49 image regions, concretely, we generate the contextual representations H_{v_j} of the 49 local spatial region features by a conventional Transformer-encoder, which can be expressed as,

$$\mathrm{H}_{v_j} = \mathrm{Multihead}(\mathrm{E}_j^{v}, \mathrm{E}_j^{v}, \mathrm{E}_j^{v}) \tag{6}$$

$$\mathrm{F}_{v_j} = \mathrm{FFN}(\mathrm{H}_{v_j}) \tag{7}$$

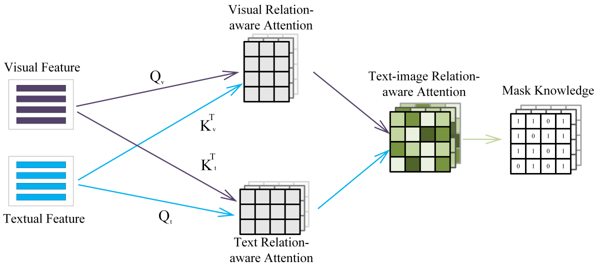


Figure 3: cross-modal interaction attention mask mechanism module.

3.2.2 Cross-modal mask mechanism for visual presentation

Inspired by Li et al. (2021c), in this section, we will introduce the proposed cross-modal visual encoder with mask module. Specifically, to mask irrelevant visual information before cross-modal fusion, we propose a cross-modal interaction attention mask mechanism, as shown in figure 3. First, cross-modal interaction of textual features and visual features is performed to compute the correlations between 49 regional features and textual features as follows:

$$\mathrm{Matrix}_{v_j} = \mathrm{softmax}\left(\frac{\mathrm{F}_{v_j} \times (\mathrm{F}_{x_j}^{l})^{\mathrm{T}}}{\sqrt{d}}\right) \tag{8}$$

$$\mathrm{Matrix}_{x_j} = \mathrm{softmax}\left(\frac{\mathrm{F}_{x_j}^{l} \times (\mathrm{F}_{v_j})^{\mathrm{T}}}{\sqrt{d}}\right) \tag{9}$$

where, $\mathrm{Matrix}_{v_j} \in R^{49\times n}$ denotes the attention of 49 local region features to each word of the paired source sentence, $\mathrm{Matrix}_{x_j} \in R^{n\times 49}$ denotes the attention of each word of the source sentence on the 49 local regions of the paired image.

Then we interactively compute Matrix_{v_j} and Matrix_{x_j} as follows,

$$\mathrm{Mask}_j = \mathrm{Matrix}_{v_j} \times \mathrm{Matrix}_{x_j} \tag{10}$$

where, Mask_j denotes the correlation matrix between the 49 local regions of the image and corresponding source sentences.

The mask matrix is generated according to the importance of the local region information of the image, which we set a threshold $prob_r$ to control the image region that needs to be masked. Thus we have that

$$m_r = \begin{cases} 1, prob_r \geq p, (r = \{1, 2, \cdots, 49\}) \\ 0, prob_r < p \end{cases} \tag{11}$$

where, p is a hyper-parameter, which is leveraged to mask unimportant visual region features, and it

Model	Multi30K En→De					
	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Existing MNMT Systems						
VMMT (Calixto et al., 2019)	37.7	56.0	30.1	49.9	25.5	44.8
VAG-NMT (Zhou et al., 2018)	-	-	31.6	52.2	28.3	48.0
Del+Obj (Ive et al., 2019)	38.0	55.6	-	-	-	-
DCCN (Lin et al., 2020)	39.7	56.8	31.0	49.9	26.7	45.7
MNMT+SVA (Nishihara et al., 2020)	39.9	58.1	-	-	-	-
OVC+L_v (Wang and Xiong, 2021)	-	-	32.4	52.3	28.6	48.0
WRA-guided (Zhao et al., 2021)	39.3	58.3	32.3	52.8	28.5	48.5
Our Transformer-Based Systems						
Transformer (NMT) (Vaswani et al., 2017)	40.96	58.35	32.59	51.21	29.16	48.37
Doubly-ATT (Arslan et al., 2018) †	41.44	59.08	33.15	52.34	29.22	48.41
Multimodal self-att (Yao and Wan, 2020) †	41.50	58.52	32.51	51.33	29.10	48.48
Gated Fusion MNMT (Yin et al., 2020) †	41.58	58.88	33.01	51.90	30.04	48.95
Our model	**42.56**	**59.98**	**35.09**	**54.51**	**31.09**	**50.46**

Table 1: Comparison results on Multi30k En→De task on BLEU and METEOR metrics. † means to reproduce previous multi-modal fusion method based on our Transformer systems. Best results are highlighted in bold.

is a scalar. Our strategy ensures that each image always presents the most relevant visual region to the corresponding source textual. Then convert the image area of $m_r = 0$ to false, and $m_r = 1$ to true, which we construct a mask knowledge matrix.

Finally, we employ cross-modal visual encoder with mask to obtain effective visual information, thus we have

$$\widehat{\mathrm{H}}_{v_j} = \text{Multihead-mask}(\mathrm{F}_{v_j}, \mathrm{F}_{v_j}, \mathrm{F}_{v_j}) \quad (12)$$

$$\widehat{\mathrm{F}}_{v_j} = \mathrm{FFN}(\widehat{\mathrm{H}}_{v_j}) \quad (13)$$

where Multihead-mask$(*)$ denote the self-attention with mask knowledge, the purpose of Multihead-mask is to mask weakly correlated visual information.

3.3 Cross-modal Gated Fusion Module

In this section, we employ cross-modal gated fusion method to fuse textual features and extracted helpful visual features, which is a popular multi-modal fusion method for many recent MNMT, as shown in Figure 2. Formally, we have that

$$\Omega = \mathrm{Sigmoid}(\mathrm{W}_\Omega \widehat{\mathrm{F}}_{v_j} + \mathrm{U}_\Omega \mathrm{F}_{x_j}) \quad (14)$$

$$\mathrm{H}_{g_j} = \mathrm{F}_{x_j} + \Omega \widehat{\mathrm{F}}_{v_j} \quad (15)$$

where, W_Ω and U_Ω are trainable model parameters. The final output H_{g_j} is directly fed into our target sentence decoder (See Figure 2 right) to predict the translation.

4 Experiments

Datasets: We conduct experiments on En→De, En→Fr and En→Cs tasks of the widely used Multi30K [2] benchmark dataset, in which the training and validation sets contains 29k and 1014 text-image pairs, respectively. Furthermore, we adopt four test sets to evaluate our MNMT model, 1) the Test2016 test set with 1,000 examples contained in Multi30K; 2) the Test2017 test set with 1,000 examples in WMT2017, which contains more difficult source sentences to translate and understand; 3) we also use ambiguous COCO dataset as out-domain test data, which contains 461 examples with ambiguous verbs and encourages to use image for disambiguation; and 4) the Test2018 test set contains 1071 instances with more entity words and more low frequency words.

Data Pre-processing: We directly use the pre-processed sentence pairs via byte pair encoding (BPE) segmentation with 6k bpe vocabulary, the resulting vocabulary sizes of each language pair were 5,644→5,876 tokens for En→De, 5,644→5,684 tokens for En→Fr, 5,644→5,972 tokens for En→Cs. For each image, which is extracted through the pre-trained Resnet-101 model, the spatial features are 7x7x2048-dimensional vectors with 49 local spatial region features.

Metrics: We evaluate the quality of translations with two metrics, 1) 4-gram BLEU metrics (Pap-

[2] https://github.com/multi30k/dataset

Model	Multi30K En→Fr					
	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Existing MNMT Systems						
VAG-NMT (Zhou et al., 2018)	-	-	53.8	70.3	45.0	64.7
Del+Obj (Ive et al., 2019)	59.8	74.4	-	-	-	-
DCCN (Lin et al., 2020)	61.2	76.4	54.3	70.3	45.4	65.0
OVC+L_v (Wang and Xiong, 2021)	-	-	54.2	70.5	45.2	64.6
WRA-guided (Zhao et al., 2021)	61.8	76.3	54.1	70.6	43.4	63.8
Our Transformer-Based Systems						
Transformer (NMT) (Vaswani et al., 2017)	60.33	75.64	53.45	71.57	43.61	65.72
Doubly-ATT (Arslan et al., 2018) †	60.94	75.99	53.63	71.56	44.78	65.35
Multimodal self-att (Yao and Wan, 2020) †	61.44	75.77	54.56	71.62	44.59	65.08
Gated Fusion MNMT (Yin et al., 2020) †	61.24	76.26	54.15	71.77	44.29	64.91
Our model	**63.24**	**77.54**	**55.48**	**72.62**	**46.34**	**67.40**

Table 2: Comparison results on the En→Fr translation task on the Multi30k dataset.

ineni et al., 2002), which measures the quality of translations in terms of accuracy and fluency. 2) METEOR metrics (Denkowski and Lavie, 2014), which takes into account both precision and recall for translation quality.

4.1 Settings

We conduct our proposed models based on Transformer framework, with only stack 4-layer encoder-decoder, so the amount of parameters required by our model is small. Concretely, we set the dimensions of the encoder and decoder hidden states at d_{model}=128, the inner-layer of feed-forward network is set as d_{ffn}=256. The learning rate is set to 0.005. The max tokens is set to 4096, the learning rate is varied under a warmup-updates with 2,000 steps, and the label smoothing with value set as 0.1. We use adam optimizer with β_1, β_2 = (0.9, 0.98). We adopt 4 heads here and the dropout is set to 0.3 to avoid the over-fitting. The width of beam size is set to 5. We train our models on a single GTX 3090 GPU with fp16.

4.2 Baseline Models

To empirically verify the advantages of our proposed MNMT model, we show the performance of the following recent state-of-the-art MNMT models for comparison on the En→De and En→Fr translation task, namely: VMMT (Calixto et al., 2019), VAG-NMT (Zhou et al., 2018), Del+Obj (Ive et al., 2019), DCCN (Lin et al., 2020), MNMT+SVA (Nishihara et al., 2020), OVC+L_v (Wang and Xiong, 2021), WRA-guided (Zhao et al., 2021). Furthermore, to more fairly demonstrate the superiority and validity of our proposed model, we reproduce the recent state-of-the-art methods for comparison based on the same parameter settings and training equipment, we experiment with the following: 1) Gated Fusion MNMT (Yin et al., 2020): An efficient multi-modal fusion method to enhance machine translation. 2) Multimodal self-att (Yao and Wan, 2020): A image-aware multi-modal transformer model is proposed to extract image information to improve machine translation performance. 3) Doubly-ATT (Arslan et al., 2018): At the decoder, the visually evoked attention weights and the source language attention weights are added up as doubly-attention weights.

4.3 Results on the En→De Translation Task

As shown in Table 1, we present experimental results of our proposed model and other SOTA models on the En→De translation task. We summarize and compare the existing models in the three aspects as follows:

1) *Compare with Existing MNMT Systems:* Experimental results show that our proposed model outperforms existing SOTA models, and enhances BLEU and METEOR metrics by 3∼4 points on most of the test sets. The underlying reason is that our proposed method can effectively filter vision noise contained in the image.

2) *Compare with Text-to-text NMT:* Our MNMT model outperforms NMT baselines significantly on BLEU and METEOR metrics, which enhances about 2 points on all test sets. This indicates that our proposed MNMT model can utilize image information to improve machine translation.

3) *Compare with Reproduce Methods:* As we can observe that our proposed methods achieves

	Test2016		Test2017		MSCOCO	
p	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Multi30K **En→De**						
$p = 0$	41.40	59.19	34.37	53.77	30.01	50.17
$p = 0.01$	41.45	59.02	34.73	54.00	30.77	50.12
$p = 0.015$	41.68	59.40	34.69	54.13	30.96	50.39
$p = 0.02$	42.58	59.98	35.09	54.51	31.09	50.46
$p = 0.025$	41.71	59.62	33.68	53.39	30.82	50.27
$p = 0.03$	40.99	58.75	33.44	53.47	30.36	49.66
Multi30K **En→Fr**						
$p = 0$	61.16	76.33	54.49	72.51	44.91	65.81
$p = 0.01$	62.29	76.77	55.31	72.34	44.87	65.46
$p = 0.015$	62.67	76.96	55.36	73.00	45.48	66.76
$p = 0.02$	63.24	77.54	55.48	72.62	45.82	67.17
$p = 0.025$	62.57	76.84	55.39	72.46	46.34	67.40
$p = 0.03$	62.14	76.99	54.89	72.63	45.62	66.31

Table 3: Ablation study on hyper-parameter p on the En→Fr and En→De tasks.

a significant improvement over the SOTA method on all the evaluation metrics, which demonstrates that masking irrelevant visual information helps improve translation performance.

4.4 Results on the En→Fr Translation Task

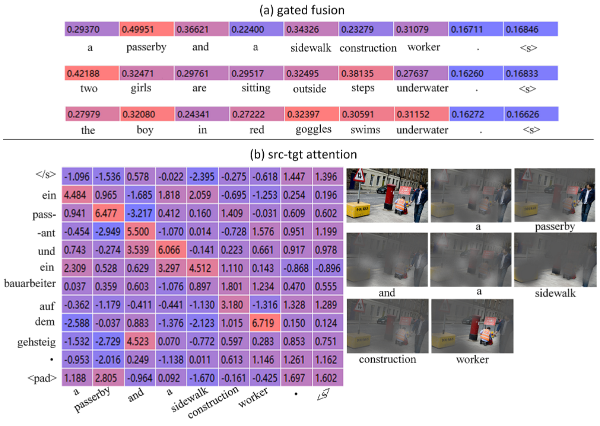
To explore the robustness of the proposed model, we also guide experiments on the Multi30K En→Fr translation task, the results are illustrated in Table 2. Concretely, we draw the following interesting conclsions:

First, comparing with existing models, our proposed model still achieves significant improvement on two evaluation metrics, which is consistently with the result of the En→De task. In addition, comparing with text-only NMT baseline models, MNMT with image information achieves superior results, which demonstrates that our MNMT model can effectively and efficiently interact with visual information to enhance machine translation.

Second, reproducing recent competitive methods based on the same NMT strong baseline model on En→Fr task, results are shown in Table 2. It is obviously that our method outperforms the SOTA methods and achieves strong competitive results among all the existing MNMT models. The results on the En→Fr translation task once again demonstrate the effectiveness and generalizability of the proposed method.

4.5 Ablation Study

To further determine the effectiveness of our proposed method, we show the following sets of ablation experiments on both the En→Fr and En→De tasks, 1) Ablation study on hyper-parameter p; 2) Ablation study on different components of model.


(a) gated fusion

0.29370	0.49951	0.36621	0.22400	0.34326	0.23279	0.31079	0.16711	0.16846
a	passerby	and	a	sidewalk	construction	worker	.	<s>
0.42188	0.32471	0.29761	0.29517	0.32495	0.38135	0.27637	0.16260	0.16833
two	girls	are	sitting	outside	steps	underwater	.	<s>
0.27979	0.32080	0.24341	0.27222	0.32397	0.30591	0.31152	0.16272	0.16626
the	boy	in	red	goggles	swims	underwater	.	<s>

(b) src-tgt attention

	a	passerby	and	a	sidewalk	construction	worker	.	<s>
</s>	-1.096	-1.536	0.578	-0.022	-2.395	-0.275	-0.618	1.447	1.396
ein	4.484	0.965	-1.685	1.818	2.059	-0.695	-1.253	0.254	0.196
pass-	0.941	6.477	-3.217	0.412	0.160	1.409	-0.031	0.609	0.602
-ant	-0.454	-2.949	5.500	-1.070	0.014	-0.728	1.576	0.951	1.199
und	0.743	-0.274	3.539	6.066	-0.141	0.223	0.661	0.917	0.978
ein	2.309	0.528	0.629	3.297	4.512	1.110	0.143	-0.868	-0.896
bauarbeiter	0.037	0.359	0.603	-1.076	0.897	1.801	1.234	0.470	0.555
auf	-0.362	-1.179	-0.411	-0.441	-1.130	3.180	-1.316	1.328	1.289
dem	-2.588	-0.037	0.883	-1.376	-2.123	1.015	6.719	0.150	0.124
gehsteig	-1.532	-2.729	4.523	0.070	-0.772	0.597	0.283	0.853	0.751
.	-0.953	-2.016	0.249	-1.138	0.011	0.613	1.146	1.261	1.162
<pad>	1.188	2.805	-0.964	0.092	-1.670	-0.161	-0.425	1.697	1.602

Figure 4: Examples of attention maps on the En→De task. (a) Gated fusion (Ω visualize): attention weights for visual information and source sentences. (b) src-tgt attention: source and target sentence attention weights with visual guidance.

Ablation study on hyper-parameter p As shown in Table 3, we report the effect of hyper-parameter p on model translation performance, where p represents the threshold that controls the effective visual similarity weights. We summarize several interesting conclusions:

First, in general, it can be observed that at $p = 0.02$, the experimental results of the proposed model on most test sets achieve the best results on the En→De and En→Fr tasks. Furthermore, gradually increase or decrease the threshold p, it is obvious that the experimental results also gradually decrease on the BLEU and METEOR metrics. We consider that there are two main reasons. On the one hand, with the decrease of the threshold p, the masked noise information decreases, and the captured visual information contains more noise, and the introduction of noise leads to a decrease in the performance of the model. On the other hand, with the increase of the threshold p, more visual information is masked, and even a lot of helpful visual information is masked, which causes the performance of the model to decline.

Secend, in more detail, when the threshold $p = 0$, which means that the model fuses all visual information, compared with NMT model, our model achieves better translation performance, but there are no significant BLEU and METEOR gains. The prove the assumption that masking visual noise information helps improve machine translation.

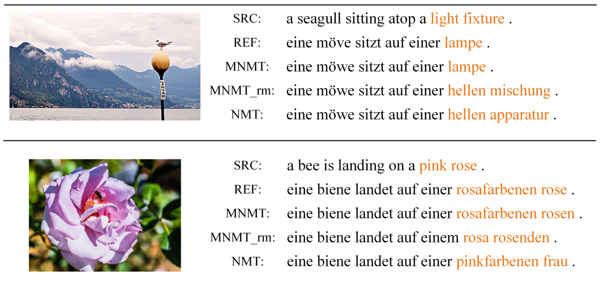
Ablation study on different components of model To investigate the effectiveness of different components in our proposed MNMT model, we

	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Multi30K **En→De**						
Complete model	42.58	59.98	35.09	54.51	31.09	50.46
MNMT_rg	41.85	59.76	34.24	54.16	29.91	50.11
MNMT_re	41.96	59.59	34.37	53.94	30.68	50.21
MNMT_rvm	41.53	59.01	34.56	53.89	30.33	50.00
Multi30K **En→Fr**						
Complete model	63.24	77.54	55.48	73.00	46.34	67.40
MNMT_rg	62.49	76.78	55.34	72.34	46.06	66.68
MNMT_re	62.85	77.37	55.41	72.84	45.73	67.30
MNMT_rvm	62.22	76.94	55.06	72.14	45.57	66.67

Table 4: Ablation study on different components of model on the En→Fr and En→De tasks. MNMT_rvm means to remove cross-modal visual encoder with mask, MNMT_re means to remove conventional Transformer encoder, MNMT_rg means to remove cross-modal gated fusion module.

En→Cs				
	Test2016		**Test2018**	
Model	BLEU	METEOR	BLEU	METEOR
Our model	**35.09**	**33.52**	**31.40**	**31.26**
Transformer (NMT)	32.70	32.34	27.62	29.03
Doubly-ATT (Arslan et al., 2018) †	33.25	32.28	29.12	29.87
Multimodal self-att (Yao and Wan, 2020) †	33.12	32.01	28.75	29.51
Gated Fusion MNMT (Yin et al., 2020) †	33.77	32.24	29.43	29.41

Table 5: Experiment results on En→Cs task.



SRC: a seagull sitting atop a light fixture .
REF: eine möve sitzt auf einer lampe .
MNMT: eine möwe sitzt auf einer lampe .
MNMT_rm: eine möwe sitzt auf einer hellen mischung .
NMT: eine möwe sitzt auf einer hellen apparatur .

SRC: a bee is landing on a pink rose .
REF: eine biene landet auf einer rosafarbenen rose .
MNMT: eine biene landet auf einer rosafarbenen rosen .
MNMT_rm: eine biene landet auf einem rosa rosenden .
NMT: eine biene landet auf einer pinkfarbenen frau .

Figure 5: Examples of successful translation with remove of visual noise. Improved translations are highlighted in color.

further conduct experiments to compare with the following variants models in Table 4:

1) *Effectiveness of cross-modal gated fusion.* The result in row 2 indicates that removing the gated fusion leads to a significant performance decline on BLEU and METEOR metrics. It suggests that gated fusion is an efficient method for fusing multimodal features, which is helpful in order to enhance translation performance.

2) *Effectiveness of conventional Transformer encoder.* To verify the usefulness of establishing the intra-modal correspondences before interacting with multimodal features, we remove the conventional Transformer encoder component. The result in row 3 shows that this change causes a light performance drop. The underlying reason is the lack of visual contextual semantic information in visual information without intra-modal correspondences.

3) *Effectiveness of cross-modal visual encoder with mask.* To construct this variant, we directly remove the cross-modal visual encoder and then employ gated fusion to incorporate full visual features and textual features. Apparently, the performance drop reported in line 4 demonstrates the validity of our proposed cross-modal visual encoder with mask module. Furthermore, it also validates our hypothesis that masking irrelevant visual information before fusing multimodal features is favourable to improve translation performance.

4.6 Visual analysis

As shown in Figure 4, to further understand and verify our model, we visualize the gated fusion and src-tgt attention weights. 1) Gated fusion: the results show that our model can effectively focus on the consistent visual regions corresponding to the source text. 2) Src-tgt attention: useful visual information as a bridge can effectively align source and target sentences to help translation.

4.7 Results on the En→Cs Translation Task

We further verify the effectiveness and generalization of the proposed method on the En→Cs task, the results shown in Table 5. Our model still achieves excellent performance compared with all baselines, which again proves that our model is effective and general for different language pairs.

4.8 Case Study

As shown in Figure 5, we further confirm the effectiveness of our proposed method. It can be observed that the two words *'light fixture'* and *'pink rose'* can be correctly translated by the MNMT model, while the MNMT_rvm model is not fully translated, and the NMT model is translated incorrectly. The underlying reason is that the complete image information introduces noise into MNMT model and distracts the model. This reveals that the proposed encoder is able to learn more efficient representations.

5 Conclusion

In this paper, we propose a noise-robust multimodal interactive fusion approach with cross-modal relation-aware mask mechanism to address image noise in MNMT. Experiment results and analysis on three benchmark translation tasks demonstrate the effectiveness and superiority of our proposed method. Further ablation experiments demonstrate that masking irrelevant visual information helps machine translation. In future work, we will continue to explore how to more effectively remove noisy information in vision.

Acknowledgments

This work was supported by National Key Research and Development Program of China (No. 2020AAA0107904). National Natural Science Foundation of China (Nos. 61866020, 61732005), Key Research and Development Program of Yunnan Province (Nos. 202203AA080004, 202103AA080015). Natural Science Foundation Project of Yunnan Science and Technology Department (No. 2019FB082)

References

Hasan Sait Arslan, Mark Fishel, and Gholamreza Anbarjafari. 2018. Doubly attentive transformer machine translation. *arXiv:1807.11605*.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *THIRD CONFERENCE ON MACHINE TRANSLATION (WMT18)*, volume 2, pages 308–327.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Incorporating global visual features into attention-based neural machine translation. *In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9–11, 2017, pp. 992–1003. doi:10.18653/v1/d17-1105*.

Iacer Calixto, Miguel Rios, and Wilker Aziz. 2019. Latent variable model for multi-modal translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6392–6405, Florence, Italy. Association for Computational Linguistics.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017. An empirical study on the effectiveness of images in multimodal neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 910–919, Copenhagen, Denmark. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics*.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. *In: Proceedings of the Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 215–233*.

Zhiyi Fu, Wangchunshu Zhou, Jingjing Xu, Hao Zhou, and Lei Li. 2022. Contextual representation learning beyond masked language modeling. *arXiv preprint arXiv:2204.04163*.

Baban Gain, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2021. Experiences of adapting multimodal machine translation techniques for hindi. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 40–44.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 639–645.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6538, Florence, Italy. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Soonmo Kwon, Byung-Hyun Go, and Jong-Hyeok Lee. 2020. A text-based visual context modulation neural model for multimodal machine translation. *Pattern Recognition Letters*, 136:212–218.

Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. 2022. On vision features in multimodal machine translation. *arXiv preprint arXiv:2203.09173*.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344.

Jiaoda Li, Duygu Ataman, and Rico Sennrich. 2021a. Vision matters when it should: Sanity checking multimodal machine translation models. *In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8556–8562*.

Jiaoda Li, Duygu Ataman, and Rico Sennrich. 2021b. Vision matters when it should: Sanity checking multimodal machine translation models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8556–8562, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, et al. 2021c. Mst: Masked self-supervised transformer for visual representation. *Advances in Neural Information Processing Systems*, 34.

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020. Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1320–1329.

Tetsuro Nishihara, Akihiro Tamura, Takashi Ninomiya, Yutaro Omote, and Hideki Nakayama. 2020. Supervised visual attention for multimodal neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4304–4314.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Puyuan Peng and David Harwath. 2022. Self-supervised representation learning for speech using visual grounding and masked language modeling. *arXiv preprint arXiv:2202.03543*.

Andrew Shin, Masato Ishii, and Takuya Narihira. 2022. Perspectives and prospects on transformer architecture for cross-modal tasks with language and vision. *International Journal of Computer Vision*, pages 1–20.

Yuqing Song, Shizhe Chen, Qin Jin, Wei Luo, Jun Xie, and Fei Huang. 2021. Enhancing neural machine translation with dual-side multimodal awareness. *IEEE Transactions on Multimedia*.

Hiroki Takushima, Akihiro Tamura, Takashi Ninomiya, and Hideki Nakayama. 2019. Multimodal neural machine translation using cnn and transformer encoder. In *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2019)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Dexin Wang and Deyi Xiong. 2021. Efficient object-level visual context modeling for multimodal machine translation: Masking irrelevant objects helps grounding. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, pages 2–9.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2021. Simmim: A simple framework for masked image modeling. *arXiv preprint arXiv:2111.09886*.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350.

Junjie Ye and Junjun Guo. 2022. Dual-level interactive multimodal-mixup encoder for multi-modal neural machine translation. *Applied Intelligence*, pages 1–10.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. *In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3025–3035*.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2021. Word-region alignment-guided multimodal neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2022. Region-attentive multimodal neural machine translation. *Neurocomputing*.

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. 2018. A visual attention grounding neural model for multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3643–3653, Brussels, Belgium. Association for Computational Linguistics.

Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. 2021. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4155–4165.