

THE PROGRESSIVE ALIGNMENT-AWARE MULTI-MODAL FUSION WITH EASY2HARD STRATEGY FOR MULTIMODAL NEURAL MACHINE TRANSLATION

Anonymous authors

Paper under double-blind review

ABSTRACT

Multimodal 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 have only focused on either multimodal feature fusion or noise multi-modal representations based on full visual and textual features, however, the degree of multi-modal alignment is often ignored. Generally, the fine-grained multi-modal information, such as visual object, textual entity, is easy to align, but the global-level semantic alignment is always difficult. In order to alleviate the challenging problem of multi-modal alignment, this paper proposes a novel progressive multimodal fusion approach with the easy-to-hard (easy2hard) cross-model alignment strategy by fully exploiting visual information for MNMT. We first extract both visual and textual features with modal-specific pre-trained models, respectively, and the fine-grained features (e.g., the regional visual features, the entity features) are roughly aligned as multi-modal anchors based on cross-modal interactive module. Then a progressive multi-modal fusion framework is employed for MNMT by gradually narrowing the global-level multi-modal semantic gap based on the roughly aligned anchors. 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 (SOTA) scores in all En-De, En-Fr and En-Cs translation tasks.

1 INTRODUCTION

Multimodal Neural Machine Translation (MNMT) Elliott et al. (2017); Barrault et al. (2018); Grönroos et al. (2018); Elliott (2018); Wu et al. (2021); Caglayan et al. (2021) aims to optimize the conventional text-only machine translation performance by fusing multimodal features (eg., image, video, sound), and it has received growing research attentions in the fields of CV and NLP, respectively. A reasonable assumption of multimodal fusion is that visual information is helpful to improve textual-level machine translation Elliott et al. (2017); Barrault et al. (2018); Ye & Guo (2022); Chen et al. (2022), 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. (2021b); Su et al. (2018); Gong et al. (2022). As expected, fusion of visual information actually improves the performance of machine translation Caglayan et al. (2019); Yawei & Fan (2021); Calixto et al. (2019); Su et al. (2021).

Image is a kind of language-independent information that can be understood by people who speaks different languages, therefore, it seems possible that visual information might serve as pivot information to narrow the gap between different languages. As shown in Figure 1, entities in different languages share the same image regions, for example: 'women' and 'frauen' can be aligned into the same red box area in the image. However, there is a significant semantic gap between visual and textual information. There are two types of definitions for multi-modal fusion, 1) local-level multi-modal alignment (the fine-grained multi-modal alignment), both visual objects and textual entities are aligned into multi-modal feature space; 2) global-level multi-modal alignment, global textual semantic features and visual features are aligned into multi-modal feature space.

Most of existing MNMT approaches mainly focus on multi-modal fusion strategies, how to extract and utilize visual information effectively and efficiently is one of the core issues for MNMT, there

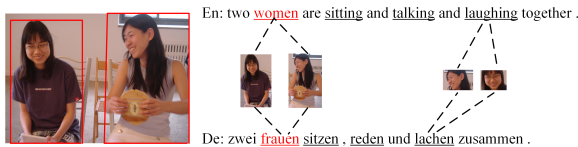


Figure 1: An example of an En→De translation that illustrates.

are three main multi-modal fusion strategies, 1) Multi-modal attention mechanism, such as cross-modal interactive attention mechanism Kwon et al. (2020); Song et al. (2021); Zhao et al. (2021); Delbrouck & Dupont (2017b) and adaptive visual-textual feature selection mechanism Wang & Xiong (2021); Zhao et al. (2022); Li et al. (2022). 2) Multi-modal Transformer fusion, Transformer framework is leveraged to encode textual features and visual features, respectively Takushima et al. (2019); Nishihara et al. (2020), and then a multi-head cross-modal attention mechanism Yao & Wan (2020); Gain et al. (2021); Li et al. (2021b); Ive et al. (2019) is adopted to integrate multi-modal features for MNMT, and 3) multi-modal gated fusion, Yin et al. (2020); Lin et al. (2020); Li et al. (2021a); Wu et al. (2020), the gating mechanism is leveraged to ensure both textual representations and visual representations are consistent with each other. However, there is a large gap between image and text, and it is difficult to directly align text information and image information by only leveraging above image-text fusion strategies. Generally, sentence-level alignment is much harder than entity-level multi-modal alignment. among global-level image-text information, non-entity clue information, such as 'sitting', 'talking' or 'laughing', is difficult to be aligned.

This paper endeavors to enhance cross-modal semantic consistency between the text-image pairs, and we propose an easy-to-hard (Easy2Hard) visual-textual fusion strategy by considering the degree of multi-modal alignment for MNMT. we first roughly align local-level object information as multi-modal anchors, and then employ a progressive cross-modal aligning mechanism to facilitate global-level multi-modal alignment based on the aligned local-level anchors. Compared with previous works, the major contributions of our paper are three-fold:

- We present a novel Easy2Hard multi-modal fusion approach by progressively narrowing the modality gap between image and text for MNMT. A two-stage visual-textual fusion strategy is adopted to improve the textual machine translation performance by fully utilizing visual information in seq2seq framework.
- An easy-to-hard visual-textual fusion strategy is adopted to capture multi-modal semantic consistency for image-text pairs. A local-level alignment module is first presented to bridge local-level semantic gap between image and text, and then a cross-modal interactive fusion module is employed for global-level semantic alignment based on the aligned local-level multimodal anchors.
- The extensive experimental results show that our proposed model outperforms other state-of-the-art MNMT approaches and significantly improves machine translation performance on all English-German, English-French and English-Czech translation tasks.

2 BACKGROUND

Early multimodal fusion methods are mainly based on the seq2seq framework of recurrent neural network (RNN), which employs visual features to initialize the hidden state of the RNN encoder-decoder Calixto et al. (2017b); Caglayan et al. (2017); Huang et al. (2016); Zhang et al. (2019), or adopts visual features to enhance the ability of text semantic representation and improve the performance of machine translation Huang et al. (2016). Although these methods improve the performance of machine translation, the visual features are not actually aligned with the textual features. To better align visual and textual semantic features, Calixto et al. (2017a) adopted two modality-specific attention mechanisms for source sentence words and images, respectively, to better align visual and textual features. Caglayan et al. (2016b;a) used a multimodal attention mechanism to simultaneously pay attention to images and their corresponding texts to align visual and textual semantic features; Delbrouck & Dupont (2017a) proposed a local visual attention mechanism that combines Local visual features are aligned with corresponding text features.

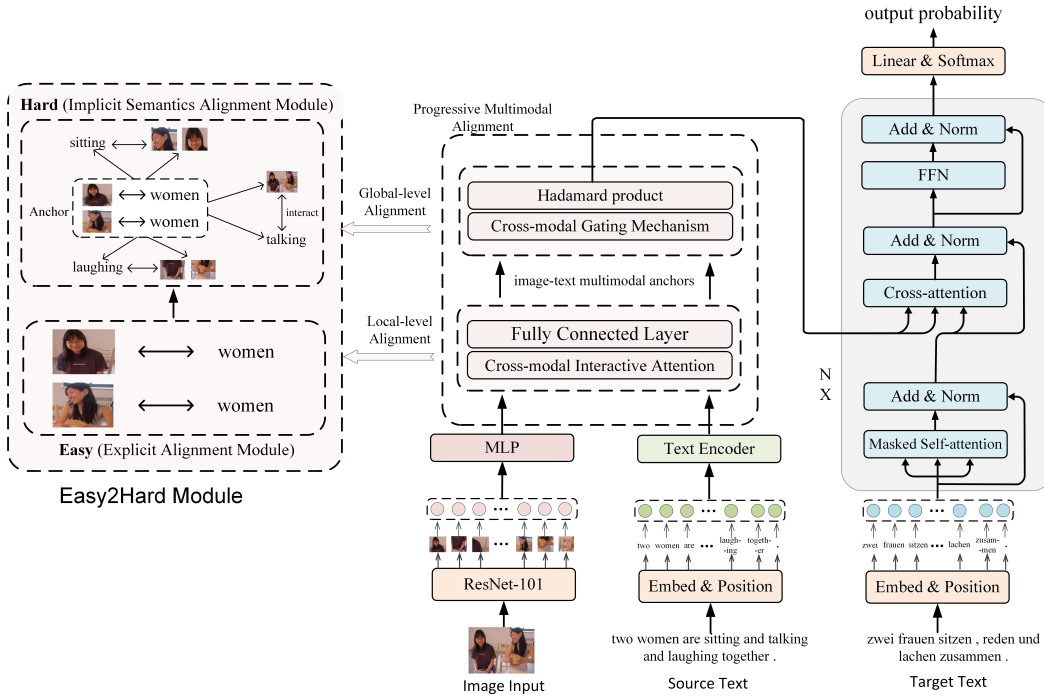


Figure 2: The overview of our proposed model.

With the development of machine translation technology, a multimodal neural machine translation method based on Transformer structure is proposed, and Transformer-based seq2seq framework has achieved significant improvement for MNMT. Yao & Wan (2020) used multi-modal Transformer to align both visual features and textual features; Nishihara et al. (2020) presented a supervised cross-modal attention module to align textual features and visual features; Yin et al. (2020) proposed a graph-based MNMT approach to extract multi-modal features through text-image gating attention mechanism; Zhao et al. (2021) based on a multimodal Transformer, propose a word domain alignment-guided method to establish semantic correlations between textual and visual features; Song et al. (2021) employed a co-attention graph updating module at each Transformer encoder layer to align multi-modal features; Zhao et al. (2022) used object detection features and additional region-related attention mechanisms to fuse visual region features and textual features; Lin et al. (2020) adopted a gating mechanism to fuse visual features extracted by a dynamic context-guided capsule network. Although these methods above improve the performance of machine translation, they do not fully consider the ease of multimodal fusion of image-text.

3 METHODOLOGY

In this section, we introduce our proposed easy2hard multi-modal alignment approach for multi-modal neural machine translation, the framework of our proposed approach is illustrated in Figure 2. Our proposed model is based on the structure of Transformer, which contains four subnetworks, 1) image encoder, 2) source sentence encoder, 3) easy2hard multi-modal fusion module, 4) target sentence decoder.

3.1 SOURCE SENTENCE ENCODER

Without loss of generality, input words are embedded by traditional embedding layer with position embedding. Denote by $S_j = \{S_1^j, \dots, S_n^j\}$ 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 S_j . Formally, the source

sentence representation E_j^S is calculated as follows:

$$E_j^S = \text{Emb}_S(S_j) + \text{PE}_S(S_j) \quad (1)$$

where, Emb_S is the source embedding layer, PE_S is position embedding layer, and $E_j^S \in R^{n \times 128}$.

The text encoder is a traditional multi-head Transformer encoder, and each encoder layer consists of two sub-layers: multi-head self-attention layer and position feed-forward network (FFN) layer. Concretely, the multi-head self-attention module is used to establish word-to-word interconnection with the source text representation as a query/key/value matrix, which can be expressed as,

$$H_{S_j}^l = \text{Multihead}(E_j^S, E_j^S, E_j^S) \quad (2)$$

$$= \text{Concat}(\text{head}_k^1, \dots, \text{head}_k^M) \quad (3)$$

where, M represents the number of heads, $\text{Multihead}(\cdot)$ is the multi-head attention layer, and $l = \{0, \dots, 3\}$ is the Transformer layer index.

The position-wise Feed-Forward (FFN) neural network is employed to update the state at each position of the sequence and get F_{S_j} as follows:

$$F_{S_j} = \text{FFN}(H_{S_j}^l) \quad (4)$$

3.2 IMAGE ENCODER

Image features are extracted by the pre-trained ResNet-101 vision model He et al. (2016), and the visual representation vector E_j^V can be expressed as,

$$E_j^V = \text{Emb}_V(V_j) \quad (5)$$

where, Emb_V is the visual feature extraction layer with Resnet-101, and $E_j^V \in R^{7 \times 7 \times 2048}$.

Then, map the regional features of each image into the same representation space as visual-textual representations, which can be expressed as,

$$F_{V_j} = \text{MLP}(E_j^V) \quad (6)$$

where, $\text{MLP}(\cdot)$ is a multilayer perception, and $F_{V_j} \in R^{49 \times 128}$.

3.3 MULTI-MODAL INTERACTIVE ATTENTION MODULE WITH EASY2HARD STRATEGY

Local-level Alignment: Inspired by Nishihara et al. (2020), we adopt the cross-modal interactive attention mechanism to facilitate visual-textual fusion, and leverage the source representations as query matrix and visual representations as key/value matrix to capture local-level multi-modal semantic consistency,

$$H_j = \text{Multihead-Local}(F_{S_j}, F_{V_j}, F_{V_j}) \quad (7)$$

$$= \sum_{i=1}^m \hat{\alpha}_{fi} (F_{V_{j,i}} \mathbf{W}_1^V) \quad (8)$$

$$\hat{\alpha}_{fi} = \text{softmax} \left(\frac{(F_{S_{j,f}} \mathbf{W}_2^Q)(F_{V_{j,i}} \mathbf{W}_3^K)^T}{\sqrt{d}} \right) \quad (9)$$

where, $\text{Multihead-Local}(\cdot)$ represents the cross-modal attention mechanism of text semantics and image semantics, $\hat{\alpha}_{ni}$ is the similarity weight of text words and image regions, which represents the similarity between the f -th word and the i -th image regions, $f \in \{1, \dots, n\}$, n is the source length of S_j , m is the number of image partition regions, in this paper m is 49. \mathbf{W}_1^V , \mathbf{W}_2^Q , \mathbf{W}_3^K are parameter matrices.

Global-level Alignment: In order to further align and fuse the deep hidden semantic information of images, we take the roughly aligned fine-grained features as multi-modal anchors, and further

interact the text modal semantic and image modal semantic based on the aligned semantic information. Concretely, a novel cross-modal gating strategy is employed to achieve further interaction fusion as follow,

$$\Omega = \text{Sigmoid}(W_{\Omega}(H_j \parallel F_{S_j})) \quad (10)$$

$$\hat{H}_j = F_{S_j} \otimes \Omega \quad (11)$$

where, \otimes is the Hadamard product (outer product), W_{Ω} is the model parameter matrix, \parallel represents the concatenating operation, the image and text features are concatenated in the last dimension.

Then, the multi-modal features is fused by concatenate, which is shown as follows:

$$O_j = \hat{H}_j \parallel F_{S_j} \quad (12)$$

The output of the encoder O_j is finally feed to the decoder for target sentence generation.

3.4 TARGET SENTENCE DECODER

We define $t_j = \{t_1^j, \dots, t_s^j\}$ as the target sentence sequence of the corresponding source sentence S_j , where s is the sentence length of t_j , and the target sentence representation is $E_j^t = \text{Emb}_t(t_j) + \text{PE}_t(t_j)$. As shown in Figure 2, we employ the traditional multi-head Transformer framework as the decoder, and each decoder layer consists of three sub-layers: 1) masked multi-head self-attention layer; 2) cross-language multi-head attention layer; and 3) Feed-forward network layer. There can be expressed as,

$$A_j^l = \text{Multihead}(E_j^t, E_j^t, E_j^t) \quad (13)$$

$$Y_j = \text{Cross-att}(A_j, O_j, O_j) \quad (14)$$

$$F_{d_j} = \text{FFN}(Y_j) \quad (15)$$

$$P = \text{Softmax}(W_p F_{d_j} + b) \quad (16)$$

4 EXPERIMENTS

Experimental Dataset We conduct experiments on En→De, En→Fr and En→Cs tasks of the widely adopted Multi30K benchmark dataset ¹, in which the training and validation sets contains 29,000 and 1014 text-image pairs, respectively. Moreover, we employe four test sets to evaluate proposed MNMT model, 1) the Test2016 test set with 1,000 text-image examples included in Multi30K; 2) the Test2017 test set with 1,000 text-image examples from WMT2017, which contains more difficult source sentences to translate and understand; 3) we also employe ambiguous COCO dataset as out-domain test data, which contains 461 text-image examples with ambiguous words and encourages to adopt image for disambiguation; and 4) the Test2018 test set includes 1,071 examples with many entity words and many low frequency words.

Data Pre-processing We directly adopt the preprocessed sentence pairs by 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, 4-gram BLEU Papineni et al. (2002) metrics, which measures the quality of translations in terms of accuracy and fluency. METEOR Denkowski & Lavie (2014) metrics, which takes into account both precision and recall for translation quality.

4.1 SETTINGS

We conduct our proposed models based on Transformer framework Vaswani et al. (2017), with only stack 4-layer encoder-decoder. Concretely, we set the dimensions of the encoder and decoder hidden

¹<https://github.com/multi30k/dataset>

states at $d_{hidden}=128$, the inner-layer of feed-forward network is set as $d_{ffn}=256$. The learning rate is set to 0.008 for En→De, 0.006 for En→Fr, 0.005 for En→Cs, respectively. 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 for En→De, 0.2 for En→Fr and En→Cs. We use adam optimizer with $\beta_1, \beta_2 = (0.9, 0.98)$. 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 visually verify the advantages of our proposed multimodal neural machine translation model, the paper is compared with the following recent state-of-the-art multimodal neural machine translation models, 1) VAG-NMT Zhou et al. (2018): A background attention mechanism is employed to leverage visual information to enhance model translation performance. 2) DCCN Lin et al. (2020): A Dynamic Context-Guided Capsule Network (DCCN) is proposed to guide visual feature extraction to improve machine translation performance. 3) MNMT+SVA Nishihara et al. (2020): A supervised visual attention mechanism for capturing text-related visual regions for machine translation. 4) OVC+ L_v Wang & Xiong (2021): An object-level visual context semantic framework is constructed to efficiently explore and capture visual information to guide machine translation. 5) WRA-guided Zhao et al. (2021): Based on the multimodal Transformer, a word domain alignment guided method is proposed to establish the semantic correlation between textual and visual features. 6) IO-MMT Song et al. (2021): A relation-aware graph encoder is built to fully exploit the relation between images and source sentences, and an efficient multi-modal reward function is proposed at the target side to improve the visual consistency of translations. 7) DLMulMix Ye & Guo (2022): A novel two-stage interactive multimodal hybrid encoder (DLMulMix) is proposed to extract useful visual features to enhance text-level machine translation.

Further, in order to more fairly demonstrate the superiority and effectiveness of the model proposed in this paper, three most popular multimodal fusion methods are reproduced in this paper on the basis of the same parameter settings and training equipment, 1) Gated Fusion MNMT Li et al. (2021a): An efficient multimodal fusion method to improve machine translation performance by enhancing important information in text. This method is widely used in multimodal neural machine translation and other multimodal tasks in the field of natural language processing. 2) Multimodal self-att Yao & Wan (2020): An image-aware multimodal Transformer model is proposed to extract useful image information to improve machine translation performance. This method mainly connects text features and visual features for multimodal cross attention. 3) Doubly-ATT Arslan et al. (2018): An additional visual attention sub-layer is used between the source-target cross-attention sub-layer and the self-attention sub-layer of the decoder, and the visually evoked attention weights and the source language attention weights are added as dual attention weights.

Model	Multi30K En→De					
	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Existing MNMT Systems						
VAG-NMT Zhou et al. (2018)	-	-	31.6	52.2	28.3	48.0
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 & 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
IO-MMT Song et al. (2021)	41.3	59.2	33.5	52.8	-	-
DLMulMix Ye & Guo (2022)	41.77	58.93	33.07	51.85	29.90	49.09
Our Transformer-Based Systems with Fairseq						
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 & Wan (2020) †	41.50	58.52	32.51	51.33	29.10	48.48
Gated Fusion MNMT Li et al. (2021a) †	41.58	58.88	33.01	51.90	30.04	48.95
Our model	42.84	60.16	35.60	55.00	30.56	50.91

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.

4.3 RESULTS ON THE EN→DE TRANSLATION TASK

Table 1 shows the translation results of our proposed model and other state-of-the-art models on the Ee→De MNMT tasks. Our proposed model outperforms both strong baselines and all the existing MNMT systems on the En-De translation tasks. Concretely, We summarize and contrast the existing MNMT models in the three aspects as follows:

1) *Compare with NMT Baselines*: Our MNMT approach with easy2hard fusion strategy outperforms text-only NMT baselines significantly on BLEU and METEOR evaluation metrics, which enhances about 1–3 points on three test sets. The experimental results show that our proposed multimodal machine translation model can effectively extract image information to enhance machine translation performance.

2) *Compare with Existing MNMT Model*: In order to intuitively show the superiority of our proposed model, compare with the recent SOTA MNMT model, the experimental results show that our proposed model outperforms existing MNMT models, and enhances BLEU and METEOR metrics by 1-2 points on most of the test sets. This demonstrates that our proposed model can better extract visual information to improve machine translation.

3) *Compare with Reproduce Model*: To more fairly demonstrate the superiority of our proposed method, we reproduce three recent multimodal fusion methods based on the same training environment. The experimental results show that the proposed method achieves significant improvements over the reproduce multimodal fusion methods on all evaluation metrics. This proves that the experimental results show can better achieve the fusion of multi-modal image and text to enhance machine translation.

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
DCCN Lin et al. (2020)	61.2	76.4	54.3	70.3	45.4	65.0
OVC+ L_v Wang & 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
IO-MMT Song et al. (2021)	62.5	76.9	54.9	71.7	-	-
DLMulMix Ye & Guo (2022)	62.23	76.85	55.18	73.37	44.42	66.41
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 & 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	63.36	77.29	56.35	72.76	47.04	67.36

Table 2: Comparison results on the En→Fr translation task on the Multi30k dataset. † means to reproduce previous multi-modal fusion method based on our Transformer systems. Best results are highlighted in bold.

4.4 RESULTS ON THE EN→FR TRANSLATION TASK

To explore the robustness of our proposed model, we further conduct En→Fr translation task on our proposed easy2hard MNMT model, machine translation results are listed in Table 2. Similar to the En→De task, the proposed model is compared with existing MNMT models, text-only NMT models and reproduce MNMT models on the En→Fr task. Compared with the baseline model of text-only machine translation, the multimodal machine translation model with image information has achieved excellent results. In addition, compared with existing MNMT models, our proposed model with easy2hard strategy achieves significant improvement on all the evaluation metrics, which is consistent with the results on the En→De translation task. Second, comparing the reproduce SOTA MNMT models on the En→Fr task, the results show that our proposed model outperforms all the reproduce MNMT models, which once again proves the superiority of the proposed model. The above experimental results on the En→Fr translation task once again demonstrate the effectiveness and generality of the proposed method.

Model	En→Cs			
	Test2016		Test2018	
	BLEU	METEOR	BLEU	METEOR
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 & Wan (2020) †	33.12	32.01	28.75	29.51
Gated Fusion MNMT Li et al. (2021a) †	33.77	32.24	29.43	29.41
Our model	35.18	33.39	31.29	30.82

Table 3: Experimental results on the English→Czech (En→Cs) multimodal translation task.

4.5 RESULTS ON THE EN→CS TRANSLATION TASK

To further verify the effectiveness and robustness of the proposed method on different language pairs, we evaluate the model on the En→Cs multimodal translation task. Table 3 presents the BLEU and METEOR values of the reproduced multimodal fusion methods and the proposed multimodal fusion method. As can be observed, our proposed method achieves the best results, achieving +2.48, +3.67 BLEU improvement and +1.05, +1.79 METEOR improvement over the baseline model. Compared with the reproduce method, the proposed method achieves more than +1 point improvement in BLEU value and METEOR value on the two evaluation metrics, and the translation result is significantly improved. The experimental results on the En→Cs multimodal translation task demonstrate that the proposed method is effective and general for different language pairs.

#	Model	Test2016		Test2017		MSCOCO	
		BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
En→De Task							
1	MNMT_reLocal	41.86	59.47	34.10	53.34	30.28	50.09
2	MNMT_reGlobal	41.75	59.63	34.68	54.15	29.47	49.81
3	MNMT_full	42.84	60.16	35.60	55.00	30.56	50.91
En→Fr Task							
4	MNMT_reLocal	62.64	76.41	55.51	72.67	45.90	66.83
5	MNMT_reGlobal	62.90	76.95	55.87	72.41	46.25	67.05
6	MNMT_full	63.36	77.29	56.35	72.76	47.04	67.36

Table 4: Ablation experiments of different components of the model, MNMT_easy is the model that removes the easy-stage strategy, MNMT_hard is the model that removes the hard-stage strategy, and MNMT is complete model.

4.6 ABLATION STUDY

To further investigate the effectiveness of our proposed method, we remove different components of the model for ablation studies on both the En→De translation task and the En→Fr translation task, and the results are reported in Table 4.

1) *Effectiveness of Local-level Alignment*: To demonstrate the effectiveness of our proposed local-level alignment strategy for image and source sentence, we only remove the local-level alignment module (MNMT_reLocal) and do not introduce image-text multi-modal anchor point information. In this variant, note that the global-level alignment strategy is not removed here. Comparing machine translation results in row 1 and row 3 (row 4 and row 6) in Table 4 show that the local-level alignment strategy plays an important role in our proposed model. Further analysis shows that multimodal fusion of global-level alignment image-text without local-level multi-modal information as anchors leads to significant performance degradation, especially on Test2017 on En→De tasks, which contains more difficult source sentences to understand and translate. Our MNMT model benefits from roughly aligned local-level multi-modal anchors, which further demonstrates the effectiveness of our proposed local-level alignment strategy.

2) *Effectiveness of Global-level Alignment*: To demonstrate the importance of the global-level semantic alignment strategy, we conduct an ablation study by removing cross-modal global-level interactive fusion (MNMT_reGlobal). Note that the local-level alignment strategy is not removed here. The experimental results in row 2 (row 5) show that it achieves inferior results compared with MNMT_full model in row 3 (row 6), which demonstrates the importance of local-level alignment





 <p>Source: people are watching others play a game of <u>tennis</u> . MNMT_reLocal: leute sehen zu , wie andere ein spiel <u>spielen</u> . (people are watching others play a game .) MNMT_full: leute schauen anderen bei einem spiel beim <u>tennisspielen</u> zu . (people are watching others play <u>tennis</u> in a game .) Target: zuschauer sehen bei einem <u>tennisspiel</u> zu .</p>	 <p>Source: there is a black car on a <u>race track</u> . MNMT_reGlobal: ein schwarzes auto auf einer <u>rennstrecke</u> . (there is a black car on a <u>racecourse</u> .) MNMT_full: ein schwarzes auto auf einer <u>rennbahn</u> . (there is a black car on a <u>race track</u> .) Target: ein schwarzes auto auf einer <u>rennbahn</u> .</p>
 <p>Source: a man showing a <u>female</u> a large birthday cake . MNMT_reLocal: ein mann zeigt <u>einer großen geburtstagstorte</u> einen großen geburtstagskuchen . (a man shows a <u>big birthday</u> cake to a big birthday cake .) MNMT_full: ein mann zeigt <u>einer frau</u> einen großen geburtstagskuchen . (a man shows a <u>woman</u> a big birthday cake .) Target: ein mann , der <u>einer frau</u> einen großen geburtstagskuchen zeigt .</p>	 <p>Source: a police officer training a <u>black police dog</u> . MNMT_reGlobal: ein polizist trainiert einen <u>schwarzen hund</u> . (a police officer trains a <u>black dog</u> .) MNMT_full: ein polizist trainiert einen <u>schwarzen polizeihund</u> . (a police officer trains a <u>black police dog</u> .) Target: ein polizist trainiert einen <u>schwarzen polizeihund</u> .</p>

Figure 3: (left) Two qualitative translation examples from test set for verify the effectiveness of the local-level alignment strategy. Underscore indicates enhanced translation.(right) Two qualitative translation examples from test set for verify the effectiveness of the global-level alignment strategy. Underscore indicates enhanced translation.

strategy deep interactive and the effectiveness of our proposed cross-modal global-level interactions fusion. Furthermore, further analysis shows that the model translation performance declines when the hard strategy is discarded, indicating that the proposed method can sufficiently extract useful visual information to enhance machine translation.

4.7 CASE STUDY

In order to verify that the proposed method indeed effectively guides the generation of target sequences during the translation process, we verify the effectiveness of the proposed progressive multimodal fusion approach with the easy2hard cross-model alignment method through some qualitative translation examples. Figure 3 (left) shows two translation examples of the En→De test sets selected to demonstrate the effectiveness of the local-level alignment strategy. In the first example, the source word "tennis" is mistranslated as "spielen" by the MNMT_reLocal model, however, the MNMT_full model translated the word basically correctly. In the second example, the source phrase "a female" is mistranslated as "einer großen geburtstagstorte" by the MNMT_reLocal model, but our MNMT_full model correctly translates this phrase as "einer frau". The above two examples illustrate that our proposed local-level fusion strategy is beneficial for roughly aligning visual object-level and textual entity-level information.

Two translation examples selected from En→De test set demonstrate the effectiveness of our global-level alignment strategy, as shown in Figure 3 (right). In the first example, the source phrase "race track" is mistranslated as "rennstrecke" by the MNMT_reGlobal model. However, our model accurately translates it, the underlying reason is that MNMT_full model interacts with image-text semantic global information such as "black car" to predict the target word. In the second example, the source phrase "black police dog" is mistranslated as "schwarzen hund" by the MNMT_reGlobal model, we can observe that the model is only successful in roughly aligning image and text information. With the addition of a global-level fusion module, our MNMT_full model further interacts with image-text global semantic information, such as the police information is captured from the global information, and "black police dog" is successfully translated into "schwarzen polizeihund". The above two examples demonstrate that our proposed global-level alignment strategy can interact with deep implicit semantic information to improve translation performance.

5 CONCLUSION

In this paper, we propose a novel progressive multimodal fusion approach with the easy2hard cross-model alignment strategy by fully exploiting the visual information from images for multimodal neural machine translation. Concretely, we first roughly align local-level object information as multi-modal anchors, and then employ a progressive cross-modal aligning mechanism to facilitate global-level multi-modal alignment based on the aligned local-level anchors. The experimental results on three benchmark translation tasks demonstrate that superiority of our proposed easy2hard cross-modal alignment strategy, and achieve the new state-of-the-art result on Test2016, Test2017, Test2017mscoco, Test2018 test datasets. Furthermore, ablation experiments verify the effectiveness of our proposed two-stage progressive alignment method, and the analysis shows that the local-level fusion strategy can effectively and efficiently align entity semantic information with rough alignment.

REFERENCES

- Hasan Sait Arslan, Mark Fishel, and Gholamreza Anbarjafari. Doubly attentive transformer machine translation. *arXiv:1807.11605*, 2018.
- Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In *THIRD CONFERENCE ON MACHINE TRANSLATION (WMT18)*, volume 2, pp. 308–327, 2018.
- Ozan Caglayan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, and Joost Van de Weijer. Does multimodality help human and machine for translation and image captioning? *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 627–633, Association for Computational Linguistics*, 2016a.
- Ozan Caglayan, Loïc Barrault, and Fethi Bougares. Multimodal attention for neural machine translation. *arXiv:1609.03976*, <http://arxiv.org/abs/1609.03976>, 2016b.
- Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost Van de Weijer. Lium-cvc submissions for wmt17 multimodal translation task. *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7–8, 2017*, pp. 432–439. doi:10.18653/v1/w17-4746., 2017.
- Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 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)*, pp. 4159–4170, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1422. URL <https://aclanthology.org/N19-1422>.
- Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. Cross-lingual visual pre-training for multimodal machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1317–1324, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.112. URL <https://aclanthology.org/2021.eacl-main.112>.
- Iacer Calixto, Qun Liu, and Nick Campbell. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1913–1924, Vancouver, Canada, July 2017a. Association for Computational Linguistics. doi: 10.18653/v1/P17-1175. URL <https://aclanthology.org/P17-1175>.
- Iacer Calixto, Qun Liu, and Nick Campbell. 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, 2017b.
- Iacer Calixto, Miguel Rios, and Wilker Aziz. Latent variable model for multi-modal translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6392–6405, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1642. URL <https://aclanthology.org/P19-1642>.
- Shiyu Chen, Yawen Zeng, Da Cao, and Shaofei Lu. Video-guided machine translation via dual-level back-translation. *Knowledge-Based Systems*, 245:108598, 2022. ISSN 0950-7051. doi: <https://doi.org/10.1016/j.knsys.2022.108598>. URL <https://www.sciencedirect.com/science/article/pii/S0950705122002684>.
- Jean-Benoît Delbrouck and Stéphane Dupont. 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*, pp. 910–919, Copenhagen, Denmark, September 2017a. Association for Computational Linguistics. doi: 10.18653/v1/D17-1095. URL <https://aclanthology.org/D17-1095>.

- Jean-Benoit Delbrouck and Stephane Dupont. Multimodal compact bilinear pooling for multimodal neural machine translation. *arXiv preprint arXiv:1703.08084*, 2017b.
- Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pp. 376–380, 2014.
- Desmond Elliott. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2974–2978, Brussels, Belgium, October–November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1329. URL <https://aclanthology.org/D18-1329>.
- Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 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*, 2017.
- Baban Gain, Dibyanayan Bandyopadhyay, and Asif Ekbal. Experiences of adapting multimodal machine translation techniques for hindi. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMLRL 2021)*, pp. 40–44, 2021.
- Longchao Gong, Yan Li, Junjun Guo, Zhengtao Yu, and Shengxiang Gao. Enhancing low-resource neural machine translation with syntax-graph guided self-attention. *Knowledge-Based Systems*, 246:108615, 2022. ISSN 0950-7051. doi: <https://doi.org/10.1016/j.knosys.2022.108615>. URL <https://www.sciencedirect.com/science/article/pii/S0950705122002763>.
- Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. The MeMAD submission to the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pp. 603–611, Belgium, Brussels, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6439. URL <https://aclanthology.org/W18-6439>.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 639–645, 2016.
- Julia Ive, Pranava Madhyastha, and Lucia Specia. Distilling translations with visual awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6525–6538, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1653. URL <https://aclanthology.org/P19-1653>.
- Soonmo Kwon, Byung-Hyun Go, and Jong-Hyeok Lee. A text-based visual context modulation neural model for multimodal machine translation. *Pattern Recognition Letters*, 136:212–218, 2020.
- Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. On vision features in multimodal machine translation. *arXiv preprint arXiv:2203.09173*, 2022.
- Jiaoda Li, Duygu Ataman, and Rico Sennrich. Vision matters when it should: Sanity checking multimodal machine translation models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8556–8562, Online and Punta Cana, Dominican Republic, November 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.673. URL <https://aclanthology.org/2021.emnlp-main.673>.
- Jiaoda Li, Duygu Ataman, and Rico Sennrich. 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*, 2021b.

- Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1320–1329, 2020.
- Tetsuro Nishihara, Akihiro Tamura, Takashi Ninomiya, Yutaro Omote, and Hideki Nakayama. Supervised visual attention for multimodal neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4304–4314, 2020.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.
- Yuqing Song, Shizhe Chen, Qin Jin, Wei Luo, Jun Xie, and Fei Huang. Enhancing neural machine translation with dual-side multimodal awareness. *IEEE Transactions on Multimedia*, 2021.
- Jinsong Su, Biao Zhang, Deyi Xiong, Yang Liu, and Min Zhang. Alignment-consistent recursive neural networks for bilingual phrase embeddings. *Knowledge-Based Systems*, 156:1–11, 2018. ISSN 0950-7051. doi: <https://doi.org/10.1016/j.knosys.2018.05.003>. URL <https://www.sciencedirect.com/science/article/pii/S0950705118302119>.
- Jinsong Su, Jinchang Chen, Hui Jiang, Chulun Zhou, Huan Lin, Yubin Ge, Qingqiang Wu, and Yongxuan Lai. Multi-modal neural machine translation with deep semantic interactions. *Information Sciences*, 554:47–60, 2021.
- Hiroki Takushima, Akihiro Tamura, Takashi Ninomiya, and Hideki Nakayama. 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)*, 2019.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.
- Dexin Wang and Deyi Xiong. Efficient object-level visual context modeling for multimodal machine translation: Masking irrelevant objects helps grounding. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, pp. 2–9, 2021.
- Changxing Wu, Chaowen Hu, Ruochen Li, Hongyu Lin, and Jinsong Su. Hierarchical multi-task learning with crf for implicit discourse relation recognition. *Knowledge-Based Systems*, 195:105637, 2020. ISSN 0950-7051. doi: <https://doi.org/10.1016/j.knosys.2020.105637>. URL <https://www.sciencedirect.com/science/article/pii/S0950705120300952>.
- Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 6153–6166 August 1–6, 2021.*, 2021.
- Shaowei Yao and Xiaojun Wan. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4346–4350, 2020.
- Kong Yawei and Kai Fan. Probing multi-modal machine translation with pre-trained language model. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3689–3699, 2021.
- Junjie Ye and Junjun Guo. Dual-level interactive multimodal-mixup encoder for multi-modal neural machine translation. *Applied Intelligence*, pp. 1–10, 2022.
- Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 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, 2020.*

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*, 2019.

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

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

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. A visual attention grounding neural model for multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3643–3653, Brussels, Belgium, October–November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1400. URL <https://aclanthology.org/D18-1400>.

A APPENDIX

You may include other additional sections here.