Low-resource Neural Machine Translation with Cross-modal Alignment

Zhe Yang[1,2], **Qingkai Fang**[1,2], **Yang Feng**[1,2*]
[1] Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)
[2] University of Chinese Academy of Sciences, Beijing, China
{yangzhe22s1,fangqingkai21b,fengyang}@ict.ac.cn

Abstract

How to achieve neural machine translation with limited parallel data? Existing techniques often rely on large-scale monolingual corpora, which is impractical for some low-resource languages. In this paper, we turn to connect several low-resource languages to a particular high-resource one by additional visual modality. Specifically, we propose a cross-modal contrastive learning method to learn a shared space for all languages, where both a coarse-grained sentence-level objective and a fine-grained token-level one are introduced. Experimental results and further analysis show that our method can effectively learn the cross-modal and cross-lingual alignment with a small amount of image-text pairs and achieves significant improvements over the text-only baseline under both zero-shot and few-shot scenarios. Our code could be found at https://github.com/ictnlp/LNMT-CA.

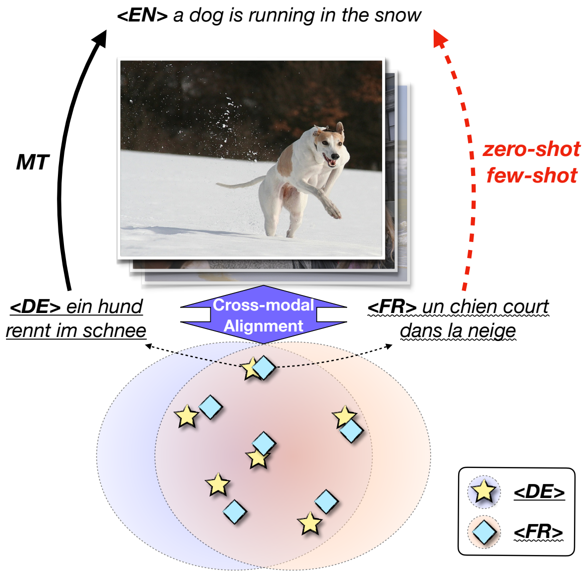


Figure 1: We aim at realizing zero-shot and few-shot machine translation for the low-resource language. Different languages with the same meanings are projected to a shared space by cross-modal alignment.

1 Introduction

Neural machine translation (NMT) has shown excellent performance and becomes the dominant paradigm of machine translation. However, NMT is a data-driven approach, which requires a large amount of parallel data. When the data is insufficient, it is impractical to train a reasonable NMT model. Unfortunately, there are many languages in the world for which sufficient training data is not available, and sometimes there is no parallel data at all. Therefore, the translation of low-resource languages is a vital challenge for NMT.

In recent years, researchers have attempted to improve the performance of NMT for low-resource languages. Lample et al. (2018a) proposed an unsupervised approach to learn weak mappings between languages with large amount of monolingual data (>1M), which is also costly for low-resource languages. Liu et al. (2020); Lin et al. (2020b); Pan et al. (2021); Gu and Feng (2022) proposed multilingual NMT models, which learn a shared space of multiple languages to achieve translations between languages that appear in the training set but do not have the corresponding parallel data. However, they still require auxiliary parallel data of source and target languages along with many other languages, which is still infeasible for low-resource languages.

In recent years, with increasing attention of multi-modal tasks, resource of image-text pairs have become more abundant. Inspired by recent efforts on cross-modal alignment (Radford et al., 2021; Li et al., 2021; Fang et al., 2022), in this paper, we propose a cross-modal contrastive learning method, which align different languages with images as the pivot to enable zero-shot and few-shot translations for low-resource languages. With parallel sentence pairs between one high-

* Corresponding author: Yang Feng.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10134–10146
December 7-11, 2022

resource auxiliary language and the target language, we can achieve the translation from low-resource languages to the target language only by obtaining small amounts of image-text pairs (<0.1M) for those languages. The parallel sentence pairs are used to learn the mapping from the high-resource language to the target language, and the image-text pairs are used to learn a shared space for all languages through cross-modal alignment. With images as the pivot, the mapping from the low-resource languages to the target language are learned, thus achieving zero-shot translation without any parallel sentence pairs between them. As shown in Figure 1, the high-resource language German and the low-resource language French are brought together by cross-modal alignment, which transfers the translation ability from DE→EN to FR→EN. Experiments and analysis show that our method consistently outperforms the baseline under both zero-shot and few-shot scenarios. Furthermore, our method can effectively realize cross-modal and cross-lingual alignment.

2 Method

In this section, we present our proposed cross-modal contrastive learning method, which includes both sentence-level and token-level objectives.

2.1 Task Definition

Our goal is to achieve zero-shot or few-shot translation from T low-resource languages $L_1, L_2, ..., L_T$ to the target language L_y with the help of a particular high-resource language $\widehat{L}$. For the high-resource language $\widehat{L}$, there are triples of data $\mathcal{D}_{\widehat{L}} = \{(\mathbf{i}, \mathbf{x}, \mathbf{y})\}$, where $\mathbf{i}$ is the image and $\mathbf{x}$ and $\mathbf{y}$ are the descriptions in $\widehat{L}$ and L_y respectively. For each low-resource language L_i, only paired data $\mathcal{D}_{L_i} = \{(\mathbf{i}, \mathbf{x})\}$ are available. Note that different languages never share the same images.

2.2 Model Framework

As shown in Figure 2, our model consists of four sub-modules: *image encoder*, *source encoder*, *target decoder* and *contrastive module*.

We use Vision Transformer (ViT) (Dosovitskiy et al., 2021) as the *image encoder* to extract visual features. ViT first splits the image into several patches, and then feed the sequence of embed patches with a special `[class]` token into Transformer (Vaswani et al., 2017). Finally, the image is encoded as a sequence of vectors $\mathbf{v} = (v_0, v_1, ..., v_m)$, where v_0 is the representation of `[class]` token which can be regarded as the global representation of the image, and $\mathbf{v}^p = (v_1, ..., v_m)$ are the patch-level representations. In next sections, we use v_0 for sentence-level contrastive learning and $\mathbf{v}^p$ for token-level contrastive learning.

The *source encoder* consists of N Transformer encoder layers, which is shared across all languages ($L_{1...T}$ and $\widehat{L}$). For the input sentence $\mathbf{x} = (x_1, ..., x_n)$, the output of *source encoder* is denoted as $\mathbf{w} = (w_1, ..., w_n)$. The *target decoder* consists of N Transformer decoder layers. For the sentence pairs $(\mathbf{x}, \mathbf{y})$, the cross-entropy loss is defined as:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{|\mathbf{y}|} \log p(y_i^* | \mathbf{y}_{<i}, \mathbf{x}). \tag{1}$$

The *contrastive module* aims to align the output of *image encoder* and *source encoder*, which contains both sentence-level and token-level parts. We will introduce them in Section 2.3 and 2.4.

2.3 Sentence-level Contrastive Learning

We start with the sentence-level contrastive learning objective, which aims at learning coarse alignment between image and text.

Contrastive Learning The idea of contrastive learning (Sohn, 2016) is to make the representations of corresponding pairs closer and, on the contrary, to make the irrelevant pairs farther.

Given two sets $\mathbf{X} = \{x_i\}_{i=1}^M$ and $\mathbf{Y} = \{y_i\}_{i=1}^M$, for each x_i, the positive example is (x_i, y_i) and the remaining $M-1$ irrelevant pairs $(x_i, y_j)(i \neq j)$ are considered as negative examples. The contrastive loss between $\mathbf{X}$ and $\mathbf{Y}$ is defined as:

$$\mathcal{L}_{\text{ctr}}(\mathbf{X}, \mathbf{Y}) = -\sum_{i=1}^{M} \log \frac{\exp(s(x_i, y_i)/\tau)}{\sum_{j=1}^{M} \exp(s(x_i, y_j)/\tau)}, \tag{2}$$

where $s()$ is the cosine similarity function $s(a, b) = a^\top b / \|a\| \|b\|$. τ is the temperature hyper-parameter to control the strength of penalties on hard negative samples (Wang and Liu, 2021).

Sentence-level Contrast Sentence-level contrastive learning aims to align the sentence-level representations across modalities, which are de-

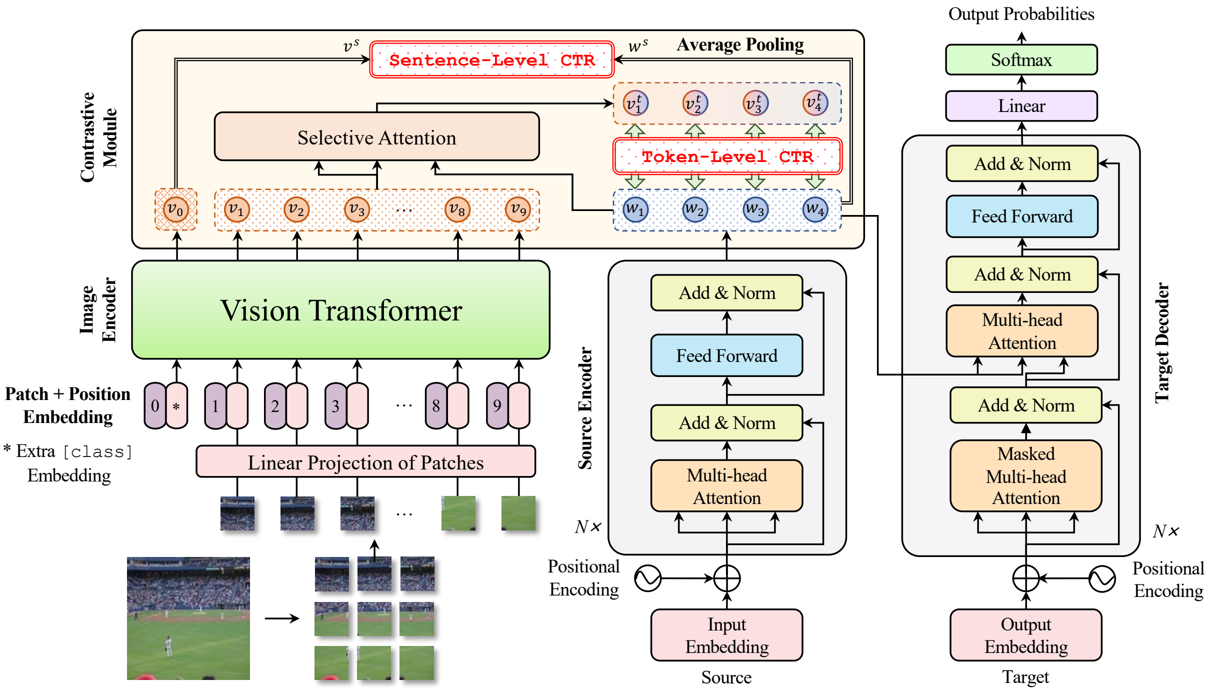


Figure 2: Overview of our proposed model.

fined as follows:

$$w^s = \frac{1}{n}\sum_{i=1}^{n} w_i, \tag{3}$$

$$v^s = v_0. \tag{4}$$

We then calculate the contrastive loss within a batch of size B, whose textual representations and visual representations are $\mathbf{W}^s = \{w_1^s, ...w_B^s\}$ and $\mathbf{V}^s = \{v_1^s, ..., v_B^s\}$, respectively. The corresponding pairs of images and captions (w_i^s, v_i^s) are positive examples, and other pairs $(w_i^s, v_j^s)(i \neq j)$ are considered as negative examples. Finally, the loss function of sentence-level contrastive learning is defined as follows:

$$\mathcal{L}_{\text{s-ctr}}(\mathbf{W}^s, \mathbf{V}^s) = \mathcal{L}_{\text{ctr}}(\mathbf{W}^s, \mathbf{V}^s) + \mathcal{L}_{\text{ctr}}(\mathbf{V}^s, \mathbf{W}^s). \tag{5}$$

Since we have image-text pairs in different languages within a batch, we first separate the batch into several mini-batches according to the language, and then calculate the contrastive loss for every language respectively. It is worth mentioning that we also calculate contrastive loss for target language L_y with paired data $\{(\mathbf{i}, \mathbf{y})\}$ in $D_{\widehat{L}}$. We will analyze its effect in Section 4.3.

2.4 Token-level Contrastive Learning

Though sentence-level contrastive learning can learn coarse-grained alignment between modalities, it may ignore some detailed information, which is crucial for predicting translations. To achieve better alignment between modalities, we propose token-level contrastive learning to learn fine-grained correspondences between images and text.

Selective Attention To model the correlations between image patches and words, we use selective attention (Li et al., 2022) to learn the patch-level contribution of images. For patch-level visual representations $\mathbf{v}^p = (v_1, ...v_m)$ and word-level textual representations $\mathbf{w} = (w_1, ..., w_n)$, the query, key and value of selective attention are $\mathbf{w}, \mathbf{v}^p, \mathbf{v}^p$, respectively:

$$\mathbf{v}^t = \text{Softmax}\left(\frac{(W_Q \cdot \mathbf{w})(W_K \cdot \mathbf{v}^p)^\top}{\sqrt{d_k}}\right)(W_V \cdot \mathbf{v}^p), \tag{6}$$

where W_Q, W_K and W_V are learnable matrix parameters.

Token-level Contrast After the selective attention, we obtain two sequences $\mathbf{w} = (w_1, ..., w_n)$ and $\mathbf{v}^t = (v_1^t, ..., v_n^t)$ with the same length of n. We then calculate the token-level contrastive loss within each pair of sequences. Tokens with same index (w_i, v_i^t) are positive examples, and other pairs of tokens $(w_i, v_j^t)(i \neq j)$ are negative examples. The token-level contrastive loss is as follows:

$$\mathcal{L}_{\text{t-ctr}}(\mathbf{w}, \mathbf{v}^t) = \mathcal{L}_{\text{ctr}}(\mathbf{w}, \mathbf{v}^t) + \mathcal{L}_{\text{ctr}}(\mathbf{v}^t, \mathbf{w}). \tag{7}$$

The token-level contrastive loss of all image-text pairs will be summed together.

2.5 Coarse-to-fine Training Strategy

To combine sentence-level and token-level objectives together, we propose a 2-stage *coarse-to-fine training strategy*, the intuition behind which is to first learn coarse-grained alignment through the sentence-level objective, and then add fine-grained alignment with the token-level objective.

Stage 1 For the first stage of training, the model is trained with cross-entropy loss of the high-resource language $\widehat{L}$ and sentence-level contrastive loss of all languages (including target language L_y):

$$\begin{aligned}\mathcal{L}_{\text{coarse}} = {} & \mathbb{E}_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_{\widehat{L}}}\mathcal{L}_{\text{CE}}(\mathbf{x},\mathbf{y}) \\ & + \lambda_s\mathbb{E}_{(\mathbf{i},\mathbf{y})\in\mathcal{D}_{\widehat{L}}}\mathcal{L}_{\text{s}-\text{ctr}}(\mathbf{i},\mathbf{y}) \\ & + \lambda_s\mathbb{E}_{(\mathbf{i},\mathbf{x})\in\mathcal{D}_{\widehat{L}}}\mathcal{L}_{\text{s}-\text{ctr}}(\mathbf{i},\mathbf{x}) \\ & + \lambda_s\sum_{i=1}^{T}\mathbb{E}_{(\mathbf{i},\mathbf{x})\in\mathcal{D}_{L_i}}\mathcal{L}_{\text{s}-\text{ctr}}(\mathbf{i},\mathbf{x}),\end{aligned} \tag{8}$$

where λ_s is the weight hyper-parameter of sentence-level contrastive loss.

Stage 2 For the second stage of training, we add the token-level contrastive loss to Eq. 8, which can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{fine}} = {} & \mathcal{L}_{\text{coarse}} \\ & + \lambda_t\mathbb{E}_{(\mathbf{i},\mathbf{y})\in\mathcal{D}_{\widehat{L}}}\mathcal{L}_{\text{t}-\text{ctr}}(\mathbf{i},\mathbf{y}) \\ & + \lambda_t\mathbb{E}_{(\mathbf{i},\mathbf{x})\in\mathcal{D}_{\widehat{L}}}\mathcal{L}_{\text{t}-\text{ctr}}(\mathbf{i},\mathbf{x}) \\ & + \lambda_t\sum_{i=1}^{T}\mathbb{E}_{(\mathbf{i},\mathbf{x})\in\mathcal{D}_{L_i}}\mathcal{L}_{\text{t}-\text{ctr}}(\mathbf{i},\mathbf{x}),\end{aligned} \tag{9}$$

where λ_t is the weight hyper-parameter of token-level contrastive loss.

Zero-shot and Few-shot Translation After 2-stage training with contrastive loss, we can directly evaluate the performance of the trained model on zero-shot translation. Furthermore, we can use small amount of additional parallel data of low-resource languages $\mathcal{D}_L = \{(\mathbf{x},\mathbf{y})\}$ to finetune the model, and then evaluate the performance on few-shot translation. During finetuning, only cross-entropy loss is used:

$$\mathcal{L}_{\text{finetune}} = \mathbb{E}_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_L}\mathcal{L}_{\text{CE}}(\mathbf{x},\mathbf{y}). \tag{10}$$

Directions	Multi30K	MsCOCO	VizWiz	Total
DE→EN	10,000	40,000	10,136	60,136
FR→EN	10,000	40,000	10,136	60,136
CS→EN	9,000	41,000	10,136	60,136

Table 1: Detailed dataset statistics.

3 Experiments

3.1 Datasets

In our experiments, we select German (DE) as the high-resource language and English (EN) as the target language. We choose French (FR) and Czech (CS) as two low-resource languages and test the performance of FR→EN and CS→EN on zero-shot and few-shot translation. Due to the scarcity of image-text pairs in German, French, and Czech, we create pseudo data with machine translation models from two image captioning datasets in English.

Multi30K Multi30K (Elliott et al., 2016) dataset contains images with annotations in four languages: English, German, French, and Czech. The training and validation sets consist of 29,000 and 1,014 instances, respectively. We evaluate our model on Test2016, Test2017, and MsCOCO test sets, which contain 1,000, 1,000, and 456[1] instances. For Czech→English task, only Test2016 is available.

MsCOCO MsCOCO (Lin et al., 2014) dataset contains images with English captions. We use the Captioning 2015 set for our experiments. After filtering out the unannotated images, there are 121,000 image-text pairs in total.

VizWiz VizWiz (Gurari et al., 2020) dataset also contains images with English captions. There are 30,408 image-text pairs in total.

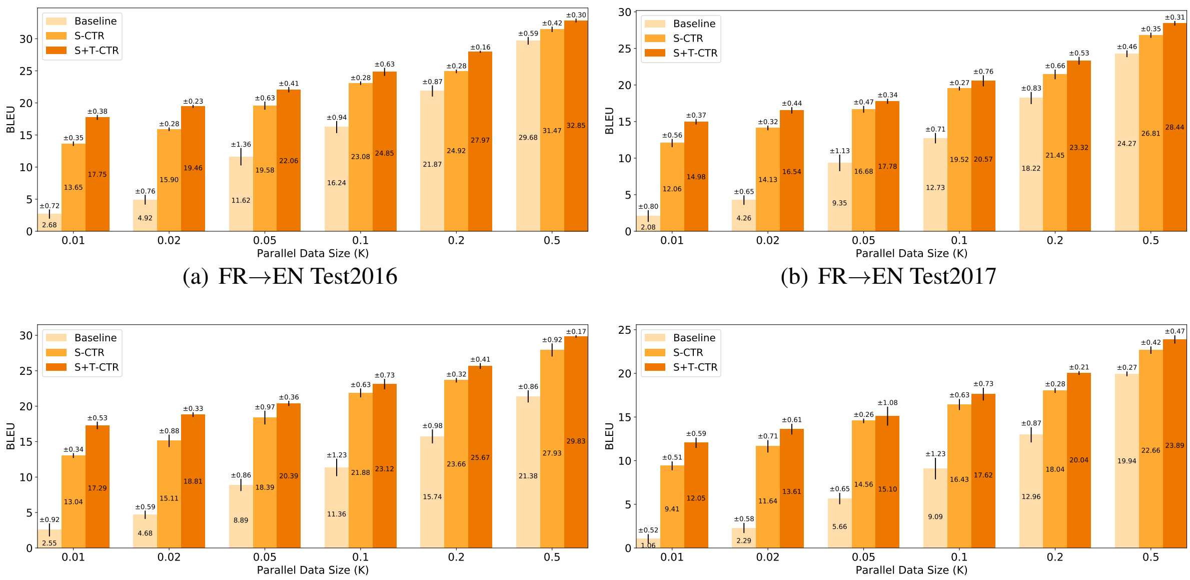
Pseudo Data Since the MsCOCO and VizWiz datasets only have English captions of images. We use pretrained machine translation models to translate English captions into German, French and Czech. The detailed information of the machine translation models can be seen in Appendix A.

Dataset Composition After creating the pseudo data, we divide the above three datasets into three equal parts for DE→EN, FR→EN, and CS→EN, respectively. As shown in Table 1, each source language has 60,136 image-text pairs with annotations

[1] 5 sentences are removed because they appear in the MsCOCO dataset, which is part of our training set.

Models	FR→EN Test2016	FR→EN Test2017	FR→EN MsCOCO	CS→EN Test2016	Average
Baseline	0.30	0.14	0.29	0.09	0.21
S-CTR	8.95	7.88	9.32	7.23	8.35
S+T-CTR	**17.76**	**14.74**	**16.97**	**13.58**	**15.76**

Table 2: BLEU scores of FR→EN and CS→EN on zero-shot translation.



(a) FR→EN Test2016 (b) FR→EN Test2017

(c) FR→EN MsCOCO (d) CS→EN Test2016

Figure 3: BLEU scores of FR→EN and CS→EN on few-shot translation.

in its own language, which are used for cross-modal contrastive learning. At the same time, the 60,136 German→English sentence pairs are used for training of translation task. All sentences are segmented into subword units using byte-pair encoding (BPE) (Sennrich et al., 2016). The vocabulary is shared for all source languages and the target language, with a size of 18K.

3.2 System Settings

We use vision transformer in pre-trained CLIP (Radford et al., 2021) model as the *image encoder*. The patch size is 16×16, and the resolution size is 224. The sequence length is 50, which contains a special `[class]` token and 49 feature tokens. The *source encoder* and *target decoder* are based on Transformer (Vaswani et al., 2017) architecture. Both the encoder and decoder have $N = 6$ layers. The number of attention heads is set to 4. The dropout is set to 0.3, and the value of label smoothing is 0.1. For training, we use Adam optimizer (Kingma and Ba, 2015) and 2000 warm-up updates. The learning rate is 5e-4. Each batch contains up to 16K tokens. We train the model for up to 70 epochs. For our 2-stage training strategy, the first half of training is Stage 1, and the rest is Stage 2.

For sentence-level contrastive learning, the temperature hyper-parameter τ_s is set to 0.007 and the weight hyper-parameter λ_s is set to 5. For token-level contrastive learning, τ_t is 0.1 and λ_t is 1.

For evaluation, we average the last 5 checkpoints and use beam search with a beam size of 5. We use sacreBLEU[2] (Post, 2018) to compute the BLEU (Papineni et al., 2002) scores on detokenized instances[3]. For few-shot translation, we randomly sample 5 groups of parallel data from the training set of Multi30K and report the means and standard deviations. All experiments are done on 4 TITAN Xp GPUs. We implement our system based on *fairseq*[4] (Ott et al., 2019).

[2] `https://github.com/mjpost/sacrebleu`

[3] sacreBLEU signature: nrefs:1 | bs:1000 | seed:12345 | case:lc | eff:no | tok:13a | smooth:exp | version:2.0.0

[4] `https://github.com/pytorch/fairseq`

Figure 4: Attention maps of the selective attention module of two cases.

3.3 Baseline Systems

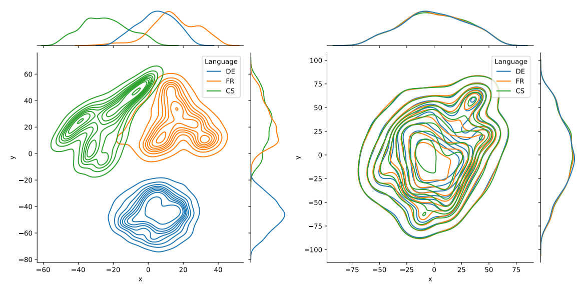
Our baseline is text-only Transformer trained with DE→EN sentence pairs. For zero-shot translation, we directly evaluate the baseline model. For few-shot translation, we finetune the baseline model with the same parallel corpus in low-resource languages as our model. All the configurations of the baseline are the same as our model.

3.4 Results

We evaluate the baseline, our model with only sentence-level contrastive loss (**S-CTR**), and our model with both sentence-level and token-level contrastive loss (**S+T-CTR**) under zero-shot and few-shot scenarios.

Zero-shot Translation Table 2 shows the results on zero-shot translation. The baseline without contrastive learning does not have the capability of zero-shot translation. On the contrary, S-CTR and S+T-CTR gain significantly improvements over the baseline. Compared with S-CTR, the S+T-CTR model has a further improvement of 7.41 BLEU score on average, which proves that more fine-grained alignment can significantly improve the performance on zero-shot machine translation.

Few-shot Translation Figure 3 shows the results on few-shot translation on four test sets. The S+T-CTR model consistently outperforms the baseline and the S-CTR model under different amounts of parallel data, demonstrating the effectiveness of our method in few-shot scenarios.



(a) Baseline (b) S+T-CTR model

Figure 5: Visualization of source representations for DE, FR, and CS under the zero-shot scenario. (a) baseline. (b) S+T-CTR model. Sentences are from Multi30K Test2016 sets of DE→EN, FR→EN, and CS→EN.

Models	R@1↑	R@5↑	R@10↑
Baseline	0.2	0.8	1.3
S-CTR	34.4	65.0	75.4
S+T-CTR	**36.3**	**66.5**	**76.1**

Table 3: Text-to-image retrieval on FR→EN Test2016.

4 Analysis

4.1 Cross-modal Alignment

The main idea of our method is to align multilingual text and images in their representation space. To verify this alignment, we conduct the text-to-image retrieval experiment and visualize the attention map of the selective attention module.

Text-to-image Retrieval Text-to-image retrieval means finding the top-K nearest images to the text. We compute the Recall@K score for $K = 1, 5, 10$. As shown in Table 3, S-CTR gains a substantial 34.2/64.2/74.1% increase in R@1/5/10 over the baseline, which proves the effectiveness of contrastive learning for cross-modal alignment. In addition, S+T-CTR gains an extra 1.9/1.5/0.7% increase in R@1/5/10, proving that the fine-grained learning objective enables better alignment.

Attention Maps To further verify the effect of token-level contrastive learning for cross-modal alignment, we extract attention maps of the selective attention module. Figure 4 demonstrates that the selective attention module successfully notices the semantically related areas. For example, the French word "gens" (means "people") corresponds to the three people and the word "maison" (means "roof") corresponds to the roof area.

4.2 Cross-lingual Alignment

Section 4.1 analyses the effectiveness of contrastive learning on cross-modal alignment. However, our ultimate goal is to achieve cross-lingual alignment through cross-modal alignment, which means to learn a shared space for all languages.

To analyze, we compare the baseline and S+T-CTR model under the zero-shot scenario, which means no FR→EN or CS→EN parallel data is available. We average the output of the *source encoder* and use T-SNE (Laurens and Hinton, 2008) to reduce the dimension into two for visualization. As shown in Figure 5, without contrastive learning, there is a clear distinction between different source languages. On the contrary, with contrastive learning, the representations of three languages have obviously overlapped, which proves that our method learned good cross-lingual alignment.

4.3 Ablation Studies

Target Language Contrast The target language is generally isolated from the source language in standard machine translation. However, we found that adding the target language into contrastive learning is effective. As shown in Table 4, models without contrastive learning of the target language have a significant drop in BLEU score under both zero-shot and few-shot situations. We conclude that contrastive learning of the target language can help establish connections between source and target languages, which will be beneficial for translation.

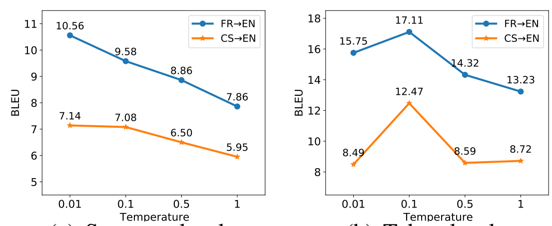
Contrastive Loss vs. L2 Loss Contrastive loss is not the only way to draw the distance between modalities. We try to replace the contrastive loss

Models	Target	FR→EN ZS	FR→EN FS100	CS→EN ZS	CS→EN FS100
Baseline	-	0.24	13.44	0.09	9.10
S-CTR	×	7.81	12.55	6.93	8.67
	✓	8.71	21.49	7.23	16.43
S+T-CTR	×	14.47	13.16	10.97	8.24
	✓	**16.49**	**22.85**	**13.58**	**17.62**

Table 4: Ablation study on contrastive learning of the target language. ZS means zero-shot translation, FS100 means few-shot translation with 100 parallel sentences.

Models	Loss	FR→EN ZS	FR→EN FS100	CS→EN ZS	CS→EN FS100
Baseline	-	0.24	13.44	0.09	9.10
S-level	L2	8.45	19.60	6.83	15.05
	CTR	8.71	21.49	7.23	16.43
S+T-level	L2	6.02	15.61	5.94	12.76
	CTR	**16.49**	**22.85**	**13.58**	**17.62**

Table 5: BLEU scores of models with L2 loss and contrastive loss. ZS means zero-shot translation, FS100 means few-shot translation with 100 parallel sentences.



(a) Sentence-level (b) Token-level

Figure 6: BLEU scores on Multi30K validation set against different temperatures for sentence-level and token-level contrastive learning.

with L2 loss:

$$\mathcal{L}_{L2} = \sum_{i=1}^{M} \|x_i - y_i\|^2. \tag{11}$$

As shown in Table 5, the contrastive loss performs better than the L2 loss. We believe it is because the contrastive loss can not only bring the corresponding pairs closer but also push the irrelevant pairs farther with negative examples.

4.4 Temperature Hyper-parameter

The temperature τ is an important hyper-parameter in contrastive learning. A lower temperature can help the model distinguish positive example from negative ones. Here we choose 0.01, 0.1, 0.5 and 1 for experiments. Figure 6 shows the BLEU scores against different temperatures on validation set.

For sentence-level contrastive learning, we observe that lower temperatures obtain better results. We try to choose the temperature as low as possible. However, a temperature lower than 0.007 may lead to gradient explosion. So we finally select $\tau_s = 0.007$. For token-level contrastive learning, we found $\tau_t = 0.1$ achieved best results on validation set. We think it is because that tokens in a sentence should not be excessively distinguished.

4.5 Case study

In this section, we make a qualitative analysis with several examples. Table 6 shows the references and translation results of different models. First, we compare S-CTR and S+T-CTR under the zero-shot scenario. In Case 1, "two trees" have not been translated by the S-CTR model, while the S+T-CTR model translates it correctly. A similar issue occurs in Case 2 (missing "a man in a dark blue shirt"). Both cases suggest that **fine-grained token-level alignment could avoid missing translation**.

However, both S-CTR and S+T-CTR may have grammar problems under the zero-shot scenario, which can be solved by finetuning with a few parallel data. In Case 1, the phrase "playing a game of dirt" is obviously illogical, while the additional 100 parallel data corrects the preposition "of" to "in", which is more grammatical. This phenomenon shows that **it is difficult to learn grammar knowledge with contrastive learning, but only a few parallel data can compensate for this**.

5 Related Work

Multimodal Machine Translation Multimodal Machine Translation aims to introduce visual modality to enhance NMT. Early methods (Caglayan et al., 2016; Huang et al., 2016; Calixto et al., 2016; Delbrouck and Dupont, 2017a; Caglayan et al., 2017; Calixto and Liu, 2017; Delbrouck and Dupont, 2017b; Calixto et al., 2017; Libovický and Helcl, 2017; Caglayan et al., 2018; Zhou et al., 2018; Helcl et al., 2018) are mainly based on RNN architecture with attention. Recent methods (Ive et al., 2019; Yao and Wan, 2020; Yin et al., 2020; Liu et al., 2021; Lin et al., 2020a; Caglayan et al., 2021; Zhang et al., 2020; Fang and Feng, 2022; Li et al., 2022) based on Transformer further improve the performance. However, recent studies (Caglayan et al., 2019; Wu et al., 2021) found that visual information is often ignored when parallel corpus is sufficient. Therefore, in this paper, we turn to investigate the contribution of visual modality when the parallel corpus is not sufficient.

Zero-shot and Few-shot MT Since NMT strongly relies on large scale of parallel data, researchers begin to focus on situations with limited parallel data. Previous methods like unsupervised machine translation (Lample et al., 2018d,b,c; Ren et al., 2019; Sennrich and Zhang, 2019; Ruiter et al., 2019) achieve this with abundant monolingual data. Multilingual machine translation (Aharoni et al., 2019; Liu et al., 2020; Lin et al., 2020b; Pan et al., 2021) achieve this with parallel corpus of many other directions. Another line of research is to achieve zero-shot or few-shot translation with the help of visual modality (Nakayama and Nishida, 2017; Li et al., 2020), but they failed to achieve satisfactory performance with extremely limited data. We extend this research line and achieve better performance with less data.

Cross-modal Contrastive Learning Contrastive learning has lead to a great success in multimodal tasks like cross-lingual transfer (Huang et al., 2021), video-text understanding (Xu et al., 2021), and so on. One of the most representative methods is CLIP (Radford et al., 2021), which learns good alignment between images and text with contrastive learning. Recent work also shows the power of cross-modal contrastive learning in speech translation (Ye et al., 2022). Inspired by these efforts, we propose a cross-modal contrastive learning method to achieve zero-shot and few-shot translation.

6 Conclusion

In this paper, we propose a cross-modal contrastive learning method including sentence-level and token-level objectives, which realizes zero-shot and few-shot translation. Experimental results show that our method gains significant improvements over baseline under both scenarios. Further analysis demonstrate that our method learns good cross-modal and cross-lingual alignment. In the future, we will explore how our method enables cross-lingual transfer on more tasks.

Limitations

One limitation of our work is the pseudo data we used. Limited by the fact that most of existing image captioning datasets are annotated in English, we have to use additional translation models to

Models		
Case 1 FR→EN		
Ref.	SRC	Des enfants sont dehors , jouant dans la terre à côté de deux arbres.
	TGT	Some children are outside playing in the dirt where two trees are.
S-CTR (ZS)		The young children are playing a game of dirt. *(two trees)*
S+T-CTR (ZS)		The children are outside playing a game of dirt next to two trees.
S+T-CTR (FS100)		The children are outside playing a game in the dirt near two trees.
Case 2 CS→EN		
Ref.	SRC	Muž ve žluté košili a muž v tmavém modrém tričku si povídají.
	TGT	A man in a yellow shirt and a man in a dark blue shirt talking.
S CTR (ZS)		Man in yellow shirt is crying. *(a man in a dark blue shirt)*
S+T-CTR (ZS)		Man in yellow shirt *(and)* a man in a blue shirt is smiling.
S+T-CTR (FS100)		A man in a yellow shirt and a man in a blue shirt is talking.

Table 6: Qualitative examples from Multi30K Test2016 set. The red text indicates the grammar or vocabulary error, *(words in brackets)* indicate the missing words, and the green text indicates the correct translations.

generate pseudo captions in German, French and Czech. The lack of real data may impact the performance of our method.

Acknowledgements

The authors would like to thank all the anonymous reviewers for their insightful and valuable comments.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3874–3884. Association for Computational Linguistics.

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. 2017. LIUM-CVC submissions for WMT17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*, pages 432–439, Copenhagen, Denmark. Association for Computational Linguistics.

Ozan Caglayan, Adrien Bardet, Fethi Bougares, Loïc Barrault, Kai Wang, Marc Masana, Luis Herranz, and Joost van de Weijer. 2018. LIUM-CVC submissions for WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 597–602, Belgium, Brussels. Association for Computational Linguistics.

Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016. Multimodal attention for neural machine translation. *CoRR*, abs/1609.03976.

Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. 2021. 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*, pages 1317–1324, Online. Association for Computational Linguistics.

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, Desmond Elliott, and Stella Frank. 2016. DCU-UvA multimodal MT system report. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 634–638, Berlin, Germany. Association for Computational Linguistics.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003, Copenhagen, Denmark. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. 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)*, pages 1913–1924, Vancouver, Canada. Association for Computational Linguistics.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017a. 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.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017b. Multimodal compact bilinear pooling for multimodal neural machine translation. *CoRR*, abs/1703.08084.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, VL@ACL 2016, August 12, Berlin, Germany*. The Association for Computer Linguistics.

Qingkai Fang and Yang Feng. 2022. Neural machine translation with phrase-level universal visual representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. STEMM: Self-learning with Speech-text Manifold Mixup for Speech Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Shuhao Gu and Yang Feng. 2022. Improving zero-shot multilingual translation with universal representations and cross-mappings. In *EMNLP 2022 Long findings*.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVII*, volume 12362 of *Lecture Notes in Computer Science*, pages 417–434. Springer.

Jindřich Helcl, Jindřich Libovický, and Dušan Variš. 2018. CUNI system for the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 616–623, Belgium, Brussels. Association for Computational Linguistics.

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, Berlin, Germany. Association for Computational Linguistics.

Poyao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alex Hauptmann. 2021. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2443–2459. Association for Computational Linguistics.

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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018c. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018d. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 5039–5049. Association for Computational Linguistics.

Van Der Maaten. Laurens and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2605):2579–2605.

Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. 2022. On vision features in multimodal machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6327–6337, Dublin, Ireland. Association for Computational Linguistics.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. UNIMO: towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2592–2607. Association for Computational Linguistics.

Yaoyiran Li, Edoardo Maria Ponti, Ivan Vulic, and Anna Korhonen. 2020. Emergent communication pretraining for few-shot machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4716–4731. International Committee on Computational Linguistics.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada. Association for Computational Linguistics.

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020a. Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1320–1329, New York, NY, USA. Association for Computing Machinery.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020b. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2649–2663. Association for Computational Linguistics.

Pengbo Liu, Hailong Cao, and Tiejun Zhao. 2021. Gumbel-attention for multi-modal machine translation. *arXiv preprint arXiv:2103.08862*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

Minh-Thang Luong and Christopher D. Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1054–1063, Berlin, Germany. Association for Computational Linguistics.

Hideki Nakayama and Noriki Nishida. 2017. Zero-resource machine translation by multimodal encoder-decoder network with multimedia pivot. *Mach. Transl.*, 31(1-2):49–64.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 1–9. Association for Computational Linguistics.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 244–258, Online. Association for Computational Linguistics.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Unsupervised neural machine translation with SMT as posterior regularization. In *The*

Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pages 241–248. AAAI Press.

Dana Ruiter, Cristina España-Bonet, and Josef van Genabith. 2019. Self-supervised neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1828–1834, Florence, Italy. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2495–2504.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6153–6166, Online. Association for Computational Linguistics.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6787–6800. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Rong Ye, Mingxuan Wang, and Lei Li. 2022. Cross-modal contrastive learning for speech translation. In *Proc. of NAACL*.

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, Online. Association for Computational Linguistics.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*.

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.

A Translation Models for Pseudo Data

In this section, we introduce the detailed information of translation models we use to construct the pseudo data. For EN→DE and EN→FR directions, we use the pretrained model from Ott et al. (2018)[5], which consist of 6 encoder and decoder layers. The number of attention heads is set to 16. The dropout is set to 0.3 for EN-DE and 0.1 for EN-FR. The label smoothing is set to 0.1.

For EN→CS, we train a Transformer-base model on the WMT2015 EN→CS training set, which contains about 15M parallel data. The model contains 6 encoder and decoder layers. The number of attention heads is set to 8. The dropout and the label smoothing is set to 0.1.

We evaluate the EN→DE and EN→FR models on the WMT test set `newstest2014`, and evaluate the EN→CS model on `newstest2015`. As shown in Table 7, the performance of our models is reliable.

Languge	Model	BLEU
EN→DE	Vaswani et al. (2017)	28.4
	Ours (Ott et al., 2018)	29.3
EN→FR	Vaswani et al. (2017)	41.0
	Ours (Ott et al., 2018)	43.2
EN→CS	Luong and Manning (2016)	20.7
	Ours (Vaswani et al., 2017)	25.2

Table 7: BLEU scores of translation models for constructing the pseudo data.

[5]https://github.com/facebookresearch/fairseq/tree/main/examples/translation