

m3P: Towards Multimodal Multilingual Translation with Multimodal Prompt

Jian Yang¹, Hongcheng Guo¹, Yuwei Yin², Jiaqi Bai¹, Bing Wang¹,

Jiaheng Liu¹, Xinnian Liang¹, Linzheng Chai¹, Liqun Yang¹ and Zhoujun Li¹

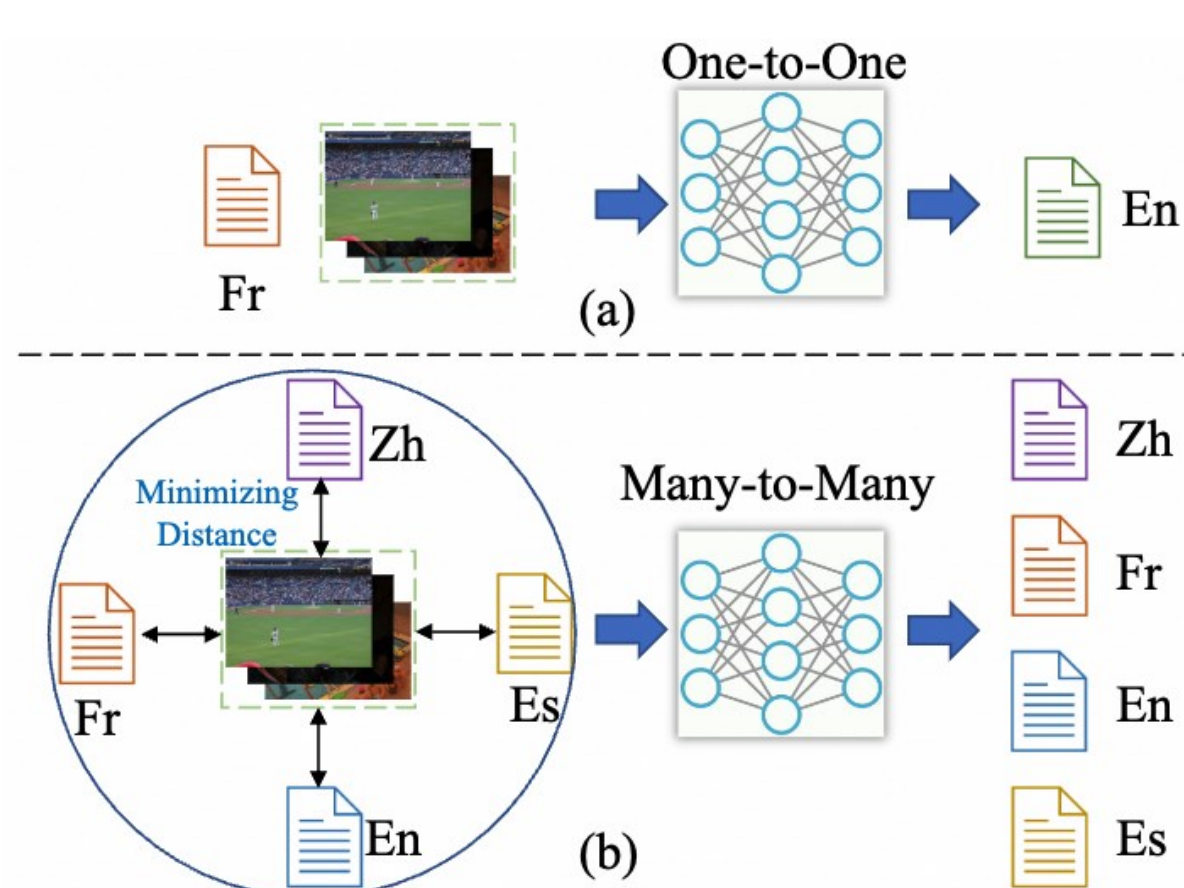
¹State Key Lab of Software Development Environment, Beihang University, Beijing, China

² Department of Computer Science, University of British Columbia

Abstract: Multilingual translation supports multiple translation directions by projecting all languages in a shared space, but the translation quality is undermined by the difference between languages in the text-only modality, especially when the number of languages is large. To bridge this gap, we introduce visual context as the universal language-independent representation to facilitate multilingual translation. In this paper, we propose a framework to leverage the multimodal prompt to guide the Multimodal Multilingual neural Machine Translation (m3P) which aligns the representations of different languages with the same meaning and generates the conditional vision-language memory for translation. We construct a multilingual multimodal instruction dataset (InstrMulti-102) to support 102 languages. Our method aims to minimize the representation distance of different languages by regarding the image as a central language. Experimental results show that m3P outperforms previous text-only baselines and multilingual multimodal methods by a large margin. Furthermore, the probing experiments validate the effectiveness of our method in enhancing translation under the low-resource and massively multilingual scenario.

1. Introduction

1.1 Motivation



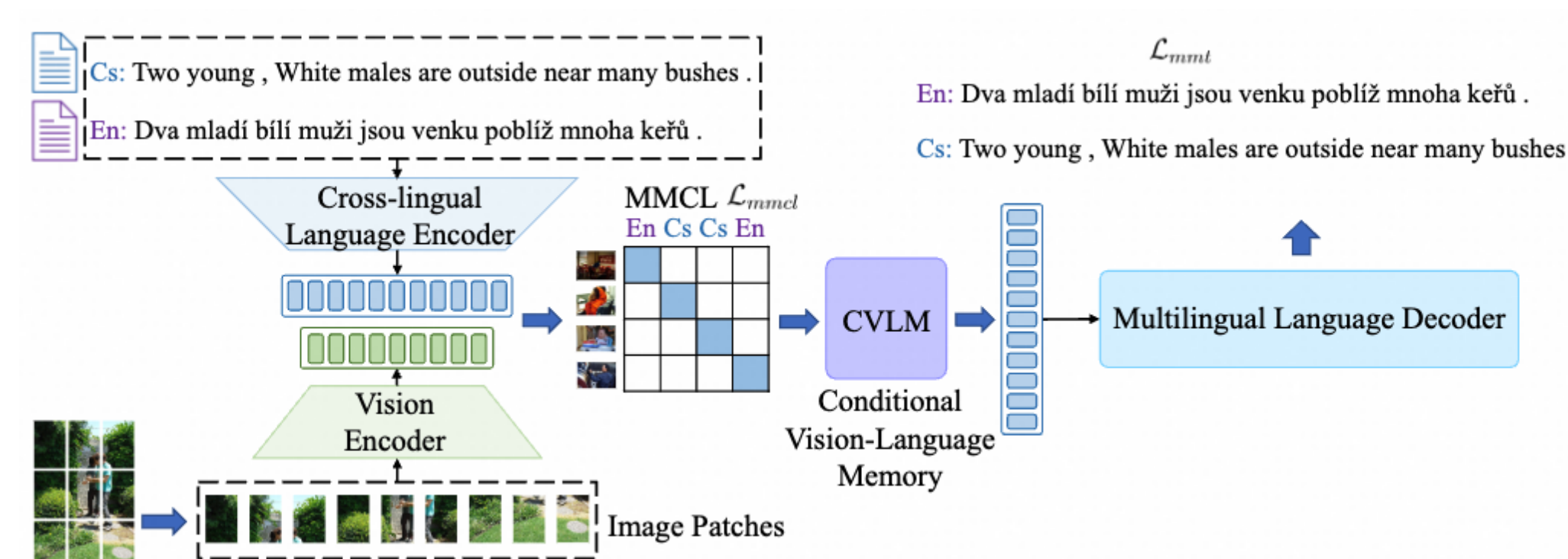
- Previous multimodal NMT works mainly focus on the bilingual translation supervised by the image-sentence training data. In Figure (a), each bilingual model can only handle a single translation direction compared to existing thousands of languages in the world. MNMT involves more languages using available linguistic resources but only implicitly brings different languages together by sharing the same parameters. There still exists a gap between different translation directions.
- Some previous works propose to leverage the aligned augmentation and contrastive learning across multiple languages only on the language modality. Meanwhile, images are regarded as the universal language to communicate ideas and concepts effectively across linguistic and cultural barriers. Hence, minimizing the difference across diverse directions by vision-language pair requires further exploration.

1.2 Our Contributions

- Propose a pre-training framework m3P, using multilingual multimodal contrastive learning to improve the transferability of different languages for the natural language generation.
- Our method is effective for multilingual translation even for the massively translation of 102 languages. Experimental results on the supervised translation directions demonstrate that our method substantially outperforms previous text-only and multilingual multimodal methods by nearly +1~+4 BLEU points.
- We can construct strMulti102 or fine-tune the encoder-decoder pre-trained models and decoder-only models to verify the effectiveness of our method. Plus, by conducting analytic experiments, we emphasize the importance of MMCL and CVLM.

2. Methods

2.1 Model Overview



2.2 Multilingual Multimodal Alignment

Sequence-to-sequence Span Corruption:

$$\mathcal{L}_m = - \sum_{m=1}^M \mathbb{E}_{x^k, y^k, z^k \in D_m} [\log P(y^k | x^k, z^k; \Theta)]$$

Replaced Token Detection:

$$\mathcal{L}_c = \sum_{x^k, z^k \in D_{all}} (f(x^k, z^k) + f(z^k, x^k))$$

Replaced Token Denoising:

$$f(x^k, z^k) = - \log \frac{\exp(z^k \cdot x^k / \tau)}{\sum_{x \in \{x^k, x^-\}} \exp(z^k \cdot x / \tau)} \quad f(z^k, x^k) = - \log \frac{\exp(x^k \cdot z^k / \tau)}{\sum_{z \in \{z^k, z^-\}} \exp(x^k \cdot z / \tau)}$$

Multi-task Training:

$$\mathcal{L}_{all} = \mathcal{L}_m + \lambda \mathcal{L}_c$$

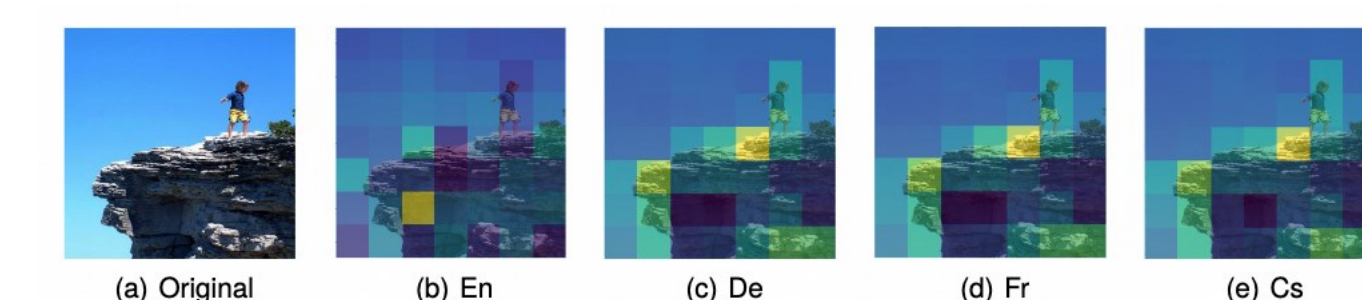
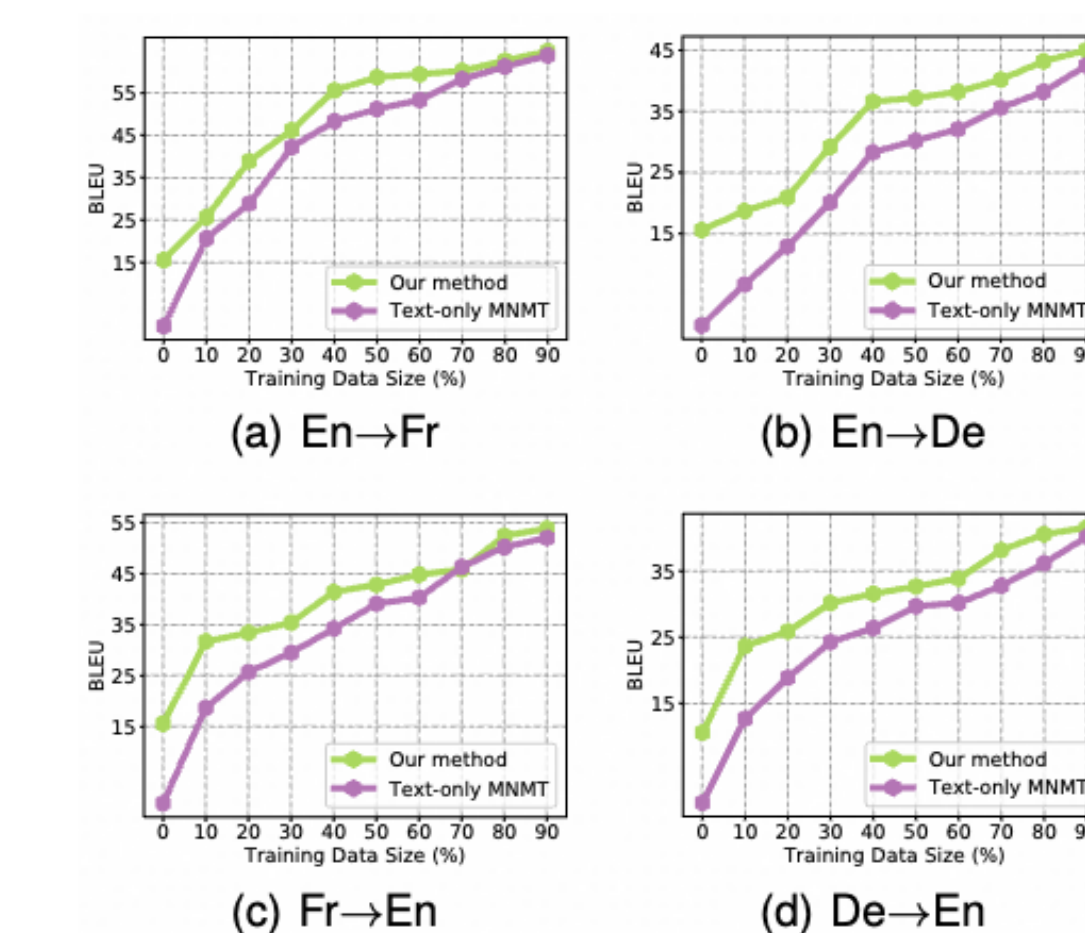
2.3 Ablation Study

ID	Flickr2016	En→De	De→En
①	m ³ P (our method)	41.6	45.0
②	① - MMCL	41.2	44.6
③	② - CVLM	40.8	44.0
④	③ - MDropNet	40.5	43.8
⑤	④ - Multilingual Training	40.1	43.2

Model	En→Fr	En→De	Fr→En	De→En
Text-only MNMT	63.8	40.2	52.0	42.5
ResNet50	64.2	40.6	52.3	43.1
ResNet101	64.4	40.8	52.4	43.4
ViT-B/32	64.8	41.6	53.8	45.0
ViT-B/16	65.1	41.8	53.6	44.8
ViT-B/14	65.2	41.9	53.4	45.2

3. Experimental Results

3.1 Analysis



3.1 Main Experiments

		En→Fr	En→Cs	En→De	Fr→En	Cs→En	De→En	Avg ₄
Only Trained on Text Data								
1→1	BINMT (Vaswani et al., 2017)	63.3	33.4	39.9	54.0	41.1	43.8	45.9
N→N	MNMT (Fan et al., 2021)	63.8	34.0	40.2	52.0	41.3	42.5	45.6
Trained on Text and Vision Data								
1→1	BINMT (Vaswani et al., 2017)	63.5	33.0	40.3	55.1	41.8	44.1	46.3
N→N	MNMT (Gated Fusion) (Li et al., 2021a)	63.8	34.4	41.0	51.5	41.1	43.3	45.8
	MNMT (Concatenation) (Li et al., 2021a)	63.0	33.8	38.8	53.3	43.6	44.0	46.1
	mRASP2 (Pan et al., 2021)	63.8	34.4	41.3	53.2	44.0	44.5	46.9
	Selective Attn (Li et al., 2022)	63.5	34.4	41.3	53.2	44.0	44.5	46.8
	LVP-M ³ (Guo et al., 2022b)	63.4	34.1	41.4	53.2	44.0	44.5	46.8
	m ³ P (Encoder-Decoder)	64.8	35.2	41.8	53.8	44.8	45.0	47.6
m ³ P (Decoder-only)	66.4	38.1	43.5	56.7	46.9	48.1	49.9	

		En→Fr	En→De	De→En	Fr→En	Avg ₄	En→Fr	En→De	Fr→En	De→En	Avg ₄
Flickr2017											
Only Trained on Text Data											
1→1	BINMT (Vaswani et al., 2017)	55.4	34.1	39.2	43.4	43.0	45.8	32.1	40.6	34.3	38.2
N→N	MNMT (Fan et al., 2021)	56.8	34.9	40.3	44.6	44.2	45.9	31.9	41.6	34.6	38.5
Trained on Text and Vision Data											
1→1	BINMT (Vaswani et al., 2017)	55.8	34.6	39.6	43.6	43.4	45.8	32.3	41.6	34.4	38.5
N→N	MNMT (Gated Fusion) (Li et al., 2021a)	56.8	34.3	40.3	44.2	43.9	46.8	32.5	42.2	34.5	39.0
	MNMT (Concatenation) (Li et al., 2021a)	56.4	34.0	39.4	43.8	43.4	46.4	32.6	42.4	34.1	38.9
	mRASP2 (Pan et al., 2021)	57.0	35.1	39.6	44.1	43.9	47.1	32.7	42.3	34.8	39.2
	Selective Attn (Li et al., 2022)	56.6	34.2	40.3	44.4	43.9	46.8	32.5	42.5	34.3	39.0
	LVP-M ³ (Guo et al., 2022b)	57.4	34.4	40.4	44.7	44.2	46.6	32.5	42.6	34.5	39.1
	m ³ P (Encoder-Decoder)	57.4	35.3	41.0	45.6	44.8	46.8	33.1	43.2	35.2	39.6
m ³ P (Decoder-only)	58.3	37.2	42.2	46.5	46.1	47.4	34.2	44.5	36.2	40.6	