South China Normal University
School of Computer Science
Text Analytics and Mining Lab

MTA: A Lightweight Multilingual Text Alignment Model for Cross-language Visual Word Sense Disambiguation

Qihao Yang [1] Xuelin Wang [2] Yong Li [1] Lap-Kei Lee [3] Fu Lee Wang [3] Tianyong Hao [1]

[1]School of Computer Science, South China Normal University, Guangzhou, China
[2]College of Chinese Language and Culture, Jinan University, Guangzhou, China
[3]School of Science and Technology, Hong Kong Metropolitan University, Hong Kong

Abstract

Visual Word Sense Disambiguation (Visual-WSD), as a subtask of fine-grained image-text retrieval, requires a high level of language-vision understanding to capture and exploit the nuanced relationships between text and visual features. However, the cross-linguistic background only with limited contextual information is considered the most significant challenges for this task. In this paper, we propose MTA, which employs a new approach for multilingual contrastive learning with self-distillation to align fine-grained textual features to fixed vision features and align non-English textual features to English textual momentum features. It is a lightweight and end-to-end model since it does not require updating the visual encoder or translation operations. Furthermore, a trilingual fine-grained image-text dataset is developed and a ChatGPT API module is integrated to enrich the word senses effectively during the testing phase. Extensive experiments show that MTA achieves state-of-the-art results on the benchmark English, Farsi, and Italian datasets in SemEval-2023 Task 1 and exhibits impressive generalization abilities when dealing with variations in text length and language.

Contributions

1) A new lightweight model is proposed, in which a text encoder is updated to flexibly adapt to multilingual contexts, while a visual encoder remains to provide fixed vision representations.
2) A new trilingual image-text dataset is created and is applied to the Visual-WSD task, encompassing a fine-grained network of 85,754 word-sense associations and 120,131 images.
3) The ChatGPT API is introduced to effectively elaborate the contextual information for brief phrases, enhancing the performance of fine-grained disambiguation tasks.

1 Method

MTA contains a fixed image encoder to generate fixed visual representations and a text encoder shared by English, Farsi, and Italian to generate cross-lingual textual representations, as illustrated in Figure 1. We employ a 24-layer vision transformer as the fixed image encoder. The text encoder and its momentum version are a 12-layer transformer, and both are updated and distilled with the involvement of momentum. Inspired by MoCo, MTA maintains an image queue and a text queue separately to ensure the model retains the consistency of critical representations. MTA is built upon CLIP-ViT-L/14 (a monolingual version of CLIP) and is fine-tuned on trilingual parallel data.

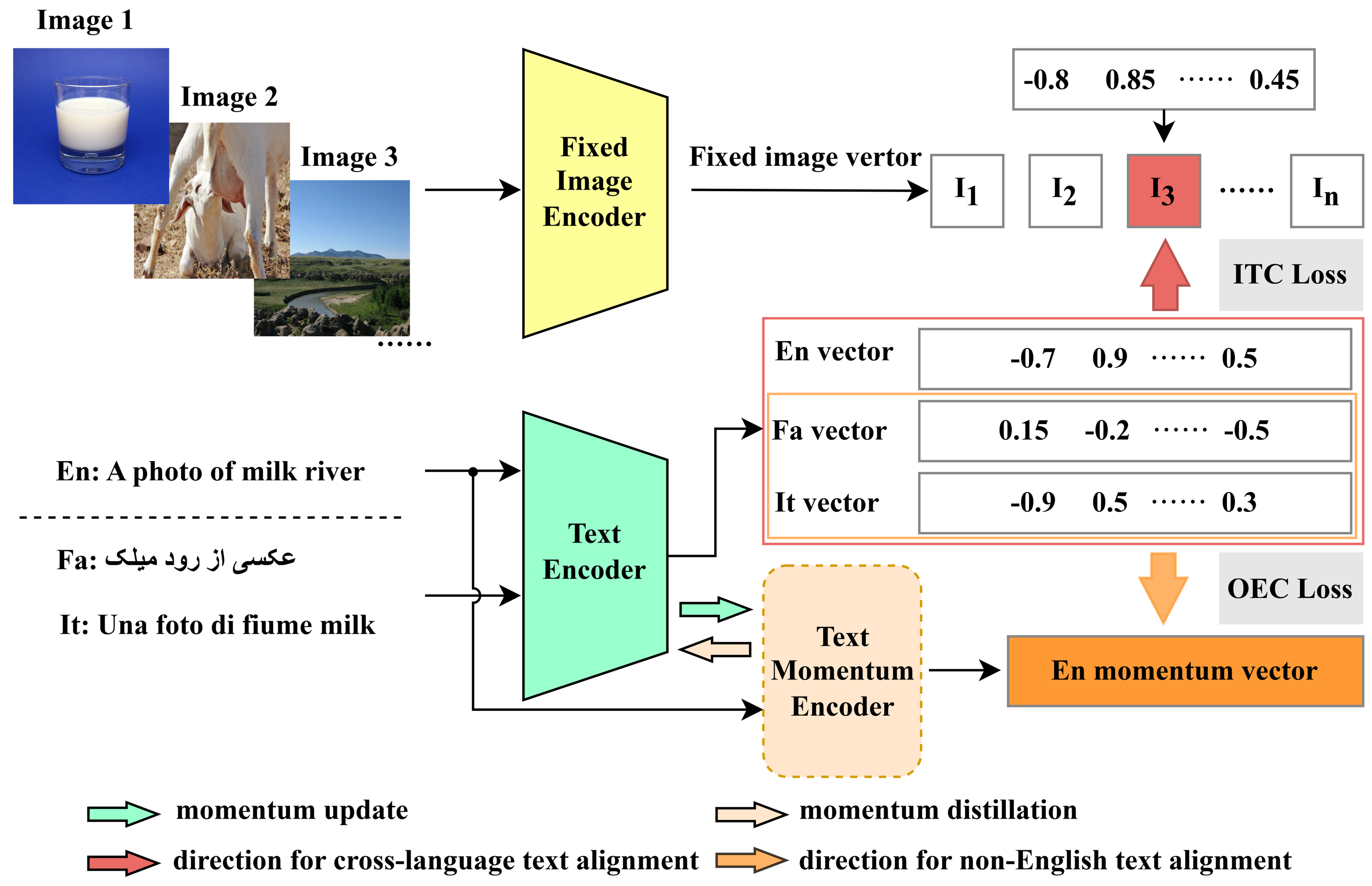

Figure 1: Overview architecture of the proposed MTA.

ChatGPT-3.5 is guided to function as a WSD assistant, generating longer sentences in the target language and phrase prompts, as shown in Figure 2.

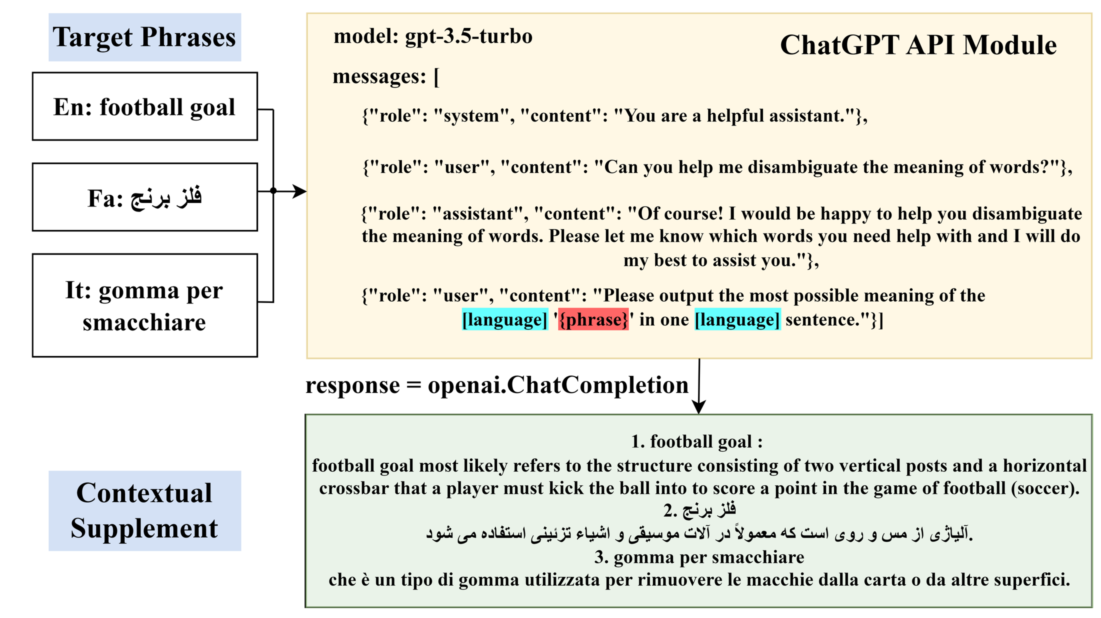

Figure 2: Operation of the ChatGPT API module.

2 Results

Table 1 displays the evaluation results of MTA and baselines on the benchmark test set from VWSD-2023. "Prompt 1" represents translating both the original phrase and gloss supplemented by ChatGPT into English and using their combination as the textual prompt. "Prompt 2" indicates directly combining the original phrase and gloss supplemented by ChatGPT as the textual prompt. "Prompt 3" denotes the use of only the original phrase as the textual prompt. Table 2 shows the results of the ablation study.

1) MTA achieves state-of-the-art performance in handling phrases and longer sentences and can even handle multiple languages without translation.
2) The image-text alignment module (with ITC loss) is the most crucial component for MTA.
3) The language alignment module (with OEC loss) can further contribute to the performance in understanding cross-language image-text knowledge.

Models	Parameters	English (%)		Farsi (%)		Italian (%)		Total (%)		Average (%)	
		HR@1	MRR@10	HR@1	MRR@10	HR@1	MRR@10	HR@1	MRR@10	HR@1	MRR@10
Prompt 1: (translated) phrase + (translated) gloss [with ChatGPT Enhancement]											
FCLL	189M	80.345	87.349	60.750	73.290	76.798	84.018	75.235	83.044	72.631	81.552
BLIP	447M	71.922	82.034	56.000	68.959	75.409	83.992	69.731	79.949	67.777	78.328
CLIP-ViT-L/14	427M	73.194	83.079	59.500	71.801	74.686	81.997	71.296	80.103	69.127	78.959
CLIP-ViT-L/14@336px	427M	73.650	83.206	59.000	71.490	75.032	83.256	71.314	80.953	69.227	79.317
CLIP-ViT-B-multilingual	286M	64.146	77.063	51.000	66.023	61.967	75.537	60.743	74.301	59.038	72.874
MTA	**85M**	**83.585**	**90.078**	**62.000**	**72.758**	**80.327**	**87.346**	**78.099**	**85.639**	**75.304**	**83.394**
Prompt 2: (original) phrase + (original) gloss [with ChatGPT Enhancement]											
FCLL	189M	80.345	87.349	9.000	27.178	33.442	51.644	50.826	63.667	40.929	55.390
BLIP	447M	71.922	82.034	8.000	26.360	27.213	45.876	44.628	59.138	35.711	51.423
CLIP-ViT-L/14	427M	73.194	83.079	8.500	27.865	43.934	61.804	51.756	65.576	41.876	57.583
CLIP-ViT-L/14@336px	427M	73.650	83.206	8.000	26.645	42.295	60.511	50.206	64.369	41.315	56.787
CLIP-ViT-B-multilingual	286M	64.146	77.063	40.000	56.277	49.180	66.109	54.442	69.317	51.109	66.483
MTA	**85M**	**83.585**	**90.078**	**48.000**	**61.013**	**76.065**	**83.954**	**73.863**	**82.143**	**69.216**	**78.348**
Prompt 3: (original) phrase [without ChatGPT Enhancement]											
FCLL	189M	60.475	74.953	6.500	24.674	17.704	37.543	35.847	52.777	28.226	45.723
BLIP	447M	60.259	73.696	8.000	26.830	20.327	39.236	36.880	53.155	29.528	46.587
CLIP-ViT-L/14	427M	57.451	72.660	9.000	28.501	30.491	49.278	38.946	56.169	32.314	50.146
CLIP-ViT-L/14@336px	427M	59.395	73.058	7.500	26.736	30.819	49.210	39.669	55.973	32.571	49.668
CLIP-ViT-B-multilingual	286M	46.868	65.062	21.500	41.080	28.524	48.357	35.847	54.844	32.297	51.500
MTA	**85M**	**64.634**	**76.967**	**40.000**	**55.976**	**52.459**	**67.915**	**54.698**	**68.286**	**52.364**	**66.952**

Table 1: Evaluation results on the benchmark test set.

Models	Image-text Alignment	Other language-English Alignment	Average HR@1	$\triangle$
MTA	✓	✓	52.364	0
$\mathrm{MTA}_{w/o-ITC}$	✗	✓	11.424	-40.940
$\mathrm{MTA}_{w/o-OEC}$	✓	✗	45.542	-6.822

Table 2: Ablation Study of MTA for "Prompt 3" on the benchmark test set.

3 Conclusion

This paper proposes a new lightweight multilingual text alignment model for cross-language visual word sense disambiguation. Employing a multilingual contrastive learning and self-distillation mechanism, it achieves state-of-the-art performance through two alignment processes. We conduct an in-depth analysis of the limitations presented in the VWSD-2023 training set and confirm the effectiveness of our T-VWSD dataset in improving model performance. High-quality multilingual fine-grained image-text data is essential for visual word sense disambiguation. The publicly accessible T-VWSD dataset may provide a promising prospect for future research. Source codes and the datasets are publicly released at: https://github.com/CharlesYang030/MTA.