University of Glasgow | School of Computing Science

Terrier Team | Information Retrieval Group

MultiWay-Adapter: Adapting Multimodal Large Language Models for scalable image-text retrieval

Zijun Long, George Killick, Richard McCreadie, Gerardo Aragon Camarasa
z.long.2@research.gla.ac.uk

Overview

As Multimodal Large Language Models (MLLMs) grow in size, adapting them to specialized tasks becomes increasingly challenging due to high computational and memory demands. While efficient adaptation methods exist, in practice they suffer from shallow inter-modal alignment, which severely hurts model effectiveness. To tackle these challenges, we introduce the MultiWay-Adapter (MWA), which deepens inter-modal alignment, enabling high transferability with minimal tuning effort.

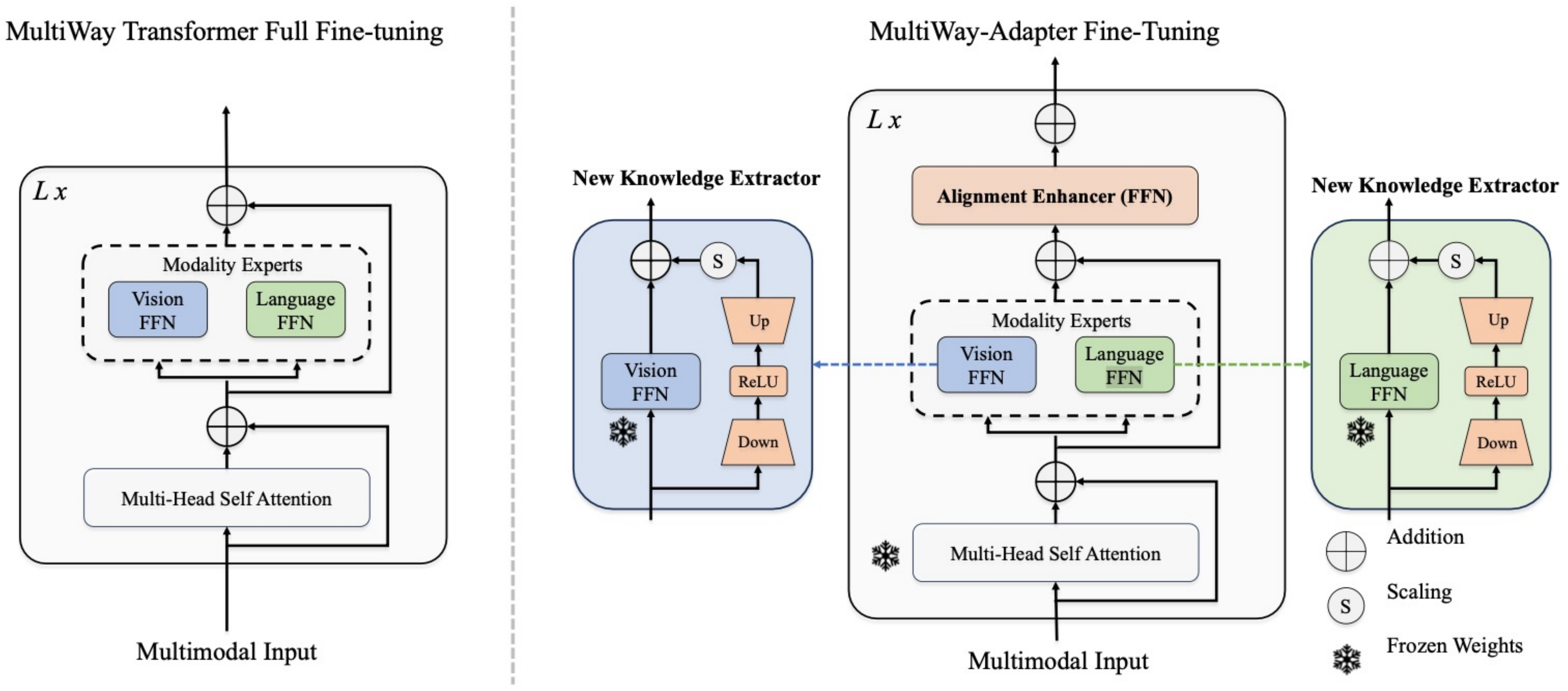
1. Motivations

- **Increasing Computational demands of MLLMs:**
  - ➢Traditional full fine-tuning requires isolated, exhaustive retraining for each new task, demanding intensive computational resources and thus limiting practical applications.
  - ➢For instance, training BLIP2-Giant on an Nvidia A100 GPU takes 144 days [1].
- **Under-explored of efficient transfer learning methods in multi-modal domain:**
  - ➢Most existing methods focus on unimodal models.
- **Existing methods suffers from shallow intra-modal alignment:**
  - ➢Existing adaptation methods for MLLMs [1] focus on information extraction from down- stream datasets but neglect the critical need for inter-modal alignment.
  - ➢With shallow alignment, the model would fail to capture the complex inter- relations between different modalities, thereby impacting its effectiveness in multi-modal tasks.

PAPER CODE

3. Proposed: MultiWay-Adapter (MWA)

- **Dual-component approach:**
  - ➢ *New Knowledge extractor*
  - ➢ *Modality Enhancer*
- **This New Knowledge extractor comprises two branches: the left branch, identical to the original network, and an additional right branch introduced for task-specific fine-tuning. The latter utilizes a bottleneck structure to limit the number of parameters and includes a down-projection layer and an up-projection layer.**
- **Alignment Enhancer module is added atop the pool of feed-forward networks. This module mimics the architecture of the New Knowledge Extractor but uses a larger middle dimension to facilitate better feature fusion and alignment.**



4. Experimental Results

- **Fine-Tuning Performance**
- **Efficiency:**
  - ➢ Fine-tuning MWA with the BEiT-3 Base model is reduced by 57% compared to full fine-tuning.
- **Effectiveness:**
  - ➢ The performance decrement when utilizing MWA is statistically insignificant for both the Base and Large BEiT-3 variants, with deviations falling within a margin of less than 1%.
- **Zero-shot performance :**
  - ➢In this setting, the model is evaluated on Flickr30k (1k test set), with which it has no prior knowledge of, thereby necessitating reliance on intrinsically learned knowledge to simulate the handling of previously unseen samples.
  - ➢MWA surpasses the performance of full fine-tuning when employed with the BEiT-3 Large model.

		Flickr30k	
Model	FT-Way	IR@1	TR@1
BEiT-3-Large	Full fine-tune	85.99	95.48
BEiT-3-Large	MultiWay-Adapter	86.26	95.51

					MSCOCO (5k test set)		Flickr30k (1k test set)	
Model	FT-Way	Tunable params (M)	GPU Mem (GB)	Time (Min)	IR@1	TR@1	IR@1	TR@1
ALBEF [24]	Full Fine-tune	196	N/A	N/A	60.7	77.6	85.6	95.9
ALIGN [25]	Full Fine-tune	825	N/A	N/A	59.9	77.0	84.9	95.3
BEiT-3-Base	Full Fine-tune	222 (100%)	37GB	225	61.4	79.0	86.2	96.3
BEiT-3-Large	Full Fine-tune	675 (304%)	45GB	353	63.4 (+2.0)	82.1 (+3.1)	88.1 (+1.9)	97.2 (+0.9)
BEiT-3-Base	MultiWay-Adapter	7.13 (**3.21%**)	**30GB**	**130**	60.7 (-0.7)	78.3 (-0.7)	85.4 (-0.8)	95.4 (-0.9)
BEiT-3-Large	MultiWay-Adapter	17.40 (**2.58%**)	**36GB**	**194**	63.3 (+1.9)	82.1 (+3.1)	88.0 (+1.8)	97.1 (+0.8)

5. Conclusions

- We introduce the MultiWay-Adapter (MWA), an effective framework designed for the efficient adaptation of MLLM to downstream tasks.
- Our MWA the issue of shallow inter-modal alignment in existing methods, MWA employs a dual-component approach, utilizing both the New Knowledge Extractor and the Alignment Enhancer.
- This strategy enables MWA to not only extract novel information from downstream datasets but also to secure deep inter-modal alignment. Our empirical findings reveal that there is no statistically significant decline in performance across all tested settings while reducing the fine-tuning time by up to 57%.

References

[1] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal, “VL- ADAPTER: parameter-efficient transfer learning for vision-and-language tasks,” in *Proc. CVPR, 2022*.