

Peer Review

Review of: "CM3T: Framework for Efficient Multimodal Learning for Inhomogeneous Interaction Datasets"

Yulai Xie¹

1. Independent researcher

The paper introduces the CM3T framework, an efficient approach for multimodal learning, addressing challenges like modality differences and data imbalance. It features Multi-Head Vision Adapters (MHVA) and Cross-Attention Adapters (CAA) to enhance adaptability and efficiency, along with prefix tuning to reduce training parameters. The framework allows effective learning without altering the backbone network, demonstrating strong performance across various datasets and tasks. CM3T offers a flexible and resource-efficient solution for multimodal learning, suitable for environments with limited resources.

There are some comments on this paper.

- The paper does not effectively showcase technical insights, appearing more like a combination of existing modules.
- For instance, the main contribution, Multi-Head Vision Adapters (MHVA), is introduced with the statement: "we introduce Multi-Head Vision Adapter (MHVA). This is different from multi-head attention as the input is divided along the window dimension of Video Swin transformers (or spatial patch dimension in ViViT) and not the channel dimension." However, the innovation and its benefits are not well-emphasized, especially in the introduction.
- Similarly, for Cross-Attention Adapters (CAA), the distinction between the designed structure and typical cross-attention modules needs to be highlighted.
- Figures 1 and 2 do not effectively convey the application scenarios of the method (Multimodal Multi-dataset Multitask), making it difficult for readers to understand. It might be beneficial to graphically represent the 3M concept.

- Figure 2 could be further improved, as it does not clearly indicate the backbone. Additionally, the plugins appear much larger than the backbone, which is counterintuitive.
- The labels in Figure 2 should be consistent with those in the introduction and method sections; for example, CAA is not identifiable in the figure.
- The Method section needs reorganization, as it is currently difficult to distinguish between different modules. Consider structuring it with sections like Foundations, Method Overview, PT, MHVA, and CAA.

Declarations

Potential competing interests: No potential competing interests to declare.