

Peer Review

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

Yushan Pan¹

1. Xi'an Jiaotong-Liverpool University, Suzhou, China

The manuscript presents a comprehensive framework for efficient multimodal learning, particularly focusing on the CM3T (Cross Multimodal Multi-dataset Multitask Transformer) model. The model is well-structured; however, there are some concerns regarding the writing and technical issues.

The writing issue:

I would like to suggest the author consider improving the abstract and introduction. The abstract is concise. However, it could benefit from a clearer statement of the problem and the specific contributions of the work. For instance, it could explicitly mention the challenges of inhomogeneous datasets and the specific improvements CM3T offers over existing methods. Similar to the introduction, it provides a good overview of the challenges in video classification and the need for efficient transfer learning. However, it could be more engaging by highlighting the real-world implications of the proposed framework. For example, how does CM3T improve practical applications in video classification?

In addition, the manuscript provides a good overview of the datasets used for experiments. However, it could include more details about the **specific challenges posed by each dataset** and how CM3T addresses these challenges. This would help readers understand the practical relevance of the experiments.

Moreover, the manuscript briefly mentions the limitations of the approach but could expand on this discussion. For instance, what are the potential drawbacks of using frozen pretrained models, and how might these be mitigated in future work?

Further, a few sentences are overly complex and could be simplified for better readability. For example, the sentence "The basic idea consists in adding adapters (i.e., plugin architectures of very few trainable parameters) to a backbone and only train these while keeping the backbone frozen" could be rephrased for clarity. Please also ensure consistent use of terminology throughout the article. For example, the term "adapters" is sometimes referred to as "plugins," which might confuse readers.

Last, the conclusion is concise but could be more impactful by summarizing the key takeaways and emphasizing the broader implications of the research. For example, how does CM3T advance the field of multimodal learning, and what are the potential long-term benefits?

Technical concerns:

1. Model architecture. The paper provides a detailed description of the CM3T framework; some parts of the technical explanation could be more accessible. For instance, the description of multi-head vision adapters and cross-attention adapters could be supplemented with more intuitive explanations or analogies to help readers who may not be familiar with these concepts.
2. The mathematical equations are well-presented, but the article could benefit from more detailed annotations or explanations of the variables and their significance. Additionally, Figure 2, which illustrates the architecture, could be more detailed or accompanied by a step-by-step explanation to help readers better understand the flow of information through the model.
3. The comparison with state-of-the-art methods is thorough, but the article could provide more insights into why CM3T outperforms or matches these methods. For example, what specific features of CM3T contribute to its superior performance on certain datasets?
4. The article mentions ablation studies but does not provide detailed results or analysis. Including a section dedicated to ablation studies would help readers understand the contribution of each component of the CM3T framework.

Minor issues:

1. The ethical statement is well-written, but it could be expanded to discuss potential ethical implications of the proposed framework. For example, how might CM3T be used in surveillance or other sensitive applications, and what ethical considerations should be taken into account?
2. The references are comprehensive, but the article could benefit from more recent citations, particularly in the discussion of related work. This would help position CM3T within the latest

developments in the field.

Declarations

Potential competing interests: No potential competing interests to declare.