Outlook

This work has demonstrated the feasibility of generating a vision-language model from unimodal components through an end-to-end self-supervised learning approach. The resulting model, while not reaching the performance of state-of-the-art vision-language models, shows promising results across various benchmarks.

Towards a General Framework While this work introduces efficient multimodal learning on the example of vision-language models, the question arises whether this approach can be extended to other modalities, such as audio or video. While the methodology presented is not restricted to vision and language, as it can be adapted by using a teacher and pretrained models (for initialization) from other modalities, the success of this approach still needs to be demonstrated.

Positional Encoding

Positional Encoding