

Abstract

Multimodal models, especially vision-language models, have gained increasing popularity due to their wide range of applications on both unimodal and multimodal tasks. However, existing approaches often require large-scale models, extensive data, and substantial compute, limiting their accessibility for smaller research groups and individuals. This thesis address this issue by introducing an efficient self-supervised vision-language model that is significantly cheaper to train and smaller in size. We leverage pretrained unimodal encoders and introduce a randomly initialized shared encoder to align representations using a contrastive loss function. A self-supervised image model is employed for simultaneous knowledge distillation, guiding the alignment through high-level image representations. Our proof-of-concept demonstrates competitive performance with popular vision-language models like CLIP and FLAVA on retrieval tasks, outperforming them on certain metrics while using only 0.75% of the data used by CLIP and 4.3% by FLAVA. These finding underscore the potential for designing efficient multimodal models, and therefore lay the foundation for future research on financially accessible models, promoting broader participation in multimodal learning.

