Limitations

While the proposed methods is an efficient way to traing comparatively small multimodal models, and can easily be adapted to other modalities, e.g. audio, it has two main limitations.

First, our method relies on knowledge distillation of a self-supervised image model as the teacher. The fact that there has been no incentive for the teacher to learn a representation that is independent of the image modality makes it difficult to learn a representation that is truly modality-invariant and aligned across the modalities of the student model. This has repeatedly been shown when comparing the loss between image-to-image and text-to-image distillation, where the former is consistently lower. Interestingly, we were still able to outperform the approach of a supervised teacher, showing that while the predicted ImageNet-1K classes (with KL-Divergence, see)