Saathi

03/07-06-2023

Topic : One-Peace : Exploring one general representation model toward unlimited modalities

Abstract :

In this work, we explore a scalable way of building a general representation model toward unlimited modalities. We release One-Peace, a highly extensible model with 4B parameters that can seamlessly align and integrate representations across vision, audio and language modalities. The architecture of One Peace comprises modality, adapters, shared self-attention layers, and modality FFNs. This design allows for the easy extension of new modalities by adding adapters and FFNs while also enabling multi model fusion through self-attention layers. To pretrain One Peace, we develop two modality-agnostic pretraining tasks, cross-modal aligning and intra-modal denoising contrast, which align the semantic space of different modalities and capture fine grained details within modalities concurrently. With the scaling friendly architecture and pretraining tasks, One Peace has the potential to expand to unlimited modalities. Without using any vision or language pretrained models

for initialization, One Piece achieves leading results on a wide range of uni-modal and multi-modal tasks, including image classification (ImageNet), semantic segmentation (ADE20K), audio-text retrieval (AudioCaps, Clotho), audio classification (ESC-50, FSD50K, VGGSound), audio question answering (AVQA), image-text retrieval (MSCOCO, Flickr30K) and visual grounding (RefCOCO/+/g).

Conclusion :

In this work, we explore a scalable way for building a general representation model across different modalities. Based on the flexible architecture and modality agnostic pretraining tasks, we release One Peace, a general representation model that can seamlessly align and integrate representations across vision, audio and language modalities. We conduct a series of experiments across 8 modalities, 11 tasks and 16 datasets. The experimental results demonstrate that One Peace achieves leading results in a wide range of tasks, including image classification, semantic segmentation, audio-text retrieval, audio classification, audio question answering, image text retrieval and visual grounding. Furthermore, we show that One Peace possesses a strong

emergent zero-shot retrieval capability, enabling ~~to~~ it to align modalities that are not paired in the training data.