Topic: Image as a foreign language: BEIT Pretraining for all Vision and vision-Language tasks.

05/08 June23

Abstract:

A big convergence of language, vision and multimodal pretraining is emerging. In this work, we introduce a general purpose multimodal foundation model BEIT-3 which achieves state-of-the-art transfer ~~programme~~ performance on both vision and vision-language tasks. Specifically, we ~~com~~ advance the big convergance from three aspects: backbone architecture, pretraining task and model scaling up. We introduce Multi way transformers for general purpose modelling, where the modular architecture enables both deep fusion and modality-specific modelling (encoding). Based On the shared backbone, we perform masked 'language' modelling on images (Imglish), texts (English) and image-text pairs ("parallel sentences") in a unified manner. Experimental results show that BEIT-3 obtains state-of-the-art performance on object detection (COCO), semantic segmentation (ADE20K), image classification (Image Net)

, visual reasoning (NLVR2), visual question answering (VQAv2), image captioning (COCO) and cross model retrieval (Flickr 30k, COCO)

Conclusion

In this paper, we present BEIT-3, a general purpose multimodal foundation model, which achieves state-of-the-art performance across a wide range of vision and vision language benchmark the key idea of BEIT-3 is that image can be modelled as a foreign language so that we can conduct masked language modelling over images, text and image-text pairs in a unified way. We also demonstrate that Multiway Transformers can effectively model different vision and vision language tasks, making it an intiguing option for for general purpose modelling. BEIT-3 is simple and effective and is a promising direction for scaling up multimodal foundation models. For future work, we are working on pretraining multilingual BEIT-3 and including more modalities in BEIT-3 facilitate the cross-lingual and advance the big convergance of large scale pretraining across tasks, languages and modalities. We are also interested in enabling in context learning capability for multimodel foundation models by combiny the strength of BEIT-3 and MetaLM