Date __/__/____

Saathi

11 | 11 June 2023

Topic: Contrastive captioners are Image Text foundation models : CoCa

Abstract:

Exploring large scale pretrained foundation models is of significant intrest in computr vision because these models can be quickly transferred to many down-stream tasks. This paper presents (CoCa) contrastive captioners, a minimalistic design to pretrain an image text encoder decoder foundation model jointly with contrastive loss and captioning loss, thereby subsuming model capabilities from contrastive approaches like CLIP and generative methods like SimVLM. In contast to standard encoder decoder transformers where all decoder layers attend to encoder outputs, CoCa omits cross attention in first half of decoder layers to encode unimodal text representations, cascades the remaining decoder layers which cross attend to the image encoder for multimodal image text representations. We apply a contrastive loss between unimodal image and text embeddings, in addition to a captioning loss on the multimodal decoder outputs which predicts text tokens autoregressively. By sharing the same computational graph, the two training

objectives are computed efficiently with minimal overhead. CoCa is pretrained end to end and from scratch on both web-scale alt-text data and annotated images by treating all labels simply as text, seamlessly unifying natural language supervision for representation learning. Empirically, CoCa achieves state of the art performance with zero-shot transfer or minimal task-specific adaptation on a broad range of downstream tasks, spanning visual recognition (ImageNet, Kinetics 400/600/700), cross-modal retrieval, multimodal understanding and image captioning. Notably on ImageNet classification, CoCa obtains 86.3% zero-shot top-1 accuracy, 90.6% with a frozen encoder and learned classification head and new state of the art 91.0% top-1 accuracy on ImageNet with a fine-tuned encoder.

Conclusion

In this work, we present Contrastive Captioners, a new image-text foundation model family that subsumes existing vision pretraining paradigms with natural language supervision. Pretrained on image-text pairs from various data sources in a single stage, CoCa efficiently combines contrastive and captioning objectives in an encoder-decoder model. CoCa

Date __/__/____

Saathi

obtains a series of state of the art performance with a single check point on a wide spectrum of vision and vision language problems. Our work bridges the gap among various pretraining approaches and hope it motivates new directions on image text foundation models.