CoCa: Contrastive Captioners are Image-Text Foundation Models [4]

Minji Kim (minji@snu.ac.kr; 2020-28702), Dept. of Electrical and Computer Engineering, Seoul National University

1. Introduction

Large-scale Vision-Language Pretraining (VLP) models is getting significant attention because of their scalability to downstream tasks. This paper presents Contrastive Captioner (CoCa) to pretrain an image-text encoder-decoder foundation model with joint training with contrastive loss (like CLIP [2]) and captioning loss (like SimVLM [3]). The pretrained CoCa can be applied to various downstream tasks including visual recognition, vision-language alignment, image captioning and multi-modal understanding with zero-shot transfer. The overview of CoCa is illustrated in fig. 1.

2. Preliminaries

Vision-language foundation models can be categorized into three: (1) **Single-encoder models** are pretrained with cross-entropy loss on image classification datasets, *e.g.*, ImageNet. The image encoder provides generic *visual representations* for various downstream tasks, but it regards image annotations as labeled vectors and cannot explore the free-form natural language labels. (2) **Dual-encoder models** are pretrained with two parallel encoders with a contrastive loss on web-scale noisy image-text pairs. These models enable the tasks to infer the *cross-modal alignment* such as zero-shot image classification and image-text retrieval, but they are not directly applicable for joint vision-language understanding tasks, *e.g.*, visual question answering (VQA), because joint multi-modal representation learning is not included in their pipelines. (3) **encoder-decoder models** takes images on the encoder side and applies language modeling loss on the decoder outputs to attain text representations for *multi-modal understanding* tasks. However, these methods do not produce text-only representations that are aligned with image embeddings.

Unlike existing approaches, CoCa is a unified approach for single-encoder, dual-encoder, and encoder-decoder paradigms, and thus is applicable to more wide range of downstream tasks.

3. Contrastive Captioner (CoCa)

Similar to general image-text encode-decoder models, CoCa encodes images to latent embeddings through a neural network encoder such as ViT or ConvNets, followed by decoding texts with standard Transformer decoders. The difference with existing decoder is that CoCa removes cross-attention in the first half of the decoder layers for processing unimodal text representations, while the last half cross-attends to the image encoder for multi-modal representations. Detailed illustrations can be found in Fig. 2.

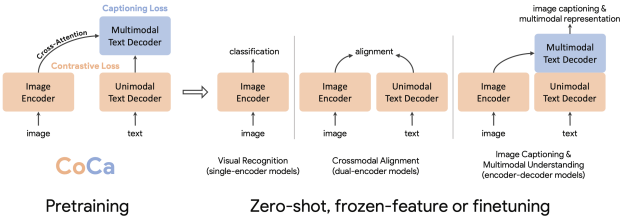

Figure 1. Overview of CoCa.

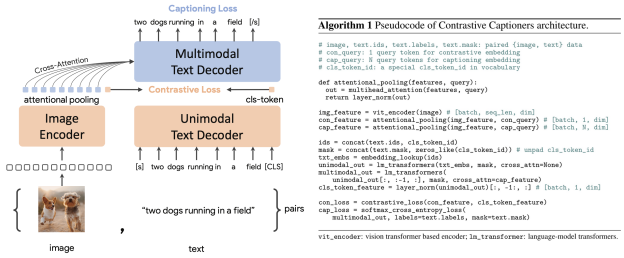

Figure 2. Detailed illustration and pseudo code of CoCa.

Model	ImageNet
ALIGN [13]	88.6
Florence [14]	90.1
MetaPseudoLabels [51]	90.2
CoAtNet [10]	90.9
ViT-G [21]	90.5
+ Model Soups [52]	90.9
CoCa (frozen)	90.6
CoCa (finetuned)	**91.0**

Model	K-400	K-600	K-700	Moments-in-Time
ViViT [53]	84.8	84.3	-	38.0
MoViNet [54]	81.5	84.8	79.4	40.2
VATT [55]	82.1	83.6	-	41.1
Florence [14]	86.8	88.0	-	-
MaskFeat [56]	87.0	88.3	80.4	
CoVeR [11]	87.2	87.9	78.5	46.1
CoCa (frozen)	88.0	88.5	81.1	47.4
CoCa (finetuned)	**88.9**	**89.4**	**82.7**	**49.0**

Figure 3. Experimental results on image classification and video action recognition tasks.

4. Some Interesting Parts in Experiments

CoCa is also applied to video recognition task in the experimental section. Multiple frames of videos are fed into the shared image encoder individually, and the additional pooler is trained on top of the spatial and temporal feature tokens. As can be seen in the right table of Fig. 3, CoCa achieves better performance than a video-based Transformer such as ViViT [1] even without video-level representation learning.

References

[1] A. Arnab, et al. Vivit: A video vision transformer. In *ICCV*, 2021. 1

[2] A. Radford, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[3] Z. Wang, et al. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 1

[4] J. Yu, et al. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1