CoCa: Contrastive Captioners are Image-Text Foundation Models

Abstract

- 본 논문에서는 Constrastive Captioner(CoCa) - contrastive loss와 captioning loss를 통한 image-text encoder-decoder foundation model을 pretrain하는 최소한의 디자인 - 를 통해 CLIP이나 SimVLM과 같이 대조적인 접근 방식에서 모델의 기능을 포함하게 한다.
- CoCa는 일반적 트랜스포머와 달리 first half of decoder layers에서 cross-attention을 생략하고, 남은 decoder layer을 cascade한다.
- 본 논문의 저자들은 unimodal image와 text embedding에 및 multimodal decoder output에 contrastive loss를 apply 하였다.

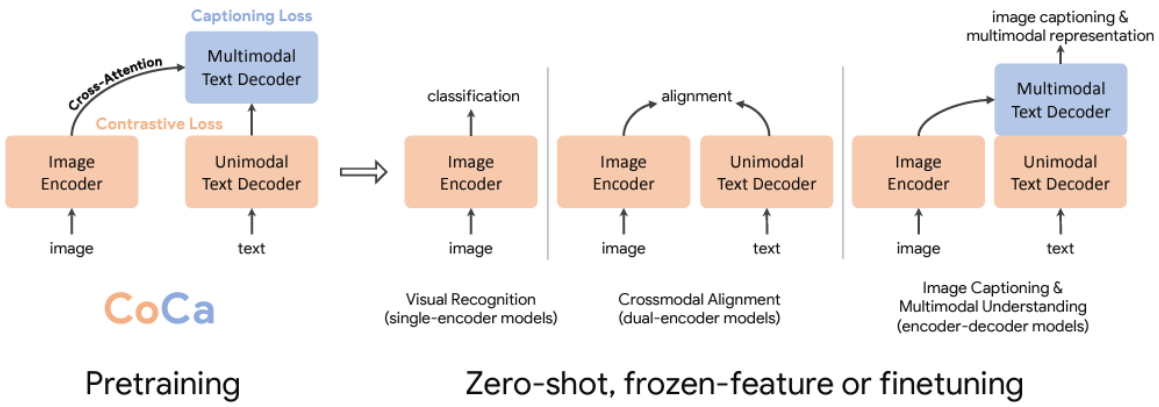


Figure 1: Overview of Contrastive Captioners (CoCa) pretraining as image-text foundation models. The pretrained CoCa can be used for downstream tasks including visual recognition, vision-language alignment, image captioning and multimodal understanding with zero-shot transfer, frozen-feature evaluation or end-to-end finetuning.

Approach

Natural Language Supervision

Single-Encode Classification

이전의 single-encoder approach는 대량의 데이터에 기반한 image classification을 통한 visual encoder을 pretrain하는 것이었다(annotation texts의 vocabulary가 고정된 상태). 이러한 image annotation은 일반적으로 아래 수식인 cross-entropy loss를 통한 학습을 위해 discrete(이산) class vector으로 매핑된다.

$$\mathcal{L}_{\text{Cls}} = -p(y) \log q_\theta(x),$$

- $p(y)$: one-hot, multi-hot or smoothed label distribution from ground truth label y.

학습된 image encoder은 다운스트림 작업을 위해 generic visual representation extractor을 사용한다.

Dual-Encoder Contrastive Learning

Single-encdoder classification과 비교해 dual-encoder은 noisy web-scale text description을 exploit하고 text tower을 통해 free-form text를 encode한다. encoder들은 paired text를 샘플 배치의 나머지들과 비교하여 optimize된다.

$$\mathcal{L}_{\text{Con}} = -\frac{1}{N}\Big(\underbrace{\sum_i^N \log \frac{\exp(x_i^\top y_i/\sigma)}{\sum_{j=1}^N \exp(x_i^\top y_j/\sigma)}}_{\text{image-to-text}} + \underbrace{\sum_i^N \log \frac{\exp(y_i^\top x_i/\sigma)}{\sum_{j=1}^N \exp(y_i^\top x_j/\sigma)}}_{\text{text-to-image}}\Big),$$

- x_i, y_j : i 번째 pair의 normalized embedding of image, j 번째 pair의 text.
- N :Batch size.
- σ : Temperature to scale logits.

더해서 dual-encoder approach는 aligned text encoder(that enables crossmodal alignment applications)을 학습한다.

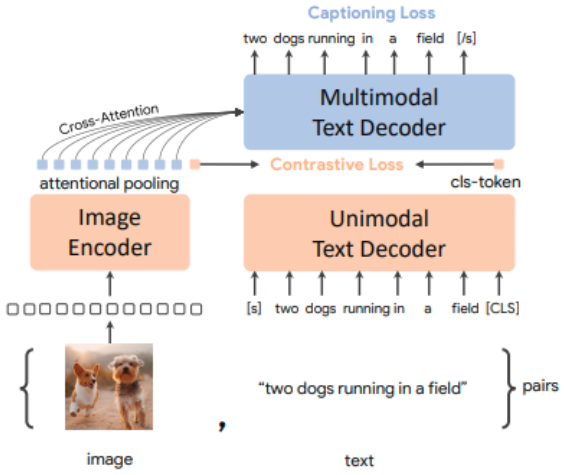
Encoder-Decoder Captioning

dual-encoder approach가 text를 encode 할 때, generative approach는 세분화된 granularity와, 모델이 y 의 extract tokenized text를 예측하는 것을 목표한다(?). 일반적인

encoder-decoder architecture과 같이, image encoder은 latent encoded features를 제공하고, text decoder은 paired y 에 대한 conditional likelihood를 최대화 하려 한다.

$$\mathcal{L}_{\text{Cap}} = -\sum_{t=1}^{T} \log P_\theta(y_t | y_{<t}, x).$$

Contrastive Captioners Pretraining



Algorithm 1 Pseudocode of Contrastive Captioners architecture.

```
# image, text.ids, text.labels, text.mask: paired {image, text} data
# con_query: 1 query token for contrastive embedding
# cap_query: N query tokens for captioning embedding
# cls_token_id: a special cls_token_id in vocabulary

def attentional_pooling(features, query):
  out = multihead_attention(features, query)
  return layer_norm(out)

img_feature = vit_encoder(image) # [batch, seq_len, dim]
con_feature = attentional_pooling(img_feature, con_query) # [batch, 1, dim]
cap_feature = attentional_pooling(img_feature, cap_query) # [batch, N, dim]

ids = concat(text.ids, cls_token_id)
mask = concat(text.mask, zeros_like(cls_token_id)) # unpad cls_token_id
txt_embs = embedding_lookup(ids)
unimodal_out = lm_transformers(txt_embs, mask, cross_attn=None)
multimodal_out = lm_transformers(
    unimodal_out[:, :-1, :], mask, cross_attn=cap_feature)
cls_token_feature = layer_norm(unimodal_out)[:, -1:, :] # [batch, 1, dim]

con_loss = contrastive_loss(con_feature, cls_token_feature)
cap_loss = softmax_cross_entropy_loss(
    multimodal_out, labels=text.labels, mask=text.mask)
```

`vit_encoder`: vision transformer based encoder; `lm_transformer`: language-model transformers.

Figure 2: Detailed illustration of CoCa architecture and training objectives.

위 그림은 CoCa를 나타낸다. Image representation을 neural network가 encode 하고, casual masking transformer decoder을 통해 text로 decode한다. CoCa는 일반적 트랜스포머와 달리 first half of decoder layers에서 cross-attention을 생략하고, 남은 decoder layer을 cascade한다. 따라서 아래 이미지와 같은 loss가 나온다.

$$\mathcal{L}_{\text{CoCa}} = \lambda_{\text{Con}} \cdot \mathcal{L}_{\text{Con}} + \lambda_{\text{Cap}} \cdot \mathcal{L}_{\text{Cap}},$$

- $\lambda_{Con}, \lambda_{Cap}$: Loss weighting hyper-parameters

Decoupled Text Decoder and CoCa Architecture

Captioning approach는 unconditional text representation을 사용하는데 text의 conditional likelihood를 optimize한다. 이러한 딜레마를 해결하고 두 모델을 하나로 합치기 위해 본 논문의 저자들은 decoder을 unimodal, multimodal component로 분리한 simple decoupled decoder 디자인을 제안하였다. 이는 unimodal decoder layers의 cross-attention mechanism을 스킵함으로 구현된다. 결국 bottom n_{uni} unimodal decoder

layers가 input text를 casual-masked self-attention을 통해 encode하고, top n_{multi} multimodal layers가 casual-masked self-attention을 적용하며, 둘 모두가 visual endocer의 output의 cross-attention을 진행한다. 모든 decoder layers는 tokens가 future tokens에 영향을 주는 것을 방지하고, captioning objective $\mathcal{L}_{Cap}$ 을 위해 muktimodal text decoder output을 사용하기 위해 straightforward된다.

Constrastive objective $\mathcal{L}_{Con}$ 에는 learnable [cls] token을 문장 끝에 더하여준다(세부사항은 논문 참고).

Attentional Poolers

본 논문의 예비 실험은 single pooled image embedding이 visual recognition task에서 global representation으로 도움을 준다는 것을 확인하였다.

CoCa는 task-specific attentional pooling을 통해 visual representations을 customize하는 것을 택하였다.

간단한 실험 결과는 다음과 같다(다른 결과는 논문 참고).

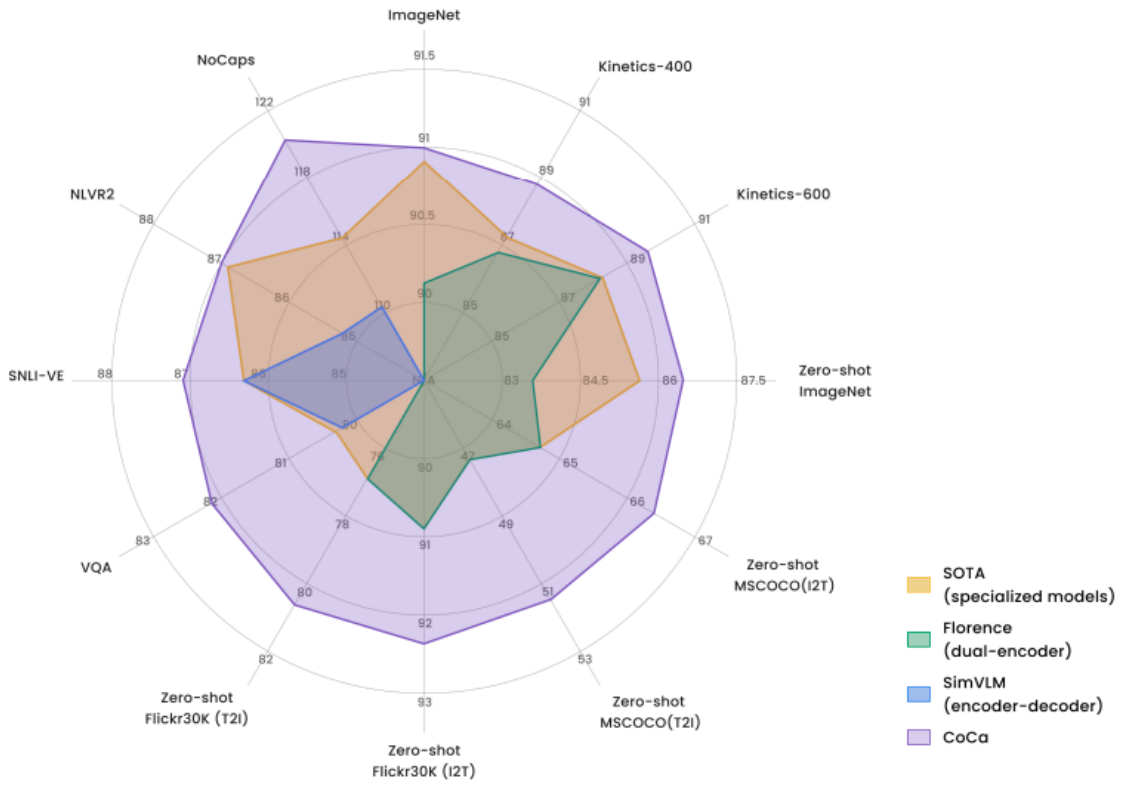


Figure 4: Comparison of CoCa with other image-text foundation models (without task-specific customization) and multiple state-of-the-art task-specialized models.

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**

Table 2: Image classification and video action recognition with frozen encoder or finetuned encoder.

Conclusion

In this work we present Contrastive Captioners (CoCa), 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 obtains a series of state-of-the-art performance with a single checkpoint on a wide spectrum of vision and vision-language problems. Our work bridges the gap among various pretraining approaches and we hope it motivates new directions for image-text foundation models.