CoCa: Contrastive Captioners are Image-Text Foundation Models

Определение

Contrastive Captioners - семество нейросетевых моделей, основанных на объединении single-encoder, dual-encoder и encoder-decoder парадигм, использующих при обучении два вида функций потерь: contrastive loss и captioning loss.

Single-encoder models

- Используются для vision и vision-language задач
- Тренируются с CrossEntropy функцией потерь

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

p(y) - one-hot или multy-hot закодированные метки изображений

- На выходе модель выдает представления, которые могут быть использованы для image или video understanding задач
- Недостаток: требуют строгой разметки датасета с фиксированным числом классов, что не позволяет обрабатывать произвольную разметку изображений

Dual-encoder models

- В отличие от Single-Encoder моделей может обрабатывать произвольный текст
- Состоит из двух энкодеров, один из которых обрабатывает изображения, второй - текстовые описания этих изображений.
- Энкодеры обучаются совместно, сопоставляя пары текст-изображение с другими парами в батче:

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

x_i и y_j - нормализованные эмбеддинги фото из i-ой пары и текста из j-ой пары

- Недостаток: отсутствие компонент для обучения на объединенных парах изображений и текстовых представлений

Encoder-Decoder models

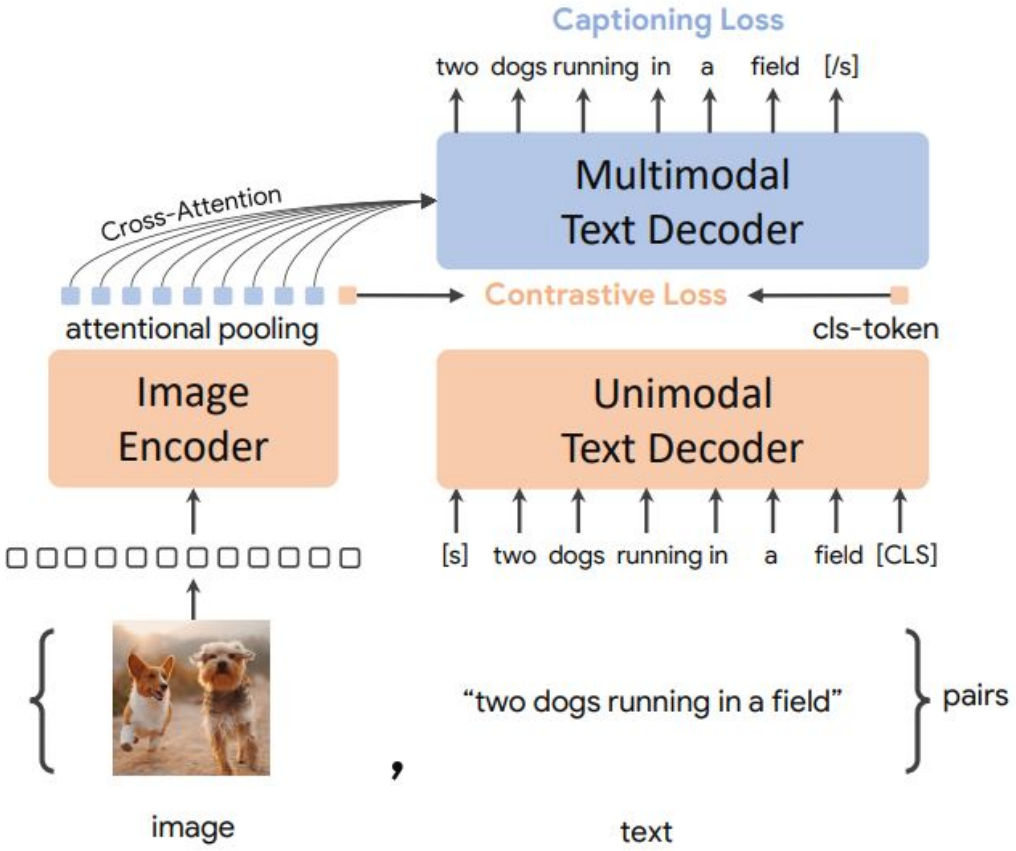
- В отличие от dual-encoder подхода, здесь мы кодируем не весь текст в одно представление, а предсказываем каждый токен текста авторегрессионно
- Image энкодер подает на вход декодеру представления изображений, а тот в свою очередь учится максимизировать условное правдоподобие представления текстового описания изображения:

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

- Обучаются с помощью teacher-forcing подхода
- Выдают совместные представления текста и изображения, которые уже могут использоваться для vision-language understanding и image-captioning задач

CoCa Особенности

- Архитектура состоит из трех крупных компонент:
  1. Энкодер изображений
  2. Унимодальный текстовый декодер
  3. Мультимодальный текстовый декодер



CoCa Псевдокод

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.

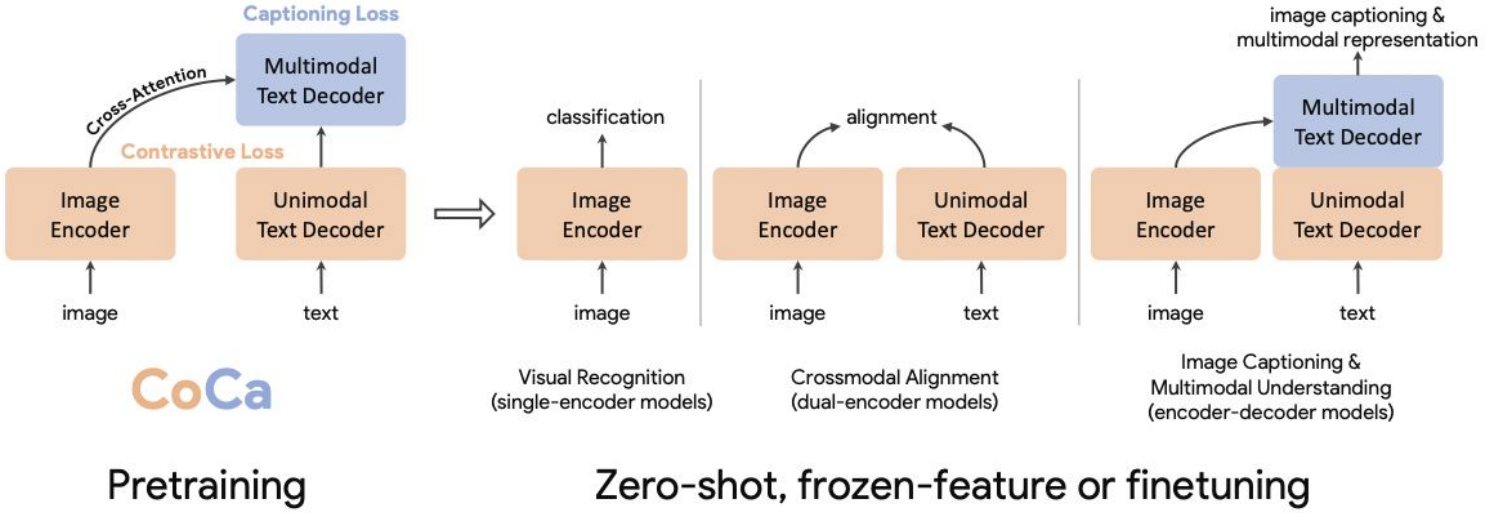
CoCa Особенности

- В унимодальном декодере текста не применяется cross-attention с эмбеддингами изображений. Этот декодер обучается выдавать унимодальные текстовые эмбеддинги.
- Cross-attention применяется в мультимодальном декодере для обучения декодера выдавать представления объектов вида изображение-текст
- Такой подход с разделением декодера на унимодальный и мультимодальный позволяет архитектуре выдавать как унимодальные, так и мультимодальные эмбеддинги
- Общие потери вычисляются следующим образом:

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

где параметры λ являются весами-гиперпараметрами модели для функций потерь

- Компоненты могут использоваться по отдельности для разных типов задач: visual recognition, vision-language alignment, image captioning и multimodal understanding

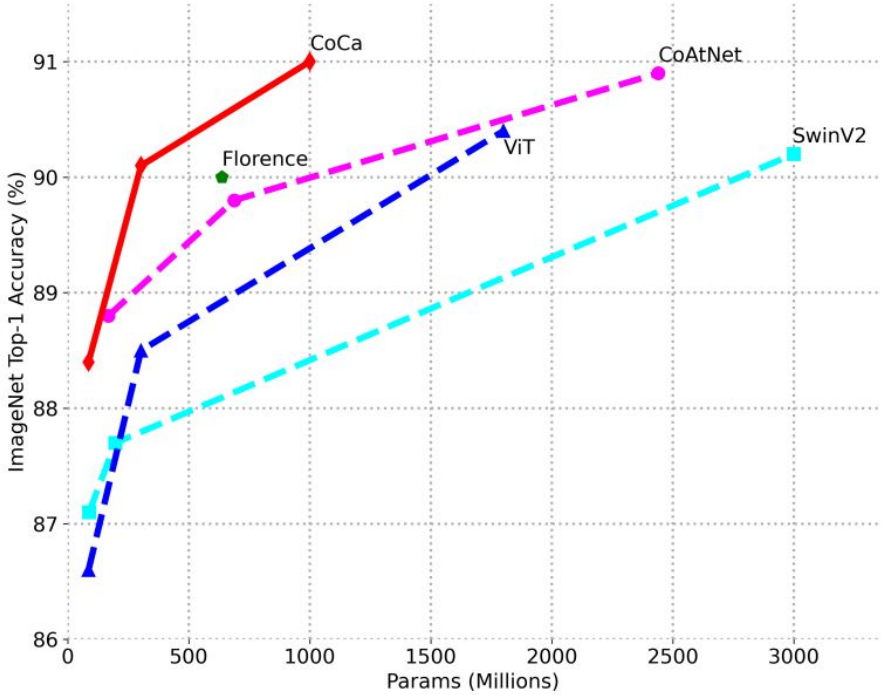


CoCa Особенности

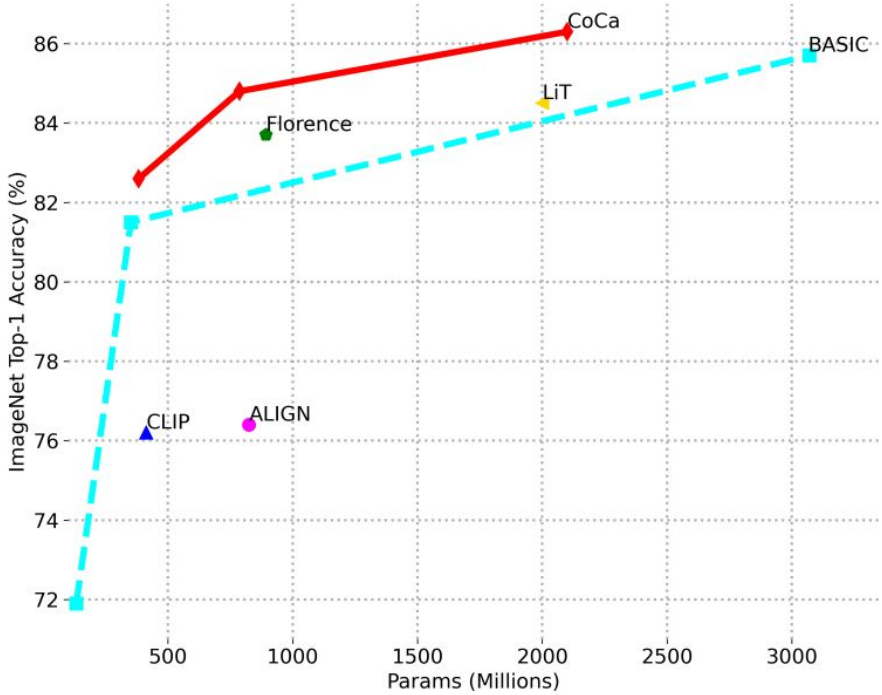
- Благодаря разделенному декодеру, модель может эффективно вычислять Contrastive и Captioning loss при обучении и таким образом обучать сразу две отдельные модели за один проход
- Однако CoCa модели при этом могут обучаться только end-to-end полностью с нуля
- CoCa архитектуры могут использоваться для zero-shot задач, но с некоторой информацией об обучающей выборке (Zero-Shot Transfer with Locked-image text Tuning)
- CoCa хорошо показывает себя во frozen-feature evaluation. В таких методах у архитектуры обучается только attention poolers

	Image Encoder			Text Decoder				Image / Text		
Model	Layers	MLP	Params	n_{uni}	n_{multi}	MLP	Params	Hidden	Heads	Total Params
CoCa-Base	12	3072	86M	12	12	3072	297M	768	12	383M
CoCa-Large	24	4096	303M	12	12	4096	484M	1024	16	787M
CoCa	40	6144	1B	18	18	5632	1.1B	1408	16	2.1B

Table 1: Size variants of CoCa. Both image encoder and text decoder are Transformers [19, 39].



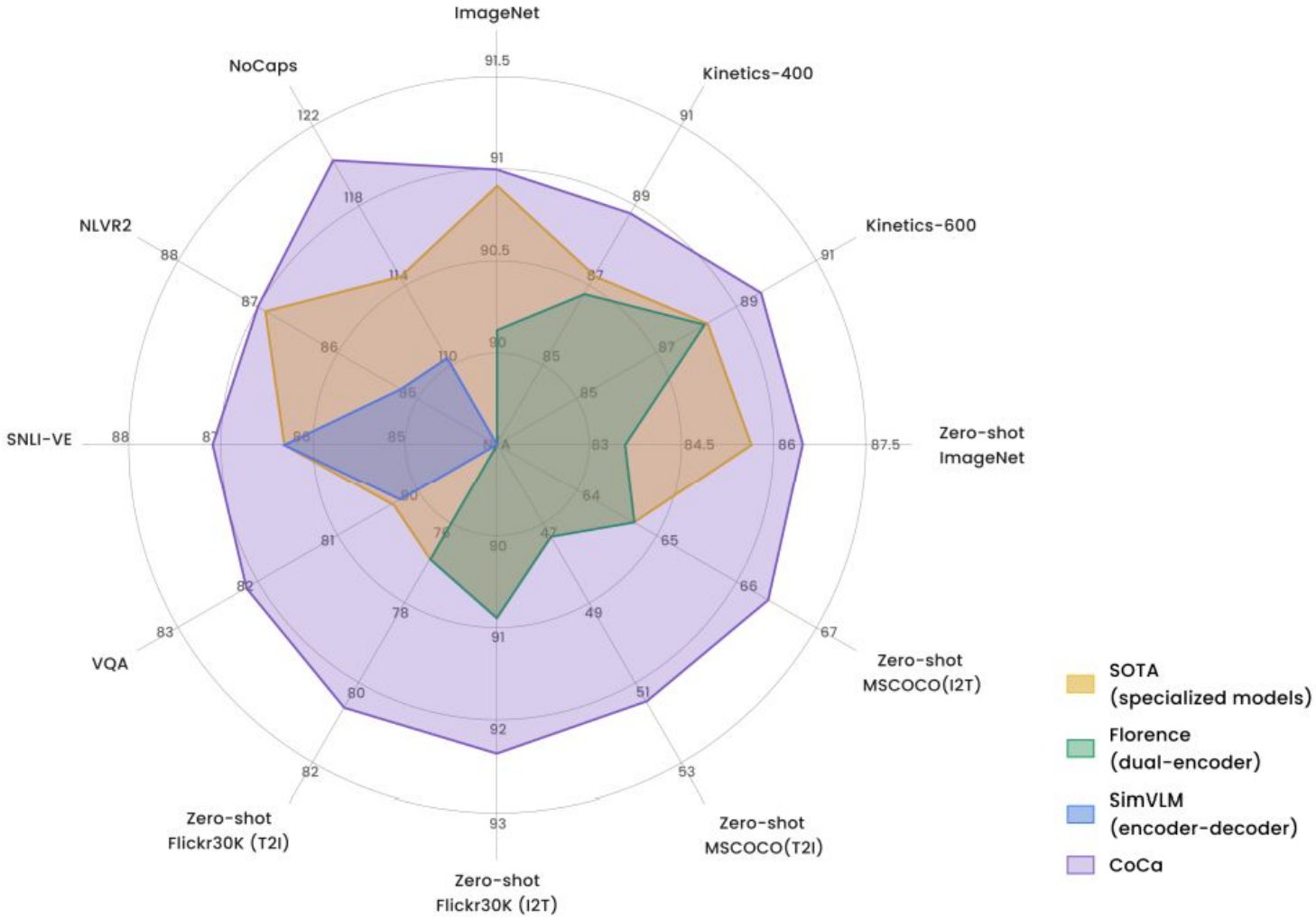
(a) Finetuned ImageNet Top-1 Accuracy.



(b) Zero-Shot ImageNet Top-1 Accuracy.

Эксперименты. Training setup

- CoCa модель обучалась с нуля за один этап как на данных изображение-текст, собранных с веб-ресурсов, так и на размеченных изображениях, обрабатывая все метки как обычный текст
- Датасеты: JFT-3B с именами меток в качестве парных текстов, ALIGN с зашумленными текстовыми описаниями
- Батч: 65536 пар изображение-текст, половина из JFT, половина из ALIGN
- Обучение длилось 500 тысяч шагов, что соответствует примерно 5 эпохам на JFT и 10 эпохам на ALIGN
- Cap_λ = 2.0, Con_λ = 1.0


ImageNet
91.5
91
90.5
90
Kinetics-400
91
89
87
85
Kinetics-600
91
89
87
85
Zero-shot
ImageNet
87.5
86
84.5
83
Zero-shot
MSCOCO(I2T)
67
66
65
64
Zero-shot
MSCOCO(T2I)
53
51
49
47
Zero-shot
Flickr30K (I2T)
93
92
91
90
Zero-shot
Flickr30K (T2I)
82
80
78
76
VQA
83
82
81
80
SNLI-VE
88
87
86
85
NLVR2
88
87
86
85
NoCaps
122
118
114
110
SOTA
(specialized models)
Florence
(dual-encoder)
SimVLM
(encoder-decoder)
CoCa

Решение простых image classification задач происходит с помощью энкодера с подстановкой attentional poolers и cross-entropy loss на его выходе

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.

Добавляем унимодальный декодер - решаем crossmodal alignment задачи

	Flickr30K (1K test set)						MSCOCO (5K test set)					
	Image → Text			Text → Image			Image → Text			Text → Image		
Model	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP [12]	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.8	62.4	72.2
ALIGN [13]	88.6	98.7	99.7	75.7	93.8	96.8	58.6	83.0	89.7	45.6	69.8	78.6
FLAVA [35]	67.7	94.0	-	65.2	89.4	-	42.7	76.8	-	38.4	67.5	-
FILIP [61]	89.8	99.2	99.8	75.0	93.4	96.3	61.3	84.3	90.4	45.9	70.6	79.3
Florence [14]	90.9	99.1	-	76.7	93.6	-	64.7	85.9	-	47.2	71.4	-
CoCa-Base	89.8	98.8	99.8	76.8	93.7	96.8	63.8	84.7	90.7	47.5	72.4	80.9
CoCa-Large	91.4	99.2	99.9	79.0	95.1	97.4	65.4	85.6	91.4	50.1	73.8	81.8
CoCa	**92.5**	**99.5**	**99.9**	**80.4**	**95.7**	**97.7**	**66.3**	**86.2**	**91.8**	**51.2**	**74.2**	**82.0**

Table 3: Zero-shot image-text retrieval results on Flickr30K [62] and MSCOCO [63] datasets.

Model	ImageNet	ImageNet-A	ImageNet-R	ImageNet-V2	ImageNet-Sketch	ObjectNet	Average
CLIP [12]	76.2	77.2	88.9	70.1	60.2	72.3	74.3
ALIGN [13]	76.4	75.8	92.2	70.1	64.8	72.2	74.5
FILIP [61]	78.3	-	-	-	-	-	-
Florence [14]	83.7	-	-	-	-	-	-
LiT [32]	84.5	79.4	93.9	78.7	-	81.1	-
BASIC [33]	85.7	85.6	95.7	80.6	76.1	78.9	83.7
CoCa-Base	82.6	76.4	93.2	76.5	71.7	71.6	78.7
CoCa-Large	84.8	85.7	95.6	79.6	75.7	78.6	83.3
CoCa	**86.3**	**90.2**	**96.5**	**80.7**	**77.6**	**82.7**	**85.7**

Table 4: Zero-shot image classification results on ImageNet [9], ImageNet-A [64], ImageNet-R [65], ImageNet-V2 [66], ImageNet-Sketch [67] and ObjectNet [68].

visual entailment

\+

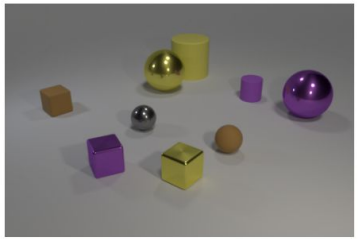
- *Two woman are holding packages.*
- *The sisters are hugging goodbye while holding to go packages after just eating lunch.*
- *The men are fighting outside a deli.*

=

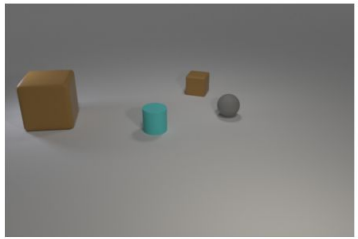
- *Entailment*
- *Neutral*
- *Contradiction*

Premise *Hypothesis* *Answer*

visual reasoning



(a) *What number of cylinders are small purple things or yellow rubber things?*
Predicted: 2



(b) *What color is the other object that is the same shape as the large brown matte thing?*
Predicted: Brown

После добавления мультимодального декодера модель начинает решать multimodal understanding задачи: visual question answering, visual entailment, visual reasoning

	MSR-VTT Full					
	Text → Video			Video → Text		
Method	R@1	R@5	R@10	R@1	R@5	R@10
CLIP [69]	21.4	41.1	50.4	40.3	69.7	79.2
Socratic Models [70]	-	-	-	44.7	71.2	80.0
CLIP [69] (subset)	23.3	44.2	53.6	43.3	73.3	**81.8**
Socratic Models [70] (subset)	-	-	-	46.9	**73.5**	81.3
CoCa (subset)	**30.0**	**52.4**	**61.6**	**49.9**	73.4	81.4

Table 5: Zero-shot Video-Text Retrieval on MSR-VTT Full test set.

	VQA		SNLI-VE		NLVR2	
Model	test-dev	test-std	dev	test	dev	test-p
UNITER [26]	73.8	74.0	79.4	79.4	79.1	80.0
VinVL [27]	76.6	76.6	-	-	82.7	84.0
CLIP-ViL [73]	76.5	76.7	80.6	80.2	-	-
ALBEF [36]	75.8	76.0	80.8	80.9	82.6	83.1
BLIP [37]	78.3	78.3	-	-	82.2	82.2
OFA [17]	79.9	80.0	90.3†	90.2†	-	-
VLMo [30]	79.9	80.0	-	-	85.6	86.9
SimVLM [16]	80.0	80.3	86.2	86.3	84.5	85.2
Florence [14]	80.2	80.4	-	-	-	-
METER [74]	80.3	80.5	-	-	-	-
CoCa	**82.3**	**82.3**	**87.0**	**87.1**	**86.1**	**87.0**

Table 6: Multimodal understanding results comparing vision-language pretraining methods. †OFA uses both image and text premises as inputs while other models utilize the image only.

Image Captioning

a hand holding a san francisco 49ers football

a row of cannons with the eiffel tower in the background

a white van with a license plate that says we love flynn

a person sitting on a wooden bridge holding an umbrella

a truck is reflected in the side mirror of a car

	MSCOCO				NoCaps			
					Valid		Test	
	B@4	M	C	S	C	S	C	S
CLIP-ViL [73]	40.2	29.7	134.2	23.8	-	-	-	-
BLIP [37]	40.4	-	136.7	-	113.2	14.8	-	-
VinVL[27]	41.0	31.1	140.9	25.4	105.1	14.4	103.7	14.4
SimVLM [16]	40.6	33.7	143.3	**25.4**	112.2	-	110.3	14.5
LEMON [80]	**41.5**	30.8	139.1	24.1	117.3	15.0	114.3	14.9
$\text{LEMON}_{\text{SCST}}$ [80]†	42.6	31.4	145.5	25.5	-	-	-	-
OFA [17]†	43.5	31.9	149.6	26.1	-	-	-	-
CoCa	40.9	**33.9**	**143.6**	24.7	**122.4**	**15.5**	**120.6**	**15.5**

Table 7: Image captioning results on MSCOCO and NoCaps (B@4: BLEU@4, M: METEOR, C: CIDEr, S: SPICE). †Models finetuned with CIDEr optimization.

Эффективность использования CoCa компонент по отдельности

- CoCa модели, тренирующиеся на captioning loss выдают примерно такие же результаты как и обычный image энкодер, обученнный с помощью CrossEntropy, что означает, что generative обучение включает в себя и обучение задаче классификации, вследствии чего CoCa модели не требуют предобученного image энкодера.
- Эксперименты показали, что captioning loss улучшают качество на VQA и zero-shot multimodal alignment задачах, при том, что в отличие от contrastive моделей CoCa не требует дополнительных слияний представлений для этих задач

Вывод

- CoCa представляет собой универсальную архитектуру, способную помимо image captioning решать большое количество down-stream задач на уровне топовых моделей, а в большинстве задач даже обгоняя их по качеству.