

# CoCa: Contrastive Captioners are Image-Text Foundation Models

Рецензия

# Авторы

Jiahui Yu, Zirui Wang,









Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, Yonghui Wu

Google Research

Pretraining CoCa takes about 5 days on 2,048 CloudTPUv4 chips

Ранние совметсные статьи: SimVLM: Simple Visual Language Model Pretaining with Weak Supervision

# А что есть в открытом доступе из топа лидерборда ImageNet?

1	<b>CoCa</b> (finetuned)	91.0%	2100M		CoCa: Contrastive Captioners are Image-Text Foundation Models	 	2022	<div>ALIGN</div> <div>Transformer</div> <div>JFT-3B</div>
2	<b>Model soups</b> (BASIC-L)	90.98%	2440M		Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time	 	2022	<div>Conv+Transformer</div> <div>JFT-3B</div> <div>ALIGN</div>
3	<b>Model soups</b> (ViT-G/14)	90.94%	1843M		Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time	 	2022	<div>JFT-3B</div> <div>Transformer</div>
11	<b>DaViT-H</b>	90.2%	362M	334	DaViT: Dual Attention Vision Transformers	 	2022	<div>Transformer</div>

# Парные энкодера

- CLIP обучался на шумных данных из интернет
- ALIGN взяли больше данных и почистили их
- BASIC ещё больше данных, модель и длину батча

	ALIGN ( <a href="#">Jia et al., 2021</a> )	CLIP ( <a href="#">Radford et al., 2021</a> )	BASIC (ours)
ImageNet	76.4	76.2	<b>85.7 (+9.3)</b>
ImageNet-A	75.8	77.2	<b>85.6 (+8.4)</b>
ImageNet-R	92.2	88.9	<b>95.7 (+3.5)</b>
ImageNet-V2	70.1	70.1	<b>80.6 (+10.5)</b>
ImageNet-Sketch	64.8	60.2	<b>76.1 (+11.3)</b>
ObjectNet	72.2	72.3	<b>82.3 (+10.1)</b>
Average	74.5	74.2	<b>84.3 (+10.1)</b>

# Применимость парных энкодеров

- Есть обученный CLIP
- Данные для обучения CLIP можно найти
- Данные используемые в статьях для ALIGN, BASIC, CoCa - закрытые

# Енкодер-Декодер

- SIMVLM обучали на большом шумном датасете, кода нет
- METER обучали на COCO, код есть
- OFA обучали на COCO, код есть

:(

- no code, loss to community
- from scratch
- visual tasks with spatial localization