# CoCa: Contrastive Captioners are Image-Text Foundation Models

Рецензия

## Авторы

Jiahui Yu, Zirui Wang,

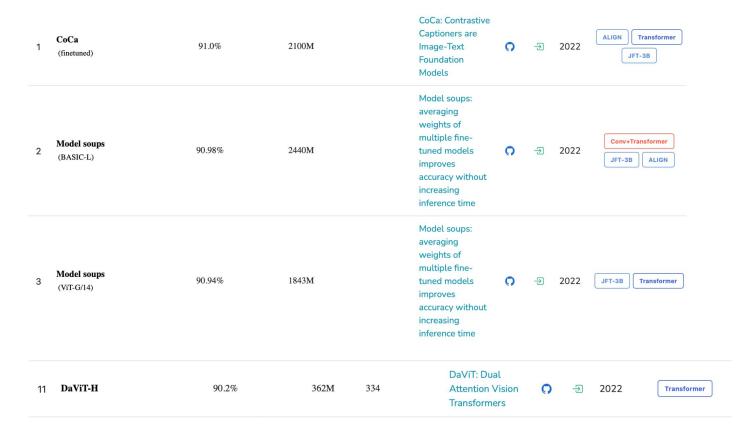
Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, Yonghui Wu

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?



#### Парные енкодера

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

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

### Применимость парных енкодеров

- 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
- МЕТЕК обучали на СОСО, код есть
- ОFA обучали на СОСО, код есть

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