Faculty of Computer Science

Научно-исследовательский семинар

2022

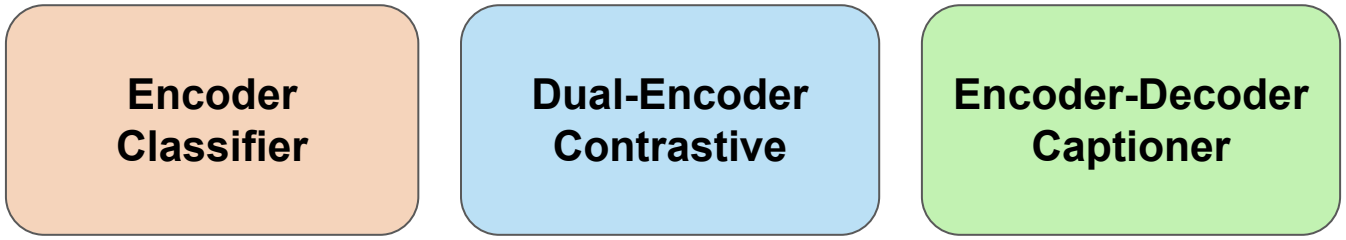
CoCa: Contrastive Captioners are Image-Text Foundation Models

Review

Akulov Dmitry

Суть работы

- Может решать большой спектр задач
- Легко делать Fine-tuning и zero-shot learning

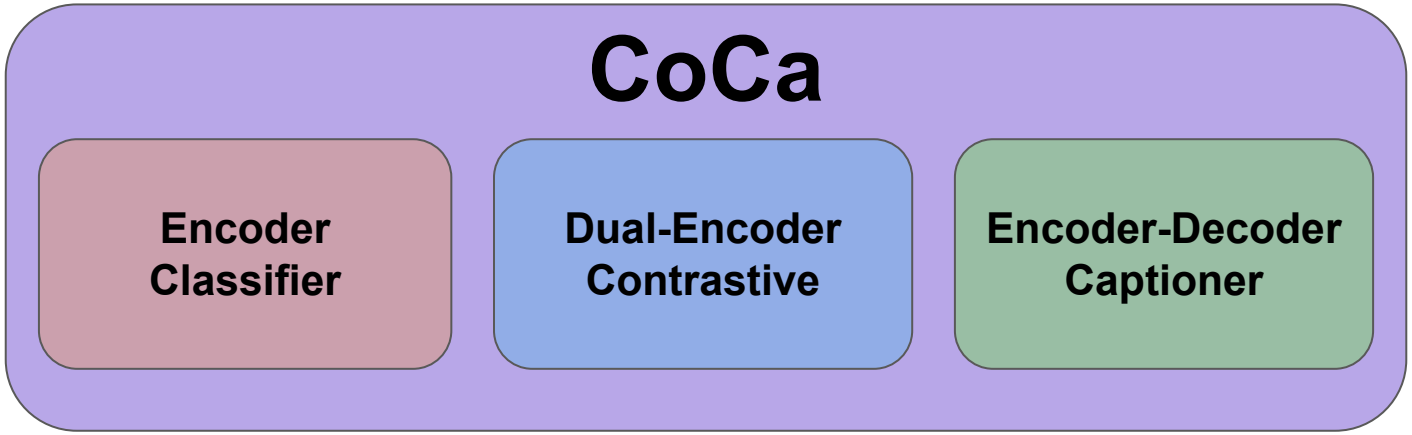


Публикация

- Залито на arXiv 04.05.2022
- Опубликовано в журнале Transactions on Machine Learning Research
- Цитирования: 63
  - Продолжения нет
  - Одно интересное упоминание в качестве ImageNet SOTA

Суть работы

- Может решать большой спектр задач
- Легко делать Fine-tuning и zero-shot learning



Публикация

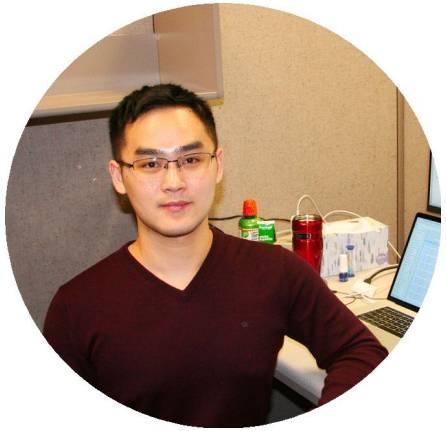
- Залито на arXiv 04.05.2022
- Опубликовано в журнале Transactions on Machine Learning Research
- Цитирования: 63
  - Продолжения нет
  - Одно интересное упоминание в качестве ImageNet SOTA

Авторы

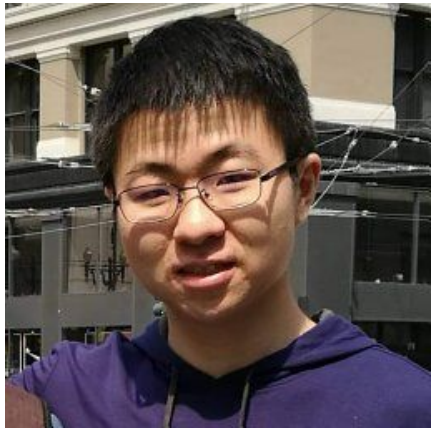
Google

С 2017 года
60 статей
8446 цитирований

CV
NLP
Audio



Jiahui Yu



Zirui Wang

С 2019 года
13 статей
775 цитирований

NLP
Opt
CV

Совместная статья (08.2021): Encoder-Decoder Captioner

Simvlm: Simple visual language model pretraining with weak supervision

Аналог

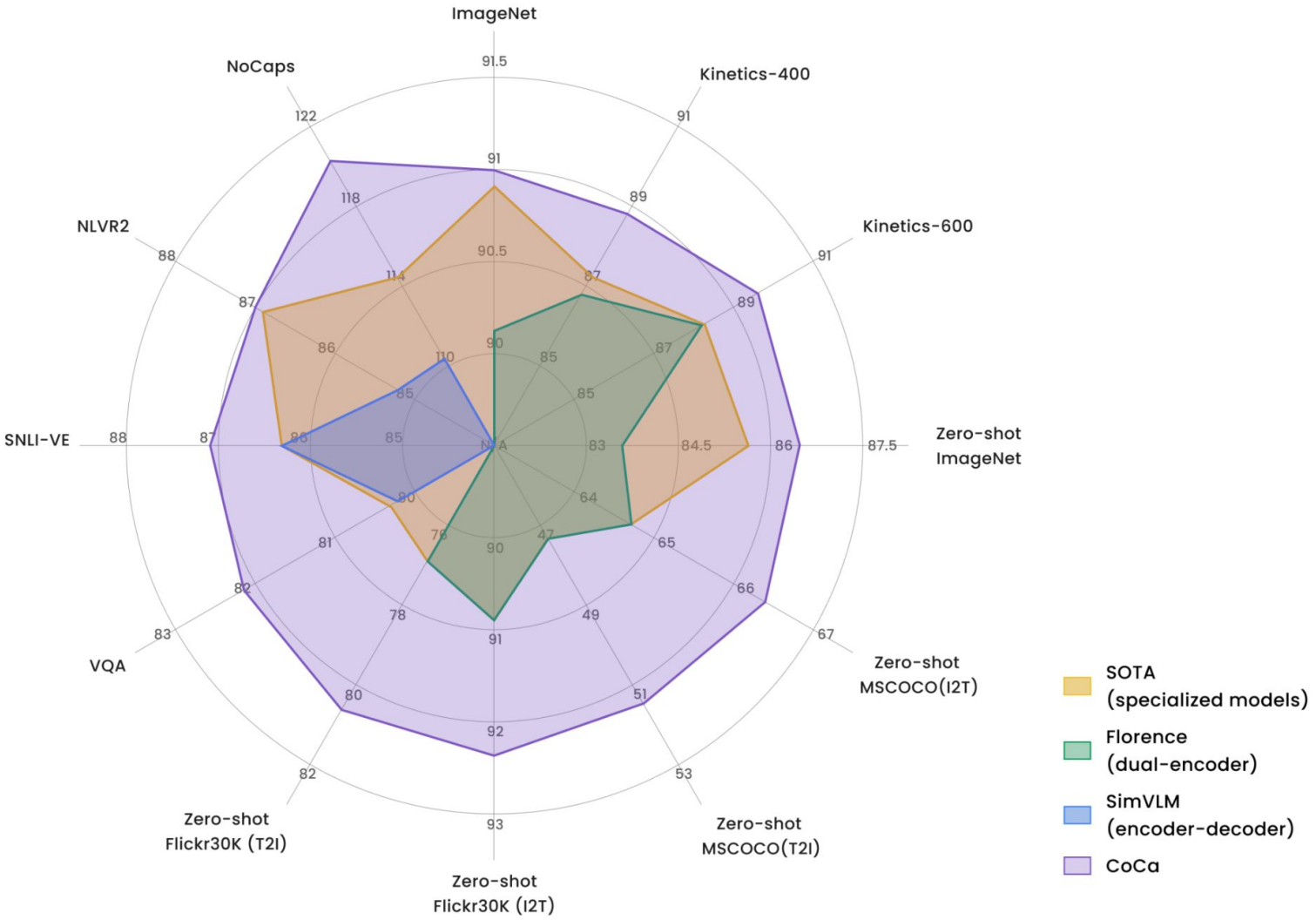
Florence: A New Foundation Model for Computer Vision (22.11.2021)

Аналог

Florence: A New Foundation Model for Computer Vision (22.11.2021)

Datasets		CoCa	Datasets		Florence
ImageNet Kinetics	1.	visual recognition	ImageNet Kinetics CIFAR10 Cross-Domain Few-Shot learning **benchmark** (FS)	1.	Classification
MSCOCO (ZH) Flickr30K (ZH) MSR-VTT (ZH)	2.	crossmodal alignment	MSCOCO (ZH+FT) Flickr30K (ZH+FT)	2.	Retrieval
VQA v2 SNLI-VE NLVR2 MSCOCO	3.	image captioning and multimodal understanding	VQA v2	3.	VL Representation Learning
			…	4.	Object Detection Text-to-Video Video Action Recognition

Результат


ImageNet
Kinetics-400
Kinetics-600
Zero-shot ImageNet
Zero-shot MSCOCO(I2T)
Zero-shot MSCOCO(T2I)
Zero-shot Flickr30K (I2T)
Zero-shot Flickr30K (T2I)
VQA
SNLI-VE
NLVR2
NoCaps
SOTA (specialized models)
Florence (dual-encoder)
SimVLM (encoder-decoder)
CoCa

Плюсы

- Хорошее техническое улучшение
- Сами подходы, которые были объединены, хорошо описаны

Минусы

- Были описаны задачи, которые уже покрываются теми моделями, что были объединены

Идеи по улучшению от авторов

- Может оказаться чувствительной к corrupted(~поврежденным) изображениям.
- Исследовать ширину используемости модели

Идеи по улучшению от меня

- Суть работы - объединение подходов. Как много еще можно наобъединять?
  - Что было бы полезно добавить именно к их модели?
  - Почему выбрали именно такие? Можно ли что-то выкинуть?

Источники

- CoCa: Contrastive Captioners are Image-Text Foundation Models https://arxiv.org/abs/2205.01917v2
- SimVLM: Simple Visual Language Model Pretraining With Weak Supervision https://arxiv.org/abs/2108.10904
- Florence: A New Foundation Model for Computer Vision https://arxiv.org/abs/2111.11432