

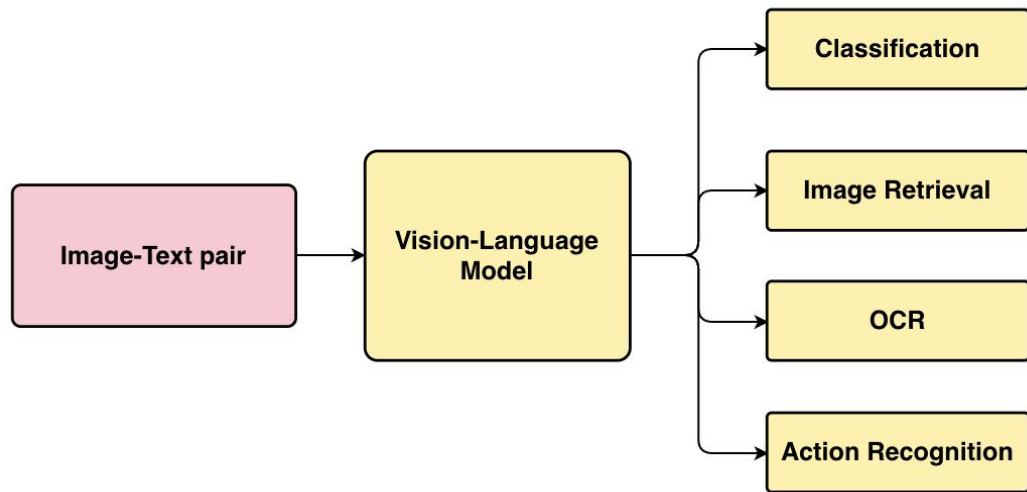
CLIP: Connecting Text and Images

Матосян А. А.

План доклада:

1. Vision-Language Models
2. Основная идея CLIP
3. Архитектура CLIP
4. Zero-shot Learning
5. Эксперименты с CLIP
6. ALIGN

Vision-Language Models



- Мультимодальные модели, которые извлекают взаимосвязи между изображениями и текстом
- Могут быть адаптированы под решение различных задач
- Часто используются для transfer-learning
- Часто weakly или self supervised

CLIP: Contrastive Language-Image Pre-training

- VLM от OpenAI
- Представлена в 2021 году
- Предобучена решать задачу сопоставления заголовков и изображений



[“Lion and cheetah”,
“Cat and dog”,
“Rabbit and bird”]

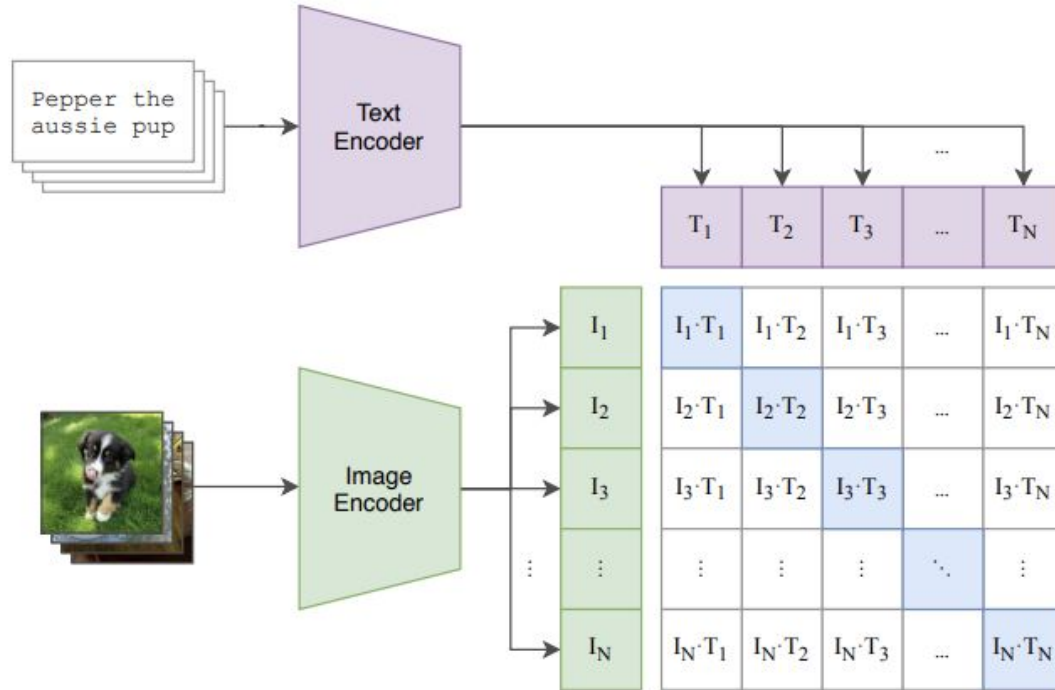


```
[{'score': 0.9950, 'label': 'cat and dog'},  
{ 'score': 0.0048, 'label': 'rabbit and lion'},  
{ 'score': 8.81e-05, 'label': 'lion and cheetah'}]
```

Мотивация CLIP

- Разметка данных требует много усилий, человеческого труда и дорого обходится
- Стандартные модели компьютерного зрения не универсальны, не могут быть использованы для решения разных задач
- Неустойчивость известных моделей к смещениям в данных

Архитектура CLIP: Contrastive pre-training



Архитектура CLIP: Contrastive pre-training

- **Text-encoder:** CBOW или Transformer (encoder block)
- **Image-encoder:** ResNet или Vision Transformer
- В **матрице схожести** записаны cosine similarity векторов полученных из encoder-ов
- Во время обучения максимизируются диагональные элементы матрицы

$$L_{i2t} = -\frac{1}{N} \sum_i \log \frac{\exp(x_i^\top y_i / \sigma)}{\sum_{j=1}^N \exp(x_i^\top y_j / \sigma)}$$

$$L_{t2i} = -\frac{1}{N} \sum_i \log \frac{\exp(y_i^\top x_i / \sigma)}{\sum_{j=1}^N \exp(y_i^\top x_j / \sigma)}$$

Датасет для предобучения CLIP



WIKIPEDIA
The Free Encyclopedia



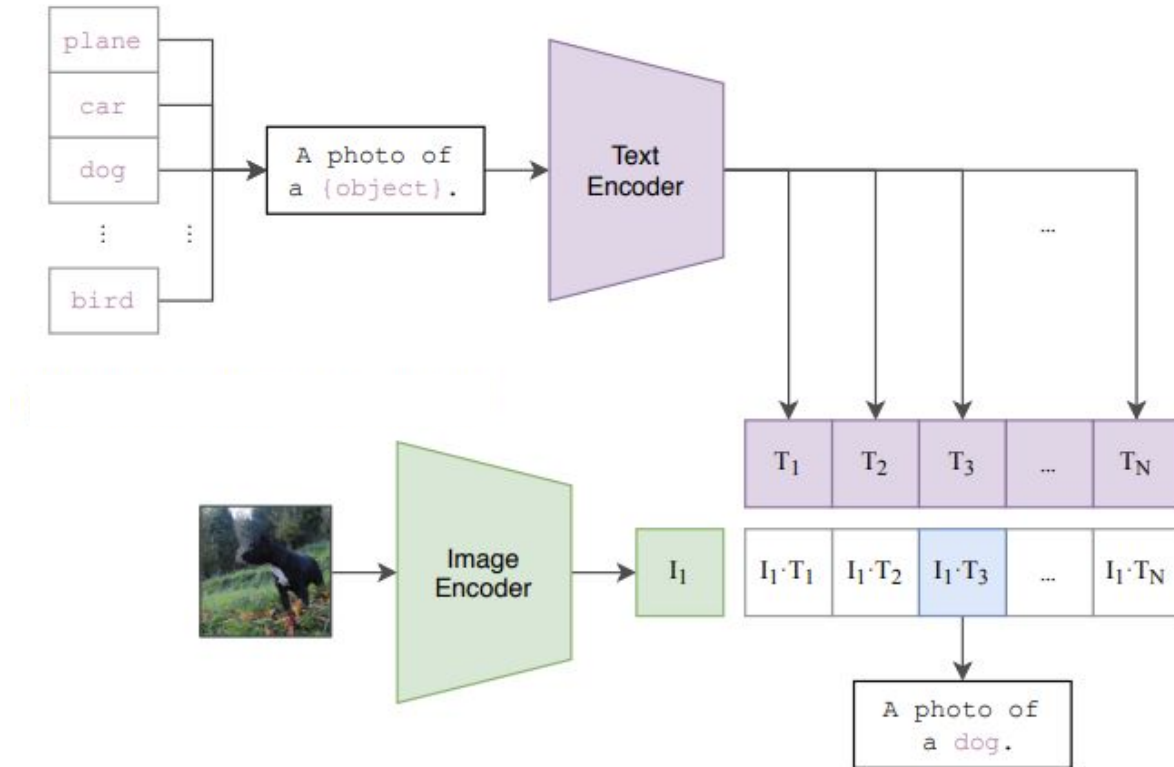
Queries:

- Собрали 500 тыс. самых часто встречаемых слов (> 100 раз)
- Аугментация: bi-gram

Итоговый датасет:

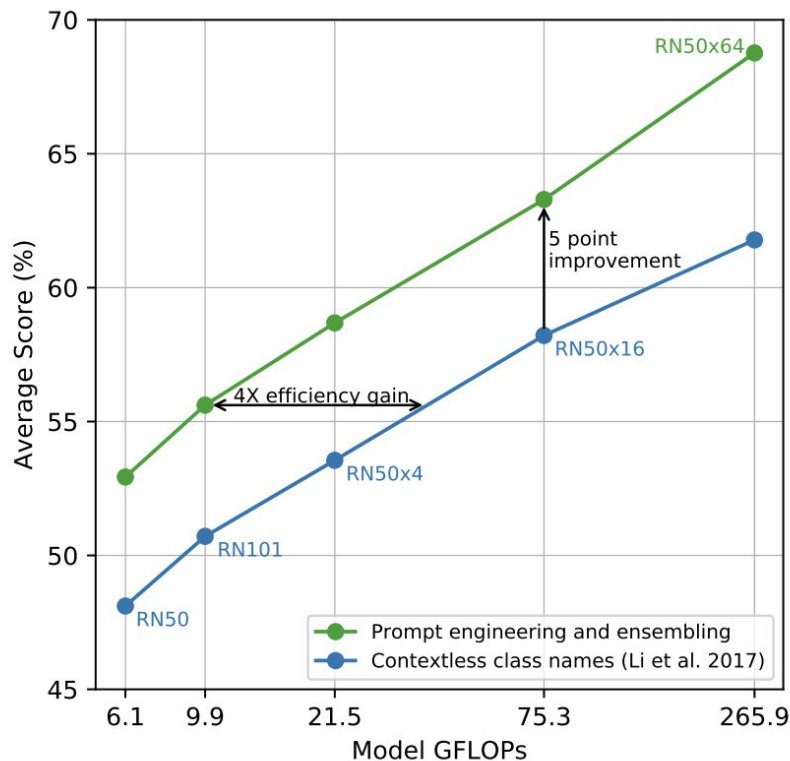
- 400 млн. пар <text, image>
- Отбирались пары, с текстом содержащим хотя бы один query
- ~20000 пар для каждого query

Архитектура CLIP: Zero-shot learning

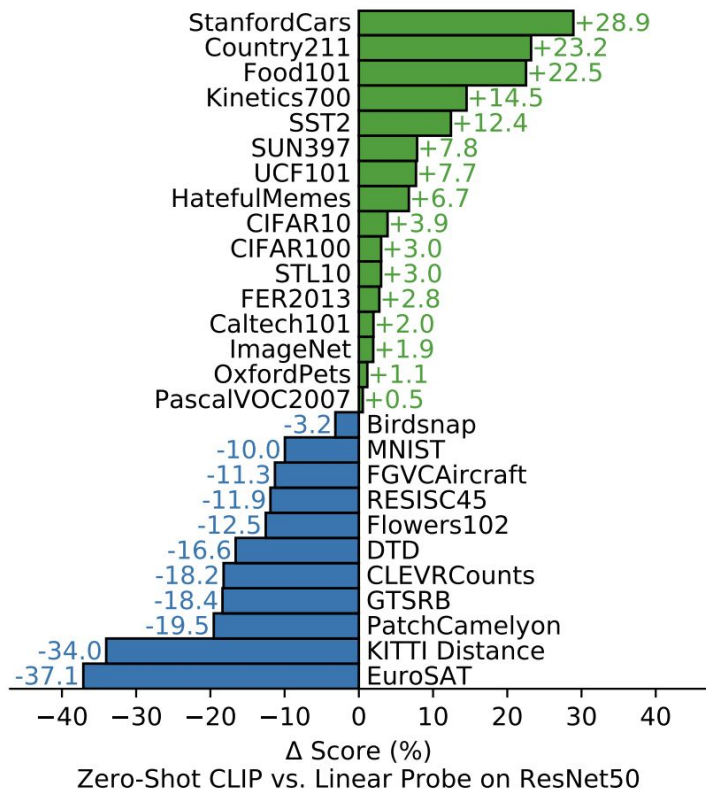


Ансамблирование и prompt engineering в CLIP

- Базовый промт: “A photo of a {label}.”
- Промт с уточнением: например, для датасета Oxford-IIIT Pet “A photo of {label}, a type of pet.”
- Для ансамблирования: несколько промптов с разными уточнениями



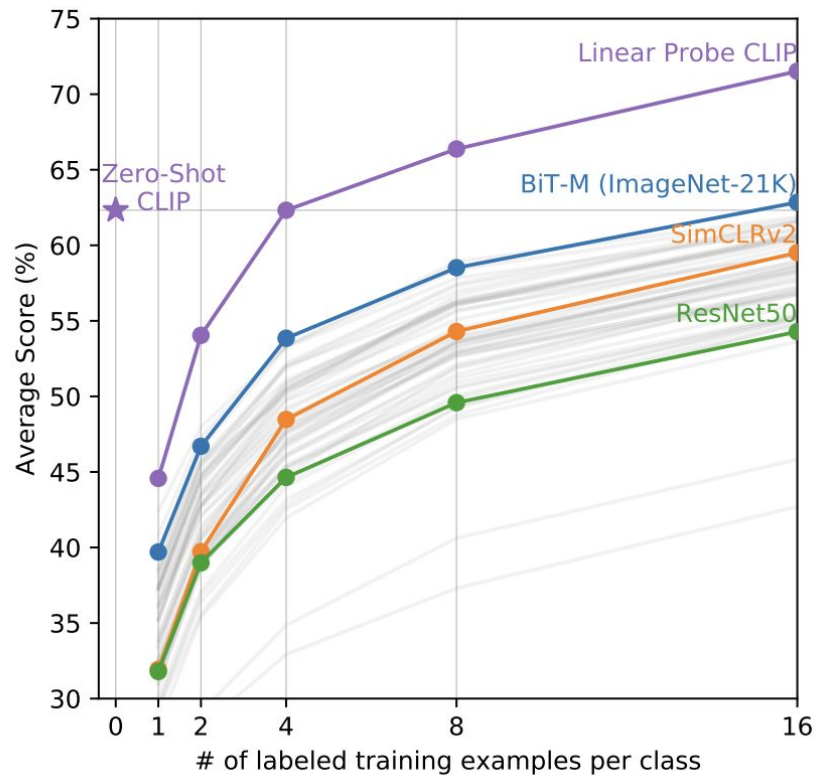
Zero-shot VS Linear-probe



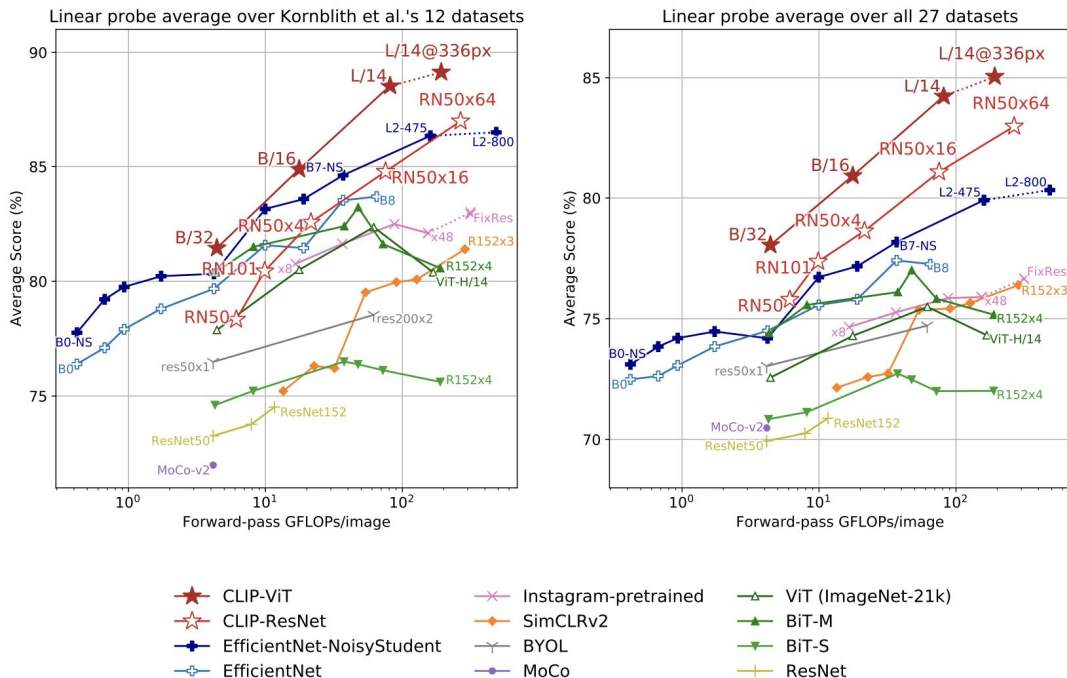
- Zero-shot CLIP показывает на 16/27 датасетов качество лучше, чем fine-tuned ResNet-50
- CLIP проигрывает по качеству на специализированных датасетах, например на EuroSAT, который предназначен для задачи классификации спутниковых снимков

Few-shot learning

- Zero-shot CLIP сопоставим с 4-shot Linear Probe CLIP, а также с 16-shot Linear Probe BiT-M
- 16-shot Linear Probe CLIP показал лучшее качество

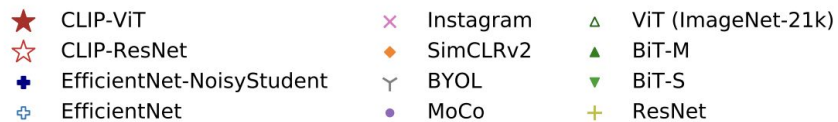
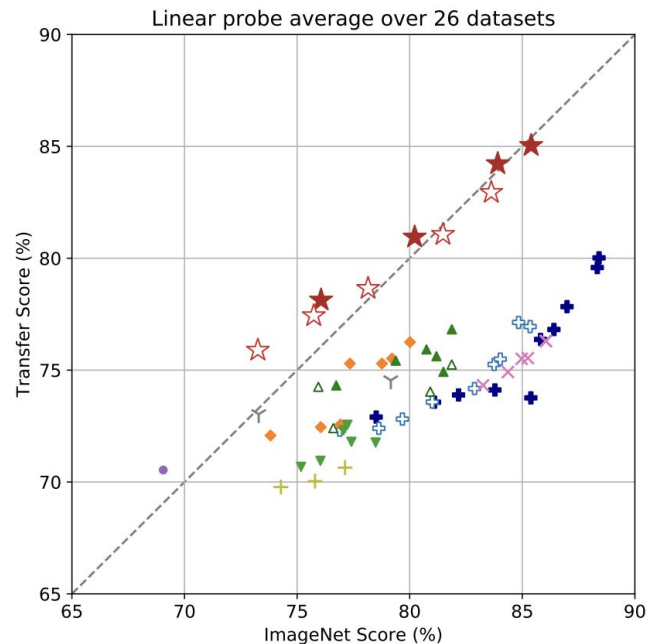
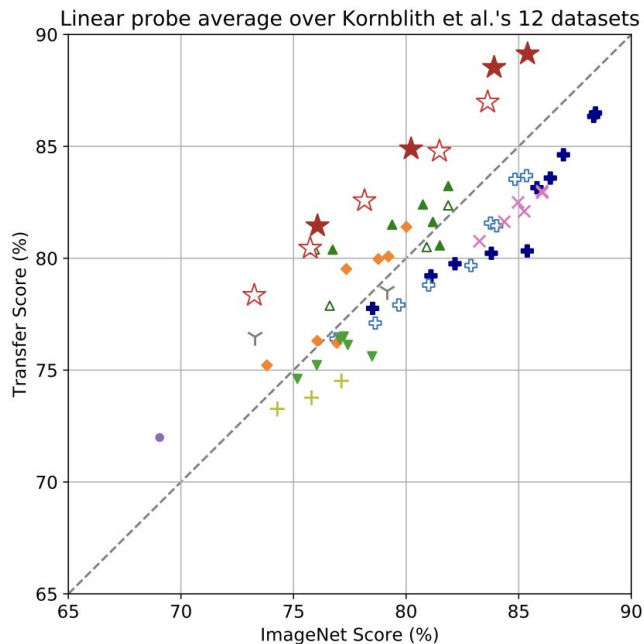


Обобщающая способность модели









- Linear probe CLIP сравнивается с SOTA моделями на ImageNet
- CLIP с Vision Transformer-ом в 3 раза эффективнее, чем с ResNet и показывает качество выше

Устойчивость



Устойчивость

	Dataset Examples	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet		76.2	76.2	0%
ImageNetV2		64.3	70.1	+5.8%
ImageNet-R		37.7	88.9	+51.2%
ObjectNet		32.6	72.3	+39.7%
ImageNet Sketch		25.2	60.2	+35.0%
ImageNet-A		2.7	77.1	+74.4%

Ограничения

- Плохое качество в сложных, абстрактных задачах. Например: задача предсказания расстояния до ближайшей машины на изображении (KITTI Distance)
- CLIP испытывает сложности в некоторых fine-grained classification задачах. Пример: задача определения модели автомобиля
- CLIP все еще плохо обобщается на изображения, не входящие в его обучающий набор данных. Например, на датасете рукописных цифр MNIST, CLIP достигает точности 88%
- Zero-shot CLIP чувствителен к контексту, поэтому требуются prompt engineering эксперименты для новых задач

Выводы по CLIP

- Vision-Language модель, которая сопоставляет текст изображению
- Natural Language Supervised
- Zero-Shot learning CLIP показывает хорошие результаты, лучше чем другие модели
- Благодаря contrastive learning, CLIP понимает сложные визуальные концепты, что делает эту модель очень гибкой и устойчивой к сдвигам распределений
- Не подходит для специфических задач

ALIGN: A Large-scale Image and Noisy-text embedding

- Собран новый датасет на 1.8 млрд пар
- Пожертвовали качеством данных в пользу количества
- **Text-encoder:** BERT
- **Image-encoder:** EfficientNet-L2



"motorcycle front wheel"



"thumbnail for version as of 21 57 29 june 2010"



"file frankfurt airport skyline 2017 05 jpg"



"file london barge race 2 jpg"



"moustache seamless wallpaper design"



"st oswalds way and shops"

Количественные результаты

		Flickr30K (1K test set)						MSCOCO (5K test set)					
		image → text			text → image			image → text			text → image		
Zero-shot	ImageBERT	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
	UNITER	70.7	90.2	94.0	54.3	79.6	87.5	44.0	71.2	80.4	32.3	59.0	70.2
	CLIP	83.6	95.7	97.7	68.7	89.2	93.9	-	-	-	-	-	-
	ALIGN	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.8	62.4	72.2
Fine-tuned		88.6	98.7	99.7	75.7	93.8	96.8	58.6	83.0	89.7	45.6	69.8	78.6
	GPO	88.7	98.9	99.8	76.1	94.5	97.1	68.1	90.2	-	52.7	80.2	-
	UNITER	87.3	98.0	99.2	75.6	94.1	96.8	65.7	88.6	93.8	52.9	79.9	88.0
	ERNIE-ViL	88.1	98.0	99.2	76.7	93.6	96.4	-	-	-	-	-	-
	VILLA	87.9	97.5	98.8	76.3	94.2	96.8	-	-	-	-	-	-
	Oscar	-	-	-	-	-	-	73.5	92.2	96.0	57.5	82.8	89.8
	ALIGN	95.3	99.8	100.0	84.9	97.4	98.6	77.0	93.5	96.9	59.9	83.3	89.8

Model	ImageNet	ImageNet-R	ImageNet-A	ImageNet-V2
CLIP	76.2	88.9	77.2	70.1
ALIGN	76.4	92.2	75.8	70.1

Интересное наблюдение



- На пространстве полученных эмбедингов хорошо определяются операции сложения и вычитания

ИСТОЧНИКИ

[Learning Transferable Visual Models From Natural Language Supervision](#)

[Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision](#)