

# CLIP: Connecting Text and Images

Терехова Юлия, 182

# Мотивация использовать NLP+CV

CLIP (*Contrastive Language–Image Pre-training*)

FOOD101

**guacamole** (90.1%) Ranked 1 out of 101 labels



✓ a photo of **guacamole**, a type of food.

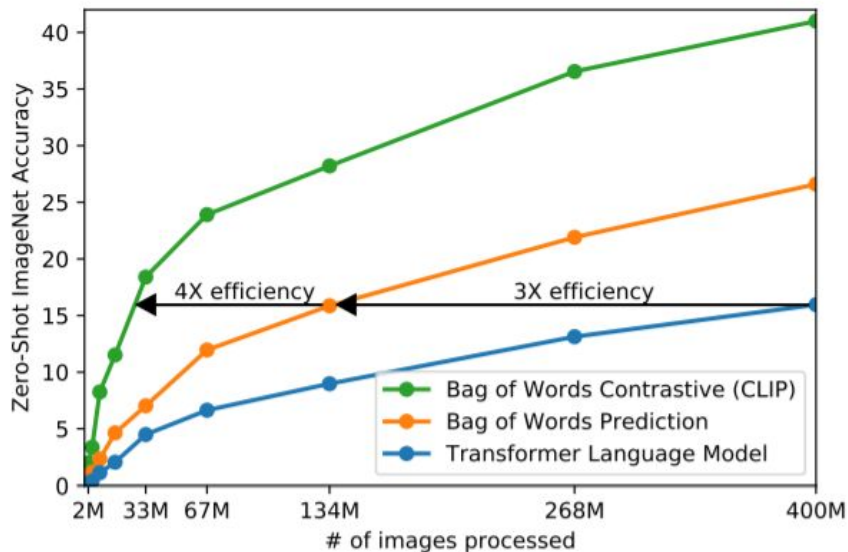
✗ a photo of **ceviche**, a type of food.

✗ a photo of **edamame**, a type of food.

✗ a photo of **tuna tartare**, a type of food.

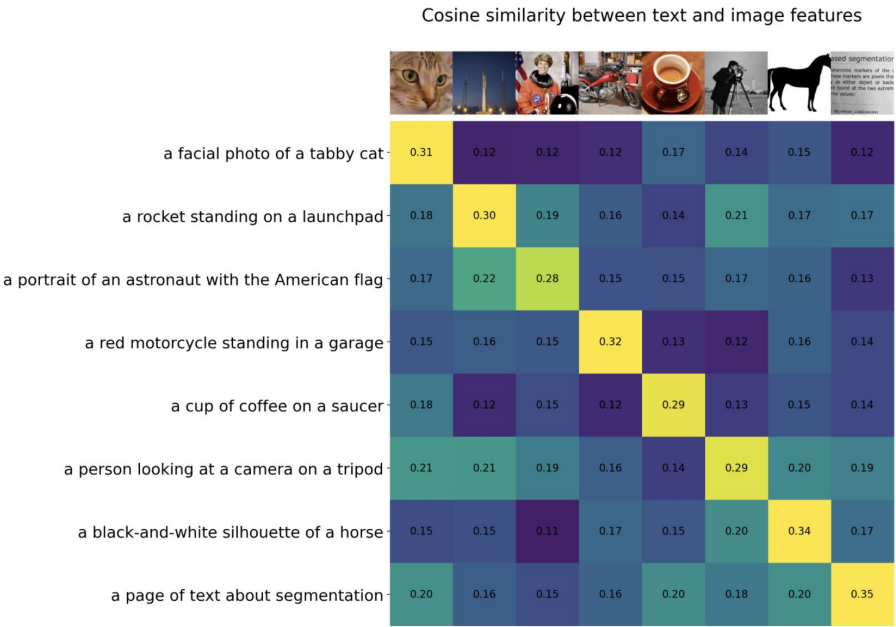
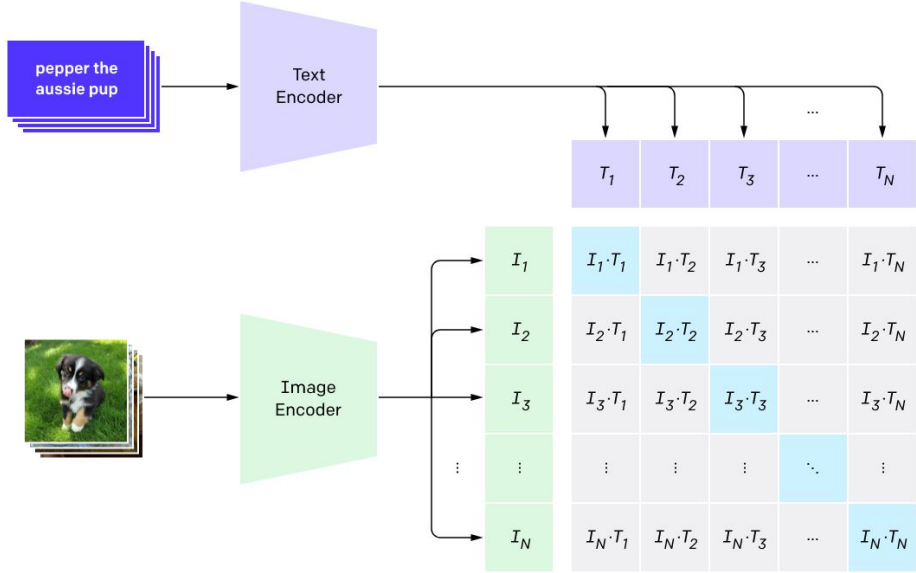
✗ a photo of **hummus**, a type of food.

# Метод



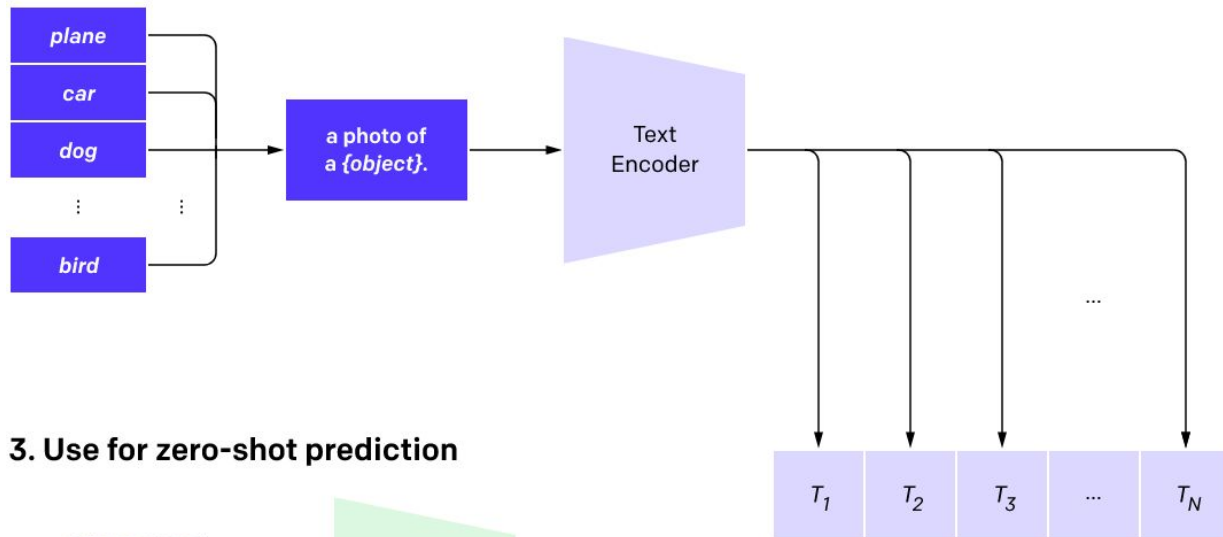
- Создание большого набора данных (400млн.пар)
- предобучение и внесение изменений
- обучение без обучения

1. Contrastive pre-training

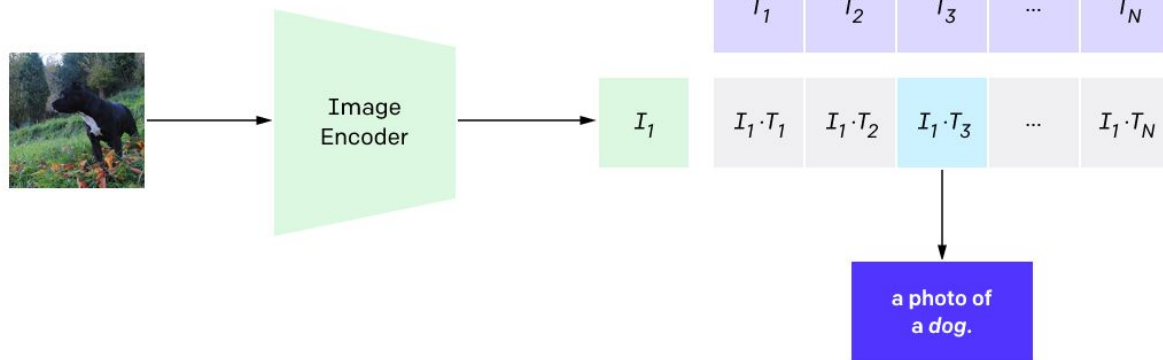


Матрица поэлементных cosine similarity между парами векторных репрезентаций изображений и текстовых описаний

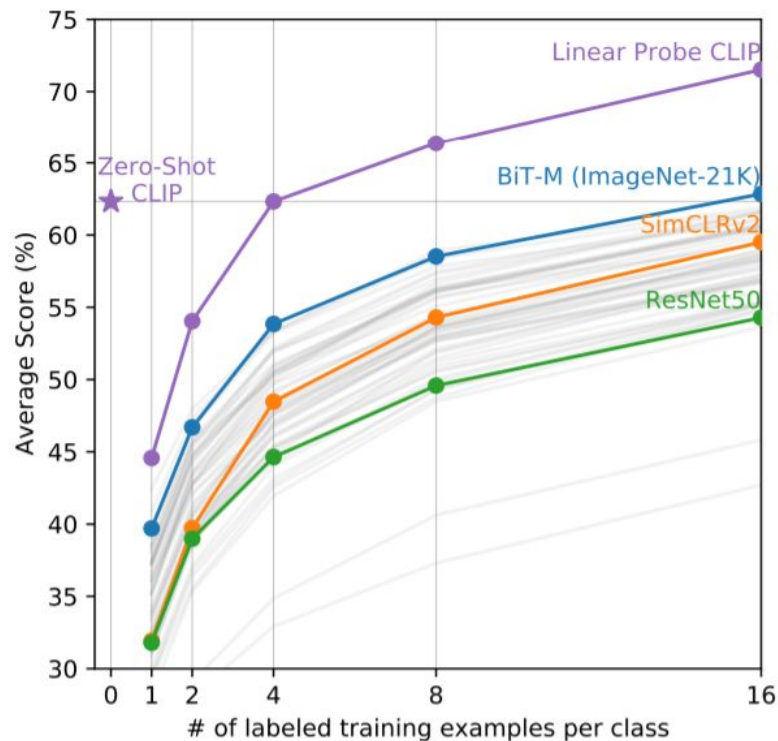
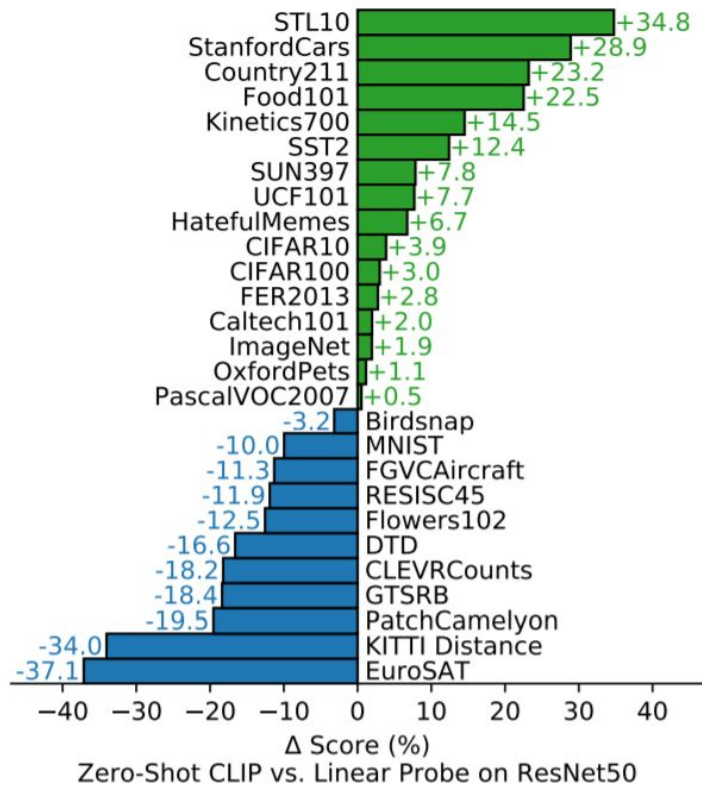
## 2. Create dataset classifier from label text



## 3. Use for zero-shot prediction



# Эксперименты




Zero-shot CLIP vs few-shot linear probes

# Ограничения CLIP


- типографические атаки
- zero-shot иногда
- data-неэффективность не решается,
- социальные предвзятости

NO LABEL



Granny Smith	85.61%
iPod	0.42%
library	0%
pizza	0%
rifle	0%
toaster	0%

LABELED "IPOD"



Granny Smith	0.13%
iPod	99.68%
library	0%
pizza	0%
rifle	0%
toaster	0%

# Ключевые выводы

- высокоэффективный
- обобщаемый и гибкий





# ИСТОЧНИКИ

<https://arxiv.org/abs/2103.00020>

<https://openai.com/blog/clip/>

<https://distill.pub/2021/multimodal-neurons/>

<https://habr.com/ru/post/539312/>