CLIP : Connecting Text and Images

Цыганов Артем, БПМИ-181

Мотивация

- Современные методы предсказывают только фиксированные категории
- Если надо усовершенствовать, прибегают к использованию, например, linear probe
- Но это не всегда удобно

- Хотим научиться получать информацию об изображениях на основе их описания
- А затем применять zero-shot transfer

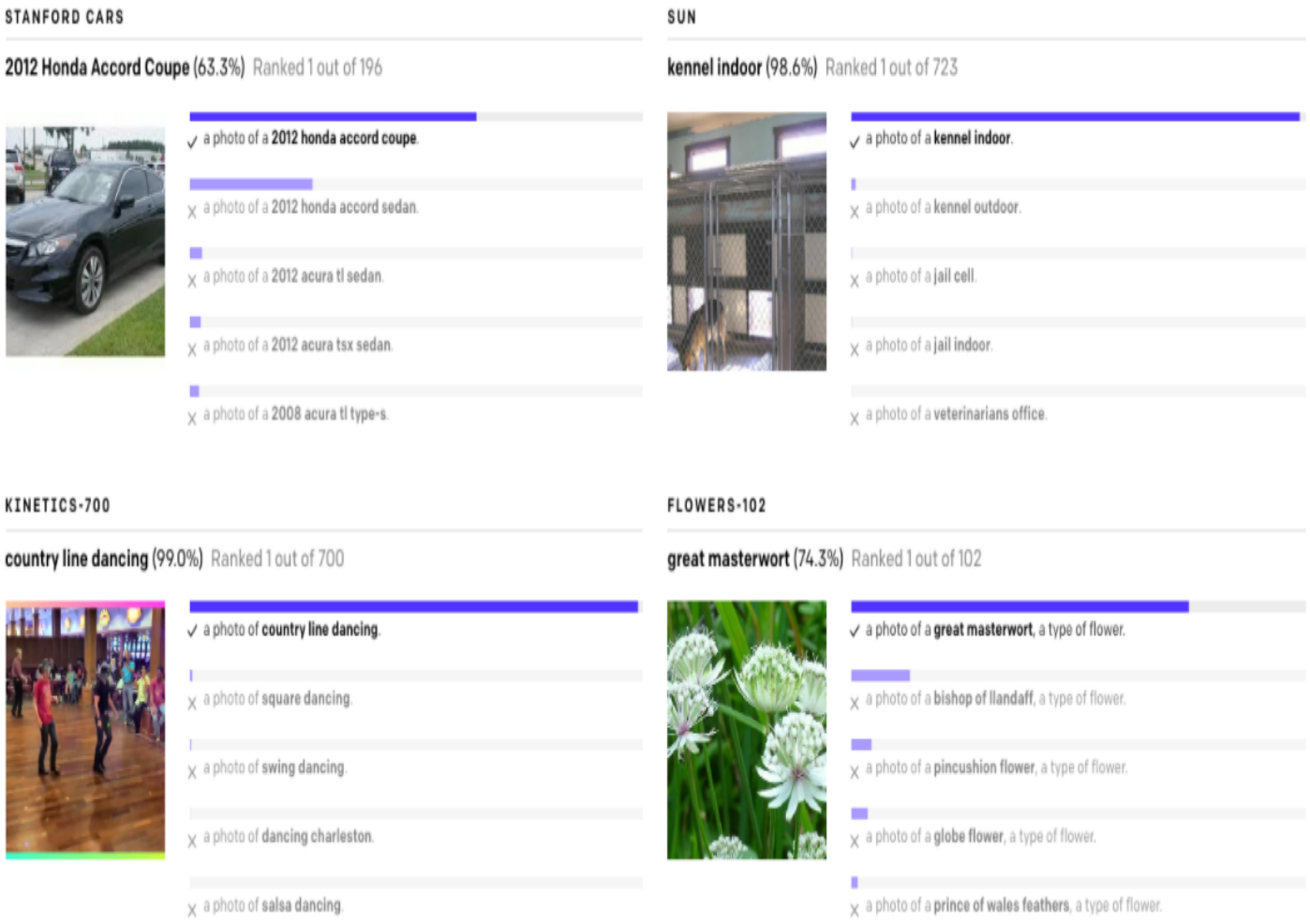
Natural language supervision

- Проще масштабировать
- Связываем векторные представления для изображений с их текстовым описанием

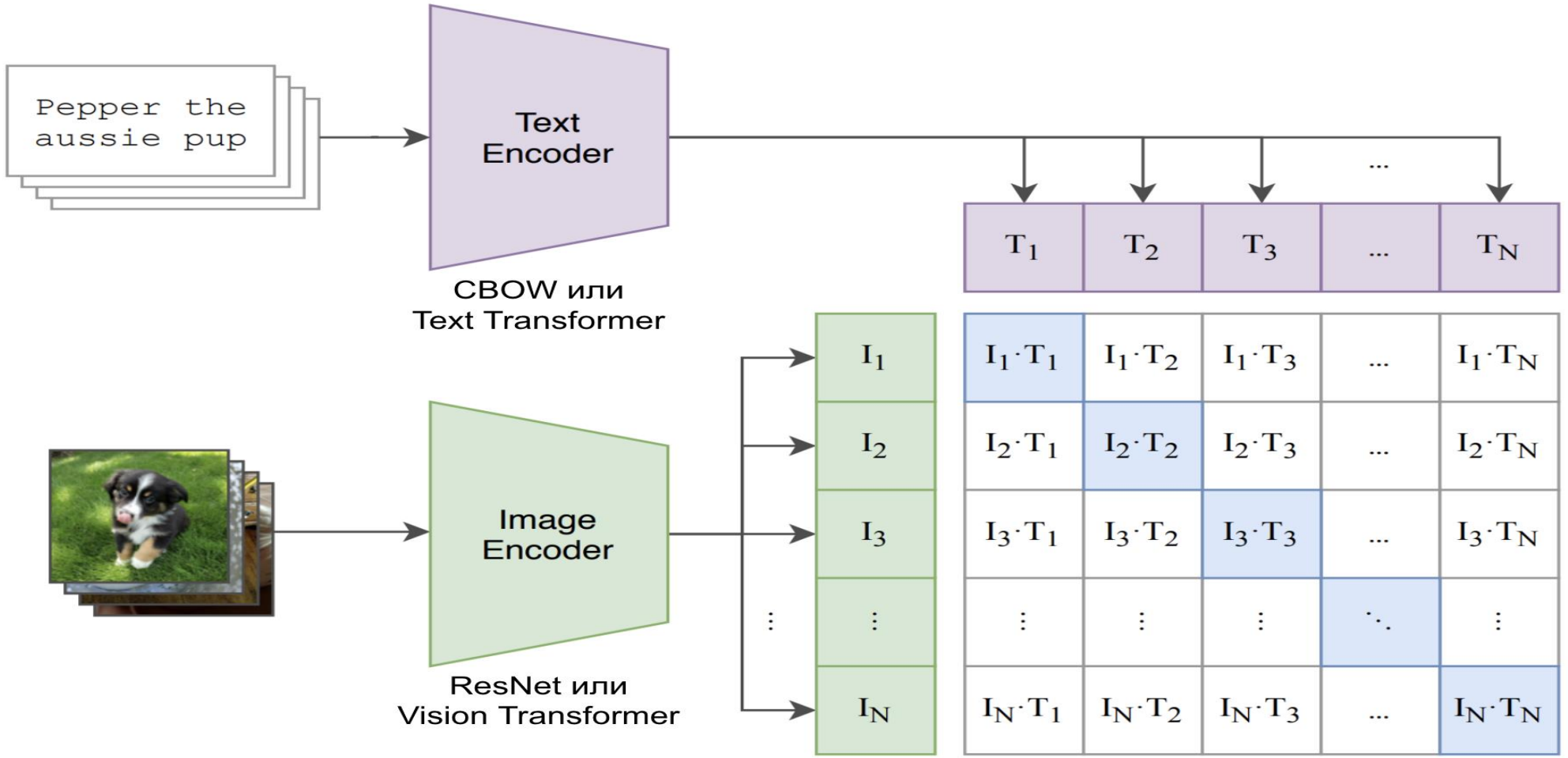
Новый набор данных

Раньше:

- MS-COCO (100 000)
- Visual Genome (100 000)
- YFCC100M (100M)



Архитектура



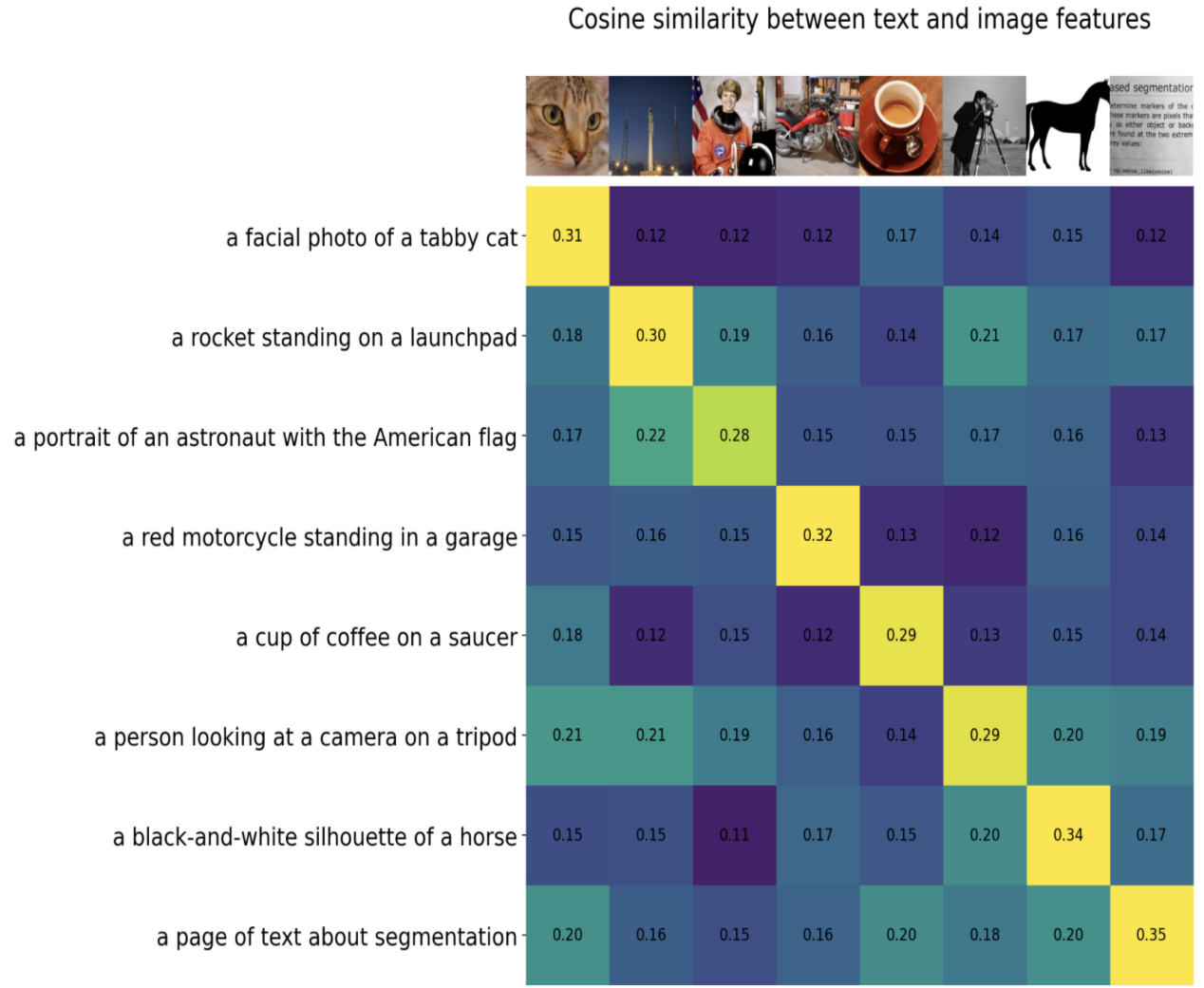
Предобучение

- Основной критерий – временные и вычислительные затраты
- Первоначальный подход работал слишком долго

- Упрощенная задача – как весь текст в целом подходит в качестве описания изображения

Предобучение

- Батч размера N
- Encoders
- Согласование размеров
- Матрица cosine similarity
- Cross-entrapy loss

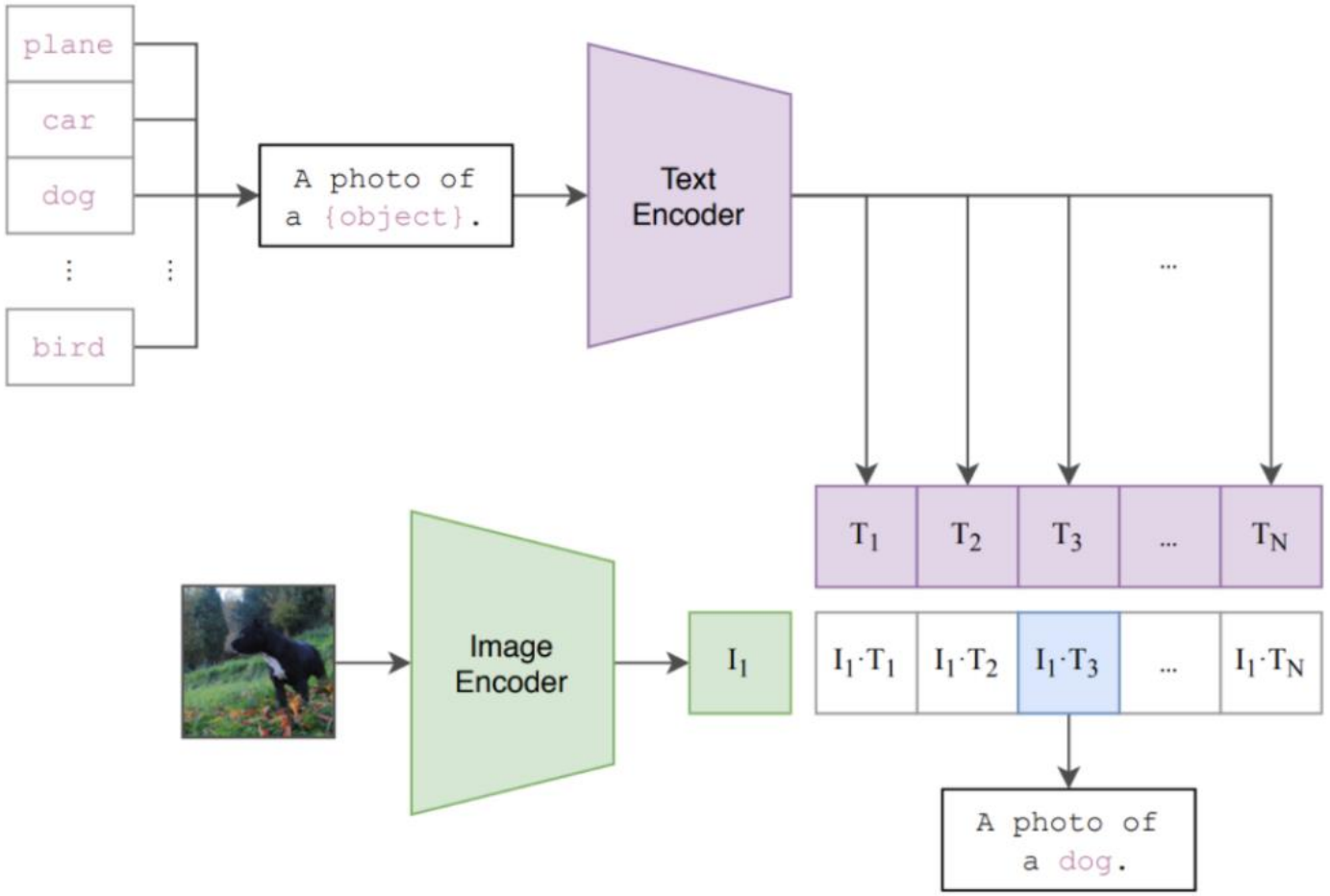


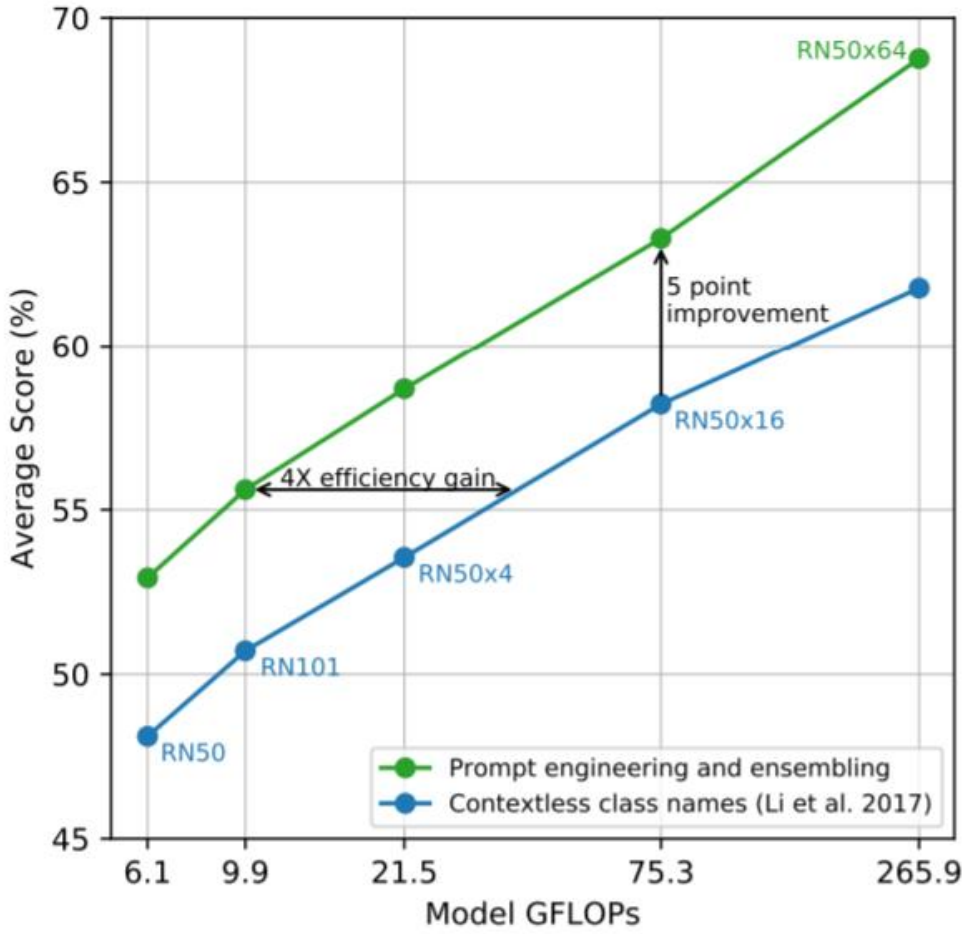
Предобучение

```
# image_encoder - ResNet or Vision Transformer
# text_encoder - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l] - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T) #[n, d_t]
```

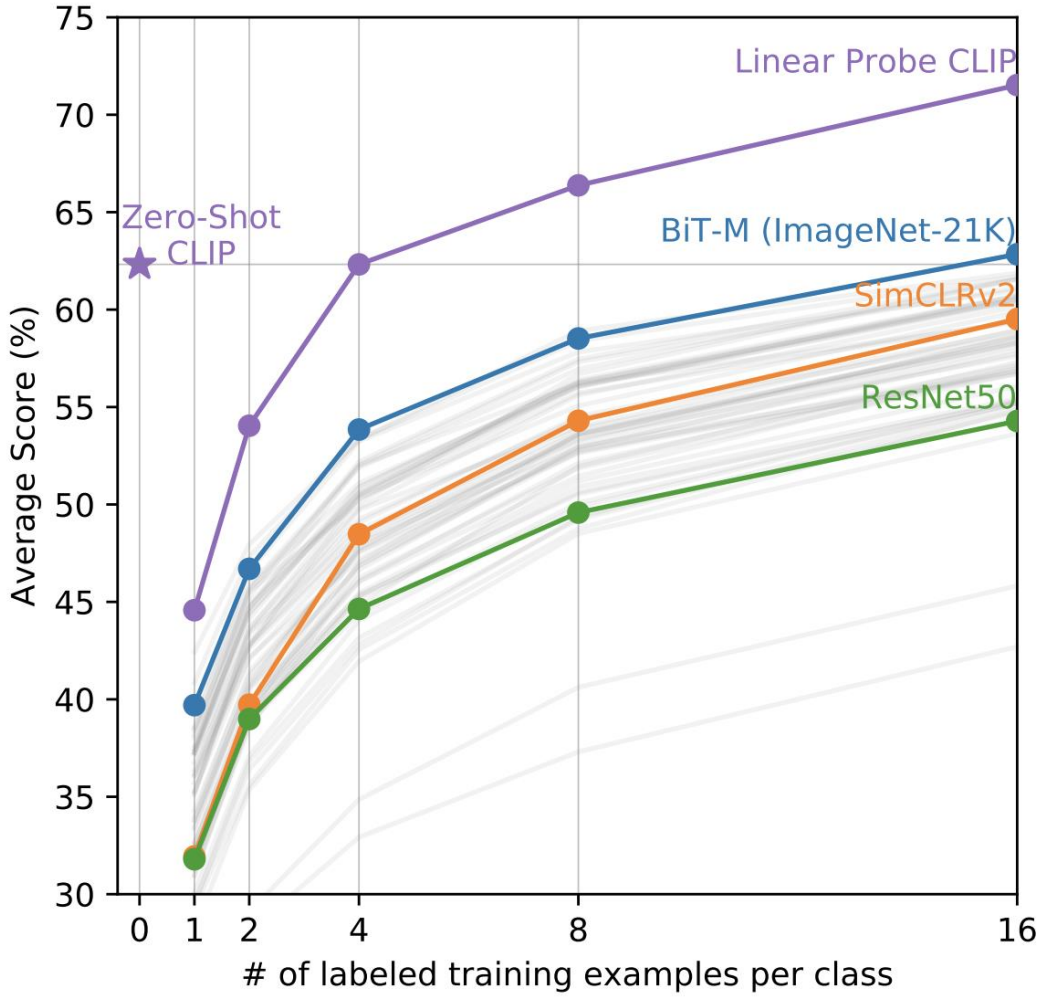
```
# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss = (loss_i + loss_t)/2
```

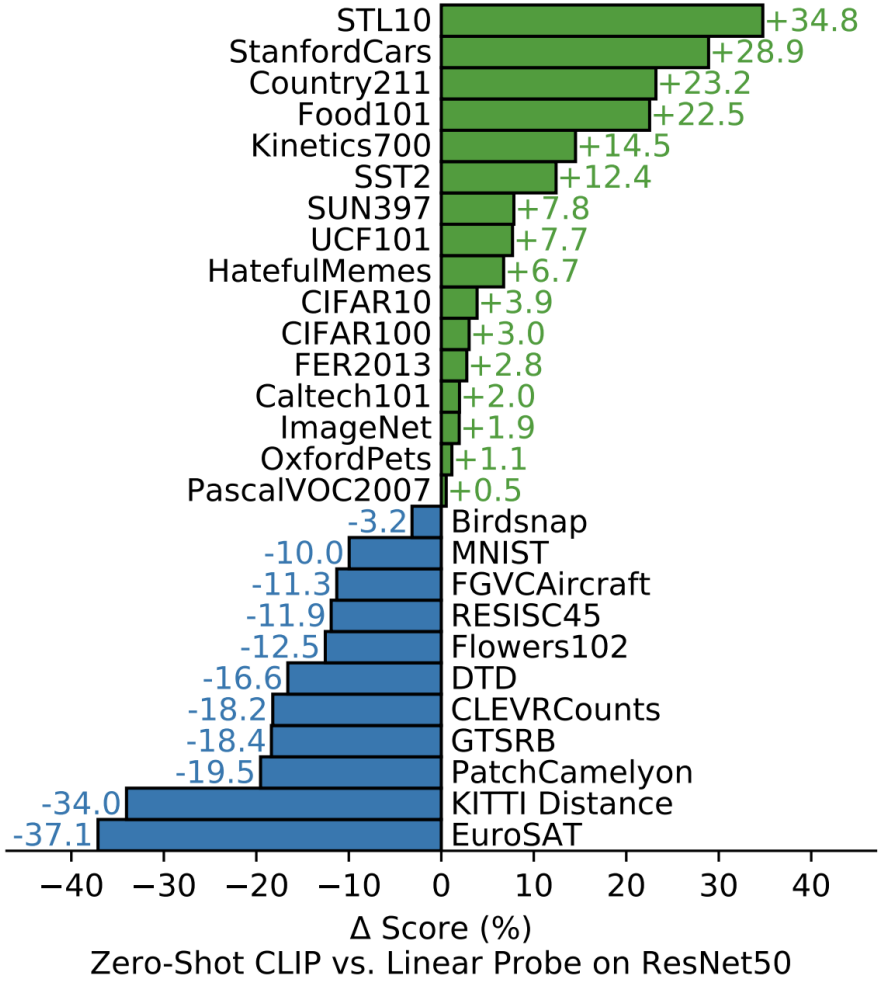
Transfer learning



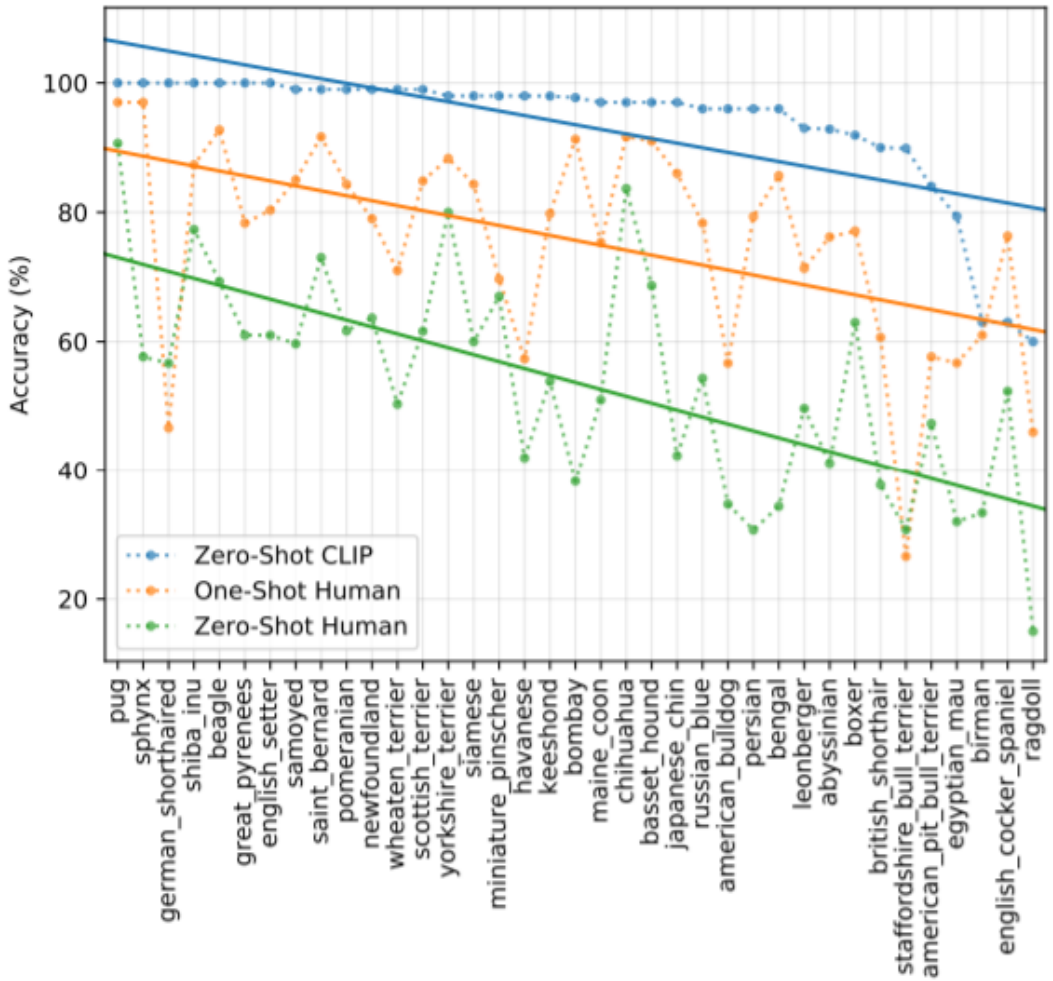


Результаты


Linear Probe CLIP
Zero-Shot CLIP
BiT-M (ImageNet-21K)
SimCLRv2
ResNet50
Average Score (%)
# of labeled training examples per class


STL10 +34.8
StanfordCars +28.9
Country211 +23.2
Food101 +22.5
Kinetics700 +14.5
SST2 +12.4
SUN397 +7.8
UCF101 +7.7
HatefulMemes +6.7
CIFAR10 +3.9
CIFAR100 +3.0
FER2013 +2.8
Caltech101 +2.0
ImageNet +1.9
OxfordPets +1.1
PascalVOC2007 +0.5
-3.2 Birdsnap
-10.0 MNIST
-11.3 FGVCAircraft
-11.9 RESISC45
-12.5 Flowers102
-16.6 DTD
-18.2 CLEVRCounts
-18.4 GTSRB
-19.5 PatchCamelyon
-34.0 KITTI Distance
-37.1 EuroSAT
Δ Score (%)
Zero-Shot CLIP vs. Linear Probe on ResNet50

Недостатки


Accuracy (%)
100
80
60
40
20
Zero-Shot CLIP
One-Shot Human
Zero-Shot Human
pug
sphynx
german_shorthaired
shiba_inu
beagle
great_pyrenees
english_setter
samoyed
saint_bernard
pomeranian
newfoundland
wheaten_terrier
scottish_terrier
yorkshire_terrier
siamese
miniature_pinscher
havanese
keeshond
bombay
maine_coon
chihuahua
basset_hound
japanese_chin
russian_blue
american_bulldog
persian
bengal
leonberger
abyssinian
boxer
british_shorthair
staffordshire_bull_terrier
american_pit_bull_terrier
egyptian_mau
birman
english_cocker_spaniel
ragdoll


Granny Smith 85.6%
iPod 0.4%
library 0.0%
pizza 0.0%
toaster 0.0%
dough 0.1%
Granny Smith 0.1%
iPod 99.7%
library 0.0%
pizza 0.0%
toaster 0.0%
dough 0.0%
iPod

Итоги

Плюсы

- Высокая производительность
- Широкое применение
- Обобщаемость и гибкость

Минусы

- Не использует уже полученные знания при повторном обучении
- Типографические атаки
- zero-shot не применим для узконаправленных задач

Источники

- https://cdn.openai.com/papers/Learning_Transferable_Visual_Models_From_Natural_Language_Supervision.pdf
- https://habr.com/ru/post/539312/
- https://distill.pub/2021/multimodal-neurons/#person-neurons
- https://habr.com/ru/post/540312/
- https://openai.com/blog/clip/
- https://habr.com/en/post/537334/