



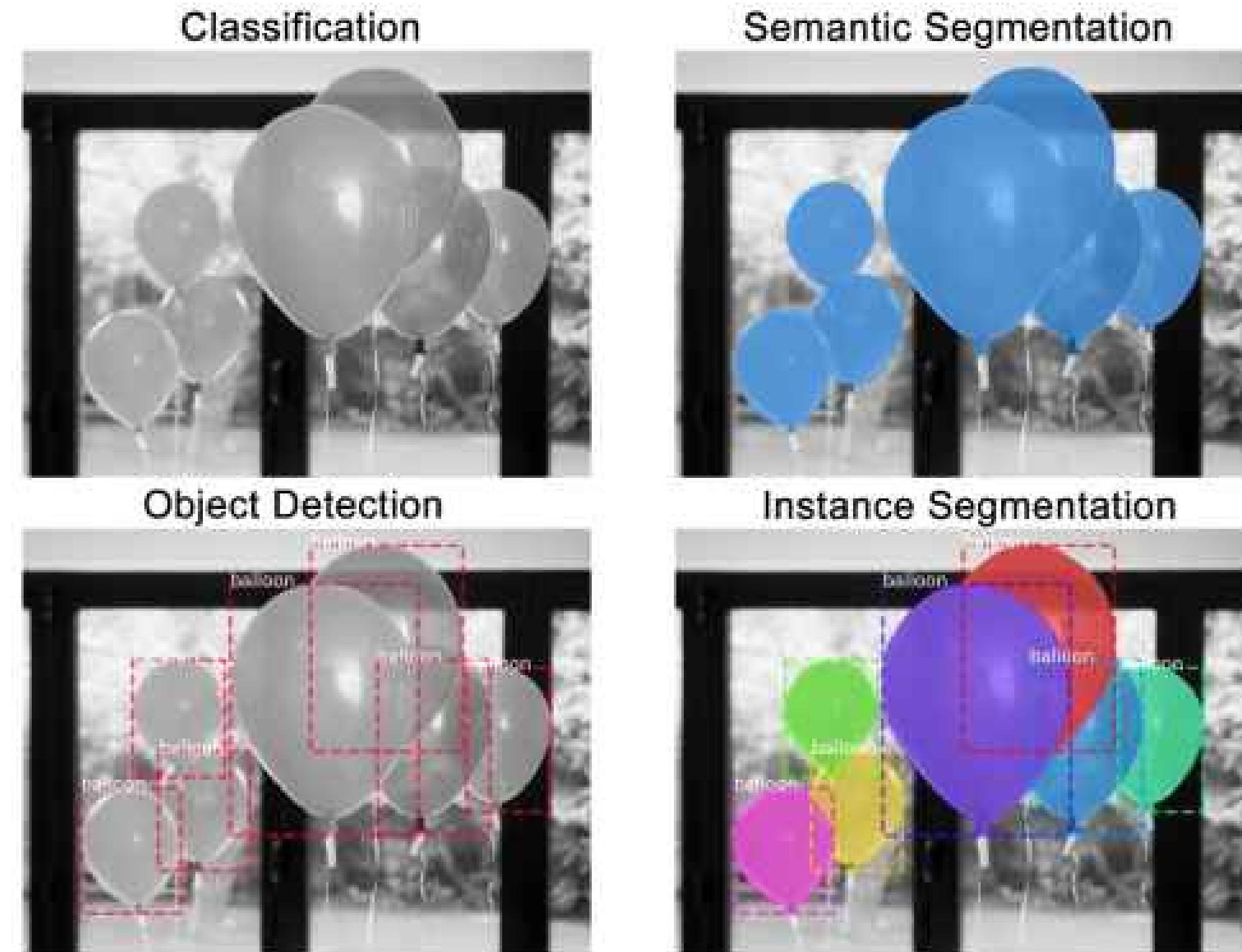
Факультет компьютерных наук

Искусственный
интеллект

Москва
2025

Контрастивное обучение CLIP/CLAP

- Классификация
- Смысловая сегментация
- Детекция объектов
- Сегментация объектов





Ограничения классического CV

3

- Ограничено число классов
- Сложно добавлять новые классы

Могут ли модели обучаться на данных без ручной разметки?

Проблема масштабируемости supervised озвучена в статье

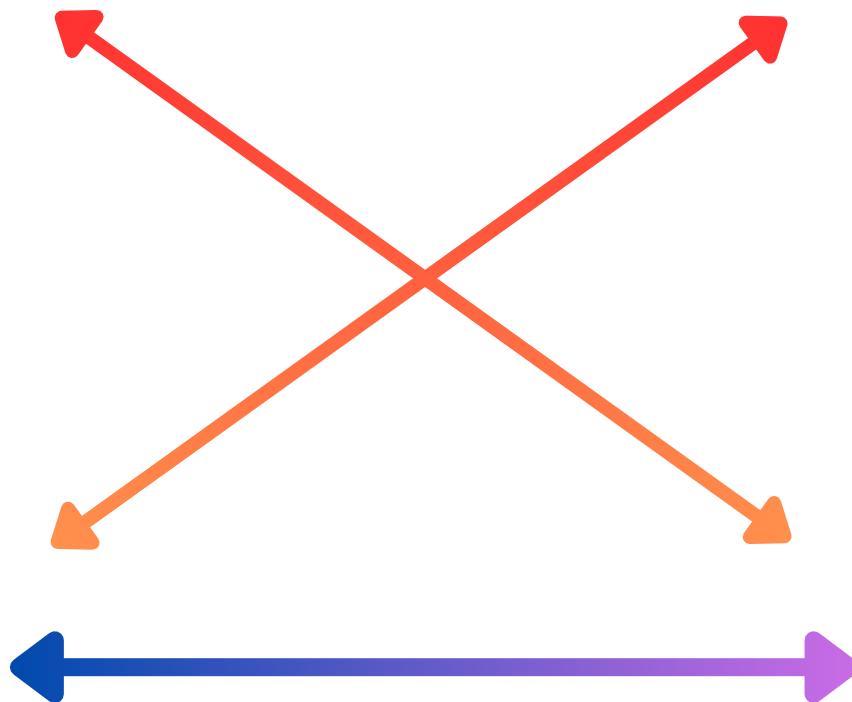
[*Learning Transferable Visual Models From Natural Language Supervision*](#)



Идея контрастивного обучения



милый котик сидит



лабrador красавчик

$$\mathcal{L}_{\text{cont}}(x_i, x_j) = 1[y_i = y_j] |f(x_i) - f(x_j)|^2 + \\ 1[y_i \neq y_j] \max(0, \varepsilon - |f(x_i) - f(x_j)|)^2$$

пара из одного класса → стягиваем точки

пара из разных классов → раздвигаем дальше порога ε

Batch-wise / InfoNCE loss / NT-Xent softmax-contrastive loss

Для батча из $2N$ аугментированных примеров (по 2 на каждое исходное изображение). Для положительной пары (i, j) :

$$\ell_{i,j} = \log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}$$

где

z_i - L2-нормированные эмбеддинги,

sim - скалярное произведение (по сути косинус),

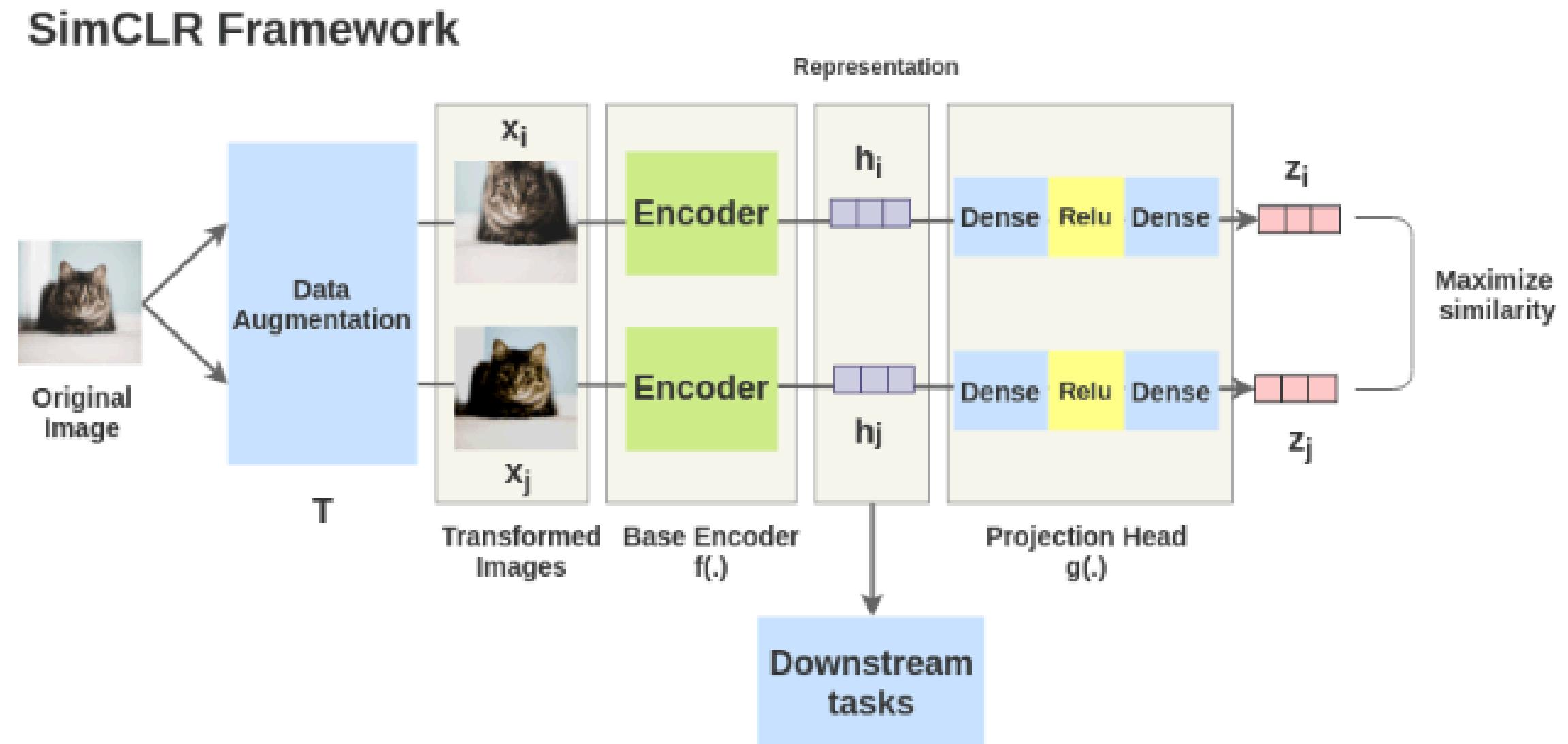
τ - температура.

$$\mathcal{L} = \frac{1}{2N} \sum_{i=1}^{2N} \sum_{j \in P(i)} \ell_{i,j}.$$



The Illustrated SimCLR Framework

7

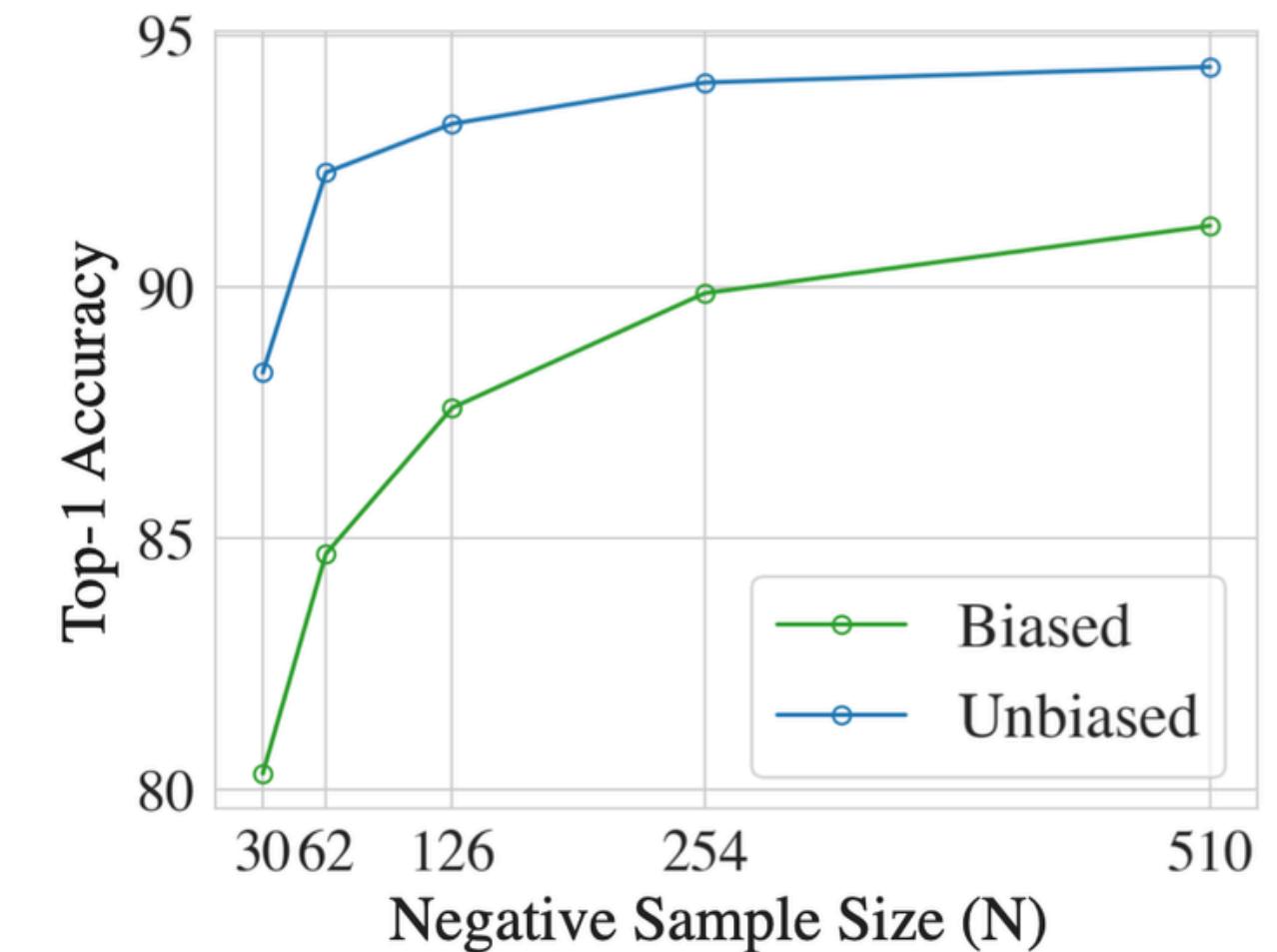
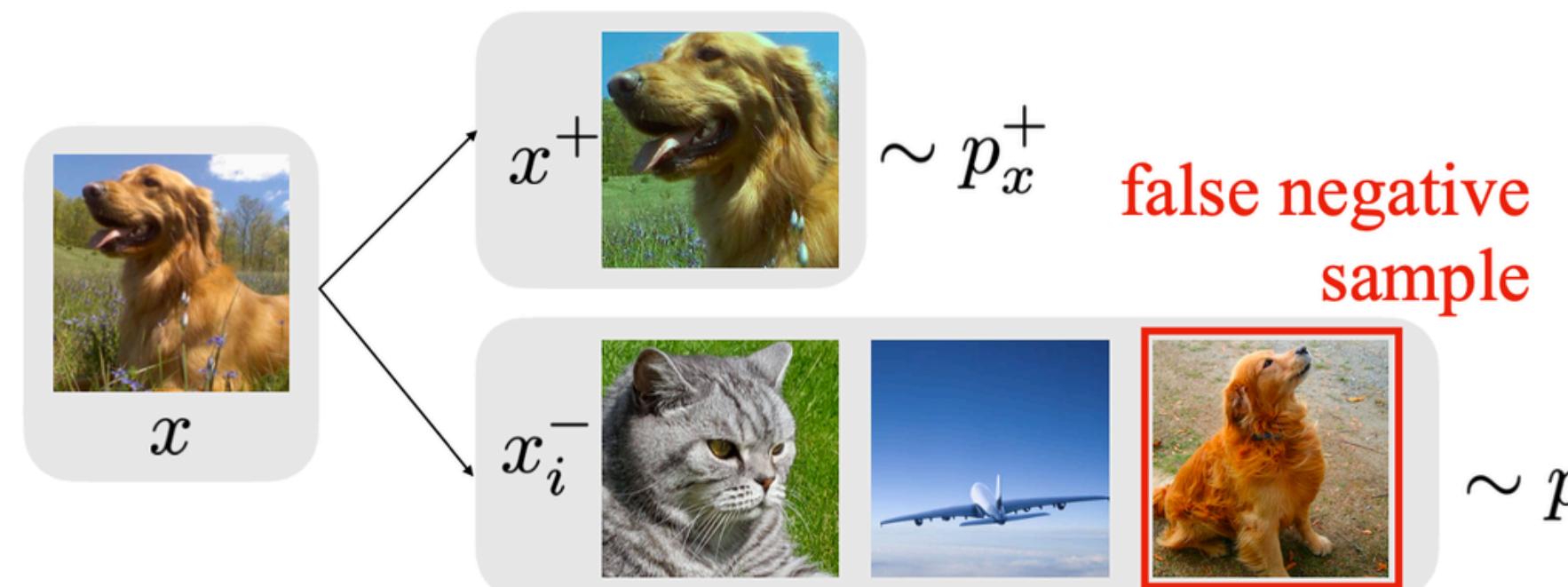


Прыгаем по ссылке



The Illustrated SimCLR Framework

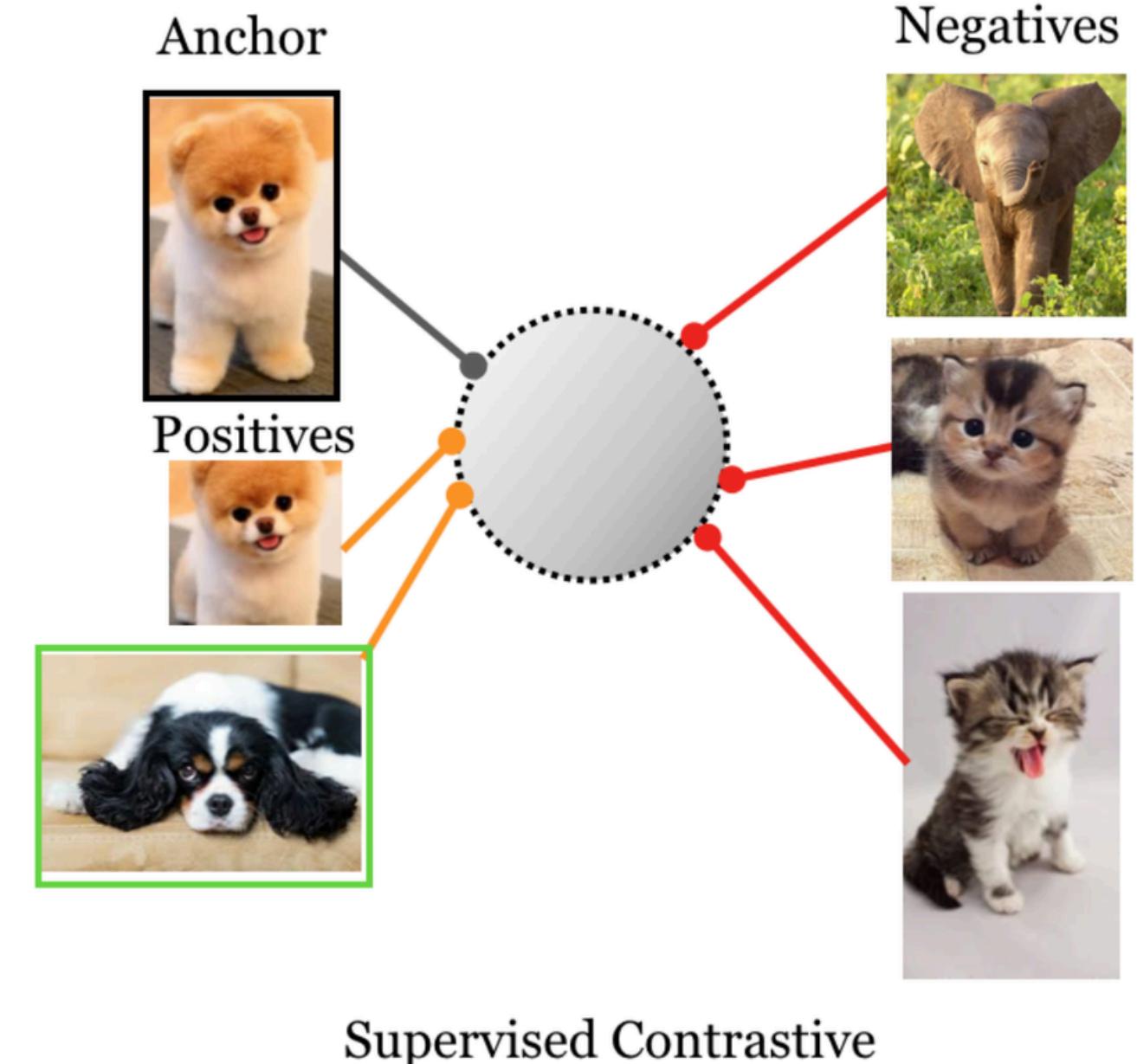
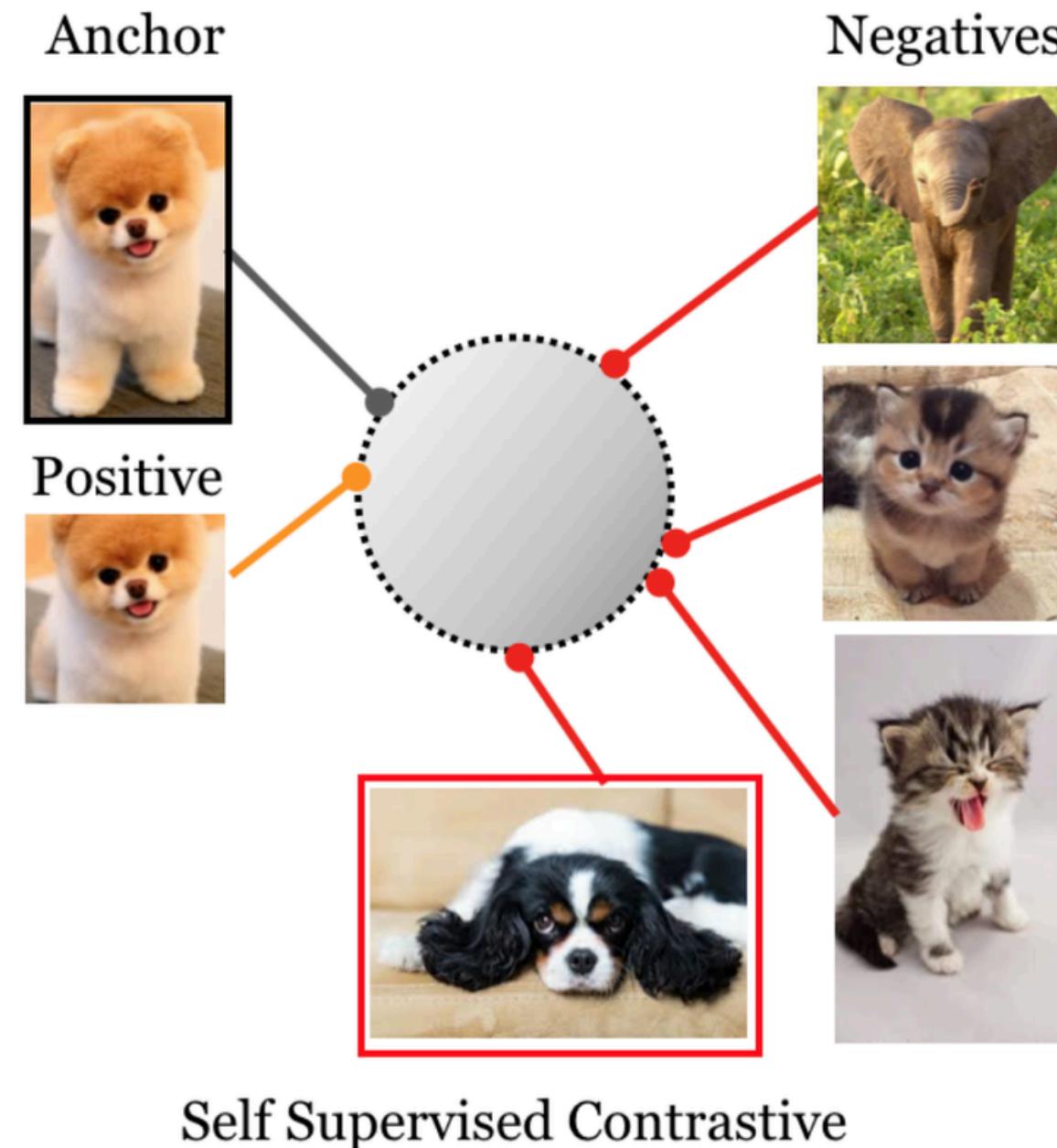
False negatives в batch-wise contrastive learning



debiased-вариант лосса вычитает вклад вероятных false negatives
(используя оценки распределения позитивов/негативов)

Debiased Contrastive Learning

Supervised contrastive learning (SupCon)



Supervised Contrastive Learning

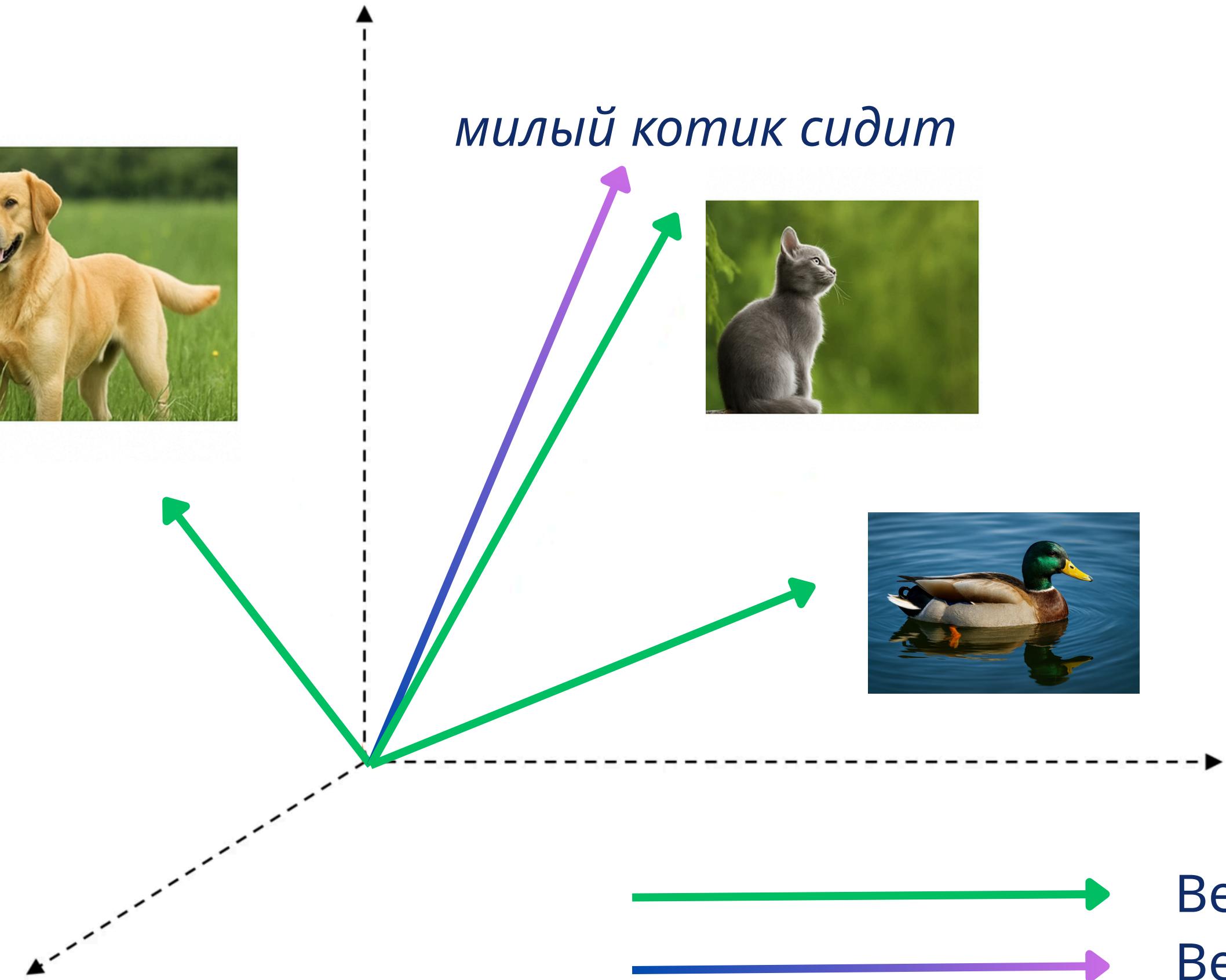


Картинки ↔ текст

10



милый котик сидит

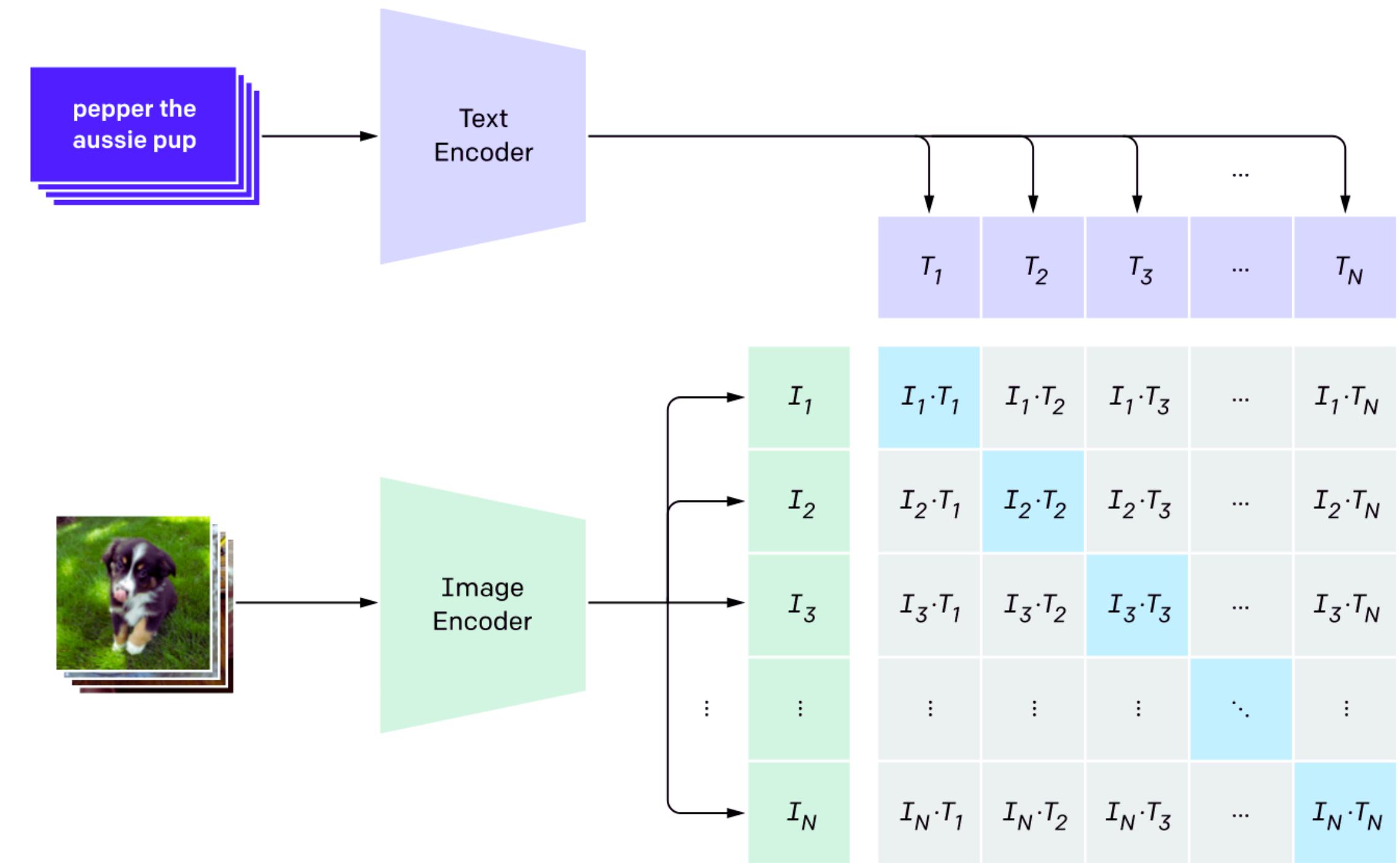


Вектор картинки
Вектор текста



Contrastive language-image pre-training

1. Contrastive pre-training



[Источник](#)



Это фото котика	Селезень, а не утка	Милый золотой лабрадор	...	Это точно слон
			...	
1			...	
	1		...	
		1	...	
:	:	:	..	:
			...	1

$$v_i = f_{\text{img}}(I_i) \quad u_j = f_{\text{text}}(T_j)$$

$$M_{ij} = \frac{\langle v_i, u_j \rangle}{\tau}$$

$$L_{\text{img} \rightarrow \text{text}} = -\frac{1}{N} \sum_{i=1}^N \log p_{i \rightarrow i} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(M_{ii})}{\sum_j \exp(M_{ij})}$$

$$L_{\text{text} \rightarrow \text{img}} = -\frac{1}{N} \sum_{j=1}^N \log p_{j \rightarrow j} = -\frac{1}{N} \sum_{j=1}^N \log \frac{\exp(M_{jj})}{\sum_i \exp(M_{ij})}$$

$$L_{\text{CLIP}} = \frac{1}{2} (L_{\text{img} \rightarrow \text{text}} + L_{\text{text} \rightarrow \text{img}})$$



Алгоритм CLIP

```
# 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)

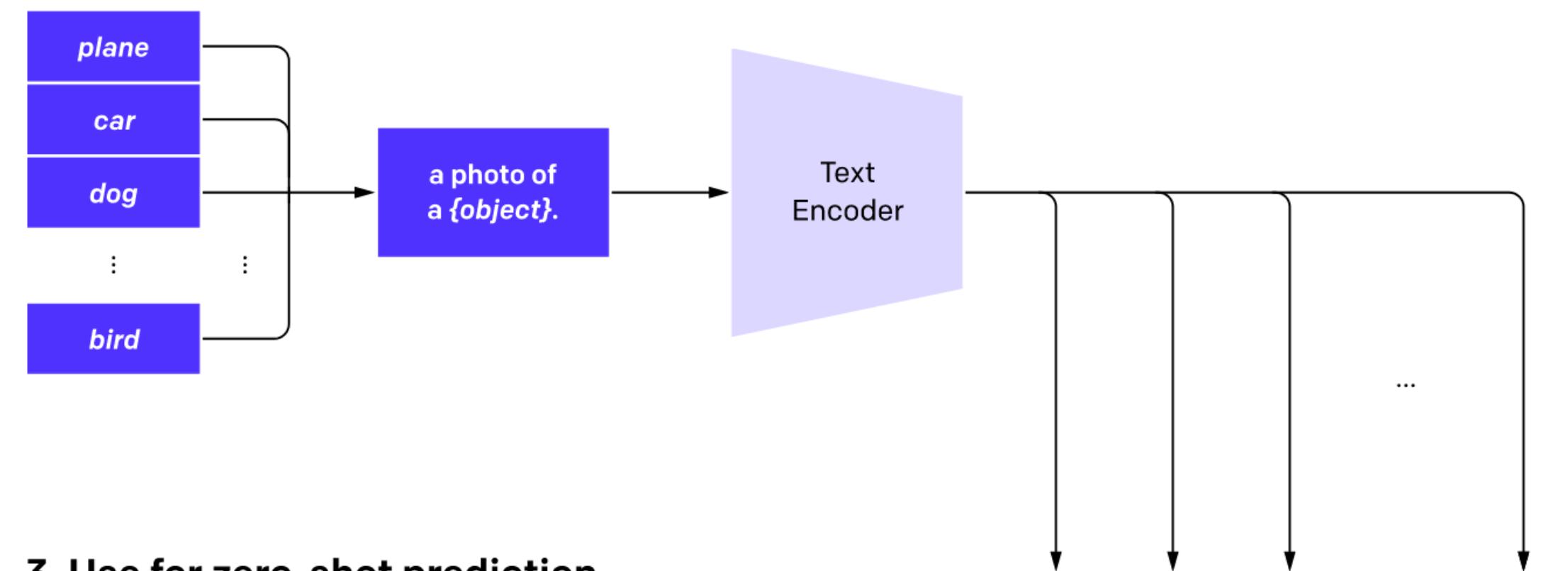
# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Figure 3. Numpy-like pseudocode for the core of an implementation of CLIP.

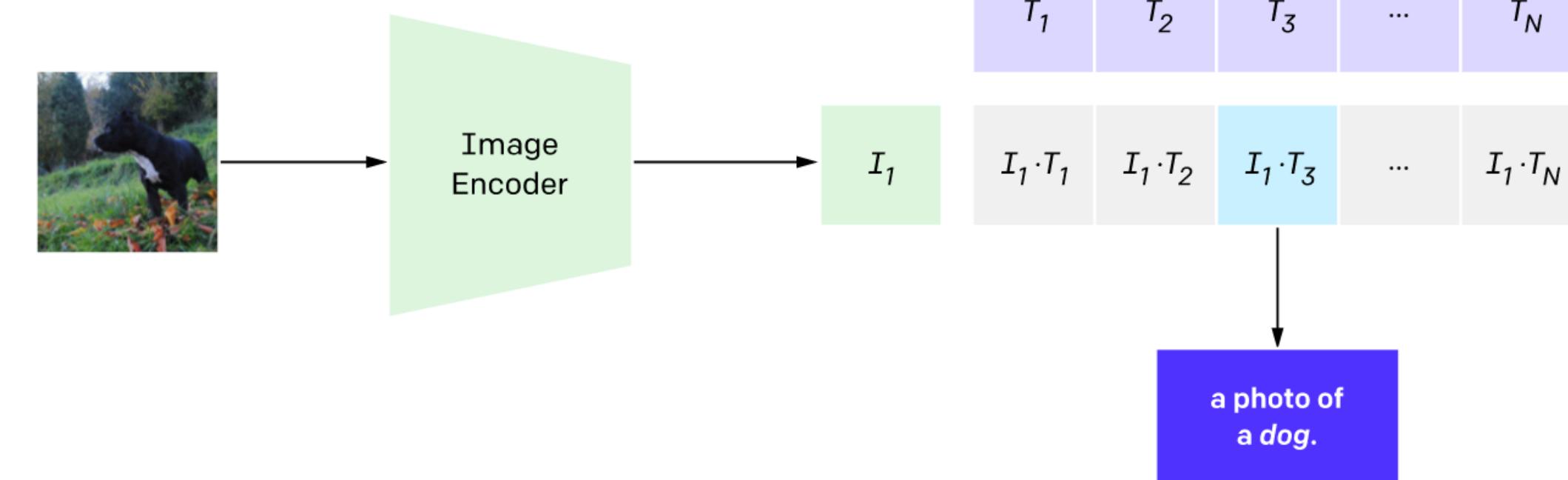


CLIP. Архитектура

2. Create dataset classifier from label text



3. Use for zero-shot prediction





Спецификация обучения

- Датасеты:

MS-COCO (100k),
Visual Genome (100k),
YFCC100M (100M),

Проприетарные сети (400M)

- Бекбоны:

Image ResNet/ViT
Text transformer

- Батч сайз 32,768

- Время обучения:

18 дней 592 V100 ResNet,
12 дней 256 V100 ViT

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.

SUN397

television studio (90.2%) Ranked 1 out of 397



- ✓ a photo of a **television studio**.
- ✗ a photo of a **podium indoor**.
- ✗ a photo of a **conference room**.
- ✗ a photo of a **lecture room**.
- ✗ a photo of a **control room**.

YOUTUBE-BB

airplane, person (89.0%) Ranked 1 out of 23



- ✓ a photo of a **airplane**.
- ✗ a photo of a **bird**.
- ✗ a photo of a **bear**.
- ✗ a photo of a **giraffe**.
- ✗ a photo of a **car**.

EUROSAT

annual crop land (12.9%) Ranked 4 out of 10

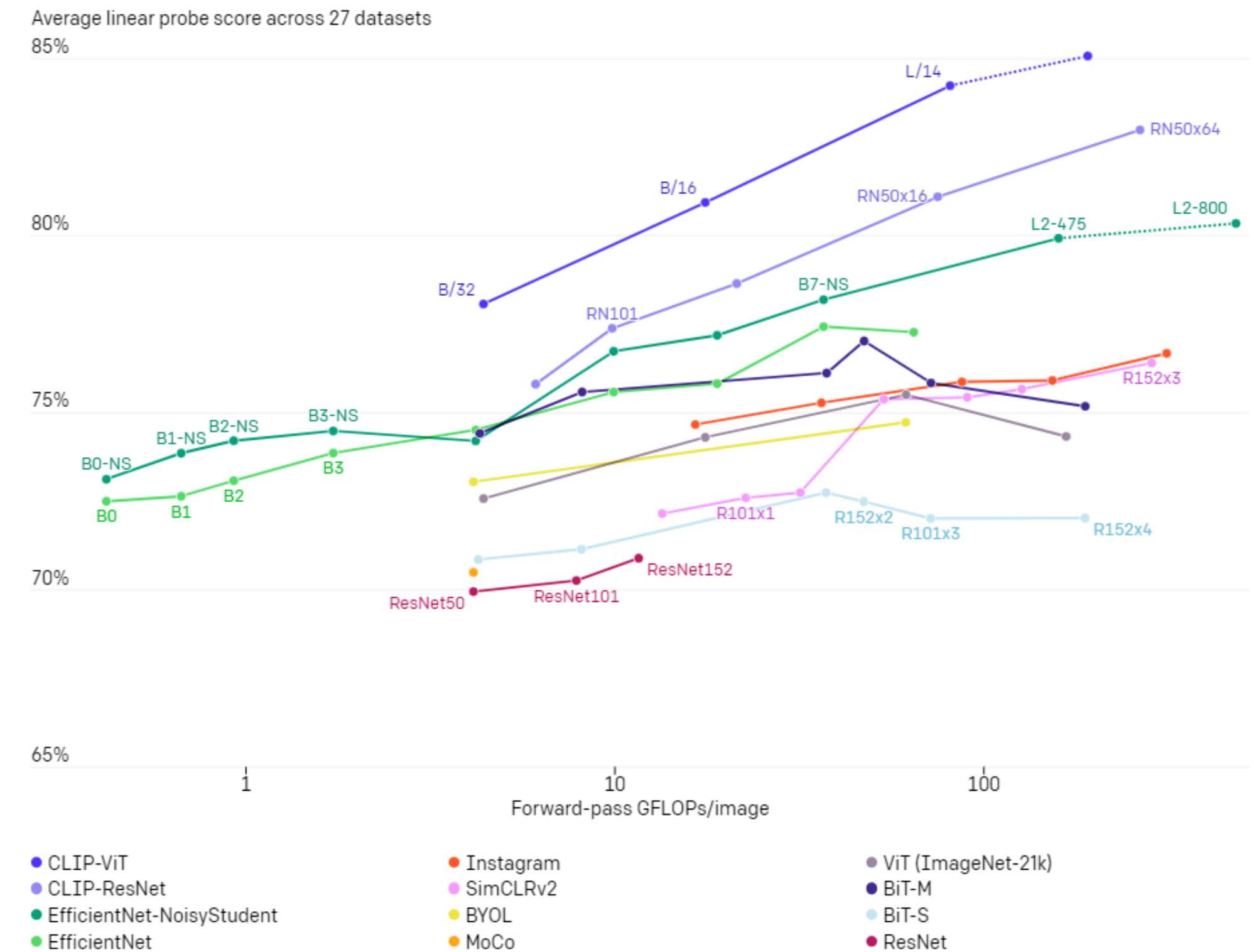


- ✗ a centered satellite photo of **permanent crop land**.
- ✗ a centered satellite photo of **pasture land**.
- ✗ a centered satellite photo of **highway or road**.
- ✓ a centered satellite photo of **annual crop land**.
- ✗ a centered satellite photo of **brushland or shrubland**.



Производительность

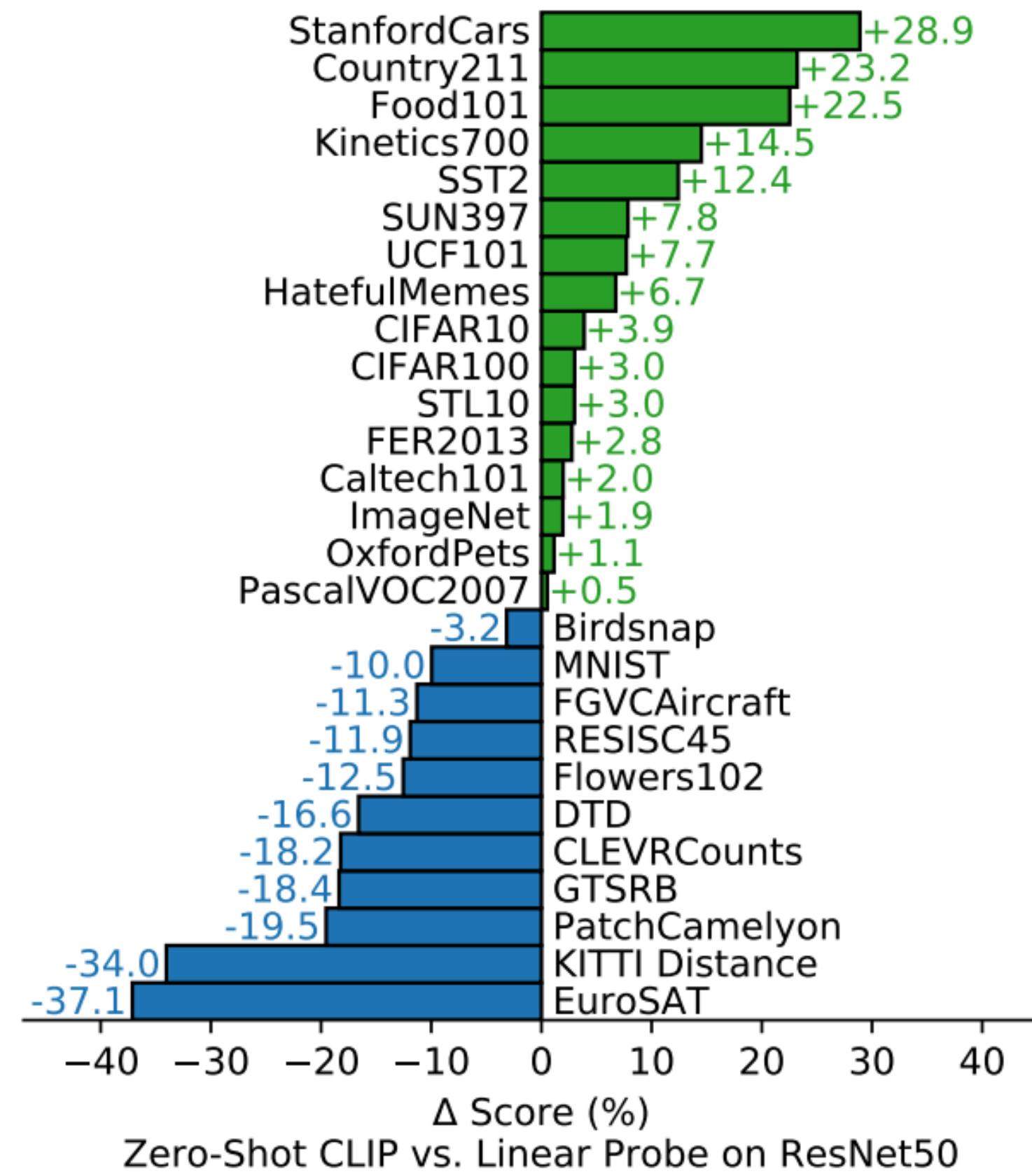
17



Across a suite of 27 datasets measuring tasks such as fine-grained object classification, OCR, activity recognition in videos, and geo-localization, we find that CLIP models learn more widely useful image representations. CLIP models are also more compute efficient than the models from 10 prior approaches that we compare with.



Производительность



EuroSAT. Классификация спутниковых снимков (типы землепользования).

KITTI Distance. Оценка расстояния до ближайшей машины по виду с машины (self-driving).

PatchCamelyon. Гистопатология, опухоль/не опухоль на микроснимках лимфоузлов.

GTSRB. Немецкие дорожные знаки.

CLEVRCounts. Синтетические 3D-сцены, задача — счёт объектов.

DTD - Describable Textures Dataset. Классификация текстур по прилагательным («dotted», «braided», «veined» и т.п.).

Flowers102. Тонко-детализированная классификация видов цветов.

RESISC45. Ещё один датасет спутниковых сцен (45 классов).

FGVC Aircraft. Тонкая классификация типов самолётов.

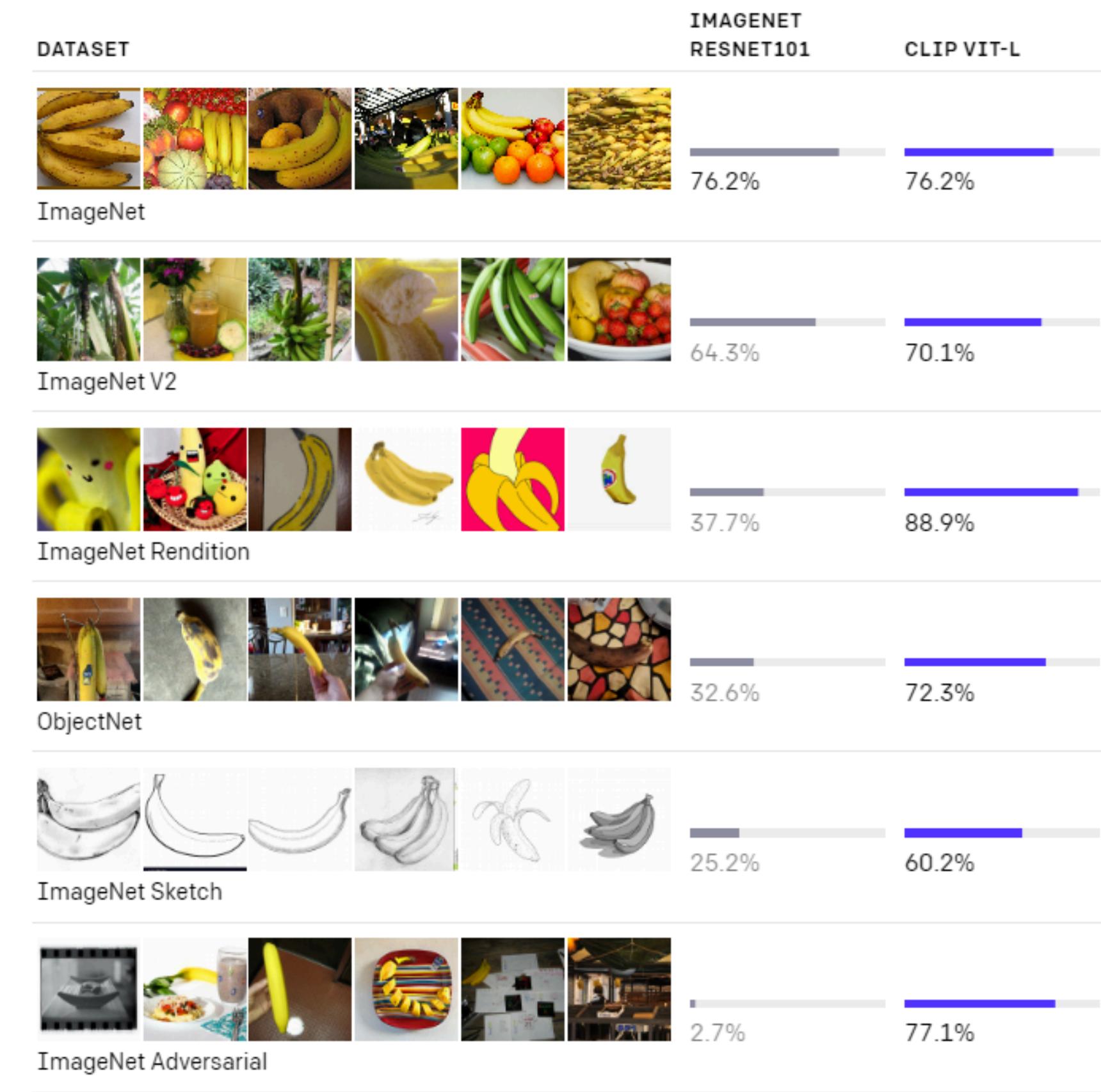
MNIST. Рукописные цифры 28×28.

Birdsnap. Тонкая классификация птиц.



Производительность

19





Бабушка всех моделей

19

LLaVA

использует буквально OpenAI CLIP (ViT-L/14-336) как vision encoder

Stable Diffusion

использует CLIP / OpenCLIP как text encoder

Qwen-VL

использует свой ViT-энкодер, предобученный по CLIP-подобному
контрастивному рецепту, а не конкретный чекпойнт OpenAI



Внуки обгоняют

OpenCLIP / LAION CLIP

такой же CLIP, но посильнее

EVA-CLIP (BAAI)

дорогие по инференсу SOTA-CLIP-подобные модели

SigLIP (Google) / SigLIP 2 (2025)

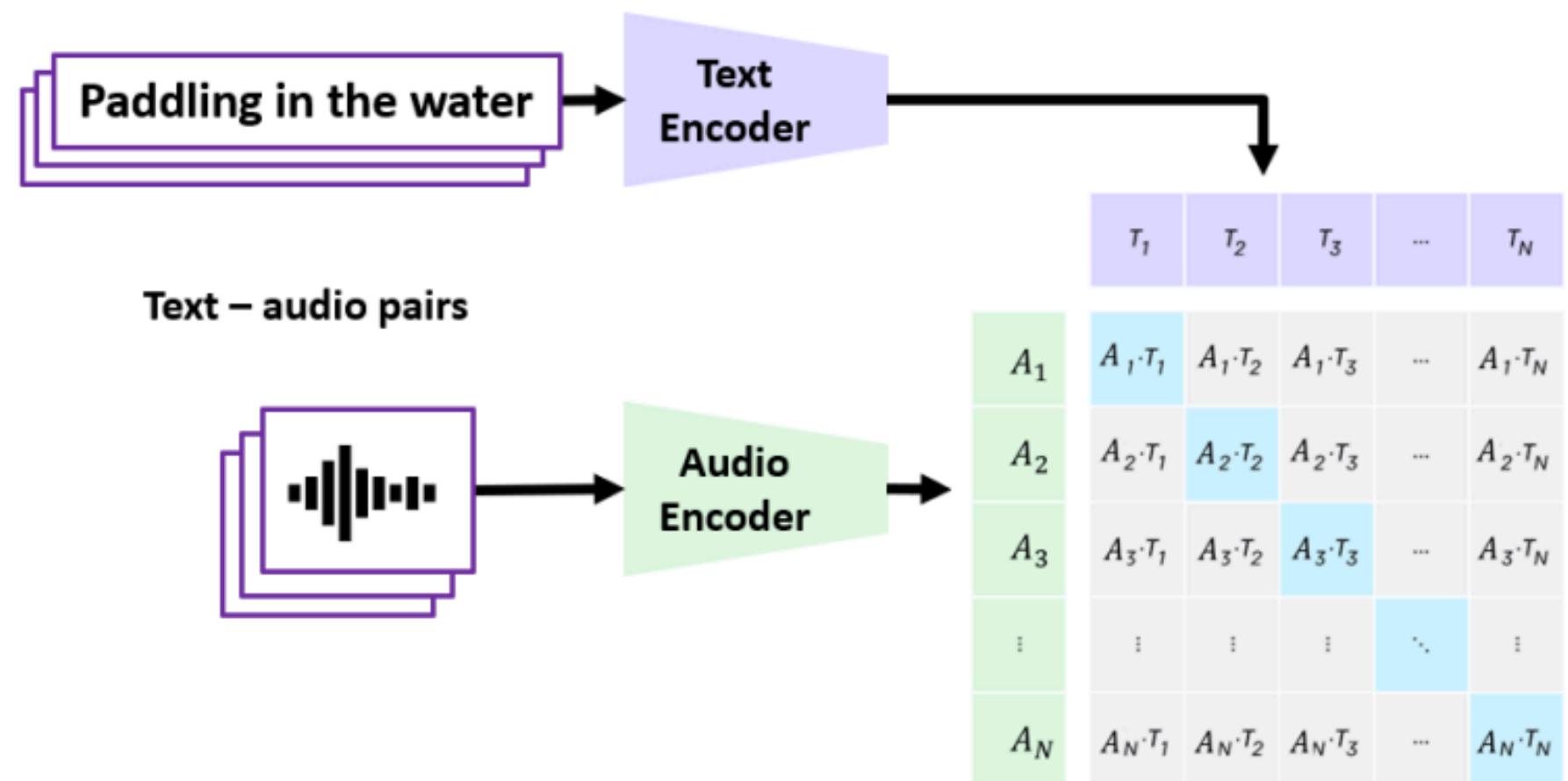
вместо softmax-контрастивки используют sigmoid-based loss по парам

CLIP-Rocket (Meta)

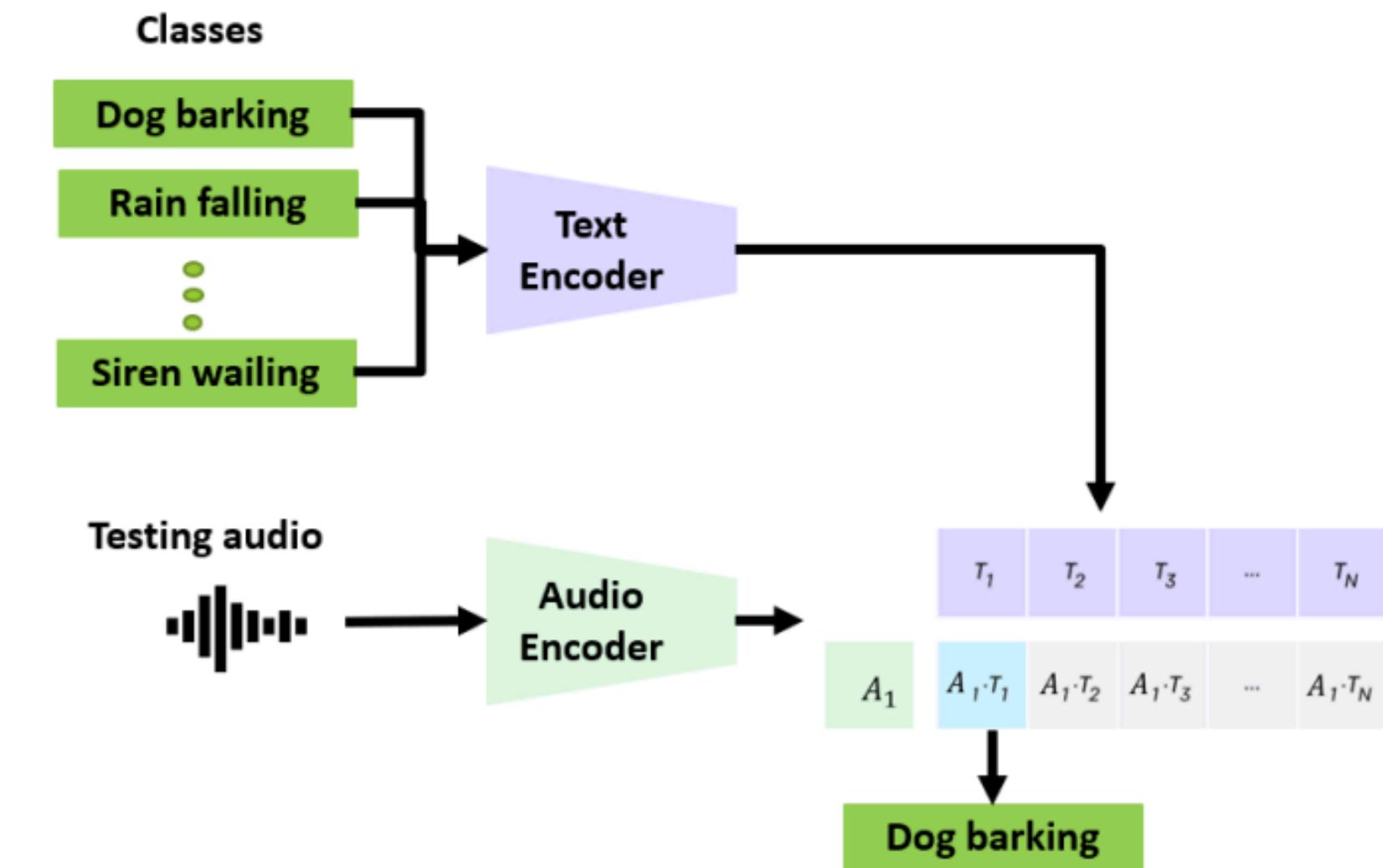
усилили training recipe для CLIP - показывают +25% относительного

прироста на downstream-тасках

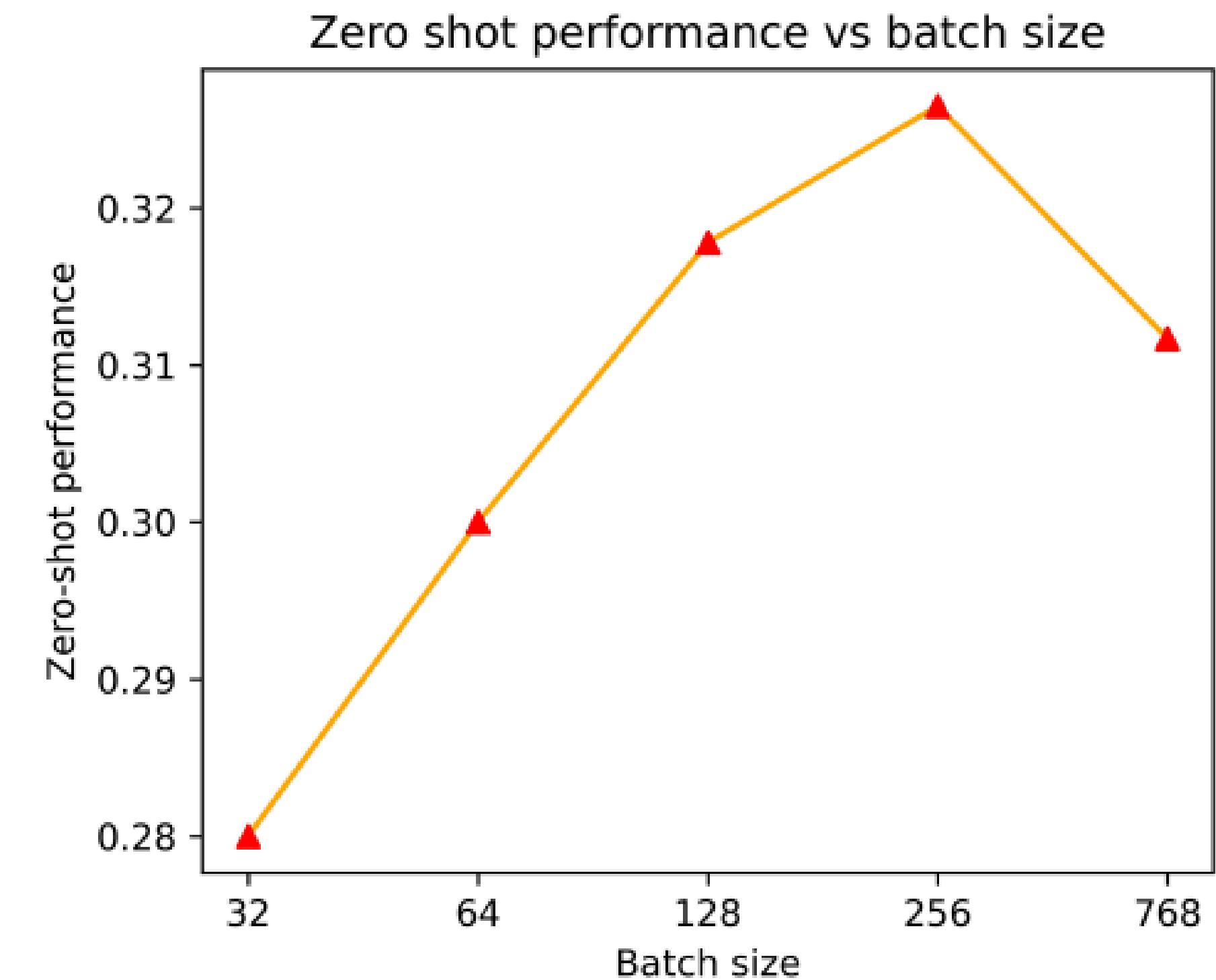
1. Contrastive Pretraining



2. Use pretrained encoders for zero-shot prediction in a new dataset or task



Dataset	Pairs	Unique audios	Unique captions
FSD50k	36,796	36,796	36,796
ClothoV2	29,646	5,929	29,646
AudioCaps	44,292	44,292	44,292
MACS	17,276	3,930	17,276
	128,010	90,947	128,010





- Бекбоны: Audio CNN14. Text BERT.
- Батч сайз 32 - 768
- Время обучения 40 эпох 8-24 V100

Prompt	ESC50 (acc)
<i>'i can hear [class label]'</i>	0.786
<i>'this is an audio of [class label]'</i>	0.8005
<i>'[class label]'</i>	0.812
<i>'this is [class label]'</i>	0.8135
<i>'this is a sound of [class label]'</i>	0.826

Audio encoder (frozen)	Text encoder (frozen)	Avg. ZS score	ESC50 (acc)
✓	✓	0.2809	0.5555
✗	✓	0.2818	0.6415
✓	✗	0.3109	0.7631
✗	✗	0.3265	0.826

