

CLIP: Connecting text and images

Дегтярев Роман
20.02.2024

План

- Определение CLIP
- Мотивация
- Архитектура модели
- Как выглядит предобучение и его детали
- Результат: концепция Zero-shot learning
- Эксперименты
- Выводы







Определение

CLIP (Contrastive Language–Image Pre-training) - нейронная сеть, которая позволяет классифицировать изображения, используя взаимосвязь между изображениями и текстом.

Мотивация

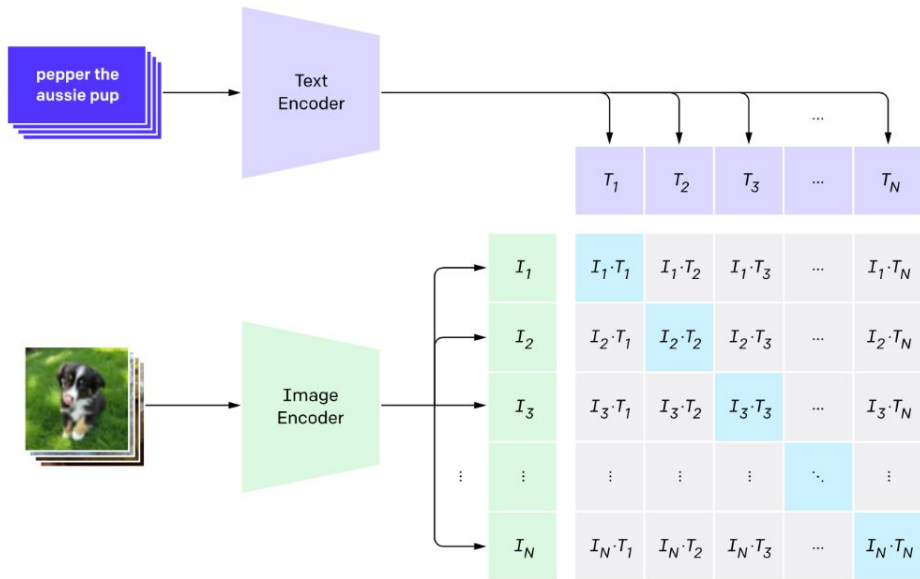
Проблемы в современном DL:

- дорогая разметка данных
- известные модели не универсальны (не могут предсказать то, чего не было при обучении)
- модели не устойчивы

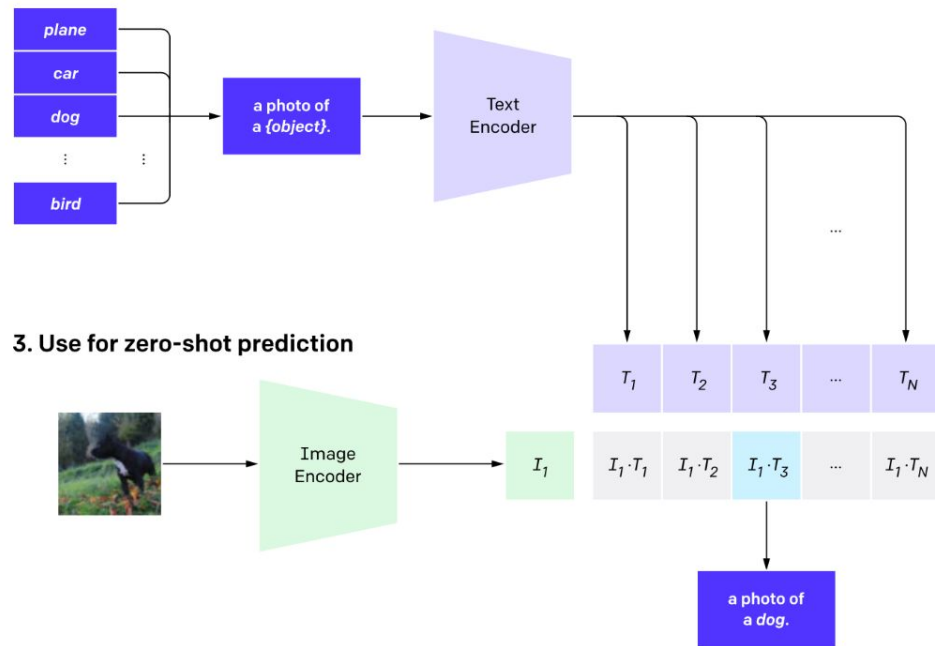
Dataset	ImageNet ResNet101	CLIP ViT-L
	76.2%	76.2%
ImageNet		
	64.3%	70.1%
ImageNet V2		
	37.7%	88.9%
ImageNet Rendition		
	32.6%	72.3%
ObjectNet		
	25.2%	60.2%
ImageNet Sketch		
	2.7%	77.1%
ImageNet Adversarial		

Архитектура

1. Contrastive pre-training



2. Create dataset classifier from label text



Предобучение

1. Contrastive pre-training

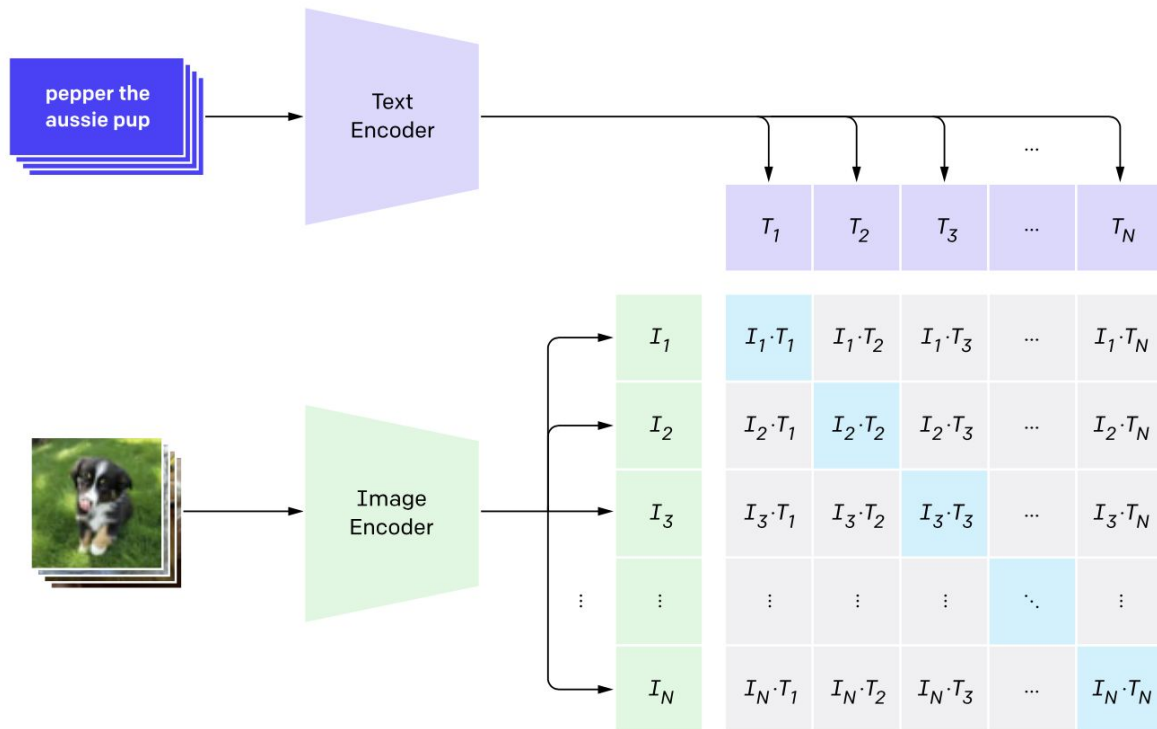


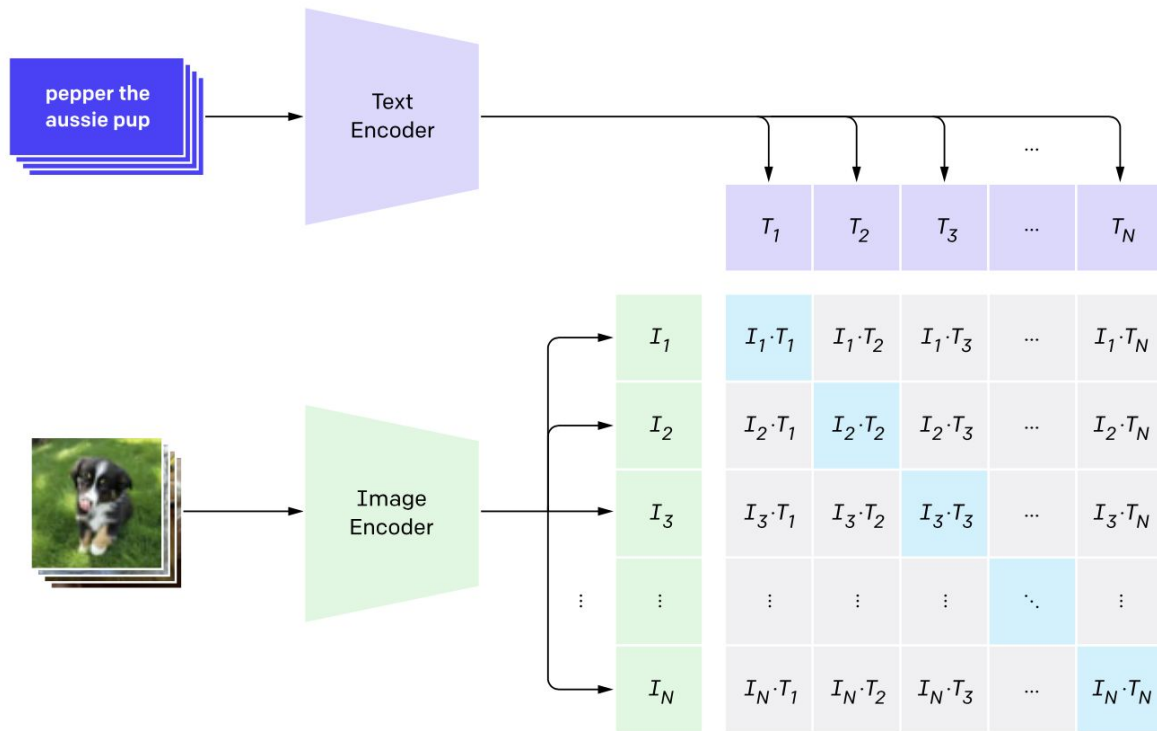
Image Encoder – ResNet или Vision Transformer

Text Encoder – Transformer (GPT-2)

Вход: пара <text, image> и одновременно обучается text / image encoder

Предобучение

1. Contrastive pre-training



$T_1..T_N$ - визуальные
репрезентации

$I_1 \dots I_N$ - текстовые
представления

Матрица похожести –
поэлементное cosine similarity
текстовых и визуальных
репрезентаций

На диагонали – правильные
пары

Максимизируем диагональные
элементы и минимизируем
внедиагональные с помощью
перекрестной энтропии


```

# 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

```

t - параметр температуры

loss_i , loss_t - перекрестные
энтропии по изображениям и
текстовым представлениям

loss - средняя энтропия

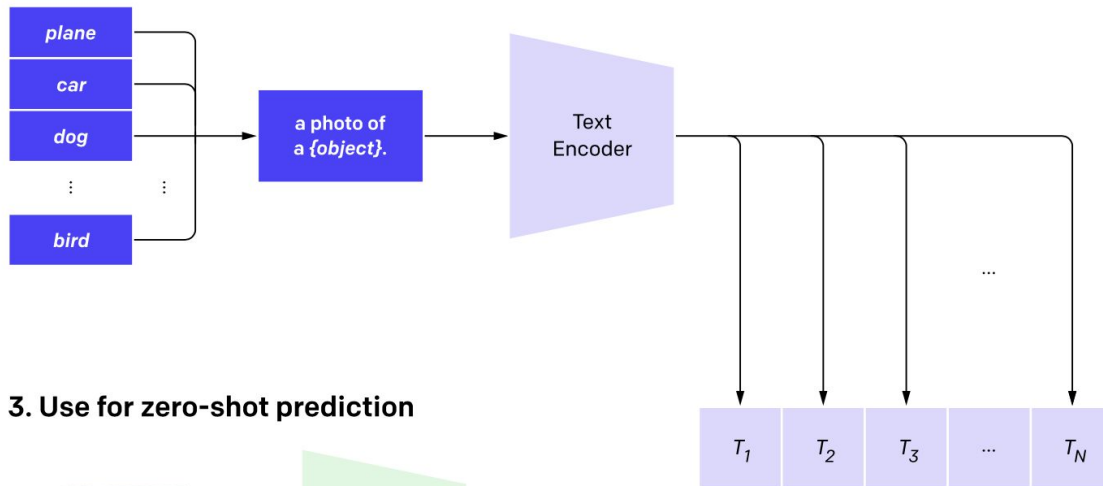
Figure 3. Numpy-like pseudocode for the core of an implementation of CLIP.

Детали обучения

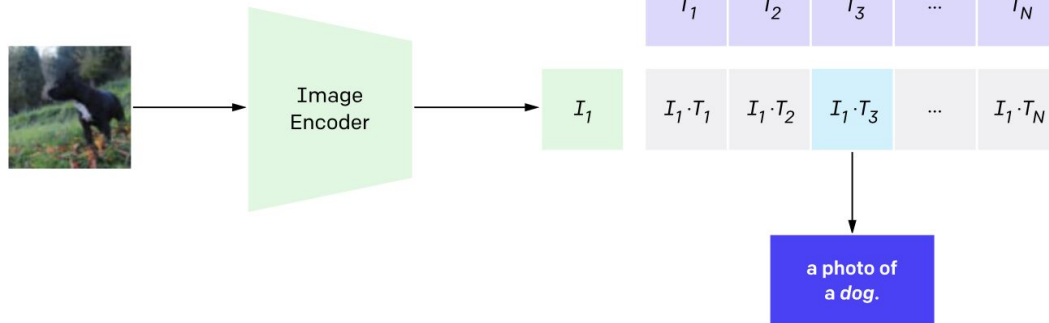
- собран датасет из 400 млн пар <text, image>
- Авторы статьи обучали 5 ResNet и 3 Vision Transformer
- 32 эпохи
- Adam
- learning rate - используя cosine schedule
- размер батча - 32 768
- обучение на ResNet заняло 18 дней
- обучение на Vision Transformer 12 дней

Результат: концепция Zero-shot learning

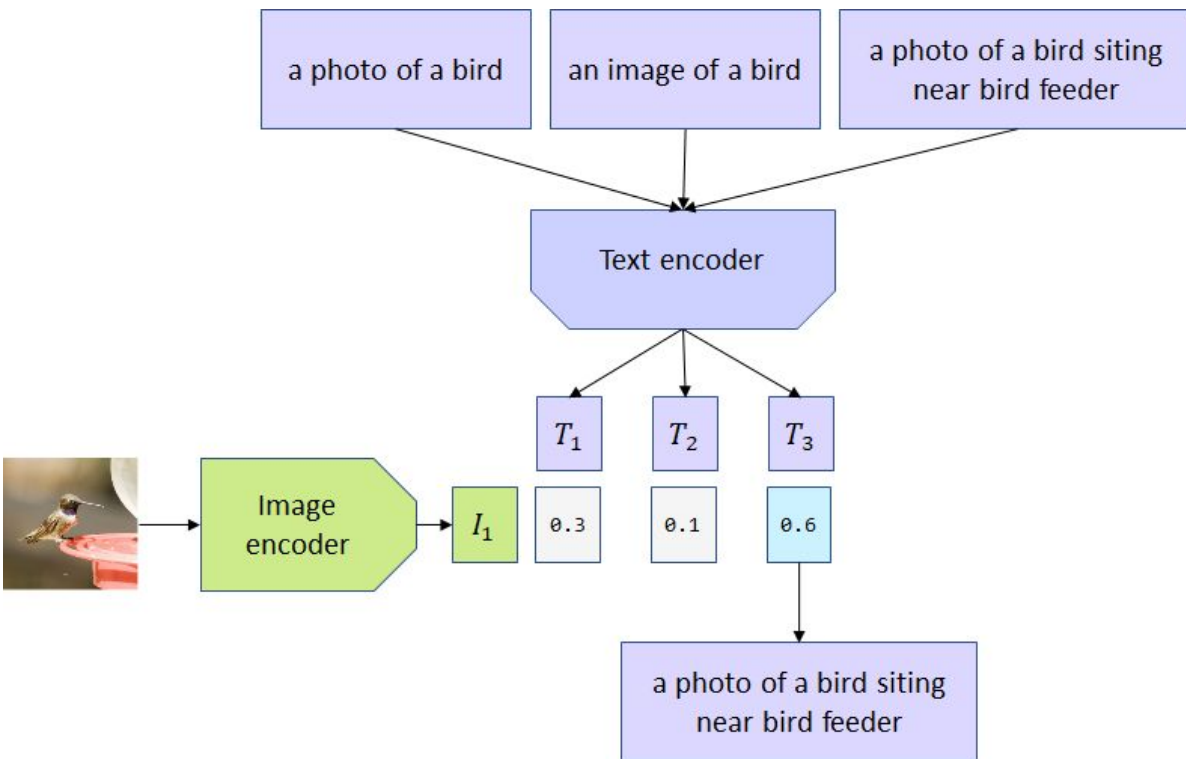
2. Create dataset classifier from label text



3. Use for zero-shot prediction



Модель учится
распознавать объекты,
которые никогда не видела,
используя взаимосвязи
между парами изображения
и текста



Деталь: модель достаточно чувствительна к окружающему контексту.

Эксперименты

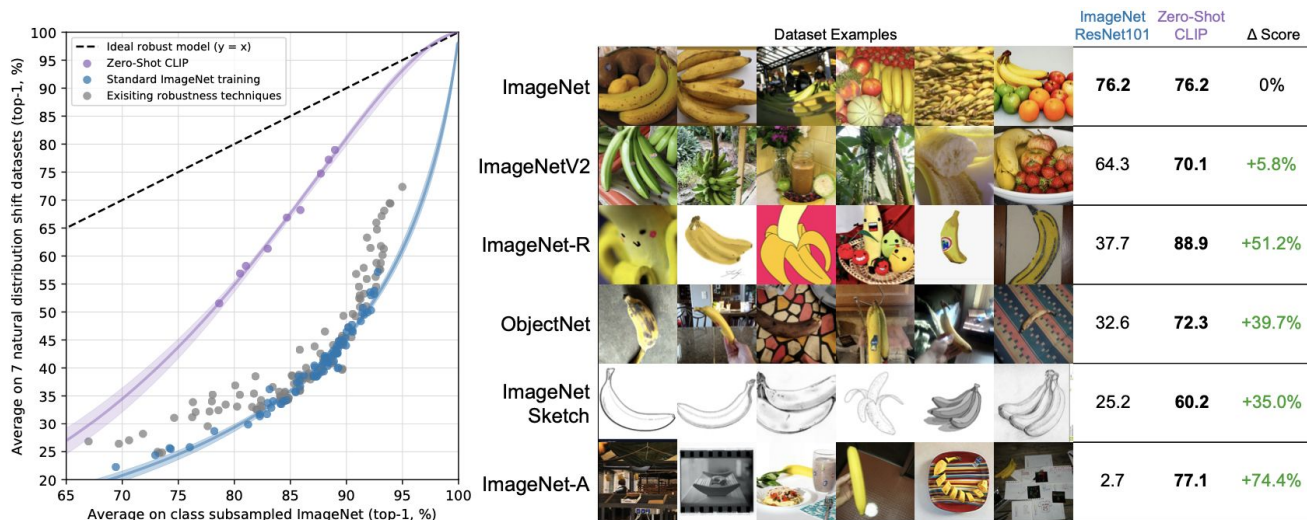
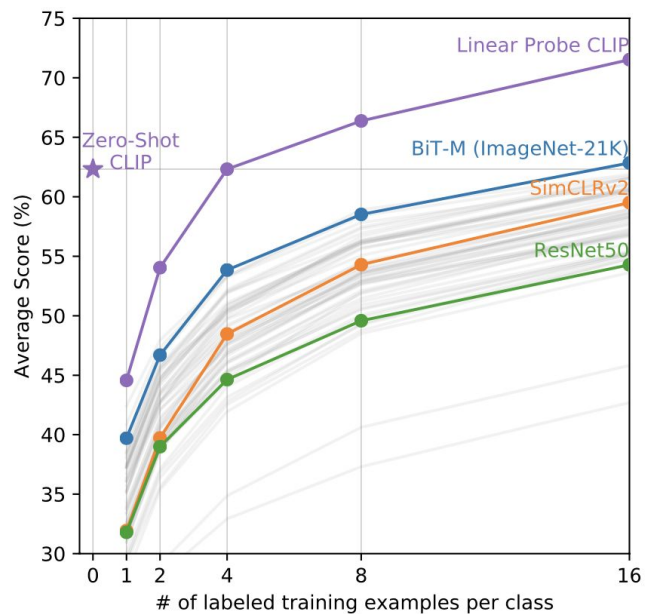


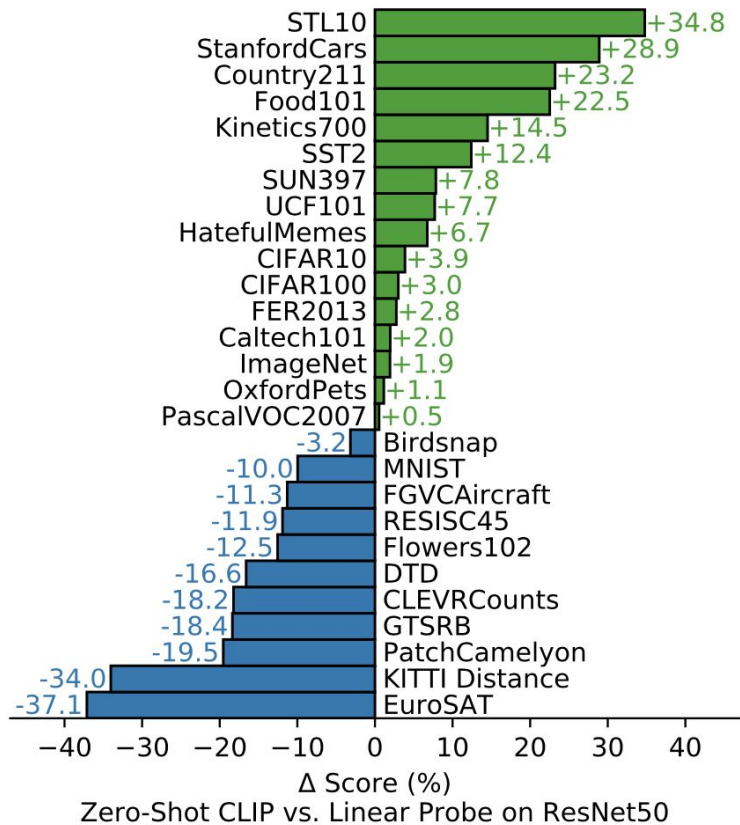
Figure 13. Zero-shot CLIP is much more robust to distribution shift than standard ImageNet models. (Left) An ideal robust model (dashed line) performs equally well on the ImageNet distribution and on other natural image distributions. Zero-shot CLIP models shrink this “robustness gap” by up to 75%. Linear fits on logit transformed values are shown with bootstrap estimated 95% confidence intervals. (Right) Visualizing distribution shift for bananas, a class shared across 5 of the 7 natural distribution shift datasets. The performance of the best zero-shot CLIP model, ViT-L/14@336px, is compared with a model that has the same performance on the ImageNet validation set, ResNet-101.

более устойчива



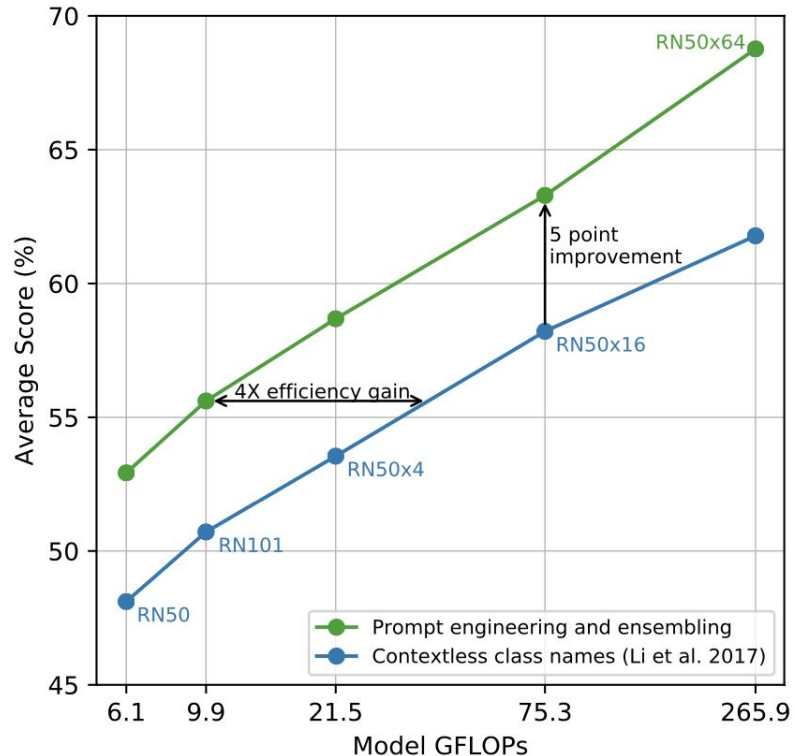
А может дообучить известные модели под новый класс?

Figure 6. Zero-shot CLIP outperforms few-shot linear probes. Zero-shot CLIP matches the average performance of a 4-shot linear classifier trained on the same feature space and nearly matches the best results of a 16-shot linear classifier across publicly available models. For both BiT-M and SimCLRv2, the best performing model is highlighted. Light gray lines are other models in the eval suite. The 20 datasets with at least 16 examples per class were used in this analysis.



Модель не подходит для
специализированных сфер
(распознавание модели машин,
рукописный текст, медицинская
сфера)

Figure 5. Zero-shot CLIP is competitive with a fully supervised baseline. Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.



prompt и ансамблирование
улучшают показатели на 5%

Figure 4. Prompt engineering and ensembling improve zero-shot performance. Compared to the baseline of using contextless class names, prompt engineering and ensembling boost zero-shot classification performance by almost 5 points on average across 36 datasets. This improvement is similar to the gain from using 4 times more compute with the baseline zero-shot method but is “free” when amortized over many predictions.

Выводы

- CLIP - модель, объединяющая в себе text, image encoder, использующая матрицу схожести для максимизации текстового описания изображения
- CLIP - это пример концепции zero-shot learning
- CLIP более робастный, устойчив в сдвигу распределений
- сбор данных легче и дешевле (пары <текст, изображение>)
- не подходит для специализированных сфер (марки машин, медицина, рукописный текст)

ИСТОЧНИКИ

- [CLIP: Connecting text and images](#)
- [Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision](#)