CLIP: Connecting Text and Images

Learning Transferable Visual Models From Natural Language Supervision

Петров Михаил, 193
апрель 2022

Zero-shot transfer в классификации изображений

Хотим предсказывать классы, которые не наблюдались при обучении.

Но это очень не просто: например, обычный линейный классификатор на эмбеддингах изображений не подстроится под новый класс без длительного дообучения.

Zero-shot transfer в классификации изображений

Хотим предсказывать классы, которые не наблюдались при обучении.

Но это очень не просто: например, обычный линейный классификатор на эмбеддингах изображений не подстроится под новый класс без длительного дообучения.

В идеале, хотим так построить и обучить модель, чтобы впоследствии уметь быстро обрабатывать любой датасет!

Идея: обучаться на текстовых описаниях!

Если обучаться на парах (картинка, описание) вместо пар (картинка, метка), то:

- таких данных в интернете на порядок больше;
- модель свяжет изображения не просто с метками, а с осмысленным текстом, что поможет в zero-shot transfer.

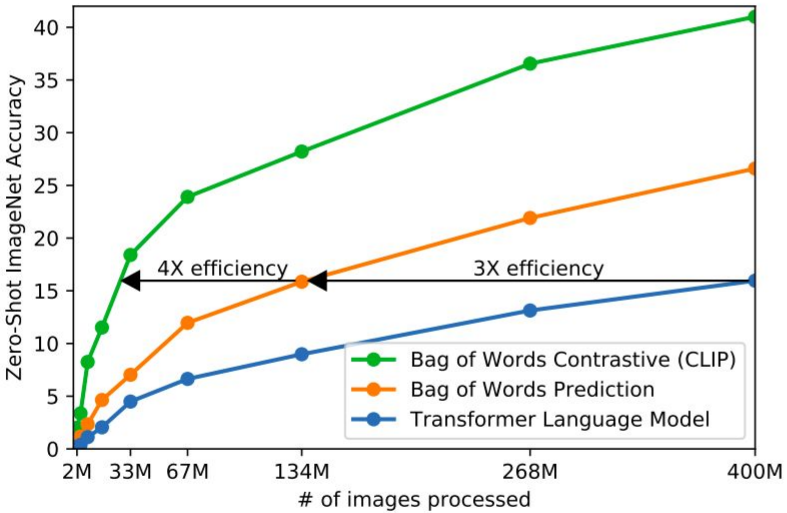
Идея: обучаться на текстовых описаниях!

Если обучаться на парах (картинка, описание) вместо пар (картинка, метка), то:

- таких данных в интернете на порядок больше;
- модель свяжет изображения не просто с метками, а с осмысленным текстом, что поможет в zero-shot transfer.

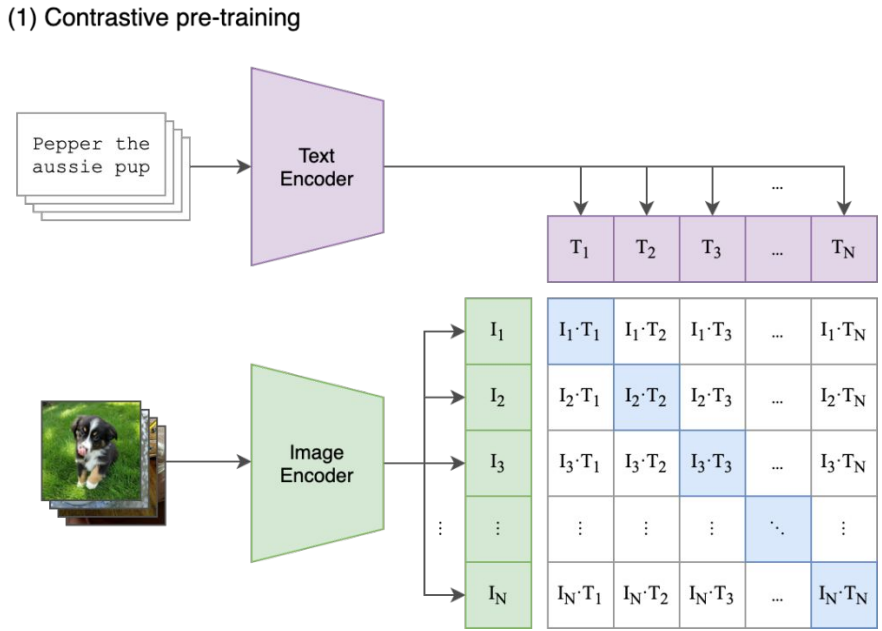
Поэтому вместе с энкодером для изображений (ResNet) будем использовать энкодер для текстов (Transformer). Но вместо *предиктивной* задачи мы будем решать **контрастивную**!

Предиктивная и контрастивная задачи



Предиктивная задача заключается в том, что мы на каждом объекте требуем попадания в соответствующее описание.
Контрастивная задача заключается в том, чтобы для данных N картинок и N описаний составить наилучшее попарное соответствие.

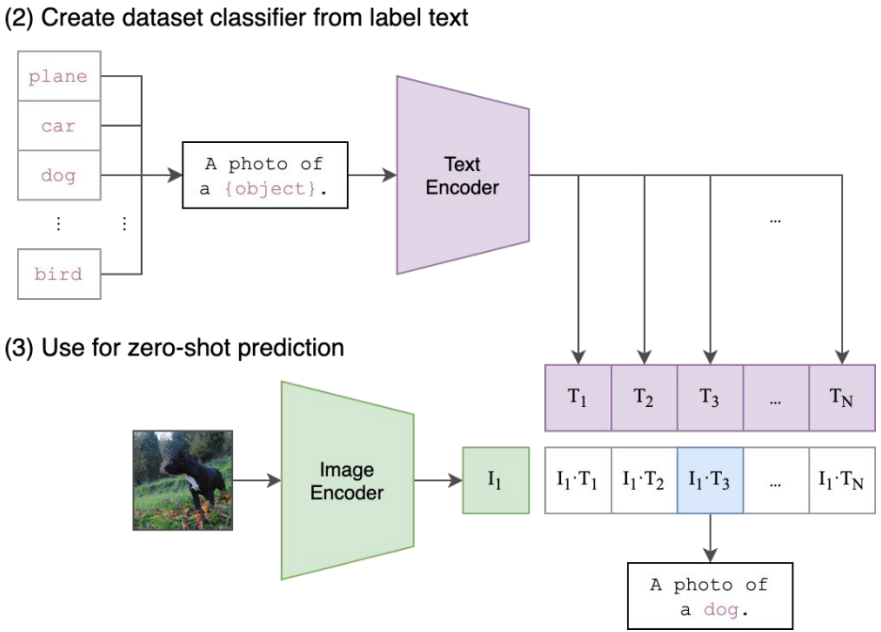
CLIP: Contrastive Language-Image Pre-training



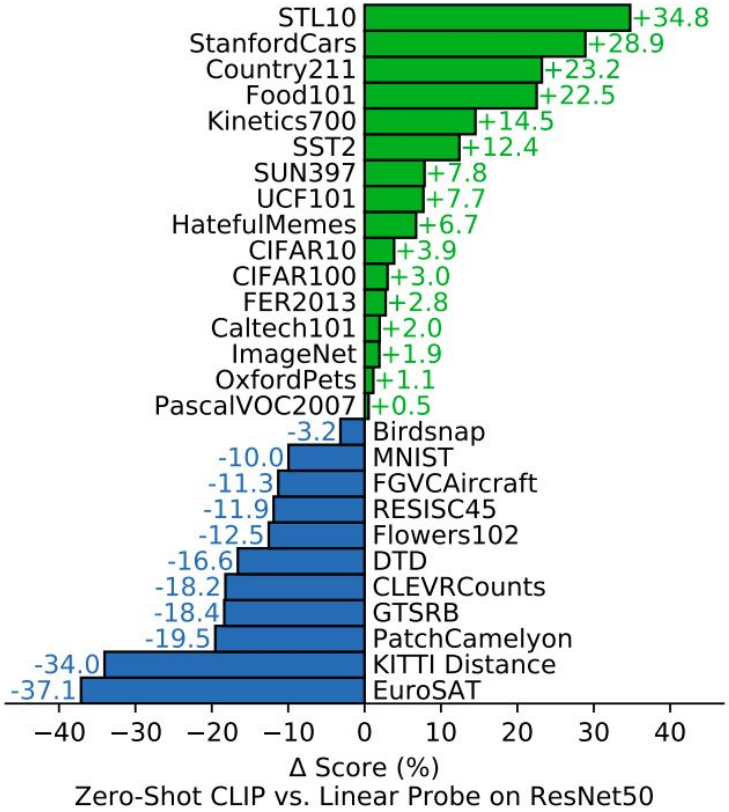
I_1, …, I_N, T_1, …, T_N (N = 32768) – векторы-представления изображений и текстов в батче (все одного размера). Считаем скалярные произведения $I_j \cdot T_k$, затем максимизируем те, для которых j = k, и минимизируем все остальные.

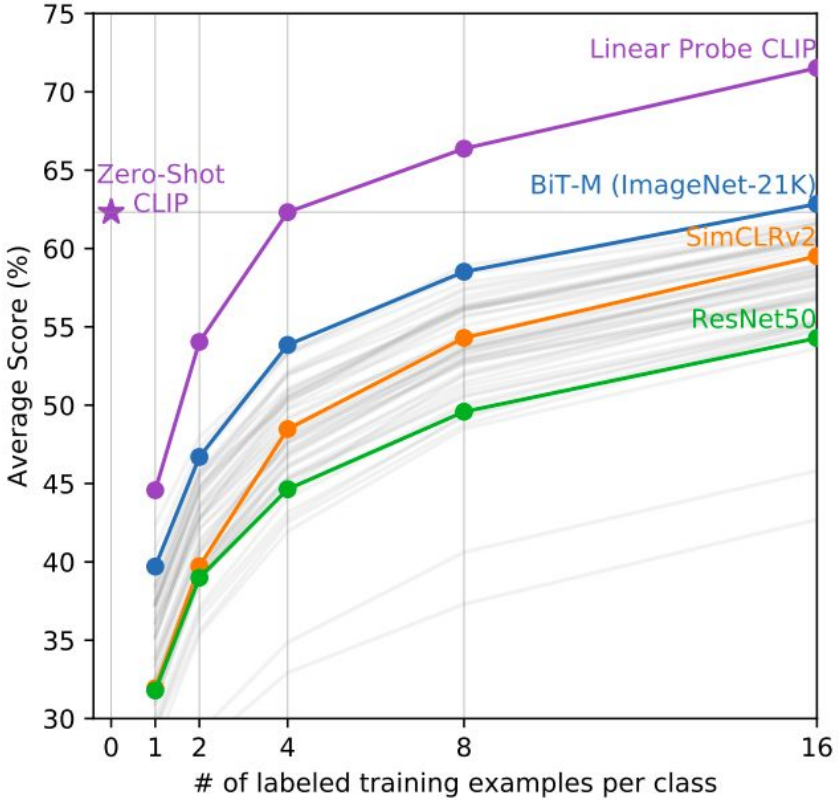
CLIP: Contrastive Language-Image Pre-training

Оборачиваем названия классов в более развёрнутые предложения (контекст зависит от датасета). Полученные скалярные произведения оборачиваем в softmax и подаём на выход как итоговые вероятности.

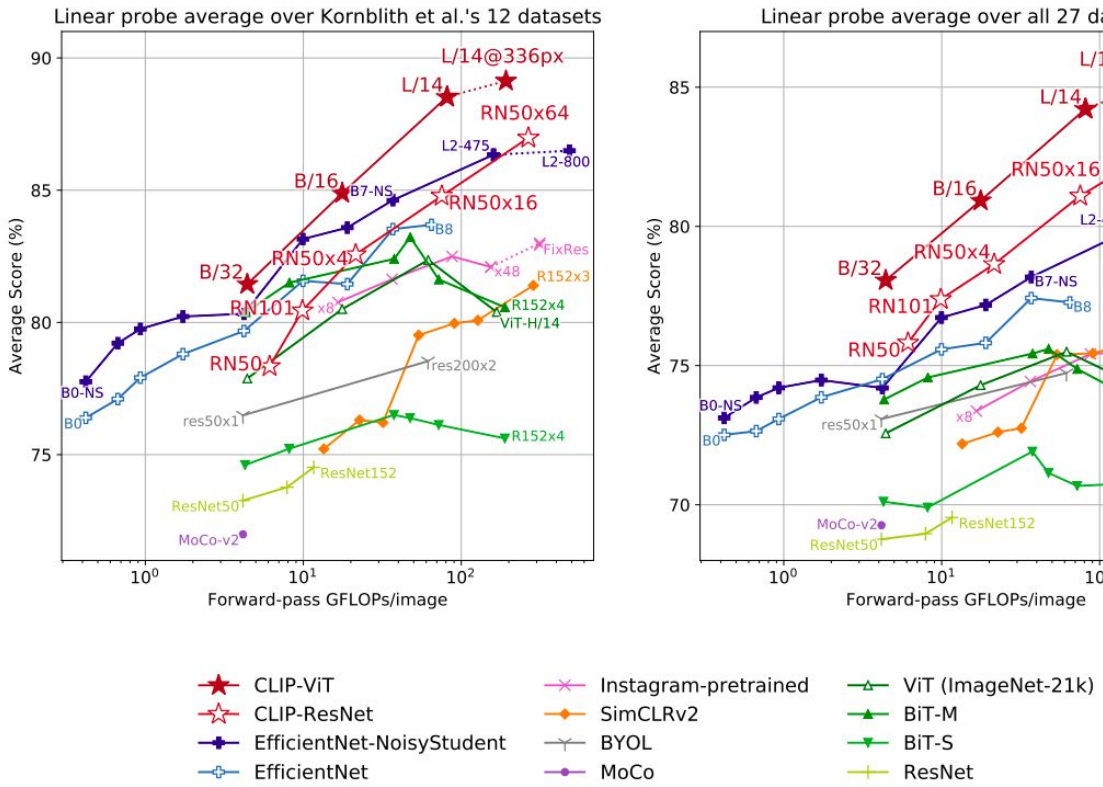


Эксперименты: zero-shot и few-shot learning

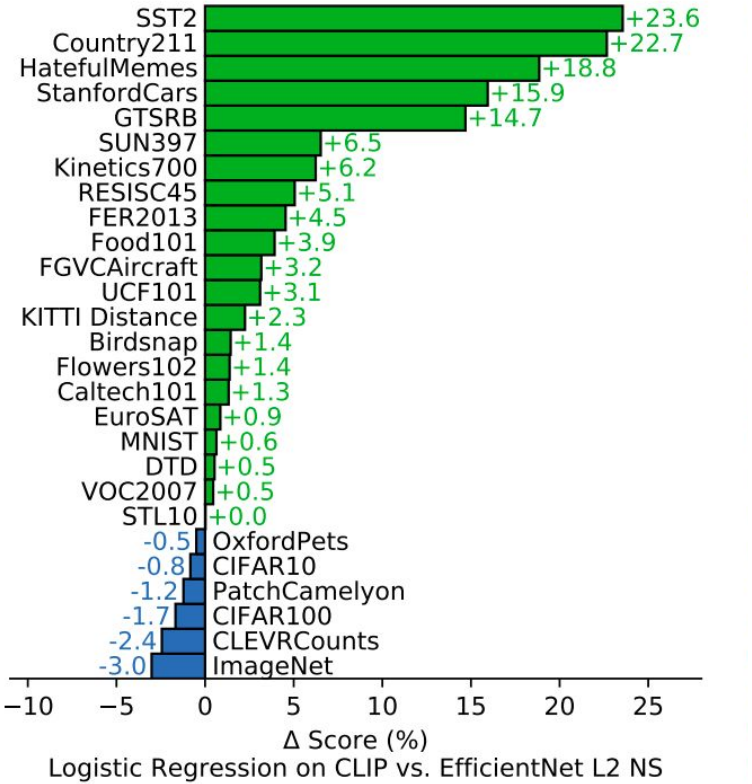




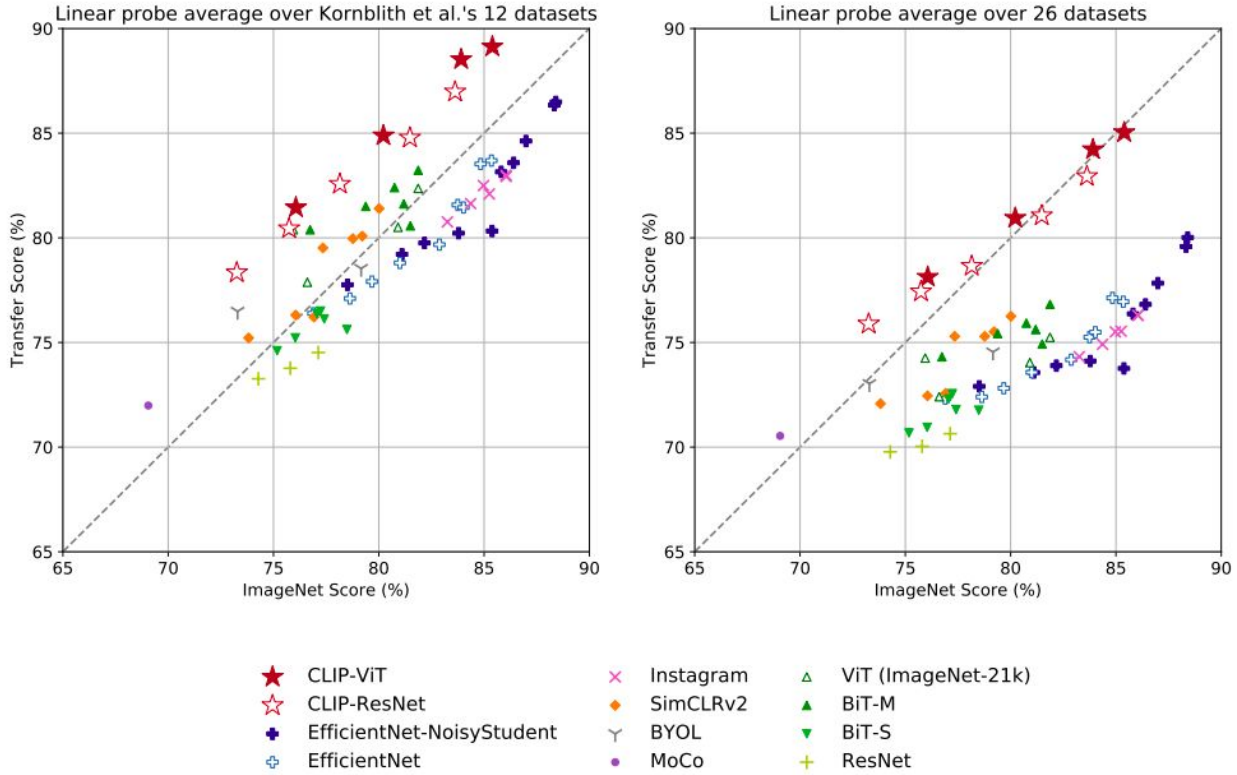
Эксперименты: representation learning



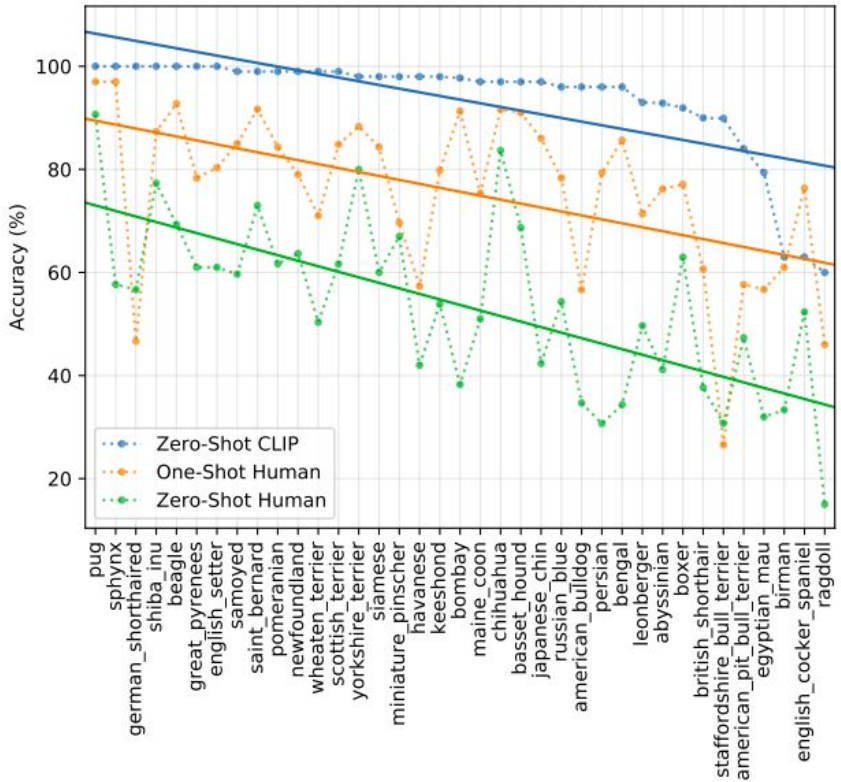
Эксперименты: representation learning



Эксперименты: task shift



Эксперименты: сравнение с человеком



Итоги

- Чтобы построить универсальный классификатор изображений, достаточно обучить модель на контрастивной задаче с использованием подробных текстовых описаний вместо обычных меток.
- Модель можно использовать несколькими способами: unsupervised (pre-trained), supervised, pre-trained and supervised.

Список литературы

Сама работа: https://arxiv.org/abs/2103.00020

Репозиторий с примерами: https://github.com/openai/CLIP

InfoNCE Loss (функция потерь, использованная в контрастивной задаче): https://arxiv.org/abs/1807.03748 (пункт 2.3)