

CLIP

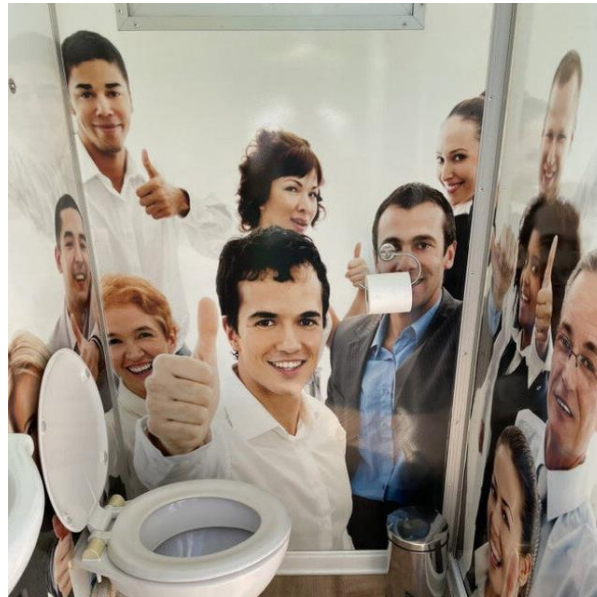
Сапожников Денис, БПМИ-192

Few-shot image classification

- Как бы вы решали задачу сейчас?
 - Image Embeddings + Linear Probe. Есть ли проблемы при few-shot?
 - Добавим L2-loss от Linear Probe
- Почему приходится кардинально менять Linear Probe при смене классов, а Image Embeddings – нет?
 - Эмбеддинги картинок хорошо описывают пространство всех возможных картинок
 - В то же время, пространство различных классов вообще никак не описано

Как описать пространство классов?

- Эмбединги классов?



- Некоторые картинки невозможно описать одним классом
- Решение: текстовое описание картинки вместо классов!

Где взять большой датасет?

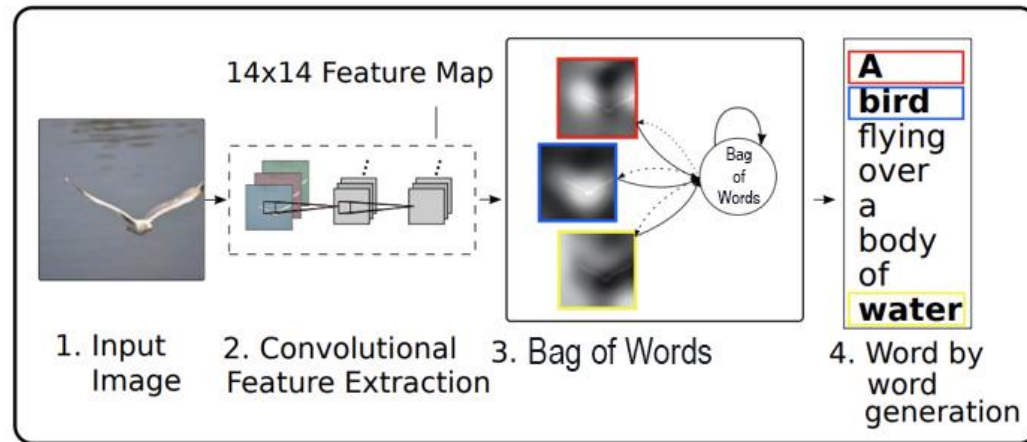
- Обычные датасеты вида (image, label)
 - слишком маленькая выразительная способность у label
- Instagram dataset
 - 3.5 миллиардов картинок с описаниями
 - плохие подписи
- Yahoo Flickr Creative Commons 100M (YFCC100M)
 - 100 миллионов фоток с подписями
 - некоторые имеют вид 20160126_135930.jpg
 - После обработки – 15 миллионов картинок – мало

Майним свой датасет

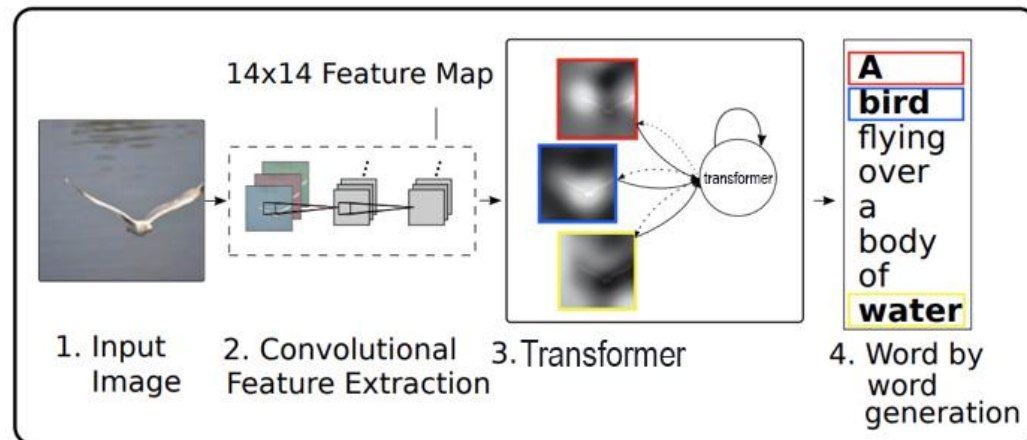
- Возьмем 500'000 достаточно популярных слов
- Возьмем порядка 20к фоток по этому слову из Интернета
- Получаем датасет размера 400 миллионов

Image Captioning

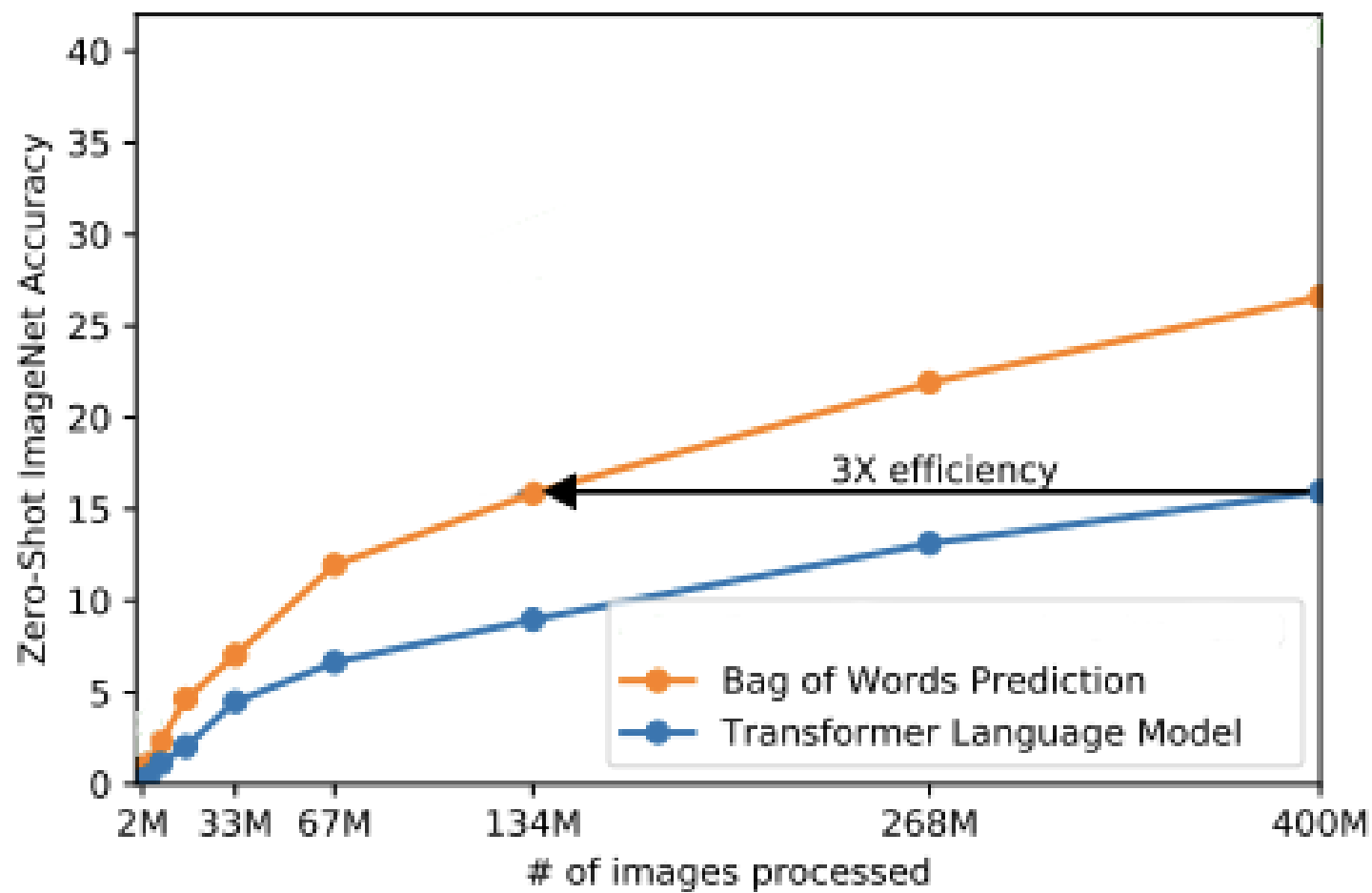
- Bag of Words



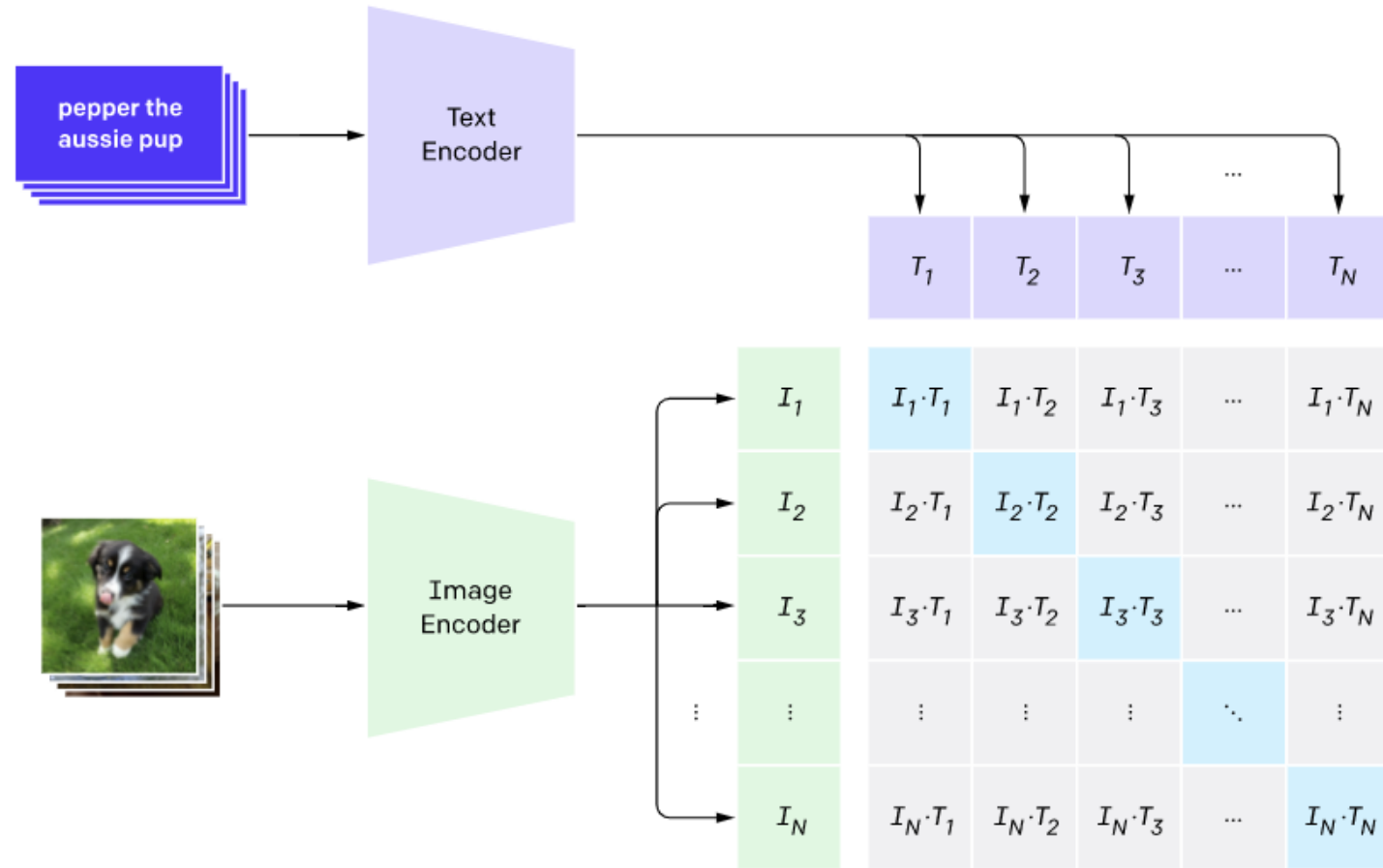
- Transformer



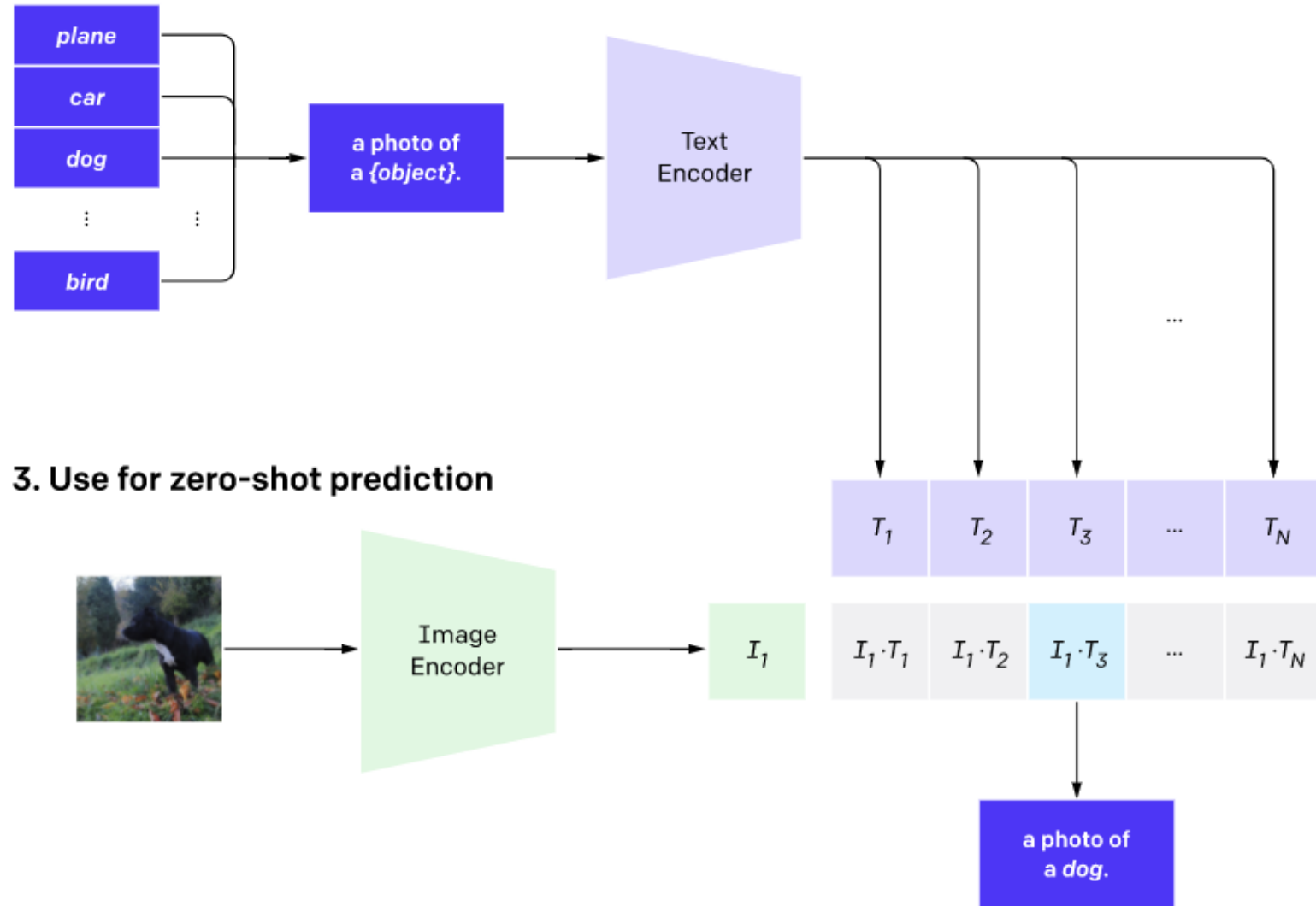
Результаты Image Captioning



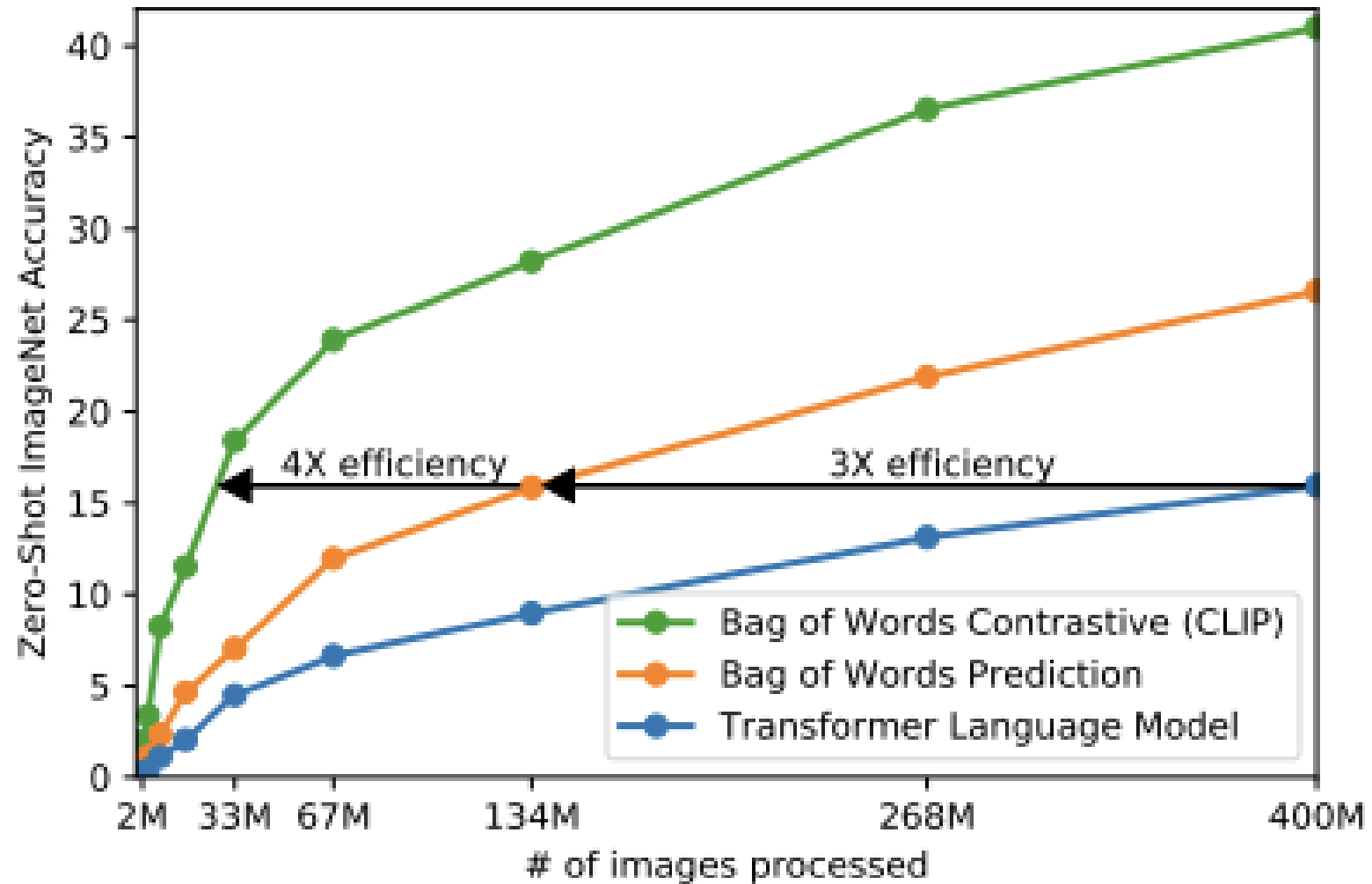
Contrastive Loss



Classification

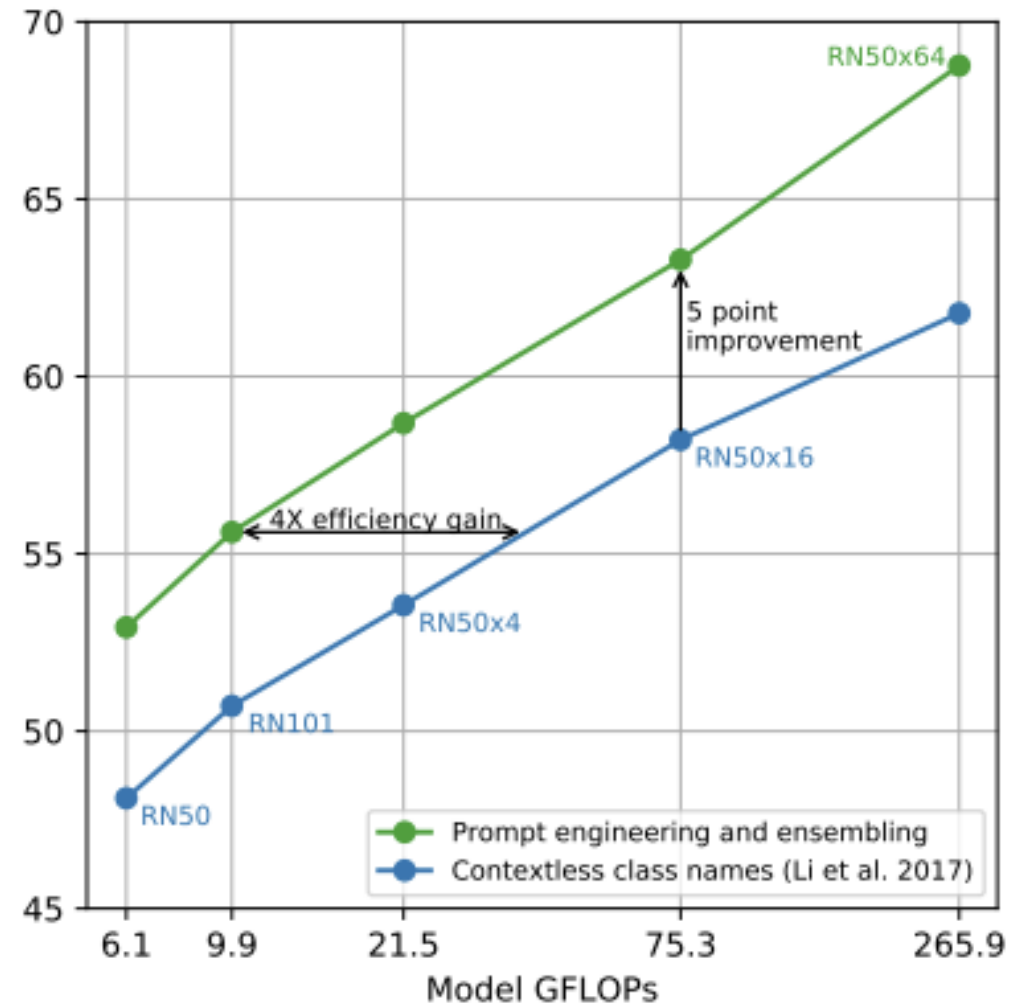


Сравнение с Image Captioning



Что в итоговой модели?

- Image Encoder – ResNet50
- Text Encoder – transformer



Сравнение с fully supervised подходом

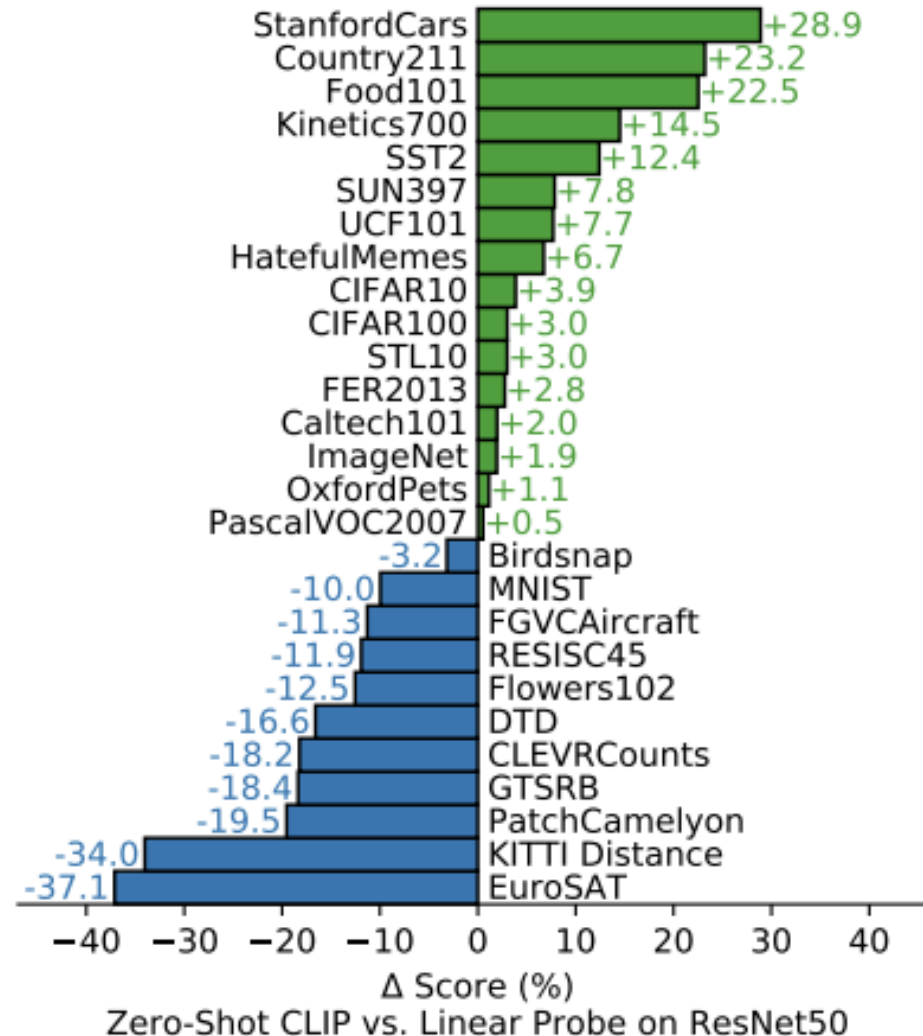






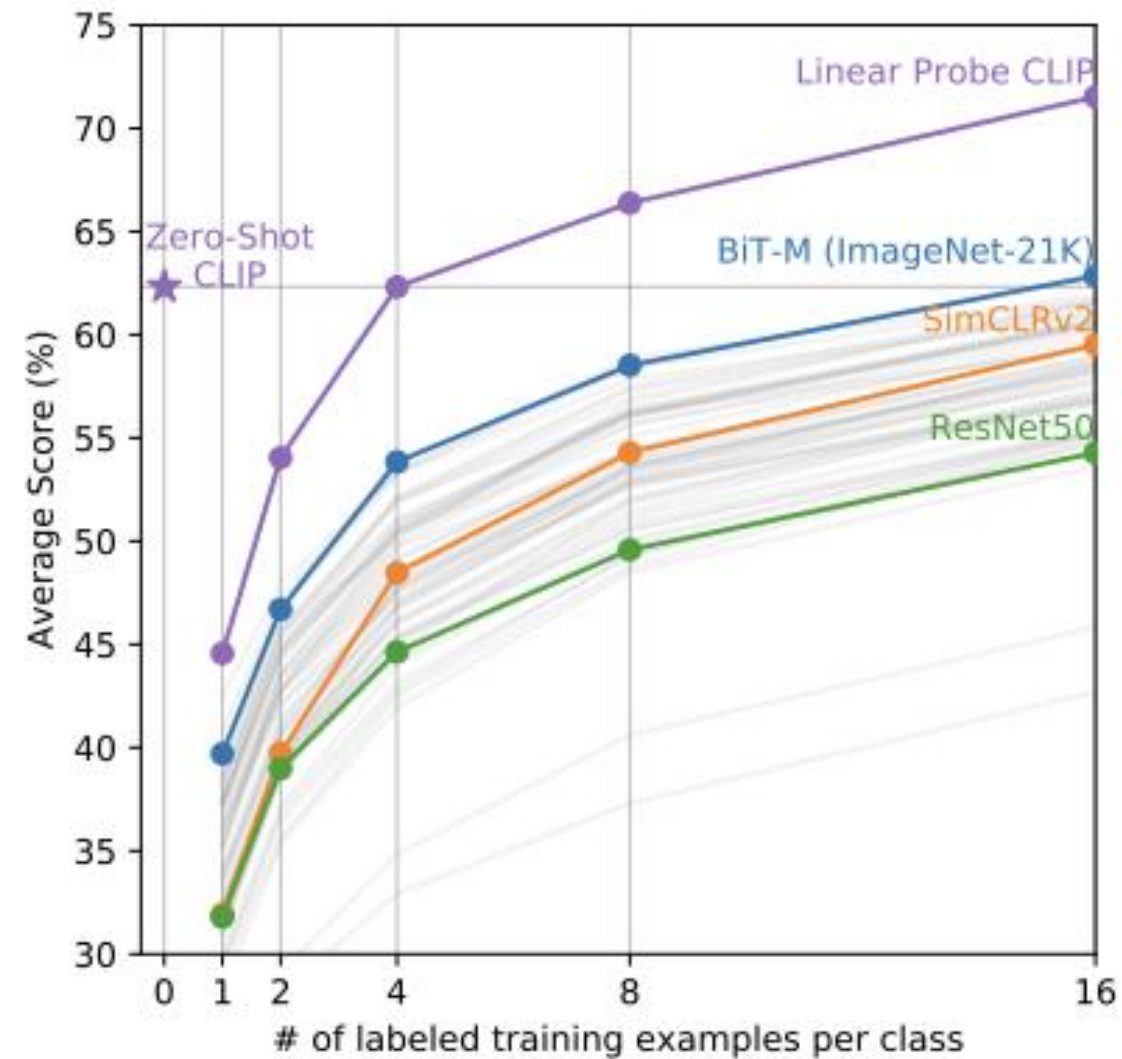


Image Net

DATASET	IMAGENET RESNET101	CLIP VIT-L
 ImageNet	<div><div></div></div> 76.2%	<div><div></div></div> 76.2%
 ImageNet V2	<div><div></div></div> 64.3%	<div><div></div></div> 70.1%
 ImageNet Rendition	<div><div></div></div> 37.7%	<div><div></div></div> 88.9%
 ObjectNet	<div><div></div></div> 32.6%	<div><div></div></div> 72.3%
 ImageNet Sketch	<div><div></div></div> 25.2%	<div><div></div></div> 60.2%
 ImageNet Adversarial	<div><div></div></div> 2.7%	<div><div></div></div> 77.1%

Few-Shot vs Zero-Shot

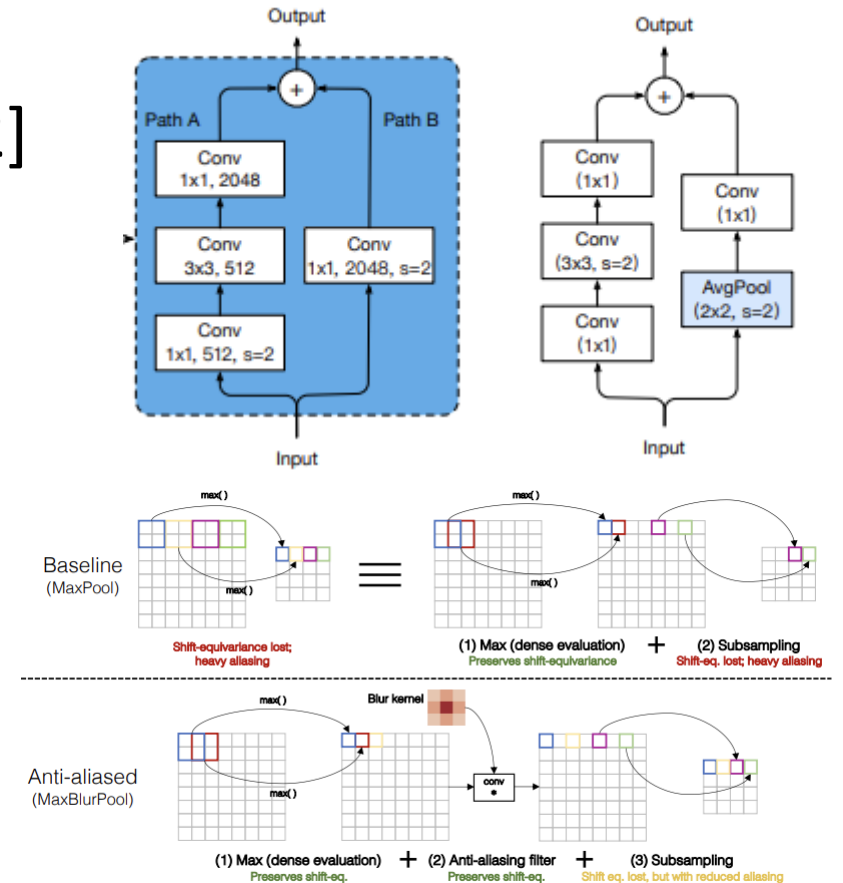


Бонус №1: Как модифицировали ResNet?

- На самом деле за основу взят ResNet-D [2]

- Заменены Max Pooling на Blur Pooling [3]

- Заменены Global Average Pooling на Attention Layer



Бонус №2: Факты

- Авторы во введении жалуются во введении на то, что SoTA модели в CV обучаются 19 GPU-лет
- На обучение потратили 592 V100 в течение 18 дней (29 GPU-лет)
- Чтобы добиться высокого качества на разных датасетах, авторы подбирали разные фразы, чтобы превратить label в предложение, например:
 - На датасете Oxford Pets: A photo of a {label}, a type of pet.
 - На датасете со спутников: A satellite photo of a {label}.
 - Иногда помогало оборачивание label в кавычки

Список литературы

- [1] [Radford et al. Learning Transferable Visual Models From Natural Language Supervision \(2021\)](#)
- [2] [He et al. Bag of Tricks for Image Classification with Convolutional Neural Networks \(2019\)](#)
- [3] [Zhang. Making Convolutional Networks Shift-Invariant Again \(2019\)](#)