Факультет Компьютерных Наук

Стриженок Сергей
БПМИ-191

Москва
2022

Learning Transferable Visual Models From Natural Language Supervision

Проблемы классических решений компьютерного зрения

* Предсказываем ограниченное число заранее определенных категорий
* Нужна дополнительная разметка для новых данных

Идея

* Будем пытаться предсказать соответствие описания картинке
* Идея не новая, но простые подходы дают лишь 11.5% точности против 88.4% у лучшего решения, 50% у классических решений
* Описание картинок это хороший сигнал для обучения

Как собрать данные для обучения? Какие есть проблемы?

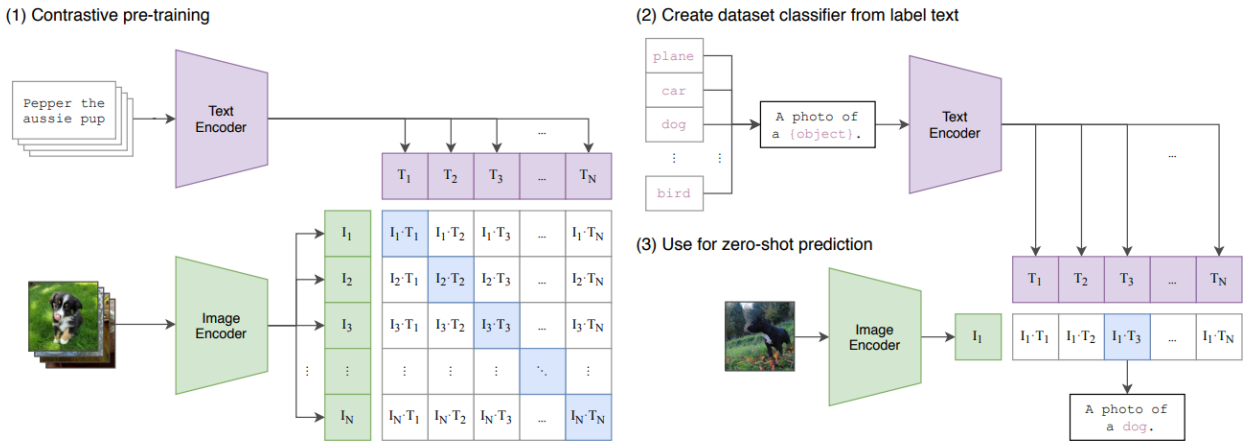
* Нужна большая тренировочная выборка
* MS-COCO и Visual Genome имеют размер всего в 100 тысяч фото
* Обычные модели тренируются на 3.5 миллиардах фото
* YFCC100M — уже 100 миллионов, но после фильтрации всего 6-15 миллионов

Как собрать данные для обучения?

* Собрали выборку из 400 миллионов примеров пар (описание, картинка)
* Данные брались из интернета
* Для большей общности брали слова, которые встречались в википедии не менее 100 раз
* Число слов оказалось примерно сопоставимо с WebText

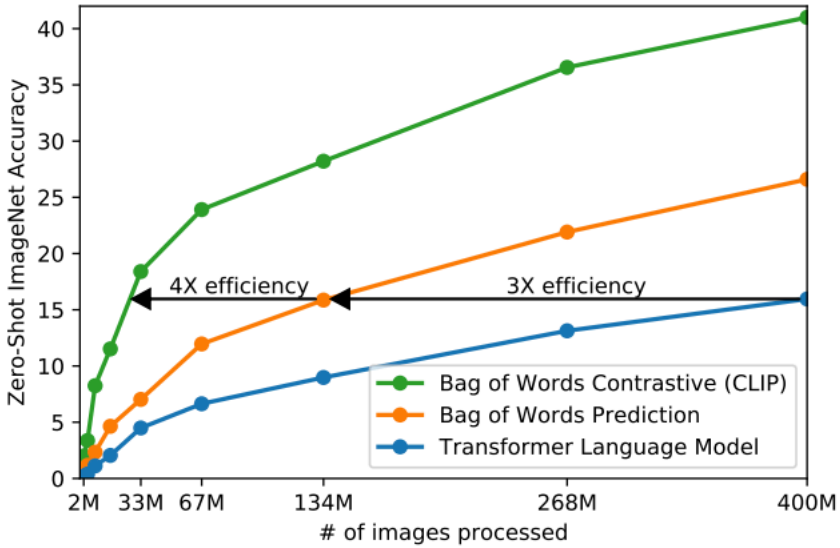
Выбор модели для предобучения

* Изначальный подход провалился
* Пытались предсказывать точные слова в описании
* Финальный подход предсказывает соответствие описания картинке



Скорость обучения

* Изначальный работал в 3 раза медленнее, чем самая простая модель
* Финальный подход работает в 4 раза быстрее



```
# 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
```

Модель: архитектуры

* Для кодировщика изображений: ResNet и Vision transformer
* Для кодировщика текста: Transformer(Vaswani et al., 2017)

Модель: обучение

* 5 ResNet’ов и 3 Vision Transformer’a
* Использовался Adam
* Гиперпараметры подбирались с помощью grid search и random search на ResNet-50 с эпохой 1

Zero-shoot

Проблема в машинном обучении, когда на вход модели поступают данные, которые она не видела при обучении

Zero-shoot

* Используем в качестве описания названия классов
* Считаем представления картинок
* Считаем представления текстов
* Считаем косинус между векторами

CLIP vs Visual N-grams

* На zero-shoot CLIP оказался лучше, достиг точности в 76.2%
* Оказался сравним с ResNet-50, не используя ни один пример из выборки
* Не все так однозначно: в 10 раз больше обучающая выборка, в в 100 раз дольше предсказание, в 1000 раз больше ресурсов на обучение
* Но на более маленьких выборках CLIP тоже победил

	aYahoo	ImageNet	SUN
Visual N-Grams	72.4	11.5	23.0
CLIP	**98.4**	**76.2**	**58.5**

Сравнение модели и человека

* CLIP показывает лучшие результаты на zero-shoot
* Люди быстро обучаются и уже один пример сильно растит точность
* Улучшения в CLIP: few-shoot learning

	Accuracy	Majority Vote on Full Dataset	Accuracy on Guesses	Majority Vote Accuracy on Guesses
Zero-shot human	53.7	57.0	69.7	63.9
Zero-shot CLIP	**93.5**	**93.5**	**93.5**	**93.5**
One-shot human	75.7	80.3	78.5	81.2
Two-shot human	75.7	85.0	79.2	86.1

Table 2. Comparison of human performance on Oxford IIT Pets. As in Parkhi et al. (2012), the metric is average per-class classification accuracy. Most of the gain in performance when going from the human zero shot case to the human one shot case is on images that participants were highly uncertain on. "Guesses" refers to restricting the dataset to where participants selected an answer other than "I don't know", the "majority vote" is taking the most frequent (exclusive of ties) answer per image.