

Обзор статьи «Image Transformers»

Поконечный Эдуард
ФИБТ МФТИ, группа M05-014г
@celidos
pokonechnyy.er@phystech.edu

15 января 2021 г.

Аннотация

Ссылка на статью: <https://arxiv.org/abs/1802.05751>.

Авторы исходной статьи: Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, Dustin Tran.

1 Введение

Существует множество архитектур для моделирования естественно выглядящего распределения изображений. Например PixelRNN, PixelCNN. PixelRNN плохо параллелизуются; PixelCNN имеют ограниченное рецептивное поле, особенно если в модели малое число слоев.

Авторы работы хотят показать, что механизм self-attention представляет собой баланс между бесконечным рецептивным полем у PixelRNN и ограниченным рецептивным полем у PixelCNN. Авторы проверяют качество работы на двух задачах: условная генерация изображений (сигнал - класс изображения) и superresolution.

2 Представление картинки

Используется два представления:

1. категориальное — интенсивность каждого пикселя представлена одним из 256 векторов размерности d (для всех трех цветов)
2. численное — изображение обрабатывается сверткой для получения размера $[h, w, d]$

К представлениям пикселей также добавляется positional encoding, он также бывает двух видов:

1. \sin/\cos [3]. Общий вид функций:

$$PE_{(pos, 2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right)$$

$$PE_{(pos, 2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right)$$

2. позиционные эмбединги.

3 Механизм self-attention

Архитектура энкодер-декодер.

Как и в задачах обработки естественного языка, трансформер на изображении – это чередование слоев self-attention и обычных полносвязных слоев, применяемых поэлементно.

Общая схема работы: есть вектора запросов (q , query) и вектора ключей (M , она же память, memory). Для вектора q и для всех векторов из памяти M вычисляются веса attention – уровень релевантности данного вектора из M текущему вектору q . Затем эти веса используются для вычисления взвешенной суммы всех преобразованных с помощью матрицы W_v векторов из блока памяти M . Полученный вектор считается новым представлением текущей позиции и отправляется дальше (в данном случае, направляется через слой dropout в residual-соединение).

$$q_a = \text{LN}(q + \text{DO}(\text{softmax}\left(\frac{W_q q (MW_k)^T}{\sqrt{d}}\right) MW_v))$$

$$q' = \text{LN}(q_a + \text{DO}(W_1[W_2 q_a]_+))$$

где DO – дропаут, LN – layer norm. Общий вид layer norm [1]:

$$\bar{a}_i^l = \frac{\text{gain}}{\sigma_i^l} (a_i^l - \mu_i^l)$$

$$\mu^l = \frac{1}{H} \sum_{i=1}^H a_i^l$$

$$\sigma^l = \sqrt{\frac{1}{H} \sum_{i=1}^H (a_i^l - \mu^l)^2}$$

блок и блок памяти. В случае двумерного внимания, изображение не вытягивается в строку, а выбор блоков запросов и блоков памяти сохраняет пространственную структуру.



5 Эксперименты и выводы

На безусловной генерации на CIFAR-10 результаты сравнимы с PixelCNN++ и PixelSNAIL (еще одна основанная на attention модель). На ImageNet показала лучшие результаты по состоянию на 2018 год. Увеличение рецептивного поля ведет к значительному улучшению перплексии.

При условной генерации сигналом является обучаемый эмбединг класса. Проверяли на CIFAR-10, лог. правдоподобие примерно такое же, как и при безусловной генерации, но выше качество для восприятия человеком.

На задаче superresolution занимались увеличением картинок с размера 8×8 до 32×32 , проверяли на наборе данных CelebA и CIFAR-10.

В целом сейчас много еще более любопытных и многообещающих статей по трансформерам [2][4], эта уже достаточно старая. Но авторы в этой статье показали принципиальную применимость трансформеров для ряда задач, считаю, что со своей задачей они справились.

Также беспокоит, что большая часть экспериментов была поставлена на картинках очень небольшого размера (до 32×32). Скорее всего, это говорит о том, что трансформеры в чистом виде слишком тяжелы для применения к крупным картинкам непосредственно. Большая часть имеющихся моделей с трансформерами так или иначе пытается сократить «зону ответственности» и уменьшить число позиций, которые должен посетить attention.

Список литературы

- [1] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer Normalization. 2016.
- [2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: transformers for image recognition at scale. 2020.
- [3] S. Takase and N. Okazaki. Positional Encoding to Control Output Sequence Length. 2019.
- [4] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers distillation through attention. 2020.