УДК 004.89

Использование матричных разложений для токенизации изображений в задаче обучения моделей типа Transformer

Прокопов Е.М. (ИТМО)
Научный руководитель – кандидат физико-математических наук, доцент Бойцев А.А. (ИТМО)

Введение. В последние годы модели архитектуры Transformer получили широкое распространение в задачах обработки естественного языка, а с появлением Vision Transformer (ViT) - в задачах компьютерного зрения. Для обработки изображений, алгоритм ViT предварительно разбивает их на малые части, чаще всего 16 на 16 пикселей, после чего подает на вход кодировщика. Такой подход может привести к избыточной длине входной последовательности, что увеличивает вычислительную сложность. Также это ухудшает восприятие локальных признаков, поскольку размер разбиения фиксирован вне зависимости от размера исходного изображения.

В данной работе предложено использование матричных разложений (таких как сингулярное разложение или преобразование Фурье) для создания более гибких методов токенизации, позволяющих адаптировать длину контекста, сохраняя при этом основную информацию изображения.

Основная часть. С помощью матричных разложений, изображение разбивается на несколько матриц меньшего порядка. В частности, рассматриваются такие методы как:

1) Сингулярное разложение (SVD) – разложение матрицы прямоугольной формы в произведение трех матриц: левых сингулярных векторов, диагональной матрицы сингулярных чисел и правых сингулярных векторов. По теореме Эккарта-Янга, наилучшая матрица для приближения заданной матрицы с заранее заданным рангом получается из сингулярного разложения исходной матрицы. Таким образом, возможно значительно сократить длину контекста, пожертвовав малой долей информации изображения.
2) Преобразование Фурье – преобразование, при котором сигнал (изображение в частности) раскладывается на гармонические составляющие. Данный метод позволяет уменьшить длину входной последовательности, снизив влияние шумов (если они присутствуют), а также пожертвовав мало информативными высокочастотными признаками.
3) Вейвлет-преобразование – преобразование, позволяющее проанализировать сигнал с одновременным учетом как частотных, так и пространственных характеристик. При анализе изображений данные разделяются на крупные структуры и детали на разных масштабах. Глубина разложения может быть настраиваемой, что также позволяет уменьшить длину контекста, пожертвовав малыми деталями.

Далее рассматриваются методы, позволяющие получить набор векторов входной последовательности, перед обработкой ее кодировщиком модели. В конце обученная модель Transformer с разобранными методами токенизации сравнивается с обученной моделью Vision Transformer.

Выводы. Проведено сравнение различных матричных разложений для токенизации изображений и стандартного метода разбиения изображения на малые части при обучении модели типа Transformer.

Список использованных источников:

1. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Houlsby, N.

(2021).
2. Wavelets Are All You Need for Autoregressive Image Generation, Mattar, W., Levy, I., Sharon, N., & Dekel, S. (2024).