ВВЕДЕНИЕ

В последние годы архитектуры нейронных сетей, основанные на модели Transformer, изначально разработанной для решения задачи машинного перевода, получили широкое распространение. Ключевой механизм данной архитектуры – механизм внимания, который позволяет обрабатывать все элементы входной последовательности параллельно. Такой подход ускоряет обучение при использовании графических ускорителей, значительно улучшает понимание нейронной сетью глобального контекста, а также положительно влияет на масштабируемость системы, позволяя более эффективно обучать нейронные сети с большим числом параметров и на больших наборах данных. Проблемой механизма внимания является вычислительная сложность и требуемая память $O(n^2)$, где n – длина входной последовательности.

Изначально модели типа Transformer применялись исключительно для обработки последовательных данных, в основном текста. Однако с появлением модели Vision Transformer (ViT), стало возможным их применение для решения задач компьютерного зрения. Архитектурно Vision Transformer повторяет кодировщик оригинальной модели Transformer. Основное отличие заключается в способе преобразования исходных данных в набор векторов. При решении задач обработки естественного языка данный процесс выполняется посредством разбиения единого текста на последовательность частей слов и иных символов – токенов. После каждый токен соотносится с определенным вектором с помощью специальной обучаемой матрицы. Этот процесс называется токенизацией. В Vision Transformer она реализуется через разбиение изображения на фиксированные по размеру части – патчи (чаще всего, 16 на 16 пикселей). После этого, каждый патч выпрямляется в вектор и проецируется в пространство определенной размерности.

Базовый метод токенизации изображений Vision Transformer, хотя и прост в реализации, имеет существенные ограничения. Во-первых, фиксированный размер патчей приводит к избыточной длине входной последовательности, что может быть критичным при обработке изображений высокого разрешения или нескольких изображений одновременно. Во-вторых, статичная сетка разбиения игнорирует структуры объектов на изображении и их семантическую важность, а также негативно сказывается на понимании нейронной сетью локального контекста. Эти ограничения делают актуальным поиск аль-

тернативных методов токенизации, которые сохранят преимущества моделей Transformer, но при этом обеспечат гибкость и вычислительную эффективность.

В данной работе исследуется применение матричных разложений для создания адаптивного подхода преобразования изображений в последовательность токенов. Данный подход должен позволить уменьшить длину входной последовательности (длину контекста) без потери критически важной информации, а также сохранить или улучшить способность модели к пониманию как глобальных, так и локальных визуальных признаков.

Основная идея – использовать матричные разложения для получения матриц меньшего порядка. В частности, рассматриваются такие методы как:

а) Сингулярное разложение – 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.

б) Преобразование Фурье – преобразование, при котором сигнал (изображение в частности) раскладывается на гармонические составляющие. Данный метод позволяет уменьшить длину входной последовательности, пожертвовав малоинформативными высокочастотными признаками.

Для оценки качества работы, рассмотренные методы токенизации сравниваются с базовым методом токенизации ViT на задаче классификации изображений. Также, для общей оценки методов, рассматривается их применение в других задачах, в том числе и авторегрессионной генерации изображений.