Разработка методов кеширования диффузионных трансформеров для генерации изображений/видео

Руководитель
Соболев Константин Викторович

Закончил аспирантуру CDSE Сколтеха в 2022 году, занимался исследованиями по теме разработки эффективных нейросетевых архитектур. Во время обучения участвовал в нескольких крупных индустриальных проектах с Huawei, а также в международном исследовательском проекте совместно с университетом Sharjah, ОАЭ. За период аспирантуры опубликовал ряд работ на конференциях Core A* и в журналах Q1. С 2022 по 2024 года работал в Samsung AI Center Moscow на позиции ведущего инженера, фокусируясь на разработке моделей генерации изображений и видео. В частности, разработал диффузионную модель ViGen, способную генерировать Full HD видео анимированных пейзажей, модель генерации последовательности ключевых точек для создания анимированных лиц и модель редактирования эмоций и переноса домена портретов людей. На основе этих разработок были получены патенты A1 и A2. С 2024 года возглавляет группу Video Generative AI в лаборатории Fusion Brain института AIRI. Команда занимается исследованиями в области диффузионных моделей генерации изображений и видео, а также участвует в разработке диффузионных моделей генерации изображений и видео линейки Kandinsky.

Зачем делаем?
Diffusion Transformer (DiT) — современная передовая архитектура (SoTA) для генерации изображений и видео. Однако процесс инференса в таких моделях требует значительных вычислительных ресурсов, что ограничивает их применимость в условиях ограничений по времени или ресурсам.

Одним из перспективных подходов для решения этой проблемы является использование кэширования, которое позволяет переиспользовать данные, пересчитываемые на каждом шаге в процессе итераций инференса. Снижение избыточных вычислений может значительно ускорить работу модели без значительной потери качества сгенерированного контента.

Цель проекта: провести анализ архитектуры DiT и существующих методов кэширования, а затем разработать новый подход, который позволит существенно снизить вычислительные затраты при генерации изображений и видео, сохранив высокое качество результатов.

Что делаем?
Процесс генерации в диффузионных трансформерах базируется на итерационном процессе обратной диффузии. Недавние исследования показали, что внутренние представления модели между соседними итерациями изменяются незначительно. Это открывает возможность повторного использования данных с предыдущих шагов и исключения пересчета дублирующихся операций.

В рамках проекта:
1. Разрабатываем новый метод кэширования внутренних представлений диффузионных трансформеров между итерациями инференса.
2. Сосредотачиваемся на выявлении избыточных вычислений и определяем представления, которые можно либо переиспользовать, либо исключить.
3. Оптимизируем процесс инференса, снижая количество операций без ущерба для качества генерации.

Проект нацелен на достижение более высокой эффективности по сравнению с существующими методами.

Как достигнем цель?

1. Анализ архитектуры современных моделей генерации:
    a. Изучаем принцип работы ведущих моделей, таких как Flux (Text-to-Image, T2I) и Wan-2.1 (Text-to-Video, T2V).
    b. Изучаем процесс итеративной диффузии и выявляем избыточные вычисления.
2. Поиск точек оптимизации:
    a. Анализируем внутренние представления модели между итерациями инференса.
    b. Ищем представления, которые можно кэшировать или исключить из пересчета.
3. Разработка нового метода:
    a. На основе полученных данных создаем более эффективный алгоритм кэширования и интегрируем его в процесс инференса.
4. Тестирование и валидация:
    a. Проверяем предложенные изменения на современных моделях.
    b. Оцениваем влияние на ускорение вычислений и качество результатов.

Как измерить достижение результата?

Для оценки эффективности предложенного подхода будут использоваться следующие метрики:

1. *Эффективность модели*: MACs reduction = MACs(original models) / MACs(cached model) (MACs - количество multiply-add операций необходимых для инференса) и Speedup (соотношение времени инференса оригинальной модели к ускоренной)
2. *Качество генерации модели*:
    a. Для T2I моделей: FID-30k и Text CLIP-SIM на MS-COCO2017
    b. Для T2V моделей: VBench Total Score

Целевые метрики:

Добиться как минимум 2-кратного увеличения эффективности вычислений (по метрикам MACs reduction и Speedup) при сохранении падения качества результатов на уровне не более 1% по ключевым метрикам.