Training-free повышение качества 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) является SoTA архитектурой для генерации изображений и видео. Однако, существующие диффузионные модели зачастую обладают недостаточно высоким качеством результатов в отдельных задачах.

Цель проекта:
Исследование архитектуры DiT, а также разработка новых подходов повышения качества генерации, не требующих дополнительного обучения модели (training-free methods). Мы фокусируемся на анализе текущих методов и создании оригинального подхода, чтобы увеличить качество генерации изображений и видео.

Что делаем?
Недавние исследования показали, что некоторые блоки архитектуры диффузионных трансформеров оказывают минимальное влияние на процесс генерации, тогда как другие играют критическую роль. Это открывает новые возможности для оптимизации, не требующей переобучения модели.

Как достигнем цель?
1. Анализ современных архитектур:
    a. Изучение ведущих моделей генерации, включая Flux / CogView 4 (Text-to-Image, T2I) и Wan-2.1 (Text-to-Video, T2V).
    b. Определение вклада отдельных слоев в итоговое качество генерации.
2. Поиск точек оптимизации:
    a. Исследование методов улучшения или ослабления моделей путем исключения отдельных слоев из процесса инференса..
    b. Исследуем подходы улучшения/ослабления модели путем исключения различных слоев из процесса инференса.
    c. Изучаем подход Autoguidance - метод улучшения качества генерации с помощью гайденса ослабленной моделью.

3. Эксперименты с Autoguidance:
    a. Анализ подхода Autoguidance, основанного на взаимодействии между "ослабленной" и "усиленной" версиями модели.
    b. Изучение возможности адаптации данного метода, чтобы улучшить качество генерации без дообучения.
4. Разработка нового метода:
    a. На основе полученных данных создаем более эффективный алгоритм инференса - Autoguidance с усиленной и ослабленной версиями модели.
5. Тестирование и валидация:
    a. Проведение экспериментов на современных T2I и T2V моделях.
    b. Оценка изменений в качестве генерации через сравнительный анализ предложенного метода с альтернативными решениями.

Как измерить достижение результата?
Для оценки эффективности предложенного подхода будут использоваться следующие метрики *качества генерации модели*:

1. Для T2I моделей: FID-30k и Text CLIP-SIM на MS-COCO 2017
2. Для T2V моделей: VBench Total Score

Целевые метрики:
Повышение качества генерации не менее чем на 5% по ключевым метрикам (FID, CLIP-SIM, VBench), без дополнительного обучения моделей.