Открытый код проекта доступен на GitHub, а исследования — на Hugging Face и Model Scope.

Инженеры из китайской компании Alibaba представили новую мультимодальную модель машинного обучения под названием mPLUG-Owl3. Эта модель способна эффективно анализировать текст, изображения и видео. Разработчики уделяют особое внимание скорости работы нейросети, утверждая, что на обработку двухчасового видео требуется всего четыре секунды.

mPLUG-Owl3 базируется на модели Qwen2, которая была существенно доработана и оптимизирована. Благодаря этим изменениям время ожидания первого токена сократилось в шесть раз, а одна видеокарта A100 теперь может обрабатывать до 400 изображений в секунду. Также в модели был использован специальный блок HATB (Hyper Attention Transformer), который интегрирует визуальные и текстовые признаки, позволяя, например, искать визуальные образы на основе текстовых запросов.

Код проекта открыт и размещён на GitHub. Дополнительно разработчики предоставили все необходимые материалы для работы на платформах Hugging Face и китайском аналоге Model Scope . В полном тексте исследования подробно описан процесс разработки и работы модели mPLUG-Owl3.

На перекрестке науки и фантазии — наш канал