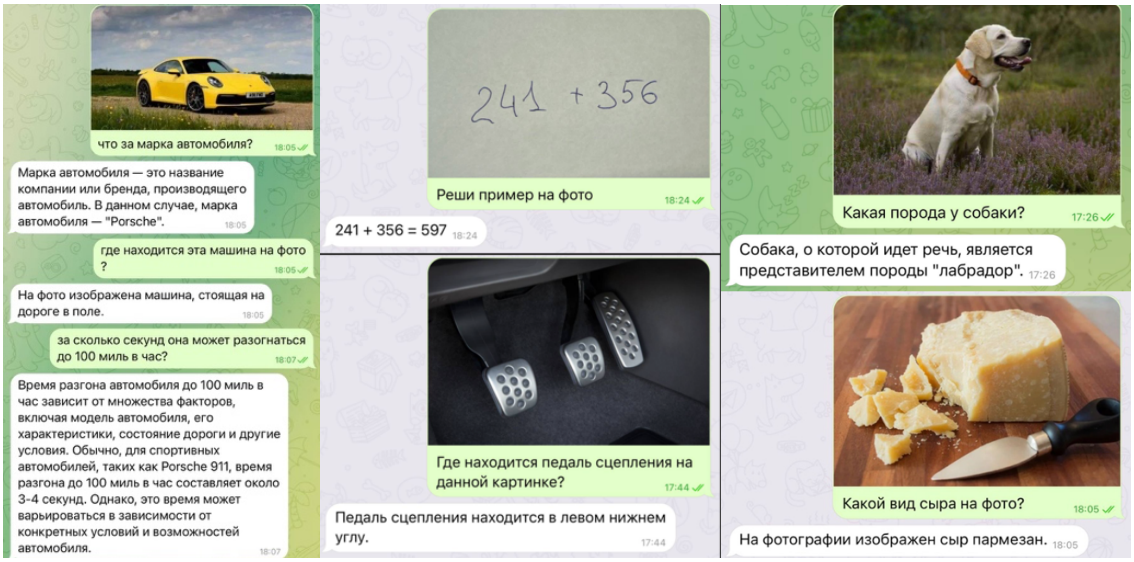
Первая в России мультимодальная модель искусственного интеллекта — в открытом доступе

Институт искусственного интеллекта AIRI представил открытую версию модели OmniFusion – OmniFusion 1.1. Первая в России мультимодальная языковая модель, способная поддерживать визуальный диалог и отвечать на вопросы пользователей по картинкам, теперь поддерживает русский язык. Open-source-код для обучения и веса доступны к использованию и могут быть применены в том числе при разработке коммерческих продуктов.

OmniFusion — это передовая мультимодальная модель искусственного интеллекта, предназначенная для расширения возможностей традиционных систем обработки языка за счет интеграции дополнительных модальностей данных, например, изображений, а в перспективе — аудио, 3D- и видеоконтента.

Модель распознает и описывает изображения. С ее помощью можно объяснить, что изображено на фото, узнать рецепт для приготовления блюда по фотографии ингредиентов, проанализировать карту помещения или узнать, как собрать устройство по фото отдельных его частей. Модель также умеет распознавать текст и решать задачи. Например, с её помощью можно решить логические задачи, написанный на доске математический пример или распознать формулу, а также получить их представления в формате LaTeX. Спектр возможностей широкий: уже сейчас модель может проанализировать медицинское изображение и указать на нем какую-то проблему. Разумеется, для того, чтобы подобная модель помогала ставить диагнозы, ее необходимо дополнительно обучать на профильных датасетах с привлечением экспертов из медицины.



OmniFusion — это первая в России мультимодальная модель. Среди зарубежных аналогов на рынке представлены, например, LLaVA, Gemini, GPT4-Vision, а также китайские модели Qwen, DeepSeek и LVIS. Часть из этих моделей относится к числу проприетарных, то есть находится в закрытом доступе, и судить о метриках качества таких моделей можно только на основе опубликованных компаниями цифр или посредством платных API. GPT4-Vision и Gemini уже встроены в продуктовую линейку чат-ботов от OpenAI и Google. В отличие от платных моделей, среди open-source решений можно также найти достойные аналоги, такие как LLaVA и Multimodal-GPT.

Всего качество модели в разных вариантах её архитектуры оценили при помощи 8 известных бенчмарков — специализированных тестов для анализа работоспособности AI-моделей в ответах на визуальные вопросы. В науке этот тип задач называется VQA, или Visual Question Answering.

Среди них, например, были проведены тесты на TextVQA — бенчмарке для оценки качества ответов на вопросы по изображениям, содержащим какой-то текст, POPE — бенчмарке для оценки галлюцинаций (когда модель начинает выдумывать несуществующие данные в ответах), а также ScienceQA — бенчмарке с вопросами, основанными на лекциях и вопросах на различные научные темы.

Эксперименты по оценке качества показали: OmniFusion достигает высоких результатов в большинстве бенчмарков, не уступая зарубежным моделям, которые в том числе построены на более крупных языковых моделях (например, LLaVA-13B). Следует отметить, что для таких известных бенчмарков как MMMU, GQA и TextVQA, модель OmniFusion показывает лучшие результаты в сравнении с LLaVA-7B и LLaVA-13B.

В основе архитектуры модели лежит методика совмещения предварительно обученной большой языковой модели и ее «глаз» — визуальных энкодеров, которые позволяют кодировать информацию на изображении в числовой вектор, называемый эмбеддингом. Обучением OmniFusion занимается научная группа FusionBrain Института AIRI при участии учёных из Sber AI и SberDevices.

Открытый исходный код и веса модели можно найти по ссылке https://github.com/AIRI-Institute/OmniFusion

«Публикуя открытый исходный код OmniFusion, включая веса модели и скрипты для обучения, мы стремимся внести вклад в сообщество исследователей искусственного интеллекта и поспособствовать дальнейшему развитию мультимодальных архитектур, созданию новых приложений на их основе. Кроме того, мы уже начали эксперименты, которые помогут обучить ее понимать видео и 3D-контент. Наш коллектив также активно сотрудничает с коллегами-учеными из области медицины. Надеемся, что в будущем эти изыскания приведут к созданию принципиально новых инструментов для помощи врачам»

Иван Оселедец, доктор физико-математических наук, Профессор РАН, генеральный директор Института AIRI

Вопросы: pr@airi.net

Научно-исследовательский Институт искусственного интеллекта AIRI — автономная некоммерческая организация, занимающаяся фундаментальными и прикладными исследованиями в области искусственного интеллекта. На сегодняшний день более 150 научных сотрудников AIRI задействовано в исследовательских проектах Института для работы совместно с глобальным сообществом разработчиков, академическими и индустриальными партнерами.