Разработан метод оценки эффективности нейросетей в работе с длинными текстами

Он будет представлен в Ванкувере на одной из крупнейших мировых ИИ-конференций.

Исследователи из Института AIRI, МФТИ и Лондонского института математических наук (LIMS) создали бенчмарк BABILong — инструмент для оценки реальной производительности языковых моделей при работе с большими объемами данных. Он включает 20 задач, которые требуют поиска и обработки разрозненных фактов в крупных текстах. Среди них — связывание и комбинирование информации из нескольких фактов, индукция, дедукция, простейший подсчет и работа со списками и множествами. BABILong выложен в публичный репозиторий для поддержки научного сообщества, а также будет представлен на конференции NeurIPS 2024 в Ванкувере.

Длина контекста — это объем информации, которую нейросеть удерживает в уме для решения конкретной задачи. Чем она выше, тем потенциально лучше результат работы модели. Несмотря на то, что этот параметр растет, фактически популярные модели используют лишь 10-20% данных, преимущественно фокусируясь на информации из первых и последних абзацев. Кроме того, производительность моделей резко снижается с увеличением сложности задач.

Созданный учеными бенчмарк оценивает две метрики: качество ответа и зависимость точности от длины контекста. В основу BABILong легли задачи из датасета BABI — 20 ключевых операций, направленных на демонстрацию понимания базовой логики и арифметики. Второй частью обновленного датасета стали массивы данных художественной литературы. Далее задачи, изначально рассчитанные на понимание коротких текстов, как «иголки в стоге сена» были раскиданы по литературным произведениям, требуя от моделей не просто найти нужную информацию, а выполнить ее анализ для получения правильного ответа.

В ходе экспериментов исследователи применили бенчмарк для анализа популярных open-source моделей при различных длинах контекста. Нейросети оперируют токенами — это базовые единицы текста, которые, как правило, представляют собой несколько символов, часть слова. Так, в рамках исследования команда провела анализ эффективности нейросетей в задачах с контекстом от тысячи до 50 миллионов токенов. Результаты показали, что производительность моделей сильно падает, когда объем данных превышает 25% от заявленной длины контекста. Это подчеркивает необходимость улучшения механизмов обработки контекстной информации.

Ученые также представили адаптацию бенчмарка BABILong для русского языка — Libra, разработанную в сотрудничестве с командой R&D SberDevices. Как и оригинал, Libra тестирует языковые модели на длинных контекстах, предлагая аналогичные задачи для оценки их работы с русскоязычными текстами.

«Разработка BABILong — это важный шаг в оценке реальной эффективности языковых моделей. Бенчмарк не только позволяет сравнивать корректность работы моделей на разной длине контекста, но и служит индикатором их качества, что демонстрирует, в каких аспектах требуется улучшение. Это значительно поможет разработчикам новых моделей», — подчеркнул **Юрий Куратов, кандидат физико-математических наук, кандидат физико-математических наук, руководитель группы** «**Модели с памятью**» **лаборатории** «**Когнитивные системы ИИ**» **Института AIRI**

……

Вопросы: pr@airi.net

Институт *AIRI — автономная некоммерческая организация, занимающаяся фундаментальными и прикладными исследованиями в области искусственного интеллекта. На сегодняшний день более 180 научных сотрудников AIRI задействовано в исследовательских проектах Института для работы совместно с глобальным сообществом разработчиков, академическими и индустриальными партнерами.*