https://secrets.tbank.ru/blogi-kompanij/iskusstvennyj-intellekt-iz-oblaka/?utm_referrer=https%3A%2F%2Fyandex.ru%2F

Василий Степаненко

20.12.2024

Облака стали не только подходом к использованию ИТ. Часто облачные платформы позволяют заказчикам получить доступ к технологиям, которые им еще недоступны в рамках собственных инфраструктур. И здесь на одном из первых мест в перечне находятся технологии искусственного интеллекта.

Как все начиналось

О своей приверженности «курсу на ИИ» уже несколько лет периодически заявляют все крупнейшие облачные провайдеры: от Amazon, Google и Microsoft до локальных игроков.

В России о работе в этом направлении в своих облаках еще в 2018 году объявил VK (тогда — Mail.ru Cloud Solutions). Тогда компания запустила сервис Cloud Machine Learning, предназначенный для разработки решений на основе машинного обучения.

Годом позже облачной платформой обзавелся «Сбер», который декларировал, что на ее базе будут предоставляться и ИИ-сервисы. Еще через год в России появились полноценные платформы машинного обучения, от тех же «Сбера» и «Яндекса».

С тех пор новости о российском искусственном интеллекте появляются постоянно. Целый ряд сервисов на базе ИИ предлагает банкам и ритейлу компания oneFactor (входит в состав «МегаФона»), «Сбер» открыл клиентам возможность встраивать в их приложения свой ИИ, платформа машинного обучения создается «Ростелекомом».

Действительно, при всем скепсисе заказчиков в отношении облаков, искусственный интеллект остается одной из тех востребованных технологий, ради которых они готовы идти наперекор своим предубеждениям.

Зачем нужна связка облака и ИИ

В проблематике совместного использования систем искусственного интеллекта и облачных сервисов выделяются несколько аспектов.

Первый — возможность обработки в облаке данных, отобранных при помощи ИИ-моделей. Здесь в качестве примера можно привести схему работы камер наблюдения, использующих механизмы ИИ. Они, в зависимости от запроса, могут анализировать входящий видеопоток, выделять в нем фрагменты, которые удовлетворяют условиям (например, найти человека в красной куртке), вырезать их и затем отправлять для последующего глубокого анализа в облако.

Благодаря такой первоначальной обработке, проведенной ИИ, облачной аналитической системе нет нужды задействовать значительные ресурсы для анализа данных огромного объема — останется только проанализировать отобранные фрагменты и детально понять, насколько выбранное камерой видео способно удовлетворить поисковому запросу.

Приведенный пример иллюстрирует схему, когда искусственный интеллект используется для пограничных вычислений (обработки данных непосредственно на устройстве). Но для того, чтобы такой сценарий совместного использования ИИ и облака был реализуем, необходимо провести предварительное обучение модели на множестве примеров — так же, как аналитические ИИ-системы в медицине сегодня учатся отбирать снимки МРТ.

Проводить такое обучение необходимо непосредственно в облаке по «накопительному» принципу. В таком случае ИИ-модель (возвращаясь к примеру с камерами видеонаблюдения) постепенно приобретает умение определять цвета, тип объекта — и так далее, вплоть до анализа мимики и считывания эмоций.

Карты такого обучения могут составляться и на корпоративном уровне, — в том случае, если компания обладает необходимым массивом данных, ресурсами, собственным облаком. Они могут и поставляться как сервис облачным провайдером, который дает возможность заказчикам обучать их собственные ИИ-системы на базе своих облачных систем.

Такое использование облаков, в качестве ML-базы для систем искусственного интеллекта — второй аспект взаимодействия облака и ИИ. Этот сценарий позволяет оптимизировать, в первую очередь, экономику «облачного ИИ» за счет централизации развития моделей или совместного использования многими компаниями ресурсов одного провайдера.

Специфические российские проблемы

В России сегодня существуют только два проекта, развивающих генеративный искусственный интеллект, — Yandex GPT и GigaChat от «Сбера». Эти модели хорошо развиваются, знают русский язык, имеют хорошую аппаратную основу. Все другие системы так или иначе используют возможности этих моделей.

Преимущество GPT-моделей (а как раз к ним и относятся российские ИИ-проекты) — в том, что они получают значительный прирост производительности в результате разработки и выпуска новых GPU, — те проектируются сегодня в расчете на использование именно с GPT. Результат — ускоренное развитие таких моделей. Но актуально это, увы, для иностранных проектов. Наши, российские, продолжают использовать уже устаревшее оборудование и это обуславливает постепенное, но постоянное отставание российских систем.

Второе обстоятельство — постепенное угасание интереса инвесторов к технологиям искусственного интеллекта. Ведущие мировые технологические корпорации неоднократно заявляли о своем интересе к ИИ, о своих разработках в этой области. Но у инвесторов свой интерес. Они уже представляют себе, какие инвестиции могут потребоваться для развития ИИ, но слабо представляют, какой экономический эффект он может дать, какую прибыль принести. Очевидно, что для получения прибыли от ИИ через десять лет, сделать инвестиции нужно уже сегодня.

Эти вложения необходимы не только для закупок постоянно обновляющегося оборудования, но и для найма специалистов, которые разрабатывают ИИ-системы, дата-инженеров. И это — очень серьезные капитальные затраты. Выжидательная позиция инвесторов сдерживает инвестиции, и это становится серьезной проблемой для корпораций, развивающих свои ИИ-проекты.

В России существует еще одна проблема, связанная с оборудованием. Мы не можем приобрести не столько сами новейшие карты NVIDIA, сколько лицензии на софт, который необходим для их работы. Поэтому для работы российских GigaChat и YandexGPT используются совсем не новые карты на GPU H100. К слову, виноваты в том, что такая ситуация сложилась, не только санкции против России.

В первую очередь это следствие стремления США ограничить развитие систем ИИ в Китае. Хорошей новостью здесь является то, что дефицит аппаратных средств ограничивает возможности развития только GPT. Нейронные сети можно обучать и на «старом» оборудовании.

Дефицит современных видеокарт усугубляется особенными требованиями к проектированию серверов, в которых такие карты должны использоваться. Тот же «Сбер» в свое время покупал сервера, рассчитанные на использование восьми мощных карт A100. А это — весьма специфическое оборудование, которое должно быть спроектировано с учетом необходимости правильно питать и охлаждать этот массив карт.

Производителей таких серверов немного. У нас же, как известно, сегодня на рынке представлены, по сути, только российские вендоры серверов, которые соответствующих компетенций, а главное, производственных возможностей, не имеют.

Есть и еще одна проблема, характерная для России. Связана она с отечественными системами виртуализации. Их очень много, но они не способны решить задачу, связанную с использованием видеокарт — разделение мощности видеокарты между виртуальными машинами. Они способны лишь на переброс карты.

Да, такая возможность имеется у VMware, но для этого необходимо решить множество проблем, связанных с драйверами, которые, в свою очередь, нерешаемы без технической поддержки вендора.

Кто использует искусственный интеллект в облаке

Эти специфические проблемы, связанные с искусственным интеллектом, оказывают значительное влияние на состав аудитории облачных ИИ-проектов. В первую очередь ее составляют те компании, которые так или иначе стремятся использовать нейронные сети в работе своих систем и сервисов. Это могут быть компании, которые работают буквально в любой сфере, от производства и добычи полезных ископаемых до маркетинга и дизайна.

Значительное количество таких компаний работают на проектной основе, когда ресурсы используются не постоянно, а необходимы временно. В таких условиях нейронные сети в сочетании с облачными сервисами позволяют принципиально сократить издержки, связанные и с затратами на оборудование и средства разработки. В качестве примера можно привести сервисную компанию, которая ведет разработки трехмерных моделей для производственных компаний.

Подобная организация всегда стремится к тому, чтобы максимально сократить объем капитальных затрат. То же самое можно сказать о стартапах, которые разрабатывают MVP своего продукта. В качестве примера можно привести стартап, разрабатывающий аналитические системы для ритейлеров, которым необходимо анализировать поведение покупателей, состав полок в магазине, оборачиваемость товара и т.п.

В момент перехода от MVP к следующим стадиям возникает необходимость в масштабировании, и здесь возможности искусственного интеллекта, используемого в разработке, прекрасно сочетаются с таким качеством облака, как простая масштабируемость.

Резюмируем

Потребитель ИИ-облаков — компания, которая может работать в любой отрасли, использует инструменты ИИ и нуждается в их простом масштабировании. При этом тип используемого облака может быть любым, в том числе частным или гибридным. Но решить задачу масштабирования ИИ публичное облако может быстрее, проще и несравненно дешевле.

На что обратить внимание заказчикам

Технические особенности нейронных сетей и ИИ-платформ определяют круг вопросов, на которые необходимо обратить внимание компании, которая стремится к использованию облачного ИИ.

В первую очередь это совместимость аппаратной части нейронной сети, все тех же видеокарт, которые используются для ее работы. Если ее нет, то, скорее всего, заказчику придется серьезно перекроить программный код.

Второй объект внимания — SLA, в первую очередь, в той части, которая описывает доступность и производительность систем облачного провайдера. Конечно, для ИИ-облака уровень SLA должен быть максимально высоким.

Наконец, стоит обращать внимание на то, какие дополнительные, сопутствующие сервисы может предложить облачный провайдер. В качестве примера можно назвать Kubernetes, который практически всегда необходим компаниям-разработчикам.

Технологии искусственного интеллекта могут буквально преобразить бизнес. Но их использование — серьезная технологическая задача, решить которую может не каждый заказчик даже класса enterprise. Даже если компания не доверяет облакам, ИИ может стать целью, ради которой стоит уйти в облако, пусть даже не полностью. Ради технологии, которая может не простить опоздания.