(планируется выпуск целой серии продуктов в период с 2017 по 2020 г.) ускорить тренинг сеток для машинного обучения до 100 раз (рис. 7).

Для того чтобы существующее ПО (в частности, например, работающее на базе фреймворка neon – одного из самых быстрых фреймворков для развития моделей для глубокого обучения и представляющего собой набор библиотек с открытым исходным кодом на языке Python; Neon развивается Intel) было оптимизировано под продуктовое семейство Intel® Nervana™, Intel разрабатывает отдельный слой ПО Intel® Nervana™ Graph (появление на рынке – 2017 г.), который будет представлять собой API для функций/моделей/решений глубокого обучения и даст возможность масштабирования алгоритмов на сотни машин (рис. 8).

На рис. 9 представлен весь портфолио – инструментарий, фреймворки, библиотеки и аппаратные компоненты, развиваемые Intel для решений AI.

Дополнительные возможности

Уровень Big Data видится Intel в качестве одного из более низких, откуда грузятся данные для AI-уровня и который служит для аккумуляции необходимых данных из множества источников (как внешних, например, публичных облаков, так и внутренних). Однако hadoop-кластеры также могут использоваться для глубокого обучения в тех случаях, когда требования к производительности не так высоки или когда алгоритмы адаптированы для архитектуры hadoop (с ориентацией на Hive, HBase и др.). Реализация машинного обучения на hadoop-кластерах обеспечивается Intel за счет поддержки фреймворка Spark с библиотеками BigDL и MLib (рис. 10).

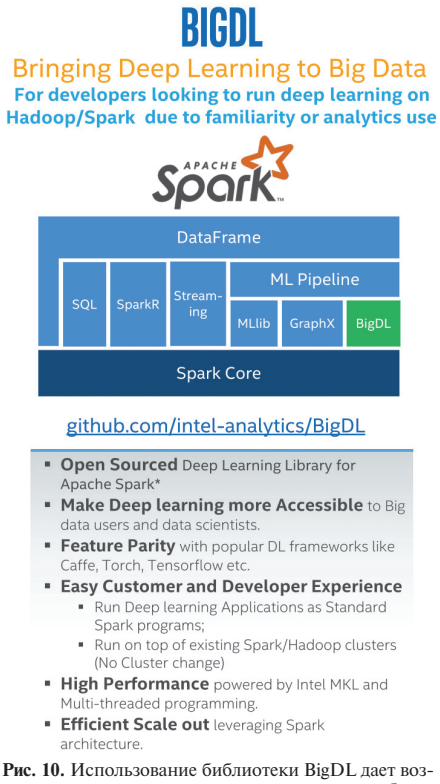

Рис. 10. Использование библиотеки BigDL дает возможность выполнения алгоритмов глубокого обучения непосредственно на hadoop-кластерах.

Cisco: машинное обучение для ИБ

Алексей Лукацкий – менеджер по развитию бизнеса, Cisco.

В 2000-м году, когда я писал свою первую книгу "Обнаружение атак", я посвятил много времени изучению не только традиционных подходов по обнаружению вторжений по сигнатурам или аномалиям, но и многим другим методам, среди которых были и нейросети, призванные автоматизировать процесс принятия решения по событиям безопасности, подаваемым на вход нейросети. Однако на тот момент времени основным препятствием на победоносном пути нейросетей была нехватка обучающего материала. Сегодня, в век Больших Данных (пусть и мало кто понимает, что это такое на самом деле), ситуация стала кардинально иной – данных для анализа (в том числе и с точки зрения безопасности) стало настолько много, что прежние подходы, те же сигнатуры, начинают потихоньку сдавать свои позиции. Неслучайно, еще в начале 2000-х годов исследовательская лаборатория HP обосновала скорый закат традиционных сигнатурных антивирусов, которые подойдут к своему потолку возможностей – нельзя будет хранить всю базу сигнатур на защищаемом устройстве. В качестве примера приведу слова Евгения Касперского, который в своем блоге пишет, что антивирусная база на компьютере содержит всего около 5 миллионов записей, в то время как только два миллиона новых записей в общую, облачную базу добавляется еженедельно (к слову сказать, в базу вредоносных программ Cisco такое количество добавляется ежедневно), а вся облачная база "Лаборатории Касперского" в декабре 2016 года насчитывала около миллиарда записей.

Учитывая такой рост вредоносных программ, уже можно говорить о достаточном объеме информации для анализа с помощью специальных алгоритмов, которые в последний год-два так активно стали упоминаться при описании различных защитных технологий под названиями "машинное обучение", "нейросети", "искусственный интеллект". Когда в 2012 году Cisco покупала компанию Cognitive Security, мало кто еще понимал, что это такое и "с чем едят" ту математику, которая лежала в основе анализа огромных объемов Web-логов, которые и анализировались с помощью технологий Cognitive Security. Сегодня технологии машинного обучения применяются не только в Cisco Cognitive Threat Analytics, но и во многих других решениях Cisco по ИБ. Спустя 4 года термин "Cognitive Security" стала использовать компания IBM, рассказывая о "новой эре" информационной безопасности, а различные "умные" технологии прочно обосновались в портфолио многих вендоров по ИБ. Хотя и тут надо признать, что первые примеры машинного обучения в информационной безопасности стали применяться гораздо раньше. Например, в Cisco Talos (еще когда это исследовательское подразделение называлось иначе) с помощью такой математики автоматизировались многие рутинные задачи аналитиков – классификация вредоносной активности, категоризация URL с выделением из них опасных и подозрительных, разбор почтовых миллиардов сообщений (а сегодня мы анализируем уже 600 миллиардов сообщений e-mail в сутки) с точки зрения анализа действий спамеров и т.п.

Первоначально речь шла просто об анализе и классификации Больших Данных, затем, по мере развития технологий, Cisco стала сопоставлять URL с другой получаемой информацией – Интернет-доменами, контрольными суммами (хешами) вредоносных файлов, IP-адресами и автономными системами. Это позволило выявлять более сложную и многовекторную вредоносную активность и при этом делать это практически в реальном времени. Раньше для классификации кого-либо сайта надо было дождаться вредоносной активности с него и внести в черный список, а сегодня вывод о вредоносности можно сделать по комбинации параметров (дата создания сайта, место его регистрации, IP-адреса, которые скрываются за доменом, автономные системы, владелец домена и т.п.). Дальше, опираясь на «схожие аналоги», делается вывод о вредоносности, который редко дает сбой. А учитывая, что часто такие сайты являются однодневками и срок их жизни составляет всего несколько часов, а используются в хакерских кампаниях они всего несколько раз (а не тысячи и десятки тысяч, как раньше), только с помощью машинного обучения и анализа Big Data становится возможным нейтрализовать такие угрозы – никакие «белые» и «черные» списки тут не помогут.

Я не буду сейчас вдаваться в дискуссию о том, чем отличается нейросеть от машинного обучения и корректно ли делать ссылки на искусственный интеллект применительно к ИБ, но факт остается фактом – на последних мировых конференциях RSAC и InfoSecurity Europe эти термины эксплуатировались всеми, кому не лень. И это не то, чтобы дань моде (хотя и такое тоже бывает). Просто технологии действительно достигли такого уровня, чтобы снять нагрузку с человека, принимающего

решения. Возьмем совсем недавнюю новость о том, что программа победила пилота ВВС США в воздушном бою. В переводе говорится об искусственном интеллекте, хотя в самих результатах упоминается немного иной математический аппарат – и нечеткая логика, и генетические алгоритмы. Но суть не в терминах, а в том, что то, что еще совсем недавно считалось невозможным – компьютер обыграл человека в пусть и учебном, но все-таки бою.

Читали ли вы роман Сергея Лукьяненко "Лабиринт отражений"? В нем (а раньше аналогичная идея была описана Гиббсоном в его известнейшем романе "Нейромансер") приводится пример системы защиты киберпространства, которая действует сама, адаптируясь к атакам, на нее направленным. С увеличением числа и сложности атак, растет мощность и системы защиты, тем самым превращая попытку проникновения (и защиты) в бой с непредсказуемым результатом, зависящим от того, у кого (атакующих или обороняющихся) лучше алгоритм самообучения. Раньше я думал, что это фантастика и при мне таких технологий не появится. Но сейчас я понимаю, что, возможно, я ошибался. Я вижу то, что делается у нас в Cisco, и какие исследования мы проводим в области новых технологий ИБ. Я примерно представляю, что происходит у других игроков рынка ИБ. Я слежу за ИБ-стартапами. В США (а где же еще?) стала проводиться ежегодная специализированная конференция по применению искусственного интеллекта в ИБ – AICS.

Я понимаю, что сегодня многие ринулись в эту сферу и скоро картина используемых в ИБ технологий сильно преобразится (возможно, это даст даже перевес в борьбе со злоумышленниками, которые пока не замечены в активном поиске научных исследований на эту тему). Например, недавно специалисты из лаборатории компьютерных наук и искусственного интеллекта Массачусетского технологического института и стартапа PatternEx представили платформу AI2, построенную как раз на данном подходе, которая позволяет обнаруживать до 85% атак, не имея никакой базы сигнатур (кстати, как сертифицировать такие решения, не имеющие базы решающих правил, не совсем понятно).

Но независимо от конкретного математического аппарата почти все такие системы базируются на анализе большого числа, в том числе разрозненных и неструктурированных данных, моделирующими человеческий процесс принятия решения, но делающие это более быстро. И чем больше устройств в Интернете будет появляться, а с приходом Интернета вещей это уже становится проблемой, тем важнее станет активное использование машинного обучения и его интеграция с анализом Больших Данных. Основным же преимуществом различных технологий машинного обучения является способность к самообучению и исправлению ошибок. Полностью, конечно, ошибки исключить нельзя, но так и человек их тоже совершает и далеко не всегда извлекает уроки из них. Помимо анализа большого объема данных (а в ИБ их действительно много), технологии машинного обучения (читайте – нейросети, читайте – искусственные интеллект) также применяются в прогнозировании и принятии решений – в задачах, которые так важны и в кибербезопасности.

По сути можно говорить о том, что машинное обучение становится насущной необходимостью в информационной безопасности, а не блажью. Да, есть еще направления в ИБ, где искусственный интеллект пока не так востребован (например, в VPN или межсетевых экранах), но там, где имеет место динамический ландшафт угроз без современных математических методов уже не обойтись. Понятно, что позволить себе это могут только крупные игроки рынка кибербезопасности, которые, инвестируя немалые средства в такую аналитику, затем предоставляют ее труды своим клиентам, самостоятельно не способным выстроить мощнейшие ЦОДы, по всему миру с сотнями серверов, которые пропускают через себя огромные потоки данных. Такая «зависимость» от крупных разработчиков заставляет некоторых заказчиков нервничать, и они пытаются выстраивать собственные математические модели и «учить» их на получаемых откуда-то Больших Данных. Но этот путь доступен единицам (в России таких компаний всего 5–6). Но это не значит, что решения нет. Стали появляться так называемые технологии автономных агентов, которые построены по принципу пиринговых, децентрализованных сетей и обмениваются данными об угрозах между собой, минуя центр. Таким образом, например, работает Cisco Stealthwatch Learning Network, где сами машршрутизаторы обмениваются данными о сетевых аномалиях, а встроенные в каждое сетевое устройство алгоритмы машинного обучения учатся на этом трафике выявлять несанкционированную активность и потом ее осуществляют уже в «боевом» режиме.

Огорчает только, что российские игроки рынка ИБ тему машинного обучения пока почему-то обходят стороной. Возможно, это связано с нехваткой данных для обкатки технологий. Все-таки этот барьер преодолеть не так-то легко. Западным компаниям проще – у них доступ к гораздо большему объему данных, чем у россиян. Например, Cisco OpenDNS анализирует 90 миллиардов DNS-запросов ежедневно (!), а Cisco Talos пропускает через себя 16 миллиардов URL, 600 миллиардов сообщений e-mail и 18,5 миллиарда файлов в день (!). На таких объемах, действительно, проще отрабатывать новую математику. В России, пожалуй что только Касперский с его KSN может выполнить аналогичную задачу, а также "Яндекс", Mail.ru, Qiwi, "Сбербанк" и другие не ИБ-компании с большими потоками данных для анализа и своими разработчиками. Хотя все, конечно, зависит от конкретной задачи. Например, для анализа защищенности Web-сайтов есть весь Интернет, а для обнаружения атак с помощью машинного обучения можно использовать различные конкурсы – те же CTF, в рамках которых собирать и анализировать методы злоумышленников. Было бы желание...

Tarantool IIoT – СУБД для интернета вещей

Февраль 2017 г. **–** Mail.Ru Group представила распределенную программную платформу Tarantool IIoT, разработанную для промышленного интернета вещей. Новый продукт позволит собирать данные с миллионов датчиков, расположенных на производственных площадках, транспорте, сельскохозяйственных полях – и пересылать в датацентры для онлайн-аналитики.

Передача данных в Tarantool IIoT (Industrial Internet of Things) осуществляется с помощью механизма репликации, предоставляемого СУБД Tarantool. Этот способ гарантирует надежную доставку данных даже в сложных случаях – например, когда для работы используется ненадежный интернет-канал, а для приема и пересылки информации в качестве IIoT-хаба применяются самые дешевые локальные миникомпьютеры.

В отличие от большинства промышленных СУБД, которые требовательны к объему дискового пространства, производительности дисков и памяти, количеству ядер на процессорах и работают медленно, Tarantool IIoT может устанавливаться даже на недорогие миникомпьютеры стоимостью $30–50 долл.

При этом СУБД Mail.Ru Group обеспечивает на этих недорогих устройствах высокую скорость работы – до 10–50 тысяч транзакций в секунду. Кроме того, продукт способен собирать информацию с миллионов датчиков и поддерживает популярные протоколы для работы с ними.

Разработка Mail.Ru Group может применяться в различных отраслях. Например, заводы, собирая данные с помощью Tarantool IIoT и анализируя их, смогут судить о техническом состоянии машин и агрегатов, предсказывать поломки и уменьшать время простоя, избегая, таким образом, многомиллионных финансовых потерь. Сельскохозяйственные организации смогут применять Tarantool IIoT для выявления порчи растений и своевременного реагирования. Продукт может также поставляться в крупные розничные сети, где на основе информации от датчиков движения и eye-tracking можно следить за траекторией движения и направлением взгляда покупателей. Это позволит оптимизировать расположение товаров на полках и пространство между стеллажами.

«Tarantool IIoT расширяет границы IT-ландшафта предприятия за пределы датацентров, на индустриальные площадки. Наша СУБД позволяет легко собирать информацию и доставлять ее в аналитические системы, даже если источники этой информации расположены локально на предприятиях и не поддерживают общепринятые интернет-протоколы, – говорит Денис Аникин, технический директор почтовых и облачных сервисов Mail.Ru Group. – Поскольку Tarantool IIoT – полностью программируемая и расширяемая платформа, построенная на решении open source, его легко кастомизировать в соответствии с потребностями бизнеса – а это, в свою очередь, позволяет снизить стоимость средств производства».