УДК 004.5:004.8:621.396

АНАЛИЗ ОТЕЧЕСТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ В СМАРТФОНАХ И ПЛАНШЕТАХ НА СЕТЯХ МОБИЛЬНОЙ СВЯЗИ

Шепелев С.В., Бабин А.И.

АО «МТУ Сатурн», Москва, e-mail: SSHepelev@k-tech.ru, ABabin@k-tech.ru

В представленной статье проведен анализ особенностей использования нейронных сетей в смартфонах и планшетах для сценариев и решений на основе технологий искусственного интеллекта и машинного обучения на сетях мобильной связи. Исследован мировой и отечественный уровень разработки нейронных процессоров, отечественные разработки нейронных сетей на смартфонах и планшетах. Применение искусственного интеллекта в ближайшие годы расширит сетевые возможности, пропускную способность и качество предоставления услуг мобильной связи. Благодаря появлению нейронных процессоров NPU для мобильных устройств они могут помочь понять запросы на естественном языке и ответить на них, получить соответствующую информацию из интернета и предоставить персонализированные предложения и рекомендации. При разработке отечественных приложений нейронных сетей для смартфонов (планшетов) необходимо уделять внимание взаимодействию их с отечественными операционными системами: «Аврора», Kaspersky OS, «Ред ОС М», «Роса Мобайл» отечественных смартфонов. В профессиональной сфере нейронные сети смартфонов и планшетов планируют участвовать в анализе данных, видеоаналитике, обработке естественных языков, в работе прогнозных и рекомендательных систем, сегментации и кластеризации данных, поиске аномалий и закономерностей в различных данных и генерации контента.

Ключевые слова: нейронные сети, нейронный процессор, ANN, NPU, приложения AI и ML

ANALYSIS OF DOMESTIC NEURAL NETWORKS IN SMARTPHONES AND TABLETS ON MOBILE NETWORKS

Shepelev S.V., Babin A.I.

Joint-stock company "MTU Saturn", Moscow, e-mail: SSHepelev@k-tech.ru, ABabin@k-tech.ru

The presented article analyzes the features of using neural networks in smartphones and tablets for scenarios and solutions based on artificial intelligence and machine learning technologies on mobile communication networks. The world and domestic level of development of neural processors, domestic developments of neural networks on smartphones and tablets are studied. The use of artificial intelligence in the coming years will expand network capabilities, bandwidth and quality of mobile communication services. Thanks to the advent of neural processors NPU for mobile devices, they can help understand and respond to natural language queries, get relevant information from the Internet and provide personalized suggestions and recommendations. When developing domestic neural network applications for smartphones (tablets), it is necessary to pay attention to their interaction with domestic operating systems: Aurora, Kaspersky OS, Red OS M, Rosa Mobile of domestic smartphones. In the professional field, neural networks of smartphones and tablets plan to participate in data analysis, video analytics, natural language processing predictive and recommendation systems, data segmentation and clustering, the search for anomalies and patterns in various data and content generation.

Keywords: neural networks, neural processor, ANN, NPU, AI and ML applications

Нейронные сети на персональных устройствах (смартфонах, планшетах, смарт-устройствах) применяют во множестве сфер. Наиболее популярные примеры – это общение голосовыми командами с Siri и Алисой, распознавание объектов и людей на фотографиях в облачных хранилищах, поиск, систематизация и обработка фотографий (размытие фона, изменение освещения и исправления дефектов) или таргетинг рекламы по запросам в интернете, онлайн-переводчики (транскрипция и перевод, создание титров и субтитров) и дополненная реальность. Кроме смартфонов и планшетов технология на основе искусственного интеллекта реализуется в смарт-устройствах (умные камеры, Интернет вещей, промышленная и городская безопасность), устройствах автономных машин (Advanced Driver-Assistance Systems, ADAS), беспилотных транспортных средствах, крупных пользовательских центрах (обработка фото- и видеоданных, текста, голоса, информационная безопасность).

Целью исследования является анализ применения различных нейронных сетей на смартфонах (планшетах), анализ мировых и отечественных разработок нейронных процессоров и нейронных приложений для смартфонов (планшетов), сделаны выводы о направлениях разработок мобильных приложений для технологий искусственных нейронных сетей (Artificial Neural Network, далее *ANN*) в России на смартфонах (планшетах).

Материалы и методы исследования

Применены методы исследования: теоретический (анализ, синтез, аналогия, обоб-

щение), монографический, цифровой анализ с применением приемов сравнения и др.

Нейронная сеть – это математическая модель, работающая по принципу человеческого мозга. Она обучается путем первичной обработки большого набора данных, не требуя написания отдельного кода под конкретную задачу [1]. Нейросети являются одним из способов машинного обучения, подраздела искусственного интеллекта (AI), и лежат в основе алгоритмов глубокого машинного обучения (ML) [2, с. 32].

Базовыми видами архитектур ANN являются: нейронная сеть Хопфилда (Hopfield Network, HN); цепи Маркова (Markov Chains, MC); машина Больцмана (Boltzmann Machine, BM); сеть типа «Deep Belief» (Deep Belief Networks, DBN).

Среди более сложных архитектур ANN: сверточные нейронные сети (Convolutional Neural Networks, CNN) и глубинные сверточные нейронные сети (Deep Convolutional Neural Networks, DCNN); глубинные сверточные обратные графические сети (Deep Convolutional Inverse Graphics Networks, DCIGN); генеративные состязательные сети (Generative Adversarial Networks, GAN); рекуррентные нейронные сети (Recurrent Neural Networks, RNN); нейронные машины Тьюринга (Neural Turing Machines, NTM) и многие другие [2, с. 113]. С развитием мобильных устройств и возросшим спросом на различные приложения, включающие использование ANN, оптимизация этих сетей для мобильных платформ стала одной из актуальных задач для разработчиков приложений. Нейронные сети могут быть невероятно полезными инструментами для мобильных устройств, но их эффективная работа требует определенной модификации и оптимизации.

Существует несколько методов и техник оптимизации ANN для мобильных устройств. Одним из них является *квантизация*, которая позволяет уменьшить размер нейронной сети и требуемую память для ее работы [3]. Это достигается путем снижения точности представления чисел в сети. За последние годы вычислительная мощность мобильных устройств, таких как смартфоны и планшеты, резко возросла, достигнув уровня настольных компьютеров, доступных не так давно. Хотя стандартные приложения для смартфонов больше не являются для них проблемой, по-прежнему существует группа задач, которые могут легко бросить вызов даже устройствам высокого класса, а именно запуск алгоритмов искусственного интеллекта.

Еще одним методом оптимизации является *сокращение (pruning)* ANN [3]. В этом случае нейроны, которые не имеют значительного вклада в работу сети, удаляются, что позволяет уменьшить ее размер и улучшить производительность. Сложные глубокие нейронные сети часто весят несколько гигабайт. При интеграции нейронной сети в мобильное программное обеспечение происходит некоторое сжатие, но этого все равно недостаточно для комфортной работы. Основная рекомендация разработчикам – максимально минимизировать размер приложения на любой мобильной платформе для улучшения пользовательского интерфейса.

Также стоит упомянуть о важности *выбора подходящей архитектуры ANN* для мобильных устройств. Некоторые архитектуры могут быть более подходящими для мобильной платформы, так как требуют меньше вычислительных ресурсов. В табл. 1 приведена структура и описание ANN на основе основных параметров и типов ANN.

Таблица 1

Структура и описание ANN мобильных приложений

Параметр	Типы	Описание
На основе шаблона подключения	Прямая связь, рекуррентные	Прямая связь – В графиках нет циклов. Рекуррентная – Циклы возникают из-за обратной связи
На основе количества скрытых слоев	Однослойные, многослойные	Однослойный – Имеющий один секретный слой. Например, одиночный персептрон. Многослойный – Имеющий несколько секретных слоев. Многослойный персептрон
На основе природы весов	Фиксированные, адаптивные	Фиксированный – Веса имеют фиксированный приоритет и не изменяются вообще. Адаптивный – обновляет веса и изменяется во время обучения
На основе блока памяти	Статические, динамические	Статический модуль без памяти. Текущий выходной сигнал зависит от текущего входного сигнала. Например, сеть прямой связи. Динамический блок памяти – выходные данные зависят как от текущего входного сигнала, так и от текущего выходного сигнала. Например, рекуррентная нейронная сеть

Таблица 2

Области применения мобильных приложений ANN

Область применения	Приложения ANN
1. Создание контента	Создание статей, блогов, постов в социальных сетях. Создание рекламных копий и маркетинговых материалов. Создание стихов, рассказов и другой творческой литературы
2. Виртуальные помощники	Обеспечение поддержки клиентов с помощью чат-ботов и голосовых помощников. Предоставление персонализированных рекомендаций и помощи. Помощь в управлении задачами, планировании и напоминаниях
3. Дизайн и искусство	Создание визуальных дизайнов, таких как логотипы и графика. Создание художественных работ, включая картины и иллюстрации. Разработка 3D-моделей и виртуальных сред
4. Развлечения и игры	Разработка персонажей, уровней и сценариев видеоигр. Создание сценариев фильмов и сюжетных линий. Сочинение музыки и создание звуковых эффектов
5. Расширение данных и моделирование	Генерация синтетических данных для обучения модели машинного обучения Моделирование реалистичных сценариев исследований и разработок. Повышение конфиденциальности данных с помощью анонимизированных наборов данных
6. Языковой перевод и обработка естественного языка	Перевод текста с одного языка на другой. Обобщение длинных статей и документов. Анализ настроений и тематическое моделирование

Таблица 3

ANN для мобильных приложений с большим объемом данных

Применение	Архитектура / Алгоритм
Моделирование процессов и управление ими	Радиальная базисная сеть
Машинная диагностика	Многослойный персептрон
Управление портфолио	Алгоритм с контролем классификации
Распознавание целей	Модульная нейронная сеть
Медицинская диагностика	Многослойный персептрон
Кредитный рейтинг	Логистический дискриминантный анализ с помощью ANN и машины опорных векторов
Целевой маркетинг	Алгоритм обратного распространения
Распознавание голоса	Многослойный персептрон, глубокие нейронные сети (сверточные нейронные сети)
Финансовое прогнозирование	Алгоритм обратного распространения
Интеллектуальный поиск	Глубокая нейронная сеть
Обнаружение мошенничества	Алгоритм градиентного спуска и алгоритм наименьших квадратов (LMS).

В табл. 2 приведены направления и области применения приложений ANN**.**

В табл. 3 приведены варианты применения и архитектура/алгоритм ANN для мобильных приложений с большим объемом данных.

Результаты исследования и их обсуждение

Работа с нейронными сетями проходит в несколько этапов: подготовка ANN (выбор архитектуры, топологии и алгоритмов обучения); загрузка входных данных в нейронную сеть; обучение ANN; проверка адекватности обучения. Далее происходит этап использования нейронной сети – разработчики интегрируют обученную модель в приложение. Именно здесь происходит основа AI и ML: вместе с объемом входных данных, поступающих в нейросеть, поступает информация об ожидаемом результате. Результат, полученный на выходном уров-

не нейронной сети, сравнивается с ожидаемым. Если они не совпадают, нейронная сеть определяет, какие нейроны повлияли на конечное значение в большей степени, и корректирует веса соединений с этими нейронами (так называемый алгоритм обратного распространения ошибок).

Ограничения нейронных сетей на мобильных устройствах

Ограничения оперативной памяти. Большинство мобильных устройств среднего и бюджетного класса, доступных на рынке, имеют от 2 до 4 ГБ оперативной памяти. И обычно 1/3 этой емкости зарезервирована операционной системой. Система может «убивать» приложения с нейронными сетями по мере их запуска при приближении к ограничению оперативной памяти.

Размер приложения. Сложные глубокие нейронные сети часто «весят» несколько гигабайт. При интеграции нейронной сети в мобильное программное обеспечение происходит некоторое сжатие, но этого все равно недостаточно для комфортной работы. Основная рекомендация разработчикам – максимально минимизировать размер приложения на любой платформе для улучшения пользовательского интерфейса.

Время выполнения. Простые нейронные сети часто возвращают результаты почти мгновенно и подходят для приложений реального времени. Однако глубоким нейронным сетям могут потребоваться десятки секунд для обработки одного набора входных данных. Современные мобильные процессоры пока не такие мощные, как серверные, поэтому обработка результатов на мобильном устройстве может занять несколько часов.

Работа с одним приложением на нескольких устройствах. В качестве примера на телефоне и планшете пользователя установлено приложение для распознавания лиц. Оно не сможет передавать данные на другие устройства, поэтому обучение нейронной сети будет происходить отдельно на каждом из них.

Чтобы разработать мобильное приложение с использованием нейронных сетей на смартфонах, сначала необходимо создать и обучить нейронную сеть на сервере или ПК, а затем реализовать ее в мобильном приложении, используя готовые фреймворки. Когда используются нейронные сети с меньшим количеством процессорных блоков и весов, программное моделирование выполняется непосредственно на компьютере. Например, распознавание голоса и т.д. Когда алгоритмы нейронных сетей разовьются до такой степени, что полезные действия можно будет выполнять с 1000 нейронами и 10000 синапсами, высокопроизводительное аппаратное обеспечение нейронных сетей станет необходимым для практической работы в мобильной сети [3].

Нейронный процессор или блок нейронной обработки (Neural Processing Unit, далее *NPU*) – это специализированная схема, которая реализует все необходимые элементы управления и арифметическую логику, необходимые для выполнения алгоритмов машинного обучения, обычно путем работы с моделями прогнозирования, такими как ANN или случайные леса (Random forest, RF) метода машинного обучения AI [4, с. 142].

Мировыми OEM-производителями процессоров и мобильных чипов ИИ и ML смартфонов являются компании: Qualcomm; AMD; MediaTek (MTK); Apple; Google; Intel; Samsung; HiSilicon (Huawei); Amazon Web Services (AWS); Oppo (BBK Electronics); Nvidia; Unisoc (Spreadtrum); Kneron. На современном этапе к классу нейронных процессоров типа NPU могут относиться и другие, разные по устройству и специализации типы чипов, например, нейроморфные процессоры (Biotic Processing Unit, BPU), тензорные процессоры (Tensor Processing Unit, TPU), интеллектуальные процессоры искусственного интеллекта (Intelligence Processing Unit, IPU) и др. Постепенно все задачи нейронных сетей сегодня будут выполняться выделенными блоками NPU [4, с. 144]. Среди отечественных разработок нейронных процессоров на сегодня назовем: NPU-процессор «Алтай», совместной работы «Лаборатории Касперского» и компании «Мотив нейроморфные технологии»; тензорный процессор *TPU IVA* от компании IVATechnologies группы компаний «Хайтек», нейронный процессор Module NM6408 от компании АО НПЦ «Модуль».

Для создания и внедрения нейронных сетей для Android приложений применяют различные инструменты, наиболее популярные: библиотеки для мобильных устройств TensorFlow Lite, Keras и PyTorch, готовые решения ML Kit для распознавания изображений и звука, анализ текста, аппаратное ускорение для вычислений NNAPI (Neural Networks API). Помимо готовых методов есть поддержка пользовательских моделей. Мобильные устройства накладывают свои ограничения на работу нейронных сетей. Если вы решите их использовать, лучшим выбором будет готовое решение от Google (ML Kit) или разработка и внедрение своей собственной нейронной сети с TensorFlow Lite.

Популярными мировыми нейронными сетями для Android приложений на сегодня являются: ChatGPT 3.5; Midjourney; Silero TTS; Remini; Google Assistent и другие.

Наиболее востребованными вариантами отечественных нейронных сетей – приложений для смартфонов (планшетов) России являются: GigaChat (NeONKA ruGPT-3.5) и Kandinsky 2.2 (Fusion Brain) от компании Сбербанк; YandexGPT 2 («Алиса, давай придумаем») и Шедеврум от компании Яндекс и ряд других мобильных приложений.

Вариантами отечественных смартфонов России на сегодня являются:

− MIG S6 от компании «Адванст Мобилити Солюшинз», г. Санкт-Петербург;

− Р-фон модель RT001 от Группа компаний АО «Рутек», г. Москва;

− F+ tech модель R570E от компании ООО «Ф-Плюс оборудование и разработки», г. Санкт-Петербург;

− AYYA T1 от компании ООО «Смартэкосистема» концерна «Автоматика», г. Москва.

Все смартфоны работают на ОС Android и различных отечественных операционных системах: «Astra Linux Mobile»; «Аврора»; Kaspersky OS; «Роса Мобайл» и «Ред ОС М». Взаимодействие разработанных нейронных сетей с операционной системой (ОС) смартфонов Android проверяется на этапе разработки. Проведены тесты с ОС «Аврора». Однако с отечественными ОС: Kaspersky OS, «Ред ОС М», «Роса Мобайл» тестирование мобильных устройств не проводится, что может критично отозваться в дальнейшем. Согласно статистике GS Group, в 2024 г. пользователей российских ОС сейчас насчитывается в B2B-сегменте (business to business) 1,2 млн, через год их будет 1,3 млн, а через два года – 1,4 млн. В сегменте B2G (business to government) количество потенциальных пользователей российских ОС около 85 тыс. [5].

Заключение

Мобильные устройства накладывают свои ограничения на работу нейронных сетей. Возможности современных смартфонов с их многоядерными процессорами, выделенными графическими процессорами и гигабайтами оперативной памяти уже вышли далеко за рамки запуска стандартных встроенных в телефон приложений или простых мобильных игр. Процессоры смартфонов находятся на грани запуска некоторых мощных нейронных сетей в качестве программного обеспечения. Авторы рассмотрели основные направления разработки приложений ANN, показали критические вопросы реализации ANN в России. Основная рекомендация разработчикам – максимально минимизировать размер приложения ANN на любой платформе для улучшения пользовательского интерфейса, особенно отечественного смартфона (планшета).

Список литературы

1. Светашов А.К. Использование искусственных нейронных сетей для их применения в существующих и перспективных радиосистемах: тематическое исследование // Молодой ученый. 2023. № 22 (469). С. 52–57.

2. Галушкин А.И. Нейронные сети: основы теории. М.: Горячая линия – Телеком, 2012. 496 с.

3. Martin Saint Laurent, Pierre Bassett, Ulyan Andersen. 28nm DSP based on embedded LDO for high-performance and energy-efficient mobile applications. [Электронный ресурс]. URL: https://chipsandcheese.com/2023/10/04/qualcomms/34QIT1172.pdf (дата обращения: 28.10.2023).

4. Вэнь Тонг, Пейин Чжу, Сети 6G. Путь от 5G к 6G глазами разработчиков. От подключенных людей и вещей к подключенному интеллекту / Пер. с англ. В.С. Яценкова. М.: ДМК Пресс, 2022. 624 с.

5. Дорофеев Г.И. Россияне создали полноценный отечественный Android. Госсектор получил достойную альтернативу iPhone // CNews. [Электронный ресурс]. URL: https:// https://importfree.cnews.ru/news/top/ysclid=ls5sbgkbil629230347.pdf (дата обращения: 22.11.2023).