ОСОБЕННОСТИ МАШИННОГО ОБУЧЕНИЯ СРЕДСТВАМИ CNN В РАМКАХ СИНТЕЗА РЕЧИ

© Э. С. Израилова

ГГНТУ им. акад. М. Д. Миллионщикова, г. Грозный, Россия

АН Чеченской Республики, г. Грозный, Россия

В статье описан процесс моделирования системы синтеза речи, основанный на имеющихся акустических данных, используемых для машинного обучения с целью получения модели, соответствующей естественным характеристикам речи. Представлены характерные особенности и архитектуры сверточных нейронных сетей CNN, в которых данные реализованы в виде набора изображений, применяются локальные операции-свертки, модифицирующие и комбинирующие данные изображений карты друг с другом. Рассмотрены актуальные способы машинного обучения и применение ускоренной на GPU библиотеки для глубоких нейронных сетей NVIDIA CUDA. Дана информация о подготовке обучающей экспериментальной базы данных, процессе машинного обучения системы, настройке параметров нейронной сети, показана динамика улучшения показателей матрицы внимания и результат обучения глубокой сверточной сети CNN.

Ключевые слова: система синтеза речи, машинное обучение, глубокая сверточная нейронная сеть, CNN сети.

В настоящее время синтез речи применяется в самых разных областях: голосовые помощники, информационно-справочные системы, военная и космическая техника, робототехника, акустический диалог человека с компьютером. До недавнего времени системы синтеза речи представляли из себя многомодульные сложные приложения, в которых реализуется множество алгоритмов и подходов. Современные системы синтеза речи основаны на использовании методов нейросетевого программирования и машинного обучения, благодаря которым удалось автоматизировать работу большинства компонентов таких систем.

Машинное обучение – это раздел искусственного интеллекта, характерной чертой которого является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач. Машинное обучение находится на стыке математической статистики, методов оптимизации и классических математических дисциплин, но имеет также и собственную специфику, связанную с извлечением информации, интеллектуальным ана-

лизом данных и проблемами вычислительной эффективности [1].

Сфера применения технологий машинного обучения широка и при этом постоянно увеличивается, это и распознавание образов, медицинская и техническая диагностика, статистический анализ, задачи прогнозирования, управления и принятия решений, обработка текстовых массивов, синтез речи. Например, в области лингвистики машинное обучение помогает определять грамматические характеристики слов; в фонетике машинное обучение используется для предсказания частоты основного тона и длины фонем. Обучению поддаются также и просодические характеристики, такие как определение длины и местонахождения пауз, предсказание интонаций.

В данной статье описан процесс моделирования системы статистического синтеза речи, основанный на имеющихся акустических данных, используемых для машинного обучения системы с целью получения модели соответствия характеристикам речи. Машинное обучение строится на данных, а от их качества и

Рис. 2. Архитектура многомерной CNN сети

количества напрямую зависит эффективность распознавания и синтеза речи.

Поэтому необходимым условием для реализации методов машинного обучения является наличие большого количества данных в виде корпуса конкретного языка, основанного на использовании акустико-фонетических, текстовых, речевых баз данных.

Сверточные нейронные сети CNN для синтеза речи

Наиболее распространенными моделями машинного обучения являются искусственные нейронные сети, использующиеся для обработки входных данных комбинации распределенных простых операций, зависящие от обучаемых параметров. Современные нейронные сети применяются в задачах прогнозирования, распознавания образов и речи, генерации текстов и многих других. Наиболее оптимальными их видами являются сверточные сети и рекуррентные нейросети.

Для создания описываемой в данной статье системы синтеза речи использовалась глубокая сверточная нейронная сеть, успешно используемая и другими исследователями в распознавании образов, обработке аудиосигналов и речи.

Характерной особенностью сверточных нейросетей является использование представления данных в виде набора изображений (карт) и применение локальных операций-сверток, модифицирующих и комбинирующих данные карты друг с другом.

На рисунке 1 показана модель сверточной нейронной сети, состоящей только из двух 2D сверточных слоев, за которыми следует глобальный слой максимального пула (GlobalMaxPool). На вход модели представляется матрица признаков X размером $N \times M$. Первый сверточный слой состоит из K фильтров размером $1 \times M$ с функцией активации ReLu. Второй сверточный слой имеет только 1 фильтр размером 1×1 . Он объединяет функции, выбранные фильтрами свертки на предыдущем слое для каждого из N кандидатов в отдельности. Этот слой не имеет функции активации. После чего GlobalMaxPool слой выбирает кандидата с максимальной активацией, тем самым предполагая, что кандидат содержит гармонику в качестве компонента входного сигнала [2]. Далее выбранный кандидат проходит через сигмоидальную функцию активации, в результате на выходе сети будет представлена вероятность, что входной сигнал является голосом.

Рис. 1. Архитектура 2-мерной сверточной сети

Эффективность и качество работы нейронной сети зависят от выбора её архитектуры. При моделировании систем синтеза речи часто используются глубокие сверточные нейронные сети, которые состоят из 1000-2000 нейронов в каждом из 6-8 скрытых слоев (рис. 2).

Подготовка обучающих данных

Основой методики создания речевых баз данных естественных языков, используемых для машинного обучения, является наличие большого объема записанной речи (от 15 часов), кроме того, корпус должен быть сбалансированным и по возможности полным, то есть содержать все необходимые единицы во всех возможных контекстах с различными характеристиками. Предварительно была скомпонована экспериментальная обучающая база данных (5 часов речи), при этом были использованы аудиоматериалы, находящиеся в свободном доступе.

При построении систем синтеза речи одной из наиболее важных задач является сегментация и маркировка баз данных речевых сигналов на семантически и фонетически значимые единицы речи, а в нашем конкретном случае это фонемы. Полученные сегменты сохраняются в базе и служат для машинного обучения акустических моделей в интегральной системе, с последующей генерацией голоса диктора в системе синтеза речи по тексту. Одним из специфических методов работы с обучаемыми системами является настройка параметров системы под конкретный язык, подбор алфавита. Так как модель синтезатора чеченской речи использует в качестве символов входного алфавита фонемы, необходимо было решение задачи транскрибирования текстов по правилам графема-к-фонеме с учетом фонетических особенностей чеченского языка. Поставленная задача была решена путем создания модуля фонетической транскрипции, и затем подготовленная текстовая обучающая выборка была транскрибирована [3].

Таким образом была создана обучающая экспериментальная база, состоящая из 3500 предложений, а каждому предложению соответствует звуковой файл в формате wav с частотой дискретизации 22050 Гц. После этого был задействован модуль системы, отвечающий за получение спектрограмм звуковых файлов, на основе которых и происходит глубокое обучение сетей (рис. 3).

Рис. 3. Образец предложения из обучающей базы

Машинное обучение системы строилось на основе библиотек и модулей языка программирования Python (Tensorflow, matplotlib, numpy, scipy).

Основной этап обучения системы синтеза речи состоит из двух сетей: Text²Mel, которая синтезирует мел-спектрограмму из входного текста, и Spectrogram Super-resolution Network (SSRN), которая преобразует мел-спектрограмму сигнала в амплитудную спектрограмму STFT, учитывая пропуски кадров и восстанавливая частоту дискретизации.

Описание процесса машинного обучения

Машинное обучение реализуется как процесс компьютерного анализа подготовленных статистических данных с целью поиска закономерностей и создания на их основе нужных алгоритмов, настройки параметров нейронной сети и дальнейшего прогнозирования работы системы.

Существует 3 основных способа машинного обучения [4]:

- обучение с учителем;
- обучение с подкреплением;
- обучение без учителя (самообучение).

В обучении с учителем используются специально отобранные данные, в которых уже известны и надежно определены правильные ответы, а параметры нейронной сети подстраиваются так, чтобы минимизировать ошибку. В этом способе система искусственного интел-

лекта может сопоставить правильные ответы к каждому входному примеру и выявить возможные зависимости ответа от входных данных. Можно обучить систему на очень больших объемах входных данных и варьировать параметры нейронных сетей, пока на выходе не получатся ожидаемые результаты. После этого надо проверить, как система прогнозирует результат для контрольных данных, с которыми компьютер еще не сталкивался.

Чаще всего обучение с учителем применяется для решения задач классификации, регрессии и прогнозирования. Например, существуют алгоритмы классификации для анализа текста и определения частей речи конкретного языка в текстах различной тематики, регрессия может применяться для категоризации текстовых документов.

Обучение с подкреплением основывается на регламентированных процессах обучения, при которых алгоритм машинного обучения снабжен набором действий, параметров и конечных значений. Определив правила, алгоритм машинного обучения пытается изучить различные варианты и возможности, отслеживая и оценивая каждый раз результат, чтобы определить, какой из вариантов является оптимальным. Подкрепляемое обучение – это метод проб и ошибок для машины. Она учится на прошлом опыте и меняет свой подход, реагируя на новую ситуацию, пытаясь достичь наилучшего возможного результата.

В случае использования способа обучения без учителя алгоритм машинного обучения изучает данные с целью выявления закономерностей (паттернов). Не существует готовых данных с ответами или оператора, который мог бы обучить машину. Напротив, программа сама определяет корреляции и связи на основе анализа доступных данных. При обучении без учителя алгоритму машинного обучения позволено самостоятельно интерпретировать большие наборы данных и делать на их основе выводы. Алгоритм пытается каким-либо образом упорядочить данные и описать их структуру. Это может выглядеть как группировка данных в кластеры, или это такое упорядочивание данных, при котором они начинают выглядеть систематизированно.

По мере поступления данных для анализа растёт способность алгоритма принимать решения на основе этих данных, а также точность этих решений. Ярким примером самообучения является кластеризация, подразумевающая группировку похожих данных на основе определенных параметров. Это может быть использовано для сегментации данных на несколько групп и проведения анализа на основе каждого набора данных по отдельности для поиска закономерностей. Обучение без учителя также можно использовать для анализа тональности высказываний, чтобы определять эмоциональное состояние людей на основе их устной и письменной речи.

Механизмы глубокого машинного обучения (deep learning) используют, как правило, многослойные нейросети и очень большое число экземпляров объектов для тренировки нейронной сети. Число записей в обучающей выборке может насчитывать сотни тысяч или даже миллионы примеров, а когда ресурсы не ограничены – еще больше.

От размера обучающей базы данных и количества записей входных данных зависит создание машиной необходимых правил классификации. Чем больше разнородных данных будет загружено в систему на этапе машинного обучения, тем точнее будут выявлены эти правила, и тем в конечном итоге точнее будет результат работы системы синтеза речи, а генерируемый на выходе речевой сигнал будет более качественный.

Начинающим исследователям в сфере машинного обучения и его применения для синтеза речи целесообразно использовать методы машинного обучения с учителем. Это потребует меньше ресурсов как временных, так и финансовых при создании прототипа работающей системы и практическом освоении методик искусственных нейронных сетей. Функционирующую систему искусственного интеллекта под конкретную задачу в этом случае можно получить быстрее. В настоящее время есть большое число находящихся в свободном доступе качественных библиотек программного кода для искусственных нейросетей, таких как TensorFlow <https://www.tensorflow.org/> для математического моделирования, OpenCV

<http://opencv.org/> для задач распознавания изображений.

Процесс машинного обучения был реализован с соблюдением основного требования системы к оборудованию: наличие компьютера с двумя графическими процессорами, видеопамятью емкостью 12 Гб и оперативной памятью 64 Гб. Аппаратное обеспечение процесса обучения состояло в использовании видеокарты NVIDIA GeForce GTX 1080 Ti. Для обучения модели требовалась поддержка CUDA. Исходя из этого на компьютер была установлена библиотека NVIDIA CUDA (cuDNN), которая представляет собой ускоренную на GPU библиотеку примитивов для глубоких нейронных сетей. CUDNN предоставляет хорошо настроенные реализации для стандартных подпрограмм, таких как прямая и обратная свертка, пул, нормализация и уровни активации.

При этом были учтены особенности работы видеоадаптеров NVIDIA и возможности, предоставленные разработчику в платформе CUDA. В отличие от CPU, применяемых для вычислений общего характера, узконаправленность развития видеокарт сформировала особую архитектуру, ориентированную на максимально эффективную обработку графики. Подобные задачи характеризуются выполнением однотипных действий на множественном объеме данных, поэтому GPU реализует SIMD архитектуру, нацеленную на использование наибольшего количества параллельных потоков, до нескольких тысяч, для расчетов. Одна из основ технологии CUDA – использование GPU в качестве сопроцессора CPU. В силу различий архитектуры процессоры отличаются производительностью при решении определенного спектра задач. Например, CPU намного эффективнее выполняет расчеты, связанные с множественными переходами и ветвлениями. В свою очередь, GPU отлично справляется с массовым параллелизмом, обрабатывая большие объемы информации, и не должен применяться к последовательным алгоритмам [5].

Генерация речи с предварительно обученной моделью может выполняться на GPU и CPU, но при этом скорость синтеза речи на CPU в нашем случае снизилась в десятки раз.

Обучение системы продолжалось 27 часов, было выполнено 500 тысяч итераций для

нейронной сети Text²Mel, 270 тысяч итераций для сети SSRN.

Рис. 4. Динамика обучения матрицы внимания

На рисунке 4 можно проследить динамику улучшения показателей матрицы внимания при обучении нейронной сети (5 тысяч итераций, 50 тысяч итераций, 400 тысяч итераций).

Сеть TextEnc сначала кодирует входное предложение $L = [l_1, \dots, l_N] \in \text{Char}^N$, состоящее из N символов, в двух матрицах $K, V \in \mathbb{R}^{d \times N}$. С другой стороны, сеть AudioEnc кодирует круп-

номасштабную спектрограмму $S(i; S_{1:F, 1:T}) \in R^{FT}$ ранее записанной речи, длина которой равна T , в матрицу $Q \in R^{dT}$.

$$(K, V) = \text{TextEnc}(L)$$

$$Q = \text{AudioEnc}(S_{1:F, 1:T})$$

Матрица внимания $A \in R^{NT}$, определяемая следующим образом, оценивает, насколько сильно связаны n -й символ l_n и t -й временной интервал $S_{1:F, t}$:

$$A = \text{softmax}_{n\text{-axis}}\left(\frac{K^T Q}{\sqrt{d}}\right)$$

$A_{nt} \sim 1$ подразумевает, что модуль смотрит на n -й символ l_n на временном интервале t , и он будет смотреть на l_n или l_{n+1} или символы вокруг них на последующем временном интервале $t + 1$. Что бы ни случилось, необходимо ожидать, что они будут закодированы в n -м столбце V . Таким образом, начальное число $R \in R^{dT}$, декодированное в последующие кадры: $S_{1:F, 2:T+1}$, получают как:

$$R = \text{Att}(Q, K, V) := VA$$

Результат R объединяется с кодированным звуком Q , как $R' = [R, Q]$.

Затем конкатенированная матрица $R' \in R^{2dT}$ декодируется модулем Audio Decoder для синтеза крупномасштабной спектрограммы:

$$Y_{1:F, 2:T+1} = \text{AudioDec}(R')$$

На стадии синтеза матрица внимания A иногда может неправильно считывать некоторые символы. Типичными ошибками, которые наблюдаются чаще всего, являются:

- иногда пропускаются несколько символов;
- неоднократно читается одно и то же слово дважды или более.

Чтобы сделать систему более надежной, необходимо эвристическим образом модифицировать матрицу A как «почти диагональную». После этого наблюдается, что устройство иногда начинает смягчать такие ошибки [6].

Выводы. Результат эксперимента по машинному обучению и моделированию системы синтеза чеченской речи оказался успешным, удалось синтезировать речевой сигнал. По нашим предварительным оценкам качество синтезируемой речи среднее. Некоторые слова система еще плохо читает, не всегда правильно озвучиваются долгие и краткие гласные, голос системы – роботизированный, но на данном первоначальном этапе, при использовании минимальной обучающей базы, полученный результат является в большей мере положительным: система синтеза чеченской речи создана, остается улучшить качество.

ЛИТЕРАТУРА

1. Абдрахманова Э.Р. Машинное обучение: Обзор и применение // Техника и технологии: Пути инновационного развития: сборник научных трудов 8-й Международной научно-практической конференции. Юго-Западный государственный университет. Курск, 2019. С. 9-12.
2. Vashkevich, R. Convolutionalneuralnetworkwithsemanticallysignificantactivationsforspeechanalysis // Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2018): материалы Международной научно-технической конференции (Минск, 15-17 февраля 2018 года). Минск: БГУИР, 2018. С. 227-230.
3. Израилова Э.С. Моделирование системы синтеза речи на основе глубоких сверточных нейронных сетей // Известия Кабардино-Балкарского научного центра РАН. 2018. № 6 – II (86). С. 181-186. <https://elibrary.ru/item.asp?id=37094528>
4. Климов М.И., Меньшов И.С. Эффективный параллельный расчет задач механики сплошной среды для систем гибридной архитектуры // Научный Вестник МГТУ ГА. М.: МГТУ ГА, 2012. № 194. С. 31-39.
5. Hideyuki Tachibana, Katsuya Uenoyama, Shunsuke Aihara. Efficientlytrainabletext-to-speechsystembasedondeepconvolutionalneuralnetworks with guided attention/ [Электронный ресурс]. URL: <https://arxiv.org/pdf/1710.08969.pdf> (дата обращения: 09.08.2019).

FEATURES OF MACHINE LEARNING BY CNN WITHIN THE SPEECH SYNTHESIS

© E. S. Izrailova

GSTOU named after acad. M. D. Millionshchikov, Grozny, Russia

AS of the Chechen Republic, Grozny, Russia

The article describes the process of modeling a speech synthesis system based on the available acoustic data used for machine learning in order to obtain a model corresponding to the natural characteristics of speech. The characteristic features and architectures of convolutional neural networks CNN, in which the data is implemented as a set of images, local operations are used-convolution, modifying and combining the data of images of the map with each other. Actual methods of machine learning and application of GPU-accelerated library for deep neural networks NVIDIA CUDA are considered. Given information about the preparation of training experimental database, the machine learning process of the system configuration parameters of the neural network, shows the dynamics of improvement of the matrix of attention and the result of learning a deep convolutional network CNN.

Keywords. Speech synthesis system, machine learning, deep convolutional neural network, CNN networks.

REFERENCES

1. Abdrahmanova, Je. R. (2019) 'Mashinnoe obuchenie: Obzor i primeneniye'. *Tehnika i tehnologii: Puti innovacionnogo razvitiya: sbornik nauchnykh trudov 8-j Mezhdunarodnoj nauchno-prakticheskoy konferencii*. Jugo-Zapadnyj gosudarstvennyj universitet [Machine Learning: Overview and Application. Technique and Technology: Ways of Innovative Development: a collection of scientific papers of the 8th International Scientific and Practical Conference. Southwestern State University]. Kursk. Pp. 9-12.
2. Vashkevich, R. (2018) 'Convolution alneural network with semantically significant activations for speech analysis'. *Otkrytye semanticheskie tehnologii proektirovaniya intellektual'nykh sistem = Open Semantic Technologies for Intelligent Systems (OSTIS-2018): materialy Mezhdunarodnoj nauchno-tehnicheskoy konferencii* [Materials of the international scientific and technical conference. Minsk, February 15-17, 2018) Minsk: BSUIR, Pp. 227-230.
3. Izrailova, Je. S. (2018) 'Modelirovaniye sistemy sinteza rechi na osnove glubokikh svertochnykh nejronnykh setej' *Izvestiya Kabardino-Balkarskogo nauchnogo centra RAN*. [Bulletin of the Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences]. № 6 – II (86). Pp. 181-186. Available at: <https://elibrary.ru/item.asp?id=37094528>
4. Klimov, M. I. and Men'shov I. S. (2012) Effektivnyj parallel'nyj raschet zadach mehaniki sploshnoj sredy dlja sistem gibridnoj arhitektury' *Nauchnyj Vestnik MGTU GA – M. MGTU GA*. [Effective parallel calculation of continuum mechanics problems for hybrid architecture systems // Scientific Herald of the Moscow State Technical University GA – M. MSTU GA] № 194. Pp. 31-39.
5. Hideyuki, Tachibana, Katsuya, Uenoyama and Shunsuke, Aihara (2019) Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. Available at: <https://arxiv.org/pdf/1710.08969.pdf>