ВЕСТНИК ВОЕННОГО ИННОВАЦИОННОГО ТЕХНОПОЛИСА «ЭРА», 2023, том 4, № 3, с. 264–268

ЭЛЕКТРОННАЯ КОМПОНЕНТНАЯ БАЗА И ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ УПРАВЛЕНИЯ

УДК 004.27

КЛАССИФИКАЦИЯ АУДИОСИГНАЛОВ ИМПУЛЬСНОЙ СВЕРТОЧНОЙ НЕЙРОННОЙ СЕТЬЮ

© 2023 г. Р. Б. Рыбка[1], А. В. Серенко[1], А. В. Наумов[1], А. Г. Сбоев[1,2,*]

[1]Национальный исследовательский центр "Курчатовский институт", Москва, Россия

[2]Национальный исследовательский ядерный университет "МИФИ", Москва, Россия

**E-mail: Rybka_RB@nrcki.ru*

Поступила в редакцию 19.10.2023 г.
После доработки 26.10.2023 г.
Принята к публикации 26.10.2023 г.

Исследована эффективность импульсной сверточной нейронной сети для решения задач классификации звукового сигнала. Предложена модель нейронной сети на базе сверточной архитектуры для обработки звукового сигнала в потоковом режиме. Обучение модели выполняется с применением аргументированного набора данных аудиозаписей. Выявлено, что сверточная нейронная сеть показывает значительно лучшие точности по сравнению с полносвязной архитектурой. Проведено построение на ее основе импульсной сверточной нейронной сети методом переноса весов. Оценка точностей импульсной модели показывает незначительную потерю точности после конвертации (3–4%).

DOI: 10.56304/S2782375X23030142

ВВЕДЕНИЕ

Импульсные нейронные сети, в которых информация представлена в форме наличия или отсутствия импульсов-спайков в каждый момент времени, перспективны для применения в задачах машинного обучения благодаря чрезвычайно низкому энергопотреблению, которое такие сети способны демонстрировать при их аппаратной реализации на специализированных нейроморфных вычислительных архитектурах. В то же время на точности решения задач машинного обучения негативно сказывается дискретизация информации при кодировании ее последовательностями спайков. Поэтому остается актуальным исследование эффективности решения импульсными нейронными сетями различных классификационных задач.

В данной работе анализируется точность решения импульсной нейронной сетью задачи классификации звуков набора данных UrbanSound8K. При этом оценивается влияние на точность классификации различных конфигураций сетей и параметров кодирования данных спайковыми последовательностями. Ранее [1] был предложен метод классификации аудиозаписей в потоковом режиме с проверкой на данном наборе данных. В настоящей работе предлагается развитие этого метода за счет улучшения качества нейросетевой модели в составе алгоритма классификации. Рассматривается сверточная нейронная сеть с двумерным окном свертки, на вход которой поступают звукозаписи, представленные спектральными признаками. Импульсная нейронная сеть формируется путем переноса синаптических весов с предварительно обученной формальной нейронной сети.

Таким образом, итоговая точность решения задачи может быть увеличена на 10% при использовании сверточной архитектуры при существенном снижении общего количества связей в сети.

МЕТОДЫ И ПОДХОДЫ

Синаптические веса импульсной нейронной сети получены путем обучения эквивалентной формальной нейронной сети методом обратного распространения ошибки. Для рассматриваемой формальной нейронной сети выбрана сверточная топология с двумерным окном свертки (2dCNN), способная извлекать сложные паттерны и временные зависимости из аудиосигнала. Она выбрана благодаря успешному использованию как в задачах выделения ключевых слов [2], так и в смежных задачах обработки аудиозаписей. Например, в [3] была предложена модель Whisper, которая превосходит лучшие опубликованные модели для решения задачи распознавания речи. В ней для обработки исходного аудиосигнала используют Мел-спектрограммы, которые подают в два сверточных слоя с шириной фильтра 3 и активационной функцией Gaussian Error Linear Unit (**GELU**).

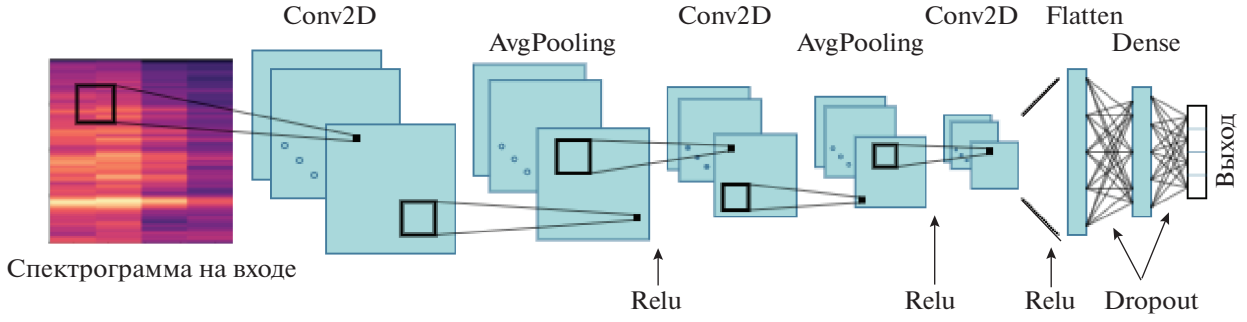

Рис. 1. Топология нейронной сети на базе сверточных слоев.

В работе использовали сеть с четырьмя внутренними слоями, из которых первые три слоя являлись сверточными с двумерным окном свертки 5 × 5 и количеством фильтров 24, 48 и 48 соответственно. После каждого сверточного слоя имеется слой AveragePooling, выход последнего из них преобразовывался в одномерный вектор с помощью метода Flatten и передавался в полносвязный слой Dense из 64 нейронов. Все слои, включая AveragePooling, имели функции активации Rectified Linear Unit (**ReLU**). Входные связи полносвязных слоев – последнего скрытого слоя из 64 нейронов и выходного из 10 нейронов – при обучении подвергали регуляризации методом Dropout с вероятностью 0.5 исключения каждой связи из обучения.

Полная схема используемой топологии представлена на рис. 1.

Обучение нейронной сети проводилось оптимизатором Adam с параметрами learning rate 10^{-5} и batch size 10; в качестве оптимизируемой функции потерь использовалась categorical cross-entropy. Останов обучения проводился по достижении 100 эпох или при неулучшении функции потерь, вычисленной на валидационной выборке данных, в течение 10 эпох.

На основе полученной сверточной нейронной сети после ее обучения формировалась эквивалентная импульсная нейронная сеть. Для этого сверточные и полносвязные слои заменялись слоями импульсных нейронов с той же топологией и теми же весами синаптических связей; этап преобразования выхода Flatten реализовывался путем установления связей между соединяемыми сверточным и полносвязным слоями; слои AveragePooling заменялись слоями импульсных нейронов с соответствующими синаптическими весами. Импульсные нейроны были реализованы моделью “пороговый интегратор” (Integrate-and-Fire) без утечки трансмембранного потенциала и рефрактерности. В этой модели каждый входной спайк увеличивает безразмерную переменную состояния – трансмембранный потенциал – на величину веса своего синапса, и, как только накопленный потенциал достигает порогового значения, он уменьшается на величину порога, а нейрон испускает выходной спайк. Таким образом, количество выходных спайков такого нейрона линейно зависит от общего количества входных спайков, умноженных на веса входов, что позволяет рассматривать данные нейроны как эквивалентные формальным нейронам с функцией активации ReLU.

Импульсную нейронную сеть получали численным моделированием с дискретным временем, что соответствует принципу функционирования существующих нейроморфных процессоров. Время исчисляется дискретными шагами по 1 мс, и на каждом временном шаге на каждый входной синапс нейрона поступает либо не поступает спайк, после чего нейрон испускает или не испускает выходной спайк.

Входные данные подавались на импульсную нейронную сеть закодированными частотным способом: один входной пример подавался на сеть в течение 200 мс, в которые на каждый вход сети подавалось количество спайков, пропорциональное значению соответствующего компонента обрабатываемого входного вектора, так, что минимальное входное значение (равное нулю для рассматриваемого набора данных) соответствовало отсутствию спайков на соответствующем входном синапсе в течение всех 200 временных шагов, а максимальное (равное 80) соответствовало подаче спайка на каждом из 200 временных шагов. В качестве выхода импульсной нейронной сети рассматривается вектор из количеств спайков, испущенных каждым из нейронов выходного слоя.

Пороговые значения потенциалов нейронов выбирали такими, чтобы в каждом слое сети реализовался весь возможный диапазон количества выходных спайков нейронов, от полного отсутствия спайков до испускания спайка на каждом временном шаге, и таким образом минимизировалась потеря точности, вызванная дискретизацией данных при представлении их количеством спайков. Для этого для каждого слоя нейронной сети было найдено максимальное количество

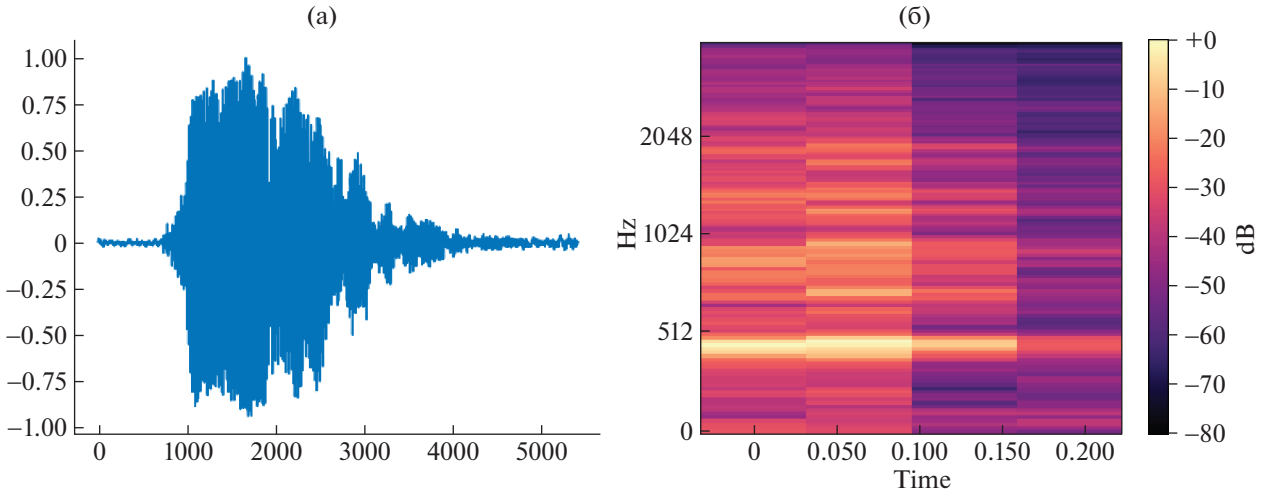

Рис. 2. Исходный аудиосигнал (а) и его Мел-спектрограмма в масштабе децибел (б).

спайков, которое какой-либо нейрон данного слоя испускает в ответ на какой-либо пример тренировочной выборки (после отбрасывания 1% примеров, вызывающих самое большое количество спайков), и затем пороговые потенциалы данного слоя устанавливались так, чтобы найденный нейрон в ответ на найденный входной пример испускал спайк на каждом временном шаге.

НАБОР ДАННЫХ

UrbanSound8K является открытым общедоступным набором данных [4], состоящим из 8732 звуковых файлов с частотой дискретизации 8 кГц, записанных в городской среде различными устройствами записи, со значениями длительности не более 4 с. Он используется для обучения и оценки различных моделей, решающих задачу классификации аудиосигнала на 10 классов (звук двигателя, дрели, кондиционера и т.д.). Каждый класс содержит не менее 800 звуковых файлов.

Для дальнейшей обработки набор данных был разделен на тренировочную, валидационную и тестовую части в пропорции 80, 10 и 10% соответственно, с равномерным распределением классов в каждой. Валидационная часть применялась для подбора гиперпараметров сети, а также для контроля переобучения сети.

Для извлечения признаков из исходного аудиосигнала используется библиотека с открытым исходным кодом librosa [6]. Аудиозаписи различной длины выравнивались до 3 с: имевшие меньшую длину удлинялись циклическим повторением, имевшие большую длину обрезались. Полученный аудиосигнал разделялся на небольшие фрагменты (фреймы) длиной 0.5 с и перекрытием в 0.25 с. Для каждого примера аудиофайла получалось по 12 фреймов.

Фреймы поступали на нейронную сеть в качестве входных примеров, закодированных векторами признаков, полученными следующим образом:

– преобразование звукового сигнала в спектрограмму – матричное представление амплитуд спектра звукового сигнала в различных дискретных моментах времени;

– применение мел-шкалы к спектрограмме – это позволяет учесть особенности слухового восприятия человека и выделить значимые частоты, которые имеют большое значение для анализа;

– перевод мел-шкалы в логарифмический масштаб – это позволяет сделать анализ более точным и устойчивым к шуму.

Для каждого фрейма вычисляется 128-канальное представление мел-спектрограммы по 128-миллисекундным окнам (window_length) с шагом 128 миллисекунд (hop_length) с помощью библиотеки Librosa.

Полученная мел-спектрограмма конвертируется из масштаба мощности в масштаб децибел. Далее значения (по модулю) используются в качестве входных данных для нейросетевой модели.

На рис. 2 представлены визуализации исходного аудиосигнала и его мел-спектрограмма в масштабе децибел.

Для расширения набора данных были проведены эксперименты с четырьмя различными методами расширения набора аудиоданных (аугментациями) из работы [5], в результате чего получаются пять наборов дополнений, как подробно описано ниже. Каждая аугментация применяется непосредственно к аудиосигналу перед преобразованием его в заданный набор признаков, используемый для обучения сети. Важным является то, что для каждой деформации выбраны параметры, обес-

Таблица 1. Точности классификации аудиосигналов

Модель	most_common		sum_argmax	
	*F*1-*micro*	*F*1-*macro*	*F*1-*micro*	*F*1-*macro*
MLP	0.54	0.57	0.56	0.59
2dCNN	0.70	0.68	0.71	0.70
Импульсная 2dCNN (100 мс)	0.65	0.66	0.67	0.67
Импульсная 2dCNN (200 мс)	0.65	0.66	0.67	0.68
Импульсная MLP [1]	0.54	0.55		

печивающие сохранение семантической достоверности метки. Используемые аугментации:

– растяжение по времени: замедление или ускорение аудио (при сохранении высоты звука неизменной). Каждый аудиосигнал был растянут/сжат во времени на четыре значения: {0.81, 0.93, 1.07, 1.23};

– изменение высоты тона: увеличение/уменьшение высоты аудиосигнала с сохранением длительности. Высота каждого аудиосигнала была увеличена/уменьшена на четыре значения (в полутонах): {–2, –1, 1, 2};

– изменение высоты тона: так как изменение высоты тона является особенно полезным дополнением, был создан второй набор дополнений. На этот раз высота каждого аудиосигнала была увеличена/уменьшена на четыре значения (в полутонах): {–3.5, –2.5, 2.5, 3.5};

– сжатие динамического диапазона: динамический диапазон аудиосигнала был пропорционально сжат с использованием четырех паттернов {музыкальный стандарт, стандарт фильмов, речь, радио}. Каждый из паттернов задает собой максимально и минимально допустимые значения амплитуды аудиосигнала;

– добавление фонового шума: смешивание аудиосигнала с записью, содержащей фоновые звуки из различных типов акустических сцен. Каждый аудиосигнал был смешан с четырьмя акустическими сценами: {звуки парка, звуки разговоров людей, звуки рабочих, звуки оживленной улицы}. Каждый аудиосигнал был изменен с помощью формулы

$$z = (1 - w) \cdot x + w \cdot y, \tag{1}$$

где z – аудиосигнал с добавленным шумом, x – аудиосигнал исходного аудио, y – аудиосигнал фоновой сцены, w – весовой параметр, который выбирался случайным образом из равномерного распределения в диапазоне [0.1, 0.5].

После применения аугментации удалось увеличить размер обучающей выборки в 20 раз.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

В качестве метрик для оценки модели использовались F1-micro и F1-macro:

$$F1 - micro = \\ = 2 \times (precision \times recall)/(precision + recall), \tag{2}$$

$$precision = TP/(TP + FP), \tag{3}$$

$$recall = TP/(TP + FN), \tag{4}$$

где *precision* – доля истинных положительных примеров относительно всех предсказанных положительных примеров, *recall* – доля истинных положительных примеров относительно всех положительных примеров, *TP* – количество верно классифицированных положительных примеров, *FP* – количество неверно классифицированных положительных примеров, *FN* – количество неверно классифицированных отрицательных примеров.

$$F1 - macro = \frac{\sum_0^N F1 - micro_n}{N}, \tag{5}$$

где N – количество классов, $F1 - micro$ – это $F1 - micro$ для класса n.

Дополнительно было проведено сравнение с предложенной ранее полносвязной моделью [1] типа многослойный персептрон (MLP), которая состоит из трех скрытых слоев с размерностями 1000, 500 и 100 нейронов с функцией активации ReLU, функцией оптимизации SGD и learning rate 0.0002. Входные признаки для этой модели из матрицы преобразовывались в одномерный вектор с помощью метода Flatten.

В табл. 1 представлены полученные точности моделей на тестовом множестве с двумя вариантами агрегации фреймов:

– по наиболее представительному классу среди всех фреймов одной аудиозаписи (most_common). Если для одной записи было несколько одинаковых классов по максимальной сумме голосов, то итоговым классом выбирался один из них случайно;

– по наибольшей суммарной активности среди выходов по всем фреймам (sum_argmax). У каждого

фрейма есть набор выходных активностей модели для каждого класса. Активности соответствующих классов суммировались между собой и выбирался класс с наибольшей суммарной активностью.

ЗАКЛЮЧЕНИЕ

Показано, что сверточная нейросетевая архитектура позволяет улучшить качество классификации по метрикам F1-micro и F1-macro по сравнению с полносвязной моделью прямого распространения. Реализация импульсной нейронной сети с применением метода переноса весов из обученной формальной сети снижает точность классификации на 3–4%. Полученная методом переноса весов импульсная сверточная нейронная сеть позволяет достичь точности 0.68 по метрике F1-macro на наборе данных классификации городских шумов, что превосходит результаты других импульсных нейронных сетей, применяемых для этой задачи.

Работа проведена в рамках выполнения государственного задания НИЦ “Курчатовский институт”.

СПИСОК ЛИТЕРАТУРЫ

1. *Сбоев А.Г., Рыбка Р.Б., Серенко А.В. и др.* // Вестник ВИТ Эра. 2022. Т. 3. Вып. 3 С. 314.
2. *Rybakov O., Kononenko N., Subrahmanya N. et al.* // arXiv preprint arXiv:2005.06720. 2020.
3. *Radford A., Wook Kim J., Tao X. et al.* // International Conference on Machine Learning. PMLR. 2023. P. 28492.
4. *Salamon J., Jacoby C., Bello J. P.* // Proc. 22 ACM Int. Conf. Multimed. 2014. P. 1041.
5. *McFee B., Raffel C., Liang D. et al.* // Proc. Python Sci. Conf. 2015. V. 8. P. 18.
6. *Salamon J., Bello J.P.* // IEEE Signal Process. Lett. 2017. V. 24. № 3. P. 279.