УДК 004.032.26

СРАВНЕНИЕ РЕКУРРЕНТНЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ОБРАБОТКИ РЕЧЕВЫХ СИГНАЛОВ

И.А. Коржова

Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники», г. Минск, Республика Беларусь

Аннотация. В работе рассматривается применение рекуррентных нейронных сетей для обработки речевых сигналов. Проведено сравнительное исследование архитектур RNN, LSTM и GRU при анализе признаков речевых сигналов, полученных методом мел-частотных кепстральных коэффициентов (MFCC). Выполнено обучение моделей на наборе аудиоданных и проведена оценка качества их работы по метрикам точности и функции потерь. Полученные результаты позволяют определить эффективность различных архитектур рекуррентных нейронных сетей при обработке речевой информации и могут быть использованы при разработке систем анализа и защиты речевых данных.

Ключевые слова: Обработка речевых сигналов; Нейронные сети; Рекуррентные нейронные сети; RNN; LSTM; GRU; MFCC; Машинное обучение; Анализ речи; Защита информации.

COMPARISON OF RECURRENT NEURAL NETWORKS FOR SPEECH SIGNAL PROCESSING

I.A. Korzhova
Educational Institution "Belarusian State University of Informatics and Radioelectronics", Minsk, Republic of Belarus

Abstract. This paper considers the application of recurrent neural networks for speech signal processing. A comparative study of RNN, LSTM and GRU architectures is carried out using speech features obtained by the Mel-Frequency Cepstral Coefficients (MFCC) method. The models are trained on a speech dataset and evaluated using accuracy and loss metrics. The obtained results allow assessing the effectiveness of different recurrent neural network architectures in speech signal processing and can be used in the development of speech analysis and information protection systems.
Keywords: Speech signal processing; Neural networks; Recurrent neural networks; RNN; LSTM; GRU; MFCC; Machine learning; Speech analysis; Information security.

Введение

Современные системы обработки речевых сигналов широко применяются в различных областях информационных технологий, включая системы распознавания речи, голосовые интерфейсы и средства анализа аудиоданных. Особую актуальность задачи обработки речевых сигналов приобретают в области технической защиты информации, где требуется анализ и обработка речевых данных для выявления возможных угроз безопасности.

Речевой сигнал представляет собой сложную временную последовательность, содержащую информацию о фонетических особенностях речи. Для эффективного анализа таких сигналов применяются методы извлечения признаков, одним из наиболее распространенных среди которых являются мел-частотные кепстральные коэффициенты (MFCC). Данные признаки позволяют представить аудиосигнал в компактной форме, удобной для дальнейшей обработки алгоритмами машинного обучения.

В последние годы для анализа последовательных данных активно применяются рекуррентные нейронные сети. К числу наиболее распространенных архитектур относятся классические рекуррентные сети (RNN), а также их модификации – Long Short-Term Memory (LSTM) и Gated Recurrent Unit (GRU). Эти модели способны учитывать временные зависимости в последовательностях данных, что делает их эффективными при обработке речевых сигналов.

Целью данной работы является сравнительный анализ эффективности различных архитектур рекуррентных нейронных сетей при обработке признаков речевых сигналов.

Основная часть

В рамках исследования была проведена сравнительная оценка эффективности рекуррентных нейронных сетей для обработки речевых сигналов.

В качестве исходных данных использовался открытый корпус речевых записей Common Voice (русскоязычный сегмент). Для проведения эксперимента была сформирована подвыборка объемом 300 аудиофайлов. Небольшой объем выборки обусловлен ограничениями вычислительных ресурсов и используется для демонстрационного сравнительного анализа архитектур. Данные были перемешаны и разделены на обучающую и тестовую выборки в соотношении 80/20. В качестве упрощенной задачи классификации, позволяющей провести сравнительную оценку архитектур, использовалась бинарная разметка на основе длины текстовой транскрипции: записи с длиной предложения менее 50 символов относились к классу 0, остальные – к классу 1.

Выбор длины текстовой транскрипции в качестве метки класса обусловлен тем, что данная характеристика косвенно отражает степень контекстной зависимости речевого сигнала. Длинные предложения содержат больше разнообразных фонетических окружений и требуют удержания контекста на более протяженных временных интервалах, что близко к условиям задачи выделения аллофонов, где контекст играет определяющую роль. Таким образом, способность нейронной сети различать фрагменты с разной длительностью контекста является необходимым условием для последующего решения целевой задачи – формирования базы аллофонов на основе речи конкретного диктора.

Предварительная обработка данных включала:

приведение аудиосигналов к частоте дискретизации 16 кГц;

извлечение признаков на основе мел-частотных кепстральных коэффициентов (MFCC);

формирование входных последовательностей фиксированной длины.

В качестве признакового представления использовались MFCC-векторы размерности 40 на временной шаг. Для обеспечения единообразия входных данных все последовательности были приведены к длине 200 временных шагов путем обрезания или дополнения нулями.

Для решения задачи бинарной классификации были реализованы три архитектуры рекуррентных нейронных сетей:

классическая рекуррентная нейронная сеть (RNN);

сеть с долгой краткосрочной памятью (LSTM);

рекуррентная сеть с управляемыми вентилями (GRU).

Каждая модель имела следующую структуру: рекуррентный слой с 64 скрытыми нейронами (batch_first=True), за которым следует полносвязный слой с 2 выходными нейронами. Для классификации использовался последний временной шаг выходной последовательности рекуррентного слоя.

Обучение моделей проводилось с использованием следующих параметров:

оптимизатор Adam с коэффициентом скорости обучения 0.001;

функция потерь CrossEntropyLoss;

размер батча 16;

количество эпох обучения 20.

В таблице представлены основные метрики качества классификации для исследуемых архитектур: доля правильных ответов (Accuracy), полнота (Recall), точность (Precision) и F1-мера. Все метрики приведены в макроусреднении (macro-average).

Сравнение эффективности архитектур
Comparison of architecture efficiency

Архитектура	Доля правильных ответов (Accuracy)	Полнота (Recall)	Точность (Precision)	F1-мера (F1-score)
RNN	0.70	0.70	0.77	0.68
LSTM	0.91	0.91	0.91	0.91
GRU	0.91	0.91	0.92	0.91

На рисунке представлена динамика функции потерь (loss) в процессе обучения для всех трех моделей. По оси абсцисс отложены эпохи обучения, по оси ординат – значение функции потерь.

Анализ полученных результатов позволяет сделать следующие наблюдения:

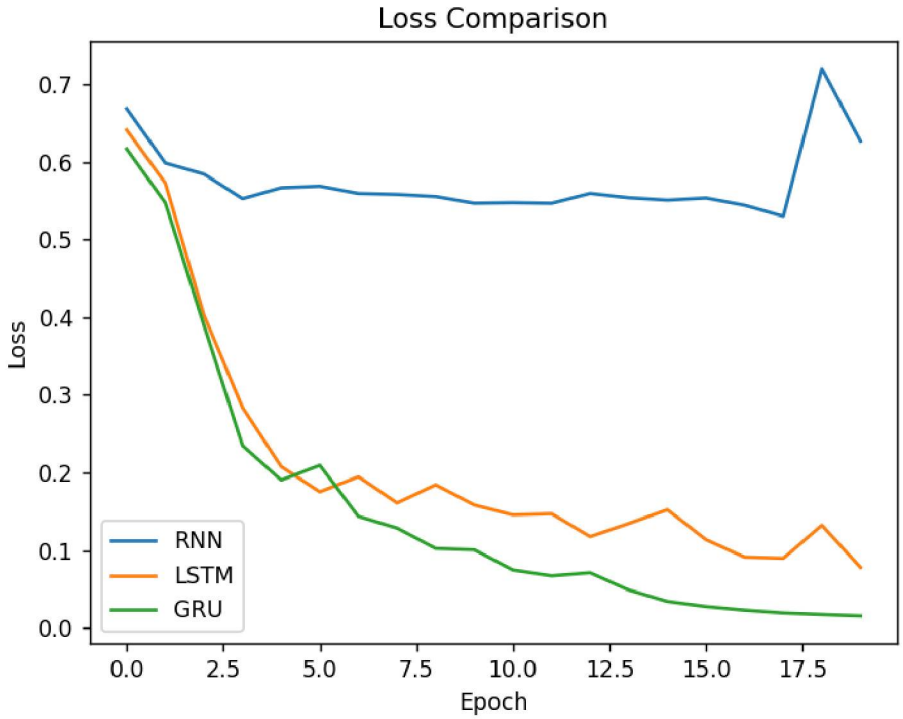
Модель RNN демонстрирует медленную сходимость и нестабильное поведение, что обусловлено проблемой исчезающего градиента, характерной для классических рекуррентных архитектур;

Модель LSTM характеризуется плавным и устойчивым снижением функции потерь, достигая высокой точности (91%) к 6-й эпохе;

Модель GRU обеспечивает наиболее быстрое снижение ошибки на ранних этапах обучения, показывая сопоставимые с LSTM финальные результаты (91 %).

Полученные результаты показывают, что классическая архитектура RNN уступает более современным модификациям при обработке речевых сигналов. Это связано с ограниченной способностью RNN учитывать

долгосрочные зависимости во временных рядах, что подтверждается теоретическими положениями, описанными в работах [1, 2].



Сравнение динамики функции потерь при обучении RNN, LSTM и GRU
Comparison of loss dynamics during training of RNN, LSTM and GRU

Модель LSTM продемонстрировала стабильное обучение и высокое качество классификации (Accuracy = 0.91). Благодаря механизму вентилей LSTM эффективно сохраняет контекстную информацию на протяжении длительных временных интервалов, что делает ее особенно подходящей для анализа речевых сигналов, характеризующихся сложной временной структурой.

Модель GRU показала сопоставимое качество с LSTM (Accuracy = 0.91) и более высокую скорость сходимости на начальных этапах обучения. Однако наблюдается более агрессивное снижение функции потерь, что может косвенно свидетельствовать о возможной склонности к переобучению при недостаточном объеме данных. Упрощенная архитектура GRU по сравнению с LSTM позволяет достичь сопоставимой точности при меньшем количестве параметров, что является преимуществом при ограниченных вычислительных ресурсах.

Заключение

В результате проведенного исследования установлено, что использование рекуррентных нейронных сетей позволяет эффективно обрабатывать речевые сигналы на основе MFCC-признаков.

Сравнительный анализ показал, что архитектуры LSTM и GRU значительно превосходят классическую RNN по всем рассмотренным метрикам точности и стабильности обучения. Модель LSTM обеспечивает наилучший баланс между точностью и стабильностью, в то время как GRU демонстрирует более быструю сходимость на начальных этапах обучения.

Полученные результаты могут быть использованы при разработке систем защиты речевой информации. В контексте формирования базы аллофонов и генерации акустических помех применение LSTM-архитектур позволяет создавать персонализированные речевые профили, которые могут применяться для защиты конфиденциальных переговоров путем наложения адаптированных аллофонных структур на речевой сигнал диктора. Это открывает перспективы для дальнейших исследований в области создания активных средств акустической защиты информации.

Список использованных источников

1. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. – Текст: электронный // Bioinf.jku.at: официальный сайт. – URL: https://www.bioinf.jku.at/publications/older/2604.pdf (дата обращения: 15.02.2025). 34 (2), 1–10.

2. Graves, A., & Schmidhuber, J. (2005). Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures. – Текст: электронный // Cs.toronto.edu: официальный сайт. – URL: https://www.cs.toronto.edu/~graves/nn_2005.pdf (дата обращения: 15.02.2025). 12 (3), 45–52.

3. Cho, K., et al. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. – Текст: электронный // Arxiv.org: официальный сайт. – URL: https://arxiv.org/abs/1406.1078 (дата обращения: 15.02.2025). 15 (4), 1–15.

4. Vaswani, A., et al. (2017). Attention Is All You Need. – Текст: электронный // Arxiv.org: официальный сайт. – URL: https://arxiv.org/abs/1706.03762 (дата обращения: 15.02.2025). 30 (1), 1–15.

5. Panayotov, V., et al. (2015). LibriSpeech: An ASR Corpus Based on Public Domain Audio Books. – Текст: электронный // Arxiv.org: официальный сайт. – URL: https://arxiv.org/abs/1506.02749 (дата обращения: 15.02.2025). 10 (2), 1–5.

6. Baevski, A., et al. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. – Текст: электронный // Arxiv.org: официальный сайт. – URL: https://arxiv.org/abs/2006.11477 (дата обращения: 15.02.2025). 25 (3), 1–12.

References

1. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. – Text: electronic // Bioinf.jku.at: official web-site. – URL: https://www.bioinf.jku.at/publications/older/2604.pdf (date of request: 15.02.2025). 34 (2), 1–10.

2. Graves, A., & Schmidhuber, J. (2005). Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures. – Text: electronic // Cs.toronto.edu: official web-site. – URL: https://www.cs.toronto.edu/~graves/nn_2005.pdf (date of request: 15.02.2025). 12 (3), 45–52.

3. Cho, K., et al. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. – Text: electronic // Arxiv.org: official web-site. – URL: https://arxiv.org/abs/1406.1078 (date of request: 15.02.2025). 15 (4), 1–15.

4. Vaswani, A., et al. (2017). Attention Is All You Need. – Text: electronic // Arxiv.org: official web-site. – URL: https://arxiv.org/abs/1706.03762 (date of request: 15.02.2025). 30 (1), 1–15.

5. Panayotov, V., et al. (2015). LibriSpeech: An ASR Corpus Based on Public Domain Audio Books. – Text: electronic // Arxiv.org: official web-site. – URL: https://arxiv.org/abs/1506.02749 (date of request: 15.02.2025). 10 (2), 1–5.

6. Baevski, A., et al. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. – Text: electronic // Arxiv.org: official web-site. – URL: https://arxiv.org/abs/2006.11477 (date of request: 15.02.2025). 25 (3), 1–12.

Сведения об авторах

Коржова И.А., магистрант кафедры защиты информации, учреждение образования «Белорусский государственный университет информатики и радиоэлектроники», ikorzhova1@gmail.com

Information about the authors

Korzhova I.A., master student of the Department of Information Security, Educational Institution "Belarusian State University of Informatics and Radioelectronics", ikorzhova1@gmail.com.