DOI:

РАСПОЗНАВАНИЕ ЭМОЦИЙ ИЗ РЕЧИ ЧЕЛОВЕКА С ПРИМЕНЕНИЕМ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ

Щетинин Е.Ю., Пьянков Г.И.

Финансовый университет при правительстве Российской Федерации,
Россия, г. Москва, Ленинградский пр-т, д.49
riviera-molto@mail.ru

Аннотация: В работе исследованы архитектуры глубоких нейронных сетей для распознавания эмоций человека по голосу. В качестве моделей глубоких нейронных сетей были использованы глубокие сверточные сети, а также рекуррентные нейронные сети с LSTM ячейкой памяти. Проведены компьютерные эксперименты по применению построенных нейронных сетей для распознавания эмоций в речи человека, содержащихся в базах RAVDESS. Полученные результаты показали высокую эффективность выбранного подхода, а оценки точности по отдельным эмоциям составили 90%.

Ключевые слова: паралингвистический анализ, классификация эмоций, рекуррентные нейронные сети, BLSTM модель.

Введение

Паралингвистика – область лингвистики, объектами исследований которой являются различные невербальные аспекты речи, такие как, эмоции, интонации, особенности произношения и другие характеристики голоса человека [1]. Компьютерная паралингвистика является одной из самых актуальных и динамично развивающихся областей современных речевых технологий, а распознавание эмоций в речи (РЭР) человека является наиболее востребованной их частью [2, 3]. Компьютерная классификация эмоций ставит перед собой задачу выделения признаков эмоциональной речи человека на основе аудиозаписей, видеозаписей людей, произнесших данное высказывание, и других модальностей. Сложность поставленной задачи состоит в необходимости определения таких признаков, которые являются достаточно устойчивыми к выбросам и шуму, при этом сохраняя все основные характеристики и особенности голоса. Также, используемая модель должна учитывать динамику признаков во времени для эффективного анализа изменений в голосе.

Наиболее распространенными методами моделирования и классификации в области РЭР являются смеси гауссовских распределений (Gaussian Mixture Models, GMM), скрытые марковские модели (Hidden Markov Models, HMM), метод опорных векторов (Support Vector Machines, SVM) и искусственные нейронные сети (Artificial Neural Networks, ANN) [4, 5]. С появлением методов глубокого обучения и созданием глубоких нейронных сетей (Deep Neural Networks, DNN) исследования в области компьютерного анализа эмоций приобрели качественно новое направление развития.

В настоящей работе предложена компьютерная паралингвистическая модель распознавания эмоций на основе двунаправленной рекуррентной нейронной сети с LSTM ячейкой памяти. На базе аудиозаписей RAVDESS проведены компьютерные эксперименты по классификации эмоций с использованием предложенной модели и сравнительный анализ полученных результатов с другими моделями нейронных сетей.

1 Описание и предварительная обработка используемых данных

В настоящей работе проведены компьютерные исследования базы RAVDESS, содержащей эмоциональную речь человека. RAVDESS - набор данных, содержащий 7356 файлов (общий размер: 24,8 ГБ). Каждый из 24 актеров (мужчины и женщины) состоит из трех форматов модальности: только аудио (16 бит, 48 кГц .wav), аудио-видео (720p H. 264, AAC 48kHz,. mp4) и только видео (без звука) [9]. Записи содержат следующие эмоции: 0 – нейтральный, 1 – спокойствие, 2 –счастье, 3 – грусть, 4 – злость, 5 –испуг, 6 – отвращение, 7 – удивление. Всего существует 16 классов (8 эмоций, разделенных на мужские и женские.) в общей сложности для 1440 образцов (только речь). Подробное описание данных можно найти в работе [6]. Прежде чем переходить к обучению алгоритмов необходимо провести извлечение признаков из описанных выше данных, необходимых для передачи их в нейронную сеть. В настоящей работе использовались акустические признаки, полученные в результате анализа длительности звукового сигнала и его спектра, а также мел-кепстрального анализа аудиосигнала [7].

2 Исследование нейронных сетей для распознавания эмоций

В работе проведены компьютерные исследования различных моделей нейронных сетей для классификации эмоций на примере описанных выше данных. Для этого были использованы сверточные сети CNN, а также рекуррентные сети с LSTM-ячейкой памяти [8]. В работе [9] была предложена модель двунаправленной рекуррентной нейронной сети с LSTM-ячейкой памяти. Двунаправленные LSTM сети работают в обоих направлениях, комбинируя выходные данные двух скрытых LSTM-слоёв, передающих информацию в противоположных направлениях — один по ходу времени, другой против него, и, тем самым, одновременно получая данные из прошлого и будущего состояний. На Рис.1 приведена схема двунаправленной рекуррентной нейронной сети с LSTM ячейками памяти (BLSTM-архитектура). Метрикой для обучения и тестирования моделей была выбрана точность классификации (Accuracy). Для того, чтобы применить сверточные сети, звук представляют в виде спектрограмм в линейной или мел-шкале, после чего с полученными спектрограммами оперируют как с обычными двумерными изображениями.

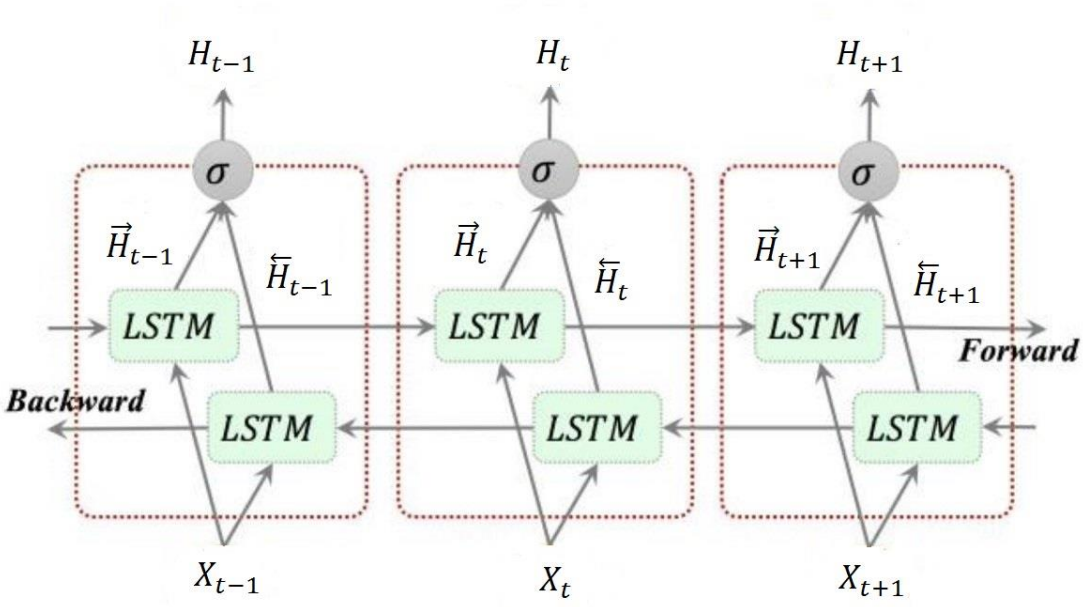

Рис.1. Архитектура BLSTM сети.

Для описанного выше набора данных модель глубокой сверточной сети CNN оказалась наиболее точной по сравнению с LSTM моделью. Она имеет следующие характеристики: 18 скрытых слоев, softmax-функция активации, 32 batch size, 1000 epochs. Ее точность на тестовой выборке составила 71%, тогда как многослойная нейронная сеть, имеющая характеристики 10 скрытых слоев и 32 batch size, 1000 epochs, достигла лишь около 43% точности. Также была применена модель рекуррентной сети с LSTM ячейкой и параметрами: 5 скрытых слоев, 32 batch size, 1000 epochs, softmax-функция активации. Эксперименты с ней показали 66.7% точности на тестируемых данных. Как видим, точность наилучшей модели нейронной сети оказалась не слишком высокой и едва превысила 70%. Аналогичные исследования этих данных, использующие глубокие нейронные сети, например, [10],[11], сообщают о достигнутой точности классификации 58.6% и 64% соответственно.

В дальнейших экспериментах из анализируемых данных были удалены записи, содержащие эмоции отвращение, удивление и нейтралитет для обоих полов, что привело к 10 классам эмоций. Это позволило повысить точность классификации до 77%. Также, для этой базы данных были проведены компьютерные эксперименты по классификации пола актера, и, кроме того, позитивности (негативности) высказываемых эмоций. В этих случаях, точность классификации составила 91.4% и 93.7% соответственно. Очевидно, что сокращение классов эмоций или их бинаризация приводит к ожидаемому существенному повышению точности классификации.

Заключение

Эмоции играют важную роль в человеческих коммуникациях, отличаются сложностью и оказывают значительное влияние на процессы принятия решений в различных сферах деятельности

человека. Эмоциональная речь плохо поддается научному пониманию и ее трудно встроить в процедуры автоматизации технологических процессов. Вопрос применения искусственного интеллекта к распознаванию эмоций в реальном мире остается открытым. Прежде всего, это связано с неоднозначностью в оценке эмоциональной речи. Некоторые высказывания могут быть по-разному классифицированы экспертами, следовательно, неоднозначно размечены в корпусе данных. В целом задача автоматического распознавания эмоций еще далека от решения, несмотря на то, что в последние годы были достигнуты значительные успехи в этой области.

Литература

1. *Rabiner L., Juang B.* Fundamental of Speech Recognition. Englewood Cliffs: Prentice-Hall N.J.,1993.
2. *Schuller B.* The Computational Paralinguistics Challenge // IEEE Signal Processing Magazine, Vol. 29, 2012, № 4, -P. 1264-1281.
3. *Schuller B., Batliner A.* Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing, Wiley, 2013.
4. *Карпов А.А., Кайа Х., Салах А.А.* Актуальные задачи и достижения паралингвистического анализа речи // Научно-технический вестник информационных технологий, механики и оптики, 2016, т.16(4), с. 581–592.
5. *Batliner A., Schuller B.* Computational Paralinguistics. Emotion, Affect and Personality in Speech and Language Processing. John Wiley & Sons Limited. 2015.
6. *Livingstone S.R., Russo F.A.* The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. // PLoS ONE, 2018, 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.
7. *Hasan R., Jamil M., Rabbani G., Rahman S.* Speaker identification using mel frequency cepstral coefficients // 3rd International Conference on Electrical and Computer Engineering, 2004. P. 28–30.
8. *Hochreiter S., Schmidhuber J.* Long Short-Term Memory // Neural Computation, 9(8) 1997- P.1735-1780.
9. *Schuster M., Paliwal Kuldip K.* Bidirectional Recurrent Neural Networks IEEE transactions on signal processing, Vol. 45, 1997, 11.
10. *Popova A., Rassadin A., Ponomarenko A.* Emotion Recognition in Sound, in: Advances in Neural Computation, Machine Learning, and Cognitive Research // Selected Papers from the XIX International Conference on Neuroinformatics, October 2-6, 2017, Moscow, Russia, Cham: Springer International Publishing, Vol. 736, 2017. P. 117-124.
11. *Стерлинг Г., Приходько П.* Глубокое обучение в задаче распознавания эмоций из речи. Труды конференции «Информационные технологии и системы 2016» // ИППИ РАН. – 2016. – С. 451-456.