УДК 004.934.2

А. С. Попова, А. Г. Рассадин, А. А. Пономаренко

Детектирование эмоций в речи с использованием долгой краткосрочной памяти

Национальный исследовательский университет Высшая школа экономики – Нижний Новгород

Рассматривается задача автоматической классификации эмоций в цифровом аудио сигнале. В работе рассматривается и верифицируется подход, в котором классификация звукового фрагмента производится с помощью рекуррентной нейронной сети с долговременно-кратковременной памятью. В качестве признаков использовались мел-кепстральные коэффициенты. Произведен численный эксперимент на открытом наборе данных Ravdess, включающий 8 различных эмоций: “нейтральный”, “спокойный”, “счастливый”, “грустный”, “злой”, “испуганный”, “отвращение”, “удивление” и проведено сравнение разных наборов признаков и разных архитектур сети.

Ключевые слова: глубокое обучение, классификация, мел-кепстральные коэффициенты, рекуррентные нейронные сети, нейронные сети с долговременно-кратковременной памятью, анализ звуков, анализ речи, эмоции.

Введение

Классификация человеческих эмоций в потоке мультимедийных данных актуальная и активно разрабатываемая область компьютерных наук. Задача классификации эмоций имеет большой потенциал к использованию во многих прикладных отраслях, таких как робототехника, системы слежения и других системах в которых происходит интерактивное взаимодействия с пользователем. Решение этой задачи позволяет получить от пользователя обратную связь естественным образом, без требования со стороны пользователя каких-либо дополнительных действий, тем самым упрощая и ускоряя взаимодействия вычислительной техники и человека.

С математической точки зрения задачу классификации можно рассматривать в качестве построения функции, $y : X \to Y$, где X – это множество всевозможных описаний объектов, Y – конечное множество классов. Таким образом существует неизвестная целевая зависимость y – отображение, значения которой известны только на объектах конечной обучающей выборки $X^m = \{(x_1, y_1), ..., (x_m, y_m)\}$. Требуется построить алгоритм $A : X \to Y$, способный классифицировать произвольный объект $x \in X$. Для нашей задачи искомое отображение y, будет выглядеть как $y : R^n \to Y$, где n – в случае распознавания аудио сигнала, число сэмплов в окне распознавания.

Существуют две основные теории описания эмоций: дискретная теория [11] основывается на существовании универсальных базовых эмоций, которые могут различаются количеством и типом выделяемых базовых эмоций; пространственная теория предполагает, что эмоции раскладываются на базисы, тем самым эмоцию можно представлять как точку в векторном пространстве [12,13]. Как правило выделяют 6 основных эмоциональных состояний: нейтральное состояние, злость, страх, счастье, печаль и удивление. В данной работе мы будем придерживаться дискретной теории.

Ранее в литературе был предложен ряд методов классификации эмоций человека как отдельно для аудио или изображений [16,17], так и одновременно для видео. Большая часть этих методов связанна с явным выделением признаков, функция обнаружения которых в явном виде присутствует в алгоритме. Например, на изображениях ищут улыбку, оценивают положение и форму рта, широту глаз, угол бровей. В свою очередь в звуковом сигнале оценивают уровень энергии, средний уровень, дисперсию, изменение высоты голоса.

В данной работе мы, вдохновленные последними успехами в области рекуррентных нейронных сетей, ставим своей целью улучшить прошлый результат, полученный на основе последних разработок в области компьютерного зрения и сверточных нейронных сетей [1][2].

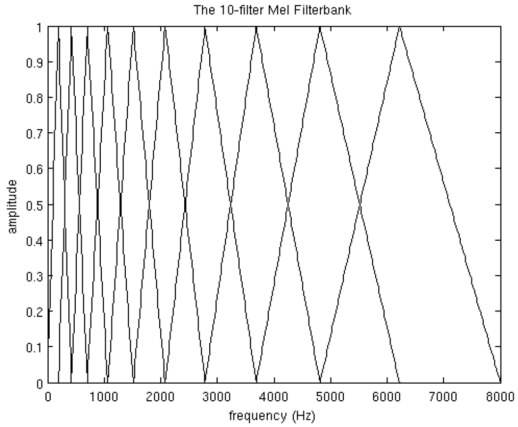

Рис 1. Визуализация мелфильтров

На сегодняшний день уже успешно решены множество задач, связанных с обработкой звука, существует множество алгоритмов работы со звуковыми файлами и методов их классификации, имеющей различную точность. Тем не менее, сравнение алгоритмов классификации звука является условным, так как во многих случаях эксперименты проводились на различных тестовых данных и с различными задачи распознавания.

Базовой техникой обработки аудио сигнала является быстрое преобразование Фурье [9][10]. Например, популярные алгоритмы J. Haitsma[8] и A. Wang [7] оба основаны на анализе частотно-временных признаков, полученных с помощью оконного преобразования Фурье.

Оба алгоритма хорошо себя проявили на задаче распознавания музыкальных треков по предъявленному фрагменту имея около 75% точности. Более того, эти методы извлечения признаков используются в мобильном приложении Яндекс.Музыка [6] в разделе «Распознавание», а так же в широко известном мобильном приложении «Shazam» поиска музыкального произведение по записанному фрагменту.

Подобный подход, на основе которого строилось визуализация быстрого преобразования Фурье и использовались сверточные нейронные сети, был опробован мной ранее и описан в статьях [1][2].

На данный момент для наилучшего результата были использованы мел-кепстральные коэффициенты [3][5] и рекуррентная нейронная сеть с долгосрочной и краткосрочной памятью (LSTM) [4].

О этом поподробнее. Первым шагом в любой задаче машинного обучения является извлечение признаков. Одним из классических подходов для работы с записью человеческой речи является выделение мел-кепстральных коэффициентов. Основная идея работы с речью человека заключается в том, что звуки, генерируемые человеком, фильтруются формой голосового тракта, включая язык, зубы и т. д. Набор этих характеристик и можно представить с помощью мел-кепстральных коэффициентов . Мел — психофизическая единица высоты звука, количесвенная оценка звука по высоте, основанная на статистической обработке большого числа данных о субъективном восприятии человеком высоты звуковых тонов. Зависимость мела от герца логарифмическая: $\mathrm{m} = 2595\ \log_{10}(1 + \frac{\mathrm{h}}{700}) = 1127\ \ln(1 + \frac{\mathrm{h}}{700})$, где m – количество мелов, а h – количество герц.

Этапы вычисления мел-кепстральных коэффициентов:

1) Деление сигнала на короткие части (около 25мс длиной каждая)
2) Вычисление оценки периодограммы спектра мощности для каждой части (быстрое преобразование Фурье)[9][10].
3) Применение мелфильтров к спектрам мощности и суммирование энергию в каждом фильтре. Мелфильтры представляют из себя набор из 20-40 треугольных фильтров
4) Вычисление логарифма всех энергий.

5) Вычисление дискретного косинусного преобразования энергии для каждого блока фильтра.
6) Необходимы 2-13 коэффициенты этого преобразования

Поскольку традиционные нейронные сети не способны хранить информацию о предыдущих входах, а идея вынесения вердикта об эмоции в аудиофайле на основе предыдущей информации является ключевой в нашем решении, было решено перейти к рекуррентным сетям[18], а именно сетям LSTM. Для лучшего понимания РНС можно представить, как множество копий одной и той же нейронной сети, каждая из которых передает информацию последующей. Структура, имеющая вид цепочки, говорит о том, что РНС тесно связаны с последовательностями. За последние несколько лет РНС с большим успехом были использованы для решения различных задач, таких как распознавание речи, моделирование языка, перевод, создание описаний к изображениям и др. Большая часть успехов была достигнута с помощью особого типа РНС, называемого LSTM-сетью (long short-term memory, долговременно-кратковременная память), который при решении различных задач значительно превосходит стандартный вариант.

Одной из ключевых идей РНС является то, что они должны быть способны использовать предшествующую информацию для решения текущей задачи, например, использовать информацию о предыдущем кадре видео для обработки текущего кадра. В теории, РНС должны справляться с «долговременными зависимостями», однако в отношении практических задач, к сожалению, реализовать это невозможно[15][16]. В рамках этих исследований были обнаружены фундаментальные причины подобного рода ограничений РНС. Однако, LSTM-сети лишены такого недостатка.

LSTM-сеть [17] – это особый тип РНС, способный обучаться долговременным зависимостям. Такие сети прекрасно справляются с решением многих задач и находят широкое применение в данное время.

LSTM-сети разработаны специально для того, чтобы решить проблему долговременных зависимостей.

Все РНС имеют форму цепочки повторяющихся модулей. Повторяющийся модуль стандартной РНС имеет очень простую структуру, например, единственный tanh-слой (функция активации – гиперболический тангенс). LSTM-сеть представляет собой аналогичную цепочку, но повторяющийся модуль имеет другую структуру. Вместо одного слоя он содержит четыре слоя, которые взаимодействуют особым образом.

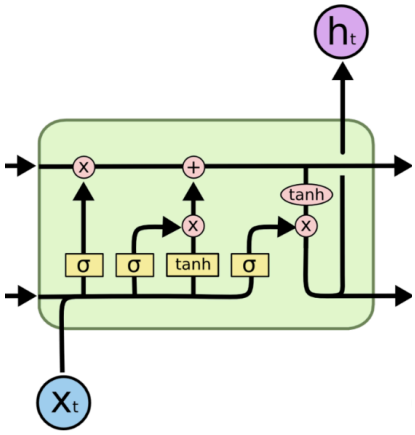

Рис 2. Слой сети LSTM

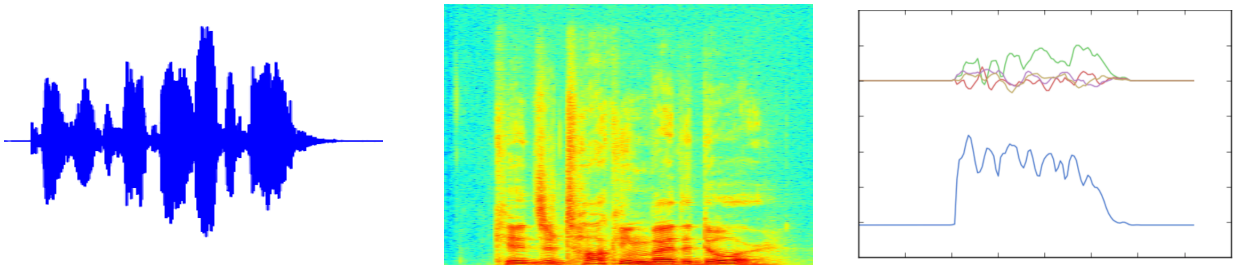

Рис. 3. Слева – осциллограмма - последовательность уровней, соответствующая колебанию мембраны микрофона, записи фразы «Kids are taking by the door» произнесенной актёром с эмоцией счастья. В середине– соответствующая ей мелспектрограмма. Справа – первые 5 мел-кепстральных коэффициентов.

Апробируемый подход

В настоящее время ряд задач стали эффективно решаться рекуррентными нейронными сетями с долгосрочной и краткосрочной памятью. С их использованием возросло качество решений задач, где необходимо обрабатывать последовательности данных, например, работа со звуком, распознавание речи, алгоритмы, работающие с текстами и многие другие. В своей работе мы пытались улучшить точность распознавания эмоций с помощью нейронной сети LSTM.

Прежде всего все аудиофайлы должны быть предворительно обработанны. А именно, отнормирована громкость (относительно максимума), применены lowpass и highpass фильтры (отсечены шумы с частотой ниже 30 Герц и выше 2700 Герц поскольку частота человеческого голоса лежит в диапазоне от 30 Герц до 2700 Герц) , каждый звуковой файл пропускался через Voice Activity Detection [19] и в конце концов были получены MFCC коэффициенты.

Данная осциллограмма в явном виде изображает последовательность значения уровней напряжения снятого с мембраны микрофона, дискретизированная через одинаковые малые промежутки времени.

Данная последовательность уровней напряжения в явном виде хранится в аудиофайлах формата wav, как последовательность двухбайтовых или трехбайтовых целых чисел, соответствующих различным 65535 либо 32 миллионам значений уровням напряжения.

Однако, как показано в работах [1][2] осциллограмма и мелспектрограма не дали очень хороших результатов классификации аудио-фрагментов на эмоции. Поэтому в этой работе мы перешли к использованию мел-кепстральных коэффициентов, которая подавались на вход классификатору. А архитектура нейронной сети была заменена на сеть LSTM, поскольку она наиболлее подходит для признаков, которые представляют собой некую последовательность связанных наблюдений.

Вычислительный эксперимент.

В качестве обучающей выборки использовался открытый размеченный набор данных “RAVDESS” [14] включающий записи 24 актеров изображающих 8 эмоций: “neutral”, “calm”, “happy”, “sad”, “angry”, “fearful”, “disgust”, “surprised” (96 экземпляров для “neutral”, 120 для “surprised” и по 192 экземпляра для остальных эмоций).

Аудиозаписи прошли описанную выше предобработку, а каждый мел-кепстральный коэффициент представлялся последовательностью из 378 чисел. Для кроссвалидации выборка делилась в соотношении 70% к 30% равномерно по классам.

Наши эксперименты показали, что наиболее оптимальной является модель с двумя LSTM слоями и одним мел-кепстральным коэффициентом (№2) поскольку процесс обучения проходил наиболее равномерно, а модель достаточно простая, имеет хорошее качество и высокую скорость обучения.

Модель	Точность
Генератор случайных чисел	12.5%
Knn + обработанный сигнал	24%
Random forest + обработанный	29%
Svm + обработанный сигнал	31%
Vgg11 + spectrogram	64%
Vgg16 + melspectrogram	71%

Таблица 1. Точность разных классификаторов

	2 mfcc	2,3 mfcc	2,3,4 mfcc
1 lstm layer	91.72%	92.19%	91.06%
2 lstm layer	97.88%	94.50%	97.27%
3 lstm layer	98.45%	99.86%	99.31%

Таблица 2. Наилучшая точность классификаторов

Заключение

В настоящей работе был предложен и апробирован подход к классификации эмоций человека в звуковом фрагменте. Проведён численный эксперимент и сравнены результаты работы различных классификаторов, сверточных сетей (VGG-11 и VGG-16) и LSTM сетей с различными наборами признаков. Данный эксперимент показал неожиданный результат – возможность классификации эмоций человека на 8 классов с точностью более 99%.

Статья подготовлена в результате проведения исследования (№ 17-05-0007) в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2018 г. и в рамках государственной поддержки ведущих университетов Российской Федерации "5-100".

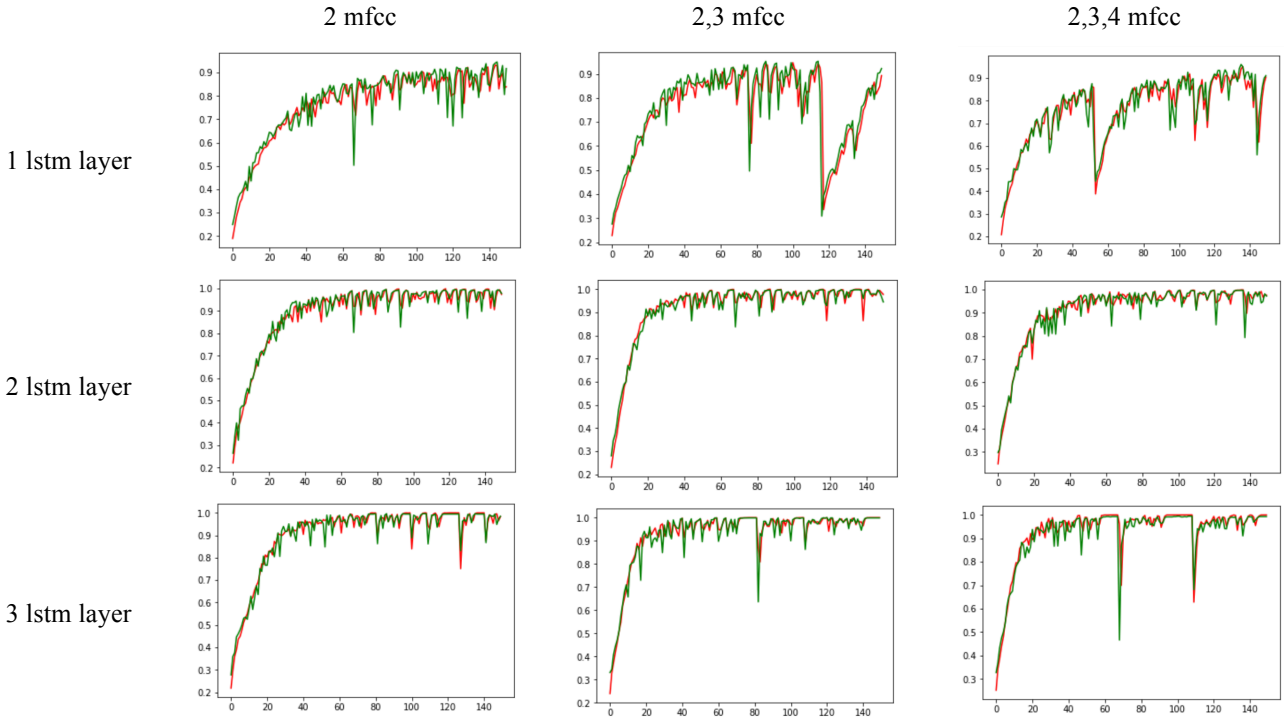

Таблица 3. Графики зависимости точности классификации от количества эпох обучения. (зеленое – обучающая выборка, красное – тестовая)

Библиографический список

1. А. С. Попова, А. Г. Рассадин, А. А. Пономаренко Детектирование эмоций в мультимедиа контенте //Международная научно-техническая конференция «Информационные системы и технологии» ИСТ-2017 , СЕКЦИЯ 5.3 Техническая кибернетика (информационное моделирование когнитивных процессов) – 2017. – С. 852-857.
2. Popova A. S., Alexandr G. Rassadin, Alexander A. Ponomarenko. 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 Vol. 736. Cham : Springer International Publishing, 2017. P. 117-124.
3. Muda, Lindasalwa, Mumtaj Begam, and Irraivan Elamvazuthi. "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques." arXiv preprint arXiv:1003.4083 (2010).
4. Karpathy, Andrej. "The unreasonable effectiveness of recurrent neural networks." Andrej Karpathy blog (2015).
5. Zheng, Fang, Guoliang Zhang, and Zhanjiang Song. "Comparison of different implementations of MFCC." Journal of Computer Science and Technology 16.6 (2001): 582-589.
6. Евгений Крофто «Как Яндекс распознает музыку с микрофона» //Доклад в рамках конференции «Yet another Conference 2013»
7. Wang A. et al. An Industrial Strength Audio Search Algorithm //ISMIR. – 2003. – Т. 2003. – С. 7-13.
8. Haitsma J., Kalker T. A highly robust audio fingerprinting system with an efficient search strategy //Journal of New Music Research. – 2003. – Т. 32. – №. 2. – С. 211-221. \
9. Oppenheim, Alan V. "Speech spectrograms using the fast Fourier transform." IEEE spectrum 7.8 (1970): 57-62.
10. Cooley J. W., Tukey J. W. An algorithm for the machine calculation of complex Fourier series //Mathematics of computation. – 1965. – Т. 19. – №. 90. – С. 297-301. Ortony A., Turner
11. T. J. What's basic about basic emotions? //Psychological review. – 1990. – Т. 97. – №. 3. – С. 315.
12. Scherer K. R. What are emotions? And how can they be measured? //Social science information. – 2005. – Т. 44. – №. 4. – С. 695-729.
13. Russell J. A., Ward L. M., Pratt G. Affective quality attributed to environments: A factor analytic study //Environment and behavior. – 1981. – Т. 13. – №. 3. – С. 259-288.
14. Livingstone S. R., Peck K., Russo F. A. Ravdess: The ryerson audio-visual database of emotional speech and song //Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science. – 2012.
15. Hochreiter, Sepp, J. urgen Schmidhuber, and Corso Elvezia. "LONG SHORT-TERM MEMORY."
16. Bengio, Yoshua, Patrice Simard, and Paolo Frasconi. "Learning long-term dependencies with gradient descent is difficult." IEEE transactions on neural networks 5.2 (1994): 157-166.
17. Hochreiter, Sepp, and Jürgen Schmidhuber. "LSTM can solve hard long time lag problems." Advances in neural information processing systems. 1997.
18. Graves, Alex, and Navdeep Jaitly. "Towards end-to-end speech recognition with recurrent neural networks." International Conference on Machine Learning. 2014.
19. Sohn J., Kim N. S., Sung W. A statistical model-based voice activity detection //IEEE signal processing letters. – 1999. – Т. 6. – №. 1. – С. 1-3.

A.S. Popova, A.G.Rassadin, A.A. Ponomarenko

EMOTION RECOGNITION IN MULTIMEDIA CONTENT

National Research University Higher School of Economics – Nizhniy Novgorod

In this paper we consider the automatic emotions recognition problem, especially the case of digital audio signal processing. We consider and verify an approach in which the classification of a sound fragment can be solved using long short-term memory neural network. The computational experiment was done based on Ravdess open dataset including 8 different emotions: "neutral", "calm", "happy," "sad," "angry," "scared", "disgust", "surprised". The best accuracy result was 99%, which was produced by using LSTM network and MFCC.

Key words: deep learning, classification, MFCC, recurrent neural networks, Long Short Term Memory networks (LSTM), audio recognition, emotion recognition, speech recognition, emotions