УДК 004.032.26:581.522-047.44

НЕЙРОСЕТЕВОЕ РАСПОЗНАВАНИЕ ЖАНРА МУЗЫКАЛЬНЫХ ПРОИЗВЕДЕНИЙ

MUSIC GENRE RECOGNITION WITH NEURAL NETWORKS

В.И. Дектярёв
Магистрант кафедры информатики БГУИР, инженер-программист

А.В. Жвакина
Доцент кафедры информатики БГУИР, кандидат технических наук, доцент

V.I.Dektiarev, A.V.Zhvakina

Белорусский Государственный Университет Информатики и Радиоэлектроники
Belarusian State University of Informatics and Radioelectronics
E-mail: *eclipter97@gmail.com, zhvakina@bsuir.by*

Аннотация. В докладе рассматривается возможность использования нейронных сетей для классификации музыкальных композиций по жанрам. Демонстрируется полученный результат и оценка решения задачи. Использовались следующие виды нейронных сетей: сети прямого распространения, сверточные сети и реккурентные сети LSTM.

Absract. The report considers the possibility of using neural networks to classify music by genre. The reached result is demonstrated along with whole solution evaluation. The following types of neural networks were used: direct propagation network, convolutional network and recurrent LSTM network.

Ключевые слова: классификация, музыка, жанр, нейронные сети, LSTM, сверточные сети.
Key words: classification, music, genre, neural networks, LSTM, convolutional networks.

В настоящее время в виду постоянно развивающихся веб-технологий и закономерно растущей численностью разнообразных сервисов, предоставляющих пользователям некоторую информацию, растёт конкуренция в различных сферах. В частности, появляется всё больше музыкальных сервисов. Для того, чтобы быть успешными на рынке, подобные средства должны предоставлять возможность не только прослушивать музыкальные композиции, но и обеспечить дополнительные функции. К примеру, полезной является возможность анализа жанровых предпочтений пользователя по прослушанным композициям для последующего составления списка рекомендаций. Для анализа композиций в разработанном программном продукте используются нейронные сети.

При данном подходе особый интерес представляет качество подготовки исходных данных. Прямой анализ звуковых сигналов во временной области может потенциально занять много времени в зависимости от длительности и качества записи, и он сам по себе не является наиболее эффективным методом, так как данный объём информации будет избыточным, если анализировать его на какие-либо общие характеристики. Составление спектрограмм и их дальнейший анализ будет быстрее, но всё же не вполне эффективным.

На текущий момент одним из наиболее рациональных представлений записи для дальнейшего анализа является метод мел-частотных кепстральных коэффициентов, которое широко применяется для составления характеристик речевых сигналов и получило широкое распространение в сфере задач распознавания речи. Данная метрика больше приближена к тому, как человек оценивает музыкальную композицию на слух, так как мы воспринимаем высоту звука в определённый момент времени, а не данные о частотах, которые являются основой спектрограмм.

Кроме того, данные, представленные в этой метрике, в среднем в 4 раза меньше по объёму, чем данные с исходных спектрограмм. Именно во столько раз удалось сжать обучающую выборку из 1000 композиций, используемых для обучения полученной модели сети.

Данные коэффициенты представляют лог мощности спектра в мел частотной области. Описывают мощность огибающей спектра, которая характеризует модель речевого тракта. Получаются путём преобразования Фурье исходного сигнала, отображения значений спектра на мел-шкалу и последующего дискретного косинусного преобразования значений на мел-шкале. Полученные значения амплитуд спектра и будут являться целевыми коэффициентами.

В настоящее время большую популярность и распространение получили сверточные нейронные сети в контексте решения задач классификации. Они работают по принципу выделения особенностей различных уровней абстракции в зависимости от слоя сети и формирования на их основе предположения о принадлежности какому-либо классу (если данное требуется для решения задачи)[1].

При этом в данных сетях используется несколько наборов весов, изначально выбираемых случайным образом и формируемых в дальнейшем в результате обучения сети [5].

В контексте решаемой задачи, кроме выделения признаков композиций на различных временных промежутках, также желательно проанализировать данные признаки в совокупности, чтобы определить жанр композиции, основываясь на её структуре и последовательной характеристики признаков, а не только по факту наличия этих признаков. Для решения подобного рода задач используются сети, способные к обучению долговременным зависимостям, один из наиболее популярных представителей которых – рекуррентные LSTM-сети [4].

Долгая краткосрочная память (англ. Long short-term memory; LSTM) − разновидность архитектуры рекуррентных нейронных сетей, предложенная в 1997 году Сеппом Хохрайтером и Юргеном Шмидхубером. Характерной особенностью LSTM-сети является возможность выполнения вычислений, характерных для компьютера, при наличии матрицы весов, рассматриваемой как программа. В отличие от традиционных рекуррентных нейронных сетей, LSTM-сеть хорошо приспособлена к обучению на задачах классификации, обработки и прогнозирования временных рядов в случаях, когда важные события разделены временными лагами с неопределённой продолжительностью и границами. В контексте поставленной задачи данное свойство особенно эффективно, так как жанр музыкальных композиций может быть опознан не только по всей композиции целиком, но также и по её ключевым фрагментам.

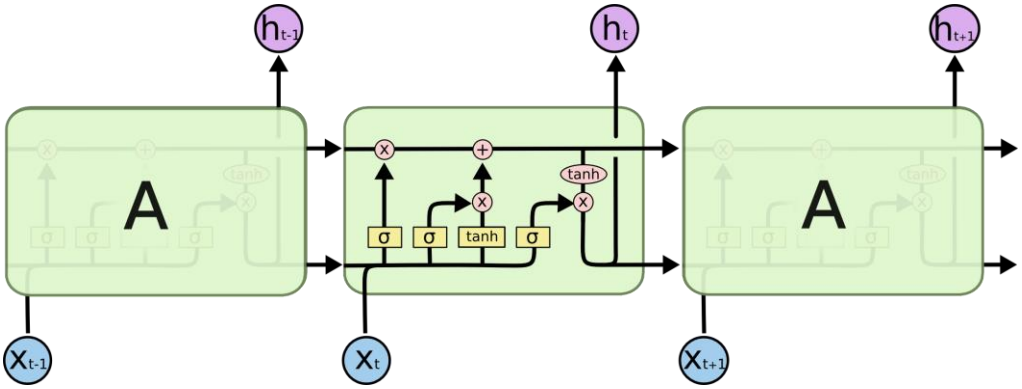


Рис. 1. Схема нейронной сети LSTM

LSTM состоит из следующих частей [2] (рис.1):

- вход (*input*) нейронной сети;
- выход (*output*) нейронной сети;
- внутреннее состояние нейронной сети или запоминающая ячейка (*memory cell*);
- фильтр очистки памяти или фильтр забывания (*forget gate*);
- входной фильтр или фильтр обновления памяти (*input gate*);
- выходной фильтр или фильтр выдачи результата (*output gate*).

На представленной выше схеме LSTM-сети каждая линия обозначает передачу вектора с выхода одного узла на входы других. Розовые круги обозначают поточечные операции, например, сложение векторов, а желтые прямоугольники – обученные слои. Слияние линий предполагает объединение, а разветвление линии говорит о том, что информация копируется, и копии направляются в разные точки назначения. LSTM-сеть имеет возможность удалять и добавлять информацию в состояние ячейки. Этот процесс регулируется специальными структурами, называемыми гейтами. Гейт – это механизм, позволяющий пропускать информацию избирательно. Он состоит из sigmoid-слоя и операции поточечного умножения. Выходом sigmoid-слоя является число от 0 до 1, которое определяет уровень пропускания. Одна ячейка имеет три гейта, управляющих ее состоянием. На первом этапе необходимо решить, какую информацию следует удалить из состояния ячейки. Это решение принимает sigmoid-слой, называемый «гейтом забывания», принимающий на вход данные из предыдущей ячейки и часть исходных данных. На выходе получается 0 или 1, что означает «забыть» или «не забыть». На следующем этапе необходимо решить, какую новую информацию следует записать в состояние ячейки. Этот этап делится на две части. На первом этапе sigmoid-слой, называемый «входным гейтом» (input gate), решает, какие значения необходимо обновить. Затем tanh-слой создает вектор новых значений-кандидатов, которые могут быть добавлены в состояние ячейки. Затем обновляется предыдущее состояние ячейки до текущего состояния. Состояние ячейки умножается на выход гейта забывания, «забывая» таким образом то, что ранее было решено «забыть». Затем прибавляется новые значения-кандидаты, отмасштабированные соответствующим образом. Наконец, необходимо решить, что следует отправить на выход. Выход будет представлять собой отфильтрованное состояние ячейки. Сначала sigmoid-слой решает, какие элементы состояния ячейки необходимо передать на выход. Затем состояние ячейки преобразуется с помощью tanh-слоя к интервалу от -1 до 1 и умножается на выход sigmoid-слоя, чтобы вывести только то, что было решено вывести [3].

Таким образом, необходимо реализовать модель нейронной сети, которая способна принимать на вход музыкальный аудиофайл и распознавать его жанр. Данная задача является

примером задачи классификации, где классы – это 10 жанров музыки: блюз, классическая, кантри, диско, хип-хоп, джаз, металл, поп, регги, рок.

В качестве обучающей выборки была использована база аудиозаписей GTZAN, представляющая собой 1000 30-секундных аудиофайлов, по 100 на каждый жанр из 10. Перед подачей записи на входной слой нейронной сети, каждая из них предварительно преобразовывалась в значения коэффициентов на мел-спектрограмме согласно методу, описанному выше.

Для решения задачи распознавания жанра музыкального аудиофайла рассматривались следующие нейронные сети:

− сеть прямого распространения, обученная для классификации входных данных в соответствии с целевыми классами (patternnet);

− самоорганизующаяся карта (selforgmap), распределяющая набор данных на количество классов, опираясь на сходство шаблонов;

− составная модель нейронной сети из комбинации сверточных и MaxPooling-слоев.

Сеть прямого распространения имеет следующую топологию (рис. 2).

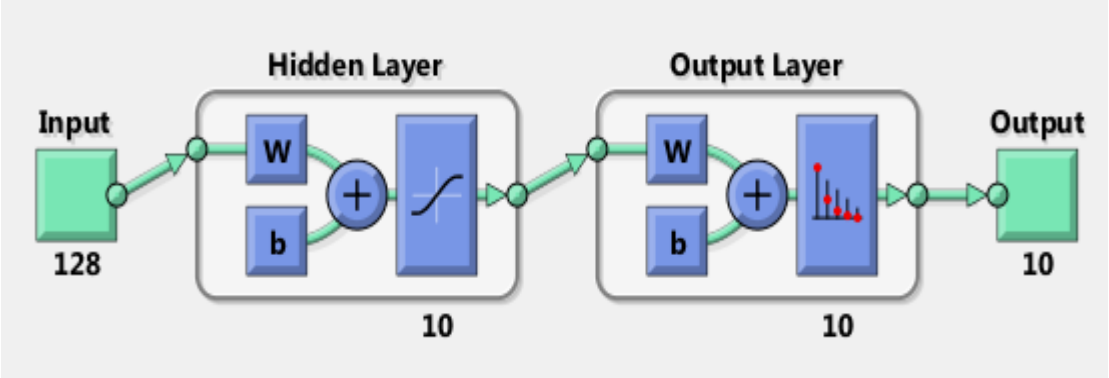


Рисунок 2

С помощью Simulink (Matlab) исследована структура сети (рис. 3).

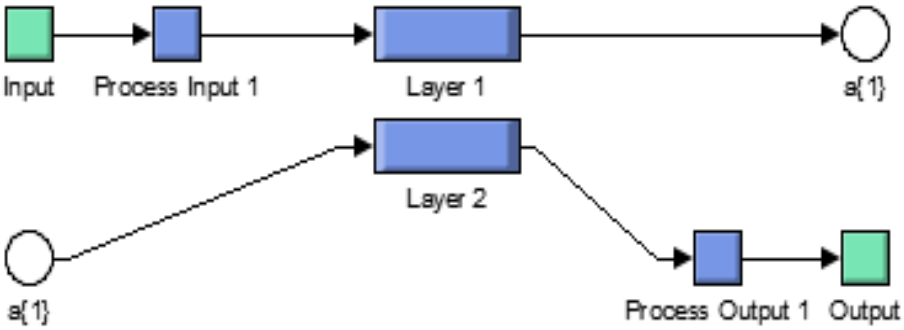


Рисунок 3

В первом слое использована функция активации tansig(рис. 4).

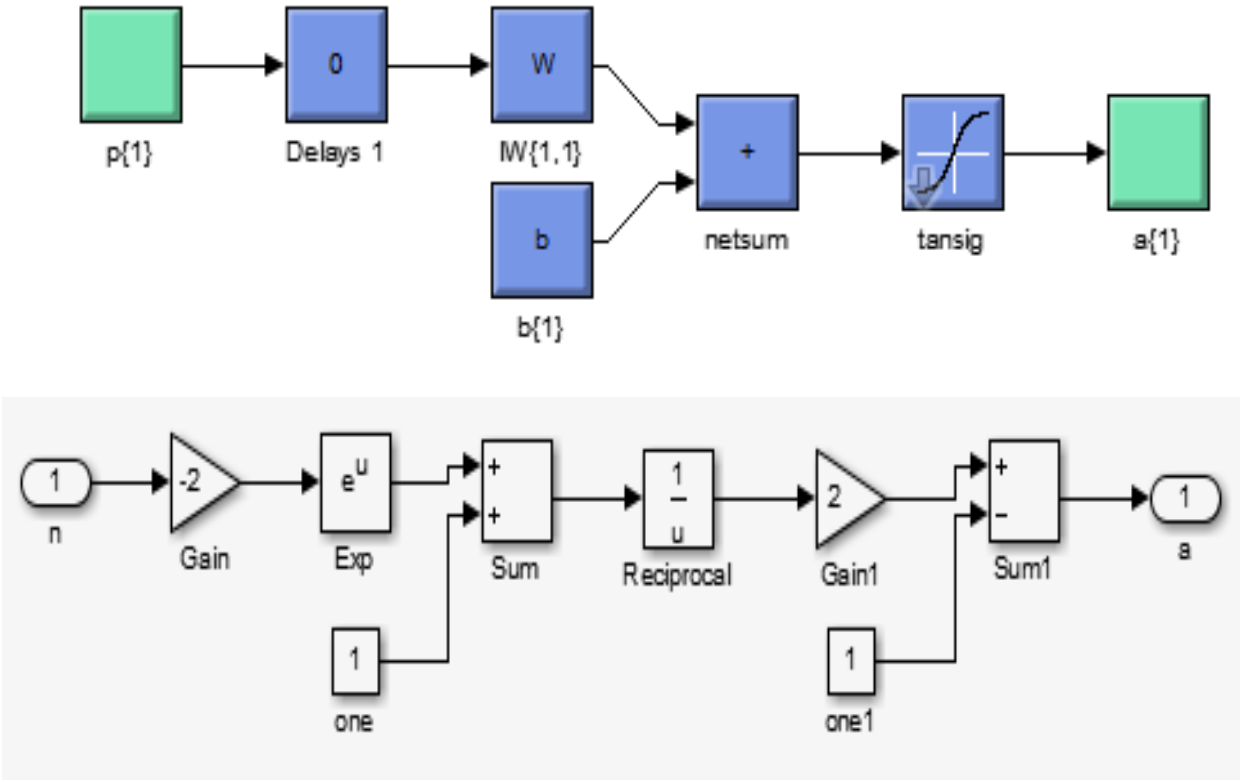
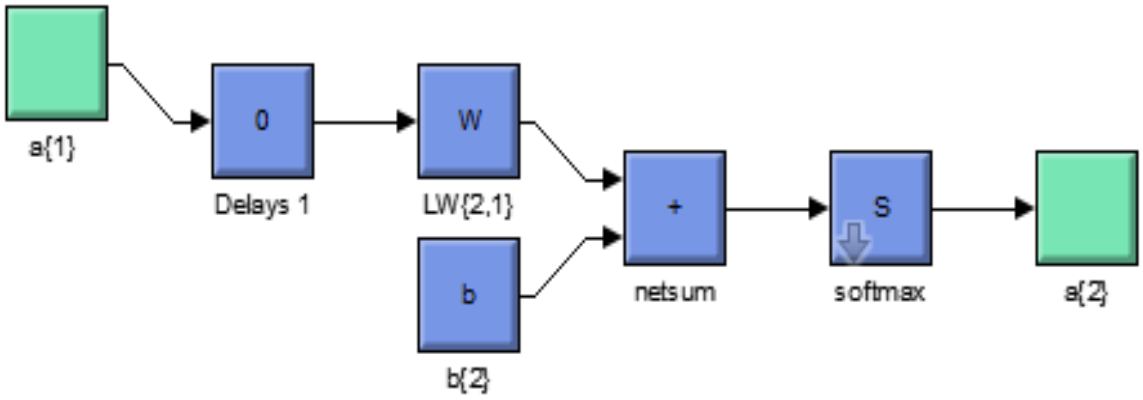


Рисунок 4

Во втором слое – функция активации softmax (рис. 5).



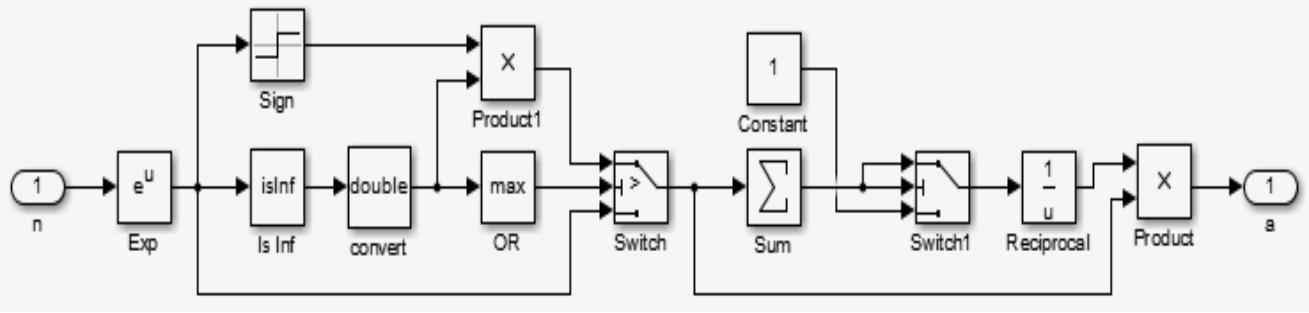
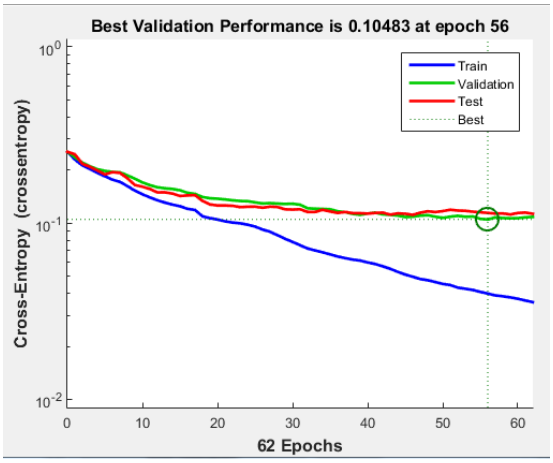
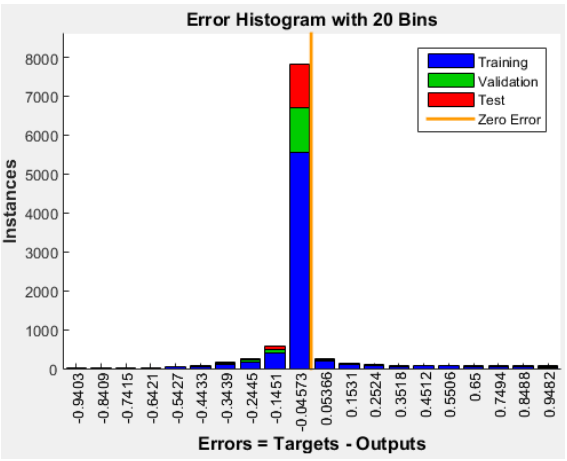
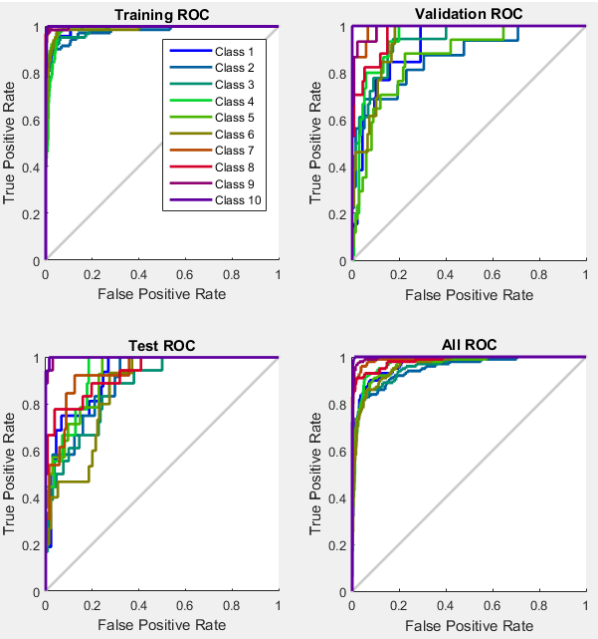


Рисунок 5

При обучении сети исследовались показатели кросс-энтропии и процент ошибок классификации. Результаты обучения, валидации и тестирования сети представлены на рис. 6.







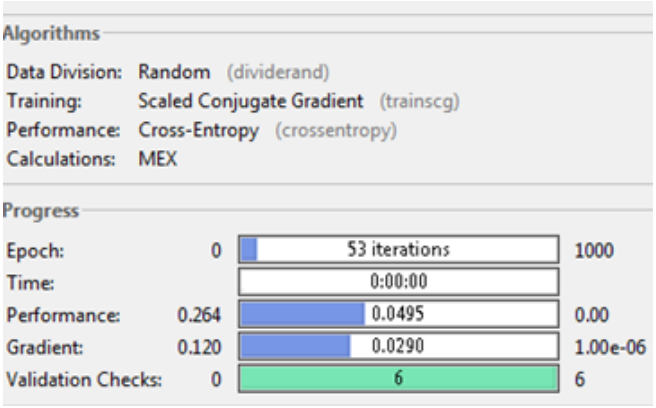


Рисунок 6

В Confusion-матрице, представленной на рис. 7, строки соответствуют прогнозируемому классу, а столбцы – истинному классу. Диагональные ячейки представляют правильно классифицированные наблюдения, остальные ячейки соответствуют неправильно классифицированным наблюдениям.

Столбец в правой части графика содержит процентное соотношение примеров, относящихся к каждому классу, которые правильно и неправильно классифицированы (истинно положительный показатель или отзыв).

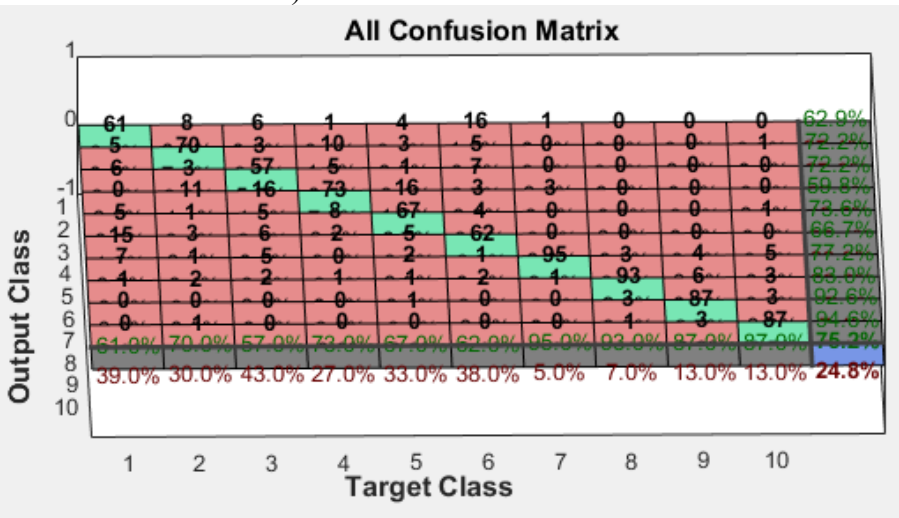


Рисунок 7

Самоорганизующаяся карта (selforgmap) имеет следующую топологию (рис. 8):

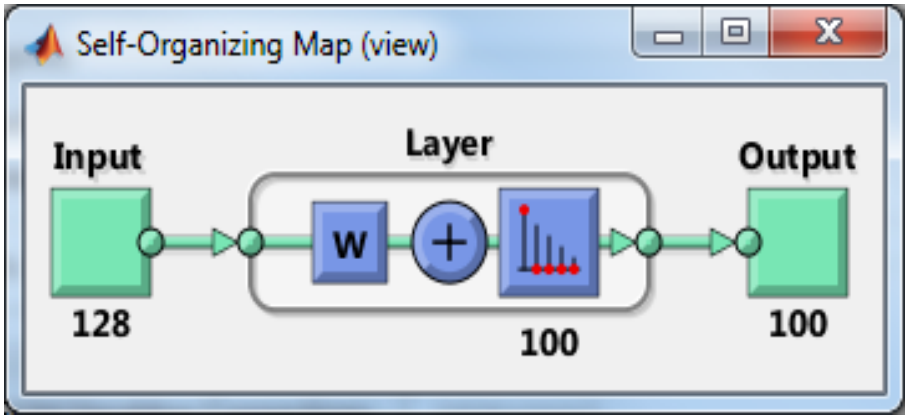


Рисунок 8

С помощью Simulink (Matlab) исследована структура сети (рис. 9).

Рисунок 9

Используется функция активации compet (рис. 10).

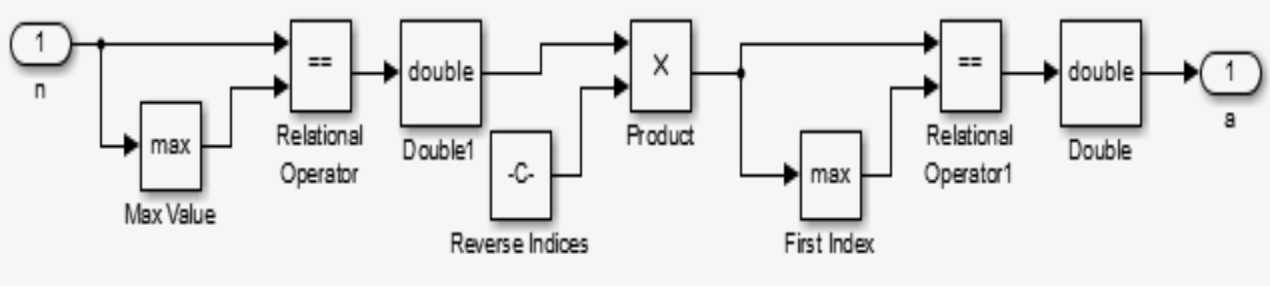


Рисунок 10

Сеть обучена с использованием пакетного алгоритма SOM (trainbu, learnsomb). Результаты работы сети представлены на рис.11.

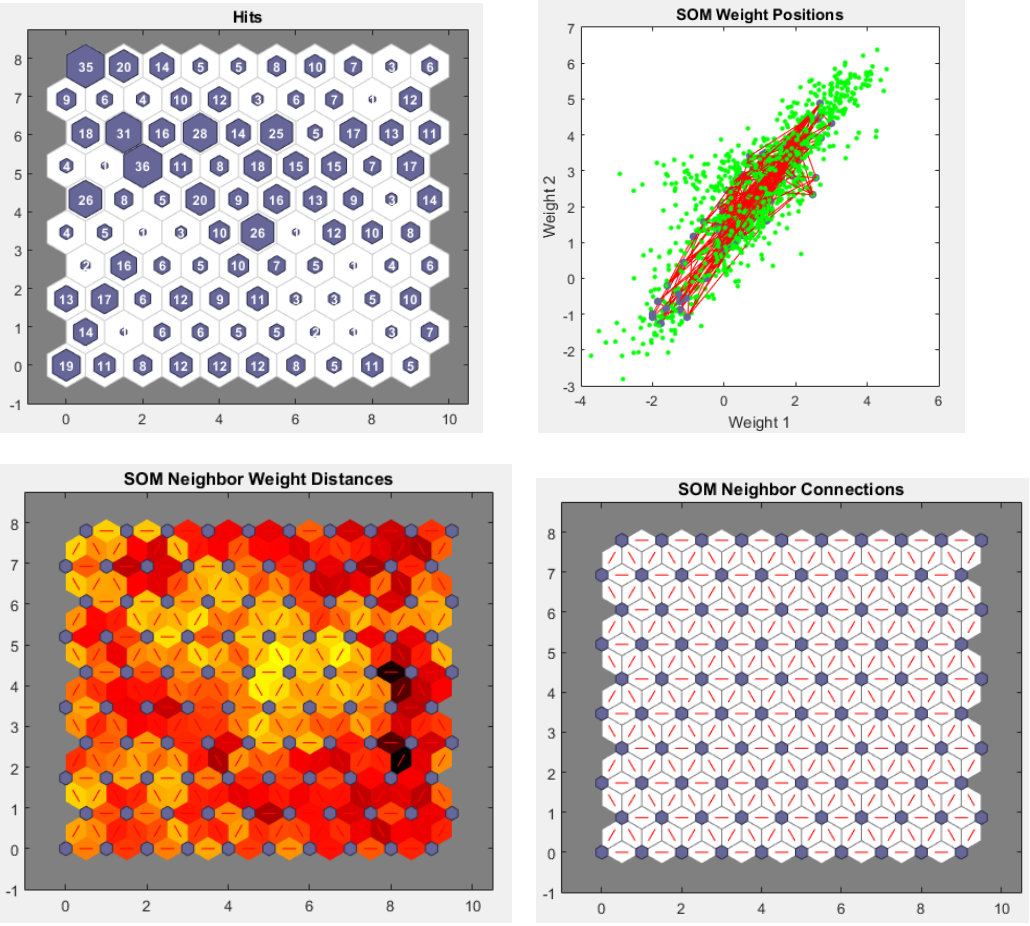


Рисунок 11

Рассмотренные выше сети могут быть использованы для решения задач классификации музыкальных произведений по жанрам и их кластеризации, однако не позволяют достичь необходимых показателей качества.

Для распознавания музыкального жанра была разработана комбинированная модель нейронной сети, состоящая из следующих слоёв (рис.12):

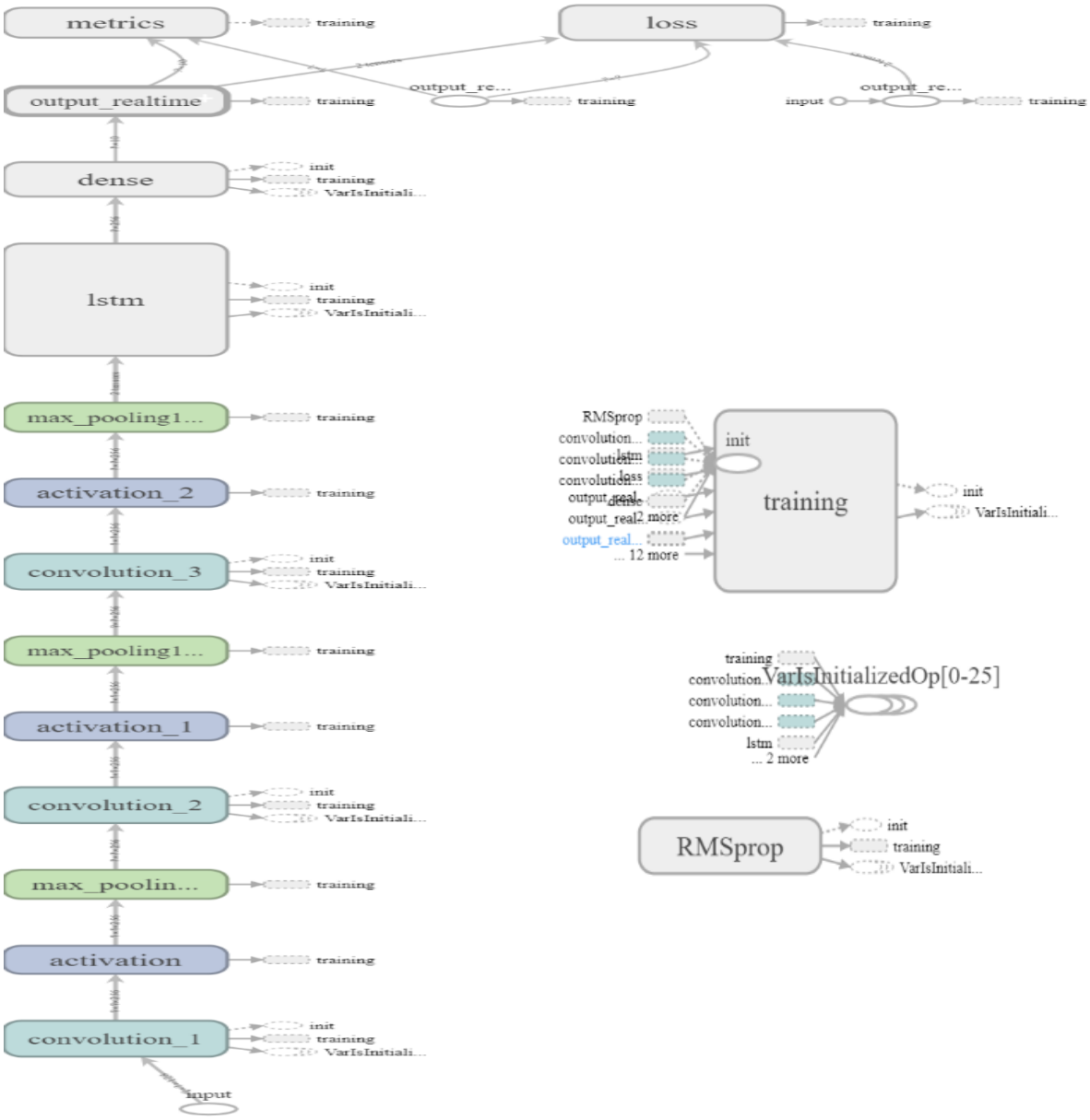


Рис. 12. Схема модели полученной нейронной сети

1. Input (входной слой): размерность – 128
2. Сверточный (1): размерность: 1, количество фильтров: 256, функция активации: ReLU, размер ядра: 5.
3. MaxPooling слой (1): размер пула: 2.
4. Сверточный (2): размерность: 1, количество фильтров: 256, функция активации: ReLU, размер ядра: 5.

5. MaxPooling слой (2): размер пула: 2.
6. Сверточный (3): размерность: 1, количество фильтров: 256, функция активации: ReLU, размер ядра: 5.
7. MaxPooling слой (3): размер пула: 2.
8. LSTM-слой: выходная размерность: 256, функция активации: гиперболический тангенс.
9. Dense-слой: выходная размерность: 10, функция активации: softmax.

Эмпирическим путём были также добавлены Dropout-слои с вероятностным коэффициентом 0.5 после каждого сверточного слоя нейронной сети, однако с ними точность сети ухудшилась на ~20% на обучающей выборке.

В качестве оптимизатора был использован метод среднеквадратичного распространения с коэффициентом обучения 0.00001. В ходе разработки архитектуры сети был также протестирован оптимизатор Adam, однако с ним точность на обучающей выборке ухудшилась на 15%.

В качестве функции потерь была использована функция перекрестной энтропии. В качестве метрики качества модели была использована «точность» (доля правильных ответов) – так как потенциальному музыкальному сервису важно, чтобы как можно больше композиций определялись с правильным жанром, а любое отклонение недопустимо.

Точность модели на тренировочной выборке составила 95%, на тестовой – 60%.

Ниже представлены результаты обучения нейронной сети (рис. 13 − 16).

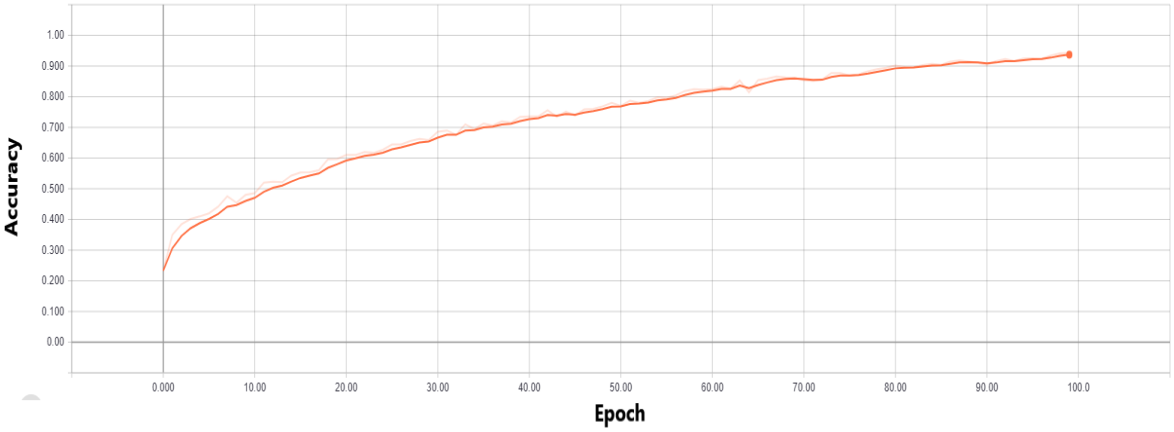


Рисунок 13. График зависимости точности нейронной сети от эпохи обучения на тренировочной выборке

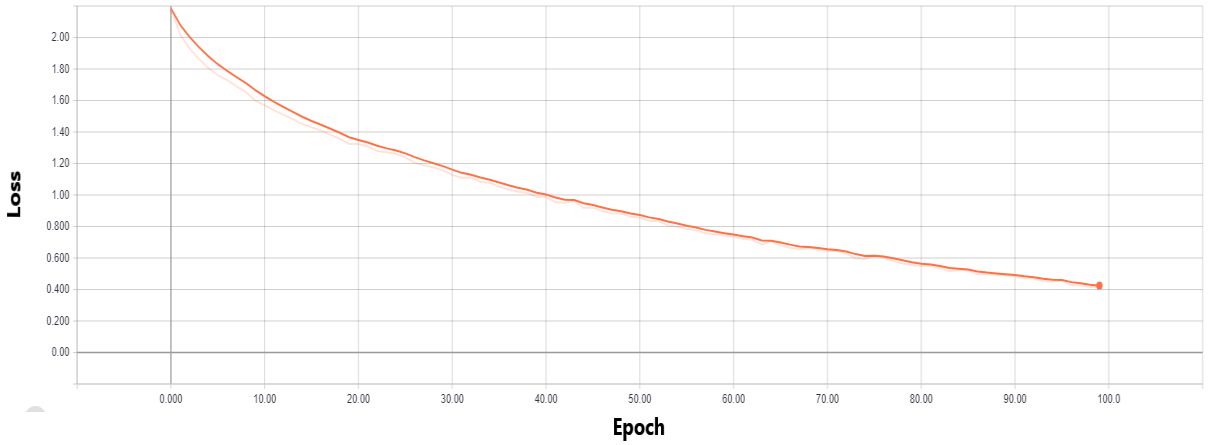


Рисунок 14. График зависимости значения функции потери от эпохи обучения на тренировочной выборке

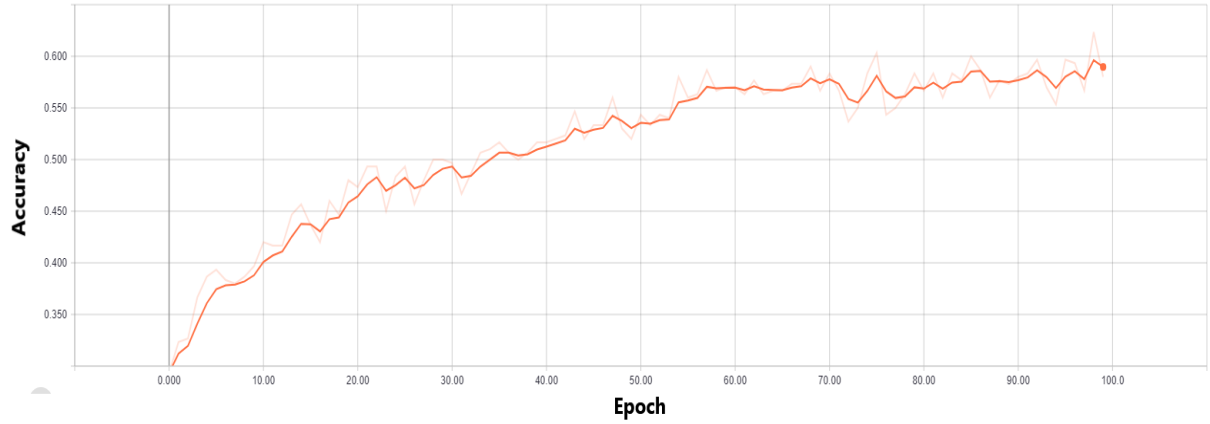


Рисунок 15. График зависимости точности нейронной сети от эпохи обучения на тестовой выборке

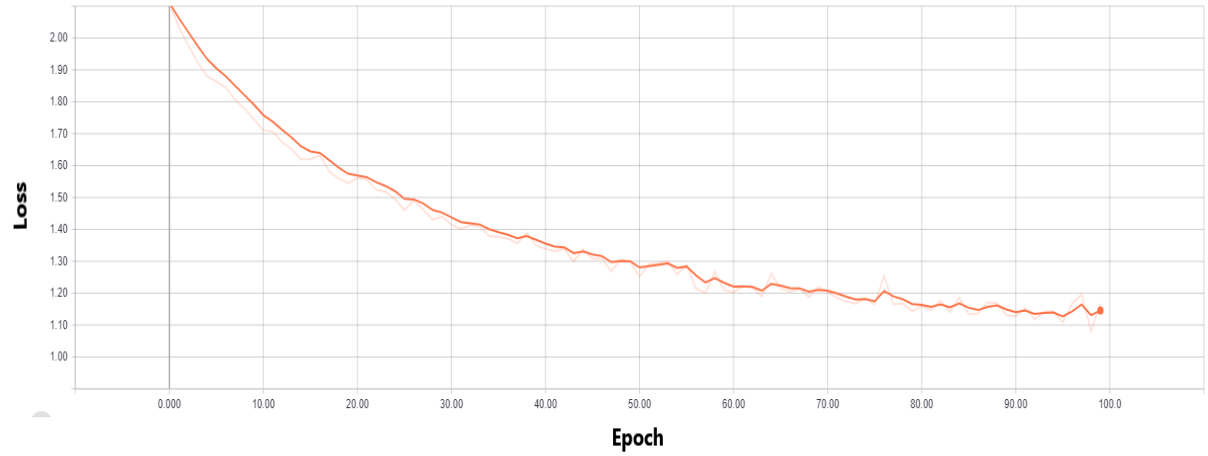


Рисунок 16. График зависимости значения функции потери от эпохи обучения на тестовой выборке

Результатом работы программы, распознающей жанр композиции с помощью нейронной сети, является следующая диаграмма:

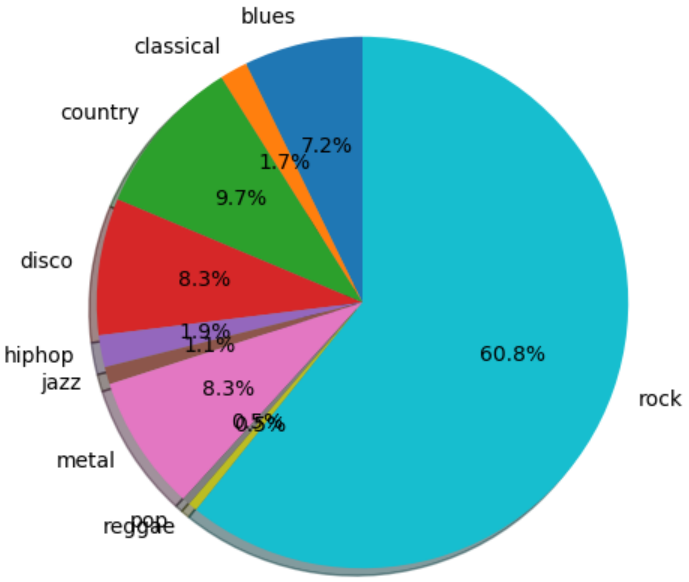


Рисунок 17. Диаграмма распределения предположения о жанре музыки программой для тестового примера

Таким образом, разработана модель нейронной сети для распознавания музыкального жанра, сформированы данные, с помощью которых данная сеть обучена, а также создан программный продукт на языке Python с использованием библиотеки Tensoflow, способный обработать аудиозапись с помощью заранее обученной нейронной сети и визуализировать предположение о жанре.

Литература

[1]. И. Заенцев. Нейронные сети: основные модели – Воронеж, 1999. – 74 с.

[2]. Долгая краткосрочная память [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/Долгая_краткосрочная_памятью - Дата доступа: 25.02.2018

[3]. LSTM – сети долгой краткосрочной памяти [Электронный ресурс]. – Режим доступа: https://habr.com/ru/company/wunderfund/blog/331310/ - Дата доступа: 16.01.2019

[4]. Рекуррентная нейронная сеть. [Электронный ресурс] – Режим доступа: http://mechanoid.kiev.ua/neural-net-lstm.html. – Дата доступа: 14.03.2018.

[5]. Что такое свёрточная нейронная сеть [Электронный ресурс]. – Режим доступа: https://habr.com/ru/post/309508/ - Дата доступа: 16.01.2019