

# Проект по МИИАД

## Классификация музыкальных произведений по жанрам

Дживеликян Е.А.  
Латышев А.К.  
Сизов В.С.

Национальный исследовательский университет  
"Московский физико-технический институт"

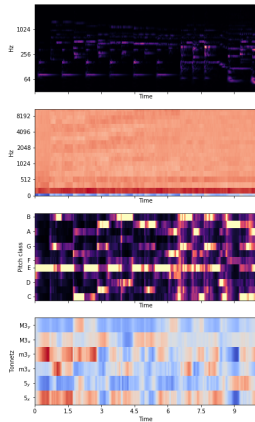
4 ноября 2020 г.

- 8000 треков по 30 секунд каждый, в формате .mp3
- 8 жанров, 1000 треков для кадного жанра

International  
Rock Folk  
Electronic  
Instrumental  
Experimental  
Pop Hip-Hop

# Инструменты

Библиотека инструментов  
для обработки звука



# Признаки

В данной работе были использованы признаки:

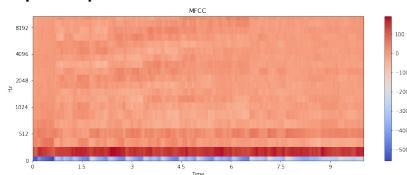
- MFCC(Мел-частотные кепстральные коэффициенты)
- Tonnetz
- Средний темп произведения
- Мощность гармонической и перкуSSIONной компоненты

Спектр спектра, но по мел-шкале.

Мел-шкала



Пример MFC



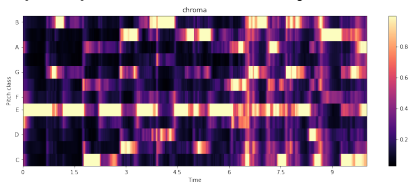
В датасете посчитаны 20 коэффициентов по бинам, на которые разбита песня.

И для каждой последовательности коэффициента рассчитаны статистики: mean, standard deviation, skew, kurtosis, median, minimum and maximum

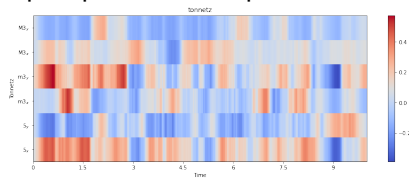
# Tonnetz

Данный признак позволяет оценить наличие гармонии в сигнале, выделить характерные интервалы путём преобразования пространства классов высоты звука.

## Пространство высот звука



## Пространство интервалов



В данной работе используются различные статистики, вычисленные для этого признака по всем фреймам трека.

## Темпоральный спектр произведения



# ГП разделение



ideal harmonic signal



ideal percussive signal



violin



castanets

Вычислены мощности гармонической и перкуSSIONной составляющих треков.

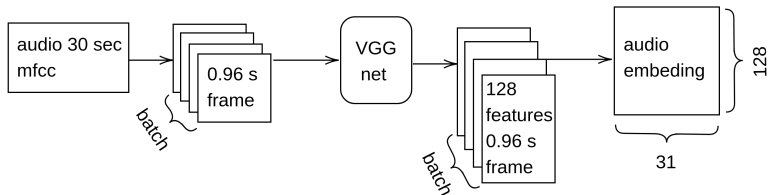


# Результаты. Часть 1

Модель	F1	Параметры	Время обучения	ЭВМ
SVC	59.92	kernel='rbf' C=3	20.2 секунды	Intel(R) Xeon(R) CPU @ 2.30GHz Google Colaboratory
Random Forest Classifier	56.23	n_estimators=500 class_weight='balanced'	28 секунд	Intel Core i9 2400 GHz
Gradient Boosting Classifier	57.13	learning_rate=0.05 max_depth=5 n_estimators=200 subsample=0.5	3 минуты 32 секунды	AMD Razen 5 3500U 2100 MHz
Logistic Regression	53.22	solver='liblinear' class_weight='balanced' multi_class='ovr'	46 секунд	Intel Core i9 2400 GHz

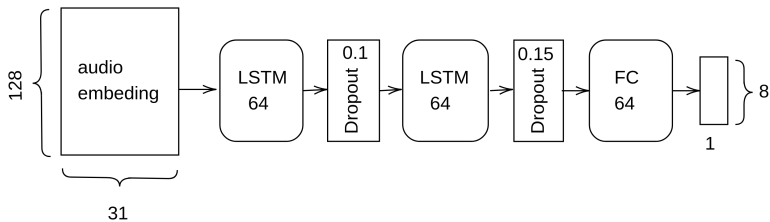
# VGG эмбеддинги

Для выделения признаков высокого уровня использовалась предобученная на Audioset VGG net.



VGG обучалась определять множество разных меток на 0.960 секундных отрывках на датасете Audioset, полученном из роликов youtube.

# LSTM



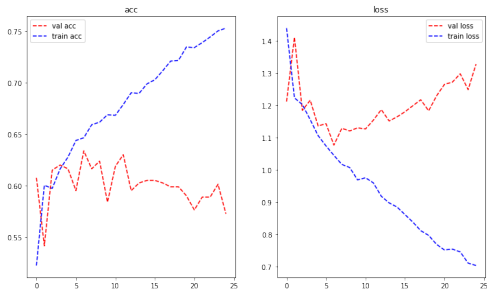
# Пространство поиска

Сэмплировано случайным образом 200 конфигураций с помощью Ray Tune

размер скрытого слоя	от $2^3$ до $2^9$ с шагом степени 1
число слоёв	{1, 2, 3, 4, 5}
скорость обучения	$(10^{-4}; 10^{-1})$
размер батча	{16, 32, 64, 128, 256}
дропаут между LSTM	$(0; 25 * 10^{-2})$
дропаут на выходе	$(0; 25 * 10^{-2})$

Использовался ранний останов по validation accuracy и по алгоритму ASHA.

## Обучение модели с лучшими параметрами



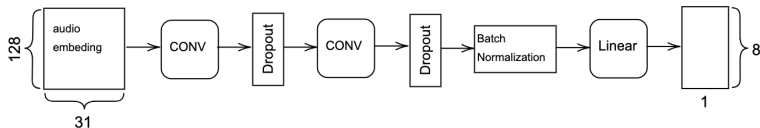
размер скрытого слоя	64
число слоёв	2
скорость обучения	0.006
размер батча	64
дропаут между LSTM	0.1
дропаут на выходе	0.15

Оптимизатор: Adam

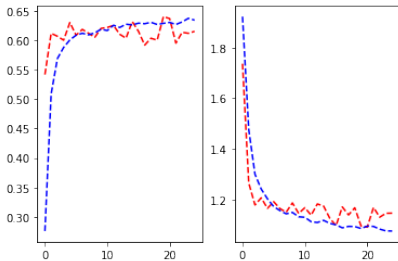
Функция потерь: Cross Entropy на softmax

# FCNN

## Результаты. Часть 2. CNN



Построенная архитектура состоит из двух сверточных слоёв и одного линейного преобразования



Слева—accuracy; справа—loss  
(красный—валидация, синий—трейн)

Количество батчей: 25

Результат на трейне:

Loss: 1.2092

Accuracy: 59.3333

Результат на тесте:

loss: 1.299

accuracy: 0.55

# Основные слайды

① Датасет и инструменты

② Признаки

③ Результаты. Часть 1

④ Результаты. Часть 2

RNN

FCNN

CNN