

# Проект по МИИАД

## Классификация музыкальных произведений по жанрам

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

Национальный исследовательский университет  
“Московский физико-технический институт”

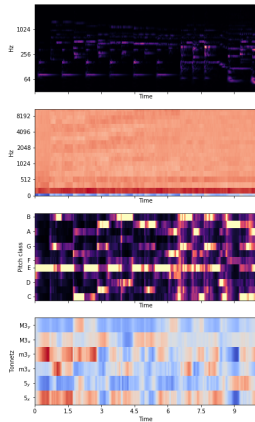
4 ноября 2020 г.

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

International  
Rock Folk  
Electronic  
Instrumental  
Experimental  
Pop Hip-Hop

# Инструменты

Библиотека инструментов  
для обработки звука



# Признаки

В данной работе были использованы признаки:

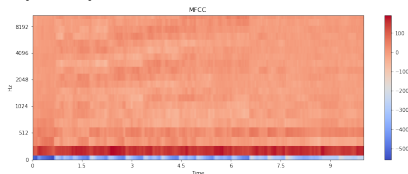
- MFCC(Мел-частотные кепстральные коэффициенты)
- Tonnetz
- Средний темп произведения
- Мощность гармонической и перкуSSIONной компоненты

Спектр спектра, но по мел-шкале.

Мел-шкала



Пример MFC



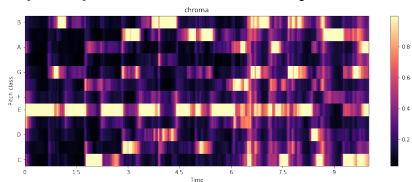
В датасете посчитаны 20 коэффициентов по бинам, на которые разбита песня.

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

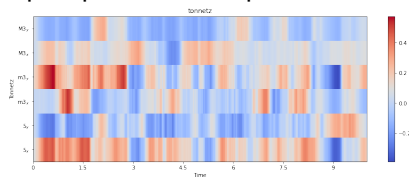
# Tonnetz

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 (те же, что и для MFCC), вычисленные для этого признака по всем фреймам трека.

## Темпоральный спектр произведения



# Гармоника и перкуссия



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 Razer 5 3500U 2100 MHz
Logistic Regression	53.59	C=0.01 solver='lbfgs' multi_class='multinomial'	516 миллисекунд	AMD Razer 5 3500U 2100 MHz
CatBoost	59.34	iterations=800 depth=6 bagging_temperature=0.05 l2_leaf_reg=0	3 минуты 28 секунд	AMD Razer 5 3500U 2100 MHz

# Вклад участников

Дживеликян Е.А.

Разбор признаков Tonnetz. Настройка и работа с Gradient Boosting Classifier.

Латышев А.К.

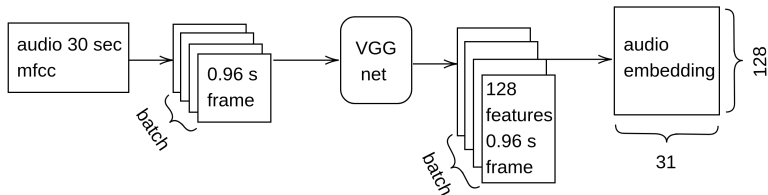
Разбор признаков MFCC и вычисление гармонической и перкуссионной компонент. Настройка и работа с SVC, Logistic Regression, CatBoost

Сизов В.С.

Разбор признаков Temp. Настройка и работа с Настройка и работа с Random Forest Classifier.

# 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.

# LogReg

В качестве baseline использовалась логистическая регрессия, на вход которой подавались эмбединги.

В результате поиска параметра  $C$  в диапазоне от  $10^{-5}$  до 1. Была найдена лучшая модель: `solver='newton-cg'`,  $C=0.001$ .

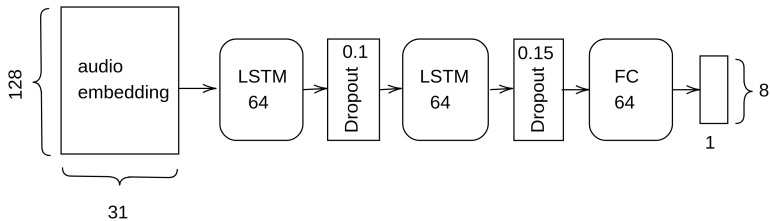
Accuracy: 53.12

F1 = 52.63

Time: 2 минуты 15 секунд

AMD Razen 5 3500U 2100 MHz

# LSTM



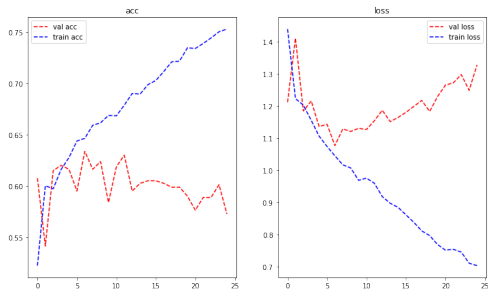
# Пространство поиска

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

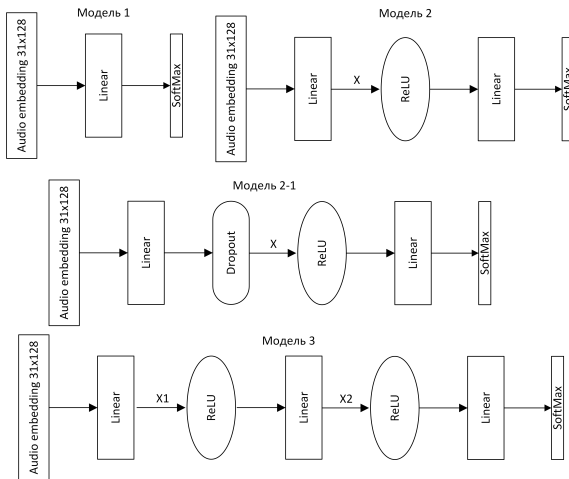
Функция потерь: Cross Entropy на softmax

**Результаты на тесте:**

Accuracy: 0.55

F1(macro): 0.54

# Fully Connected NN





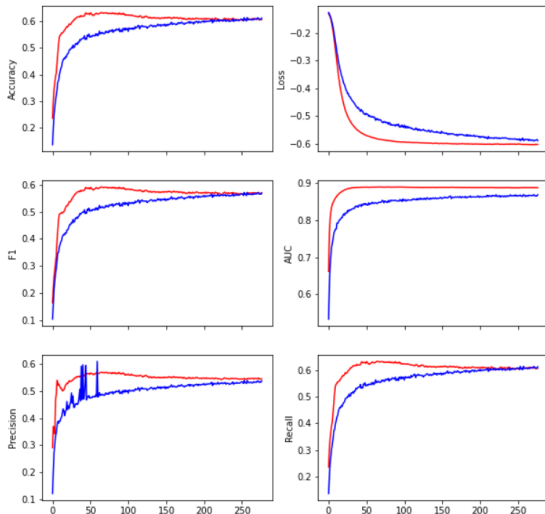
# Результаты FCNN

Для трех предложенных моделей использовался оптимизатор SGD ( $\text{lr}=0.005$ ). И лосс функции NLLLoss и CrossEntropyLoss (значимой разницы они не показали).

В моделях 2, 2-1 и 3 подбирались размеры скрытых слоев в диапазоне от 3000 до 50.

Значимых различий для этих моделей и модели 1 не наблюдалось, но в среднем модель 2-1 показала лучший результат (вероятность dropout 0.5).

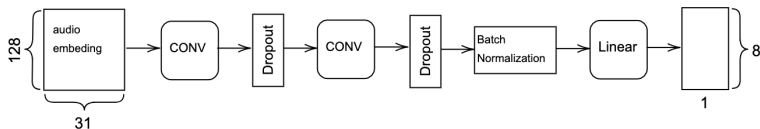
# Лучшая FCNN



красный—валлидация, синий—трейн

Для модели 2-1 был проведен более подробный анализ размера внутреннего слоя. Значения были в диапазоне от 100 до 20 с шагом 10. Было прогнано 3 модели для каждого параметра. В результате лучшим оказалось 30 скрытых нейронов. Количество батчей: 100  
 $F1 : 0.4832 \pm 0.0024$   
 $AUC : 0.8388 \pm 0.0059$   
 $Epoch : 360 \pm 17$

# Convolutional NN



Построенная лучшая архитектура состоит из двух сверточных слоёв и одного линейного преобразования

Количество эпох: 60

Размер батча: 300

Лосс функция: CrossEntropyLoss

Результат на трейне:

Loss: 1.147

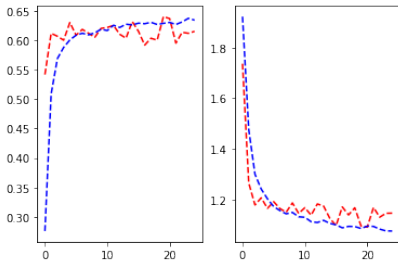
Accuracy: 0.62

Результат на тесте:

Loss: 1.289

Accuracy: 0.57

F1(macro): 0.5328



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

## Результаты. Часть 2

Модель	F1	Время обучения	ЭВМ
LofReg (BaseLine)	52.63	2min 15s	AMD Razen 5 3500U 2100 MHz
LSTM	54.01		NVIDIA Tesla T4 Google Colaboratory
Fully Connected NN	48.32		NVIDIA Tesla T4 Google Colaboratory
Convolutional NN	49.01	1min 55s	NVIDIA Tesla T4 Google Colaboratory

# Вклад участников

Дживеликян Е.А.

Подготовка эмбедингов. Обучение архитектур с рекуррентными слоями.

Латышев А.К.

Обучение полносвязных глубоких сетей и логистической регрессии.

Сизов В.С.

Обучение свёрточных архитектур.

*Все участвовали в оформлении репозитория и презентации.*

# Основные слайды

① Датасет и инструменты

② Признаки

③ Результаты. Часть 1

④ Результаты. Часть 2

RNN

FCNN

CNN