Machine Intelligence Lab Hackathon: Learning to Recognize Musical Genre

Matthias Büchi

January 12, 2018

Abstract

In the course of a final project in the *ZHAW Machine Intelligence Lab* the challenge *WWW 2018 Challenge: Learning to Recognize Musical Genre* was tackled. With the increasing popularity of music streaming services and large music databases, an automatic system for managing the data is essential. The challenge exactly targets this topic, specifically classifying musical audio into genres (e.g. rock, pop, etc.). During the 3 days of work in this challenge the main focus was on implementing an approach using convolutional neural networks on raw audio signal.

1 Introduction

The task of the challenge was to recognize the genre from a piece of music. Given a piece of audio, with a length of 30 seconds, one of 16 genres should be predicted. For this purpose a dataset of musical audio was provided in form of the FMA Dataset [1]. But only the *medium* subset must be used for training, consisting of 25000 tracks. Furthermore a test set with 35000 tracks without labels was given, which have to be predicted.

The performance of the submitted result was evaluated using **Mean Log Loss** and as a second metric the **Mean F1 Score**. The baseline system in form of a SVM, available in the start-kit of the challenge ([2]), was first reproduced and achieved a Mean Log Loss of 0.985 and a Mean F1 Score of 0.6922.

In a next step the data was explored and prepared for the training of a neural network. From overlapping windows of the raw audio signal a **Convolutional Neural Network** with subsequent fully-connected layers was trained to predict the genre of the given window.

2 Findings

2.1 Data preparation

The training data consisted of 25000 music tracks, but unbalanced with respect to the genres. While the biggest genre had 7097 samples, the smallest one had only 21 samples. For training the data was further split into a training and a validation set, where the training part consists of 80% of tracks for every genre.

2.2 Recognition System

As input to the recognition system, 1.0 second windows shifted by a 0.25 seconds from the raw audio signal were used. Samples smaller than a second were padded with zeroes.

To extract features from the signal three convolutional layers with average pooling were used. The first two layers small filter sizes were assigned and a bigger filter size to the third, intended to model long temporal properties. Subsequent fully-connected layers with softmax as final activation were used to predict the probabilities for the 16 musical genres. Except for the last layer batch normalization and ReLU activation were used.

Layer	Size	Stride	Activation
conv-32	5	1	ReLU
avgpool	4	4	-
conv-64	5	1	ReLU
avgpool	4	4	-
conv-128	100	20	ReLU
avgpool	40	30	-
fc	70	-	ReLU
fc	30	-	ReLU
fc	16	-	Softmax

Table 1: Layers and their properties used in the recognition system.

2.2.1 Training

The system was trained using Adam with a learning rate of 0.001 for two iterations over the training data. It was trained to optimize the binary cross entropy loss.

2.2.2 Prediction

For prediction non-overlapping 1 second windows were used. The output of all windows of a single track was averaged to represent the final prediction. The system achieved a Mean Log Loss of **1.098** and a Mean F1 Score of **0.6672**.

3 Outlook

References

[1] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis. In *18th International Society for Music Information Retrieval Conference*, 2017.

[2] S.P. Mohanty and Michaël Defferrard. crowdai-musical-genre-recognition-starter-kit. `https://github.com/crowdAI/crowdai-musical-genre-recognition-starter-kit`, 2017. [Online; accessed 10-January-2018].