Semi-Supervised Music Tagging Transformer

Zineb Lahrichi

December 15, 2021

1 Introduction

Groover has enhanced the way musicians are promoted, creating a new ecosystem where they can broadcast their music to professionals and receive valuable feedback. With a total of 60 000 artists and 1200 curators, relevant recommendations need to be made for artists, to make the most of their experience on the platform.

However, relying exclusively on user explicit meta-data does not allow to build a robust recommender system. **Thus, there is need to develop intelligent Music tagging models in order to classify each song into relevant categories, based on its acoustic properties.** In this report, we will briefly review the paper *Semi-Supervised Music Tagging Transformer*, understanding how we can integrate their approach into Groover's recommender system pipeline.

***keywords**: Transformers, Semi-supervised learning, Multiple instance Learning, Self-Attention, Data Augmentation, Teacher-Student Learning*

2 Proposed approach

Network Architecture The proposed model takes mel-spectrograms of each labeled song as inputs. They are divided in 3.69-second short instances called *chunks*. **The model architecture combines a CNN that captures local features and a transformer that adds global context to the learned information**. Single-instance labels are then aggregated to make a global prediction.

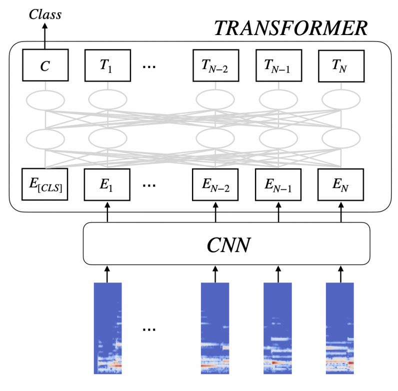

Figure 1: Music Tagging Transformer Architecture

Semi-Supervised Learning Given the scale of music libraries, a large amount of unlabeled audio tracks is often encountered. As a solution, they propose to leverage unlabeled data with a noisy teacher-student learning technique. This method is combined with data augmentation, to avoid overfitting.

Data augmentation can be easily implemented and improve the performance considerably. It consists in creating new labeled samples, by adding perturbations to available tracks, conserving their inherent global characteristics. Common perturbation methods are : *random noise, random gain, audio filters...*

3 Tackled Challenges

1. The proposed approach models each track representation as a sequence, instead of learning a simple bag-of-features representation like many methods found in the literature. This helps to learn global contextual information in addition to local features, and to make better predictions.

2. A Deep Learning model extracts patterns on its own, allowing to use unstructured data to learn feature representations. However, depending on the task, feature interpretability might be required.

3. Data augmentation is very useful to tackle overfitting problems. It allows to be robust to different mixing filters and recording settings.

The issues tackled in this paper are relatable to many other supervised learning task using audio inputs. As a result, interesting ideas can be incorporated in Groover's Music tagging pipeline. For example, the noisy-student learning technique can handle imprecise or missing labels, which might occur on the platform.