Groover technical test: Article analysis

Julien Guinot
jul.guinot@gmail.com

1. Introduction

This document summarizes the research and proceedings displayed by *Won et Al.* [1] in the paper **Semi-supervised music tagging transformer** in the Context of the Groover technical test. A novel combination of Convolutional Neural Networks, transformers and semi-supervised learning is leveraged to attain SOTA ROC-AUC measurements.

2. Task overview

Music tagging is a task in which, provided with a raw audio sample, a set of musical tags such as *rock, sad, electric guitar, etc...* are generated. It is a Multi-label classification learning task which is tackled here by exploiting the Million-song dataset, a well-known dataset in the domain of music-information retrieval comprised of a audio clips taken from a million songs.

2.1. Past work

Music tagging is generally adressed in one of two ways [1]: by extracting features at a global level, or at instance-level, then classifying the instance through features extracted for each instance and using pooling or voting strategies to determine the global class.

Past works Leverage 1D [2] or 2D CNNs [3] to classify audio extracts , and have been shown to work better on frame-level clips (a few seconds of audio), and to approach the problem using the previous pooling/voting approaches [3]. This is coherent from a human standpoint, as a listener does not need a whole song to notice it is happy or sad.

CNN Models were improved further by treating frame-level extracted features as a sequence of features and feeding them to a recurrent neural network, which shows great performance with sequential data. The advent of SOTA transformer models [4],[5] further improved this model by leveraging self-attention mechanisms, which is the model [1] focuses on.

2.2. Final model

The model *Won et. Al* use to conduct their experiments is the previously-described CNN-Transformer architecture, which is summarized in Fig. 1 [1]:

The CNN front-end is chosen to be a shallow ResNet architecture with granular *3x3* filters as per previous work with dimension reduction applied before processing through the transformer backend.

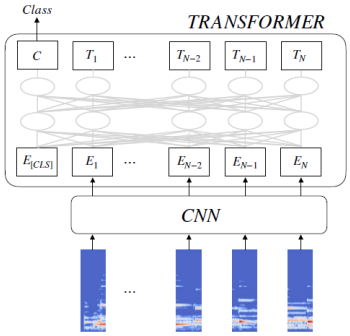

Figure 1: Proposed Music Tagging Transformer.

Figure 1: [1] final model

3. Training

[1] leverages a semi-self-supervised learning scheme to work with massive unlabeled data amounts in the million songs dataset : **Noisy student learning** [6], which combines:

- **Student learning**, in which a parent model is trained on labeled data and generates pseudo-labels for unlabeled data to train a student model on. (semi-supervised)
- **Noise-invariant training**, in which data augmentation is applied to the labeled data to generate more labeled data to train the model on (self-supervised)

3.1. Results

Pre-processing for this task consists of stratifying the MSD into new artist-level stratas when compared with previous splits, generating Mel-spectrograms for training data, and applying audio data augmentation (noise, filters , reverb, etc...).

Classification metrics used to evaluate the model are receiving operator characteristic area under curve and Precision-recall area under curve, as standard with music tagging models. *Won et. Al.* show best results are obtained with data augmentation and noisy student training implemented.

The best model is obtained with a smaller student model than the parent model (Knowledge distillation), and reaches **89.72%** ROC-AUC on the benchmark split, beating out previous SOTA by **2%**. It further achieved **92.17%** on the custom split. It it also shown that the model exhibits resistance to higher-length audio input, compared to pure-CNN approaches.

References

[1] M. Won, K. Choi, and X. Serra, "Semi-supervised music tagging transformer," *arXiv preprint arXiv:2111.13457*, 2021.

[2] J. Lee, J. Park, K. L. Kim, and J. Nam, "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms," *arXiv preprint arXiv:1703.01789*, 2017.

[3] M. Won, A. Ferraro, D. Bogdanov, and X. Serra, "Evaluation of cnn-based automatic music tagging models," *arXiv preprint arXiv:2006.00751*, 2020.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[6] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10687–10698, 2020.