Title: Timbre Classification of Musical Instruments Using A Conformer-Based Architecture

Members:

- Andrew Gao: agao31
- Sarah Bao: scbao
- Luke Zhao: lzhao40
- Ziqi Fang: zfang31

Implementation URL(s):

- https://github.com/carlosholivan/timbre-classification-multiheadattention
- https://github.com/sooftware/conformer

Introduction:

We are reimplementing the paper "Timbre Classification of Musical Instruments WIth a Deep Learning Multi-Head Attention-Based Model". The purpose of this paper is to create a deep learning model using multi-headed attention that can classify/differentiate various instrument timbres using as few parameters as possible. The paper described various methods of timbre classification, ranging from non-deep-learning approaches to CNN's to the paper's multi-head attention model. We hope to draw inspiration from this paper by implementing similar preprocessing steps and having the same overall objective, but instead using a model that combines CNNs and Transformers in conformer blocks rather than just applying multi-headed attention. We will evaluate the results of our model using metrics like Precision (P), Recall (R), and F1 score, and compare them to the two models in the original paper with respect to these metrics. We will also analyze the confusion matrix to find interesting and meaningful relationships between timbres of instruments of different families. As musicians ourselves, we chose this paper because we are very interested in music-related deep learning projects.

Related Work:

- ChordFormer: A Conformer-Based Architecture for Large-Vocabulary Audio Chord Recognition: https://arxiv.org/pdf/2502.11840
- Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music: https://ieeexplore.ieee.org/document/7755799
- MedleyDB: A MULTITRACK DATASET FOR ANNOTATION-INTENSIVE
- MIR RESEARCH: https://archives.ismir.net/ismir2014/paper/000322.pdf
- Conformer: Convolution-augmented Transformer for Speech Recognition: https://arxiv.org/abs/2005.08100, https://github.com/sooftware/conformer

Chordformer seeks to apply a Convolutional Neural Network with multi-headed attention for large-vocabulary chord recognition. This conformer architecture captures both local audio features like formants and harmonics as well as longer-term musical context, like chord progressions. For our implementation, we seek a similar approach of using a conformer architecture for timbre classification.

Data:
The London Philharmonic Orchestra Dataset consists of ~14000 monophonic sound samples from 58 instruments, grouped into 4 classes: woodwind (8), percussion (39), brass (4), and strings (7). Each sample contains an annotation of pitch, instrument type, dynamic, and playing technique. We will take a similar approach to preprocessing as the paper, extracting a clip of the sound sample beginning at the point where the energy level exceeds a specific threshold and generating normalized spectrograms using the first 500 ms of instruments playing.

Methodology:
The architecture has two main components: conformer blocks and a decoding model. After preprocessing as described in the data section, we plan to input the spectrograms into a series of conformer blocks, which blend together an overall Transformer architecture that implements multi-head self-attention to capture global interactions with CNNs that can better capture local features. Then, the decoding model will apply a linear unit that projects the hidden state into an output with size equal to the instrument classes in order to get the scores for each instrument class, and therefore the final prediction.

Our overarching motivation for following this architecture is that both CNNs and attention-based models seem to perform well at recognizing instruments, so we thought that combining them into a conformer block would be effective. As for training the model, we plan to train on 70% of the data, validate on 10%, and test on 20% – all the data is already labelled, so we would be implementing a supervised deep learning model.

Metrics:
In the study by Hernandez-Olivan and Beltran, the results were evaluated with Precision (P), Recall (R), and F-score metrics. The Freq. FC model was able to achieve an overall Precision of 0.45, Recall of 0.41, and F1 score of 0.38. On the other hand, the Freq. Attention model was able to achieve an overall Precision of 0.66, Recall of 0.62, and F1 score of 0.62. In our implementation of the conformer architecture, we will be calculating the Precision, Recall, and F1 scores of individual instruments, as well as the overall Precision, Recall and F1 score.
The original study also analyzed the confusion matrices for the test set with both models and gathered valuable information about the models. For instance, one essential result was that seemingly different instruments can have similar timbres when played with different techniques and in different registers (for example, the timbre of bowed string instruments when playing in high registers with soft dynamics have very similar timbres as woodwind instruments). In our implementation of the conformer architecture, we want to perform a detailed analysis of the confusion matrices with respect to all the instruments, and aim to discover interesting and meaningful relationships between instruments of different families.

Another important experiment that the original study performed was analyzing the weights and the activation maps for different inputs. The authors concluded that the model is able to learn timbre and that it is able to distinguish between different instruments of the same family. For our implementation of the conformer architecture, the architecture will include more features from both convolutional neural networks as well as transformers, so we want to visualize weights and activation maps as well and compare those to the ones presented in the paper. This can help us discover why the performance of the different models differ.

For our implementation of the conformer architecture, the base goal is to achieve the performance of the Freq. FC model (overall Precision of 0.45, Recall of 0.41, and F1 score of 0.38). The target goal is to achieve the performance of the Freq. Attention model (overall Precision of 0.66, Recall of 0.62, and F1 score of 0.62). The stretch goal is to surpass the performances of both of the models implemented in the original study (overall Precision of 0.8, Recall of 0.8, and F1 score of 0.8).

Ethics:

1. Why is Deep Learning a good approach to this problem?

While there have been non deep learning approaches to this problem, these methods require a lot of computational power due to the high number of audio features to distinguish between various timbres. A deep learning model learns these features automatically.

2. What is your dataset? Are there any concerns about how it was collected, or labeled? Is it representative? What kind of underlying historical or societal biases might it contain?

We will be using the London Philharmonic Dataset. The data was collected and specially recorded by members of the Philharmonia, so it should be legitimate. One concern of the dataset is that it may not be very representative of all the different instruments in an orchestra. For example, there are many samples for string and wind instruments, but very few for percussion instruments. Even within the string section, the dataset is dominated by traditional string instruments (violin, viola, cello, bass), while instruments like the guitar and mandolin have very few samples. It also does not contain cultural instruments or instruments not usually featured in Western orchestras. While this bias is not a big problem for what we are trying to achieve, it may be a concern when trying to classify a broader scope of instruments.

Division of Labor:

Luke – Preprocessing
Ziqi – Model Implementation & Tuning
Andrew – Model Implementation & Tuning
Sarah – Interpretability & Visualizations