Classification of Instrument Timbres in Classical Music Sound Samples Using a Conformer Model Architecture

by Sarah Bao (scbao), Ziqi Fang (zfang31), Andrew Gao (agao31), Luke Zhao (lzhao40)

Introduction

We reimplemented the paper "Timbre Classification of Musical Instruments WIth a Deep Learning Multi-Head Attention-Based Model" by Carlos Hernandez-Olivan and Jose R. Beltran (2021). 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 hoped 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 evaluated the results of our model on the London Philharmonic Orchestra dataset which contains simple, monophonic data, and compared our precision, recall, and f1 metrics to the two models in the original paper. We also ended up applying our model to a separate, polyphonic IRMAS dataset, which proved to be much more difficult to classify. At the end, we analyzed the confusion matrices to find interesting and meaningful relationships between timbres of instruments of different families. As musicians ourselves, we chose this paper because we were very interested in music-related deep learning projects.

Methodology

Preprocessing & Training

The two datasets that we trained our model on are the London Philharmonic Orchestra dataset and the IRMAS dataset. The LPO dataset contains professionally recorded audio samples curated from the archives of the LPO with 13,681 mp3 files labeled across 20 instruments grouped into 4 categories: woodwind, percussion, brass, and strings. The IRMAS dataset consists of 9,579 wav files sampled at 44.1 kHz labeled with the predominant instrument (11 in total). All audio files were downsampled, converted to mono, and segmented into 3-second clips. Log-mel spectrograms with 80 frequency bins and 300 time frames were computed for each clip, providing a compact but expressive time-frequency representation.

As for training the model, we used the Adam optimizer with an initial learning rate of 1e-3 and categorical cross-entropy loss with label smoothing of 0.1 to improve generalization. The training was conducted over 20 epochs with a batch size of 32, using 80-dimensional log-mel spectrograms extracted from audio recordings (LPO and IRMAS datasets). The model was validated on a 10% subset of the data and ultimately tested on our held-oud test set containing 20% of the data.

Model Architecture

Our model is based on the Conformer, a hybrid design that combines convolutional layers with self-attention to capture both local and global dependencies in the audio signal (we used and applied the code in this repo). Each input spectrogram is first linearly projected to a

256-dimensional feature space, then passed through a stack of four Conformer blocks, each comprising feedforward modules, multi-head attention with relative positional encoding, and convolutional sublayers with gated linear units. The output is aggregated over the time dimension and passed through a final dense layer followed by a log-softmax activation, producing class probabilities across all instrument categories. 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.

Results

With the Conformer Model training and testing on the LPO dataset, we observed excellent performance across all metrics in classifying timbres across 16 instrument classes. We achieved an overall Precision (P) of 0.960, Recall (R) of 0.917, F-1 Score of 0.926, and Accuracy of 0.967. For the IRMAS dataset, we observed significantly worse performance, likely due to data-related challenges. Because the IRMAS dataset is based on real musical excerpts, there are many variables that confuse standard classifiers, such as multiple instruments, background noise, reverb, etc. However, we achieved overall Precision (P) of 0.555, Recall (R) of 0.529, F-1 Score of 0.511, and Accuracy of 0.516, which met our base goal. The use of conformer modules that combined convolution and attention addressed the problems brought up in the original paper that only implemented MLPs and FC models with Self-Attention components.

Analysis of the Confusion Matrix allowed us to confirm a phenomenon that previous authors have discovered - instruments are more likely to be misclassified as other instruments in the same family/class. For instance, the violin/cello and clarinet/saxophone pairings are the most obvious from the confusion matrices.

Challenges

The IRMAS dataset consisted of sound files with multiple instruments but labeled with a single instrument. The task of predicting a predominant instrument in this situation is a task that is even challenging to a classically trained musician, which explains why we could not achieve our target and stretch goals after days of extensive tuning.

Reflection

How do you feel your project ultimately turned out? How did you do relative to your base/target/stretch goals?

We are overall satisfied with the result of our project. We managed to reach all of our original goals, although this was mainly because the original paper we looked at had very low accuracy scores. This made it fairly easy to reach our target, as we even exceeded our stretch goals by over 0.1 in precision, recall, and F1. Thus, we set a new challenge for ourselves, which was to classify the more difficult IRMAS dataset. We managed to achieve an accuracy of about 0.5, which fell short of the paper's accuracy of about 0.66. While we tried to fine tune our conformer model, if we had more time to work on this project, we would have experimented even more with tweaking the conformer blocks to try to achieve higher accuracy.

Did your model work out the way you expected it to?
Yes, generally our model worked the way we expected. It was complex enough to perform very well on the simple, monophonic data, but not fine tuned enough to perform as well on the polyphonic, IRMAS data.

How did your approach change over time? What kind of pivots did you make, if any? What would you have done differently if you could do your project over again?
At the start, we had only focused on achieving high accuracy on the LPO dataset, but once we found that we were smoothly able to achieve this with only 5 epochs of training, we wanted to try something more difficult. That's when we found the IRMAS dataset. Initially, we had hoped that our conformer model was just "OP" and could classify instruments very well, but we soon found out that wasn't true. The IRMAS dataset was incredibly hard to classify – our very first run on the dataset achieved less than 30% accuracy, and as we were listening to various clips in the dataset, we found that we also struggled to identify the predominant instrument in multiple clips.

We then all tried to experiment in different ways to see if we could increase the accuracy. Andrew tried to tune hyperparameters, playing with the learning rate, hidden size dimension, and dropout. Luke tried to edit the architecture itself, adding more conformer blocks and convolution layers. Ziqi experimented with feature engineering, trying to see if adding spectral metrics such as centroid, roll-off, and flux or augmenting the data would help. Sarah played around with a transfer model where she tried to train for a few epochs on the LPO dataset, freeze certain layers of the model, and then continue training on the IRMAS dataset. Our hypothesis here was that the model could first identify general trends present in the monophonic data and calibrate against it before having to work with the more complicated IRMAS data. This ended up not really working, likely because the LPO and IRMAS datasets were just too different. But then we noticed that after training on the IRMAS dataset, there were cases when the model would primarily predict one class for all the examples even though the distribution of instruments in the dataset was relatively uniform. To prevent this, we kept track of the training accuracy of each class and weighted the classes that were more poorly predicted on higher. This ended up being effective, and our best model was able to achieve an accuracy of 51.6%, a significant improvement from our first run.

What do you think you can further improve on if you had more time?
The LPO dataset is also labeled with dynamics (on a spectrum from fortissimo to pianissimo) and bowing techniques (normal, arco, sul tasto). We could expand our model to classify the data with these labels. We also could've spent more time working on model interpretability. In implementing the Multi-Headed Attention Layers, we could extract the attention weights, and overlay them with the Mel-Spectrograms in order to visualize parts of the input data that the attention layers focused on. We could also visualize the weights in the convolution layers, all in hopes of developing a better understanding of why our model performs the way that it does.

What are your biggest takeaways from this project/what did you learn?

Our biggest takeaway was that when working with datasets like IRMAS and being faced with more difficult classification problems, models that utilize self-attention components achieve far more success than models that do not. This is because the attention mechanism allows the model to extract the most pertinent features in the timbral space. The model was able to use the various spectral, temporal and spectrotemporal properties of sound events that most effectively distinguished one instrument timbre from another.

We also realized that domain knowledge is very helpful when trying to handle difficult classification problems – if we had a better idea of how dominant instruments can be identified, we could've done a better job adding features to our model.

Another important takeaway is that when working in a team, effective and clear communication is a crucial factor for successful execution of the project. Whenever a member ran into difficulty in an intermediate step, we all came together to brainstorm solutions and offer alternative strategies. In this process we built mutual trust and crafted a project that all of us are proud of.

References

Github Repos

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

Academic Papers

Akram, Muhammad Waseem, et al. "ChordFormer: A Conformer-Based Architecture for Large-Vocabulary Audio Chord Recognition." arXiv preprint arXiv:2502.11840 (2025).

Bosch, Juan J., Jordi Janer, Emilia Gómez, and Perfecto Herrera. "A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals." Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR), 2012.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. arXiv. https://arxiv.org/abs/2005.08100

Hernandez-Olivan, Carlos, and Jose R. Beltran. "Timbre classification of musical instruments with a deep learning multi-head attention-based model." arXiv preprint arXiv:2107.06231 (2021).

Y. Han, J. Kim and K. Lee, "Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 1, pp. 208-221, Jan. 2017