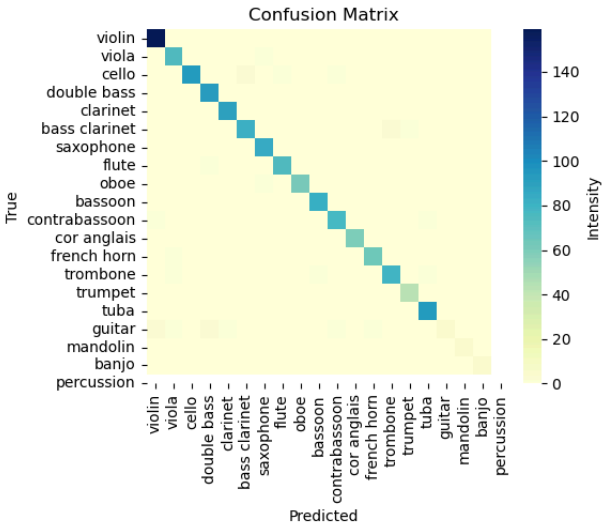
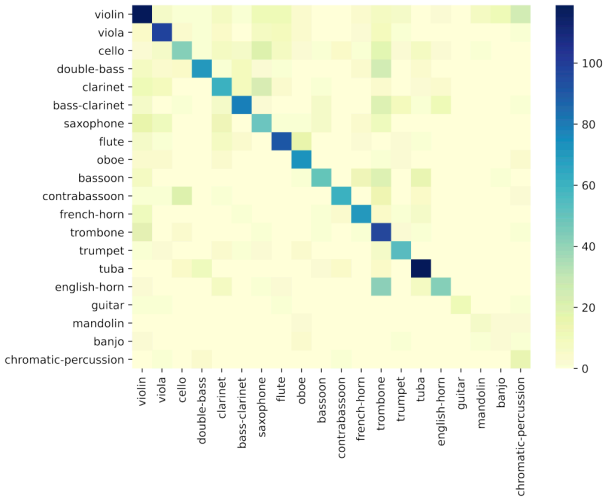
Title: Timbre Classification of Musical Instruments Using A Conformer-Based Architecture

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.

Challenges:
Since we're working with the dataset and preprocessing steps from one paper but trying to implement the conformer-based model architecture of a different paper, it's been difficult to make the code compatible with each other. Also, a lot of the code that we pulled initially wouldn't run (e.g. we were having a lot of trouble with the import statements), so we spent a bulk of our time debugging. After we were able to get the code running though, the conformer model worked shockingly well on the dataset (more details below). Since we're a little surprised at how well the model works, we're shifting our focus to do more interpretability analyses, which also prove to be quite challenging. Since the input data is 2 dimensional, which is more complicated than just one dimensional sequences, we're trying to brainstorm ways to effectively visualize the attention and convolution layers. We've been looking into visualization tools such as BertViz for transformers and other tools, but we suspect that applying these visualization techniques is going to be quite difficult.

Insights:
Our model outperforms the best model in the original paper by over 40% (the paper's best model has a precision of 0.66 and the conformer model we implemented has a precision of 0.985). Below is a side by side of the confusion matrix produced from our model (left) and the matrix produced by the paper's model (right):





Plan:

We think that we're on track to finish by DL Day. The only next steps that we have are interpretability analysis (visualizing + conducting an ablation study) and maybe fine-tuning our model a bit more. We currently only use one conformer block in our model, but we can experiment with a ConformerEncoder architecture that stacks multiple conformer blocks; we're also thinking about fine-tuning different parameters such as number of attention heads, size of kernel, pooling metrics, etc.