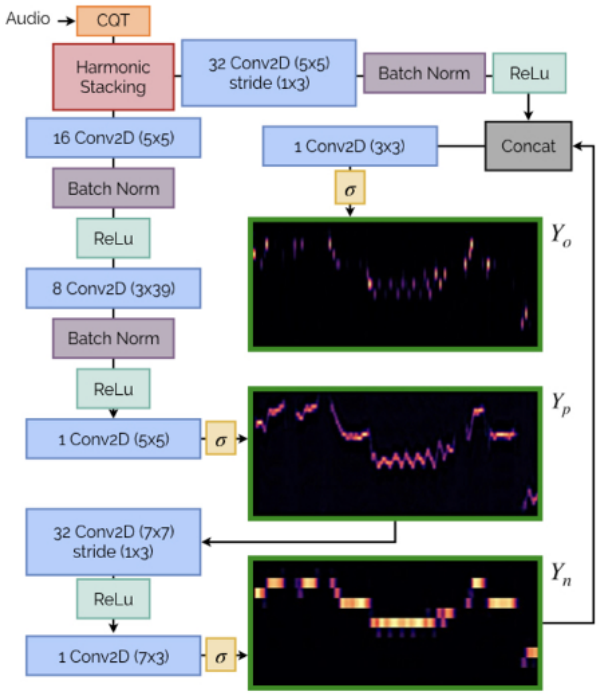
Noah Adhikari, Mrunali Manjrekar, Eran Kohen Behar, Shalin Brahmbhatt

CS182 Neural Networks / Deep Learning

7 December 2022

Final Commentary

The application of technology to music transcription has long been a curiosity in the field of computer science. Generally, note estimation has been proposed via a variety of formats including median filtering and neural networks. Our assignment seeks to delve deeper into one model in particular; Spotify's instrument agnostic music transcription model known as "Notes and Multipitch" (NMP), or "Basic Pitch". This model seeks to translate raw audio files into computerized sheet music (MIDI). The paper can be found here and the associated GitHub repository here. Our repository is here.



Harmonic Stacking, and multiple convolutional and normalization layers, all of which are explored in the assignment.

The model is designed to be extremely lightweight, only consisting of roughly 17,000 parameters. It uses a signal transform followed by a series of batch normalization and convolutional layers to output three separate annotation matrices: pitches (frequencies), onsets (when notes begin), and contours (inflections and nuances within notes). In the paper, researchers indicate a convolutional layer of size 5x5, a batch norm, and a ReLU are used after harmonic stacking; however, they appear to have made a mistake in the source code - some layers are skipped, so we have students implement the missing layers themselves. Additionally, to ensure no barriers to entry, students are given descriptions and code for constant Q-transforms and harmonic stacking (signal processing steps), as well as reading and writing audio files; all of which are pertinent to the model yet not important for students to understand at a deeper level.

For students to properly grapple with the intricacies of the model, we have implemented visualization and audiation methods to allow the comparison of predictions and expected outputs via sensory feedback. The paper augments the input audio via reverb, white noise, and equalization in its training phase. As such, our assignment encourages students to create a more robust model and explores other options beyond the paper, such as transposition of key and adding irrelevant waveforms. Students are able to hear the unmodified audio files in comparison to the augmented ones. This part of the homework connects with denoising autoencoders, where noise is added to try and restore the original input, and “mixup” augmentation, where two images are combined together. It also helps students think about data augmentation in general, where we must consider invariances that we want our augmentations to capture. As a (non)example, we

highlight the issue with masking in this context because a note with the center of it masked, or muted, will come across as two notes, changing the expected output significantly.

The original paper used several datasets to train the model, but we have selected only one - the MAESTRO piano dataset, found here, due to both availability constraints and pedagogical instruction. Though NMP is instrument-agnostic, the visualization and audiation methods used in our notebook are most easily absorbed through the piano samples provided in MAESTRO.

To experiment with the original model, students engage in ablation studies and gauge performance differences. The two main ablation studies conducted in the paper are the removal of harmonic stacking as an input representation and the removal of Y_p as a supervised bottleneck layer. If we had more time, we would have liked to bring light to the importance of iterating through model ideas and understanding the importance of each layer with similar ablation studies. Another idea on our wish list is that of 2-second sliding samples across the input files as a means to train across many more short samples, which the paper implements - we would have liked to compare training performance across this technique and simply taking the inputs without any sort of duration preprocessing.

Students are also able to strengthen their mechanical foundations through analytical problems. One of our problems explores median filtering with visual yet mathematical reinforcement through pictures and matrices. It illustrates some conceptual ideas and mathematical reasoning behind using median filtering for denoising purposes in comparison to box-blur filters and connects the technique to CNNs.

Another topic relevant to the paper and the course is weighted cross-entropy, which makes traditional cross-entropy loss more robust to unbalanced distributions. We have another

analytical problem exploring this and why it can perform much better than traditional cross-entropy, as well as what hyperparameter setting performs best for a given unbalanced distribution when wanting to give negative and positive samples an equal vote in the loss.

As students complete training the model, we ask them to consider the original paper's result and how the paper went about training the model. There are three methods by which to benchmark the model; accuracy, F-measure (F), and note-level F-measure-no-offset (Fno). For the purposes of the project, we decided to benchmark using the F value. Not only was this value easier to calculate with built-in libraries, the fact that offsets are less objective made for a more concrete benchmark. Future work may include the application of a transformer architecture to audio transcription as well.