System Description for the Audio Chord Estimation Task at MIREX 2025

Yiwei Ding and Christof Weiß

Center for Artificial Intelligence and Data Science, University of Würzburg, Germany

Overview

Our system is an ensemble of a transformer-based architecture and a CRNN-based architecture. The training target includes is a combination of chord change, structured chord representation and chord symbols. The model is trained in two-steps, first step on a big classical music dataset, and then fine-tuned on a smaller set of pop music. Our results on a held-out dataset show that the transformer-based architecture is more effective than the CRNN-based architecture, but the ensemble still gives a small boost.

We go over these setups in the following sections.

Model Architecture

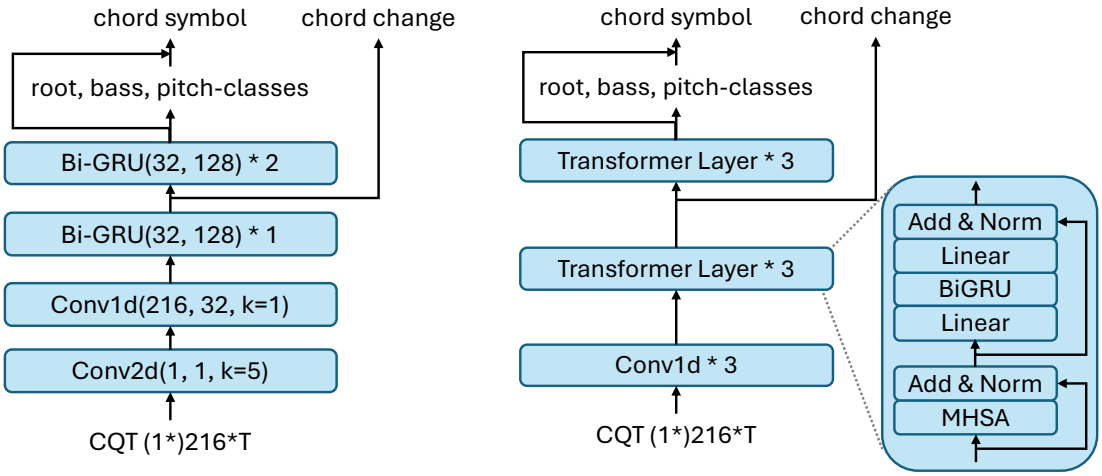

Figure 1: Illustration of the CRNN-based architecture (left) and the transformer-based architecture (right).

The architectures of the models we use are illustrated in Figure 1.

The CRNN-based architecture is based on [1], except that we add one more Bi-GRU layer for higher capacity. The output of the first Bi-GRU layer is branched out to predict chord changes (specified in the next section).

The transformer-based architecture is inspired by, but very different from [2]. We first apply three 1-D convolutions to the input CQT. Then we use six transformer layers. In the transformer layers, we use post-norm and relative positional encoding, as in [2]. We add a Bi-GRU layer in the feedforward block in normal transformers, replacing the convolutional feedforward in [2]. After three layers we branch out the output to predict chord changes.

During inference, we use only the chord symbol prediction. To ensemble two models we simply average the output probabilities.

Training target (loss function)

The output of the CRNN (or transformer) firstly predict structured chord representations, which are the root note, the bass note, and the activated pitch-classes in the chord, with linear layers, as in [1]. After that, these predictions, as well as the original output feature, are concatenated together and predict the chord symbol with a linear layer. The root note and the bass note are trained with cross-entropy losses, the pitch-class prediction is trained with a binary cross-entropy loss, and the chord symbol prediction is also trained with a cross-entropy loss.

We use the chord vocabulary of sevenths with inversions during training. We map all the other chords to the vocabulary and map those untransferable chords as “no chord”. We also ignore the “no chord” label, i.e., we do not calculate losses on these frames.

Predicting chord changes is found to be helpful in [3]. We let the model predict chord change at each frame and use binary cross-entropy as the loss.

These losses are simply added together during training.

Datasets and two-step training

The datasets we use are summarized in Table 1.

Training Stage	Dataset	Duration (hh:mm)
Pre-training	BPSD [4]	41:07
	BSQD [t.b.p]	62:12
Finetuning	Beatles [5]	8:09
	Isophonics (excl. Beatles) [5]	7:53
Test	RWCPop [6]	6:47

Table 1: Datasets used in our experiments

The models are first (pre-)trained on a combination of two classical music datasets: Beethoven Piano Sonata Dataset (BPSD) [4], and Beethoven String Quartet Dataset (BSQD)[1] . We pre-train on classical music because these datasets are bigger in size (~100h). We pre-train the model for 50 epochs and save the model with highest CSR on the validation set.

After pre-training, we fine-tune the model on pop music. We use the Isophonics dataset [5], including 18 songs by Zweieck, 20 songs by Queen, 7 songs by CaroleKing, and the Beatles dataset including 180 songs by the Beatles. Both datasets are split into training, validation and test set. We fine-tune the model for 25 epochs with a smaller learning rate and save both the model with highest validation CSR and the last model.

We use RWC Pop dataset [6] as a held-out test set.

Experiments and results

We evaluate our models on the RWC Pop dataset. We ignore the "no chord" labels in the annotations. The results are shown in Table 2.

Model	Root	Majmin	+Inv	7ths	+Inv	UnderS	OverS
1-CRNN-best	82.33	81.90	77.86	59.81	56.55	87.30	85.41
2-CRNN-last	82.61	82.07	78.22	60.59	57.50	87.35	85.30
Ensemble 1+2	82.54	82.05	78.10	60.18	57.00	87.33	85.39
3-Transformer-best	84.09	83.45	80.33	60.94	58.38	89.28	**86.78**
4-Transformer-last	84.30	83.58	80.49	65.13	62.65	**89.86**	85.92
Ensemble 3+4	84.60	84.09	81.08	63.73	61.28	89.75	86.65
Ensemble 2+4	**84.94**	**84.70**	**81.46**	**65.26**	**62.64**	89.17	86.70

Table 2: Results on RWC Pop dataset. The numbers are averaged across songs (instead of across frames. The second to sixth columns are the chord symbol recall (CSR) on different vocabularies, and the last two columns are the segmentation scores.

For this submission, we submit the "Ensemble 2+4" system because it shows consistently higher CSR than other systems. We submit the "4-Transformer-last" system as the best single-model system in a separate submission.

[1] This dataset will be published soon and comprises multiple versions of all movements from all Beethoven's string quartets. Chord annotations are derived from the symbolic ABC dataset [7].

References

[1] B. McFee and J. P. Bello, "Structured Training for Large-Vocabulary Chord Recognition," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, 2017.

[2] T.-P. Chen and L. Su, "Attend to Chords: Improving Harmonic Analysis of Symbolic Music Using Transformer-Based Models," *Transactions of the International Society for Music Information Retrieval (TISMIR),* 2021.

[3] T.-P. Chen and L. Su, "Harmony Transformer: Incorporating Chord Segmentation into Harmony Recognition," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019.

[4] J. Zeitler, C. Weiß, V. Arifi-Müller and M. Müller, "BPSD: A Coherent Multi-Version Dataset for Analyzing the First Movements of Beethoven's Piano Sonatas," *Transactions of the International Society for Music Information Retrieval (TISMIR),* 2024.

[5] M. Mauch, C. Cannam, M. Davies, S. Dixon, C. Harte, S. Kolozali, D. Tidhar and M. Sandler, in *Late-breaking Demo at International Conference on Music Information Retrieval (ISMIR LBD)*, Kobe, Japan, 2009.

[6] M. Goto, H. Haishiguichi, T. Nishimura and R. Oka, "RWC Music Database: Popular, Classical and Jazz Music Databases," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2002.

[7] M. Neuwirth, D. a. M. F. C. Harasim and M. Rohrmeier, "The Annotated Beethoven Corpus (ABC): A dataset of harmonic analyses of all Beethoven string quartets," *Frontiers in Digital Humanities,* 2018.