In music, harmonic analysis is the process of finding the underlying relationship among the notes and joining them together. The process is two-fold: recognizing harmony labels and finding their time boundaries. Most previous works only focused on the first component, which may lead to segmentation errors.

In this paper, we introduce a novel approach to jointly detect the labels and time boundaries with neural semi-CRF (Conditional Random Field).

**Dataset:** Four classical music datasets of different styles are included in the experiments. In total, there are 321 pieces and 44K harmony labels

**Input Representations:** MIDI-like symbolic music are sliced into fixed-length frames (eighth note). The pitch information in each frame is processed and transformed to a 24-dimensional vector that encodes the pitch-class activations as well as the bass note.

1. **Frame-level estimation:** CRNN-based encoder to compute the probability distribution of harmony types at the frame-level.
2. **Attention-based score function:** For each candidate harmony region, attention is used to aggregate multiple frame-level estimations to a single distribution of harmony types.
3. **Absence Score:** An additional score component is used to explicitly penalize incomplete chord profiles and extra chordal notes in the estimations.

**Neural Score Function**

\[ S(Y, X) = \sum_{i=1}^{m} S_i(Y_i, X_i) \]

**Score Function:** \[ S(X, Y) = \sum_{i=1}^{m} S_i(X_i) \]

**Training Objective:** Given \( X \) and \( Y \), find the model parameters \( \theta \) that minimizes \( L(\theta) = -\log(P(\theta | Y | X)) \)

**Inference:** Given \( X \), find \( Y \) that maximizes \( P(Y | X) \)

Support Audio Input
Design more efficient training strategies to alleviate the time complexity of semi-CRF
Enable real-time processing