Input: Single-source as well as mixes

↓ (Mel-Spectr. or „Harmonic Tensor")

Model (pre-trained) + fine-tuned

- Harmonic CNN
- Short-Chunk CNN
- PANNs

separate parallel layers (later on attention might be attention layers)

↓ ↓ ↓ ↓ ↓ ↓ ↓

| Git | Vox | Strings | | | Drums Perc |

Instrument-Family space

Activity/ Loudness

256

Activity/Loudness/ Event Detection

Event-Det.: Norm

Norm

Normalize

Violin   Viola   32

Hi-Hat

Snare

String Orchestra

Drums

Cello   CB

Kick

Marimba

MFCCs (single-frame?) Pooling

Zero Crossing Rate

Spectral-Centroid

Deep Clustering or Neighborhood Component Analysis:

$$\mathcal{L} = \frac{1}{|B|} \sum_{i \in B} \log \frac{\sum_{\substack{j \neq i \\ y_j = y_i}} \exp(-\|z_i - z_j\|^2)}{\sum_{k \neq i} \exp(-\|z_i - z_k\|^2)}$$

(how to sample Batches?)