Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching

Colin Raffel
May 16, 2016
The Goal

- artist: 'Tori Amos'
- release: 'LIVE AT MONTREUX'
- title: 'Smells Like Teen Spirit'
- id: 'TRKUYPW128F92E1FC0'
- duration: 216.4502
- sample_rate: 22050
- audio_md5: '8'
- 7digitalid: 5764727
- year: 1992
Sequence Matching
Dynamic Time Warping
Comparing MIDIs with DTW
Downsampled Hash Sequences
Similarity-Preserving Hashing
Network Structure
Loss Function

\[ \mathcal{L} = \frac{1}{|\mathcal{P}|} \sum_{(x,y) \in \mathcal{P}} \| f(x) - g(y) \|_2^2 + \frac{\alpha}{|\mathcal{N}|} \sum_{(x,y) \in \mathcal{N}} \max(0, m - \| f(x) - g(y) \|_2)^2 \]
Raw Distance Distributions
Output Distance Distributions
Hash Distance Distributions
Match Ranks

![Graph showing the percentage below different ranks. The x-axis represents the rank, ranging from $10^0$ to $10^6$, and the y-axis represents the percentage below, ranging from 0 to 100. The graph shows a curve that levels off as the rank increases.]
Pairwise Sequence Embedding
Pairwise Sequence Embedding
Feed-Forward Attention

Raffel & Ellis, “Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems”, ICLR 2016
Embedding Network

Convolution and Pooling

Feedforward Attention

Dense Layers
Example Embeddings
Match Ranks

- Y-axis: Percentage below
- X-axis: Rank

Graph shows the cumulative percentage of ranks below a certain level.
Combined Match Ranks

Combined Match Ranks

- Combined
- DHS
- PSE

Rank

Percentage below

10^0 10^1 10^2 10^3 10^4 10^5 10^6
References


[2] Raffel & Ellis, “Large-Scale Content-Based Matching of MIDI and Audio Files”, ISMIR 2015