Accelerating Large-Scale Sequence Retrieval with Convolutional Networks

Colin Raffel
IIT Bombay
December 23rd, 2015
Sequence Retrieval
Dynamic Time Warping
Goal

artist: 'Tori Amos'
release: 'LIVE AT MONTREUX'
title: 'Smells Like Teen Spirit'
id: 'TRKUYPW128F92E1FCO'
duration: 216.4502
sample_rate: 22050
audio_md5: '8'
7digitalid: 5764727
year: 1992
Matching and Aligning

Beatles/hello, goodbye.mid  ⇔  The Beatles/Hello.mp3
Matching by Metadata Won’t Work

J/Jerseygi.mid
V/VARIA180.MID
Carpenters/We'veOnly.mid
2009 MIDI/handy_man1-D105.mid
G/Garotos Modernos - Bailanta De Fronteira.mid
Various Artists/REWINDNAS.MID
GoldenEarring/Twilight_Zone.mid
Sure.Polyphone.Midi/Poly 2268.mid
d/danza3.mid
100%sure.polyphone.midi/Fresh.mid
rogers_kenny/medley.mid
2009 MIDI/looking_out_my_backdoor3-Bb192.mid
DTW: Natural, and Too Slow
Hash Sequences

\[ \text{distance}[m, n] = \text{bits.set}[x[m] \oplus y[n]] \]
Similarity-Preserving Hashing
Similarity-Preserving Hashing
Collecting Data

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIDI</td>
<td>140,910</td>
</tr>
<tr>
<td>Hand</td>
<td>24,850</td>
</tr>
<tr>
<td>Test</td>
<td>17,243</td>
</tr>
<tr>
<td>Test</td>
<td>26,311</td>
</tr>
<tr>
<td>Train</td>
<td>25%</td>
</tr>
<tr>
<td>Development</td>
<td>25%</td>
</tr>
<tr>
<td>Validate</td>
<td>10%</td>
</tr>
</tbody>
</table>
Network Structure

5x12 convolution, 2x2 max-pool

3x3 convolution, 2x2 max-pool

Dense ReLU

Dense ReLU

Dense tanh
Loss Function

\[ \mathcal{L} = \frac{1}{|\mathcal{P}|} \sum_{(x,y) \in \mathcal{P}} \| f(x) - g(y) \|_2^2 - \frac{\alpha}{|\mathcal{N}|} \sum_{(x,y) \in \mathcal{N}} \max(0, m - \| f(x) - g(y) \|_2)^2 \]
Validation Distance Distribution

![Graph showing the distribution of distances between similar and dissimilar items, with bars representing proportions at different distances.](image-url)
Example Sequence

7 digital audio CQT

Synthesized MIDI CQT

Audio hash sequence

MIDI hash sequence

CQT distance matrix

Hash sequence Hamming distance matrix
Side Note: Cross-Modal
First Layer Filters
Test Set Matching Results
Sequence Embedding
Sequence Embedding
Sequence Embedding
Attention
Feed-Forward Attention

$\alpha(h_t)$

$\alpha_1, \alpha_2, \alpha_3, \alpha_T$

$h_1, h_2, h_3, \ldots, h_T$

$C$
Side Note: Toy Problems

from Ilya Sutskever et al., “On the Importance of Initialization and Momentum in Deep Learning”

<table>
<thead>
<tr>
<th>Sequence length ($T_0$)</th>
<th>Addition</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
<td>100</td>
<td>500</td>
<td>1000</td>
<td>5000</td>
<td>10000</td>
</tr>
<tr>
<td>Attention</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Unweighted</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>8</td>
<td>17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence length ($T_0$)</th>
<th>Multiplication</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
<td>100</td>
<td>500</td>
<td>1000</td>
<td>5000</td>
<td>10000</td>
</tr>
<tr>
<td>Attention</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>Unweighted</td>
<td>2</td>
<td>2</td>
<td>8</td>
<td>33</td>
<td>89.8%</td>
<td>80.8%</td>
</tr>
</tbody>
</table>
Validation Distance Distribution

The graph shows the distribution of distances for similar and dissimilar cases. The y-axis represents the proportion, while the x-axis represents distance. The pink bars indicate similar cases, and the green bars indicate dissimilar cases.
Example Embeddings
Embedding Distance Matrix
MIDI-to-MSD Matching

![Graph showing MIDI-to-MSD matching with two lines representing Embedding and Mean and Standard Deviation.]
References

1. “Large-Scale Content-Based Matching of MIDI and Audio Files”, 16th International Society for Music Information Retrieval Conference, 2015

2. “Pruning Subsequence Search with Attention-Based Embedding”, 2016 IEEE International Conference on Acoustics, Speech and Signal Processing


4. “Feedforward Networks with Attention Can Solve Some Long-Term Memory Problems”, in preparation

http://github.com/craffel/midi-dataset
http://github.com/craffel/sequence-embedding

craffel@gmail.com