Learning Tags that Vary Within a Song

Michael Mandel, Douglas Eck, and Yoshua Bengio

LISA Lab, DIRO, Université de Montréal
mandelm@iro.umontreal.ca

International Society of Music Information Retrieval Conference
August 11th, 2010
Automatically describe music from its sound (autotag)
  ▶ in terms of genre, instrument, vocals, production, rhythm, ...

In order to
  ▶ find music by description
  ▶ browse music with similar descriptions
  ▶ summarize many songs simultaneously
News

- Amazon’s Mechanical Turk can collect useful training data
- People agree more when describing “closer” musical excerpts
- Correlations over time can be used to “smooth” MTurk tags
- Smoothing improves overall classification accuracy from 57% to 61%
Overview

Trex

MTurk

T1: chorus male vocals rock driving drumset happy
T2: chorus male vocals rock drum female vocals
T3: club dance electric male rap

Feat Ext

Aggregate

SVMs

Smooth

Covariance
Envelope Cepstrum

Male vocals
Male chorus
drumset
electric
dance

Happy

Rock

Driving

Male vocals
Female vocals

Rap

Modern

Fun

Country

Excited

Indie

Alternative

Guitar

Violin

Party

Upbeat

Club

Dance
Outline

1. Collecting tags
2. Tag language model
3. Autotagging experiment
4. Summary
Outline

1. Collecting tags
2. Tag language model
3. Autotagging experiment
4. Summary
Mechanical Turk

Instructions (click to expand/collapse)

1. Listen to the clip

2. Describe its unique qualities with 5-10 tags total (hover over label for definition)
   Styles/Genres:
   Vocals/Instruments:
   Overall sound/feel:
   Moods/Emotions:
   Other:

3. (Optional) How can we improve this HIT? What would make it faster or easier?

Submit
MTurk data collected

- 5 clips from each of 185 random blog tracks = 925 clips
  - each seen by 3 turkers and described with 18 tags on average
- Total of 2,500 (user,clip) pairs, 15,500 (user,clip,tag) triples
- Paid $0.03–$0.05 per clip, total of about $100
- Rejected 11% of responses as spammy or incomplete
MTurk data analysis: Tag co-occurrence

- How similar are the tag vectors from two different (user, clip) pairs?
- How does that vary for pairs with different relationships?
- For all pairs of observations (or a random subset)
  - count number of shared tags
  - add this count to a bucket based on the observations’ relationship
Tag co-occurrence vs “separation” (MajorMiner)
Tag co-occurrence vs “separation” in one track

Mandel, Eck, Bengio (LISA Lab)
Outline

1. Collecting tags
2. Tag language model
3. Autotagging experiment
4. Summary
Tag language model

- Predict joint distribution of tags for each clip
  - using knowledge of its proximity to other clips
- Up-weight false negative tags, down-weight false positives
- Conditional restricted Boltzmann machine
  - unsupervised model of $p(\text{tags} \mid \text{user}, \text{track}, \text{clip})$
  - trained on all tags of all clips
  - smooth by computing $E\{\text{tags} \mid \text{user}, \text{track}, \text{clip}\}$ for each clip
Restricted Boltzmann machine
Conditional restricted Boltzmann machine

Hidden

Visible (tags)

Aux (user, track, clip)
Conditional restricted Boltzmann machine
Conditional restricted Boltzmann machine
Conditional restricted Boltzmann machine
Conditional restricted Boltzmann machine
Conditional restricted Boltzmann machine
Outline

1. Collecting tags
2. Tag language model
3. Autotagging experiment
4. Summary
Experiment

T1: chorus male vocals rock driving drumset happy
T2: chorus male vocals rock drum female vocals
T3: club dance electric male rap

MTurk

TReX

Feat ext

SVMs

Aggregate

Smooth

Covariance

Envelope Cepstrum

Male vocals

Male chorus

Drum set

Electric

Club

Dance

Happy

Rock

Driving

Female vocals

Rap

Fun

Country

Harmony

Ambient

Excited

Indie

Modern

Party

Upbeat

Violin

Guitar

Rock

Dance

Pop

Alternative
Experiment

- Two datasets
  - MTurk: 925 clips, 100 tags
  - MajorMiner: 2500 clips, 100 tags
- Each tag is treated as a separate classification problem
  - Train using smoothed or aggregated data
  - Test using smoothed or aggregated data
- Evaluate using classification accuracy on a balanced test sets
### Accuracy on balanced test set, averaged over tags

<table>
<thead>
<tr>
<th></th>
<th>Mechanical Turk</th>
<th></th>
<th>MajorMiner</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Trained</td>
<td>Tested</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw</td>
<td>56.87 ± 0.52</td>
<td>56.56 ± 0.36</td>
<td>Raw</td>
<td>65.97 ± 0.49</td>
</tr>
<tr>
<td>Smoothed</td>
<td>61.43 ± 0.51</td>
<td>63.40 ± 0.35</td>
<td>Smoothed</td>
<td>66.67 ± 0.49</td>
</tr>
</tbody>
</table>
MTurk data, tested on raw tags

Mandel, Eck, Bengio (LISA Lab)

Learning tags

August 11th, 2010 20 / 26
MajorMiner data, tested on raw tags

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure.png}
\caption{Smoothed vs Raw tags comparison for various musical genres and instrument types.}
\end{figure}

---

*Mandel, Eck, Bengio (LISA Lab)*

*August 11th, 2010*
Outline

1. Collecting tags
2. Tag language model
3. Autotagging experiment
4. Summary
Amazon’s Mechanical Turk can collect useful training data
People agree more when describing “closer” musical excerpts
Correlations over time can be used to “smooth” MTurk tags
Smoothing improves overall classification accuracy from 57% to 61%
Future work

- Combine audio with joint tag model
- Exploit unlabeled and weakly labeled data
- Analyze song structure using tags more systematically
Thanks!
Thanks!

Questions?