Advances in Multimodal Datasets
CMU Multimodal SDK

Paul Pu Liang
Advancements in Modeling

- Human multimodal language has seen a surge of interest in fine-grained modeling.

- Wouldn’t account for changes or temporal co-occurrences.

I liked the movie I saw last night

Overall smile

Excited voice
Advancements in Modeling

- New approaches rely on word-level alignments to build a co-occurrence model (Gu et al. ACL 2018, Zadeh et al. AAAI 2018, Chen et al. ICMI 2017)
Advancements in Modeling

- New approaches rely on word-level alignments to build a co-occurrence model (Gu et al. ACL 2018, Zadeh et al. AAAI 2018, Chen et al. ICMI 2017)
- These can account for changes in each modality.

\[ \text{I liked the movie I saw last night} \]

\[ \text{Smile} \quad \text{Neutral} \quad \text{Frown} \]

\[ \text{Excited} \quad \text{Neutral} \quad \text{Excited} \]
Advancements in Modeling

- New approaches rely on word-level alignments to build a co-occurrence model (Gu et al. ACL 2018, Zadeh et al. AAAI 2018, Chen et al. ICMI 2017)
- These can account for changes in each modality.

- These approaches complicate datasets and data processing.
CMU Multimodal SDK

• Publicly available SDK for loading and training multimodal temporal data.
• Call Sequence:

  Multimodal Scientist

  Feature Request

  Computational Sequence

  Validations Parameters

  SDK Server
Computational Sequence

<Word Vectors>

<Video ID #1>

<Video ID #2>

<Video ID #3>

.
### Computational Sequence

#### Multimodal Computational Descriptors in Hierarchical Format

<table>
<thead>
<tr>
<th>Intervals</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>S(0) E(0)</td>
<td>f(0,0) f(0,1) f(0,2) f(0,3) f(0,4)</td>
</tr>
<tr>
<td>S(1) E(1)</td>
<td>f(1,0) f(1,1) f(1,2) f(1,3) f(1,4)</td>
</tr>
<tr>
<td>S(2) E(2)</td>
<td>f(2,0) f(2,1) f(2,2) f(2,3) f(2,4)</td>
</tr>
<tr>
<td>S(T) E(T)</td>
<td>f(T,0) f(T,1) f(T,2) f(T,3) f(T,4)</td>
</tr>
</tbody>
</table>
Computational Sequence

• Suitable for temporal data from multiple modalities.

• Compatible with hdf5 (hierarchical data format) protocol.

• Stored using binary values and validated using trust server using sha256 hash – allowing for feature sharing among community.

• Storage using lossless floating point and string compression.
Computational Sequence

- Community can share their extracted features across network.

Multimodal Scientist

Deploy Request
Set Private Key

Acknowledgment
Using UUID and SHA256

SDK Server
Current Datasets

- CMU-MOSEI – 23453 samples, sentence level sentiment and emotions
- CMU-MOSI – 2199 samples, sentence level sentiment
Current Datasets

- CMU-MOSEI – 23453 samples, sentence level sentiment and emotions
- CMU-MOSI – 2199 samples, sentence level sentiment
- ICT-MMNO – 340 samples, video level sentiment
Current Datasets

- CMU-MOSEI – 23453 samples, sentence level sentiment and emotions
- CMU-MOSI – 2199 samples, sentence level sentiment
- ICT-MMMO – 340 samples, video level sentiment
- MOUD – 386 samples, videos in Spanish, sentence level sentiment
Current Datasets

- CMU-MOSEI – 23453 samples, sentence level sentiment and emotions
- CMU-MOSI – 2199 samples, sentence level sentiment
- ICT-MMMO – 340 samples, video level sentiment
- MOUD – 386 samples, videos in Spanish, sentence level sentiment
- POM – 903 samples, video level personality traits
Current Computational Sequences

- Language
  - Glove word embeddings

- Vision
  - FACET descriptors
  - OpenFace descriptors

- Acoustic
  - Covarep
  - OpenSmile
Future Computational Sequences

- Vision
  - Face Embeddings
  - VGG-Face

- Acoustic
  - Low-level features
  - Phoneme Embeddings
Multisequence Alignment

Expected cross-sequence Alignment

<table>
<thead>
<tr>
<th>Visual</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Visual Alignment" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Audio</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image2.png" alt="Audio Alignment" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language</th>
<th>l</th>
<th>have</th>
<th>been</th>
<th>umm</th>
<th>really</th>
<th>well</th>
<th>since</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image3.png" alt="Language Alignment" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Multisequence Alignment

Expected cross-sequence Alignment

<table>
<thead>
<tr>
<th>Hierarchy</th>
<th>S</th>
<th>E</th>
<th>S</th>
<th>E</th>
<th>S</th>
<th>E</th>
<th>S</th>
<th>E</th>
<th>S</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audio</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language</td>
<td>I</td>
<td>have</td>
<td>been</td>
<td>umm</td>
<td>really</td>
<td>well</td>
<td>since</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Multisequence Alignment

Cross-sequence Alignment

Non-flatten

Hierarchy

Visual

Audio

Language

I have been

Hierachy

Visual

Audio

Language

I have been

Flatten
Multisequence Alignment

Cross-sequence Alignment

<table>
<thead>
<tr>
<th>Hierarchy</th>
<th>Visual</th>
<th>Audio</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-flatten</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flatten</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hierarchy</th>
<th>Visual</th>
<th>Audio</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-flatten</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flatten</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Carnegie Mellon University
Future Work

- Automated evaluation of submitted models
- Public leaderboard on the CMU-MultimodalSDK
The End!

Data: https://github.com/A2Zadeh/CMU-MultimodalSDK
Website: www.cs.cmu.edu/~pliang, www.amir-zadeh.com
Email: pliang@cs.cmu.edu, abagherz@cs.cmu.edu
Twitter: @pliang279