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.
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.
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

<Word Vectors>

<Video ID #1>
<Video ID #2>
<Video ID #3>

\[ (S(0), E(0)) \]
\[ (S(1), E(1)) \]
\[ (S(2), E(2)) \]
\[ \ldots \]

\[ (S(T), E(T)) \]

<Intervals>

<features>

\[ f(0,0) \quad f(0,1) \quad f(0,2) \quad f(0,3) \quad f(0,4) \]
\[ f(1,0) \quad f(1,1) \quad f(1,2) \quad f(1,3) \quad f(1,4) \]
\[ f(2,0) \quad f(2,1) \quad f(2,2) \quad f(2,3) \quad f(2,4) \]

\[ \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \]
\[ \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \]

Multimodal Computational Descriptors in Hierarchical Format
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-MMMO – 340 samples, video level sentiment
Current Datasets

- CMU-MOSEI – 23453 samples, sentence level sentiment and emotions
- CMU-MOSI – 2199 samples, sentence level sentiment
- ICT-MMMMO – 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>
<th>Audio</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

```
I have been really well since
```

Carnegie Mellon University
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

Hierarchy

Visual

Audio

Language

I have been umm really well since
Multisequence Alignment

Cross-sequence Alignment

<table>
<thead>
<tr>
<th>Non-flatten</th>
<th>Flatten</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hierarchy</strong></td>
<td><strong>Hierarchy</strong></td>
</tr>
<tr>
<td><strong>Visual</strong></td>
<td><strong>Visual</strong></td>
</tr>
<tr>
<td><strong>Audio</strong></td>
<td><strong>Audio</strong></td>
</tr>
<tr>
<td><strong>Language</strong></td>
<td><strong>Language</strong></td>
</tr>
<tr>
<td>I</td>
<td>I</td>
</tr>
<tr>
<td>have</td>
<td>have</td>
</tr>
<tr>
<td>been</td>
<td>been</td>
</tr>
</tbody>
</table>
Multisequence Alignment

Cross-sequence Alignment

Non-flatten

<table>
<thead>
<tr>
<th>Hierarchy</th>
<th>Visual</th>
<th>Audio</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Flatten

<table>
<thead>
<tr>
<th>Hierarchy</th>
<th>Visual</th>
<th>Audio</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

I have been really well since I have been umm

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