Multimodal recognition of behavior and affect

USC Viterbi
School of Engineering

Guest lecture for Affective Computing
Multimodal emotion recognition

Instructor: Mohammad Soleymani
Multimodal emotion recognition
Affective states

- Observing external manifestations in short episodes
Emotions as componential constructs

- GSR
- Blood pressure
- Respiration pattern
- EMG Frontalis
Why multimodal?

- What does he feel?

Slide credit: Nicole Nelson
Modality

• “Particular mode in which something exists or is experienced or expressed.”
• “a particular form of sensory perception.” for example, auditory, visual, touch
• Multimodal
• Examples:

• We also include other perceptual channels, for example, language
Why multimodal?

• Complementary information

• Multimodal interaction
  • McGurk effect

• Robustness
  • Missing/noisy channels
Human behavior sensing modalities

- Behavior
  - Audio
  - Visual
  - Physiological response
    - Peripheral
    - Central

- Nonverbal
  - Language
  - Prosody
  - Face
  - Body
  - Gaze

- Verbal
Unimodal, bimodal and trimodal interactions

- **Unimodal**
  - Speaker’s behaviors: "This movie is sick" or "This movie is fair"
  - Sentiment Intensity: +, -
  - Examples:
    - "This movie is sick" with a smile: Ambiguous!
    - "This movie is fair" with a loud voice: Unimodal cues

- **Bimodal**
  - Speaker’s behaviors: "This movie is sick" with smile or frown, loud voice
  - Sentiment Intensity: +, -
  - Examples:
    - "This movie is sick" with a smile and loud voice: Resolves ambiguity (bimodal interaction)
    - "This movie is sick" with a frown and loud voice: Still Ambiguous!

- **Trimodal**
  - Speaker’s behaviors: "This movie is sick" with smile, loud voice, "This movie is fair" with smile, loud voice
  - Sentiment Intensity: +, ++, +++
  - Examples:
    - "This movie is sick" with smile and loud voice: Different trimodal interactions!
    - "This movie is fair" with smile and loud voice: Different trimodal interactions!

Slide credit: LP Morency
Multimodal representation learning for emotion recognition
What are representations and why they matter?
Features are representation
Learning representations – neural networks

- Perceptron
- Multi-layer perceptron

Learns representations
Face encoders ConvNets – holistic methods

Expression of emotion

Levi and Hassner, ICMI 2015
ConvNets – patch based

- Ertugrul et al.,
- Use z-face for 3D registration
- Create overlapping patches
- Pass them through CNN/3DCNN
Convolutional nets—self-supervised

- **Input:** Sequence of frames extracted from a video.

- **Network:** ResNet-18 encoders with shared weights

- **Loss:** Triplet losses between adjacent frames. Summing up as ranking triplet loss.

- **Learning to rank,** Lu et al., BMVC 2020
Language encoders

• Language is sequential
• It has structure (grammar)
• It is also full of ambiguity
• Typical approach is to use a sequential encoder:
  • CNNs
  • Causal CNN
  • RNNs
  • Transformers
How to learn word representations?

• Input and output are one-hot coded

He was walking away because …
He was running away because …

• The n-dim hidden layer learns a compact representation, e.g., 300d

Word2vec [https://code.google.com/archive/p/word2vec/](https://code.google.com/archive/p/word2vec/)
GloVe [https://nlp.stanford.edu/projects/glove/](https://nlp.stanford.edu/projects/glove/)
Word embedding

• Represent words as vectors
• Unsupervised method that learns the neighboring words in text
• Word2vec and GloVe are the popular examples
CNN for text analysis

*Kim, Y. “Convolutional Neural Networks for Sentence Classification”, EMNLP (2014)*
Recurrent neural networks

- You can use RNNs such as LSTM or GRU to encode language
- A famous example is sequence to sequence learning for translation
- Encoder-decoder architecture
The age of Transformers!

• Unlike recurrent you can run it in parallel
• Main ideas is to use multi-head self-attention
• Self-attention looks into the similarity in the input space to see which one should be taken into account
• Multiple attention can encode different information
• Position embedding helps remembering the order – otherwise it becomes bag of words!

Figure 1: The Transformer - model architecture.
BERT – Devlin et al 2019

• Unsupervised pre-training on multi-layer transformer
  • Mask 15% of the words and train a model to predict them
  • Predict the next sentence
• Can give sentence and contextualized embedding
• Difference between BERT-base and BERT-large is the depth of the model
Voice prosodic measurements

- Pitch/f0 tracking and contour
- Articulation rate and pause timings
- Mel frequency cepstral coefficients (MFCC)
  - Compact representation of the spectrum
  - Emulates human hearing
  - Popular for speech recognition
Deep spectrum features (voice)


Questions?

- Representation learning application for you?
- When is it useful and when do you just use handcrafted features?
Multimodal fusion
Multimodal fusion

- Fusing information from multiple modalities
  - Examples:
    - Audiovisual speech recognition
    - Audiovisual emotion recognition
    - Multimodal biometrics (e.g., face and fingerprint)

- Fusion techniques
  - Model free
    - Early, late and hybrid
  - Model-based
    - Multiple kernel learning
    - Neural networks
Model free approaches – early fusion

- Easy to implement – just concatenate the features
- Exploit dependencies between features
- Can end up very high dimensional
- More difficult to use if features have different framerates
Model free approaches – late fusion

- Train a unimodal predictor and a multimodal fusion one
- Requires multiple training stages
- Do not model low level interactions between modalities
- Fusion mechanism can be voting, weighted sum or an ML approach
Model free approaches – hybrid fusion

- Combine benefits of both early and late fusion mechanisms
Model-based: Joint Representation

• For supervised learning tasks

• Joining the unimodal representations:
  • Simple concatenation
  • Element-wise multiplication or summation
  • Multilayer perceptron

• How to explicitly model both unimodal and bimodal interactions?
Model-based: Canonical correlation analysis

- Projecting different modalities/views into spaces where the correlation is maximized
- Can be sensitive to noise
- Not ideal if the information is complementary

\[(a', b') = \arg \max_{a, b} \text{corr}(a^T X, b^T Y)\]
Model-based: Multimodal Tensor Fusion Network (TFN)

Can be extended to three modalities:

\[ h_m = \begin{bmatrix} h_x \\ 1 \end{bmatrix} \otimes \begin{bmatrix} h_y \\ 1 \end{bmatrix} \otimes \begin{bmatrix} h_z \\ 1 \end{bmatrix} \]

Explicitly models unimodal, bimodal and trimodal interactions!

[Zadeh, Jones and Morency, EMNLP 2017]
Model-based: Multimodal Transformer

- Cross-modal attention for alignment
- Attention mechanism helps with (long-term) temporal dependency
- What is the limitation here?
Question

- Imagine modalities have a high level of correlation/interaction – which fusion approach is better?
Case study 1

EEG Emotion recognition
Brain and emotions

- The limbic system
  - Emotional significance
  - Coordination of emotional behavior
- Frontal brain lateralization
  - Right frontal: withdrawal
  - Left frontal: approach
- Other patterns
  - Synchronization of different neuronal populations
Electroencephalogram (EEG)

Weak electrical activity from postsynaptic potentials generated in superficial layers of the cortex
Convolutional neural nets for EEG

Based on Bashivan et al., ICLR, 2016

No pooling
Cross-modal learning

- What if we use one modality with stronger association for alignment
- Behavior (facial expression) has a better performance for emotion recognition

Smile and EEG Correlation
Cross-modal representation learning

- Jointly learn the other modality + class labels
- Representation can be applied to datasets without the behavioral modality
Cross-modal representation learning

<table>
<thead>
<tr>
<th></th>
<th>MAHNOB</th>
<th>MAHNOB</th>
<th>MAHNOB</th>
<th>MAHNOB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Between-database</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deep CNN</td>
<td>64.8</td>
<td>62.9</td>
<td>50.3</td>
<td>50.0</td>
</tr>
<tr>
<td>Expression</td>
<td>64.6</td>
<td>63.4</td>
<td>53.0</td>
<td>52.7</td>
</tr>
<tr>
<td>Emo-expression</td>
<td><strong>65.4</strong></td>
<td><strong>64.5</strong></td>
<td><strong>55.8</strong></td>
<td><strong>55.3</strong></td>
</tr>
<tr>
<td><strong>Transfer learning and domain adaptation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADDA</td>
<td>67.2</td>
<td>65.3</td>
<td>54.0</td>
<td>53.9</td>
</tr>
<tr>
<td>Deep CNN</td>
<td>71.8</td>
<td>70.7</td>
<td>57.9</td>
<td>51.8</td>
</tr>
<tr>
<td>Expression</td>
<td>71.8</td>
<td>70.8</td>
<td>55.1</td>
<td>54.4</td>
</tr>
<tr>
<td>Emo-expression</td>
<td><strong>73.5</strong></td>
<td><strong>72.5</strong></td>
<td><strong>56.1</strong></td>
<td><strong>55.7</strong></td>
</tr>
</tbody>
</table>

Multimodal Gated Fusion

\[ H_{\text{fusion}} = [H_{\text{EEG}} \odot W_{\text{EEG}} \oplus H_{\text{Face}} \odot W_{\text{Face}}] \]

## Fusion results – within database

### Within-database

<table>
<thead>
<tr>
<th>Classifier Type</th>
<th>Test</th>
<th>Valence</th>
<th>Arousal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CR</td>
<td>F1</td>
</tr>
<tr>
<td><strong>EEG CNN</strong></td>
<td>DAI-EF</td>
<td>69.5</td>
<td>67.2</td>
</tr>
<tr>
<td><strong>MLP on face features</strong></td>
<td>DAI-EF</td>
<td>73.1</td>
<td>68.2</td>
</tr>
<tr>
<td><strong>Concatenate fusion</strong></td>
<td>DAI-EF</td>
<td>74.1</td>
<td>73.1</td>
</tr>
<tr>
<td><strong>Tensor fusion</strong></td>
<td>DAI-EF</td>
<td>74.1</td>
<td>73.0</td>
</tr>
<tr>
<td><strong>Gated Fusion</strong></td>
<td>DAI-EF</td>
<td>74.8</td>
<td>73.4</td>
</tr>
<tr>
<td><strong>Coordinated (Cosine)</strong></td>
<td>DAI-EF</td>
<td>71.1</td>
<td>68.8</td>
</tr>
<tr>
<td><strong>Gated coordinated fusion</strong></td>
<td>DAI-EF</td>
<td>75.4</td>
<td>74.1</td>
</tr>
</tbody>
</table>
Summary - EEG

• Behavior is a strong emotional signal
• Behavioral activity shows up in EEG signals
• Cross-modal relationship can be leveraged to improve emotion recognition from EEG
Question

- Where do you think EEG emotion recognition is useful?
Case study 1

Hierarchical fusion for detecting humorous utterances

Detecting humorous utterances

• A context-aware hierarchical multi-modal fusion network for the task of punchline detection

“Nervous, I went down to the street to look for her. Now, I did not speak Portuguese. I did not know where the beach was. I could not call her on a cell phone because this was 1991, and the aliens had not given us that technology yet”
Data and evaluation

• UR-FUNNY database
• 8257 humorous and non-humorous punchlines from TED talks
• Diverse in terms of topics and speakers
• 1866 videos, 1741 Speakers, 417 topics.
• Multimodal involving text, audio and visual modalities
• Each punchline is labelled humorous/non-humorous.
• Around 64%, 16%, and 20% of data was used for training, validation, and testing
Hierarchical fusion

Figure 1: Hierarchical Trimodal Encoder
Context modelling

Figure 2: Context Aware Hierarchical Trimodal Network
Results

- MFN: memory Fusion Network
- TFN: Tensor Fusion Network
- EF: Early Fusion
- FF: Flat Fusion
- MF: Merge Fusion

<table>
<thead>
<tr>
<th>Modality</th>
<th>T</th>
<th>A+V</th>
<th>T+A</th>
<th>T+V</th>
<th>T+A+V</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-MFN (P)</td>
<td>62.85</td>
<td>53.3</td>
<td>63.28</td>
<td>63.22</td>
<td>64.47</td>
</tr>
<tr>
<td>C-MFN (P + C)</td>
<td>64.44</td>
<td>57.99</td>
<td>64.47</td>
<td>64.22</td>
<td>65.23</td>
</tr>
<tr>
<td>TFN (P)</td>
<td>64.52</td>
<td>56.47</td>
<td>66.47</td>
<td>66.53</td>
<td>66.77</td>
</tr>
<tr>
<td>EF (P)</td>
<td>64.64</td>
<td>56.43</td>
<td>65.83</td>
<td>65.50</td>
<td>65.92</td>
</tr>
<tr>
<td>MF (P)</td>
<td>64.64</td>
<td>57.32</td>
<td>66.44</td>
<td>65.74</td>
<td>67.05</td>
</tr>
<tr>
<td>FF (P)</td>
<td>64.64</td>
<td>57.24</td>
<td>66.50</td>
<td>66.41</td>
<td>66.68</td>
</tr>
<tr>
<td>HF (P)</td>
<td>64.64</td>
<td>57.24</td>
<td>66.50</td>
<td>66.41</td>
<td>67.08</td>
</tr>
<tr>
<td>HF (P + C)</td>
<td><strong>64.72</strong></td>
<td><strong>57.42</strong></td>
<td><strong>66.68</strong></td>
<td><strong>66.48</strong></td>
<td><strong>67.84</strong></td>
</tr>
</tbody>
</table>
Summary

• Multimodal model better captures humor

• Language performs the best in unimodal models

• Hierarchical fusion better captures the inter-modality interactions for humor – maybe!

• Incorporating the context of punchline can boost the accuracy of prediction
Summary

- Emotions are multi-faceted and have manifestations in different modalities
- Representation learning enables machine learning models to learn a useful representation without the need to handcraft new features
- Multimodal fusion increases robustness and take advantage of complementary information
- Often times multimodal fusion yields superior performance