Topics in AI (CPSC 532S):
Multimodal Learning with Vision, Language and Sound

Lecture 13: Coordinated Representations and Joint Embeddings
Multimodal Representations

What is a good multimodal representation?

— **Similarity** in the representation (somehow) implies similarity in corresponding concepts (we saw this in word2vec)

— **Useful** for various discriminative tasks (retrieval, mapping, fusion, etc.)

— Possible to obtain in absence of one or mere modalities

— **Fill in missing modalities** given others (map or translate between modalities)

*slide from Louis-Philippe Morency*
**Multimodal Representation Types**

**Joint** representations:

- Simplest version: *modality concatenation* (early fusion)
- Can be learned *supervised* or *unsupervised*

*slide from Louis-Philippe Morency*
Multimodal Representation Types

**Joint** representations:

- Simplest version: *modality concatenation* (early fusion)
- Can be learned *supervised* or *unsupervised*

**Coordinated** representations:

- *Similarity-based* methods (e.g., cosine distance)
- *Structure constraints* (e.g., orthogonality, sparseness)
- Examples: CCA, joint embeddings

*slide from Louis-Philippe Morency*
Multimodal Representation Types

**Joint** representations:

- Simplest version: modality concatenation (early fusion)
- Can be learned supervised or unsupervised

*slide from Louis-Philippe Morency*
Joint Representation: Simple Multimodal Autoencoders

**Concatenating** modalities is fine, but requires both modalities at test time.

No ability to ensure there is indeed **sharing** in the representations space.
Each \textbf{modality} can be pre-trained
  - using denoising autoencoder

To train the model, \textbf{reconstruct both modalities} using
  - both Audio & Video
  - just Audio
  - just Video

*slide from Louis-Philippe Morency*
Multimodal Research: Historical Perspective

The McGurk Effect

McGurk Effect (1976)

* video credit: OK Science

* Adopted from slides by Louis-Philippe Morency
Joint Representation: Deep Multimodal Autoencoders

[Ngiam et al., 2011]

*slide from Louis-Philippe Morency

---

**Table 3: McGurk Effect**

<table>
<thead>
<tr>
<th>Audio / Visual Setting</th>
<th>Model prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ga/ Audio /ga/</td>
<td>82.6% 2.2% 15.2%</td>
</tr>
<tr>
<td>/ba/ Audio /ba/</td>
<td>4.4% 89.1% 6.5%</td>
</tr>
<tr>
<td>/ga/ Audio /ba/</td>
<td>28.3% 13.0% 58.7%</td>
</tr>
</tbody>
</table>
Joint Representation: Deep Multimodal Autoencoders

[Ngiam et al., 2011]

Useful when you know you may only be conditioning on one modality at test time

Can be regarded as a form of regularization

*slide from Louis-Philippe Morency*
Supervised Joint Representation

For supervised learning tasks, we need to join unimodal representations

- Simple concatenation
- Element-wise multiplicative interactions
- Many many others

Encoder-decoder Architectures

*e.g. Sentiment*
For supervised learning tasks, we need to join unimodal representations

- Simple **concatenation**

\[
\mathbf{h}_m = \sigma(\mathbf{W} \cdot [\mathbf{h}_x, \mathbf{h}_y, \mathbf{h}_z]^T)
\]

**Multi-modal Sentiment Analysis**

MOSI dataset (Zadeh et al, 2016)

- 2199 subjective video segments
- Sentiment intensity annotations
- 3 modalities: text, video, audio

*slide from Louis-Philippe Morency*
Bilinear Pooling

For supervised learning tasks, we need to join unimodal representations:

- Simple **concatenation**
- Element-wise **multiplicative** interactions

\[ h_m = h_x \otimes h_y \]

[ Tenenbaum and Freeman, 2000 ]

*slide from Louis-Philippe Morency*
Multimodal Tensor Fusion Network (TFN)

For supervised learning tasks, we need to join unimodal representations

- Simple **concatenation**
- Element-wise **multiplicative** interactions

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

[ Zadeh, Jones and Morency, EMNLP 2017 ]
Multimodal Tensor Fusion Network (TFN)

For supervised learning tasks, we need to join unimodal representations

- Simple **concatenation**
- Element-wise **multiplicative** interactions

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

[Zadeh, Jones and Morency, EMNLP 2017]

*slide from Louis-Philippe Morency*
Low-rank Tensor Fusion

Tucker tensor decomposition leads to MUTAN fusion

[Ben-younes et al., ICCV 2017]
Supervised Joint Representation

For supervised learning tasks, we need to join unimodal representations

- Simple **concatenation**
- Element-wise **multiplicative** interactions

**Encoder-decoder** Architectures