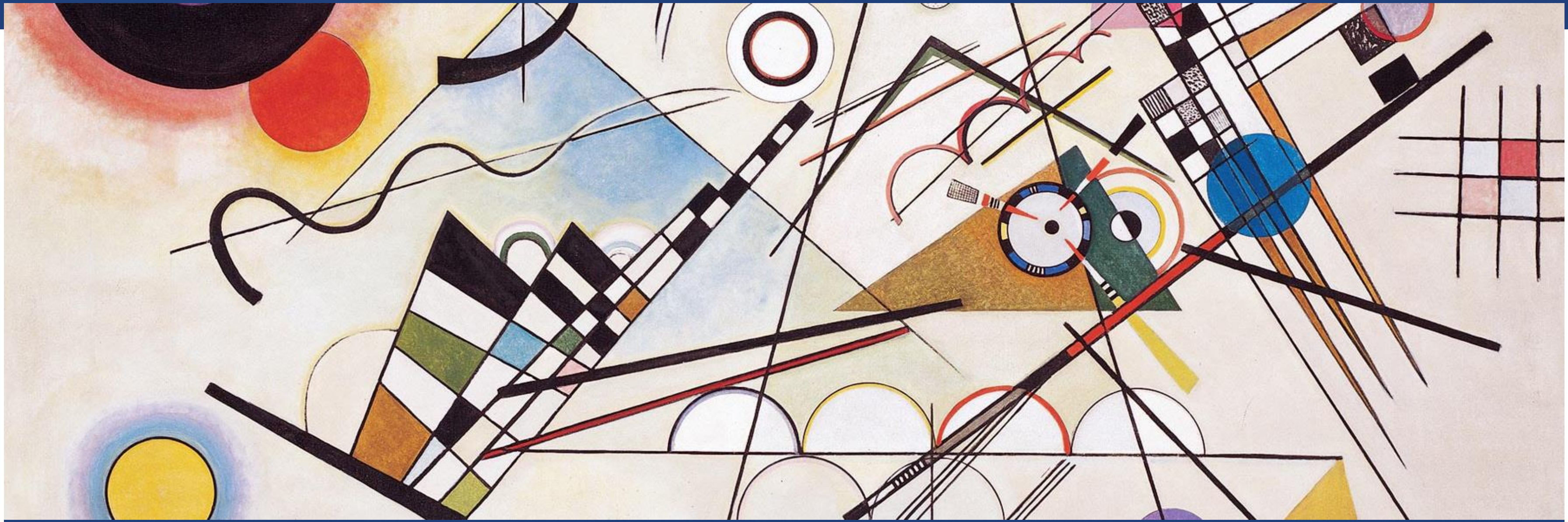


Multimodal Emotion Recognition using Lexico-Acoustic Language Descriptions

Gustavo Cid Ornelas

Motivation

- Emotions are ubiquitous in human communication
- Emotionally-aware intelligence for more natural human-computer interaction



Motivation

- Human experience is multimodal

- Incorporate multimodality to machine learning models
- **Our problem:** consider **text** and **audio** to recognize the emotion (e.g. *happy, sad,* …)

Agenda

1. Challenges in multimodal emotion recognition

2. From Neural Machine Translation (NMT) to multimodal emotion recognition

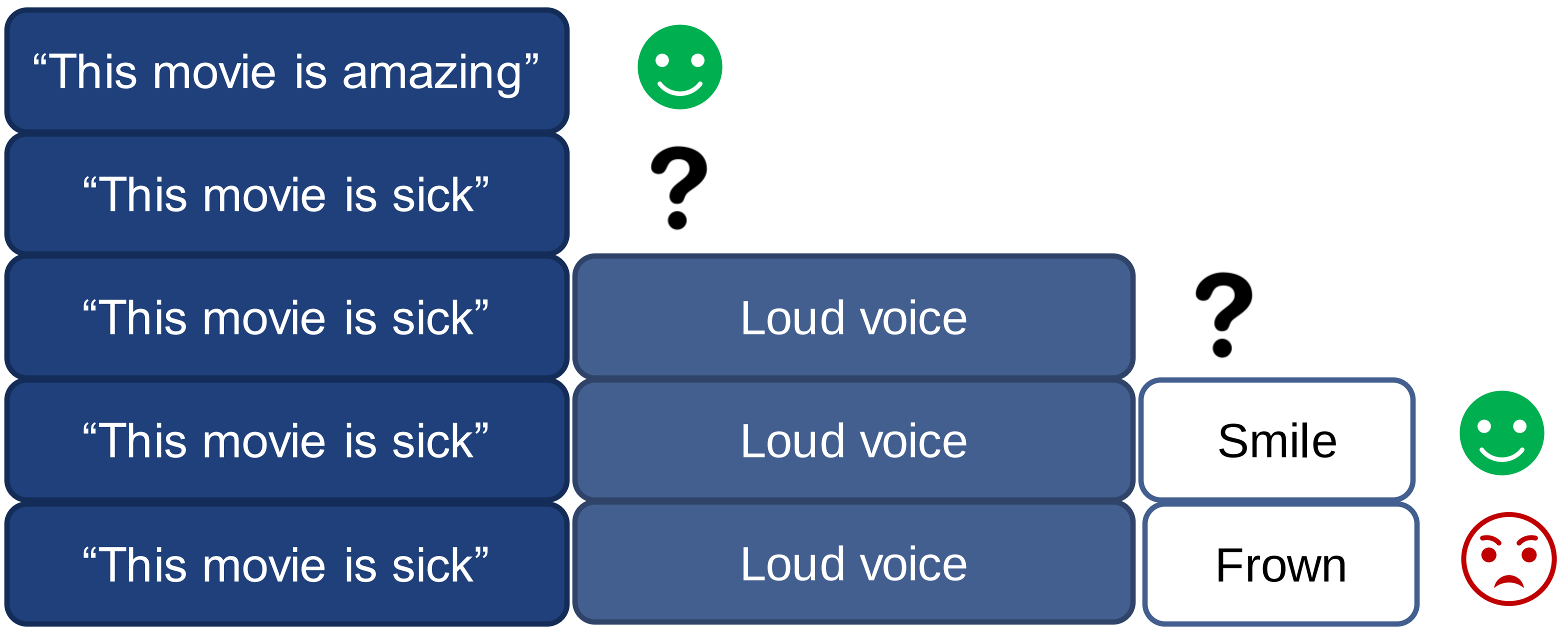
3. Results & conclusion

ETH zürich

Challenges in multimodal emotion recognition

Challenges in multimodal emotion recognition

- Single modality can be inconclusive (adapted from [Zadeh et al., 2018])
- Exploration of intra and inter-modality dynamics



Challenges in multimodal emotion recognition

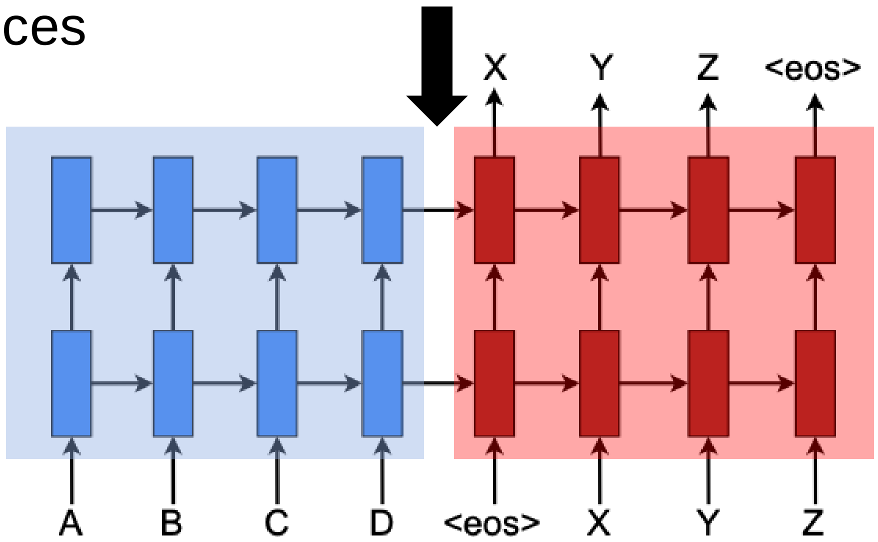
- Classical approaches
  - Feature-level fusion → intra-modality dynamics
  - Decision-level fusion → inter-modality dynamics
- Alternative approaches
  - Learning joint representations, tensor fusion, …

A good model explores to the fullest the **interplay** between the intra and inter-modality dynamics

From NMT to multimodal emotion recognition

From NMT to multimodal emotion recognition

- Encoder-decoder architecture
  - Encoder generates the context vector
  - Decoder uses the context vector to generate the translation
- Context vector
  - Bottleneck for longer sentences



[Luong et al., 2015]

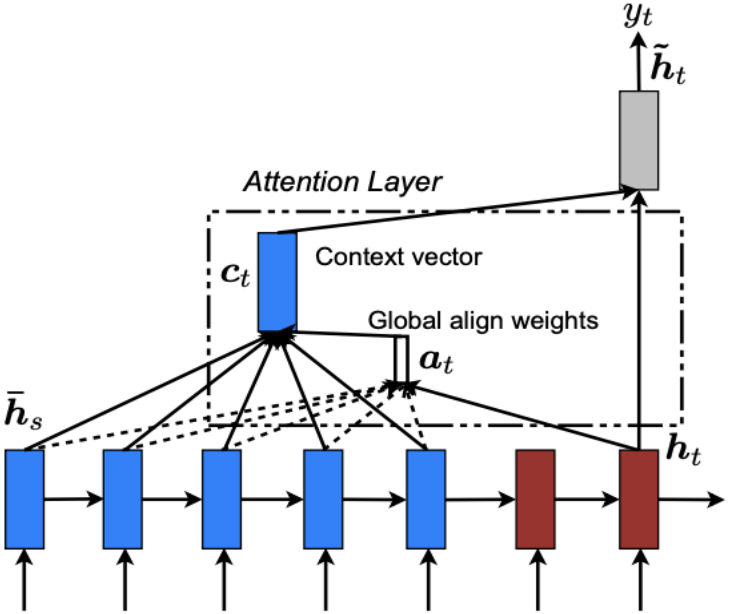
From NMT to multimodal emotion recognition

- Decoder recurrently consulting the encoders hidden states throughout the translation process
- Attention mechanism to generate the context vector
  - Global attention
  - Local attention

$$\mathbf{c_t} = \sum_{s=1}^{n} \mathbf{a_t}(s)\bar{\mathbf{h}}_\mathbf{s}$$

$$\mathbf{a_t}(s) = \mathrm{align}(\mathbf{h_t}, \bar{\mathbf{h}}_\mathbf{s}) = \frac{\exp(\mathrm{score}(\mathbf{h_t}, \bar{\mathbf{h}}_\mathbf{s}))}{\sum_{s'=1}^{n} \exp(\mathrm{score}(\mathbf{h_t}, \bar{\mathbf{h}}_{\mathbf{s'}}))}$$

$$\mathrm{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s & \textit{dot} \\ \boldsymbol{h}_t^\top \boldsymbol{W_a} \bar{\boldsymbol{h}}_s & \textit{general} \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W_a}[\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s]\right) & \textit{concat} \end{cases}$$



[Luong et al., 2015]

From NMT to multimodal emotion recognition

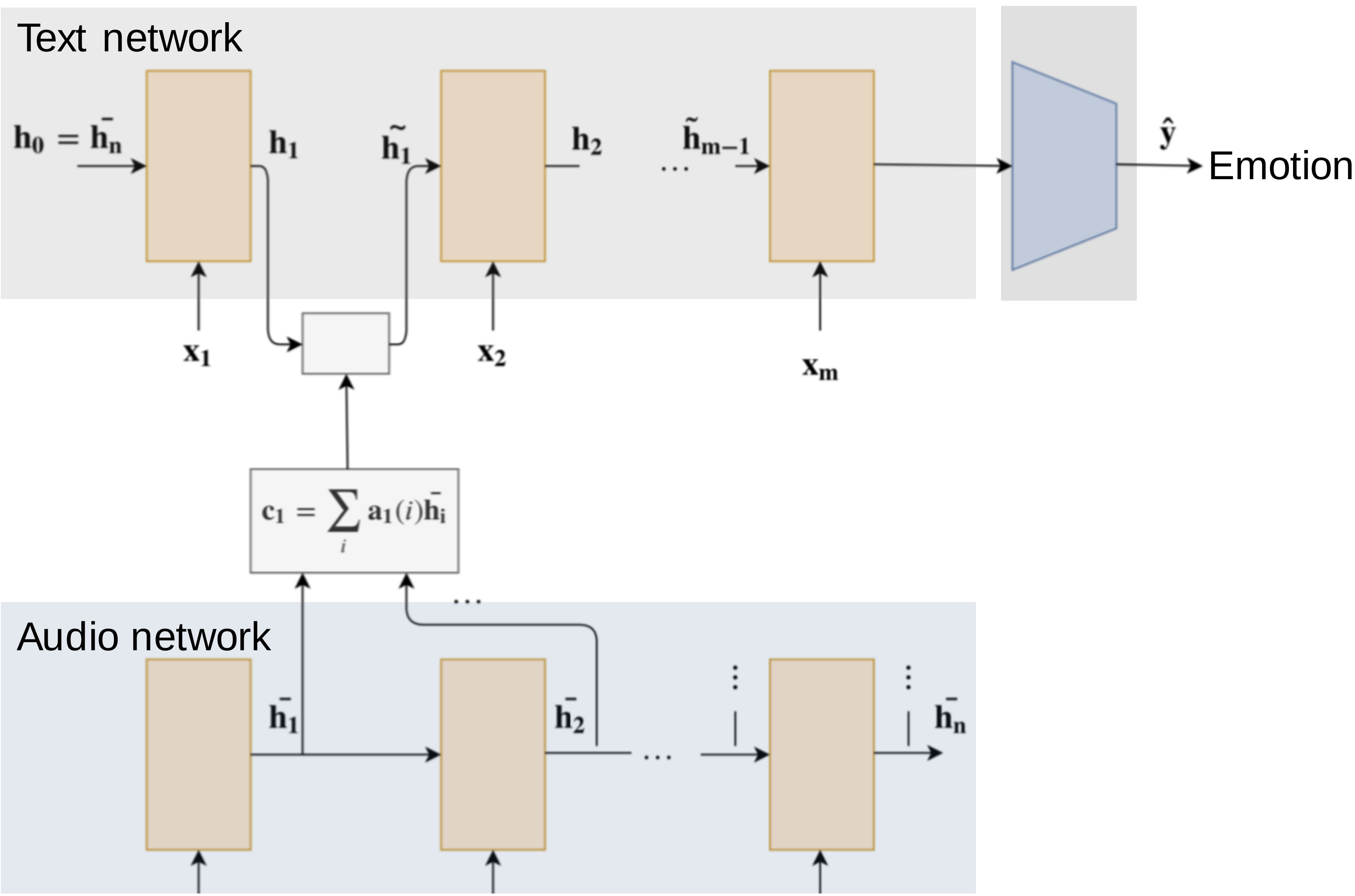
- Connections between the problems
  - Audio and text: one RNN for each
  - Encoder/decoder – Main/auxiliary modalities
  - Main modality places attention on the auxiliary modality

- Choices involved
  - Main modality: text
  - Auxiliary modality: audio
  - Attention mechanism: global attention and attention score [Bahdanau et al., 2014]

From NMT to multimodal emotion recognition

- Text network [Yoon et al., 2018]
  - Embedding layer, initialized with pre-trained GloVe embedding
  - RNN with GRU cells

- Audio network
  - Raw audio as input
  - Convolutional layers for feature extraction
  - RNN with GRU cells



Text network
$\mathbf{h_0} = \bar{\mathbf{h}}_\mathbf{n}$
$\mathbf{h_1}$
$\tilde{\mathbf{h}}_\mathbf{1}$
$\mathbf{h_2}$
$\tilde{\mathbf{h}}_\mathbf{m-1}$
$\hat{\mathbf{y}}$
Emotion
$\mathbf{x_1}$
$\mathbf{x_2}$
$\mathbf{x_m}$
$\mathbf{c_1} = \sum_i \mathbf{a_1}(i)\bar{\mathbf{h}}_\mathbf{i}$
Audio network
$\bar{\mathbf{h}}_\mathbf{1}$
$\bar{\mathbf{h}}_\mathbf{2}$
$\bar{\mathbf{h}}_\mathbf{n}$

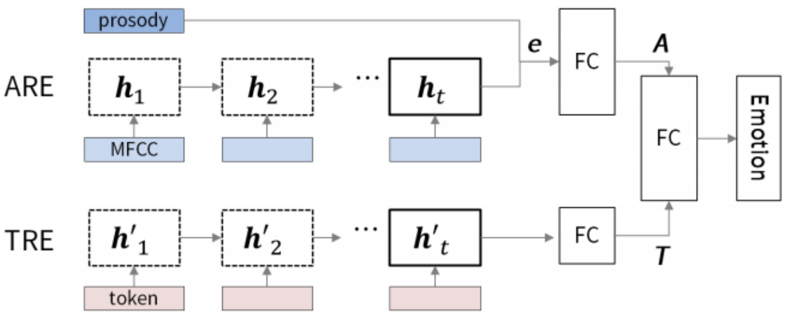
Results & conclusion

Results & conclusion

- IEMOCAP dataset
  - Raw audio files and transcriptions
    - Happy (1,636 samples)
    - Sad (1,084 samples)
    - Angry (1,103 samples)
    - Neutral (1,708 samples)

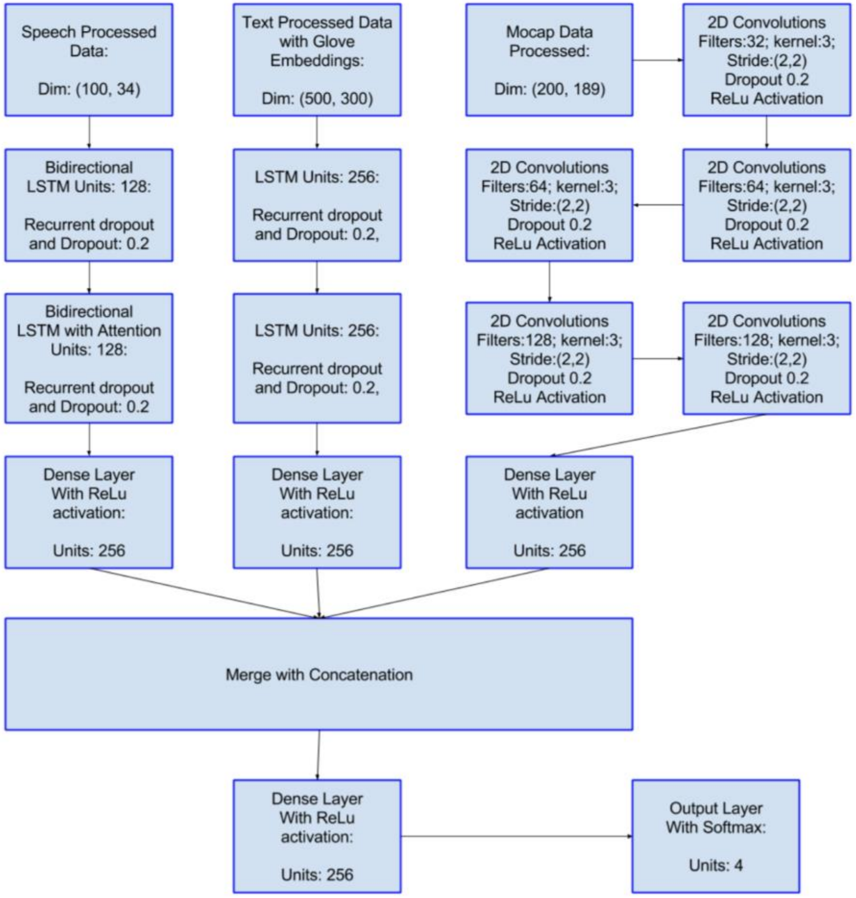
- State-of-the-art: 71% classification accuracy [Yoon et al., 2018]
- Popular models that get close to the state-of-the-art:
  - Decision-level fusion
  - Use of hand-engineered audio features, such as MFCC and pitch

Results & conclusion



[Yoon et al., 2018]

[Tripathi et al., 2018]

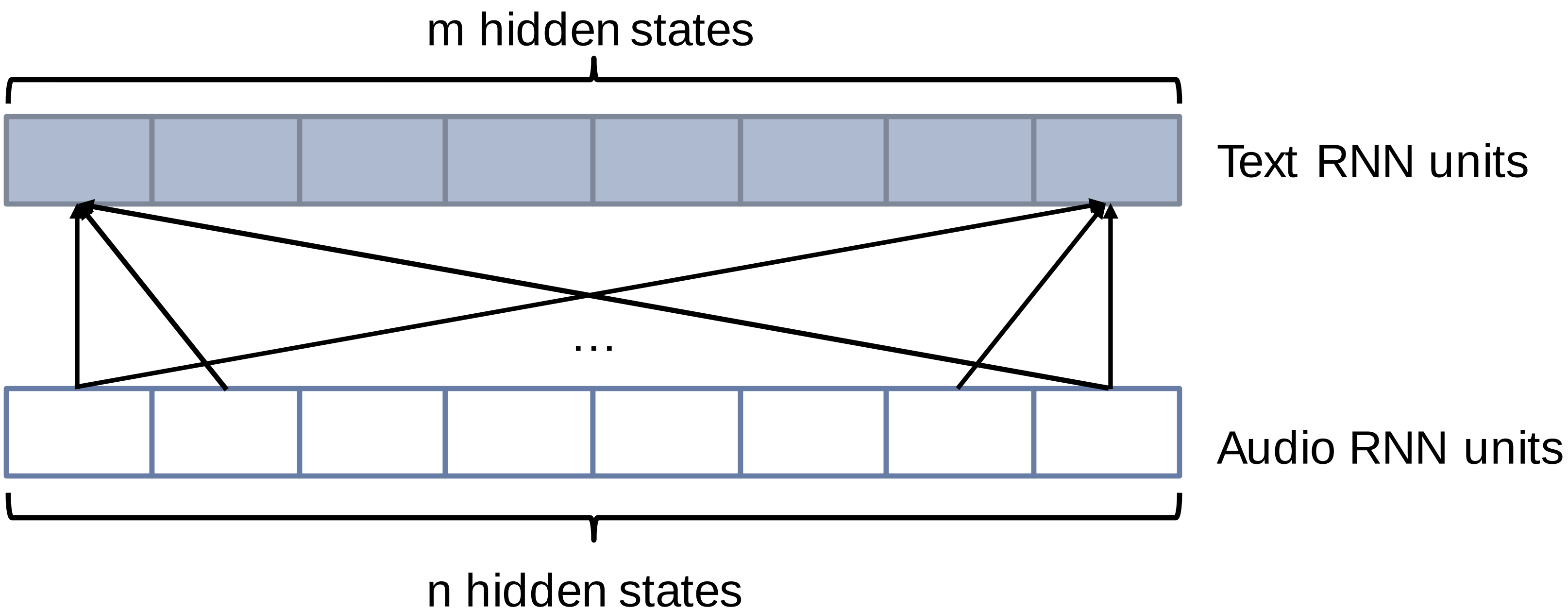


Results & conclusion

- Classification accuracy with audio only: 52.7%
  - Very close to the state-of-the-art models using only audio
  - Our model is completely end-to-end
- Classification accuracy with text only: 62.5%
  - Reproducing the results from [Yoon et al., 2018], which sets the state-of-the-art
- Classification accuracy of the multimodal model: 61%
  - Better than audio, but worse than text
  - **What's going on?**

Results & conclusion

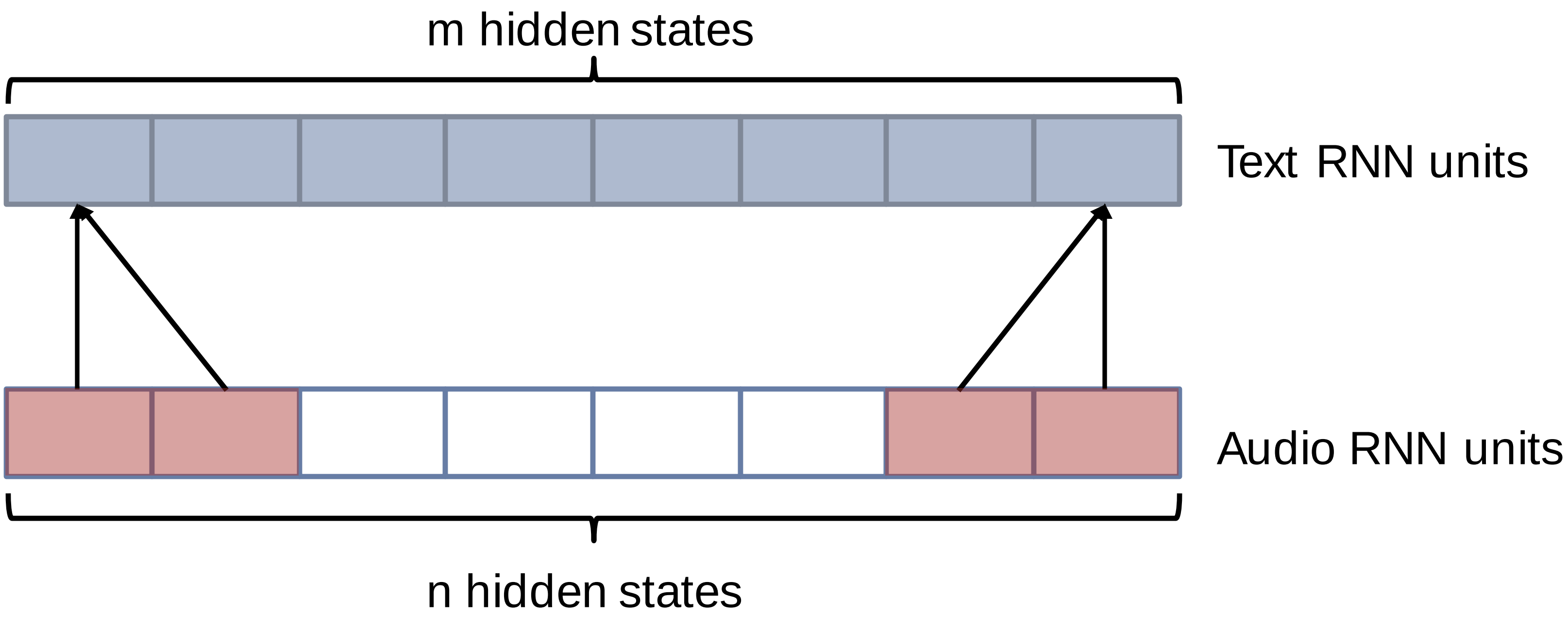
- Multimodal model quickly overfits!
- Attention mechanism with too much flexibility
- Our model: $O(nm)$ parameters for alignment



- Learning this alignment is challenging!

Results & conclusion

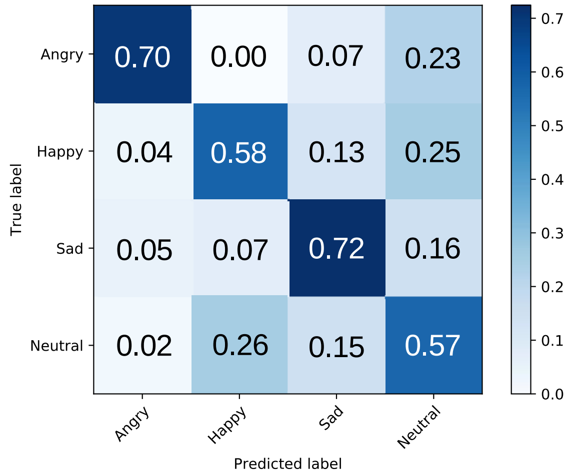
- Possible solutions: sliding window or forced alignment
- Sliding window, with window size w: $O(mw)$ parameters
  - Monotonicity in the audio and text data



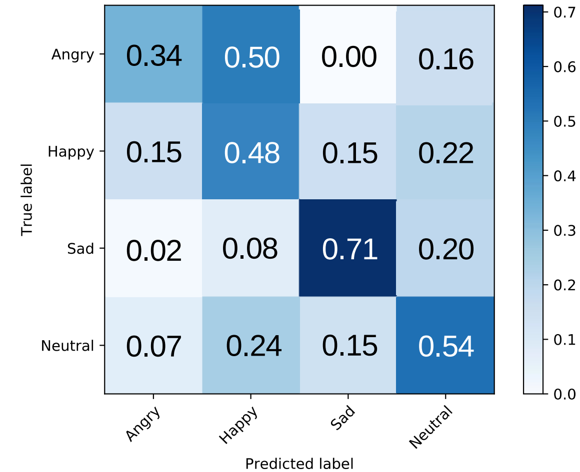
Results & conclusion

- Forced alignment: know exactly in which moment a certain word is spoken
  - Wouldn't need to learn the alignment at all

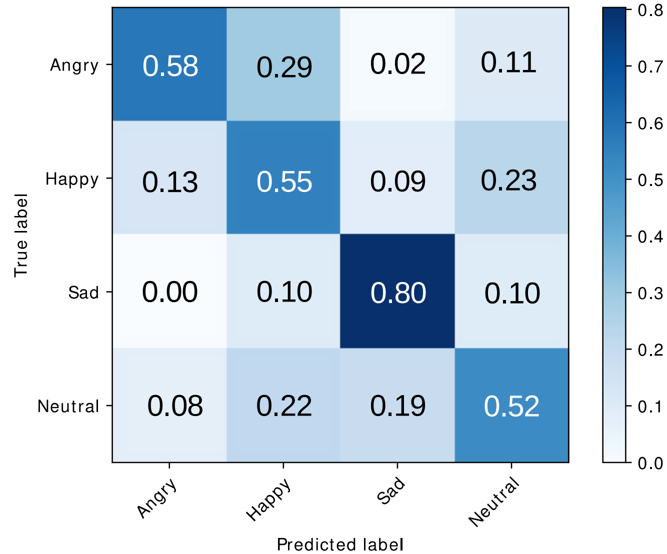
Results & conclusion



Text only



Audio only



Multimodal

Results & conclusion

- Emotion recognition with audio and text
- Proposed a novel architecture inspired on NMT
  - Better explore the interplay between the dynamics and end-to-end
- Idea has potential!
- IEMOCAP dataset is relatively small

- Future work
  - Implement local attention or forced alignment
  - Performance on a larger dataset

ETH zürich

Thank you!