Multimodal Emotion Recognition in Conversation

Carnegie Mellon University

Liyun Tu, Qichen Fu, Shengli Zhu, Xi Chen, Xiaoyu Sun

Team 7

Motivation

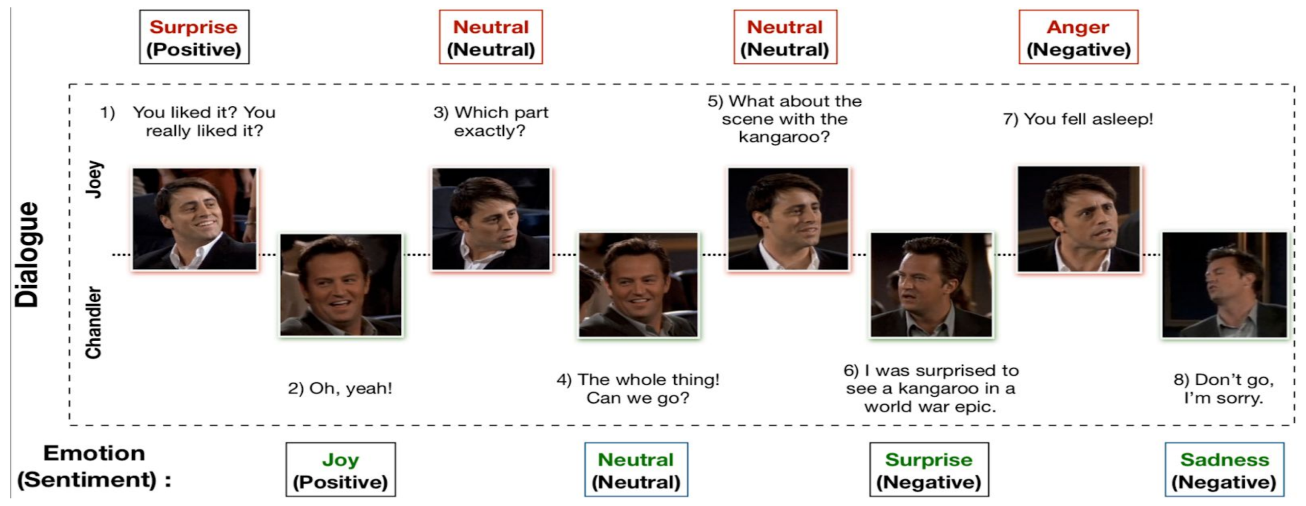
- Emotion Recognition in Conversations (ERC) has a wide variety of applications in multiple domains, including but not limited to customer intelligence, healthcare systems and education.
- The goal is to design and deploy effective and scalable conversational emotion-recognition solutions that could recognize and track people's emotions in multi-party dialogues.

Challenges

- **Multimodal emotion recognition**: to figure out how to develop and adapt algorithms/solutions on multi-party data.
- Existing SOTA solutions in ERC only leveraged textual data → remains challenges in **multimodal feature extraction**, **fusion and alignment** for deploying a multimodal solution
- **Adapting algorithms to multiple interlocutors** and **de-noising acoustic modality** since MELD data are based on TV series.

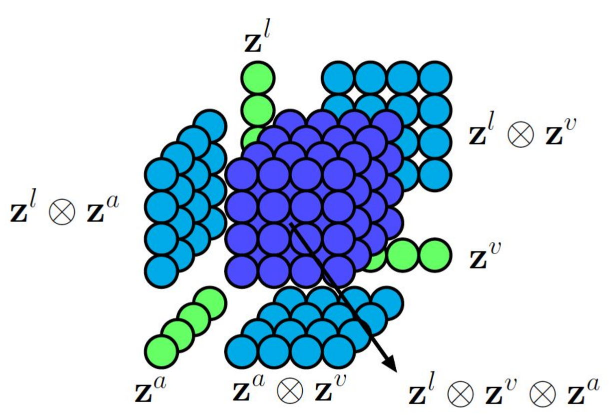
Dataset

- **MELD**: Multimodal EmotionLines Dataset" multimodal sentiment/emotion recognition dataset
  - Multi-modal data for conversations from Friends TV series
  - More than 1300 dialogues and 13000 utterances
- Two sets of labels
  - **Emotion** labels: {Anger, Disgust, Sadness, Joy, Neutral, Surprise, Fear}
  - **Sentiment** labels: {Positive, Negative, Neutral}



Research Idea: BC-LSTM with Tensor Fusion

- Vanila BC-LSTM network uses **two-level hierarchical LSTM-based structure** to first extract **contextualized** unimodal representations and then generate contextualized multimodal representations.
- Tensor Fusion captures **both intermodal and intramodal dynamics,** which can used to get more informative contextualized multimodal representations.

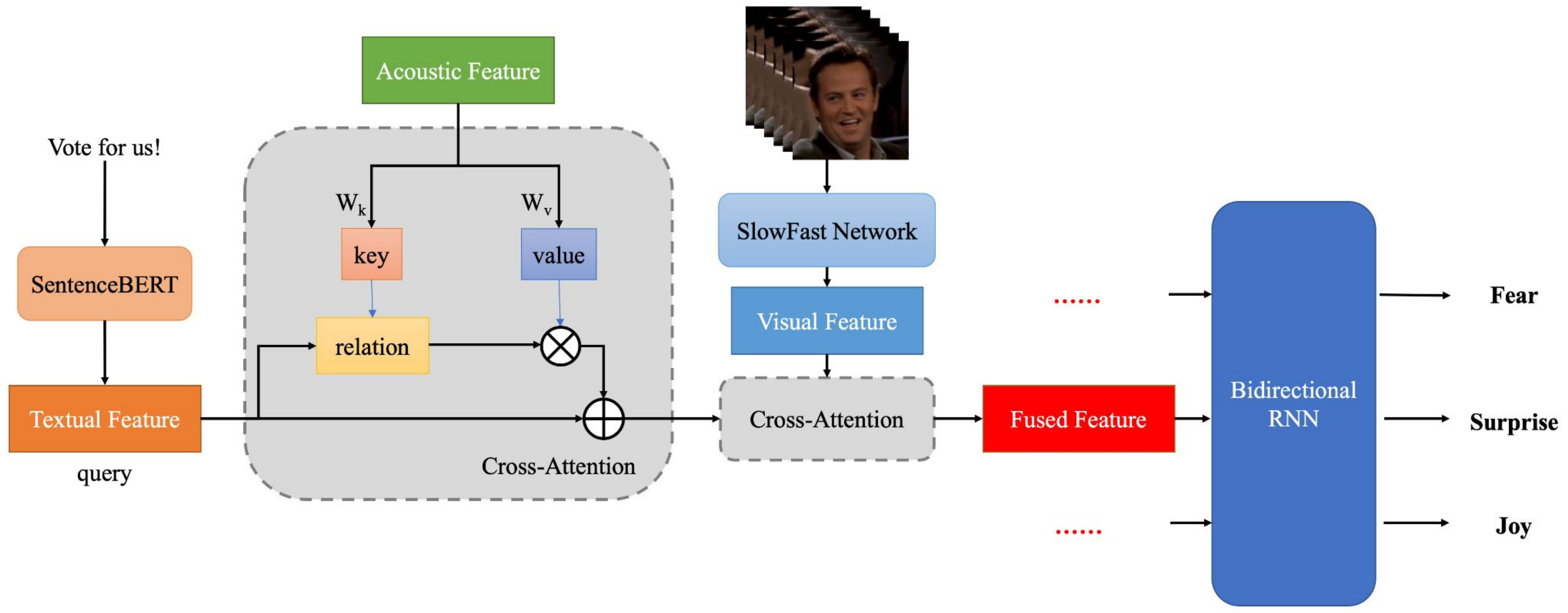


Tensor Fusion with 3 modalities
(we only use language and acoustic for BC-LSTM)

Research Idea: Incorporating Third Modality - Visual

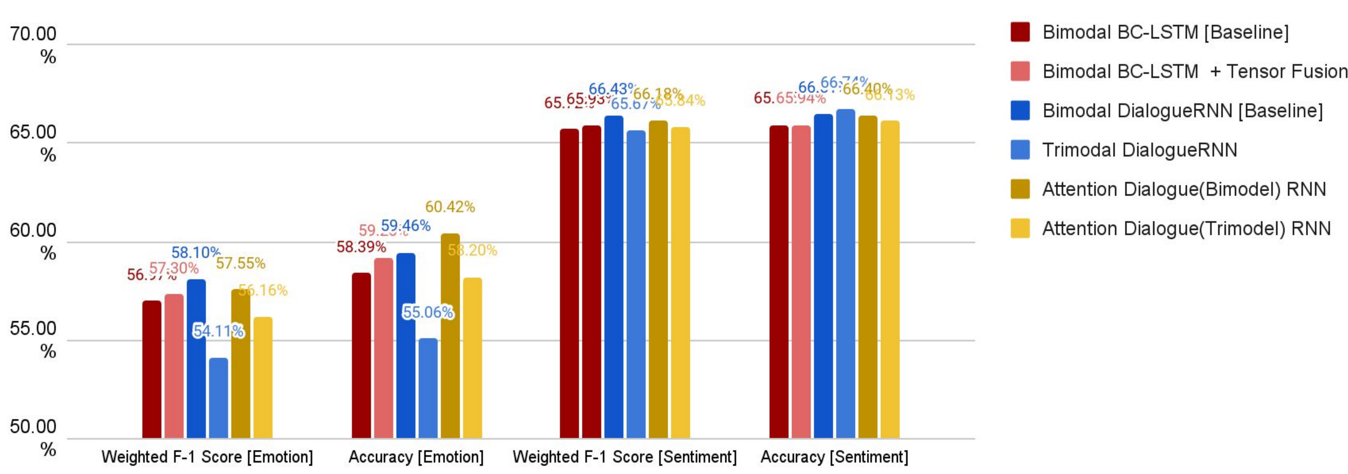
- Visual info in the scene is highly indicative to emotion. ⟹
- Use SlowFast network as CNN-based feature extractor to acquire the representation of video clips.
- Incorporate visual features by:
  - Naive fusion - Concatenation
  - Proposed fusion - **Cross Attention** following a **hierarchical** fashion

Emotion - Disgust



AttentionDialogueRNN with Hierarchical Cross Attentions

Results & Analysis



Analysis

- For both model structures (BC-LSTM and DialogueRNN), our proposed ideas achieve comparable performance compared to baseline.
- Better fusion strategy only slightly improve baseline BC-LSTM model, possibly indicating the poor quality of acoustic features is more urgent to be solve.
- Adding visual modality to Dialogue-RNN didn't help much with model's performance, probably due to the reason that video frames often contain multiple faces even only one of them is speaking.

Future Work

- **Speaker Diarization**
  - Recognizing speaker from multi-party scenes to capture the most informative visual information
- **Transformers**
  - Instead of using utterance level representations, use word-level features for language modality
  - Extracting corresponding acoustic and vision representations using word-level timestamps
  - Multiple existing transformer-based multimodal models
    - Multimodal Fusion Transformer
    - Factorized Multimodal Transformer
- **End-to-end Visual Representation Training**
  - Instead of applying model(SlowFast) pre-trained on other tasks, integrate the representation extraction process into the training process.