05/09/2025

Investigating Multimodal Fusion for Emotion Recognition Using Deep Learning

Agenda

01	Executive summary
02	Motivation
03	Background Related works
04	Method & Approach
05	Implementation Details
06	Results & Experimental evaluation
07	Conclusion

Executive summary

- **Problem:**
  - Human emotion recognition is critical for various applications, including human-computer interaction, mental health support, and customer sentiment analysis. However, existing systems often rely on single-modal data (text or audio only), which limits their understanding of complex emotional expressions.
  - Current emotion recognition models often fail to capture subtle emotional cues, especially when relying solely on text or audio. This can lead to misclassification and a lack of emotional intelligence in AI systems.

- **Goal:**
  - Develop a robust multimodal emotion recognition system that integrates text (BERT) and audio (Mel-Frequency Cepstral Coefficients MFCC) features.
  - Demonstrate the effectiveness of multimodal fusion by comparing the performance of single-modality (text or audio) models with multimodal model.

- **Technical Challenges:**
  - Efficiently preprocessing and synchronizing text and audio data from the MELD dataset.
  - Designing a scalable multimodal fusion model that effectively combines BERT text embeddings and MFCC audio features (set of features that represent the short-term power spectrum of sound).
  - Ensuring model generalization without overfitting, given the limited size of the MELD dataset.

Executive summary

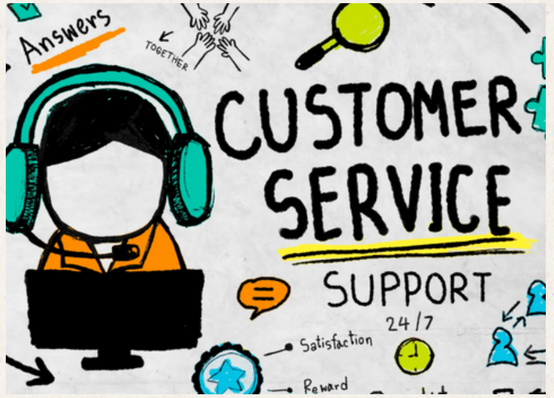
- **Solution Approach:**
  - Text Processing: Fine-tuned BERT model for text-based emotion recognition, leveraging pre-trained language representations.
  - Audio Processing: Extracted MFCC features from audio clips and processed them using a fully connected MLP (Multilayer Perceptron) for classification. The audio features are padded or truncated to a fixed length, then flattened for model input.
  - Fusion Model: Combined BERT text embeddings and MFCC audio features using a fully connected MLP for emotion classification.

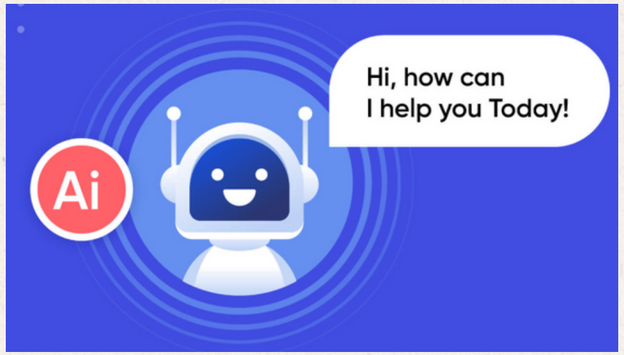
- **Value and Benefits:**
  - Enhanced emotion recognition by leveraging both text and audio modalities, providing a richer understanding of emotional context.
  - Potential applications in:
  - Customer support systems that understand user emotions.
  - Mental health monitoring tools that detect emotional distress.
  - Advanced human-computer interaction systems.
  - Provides a scalable architecture that can be extended to additional modalities (e.g., video-based emotion recognition).

Motivation

- Emotions are complex and expressed through multiple channels, including voice, text, and facial expressions. Single-modality models often miss critical emotional information.
- To bridge this gap by developing a model that can process and integrate both text and audio data simultaneously, capturing semantic and vocal emotions.

- **Real-World Use Cases:**
  - Sentiment analysis for customer feedback in call centers.
  - Mental health monitoring through speech and text analysis.
  - Enhanced chatbot systems capable of understanding user emotions.





Related works

- **Speech Emotion Recognition Using MFCC Features**

https://www.researchgate.net/publication/338987485_Speech_emotion_recognition_using_MFCC_features
Utilized Mel-Frequency Cepstral Coefficients (MFCC) for emotion recognition in speech, but lacks text understanding.

- **Emotion Detection in Text Using Convolutional Neural Network**

https://ieeexplore.ieee.org/document/9967913
Explored text emotion detection using BERT and CNN, but focused only on text data without incorporating audio.

- **Speech Emotion Recognition Using Convolutional Recurrent Neural Network**

https://www.researchgate.net/publication/361277701_Speech_Emotion_Recognition_using_Convolutional_Recurrent_Neural_Network
Combined CNN and BiLSTM for speech emotion recognition but lacked text analysis.

Related works

- **Emotion Recognition Using Convolutional Neural Network (CNN)**

https://iopscience.iop.org/article/10.1088/1742-6596/1962/1/012040
Applied CNN for facial emotion recognition but did not use text or audio for richer emotion understanding.

- **Multimodal Sentiment Analysis Based on BERT and ResNet**

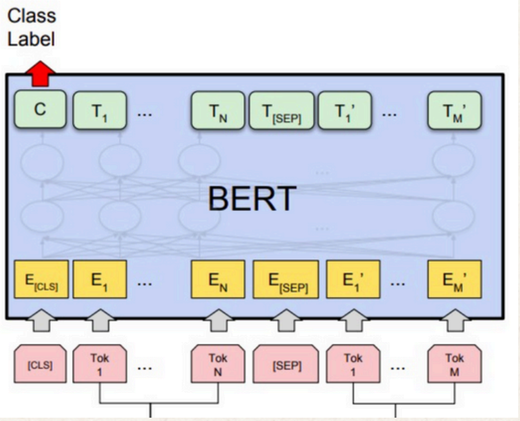
https://arxiv.org/abs/2412.03625v1
Proposes a multimodal sentiment analysis framework using BERT for text and ResNet for image processing.

Limitations of Existing Approaches:

- **Single-Modality Focus:** Most existing works focus on a single modality (text, audio, or image), limiting their ability to capture complex emotional expressions.
- **Lack of Contextual Understanding:** Text-only models lack vocal tone, while audio-only models lack semantic context.
- **High Computational Cost:** Some models employ complex architectures that are computationally expensive without significant performance gains.

Method & Approach

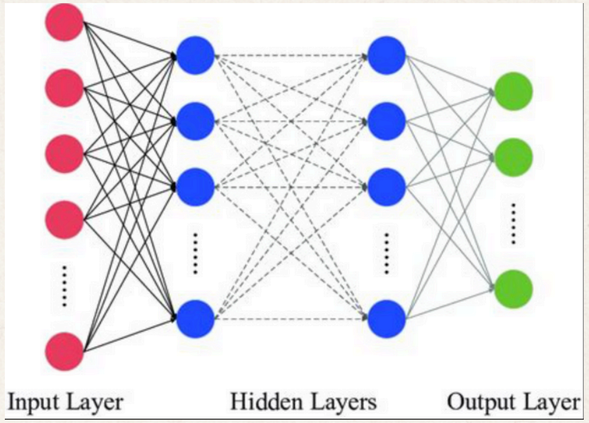
Text Processing (BERT)



- **Model: BERT**
  - A pre-trained model designed to understand text context in a bidirectional manner, meaning it captures information from both the left and right of each word.
  - Our model uses the "bert-base-uncased" version, which is a smaller yet efficient version of BERT.
- **Preprocessing:**
  - **Text Tokenization:** The input text is split into tokens using the BERT tokenizer. This tokenizer converts text into a sequence of integer IDs representing specific subwords.
  - **Maximum Sequence Length:** We set a maximum length of 128 tokens to ensure consistent input size. Shorter texts are padded, and longer texts are truncated.
  - **Embedding Extraction:** We use the CLS token output from BERT, which is a special token designed to capture the overall meaning of the input text.
  - **Embedding Dimension:** The CLS token generates a 768-dimensional vector for each text input, which serves as a dense representation of the text's semantic information.

Method & Approach

Audio Processing (MFCC + MLP)



- **Audio Features (MFCC):**
  - We extract Mel-Frequency Cepstral Coefficients (MFCC) from the audio clips. MFCC is a widely used feature in speech processing, capturing the power spectrum of sound.
  - Number of Coefficients: 13 per frame, representing the key characteristics of sound.
  - Frame Length: Each audio clip is processed into a series of frames, each with 13 MFCC values.
  - Fixed Length: We pad or truncate the audio features to a consistent length of 401 frames. This ensures that all audio inputs have the same size, making them compatible for model training.
- **Model Architecture (MLP for Audio):**
  - The flattened MFCC features (13 × 401) are passed through a Multi-Layer Perceptron (MLP).
  - **Input Layer:** 5213 dimensions.
  - **Hidden Layers:** Two fully connected layers with ReLU activation.
  - **Output Layer:** Number of emotion classes (7 in our case).
  - **Dropout Regularization:** We apply dropout (randomly setting some neurons to zero) within the MLP to prevent overfitting.

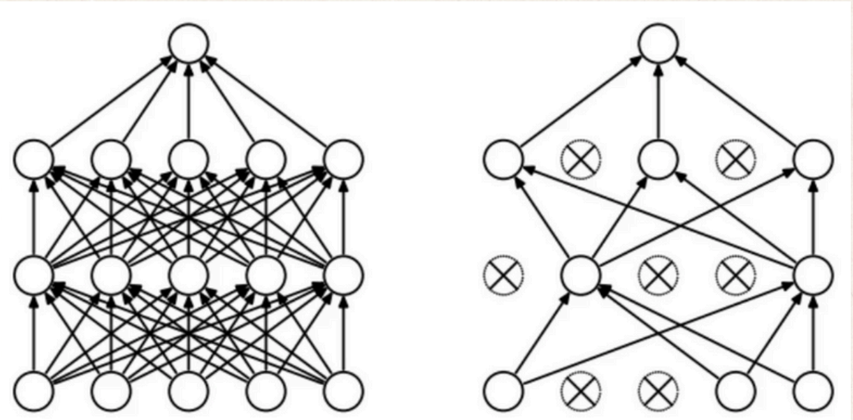
Method & Approach

Fusion Model (Text + Audio Integration)

- We combine the semantic understanding from text (BERT) and acoustic features from audio (MFCC) using a two-path architecture.
- **Text Path:**
  - BERT text embeddings (768 dimensions) are first passed through a fully connected layer that reduces them to 128 dimensions. This step reduces computational complexity and aligns the size of text and audio features.
  - These text embeddings are rich in semantic context, capturing meaning and sentiment from the text.
- **Audio Path:**
  - The raw MFCC features (13 × 401) are flattened into a single vector of 5213 dimensions.
  - These features are flattened into a single vector of 5213 dimensions for compatibility.
  - The flattened audio features are then passed through a fully connected layer (5213 → 128) for dimension reduction.

Method & Approach

Fusion Model (Text + Audio Integration)



- **Fusion Mechanism:**
  - **Dropout Application:**
    - Two dropout layers are used:
      - First Dropout (0.2) after the first ReLU activation.
      - Second Dropout (0.2) after the second ReLU activation.
    - Dropout is used to prevent overfitting by randomly setting a fraction of neurons to zero during training.
  - **Normalization Application:**
    - Before feeding audio into the model, we standardize it by subtracting the mean and dividing by the standard deviation, ensuring consistent scale and stable training.

Implementation

BERT Text Classifier

- Use a pre-trained BERT model: bert-base-uncased
- Model: 2-layer MLP on BERT's CLS tokens
  - hidden layer: ReLu activation, 128 hidden units
- Hyperparameters:
  - Loss Function: Cross Entropy Loss
  - Optimizer: AdamW
  - Learning rate: 2e-5
  - Batch Size: 8
  - Epochs: 10

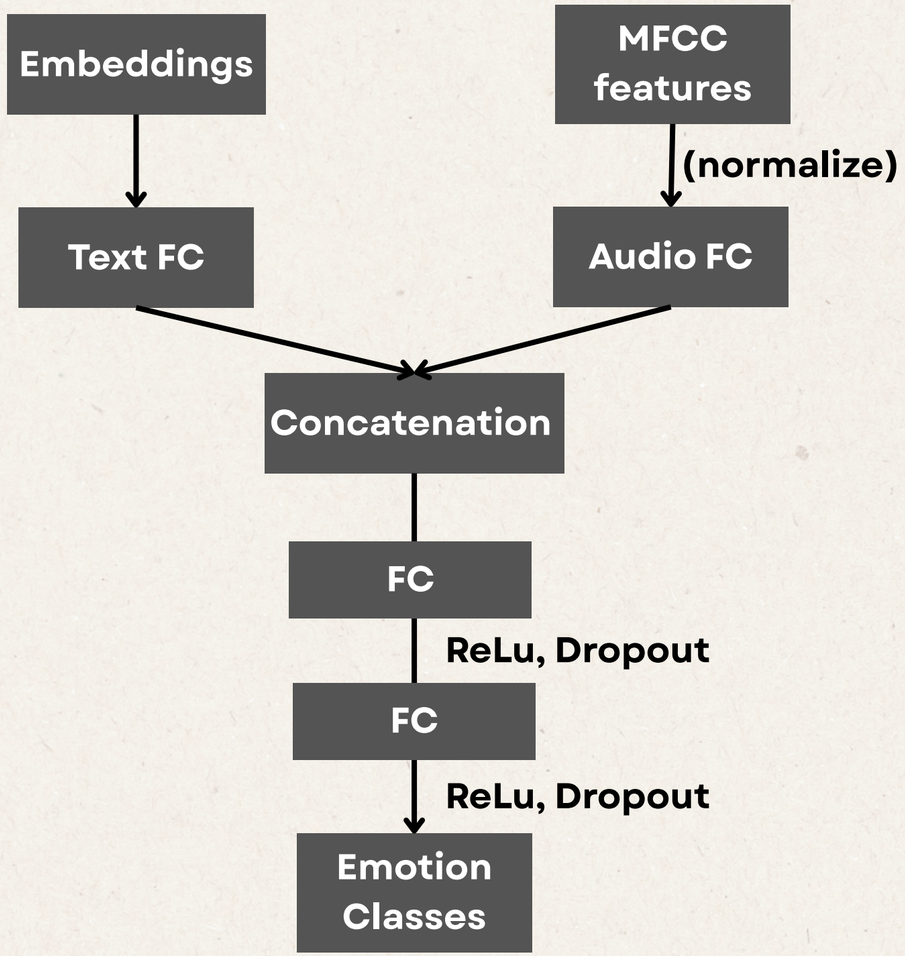
Audio Classifier

- Model: 3-layer MLP
  - 1st hidden layer: ReLu activation, 256 hidden units, dropout at 0.2
  - 2nd hidden layer: ReLu activation, 128 hidden units, dropout at 0.2
- Hyperparameters:
  - Loss Function: Cross Entropy Loss
  - Optimizer: Adam
  - Learning rate: 1e-4
  - Batch Size: 8
  - Epochs: 20

Implementation

Multimodal Classifier

- One fully connected layer for text
  - ReLu activation; 128 hidden units
- One fully connected layer for audio
  - ReLu activation; 128 hidden units
- After concatenating:
  - 1st FC: ReLu activation; 256 hidden units
  - 2nd FC: ReLu activation; 128 hidden units
- Hyperparameters:
  - Loss Function: Cross Entropy Loss
  - Optimizer: Adam
  - Learning rate: 1e-4
  - Batch Size: 8
  - Epochs: 20
  - Dropout: 0.2



Results & Experimental Evaluation

BERT Text Classifier

- Training over 10 epochs
  - loss: 3.79
  - Accuracy: 98%
- Strong ability to capture semantics from text
- But, lack access to vocal tone and prosody

```
Using device: cuda
Epoch 1: Loss = 24.29, Accuracy = 31.00%
Epoch 2: Loss = 19.67, Accuracy = 52.00%
Epoch 3: Loss = 17.48, Accuracy = 55.00%
Epoch 4: Loss = 14.40, Accuracy = 66.00%
Epoch 5: Loss = 12.06, Accuracy = 74.00%
Epoch 6: Loss = 9.83, Accuracy = 82.00%
Epoch 7: Loss = 7.77, Accuracy = 88.00%
Epoch 8: Loss = 6.04, Accuracy = 94.00%
Epoch 9: Loss = 4.79, Accuracy = 97.00%
Epoch 10: Loss = 3.79, Accuracy = 98.00%
```

Results & Experimental Evaluation

Audio-only Classifier

- Training over 20 epochs
  - Loss: 1.1309
  - Accuracy: 63%
- Test
  - Accuracy: 13%
- Bad generalization
  - overfitting ← dropout
  - limitation of using MFCC only

```
torch.Size([100, 13, 401])
Epoch 1: Loss = 9.0137, Accuracy = 10.00%
Epoch 2: Loss = 5.7171, Accuracy = 11.00%
Epoch 3: Loss = 4.2682, Accuracy = 32.00%
Epoch 4: Loss = 3.5020, Accuracy = 46.00%
Epoch 5: Loss = 4.1964, Accuracy = 41.00%
Epoch 6: Loss = 3.5363, Accuracy = 42.00%
Epoch 7: Loss = 2.7567, Accuracy = 48.00%
Epoch 8: Loss = 2.7225, Accuracy = 45.00%
Epoch 9: Loss = 2.1840, Accuracy = 48.00%
Epoch 10: Loss = 2.0531, Accuracy = 42.00%
Epoch 11: Loss = 1.8790, Accuracy = 46.00%
Epoch 12: Loss = 1.7845, Accuracy = 42.00%
Epoch 13: Loss = 1.5702, Accuracy = 55.00%
Epoch 14: Loss = 1.6945, Accuracy = 56.00%
Epoch 15: Loss = 1.3058, Accuracy = 56.00%
Epoch 16: Loss = 1.2443, Accuracy = 57.00%
Epoch 17: Loss = 1.2325, Accuracy = 60.00%
Epoch 18: Loss = 1.0825, Accuracy = 64.00%
Epoch 19: Loss = 1.1311, Accuracy = 63.00%
Epoch 20: Loss = 1.1309, Accuracy = 63.00%
```

Results & Experimental Evaluation

Multimodal Classifier - w/o normalization

```
Epoch 1: Loss = 1.9920, Accuracy = 19.00%
Epoch 2: Loss = 1.8690, Accuracy = 22.00%
Epoch 3: Loss = 1.8239, Accuracy = 34.00%
Epoch 4: Loss = 1.7019, Accuracy = 31.00%
Epoch 5: Loss = 1.7073, Accuracy = 31.00%
Epoch 6: Loss = 1.6298, Accuracy = 39.00%
Epoch 7: Loss = 1.7405, Accuracy = 39.00%
Epoch 8: Loss = 1.7225, Accuracy = 44.00%
Epoch 9: Loss = 1.5675, Accuracy = 50.00%
Epoch 10: Loss = 1.6316, Accuracy = 49.00%
Epoch 11: Loss = 1.6549, Accuracy = 49.00%
Epoch 12: Loss = 1.6426, Accuracy = 49.00%
Epoch 13: Loss = 1.6312, Accuracy = 49.00%
Epoch 14: Loss = 1.7825, Accuracy = 46.00%
Epoch 15: Loss = 1.6037, Accuracy = 50.00%
Epoch 16: Loss = 1.7242, Accuracy = 49.00%
Epoch 17: Loss = 1.5511, Accuracy = 50.00%
Epoch 18: Loss = 1.6078, Accuracy = 42.00%
Epoch 19: Loss = 1.6149, Accuracy = 48.00%
Epoch 20: Loss = 1.6569, Accuracy = 45.00%
```

Multimodal Classifier - w/ normalization on audio features

```
Epoch 1: Loss = 1.9930, Accuracy = 9.00%
Epoch 2: Loss = 1.9618, Accuracy = 9.00%
Epoch 3: Loss = 1.9332, Accuracy = 10.00%
Epoch 4: Loss = 1.9067, Accuracy = 10.00%
Epoch 5: Loss = 1.8815, Accuracy = 37.00%
Epoch 6: Loss = 1.8575, Accuracy = 50.00%
Epoch 7: Loss = 1.8349, Accuracy = 52.00%
Epoch 8: Loss = 1.8137, Accuracy = 53.00%
Epoch 9: Loss = 1.7929, Accuracy = 54.00%
Epoch 10: Loss = 1.7723, Accuracy = 53.00%
Epoch 11: Loss = 1.7516, Accuracy = 53.00%
Epoch 12: Loss = 1.7310, Accuracy = 53.00%
Epoch 13: Loss = 1.7102, Accuracy = 53.00%
Epoch 14: Loss = 1.6892, Accuracy = 53.00%
Epoch 15: Loss = 1.6682, Accuracy = 53.00%
Epoch 16: Loss = 1.6475, Accuracy = 52.00%
Epoch 17: Loss = 1.6271, Accuracy = 52.00%
Epoch 18: Loss = 1.6072, Accuracy = 52.00%
Epoch 19: Loss = 1.5878, Accuracy = 52.00%
Epoch 20: Loss = 1.5692, Accuracy = 52.00%
```

Conclusion

Key Findings:

- BERT Text-only Classifier could achieve a high accuracy, at 98%
  - Captured semantic nuances in text well
  - But, lack emotional tone or speech patterns.
- Audio-only classifier achieves 63% accuracy on training set, but 13% accuracy on test set.
  - Bad generalization → limitations of relying solely on MFCC features
- Our multimodal classifier
  - Improved overall performance from 45% to 52% by normalizing audio MFCC features
  - Demonstrated the potential of integrating text and audio

Future Works:

- Keep fine-tuning the multimodal classifier to improve its performance
- Extend to video-based facial emotion recognition
- Explore larger and more diverse datasets to check the generalization of the multimodal classifier

Thank you!