COM 6165 Topics In Software Engineering
Project Progress Report
Group: Annie Xu (jx2603) & Beijia Zhang (bz2527)

1. Overview

The topic that our group hopes to explore for the final project is multimodal systems. These systems gather information from various modalities such as text, audio, and visual input, and combine them to enhance task performance in applications like emotion recognition, video understanding, and autonomous systems. However, many systems treat all inputs the same way, which causes performance to drop when one or more inputs are unreliable. For example, poor lighting can impair visual data, while background noise can degrade audio signals. While researching different fusion models, we discovered that many traditional approaches assume all modalities are equally reliable. This implementation falls short in real-world settings since input quality may vary due to noise, occlusion, or even sensor failure.

Our project will introduce an innovative extension to existing baseline multimodal fusion systems by integrating a Dynamic Modality Weighting Module. This module evaluates the real-time quality of each modality and adapts the fusion process accordingly. By implementing this, we hope to improve system adaptability, resilience and effectiveness in unpredictable or noisy environments.

2. Research Questions

1. How does the performance of a baseline multimodal fusion model change under modality degradation (e.g., noisy audio, blurred visuals)?
2. Does integrating dynamically computed modality weights improve classification performance compared to using static weights?
3. What is the tradeoff between computational complexity and performance gains introduced by the quality-aware fusion mechanism?
4. Which modality contributes most to emotion recognition performance under different quality conditions?
5. How does modality reliability vary across different types of emotions? Are certain emotions (e.g., fear, joy) more easily recognized through specific modalities?

3. Value to User Community

This project can be cratered to a wide range of users including AI researchers, software developers working on multimodal systems and industry professionals building applications that

rely on integrating multiple sensory inputs. The below will list some specific user cases for example:

- Emotion Recognition in Human-Computer Interaction: Virtual assistants, customer service bots or mental health monitoring tools can more accurately interpret user emotions even if one signal (ex. facial video) is missing or degraded.
- Autonomous Vehicles: In poor weather or low-light conditions, visual input may be unreliable, so the vehicle can dynamically rely more on radar or LiDAR to make driving decisions safely.
- Video Surveillance and Security: In crowded or noisy environments, or during visual obstructions, the system can still reliably detect anomalies or threats by adjusting its reliance on different inputs.

4. Demo

Elevator Pitch:
Hello everyone, our project explores multimodal systems which is basically a technology that combines input from text, audio, and visual data to improve task performance in applications like emotion recognition and autonomous vehicles. While existing systems often treat all inputs equally, that assumption breaks down in real-world scenarios—think of blurred video during a Zoom call or noisy background audio. Our solution is a Dynamic Modality Weighting Module that evaluates the real-time quality of each input and adjusts its influence accordingly. This adaptive fusion approach makes the system more robust, reliable, and better suited for unpredictable environments. Whether it's a mental health tool interpreting emotions or a self-driving car navigating in fog, we aim to bring smarter decision-making to multimodal AI.

Presentation:
For our 5-minute demo, we'll compare a baseline multimodal fusion model with our enhanced version that includes a Dynamic Modality Weighting Module. We'll start with a brief overview of traditional fusion methods, then show side-by-side results under degraded input conditions like noisy audio or blurred video. Our demo will highlight how the system dynamically adjusts modality weights, along with visualizations and performance metrics demonstrating improved accuracy and robustness.

5. Delivery

We will deliver all components of the project via a GitHub Repository, which will include the following:

- Source Code: All implementations files for the baseline model, dynamic modality weighting module, and adaptive fusion system.

- Training and Evaluation Scripts: Scripts for reprocessing, training, validation, testing under different input conditions, and visualization.
- Documentation: A README.md with setup instructions, usage examples, project structure and dependencies. Additional documentation will describe the architecture, quality scoring mechanism and fusion strategy.
- Final Report: The written report answering the research questions and summarizing the findings as well as the demo on it.
- Datasets and Models: We will not host large datasets or pretrained models directly in the repository. Instead we will link the dataset we use as well as mention the fine tuned models.

6. Others

During our exploration of available resources, we discovered a comprehensive collection of multimodal datasets at https://github.com/drmuskangarg/Multimodal-datasets. We selected the Multimodal EmotionLines Dataset (MELD) (https://github.com/declare-lab/MELD) for our project. MELD is well-suited for our research focus as it provides synchronized text, audio, and visual data for each utterance in multi-party dialogues, alongside labeled annotations for both emotion (seven categories) and sentiment (positive, neutral, negative). This setup makes it an ideal benchmark for testing our Dynamic Modality Weighting Module (DMWM) in the context of emotion recognition.

MELD's data originates from real conversations in the Friends TV series which naturally introduces variability across modalities. For example, audio signals may include overlapping dialogue or background music, video frames may suffer from suboptimal angles (e.g., side profiles), and text transcripts may not always convey non-verbal cues like tone. These modality-specific inconsistencies align closely with the types of reliability issues we aim to address through our DMWM.

- Text: Each utterance is transcribed and labeled with emotion and sentiment.
- Audio: Extracted from the Friends TV show, provided in .mp4 video clips and in audio feature pickle files (e.g., audio_embeddings_feature_selection_emotion.pkl).
- Visual: Available via video clips for each utterance, and visual embeddings can be extracted for analysis.