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

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.

Implementation Plan:
1. Dataset: Utilize an existing multi-model dataset containing text, audio, and visual modalities. (used to train the multi-modal fusion model)
2. Baseline Model: Use a pre-trained or existing baseline multi-modal fusion architecture.
3. Quality Estimation Module: Calculate the quality scores for each modality using self-defined metrics or neural models. The scores reflect how clean and reliable each module is for the given dataset.
4. Adaptive Fusion Layer: Modify our baseline model integrated with the quality scores as weights to fine-tuning the model. (more reliable module will contribute more)
5. Training: Train the model using loss function: cross-entropy loss for classification and regression loss for quality estimation.
6. Evaluation: Compare the extended model against the baseline one for clean input conditions, controlled degradation and real-world noisy test samples.

Evaluation Plan
1. For quantitative: We compare accuracy, precision, recall, and F1 scores, and analyze robustness under degraded input conditions.
2. For qualitative: We visualize the modality weights and show real-time modality adoption in sample inputs.

Papers to Read (for now)

- https://www.sciencedirect.com/science/article/pii/S1566253522001634?via%3Dihub
- https://openaccess.thecvf.com/content/CVPR2023/html/Li_Efficient_Multimodal_Fusion_via_Interactive_Prompting_CVPR_2023_paper.html
- https://www.sciencedirect.com/science/article/pii/S1566253523002609
- https://www.sciencedirect.com/science/article/pii/S0925231223005507
- https://www.mdpi.com/1424-8220/23/5/2381
- https://dl.acm.org/doi/full/10.1145/3649447
- https://link.springer.com/article/10.1007/s00521-022-06913-2

Conclusion

This project improves the baseline model with constant weight for each of the modality. The weights are re-computed for each new input making the system adaptive and dynamic. It enhances both robustness and performance in realistic conditions. This experimental work will hope to contribute to the development of more adaptive and intelligent AI systems for multi-modal understanding tasks.