DSCI 565 Project Proposal

Group Member: Colin(Zixi) Wang, Nealson Setiawan, Chris Chen

**Project Objective:**Develop a robust multimodal model capable of detecting emotions from multiple data types (images, text, audio) to enhance social media content recommendation systems.

**Goal 1:**Design and implement a multimodal model that effectively classifies human emotions with high performance, evaluated using key metrics such as Accuracy, F1-score, and Area Under the Curve (AUC).

**Goal 2:**Investigate the impact of different multimodal model architectures, including early fusion, late fusion, and autoencoder-based approaches, on classification performance. Compare the outcomes of each design.

**Goal 3:**Evaluate the model's performance across various social media platforms and content contexts to ensure adaptability and reliability in real-world scenarios.

### **Current State of the Art**

The current SOTA model uses Transformers with a hierarchical gated fusion layer to combine text, audio, and visual data. Self-distillation provides output-based knowledge transfer, refining the model’s multimodal understanding.

The IEMOCAP dataset, with over 12 hours of video, audio, and textual transcriptions, is used to train models to recognize emotions. Compute-intensive training typically requires GPU clusters for handling large models and diverse data types.

**Reference Paper**: [A Transformer-Based Model With Self-Distillation for Multimodal Emotion Recognition in Conversations | Papers With Code](https://paperswithcode.com/paper/a-transformer-based-model-with-self)

### **Proposed Approach**

Our first step would be to replicate the original Transformer-based model with self-distillation and measure initial performance metrics. After that, we plan to make some enhancements to the model to boost its performance. Some directions we will try out include:

1. **Fusion Mechanism**: Current approaches mostly utilize early fusion or late fusion, while intermediate fusion techniques are rarely explored. One of our approaches would be to experiment with attention-based and intermediate fusion techniques for improved cross-modal interactions.
2. **Enhanced Self-Distillation**: The state of the art model implemented a self-distillation technique. While the model achieved decent performance, it is still worthwhile to explore different feature-based self-distillation by extracting and transferring intermediate representations within the Transformer layers.
3. **Pre-trained Modal Encoders**: Another direction for improvement would be to replace modality-specific encoders with pre-trained models like BERT and Wav2Vec for more robust feature extraction.

**Benchmark datasets used for evaluation**

To follow current state of the art benchmarks and add credibility to our project, we will be evaluating our model on the IMECOAP dataset. It is a large dataset of 17GB, and full of multimodal emotion dataset of images. Spoken by 10 speakers back to back, and recorded in USC at 2008.

The dataset consists of 302 speech, and videos of 2 speakers talking back to back to each other. We will use whisper or a transcription model to generate text for textual analysis. The classification labels in this dataset which are annotated manually includes: happy sad neutral angry excited frustrated. The dataset is recorded across 5 sessions with 5 pairs of speakers.

We have applied and got permission from the authors to use this dataset.

Links:

* <https://paperswithcode.com/dataset/iemocap>
* <https://sail.usc.edu/iemocap/>

**Training datasets**

To further increase generalizability in our model, we'll use other datasets for training as well, including pure speech, pure text, or pure image.

Current list of training datasets (but not limited to):

1. **FER+ (Face Expression Recognition Plus dataset):** <https://paperswithcode.com/dataset/fer>
2. AffectNet: <https://paperswithcode.com/dataset/affectnet>
3. **NRC Word-Emotion Association Lexicon:** <https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm>
4. [**CARER: Contextualized Affect Representations for Emotion Recognition**](https://aclanthology.org/D18-1404.pdf)
5. <https://huggingface.co/datasets/dair-ai/emotion>

**Risk:**

1. Model limitations: We will found out once implemented our model
2. Dataset limitations: Training dataset and testing dataset embedded various context/background noise might induced some bias
3. Computational resource limitations: GPU limitations on Colab

**Task Division**

* **Colin**: Responsible for building the model that detects facial expressions (Images)
* **Nealson**: Responsible for building the model that detects speech (Audio)
* **Chris**: Responsible for building the model that detects writing contents (Text)
* Collaborate on fine-tuning, validating, and testing the final multimodal model to ensure optimal performance and integration of all data types (images, audio, text).