COSE474-2024F: Final Project Proposal
“Multimodal Emotion Recognition Using CLIP”

Aliff Khuzairi 2021320101

1. Introduction

Emotion recognition is a key task in understanding human behaviour and improving human-computer interaction. While text and image modalities each contribute to emotion detection, combining both can yield richer, more accurate predictions. CLIP’s ability to jointly understand text and images makes it an ideal candidate for multi-modal emotion recognition, enhancing emotion classification in contexts such as social media analysis, video content, and human-robot interaction.

2. Problem definition & challenges

The goal of this project is to see how well CLIP can be fine-tuned for emotion recognition when using both text and image inputs. We will train the model on a dataset that contains both images and text and compare how well it performs against other advanced models. Identifying the optimal way to integrate emotional cues from graphics and text is one of the challenges. The two may occasionally send out different emotional cues, thus the model needs to coordinate them carefully. Optimizing CLIP especially for emotion recognition while maintaining its generalizability across many text and image types presents another problem.

3. Related Works

The majority of previous studies have concentrated on either text-based emotion identification (sentiment analysis) or image-based emotion recognition (facial expressions, situations, etc.) For example, *Radford et al. (2021)* presented **CLIP**, which uses a contrastive method to learn a joint representation of text and images. This model provides a strong basis for emotion detection tasks and has demonstrated encouraging performance in a variety of tasks, such as multimodal categorization and zero-shot learning (**?**).

4. Datasets

The Emotion Dataset from DAIR-AI, which consists of a variety of textual material labelled with various emotions, will be used for Emotion Recognition Tasks. This dataset can be used to train our model to identify emotions from written input since it contains a variety of text samples that have been grouped into several emotional labels. For the imaging modality, the FER-2013 (Facial Expression Recognition 2013) dataset will be used. This dataset includes gray-scale portraits of faces labelled with various emotions, including surprise, rage, contempt, fear, happiness, sadness, and neutral. The photographs will offer the required visual input to support the written material because they are well-annotated.

5. Schedule & Roles

This is an individual project, and I will be responsible for data preprocessing, model training, evaluation, and reporting.

- Weeks 1-2: Initial Setup and Data Preparation
- Weeks 3-5: Model Development and Training
- Weeks 6-7: Results and Report Writing
- Week 8: Finalization and Submission

6. Comparison with SOTA (State-of-the-Art)

- **Text-Only Model:** A simple LSTM or GRU model will be used to classify emotions based solely on the textual data from the Emotion Dataset. This model will provide a baseline for evaluating the performance of text processing without visual context.
- **Image-Only Model:** A Convolutional Neural Network (CNN) will be implemented to recognize emotions based only on the FER-2013 images. This model will serve as a comparison point for visual emotion recognition without textual input.