COMP8420 Major Project Report

# Project Title: Emotion Feedback Generator

A Visual-Language Hybrid System for Emotion-Aware Text Generation

# Team Members

Md Raihanul Islam Bhuiyan  
Student ID: 48401706

Rakibul Alam Nahin  
Student ID: 48442119

# Introduction: Scope and Motivation

Emotional expression is a vital component of human communication, yet it's often lost in virtual interactions like video meetings, digital therapy, or e-learning environments. This project proposes a system that combines computer vision and natural language processing to identify emotions in facial images and generate supportive, empathetic feedback. The aim is to improve communication quality and user engagement by providing context-aware, emotion-aligned responses. Such technology can be useful in applications ranging from mental health support to AI-based tutors or communication aids.

# Project Objectives

* Create custom dataset of 3 emotions (Confident, Scared/Nervous, Confused)
* Data analysis.
  + Image: Sample image for analysis
  + Text: Sample text for analysis. Remove punctuation and stop words.
  + Data augmentation and transformation.
* Using a hybrid approach to make a custom VLM system. Fusion of CLIP vision encoder and Llama model.
* Using the model to generate emotional meaningful feedback by taking emotional images as input.
* Explore the feasibility of real-time emotion-response interaction tools.

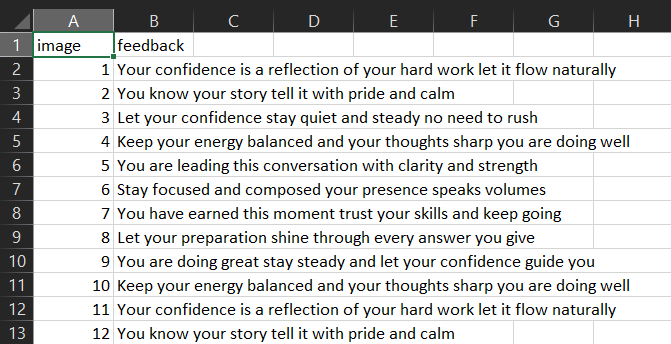
# Dataset Description

We created a custom dataset by scraping images from public sources and categorizing them into three emotional classes: 'Confident', 'Confused', and 'Scared'. Each image was paired with a feedback sentence intended to reflect a helpful or empathetic response. The dataset consists of three CSV files, one per emotion class, with each entry containing the path to the image and the associated textual response. Approximately 150 entries were compiled (50 per class), ensuring diversity across age, gender, and background. Also, due to lack of data we upscaled our data using repeated DataLoader to make the make training data and as the data is transformed continuously it might help with creating different images.

# Sample Dataset Images

Below are some examples from the curated dataset, representing the three emotion classes: Confident, Confused, and Scared.  
  
• Each image is paired with an emotion label and a handcrafted feedback response.  
• These examples illustrate the diversity in visual features and emotional cues.

|  |  |  |
| --- | --- | --- |
| Label: Scared/Nervous | Label: Confident | A person with glasses and a beard  AI-generated content may be incorrect.  Label: Confused |



Label: confident\_feedback.csv sample

# Data Pre-processing

**Image preprocessing**

Images collected first went through various augmentation techniques. With various experiments we found optimum augmentations and then image tensors were formed – note normalization was not done due to it is done in the next layers:

1. RandomResizedCrop(224)
2. RandomHorizontalFlip()
3. ColorJitter(0.2, 0.2, 0.5, 0.2, 0.1)
4. RandomGrayScale(p=0.5)

**Text preprocessing**

While collecting text data texts were cleaned by removing punctuation and stop-words. Text collected with were merged with the emotion and the feedback to form one prompt which is used in the later layers.

# Methodology

**Model architecture**

A diagram of a computer program

AI-generated content may be incorrect.

Firstly, using Llama-7b-chat-hf for our LLM model and tokenizer to tokenize the text into a token embedding. Next using pre-trained CLIP Image encoder and processor from, openai/clip-vit-base-path32, to encode the images to make image embeddings. Then concat the image embeddings with the text embeddings. Doing the same with attention mask. The labels are adjusted the same way that the labels are shifted to the right. This creates a so each token is predicted from the previous one — this is **teacher forcing** in CLM. Thus creating the latent space of embeddings this embeddings are used to train Llama model with 4-bit LoRA quantization so that model is trained easily with minimal hardware requirements.

# Experimental Results and Evaluation

Result:

|  |  |
| --- | --- |
| Image | Generated Feedback |
| A collage of a person smiling  AI-generated content may be incorrect. | 🧠 Generated Feedback:  You look confident. you're doing great, remember to take it one step at a time. you’ve got this, at university of missouri health. |
|  | 🧠 Generated Feedback:  scared, you will be proud of yourself, but you are a good at this scary environment. nobody is perfect quote. |
|  | 🧠 Generated Feedback:  You look scared, it ok to be confused stuck pause for a moment, and take a break. |

Sample generation of the Feedback

**BERTScore:**

Average Precison: 0.8734 | Average Recall: 0.70465 | Average F1-Score: 0. 8881

We qualitatively evaluated generated responses for short text and using BERTScore between original text and the generated text to measure the performance of the model.  
From the above images and model generations it can be said that although the text generations are not semantically sensible but it does give out a sense of generations with respect to the image given.

# Challenges and Limitations

• Dataset Creation: Gathering and annotating emotion-based images with high-quality feedback was time-intensive.  
• Resource Constraints: Limited compute resources made fine-tuning infeasible.  
• Feedback Diversity: Crafting emotionally appropriate feedback texts manually is slow and not scalable.

# Future Work

• Increase dataset size and demographic diversity for better generalization.  
• Incorporate automated response scoring using emotion classifiers.  
• Enable speech and video integration for real-time emotion feedback.  
• Add more team members or crowdsource annotation tasks to scale feedback writing.

# Conclusion

Our hybrid system demonstrates that combining CLIP and LLaMA can effectively generate empathetic feedback based on facial expressions. The results show promise in improving digital communication through emotionally-aware AI. Although there are challenges in data collection and evaluation, this project lays the groundwork for further development of emotion-enhanced human-computer interaction tools.

# Contribution Breakdown

[Md Raihanul Islam Bhuiyan]:  
Image collection and emotion tagging.  
Writing feedback responses.  
Code review and correction.

Ideation

Report writing

[Rakibul Alam Nahin]:  
Ideation

Image collection and emotion tagging.  
Writing feedback responses.  
Model creation and evaluation

Report writing

Github Repository:

https://github.com/rakibulnahin/nlp\_project.git