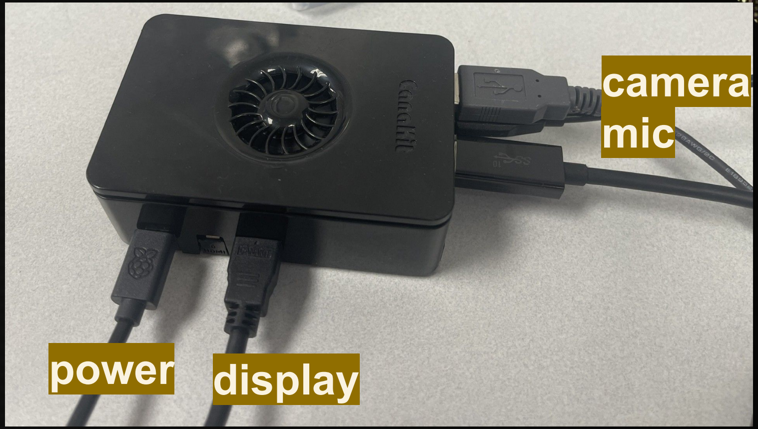
Vibe Check: Multimodal Emotion Recognition at the Edge

Regan Willis, Haley Lind, Josh Moorehead

CSCE790-007 Spring 2025

Emotion Recognition – Motivation

- Potential use cases: improved health care, awareness of customer opinions, and gauging political opinions
- **Verbal + non-verbal cues give a complete picture of a person's current emotion**
- In privacy-sensitive applications emotions should be predicted at the edge
- Emotions can shift and change rapidly so they must be predicted in a timely manner



Expression Recognition

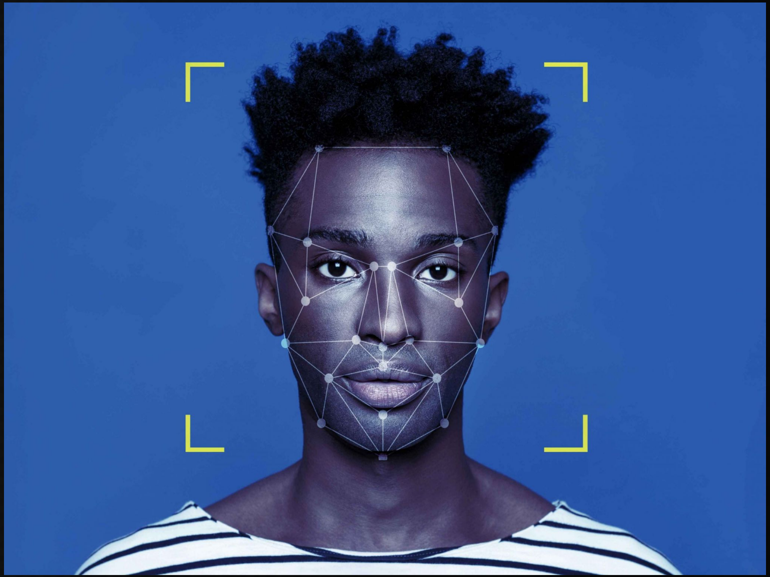
What is FER?

Goal: Detect and classify **emotions** from human faces.

Emotions: Anger, Disgust, Fear, Happiness, Neutral, Sadness, Surprise.

Applications:

- Human-Computer Interaction.
- Mental Health Monitoring.
- Sentiment Analysis in Social Media.



Source: *The Problem with Emotion Detection Technology*, Charlotte Gifford, The New Economy, June 15, 2020. Link

FER2013 Dataset Overview

- **Purpose**: Benchmark dataset for Facial Expression Recognition (FER).
- **Size**: 35,887 grayscale images (48x48 resolution).
- **Emotions**: Angry, Disgust, Fear, Happy, Neutral, Sad, Surprise.
- **Split**: 28,709 training, 3,589 validation, 3,589 test images.
- **Challenges**:
  - Low resolution and real-world variability.
  - Class imbalance (e.g., few Disgust samples).
  - Noisy labels and diverse facial angles.

Source: Kaggle Notebook – Face Emotion Detection

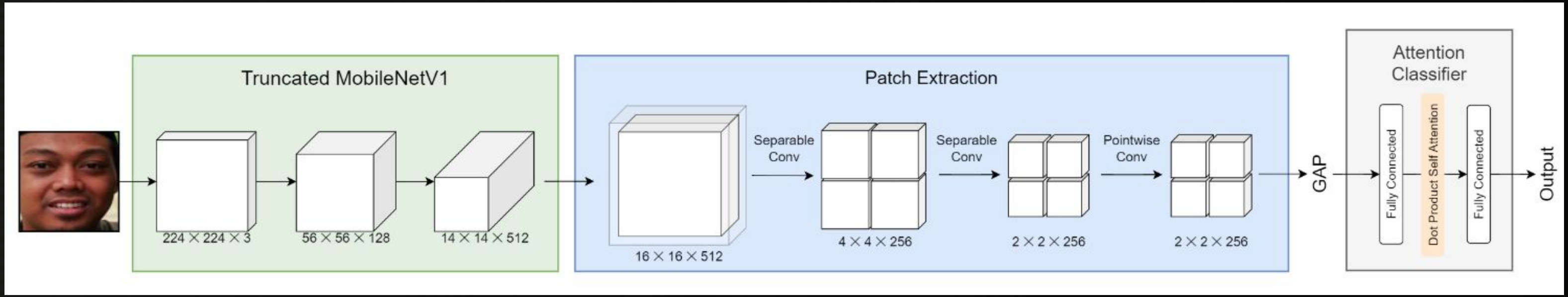
Patt-Lite Overview

Lightweight FER model for real-time edge deployment.

Combines:

- **Truncated MobileNetV1 CNN** for low-complexity global features.
- **Patch Extraction Block** for robust local feature focus.
- **Self-Attention** for enhanced classification from minimal data.

Efficient: Only 1.1M parameters vs. 40M+ in other models.



Source: Ngwe, J. L., et al. "PAtt-Lite: Lightweight Patch and Attention MobileNet for Challenging Facial Expression Recognition," IEEE Access, 2024.

Patt-Lite Results

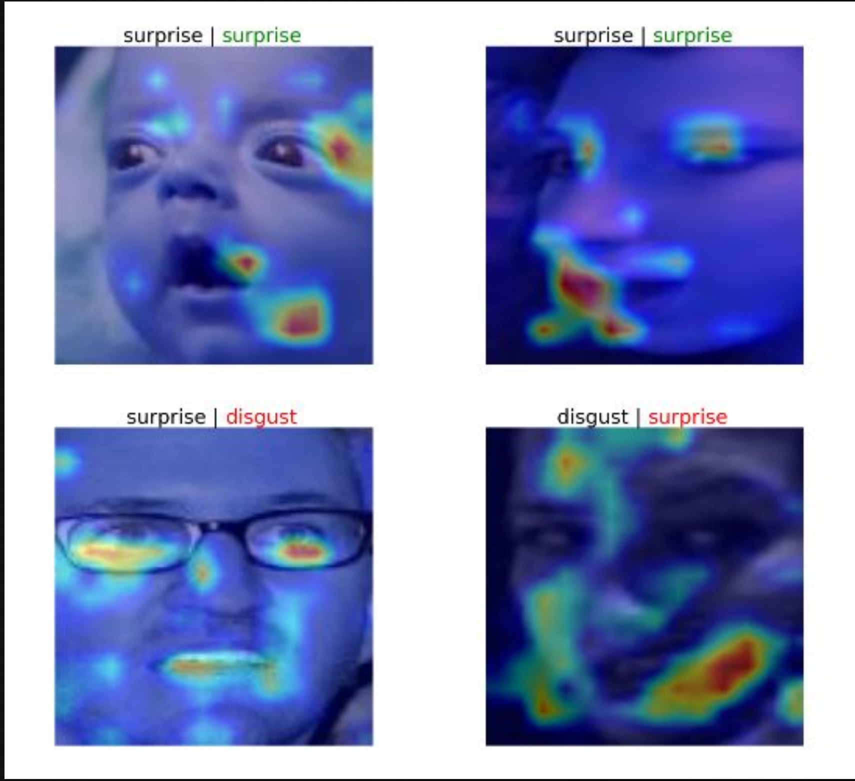
Outperforms state-of-the-art on:

- **RAF-DB:** 95.05%
- **FER2013:** 92.5%
- **FERPlus:** 95.5%

Handles real-world challenges:

- Occluded faces
- Varied lighting/angles
- Class imbalance (rare emotions)

Edge Ready: Runs on constrained devices with high accuracy.



Source: Ngwe, J. L., et al. "PAtt-Lite: Lightweight Patch and Attention MobileNet for Challenging Facial Expression Recognition," IEEE Access, 2024.

Our FER Model Results

Differences from Original Model:

- **Attention Mechanism Removed** → Simplified architecture, but maintained similar performance.
- Kept **MobileNet backbone** and **patch-based feature extraction** for lightweight inference.
- Designed for **Edge Deployment** (e.g., Raspberry Pi) with **minimal resource usage**.

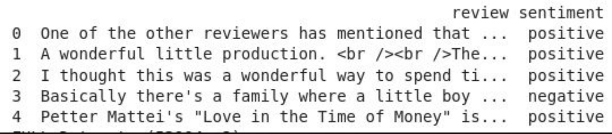
Performance Comparison:

- **Accuracy**: ~60% (with or without attention).
- **Reason for Similar Accuracy**:
  - The attention layer didn't significantly boost performance, suggesting the **core feature extraction** handled most of the learning.
  - Model benefits more from **pretrained MobileNet** and **data augmentation** than additional complexity.

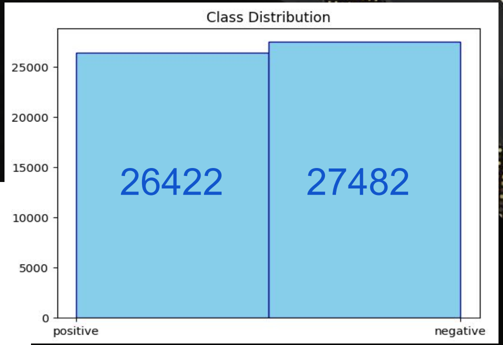
Sentiment Analysis

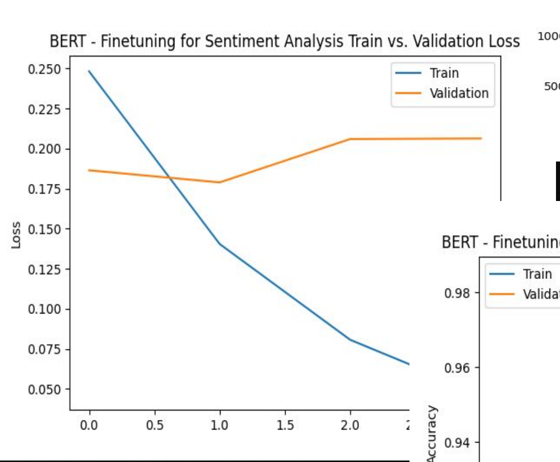
Sentiment Analysis

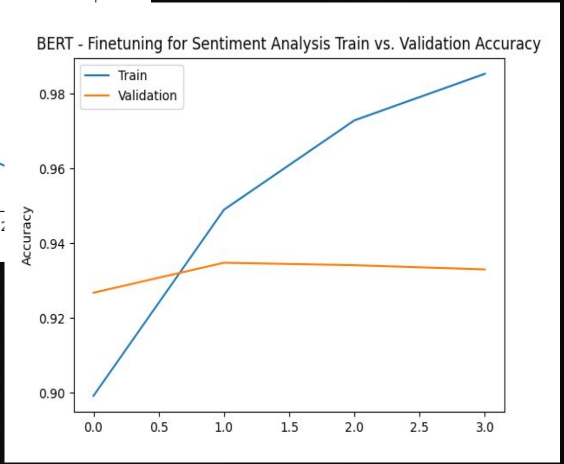
- Model architecture: BERT
- Datasets:
  - Sentiment Analysis Datasets [3]:
    - 2014 Twitter Data,
    - Archeage (MMORPG) reviews,
    - Ntua
  - IMDB Dataset [4]
- Our average accuracy on test dataset: **93.01%**
- Average inference time on Raspberry Pi 5: **~313 ms**



	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive

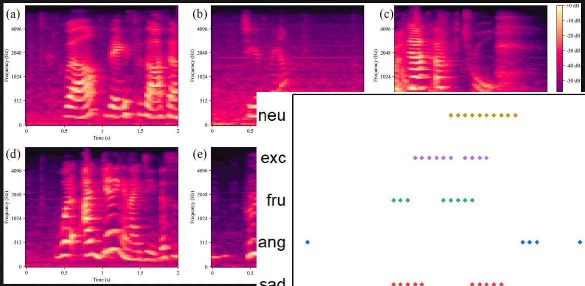


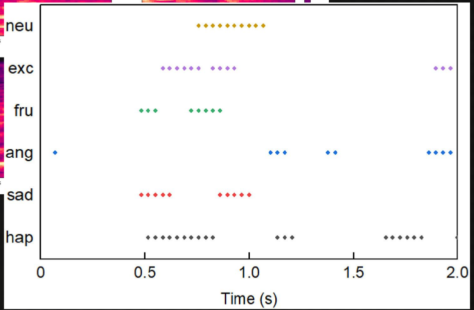




Speech Emotion Recognition based on Spiking Neural Network and Convolutional Neural Network (2025) [2]

- Text and images alone may not have enough information to convey emotion at a high accuracy
- Claim: temporal information matters in Speech Emotion Recognition (SER)
- Dataset: IEMOCAP - information about the speech signals, facial expressions, and hand movements of ten actors
- Accuracy of **65.3%**, beating current SOTA SER methods





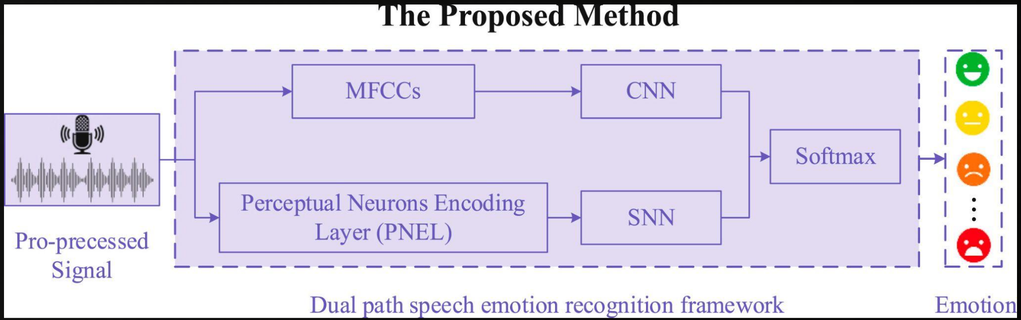


Figure 1 (left), 2 (middle), 3 (right) from [3] show the output from MFCCs, the output from PNEL, and the proposed framework, respectively.

Multimodal Data Fusion

A Short Survey on Multimodal Data Fusion in Image Classification [6]

Paper:

- As many classification tasks require multiple streams of data, there has been a rise in the need for multimodal fusion.
  - Featured-based
  - Intermediate-level
  - Decision-level

Relevance:

- Image classification + Text applicable to emotion recognition

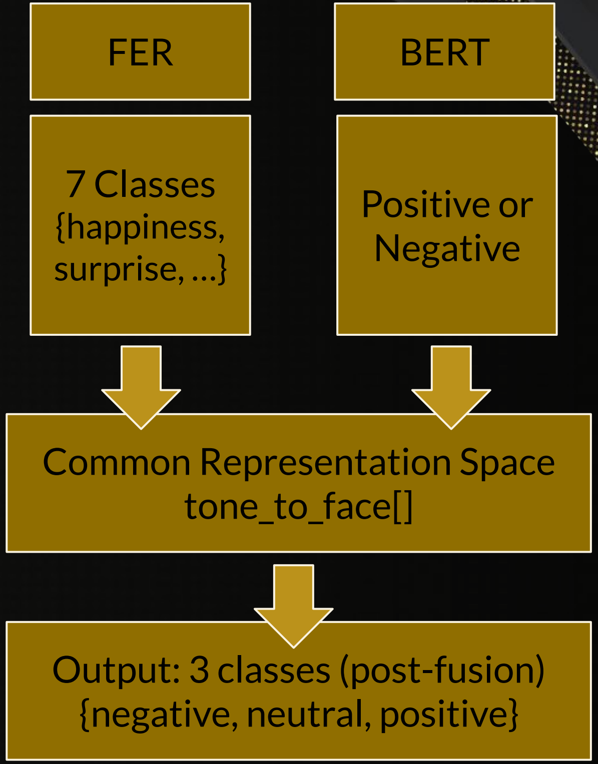
"The significance of multimodal fusion lies in its ability to address the shortcomings of unimodal approaches, leading to improved performance, reliability, and adaptability" [6].

Ref	*Technique*	*Accuracy*	*Advantages*	*Disadvantages*
[9]	Feature fusion using Histogram of Oriented Gradient + Local Phase Quantization	97,15%	- Best performance metrics	- Complexity and execution time
[10]	Fuse both the chest X-ray and cough (audio) model + CNN	98.91%	-Early diagnosis, non-invasive, fast prediction	- Need devices for the early diagnosis of non-communicable diseases in rural and remote areas.
[11]	early data fusion + late decision fusion SVM, Decision tree, KNN, MLP, RF, XGBoost	89.15%	- Long term prediction - Low cost implementation	- Model complexity
[12]	intermediate fusion + Self attention	99.78%	- High performance metrics	- Model not generalized - Small dataset
[13]	Coupled Adversarial Feature Learning (CAFL) Sub-network. - Supervised Multi-Level Feature Fusion Classification	99%	- Preservation of Detail information - Adaptive Probability Fusion - higher score classification	- Computational Complexity - Sensitivity to Hyperparameters
[14]	Combining TextCNN , ResNet50 with weight adaptive decision level fusion model	87.6%	- Applicability to Multimodal Environments - Improved Classification Accuracy	- Data Dependency - Sensitivity to Noise
[23]	Late fusion + intermediate fusion + deep learning	93.15%	- Improved diagnosis accuracy - Adaptive Batch Size	- Complexity and Resource Requirements - Optimal fusion strategy

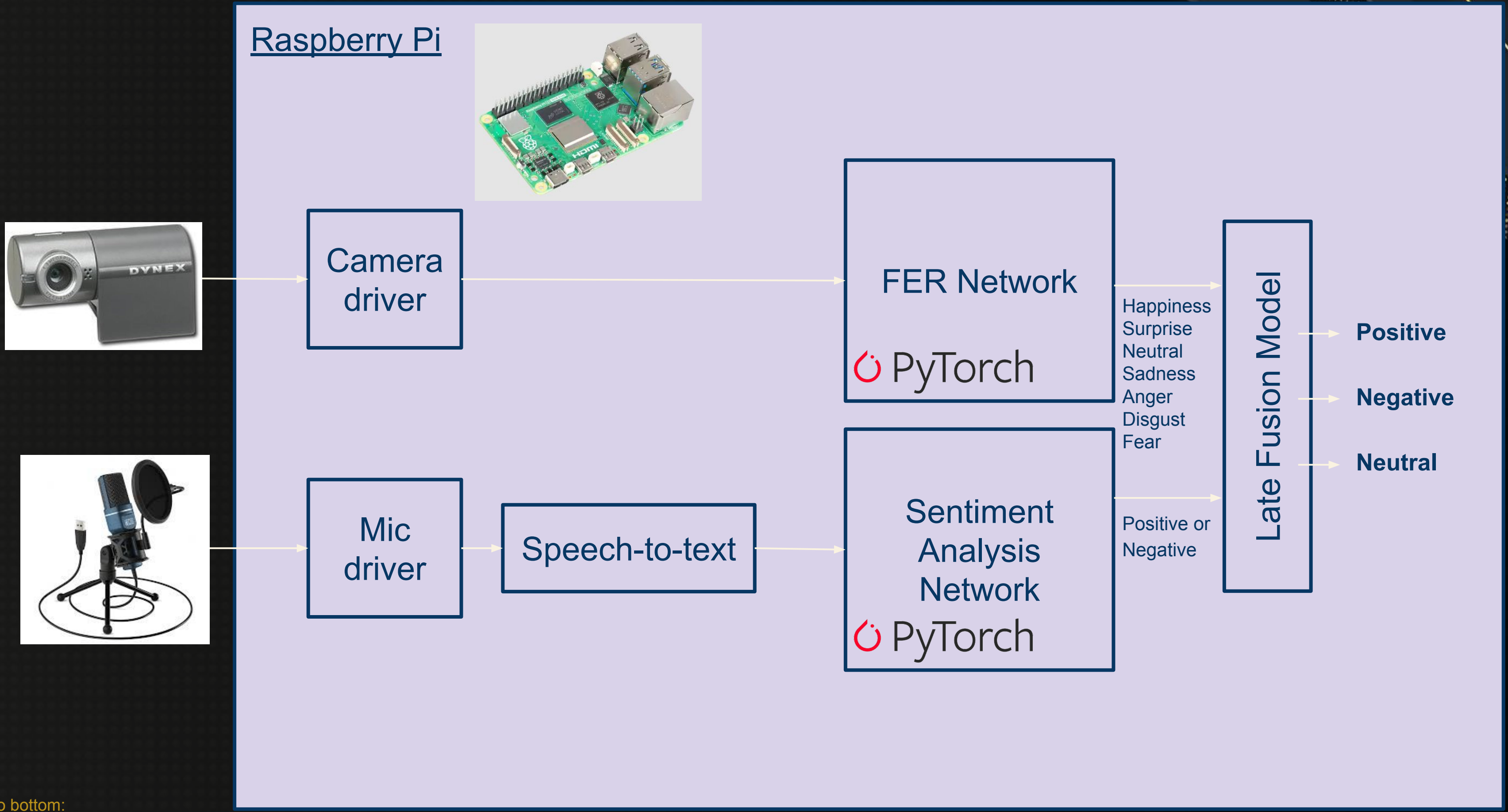
Figure: Comparative analysis of models from [6].

Inside Late Fusion

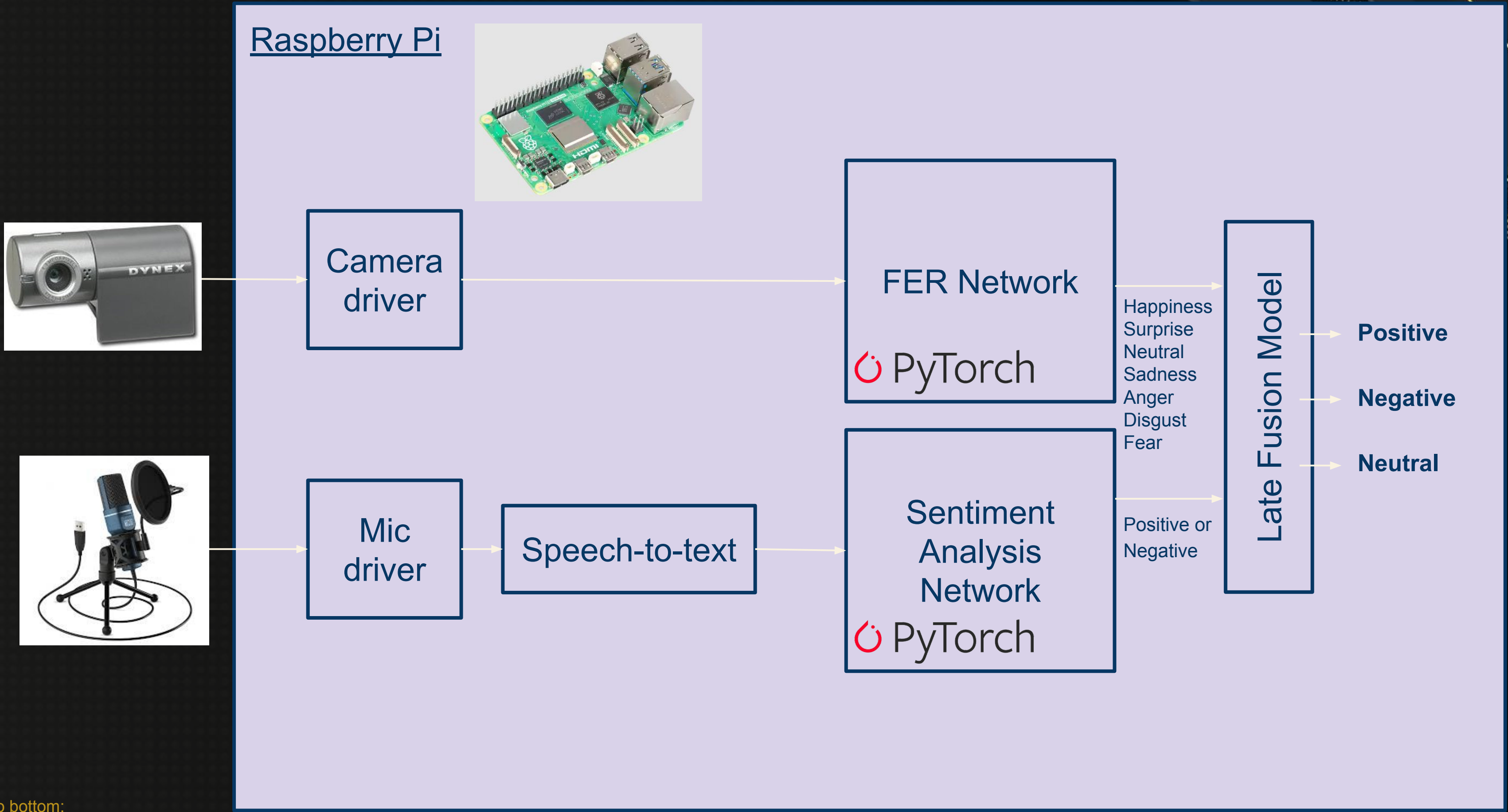
```
# Pseudocode

Initialize model:
   fc = Linear(9 → 3)
   softmax(dim=1)

Forward(bert_pred [1x2], fer_pred [1x7]):
   sentiment_class = argmax(bert_pred)
   class_weights = tone_to_face[sentiment_class]

   Weighted_FER = []
   for each class in fer_classes:
      Weighted_FER.append(fer_pred[class] * class_weights[class])
   fer_tensor = tensor(Weighted_FER)

   input = concat(bert_pred, fer_tensor)  # shape [ 1 x 9 ]
   output = softmax(fc(input))            # shape [ 1 x 3 ]  (-1, 0, 1)

   return output
```



Demo



Images from top to bottom:
https://www.raspberrypi.com/products/raspberry-pi-5/ ;
https://www.amazon.com/Dynex-DX-WEB1C-1-3MP-Webcam/dp/B001AO2Q5W ;
https://www.tonormic.com/products/tonor-tc-777-usb-microphone



Images from top to bottom:
https://www.raspberrypi.com/products/raspberry-pi-5/ ;
https://www.amazon.com/Dynex-DX-WEB1C-1-3MP-Webcam/dp/B001AO2Q5W ;
https://www.tonormic.com/products/tonor-tc-777-usb-microphone

Sentiment Analysis Output

Input	Prediction
the weather is beautiful today	positive
i’m so disappointed	negative
i love you	positive
this is the worst	negative
great! this is just what i needed today	positive
it’s raining cats and dogs	positive

Results

```
readying recording devices..
Capture Completed.

Analyzing image: ./tmp/vid/0_1713952790.123456.png
Analyzing text: ./tmp/1713952790.123456.txt

=== Facial Expression Analysis ===
Detected emotion: happiness
Confidence: 0.75
Inference time: 1.5s

=== Sentiment Analysis ===
Text content: "this is the worst"
Detected sentiment: negative (-1)
Inference time: 0.3s

=== Running Multimodal Fusion ===

=== Final Multimodal Result ===
Combined sentiment: neutral
Sentiment value: 0 (-1=negative, 0=neutral, 1=positive)
Confidence: 0.60
Inference time: 0.2s
```

Results

```
readying recording devices..
Capture Completed.

Analyzing image: ./tmp/vid/0_1713952790.123456.png
Analyzing text: ./tmp/1713952790.123456.txt

=== Facial Expression Analysis ===
Detected emotion: surprise
Confidence: 0.85
Inference time: 1.78s

=== Sentiment Analysis ===
Text content: "the weather is beautiful"
Detected sentiment: positive (1)
Inference time: 0.33s

=== Running Multimodal Fusion ===

=== Final Multimodal Result ===
Combined sentiment: positive
Sentiment value: 1 (-1=negative, 0=neutral, 1=positive)
Confidence: 0.65
Inference time: 0.25s
```

Conclusion

Key Takeaways:

- Multimodal emotion recognition improves accuracy over unimodal methods.
- Edge deployment is feasible with lightweight FER models and optimized sentiment analysis.
- Fusion of visual and textual cues provides a more complete emotional context.

Future Work:

- Improve FER model accuracy and enable real-time analysis of multiple frames.
- Incorporate speech pattern analysis (pitch, loudness, pauses) for richer multimodal input.
- Explore fusion at intermediate model layers for tighter integration.
- Train an end-to-end multimodal fusion model for a stricter and adaptive emotion prediction.

References

[1] Ngwe, J. L., Lim, K. M., Lee, C. P., Ong, T. S., & Alqahtani, A. (2024). *PAtt-Lite: Lightweight Patch and Attention MobileNet for Challenging Facial Expression Recognition*. IEEE Access, 12, 79327–79341. https://doi.org/10.1109/ACCESS.2024.3407108

[2] Singh, Upendra and Abhishek, Kumar and Azad, Hiteshwar Kumar. A Survey of Cutting-edge Multimodal Sentiment Analysis. September 2024. Association for Computing Machinery, vol. 56, no.9. https://doi.org/10.1145/3652149

[3] Chengyan Du, Fu Liu, Bing Kang, Tao Hou. Speech emotion recognition based on spiking neural network and convolutional neural network, Engineering Applications of Artificial Intelligence, Volume 147, 2025, https://doi.org/10.1016/j.engappai.2025.110314.

[4] Bashiri, H., Naderi, H. Comprehensive review and comparative analysis of transformer models in sentiment analysis. Knowl Inf Syst 66, 7305–7361 (2024). https://doi.org/10.1007/s10115-024-02214-3

[5] Maas, A., Large Movie Review Dataset. http://ai.stanford.edu/~amaas/data/sentiment/

[6] T. Datsi, K. Aznag, B. A. BenAli, K. Karbout, A. El Oirrak and E. K. Khayya, A Short Survey on Multimodal Data Fusion in Image Classification, 2024 International Conference on Global Aeronautical Engineering and Satellite Technology (GAST), Marrakesh, Morocco, 2024

REGAN, HALEY, JOSH

Milestones

Preliminary Research
March 3

- Literature review to find a model architecture and dataset for FER
- Literature review to find a model architecture and dataset for sentiment analysis
- **Literature review on data fusion with neural networks**

Set Up
March 17

- Write FER model training script and begin training
- **Write sentiment analysis model training script and begin training**
- Preprocess / clean FER dataset
- **Preprocess / clean sentiment analysis dataset**
- **Research freezing layers in CNNs and Transformers**

Initial Results
March 31

- **Write sensor driver / infrastructure scripts**
- Write FER model inference script and test model
- **Write sentiment analysis model inference script and test model**
- **Write data fusion script and test prediction**

Iterate
April 14

- Test system performance on edge device
- Improve prediction accuracy

Final Testing
April 28

- Re-test system performance

Vibe Check: Multimodal Emotion Recognition at the Edge

Regan Willis, Haley Lind, Josh Moorehead

CSCE790-007 Spring 2025