Transformer-Based Multimodal Pain Assessment Using Facial Expressions

Meysam Safarzadeh, Maoqin Zhu, Shishir Prasad, Sudeshna Das, Peter Xie, Joohyun Chung, Xian Du
University of Massachusetts Amherst
MassAITC [AD/ADRD *or* Aging] Focus Pilot Core

Introduction

Traditional pain assessment relies on self-reports or caregiver evaluations, which are subjective and inconsistent. To improve accuracy, we propose a **transformer-based multimodal model** using **RGB, depth, and thermal images** from the **MIntPAIN dataset**. Unlike prior methods, our approach captures spatiotemporal patterns, enhancing pain prediction.

Key Contributions:

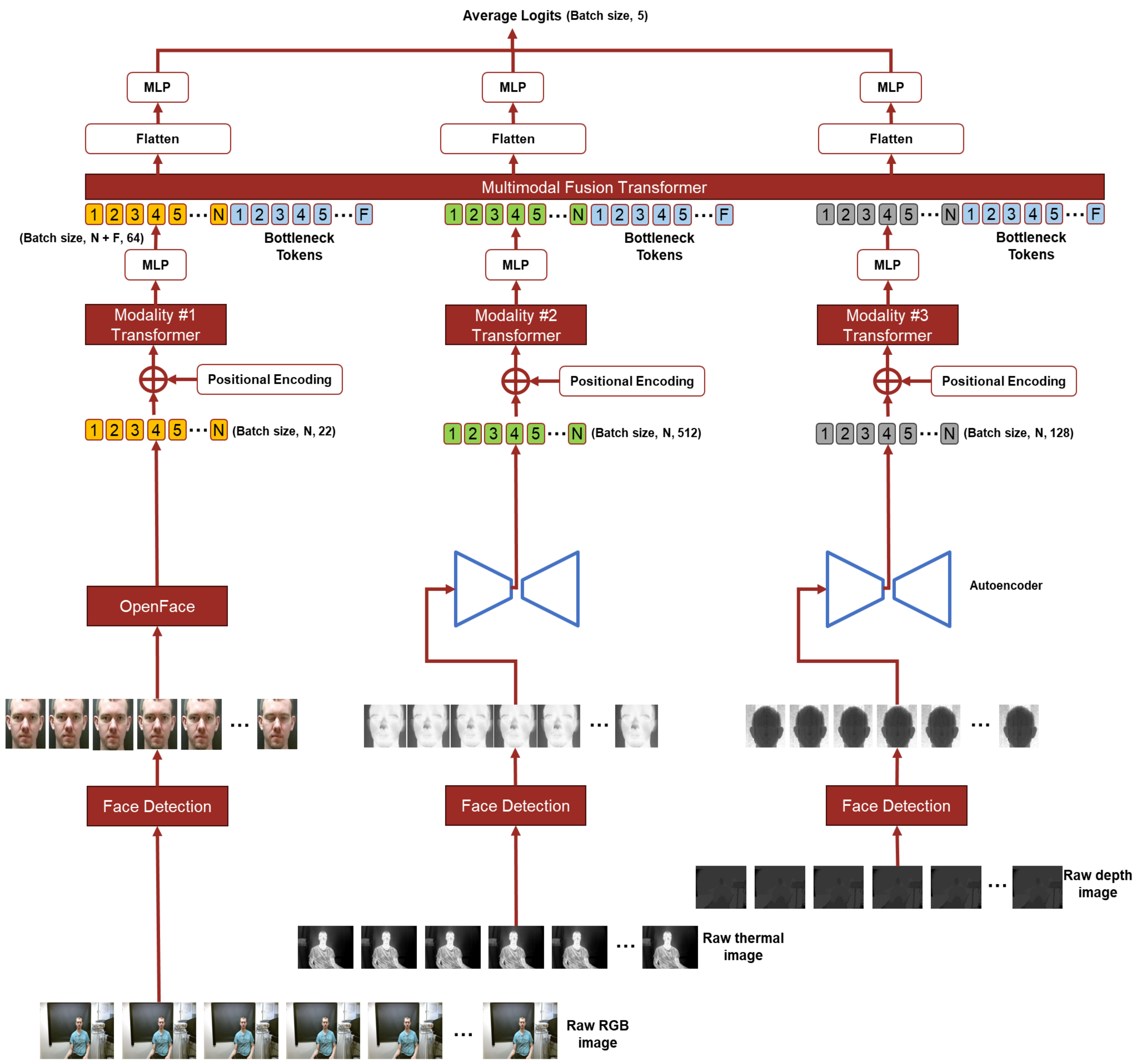
- **Efficient Multimodal Fusion:** Uses bottleneck tokens for effective cross-modal interaction.
- **Temporal Feature Extraction:** Leverages multi-head attention to identify critical pain indicators.
- **Improved Explainability:** Attention maps and integrated gradients provide insight into model decisions.

Our approach offers an **objective, interpretable, and intelligent** pain assessment method for improved clinical decision-making.

Objectives

- Develop a transformer-based multimodal model for pain assessment using RGB, depth, and thermal imaging.
- Capture spatiotemporal relationships in facial expressions to improve pain intensity prediction.
- Enhance multimodal fusion using bottleneck tokens for efficient cross-modal interactions.
- Utilize multi-head attention to extract critical temporal pain indicators.
- Provide an objective, interpretable, and clinically useful pain assessment method.

Method: A transformer-based model for multimodal pain assessment using RGB, thermal, and depth images from the MIntPAIN dataset.



Key Steps:

- **Face Feature Extraction**
- **Transformer-Based Encoding:** Positional encoding, Modality-specific transformers, MLP standardization
- **Multimodal Fusion:** Bottleneck tokens and Iterative transformer layers
- **Classification:** MLP and SoftMax layer

This method **captures spatiotemporal patterns**, enhances **multimodal integration**, and improves **pain prediction accuracy**.

Conclusions &Takeaways

Our transformer-based model **outperforms prior deep multimodal methods** in pain assessment, achieving higher accuracy across **single and multimodal settings**.

Key Findings:

- **Modality Importance: RGB** is the most informative, followed by **depth** and **thermal,** due to resolution and feature limitations.
- **Multimodal Fusion Boost:** Combining modalities improves accuracy, with **RGB + Depth** performing best.
- **Comparison with Nurses:** Our model achieves **42.6% accuracy** in five-class classification, comparable to nurses' **63.5% accuracy** in three-class assessment.
- **Feature Importance:** Key **facial action units (AUs)** for pain prediction include **blink (AU 45), nose wrinkler (AU 9), lip tightener (AU 23), inner brow raiser (AU 1), and jaw drop (AU 26).**
- **Attention maps** show different heads capture **global (positional) and local (content-driven) features**, improving interpretability.

Our model **effectively captures spatiotemporal features**, improving **pain prediction, interpretability, and clinical utility**.

Acknowledgement

This project was partially supported by **Mass AI and Technology Center for Connected Care in Aging and Alzheimer's Disease**, the **IALS Translational Graduate Student Assistantship, Pilot Grant Award from the Elaine Marieb Center for Nursing and Engineering Innovation,** and the **Center for Personalized Health Monitoring**.