Isabella Smith

Focus Tracker: Engagement Detection

CSCI 4050U

Agenda

The Problem

Students often lose focus during online learning. The goal is to detect emotional state (FER-2013) and engagement level (DAiSEE) in real-time and send notifications when attention drops.

Why this matters:

- Real-time feedback can improve focus and learning outcomes
- Combines facial expression detection and engagement prediction

FER-2013 Dataset

Full Name: Facial Expression Recognition 2013.
Purpose: Standard benchmark for detecting basic human emotions.
Data Size: Approximately **35,887** images.
Format:

- **48x48 pixel** resolution.
- **Grayscale** (1 channel).
- Cropped specifically to the face.

Classes (7 Emotions):

- Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral.

DAiSEE Dataset

Full Name: Dataset for Affective States in E-Environments.
Purpose: Specifically designed to recognize user engagement in online learning settings.
Data Size: 9,068 video snippets (extracted frames for this project).
Format:

- Full-color (RGB) videos.
- Captures diverse lighting and head poses typical of webcam usage.

Classes (4 States):

- **Engagement** (The primary target).
- Boredom, Confusion, Frustration.

Labels: Ranked by intensity levels **0 (Very Low)** to **3 (Very High)** rather than simple Yes/No.

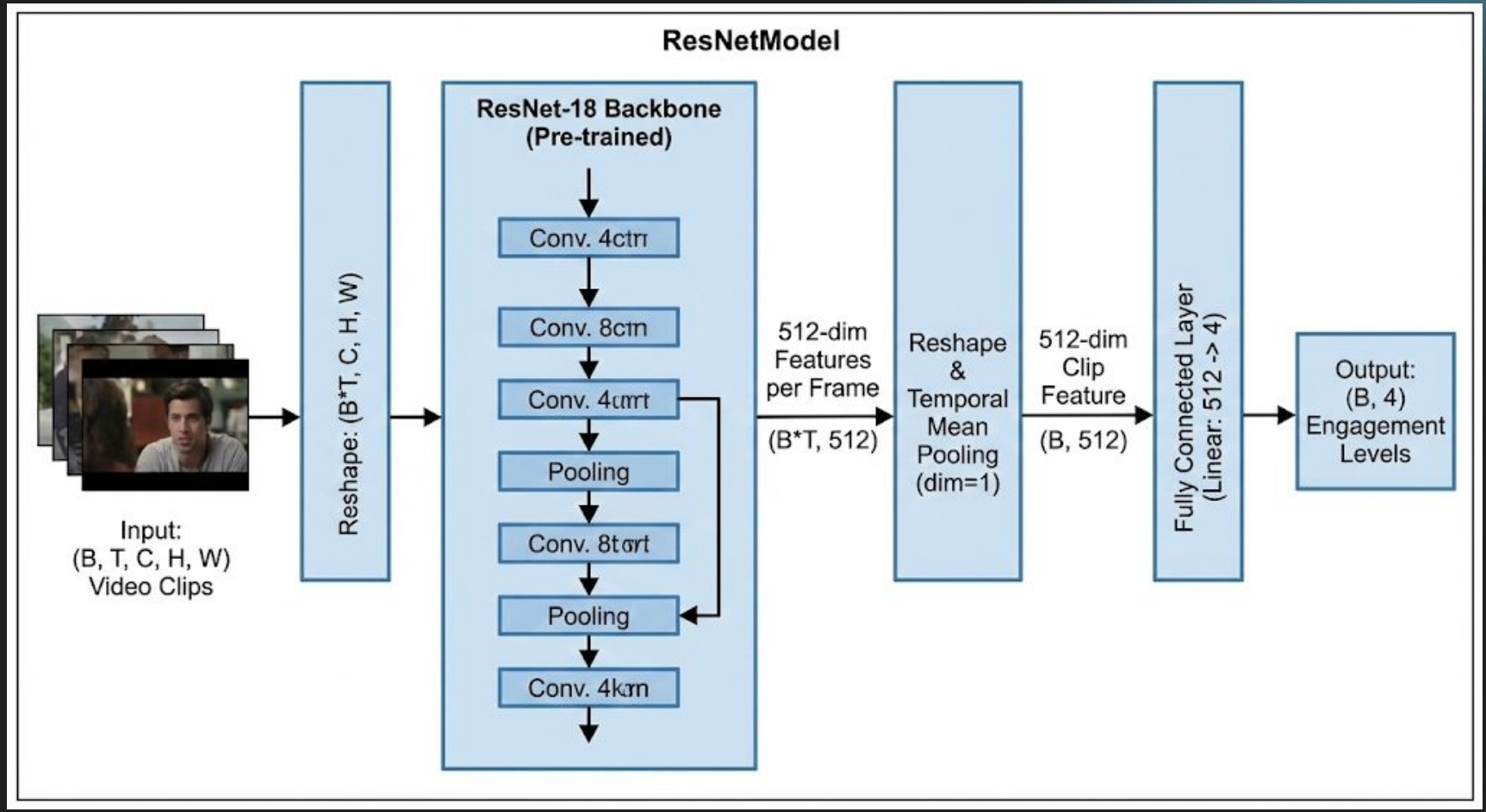
The Model

Backbone: Standard ResNet-18, pre-trained on ImageNet (transfer learning).

Input Handling: Designed to process video clips, not just single images. It expects a batch of inputs with shape (Batch, 5, 3, 64, 64) — representing 5 frames of video, each 64x64 pixels, in RGB color.

- Instead of using heavy 3D convolutions or LSTMs, this model:

1. **Flattens:** It treats the 5 video frames as independent images (Batch * 5).
2. **Extracts:** The ResNet backbone extracts features for all 5 frames simultaneously.
3. **Averages:** It calculates the mean of the features across the time dimension (.mean(dim=1)).


ResNetModel
Input:
(B, T, C, H, W)
Video Clips
Reshape: (B*T, C, H, W)
ResNet-18 Backbone
(Pre-trained)
Conv. 4ctn
Conv. 8ctn
Conv. 4ctrt
Pooling
Conv. 8tart
Pooling
Conv. 4kan
512-dim
Features
per Frame
(B*T, 512)
Reshape
&
Temporal
Mean
Pooling
(dim=1)
512-dim
Clip
Feature
(B, 512)
Fully Connected Layer
(Linear: 512 -> 4)
Output:
(B, 4)
Engagement
Levels

Data Remapping

- **DAiSEE:** Video clips labeled with **Engagement** (0-3).
- **FER-2013:** Static images labeled with **7 Emotions** (Happy, Sad, Angry, etc.).

The Solution (The "Remap"):

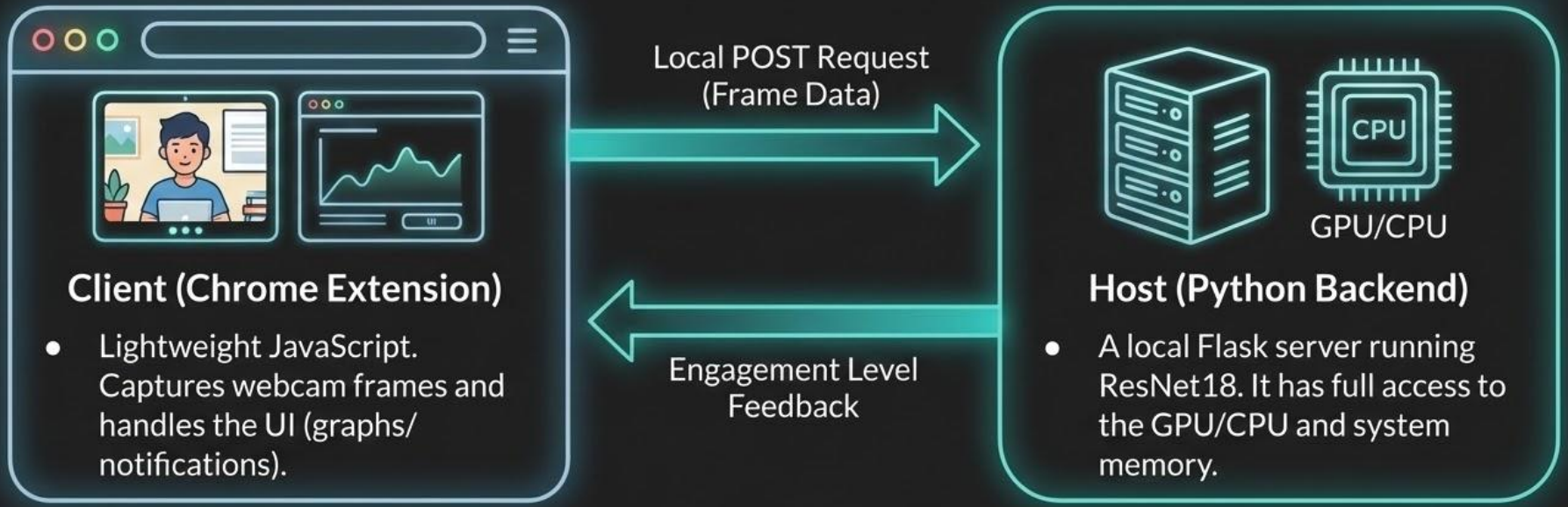
- *Happy / Surprise* → **Level 3 (High Engagement)**
- *Sad* → **Level 2 (Medium Engagement)**
- *Fear* → **Level 1 (Low Engagement)**
- *Angry / Disgust* → **Level 0 (Very Low Engagement)**

The "Fake Video" Technique:

- Since the model expects 5 video frames, but FER-2013 only has static images, the code uses torch.stack([image] * 5).
- This duplicates the single emotion image 5 times to simulate a "static video," allowing the model to train on both datasets simultaneously without crashing.

Deployment

Overview:



Deployed model as a Google Chrome Extension to provide real-time, non-intrusive monitoring directly within the browser. This 'Focus TrackeTracker' runs locally, analyzing webcam data without sending video to a cloud server (ensuring privacy).

Data Flow

Capture: The extension captures a frame from the webcam every few seconds via JavaScript.

Transmit: The frame is converted to a base64 string and sent to the local Python server via a POST request.

Analyze: The Python server pre-processes the image (resize/normalize) and feeds it into the **Model.**

Feedback: The model returns Engagement Level.

Display: The extension updates the UI with a live graph and sends a notification if engagement drops too low (e.g., "Distracted" or "Low Energy").

DEMO

(Refer to video)