Human Activity Recognition Using 3D CNN and Transformer Encoder

Aria Sokhangoo
Deep Learning, AI Engineering

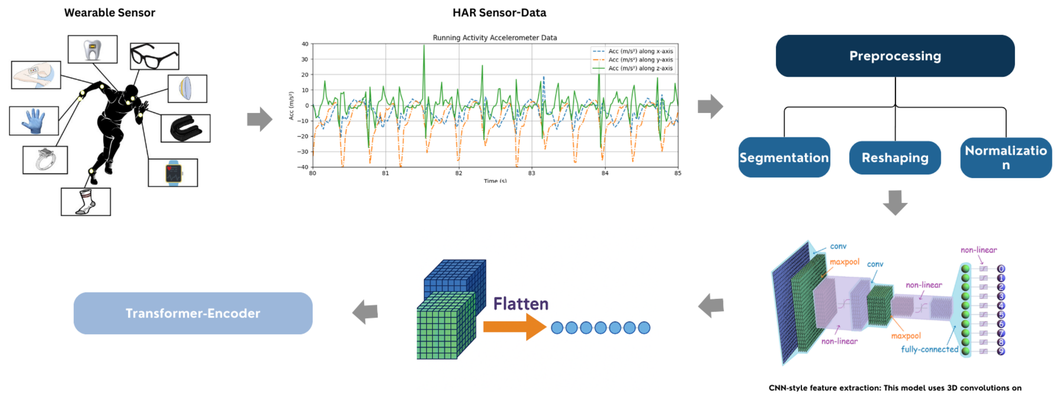
Abstract

- This project focuses on recognizing human activities using motion sensor data from smartphones, such as accelerometer and gyroscope readings.
- We combine **3D Convolutional Neural Networks (3D CNN)** with a **Transformer Encoder** to capture both spatial and temporal features of movement.
- The model was trained on the public **UCI HAR dataset**, which includes six labeled activities like walking, sitting, and laying.
- Results show high accuracy: **97.5% training** and **96.5% validation**, with strong generalization and low overfitting.
- This approach can be applied in health monitoring, fitness tracking, or smart environments using real-world wearable devices.

Motivation

- Understanding human physical activities has real-world value in healthcare, fitness apps, and smart environments.
- Traditional activity recognition systems often rely on hand-crafted features or basic models, limiting their accuracy and adaptability.
- [*Challenge*] • Sensor data is noisy and complex • Needs both spatial (movement patterns) and temporal (time sequence) modeling.
- Our approach: Combine **3D CNN** to capture spatial signal features and a **Transformer Encoder** to learn activity patterns over time.
- Goal: Build an accurate and generalizable deep learning model using only raw sensor input — no hand-engineering needed.

Flow-Chart

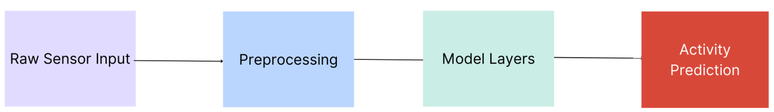


Dataset

- I used the publicly available **UCI HAR Dataset**, which contains motion sensor data collected from a smartphone worn on the waist by 30 subjects performing 6 activities.
- The sensors include a 3-axis accelerometer and gyroscope sampled at 50 Hz.
- The dataset is segmented using a sliding window of 2.56 seconds with 50% overlap, resulting in 561 features per time step.

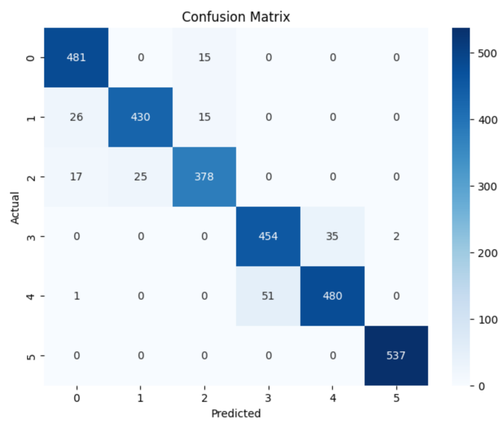
Experiment Setup

- **Input:** Preprocessed time-series sensor data shaped into 32×6 for each sample.
- **Model:** Combined architecture of 3D CNN (for spatial features) and Transformer Encoder (for temporal sequence learning).
- **Training:** 18 epochs, Adam optimizer, categorical cross-entropy loss.
- **Output:** One of six activity classes: Walking, Walking Upstairs, Walking Downstairs, Sitting, Standing, Laying.

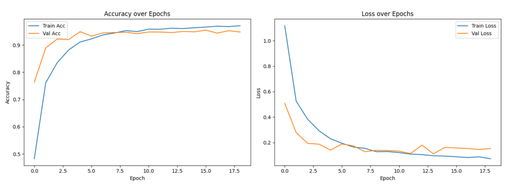


Results

- The combined **3D CNN + Transformer Encoder** model achieved **97.5% training** and **96.5% validation accuracy**.
- Confusion matrix highlights high precision in static activities (e.g., **Laying**, **Standing**) with few misclassifications in dynamic ones.
- Accuracy and loss curves show steady training with minimal overfitting.



Confusion Matrix of activity classification.



Accuracy and loss across 18 epochs.

Contribution

- Combined 3D CNN and Transformer for HAR.
- Used raw sensor data (accel + gyro).
- Achieved 96.5% validation accuracy.
- Built clear visual flow and result plots.
- Ready for real-world wearable deployment.