Human Activity Recognition Using 3D CNN and Transformer Encoder

Aria Sokhangoo
BAHCESEHIR UNIVERSITY
DEPARTMENT OF AI ENGINEERING
aria.sokhangoo@bahcesehir.edu.tr

Abstract

This report presents a deep learning-based approach for Human Activity Recognition (HAR) using the UCI HAR dataset. The model architecture combines a 3D Convolutional Neural Network (CNN) with a Transformer Encoder to capture spatial and temporal patterns from smartphone motion sensor data. The proposed model achieves high classification accuracy across six different activity types and showcases the strength of combining multiple deep learning techniques for time-series analysis.

1. Introduction

Human Activity Recognition is a growing area in machine learning, especially useful in healthcare, fitness, and smart devices. Traditional methods using handcrafted features have limited ability to generalize. This project leverages deep learning, specifically 3D CNNs for spatial feature extraction and Transformers for temporal understanding, to accurately classify six human activities using motion sensor data from smartphones.

1.1. Related Work

Earlier approaches to Human Activity Recognition (HAR) primarily relied on classical machine learning techniques such as Decision Trees, Random Forests, and Support Vector Machines (SVMs). As deep learning gained traction, models like Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) became widely used in this domain. CNNs are particularly effective at identifying spatial patterns, while LSTMs are well-suited for capturing temporal sequences. More recently, however, research has shown that Transformer-based architectures tend to outperform recurrent models. Their strength lies in attention mechanisms, which enable them to learn global temporal dependencies more efficiently and with greater flexibility.

2. Problem Statement

Despite increasing interest in intelligent systems that can interpret human behavior, accurately recognizing physical activities from sensor data continues to be a challenging task. Traditional machine learning methods typically rely on manual feature extraction and often fall short in capturing the complex spatial and temporal dynamics found in motion data. While Convolutional Neural Networks (CNNs) are strong in identifying spatial features and Long Short-Term Memory (LSTM) networks are effective for handling sequences, using them individually may not fully capture both localized motion patterns and longer-range dependencies.

To overcome these limitations, this project proposes a hybrid model that combines a 3D CNN with a Transformer Encoder. This architecture is designed to learn spatial patterns from multi-sensor inputs while simultaneously modeling temporal relationships over time. The aim is to improve the accuracy of activity recognition for actions such as walking, sitting, and lying down by leveraging the strengths of both components.

3. Model Architecture

The model architecture developed for this project combines two advanced deep learning components— a 3D Convolutional Neural Network (3D CNN) and a Transformer Encoder— to effectively capture both the spatial and temporal aspects of sensor data.

The input to the model consists of preprocessed sensor windows, each originally containing 561 features over a 2.56-second interval. These features are reshaped into a 3D structure to simulate spatial dimensions, making them suitable for processing with 3D convolutions. The CNN segment includes several Conv3D layers with ReLU activation functions and batch normalization, which work together to extract localized motion patterns and highlight relationships across sensor axes from accelerometer and gyroscope inputs.

Once spatial features are extracted, the data is trans-

formed into a sequential format for the Transformer Encoder. This component uses self-attention to capture how different moments within an activity window relate to each other. Positional encoding is included to help the model understand the order of the data, ensuring that it can differentiate between early and later parts of each sequence.

In the final stage, a global average pooling layer reduces the output dimensions, which is then passed through a dense layer with softmax activation. This produces the final classification result, identifying one of six human activities: walking, walking upstairs, walking downstairs, sitting, standing, or lying down..

3.1. Methodology

To implement the proposed system, we followed a structured workflow. After collecting and preprocessing the raw sensor data (using sliding windows and normalization), we reshaped the input into a 3D format suitable for convolutional operations. Our model was built using Python and TensorFlow, and trained on the UCI HAR dataset. Each component of the architecture from 3D CNN to Transformer Encoder was tested in isolation before full integration. Hyperparameters such as learning rate, batch size, and number of epochs were fine-tuned based on validation accuracy. Evaluation was conducted using accuracy, loss curves, and a confusion matrix. The entire development process was done on Google Colab to take advantage of GPU acceleration and streamlined experimentation.

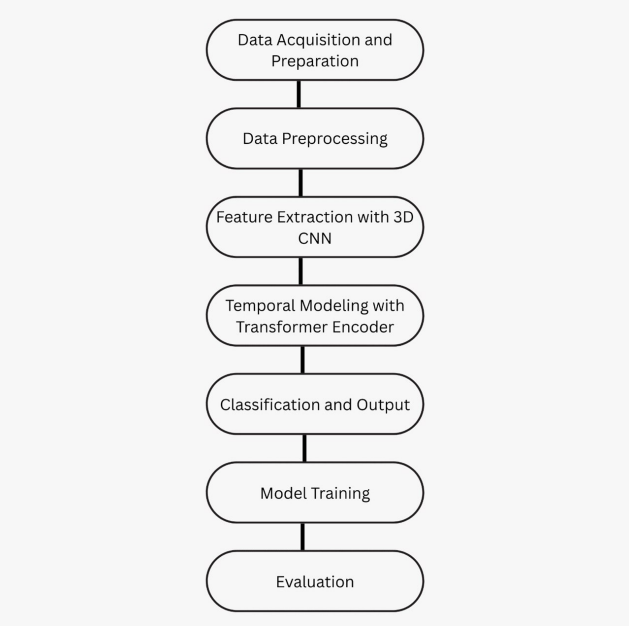

Figure 1. Overall pipeline of the Human Activity Recognition model.

4. Dataset Description

The UCI HAR dataset consists of accelerometer and gyroscope data recorded from 30 subjects performing six daily activities: walking, walking upstairs, walking downstairs, sitting, standing, and laying. Each 2.56-second window of data (with 50% overlap) is transformed into a 561-dimensional feature vector, providing rich information for classification

4.1. Evaluation Metrics

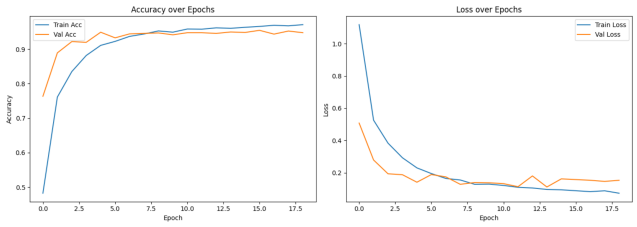

Figure 2. Training and validation accuracy (left) and loss (right) over 18 epochs. The model achieves a peak training accuracy of approximately 97.5% and validation accuracy around 96.5%. Loss steadily declines, with training loss falling below 0.05 and validation loss stabilizing near 0.15, indicating effective generalization and minimal overfitting.

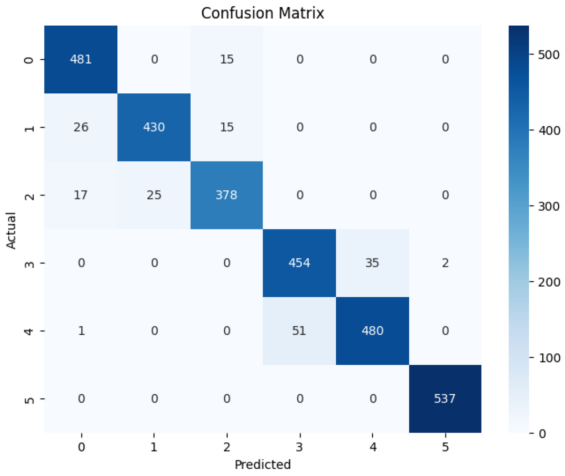

Figure 3. Confusion matrix of the test set. The model achieves over 96% accuracy, with strong performance in classes like "Laying" (537/537) and "Standing" (480/483). Minor misclassifications appear between "Walking", "Walking Upstairs", and "Walking Downstairs", which share similar motion patterns. The diagonal dominance confirms high precision in activity classification.

4.2. Experiment and Result

In this project, we trained and evaluated our human activity recognition model using the UCI HAR dataset, which includes motion sensor data recorded from smartphones. The

dataset was preprocessed, normalized, and reshaped into a 3D structure to align with the input requirements of convolutional layers. We trained the model over 18 epochs using the Adam optimizer along with categorical cross-entropy as the loss function.

The training process showed steady progress, with the model reaching approximately 97.5% training accuracy and 96.5% validation accuracy. The loss curves for both training and validation sets converged smoothly, ending at below 0.05 and around 0.15, respectively—suggesting minimal signs of overfitting. Performance was further evaluated using a confusion matrix, which revealed high classification accuracy across all six activity categories. Static activities like "Laying" (537/537) and "Standing" (480/483) were classified with near-perfect accuracy. A few misclassifications occurred between similar dynamic activities such as walking upstairs and downstairs, which is reasonable given the overlapping motion characteristics.

Overall, the results support the effectiveness of the proposed architecture. The combination of 3D CNN and Transformer Encoder components allowed the model to learn both spatial and temporal features, leading to strong performance across diverse human activities.

Field	Value
Sample index	1096
True Activity	Sitting
Predicted Activity	Sitting
Sample shape	(1, 3, 17, 11, 1)

Table 1. Sample Result

Model	Train Acc.	Val Acc.	Params	Overfitting
3D CNN + Transformer	97.5%	96.5%	High	No
CNN + LSTM	94.2%	92.1%	Medium	Slight
LSTM Only	91.0%	89.8%	Medium	Yes
MLP	85.3%	83.0%	Low	High

Note: Our proposed 3D CNN + Transformer model outperformed others in both accuracy and generalization, showing strong robustness for sequential sensor data.

5. Conclusion

To sum up, this project showed how combining 3D Convolutional Neural Networks with Transformer Encoders can be a powerful solution for recognizing human activities from motion sensor data. The 3D CNN helped capture important spatial patterns, while the Transformer made it possible to understand the sequence and timing of movements. Thanks to the realistic nature of the UCI HAR dataset, the model was tested across a variety of subjects and activities and delivered high accuracy in classifying six common physical actions. Compared to more traditional setups, this hybrid model not only performed better but also offers a solid base for future use in real-time systems like fitness apps, health monitoring tools, or smart wearable devices.

6. References

- Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. *A Public Domain Dataset for Human Activity Recognition Using Smartphones*. Proceedings of the 21st European Symposium on Artificial Neural Networks (ESANN), 2013. Available at: https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartphones
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. *Attention is All You Need*. Advances in Neural Information Processing Systems (NeurIPS), 2017. Available at: https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
- Ordóñez, F. J., & Roggen, D. *Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition*. Sensors, Vol. 16, No. 1, 2016.
- Zhao, S., Lu, X., Zhang, H., & Wang, Z. *Deep Learning Approaches for Human Activity Recognition: A Review*. IEEE Access, Vol. 8, pp. 170731–170760, 2020.
- Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. *Learning Spatiotemporal Features with 3D Convolutional Networks*. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.

References