Spatio-Temporal Dual Attention with Cross-Sensor Attention for Enhanced IMU-based Human Activity Recognition

Index Terms: Human Activity Recognition, IMU Data, Cross-Sensor Attention, Transformers

Human Activity Recognition (HAR) using Inertial Measurement Unit (IMU) sensor data has emerged as a critical research domain with applications spanning healthcare monitoring, surveillance systems, and sports analytics. Despite significant advances in deep learning architectures, existing approaches exhibit limitations in capturing complex spatiotemporal dependencies inherent in human activities and fail to exploit coordinated movements across multiple body segments. This work presents the Spatio-Temporal Dual Attention Transformer with Cross-Sensor Attention (STDual-X), a novel framework for human activity recognition (HAR). STDual-X processes sensor readings from each IMU placed on a body part using a Spatio-Temporal Dual Attention Transformers (STDAT) (1) to generate a vector that encodes temporal and spatial movement patterns. The resulting vectors from multiple body-worn IMUs are subsequently input to the Cross-Sensor Attention (CSA) module, which applies cross-attention to model inter-sensor relationships, generating Cross-Sensor Relation (CSR) vectors that encode dependencies between sensors. The final embeddings, enriched with cross-sensor information, are processed through a fully connected neural network with a softmax layer to classify human activities. The model is optimized using a composite loss function combining Cross-Entropy Loss for accurate activity classification and a Contrastive Loss, based on triplet loss, to enhance the discriminative quality of the feature embeddings.

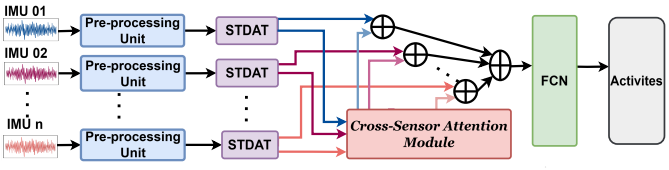

Figure 1: Overview of STDual-X

Extensive experiments on three benchmark datasets demonstrate the superior performance of STDual-X, achieving state-of-the-art accuracies of 97.02% on PAMAP2(2), 95.43% on Opportunity(3), and 98.80% on UCI-HAR (4) datasets. Comprehensive ablation studies validate the contribution of each architectural component, revealing that removal of STDAT reduces accuracy by 7.74%, elimination of CSA decreases performance by 2.34%, and exclusion of contrastive loss results in a 4.64% accuracy reduction. Quantitative analysis using t-SNE visualization demonstrates superior feature separability with a Silhouette score of 0.5826, significantly outperforming baseline approaches. Cross-sensor attention visualization reveals meaningful inter-sensor relationships that align with biomechanical principles of human movement, including high hand-ankle coordination during walking activities and uniform attention distribution during static postures.

The framework exhibits robust performance under adverse conditions, maintaining 96.78% accuracy with 10% missing data and demonstrating graceful degradation with noise injection. Computational efficiency analysis reveals an inference time of 1.6 milliseconds per sample with 144.72 MB peak memory usage on NVIDIA Tesla P100 hardware, indicating suitability for edge deployment scenarios. The proposed STDual-X framework advances the state-of-the-art in IMU-based HAR by effectively addressing spatiotemporal feature extraction challenges and establishing a new paradigm for IMU-based activity recognition systems. These contributions have significant implications for real-world applications requiring accurate, efficient, and privacy-preserving activity monitoring capabilities.

References

[1] D. Senarath, S. Tharinda, M. Vishvajith, S. Rasnayaka, S. Wickramanayake, and D. Meedeniya, "Behaveformer: A framework with spatio-temporal dual attention transformers for imu-enhanced keystroke dynamics," in *2023 IEEE International Joint Conference on Biometrics (IJCB)*, 2023, pp. 1–9.

[2] A. Reiss, "PAMAP2 Physical Activity Monitoring," UCI Machine Learning Repository, 2012.

[3] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Förster, G. Tröster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Sagha, H. Bayati, M. Creatura, and J. del R. Millán, "Collecting complex activity datasets in highly rich networked sensor environments," *Seventh International Conference on Networked Sensing Systems (INSS)*, pp. 233–240, 2010.

[4] A. D. G. A. O. L. Reyes-Ortiz, Jorge and X. Parra, "Human Activity Recognition Using Smartphones," UCI Machine Learning Repository, 2013.