Skeleton-Based Human Action Recognition Using Graph Neural Networks

Author: Khuloud Halimeh
M.Sc. in Computer Engineering – OSTİM Teknik Üniversitesi, Ankara, Turkey
Year: 2025

1. Introduction

Human Action Recognition (HAR) is a fundamental issue in computer vision, allowing machines to interpret human actions for applications in healthcare, intelligent environments, robots, and human-computer interaction. Conventional HAR techniques depend on RGB video, leading to issues such as background interference, lighting fluctuations, privacy issues, and substantial processing demands. Skeleton-based Human Activity Recognition provides a more efficient and privacy-conserving option by depicting human mobility via joint coordinates. Recent advancements in graph neural networks (GNNs) have markedly enhanced skeleton-based human activity recognition (HAR). Models like ST-GCN, AGCN, CTR-GCN, and ShiftGCN acquire spatio-temporal correlations among joints. This thesis conducts a comprehensive examination of these designs using a single GPU, analyzing their accuracy, efficiency, and feasibility of deployment within realistic computing limitations.

2. Research Objective

This thesis aims to do a thorough, reproducible comparison of cutting-edge graph-based human activity recognition models on a single GPU, while examining the trade-offs among model complexity, accuracy, and computational resources. The research evaluates AGCN (2-stream), CTR-GCN, and ShiftGCN on the NTU RGB+D 60 dataset employing Cross-Subject (X-Sub) and Cross-View (X-View) protocols. Essential inquiries encompass the performance of prominent GCN designs under constrained hardware, the resultant accuracy–efficiency trade-offs, and the optimization of training pipelines.

3. Methodology

The NTU RGB+D 60 skeleton dataset, comprising 60 motion categories and various camera perspectives, was utilized for all studies. Preprocessing encompassed skeleton normalization, frame alignment, joint re-indexing, class balance, and GPU-optimized batching.

Assessed Models:

- AGCN (2-Stream): Acquires adaptive adjacency matrices to enhance graph topology.
- CTR-GCN: Employs channel-specific topology refinement to improve feature extraction.
- ShiftGCN: An efficient architecture employing shift operations with few FLOPs.

Training Configuration: The experiments utilized a single NVIDIA V100 GPU, implemented mixed-precision training, employed gradient accumulation, utilized memory-efficient checkpointing, and incorporated batch sizes of 64 and 72.

4. Key Results

AGCN (2-Stream) attained 88.14% X-Sub and 88.27% X-View accuracy, demonstrating robust overall performance with little accuracy declines relative to multi-GPU benchmarks.
ShiftGCN attained 83.18% X-Sub / 84.87% X-View at a batch size of 64 and 86.34% X-Sub / 87.14% X-View with a batch size of 72, indicating remarkable efficiency and appropriateness for real-time applications.

CTR-GCN attained 81.52% on X-Sub and 84.15% on X-View with the Lion optimizer, resulting in rapid convergence but increased memory consumption. Single-GPU training across all models led to a 1.4%–3% decrease in accuracy attributable to reduced batch sizes, with evident trade-offs between accuracy, efficiency, and computational expense.

5. Contributions

This thesis presents an integrated training and preprocessing pipeline for skeleton-based human activity recognition, benchmark results for various graph convolutional network topologies under single-GPU limitations, optimization techniques including gradient accumulation, and analyses of model deployment trade-offs. It also tackles dataset issues such as class imbalance and irregular skeleton encoding.

6. Future Work

Future endeavors may involve the development of hybrid architectures that integrate AGCN adaptability with ShiftGCN efficiency, the investigation of attention-enhanced GCNs, the implementation of self-supervised learning to diminish reliance on labeled data, and the optimization of models via pruning, quantization, and neural architecture search for real-time and embedded applications.

7. Conclusion

This thesis shows that modern graph-based deep learning models can get superior performance in skeleton-based human activity recognition, even when limited to a single GPU. AGCN delivers superior overall accuracy, ShiftGCN ensures optimal efficiency, and CTR-GCN achieves a balance between refinement and speed. This study establishes a robust basis for subsequent investigations in multimodal human activity recognition, efficient real-time systems, and artificial intelligence-enhanced healthcare monitoring.