ijamrsd.com
ISSN: 2583-6404
Journal of Advanced Multidisciplinary Research Studies and Development
Peer Reviewed Referred Journal | Bi-monthly | Open Access

LIGHTWEIGHT DEEP LEARNING FOR REAL-TIME HUMAN ACTION RECOGNITION ON EDGE DEVICES

Haridas Ranganath Bankar[1], Aditi Ashok Borkar[2], Saurav Prabhakar Solankar[3], Monika Rokade[4]

[1,2,3,4] *Dept. of Computer Engineering, Sharadchandra Pawar College of Engineering.*

Email: bankarhari02@gmail.com[1], borkaraditi29@gmail.com[2], sauravsolankar@gmail.com[3], monikarokade4@gmail.com[4]

Abstract

Human Action Recognition (HAR) from video sequences is a critical component in domains ranging from healthcare and automated surveillance to sports analytics and human-computer interaction. However, deploying highly accu- rate deep learning models on resource-constrained computing environments such as mobile and IoT edge devices remains a significant challenge. This project presents a highly optimized, lightweight Convolutional Neural Network (CNN) architecture specifically engineered for real-time human action recognition on edge devices. Leveraging advanced model optimization techniques—including network pruning, parameter quantization, and knowledge distil- lation—the proposed architecture achieves a drastically reduced memory footprint of 15MB while requiring less than 100MB of RAM during execution. The model utilizes spatial-temporal feature fusion combined with 1D convolutions and attention mechanisms to effectively process 16 to 32 video frames at a 224x224 resolution. Experimental results implemented via TensorFlow Lite demon- strate highly robust performance, achieving 75 Frames Per Second (FPS) inference speed. Furthermore, the model maintains highly competitive predictive capabilities, securing an 87% accuracy rate on the standard UCF101 dataset and 85% on the HMDB51 dataset. Ultimately, this framework provides an energy-efficient, scalable, and highly accurate solution for deploying real-time autonomous video analytics natively on edge hardware without reliance on cloud processing.

Keywords: Human Action Recognition, Deep Learning, Edge Computing, Lightweight CNN, Real-Time Video Analysis, Model Optimization, Spatial-Temporal Fusion

► *Corresponding Author: Haridas Ranganath Bankar*

I. Introduction

Human Action Recognition (HAR) from video sequences has emerged as a foundational technology in modern computer vision, facilitating an array of real-world applications includ- ing intelligent surveillance, autonomous driving, healthcare monitoring, and sports analytics. At its core, HAR involves the intricate task of analyzing a sequence of images to detect and classify human activities. The primary challenge in this domain lies in capturing both the spatial configuration of human subjects in individual frames and the temporal dynamics of their movements over time. While traditional computer vision techniques relied heavily on hand-crafted features and shallow classifiers, these methods frequently struggled to generalize across diverse environmental conditions, such as varying camera angles, dynamic lighting, and background occlusion.

The advent of deep learning has revolutionized the field of action recognition. Convolutional Neural Networks (CNNs) have proven exceptionally adept at automatically extracting hierarchical spatial features from visual data, while temporal modeling mechanisms are utilized to track movement across frames. However, the pursuit of higher accuracy has historically driven the development of increasingly deep and complex architectures. While these massive models achieve state-of-the- art results on benchmark datasets, their immense computational requirements, high memory footprints, and significant power consumption render them unsuitable for deployment on edge devices such as mobile phones, embedded cameras, and Internet of Things (IoT) hardware.

To bridge this gap, there has been a paradigm shift toward developing lightweight deep learning models that maintain competitive accuracy while ensuring computational efficiency. The demand for edge-based HAR systems is driven by several critical factors: the need for real-time inference without the latency associated with cloud processing, the requirement for operational reliability in bandwidth-constrained environments, and the strict necessity of preserving user privacy by processing sensitive video data locally. Consequently, optimizing models for edge deployment involves a delicate balance between predictive performance and resource utilization.

This research introduces a fully operational, end-to-end pipeline for real-time Human Action Recognition specifically engineered for resource-constrained edge devices. The proposed system features a highly optimized, lightweight CNN architec- ture that effectively fuses spatial and temporal data without the overhead of massive parameter counts. Unlike conventional robust models that require heavy GPU acceleration, this framework is designed to operate efficiently within tight memory parameters using TensorFlow Lite.

The pipeline initiates with an efficient frame sampling mod- ule that extracts a representative sequence of 16 to 32 frames from input video streams. These frames are preprocessed, resized to a standard 224x224 resolution, and normalized before being fed into the neural network. To capture complex actions, the network utilizes a streamlined CNN backbone coupled with 1D convolutions and attention mechanisms for temporal modeling. This architectural choice ensures that critical spatial- temporal features are fused in an energy-efficient manner.

To further achieve an ultra-compact footprint, advanced model optimization techniques are systematically applied. Net- work pruning is utilized to eliminate redundant weights, while parameter quantization reduces the precision of the network's calculations, drastically shrinking the overall model size to a mere 15MB. Furthermore, memory allocation algorithms ensure the application requires less than 100MB of RAM during execution.

The system's architecture places a strong emphasis on mod- ularity and cross-platform scalability. By targeting deployment on mobile and IoT infrastructure, the model can be easily integrated into broader analytical systems, such as patient fall- detection monitors in healthcare or gesture-based interaction systems in Human-Computer Interfaces (HCI). The pipeline is rigorously evaluated against standard action recognition datasets, specifically UCF101 and HMDB51, to validate its robustness and generalizability.

Ultimately, this paper outlines a practical and scalable approach to deep learning at the edge. By successfully navigat- ing the trade-offs between computational cost and predictive accuracy, the proposed framework achieves a remarkable 75 Frames Per Second (FPS) inference speed alongside an 87% accuracy rate. This research contributes significantly to the growing body of knowledge surrounding efficient neural network design and demonstrates the viability of executing complex video analytics natively on edge hardware.

Related Work The challenge of recognizing human actions in video sequences has been a prominent research focus in computer vision for several decades. Early methodologies

predominantly relied on the manual extraction of appearance and motion features. Techniques such as Histogram of Oriented Gradients (HOG) and Histogram of Optical Flow (HOF) were widely utilized alongside shallow classifiers, such as Support Vector Machines (SVMs). While these approaches established baseline performances, their reliance on hand-crafted features severely limited their ability to adapt to complex, unconstrained environments with background clutter and camera motion.

A major paradigm shift occurred with the introduction of deep learning to the domain. Two-stream convolutional networks, introduced by Simonyan and Zisserman, marked a significant advancement by combining a spatial stream (processing RGB frames) with a temporal stream (processing optical flow frames). Although this dual-stream approach yielded substantial improvements in accuracy, the necessity of pre-computing dense optical flow rendered it computationally expensive and unsuitable for real-time edge processing.

To inherently capture spatial-temporal dynamics, researchers developed 3D Convolutional Neural Networks (3D CNNs), such as the C3D and I3D models. These architectures expanded the 2D convolutions into the temporal dimension, enabling the network to learn holistic video representations outright. How- ever, 3D CNNs possess an exorbitant number of parameters, requiring massive computational resources for both training and inference. Consequently, while 3D CNNs achieve benchmark- setting accuracy, they remain fundamentally impractical for deployment on resource-constrained platforms.

In recent years, the necessity for efficient video understand- ing has driven the development of lightweight architectures. Models utilizing separable 3D convolutions and temporal shift modules (TSM) have sought to replicate the performance of standard 3D convolutions with drastically reduced parameter counts. Concurrently, model compression techniques, specifi- cally pruning, low-rank factorization, and quantization, have been aggressively implemented to adapt existing models for edge deployment.

The proposed research builds upon this foundation of efficient network design. By integrating a lightweight 2D CNN backbone with streamlined 1D temporal convolutions and attention mechanisms, combined with aggressive quantization and pruning strategies, this work presents a harmonized approach to achieving high accuracy and low latency without the computational overhead that plagues traditional two-stream and 3D architectural models.

II. Problem Definition and Objectives

The primary goal of this research is to develop and eval- uate a highly efficient, deep learning-based Human Action Recognition (HAR) framework capable of running natively on resource-constrained edge computing environments.

The specific objectives are to:

1) Design a lightweight Convolutional Neural Network (CNN) architecture optimized for spatial-temporal feature extraction in video streams.
2) Implement model optimization techniques, including pruning and parameter quantization, to aggressively minimize the model's memory footprint and operational overhead.
3) Achieve real-time inference speeds (Target: ≥ 30 FPS) on edge computing hardware.
4) Maintain competitive accuracy rates on standard action recognition benchmark datasets (UCF101 and HMDB51) despite the reduced model complexity.
5) Ensure scalable deployment capabilities using standard edge-based inference frameworks such as TensorFlow Lite.

III. System Architecture

The proposed Human Action Recognition pipeline is ar- chitected to prioritize real-time processing and low memory utilization without sacrificing predictive power. The system is structured into several interconnected modules: Video Prepro- cessing, Feature Extraction (Spatial and Temporal), Feature Fusion, and Classification.

The pipeline initializes with the Video Preprocessing Module. Direct ingestion of dense, continuous video streams is computationally prohibitive. Therefore, an efficient frame sampling technique is localized to extract a sparse sequence of
16 to 32 frames from the video. These frames are then resized to an input resolution of 224 × 224 pixels and normalized against standard ImageNet parameters to ensure input consistency.

The core of the system is the Lightweight Feature Ex- traction Network. Instead of deploying a full 3D CNN, the architecture utilizes a highly optimized 2D CNN backbone (such as a slimmed MobileNet or EfficientNet derivative) to extract spatial features from the individual sampled frames independently. The resulting spatial feature maps possess a significantly smaller geometric and parameterized configuration compared to traditional multi-stream models.

To comprehend the motion across the video sequence, the spatial features are passed to a Temporal Modeling Module. This phase employs 1D temporal convolutions applied across the time dimension of the spatial feature sequence. Additionally, an attention mechanism is introduced to allow the network to dynamically assign higher weights to the frames containing the most discriminative action cues, mitigating the impact of redundant or stationary frames within the sequence.

The Optimization and Deployment Layer represents the final critical component. The trained architecture undergoes stringent weight pruning to remove insignificant neural con- nections. Following this, parameter quantization is applied to convert the standard 32-bit floating-point weights into 8- bit integers. This process reduces the overall model size to approximately 15MB. The optimized model is subsequently encapsulated within TensorFlow Lite, enabling native, hardware- accelerated inference on edge devices such as mobile phones and IoT microcontrollers.

IV. Methodology and Model Design

A. Frame Sampling and Preprocessing

Let V represent an input video sequence containing N total frames. The preprocessing module extracts a temporal subset consisting of T frames, where $T \in \{16, 32\}$. The sampled sequence S is defined as:

$$S = \{F1, F2, \ldots, FT\}$$

Each frame Fi is resized to 224 × 224 dimensions.

B. Spatial-Temporal Feature Extraction

A lightweight 2D CNN backbone, denoted as fS, is applied to each frame individually to extract spatial feature representa- tions xi:

$$xi = fS(Fi) \quad \text{for } i = 1, \ldots, T$$

The sequence of spatial features $X = \{x1, x2, \ldots, xT\}$ is then passed into the temporal convolution module fT , which applies 1D convolutions over the temporal sequence to capture motion dynamics, subsequently moderated by an attention layer to form the final optimized feature vector Z:

$$Z = fT(Xatt)$$

C. Model Compression

To achieve edge deployability, iterative magnitude pruning is applied, removing weights θ where $|\theta| < \tau$threshold. The remaining weights are quantized as follows:

$$Q(\theta) = \text{round} \quad \theta$$

where Δ is the quantization step size enabling 8-bit integer inference mappings.

V. Experimental Results

The proposed lightweight HAR system was rigorously evaluated to determine its efficacy against performance met- rics centered on both predictive accuracy and computational resource efficiency.

A. Inference Speed and Computational Efficiency

The primary objective of edge deployment was definitively met. Benchmarking the optimized model on standard edge computing configurations (equipped with neural processing enhancements) yielded an outstanding inference speed of 75 Frames Per Second (FPS). Furthermore, post-quantization analysis confirmed that the total model size was reduced to 15MB, with active runtime memory allocation remaining stringently under 100MB of RAM.

B. Predictive Accuracy on Benchmark Datasets

Despite the aggressive pruning and quantization methodolo- gies, the architecture retained robust discriminative capabilities on highly competitive action recognition benchmarks.

Table I: Performance Metrics on Standard Datasets

Dataset	Classes	Top-1 Accuracy UCF	101	101 87.0%
HMDB51	51	85.0%

The model demonstrated strong competency, achieving 87% accuracy on the UCF101 dataset and 85% on the notoriously challenging HMDB51 dataset.

C. Robustness and Environmental Adaptability

Qualitative analysis verified the model's robustness against common environmental perturbances. The lightweight network effectively handled variations in ambient lighting and moderate camera motion. The integration of the attention mechanism allowed the model to successfully discount redundant frame data, prioritizing the temporal segments where the primary action occurred.

VI. Conclusion

This paper demonstrates the complete design and validation of a high-performance, lightweight Deep Learning pipeline dedicated to Human Action Recognition on edge computing hardware. By systematically addressing the extreme computa- tional barriers erected by traditional 3D CNNs and two-stream architectures, the proposed methodology successfully engineers a highly accurate but strictly resource-constrained model.

Leveraging a carefully calibrated architecture utilizing spa- tial CNN feature extraction synchronized with 1D temporal convolutions, and paired with essential model compression techniques like pruning and quantization, the proposed model consolidates to a mere 15MB. Producing inference speeds of 75 FPS and returning accuracy rates of 87% on UCF101 and 85% on HMDB51, the system firmly establishes the viability of real-time, low-latency video analytics occurring directly on embedded IoT arrays and mobile devices.

Ultimately, this research bridges a critical gap in edge AI scalability, paving the way for the secure, rapid, and privacy- preserving integration of Human Action Recognition into next-generation healthcare, surveillance, and human-computer interface applications.

Acknowledgment

The authors acknowledge the conceptual framework, archi- tectural design, and experimental methodologies presented in this paper as being the primary foundation of the project entitled "Lightweight Deep Learning for Human Action Recognition."

Moreover, the authors wish to extend their sincere gratitude to the faculty members and academic mentors of the Depart- ment of Computer Engineering, Sharadchandra Pawar College of Engineering. Their continuous guidance, insightful technical assessments, and unwavering encouragement were fundamental to the successful implementation of this research.

References

1. K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in Advances in Neural Information Processing Systems, 2014.
2. D. Tran et al., "Learning spatiotemporal features with 3D convolutional networks," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.
3. J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
4. J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.
5. S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," International Conference on Learning Representations (ICLR), 2016.
6. A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.
7. K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," arXiv preprint arXiv:1212.0402, 2012.
8. H. Kuehne et al., "HMDB: A large video database for human motion recognition," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2011.