LAST: Lightweight Adaptive-Shift Transformer for Real-Time Human Action Recognition on Edge Devices

Research Proposal: Efficiency-First Human Action Recognition

---

1. Abstract

Current State-of-the-Art (SOTA) models for Human Action Recognition (HAR), such as VideoMAE and SlowFast, achieve high accuracy but are computationally prohibitive for real-time deployment on low-power edge devices. This research proposes **LAST**, a hybrid architecture that minimizes Floating Point Operations (FLOPs) by shifting from dense video processing to sparse skeletal representations. By combining **Adaptive Graph Convolutions (A-GCN)**, **Temporal Shift Modules (TSM)**, and **Linear Transformers**, this model aims to achieve SOTA-level performance with a fraction of the traditional computational cost.

2. Problem Statement

The primary challenges in modern HAR research are:

- **Computational Overload:** RGB-based 3D-CNNs require massive GPU memory.
- **Temporal Inefficiency:** Standard transformers have quadratic complexity \$O(N^2)\$ relative to video length.
- **Static Graph Limitations:** Traditional skeleton models use fixed physical connections, failing to capture complex interactions between non-adjacent joints.

3. Proposed Methodology

The LAST model follows a "More with Less" philosophy through three core pillars:

A. Adaptive Spatial Modeling (A-GCN)

Instead of a fixed skeleton, we implement a dynamic adjacency matrix. The model learns to create "virtual edges" between joints (e.g., hand-to-head during a phone call) based on the context of the action.

B. Zero-Parameter Temporal Modeling (TSM)

To capture movement through time without the cost of 3D convolutions, we utilize the **Temporal Shift Module**. By shifting a portion of the feature channels along the time dimension, we achieve temporal information exchange at **zero computational cost**.

C. Linearized Global Context

We employ a Transformer head using **Linear Attention**. This reduces the complexity to $O(N)$, allowing the model to process long action sequences (e.g., 30+ seconds) on devices with limited RAM.

D. Training via Knowledge Distillation

We will use a **Teacher-Student paradigm**. A high-performance, heavy-duty RGB model (e.g., VideoMAE V2) will serve as the "Teacher," guiding the "Student" (LAST) to learn rich semantic features from simple skeletal data.

---

4. Mathematical Framework

The core operation of a LAST block for feature X is defined as:

1. **Adaptive Graph Conv:** $X_{spatial} = \sigma(\sum(A_{fixed} + B_{learned} + C_{sample})XW)$
2. **Temporal Shift:** $X_{temp} = \text{Shift}(X_{spatial}, \pm 1)$
3. **Linear Attention:** $Output = $\phi(Q)(\phi(K)^T V)/(\phi(Q)\sum\phi(K)^T)$

This sequence ensures that spatial, temporal, and global relationships are computed in linear time.

---

5. Feasibility and Resource Requirements

Data & Tools

- **Datasets:** NTU RGB+D 120 (Skeletal) and Kinetics-400/700 for pre-training.
- **Hardware:** Training can be completed on a single mid-range GPU (e.g., RTX 30-series/40-series). Deployment target is a mobile CPU or Raspberry Pi.
- **Frameworks:** PyTorch, PyTorch Geometric, and MediaPipe for real-time skeleton extraction.

Feasibility Score: High

The approach is highly feasible because it leverages **pre-existing SOTA weights** for distillation and focuses on **memory-efficient operations** that are already supported by standard inference engines (ONNX, TFLite).

---

6. Expected Outcomes

- **Latency:** $<10ms$ inference time on standard mobile processors.
- **Accuracy:** Competitive with SOTA on NTU-120 ($>90\%$).
- **Contribution:** A novel architecture that enables high-accuracy HAR in privacy-sensitive and power-constrained environments (e.g., home healthcare or factory safety).

---

7. Supporting Previous Works

- **Yan et al. (2018):** Established the foundation of Spatial-Temporal GCNs.
- **Lin et al. (2019):** Proved the efficiency of Temporal Shift Modules in video.
- **Katharopoulos et al. (2020):** Introduced Linear Transformers for efficient sequence modeling.
