

Deep Learning Models Optimizations for Real-Time Intelligent Video Analytics

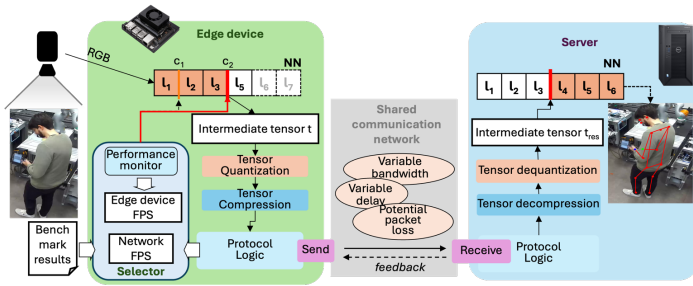
Michele Boldo, Nicola Bombieri

Dept. of Engineering for Innovation Medicine, University of Verona, Italy
michele.boldo@univr.it - nicola.bombieri@univr.it

- Real-time video stream **Deep Learning (DL)** based analysis is gaining importance
- Shift from cloud-based computing to **edge-based** processing to:
 - Reduce **latency**
 - Protect **privacy**
- DL models require substantial computational resources to achieve high accuracy.
- Challenges:** computation, transmission and adaptation

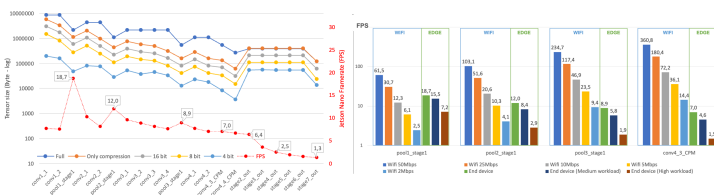
Collaborative Deep Inference

- High accuracy** often is related to **computational heavy DL** model
 - Deploying such models on edge devices poses challenges due to: latency constraints, memory limitations
 - Collaborative Deep Inference** mitigates these issues by:
 - Partitioning the DL model across multiple devices
 - Balancing computation between an edge device and a server
- CONTRIBUTION [1]:**
- Adaptive** framework dynamically selects a split point from candidate layers based on real-time performance metrics.
 - The edge device processes the initial layers up to the selected split point, ensuring:
 - Predefined latency constraints are met
 - Computational load and transmission time are optimized
 - Network bandwidth and workload are monitored using a custom UDP-based protocol
 - This approach maintains accuracy without requiring model retraining and architectural modifications



RESULTS:

- Tensor quantization and compression scheme has a minimal effect on the accuracy of human pose estimation (HPE).
- In gait analysis, small discrepancies in knee flexion angle estimation between our adaptive HPE framework and the marker-based motion capture system are unlikely to result in significant clinical differences
- Our framework meets the necessary computational performance requirements, keeping the maximum angular error within acceptable limits (i.e., 7 FPS).
- Low complexity of the runtime split point selector



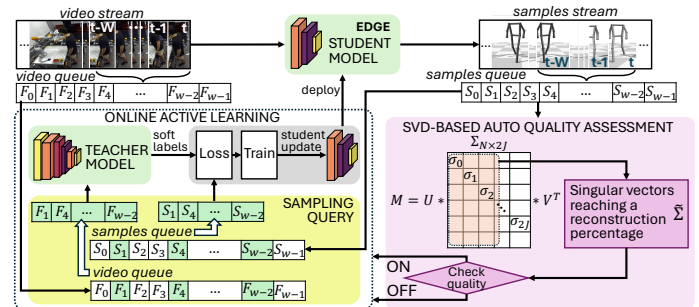
Online Domain Adaptation

- DL models for edge devices are usually **pre-trained**
- Low generalization capability due to:
 - Variations in real-world data
 - Out-of-Distribution (OOD) issues
- Fine-tuning** is essential to adapt models to deployment conditions
- Spatio-temporal models, used in applications like autonomous driving, must dynamically adapt
- Online fine-tuning** is challenging due to lack of ground truth in real-time data

CONTRIBUTION [2]:

Efficient training scheduler for spatio-temporal tasks that:

- Eliminates dependence on ground truth
- Determine the quality of the model prediction using the reconstruction percentage of the **Singular Value Decomposition (SVD)**
 - If reconstruction quality drops below a threshold, the teacher model generates soft labels and start the fine-tuning process
 - Otherwise, student predictions are accepted

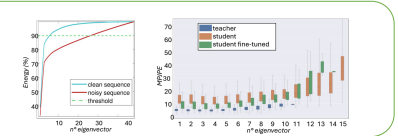


SVD-based prediction quality assessment

$$M = U \Sigma V^T$$

$$\hat{\Sigma} = [\sigma_1, \sigma_2, \dots, \sigma_r, 0, \dots, 0] \quad \hat{M} = U \hat{\Sigma} V^T \quad r < 2J$$

$$\frac{\|M - \hat{M}\|_F}{\|M\|_F} < \epsilon$$

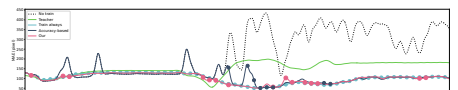


RESULTS:

- Evaluation on two spatiotemporal tasks: Human pose estimation (HPE) and Object Detection (OD)
- Three scenarios: dynamic, semi-repetitive, repetitive
- State-of-the-art accuracy
- Reduced training iterations by up to 90.6% in semi-repetitive and repetitive cases

Car Detection

Metric	MAE	% update	% teacher execution
Teacher	155.5	-	-
No train	245.4	0.0	0.0
Train always	100.2	100.0	100.0
Accuracy-based	108.7	5.3	100.0
Our	100.8	9.2	9.2



References

[1] M. Boldo, D. Carra, D. Quaglia, and N. Bombieri. "A Dynamic and Collaborative Deep Inference Framework for Human Motion Analysis in Telemedicine". In: IEEE International Conference On EDGE COMPUTING & COMMUNICATIONS (EDGE). 2023

[2] M. Boldo, M. De Marchi, E. Martini, S. Aldegheri, and N. Bombieri. "Domain-Adaptive Online Active Learning for Real-Time Intelligent Video Analytics on Edge Devices". In: IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 43.11 (2024), pp. 4105–4116.