b
UNIVERSITÄT
BERN

UNIVERSITÄT BERN

Resource-Aware Distributed Machine Learning on Heterogeneous IoT Devices

Eric Samikwa, Torsten Braun
October 6, 2022

Communication and Distributed Systems
Institute of Computer Science
University of Bern – Switzerland

eric.samikwa@unibe.ch

Agenda

Motivation

Cloud-based Machine Learning

- Raw user data transmission
- **Latency, Energy, Bandwidth issues**

Concerns of privacy

- Network transmission, central storage

Moving ML towards edge IoT devices

Rising demand for intelligent IoT systems

UNIVERSITÄT BERN

Motivation

On-Device ML implementation?

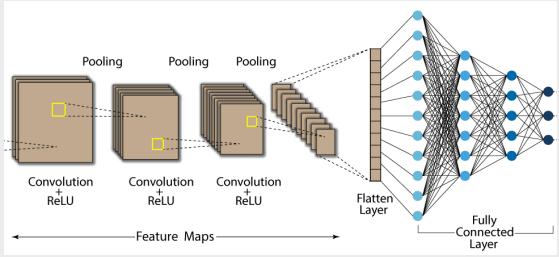
State-of-the-art models have significant resource demands

- Memory, computation, energy

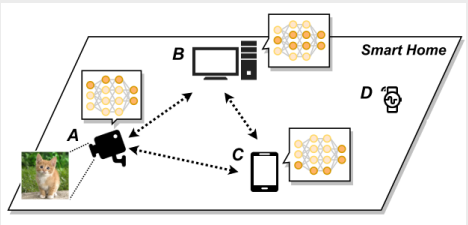
Resource-constrained IoT devices

- Limited energy budget, memory, computation

Towards collaborative Machine Learning



Collaborative Machine Learning



Advantages and opportunities

- Less dependence on distant cloud
- More privacy for sensitive data
- Real-time execution performance
- Reduced bandwidth requirements

Applications

- Smart health (IoMT), Industrial IoT, smart home, multimedia IoT, etc.

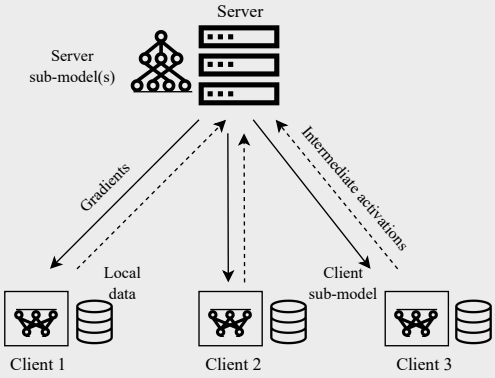
Split Learning

Neural Network = a sequential connection of independently executable layers

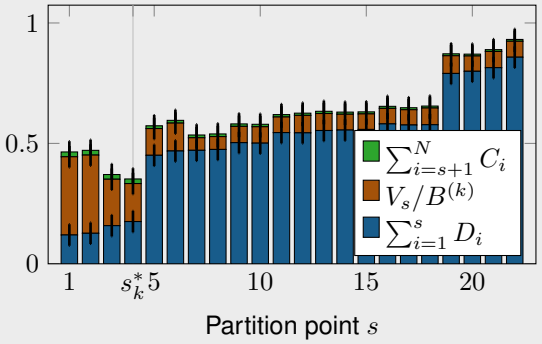
Neural network is split into atleast two sub-networks

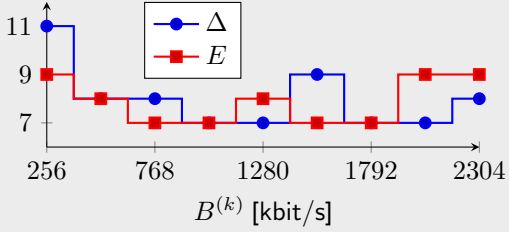
- Client/server side sub-model
- **Reduced resource requirements**
- Improved privacy

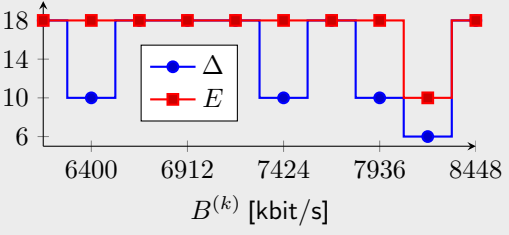
Decouples model training from the need for direct access to the raw data



Where to Split?







Early exit of computation (device)

Challenges for Split Learning

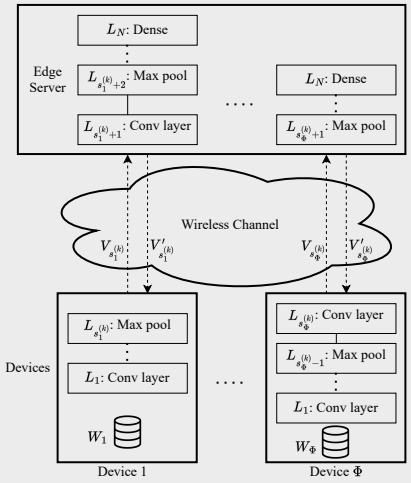
- Stragglers arising from resource heterogeneity of IoT devices that slow down other devices during training
- Variable network throughput and computing resources on devices and server that affect the training time and energy consumption
- Conflicting optimisation objectives: tradeoffs in training time and energy consumption achieved

ARES: Adaptive REsource-aware Split-learning

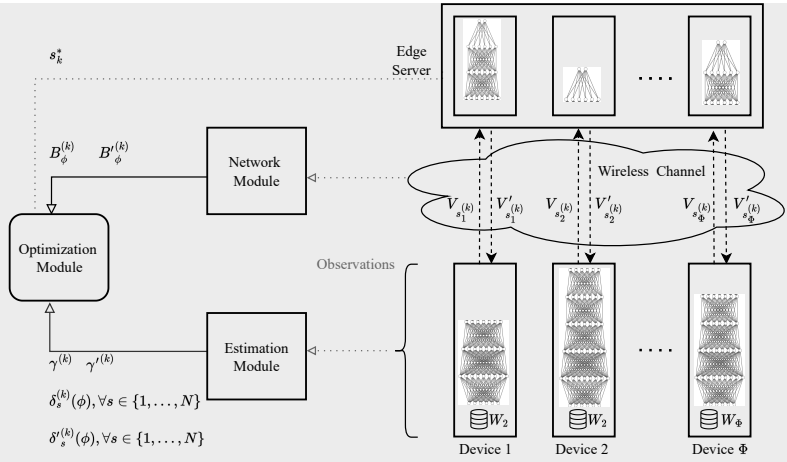
Jointly accelerates model training time and minimizes energy consumption

- Device-targeted (personalised) adaptive model splitting
- Training time $\Delta^{(k)}(\phi)$; Energy $E^{(k)}(\phi)$
- Energy sensitivity parameter $\alpha \in [0, 1]$

$$\underset{s \in \{L\}^{\Phi \times R}}{\text{minimize}} \quad \sum_{k=1}^{R} \left(\alpha \max_{\phi \in \Phi} \{\Delta^{(k)}(\phi)\} + (1-\alpha) \sum_{\phi \in \Phi} E_s^{(k)}(\phi) \right)$$



ARES framework



Testbed Setup

NVIDIA Jetson Nano Developer Kit

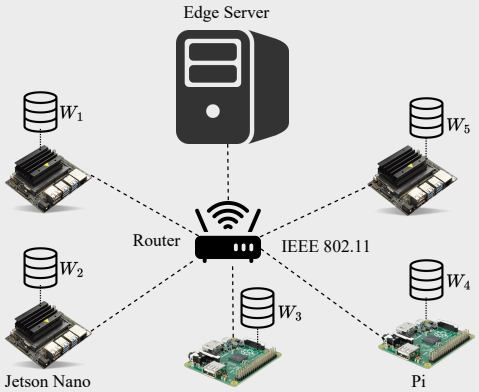
- Power modes: MAXN/5W
- Custom mode: flexible resources

Raspberry Pi

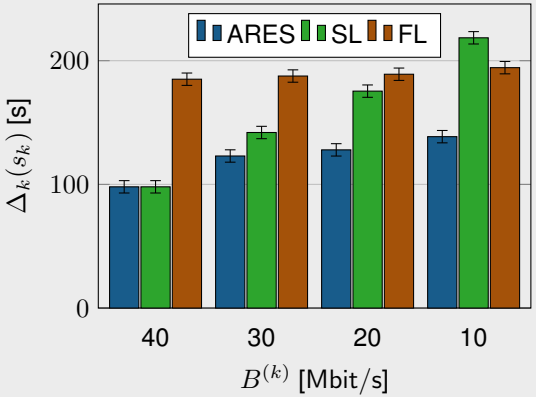
- Model A/3B+

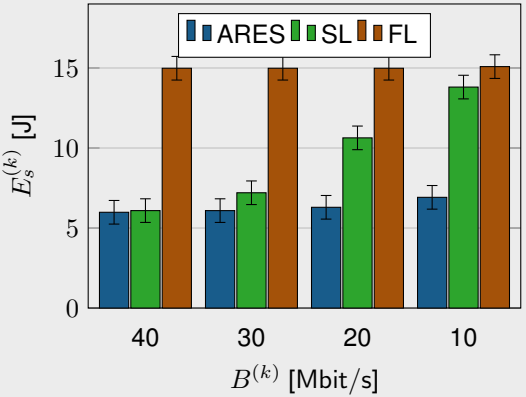
Edge server

Traffic shaping - Linux tc, wondershaper

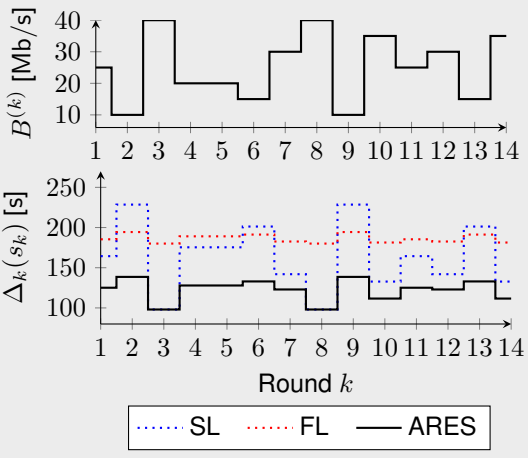


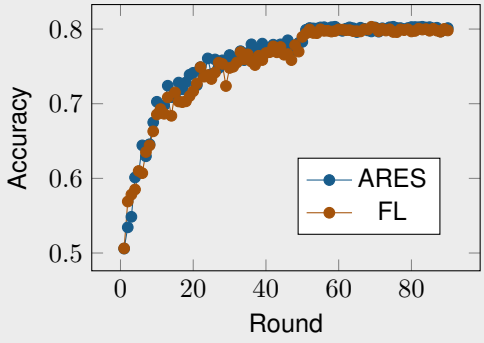
Evaluation Results





Evaluation Results



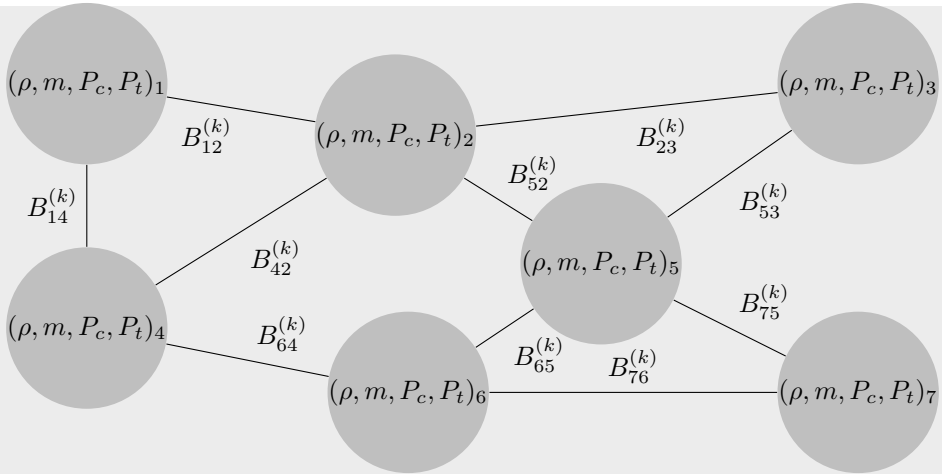


Current work

Fine-grained Distributed Split Neural Networks for Cooperative Deep Learning over Heterogeneous IoT Devices

- Cooperative execution of DNN model in resource-constrained IoT devices
- Considering both layer-based and horizontal DNN partitioning
- Utilising the collective computing power of heterogeneous ubiquitous devices
- Less server dependency - keeping data within local networks

Current work



Conclusion

Distributed Machine Learning in IoT edge environments

- More privacy, low response time, less cloud dependency, etc.

Challenges for Split Learning

- Heterogeneity of IoT devices
- Variable network throughput and computing resources, etc.

ARES: Adaptive REsource-aware Split-learning

- Jointly accelerates model training time and minimizes energy consumption
- Device-targeted (personalized) adaptive model splitting

References

Eric Samikwa, Antonio Di Maio, and Torsten Braun.
Adaptive early exit of computation for energy-efficient and low-latency machine learning over iot networks.
In 2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC), pages 200–206. IEEE, 2022.

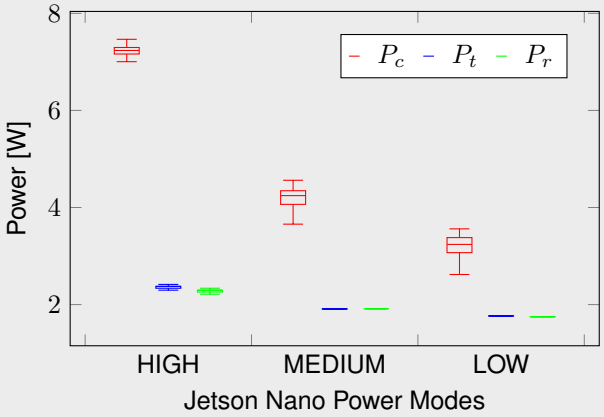
Eric Samikwa, Antonio Di Maio, and Torsten Braun.
Ares: Adaptive resource-aware split learning for internet of things.
Computer Networks, page 109380, 2022.

Thank you for your attention!

Questions

eric.samikwa@unibe.ch

Appendix



Appendix

$$\Delta_s^{(k)}(\phi) = \frac{W_\phi}{\xi}\left(D_s^{(k)}(\phi) + C_s^{(k)} + \frac{V_s}{B_\phi^{(k)}} + \frac{V'_s}{B_\phi^{'(k)}} + \Theta^{(k)}\right)$$

$$E_s^{(k)}(\phi) = \frac{W_\phi}{\xi}\left(P_{\mathsf{c}}(\phi)D_s^{(k)}(\phi) + P_{\mathsf{t}}(\phi)\frac{V_s}{B_\phi^{(k)}} + P_{\mathsf{r}}(\phi)\frac{V'_s}{B_\phi^{'(k)}}\right)$$