Efficient On-Device Fine-Tuning for Edge Intelligence

L. Pes[(1)], S. Stuijk[(1)] and F. Corradi[(1)]

[(1)] Eindhoven University of Technology

Spiking neural networks (SNNs) offer event-driven computation for low-latency, energy-efficient inference on resource-constrained hardware. However, efficient inference alone is insufficient for edge intelligence, as deployed devices must also adapt to evolving real-world data distributions. With billions of edge devices spanning applications from personal electronics to industrial systems, such adaptation must be fast, energy-efficient, and private. As illustrated in Fig. 1a), a keyword spotting model trained on Google Speech Commands (GSC) may require post-deployment adaptation to user-specific and environmental factors, since variations in speaker characteristics, pronunciation, or background noise can cause a spoken "no" command to be misclassified as "yes." Offloading this adaptation to the cloud increases latency and energy consumption while raising privacy concerns, motivating efficient on-device fine-tuning. Meeting these requirements calls for learning rules with spatial and temporal locality, since the non-local credit assignment of backpropagation limits efficient on-device adaptation. Although fully local methods such as DECOLLE [1], ETLP [2], OSTTP [3], and TESS [4] have been proposed, scaling to deeper architectures remains challenging due to auxiliary-matrix overhead, and only TESS has demonstrated applicability beyond shallow networks. To address this issue, we propose Traces Propagation (TP) [5], a fully local learning rule that eliminates auxiliary matrices and scales to deeper architectures such as VGG-9. To showcase its effectiveness in a realistic setting, we evaluate TP for fine-tuning on GSC. As illustrated in Fig. 1b), the deployed model attains only 81.62% accuracy on an excluded user, highlighting the need for user-specific adaptation. Fine-tuning with TP for 3 epochs using limited user data in 1-shot, 5-shot, and all-shot settings, corresponding to 1, 5, or all available recordings per class, improves accuracy by +10 pp, +11.76 pp, and +15.68 pp, respectively, demonstrating effective adaptation.

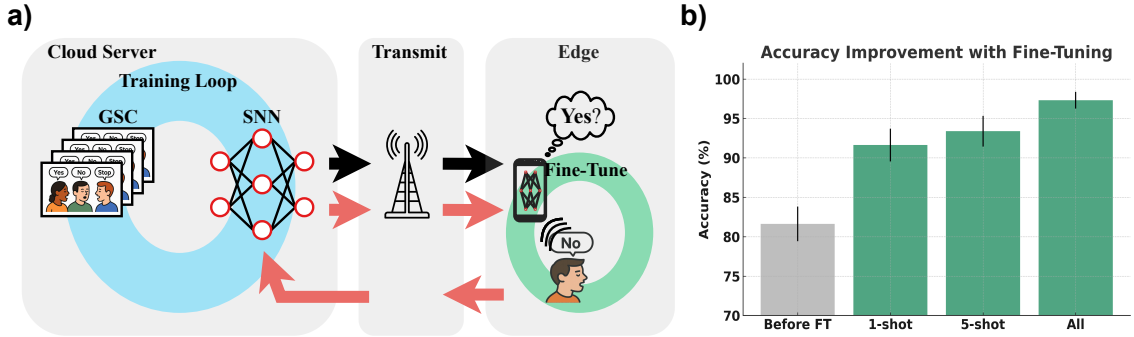

Figure 1: **Efficient GSC On-Device Fine-Tuning: a)** Adaptation to user data requires on-device fine-tuning to avoid cloud dependency (red lines), improving privacy, energy, and latency. **b)** Fine-tuning with TP achieve significant accuracy gains with few local samples.

[1] J. Kaiser et al., Frontiers in Neuroscience. 14, 424, 2020.

[2] F.M. Quintana, Neuromorph. Comput. Eng., 6, 034006, 2024.

[3] T. Ortner et al., IEEE 5th AICAS, 1-5, 2023.

[4] M.P.E. Apolinario et al., Int. Joint Conf. on Neural Networks, 1-9, 2025.

[5] L. Pes et al., Neuromorph. Comput. Eng., 6, 014002, 2026.