This thesis explores the deployment and optimization of Transformer-based models on low-power devices. While Transformers represent the state-of-the-art (SoA) across numerous applications, their high computational cost and memory footprint make them unsuitable for inference and training on resource-constrained hardware. We developed optimized kernels for Transformer models to address these challenges without relying on aggressive quantization. We introduce a structured pruning methodology for deployment that reduces model complexity and size while maintaining controlled accuracy loss. We validate this approach on three mobile-class tiny Transformer architectures: TinyViT, MobileBERT, and TinyLLAMA.
Beyond inference, we investigate the feasibility of on-device learning, specifically through Continual Learning (CL) techniques. In particular, we explore latent replay to reduce memory and computational demands, enabling transformers to be trained and specialized in new data during runtime. Additionally, we extend our study to on-device training for Spiking Neural Networks (SNNs)—a class of models designed for spiked data, which naturally aligns with low-power, resource-constrained devices. We apply Continual Learning to SNNs, an approach that remains relatively unexplored in the literature, demonstrating its potential for efficient adaptation in embedded systems.