Design Space Exploration and Selection of Optimal Model and Hardware Pairs

A Tool-Based Approach for Embedded Machine Learning Deployment

Graduate

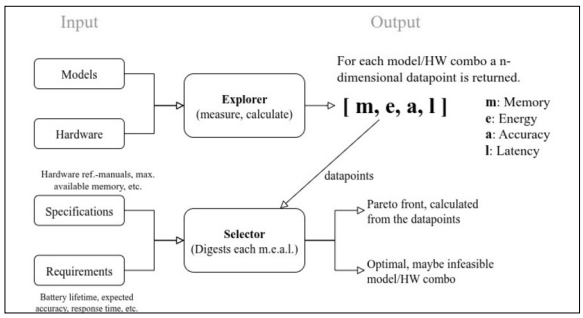
Andri Trottmann

Introduction: Deploying deep learning models on resource-constrained devices requires careful consideration of multiple and often conflicting constraints, including memory, energy consumption, accuracy, and latency. Existing TinyML deployment frameworks focus primarily on individual metrics or single-model optimization, leaving developers without systematic support for multi-constraint decision-making.

Approach: In this work, we present TinyML Design Space Explorer (TDSE), a framework that simplifies AI model deployment across diverse embedded devices, improving flexibility and accessibility for edge AI. TDSE evaluates candidate model-hardware pairs across four key metrics and provides data-driven recommendations for optimal deployment under user-specified constraints. It leverages microTVM for model compilation and integrates both measured and estimated metrics to generate actionable deployment decisions. The framework implements a two-stage evaluation pipeline: (1) direct measurement or estimation of competing metrics for model-hardware pairs, and (2) selection of solutions according to hardware or user-specified constraints. A web-based interface enables iterative exploration and configuration of new models and target platforms. Empirical evaluation on ARM Cortex-M4 and M7 microcontrollers demonstrates viable deployments of neural networks varying in architectural complexity.

Conclusion: Deployed models maintain Top-1 accuracy between 70% and 87%, and latency estimates exhibit a mean relative error of 9.58% over measured latency, while energy estimates show an approximate error of 20% over actual energy measurements. Our findings underscore the importance of integrated tool support for real-world embedded AI applications and highlight the potential for wider application of systematic trade-off exploration within data-driven systems engineering.

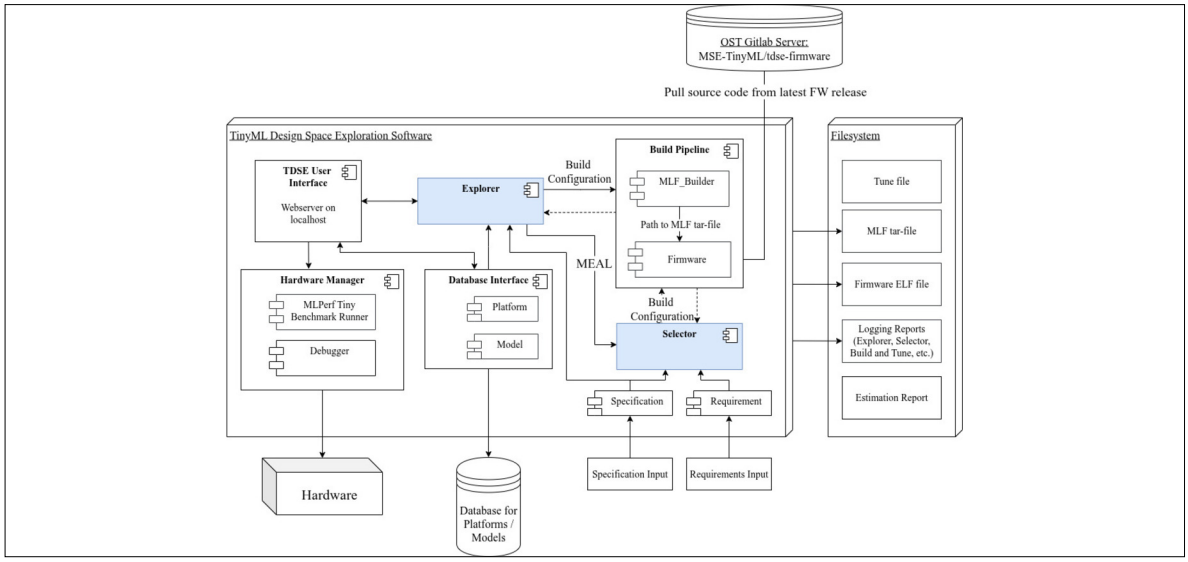
The core exploration and selection pipeline; searching for solution candidates, while concurrently applying constraints.
Own presentment



Found model-hardware pairs, which achieve at maximum 50% memory ratio and at least 71% inference accuracy.
Own presentment

Platform	Model	Memory Ratio Flash / RAM	Accuracy
Nucleo-L4R5ZI	ResNet, float32	0.161 / 0.311	0.87
	ResNet, int8	0.088 / 0.206	0.855
	LeNet-5, float32	0.169 / 0.048	0.715
	LeNet-5, int8	0.047 / 0.054	0.715
	MicroCNN, float32	0.087 / 0.227	0.75
	MicroCNN, int8	0.035 / 0.056	0.74
Nucleo-F746ZG	ResNet, int8	0.178 / 0.412	0.855
	LeNet-5, float32	0.338 / 0.096	0.715
	LeNet-5, int8	0.095 / 0.108	0.715
	MicroCNN, float32	0.173 / 0.454	0.75
	MicroCNN, int8	0.07 / 0.113	0.74

Component diagram of the TDSE software, integrating external dependencies such as models and the target platforms.
Own presentment



Advisors
Prof. Dr. Mitra Purandare, Prof. Dr. Andreas Breitenmoser

Co-Examiner
Sebastian Stenzel, Sonova

Subject Area
Computer Science