SimQ-NAS: Simultaneous Quantization Policy and Neural Architecture Search

Sharath Nittur Sridhar* [1], Maciej Szankin* [1], Fang Chen [2], Sairam Sundaresan [1], Anthony Sarah [1]

(*=Equal contribution authors)

[1] Intel Labs, San Diego, USA

[2] University of California, Merced

AAAI Association for the Advancement of Artificial Intelligence

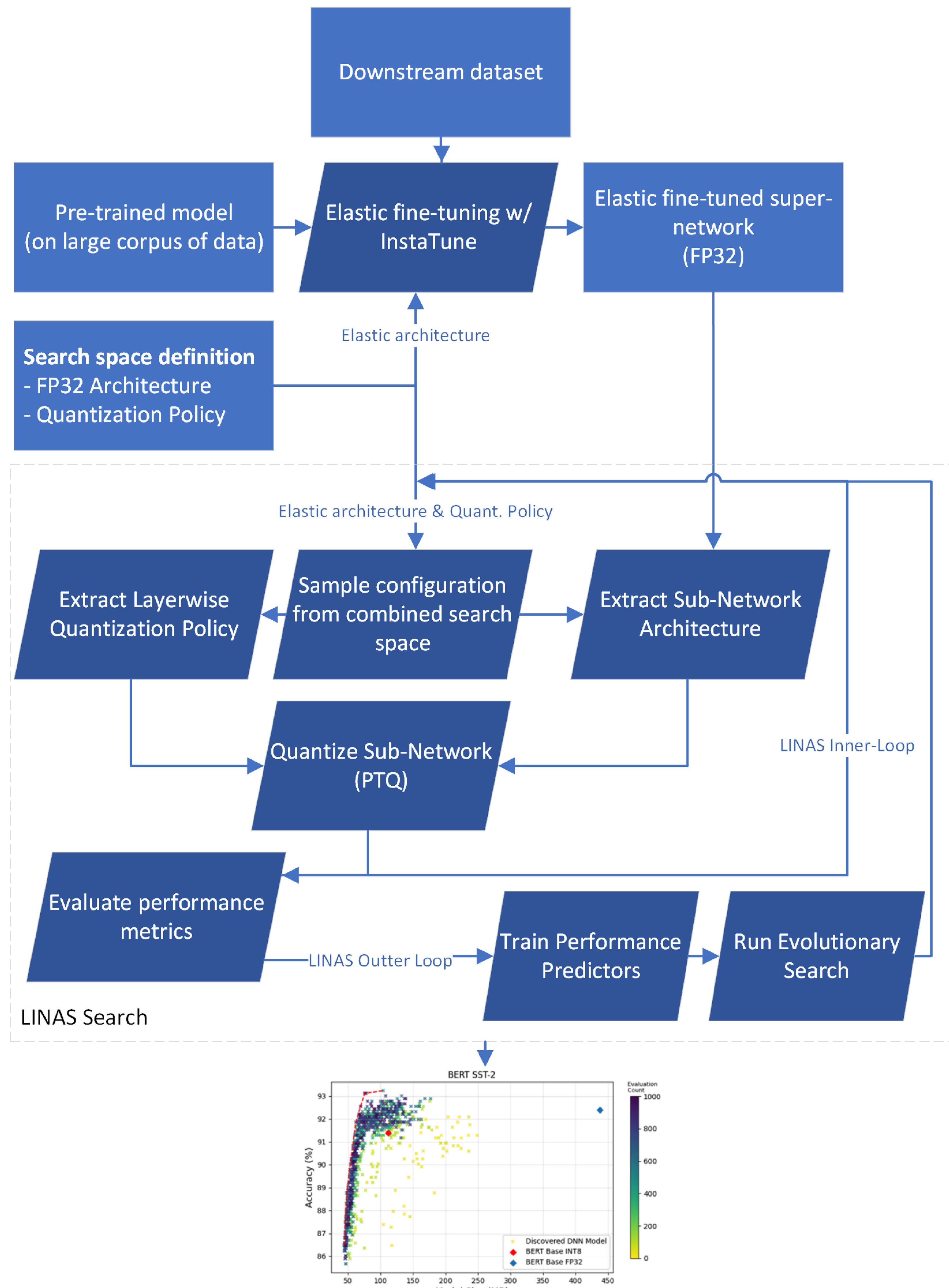
Edge Intelligence Workshop

Introduction & Motivation

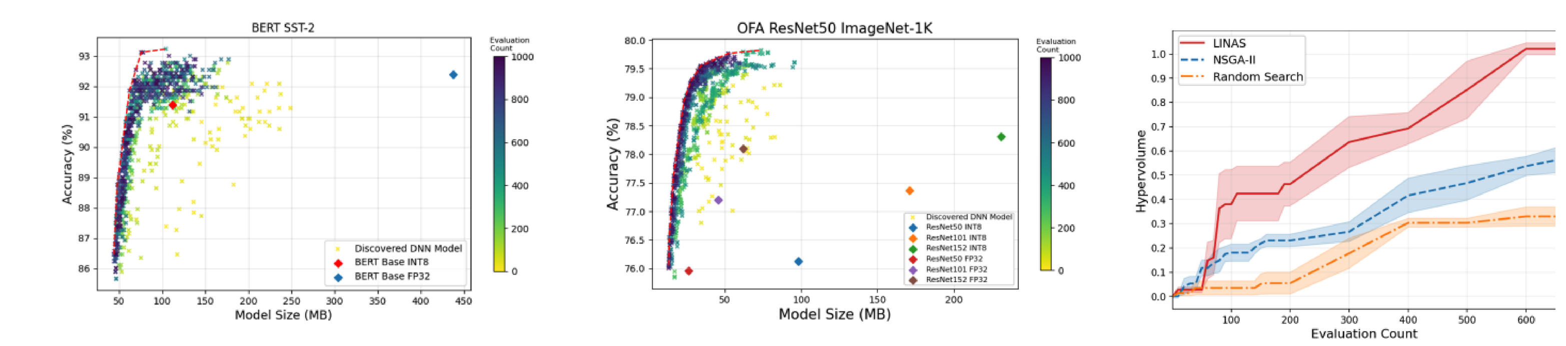
- **Rising Complexity in Neural Networks:** As neural networks grow in complexity, optimizing them for diverse hardware platforms becomes increasingly challenging.
- **Need for Hardware-Agnostic Solutions:** Traditional Neural Architecture Search (NAS) methods focus on finding efficient architectures without considering hardware constraints, leading to suboptimal performance on specific platforms.
- **Quantization as a Key Optimization:** Quantization reduces model size and latency by approximating high-precision weights with lower precision, but it often requires careful tuning to maintain accuracy.
- **Gap in Quantization Policy Search:** Existing approaches to quantization policy search are well-established for CNNs but less so for transformer-based models, including foundation models.
- **Objective:** To develop a method that simultaneously optimizes for neural network architecture and quantization policy, catering to the specific needs of various hardware platforms without compromising on model performance.

Proposed Solution: SimQ-NAS

- **Unified Framework**: SimQ-NAS integrates the search for optimal neural network architectures with quantization policies into a single, cohesive framework.
- **Multi-Objective Optimization**: Utilizes multi-objective search algorithms to navigate the trade-offs between model accuracy, size, and latency effectively.
- **Lightly Trained Predictors**: Employs predictors that are trained with minimal computational overhead to estimate the performance of different architecture-quantization combinations.
- **Broad Applicability**: Demonstrates effectiveness across a range of architectures, including uni-modal (ViT, BERT), multi-modal (BEiT-3) transformers, and CNNs (ResNet).
- **Significant Performance Gains:** Achieves up to 4.80x improvement in latency and 3.44x reduction in model size for certain networks, without degrading accuracy compared to fully quantized INT8 baselines.
- **Adaptability:** SimQ-NAS's flexible approach allows for adaptation to emerging neural network models and evolving hardware specifications.



SimQ-NAS: Experimental Results

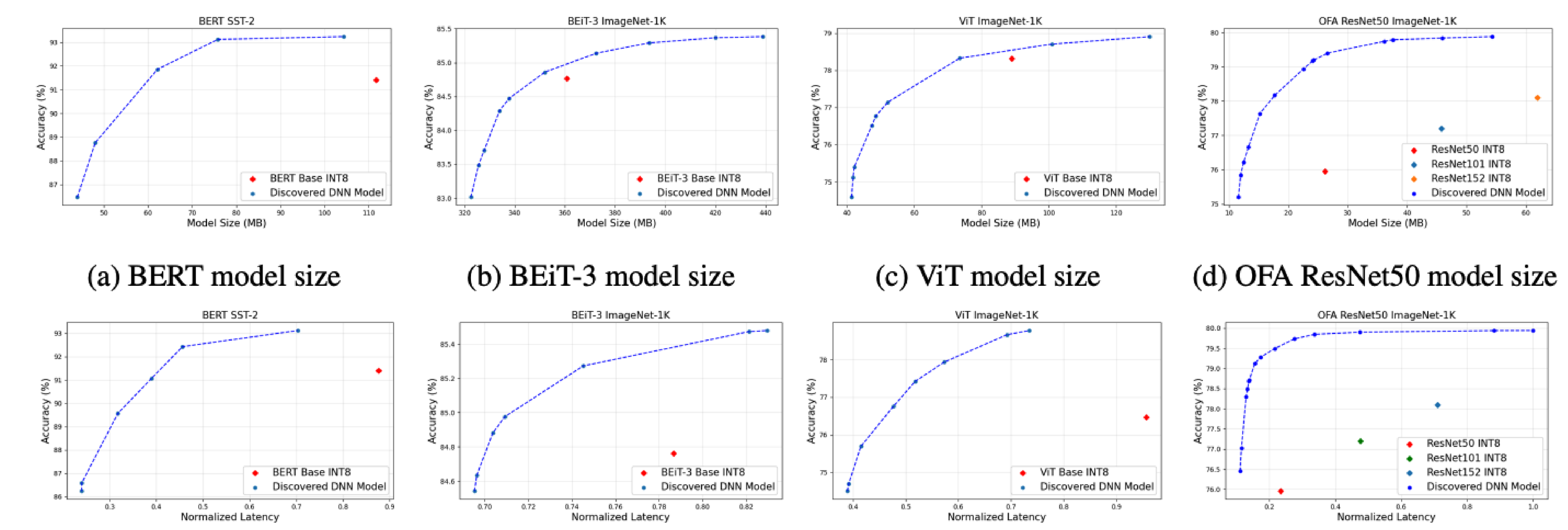


(a) BERT SST-2 model size (b) OFA ResNet50 model size (c) BERT SST-2 hypervolume

Search progression using LINAS [1] for BERT Base (a) and OFA ResNet50 (b) super-networks.

➢ A clear progression towards the near-optimal Pareto front can be seen with an increase in evaluation count.

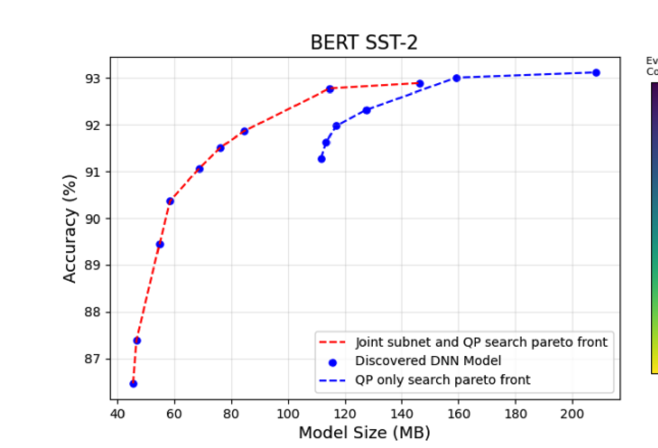
➢ A significant improvement in model size is observed when compared to the FP32 baselines.



(a) BERT model size (b) BEiT-3 model size (c) ViT model size (d) OFA ResNet50 model size

(e) BERT latency (f) BEiT-3 latency (g) ViT latency (h) OFA ResNet50 latency

Joint architecture and quantization policy search Pareto fronts on BERT Base, BEiT-3 Base, ViT Base and OFA ResNet50 super-networks using model size and normalized latency as the search objectives.



Comparison of quantization policy only search and joint architecture and quantization policy search with accuracy and model size on BERT Base fine-tuned with InstaTune [2] on SST2 dataset.

Performing joint search using our combined InstaTune+SimQ approach significantly improves the model size for similar accuracy.

Summary

➢ Demonstrated the use of multi-objective search algorithms with lightly trained predictors for efficient search of sub-network architecture and quantization policy.

➢ Comprehensive Applicability: Demonstrated effectiveness across a variety of architectures, including ViT, BERT for uni-modal, BEiT-3 for multi-modal, and ResNet for convolutional models.

➢ Significant Performance Gains: Achieved up to 4.80x improvement in latency and 3.44x reduction in model size, maintaining accuracy compared to fully quantized INT8 baselines.

References

[1] D. Cummings et al., "A hardware-aware framework for accelerating neural architecture search across modalities", AutoML Workshop 2022.

[2] Sridhar, Sharath Nittur, et al. "InstaTune: Instantaneous Neural Architecture Search During Fine-Tuning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.