ActNAS : Generating Efficient YOLO Models using Activation NAS

Sudhakar Sah, Ravish Kumar, Darshan C. Ganji, Ehsan Saboori

Deeplite, Toronto, Canada

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

Motivation: Standard YOLO models use a single activation function, limiting optimization for specific hardware. Mixed activations can improve latency and memory efficiency.

Contribution:

- HW-aware optimization for YOLO models with layer-specific activations.
- ActNAS models show minimal mAP drop and significant latency improvement over baseline models on various edge devices.
- Comprehensive analysis of ReLU, SiLU, and HardSwish activation functions for improved performance.

Motivation

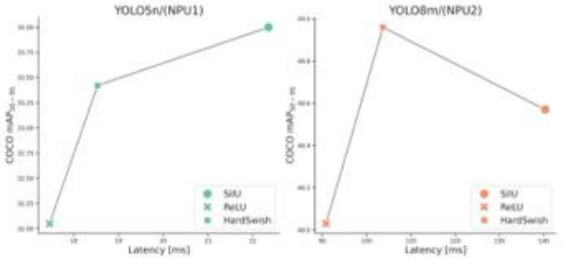

Figure 1: Latency-mAP plot of YOLO5n and YOLO8m for two different NPUs

- Choosing the right activation function is crucial for optimizing model performance, where faster functions like ReLU reduce latency but may result in slightly lower accuracy compared to slower functions like SiLU.
- Hardware aware model for one device may not necessarily perform well on other type of device.
- Faster activation functions can help reduce inference time, but the challenge lies in maintaining an acceptable level of accuracy for the model's tasks.

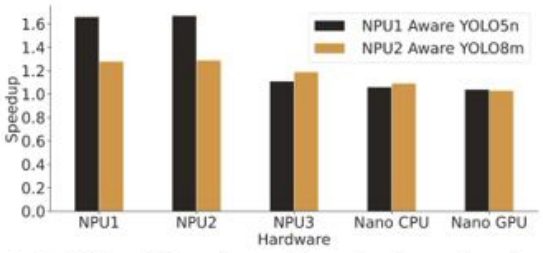

Figure 2: ActNAS model speedup on different hardware where the model is optimized for NPU1 (NPU1 and NPU2 have similar configuration)

Activation NAS Methodology

Search Space

Replacing activations one layer at a time, with multiple candidates per layer, allows for a detailed exploration of different activation functions impact on model performance.

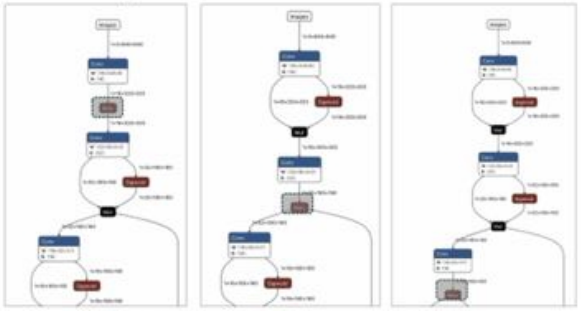

Figure 3: Replace one activation at a time

Model Benchmarking

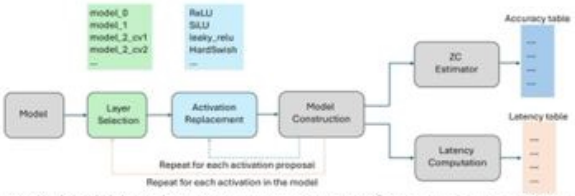

Figure 4: Model benchmarking using training free estimators and on device inference

- ActNAS pipeline integrates model construction, zero-cost estimations, and latency evaluation, streamlining activation selection to deliver optimal performance across diverse edge devices.
- ActNAS benchmarks various activation functions (ReLU, SiLU, Hardswish) and custom proposals for each model layer, ensuring accuracy and latency improvements tailored to specific hardware.

Activation NAS Process

Figure 5: Combines precomputed accuracy and latency tables using Integer Linear Programming (ILP) to efficiently identify the best activation configurations.

Experimental Results

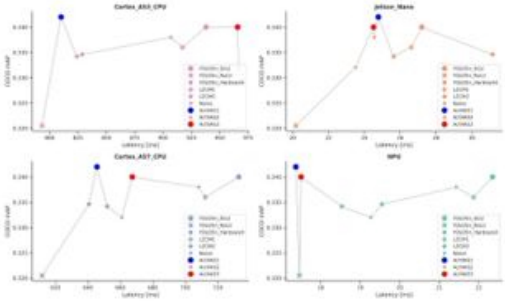

Figure 6:Performance of YOLO5n reference models and ActNAS models on different edge devices

- ActNAS adapts to different hardware, achieving low latency and efficient mAP-latency trade-offs on Cortex CPUs and NPUs.
- ActNAS consistently achieves higher mAP with competitive latency across diverse hardware platforms, outperforming YOLO5n variants.

TABLE 1: Performance of ActNAS models compared to baseline models on GPU and CPU

Model	mAP	Jetson Nano		Cortex A-53	
		Hswish	SiLU	Hswish	SiLU
YOLO5n_SiLU	**0.3400**	-	-	-	-
YOLO5n_ReLU	**0.3205**	-	-	-	-
YOLO5n_Hardswish	**0.3342**	-	-	-	-
LZCM1(SiLU/ReLU)	0.3360	-21.54%	2.17%	-0.58%	2.23%
LZCM2(SilU/Hswish)	0.3346	-21.54%	-14.52%	-0.58%	11.64%
Naive(ReLU/SiLU)	0.3380	4.21%	9.74%	-10.06%	3.32%
ActNAS1(Mixed)	**0.3420**	**3.32%**	**8.90%**	**1.69%**	**13.64%**
ActNAS2(Mixed)	0.3320	**8.31%**	**13.60%**	-17.34%	-3.08%
ActNAS3(Mixed)	**0.3400**	**4.37%**	**9.89%**	-17.20%	-2.96%

Summary

- ActNAS introduces hardware-aware neural architecture search for optimizing YOLO models with mixed activation functions.
- Achieves up to 1.67× faster inference and 64.15% memory savings with negligible accuracy loss.
- Utilizes zero-cost estimators to efficiently configure layer-specific activations for edge devices.

References:

[1] Yevgeniy Bodyanskiy and Serhii Kostiuk. Adaptive hybrid activation function for deep neural networks.

[2] Mohamed S Abdelfattah, Abhinav Mehrotra, Łukasz Dudziak, and Nicholas D Lane. Zero-cost proxies for lightweight nas.