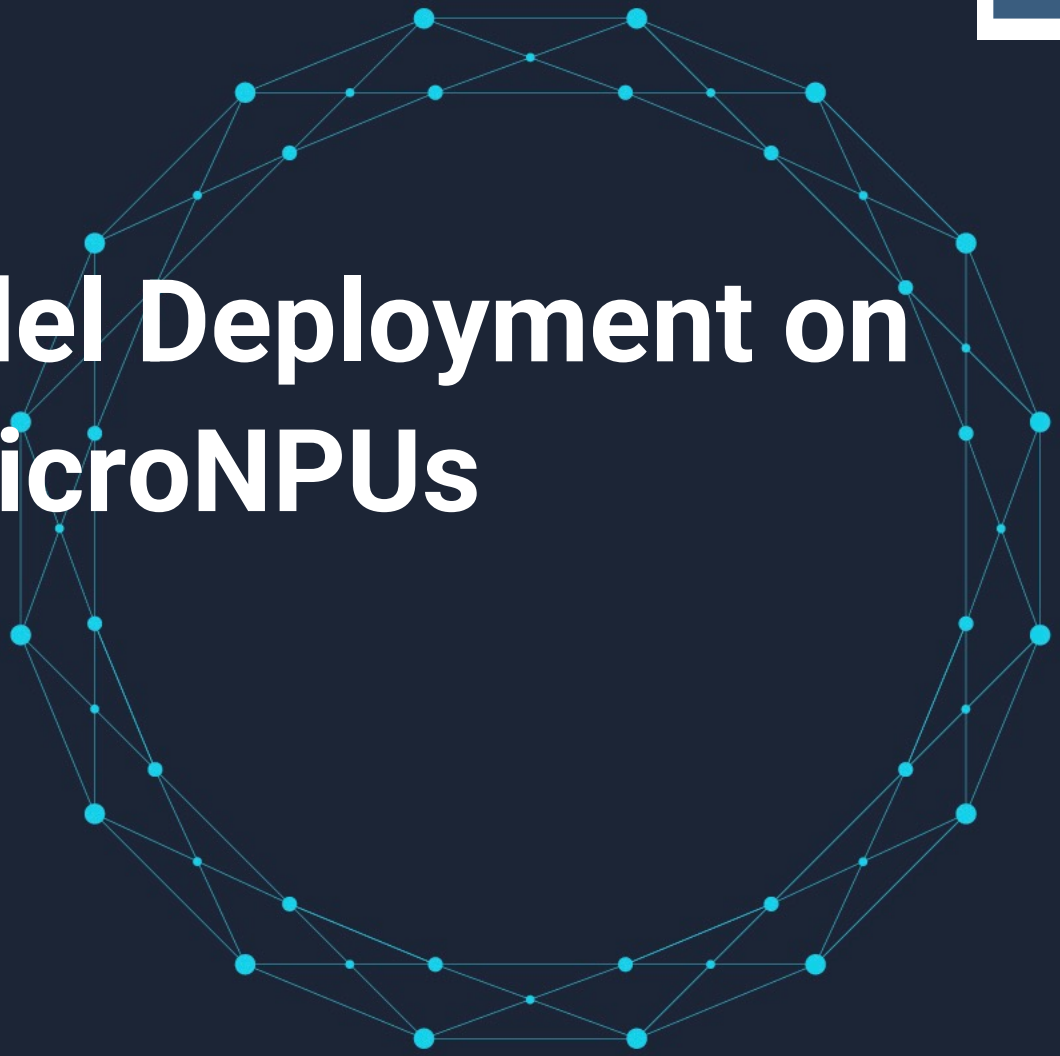
tinyML® Foundation

Enabling Ultra-low Power Machine Learning at the Edge

tinyML Summit April 22 - 24, 2024

www.tinyML.org

Nota AI



Transformer-Based Model Deployment on Edge Devices through MicroNPUs Operator Converter

Shinkook Choi | Core Research Lead | April 23, 2024

Nota AI®

Nota AI®

Corporate Journey: Pioneering the AI Industry

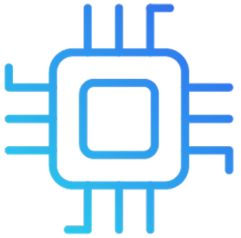
- Nota AI®'s journey showcases growth, partnerships, events, and awards since inception.

Investment Funding

U$ 22.7M

Global Customers

101+



Supported Devices

30+

IP Rights

69+

Nota AI®

Global Partnerships: Collaborating with Tech Leaders

- Strategic alliances with industry leaders drive Nota AI®'s global expansion.

Partners

Clients

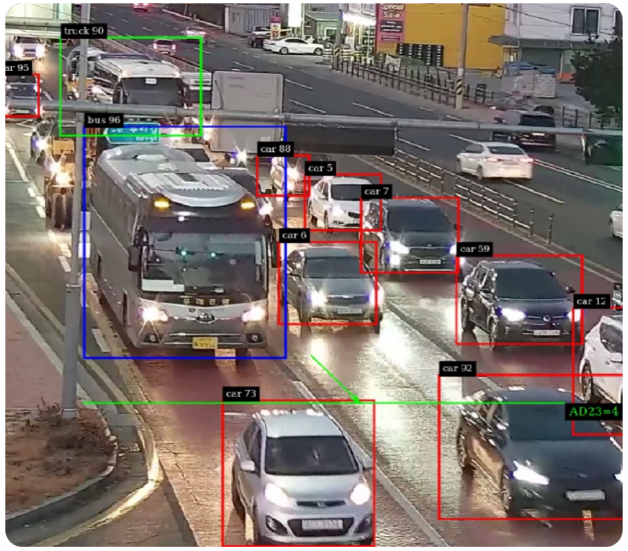
Nota AI®

Platform & Edge Solutions: Elevating Excellence

- Continuously advancing technologically through the utilization of NetsPresso®, Nota ITS, and Nota DMS.

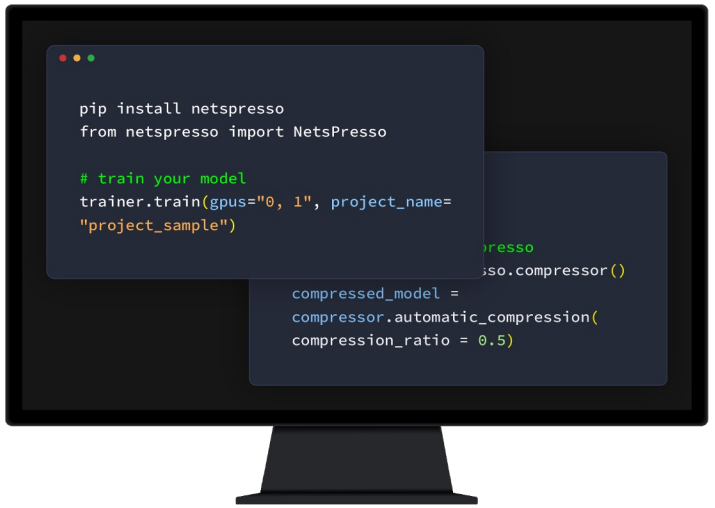
Solution

Nota ITS



Platform

NetsPresso®



Solution

Nota DMS

Table of Contents

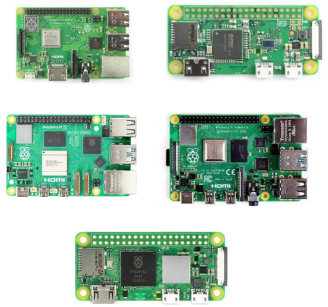
Problem Statement

Problem statement

Various device support

- Nota AI® supports a diverse range of devices such as CPUs, GPUs, NPUs, MCUs, etc.

Raspberry Pi



NVIDIA

arm AI

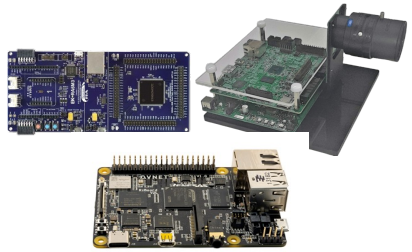
ALIF SEMICONDUCTOR

NXP

intel

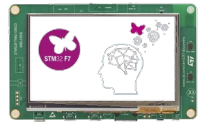
intel XEON

RENESAS



SONY

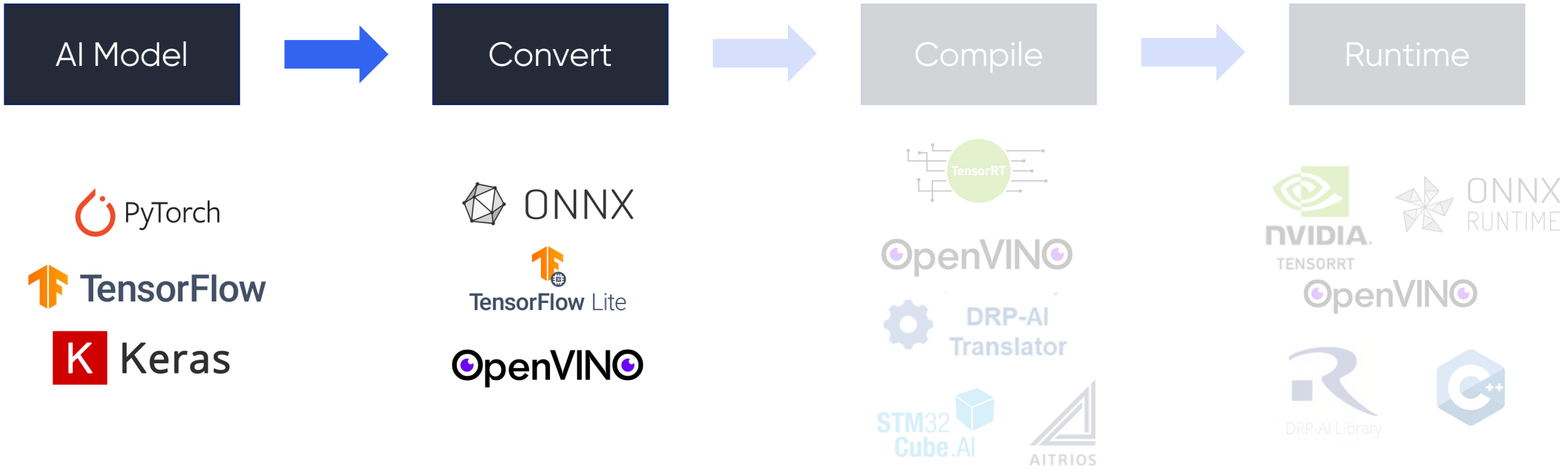
life.augmented



Problem statement

Runtime-specific IR Converter

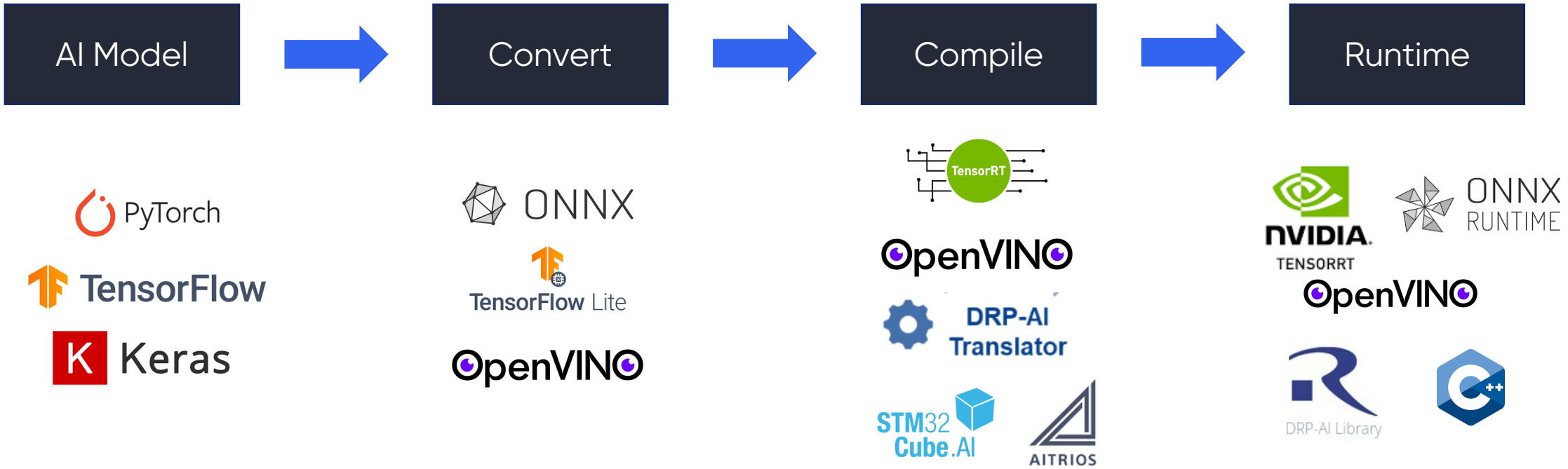
- To deploy AI models onto various devices, converting and compiling the AI model from deep learning frameworks into a runtime-specific Intermediate Representation(IR) is necessary.



Problem statement

Limitation of compiler

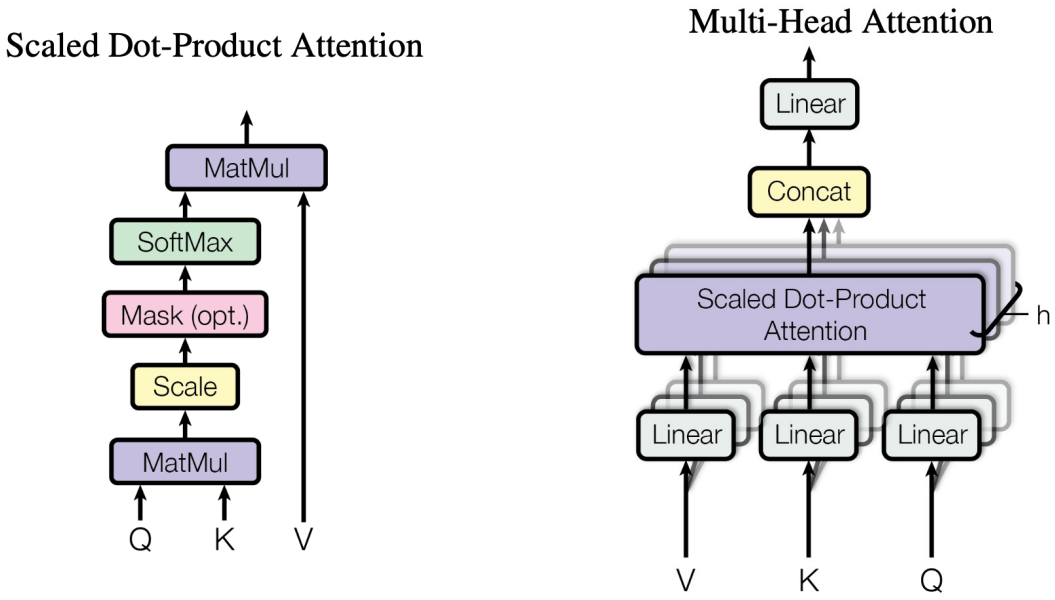
- Each compiler can only support the operators that it has implemented.
- If an AI model contains unsupported operators, compilation may fail, or the operations may be executed on the CPU or MCU, resulting in reduced performance.

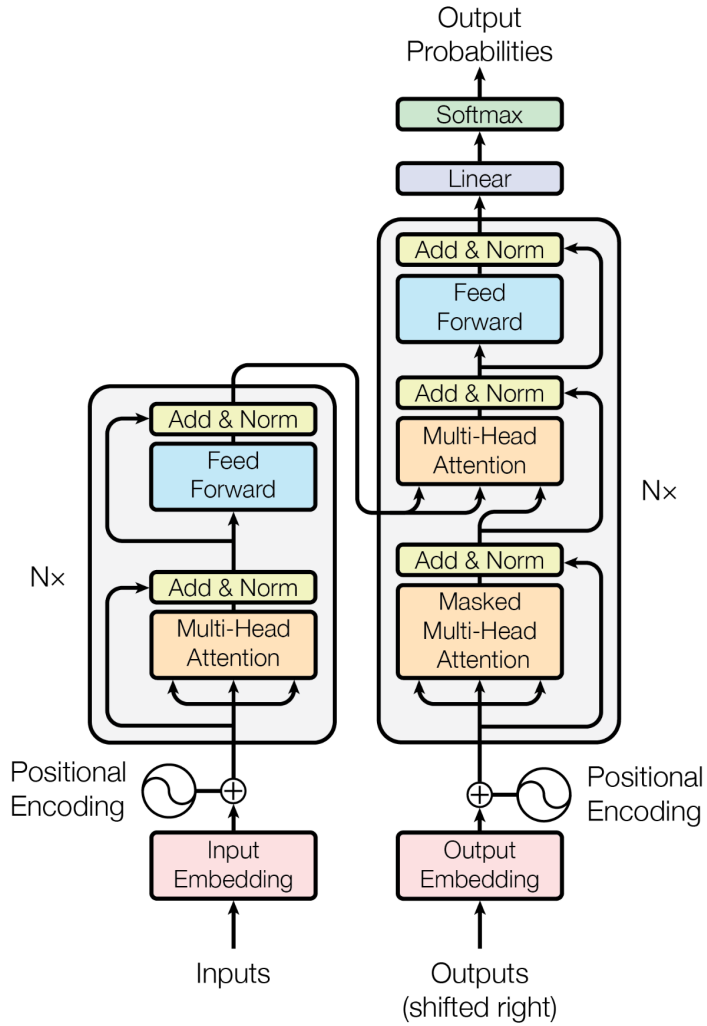


Problem statement

Transformer

- Transformers have revolutionized deep learning, dominating natural language processing (NLP).
- Recent advancements have shown their remarkable performance in diverse domains, marking a paradigm shift in AI.



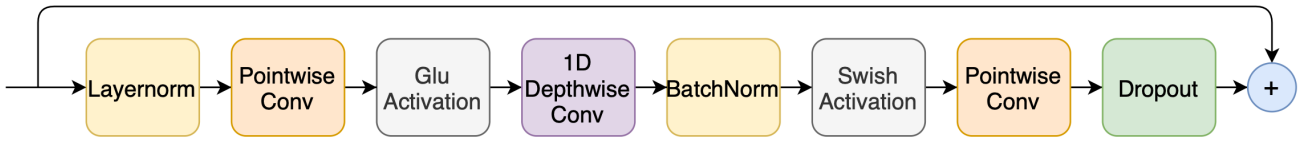


Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

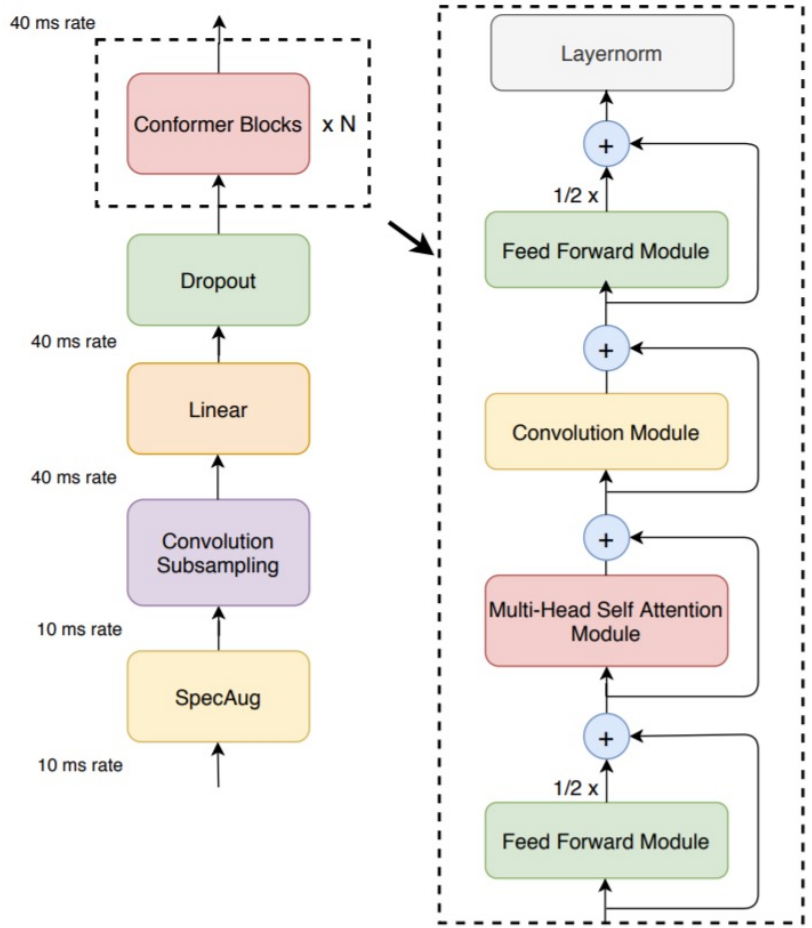
Problem statement

Conformer

- Conformer enhances Transformers for speech and audio tasks, providing effective solutions in voice recognition and audio processing.
- Conformer combines CNNs with self-attention to improve speech-to-text accuracy, harnessing local and global data patterns.



Convolution Module



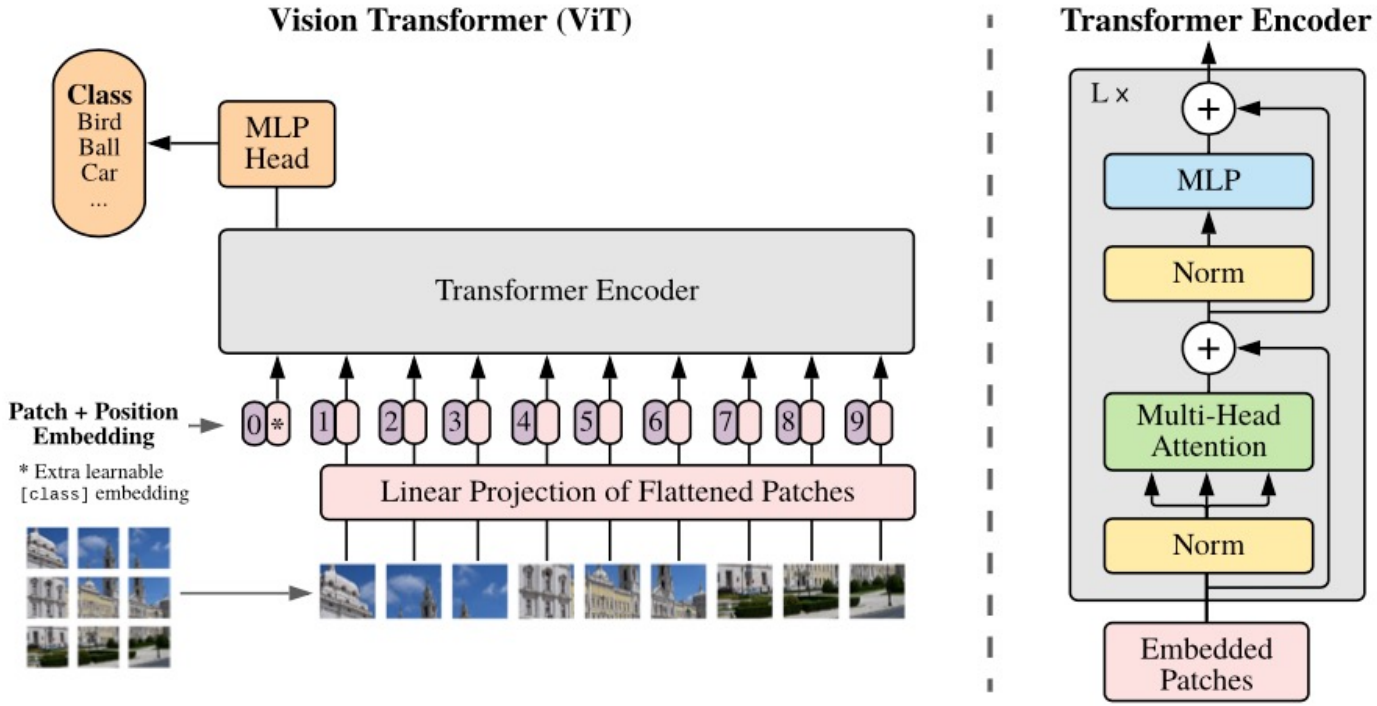
Conformer

  Gulati, Anmol, et al. "Conformer: Convolution-augmented transformer for speech recognition." *arXiv preprint arXiv:2005.08100* (2020).

Problem statement

Vision Transformer (ViT)

- Vision Transformer (ViT) expands Transformer's capabilities to computer vision tasks, achieving remarkable results in image classification and object detection.
- ViT processes images by dividing them into patches.



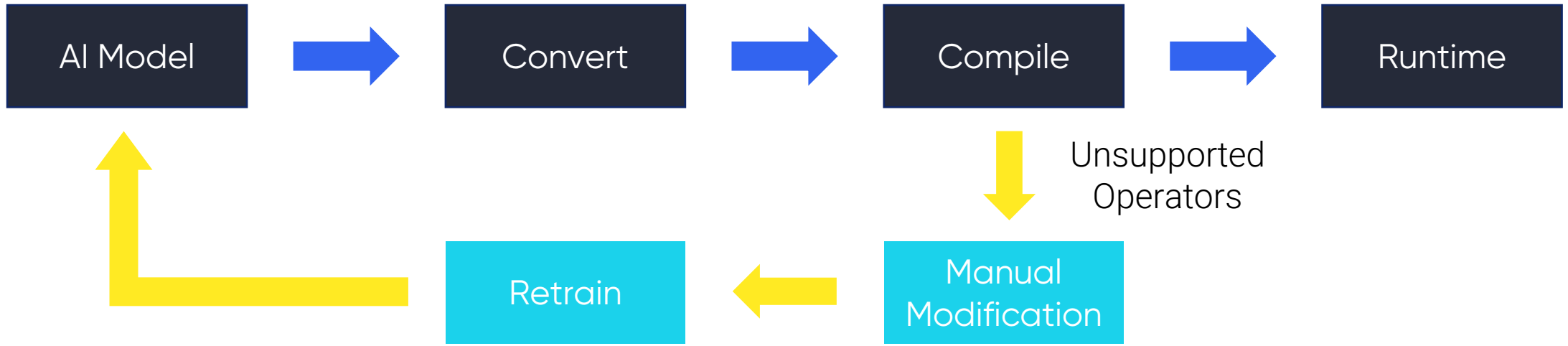
  Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

Operator Converter

Operator Converter

Challenges of Manual Model Modification

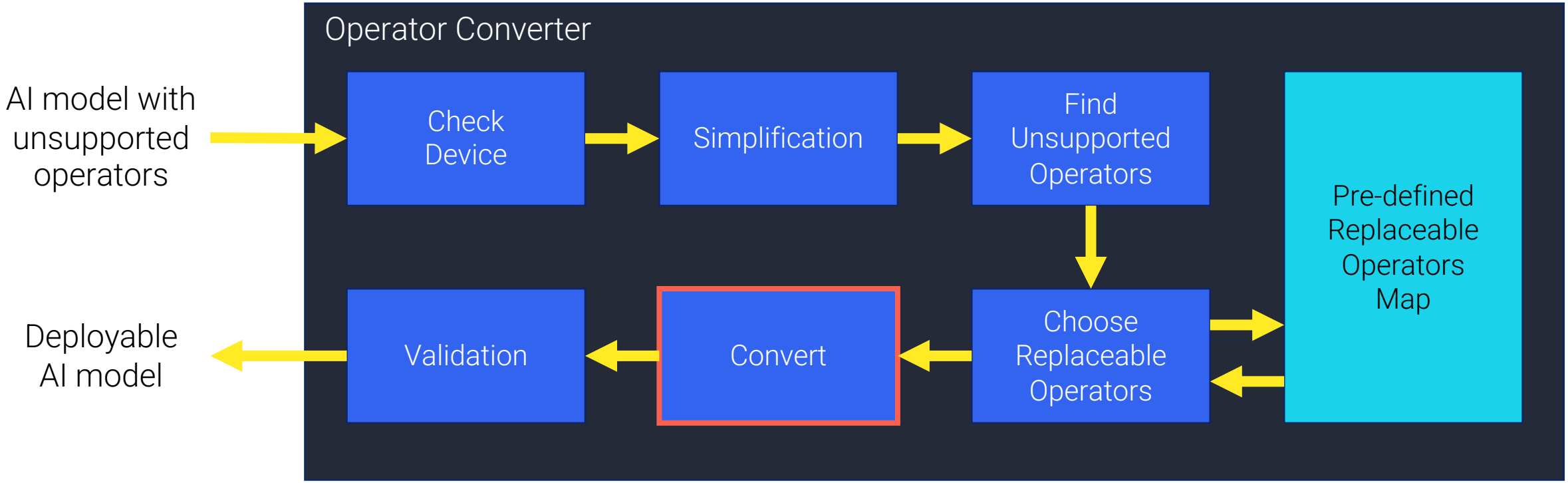
- AI models with unsupported operators require replacement or removal for deployment.
- Manual model modification in deep learning frameworks is cumbersome.



Operator Converter

Operator Converter

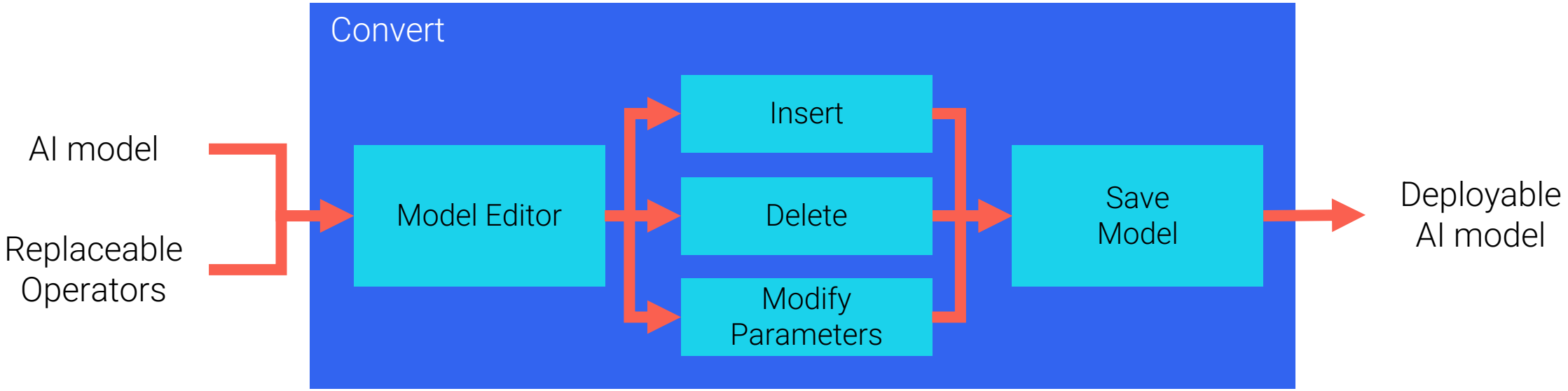
- The Operator Converter automates unsupported operation replacement, easing deployment.



Operator Converter

Model Editor

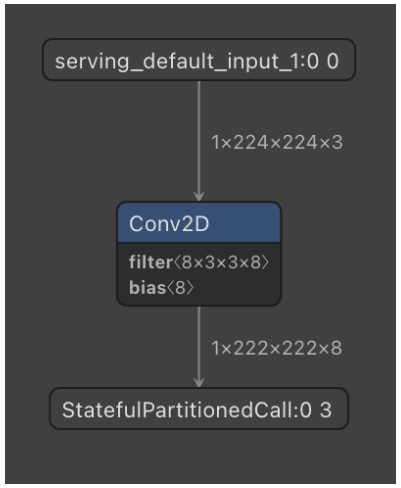
- The convert functionality within the operator converter utilizes a model parser equipped with insert, delete, and modify functions to alter the graph.



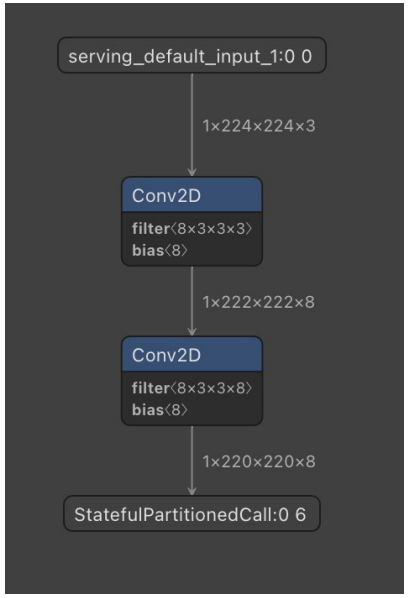
Operator Converter

Model Editor

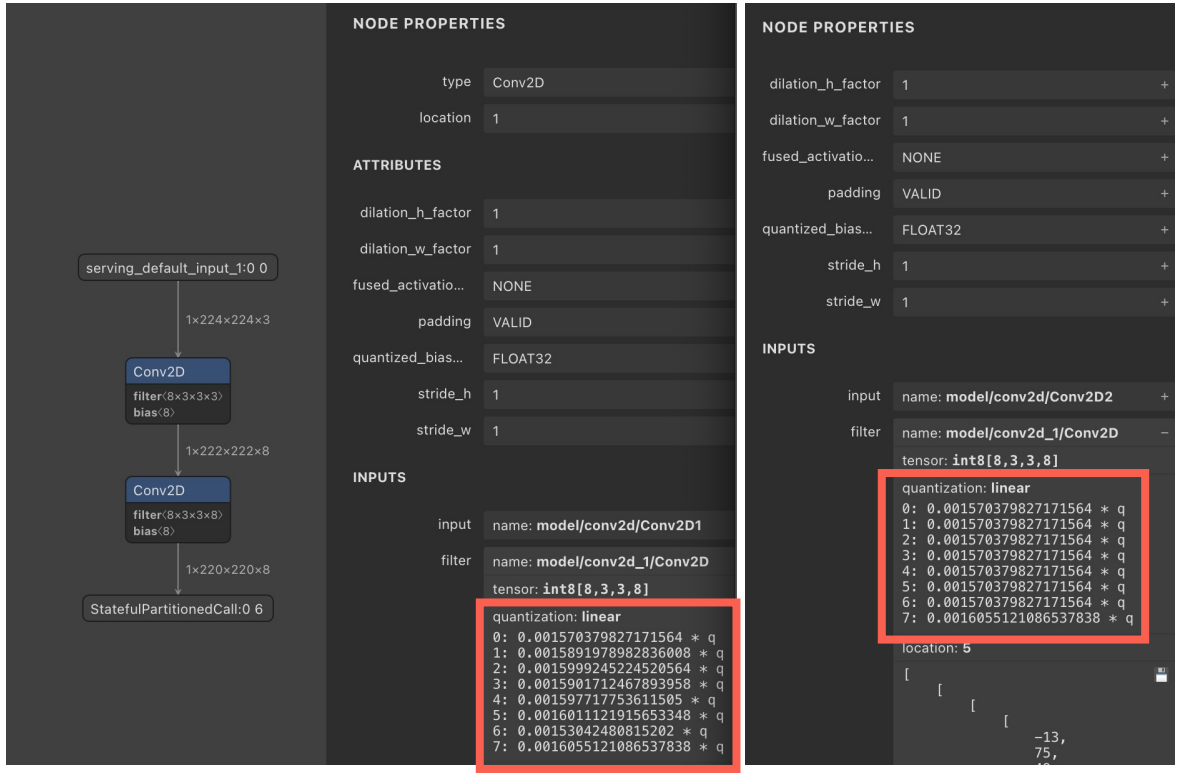
- Example



Subtract layer



Original



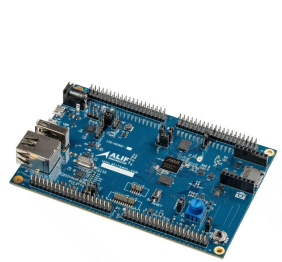
Modify quantization params

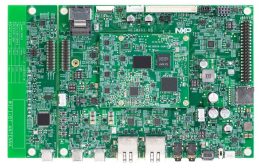
Results

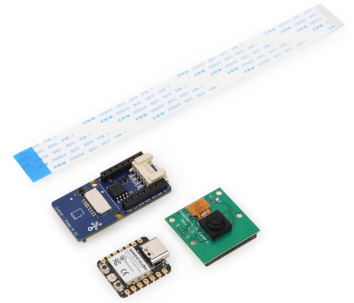
Results

Target Device

- Alif Ensemble E7 DevKit Gen2
  - Arm Cortex-M55 + Arm Ethos-U55 NPU
- NXP i.MX 93 Evaluation Kit
  - Arm Cortex-M33 + Arm Ethos-U65 NPU
- Seeed Studio Grove Vision AI V2
  - Arm Cortex-M55 + Arm Ethos-U55 NPU







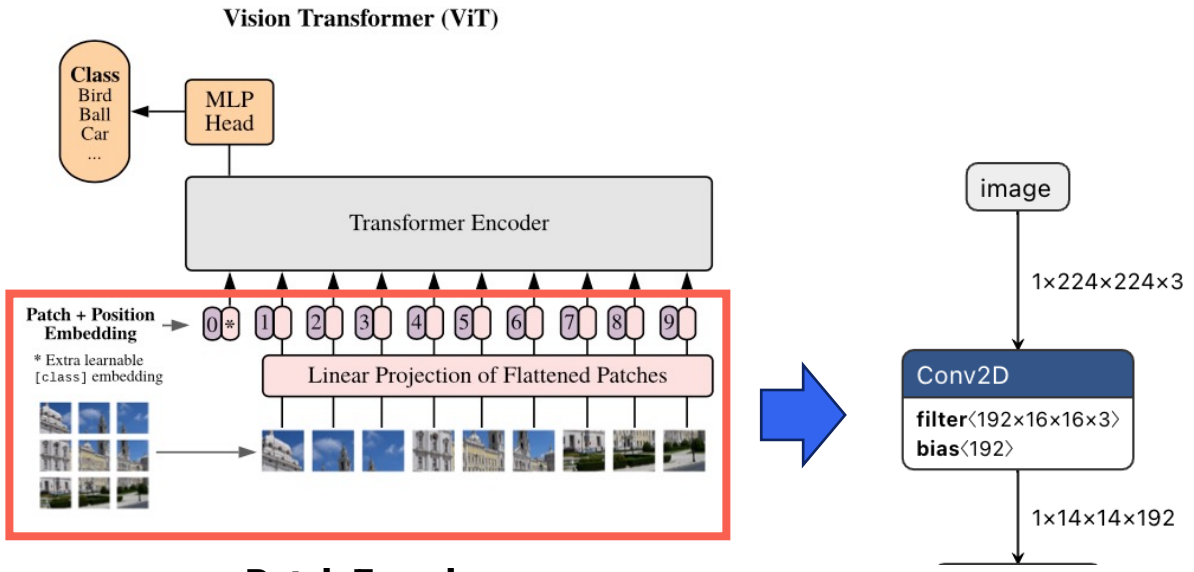
- Convert the TensorFlow Lite model with full INT8 quantization using the Vela compiler

  https://review.mlplatform.org/plugins/gitiles/ml/ethos-u/ethos-u-vela/

Results

Constraints with Vela Compiler

- The patch encoder in ViT uses a convolution with kernel size and stride matching the patch size.
- This particular convolution configuration is not currently supported by the Vela compiler.



Patch Encoder

TFLite CONV_2D Constraints

This is a list of constraints that the CONV_2D operator must satisfy in order to be scheduled on the NPU.

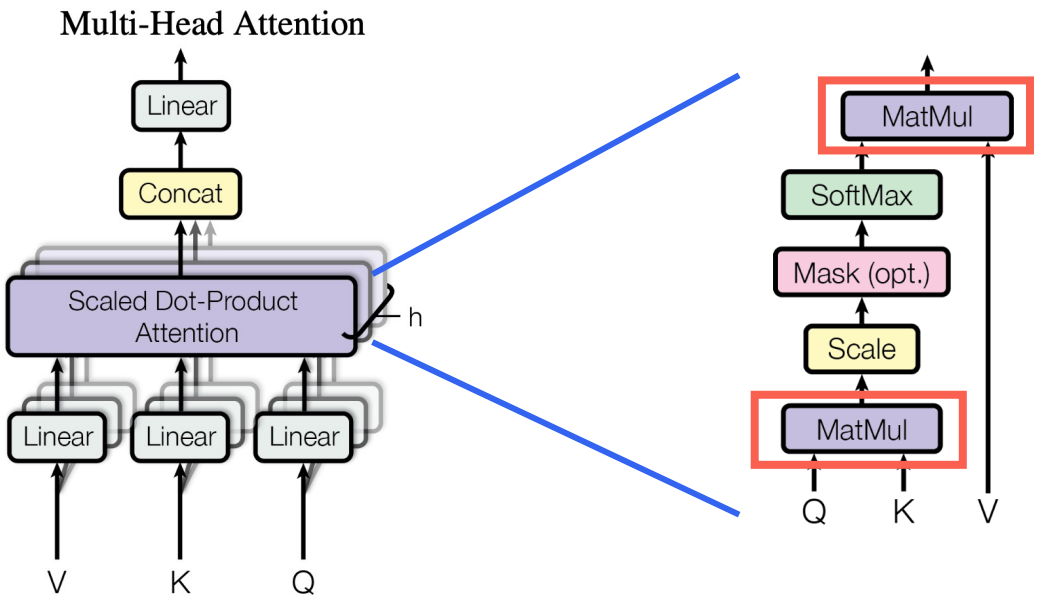
- Stride values for both width and height must be integer types
- IFM depth must be a whole multiple of the filter kernel depth
- Number of filter kernels must be equally divisible by the number of convolution groups
- Dilation factor values for both width and height must be integer types
- Strides must fulfil the following criteria:
  - Stride h must be between 1 and 3 when ofm height is greater than 1
  - Stride w must be between 1 and 3 when ofm height is greater than 1 or stride w must be divisible by 2 or 3 and ifm width must be divisible by stride_w/2 or stride_w/3
- Dilated kernel height must be in the range [1, 64]
- Product of dilated kernel width and height must be in the range [1, 4096]
- Weight tensor must be 8-bit
- Weight tensor must be constant
- The sum of the weights cannot exceed 8323072
- Optional Bias tensor must be of shape: 1D
- Optional Bias tensor must be of type: int32, int64
- Optional Bias tensor values must fit within 40-bits

https://review.mlplatform.org/plugins/gitiles/ml/ethos-u/ethos-u-vela/+/HEAD/SUPPORTED_OPS.md

Results

Constraints with Vela Compiler

- MatMul in Multi-Head Attention is not supported by the Vela compiler.
- Addressing NPU compatibility issues with Transformer blocks by representing unsupported operators with supported operator combinations.



Operator	TFLite Constraints
ABS	Generic, Specific
ADD	Generic, Specific
ARG_MAX	Generic, Specific
AVERAGE_POOL_2D	Generic, Specific
CONCATENATION	Generic, Specific
CONV_2D	Generic, Specific
DEPTHWISE_CONV_2D	Generic, Specific
EXP	Generic, Specific
EXPAND_DIMS	Generic, Specific
FULLY_CONNECTED	Generic, Specific
HARD_SWISH	Generic, Specific
LEAKY_RELU	Generic, Specific
LOGISTIC	Generic
MAXIMUM	Generic, Specific
MAX_POOL_2D	Generic, Specific
MEAN	Generic, Specific
MINIMUM	Generic, Specific
MIRROR_PAD	Generic, Specific
MUL	Generic, Specific

https://review.mlplatform.org/plugins/gitiles/ml/ethos-u/ethos-u-vela/+/HEAD/SUPPORTED_OPS.md

Results

Constraints with Vela Compiler

BatchMatmul

Unsupported case in Ethos-U NPU

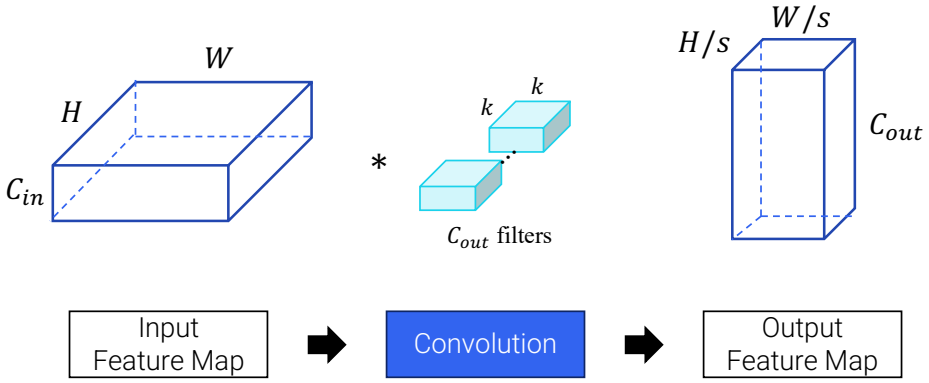
```
Warning: Using internal-default values for system configuration
Warning: Using internal-default values for memory mode
Warning: Unsupported TensorFlow Lite semantics for BATCH_MATMUL 'PartitionedCall:0'. Placing on CPU instead
 - Input(s), Output and Weight tensors must have quantization parameters
   Op has tensors with missing quantization parameters: PartitionedCall:0

Network summary for only_matmul
Accelerator configuration              Ethos_U55_256
System configuration                 internal-default
Memory mode                          internal-default
Accelerator clock                                500 MHz


CPU operators = 1 (100.0%)
NPU operators = 0 (0.0%)

Neural network macs                                0 MACs/batch
Network Tops/s                                   nan Tops/s

NPU cycles                                         0 cycles/batch
SRAM Access cycles                                 0 cycles/batch
DRAM Access cycles                                 0 cycles/batch
On-chip Flash Access cycles                        0 cycles/batch
Off-chip Flash Access cycles                       0 cycles/batch
Total cycles                                       0 cycles/batch

Batch Inference time                 0.00 ms,     nan inferences/s (batch size 1)
```


Results

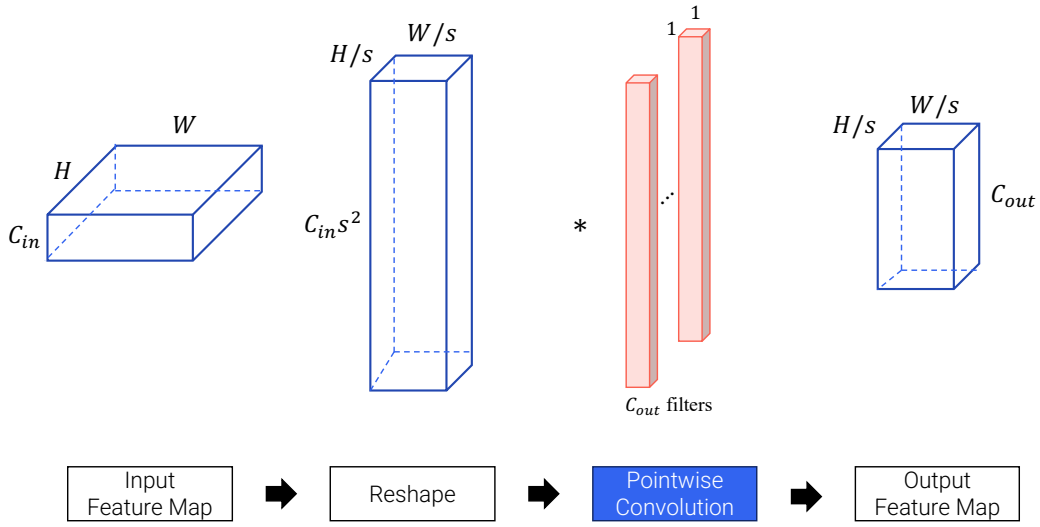
Operator Converter for Patch Encoder

- Patch Encoder can be expressed using combinations of supported operators such as Slice, Reshape, and Pointwise Convolution.

Without Operator Converter



With Operator Converter

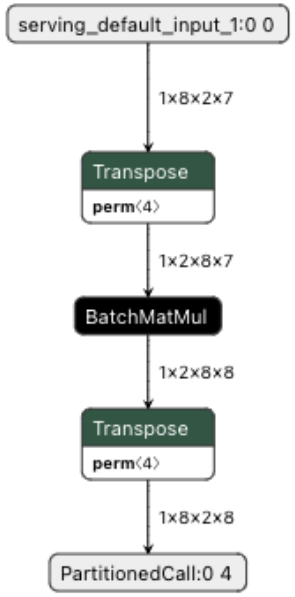


Results

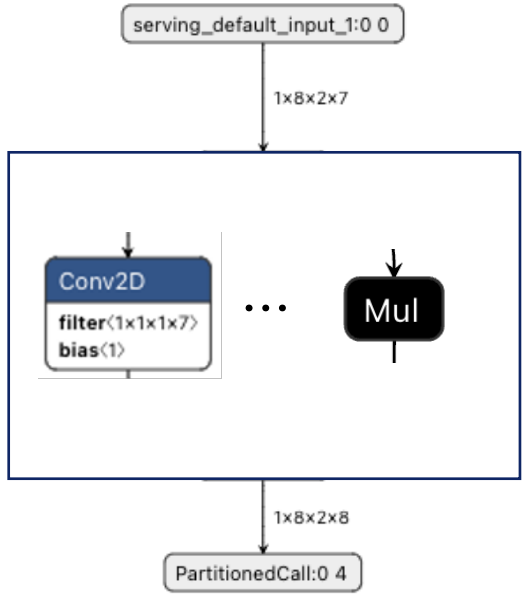
Operator Converter for BatchMatMul

- Batchmatmul can be expressed using combinations of supported operators such as Convolution, Multiplication, and so on.
- Quantization values have also been appropriately filled in.

Without Operator Converter



With Operator Converter

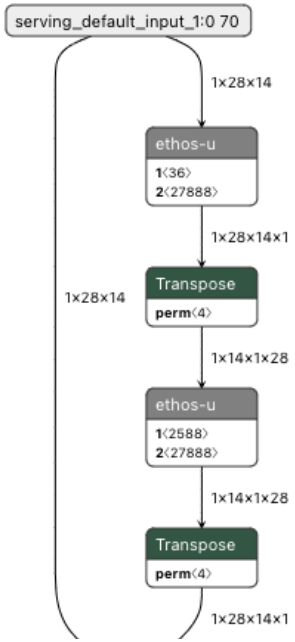


Results

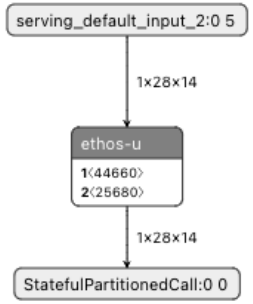
Optimizing NPU Utilization

- After the operator converter, the model consists solely of supported operators, condensing into a single layer through the Vela compiler.
- This setup facilitates full NPU acceleration.

Without Operator Converter



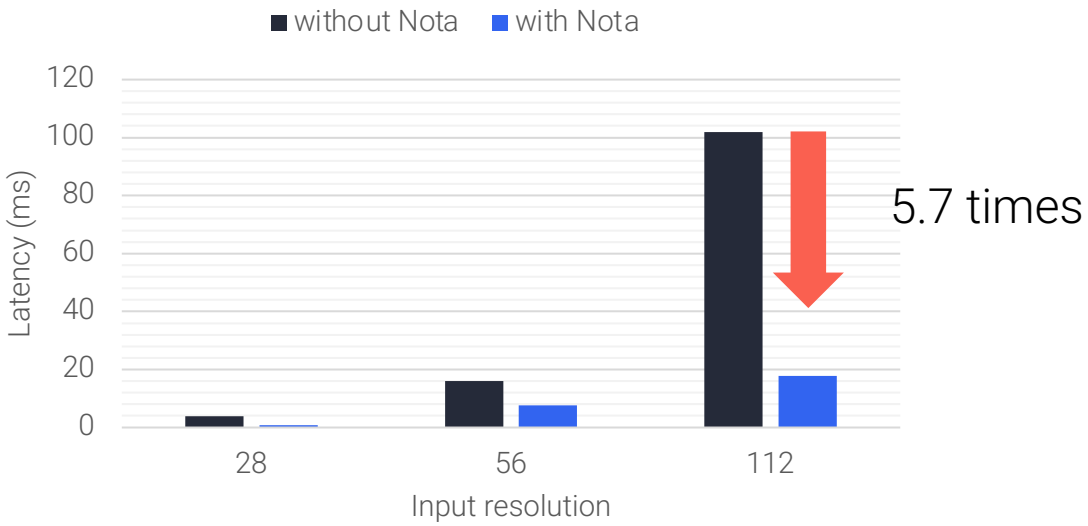
With Operator Converter



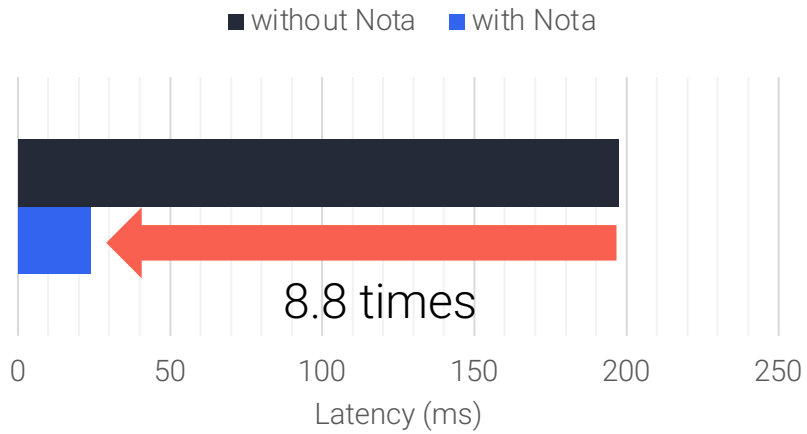
Results

Benchmark

- Ethos-U NPUs don’t currently support Conformer & Vision Transformer blocks.
- Ran AI model with Nota AI® using only Ethos-U NPUs
- On Alif Ensemble E7 DevKit Gen2, Conformer is 5.7 times faster, and ViT is 8.8 times faster



Conformer

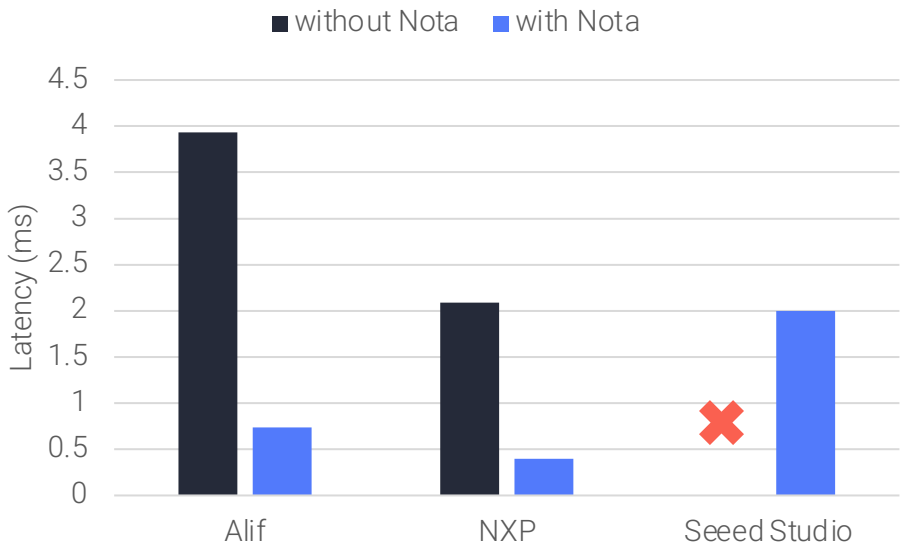


Vison Transformer

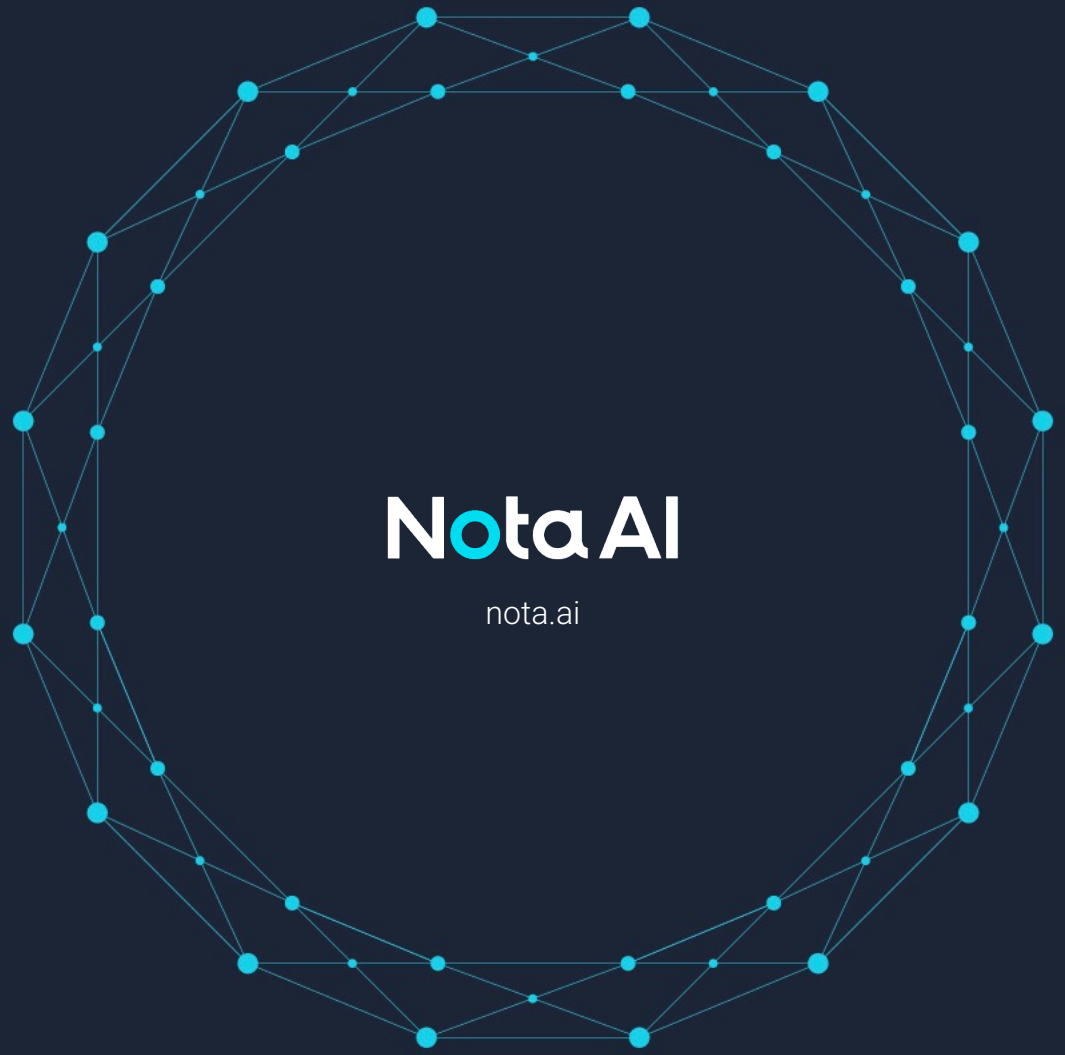
Results

Benchmark

- All devices with Arm Ethos NPUs(Alif, NXP, and Seeed Studio) demonstrated accelerated model inference speeds with Nota AI®.
- Using only NPUs, the Conformer model achieves notable performance improvements on each device.



Block	Type	Latency (ms)		
		Alif Ensemble E7 DevKit Gen2	NXP i.MX 93 Evaluation Kit	Seeed Studio Grove Vision AI V2
Conformer	Original	3.93	2.09	X
	with Nota	0.73	0.40	2.00



Copyright Notice

www.tinyml.org