Implementing Transformer Neural Networks for Visual Perception on Embedded Devices

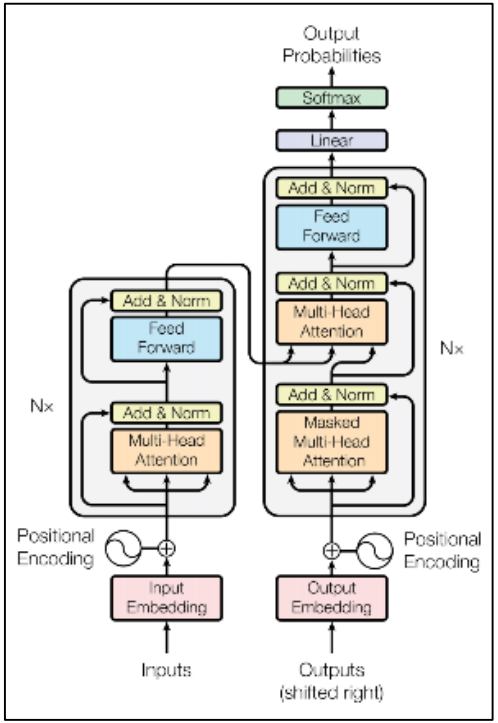
Shang-Hung Lin

VP, NPU IP

VeriSilicon Inc.

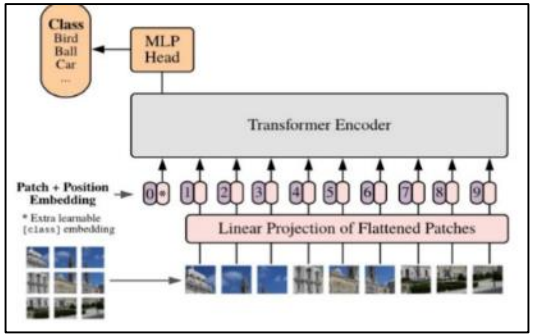
VeriSilicon

Transformer Everything

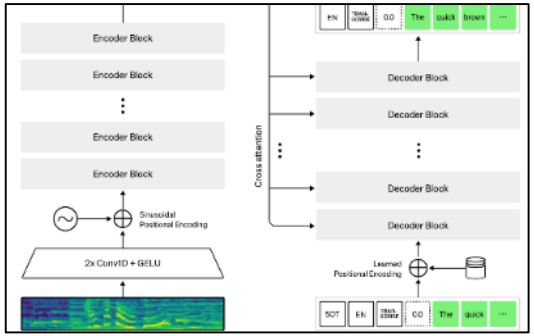


Multi-Head Attention Mechanism

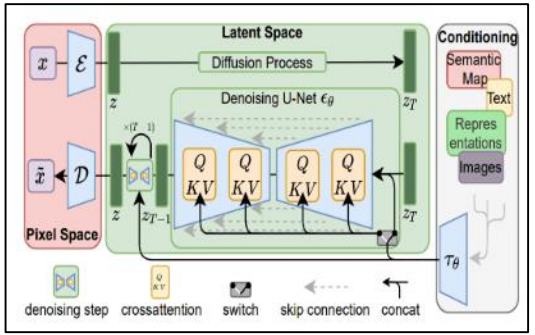
ViT (AI Vision)



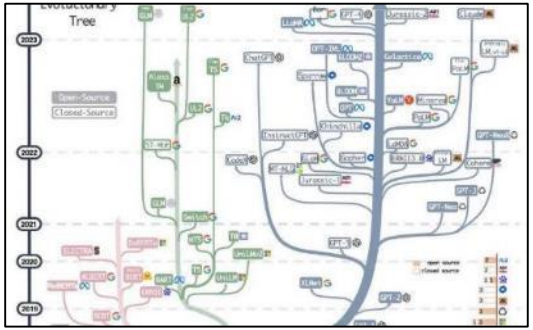
Whisper (AI Voice)



Stable Diffusion (AI Pixel)



LLaMa2 (AI Language)



VeriSilicon: Leadership In Embedded NPU Over 7+ Years

AI LANGUAGE

AI VISION

AI VOICE

AI PIXEL

72 Customers

128+ SoC

More Than 10 Markets

Smart Home

AI Wearable
ARVR/Watch

AI Server

Automotive
ADAS

Surveillance
IPC Cameras

AIoT
Smartphone

PC

Robotics

VeriSilicon NPU IP is Shipped in Over 100 Million AI-Enabled Chips Worldwide

Enabling efficient execution of applications ranging from AI voice, AI vision, AI pixel to AIGC on embedded devices

February 28, 2024 07:00 PM Eastern Standard Time

businesswire
A BERKSHIRE HATHAWAY COMPANY

VeriSilicon

VIP9000 Series: World-Class Performance for Generative AI

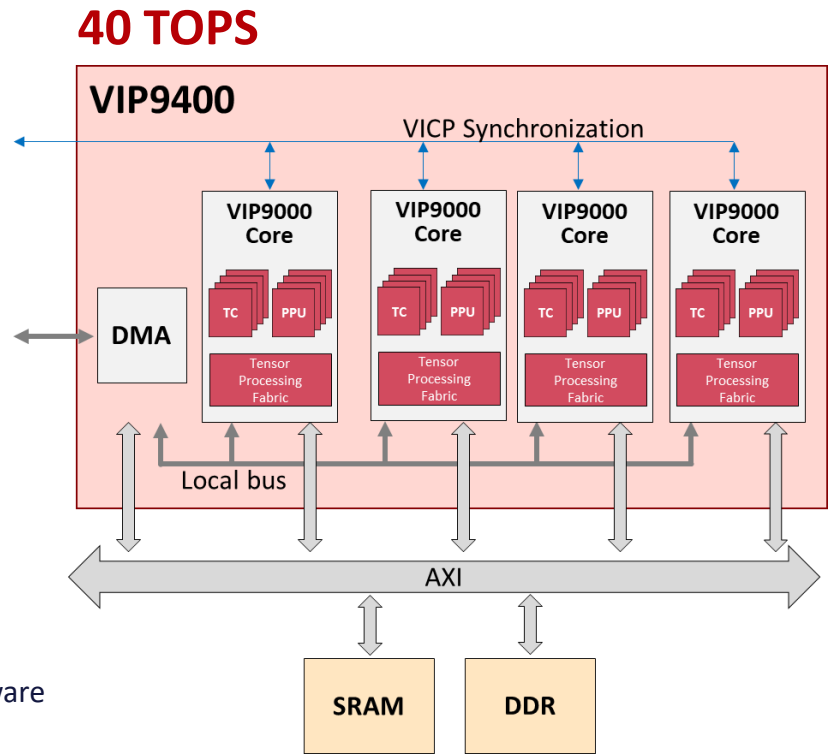
Stable Diffusion 1.5

20 steps under 2 seconds

LLaMA2 7B

20 Tokens/s

- Off-the-shelf models. No tailored modification for hardware



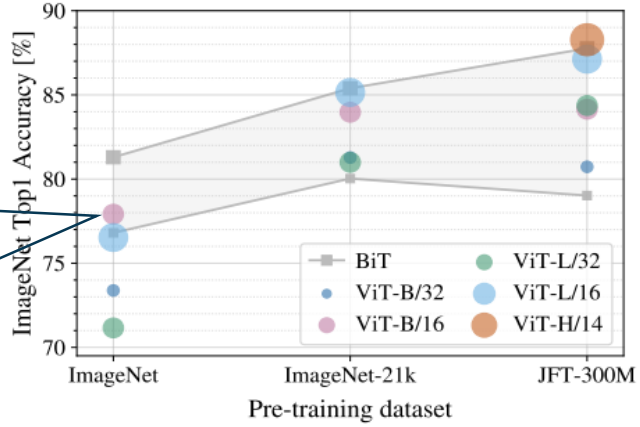
Challenges of Deploying ViT on Embedded Devices

- ViT is good at scaling up; tends to grow a lot bigger than CNNs
  - But the resource on embedded devices is limited
- ViT needs a lot more data to train well
  - Lacking inductive biases is a double-edged sword

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Table 1: Details of Vision Transformer model variants.

Network	***Input Img Size***	***ImageNet1K Top-1 (%)***	***# Param (M)***	***MACs (G)***
ViT-B/16	384^2	77.9	87	55.5
ResNet-101	224^2	77.4	45	7.9
EfficientNet-B0	224^2	77.1	5.3	0.4



From ViT original paper (arxiv 2010.11929)

How to “Squeeze ViT In” Without Sacrificing Performance?

- Knowledge distillation
- Pruning
- Weight sharing
- Quantization
- Hybrid architecture
- HW accelerator for embedded devices

Knowledge Distillation

- Transfer the knowledge of a pre-trained “teacher” model to a smaller “student” model
- Data Efficient Image Transformer (DeiT, arXiv:2012.12877):
  - Use identical ViT architecture and learn inductive biases from a large CNN “teacher”
  - Achieve 5+% accuracy improvement with a small training dataset
  - Use Knowledge Distillation to create smaller ViT model variants

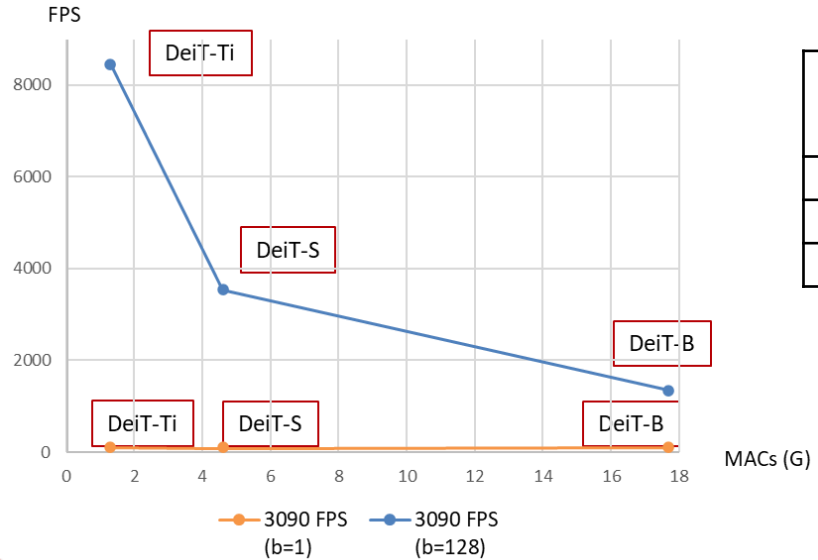
Network	*Input Img Size*	*ImageNet1K Top-1 (%)*	*# Param (M)*	*MACs (G)*
ViT-B/16	384^2	77.9	**87**	55.5
DeiT-B	224^2	83.4	87	17.7
DeiT-S	224^2	81.2	**22**	4.6
DeiT-Ti	224^2	74.5	5.9	1.3

4x model size reduction with better Top1 %

- Training dataset: ImageNet
- Teacher: RegNetY-16GF
  - 84M parameters
  - Top1 82.9%

Memory Bandwidth Dictates ViT Inference Speed

- Embedded applications may not allow batch processing
- Let's continue to compress ViT



Network	*# Param (M)*	*Model Size (MB)*	*MACs (G)*	*RTX3090 FPS (batch=1)*	*RTX3090 FPS (batch=128)*
DeiT-B	87	348	17.7	103	1343
DeiT-S	22	88	4.6	101	3538
DeiT-Ti	5.9	23.6	1.3	102	8453

EVS2024

More Model Reduction Techniques

- Pruning: set insignificant weights to zero
- Weight sharing or weight multiplexing: reuse weights from one layer to other layers
- May need special hardware (e.g., NPU) to take full advantage

Network	*Input Img Size*	*ImageNet1K Top-1 (%)*	*# Param (M)*	*MACs (G)*	
ViT-B/16	**384^2**	**77.9**	**87**	**55.5**	
DeiT-B	224^2	83.4	87	17.7	
DeiT-S	**224^2**	**81.2**	**22**	**4.6**	
DeiT-Ti	224^2	74.5	5.9	1.3	
X-Pruner-DeiT-S	**224^2**	**78.9**	**22**	**2.4**	Pruning
X-Pruner-DeiT-Ti	224^2	71.1	5.9	0.6	
Mini-DeiT-S	**224^2**	**80.9**	**11**	**4.7**	Weight multiplexing
Mini-DeiT-Ti	224^2	72.8	3	1.3	

Quantizing ViT to Lower Bits

- Direct saving on memory footprint, bandwidth, and power
- Post-Training Quantization (PTQ)
  - Quantizes a pre-trained model with a small calibration set (fast)
  - Can deliver 8-bit ViT with decent accuracy
- Quantization-Aware Training (QAT)
  - Interleaves quantization during model training phase (costly)
  - For 4-bit or lower

PTQ Comparison	# Bit (W | A)	ImageNet1K Top-1 (%)	# Param (M)	Model Size (MB)
DeiT-B	32 | 32	83.4	87	348
DeiT-S	32 | 32	81.2	22	88
APQ-ViT-DeiT-B	8 | 8	81.7	87	87
APQ-ViT-DeiT-S	8 | 8	79.8	22	22
APQ-ViT-DeiT-B	6 | 6	80.4	87	64
APQ-ViT-DeiT-S	6 | 6	77.8	22	16.6
APQ-ViT-DeiT-B	4 | 4	67.5	87	43.5
APQ-ViT-DeiT-S	4 | 4	43.5	22	11

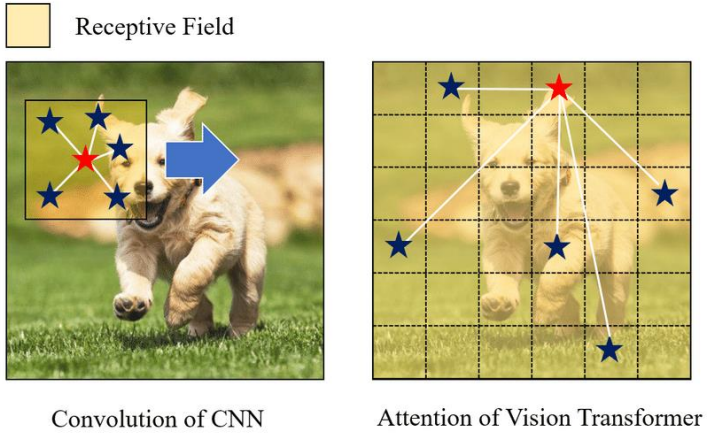
APQ-ViT – arxiv:2303.14341

QAT Comparison	# Bit (W | A)	ImageNet1K Top-1 (%)	# Param (M)	Model Size (MB)
DeiT-B	32 | 32	83.4	87	348
DeiT-S	32 | 32	81.2	22	88
Q-ViT-DeiT-B	4 | 4	83.0	87	43.5
Q-ViT-DeiT-S	4 | 4	80.9	22	11
Q-ViT-DeiT-B	3 | 3	81.0	87	33.4
Q-ViT-DeiT-S	3 | 3	79.0	22	8.7
Q-ViT-DeiT-B	2 | 2	74.2	87	21.8
Q-ViT-DeiT-S	2 | 2	72.1	22	6

Q-ViT – NeurIPS 2022

Hybrid Architecture – The Motivation

- Transformer is good at capturing long range dependency
- Convolution can extract local information efficiently due to its inductive biases
- ViT learns the meaning from image patches; why not learn from the feature maps extracted by CNN?

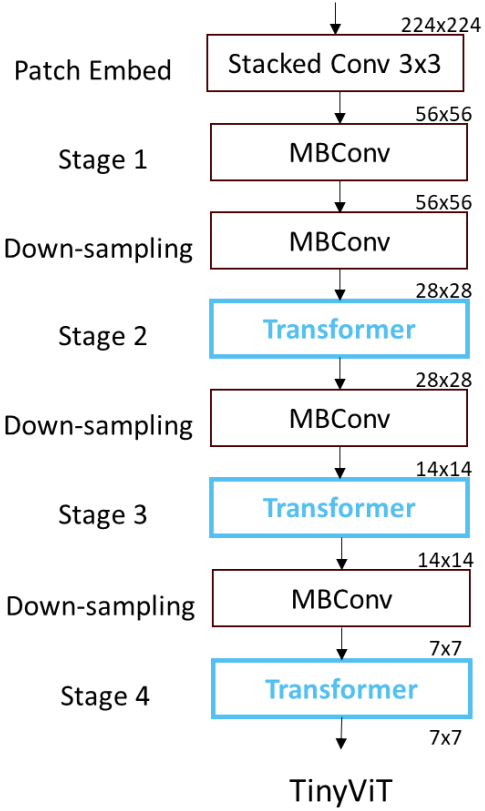


Hybrid Architecture (cont'd)

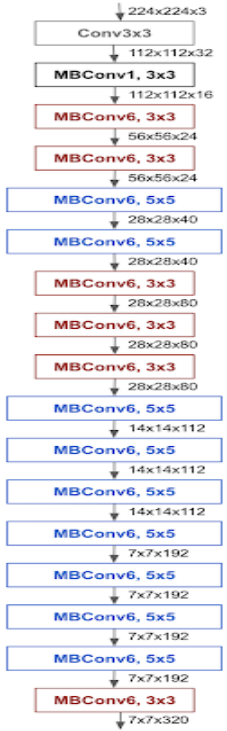
- TinyViT (ECCV 2022)
  - Extracts local information by Conv3x3 and MBConv (inverted residual block) from the beginning
  - Trained by knowledge distillation
  - Comparison with EfficientNet-B0 (similar parameter size):

Network	*Input Img Size*	*ImageNet1K Top-1 (%)*	*# Param (M)*	*MACs (G)*
EfficientNet-B0	224^2	77.1	5.3	0.4
TinyViT-5M	224^2	80.7	5.4	1.3

EVS2024



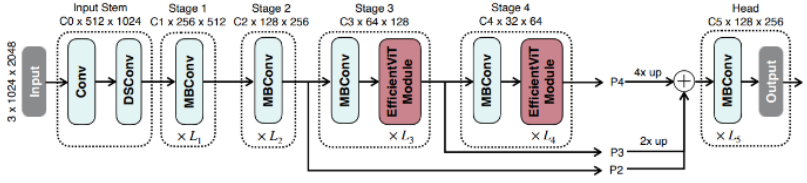
TinyViT

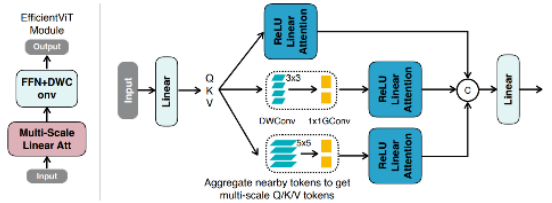


EfficientNet-B0

Hybrid Architecture (cont'd)

- EfficientViT (ICCV 2023)
  - Also extracts local info by convolutions
  - "Multi-scale linear attention" for speedup
    - Replacing softmax with ReLu
    - Add multi-scale depthwise and 1x1 convolutions to improve receptive field





arxiv:2205.14756

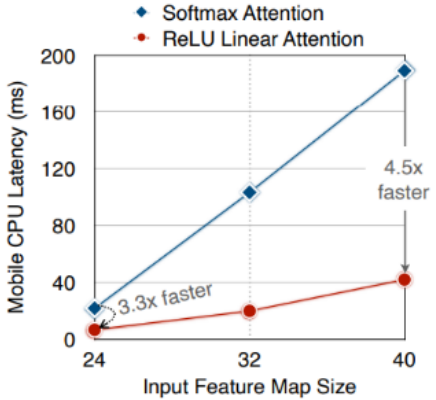
Linear Attention

- Attention:

$$V'_i = \frac{\sum_{j=1}^{N} sim(Q_i, K_j)V_j}{\sum_{j=1}^{N} sim(Q_i, K_j)}, \quad sim(q,k) = exp\left(\frac{q \cdot k^T}{\sqrt{D}}\right)$$

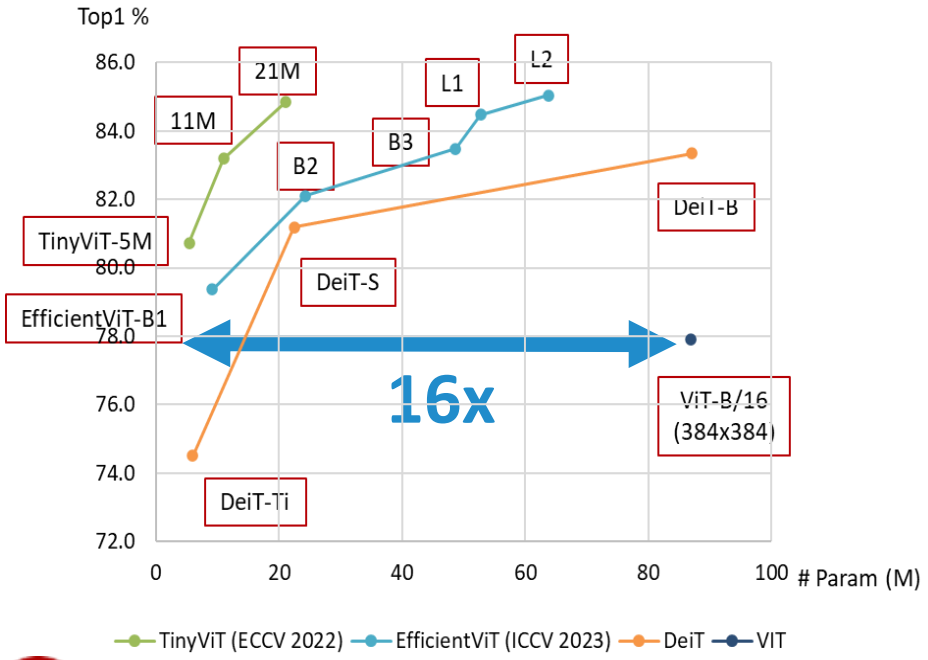
- Replace *sim(q,k)* with separable $\phi(q)\phi(k)$ (e.g. $\phi(x)=ReLu(x)$)

$$V'_i = \frac{\sum_{j=1}^{N} \phi(Q_i)^T \phi(K_j)V_j}{\sum_{j=1}^{N} \phi(Q_i)^T \phi(K_j)} \Rightarrow V'_i = \frac{\phi(Q_i)^T \left(\sum_{j=1}^{N} \phi(K_j)V_j\right)}{\phi(Q_i)^T \left(\sum_{j=1}^{N} \phi(K_j)\right)}$$

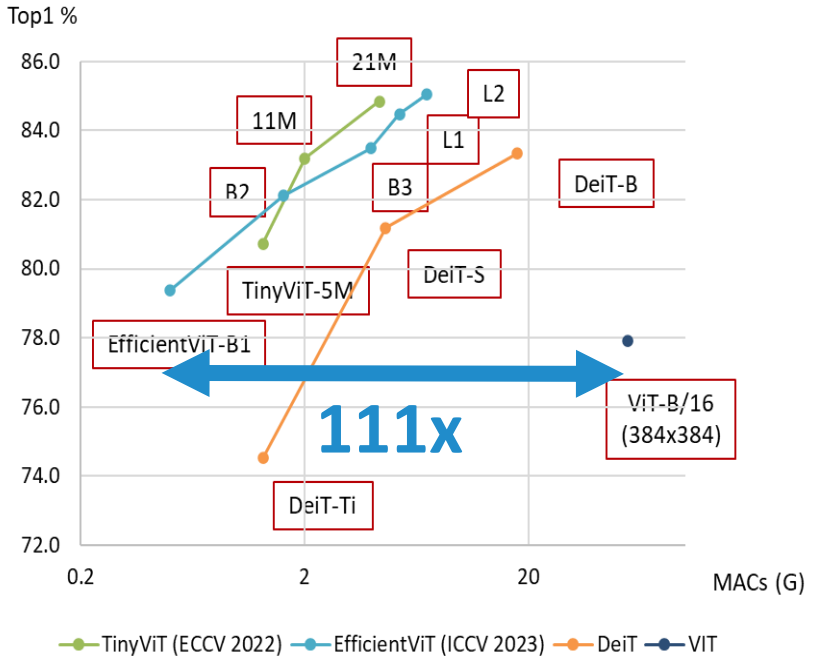


Hybrid Architecture (cont'd)

- Parameter Size Comparison

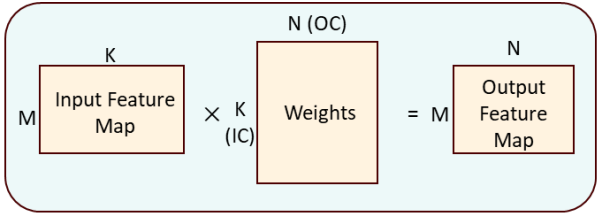


- MAC Count Comparison



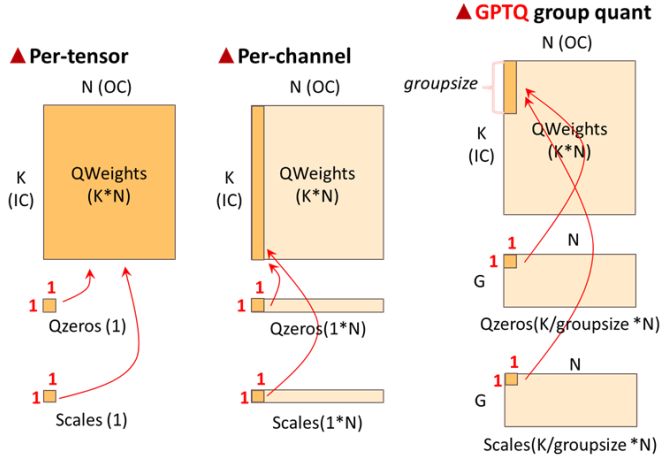
Quantizing Small ViTs

- PTQ weight quantization
  - 8-bit weights: Per-channel quantization
  - 4-bit weights: GPTQ group quantization
- Mixed-precision activations
  - Activations are more sensitive than weights
  - Static range PTQ: per-layer assigning bits based on KL divergence



Network	*ImageNet1K Top-1 (%)*	*Model Size (MB)*
ViT-B/16 (Full Precision)	77.9	348
TinyViT-5M (Full Precision)	80.7	21.6
TinyViT-5M (W: INT8, A: $INT8_{MP}$)	80.4	5.4
TinyViT-5M (W: INT4, A: $INT8_{MP}$)	**78.8**	**2.7**
EfficientViT-B1 (Full Precision)	79.4	36.4
EfficientViT-B1 (W: INT8, A:$INT8_{MP}$)	78.5	9.1
EfficientViT-B1 (W: INT4, A:$INT8_{MP}$)	76.5	4.6

128x



EVS2024

Key NPU Technologies to Enable ViT on Embedded Devices

- High bandwidth / throughput
- Highly efficient matrix engine
- In-place transpose
- 4-bit and mixed precision HW & SDK
- Weight compression
- Efficient accelerator for “hidden devil” operators
- AI compiler to optimize the graph and take full advantage of hardware accelerations
- Decent CNN performance (still need it)

EfficientViT for Semantic Segmentation

VeriSilicon
VIP9400
64 TOPS

NVIDIA
Jetson AGX Orin
275 TOPS

- Cityscapes 2048x1024
- EfficientViT L1, mIoU 82.7

Squeezing ViT Into Embedded Devices – Summary

Compressing baseline ViT-B/16:

- 128x model size reduction
- 111x MACs reduction
- No top-1 accuracy loss

Network	*ImageNet1K Top-1 (%)*	*Model Size (MB)*	*MACs (G)*
ViT-B/16 (Full Precision)	77.9	348	55.5
DeiT-S (Full Precision)	81.2	88	4.6
Q-ViT-DeiT-S (W4A4)	80.9	11	4.6*
TinyViT-5M ($W4A8_{MP}$)	78.8	**2.7**	1.3*
EfficientViT-B1 ($W8A8_{MP}$)	78.5	9.1	**0.5***

EVS2024

Let's keep watch as the technology evolves:

- Compressing higher precision ViTs
- 4-bit (INT4, FP4) or less
- New training methods, quantization techniques, model architectures
- HW acceleration and lower power on embedded devices

Resources

Vision Transformer (ViT)
https://arxiv.org/pdf/2010.11929.pdf

Knowledge Distillation (DeiT)
https://arxiv.org/pdf/2012.12877.pdf

Pruning, Weight Multiplexing
https://arxiv.org/pdf/2303.04935.pdf
https://arxiv.org/pdf/2204.07154v1.pdf

Quantization
https://arxiv.org/pdf/2303.14341.pdf
https://arxiv.org/pdf/2210.06707.pdf

Hybrid Architecture
https://arxiv.org/pdf/2207.10666.pdf
https://arxiv.org/pdf/2205.14756.pdf

2024 Embedded Vision Summit

VeriSilicon booth is at 509. Welcome to visit us!