ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

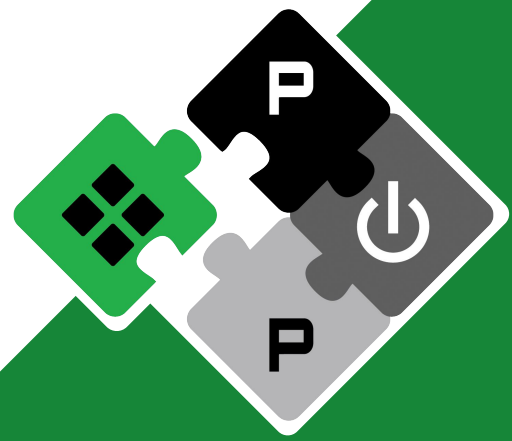
KU LEUVEN

Tiny CNNs or Tiny Transformers?

An Analysis of Latency, Accuracy and Domain Shift Robustness in Extreme-Edge Image Classification

Călin Diaconu - DEI, University of Bologna
Luca Bompani - DEI, University of Bologna
Alberto Dequino - DEI, University of Bologna
Davide Nadalini - PSI, Katholieke Universiteit Leuven
Francesco Conti - DEI, University of Bologna

PULP Platform
Open Source Hardware, the way it should be!



pulp-platform.org
@pulp_platform
company/pulp-platform
youtube.com/pulp_platform

Background

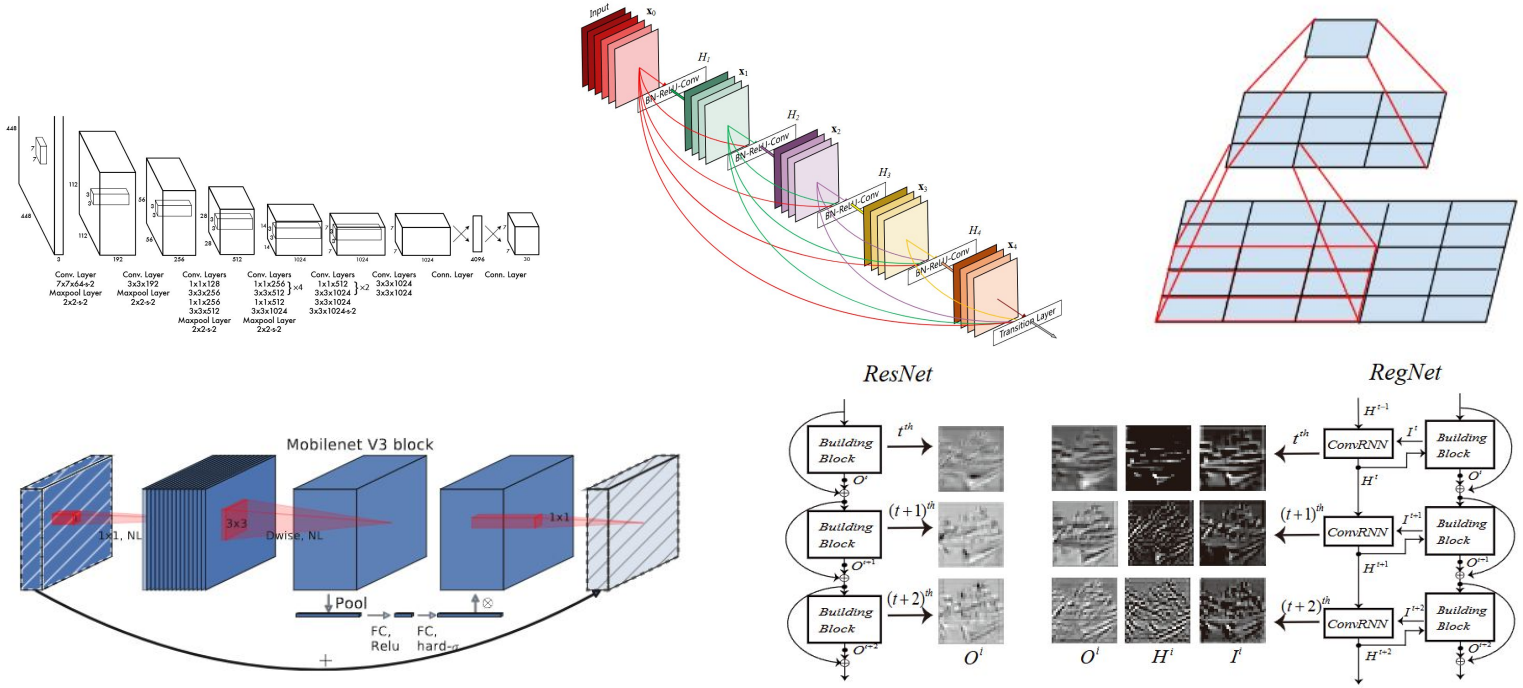


Fig. 1 - CNN architectures. Sources: [1], [2], [3], [4], [5]

CNNs traditionally used for **ML Computer Vision** in all types of tasks:

- classification
- segmentation
- object detection

Background

- The **Attention** mechanism and the **Transformer** took the stage by storm **9 years ago**
- **Replaced RNNs** with **self-attention** allowing every sequence element to **parallely relate to every other one**
- Initially **designed as NLP** solutions

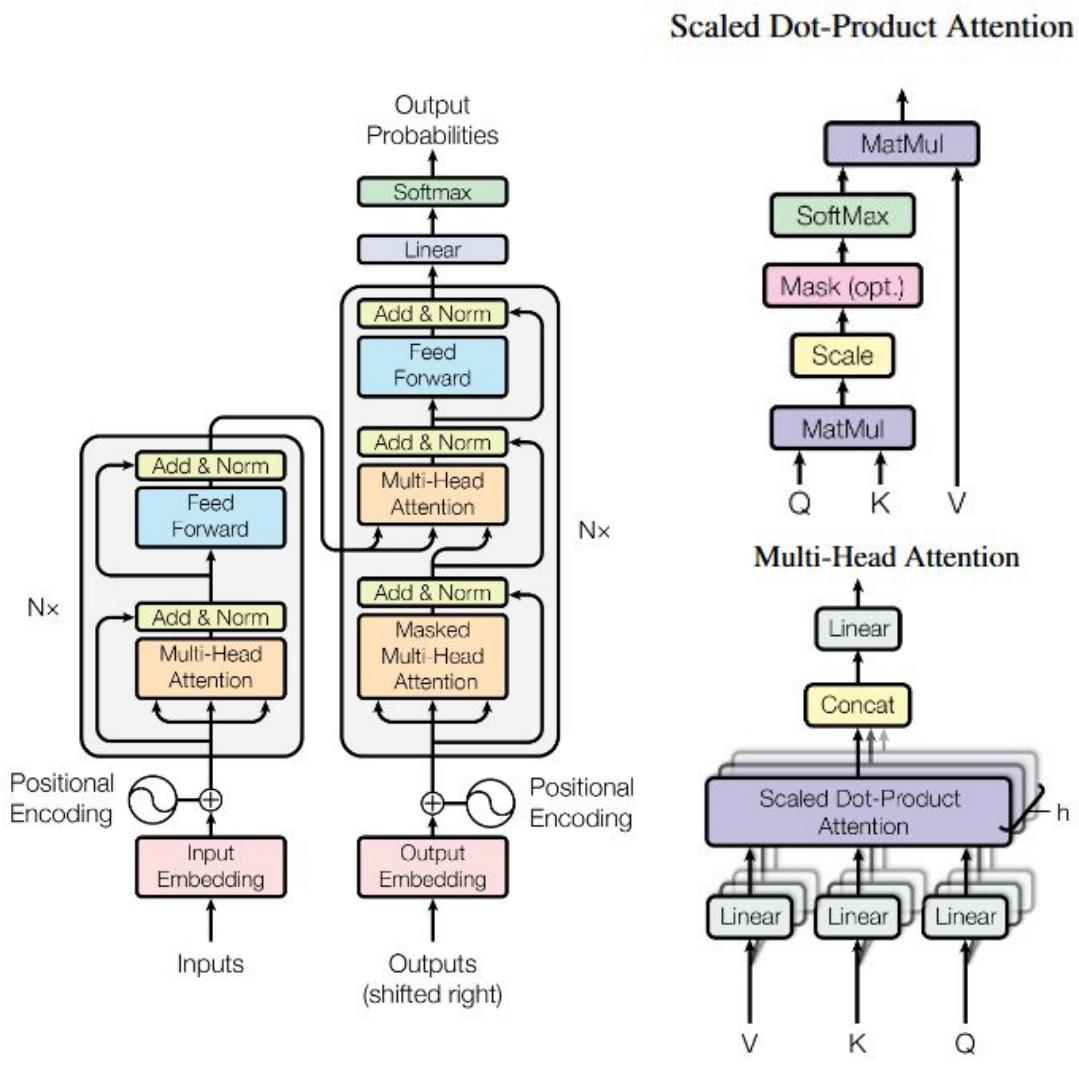


Fig. 2 - Transformer & Attention. Source: [7]

Background

- Quickly **ported to vision** problems
- Dosovitskiy's **Vision Transformer (ViT)** became a landmark
- Adopted architectures - **pure attention**, but mostly **hybrid** (Conv. & att. layers)

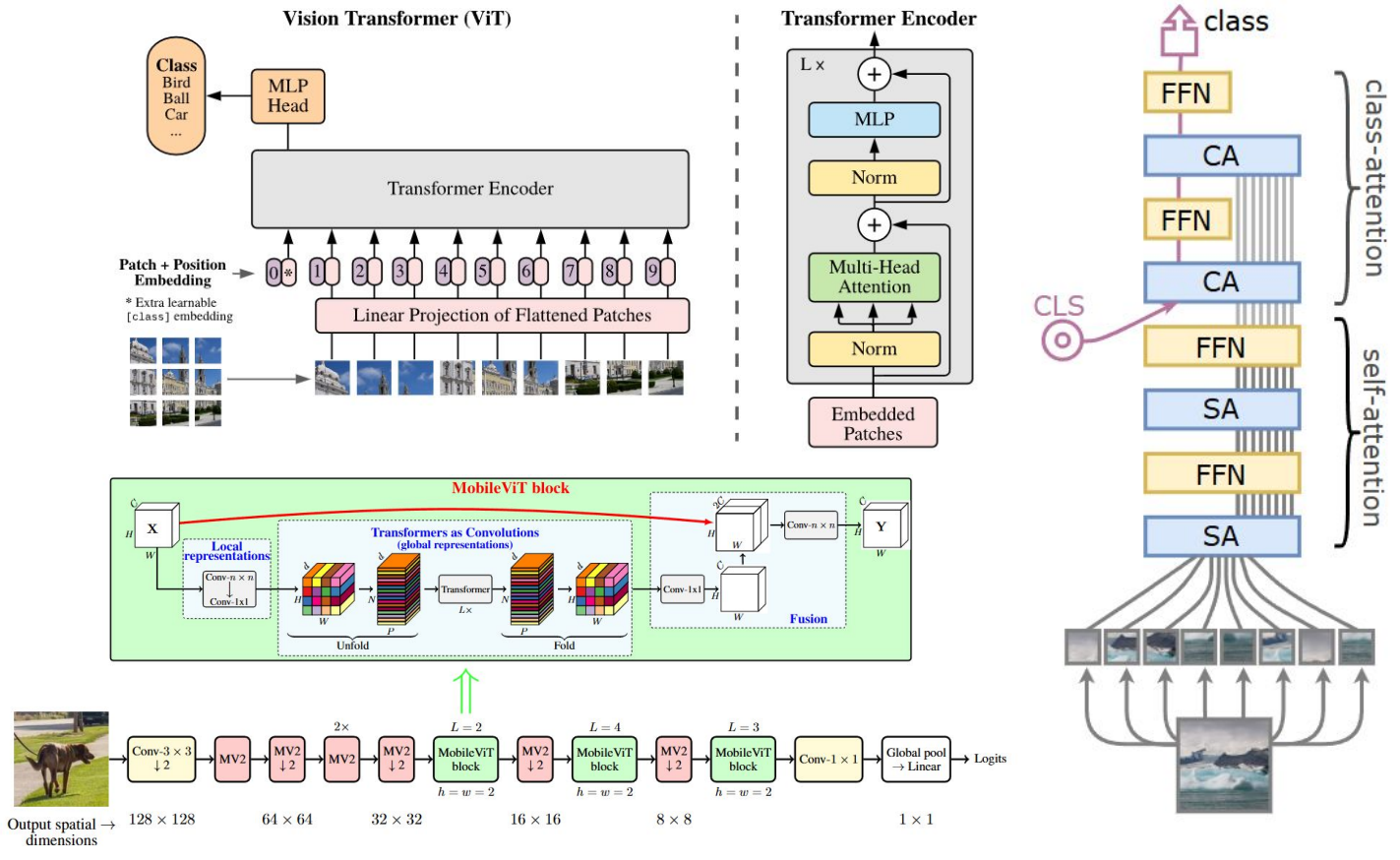


Fig. 3 - ViT architectures. Sources: [8], [9], [10]

The Gap?

ViT Large
85.15% accuracy

CaiT M48
86.5% accuracy

DINO v2 Large
86.3% acc.

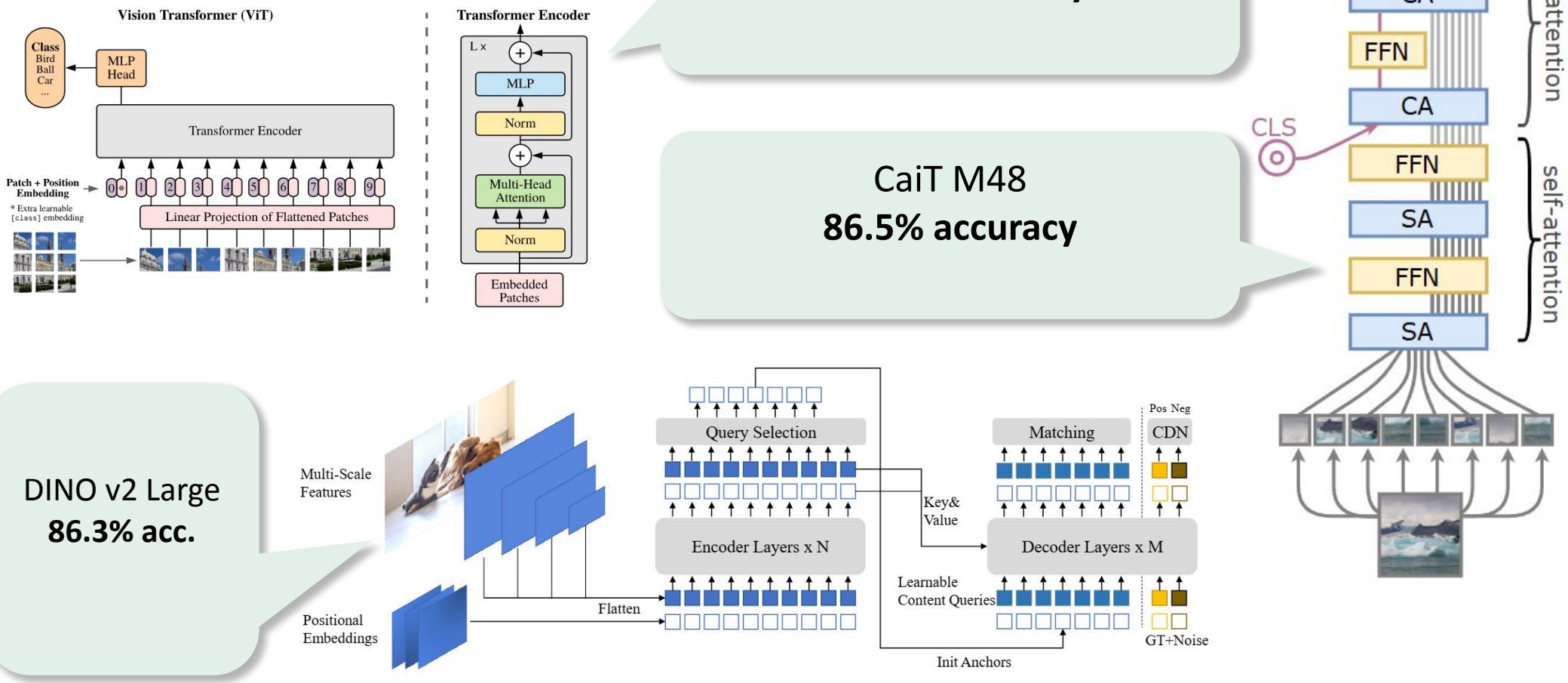


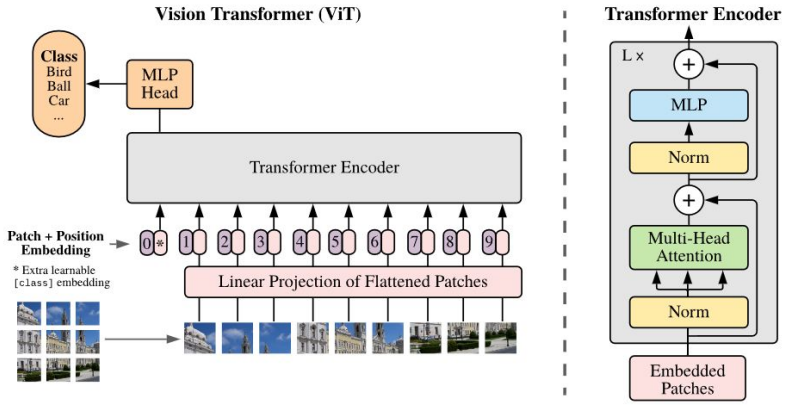
Fig. 4 - ViT [8], DINO [13], and CaiT [9] architectures

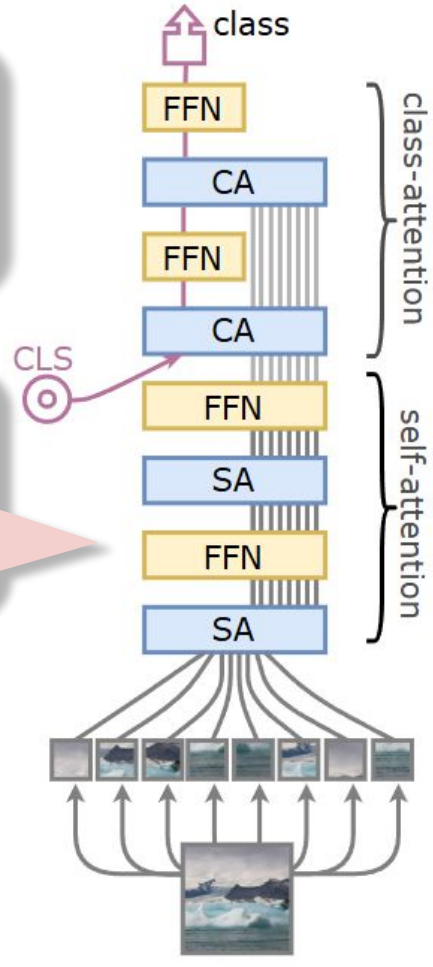
The Gap

ViT Large
85.15% accuracy
307M params

CaiT M48
86.5% accuracy
356M params

DINO v2 Large
86.3% acc.
306M params.





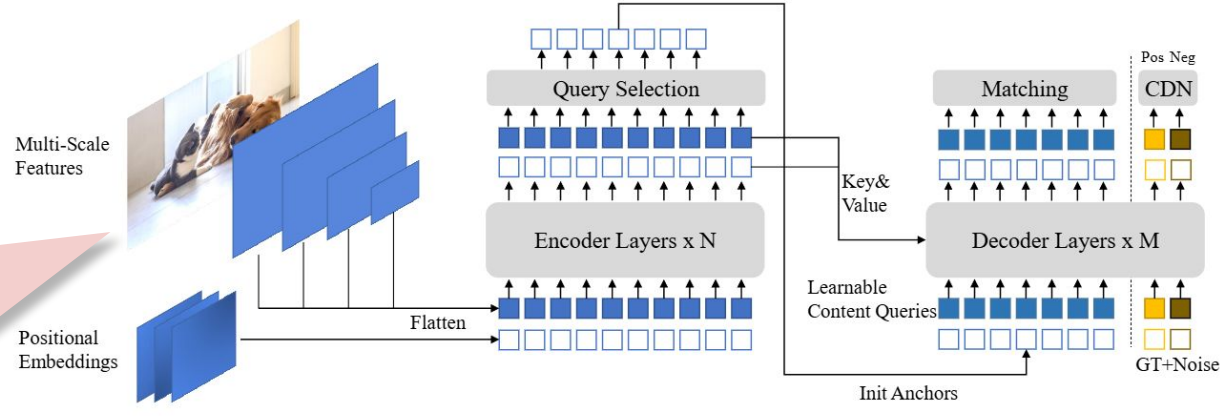


Fig. 4 - ViT [8], DINO [13], and CaiT [9] architectures

The Gap

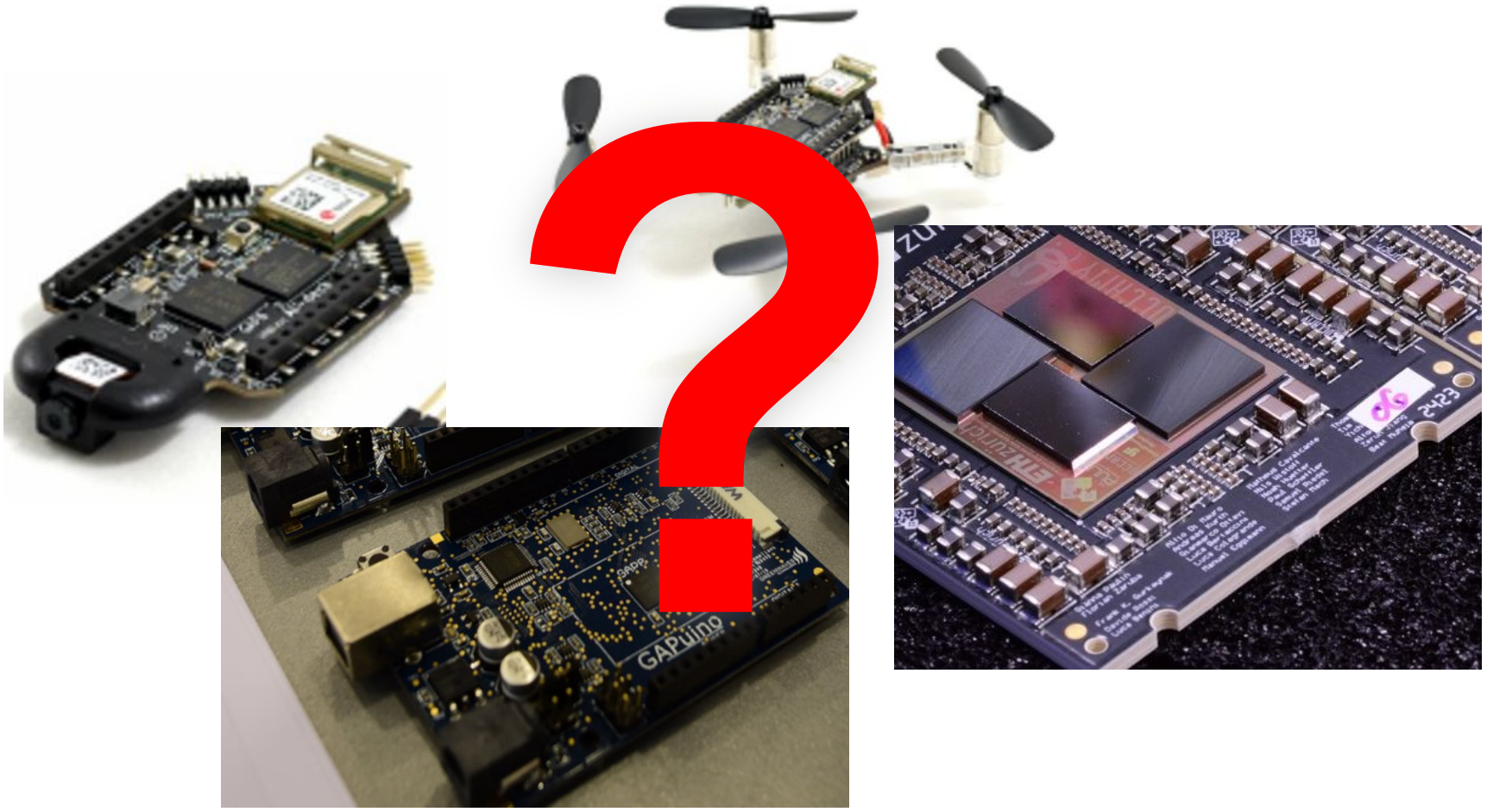


Fig. 5 - Ultra-Low Power Platforms [14]

The Analysis Method

Hardware-agnostic

Hardware-aware

The Analysis Method

Hardware-agnostic

- **Computational complexity** - floating point operations (FLOPs) per inference
- **Model size** - parameter count
- **Task performance** - top-1 classification accuracy, evaluated on the ImageNet evaluation subset
- **Domain shift robustness**

Hardware-aware

The Analysis Method

Hardware-agnostic **Hardware-aware**

- **Computational** point operatio inference
- **Model size** - p
- **Task performa** classification a the ImageNet
- **Domain shift**

Domain shift robustness

- Backbone **pre-trained** on one set (ImageNet)
- Head **fine-tuned** and model **evaluated** on other sets: CIFAR 10 and Visual Wake Words (MLPerf Tiny)
- Two types of evaluation:
  - **Domain shift only** - input resolution identical between pre-train and fine-tune + evaluation
  - **Domain and resolution shift** - fine-tune + evaluation on different input resolution relative to pre-training

The Analysis Method

Hardware-agnostic

- **Computational complexity** - floating point operations (FLOPs) per inference
- **Model size** - parameter count
- **Task performance** - top-1 classification accuracy, evaluated on the ImageNet evaluation subset
- **Domain shift robustness**

Hardware-aware

- **Working memory**

The Analysis Method

Hardware-agnostic **Hardware-aware**

- **Computation** point operat inference
- **Model size** -
- **Task perforr** classificatior the ImageNe
- **Domain shif**

Working memory

- **Peak memory** required during inference
- Computed as **worst-case per-layer memory**
- Offers **finer-grained profiling** compared to end-to-end
- Inputs, outputs, weights, biases, other buffers considered
- **Auxiliary elements ignored**, since they are implementation-dependent

The Analysis Method

Hardware-agnostic

- **Computational complexity** - floating point operations (FLOPs) per inference
- **Model size** - parameter count
- **Task performance** - top-1 classification accuracy, evaluated on the ImageNet evaluation subset
- **Domain shift robustness**

Hardware-aware

- **Working memory**
- **Inference latency**

The Analysis Method

Hardware-agnostic **Hardware-aware**

- **Computational** point operatio inference
- **Model size** - p
- **Task performa** classification a the ImageNet
- **Domain shift** r

Inference latency

- Measured in **clock cycles**
- Measured with on-board **performance counters**, directly on the **target platform**:
  - **PULP SoC**
  - MCU + **8 RISC-V cores**
  - **L2** memory limited to **2 MiB**

The Analysis Method

- **16 model families** chosen
- Up to **12M parameters**
- Most “attention” models **actually hybrid**
- Most **top positions** belong to attention or hybrid

Model	Year	Params	FLOPs	Top-1 Acc	Type
TinyViT-5M	2022	5.4M	1.3G	79.1–80.7%	Hybrid
CaiT-XXS-24	2021	12M	2.5G	78.4%	Attention
FastViT-T8	2023	4M	0.7G	77.2%	Hybrid
EfficientNet-B0	2019	5.3M	0.39G	77.1%	CNN
EfficientFormerV2-S0	2023	3.5M	0.4G	76.2%	Hybrid
DeiT-Tiny	2020	5.7M	1.26G	72.2–74.5%	Attention
EdgeNeXt-XXS	2022	1.3M	0.2G	71.2%	Hybrid
PVT-v2-B0	2022	3.7M	0.6G	∼70%	Hybrid
MobileViTv2-0.5x	2022	∼1.5M	0.5G	∼70%	Hybrid
MobileViT-XXS	2021	1.3M	0.4G	69.0%	Hybrid
MobileNetV3 Small 0.75x	2019	2.4M	44M	65%	CNN
ShuffleNetV2 0.5x	2018	1.4M	41M	61.1%	CNN
MobileNetV2 0.35x	2018	1.7M	66M	60.3%	CNN
SqueezeNet V1.1	2016	1.2M	0.35G	58.2%	CNN
ShuffleNetV1 0.5x	2018	0.7M	38M	57.3%	CNN
MobileNetV1 0.25x	2017	0.5M	41M	50.0%	CNN

Results - Accuracy at the Same Num. of Params.

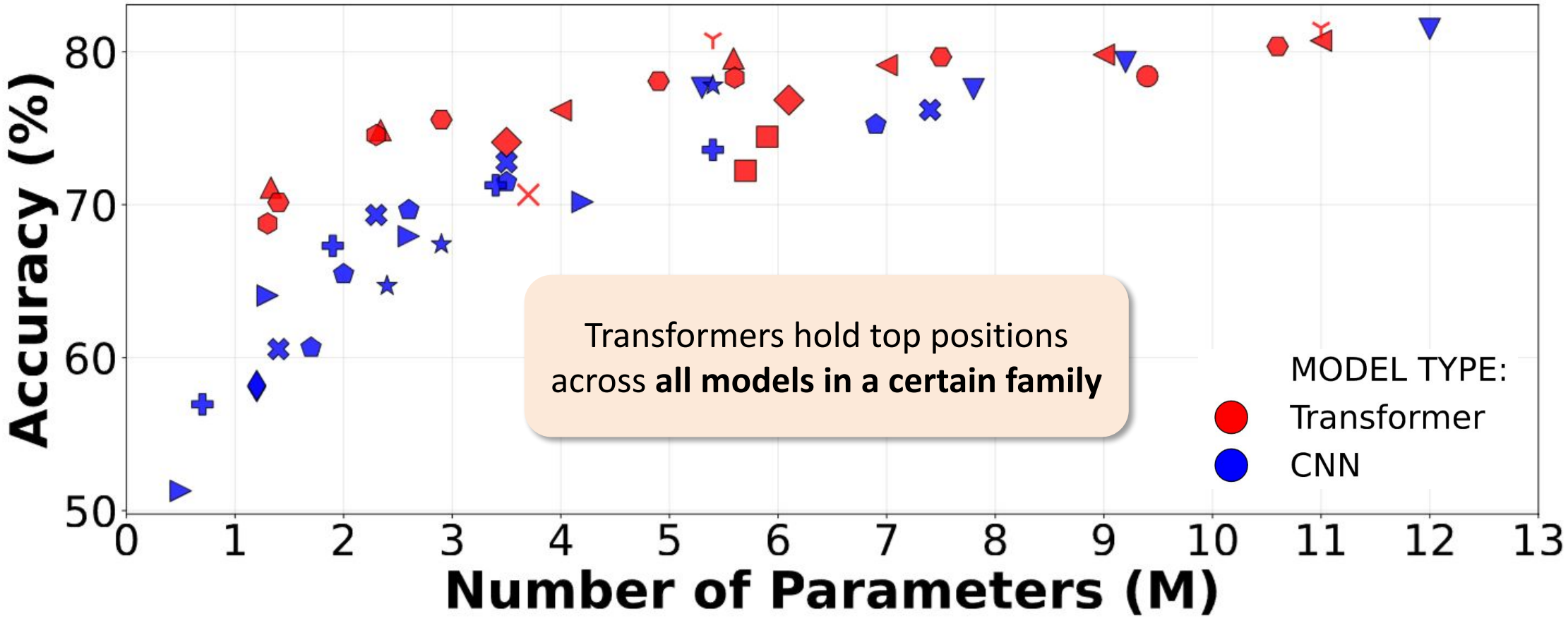


Fig 6: ImageNet - Number of Parameter / Accuracy @ 224x224 input

Results - Accuracy at the Same Num. of Params.

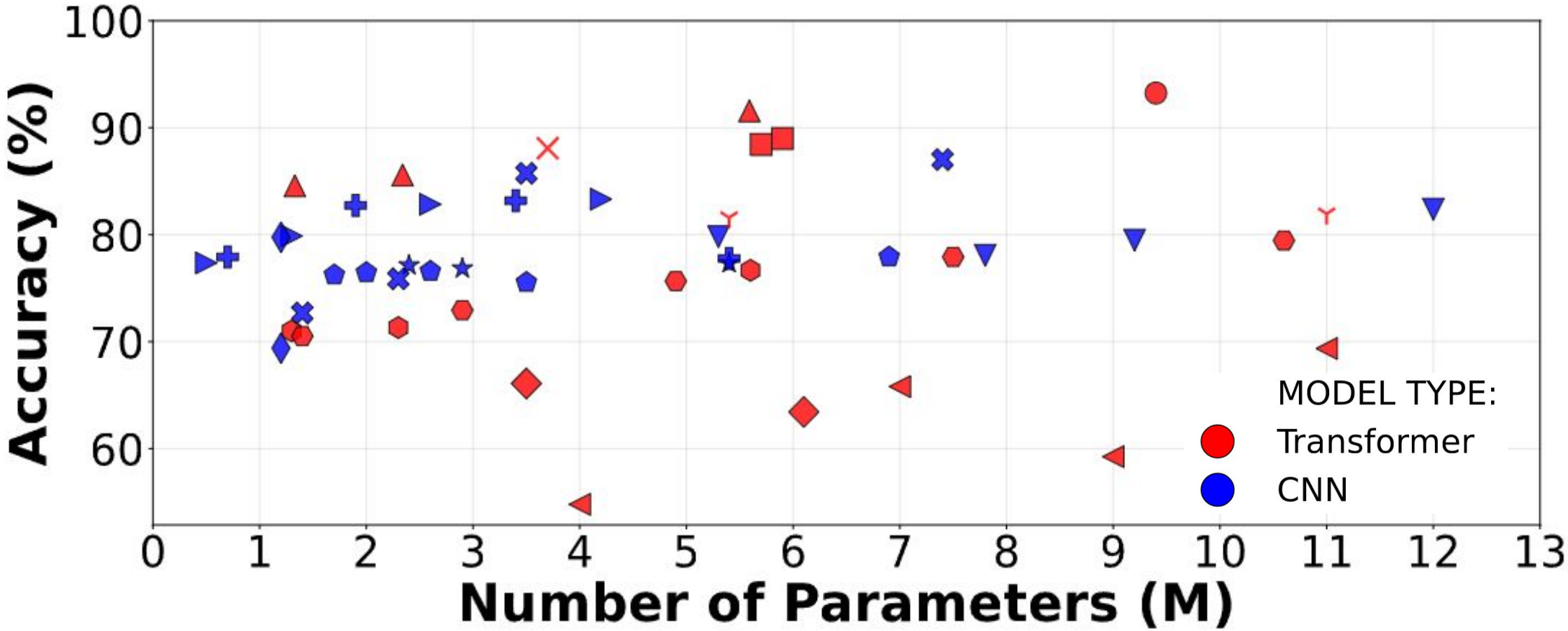


Fig 7: CIFAR-10 - Number of Parameter / Accuracy @ 224x224 input

Results - Accuracy at the Same Num. of Params.

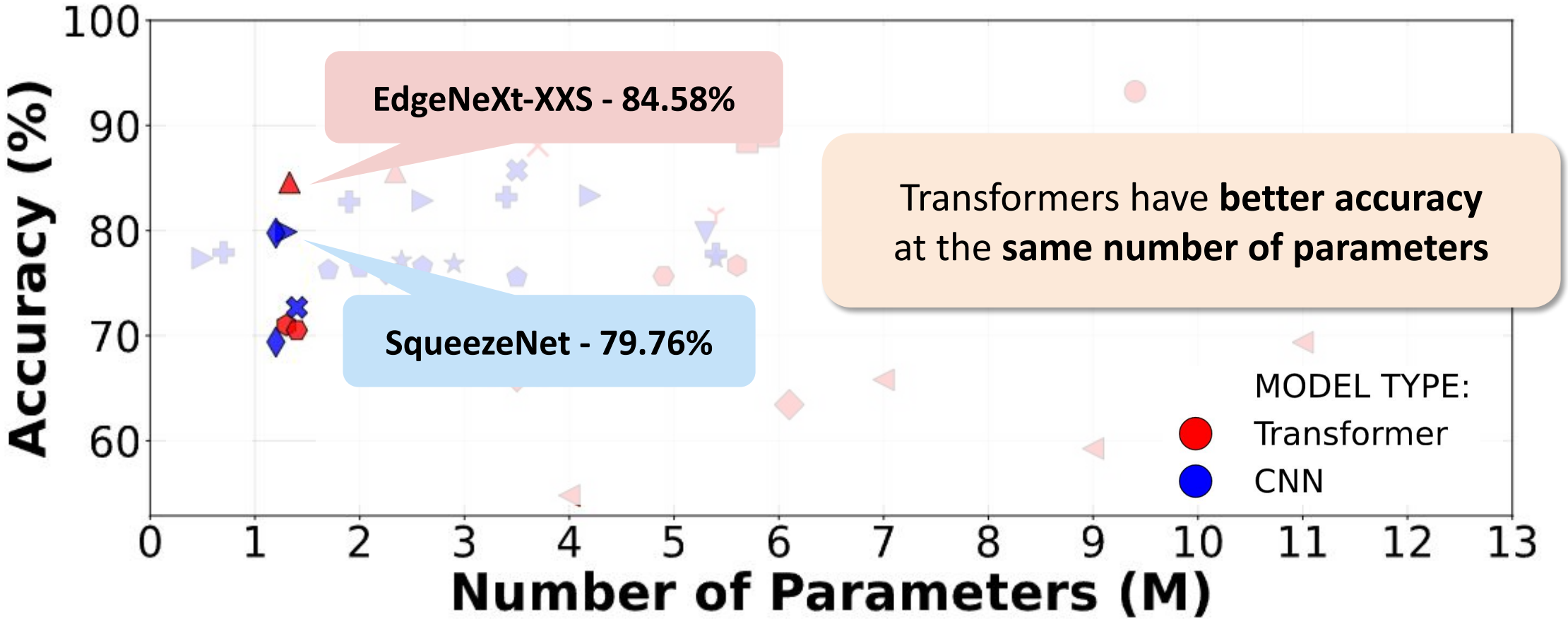


Fig 7: CIFAR-10 - Number of Parameter / Accuracy @ 224x224 input

Results - Computation at the Same Accuracy

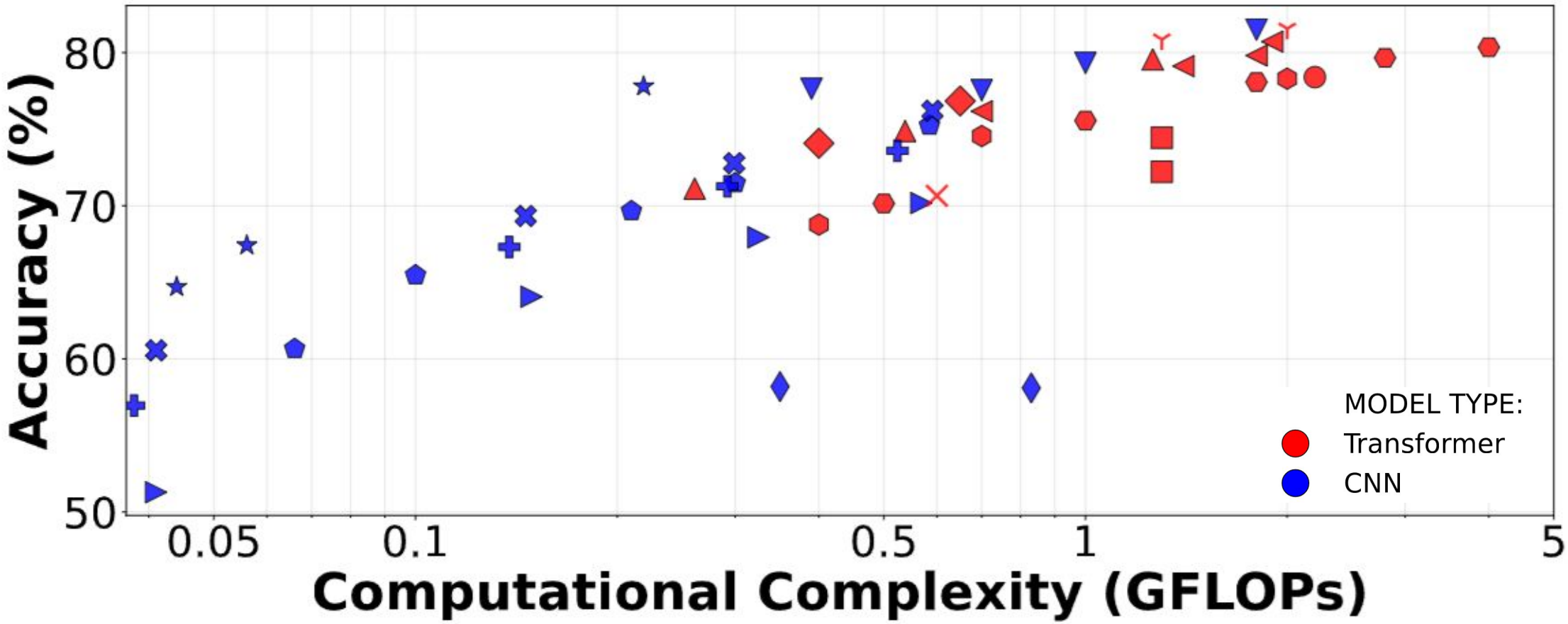
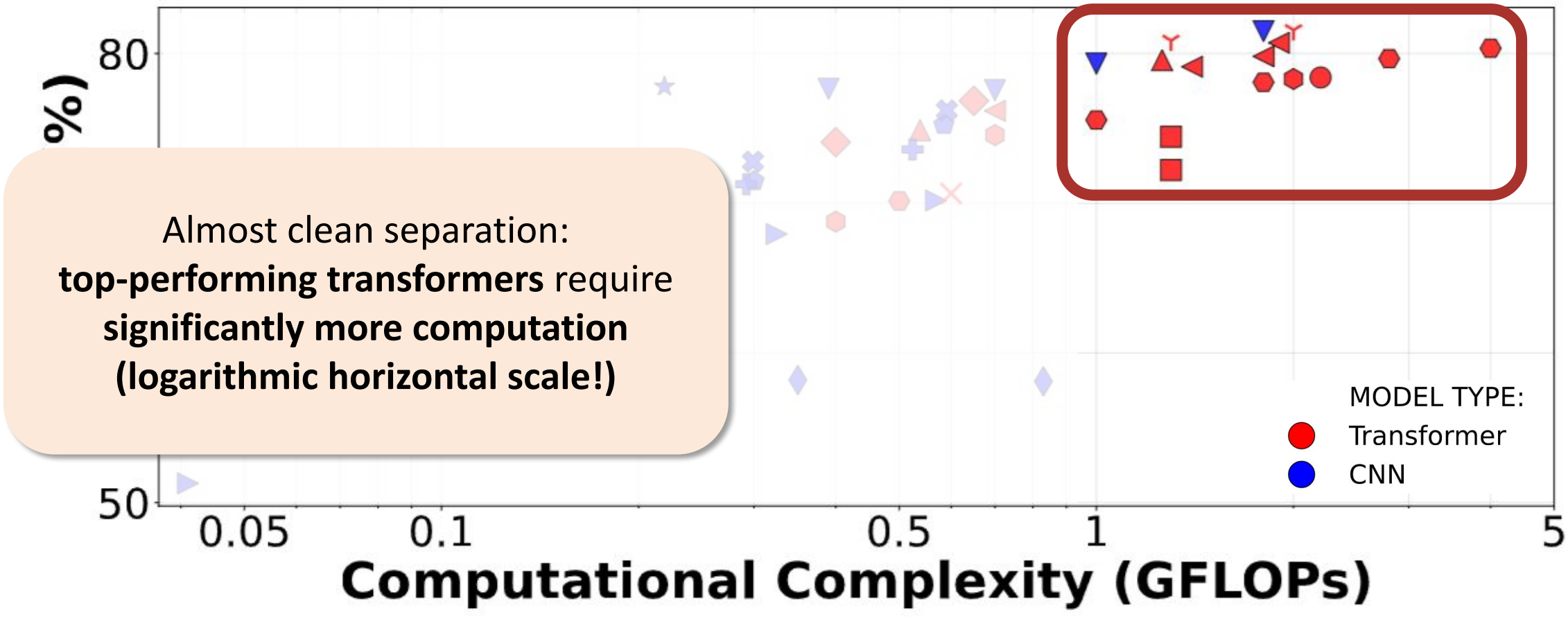


Fig 9: ImageNet - Computational Complexity / Accuracy @ 224x224 input

Results - Computation at the Same Accuracy



Almost clean separation:
top-performing transformers require
significantly more computation
(logarithmic horizontal scale!)

Fig 9: ImageNet - Computational Complexity / Accuracy @ 224x224 input

Results - Computation at the Same Accuracy

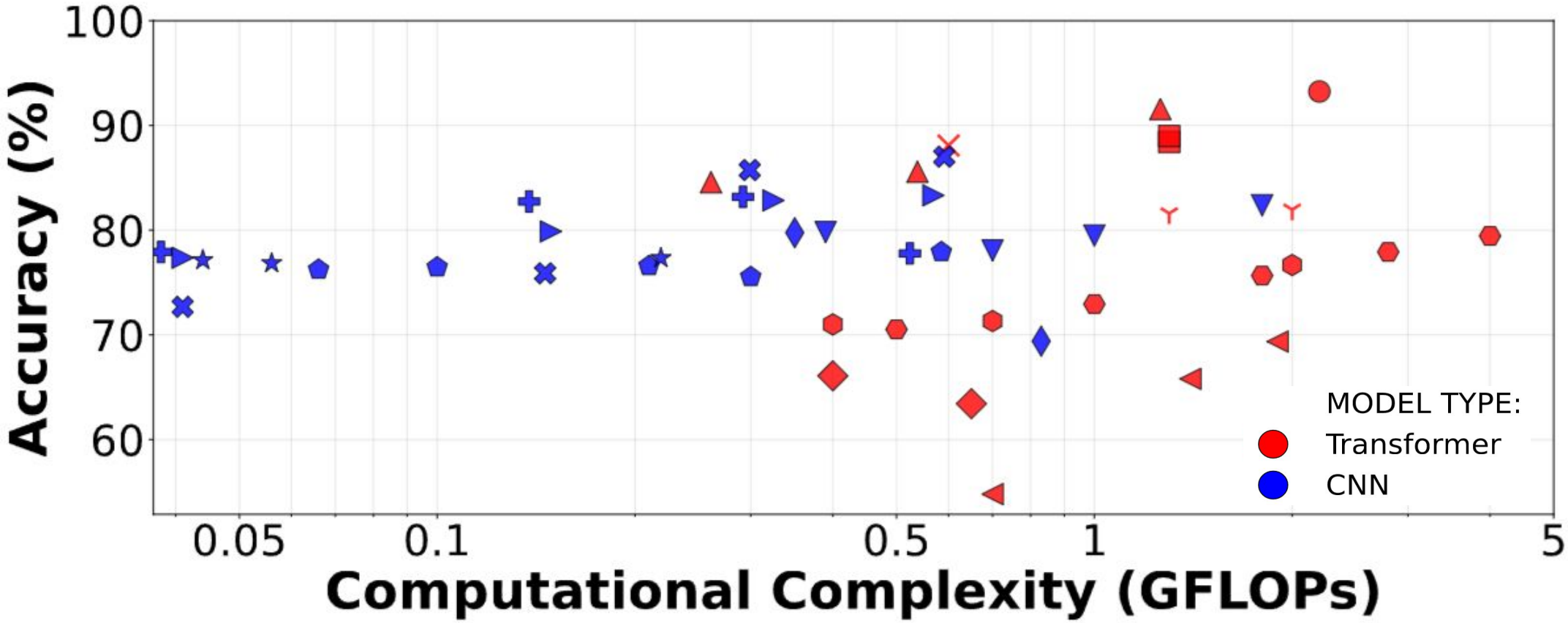


Fig 10: CIFAR-10 - Computational Complexity / Accuracy @ 224x224 input

Results - Computation at the Same Accuracy

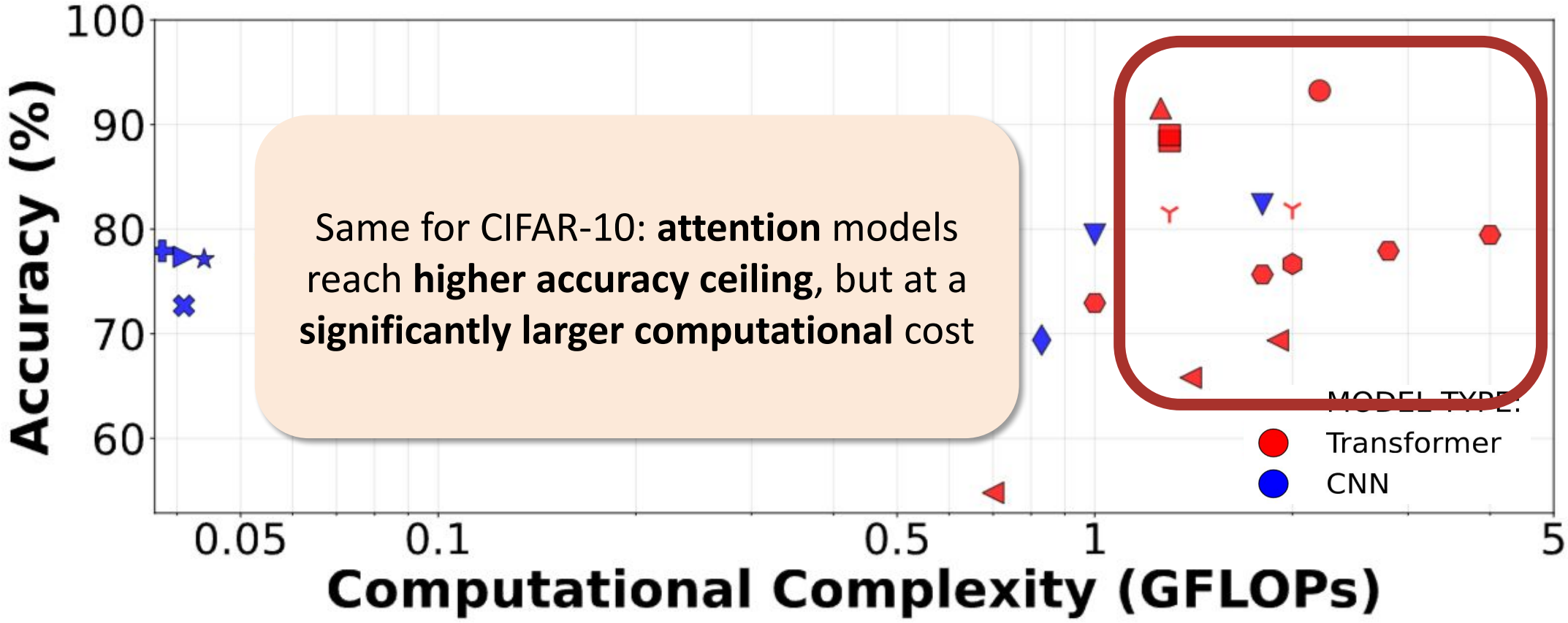


Fig 10: CIFAR-10 - Computational Complexity / Accuracy @ 224x224 input

Results - Computation at the Same Accuracy

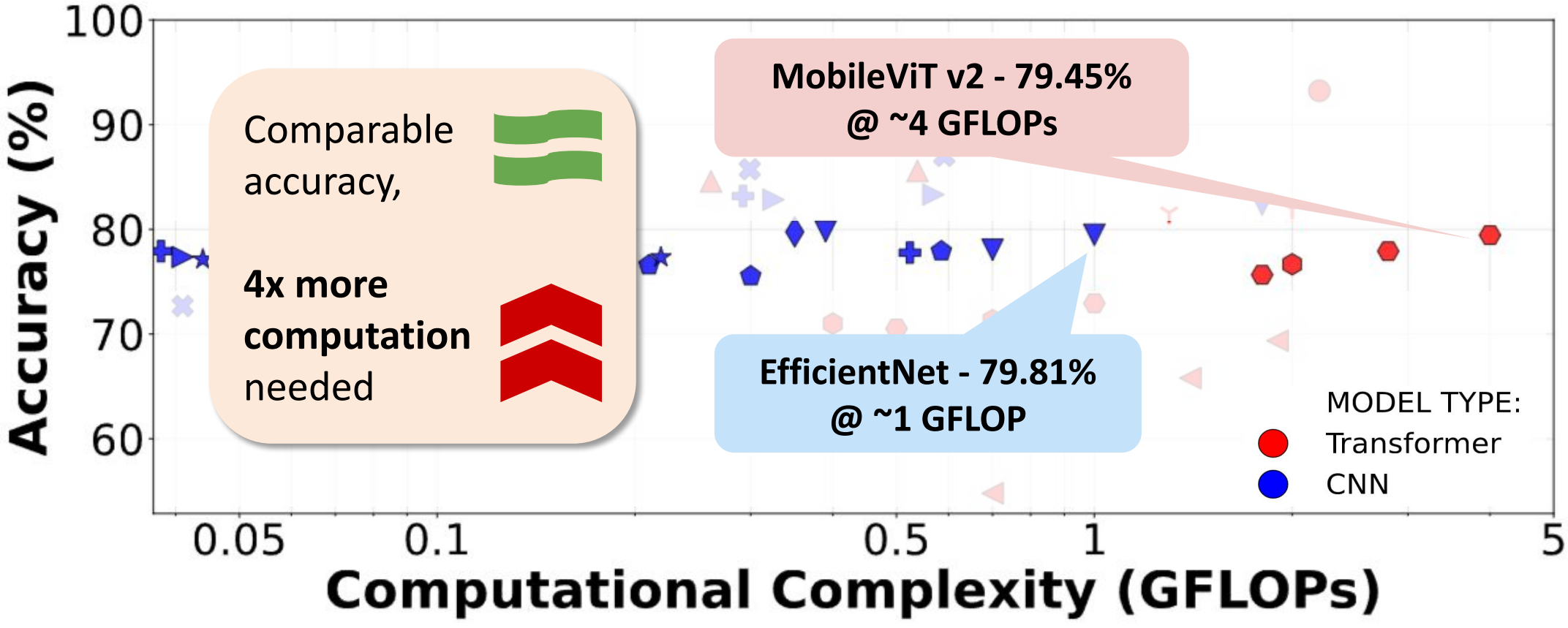


Fig 10: CIFAR-10 - Computational Complexity / Accuracy @ 224x224 input

Results - Computation at the Same Accuracy

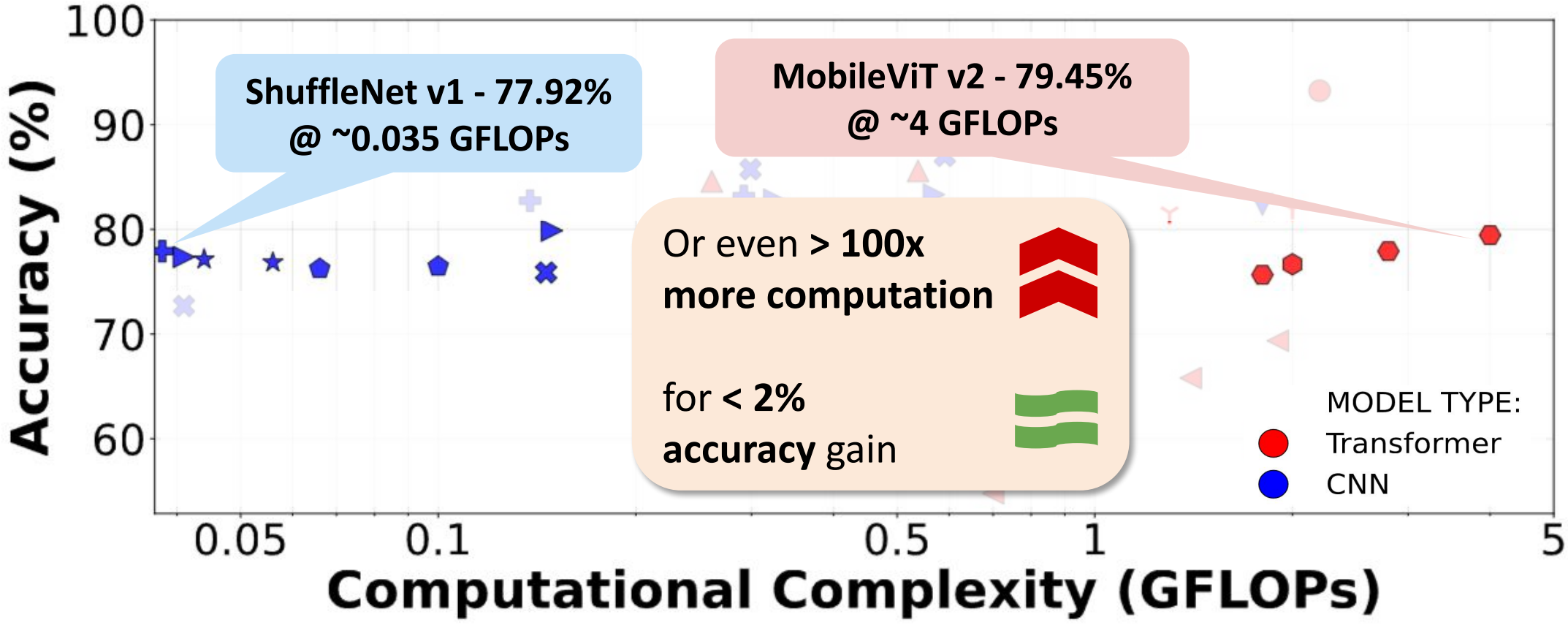


Fig 10: CIFAR-10 - Computational Complexity / Accuracy @ 224x224 input

Results - Resolution Shift Robustness

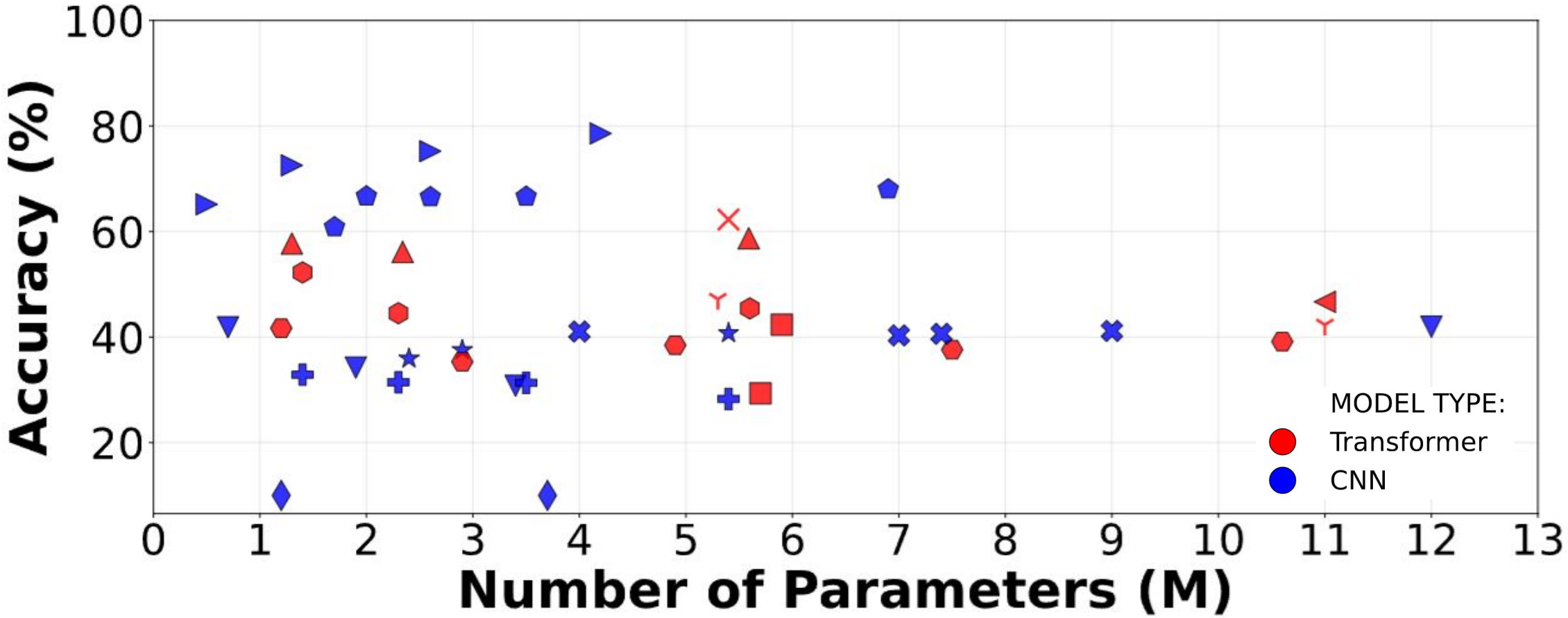


Fig 11: CIFAR-10 - Num. of Params / Accuracy @ **32x32** input

Results - Resolution Shift Robustness

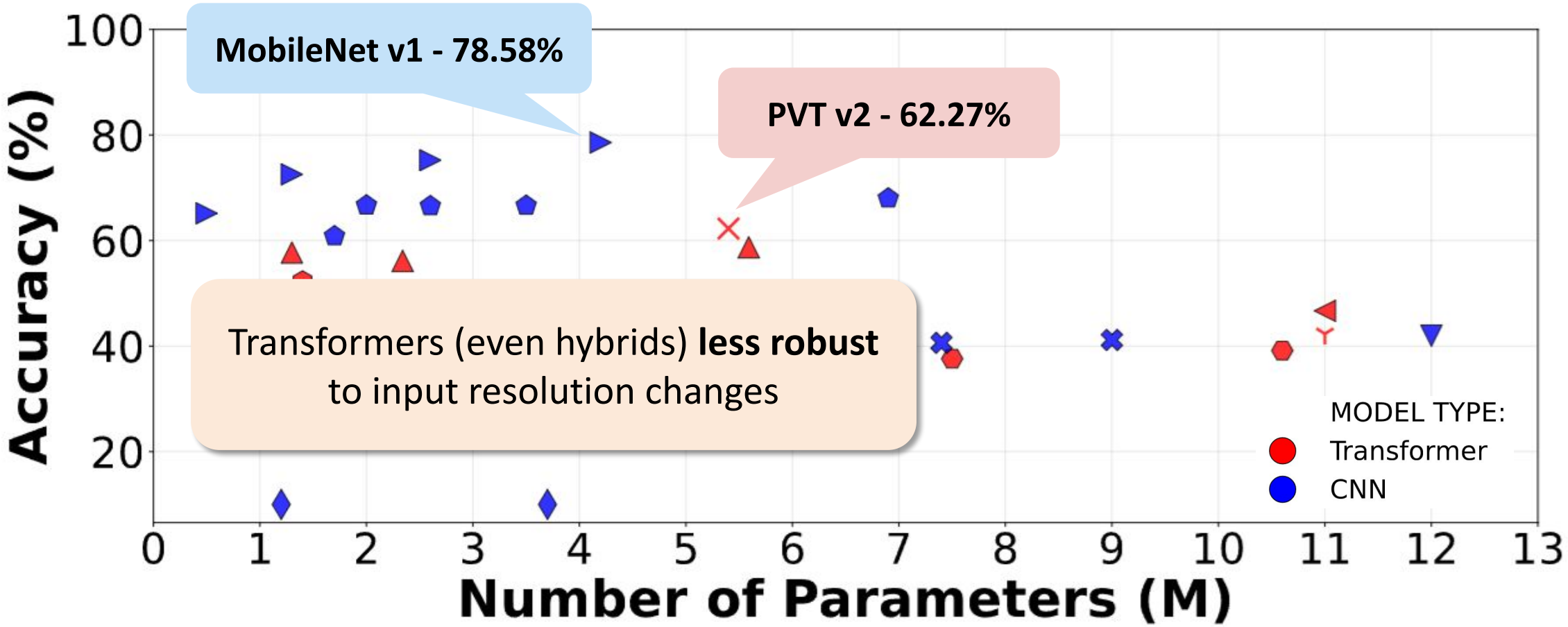
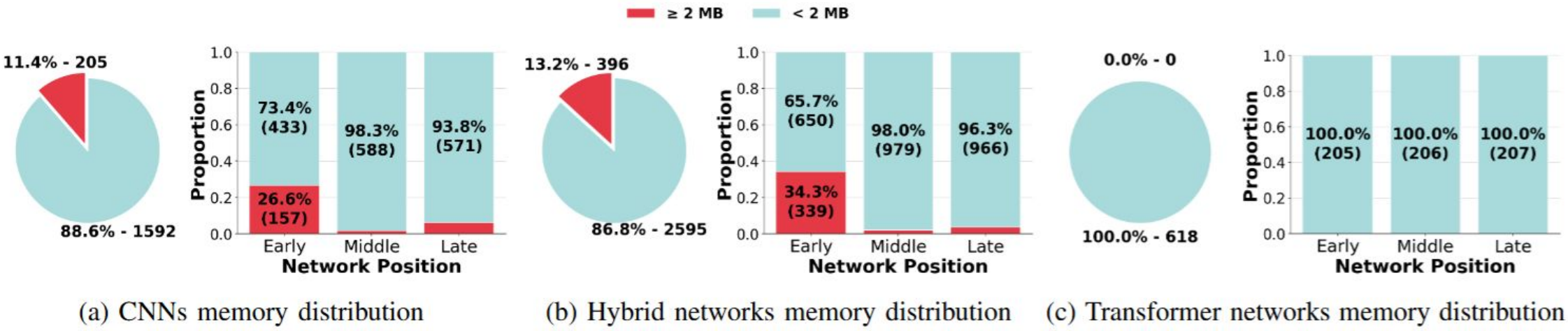


Fig 11: CIFAR-10 - Num. of Params / Accuracy @ **32x32** input

Results - Working Layer Memory Distribution

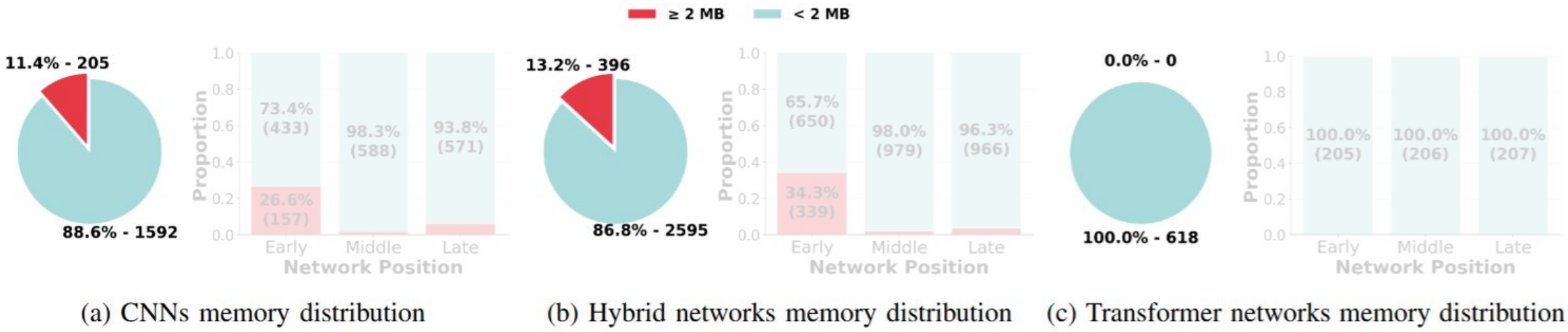


(a) CNNs memory distribution (b) Hybrid networks memory distribution (c) Transformer networks memory distribution

Fig 12: Layer-wise distribution of required working memory @ 224x224 input

Results - Working Layer Memory Distribution

11.4% layers of CNN models and **13.2% of hybrid** ones **exceed the 2 MB limit**
Pure transformer models - 0%!

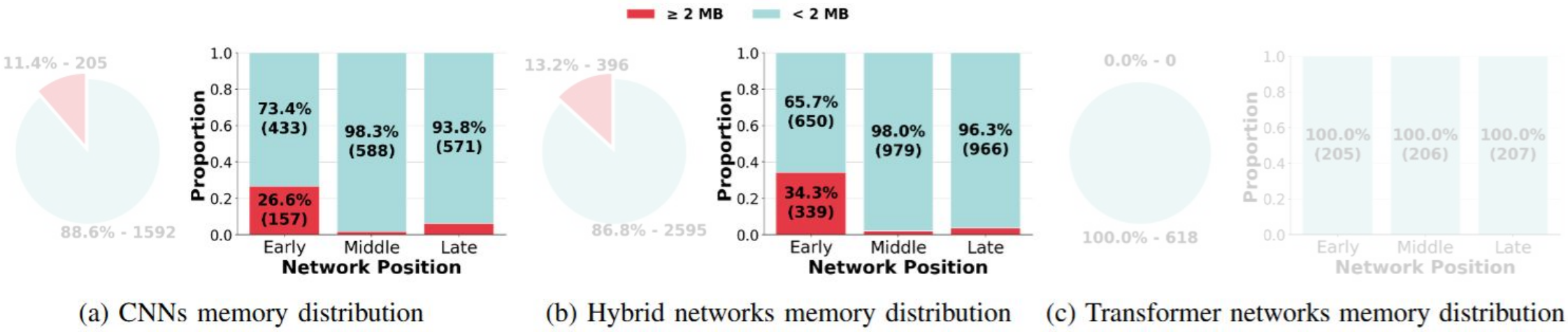


(a) CNNs memory distribution (b) Hybrid networks memory distribution (c) Transformer networks memory distribution

Fig 12: Layer-wise distribution of required working memory @ 224x224 input

Results - Working Layer Memory Distribution

CNN and hybrid architectures exhibit a **less uniform memory footprint along depth**, with pronounced **peaks**, especially in the **early layers**

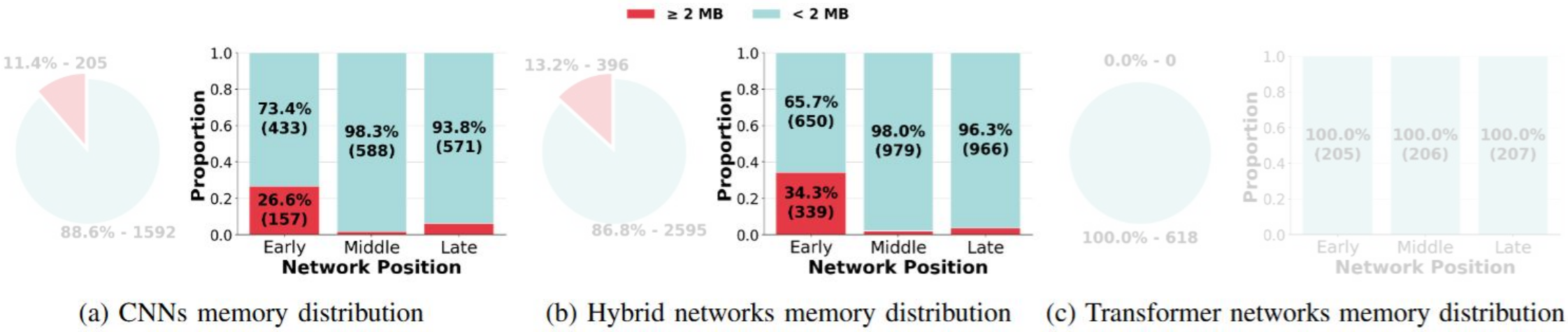


(a) CNNs memory distribution (b) Hybrid networks memory distribution (c) Transformer networks memory distribution

Fig 12: Layer-wise distribution of required working memory @ 224x224 input

Results - Working Layer Memory Distribution

This **regularity** makes **transformers favorable** to deployment on **memory-constrained platforms**, as it enables **buffer reuse** and **simpler memory management**



(a) CNNs memory distribution (b) Hybrid networks memory distribution (c) Transformer networks memory distribution

Fig 12: Layer-wise distribution of required working memory @ 224x224 input

Results - Deployment Latency Analysis

Model	Working Mem. (MiB)	Cycles (G)	FLOPS (M)	ImageNet Acc.
CaiT-XXS-24	**1.28**	9.41	2,500	**0.78**
EdgeNeXt-XS	1.93	4.72	538	0.75
DeiT-Ti	1.29	40.43	1,300	0.74
EdgeNeXt-XXS	1.44	2.06	261	0.71
MobileNet v3 Small 0,75x	1.69	1.46	44	0.65
ShuffleNetV1 0,5x (g=3)	1.84	**1.37**	**41**	0.57
MobileNet v1 0,25x	1.53	1.55	44	0.51

Fig 13: Model memory and computational cost @ 224×224 input

Results - Deployment Latency Analysis

Model	Working Mem. (MiB)	Cycles (G)	FLOPS (M)	ImageNet Acc.
CaiT-XXS-24		9.41	2,500	0.78
EdgeNeXt-XS		4.72	538	0.75
DeiT-Ti		40.43	1,300	0.74
EdgeNeXt-XXS		2.06	261	0.71
MobileNet v3 Sn		1.46	44	0.65
ShuffleNetV1 0,5		1.37	41	0.57
MobileNet v1 0,2		1.55	44	0.51

FLOPs don't tell the full story:

CaiT is 4× faster than DeiT, even though DeiT reports about **half the FLOPs**

Fig 13: Model memory and computational cost @ 224×224 input

Results - Deployment Latency Analysis

Model	Working Mem. (MiB)	Cycles (G)	FLOPS (M)	ImageNet Acc.
CaiT-XXS-24		9.41	2,500	0.78
EdgeNeXt-XS		4.72	538	0.75
DeiT-Ti		40.43	1,300	0.74
EdgeNeXt-XXS		2.06	261	0.71
MobileNet v3 Sn		1.46	44	0.65
ShuffleNetV1 0,5		**1.37**	**41**	0.57
MobileNet v1 0,2		1.55	44	0.51

CaiT - 61× more FLOPs reported than **ShuffleNet**, but only **~7x fewer cycles**

Thus, **target platform evaluation** necessary for true picture

Fig 13: Model memory and computational cost @ 224×224 input

Results - Deployment Latency Analysis

Model	Working Mem. (MiB)	Cycles (G)	FLOPS (M)	ImageNet Acc.
CaiT-XXS-24		9.41	2,500	**0.78**
EdgeNeXt-XS		4.72	538	0.75
DeiT-Ti		40.43	1,300	0.74
EdgeNeXt-XXS		2.06	261	0.71
MobileNet v3 Sn		1.46	44	0.65
ShuffleNetV1 0,5		**1.37**	**41**	0.57
MobileNet v1 0,2		1.55	44	0.51

Slowest CNN >1.3x faster than **fastest attention** model

Attention models may benefit in the future from **better hw-sw co-optimization**

Fig 13: Model memory and computational cost @ 224×224 input

Conclusions:

- **Transformers win on accuracy** (up to +6.67%) and **memory predictability**, but require **more cycles**
- Resolution **kills the advantage**: at 32×32, **CNNs beat transformers** by 16.4%
- **FLOPs can be misleading**, making **on-device profiling mandatory** for choosing an architecture

PULP Platform
Open Source Hardware, the way it should be!

Thank you!

pulp-platform.org
@pulp_platform
company/pulp-platform
youtube.com/pulp_platform

PSI, Katholieke Universiteit Leuven
Kasteelpark Arenberg 10
Leuven, Belgium

DEI, Università di Bologna
Viale del Risorgimento 2
Bologna, Italy

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

KU LEUVEN

Bibliography

- [1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, 25.
- [2] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
- [3] Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
- [4] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., ... & Adam, H. (2019). Searching for MobileNetV3. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)
- [5] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
- [6] Xu, J., Pan, Y., Pan, X., Hoi, S., Yi, Z., & Xu, Z. (2021). RegNet: Self-Regulated Network for Image Classification. IEEE Transactions on Neural Networks and Learning Systems
- [7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS)
- [8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations (ICLR)
- [9] Touvron, H., Cord, M., Sablayrolres, A., Synnaeve, G., & Jégou, H. (2021). Going Deeper with Image Transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)
- [10] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)
- [11] Mehta, S., & Rastegari, M. (2022). MobileViT: Light-Weight, General-Purpose, and Mobile-Friendly Vision Transformer. International Conference on Learning Representations (ICLR)
- [12] Dequino, A. (2024). Convolutional Neural Networks on Ultra-Low-Power Devices. PhD Thesis, University of Bologna, Department of Electrical, Electronic, and Information Engineering "Guglielmo Marconi."
- [13] Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. M., & Shum, H.-Y. (2022). DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. arXiv preprint arXiv:2203.03605
- [14] PULP Platform. Parallel Ultra Low Power Platform [Online]. Available: https://pulp-platform.org/ [Accessed: Jun. 2026]