DNNShifter: Compressing Large Neural Networks for Edge Systems

Bailey Eccles and Blesson Varghese

bje1@st-andrews.ac.uk

Seventh Annual UK Systems Research Challenges Workshop

Edge Computing

- **~175 ZB** of data by 2025 and **~25 billion** IoT devices by 2030
- Computation and data moved towards the edge of the network
- Reduced latency, bandwidth, and energy consumption
- Ideal for real-time, and privacy preserving applications

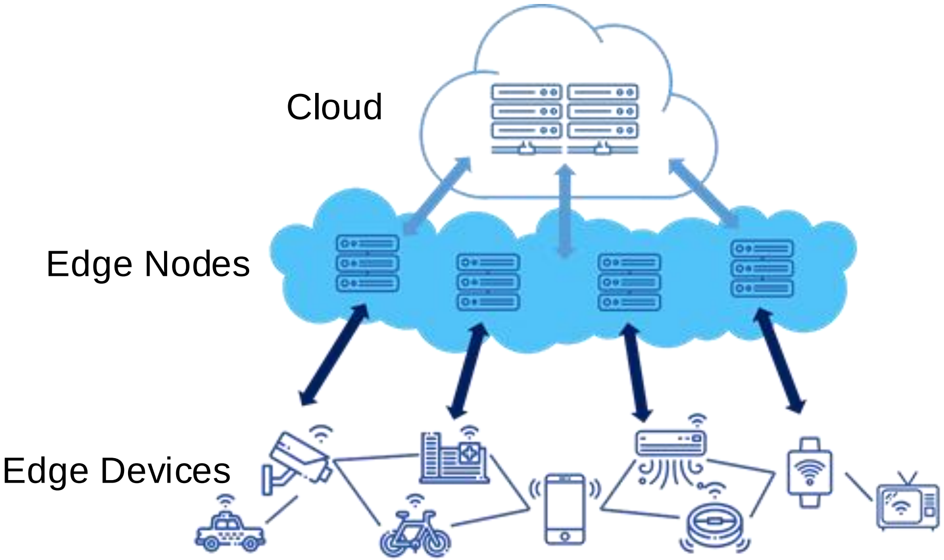
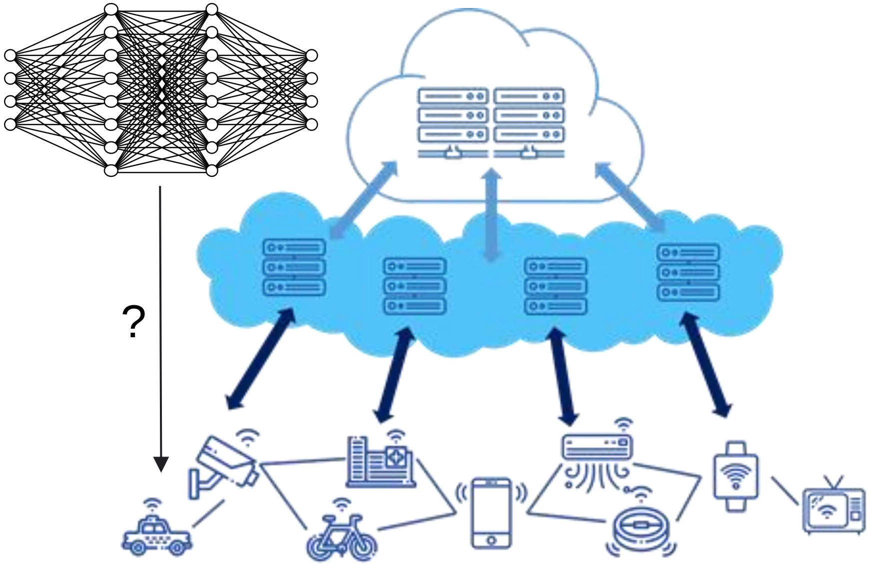


Image: alibabacloud.com/knowledge/what-is-edge-computing

Edge Machine Learning

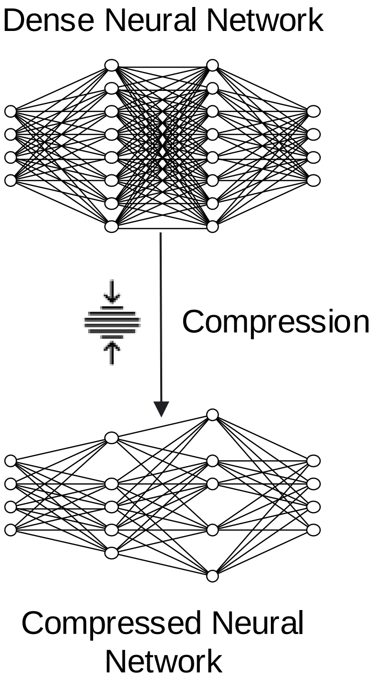
- Neural networks are designed for cloud resources
- Neural network training relies on large over-parameterization models and hardware accelerators (e.g. GPUs)
- **Compute**, **memory**, and **energy** constraints of edge devices limit deployment to the edge



Neural Network Compression

- Reduces network complexity
- Improves runtime performance
- Degrades model accuracy
- Many existing methods
  - Neural Architecture Search (NAS)
  - Quantisation
  - Distillation
  - **Pruning**

Accuracy: 93%	
Latency: 100ms	
Size: 500MB	



Accuracy: 91%	
Latency: 40ms	
Size: 150MB	

System Challenges

How can we automate and find suitable smaller neural networks **quickly**?

- NAS can take up to **2000 GPU days**

How can we **preserve** accuracy in compressed neural networks?

- Risk of neural network **collapse**

How can we **adapt** neural networks to changing operational conditions?

- Edge runtime conditions **frequently change**

Solution

DNNShifter: An Efficient DNN Pruning System

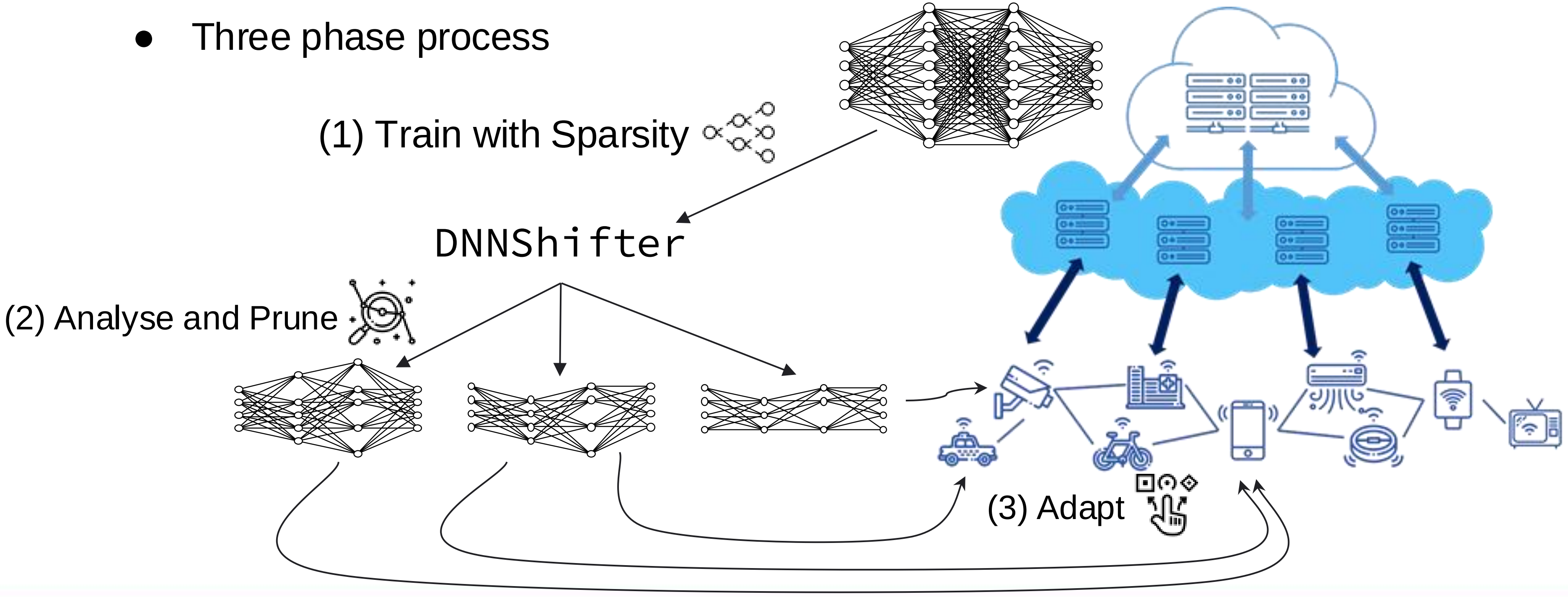
Bailey J. Eccles, Philip Rodgers, Peter Kilpatrick, Ivor Spence, and Blesson Varghese

- Lightweight framework for converting cloud neural networks into a range of compressed edge deployable neural networks
- Under review, IEEE Internet of Things Journal
- Funded by Rakuten Mobile, Japan
  (Patent: Edge-Masking Guided Node Pruning. US2022/053590)

Rakuten Mobile

DNNShifter

- Three phase process



Types of Neural Network Pruning

- Unstructured Pruning
  - **Maintains (most of) Accuracy**
  - **No runtime improvements**

- Structured Pruning
  - **Degrades Accuracy**
  - **Runtime improvements**

DNNShifter combines unstructured and structured pruning to leverage the benefits of both methods.

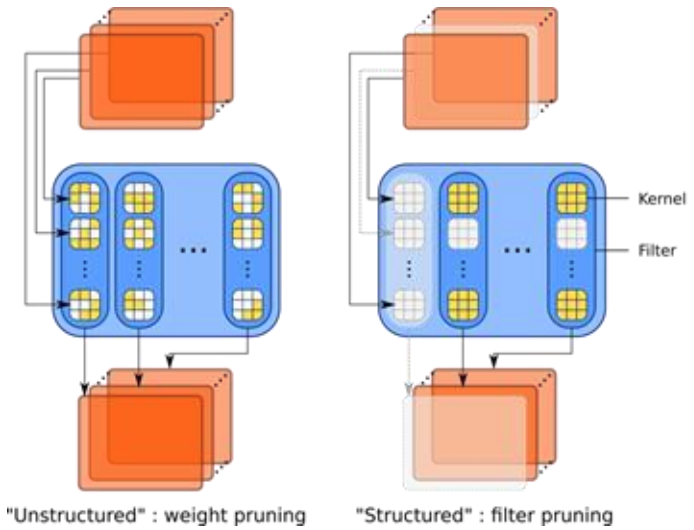
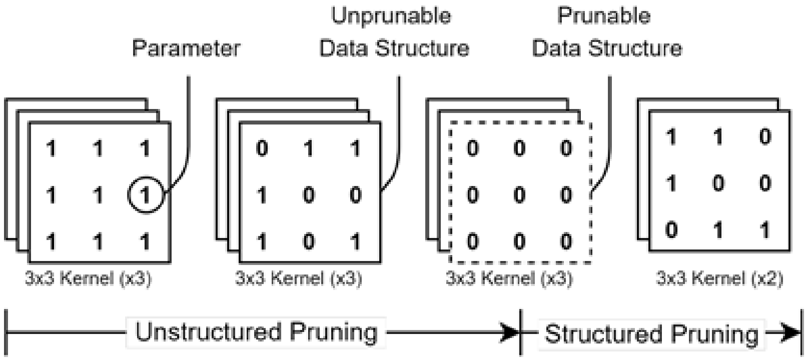


Image: towardsdatascience.com/neural-network-pruning-101-af816aaea61

Lossless Structured Pruning

1. Train with Sparsity (**Unstructured Pruning**) [1]
2. Identify **structured pruning** opportunities and prune (takes less than **200ms**)
3. Create a range of neural networks with different degrees of both pruning types



[1]: Frankle, J. and Carbin, M., 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *ICLR*.

Image: Eccles, B. J., Rodgers, P., Kilpatrick, P., Spence, I. and Varghese, B., 2023. DNNShifter: An Efficient DNN Pruning System. *IEEE Internet of Things Journal* (Under review).

Training Time

System	Method	Params. Trained (M)	GPU-days	Accuracy (%)
DNNShifter	Pruning	**132.30**	**0.28**	93.25 ± 0.66
DARTS	NAS - Automatic	165.00	1.77	74.01 ± 16.9
RepVGG	NAS - Manual	12329.00	26.09	94.96 ± 0.16

- **93x** faster than exhaustive manual searches (RepVGG)
- **6.3x** faster than accelerated NAS methods (DARTS)
- **315x** faster than legacy NAS methods (Google's NASNet)

Table (adapted from): Eccles, B. J., Rodgers, P., Kilpatrick, P., Spence, I. and Varghese, B., 2023. DNNShifter: An Efficient DNN Pruning System. *IEEE Internet of Things Journal* (Under review).

Runtime Performance

- Up to **1.67x** faster (CPU)
- Up to **1.42x** faster (GPU)
- Up to **5.14x** smaller (MB)
- No accuracy loss

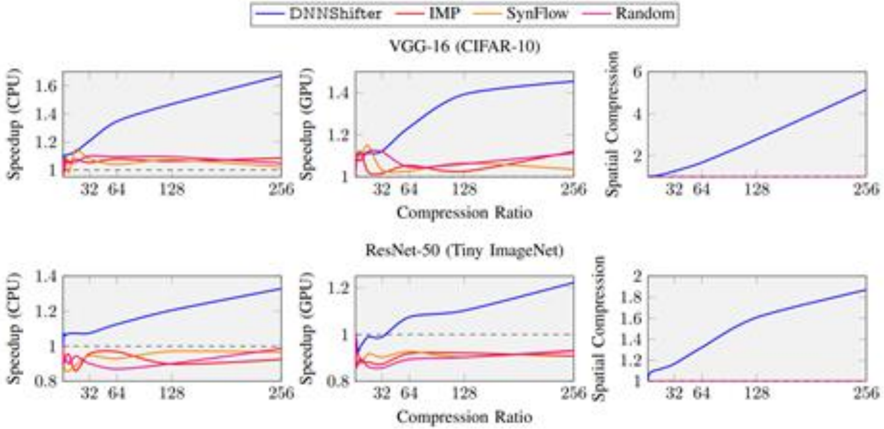


Figure: Eccles, B. J., Rodgers, P., Kilpatrick, P., Spence, I. and Varghese, B., 2023. DNNShifter: An Efficient DNN Pruning System. *IEEE Internet of Things Journal* (Under review).

Switching Neural Networks at Runtime

System	Mem. Util. (MB)	Decision Overhead (ms)
DNNShifter	**47.0 ± 15.5**	**43**
Model Ensemble	115.8	49
Dynamic-OFA	56.2	512

- Low as **0.5x** memory utilisation
- Low as **43ms** decision overhead

Table (adapted from): Eccles, B. J., Rodgers, P., Kilpatrick, P., Spence, I. and Varghese, B., 2023. DNNShifter: An Efficient DNN Pruning System. *IEEE Internet of Things Journal* (Under review).

University of St Andrews

Thank you and Questions

edgehub.co.uk

bje1@st-andrews.ac.uk
linkedin.com/in/bailey-eccles/

www.st-andrews.ac.uk