EFFICIENT EDGE INFERENCE BY SELECTIVE QUERY

ANIL KAG [1], IGOR FEDOROV[3], ADITYA GANGRADE [2], PAUL WHATMOUGH[3], VENKATESH SALIGRAMA [1]

1 BOSTON UNIVERSITY 2 Carnegie Mellon University 3 arm

ICLR 2023

ML Inference at Edge

- Problem Setup: ImageNet Classification on Edge

Tiny Models — MobileNetV3 — MCUNet

Edge Devices

Constraints

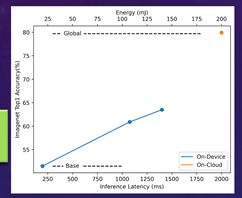
- ❑ Model Size
- ❑ Compute Usage
- ❑ Battery Drain
- ❑ Inference Latency

MCU

- ❑ 320KB SRAM
- ❑ 1MB Flash Storage

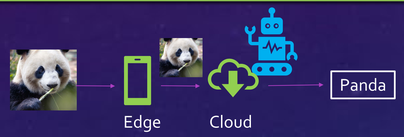
- Existing Methods: On-Device

✓ Deploy best performing model on device: MCUNet
✓ Include compression, quantization, distillation, NAS



- Existing Methods: On-Cloud

✓ Send inputs to cloud with high-capacity model: OFA
✓ Communication cost is very high

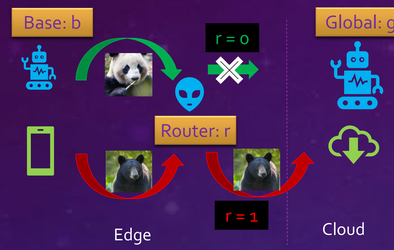


- Leverage Cloud Intelligence on Edge

✓ Depending on input hardness, use cloud or edge model

Hybrid Models

- Hybrid Model Setup



Coverage = $P(r(x) = 0)$

Accuracy = $P(r(x) = 0, b(x) = y) + P(r(x) = 1, g(x) = y)$

Optimization:
Maximize Accuracy
s.t. Coverage $\geq c$

- ❑ *r is computationally cheap*
- ❑ *r takes base features as input*
- ❑ *r measures base prediction confidence*

- ❑ *Abstaining classifier w/o Global*
- ❑ *High coverage → Most predictions on Edge*
- ❑ *Low coverage → Most predictions on Cloud*

- Learning a Router given Pre-trained Base & Global

$$\max_r \; E[r(x)1_{g(x)=y} + (1 - r(x))1_{b(x)=y}] \quad \text{s.t.} \quad E[1 - r(x)] \geq c$$

$$\min_r \; \ell(r(x), o(x)) \quad \text{s.t.} \quad E[1 - r(x)] \geq c$$

$$\lambda \left(\frac{1}{N} \sum_x 1(r(x) = 0) - c \right)_+$$

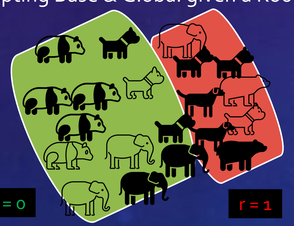
Without constraint

$$r(x) = \begin{cases} 0, & b(x) = y \\ 1, & b(x) \neq y, g(x) = y \\ ?, & b(x) = g(x) \neq y \end{cases}$$

Oracle

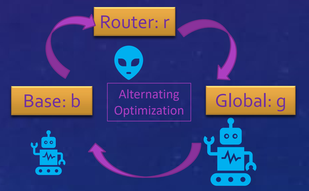
$$o(x) = 1_{b(x) \neq g(x) = y}$$

- Adapting Base & Global given a Router



$$\mathcal{L}_{base} = \sum (2 - r(x))\, \ell(b(x), y)$$

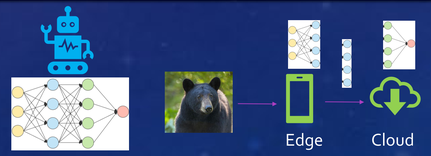
$$\mathcal{L}_{global} = \sum (1 + r(x))\, \ell(g(x), y)$$

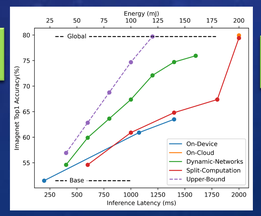


- Dynamic Baselines

Split Computation

✓ Split Global between edge & cloud: BranchyNet
✓ Communication cost is still high due to feature size



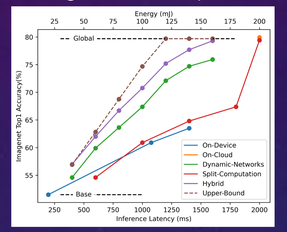


Dynamic Neural Networks

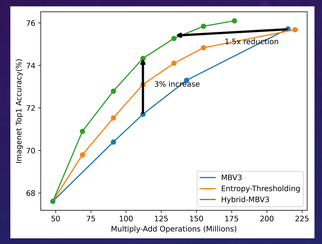
✓ Entropy threshold on edge: Adaptive Inference
✓ Still a significant gap between Edge & Cloud

Empirical Evaluation

- Edge-Cloud Setup



Same Device Setup



- Hybrid Models Pareto Dominate Baselines (Global = OFA)

MobileNetV3 Base MACs	Base Accuracy	Method	Accuracy @ 90% Coverage	Accuracy @ 70% Coverage
48M	67.6	Entropy	70.7	74.9
		Hybrid	71.6	76.8
143M	73.3	Entropy	75.1	77.6
		Hybrid	75.9	79.0
215M	75.7	Entropy	77.1	78.9
		Hybrid	77.6	79.6

- Hybrid Models Resource Usage on MCUs

Latency (ms)	1000	1400	1600	2000
On-Cloud	-	-	-	79.9
On-Device	60.9	63.5	-	-
Entropy	67.4	74.7	76.93	-
Hybrid	70.8	77.7	79.5	-

- Hybrid Models w/o Cloud : Abstaining Classifier

MobileNetV3 Base MACs	Base Accuracy	Accuracy @ 90% Coverage	Accuracy @ 80% Coverage	Accuracy @ 70% Coverage
48M	67.6	73.3	78.6	83.4
143M	73.3	79.0	83.9	88.4
215M	75.7	81.3	86.1	90.1

- https://github.com/anilkagak2/Hybrid_Models/