Distributed Inference at the Edge: Variant Parallelism

Navidreza Asadi, Chair of Communication Networks, Technical University of Munich
Maziar Goudarzi, Computer Engineering Department, Sharif University of Technology

6G-life

Distributed Edge AI

Why Deep Learning on Edge?

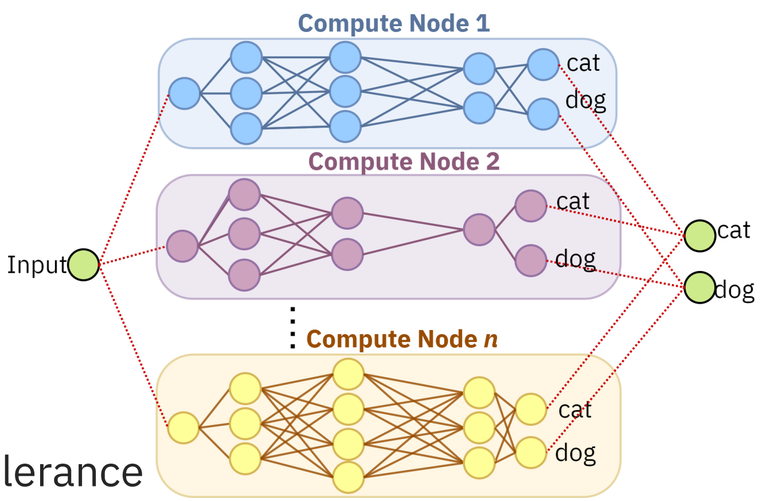
- Lower latency (processing near data sources)
- Data privacy concerns
- Higher availability

Why Distributed Inference?

- AI trends are towards more complicated tasks and larger deep learning models.
- Edge devices have limited compute, memory and storage resources.
- It may **not** be **feasible** to deploy the whole neural network model on a single device.
- **Faster response time** might be achievable by applying **parallelism** techniques across multiple compute nodes.

1

Variant Parallelism*



Main Objectives:

1. Higher fault-tolerance
2. Better parallelization across multiple (heterogenous) compute nodes
3. Maintaining accuracy as much as possible

How?

- We propose a bottom-up ensemble-based model distribution.
- Each variant is a lighter pruned version of the reference model that has its own characteristics.
- Variants are independent of each other and can produce complete prediction vectors.
- Each node returns only the top k (e.g., $k = 2$) elements of its prediction vector with the highest probability.
- Variants can be flexibly selected or generated based on a device characteristics.

3

Future Directions

- Distributed Inference in untrusted environments
  - E.g., Byzantine attacks
- Optimization of input data communication
- Automated generation of variants for more complicated tasks

5

References

[*] N. Asadi et al., “Variant Parallelism: Lightweight Deep Convolutional Models for Distributed Inference on IoT Devices” IEEE Internet of Things Journal, 2023.
[1] G. Wang et al., “sensAI: ConvNets Decomposition via Class Parallelism for Fast Inference on Live Data” MLSys ’21, 2021.

Contact: navidreza.asadi@tum.de

Need for a Better Parallelism

Model Parallelism

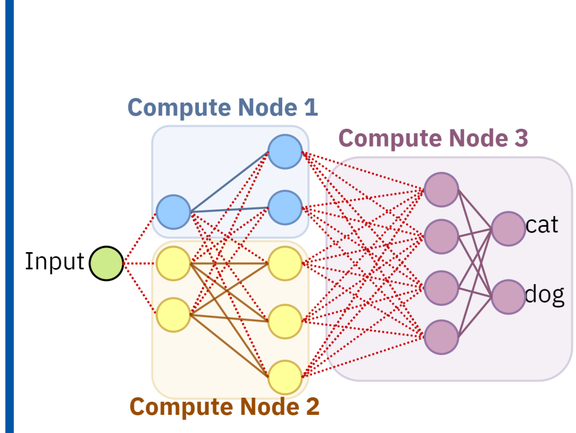
× Huge communication cost
× Multiple instances of single point of failure
× Inherent sequential blockages.

Class Parallelism

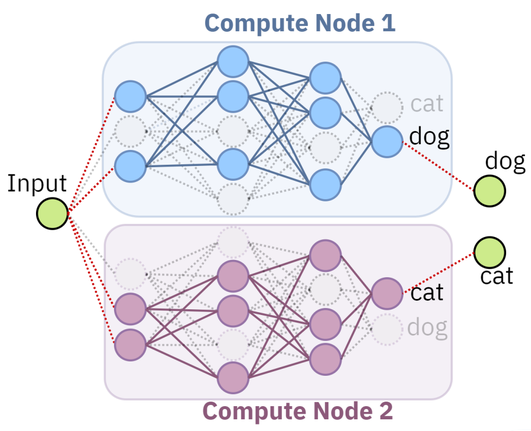
✓ Improved parallelism and communication cost
× Multiple instances of single point of failure
× Limited flexibility
× Homogeneity: all sub-models have similar characteristics.

Data Parallelism

× Atomic data can not be split into further pieces.



Model Parallelism



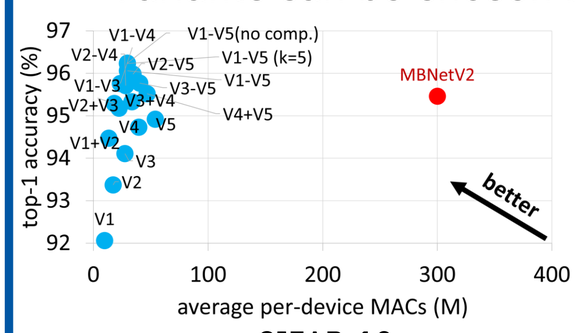
Class Parallelism

2

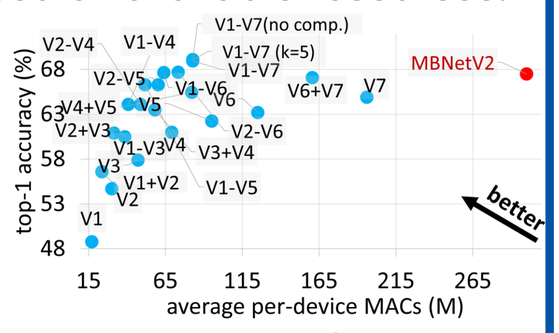
Results

Single Variant Accuracy and Scaling Accuracy

- Different accuracy vs. MACs can be achieved using different combinations of variants.
- Variants can be chosen based on available resources.



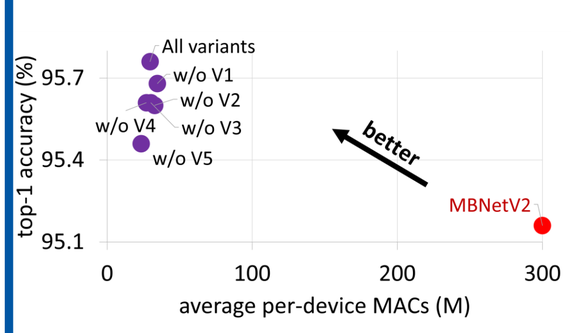
CIFAR-10



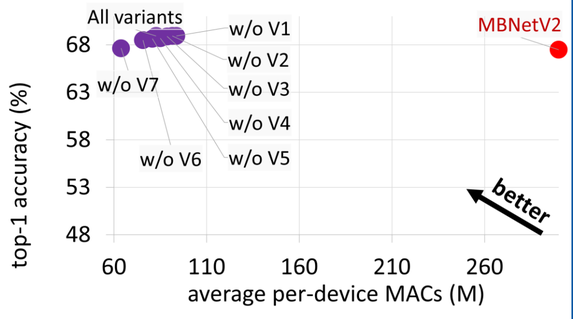
ImageNet-1k

Impact of Each Variant

- Each point shows aggregated accuracy of all variants when omitting one of them.



CIFAR-10



ImageNet-1k

Variants Performance Characteristics

Model	Response Time (ms)	Speedup	MACs Gain	Params Gain
MobileNetV2 (Baseline)	226	1x	1x	1x
V_5	**71**	**3.2x**	**5.6x**	**6.1x**
V_4	**53**	**4.2x**	**7.6x**	**6.3x**
V_3	**38**	**5.9x**	**10.9x**	**6.7x**
V_2	**26.5**	**8.6x**	**17.5x**	**6.9x**
V_1	**17**	**13.2x**	**31x**	**7.1x**
senAI[1]	-	2x	3.5x	5x

4