**Background and Motivation**

- Deep neural networks’ (DNNs) high performance comes with **large DNNs and powerful computers**
  - Powerful computers + Large DNNs + High DNN performance
- Deep neural networks (DNNs) are **costly**:
  - Prohibitive training cost:
    - 10^4 FLOPs for training ResNet-50@ImageNet
  - Excessive inference cost:
    - 10^4 FLOPs for single-image inference with ResNet-50@ImageNet

Growing demand for on-device training & inference: Large DNNs with prohibitive training & inference cost

**Low-precision Method:** a promising direction to narrow the gap

**Existing Low-precision Methods**

- **Static low-precision training**: [J. Rennie, NeurIPS'18]
  - Use same precision during training process
  - Large accuracy gap under low-precision
- **Temporal dynamic low-precision training**: A promising direction [Y. Fu, NeurIPS'20] [Y. Fu, ICLR'21]
  - Assign different precisions for different training stages for better accuracy-efficiency trade-off
  - Only consider **temporal** dynamic precision
  - Need extra efforts in hyperparams finetuning

**Motivating Observations**

- Is only the **temporal** dynamic precision enough?
- Inspirations from previous works:
  - Different layers have different sensitivities [Y. Fu, KDD'17]
  - Precision has similar effect as learning rate [Y. Fu, ICLR'20]

**Spatial dynamic precision allocation is also important**

**Exploration on the importance of **spatial and temporal** precision allocation**

- Settings:
  - **Spatial**: Assign a,b,c-bit to first three blocks, respectively
  - **Temporal**: Change precision at 30, 60, 90 epochs
- Insights:
  - Both **temporal** and **spatial** precision allocations impact the training accuracy-efficiency trade-off.
  - Different combination lead to 0.75% accuracy gap.

**How to automatically generate the **spatial and temporal** precision allocation during training?**

- Learnable dynamic precision (LDP): a framework to automatically learn the **spatial and temporal** precision allocation during training

**Contributions**

- Learnable dynamic precision (LDP): a framework to automatically learn the **spatial and temporal** precision allocation during training
- Develop a differentiable method to enable **end-to-end** learnable dynamic precision DNN training
- Achieve the SOTA accuracy-efficiency trade-off on seven DNNs, five datasets and three tasks in both training and inference

**LDP: Learning Dynamic Precision for Efficient Deep Neural Network Training and Inference**

Zhongzhi Yu, Yonggan Fu, Shang Wu, Mengquan Li, Haoran You, Yingyan Lin
Rice University

**LDP: Method**

- Automatically learn the **spatial and temporal** precision allocation during training
- **Enabler 1**: Differentiable learnable precision
  - Challenge: How to achieve a differentiable precision learning on top of the discrete precision
  - Vanilla quantization process:
    - Quantization Output = Roundn
    - Quantization Step = \( \text{Precision} / \text{Number of quantization steps} \)
    - Use a learnable quantization step with a layer-wise learnable parameter \( \beta \)

**LDP: Evaluation**

- **Seven models on five datasets from three tasks**:
  - ResNet@CIFAR for image classification
  - ResNet8/DeiT-Tiny@ImageNet for image classification
  - PAN@Urban-100 for image super-resolution
  - Transformer@Wiki-101 for language modeling
- **Three baselines**:
  - Static low-precision training: SBM [Y. Fu, NeurIPS'18]
  - Dynamic low-precision training: PFQ [Y. Fu, NeurIPS'19]
  - & CPT [Y. Fu, ICLR'20]

**LDP: Visualization**

- **Vis. 1**: Learned precision is consistent with manual design
  - Strong, APA2020/F Wang, SP'20
  - Higher precision in
    - Blocks after downsampling
    - Deep blocks with lowest spatial resolution
- **Vis. 2**: Learned precision can guide model design
  - Decreased precision (higher redundancy) in the last two FC layers
  - Consistent with the work studying FC layers [G. Gra, arXiv'17]