CHARACTERIZING THE PERFORMANCE OF ACCELERATED JETSON EDGE DEVICES FOR TRAINING DEEP LEARNING MODELS

KEDAR DHULE
(Research Fellow)
IISC Bengaluru

PROBLEM STATEMENT

TO ADDRESS THE CHALLENGES AND LIMITATIONS OF TRAINING DEEP NEURAL NETWORKS (DNNS) ON EDGE COMPUTING DEVICES,SPECIFICALLY FOCUSING ON NVIDIA JETSON DEVICES.

RESEARCH AIM

- Through previous work we address that the training of DNNs on the edge is not well-explored. So research aim is to investigate and understand the complexities of training Deep Neural Networks in edge computing environments .

- The Research aim to accurate modelling on training time and energy usage of edge devices by applying various training strategies and changing in configuration of devices .

- The principal aim is to establish the viability of using empirical data and simple regression techniques to make a model for predicting training time and energy consumption for DNN architecture.

- The proposed research aims to conduct a principled study of DNN training on three Jetson device types (AGX Xavier, Xavier NX, and Nano) using DNN models such as Lenet-5,MobileNet v3 ,ResNet-18 and VGG-11 on respective datasets.

- Investigate the variation of factors which having impact on training time and energy usage for DNN training such as
  a) Hardware Platforms
  b) Hardware configuration
  c) Software configuration
  d) Training techniques

- So that we can predict both the time and energy required to train a DNN model for any custom configuration of a given device . This prediction can assist in selecting the best configuration to balance time and energy trade-offs during model training

ADDRESSING KEY CHALLENGES

1. CPU and GPU utilisation

Maximizing performance and efficiency in DNN training on Jetson devices poses a challenge in effectively utilizing resources like GPU and CPU. Idle resources lead to increased end-to-end training time, which negatively impacts performance .

2. Optimizing Performance

Understanding the impact of factors such as caching, pipelining, and parallelism on the overall training time. It aims to find the right balance between CPU, GPU, and disk speeds, as well as the size of training data and model complexity .

3. Addressing Counter-Intuitive Results

Encountering scenarios where regular results may not hold for different models or configurations of the device . So the research aims to encourage a rethink of system design and tuning for DNN workloads on edge devices

4 .Suitable device for investigate

The suitability of a model and device configuration can significantly impact the efficiency of DNN training .Certain experiments can be done only with specific device or model due to limitation in other devices to give efficient result in training .

PURPOSE OF THIS RESEARCH

- These predictive models can assist in making informed decisions about device configuration and selection for DNN training, contributing to more efficient and sustainable training processes.

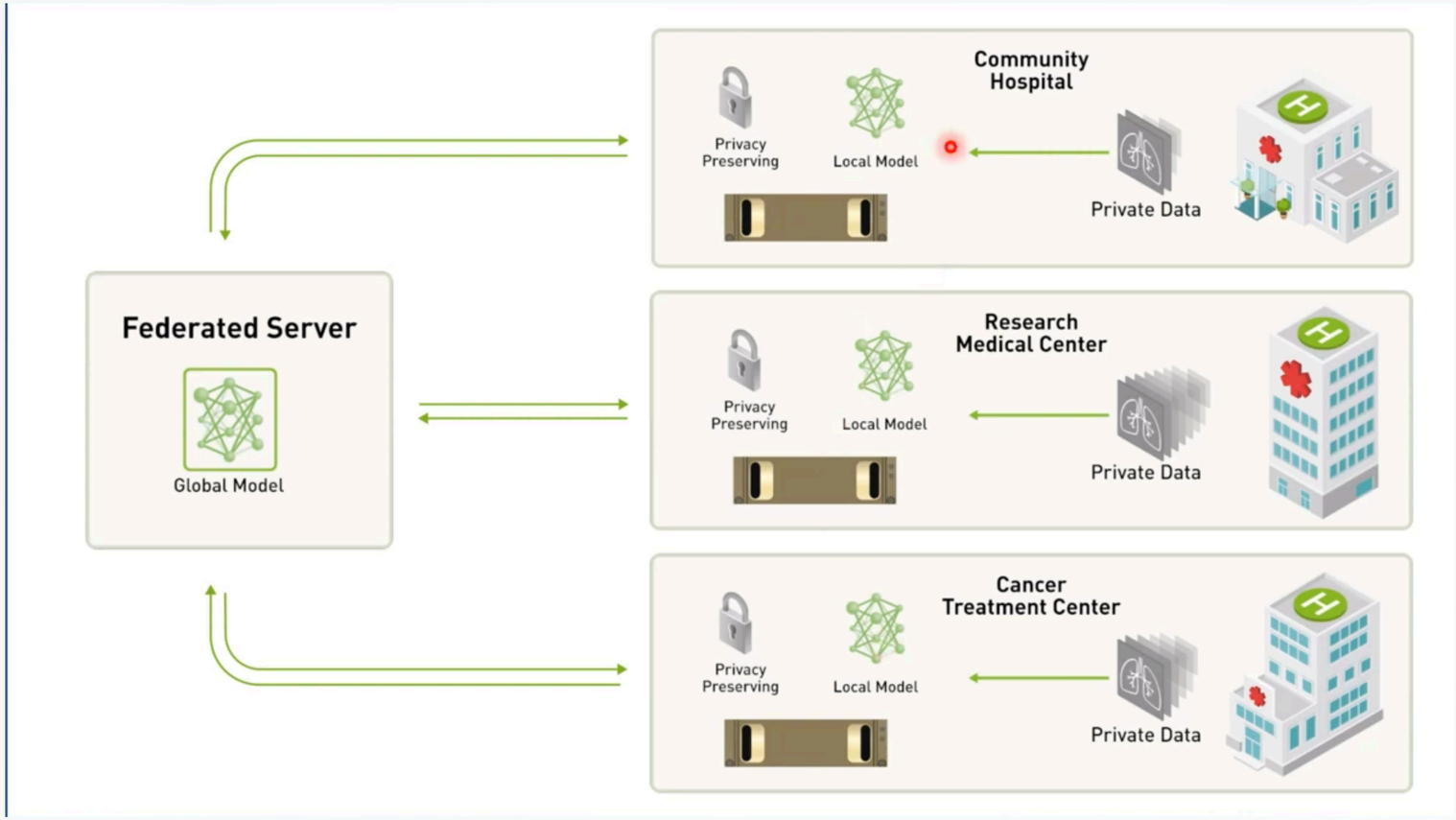
- This approach enables users to make informed decisions about hardware and software setups for DNN training, balancing time and energy considerations efficiently.

- Scalability and Sustainability is one of the purpose as this training is essential for scaling up machine learning applications and ensuring sustainability by reducing energy consumption .

- Research aim to acurate modelling on training time and energy usage for different power modes of edge devices .

- The techniques like federated and geo-distributed learning require synchronized training rounds because in these techniques, we are sending the model of a particular device to a centralized model, not the data of the device to the centralized model.

- In this technique, devices in the training round need to complete training at about the same time.



- Fulfilling the notable gaps in edge-based DNN training includes addressing the lack of focus on evaluating performance specifically in edge environments. While much attention has been given to assessing DNN training performance in cloud environments, there is a significant gap in understanding how DNN training performs on edge servers.

- We can optimize the edge environment as edge devices have limited resources. With profiling, we can identify efficient use of these resources and unlock the full potential of edge computing devices .

- So that we can predict both the time and energy required to train a DNN model for any custom configuration of a given device . This prediction can assist in selecting the best configuration to balance time and energy trade-offs during model training

RESULT OF THIS RESEARCH

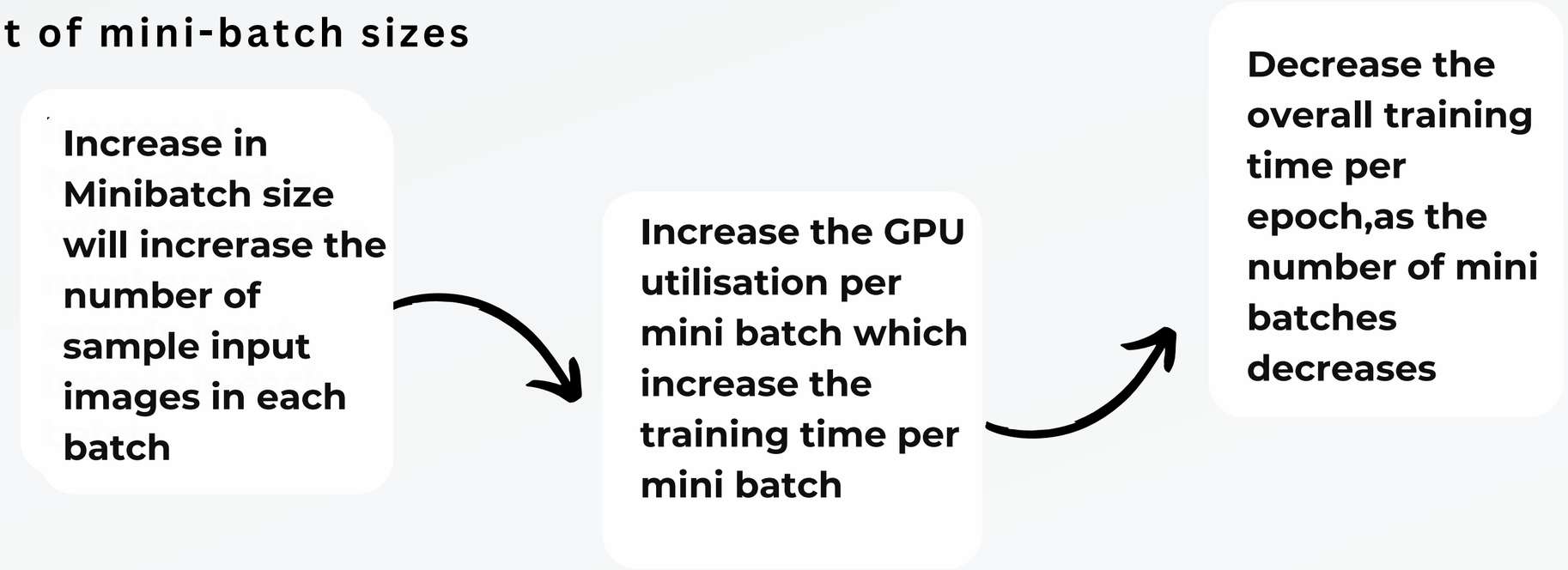
- Benefit of Disc caching ,Pipelining and Parallelizing on different devices and on different models in decreasing the end to end training time which helps in training efficiency and in energy conservation .

- This methods are not worked always beneficially in all the scenarios of models and devices ,so need to select a perfect method by considering factors like size of fetching data ,device capacity ,etc .

- How Faster storage media can reduce the stall time by faster accessing the input data samples to CPU ,But Even with slower storage media, the stall times may be small or negligible when pipelining and parallelization techniques are employed.

- Among SSD ,HDD and SD card ,Faster storage media SSDs, can lead to reduced stall times and faster training times, particularly for models with larger data requirements like MobileNet trained on large datasets but no stall time benefits for models like LeNet and ResNet .

- Effect of mini-batch sizes



- why we are dropping epoch 0 due to its bootstrapping overheads ?

- DNN models are trained on 5 devices each of AGX, NX, and Nano For each device type, so found out the significant variability among all the devices and the largest model it can train is selected –

LENET	on NANO
MOBILENET	on Xavier NX
RESNET	on AGX Xavier

- Effect of DVFS(Dynamic Voltage and Frequency Scaling) is negligible on end to end training time and energy usage across all the devices ,But observed change in GPU and CPU frequency on enabling and disabling of DVFS .

- Understanding the baseload under different states of idleness including conditions with no applications running and various configurations of Dynamic Voltage and Frequency Scaling (DVFS), as well as Wake on LAN (WoL) state.

- WoL state has the lowest power consumption for all devices, substantially lower than other idle states with DVFS on .Enabling DVFS results in significant power savings for more powerful devices like Orin and AGX but has a negligible impact on NX or Nano .Logging has minimal impact on faster devices but leads to higher power load for Nano

- Impact of different Power Modes on training time and energy consuption . This will eventually helps selecting the appropriate power mode for achieving an optimal trade-off between training time and energy consumption on Jetson AGX.

- The default power mode may not always be the best choice, and considering other power modes can lead to more efficient training processes .We got to know that energy consumed per epoch is not varying significantly on different power modes .

CUSTOM POWER MODES

- GPU frequency and compute time are often inversely proportional, meaning as the frequency increases, the compute time decreases, and vice versa. This relationship is especially noticeable in larger models like MobileNet and ResNet but not on smaller models like lenet .

- Same for the CPU frequency and Core counts are inversely proportional to the stall time ,i.e the stall time will decrease on increasing the CPU frequency and number of core counts .

- On comparing both CPU and GPU frequency ,GPU frequency have lesser impact than CPU frequency and core counts on GPU compute time in case of smaller models like lenet

Questions

- How we are deciding size of model ,will it suitable for train on device based on size ,As Resnet having disk size of 150 MB even it is not trained on Nano and NX xavier ?

- What is role of shared memory in DNN training stages ?

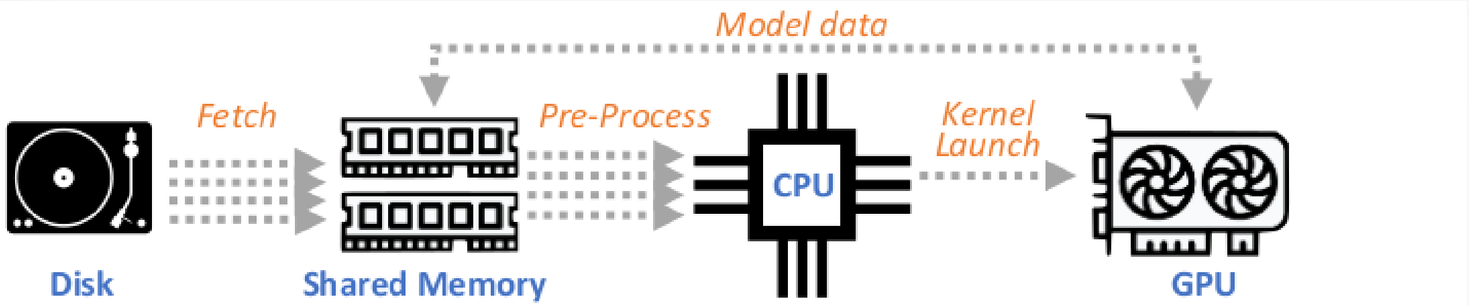


Fig. 1. DNN training stages using PyTorch on the Edge

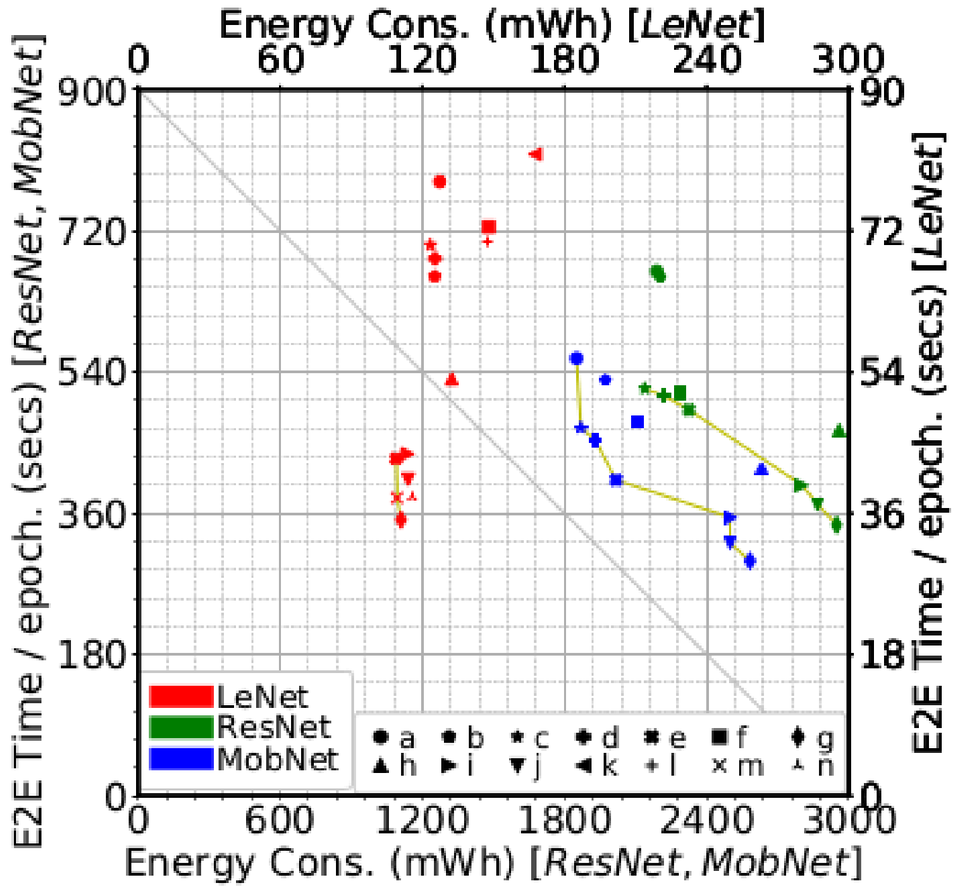


Fig. 8. Scatter plot of *E2E time vs. Energy consumed per epoch (1+)*, for power modes *a*–*n* of AGX for LeNet (secondary axes) and ResNet/MobileNet (primary axes). The yellow lines indicate the *Pareto front* per DNN.

THANK YOU