Energy Efficiency of Quantized Neural Networks in Medical Imaging

Priyanshu Sinha PRISINHA@IU.EDU
Sai Sreya Tummala TUMMALAS@IU.EDU
Saptarshi Purkayastha SAPTPURK@IUPUI.EDU
Indiana University Purdue University Indianapolis, Indianapolis, IN, USA.
Judy W. Gichoya JUDYWAWIRA@EMORY.EDU
Emory University, Atlanta, GA, USA

Editors: Under Review for MIDL 2022

Abstract

The main goal of this paper is to compare the energy efficiency of quantized neural networks to perform medical image analysis on different processors and neural network architectures. Deep neural networks have demonstrated outstanding performance in medical image analysis but require high computation and power usage. In our work, we review the power usage and temperature of processors when running Resnet and UNet architectures to perform image classification and segmentation respectively. We compare Edge TPU, Jetson Nano, Apple M1, Nvidia Quadro P6000 and Nvidia A6000 to infer using full-precision FP32 and quantized INT8 models. The results will be useful for designers and implementers of medical imaging AI on hand-held or edge computing devices.

Keywords: medical imaging, segmentation, classification, TPU, Jetson Nano, Apple M1.

1. Introduction

In the realm of medical imaging, image classification and segmentation are important clinical tasks. However, deploying high-precision models consumes a lot of resources (AskariHemmat et al., 2019). Quantization of these deep neural networks reduces both computational and memory load. As shown in our previous work (Abid et al., 2021; Sinha et al., 2022), many quantized models demonstrate similar performance metrics compared to full precision models. However, there is little known evidence, other than advertised peak power usage of processors (CPU/GPU), about the efficiency of neural networks for medical imaging tasks.

Various strategies for optimizing the energy usage of deep neural networks have been presented in the literature, including custom hardware designs, or optimizing the type, arrangement, and hyperparameters of the individual layers in the deep neural network (Young et al., 2019). To get energy-efficient 3D CNN processing for embedded system action recognition, systematic quantization can be used (Lee et al., 2018). Energy estimation of a model can be measured (in joules) once we have estimated the inference in FLOPs for a range of models and the GFLOPS per Watt for different GPUs (Desislavov et al., 2021).

In this paper, we optimized commonly-used neural network models using quantized aware training (QAT) and compared the energy efficiency of the models on different hardware through temperature and power consumption. The Jetson Nano and Coral Edge TPU

© 2022 P. Sinha, S.S. Tummala, S. Purkayastha & J.W. Gichoya.

are now easy to integrate with handheld devices, and our ultimate goal is to create portable medical AI support devices that can improve diagnosis at the point-of-care. For this, we need to reduce the computation for saving energy, cost, and time. The major goal is to enable medical AI in low-resource contexts.

2. Methods

We review the energy usage of Google Edge TPU, Nvidia Jetson Nano, Apple M1, Nvidia Quadro P6000, and Nvidia A6000. The Edge TPU is a Google-designed application-specific integrated circuit (ASIC) for ML inference on low-powered devices. It is capable of 2 TOPS/watt. We run our experiments on the Edge TPU instead of the Vivante GC7000Lite GPU available on the Coral Dev Board. The Nvidia Jetson Nano contains a Maxwell architecture GPU with 128 CUDA cores combined with a Quad-core ARM Cortex-A57 CPU and 4GB RAM. Its GPU is capable of 0.47 TOPS/watt but can perform 32-bit floating-point operations and not 8-bit int. The Apple M1 is part of the Macbook Air, containing an A12Z Armv8 CPU and a 7-core GPU. ML acceleration is through a SOC integrated Neural Engine, capable of 1.1 TOPS/watt. However, our experiments were run using tensorflow-metal on the M1 GPU, with a max throughput of 2.6 TFlops. The Nvidia Quadro P6000 is a Pascal architecture GPU with 24GB VRAM with 3840 CUDA cores. It has a 250W peak power draw to produce 12 TFlops with 0.026-0.048 TOPS/watt. The Nvidia A6000 is an Ampere architecture GPU with 48GB VRAM with 10752 CUDA cores. It has a 300W peak power draw to produce 38.7 TFlops with 0.096-0.129 TOPS/watt.

While each CPU/GPU is architecturally very different, and comparisons are hard to make, the objective is to review off-the-shelf products for anyone deploying medical AI models. The advertised peak performance is in floating-point operations per second (FLOPS) or trillion operations per second (TOPS). However, in practice, actual performance depends on various factors such as the ML model architecture, ML framework, power source, environmental temperature, OS drivers, and other running processes. We have tried our best to abstract these factors and report the change in temperature and power draw, keeping all other factors the same before and after the inference. The QAT training and model optimization is similar to our previous work reported elsewhere (Sinha et al., 2022). We capture the power draw using a power meter at the inlet for Edge TPU and Jetson Nano with only a network cable and no other peripherals connected. For others, we capture power draw on Nvidia P6000 and A6000 using *powertop*, and M1 using *powermetrics*. We capture the temp using an infrared camera for Edge TPU and Jetson Nano, *TGPro* for M1, and *powertop* for others.

3. Results

The Tables 1 and 2 show the power and temperature usage (difference in post and pre evaluation values) when running the most relevant models for inference on each platform.

4. Conclusion

The Edge TPU is excellent in accuracy-power performance, but complex models will need to be quantized to INT8. Jetson Nano does excellent in classification tasks compared to

Table 1: Model performance

	FP32	INT8
Ultrasound Nerve Segmentation (Dice)	0.617	0.646
Brain MRI Segmentation (Dice)	0.96	0.94
Chest X-ray Classification (AUC-ROC)	0.81	0.77

Table 2: Power and Temperature usage across platforms

	Measure	Google Edge TPU	Nvidia Jetson Nano	Apple M1	Nvidia P6000	Nvidia A6000
Ultrasound Nerve Segmentation	Temp.(C)	0.9	4.5	21.0	31.0	15.0
	Power(mW)	1300	3376	1286	172000	142000
Brain MRI Segmentation	Temp.(C)	2.5	1.0	8.0	19.0	9.0
	Power(mW)	1490	3703	4004	171000	65000
Chest X-ray Classification	Temp.(C)	3.1	0.5	20.0	18.0	12.0
	Power(mW)	780	3520	720	168000	71000

segmentation and likely due to Resnet optimized for CUDA cores. However, the biggest surprise is Apple M1 with high performance and low power usage with very little heating and fast cooling with passive cooling.
Source Code: deep_models_energy_consumption.

References

Areeba Abid et al. Optimizing medical image classification models for edge devices. In *International Symposium on Distributed Computing and Artificial Intelligence*, pages 77–87. Springer, 2021.

MohammadHossein AskariHemmat et al. U-net fixed-point quantization for medical image segmentation. In *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention*, pages 115–124. Springer, 2019.

Radosvet Desislavov et al. Compute and energy consumption trends in deep learning inference. *CoRR*, abs/2109.05472, 2021. URL https://arxiv.org/abs/2109.05472.

Hyunhoon Lee et al. Fixed-point quantization of 3d convolutional neural networks for energy-efficient action recognition. In *2018 International SoC Design Conference (ISOCC)*, pages 129–130, 2018. doi: 10.1109/ISOCC.2018.8649987.

Priyanshu Sinha et al. Leapfrogging medical ai in low-resource contexts using edge tensor processing unit. In *2022 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT)*, pages 67–70, 2022. doi: 10.1109/HI-POCT54491.2022.9744071.

Steven R. Young et al. Evolving energy efficient convolutional neural networks. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 4479–4485, 2019. doi: 10.1109/BigData47090.2019.9006239.