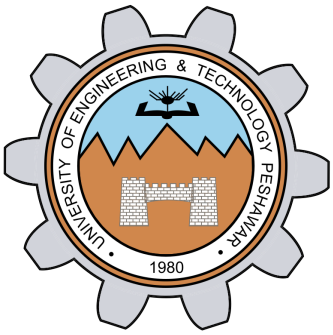
Assignment # 2



Fall 2024
CSE-420 Embedded Systems

Name: Ali Asghar
Registration: 21PWCSE2059
Section: A

Submitted to:
Dr.Asif Ali Khan

Date:
7th November, 2024

Department of Computer Systems Engineering
University of Engineering and Technology, Peshawar

Apple Neural Engine

The Apple Neural Engine (or ANE) is a type of NPU, which stands for Neural Processing Unit. It's like a GPU, but instead of accelerating graphics an NPU accelerates neural network operations such as convolutions and matrix multiplies[1]. It was introduced with the A11 Bionic chip, and is a dedicated hardware block within Apple's system-on-chip (SoC) designed specifically for accelerating machine learning and artificial intelligence tasks.

The ANE isn't the only NPU out there — many companies besides Apple are developing their own AI accelerator chips. Besides the Neural Engine, the most famous NPU is Google's TPU (or Tensor Processing Unit). In Figure 1 below, we can see the ANE in recently launched Apple A17 Pro Chip having 16 cores.

Figure 1: Neural Engine in Apple A17 Pro Chip

1 General Overview

1. Micro-architecture

- Apple's Neural Engine, first introduced with the A11 chip, is a specialized micro-architecture within Apple Silicon (like the M1 and M2 chips) designed to accelerate machine learning tasks. It includes multiple cores optimized for tensor processing, designed to perform matrix multiplications efficiently. The Neural Engine is tightly integrated within Apple's SoCs, allowing fast memory access and low latency.
- The architecture emphasizes energy efficiency and performance balance for real-time AI tasks on mobile and desktop devices.

2. Programming Model

- Developers primarily program the Neural Engine using Apple's Core ML framework, which enables high-level APIs for deploying machine learning models on Apple devices. Core ML automatically optimizes models to take advantage of the Neural Engine, CPU, and GPU, depending on task requirements.
- Core ML can import models trained in popular frameworks like TensorFlow or PyTorch, converting them into formats optimized for Apple's hardware.

3. Type of Parallelism

- The Neural Engine uses both data and task-level parallelism. The engine is composed of multiple cores that can handle parallel processing of tasks, which is especially beneficial for tasks like image processing and other real-time AI tasks on-device.
- Apple's design is optimized for low-power parallel processing, making it ideal for mobile applications where energy efficiency is crucial.

4. Comparison with Google's TPU

- **Micro-architecture Comparison:**
  - Apple's NPU is designed for low-power, on-device computations, making it suitable for mobile and desktop AI applications. In contrast, Google's TPU is designed for high-performance, high-throughput tasks in a data center environment, focusing less on power efficiency and more on computation speed.
- **Programming Differences:**
  - Apple's Neural Engine uses Core ML for programming, aiming to make machine learning integration seamless for developers on Apple platforms. Google's TPU, on the other hand, is closely integrated with TensorFlow, requiring a more specialized setup in Google's ecosystem.

- **Parallelism and Optimization:**
  - While both employ parallelism, Apple's Neural Engine is optimized for task-level parallelism and energy efficiency, making it suitable for real-time applications on personal devices. Google's TPU focuses on high-throughput data parallelism, suitable for large-scale neural network training.
- **Use Case Differences:**
  - Apple's NPU is aimed at enhancing AI-driven features on personal devices (like photo processing, augmented reality, and voice recognition), while Google's TPU is intended for large-scale AI model training and inference in data centers, particularly for tasks that require extensive computational power without strict power constraints.

Summary

Apple's Neural Engine and Google's TPU serve distinct purposes. Apple's NPU is optimized for on-device, energy-efficient processing for user-facing applications, while Google's TPU is designed for large-scale AI processing in cloud environments where power constraints are less critical [2].

References

[1] hollance, "neural-engine," https://github.com/hollance/neural-engine, 2024.

[2] OpenAI, "Chatgpt," https://www.openai.com, 2024.