In our past work we described offloading an approximable region of code to a fast and efficient neural processing unit made, naturally, in silicon. Here we envision a similar system, but instead using the most powerful neural network of all – a biological neural network as the accelerator.

1. Introduction

As the historical trend of speed and energy efficiency improvements diminishes [2], radical departures from conventional approaches are becoming critical to improving the performance and energy efficiency of general-purpose processors. Inspired by biological nervous systems, neuromorphic computing delivers high performance within a very low power envelope. In fact, human brains can use only 20 Watts to carry out tasks which require a warehouse of processors to accomplish. Despite these advantages, neuromorphic computers are not easily programmable in the way that traditional von Neumann machines are. Furthermore, such models have inherent inaccuracy in their computation, which must be addressed by the programmer. Towards reconciling these differences, our past research has focused on bridging neuromorphic and von Neumann computing models through intuitive programming models and architectural interfaces, while embracing the notion of approximate execution [3]. Of course, this work has only used artificial digital or analog neural networks. We seem to have forgotten the most powerful neural network of all — a biological neural network! We imagine a future technology in which computational neuroscience and computer architecture intersect, leading to a programming environment which offloads approximable regions of code onto the very brain of its users. In this framework, the biological nervous tissue becomes an accelerator for a code written in conventional programming languages. We refer to these accelerators as bio-accelerators, and explore their function, strengths and limitations.

2. Application

Human brains are capable of performing tremendously complicated tasks while consuming minimal energy. A human can complete a facial recognition task in only 100 ms, all while processing dozens of other thoughts and consuming only the energy found in a peanut butter sandwich. Imagine a wearable device, similar perhaps to Google Glass, which is capable of offloading spatial, visual, and audio information storage and processing to the brain of its wearer. Such a device could consumer far less power, allowing it to run for days while performing computation tasks far beyond the capabilities of today’s devices. A camera on the device could feed images directly to the brain and gain classification for free using preexisting circuitry designed for the same task. Some type of brain to brain interface may even be possible, through which information is directly communicated between individuals.

3. Bio-Acceleration

Computation. Newer generations of neural networks are becoming increasingly similar to their biological counterparts [5]. Early neural networks consisted of neurons which could only send binary signals. They would send a “one” only when the weighted sum of their inputs crossed some threshold value. Later networks used a continuous activation function instead of a step function, branching into the analog domain. Newer networks use spikes or pulses to encode information much as a real neuron would. Perhaps if this trend towards biological realism continues, the task of offloading code will not seem very daunting. However, there are some important differences between a biological neural network and the types of neural networks constructed today. Most obviously, the sheer number of neurons available in a biological system is far beyond what is typically used. For example, a perceptron consisting of four layers and a dozen neurons per layer might be reasonable. In a system with 100 billion neurons, however, part of the challenge may be simply utilizing all of them. Additionally, each neuron in a biological system may send signals to as many as 10,000 other neurons [5], making the task of utilizing such a topology difficult. To further complicate matters, the neurons themselves may be slower than the ones we design today. Chemical signals are sent across synapses which are slower than electrical signals, so a more complicated network may be necessary to make up for these slow interactions.

Storage. While storage is intuitively commonplace in neural networks – after all brains are clearly capable of remembering things – the exact mechanisms are not completely understood, let alone controllable. The primary mechanism of learning derives from Hebbian theory, which at its core states that “cells which fire together wire together.” In other words, the connection between neurons becomes stronger when there is a correlation between activity in the presynaptic and postsynaptic neurons. This lasting connection between neurons is called long-term potentiation. In a simplistic perceptron model, the potentiation of a synapse might be compared to the “weight” of an artificial neuron. Storage of information in a brain would, of course, be very different than our current random access memory model. Instead, information appears to be stored in an associative manner, perhaps more analogous to a key-value store. For example, when your visual cortex is presented with the “key” of a face, you might retrieve the “value” of a name, along with dozens of other pieces of information about the person. Although the memory interface would differ from a traditional random access scheme, today’s programmers may actually feel comfortable in the new environment. Such a key-value interface is actually analogous to many database systems in existence already, with the caveat that any operations done within the brain are inherently approximate.

Programming. We have demonstrated the viability of methods which allow certain regions of code to be offloaded to a neural processing unit without disruptive changes to the traditional programming model [3]. The programmer marks certain regions of code as “approximable” using a simple keyword. We can then utilize an algorithmic transformation which automatically converts these regions of code from a traditional von Neumann model to a neural model. We
Although the effects are under scrutiny, limited use seems to have been the short and long term safety of such an activity. Past research has demonstrated the possibility of using the brain as a computing substrate. However, the ethical concerns about the invasiveness of neurostimulation and the potential for long term harm must be addressed. Today, a large body of research exists concerning electrocorticography — a method for recording brain activity using an array of sensors. With this procedure, brain signals can be monitored and transmitted to other devices, allowing basic control of robotics or other interfaces. Although using the brain as a computing substrate seems impossible today, future research may open new doors. For example, the BRAIN initiative is a proposed collaborative project aiming to map every neuron of the human brain, much like the human genome project did for DNA. This project could close an important gap between our understanding of high level brain function and cell level neuron function. Perhaps this will lead to breakthroughs which will make bio-accelerator research possible in the future.

6. Conclusion
Although this idea is wacky and a little strange, we believe that the technology may one day exist to make a bio-accelerator reality. We recognize that there would be a long list of ethical concerns about creating such a device, but we delegate that discussion to the future society capable of building it.

References