Bio-Accelerators: Bridging Biology and Silicon for General-Purpose Computing

Bradley Thwaites Amir Yazdanbakhsh Jongse Park Hadi Esmaeilzadeh
Georgia Institute of Technology
{bthwaites,a.yazdanbakhsh,jspark}@gatech.edu hadi@cc.gatech.edu

In our past work we described offloadling 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 programer. 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

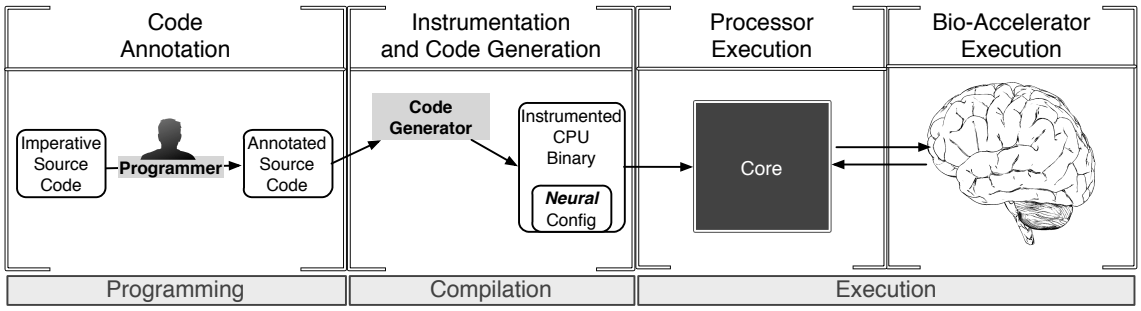

Figure 1: The neural transformation: from annotated code to accelerated execution on a bio-accelerator.

envision a system similar to what we have already demonstrated, but instead of a silicon neural processing unit we envision using biological neural networks. Figure 1 demonstrates a possible workflow from source code to bio-accelerator invocation. We believe the following criteria are enough to offload computation to a neural network: 1) it must be approximable, since the transformation is not guaranteed to exactly mimic the target region, and 2) there must be well defined inputs and outputs. While the transformation can, in theory, create code which can run on an arbitrary neural processor, an additional layer might be necessary when working with a real brain. There might be limitations to what kind of networks can be programmed, or there might be limits to how long a living brain can be used for general purpose computing before there are lasting health effects. The final requirement is an architectural interface between computer and brain. While silicon NPUs are tightly coupled with the processor, here due to the high latency of interfacing with biological nervous tissue we envision that the bio-accelerator would be of a loosely coupled form. They might be similar to GPUs, where large granularities of computation are offloaded to the bio-accelerator. Finally, unlike the code that is offloaded to a silicon NPU, a bio-accelerator could handle code that is stateful thanks to its storage capabilities.

4. Safety

One of the concerns of using bio-accelerator as a computing substrate would be the short and long term safety of such an activity. Past brain-computer interfaces have focused on measuring brain activity and using this information to control external devices. This proposal, however, would involve reprogramming groups of neurons for general purpose computation. Currently, a technique called transcranial magnetic stimulation can be used to depolarize neurons in a specific area of the brain for treatment of certain neurological disorders [6]. Although the effects are under scrutiny, limited use seems to have mild effects, whereas long term use could cause gradual reprogramming of the synapses. Perhaps there is a balance to be struck. If interactions with the bio-accelerator for only fractions of a second do not cause lasting damage, useful work could theoretically still be performed. Another safety concern would be the invasiveness of procedures required to install any necessary devices. While a small outpatient procedure may be allowable, few users would accept a major surgery. In addition, hardware needs to be general enough to not require frequent, if any, hardware updates.

5. Feasibility

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 monitored and transmitted to other devices, allowing basic control of robotics or other interfaces [1]. However, while observing signals in the brain and acting upon them is an important step, it is far from the level of technology which would be required to accomplish what we propose here. Although using the brain as a computing substrate seems impossible today, future research may open new doors. For example, the BRAIN initiative [4] 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

[1] T. Blakely, K. Miller, S. Zanos, R. Rao, and J. Ojemann, "Robust, long-term control of an electrocorticographic brain-computer interface with fixed parameters," *Neurosurg. Focus*, vol. 27, no. 1, p. E13, Jul. 2009.

[2] H. Esmaeilzadeh, E. Blem, R. S. Amant, K. Sankaralingam, and D. Burger, "Power challenges may end the multicore era," *Communications of the ACM*, vol. 56, no. 2, pp. 93–102, February 2013.

[3] H. Esmaeilzadeh, A. Sampson, L. Ceze, and D. Burger, "Neural acceleration for general-purpose approximate programs," in *International Symposium on Microarchitecture (MICRO)*, 2012, pp. 449–460.

[4] J. Markoff, "The New York Times, Obama seeking to boost study of human brain," Feb. 2013. Available: http://www.nytimes.com/2013/02/18/science/project-seeks-to-build-map-of-human-brain.html

[5] J. Vreeken, "Spiking neural networks, an introduction," Utrecht University: Information and Computing Sciences, Tech. Rep. UU-CS-2003-008, 2003.

[6] Wikipedia, "Transcranial magnetic stimulation," Jan. 2014. Available: http://en.wikipedia.org/wiki/Transcranial_magnetic_stimulation