DISS. ETH NO. 24905

Event-Based CMOS Circuits for a Class of Belief-Propagation Models

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

presented by

Chen-Han Chien

MSc. Electrical Engineering, National Tsing Hua University, Taiwan

born on *27.05.1984*

citizen of Taiwan, Republic of China (R.O.C)

accepted on the recommendation of

PD. Dr. Shih-Chii Liu
Prof. Dr. Tobi Delbruck
Prof. Dr. Hans-Andrea Loeliger

2018

Abstract

Abstract Bayesian networks are often used to describe how brains can perform inference. Methods of transforming these abstract models to spiking neural networks that can perform inference are still scarce. A recently proposed model called the event-based Belief-Propagation (BP) model shows how inference can be carried out by using the distribution of interspike intervals in spike trains as the messages. Because the simulation times of a factor graph that uses this model can be very long, this thesis proposes an analog Very-Large-Scale Integration (aVLSI) version of this model as one method of speeding up the computation times. The electronic model will be a useful addition to the neuromorphic effort in building spiking neural network systems.

This thesis describes one hardware implementation of this event-based BP model, which uses both a Field-Programmable Gate Array (FPGA) and a mixed analog-digital Application-Specific Integrated Circuit (ASIC) chip developed in a 0.35um CMOS process technology. It describes the challenges in implementing the various circuit blocks of this stochastic model which includes the critical hazard function needed for the neuron to generate stochastic spikes following a defined probability distribution. Many of these circuit blocks did not exist in any form at the start of the thesis because most of the focus in the neuromorphic community is on spiking neural network chips that do not include a stochastic component. Therefore, this thesis presents possible solutions for implementing the event-based stochastic model in hardware.

The hardware system developed in this work is based on an architecture of the event-based BP model that is partitioned into a Landscape Sampling (LS) block and a Random Sampling (RS) block. The input spike trains carrying the BP messages are processed by the LS block that implements the constraint function of a defined factor node. The LS block outputs a message-combined probability distribution that is used by the RS block to produce the stochastic output spikes using the implemented hazard function.

The thesis considers the practical challenges of mapping the theoretical model to aVLSI circuits, the possible methods for generating on-chip noise sources, and the subsequent partitioning of the hardware system into an FPGA and an ASIC. The factor graphs constructed by the event-based BP model under the constraints of the hardware are validated through simulations and then applied in two tasks 1) object tracking using an event-based Kalman filter and 2) data reconstruction using the event-based Continuous Restricted Boltzmann Machine (CRBM). These applications are examples of possible applications of the hardware system. The thesis shows the capability of the final hardware system in implementing graphs with arbitrary variable distributions for its inputs and using constraint functions such as “plus” and “equality”. Measured results show that the BP hardware consumes 6.32 mW of power with 0.046 mW of power per RS channel on the ASIC.

Zusammenfassung

Abstrakte Bayessche Netze werden häufig verwendet, um Inferenz in Gehirnen zu erklären. Methoden, die solche abstrakten Modelle in gepulste neuronale Netze übersetzen, sind jedoch noch selten anzutreffen. Ein kürzlich vorgeschlagenes Modell, welches unter dem Namen ereignisbasiertes Belief-Propagation (BP) bekannt ist, zeigt auf, wie Inferenz erfolgen kann, wobei die Zeitintervall-Verteilung zwischen konsekutiven Pulsen die zu übertragende Nachricht darstellt. Da Faktorgraphen, die solche Modelle verwenden, eine lange Simulationszeit aufweisen, wird in dieser Arbeit eine analoge Very-Large-Scale Integration (VLSI) Variante dieses Modells eingeführt, um die Berechnungszeit zu verkürzen. Das elektronische Modell ist eine nützliche Ergänzung zu neuromorphen Ansätzen, gepulste neuronale Netze zu entwickeln.

Diese Arbeit beschreibt eine Hardware-Implementierung dieses ereignisbasierten BP-Modells, wobei sowohl ein Field-Programmable Gate Array (FPGA) als auch ein Mixed Analog-Digital Application-Specific Integrated Circuit (ASIC) Chip verwendet werden; der ASIC Chip ist für die 0.35um CMOS Prozesstechnologie entworfen. Die Arbeit beschreibt weiterhin die Schwierigkeiten, die bei der Implementierung der verschiedenen Schaltkreiskomponenten dieses stochastischen Modells entstanden; dies beinhaltet die wichtige Ausfallrate, die vom Neuron verwendet wird, um stochastische Pulse zu generieren, die einer definierten Wahrscheinlichkeitsverteilung folgen. Viele dieser Schaltkreiskomponenten existierten zu Beginn dieser Arbeit nicht, da ein Schwerpunkt innerhalb der neuromorphen Entwicklergemeinschaft darin liegt, gepulste neuronale Netze zu entwickeln, welche keine stochastischen Komponenten enthalten. Diese Arbeit zeigt daher Möglichkeiten auf, um ereignisbasierte stochastische Modelle in Hardware zu implementieren.

Das im Rahmen dieser Arbeit entwickelte Hardware-System basiert auf einer Architektur des ereignisbasierten BP-Modells, welche in einen Landscape Sampling (LS) Block und einen Random Sampling (RS) Block zerlegt ist. Die Eingangspulszüge, die die BP-Nachrichten tragen, werden vom LS-Block verarbeitet, welcher die

Nebenbedingungsfunktion eines definierten Faktorknotens implementiert. Der LS-Block gibt eine Wahrscheinlichkeitsverteilung aus, die den verknüpften Eingangsnachrichten in einen Knoten entspricht, und die im RS-Block dazu verwendet wird, um stochastische Ausgangspulse mittels der implementierten Ausfallrate zu erzeugen.

Diese Arbeit untersucht die praktischen Herausforderungen, die bei der Abbildung des theoretischen Modells auf einen aVLSI-Schaltkreis entstehen, die Möglichkeiten, um On-Chip Rauschquellen zu generieren, sowie die nachfolgende Zerlegung des Hardware-Systems in eine FPGA- und eine ASIC-Komponente. Die Faktorgraphen, welche im Rahmen des ereignisbasierten BP-Modells entwickelt werden, und die Hardware-Beschränkungen berücksichtigen, werden mittels Simulationen validiert. Sodann werden sie auf zwei Aufgaben angewendet: 1) Objektverfolgung mittels eines ereignisbasierten Kalmanfilters; 2) Datenrekonstruktion mittels einer ereignisbasierten Continuous-Restricted Boltzmann-Maschine (CRBM). Diese Anwendungen sind Beispiele möglicher Applikationen des Hardware-Systems. Diese Arbeit zeigt die Fähigkeit des entwickelten Hardware-Systems auf, Graphen mit beliebigen Eingangs-Zufallsvariablen-Wahrscheinlichkeitsverteilungen zu simulieren, die Nebenbedingungsfunktionen wie "Plus" und "Gleichheit" benutzen. Messungen ergeben, dass die BP-Hardware 6.32 mW Leistung aufnimmt, wobei 0.046 mW Leistung pro RS-Kanal des ASICs aufgenommen werden.