Project Proposal

Andrew Davis

CS 461 Oregon State University

Fall 2018

Abstract

Artificial neural networks mirror neural networks that take place in brains. They allow machines to learn on their own and come to decisions based on the strength of connections between neurons (similar to synapses in the brain). For artificial networks to learn and make decisions, the weights between all of the artificial neurons must be traversed in both decision making and updating information. There are many ideas out there attempting to pioneer faster processes in this regard. This research and design / development project will look at ways to make the power output of artificial neural networks smaller, allowing for the use of these systems to reach all levels of society.

1 Definition and Description

The following will cover: a brief overview of artificial neural networks and inference engines, the motivation behind the project, proposed solutions, and ways to measure performance.

1.1 Brief Overview of Artificial Neural Networks

Within an artificial neural network are two main components: artificial neurons (similar to neurons in the brain), and edges connecting these artificial neurons to one another (similar to synapses in the brain). These edges connect artificial neurons from the output of one, to the input of another, ultimately ending at a set of nodes. Each edge carries a weight that is used to determine a number of things throughout the neural network.

Initially, the artificial neural network must be trained. That is, given a set of data, the artificial neural network takes the information in and creates a network of nodes that lead to decisions. As more data is introduced, the artificial neural network reframes the network, changing the weights (also known as coefficients) of the system, leading to different decisions being made based on a better knowledge base. The number of nodes that can be present in a single system can range from the thousands to the millions.

1.2 Brief Overview of an Inference Engine

An inference engine is a portion or step taken in artificial intelligence systems such as neural networks. It uses logical rules to come to conclusions of a given knowledge base. That is, it makes inferences on which direction to traverse on a graph based on what it already knows. This is different from conventional control flow methods that make decisions on a step by step basis. The prior is often significantly faster in performing tasks and deducing the best action to take. As a result, we will look into the benefits of inference engines over specifying control flows.

2 Motivation

Due to the complex connectivity among different artificial neurons, and the scale to which artificial neural networks can reach, time and space are very demanding, limiting resources. This demand negatively affects the processing time of artificial neural networks, which are used in the financial realm, medicine, image processing, compression, and other fields. While these areas of expertise do not always require fast results, there is not a need for immediate results to the user. There are applications that many are attempting to apply artificial neural networks to that do require immediate results. Therefore, the need for artificial neural networks to perform their job quicker is an important field that has led to breakthroughs such as the use of inference engines.

The efficiency requested is applicable to fields accessible to everyday users, rather than being limited to industry and professionals. In order to achieve industrial levels of artificial neural networks on everyday lives, such as cell phones, autonomous vehicles, personal computers, and many more, the power consumption and time for calculations in artificial neural networks require them to be made more efficient.

3 Solution

The proposed solution is not entirely set in stone, as the specifics of the project are still somewhat unclear. Upon research, one solution to creating faster artificial neural networks is through the compression of the lookup table by pruning redundant edges which share the same weights [2]. This solution is a prominent route in the field of artificial

intelligence, especially in neural networks, as these systems can consist of thousands to millions of nodes. Once this modified artificial neural network is reduces to a more usable state, an inference engine can go in and perform its duties to speed up the process in deducing information. Another research solution propose by [1] uses a JPEG image encoding algorithm. This solution proposes a designed inference engine on data that has already been pruned, similar to [2].

4 Performance Measures

Measuring success of our research and design / development project may be tough. It seems that a measure of success comes from the speed and power use of artificial neural networks. In industry and the research realm, it seems this is all measure in percentages of new systems versus old ones.

As far as memory performance goes, there is clearly a difference between access to memory using static RAM and dynamic RAM. A way to possibly test the effectiveness of the desire artificial neural network is to assess the memory usage between static RAM and dynamic RAM. Our goal should be to take an existing system and minimized the memory usage between the two types of RAM listed (still pending).

As far as time performance goes, this can be related to memory. The outputs of an artificial neural network should be computed within a reasonable amount of time. Since there are possibly millions of artificial neurons within a single system, it is not feasible to traverse the entire tree. Instead, there should be a way to smartly choose paths, therefore lessening the time the artificial neural network spends coming to conclusions and updating the weights of the edges.

5 Conclusion

Up to this point, the project is still somewhat unclear based only on the project description given. We have yet to receive contact from out client but will reach out again to establish a concrete baseline for what is desired of us. The previously proposed problem statement, again, is based on my interpretation of the project description and research done on the subject.

This problem statement gives a very brief overview of artificial neural networks and inference engines. It also outlines the need for faster artificial neural network systems as they apply to commercial, everyday use by the general populace. Proposed solutions to creating more efficient systems are given through research done by other entities. Finally, ways in which we will measure the performance of our system are reviewed, focusing on the two most important factors: memory usages and time complexity.

REFERENCES

[1] Jong Hwan Ko, Duckhwan Kim, Taesik Na, Saibal Mukhopadhyay. Design and Analysis of a Neural Network Inference Engine based on Adaptvie Weight Compression. February 2 2018.
`https://ieeexplore.ieee.org/abstract/document/8279481`

[2] Song Han, Xingyu Liu, Huizi Mao, Jing Pu, Ardavan Pedram, Mark A. Horowitz, William J. Dally. EIE: Efficient Inference Enginge on Compressed Deep Neural Network. February 4 2016.
https://arxiv.org/abs/1602.01528