Comments on thesis draft 8th January 2023 version chapter 1

Section 1.1

Change title to "Neural Network Model"

Change the last paragraph to
"On one hand there are single-layer NNs that consist of just two layers - an input and an output layer, with only the output layer performing the computation. Self-Organizing Maps (SOMs) are examples of single-layer NNs and are used in dimensionality reduction and clustering applications. On the other hand, there are modern NN architectures that are very deep and involve many layers and millions of parameters. Multilayer NNs with more than three layers are referred to as deep neural networks (DNNs). These DNNs are capable of learning complex functions from the data."

Section 1.1.1

Change the section number to 1.2 and adjust subsequent numbers

Insert "For example," before "A popular CNN, VGG16, performs ...."

Delete the text "The recent growth of deep learning, advancement in deep learning tools, and recent research in the field of efficient edge AI accelerators have enabled several intelligent applications for consumer and edge devices."

Section 1.2.1

Let this not be a separate subsection. Merge it with the rest of 1.2.

Change "desired to improve" to "aimed at improving"

Delete "We aimed to provide energy-efficient solutions for NN accelerators by reducing the off-chip memory accesses by applying novel data reuse schemes and representing data in low resolution for FFNNs and RNNs."

Section 1.3.1

Replace text
"Since memory accesses dominate . . . as shown in Figure 1.5."
with
"NN computations are memory intensive. With limited on-chip memory on the accelerators and large difference between latencies and energy consumption of off-chip and on-chip memories, off-chip memory accesses dominate the performance and energy consumption of these accelerators. As much as 80% of the overall energy consumption of an NN accelerator could be due to off-chip memory accesses [6]. Therefore, reducing the off-chip memory is the key to improving the throughput and energy efficiency of DNN accelerators. This has led several researchers to focus on reducing the off-chip memory accesses [6, 9, 49]. Some approaches [14, 27, 37] have used on-chip memory to

store all the weights. However, since sizes of weights in modern NN models can be several MBs, these approaches are not scalable and are effective only for small NN models. Approaches attempting to reduce the off-chip memory accesses of NN accelerators can be classified into two broad categories, as shown in Figure 1.5. "

In the 2nd paragraph, replace the text
"Also, the number of parameters . . . memory access further."
with
"These techniques have been explored only for very limited domains such as image processing and computer vision {check the accuracy of this statement}. The number of parameters in modern DNNs is significantly large. For these DNNs, besides quantization and pruning, additional techniques may be required to reduce the off-chip memory accesses further."

Replace the text
"Data-reuse schemes are the other line of approach that does not affect the accuracy of the network."
with
"The second category of approaches, that do not affect the accuracy of the network, are data-reuse schemes."

Section 1.3.2

Change title to "Thesis objectives and contributions"

Delete the text "Accessing data from the off-chip memory . . . key to improving the throughput and energy efficiency of DNN accelerators."

Replace the text
"In this thesis, we have explored both bit-accurate and approximate techniques on various NNs and contributed some new ideas."
with
"Objective of this thesis has been to provide novel energy-efficient solutions for NN accelerators by reducing the off-chip memory accesses. The proposed techniques fall within the classification of figure 1.5, but carry some new ideas. Since every memory access reduction technique is not necessarily applicable to the entire range of NN algorithms, we have considered three NN algorithms which present quite differeent characteristics from each other. These are Self Organizing Maps (SOMs), Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs)."

Replace the text
"We have applied quantization techniques on Self Organizing Maps (SOMs), a single layer FFNN, to analyze the impact on NN accuracy and benefits on improving energy efficiency."
with
"We have applied quantization techniques on Self Organizing Maps (SOMs) to analyze the impact on NN accuracy and benefits on improving energy efficiency. These are single layer FFNNs, not studied in this context previously."

Replace the text
"We also proposed novel data reuse approaches for multilayer feedforward and recurrent NNs."
with
" For multilayer feedforward and recurrent NNs, we have proposed novel data reuse approaches."
and move it from the end of the existing paragraph to the beginning of the next paragraph.

Section 1.4

The first paragraph is a repetition of what has already been said. This can be removed.