Answer to Reviewers

We sincerely thank all the reviewers and editors for providing us so many valuable suggestions to improve the quality of this paper. All the review comments are addressed and answered. Please see the answers below and the corresponding contents in the paper.

The newly added contents and the major revised contents are marked in blue in the paper.

The major improvements are listed as follows:

Reviewer 1

The goal of this paper is to provide a survey of neural networks built using FPGAs. The networks are categorized according to several approaches and then some comparison is attempted.

The paper does cover the many techniques that have been applied to FPGA-based designs, but it needs more and better organization. The usual flow is to mention an approach and then list various papers that have applied that technique. More analysis could be provided. Based on the listed works, did any trend appear? Did anyone do it better? If so, why? I realize this may often be hard, but that is the value that you will contribute with this paper.

I do not recommend publication of this paper until the issues described here have been addressed. I think significant work is required.

There are numerous spelling errors that could have easily been corrected with a spell checker. There are many grammatical errors that need to be addressed.

Answer:

Section 2

1. The paper begins with a very brief overview of neural networks and only tries to explain several types of layers. I believe you are just trying to show that there are several types of layers that need to be computed, but you also need to give some idea about the overall system, especially since that is important for later discussions. Figures would be helpful. Consider that many people new to the area would want to start with a survey paper. They need a clear explanation of the structure. As currently written, this section will make sense to someone that has worked in the area, but such a reader does not need this introduction. Someone new to area will not be able to

understand the bigger picture and challenges.

2. You have not defined what "weights" are when you first mention them. "The weights of this layer..."

3. You should briefly discuss the difference between inference and training. You also need to discuss accuracy. These are important concepts. Your work focuses on using FPGAs for inference. You don't seem to consider accuracy in your overall comparisons.

Section 3

4. You begin the section stating that accuracy, throughput, efficiency and flexibility are important. You need to more carefully define what you mean for each of those properties. For example, you give an equation for throughput, but have not defined the terms in the equation. How are each of the terms measured? There is no definition of flexibility.

5. You make the following statement, "Different network structures like the ones in [17, 22, 46] surely affect model accuracy, but is out of the discussion of this paper." Why? I don't think you have specifically defined the focus of this paper and you need to say more here why these papers are not relevant to this paper.

6. "model compression methods" has not been defined when you first use the term.

Section 4

7. It's not clear to me why this section includes "Software Design" in its title. These techniques are also relevant to any hardware design.

8. Section 4.1.3 - What are BW_weight and BW_neurons? I'm guessing BW means "bit width"? Explain how the comparison is being made. Where is the data coming from?

9. "... we see that..." How do we see this?

10. Section 4.2 - What is the "lasso object function"?

Section 5

11. Figure 2 - Explain more what you are doing here. What is the point of this comparison? Why not do comparisons with and without DSPs? It looks like you have synthesized multipliers and adders in different ways to get their resource usage. You need to provide more details on how you did this. There's not enough information for me to determine whether the

comparisons are fair. How did you describe the functional units? Did you just use A + B in Verilog with appropriate signal bit widths and types? For floating point, did you instantiate the cores from the library? What if you were to use an Altera part with the hardened FP in the DSP?

12. In section 5.1.1, "Operations with 32-bit fixed point data consumes similar resource as 32-bit floating point operations." This does not make sense to me. Needs more explanation.

Section 6

13. The challenge for this paper is that there are so many different FPGA platforms used combined with many different networks that have been implemented. It becomes very hard to make any comparisons. While it is good that you are trying to gain some overall understanding based on the results of your survey, what you have is not well-justified and insufficient. I think you should begin by first discussing what are important trends that you want to observe and then explain how you will gather the data and show the results. You have chosen to look at energy efficiency, but what about accuracy? What features affect that? For that matter, your energy efficiency comparison has no consideration for accuracy. In the extreme case I could build a circuit that has 1% accuracy and is very energy efficient, but is of no practical use so I don't think your current comparison is meaningful without more context.

14. Why did you pick those particular designs for Table 1?

15. In the text and Table 1 you refer to the various designs by their reference number, but then in Fig. 6 you use author names. This makes it very hard to figure out what the text is referring to in the graph. A similar issue is a statement like, "1-2 bit based designs show...". Which points are those on the figure?

16. Fig. 6: Y label should be GOP/s, not GOP.

17. Section 6.0.4 - estimate of achievable performance of an ideal design. Needs more discussion. Why is this a valid estimate? Why are we not anywhere close to this?

Reviewer 2

This paper is well written and presents the great effort of the authors with surveying the existing FPGA-based neural network accelerator designs. There are several minor issues that require a minor revision.

1. The last paragraph of Section 1, Section 3 is missing in the paper structure introduction.

2. Section 4.1.3, for the comparison of linear and non-linear quantization methods, I suggest either more examples should be collected for non-linear quantization method or the explanation of the observation need to be revised, e.g. the bit-width of neurons are bounded with 32, why?

3. Section 5.1.1 paragraph 4, for the declaration "All the IPs are .... actual hardware cost". If the IPs are required to avoid using DSPs, this comparison is only providing an ideal hardware cost in terms of logic resources. This need to be modified.

4. Section 7.1, DnnWeaver should be either a mixed method or instruction based.

Several minor typos: 1. Section 4.4, paragraph 2, "The left weights are then fined-tuned are...." should be "The left weights are then fined-tuned with". 2. Section 5.1, paragraph 1, "then the this also A", seems a sentence is missing here. 3. Section 5.2, paragraph 1, "and can further accelerated in frequency domain" should be "can be ...". 4. The figures in this paper need further arrangement to make it more clear and beautiful.

Reviewer 3

This paper presents a survey of FPGA-based neural network accelerators. The survey is of limited value though because it focuses only on CNN. Moreover, the survey does not provide a high-level view of the different architectural choices in designing FPGA-based CNNs. Instead, it looks into low-level design optimizations. However, the evaluation section is well done and provides a comprehensive summary of the quantitative performance of the different proposals.

1. The design methodology considers model accuracy, throughput, energy-efficiency but does not consider latency as an important optimization function (specially in non-batched mode).

2. Please explain Equation 3 on throughput. Why is utilization included in the equation?

3. Energy-total in Equation 4 is not exactly energy efficiency. It is simply total energy. I would have liked to see performance per watt as a metric instead.

4. The hardware design section can benefit from a high-level classification of the architectures used in FPGA implementation and their summary: for

example systolic array, daisy-chain like Intel-DLA architecture etc. Currently the focus is on simple low-level optimizations.

5. The description of Figure 3 should be improved. I read this part many times and still could not understand what Figure 3(a) and 3(b) are representing.

6. The approaches to efficiently utilize on-chip memory is missing from the survey, even though it is an important consideration in most designs.