

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.

Reviewer 1

Reviewer 2

The paper is much improved. My main concern is that there are still numerous grammatical issues and still some typos in the paper. A significant amount of copyediting is required.

I have a number of small issues remaining that should be addressed:

1. p 11.4: SRAM is not defined
2. Figure 2b needs more explanation. Be more explicit on what the x and y axes represent and the dotted lines.
3. p 11.5: I don't think batch, batch size, layer pipeline have been defined. Remember that the reader is not necessarily fluent in all the terminology.
4. p 11.10: sect 5.1.2 we estimate the theoretical benefit. sect 5.1.3 we estimate the theoretical hardware performance gain as $2x$. How are you making these estimations? You have to say enough for the reader to make a judgement as to whether your estimations are valid.
5. p. 11.16: Fig. 6 - I suggest putting FP in squares and GPU in triangles to make them stand out even better and make it visually easier to see where the different approaches lie. It is especially important for those that are color challenged!
6. "Overall, the improvement does not match the estimation ..." What estimation?
7. p. 11.17: Fig. 7. You do not comment on the Logic/BRAM chart. If you do not say anything about it in the text, then it should be removed.

Reviewer 3

A survey paper needs to add value to its readers. The readership of a survey paper are researchers who want a snapshot of the state-of-the-art, understand the trends, as well as challenges and opportunities. The original manuscript failed to provide this value. The revised manuscript adds some superficial changes and a nice Section 2.1. But beyond that the authors simply ignored most of my concerns. As such I cannot recommend this paper for publication.

In particular, the survey only focuses on a narrow topic, namely CNN inference with primarily quantization and unrolling as the only optimizations. These optimizations are only at the kernel level; but the survey does not consider higher design architectural design issues such as how these kernels are put together or even CNN consisting of multiple layers and communications among layers are put together. The survey is completely missing the big picture, the system level design issues, memory bandwidth and storage issues, major architectural trends etc.

1. **As your survey focuses on CNN and inference, please state that directly in the tile and abstract.**
2. **Section 3 remains as confusing as before. In Equation (1) what is actual performance? What's the metric? What is workload? What's the metric again. How is utilization measured? What is the unit of throughput?**
3. **As there has been some time between the submission and current time in a fast moving field, the authors need to include newer systems, for example, CHaiDNN etc.**
4. **There are numerous typos and grammatical mistakes even in this version.**