Reviewer #1:

First of all, I want to thank the authors sincerely. They do make some changes to the manuscript and updated quite some part of it. However, I think there are still major issues that need to be addressed.

Some of my questions/comments are list below, please consider carefully.

1. As one of the critical contribution claimed by the authors, the quantization related design, including 1) the design space searching for optimal quantization decision; b) hyper-parameter quantization, have been presented in Section 3. However, Section 3 does not provide any specific design solution rather than just pointing to the experimental sections as for the result based analysis. I did not obtain any 'decision making' information from Section 3 for the quantization for the different data for the model.

2. From Fig.3, the the conv-1, conv-2 and FC-1 layers in the accelerator is pipelined for the execution, then how about the rest of the modules (Norm, Vec. Pro. for FC-2, FC-3)?

3. Another major issue is the style of the Figures and Tables in this manuscript. I hope the authors could carefully refer to other scientific paper for the presentation of the Figures and Tables.

4. For the quantization results shown in Figure 8, what is the baseline floating point counterpart?

5. The authors may want to notice the format requirement of the journal. It become hard to follow the content when the manuscript has to breakpoint in the middle.