You are being carbon copied ("cc:'d") on an e-mail "To" "Jingfei Jiang" jingfeijiang@nudt.edu.cn

CC: litao@nankai.edu.cn, "Dong WEN" 310-we-aaa-1@163.com, "Yong Dou" yongdou@nudt.edu.cn, "Jinwei Xu" jinwei200911@163.com, "Tao Xiao" xt@nudt.edu.cn

Dear Dr. Jiang,

We have received the reports from our advisors on your manuscript, "An Energy-efficient Speech Classification Convolution Neural Network Accelerator Based on FPGA and Quantization", which you submitted to CCF Transactions on High Performance Computing.

Based on the advice received, I have decided that your manuscript could be reconsidered for publication should you be prepared to incorporate major revisions. When preparing your revised manuscript, you are asked to carefully consider the reviewer comments which can be found below, and submit a list of responses to the comments. You are kindly requested to also check the website for possible reviewer attachment(s).

While submitting, please check the filled in author data carefully and update them if applicable - they need to be complete and correct in order for the revision to be processed further.

In order to submit your revised manuscript, please access the following web site:

https://www.editorialmanager.com/thpc/

Your username is: \*\*\*\*\*\*\*\*

If you forgot your password, you can click the 'Send Login Details' link on the EM Login page.

We look forward to receiving your revised manuscript before 11 Sep 2020.

With kind regards,

Tao Li, Ph. D.

Lead guest editor

CCF Transactions on High Performance Computing

Comments to the author (if any):

Reviewer #1:

1.This paper need a major revision. The current expression of the novel ideas are not clear**. Quantization is a commonly used technique for FPGA based implementations and multi-PE is also not totally novel. I believe the systematic design of the paper should be a good one, but it is not convincible enough now.**

2. The 'level' in the paper is used wrongly, it should be 'layer' according to the architecture of the CNNs. Please double check all the scientific naming of the key words and revise them.

3. Fig. 1 is not referred and also not helpful to the main context.

4. It is not clear through the entire paper why the CNNs show advantage to the temporal models on the temporal data, which is the speed data. Please further extend is and explain it clearly.

5. The idea that Fig.2 is presenting is too simple, should not use a figure but just explanation is enough to understand.

6. Does the 'level-by-level' pipeline actually pipelining the 'layer'? if so, it should be 'layer-by-layer' pipelining. Again, please verify the correctness of all the names of the proper nouns.

7. Fig. 10 is hard to understand, how about using the bitwidth as the x-axis?

Reviewer #2: Overall, this paper focuses on speech classification model with quantization. The authors also provided FPGA implementations.

I have two major concerns about this submission.

First, while this paper focuses on RNN with quantization on FPGAs, it is important to understand what are the major contributions and what are the key advances. **The two compared baselines are all published in 2015, which is already five years. I encourage the authors to compare with more recent and state-of-the-art FPGA implementations to clear show the advances of this paper.** Also, the **quantization has been studied widely, what are the differences in this paper? How does the FPGA implementation compared against other ASIC solutions?**

Second, the writing could be improved largely, including the figures and tables. For example, The Figure 2 is quite confusing. Where do the extra mantissa come from in the two fixed-pointed data? What does the accuracy mean? In the caption, what does the "bitwise" mean?

Reviewer #3: The authors propose a convolution neural network accelerator for speech classification. Weight and activation are represented by binary and fix-point data, respectively, which is faster and more energy-efficient than floating point. The authors also design the accelerator with shared weight storage and balanced pipeline structure, eliminating accesses to DRAM. In a word, the performance and energy-efficiency for this specified convolution neural network are improved notably over CPU platforms and previous RNN accelerator.

The idea of this paper is not that novel although he improvement of performance and energy-efficiency of this accelerator is notable. Quantization has been widely used in conversion from floating point to fix-point number. **Are there any new ideas in your quantization method?**

In Section 4.3, the authors said FC-1 is layer is the bottleneck of this accelerator, is this because of the massive computation on parameters? **If so, why does the bottleneck become Conv-2 layer in Section 5.1?**

The title of Section 5.1 is "Quantized Model's Performance". However, the authors firstly show the accuracy result with various bitwise. Then the result of non-quantization version on CPUs is listed in Table 2. **Why is the result for quantization not shown, or what is the relationship of Fig.10 and Table 2? In addition, in Fig. 10,** the accuracy is best when the decimal bitwise is 8 and 9 for middle results and normalized results, respectively. **The question is why the accuracy with more bits is worse, e.g., the accuracies of the third, fourth and fifth groups in Fig.10.**

In Section 5.2, the authors compares the results with previous accelerators based on RNN. Firstly, both of these two RNN works are not the most recent. **Secondly, I think it is better to compare the result with CNN based accelerators.**

These is no section about related work.

Many figures are not cited in the paper, e.g. Fig.1, Fig.5, Fig. 6, Fig. 8, Fig.9, and there are no specified descriptions about these figures, either. Some figures are not clear, which needs to be enhanced. There is no Table 3 in this paper.

\*\*\*\*\*\*\*\*

\*\*Our flexible approach during the COVID-19 pandemic\*\*

If you need more time at any stage of the peer-review process, please do let us know. While our systems will continue to remind you of the original timelines, we aim to be as flexible as possible during the current pandemic.

This letter contains confidential information, is for your own use, and should not be forwarded to third parties.

Recipients of this email are registered users within the Editorial Manager database for this journal. We will keep your information on file to use in the process of submitting, evaluating and publishing a manuscript. For more information on how we use your personal details please see our privacy policy at https://www.springernature.com/production-privacy-policy. If you no longer wish to receive messages from this journal or you have questions regarding database management, please contact the Publication Office at the link below.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

In compliance with data protection regulations, you may request that we remove your personal registration details at any time. (Use the following URL: https://www.editorialmanager.com/thpc/login.asp?a=r). Please contact the publication office if you have any questions.