(R1Q1) More extensive and thorough experiments are needed.

(A) Since our major contribution is to enable fractional sub 1-bit weight quantization, we selected models that have been previously quantized by ‘1’ bit/weight for comparisons on the model accuracy. As a result, unfortunately, the range of model selections is somewhat limited as we present in Table 1 and 3.

(R1Q2) What about average ratio when different layers are quantized with distinct integer ratio of bits per weights?

(A) We acknowledge that fractional numbers can be obtained when considering the average number of quantization bits. The number of quantization bits per layer is, however, still limited to be an integer such that the minimum number of quantization bits is ‘1’ while we proved that some layers can be quantized with 0.35 bits/weight without noticeable accuracy degradation (as shown in Table 2). Sub 1-bit quantization is only available through FleXOR.

(R2Q1) Every weight utilizes less than 1b? Or do some weights use >1b while other can use much less?

(A) It would be an exciting research topic to study the distribution of the optimal number of quantization bits for each weight. We believe that such distribution would be wide and some weights require >1b while numerous weights need <1b because 1) increasing N_{in} and N_{out} allows such distributions to be wider and enhances model accuracy even for the same compression ratio and 2) as shown in Table 1, model accuracy of 1-bit quantization with FleXOR is higher than other quantization schemes that do not include encoding schemes. We will add related discussions with supporting experiments in the final manuscript.

(R2Q2) The paper doesn’t discuss any other possible methods for accomplishing a similar thing.

(A) We understand that there are numerous methods to reduce the neural networks in size. For example, low-rank approximation and parameter pruning could be additionally performed to reduce the size further. We believe that such methods are orthogonal to our proposed method while we feel that our contributions in this work are 1) rethinking the limits of quantization method through encoding schemes and 2) providing a new compression method using the techniques that have been widely studied in digital communication or cryptography.

(R2Q3) Issues on clarity

(A) We appreciate your careful reviews. We will revise the abstact, correct typos, and introduce some definitions clearly in the revised manuscript.

(R3Q1) Figure 1 uses binarized inputs. The reviewer did not find results in the paper that used quantized inputs.

(A) We will fix Figure 1 since inputs are not quantized. “Input weight format” should read “Internal weight format.”

(R3Q2) Algorithm 1 considers flattened tensors. Encoded weights risked inference between consecutive filters.

(A) Algorithm 1 describes hardware operations (that are best implemented by ASIC or FPGA) rather than instructions to be operated by CPUs or GPUs. We acknowledge that Algorithm 1 may lead to significant overhead when implemented by general-purpose computing systems. We will add discussions to address your concern.

(R3Q3) In both Tables 1 and 3, the authors reported 1-bit weight with full-precision activation results from other methods. The reviewer believes that this is not sufficient.

(A) We want to point out that the goal of our work is not to present the best model accuracy with 2-3 quantization bits. Previously, it has been known that the form of binary neural networks is the minimum for quantization. On the other hand, throughout this work, we show that fractional sub 1-bit quantization is possible by using XOR gates. Accordingly, we selected models that have been quantized by 1-bit per weight while we could not quantize activations because once activations are quantized, model accuracy is affected by the choice of activation quantization method. Activations can be quantized additionally in our work.

(R4Q1) FleXOR can achieve better memory saving with similar accuracy as previous SOTA, but the gain is marginal.

(A) Outperfoming previous models even with sub 1-bit quantization can be important but not our major target in this work. As we include the results with 0.63 or 0.6 bits/weight in Table 3, our goal is to demonstrate that even sub 1-bit quantization is enabled by FleXOR with graceful accuracy degradation. We believe that FleXOR presents a new inference design paradigm with much wider trade-off search space between model accuracy and compression ratio.

(R4Q2) Clarify how to structure XOR-gate networks.

Since an XOR-gate network is shared by many weights (such as 1 million), it is difficult (if not impossible) to manually optimize an XOR-gate network. Hence, random network configuration is enough to fulfill the purpose of random number generation. In short, XOR-gate network design is simple and straightforward. We will add discussions to address your concern in the revised manuscript.