Reviewer-Two

Summary:

This paper presents an ANN algorithm on TPU and analyzes the memory and instruction bandwidth of ANN algorithms.

Strengths:

1) The problem is well-motivated, 2) The related work is comprehensively studied and discussed. And the entire story is easy to follow for readers, 3) The methodology to study the algorithm from the hardware bottlenecks is interesting.

Weaknesses:

1) I suggest the authors to include a brief discussion of TPU, e.g., what it can be done and what it can not be done (efficiently). 2) The experimental evaluation is a bit problematic. How about other methods on GPUs/TPUs, e.g., hashing and graph-based methods besides the FAISS baseline? 3) In Figure 3, the highest recall shown for Glove is 0.9. Is there a recall limitation for the proposed algorithm? 4) The technical contribution of the bi-stage partial reduction and scoring is limited.

Questions:

1) Could we have other methods on GPUs/TPUs, e.g., hashing and graph-based methods besides the FAISS baseline, on more measures, e.g. cosine? 2) Could we show that the proposed algorithm can achieve high recalls for most ANN datasets?

Limitations:

I suggest the authors to discuss the limitation of the algorithm, e.g., how to adapt the problem to other common similarity measures.

Ethics Flag: No

Soundness: 2 fair

Presentation: 4 excellent

Contribution: 2 fair

Rating: 5: Borderline accept

Confidence: 5

Code Of Conduct: Yes