1 Summary

The Architectural Implications of Facebook's DNN-based Personalized Recommendation (for the 2020 IEEE International Symposium on High-Performance Computer Architecture): The authors of this paper give an in-depth description and characterization of performance metrics of production-level recommendation systems using FB's open sourced DLRM benchmarks. The rationale for this is that publiclaly available deep neural network (DNN) based recommandation benchmarks do not represent those actually used on production-scale recommendation systems in data centers. The problem, therefore, is this gap between currently available benchmarks and realistic production scale benchmarks, despite these deep learning recommendation systems being used so widely and incresingly in industry. Using their performance analysis of 3 major recommendation models (RMC1, RMC2, RMC3) and using FB's open source suite of synthetic models, the authors aim to provide realistic production benchmarks; this information would ideally motivate future optimization and innovation for recommendation system designs by providing more insight into the computing requirements, storage capacity, and memory access patterns of real production-scale recommendation systems, which these models show have differing efficiency challenges than other traditional models.

2 Strengths

- The context the authors added in the intro/background and mentioning familiar use cases on what function a recommendation model serves or what a feature accomplishes (ex. curating news feed/video rec/social media posts, data sources from clicks to demographic information, etc.) helped frame a lot of the purpose of their experiments and the purpose of the performance metric analyses that they decided to do. This relation between some low level mechanism to a broader use case in layman's term demonstrates significance of what they are studying and makes it more understandable to someone not familiar with recommendation systems or ones at server-level.

- Most of the figures and tables the authors included to explain the differences between the 3 recommendation models they examined and the differences in varying performance areas were simple and clear, and seemed appropriate for the given audience (in a IEEE Symposium).

3 Weaknesses

- The way they presented their observations in two sections, under "Understanding XYZ Model(s)", as several "takeaways" seemed like a disorganized way to discuss their results, but I understood the intention to discuss particular metrics distinctly and apply what the specific "implication" of their observations were.

- The authors did not include performance data of the 3 models they looked at using non-representative benchmarks, in order to compare against the DLRM benchmarks.

4 Rating: 4

5 Comments

While I am not particularly familiar with recommendation systems, much less how they scale to the size of data centers, this paper does a fairly good job of framing their observations and suggestions for optimizations in lay terms, while also keeping their discussion at a level appropriate for the given audience (at a IEEE Symposium). For example, their examination of embedding tables, effects of co-locating, the differences in performance behaviors of the 3 models, etc. remained technical and used data they observed, but was often linked with their interpretted implications and/or explanation of cause and effect, which were succinct and clear. That being said, how they presented their results could have been more organized. The authors used brief "takeaway-messages" to introduce difference metric discussions by highlighting a key observation or problem, and often finished them with suggested solutions. This happens several times under two section headings, "Understanding XYZ Model(s)", but I think several "takeaways" could have been better presented as a traditional Results/Discussion section (which I'm unsure is the norm in computer architecture papers the same way chemistry or biology papers expect). Another way the authors could improve on their paper would be the inclusion of the same experiment they did here but using the publicly available benchmarks they said were not representative of production use case. They briefly describe and cite other sources of how they differ from the benchmarks that was used, but a comparison of the DLRM against something like the MLPerf-NCF on the 3 recommendation models they chose to examine would have made their point (that ones like the MLPerf aren't good benchmarks for scaled systems) better. The topic of the paper is interesting and extremely relevant, as plenty of other areas of research closely focus on the designs and systems in data centers, specifically (ex. networking research in DCs); the services and work done in industry data centers is prevalent. Since these services are likely to continue to grow (and therefore their resource demands increase, as well), it is important to examine and address the unique challenges, resource requirements, and direction of growth for computers of these scale; the authors make a compelling point that there should be more reliable representative resources available that help build these solutions.