Amy Bui

1 Summary

A Hierarchical Neural Model of Data Prefetching (2021): The authors in this paper introduce Voyager, a neural network model trained for data prefetching. Their neural model has less overhead than previous models; it is 110-200× more compact in terms of storage, and 15-20× less computationally expensive. In their approach, they re-frame the data prefetching problem as a probabilistic prediction problem, and the outputs (address prefetches) are represented as probability distributions, rather than prefetching from grounded, labeled truths. Some of the unique benefits of their design, compared to previous prefetchers, is that Voyager can learn both address correlations (temporal prefetching) and delta patterns (strides) because of the hierarchical structure of the implemented embedding layers. The techniques they introduce to their model address some of the common problems found in using data prefetchers. The class explosion problem in data prefetching (having to predict addresses from among 2^{64} unique addresses) is dealt with by decomposing memory address prediction into two subproblems (embedding layers) in their hierarchical model, called page prediction and page-aware offset prediction. For compulsory misses for addresses requested at low frequency, the model is able to learn the delta patterns to prefetch them, rather than spend the work learning address correlations. And Voyager addresses the labeling problem by predicting the most predictable addresses when given multiple possible labels using the authors' multi-label training scheme. The authors show comparisons against other data prefetching neural models, as well as hardware prefetchers; while they conclude that their neural network, like previous models, is still too computationally expensive to be in hardware prefetchers, Voyager is a significant contribution towards the development of more practical prefetchers.

2 Strengths

- The paper picks a good variety of other baseline prefetchers to compare Voyager against, as well as discussing the metrics or parameters that are specifically relevant to data prefetching.

3 Weaknesses

- Certain metrics and applications the authors selected for analyzing Voyager's performance against baseline prefetchers appear biased, and much of the results discussion focused more on comparing Voyager to one other (arbitrarily) picked baseline prefetcher (ISB) without explanation, despite earlier mentions of 4 other baseline prefetchers.

4 Rating: 4

5 Comments

This work yields a model that not only performs as well as, if not better than, previous models for irregular access patterns, but also works with more reduced overhead than any previous model. While the paper details a design that sensibly addresses problems of class explosion, offset aliasing, and labeling, the way in which they present overall performance compared to the 5 other baseline prefetchers make their experiments seem biased. For example, to compare performance in terms of accuracy, coverage, and IPC, the paper only picks specific sets of irregular benchmarks from SPEC06 and GAP; the paper did not include other applications where the other baseline prefetchers' features are tuned to or are known to have previously performed well. Second, when discussing how Voyager performed when varying the prefetching degree (Fig. 9), the paper only compares it to ISB and a hybrid ISB+BO; the authors are not clear why the temporal prefetcher, ISB, was selected, unless the audience should know or assume ISB was the "best" of the 5 baseline prefetchers. Likewise, a similar preference for only comparing Voyager to ISB is seen when they discuss performance on spatial/non-spatial patterns, but no explanation is given as to why BO, STMS, Domino, and Delta-LSTM are excluded. And finally, the results in Figure 15, which compares performance (given by their own unified definition of accuracy and coverage) of individual labeling schemes to Voyager's hybrid multi-labeling scheme, they do not clarify how they arrive at those aggregated numbers; moreover, it is mentioned that different benchmarks may have their preferred labeling schemes, but the paper did not show or discuss how the other models performed on benchmarks using these preferred schemes versus the hybrid scheme. Other than the way certain results are presented, their methods do a good job of isolating problems in the area of neural data prefetching seen in previous works and systematically addressing them in specific components of their model. The authors' Voyager model is a new and exciting contribution to data prefetching neural networks, even if it remains impractical.