Mark Zhao
Meta Ph.D. Fellowship Research Statement

Hardware/Software Co-design for Training Data Management at Scale

Industry-scale ML training relies on massive data storage and ingestion (DSI) infrastructure that generate, store, and ingest exabytes of training data to feed scale-out trainers. My recent research [DSI, ISCA'22] highlights the critical need to improve the performance, scalability, and cost-effectiveness of this infrastructure for data-intensive recommendation models. Because training infrastructure is highly inter-dependent, these improvements require co-designed hardware and software optimizations across the DSI and training pipeline. Unfortunately, current production DSI infrastructure relies on a patchwork of interconnected systems, resulting in poor abstractions between DSI systems, training clusters, and ML engineers that hamper these cross-cutting optimizations.

To address these large-scale system challenges, my research proposes a framework for a unified DSI ecosystem. It aims to lift ML abstractions (datasets, features, transformations) up from their underlying infrastructure details (Hive tables, schemas, ingestion pipelines), and enable broad and independent optimizations to model development velocity and training infrastructure performance. I plan to pursue my research in two phases.

In the first phase, I will investigate proper abstractions for recommendation model datasets. ML engineers need to perform myriad tasks, such as analytics and preprocessing, over the same structured categorial and embedding features. As a result, engineers use multiple interfaces (SQL, PyTorch, dataframes) and engines (DPP, Spark, Presto), resulting in duplicate efforts and wasted resources. I will build a framework that allows users to specify queries and logical ingestion pipelines, from raw data inputs to tensor outputs, in an easy-to-use interface (e.g., dataframes). Users will not need to define infrastructure details such as execution frameworks or intermediate tables and schemas, allowing me to extensively optimize the underlying systems.

In the second phase, I plan to explore novel hardware/software co-designed optimizations that are unlocked by this logical view of the end-to-end DSI pipeline. For example, the framework can analyze how training jobs share transformations and transformations' performance characteristics across heterogeneous hardware. It can then choose the right execution engine (e.g., Spark or DPP), schedule compute on optimal hardware (e.g., distributed CPUs, trainer-local CPUs, or trainer GPUs), and decide how to manage intermediate data (e.g., Hive tables, DRAM, or SSD caches). By understanding training jobs' ordering, feature length, and quantization requirements, my framework can further enable end-to-end compute and storage optimizations by transparently clustering (i.e., compressing), truncating, and adjusting precision without affecting model accuracy.

My research is a good fit for the AI System HW SW Co-Design fellowship because it spans Meta's end-to-end training pipeline. It can enhance experiences for thousands of ML engineers and directly improve the performance and scalability of Meta's gigawatt-scale training infrastructure. My recent experience at Meta provided me the tools and understanding to tackle these impactful problems, as evidenced by two high-impact projects I helped conceive, develop,

and deploy in Meta's production training infrastructure: [TectonicShift, under submission] and [BatchReuse, in preparation]. Furthermore, I plan to leverage Meta's open-source software including Velox, TorchArrow, and TorchRec in my research, allowing the cross-cutting techniques I discover to further benefit the wider community.