Querying the World Efficiently and Reliably using Machine Learning

Daniel Kang

Machine learning (ML) has the potential to enable analytics about the real world by querying unstructured data (e.g., videos, text) with high accuracy ML methods, e.g., expensive deep neural networks (DNNs). Scientists and analysts can use these capabilities to understand the real world: city planners can ask how many cyclists passed through an intersection [2], ecologists can perform ecological analysis on hummingbird feeding patterns [4], and firefighters can use cameras for early detection of wildfires [7].

Unfortunately, there are two key barriers to the adoption of ML at scale: cost and reliability. ML methods can be computationally expensive, costing up to hundreds of billions of floating-point operations. This computational cost directly translates to dollar costs that are infeasible for many organizations, e.g., analyzing a year of video can cost over $200,000. Furthermore, ML methods can be unreliable, causing errors in queries or other downstream applications. These barriers require rethinking standard data management techniques when deploying ML for analytics.

To address these barriers, I study and build data management systems that facilitate the use of ML at scale. I have focused on accelerating common ML-based queries using cheap approximations *while* providing statistical guarantees on query results. My algorithms can improve query processing speeds by orders of magnitude. Furthermore, I have applied insights from my work to accelerate traditional ML pipelines and studied broader robustness guarantees.

In my research, I have collaborated with and deployed my research with scientific and industrial partners, including Stanford biologists and researchers at Toyota Research Institute, an autonomous vehicle company. By collaborating with practitioners, I have been able to understand real-world problems and find generally applicable solutions. I have focused on carefully benchmarking and understanding *end-to-end* application semantics. Given this understanding, I have designed algorithms, built systems, and constructed programming abstractions as the problem at hand demands. I believe efficient and reliable queries have the potential for good, including applications ranging from early wildfire detection to ecological analysis.

1 Query Processing Algorithms for ML-based Queries

Unlike in standard queries over structured data, the primary cost in querying unstructured data is extracting the structured information via expensive, *target DNNs*. In many applications, it is infeasible to exhaustively extract this information, so this extraction must be done at query time. As a result, many standard query processing techniques cannot be applied and data management with ML must be rethought.

In one line of my dissertation research, I have focused on generating and using cheap approximations, called proxy models, to accelerate ML-based queries. Proxy models are substantially cheaper than expensive ML models, but can be inaccurate, which is not acceptable in many applications. To rectify this, I have developed algorithms to accelerate general classes of queries: selection, aggregation, and limit queries *with statistical guarantees on query results*. I further show how to efficiently generate these approximations.

Selection queries (classification). An important class of queries are selection queries, in which the user wishes to select records matching a predicate, e.g., frames of a video containing a hummingbird. I explored generating cheap approximations (which I refer to as proxy scores) in the NoScope system [10]. NoScope trains a proxy model to approximate whether or not a data record satisfies the target DNN-based predicate. The proxy model is used to generate a proxy score per data record, which is combined with the target DNN to answer queries. NoScope can improve approximate selection by orders of magnitude compared to the solution of exhaustive labeling. NoScope has inspired other research for ML-based data analytics [1, 3, 19]. Furthermore, I have shown that proxy models can accelerate general classes of traditional ML workloads [18], e.g., data-transformation bound workloads.

Selection queries with guarantees. While NoScope and other systems [1, 3, 19] can accelerate approximate selection queries, they does not provide *statistical guarantees on the recall of the returned set*. These guarantees are critical for scientific rigor. For example, our collaborators in the Stanford biology

department wish to find rare events of hummingbirds feeding in wildlife video. To ensure scientifically valid inferences, they require statistical guarantees on the recall of the discovered hummingbirds. I am actively working on deploying SUPG to this application.

To obtain statistical guarantees, I have developed SUPG, query semantics and sampling algorithms for approximate selection queries with guarantees [11]. Our algorithms selectively sample the target DNN and optimize confidence intervals over the samples, which provides the statistical guarantees. Prior work uses uniform sampling, which I show results in poor quality results (i.e., returned sets with low precision). To improve the sampling efficiency, SUPG instead use a novel set of weights for importance sampling. I show that our algorithms can improve query quality by up to 30× for a fixed budget.

Other queries with guarantees. I have also developed algorithms to optimize aggregation (computing a statistic over the data records), aggregation with predicate (computing a statistic over a subset of the data records that satisfy a condition), and limit (finding a limited number of records that match a set of predicates) queries [9, 12]. Perhaps surprisingly, we show that these different queries require different algorithms.

BlazeIt, which optimizes aggregation and limit queries, reduces variance in sampling for aggregation queries and ranks rare events for limit queries. In contrast, ABae, which optimizes aggregation with predicate queries, uses stratified sampling based on proxy models to avoid sampling records not satisfying the predicates. We show that the convergence of ABae requires novel analysis of stratified sampling with stochastic draws. Both systems can improve query execution times by orders of magnitude compared to baselines.

Efficient indexes for proxy scores. While proxy scores can accelerate many query types, they can be inefficient to deploy. A common method of generating proxy scores is to train a new, cheap model *per query* to approximate the expensive target DNN. Unfortunately, this method does not share work across queries, requires ad-hoc training methods, and requires many target DNN annotations for training data.

To address these issues, I have developed TASTI, a general-purpose method for constructing proxy scores for unstructured data via an embedding index [13]. TASTI pre-computes embeddings that can be used to place records that are close under target DNN outputs together and annotates a small fraction of the records. To generate scores, TASTI assigns close records (by embedding distance) to the value of the nearest annotated record. Because these embeddings are pre-computed and are designed to work for any query over the target DNN output, they can be reused across queries and query types (including every query type I described above). I show that TASTI is simultaneously over 10× cheaper at index construction time and can return query results up to 24× better than ad-hoc proxy models.

Efficiently executing visual analytics [15]. Recent research, e.g., new accelerators, has greatly improved the throughput of DNNs by up to 150×. While this work has improved the throughput of *DNN execution*, it ignores other costs. I show that the *preprocessing* of visual data (e.g., image decoding) now bottlenecks end-to-end DNN inference for visual analytics systems by up to 23×, in the first measurement study of its kind [15]. To address this bottleneck, I built Smol, a system that jointly optimizes preprocessing and DNN execution for improved end-to-end DNN inference. Smol leverages low resolution visual data, partial decoding, and preprocessing-aware cost-based optimzation to balance preprocessing and DNN execution. These optimizations can improve throughput by up to 5.9× at a fixed accuracy.

2 Model Robustness and Quality Control

Robustness of ML is a key barrier to widespread adaption. In mission-critical systems, errors in models can have cascading effects, e.g., safety violations in autonomous vehicles. In query processing, errors in the target DNN will be reflected in results, as guarantees are with respect to the target DNN. I have developed methods to monitor ML methods, improve training data quality, and methods of measuring robustness. My work on model assertions is being deployed at an autonomous vehicle company.

Assertions for ML and retraining. As ML methods continue to improve on benchmark tasks, they are increasingly being deployed in mission-critical settings, such as autonomous vehicles. However, average-case measures of performance can hide potentially critical errors. While software testing has developed many tools for testing critical software, but is not directly applicable to ML.

My work has taken steps to bridge these two views. I've developed two abstractions, model assertions [16] and learned observation assertions [8]. Both abstractions are used to find potential errors in ML model predictions and human labels.

Model assertions allow users to specific specific forms of potential errors. For example, consider a state-of-the-art object detection DNN deployed over video to detect cars. Even state-of-the-art can fail simple assertions, such as temporal consistency, e.g., that a car should not appear and disappear rapidly in a video. Learned observation assertions (LOA) leverage existing human labels to learn when there may be discrepancies in ML model predictions or new, possibly erroneous, human labels. Model assertions and LOA can find errors with high true positive rate, at least 75% in all cases we studied.

Furthermore, I showed that assertions can be used in retraining ML models. Organizations continuously collect data to retrain ML models as they are deployed over new scenarios, e.g., autonomous vehicles seeing new streets. It is critical to select data that will improve the model as the majority of data is uninteresting. I showed that model assertions can be used to select "difficult" data (i.e., data that the model fails on), which improves ML model quality more than baselines. Assertions can reduce labeling costs by up to 40% at a fixed budget by finding such data.

Broader robustness guarantees. Robustness of ML models is also of broader interest. I have studied broader robustness guarantees by understanding end-to-end application concerns. First, I have developed a method for training natural language generation (NLG) models to be robust against noise in training data [14]. Second, I have developed a method of measuring the robustness of models against adversaries not seen at training time, which is more reflective of reality [17].

3 Benchmarking ML Pipelines

As ML systems have improved, it has become increasingly difficult to compare the performance of these systems. Existing work has measured proxy metrics, such as the time to process a single minibatch of data, but these metrics are not indicative of producing a high quality result, e.g., DNNs with high accuracy.

I was part of the founding team for DAWNBench [5, 6] and MLPerf [20], which have set standards for comparing DNN systems and is now widely used in industry and academia. We introduced the *time-to-accuracy* (TTA) metric, which measures DNN training systems by the *end-to-end* training time required to achieve a state-of-the-art accuracy. Using TTA, we showed that optimizations can interact in non-trivial ways, e.g., producing lower speedups, demonstrating that proxy metrics are not sufficient for measuring DNN systems [5]. I also helped develop MLPerf, an industry and research consortium that uses TTA a metric to measure DNN systems [20].

4 Applications and Future Research

While my work has shown the promise of ML-based analytics, I believe we have only begun to enable scientists and organizations to use ML. Increasingly, the data we collect can be used in high impact ways, particularly in analytics in rare events and high-stakes analytics.

ML-based analytics over rare events. Analytics over rare events are increasingly important and will require new techniques. For example, I am collaborating with Stanford biologists to find rare events of hummingbird visits: their prevalence is under 0.1%. Unfortunately, off-the-shelf models perform poorly and it is difficult to obtain enough samples of hummingbirds to train a proxy model as the vast majority of the video is empty. Furthermore, scientific analyses require statistical guarantees on accuracy, i.e., recall of hummingbird visits, which off-the-shelf methods do not provide.

This applications is indicative of a larger class of analytics: analytics over rare events, which is challenging as standard methods of data collection and training DNNs do not perform well in sample-limited regimes. I plan to develop methods that bootstrap DNNs for end-to-end analytics over rare events. While seemingly simple, I believe there are many research questions that will arise. For instance, how should we select which data records to label? How should we split between iteratively training DNNs and performing inference over large quantities of data? How can we achieve statistical guarantees when combined with iterative data exploration?

I plan to leverage my expertise in ML-based query processing to answer these questions.

ML-based analytics for decision making. Another important class of analytics will be analytics that inform high-stakes decisions. As an example, I am collaborating with staff at a nature preserve for early wildfire detection via cameras. In conjunction with ALERTWildfire, Jasper Ridge nature preserve has installed cameras to detect wildfires, as they do not have the resources to patrol in person. This application has several features that I believe will becoming increasingly common. Similarly to the settings I have studied, there are limited resources, but there are several key differences. First, many applications require *low latency* responses as actions must be taken to mitigate potential hazards, in contrast to the batch setting where throughput is the primary metric. Second, conditions can constantly change, e.g., because of seasonality or changing wildfire conditions. This introduces problems of data drift, which is also not present in the batch analytics setting. I plan to develop systems and algorithms to reason about how analytics inform decision making, including using my expertise in monitoring models to understand when to trigger retraining and my expertise in building high performance systems to ensure low latency in the face of resource constraints.

Other query types. While I have shown that certain ML-based queries can be accelerated, there is much work to be done. For example, many applications would benefit from accelerated systems for joins, group bys, and nested queries. Furthermore, more complex queries will require query optimization for efficient execution. I plan to explore these queries by developing new algorithms and query processing techniques.

My research approach has been to understand end-to-end applications and develop widely applicable algorithms, systems, and programming abstractions to solve the general problems that have arisen from these applications. While my research has shown the promise of ML-based queries, I believe there will be a range of new, impactful applications. I plan to continue to take my principled approach towards solving problems to enable these applications.

References

[1] Michael R Anderson, Michael Cafarella, Thomas F Wenisch, and German Ros. Predicate optimization for a visual analytics database. *ICDE*, 2019.
[2] C. Bautista et al. Convolutional neural network for vehicle detection in low resolution traffic videos. In *TENSYMP*. IEEE, 2016.
[3] C. Canel, T. Kim, G. Zhou, C. Li, H. Lim, D. Andersen, M. Kaminsky, and S. Dulloor. Scaling video analytics on constrained edge nodes. *SysML*, 2019.
[4] Callie R Chappell and Tadashi Fukami. Nectar yeasts: a natural microcosm for ecology. *Yeast*, 35(6):417–423, 2018.
[5] C. Coleman, D. Narayanan, Daniel Kang, T. Zhao, J. Zhang, L. Nardi, P. Bailis, K. Olukotun, C. Ré, and M. Zaharia. Dawnbench: An end-to-end deep learning benchmark and competition. *NeurIPS MLSys Workshop*, 2017.
[6] Cody Coleman, Daniel Kang, Deepak Narayanan, Luigi Nardi, Tian Zhao, Jian Zhang, Peter Bailis, Kunle Olukotun, Chris Re, and Matei Zaharia. Analysis of dawnbench, a time-to-accuracy machine learning performance benchmark. *ACM SIGOPS*, 2019.
[7] K. Govil et al. Preliminary results from a wildfire detection system using deep learning on remote camera images. *Remote Sensing*, 2020.
[8] Daniel Kang, Nikos Arechiga, Sudeep Pillai, Peter Bailis, and Matei Zaharia. Finding label and model errors in perception data with learned observation assertions. *SIGMOD*, 2022.
[9] Daniel Kang, Peter Bailis, and Matei Zaharia. Blazeit: Optimizing declarative aggregation and limit queries for neural network-based video analytics. *PVLDB*, 2019.
[10] Daniel Kang, John Emmons, Firas Abuzaid, Peter Bailis, and Matei Zaharia. Noscope: optimizing neural network queries over video at scale. *PVLDB*, 2017.
[11] Daniel Kang, Edward Gan, Peter Bailis, Tatsunori Hashimoto, and Matei Zaharia. Approximate selection with guarantees using proxies. *PVLDB*, 2020.
[12] Daniel Kang, John Guibas, Peter Bailis, Tatsunori Hashimoto, Yi Sun, and Matei Zaharia. Accelerating approximate aggregation queries with expensive predicates. *PVLDB*, 2021.
[13] Daniel Kang, John Guibas, Peter Bailis, Tatsunori Hashimoto, and Matei Zaharia. Task-agnostic indexes for deep learning-based queries over. *SIGMOD*, 2022.
[14] Daniel Kang and Tatsunori Hashimoto. Improved natural language generation via loss truncation. *ACL*, 2020.
[15] Daniel Kang, Ankit Mathur, Teja Veeramacheneni, Peter Bailis, and Matei Zaharia. Jointly optimizing preprocessing and inference for dnn-based visual analytics. *PVLDB*, 2021.
[16] Daniel Kang, Deepti Raghavan, Peter Bailis, and Matei Zaharia. Model assertions for monitoring and improving ml model. *MLSys*, 2020.
[17] Daniel Kang, Yi Sun, Dan Hendrycks, Tom Brown, and Jacob Steinhardt. Testing robustness against unforeseen adversaries. *ICLR (under review)*, 2021.
[18] P. Kraft, Daniel Kang, D. Narayanan, S. Palkar, P. Bailis, and M. Zaharia. Willump: A statistically-aware end-to-end optimizer for machine learning inference. *MLSys*, 2019.
[19] Yao Lu, Aakanksha Chowdhery, Srikanth Kandula, and Surajit Chaudhuri. Accelerating machine learning inference with probabilistic predicates. In *SIGMOD*, 2018.
[20] P. Mattson, C. Cheng, C. Coleman, G. Diamos, P. Micikevicius, D. Patterson, H. Tang, G. Wei, et al. Mlperf training benchmark. *MLSys*, 2020.