Querying the World Efficiently and Reliably Using Machine Learning

Daniel Kang, ddkang@stanford.edu, ddkang.github.io

I build data management systems to deploy machine learning (ML) efficiently and reliably. ML has the potential to enable scientists and analysts to answer queries about the real world: city planners can ask how many cyclists passed through an intersection [2], ecologists can perform analysis on hummingbird feeding patterns [4], firefighters can use cameras for early detection of wildfires [7], etc. I aim to enable these ML-based applications by reimaging the standard data management stack.

Unfortunately, there are two key barriers to deploying ML-based systems: cost and reliability. Executing a state-of-the-art ML model over a year of video can cost $200,000 on the cloud, which is infeasible for resource-constrained organizations, as is the case with my scientific collaborators. ML models can also be unreliable, causing potentially fatal failures (e.g., in autonomous vehicles) or errors in scientific inferences. As a result, it is difficult to apply standard data management techniques for ML-based systems. Many standard techniques (e.g., indexes, predicate pushdown) assume that data is structured and error-free. In contrast, an expensive and error-prone ML model is required to extract structured information in ML-based systems. These unique barriers to ML-based systems require rethinking standard data management techniques.

To address these barriers, my research has taken two broad approaches. First, I have reduced the cost of common ML-based queries and traditional ML workloads by using new systems that leverage cheap approximations *while* providing statistical guarantees on results. Second, I have proposed new techniques to improve the reliability of ML models by improving training data quality and training methods.

My research on query processing has been cited over 250 times and has been extended or used by groups at MIT, CMU, MSR, etc. [1, 3, 17]. It has demonstrated that common ML-based queries can be accelerated by orders of magnitude. To deploy my research for impact, I have been collaborating with Stanford scientists for ecological analysis and to detect wildfires early. I was also on the founding team for DAWNBench and MLPerf (which was inspired by DAWNBench) to benchmark the computational performance of ML systems. MLPerf is now the industry standard used by Google, Facebook, NVIDIA, and academic research. My work on robustness is being used at the Toyota Research Institute, an autonomous vehicle (AV) company, to vet their training data for their mission-critical ML models. I plan to continue to collaborate with industrial and academic partners to deploy my research to enable scientists and analysts to be more efficient.

My research approach has been to collaborate closely with academic and industrial partners to understand real-world problems and find generally applicable solutions. I have focused on benchmarking and understanding *end-to-end* application semantics. Leveraging this understanding, I have built systems, constructed programming abstractions, and designed algorithms as the problem at hand demands.

1 Query Processing Systems for ML-based Queries

Unlike in queries over structured data, the primary cost in querying unstructured data is extracting structured information via expensive methods, which we call *target methods*. It is infeasible to exhaustively extract this information for many resource-constrained applications, so this extraction must be done at query time. Thus, standard query processing techniques that assume the structured data is present 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 can be orders of magnitude cheaper than expensive ML models, but can be inaccurate, which is not acceptable in many applications. To rectify this, I have built systems to accelerate general classes of queries: selection, aggregation, and limit queries *with statistical guarantees on query results*. Specifically, these guarantees are with respect to target methods. While target methods are usually expensive deep neural networks (DNNs), they can also be human annotators for mission-critical or scientific use cases. I further show how to efficiently generate these approximations.

Selection queries and 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 proxy models for selection in NoScope [9]. 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 are cascaded with the target DNN to answer queries. NoScope can improve approximate selection by orders of magnitude, compared to exhaustive labeling.

Since published in VLDB 2017, NoScope has inspired new work [1, 3, 17] and has been cited over 180 times. Furthermore, I have shown that proxy models can accelerate general classes of traditional ML workloads [16], e.g., data-transformation bound workloads.

Proxy model with guarantees. While proxy models can be used to accelerate queries, they do not achieve guarantees on query results. These guarantees are critical for robust conclusions, in particular for scientific inferences. For example, our Stanford biologist collaborators wish to find footage of hummingbirds in wildlife video to analyze interactions between bacteria and feeding patterns. They require guarantees on the fraction of hummingbirds discovered (i.e., recall) for scientifically valid inferences. I have focused on accelerating selection, aggregation, and limit queries with guarantees to address these needs.

While NoScope can accelerate selection queries, it does not provide *statistical guarantees on the recall of query results*. To address this, I developed SUPG, a system with query semantics and sampling algorithms for approximate selection queries with guarantees [10]. It selectively samples the target DNN to form confidence intervals over the samples, which provides guarantees on recall. Prior work uses uniform sampling, which results in poor quality (i.e., returned sets with low precision). To improve sampling efficiency, SUPG instead uses a novel set of weights for importance sampling. Standard importance sampling uses weights proportional to proxy scores, but for selection queries, the ground truth is binary. As such, we use square root weights, which "downweighs" the proxy confidence, which we show is optimal. SUPG can improve query quality by up to 30× at a fixed budget over uniform sampling.

In addition to selection queries, analysts are interested in accelerating other queries with guarantees on results. To provide these guarantees, I developed BlazeIt to optimize aggregation (computing a statistic over the data records) and limit (finding a limited number of records matching predicates) queries [8]. BlazeIt trains proxy models to approximate arbitrary statistics via annotations from the target DNN. It uses these proxy models with algorithms to reduce sampling variance for aggregation queries and rank rare events for limit queries. BlazeIt 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, which can generate proxy scores across many queries and query types efficiently via an embedding index [11]. 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. These embeddings can be reused across query types (including every query type I described above) since they are pre-computed and designed to work for any query over the target DNN output. 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 [13]. Recent research, e.g., new accelerators, has greatly improved DNN throughputs by up to 150×. While this work has improved the throughput of *DNN execution*, it ignores other costs. In the first measurement study of its kind, I showed that the *preprocessing* of visual data (e.g., image decoding) can bottleneck end-to-end DNN inference for visual analytics systems by up to 23× [13]. 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 optimization to balance preprocessing and DNN execution. Smol can improve throughput by up to 5.9× at a fixed accuracy over recent work in visual analytics.

2 Systems for ML Quality Assurance and Robustness

Robustness of ML is a key barrier to widespread adoption. In mission-critical systems, errors in models can have cascading effects, e.g., causing safety violations in AVs. In query processing, query guarantees are with respect to target DNNs. Thus, errors in these target DNNs will be reflected in query results. I have developed methods to monitor ML methods, improve training data quality, and measure robustness. My work on model assertions is being deployed at the Toyota Research Institute, an AV company.

Model assertions and retraining. As ML methods continue to improve on benchmark tasks, they are increasingly being deployed in mission-critical settings, such as AVs. However, average-case measures of performance can hide potentially critical errors. 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 via model assertions, which is a method for allowing users to specify when ML methods may be causing errors [14]. For example, consider a state-of-the-art object detection DNN deployed over video to detect cars. Even state-of-the-art models can fail simple assertions, such as temporal consistency, e.g., that a car should not appear and disappear rapidly in a video. Model assertions can find these errors and others with high true positive rate, at least 88% in all cases we studied.

Furthermore, I showed that model 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., AVs 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): by retraining on this data, model assertions can reduce labeling costs by up to 40% at a fixed budget.

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, particularly how training data and procedures affect downstream tasks. First, I developed a method for training natural language generation models to be robust against noise in training data [12]. Second, I developed a method of measuring the robustness of models against adversaries not seen at training time, as is more reflective of reality [15].

3 Benchmarking ML Pipelines

It has become increasingly difficult to measure the performance of ML systems as they have improved and expanded in diversity. Existing work measures proxy metrics, such as the time to process a minibatch of data. Unfortunately, these metrics are not indicative of producing a high-quality result, e.g., high accuracy.

I was part of the founding team for DAWNBench [5, 6] and MLPerf [18], which have set standards for comparing DNN systems and are 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, showing that proxy metrics are insufficient for measuring DNN systems.

4 Future Research

I am excited to continue my research in data management for and using ML. The deployment of these systems raises a number of research challenges, including:

Training models for analytics. As ML is increasingly deployed in bespoke environments, off-the-shelf methods will be unsuitable for many applications. For example, my collaborators at Stanford require a hummingbird detector: a state-of-the-art model trained on MS-COCO fails on our data (due to domain shift). Thus, analytics systems will need to train ML models to answer queries.

Training models for analytics raises several research questions. Systems must balance between iteratively collecting training data and inference of models while achieving guarantees on query results. Standard approaches (e.g., active learning) simply focus on training a model with high accuracy, which I have shown is unnecessary to answer queries. Instead, we can stop once the desired accuracy is achieved by balancing

between improving a model and iterative query processing. I will combine my expertise in managing training data and ML-based query processing with guarantees to build systems that navigate these tradeoffs.

Low-latency analytics for decision making. As unstructured data volumes grows, ML-based analytics will increasingly be used for decision making. For example, I am collaborating with Jasper Ridge Nature Preserve to find wildfires with remote pan-tilt-zoom cameras. We require *low-latency* responses over these video streams as corrective action (i.e., firefighting) must occur rapidly.

Low-latency ML-based analytics has several unique challenges. Many organizations have limited computational and human-attention resources to monitor fleets of sensors. In addition, these applications face constantly changing conditions (e.g., weather changes, seasonal changes). From a resource perspective, analytics systems must allocate computational and human-attention resources to different sensors as necessary: only a tiny fraction of the data needs corrective action. From a statistical perspective, analytics systems must decide when to retrain both cheap and expensive ML models in the face of changing conditions. I will develop systems that allocate resources across sensors in these low-latency settings by leveraging my expertise in navigating resource tradeoffs via cheap approximations while maintaining guarantees.

More complex query types. While I have shown that broad classes of ML-based queries can be accelerated, there are many other query types. For example, many applications would benefit from accelerated systems for joins, group bys, and nested queries. Furthermore, complex queries beyond simple aggregates or selection will require query optimization for efficient execution. I plan to explore these queries by developing new algorithms and query processing techniques.

New advances in ML capabilities will enable impactful applications, but cost and reliability will be of paramount concern. My research has shown the promise of systems thinking for deploying ML systems by reducing query costs by orders of magnitude and approaches for queries with guarantees. I plan to continue my principled approach towards solving problems to enable these new, impactful applications.

References

[1] Favyen Bastani, Oscar Moll, and Sam Madden. Vaas: video analytics at scale. *Proceedings of the VLDB Endowment*, 13(12):2877–2880, 2020.

[2] Carlo Migel Bautista, Clifford Austin Dy, Miguel Iñigo Mañalac, Raphael Angelo Orbe, and Macario Cordel. Convolutional neural network for vehicle detection in low resolution traffic videos. In *2016 IEEE Region 10 Symposium (TENSYMP)*, pages 277–281. IEEE, 2016.

[3] Christopher Canel, Thomas Kim, Giulio Zhou, Conglong Li, Hyeontaek Lim, David Andersen, Michael Kaminsky, and Subramanya 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] 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.

[6] Cody Coleman, Deepak Narayanan, Daniel Kang, Tian Zhao, Jian Zhang, Luigi Nardi, Peter Bailis, Kunle Olukotun, Chris Ré, and Matei Zaharia. Dawnbench: An end-to-end deep learning benchmark and competition. *NeurIPS MLSys Workshop*, 2017.

[7] Kinshuk Govil, Morgan L Welch, J Timothy Ball, and Carlton R Pennypacker. Preliminary results from a wildfire detection system using deep learning on remote camera images. *Remote Sensing*, 12(1):166, 2020.

[8] Daniel Kang, Peter Bailis, and Matei Zaharia. Blazeit: Optimizing declarative aggregation and limit queries for neural network-based video analytics. *PVLDB*, 2019.

[9] Daniel Kang, John Emmons, Firas Abuzaid, Peter Bailis, and Matei Zaharia. Noscope: optimizing neural network queries over video at scale. *PVLDB*, 2017.

[10] Daniel Kang, Edward Gan, Peter Bailis, Tatsunori Hashimoto, and Matei Zaharia. Approximate selection with guarantees using proxies. *PVLDB*, 2020.

[11] Daniel Kang, John Guibas, Peter Bailis, Tatsunori Hashimoto, and Matei Zaharia. Task-agnostic indexes for deep learning-based queries over. *MLSys (under review)*, 2021.

[12] Daniel Kang and Tatsunori Hashimoto. Improved natural language generation via loss truncation. *ACL*, 2020.

[13] Daniel Kang, Ankit Mathur, Teja Veeramacheneni, Peter Bailis, and Matei Zaharia. Jointly optimizing preprocessing and inference for dnn-based visual analytics. *PVLDB*, 2021.

[14] Daniel Kang, Deepti Raghavan, Peter Bailis, and Matei Zaharia. Model assertions for monitoring and improving ml model. *MLSys*, 2020.

[15] Daniel Kang, Yi Sun, Dan Hendrycks, Tom Brown, and Jacob Steinhardt. Testing robustness against unforeseen adversaries. *ICLR (under review)*, 2021.

[16] Peter Kraft, Daniel Kang, Deepak Narayanan, Shoumik Palkar, Peter Bailis, and Matei Zaharia. Willump: A statistically-aware end-to-end optimizer for machine learning inference. *MLSys*, 2019.

[17] Yao Lu, Aakanksha Chowdhery, Srikanth Kandula, and Surajit Chaudhuri. Accelerating machine learning inference with probabilistic predicates. In *SIGMOD*, pages 1493–1508. ACM, 2018.

[18] Peter Mattson, Christine Cheng, Cody Coleman, Greg Diamos, Paulius Micikevicius, David Patterson, Hanlin Tang, Gu-Yeon Wei, Peter Bailis, Victor Bittorf, et al. Mlperf training benchmark. *MLSys*, 2020.