Accelerating Queries over Unstructured Data with ML

Daniel Kang

ACM Reference Format:
Daniel Kang. 2020. Accelerating Queries over Unstructured Data with ML. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 ABSTRACT

Queries over unstructured data (e.g., videos and text) are becoming increasingly common due to large data volumes and machine learning (ML). A common method of answering queries is to use a deep neural network (DNN) or human labeler (which we collectively refer to as oracle methods) to extract structured information from this unstructured data. For example, an ecologist may be interested in understanding hummingbird behavior, so extracts all bird positions from a video.

Unfortunately, these methods can be costly: labeling 100 days of video via human annotators can cost hundreds of thousands of dollars. Thus, to reduce the cost of executing queries, recent work has proposed using proxy models: cheap approximations to oracle methods. They have primarily been studied in the context of approximating binary predicates, in which the proxy model produces a score between 0 and 1 and records above some ad-hoc score threshold are assumed to satisfy the predicate [1, 3, 6].

However, this prior work on binary predicates leaves major concerns unaddressed: 1) existing query processing algorithms do not provide statistical guarantees on query results and 2) they cannot share work between queries efficiently.

To address these issues, we have been developing indexing and query processing algorithms for unstructured data using ML in a system MEME. We describe our recent developments and some applications below.

Applications. We describe two concrete applications to motivate our query processing and indexing algorithms.

First, we are collaborating with biologists at Stanford University to do ecological analysis. They have collected around 200 camera-days of field videos and are interested in analyzing hummingbird feeding patterns. To do this analysis, they are interested in finding at least 80% of the frames containing hummingbirds. Prior work on accelerating this query misses a key requirement that arises in scientific analysis: the need for *statistical guarantees* on query results (i.e., finding 80% of the hummingbird frames with the query succeeding 95% of the time). We are actively deploying our algorithms (described below) to this application.

Second, consider a hypothetical urban planner analyzing traffic from street camera footage. To understand bulk patterns, the urban planner issues an aggregation query, counting the average number of cars per frame. To understand congestion, the urban planner issues a limit query searching for frames of at least 5 cars and one bus. While our prior work can accelerate such queries [2], it cannot easily share work between queries, requiring new proxy models for each query.

Query processing with proxy scores with guarantees. As illustrated by the applications above, two key requirements are statistical guarantees and extended query types.

To address the challenge of statistical guarantees, we propose new algorithms that sample from the oracle methods via the proxy scores and use these samples to achieve statistical guarantees on query accuracy [2, 4]. For selection queries, we show that ad-hoc methods fail to achieve statistical guarantees. We fix this by using confidence intervals over the recall/precision of the returned set. Additionally, we propose importance sampling algorithms that use the proxy score to more efficiently sample rare events [4].

To address the challenge of sharing work between queries, we have designed indexing algorithms for unstructured data [5]. We propose a general method of clustering unstructured data records so that similar records from the perspective of *all* downstream queries are close. Given this clustering, we show how to generate proxy scores using cached samples from the oracle method. These proxy scores can then be used in a variety of downstream query processing algorithms, including for answering approximate selection [3, 4], approximate aggregation [2], and limit queries [2].

While our query processing and indexing algorithms show promise for accelerating queries over unstructured data records, we believe there is much work to be done. For example, MEME is primarily designed for batch queries, as opposed to streaming data. Real time answer to queries are critical for other applications; we are actively exploring ways to extend MEME for live wildfire detection. We hope that MEME can serve as a starting point for accelerating queries across a wider range of settings and applications.

REFERENCES

[1] Michael R Anderson, Michael Cafarella, Thomas F Wenisch, and German Ros. 2019. Predicate Optimization for a Visual Analytics Database. *ICDE* (2019).
[2] Daniel Kang, Peter Bailis, and Matei Zaharia. 2019. BlazeIt: Optimizing Declarative Aggregation and Limit Queries for Neural Network-Based Video Analytics. *PVLDB* (2019).
[3] Daniel Kang, John Emmons, Firas Abuzaid, Peter Bailis, and Matei Zaharia. 2017. NoScope: optimizing neural network queries over video at scale. *PVLDB* 10, 11 (2017), 1586–1597.
[4] Daniel Kang, Edward Gan, Peter Bailis, Tatsunori Hashimoto, and Matei Zaharia. 2020. Approximate Selection with Guarantees using Proxies. *PVLDB* (2020).
[5] Daniel Kang, John Guibas, Peter Bailis, Tatsunori Hashimoto, and Matei Zaharia. 2021. Task-agnostic Indexes for Deep Learning-based Queries over. *PVLDB (under review)* (2021).
[6] Yao Lu, Aakanksha Chowdhery, Srikanth Kandula, and Surajit Chaudhuri. 2018. Accelerating Machine Learning Inference with Probabilistic Predicates. In *SIGMOD*. ACM, 1493–1508.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn