Technical Perspective: Query Optimization for Faster Deep CNN Explanations

Sebastian Schelter
University of Amsterdam
s.schelter@uva.nl

Machine learning (ML) is increasingly used to automate decision making in various domains. In recent years, ML has not only been applied to tasks that use structured input data, but also, tasks that operate on data with less strictly defined structure such as speech, images and videos. Prominent examples are speech recognition for personal assistants or face recognition for boarding airplanes.

Responsible data management. There exists a variety of challenges with respect to the fairness, accountability and transparency of the resulting automating decision-making systems, whose data specific aspects are addressed by research under the umbrella of "responsible data management" [2]. A particular challenge in this area is the explainability of the predictions of an ML model. Common approaches derive local explanations for a model's predictions by perturbing the features of a single example and calculating how much these perturbations affect the prediction outcome as a measure of the explanatory power of the perturbed feature [7].

Data management for machine learning. ML poses additional challenges apart from responsibility. ML models are part of larger end-to-end ML pipelines, which include the integration, validation, and cleaning of data, as well as the training, deployment and analysis of models. The definition, maintenance and efficient execution of such pipelines pose various data management challenges [4, 6, 8].

Unfortunately, it is often very difficult to apply established data management techniques, such as query optimization or provenance tracking, as they rely on an abstract algebraic specification of the computation, which is typically lacking for end-to-end ML pipelines. These pipelines comprise of the previously mentioned heterogeneous stages, for which different systems are often "glued" together in the real world. This results in tedious work, complex environments, and a loss of potential for optimisation. It is an ongoing research challenge to find well working abstractions for ML computations that incorporate data and operations from both relational algebra and linear algebra [5].

Convolutional neural networks. Deep neural networks are the current state-of-the-art in machine learning, and heavily influence adjacent domains such as computer vision and natural language processing. Convolutional neural networks (CNNs), designed to learn features from image data, started the triumph of deep neural networks with their outstanding performance in image recognition tasks [3], and earned their inventor Yann LeCun a Turing award. Efficient training and inference for deep neural networks is gaining a lot of attention recently, e.g., in the form of specialized optimizing compilers [1].

The highlighted paper. The work by Nakandala et al. concentrates on the problem of efficiently computing occlusion-based explanations of the predictions of a CNN. This explanation method repeatedly occludes small regions of an input image, and measures the corresponding changes in the predictions. This approach requires a very large number of inference requests to the CNN, and the paper presents efficient methods to drastically reduce the runtime of the corresponding computation.

The beauty of this research lies in the fact that it elegantly connects all three previously discussed areas: responsible data management (explaining the predictions of an ML model), deep learning (with its focus on convolutional neural networks) and data management (query optimisation for ML inference).

The paper is based on the observation that deep neural networks open up many opportunities for applying established optimisations from relational query processing, as they also build on a strict algebraic foundation: their underlying computations are modeled as directed acyclic graphs of linear algebra operators, which exchange data in the form of tensors. The authors take established techniques from the data management space (incremental view maintenance and multi-query optimisation), and reinvent them for the ML related context of efficiently executing a large number of related inference requests. For that, the paper treats inference requests as "queries", the CNN as a "query plan" with operators like max-pooling, and tensors (which represent the input images and operator outputs) as "relations".

This work is an important step forward towards bridging the gap between classical relational data management and modern machine learning workloads. I hope that we will see generalizations of the applied techniques for a wide variety of related ML tasks in the future.

REFERENCES

[1] Tianqi Chen, Thierry Moreau, Ziheng Jiang, Lianmin Zheng, Eddie Yan, Haichen Shen, Meghan Cowan, Leyuan Wang, Yuwei Hu, Luis Ceze, et al. 2018. TVM: An Automated End-to-End Optimizing Compiler for Deep Learning. *OSDI*, 578–594.

[2] HV Jagadish, Francesco Bonchi, Tina Eliassi-Rad, Lise Getoor, Krishna Gummadi, and Julia Stoyanovich. 2019. The Responsibility Challenge for Data. *SIGMOD*, 412–414.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet Classification with Deep Convolutional Neural Networks. *NeurIPS*, 1097–1105.

[4] Arun Kumar, Robert McCann, Jeffrey Naughton, and Jignesh M Patel. 2016. Model Selection Management Systems: The Next Frontier of Advanced Analytics. *SIGMOD Record* 44, 4 (2016), 17–22.

[5] Andreas Kunft, Alexander Alexandrov, Asterios Katsifodimos, and Volker Markl. 2016. Bridging the gap: towards optimization across linear and relational algebra. *Workshop on Algorithms and Systems for MapReduce and Beyond at SIGMOD*, 1–4.

[6] Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. 2018. Data Lifecycle Challenges in Production Machine Learning: A Survey. *SIGMOD Record* 47, 2 (2018), 17–28.

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the Predictions of Any Classifier. *KDD*, 1135–1144.

[8] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden Technical Debt in Machine Learning Systems. *NeurIPS*, 2503–2511.