Neural Information Processing Systems (NeurIPS) 2020

Test of Time Award Winner

Hogwild: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent
Authors: Benjamin Recht, Christopher Re, Stephen Wright, Feng Niu
Institution: University of Wisconsin-Madison (NeurIPS 2011)

Abstract:
Machine learning is the problem of turning exemplar data into a model, stored in a computer, that can be used to make decisions or take actions. At the core of modern machine-learning systems is the stochastic gradient method - usually known as “SGD” - which searches the space of possible models to find one that matches up well with the exemplar data. This paper described an implementation of SGD that can be run in parallel across a collection of fast computers, all of them making repeated small changes to the model without any coordination or synchronization. This approach, which the authors dubbed Hogwild!, outperformed alternative parallelization schemes that required synchronization. The paper also presented a theoretical analysis of Hogwild!’s convergence rate, showing that linear speedup in the number of processors could be attained (to within a constant factor) even when a large number of processors were used. The paper has been cited almost 2000 times, attesting to its influence not only on machine learning but also on the fields of computer systems and optimization, both of which contributed to the development and understanding of the Hogwild! approach.

Three Best Paper Award Winners

Language Models are Few-Shot Learners
Authors: Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei
Institution: OpenAI

Abstract:
Artificial intelligence systems trained to estimate the likelihood of the next word in a sequence are known as “language models”. Language models were first described in the 1950s as a theoretical construct for connecting the then-new field of information theory with natural language. This paper describes GPT-3, the largest and most sophisticated language model ever constructed. It demonstrates that, if you make a language model accurate enough by using unprecedented amounts of compute and data, it gains the ability to solve a wide variety of tasks

without additional training, using only simple, natural language prompts. Example tasks include answering trivia questions, generating essays, determining if a movie review is positive or negative, and translating between French and English. The authors note that GPT-3 is better at some tasks than others, and devote most of the paper to carefully cataloging its strengths and weaknesses. The authors also consider potentially harmful implications of the technology, such as cheap generation of almost undetectable fake news and the model's tendency to reflect the biases of its training data on sensitive topics such as race, gender, and religion.

No-Regret Learning Dynamics for Extensive-Form Correlated Equilibrium
Authors: Andrea Celli (Polimi), Alberto Marchesi (Polimi), Gabriele Farina (CM) and Nicola Gatti (Polimi)
Institutions: Politecnico di Milano and Carnegie Mellon University

Abstract:
Our decisions impact others and their decisions impact us. To settle on a rational way to behave, we need to cut through this interdependence to reach what economists call an equilibrium. Creating automated procedures for finding equilibria is notoriously difficult. This paper provides the first approach for finding so-called correlated equilibria for general interactions using a learning approach. Correlated equilibria require a trusted external mediator that makes decision recommendations to the decision-makers. The canonical example of a correlated equilibrium is a stoplight. The stoplight tells approaching cars whether it is safe to go. Even in the absence of relevant laws, we should follow the stoplight's recommendations because we know that everyone can reason that it is in their best interest to do so---driving through the red light is a risky proposition. The paper shows that such equilibria can be arrived at by learning algorithms acting completely independently---no external traffic engineer is needed---even when the decisions involve multiple steps and the decision-makers are partly in the dark about the state of the world. Such an approach could have powerful implications in the modern "gig economy", where centralized supervision of self-interested actors is the norm.

Improved guarantees and a multiple-descent curve for Column Subset Selection and the Nystrom method
Authors: Michał Dereziński, Rajiv Khanna, Michael W. Mahoney
Institution: University of California, Berkeley

Abstract:
As the availability of large datasets expands, so does society’s dependence on being able to summarize complex data succinctly. Data summarization is the problem of identifying important examples and attributes in data to help characterize it efficiently. It can be used to select a representative subset of gene variants from a genetics dataset or the most informative documents from a text database. Prior work has shown that data summarization is an intractable problem---there are data sets for which no known algorithm can provide a good summary in a reasonable time frame. This paper shows that these analyses are far too

pessimistic. The datasets that make the data summarization problem intractable are pathological and, in fact, interpretable summaries can be generated far more cheaply for real-world data. The work suggests that future systems will be able to create data summaries that are accurate, interpretable, and efficiently generated, greatly aiding our ability to absorb and process complex datasets.