Google Research Identifies a Bottleneck in Hyperscale Approaches to AI

Originally published at https://www.unite.ai/google-research-identifies-a-bottleneck-in-hyperscale-approaches-to-ai/

A new paper from Google Research indicates that the current trend towards the curation of very high-volume datasets may be counterproductive to developing effective artificial intelligence systems. In fact, the research indicates that better machine learning products may emerge from being trained on *less* accurate (i.e. technically 'worse') datasets.

If the principles obtained by the researchers are valid, it means that 'hyperscale' datasets such as the recently-released LAION-400M (which contains 400 million text/image pairs), and the data behind the GPT-3 neural language engine (containing 175 billion parameters), are potentially subject to a kind of 'thermal limit' in traditional and popular machine learning architectures and methodologies, whereby the sheer volume of data 'saturates' downstream applications and prevents them generalizing in a useful way.

The researchers also propose alternate methods to rethink hyperscale dataset architecture, in order to redress the imbalance.

The paper states:

> *'Delving deeper to understand the reasons that give rise to these phenomena, we show that the saturation behavior we observe is closely related to the way that representations evolve through the layers of the models. We showcase an even more extreme scenario where performance on upstream and downstream are at odds with each other. That is, to have a better downstream performance, we need to hurt upstream accuracy.'*

The study is titled *Exploring the Limits of Large Scale Pre-training*, and comes from four authors at Google Research.

Investigating 'Saturation'

The authors challenge the prevailing assumptions of machine learning>data relationships in the hyperscale data age: that scaling models and data size notably improves performance (a belief that has been cemented in the hype over GPT-3 since its launch); and that this improved performance 'passes through' to downstream tasks in a linear (i.e. desirable) way, so that the on-device algorithms that are eventually launched to market, derived from the otherwise ungovernably huge datasets and undistilled trained models, benefit completely from the insights of the full-sized, upstream architectures.

'These views,' the researchers note *'suggest that spending compute and research effort on improving the performance on one massive corpus would pay off because that would enable us to solve many downstream tasks almost for free.'*

But the paper contends that a lack of computing resources and the subsequent 'economical' methods of model evaluation are contributing to a false impression of the relationship dynamics between data volume and useful AI systems. The authors identify this habit as 'a major shortcoming', since the research community typically assumes that local (positive) results will translate into useful later implementations:

> *'[Due] to compute limitations, performance for different choices of hyper-parameter values is not reported. Scaling plots seem more favorable if the hyper-parameter chosen for each scale is fixed or determined by a simple scaling function.'*

The researchers further state that many scaling studies are measured not against absolute scales, but as incremental improvements against the state-of-the-art (SotA), observing that 'there is no reason, a priori, for the scaling to hold outside of the studied range'.

Pre-Training

The paper addresses the practice of 'pre-training', a measure designed to save compute resources and cut down on the often horrendous timescales needed to train a model on large-scale data from zero. Pre-training snapshots handle the 'ABCs' of the way that data within one domain will become generalized during training, and are commonly used in a variety of machine learning sectors and specialties, from Natural Language Processing (NLP) through to deepfakes.

Previous academic research has found that pre-training can notably improve model robustness and accuracy, but the new paper suggests that the complexity of features, even in relatively short-trained pre-training templates, might be of more benefit if shunted down the line to later processes in the pipeline.

However, this can't happen if researchers continue to depend on pre-trained models that use current best practice in application of learning rates, which, the research concludes, can notably affect the ultimate accuracy of the final applications of the work. In this respect, the authors note that 'one cannot hope to find one pre-trained checkpoint that performs well on all possible downstream tasks'.

The Study

To establish the saturation effect, the authors conducted 4800 experiments on Vision Transformers, ResNets and MLP-Mixers, each with a varying number of parameters, from 10 million to 10 billion, all trained on the highest-volume datasets available in the respective sectors, including ImageNet21K and Google's own JFT-300M.

The results, the paper claims, show that *data diversity* should be considered as an additional axis when attempting to 'scale up' data, model parameters and compute time. As it stands, the heavy concentration of training resources (and researcher attention) on the upstream section of an AI pipeline is effectively blasting downstream applications with an avalanche of parameters up to a point of 'saturation', lowering the capability of deployed algorithms to navigate through features and perform inference or effect transformations.

The paper concludes:

> *'Through an extensive study, we establish that as we improve the performance of the upstream task either by scaling up or hyper-parameter and architectural choices, the performance of downstream tasks shows a saturating behaviour. In addition, we provide strong empirical evidence that, contrary to the common narrative, scaling does not lead to a one-model-fits-all solution.'*

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More