AICONIC

Big and Tiny Machine Learning in 2020: The Promise of Mainstream AI

In many ways, 2020 was the year in which the promise of AI became mainstream. While progress in deep machine learning has been steadily accelerating this decade, the type of AI dreamed up in Hollywood did not quite materialize. While many fascinating results were published and presented in research circles, they mostly failed in making the ultimate step toward consumer product improvement. In this discussion, we will focus on two seemingly opposite research directions that sought (and succeeded) in changing this: very big and tiny machine learning.

2020 Summary:

- Significant victories in machine learning research that pushes model size to its limits are likely to result in *ML-as-a-Service APIs* facilitating a new generation of ML products and companies
- Progress in reducing data dependency and computation costs, i.e. 'tiny' AI, continues to *lower the cost barrier-to-entry* for startups and will lead to a wave of new AI companies
- The new ML-as-a-Service industry will be *dominated by the few selected companies* that successfully cornered the market of AI talent with the resources to make deep long-term bets on multiple moon-shot explorations
- Significant AI results, combined with lower costs of failure and increased competition, will lead to faster iteration and *force corporations and governments to invest more broadly in venture capital*

Big ML Possibly the two most important drivers in the (deep) machine learning renaissance of the last decade are access to (very) cheap computation power and (very) large datasets. Ostensibly, the bigger the model architecture and the larger the dataset, the better the model performance. Hence, various researchers try to follow these insights to their logical conclusion: how big can a model be? The first target was computer vision, where in rapid succession, ever increasing deep machine learning models rendered the biggest image classification competition obsolete. While impressive, the results largely depended on large amounts of *human-labeled* image data. It was therefore not quite clear how such results could be transferred over to, e.g. the domain of natural language understanding (NLP), where labeling targets are less clearly defined.

Two breakthroughs set off a series of papers by the Elon Musk + Silicon Valley all-star investor ensemble funded OpenAI that would ultimately result in the much publicized GPT-3 work. The first was the Transformer model, a new type of machine learning architecture that allowed for much bigger NLP based modeling. The second was the idea of using *generative pre-training* (hence the name 'GPT'), which seeks to exploit natural structure in language data, by reinterpreting it as the result of a generative model - in doing so, they cleverly side-stepped the need for large amounts of labeled data. In 2018, they presented a proof-of-concept combining the two, in which they showed that large semi-supervised language models could be used for various classic downstream NLP tasks (GPT-1). In the follow up works, GPT-2 and GPT-3, they introduced various engineering improvements, extended the training data to include most of the internet and more, and increased their model size to a staggering 175 billion parameters. While the costs of training such a model from scratch, even ignoring all the research and experimentation time to conceive it, runs in the millions of dollars, the results and demo-api released to the public are astounding.

The second giant-model breakthrough of 2020 comes from the domain of molecular biology, for the task of protein-structure prediction also known as 'the protein folding problem'. To better understand why this specific problem is deemed so important see the following quote from Google's Deepmind:

> *"Scientists have long been interested in determining the structures of proteins because a protein's form is thought to dictate its function. Once a protein's shape is understood, its role within the cell can be guessed at, and scientists can develop drugs that work with the protein's unique shape."*

Especially in the current pandemic climate, it is not hard to see why solutions would be welcomed. Similar to OpenAI, Deepmind had their proof-of-concept moment in 2018, when their deep learning based entry, AlphaFold 1, won the main competition for this problem (CASP). The model combined problem domain knowledge and machine learning techniques to predict protein structure from scratch. However, it wasn't until their recent November 2020 presentation of AlphaFold 2 in which they appear to fully 'solve' the protein folding problem, that they secured widespread recognition of the academic community - some going as far as calling for the first machine learning based Nobel prize. While much is unknown about the AlphaFold 2 architecture, two main ingredients are: Transformers and Data Symmetry exploitation. The first we encountered already, the second requires some clarification. Research in data symmetries attempts to incorporate prior knowledge of natural data symmetries, to make neural networks more robust and effectively restrict solution spaces, e.g. rotating a picture of a cat should not change it's classification, nor should it require 360 cat pictures to learn this. While less audacious as GPT-3, AlphaFold 2 still takes weeks to train on hundreds of GPUs.

Tiny ML With these two massive giant-model successes, one wonders: is bigger really always better? Not necessarily - how should we fit all these giant models on the tiniest of machines, say a mobile phone? Enter tiny machine learning research: how to use less data and smaller model architectures to still obtain great results. We will restrict our discussion to one example of each: The first we already encountered in AlphaFold 2; using natural data structure. The past 5 years we have seen an explosion into ML research focused on using concepts such as 'equivariance', or more generally the concept of data as having some distinct structural properties, that in some cases can be known in advance. When this is the case, it reduces the need for large amounts of data, by instead using small amounts of data more efficiently. A second direction is that of model compression, in the forms of quantization and distillation The ideas being that we can reduce a model's parameter precision, i.e. the number of bits needed to represent the values of weights, or reduce the number of parameters all together by taking a large, pre-trained model, and investigating which of the parameters truly contribute to the end result, all while maintaining a certain performance accuracy.

Productionized ML How is all of this related to fulfilling the promise of mainstream, transformative, productionized AI? While the giant machine learning models of OpenAI and DeepMind are entirely unfeasible to create for most companies or mortals, they represent a distinct shift in what the next generation of companies might look like. Just as how AWS alleviated most of the tech infrastructure burden for tech startups through cloud services, how Shopify provided the go-to online sales channel for small retailers, and how Spotify gave a voice to millions of voice artists, ML-as-a-Service accessible through APIs can offer entrepreneurs crucial building blocks to build new products previously considered unfeasible to build from scratch. Look no further than the ease of accessing a 'Translation' ML-as-a-Service API, used by specialized high-tech startups as well as by any every-day users through a Google Sheets plugin, i.e. =GoogleTranslate(sentence, source-language, target-language).

Similarly, as research in tiny ML progresses, two outcomes materialize: more machine learning applications *on the edge*, e.g. mobile phones, small devices, and an ever lowering cost barrier to entry for small startups to conduct competitive machine learning R&D. Whereas most breakthroughs are currently restricted to large academic or industry labs due to the high compute and/or data requirements, a future in which both no longer form a necessary condition should speed up innovative product development.

Investing in ML One obvious question is how the developments outlined above might influence the market. We summarize a selection of predictions on Cloud Providers, Large AI companies, and Venture Capital below:

Cloud Providers are already becoming increasingly more powerful as existing (non-ML) companies transition their workflows to the cloud. The productionization of ML will add to this in various ways:

- Non-ML companies will begin integrating ML components into their workflow, increasing the demand for scalable computation resources.
- More startups will develop in the ML-sphere, producing more ML products, which would further increase the demand for cloud services.

AI Giants like Google, Microsoft, Facebook, Amazon & co, have been the big winners of 2020 so far. Their market robustness will only increase having successfully cornered the market of AI talent.

- ML-as-a-Service will become a new non-trivial source of revenue that is likely to increase the market power of the select group with the ability to offer crucial building blocks at scale.
- ML related acquisitions will strongly increase in small companies that prove themselves most successful in either (a) building profitable products on top of ML building blocks, or (b) showing ML-as-a-Service proof-of-concepts based on particular domain knowledge too time-consuming to build from scratch internally

Venture Capital has long been crucial to promote innovation. However, it is often seen (with reason) as a risky form of investing, only reserved for those with a high tolerance for risk. Yet when it comes to AI, in recent years this has started to slowly change: Large companies with natural access to vast amounts of data realized their inability to successfully capitalize on this resource and began spinning off companies focused on narrow use-cases. Governments, concerned with missing out on the next generation of jobs, have begun to establish 'innovation investment funds' in an attempt to keep or attract new companies within their national borders.

- With a new generation of AI-enabled companies about to surface, investing early and broadly will increasingly become the norm for corporations and governments alike.
- Cheaper resources, less dependency on data, and higher availability of specialized ML building blocks through APIs will create a wave of new companies. Even though many may fail, as the cost of iteration reduces, and the competitive landscape intensifies, investors will likely see quicker outcomes.

Conclusion 2020 will likely go down as one of the most tragic years in modern history for most people; death, sickness, and large economic damage will make a long lasting impact for all. However, in the unique world of machine learning, it was a year of meaningful breakthroughs. While uncertainty looms over 2021 still, we see enough evidence to support a bullish view on everything Cloud + ML. The expanding ML-as-a-Service industry and ever lowering costs of research are likely to empower a new generation of ML products and companies. Lower failure costs combined with increased competition will accelerate the startup iteration cycle, providing investors with faster outcomes. Finally, corporations and governments will drastically increase their venture investment targets in an attempt to stay relevant in a rapidly digitalizing and automating economic environment.

Wishing you a very Happy New Year,

Tim Davidson
CEO, Aiconic B.V.