ML in 2022: Trends and Projections

November 27, 2022

Abstract

Having been working in the ML industry with a primary focus on tabular embedding for one year, I did observe my mindset changing constantly from academia and hence would like to share some perspectives from my personal stream. I will first synthesize the mainstreams of ML in 2022 based on the papers read and practices made, and then forecast some unnoticed area that might be trending topics in the coming year.

Trends & Observations in 2022

One dominates all

Transformer and its underlying self-attention module stemmed from text and gradually conquered vision, speech, graph and even tabular domain (with tabular embedding based on my exploration) within five years. It reminds me of ResNet in 2016, where the residual connection has been widely utilized by applications beyond vision. The two observations above also match our experience in real life that FAANG dominates the rest in QQQ and SPY. In literature you can expect different variations of transformer (batch, long-former to name a few), but you can hardly come across any ultraman or any other super hero.

ML of the art

As an omnipotent model requires less effort on model development, more and more novelties are accredited to improving model performance under different scenarios in real world, making ML scientists work like engineers. For example, there are various works on improving generalizability of large pre-trained models under {x}-shot learning scenarios, and on assembling different modules to enable multi-modal learning given input from heterogeneous sources. The majority of these efforts are spent on designing heuristic operations that can ameliorate the gaps between upstream and downstream tasks. Sometimes it feels like the improvements come as the fruit of heavy engineering efforts, which looks like art unless you deep dive the details in appendix.

The neuroscientist

Great human ideas come from inspirations, and great ML innovations , to some degree, come from replicating humans. Although explainability is not readily available for every SOTA ML models, researchers tend to affirm the effectiveness of their ideas at the very beginning of the articles, building a connection between their inspirations and the functionality of humans. It is interesting to see some of the works directly refer to articles from neuroscience, making this interdisciplinary knowledge necessary for notable ML progress. For example, in Yann LeCun's vision on autonomous AI, it is clear to see the system is constructed as human which rely on consciousness, subconsciousness and memory to build intelligence. Personally I also believe this is the right way to follow since the upper bound of ourselves is ourselves. However, this approach remains challenging given the system to be replicated is so complex that we might only be able to make partial observations. Still, I am optimistic at least we can get some sketches if not the portraits.

Nothing comes from nowhere

You can expect new wines in old bottles or vice versa, but honestly not new wines in new bottles. For example, diffusion models are built upon Markov chains, stacked VAE and stochastic differential equations. Sometimes you might run into fancy terms but eventually you find out that they might be just the rephrasings of hidden views in the 1980s, 1990s, or 2000s. For example, you might understand Elon Musk's vision better if you happen to have read Nicola Tesla's stories. Another reason to try something canonical is that ML progresses so fast that it is risky, suboptimal and time-consuming to try something brand-new. This phenomenon also explains why we get more papers with small steps rather than fewer papers with leaps.

Projections in 2023

Memory

To equip ML models with memory is analogous to attending open book exams with references. Ideally, this module might not only relieve the burden of large pre-train models but also improve model performance through direct inclusion of the underrepresented samples. It also matches the aforementioned neurological methodology that human can abstract as well as memorize. Intuitively, the memory module should not differ significantly from the model, given the fact that the weights in the model are boundary function guided by the samples. This idea has been heuristically explored in NLP papers with the "retriever" keyword, but can be improved further from many perspectives.

Metrics learning

Neural networks (NN), in whatever manners, comprises of an encoder to transform the raw samples into representations, and a projector to draw boundary that makes the representations separable. The nonlinearity of NN lies in the encoder while the projector are linear most of the time. In practice, much of the attention has been paid to learn a complex encoder, while very few realize the simple projector can be the actual bottleneck that restricts NN performance. Metrics learning, on the other hand, targets towards this limitation through customized loss function which explicitly reformulates the boundary. This technique has been explored and found effective in representation learning works such as contrastive learning, but is far from mature given no omnipotent loss has been found and not all representation spaces (i.e., unit hypersphere) have been fully explored.