Large Language Models: Machine Learning Down to Earth

Paulo H. Leocadio

Zinnia Holdings LLC

Author Note

The article was written exclusively for the Brainz Magazine

Abstract

The business marketplace has been flooded with waves of technology trends that periodically surface and become present on every other sales pitch from technology vendors and build up as the ultimate necessity in the minds of many CIOs.

There is a variety of examples to mention: downsizing, rightsizing, outsourcing, offshoring, consumerization of IT, and the adoption of Cloud technologies, among others. And like virtual tsunamis, these trends come impacting, sometimes disrupting, and even influencing the performance or changing the perception of a corporation (or government) in the eyes of the consumer.

It is not different today: tailgating the Cloud adoption, and the spreading of a Big Data mindset to solve problems and build solutions comes Artificial Intelligence (AI) and Machine Learning (ML). Vendors quickly packaged “AI” offerings, often pre-existing call-center-related scripts and chatbots, the door to the CIOs’ IT team was opened, and the offerings began to mature beyond call centers and the diverse medical and micro-biological use.

University and Academic AI-related research were reinforced by highly specialized and well-funded dedicated groups in companies like Google, Facebook, and IBM. Smaller organizations, collectives, and startups are also receiving funding and investments.

Open AI became an important presence under many aspects and circumstances. Anyone can interact online and in real-time with its chatbot, obtaining real-life value as a work tool for many areas.

When asked “please explain what Large Language Models are?”, the OpenAI chatbot responded (OpenAI, 2023): “Large Language Models (LLMs) are a type of artificial intelligence (AI) model that is designed to understand and generate human-like language. These models are

trained on massive amounts of text data, such as books, articles, and web pages, and use advanced algorithms to learn the patterns and structures of language.

LLMs can perform a wide range of natural language processing (NLP) tasks, such as language translation, sentiment analysis, text classification, and question-answering. They can also generate human-like text, such as short stories, news articles, and even poetry.

Some examples of LLMs include GPT-3 (Generative Pre-trained Transformer 3), BERT (Bidirectional Encoder Representations from Transformers), and T5 (Text-to-Text Transfer Transformer). These models have achieved significant breakthroughs in NLP and are widely used in industry and academia for various applications."(grammar errors found in the AI-generated text were left unchanged on purpose).

Keywords: Artificial Intelligence; AI; Machine Learning; Large Language Models, MLL

Large Language Models: Machine Learning Down to Earth

Large Language Models (MLLs) are potentially the closest Machine Learning models a random person can interact with. It is likely the best AI set of tools that enable the building of applications that would come to our dreamer minds when visualizing science fiction stories in real life down to Earth.

Large Language Models

This article superficially explores current advances in Neural networks for large-scale language modeling. The text introduces current and potential real-life applications, as well as discusses the potential sources of risk (and concerns) related to some of the currently used techniques.

Understanding the Limits of LLMs

In a simplistic description, Language Modeling (LM) is a set of core tasks for sentence building, encoding grammatical structure, and also to digest information about the knowledge a data set (or text) may be composed of, in other words, when very large amounts of data are available to train language models, they can extract a compact version of the knowledge encoded in that training data (Vaswani, Zhao, Fossum, & Chiang). Currently, a considerable amount of effort is dedicated to smaller sets of data to train LMs. The performance of large tasks (on larger sets of data) demonstrates to be very relevant for extracting what matters, many ideas work well in smaller sets but do not improve further when applied in larger sets (Jozefowicz, Vinyals, Schuster, Shazeer, & Wu, 2016).

It is worth noting that, based on how fast hardware is improving currently, combined with unimaginable amounts of text (data) available in the globalized connected world, dealing with

large-scale modeling is not as problematic as it used to be (Jozefowicz, Vinyals, Schuster, Shazeer, & Wu, 2016).

Extracting Data from LLMs.

One level deeper than the previous section, Language Models are statistical models assigning a probability to a sequence of words. Contemporary neural-network-based LMs use models with very large data architecture (in the hundreds of billions of parameters) and are exposed to training on terabytes of datasets of English texts. This very large-scale scenario expands the ability of LMs to generate fluent natural language, also enabling them to be applied to a multitude of other activities (Carlini, et al., 2021).

The concern here is the possibility of exposing information related to the training data, which can contain private data, an existing scenario when training language models. It can be exploited to predict the presence (or not) of any specific content in the training data (Carlini, et al., 2021).

LLMs and the risk of creating stochastic parrots.

How to know when a Large Language Model becomes too big? For the English language alone, the last three years of development and deployment of LLMs have been characterized by the advent of continuously larger LMs. It is necessary to evaluate the impacts of scale in the understanding of what is in the training data. Multiple sources of data will also be sources of different biases (cultural, regional, and interpretation) (Bender, Gebru, McMillan-Major, & Shmitchell, 2021). Consider the anecdote of a random corporation working with their business quarterly data to interpret performance. It is not unusual for two competing areas to use the same data with a strong selection bias to show positive or negative performance based on the same data set for the month. As an example, two competing departments use the same data, from the

same source, with the same quantities, results, and comments. They are evaluating the trends of customer satisfaction with the services provided by one of the departments to the customers of the competing one. During the most recent quarter, the results show 100% satisfaction, 50% satisfaction, and 0 satisfaction. The customer management team defends its position to senior management that the delivery team is doing a poor job which is supported by the data with not only a negative trend but a concerning drop in satisfaction to zero. The delivery team argues that, for the first month, only one of the clients responded to the survey, two clients responded in the second month, and finally no client responded to the survey in the third month of that quarter.

There are mitigating actions to deal with most sources of risk. Extrapolating our example to large scale, curation and documentation of the data sources, including every aspect possible (geography, population, locale, number of surveyed unities – individuals, families, properties, among others) making sure the project starts with datasets created as large as they can be properly documented (Bender, Gebru, McMillan-Major, & Shmitchell, 2021).

It is important to understand the limitations of LMs and have a solid context for the achievements. At the top of the priorities, the list must be avoiding misleading the target audience. Large-scale data sets do not signify success in training the LMs. They are not executing natural language understanding (Bender & Koller, Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data, 2020), and their success is only achieved in tasks where linguistic form can be manipulated (Bender, Gebru, McMillan-Major, & Shmitchell, 2021).

The Massively Pretrained Language Models Storytellers

A recent Stanford University study (See, Pappu, Saxena, Yerukola, & Manning, 2019) dissects the strategy of training large neural language models on “massive amounts of text” for

tasks related to natural language understanding. Their work aimed at understanding the strength of these models as natural language generators instead, given that only anecdotal evidence exists to suggest that these models generate better text quality, with a lack of detailed studies aiming at characterizing their storytelling capabilities. Their findings demonstrate that a general-purpose model architecture can outperform a complex task-specific architecture conceived to augment the story-prompt relevance. In other words, a general-purpose model architecture can sometimes outperform a task-specific complex architecture when sufficient pretraining data is provided. For scenarios where the quality/diversity ratio is small, the text generated is lexically under-diverse using any of the models mentioned above.

Ethical and Social Impacts of LLMs

As AI and ML adoption grows, the possibilities of use and application of the tools and other benefits they bring to society, also represent risks of malicious use of the different LMs available. The list is not short, from discrimination to malicious use, from misinformation to forbidden access. It is expected that LM training tends to reinforce stereotypes and discrimination by default since they are optimized to mirror language with a high degree of accuracy by detecting statistical patterns present in living natural languages, and with it comes, for example, sets of text (data) representing specific communities (Weidinger, et al., 2021).

When training and implementing neural language models, it is expected assumption by the professionals that LMs are more likely to generate racist, sexist, or other toxic languages than not, directing them to build a comprehensive revision and mitigation plan (Gehman, Gururangan, Sap, Choi, & Smith, 2020).

We are already dealing with scale-related impacts on the environment. Working with massive data sets concerns not only the precision and quality of what the adopted LM is

generating but also the consequences to the environment given the large investments needed (cost associated with LM training is also an ongoing concern) in hardware and process capacity to obtain and train the LMs given the chosen datasets (Bender, Gebru, McMillan-Major, & Shmitchell, 2021).

Working with LMs also means organizational responsibility to evaluate and assess potential risk sources, It is necessary to know with precise detail the roots of the risks, the sources of the data, and what each set of data represents, until the point where the selection of the training data includes proper curation resulting in detailed documentation supporting the next decisions (Weidinger, et al., 2021).

This preparedness can -and must – be considered from the scope and design of the LM. Instead of releasing a full final version of a model, the phased (or staged) release is a strategy the implementation team can use to construct each stage with necessary risk assessment and curation, eliminating potential negative social impacts from the very early stages (Solaiman, et al., 2019).

Wrapping up: current application of LLMs

In a possible real-life scenario, it would be automatic to come to the assumption that the use of LMs in machine translation is a low-hanging fruit. Indeed from the perceived precision of translating from one source language to the target language, given that both source and target datasets for training are well-known and documented, the training will tackle a vocabulary map (that includes a pre-determined threshold for more or less occurring words for example) and text reading training (Brants, Popat, Xu, Och, & Dean, 2007).

LLMs are capable of creating narratives of events with segmentation similar to what humans use. Populations naturally develop their storytelling models throughout time. Given the

proximity and the ubiquitous presence of communication vehicles, these models are likely to be constant in the language spoken in determined regions, and thus the perception of life events as a continuous experience. Training LM in studying event cognition at naturalistic scales is enabling more powerfully the prediction of the next word over the backlog of billions of texts used in the training, using the probability of the next N predicted segments of words (Michelmann, Kumar, Norman, & Toneva, 2023).

Finally, the official use of LLMs in assisted medical education is approved for the licensing exams (Kung, et al., 2023). The adoption of these tools in different fields of medicine and, broadly, biology is a promising (and continuously) successful field. Protein sequencing (Rao, Meier, Sercu, Ovchinnikov, & Rives, 2020) for example, the set of evolutionarily related protein sequences known as Multiple Sequence Alignments (MSAs) benefit from using LLMs tools (nominally Jakkckhmmer and HHblits) increases the number and the diversity of sequences returned when executing the iterative search and alignment steps.

Artificial Intelligence and Machine Learning have the potential to be the new and the modern and the disruptor for a long time in technology evolution and technology solutions adoption. It is certainly today a trend like many other Information Technologies that came before, however, it is likely to remain at the focal point for pre-existing and future coming trends. From the deployment of Cloud technologies to the implementation of Big Data strategies, AI and ML will have their role.

How far we will advance with AI is limited only by the human mind, with potential hazards surfacing from the malicious use or poorly curated information used to train the wide and diverse existing models and those being built.

References

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics*, (pp. 5185–5198). doi:10.18653/v1/2020.acl-main.463

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *2021 ACM conference on fairness, accountability, and transparency*, (pp. 610-623).

Brants, T., Popat, A. C., Xu, P., Och, F. J., & Dean, J. (2007). Large Language Models in Machine Translation. In Google (Ed.), *Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 858-867). Prague: Association for Computational Linguistics.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., . . . Raffel, C. (2021). Extracting Training Data from Large Language Models. *30th USENIX Security Symposium* (pp. 2633-2650). USENIX Association.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). *Realtoxicityprompts: Evaluating neural toxic degeneration in language models.* Paul G. Allen School of Computer Science & Engineering, University of Washington, Allen Institute for Artificial Intelligence. Seattle: University of Washington. doi:10.48550/arXiv.2009.11462

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). *Exploring the Limits of Language Modeling.* Google. ArXiv. doi:10.48550/arXiv.1602.02410

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., . . . Diaz-Candido, G. (2023, February 9). Performance of ChatGPT on USMLE: Potential for AI-

assisted medical education using large language models. (A. Dagan, Ed.) *PLOS Digital Health, 2*(2). doi:10.1371/journal.pdig.0000198

Michelmann, S., Kumar, M., Norman, K. A., & Toneva, M. (2023). *Large language models can segment narrative events similarly to humans.* Princeton University. ArXiv. doi:10.48550/arXiv.2301.10297

OpenAI. (2023, February 4). *What are Large Language Models?* Retrieved from OpenAI: https://chat.openai.com/chat/fe928d57-26c4-4944-aae9-28ce36423d9d/

Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., & Rives, A. (2020). *Transformer protein language models are unsupervised structure learners.* Facebook AI Research. bioRxiv. doi:10.1101/2020.12.15.422761

See, A., Pappu, A., Saxena, R., Yerukola, A., & Manning, C. D. (2019). *Do Massively Pretrained Language Models Make Better Storytellers?* Stanford University. ArXiv. doi:10.48550/arXiv.1909.10705

Solaiman, I. B.-V., Krueger, G., Kim, J. W., Kreps, S., McCain, M., Newhouse, A., . . . Wang, J. (2019). *Release strategies and the social impacts of language models.* OpenAI Report. doi:10.48550/arXiv.1908.09203

Vaswani, A., Zhao, Y., Fossum, V., & Chiang, D. (n.d.). *Decoding with large-scale neural language models improves translation.* Citeseer.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P., . . . Isaac, W. (2021). *Ethical and social risks of harm from Language Models.* Cornell University. arXiv. doi:10.48550/arXiv.2112.04359