Project Wenju Manifesto

Motivation

Some thought leaders have repeatedly warned about the potential danger of artificial intelligence and expressed fear that AI may annihilate humans someday. Although such fear has not been shared by the vast majority of computer scientists, a different AI crisis is upon us now, and is having a huge impact on the business world.

As much as enterprises are eager to embrace AI to innovate products, transform business, reduce costs, and improve competitive advantages, they find it very difficult to productionize AI and realize its full benefits, due to the time, budget, and skills required. As a result, the rate of AI adoption has significantly lagged the level of interest, particularly for small- and medium-sized enterprises, which are more resource-constrained. Despite a good number of AI pilot projects for evaluation purposes, only a small portion of those have turned into full-scale, revenue-bearing production. Some industry analysts have pegged enterprise adoption at less than 20% so far. The world is still far away from AI democratization.

Although modern-era AI is centered around machine learning technologies, ironically the crisis of AI production does not have much to do with the adequacy of machine learning algorithms or engines. Consequently, progresses in machine learning platforms have provided little relief in solving the problem. The challenges for the current AI crisis stem from what is needed to develop and execute AI systems end-to-end, of which machine learning is merely a small part.

Objectives

Project Wenju is an initiative focused on the acceleration of production AI. It advocates a holistic and engineering approach to enterprise AI, facilitates the end-to-end integration of production AI systems, streamlines the lifecycle management of AI solutions, and significantly reduces skills requirements and time to value.

Specifically, Wenju seeks to tackle the following challenges in production AI.[1]

The Challenge of Infrastructure: A company's ultimate success with AI depends on how suitable its infrastructure is for its AI applications. Provisioning and managing AI infrastructure requires key insights for technology selection, topology design, configuration engineering, system interoperation, and resource optimization. It must be performed expediently and effectively in order to meet the business needs and maximize the return-on-investment of AI initiatives.

The Challenge of Data: Data is the fuel that powers AI, because machine learning algorithms count on extremely large datasets to reveal patterns, trends, and associations. Big data, as it is popularly called, has four important attributes: volume, velocity, variety, and veracity. These four V's have given rise to many problems such as data quality, data heterogeneity, data silos, data cataloging, data consistency, data performance, and data privacy. If not adequately addressed, these problems will hinder the timely delivery of big data's true value, the fifth V, by AI systems.

The Challenge of Skills: Although AI is intended to automate things as much as possible, the development of AI itself requires extensive human engagement, not counting the new blue-

[1] For a more in-depth discussion on challenges in production AI, see The Real AI Crisis.

collar job of data labeling. AI development requires new skills of data science and machine learning. In addition, software engineers have to relearn a lot of what they take for granted about how to program. AI-related skills are rare and in high demand. The general shortage of skilled resources in the industry calls for reducing the entry barriers to AI.

The Challenge of Trust: Broad adoption of AI will heavily depend on the ability to trust the behavior and output of AI systems. People need assurance that AI is reliable and accountable to people, can explain its reasoning and decision-making, will cause no harm, and will reflect the values and norms of our societies in its outcomes. There is currently a substantial trust gap for AI, which is obstructing an effective path for economic growth and societal benefit.

The Challenge of Operationalization: Operationalizing one machine learning model may not be a big deal, but it is a completely different beast to consistently and effectively operationalize hundreds of AI applications in an enterprise, where the applications are frequently updated and stringent service-level objectives in terms of availability, performance, and prediction quality must be met. In addition, AI systems consist of software, data, and learning components that are intertwined. The interdependencies among those components complicate the continuous integration and delivery of end-to-end AI systems.

Technical Directions

Wenju applies a holistic approach to addressing the challenges in production AI. It will provide a first-of-a-kind overarching platform for the development and management of enterprise AI applications. A key observation is that there are three main building blocks in an AI application: data pipelines that integrate heterogeneous data sources and produce coherent datasets to be used by machine learning, model pipelines that train and tune machine learning models based on the integrated datasets, and insight pipelines that generate business insights from data analytics and model inferencing. Wenju intends to provide unified treatment of all three kinds of pipelines. In addition, it addresses issues that cut across these pipelines, such as infrastructure automation and optimization, end-to-end governance of data and models, continuous integration and delivery of AI application, and hybrid multicloud enablement.

As an integral platform for AI solutioning, Wenju plans to provide a one-stop shop and a unified experience for the entire solution team: data scientists, data engineers, machine learning engineers, software engineers, and IT engineers. It will support the development and operations of data pipelines, model pipelines, and insight pipelines in a consistent manner. That will translate into a common set of programming abstractions for all pipelines, including higher-level APIs and reusable templates to simplify the use of various data and machine learning frameworks. Wenju will also enable the unified management and governance of AI assets, including datasets, models, and code. AI asset management will be based on a common asset catalog and an overarching metadata schema, in order to offer a big-picture view on lineage, trust, and quality. In addition, Wenju will support self-service AI infrastructure with flexibility in technology choices across bare metal, virtual machines, containers, and serverless functions. It will enable turnkey provisioning and configuration of experiment, development, testing, staging and production environments with a wide range of data and AI services.

It is important to note that there has been a proliferation of tools for building AI applications. These tools span many areas, such as infrastructure management, data management, data analytics, machine learning, and domain-specific services for natural language processing, speech, computer vision, and business intelligence. Unfortunately, the current AI tooling landscape is siloed, crowded and confusing to developers. New tools keep surfacing, existing

tools keep evolving, and no tool is suitable for all use cases. These tools have a long learning curve and require knowledge and skills not readily available in enterprises, hampering productivity. Further, different roles on an AI development team have to use different tools, and integrating the output from those tools relies on glue code that often incurs large overhead and technical debt. Wenju will not reproduce the capabilities already available in existing AI tools. Instead, it will build on the existing tools and surface their capabilities through an integrated and simplified developer experience and interface that embeds best engineering practices. Ultimately Wenju holds promise to bridge skill gaps, facilitate collaboration, and shorten time to value for enterprise AI.

Conclusion

An analogy may be drawn between the crisis of AI production today and the software crisis that stemmed in the late 1960s. The term software crisis referred to difficulties in writing high-quality and efficient computer programs within the required time and budget. The major cause of the software crisis was that computers had become several orders of magnitude more powerful, giving rise to opportunities for much larger and much more complex software programs. Unfortunately, the same methods used to build small software systems were not applicable to the development of large-scale software. In response to the software crisis, software engineering emerged as a discipline for the establishment and application of well-defined engineering principles and procedures for software production. Over the years, many software engineering practices have been developed to address the growing demands of enterprises. Those practices, ranging across information hiding, model-driven architecture, object-oriented design, agile development, and software as a service, have exerted a very positive impact on the industry and society.

The current AI crisis arises from the advances in hardware technologies, the breakthroughs in machine learning algorithms, and the explosion of digital data, which in combination have made it feasible to incorporate AI in business operations and processes. However, it takes a huge leap to move from the development of machine learning prototypes in lab settings to the development of enterprise AI systems for production. The AI crisis calls for AI engineering, i.e., applying a systematic, disciplined and quantifiable approach to AI production. AI systems are constructed differently from conventional programmable software. AI systems are based on machine learning from big data. They require a variety of personas working together to generate distinct artifacts like datasets, models, and code modules. Existing software engineering techniques are insufficient for AI development. New AI engineering methodologies and platforms are needed to solve the AI crisis and to unlock AI’s potential to businesses and society.

Project Wenju is an open collaboration project with open design, open APIs, open communication, and open governance. It welcomes contributions from the broad community and in all forms. It looks forward to close partnership with both academia and industry to impact carry forward the industrialization of AI, through a strong and open ecosystem.