The Role of Open-Source Software in Artificial Intelligence

Jim Spohrer

■ *With this publication, we launch a new column for AI Magazine on the role of open-source software in artificial intelligence. As the column editor, I would like to extend my welcome and invite AI Magazine readers to send short articles for future columns, which may appear in the traditional print version of AI Magazine, or on the AI Magazine interactive site currently under development. This introductory column serves to highlight my interests in open-source software and to propose a few topics for future columns.*

My Surprise

The field of artificial intelligence (AI) is arguably 64 years old now, as measured from the summer of 1956 Dartmouth Workshop. What is most surprising to you about AI today? For me, the answer is simple — it's the extent to which open-source software has advanced AI. During the 1970s, I was a student taking AI courses at the Massachusetts Institute of Technology. After graduating, I was working at an early AI startup doing speech recognition, where we protected our code as an important proprietary secret. In the 1980s, while I was a graduate student at Yale working toward my PhD in computer science, even more AI startups were forming with the same mentality toward protecting their code; back then, except for a few LISP companies, few took an open-source first approach.

Fast forward to 2020, where companies like Databricks (Spark for scaling data and AI workflows) to Seldon (Seldon Core for AI model serving) are using open-source first approaches to build mindshare and market share. In fact, as part of the work at IBM's Center for Opensource Data and AI Technologies,[1] we are encouraged to participate in open-source data science and AI communities. Recently, I was elected chairperson of the Technical Advisory Board at the LF AI & Data Foundation,[2] which helps incubate and launch open-source AI projects.

The AI Landscape

The LF AI Landscape tool tracks approximately three-hundred AI-related open-source community projects such as TensorFlow, PyTorch, SciKit-learn, and many others that collectively generate approximately one-million lines of code changes every two weeks. Each project has its own information card that includes information such as the number of GitHub stars — a common measure of project popularity.

 ISSN 0738-4602

The cards also contain information about the organizations that host code repositories (repos) on GitHub. Currently, the projects collectively have more than 1.5 million stars; the public companies with projects have a combined market cap of $13.94 trillion dollars, and the startup companies with projects have a combined funding of $53.9 billion dollars (based on Crunchbase[3]). If you check now, the numbers will probably be higher still!

The landscape of AI-related projects continues to grow, adding a few new popular projects each month. The landscape groups projects into twelve categories: data (the largest category), model (the second largest category), machine learning, deep learning, reinforcement learning, programming, notebook environment, trusted and responsible AI, distributed computing, security and privacy, natural language processing, and education.

Developers

Why is there all this interest in open source? In a word, it is because of developers. The GitHub Octoverse[4] continues to grow with no end in sight.

All the major information technology vendors provide developer websites to help developers such as data scientists, students, faculty, researchers, practitioners, and really anyone, interested in using software to build solutions. These developer portals provide open-source on-ramps and starter kits for those interested in developing AI projects as well as using AI in solutions, and typically include other advanced technologies such as blockchain, internet of things, analytics, and more. For example, Microsoft[5] has quick links to AI, bots, and machine learning, as well as other areas. At Facebook,[6] an AI quick link takes you to tools, research, and Wit.ai,[7] and of course Google[8] has a quick link to TensorFlow. At IBM,[9] there are forty-seven technologies, including COBOL, as well as a dozen AI-related links.

My team at the Center for Open-source Data and AI Technologies provides the data asset exchange, model asset exchange, and a range of code patterns to get started quickly, as well as annual challenges with prizes to use AI for good. Recently, Trusted AI (fairness, explainability, robustness) projects from IBM Research were donated to the LF AI & Data Foundation. IBM's approach to open source is founded on the advantages of vendor-neutral, open governance in a foundation for long-term success.

When students and independent developers around the world get access to the code, learn the code, and help improve the code, we then get documentation and test cases; the supply of talent increases faster than for proprietary code. Add to this talent factor that companies can then use their scarce resources on higher-level innovations, and the advantage of open source over proprietary software becomes clearer. Everyone wins, because easy access helps more AI startups get up and running, offering customers a wider variety of applications and solutions.

Accelerating Innovation

Research best-practice is to ensure others can replicate your results. Therefore, open-source software acts as an important common building block and has played a significant role in accelerating AI research and innovation, especially with respect to recent advances in deep learning. One might say that the three R's of open-source AI are *read*, *redo*, and *report*: read a paper, get the code to redo the experiment, then report any new insights or improvements. For example, PapersWithCode[10] is a website that tracks AI innovations across a range of categories that includes computer vision, natural language processing, medical applications, speech, game play, time series, robots, knowledge base applications, and reasoning. All this code needs datasets, and Kaggle[11] is a website and challenge platform for data scientists to sharpen their skills. The final ingredient, beyond code and data, is computing power. Investments in computing power tend to both improve speed and lower energy requirements, and unlike the software stack, are mostly proprietary in nature.

Concluding Remarks

Are you interested in contributing to this column? Here are some topics that could be explored in future columns: What are the "hot" emerging open-source AI-related projects? What open communities exist for open-source AI, and what are their best practices, such as AI resources, and so on? How might an analysis of open-source projects be used in measuring AI progress? I look forward to your submissions, which should be sent to the attention of Jim Spohrer at aimag-ed@aaai.org.

Notes

1. www.ibm.com/opensource/centers/codait/
2. lfaidata.foundation/
3. www.crunchbase.com/
4. octoverse.github.com/
5. developer.microsoft.com
6. developer.facebook.com
7. wit.ai/
8. developer.google.com
9. developer.ibm.com
10. paperswithcode.com/
11. www.kaggle.com/

James C. Spohrer is the Director of IBM Cognitive OpenTech. He was formerly the Director of IBM Global University Programs Worldwide. Between 2003 and 2009, he was the Director of Almaden Services Research with IBM at the IBM Almaden Research Center. He was an advocate of the service science, management, and engineering initiative across companies, governments, and academics. His research group received IBM awards for customer modeling and mapping of global service systems including performance measures; costing and pricing of complex, interorganizational service projects; analytics and information service innovations; and process improvement methods and innovation. He had earlier served in a number of roles at IBM. He received a PhD in Computer Science and AI from Yale University in 1988.