Open Fabric for Deep Learning Models

Falk Pollok
IBM Research
falk.pollok@ibm.com

Scott Boag
IBM Research
scott_boag@us.ibm.com

Maria-Irina Nicolae
IBM Research
maria-irina.nicolae@ibm.com

Abstract

We will show the advantages of using a fabric of open source AI services and libraries, which have been launched by the AI labs in IBM Research, to train, harden and de-bias deep learning models. The motivation is that model building should not be monolithic. Algorithms, operations and pipelines to build and refine models should be modularized and reused as needed. The componentry presented meets these requirements and shares a philosophy of being framework- and vendor-agnostic, as well as modular and extensible. We focus on multiple aspects of machine learning that we describe in the following. To train models in the cloud in a distributed, framework-agnostic way, we use the Fabric for Deep Learning (FfDL). Adversarial attacks against models are mitigated using the Adversarial Robustness Toolbox (ART). We detect and remove bias using AI Fairness 360 (AIF360). Additionally, we publish to the open source developer community using the Model Asset Exchange (MAX). Overall, we demonstrate operations on deep learning models, and a set of developer APIs, that will help open source developers create robust and fair models for their applications, and for open source sharing. We will also call for community collaboration on these projects of open services and libraries, to democratize the open AI ecosystem.

1 Introduction

The AI labs of IBM Research have imagined and launched a series of AI componentry for operations on deep neural network models. The motivation is that model building should not be monolithic. Algorithms and operations to build and refine models should be modularized and reused as needed. The componentry presented share a philosophy of being framework and vendor agnostic as well as extensible.

The Fabric for Deep Learning (FfDL, pronounced "fiddle") is a platform for training deep learning models in the cloud via Kubernetes. It is largely framework agnostic and thus supports Tensorflow, PyTorch, Keras, Caffe2 and others, while making it easy to add new frameworks as well. It is completely open source and can be cloned from Github [1]. The FfDL platform uses a microservices architecture to reduce coupling between components, keep each component simple and as stateless as possible, isolate component failures, and allow each component to be developed, tested, deployed, scaled, and upgraded independently. Leveraging the power of Kubernetes, FfDL provides a scalable, resilient, and fault tolerant deep-learning framework. A resource provisioning layer enables flexible job management on heterogeneous resources, such as GPUs and CPUs on top of Kubernetes. In addition to standard training, FfDL can be integrated with other libraries to deliver specific properties into models, like fairness or security robustness.

Adversarial attacks against machine learning models have proven to be an important risk to AI reliability [2, 3]. Protecting models against attacks is vital, since otherwise the results can be influenced to benefit attackers by evasion and poisoning attacks, with disastrous effects for many applications (e.g. CCTV, voice assistants, autonomous cars). To provide robustness for models, we use the Adversarial Robustness Toolbox (ART [4, 5, 6, 7]). ART is a framework-agnostic library

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

providing implementations of state-of-the-art attacks, hardening and robustness evaluation methods for classifiers. Its modular design allows for easy extension to new approaches and composition of methods as building blocks via interfaces. Like FfDL, ART can be used with a wide range of machine learning frameworks. Overall, this library provides a comprehensive foundation of defense mechanisms for real-world AI systems, cmp. [8].

Furthermore, we show how to find and remove bias in datasets and models using the AI Fairness 360 toolkit (AIF360), cmp. [9, 10]. AIF360 is a Python package that uses fairness metrics like statistical partity difference, equal opportunity difference, average odds difference, disparate impact or the Theil index to detect bias. It can leverage explainers to report its findings in natural language and finally use bias mitigators to remove bias from the dataset which can be verified using the same fairness metrics from the first step. Besides the obvious ethical reason to strive for fairness, a company might want to do so to adapt its models to its strategic goals (e.g. to sell to all income groups equally) or to avoid legal traps like deciding based on age which might be considered age discrimination in the country the model is supposed to be used in and thus illegal.

We publish to the open source developer community using the Model Asset Exchange (MAX). MAX can be described as an app store to discover, share and rate models that can be implemented using popular machine learning frameworks and that provides a standardized way to classify, annotate and deploy them, cmp. [11].

Finally, we will give a brief overview of external collaborations. The FfDL project is targeted at model training, and we chose to partner with Seldon [12] as a serving solution for the obtained models. We also integrated Uber's Horovod [13] as an alternative to parameter servers for distributed training, as well as the machine learning and predictive analytics platform H2O [14].

2 Related Work

2.1 Publications

With the success of deep learning several cloud and AI companies have released platforms to offer deep learning as a service, e.g. IBM Watson Machine Learning [15], AWS SageMaker [16], Microsoft Azure Machine Learning [17], Google TFX [18], Uber Michelangelo [19] and Facebook FBLearner [20]. The main idea is to isolate infrastructure challenges and hide the complexity of multi-tenancy, metrics collection, fault tolerance and cloud provisioning from data scientists such that they can concentrate on their actual work. Besides these products, there have also been several open source efforts, the most popular of which is Kubeflow [21] followed by FfDL [1] and Microsoft Deep Learning Workspace [22]. While solving the same problem, there are differences between the projects in terms of job scheduling and distribution, framework support, ecosystem and general architecture.

Our team has published on IBM's Deep Learning as a Service [23], Scalable Multi-Framework Multi-Tenant Lifecycle Management of Deep Learning Training Jobs [24] and Dependability in a Multi-tenant Multi-framework Deep Learning as-a-Service Platform [25], but here we present the open source counterpart and wider ecosystem. These publications and links to the product allow us to not only speak about open source technology, but also real-world challenges regarding multi-tenancy, fault tolerance, scalability and open sourcing itself, since we encountered several complications when moving from code deeply integrated with internal infrastructure to open code everyone can deploy to on-premise or public cloud targets.

2.2 Open AI Fabric in Media and at Summits

After releasing FfDL at Think earlier this year, we have already published multiple blog posts about it [26, 27, 28, 29, 30] and were also published about by diverse external sources like TechCrunch, IT World and InfoQ, e.g. [31, 32, 33]. Furthermore, we already presented the platform at several venues like the IEEE Computer Society Silicon Valley [34] and have published additional video presentations, e.g. [35, 36].

3 Structure of the Talk

We will first discuss deployment options, since FfDL can be setup both on-premise and in public clouds making it relevant for many industries which are reluctant to compute confidential models on public infrastructure. This also includes aspects like the storage setup (FfDL currently uses cloud object storage via S3 APIs for data access, as well as NFS for logs) and GPU support in Kubernetes.

Once set up, we will demonstrate how to train a model based on a manifest that specifies the framework, data access and requested resources. We will briefly describe the architecture and APIs, how we expose a public REST endpoint for the end user while internally using gRPC to reduce latency as well as how to distribute training with Horovod or parameter servers.

We will also present how to train a model, launch an adversarial attack against it in order to evaluate its robustness, and subsequently demonstrate how to protect a model by employing the Adversarial Robustness Toolbox. Moreover, we will show bias detection and how to submit to MAX.

Finally, we can discuss potential collaborative future ideas for open source AI and share some thoughts about the state of AI systems in the open source world.

References

[1] Fabric for deep learning (FfDL) on github. https://github.com/IBM/FfDL, 2018.

[2] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013.

[3] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 387–402. Springer, 2013.

[4] Adversarial robustness toolbox. https://developer.ibm.com/code/open/projects/adversarial-robustness-toolbox/, 2018.

[5] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, and Ben Edwards. Adversarial robustness toolbox v0.3.0. *CoRR*, 1807.01069, 2018.

[6] Maria-Irina Nicolae and Mathieu Sinn. The adversarial robustness toolbox: Securing ai against adversarial threats. https://www.ibm.com/blogs/research/2018/04/ai-adversarial-robustness-toolbox/, 2018.

[7] Maria-Irina Nicolae, Mathieu Sinn, Nathalie Baracaldo, and Heiko Ludwig. The adversarial robustness toolbox v0.3.0: Closing the backdoor in AI security. https://www.ibm.com/blogs/research/2018/08/art-v030-backdoor/, 2018.

[8] Adversarial robustness toolbox on github. https://github.com/IBM/adversarial-robustness-toolbox, 2018.

[9] AI fairness 360 toolkit on github. https://github.com/IBM/AIF360, 2018.

[10] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2017.

[11] Ibm code model asset exchange (max). https://developer.ibm.com/code/exchanges/models/, 2018.

[12] Animesh Singh and Clive Cox. Serve it hot! deploy your FfDL trained models using seldon. https://developer.ibm.com/code/2018/06/12/serve-it-hot-deploy-your-ffdl-trained-models-using-seldon/, 2018.

[13] Animesh Singh and Alex Sergeev. Scalable distributed training using horovod in ffdl. https://developer.ibm.com/code/2018/07/18/scalable-distributed-training-using-horovod-in-ffdl/, 2018.

[14] Animesh Singh, Nicholas Png, Tommy Li, and Vinod Iyengar. H2O-3 on FfDL: Bringing deep learning and machine learning closer together. https://developer.ibm.com/code/2018/06/25/h2o-on-ffdl-bringing-deep-learning-and-machine-learning-closer-together/, 2018.

[15] Ibm watson machine learning. https://www.ibm.com/cloud/machine-learning, 2018.

[16] Amazon Web Services. Amazon Sagemaker. https://aws.amazon.com/sagemaker/, 2017.

[17] Microsoft Azure. Microsoft Azure Machine Learning. https://azure.microsoft.com/en-us/overview/machine-learning/, 2018.

[18] Denis Baylor, Eric Breck, Heng-Tze Cheng, Noah Fiedel, Chuan Yu Foo, Zakaria Haque, Salem Haykal, Mustafa Ispir, Vihan Jain, Levent Koc, Chiu Yuen Koo, Lukasz Lew, Clemens Mewald, Akshay Naresh Modi, Neoklis Polyzotis, Sukriti Ramesh, Sudip Roy, Steven Euijong Whang, Martin Wicke, Jarek Wilkiewicz, Xin Zhang, and Martin Zinkevich. Tfx: A tensorflow-based production-scale machine learning platform. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 1387–1395, New York, NY, USA, 2017. ACM.

[19] Jeremy Hermann and Mike Del Baso. Meet michelangelo: Uber's machine learning platform. https://eng.uber.com/michelangelo/, 2017.

[20] Jeffrey Dunn. Introducing fblearner flow: Facebook's AI backbone. https://code.fb.com/core-data/introducing-fblearner-flow-facebook-s-ai-backbone/, 2016.

[21] Kubeflow. https://github.com/kubeflow/kubeflow, 2018.

[22] Microsoft deep learning workspace. https://github.com/Microsoft/DLWorkspace, 2018.

[23] Bishwaranjan Bhattacharjee, Scott Boag, Chandani Doshi, Parijat Dube, Ben Herta, Vatche Ishakian, K. R. Jayaram, Rania Khalaf, Avesh Krishna, Yu Bo Li, Vinod Muthusamy, Ruchir Puri, Yufei Ren, Florian Rosenberg, Seetharami R. Seelam, Yandong Wang, Jian Ming Zhang, and Li Zhang. IBM deep learning service. *CoRR*, abs/1709.05871, 2017.

[24] Scott Boag, Parijat Dube, Benjamin Herta, Waldemar Hummer, Vatche Ishakian, K. R. Jayaram, Michael H. Kalantar, Vinod Muthusamy, Priya Nagpurkar, and Florian Rosenberg. Scalable multi-framework multi-tenant lifecycle management of deep learning training jobs. In *NIPS Workshop on Systems for Machine Learning*, 2017.

[25] Scott Boag, Parijat Dube, Kaoutar El Maghraoui, Benjamin Herta, Waldemar Hummer, K. R. Jayaram, Rania Khalaf, Vinod Muthusamy, Michael H. Kalantar, and Archit Verma. Dependability in a multi-tenant multi-framework deep learning as-a-service platform. *CoRR*, abs/1805.06801, 2018.

[26] FfDL IBM code open project page. https://developer.ibm.com/code/open/projects/fabric-for-deep-learning-ffdl/, 2018.

[27] Angel Diaz, Ruchir Puri, and Rania Khalaf. Fabric for deep learning. https://developer.ibm.com/code/2018/03/20/fabric-for-deep-learning/, 2018.

[28] Animesh Singh and Scott Boag. Democratize AI with fabric for deep learning (FfDL). https://developer.ibm.com/code/2018/03/20/democratize-ai-with-fabric-for-deep-learning/, 2018.

[29] Animesh Singh and Scott Boag. Introducing fabric for deep learning (FfDL). https://medium.com/ibm-watson/introducing-fabric-for-deep-learning-ffdl-542522774775, 2018.

[30] Deep learning advances from ibm research. https://www.ibm.com/blogs/research/2018/03/deep-learning-advances/, 2018.

[31] Rags Srinivas. Q&a on ibm's fabric for deep learning with chief architect of watson (infoq). https://www.infoq.com/news/2018/04/ffdl-ruchir-puri, 2018.

[32] Peter Sayer. Ibm wants to open up the deep learning expertise bottleneck (it world). https://www.itworld.com/article/3263454/cloud-computing/ibm-wants-to-open-up-the-deep-learning-expertise-bottleneck.html, 2018.

[33] Frederic Lardinois. Ibm launches deep learning as a service inside its watson studio (tech crunch). https://techcrunch.com/2018/03/19/ibm-launches-deep-learning-as-a-service-inside-its-watson-studio/, 2018.

[34] FfDL at IEEE computer society silicon valley. https://www.youtube.com/watch?v=wPKin0mN9LA&t=743, 2018.

[35] Introduction to fabric for deep learning (FfDL). https://www.youtube.com/watch?v=aKOqFL7VWhI, 2018.

[36] Fabric for deep learning. https://www.youtube.com/watch?v=nQsYWmkfLP4, 2018.