Acumos Day – NYC Nov 5, 2018
https://events.linuxfoundation.org/events/acumosaiday-ny/

AI Today and Tomorrow

Susan Malaika @sumalaika malaika@us.ibm.com
developer.ibm.com/opentech/category/susan-malaika/
developer.ibm.com/open

November 2018
Agenda

Some Remarks on AI

Some AI Projects

Sharing Data and Models – Some Considerations

Trusting AI Systems – Some Projects

What’s Coming for AI

Backup Materials
Artificial Intelligence, Machine Learning & Deep Learning

- Training Computers instead of Programming Them

Fact Sheets for Data and AI
- Data
- Data & Model Licenses

Artificial Intelligence
A program that can sense, reason, act, and adapt

Machine Learning
Algorithms whose performance improve as they are exposed to more data over time

Deep Learning
Subset of machine learning in which multilayered neural networks learn from vast amounts of data

Model Marketplaces
- Models
- Model Exchange

Wired
Soon, we won't program computers. We'll train them.

/* the end of code */

What that means for us.

IBM
Building AI is still hard

- Lack of massive labeled data sets in enterprises
- Lack of reliable tools for monitoring
  - No feedback loop to improve models in situ
- No standards to inject AI
  - No automation
  - Concerns over security, bias, and ethics
- Many point solutions
- No patterns or abstractions
- Black box models
- Rapid innovation down to the hardware
- No testing and CI/CD methodologies for AI
- Extend to edge, private and public clouds
- APIs are not enough
- Composition needs to be flexible and reactive
- Lack of massive labeled data sets in enterprises
- No feedback loop to improve models in situ
- Concerns over security, bias, and ethics
- No standards to inject AI
  - No automation
Challenges of Building and deploying AI Models Today

### Training
- “Roll your own” home-brew environments
- Stateful, compute-intensive execution at odds with cloud-native design
- Stresses cloud networking, storage, and hardware
- Open source components evolving at different rates and speed

### Deployment
- Testing and debugging neural nets is an active research topic
- Model evolution based on feedback is unavailable
- Enterprise readiness for compliance and traceability is not well understood

### Inference
- Must handle streaming data
- Near-real-time response required though inferencing on large deep learning networks is compute intensive
- Must be able to run in the cloud and at the edge
AI Engineering: An emerging discipline

Data handling tools
- Image/Video
- Audio
- Text
- Language

DLaaS Cloud Platform & Access to Frameworks
- TensorFlow
- Caffe
- Theano
- OpenCV
- torch

Visualization & Human-Computer interaction

Network optimization tools

Computation & Distributed Learning
- Power Systems
- NVIDIA

AI Model Lifecycle Management

AI Open Scale
Three approaches for building AI Models

1. **pre-trained AI**
   - pre-trained model
   - app developer or SME

2. **transfer learning**
   - pre-trained model
   - app developer or SME
   - your domain data

3. **custom AI**
   - data scientist
   - your domain data
   - custom model

Watson Visual Recognition
Natural Language Understanding
Watson Speech to Text
Watson Text to Speech
...

Watson Visual Recognition
Natural Language Classifier
Watson Speech to Text
...

Watson Studio (Mostly open source)
Watson Machine Learning
Deep Learning as a Service (FfDL is open source project)
...

Center for Open Source Data and AI Technologies

Code - Build and improve practical frameworks to enable more developers to realize immediate value (e.g. FfDL, MAX, Tensorflow, Jupyter, Spark)

Content – Showcase solutions to complex and real world AI problems

Community – Bring developers and data scientists to engage with IBM (e.g. MAX)

CODAIGHT

codait.org

https://m.interglot.com/fr/en/codait

codait (French) = coder/coded

Improving Enterprise AI lifecycle in Open Source
The Emergence of LeaderBoards in AI

Leaderboards – such as those in Kaggle - the home of data science contests that utilize open tech & datasets for predictive modeling – resulting in:

• Ranking of data scientists world-wide
• Fresh datasets, data science models, methods, and education - all in open source
• New coursera class https://www.coursera.org/learn/competitive-data-science/home/welcome
• Companies (that sponsor datasets and contests) who benefit through:
  • Recruitment of great employees ; Eminence of own employees; Excellent publicity ; Better understanding of what can be done with their data, Being part of a global AI conversation around open technologies

Sample contest:

Women in Data Science Datathon Feb 2018

Xi Lui and Ye Wang, Worcester Polytechnic Institute’s data science graduate program, beat 230 teams composed of students, faculty, and professional data scientists from 26 countries.

IBM had 12 wonderful teams in the contest (more than any other institution) - the highest ranked was at 7

The contest goal was to predict if a person is male or female by examining the responses the people gave to some questions.

Kaggle: "a way to organize the brainpower of the world’s most talented data scientists and make it accessible to organizations of every size“ - Hal Varian, Google
Open Source & AI in the MENA region (Middle East & North Africa)

Minister of AI in UAE:

Of course, no plan for the future can be complete without considering the role that artificial intelligence (AI) will play, so on October 19, the UAE became the first nation with a government minister dedicated to AI. Yes, the UAE now has a minister for artificial intelligence.

“We want the UAE to become the world’s most prepared country for artificial intelligence,” UAE Vice President and Prime Minister and Ruler of Dubai His Highness Sheikh Mohammed bin Rashid Al Maktoum said during the announcement of the position, according to Gulf News.

The first person to occupy the state minister for AI post is H.E. Omar Bin Sultan Al Olama. The 37-year-old is currently the Managing Director of the World Government Summit in the Prime Minister’s Office at the Ministry of Cabinet Affairs and the Future, and he holds degrees in Project Management and Excellence from the American University of Sharjah and a Bachelor of Business Administration from the American University of Dubai.

Github Partnerships in MENA:

From Report by Prof Manar AbuTalib UAE:

See AI projects at NYU AD social good hackathon in May 2018 (which may inspire code patterns)


https://github.com/NYUAD-Hackathon-2018

© 2018 IBM Corporation
Sharing Data and Models – Some Considerations

Separator Page
Community Data License Agreement

There are two CDLA license agreements:

- “Sharing” based on a form of copyleft designed to encourage recipients to participate in reciprocal sharing of data
- “Permissive” an approach similar to permissive open source licenses (e.g. Apache, BSD or MIT) where recipients are not required to share any changes

Candidate users of CDLA:
- Communities training AI and ML systems
- Public-private infrastructure (e.g. data on traffic)
- Researchers
- Companies with mutual interests in sharing data

License Announced in November 2017 by Linux Foundation
Egeria Design philosophy – Share metadata across repositories

Open Metadata Access Services

Open Metadata Repository Services

Use cases, Personas, Practitioners input

Data integration, availability and integrity best practices

Egeria - Realizing open metadata and governance

- Delivering core technology
- Recruiting vendors
- Assisting practitioners

https://github.com/odpi/egeria
Datasheets Proposal

• The machine learning community has no standardized way to document how and why a dataset was created, what information it contains, what tasks it should and should not be used for, and whether it might raise any ethical or legal concerns. To address this gap, we propose the concept of datasheets for datasets.

• In the electronics industry, it is standard to accompany every component with a datasheet providing standard operating characteristics, test results, recommended usage, and other information. Similarly, we recommend that every dataset be accompanied with a datasheet documenting its creation, composition, intended uses, maintenance, and other properties.

• Datasheets for datasets will facilitate better communication between dataset creators and users, and encourage the machine learning community to prioritize transparency and accountability.

Sample questions:
• Why was the dataset created? (e.g., was there a specific intended task gap that needed to be filled?)
• Who funded the creation of the dataset?
• What preprocessing/cleaning was done? (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances)
• If it relates to people, were they told what the dataset would be used for and did they consent? If so, how? Were they provided with any mechanism to revoke their consent in the future or for certain uses?
• Will the dataset be updated? How often, by whom?

What is ONNX?

ONNX is an open format to represent deep learning models. With ONNX, AI developers can more easily move models between state-of-the-art tools and choose the combination that is best for them. ONNX is developed and supported by a community of partners.

ONNX tutorials: import and export from frameworks

<table>
<thead>
<tr>
<th>Framework/tool</th>
<th>Installation</th>
<th>Exporting to ONNX (frontend)</th>
<th>Importing ONNX models (backend)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caffe2</td>
<td>part of caffe2 package</td>
<td>Exporting</td>
<td>Importing</td>
</tr>
<tr>
<td>PyTorch</td>
<td>part of pytorch package</td>
<td>Exporting, Extending support</td>
<td>coming soon</td>
</tr>
<tr>
<td>Cognitive Toolkit (CNTK)</td>
<td>built-in</td>
<td>Exporting</td>
<td>Importing</td>
</tr>
<tr>
<td>Apache MXNet</td>
<td>part of mnn package docs github</td>
<td>Exporting</td>
<td>Importing</td>
</tr>
<tr>
<td>Chainer</td>
<td>chainer/onnx-chainer</td>
<td>Exporting</td>
<td>Importing</td>
</tr>
<tr>
<td>TensorFlow</td>
<td>onnx/onnx-tensorflow</td>
<td>Exporting</td>
<td>Importing [experimental]</td>
</tr>
<tr>
<td>Apple CoreML</td>
<td>onnx/onnx-coreml and onnx/onnxmtools</td>
<td>Exporting</td>
<td>Importing</td>
</tr>
<tr>
<td>SciKit-Learn</td>
<td>onnx/onnxmtools</td>
<td>Exporting</td>
<td>n/a</td>
</tr>
<tr>
<td>ML.NET</td>
<td>built-in Convert to ONNX-ML</td>
<td>Exporting</td>
<td>n/a</td>
</tr>
<tr>
<td>Menoh</td>
<td>pfnet-research/menh</td>
<td>n/a</td>
<td>Importing</td>
</tr>
</tbody>
</table>

What is PFA for?

Hardening a data analysis

Data analysis is not software development: a different set of best practices apply. For a large software project, one should start by designing a maintainable architecture, but for data analysis, one should start by examining the dataset in as many ways as possible. Sometimes, a simple observation in this exploratory phase dramatically changes one's analysis strategy.

The worlds of data analysis and software development clash when a poorly structured analytic procedure must be scaled up to a large production workflow. The “try anything, get feedback quickly” mindset that was an asset in the development phase leads to failures in production. As data analysts mature, they must be hardened— they must have fewer dependencies, a more maintainable structure, and they must be robust against errors.
What does it take to trust a decision made by a machine?
(Other than that it is 99% accurate)

Is it fair?

Is it easy to understand?

Did anyone tamper with it?

Is it accountable?

© 2018 IBM Corporation
Tusting AI Systems – Some Projects

Factsheets for AI Services

Concerns about safety, transparency, and bias in AI are widespread, and it is easy to see how they erode trust in these systems. Part of the problem is a lack of standard practices to document how an AI service was created, tested, trained, deployed, and evaluated, how it should operate, and how it should (and should not) be used. To address this need, my colleagues and I recently proposed the concept of factsheets for AI services. In our paper [1], we argue that a Supplier’s Declaration of Conformity (SDoC, or factsheet, for short) be completed and voluntarily released by AI service developers and providers to increase the transparency of their services and engender trust. As in nutrition labels for foods or information sheets for appliances, factsheets for AI services provide information about the product’s important characteristics. Standardizing and publicly providing this information is key to building trust in AI services across the industry.

Supplier’s Declaration of Conformity

Aleksandra Mojsilovic
IBM Fellow, IBM Research

Adversarial Robustness Toolkit

Adversarial Attack Example – Pandas & Capuchins

- Perturb model inputs with crafted noise
- Model fails to recognize input correctly
- Attack undetectable by humans
- Random noise does not work

AI Fairness 360

The AI Fairness 360 toolkit (AI360) is an open source software toolkit that can help detect and remove bias in machine learning models.

© 2018 IBM Corporation
Adversarial Attack Example – Self Driving Cars

Image segmentation

Attack noise hides pedestrians from the detection system.

AI Fairness 360

Datasets

Toolbox

  Fairness metrics (30+)
  Fairness metric explanations
  Bias mitigation algorithms (9+)

Guidance

Industry-specific tutorials


Differentiation

Comprehensive bias mitigation toolbox (including unique algorithms from IBM Research)

Several metrics and algorithms that have no available implementations elsewhere

Extensible

Designed to translate new research from the lab to industry practitioners (e.g. scikit-learn’s fit/predict paradigm)
Meetup Nov 7 2018 – 6:15 for 6:30pm start at 303 Spring Street, Manhattan: Fairness in Artificial Intelligence

Register here:

https://www.meetup.com/Big-Data-Developers-in-NYC/events/255613362/

https://www.meetup.com/ibmcodenyc/events/256005884/


Attend or present at weekly AI talks Thursdays @ 10:30 ET: http://cognitive-science.info/community/weekly-update/

Next call on Thursday Nov 8 is on the CDLA
What’s next for AI?

**AI Everywhere**
- Healthcare
- Finance
- Agriculture
- Government
- Education
- Energy
- Science
- Business solutions

**Deeper Insights**
- Data-centric systems
- Distributed Deep Learning
- Neuromorphic systems
- Quantum computing
- Homomorphic encryption
- Machine foresight
- Cognitive discovery

**Engagement Reimagined**
- Human-machine collaboration
- New AI modalities
- Augmented reality
- Global trade logistics
- Blockchain for payments

**Personalization at Scale**
- Personalized healthcare
- Micro-segmentation
- Personalized finance
- Targeted marketing
- Personalized learning
- Individualized solutions

**Instrumented Planet**
- Environmental solutions
- Digital agriculture
- Connected cars
- Geospatial-temporal data and analytics
- Smart sensors

IBM Research / © 2018 IBM Corporation
Handy Links – Sharing Data, Metadata, & Models

- Community Data License Agreement
  - https://www.cdla.io/

- ODPi Egeria - for sharing metadata
  - https://github.com/odpi/data-governance
  - https://github.com/odpi/egeria
  - https://www.odpi.org/

- Sharing Models – Model storage &/or exchange formats
  - PFA - Portable Format for Analytics http://dmg.org/pfa/docs/motivation/
  - ONNX - Open Neural Network Exchange https://onnx.ai/

- Open Source Model Marketplaces – Examples:
Handy Links – Trusting AI & More

• Wired Magazine – The end of code https://www.wired.com/2016/05/the-end-of-code/
  – Soon We Won't Program Computers. We'll Train Them Like Dogs

• Women in Data Science 2019
  – Conference https://www.widsconference.org/
  – Datathon https://www.widsconference.org/datathon.html

• AI Fairness 360 https://github.com/IBM/AIF360

• AI Adversarial Robustness Toolkit https://github.com/IBM/adversarial-robustness-toolbox


• Increasing Trust in AI Services through Supplier's Declarations of Conformity https://arxiv.org/abs/1808.07261


This paper reviews the challenges and metrics for enterprise workloads, the benchmark tests that are available, and the gaps which need to be filled.

The paper identifies the following areas as important to enterprises concerned about performance:

1. **Model training performance**
   - data labeling / preparation
   - time-to-accuracy
   - computational time / cycles
   - throughput-to-accuracy

2. **Hyper-parameter optimization performance**

3. **Inference runtime performance**

The paper offers a summary table of the main three AI areas important to enterprises, alongside:

- Workload profile,
- Important performance indicators to assess the task’s efficiency
- Potential technical bottlenecks to look out for that could limit the AI tasks performance delivered by a given solution.
# AI tasks: Model training

<table>
<thead>
<tr>
<th>AI Task</th>
<th>Workload Profile</th>
<th>Important Performance Indicators</th>
<th>Potential Technical Bottlenecks in Standalone Scenarios</th>
<th>More Potential Technical Bottlenecks in Concurrent Scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Training</td>
<td>Batch task</td>
<td>Trained models, duration of the training process, scalability of the training mechanism if any</td>
<td>GPU memory capacity and latency/bandwidth, GPU compute capabilities and capacities</td>
<td>Platform ability to efficiently manage systems resources and schedule AI training workload (similar to HPC workload management tool benchmark)</td>
</tr>
<tr>
<td></td>
<td>GPU intensive workload</td>
<td>Price-Performance metrics: in regard to TCA (Total Cost of Acquisition) and or TCO (Total Cost of Ownership)</td>
<td>GPU-CPU and CPURAM communication characteristics could matter for large dataset training and/or Out-of-GPU-memory training</td>
<td>CPU-CPU communication characteristics could matter for large datasets</td>
</tr>
<tr>
<td></td>
<td>Minutes to days</td>
<td>For concurrent scenarios, the level of model training concurrency (similar to batch concurrency benchmarks)</td>
<td>Server-server communication characteristics could matter for intra-parallelism model training (Training a single model across multiple servers)</td>
<td>CPU-RAM communication characteristics could matter for Out-of-Core training</td>
</tr>
</tbody>
</table>
## AI Tasks: Hyper-Parameter Optimization

<table>
<thead>
<tr>
<th>AI Task</th>
<th>Workload Profile</th>
<th>Important Performance Indicators</th>
<th>Potential Technical Bottlenecks in Standalone Scenarios</th>
<th>More Potential Technical Bottlenecks in Concurrent Scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyper-parameter Optimization</td>
<td>Batches tasks managed by a workload orchestrator and hyper-parameters solver</td>
<td>Hyper-parameter combinatorial values to cover</td>
<td>Solver Algorithm limitations</td>
<td>All the Model Training potential bottlenecks apply here as well</td>
</tr>
<tr>
<td></td>
<td>GPU-tasks</td>
<td>Optimum value found for the model Accuracy</td>
<td>All the Model Training potential bottlenecks apply here</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Minutes to days</td>
<td>Overall duration to find the model with the best hyper-parameters</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## AI Tasks: inference run-time performance

<table>
<thead>
<tr>
<th>AI Task</th>
<th>Workload Profile</th>
<th>Important Performance Indicators</th>
<th>Potential Technical Bottlenecks in Standalone Scenarios</th>
<th>More Potential Technical Bottlenecks in Concurrent Scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deployed Model Inference Run-time</td>
<td>Online Service or library API</td>
<td>Latency of inference</td>
<td>GPU latency, GPU compute capabilities and capacities</td>
<td>Platform ability to efficiently manage the systems</td>
</tr>
<tr>
<td></td>
<td>Aiming for real-time request response-time in most cases. Milli-seconds to seconds.</td>
<td>Price performance metrics: in regard to TCA (Total Cost of Acquisition) and or TCO (Total Cost of Ownership)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mostly hardware AI accelerator intensive workload (GPU, FPGA, neuromorphic chip, Embedded Solutions)</td>
<td>For embedded and/or autonomous systems: Energy consumption/performance metric</td>
<td>GPU-CPU and CPU-RAM latency and throughput</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>For concurrent scenario, the level of model inference concurrency (similar to OLTP metrics)</td>
<td>Infrastructure network communication characteristics</td>
<td></td>
</tr>
</tbody>
</table>
The IBM open source way
https://developer.ibm.com/open/culture/

Training
Open Source @ IBM Program touches

78,000 IBMers annually

Recognition
We recognize our open source leaders with

300+ cash awards annually

Tooling
Our open source management tool suite is used over

30,000+ times per month

Organization
Our Open Source Core Team includes

~12 FTEs supporting all of IBM

Consuming
Virtually all of our products contain open source

3000+ packages reviewed every month

Contributing
We invest in community code & innovation

1500+ GitHub repos
Over time, IBM engaged as a key open source supporter, contributor as well as consumer.
FfDL – Fabric for Deep Learning

- FfDL uses REST APIs to access multiple deep learning libraries.
- Data Scientists or admins can deploy FfDL by launching a single command.
- Once deployed, there are four steps that data scientists perform to use FfDL:
  1. Prepare their deep learning model
  2. Upload the model and training data
  3. Start the training job and monitor its progress
  4. Download the trained model once the training job is complete

Get the code: https://github.com/ibm/ffdl
Try the Code Pattern on IBM Code
MAX - Model Asset Exchange

MAX is a one-stop exchange for data scientists and AI developers to consume models created using their favorite machine learning engines like TensorFlow, PyTorch, and Caffe2, and provides a standardized approach to classify, annotate, and deploy these models for prediction and inferencing.

Visit the Model Asset Exchange at: https://developer.ibm.com/code/exchanges/models/