Innovation Crossover Preliminary Research Report
IT/Cyber – Machine Learning/Artificial Intelligence

Context/Scope

This paper represents research conducted by OVO Innovation for the NSWC Crane Innovation Crossover event October 12-13, 2016. This research is intended to provide more insight into key challenges that were identified within the four technology clusters (Advanced Manufacturing, Cyber/IT, Life Sciences and DoD Technologies) first documented in the Battelle report. OVO consultants interviewed subject matter experts (SMEs) from the private sector, academia and the government identified by NSWC Crane to gather insights into key challenges in each cluster. This report is meant to inform the participants of the Innovation Crossover event and identify new research and new technologies that might address the key challenges.

This research was collected during August and September, 2016. The reports were submitted by OVO to NSWC Crane in late September 2016.

Introductory Narrative

The Innovation Crossover event, scheduled for 12-13 October 2016 in Bloomington is the culmination of months of planning and hard work. Some of this preparatory work involved the initial Battelle study which identified key technology clusters (Advanced Manufacturing, Life Sciences, Cyber/IT and DoD Technologies) in southern Indiana. From these clusters NSWC Crane and its contractor OVO Innovation conducted further, more detailed research, to examine detailed challenges and opportunities in each technology cluster. The reports attached document the research OVO conducted with subject matter experts identified by NSWC Crane in academia, industry and in the government. The reports are meant to document specific challenges within each technology cluster that could become areas of joint research and cooperation across the three constituents in southern Indiana. The reports are provided to you to help you prepare for your participation in the upcoming Innovation Crossover event and to frame both the challenges and active research underway to address these challenges.

Problem Statement

Machine Learning/Artificial Intelligence Challenge

Crane Problem Definition

- Context: A key challenge of moving deep learning forward is to make it more computationally feasible as the current state-of-the-art requires significant amount of training data and CPU cycles. These should also be discussed on where we draw the boundary on what AI should or should not be used in certain decisions. Further, applying new machine learning techniques/technologies to real-world applications where the system currently depends on continual operator input, database access, data downloads, etc.

Problem Context

Machine Learning/AI Challenges divide into three themes

- Inputs—the data collected and presented to a machine for learning
- Processing—challenges dealing with how the machine processes data to learn and produce useful results
- Outputs—confidence in the answers the machine produces

Distribution Statement A: Approved for Public Release

Problem Context

Input Challenges

- Dealing with multiplicity of data sources and data types in big data and machine learning
  - How do we integrate the diverse modalities such as numerical, categorical, text, transactional, audio, video, etc. when building models and mining patterns?
  - Where do we get the data? What open data is relevant?
  - There has been an explosion of large publically available data sets (e.g., 10 years ago overhead imagery was very expensive; today you can get it from Google Maps)—they need to be tuned to the needs of the machine that will be learning.

Problem Context

Input Challenges

- Having enough data collected and labeled so it can be applied by machine learning
  - Machine learning is trying to learn good from bad. This requires enough and known data to provide to the machine.
  - How do we take data and put it in a format that we can use?
  - Scale: Google, Facebook, Microsoft have billions of data points. However, if we have a highly dimensional space we actually have very sparse data. We need think of data at a different scale. Humans and can’t handle complex large data sets, so we ask machines to look for patterns. They need many more data points—and that data has to be understood to make sure we provide appropriate data for the machine to learn.

Problem Context

Input Challenges

- Generating data when we don’t have enough. Taking small data sets that we are confident about and adding data to them
  - Need to rethink the vastness of the data that is required. Humans can handle small problems. Machine learning is used to discover patterns incomprehensible to humans, which means we need much more data. How can we take small data sets we can understand and add data to them that we are confident will work? (See bootstrapping in processing)
  - To generate data, we sometimes use simulation. How can we create good enough simulators to create data?

Problem Context

Input Challenges

- Distributed data storage and I/O (this is an input and output challenge)
  - Data is stored across many machines in the cloud. Data is not in the same location, the hardware is different, different latencies, different storages and I/O. Each operation shuffling data in and out of memory takes 8x the operations step. Shuffling data in and out of disks will be a next big challenge.
  - Getting and storing data accounts for a huge portion of any computation effort and varies depending on the type of computation/domain.
  - We need to design adaptive, intelligent (learning) optimal data placement and retrieval strategies from distributed duplicated storage devices, especially hybrid architectures including conventional, flash, solid state, etc. drives in order to store and access massive amounts of data for machine learning.

Distribution Statement A: Approved for Public Release

Problem Context

Process Challenges

- Overview
  - The systems that we are engineering are so complex that it's becoming impossible to preengineer the systems. To sit down and define the algorithms is beyond the human engineering capabilities to program them. This is why we have to develop machines that learn. Humans were never programmed—we went to school. We weren't rewired, we were taught. The engineer is not programming a machine—he is teaching a machine. This is a major paradigm shift and presents new challenges for processing

Problem Context

Process Challenges

- Taking small certified data sets and combining with other data that is relevant to the problem
  - This relates to the input challenge of not having enough data—humans can handle small problems; machine learning is used to discover patterns incomprehensible to humans and require vast data.
  - How can we take small data sets about which we are confident and add data to it
  - This is sometimes called bootstrapping and is used when the machine learns from relevant data but not sensitive data (e.g., classified) that can later be applied.
  - How can we make sure that we can add new data in a meaningful way to data that already gets us close?

Problem Context

Process Challenges

- Learning in the field
  - Supervised learning in many domains is inadequate because we can’t generate sufficient training. The machine goes to school to learn on training data. Then it goes to the field, and there it doesn’t get new data, so it stops learning.
  - The challenge becomes how a machine in the field can learn on the job—while engaged in field.
  - This is sometimes called reinforcement learning.

Distribution Statement A: Approved for Public Release

Problem Context

Process Challenges

- Choosing the right machine learning algorithm
  - How do we keep track of this rapidly changing field? There are many machine learning algorithms and more being developed—how do we pick the appropriate one?
  - How do you explore the space and understand which machine learning would work best?
  - How as a community do you develop and track the rapidly state of the art?
  - How to keep up to date and not reinvent?
  - After selecting the algorithm, there are lots of different parameters that need to be set—how do you do that?

Distribution Statement A: Approved for Public Release

Problem Context

Process Challenges

- Learning from the operator.
  - One goal for machine learning is to process data automatically. However, there are many cases where the machine could learn from the operator if it could query the operator in a way that the operator could respond and help it learn.
  - How can we make the machine smart enough to have a symbiotic relationship with the operator and ask the operator—how do I do this, what do I do in this case? How can the machine prompt the operator?
- Reinforcement learning: merging sensing and control in complex physical systems
  - This is, in some senses, opposite of learning from the operator. How can we remove human biases about how to solve a problem? For example, in many robotic applications, we teach machines to find the edges before picking something up. But what if finding edges isn't important? How can we use sensing to help figure out to learn to control? How do we present the problem in a way that is appropriate?

Distribution Statement A: Approved for Public Release

Output Challenges

- Explainability
  - A huge challenge in machine learning using AI is explaining how the machine made the decision it made.
  - The machine might be very successful at addressing a problem, but it can't explain how it did it or explain a new decision point.
  - Most operators need a justification, not just a black box decision, especially if the situation is new.
  - Explanations help humans make better decisions. A recommendation needs to be scrutable—capable of being understood. Otherwise humans cannot tell if there is a flaw in the explanation.
  - How do we achieve greater explainability and predict the bounds under which the algorithm will work?
  - What performance can we sacrifice for confident answers?

Distribution Statement A: Approved for Public Release

Technologies

- Graphical Processing Units (GPUs), originally developed for accelerating gaming (and, as a result, became cost effective) are popular for machine learning.
- Manufacturers (e.g., Intel, NVIDIA), and companies (e.g., IBM, Facebook, Amazon) are investing in hardware to accelerate machine learning
- Most researchers do not build custom machines for machine learning
- Optimization in distributed data storage and I/O was identified as an input challenge

Relevance

- Machine learning is very important because it will allow machines to make suggestions for decisions in highly complex environments/problems. Decisions with explanations allow humans to make better decisions, which can extend the expertise of humans because we not only have a decision to something complex but we know why. This means we can also teach humans to make better decisions.
- Machine learning can help in almost any domain: from medical decisions to commercial (e.g., shopping, recommendations), machine learning can help sift through vast amounts of data to make relevant decisions.
- Who benefits? The human race through more lives saved, lower costs, greater efficiencies, etc.

Relevance

- DOD perspective 1: we cannot afford to put the same amount of man power as our adversaries are. A game changer is the combination of humans and machines—and machine learning is one of the biggest drivers. How do we make machines more intelligent that used to require human expertise? That allow us to do it better or faster than thousands of people? Doing so would make us safer and meet commitments with resources we can apply.
- DOD perspective 2: The only way to overcome anti-access/area-denial (A2AD) is through a highly integrated sensing and weapons approach—using many coordinated sensors and many coordinated weapons. We really can't engineer this—we need machine learning. So from a defense standpoint, if the US is maintain its peacekeeper role for open shipping of the seas, we have to develop machine learning to do it.

Distribution Statement A: Approved for Public Release

Scope

- For the purpose of this challenge, the technical scope was primarily the inputs (data), outputs (decisions with explanation), and processing of machine learning, not hardware
- Machine learning is applicable to a huge number of domains from DOD to medical to energy to commercial—anywhere where decisions must be made in complex environments, data, and/or problems.

Work/Research Underway

- General: Research underway parallels the challenge areas of input, processing, and output
  - How to get more and better data (e.g., data sets, labeling, open data initiatives)
  - How to deal with data to make it more useful (e.g., bootstrapping, reinforcement learning)
  - Changing the structure of deep learning models to allow explainability (including tradeoffs between accuracy and being able to explain)
  - Deep learning—improving and creating new algorithms
  - Human machine interactions (including Natural Language Processing)
  - Optimizing data storage and I/O

Distribution Statement A: Approved for Public Release

Work/Research Underway

- Academic and Consortium organizations identified by interviewed SMEs
  - Neural Information Processing Systems (NIPS) foundation provides a good survey of research. https://nips.cc/About
  - The International Conference on Machine Learning is a leading conference on machine learning. http://icml.cc
  - The iSchools organization is a consortium of Information Schools dedicated to advancing the information field—many universities doing machine learning research are members. http://ischools.org/
  - Knowledge Discovery & Web Mining Lab at University of Louisville. http://webmining.spd.louisville.edu/
  - Machine Learning Department at Carnegie Mellon University http://www.ml.cmu.edu/
  - Machine Learning at Berkeley https://ml.berkeley.edu/
  - Machine Learning at University of Washington https://www.cs.washington.edu/research/ml
  - Center for Machine Learning and Applications (CMLA) at The Pennsylvania State University. http://www.cse.psu.edu/research/cmla

Work/Research Underway

- Commercial Organizations identified by interviewed SMEs
  - Amazon
  - Google
  - Facebook
  - Microsoft
  - IBM Watson
  - Yahoo
- Funding Organizations identified by interviewed SMEs
  - National Science Foundation
  - Office of Naval Research
  - Department of Defense
  - Department of Energy
  - DARPA
    - On August 10, 2016 DARPA released the BAA for Explainable Artificial Intelligence Program
    - http://www.darpa.mil/program/explainable-artificial-intelligence

Summary

- Machine learning systems seek to understand vast amounts of data and provide decision making at a revolutionary level. Many people are already familiar with popular examples such as recommendation engines for movies, books and IBM's Watson. The ramifications for defense, medical, information, safety, commercial, and many other applications are astounding.
- Machine language has had a number of successes, yet faces many challenges. The Subject Matter Experts identified by NSWC Crane identified challenges in three major areas (see following slide)

Summary

- The Subject Matter Experts identified by NSWC Crane identified challenges in three major areas
  - Input Challenges
    - Dealing with multiplicity of data sources and data types in big data and machine learning
    - Having enough data collected and labeled so it can be applied by machine learning
    - Generating data when we don’t have enough. Taking small data sets that we are confident about and adding data to them
    - Distributed data storage and I/O
  - Process Challenges
    - Choosing the right machine learning algorithm
    - Taking small certified data sets and combining with other data that is relevant to the problem
    - Learning in the field
    - Learning from the operator
    - Reinforcement learning: merging sensing and control in complex physical systems
  - Output Challenges
    - Explainability: Explaining how the machine made the decision it made

Summary

- Addressing these challenges has the potential to
  - Significantly enhance US defense
  - Significantly advance decisions for medical, commercial, safety, and countless other domains
  - Advance the human race
- These challenges and the their potential solutions have been widely recognized and funded by governments and commercial companies and research, including at universities around NSWC Crane, abounds

Sources

Subject Matter Experts consulted / interviewed

- Dr. Robert Cruise, NSWC Crane
- Dr. Mark H. Linderman, AFRL
- Dr. Olfa Nasraoui, University of Louisville
- Dr. Lee Seversky, AFRL

Distribution Statement A: Approved for Public Release