Accelerated Deep Learning Discovery in Fusion Energy Science

William M. Tang
Princeton University/Princeton Plasma Physics Laboratory (PPPL)

NVIDIA GPU TECHNOLOGY CONFERENCE
GTC-2018

San Jose, CA

March 19, 2018

CNN’s “MOONSHOTS for 21st CENTURY” (Hosted by Fareed Zakaria)

— Five segments (Spring, 2015) exploring “exciting futuristic endeavors in science & technology in 21st Century”

(1) Human Mission to Mars
(2) 3D Printing of a Human Heart
(3) Creating a Star on Earth: Quest for Fusion Energy
(4) Hypersonic Aviation
(5) Mapping the Human Brain

“Creating a Star on Earth” → “takes a fascinating look at how harnessing the energy of nuclear fusion reactions may create a virtually limitless energy source.”

Stephen Hawking: (BBC Interview, 18 Nov. 2016)

“I would like nuclear fusion to become a practical power source. It would provide an inexhaustible supply of energy, without pollution or global warming.”
APPLICATION FOCUS FOR DEEP LEARNING STUDIES: FUSION ENERGY SCIENCE

Most Critical Problem for Fusion Energy ➔

*Accurately predict and mitigate large-scale major disruptions in magnetically-confined thermonuclear plasmas such as the ITER – the $25B international burning plasma “tokamak”*

- **Most Effective Approach**: Use of big-data-driven statistical/machine-learning predictions for the occurrence of disruptions in world-leading facilities such as EUROFUSION “Joint European Torus (JET)” in UK, DIII-D (US), and other tokomaks worldwide.

- **Recent Status**: 8 years of R&D results (led by JET) using Support Vector Machine Machine Learning on zero-D time trace data executed on CPU clusters yielding success rates in mid-80% range for JET 30 ms before disruptions, 
  **BUT > 95% accuracy with false alarm rate < 5% at least 30 milliseconds before actually needed for ITER!** Reference – P. DeVries, et al. (2015)
CURRENT CHALLENGES FOR DEEP LEARNING/AI STUDIES:

• Disruption Prediction & Avoidance Goals include:

(i) improve **physics fidelity** via development of new **ML multi-D, time-dependent software including improved classifiers**;

(ii) develop **“portable”** (cross-machine) predictive software beyond JET to other devices and eventually ITER; and

(iii) enhance **accuracy & speed** of disruption analysis for very large datasets via HPC

→ TECHNICAL FOCUS: development & deployment of advanced Machine Learning Software via Deep Learning/AI Neural Networks

• **both Convolutional & Recurrent Neural Nets** included in Princeton’s “Fusion Recurrent Neural Net (FRNN) Software

• Julian Kates-Harbeck (chief architect)
CLASSIFICATION

● Binary Classification Problem:
  ○ *Shots are Disruptive or Non-Disruptive*

● *Supervised* ML techniques:
  ○ *Domain fusion physicists combine knowledge base of observationally validated information with advanced statistical/Machine Learning predictive methods.*

● **Machine Learning Methods Engaged:**
  Shallow Learning “SVM” approach initiated by JET team with “APODIS” software has led now to Princeton’s **New Deep Learning Fusion Recurrent Neural Net (FRNN) code including both Convolutional & Recurrent NN)**

● **Challenge:**
  → Multi-D data analysis requires **new signal representations**;
  → FRNN software’s Convolutional Neural Nets (CNN) **enables – for first time** – capability to deal with dimensional (beyond Zero-D) data

- **14 Feature vectors** are extracted from raw time series data
  - 7 signals* (O7) x 2 representations+

*S Signals: (“ZERO-D Time Traces”)
1. Plasma current [A]
2. Mode lock amplitude [T]
3. Plasma density [m⁻³]
4. Radiated power [W]
5. Total input power [W]
6. d/dt Stored Diamagnetic Energy [W]
7. Plasma Internal Inductance

+Representations:
1. Mean
2. Standard deviation of positive FFT spectrum (excluding first component)

Feature vectors are remapped to higher-D space → “hyper-plane” maximizing distance between classes of points
**APODIS ("Advanced Predictor of Disruptions"): Multi-tiered SVM Code**

→ *separate SVM models trained for separate consecutive time intervals preceding disruption*


**Incoming real-time data**

<table>
<thead>
<tr>
<th>M3 (SVM)</th>
<th>X(t-64)</th>
<th>D(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M2 (SVM)</td>
<td>X(t-32)</td>
<td>D(2)</td>
</tr>
<tr>
<td>M1 (SVM)</td>
<td>X(t)</td>
<td>D(1)</td>
</tr>
</tbody>
</table>

**Decision Function (SVM classifier)**

| Prediction: -1 (no alarm) or 1 (alarm) |

**BUT – UNABLE TO DEAL WITH 1D PROFILE SIGNALS!**
Background/Approach for DL/AI

• **Deep Learning Method**: distributed data-parallel approach to train deep neural networks → *Python Framework using high-level Keras library with Google Tensorflow backend*

Reference: Deep Learning with Python, François Chollet (Nov. 2017, 384 pages)

*** Major contrast with “Shallow Learning” approaches including SVM’s, Random Forests, Single Layer Neural Nets, & modern Stochastic Gradient Boosting (“XG-BOOST”) methods by enabling moving from ML software deployment on clusters to supercomputers:
→ *Titan (ORNL), Summit (ORNL); Tsubame-3 (TiTech); Piz Daint (CSCS); .. Also other architectures, e.g. – Intel Systems: KNL currently + promising new future designs*

-- **stochastic gradient descent (SGD)** used for large-scale (i.e., optimization on supercomputers) with parallelization via mini-batch training to reduce communication costs

-- **DL Supercomputer Challenge**: need large-scale scaling studies to examine if convergence rate saturates with increasing mini-batch size (to thousands of GPU’s)
Machine Learning Workflow

- Identify Signals
  - Classifiers

- Preprocessing and feature extraction

- Normalization
  - All data placed on appropriate numerical scale $\sim O(1)$
  - e.g., Data-based with all signals divided by their standard deviation

- Train model, Hyper parameter tuning
  - All available data analyzed;
  - Train LSTM (Long Short Term Memory Network) iteratively;
  - Evaluate using ROC (Receiver Operating Characteristics) and cross-validation loss for every epoch (equivalent of entire data set for each iteration)

- Use model for prediction
  - Apply ML/DL software on new data

Princeton/PPPL DL predictions now advancing to multi-D time trace signals (beyond zero-D)

Measured sequential data arranged in patches of equal length for training
### JET Disruption Data

<table>
<thead>
<tr>
<th># Shots</th>
<th>Disruptive</th>
<th>Nondisruptive</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbon Wall</td>
<td>324</td>
<td>4029</td>
<td>4353</td>
</tr>
<tr>
<td>Beryllium Wall (ILW)</td>
<td>185</td>
<td>1036</td>
<td>1221</td>
</tr>
<tr>
<td>Totals</td>
<td>509</td>
<td>5065</td>
<td>5574</td>
</tr>
</tbody>
</table>

**JET produces ~ Terabyte (TB) of data per day**

**JET studies → 7 Signals of zero-D (scalar) time traces, including**

<table>
<thead>
<tr>
<th></th>
<th>Data Size (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plasma Current</td>
<td>1.8</td>
</tr>
<tr>
<td>Mode Lock Amplitude</td>
<td>1.8</td>
</tr>
<tr>
<td>Plasma Density</td>
<td>7.8</td>
</tr>
<tr>
<td>Radiated Power</td>
<td>30.0</td>
</tr>
<tr>
<td>Total Input Power</td>
<td>3.0</td>
</tr>
<tr>
<td>d/dt Stored Diamagnetic Energy</td>
<td>2.9</td>
</tr>
<tr>
<td>Plasma Internal Inductance</td>
<td>3.0</td>
</tr>
</tbody>
</table>

**→ Well over 350 TB total amount with multi-dimensional data just recently being analyzed**

**~55 GB data collected from each JET shot**
Deep Recurrent Neural Networks (RNNs): Basic Description

- “Deep”
  - Learn salient representation of complex, higher dimensional data

- “Recurrent”
  - Output $h(t)$ depends on input $x(t)$ & internal state $s(t-1)$

*Internal State ("memory/context")*
Deep Learning/AI FRNN Software Schematic

FRNN Architecture:
- LSTM
- 3 layers
- 300 cells per layer

Signals:
- Plasma Current
- Locked Mode Amplitude
- Plasma Density
- Internal Inductance
- Input Power
- Radiated Power
- Internal Energy
- 1D profiles (electron temperature, density)
- ...

Output: Disruption coming?

$T = 0$ [ms]

$T = 1$

$T = t$
FRNN Code PERFORMANCE: ROC CURVES

JET ITER-like Wall Cases @30ms before Disruption

Performance Tradeoff: Tune True Positives (good: correctly caught disruption) vs. False Positives (bad: safe shot incorrectly labeled disruptive).

Data (~50 GB), 0D signals:
- Training: on 4100 shots from JET C-Wall campaigns
- Testing 1200 shots from Jet ILW campaigns
- All shots used, no signal filtering or removal of shots
RNNs: HPC Innovations Engaged

**GPU training**
- Neural networks use dense tensor manipulations, efficient use of GPU FLOPS
- Over 10x speedup better than multicore node training (CPU’s)

**Distributed Training via MPI**
Linear scaling:
- Key benchmark of “time to accuracy”: we can train a **model that achieves the same results nearly N times faster with N GPUs**
Scalable
- to 100s or >1000’s of GPU’s on Leadership Class Facilities
- TB’s of data and more
- Example: Best model training time on full dataset (~40GB, 4500 shots) of 0D signals training
  - SVM (JET) : > 24hrs
  - RNN ( 20 GPU’s) : ~40min
Scaling Summary

Communication: each batch of data requires time for synchronization

\[ T_{\text{sync}} \sim \log (N_{\text{workers}}) \]

Runtime: computation time

\[ T \sim \frac{1}{N} \left( A + B \log(N) \right) = O \left( \frac{\log(N)}{N} \right) \]

Parallel Efficiency

\[ \text{Parallel Efficiency} \sim \frac{A + B}{A + B \log(N)} = o \left( \frac{1}{\log(N)} \right) \]
FRNN Scaling Results on GPU’s

- Tests on OLCF Titan CRAY supercomputer
  - **OLCF DD AWARD**: Enabled Scaling Studies on Titan currently up to 6000 GPU’s
  - Total ~ 18.7K Tesla K20X Kepler GPUs

Tensorflow+MPI
NEW INSIGHTS/FINDINGS

- **Deep learning** performs (vs. “shallow learning) very well in raw performance and *more suited to generalizing*
- Deep RNN can demonstrably use 1D profiles effectively.

Methods to Further Improve Cross Machine Portability:

- **Hyperparameter Tuning** -- leveraging continuing HPC performance enhancements

- Obtain more & better-validated data – *properly prepared for analysis*
- Apply better Normalization (physics-based) (e.g., “Greenwald limit” for density normalization)
- Data cleaning (*eliminate non-physical inputs* – e.g., “negative density”) 
- Continue transfer learning → Train on one machine and fine tune with a few shots from other machine [e.g., *train on DIII-D (US) and apply to JET (Europe)*] → *very important for ITER!!*
Cross Machine Prediction (DIII-D to JET)  
(J. Kates-Harbeck)

Train (DIII-D)  

Test (JET)  

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN 0D &amp; RNN 1D</td>
<td>~0.80</td>
</tr>
<tr>
<td>XGBoost (shallow)</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Deep Learning (particularly with profiles) gives encouraging results!
Hyper-parameter Tuning enabled by HPC

• **Example** → random grid of 100 iterations with 100 GPUs per each trial
  -- Trials run asynchronously to convergence
  -- Distributed training performed with *data-parallel synchronous “Stochastic Gradient Descent (SGD)”* – standard approach in DL applications
  – Master loop determines the best set of parameters based on the validation level

• **Exciting New Trends Emerging** → aggressive large-scale hyper-parameter tuning trials carried out on the “Titan” exhibit very promising potential for shifting the minimum warning time before disruptions to 50 ms or even up to 100 ms and above.
  → *Strongly motivate new HPC-enabled studies enabled by deployment of new half-precision version FRNN on powerful NVIDIA Volta GPU’s on SUMMIT at ORNL*

**Significance:** Key to enabling future risk mitigation for ITER via achieving increased pre-disruption warning time
FRNN DL/AI software reliably scales to 1K P-100 GPU’s on TSUBAME 3.0
→ associated production runs contribute strongly to Hyper-parameter-Tuning-enabled physics advances !

Recent results: TSUBAME 3.0 supercomputer (TiTech, Tokyo, Japan)

Tsubame 3.0 “Grand Challenge Runs” (A. Svyatkovskii, Princeton U)
– More than 1K Tesla P100 SXM2 GPUs, 4 GPUs per node, Nvlink communication
– Tensorflow+MPI, CUDA8, CuDNN 6

![Graph showing scaling model and ideal scaling]
Fusion Energy Mission:
-- Accelerate demonstration of the scientific & technical feasibility of delivering Fusion Power
-- Most critical associated problem is to avoid/mitigate large-scale major disruptions.

ML Relevance to HPC:
-- Rapid Advances on development of predictive methods via large-data-driven “machine-learning” statistical methods
-- Approach Focus: Deep Learning/Al Neural Nets (Convolutional & Recurrent)
-- Significance: Exciting predictive approach beyond previous “hypothesis-driven/first principles” exascale methods alone

Convergence/Complementarity of Exascale HPC and Big-Data ML/Al Methods:

-- Domain Physics Exascale HPC needed to introduce & establish improved Supervised ML/Al Classifiers!
DL/AI Vision Summary in Moving from Prediction to Control

ZERO-D to HIGHER-D SIGNALS via CONVOLUTIONAL NEURAL NETS (CNN)

• Enables immediate learning of generalizable features (→ helps enable cross-tokamak portability of DL/AI software)

• Reinforcement learning enables transition from PREDICTION to CONTROL!

• Takes advantage of increasingly powerful world class HPC (supercomputing) facilities!
EARLY RESULTS from APPLICATION of FRNN to POWERFUL NVIDIA VOLTA’S on SUMMIT at ORNL

• **Initial Scalability Studies (A. Svyatkovskii, Princeton U)**
  – Strong linear runtime scaling and logarithmic communication time
  – Up to 6000 K20 GPU on Titan (CUDA 7.5, CuDNN 5.1)
  – Up to 192 V100 GPUs on Summit (CUDA 9.1, CuDNN 7.0.5)

• FRNN architecture includes LSTM, CNN and fully connected Layers
  • 2x speedup over P100
  • 8x speedup over K20

![Graph showing performance scaling with number of GPUs]
SUCCES STORY | PRINCETON UNIVERSITY: ITER FUSION ENERGY

SPEEDING THE PATH TO FUSION ENERGY WITH DEEP LEARNING
RAPID GROWTH & BROAD INVESTMENTS IN MACHINE LEARNING/DL/AI TRENDS FOR FUTURE

Business World ➔ Reference: Amazon article:
Reformation of Amazon and other top businesses incorporating ML/DL/AI:
https://www.wired.com/story/amazon-artificial-intelligence-flywheel/

Cancer Research ➔ Reference: “Candle Project” with ECP
Exascale Computing Project (DOE & NIH) to identify optimal cancer treatment strategies, by building a scalable deep neural network code called the CANcer Distributed Learning Environment (CANDLE).
→ development of predictive models for drug response, and automation of the analysis of information from millions of cancer patient records -- via developing, implementing, & testing DL/AI Algorithms and their benchmarks

• Key Applications Areas like Fusion Energy & others should enhance efforts to leverage cross-disciplinary connections to enormous worldwide investments in ML/DL/AI R&D !