HSF CWP ML

HEP Software Foundation: Community White Paper Machine Learning

Paul Seyfert

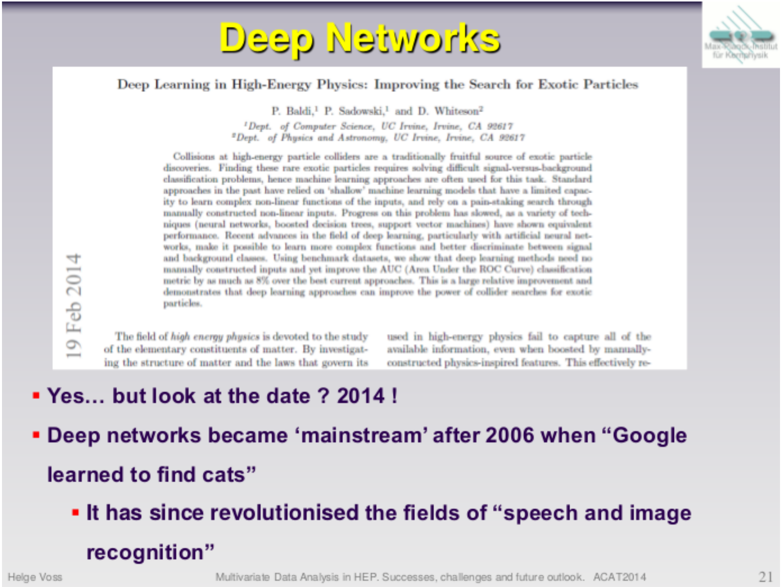
CERN

15th May 2017

timeline / facts

- first inputs collected in October 2016
- writing sessions at
  - HSF workshop, Jan 2017
  - IML workshop, Mar 2017
  - DS@HEP, May 2017
  - editing session for people “based in Europe”, this Friday
- $\mathcal{O}(103)$ “authors” accumulated
- ≤ 30 pages of text/items by now
- google docs link

we need to catch up



Deep Networks

Max-Planck-Institut für Kernphysik

Deep Learning in High-Energy Physics: Improving the Search for Exotic Particles

P. Baldi,[1] P. Sadowski,[1] and D. Whiteson[2]

[1]*Dept. of Computer Science, UC Irvine, Irvine, CA 92617*
[2]*Dept. of Physics and Astronomy, UC Irvine, Irvine, CA 92617*

Collisions at high-energy particle colliders are a traditionally fruitful source of exotic particle discoveries. Finding these rare exotic particles requires solving difficult signal-versus-background classification problems, hence machine learning approaches are often used for this task. Standard approaches in the past have relied on 'shallow' machine learning models that have a limited capacity to learn complex non-linear functions of the inputs, and rely on a pain-staking search through manually constructed non-linear inputs. Progress on this problem has slowed, as a variety of techniques (neural networks, boosted decision trees, support vector machines) have shown equivalent performance. Recent advances in the field of deep learning, particularly with artificial neural networks, make it possible to learn more complex functions and better discriminate between signal and background classes. Using benchmark datasets, we show that deep learning methods need no manually constructed inputs and yet improve the AUC (Area Under the ROC Curve) classification metric by as much as 8% over the best current approaches. This is a large relative improvement and demonstrates that deep learning approaches can improve the power of collider searches for exotic particles.

19 Feb 2014

The field of *high energy physics* is devoted to the study of the elementary constituents of matter. By investigating the structure of matter and the laws that govern its

used in high-energy physics fail to capture all of the available information, even when boosted by manually-constructed physics-inspired features. This effectively re-

- Yes… but look at the date ? 2014 !
- Deep networks became 'mainstream' after 2006 when "Google learned to find cats"
  - It has since revolutionised the fields of "speech and image recognition"

Helge Voss Multivariate Data Analysis in HEP. Successes, challenges and future outlook. ACAT2014 21

what do we need ML for

- obvious points: more efficient event selection equivalent to more luminosity (time)
- or less events to process (money)
- or better discovery potential (first single experiment with N σ)
- do something with ML that human written algorithms can't do (we lost checkers, chess, go, ...will track finding be next)

TMVA or what?

- scikit learn: 500k hits on google
- tmva: 80k
- Machine learning researchers publish their code along with papers (but for keras or tensorflow or caffe, rarely for TMVA)

→ GANs for simulation, RNNs for jet tagging all implemented in keras (afaik)

what do we need ML for

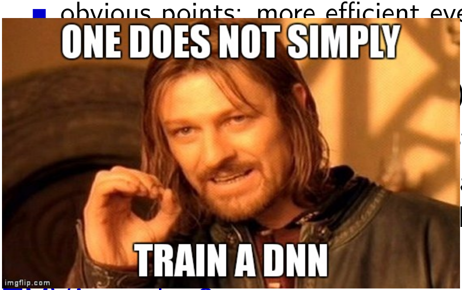
- obvious points: more efficient event selection equivalent to more

)

t single experiment with N σ)

an written algorithms can't do

ll track finding be next)



TMVA or what?

- scikit learn: 500k hits on google
- tmva: 80k
- Machine learning researchers publish their code along with papers (but for keras or tensorflow or caffe, rarely for TMVA)

→ GANs for simulation, RNNs for jet tagging all implemented in keras (afaik)

state of the paper

+ very diverse inputs
  - I/O issues (large ntuples which don't fit into RAM)
  - find new use cases for ML
  - train analysts to use state-of-the-art ML
  - facilitate usage of outside world tools
  - concerns about lifespan of tools
  - development of benchmarks
  - collaborations with industry / data science

state of the paper

+ very diverse inputs

- I/O issues (large ntuples which don't fit into RAM)
- find new use cases for ML
- train analysts to use state-of-the-art ML
- facilitate usage of outside world tools
- concerns about lifespan of tools
- development of benchmarks
- collaborations with industry / data science

− very incoherent inputs

- make TMVA a state-of-the-art tool vs. just use keras/scikit-learn
- learn outside world ML vs. work on HEP specific aspects of ML
- proliferation of non-HEP tools vs. problems non-HEP tools cannot handle

state of the paper

+ very diverse inputs
  - I/O issues (large ntuples which don't fit into RAM)
  - find new use cases for ML
  - train analysts to use state-of-the-art ML
  - facilitate usage of outside world tools
  - concerns about lifespan of tools
  - development of benchmarks
  - collaborations with industry / data science

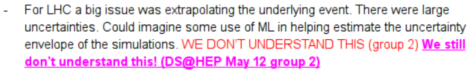
– very incoherent inputs

  - make TMVA a state-of-the-art tool vs. just use keras/scikit-learn
  - learn outside world ML vs. work on HEP specific aspects of ML
  - proliferation of non-HEP tools vs. problems non-HEP tools cannot handle

– pretty much unsorted

– no replies

– e.g. safety for real time application a bit hidden



- For LHC a big issue was extrapolating the underlying event. There were large uncertainties. Could imagine some use of ML in helping estimate the uncertainty envelope of the simulations. WE DON'T UNDERSTAND THIS (group 2) **We still don't understand this! (DS@HEP May 12 group 2)**

subsections

- External and internal ML Tools
- Applications of ML and R&D
- Bridges to other communities
- Resources and related challenges
- Training the community in ML

What is needed (in general)?

- check compatibility with other papers
- sorting
- turn notes into text
- **editorial work**
    - make the document look like the authors know what they want to say

my impression

Get rid of technical discussions if we need a “middleware” or “interface” or scripts to convert root ntuples with arrays to root ntuples with floats.

What should we do?

- I don't think we want to do the editing for the entire community now and here
- check if we agree on everything
- check if our points of interest are understandably written (spell them out)