Deep Learning Development and Deployment for Low-Latency Gravitational-Wave Astronomy

University of Minnesota

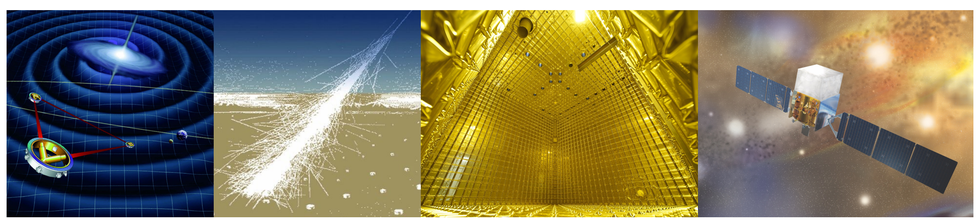
Will Benoit, University of Minnesota

a3d3 Accelerated AI Algorithms for Data-Driven Discovery

NSF

Scientific Motivation

- **Gravitational waves** – ripples in spacetime – contain information about important cosmological/astronomical questions: neutron star EoS, Hubble constant, the very early universe, etc.[7,10,6]
- But why look at only one stream of information when there are multiple – **Multi-Messenger Astrophysics!**
- Gravitational wave detectors can act as triggers for other types of observatories
- More gravitational wave detectors, greater sensitivity, and a larger parameter space,[5,6] coupled with a need for low latency, point us towards **Machine Learning** solutions



Representations of the 4 extrasolar messengers. From left to right: gravitational waves, cosmic rays, neutrinos (DUNE), gamma rays (Fermi telescope)[4,11,3,9]

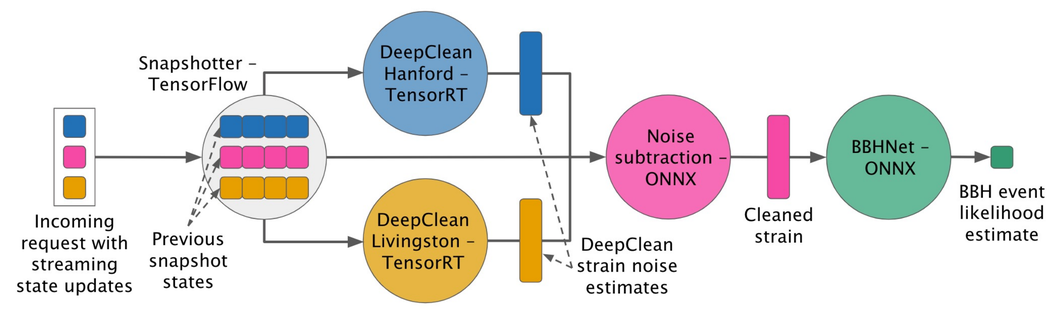
Cleaning and Detection Pipeline

DeepClean

- 1-D convolutional autoencoder
- Uses system and environmental monitoring channels to **predict and subtract noise** from strain data[8]
- Able to simultaneously subtract linear, **non-linear, and non-stationary** noise[8]
- Fully online implementation planned for O4 data collection

BBHNet

- 1-D convolutional network
- Performs binary classification of noise/binary black hole (BBH) mergers on DeepClean output
- Able to differentiate between **true signals, glitches, and background** noise in a low-latency manner[1]
- Future: online implementation, inclusion of neutron star mergers



Flow of data through the DeepClean/BBHNet pipeline in the online streaming configuration

Machine Learning for Detection of Gravitational-Waves Enables Multi-Messenger Follow-Up

Machine Learning Operations and Best Practices

- Tools need to be used to be effective – not uncommon for good software to languish due to poor adoption or implementation
- **Physicists shouldn't need to become ML experts to do analysis**
- Development is easier when collaborators can understand each others' work
- **Use established industry practices from software development and MLOps**

Organizational Practices[2]

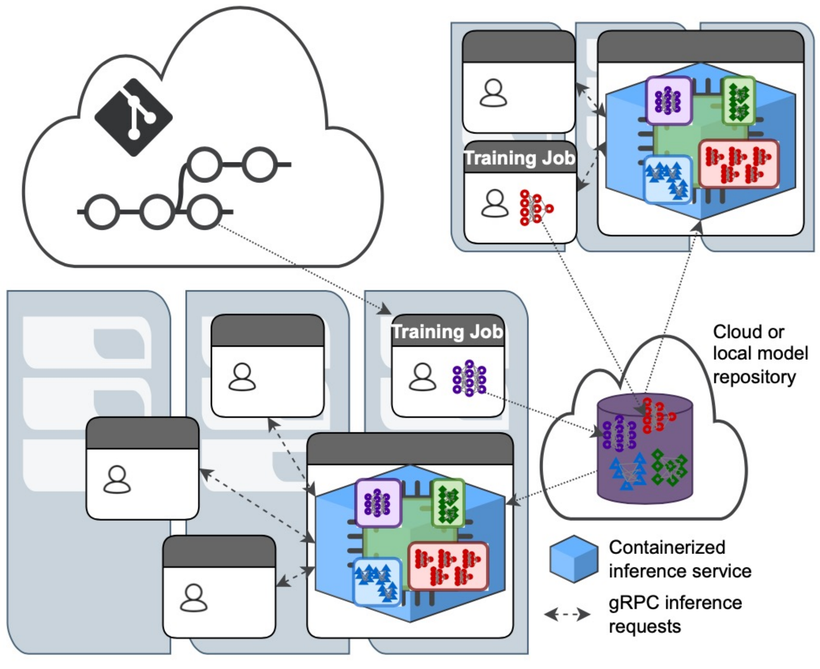
- Deliberately designed structure for code and repository: **monorepo**
- CI/CD pipelines version experiments and automate model deployment, which ensures **repeatable, meaningful results**

Inference-as-a-Service[1,2]

- Models are hosted in a centralized server, inference application loads them for user requests
- **Keeps all users in sync with latest version, and abstracts away the overhead**
- Makes efficient use of available resources – can optimize hardware for inference

Hermes Libraries[2]

- Developed to address the difficulties that come with using an inference server
- Handles model export and acceleration, asynchronous data processing and inference request generation, and cloud-based resource allocation



Inference-as-a-Service model. Users are kept in sync with the most up-to-date model, and the inference server handles job coordination and resource allocation[2]

References

[1] A. Gunny et al., Nature Astronomy **6** (2022). [2] A. Gunny et al., Proceedings of the 12th Workshop on AI and Scientific Computing at Scale using Flexible Computing Infrastructures (2022). [3] Brice, M. (2017). *Inside ProtoDune*. CERN. photograph. [4] *Searching for gravitational waves with LISA*. ESA. (2002, September 11). [5] LIGO Scientific Collaboration, Classical and Quantum Gravity 32 (2015). [6] The LIGO Scientific Collaboration & The Virgo Collaboration, Nature **460** (2009). [7] M. W. Coughlin et al., Monthly Proceedings of the Royal Astronomical Society **480** (2018). [8] R. Ormiston et al., Physical Review Research **2** (2020). [9] Simonnet, A. (n.d.). *Glast/Fermi Litho*. Fermi Gamma-ray Space Telescope. NASA E/PO. [10] T. Dietrich et al., Science **370** (2020). [11] Yumpu. (2015, November 7). Ultra-high energy cosmic rays and the Pierre Auger Observatory.

Acknowledgements

The author acknowledges support from the National Science Foundation with grant numbers OAC-1931469, OAC-1934700, PHY-2010970, OAC-2117997, and DGE-1922512. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility funded by the National Science Foundation. The author is grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants PHY-0757058 and PHY-0823459.