Preparing *DeepClean* for online noise regression at LIGO Hanford and Livingston

Saleem Muhammed, Alec Gunny, Michael Coughlin, Erik Katsavounidis

April 28, 2022
Overview of deepClean infrastructure

The deep learning infrastructure for noise reduction (DeepClean) was presented by Ormiston et al. 2020

- Generic enough to regress linear, non-linear and non-stationary couplings
- Performance assessed with LIGO O2 and O3 data.
  - Subtraction of O2 Jitter noise in agreement with results from Weiner filter.
  - Subtraction of O3 60 Hz sidebands in agreement with results from Nonsense algorithm
- Robust and safe
  - Assessed through parameter estimation of injected CBC signals.
DeepClean at low latency (a.k.a DeepClean-online)

- Ability to load, clean and write the 1s-long frames as soon as they are available, at an additional latency of ~1 second
- Use pre-trained models of the neural network (updated only once a day or so)
- The algorithm to be different from the DeepClean-offline, as the latter is designed to perform noise regression on long stretches of data at once (typically ~1 hour)
Challenge: noise predictions on short-duration data

- At the edges of the kernels, the noise prediction differs from offline predictions, for about ~0.75 s.
- That means, at time t, a reliable prediction is available till t-0.75s.
- DeepClean latency = 0.75 + inference time (0.75s here is usually referred to as aggregation latency)
Latency vs quality for 60Hz noise subtraction

- Performed on a pseudo replay of O3 data.
- The subtraction improves with increasing aggregation latency
Safety of the underlying signals

Recovery of chirp mass shown from a distribution of injected BBH signals
Subtraction of multiple noise couplings

- Along with 60 Hz noise, we also plan to implement the low latency subtraction of broadband low-frequency noise.
- Multiple noise subtraction, if performed sequentially, will lead to increase the inference latency linearly.
- For disjoint frequency bands, we expect the noise predictions to work in parallel.
Accelerated inference

- Use of Hardware accelerators, e.g., GPUs and maximize the benefits with the incorporation of computing paradigm such as Inference-as-a-Service [Gunny et al 2022]

- The newly developed HERMES toolkit can speed up the inferences by order of magnitudes compared to the inferences traditionally-performed on LDG resources.

- Concurrent execution of requests for inference, data loading and pre-processing can significantly save the latency, especially when there are multiple noises to subtract.
Plan for running on O4 MDC

- Low latency DeepClean infrastructure is ready for production-level analysis
- To be demonstrated on the ongoing O4 MDC replays
- Hardware (GPU) allocated at LDG clusters, currently waiting to get the detchar frames available along with h(t) frames of MDC.
Extra Slides
Sliding window model for online DeepClean
Performance vs aggregation latency

*aggregation latency* - the amount of data (seconds) to be excluded from the end of the segment due to quality degradation.
Low-Latency and Offline Deployment

- Local deployments at detector sites for low-latency inference on single T4 GPUs or FPGAs
  - Cleaned frames written to shared memory at LDAS for use by downstream algorithms
- Google Cloud deployment for offline analysis and validation
- All deployments synced to centralized repository for model retraining and updating
Time taken for training 600s data with GPU

It takes an average of ~8 minutes