IML Workshop

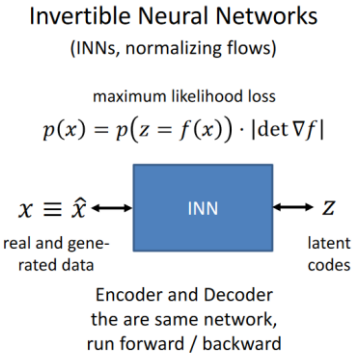
Solving Inverse Problems with Invertible Neural Networks
Distributed training

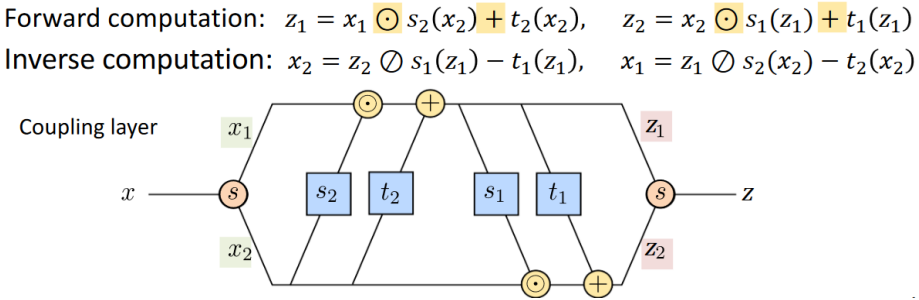
Renato Cardoso IT-DI-OPL

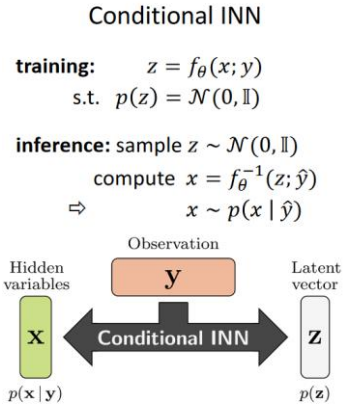
Solving Inverse Problems with Invertible Neural Networks

Inverse problem goal: given observations ŷ, determine underlying hidden parameters X̂

Use Invertible Neural Networks to interpret the results obtained by inferring the hidden parameters.





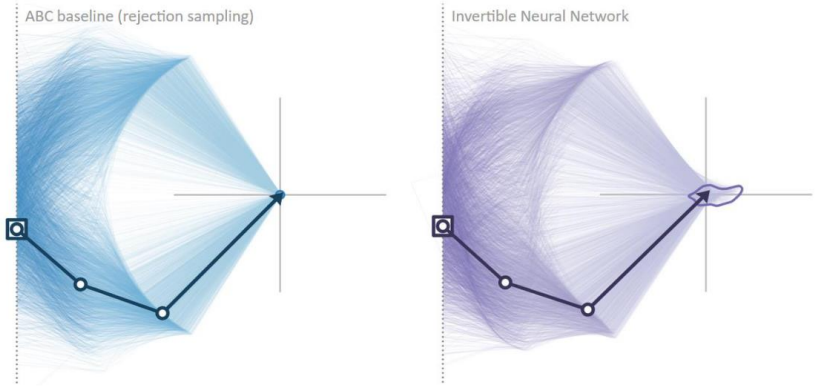


Solving Inverse Problems with Invertible Neural Networks - Example

Forward problem

- robot arm with 4 DOF: x=[x1,...,x4]
  - x1: vertical position of first joint
  - x2,x3,x4: joint angles
  - Gaussian priors: x1 prefers the center x2,x3,x4 prefer to be straight
- observation: hand position y=[y1,y2]
- geometric arm simulation y=$g(x)$ implicitly defines likelihood $p(y|x)$

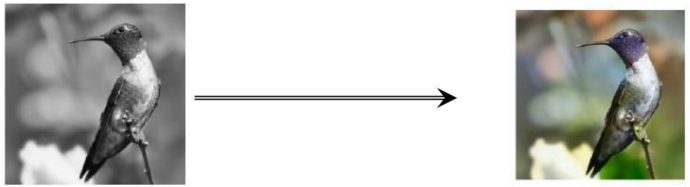
INN infers posterior p(x|ŷ) for given hand position ŷ



Solving Inverse Problems with Invertible Neural Networks

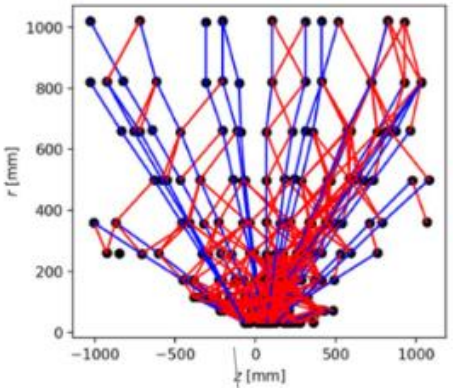
INNs are very good density estimators:

- Not yet quite as good as GANs
- Much stronger mathematical interpretation and guarantees
- Successful applications in particle and astrophysics, epidemiology, medicine, psycholog



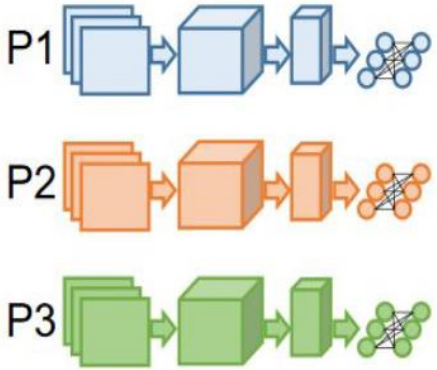
Distributed training of GNN on HPC

- Graph is irregular, sparse and dynamic in its size.
- To cope with grown graph size, we must explore different computing resources.
- Currently AI hardwares primarily focus on accelerating dense 1D or 2D arrays, to some extend neglecting sparse and irregular tensor calculations



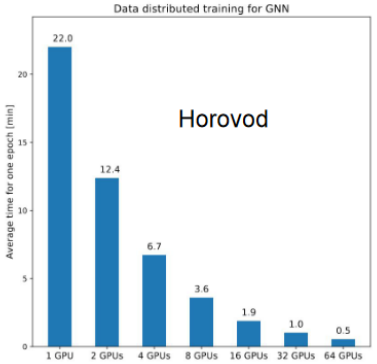
Distributed training of GNN on HPC

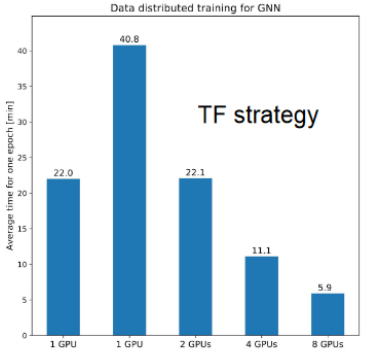
- Replication of the model in different devices
- Distribute different data for each devices
- Gradients are averaged among devices
- Update the weights

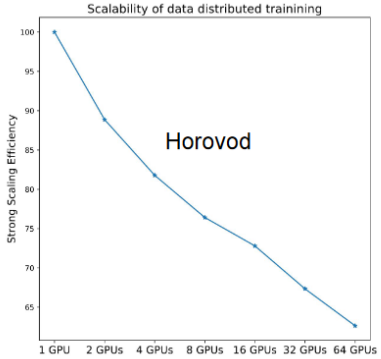


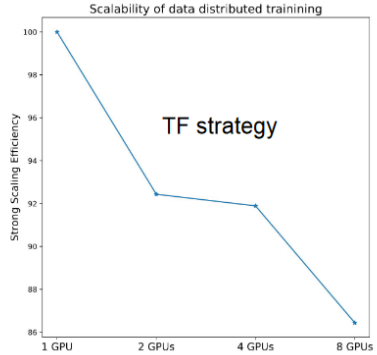
Distributed training of GNN on HPC

- Horovod - Framework for TensorFlow, Keras, PyTorch, and Apache MXNet.
- Uses MPI model to easily and efficiently parallelize the model.
- Horovod shows a stronger Scaling Efficiency than Tensorflow Strategy for Graph Neural Networks using GPUs.



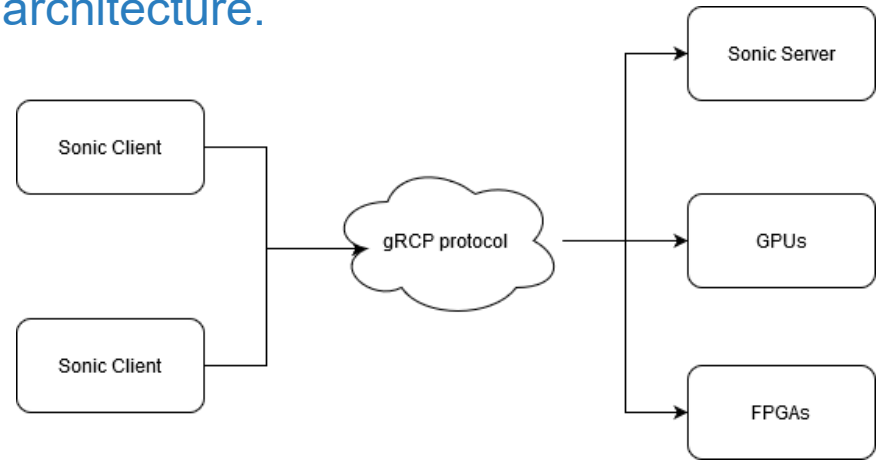






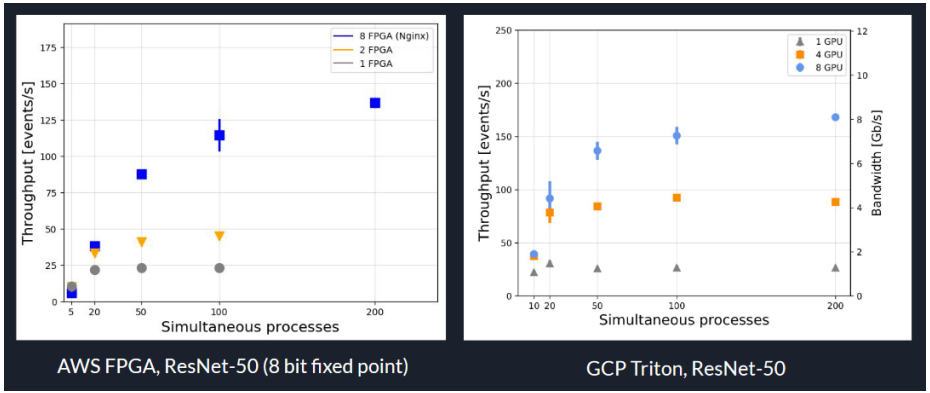
GPU and FPGA as a Service for Machine Learning Inference Accelerations

- FPGAs - Field-programmable gate array, is an integrated circuit designed to be configured by a customer or a designer after manufacturing
- Use GPUs and FPGAs as a better alternative for CPUs
- Necessity of using an heterogeneous computing, using more than one computing architecture.



GPU and FPGA as a Service for Machine Learning Inference Accelerations

- Their low power consumption and extremely fast processing make them particularly suited for applications across industry and high energy physics.
- FPGAs offer good scalability, however not as good as GPUs but with lower power consumption.
- Compared to ASICs, FPGAs are still less power and more power consuming, however they can be utilized in a wide array of domains.



Thank you for your attention!