Deep Learning with GPUs

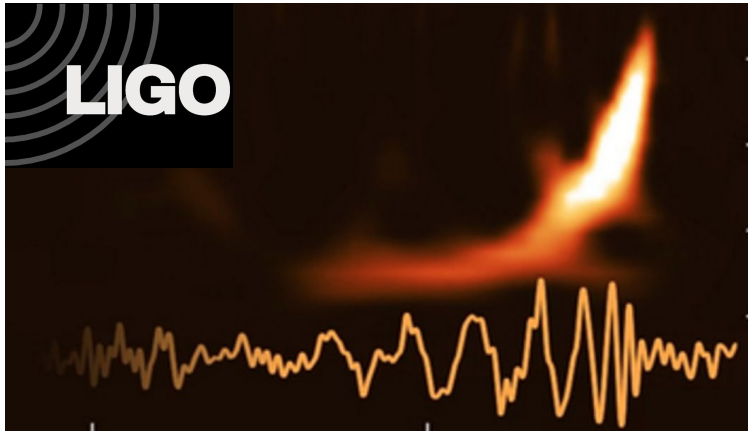
Deep Neural Networks To Enable Real-time Multimessenger Astrophysics
(arXiv:1701.00008)

Daniel George

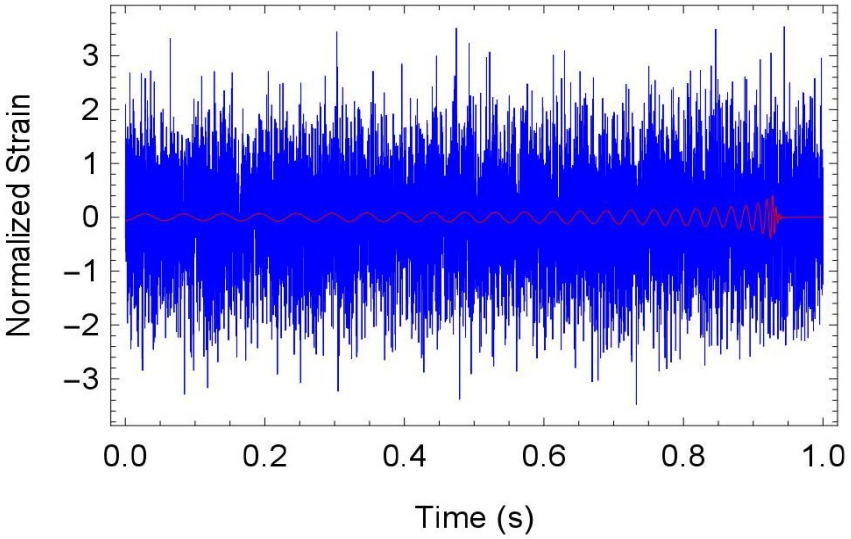
NCSA, University of Illinois at Urbana-Champaign

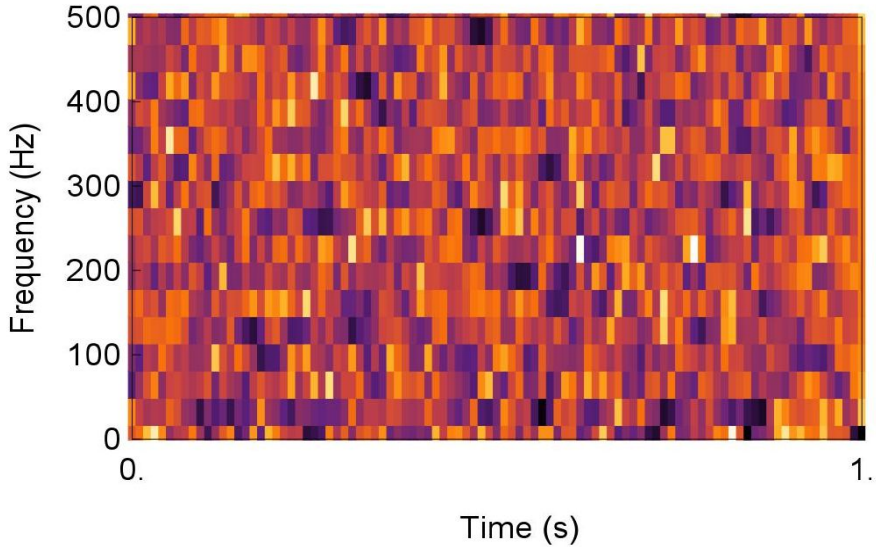
March 6, 2017

Signal Processing



- Extracting signals weaker than noise
- Traditional methods use matched-filtering (template matching)
- We developed a deep learning method trained with these templates





Artificial Neural Networks

Universality Theorem

Can model any function

Artificial Neurons

Weights (w) and bias (b)

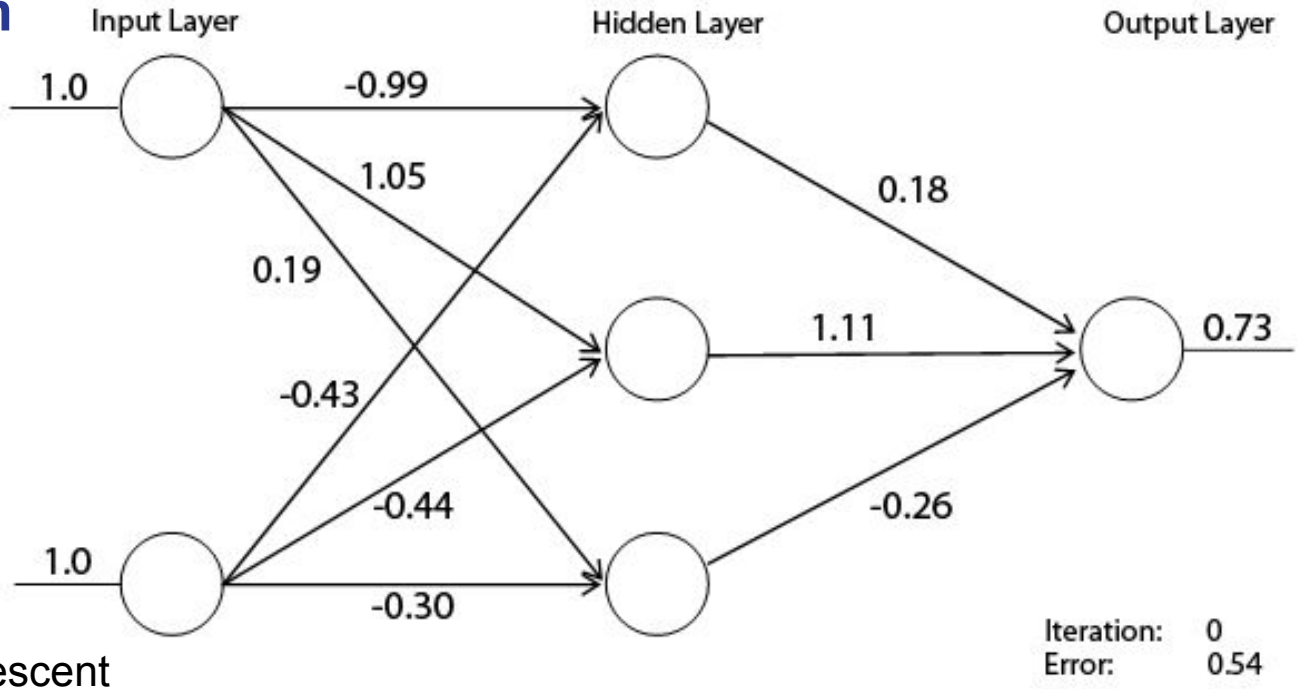
Output = f(w . Input + b)

Activation

Nonlinear function (f)

Learning Algorithm

Backpropagation, steepest descent



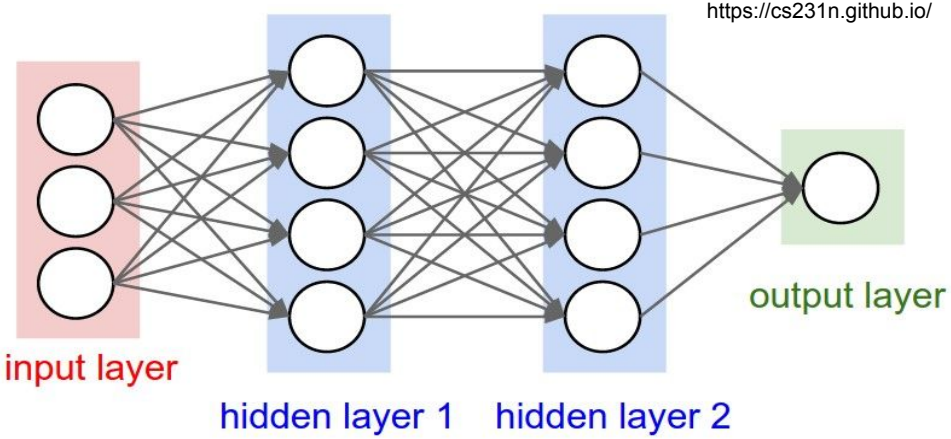
Source: codeproject.com

Deep Learning

Overview

- Very long networks of artificial neurons (dozens of layers)
- State-of-the-art algorithms for face recognition, object identification, natural language understanding, speech recognition and synthesis, web search engines, self-driving cars, games (Go) etc.

Source: https://cs231n.github.io/



- Does not require hand-crafted features to be extracted first
- Automatic end-to-end learning
- Deeper layers can learn highly abstract functions
- Optimized hardware (GPU/FPGA)

Speed-Up of Analysis

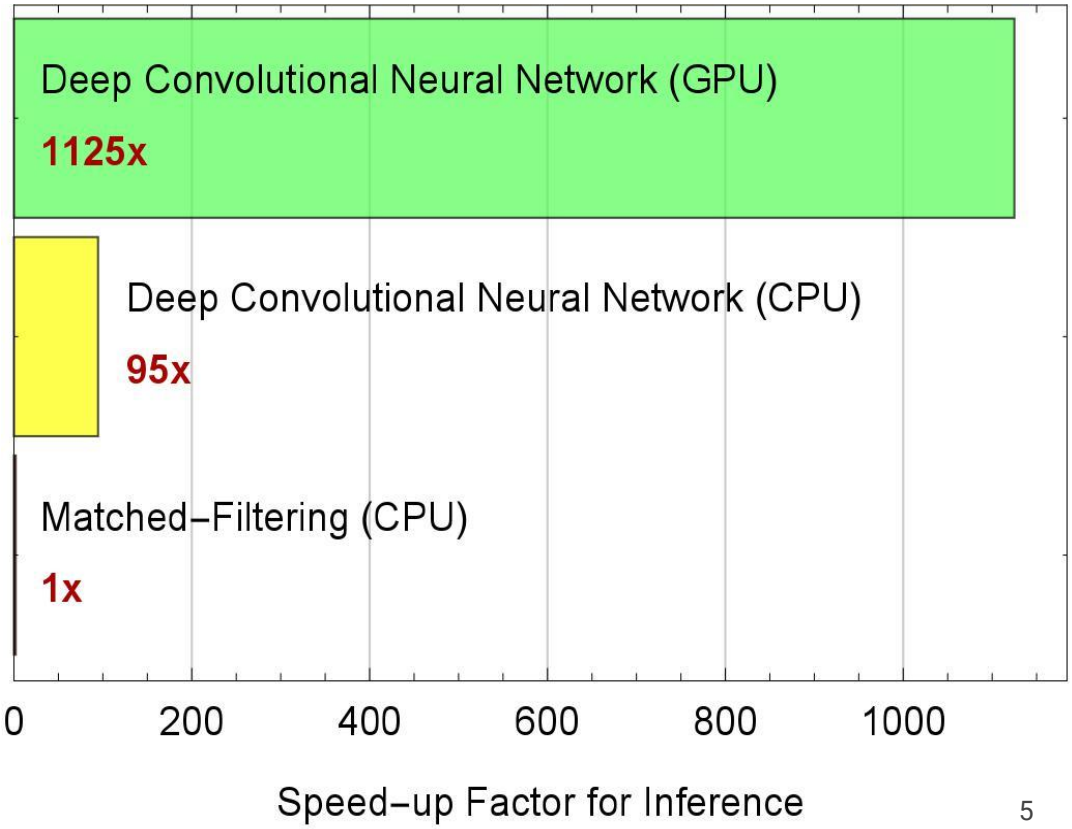
One-time intensive training process

(used Tesla & P100 GPUs at ISL)

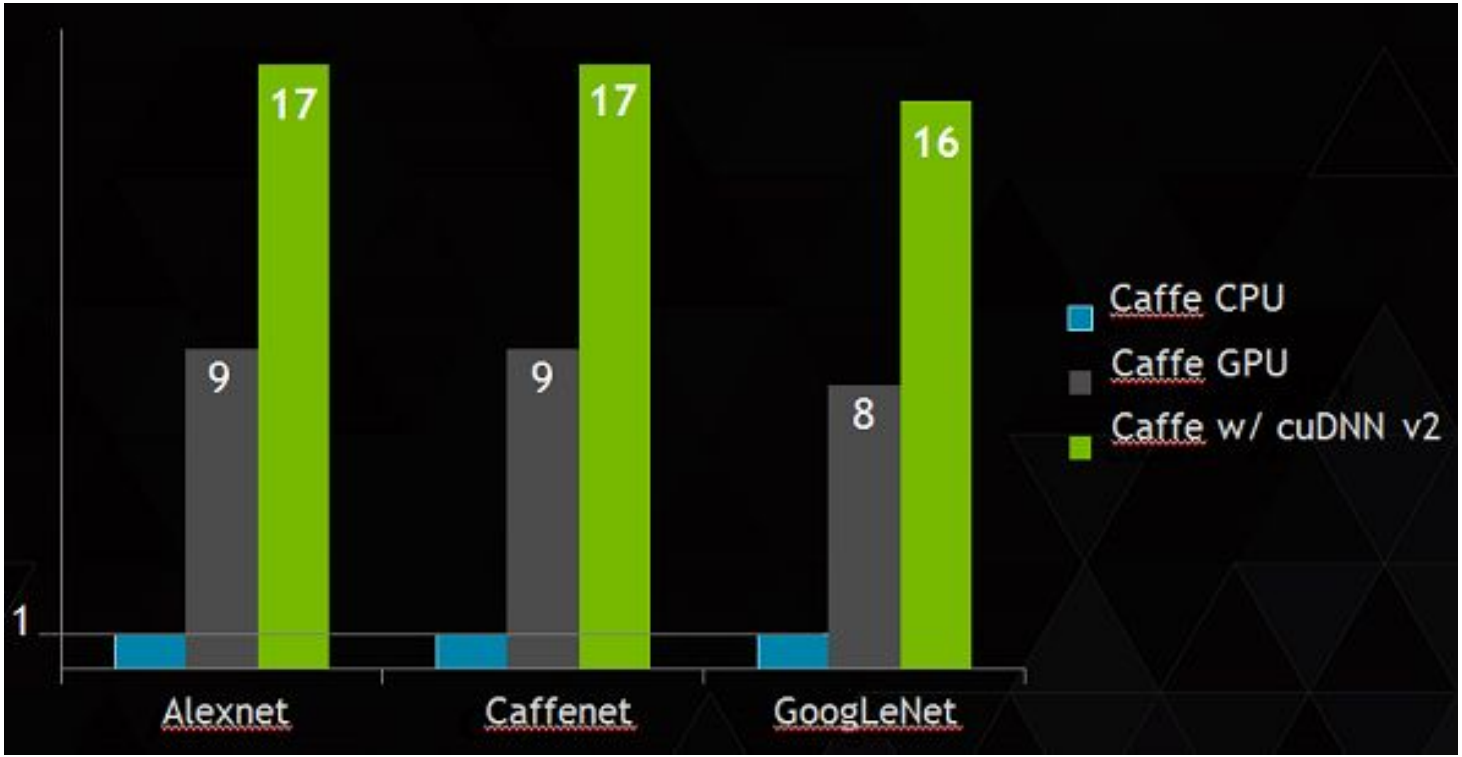
Real-time analysis (milliseconds).

Constant time of evaluation regardless of number of templates.

Thousands of inputs can be processed at once on a GPU.



Benchmarks for Training Speed-up



CPU is 16-core Intel Haswell E5-2698 2.3 GHz with 3.6 GHz Turbo. GPU is NVIDIA GeForce GTX TITAN X.

Deep Learning Frameworks

- All are optimized to automatically use NVIDIA GPUs
- Developed and/or sponsored by industries (open-source)
- Not optimized for CPUs
- Comparisons between these frameworks: https://en.wikipedia.org/wiki/Comparison_of_deep_learning_software

NVIDIA Deep Learning SDK

- **Deep Learning Primitives (cuDNN)**: High-performance building blocks for deep neural network applications including convolutions, activation functions, and tensor transformations
- **Deep Learning Inference Engine (TensorRT)**: High-performance deep learning inference runtime for production deployment
- **Deep Learning for Video Analytics (DeepStream SDK)**: High-level C++ API and runtime for GPU-accelerated transcoding and deep learning inference
- **Linear Algebra (cuBLAS)**: GPU-accelerated BLAS functionality that delivers 6x to 17x faster performance than CPU-only BLAS libraries
- **Sparse Matrix Operations (cuSPARSE)**: GPU-accelerated linear algebra subroutines for sparse matrices that deliver up to 8x faster performance than CPU BLAS (MKL), ideal for applications such as natural language processing
- **Multi-GPU Communication (NCCL)**: Collective communication routines, such as all-gather, reduce, and broadcast that accelerate multi-GPU deep learning training on up to eight GPUs

Thank you