HPC for Deep Neural Network (DNN)

Vinod Nigade

Software Engineer & Research Enthusiast

25th April, 2018

About Me

Vinod Nigade

Software Engineer

"If you optimize everything, you will always be happy" - Donald Knuth

Interests

System Software, Distributed System, HPC

Achievements

- Cum Laude in PDCS Masters at VU, Amsterdam
- VUFP scholarship for masters degree
- One publication in the Eurosys Conference
- One patent in the distributed replication area

Outline

Introduction

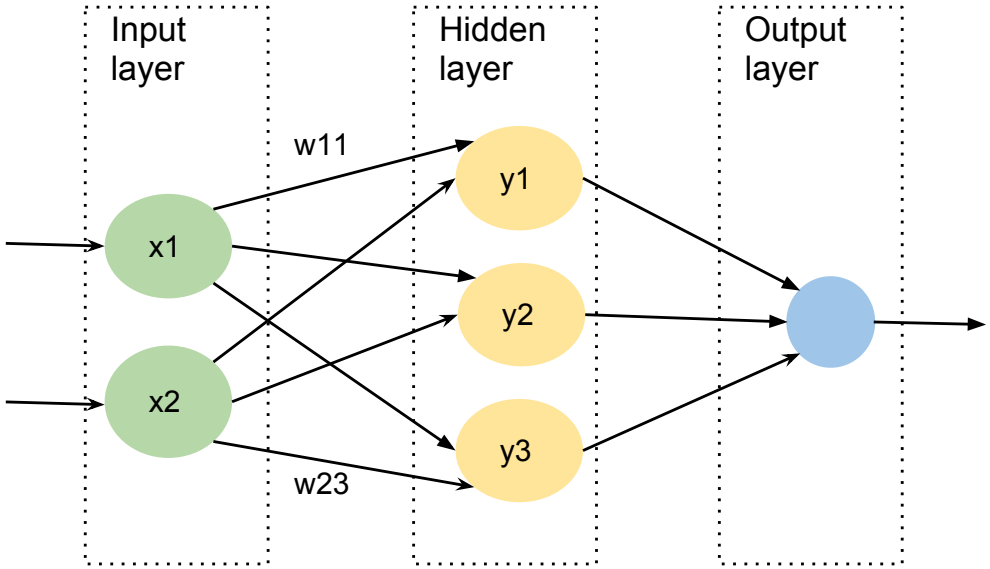
Why suddenly DNN?

DNN is getting more and more traction due to,

- Big data availability
- Huge amount of compute capacity
- Algorithmic advancement
- High growth in development resources

Introduction

What is DNN?



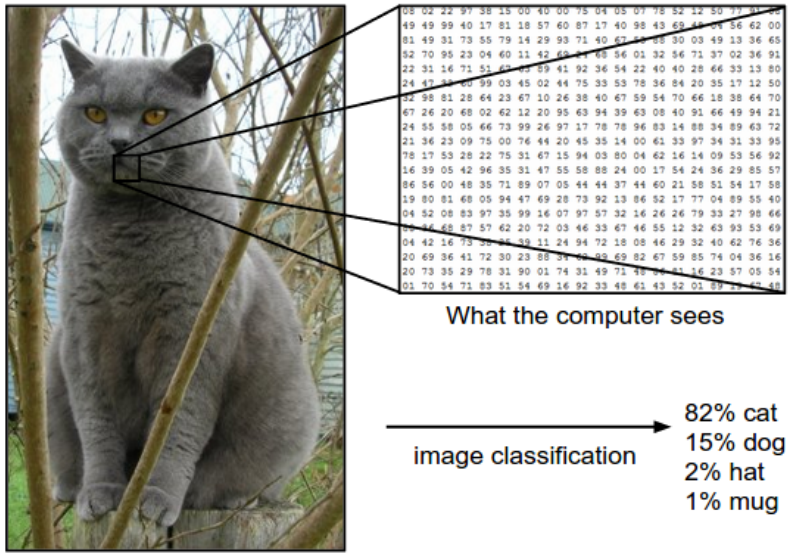
- Brain-inspired computation in machine learning

$$y_j = f(\sum_{i=1}^{2} w_{ij} \times x_i + b) \tag{1}$$

Introduction

Applications of DNN

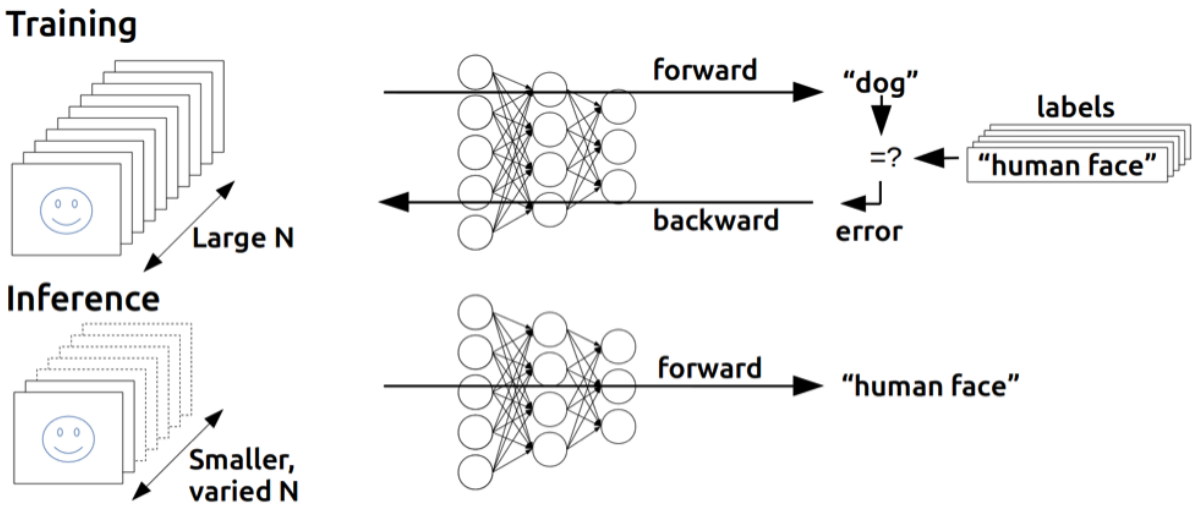
- Image and video classification
- Predictive maintenance
- Speech recognition
- Robotic and self-driving car
- Computer vision
- Health care
- Finance (e.g. prediction stock prices) etc.



Source: http://cs231n.github.io/classification/

Introduction

Training v/s Inference

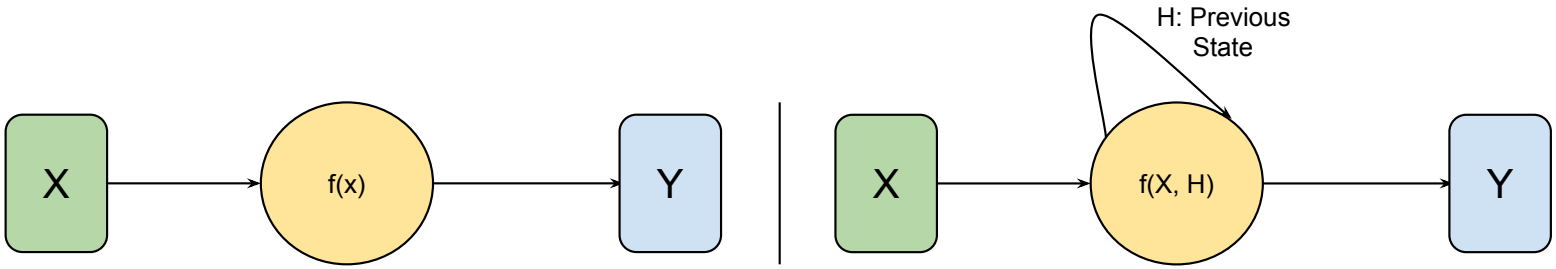


Source: https://devblogs.nvidia.com/inference-next-step-gpu-accelerated-deep-learning/

- Correct weights and biases in backward propagation
- Gradient descent optimization is commonly used in backward propagation

DNN Models

Types



Feed forward

- Convolutional neural network
- Applications in classification, identification

Recurrent

- Recurrent neural networks
- Applications in time series, sequence prediction

DNN Models

Popular models

Metrics	AlexNet	ResNet
Top-5 error (%)	16.4	3.5
Layers	8	152
Total Weights	61M	25.5M
Total MACs	724M	3.9G
Total GFLOP	1.4	22.6
DRAM Access	3000M	? ($\approx$ 100MB for weights)

larger model $\Rightarrow$ large memory access $\Rightarrow$ more energy consumption
[Sze et al., 2017]

HPC

In a nutshell

- High performance computers with huge memory and fast inter-connect
- Performance measured in FLOPS (not in MIPS)
- Cluster of CPU and GPU nodes for massively parallel computation
- Scientific applications
  - Weather forecast
  - Physics modelling and simulation
  - Defence
  - Finance etc.

Hardware

- Vector machines, GPUs, application specific accelerators like FPGA
- Infiniband, high speed Ethernet, RDMA

Software

- OpenMPI, OpenMP
- CUDA, OpenCL
- BLAS, LAPACK, Intel MKL

Co-design DNN and HPC

Why?

Goal

- Maximize accuracy and throughput
- Minimize energy and cost

How

- Faster multiple and accumulate (MAC) operations
- Reduce MAC operations
- Reduce off-chip memory access
- Parallelism

Co-design DNN and HPC

Existing techniques

Algorithm

- Adopt reduce precision
- Weights sharing
- Network pruning
- Knowledge distillation

Hardware System

- Reduced precision (FP-16 arithmetic)
- Spatial-architecture for dataflow processing
- Faster memory interconnect (e.g. HBM)

Software System

- Model Parallelism
- Data Parallelism
- RDMA support
- Overlap computation and communication

Research Questions

1. How to auto-scale system for different types of DNN models over an heterogeneous system?
2. How to design a debugger for large DNN models? Hard to find whether a low accuracy is due to a bug or model design issue?
3. How to design a performance profiler for HPC-DNN system?
4. Can we relax parameter consistency in case of DNN applications?
5. There is no randomness in DNN dataflow. How can we exploit this property for data placement and/or pre-fetching?


Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Senior, A., Tucker, P., Yang, K., Le, Q. V., et al. (2012).
Large scale distributed deep networks.
In *Advances in neural information processing systems*, pages 1223–1231.


Jacobs, S. A., Dryden, N., Pearce, R., and Van Essen, B. (2017).
Towards scalable parallel training of deep neural networks.
In *Proceedings of the Machine Learning on HPC Environments*, page 5. ACM.


Sze, V., Chen, Y.-H., Yang, T.-J., and Emer, J. S. (2017).
Efficient processing of deep neural networks: A tutorial and survey.
Proceedings of the IEEE, 105(12):2295–2329.