OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG

INF

FAKULTÄT FÜR
INFORMATIK

Performance Factors for Deep Learning and Shallow Neural Network Applications: A Beginners Guide

Johannes Wünsche

27. Januar 2018

Agenda

Motivation

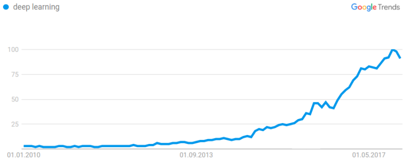
Overview

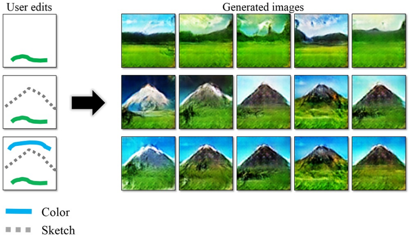
Choice of Neural Network

Choice of Processing Unit

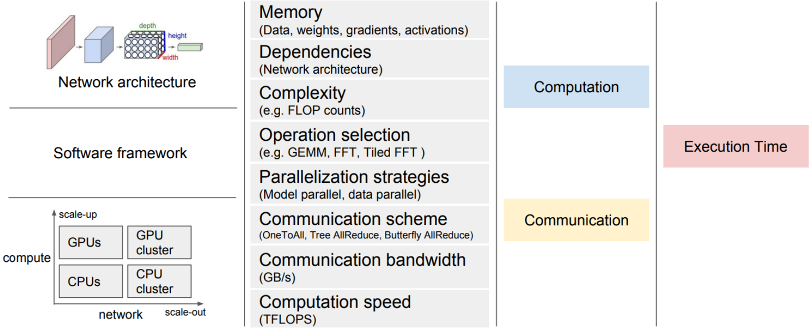
Challenges

Motivation


deep learning
Google Trends
100
75
50
25
01.01.2010
01.09.2013
01.05.2017


User edits
Generated images
Color
Sketch

Overview



Network architecture

Software framework

GPUs	GPU cluster
CPUs	CPU cluster

Memory
(Data, weights, gradients, activations)

Dependencies
(Network architecture)

Complexity
(e.g. FLOP counts)

Operation selection
(e.g. GEMM, FFT, Tiled FFT)

Parallelization strategies
(Model parallel, data parallel)

Communication scheme
(OneToAll, Tree AllReduce, Butterfly AllReduce)

Communication bandwidth
(GB/s)

Computation speed
(TFLOPS)

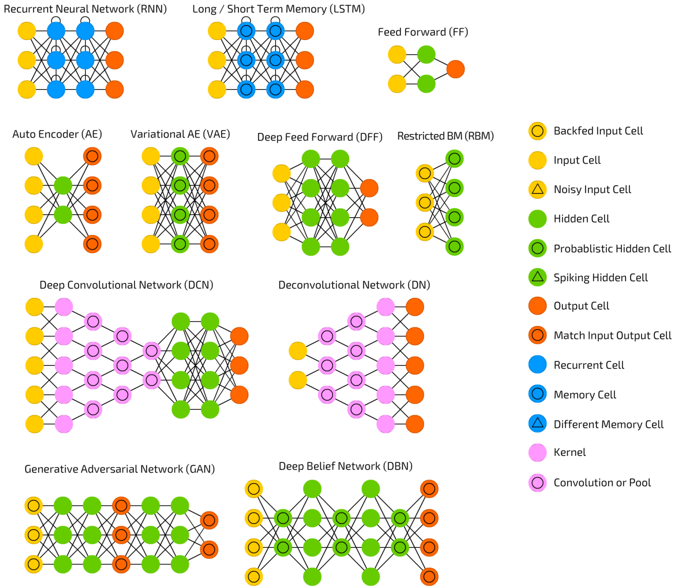
Computation

Communication

Execution Time

Choice of Neural Network

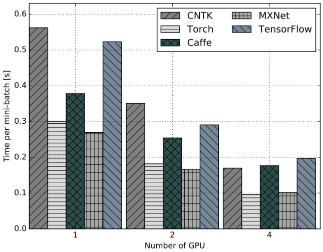
- Shallow Networks
  - FFN, AE, RBM
- Deep Networks
  - VAE, DBN, GAN, RNN e.g. LSTM, DCN, DN, RN
- Hyperparameters
- Usage of specific networks

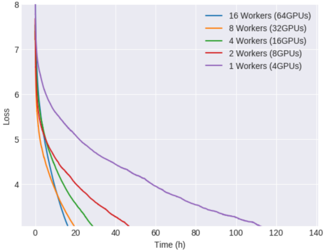

Recurrent Neural Network (RNN)
Long / Short Term Memory (LSTM)
Feed Forward (FF)
Auto Encoder (AE)
Variational AE (VAE)
Deep Feed Forward (DFF)
Restricted BM (RBM)
Deep Convolutional Network (DCN)
Deconvolutional Network (DN)
Generative Adversarial Network (GAN)
Deep Belief Network (DBN)
Backfed Input Cell
Input Cell
Noisy Input Cell
Hidden Cell
Probablistic Hidden Cell
Spiking Hidden Cell
Output Cell
Match Input Output Cell
Recurrent Cell
Memory Cell
Different Memory Cell
Kernel
Convolution or Pool

Choice of Processing Unit

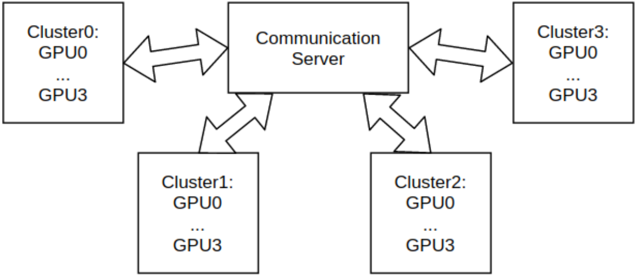
- CPU
  - Single-thread
  - Multi-thread
  - Advantage
  - Disadvantage
- GPU
  - Single unit
  - Multi-GPU
  - Advantage
  - Disadvantage
- GPU-Cluster
  - Advantage
  - Disadvantage

Framework	1 Thread	2 Threads	4 Threads	8 Threads
Caffe	1.324	0.790	0.578	15.444
Tensorflow	7.062	4.789	2.648	1.938
Torch	1.329	0.710	0.423	na





Challenges



Cluster0:
GPU0
...
GPU3

Communication
Server

Cluster3:
GPU0
...
GPU3

Cluster1:
GPU0
...
GPU3

Cluster2:
GPU0
...
GPU3

Thank you for your attention!

References I

- S. Shi, Q. Wang, P. Xu, and X. Chu, "Benchmarking state-of-the-art deep learning software tools," *arXiv preprint arXiv:1608.07249*, 2016.
- M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.
- Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.
- R. Collobert, S. Bengio, and J. Mariéthoz, "Torch: a modular machine learning software library," Idiap, Tech. Rep., 2002.
- T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang, "Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems," *arXiv preprint arXiv:1512.01274*, 2015.
- F. Sastre Cabot, "Scalability study of deep learning algorithms in high performance computer infrastructures," Master's thesis, Universitat Polit'ecnica de Catalunya, 2017.
- J. Patterson and A. Gibson, *Deep Learning: A Practitioner's Approach*. Beijing: O'Reilly, 2017. [Online]. Available: https://www.safaribooksonline.com/library/view/deep-learning/9781491924570/

References II

M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The rprop algorithm," in *Neural Networks, 1993., IEEE International Conference on*. IEEE, 1993, pp. 586–591.

H. Qi, E. R. Sparks, and A. Talwalkar, "Paleo: A performance model for deep neural networks," 2016.

S. team Chris Nicholson. Deeplearning4j.org. Access Date: 06.01.18 11:50:00. [Online]. Available: https://deeplearning4j.org/

Microsoft/cntk. Access Date: 06.01.18 11:55:34. [Online]. Available: https://github.com/Microsoft/CNTK

Keras.io. Access Date: 06.01.18 12:00:17. [Online]. Available: https://keras.io/

Mxnet: A scalable deep learning framework. Access Date: 06.01.18 12:02:43. [Online]. Available: https://mxnet.apache.org/

Junyanz, "junyanz/igan," Apr 2017, access Date: 07.01.18 10:58:57. [Online]. Available: https://github.com/junyanz/iGAN

References III

Spiglerg. (2017, Dec) spiglerg/rnn_text_generation_tensorflow. Access Date: 07.01.18 11:04:34. [Online]. Available: https://github.com/spiglerg/RNN_Text_Generation_Tensorflow

P. Zhong and Z. Gonga, “a diversified deep belief network for hyperspectral image classification,” *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 443–449, 2016.

H. Soltau, H. Liao, and H. Sak, “Neural speech recognizer: Acoustic-to-word lstm model for large vocabulary speech recognition,” *arXiv preprint arXiv:1610.09975*, 2016.

J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, “Generative visual manipulation on the natural image manifold,” in *European Conference on Computer Vision*. Springer, 2016, pp. 597–613.

I. Sutskever, J. Martens, and G. E. Hinton, “Generating text with recurrent neural networks,” in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 1017–1024.

Y. Yang, P. Xiang, J. Kong, and H. Zhou, “A gpgpu compiler for memory optimization and parallelism management,” in *ACM Sigplan Notices*, vol. 45, no. 6. ACM, 2010, pp. 86–97.

“Nvidia gpudirect,” Apr 2017, access date: 08.01.18 15:45:57. [Online]. Available: https://developer.nvidia.com/gpudirect

W. Wang, G. Chen, H. Chen, T. T. A. Dinh, J. Gao, B. C. Ooi, K.-L. Tan, S. Wang, and M. Zhang, “Deep learning at scale and at ease,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 12, no. 4s, p. 69, 2016.

References IV

I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

L. Deng, D. Yu *et al.*, “Deep learning: methods and applications,” *Foundations and Trends® in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

Feb 2017. [Online]. Available: https://beamandrew.github.io/deeplearning/2017/02/23/deep_learning_101_part1.html

F. v. Veen, “The neural network zoo,” Nov 2017. [Online]. Available: http://www.asimovinstitute.org/neural-network-zoo/

G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

R. Frank, “The perceptron a perceiving and recognizing automaton,” *tech. rep., Technical Report 85-460-1*, 1957.

A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.