Practical Challenges in training ConvNets for Vision

Soumith Chintala
Facebook AI Research
soumith@gmail.com

IIIT Hyderabad 18th February, 2015

Overview

- Brief Introduction
  - Neural Networks and Backpropagation
  - Convolutions
  - Linear Networks and Non-linear activations
  - Gradient Descent

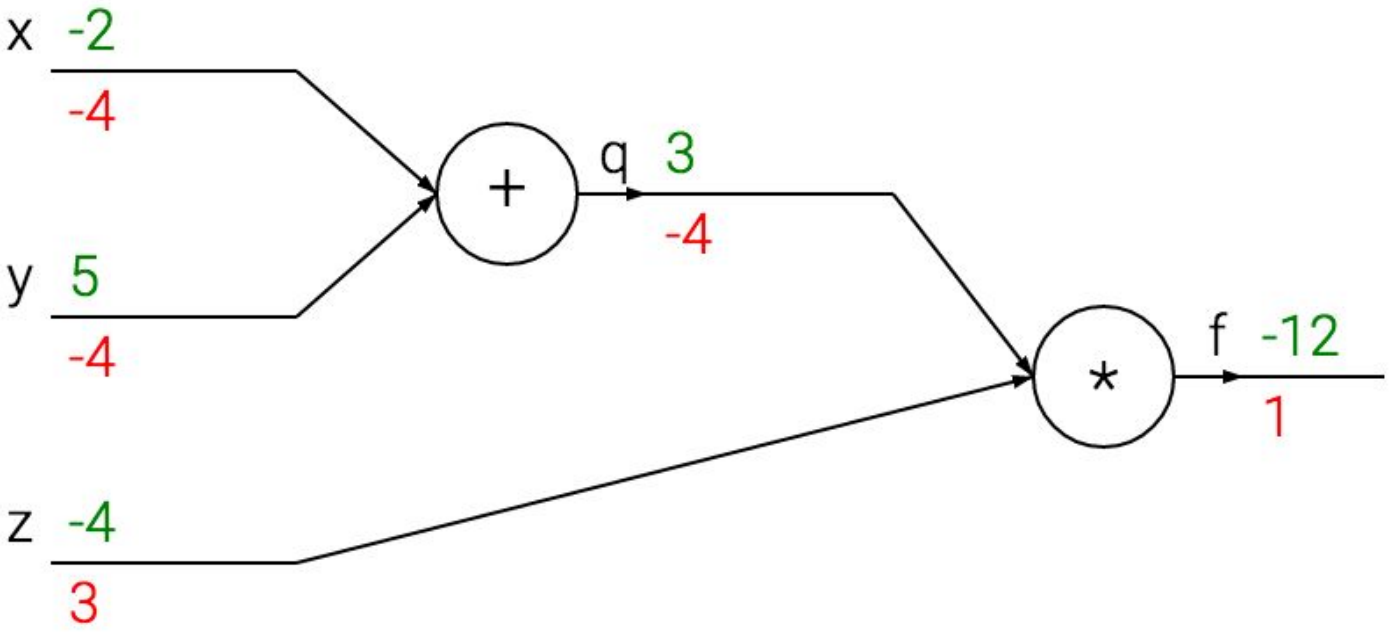
Overview

- Practical issues
  - data (in-memory? on-disk?)
  - GPUs, multiple GPUs
  - hyperparameter search
  - weight initialization
  - choosing architecture
  - Numerical stability

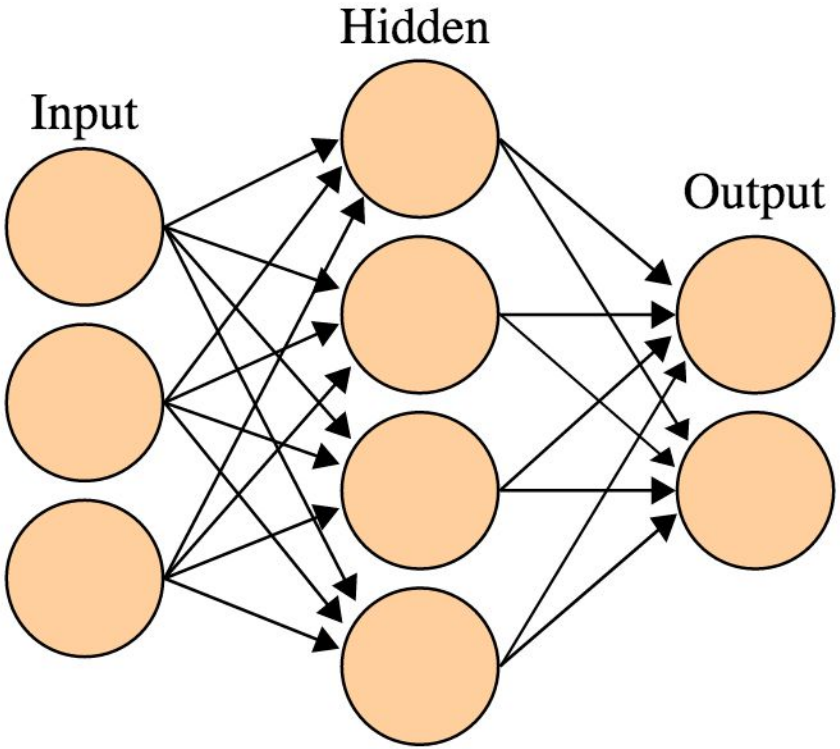
Pre-thanks

- Andrej Karpathy for his fantastic teaching material at: http://cs231n.github.io/
- If in doubt, go do the course
- Christopher Olah for his simple and intuitive blog posts
  - http://colah.github.io
  -

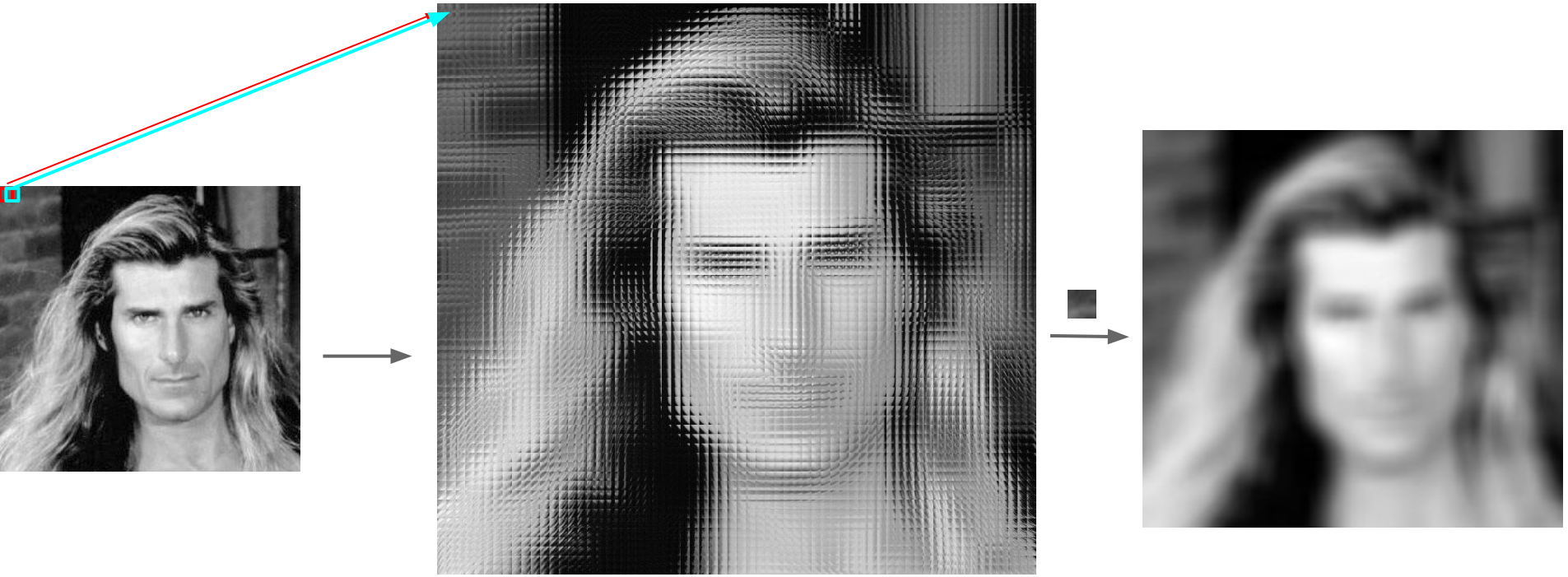
Neural networks and Backpropagation



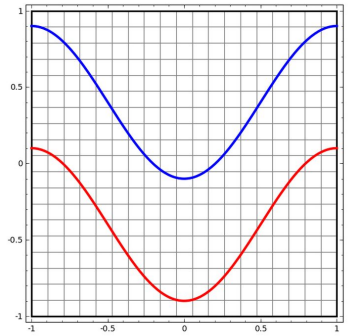
Neural Networks

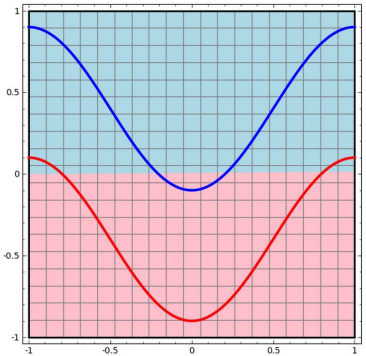


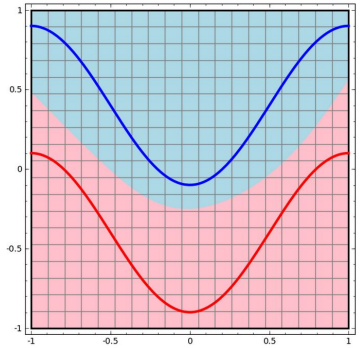
Convolutions



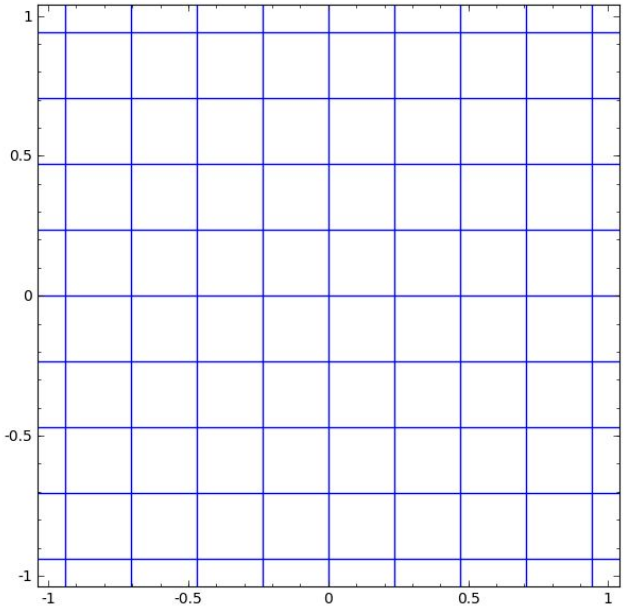
Linear vs Non-linear

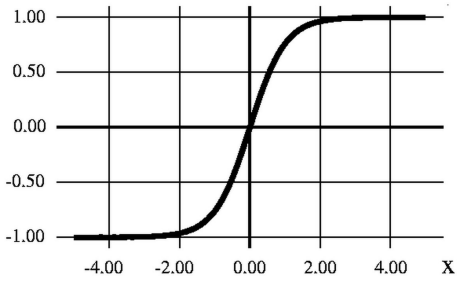


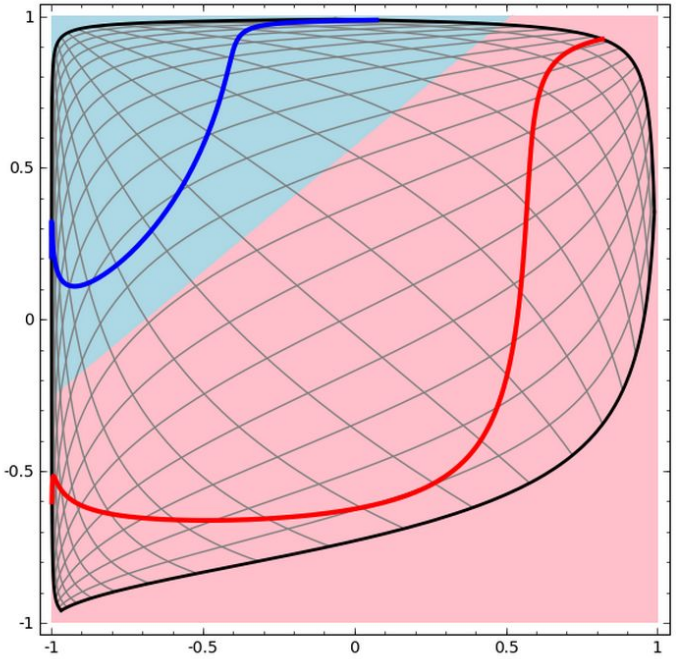




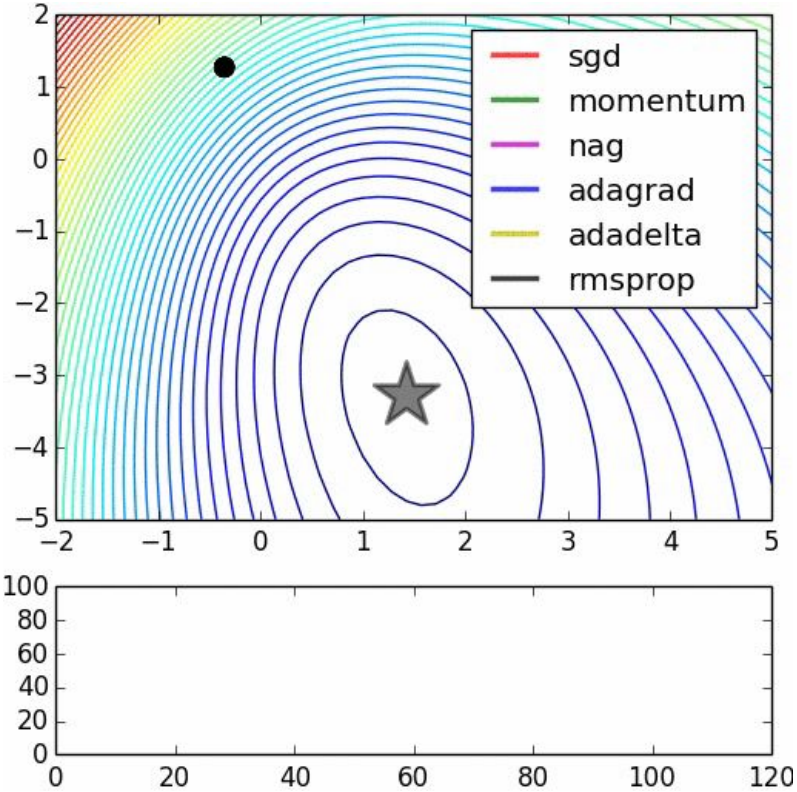
Tanh non-linearity

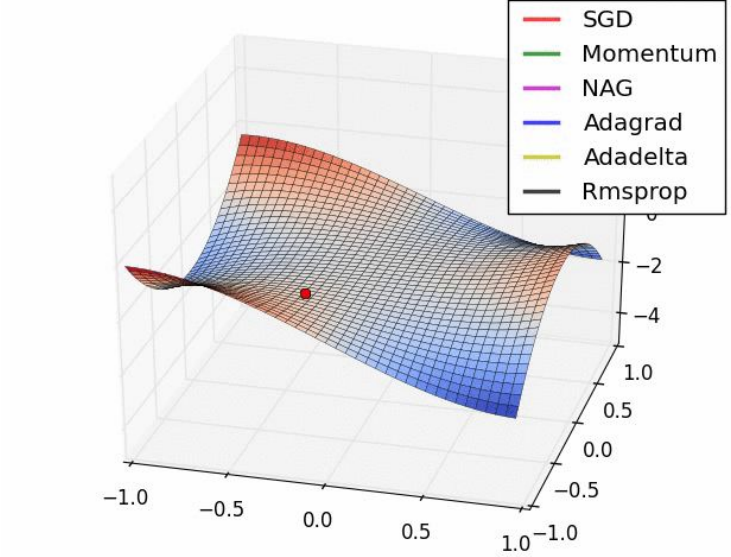


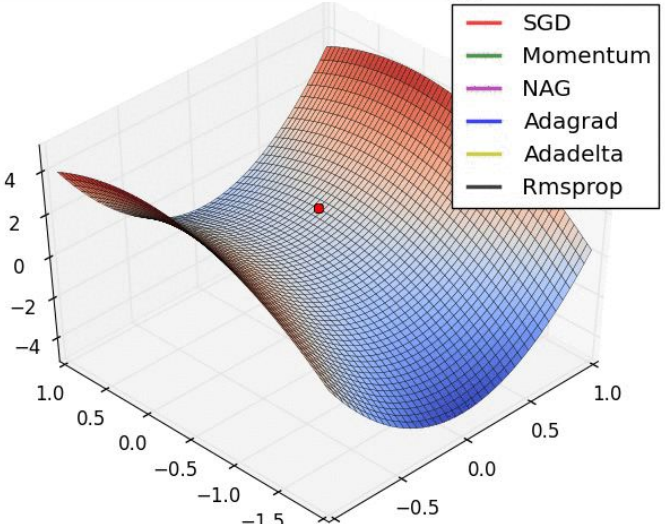




Gradient descent







Practical Issues - Data

- In-memory
- GPU Transfers
- SSD
- Distributed Storage
- LMDB, CouchDB
- Multi-threaded loading

Practical Issues - GPUs

- NVIDIA CUDA
- single vs multi-GPU

GPUs	Batch size	Cross-entropy	Top-1 error	Time	Speedup
1	(128, 128)	2.611	42.33%	98.05h	1x
2	(256, 256)	2.624	42.63%	50.24h	1.95x
2	(256, 128)	2.614	42.27%	50.90h	1.93x
4	(512, 512)	2.637	42.59%	26.20h	3.74x
4	(512, 128)	2.625	42.44%	26.78h	3.66x
8	(1024, 1024)	2.678	43.28%	15.68h	6.25x
8	(1024, 128)	2.651	42.86%	15.91h	6.16x

One Weird Trick for Paralellizing ConvNets - Alex Khrizevsky

Practical issues - Hyperparameters

- Learning rate and momentum
- Weight decay
- Number of feature maps
- Convolution size
- Dropout
- Optimization algorithm

Practical issues - Weight Initialization

- Normalized initialization
- Xavier’s distribution
- Orthogonal initialization

Practical Issues: choosing architecture

Numerical Stability

- Float precision
- NaNs from SoftMax

Discussion

- Problems being faced