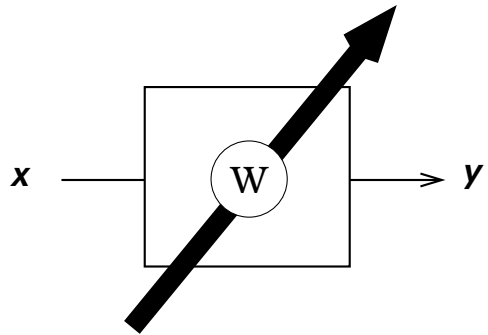
Deep Learning as compute paradigm

Roland Memisevic

University of Montreal

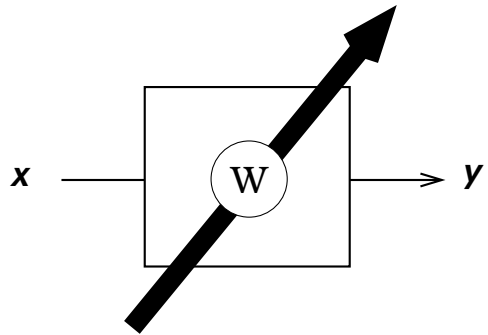
January 28, 2015

Machine Learning



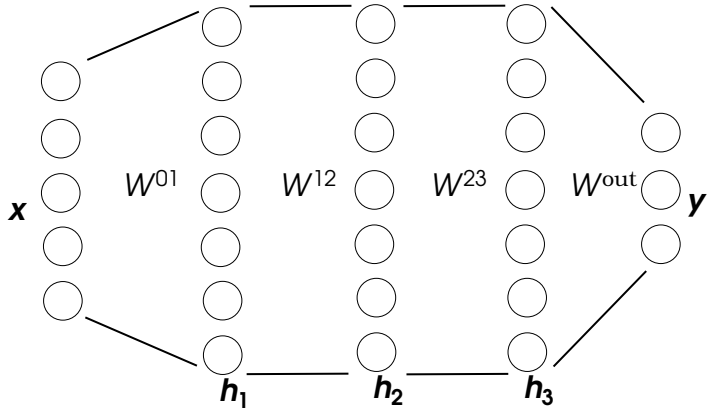
- classic view: Learning allows us to harness **data**

Machine Learning

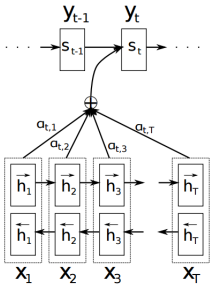


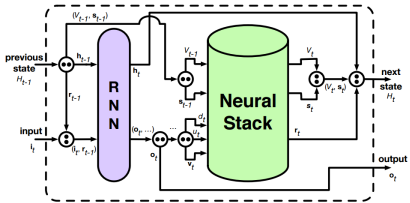
- classic view: Learning allows us to harness **data**
- better view: Learning allows us to harness **hardware**

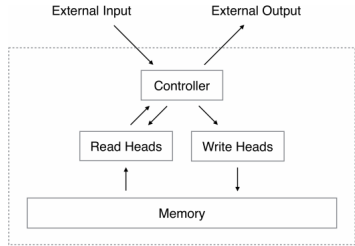
Neural networks run on parallel operations



Von Neumann via Deep Learning

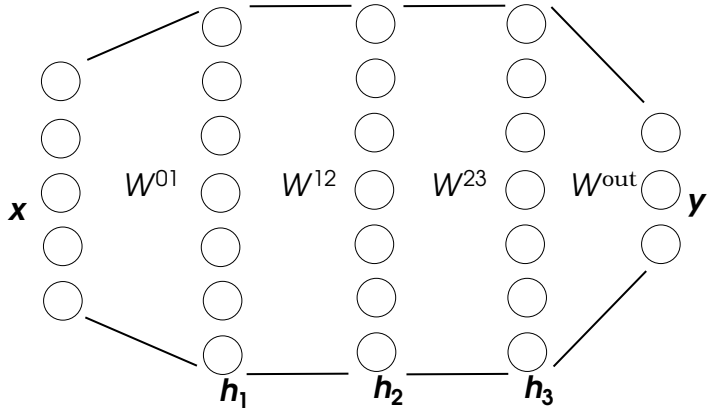






(Bahdanau et al. 2014), (Graves et al, 2014), (Weston et al, 2014), (Grefenstette et al. 2015), etc.

Neural networks run on parallel operations



Floating point multiplication

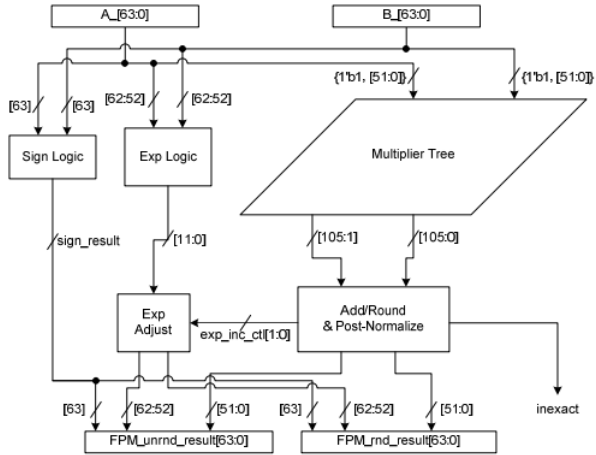


figure from: http://www.gamasutra.com/view/news/128521/Indepth_IEEE_754_Multiplication_And_Addition.php

- Waste of circuitry?

Low-precision weights

- Courbariaux, Bengio, David (2015): Stochastically binarize (or ternarize) weights during **propagation** (forward and backward) of activations:

$$p(W_{ij} = 1) = \frac{W_{ij} + 1}{2}, p(W_{ij} = -1) = 1 - P(W_{ij} = 1)$$

Method	MNIST	CIFAR-10	SVHN
No regularizer	$1.30 \pm 0.04\%$	10.64%	2.44%
BinaryConnect (det.)	$1.29 \pm 0.08\%$	9.90%	2.30%
BinaryConnect (stoch.)	$1.18 \pm 0.04\%$	**8.27%**	2.15%
50% Dropout	$1.01 \pm 0.04\%$		
Maxout Networks [29]	0.94%	11.68%	2.47%
Deep L2-SVM [30]	**0.87%**		
Network in Network [31]		10.41%	2.35%
DropConnect [21]			1.94%
Deeply-Supervised Nets [32]		9.78%	**1.92%**

- see also (Soudry et al, NIPS 2014), (Cheng et al, arXiv 2015), (Hwang & Sung, 2015)

Neural nets with few multiplications

- Lin, Courbariaux, Memisevic, Bengio (2015):

- Backprop updates:

$$\begin{aligned} \Delta W &= \left[\eta\delta \circ h'(Wx+b)\right] x^{\mathrm{T}} \\ \Delta b &= \eta\delta \circ h'(Wx+b) \\ \delta &= \left[W^{\mathrm{T}}\delta\right] \circ h'(Wx+b) \end{aligned}$$

	Full precision	Binary connect	Binary connect + Quantized backprop	Ternary connect + Quantized backprop
MNIST	1.33%	1.23%	1.29%	1.15%
CIFAR10	15.64%	12.04%	12.08%	12.01%
SVHN	2.85%	2.47%	2.48%	2.42%

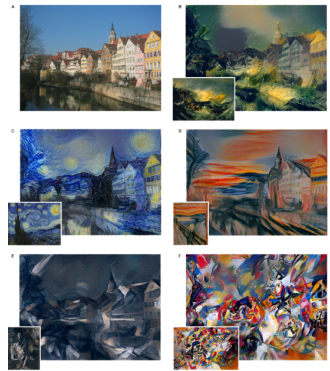
	Full precision	Ternary connect + Quantized backprop	ratio
without BN	1.7498×10^9	3.6988×10^6	0.002114
with BN	1.7554×10^9	9.2741×10^6	0.005283

- Eliminate multiplications in updates by quantizing activations to power of two
  - see also (Simard, Graf 1992)
  - low precision *activations*: (Kim, Smaragdis 2015), (Hwang, Sung 2014), (Vanhoucke et al 2011)

Transfer learning

- A sufficiently universal computational model can learn to solve many different problems.
- This allows us to pool tasks and overcome data scarcity, *without unsupervised learning*.
- Pre-trained modules are the software libraries of DL.
- (Girshick et al 2014), (Razavian et al 2014), (Luong et al 2015), etc.

An example:



Gatys, Ecker, Bethge (2015)

generic $\leftrightarrow$ specific

- according to the classic paradigm:

  specific: faster, but tedious to program

  $\leftrightarrow$

  generic: slower, but easy to program

- according to the DL paradigm:

  specific: more accurate, but more data needed

  $\leftrightarrow$

  generic: less accurate, but less data needed

Where are humans along that scale?

$$3 + x = 7$$

- We solve equations like these using "dexterous manipulation" involving our motor cortex.
- See, eg., (Hofstadter, Sander 2013), (Lakoff, 1980), or the "embodied cognition" movement

Cognition via analogy making is useful:
It enables data pooling

Twenty Billion Neurons

www.twentybn.com

Thank you

Questions?