2011 1a

Max pooling size:

~~layer size + 2·padding - filter~~

$$\frac{\text{layer} + 2\cdot\text{padding} - \text{filter}}{\text{stride size}} + 1$$

CNN: features (Image)
LSTM: time series (video and audio)
+ softmax to build classifier

Audio: MFCC

2017 2a $E = \sum_{d\in D} \frac{1}{2}(t_d - O_d)^2$

$$\frac{dE}{dw_i} = \sum_{d\in D} \frac{1}{2}\cdot 2(t_d - O_d)\cdot\frac{d}{dw_i}(-O_d)$$

$$= \sum_{d\in D}(t_d - O_d)\cdot -(w_0 + w_1 x_1 + w_2 x_i^3 \ldots)$$

$$= -\sum_{d\in D}(t_d - O_d)\cdot(w_i x_{id} + x_{id}^3)$$

$$\Delta w = \eta \sum_{d\in D}(t_d - O_d)(x_{id} + x_{id}^3)$$

~~2017 3a~~ (quadratic error) $E = \frac{1}{2}(t_d - O_d)^2 + \lambda\Sigma|w|$

L1: $\lambda\Sigma|w| \Rightarrow -\eta\lambda\,\text{sign}(w)$

L2: $\frac{1}{2}\lambda\Sigma w^2 \Rightarrow -\eta\lambda w$

Binary cross entropy: $E = -\sum_{k=1}^{K}(t_k \ln O_k + (1-t_k)\ln(1-O_k))$

used with sigmoid $\Delta w_{ij} = -\eta\frac{dE}{dO_k}\cdot\frac{d\sigma(net_k)}{dnet_k}y_j$

$$\frac{dE}{dO_k} = \frac{O_k - t_k}{O_k(1-O_k)} \qquad \frac{d\sigma(net_k)}{dnet_k} = O_k(1-O_k)$$

~~$\Delta w_{kj} = -\eta(t_k - O_k)y_j$~~

$$\Delta w_{kj} = -\eta\frac{O_k - t_k}{O_k(1-O_k)}\cdot\frac{O_k(1-O_k)}{1}y_j = \eta(t_k - O_k)y_j$$

Negative log likelihood: $E = -\sum_k t_k \ln O_k$ or $-\log O_k$ → $\frac{\partial E}{\partial t_k}\frac{t_k}{O_k}$

used with softmax

Softmax: $O^k = \frac{e^{net_k}}{\sum_k e^{net_k}} \rightarrow \frac{dE}{dw_{ij}} = y_i(O_j - t_j)$

2017 3a: $E = \frac{1}{2}(t_d - O_d)^2 + \lambda\Sigma|w|$ (refer to Mitchell P.113)

$$\frac{dE}{dw_{ji}} = \frac{dE_d}{dnet_j}\cdot\frac{dnet_j}{dw_{ji}} = \frac{dE_d}{dnet_j}\cdot x_{ji}$$

$$\frac{dE_d}{dnet_j} = \frac{dE_d}{dnet_j}\cdot\frac{dE_d}{dO_j}\cdot\frac{dO_j}{dnet_j} = \frac{dE_d}{dO_j}\cdot O_j(1-O_j) \quad \text{sigmoid}$$

$$\frac{dE_d}{dO_j} = \frac{1}{2}\cdot 2(t_d - O_d)\cdot\frac{d}{dO_j}(-O_d) = -(t_d - O_d)$$

$$\frac{dE}{dw_{ji}} = -(t_d - O_d)O_j(1-O_j)x_{ji}$$

$$\Delta w_{ji} = -\eta\frac{dE}{dw_{ji}} = \eta(t_d - O_d)O_j(1-O_j)x_{ji} - \eta\lambda\,\text{sign}(w)$$

Momentum: $\Delta w_i(t) = \mu\cdot\Delta w_i(t-1) + (\eta\frac{dE}{dw_i(t)})$

$\mu = 0.9/0.95$ (0.5 initial)

η decay: $\frac{\eta_0\cdot T}{\max(t,T)}$, $\eta_0\cdot C$, $\frac{\eta_0}{1+\frac{t}{C}}$

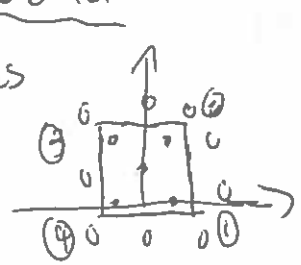
Initiation: Glorot $= \sqrt{2/(n_1+n_2)}$, He $= \sqrt{\frac{2}{n_1}}$ (for Relu)

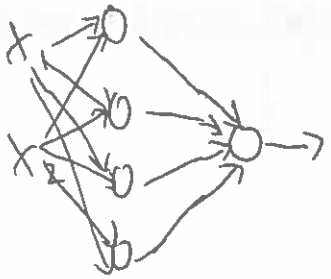
data normalization: $y = \frac{x - x_{mean}}{x_{std}}$

Relu: $y = \max(0,x) \rightarrow y' \begin{cases} 1, & \text{if } x>0 \\ 0, & \text{otherwise} \end{cases}$

2006 1a

4 lines





2006 1b refer to Mitchell p.115

$$\frac{dE_d}{dw_{ji}} = \frac{dE}{dw} \frac{dE_d}{dnet_j} \cdot \frac{dnet_j}{dw_{ji}} = \frac{dE_d}{dnet_j} \cdot x_{ji}$$

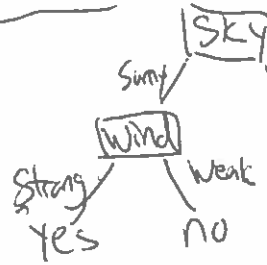
$$\frac{dE_d}{dnet_j} = \sum_{k \in Down(j)} \frac{dE_d}{dnet_k} \cdot \frac{dnet_k}{dnet_j}$$

$$\frac{dE_d}{dnet_j} = \sum_{k \in Down(j)} -\delta_k \cdot \frac{dnet_k}{dO_j} \cdot \frac{dO_j}{dnet_j}$$

$$\delta_j = \sum_{k \in Down(j)} -\delta_k \cdot w_{kj} \cdot O_j(1-O_j)$$

$$\Delta w_{ji} = \eta \delta_j x_{ji}$$

2012 1a



$$\tanh = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$$\tanh' = 1 - \tanh^2(x)$$

2012 2a,b,c: Forward, $E = \frac{1}{2}(t_\alpha - O_\alpha)^2 + \frac{1}{2}(t_\beta - O_\beta)^2$

sigmoid layer $O_\alpha = \frac{1}{1+e^{-net_\alpha}}$ $O_\beta = \frac{1}{1+e^{-net_\beta}}$

net: $Net_\alpha = W_{A\alpha} y_A + W_{B\alpha} y_B + W_{C\alpha} y_C$, $Net_\beta = W_{A\beta} y_A + W_{B\beta} y_B + W_{C\beta} y_C$

sig:

$$y_A = \frac{1}{1+e^{-net_A}} \quad y_B = \frac{1}{1+e^{-net_B}} \quad y_C = \frac{1}{1+e^{-net_C}}$$

net: $Net_A = W_{2A} x_2 + W_{1A} x_1$, $Net_B = W_{2B} x_2 + W_{1B} x_1$

$Net_C = W_{2C} \cdot x_2 + W_{1C} \cdot x_1$

Backward: $\delta_{\alpha/\beta} = -\frac{dE}{dnet_k} = -\frac{dE}{dO_\alpha} \cdot \frac{dO_\alpha}{dnet_k} = -\frac{d}{dO_\alpha}\left(\frac{1}{2}(t_\alpha - O_\alpha)^2\right) \cdot \frac{d}{dnet_\alpha}\left(\frac{1}{1+e^{-net_\alpha}}\right)$

$= \frac{1}{2}(2)(t_\alpha - O_\alpha) \cdot (-1) \frac{(-1)}{(1+e^{-net})^2} \cdot (e^{-net_2})(-1)$

$= k(t_k - O_k)(O_k)(1-O_k)$

$\forall j \in \{A,B,\} k \in (\alpha,\beta)$ $\Delta w_{ji} = -\eta \frac{dE}{dw_{ji}} = -$

2010 2b: $\frac{1}{1+e^{-x}}$ sigmoid

$$\frac{d}{dx}\sigma(x) = \frac{d}{dx}(1+e^{-x})^{-1} = -(1+e^{-x})^{-2}(-e^{-x}) = \frac{e^{-x}}{(1+e^{-x})^2}$$

$$= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} = \frac{1}{1+e^{-x}} \cdot \frac{(1+e^{-x})-1}{1+e^{-x}} = \frac{1}{1+e^{-x}} \cdot \left(\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right)$$

$$= \frac{1}{1+e^{-x}} \cdot \left(1 - \frac{1}{1+e^{-x}}\right) = \sigma(1-\sigma)$$

id3: doesn't consider all consistent hypo while C-E does

CI: $Error \pm \sqrt{\frac{Error(1-Error)}{n}}$

$(2 \cdot 3 \cdot 3 \cdot 7) : (2 \cdot 3 \cdot 3 \cdot 7) = 1 : 5876$

2016 2a: $w_i \leftarrow w_i + \eta \frac{dE}{dw_i}$, $E = \frac{1}{2}(t_d - o_d)^2$ (LMS; Quadratic Error)

$$\frac{dE}{dw_i} = -(t_d - o_d)(x_{id} + x_{id}^2 + x_{id}^3)$$

$$w_i \leftarrow w_i + \eta(t_d - o_d)(x_{id} + x_{id}^2 + x_{id}^3)$$

LMS alters weight in same proportion to SGD proving LMS performs SGD

or

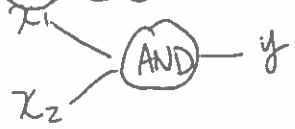
wo LMS: $w_i \leftarrow \eta(t_d - o_d)$ ... (substitution makes this linear)

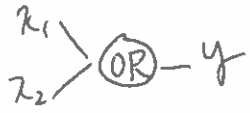
$(t_d - o_d) = \frac{dE}{dw}$ ↗ plug

$(x_{id} + x_{id}^2 + x_{id}^3)$

LMS: $w_i \leftarrow \eta(t_d - o_d) \cdot x_d = w_i \leftarrow \eta \frac{dE}{dw}$

2015 2a: w=1





x_1	x_2	$x_1w_1 + x_2w_2$	y
1	1	2	1
1	0	1	0
0	1	1	0
0	0	0	0

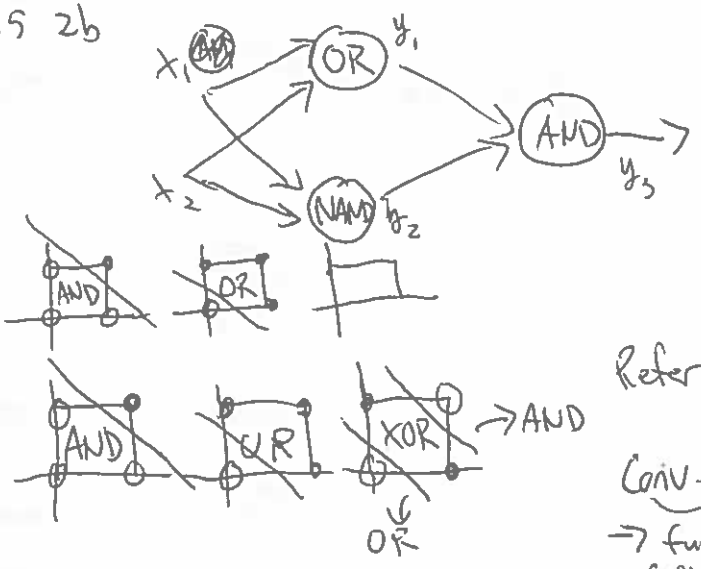
x_1	x_2	Net	y
1	1	2	1
1	0	1	1
0	1	1	1
0	0	0	0

x_1	x_2	net	y
1	1	2	0
1	0	1	1
0	1	1	1
0	0	0	1

$$Step_{NAND} = \begin{cases} 1 \text{ if net} < 2 \\ 0 \text{ otherwise} \end{cases}$$

$$Step_{AND} = \begin{cases} 1 \text{ if net} \geq 2 \\ 0 \text{ otherwise} \end{cases} \quad Step_{OR} = \begin{cases} 1 \text{ if net} \geq 1 \\ 0 \text{ otherwise} \end{cases}$$

2015 2b

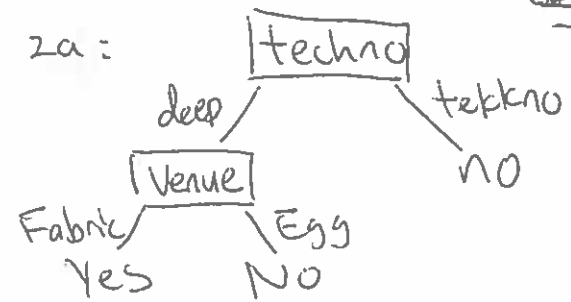


y_1	y_2	y_3
1	1	0
1	0	1
0	1	1
0	0	0

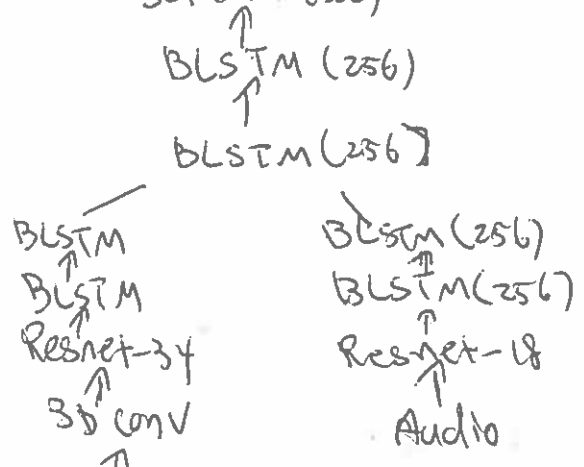
Refer to 6ltd lecture 4 slide 20, 23

ConV → pool/sample → ConV → pool → fully connected → softmax

2014 2a:



End to End Audio visual



Refer to lecture 6. page 39