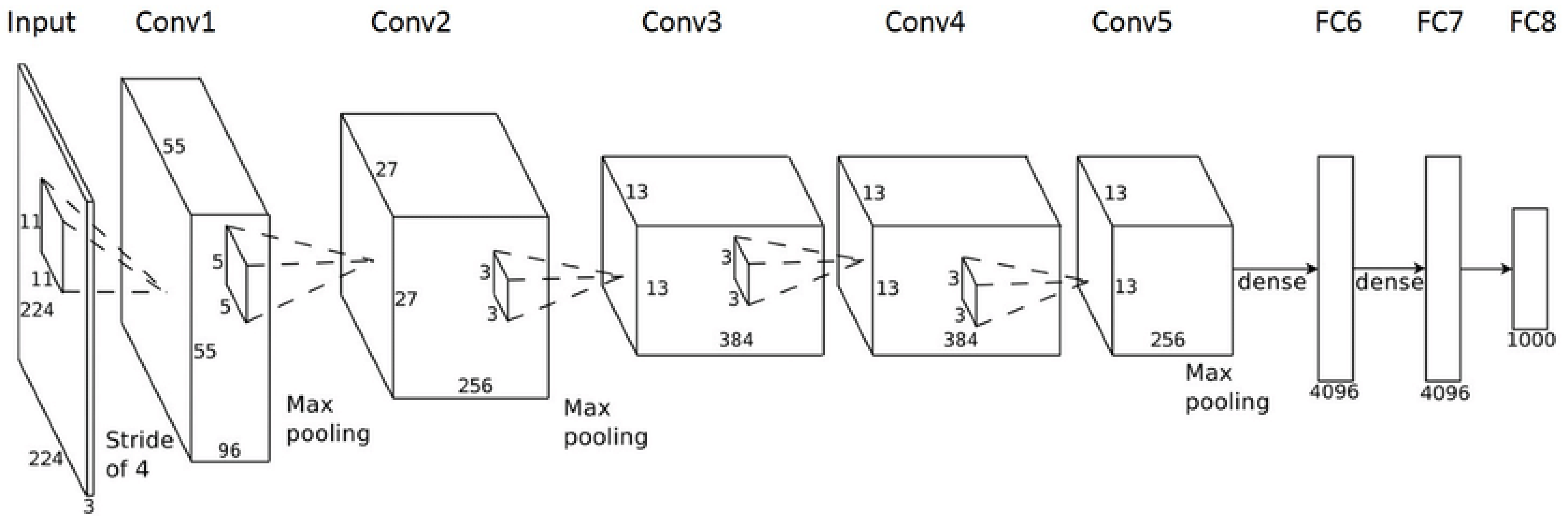
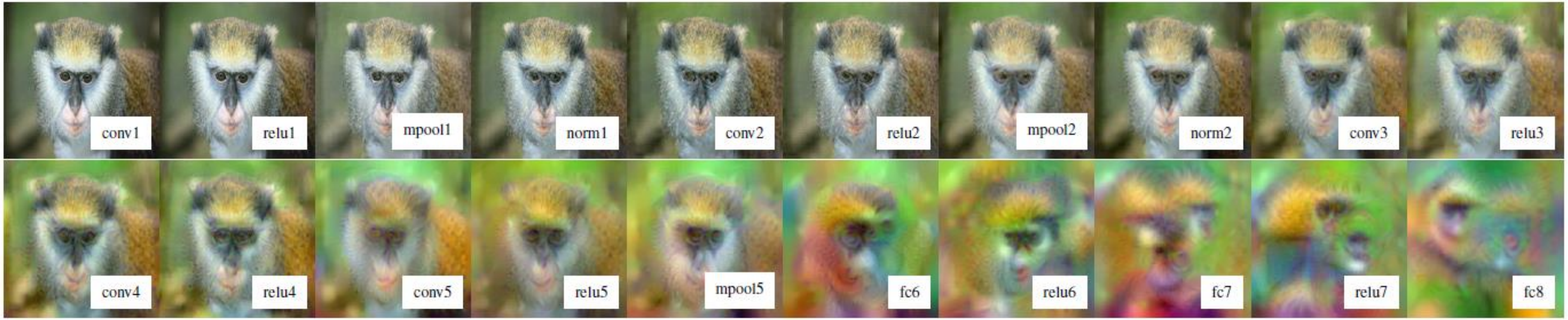
Training Convolutional Neural Network

Hyun Soo Park

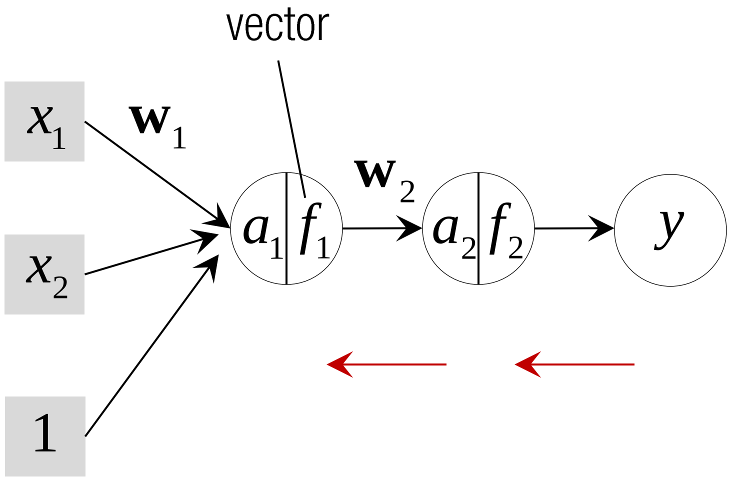


Input
Conv1
Conv2
Conv3
Conv4
Conv5
FC6
FC7
FC8
11
11
224
224
3
Stride of 4
55
5
5
55
96
Max pooling
27
3
3
27
256
Max pooling
13
3
3
13
384
13
3
3
13
384
13
13
256
Max pooling
dense
dense
4096
4096
1000



conv1
relu1
mpool1
norm1
conv2
relu2
mpool2
norm2
conv3
relu3
conv4
relu4
conv5
relu5
mpool5
fc6
relu6
fc7
relu7
fc8

Recall: Back-propagation Algorithm w/ Stochastic Gradient Descent



$$f_2 = \sigma\left(\mathbf{w}_2 \cdot \sigma\left(\mathbf{w}_1 \cdot \mathbf{x}\right)\right)$$

While until converges:

Sample mini-batch

For each data sample in mini-batch,

1. Prediction

$$f_2 = \sigma\left(\mathbf{w}_2 \cdot \sigma\left(\mathbf{w}_1 \cdot \mathbf{x}\right)\right)$$

2. Measure error

$$L(\mathbf{w}_1, \mathbf{w}_2) = \sum_i \left(\tilde{y}^i - y^i\right)^2$$

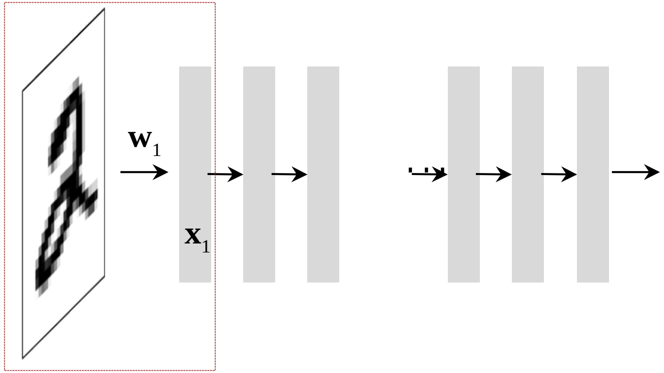
3. Back-propagation

$$\frac{\partial L}{\partial \mathbf{w}_2} += \frac{\partial L}{\partial f_2}\frac{\partial f_2}{\partial a_2}\frac{\partial a_2}{\partial \mathbf{w}_2} \qquad \frac{\partial L}{\partial \mathbf{w}_1} += \frac{\partial L}{\partial f_2}\frac{\partial f_2}{\partial a_2}\frac{\partial a_2}{\partial f_1}\frac{\partial f_1}{\partial a_1}\frac{\partial a_1}{\partial \mathbf{w}_1}$$

Update gradient

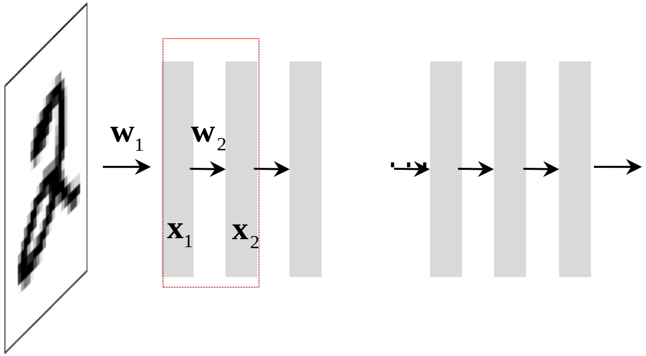
$$\mathbf{w}_1 = \mathbf{w}_1 - \gamma \frac{\partial L}{\partial \mathbf{w}_1} \qquad \mathbf{w}_2 = \mathbf{w}_2 - \gamma \frac{\partial L}{\partial \mathbf{w}_2}$$

PREDICTION



$$\mathbf{x}_1 = f(I;\mathbf{w}_1)$$

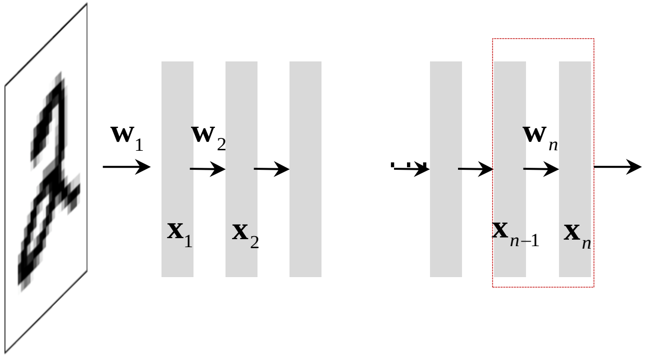
PREDICTION



$$\mathbf{x}_1 = f(I; \mathbf{w}_1)$$

$$\mathbf{x}_2 = f(\mathbf{x}_1; \mathbf{w}_2)$$

PREDICTION

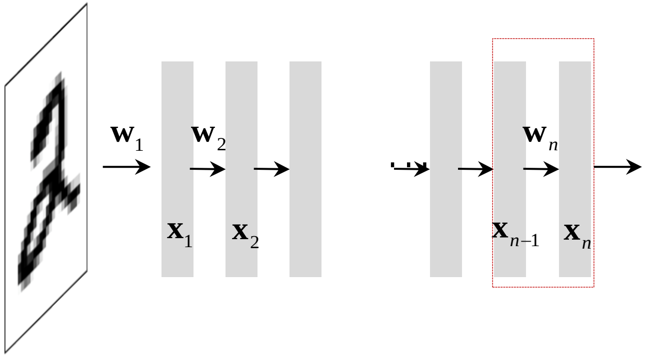


$$\mathbf{x}_1 = f(I;\mathbf{w}_1)$$

$$\mathbf{x}_2 = f(\mathbf{x}_1;\mathbf{w}_2)$$

$$\mathbf{x} = f(\mathbf{x}_{n-1};\mathbf{w}_n)$$

PREDICTION



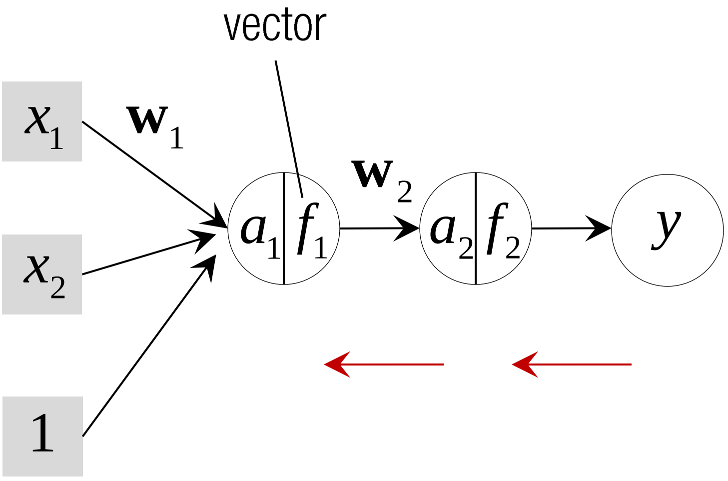
$$\mathbf{x}_1 = f(I; \mathbf{w}_1)$$

$$\mathbf{x}_2 = f(\mathbf{x}_1; \mathbf{w}_2)$$

$$\mathbf{x} = f(\mathbf{x}_{n-1}; \mathbf{w}_n)$$

function y = foo(x)

RECALL: BACK-PROPAGATION ALGORITHM W/ STOCHASTIC GRADIENT DESCENT



$$f_2 = \sigma\left(\mathbf{w}_2 \cdot \sigma\left(\mathbf{w}_1 \cdot \mathbf{x}\right)\right)$$

While until converges:

Sample mini-batch

For each data sample in mini-batch,

1. Prediction

$$f_2 = \sigma\left(\mathbf{w}_2 \cdot \sigma\left(\mathbf{w}_1 \cdot \mathbf{x}\right)\right)$$

2. Measure error

$$L(\mathbf{w}_1, \mathbf{w}_2) = \sum_i \left(\tilde{y}^i - y^i\right)^2$$

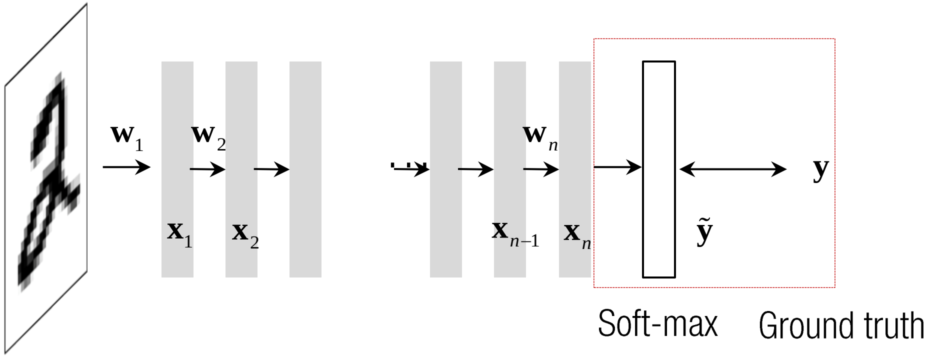
3. Back-propagation

$$\frac{\partial L}{\partial \mathbf{w}_2} += \frac{\partial L}{\partial f_2}\frac{\partial f_2}{\partial a_2}\frac{\partial a_2}{\partial \mathbf{w}_2} \qquad \frac{\partial L}{\partial \mathbf{w}_1} += \frac{\partial L}{\partial f_2}\frac{\partial f_2}{\partial a_2}\frac{\partial a_2}{\partial f_1}\frac{\partial f_1}{\partial a_1}\frac{\partial a_1}{\partial \mathbf{w}_1}$$

Update gradient

$$\mathbf{w}_1 = \mathbf{w}_1 - \gamma \frac{\partial L}{\partial \mathbf{w}_1} \qquad \mathbf{w}_2 = \mathbf{w}_2 - \gamma \frac{\partial L}{\partial \mathbf{w}_2}$$

Loss



$$\tilde{\mathbf{y}}_i = \frac{e^{\mathbf{x}_i}}{\sum_i e^{\mathbf{x}_i}}$$

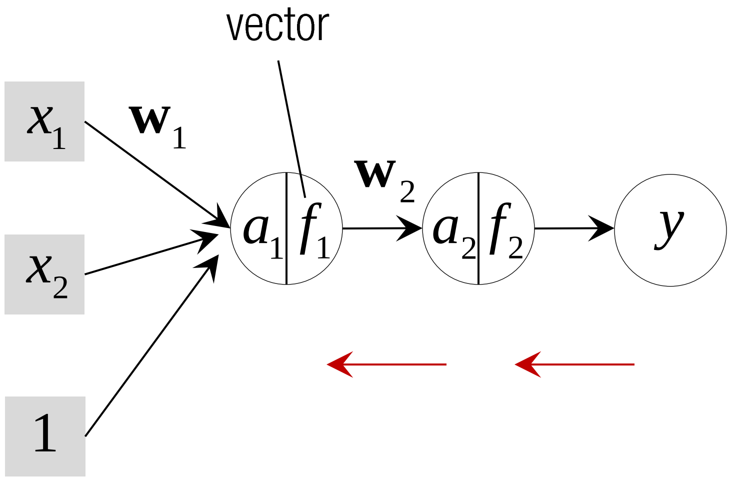
Cross-entropy loss

$$L = -\sum_i \mathbf{y}_i \log \tilde{\mathbf{y}}_i$$

function y = soft-max(x)

function l, dl_dy = loss_cross_entropy_softmax (x, y_gt)

RECALL: BACK-PROPAGATION ALGORITHM W/ STOCHASTIC GRADIENT DESCENT



$$f_2 = \sigma\left(\mathbf{w}_2 \cdot \sigma\left(\mathbf{w}_1 \cdot \mathbf{x}\right)\right)$$

While until converges:

Sample mini-batch

For each data sample in mini-batch,

1. Prediction

$$f_2 = \sigma\left(\mathbf{w}_2 \cdot \sigma\left(\mathbf{w}_1 \cdot \mathbf{x}\right)\right)$$

2. Measure error

$$L(\mathbf{w}_1, \mathbf{w}_2) = \sum_i \left(\tilde{y}^i - y^i\right)^2$$

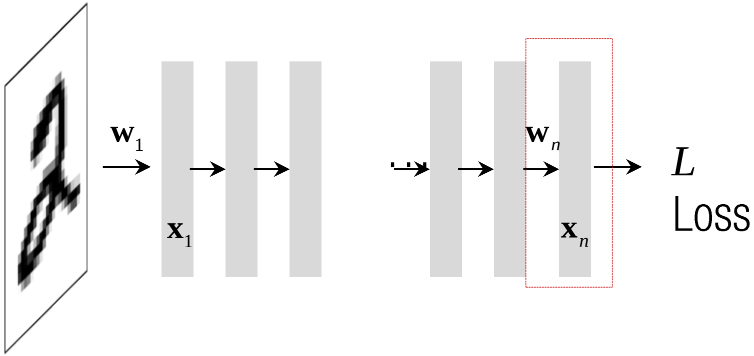
3. Back-propagation

$$\frac{\partial L}{\partial \mathbf{w}_2} += \frac{\partial L}{\partial f_2}\frac{\partial f_2}{\partial a_2}\frac{\partial a_2}{\partial \mathbf{w}_2} \qquad \frac{\partial L}{\partial \mathbf{w}_1} += \frac{\partial L}{\partial f_2}\frac{\partial f_2}{\partial a_2}\frac{\partial a_2}{\partial f_1}\frac{\partial f_1}{\partial a_1}\frac{\partial a_1}{\partial \mathbf{w}_1}$$

Update gradient

$$\mathbf{w}_1 = \mathbf{w}_1 - \gamma \frac{\partial L}{\partial \mathbf{w}_1} \qquad \mathbf{w}_2 = \mathbf{w}_2 - \gamma \frac{\partial L}{\partial \mathbf{w}_2}$$

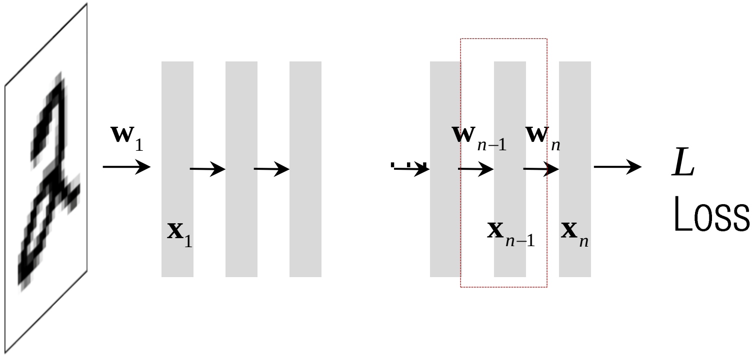
BACK-PROPAGATION



$$\frac{\partial L}{\partial \mathbf{w}_n} = \frac{\partial L}{\partial \mathbf{x}_n} \frac{\partial \mathbf{x}_n}{\partial \mathbf{w}_n}$$

$$\mathbf{w}_n = \mathbf{w}_n - \gamma \frac{\partial L}{\partial \mathbf{w}_n}$$

BACK-PROPAGATION



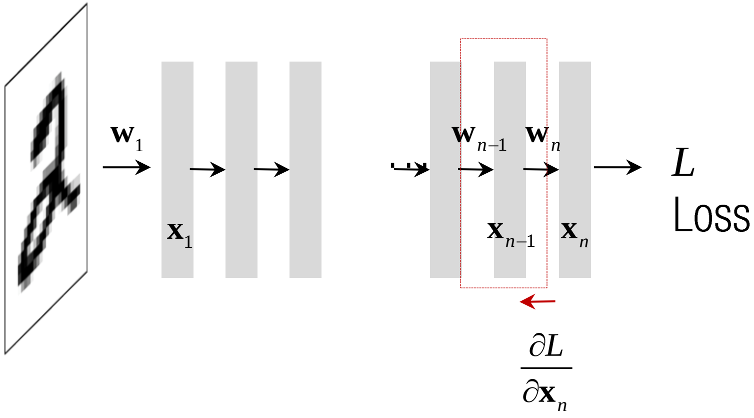
$$\frac{\partial L}{\partial \mathbf{w}_n} = \frac{\partial L}{\partial \mathbf{x}_n} \frac{\partial \mathbf{x}_n}{\partial \mathbf{w}_n}$$

$$\mathbf{w}_n = \mathbf{w}_n - \gamma \frac{\partial L}{\partial \mathbf{w}_n}$$

$$\frac{\partial L}{\partial \mathbf{w}_{n-1}} = \frac{\partial L}{\partial \mathbf{x}_n} \frac{\partial \mathbf{x}_n}{\partial \mathbf{x}_{n-1}} \frac{\partial \mathbf{x}_{n-1}}{\partial \mathbf{w}_{n-1}} = \frac{\partial L}{\partial \mathbf{x}_{n-1}} \frac{\partial \mathbf{x}_{n-1}}{\partial \mathbf{w}_{n-1}}$$

$$\mathbf{w}_{n-1} = \mathbf{w}_{n-1} - \gamma \frac{\partial L}{\partial \mathbf{w}_{n-1}}$$

BACK-PROPAGATION



$$\frac{\partial L}{\partial \mathbf{w}_n} = \boxed{\frac{\partial L}{\partial \mathbf{x}_n}} \frac{\partial \mathbf{x}_n}{\partial \mathbf{w}_n}$$

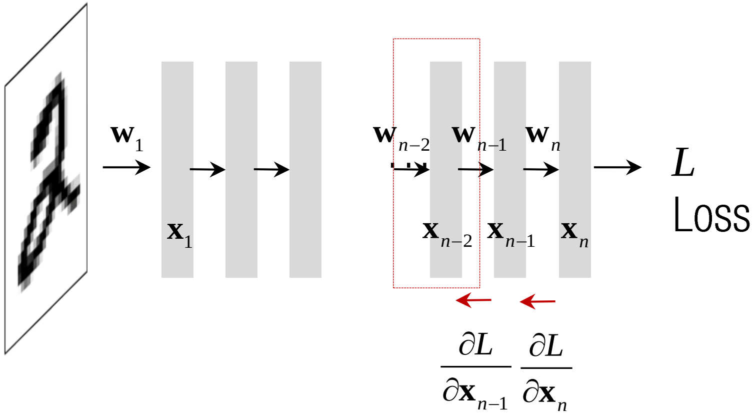
$$\frac{\partial L}{\partial \mathbf{w}_{n-1}} = \boxed{\frac{\partial L}{\partial \mathbf{x}_n}} \frac{\partial \mathbf{x}_n}{\partial \mathbf{x}_{n-1}} \frac{\partial \mathbf{x}_{n-1}}{\partial \mathbf{w}_{n-1}}$$

$$= \frac{\partial L}{\partial \mathbf{x}_{n-1}} \frac{\partial \mathbf{x}_{n-1}}{\partial \mathbf{w}_{n-1}}$$

$$\mathbf{w}_n = \mathbf{w}_n - \gamma \frac{\partial L}{\partial \mathbf{w}_n}$$

$$\mathbf{w}_{n-1} = \mathbf{w}_{n-1} - \gamma \frac{\partial L}{\partial \mathbf{w}_{n-1}}$$

Back-propagation



$$\frac{\partial L}{\partial \mathbf{w}_n} = \frac{\partial L}{\partial \mathbf{x}_n}\frac{\partial \mathbf{x}_n}{\partial \mathbf{w}_n}$$

$$\mathbf{w}_n = \mathbf{w}_n - \gamma \frac{\partial L}{\partial \mathbf{w}_n}$$

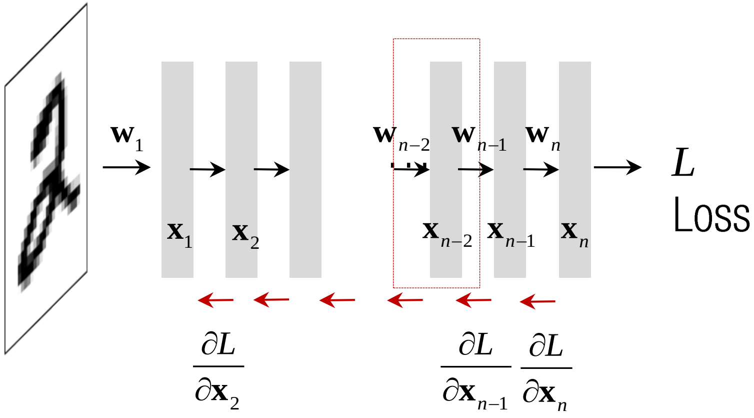
$$\frac{\partial L}{\partial \mathbf{w}_{n-1}} = \frac{\partial L}{\partial \mathbf{x}_n}\frac{\partial \mathbf{x}_n}{\partial \mathbf{x}_{n-1}}\frac{\partial \mathbf{x}_{n-1}}{\partial \mathbf{w}_{n-1}} = \boxed{\frac{\partial L}{\partial \mathbf{x}_{n-1}}}\frac{\partial \mathbf{x}_{n-1}}{\partial \mathbf{w}_{n-1}}$$

$$\mathbf{w}_{n-1} = \mathbf{w}_{n-1} - \gamma \frac{\partial L}{\partial \mathbf{w}_{n-1}}$$

$$\frac{\partial L}{\partial \mathbf{w}_{n-2}} = \boxed{\frac{\partial L}{\partial \mathbf{x}_n}\frac{\partial \mathbf{x}_n}{\partial \mathbf{x}_{n-1}}}\frac{\partial \mathbf{x}_{n-1}}{\partial \mathbf{x}_{n-2}}\frac{\partial \mathbf{x}_{n-2}}{\partial \mathbf{w}_{n-2}} = \frac{\partial L}{\partial \mathbf{x}_{n-2}}\frac{\partial \mathbf{x}_{n-2}}{\partial \mathbf{w}_{n-2}}$$

$$\mathbf{w}_{n-2} = \mathbf{w}_{n-2} - \gamma \frac{\partial L}{\partial \mathbf{w}_{n-2}}$$

BACK-PROPAGATION



$$\frac{\partial L}{\partial \mathbf{w}_n} = \frac{\partial L}{\partial \mathbf{x}_n} \frac{\partial \mathbf{x}_n}{\partial \mathbf{w}_n}$$

$$\mathbf{w}_n = \mathbf{w}_n - \gamma \frac{\partial L}{\partial \mathbf{w}_n}$$

$$\frac{\partial L}{\partial \mathbf{w}_{n-1}} = \frac{\partial L}{\partial \mathbf{x}_n} \frac{\partial \mathbf{x}_n}{\partial \mathbf{x}_{n-1}} \frac{\partial \mathbf{x}_{n-1}}{\partial \mathbf{w}_{n-1}} = \frac{\partial L}{\partial \mathbf{x}_{n-1}} \frac{\partial \mathbf{x}_{n-1}}{\partial \mathbf{w}_{n-1}}$$

$$\mathbf{w}_{n-1} = \mathbf{w}_{n-1} - \gamma \frac{\partial L}{\partial \mathbf{w}_{n-1}}$$

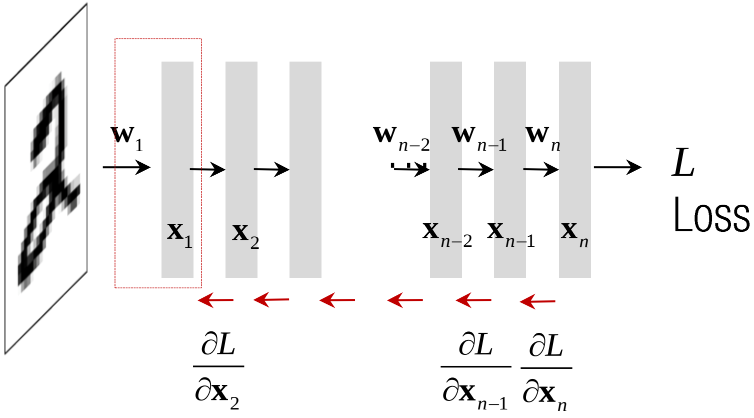
$$\frac{\partial L}{\partial \mathbf{w}_{n-2}} = \frac{\partial L}{\partial \mathbf{x}_n} \frac{\partial \mathbf{x}_n}{\partial \mathbf{x}_{n-1}} \frac{\partial \mathbf{x}_{n-1}}{\partial \mathbf{x}_{n-2}} \frac{\partial \mathbf{x}_{n-2}}{\partial \mathbf{w}_{n-2}} = \frac{\partial L}{\partial \mathbf{x}_{n-2}} \frac{\partial \mathbf{x}_{n-2}}{\partial \mathbf{w}_{n-2}}$$

$$\mathbf{w}_{n-2} = \mathbf{w}_{n-2} - \gamma \frac{\partial L}{\partial \mathbf{w}_{n-2}}$$

$$\frac{\partial L}{\partial \mathbf{w}_1} = \frac{\partial L}{\partial \mathbf{x}_n} \frac{\partial \mathbf{x}_n}{\partial \mathbf{x}_{n-1}} \cdots \frac{\partial \mathbf{x}_2}{\partial \mathbf{x}_1} \frac{\partial \mathbf{x}_1}{\partial \mathbf{w}_1} = \frac{\partial L}{\partial \mathbf{x}_1} \frac{\partial \mathbf{x}_1}{\partial \mathbf{w}_1}$$

$$\mathbf{w}_1 = \mathbf{w}_1 - \gamma \frac{\partial L}{\partial \mathbf{w}_1}$$

BACK-PROPAGATION



$$\frac{\partial L}{\partial \mathbf{w}_n} = \frac{\partial L}{\partial \mathbf{x}_n}\frac{\partial \mathbf{x}_n}{\partial \mathbf{w}_n} \qquad \mathbf{w}_n = \mathbf{w}_n - \gamma \frac{\partial L}{\partial \mathbf{w}_n}$$

$$\frac{\partial L}{\partial \mathbf{w}_{n-1}} = \frac{\partial L}{\partial \mathbf{x}_n}\frac{\partial \mathbf{x}_n}{\partial \mathbf{x}_{n-1}}\frac{\partial \mathbf{x}_{n-1}}{\partial \mathbf{w}_{n-1}} = \frac{\partial L}{\partial \mathbf{x}_{n-1}}\frac{\partial \mathbf{x}_{n-1}}{\partial \mathbf{w}_{n-1}} \qquad \mathbf{w}_{n-1} = \mathbf{w}_{n-1} - \gamma \frac{\partial L}{\partial \mathbf{w}_{n-1}}$$

$$\frac{\partial L}{\partial \mathbf{w}_{n-2}} = \frac{\partial L}{\partial \mathbf{x}_n}\frac{\partial \mathbf{x}_n}{\partial \mathbf{x}_{n-1}}\frac{\partial \mathbf{x}_{n-1}}{\partial \mathbf{x}_{n-2}}\frac{\partial \mathbf{x}_{n-2}}{\partial \mathbf{w}_{n-2}} = \frac{\partial L}{\partial \mathbf{x}_{n-2}}\frac{\partial \mathbf{x}_{n-2}}{\partial \mathbf{w}_{n-2}} \qquad \mathbf{w}_{n-2} = \mathbf{w}_{n-2} - \gamma \frac{\partial L}{\partial \mathbf{w}_{n-2}}$$

$$\frac{\partial L}{\partial \mathbf{w}_1} = \frac{\partial L}{\partial \mathbf{x}_n}\frac{\partial \mathbf{x}_n}{\partial \mathbf{x}_{n-1}}\cdots\frac{\partial \mathbf{x}_2}{\partial \mathbf{x}_1}\frac{\partial \mathbf{x}_1}{\partial \mathbf{w}_1} = \frac{\partial L}{\partial \mathbf{x}_1}\frac{\partial \mathbf{x}_1}{\partial \mathbf{w}_1} \qquad \mathbf{w}_1 = \mathbf{w}_1 - \gamma \frac{\partial L}{\partial \mathbf{w}_1}$$

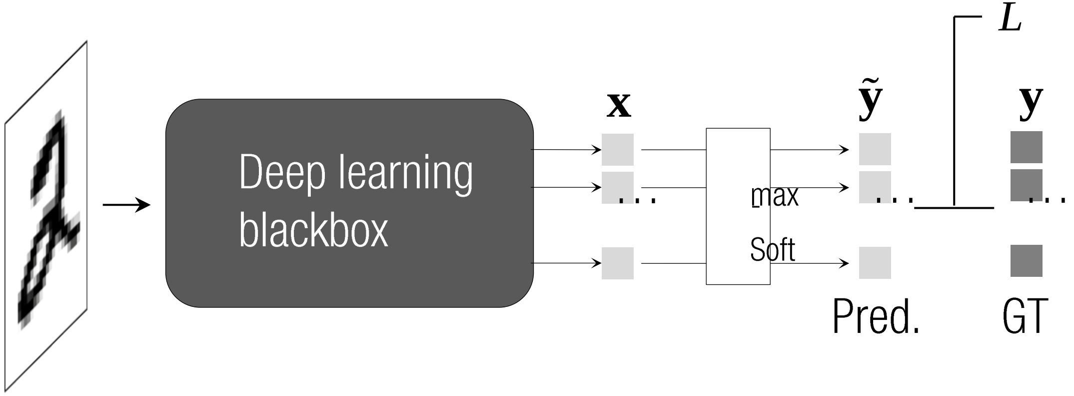
function y = foo(x)

function [dLdx dLdw dLdb] = foo_back(dLdy, x, y)

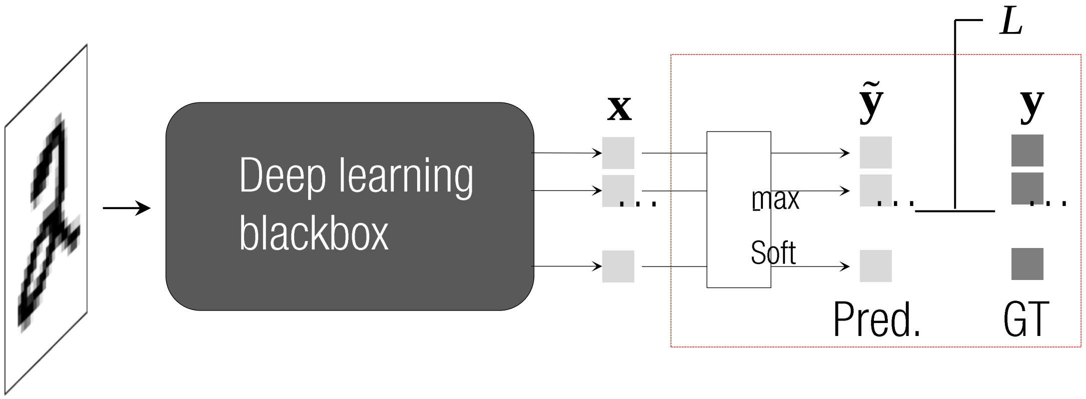
Weight update

Loss propagation

ENTROPY LOSS DERIVATIVE



ENTROPY LOSS DERIVATIVE

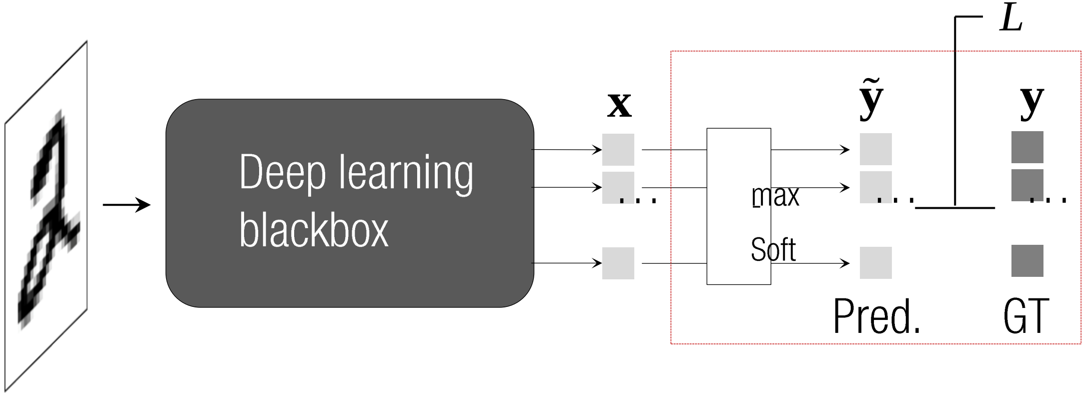


Input:

Trainable var.:

Output:

ENTROPY LOSS DERIVATIVE

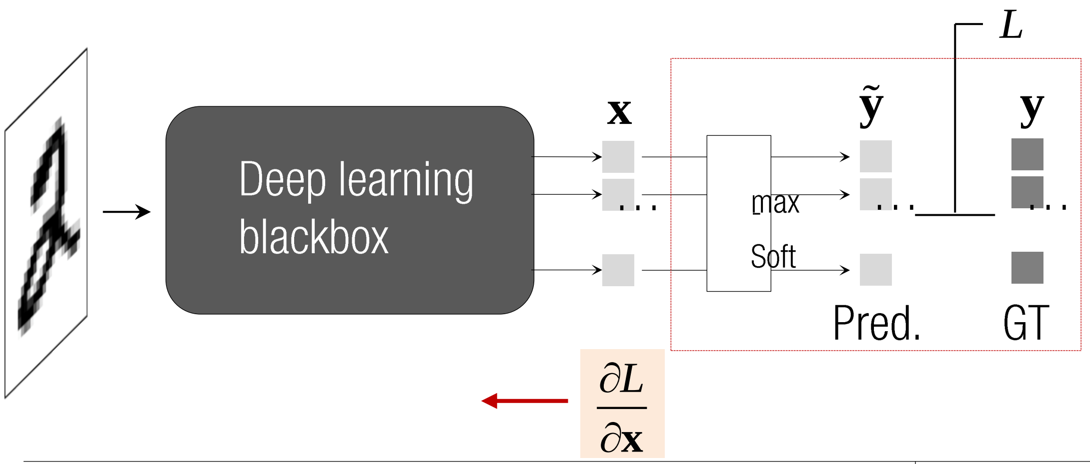


Input: $\mathbf{x}$

Trainable var.: None

Output: $L = \sum_i \mathbf{y}_i \log \tilde{\mathbf{y}}_i$ where $\tilde{\mathbf{y}}_i = \dfrac{e^{\mathbf{x}_i}}{\sum_i e^{\mathbf{x}_i}}$

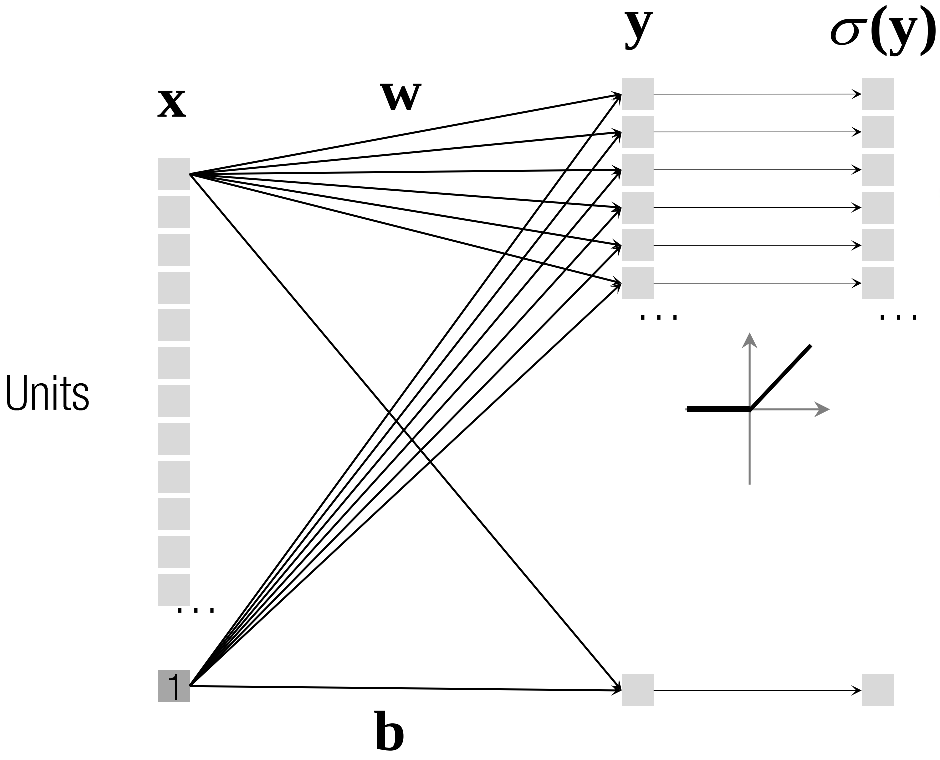
Entropy Loss Derivative

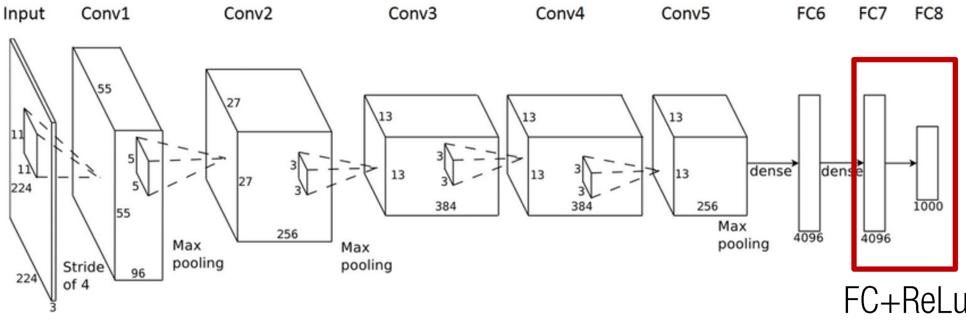


function [L, dLdx] = Loss(x, y)

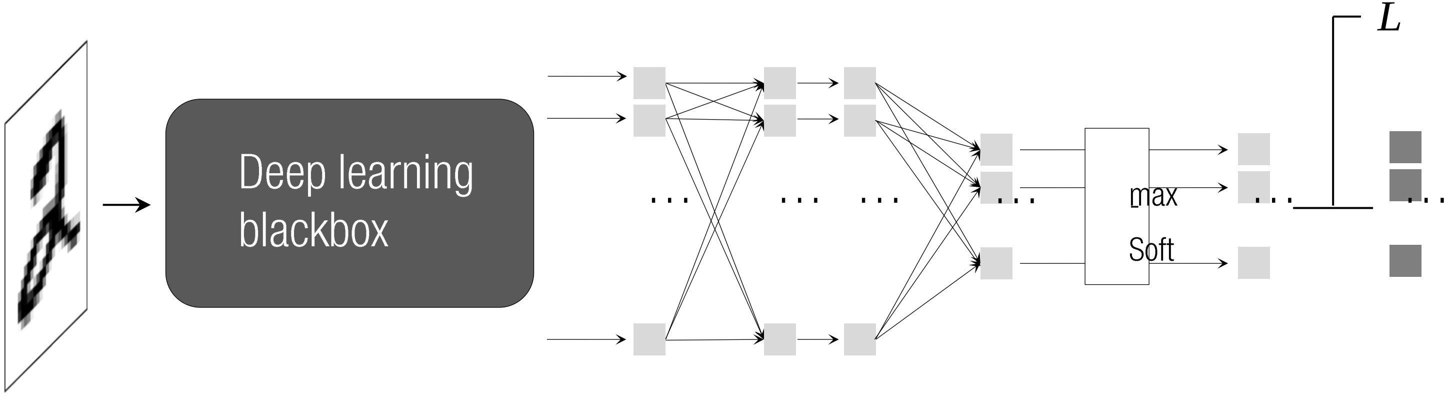
Input: $\mathbf{x}$	$\frac{\partial L}{\partial \mathbf{x}_i} = \tilde{\mathbf{y}}_i - \mathbf{y}_i$ 1 x n
Trainable var.: None	None
Output: $L = \sum_i \mathbf{y}_i \log \tilde{\mathbf{y}}_i$ where $\tilde{\mathbf{y}}_i = \frac{e^{\mathbf{x}_i}}{\sum_i e^{\mathbf{x}_i}}$	

RECALL: FC+RELU

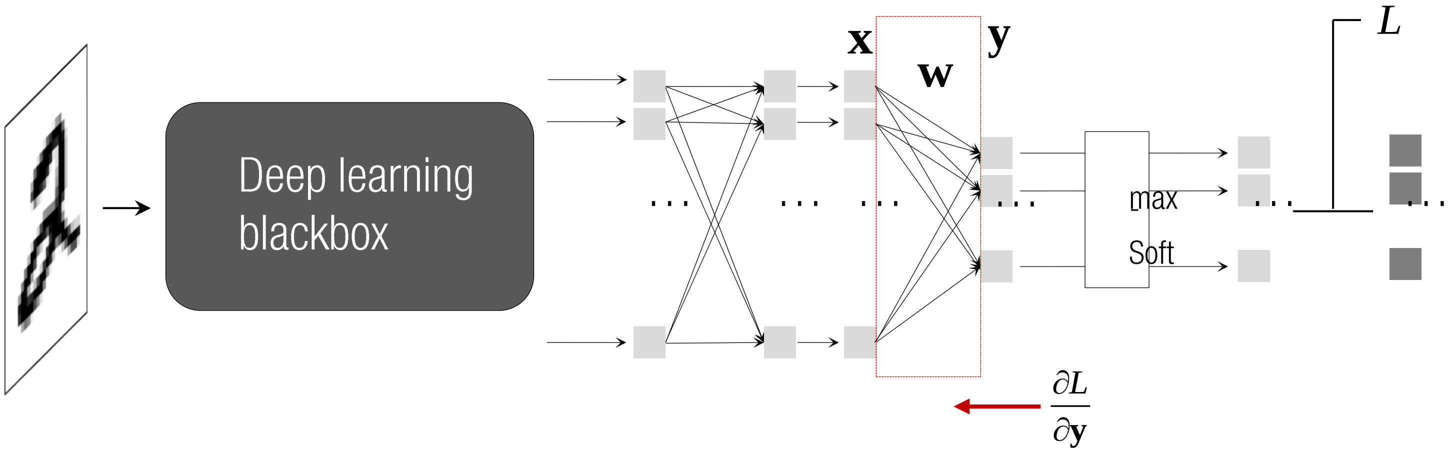




FULLY CONNECTED LAYER



Fully Connected Layer



Input: $\mathbf{x} \in \mathbb{R}^n$

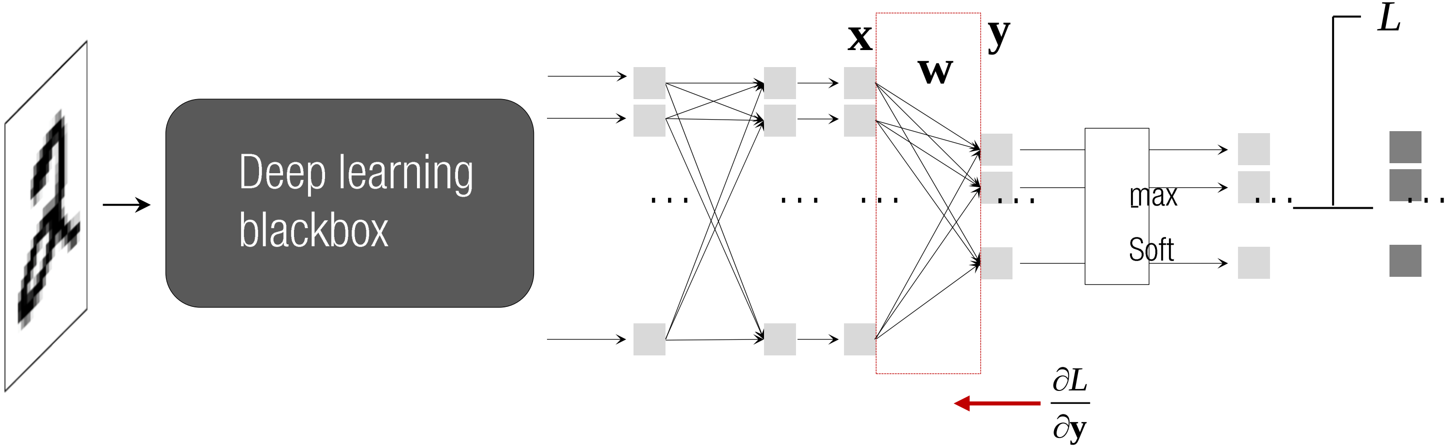
Trainable var.: $\mathbf{w} \in \mathbb{R}^{m \times n}$

Output: $\mathbf{y} = \mathbf{w}\mathbf{x}$

$\mathbf{y} \in \mathbb{R}^m$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_1} & \cdots & \frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{y}_m}{\partial \mathbf{x}_1} & \cdots & \frac{\partial \mathbf{y}_m}{\partial \mathbf{x}_n} \end{bmatrix} = \mathbf{w} \longrightarrow \frac{\partial L}{\partial \mathbf{x}} = \frac{\partial L}{\partial \mathbf{y}} \mathbf{w}$$

FULLY CONNECTED LAYER



Input: $\mathbf{x} \in \mathbb{R}^n$

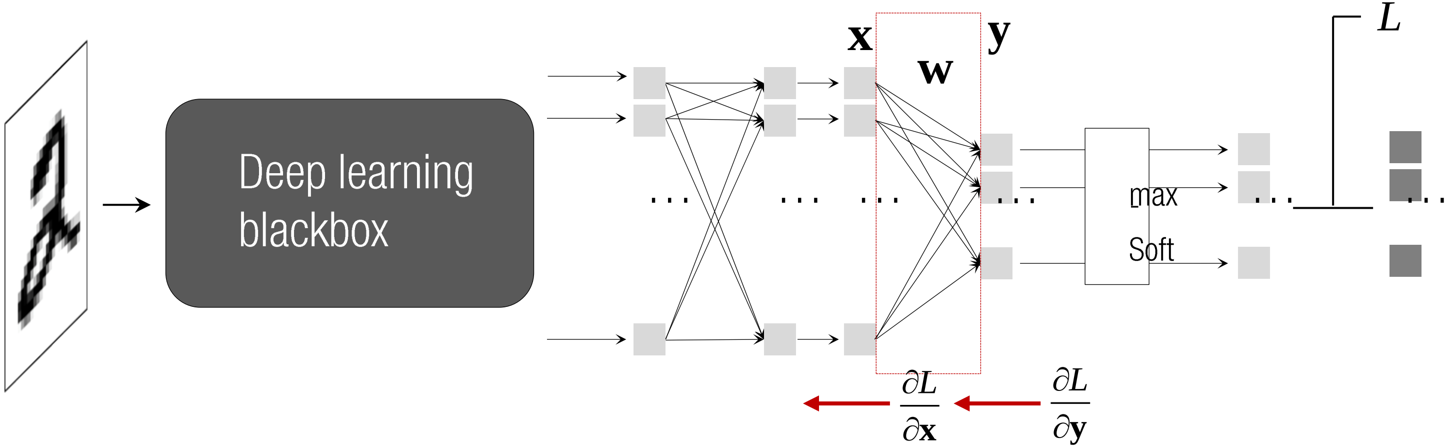
Trainable var.: $\mathbf{w} \in \mathbb{R}^{m \times n}$

Output: $\mathbf{y} = \mathbf{w}\mathbf{x}$

$\mathbf{y} \in \mathbb{R}^m$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_1} & \cdots & \frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{y}_m}{\partial \mathbf{x}_1} & \cdots & \frac{\partial \mathbf{y}_m}{\partial \mathbf{x}_n} \end{bmatrix} = \mathbf{w} \longrightarrow \frac{\partial L}{\partial \mathbf{x}} = \underbrace{\frac{\partial L}{\partial \mathbf{y}} \mathbf{w}}_{1\text{xn}}$$

FULLY CONNECTED LAYER



Input: $\mathbf{x} \in \mathbb{R}^n$

Trainable var.: $\mathbf{w} \in \mathbb{R}^{m \times n}$

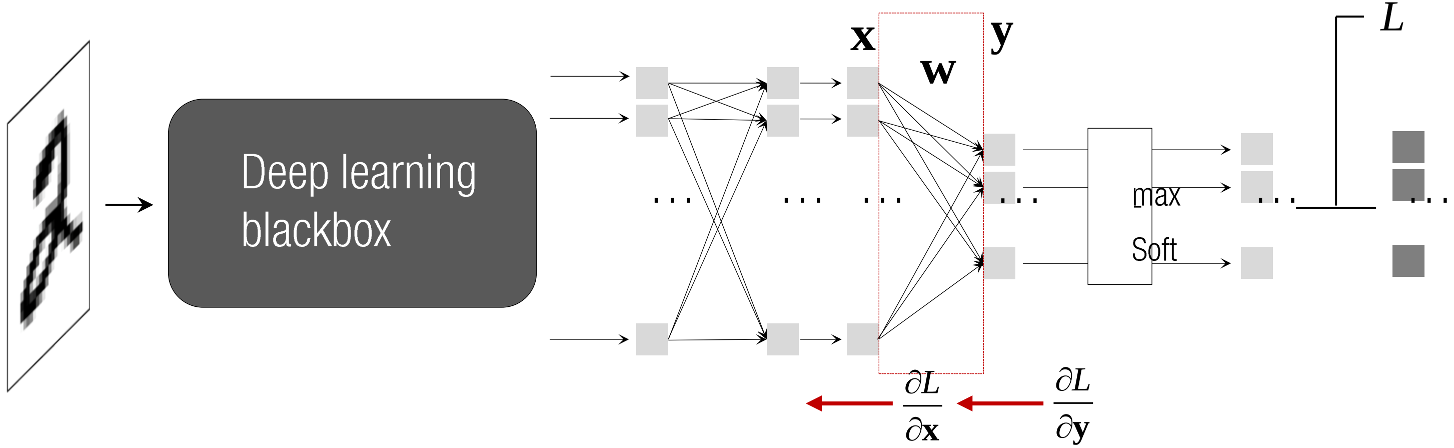
Output: $\mathbf{y} = \mathbf{w}\mathbf{x}$

$\mathbf{y} \in \mathbb{R}^m$

$$\frac{\partial L}{\partial \mathbf{x}_i} = \sum_j \frac{\partial L}{\partial \mathbf{y}_j} \frac{\partial \mathbf{y}_j}{\partial \mathbf{x}_i} =$$

1xn

Fully Connected Layer



$$\frac{\partial L}{\partial \mathbf{x}_i} = \sum_j \frac{\partial L}{\partial \mathbf{y}_j} \frac{\partial \mathbf{y}_j}{\partial \mathbf{x}_i} = \sum_j \frac{\partial L}{\partial \mathbf{y}_j} \mathbf{w}_{ji}$$

1xn

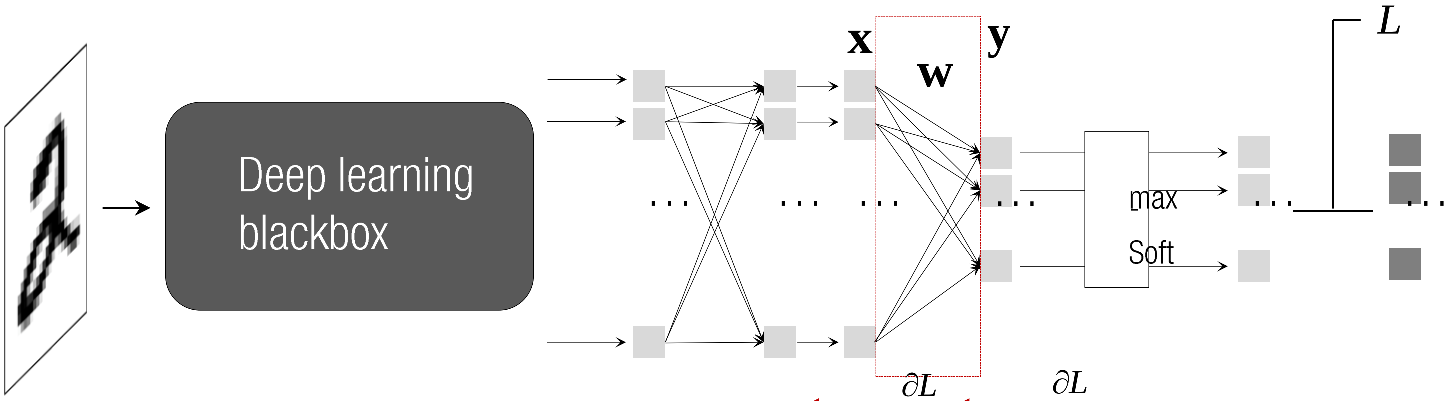
Input: $\mathbf{x} \in \mathbb{R}^n$

Trainable var.: $\mathbf{w} \in \mathbb{R}^{m \times n}$

Output: $\mathbf{y} = \mathbf{w}\mathbf{x}$

$\mathbf{y} \in \mathbb{R}^m$

Fully Connected Layer



$$\frac{\partial L}{\partial \mathbf{x}}$$ $$\frac{\partial L}{\partial \mathbf{y}}$$

Input: $\mathbf{x} \in \mathbb{R}^n$

Trainable var.: $\mathbf{w} \in \mathbb{R}^{m \times n}$

Output: $\mathbf{y} = \mathbf{wx}$

$\mathbf{y} \in \mathbb{R}^m$

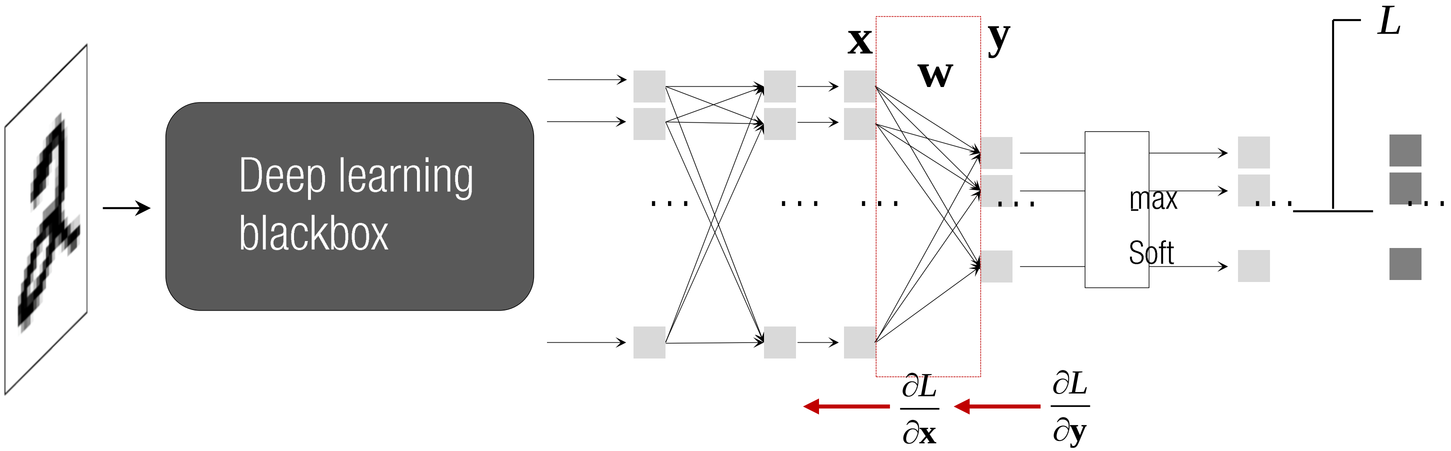
$$\frac{\partial L}{\partial \mathbf{x}_i} = \sum_j \frac{\partial L}{\partial \mathbf{y}_j} \frac{\partial \mathbf{y}_j}{\partial \mathbf{x}_i} = \sum_j \frac{\partial L}{\partial \mathbf{y}_j} \mathbf{w}_{ji}$$

1xn

$$\frac{\partial L}{\partial \mathbf{w}_{ij}} =$$

$$\mathbf{y}_i = \mathbf{w}_{i1}\mathbf{x}_1 + \cdots + \mathbf{w}_{im}\mathbf{x}_m$$

Fully Connected Layer



Input: $\mathbf{x} \in \mathbb{R}^n$

Trainable var.: $\mathbf{w} \in \mathbb{R}^{m \times n}$

Output: $\mathbf{y} = \mathbf{w}\mathbf{x}$

$\mathbf{y} \in \mathbb{R}^m$

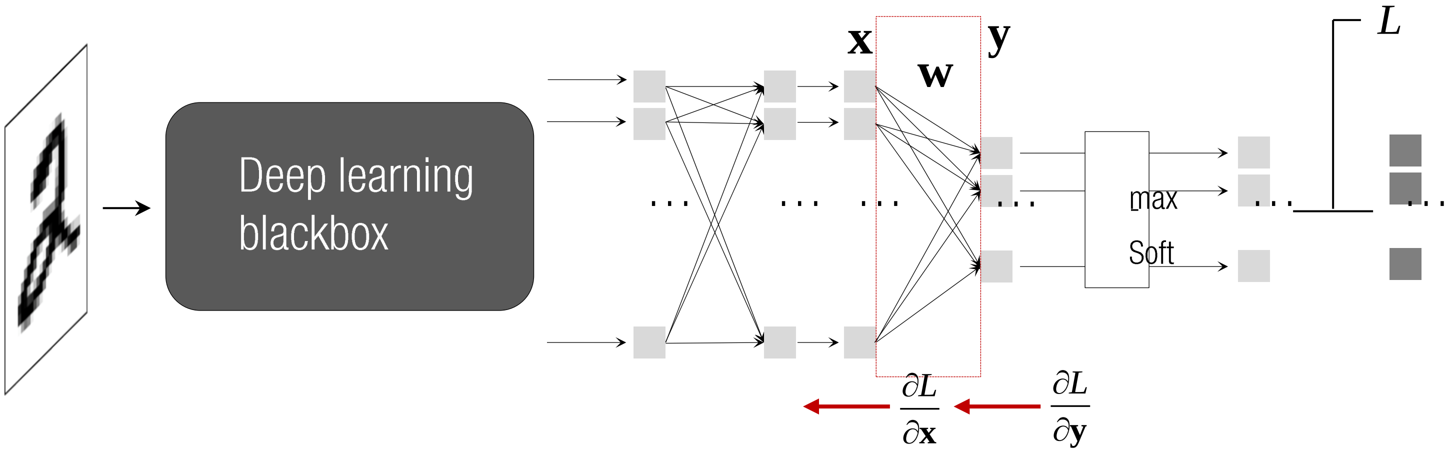
$$\frac{\partial L}{\partial \mathbf{x}_i} = \sum_j \frac{\partial L}{\partial \mathbf{y}_j} \frac{\partial \mathbf{y}_j}{\partial \mathbf{x}_i} = \sum_j \frac{\partial L}{\partial \mathbf{y}_j} \mathbf{w}_{ji}$$

1xn

$$\frac{\partial L}{\partial \mathbf{w}_{ij}} = \frac{\partial L}{\partial \mathbf{y}_i} \mathbf{x}_j$$

1x(nxm)

FULLY CONNECTED LAYER



Input: $\mathbf{x} \in \mathbb{R}^n$

Trainable var.: $\mathbf{w} \in \mathbb{R}^{m \times n}$

Output: $\mathbf{y} = \mathbf{w}\mathbf{x}$

$\mathbf{y} \in \mathbb{R}^m$

$$\frac{\partial L}{\partial \mathbf{x}_i} = \sum_j \frac{\partial L}{\partial \mathbf{y}_j}\frac{\partial \mathbf{y}_j}{\partial \mathbf{x}_i} = \sum_j \frac{\partial L}{\partial \mathbf{y}_j}\mathbf{w}_{ji}$$

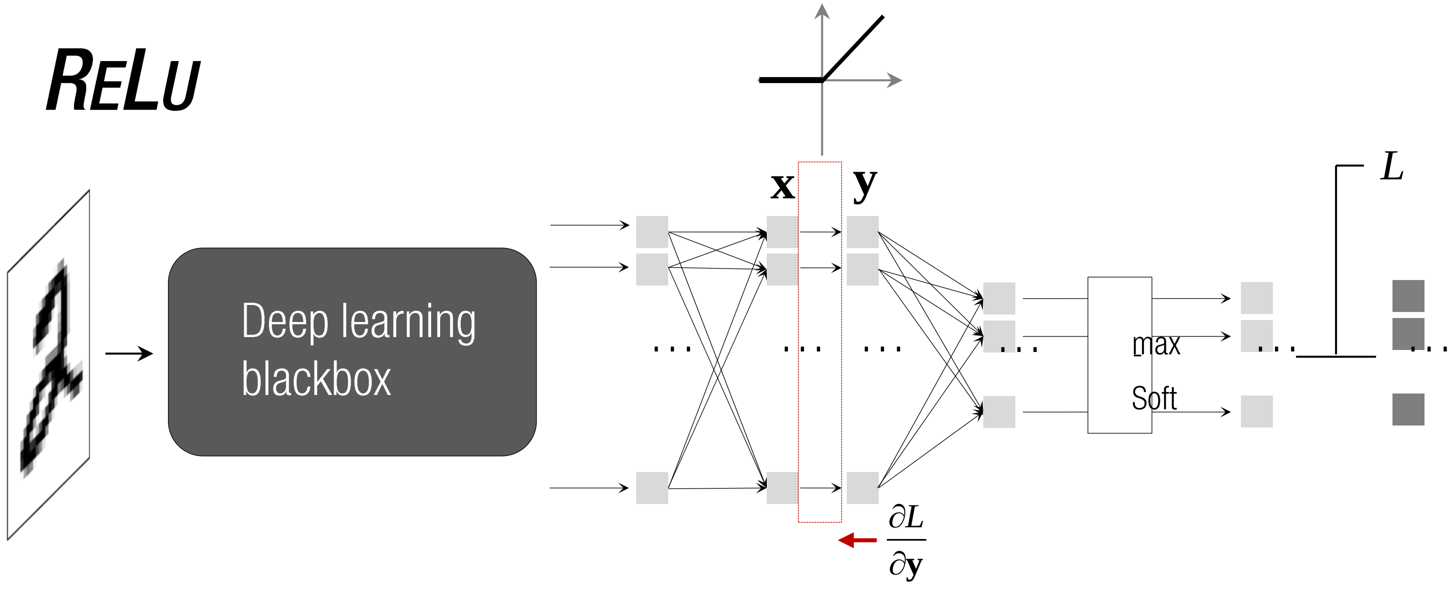
1xn

$$\frac{\partial L}{\partial \mathbf{w}_{ij}} = \frac{\partial L}{\partial \mathbf{y}_i}\mathbf{x}_j$$

1x(nxm)

function [y] = FC(x, w)
function [dLdx, dLdw, dLdb] = FC_back(dLdy, x, w, y)

ReLu

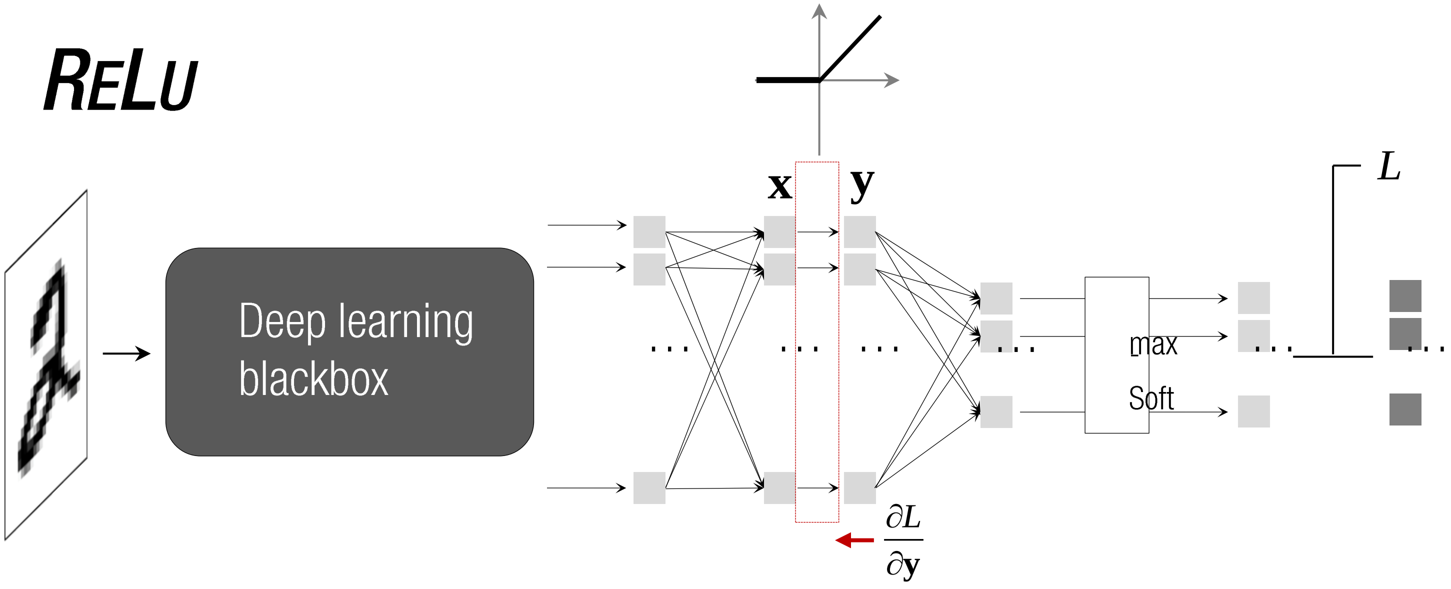


Input:

Trainable var.:

Output:

ReLu



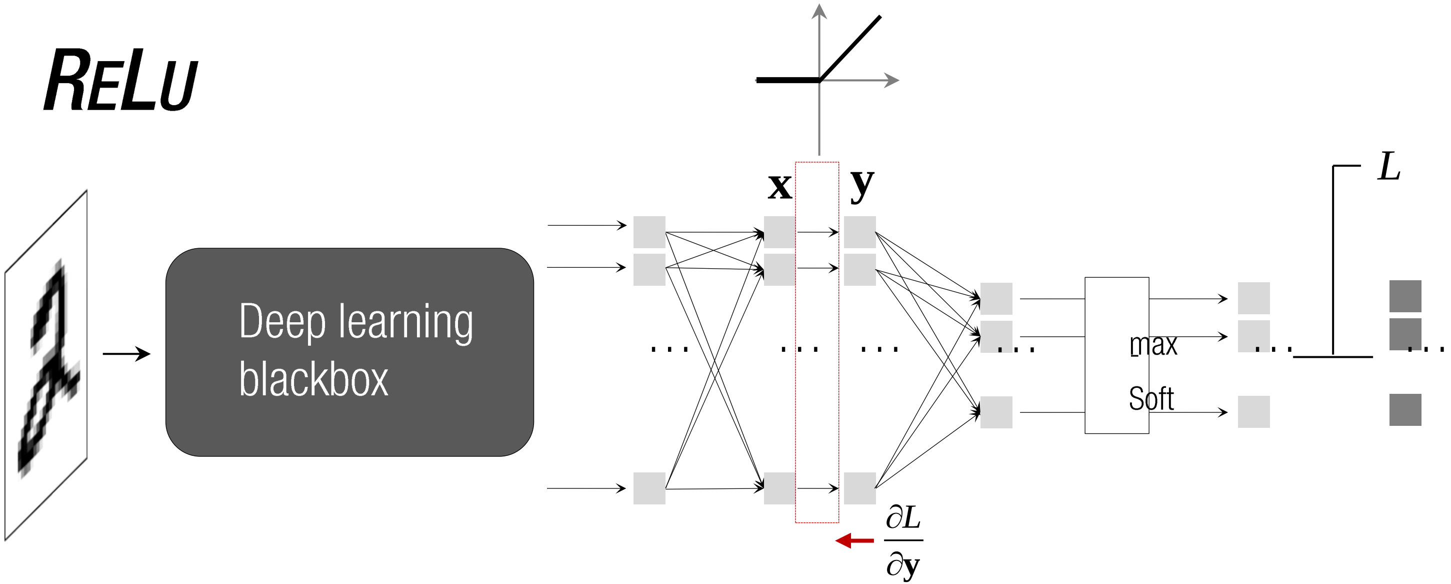
Input: $\mathbf{x} \in \mathbb{R}^n$

Trainable var.: None

Output: $\mathbf{y}_i = \max(0, \mathbf{x}_i)$

$\mathbf{y} \in \mathbb{R}^n$

ReLu



Input: $\mathbf{x} \in \mathbb{R}^n$

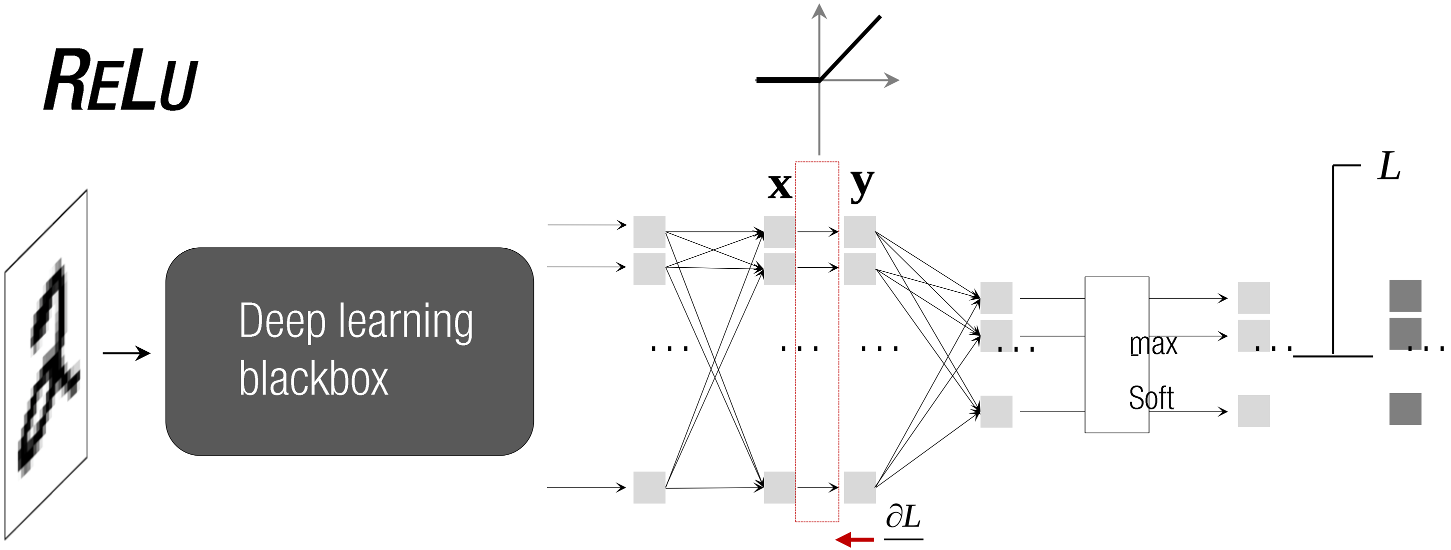
$$\frac{\partial L}{\partial \mathbf{x}_i} =$$

Trainable var.: None

Output: $\mathbf{y}_i = \max(0, \mathbf{x}_i)$

$\mathbf{y} \in \mathbb{R}^n$

ReLu



$$\frac{\partial L}{\partial \mathbf{y}}$$

$$\frac{\partial L}{\partial \mathbf{x}_i} = \sum_j \frac{\partial L}{\partial \mathbf{y}_j} \frac{\partial}{\partial \mathbf{x}_i} \max(\mathbf{0}, \mathbf{x}_j)$$

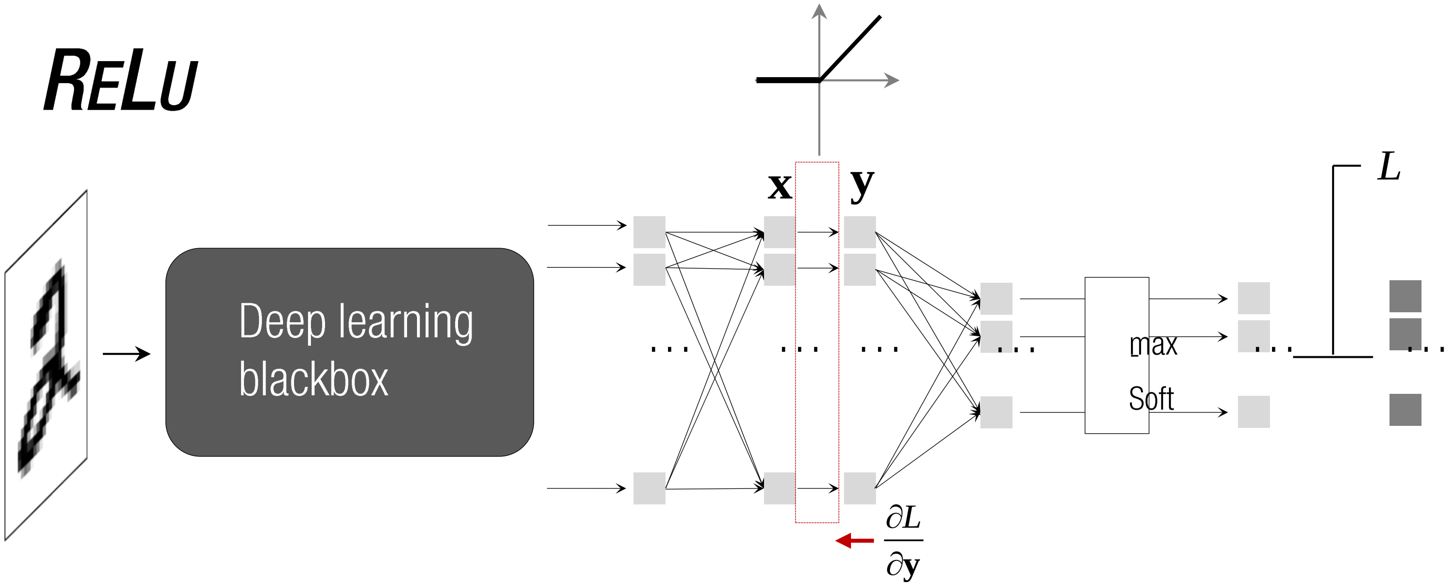
Input: $\mathbf{x} \in \mathbb{R}^n$

Trainable var.: None

Output: $\mathbf{y}_i = \max(\mathbf{0}, \mathbf{x}_i)$

$\mathbf{y} \in \mathbb{R}^n$

ReLu



$$\frac{\partial L}{\partial \mathbf{x}_i} = \sum_j \frac{\partial L}{\partial \mathbf{y}_j} \frac{\partial}{\partial \mathbf{x}_i} \max(0, \mathbf{x}_j) = \frac{\partial L}{\partial \mathbf{y}_i} \frac{\partial}{\partial \mathbf{x}_i} \max(0, \mathbf{x}_i)$$

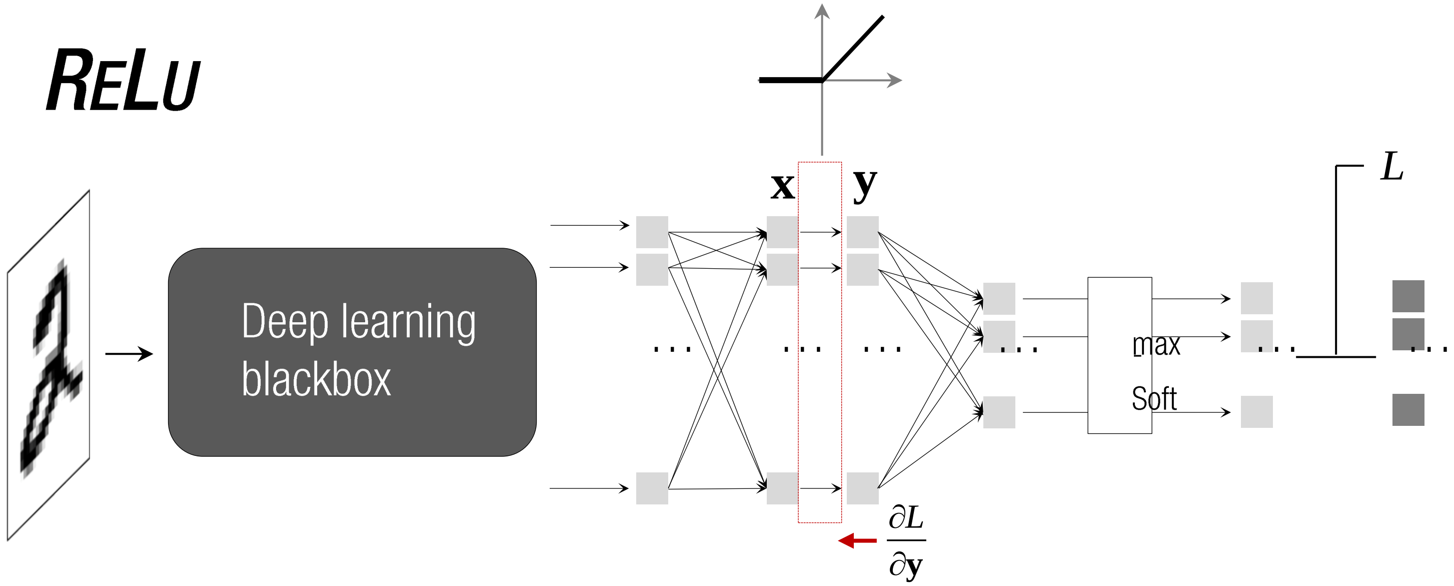
Input: $\mathbf{x} \in \mathbb{R}^n$

Trainable var.: None

Output: $\mathbf{y}_i = \max(0, \mathbf{x}_i)$

$\mathbf{y} \in \mathbb{R}^n$

ReLU



$$\frac{\partial L}{\partial \mathbf{x}_i} = \sum_j \frac{\partial L}{\partial \mathbf{y}_j} \frac{\partial}{\partial \mathbf{x}_i} \max(0, \mathbf{x}_j) = \frac{\partial L}{\partial \mathbf{y}_i} \frac{\partial}{\partial \mathbf{x}_i} \max(0, \mathbf{x}_i) = \begin{cases} \frac{\partial L}{\partial \mathbf{y}_i} & \text{if } \mathbf{y}_i \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

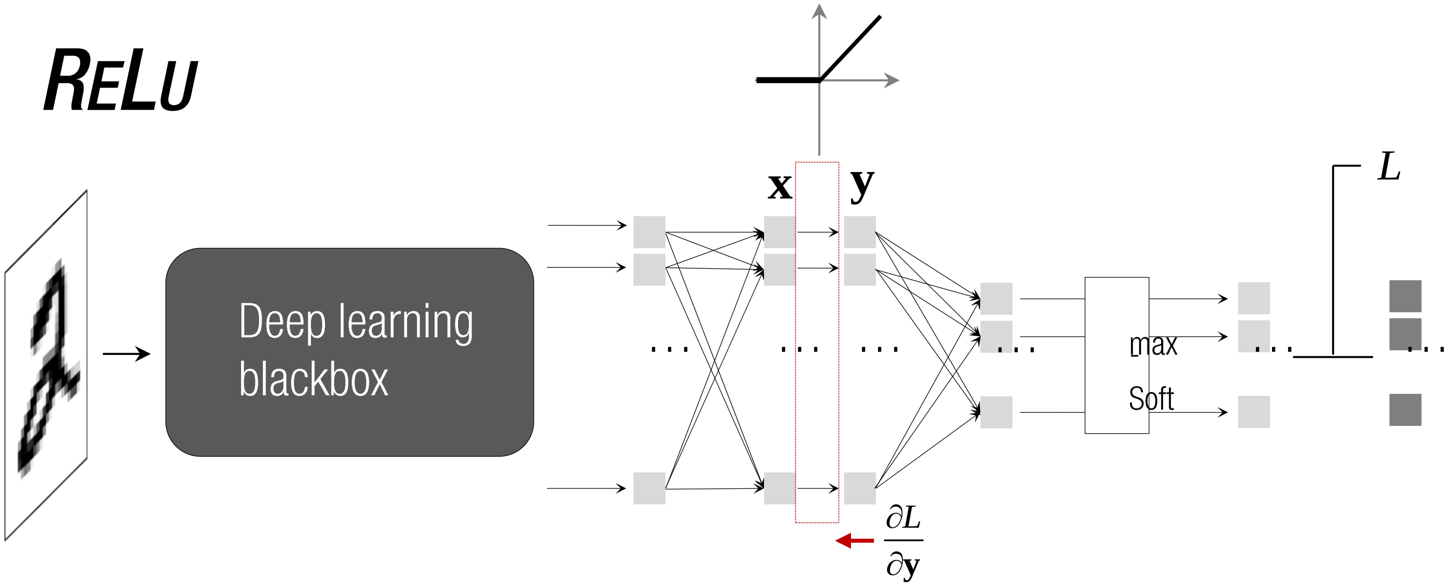
Input: $\mathbf{x} \in \mathbb{R}^n$

Trainable var.: None

Output: $\mathbf{y}_i = \max(0, \mathbf{x}_i)$

$\mathbf{y} \in \mathbb{R}^n$

ReLu



Input: $\mathbf{x} \in \mathbb{R}^n$

Trainable var.: None

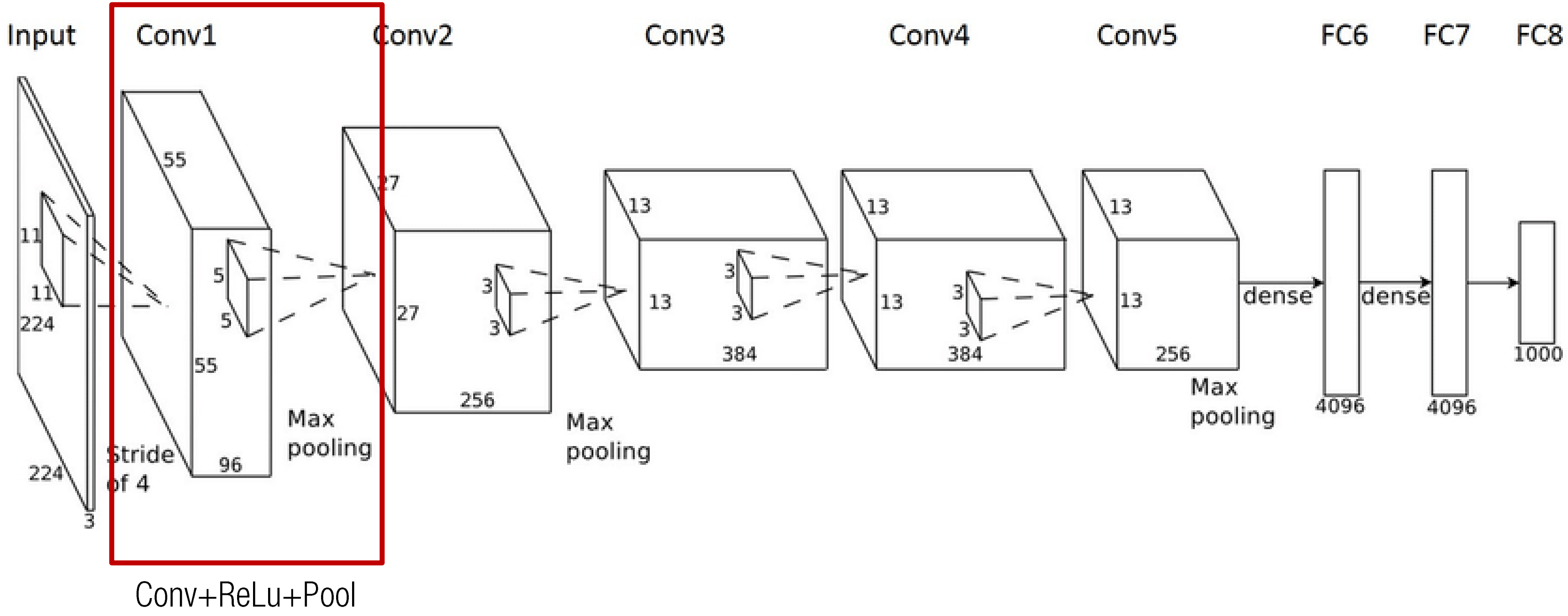
Output: $\mathbf{y}_i = \max(0, \mathbf{x}_i)$

$\mathbf{y} \in \mathbb{R}^n$

$$\frac{\partial L}{\partial \mathbf{x}_i} = \sum_j \frac{\partial L}{\partial \mathbf{y}_j} \frac{\partial}{\partial \mathbf{x}_i} \max(0, \mathbf{x}_j) = \frac{\partial L}{\partial \mathbf{y}_i} \frac{\partial}{\partial \mathbf{x}_i} \max(0, \mathbf{x}_i) = \begin{cases} \frac{\partial L}{\partial \mathbf{y}_i} & \text{if } \mathbf{y}_i \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

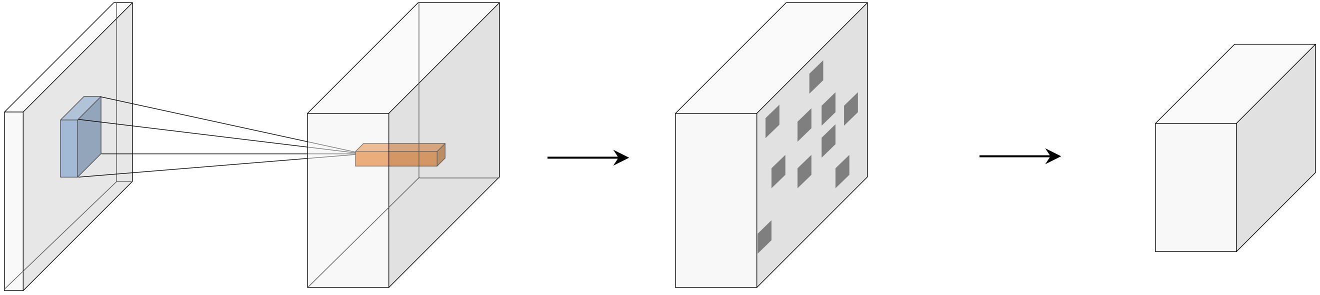
function [y] = Relu(x)
function [dLdx] = Relu_back(dLdy, x, w, y)

RECALL: ALEX NET



Recall: Conv+ReLu+Pool

Conv+ReLu layer



Operations:		Conv		ReLu		Max-pool	
# of units:	$H \times W \times C_1$		$H \times W \times C_2$		$H \times W \times C_2$		$H_1 \times W_1 \times C_2$
# of weights:		$F \times F \times C_1 \times C_2$		0		0	
# of biases:		$1 \times C_2$		0		0	

RECALL: SPATIAL POOLING

3	2	3	1
0	**5**	3	**4**
1	2	2	2
7	3	1	**7**

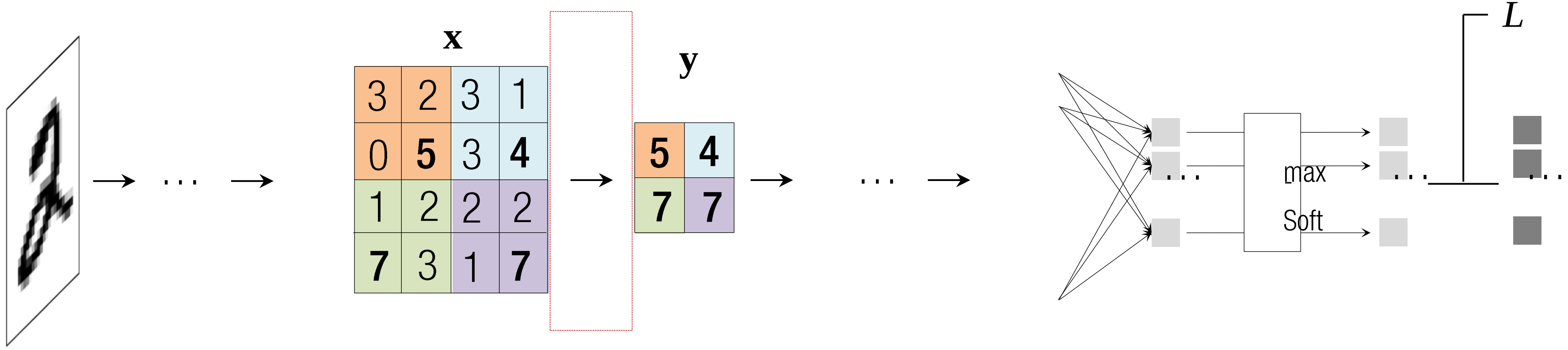
4×4

5	**4**
7	**7**

2×2

Max-pooling (window size 2x2, stride 2)

MAX-POOL

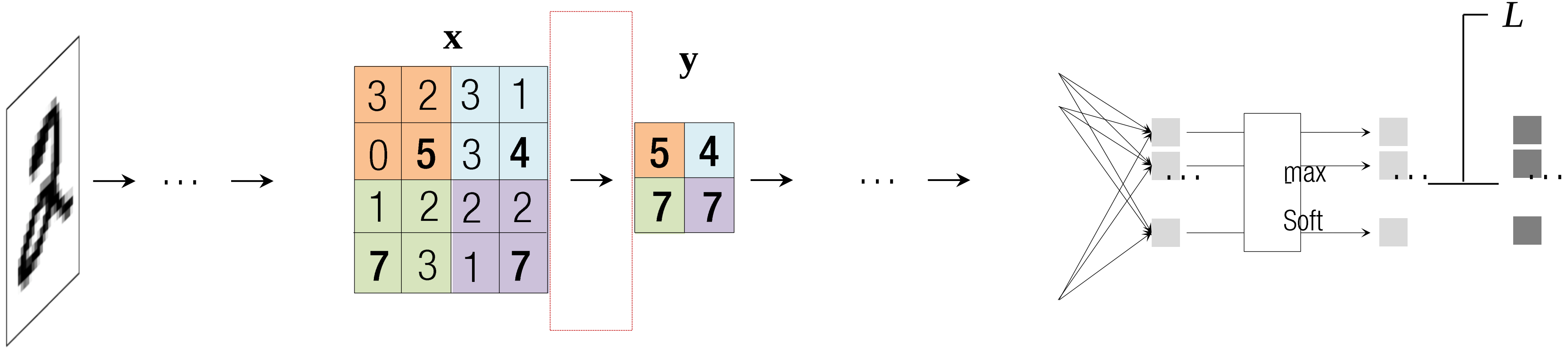


Input:

Trainable var.:

Output:

MAX-POOL

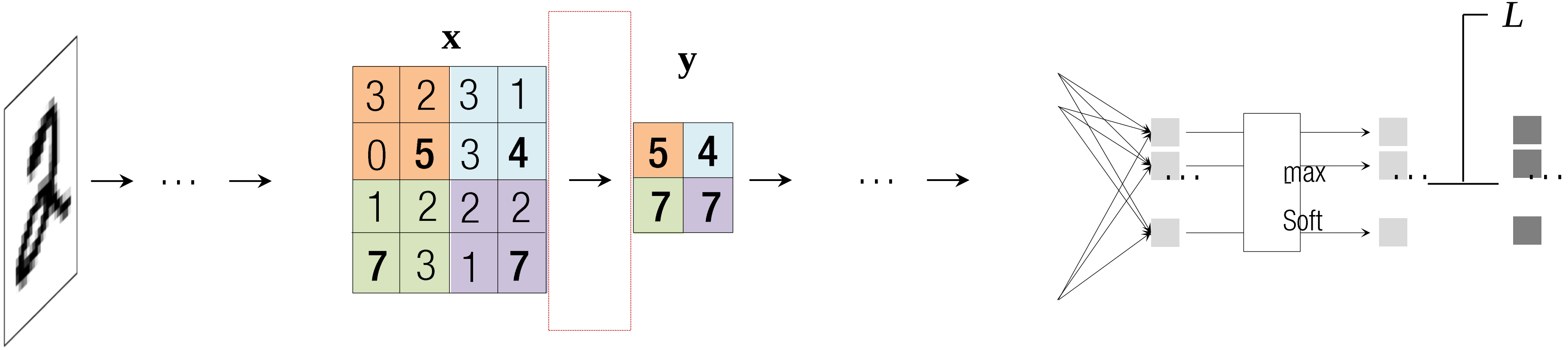


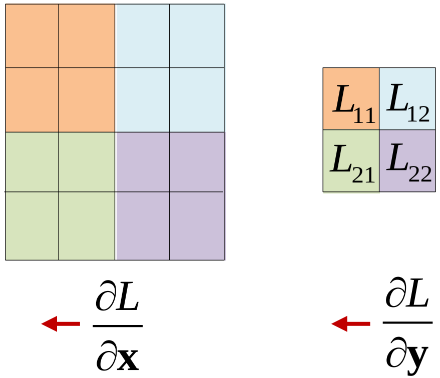
Input: $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$

Trainable var.: None

Output: $\mathbf{y} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$

MAX-POOL



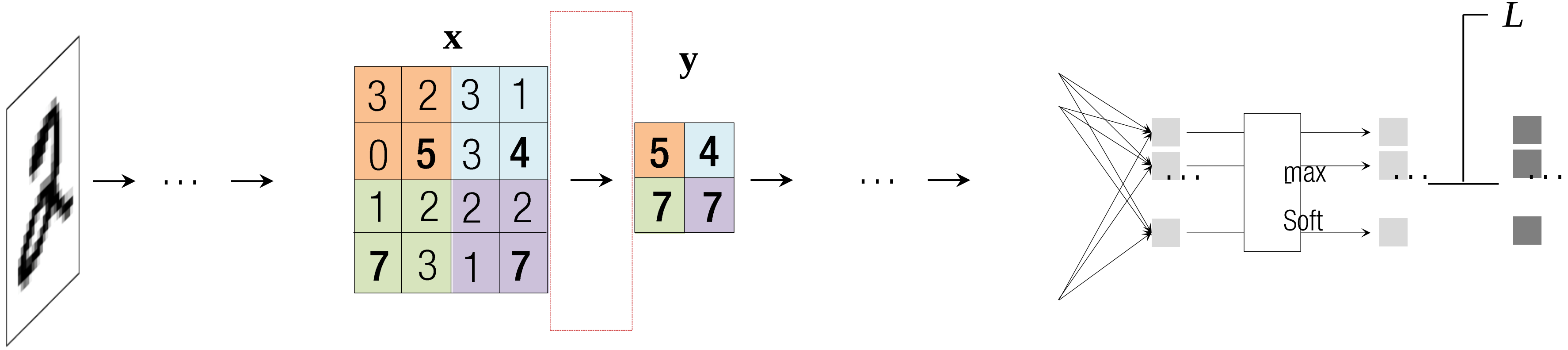


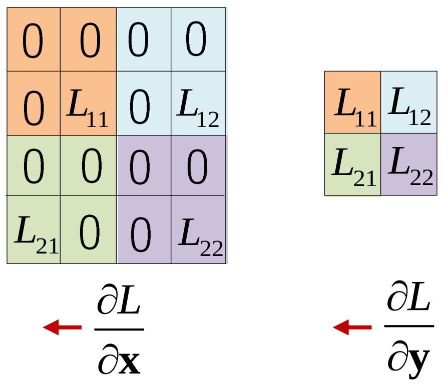
Input: $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$

Trainable var.: None

Output: $\mathbf{y} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$

MAX-POOL



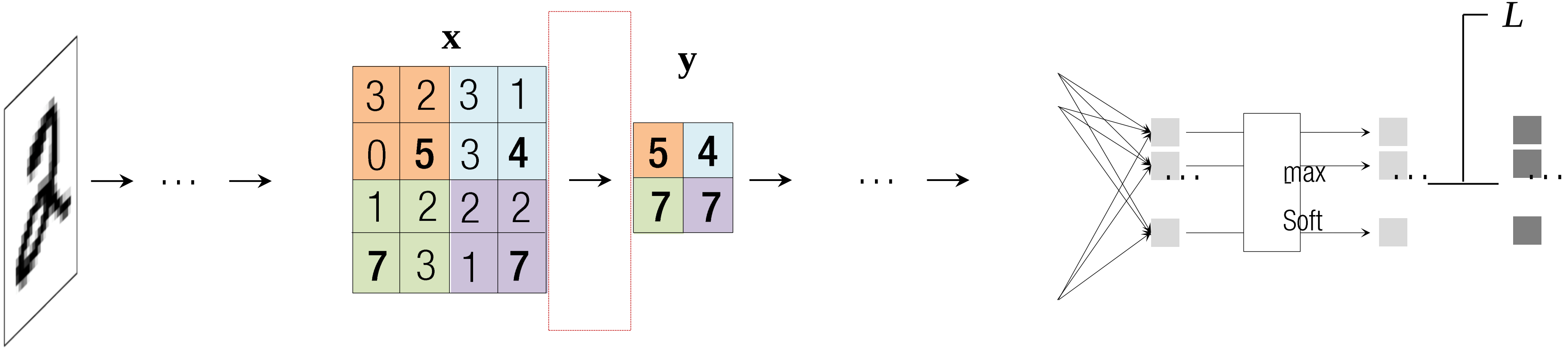


Input: $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$

Trainable var.: None

Output: $\mathbf{y} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$

MAX-POOL



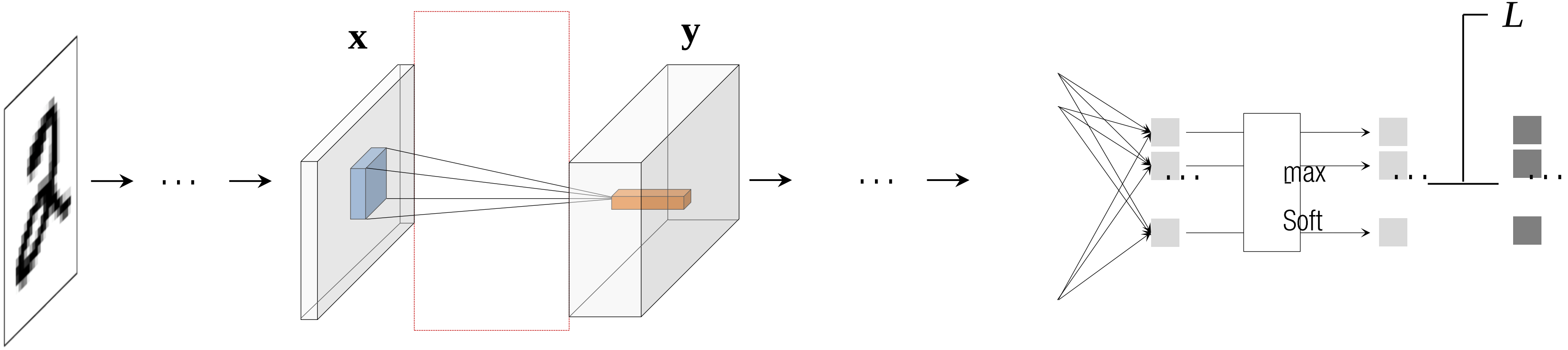
Input: $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$

Trainable var.: None

Output: $\mathbf{y} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$

function [y] = Maxpool(x, size, stride)
function [dLdx] = Maxpool_back(dLdy, x, size, stride, y)

Convolutional Operation



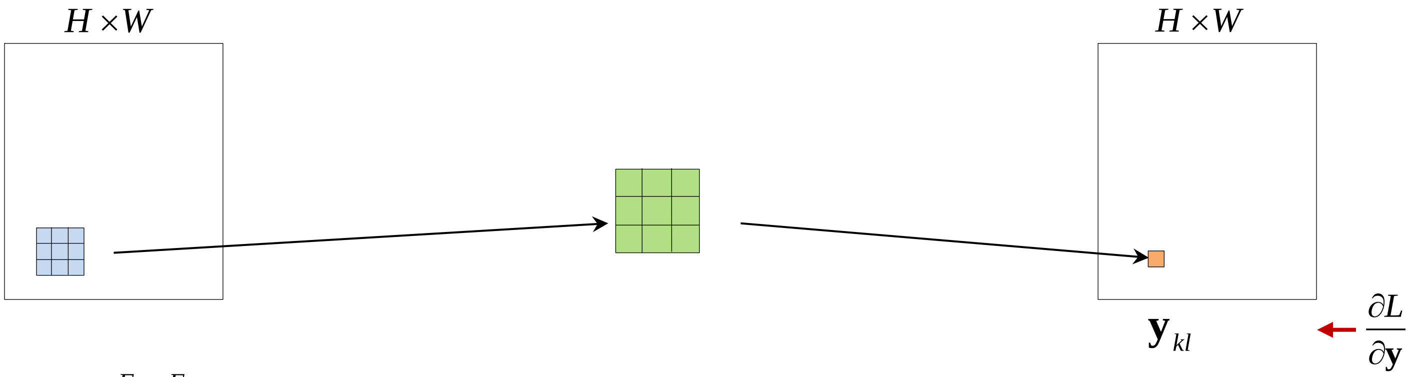
Input: $\mathbf{x} \in \mathbb{R}^{H \times W \times C_1}$

Trainable var.: $\mathbf{w} \in \mathbb{R}^{F \times F \times C_1 \times C_2}$

Output: $\mathbf{y} = \mathbf{x} * \mathbf{w}$

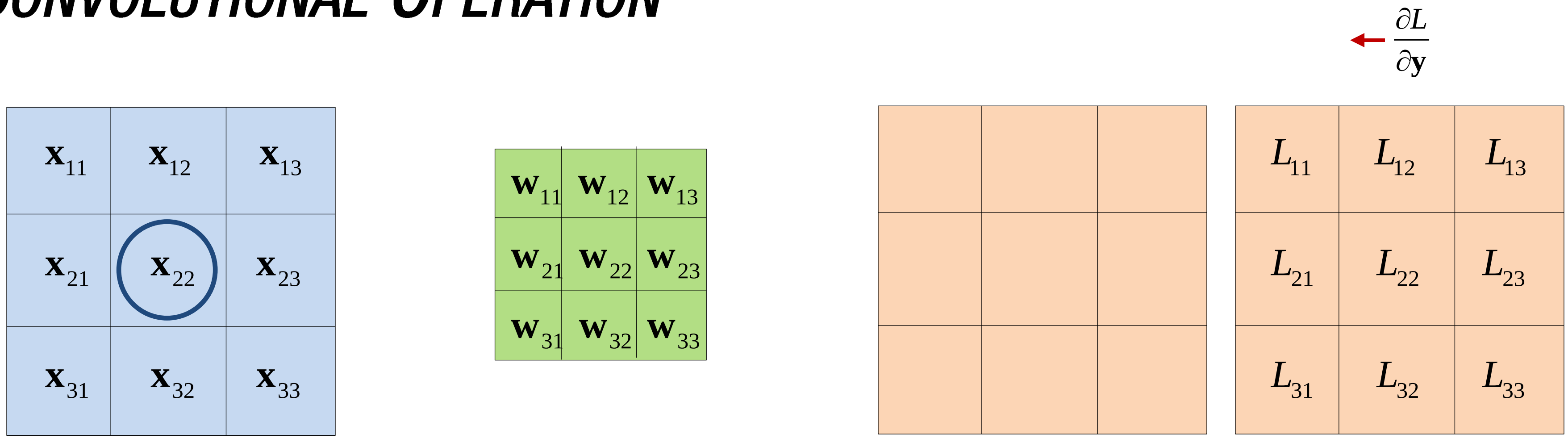
$\mathbf{y} \in \mathbb{R}^{H \times W \times C_2}$

Convolutional Operation



$$\mathbf{y}_{kl} = \sum_{i=k}^{F} \sum_{j=l}^{F} w_{i-k,j-l} x_{ij}$$

Convolutional Operation



$$\frac{\partial L}{\partial \mathbf{x}_{ij}} = \sum_k \sum_l \frac{\partial L}{\partial \mathbf{y}_{kl}} \frac{\partial \mathbf{y}_{kl}}{\partial \mathbf{x}_{ij}}$$

Convolutional Operation

$$\frac{\partial L}{\partial \mathbf{y}}$$

$$\begin{bmatrix} \mathbf{x}_{11} & \mathbf{x}_{12} & \mathbf{x}_{13} \\ \mathbf{x}_{21} & \mathbf{x}_{22} & \mathbf{x}_{23} \\ \mathbf{x}_{31} & \mathbf{x}_{32} & \mathbf{x}_{33} \end{bmatrix} \quad \begin{bmatrix} \mathbf{w}_{11} & \mathbf{w}_{12} & \mathbf{w}_{13} \\ \mathbf{w}_{21} & \mathbf{w}_{22} & \mathbf{w}_{23} \\ \mathbf{w}_{31} & \mathbf{w}_{32} & \mathbf{w}_{33} \end{bmatrix} \quad \begin{bmatrix} \mathbf{y}_{11} & & \\ & & \\ & & \end{bmatrix} \quad \begin{bmatrix} L_{11} & L_{12} & L_{13} \\ L_{21} & L_{22} & L_{23} \\ L_{31} & L_{32} & L_{33} \end{bmatrix}$$

$$\mathbf{y}_{11} = \mathbf{w}_{11}\mathbf{x}_{00} + \cdots + \mathbf{w}_{33}\mathbf{x}_{22}$$

$$\frac{\partial L}{\partial \mathbf{x}_{ij}} = \sum_k \sum_l \frac{\partial L}{\partial \mathbf{y}_{kl}} \frac{\partial \mathbf{y}_{kl}}{\partial \mathbf{x}_{ij}}$$

Convolutional Operation

$$\leftarrow \frac{\partial L}{\partial \mathbf{y}}$$

$$\begin{bmatrix} \mathbf{x}_{11} & \mathbf{x}_{12} & \mathbf{x}_{13} \\ \mathbf{x}_{21} & \mathbf{x}_{22} & \mathbf{x}_{23} \\ \mathbf{x}_{31} & \mathbf{x}_{32} & \mathbf{x}_{33} \end{bmatrix} \quad \begin{bmatrix} \mathbf{w}_{11} & \mathbf{w}_{12} & \mathbf{w}_{13} \\ \mathbf{w}_{21} & \mathbf{w}_{22} & \mathbf{w}_{23} \\ \mathbf{w}_{31} & \mathbf{w}_{32} & \mathbf{w}_{33} \end{bmatrix} \quad \begin{bmatrix} \mathbf{y}_{11} & & \\ & & \\ & & \end{bmatrix} \quad \begin{bmatrix} L_{11} & L_{12} & L_{13} \\ L_{21} & L_{22} & L_{23} \\ L_{31} & L_{32} & L_{33} \end{bmatrix}$$

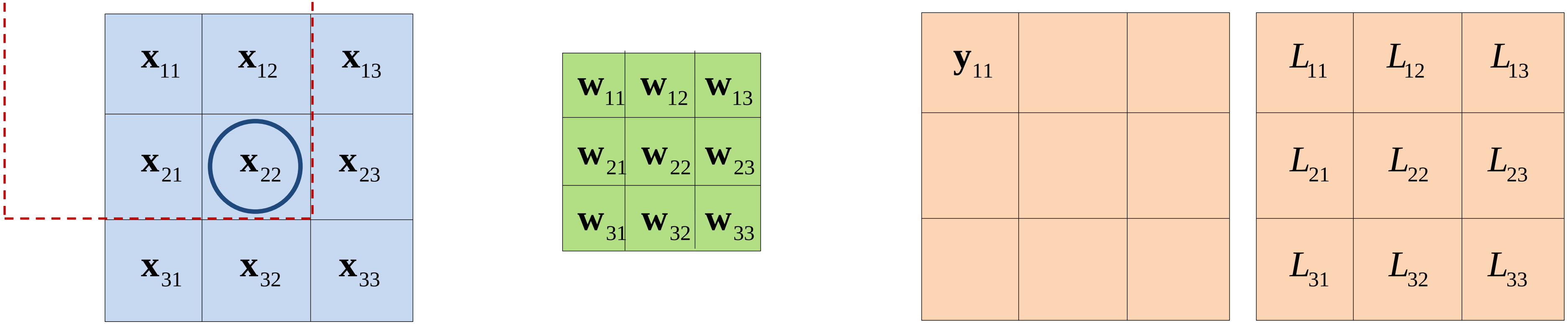
$$\mathbf{y}_{11} = \mathbf{w}_{11}\mathbf{x}_{00} + \cdots + \mathbf{w}_{33}\mathbf{x}_{22}$$

$$\frac{\partial \mathbf{y}_{11}}{\partial \mathbf{x}_{22}} = \mathbf{w}_{33}$$

$$\frac{\partial L}{\partial \mathbf{x}_{ij}} = \sum_k \sum_l \frac{\partial L}{\partial \mathbf{y}_{kl}} \frac{\partial \mathbf{y}_{kl}}{\partial \mathbf{x}_{ij}}$$

Convolutional Operation

$$\frac{\partial L}{\partial \mathbf{y}}$$



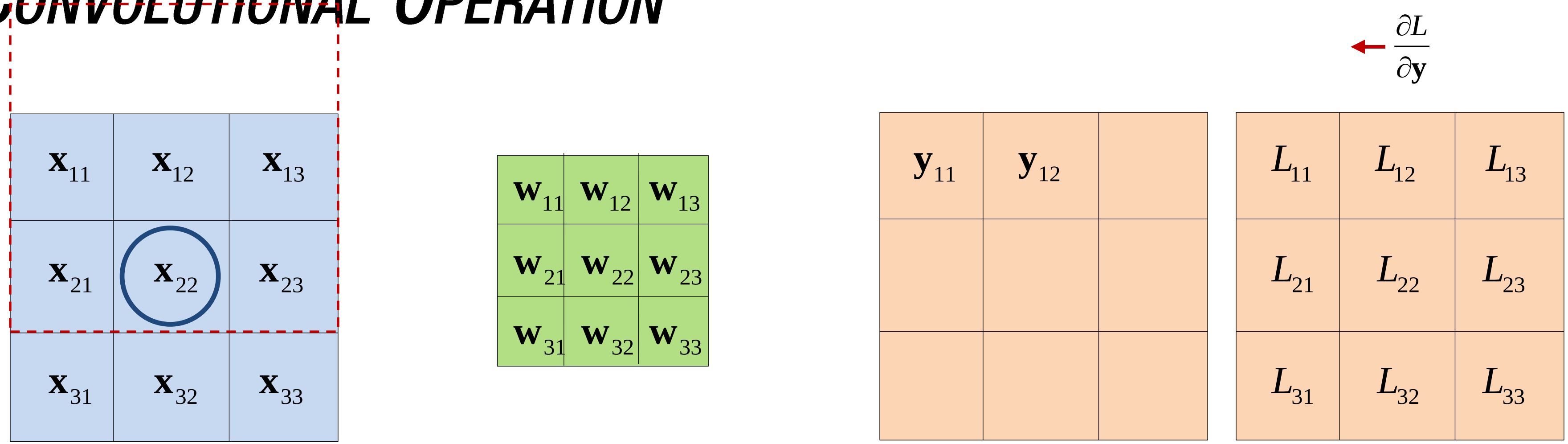
$$\mathbf{y}_{11} = \mathbf{w}_{11}\mathbf{x}_{00} + \cdots + \mathbf{w}_{33}\mathbf{x}_{22}$$

$$\frac{\partial \mathbf{y}_{11}}{\partial \mathbf{x}_{22}} = \mathbf{w}_{33}$$

$$\frac{\partial L}{\partial \mathbf{y}_{11}} \frac{\partial \mathbf{y}_{11}}{\partial \mathbf{x}_{22}} = L_{11}\mathbf{w}_{33}$$

$$\frac{\partial L}{\partial \mathbf{x}_{ij}} = \sum_k \sum_l \frac{\partial L}{\partial \mathbf{y}_{kl}} \frac{\partial \mathbf{y}_{kl}}{\partial \mathbf{x}_{ij}}$$

Convolutional Operation



$$\mathbf{y}_{11} = \mathbf{w}_{11}\mathbf{x}_{00} + \cdots + \mathbf{w}_{33}\mathbf{x}_{22}$$

$$\mathbf{y}_{12} = \mathbf{w}_{11}\mathbf{x}_{01} + \cdots + \mathbf{w}_{33}\mathbf{x}_{23}$$

$$\frac{\partial \mathbf{y}_{11}}{\partial \mathbf{x}_{22}} = \mathbf{w}_{33} \qquad \frac{\partial L}{\partial \mathbf{y}_{11}}\frac{\partial \mathbf{y}_{11}}{\partial \mathbf{x}_{22}} = L_{11}\mathbf{w}_{33}$$

$$\frac{\partial \mathbf{y}_{12}}{\partial \mathbf{x}_{22}} = \mathbf{w}_{32} \qquad \frac{\partial L}{\partial \mathbf{y}_{12}}\frac{\partial \mathbf{y}_{12}}{\partial \mathbf{x}_{22}} = L_{12}\mathbf{w}_{32}$$

$$\frac{\partial L}{\partial \mathbf{x}_{ij}} = \sum_k \sum_l \frac{\partial L}{\partial \mathbf{y}_{kl}} \frac{\partial \mathbf{y}_{kl}}{\partial \mathbf{x}_{ij}}$$

Convolutional Operation

$$\leftarrow \frac{\partial L}{\partial \mathbf{y}}$$

$$\begin{bmatrix} \mathbf{x}_{11} & \mathbf{x}_{12} & \mathbf{x}_{13} \\ \mathbf{x}_{21} & \mathbf{x}_{22} & \mathbf{x}_{23} \\ \mathbf{x}_{31} & \mathbf{x}_{32} & \mathbf{x}_{33} \end{bmatrix} \quad \begin{bmatrix} \mathbf{w}_{11} & \mathbf{w}_{12} & \mathbf{w}_{13} \\ \mathbf{w}_{21} & \mathbf{w}_{22} & \mathbf{w}_{23} \\ \mathbf{w}_{31} & \mathbf{w}_{32} & \mathbf{w}_{33} \end{bmatrix} \quad \begin{bmatrix} \mathbf{y}_{11} & \mathbf{y}_{12} & \mathbf{y}_{13} \\ \mathbf{y}_{21} & \mathbf{y}_{22} & \\ & & \end{bmatrix} \quad \begin{bmatrix} L_{11} & L_{12} & L_{13} \\ L_{21} & L_{22} & L_{23} \\ L_{31} & L_{32} & L_{33} \end{bmatrix}$$

$$\mathbf{y}_{11} = \mathbf{w}_{11}\mathbf{x}_{00} + \cdots + \mathbf{w}_{33}\mathbf{x}_{22} \qquad \frac{\partial \mathbf{y}_{11}}{\partial \mathbf{x}_{22}} = \mathbf{w}_{33} \qquad \frac{\partial L}{\partial \mathbf{y}_{11}} \frac{\partial \mathbf{y}_{11}}{\partial \mathbf{x}_{22}} = L_{11}\mathbf{w}_{33}$$

$$\mathbf{y}_{12} = \mathbf{w}_{11}\mathbf{x}_{01} + \cdots + \mathbf{w}_{33}\mathbf{x}_{23} \qquad \frac{\partial \mathbf{y}_{12}}{\partial \mathbf{x}_{22}} = \mathbf{w}_{32} \qquad \frac{\partial L}{\partial \mathbf{y}_{12}} \frac{\partial \mathbf{y}_{12}}{\partial \mathbf{x}_{22}} = L_{12}\mathbf{w}_{32}$$

$$\cdots$$

$$\mathbf{y}_{22} = \mathbf{w}_{11}\mathbf{x}_{11} + \cdots + \mathbf{w}_{33}\mathbf{x}_{33} \qquad \frac{\partial \mathbf{y}_{22}}{\partial \mathbf{x}_{22}} = \mathbf{w}_{22} \qquad \frac{\partial L}{\partial \mathbf{y}_{22}} \frac{\partial \mathbf{y}_{22}}{\partial \mathbf{x}_{22}} = L_{22}\mathbf{w}_{22}$$

$$\frac{\partial L}{\partial \mathbf{x}_{ij}} = \sum_k \sum_l \frac{\partial L}{\partial \mathbf{y}_{kl}} \frac{\partial \mathbf{y}_{kl}}{\partial \mathbf{x}_{ij}}$$

Convolutional Operation

$\leftarrow \dfrac{\partial L}{\partial \mathbf{y}}$

$\mathbf{x}_{11}$	$\mathbf{x}_{12}$	$\mathbf{x}_{13}$
$\mathbf{x}_{21}$	$\mathbf{x}_{22}$	$\mathbf{x}_{23}$
$\mathbf{x}_{31}$	$\mathbf{x}_{32}$	$\mathbf{x}_{33}$

$\mathbf{w}_{11}$	$\mathbf{w}_{12}$	$\mathbf{w}_{13}$
$\mathbf{w}_{21}$	$\mathbf{w}_{22}$	$\mathbf{w}_{23}$
$\mathbf{w}_{31}$	$\mathbf{w}_{32}$	$\mathbf{w}_{33}$

$\mathbf{y}_{11}$	$\mathbf{y}_{12}$	$\mathbf{y}_{13}$
$\mathbf{y}_{21}$	$\mathbf{y}_{22}$	$\mathbf{y}_{23}$
$\mathbf{y}_{31}$	$\mathbf{y}_{32}$	$\mathbf{y}_{33}$

L_{11}	L_{12}	L_{13}
L_{21}	L_{22}	L_{23}
L_{31}	L_{32}	L_{33}

$$\mathbf{y}_{11} = \mathbf{w}_{11}\mathbf{x}_{00} + \cdots + \mathbf{w}_{33}\mathbf{x}_{22} \qquad \frac{\partial \mathbf{y}_{11}}{\partial \mathbf{x}_{22}} = \mathbf{w}_{33} \qquad \frac{\partial L}{\partial \mathbf{y}_{11}}\frac{\partial \mathbf{y}_{11}}{\partial \mathbf{x}_{22}} = L_{11}\mathbf{w}_{33}$$

$$\mathbf{y}_{12} = \mathbf{w}_{11}\mathbf{x}_{01} + \cdots + \mathbf{w}_{33}\mathbf{x}_{23} \qquad \frac{\partial \mathbf{y}_{12}}{\partial \mathbf{x}_{22}} = \mathbf{w}_{32} \qquad \frac{\partial L}{\partial \mathbf{y}_{12}}\frac{\partial \mathbf{y}_{12}}{\partial \mathbf{x}_{22}} = L_{12}\mathbf{w}_{32}$$

$$\cdots$$

$$\mathbf{y}_{22} = \mathbf{w}_{11}\mathbf{x}_{11} + \cdots + \mathbf{w}_{33}\mathbf{x}_{33} \qquad \frac{\partial \mathbf{y}_{22}}{\partial \mathbf{x}_{22}} = \mathbf{w}_{22} \qquad \frac{\partial L}{\partial \mathbf{y}_{22}}\frac{\partial \mathbf{y}_{22}}{\partial \mathbf{x}_{22}} = L_{22}\mathbf{w}_{22}$$

$$\cdots$$

$$\mathbf{y}_{33} = \mathbf{w}_{11}\mathbf{x}_{22} + \cdots + \mathbf{w}_{33}\mathbf{x}_{44} \qquad \frac{\partial \mathbf{y}_{33}}{\partial \mathbf{x}_{22}} = \mathbf{w}_{11} \qquad \frac{\partial L}{\partial \mathbf{y}_{33}}\frac{\partial \mathbf{y}_{33}}{\partial \mathbf{x}_{22}} = L_{33}\mathbf{w}_{11}$$

$$\frac{\partial L}{\partial \mathbf{x}_{ij}} = \sum_k \sum_l \frac{\partial L}{\partial \mathbf{y}_{kl}} \frac{\partial \mathbf{y}_{kl}}{\partial \mathbf{x}_{ij}}$$

Convolutional Operation

$$
\begin{bmatrix} \mathbf{x}_{11} & \mathbf{x}_{12} & \mathbf{x}_{13} \\ \mathbf{x}_{21} & \mathbf{x}_{22} & \mathbf{x}_{23} \\ \mathbf{x}_{31} & \mathbf{x}_{32} & \mathbf{x}_{33} \end{bmatrix}
\quad
\begin{bmatrix} \mathbf{w}_{11} & \mathbf{w}_{12} & \mathbf{w}_{13} \\ \mathbf{w}_{21} & \mathbf{w}_{22} & \mathbf{w}_{23} \\ \mathbf{w}_{31} & \mathbf{w}_{32} & \mathbf{w}_{33} \end{bmatrix}
\quad
\begin{bmatrix} \mathbf{y}_{11} & \mathbf{y}_{12} & \mathbf{y}_{13} \\ \mathbf{y}_{21} & \mathbf{y}_{22} & \mathbf{y}_{23} \\ \mathbf{y}_{31} & \mathbf{y}_{32} & \mathbf{y}_{33} \end{bmatrix}
\quad
\begin{bmatrix} L_{11} & L_{12} & L_{13} \\ L_{21} & L_{22} & L_{23} \\ L_{31} & L_{32} & L_{33} \end{bmatrix}
$$

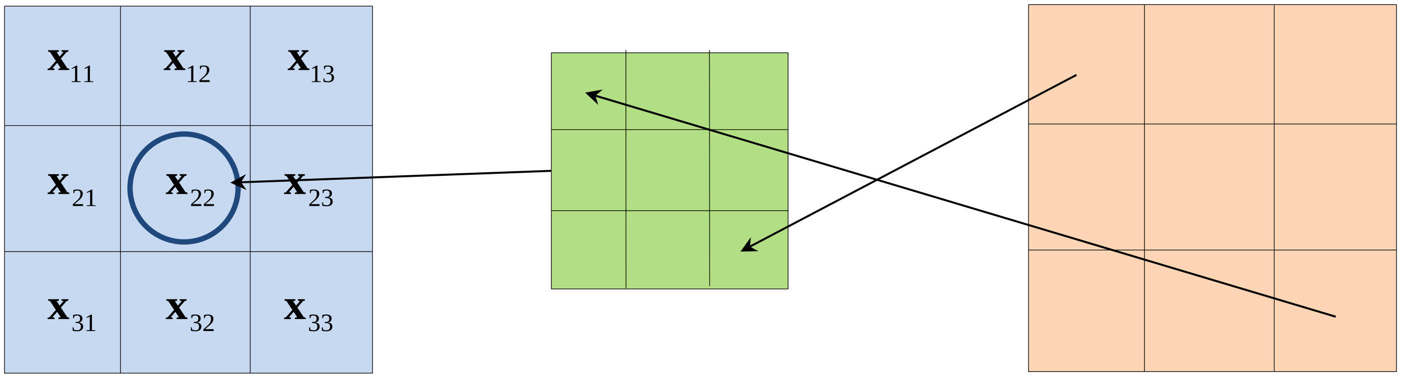
$$\frac{\partial L}{\partial \mathbf{y}_{11}} \frac{\partial \mathbf{y}_{11}}{\partial \mathbf{x}_{22}} = L_{11} \mathbf{w}_{33}$$

$$\cdots$$

$$\frac{\partial L}{\partial \mathbf{y}_{33}} \frac{\partial \mathbf{y}_{33}}{\partial \mathbf{x}_{22}} = L_{33} \mathbf{w}_{11}$$

$$\frac{\partial L}{\partial \mathbf{x}_{ij}} = \sum_m \sum_n L_{mn} \mathbf{w}_{m-i,n-j}$$

Convolutional Operation



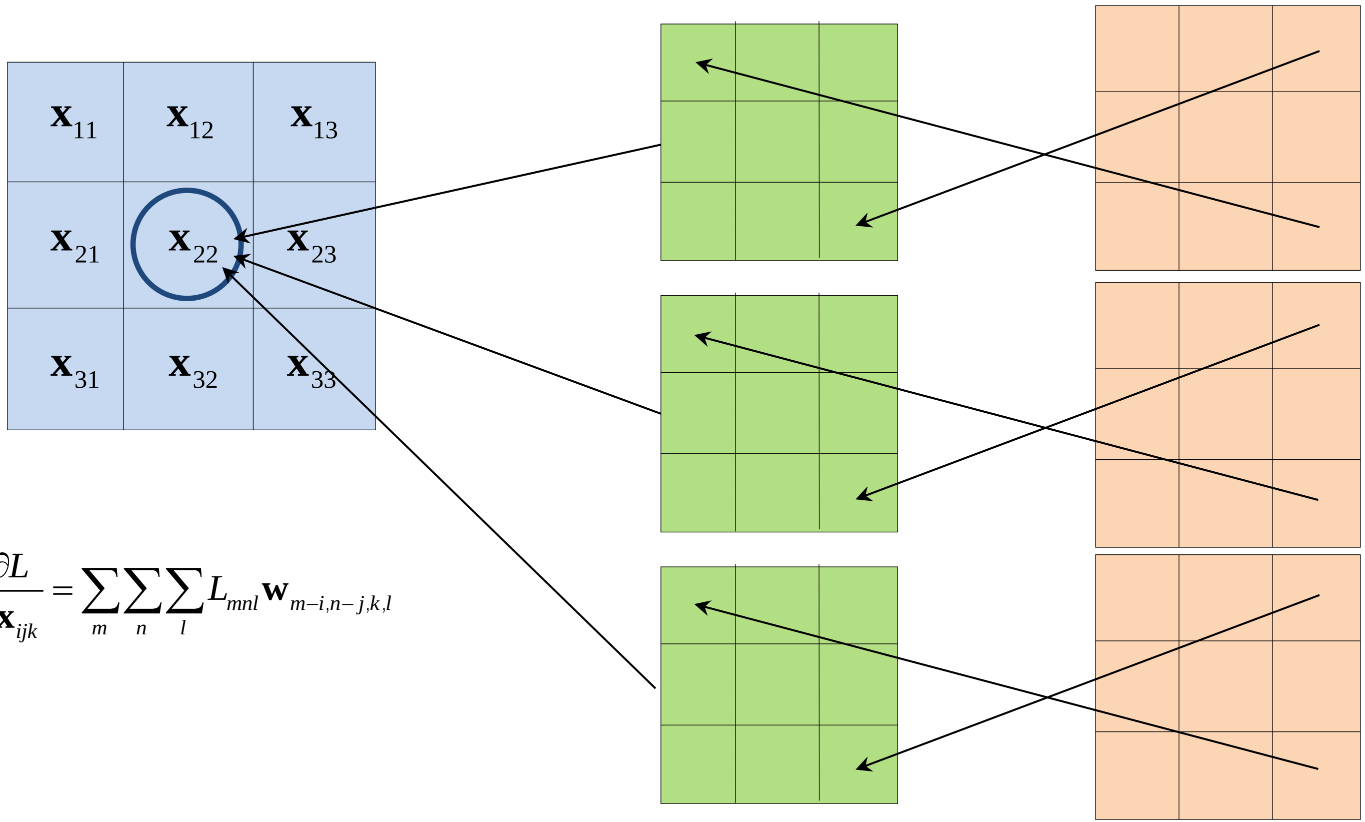
$$\frac{\partial L}{\partial \mathbf{y}_{11}} \frac{\partial \mathbf{y}_{11}}{\partial \mathbf{x}_{22}} = L_{11}\mathbf{w}_{33}$$

$$\cdots$$

$$\frac{\partial L}{\partial \mathbf{x}_{ij}} = \sum_m \sum_n L_{mn} \mathbf{w}_{m-i,n-j}$$

$$\frac{\partial L}{\partial \mathbf{y}_{33}} \frac{\partial \mathbf{y}_{33}}{\partial \mathbf{x}_{22}} = L_{33}\mathbf{w}_{11}$$

Convolutional Operation



$$\frac{\partial L}{\partial \mathbf{x}_{ijk}} = \sum_m \sum_n \sum_l L_{mnl} \mathbf{w}_{m-i,n-j,k,l}$$

Convolutional Operation

$$\leftarrow \frac{\partial L}{\partial \mathbf{y}}$$

$$\begin{bmatrix} \mathbf{x}_{11} & \mathbf{x}_{12} & \mathbf{x}_{13} \\ \mathbf{x}_{21} & \mathbf{x}_{22} & \mathbf{x}_{23} \\ \mathbf{x}_{31} & \mathbf{x}_{32} & \mathbf{x}_{33} \end{bmatrix} \quad \begin{bmatrix} \mathbf{w}_{11} & \mathbf{w}_{12} & \mathbf{w}_{13} \\ \mathbf{w}_{21} & \mathbf{w}_{22} & \mathbf{w}_{23} \\ \mathbf{w}_{31} & \mathbf{w}_{32} & \mathbf{w}_{33} \end{bmatrix} \quad \begin{bmatrix} \mathbf{y}_{11} & \mathbf{y}_{12} & \mathbf{y}_{13} \\ \mathbf{y}_{21} & \mathbf{y}_{22} & \mathbf{y}_{23} \\ \mathbf{y}_{31} & \mathbf{y}_{32} & \mathbf{y}_{33} \end{bmatrix} \quad \begin{bmatrix} L_{11} & L_{12} & L_{13} \\ L_{21} & L_{22} & L_{23} \\ L_{31} & L_{32} & L_{33} \end{bmatrix}$$

$$\mathbf{y}_{11} = \mathbf{w}_{11}\mathbf{x}_{00} + \cdots + \mathbf{w}_{33}\mathbf{x}_{22}$$

$$\frac{\partial \mathbf{y}_{11}}{\partial \mathbf{w}_{22}} = \mathbf{x}_{11}$$

$$\frac{\partial L}{\partial \mathbf{y}_{11}} \frac{\partial \mathbf{y}_{11}}{\partial \mathbf{w}_{22}} = L_{11}\mathbf{x}_{11}$$

$$\frac{\partial L}{\partial \mathbf{w}_{ij}} = \sum_k \sum_l \frac{\partial L}{\partial \mathbf{y}_{kl}} \frac{\partial \mathbf{y}_{kl}}{\partial \mathbf{w}_{ij}}$$

Convolutional Operation

$$\leftarrow \frac{\partial L}{\partial \mathbf{y}}$$

$$\begin{bmatrix} \mathbf{x}_{11} & \mathbf{x}_{12} & \mathbf{x}_{13} \\ \mathbf{x}_{21} & \mathbf{x}_{22} & \mathbf{x}_{23} \\ \mathbf{x}_{31} & \mathbf{x}_{32} & \mathbf{x}_{33} \end{bmatrix} \quad \begin{bmatrix} \mathbf{w}_{11} & \mathbf{w}_{12} & \mathbf{w}_{13} \\ \mathbf{w}_{21} & \mathbf{w}_{22} & \mathbf{w}_{23} \\ \mathbf{w}_{31} & \mathbf{w}_{32} & \mathbf{w}_{33} \end{bmatrix} \quad \begin{bmatrix} \mathbf{y}_{11} & \mathbf{y}_{12} & \mathbf{y}_{13} \\ \mathbf{y}_{21} & \mathbf{y}_{22} & \mathbf{y}_{23} \\ \mathbf{y}_{31} & \mathbf{y}_{32} & \mathbf{y}_{33} \end{bmatrix} \quad \begin{bmatrix} L_{11} & L_{12} & L_{13} \\ L_{21} & L_{22} & L_{23} \\ L_{31} & L_{32} & L_{33} \end{bmatrix}$$

$$\mathbf{y}_{11} = \mathbf{w}_{11}\mathbf{x}_{00} + \cdots + \mathbf{w}_{33}\mathbf{x}_{22}$$

$$\mathbf{y}_{12} = \mathbf{w}_{11}\mathbf{x}_{01} + \cdots + \mathbf{w}_{33}\mathbf{x}_{23}$$

$$\cdots$$

$$\frac{\partial \mathbf{y}_{11}}{\partial \mathbf{w}_{22}} = \mathbf{x}_{11} \qquad \frac{\partial L}{\partial \mathbf{y}_{11}} \frac{\partial \mathbf{y}_{11}}{\partial \mathbf{w}_{22}} = L_{11}\mathbf{x}_{11}$$

$$\frac{\partial \mathbf{y}_{12}}{\partial \mathbf{w}_{22}} = \mathbf{x}_{12} \qquad \frac{\partial L}{\partial \mathbf{y}_{12}} \frac{\partial \mathbf{y}_{12}}{\partial \mathbf{w}_{22}} = L_{12}\mathbf{x}_{12}$$

$$\frac{\partial L}{\partial \mathbf{w}_{ij}} = \sum_k \sum_l \frac{\partial L}{\partial \mathbf{y}_{kl}} \frac{\partial \mathbf{y}_{kl}}{\partial \mathbf{w}_{ij}}$$

Convolutional Operation

$\leftarrow \dfrac{\partial L}{\partial \mathbf{y}}$

$\mathbf{x}_{11}$	$\mathbf{x}_{12}$	$\mathbf{x}_{13}$
$\mathbf{x}_{21}$	$\mathbf{x}_{22}$	$\mathbf{x}_{23}$
$\mathbf{x}_{31}$	$\mathbf{x}_{32}$	$\mathbf{x}_{33}$

$\mathbf{w}_{11}$	$\mathbf{w}_{12}$	$\mathbf{w}_{13}$
$\mathbf{w}_{21}$	$\mathbf{w}_{22}$	$\mathbf{w}_{23}$
$\mathbf{w}_{31}$	$\mathbf{w}_{32}$	$\mathbf{w}_{33}$

$\mathbf{y}_{11}$	$\mathbf{y}_{12}$	$\mathbf{y}_{13}$
$\mathbf{y}_{21}$	$\mathbf{y}_{22}$	$\mathbf{y}_{23}$
$\mathbf{y}_{31}$	$\mathbf{y}_{32}$	$\mathbf{y}_{33}$

L_{11}	L_{12}	L_{13}
L_{21}	L_{22}	L_{23}
L_{31}	L_{32}	L_{33}

$$\mathbf{y}_{11} = \mathbf{w}_{11}\mathbf{x}_{00} + \cdots + \mathbf{w}_{33}\mathbf{x}_{22}$$

$$\mathbf{y}_{12} = \mathbf{w}_{11}\mathbf{x}_{01} + \cdots + \mathbf{w}_{33}\mathbf{x}_{23}$$

$$\cdots$$

$$\mathbf{y}_{22} = \mathbf{w}_{11}\mathbf{x}_{11} + \cdots + \mathbf{w}_{33}\mathbf{x}_{33}$$

$$\cdots$$

$$\mathbf{y}_{33} = \mathbf{w}_{11}\mathbf{x}_{22} + \cdots + \mathbf{w}_{33}\mathbf{x}_{44}$$

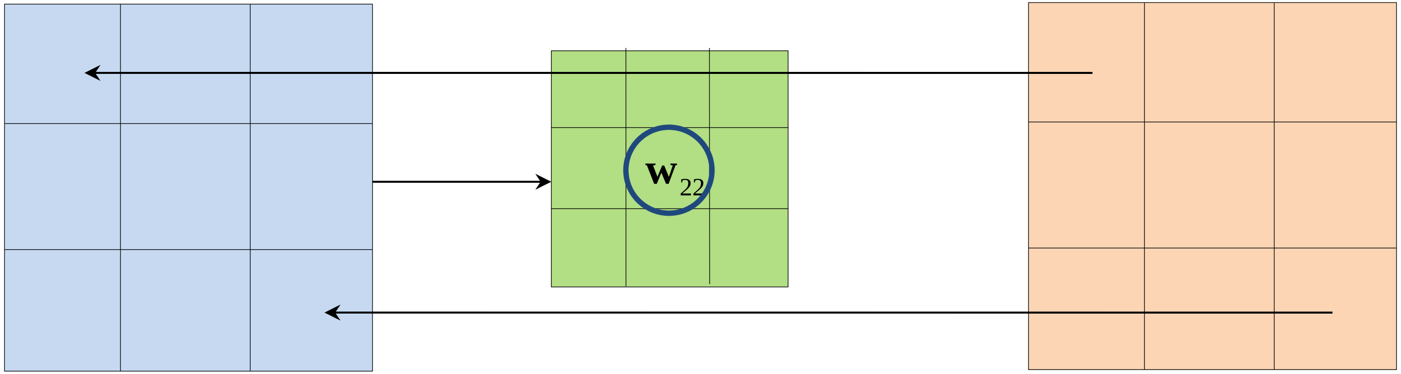
$$\frac{\partial \mathbf{y}_{11}}{\partial \mathbf{w}_{22}} = \mathbf{x}_{11} \qquad \frac{\partial L}{\partial \mathbf{y}_{11}}\frac{\partial \mathbf{y}_{11}}{\partial \mathbf{w}_{22}} = L_{11}\mathbf{x}_{11}$$

$$\frac{\partial \mathbf{y}_{12}}{\partial \mathbf{w}_{22}} = \mathbf{x}_{12} \qquad \frac{\partial L}{\partial \mathbf{y}_{12}}\frac{\partial \mathbf{y}_{12}}{\partial \mathbf{w}_{22}} = L_{12}\mathbf{x}_{12}$$

$$\frac{\partial \mathbf{y}_{33}}{\partial \mathbf{w}_{22}} = \mathbf{x}_{33} \qquad \frac{\partial L}{\partial \mathbf{y}_{33}}\frac{\partial \mathbf{y}_{33}}{\partial \mathbf{w}_{22}} = L_{33}\mathbf{x}_{33}$$

$$\frac{\partial L}{\partial \mathbf{w}_{ij}} = \sum_k \sum_l \frac{\partial L}{\partial \mathbf{y}_{kl}} \frac{\partial \mathbf{y}_{kl}}{\partial \mathbf{w}_{ij}}$$

Convolutional Operation



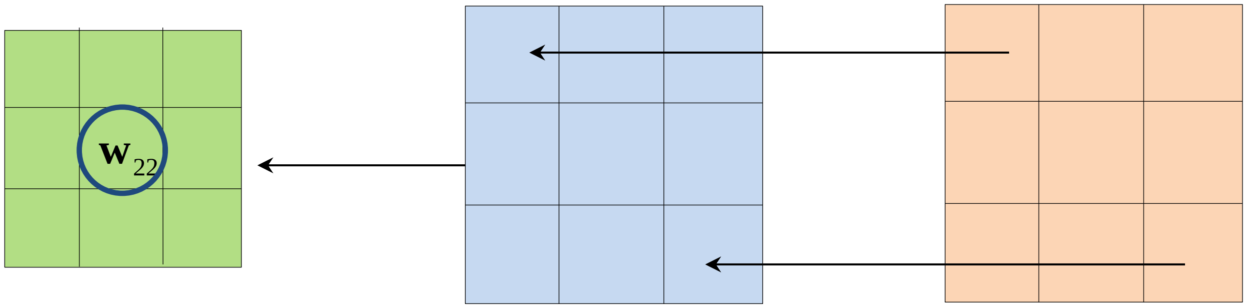
$$\frac{\partial L}{\partial \mathbf{y}_{11}} \frac{\partial \mathbf{y}_{11}}{\partial \mathbf{w}_{22}} = L_{11}\mathbf{x}_{11}$$

$$\cdots$$

$$\frac{\partial L}{\partial \mathbf{y}_{33}} \frac{\partial \mathbf{y}_{33}}{\partial \mathbf{w}_{22}} = L_{33}\mathbf{x}_{33}$$

$$\frac{\partial L}{\partial \mathbf{w}_{ij}} = \sum_k \sum_l L_{ij}\mathbf{x}_{k+i,l+j}$$

Convolutional Operation



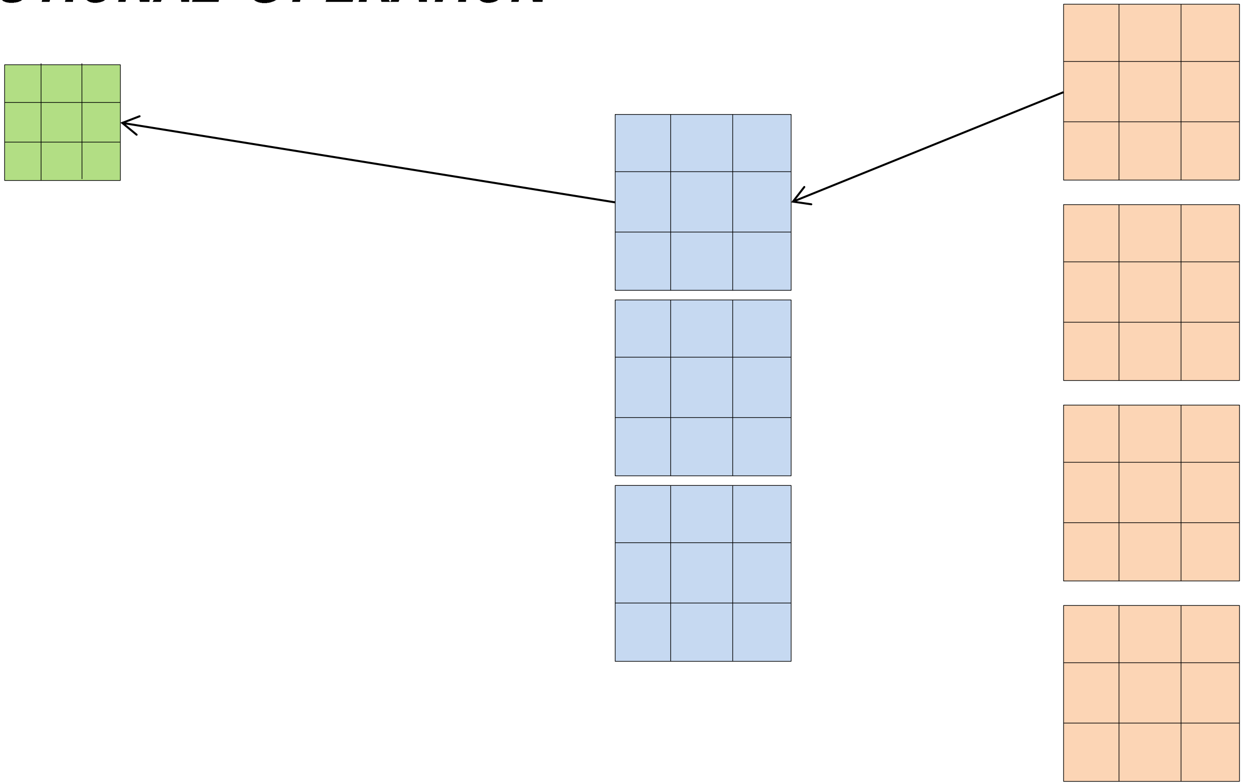
$$\frac{\partial L}{\partial \mathbf{y}_{11}} \frac{\partial \mathbf{y}_{11}}{\partial \mathbf{w}_{22}} = L_{11}\mathbf{x}_{11}$$

$$\cdots$$

$$\frac{\partial L}{\partial \mathbf{y}_{33}} \frac{\partial \mathbf{y}_{33}}{\partial \mathbf{w}_{22}} = L_{33}\mathbf{x}_{33}$$

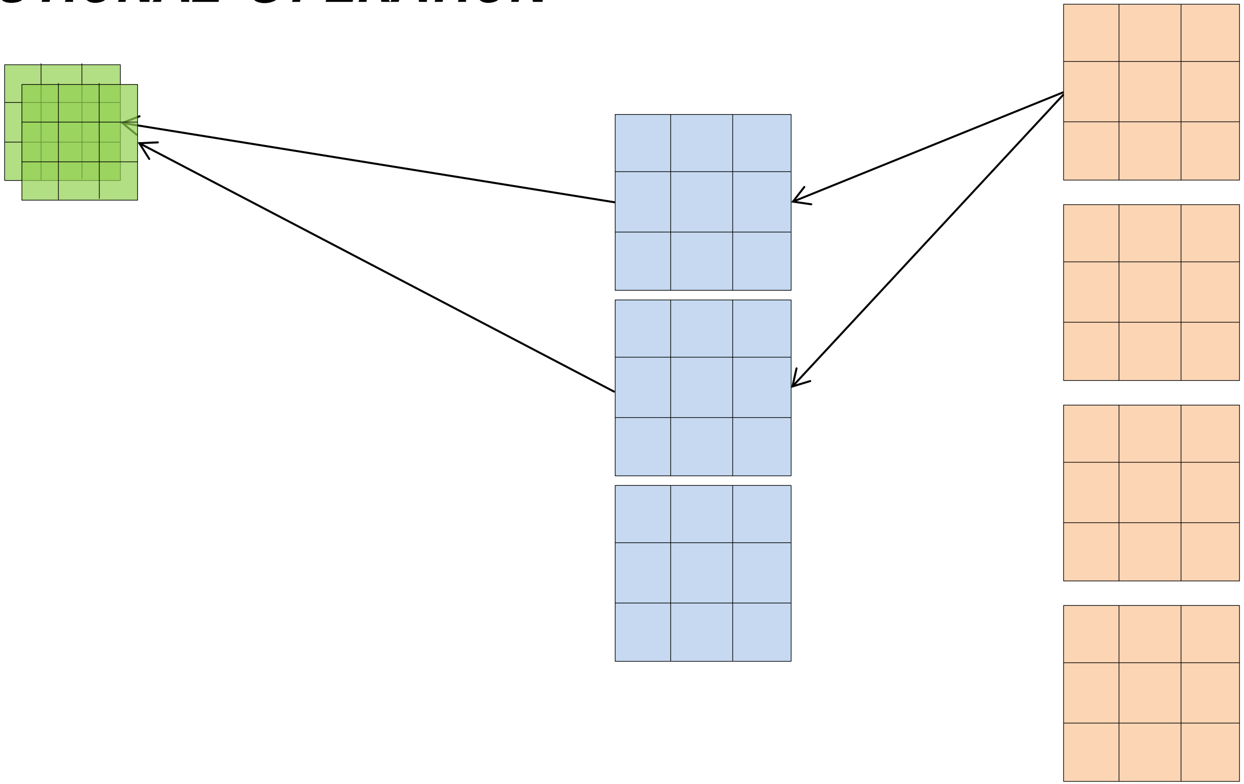
$$\frac{\partial L}{\partial \mathbf{w}_{ij}} = \sum_k \sum_l L_{ij}\mathbf{x}_{k+i,l+j}$$

Convolutional Operation



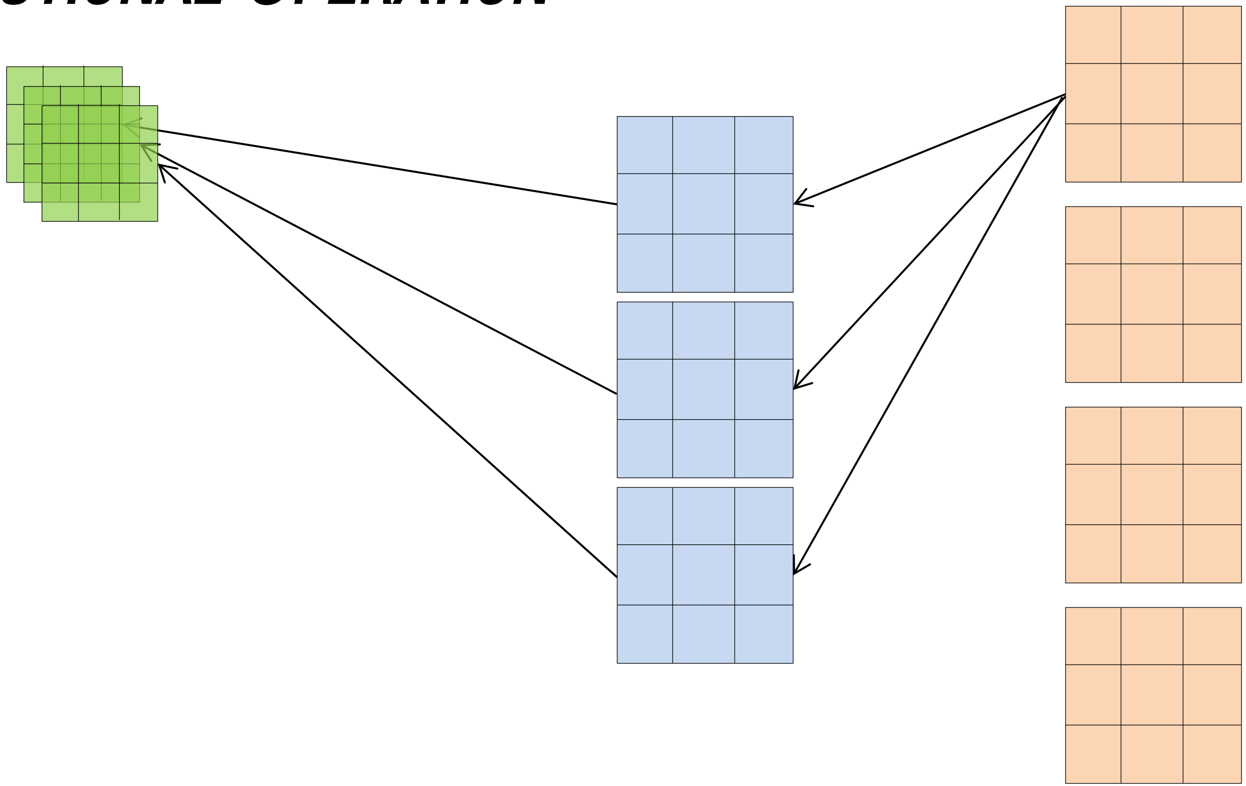
$$\frac{\partial L}{\partial \mathbf{w}_{ijcd}} = \sum_{k} \sum_{l} L_{kld} \mathbf{x}_{k+i,l+j,c}$$

Convolutional Operation



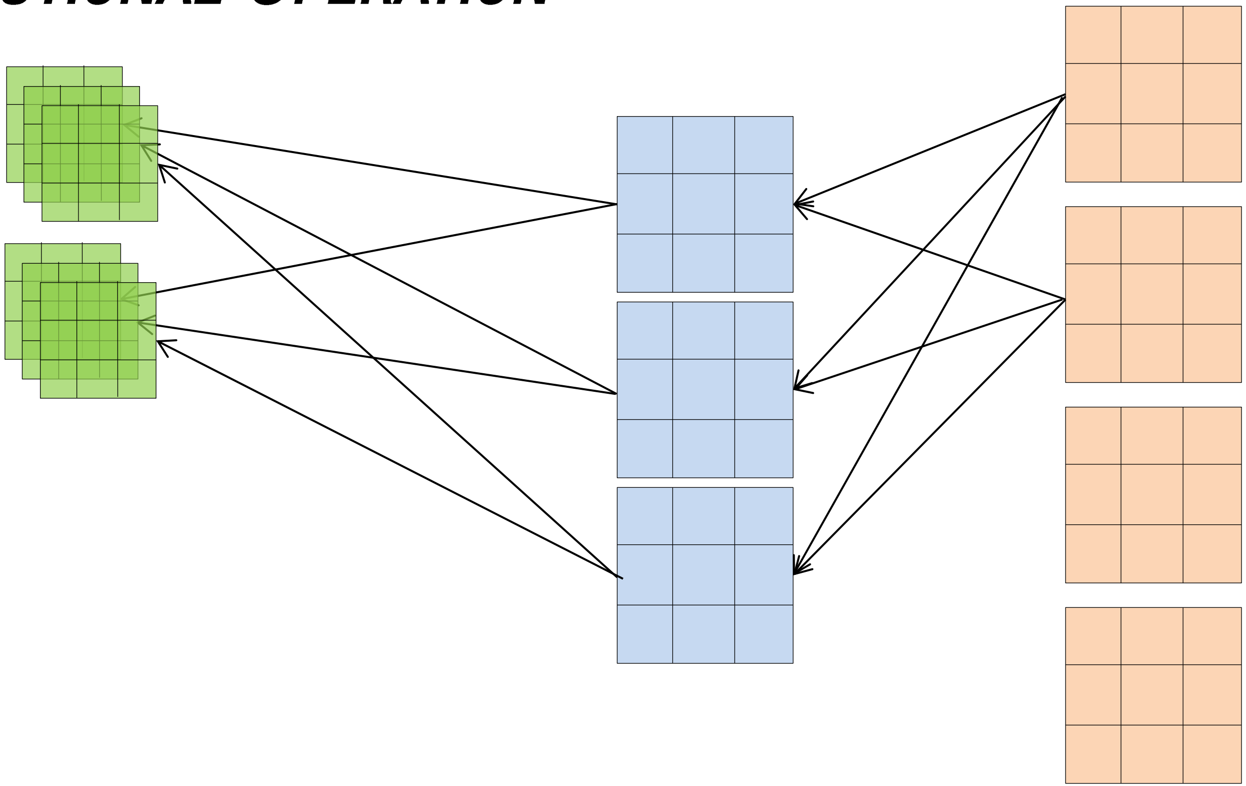
$$\frac{\partial L}{\partial \mathbf{w}_{ijcd}} = \sum_k \sum_l L_{kld} \mathbf{x}_{k+i,l+j,c}$$

Convolutional Operation



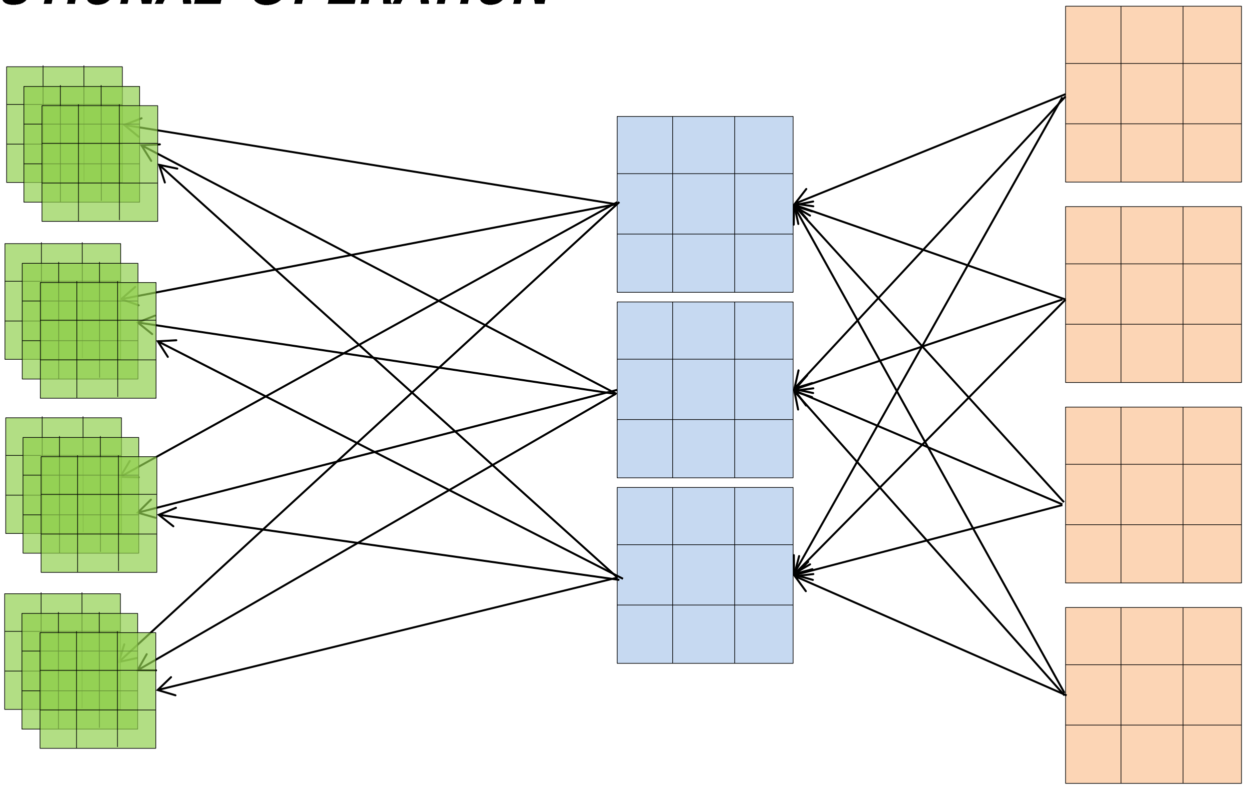
$$\frac{\partial L}{\partial \mathbf{w}_{ijcd}} = \sum_k \sum_l L_{kld} \mathbf{x}_{k+i,l+j,c}$$

Convolutional Operation



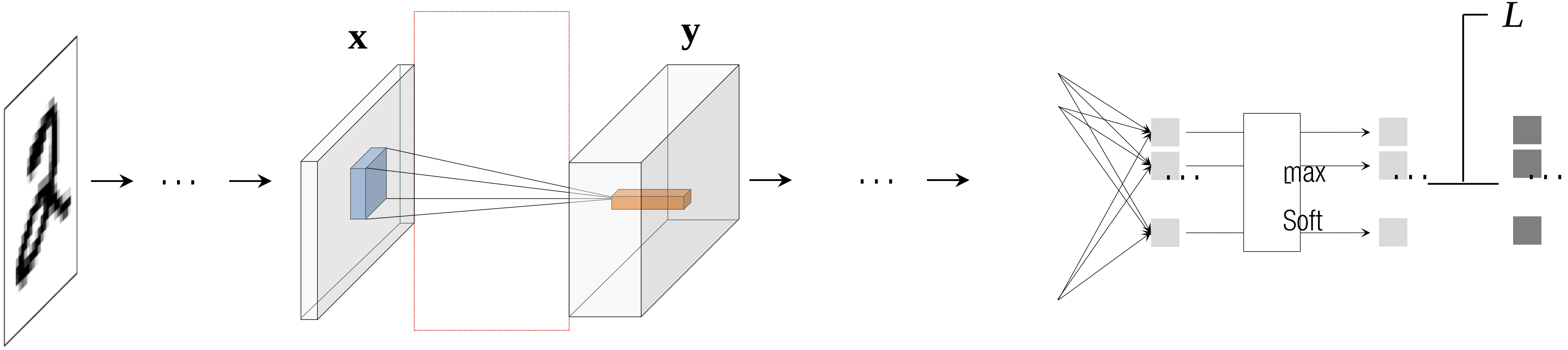
$$\frac{\partial L}{\partial \mathbf{w}_{ijcd}} = \sum_k \sum_l L_{kld} \mathbf{x}_{k+i,l+j,c}$$

Convolutional Operation



$$\frac{\partial L}{\partial \mathbf{w}_{ijcd}} = \sum_k \sum_l L_{kld} \mathbf{x}_{k+i,l+j,c}$$

CONVOLUTIONAL OPERATION



Input: $\mathbf{x} \in \mathbb{R}^{H \times W \times C_1}$

Trainable var.: $\mathbf{w} \in \mathbb{R}^{F \times F \times C_1 \times C_2}$

Output: $\mathbf{y} = \mathbf{x} * \mathbf{w}$

$\mathbf{y} \in \mathbb{R}^{H \times W \times C_2}$

$$\frac{\partial L}{\partial \mathbf{x}_{ijk}} = \sum_m \sum_n \sum_l L_{mnl} \mathbf{w}_{m-i,n-j,k,l}$$

$$\frac{\partial L}{\partial \mathbf{w}_{ijcd}} = \sum_k \sum_l L_{kld} \mathbf{x}_{k+i,l+j,c}$$

function [y] = Conv(x, w, b)
function [dLdx dLdw dLdb] = Conv_back(dLdy, x, w, b, y)

IM2COL

3	2	3	1
0	5	3	4
1	2	2	2
7	3	1	7

$\xrightarrow{\text{im2col}}$

Im2Col

3	2	3	1
0	5	3	4
1	2	2	2
7	3	1	7

Im2col(A, [2,2], ‘distinct’) →

$$\begin{matrix}3\\0\\2\\5\end{matrix}$$

Im2Col

3	2	3	1
0	5	3	4
1	2	2	2
7	3	1	7

$$\xrightarrow{\text{Im2col(A, [2,2], ‘distinct’)}}$$

$$\begin{matrix} 3 & 0 \\ 0 & 1 \\ 2 & 5 \\ 5 & 2 \end{matrix}$$

Im2Col

3	2	3	1
0	5	3	4
1	2	2	2
7	3	1	7

Im2col(A, [2,2], 'distinct') →

3	0	1
0	1	7
2	5	2
5	2	3

IM2COL

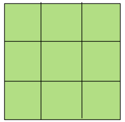
3	2	3	1
0	5	3	4
1	2	2	2
7	3	1	7

Im2col(A, [2,2], ‘distinct’) →

$$\begin{matrix} 3 & 0 & 1 & 2 \\ 0 & 1 & 7 & 5 \\ 2 & 5 & 2 & 3 \\ 5 & 2 & 3 & 3 \end{matrix} \quad \ldots$$

Convolution via im2col (Derivative w.r.t. W)

$\mathbf{w}$



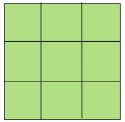
$$* \quad \begin{array}{c} \mathbf{x} \\ \begin{array}{|c|c|c|c|} \hline 3 & 2 & 3 & 1 \\ \hline 0 & 5 & 3 & 4 \\ \hline 1 & 2 & 2 & 2 \\ \hline 7 & 3 & 1 & 7 \\ \hline \end{array} \end{array} \quad = \mathbf{y}$$

$\mathbf{W}$

$$\begin{array}{c} \mathbf{X} \\ \begin{bmatrix} 3 & 0 & 1 & 2 & \\ 0 & 1 & 7 & 5 & \cdots \\ 2 & 5 & 2 & 3 & \\ 5 & 2 & 3 & 3 & \end{bmatrix} \end{array} = \mathbf{Y}$$

Convolution via im2col (Derivative w.r.t. W)

$\mathbf{w}$



$*$

$\mathbf{x}$

3	2	3	1
0	5	3	4
1	2	2	2
7	3	1	7

$= \mathbf{y}$

$$\frac{\partial L}{\partial \mathbf{w}_{ij}} = \sum_k \sum_l \frac{\partial L}{\partial \mathbf{y}_{kl}} \frac{\partial \mathbf{y}_{kl}}{\partial \mathbf{w}_{ij}}$$

$\mathbf{W}$

$\mathbf{X}$

$$\begin{bmatrix} 3 & 0 & 1 & 2 & \\ 0 & 1 & 7 & 5 & \cdots \\ 2 & 5 & 2 & 3 & \\ 5 & 2 & 3 & 3 & \end{bmatrix} = \mathbf{Y}$$

Convolution via im2col (Derivative w.r.t. W)

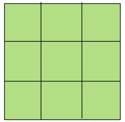
$$\mathbf{w} * \overset{\mathbf{x}}{\begin{bmatrix} 3 & 2 & 3 & 1 \\ 0 & 5 & 3 & 4 \\ 1 & 2 & 2 & 2 \\ 7 & 3 & 1 & 7 \end{bmatrix}} = \mathbf{y}$$

$$\frac{\partial L}{\partial \mathbf{w}_{ij}} = \sum_k \sum_l \frac{\partial L}{\partial \mathbf{y}_{kl}} \frac{\partial \mathbf{y}_{kl}}{\partial \mathbf{w}_{ij}}$$

$$\mathbf{W} \quad \overset{\mathbf{X}}{\begin{bmatrix} 3 & 0 & 1 & 2 & \\ 0 & 1 & 7 & 5 & \cdots \\ 2 & 5 & 2 & 3 & \\ 5 & 2 & 3 & 3 & \end{bmatrix}} = \mathbf{Y}$$

Convolution via im2col (Derivative w.r.t. W)

$\mathbf{w}$



$$* \quad \overset{\mathbf{x}}{\begin{bmatrix} 3 & 2 & 3 & 1 \\ 0 & 5 & 3 & 4 \\ 1 & 2 & 2 & 2 \\ 7 & 3 & 1 & 7 \end{bmatrix}} = \mathbf{y}$$

$$\frac{\partial L}{\partial \mathbf{w}_{ij}} = \sum_k \sum_l \frac{\partial L}{\partial \mathbf{y}_{kl}} \frac{\partial \mathbf{y}_{kl}}{\partial \mathbf{w}_{ij}}$$

$$\mathbf{X}^T\mathbf{W} = \mathbf{Y}$$

CONVOLUTION VIA IM2COL (DERIVATIVE W.R.T. W)

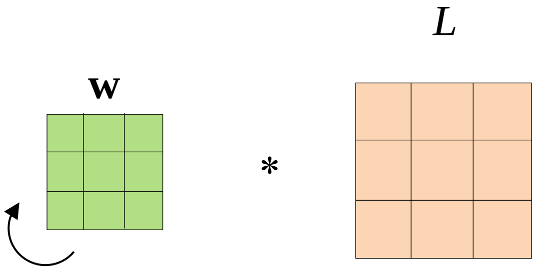
$$\mathbf{w} * \mathbf{x} = \mathbf{y}, \qquad \mathbf{x} = \begin{bmatrix} 3 & 2 & 3 & 1 \\ 0 & 5 & 3 & 4 \\ 1 & 2 & 2 & 2 \\ 7 & 3 & 1 & 7 \end{bmatrix}$$

$$\frac{\partial L}{\partial \mathbf{w}_{ij}} = \sum_k \sum_l \frac{\partial L}{\partial \mathbf{y}_{kl}} \frac{\partial \mathbf{y}_{kl}}{\partial \mathbf{w}_{ij}}$$

$$\mathbf{X}^T \mathbf{W} = \mathbf{Y}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{w}} = \mathbf{X}^T \qquad \frac{\partial L}{\partial \mathbf{w}} = \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{w}}$$

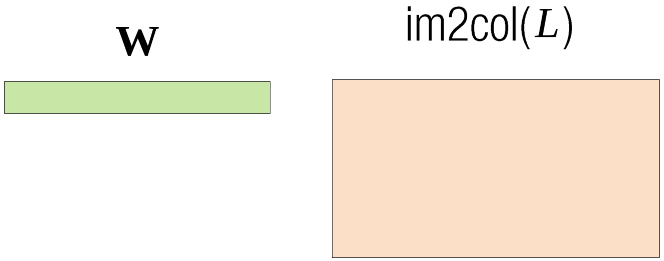
Convolution via im2col (Derivative w.r.t. X)



$= \dfrac{\partial L}{\partial \mathbf{x}}$

$$\frac{\partial L}{\partial \mathbf{x}_{ij}} = \sum_k \sum_l L_{ij} \mathbf{w}_{i-k,\,j-l}$$

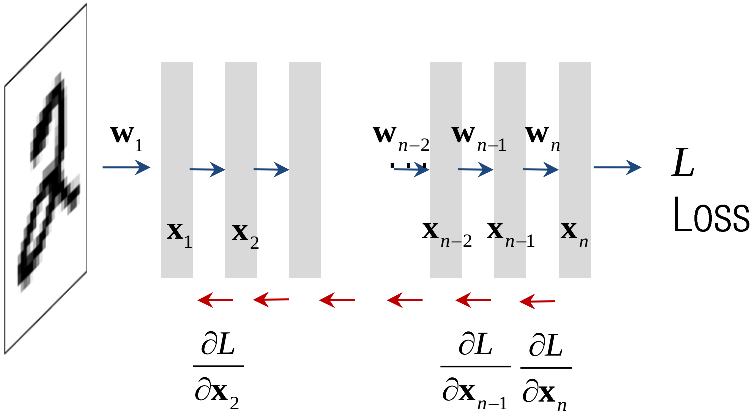
Reverse order



$= \dfrac{\partial L}{\partial \mathbf{x}}$

Row version of dLdx

SUMMARY



Forward prediction

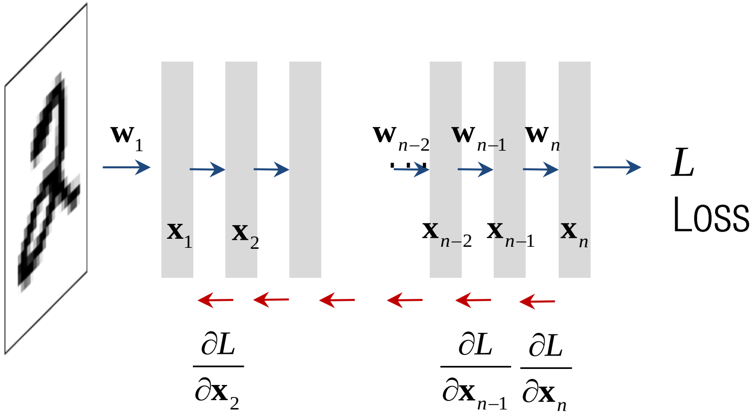
pred1 = conv(x, w1)
pred2 = relu(pred1)
pred3 = pool(pred2)

pred10 = flatten(pred9)
pred11 = fc(pred10, w10)

pred16 = loss_ce_sm(pred15, label)

function [dLdx dLdy] = foo_back(dL, x, y)

SUMMARY



function [dLdx dLdy] = foo_back(dL, x, y)

Forward prediction

pred1 = conv(x, w1)
pred2 = relu(pred1)
pred3 = pool(pred2)

pred10 = flatten(pred9)
pred11 = fc(pred10, w10)

pred16 = loss_ce_sm(pred15, label)

Back-propagation

dLdx = loss_ce_back(pred15, label)

dLdx, dLdw10 = fc_back(pred10, w10, pred11)
dLdx = flatten_back(dLdx, pred9, pred10)

dLdx = pool_back(dLdx, pred2, pred3)
dLdx = relu_back(dLdx, pred1, pred2)
dLdx, dLdw1 = conv_back(dLdx, x, w1,pred1)