CMPUT 328 Midterm Review

October 8 2025

Calc Background for BP

When solving BP questions, you need to generalize these rules to vector-valued functions.

You only need the rules highlighted, because the components you see in NNs are pretty limited!

(think of your matrix-multiplications, ReLUs, CEs, etc., all very repetitive!)

Constant Rule: $\frac{d}{dx}(c) = 0$

Constant Multiple Rule: $\frac{d}{dx}[cf(x)] = cf'(x)$

Power Rule: $\frac{d}{dx}(x^n) = nx^{n-1}$

Sum Rule: $\frac{d}{dx}[f(x)+g(x)] = f'(x)+g'(x)$

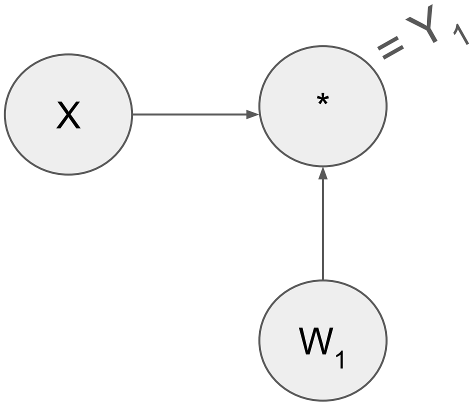
Difference Rule: $\frac{d}{dx}[f(x)-g(x)] = f'(x)-g'(x)$

Product Rule: $\frac{d}{dx}[f(x)g(x)] = f(x)g'(x)+g(x)f'(x)$

Quotient Rule: $\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] = \frac{g(x)f'(x)-f(x)g'(x)}{[g(x)]^2}$

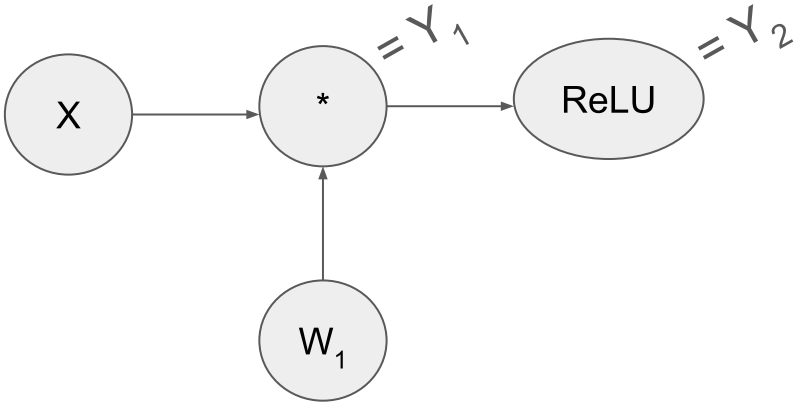
Chain Rule: $\frac{d}{dx}f(g(x)) = f'(g(x))g'(x)$

Neural Networks are Functions



Equivalently,
$Y_1 = X W_1$

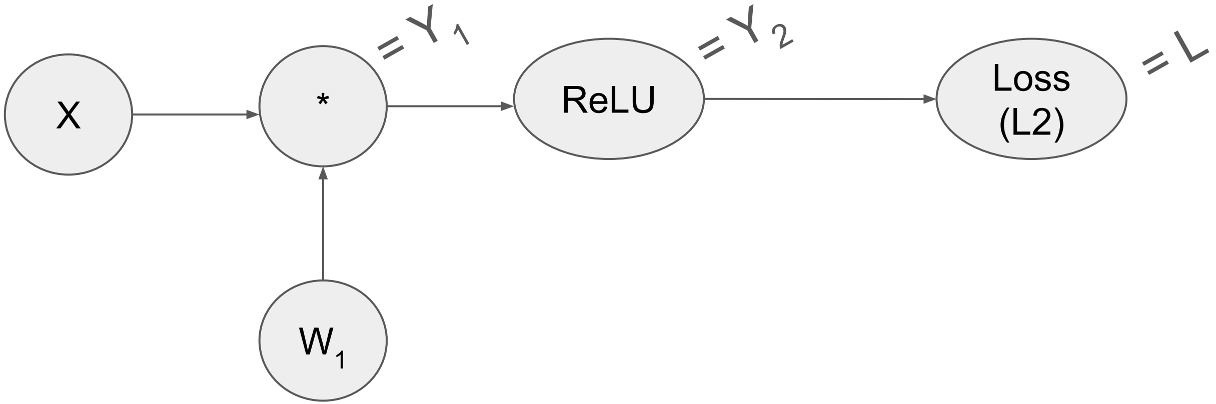
Neural Networks are (Compositions of) Functions



Equivalently,
$Y_1 = X\,W_1$

$Y_2 = \mathrm{ReLU}(X\,W_1) = \mathrm{ReLU}(Y_1)$

Neural Networks are (Compositions of) Functions

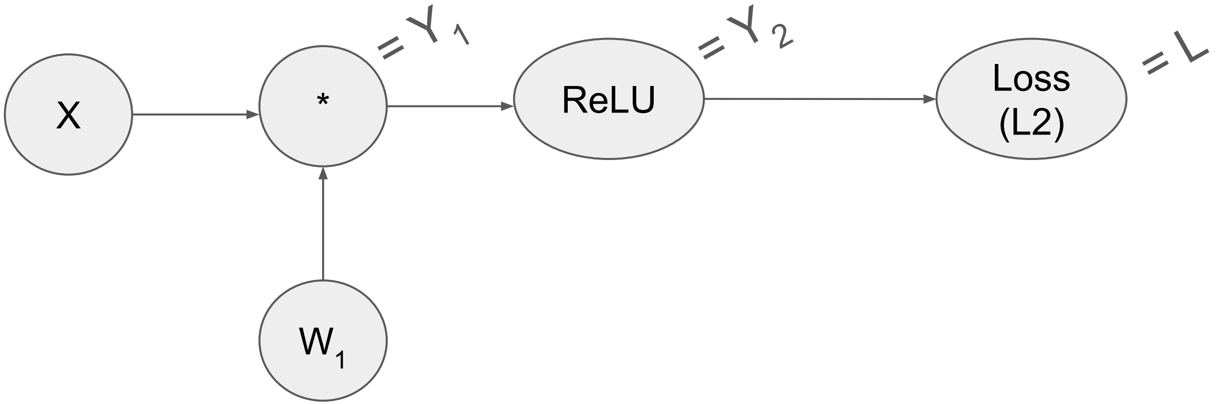


$Y_1 = X\ W_1$

$Y_2 = \mathrm{ReLU}(X\ W_1) = \mathrm{ReLU}(Y_1)$

$L = ½\ ||Y_2 - Y||^2 = ½\ ||\mathrm{ReLU}(Y_1) - Y||^2 = ½\ ||\mathrm{ReLU}(Y_1) - Y||^2 = ½\ ||\mathrm{ReLU}(X\ W_1) - Y||^2$

Partials of What?



To learn, the network needs to know how to change its parameters (W_1 in this case) to minimize the loss (L2 in this case).

The first step to this is computing the partial derivative of L with respect to W_1

Side Mark - Gradient Descent

To learn, the network needs to know how to change its parameters (W_1 in this case) to minimize the loss (L2 in this case).

The first step to this is computing the partial derivative of L with respect to W_1

$$\text{Repeat until converge } \{$$

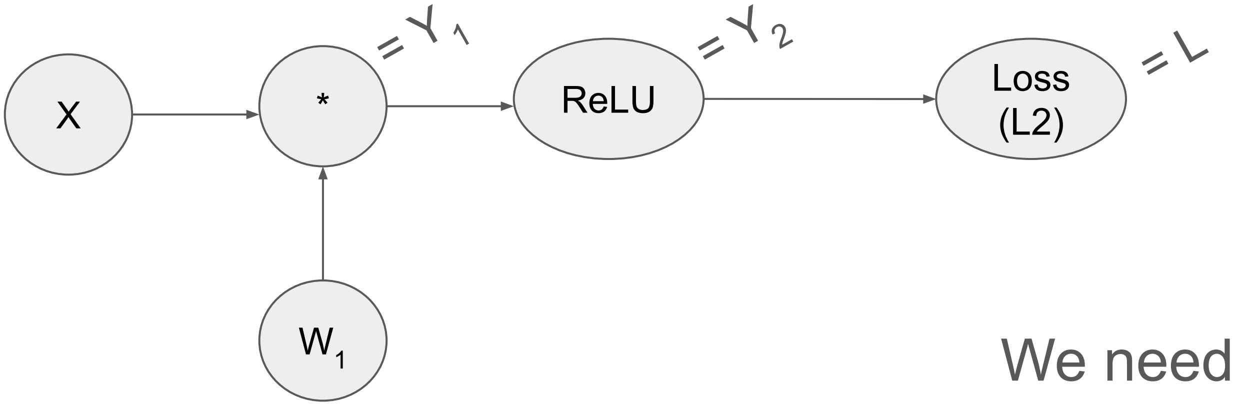
$$\mathrm{w} = \mathrm{w} - \alpha \left[\frac{\partial Loss}{\partial w}\right]$$

$$\mathrm{b} = \mathrm{b} - \alpha \left[\frac{\partial Loss}{\partial b}\right]$$

$$\}$$

Where does this update even come from?? I will tell you later. Otherwise, memorize this because you will see it all the time.

Notation

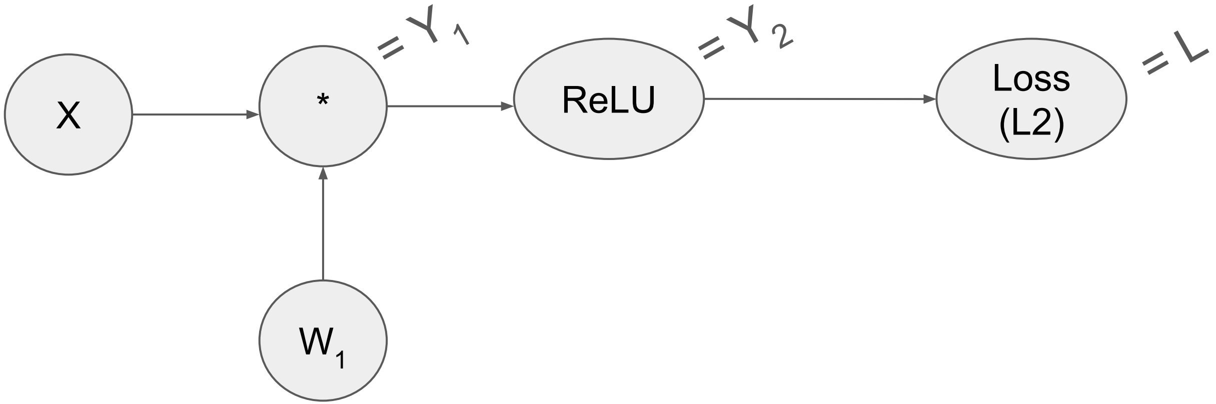


We need to solve for

$\partial L / \partial W_1 = \delta W_1$
(assume wrt L always)

“dell vs delta”

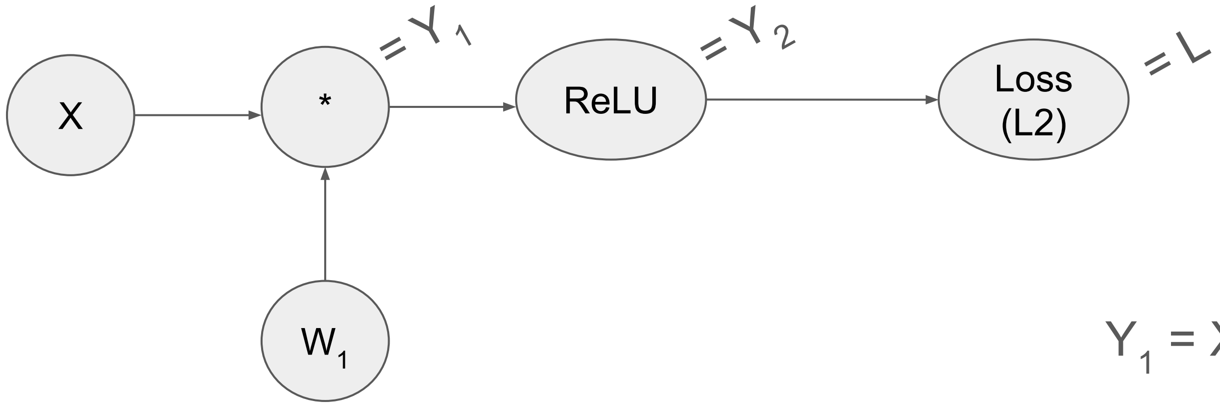
1. Draw the chain (done)



Tips:

- Nodes should be **operations** (mat-mul, activation functions, loss functions)
- Label the intermediate variables (I default to y_n)
- Draw incoming parameters and targets vertically (like how W_1 is drawn)

2. Write out Intermediate Function Definitions (optional, helps me)

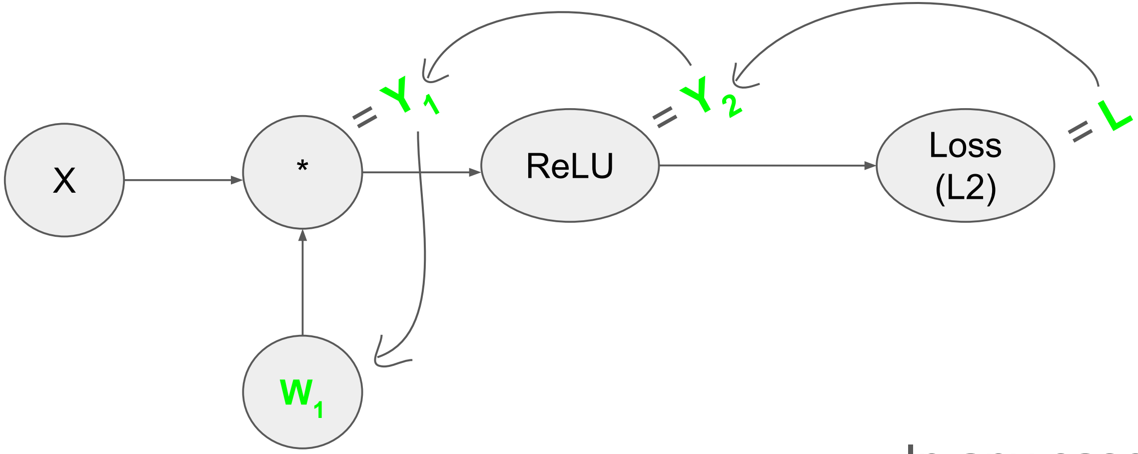


$Y_1 = X\,\boxed{W_1}$

$Y_2 = \text{ReLU}(X\,W_1) = \text{ReLU}(\boxed{Y_1})$

$L = \frac{1}{2}\,||\boxed{Y_2} - Y||^2$

3. Trace the chain/definitions for partials



$Y_1 = X\, W_1$

$Y_2 = \mathrm{ReLU}(X\, W_1) = \mathrm{ReLU}(Y_1)$

$L = ½\, ||Y_2 - Y||^2$

In any case, you should see

$\partial L/\partial W_1$

$= \partial L/\partial Y_2 \quad \partial Y_2/\partial Y_1 \quad \partial Y_1/\partial W_1$

4. Plug and Chug the Derivative Definitions (START AT THE END BC THAT'S THE POINT OF BP)

In any case, you should see

$$\partial L/\partial W_1$$

$$= \partial L/\partial Y_2 \quad \partial Y_2/\partial Y_1 \quad \partial Y_1/\partial W_1$$

1 2 3

$$Y_1 = X\, W_1$$

$$Y_2 = \mathrm{ReLU}(X\, W_1) = \mathrm{ReLU}(Y_1)$$

$$L = ½\, ||Y_2 - Y||^2$$

4. Plug and Chug the Derivative Definitions

$$= \boxed{\partial L/\partial Y_2} \quad \partial Y_2/\partial Y_1 \quad \partial Y_1/\partial W_1$$

1 2 3

I need $\partial L/\partial Y_2$

$\frac{\partial L}{\partial y_p}$

$L = ½ ||Y_2 - Y||^2 = ½ (Y_2 - Y)^2$

$\partial L/\partial Y_2 = (Y_2 - Y)$

$Y_1 = X W_1$

$Y_2 = \text{ReLU}(X W_1) = \text{ReLU}(Y_1)$

$L = ½ ||Y_2 - Y||^2$

$$\frac{\partial \frac{1}{2} ||Y_2 - Y||^2}{\partial Y_2} =$$

4. Plug and Chug the Derivative Definitions

$= \partial L/\partial Y_2 \quad \boxed{\partial Y_2/\partial Y_1} \quad \partial Y_1/\partial W_1$

1 2 3

$Y_1 = X\, W_1$

$Y_2 = \mathrm{ReLU}(X\, W_1) = \mathrm{ReLU}(Y_1)$

$L = ½\, ||Y_2 - Y||^2$

I need $\partial Y_2/\partial Y_1$

$Y_2 = \mathrm{ReLU}(X\, W_1) = \mathrm{ReLU}(Y_1)$

Wait… what's the derivative of ReLU??

Side Mark: ReLU Derivative and Jacobians

I need $\partial Y_2 / \partial Y_1$

$Y_2 = \mathrm{ReLU}(X W_1) = \mathrm{ReLU}(Y_1)$

The Jacobian is the collection of all partial derivatives of a function. It is a matrix.

Think of it as a generalization of the gradient to vector functions.

$\begin{bmatrix} 0 \\ -10 \\ 1 \end{bmatrix} \xrightarrow{\mathrm{Relu}} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

$\frac{\partial f_1}{\partial x_1}$ $\frac{\partial f_1}{\partial x_2}$

$$\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$$

$$\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})) \in \mathbb{R}^m$$

$$\mathbf{J_f} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

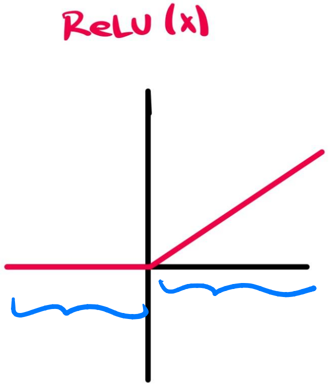
Notice how the columns are consistent across input variables

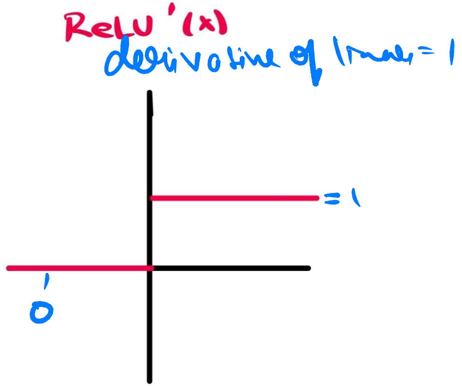
Side Mark: ReLU Derivative and Jacobians

ReLU is piecewise, so we can compute its piecewise derivative. It's technically continuous but not differentiable, but we don't really care for BP.

ReLU = {0 if x ≤ 0
x if x > 0}

ReLU' = {0 if x ≤ 0
1 if x > 0}





Side Mark: ReLU Derivative and Jacobians

say $x = [-10, 10, 0]$

so $\text{ReLU}(x) = [0, 10, 0]$

also, $\text{ReLU} : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ (in this example)

Jacobian has 3 columns 3 rows

$$\begin{bmatrix} \frac{\partial R_1}{\partial x_1} & \frac{\partial R_1}{\partial x_2} = 0 & \frac{\partial R_1}{\partial x_3} = 0 \\ \frac{\partial R_2}{\partial x_1} = 0 & \frac{\partial R_2}{\partial x_2} & \frac{\partial R_2}{\partial x_3} = 0 \\ \frac{\partial R_3}{\partial x_1} = 0 & \frac{\partial R_3}{\partial x_2} = 0 & \frac{\partial R_3}{\partial x_3} \end{bmatrix}$$

$$\text{relu}(x) = [R_1(x_1), R_2(x_2), R_3(x_3)]$$

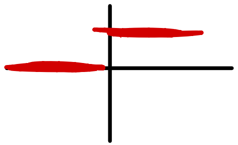
Side Mark: ReLU Derivative and Jacobians

$$\begin{bmatrix} \frac{\partial R_1}{\partial x_1} & \frac{\partial R_1}{\partial x_2} & \frac{\partial R_1}{\partial x_3} \\ \frac{\partial R_2}{\partial x_1} & \frac{\partial R_2}{\partial x_2} & \frac{\partial R_2}{\partial x_3} \\ \frac{\partial R_3}{\partial x_1} & \frac{\partial R_3}{\partial x_2} & \frac{\partial R_3}{\partial x_3} \end{bmatrix} = \begin{bmatrix} 0/1 & \cancel{0/1}\ 0 & \cancel{0/1}\ 0 \\ \cancel{0/1}\ 0 & 0/1 & \cancel{0/1}\ 0 \\ \cancel{0/1}\ 0 & \cancel{0/1}\ 0 & 0/1 \end{bmatrix}$$

since ReLU' = {0 if $x \leq 0$
1 if $x > 0$}

Side Mark: ReLU Derivative and Jacobians

relu'



say $x = [-10, 10, 0]$

so $\mathrm{ReLU}(x) = [0, 10, 0]$

also, $\mathrm{ReLU} : \mathbb{R}^3 \to \mathbb{R}^3$ (in this example)

Jacobian has 3 columns 3 rows

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

ReLU' = {0 if x ≤ 0
1 if x > 0}

Back Home: ReLU Derivative and Jacobians

I need $\partial Y_2/\partial Y_1$

$Y_2 = \text{ReLU}(X\ W_1) = \text{ReLU}(Y_1)$

Oh no, I didn't give you any values for X to compute the ReLU from!! All you do in this case is denote the partial as the Jacobian, maybe like

$$\partial Y_2/\partial Y_1 = J_{relu}$$

4. Plug and Chug the Derivative Definitions

$= \cancel{\partial L/\partial Y_2} \ \cancel{\partial Y_2/\partial Y_1} \ \boxed{\partial Y_1/\partial W_1}$

1 2 3

$Y_1 = X W_1$

$Y_2 = \text{ReLU}(X W_1) = \text{ReLU}(Y_1)$

$L = ½ ||Y_2 - Y||^2$

I need $\partial Y_1/\partial W_1$

$Y_1 = X W_1$

that's easy, we just need the mat-mul rule, ie.,

$\partial Y_1/\partial W_1 = X$

5. Put it all together!

$$= \partial L/\partial Y_2 \quad \partial Y_2/\partial Y_1 \quad \partial Y_1/\partial W_1$$

1 2 3

$Y_1 = X\, W_1$

$Y_2 = \mathrm{ReLU}(X\, W_1) = \mathrm{ReLU}(Y_1)$

$L = ½\, ||Y_2 - Y||^2$

$\partial L/\partial Y_2 = (Y_2 - Y)$ → L2 der.

$\partial Y_2/\partial Y_1 = J_{relu}$ → relu der.

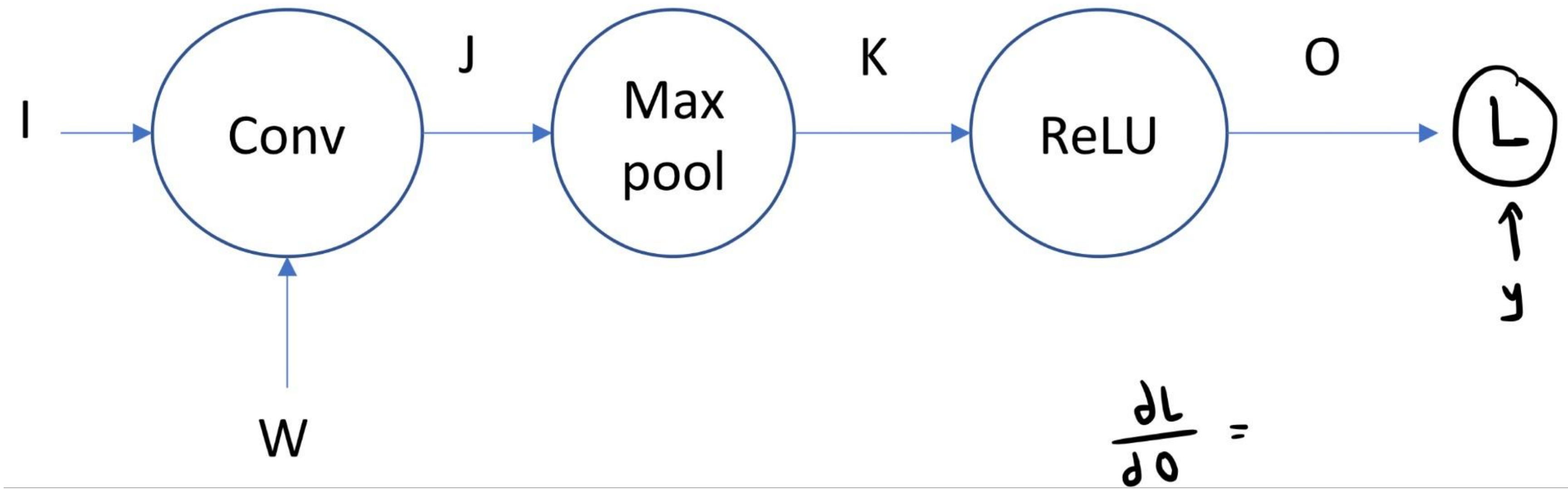
$\partial Y_1/\partial W_1 = X$

so

$\partial L/\partial W_1 = (Y_2 - Y)\, J_{relu}\, X$

How about BP for Convolutions, Max Pooling, etc.?

How about BP for Convolutions, Max Pooling, etc.?



How about BP for Convolutions, Max Pooling, etc.?

MAX POOL

$$\frac{\partial L}{\partial J} = \frac{\partial L}{\partial O} \frac{\partial O}{\partial K} \frac{\partial K}{\partial J}$$

$$K = \text{max\_pool}(J) \rightarrow \text{max\_pool}\begin{bmatrix} 2 & -4 & -6 & -1 \\ -1 & -9 & 9 & -2 \end{bmatrix}$$

zero out all not selected terms

$$\frac{\partial K}{\partial J} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

1×2 2×

$$\frac{\partial L}{\partial J} = \frac{\partial L}{\partial O} \frac{\partial O}{\partial K} \frac{\partial K}{\partial J} = \underset{1\times 2}{\begin{bmatrix} -5 & -1 \end{bmatrix}} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

CONV WRT FILTER

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial O} \frac{\partial O}{\partial K} \frac{\partial K}{\partial J} \frac{\partial J}{\partial W}$$

MAX POOL GRAD ALWAYS HAS SIZE OF INPUT

$$\begin{bmatrix} \delta k_1 & 0 & 0 & 0 \\ 0 & 0 & \delta k_2 & 0 \end{bmatrix}$$

$J = \text{conv}(x ; w)$

$$\begin{bmatrix} -5 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{bmatrix}$$

$$\frac{\partial L}{\partial W} = \text{conv}\left(x, \frac{\partial L}{\partial J}\right)$$

DO NEED $\frac{\partial L}{\partial J}$ to have 2×4 dim...

1	3	2	4	6	4
4	8	3	1	0	2
2	1	4	3	9	1
4	7	2	3	9	2

$$\begin{bmatrix} -5 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{bmatrix}$$

-5 -3

-20 -4

-15 -1

-10

-40 -3

-15 -9

$$\begin{bmatrix} -8 & -16 & -10 \\ -24 & -43 & -24 \\ -12 & -8 & -29 \end{bmatrix} = \frac{\partial K}{\partial J}$$

CONV WRT INPUT

$$\frac{\partial L}{\partial x} = \text{pad}\left(\frac{\partial L}{\partial J}\right) * \text{flip}(W)\}$$

$$\begin{bmatrix} -5 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{bmatrix}$$

ix	-1	0	1
	-1	0	1
	-1	0	1

$$\text{flip}(W) = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$$

$$\begin{bmatrix} -5 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{bmatrix}$$

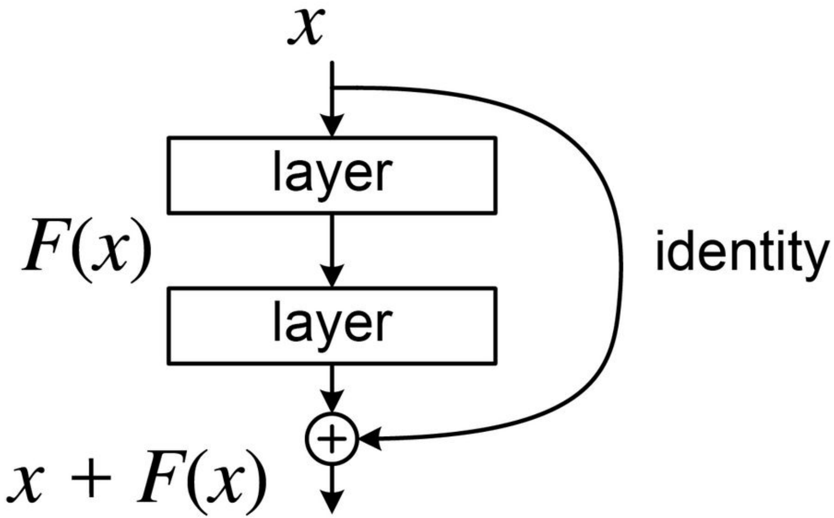
add
h-1 rows
w-1 cols

h x w
is dimension
of FILTER

$$\begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix} \quad \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$$

$$\frac{\partial L}{\partial x} = \begin{bmatrix} 5 & 0 & -5 & 0 & 0 & 0 \\ 5 & 0 & -4 & 0 & -1 & 0 \\ 5 & 0 & -4 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 \end{bmatrix}$$

A ResNet Example



Can you already see how this will affect the chain? It won't complicate the partials computation that much.

A Transformer Example

1.