Motivation for Back Propagation

A Santosh Kumar

Reference: Prof Andrew Ng DeepLearning.ai @ Coursera x 4

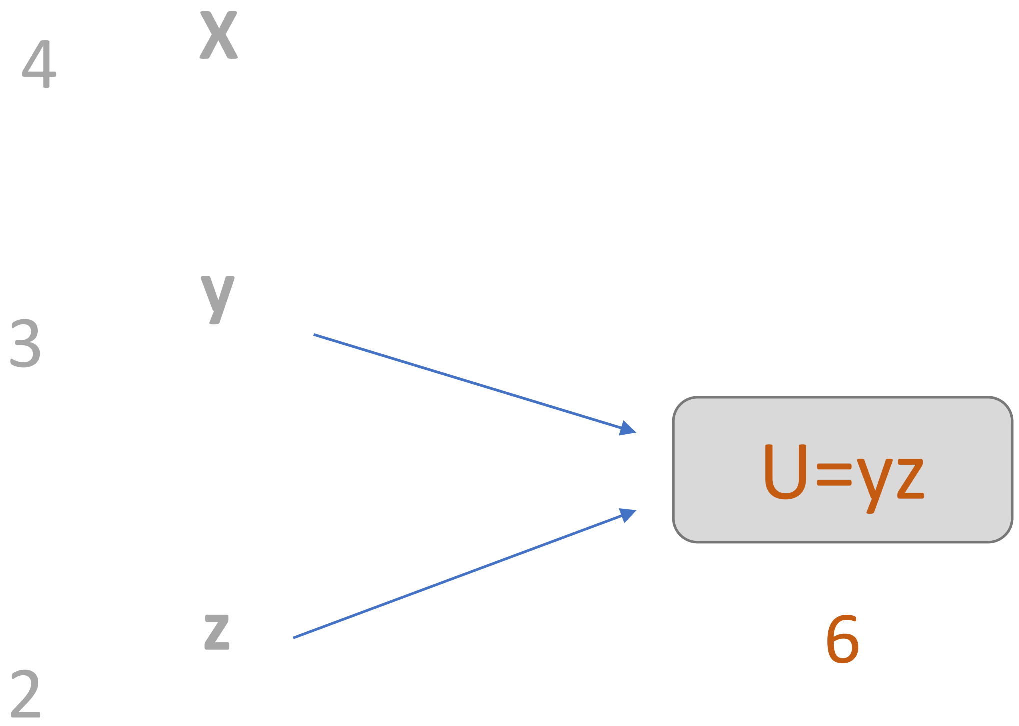
y 3

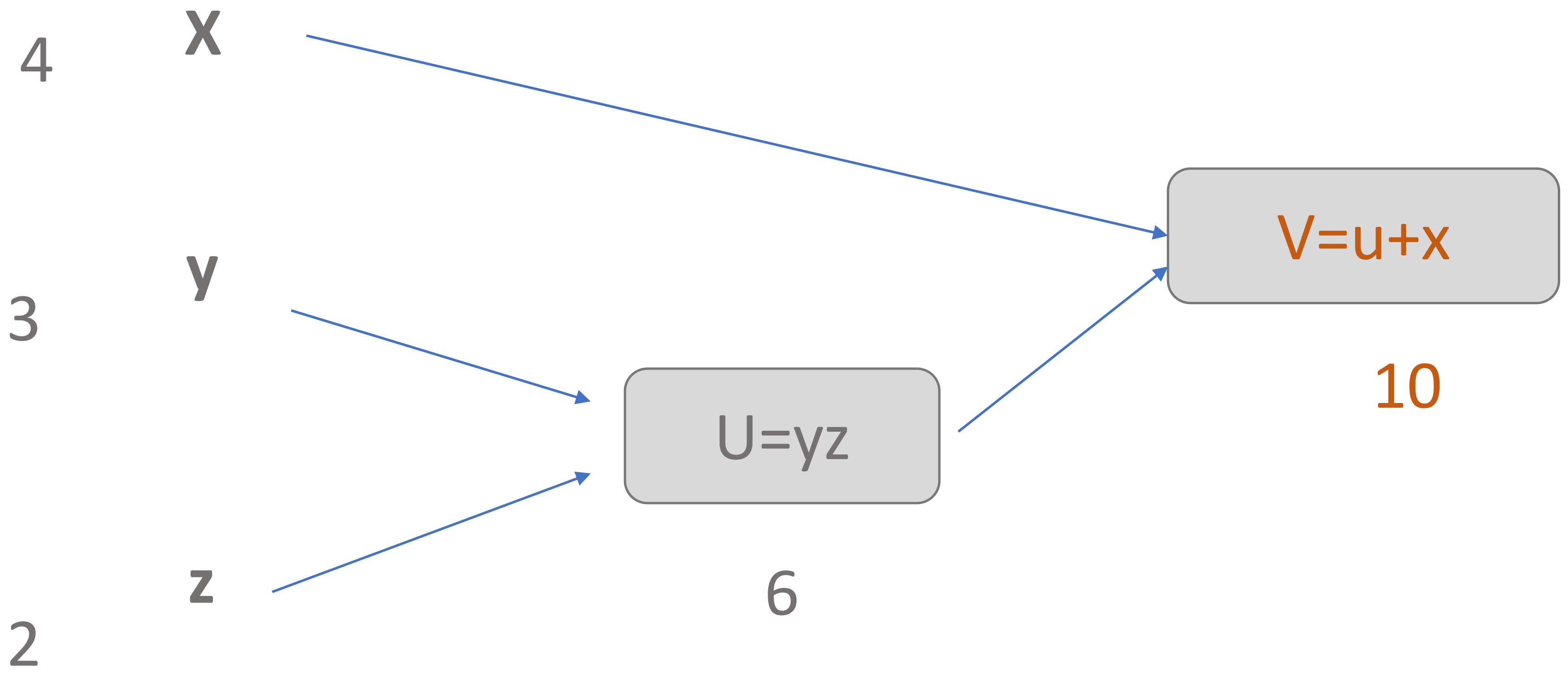
z 2

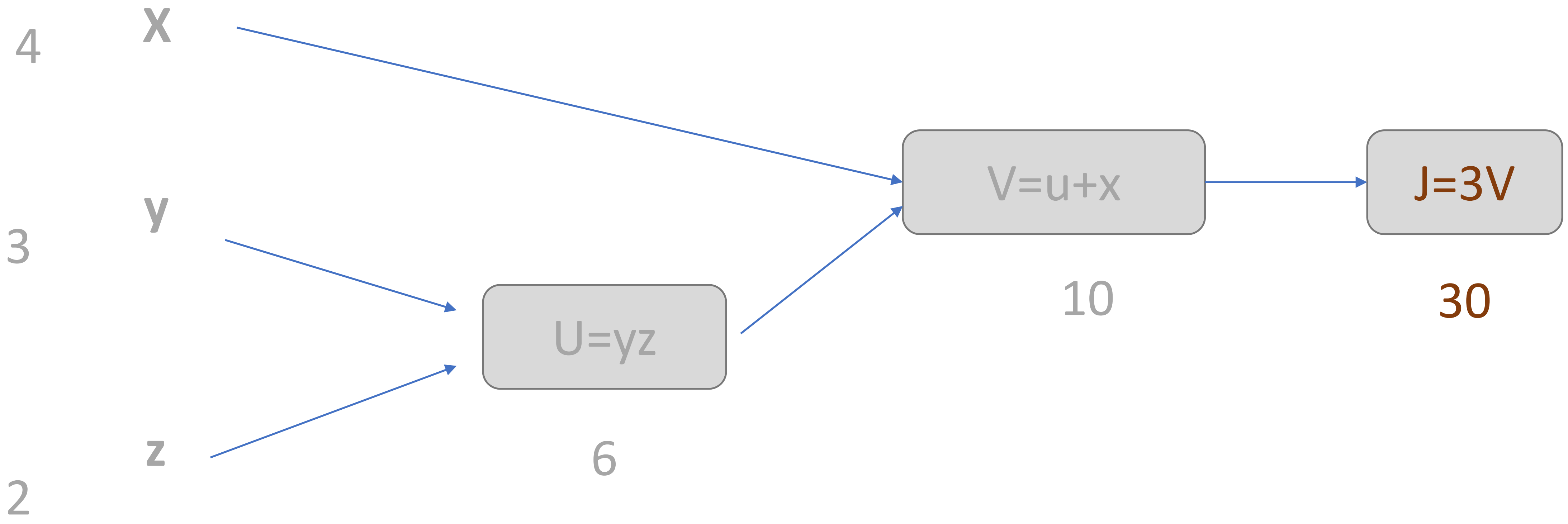
Let say we have three variables, **{x,y,z}** which determine the value of a function J.

Now, lets find value of function J using { **x, y, z**}

$$J = 3(x + yz)$$


4
x
3
y
U=yz
z
6
2


4
X
3
y
2
z
U=yz
6
V=u+x
10


4
x
3
y
2
z
U=yz
6
V=u+x
10
J=3V
30

Now, lets find how J varies with a infinitesimally small changes in { **x, y, z** }

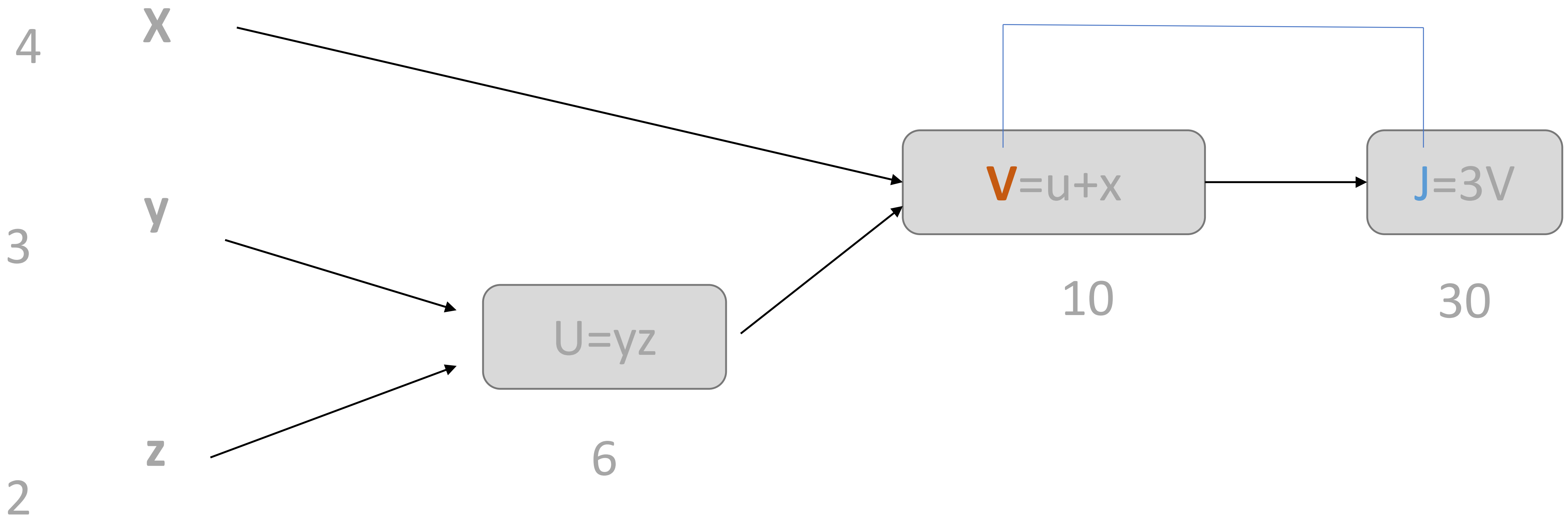
i.e

$$\frac{\partial J}{\partial x}, \frac{\partial J}{\partial y} \text{ and } \frac{\partial J}{\partial z}$$

To find $\frac{\partial J}{\partial x}$, $\frac{\partial J}{\partial y}$ and $\frac{\partial J}{\partial z}$

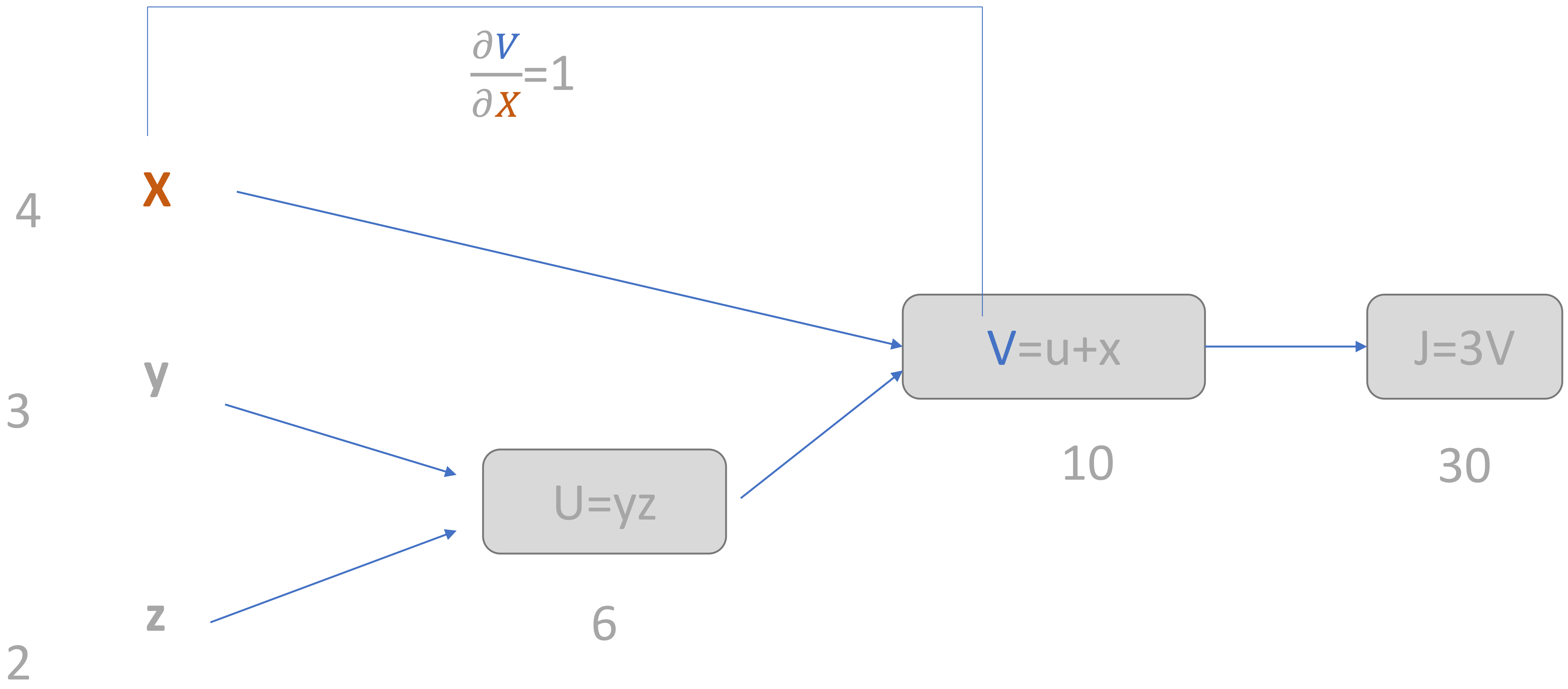
Let start with finding $\frac{\partial J}{\partial v}$

$$\frac{\partial J}{\partial V} = 3$$



So, $\frac{\partial J}{\partial V} = 3$ implies, change in J is three times the change in V.

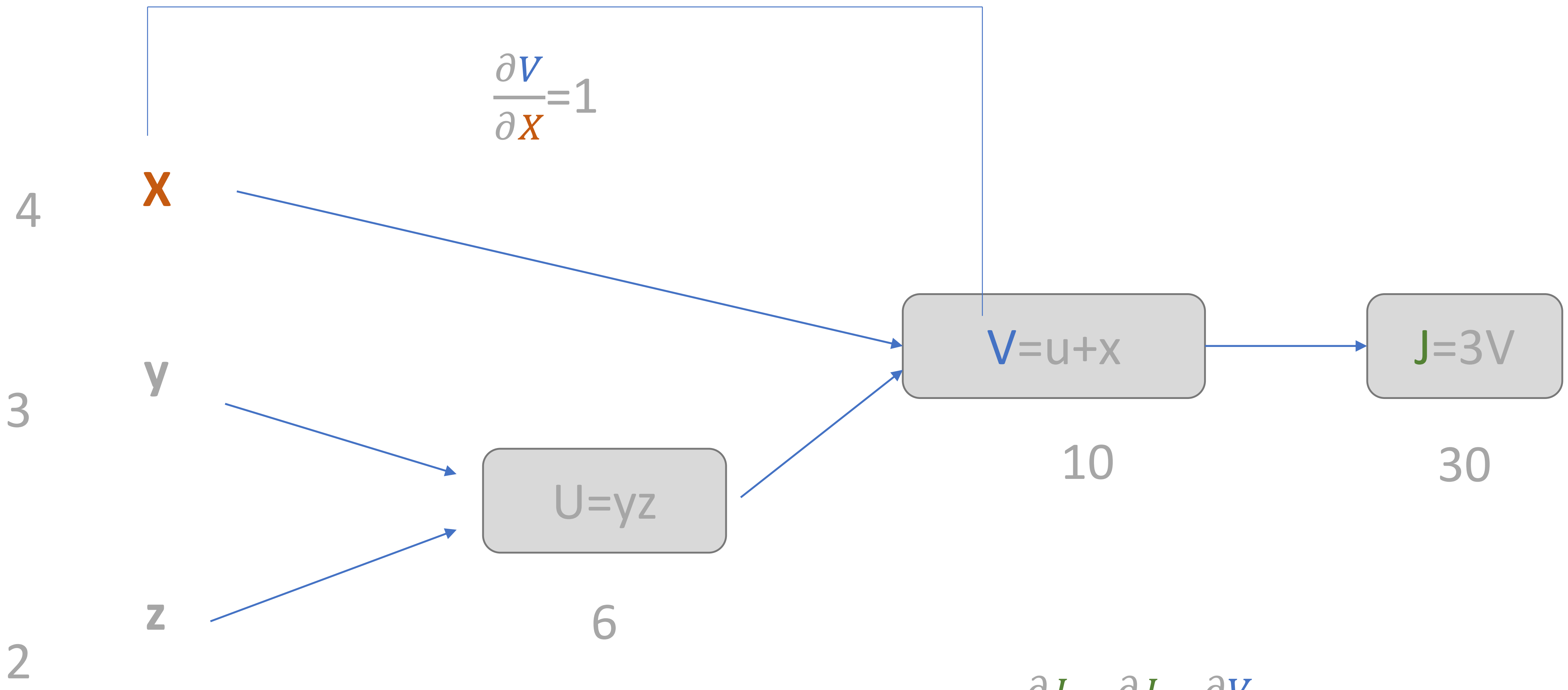
Now lets find $\frac{\partial v}{\partial x}$



$\frac{\partial V}{\partial X}=1$
4
X
3
y
2
z
U=yz
6
V=u+x
10
J=3V
30

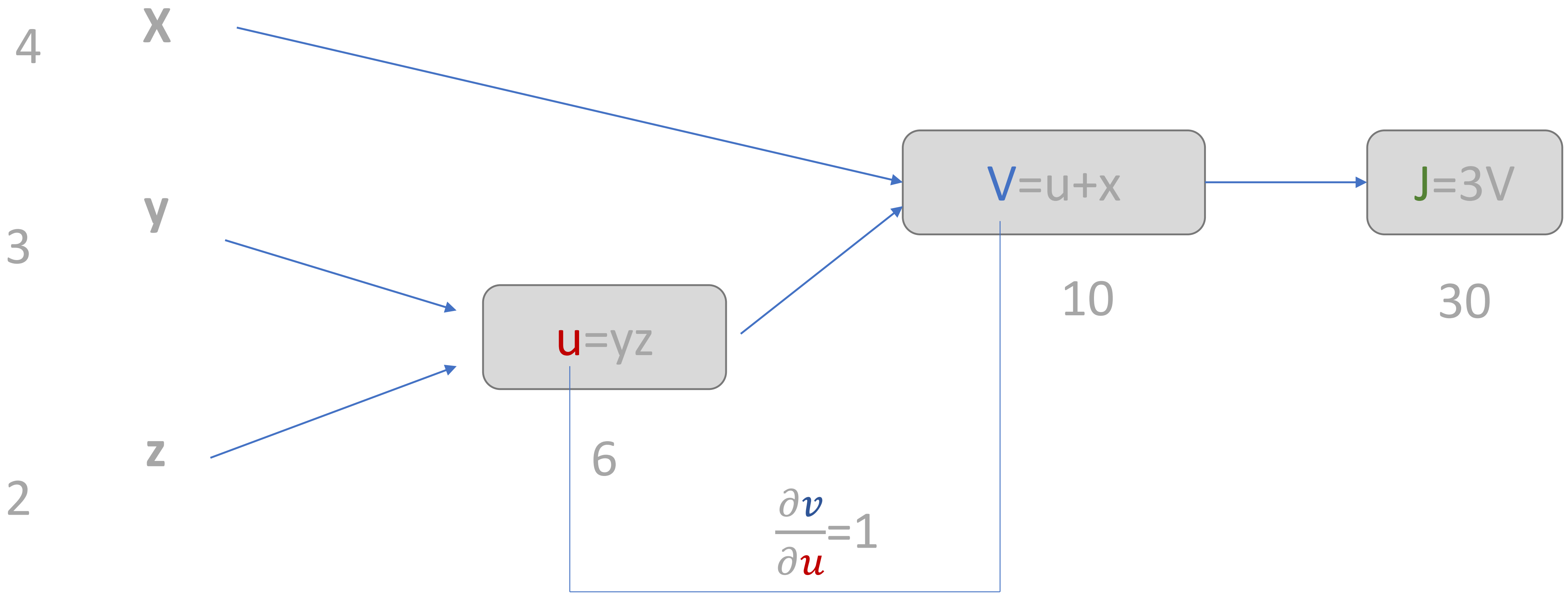
Now, we know $\frac{\partial J}{\partial v} = 3$ *and* $\frac{\partial v}{\partial x} = 1$

By using chain rule $\frac{\partial J}{\partial X} = \frac{\partial J}{\partial V} * \frac{\partial V}{\partial X} = 3*1 = 3$



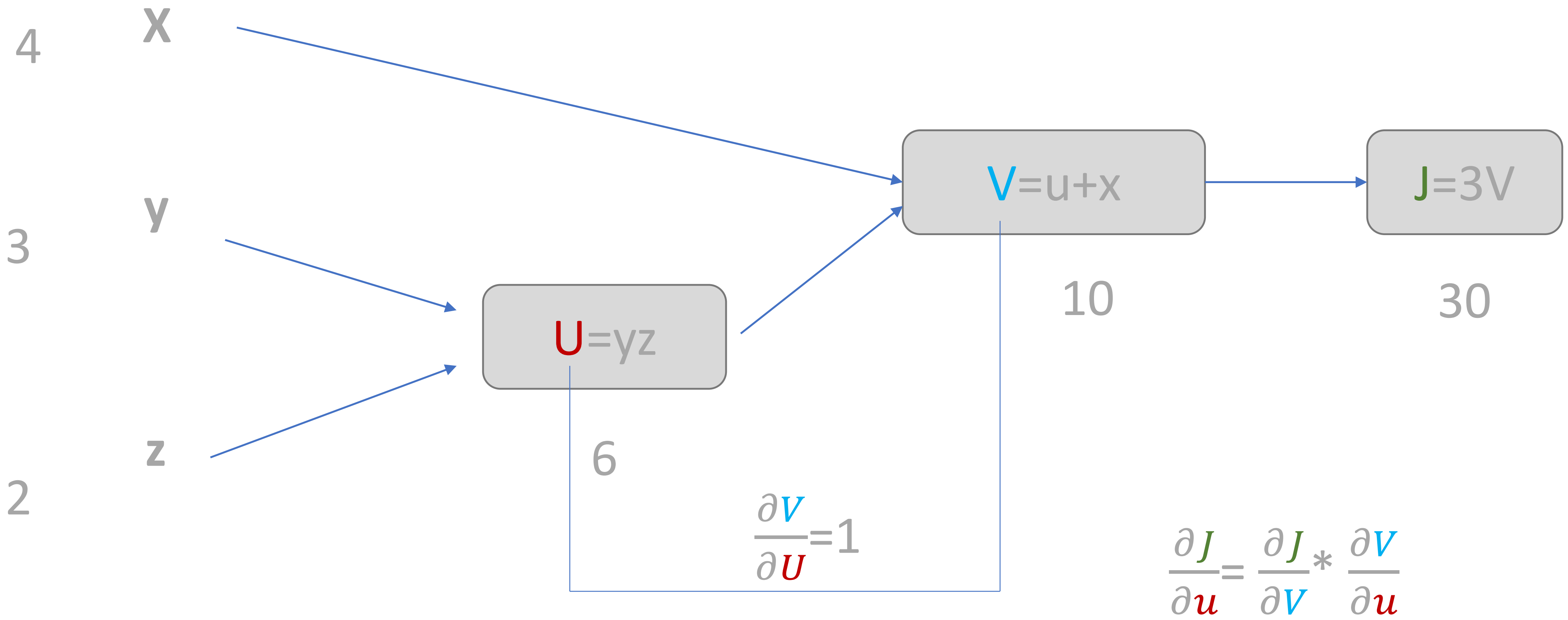
$$\frac{\partial J}{\partial X} = \frac{\partial J}{\partial V} * \frac{\partial V}{\partial X}$$
$$= 3*1 = 3$$

Now lets find $\dfrac{\partial v}{\partial u}$


4
x
3
y
2
z
u=yz
6
V=u+x
10
J=3V
30
$\frac{\partial v}{\partial u}$=1

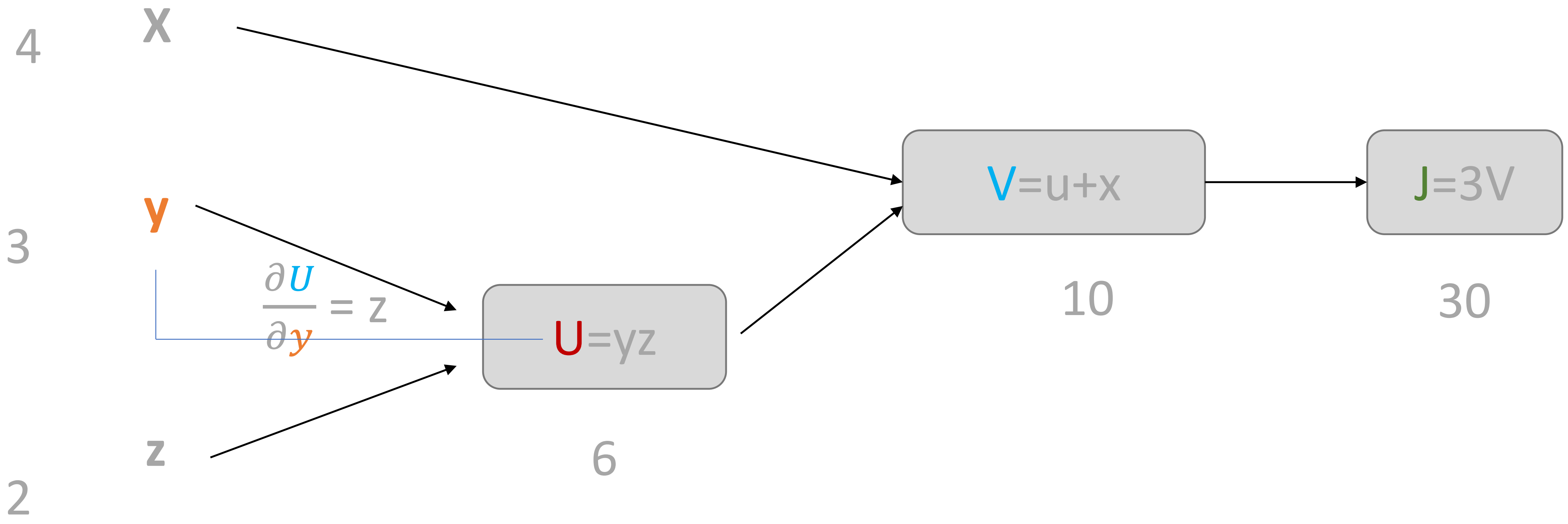
so, we found $\frac{\partial J}{\partial v} = 3 \; and \; \frac{\partial v}{\partial u} = 1$

Again, by using chain rule $\frac{\partial J}{\partial u} = \frac{\partial J}{\partial V} * \frac{\partial V}{\partial u} = 3*1 = 3$



4
x
3
y
2
z
U=yz
6
V=u+x
10
J=3V
30
$\frac{\partial V}{\partial U}=1$
$\frac{\partial J}{\partial u}=\frac{\partial J}{\partial V}*\frac{\partial V}{\partial u}$

Now lets find $\dfrac{\partial u}{\partial y}$

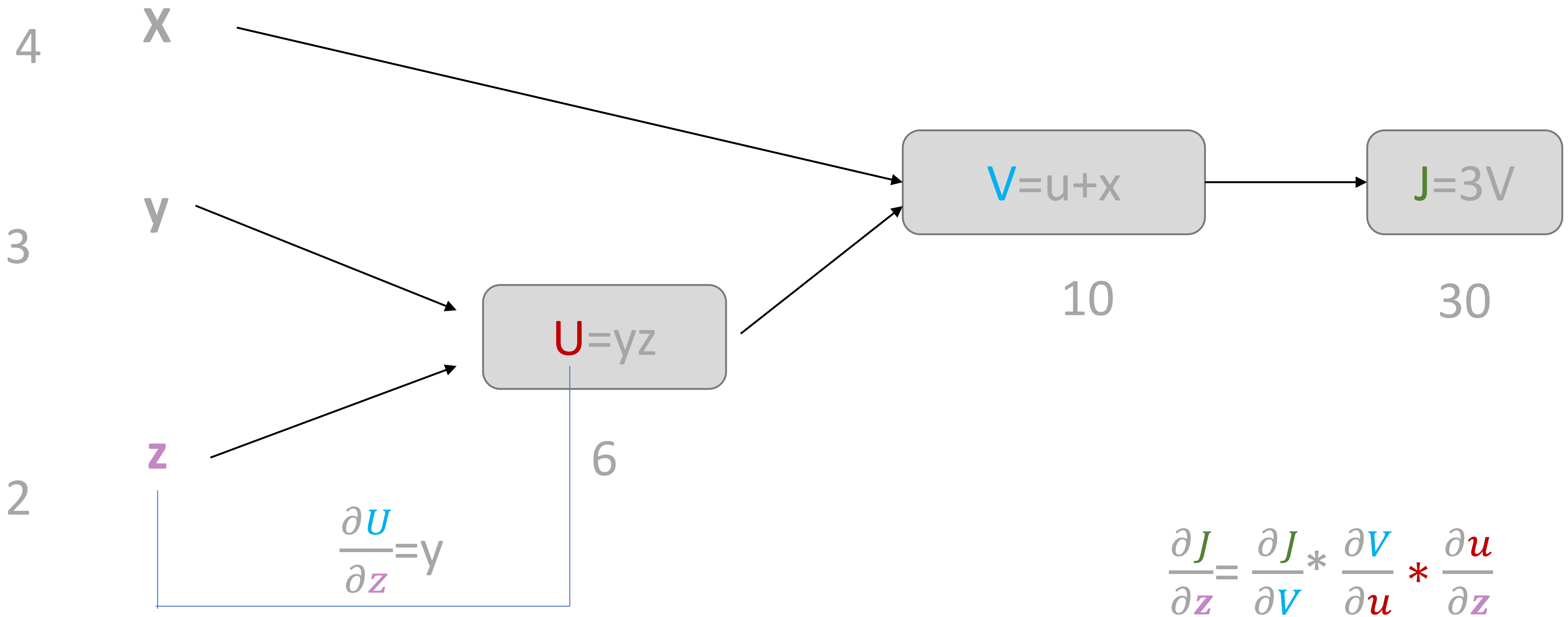


$$\frac{\partial J}{\partial y} = \frac{\partial J}{\partial V} * \frac{\partial V}{\partial u} * \frac{\partial u}{\partial y}$$

so, $\frac{\partial J}{\partial v} = 3$, $\frac{\partial v}{\partial u} = 1$ and $\frac{\partial U}{\partial y} = z$

Again, by using chain rule $\frac{\partial J}{\partial y} = \frac{\partial J}{\partial V} * \frac{\partial V}{\partial u} * \frac{\partial u}{\partial y}$ = 3*1*z =3z

Now lets find $\frac{\partial u}{\partial z}$



$$\frac{\partial J}{\partial z} = \frac{\partial J}{\partial V} * \frac{\partial V}{\partial u} * \frac{\partial u}{\partial z}$$

so, $\frac{\partial J}{\partial v} = 3$, $\frac{\partial v}{\partial u} = 1$ and $\frac{\partial U}{\partial z} =$ y

Again, by using chain rule $\frac{\partial J}{\partial z} = \frac{\partial J}{\partial V} * \frac{\partial V}{\partial u} * \frac{\partial u}{\partial z}$ = 3*1*y =y so,   we found how J varies w. r. t each of variable { x, y , z} as

$$\frac{\partial J}{\partial x} = 3 \ , \quad \frac{\partial v}{\partial y} = z \ \ and \quad \frac{\partial J}{\partial z} = y$$

$Now,\ let\ arrange\ \frac{\partial J}{\partial x} = 3\ ,\ \ \frac{\partial v}{\partial y} = z\ \ and\ \ \frac{\partial J}{\partial z}$ = y in a vector format

$$\left[\frac{\partial J}{\partial x}, \qquad \frac{\partial v}{\partial y}, \qquad \frac{\partial J}{\partial z}\right]$$

This is called gradient of function J.

Denoted as $\nabla J = \left[\frac{\partial J}{\partial x}, \ \frac{\partial v}{\partial y}, \ \frac{\partial J}{\partial z}\right]$

Gradient helps in finding out maxima or minima of function w.r.t variables.

In our example maxima or minima of **J** w.r.t **x, y and z**.

This quantity is very often used in various optimization techniques.

Single Neuron Network

lets consider a network where each training point has two features x1 and x2.

For ex. Lets consider digital AND gate

x1	x2	y
0	0	0
0	1	0
1	0	0
1	1	1

Here we have n=4 training points, each having two features x1 and x2 with label y

Lets say, for given training point i the output of network is y_{i}.

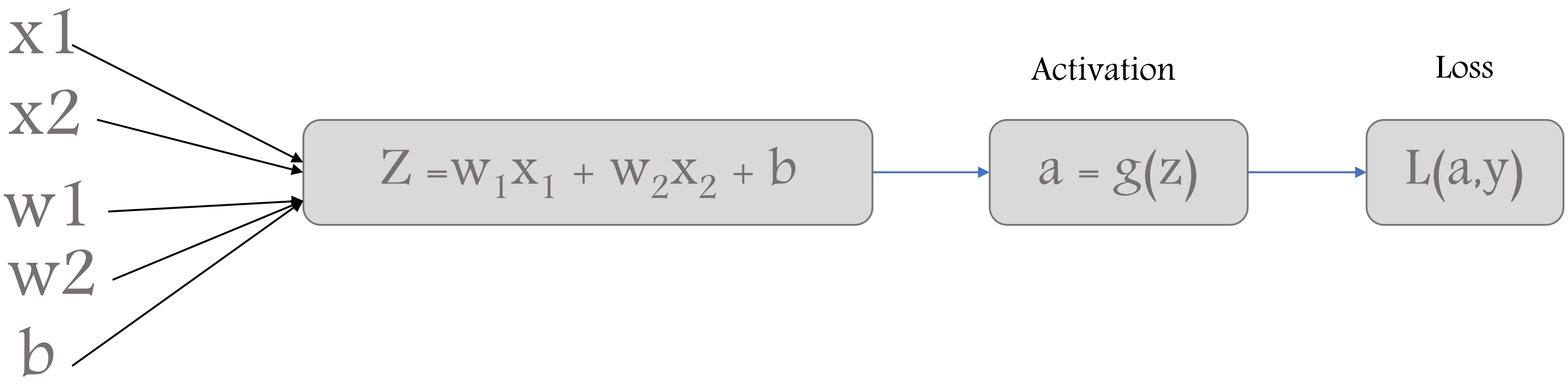
How well the neural network has done the job in predicting correctly? For that we need loss function.

Here, L is loss function, let us define for n training points L = $(y_i - y_{pi})^2$

Single Neuron Network

lets consider a network where each training point has two features x1 and x2.

For ex.



considering all the training points

$$L = \sum_{i=0}^{n} (y_i - y_{pi})^2$$

our objective is to make this loss minimum. i.e optimize L .

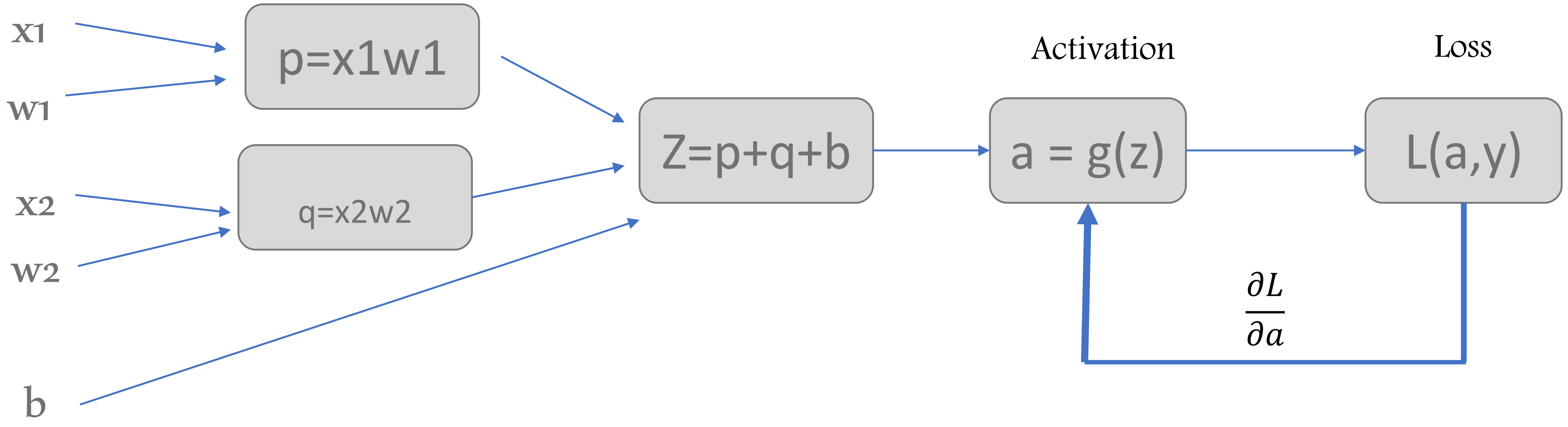
Note: we have to optimize L w.r.t weights w1, w2 and b . i.e we have to find set of { w1, w2 , b} such that our neural network has minimum loss .

So, we have to find gradient and for that we need $\left[\frac{\partial L}{\partial w_1}, \frac{\partial L}{\partial w_2}, \frac{\partial L}{\partial b}\right]$

Single Neuron Network

lets consider a network where each training point has two features x1 and x2.

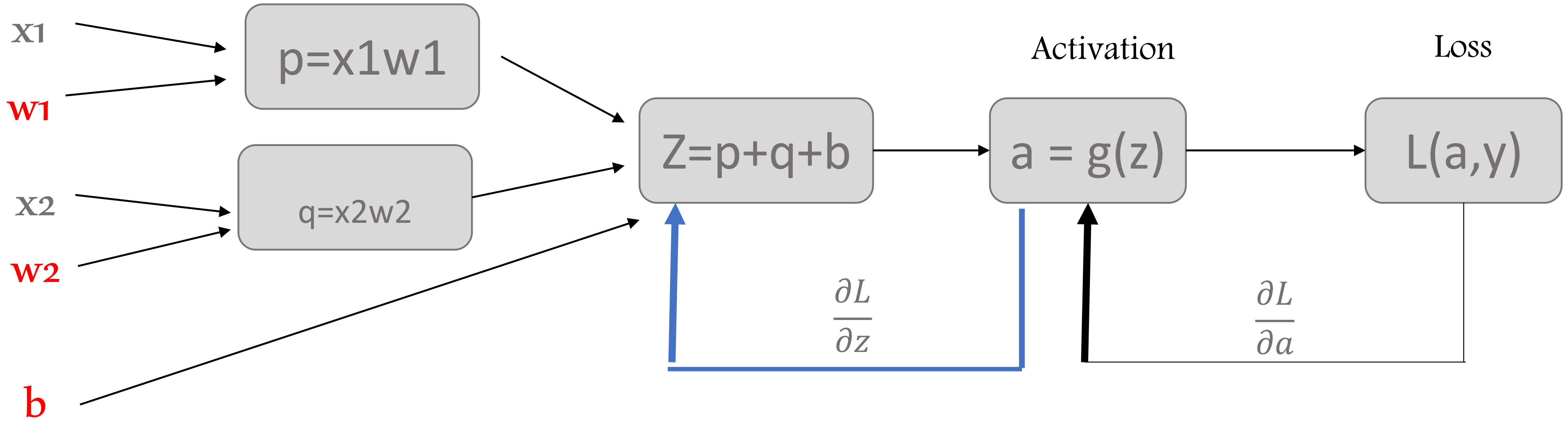
For ex.



Single Neuron Network

lets consider a network where each training point has two features x1 and x2.

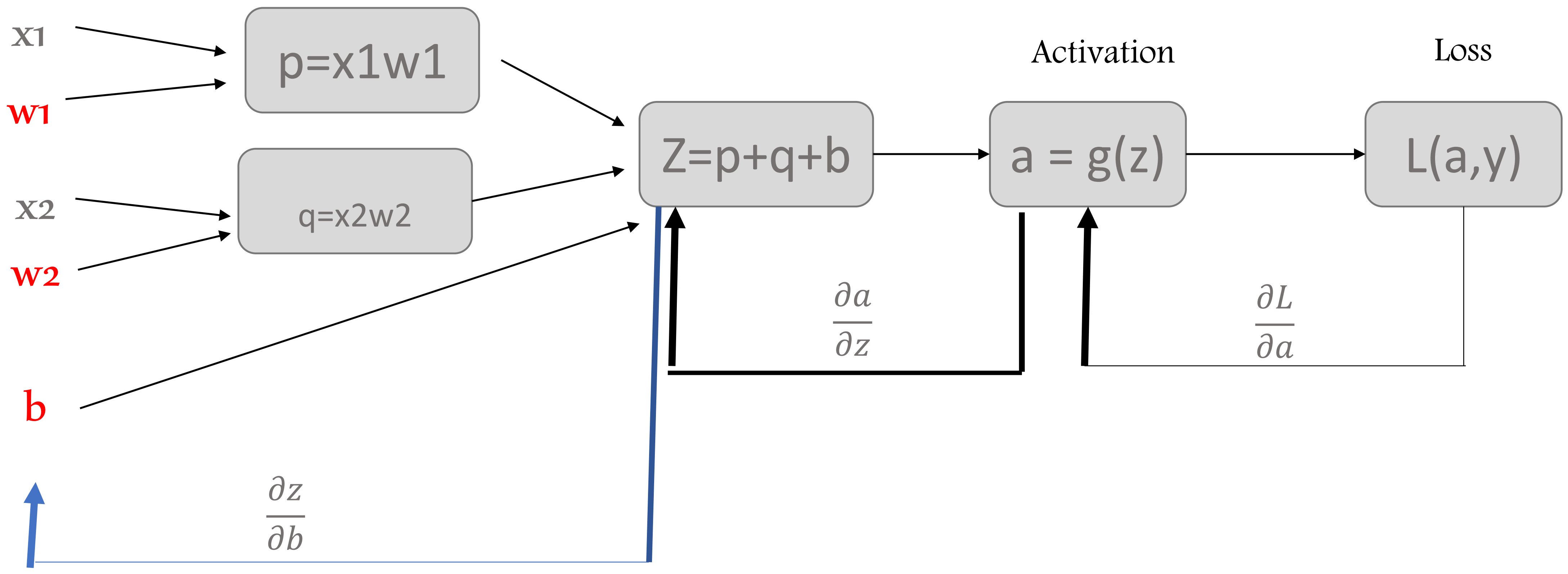
For ex.



Single Neuron Network

lets consider a network where each training point has two features x1 and x2.

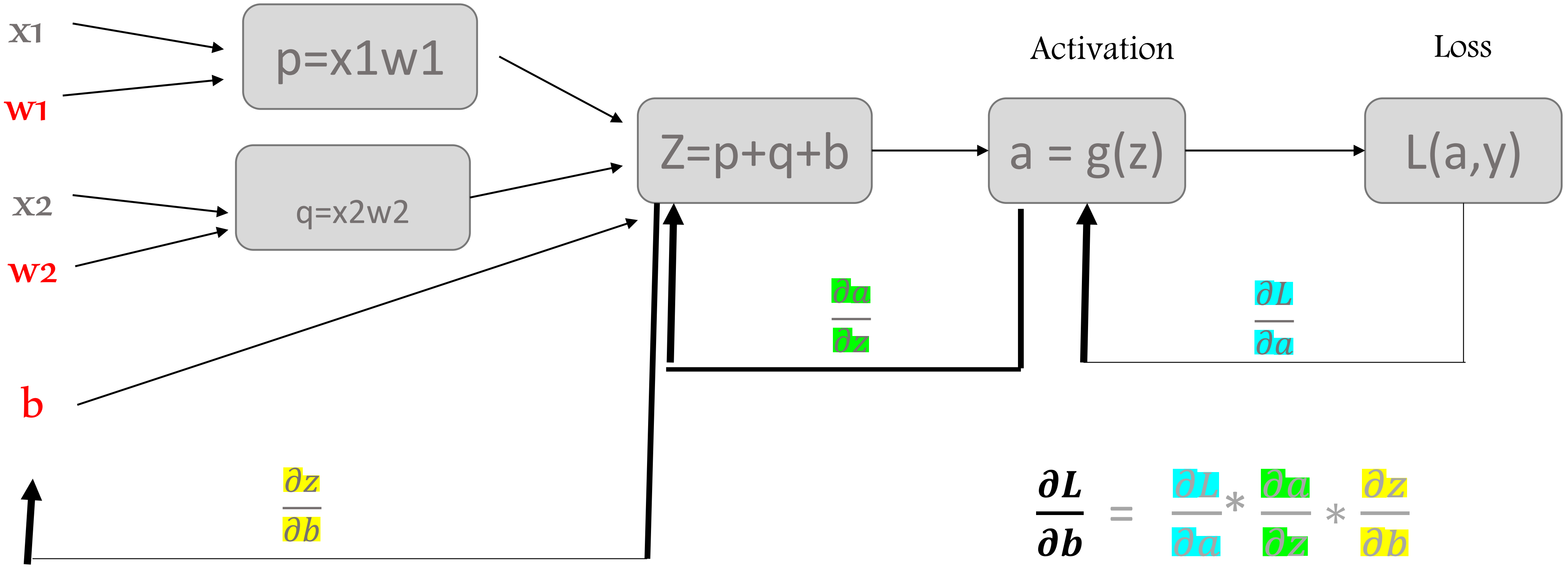
For ex.



Single Neuron Network

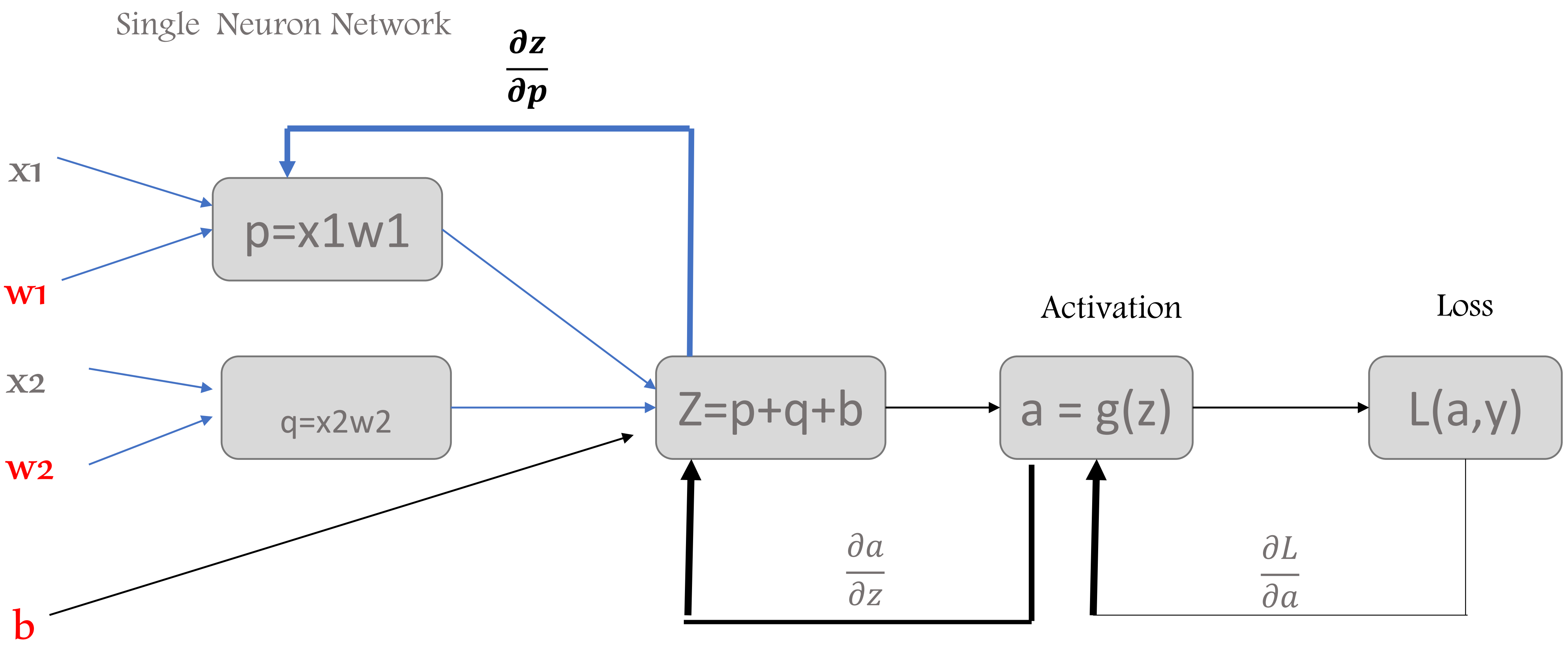
lets consider a network where each training point has two features x1 and x2.

For ex.

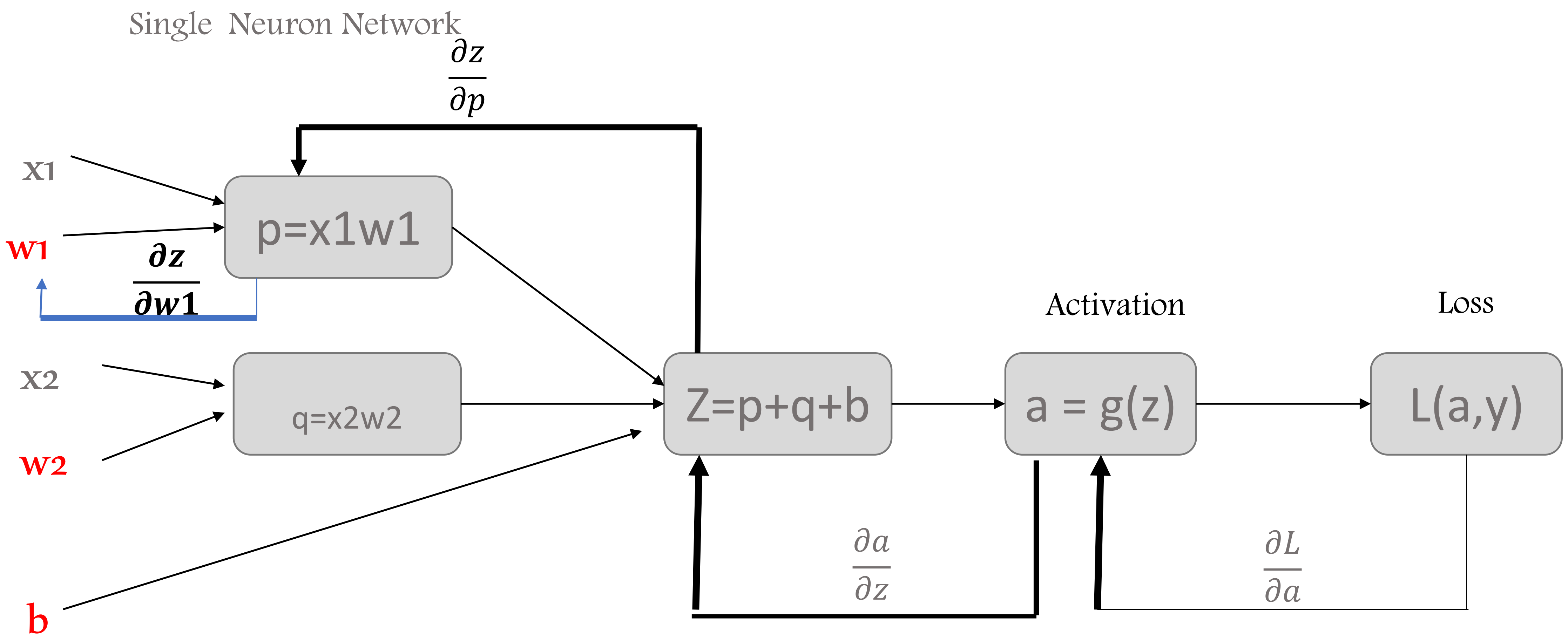


$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial a} * \frac{\partial a}{\partial z} * \frac{\partial z}{\partial b}$$

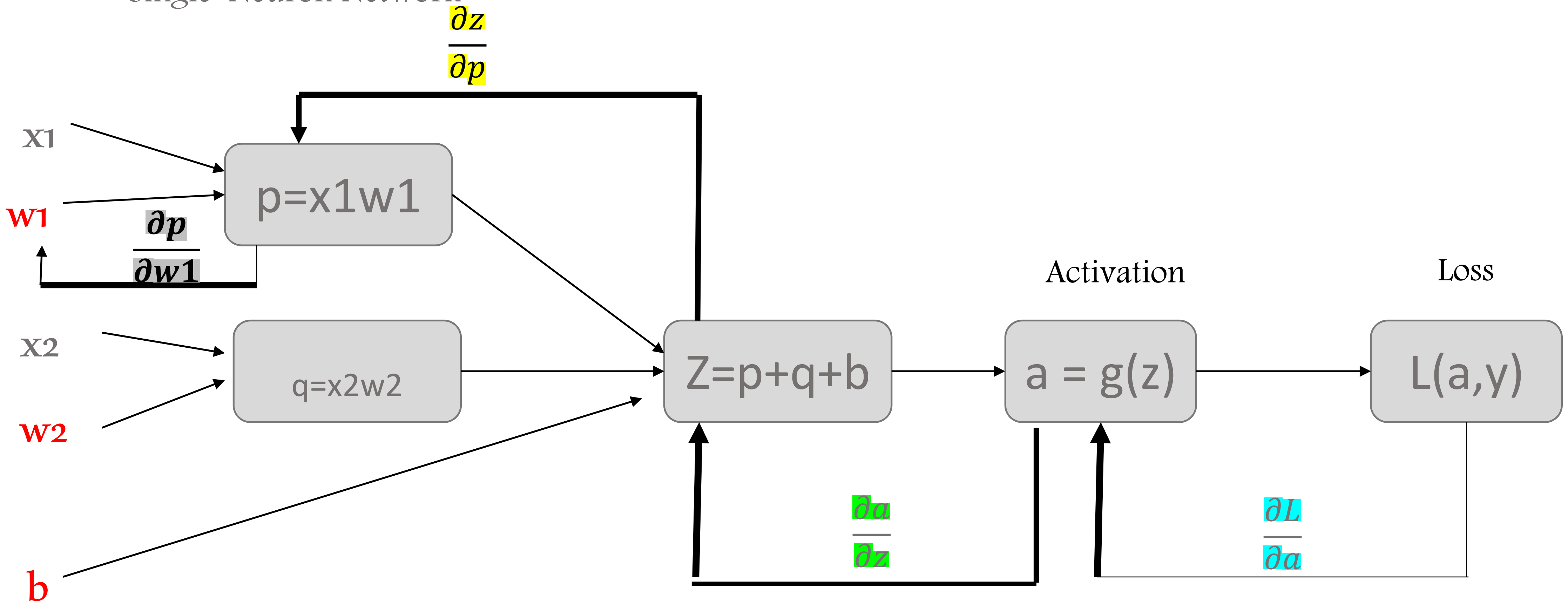
Now lets find $$\frac{\partial L}{\partial w_1}$$



Single Neuron Network
$\frac{\partial z}{\partial p}$
x1
w1
x2
w2
b
p=x1w1
q=x2w2
Z=p+q+b
Activation
a = g(z)
Loss
L(a,y)
$\frac{\partial a}{\partial z}$
$\frac{\partial L}{\partial a}$


Single Neuron Network
$\frac{\partial z}{\partial p}$
x1
w1
p=x1w1
$\frac{\partial z}{\partial w1}$
x2
w2
q=x2w2
b
Z=p+q+b
Activation
a = g(z)
Loss
L(a,y)
$\frac{\partial a}{\partial z}$
$\frac{\partial L}{\partial a}$

Single Neuron Network



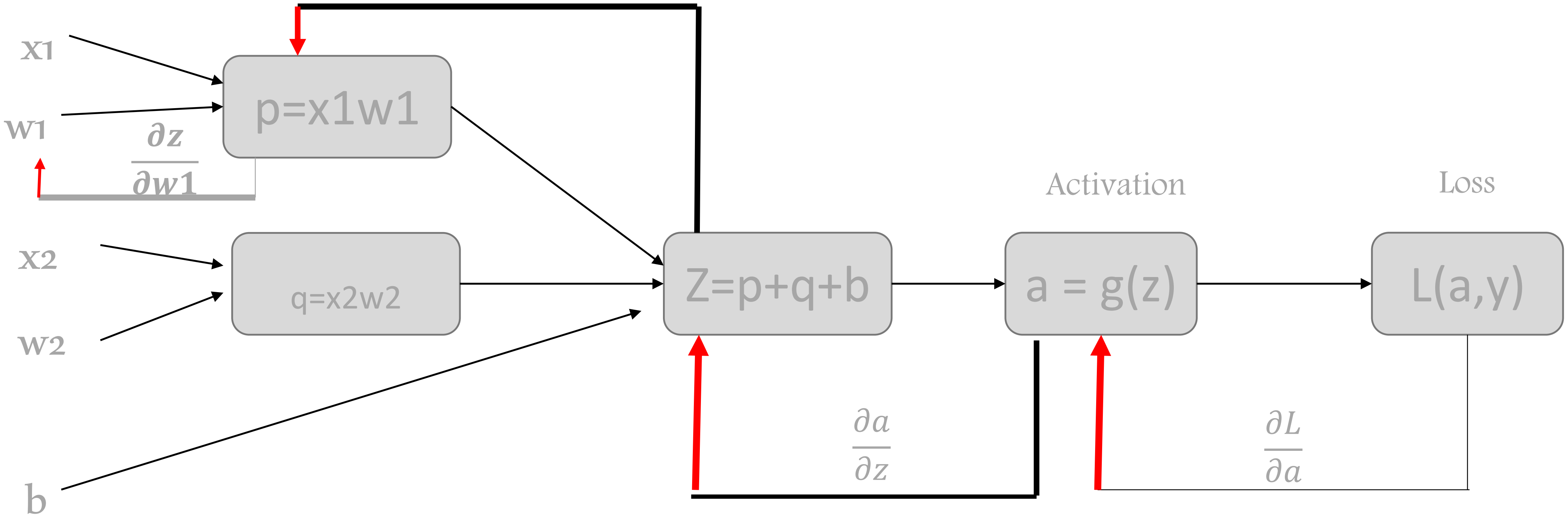
$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial a} * \frac{\partial a}{\partial z} * \frac{\partial z}{\partial p} * \frac{\partial p}{\partial w_1}$$

Similarly

$$\frac{\boldsymbol{\partial L}}{\boldsymbol{\partial w_2}} = \frac{\partial L}{\partial a} * \frac{\partial a}{\partial z} * \frac{\partial z}{\partial p} * \frac{\partial p}{\partial w_2}$$

So, that is how we find $\left[\frac{\partial L}{\partial w_1}, \quad \frac{\partial L}{\partial w_2}, \quad \frac{\partial L}{\partial b}\right]$ Just, look at the arrows for finding each of these terms.

Since, they start from output and move back towards input. It is called back propagation method.



Once we get $\left[\frac{\partial L}{\partial w_1}, \quad \frac{\partial L}{\partial w_2}, \quad \frac{\partial L}{\partial b}\right]$, gradient is known. This can be used by an

optimization method to find minima or maxima