AI VIET NAM
@aivietnam.edu.vn

Multi-layer Perception

Model Initialization

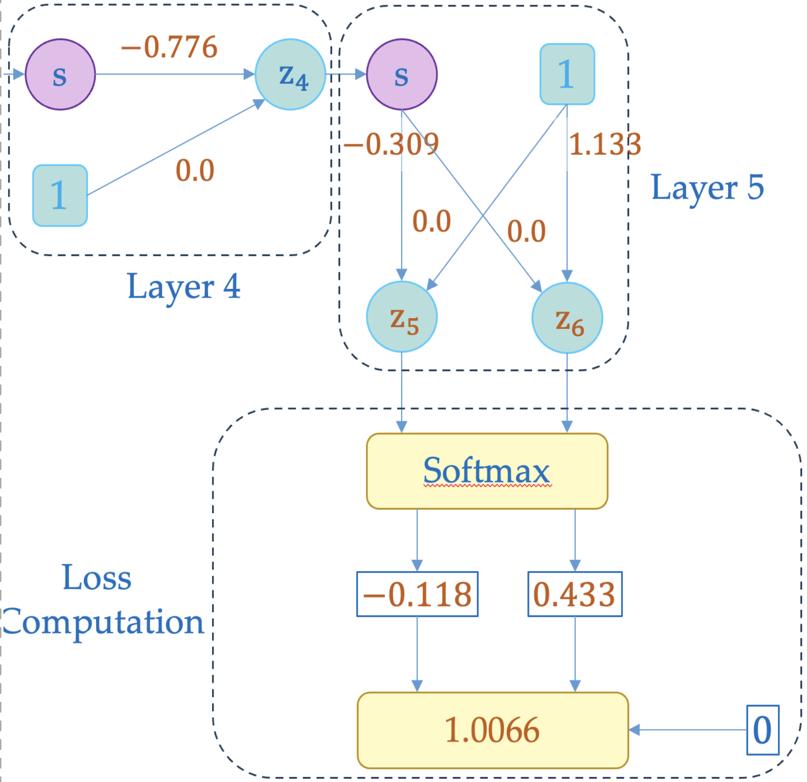
Quang-Vinh Dinh
Ph.D. in Computer Science

Year 2024

code&data

Objectives

Case Studies



Xavier Glorot Init.

$$W_i \sim U\left(-\frac{4\sqrt{3}}{\sqrt{n}}, \frac{4\sqrt{3}}{\sqrt{n}}\right)$$

$$W_i \sim N\left(0, \frac{1}{n}\right)$$

Kaiming He Init.

$$W_i \sim U\left(-\frac{\sqrt{6}}{\sqrt{n}}, \frac{\sqrt{6}}{\sqrt{n}}\right)$$

$$W_i \sim N\left(0, \frac{2}{n}\right)$$

Outline

SECTION 1

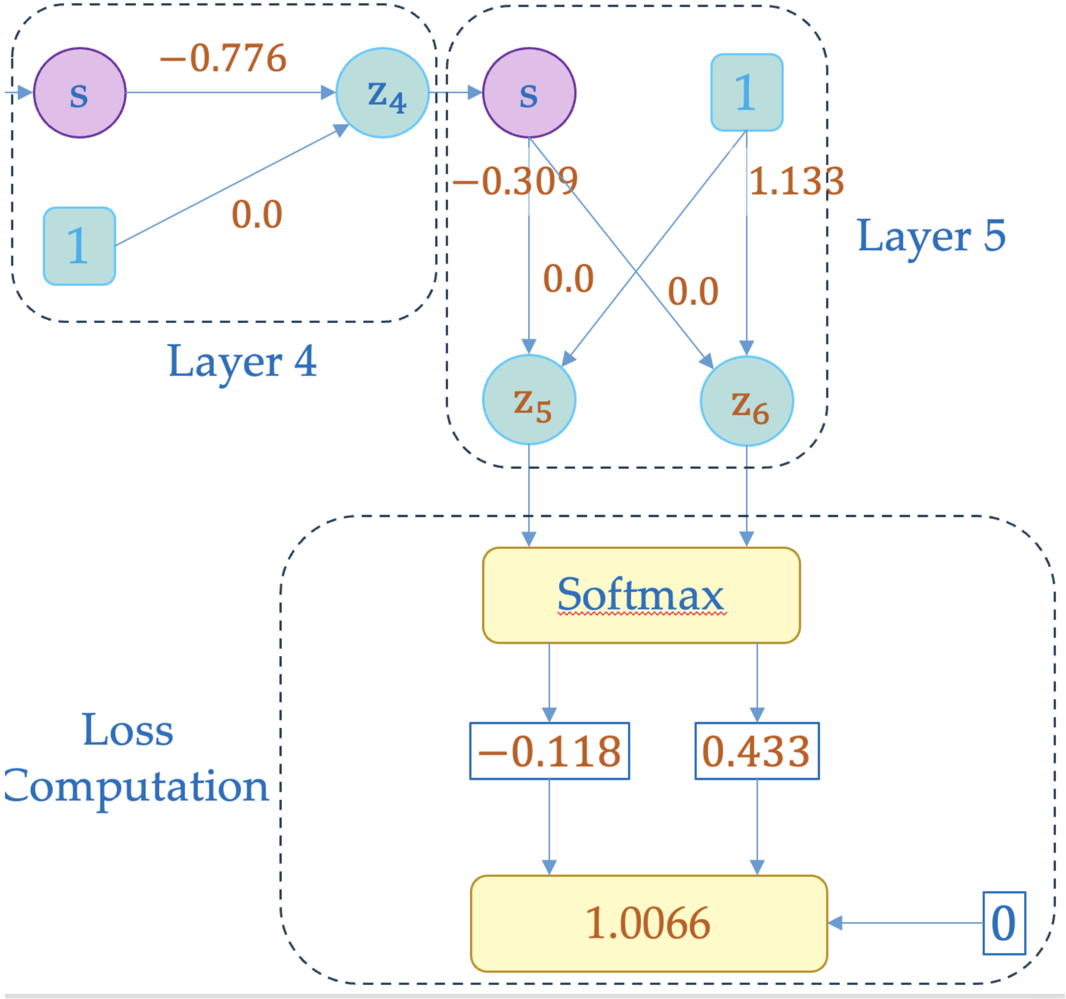
Case Studies

SECTION 2

Xavier Glorot Init.

SECTION 3

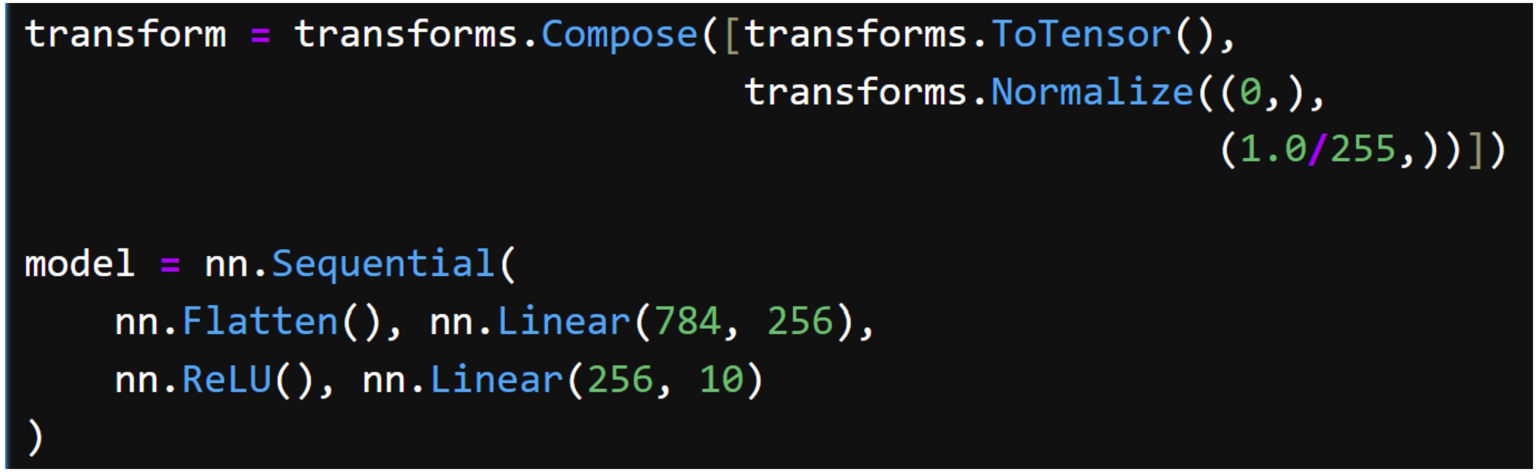
Kaiming He Init.



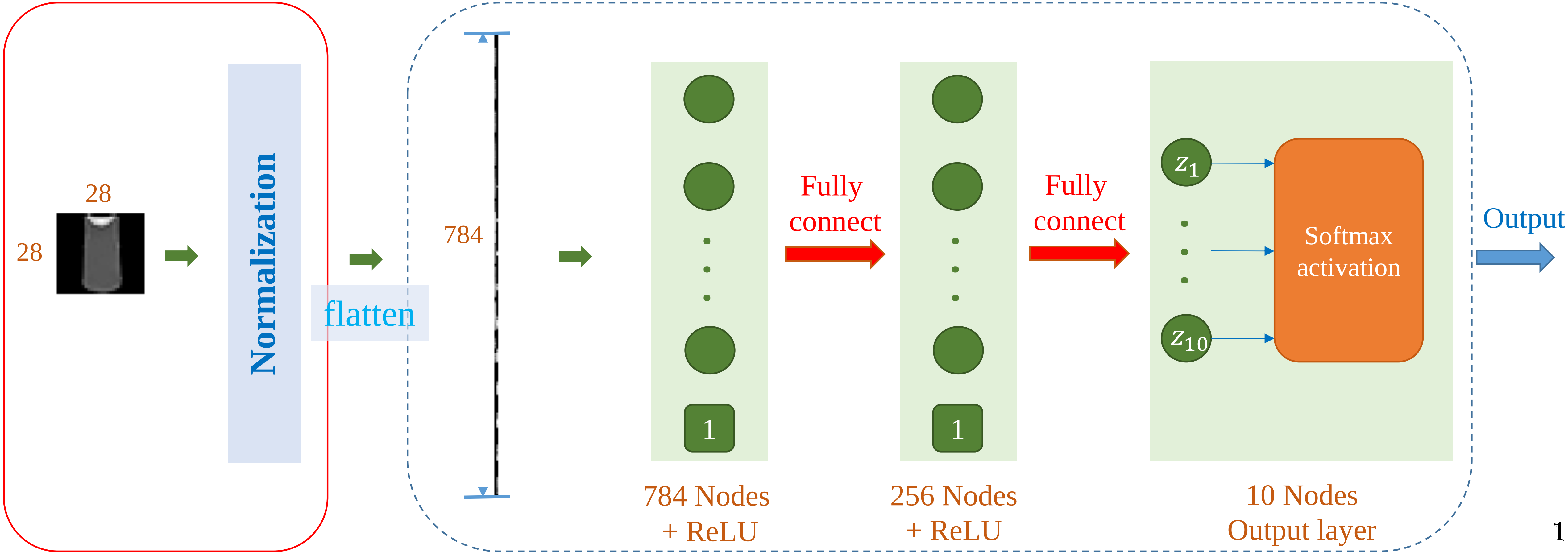
$$X \in [0, 255]$$

$$\text{Normalize}(\mathit{mean}, \text{std})$$

$$\text{Image} = \frac{\text{Image} - \mathit{mean}}{\text{std}}$$



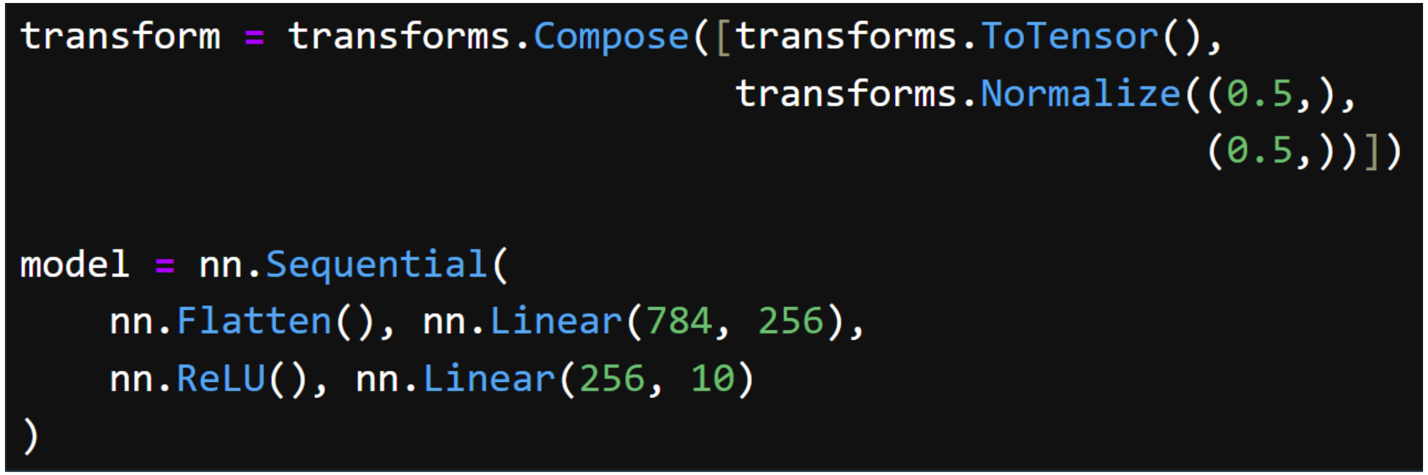
```
transform = transforms.Compose([transforms.ToTensor(),
                                transforms.Normalize((0,),
                                                     (1.0/255,))])


model = nn.Sequential(
    nn.Flatten(), nn.Linear(784, 256),
    nn.ReLU(), nn.Linear(256, 10)
)
```



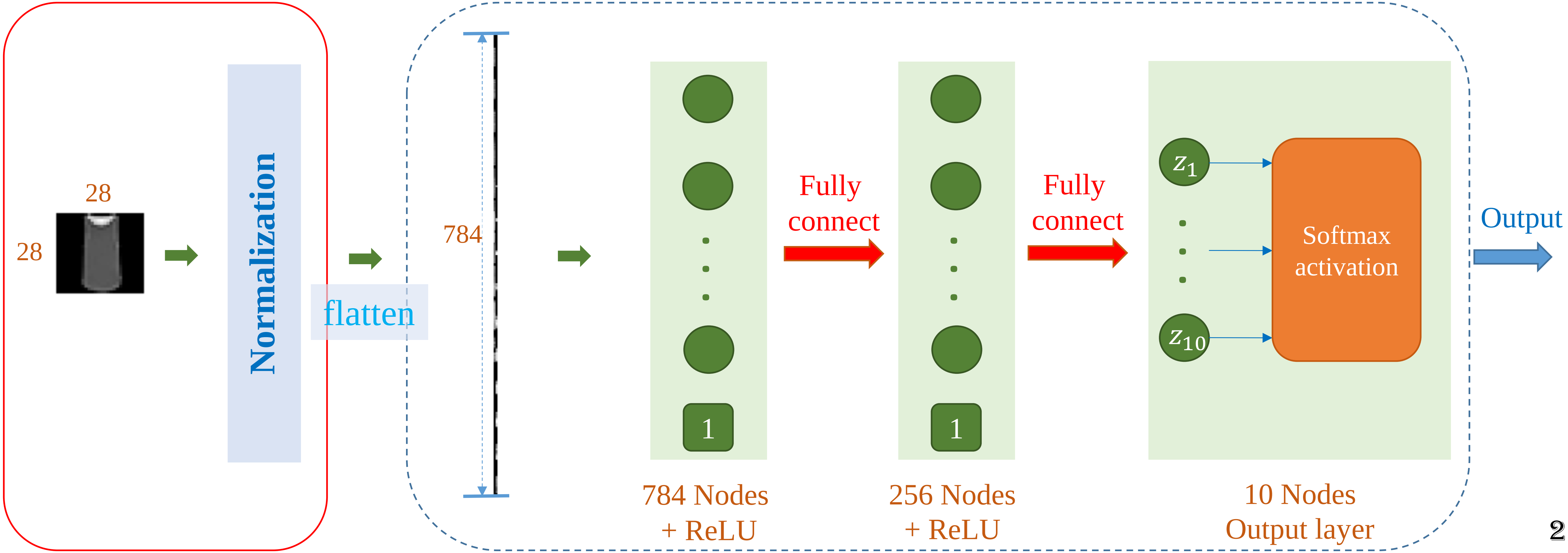
$$X \in [-1, 1]$$

$$\text{Normalize}(mean, \text{std})$$

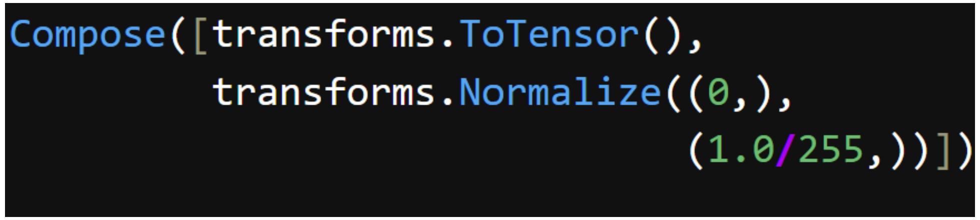
$$\text{Image} = \frac{\text{Image} - mean}{\text{std}}$$



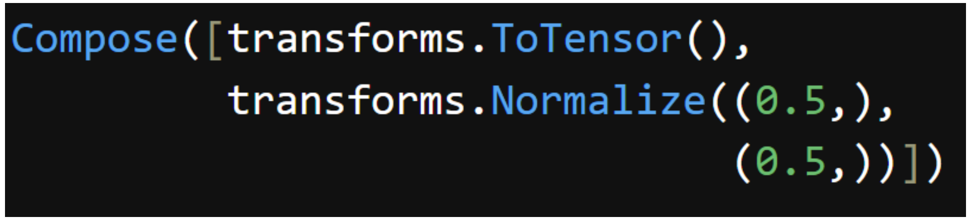
```
transform = transforms.Compose([transforms.ToTensor(),
                                transforms.Normalize((0.5,),
                                                     (0.5,))])

model = nn.Sequential(
    nn.Flatten(), nn.Linear(784, 256),
    nn.ReLU(), nn.Linear(256, 10)
)
```



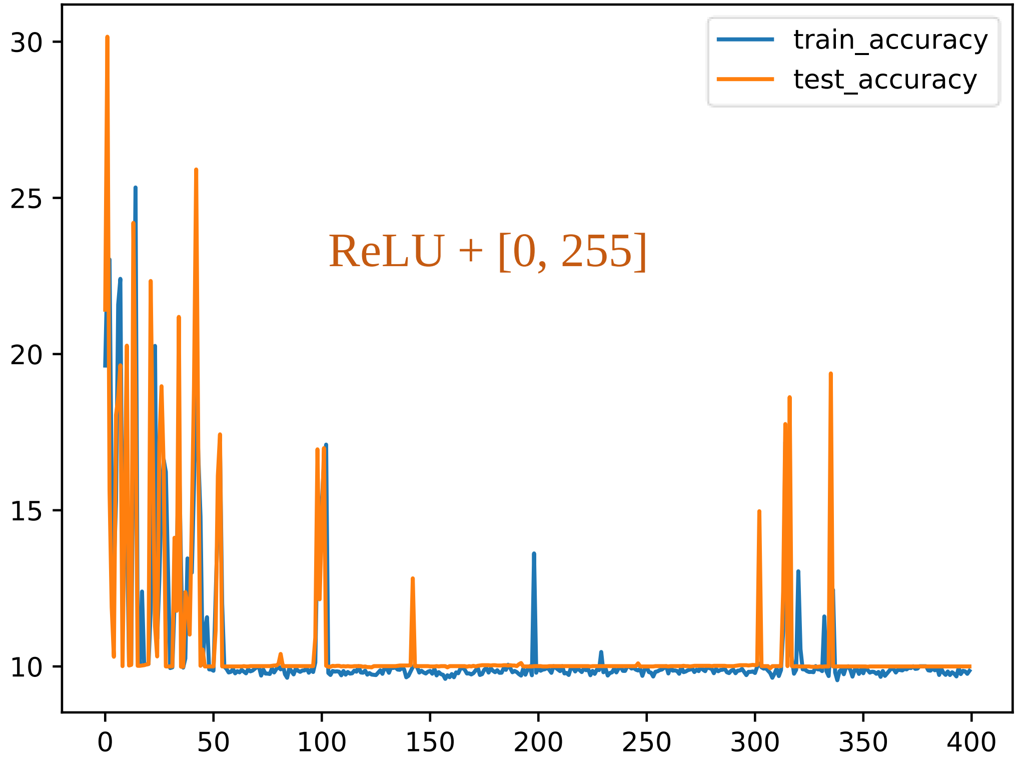
Experimental Results



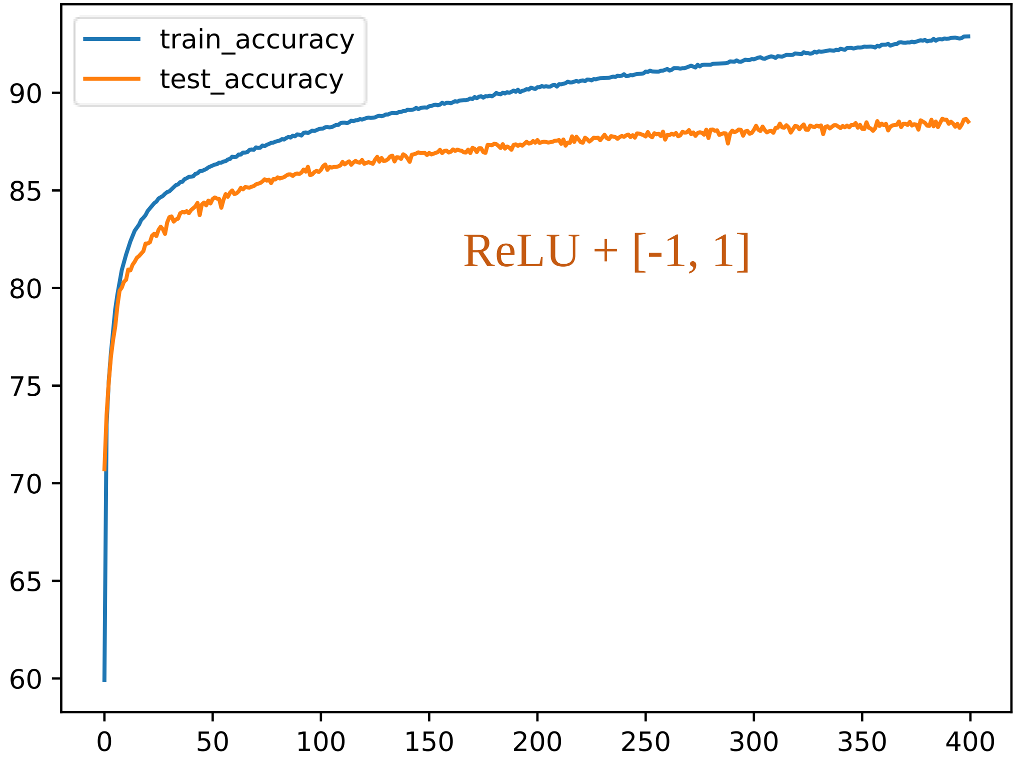
```
Compose([transforms.ToTensor(),
         transforms.Normalize((0,),
                              (1.0/255,))])
```



```
Compose([transforms.ToTensor(),
         transforms.Normalize((0.5,),
                              (0.5,))])
```



ReLU + [0, 255]

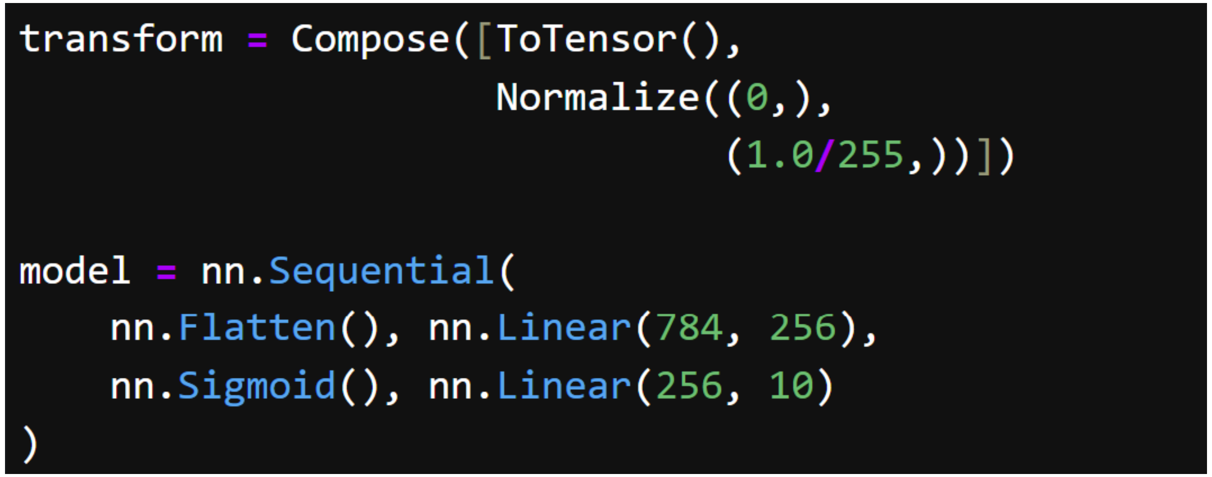


ReLU + [-1, 1]

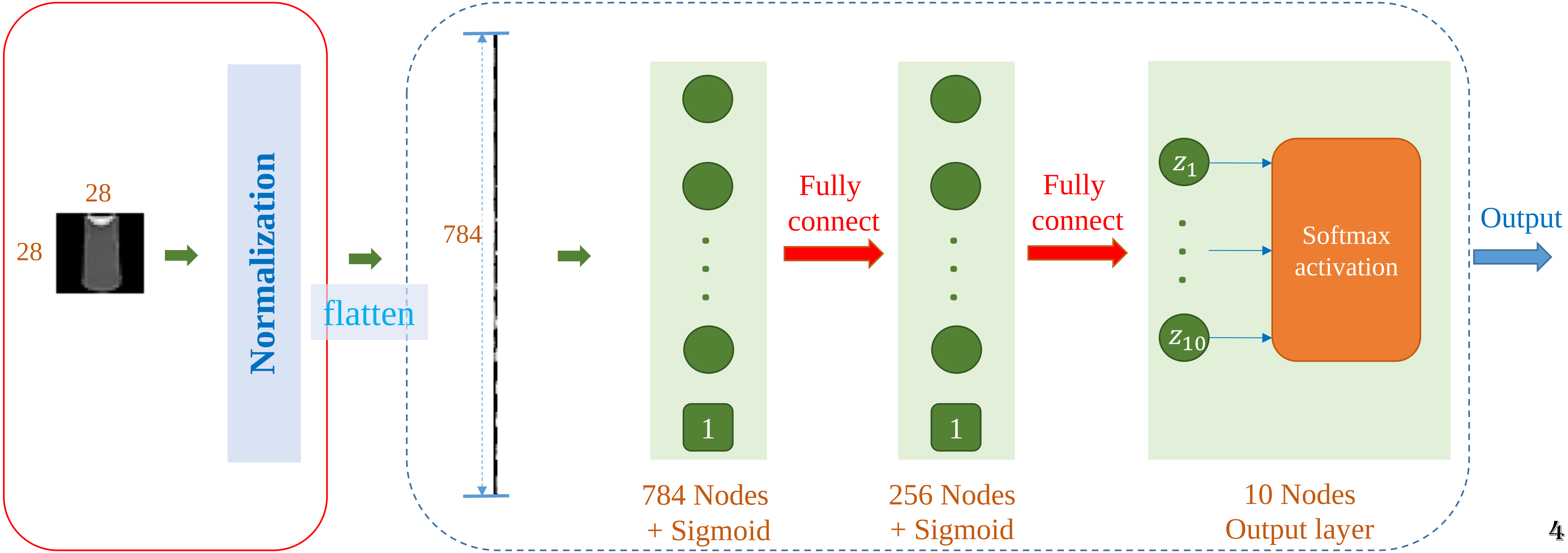
$$X \in [0, 255]$$

$$\text{Normalize}(mean, \text{std})$$

$$\text{Image} = \frac{\text{Image} - mean}{\text{std}}$$



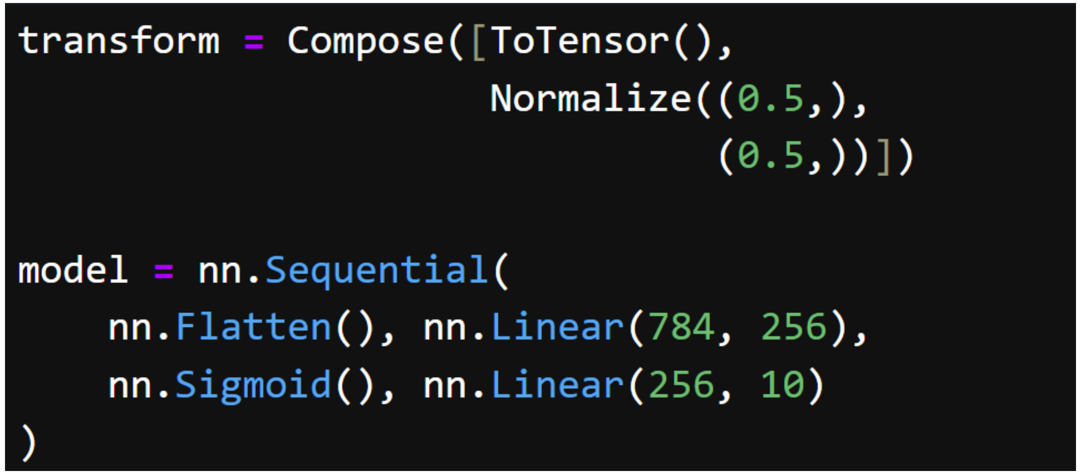
```
transform = Compose([ToTensor(),
                     Normalize((0,),
                               (1.0/255,))])

model = nn.Sequential(
    nn.Flatten(), nn.Linear(784, 256),
    nn.Sigmoid(), nn.Linear(256, 10)
)
```



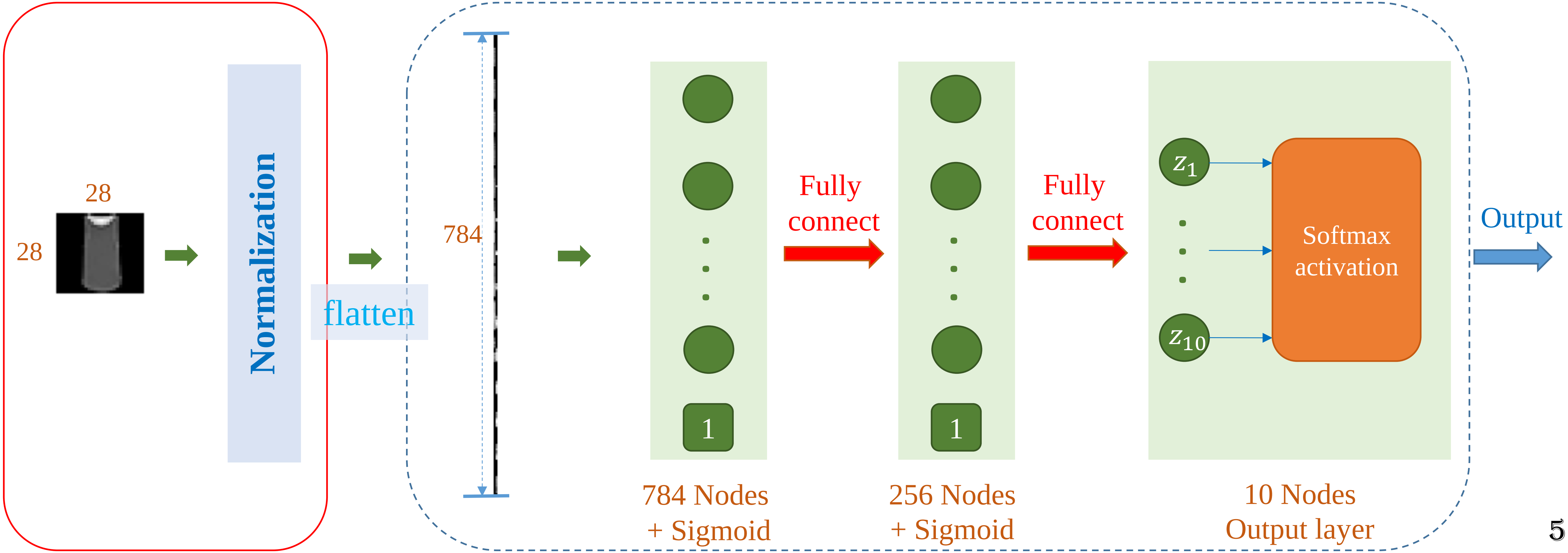
$$X \in [-1, 1]$$

$$\text{Normalize}(\mathit{mean}, \text{std})$$

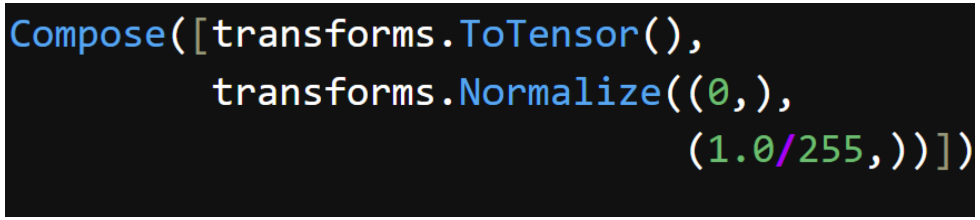
$$\text{Image} = \frac{\text{Image} - \mathit{mean}}{\text{std}}$$



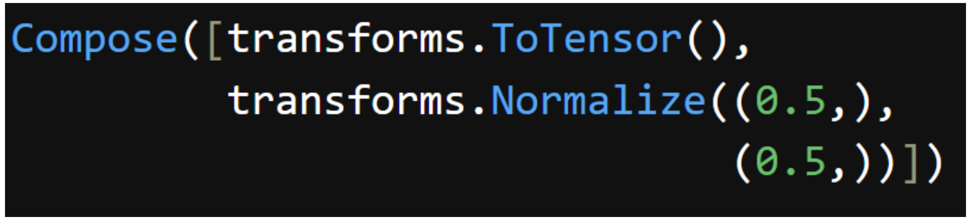
```
transform = Compose([ToTensor(),
                     Normalize((0.5,),
                               (0.5,))])

model = nn.Sequential(
    nn.Flatten(), nn.Linear(784, 256),
    nn.Sigmoid(), nn.Linear(256, 10)
)
```



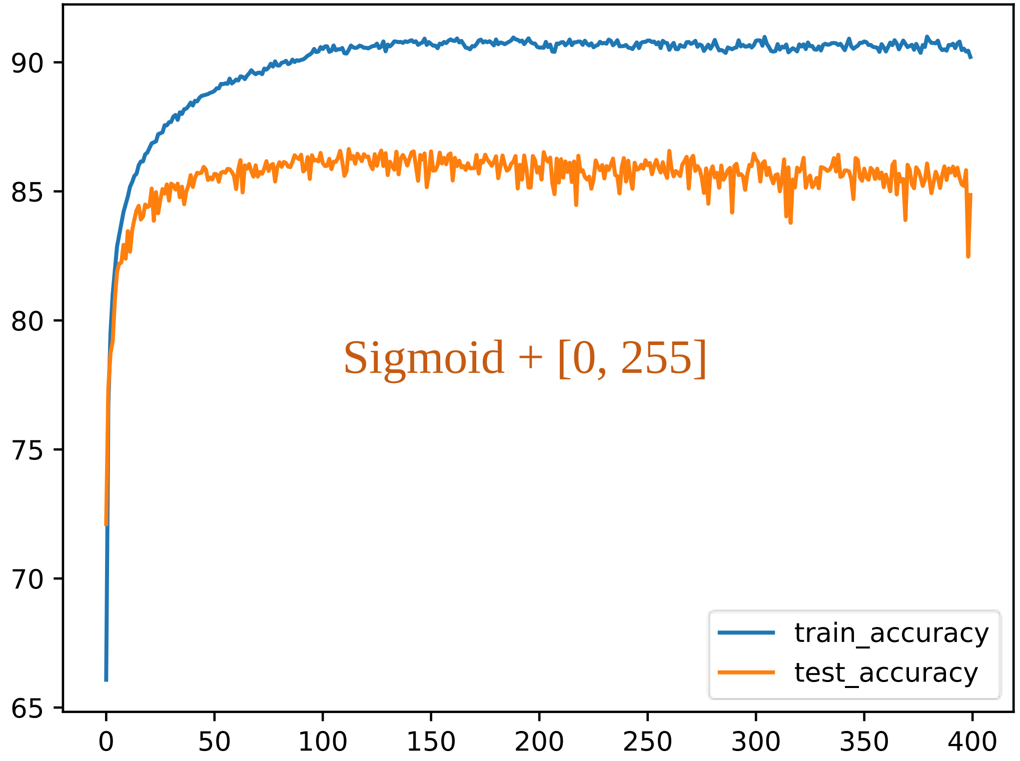
Experimental Results

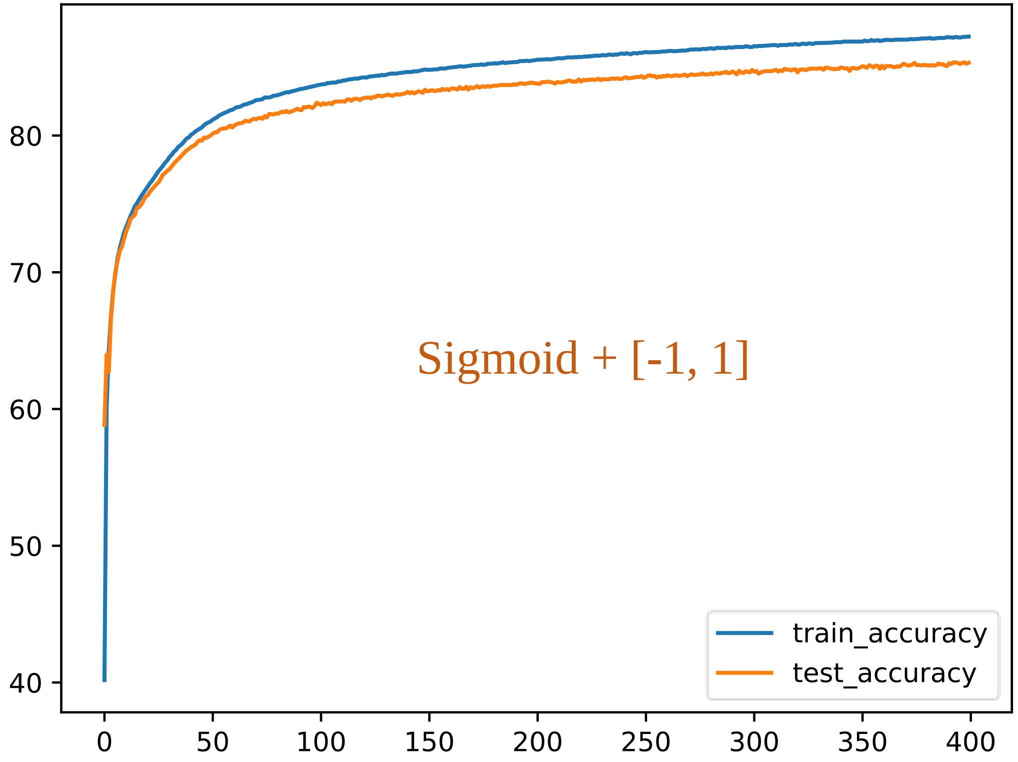


```
Compose([transforms.ToTensor(),
         transforms.Normalize((0,),
                              (1.0/255,))])
```



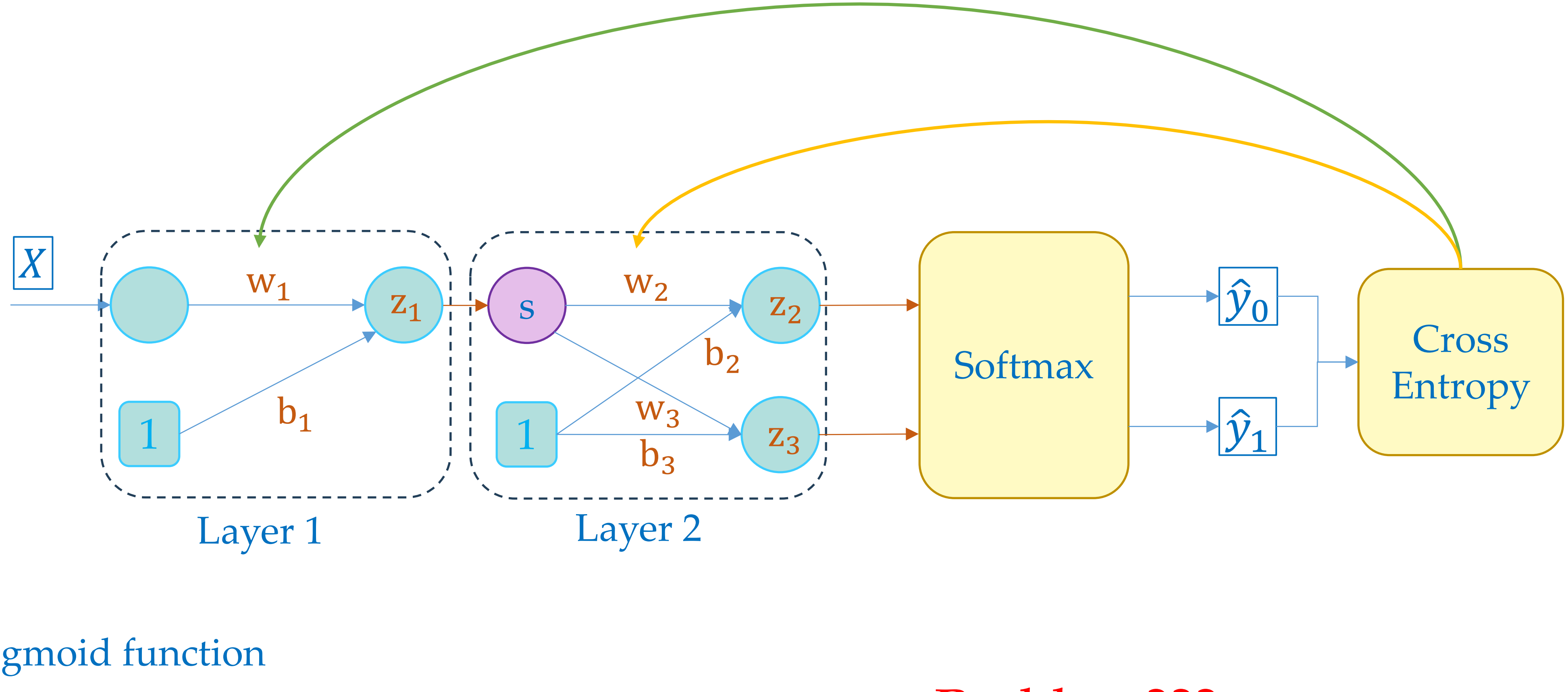
```
Compose([transforms.ToTensor(),
         transforms.Normalize((0.5,),
                              (0.5,))])
```





Gradient Vanishing

Large weight initialization



Problem???

Activation Functions

$$\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$$

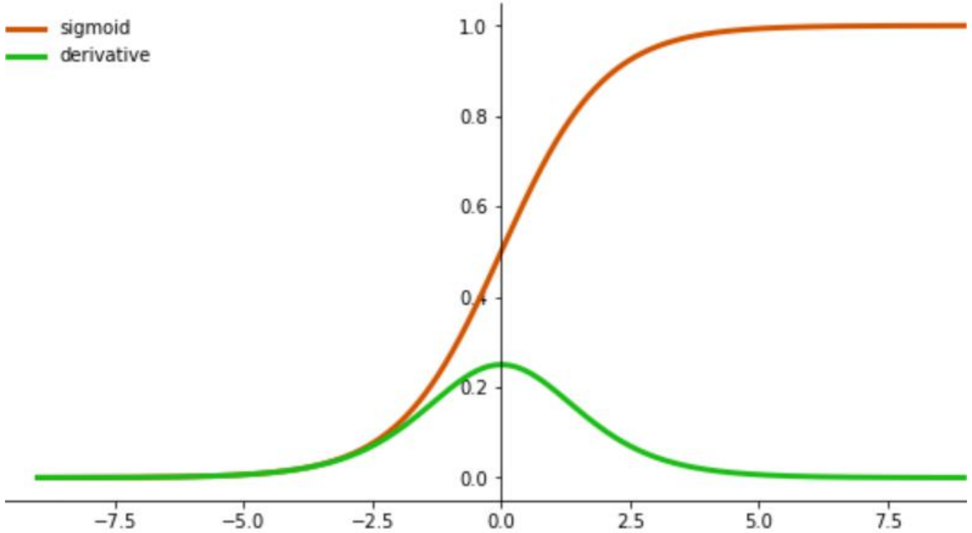
data =

1	5	-4	3	-2

data_a = sigmoid(data)

data_a =

0.731	0.993	0.017	0.95	0.119



$$\text{sigmoid}'(x) = \left(\frac{1}{1+e^{-x}}\right)' = \frac{-1}{(1+e^{-x})^2}(-e^{-x})$$

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

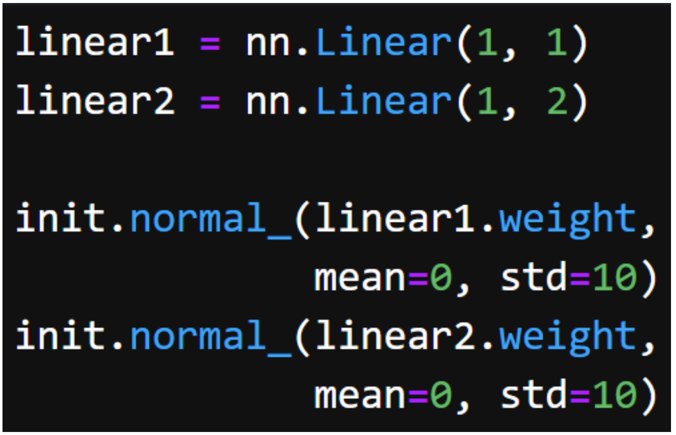
$$= \frac{1}{1+e^{-x}} - \frac{1}{(1+e^{-x})^2}$$

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

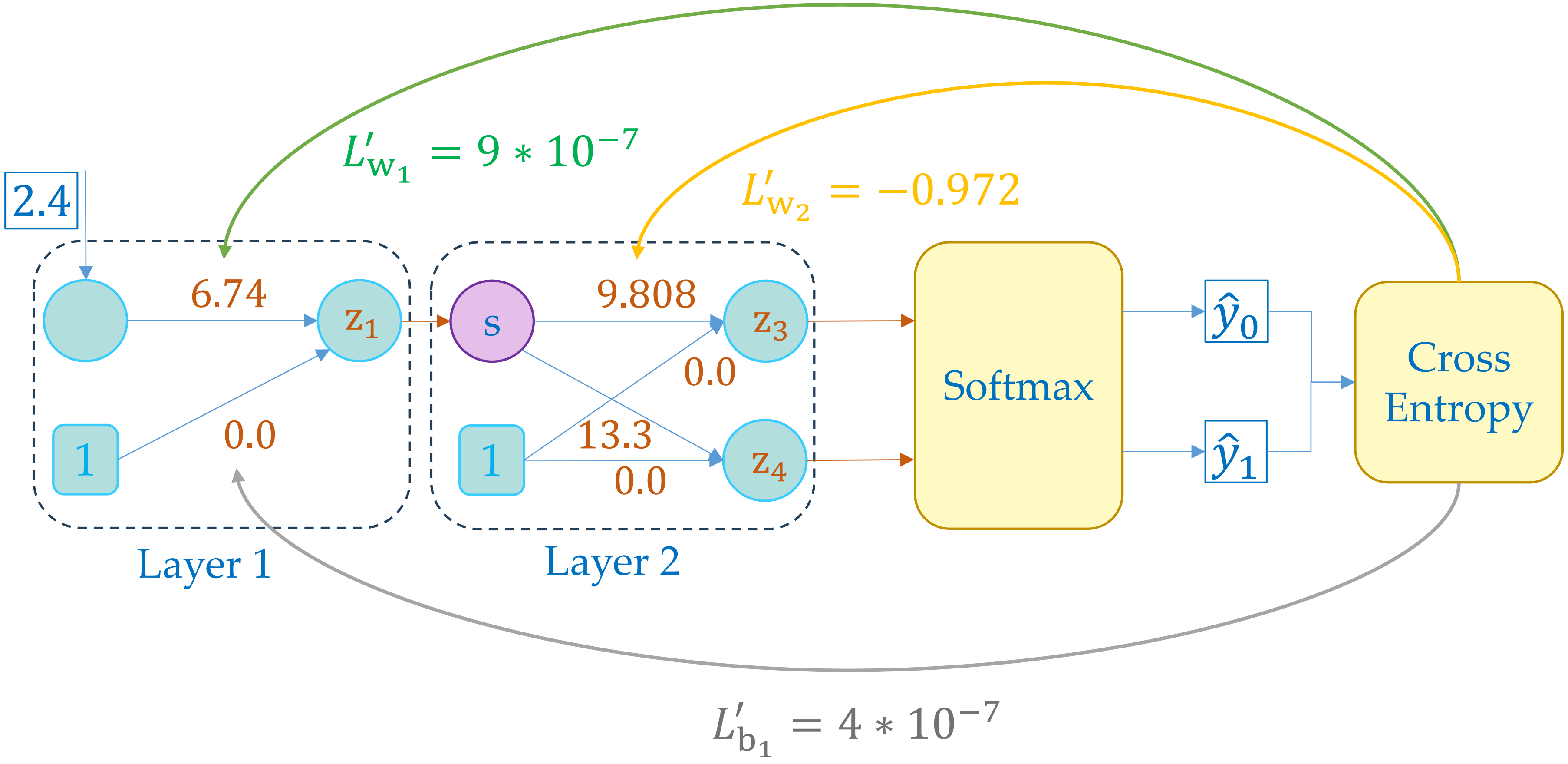
$$= \text{sigmoid}(x)\,(1-\text{sigmoid}(x))$$

Gradient Vanishing

Large weight initialization



```
linear1 = nn.Linear(1, 1)
linear2 = nn.Linear(1, 2)

init.normal_(linear1.weight,
             mean=0, std=10)
init.normal_(linear2.weight,
             mean=0, std=10)
```



with $\eta = 0.01$

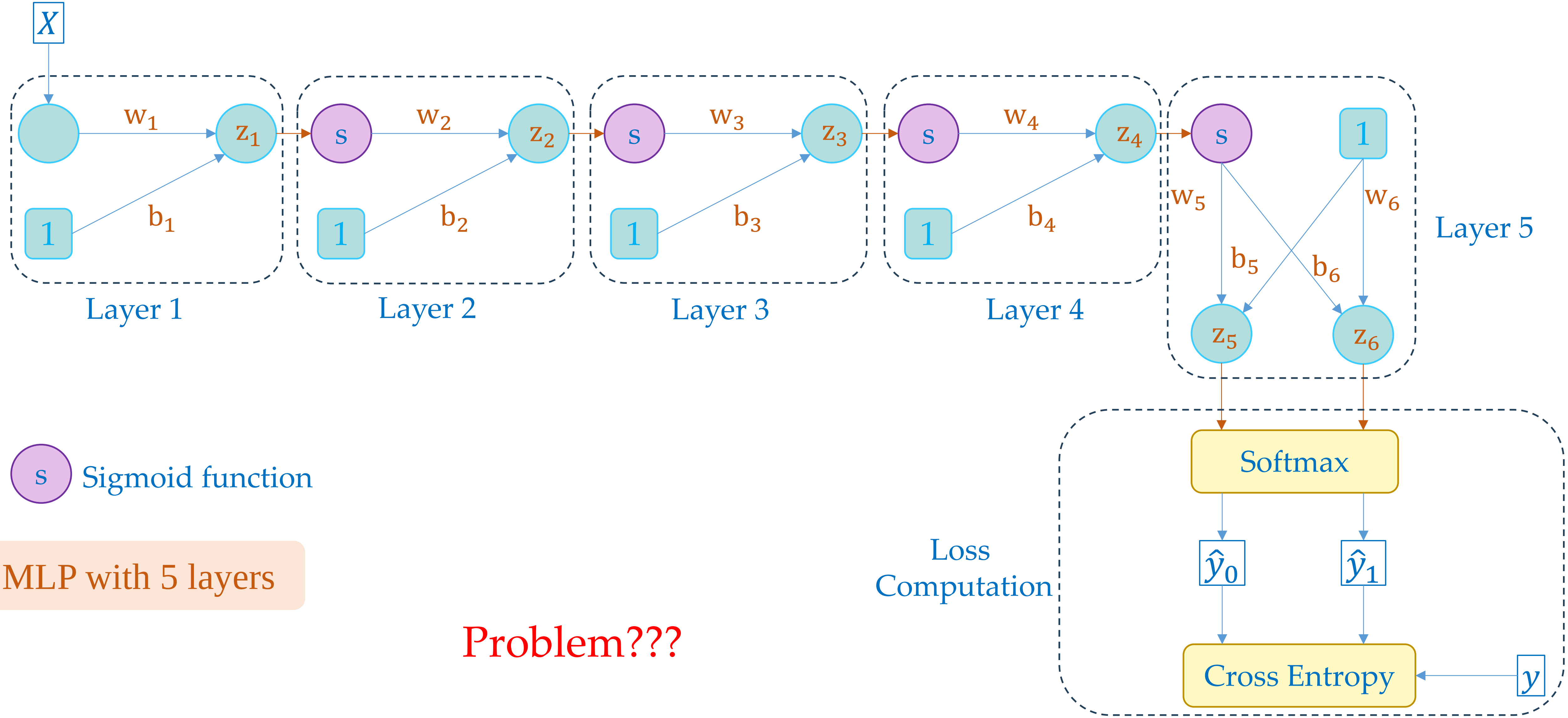
$$\eta L'_{w_1} = 9 * 10^{-9}$$

$$\eta L'_{b_1} = 4 * 10^{-9}$$

Derivative values are too small

Gradient Vanishing

Using appropriate weight initialization



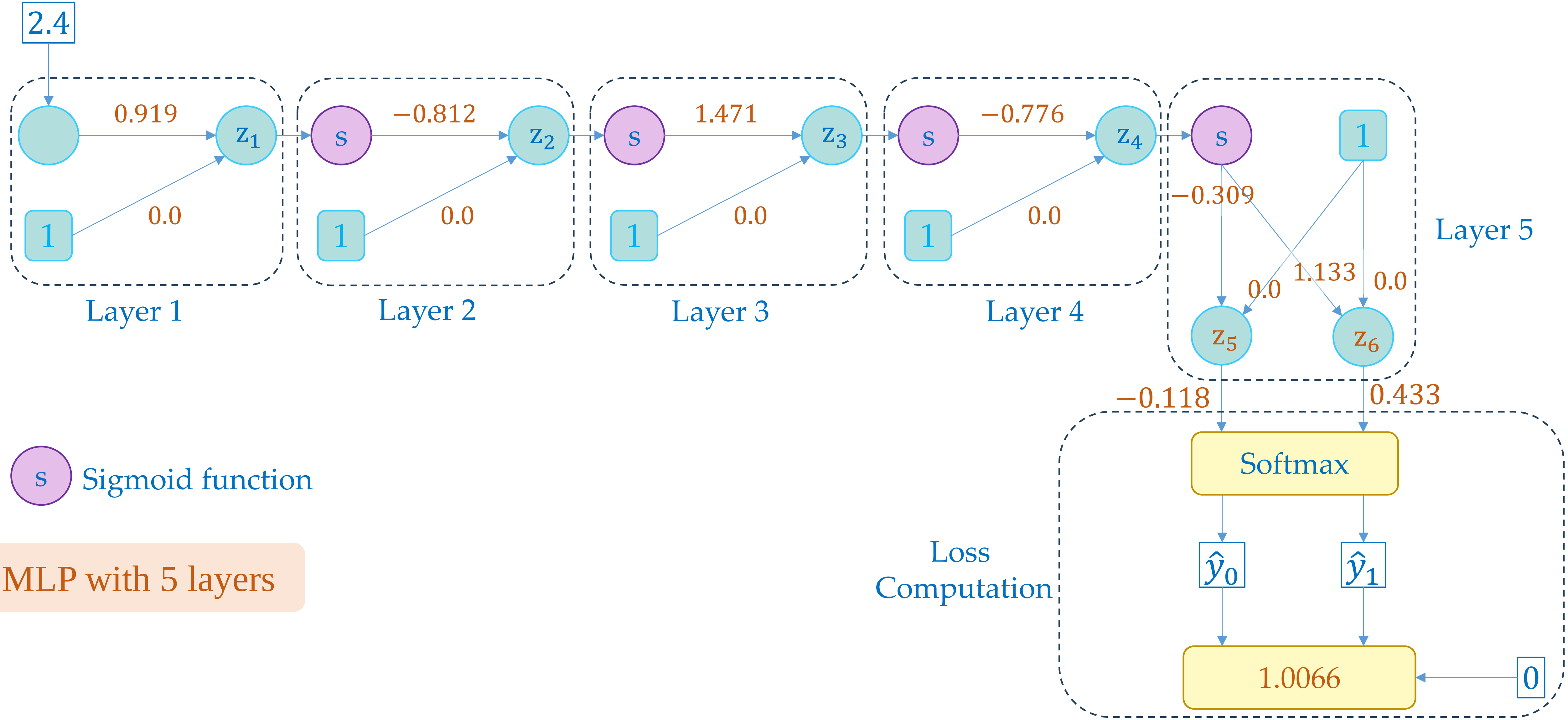
MLP with 5 layers

Problem???

AI VIET NAM
@aivietnam.edu.vn

Gradient Vanishing

Using appropriate weight initialization



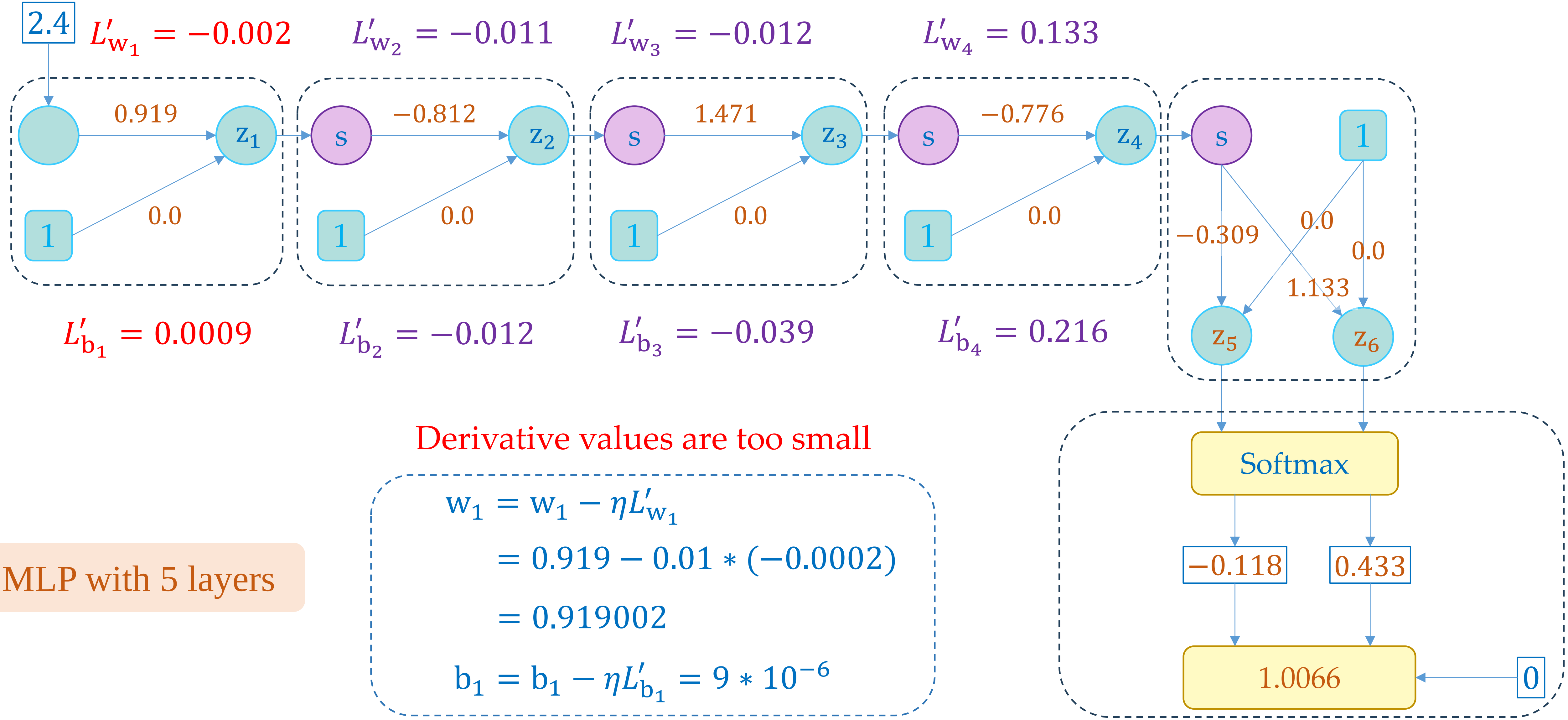
s Sigmoid function

MLP with 5 layers

AI VIET NAM
@aivietnam.edu.vn

Gradient Vanishing

Using appropriate weight initialization



Derivative values are too small

$$w_1 = w_1 - \eta L'_{w_1}$$

$$= 0.919 - 0.01 * (-0.0002)$$

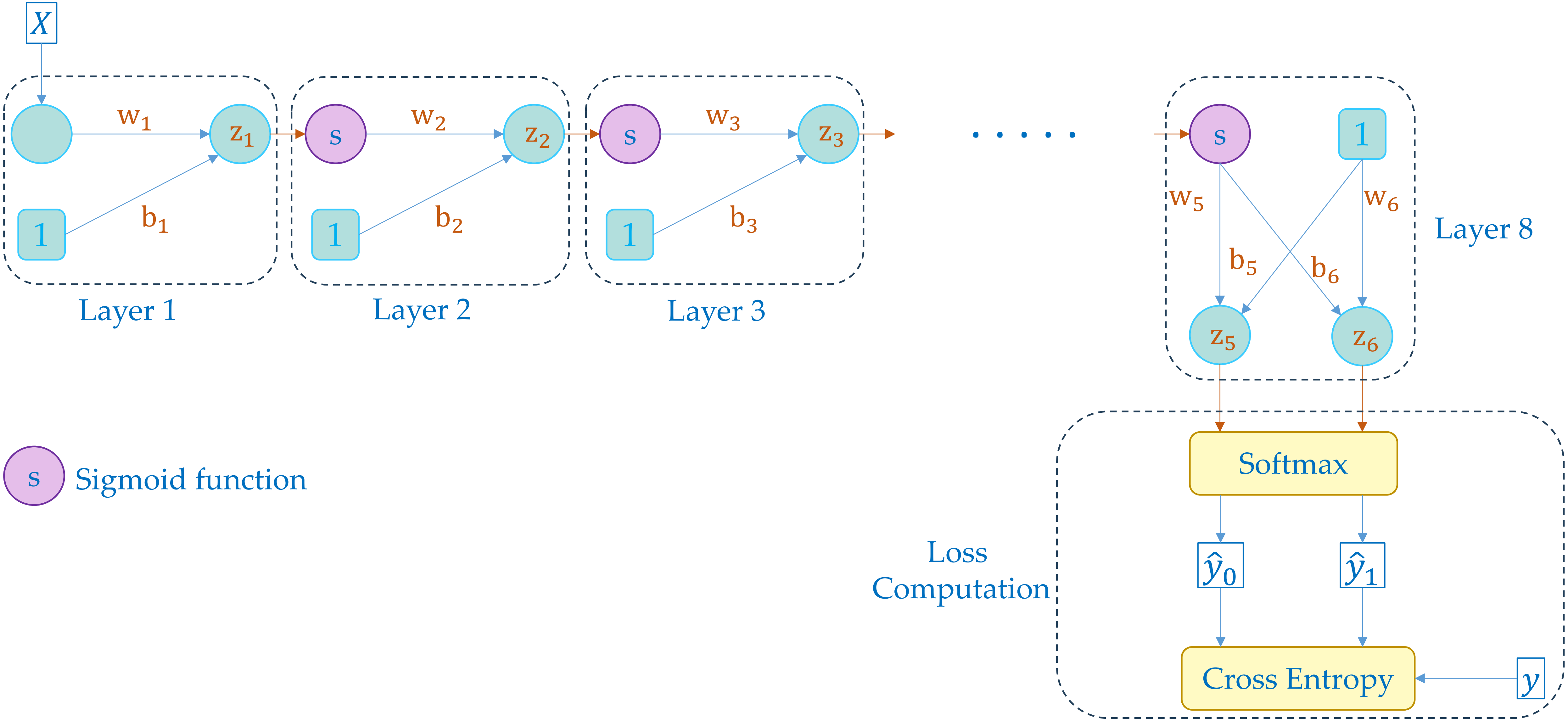
$$= 0.919002$$

$$b_1 = b_1 - \eta L'_{b_1} = 9 * 10^{-6}$$

MLP with 5 layers

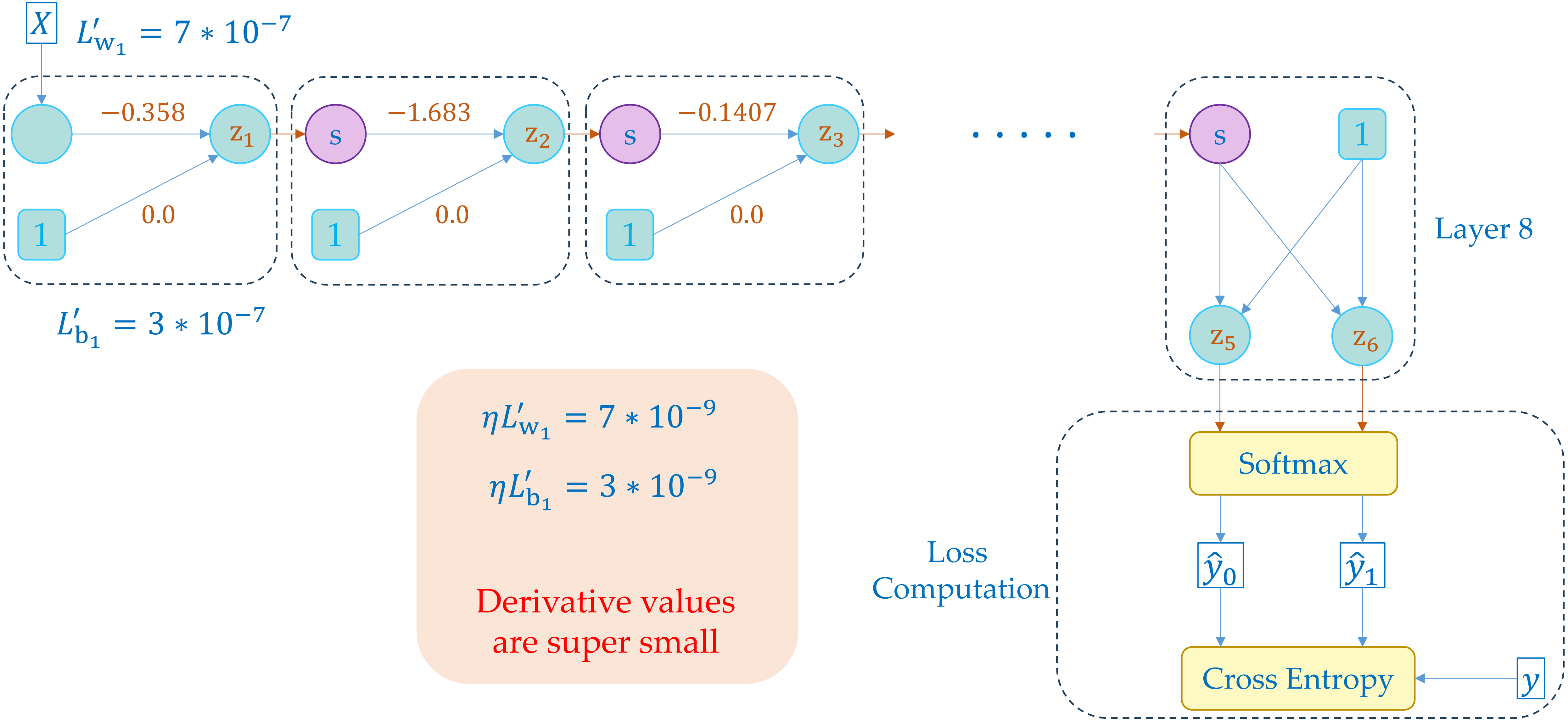
Gradient Vanishing

MLP with 8 layers



Gradient Vanishing

MLP with 8 layers



$$\eta L'_{w_1} = 7 * 10^{-9}$$

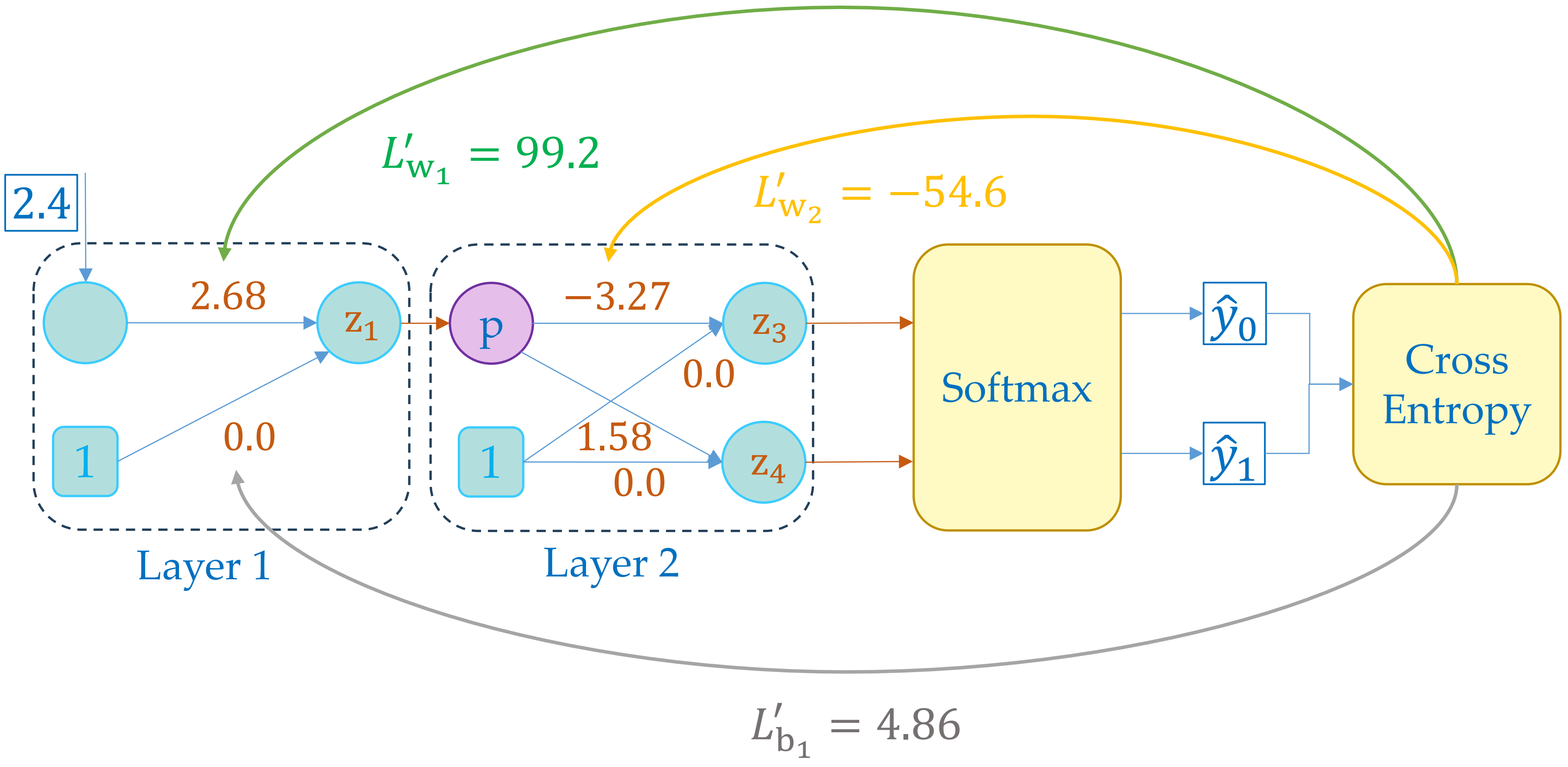
$$\eta L'_{b_1} = 3 * 10^{-9}$$

Derivative values
are super small

Gradient Explosion

Large weight initialization and large learning rate

p PReLU function



with $\eta = 1.0$

$\eta L'_{w_1} = 99.2$

$\eta L'_{b_1} = 4.86$

Derivative values are too large

Activation Functions

❖ PReLU function

$$\text{PReLU}(x) = \begin{cases} \alpha x & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$$

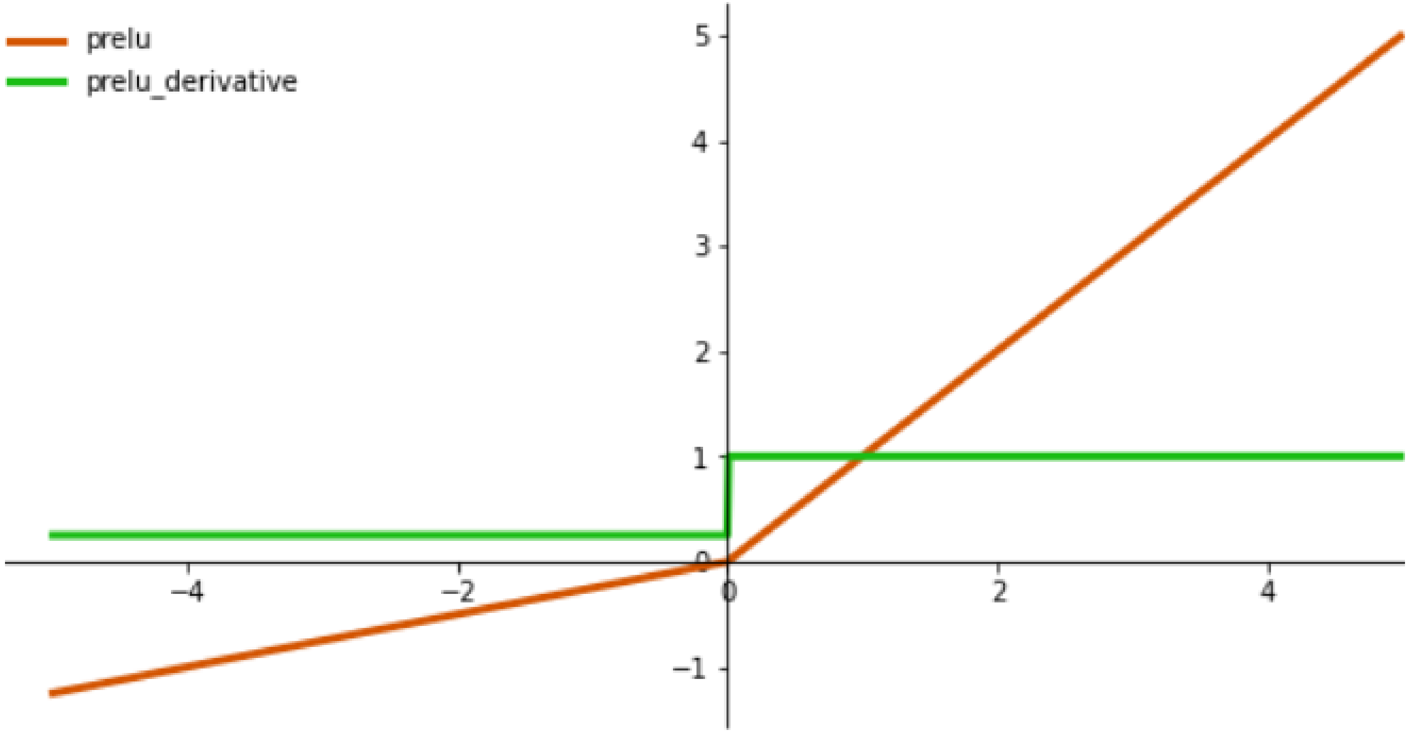
data =

1	5	-4	3	-2

data_a = PRELU(data)

data_a =

1	5	-0.4	3	-0.2



$$\text{PReLU}'(x) = \begin{cases} \alpha & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}$$

Outline

SECTION 1

Case Studies

SECTION 2

Xavier Glorot Init.

SECTION 3

Kaiming He Init.

$$W_i \sim U\left(-\frac{4\sqrt{3}}{\sqrt{n}}, \frac{4\sqrt{3}}{\sqrt{n}}\right)$$

$$W_i \sim N\left(0, \frac{1}{n}\right)$$

Mean

Data

$$X = \{X_1, \dots, X_N\}$$

Formula

$$E(X) = \sum_{i=1}^{N} X_i P_X(X_i)$$

Given the data

$$X = \{2, 8, 5, 4, 1, 4\}$$

$$N = 6$$

$$P_X(X = 2) = \frac{1}{6} \qquad P_X(X = 4) = \frac{2}{6}$$

$$P_X(X = 8) = \frac{1}{6} \qquad P_X(X = 1) = \frac{1}{6}$$

$$P_X(X = 5) = \frac{1}{6}$$

$$E(X) = 2 \times \frac{1}{6} + 8 \times \frac{1}{6} + 5 \times \frac{1}{6} + 4 \times \frac{2}{6} + 1 \times \frac{1}{6}$$

$$= \frac{2}{6} + \frac{8}{6} + \frac{5}{6} + \frac{8}{6} + \frac{1}{6} = 4$$

Mean

Data

$$X = \{X_1, \dots, X_N\}$$

Formula

$$E(X) = \sum_{i=1}^{N} X_i P_X(X_i)$$

$$E(XY) = \sum_{i=1}^{N}\sum_{j=1}^{N} X_i Y_j P(X_i, Y_j)$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N} X_i Y_j P(X_i) P(Y_j)$$

$$= \sum_{i=1}^{N} X_i P(X_i) \sum_{j=1}^{N} Y_j P(Y_j)$$

$$= E(X)E(Y)$$

Variance

Formula

mean

$$E(X) = \sum_{i=1}^{N} X_i P_X(X_i)$$

variance

$$var(X) = E\left(\left(X - E(X)\right)^2\right)$$

$$= \sum_{i=1}^{N} \left(X_i - E(X)\right)^2 P_X(X_i)$$

Standard deviation

$$\sigma = \sqrt{var(X)}$$

Example: $X = \{5, 3\ 6, 7, 4\}$

$$E(X) = 5 \times \frac{1}{5} + 3 \times \frac{1}{5} + 6 \times \frac{1}{5} + 7 \times \frac{1}{5} + 4 \times \frac{1}{5}$$

$$= 5$$

$$var(X) = \frac{1}{5}[(5-5)^2 + (3-5)^2 + (6-5)^2 + (7-5)^2 + (4-5)^2]$$

$$= \frac{1}{5}(0+4+1+4+1)=2$$

$$\sigma = \sqrt{var(X)} = 1.41$$

Variance

Formula

mean

$$E(X) = \sum_{i=1}^{N} X_i P_X(X_i)$$

variance

$$\begin{aligned} var(X) &= E\left(\left(X - E(X)\right)^2\right) \\ &= \sum_{i=1}^{N} \left(X_i - E(X)\right)^2 P_X(X_i) \end{aligned}$$

Standard deviation

$$\sigma = \sqrt{var(X)}$$

$$\begin{aligned} var(X) &= \sum_{i=1}^{N} \left(X_i - E(X)\right)^2 P_X(X_i) \\ &= \sum_{i=1}^{N} \left(X_i^2 - 2X_iE(X) + E(X)^2\right) P_X(X_i) \\ &= \sum_{i=1}^{N} X_i^2 P_X(X_i) - \sum_{i=1}^{N} 2X_iE(X)P_X(X_i) \\ &\quad + \sum_{i=1}^{N} E(X)^2 P_X(X_i) \\ &= E(X^2) - 2E(X)\left[\sum_{i=1}^{N} X_i P_X(X_i)\right] + E(X)^2 \\ &= E(X^2) - \left(E(X)\right)^2 \end{aligned}$$

Variance

$$var(X) = E(X^2) - \left(E(X)\right)^2$$

$$var(XY) = E(X^2Y^2) - \left(E(XY)\right)^2$$

$$= E(X^2)E(Y^2) - \left(E(X)E(Y)\right)^2$$

$$= \left[var(X) + \left(E(X)\right)^2\right]\left[var(Y) + \left(E(Y)\right)^2\right] - \left(E(X)E(Y)\right)^2$$

$$= var(X)var(Y) + var(X)\left(E(Y)\right)^2 + var(Y)\left(E(X)\right)^2$$

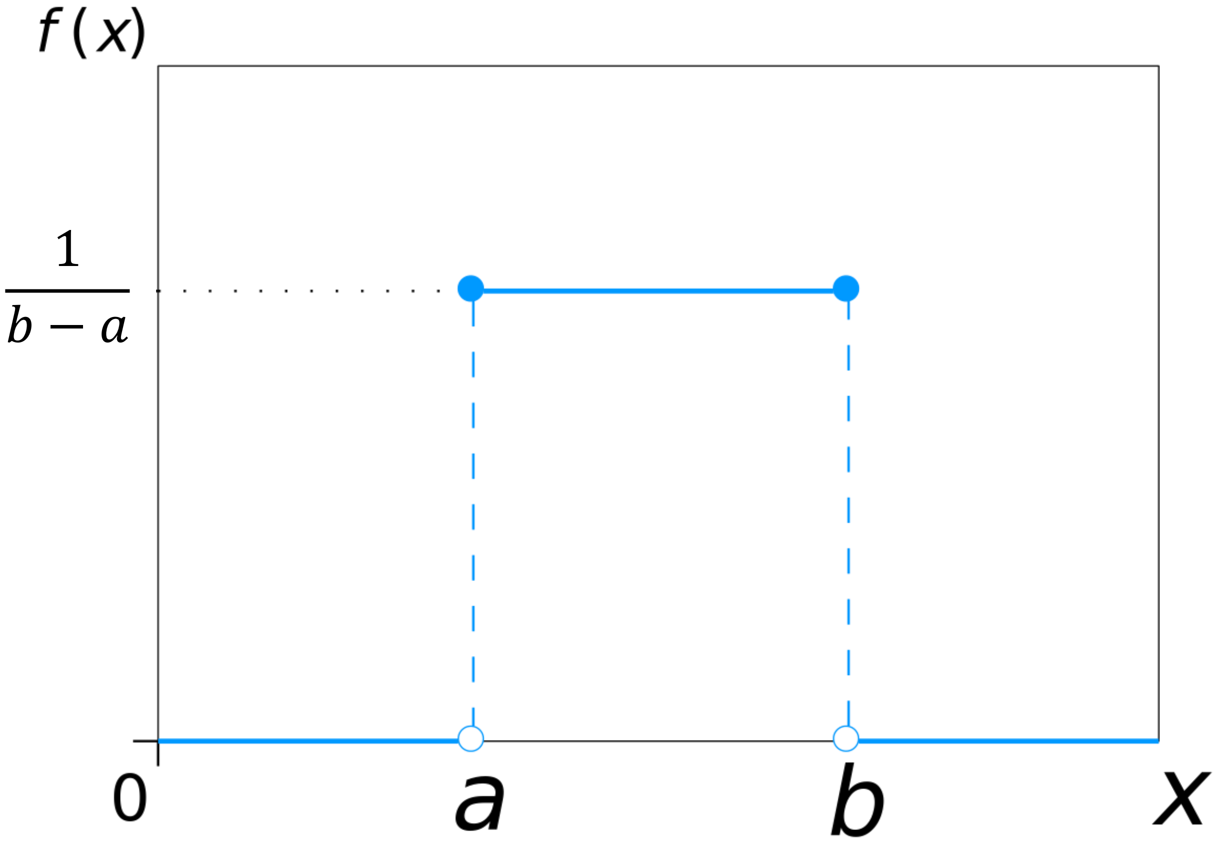
Initialization Methods

❖ Xavier Glorot Initialization

Uniform Distribution

$$X \sim U(a, b) \qquad E[X] = \frac{a+b}{2}$$

$$f(x) = \frac{1}{b-a} \qquad var[X] = \frac{(b-a)^2}{12}$$



Initialization Methods

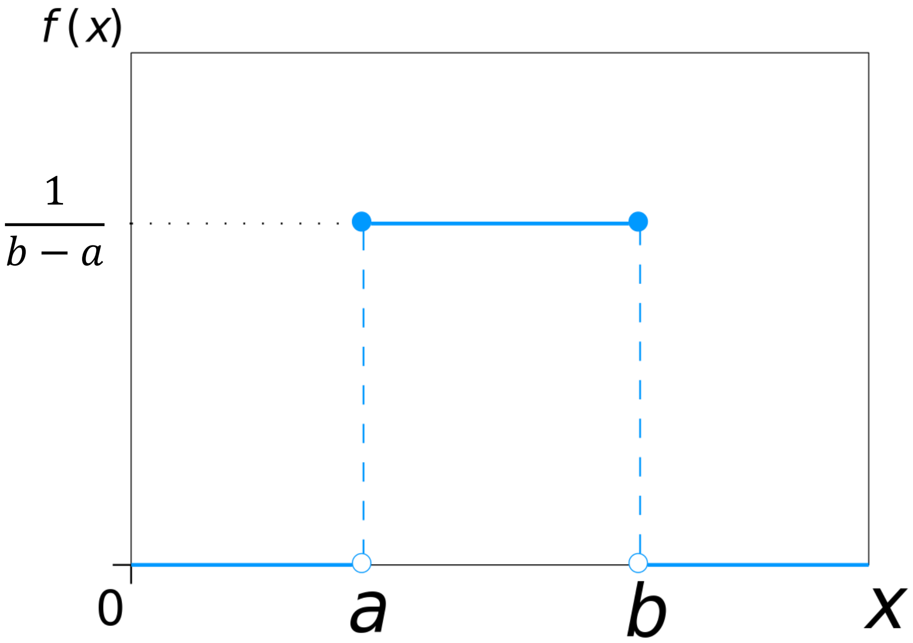
Uniform Distribution

$X \sim U(a, b)$

$$E[X] = \frac{a+b}{2}$$

$$f(x) = \frac{1}{b-a}$$

$$var[X] = \frac{(b-a)^2}{12}$$



$$E[X] = \int_{-\infty}^{\infty} x f(x) dx = \int_a^b x \frac{1}{b-a} dx$$

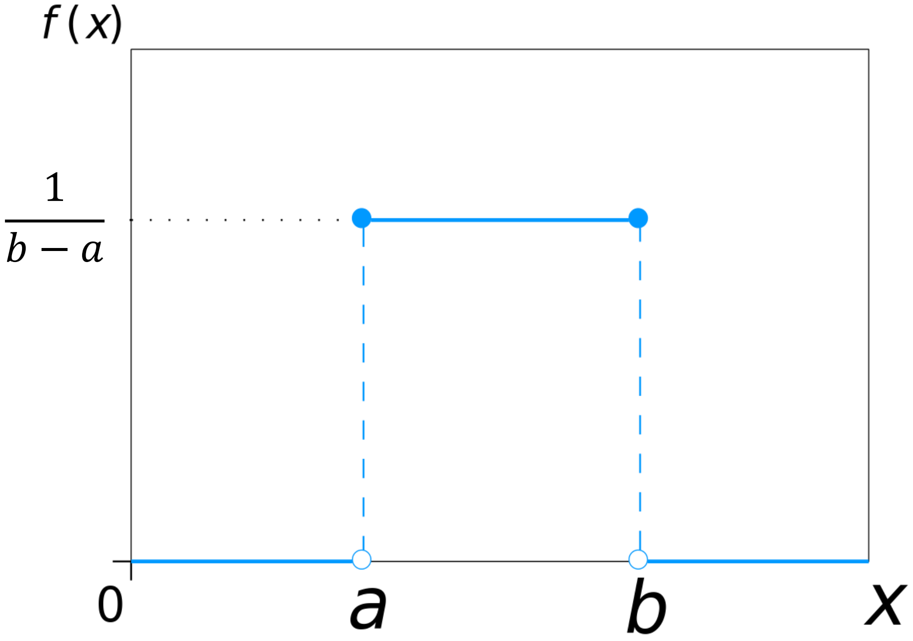
$$= \frac{x^2}{2(b-a)} \Big|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}$$

Initialization Methods

Uniform Distribution

$$X \sim U(a,b) \qquad E[X] = \frac{a+b}{2}$$

$$f(x) = \frac{1}{b-a} \qquad var[X] = \frac{(b-a)^2}{12}$$



$$var[X] = E\left(\left(X - E(X)\right)^2\right) = \int_{-\infty}^{\infty} \left(x - E(X)\right)^2 f(x)dx$$

$$= \int_a^b \left(x - \frac{a+b}{2}\right)^2 \frac{1}{b-a} dx$$

$$= \frac{1}{b-a}\left[\int_a^b x^2 dx - \int_a^b 2x\frac{a+b}{2}dx + \int_a^b \left(\frac{a+b}{2}\right)^2 dx\right]$$

$$= \frac{1}{b-a}\left[\frac{x^3}{3}\Big|_a^b - \frac{x^2(a+b)}{2}\Big|_a^b + \left(\frac{a+b}{2}\right)^2 x\Big|_a^b\right]$$

$$= \frac{1}{b-a}\left[\frac{b^3-a^3}{3} - \frac{(b^2-a^2)(a+b)}{2} + \left(\frac{a+b}{2}\right)^2 (b-a)\right]$$

$$= \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{2} + \frac{a^2+2ab+b^2}{4}$$

$$= \frac{4(a^2+ab+b^2) - 3(a^2+2ab+b^2)}{12} = \frac{(b-a)^2}{12}$$

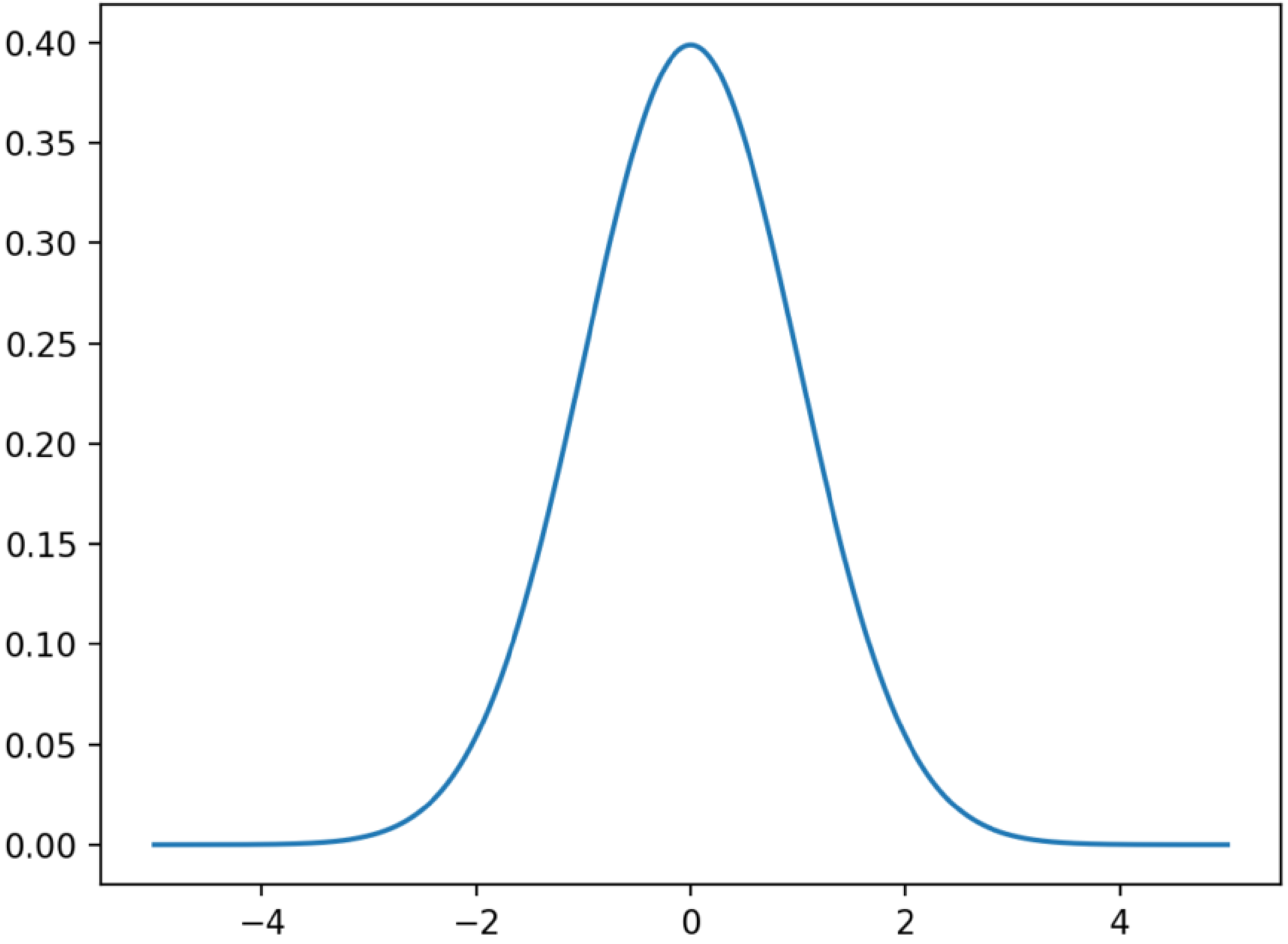
Initialization Methods

❖ Xavier Initialization

Gaussian Distribution

$$X \sim \boldsymbol{N}(\mu, \sigma^2\ )$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

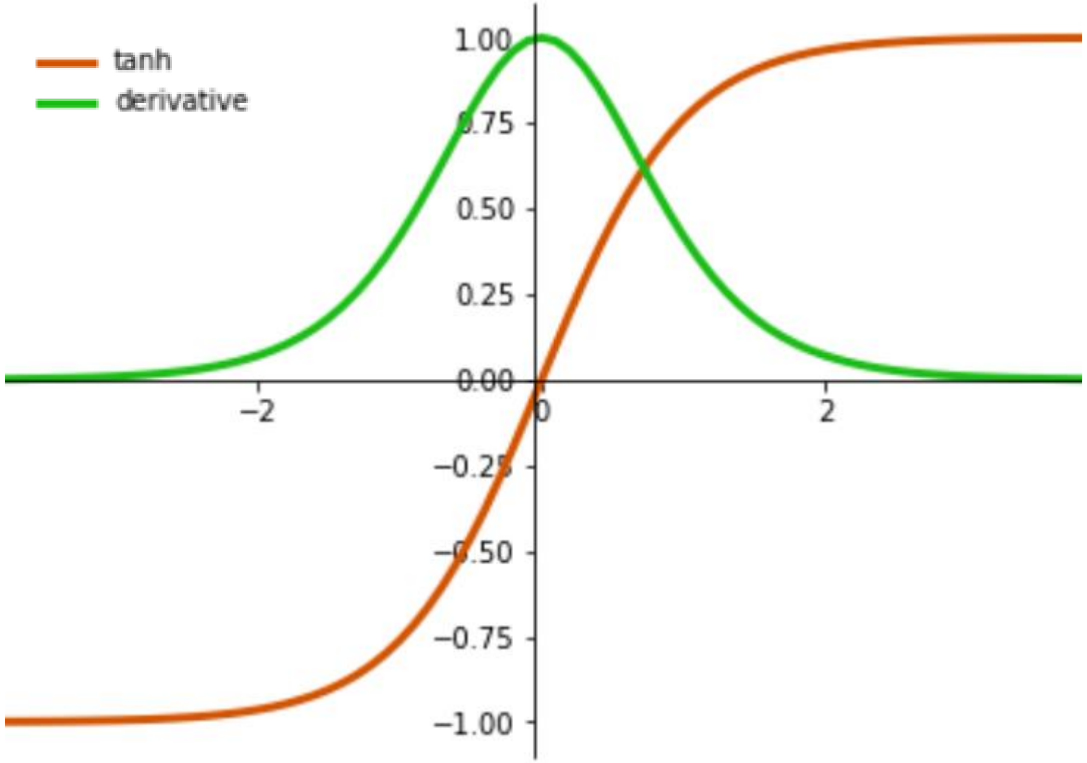


Maclaurin series

Tính giá trị xấp xỉ hàm f(x) cho những giá trị $x \approx 0$

$$f(x) = \sum_{n=0}^{\infty} f^{(n)}(0)\frac{x^n}{n!}$$

$$= f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f^{(3)}(0)}{3!}x^3 + \cdots$$



$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = 1 - \frac{2}{e^{2x}+1} = \frac{2}{e^{-2x}+1} - 1$$

$$\tanh(0) = 0$$

$$\tanh'(0) = 1 - tanh^2(0) = 1$$

$$\tanh''(0) = \left(1 - tanh^2(0)\right)'$$

$$= -2tanh(0)\tanh'(0) = 0$$

$$\tanh^{(3)}(0) = \left(-2tanh(0)\tanh'(0)\right)'$$

$$= -2[\tanh'(0)\tanh'(0) + \tanh''(0)tanh(0)] = -2$$

$$\tanh(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f^{(3)}(0)}{3!}x^3 + \cdots$$

$$= x - \frac{2x^3}{3!} + \cdots$$

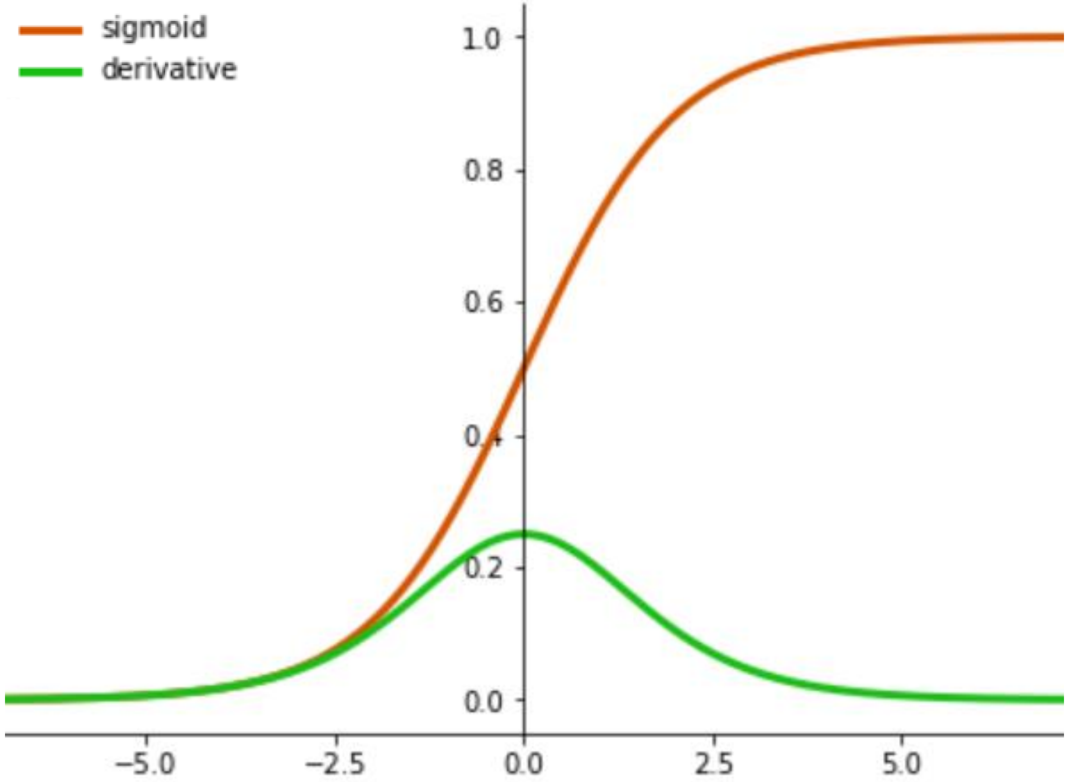
➡ $\tanh(x) \approx x$

Maclaurin series

Tính giá trị xấp xỉ hàm f(x) cho những giá trị $x \approx 0$

$$f(x) = \sum_{n=0}^{\infty} f^{(n)}(0)\frac{x^n}{n!}$$

$$= f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f^{(3)}(0)}{3!}x^3 + \cdots$$



$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

$$\text{sigmoid}(0) = \frac{1}{2}$$

$$\text{sigmoid}'(0) = \text{sigmoid}(0)\,(1 - \text{sigmoid}(0)) = \frac{1}{4}$$

$$\text{sigmoid}''(0) = [\text{sigmoid}(0)\,(1 - \text{sigmoid}(0))]'$$

$$= \text{sigmoid}'(0) - 2\,\text{sigmoid}(0)\text{sigmoid}'(0) = 0$$

$$\text{sigmoid}(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f^{(3)}(0)}{3!}x^3 + \cdots$$

$$= \frac{1}{2} + \frac{x}{4} + \cdots$$

$$\Rightarrow \quad \text{sigmoid}(x) \approx \frac{1}{2} + \frac{x}{4}$$

Initialization Methods

$x_i \sim a_i$

❖ Xavier Initialization

$$E(XY) = E(X)E(Y)$$

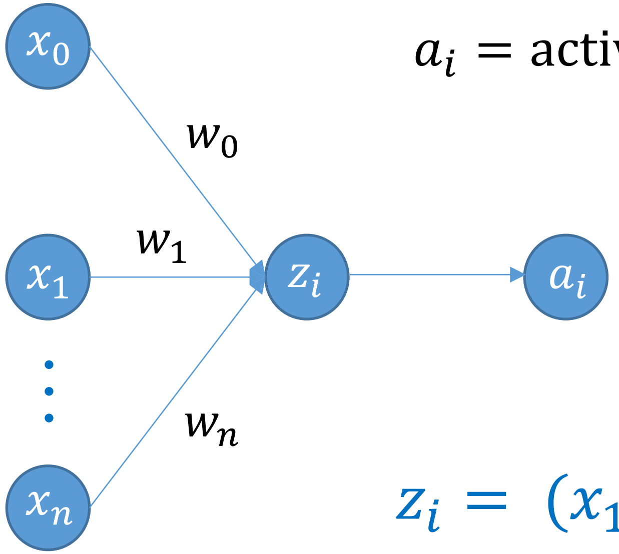
$$var(XY) = var(X)var(Y) + var(X)\left(E(Y)\right)^2 + var(y)\left(E(X)\right)^2$$

Uniform Distribution

$$X \sim U(a, b)$$

$$f(x) = \frac{1}{b-a}$$

$$var[X] = \frac{(b-a)^2}{12}$$



$$a_i = \text{activation}(z_i)$$

$$E(X) = 0$$
$$E(W) = 0$$
$$b = 0$$

$$z_i = (x_1 w_1 + \cdots + x_n w_n + b)$$

$$\text{var}(z_i) = \text{var}(x_1 w_1 + \cdots + x_n w_n + b) = n\text{var}(x_i w_i) = n\text{var}(x_i)\text{var}(w_i)$$

activation = tanh ➡ $a_i = \tanh(z_i) \approx z_i$ ➡ $\text{var}(a_i) = \text{var}(z_i)$

$\text{var(X)} = \text{var}(\mathbf{a})$ $\xrightarrow{\text{iid}}$ $\text{var}(x_i) = \text{var}(a_i)$ ➡ $n\text{var}(w_i) = 1$

➡ $\text{var}(w_i) = \frac{1}{n}$

Initialization Methods

❖ Xavier Initialization

activation = tanh

$$E(XY) = E(X)E(Y)$$

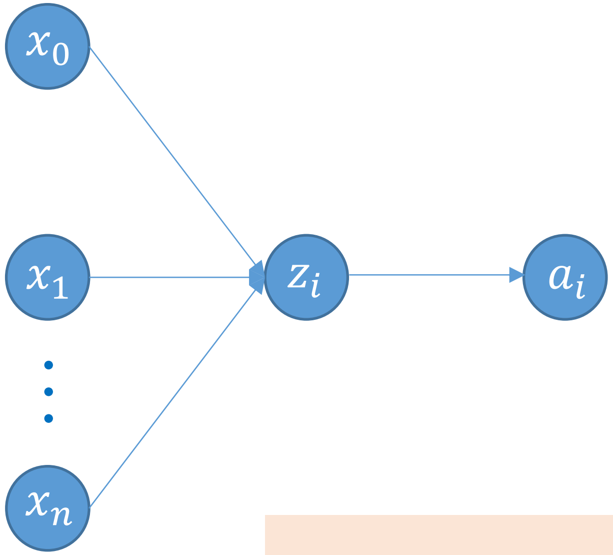
$$var(XY) = var(X)var(Y) + var(X)\left(E(Y)\right)^2 + var(y)\left(E(X)\right)^2$$

Uniform Distribution

$$X \sim U(a, b)$$

$$f(x) = \frac{1}{b-a}$$

$$var[X] = \frac{(b-a)^2}{12}$$



$$\text{var}(w_i) \approx \frac{1}{n}$$

$$w_i \sim U(-r, r)$$

$$var[w_i] = \frac{r^2}{3}$$

$$W_i \sim U\left(-\frac{\sqrt{3}}{\sqrt{n}}, \frac{\sqrt{3}}{\sqrt{n}}\right)$$

Initialization Methods

❖ Xavier Initialization

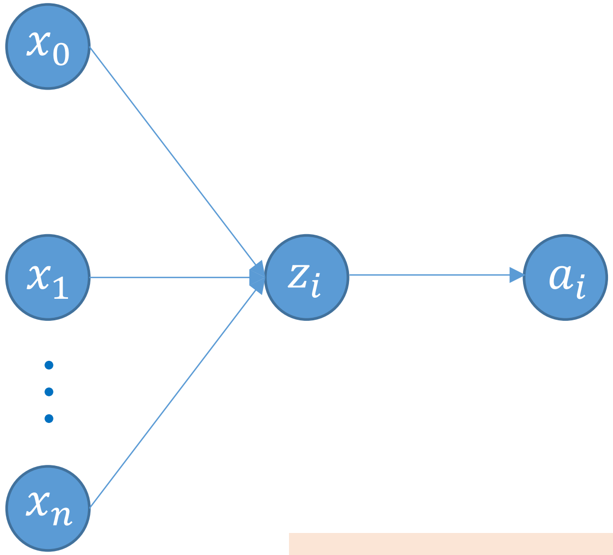
activation = tanh

$$E(XY) = E(X)E(Y)$$

$$var(XY) = var(X)var(Y) + var(X)\big(E(Y)\big)^2 + var(y)\big(E(X)\big)^2$$

Gaussian Distribution

$$X \sim N(0, \sigma^2)$$



$$\text{var}(w_i) \approx \frac{1}{n}$$

$$w_i \sim N(0, \sigma^2)$$

$$\sigma^2 = \frac{1}{n} \quad ➡ \quad \sigma = \frac{1}{\sqrt{n}}$$

$$W_i \sim N\left(0, \frac{1}{n}\right)$$

Initialization Methods

❖ Xavier Initialization

$\text{activation} = \tanh$

Uniform Distribution

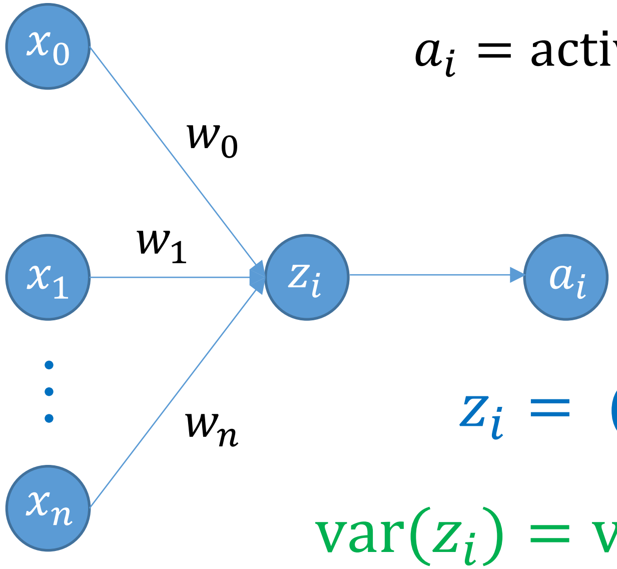
$$W_{ij} \sim U\left(-\frac{\sqrt{3}}{\sqrt{n}}, \frac{\sqrt{3}}{\sqrt{n}}\right)$$

Gaussian Distribution

$$W_{ij} \sim \boldsymbol{N}\left(0, \frac{1}{n}\right)$$

Initialization Methods

❖ Xavier Initialization



$$a_i = \text{activation}(z_i)$$

$$E(X) = 0$$
$$E(W) = 0$$
$$b = 0$$

$$E(XY) = E(X)E(Y)$$

$$var(XY) = var(X)var(Y) + var(X)\big(E(Y)\big)^2 + var(y)\big(E(X)\big)^2$$

Uniform Distribution

$$X{\sim}U(a,b)$$

$$f(x) = \frac{1}{b-a}$$

$$var[X] = \frac{(b-a)^2}{12}$$

$$z_i = (x_1 w_1 + \cdots + x_n w_n + b)$$

$$\text{var}(z_i) = \text{var}(x_1 w_1 + \cdots + x_n w_n + b)$$
$$= n\text{var}(x_i w_i) = n\text{var}(x_i)\text{var}(w_i)$$

$\text{activation} = \text{sigmoid}$ ➡ $a_i = \text{sigmoid}(z_i) \approx \frac{1}{2} + \frac{z_i}{4}$

➡ $16\text{var}(a_i) = \text{var}(z_i)$

$\text{var(X)} = \text{var}(\mathbf{a})$ $\xrightarrow{\text{iid}}$ $\text{var}(x_i) = \text{var}(a_i)$ ➡ $n\text{var}(w_i) = 16$

➡ $\text{var}(w_i) = \frac{16}{n}$

Initialization Methods

❖ Xavier Initialization

activation = sigmoid

$$E(XY) = E(X)E(Y)$$

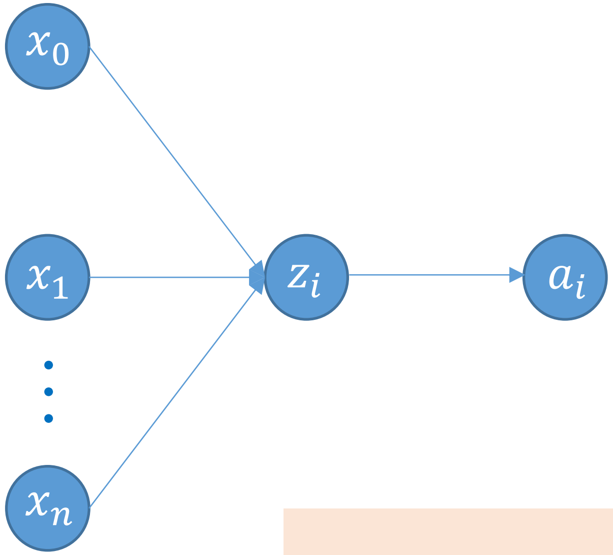
$$var(XY) = var(X)var(Y) + var(X)\big(E(Y)\big)^2 + var(y)\big(E(X)\big)^2$$

Uniform Distribution

$$X \sim U(a, b)$$

$$f(x) = \frac{1}{b-a}$$

$$var[X] = \frac{(b-a)^2}{12}$$



$$\mathrm{var}(w_i) \approx \frac{16}{n}$$

$$w_i \sim U(-r, r)$$

$$var[w_i] = \frac{r^2}{3}$$

$$W_i \sim U\left(-\frac{4\sqrt{3}}{\sqrt{n}}, \frac{4\sqrt{3}}{\sqrt{n}}\right)$$

Initialization Methods

❖ Xavier Initialization

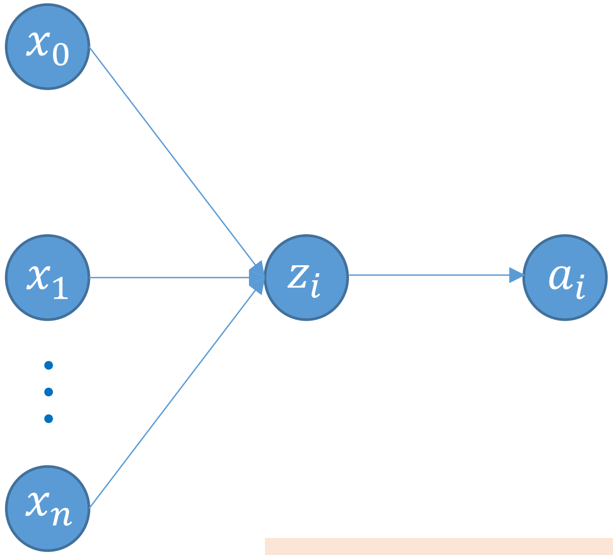
activation = sigmoid

$$E(XY) = E(X)E(Y)$$

$$var(XY) = var(X)var(Y) + var(X)\left(E(Y)\right)^2 + var(y)\left(E(X)\right)^2$$

Gaussian Distribution

$$X \sim N(0, \sigma^2)$$

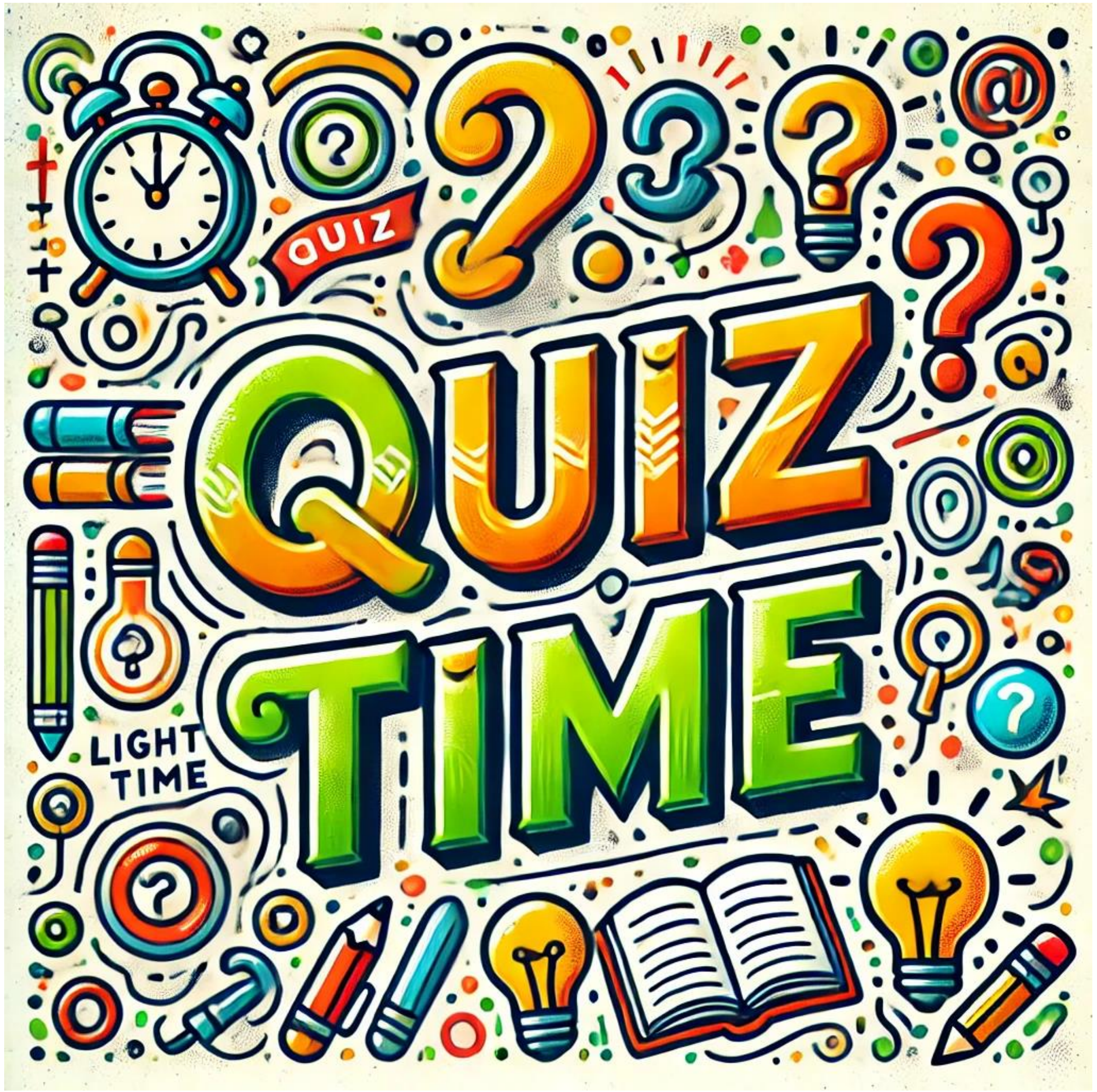


$$\text{var}(w_i) \approx \frac{16}{n}$$

$$w_i \sim N(0, \sigma^2)$$

$$\sigma^2 = \frac{1}{n}$$

$$W_i \sim N\left(0, \frac{16}{n}\right)$$


QUIZ
QUIZ
TIME
LIGHT
TIME

Question 1

❖ Chuẩn hóa dữ liệu nào nên dùng cho Glorot Initialization (chọn nhiều đáp án)?

a) Sau chuẩn hóa có range là [0, 255]

b) Có range là [0, 1]

c) Có range là [-1, 1]

d) Dạng z-score

Question 2

❖ Glorot Initialization giả định activation đang dùng là gì (chọn nhiều đáp án)?

a) Sigmoid

b) Tanh

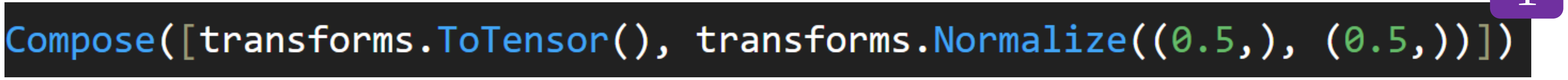
c) ReLU

d) PReLU

Question 3

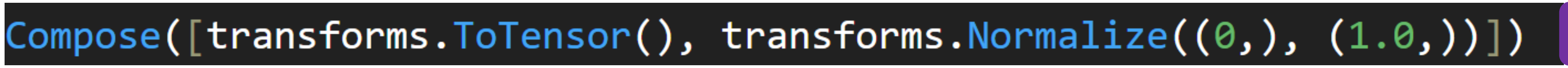
❖ Code nào nên dùng khi sử dụng với Xavier Init.?

1



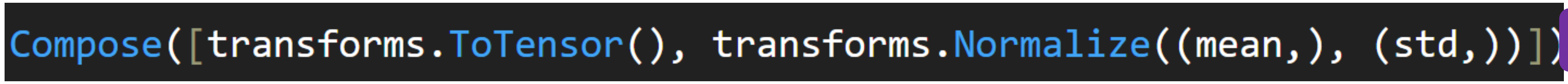
```
Compose([transforms.ToTensor(), transforms.Normalize((0.5,), (0.5,))])
```

2



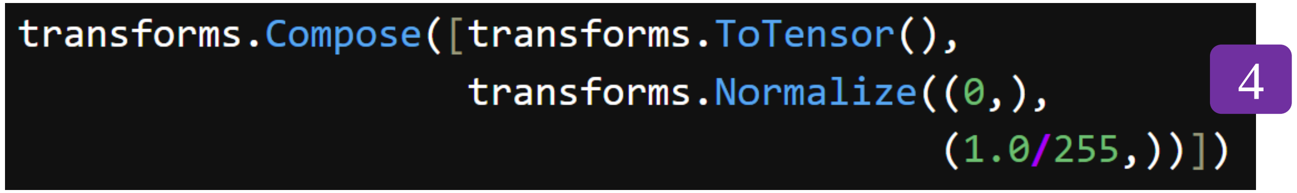
```
Compose([transforms.ToTensor(), transforms.Normalize((0,), (1.0,))])
```

3



```
Compose([transforms.ToTensor(), transforms.Normalize((mean,), (std,))])
```

4



```
transforms.Compose([transforms.ToTensor(),
                    transforms.Normalize((0,),
                                         (1.0/255,))])
```

a) Code 1

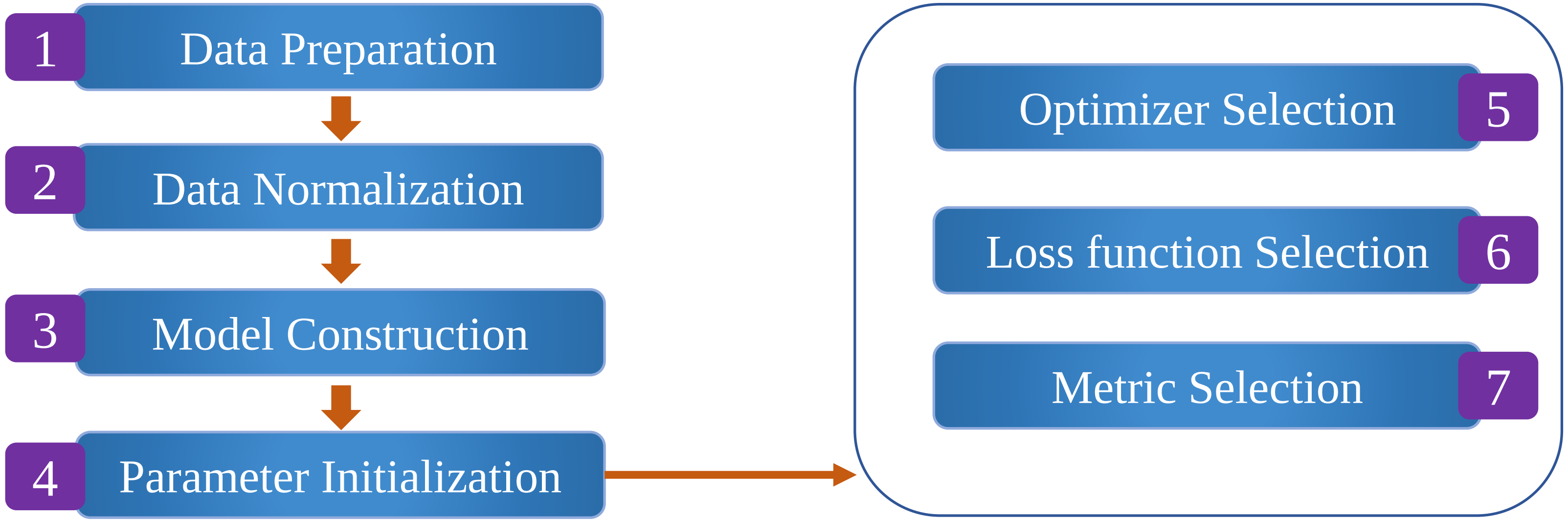
b) Code 2

c) Code 3

d) Code 4

Question 4

❖ Dựa vào kiến thức AIO tới thời điểm này, nếu chọn bước (4) là Glorot init., thì bước nào cần có hành động (gì đó) tương ứng?



a) Bước (1)

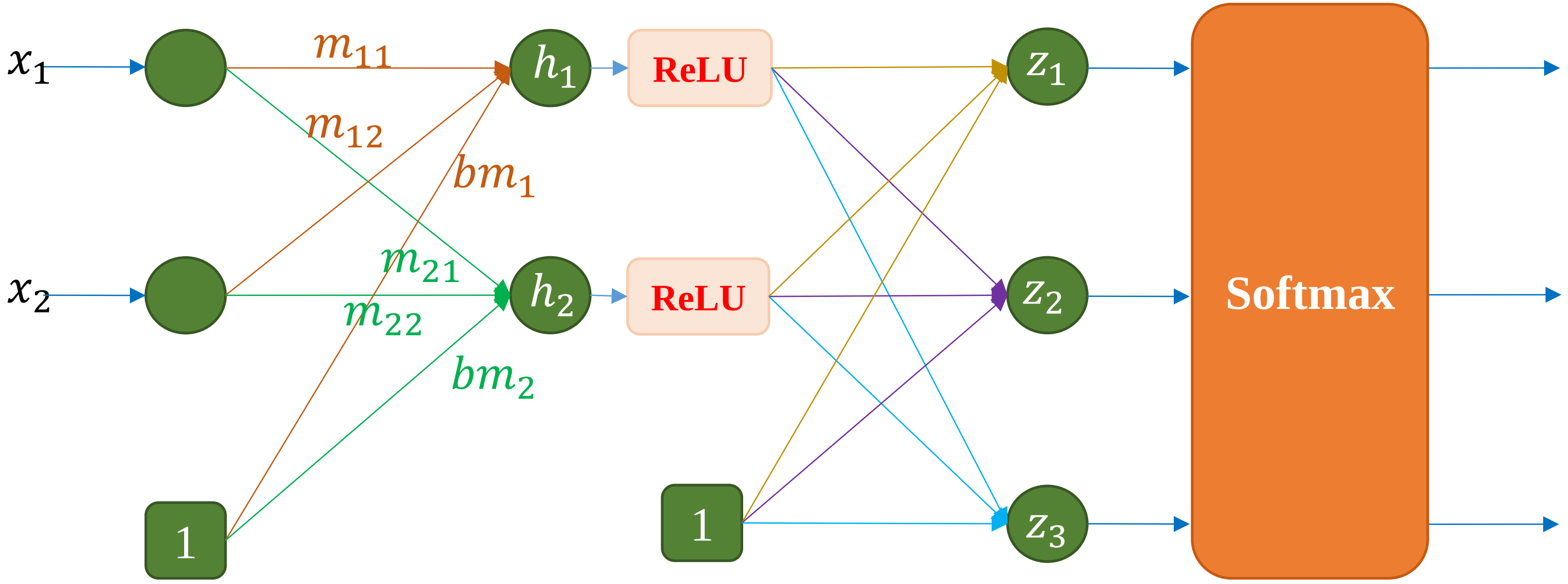
b) Bước (2)

c) Bước (3)

d) Bước (5)

Question 5

❖ Hãy chọn 1 giải pháp hay nhất để khắc phục vấn đề dying relu?



a) Thay relu bằng sigmoid

b) Thay relu bằng tanh

c) Thay relu bằng prelu

d) Giữ relu và dùng He init.

Question 6

❖ Glorot Init. có những giả định (điều kiện cho trước) nào?

a) Activation là sigmoid

b) Activation là tanh

c) Activation là relu

d) Data input có mean=0

Outline

SECTION 1

Case Studies

SECTION 2

Xavier Glorot Init.

SECTION 3

Kaiming He Init.

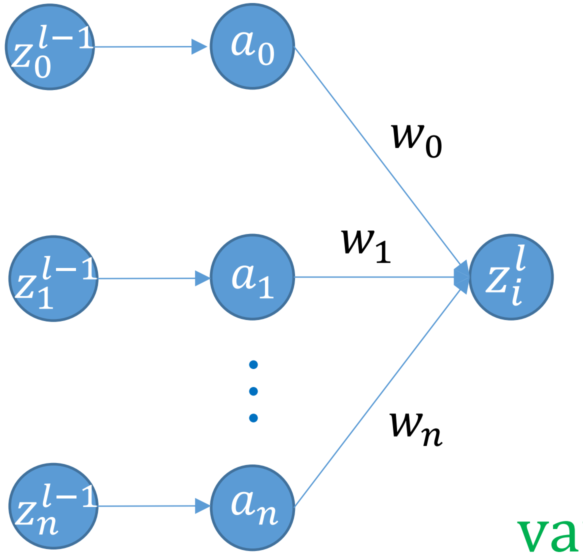
$$W_i \sim U\left(-\frac{\sqrt{6}}{\sqrt{n}}, \frac{\sqrt{6}}{\sqrt{n}}\right)$$

$$W_i \sim N\left(0, \frac{2}{n}\right)$$

Initialization Methods

$x_i \sim a_i$

❖ Kaiming He Initialization



$$a_i = \text{activation}(z_i)$$

$$E(W) = 0$$
$$b = 0$$

$$E(XY) = E(X)E(Y)$$
$$var(XY) = var(X)var(Y) + var(X)\big(E(Y)\big)^2 + var(y)\big(E(X)\big)^2$$

$$z_i = (a_1 w_1 + \cdots + a_n w_n + b)$$
$$\text{var}(z_i) = \text{var}(a_1 w_1 + \cdots + a_n w_n + b)$$

Uniform Distribution

$$X \sim U(a, b)$$
$$f(x) = \frac{1}{b-a}$$
$$var[X] = \frac{(b-a)^2}{12}$$

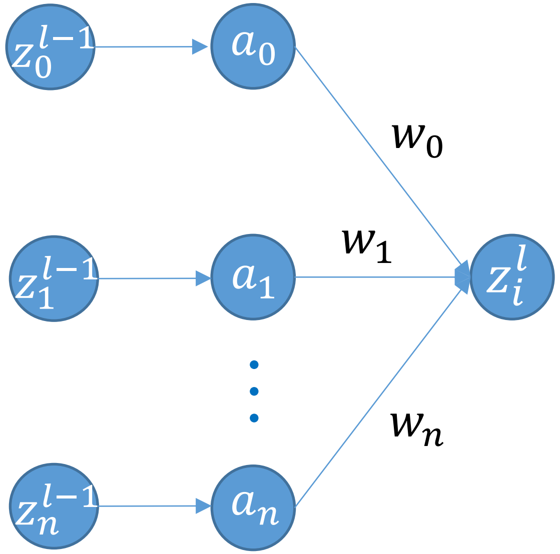
$$\text{activation} = \text{relu} \quad ➡ \quad a_i = max(0, z_i)$$

$$\text{var}(z^{l-1}) = \text{var}(z^l) \xrightarrow{\text{iid}} \text{var}(z_i^{l-1}) = \text{var}(z_i^l) \; ➡ \; \text{nvar}(w_i) = 2$$
$$➡ \; \text{var}(w_i) = \frac{2}{n}$$

Initialization Methods

❖ He Initialization

activation = relu



$$E(XY) = E(X)E(Y)$$

$$var(XY) = var(X)var(Y) + var(X)\big(E(Y)\big)^2 + var(y)\big(E(X)\big)^2$$

$$\text{var}(w_i) \approx \frac{2}{n}$$

$$w_i \sim U(-r, r)$$

$$var[w_i] = \frac{r^2}{3}$$

Uniform Distribution

$$X \sim U(a, b)$$

$$f(x) = \frac{1}{b-a}$$

$$var[X] = \frac{(b-a)^2}{12}$$

$$W_i \sim U\left(-\frac{\sqrt{6}}{\sqrt{n}}, \frac{\sqrt{6}}{\sqrt{n}}\right)$$

Initialization Methods

❖ He Initialization

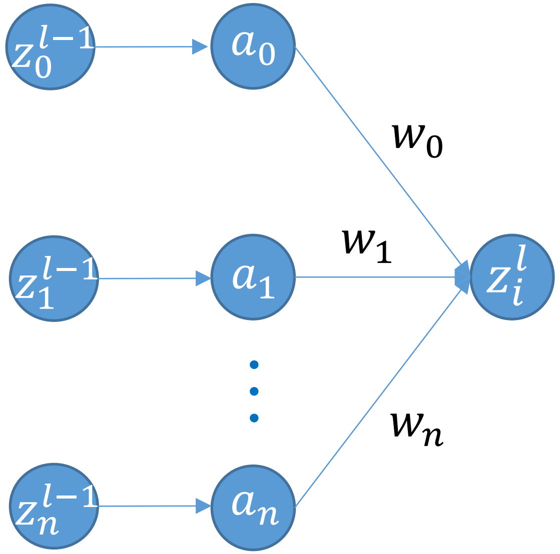
activation = he

$$E(XY) = E(X)E(Y)$$

$$var(XY) = var(X)var(Y) + var(X)\big(E(Y)\big)^2 + var(y)\big(E(X)\big)^2$$

Gaussian Distribution

$$X \sim N(0, \sigma^2)$$



$$\text{var}(w_i) \approx \frac{2}{n}$$

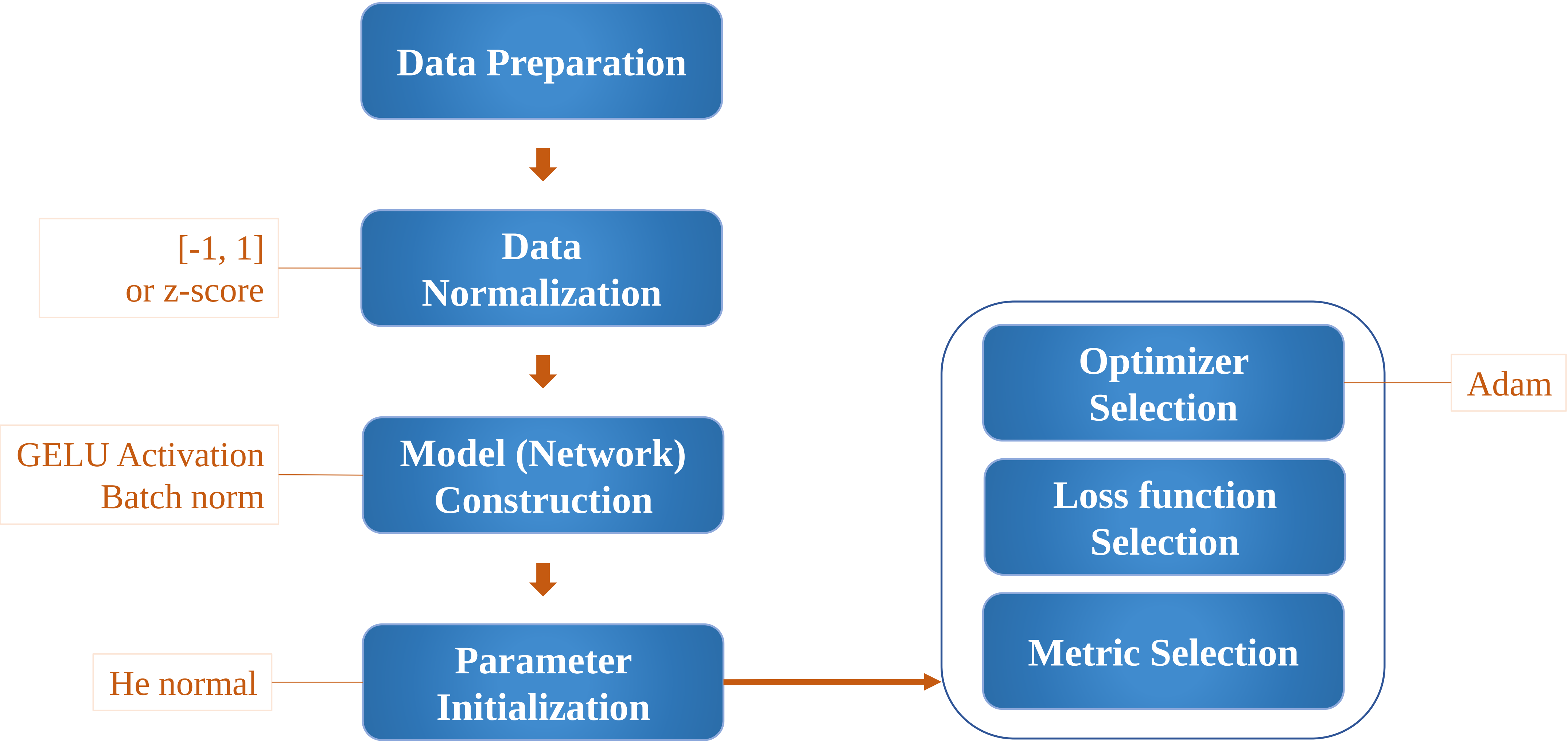
$$w_i \sim N(0, \sigma^2)$$

$$\sigma^2 = \frac{1}{n}$$

$$W_i \sim N\left(0, \frac{2}{n}\right)$$

Summary

❖ Recommendation



Further Reading

Dying ReLU

https://towardsdatascience.com/the-dying-relu-problem-clearly-explained-42d0c54e0d24

Initialization

https://www.deeplearning.ai/ai-notes/initialization/index.html