머신 러닝·딥러닝에 필요한 수학 기초 with 파이썬

Back Propagation

신경망의 미분

조준우
metamath@gmail.com

Why do we have to write the backward pass...?

- **Yes you should understand backprop**

Andrej Karpathy Follow

Dec 20, 2016 · 7 min read

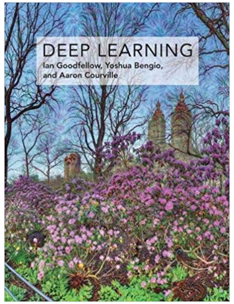
When we offered CS231n (Deep Learning class) at Stanford, we intentionally designed the programming assignments to include explicit calculations involved in backpropagation on the lowest level. The students had to implement the forward and the backward pass of each layer in raw numpy. Inevitably, some students complained on the class message boards:

> *"Why do we have to write the backward pass when frameworks in the real world, such as TensorFlow, compute them for you automatically?"*

https://medium.com/@karpathy/yes-you-should-understand-backprop-e2f06eab496b

Relationship between back-propagation and auto differentiation

- **Deep Learning Book,** Ian Goodfellow, Yoshua Bengio, Aaron Courville, page 206



In vector notation, this may be equivalently written as

$$\boxed{\nabla_{\boldsymbol{x}}} z = \boxed{\left(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right)}^{\top} \boxed{\nabla_{\boldsymbol{y}}} z, \tag{6.46}$$

where $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}$ is the $n \times m$ Jacobian matrix of g.

From this we see that the gradient of a variable $\boldsymbol{x}$ can be obtained by multiplying a Jacobian matrix $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}$ by a gradient $\nabla_{\boldsymbol{y}} z$. The back-propagation algorithm consists of performing such a Jacobian-gradient product for each operation in the graph.

Usually we do not apply the back-propagation algorithm merely to vectors, but rather to tensors of arbitrary dimensionality. Conceptually, this is exactly the same as back-propagation with vectors. The only difference is how the numbers are arranged in a grid to form a tensor. We could imagine flattening each tensor into a vector before we run back-propagation, computing a vector-valued gradient, and then reshaping the gradient back into a tensor. In this rearranged view, back-propagation is still just multiplying Jacobians by gradients.

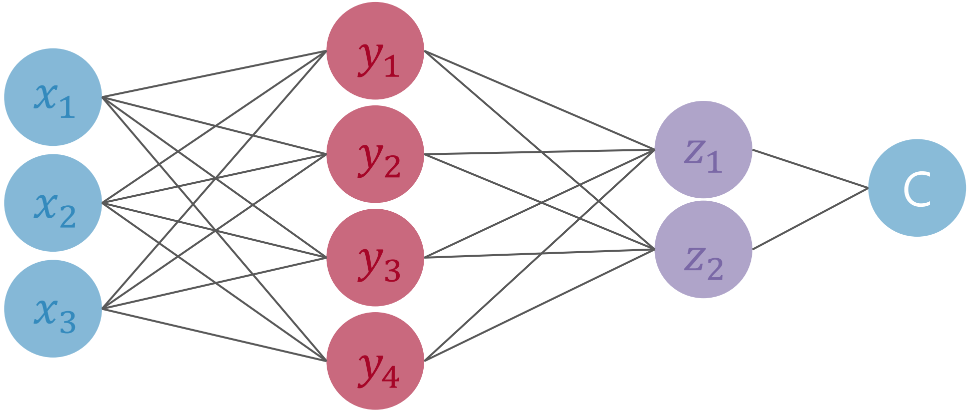
Jacobian

- 다변수 벡터함수에 대한 미분을 나타낸 행렬
- $f: \mathbb{R}^m \rightarrow \mathbb{R}^n$라는 함수가 있을 때 다음과 같은 $n \times m$ 행렬
- 행렬의 요소는 각각의 성분의 각각의 변수에 대한 편미분

$$\mathrm{J} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_m} \end{pmatrix}$$

Back-propagation

$f_1\colon \mathbb{R}^3 \to \mathbb{R}^4$ $f_2\colon \mathbb{R}^4 \to \mathbb{R}^2$ $f_3\colon \mathbb{R}^2 \to \mathbb{R}$



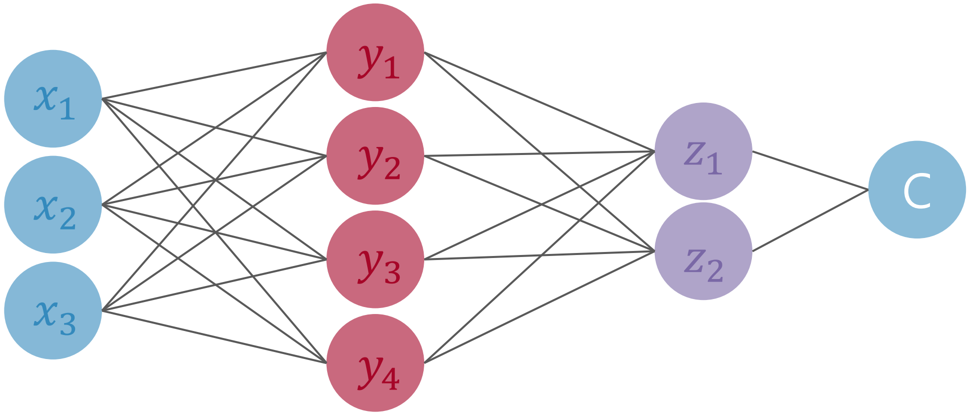
$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$: (4,3) $\frac{\partial \mathbf{z}}{\partial \mathbf{y}}$: (2,4) $\frac{\partial C}{\partial \mathbf{z}}$: (2,1)

gradient

$$\frac{\partial C}{\partial \mathbf{y}} = \left(\frac{\partial \mathbf{z}}{\partial \mathbf{y}}\right)^{\mathrm{T}} \frac{\partial C}{\partial \mathbf{z}} = (4,2)(2,1) = (4,1)$$

Back-propagation

$f_1: \mathbb{R}^3 \to \mathbb{R}^4$ $\quad f_2: \mathbb{R}^4 \to \mathbb{R}^2$ $\quad f_3: \mathbb{R}^2 \to \mathbb{R}$

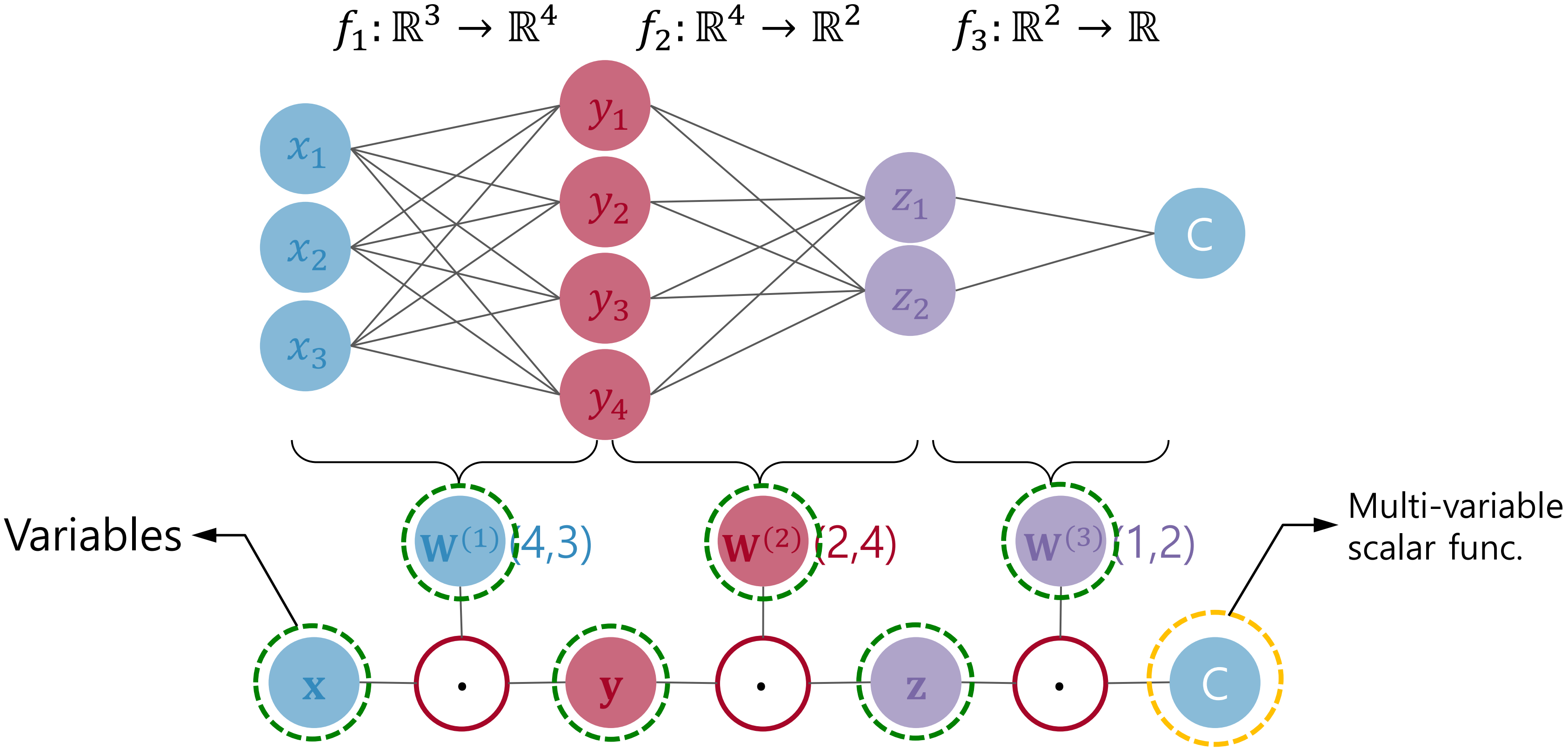


$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$: (4,3) $\qquad \frac{\partial C}{\partial \mathbf{y}}$: (4,1)

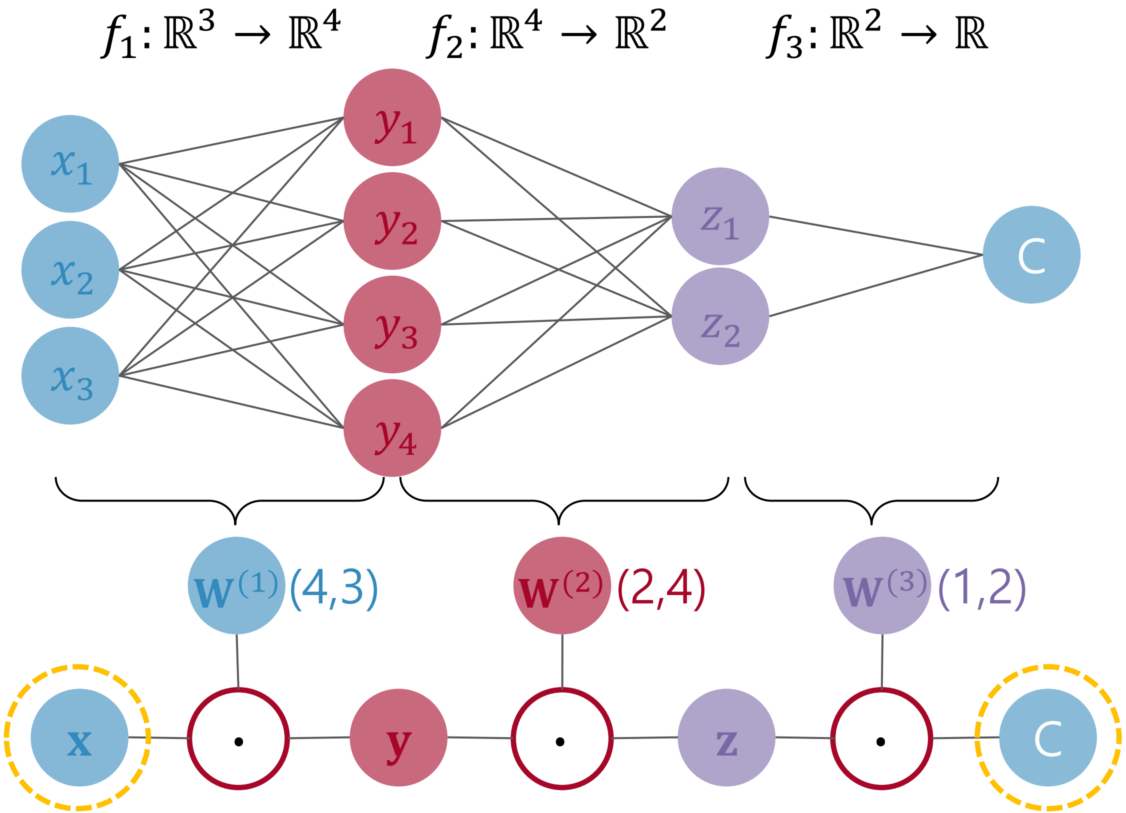
gradient

$$\frac{\partial C}{\partial \mathbf{x}} = \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right)^{\mathrm{T}} \frac{\partial C}{\partial \mathbf{y}} = (3,4)(4,1) = (3,1)$$

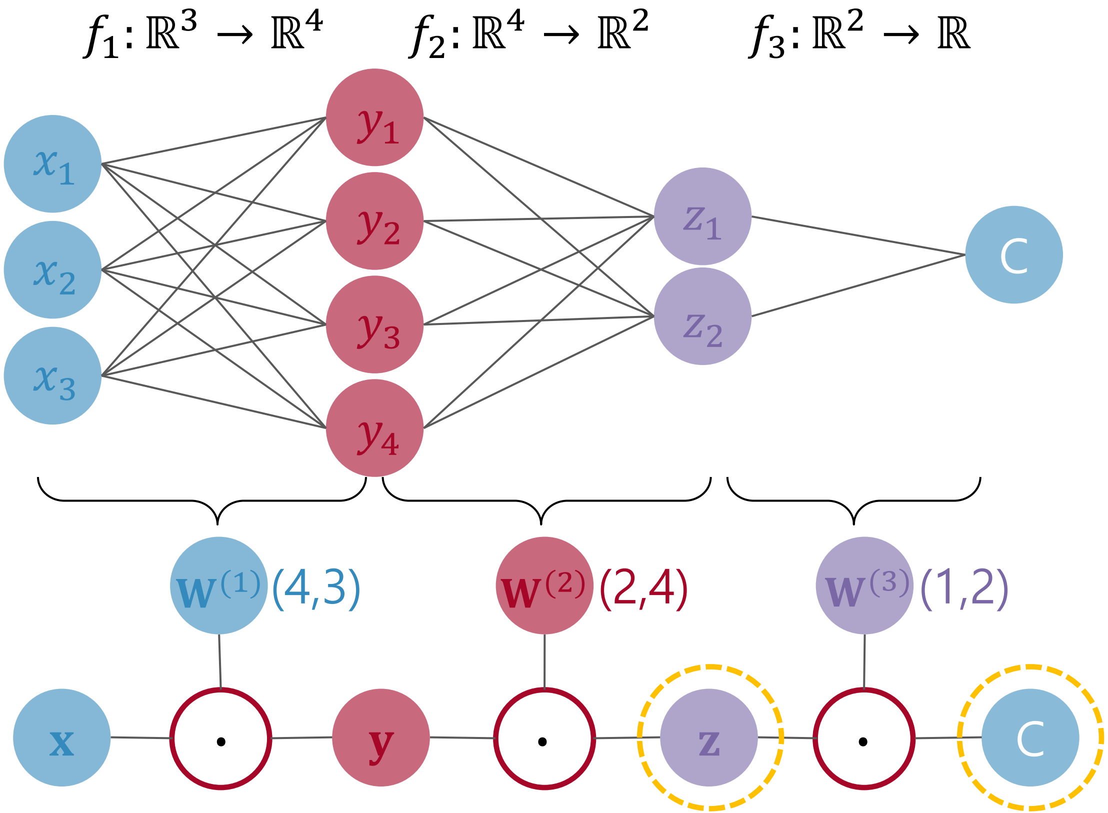
Graphs to Describe Model Structure



C를 입력 $\mathbf{x}$로 미분: $\partial C / \partial \mathbf{x}$



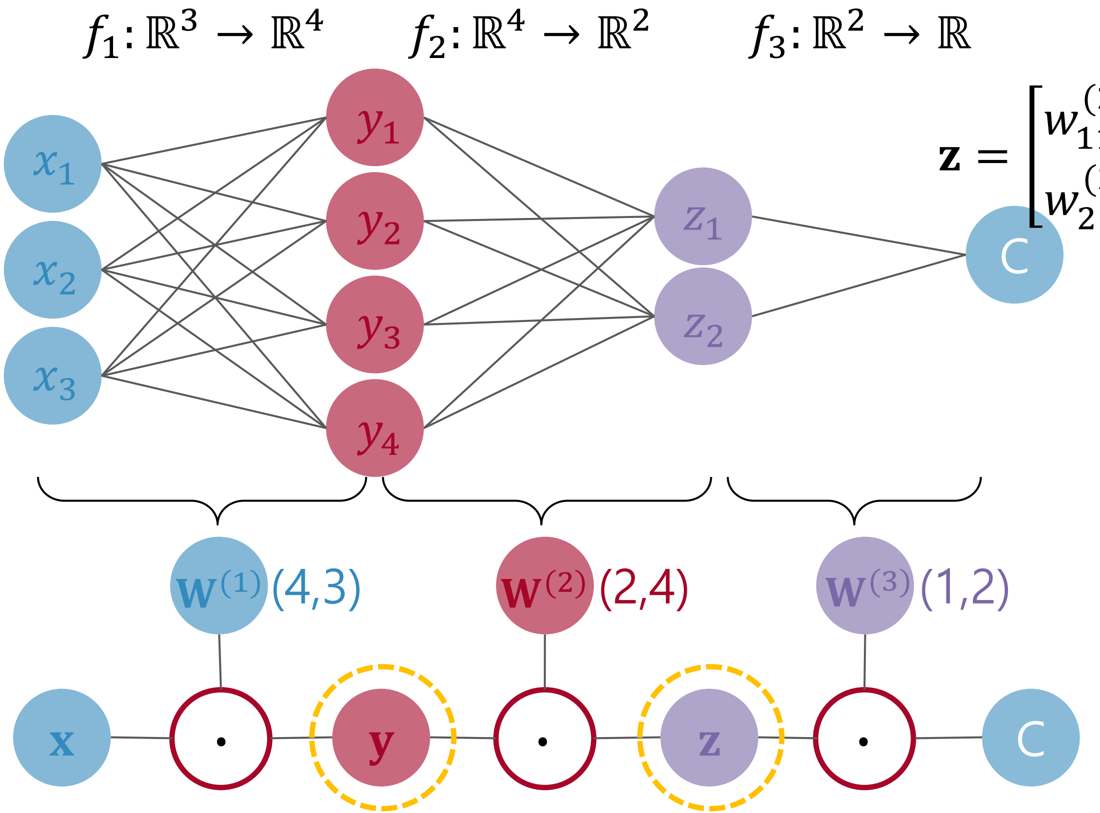
$\partial C / \partial \mathbf{z}$



$$C = w_1^{(3)} z_1 + w_2^{(3)} z_2$$

$$\frac{\partial C}{\partial \mathbf{z}} = \begin{bmatrix} w_1^{(3)} \\ w_2^{(3)} \end{bmatrix} = {\mathbf{W}^{(3)}}^{\mathrm{T}}$$

$\partial\mathbf{z}/\partial\mathbf{y}$



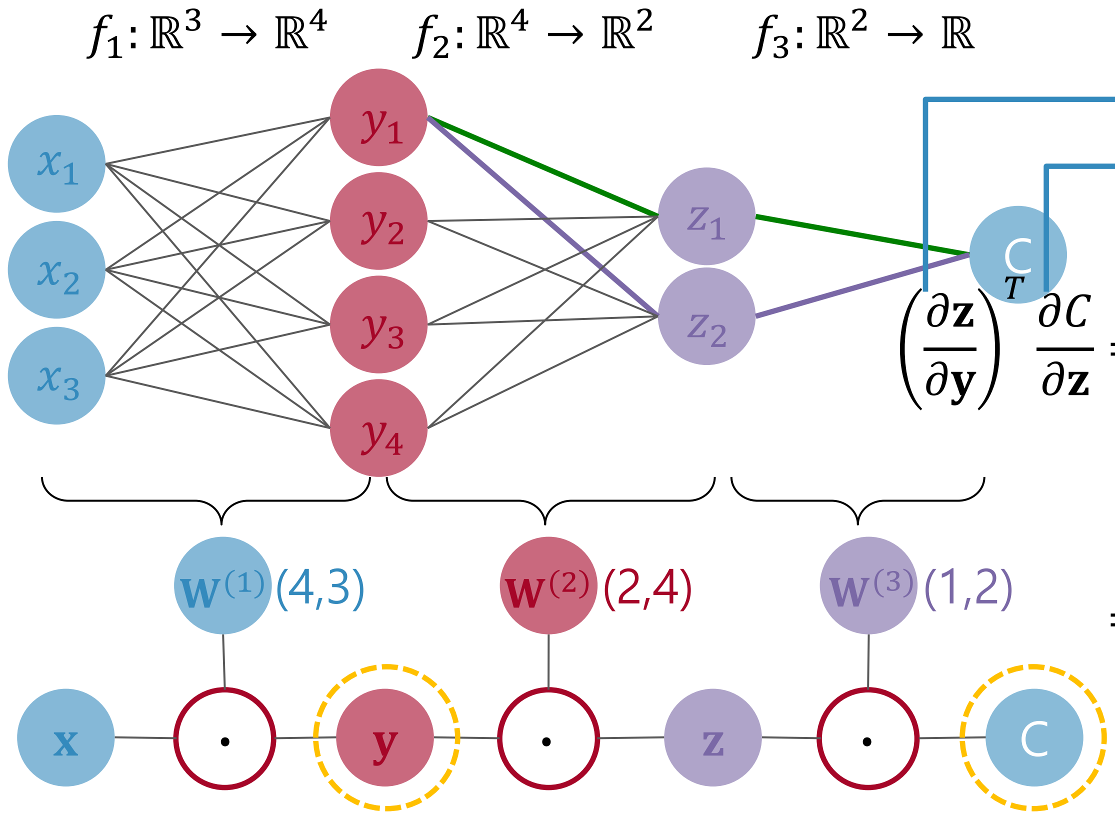
$$\mathbf{z} = \begin{bmatrix} w_{11}^{(2)} y_1 + w_{12}^{(2)} y_2 + w_{13}^{(2)} y_3 + w_{14}^{(2)} y_4 \\ w_{21}^{(2)} y_1 + w_{22}^{(2)} y_1 + w_{23}^{(2)} y_1 + w_{24}^{(2)} y_1 \end{bmatrix}$$

$$\frac{\partial \mathbf{z}}{\partial \mathbf{y}} = \begin{bmatrix} \frac{\partial z_1}{\partial y_1} & \frac{\partial z_1}{\partial y_2} & \frac{\partial z_1}{\partial y_3} & \frac{\partial z_1}{\partial y_4} \\ \frac{\partial z_2}{\partial y_1} & \frac{\partial z_2}{\partial y_2} & \frac{\partial z_2}{\partial y_3} & \frac{\partial z_2}{\partial y_4} \end{bmatrix}$$

$$= \begin{bmatrix} w_{11}^{(2)} & w_{12}^{(2)} & w_{13}^{(2)} & w_{14}^{(2)} \\ w_{21}^{(2)} & w_{22}^{(2)} & w_{23}^{(2)} & w_{24}^{(2)} \end{bmatrix}$$

$$= \mathbf{W}^{(2)}$$

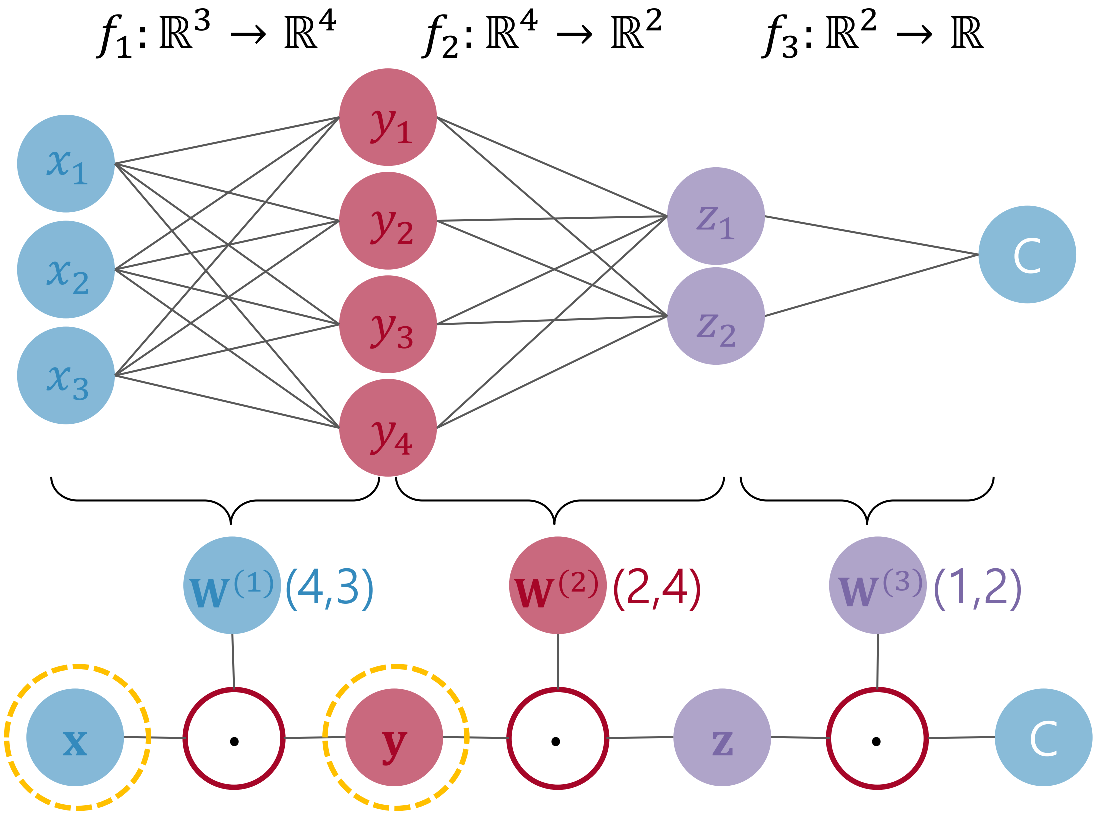
$$\frac{\partial C}{\partial \mathbf{y}} = \left(\frac{\partial \mathbf{z}}{\partial \mathbf{y}}\right)^T \frac{\partial C}{\partial \mathbf{z}}$$



$$\left(\frac{\partial \mathbf{z}}{\partial \mathbf{y}}\right)^T \frac{\partial C}{\partial \mathbf{z}} = \begin{bmatrix} \partial z_1/\partial y_1 & \partial z_2/\partial y_1 \\ \partial z_1/\partial y_2 & \partial z_2/\partial y_2 \\ \partial z_1/\partial y_3 & \partial z_2/\partial y_3 \\ \partial z_1/\partial y_4 & \partial z_2/\partial y_4 \end{bmatrix} \begin{bmatrix} \partial C/\partial z_1 \\ \partial C/\partial z_2 \end{bmatrix}$$

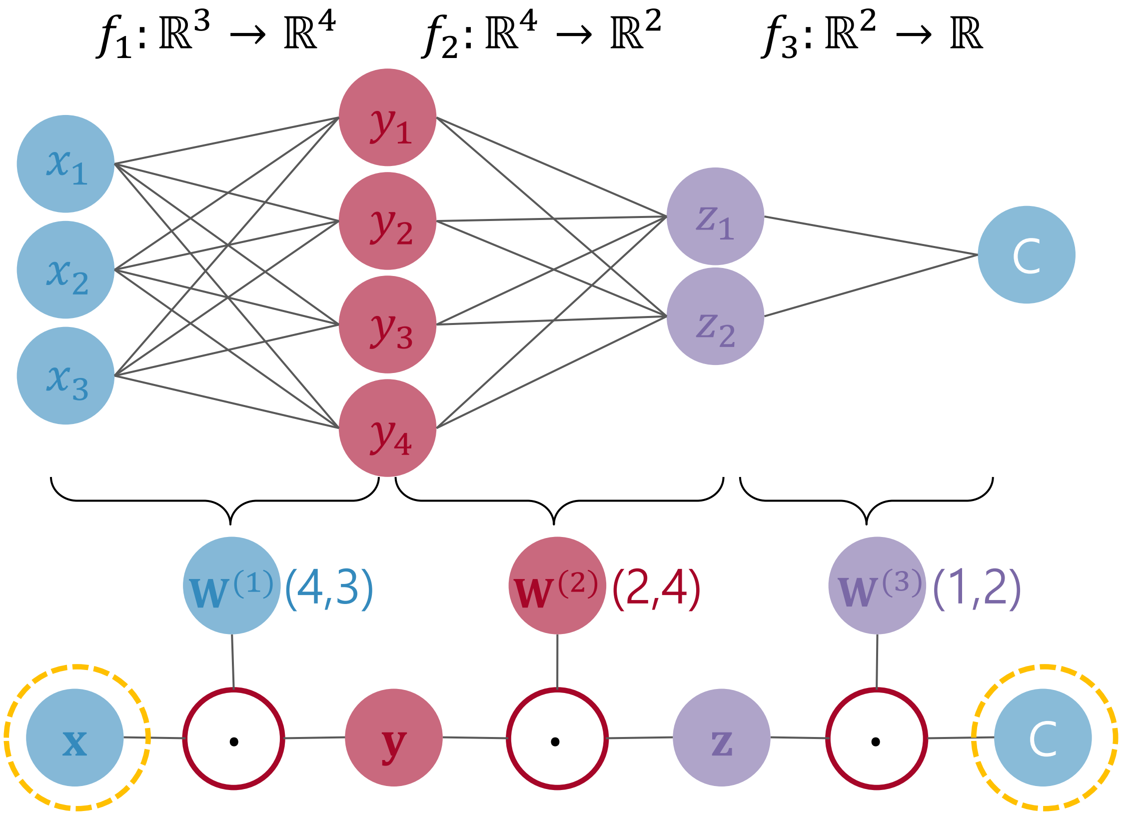
$$= \begin{bmatrix} w_{11}^{(2)} & w_{21}^{(2)} \\ w_{12}^{(2)} & w_{22}^{(2)} \\ w_{13}^{(2)} & w_{23}^{(2)} \\ w_{14}^{(2)} & w_{24}^{(2)} \end{bmatrix} \begin{bmatrix} w_1^{(3)} \\ w_2^{(3)} \end{bmatrix} = \mathbf{W}^{(2)^{\mathrm{T}}} \frac{\partial C}{\partial \mathbf{z}}$$

$\partial\mathbf{y}/\partial\mathbf{x}$



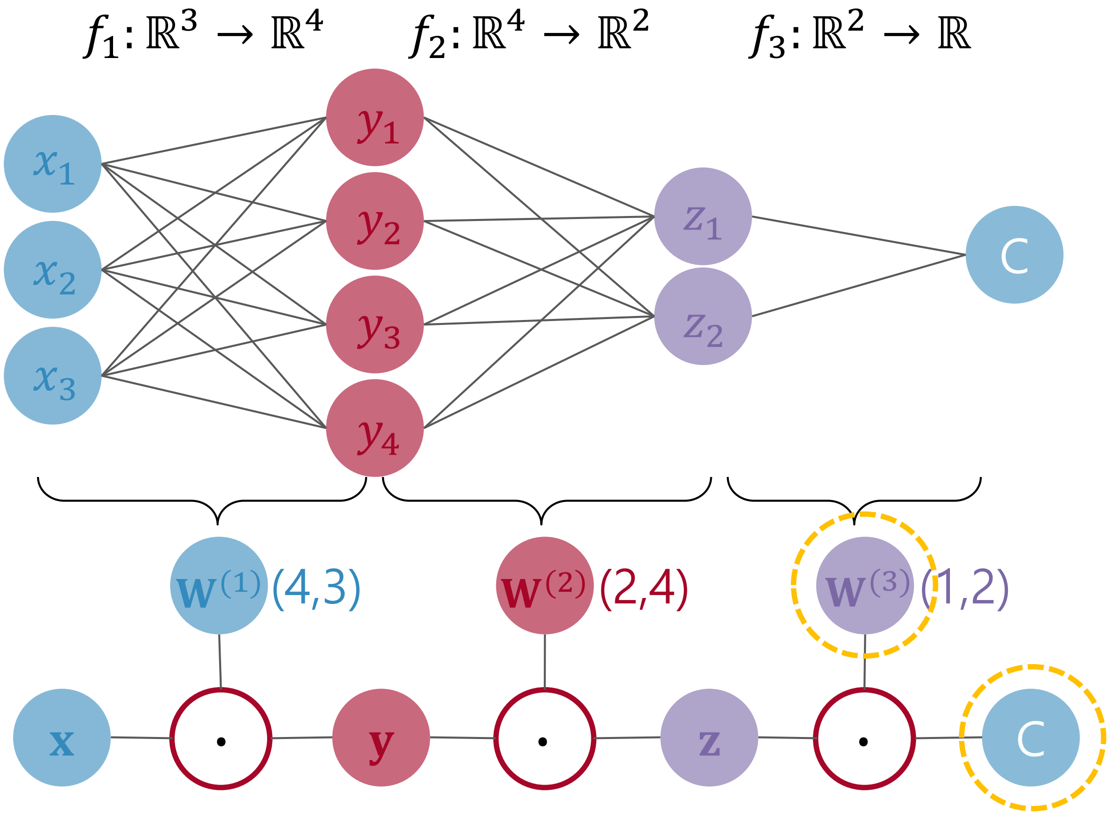
$$\frac{\partial\mathbf{y}}{\partial\mathbf{x}} = \mathbf{W}^{(1)}$$

C를 입력 $\mathbf{x}$로 미분: $\partial C/\partial \mathbf{x}$



$$\frac{\partial C}{\partial \mathbf{x}} = {\mathbf{W}^{(1)}}^{\mathrm{T}} \frac{\partial C}{\partial \mathbf{y}}$$

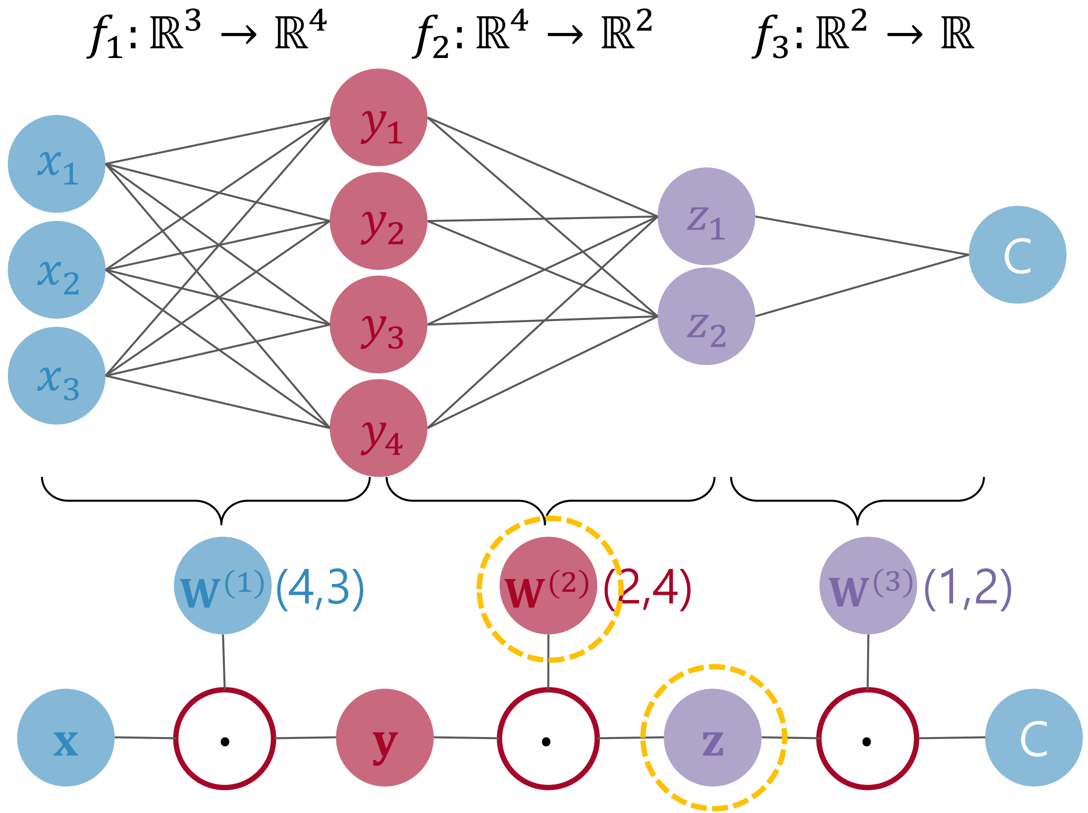
$\partial C / \partial \mathbf{W}^{(3)}$



$$C = w_1^{(3)} z_1 + w_2^{(3)} z_2$$

$$\frac{\partial C}{\partial \mathbf{W}^{(3)}} = [z_1 \quad z_2] = \mathbf{z}^{\mathrm{T}}$$

$\partial \mathbf{z} / \partial \mathbf{W}^{(2)}$



$$\underset{(2,4)}{\overset{(2,1)}{\frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(2)}}}} = ?$$

$\partial \mathbf{z} / \partial \mathbf{W}^{(2)}$

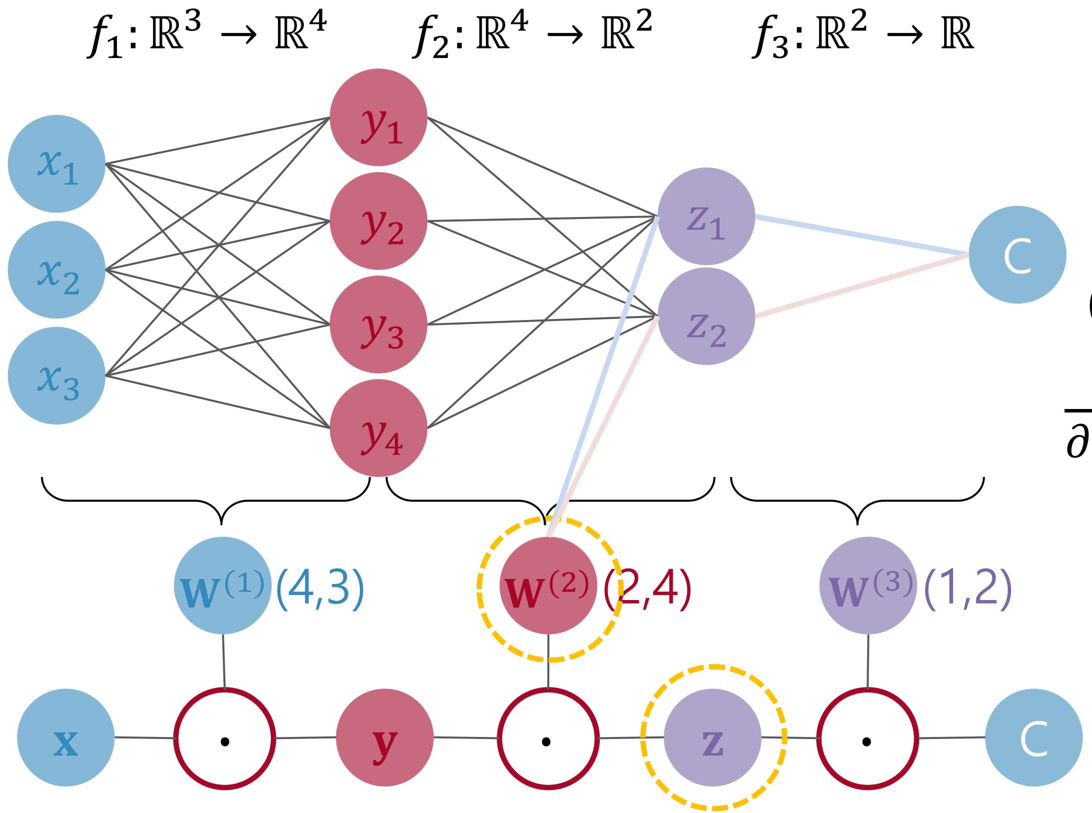
$$\frac{\partial z_2}{\partial \mathbf{W}^{(2)}} = \begin{bmatrix} \frac{\partial z_2}{\partial W_{11}^{(2)}} & \frac{\partial z_2}{\partial W_{12}^{(2)}} & \frac{\partial z_2}{\partial W_{13}^{(2)}} & \frac{\partial z_2}{\partial W_{14}^{(2)}} \\ \frac{\partial z_2}{\partial W_{21}^{(2)}} & \frac{\partial z_2}{\partial W_{22}^{(2)}} & \frac{\partial z_2}{\partial W_{23}^{(2)}} & \frac{\partial z_2}{\partial W_{24}^{(2)}} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ y_1 & y_2 & y_3 & y_4 \end{bmatrix}$$

$$\frac{\partial z_1}{\partial \mathbf{W}^{(2)}} = \begin{bmatrix} \frac{\partial z_1}{\partial W_{11}^{(2)}} & \frac{\partial z_1}{\partial W_{12}^{(2)}} & \frac{\partial z_1}{\partial W_{13}^{(2)}} & \frac{\partial z_1}{\partial W_{14}^{(2)}} \\ \frac{\partial z_1}{\partial W_{21}^{(2)}} & \frac{\partial z_1}{\partial W_{22}^{(2)}} & \frac{\partial z_1}{\partial W_{23}^{(2)}} & \frac{\partial z_1}{\partial W_{24}^{(2)}} \end{bmatrix} = \begin{bmatrix} y_1 & y_2 & y_3 & y_4 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(2)}} = \left(\frac{\partial z_1}{\partial \mathbf{W}^{(2)}}, \frac{\partial z_2}{\partial \mathbf{W}^{(2)}} \right)$$

2 2 4

$\partial \mathbf{z} / \partial \mathbf{W}^{(2)}$



$$\overset{(2,4)}{\frac{\partial C}{\partial \mathbf{W}^{(2)}}} = \overset{(2,4)}{\boxed{\frac{\partial z_1}{\partial \mathbf{W}^{(2)}}}} \overset{\text{scalar}}{\underset{\downarrow\, w_1^{(3)}}{\frac{\partial C}{\partial z_1}}} + \overset{(2,4)}{\boxed{\frac{\partial z_2}{\partial \mathbf{W}^{(2)}}}} \overset{\text{scalar}}{\underset{\downarrow\, w_2^{(3)}}{\frac{\partial C}{\partial z_2}}}$$

$\partial \mathbf{z}/\partial \mathbf{W}^{(2)}$

$$\frac{\partial z_2}{\partial \mathbf{W}^{(2)}} = \begin{bmatrix} \frac{\partial z_2}{\partial W_{11}^{(2)}} & \frac{\partial z_2}{\partial W_{12}^{(2)}} & \frac{\partial z_2}{\partial W_{13}^{(2)}} & \frac{\partial z_2}{\partial W_{14}^{(2)}} \\ \frac{\partial z_2}{\partial W_{21}^{(2)}} & \frac{\partial z_2}{\partial W_{22}^{(2)}} & \frac{\partial z_2}{\partial W_{23}^{(2)}} & \frac{\partial z_2}{\partial W_{24}^{(2)}} \end{bmatrix} \times \frac{\partial C}{\partial z_2}$$

$+$

$$\frac{\partial z_1}{\partial \mathbf{W}^{(2)}} = \begin{bmatrix} \frac{\partial z_1}{\partial W_{11}^{(2)}} & \frac{\partial z_1}{\partial W_{12}^{(2)}} & \frac{\partial z_1}{\partial W_{13}^{(2)}} & \frac{\partial z_1}{\partial W_{14}^{(2)}} \\ \frac{\partial z_1}{\partial W_{21}^{(2)}} & \frac{\partial z_1}{\partial W_{22}^{(2)}} & \frac{\partial z_1}{\partial W_{23}^{(2)}} & \frac{\partial z_1}{\partial W_{24}^{(2)}} \end{bmatrix} \times \frac{\partial C}{\partial z_1}$$

$$\frac{\partial C}{\partial \mathbf{W}^{(2)}} = \boxed{\frac{\partial z_1}{\partial \mathbf{W}^{(2)}}} \frac{\partial C}{\partial z_1} + \boxed{\frac{\partial z_2}{\partial \mathbf{W}^{(2)}}} \frac{\partial C}{\partial z_2}$$

Multiplying transposed Jacobian by gradients? 😟

Flattering tensor into a vector

- **Deep Learning Book, Ian Goodfellow, Yoshua Bengio, Aaron Courville, page 206**

We could imagine **flattering each tensor into a vector** before we run back-propagation, computing a vector-valued gradient, and then **reshaping the gradient back into a tensor**.

```
T = np.arange(16).reshape(2,4,2)

array([[[ 0,  1],
        [ 2,  3],
        [ 4,  5],
        [ 6,  7]],

       [[ 8,  9],
        [10, 11],
        [12, 13],
        [14, 15]]])
```

flattering

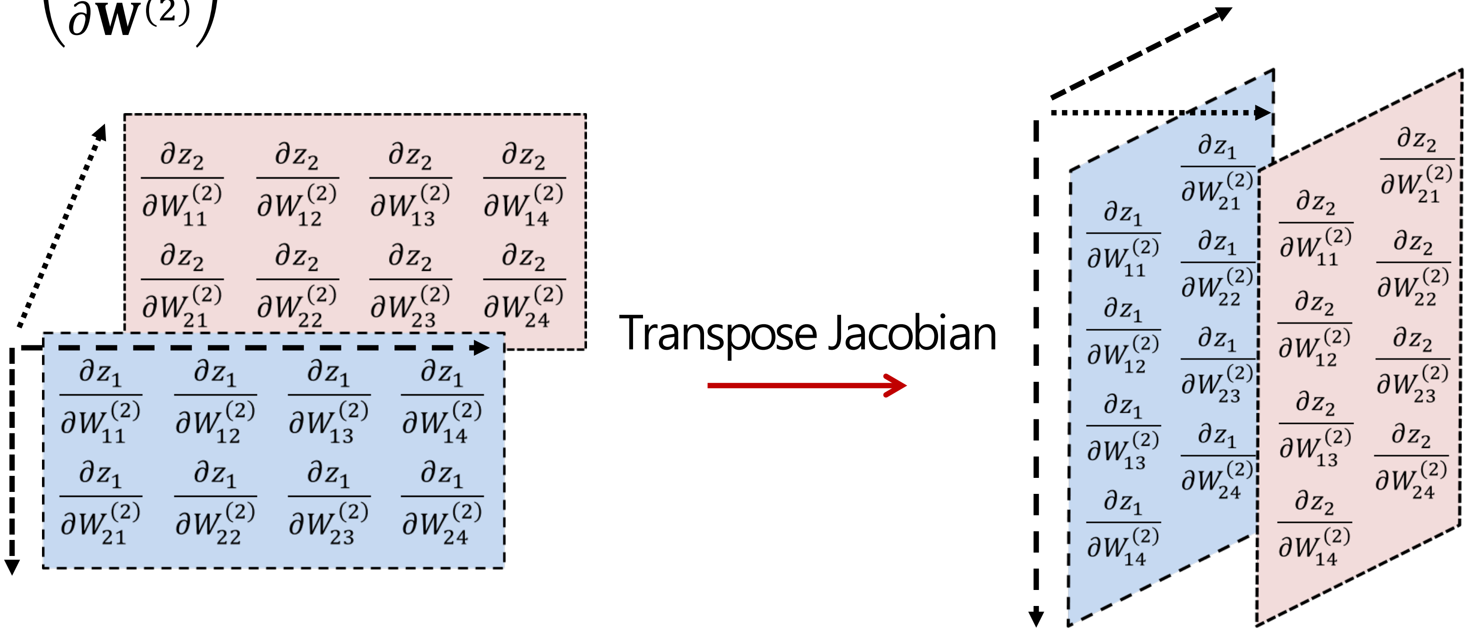
```
T.reshape(-1,2)

array([[ 0,  1],
       [ 2,  3],
       [ 4,  5],
       [ 6,  7],
       [ 8,  9],
       [10, 11],
       [12, 13],
       [14, 15]])
```

back into a tensor

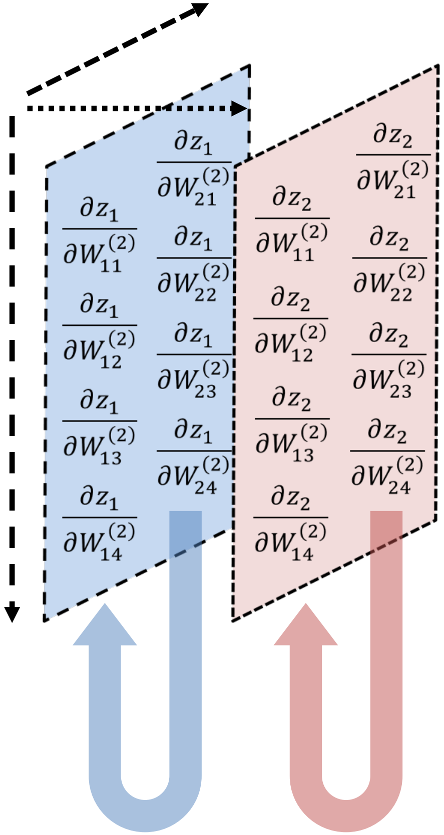
```
T.reshape(-1,2).reshape(2,4,2)

array([[[ 0,  1],
        [ 2,  3],
        [ 4,  5],
        [ 6,  7]],

       [[ 8,  9],
        [10, 11],
        [12, 13],
        [14, 15]]])
```

Step 1 : Transpose Jacobian

$$\left(\frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(2)}}\right)^{\mathrm{T}}$$



Step 2 : Flattering transpose jacobian



Fattening each tensor into a vector

$$\begin{bmatrix} \frac{\partial z_1}{\partial W_{11}^{(2)}} \\ \frac{\partial z_1}{\partial W_{12}^{(2)}} \\ \frac{\partial z_1}{\partial W_{13}^{(2)}} \\ \frac{\partial z_1}{\partial W_{14}^{(2)}} \\ \frac{\partial z_1}{\partial W_{21}^{(2)}} \\ \frac{\partial z_1}{\partial W_{22}^{(2)}} \\ \frac{\partial z_1}{\partial W_{23}^{(2)}} \\ \frac{\partial z_1}{\partial W_{24}^{(2)}} \end{bmatrix} \quad \begin{bmatrix} \frac{\partial z_2}{\partial W_{11}^{(2)}} \\ \frac{\partial z_2}{\partial W_{12}^{(2)}} \\ \frac{\partial z_2}{\partial W_{13}^{(2)}} \\ \frac{\partial z_2}{\partial W_{14}^{(2)}} \\ \frac{\partial z_2}{\partial W_{21}^{(2)}} \\ \frac{\partial z_2}{\partial W_{22}^{(2)}} \\ \frac{\partial z_2}{\partial W_{23}^{(2)}} \\ \frac{\partial z_2}{\partial W_{24}^{(2)}} \end{bmatrix}$$

Step 3 : Matrix multiply

$$\begin{bmatrix} \frac{\partial z_1}{\partial W_{11}^{(2)}} & \frac{\partial z_2}{\partial W_{11}^{(2)}} \\ \frac{\partial z_1}{\partial W_{12}^{(2)}} & \frac{\partial z_2}{\partial W_{12}^{(2)}} \\ \frac{\partial z_1}{\partial W_{13}^{(2)}} & \frac{\partial z_2}{\partial W_{13}^{(2)}} \\ \frac{\partial z_1}{\partial W_{14}^{(2)}} & \frac{\partial z_2}{\partial W_{14}^{(2)}} \\ \frac{\partial z_1}{\partial W_{21}^{(2)}} & \frac{\partial z_2}{\partial W_{21}^{(2)}} \\ \frac{\partial z_1}{\partial W_{22}^{(2)}} & \frac{\partial z_2}{\partial W_{22}^{(2)}} \\ \frac{\partial z_1}{\partial W_{23}^{(2)}} & \frac{\partial z_2}{\partial W_{23}^{(2)}} \\ \frac{\partial z_1}{\partial W_{24}^{(2)}} & \frac{\partial z_2}{\partial W_{24}^{(2)}} \end{bmatrix} \times \begin{bmatrix} \frac{\partial C}{\partial z_1} \\ \frac{\partial C}{\partial z_2} \end{bmatrix} = \begin{bmatrix} \frac{\partial z_1}{\partial W_{11}^{(2)}}\frac{\partial C}{\partial z_1} + \frac{\partial z_2}{\partial W_{11}^{(2)}}\frac{\partial C}{\partial z_2} \\ \frac{\partial z_1}{\partial W_{12}^{(2)}}\frac{\partial C}{\partial z_1} + \frac{\partial z_2}{\partial W_{12}^{(2)}}\frac{\partial C}{\partial z_2} \\ \frac{\partial z_1}{\partial W_{13}^{(2)}}\frac{\partial C}{\partial z_1} + \frac{\partial z_2}{\partial W_{13}^{(2)}}\frac{\partial C}{\partial z_2} \\ \frac{\partial z_1}{\partial W_{14}^{(2)}}\frac{\partial C}{\partial z_1} + \frac{\partial z_2}{\partial W_{14}^{(2)}}\frac{\partial C}{\partial z_2} \\ \frac{\partial z_1}{\partial W_{21}^{(2)}}\frac{\partial C}{\partial z_1} + \frac{\partial z_2}{\partial W_{21}^{(2)}}\frac{\partial C}{\partial z_2} \\ \frac{\partial z_1}{\partial W_{22}^{(2)}}\frac{\partial C}{\partial z_1} + \frac{\partial z_2}{\partial W_{22}^{(2)}}\frac{\partial C}{\partial z_2} \\ \frac{\partial z_1}{\partial W_{23}^{(2)}}\frac{\partial C}{\partial z_1} + \frac{\partial z_2}{\partial W_{23}^{(2)}}\frac{\partial C}{\partial z_2} \\ \frac{\partial z_1}{\partial W_{24}^{(2)}}\frac{\partial C}{\partial z_1} + \frac{\partial z_2}{\partial W_{24}^{(2)}}\frac{\partial C}{\partial z_2} \end{bmatrix}$$

Multiplying transposed Jacobian by gradients! ☺

Step 4 : Reshape

$$\begin{bmatrix} \frac{\partial z_1}{\partial W_{11}^{(2)}}\frac{\partial C}{\partial z_1}+\frac{\partial z_2}{\partial W_{11}^{(2)}}\frac{\partial C}{\partial z_2} \\ \frac{\partial z_1}{\partial W_{12}^{(2)}}\frac{\partial C}{\partial z_1}+\frac{\partial z_2}{\partial W_{12}^{(2)}}\frac{\partial C}{\partial z_2} \\ \frac{\partial z_1}{\partial W_{13}^{(2)}}\frac{\partial C}{\partial z_1}+\frac{\partial z_2}{\partial W_{13}^{(2)}}\frac{\partial C}{\partial z_2} \\ \frac{\partial z_1}{\partial W_{14}^{(2)}}\frac{\partial C}{\partial z_1}+\frac{\partial z_2}{\partial W_{14}^{(2)}}\frac{\partial C}{\partial z_2} \\ \frac{\partial z_1}{\partial W_{21}^{(2)}}\frac{\partial C}{\partial z_1}+\frac{\partial z_2}{\partial W_{21}^{(2)}}\frac{\partial C}{\partial z_2} \\ \frac{\partial z_1}{\partial W_{22}^{(2)}}\frac{\partial C}{\partial z_1}+\frac{\partial z_2}{\partial W_{22}^{(2)}}\frac{\partial C}{\partial z_2} \\ \frac{\partial z_1}{\partial W_{23}^{(2)}}\frac{\partial C}{\partial z_1}+\frac{\partial z_2}{\partial W_{23}^{(2)}}\frac{\partial C}{\partial z_2} \\ \frac{\partial z_1}{\partial W_{24}^{(2)}}\frac{\partial C}{\partial z_1}+\frac{\partial z_2}{\partial W_{24}^{(2)}}\frac{\partial C}{\partial z_2} \end{bmatrix}$$

$$\begin{array}{c|c|c|c} \frac{\partial z_1}{\partial W_{11}^{(2)}}\frac{\partial C}{\partial z_1}+\cancel{\frac{\partial z_2}{\partial W_{11}^{(2)}}\frac{\partial C}{\partial z_2}}^{0} & \frac{\partial z_1}{\partial W_{12}^{(2)}}\frac{\partial C}{\partial z_1}+\cancel{\frac{\partial z_2}{\partial W_{12}^{(2)}}\frac{\partial C}{\partial z_2}}^{0} & \frac{\partial z_1}{\partial W_{13}^{(2)}}\frac{\partial C}{\partial z_1}+\cancel{\frac{\partial z_2}{\partial W_{13}^{(2)}}\frac{\partial C}{\partial z_2}}^{0} & \frac{\partial z_1}{\partial W_{14}^{(2)}}\frac{\partial C}{\partial z_1}+\cancel{\frac{\partial z_2}{\partial W_{14}^{(2)}}\frac{\partial C}{\partial z_2}}^{0} \\ \hline \cancel{\frac{\partial z_1}{\partial W_{21}^{(2)}}\frac{\partial C}{\partial z_1}}^{0}+\frac{\partial z_2}{\partial W_{21}^{(2)}}\frac{\partial C}{\partial z_2} & \cancel{\frac{\partial z_1}{\partial W_{22}^{(2)}}\frac{\partial C}{\partial z_1}}^{0}+\frac{\partial z_2}{\partial W_{22}^{(2)}}\frac{\partial C}{\partial z_2} & \cancel{\frac{\partial z_1}{\partial W_{23}^{(2)}}\frac{\partial C}{\partial z_1}}^{0}+\frac{\partial z_2}{\partial W_{23}^{(2)}}\frac{\partial C}{\partial z_2} & \cancel{\frac{\partial z_1}{\partial W_{24}^{(2)}}\frac{\partial C}{\partial z_1}}^{0}+\frac{\partial z_2}{\partial W_{24}^{(2)}}\frac{\partial C}{\partial z_2} \end{array}$$

Reshape

$$\frac{\partial C}{\partial \mathbf{W}^{(2)}} = \begin{bmatrix} y_1\frac{\partial C}{\partial z_1} & y_2\frac{\partial C}{\partial z_1} & y_3\frac{\partial C}{\partial z_1} & y_4\frac{\partial C}{\partial z_1} \\ y_1\frac{\partial C}{\partial z_2} & y_2\frac{\partial C}{\partial z_2} & y_3\frac{\partial C}{\partial z_2} & y_4\frac{\partial C}{\partial z_2} \end{bmatrix} \xrightarrow{\text{Index form}} \frac{\partial C}{\partial W_{ij}^{(2)}} = y_j\frac{\partial C}{\partial z_i}$$

$$\frac{\partial C}{\partial \mathbf{W}^{(2)}} = \frac{\partial C}{\partial \mathbf{z}}\mathbf{y}^{\mathrm{T}} \qquad \frac{\partial C}{\partial \mathbf{W}^{(1)}} = \frac{\partial C}{\partial \mathbf{y}}\mathbf{x}^{\mathrm{T}}$$