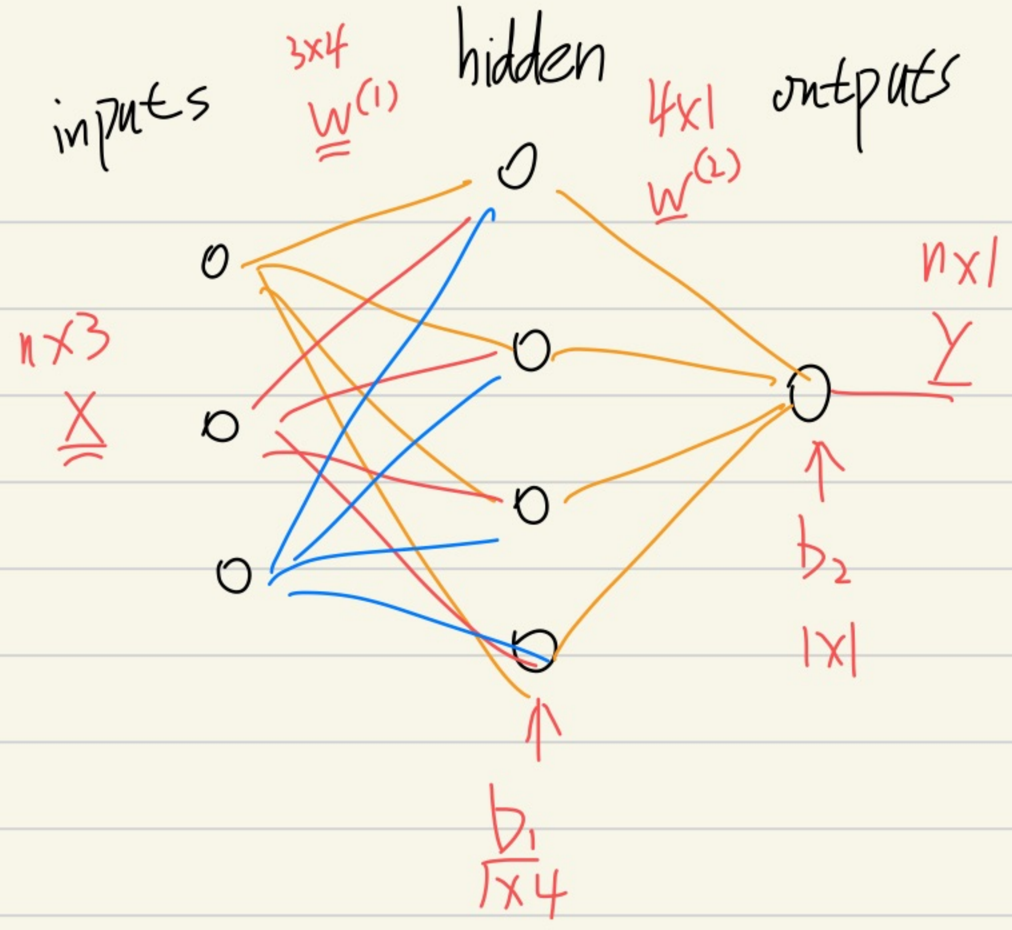


$$\underset{n\times4}{X \cdot W^{(1)}} + \underset{n\times4}{b_1} = \underset{n\times4}{S}$$

$$\underset{n\times4}{S} \cdot \underset{4\times1}{W^{(2)}} + \underset{n\times1}{b_2} = \underset{n\times1}{\hat{y}}$$

$$b_1 = \underbrace{\begin{bmatrix} b_1 \\ \vdots \\ b_1 \end{bmatrix}}_{n} \qquad b_2 = \underbrace{\begin{bmatrix} b_2 \\ \vdots \\ b_2 \end{bmatrix}}_{n}$$

$$\text{scalar } loss = MSE = \frac{\sum_i^n (\hat{y}_i - y_i)^2}{n} = \|\hat{y} - y\|_2^2 \cdot \frac{1}{n} = (\hat{y} - y)^T(\hat{y} - y) \cdot \frac{1}{n}$$

$$\left.\begin{aligned} &1\times1 \quad \text{gradient } \hat{y}_n = \nabla_{\hat{y}_n} loss = \frac{2}{n}(\hat{y}_n - y_n) \\ &1\times1 \quad \text{gradient } y_n = \nabla_{y_n} loss = -\frac{2}{n}(\hat{y}_n - y_n) \end{aligned}\right\} \text{single row}$$

①

$$\boxed{\begin{aligned} \Rightarrow \ & \underset{n\times1}{\text{gradient}}\ \hat{y} = \nabla_{\hat{y}} loss = \nabla_{\hat{y}} \frac{\hat{y}^T\hat{y} - \hat{y}^T y - y^T\hat{y} + y^T y}{n} = \left(\frac{\hat{y}}{n} + \frac{\hat{y}}{n}\right) - \frac{y}{n} - \frac{y}{n} + 0 \\ & = -\frac{2}{n}(y - \hat{y}) \\ & \underset{n\times1}{\text{gradient}}\ y = \nabla_y loss = \frac{-\hat{y} - \hat{y} + y + y}{n} = \frac{2}{n}(y - \hat{y}) \end{aligned}}$$

for each row in $\underline{\hat{y}}$:

$$\begin{cases} s_{11}\cdot w_1^{(2)} + s_{12}\cdot w_2^{(2)} + s_{13}\cdot w_3^{(2)} + s_{14}\cdot w_4^{(2)} + b_2 = \hat{y}_1 \\ \vdots \\ s_{n1}\cdot w_1^{(2)} + s_{n2}\cdot w_2^{(2)} + s_{n3}\cdot w_3^{(2)} + s_{n4}\cdot w_4^{(2)} + b_2 = \hat{y}_n \end{cases}$$

$$= \begin{cases} \overset{1\times4}{\underline{s}_1} \cdot \overset{4\times1}{\underline{w}^{(2)}} + b_2 = \hat{y}_1 \\ \vdots \\ \underline{s}_n \cdot \underline{w}^{(2)} + b_2 = \hat{y}_n \end{cases}$$

2

$$\left.\begin{aligned} &1\times4 \quad \text{gradient}_{\,\underline{s}_n} = \text{gradient}\ \hat{y}_n \cdot \nabla_{\underline{s}_n} \hat{y}_n = \overset{1\times1}{\tfrac{2}{n}(\hat{y}_n - y_n)} \cdot \overset{1\times4}{\underline{w}^{(2)T}} \\ &4\times1 \quad \text{gradient}_{\,\underline{w}^{(2)}} = \nabla_{\underline{w}^{(2)}} \hat{y}_n \cdot \text{gradient}\ \hat{y}_n = \overset{4\times1}{\underline{s}_1^T} \cdot \overset{1\times1}{\tfrac{2}{n}(\hat{y}_n - y_n)} \\ &1\times1 \quad \text{gradient}\ b_2 = \nabla_{b_2} \hat{y}_n \cdot \text{gradient}\ \hat{y}_n = 1 \cdot \tfrac{2}{n}(\hat{y}_n - y_n) \end{aligned}\right\} \text{single row}$$

$$\Rightarrow \overset{n\times4}{\text{gradient}_{\,\underline{\underline{S}}}} = \begin{bmatrix} \text{gradient}\ \underline{s}_1 \\ \vdots \\ \text{gradient}\ \underline{s}_n \end{bmatrix} = \frac{2}{n}\begin{bmatrix} \hat{y}_1 - y_1 \\ \vdots \\ \hat{y}_n - y_n \end{bmatrix} \cdot \underline{w}^{(2)T} = -\frac{2}{n}(\underline{y}-\underline{\hat{y}}) \cdot \underline{w}^{(2)T} = \overset{n\times1}{\text{gradient}\ \underline{\hat{y}}} \cdot \overset{1\times4}{\underline{w}^{(2)T}}$$

$$\overset{4\times1}{\text{gradient}\ \underline{w}^{(2)}} = \nabla_{\underline{w}^{(2)}} \underline{\hat{y}} \cdot \text{gradient}\ \underline{\hat{y}} = \begin{bmatrix} \underline{s}_1^T \\ \vdots \\ \underline{s}_n^T \end{bmatrix} \cdot -\frac{2}{n}(\underline{y}-\underline{\hat{y}}) = \overset{4\times n}{\underline{\underline{S}}^T} \cdot \overset{n\times1}{\frac{-2}{n}(\underline{y}-\underline{\hat{y}})}$$

$$\overset{1\times1}{\text{gradient}\ b_2} = \nabla_{b_2} \underline{\hat{y}} \cdot \text{gradient}\ \underline{\hat{y}} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \cdot -\frac{2}{n}(\underline{y}-\underline{\hat{y}}) = \overset{\text{scalar}}{-\frac{2}{n} \cdot \sum_{i=1}^{n}(y_n - \hat{y}_n)}$$

3

$$\overset{n\times3}{\underline{\underline{X}}} \cdot \overset{3\times4}{\underline{\underline{W}}^{(1)}} + \overset{n\times4}{\underline{b}_1} = \overset{n\times4}{\underline{\underline{S}}} \quad \begin{cases} \overset{1\times3}{\underline{x}_1}\ \overset{3\times4}{\underline{\underline{W}}^{(1)}} + \overset{1\times4}{\underline{b}_1} = \overset{1\times4}{\underline{S}_1} \\ \vdots \\ \underline{x}_n\ \underline{\underline{W}}^{(1)} + \underline{b}_1 = \underline{S}_n \end{cases}$$

$$\text{gradient } x_{ij} = \sum_{k=1}^{4} \text{gradient } S_{ik} \cdot \nabla_{x_{ij}} S_{ik} = \sum_{k=1}^{4} -\frac{2}{n}(y_i - \hat{y}_i) \cdot w_k^{(2)} \cdot w_{jk}^{(1)}$$

$$= -\frac{2}{n}(y_i - \hat{y}_i) \cdot \overset{1\times4}{\underline{w}^{(2)T}} \cdot \overset{4\times1}{\underline{w}_j^{(1)T}}$$

$$\Rightarrow \text{gradient } \overset{1\times3}{\underline{x}_n} = -\frac{2}{n}(y_i - \hat{y}_i) \cdot \overset{1\times4}{\underline{w}^{(2)T}} \cdot \overset{4\times3}{\underline{\underline{W}}^{(1)T}} = \text{gradient } \underline{S}_n \cdot \underline{\underline{W}}^{(1)T}$$

$$\boxed{\Rightarrow \text{gradient } \underset{n\times3}{\underline{\underline{X}}} = -\frac{2}{n}\underset{n\times1}{(y - \hat{y})} \cdot \underset{1\times4}{\underline{w}^{(2)T}} \cdot \underset{4\times3}{\underline{\underline{W}}^{(1)T}} = \text{gradient } \underline{\underline{S}} \cdot \underline{\underline{W}}^{(1)T}}$$

$$\text{gradient } w_{ij}^{(1)} = \sum_{k=1}^{n} \text{gradient } S_{kj} \cdot \nabla_{w_{ij}} S_{kj} = \sum_{k=1}^{n} -\frac{2}{n}(y_k - \hat{y}_k) \cdot w_j^{(2)} \cdot x_{ki}$$

$$= -\frac{2}{n} w_j^{(2)} \cdot \underset{1\times n}{\begin{bmatrix} x_{1i} \\ \vdots \\ x_{ni} \end{bmatrix}^T} \cdot \underset{n\times1}{(y - \hat{y})}$$

$$\Rightarrow \text{gradient } \overset{1\times4}{\underline{w}_n^{(1)}} = -\frac{2}{n}\ \underset{1\times4}{\underline{w}^{(2)T}} \cdot \underbrace{\begin{bmatrix} x_{1i} \\ \vdots \\ x_{ni} \end{bmatrix}^T (y - \hat{y})}_{\text{scalar}}$$

$$\boxed{\Rightarrow \text{gradient } \underset{3\times4}{\underline{\underline{W}}^{(1)}} = -\frac{2}{n} \underset{3\times n}{\begin{bmatrix} x_{11} & \cdots & x_{n1} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ x_{13} & \cdots & x_{n3} \end{bmatrix}} \cdot \underset{n\times1}{(y - \hat{y})} \cdot \underset{1\times4}{\underline{w}^{(2)T}} = -\frac{2}{n} \underset{3\times n}{\underline{\underline{X}}^T} \cdot \underset{n\times1}{(y-\hat{y})} \cdot \underset{1\times4}{\underline{w}^{(2)T}} = \underline{\underline{X}}^T \cdot \text{gradient } \underline{\underline{S}}}$$

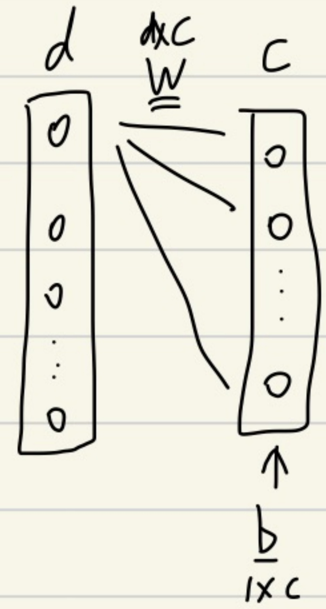
$$\text{gradient } b_{1i} = \sum_{k=1}^{n} \text{gradient } s_{ki} \cdot \nabla_{b_{1i}} s_{ki} = \sum_{k=1}^{n} -\frac{2}{n}(y_k - \hat{y}_k)\cdot w_i^{(2)} \cdot 1$$

$$= -\frac{2}{n}\cdot \sum_{k=1}^{n}(y_k - \hat{y}_k)\cdot w_i^{(2)}$$

$$\Rightarrow \boxed{\text{gradient } \underset{1\times 4}{\underline{b}_1} = -\frac{2}{n}\cdot \sum_{k=1}^{n}(y_k - \hat{y}_k)\cdot \underset{1\times 4}{\underline{w}^{(2)^T}}}$$

Brief Summary

for linear node: $\underset{n\times d}{\overset{}{\underline{\underline{X}}}} \cdot \underset{d\times C}{\underline{\underline{W}}} + \underline{b} = \underset{n\times C}{\underline{\underline{S}}}$



$$\text{gradient } \underset{n\times d}{\underline{\underline{X}}} = \text{gradient } \underset{n\times C}{\underline{\underline{S}}} \cdot \underset{C\times d}{W^T}$$

$$\text{gradient } \underset{d\times C}{\underline{\underline{W}}} = \underset{d\times n}{\underline{\underline{X}}^T} \cdot \text{gradient } \underset{n\times C}{\underline{\underline{S}}}$$

$$\text{gradient } \underline{b} = \text{sum}(\text{gradient } \underline{\underline{S}}, \text{axis}=0)$$

for activation Node: $\sigma(\underset{m\times n}{\underline{\underline{X}}}) = \underset{m\times n}{\underline{\underline{Y}}}$

Sigmoid $\sigma(x) = \frac{1}{1+e^{-x}}$ $\quad$ gradient $\sigma(x)$ = gradient $\underline{\underline{Y}}\ \sigma(1-\sigma)$

Tanh $\sigma(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ $\quad$ gradient $\sigma(x)$ = gradient $\underline{\underline{Y}} \cdot (1-\tanh^2(x))$

for Add Node:

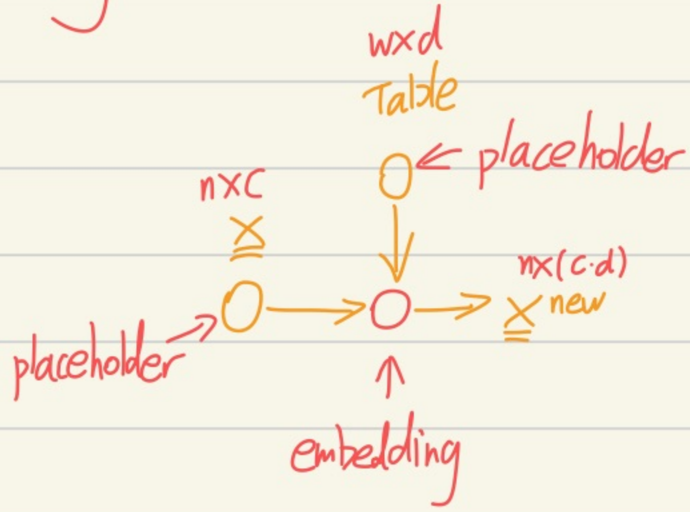
$$\overset{n\times d}{X} + \overset{n\times d}{Y} + \overset{n\times d}{b} = \overset{n\times d}{Z}$$

$$\overset{n\times d}{b} = \left.\begin{bmatrix} \overset{1\times d}{b} \\ \vdots \\ b \end{bmatrix}\right\} n$$

$$\text{gradient } X = \text{gradient } Y = \text{gradient } Z$$

$$\text{gradient } \underset{1\times d}{b} = \text{sum}(\text{gradient } \underset{n\times d}{Z}, \text{axis}=0)$$

for Embedding case in NLP



$\overset{n\times C}{X}$: $\boxed{n}$ sentences, $\boxed{C}$ words in each

$\overset{w\times d}{\text{Table}}$: Total $\boxed{w}$ instinct words in all sentences, word embedding $\boxed{d}$ dimension

In embedding node:

① change $\overset{n\times C}{X} \longrightarrow \overset{(n\cdot C)\times W}{x}$ [one hot transformation]

② $\overset{nc\times w}{x} \cdot \overset{w\times d}{\text{Table}} = \overset{nc\times d}{R} \xrightarrow[\text{reshape}]{} \overset{n\times(c\cdot d)}{X^{new}}$

$$\underset{}{\text{gradient }\overset{(nc)\times d}{\underline{\underline{R}}}} = \text{reshape gradient } \overset{n\times(c\cdot d)}{\underline{\underline{x}}^{new}} \text{ to } nc\times d$$

$$\underset{w\times d}{\text{gradient Table}} = \underset{w\times(nc)}{\underline{\underline{X}}^{T}} \cdot \underset{(nc)\times d}{\text{gradient } \underline{\underline{R}}}$$

$$\underset{(nc)\times w}{\text{gradient } \underline{\underline{X}}} = \underset{(nc)\times d}{\text{gradient } \underline{\underline{R}}} \cdot \underset{d\times w}{\text{Table}^{T}}$$

$$\underset{n\times c}{\text{gradient } \underline{\underline{x}}} = \text{reshape } \underset{nc\times 1}{\text{sum}} \big(\underset{nc\times w}{\text{gradient } \underline{\underline{X}}},\ \text{axis}=1 \big) \text{ to } n\times c$$