Softmax

given outputs from prev. layers $= \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \end{bmatrix}$

for every z_i,

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}.$$

$$= \frac{e^{z_i - \max(z)}}{\sum_{j=1}^{n} e^{z_j - \max(z)}}$$

[to stop overflow]

backward pass ⇒

$$\underset{Z}{\begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \end{bmatrix}} \xRightarrow[\text{softmax}]{} \underset{S}{\begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ \vdots \end{bmatrix}}$$

$$\frac{\partial s_i}{\partial z_j} = \begin{cases} s_i(1-s_i) & \text{when } i=j \\ -s_i s_j & \text{when } i \neq j. \end{cases}$$

$$J_s = \text{diag}(S) - SS^T..$$

Batch normalization

forward

$$\hat{x} = \frac{x-\mu}{\sqrt{\sigma^2+\epsilon}}$$

$$Y = \gamma \cdot \hat{x} + \beta.$$

backward

$$\frac{\partial L}{\partial \gamma} = \frac{\partial L}{\partial Y} \times \frac{\partial Y}{\partial \gamma}$$

$$= \sum_{batch} \hat{x}_i \times \frac{\partial L}{\partial Y_i}$$

$$= \sum \frac{\partial L}{\partial Y_i} \hat{x}_i$$

sum because gamma has effect on all the samples in the batch.

$$\frac{\partial L}{\partial \gamma} = \sum_{i=1}^{m} \frac{\partial L}{\partial Y_i} \hat{x}_i \,.$$

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{m} \frac{\partial L}{\partial Y_i} \qquad [m = \text{batch size}]$$

for a sample,

$$\frac{\partial L}{\partial \hat{x}} = \frac{\partial L}{\partial Y} \cdot \gamma .$$

$$\frac{\partial L}{\partial \sigma^2} = \frac{\partial L}{\partial \gamma} \times \frac{\partial \gamma}{\partial \hat{x}} \times \frac{\partial \hat{x}}{\partial \sigma^2}$$

$$= \frac{\partial L}{\partial \gamma} \times \frac{\partial \gamma}{\partial \hat{x}} \times (x-\mu) \times -\frac{1}{2} \times \left(\sqrt{\sigma^2+\epsilon}\right)^{-3/2}$$

for many samples $\Rightarrow$

$$\frac{\partial L}{\partial \sigma^2} = \sum_{i=1}^{m} \frac{\partial L}{\partial \gamma_i} \times \frac{\partial \gamma_i}{\partial \hat{x}_i} \times (x_i-\mu) \times -\frac{1}{2} \times \left(\sqrt{\sigma^2+\epsilon}\right)^{-3/2}$$

* same way (chain rule and partial derivative),

we can find, $\frac{\partial L}{\partial x}$ and $\frac{\partial L}{\partial \mu}$.

* sum over x_i, $\hat{x}$, or γ is important for γ, σ^2, β, μ because they are same over all samples.