Word2Vec
skip-gram

$$\text{minimize } J(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-m\le j\le m \\ j\neq 0}} \log P(w_{t+j}\mid w_t)$$

all parameters i.e. u_o, v_c

$$P(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w=1}^{V}\exp(u_w^T v_c)}$$

$$\frac{\partial}{\partial v_c}\log\frac{\exp(u_o^T v_c)}{\sum_{w=1}^{V}\exp(u_w^T v_c)}$$

f ($g(v_c)$

$\log\frac{x}{y} = \log x - \log y$

$$= \frac{\partial}{\partial v_c}\log\exp(u_o^T v_c) - \frac{\partial}{\partial v_c}\log\sum_{w=1}^{V}\exp(u_w^T v_c)$$

$\log(e^x) = x$

$$\frac{\partial}{\partial v_c} u_o^T v_c = u_o$$

$\frac{\partial \log x}{\partial x} = \frac{1}{x}$ ||

chain rule $f(g(x))$: $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial g}\cdot\frac{\partial g}{\partial x}$

$$\frac{1}{\sum_{w=1}^{V}\exp(u_w^T v_c)}\cdot\frac{\partial}{\partial v_c}\sum_{x=1}^{V}\exp(u_x^T v_c)$$

||

$$\frac{1}{\sum_{w=1}^{V}\exp(u_w^T v_c)}\cdot\sum_{x=1}^{V}\frac{\partial}{\partial v_c}\exp(u_x^T v_c)$$

f ($g(v_c)$

chain rule again

$\frac{\partial e^x}{\partial x} = e^x$

||

$$\frac{1}{\sum_{w=1}^{V}\exp(u_w^T v_c)}\cdot\sum_{x=1}^{V}\exp(u_x^T v_c)\frac{\partial}{\partial v_c}u_x^T v_c$$

$\frac{\partial}{\partial v_c}u_x^T v_c = u_x$

$$= u_o - \frac{\sum_{x=1}^{V}\exp(u_x^T v_c)\cdot u_x}{\sum_{w=1}^{V}\exp(u_w^T v_c)}$$

$P(x|c)$

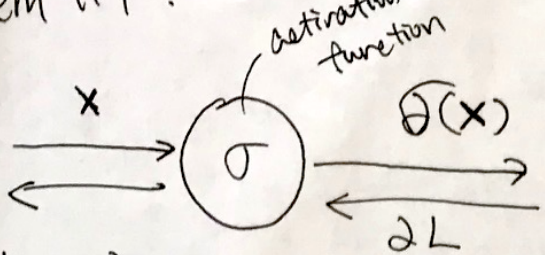
$$= u_o - \sum_{x=1}^{V}\frac{\exp(u_x^T v_c)}{\sum_{w=1}^{V}\exp(u_w^T v_c)}\cdot u_x$$

observed context word.

expectation of context word by current model

$$= u_o - \sum_{x=1}^{V} P(x|c)\cdot u_x$$

Problem #1 : e.g. tanh



$$\frac{\partial L}{\partial x} = \frac{\partial \sigma}{\partial x} \cdot \frac{\partial L}{\partial \sigma}$$

→ gradients getting close to 0 ⇒ "saturated" neurons

Problem #2 : e.g.: ReLu

not zero-centered

⇒ input to a neuron is always positive ...

$$g(wx+b) = g\left(\sum_i w_i x_i + b\right)$$

⇒ gradients on w will always be all positive or all negative

$$\frac{\partial L}{\partial w_i} = \frac{\partial L}{\partial g} \cdot \frac{\partial g}{\partial z} \cdot \frac{\partial (wx+b)}{\partial x}$$

$$= \frac{\partial L}{\partial g} \cdot \frac{\partial g}{\partial z} \cdot x_i$$