$$L(\mathbf{t}, \mathbf{s}) = -\sum_{i=1}^{n} t_i \log s_i \qquad \mathbf{s} = \text{softmax}(\underset{(n\times n^*)}{W^{(d)}}(\underset{(n^*\times n)}{W^{(e)}}\underset{(n\times 1)}{\mathbf{x}}))$$

$$W^{(e)}\mathbf{x} = \mathbf{z}$$
$$W^{(d)}\mathbf{z} = \mathbf{y}$$

center word.

"a" "fat" "cat" "sat" "on" "the" "mat"

context word.

$$\mathbf{x}_1 = a = \begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix} \quad fat = \begin{bmatrix}0\\1\\0\\\vdots\\0\end{bmatrix} \quad cat = \begin{bmatrix}0\\0\\1\\0\\\vdots\\0\end{bmatrix} \quad sat = \begin{bmatrix}0\\\vdots\\0\\1\\0\\\vdots\\0\end{bmatrix} \quad on = \begin{bmatrix}0\\\vdots\\0\\1\\0\\\vdots\\0\end{bmatrix} \quad \cdots$$

$\mathbf{X}$: 2m 개의 n차원으로 one-hot 인코딩된 context word 들의 벡터합.

$$\mathbf{X} = \begin{bmatrix}1\\1\\0\\\vdots\\0\\1\\1\\0\end{bmatrix} = a + fat + sat + on$$, m=2 인 경우.

$\mathbf{t}$: n차원으로 one-hot 인코딩된 center word 벡터.

$$\mathbf{t} = \begin{bmatrix}0\\0\\1\\0\\\vdots\\0\end{bmatrix} = cat.$$

$W^{(e)}$: n차원에서 n^* 워드임베딩 차원으로 선형변환시키는 인코딩 행렬.
$(n^* \times n)$

$$W^{(e)} = \begin{bmatrix} w^{(e)}_{11} & w^{(e)}_{12} & w^{(e)}_{13} & \cdots & w^{(e)}_{1n} \end{bmatrix}$$

$$
\begin{bmatrix}
\vdots & \ddots & \vdots \\
w^{(e)}_{n^*,1} & \cdots & w^{(e)}_{n^*,n}
\end{bmatrix}
$$

$W^{(d)}$: n^* 워드임베딩 차원에서 n차원으로 선형변환 시키는 디코딩 행렬.
$(n \times n^*)$

$$W^{(d)} = (W^{(e)})^T.$$

Softmax : 0과 1 사이 확률적으로 표현하는 방법.

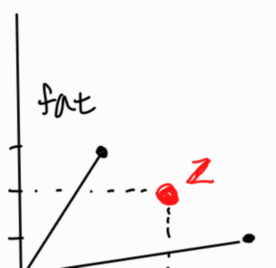
의미해석

① 워드임베딩.

$$\underset{(n^* \times n)}{W^{(e)}} \underset{(n \times 1)}{x} = \underset{(n^* \times 1)}{z}$$ ... 2m개의 context vector 들을 더한 워드임베딩.

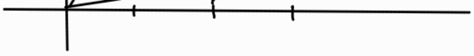
$$
\begin{bmatrix}
w^{(e)}_{11} & w^{(e)}_{12} & w^{(e)}_{13} & \cdots & w^{(e)}_{1n} \\
\vdots & & \ddots & & \vdots \\
w^{(e)}_{n^*,1} & & \cdots & & w^{(e)}_{n^*,n}
\end{bmatrix}
\begin{bmatrix} 1 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}
=
\begin{bmatrix}
w^{(e)}_{11} + w^{(e)}_{12} + w^{(e)}_{1s} + w^{(e)}_{1o} \\
\vdots \quad \vdots \quad \vdots \quad \vdots \\
\underset{\downarrow\ A}{w^{(e)}_{n^*,1}} + \underset{\downarrow\ fat}{w^{(e)}_{n^*,2}} + \underset{\downarrow\ sat}{w^{(e)}_{n^*,s}} + \underset{\downarrow\ on.}{w^{(e)}_{n^*,o}}
\end{bmatrix}
\times \underset{(2m)}{\frac{1}{4}}
=
\begin{bmatrix} z_1 \\ \vdots \\ z_{n^*} \end{bmatrix}
$$

· Context word 각각에 대응하는 워드임베딩 들을 모두 벡터 덧셈한 것.



$z \in \mathbb{R}^{n^*}$

⇒ 만약 $L(s,t)=0$ 이라고 해보자.
distributional hypothesis 에 합당함.



② 디코딩.

$$\underset{(n\times1)}{y} = \underset{(n\times n^*)}{W^{(d)}}\underset{(n^*\times1)}{z} = (W^{(e)})^T z$$

$$\begin{bmatrix} w^{(e)}_{11} & \cdots & w^{(e)}_{n^*,1} \\ w^{(e)}_{12} & & w^{(e)}_{n^*,2} \\ w^{(e)}_{13} & \ddots & w^{(e)}_{n^*,3} \\ \vdots & & \vdots \\ w^{(e)}_{1,n} & \cdots & w^{(e)}_{n^*,n} \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_{n^*} \end{bmatrix} = \begin{bmatrix} w^{(e)}_{11} & \cdots & w^{(e)}_{n^*,1} \\ w^{(e)}_{12} & & w^{(e)}_{n^*,2} \\ w^{(e)}_{13} & \ddots & w^{(e)}_{n^*,3} \\ \vdots & & \vdots \\ w^{(e)}_{1,n} & \cdots & w^{(e)}_{n^*,n} \end{bmatrix}\begin{bmatrix} w^{(e)}_{11} + w^{(e)}_{12} + w^{(e)}_{15} + w^{(e)}_{10} \\ \vdots \quad \vdots \quad \vdots \quad \vdots \\ w^{(e)}_{n^*,1} + w^{(e)}_{n^*,2} + w^{(e)}_{n^*,5} + w^{(e)}_{n^*,0} \end{bmatrix} \times \frac{1}{4}$$

(2m)

$$= y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\underset{(n\times1)}{S} = \mathrm{softmax}(\underset{(n\times1)}{y})$$ → y 벡터 0과 1 사이 확률화.

$$L(t, S) = -\sum_{i=1}^{n} t_i \log S_i$$

Cross entropy.

$$t = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad S = \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ \vdots \\ S_n \end{bmatrix}$$

Cat

---

①. t 에서 답을 나타내는 1 의 의미를 해석해보자.

그러므로 학습시 잘되려면 $L(t, S) = 0$ ...

극단적으로 학습이 잘된경우 $L(t,S)=0$ 을 의미하고

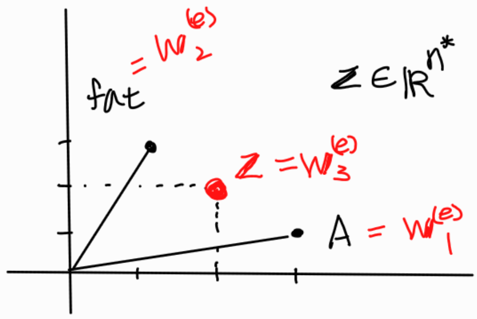
$S_3 = 1$ 임.

$$S_3 = \frac{e^{y_3}}{\sum_{i=1}^{n} e^{y_i}} = 1$$ ∴ y_3 값이 다른 y_i 값보다 상대적으로 매우커 1이되는 상황.

$$\begin{bmatrix} w_{11}^{(e)} & \cdots & w_{n^*,1}^{(e)} \\ w_{12}^{(e)} & & w_{n^*,2}^{(e)} \\ \boxed{w_{13}^{(e)} \quad \ddots \quad w_{n^*,3}^{(e)}} & & \\ \vdots & & \vdots \\ w_{1,n}^{(e)} & \cdots & w_{n^*,n}^{(e)} \end{bmatrix} \begin{bmatrix} w_{11}^{(e)} + w_{12}^{(e)} + w_{15}^{(e)} + w_{10}^{(e)} \\ \vdots \\ w_{n^*,1}^{(e)} + w_{n^*,2}^{(e)} + w_{n^*,s}^{(e)} + w_{n^*,0}^{(e)} \end{bmatrix} \times \frac{1}{4} \ (2m)$$

• Unit vector 가정.

$y_3 = \|w_3^{(e)}\| \, \|z\| \cos\theta$. $\rightarrow$ $W_3^{(e)} = z$ 같다고 해보자.



즉, $\frac{1}{4}(\underbrace{W_1^{(e)}}_{A} + \underbrace{W_2^{(e)}}_{fat} + \underbrace{W_4^{(e)}}_{sat} + \underbrace{W_5^{(e)}}_{on}) = \underbrace{W_3^{(e)}}_{cat}$.

distributional hyphoethsis on 합당

극단적으로 학습이 잘된경우 $L(t,S)=0$ 을 의미하고

S_3을 제외한 나머지 S_i 들은 모두 0임.

special 한 경우를 생각해서 $y_3 = 1$, 나머지 $y_i = 0$ 이라해보자.

$y_i = \|W_i^{(e)}\| \, \|z\| \cos\theta = 0$. 즉 orthogonal 한 경우를 의미함.

---

② $L(t,S) > \delta$ 인 경우.