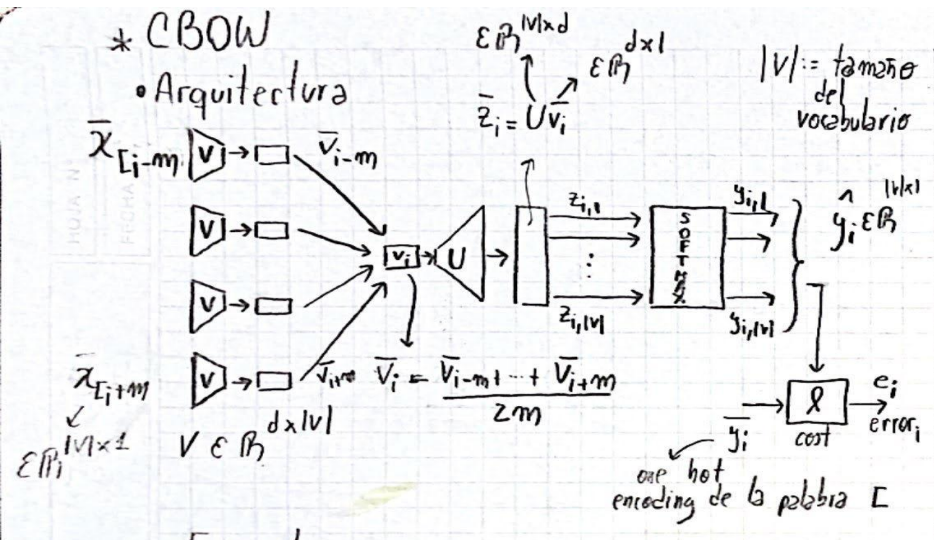
Procesamiento del Lenguaje Natural
Facultad de Ingeniería
Universidad de Buenos Aires

Ezequiel Esposito y Pablo Casas
(eesposito@optiwe.com, pcasas.biz@gmail.com)

* CBOW

- Arquitectura



- Forward

$$\bar{v}_{i-m} = V\bar{x}_{[i-m} \cdots \bar{v}_{i+m} = V\bar{x}_{[i+m}$$

$$\bar{v}_i = (\bar{v}_{i-m} + \cdots + \bar{v}_{i+m}) / 2m$$

$$\bar{z}_i = U\bar{v}_i \qquad \bar{z}_i \in \mathbb{R}^{|V|\times 1}$$

$$\hat{y}_i = \text{SOFTMAX}(\bar{z}_i) = \begin{bmatrix} e^{z_{i,1}} / \sum_{k=1}^{|V|} e^{z_{i,k}} \\ e^{z_{i,|V|}} / \sum_{k=1}^{|V|} e^{z_{i,k}} \end{bmatrix}$$

$\bar{y}_i$:= one hot encoding de la palabra central $\bar{x}_{[i}$

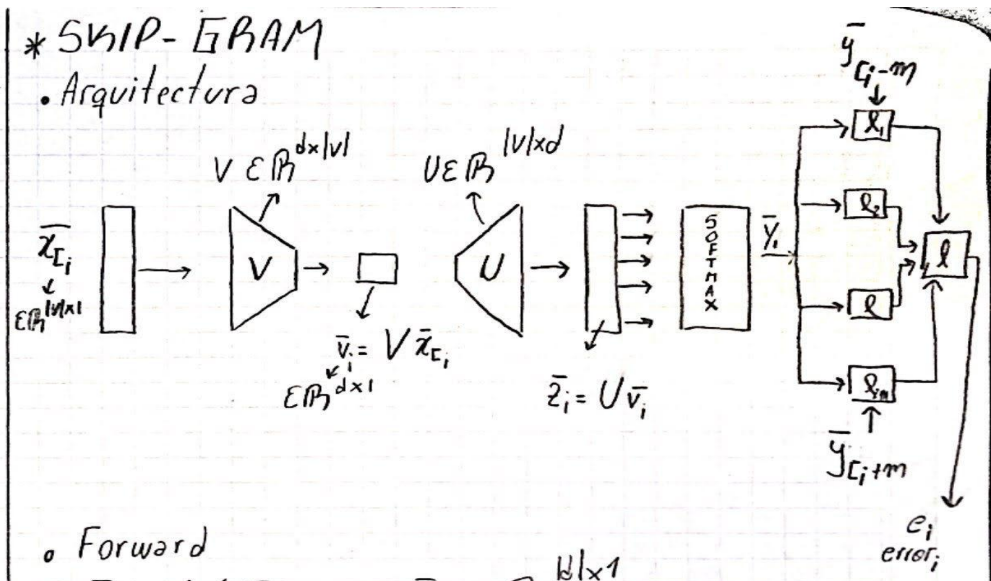
- Cost

$$J = \prod_{i=1}^{m} P(\hat{y}_i = \bar{y}_i \mid \bar{x}_{[i-m}, \ldots, \bar{x}_{[i+m})$$

multinomial vectorizada

* SKIP-GRAM

- Arquitectura



- Forward

$$\bar{v}_i = V\bar{x}_{[i} \quad ; \quad \bar{v}_i \in \mathbb{R}^{|V|\times 1}$$

$$\bar{z}_i = U\bar{v}_i \quad ; \quad \bar{z}_i \in \mathbb{R}^{|V|\times 1}$$

$$\hat{y}_i = \text{SOFTMAX}(\bar{z}_i)$$

$\bar{y}_{[i-m}$ = one hot encoding de la palabra $[i-m$

$\bar{y}_{[i+m}$ = one hot encoding de la palabra $[i+m$

- Cost

$$J = \prod_{i=1}^{m} \left[P(\hat{y}_i = y_{[i-m} \mid \bar{x}_{[i}) * \cdots * P(\hat{y}_i = y_{[i+m} \mid \bar{x}_{[i}) \right]$$

* CBOW
  - Backward

* SKIP-GRAM
  - Backward

- Recordemos sobre como calcular el cost en una clasificación de n clases
- Los labels pueden tomar valores entre $\{0, .., C-1\}$ con C la cantidad de clases. Otra manera de representarlo es con vectores one-hot-encoding

$y_i = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ para la clase 0, $y_i = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}$ para la clase 1, $\cdots$, $y_i = \begin{bmatrix} 0 \\ \vdots \\ 1 \end{bmatrix}$ para la clase $C-1$

- La salida de nuestra red neuronal es el SOFTMAX: $\hat{y}_i = \begin{bmatrix} e^{z_{i,0}} / \sum_{n=0}^{C-1} e^{z_{i,n}} \\ e^{z_{i,C-1}} / \sum_{n=0}^{C-1} e^{z_{i,n}} \end{bmatrix}$
- Donde cada componente j de $\hat{y}_i$ representa la probabilidad de que la salida de nuestro clasificador pertenezca a la clase j. Entonces podemos escribir de manera genérica
- $P\left(\hat{y}_i = \bar{y}_i \mid \bar{X}_i = x_i\right) = \sum_{j=0}^{C-1} y_{i,j} \frac{e^{z_{i,j}}}{\sum_{n=0}^{C-1} e^{z_{i,n}}}$ donde $y_{i,j}$ solo vale 1 para un j en particular

Para cada muestra solo un término de la sumatoria es distinto a cero

* CBOW

$$J = \prod_{i=1}^{n} P\left(\hat{y}_i = \bar{y}_i \mid X_{C_i-m}, \dots, X_{C_i+m}\right)$$

$$\arg\min_\theta \; -\frac{1}{n} \log \prod_{i=1}^{n} P\left(\hat{y}_i = \bar{y}_i \mid X_{C_i-m}, \dots, X_{C_i+m}\right)$$

prop del log → $$-\frac{1}{n} \sum_{i=1}^{n} \log\left[P\left(\hat{y}_i = \bar{y}_i \mid X_{C_i-m}, \dots, X_{C_i+m}\right)\right]$$

softmax → $$-\frac{1}{n} \sum_{i=1}^{n} \log\left[\sum_{j=0}^{|V|} y_{i,j}\, e^{z_{i,j}} \Big/ \sum_{k=0}^{|V|} e^{z_{i,j}}\right]$$

solo el término $j=C_i \neq 0$ → $$-\frac{1}{n} \sum_{i=1}^{n} \log\left[y_{i,C_i}^{1}\, e^{z_{i,C_i}} \Big/ \sum_{k=0}^{|V|} e^{z_{i,k}}\right]$$

prop logs → $$-\frac{1}{n} \sum_{i=1}^{n} \left[z_{i,C} - \log\left(\sum_{k=0}^{|V|} e^{z_{i,k}}\right)\right]$$

⇒ Como $z_i = U \bar{v}_i \rightarrow z_{i,C} = \bar{\mu}_{C_i}^{T} \bar{v}_i$ → embedding de palabra C_i

⇒ $$\arg\min_\theta \; \sum_{i=1}^{n} \left(\bar{\mu}_{C_i}^{T} \bar{v}_i - \log \sum_{k=0}^{|V|} e^{z_{i,k}}\right)$$

$z_{i,k} = \bar{\mu}_k^{T} \bar{v}_i$

* SKIP-GRAM

$$J = \prod_{i=1}^{n} \left[P\left(\hat{y}_i = y_{C_i-m} \mid \bar{X}_{C_i}\right) * \cdots * P\left(\hat{y}_i = y_{C_i+m} \mid \bar{X}_{C_i}\right)\right] \quad (A)$$

$$\arg\min_\theta \; -\frac{1}{n} \log \prod_{i=1}^{n} \left[\prod_{\substack{j=0 \\ j\neq m}}^{2m} P\left(\hat{y}_i = y_{C_i-m+j}\right)\right]$$ → escribí más corto la productoria

$$\arg\min_\theta \; -\frac{1}{n} \sum_{i=1}^{n} \log\left[\prod_{\substack{j=0 \\ j\neq m}}^{2m} P\left(\hat{y}_i = y_{C_i-m+j}\right)\right]$$ → (A) prop del log

$$\arg\min_\theta \; -\frac{1}{n} \sum_{i=1}^{n} \sum_{\substack{j=0 \\ j\neq m}}^{2m} \log\left(e^{z_{i,C_i-m+j}} \Big/ \sum_{k=0}^{|V|} e^{z_{i,k}}\right)$$ → soft max (solo el término $\neq 0$)

$$\arg\min_\theta \; -\frac{1}{n} \sum_{i=1}^{n} \sum_{\substack{j=0 \\ j\neq m}}^{2m} z_{i,C_i-m+j} - \log \sum_{k=0}^{|V|} e^{z_{i,k}}$$ → prop log

⇒ Como $z_{i,C_i-m+j} = \bar{\mu}_{C_i-m+j}^{T} \bar{v}_{C_i}$

⇒ $$\arg\min_\theta \; -\frac{1}{n} \sum_{i=1}^{n} \sum_{\substack{j=0 \\ j\neq m}}^{2m} \left[\bar{\mu}_{C_i-m+j}^{T} \bar{v}_{C_i} - \log \sum_{k=0}^{|V|} e^{z_{i,k}}\right]$$

$z_{i,k} = \bar{\mu}_k^{T} \bar{v}_i$