Procesamiento de Lenguaje Natural

Olivia Gutú y Julio Waissman

Maestría en Ciencia de Datos
Semana 8: Modelos continuos de bolsa de palabras

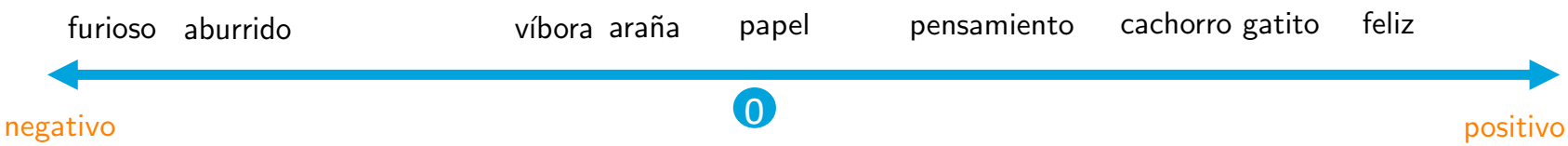
Objetivos de esta semana

- Identificar los conceptos claves de la representación de palabras
- Generar *word embedding vectors*
- Preparar el texto para el aprendizaje automático
- Implementar el modelo continuo de bolsa de palabras

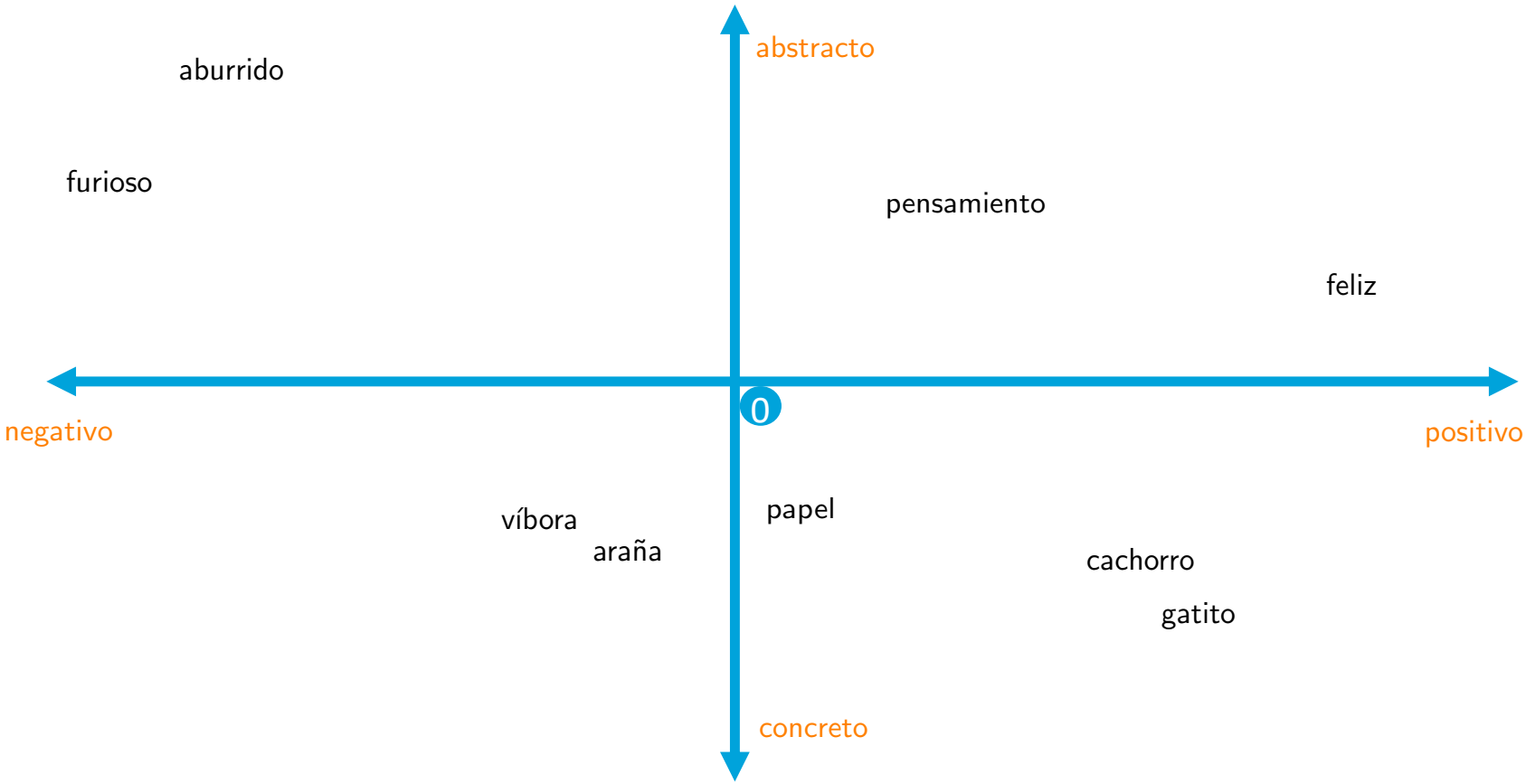
Representación vectorial básica de palabras

- Un *integer i* por palabra: muy simple pero 'cero' sentido semántico.
- *one-hot vectors*: vectores e_i en $\mathbb{R}^{|V|}$, simple, no hay orden, pero vectores enormes, sin sentido semántico, la distancia entre cualquier par de vectores es 1.
- *word embedding vectors*: baja dimensión, significado semántico, inferencia. vectores cercanos semánticamente tienen distancia coseno cercana.

Significado de los vectores



Significado de los vectores



Tipos de procesos de *word embedding*

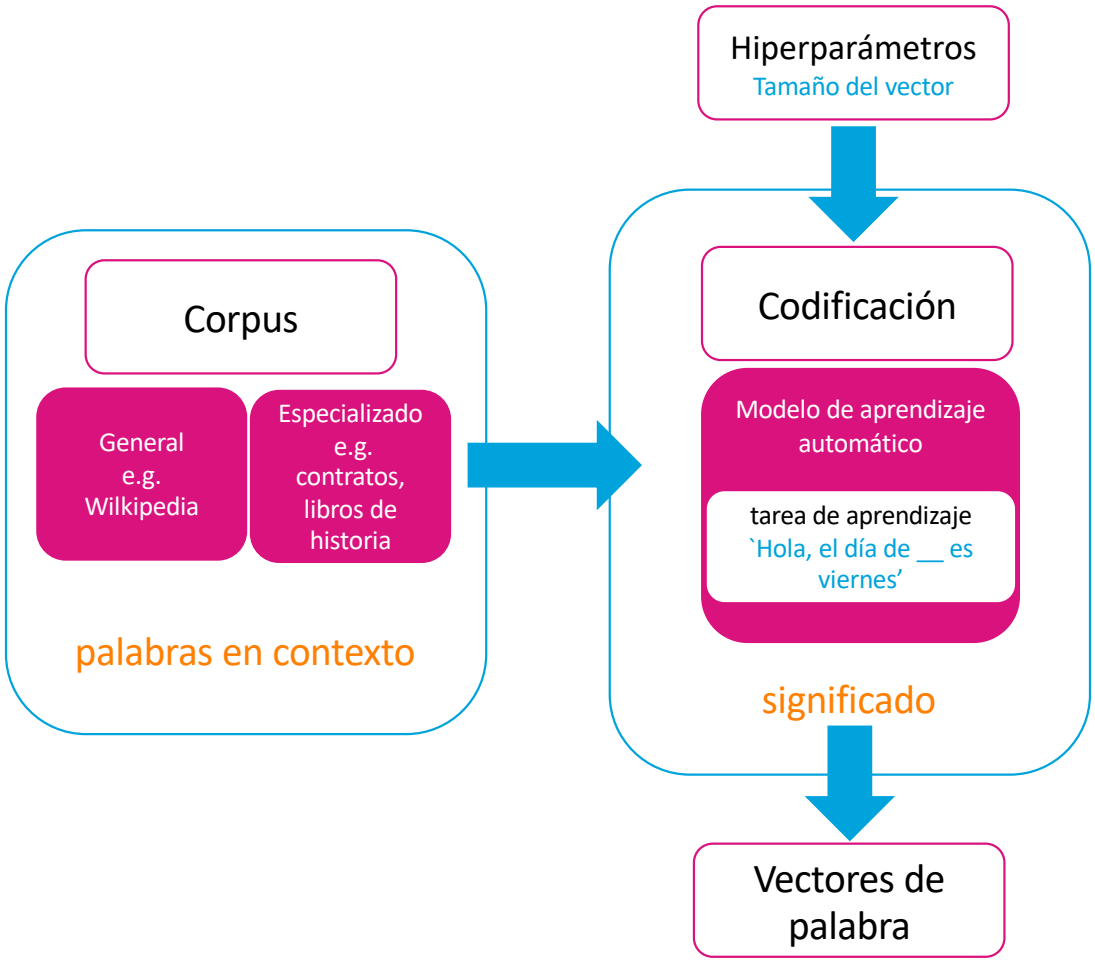
Básicos:

- word2vec (Google, 2013)
  - CBOW
  - Skip-Gram
- GloVe (Stanford, 2014)
- fastText (Facebook, 2016)
  - Soporte de OOV

Avanzados (deep learning)

- BERT (Google, 2018)
- ELMo (Allen Institute for AI, 2018)
- GPT-2 (OpenAI, 2018)

Proceso de *word embedding*



Preprocesamiento para el entrenamiento

Limpieza y tokenización

- Uniformizar mayúsculas y minúsculas canto = Canto = CANTO lowecase/uppercase
- Puntuación , ! . ? ??? !!! → . “”«»‘’ → ⊘
- Números 1 2 3 9 10 → ⊘ 3.1416 →< NUMBER >
- Caracteres especiales % $ § → ⊘
- Tokens especiales ☺ #nezahualcoyotl →: *happy* : #nezahualcoyotl

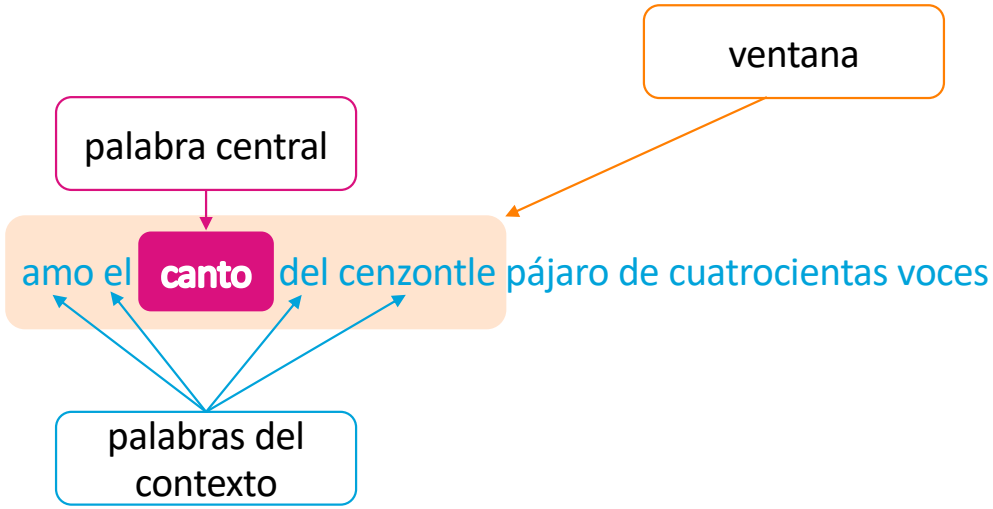
♡ 1000 el canto del cenzotle!!!!! #nezahualcoyotl

↓

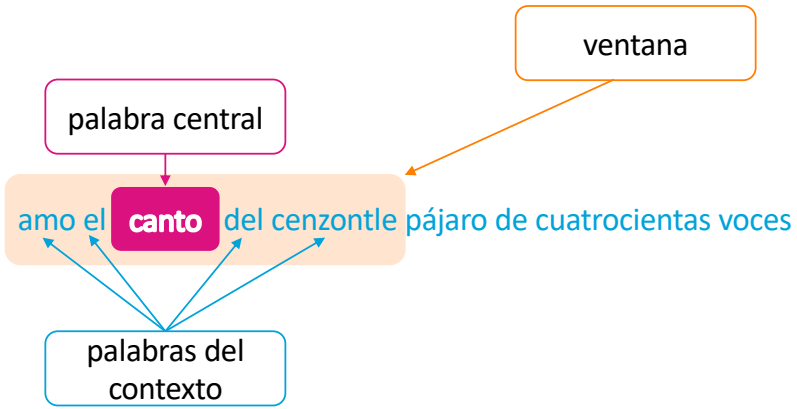
[‘♡’, ‘el’, ‘canto’, ‘del’ , ‘cenzontle’, ‘.’,‘#nezahualcoyotl’]

Preprocesamiento para el entrenamiento

- tamaño de ventana $= 5$ (window size)
- $C = 2$ (context half-size)



Preprocesamiento para el entrenamiento



contexto	palabra central
amo el del cenzontle	canto
el canto cenzontle pájaro	del
canto del pájaro de	cenzontle
del cenzontle de cuatrocientas	pájaro
⋮	⋮

Preprocesamiento para el entrenamiento

Codificación inicial de las palabras centrales

$$V = \{\text{amo, canto, cenzontle, cuatrocientas, de, del, el, pájaro, voces}\}$$

$$\begin{bmatrix}1\\0\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}\quad\begin{bmatrix}0\\1\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}\quad\begin{bmatrix}0\\0\\1\\0\\0\\0\\0\\0\\0\end{bmatrix}\quad\begin{bmatrix}0\\0\\0\\1\\0\\0\\0\\0\\0\end{bmatrix}\quad\begin{bmatrix}0\\0\\0\\0\\1\\0\\0\\0\\0\end{bmatrix}\quad\begin{bmatrix}0\\0\\0\\0\\0\\1\\0\\0\\0\end{bmatrix}\quad\begin{bmatrix}0\\0\\0\\0\\0\\0\\1\\0\\0\end{bmatrix}\quad\begin{bmatrix}0\\0\\0\\0\\0\\0\\0\\1\\0\end{bmatrix}\quad\begin{bmatrix}0\\0\\0\\0\\0\\0\\0\\0\\1\end{bmatrix}$$

Codificación inicial del contexto: 'promedio' de las palabras involucradas.

Promedio

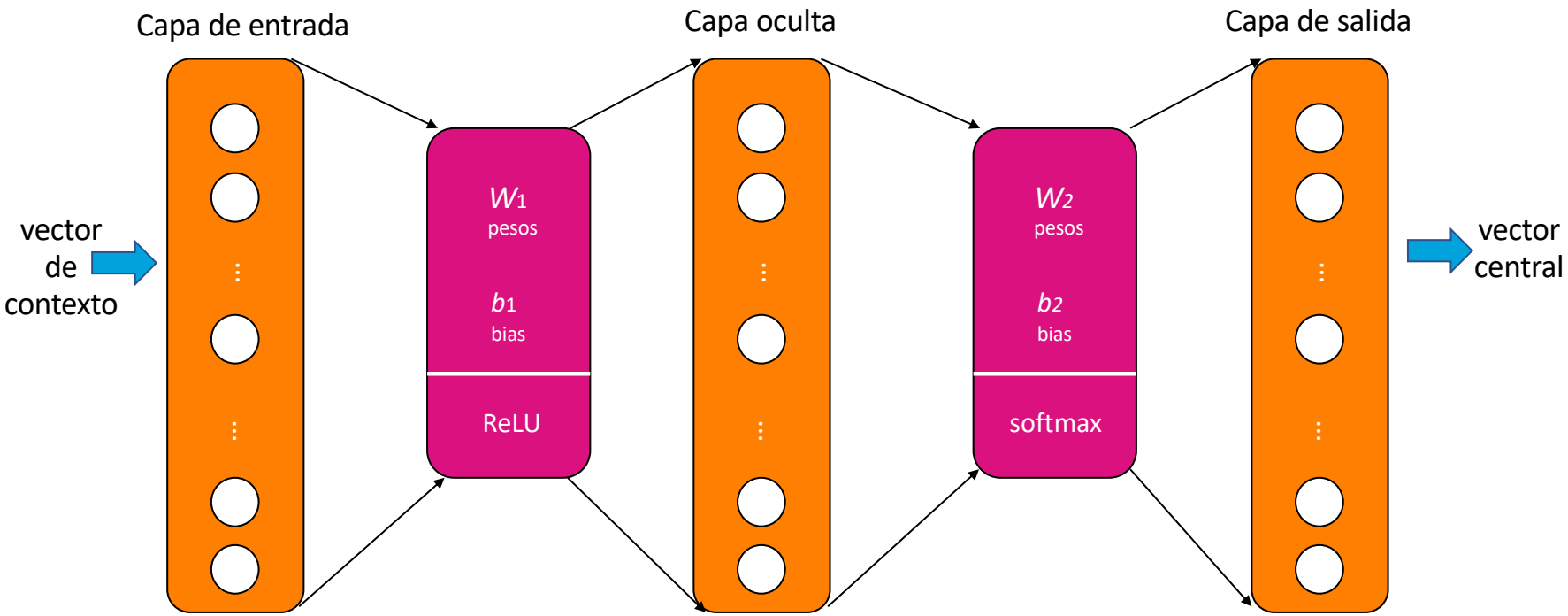
amo	el	del	cenzontle	contexto
1	0	0	0	0.25
0	0	0	0	0
0	0	0	1	0.25
0	0	0	0	0
0	0	0	0	0
0	0	1	0	0.25
0	1	0	0	0.25
0	0	0	0	0
0	0	0	0	0

Preprocesamiento para el entrenamiento

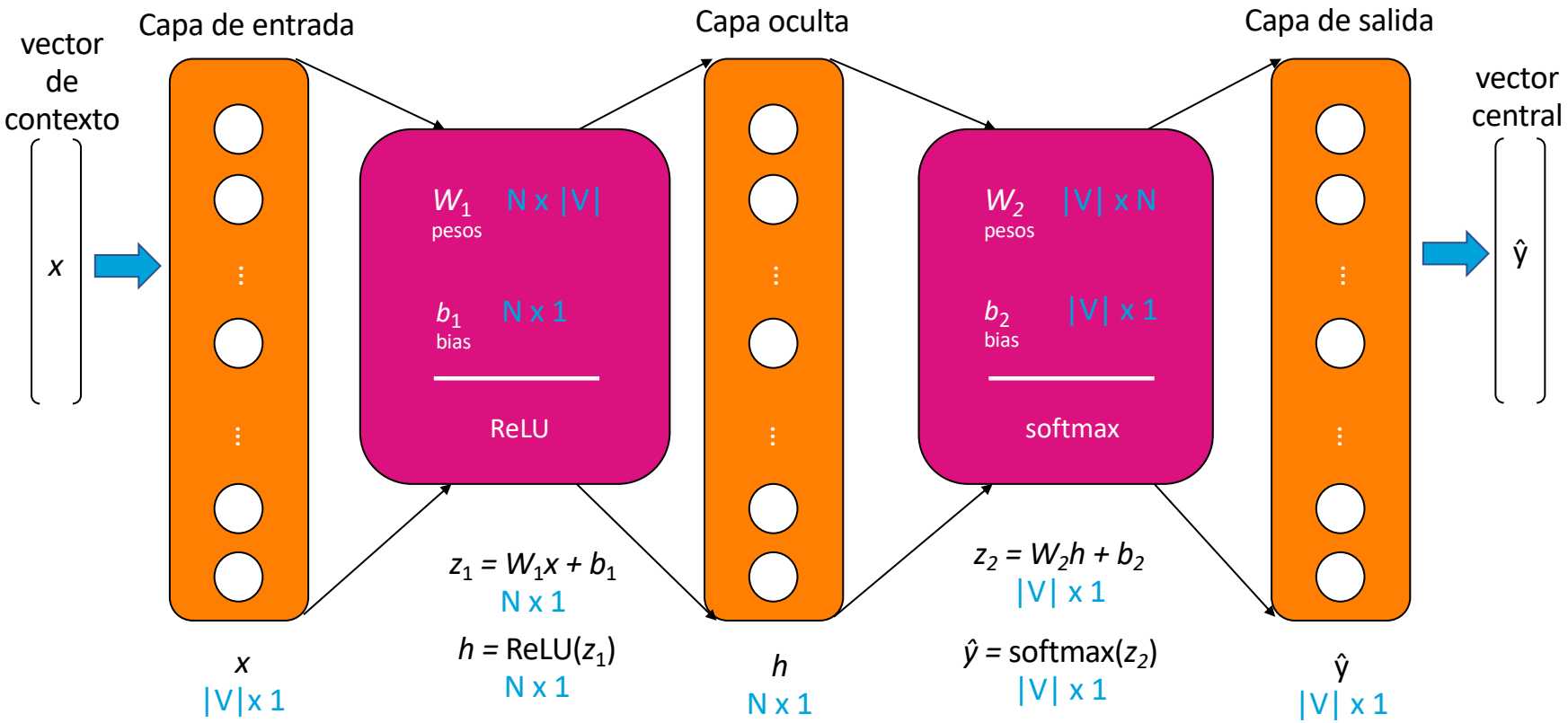
Preparación final

contexto	vector contexto	palabra central	vector palabra central
amo el del cenzontle	[0.25,0,0.25,0,0,0.25,0.25,0,0]	canto	[0,1,0,0,0,0,0,0,0]
⋮	⋮	⋮	⋮

Arquitectura del modelo CBOW

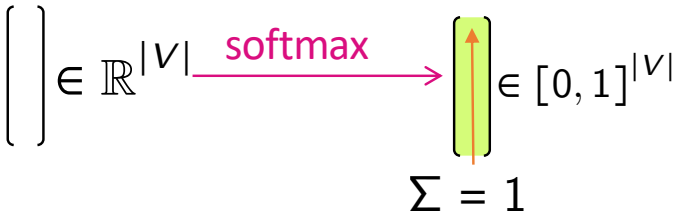


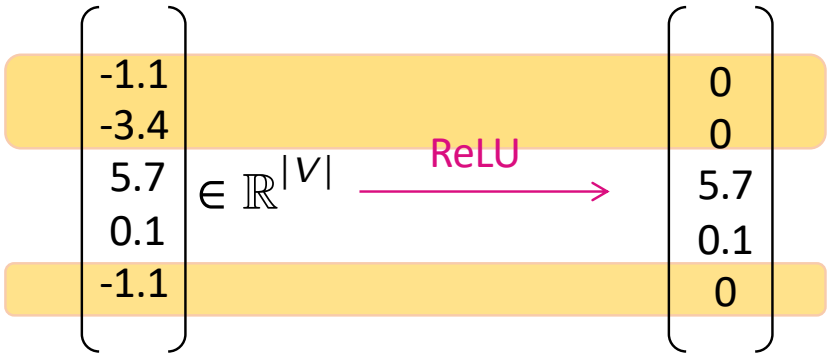
Dimensiones (una entrada)

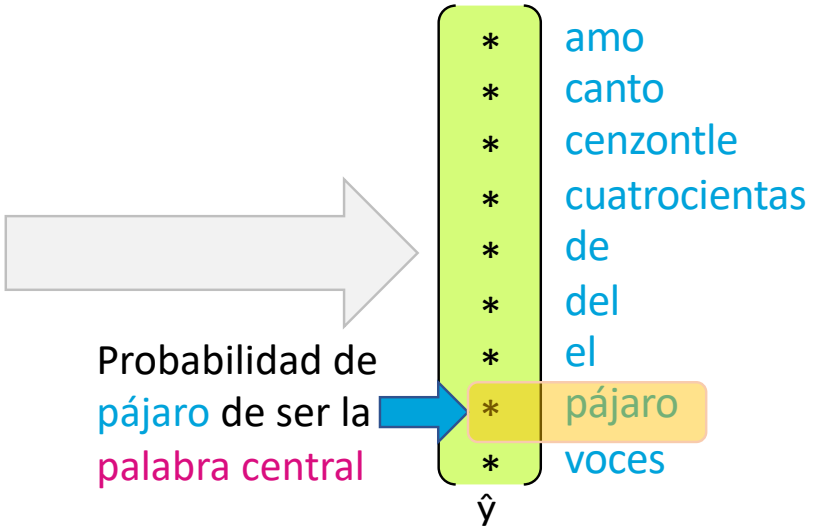


Dimensiones (una entrada)

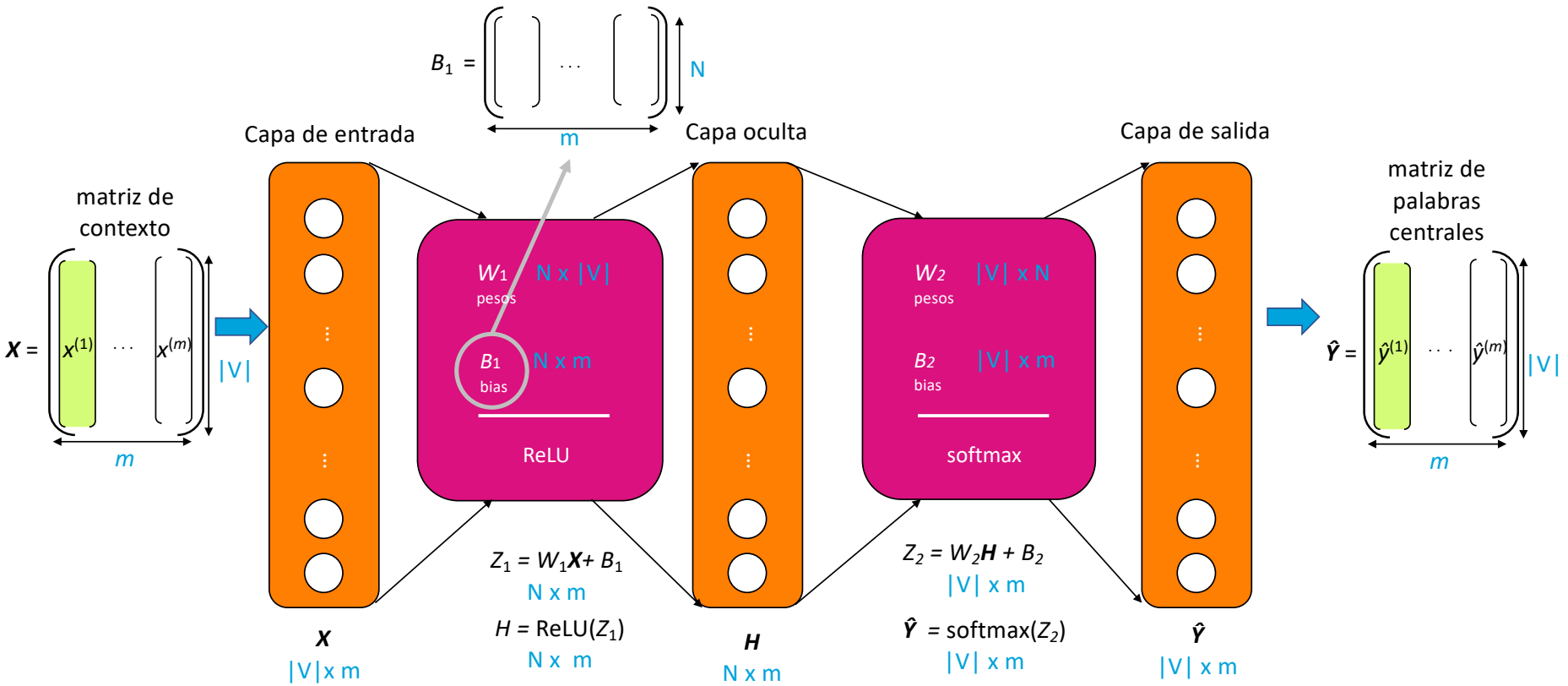
- $\text{ReLU}(z) = \max(0, z)$ entrada por entrada (*rectified linear unit*)
- $\text{softmax}(z)_i = \frac{\exp(z_i)}{\sum_{k=1}^{|V|} \exp(z_k)}$



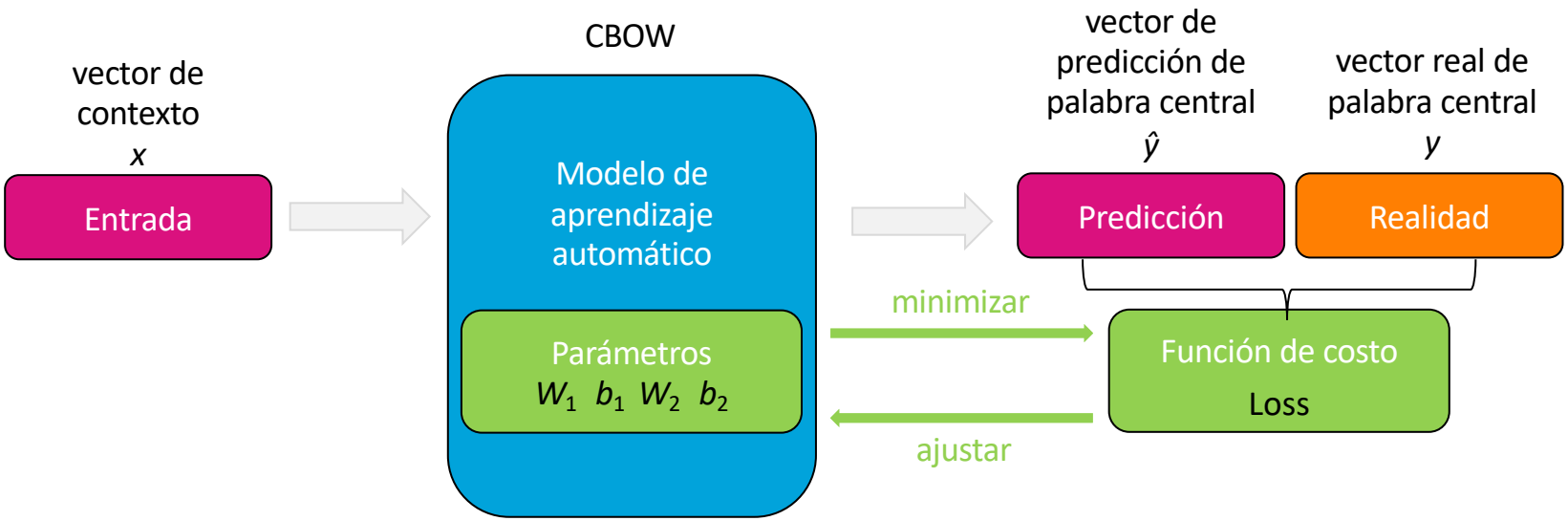




Dimensiones (entrada en *batch*)

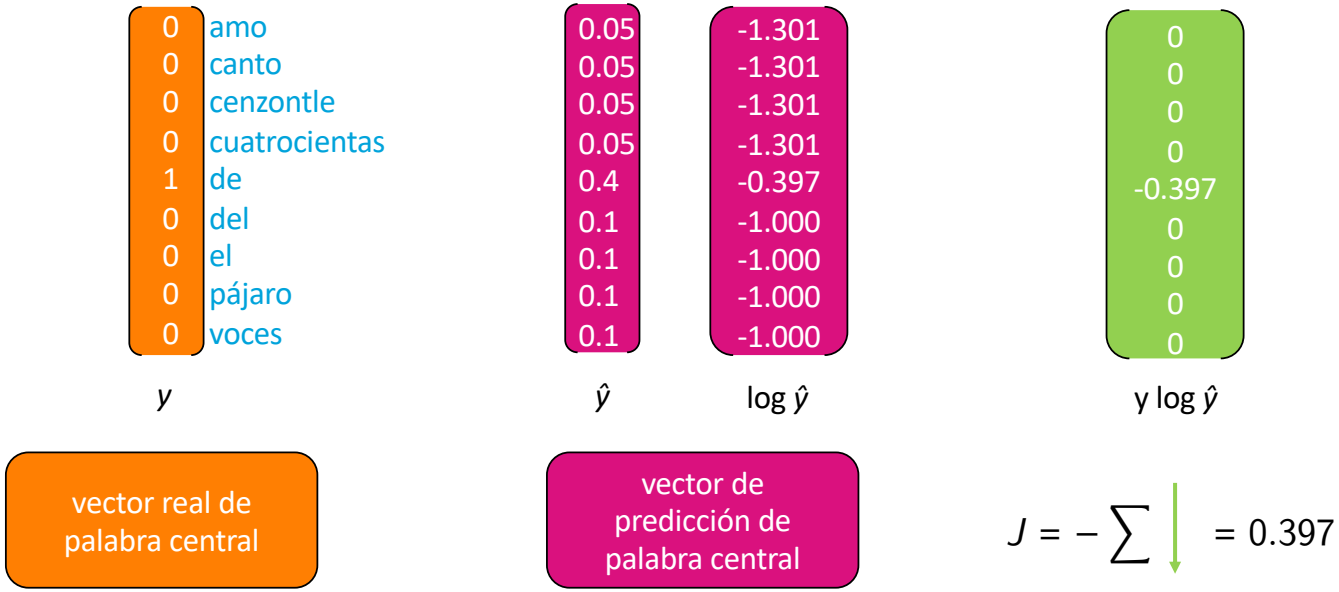


Entrenamiento



Función de costo

$$J = -(y_1 \log \hat{y}_1 + y_2 \log \hat{y}_2 + \cdots + y_{|V|} \log \hat{y}_{|V|})$$



$$J_{\text{batch}} = \frac{1}{m} \sum_{i=1}^{m} J_i$$

Minimización del costo

- Backpropagation

$$Grad\, J_{batch}$$

- Descenso de gradiente

$$(W_1, W_2, b_1, b_2) := (W_1, W_2, b_1, b_2) - \alpha \text{Grad}\, J_{batch}$$

Minimización del costo

Backpropagation

$$\frac{\partial J_{\text{batch}}}{\partial W_1} = \frac{1}{m}\text{ReLU}\left(W_2^T(\hat{Y}-Y)\right)X^T$$

$$\frac{\partial J_{\text{batch}}}{\partial W_2} = \frac{1}{m}(\hat{Y}-Y)H^T$$

$$\frac{\partial J_{\text{batch}}}{\partial b_1} = \frac{1}{m}\text{ReLU}\left(W_2^T(\hat{Y}-Y)\right)(1,\cdots,1)_m{}^T$$

$$\frac{\partial J_{\text{batch}}}{\partial b_2} = \frac{1}{m}(\hat{Y}-Y)(1,\cdots,1)_m{}^T$$

Descenso de gradiente

$$W_1 := W_1 - \alpha\frac{\partial J_{\text{batch}}}{\partial W_1}$$

$$W_2 := W_2 - \alpha\frac{\partial J_{\text{batch}}}{\partial W_2}$$

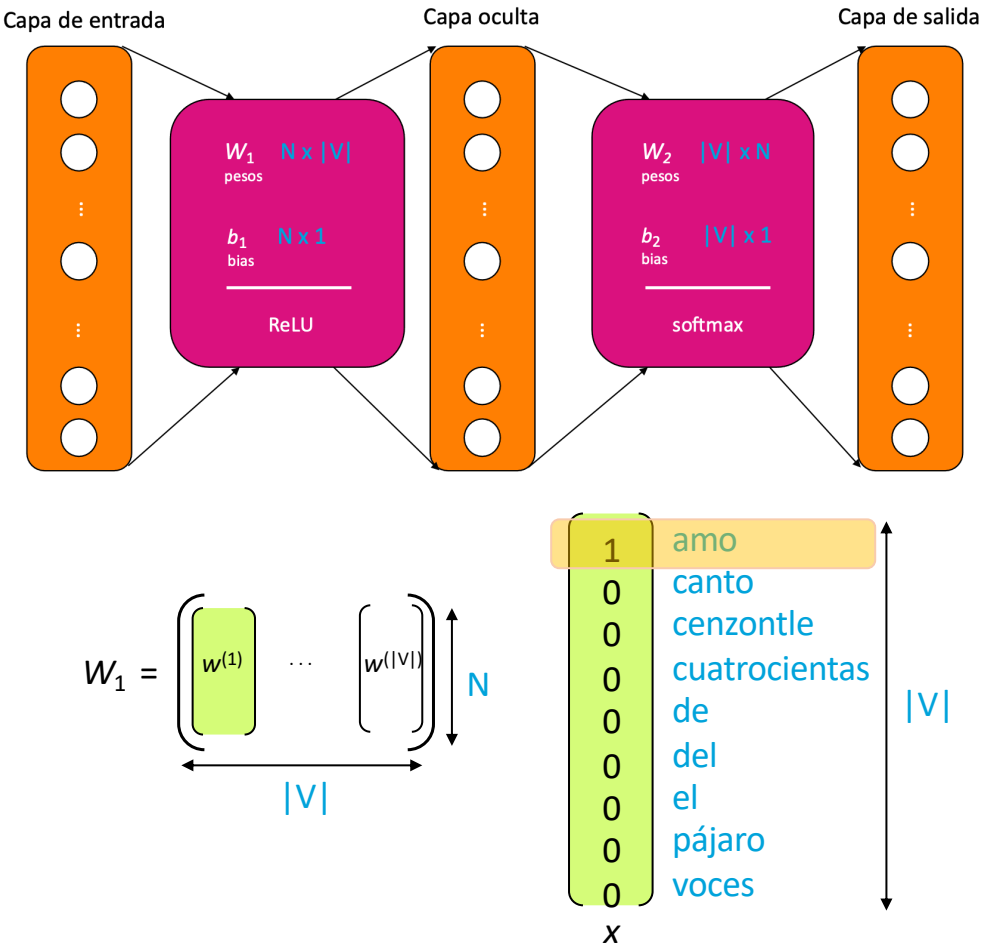
$$b_1 := b_1 - \alpha\frac{\partial J_{\text{batch}}}{\partial b_1}$$

$$b_2 := b_2 - \alpha\frac{\partial J_{\text{batch}}}{\partial b_2}$$

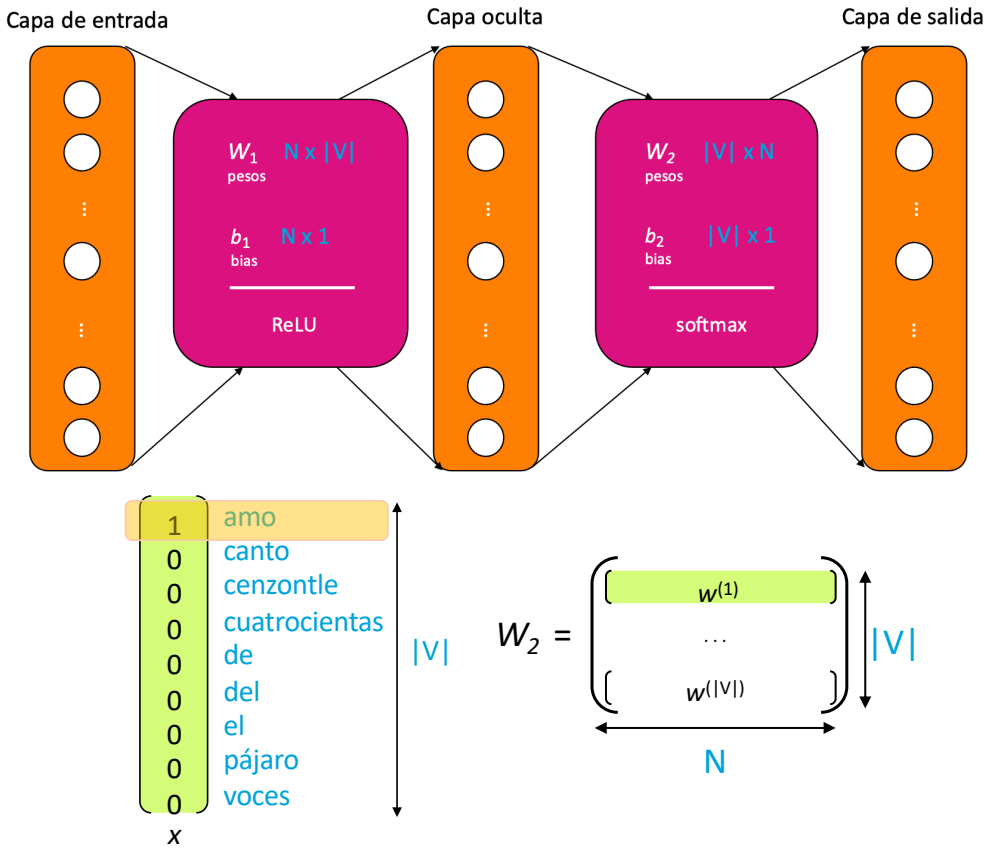
α : tasa de aprendizaje

https://www.coursera.org/lecture/neural-networks-deep-learning/derivatives-of-activation-functions-qcG1j

Extracción de vectores de palabra: opción 1

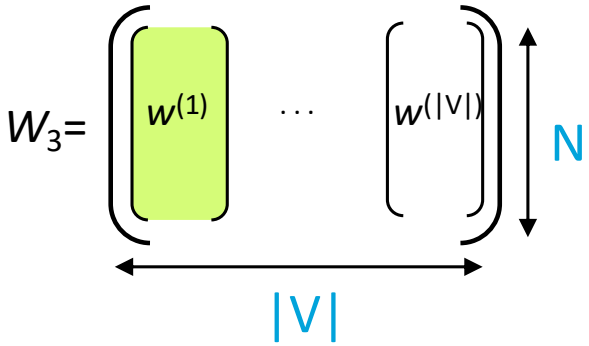


Extracción de vectores de palabra: opción 2



Extracción de vectores de palabra: opción 3

$$W_3 = \frac{1}{2}(W_1 + W_2^T)$$



Evaluación intrínseca

- Analogías
  - Semánticas: 'México' es a 'CDMX' lo que 'Argentina' es a <? >
  - Sintácticas: 'cantó' es a 'cantar' lo que 'bailó' es a <? >
- Clustering

fuente: Google news