DISTRIBUTIONAL SEMANTICS AND WORD2VEC

- BOW FEATURE REPRESENTATION IS PARTLY ONE OF CONVENIENCE! IT GIVES US A WAY TO REPRESENT VARIABLE LENGTH SEQUENCES WITH A FIXED INPUT SIZE.
- UNFORTUNATELY, LANGUAGE IS CONTEXTUAL. THE MEANING OF A WORD IS A FUNCTION OF ITS CONTEXT.
- ANY MODEL OF NATURAL LANGUAGE THEREFORE MUST ACCOUNT FOR THE MEANING OF WORDS IN CONTEXT.

∴ WE MUST MOVE AWAY FROM BOW!

- NEW FEATURE REPRESENTATION:

VOCABULARY: $X = I_N = \begin{matrix} \\ N \end{matrix}\begin{bmatrix} 1 & \cdots & 0 \\ 0 & \ddots & 1 \end{bmatrix}$ N = VOCAB SIZE

EXAMPLE

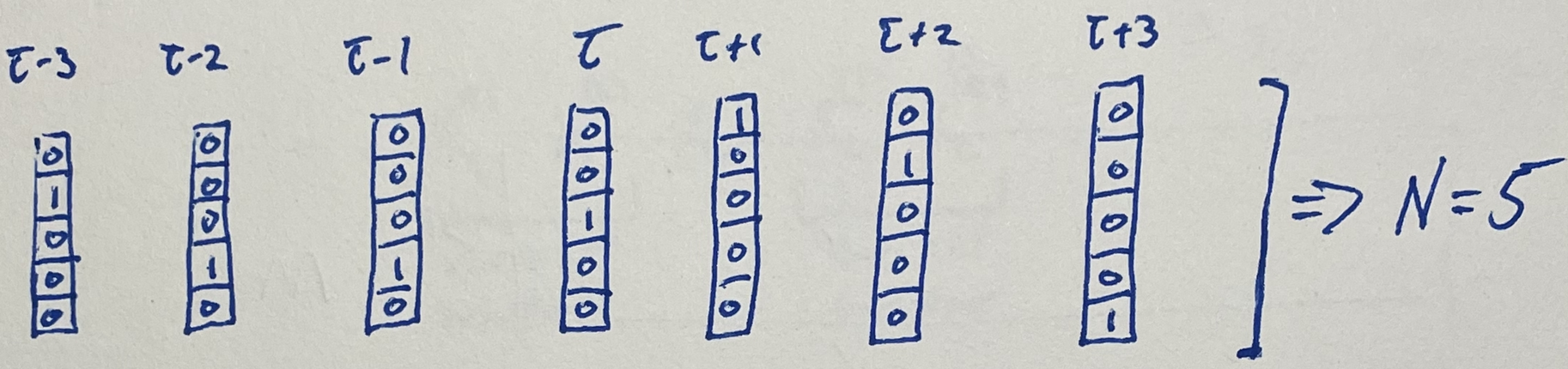
↳ $x_i = [0\ 0\ 0\ 1\ 0]$ = ONE-HOT VECTOR REPRESENTATION of i^{TH} WORD IN X.

SEQUENCE-PRESERVING FEATURE REPR.

↳ NEW FEATURE REPR.:

$$\ldots\ldots\ X^{(t-3)}\ X^{(t-2)}\ X^{(t-1)}\ X^{(t)}\ X^{(t+1)}\ X^{(t+2)}\ X^{(t+4)}\ldots\ldots\ldots$$

↳ t INDEXES THE SEQUENCE DIMENSION



↳ SEQUENCE of ONE-HOT ENCODED WORDS

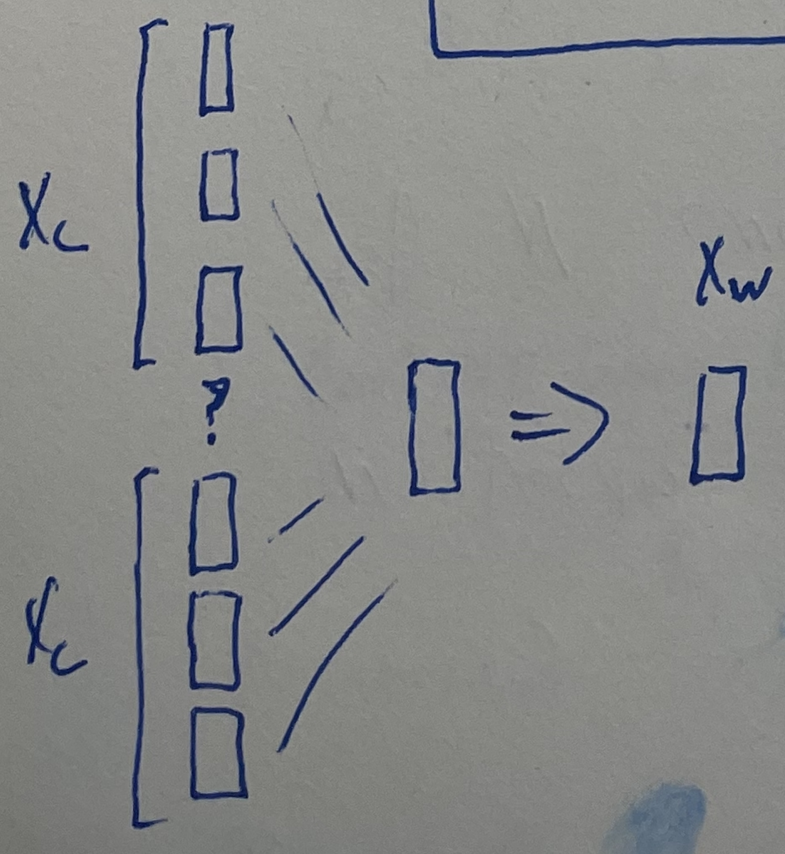
LANGUAGE MODELING APPROACHES

X_w = CENTER WORD
X_c = CONTEXT WORD

CONTINUOS BAG OF WORDS

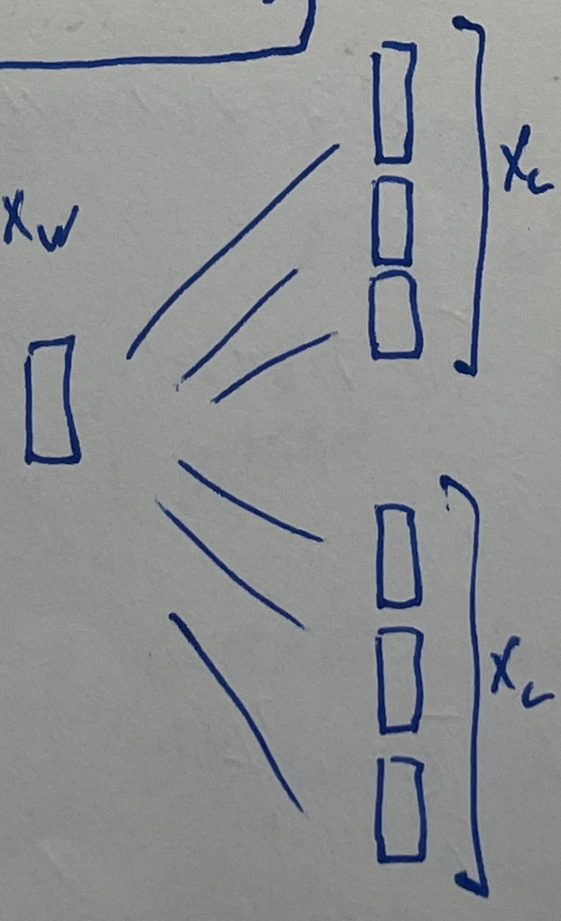
CBOW

$P(X_w \mid X_c^{(1)}, \ldots, X_c^{(c)})$



SKIP-GRAM

$P(X_c^{(i)} \mid X_w)$



SKIP GRAM WORD2VEC

DATASET: $D = \{ X_w^{(1)}, X_c^{(1)} \ \ldots \ X_w^{(M)}, X_c^{(M)} \}$

↳ M DOES NOT EQUAL NUMBER of WORDS IN CORPUS!

↳ FOR EACH (CENTER) WORD IN CORPUS:

$\_ \ \_ \ \_ \ \_ \ X_c^{(1)} \ X_c^{(2)} \ X_w \ X_c^{(3)} \ X_c^{(4)} \ \_ \ \_ \ \_ \ \_$

CONTEXT WINDOW — CONTEXT WINDOW

↳ $(X_w, X_c^{(1)}), (X_w, X_c^{(2)}), (X_w, X_c^{(3)}), (X_w, X_c^{(4)})$

$M =$ NUM WORDS × CONTEXT WINDOW × 2

WORD2VEC

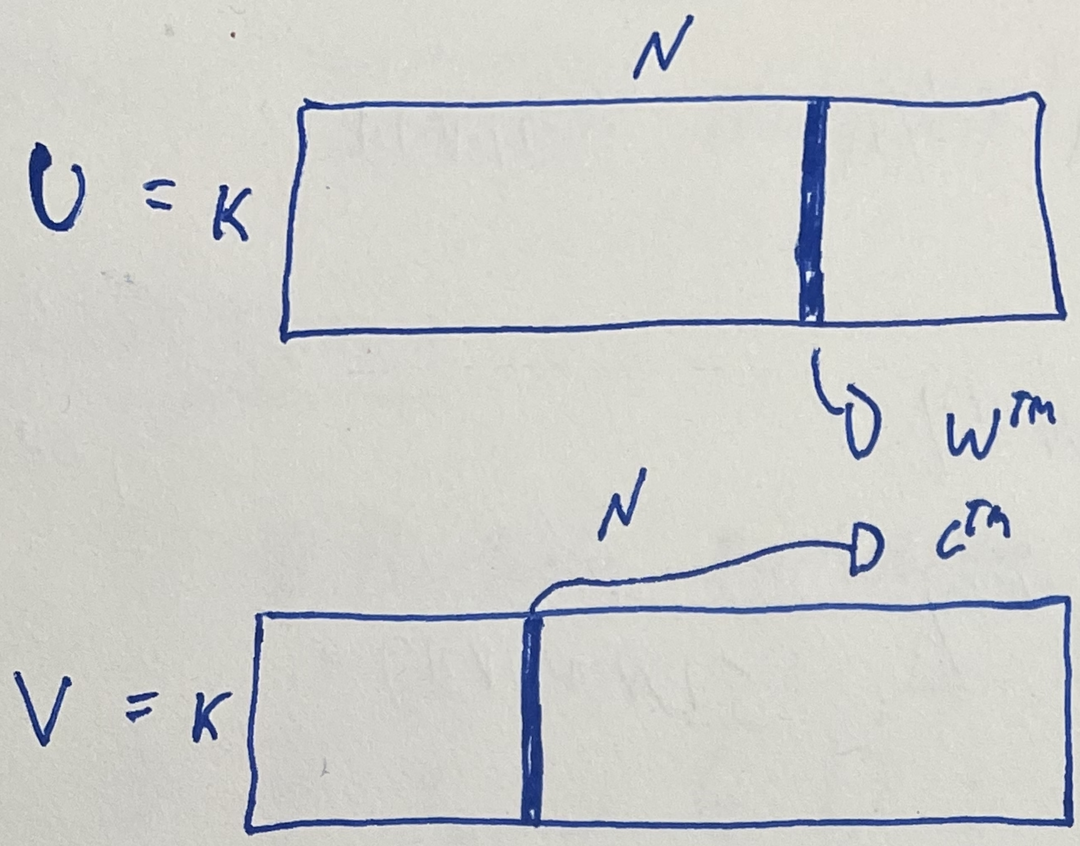
CENTER WORD EMBEDDINGS: $U \in \mathbb{R}^{K \times N}$ $\quad K << N$

↳ $X_w \cdot U = U_w \in \mathbb{R}^{K \times 1}$ = CENTER WORD EMB FOR w^{TH} WORD IN X

CONTEXT WORD EMBEDDINGS: $V \in \mathbb{R}^{K \times N}$

$X_c \cdot V = V_c \in \mathbb{R}^{K \times 1}$ = CONTEXT WORD EMB for c^{TH} WORD IN X.

SKIP-GRAM Word2Vec CONT...



↳ W^{TH} COLUMN OF $U = X_w \cdot U = u_w$

↳ C^{TH} COLUMN OF $V = X_c \cdot V = v_c$

↳ WORD2VEC OBJECTIVE:

① MAKE $u_w^T v_c$ LARGE IF X_c IS IN CONTEXT OF X_w IN D

② MAKE $u_w^T v_c$ SMALL IF X_c IS NOT IN CONTEXT OF X_w IN D

↳ HOW DOES WORD2VEC DO THIS?

↳ ANSWER: MLE OF THE DISCRIMINATIVE MODEL $P(X_c | X_w)$

↳ $P(X_c | X_w ; U, V) = \frac{e^{u_w^T v}}{\sum_{j=1}^{N} e^{(u_w^T v)_j}} \in [0,1]^N$

↳ $P_{X_c | X_w}$ FOR SHORT

SKIP-GRAM WORD2VEC ALGORITHM

L5-5

GOAL: MINIMIZE $NLL(U,V \mid D)$ USING STOCHASTIC GD

ALGORITHM:

① RANDOMLY INITIALIZE $U, V \in \mathbb{R}^{K \times N}$

② $\forall\ X_c, X_w \in D$:

- "LOOKUP" THE EMBEDDING U_w IN U POINTED TO BY X_w.
- COMPUTE: $Z = U_w^T V \quad \in \mathbb{R}^N$
- COMPUTE: $P_{X_c \mid X_w} = \dfrac{e^Z}{\sum_{j=1}^{N} e^{z_j}} \quad \in [0,1]^N$
- COMPUTE: $NLL = -X_c \cdot \log P_{X_c \mid X_w} \quad \in \mathbb{R}$
- COMPUTE: $\nabla_{U_w} NLL = V \cdot (P_{X_c \mid X_w} - X_c)^T \quad \in \mathbb{R}^{K \times 1}$
- COMPUTE: $\nabla_V NLL = U_w^T \cdot (P_{X_c \mid X_w} - X_c) \quad \in \mathbb{R}^{K \times N}$
- UPDATE: $(U^T)_w = (U^T)_w - \eta \nabla_{U_w} NLL$
- UPDATE: $V = V - \eta \nabla_V NLL$

REPEAT