1、什么是 Word2Vec

(1) 大部分的有监督 ML 模型都可概括为 $f(x) \to y$

(2) 词嵌入：映射的数学模型只接受数值型输入，所以要把词语嵌入数学空间

(3) 语言模型 f：判断 (x, y) 是否符合自然语言的法则（x和y放一起是不是人话）

(4) 目的：只关心训练后的副产物 — 参数（NN的权重）

(5) 词向量：将这些参数作为输入 x 的某种向量化的表示

2、Skip-gram 和 CBOW 模型 （上下文 ⟶ 当前词 ； 当前词 ⟶ 上下文）

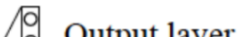
(1) x 的原始输入形式： One-hot encoder

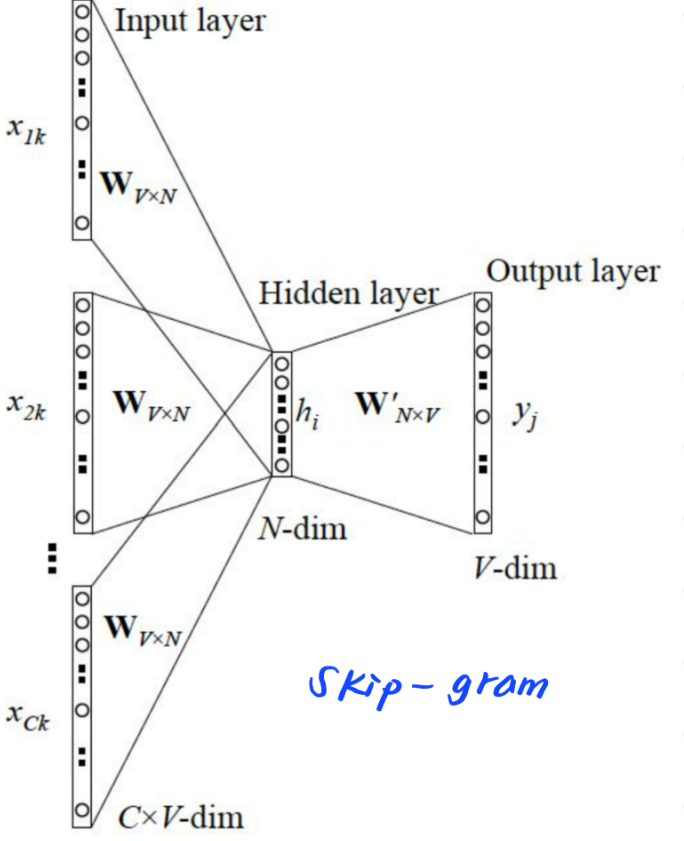
(2) Word2Vec 精髓：如果输入 [1,0,…,0]，则输入层到隐藏层的权重里，只有1对应位置的权重被激活，这些权重的个数，跟隐藏层节点数是一致的，从而这些权重组成一个向量 v_x 来表示 x（且唯一）

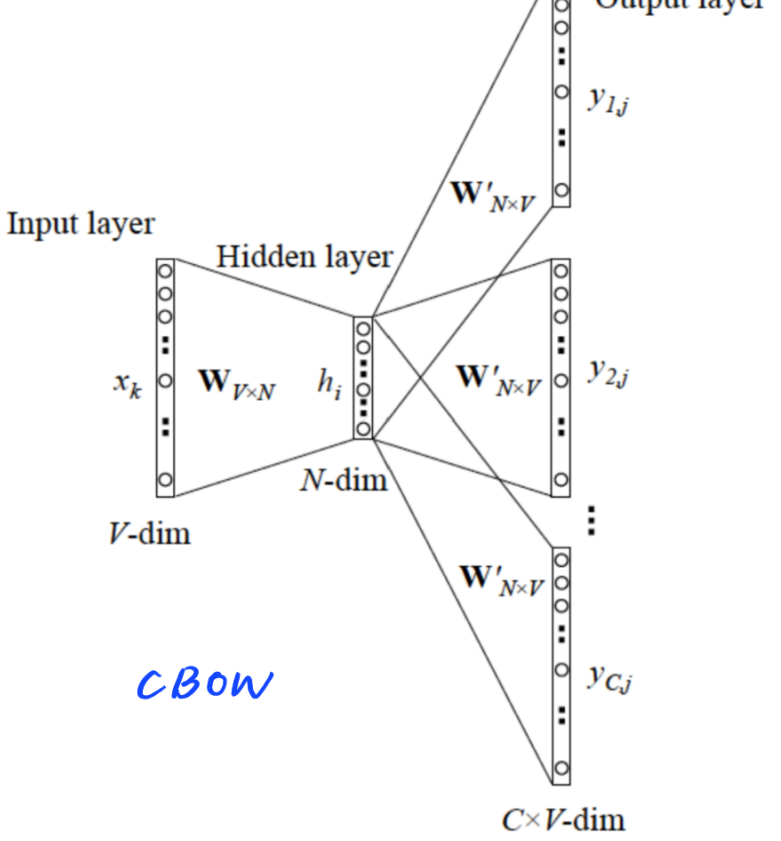
$$h_1 = \sigma(w_{11}x_1 + w_{12}x_2 + \cdots + w_{1V}x_V)$$
$$h_2 = \sigma(w_{21}x_1 + w_{22}x_2 + \cdots + w_{2V}x_V)$$
$$\vdots$$
$$h_N = \sigma(w_{N1}x_1 + w_{N2}x_2 + \cdots + w_{NV}x_V)$$

本质上是一种降维操作

(3) 上述是"输入向量"，也有"输出向量"：输出层用 One-hot 表示，同样地，隐藏层到输出层的对应权重可构成 v_y 用来唯一表示词







3、训练trick

(1) hierarchical softmax：本质是把N分类问题变成 log(N)次二分类

(2) negative sampling：本质是预测总体类别的一个子集