

Skip-gram model

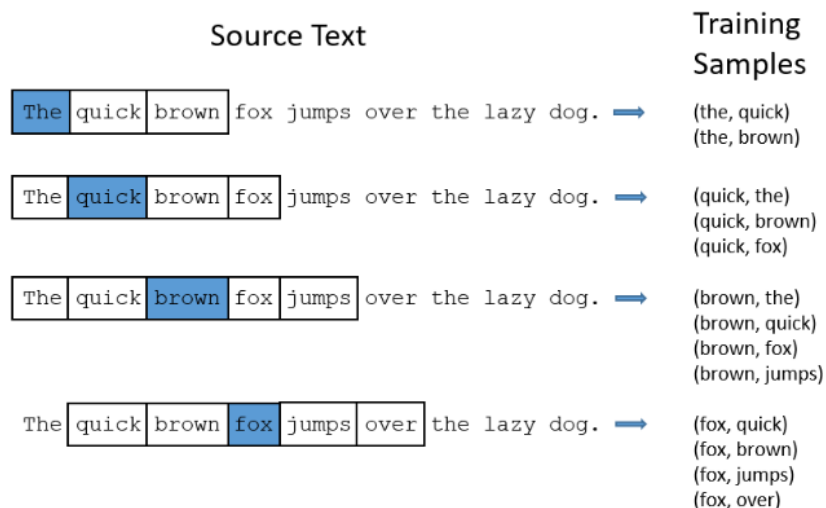
¿Qué es?

Es una arquitectura que, dado una secuencia de oraciones o corpus, se analiza las palabras en el en búsqueda del contexto de las palabras y la búsqueda de sus vecinos.

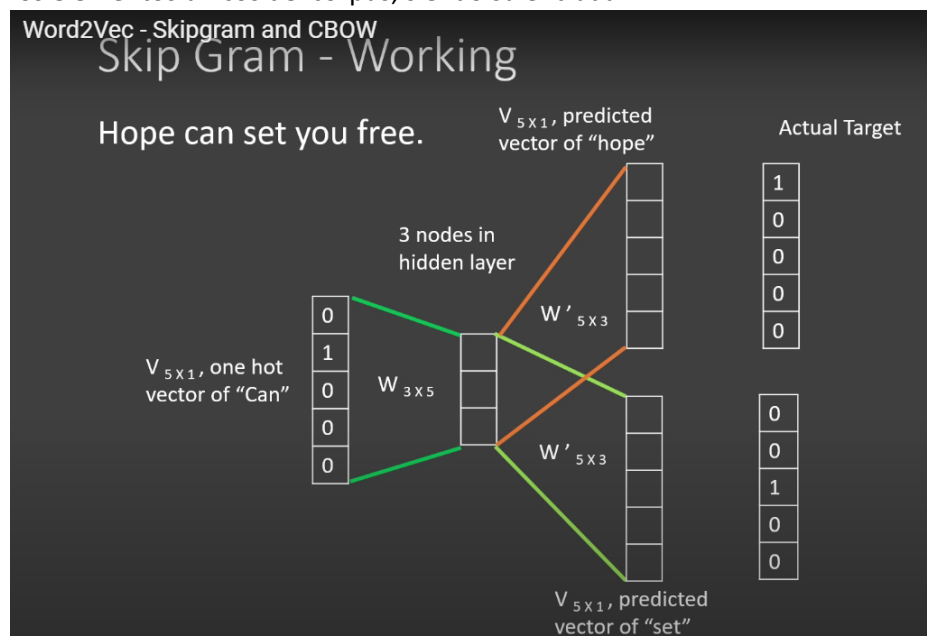
¿Como funciona?

Se utiliza una red neuronal simple con el objetivo de transformar las palabras a vectores (Word2Vect)

1. Se entrena con pares de palabras tomadas de algún documento



2. Se transforman las palabras por medio de un one hot vector, tomando como dimensión a los elementos únicos del corpus, siendo su entrada.



3. Los pesos que se generan por cada palabra en busca de acertar al target se guardan.

Weights after training

$W_{3 \times 5}$

w00	w01	w02	w03	w04
w10	w11	w12	w13	w14
w20	w21	w22	w23	w24

4. Y con los pesos podemos generar los Word vectors, utilizando los pesos de la red y sus respectivos One Hot vectors.

Word Vector for hope = $W_{3 \times 5} \times V_{5 \times 1}$

w00	w01	w02	w03	w04
w10	w11	w12	w13	w14
w20	w21	w22	w23	w24

 \times

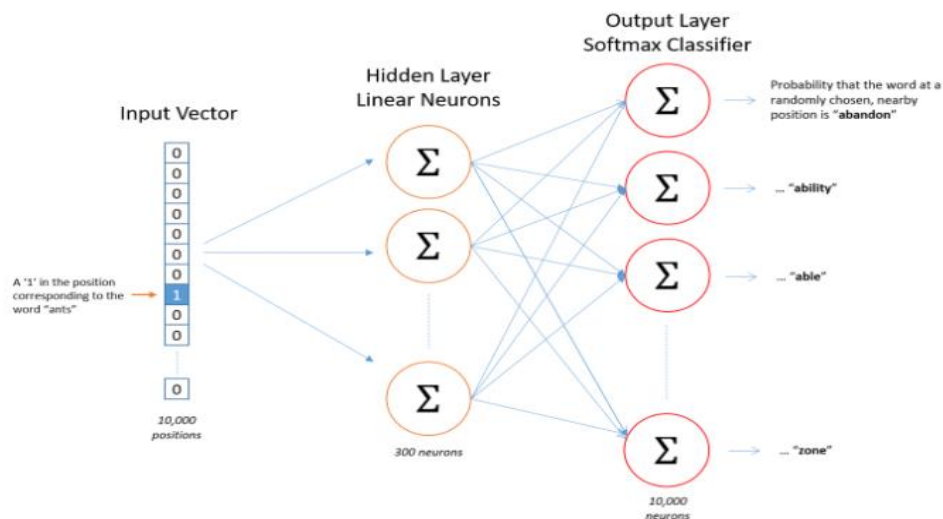
1
0
0
0
0

 $=$

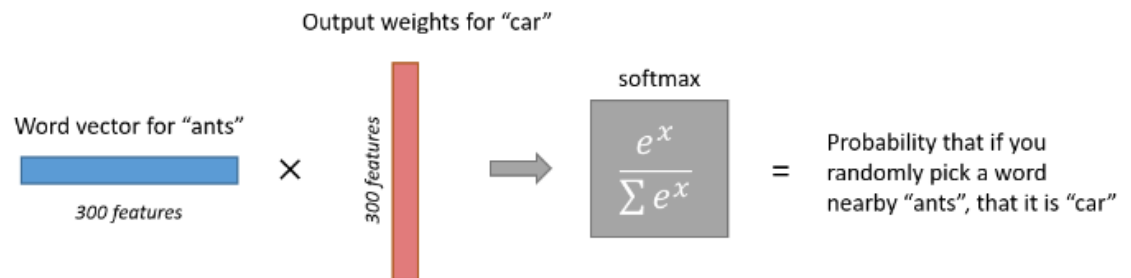
w00
w10
w20

$V_{3 \times 1}$

5. Una vez teniendo los Word vectors y la matriz de pesos, se puede obtener vectores con la probabilidad de ser vecino de la palabra en cuestión.



- a. Esto por medio de softmax, aplicando a cada palabra y así generando una probabilidad por todos los vectores.



¿Para qué funciona?

Todo este proceso sirve para dada una palabra y el contexto que se recabo del texto, se puede dar una múltiple solución sobre que palabras podrían estar antecediendo o procediendo a la entrada, esto genera una contextualización de las palabras (semántica).

Aplicación

Una sencilla sería un sistema de autocompletado mucho más eficiente y adaptativo al usuario, esto recabando la información que el mismo escribe y de los sitios que frecuenta.

En cuanto a la traducción de textos mejoraría bastante al evitar la traducción directa de la palabra y dando mayor sentido a la traducción completa, además de mayor cohesión.