Word Embedding 補充

2019/7/5

Word2Vec

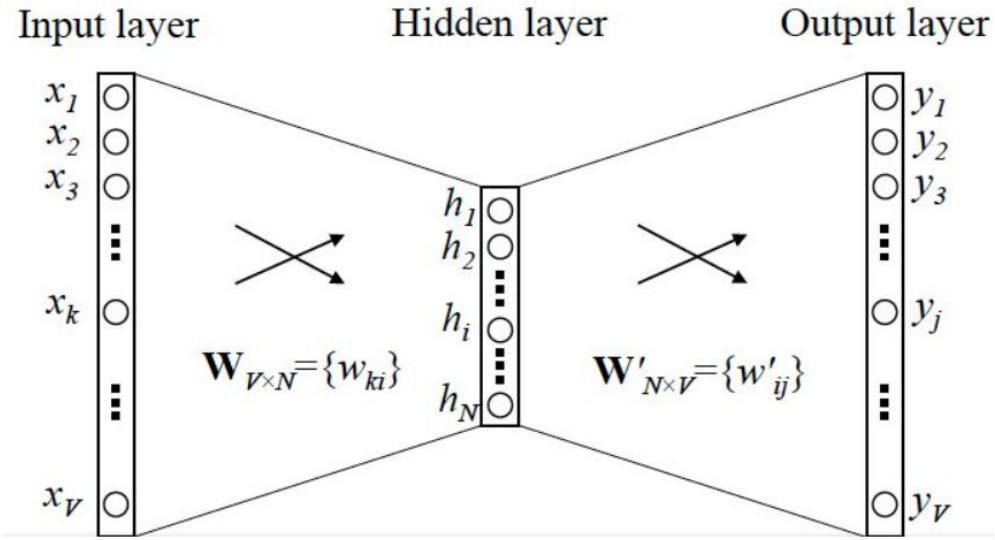
前言

f(x)->y

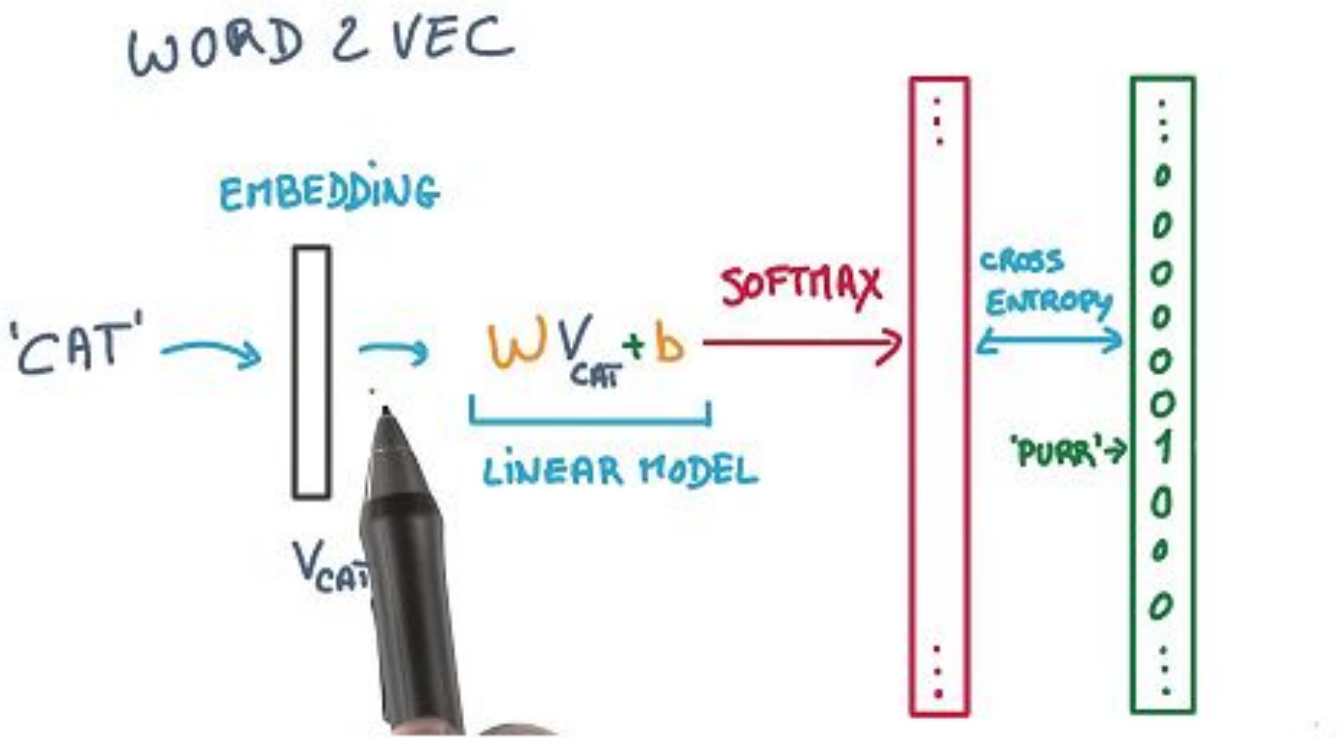
- x: 詞彙, y: 這個詞彙的上下文詞彙
- 希望NN可以讓x、y是合理的
- Word2vec 要拿到NN裡的第一個hidden layer

Word2Vec

- 輸入為一個x的one-hot encoder
- 降維—把詞彙從one-hot 形式降維到Word2vec形式



Word2Vec


WORD 2 VEC
EMBEDDING
'CAT'
V_{CAT}
$WV_{CAT}+b$
LINEAR MODEL
SOFTMAX
CROSS ENTROPY
'PURR'

Bag-of-word

What is bag-of-word?

- 字詞出現在documents的次數
- 未考慮字詞的**順序** (每個詞的出現都是獨立的)

For instance

John likes to watch movies. Mary likes too.

John also likes to watch football games.

dictionary

{"John": 1, "likes": 2, "to": 3, "watch": 4, "movies": 5, "also": 6, "football": 7, "games": 8, "Mary": 9, "too": 10}

第i個元素代表第i個單字在句子中出現的次數

[1, 2, 1, 1, 1, 0, 0, 0, 1, 1]

[1, 1, 1, 1, 0, 1, 1, 1, 0, 0]

適用場景

- 假設有一個document set, 裡面共有M個documents, 所有document的單字建立一個包含N個單字的字典, 經過Bag-of words, 每個document都可以被表示為N維向量