自然語言處理與應用

Natural Language Processing and Applications

Word Embeddings

Instructor: 林英嘉 (Ying-Jia Lin)

2025/03/03

Course GitHub

Slido # NLPA

生成式AI融合教學

- 以組為單位
  - ChatGPT Plus
  - Google Colab Pro 雲端運算付費服務
  - OpenAI API 呼叫大型語言模型進行生成式服務
- 可以開始找組員，W3 or W4 開始受理核銷 (暫定)

自然語言處理基本功 (如何表達字詞)

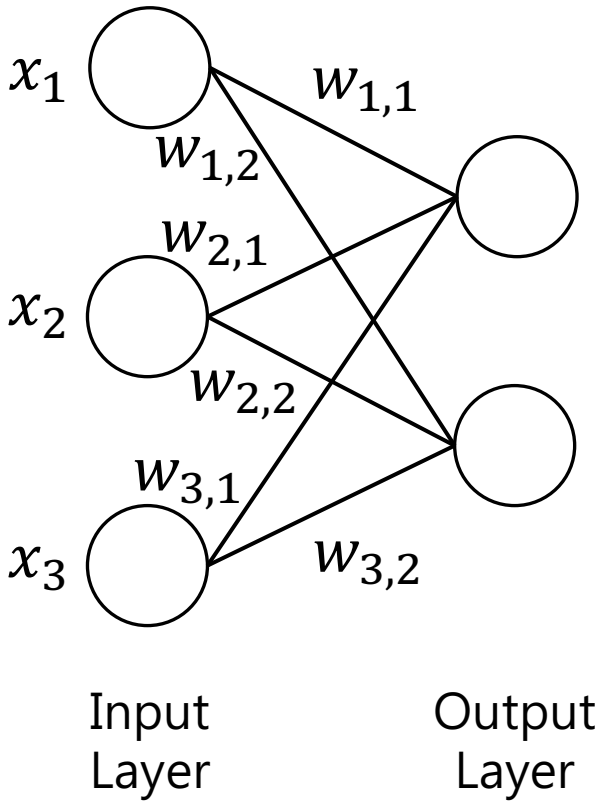
- Week 1 – Week 3
  - 自然語言處理介紹
  - 統計語言模型與基本詞向量方法
  - 詞嵌入模型

Outline

- Word2vec (Mikolov et al., 2013)
  - Softmax
  - Hierarchical softmax
  - Negative Sampling
- GloVe (Pennington et al., 2014)

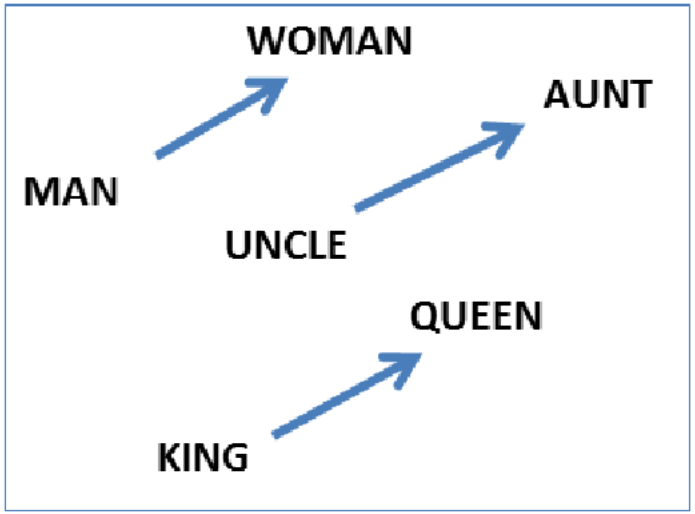
Distributed Representations

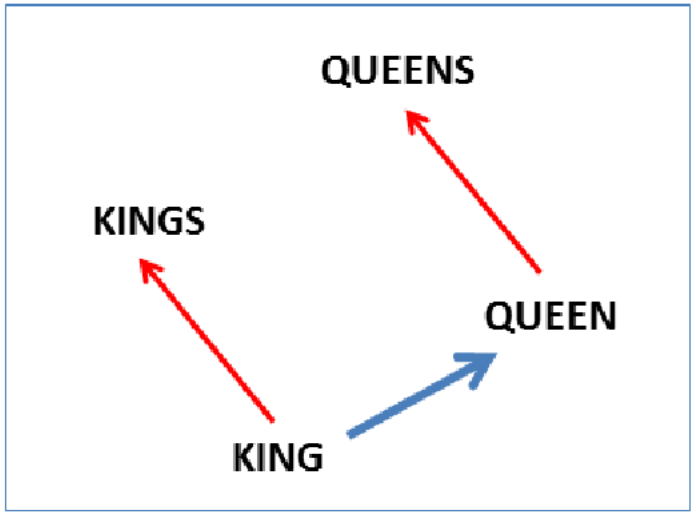
The natural expression of distributed representations in a neural net is to make each concept be a single unit and to use the connections between units to encode the relationships between concepts.



Hinton, Geoffrey E. "Learning distributed representations of concepts." Proceedings of the Annual Meeting of the Cognitive Science Society. Vol. 8. 1986.

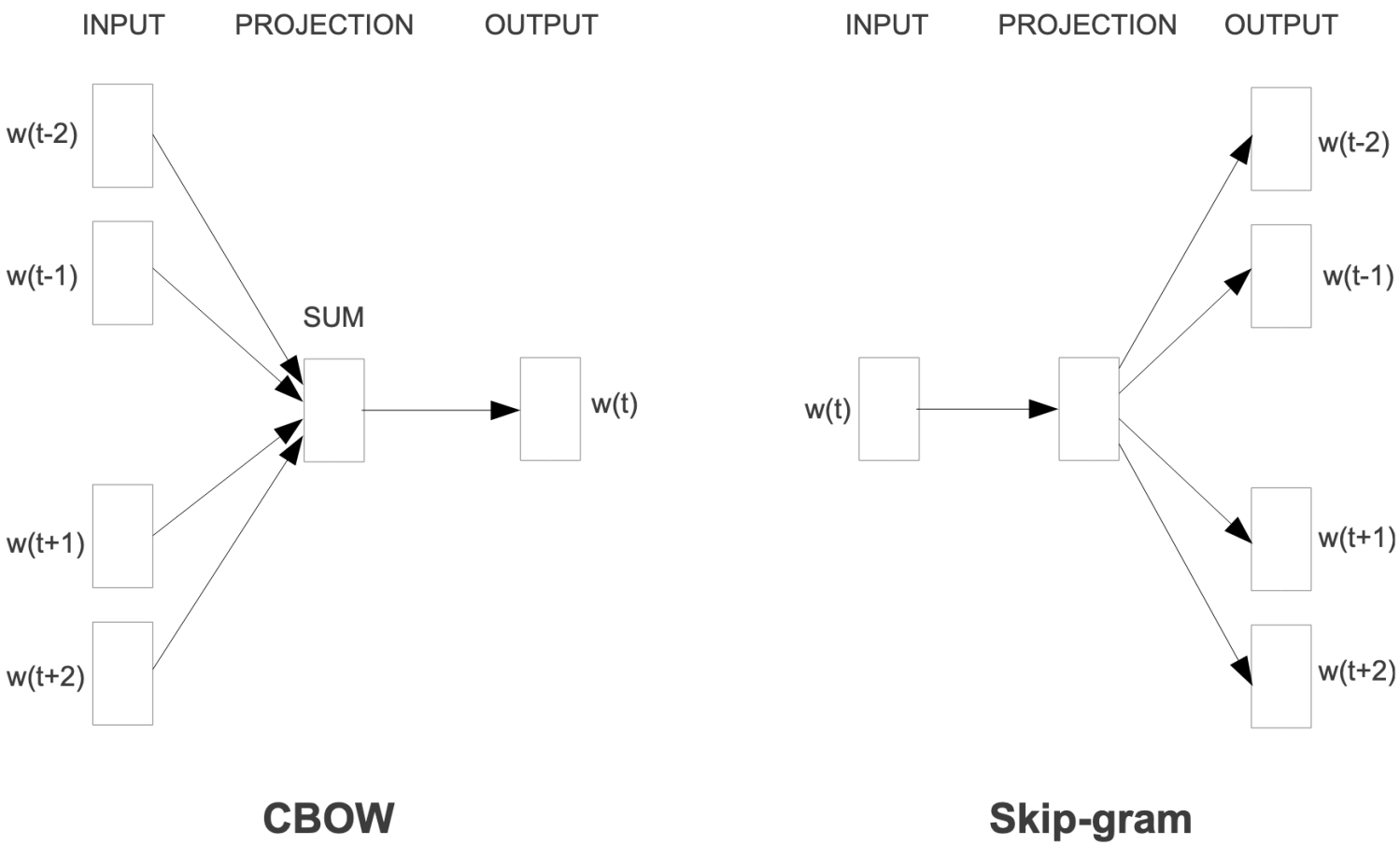
Word pairs illustrating the gender relation





Mikolov et al. “Linguistic regularities in continuous space word representations.” NAACL 2013.

[Overview] The Word2vec Approaches

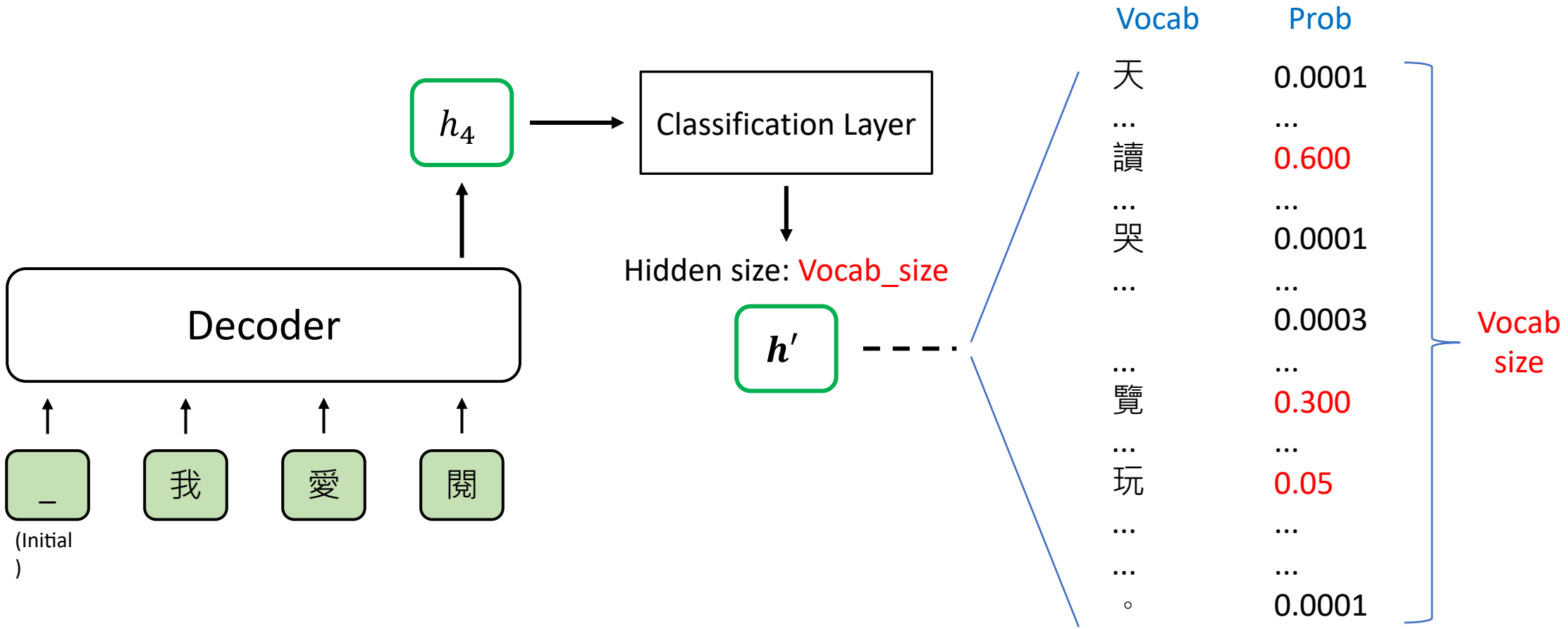


Mikolov et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

The Two Word2vec Papers

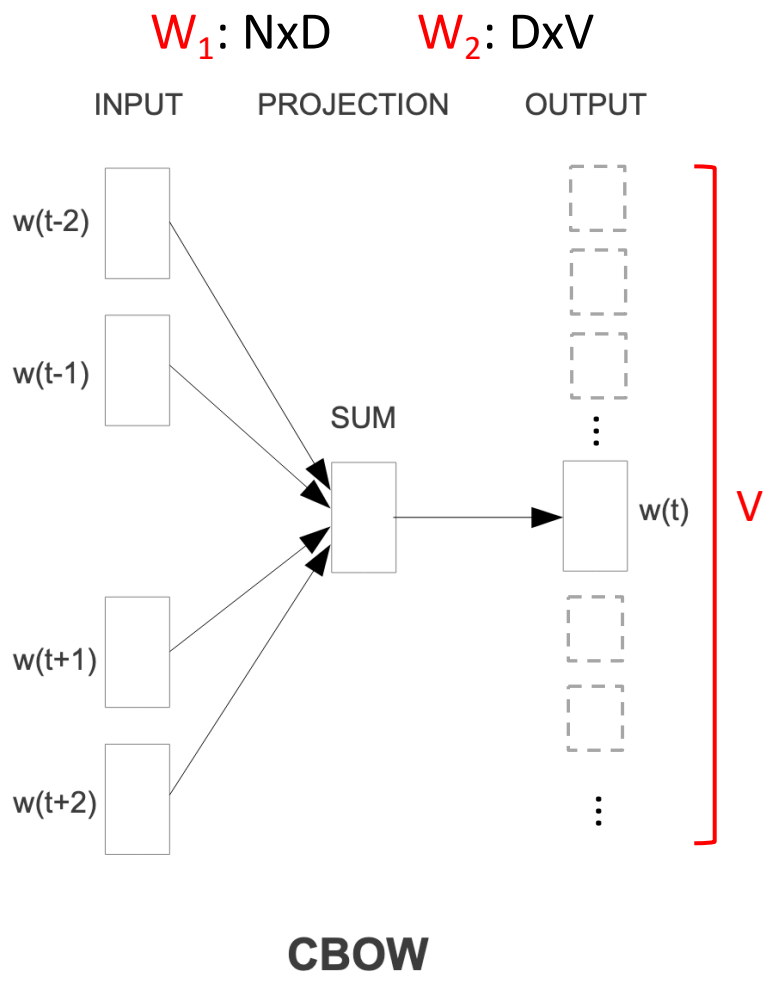
- Mikolov et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).
  - 介紹 CBOW 以及 Skip-gram
- Mikolov et al. "Distributed representations of words and phrases and their compositionality." NeurIPS 2013.
  - 針對 Skip-gram 來進行詳細說明，並且介紹加速訓練的方法：
    - (1) Hierarchical Softmax (2) Negative Sampling

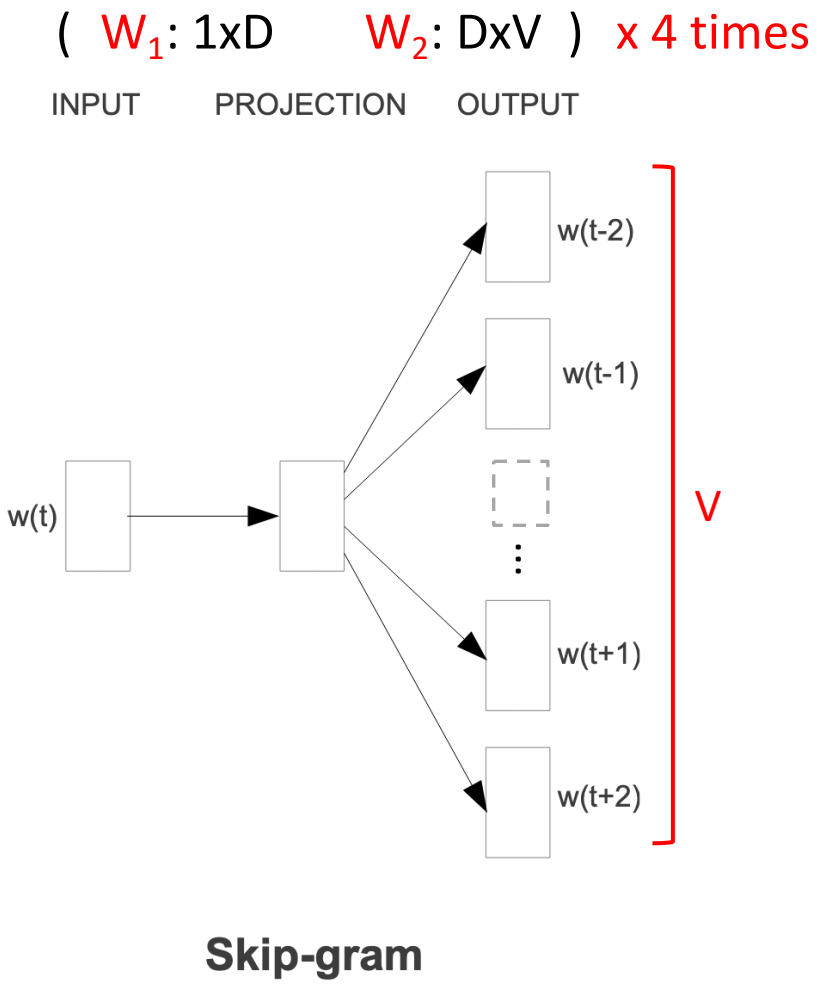
Text Generation Process



Dimensions in Word2vec

Symbol	Meaning	Example
N	輸入數量	4
D	Hidden size	100
V	字典大小	100,000





Mikolov et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

How to train? (Probabilities of Skip-gram)

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)$$

中間字預測周圍字的機率值

Symbol	Meaning
T	一個序列 $(w_1, w_2, \ldots, w_T)$ 有T個字
c	Context window 的範圍，以前一頁的例子來說：$c = 4$
w_t	中間的字 (代表 $t + j = 0$)
W	字典大小

為什麼要取 log?

考慮到一些極端數值： $10^4 - 10^3$

如果取log之後： $\log 10^4 - \log 10^3 = 4 - 3$

- 會讓數值範圍變小，使機器學習的最佳化過程穩定
- 會降低極端值的影響，避免 outliers 影響訓練結果

https://datascience.stackexchange.com/questions/40089/what-is-the-reason-behind-taking-log-transformation-of-few-continuous-variables

How to train? (Softmax of Skip-gram)

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)$$

中間字預測周圍字的機率值

Symbol	Meaning
T	一個序列 $(w_1, w_2, \ldots, w_T)$ 有T個字
c	Context window 的範圍，以前一頁的例子來說：$c = 4$
w_t	中間的字 (代表 $t + j = 0$)

Softmax function

$$p(w_O|w_I) = \frac{\exp\left({v'_{w_O}}^{\top} v_{w_I}\right)}{\sum_{w=1}^{W}\exp\left({v'_w}^{\top} v_{w_I}\right)}$$

內積

Symbol	Meaning
v'_{w_O}	輸出字的向量 (output word vector)
v_{w_I}	輸入字的向量 (Input word vector)
w_O	輸出字 (output word)
w_I	輸入字 (input word)
W	字典大小

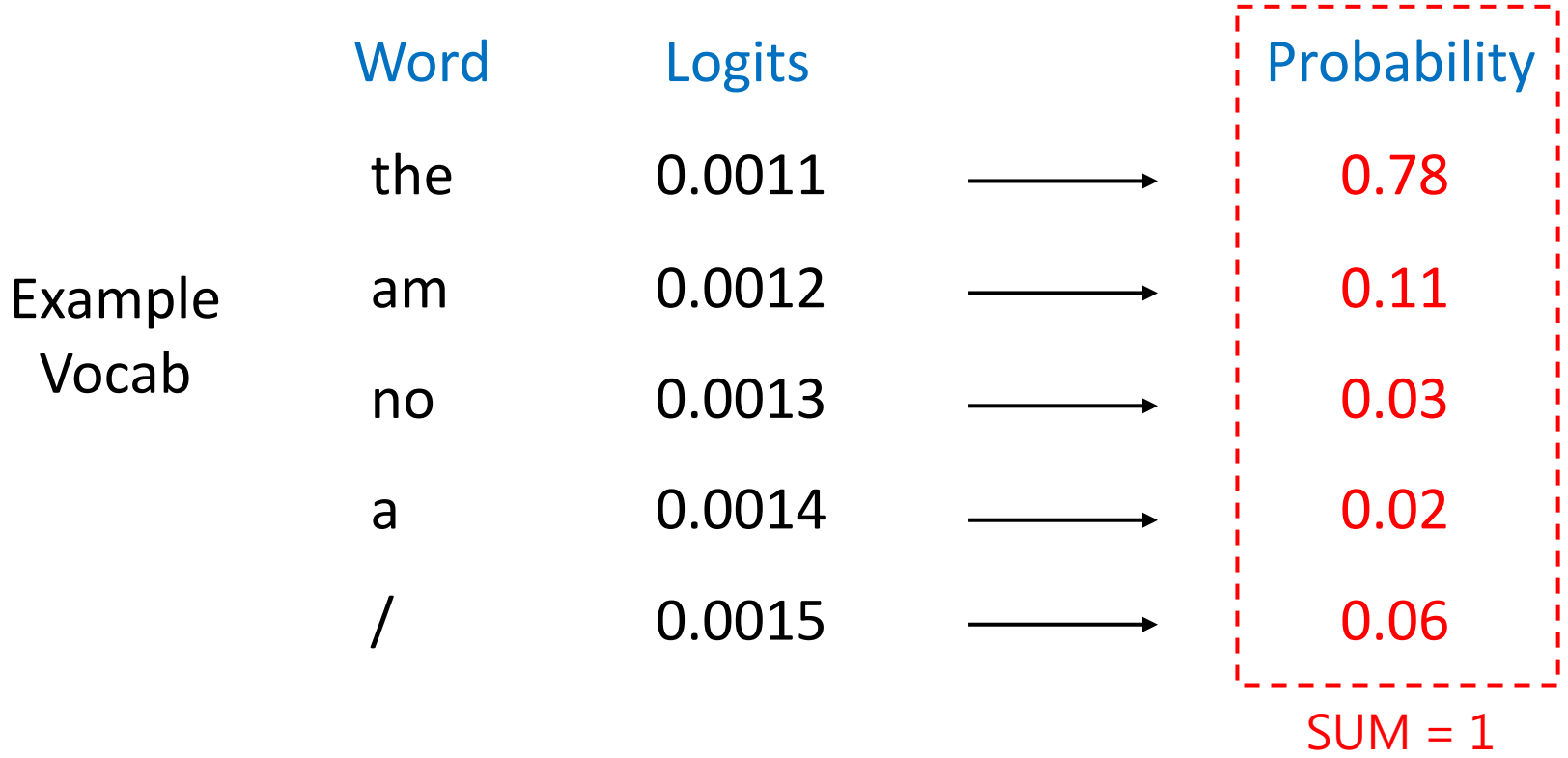
Softmax

- When generating the next word, `softmax` is performed to get the probabilities among the words in the vocabulary.
- Softmax formula:

$$\frac{\exp(u_l/\mathrm{t})}{\sum_{l'}^{|V|}\exp\left(u_{l'}/t\right)}$$

u_l: logits (model outputs before softmax)
$|V|$: size of the vocabulary
t: softmax temperature

Softmax



- Note that softmax is required for every decoding strategies since we need to find out the next word from a vocabulary.

Advanced Techniques of Word2vec

- Speed up the training process of Word2vec
- Hierarchical Softmax
- Negative Sampling

Mikolov et al. "Distributed representations of words and phrases and their compositionality." NeurIPS 2013.

Hierarchical Softmax Overview

- 總目標：近似輸出層，因為在 Vocab size 很大的情況下做 Softmax 計算量很大
- 訓練流程：
  - Step1: 把字典的字根據頻率建立出一棵 Huffman Tree
  - Step2: 以 internal nodes 的權重 (vectors) 代替輸出層，從 root node 開始計算往下走每層的機率值 Softmax(dot_product)，每層都只做二元分類
  - Step3: 調整 internal nodes 的權重

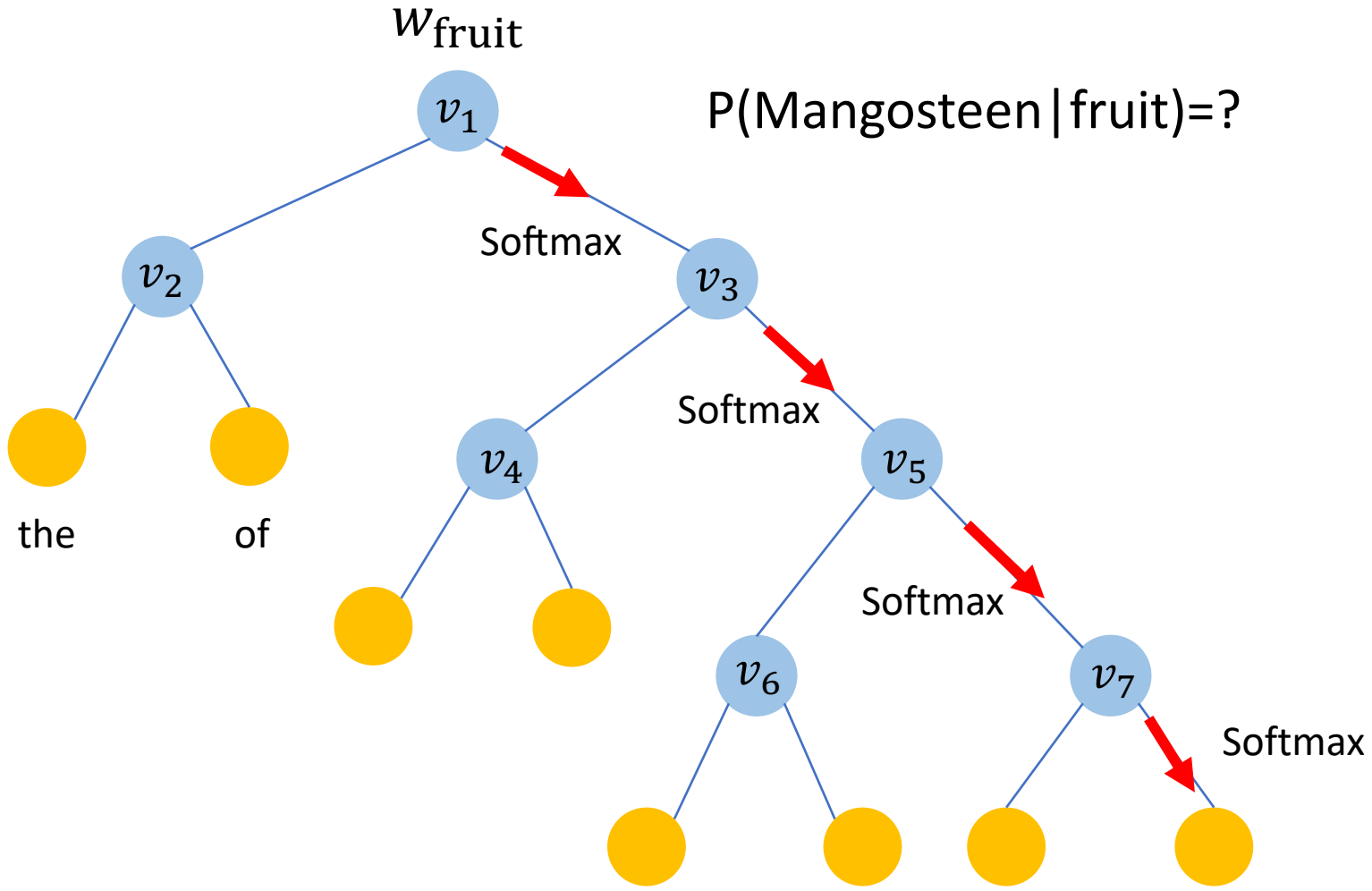
Huffman Tree 是一種最佳前綴編碼樹 (optimal prefix code tree)，其特性是：

- 頻率較高的詞會被分配較短的路徑，使得它們的預測成本較低。
- 頻率較低的詞會有較長的路徑，以減少總體的編碼長度。

Mikolov et al. "Distributed representations of words and phrases and their compositionality." NeurIPS 2013.

Hierarchical Softmax

Mikolov et al. "Distributed representations of words and phrases and their compositionality." NeurIPS 2013.



P(Mangosteen|fruit)=?

Internal nodes (帶有 weights)

字典中的字 (越上層詞頻越高，結構為 Huffman Tree)

訓練目標為最大化路徑中每個節點經過Softmax後結果相乘的機率值

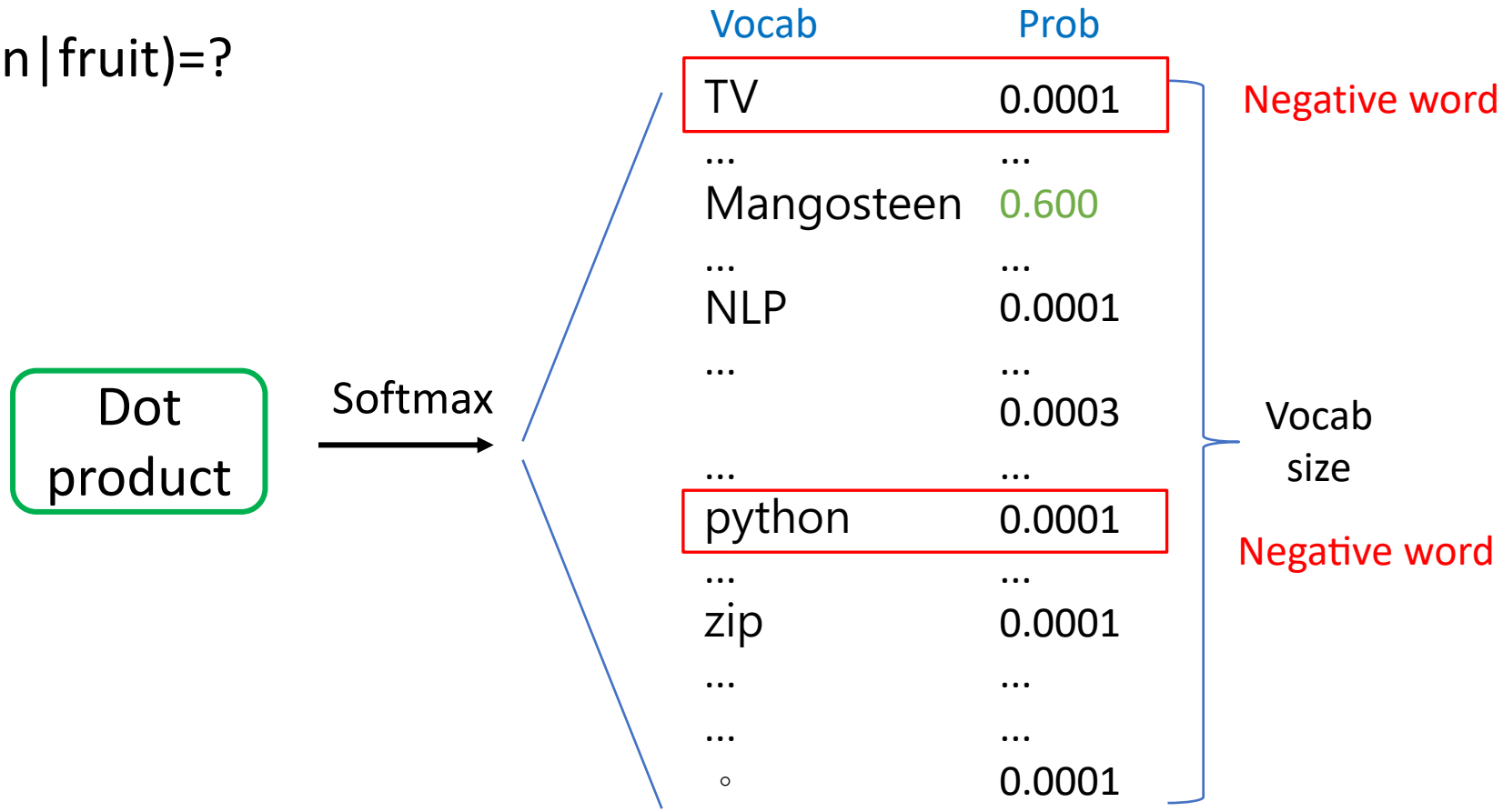
Negative Sampling Overview

- 總目標：訓練時，不更新字典中的每個字，僅更新採樣到的負向字
- P(Mangosteen|fruit)=?
  - TV 可能跟 fruit 無關
  - python 可能也跟 fruit 無關
- 如何採樣負向的字？論文中使用 Uniform Distribution 的小改版

Mikolov et al. "Distributed representations of words and phrases and their compositionality." NeurIPS 2013.

Negative Sampling

P(Mangosteen|fruit)=?



Mikolov et al. "Distributed representations of words and phrases and their compositionality." NeurIPS 2013.

Negative Sampling Details

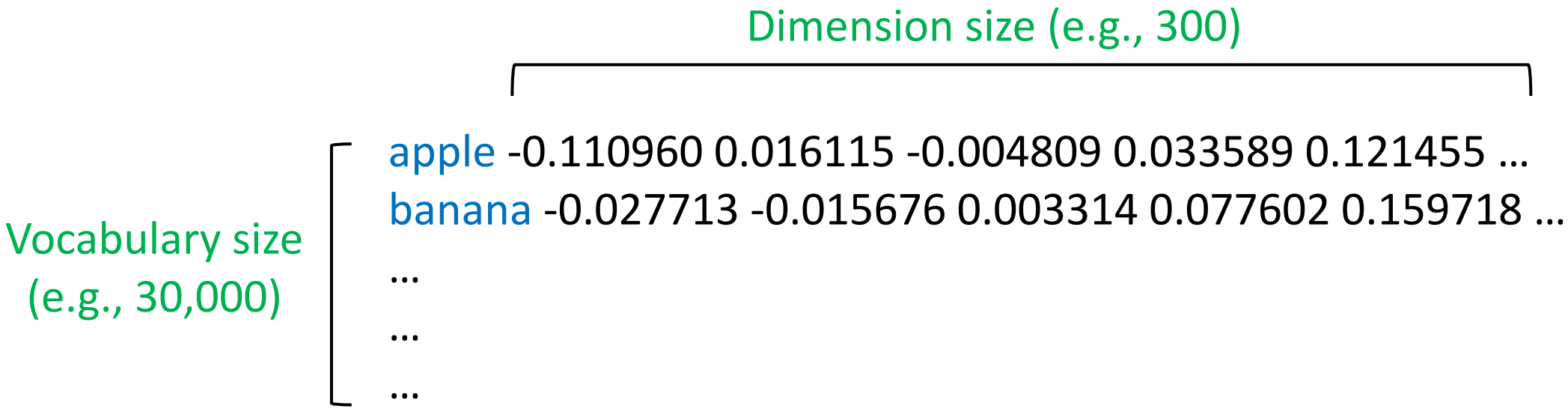
- Negative words 要設定成多少呢？
  - 5–20 are useful for small training datasets
  - 2–5 for large training datasets
  - (Empirical setting; Hyperparameter)

https://radimrehurek.com/gensim/models/word2vec.html

Mikolov et al. "Distributed representations of words and phrases and their compositionality." NeurIPS 2013.

After Training Word Embeddings

- The outputs are dense vectors with a fixed dimension size:

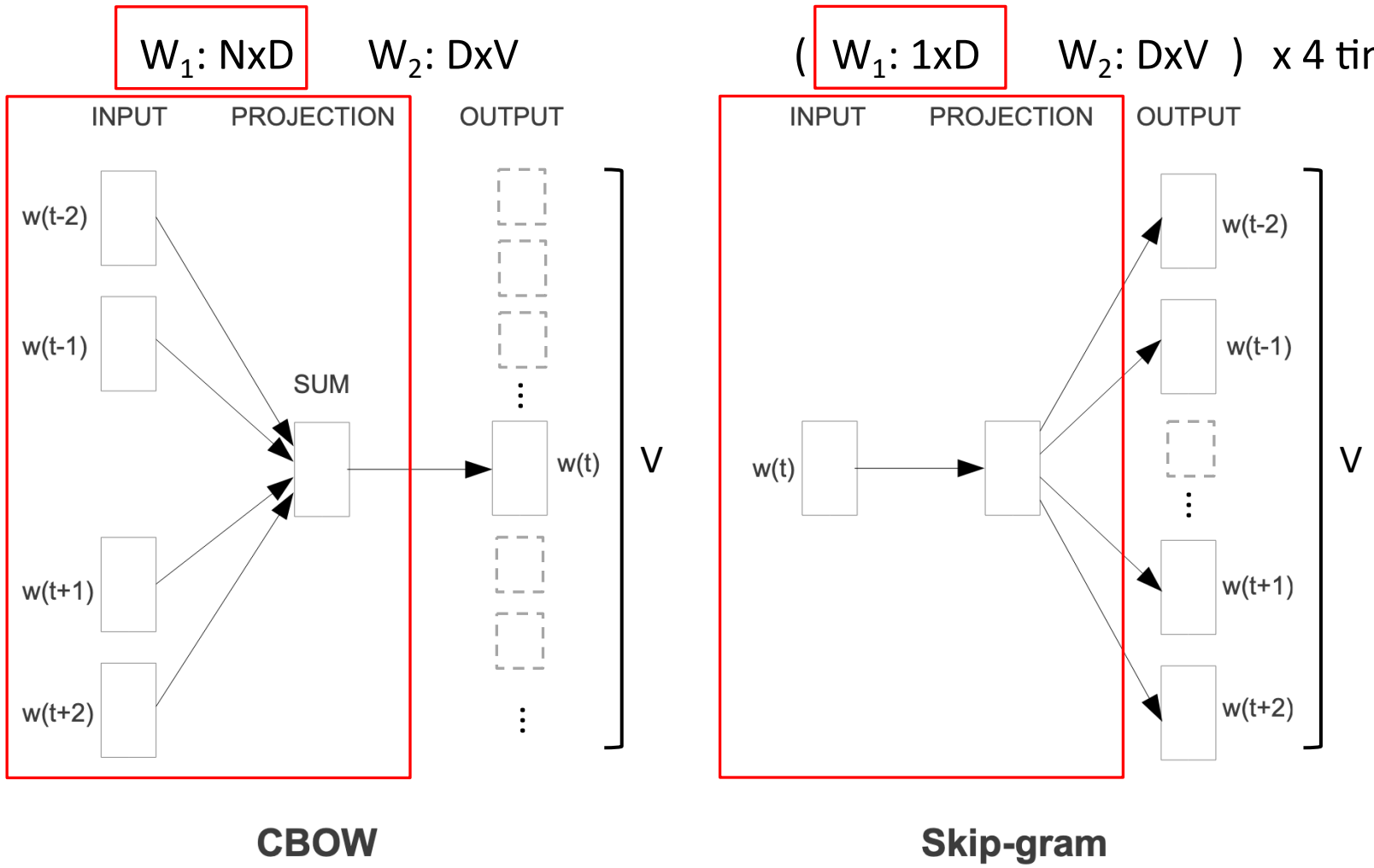


How to get these values?

Projection Matrix as Word Embeddings

Symbol	Meaning	Example
N	輸入數量	4
D	Hidden size	100
V	字典大小	100,000

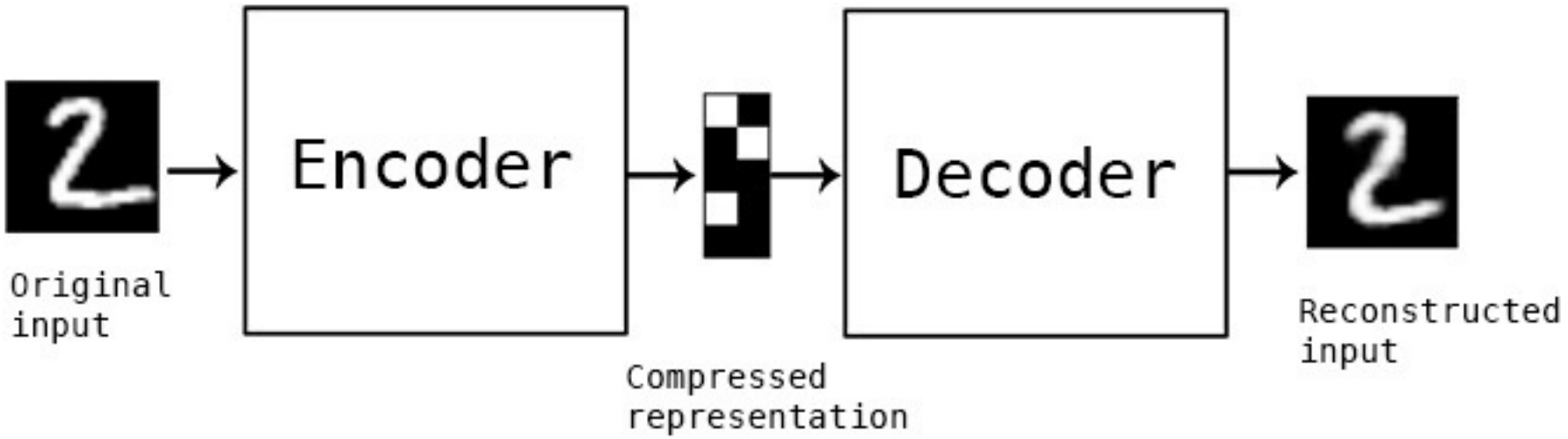
Word embeddings after training



Mikolov et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

Dimensionality Reduction

以 AutoEncoder 為例



類似於近年的 self-supervised learning

Figure source: https://blog.keras.io/img/ae/autoencoder_schema.jpg

GloVe

- Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." EMNLP 2014.

➡️ Figure source: https://nlp.stanford.edu/projects/glove/

1. Nearest neighbors

The Euclidean distance (or cosine similarity) between two word vectors provides an effective method for measuring the linguistic or semantic similarity of the corresponding words. Sometimes, the nearest neighbors according to this metric reveal rare but relevant words that lie outside an average human's vocabulary. For example, here are the closest words to the target word *frog*:

0. *frog*
1. frogs
2. toad
3. litoria
4. leptodactylidae
5. rana
6. lizard
7. eleutherodactylus


3. litoria


4. leptodactylidae

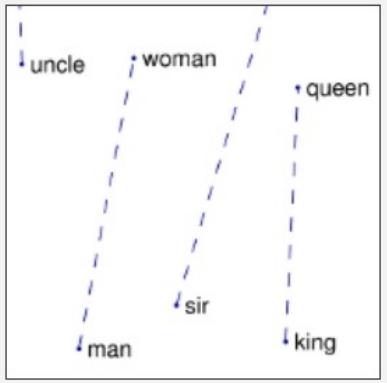

5. rana


7. eleutherodactylus

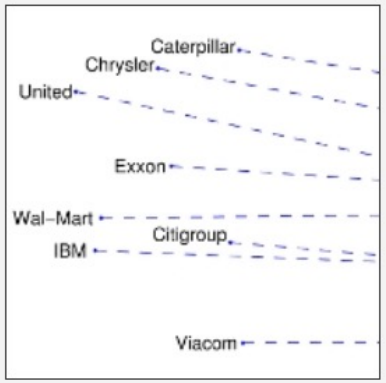
2. Linear substructures

The similarity metrics used for nearest neighbor evaluations produce a single scalar that quantifies the relatedness of two words. This simplicity can be problematic since two given words almost always exhibit more intricate relationships than can be captured by a single number. For example, *man* may be regarded as similar to *woman* in that both words describe human beings; on the other hand, the two words are often considered opposites since they highlight a primary axis along which humans differ from one another.

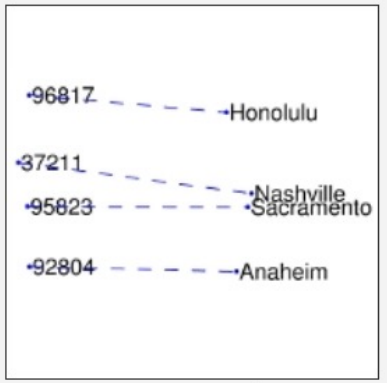
In order to capture in a quantitative way the nuance necessary to distinguish *man* from *woman*, it is necessary for a model to associate more than a single number to the word pair. A natural and simple candidate for an enlarged set of discriminative numbers is the vector difference between the two word vectors. GloVe is designed in order that such vector differences capture as much as possible the meaning specified by the juxtaposition of two words.



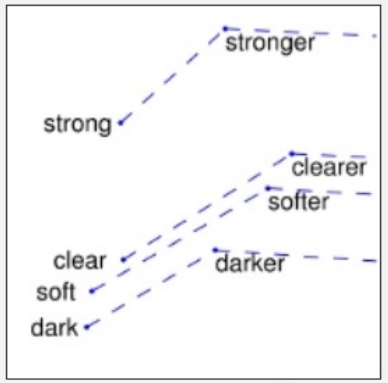
man - woman



company - ceo



city - zip code



comparative - superlative

Keypoints of GloVe

- GloVe as its name: Global Vectors
  - Word2vec uses local context window during learning.
- GloVe integrates the concepts of co-occurrence matrices into neural networks (NN).
  - Word2vec learns word vectors with NN only.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." EMNLP 2014.

Co-occurrence probabilities (1)

GloVe的設計理念

- ice + k (k=solid / gas / water / fashion)
- steam + k (k=solid / gas / water / fashion)

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

solid跟ice有關，但跟
steam較無關->相除的值大

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." EMNLP 2014.

Co-occurrence probabilities (2)

GloVe的設計理念

- ice + k (k=solid / gas / water / fashion)
- steam + k (k=solid / gas / water / fashion)

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

gas跟steam有關，但跟ice
較無關->相除的值小

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." EMNLP 2014.

Co-occurrence probabilities (3)

GloVe的設計理念

- ice + k (k=solid / gas / water / fashion)
- steam + k (k=solid / gas / water / fashion)

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

water跟ice和steam都有關
fashion跟ice和steam都無關
相除的值接近1

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." EMNLP 2014.

Summary for the Co-occurrence probabilities

GloVe的設計理念

- Compared to the raw probabilities, the ratio is better able to distinguish relevant words (solid and gas) from irrelevant words (water and fashion)
  - 強調的是「詞與詞的相對關係」而不是「詞的絕對出現頻率」
  - 詞與詞的相對關係也能夠帶來更多的語意資訊

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." EMNLP 2014.

如何設計 GloVe 的學習目標

GloVe的設計理念

w_i: 第 i 個字的 word vector

w_j: 第 j 個字的 word vector

$\tilde{w}_k$: 接在第 i 或 j 個字的 word vector，代表 context word

Index	i	j	k
Word	ice	steam	solid

Eq (2) $$F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$$

詞向量存在線性特徵，如果最後想讓模型學到 $\frac{P_{ik}}{P_{jk}}$ 的話，可以使用相減做法

Eq (3) $$F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{P_{ik}}{P_{jk}}$$

注意：等號右邊是純量
可以使用內積來兩側與 $\tilde{w}_k$ 之間的關係

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." EMNLP 2014.

如何設計 GloVe 的學習目標

GloVe的設計理念

Eq (3) $$F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{P_{ik}}{P_{jk}}$$

理論上如果把 wi 跟 wk 交換 (或 wj 跟 wk 交換)，結果應該要是一致的。但因為 wi 跟 wj 目前是減法，所以結果不一致：

Index	i	j	k
Word	ice	steam	solid
Word’	solid	steam	ice

$$(w_i - w_j)^T \widetilde{w_k} \neq (\widetilde{w_i} - w_j)^T w_k$$

Eq (4) $$F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{F(w_i^T \tilde{w}_k)}{F(w_j^T \tilde{w}_k)}$$

因此把等號左邊改一下 (這邊還不用看等號右邊)

如何設計 GloVe 的學習目標

GloVe的設計理念

Eq (3) $$F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{P_{ik}}{P_{jk}}$$

Eq (4) $$F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{F(w_i^T \tilde{w}_k)}{F(w_j^T \tilde{w}_k)}$$

Eq (5) $$F(w_i^T \tilde{w}_k) = P_{ik} = \frac{X_{ik}}{X_i}$$

X_i 代表單獨出現第 i 個字的次數
X_{ik} 代表第 i 個字和第 k 個字一起出現的次數

Eq (6) $$w_i^T \tilde{w}_k = \log(P_{ik}) = \log(X_{ik}) - \log(X_i)$$ 所以 F 是 exp

如何設計 GloVe 的學習目標

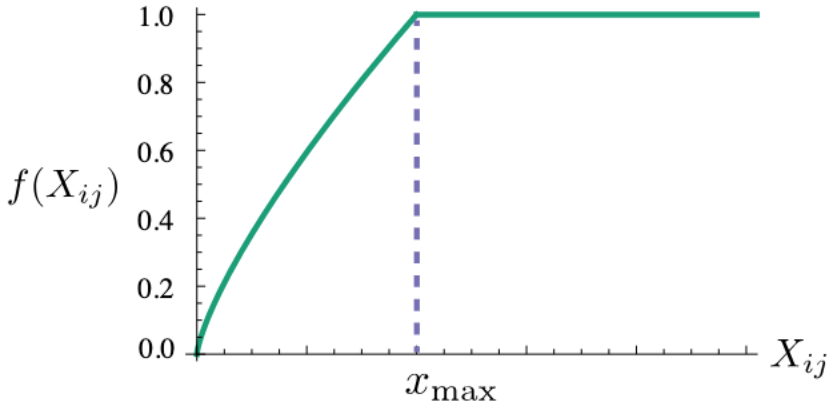
GloVe的目標函數

Eq (6) $w_i^T \tilde{w}_k = \log(P_{ik}) = \log(X_{ik}) - \log(X_i)$

推導至此 j 已經不見

真正重要的不是三個字之間的關係，只需要量測兩個字之間的關係就可以了

Eq (8) $J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$



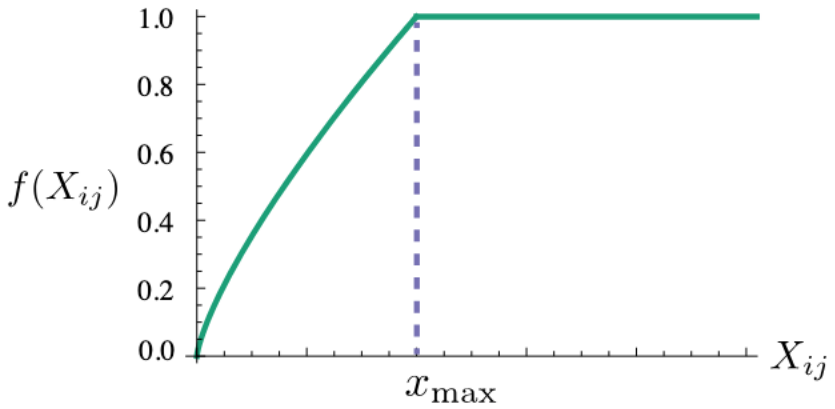
其中 b_i 和 b_j 是用來模擬 $\log(X_i)$ 的 bias terms (會被訓練的參數)
$f(X_{ij})$ 代表 X_{ij} 代表第 i 個字和第 j 個字一起出現的次數再經過某種處

如何設計 GloVe 的學習目標

GloVe的目標函數

Eq (8) $$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$$

$f(X_{ij})$ 代表 X_{ij} 代表第 i 個字和第 j 個字一起出現的次數再經過某種處理



$f(X_{ij})$ 滿足三種條件：

(1) $f(0) = 0$

(2) $f(X_{ij})$ 必須遞增，罕見詞的頻率不需要被懲罰

(3) $f(X_{ij})$ 必須趨於平滑，使常見詞的權重有上限

Summary

- Two main word embedding approaches today:
- Word2vec uses local context window during learning.
- GloVe integrates the concepts of co-occurrence matrices into neural networks (NN).
  - Word2vec learns word vectors with NN only.

Additional Resource

- Word2vec 理論
  - http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/
  - (Hierarchical Softmax) https://www.youtube.com/watch?v=pzyIWCelt_E
- Word2vec 實作
  - TensorFlow Word2vec
    - https://www.tensorflow.org/text/tutorials/word2vec#model_and_training

Thank you!

Instructor: 林英嘉

✉ yjlin@cgu.edu.tw

TA: 吳宣毅

✉ m1161007@cgu.edu.tw