# Distributed Representation

# Distributed representation : word2vec, doc2vec

**Discrete representation**

- One-hot vector/ Bag-of-word vector

$$\text{'dog'} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{'cat'} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad \text{Doc1} = \begin{bmatrix} 12 \\ 0 \\ 0 \\ 3 \\ 1 \\ 0 \\ 5 \end{bmatrix}$$

- Frequency 기반으로 표현하는 방법
- 구성 변수들을 직관적으로 이해 가능함
- 전처리 과정이 뚜렷하지 않으며 단어 빈도가 낮은 경우 중요하지 않게 판별됨

**Distributed representation**

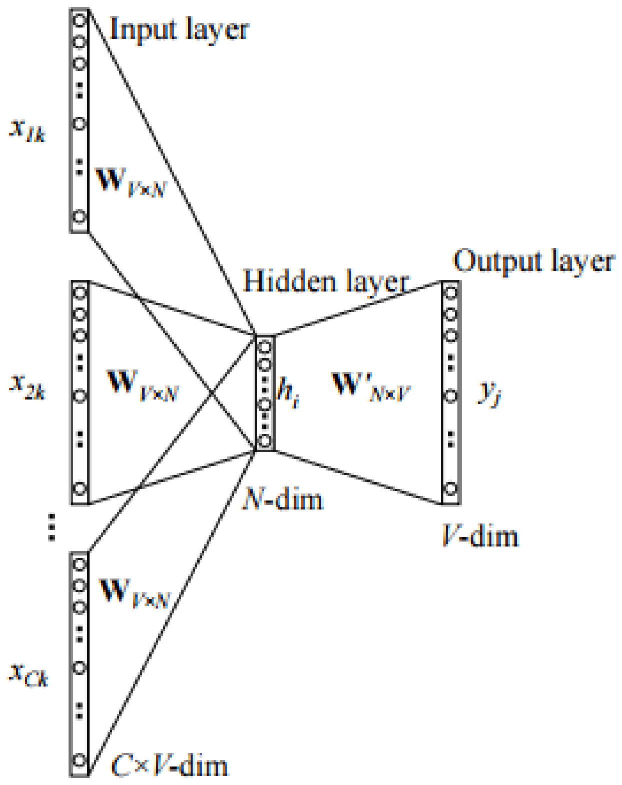
- Word2vec, Doc2vec

$$\text{'dog'} = \begin{bmatrix} 0.5 \\ 0.3 \\ -0.1 \\ 1 \end{bmatrix} \quad \text{'cat'} = \begin{bmatrix} 0.8 \\ -0.3 \\ -0.2 \\ 0.6 \end{bmatrix} \quad \text{Doc1} = \begin{bmatrix} 0.68 \\ 0.23 \\ 0.10 \\ -0.41 \\ 0.90 \\ 0.51 \\ -0.33 \end{bmatrix}$$

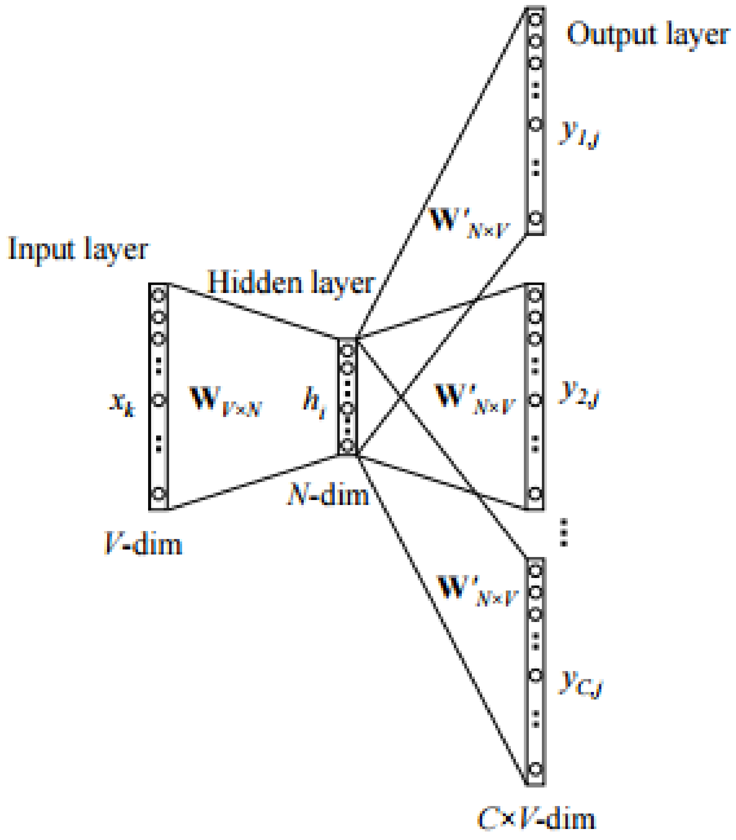
- Neural Network를 통해 continuous vector로 변환 가능
- 단어 별 유사도 계산가능
  - 'king' - 'man' + 'woman' → closest('Queen')

# Distributed representation : word2vec, doc2vec

- **Data :** All Bloomberg news articles from 2008 to 2015, # = 520,728




Continuous bag-of-word model




Skip-gram model

- $\mathrm{V} = \mathit{vocabulary\ size}$
- $\mathrm{C} = \mathit{window\ size}$
- $\mathrm{N} = \mathit{word2vec\ dimension}$

# Distributed representation : word2vec, doc2vec

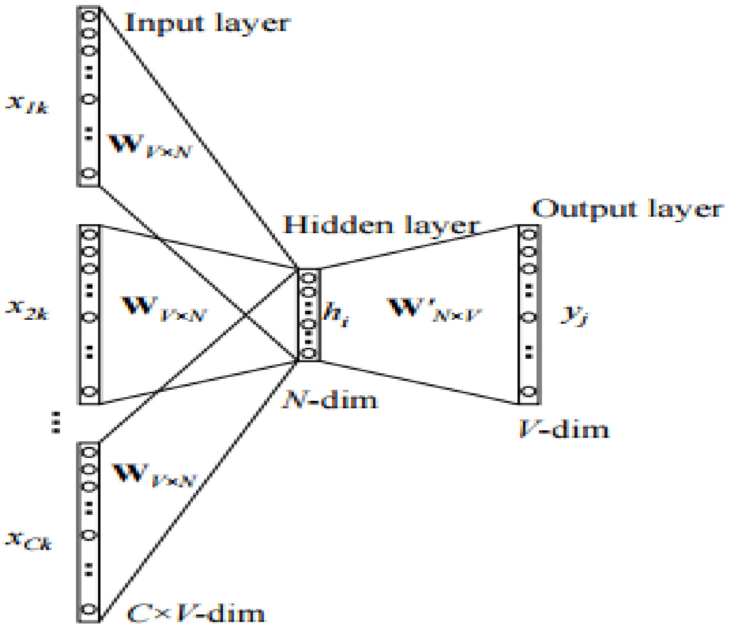
**Doc1** : { 'You are a very good boy.'

'You are a good girl too.' }

**Vocabulary** : { 'you' 'are' 'a' 'very' 'good' 'girl', 'too' }

$$\text{'you'} = \begin{bmatrix}1\\0\\0\\0\\0\\0\\0\end{bmatrix} \quad \text{'are'} = \begin{bmatrix}0\\1\\0\\0\\0\\0\\0\end{bmatrix} \quad \cdots \quad \text{'too'} = \begin{bmatrix}0\\0\\0\\0\\0\\0\\1\end{bmatrix}$$

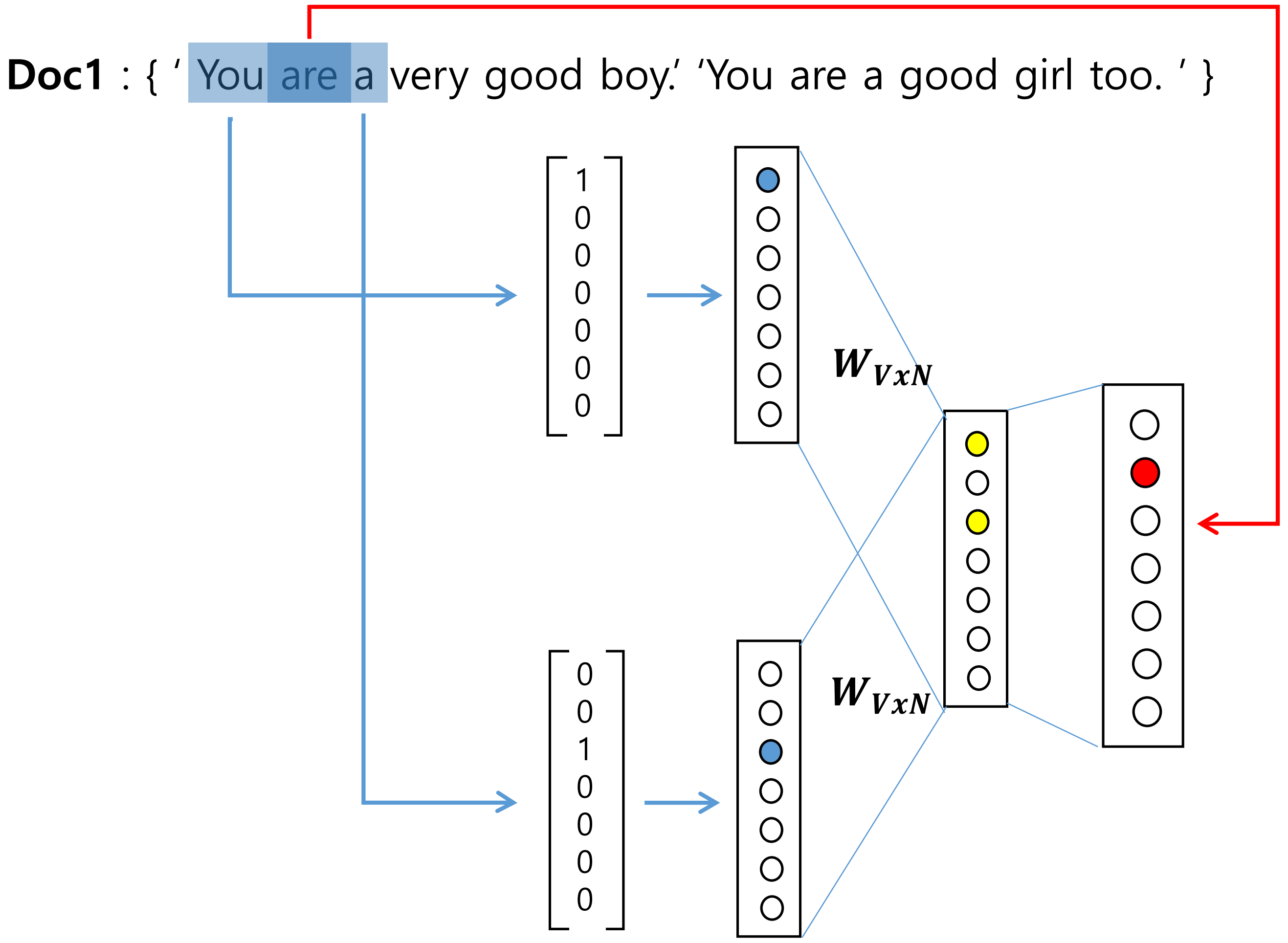
One-hot coding

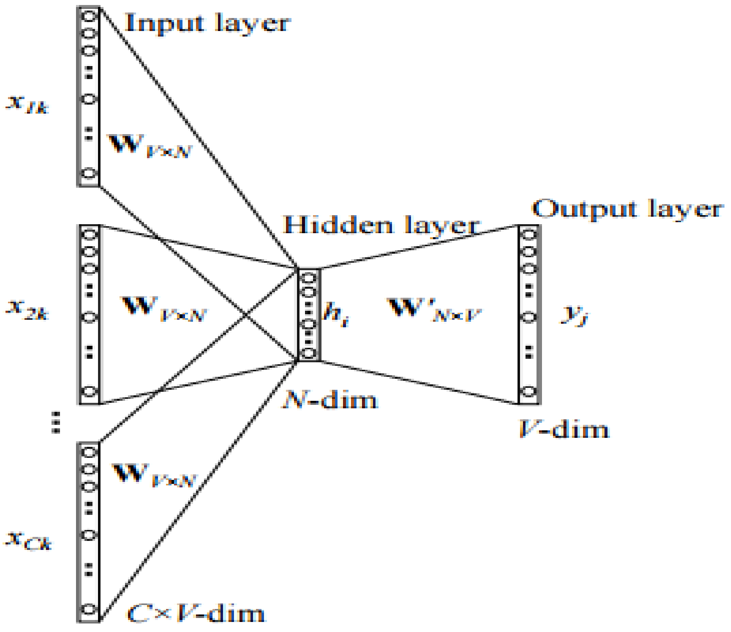



Continuous bag-of-word model

- $\mathrm{V} = vocabulary\ size$
- $\mathrm{C} = window\ size$
- $\mathrm{N} = word2vec\ dimension$

# Distributed representation : word2vec, doc2vec



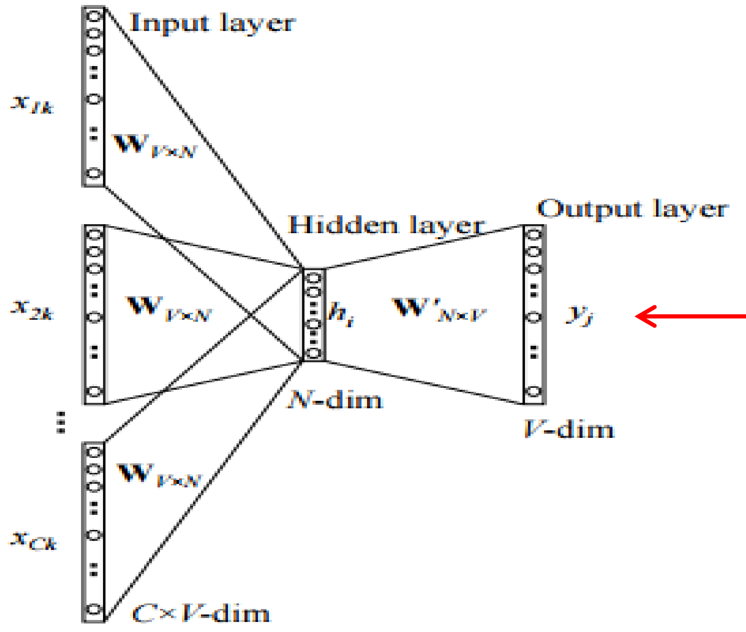




Continuous bag-of-word model

- $\mathrm{V} = vocabulary\ size$
- $\mathrm{C} = window\ size$
- $\mathrm{N} = word2vec\ dimension$

# Distributed representation : word2vec, doc2vec

**Doc1** : { ' You are a very good boy.' 'You are a good girl too. ' }

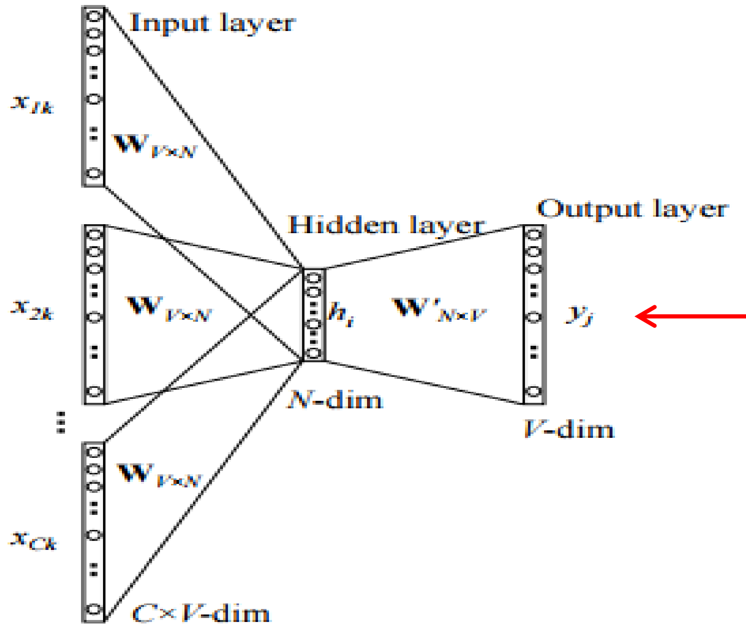



Continuous bag-of-word model

- $V = vocabulary\ size$
- $C = window\ size$
- $N = word2vec\ dimension$

# Distributed representation : word2vec, doc2vec

**Doc1** : { ' You are a very good boy.' 'You are a good girl too. ' }

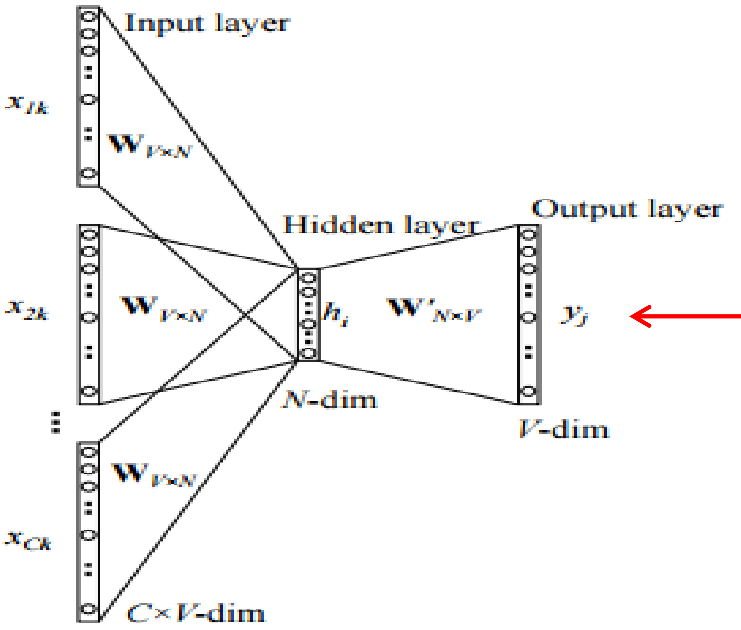



Continuous bag-of-word model

- $V = vocabulary\ size$
- $C = window\ size$
- $N = word2vec\ dimension$

# Distributed representation : word2vec, doc2vec

**Doc1** : { ' You are a very good boy.' 'You are a good girl too. ' }

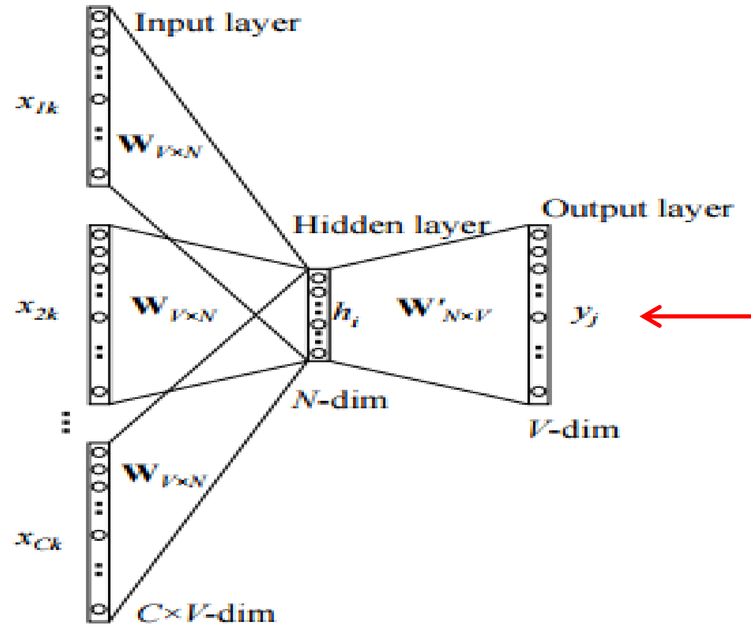



Continuous bag-of-word model

- $V = vocabulary\ size$
- $C = window\ size$
- $N = word2vec\ dimension$

# Distributed representation : word2vec, doc2vec

**Doc1** : { ' You are a very good boy.' 'You are a good girl too. ' }

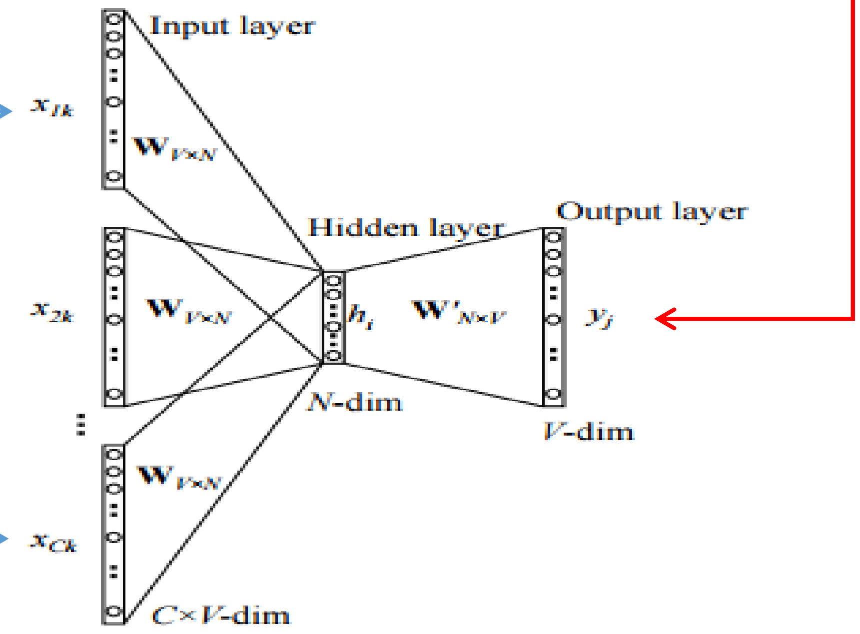



Continuous bag-of-word model

- $\mathrm{V} = vocabulary\ size$
- $\mathrm{C} = window\ size$
- $\mathrm{N} = word2vec\ dimension$

# Distributed representation : word2vec, doc2vec

**Doc1** : { ' You are a very good boy.' ' You are a good girl too. ' }

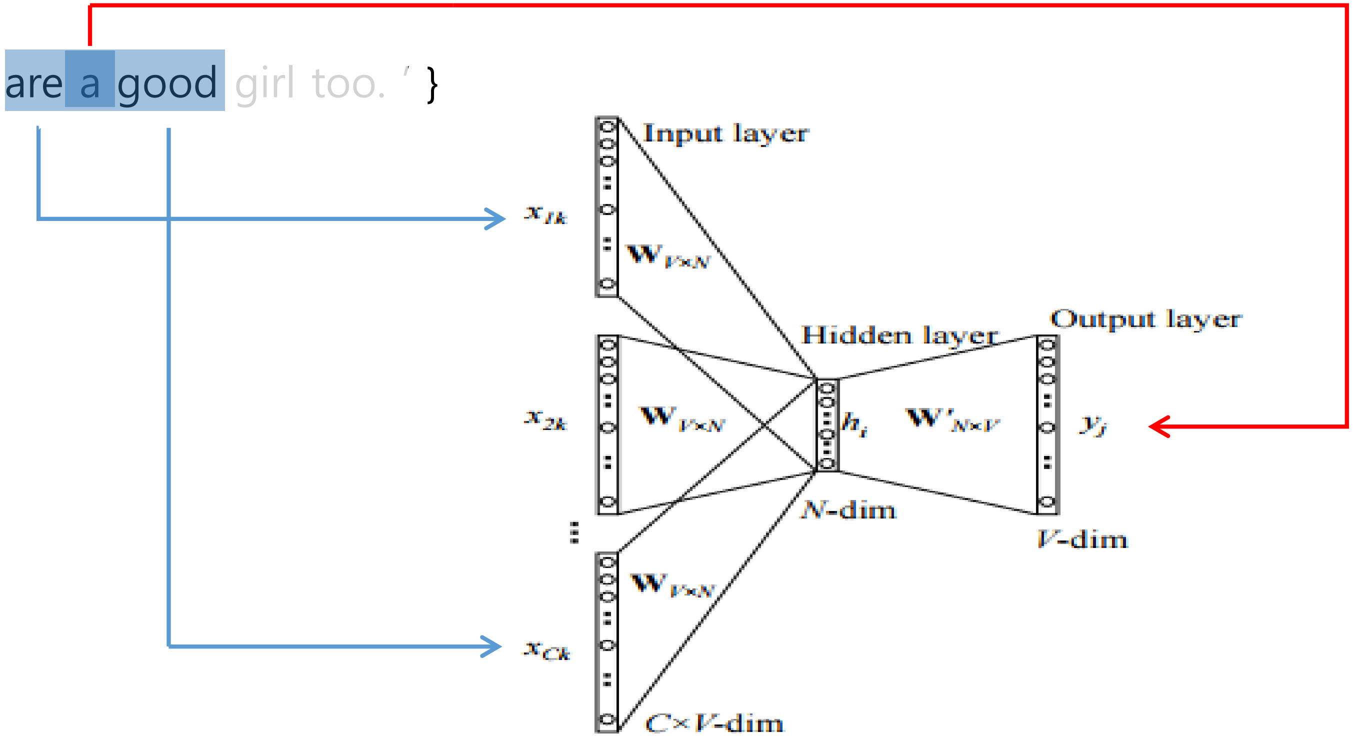



Continuous bag-of-word model

- $V = vocabulary\ size$
- $C = window\ size$
- $N = word2vec\ dimension$

# Distributed representation : word2vec, doc2vec

**Doc1** : { ' You are a very good boy.' ' You are a good girl too. ' }

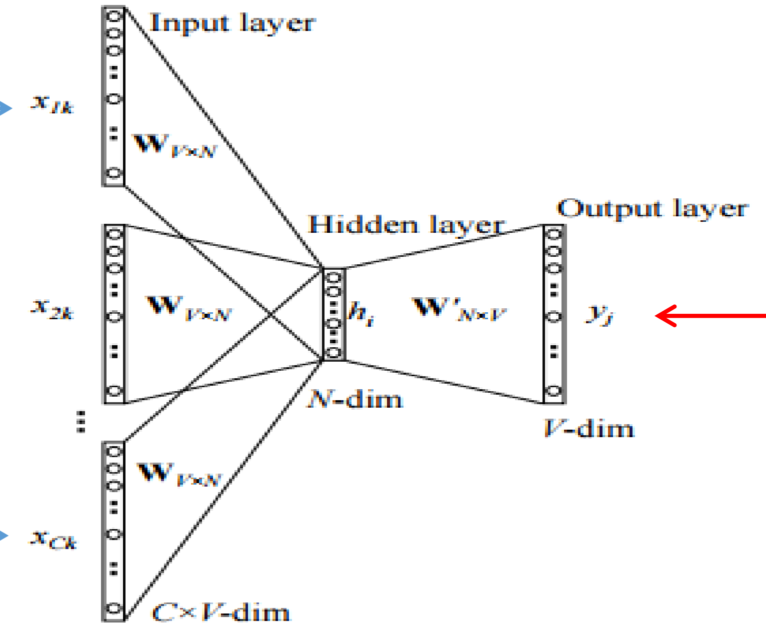



Continuous bag-of-word model

- $\mathrm{V} = vocabulary\ size$
- $\mathrm{C} = window\ size$
- $\mathrm{N} = word2vec\ dimension$

# Distributed representation : word2vec, doc2vec

**Doc1** : { ' You are a very good boy.' ' You are a good girl too. ' }

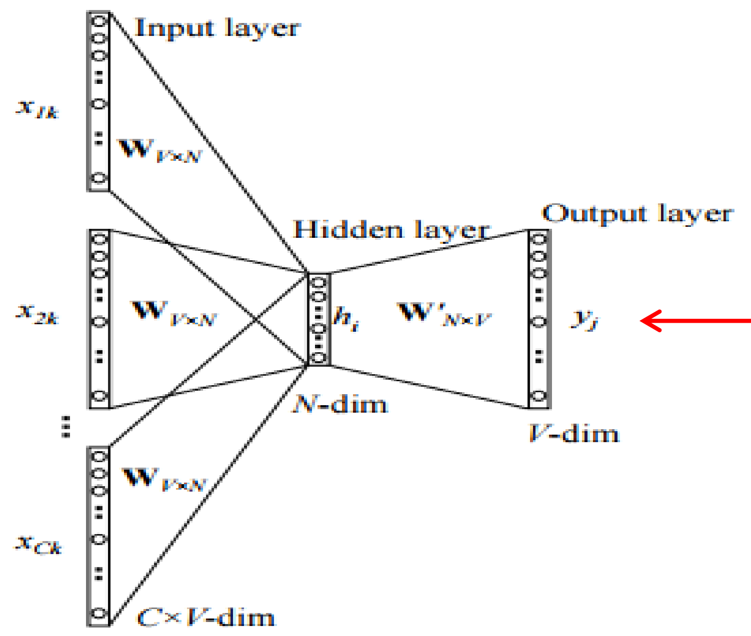



Continuous bag-of-word model

- $\mathrm{V} = vocabulary\ size$
- $\mathrm{C} = window\ size$
- $\mathrm{N} = word2vec\ dimension$

# Distributed representation : word2vec, doc2vec

**Doc1** : { ' You are a very good boy.' ' You are a good girl too. ' }




Continuous bag-of-word model

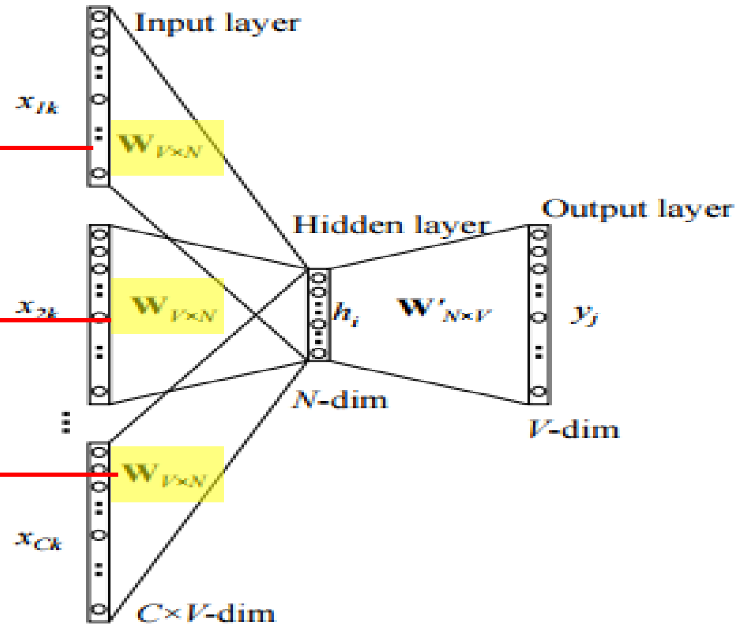
- $\mathrm{V} = vocabulary\ size$
- $\mathrm{C} = window\ size$
- $\mathrm{N} = word2vec\ dimension$

# Distributed representation : word2vec, doc2vec

**Doc1** : { ' You are a very good boy.' 'You are a good girl too. ' }

$W_{VxN}$

| | X1 | X2 | X3 | X4 | X5 | | Xn |
|---|---|---|---|---|---|---|---|
| **Word 1** | 0.345 | 0.121 | -0.538 | 1.011 | 2.011 | | 0.004 |
| **Word2** | 0.445 | 2.101 | 1.054 | -0.181 | -0.114 | | 0.764 |
| **…** | … | … | … | … | … | | … |
| **word V** | 0.334 | -0.087 | -0.407 | 1.114 | 0.554 | | 0.674 |




Continuous bag-of-word model

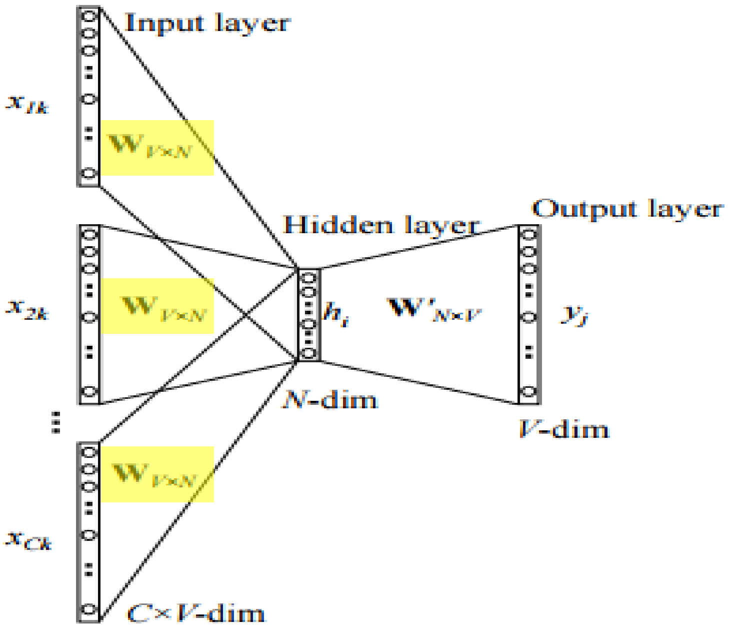
- $V = vocabulary\ size$
- $C = window\ size$
- $N = word2vec\ dimension$

# Distributed representation : word2vec, doc2vec

**Doc1** : { ' You are a very good boy.' 'You are a good girl too. ' }

$W_{VxN}$

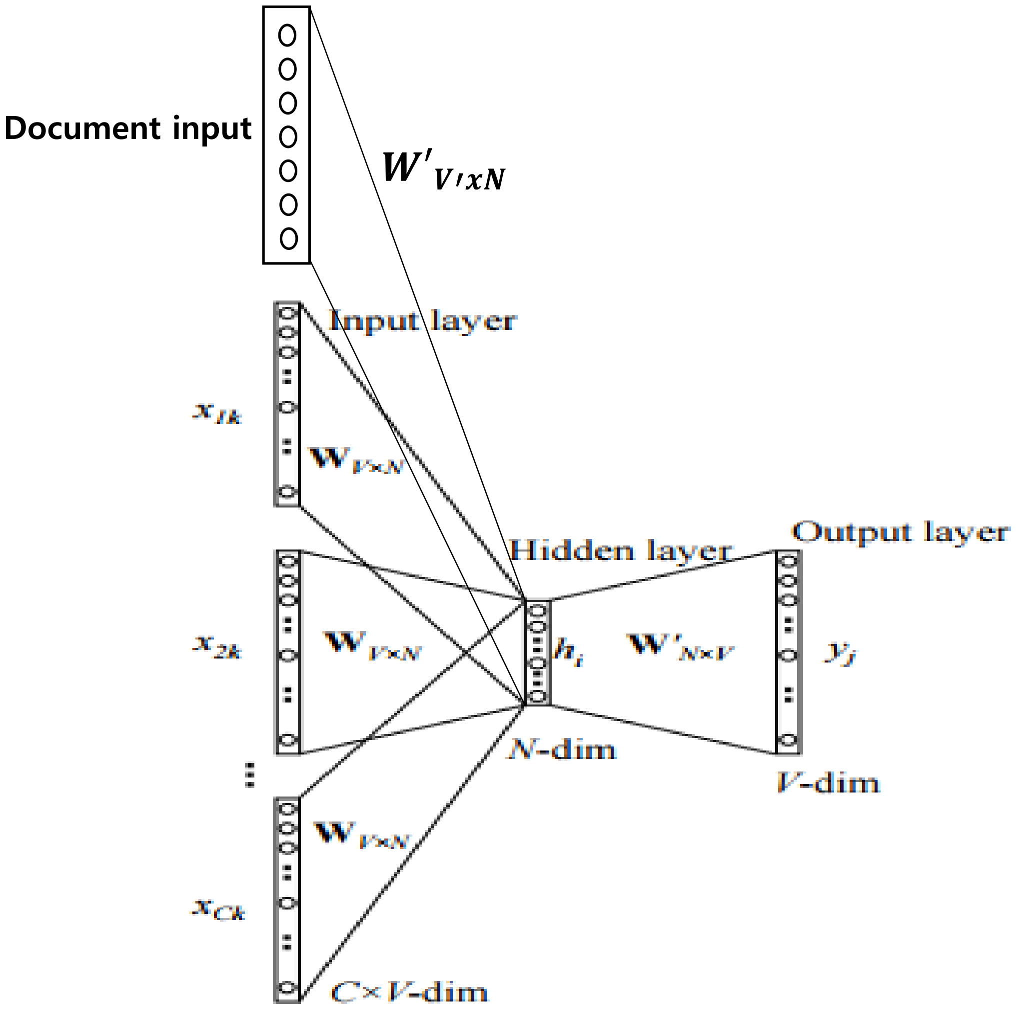
| | X1 | X2 | X3 | X4 | X5 | | Xn |
|---|---|---|---|---|---|---|---|
| **Word 1** | **0.345** | **0.121** | **-0.538** | **1.011** | **2.011** | | **0.004** |
| **Word2** | 0.445 | 2.101 | 1.054 | -0.181 | -0.114 | | 0.764 |
| **…** | … | … | … | … | … | | … |
| **word V** | 0.334 | -0.087 | -0.407 | 1.114 | 0.554 | | 0.674 |




Continuous bag-of-word model

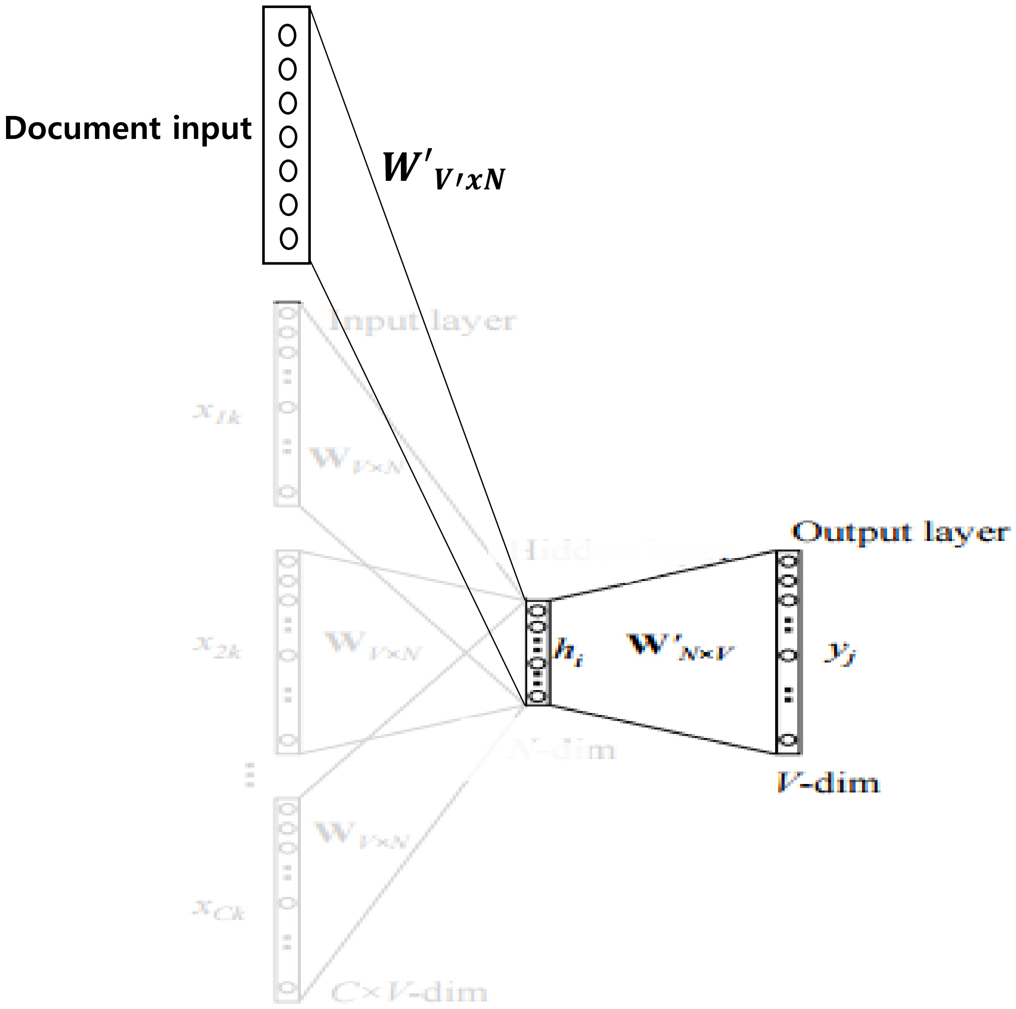
- $V = vocabulary\ size$
- $C = window\ size$
- $N = word2vec\ dimension$

# Distributed representation : word2vec, doc2vec



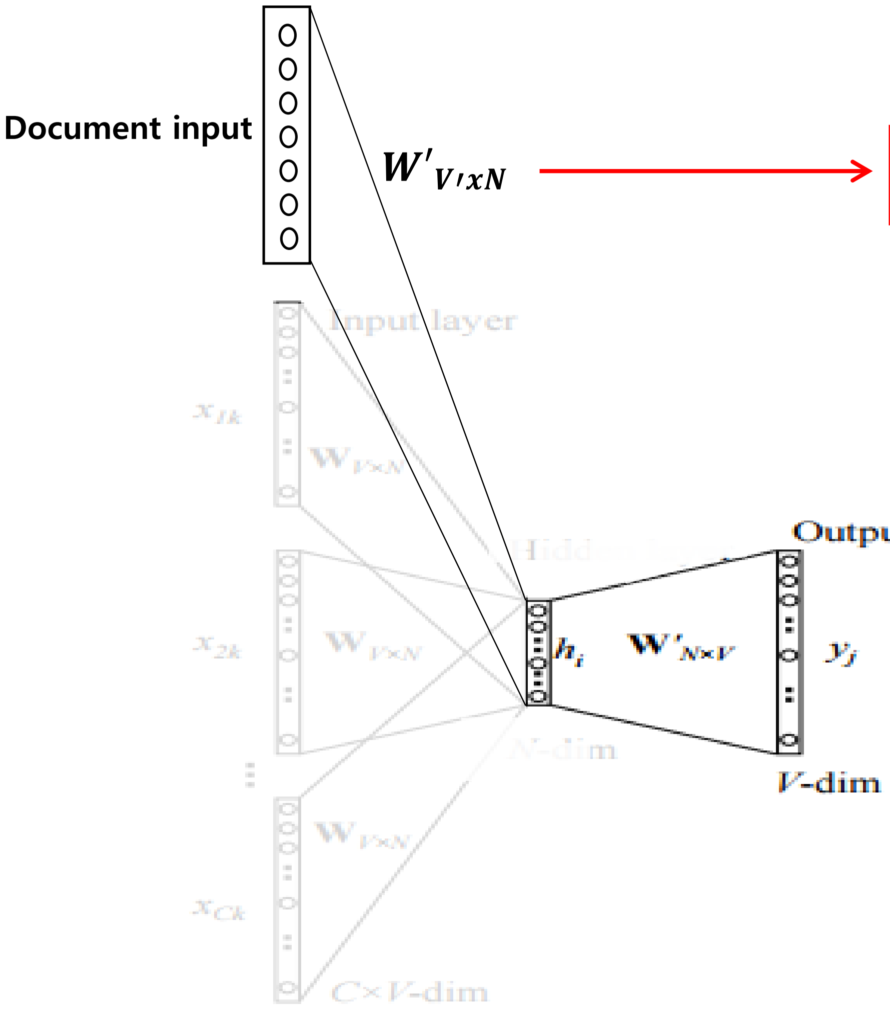

- $\mathrm{V} = vocabulary\ size$
- $\mathbf{V'} = \boldsymbol{documents\ size}$
- $\mathrm{C} = window\ size$
- $\mathrm{N} = doc2vec\ dimension$

# Distributed representation : word2vec, doc2vec




- $\mathrm{V} = vocabulary\ size$
- $\mathbf{V'} = \boldsymbol{documents\ size}$
- $\mathrm{C} = window\ size$
- $\mathrm{N} = doc2vec\ dimension$

# Distributed representation : word2vec, doc2vec




| | X1 | X2 | X3 | X4 | X5 | | Xn |
|---|---|---|---|---|---|---|---|
| **Doc 1** | **0.345** | **0.121** | **-0.538** | **1.011** | **2.011** | | **0.004** |
| **Doc 2** | 0.445 | 2.101 | 1.054 | -0.181 | -0.114 | | 0.764 |
| **…** | … | … | … | … | … | | … |
| **Doc V′** | 0.334 | -0.087 | -0.407 | 1.114 | 0.554 | | 0.674 |

- $\mathrm{V} = \textit{vocabulary size}$
- $\mathbf{V'} = \boldsymbol{documents\ size}$
- $\mathrm{C} = \textit{window size}$
- $\mathrm{N} = \textit{doc2vec dimension}$