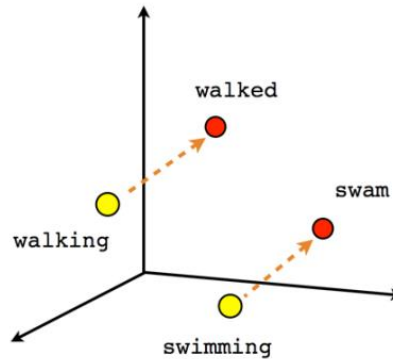
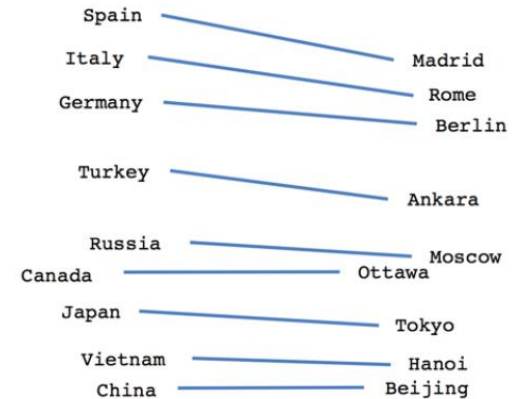


Male-Female



Verb tense



Country-Capital

# Lecture 5: Text Representation II

## Distributed Representations

Pilsung Kang

School of Industrial Management Engineering

Korea University

# AGENDA

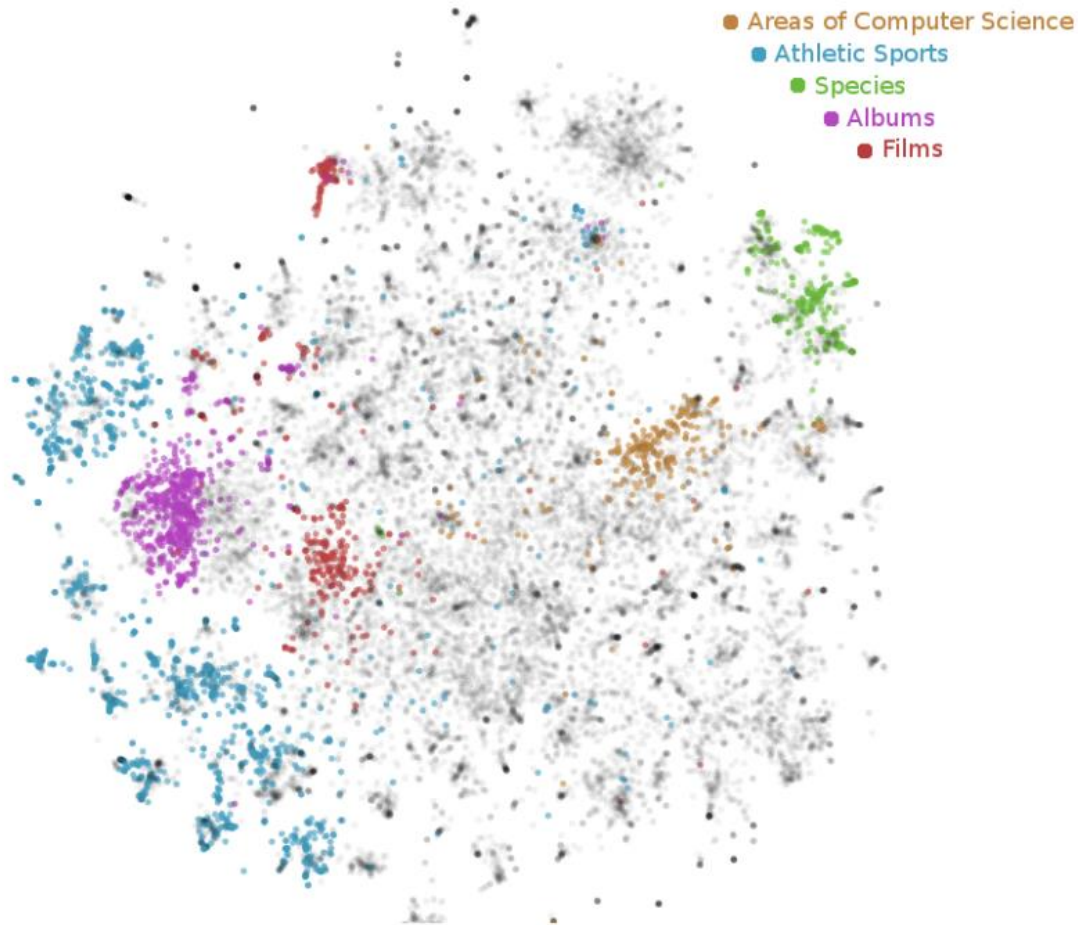
- 01 Word-level: NNLM
- 02 Word-level: Word2Vec
- 03 Word-level: GloVe
- 04 Word-level: Fasttext
- 05 Sentence/Paragraph/Document-level
- 06 More Things to Embed?

# Document Embedding

Dai et al. (2015)

- If we can embed words, why not sentences, phrases, or documents?

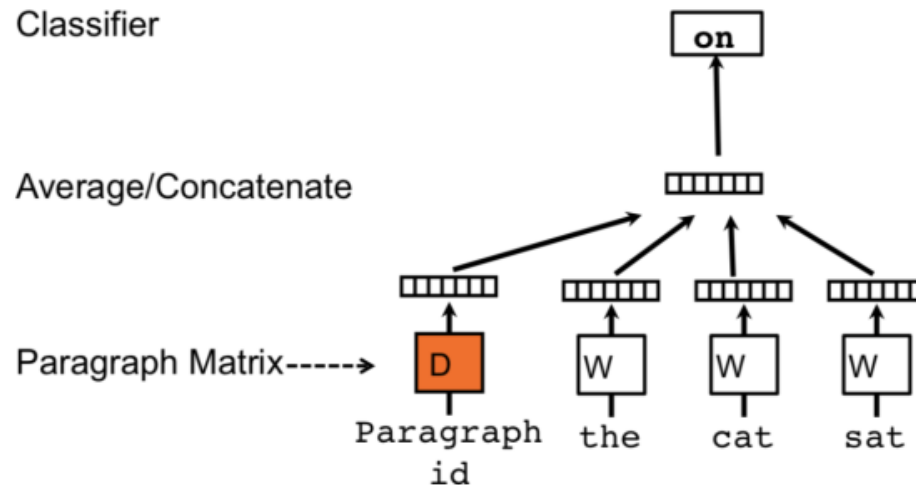
Visualization of Wikipedia paragraph vectors using t-SNE



# Document Embedding

Le and Mikolov (2015)

- Paragraph Vector model: Distributed Memory (PV-DM) model

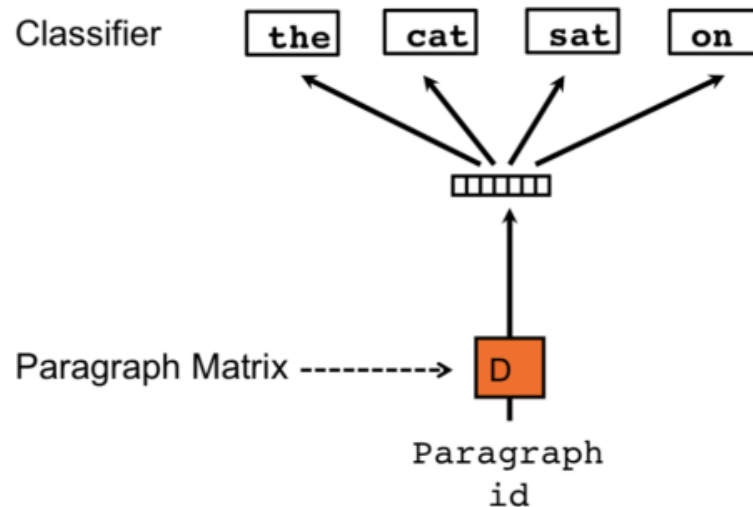


- ✓ The paragraph vectors are also asked to contribute to the prediction task of the next word given many contexts sampled from the paragraph
- ✓ Paragraph vectors are shared for all windows generated from the same paragraph, but not across paragraphs
- ✓ Word vectors are shared across all paragraphs

# Document Embedding

Le and Mikolov (2015)

- Paragraph Vector model: Distributed Bag of Words (PV-DBOW)

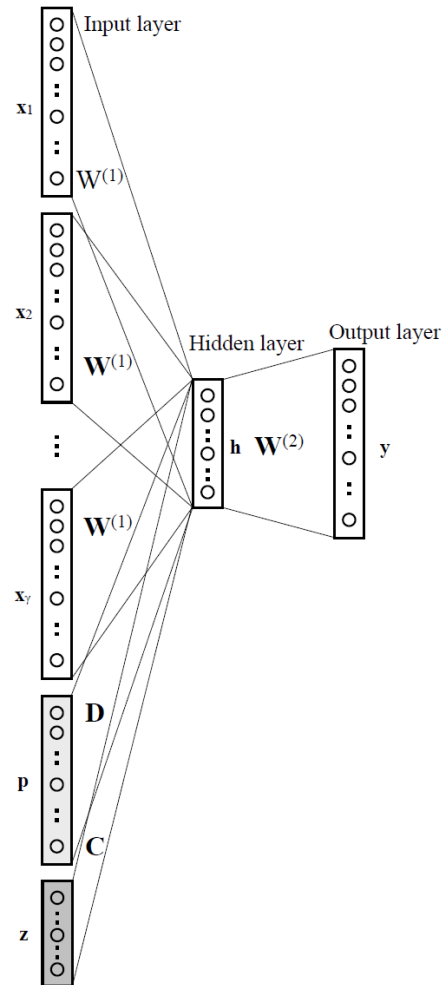


- ✓ Ignore the context words in the input, and force the model to predict words randomly sampled from the paragraph in the output
- ✓ Does not need word vectors
- ✓ PV-DM alone usually works well for most tasks, but the combination of PV-DM and PV-DBOW are recommended

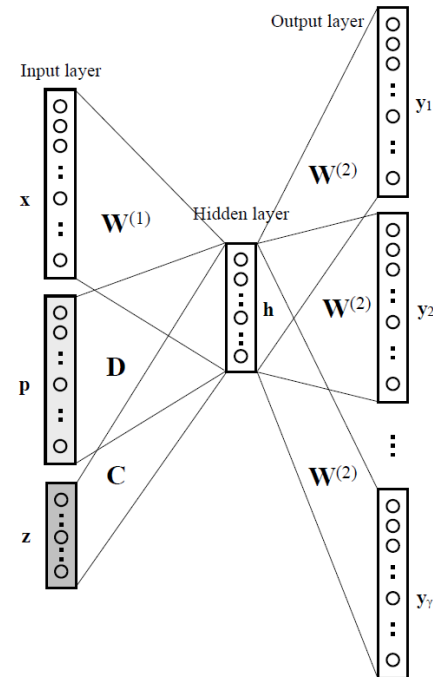
# Let's Embed Everything!

Park et al. (2016+)

- Supervised Paragraph Vector (SPV) for Class Embedding



(a)

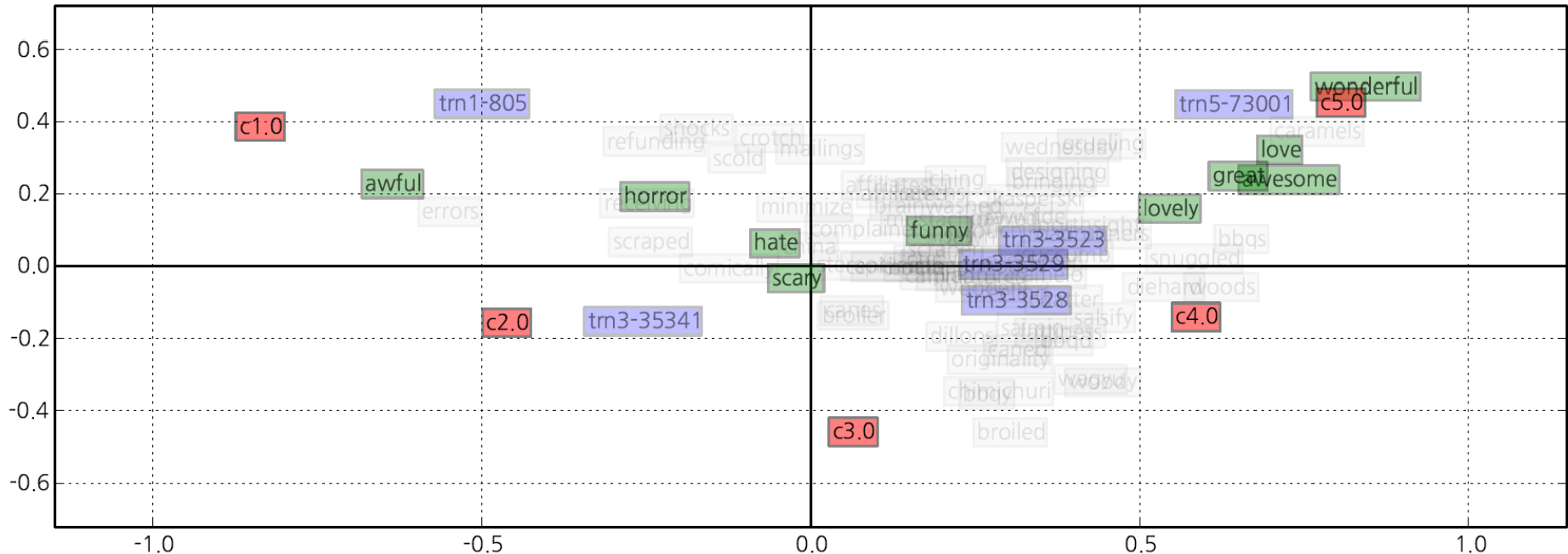


(b)

# Let's Embed Everything!

Park et al. (2016+)

- Supervised Paragraph Vector (SPV) for Class Embedding



# Let's Embed Everything!

Park et al. (2016+)

- Supervised Paragraph Vector (SPV) for Class Embedding

	imdb							
n(epochs)	1	5	10	30	50	70	100	<i>t</i>
BOW-TF	<b>85.30</b>	-	-	-	-	-	-	-
BOW-TFIDF	<b>85.55</b>	-	-	-	-	-	-	-
PV-DM	77.06	80.78	80.84	79.68	81.22	81.49	<b>82.16</b>	123.14
PV-DBOW	85.89	88.19	<b>88.47</b>	88.27	88.22	88.12	88.04	115.22
SPV-DM	82.57	81.68	82.05	<b>82.66</b>	82.42	82.53	82.61	121.51
SPV-DBOW	87.58	<b>*88.87</b>	88.69	88.53	88.51	88.56	88.49	117.33

	yelp							
n(epochs)	1	5	10	30	50	70	100	<i>t</i>
BOW-TF	<b>58.42</b>	-	-	-	-	-	-	-
BOW-TFIDF	<b>58.93</b>	-	-	-	-	-	-	-
PV-DM	50.59	51.70	52.67	<b>52.97</b>	51.73	51.81	52.70	546.10
PV-DBOW	58.53	<b>59.37</b>	58.91	59.13	59.10	59.06	59.21	534.03
SPV-DM	51.48	51.42	52.40	53.14	53.77	53.75	<b>53.84</b>	552.71
SPV-DBOW	60.13	<b>*60.21</b>	60.04	59.93	59.85	59.77	59.93	538.95

	amazon							
n(epochs)	1	5	10	30	50	70	100	<i>t</i>
BOW-TF	<b>85.91</b>	-	-	-	-	-	-	-
BOW-TFIDF	<b>85.97</b>	-	-	-	-	-	-	-
PV-DM	76.47	77.38	78.86	<b>79.26</b>	75.83	77.00	78.57	361.52
PV-DBOW	86.87	87.96	88.30	88.34	88.31	<b>88.42</b>	88.18	339.14
SPV-DM	79.05	78.35	78.66	80.26	80.36	80.62	<b>80.70</b>	371.95
SPV-DBOW	88.58	89.21	89.16	<b>*89.34</b>	89.08	89.06	89.29	342.40

	20news							
n(epochs)	1	5	10	30	50	70	100	<i>t</i>
BOW-TF	<b>58.58</b>	-	-	-	-	-	-	-
BOW-TFIDF	<b>63.43</b>	-	-	-	-	-	-	-
PV-DM	24.45	41.17	45.36	49.04	52.34	53.85	<b>54.27</b>	71.64
PV-DBOW	53.51	69.16	72.51	74.76	75.22	75.16	<b>75.40</b>	70.37
SPV-DM	44.76	64.18	67.01	67.84	68.32	<b>68.34</b>	67.70	71.88
SPV-DBOW	69.41	76.97	77.95	78.93	78.93	79.47	<b>*79.59</b>	72.18



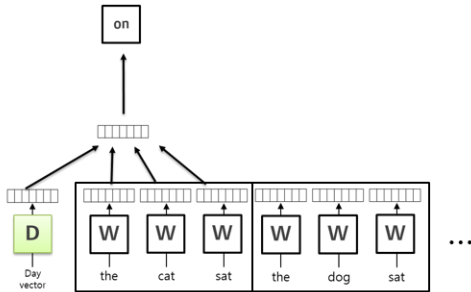
# AGENDA

- 01 Word-level: NNLM
- 02 Word-level: Word2Vec
- 03 Word-level: GloVe
- 04 Word-level: Fasttext
- 05 Sentence/Paragraph/Document-level
- 06 More Things to Embed?

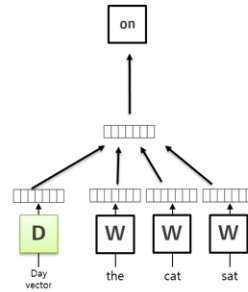
# Let's Embed Everything!

- Day Embedding in News corpus

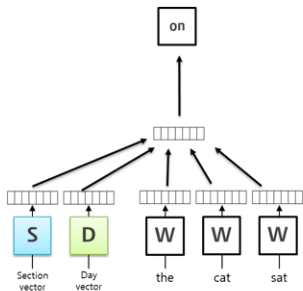
Approach 01: 하루 동안의 기사제목들을 병합 후 Day 태그



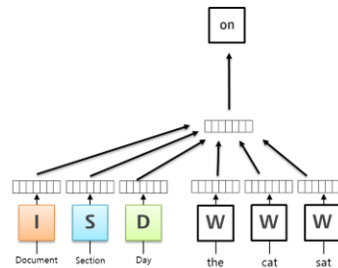
Approach 02: 각 뉴스기사 제목에 Day 태그



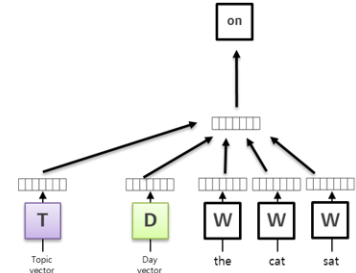
Approach 03: 각 뉴스기사제목에 Day, Section 태그



Approach 04: 각 뉴스기사제목에 Day, Section, IDX 태그



Approach 05: 각 뉴스기사제목에 Day, Topic 태그



## Host-based

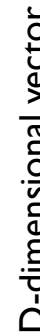
- # Network-based

## Syscall Trace

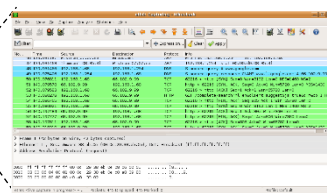


Classifier

Average/Concatenate



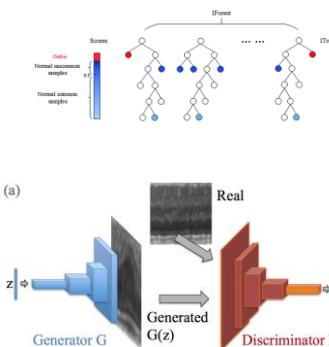
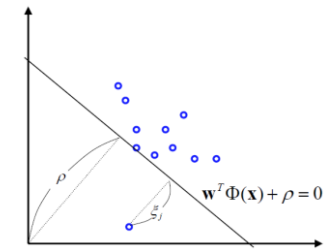
## Packet Capture



## Features

#	Flow	Flow Name	Source	Port	Estimate	Estimated Protocol	Timestamp	Flow Size	Total Flow	Total Size	Total Len	Total Len	Flow Pack
1	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
2	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
3	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
4	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
5	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
6	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
7	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
8	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
9	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
10	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
11	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
12	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
13	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
14	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
15	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
16	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
17	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
18	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
19	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
20	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
21	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
22	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
23	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
24	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
25	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
26	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
27	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
28	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
29	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	
30	192.168.1.100	192.168.1.100	192.168.1.100	49150	6.0.0.0/1	0	1	2	0	12	0	6	

# CSV Format



길이가 **긴** 시퀀스도 **10차원** 벡터로

# Let's Embed Everything!

- Sequence Embedding based on Doc2Vec

- ✓ Syscall2Vec: 하나의 System Call Trace를 Document로 취급하고, 개별 syscall을 word로 취급하여 임베딩 수행

Document

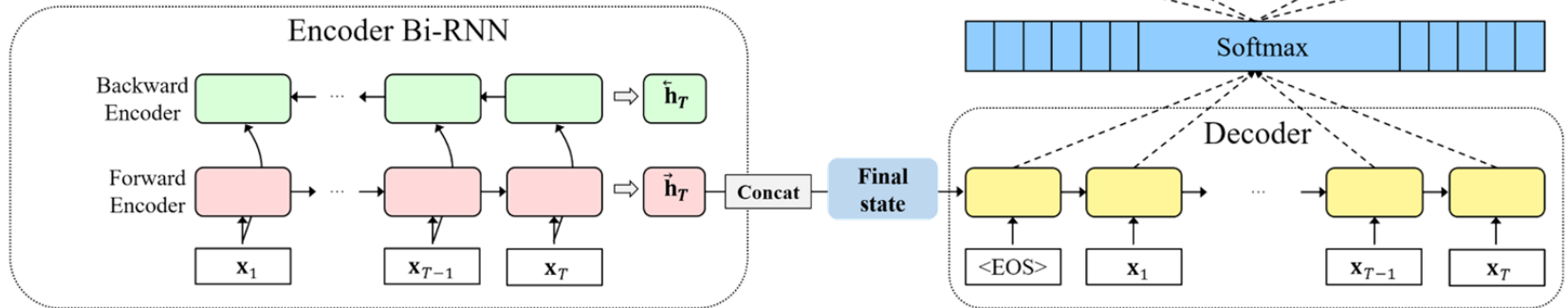
```
168 3 3 265 168 3 43 168 3 168 168 43 265 168 3 168 43 168 43 168 265 43 265 265 168  
265 265 168 168 168 3 168 3 265 168 3 168 168 168 168 3 168 168 168 3 3 168 168 265  
168 3 168 265 168 168 3 168 265 43 168 265 43 3 265 43 43 3 ...
```

Word 1   Word 2   Word 3   Word 4

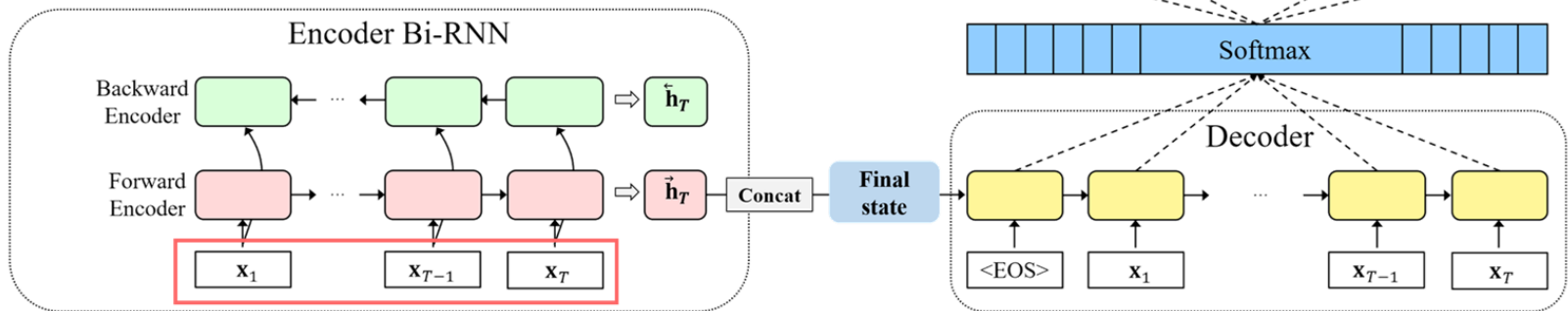
# Let's Embed Everything!

- RNN-AE 구조

- ✓ Bi-Directional RNN



- RNN-DAE 구조



## Corruption Model

- 임의의 syscall을  $p$ 의 확률로 drop
- 임의의 syscall sequence를 permutation

# Let's Embed Everything!

- Live2Vec in **afreecaTV** 



[그림10] 문장과 Live 방송의 추론의 예시

# Let's Embed Everything!

- Live2Vec in **afreecaTV**



[그림11] 유사도 측정 비교(Nearest neighbor vs Live2Vec)



A person in a dark suit and light blue striped shirt is holding a white rectangular sign. The sign has the text "ANY questions?" written on it in a black, handwritten-style font. The background is slightly blurred, showing orange and white elements.

ANY  
questions?