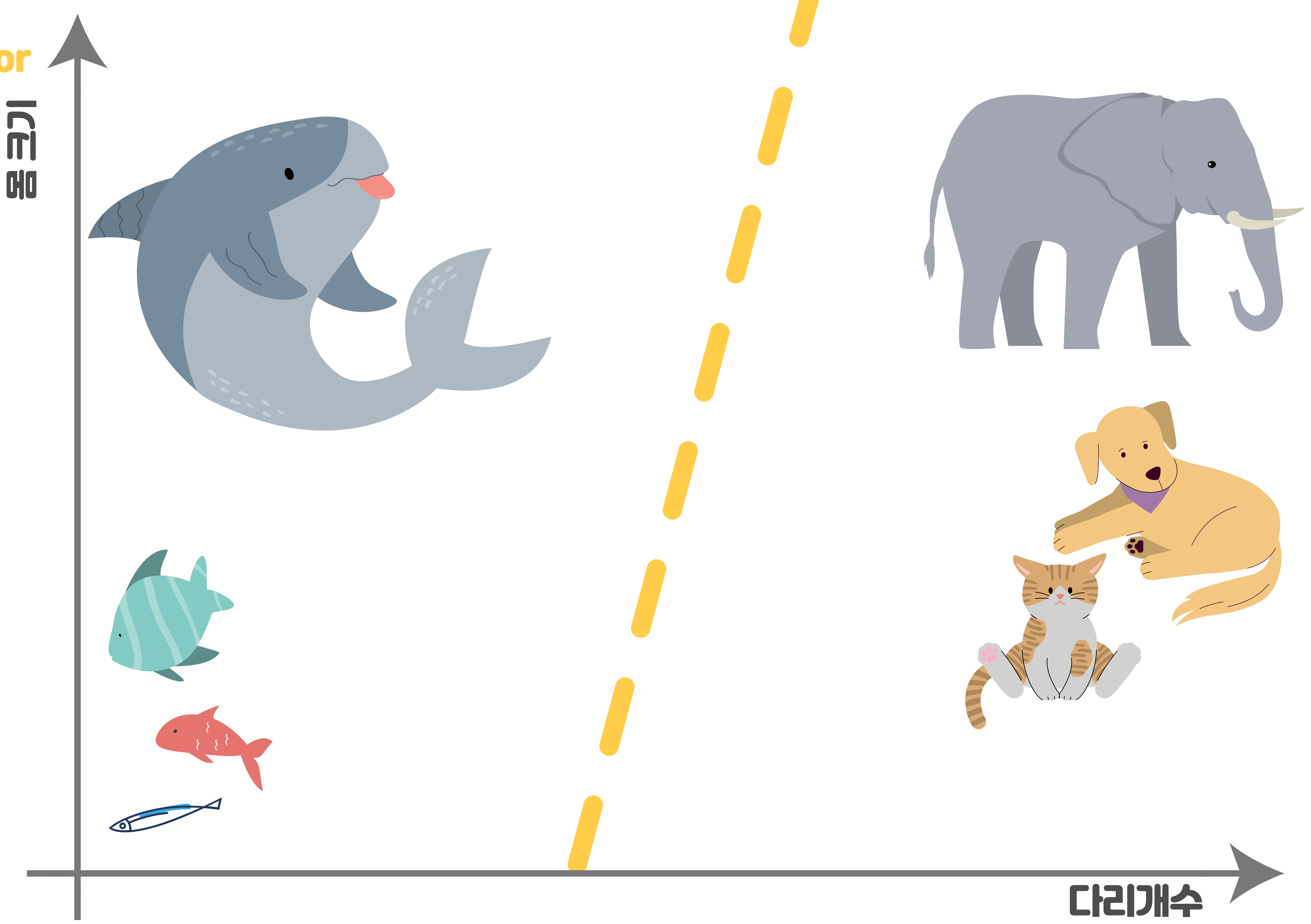
Word2Vector

정수연

00

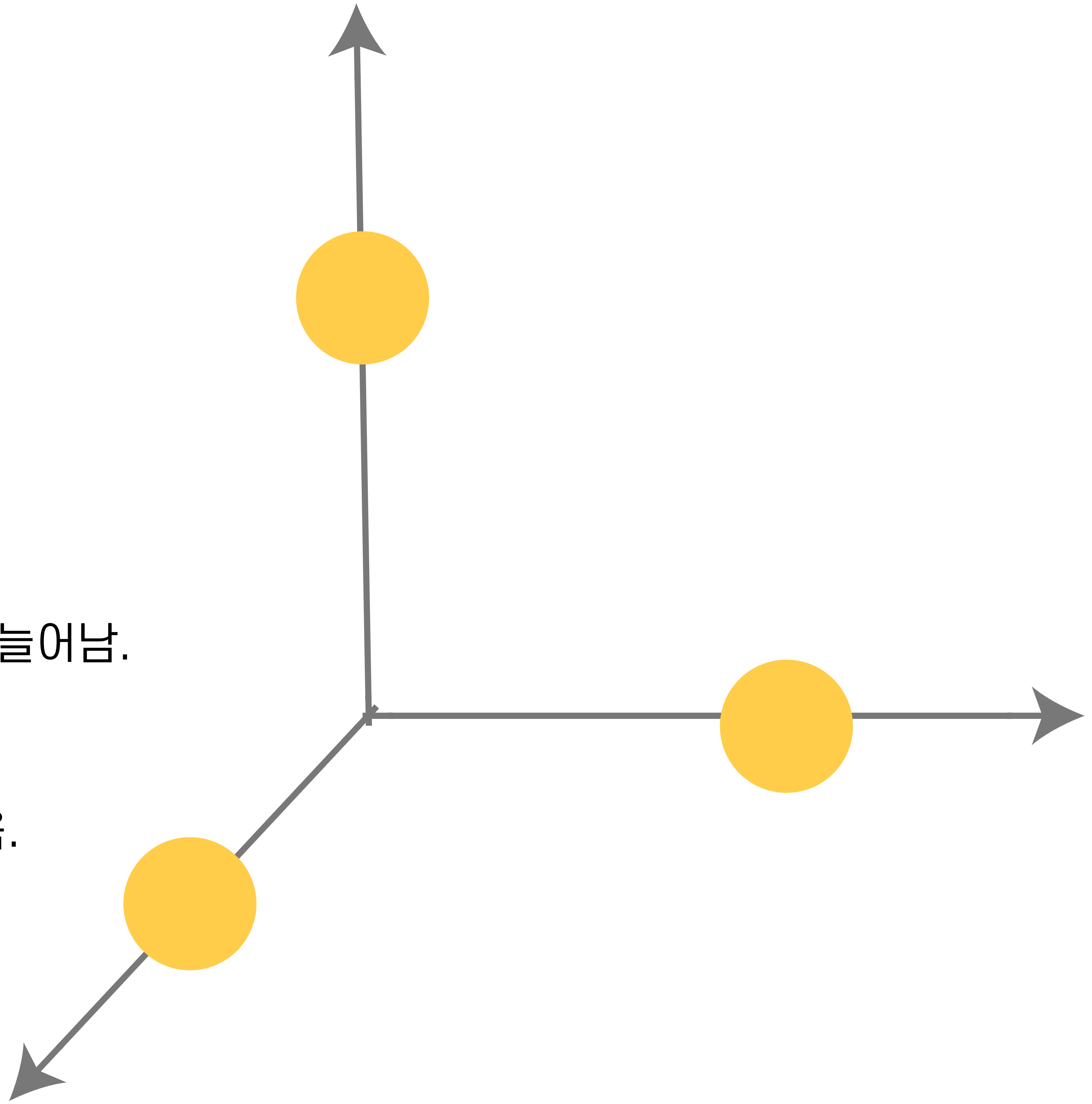
Word2Vector

00 Word2Vector



One Hot Encoding

단어	Vector
세상	[1,0,0]
모든	[0,1,0]
사람	[0,0,1]



단어의 수가 늘어날 수록, 벡터를 저장하기 위해 필요한 공간이 계속 늘어남.

즉, 단어 집합의 크기가 곧 벡터의 차원의 수가 됨.

그 외에도 희소행렬 문제, 단어 간 유사성을 알 수 없다는 단점이 있음.

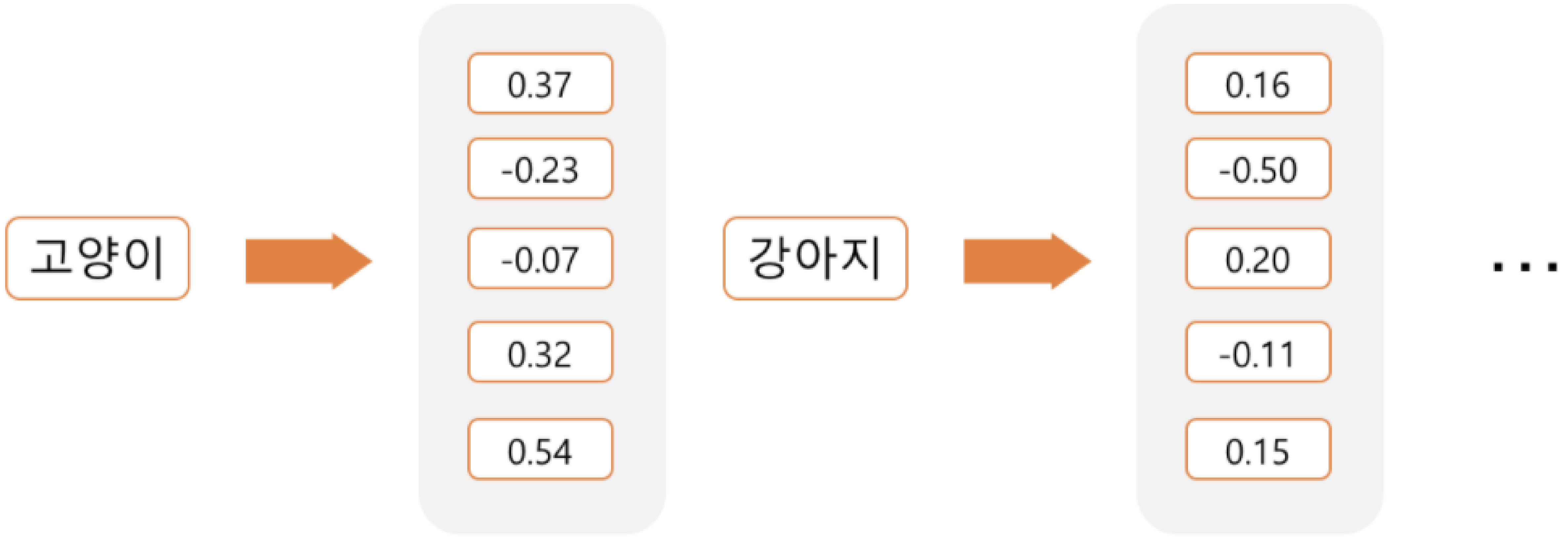
Distributed Representation

distributional hypothesis

> A word's meaning is give by the words that frequently appear close-by

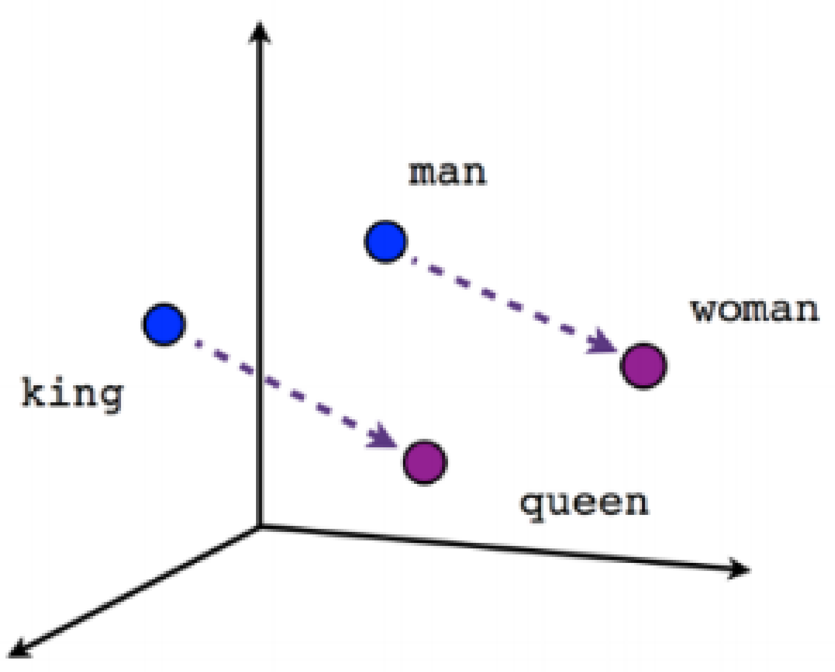
Distributed Representation

각각의 속성을 독립적인 차원으로 나타내지 않고, 우리가 정한 차원으로 대상을 대응시켜서 표현

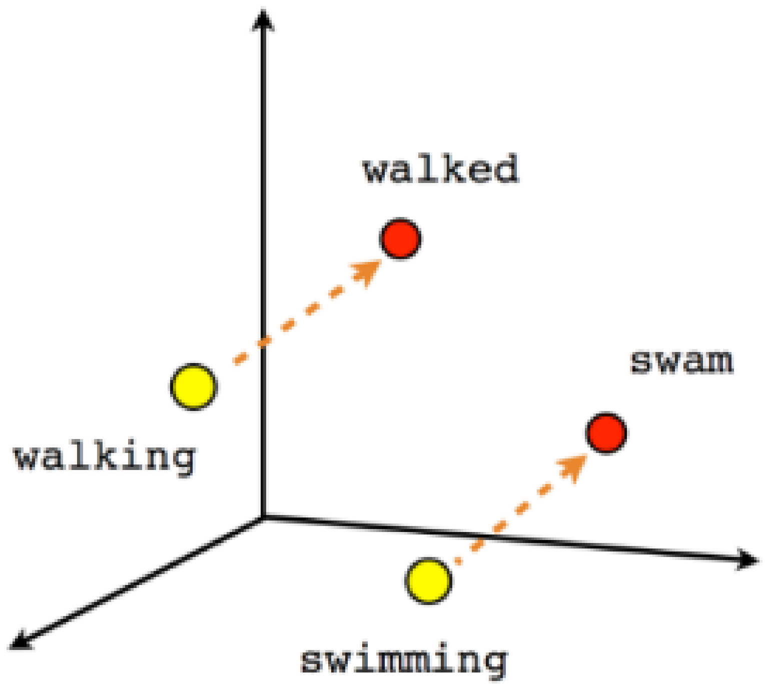


하나의 차원이 하나의 속성을 명시적으로 표현하는 것이 아니라 여러 차원들이 조합되어 나타내고자 하는 속성들을 표현

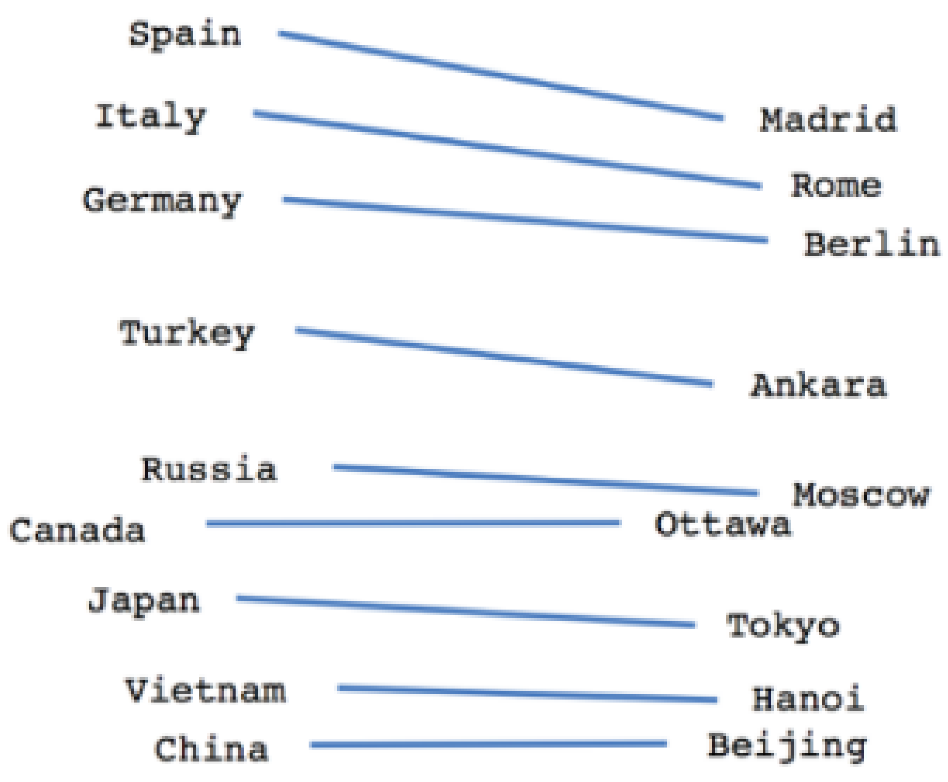
Distributed Representation



Male-Female



Verb tense

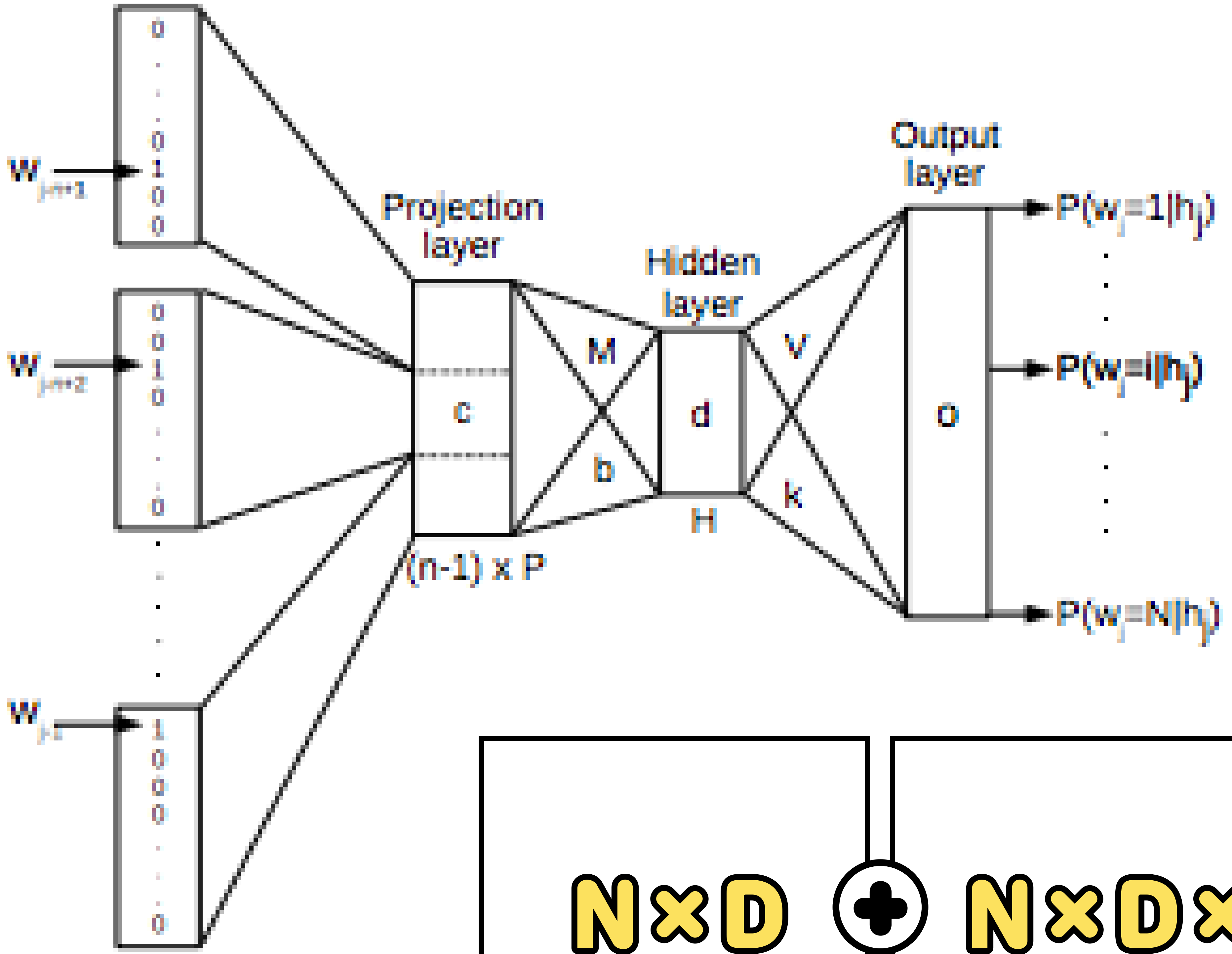


Country-Capital

단어 벡터의 요소값이 연속적이기 때문에 단어 벡터 간에 거리나 유사도를 잴 수가 있음.

Goals of Paper

- 매우 큰 데이터에서 퀄리티 높은 단어 벡터를 학습
- 이전에 제안되었던 NNLM, RNNLM 소개
- 단어 사이의 선형 규칙을 보존하는 새로운 모델 구조 제안
- 훈련 시간과 정확도

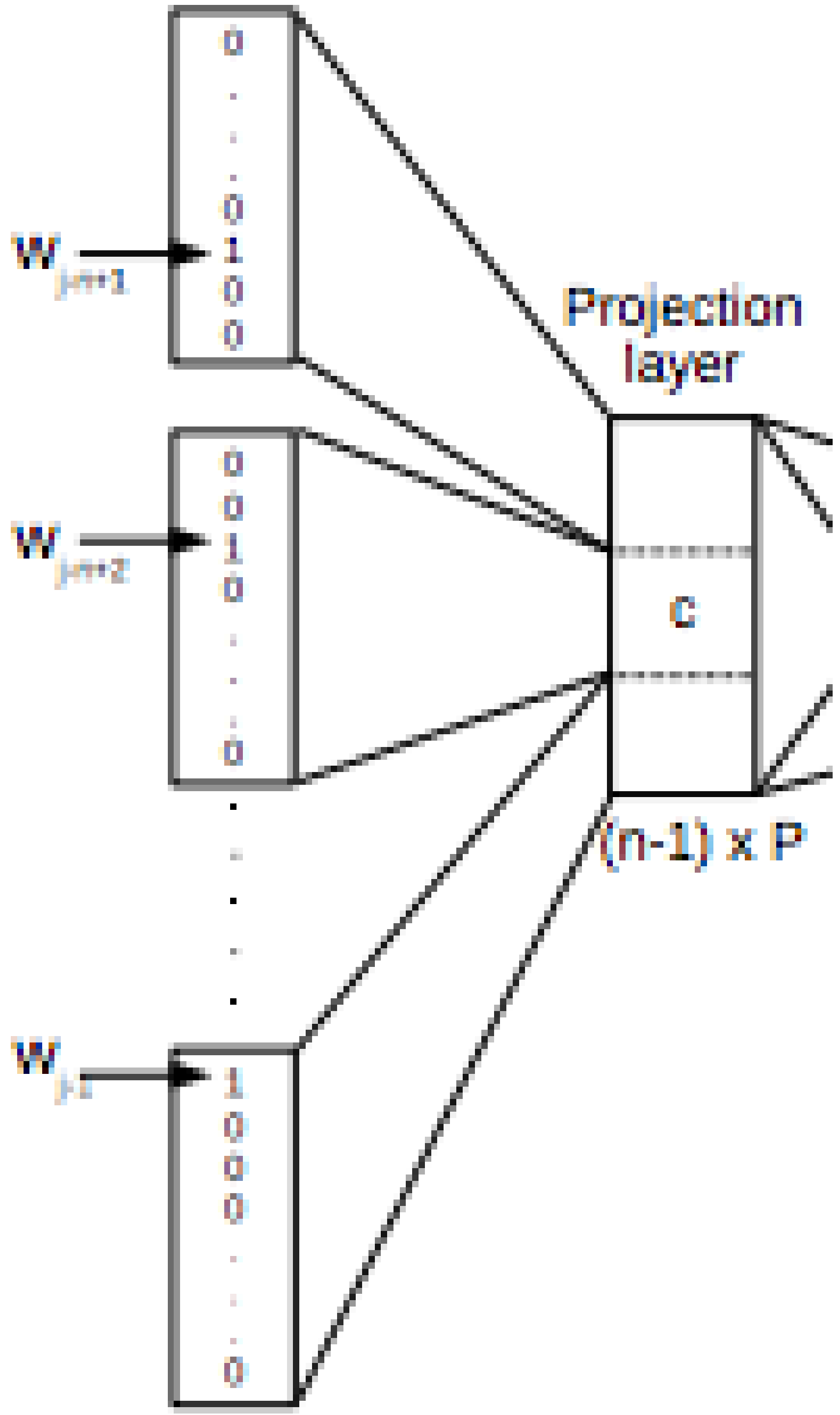


NNLM

Neural Network Language Model

n-1 개의 단어를 가지고 n번째 단어를 예측하는 n-gram 언어 모델

$$N \times D + N \times D \times H + H \times V = Q$$



- 어휘 집합에 속한 단어가 5개 뿐이고, wt가 네 번째 단어라고 가정하자.

- C(wt)는 행렬 C와 wt에 해당하는 원-핫 벡터 (One-hot Vector) 를 내적 (Inner Product) 한 것과 같다.

- 이는 C행렬에서 wt에 해당하는 행 (Row) 만 참조하는 것과 동일하다.

$$C(w_t) = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} 11 & 18 & 25 \\ 10 & 12 & 19 \\ 4 & 6 & 13 \\ 23 & 5 & 7 \\ 17 & 24 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 23 & 5 & 7 \end{bmatrix}$$

장점

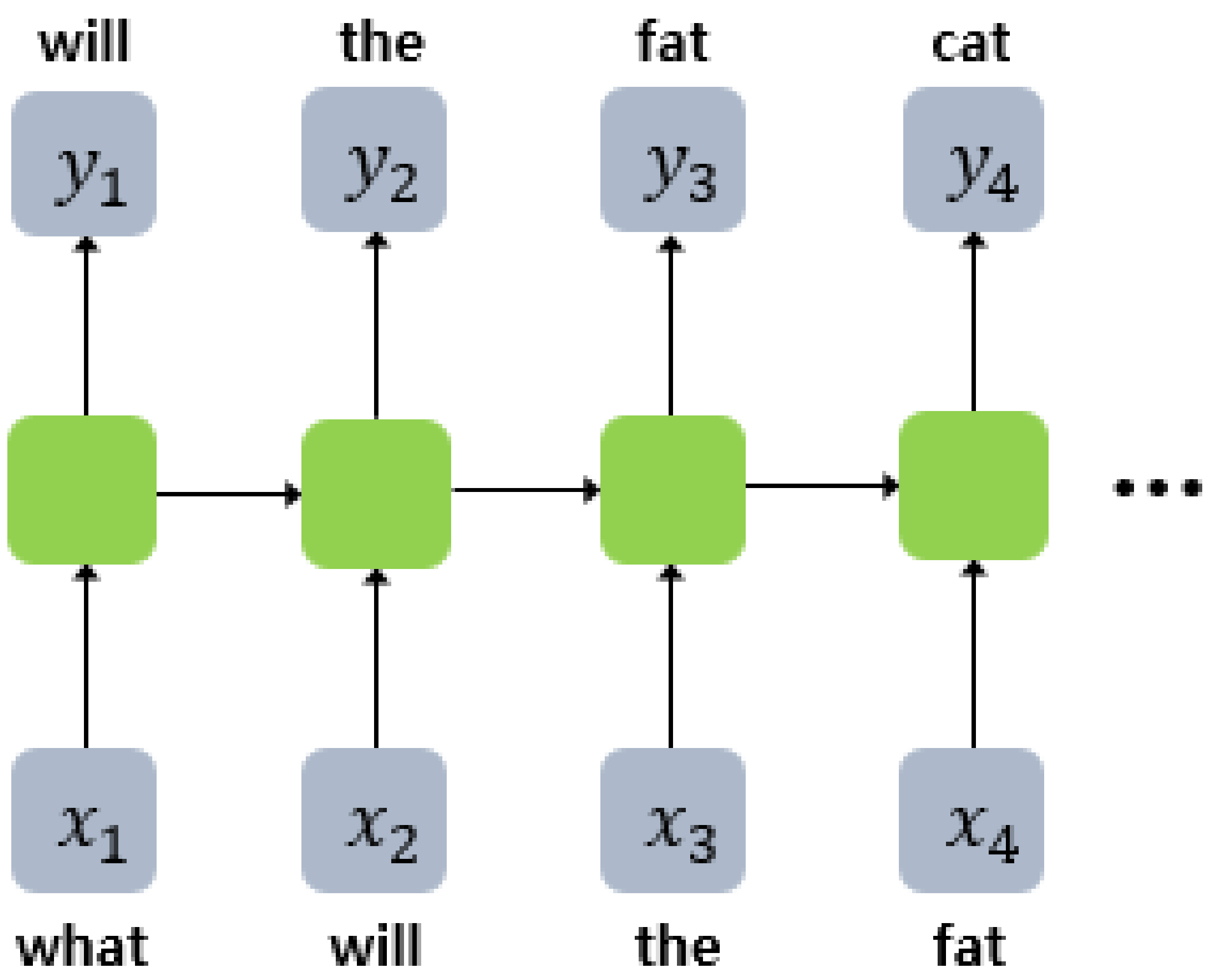
- 단어의 유사도를 표현
- 희소행렬 문제 해결

단점

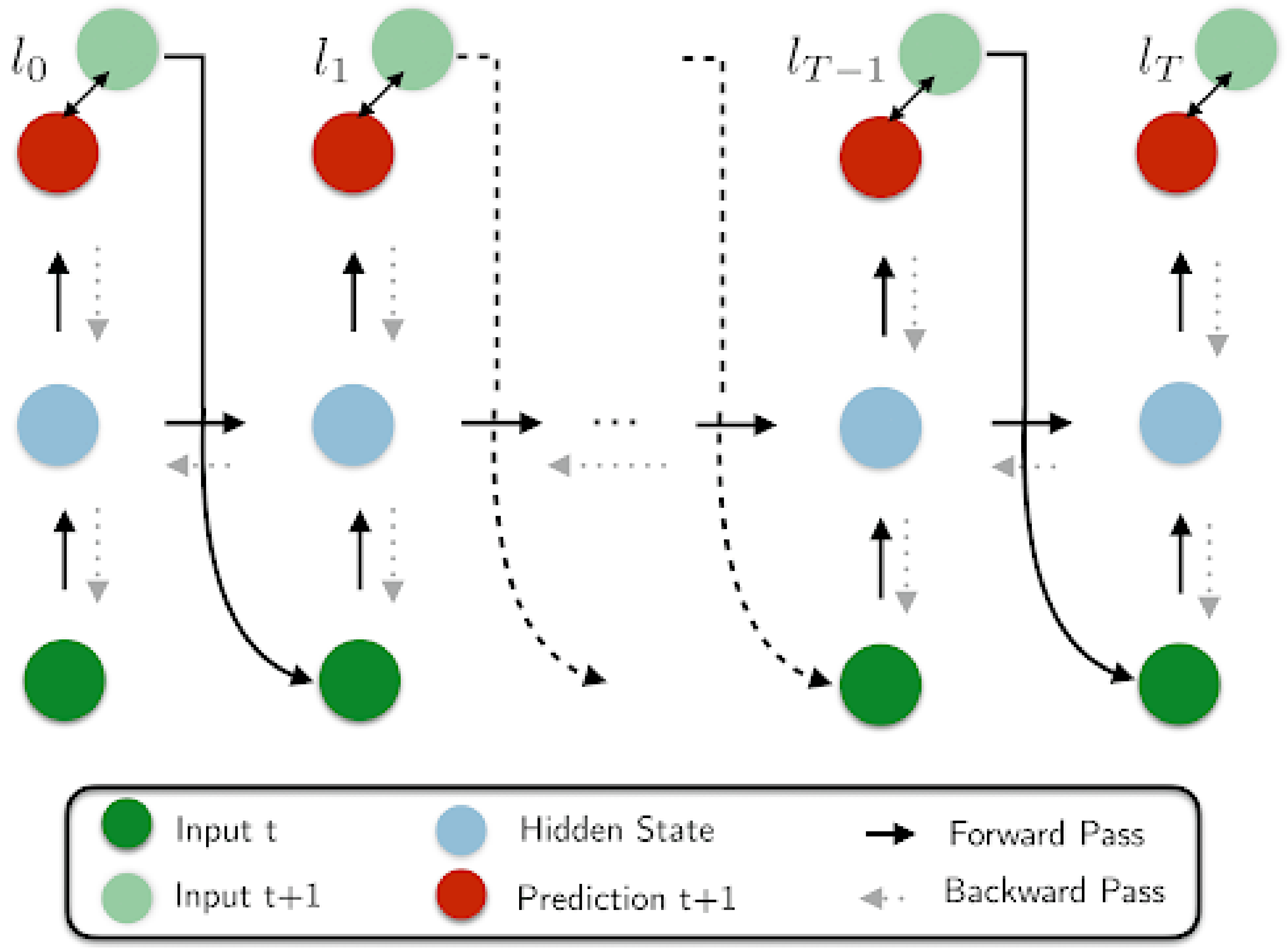
- 정해진 n개의 단어만 참고
- 미래 시점의 단어들을 고려 X

RNNLM | Recurrent Neural Network Language Model

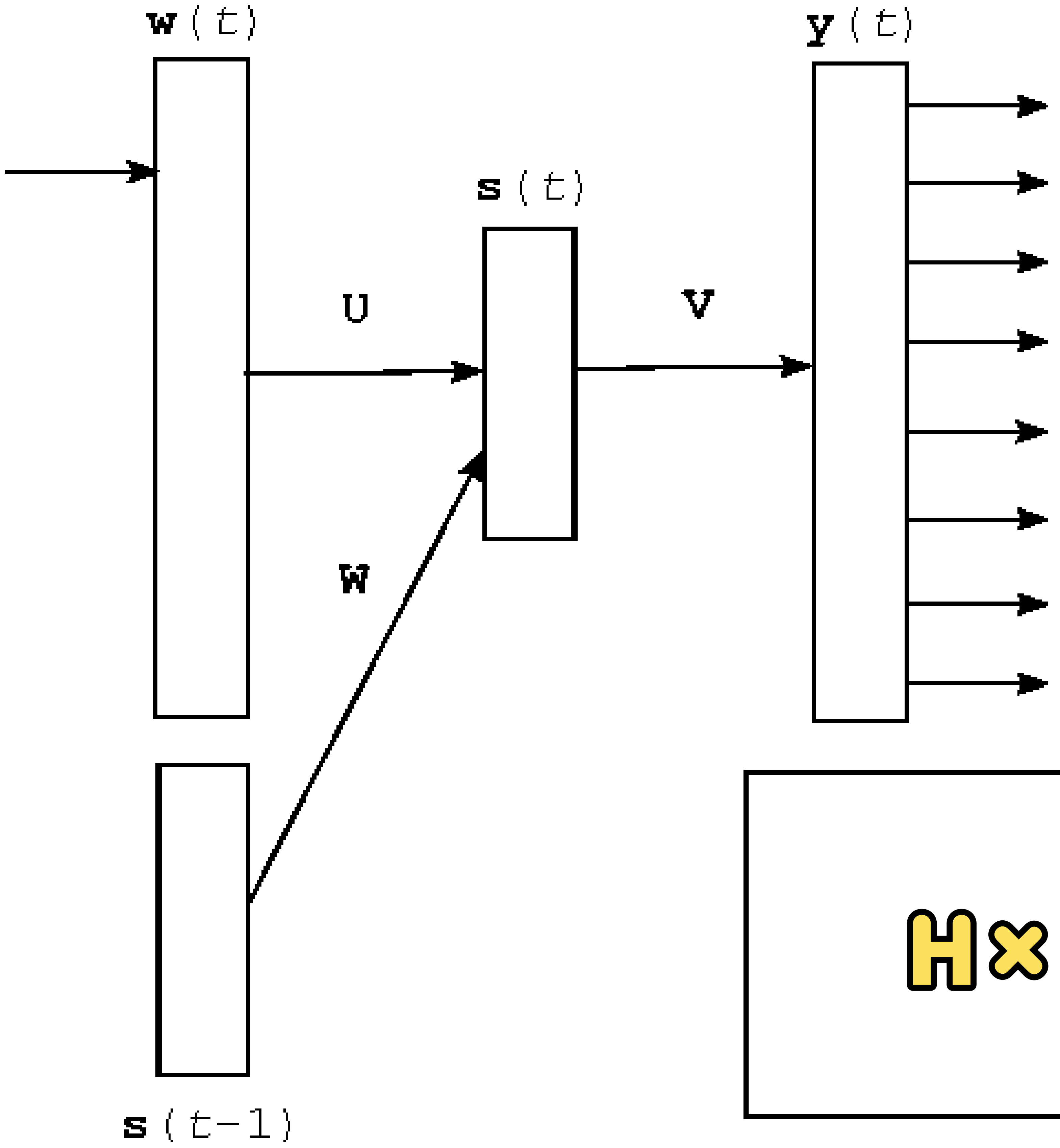
- 고정된 개수의 단어만을 입력으로 받아야하는 NNLM의 단점을 개선
- 시점(time step)이라는 개념이 도입



RNN Training with Teacher Forcing



02 Model architecture



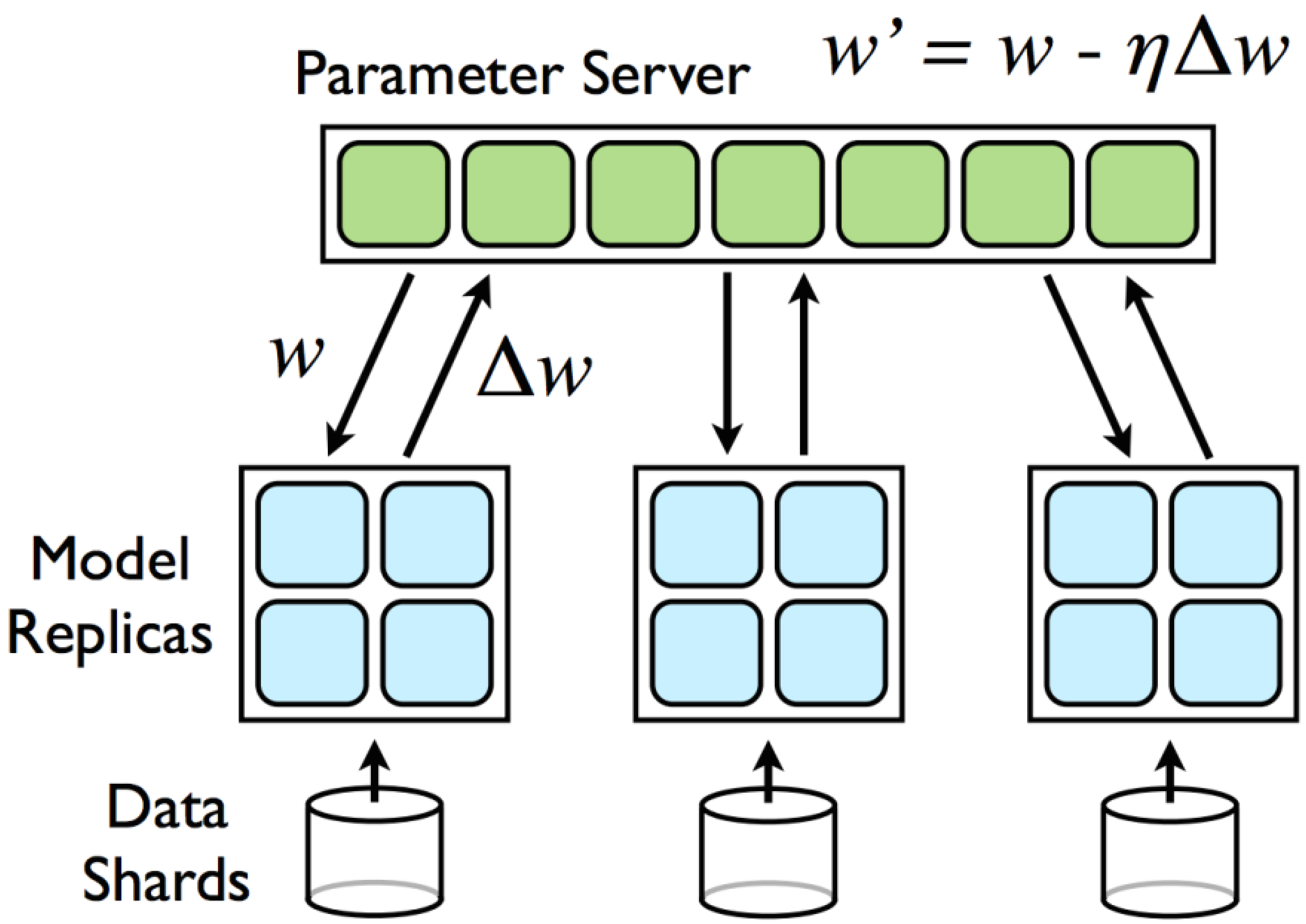
RNNLM

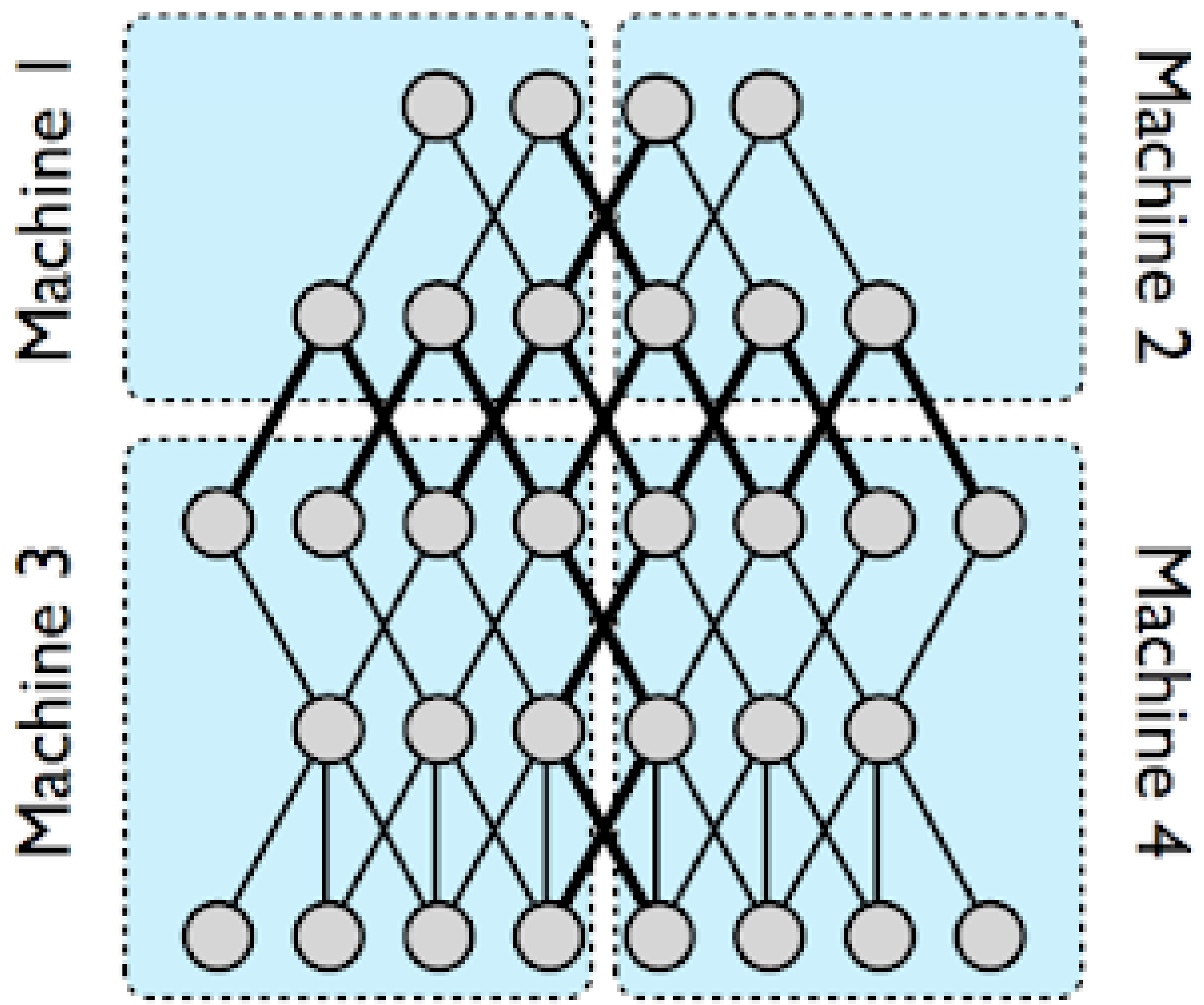
Recurrent Neural Network Language Model

$H \times H + H \times V = Q$

DistBelief

Parallel Training of Neural Networks





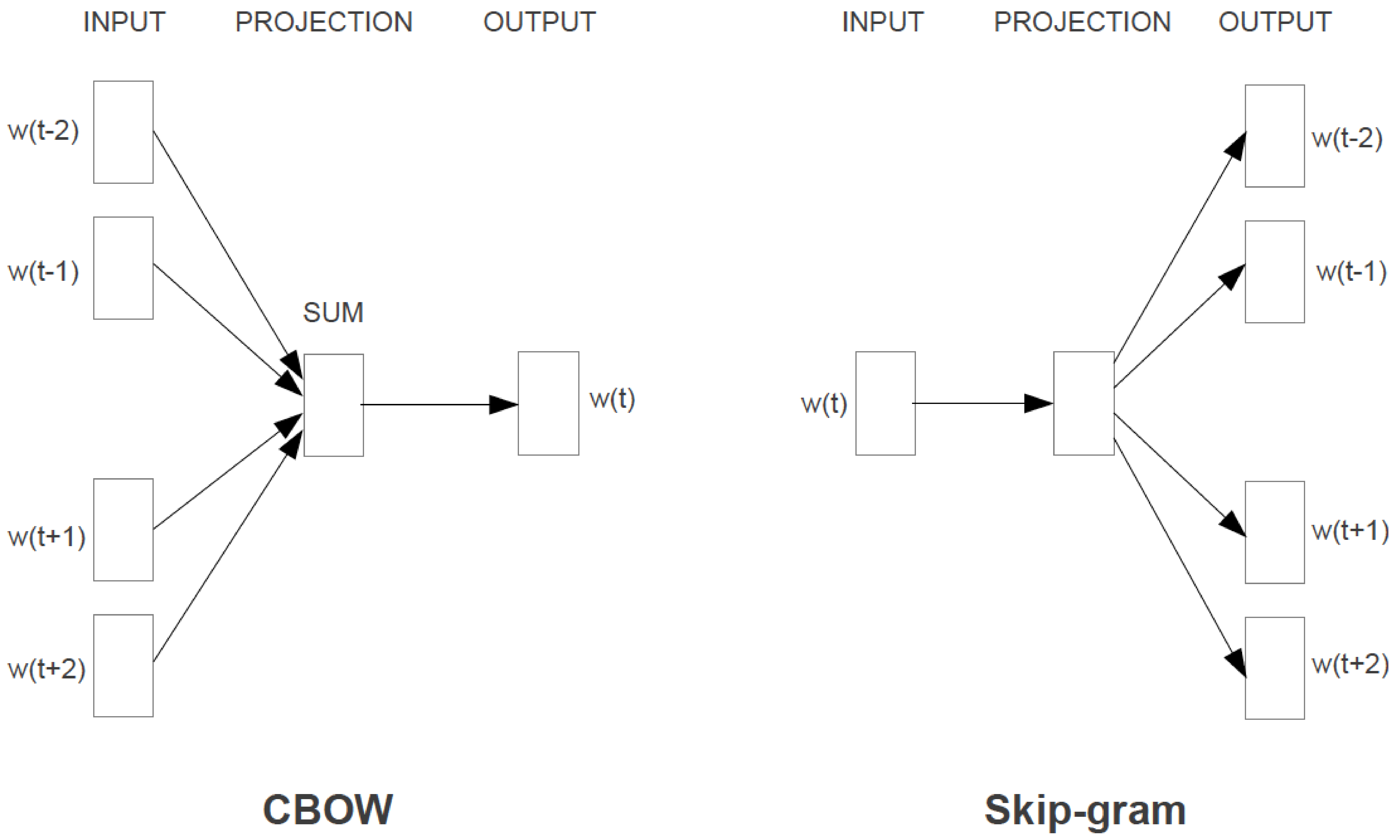
NNLM, RNNLM >> CBOW, Skip-gram

01 계산 복잡도 감소

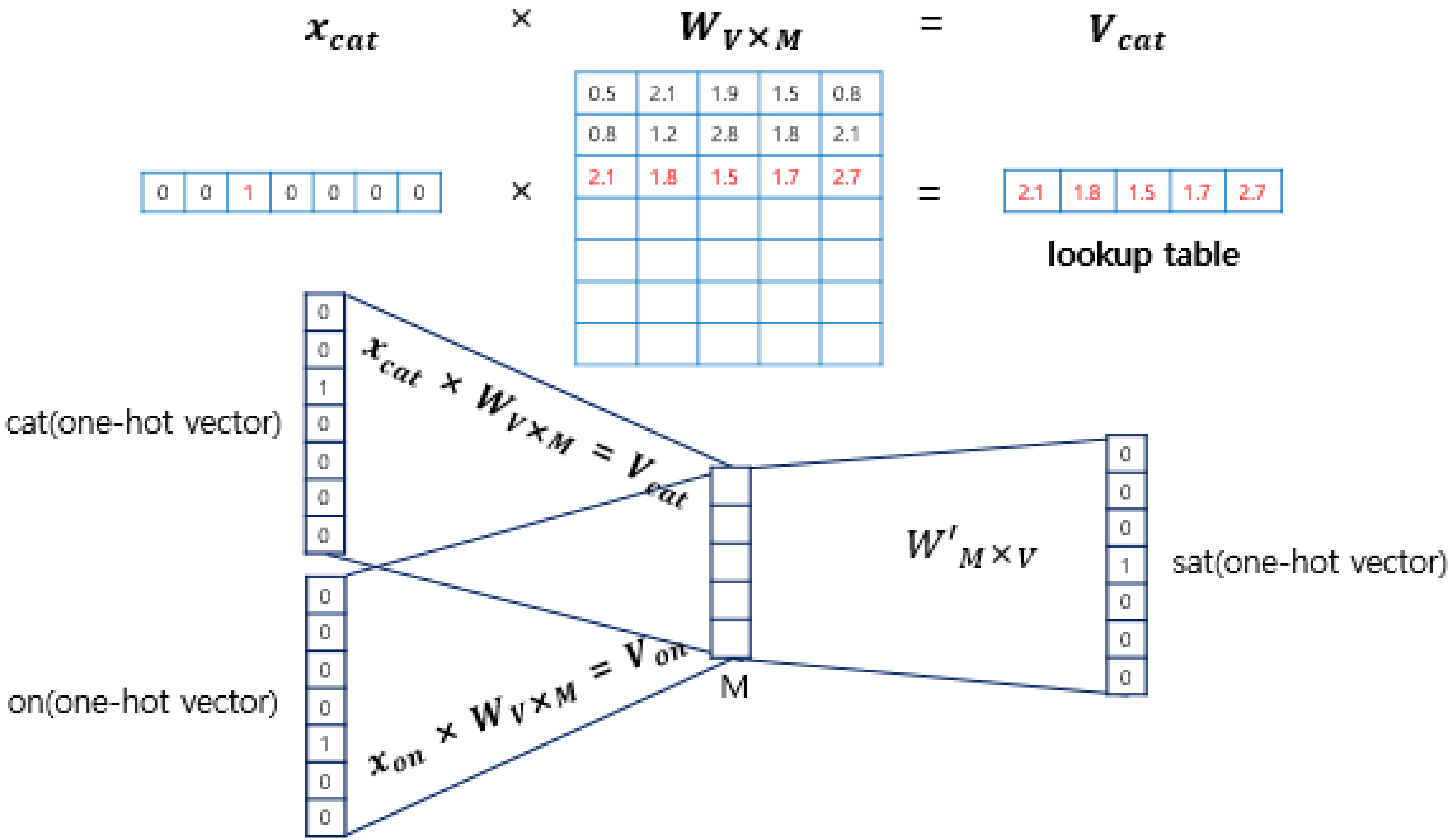
02 Future word 고려

03

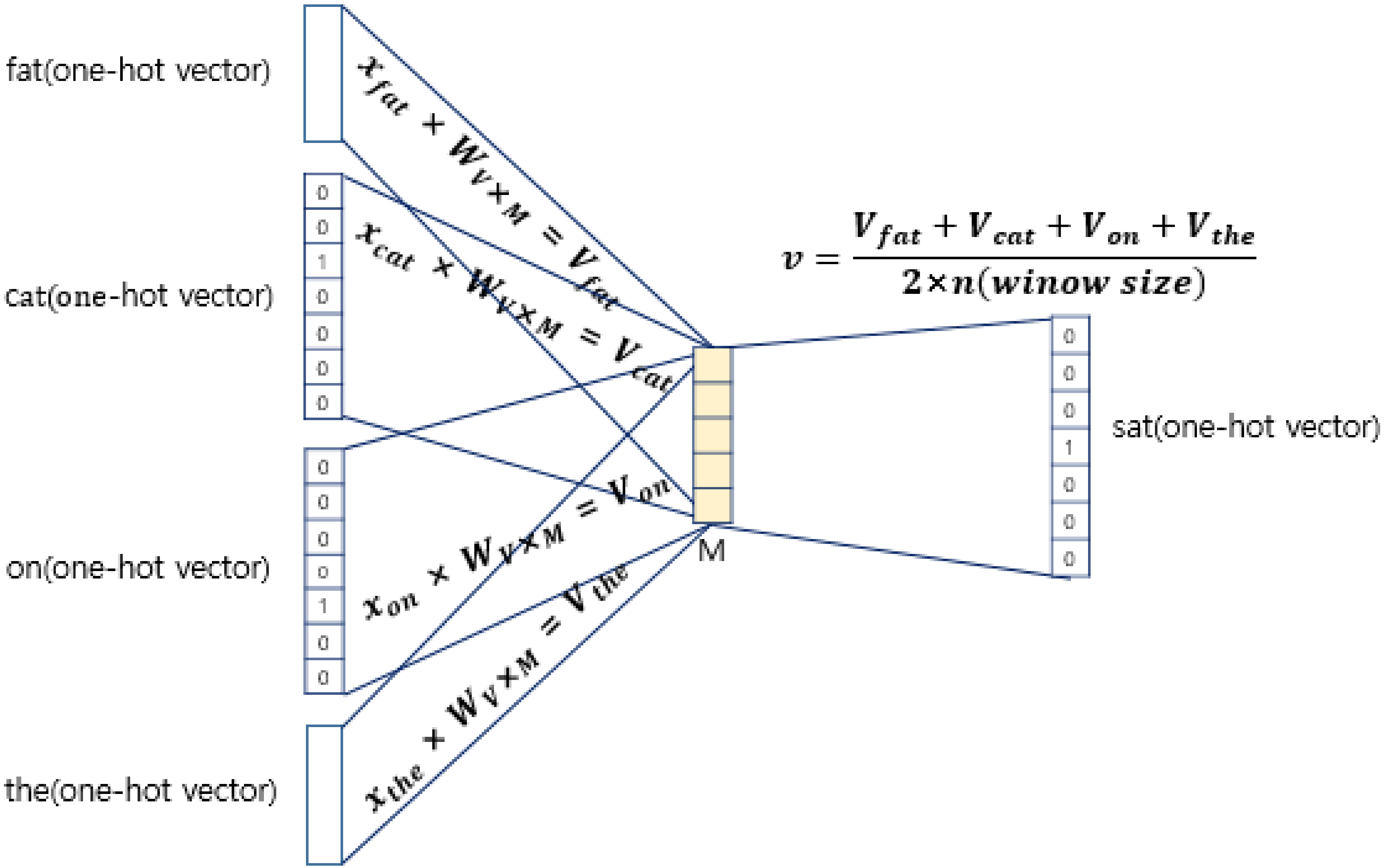
New Log-linear Model



New Log-linear Model



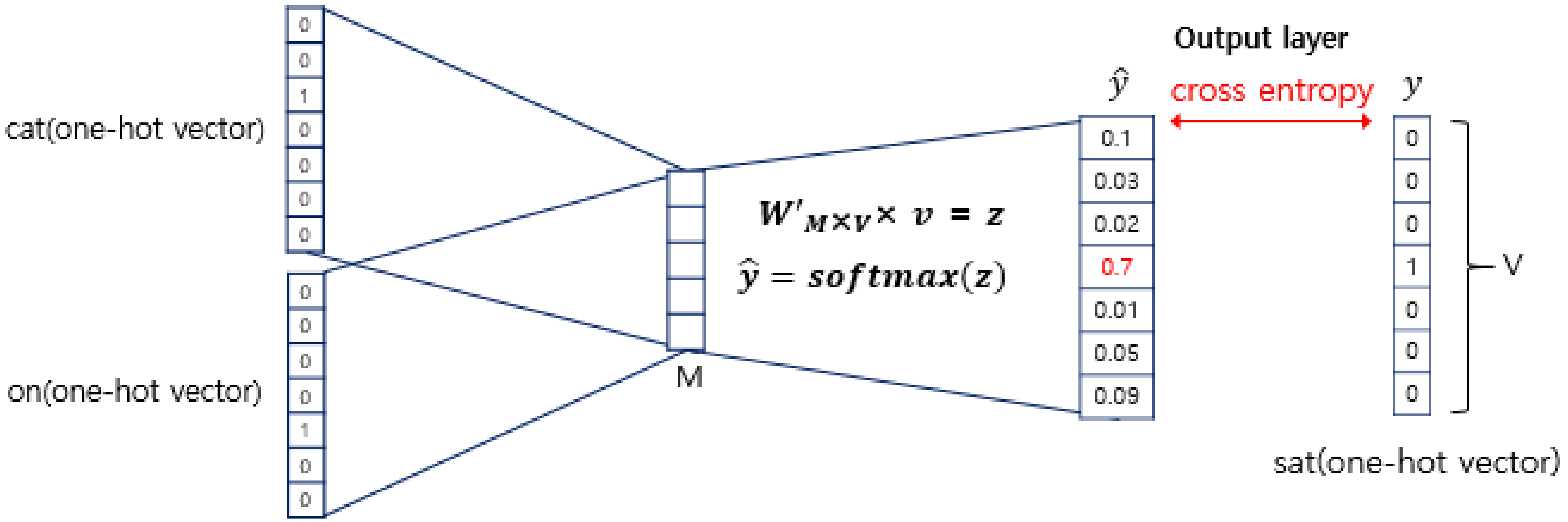
New Log-linear Model

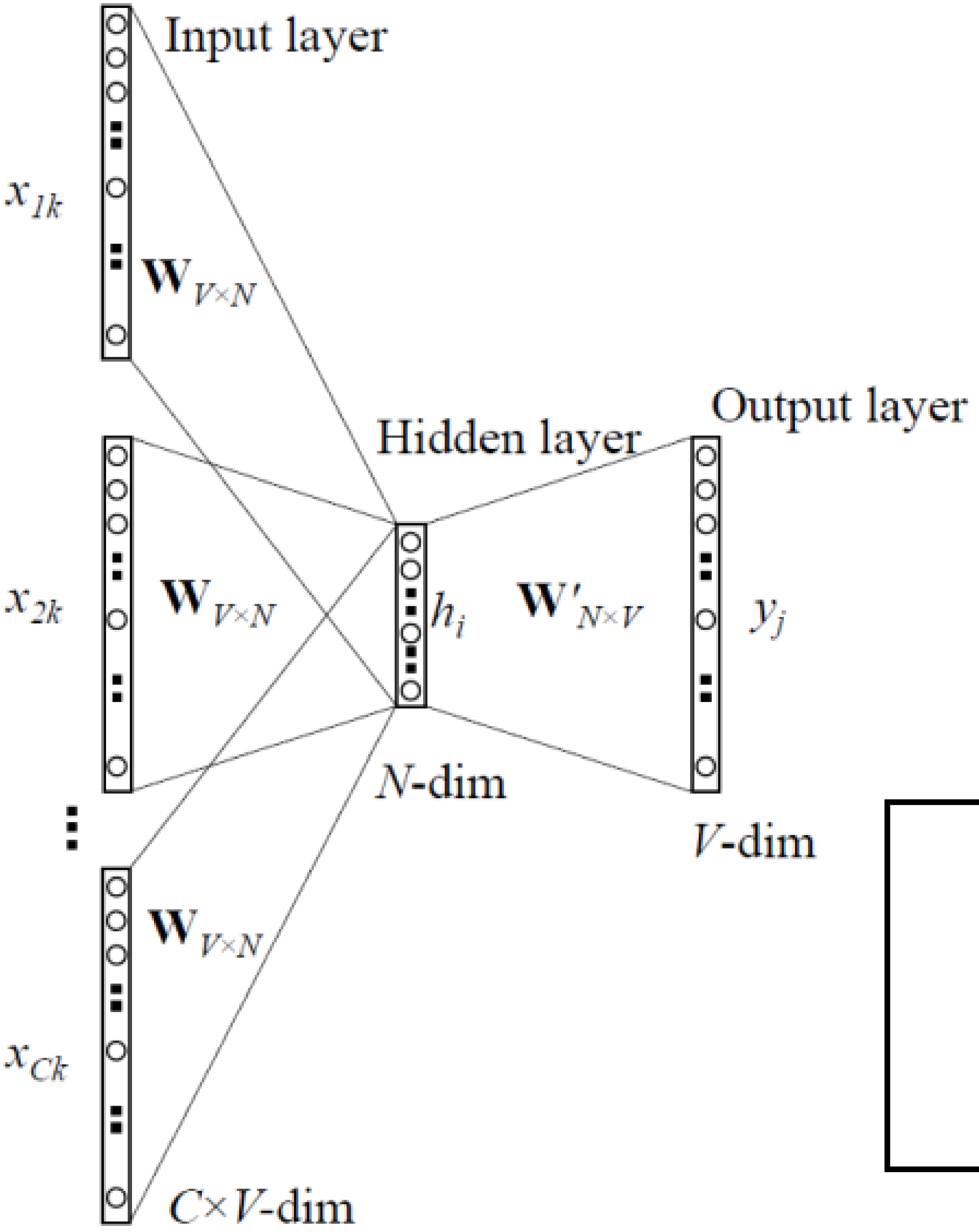


$$v = \frac{V_{fat} + V_{cat} + V_{on} + V_{the}}{2 \times n(winow\ size)}$$

03

New Log-linear Model



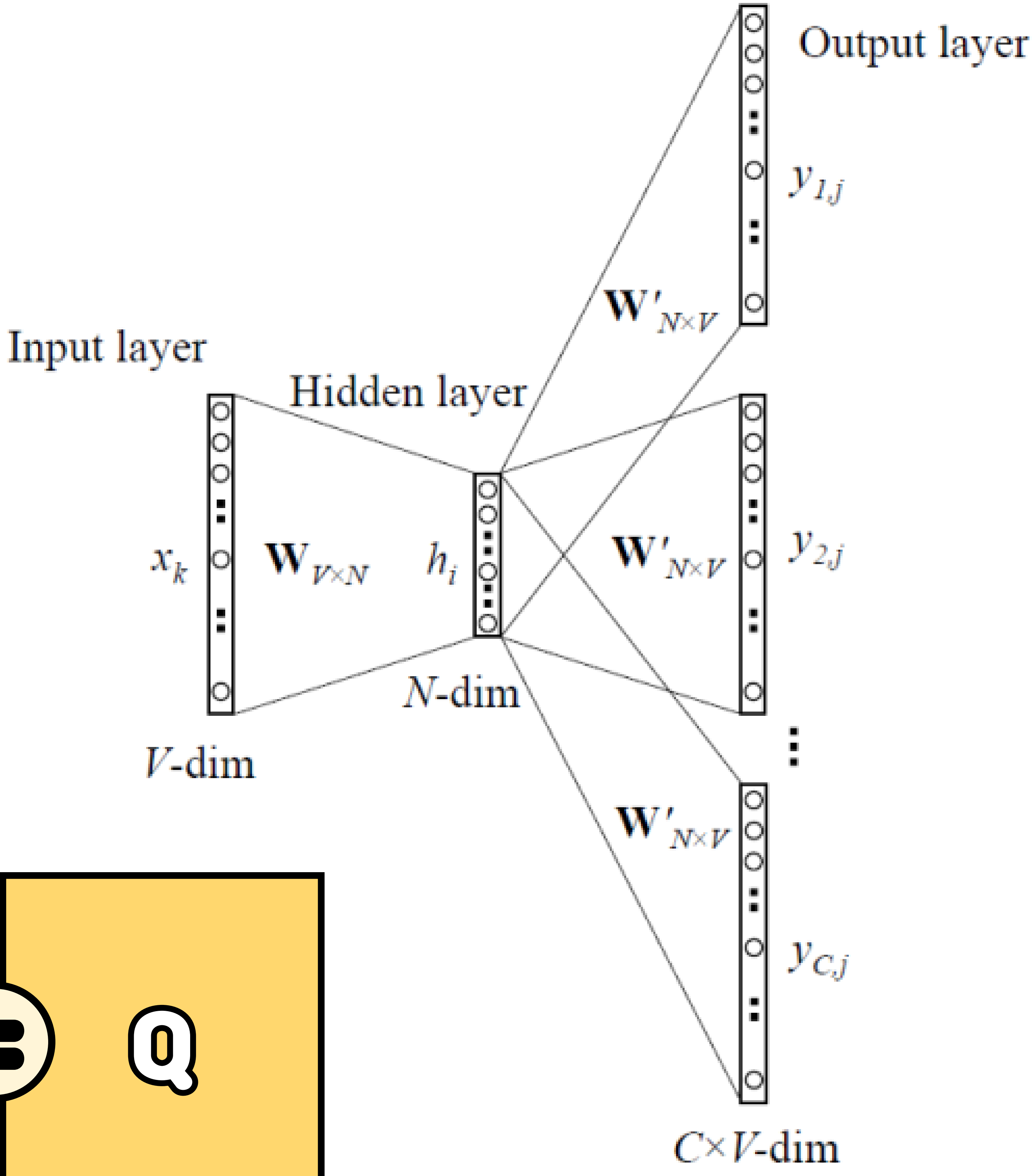


CBOW

Continous Bag-of-Words Model

$N \times D + D \times \log_2(V) = Q$

Skip-gram



$$C \times \left(D + D \times \log_2(V) \right) = Q$$

03 New Log-linear Model

- N = 10(정해진 N개의 단어 수)
  - 정해진 N개의 단어 수
- M = 500
  - 투사층의 크기
- H = 500
  - 은닉층의 크기
- C = 10
  - CBOW 모델에서 10내외
- V = 최대 1000만
  - 사전의 단어 수

$$Q_{(NNLM)} = N \times M + N \times M \times H + H \times ln(V)$$

$$= 10 \times 500 + \boxed{10 \times 500 \times 500} + 500 \times \ln(10{,}000{,}000) = \boxed{250만}$$

$$Q_{(RNNLM)} = H \times H + H \times ln(V)$$

$$= \boxed{500 \times 500} + 500 \times \ln(10{,}000{,}000) = \boxed{25만}$$

$$Q_{(Word2Vec_{CBOW})} = C \times M + M \times ln(V)$$

$$= 10 \times 500 + \boxed{500 \times \ln(10{,}000{,}000)} = \boxed{8{,}059}$$

$$Q_{(Word2Vec_{Skip-gram})} = C\ (M + M \times ln(V))$$

$$= 10 \times 500 + \boxed{10 \times (500 \times \ln(10{,}000{,}000))} = \boxed{80{,}590}$$

Analogy Reasoning Task

- 'semantic' 관계 5가지
- 'syntactic' 관계 9가지

vector(Greece)
− vector(Athens)
+ vector(Oslo)

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

Comparison of Model Architecture

Word vector dimensionality = 640

Model Architecture	Semantic-Syntactic Word Relationship test set		MSR Word Relatedness Test Set [20]
	Semantic Accuracy [%]	Syntactic Accuracy [%]	
RNNLM	9	36	35
NNLM	23	53	47
CBOW	24	64	61
Skip-gram	55	59	56

Comparison of Model Architecture

epoch = 1 or 3

Model	Vector Dimensionality	Training words	Accuracy [%]			Training time [days]
			Semantic	Syntactic	Total	
3 epoch CBOW	300	783M	15.5	53.1	36.1	1
3 epoch Skip-gram	300	783M	50.0	55.9	53.3	3
1 epoch CBOW	300	783M	13.8	49.9	33.6	0.3
1 epoch CBOW	300	1.6B	16.1	52.6	36.1	0.6
1 epoch CBOW	600	783M	15.4	53.3	36.2	0.7
1 epoch Skip-gram	300	783M	45.6	52.2	49.2	1
1 epoch Skip-gram	300	1.6B	52.2	55.1	53.8	2
1 epoch Skip-gram	600	783M	56.7	54.5	55.5	2.5

05

Examples of the Learned Relationships

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Skip-gram model trained on 784M words with 300 dimensionality

Thank you