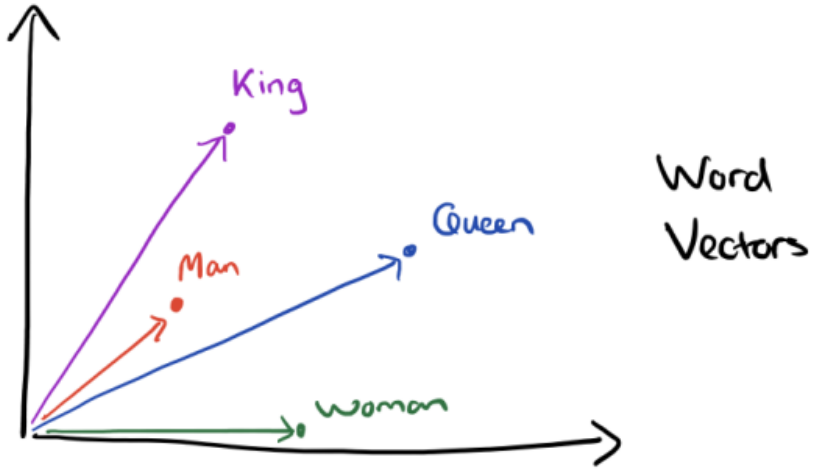
Word2Vec

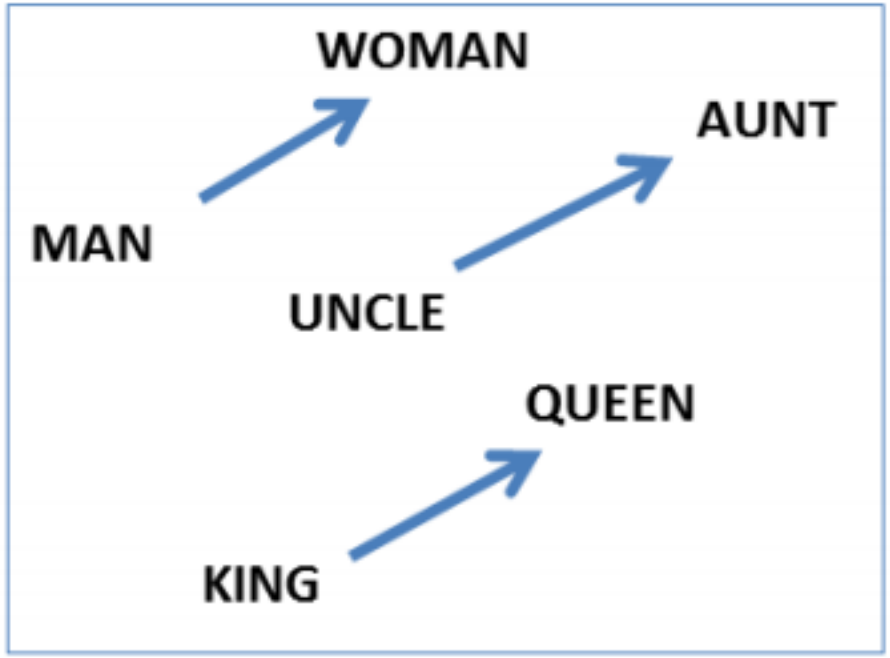


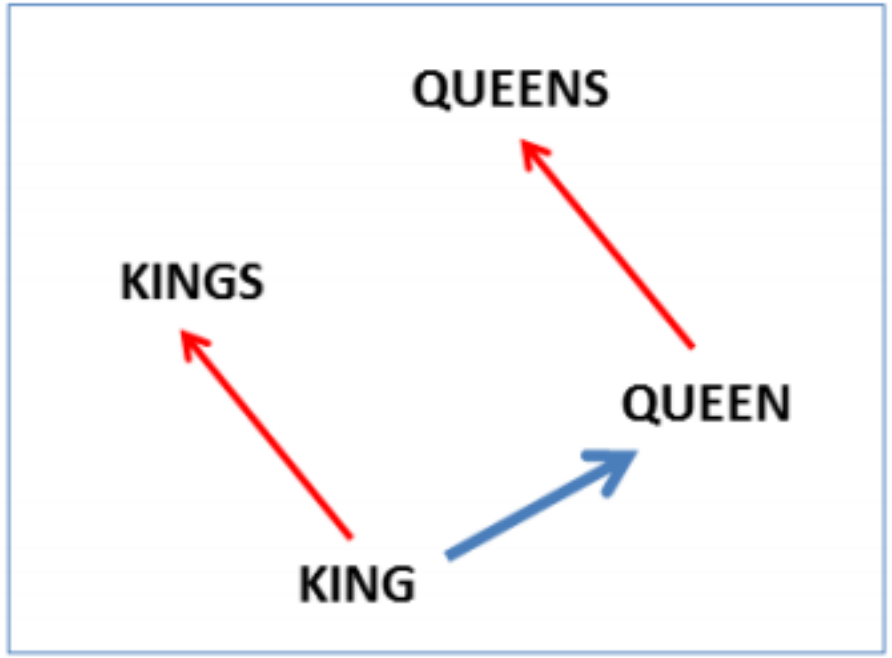
Fast Campus
Start Deep Learning with Tensorflow

Word2Vec

- RNN에 단어를 입력하려면, 단어를 숫자 즉 vector로 바꿔야 함
- Intuitive embedding – one hot encoding
  - Apple, Strawberry, Dog 세 단어가 있을 때,
  - Apple → [1, 0, 0]
  - Strawberry → [0, 1, 0]
  - Dog → [0, 0, 1]
- 장점
  - Easy!
- 단점
  - 단어들 간의 의미관계를 파악할 수 없음(apple과 strawberry, apple과 dog)
  - 단어가 많아지면?

We Want…





(Mikolov et al., NAACL HLT, 2013)

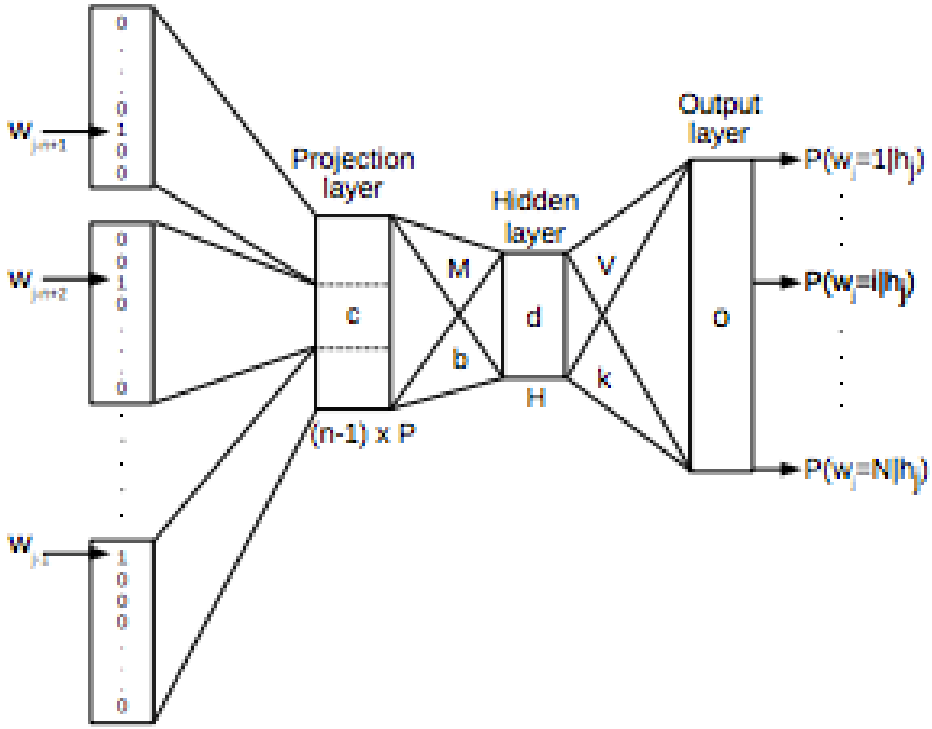
Let's Try It

- http://w.elnn.kr/search/

- How to train word vectors?????
  - MLP
  - RNN
  - ...

NNLM

- Input : 현재 단어 이전의 N개 단어의 one-hot vector들
- Output : probabilities(softmax)
- Projection layer : embedded vector(P)
- 단점
  - 이전 단어만 고려함
  - 매우 느리다



RNNLM

- NNLM을 RNN형태로 변경
- N을 정하지 않아도 됨
- 단점
  - 이전 단어만 고려함
  - 여전히 느리다

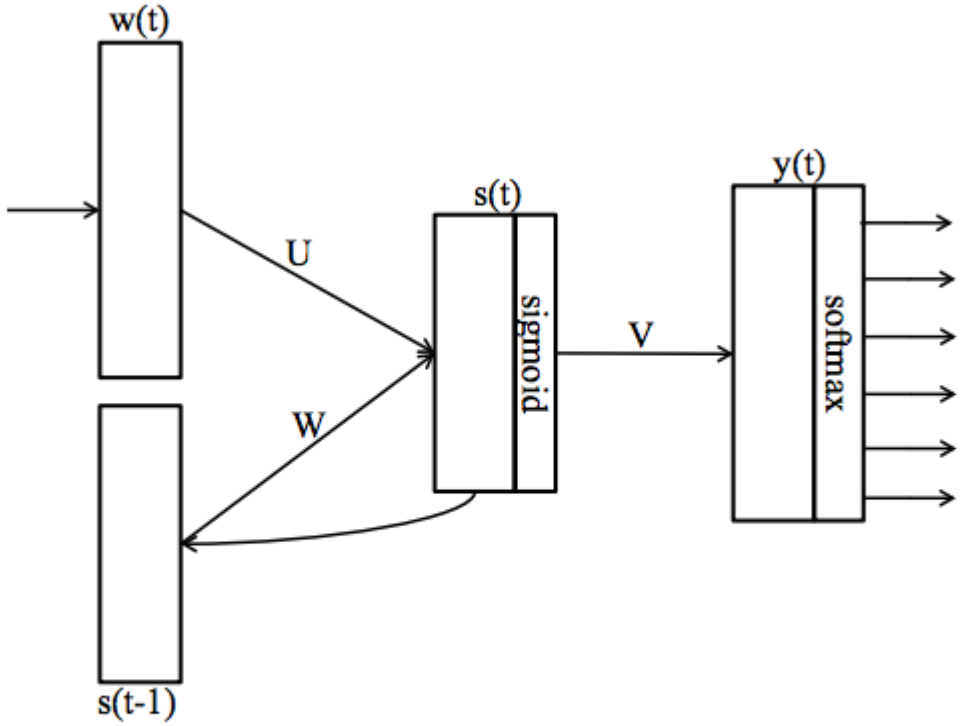
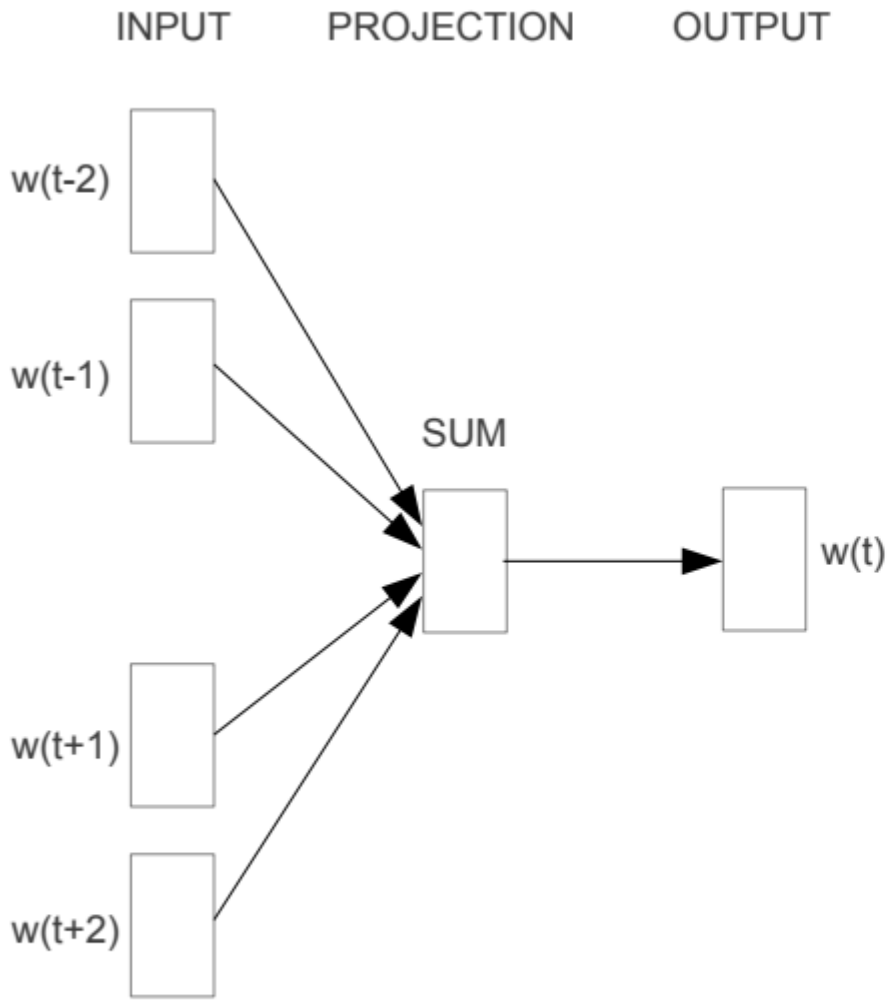
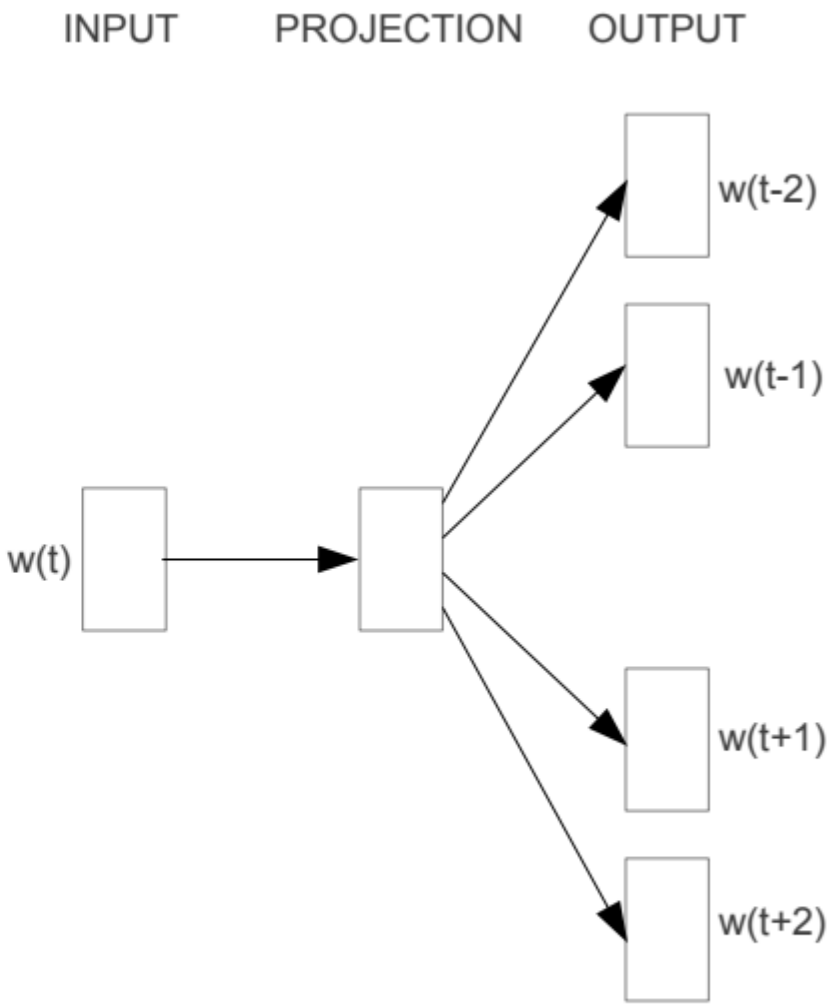


Figure 1: The architecture of the RNNLM.

CBOW & Skip-gram



CBOW



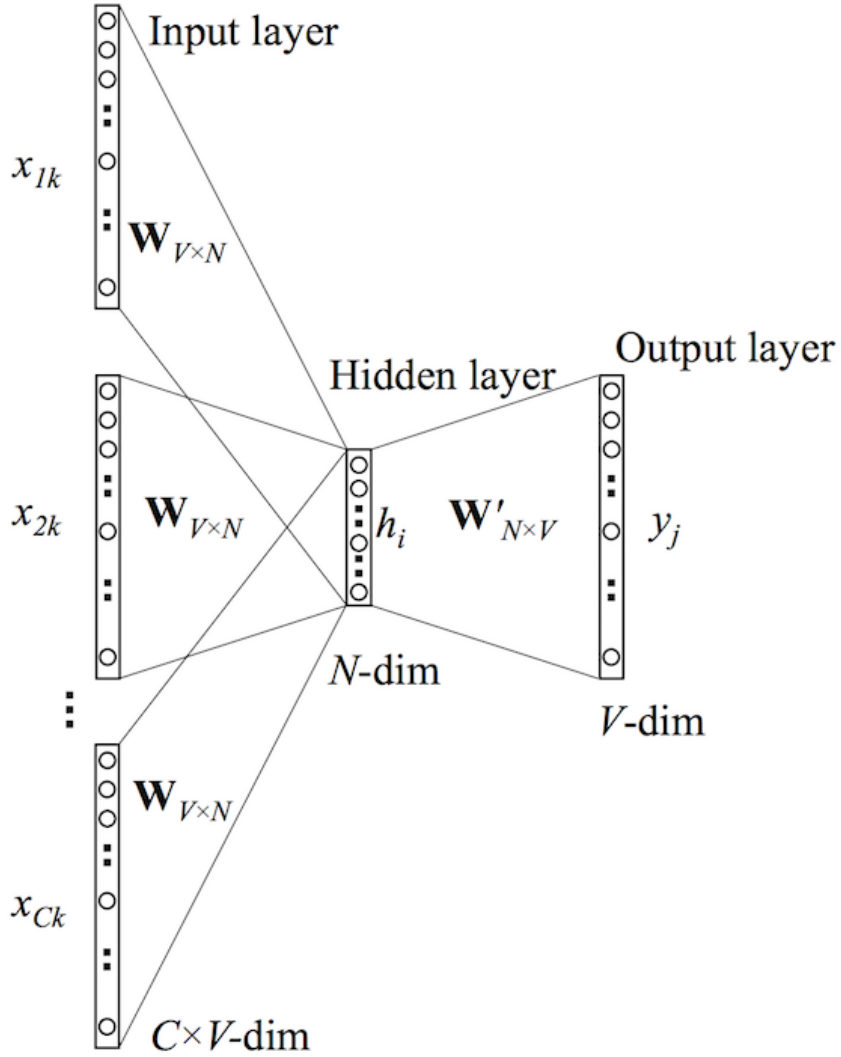
Skip-gram

CBOW – Continuous Bag of Words

- Similar to the feedforward NNLM, but
- Non-linear hidden layer removed
- Projection layer shared for all words
- Projected vectors are just averaged
- Called CBOW because the order of the words is lost
- Another modification is to use words from past and from future(window centered on current word)

CBOW – Continuous Bag of Words

- Fill the blank
  - 아이스크림을 사 먹었는데, ___ 시려서 먹기가 힘들었다.
- 앞 뒤로 C/2개의 단어를 input으로 하여 center 단어를 맞추도록 학습
- Input은 one-hot encoding
- Input → Hidden layer는 linear mapping(avg(Wx_{ik}))
- Hidden → Output layer는 Softmax($W'h_i$)



Skip-gram

- Similar to CBOW, but instead of predicting the current word based on the context
- Tries to maximize classification of a word based on another word in the same sentence
- Thus, uses each current word as an input to a log-linear classifier
- Predicts words within a certain window

Skip-gram

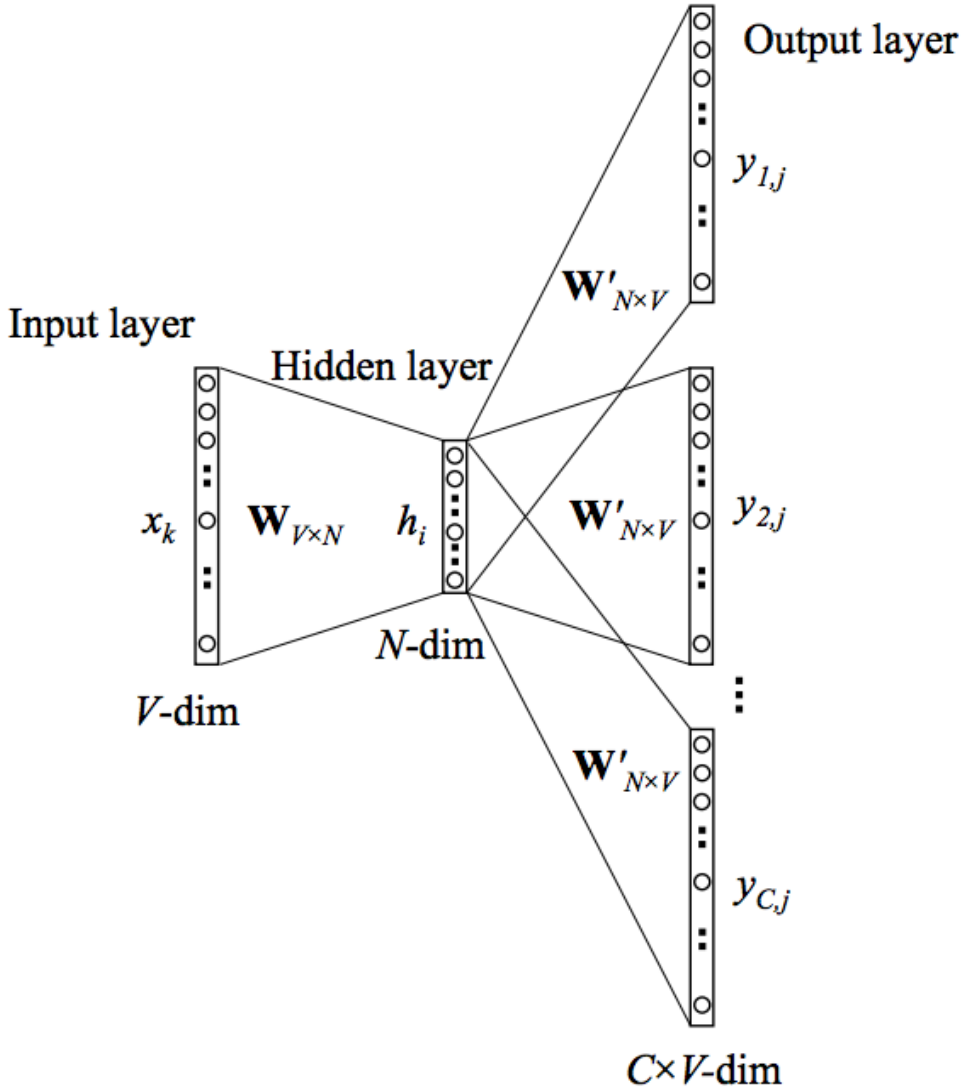
- CBOW와 반대로 중심 단어를 주고 주변 단어들에 대한 확률 값을 출력함
- Window 내에 있는 단어의 확률이 최대가 되도록 학습
- Objective function
  - Maximize

$$J'(\theta) = \prod_{t=1}^{T} \prod_{-C/2 \le j \le C/2, j \ne 0} P(x_{t+j} | x_t ; \theta)$$

→ Negative log likelihood

  - Minimize

$$J(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{-C/2 \le j \le C/2, j \ne 0} log P(x_{t+j} | x_t ; \theta)$$

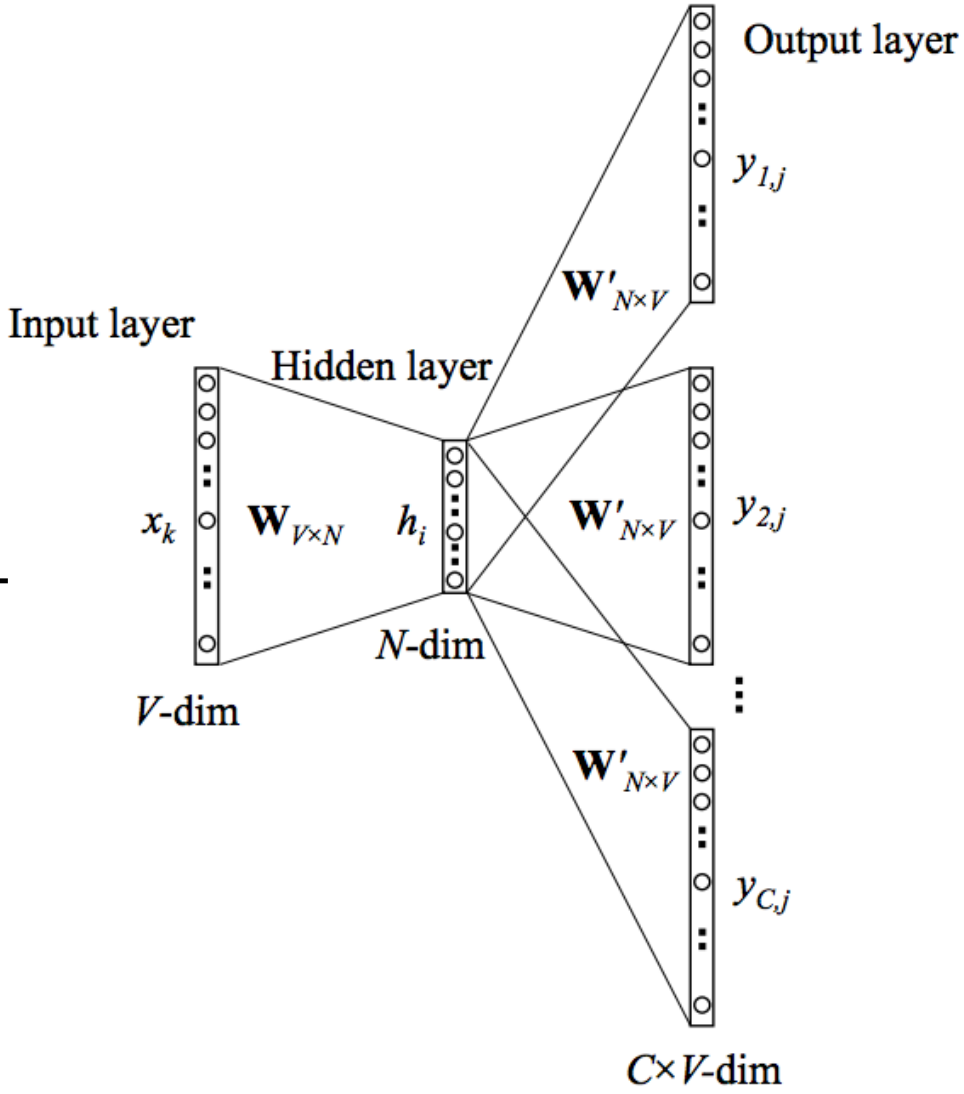


Skip-gram

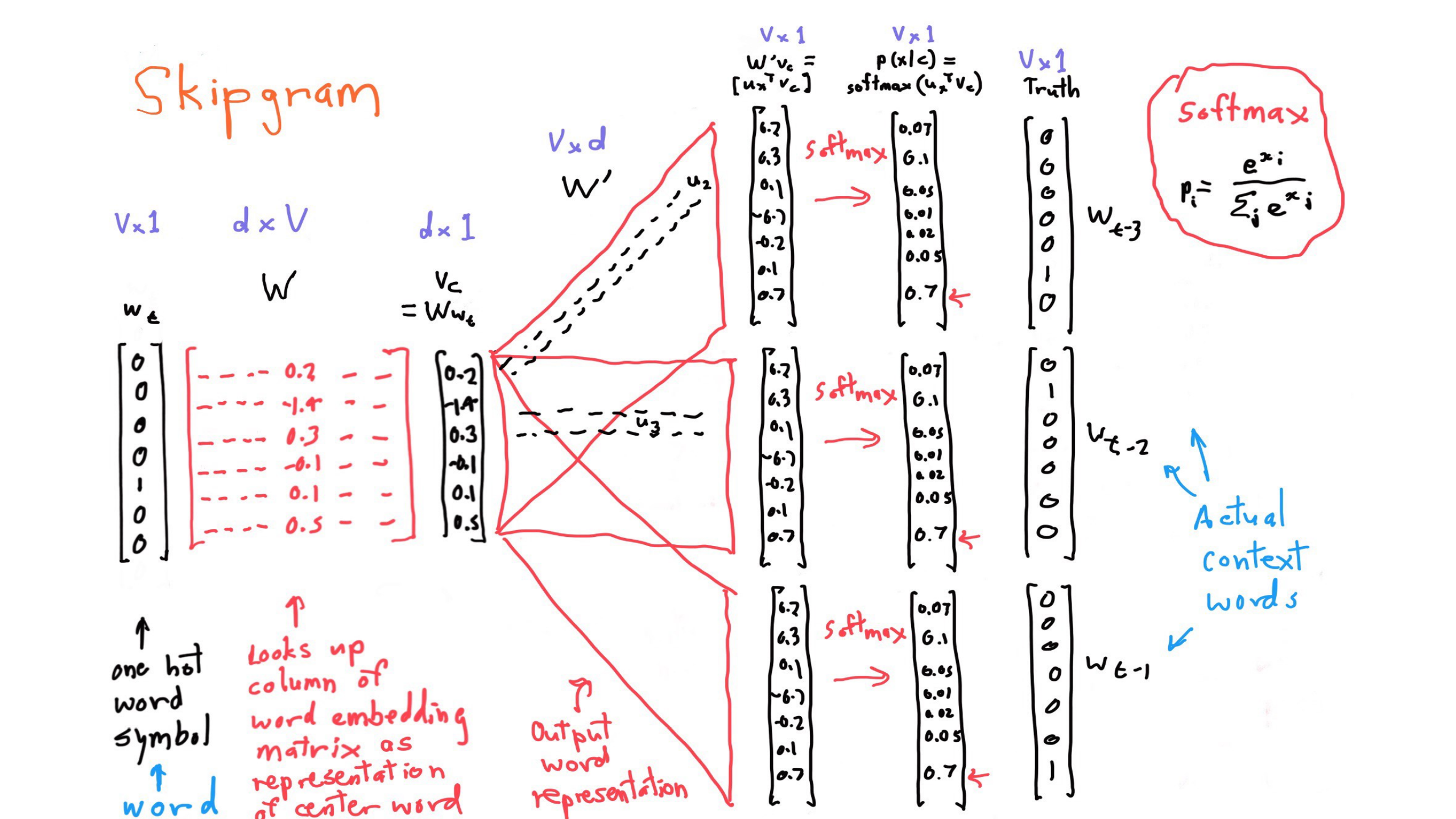
- $P(o|c) = \frac{\exp({u_o}^T \cdot v_c)}{\sum_{w=1}^{V} \exp({u_w}^T \cdot v_c)}$
- $J(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{-C/2 \le j \le C/2, j \neq 0} logP(x_{t+j}|x_t;\theta)$

에서, P의 분모를 구하려면 모든 단어에 대해 고려해야 한다

- Negative sampling 방법으로 window 밖에 있는 단어를 일부만 sampling 하여 근사한다 → NCE(Noise Contrastive Estimation)



Skipgram


V x 1
d x V
d x 1
V x d
W'
W
w_t
$v_c = W w_t$
$W' v_c = [u_x^T v_c]$
$p(x|c) = \text{softmax}(u_x^T v_c)$
V x 1
Truth
Softmax
$p_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$
u_2
u_3
Softmax
W_{t-3}
W_{t-2}
W_{t-1}
Actual context words
one hot word symbol
word
Looks up column of word embedding matrix as representation of center word
Output word representation

Result

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