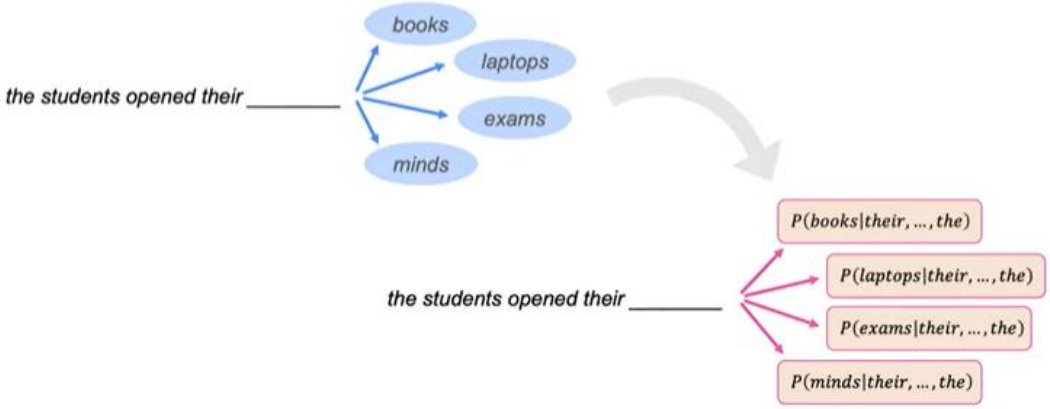
CS224N Lecture6. Language Models and Recurrent Neural Networks

* Language Model 이란?

- 단어의 시퀀스에 대해서 얼마나 자연스러운 문장인지를 확률을 통해 예측

- 주어진 단어의 시퀀스에 대해서 다음에 나타날 단어가 어떤 것인지 예측하는 작업을 Language Modeling이라고 한다.



$$w_1, w_2, \cdots, w_{t-n+1}, \cdots, w_{t-1}, w_t, \cdots, w_{T-1}, w_T$$

$$P(w_1, \ldots, w_T) = P(w_1) \times P(w_2|w_1) \times \cdots \times P(w_T|w_{T-1}, \ldots, w_1)$$

$$= \prod_{t=1}^{T} P(w_t|w_{t-1}, \ldots, w_1)$$

* n-gram Language Models

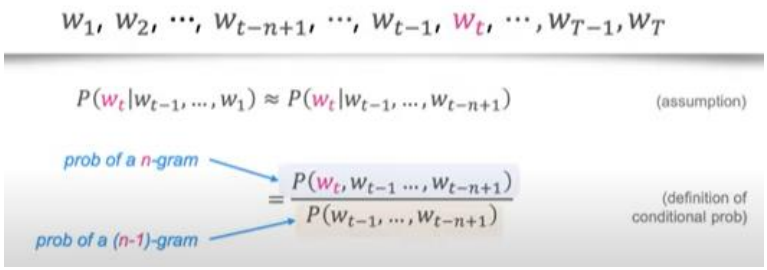
- Neural Network 이전에 사용되었던 Language Model

- 예측에 사용할 앞 단어들의 개수를 정하여 모델링하는 방법

(uni-grams, bi-grams, tri-grams, 4-grams)

(앞 단어 n개를 가지고 다음에 올 단어를 예측한다.)

- uni-grams: *"the", "students", "opened", "their"*
- bi-grams: *"the students", "students opened", "opened their"*
- tri-grams: *"the students opened", "students opened their"*
- 4-grams: *"the students opened their"*



* n-gram Language Models의 문제점

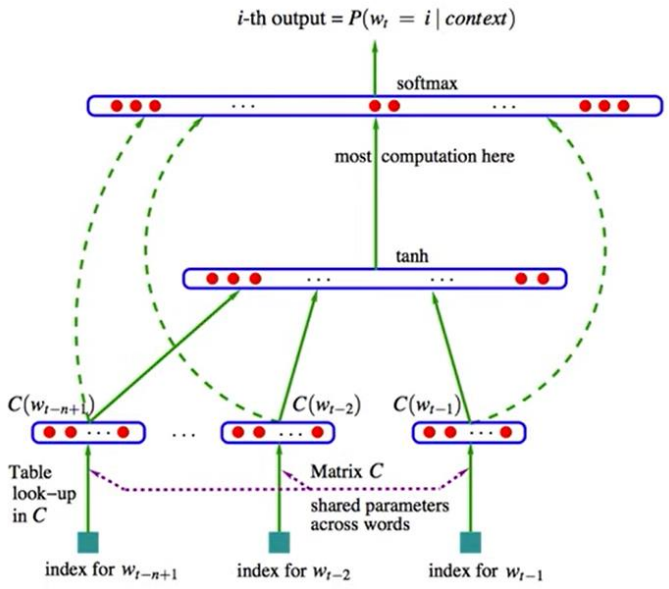
- Sparsity problems: n이 커질수록 성능이 떨어지며, 일반적으로 n 값으로 5가 되지 않는 수를 설정함.

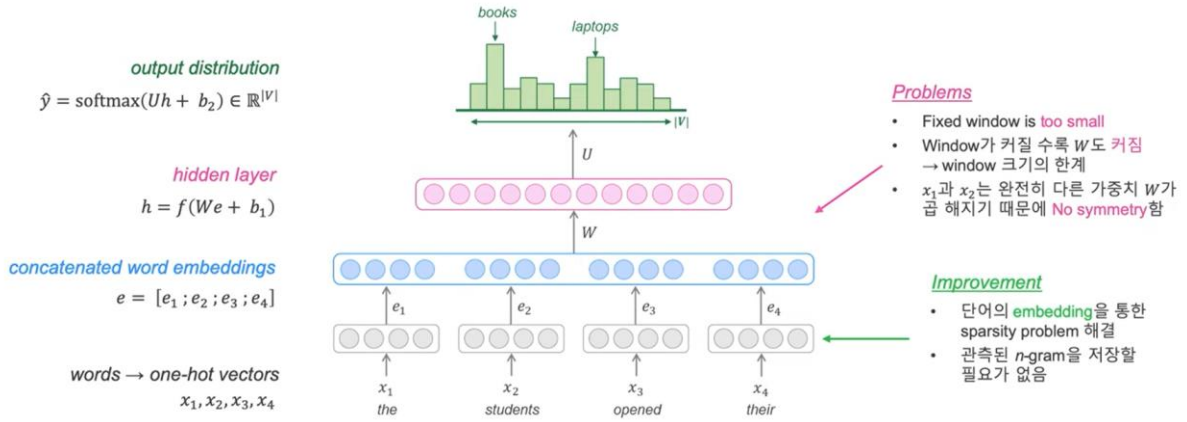
- Storage problems: n이 커지거나 corpus가 증가하면 모델의 크기가 증가함.

* Window-based Neural Network Language Model(NNLM)

- 'curse of dimensionality'를 해결하기 위해 제안된 신경망 기반 Language Model

- Language Model이면서 동시에 단어의 'distributed representation' 학습





* Recurrent Neural Networks(RNN)

- Take sequential input of any length

- Apply the same weights on each step

- Can optionally produce output on each step