Practical Issue: Masking on Attention

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

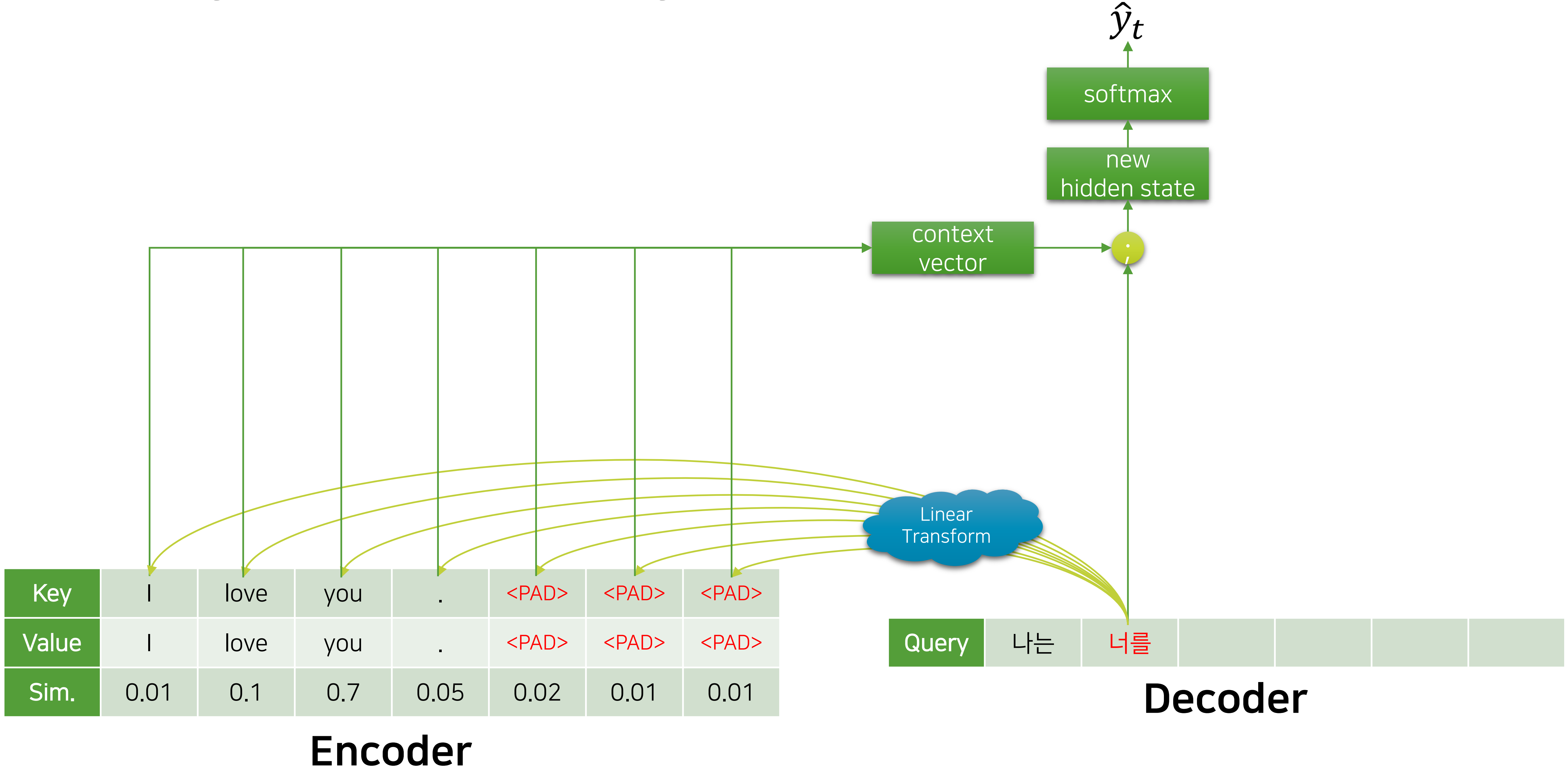
Motivation

- We always do mini-batch parallelized operations.
  - Thus, attention weight can be assigned in empty spot, too.

- This can be serious problem at inference.

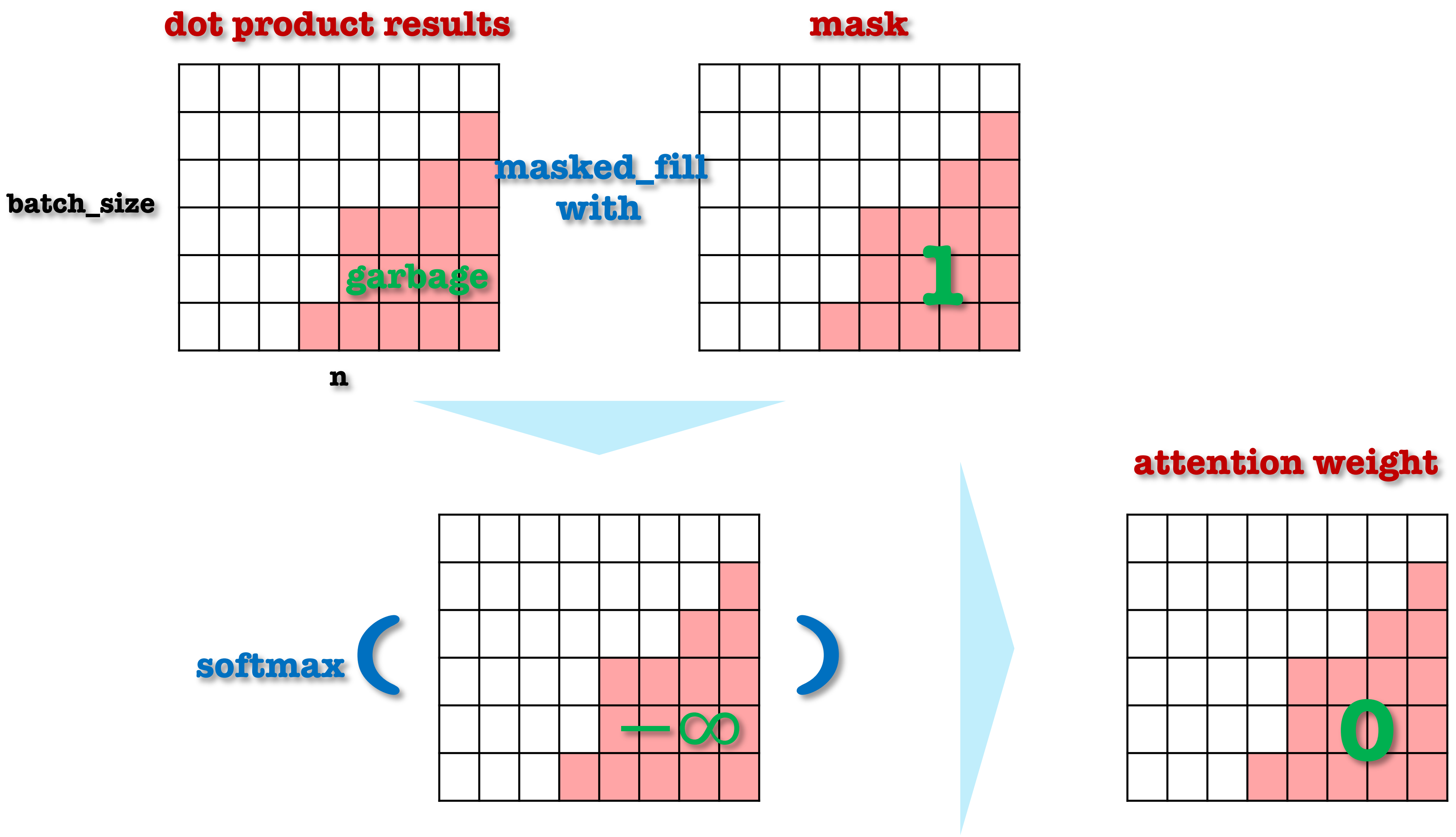
I	love	to	go	to	school	.	<PAD>	<PAD>	<PAD>
All	you	need	is	attention	.	<PAD>	<PAD>	<PAD>	<PAD>
I	ate	special	dinner	with	her	in	the	room	.
RNN	can	not	memorize	every	detail	.	<PAD>	<PAD>	<PAD>

Assign Attention Weights to <PAD>



Solution

- Using mask, assign $-\infty$ to make 0s for softmax results.



In Equations

- After dot-products, before softmax.

$$w = \text{softmax}(h_t^{\text{dec}} \cdot W_{\text{a}} \cdot {h_{1:m}^{\text{enc}}}^T)$$

$$c = w \cdot h_{1:m}^{\text{enc}},$$

$$\text{where } c \in \mathbb{R}^{\text{batch\_size} \times 1 \times \text{hidden\_size}} \text{ is a context vector, and } W_{\text{a}} \in \mathbb{R}^{\text{hidden\_size} \times \text{hidden\_size}}.$$

Summary

- Mini-batch 내의 문장 구성에 따라, <pad>가 동적으로 생성됨
  - <pad>의 hidden state에는 attention weight가 할당되면 안됨

- 따라서, Key와 Query의 dot product 이후에 (softmax 이전에), masking을 통해 <pad> 위치의 값을 음의 무한대로 변경
  - softmax 결과 <pad>에는 0이 할당됨

- 이 기법은 이후 Transformer에서도 유용하게 쓰일 것