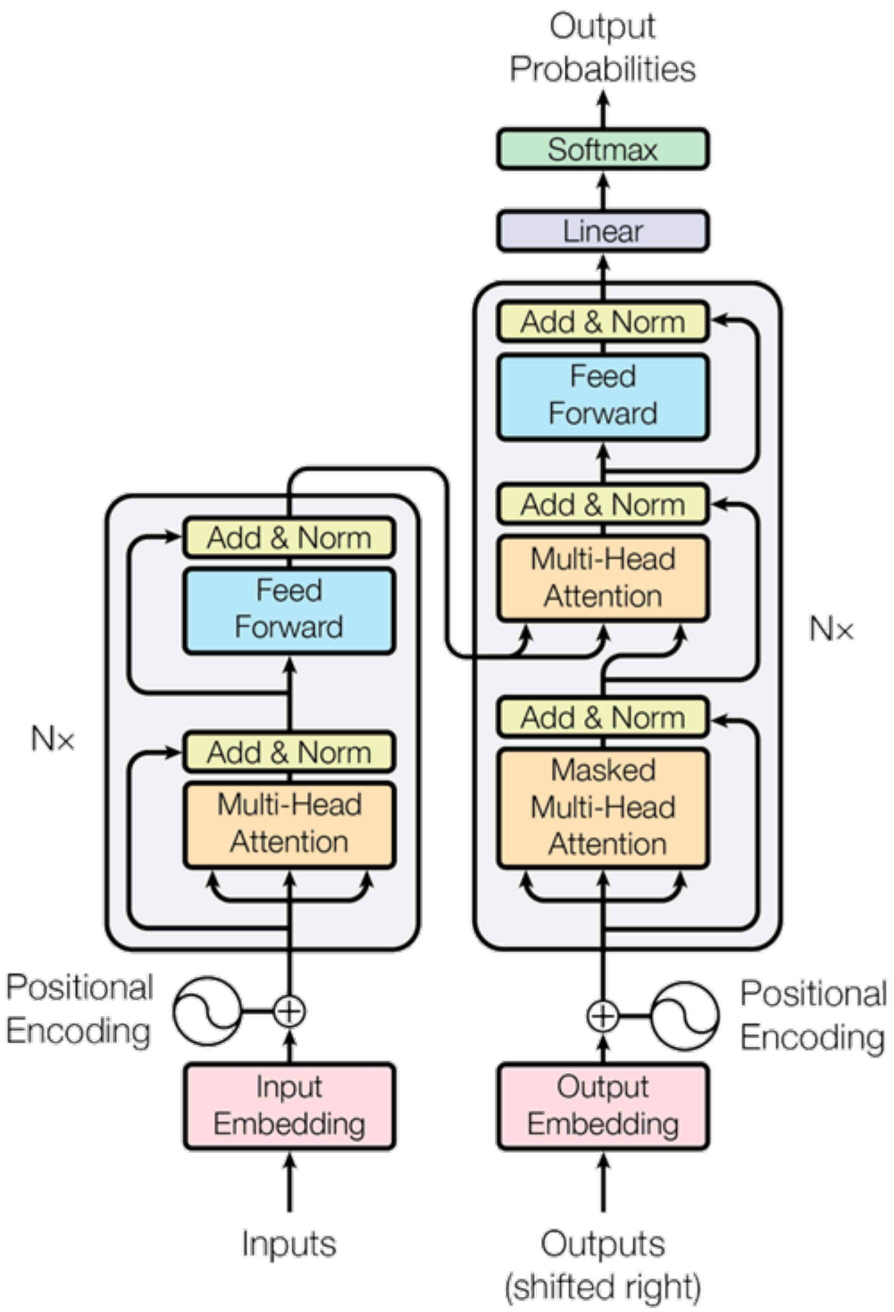
Transformer: Decoder Block with Masks

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Transformer



Equations

- Given Dataset,

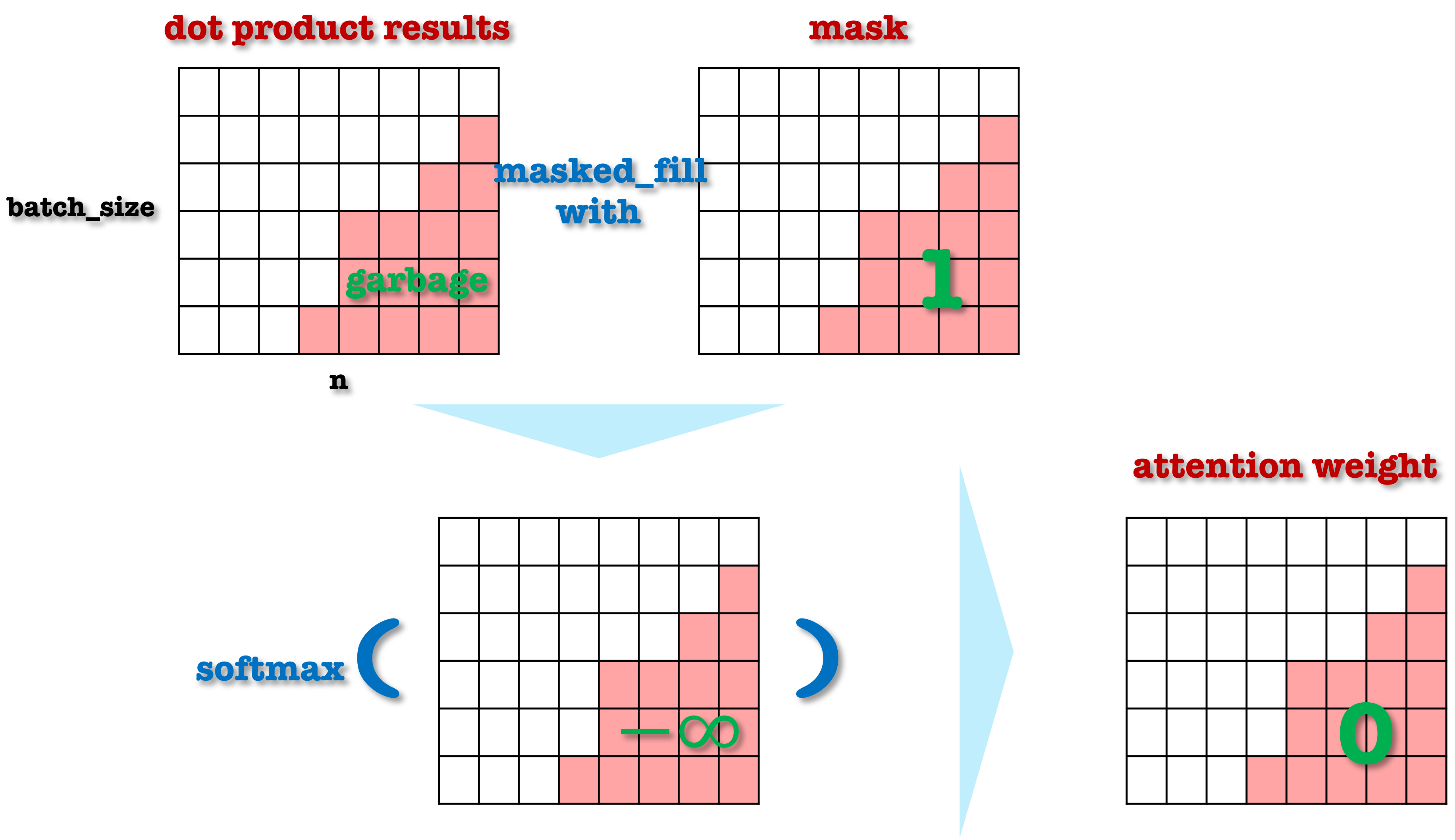
$$\mathcal{D} = \{x^i, y^i\}_{i=1}^{N}$$
$$x^i = \{x_1^i, \cdots, x_m^i\} \text{ and } y^i = \{y_0^i, y_1^i, \cdots, y_n^i\},$$
$$\text{where } y_0 = \text{<BOS>} \text{ and } y_n = \text{<EOS>}.$$

- What we want is

$$\hat{y}_{1:n} = f(x_{1:m} : \theta)$$

Before we start,

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



Equations

- Self-attention with mask

$$
\begin{gathered}
h_{0,1:n} = \text{emb}(y_{0:n-1}) + \text{pos}(0, n-1) \\
\tilde{h}^{\text{dec}}_{i,1:n} = \text{LayerNorm}(\text{Multihead}_i(Q, K, V) + h^{\text{dec}}_{i-1,1:n}), \\
\text{where } Q = K = V = h^{\text{dec}}_{i-1,1:n}.
\end{gathered}
$$

Equations

- Attention from encoder with mask for <pad>

$$\tilde{h}^{\text{dec}}_{i,1:n} = \text{LayerNorm}(\text{Multihead}_i(Q, K, V) + h^{\text{dec}}_{i-1,1:n}),$$

$$\text{where } Q = \tilde{h}^{\text{dec}}_{i,1:n} \text{ and } K = V = h^{\text{dec}}_{\ell,1:m}.$$

Equations

- FC layers

$$\mathrm{FFN}(h_{i,t}) = \mathrm{ReLU}(h_{i,t} \cdot W_i^1) \cdot W_i^2$$
$$\text{where } W_i^1 \in \mathbb{R}^{d_{\text{model}} \times d_{\text{ff}}} \text{ and } W_i^2 \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{model}}}.$$

$$h_{i,1:m}^{\text{dec}} = \mathrm{LayerNorm}([\mathrm{FFN}(\tilde{h}_{i,1}^{\text{dec}}); \cdots; \mathrm{FFN}(\tilde{h}_{i,m}^{\text{dec}})] + \tilde{h}_{i,1:m}^{\text{dec}})$$

Equations

- Decoder is stack of decoder blocks.

$$
\begin{aligned}
h^{\text{dec}}_{\ell_{\text{dec}},1:m} &= \text{Block}_{\text{dec}}(h^{\text{dec}}_{\ell_{\text{dec}}-1,1:m}) \\
&\cdots \\
h^{\text{dec}}_{1,1:m} &= \text{Block}_{\text{dec}}(h^{\text{dec}}_{0,1:m})
\end{aligned}
$$

- Generator:

$$
\hat{y}_{1:n} = \text{softmax}(h^{\text{dec}}_{\ell_{\text{dec}},1:m} \cdot W_{\text{gen}}),
$$

$$
\text{where } h^{\text{dec}}_{\ell_{\text{dec}},1:m} \in \mathbb{R}^{\text{batch\_size}\times n\times\text{hidden\_size}} \text{ and } W_{\text{gen}} \in \mathbb{R}^{\text{hidden\_size}\times|V|}.
$$

Summary

- Decoder는 2가지의 attention으로 구성됨
  - Attention from encoder:
    - K와 V는 encoder의 최종 출력 값, Q는 이전 레이어의 출력 값
  - Self-attention with mask:
    - Q, K, V는 이전 레이어의 출력 값
    - Attention weight 계산 시, softmax 연산 이전에 masking을 통해 음의 무한대를 주어, 미래 time-step을 보는 것을 방지

- 추론 때에는 self-attention의 mask는 필요 없으나, 모든 layer의 t 시점 이전의 모든 time-step($< t$)의 hidden state가 필요