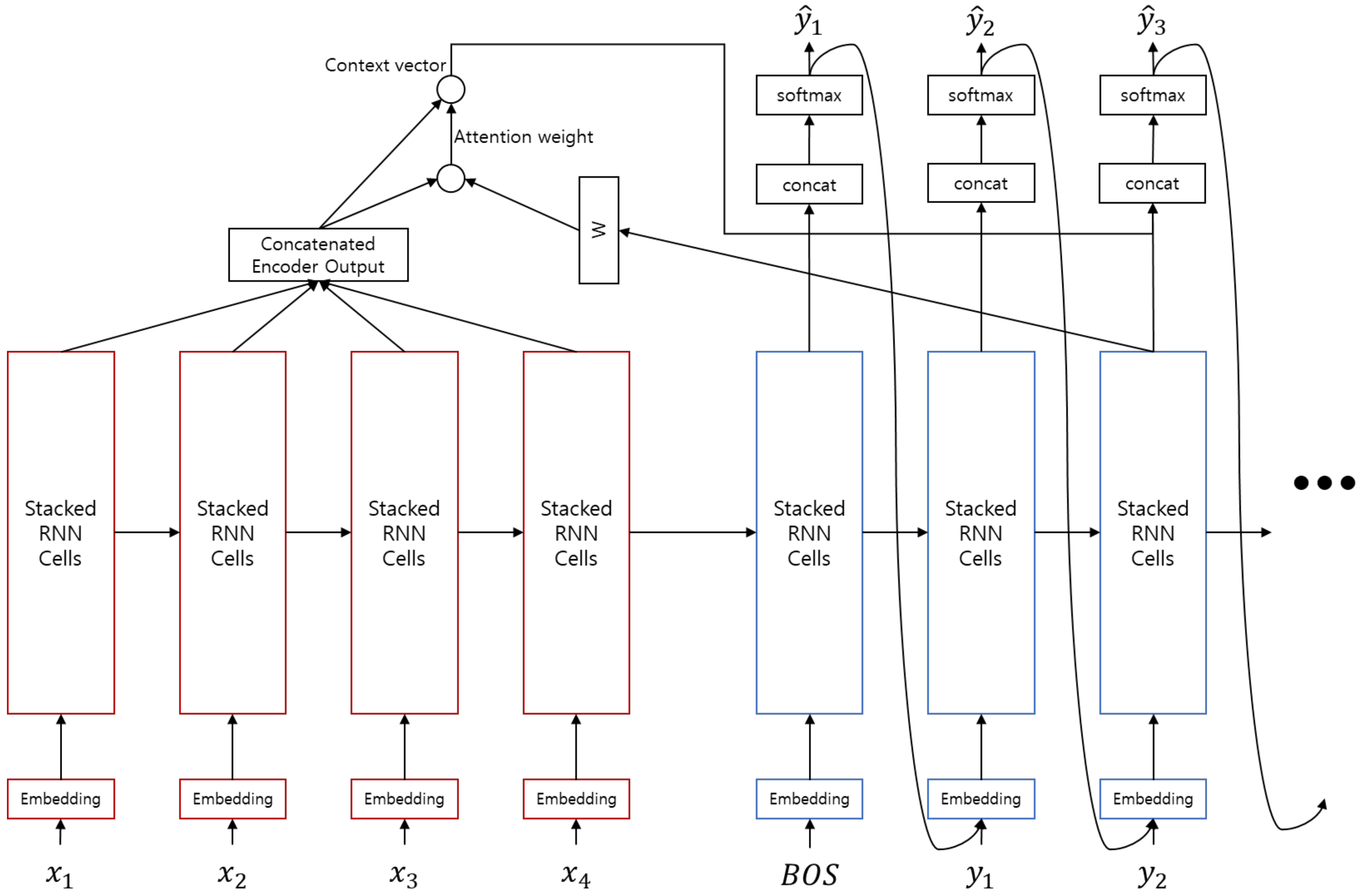
Input Feeding

Ki Hyun Kim

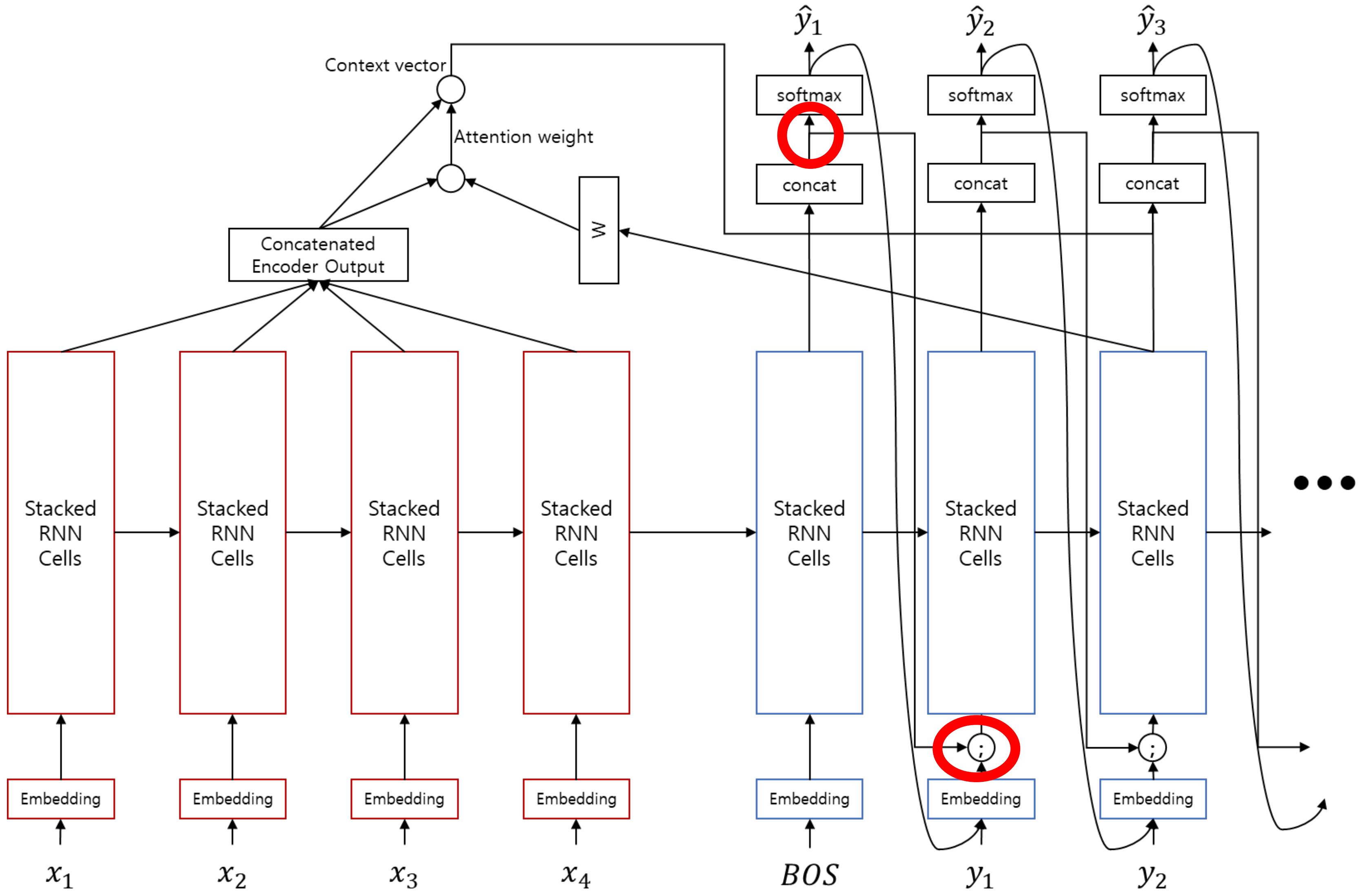
nlp.with.deep.learning@gmail.com

Motivation



$\hat{y}_1$
$\hat{y}_2$
$\hat{y}_3$
Context vector
softmax
softmax
softmax
Attention weight
concat
concat
concat
W
Concatenated Encoder Output
Stacked RNN Cells
Stacked RNN Cells
Stacked RNN Cells
Stacked RNN Cells
Stacked RNN Cells
Stacked RNN Cells
Stacked RNN Cells
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
x_1
x_2
x_3
x_4
BOS
y_1
y_2

Input Feeding



By Input Feeding,

- Sampling 과정에서 손실 되는 정보를 최소화

- Teacher Forcing으로 인한 학습/추론 사이의 괴리를 최소화

Full Equations

$$\mathcal{D} = \{x^i, y^i\}_{i=1}^N$$

$$x^i = \{x_1^i, \cdots, x_m^i\} \text{ and } y^i = \{y_0^i, y_1^i, \cdots, y_n^i\},$$

$$\text{where } y_0 = \text{<BOS> and } y_n = \text{<EOS>}.$$

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

$$h_{1:m}^{\text{enc}} = \text{RNN}_{\text{enc}}(\text{emb}_{\text{enc}}(x_{1:m}), h_0^{\text{enc}}), \text{ where } h_0^{\text{enc}} = 0.$$

$$h_t^{\text{dec}} = \text{RNN}_{\text{dec}}([\text{emb}_{\text{dec}}(y_{t-1}); \tilde{h}_{t-1}^{\text{dec}}], h_{t-1}^{\text{dec}}),$$

$$\text{where } h_0^{\text{dec}} = h_m^{\text{enc}}.$$

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

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

$$\text{where } W_{\text{a}} \in \mathbb{R}^{\text{hidden\_size} \times \text{hidden\_size}}.$$

$$\tilde{h}_t^{\text{dec}} = \tanh([h_t^{\text{dec}}; c] \cdot W_{\text{concat}})$$

$$\hat{y}_t = \text{softmax}(\tilde{h}_t^{\text{dec}} \cdot W_{\text{gen}}),$$

$$\text{where } W_{\text{concat}} \in \mathbb{R}^{(2 \times \text{hidden\_size}) \times \text{hidden\_size}} \text{ and } W_{\text{gen}} \in \mathbb{R}^{\text{hidden\_size} \times |V|}.$$

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} \sum_{t=1}^{n} \log P(y_t^i | x^i, y_{<t}^i; \theta)$$

$$= -\sum_{i=1}^{N} \sum_{t=1}^{n} y_t^{i\intercal} \cdot \log \hat{y}_t^{\,i}$$

$$\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}(\theta)$$

Wrap-up

Encoder

- 문장을 받아 context vector로 압축
- Bi-directional RNN을 통해 구현

Decoder & Generator

- Conditional Language Model
  - Encoder로부터 정보를 받아 문장을 생성
- Cross Entropy(PPL)를 통해 최적화

Attention

- Key-value 함수
- 디코더의 hidden state를 인코더의 각 hidden state에 유사도 비교
- 좋은 query를 만들어내는 과정을 학습

Input Feeding

- Sampling 과정에서 손실된 정보를 word embedding에 concatenate