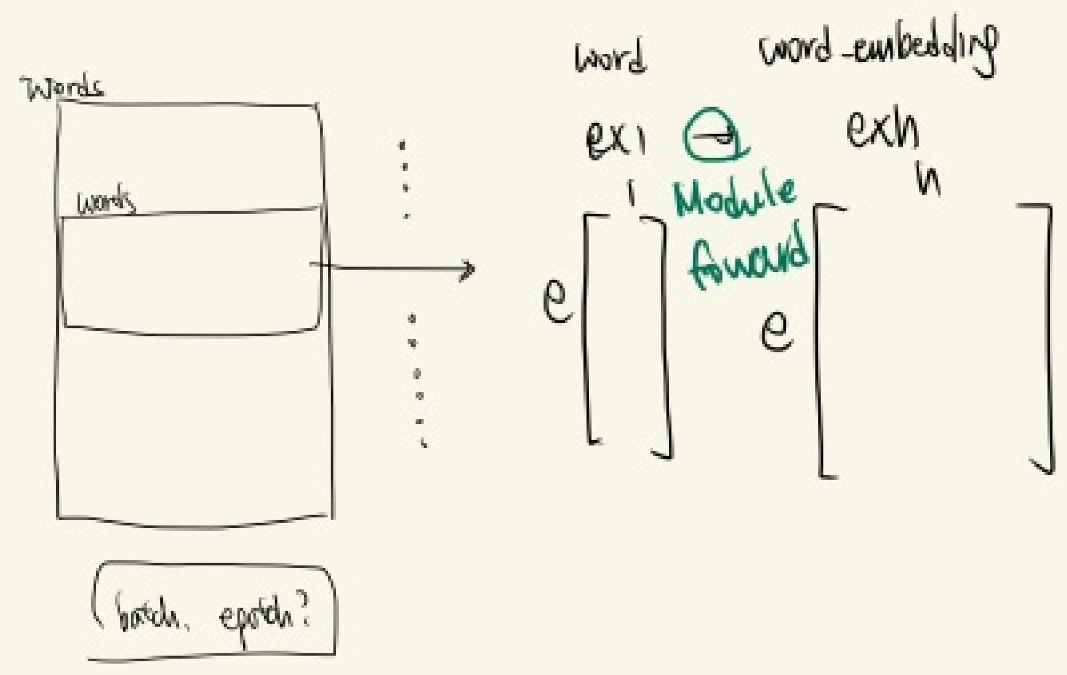
Blockdiagram_rnn_attention-1.jpg

$\mathbb{R}$



encoder

$W_h \in \mathbb{R}^{h\times 2h}$ · projection

$h^{enc} = [\overrightarrow{h}^{enc}; \overleftarrow{h}^{enc}] \in \mathbb{R}^{2h\times 1}$, $\mathbb{R}^{h\times 1}$

W_c. $c^{enc} = [\overrightarrow{c}^{enc}; \overleftarrow{c}^{enc}] \in \mathbb{R}^{2h\times 1}$, $\mathbb{R}^{h\times 1}$

2h

$$e_{t,j} = (h_t^{dec})^T W_{attproj} h_i^{enc} \rightarrow \alpha_t = \mathrm{softmax}(e_t) \rightarrow a_t = \sum_{i=1}^{m} \alpha_{t,i} h_i^{enc}$$

$m\times 1$, $h\times 1$, $h\times 2h$, $2h\times 1$; $m\times 1$; $2h\times 1$

m개의 LTSM.

? decoder를 진행중에 encoder에 넣을 순 없을까? ex, Greedy coding
이전 decoder element를 계산하는 것.

h 대신 사용하면 안되나.
분포도가 높은 성분을 더 높게 평가가능

$$h_0^{dec} = W_h[\overleftarrow{h}_1^{enc}; \overrightarrow{h}_m^{enc}]$$

$h\times 1$, $h\times 2h$, $h\times 1$

$$c_0^{dec} = W_c[\overleftarrow{c}_1^{enc}; \overrightarrow{c}_m^{enc}]$$

Bidectional (→←)
LTSM (h, c)

enc
← LTSM_m ··· LTSM_0
→ LTSM_0 ··· LTSM_m

dec
LTSM_0 ····· LTSM_M

$$h_t^{dec}, c_t^{dec} = \mathrm{Decoder}(\bar{y}_t, h_{t-1}^{dec}, c_{t-1}^{dec})$$

$h\times 1$, $h\times 1$

$$u_t = [a_t; h_t^{dec}]$$

$3h\times 1$

$$v_t = W_u u_t$$

$h\times 1$, $h\times 3h$, $3h\times 1$

$$o_t = \mathrm{dropout}(\tanh(v_t))$$

$h\times 1$

decoder

model.foward

$$P_t = \mathrm{softmax}(W_{vocab} o_t) \rightarrow J_t(\theta) = \mathrm{CrossEntropy}(P_t, g_t)$$

$V_t\times 1$, $V_t\times h$

one-hot vector

Blockdiagram_rnn_attention-2.jpg

Encoder

$W_h \in \mathbb{R}^{h \times 2h}$ projection

$h^{enc} = [\overrightarrow{h}^{enc}_i \; \overleftarrow{h}^{enc}] \in \mathbb{R}^{2h \times 1}$

$2h \begin{bmatrix} 1 \\ \\ \end{bmatrix}$

$W_c \in \mathbb{R}^{h \times 2h}$. $c^{enc} = [\overrightarrow{c}^{enc} \; \overleftarrow{c}^{enc}] \in \mathbb{R}^{2h \times 1}$

attention

$e_{t,j} = (h_t^{dec})^T \; W_{attproj} \; h_{i.}^{enc}$ — $m \times 1$ = $h \times 1$ transposed, times $h \times 2h$, times $2h \times 1$

$\alpha_t = \text{Softmax}(e_t)$ — $m \times 1$

$c_t = \sum_{i=1}^{m} \alpha_{t,i} \, h_i^{enc}$ — $2h \times 1$

mixed LSTM.

Decoder

Blockdiagram_rnn_attention-3.jpg

Encoder

W_h $\mathbb{R}^{h \times 2h}$

projection

$h^{enc} = [\overrightarrow{h}^{enc}; \overleftarrow{h}^{enc}]$ $\mathbb{R}^{2h \times 1}$ $\mathbb{R}^{h \times 1}$

W_c $C^{enc} = [\overrightarrow{C}^{enc} \overleftarrow{C}^{enc}]$ $\mathbb{R}^{2h \times 1}$ $\mathbb{R}^{h \times 1}$

$2h$ [1]

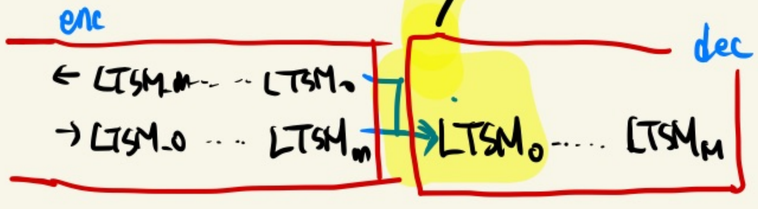
$h_0^{dec} = W_h[\overleftarrow{h}_1^{enc}; \overrightarrow{h}_m^{enc}]$

$h \times 1$ $h \times 2h$ $2h \times 1$

↳ attension/encoder. word vector

$m \times M$

$C_0^{dec} = W_c[\overleftarrow{C}_1^{enc}; \overrightarrow{C}_m^{enc}]$



Decoder

Blockdiagram_rnn_attention-4.jpg

Encoder

W_h $\mathbb{R}^{h \times 2h}$ · projection

$h^{enc} = [\overrightarrow{h}^{enc}; \overleftarrow{h}^{enc}]$ $\mathbb{R}^{2h \times 1}$ $\mathbb{R}^{h \times 1}$

W_c $\mathbb{R}^{2h \times 1}$ $c^{enc} = [\overrightarrow{c}^{enc}; \overleftarrow{c}^{enc}]$ $\mathbb{R}^{h \times 1}$

h^{enc} $\mathbb{R}^{2h \times 1}$

Words → word (ex 1) → word-embedding (e ∈ $\mathbb{R}^h$) — Module forward

batch, epoch?

Decoder_0

$h_0^{dec} = W_h[\overleftarrow{h}_1^{enc}; \overrightarrow{h}_m^{enc}]$ — $h \times 1$ = $h \times 2h$ · $h \times 1$

$c_0^{dec} = W_c[\overleftarrow{c}_1^{enc}; \overrightarrow{c}_m^{enc}]$

attention

$e_{t,i} = (h_t^{dec})^T W_{attproj} h_i^{enc}$ — $m \times 1$ = $1 \times h$ · $h \times 2h$ · $2h \times 1$

$\alpha_t = \text{softmax}(e_t)$ — $m \times 1$

$a_t = \sum_{i=1}^{m} \alpha_{t,i} h_i^{enc}$ — $2h \times 1$

mixed LTSM

h^{dec} $h \times 1$

$h_t^{dec}, c_t^{dec} = \text{Decoder}(\bar{y}_t, h_{t-1}^{dec}, c_{t-1}^{dec})$ — $h \times 1$, $h \times 1$

a_t $2h \times 1$

$u_t = [a_t; h_t^{dec}]$ — $3h \times 1$

$v_t = W_u u_t$ — $h \times 1$ = $h \times 3h$ · $3h \times 1$

$o_t = \text{dropout}(\tanh(v_t))$ — $h \times 1$

decoder

model.forward

$P_t = \text{softmax}(W_{vocab} o_t)$ — $V_t \times 1$ = $V_t \times h$ · h

$J_t(\theta) = \text{CrossEntropy}(P_t, g_t)$

g_t: one-hot vector

Blockdiagram_rnn_attention-5.jpg

Encoder

$h \times 1 \quad h \times 1$

$$h_t^{dec}, c_t^{dec} = \text{Decoder}(\bar{y}_t, h_{t-1}^{dec}, c_{t-1}^{dec})$$

$3h \times 1$

$$u_t = [a_t ; h_t^{dec}]$$

$$v_t = W_u u_t$$

$h \times 1 \quad h \times 3h \quad 3h \times 1$

$$o_t = \text{dropout}(\tanh(v_t))$$

$h \times 1$

decoder

model.foward

$$P_t = \text{softmax}(W_{vocab} o_t) \rightarrow J_t(\theta) = \text{CrossEntropy}(P_t, g_t)$$

$V_t \times 1 \quad V_t \times h$

g_t: one-hot vector