- Notation and Problem Statement

$\mathbb{X} = (\mathbb{x}^1, \mathbb{x}^2, \cdots, \mathbb{x}^n)^T$ n driving series : n features

$= (\mathbb{x}_1, \mathbb{x}_2, \cdots, \mathbb{x}_T) \in \mathbb{R}^{n \times T}$

where $\mathbb{x}^k = (x_1^k, x_2^k, \cdots, x_T^k)^T$

$$\begin{matrix} x_1^1 & x_1^2 & \cdots & x_1^n \\ x_2^1 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \ddots & \vdots \\ x_T^1 & x_T^2 & \cdots & x_T^n \end{matrix}$$

$\mathbb{x}_t = (x_t^1, x_t^2, \cdots, x_t^n)^T \in \mathbb{R}^n$

시간 t 에서 n개의 feature 에 대한 행렬

$(y_1, y_2, \cdots, y_{T-1})$ with $y_t \in \mathbb{R}$

$\hat{y}_T = \mathbb{F}(y_1, \cdots, y_{T-1}, \mathbb{x}_1, \cdots, \mathbb{x}_T)$

- Encoder with input attention

input sequence $\mathbb{X} = (\mathbb{x}_1, \mathbb{x}_2, \cdots, \mathbb{x}_T)$, $\mathbb{x}_t \in \mathbb{R}^n$

(eqn 2)

$\mathbb{h}_t = f_1(\mathbb{h}_{t-1}, \mathbb{x}_t)$, $\mathbb{h}_t \in \mathbb{R}^m$ m: size of hidden state

we use an LSTM unit as f_1 to capture long-term dependencies

LSTM unit

$\mathbb{f}_t$: forget gate, $\mathbb{i}_t$: input gate, $\mathbb{O}_t$: output gate

$$\mathbb{f}_t = \sigma(\mathbb{W}_f[\mathbb{h}_{t-1}; \mathbb{x}_t] + \mathbb{b}_f) \quad \text{eqn 3}$$
$$\mathbb{i}_t = \sigma(\mathbb{W}_i[\mathbb{h}_{t-1}; \mathbb{x}_t] + \mathbb{b}_i) \quad \text{eqn4}$$
$$\mathbb{O}_t = \sigma(\mathbb{W}_o[\mathbb{h}_{t-1}; \mathbb{x}_t] + \mathbb{b}_o) \quad \text{eqn5}$$

$$\mathbb{S}_t = \mathbb{f}_t \odot \mathbb{S}_{t-1} + \mathbb{i}_t \odot \tanh(\mathbb{W}_s[\mathbb{h}_{t-1}; \mathbb{x}_t] + \mathbb{b}_s)$$
$$\mathbb{h}_t = \mathbb{O}_t \odot \tanh(\mathbb{S}_t) \quad \text{eqn 7} \qquad \hookrightarrow \text{eqn6}$$

$[\mathbb{h}_{t-1}; \mathbb{x}_t] \in \mathbb{R}^{m+n}$: concatenation of $\mathbb{h}_{t-1}$ and $\mathbb{x}_t$

$\mathbb{W}_{f-s} \in \mathbb{R}^{m \times (m+n)}$ $\qquad$ $\mathbb{W}[\mathbb{h}_{t-1}; \mathbb{x}_t] \in \mathbb{R}^m$

$\mathbb{b}_{f-s} \in \mathbb{R}^m$

parameters to learn

Input attention

$$e_t^k = \mathbb{v}_e^\top \tanh(\mathbb{W}_e[\mathbb{h}_{t-1}; \mathbb{s}_{t-1}] + \mathbb{U}_e \mathbb{x}^k]) \qquad \text{eqn 8}$$

$$\alpha_t^k = \frac{\exp(e_t^k)}{\sum_{i=1}^{n} \exp(e_t^i)} \qquad \text{eqn 9}$$

: attention weight (softmax function applied)

where $\mathbb{v}_e \in \mathbb{R}^T$, $\mathbb{W}_e \in \mathbb{R}^{T \times 2m}$ and $\mathbb{U}_e \in \mathbb{R}$

$\mathbb{x}^k \in \mathbb{R}$

$$\tilde{\mathbb{x}}_t = (\alpha_t^1 x_t^1, \alpha_t^2 x_t^2, \cdots, \alpha_t^n x_t^n)^\top \qquad \text{eqn 10}$$

$$\mathbb{h}_t = f_1(\mathbb{h}_{t-1}, \tilde{\mathbb{x}}_t) \qquad \text{eqn 11}$$

↳ hidden state at time t

□

- Decoder with temporal attention

Temporal attention

$l_t^i = v_d^\top \tanh(W_d [d_{t-1} ; s'_{t-1}] + U_d h_i) \quad 1 \le i \le T$ eqn 12

$d_{t-1} \in \mathbb{R}^p$: previous decoder hidden state p?

$s'_{t-1} \in \mathbb{R}^p$: cell state of the LSTM unit

$[d_{t-1} ; s'_{t-1}] \in \mathbb{R}^{2p}$

$v_d \in \mathbb{R}^m$, $W_d \in \mathbb{R}^{m \times 2p}$, and $U_d \in \mathbb{R}^{m \times n}$: parameters to learn

$\beta_t^i = \dfrac{\exp(l_t^i)}{\sum_{j=1}^{T} \exp(l_t^j)}$: attention weight eqn 13

$c_t = \sum_{i=1}^{T} \beta_t^i h_i$: context vecotor eqn 14

$\tilde{y}_{t-1} = \tilde{w}^\top [y_{t-1} ; c_{t-1}] + \tilde{b}$ eqn 15

where $[y_{t-1} ; c_{t-1}] \in \mathbb{R}^{m+1}$

$d_t = f_2(d_{t-1}, \tilde{y}_{t-1})$ eqn 16

LSTM unit

$$f'_t = \sigma(\mathbb{W}'_f [\mathbb{d}_{t-1} ; \tilde{y}_{t-1}] + \mathbb{b}'_f) \quad \text{eqn 17}$$

$$i'_t = \sigma(\mathbb{W}'_i [\mathbb{d}_{t-1} ; \tilde{y}_{t-1}] + \mathbb{b}'_i) \quad \text{eqn 18}$$

$$o'_t = \sigma(\mathbb{W}'_o [\mathbb{d}_{t-1} ; \tilde{y}_{t-1}] + \mathbb{b}'_o) \quad \text{eqn 19}$$

$$s'_t = f'_t \odot s'_{t-1} + i'_t \odot \tanh(\mathbb{W}'_s [\mathbb{d}_{t-1} ; \tilde{y}_{t-1}] + \mathbb{b}'_s)$$

$$\mathbb{d}_t = o'_t \odot \tanh(s'_t) \quad \text{eqn 21} \qquad \hookrightarrow \text{eqn 20}$$

$[\mathbb{d}_{t-1} ; \tilde{y}_{t-1}] \in \mathbb{R}^{p+1}$

$\mathbb{W}'_{f-s} \in \mathbb{R}^{p \times (p+1)}$ and $\mathbb{b}'_{f-s} \in \mathbb{R}^{p}$: parameters to learn

$$\hat{y}_T = F(y_1, \cdots ; y_{T-1} ; x_1, \cdots, x_T)$$

$$= \mathbb{v}^{\top} (\mathbb{W}_y [\mathbb{d}_T ; c_T] + \mathbb{b}_w) + \mathbb{b}_v$$

$$O(y_T, \hat{y}_T) = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}^i_T - y^i_T)^2$$

$\square$

- 논문 수식 → 코드

$$e_t^k = \mathbf{v}_e^{\top} \tanh(\mathbf{W}_e[\mathbf{h}_{t-1}; \mathbf{s}_{t-1}] + \mathbf{U}_e \mathbf{x}^k]) \qquad \text{eqn 8}$$

$$\alpha_t^k = \frac{\exp(e_t^k)}{\sum_{i=1}^{n} \exp(e_t^i)} \qquad \text{eqn 9}$$

: attention weight (softmax function applied)

$\mathbf{h}_{t-1}$	(batch size, m) at time t-1
$\mathbf{s}_{t-1}$	(batch size, m) at time t-1
$[\mathbf{h}_{t-1}; \mathbf{s}_{t-1}]$	(batch size, 2m)
K.repeat / tf.concat	(batch size, n, 2m)
$\mathbf{W}_e[\mathbf{h}_{t-1}; \mathbf{s}_{t-1}]$ ↳ Dense(T)	(batch size, n, T)
$\mathbf{x}^k$	(batch size, T, n) permute
$\mathbf{U}_e \mathbf{x}^k$ ↳ Dense(T)	(batch size, n, T)

① add($\mathbf{W}_e[\mathbf{h}_{t-1}; \mathbf{s}_{t-1}]$, $\mathbf{U}_e \mathbf{x}^k$) (batch size, n, T)

② concat ($\mathbf{W}_e[\mathbf{h}_{t-1}; \mathbf{s}_{t-1}]$, $\mathbf{U}_e \mathbf{x}^k$) (batch size, n, 2T)

$$\tanh(W_e[h_{t-1}; s_{t-1}] + U_e x^k])$$

(batch size, n, $\boxed{T \text{ or } 2T}$)

Dense(1)

$$\underline{v_e^{\top}} \tanh(W_e[h_{t-1}; s_{t-1}] + U_e x^k])$$

↳ Dense(1)

(batch size, n, $\boxed{1}$)

permute

$$\alpha_t^k = \frac{\exp(e_t^k)}{\sum_{i=1}^{n} \exp(e_t^i)}$$

(batch size, 1, n)

$$\tilde{\mathbb{x}}_t = (\alpha_t^1 x_t^1,\ \alpha_t^2 x_t^2,\ \cdots,\ \alpha_t^n x_t^n)^t \qquad \text{eqn } 10$$

$$\mathbb{h}_t = f_1(\mathbb{h}_{t-1},\ \tilde{\mathbb{x}}_t) \qquad \text{eqn } 11$$