Conditional language model

- $p(Y)$, where $Y = (y_1, y_2 \cdots, y_T)$
- ...
- $p(Y|X)$ where $X = (x_1, x_2, \cdots, x_{T'})$

Machine translation

- Learning

$$\max_{\theta} \frac{1}{N} \sum_{n=1}^{N} \log p(Y^n | X^n; \theta)$$

- Inference / generation

$$\arg\max_{Y} \log p(Y|X)$$

(Y: ... ; X: given)

Parametrization

Autoregressive modeling

$$p(Y|X) = \prod_{t=1}^{T_Y} p(y_t | y_{<t}, X)$$

$p(y_t | y_{<t}, X)$

- Input: $(y_1, \cdots, y_{t-1})$ + $(x_1, x_2, \cdots, x_{T_X})$ ← concat.
- Output: $y_t \in V_Y = V$

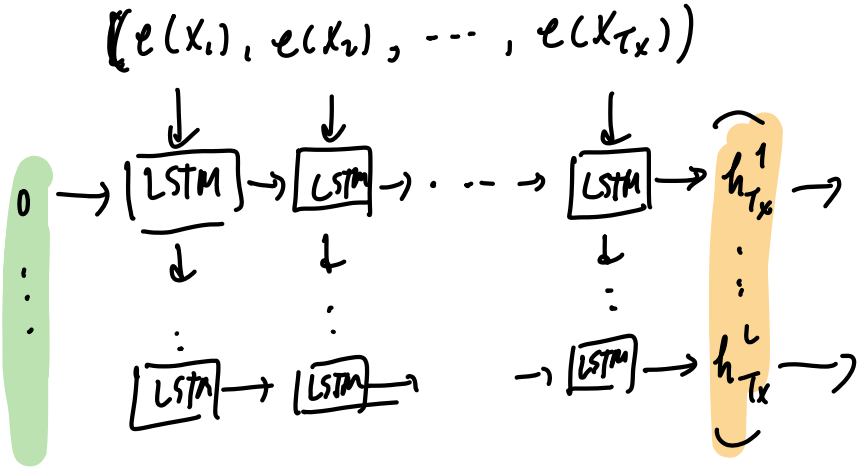
$$F(y_1, \cdots, y_{t-1}, x_1, \cdots, x_{T_X}, y_t) = b_{y_t} + e(y_t)^T G(y_1, \cdots, y_{t-1}, x_1, \cdots, x_{T_X})$$

$e(y_t) \in \mathbb{R}^d$

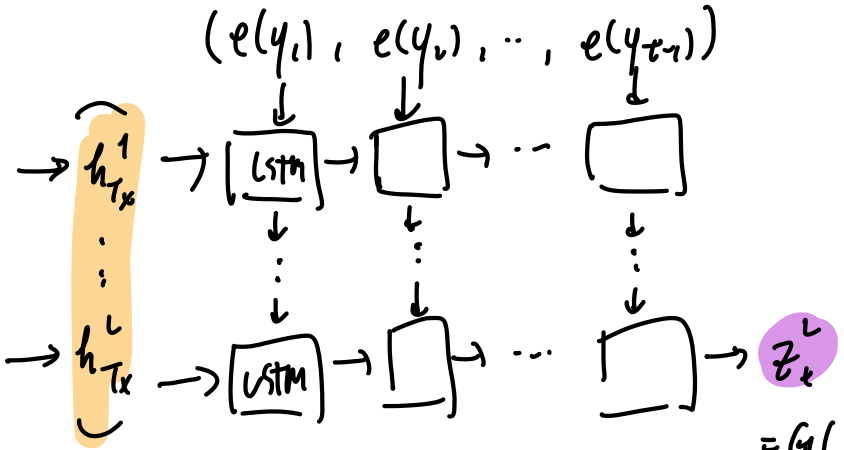
Sutskever et al. [2014] encoder-decoder

Encoder : LSTM network

- $e(x_t) \in \mathbb{R}^d$ for all $t = 1, \dots, T_x$



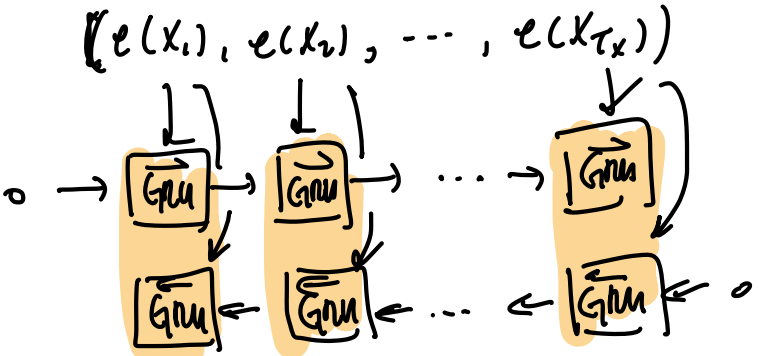
Decoder : LSTM network



$= G(x_1, \dots, x_{T_x}, y_1, \dots, y_{t-1})$

Bahdanau et al. [2014].

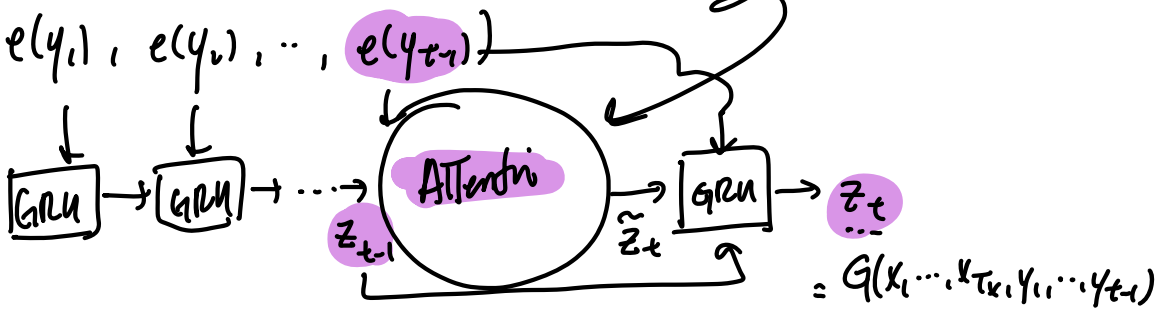
Encoder : bidirectional GRU / LSTM network.



$(h_1, h_2, \cdots, h_{T_x})$

Decoder: unidirectional GRU (LSTM) network

$(e(y_1), e(y_2), \cdots, e(y_{t-1}))$



① mix $e(y_{t-1})$ & z_{t-1} into $Q = \text{mix}(e(y_{t-1}), z_{t-1})$

② $V_{t'} = K_{t'} = h_{t'}$, $t' = 1, \cdots, T_x$

③ $\alpha_{t'} \propto \underbrace{\exp\{Q^T K_{t'}\}}_{\text{softmax}}$

④ $\tilde{z}_t = \sum_{t'=1}^{T_x} \alpha_{t'} h_{t'}$

Generation

$$\arg\max_{y \in \mathcal{L}} \log p(y|x)$$

$$p(y|x) = \prod_{t=1}^{T_y} p(y_t | y_{<t}, x)$$

log

greedy decoding

$$\hat{y}_t = \arg\max_{y_t \in V} \log p(y_t | \hat{y}_{<t}, x)$$

beam search.

$$\{\hat{y}_{<t}^1 \cdots \hat{y}_{<t}^K\} = \underset{\substack{k=1,\cdots,K \\ i=1,\cdots,|V|}}{\text{arg top-}K} \log p(\hat{y}_{<t}^k | x) + \log p(y_t = i | \hat{y}_{<t}^k, x)$$