$$h_t = u_t \odot \underbrace{\sigma(U h_{t-1} + W x_t + b)}_{= \hat{h}_t} + (1 - u_t) \odot h_{t-1}$$

- $h_1 = u_1 \odot \hat{h}_1 + (1 - u_1) \odot h_0$
- $h_2 = u_2 \odot \hat{h}_2 + (1 - u_2) \odot u_1 \odot \hat{h}_1 + (1 - u_2) \odot (1 - u_1) \odot h_0$
- $h_3 = u_3 \odot \hat{h}_3 + (1 - u_3) \odot u_2 \odot \hat{h}_2 + (1 - u_3) \odot (1 - u_2) \odot u_1 \odot \hat{h}_1$
  $+ (1 - u_3) \odot (1 - u_2) \odot (1 - u_1) \odot h_0$

$\vdots$

- $$h_t = \sum_{t'=1}^{t} \underbrace{u_{t'} \odot \left(\prod_{t''=t'}^{t-1} (1 - u_{t''}) \right)}_{\Downarrow} \odot \underbrace{\hat{h}_{t'}}_{\Downarrow}$$
  $$= \sum_{t'=1}^{t} w_{t'}(x_{t'}, h_{t'-1}) \odot \underline{\hat{h}_{t'}(x_{t'}, h_{t'-1})}$$

- Let's break the recurrence / temporal dependency.

① what if each $\underline{\hat{h}_{t'}}$ was computed independently?

$$h_t = \sum_{t'=1}^{t} w_{t'}(x_{t'}, h_{t'-1}) \odot \hat{h}(x_{t'}, t')$$

positional embedding.

② what if each $w_{t'}$ was computed independently?

$$h_t = \sum_{t'=1}^{t} w(x_{t'}, x_t, t', t) \odot \hat{h}(x_{t'}, t').$$

query key

Implement:

- $$w(x_{t'}, x_t, t', t) = \frac{\exp\{ Q(x_t, t)^T K(x_{t'}, t') \}}{\sum_{t'=1}^{t} \exp\{ Q(x_t, t)^T K(x_{t'}, t') \}}$$

$\hat{h}(x_{t'}, t') = V(\ldots$

$\rightarrow \hat{h}(x_{t'}, t') = \underbrace{V}_{\text{value}}(x_{t'} + \underbrace{p(t')}_{\text{positional embedding}})$

. Multiple attention heads

$$\underset{\dots}{h_{t,m}} = \sum_{t'=1}^{t} \underset{\dots}{W_m}(x_{t'}, x_t) \odot \hat{h}_m(x_{t'})$$

$$\cancel{h_t = [h_{t,1}; h_{t,2}; \dots; h_{t,M}]^T}$$

. Nonlinear fusion

$$h_t = \underset{\text{nonlinear}}{f}(\underbrace{[h_{t,1}; \dots; h_{t,M}]^T}_{\tilde{h}_t})$$

$$f(\tilde{h}_t) = \underbrace{\max(0, U_f \max(0, W_f \tilde{h}_t + b_f) + c_f)}\ \underbrace{+ x_t}.$$

. Self-attention : non-"causal" attention

$$h_{t,m} = \sum_{t'=1}^{T} W_m(x_{t'}, x_t, t', t) \odot \hat{h}_m(x_{t'}, t')$$

Inpainting

$$\underset{y}{\text{argmax}} \log p(\underbrace{x_{pre}}_{\substack{\text{seq} \\ \text{prefix}}}, \underbrace{y}_{\substack{\text{seq} \\ \text{missing}}}, \underbrace{x_{post}}_{\substack{\text{seq} \\ \text{suffix}}}).$$

$$\Leftrightarrow \underset{y}{\arg\max} \underbrace{\log p(y \mid x_{pre}, x_{post})} + \cancel{\log p(x_{pre}, x_{post})}$$

$$(x_1, x_2, \dots, x_t, \underset{\dots}{y}, x_{t+|y|}, \dots, x_T)$$

$$(x_1, x_2, \dots, x_t, \underbrace{\langle mask\rangle, \langle mask\rangle, \dots, \langle mask\rangle}_{|y|}, x_{t+|y|}, \dots, x_T)$$

$$\Downarrow$$

$$Y$$

① masked-out indices : $m_1, m_2, \dots, m_{T_m}$
observed indices : $o_1, o_2, \dots, o_{T_o}$
$T_m + T_o = T$

② corrupt(X) <mask>

③ $p(x_{m_1}, \dots, x_{m_{T_m}} \mid corrupt(x))$

$$= \prod_{i=1}^{T_m} p(x_{m_i} \mid corrupt(x))$$

Objective function

$$J(\theta) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{M,O \sim \gamma(T_n)} \left[\sum_{i=1}^{T_m^n} \log p(x^n_{m_i} \mid x^n_{o_1} \dots x^n_{o_{T_o}} ; \theta) \right]$$

$\Longrightarrow$ BERT, ...

$\underbrace{\hspace{8em}}$ $\boxed{\text{masked}}$ language modeling

① $$p(X) = \prod_{t=1}^{T} p(\underbrace{x_t}_{masked} \mid \underbrace{x_1, \dots, x_{t-1}}_{observed}, \underbrace{x_{t+1} = \langle mask\rangle, \dots, x_T = \langle mask\rangle}_{masked})$$

② $$F(X) = \sum_{t=1}^{T} \log p(x_t \mid x_{1:t-1}, x_t = \langle mask\rangle, x_{t+1:T}).$$

$\Uparrow$ pseudo log likelihood

neg. energ.

$$p(x) = \frac{\exp\{F(x)\}}{\sum_{x' \in L} \exp\{F(x')\}}$$

③ Implicit way.