

21H-Q4 Attention April 20th /

初級Aの n-gram

問題
↓

↓ 文脈が理解できない

RNN (LSTM) / 系列化できない

↓

Seq2Seq (Encoder-Decoder) 翻訳の尺印?

↓

文章同士の 関係関係
長い入力 → 小さな出力で情報圧縮

Attention 付き Seq2Seq

↓

Transformer (RNN なく Attention)

2(H-Q4

2

III soft attention

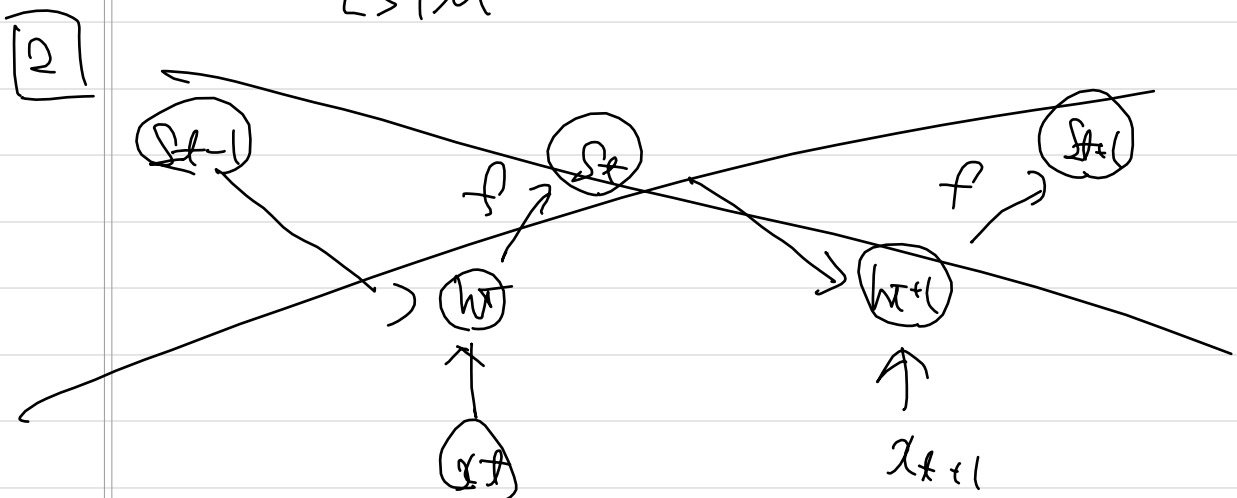
$$y = \sum_i \alpha_i x_i$$

where α is weight of importance.

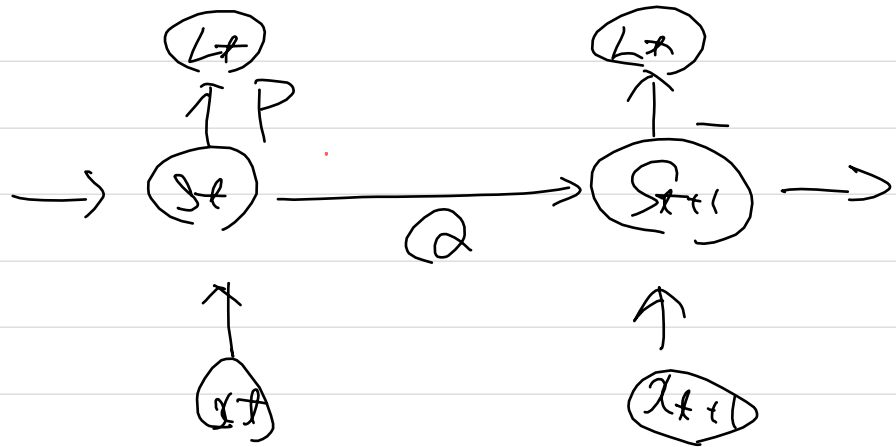
Backprop

$$\begin{aligned} \frac{\partial y}{\partial x_i} &= \frac{\partial}{\partial x_i} \left\{ \dots + \alpha_i x_i + \alpha_{i+1} x_{i+1} \dots \right\} \\ &= \underline{\alpha_i} \end{aligned}$$

LSTM



21H-Q4



let $\partial S_t L_t = P$

$$J = \frac{\partial L}{\partial S_t} = \frac{\partial L}{\partial S_{t+1}} \frac{\partial S_{t+1}}{\partial S_t} + \frac{\partial L}{\partial L_t} \frac{\partial L_t}{\partial S_t}$$

$$= \frac{\partial L}{\partial S_{t+1}} \left(\frac{\partial S_{t+1}}{\partial S_t} \right) + P$$

$$\frac{\partial L}{\partial S_t} = \frac{\partial L}{\partial S_{t+1}} (Q) + P$$

$$= \frac{\partial L}{\partial S_{t+2}} \left(\frac{\partial S_{t+2}}{\partial S_{t+1}} (Q) + P \right) + P$$

$$= \frac{\partial L}{\partial S_t} \left(Q^{T-t} + P(\dots) \right) + P$$

let $Q = U \Lambda U^T$ $\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix}$

when $t \rightarrow \infty$

$$\lambda_1 \geq \dots \geq \lambda_d$$

if $\lambda_1 > 1$ or $\lambda_d < -1$, gradient explode
if $|\lambda| < 1$ gradient vanishing

2/H-Q4

4

(3) How attention overcame
long term dependencies,

LSTM is RNN & f, g, i, o

f, g are long term dependencies

i, o are short term