

Simple & LSTM RNNs

Back Propagation Through Time (BPTT)

$$\textcircled{1} \quad \frac{\partial J(\theta)}{\partial U} = \sum_{t=1}^T \frac{\partial J^{(t)}(\theta)}{\partial U}$$

$$\textcircled{2} \quad \frac{\partial J^{(t)}(\theta)}{\partial W_h} = \sum_{i=1}^t \frac{\partial J^{(t)}(\theta)}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial h^{(i)}} \cdot \frac{\partial h^{(i)}}{\partial W_h}$$

$$\frac{\partial J(\theta)}{\partial W_h} = \sum_{t=1}^T \frac{\partial J^{(t)}(\theta)}{\partial W_h}$$

$J^{(t)}(\theta)$ 손실함수를 시작으로 \hat{y} 예측값, $h^{(t)}$, W_h 까지 연쇄적으로 연산과정

$$\textcircled{3} \quad \frac{\partial J^{(t)}(\theta)}{\partial W_e} = \sum_{i=1}^t \frac{\partial J^{(t)}(\theta)}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial h^{(i)}} \cdot \frac{\partial h^{(i)}}{\partial W_e}$$

$$\frac{\partial J(\theta)}{\partial W_e} = \sum_{t=1}^T \frac{\partial J^{(t)}(\theta)}{\partial W_e}$$

Vanishing / Exploding Gradients

- W_h 가 작은수(<1) 반복적으로 곱해지는 값이 0에 가까워져 gradient vanishing
- W_h 가 큼수(>1) 반복적으로 곱해지는 값이 기하급수적으로 커져 gradient exploding

Perplexity evaluation metric for LM

- Inverse probability of corpus, according to LM
- Equal to the exponential of the cross-entropy loss
- 낮을수록 좋음

LM: a system that predicts the next word

Gradient clipping: 만약 norm of the gradient가 threshold보다

크면 scale it down

같은 방향으로 step이나 작은 step으로 ∵ exploding gradient 해결

LSTM: Long Short-Term Memory RNN

- hidden state를 통해 short term memory 전달,

cell state를 통해 long term memory 보존

- forget, input, output 3개 gate를 통해 매 timestep의
cell state, hidden state, input에서 흐름 정보의 양 결정

① forget gate $f_t = \sigma(W_f h^{(t-1)} + U_f x^{(t)} + b_f)$

- 이전 timestep의 hidden state와 새로운 입력 seq에 시그모이드 취함

② input gate, new cell content

$$i_t = \sigma(W_i h^{(t-1)} + U_i x^{(t)} + b_i)$$

- input gate: 입력 정보에 시그모이드 취함

$$\tilde{C}_t = \tanh(W_c h^{(t-1)} + U_c x^{(t)} + b_c)$$

- new cell content: 입력 정보에 tanh 취함

③ Cell state

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

forget gate: 이전 timestep의 cell state에서 어느 정도의 정보 가져갈지 결정

input gate: 새로운 정보에서 장기 기억으로 가져갈 정보의 양 결정

④ output gate, hidden state

$$o_t = \sigma(W_o h^{(t-1)} + U_o x^{(t)} + b_o)$$

- output gate: 출력 정보에 시그모이드 취함

$$h_t = o_t * \tanh(C_t)$$

- hidden state: 현재 입력 대비 장기 기억에서 어느 정도의 단기 기억으로 사용할지 결정