Recitation 3 Feb 23

Agenda

- Recurrent Neural Network
- BP in RNN
- LSTM

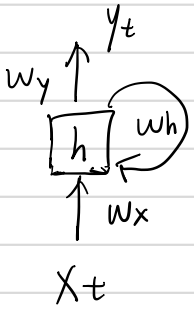
RNN

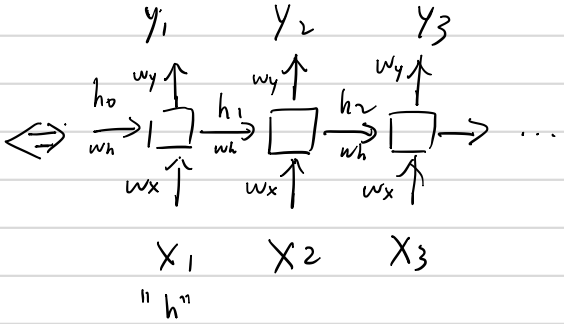
- General Structure

  x_t : input at time t

  y_t : output at time t

  h_t : hidden state of RNN at time t





$$h_t = f_1(x_t, h_t; W_h, W_x)$$

$$y_t = f_2(x_t, h_{t-1}; W_y)$$

- RNN for char-level language model (CLLM)

$$\max_{W_h, W_y} p(x_1, x_2, x_3 \cdots x_t) \quad \text{(likelyhood)}$$

p("hello") ↑

p("htktt") ↓

$$p(x_1, x_2, \cdots x_t) = \prod_{i=t}^{2} p(x_i \mid x_{i-1}, \cdots x_1)$$

⇒ • use h_{t-1} to capture history input $x_1, x_2 \cdots x_{t-1}$

- $y_t = p(x_t \mid h_{t-1}) \approx p(x_t \mid x_{t-1}, \cdots x_1)$
- Evaluate the loss by cross entropy $\sum_t E(y_t, \hat{y}_t)$

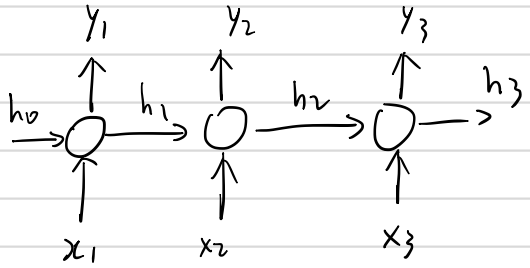
- Canonical RNN (for CLLM)

$f_1 = \tanh(\cdot; W_h, W_x) \Rightarrow h_t = \tanh(W_{xh} x_t + W_{hh} h_{t-1})$

$f_2 = \text{softmax}(\cdot; W_y, b_y) \Rightarrow y_t = \text{softmax}(W_{hy} h_t + b_y)$

embedding

BP in RNN



$$E = E_1 + E_2 + E_3 \quad ①$$

$$E_t = E(y_t, \hat{y}_t) \quad ②$$

$$y_t = \text{softmax}(W_y h_t + b_y) \quad ③$$

$$h_t = \tanh(W_h h_{t-1} + W_x x_t) \quad ④$$

- $\frac{\partial E}{\partial W_y}$

$$\frac{\partial E}{\partial W_y} \overset{①}{=} \sum_t \frac{\partial E_t}{\partial W_y}$$

②③

$\frac{\partial E_t}{\partial y_t}$ ✓ $\frac{\partial y_t}{\partial W_y}$ ✓

- $\frac{\partial E}{\partial W_x}$

$$\frac{\partial E}{\partial W_x} = \sum_t \frac{\partial E_t}{\partial W_x}$$

②③④

$\frac{\partial E_t}{\partial y_t}$ ✓ $\frac{\partial y_t}{\partial h_t}$ ✓ $\frac{\partial h_t}{\partial W_x}$?

$$h_3 = f(x_3, h_2, W_h, W_x)$$
$$h_2 = f(x_2, h_1, W_h, W_x)$$
$$h_1 = f(x_1, h_0, W_h, W_x)$$

$$\Rightarrow \frac{\partial h_3}{\partial W_x} = \frac{\partial h_3}{\partial W_x} + \frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial W_x} + \frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial W_x}$$

BP through time (BPTT)

can vanish or explode

- $\frac{\partial E}{\partial W_h}$ similar

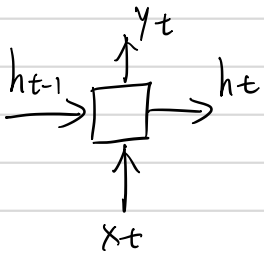
- Fight against Vanishing / Exploding gradient
  - Good activation (ReLU)
  - Clip the gradient
  - LSTM / GRU cell instead of tanh
  - TBPTT (Truncated BPTT)
- Embedding

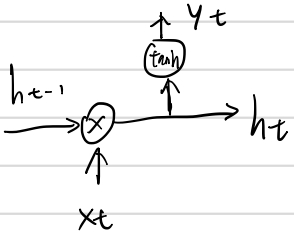
LSTM

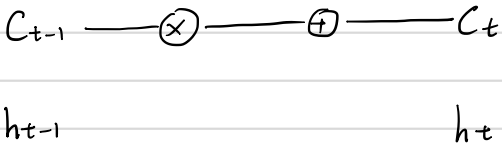
Motivation: Learn longer dependecies (RNN in theory can, but doesn't in practice)
Hochreiter 1991 / Bengio 1994

e.g 1: "Today is a good ___"

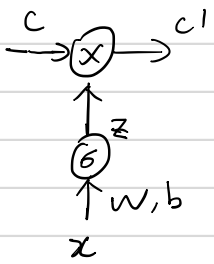
eg 2: "I live in France. ..... I speak ___"







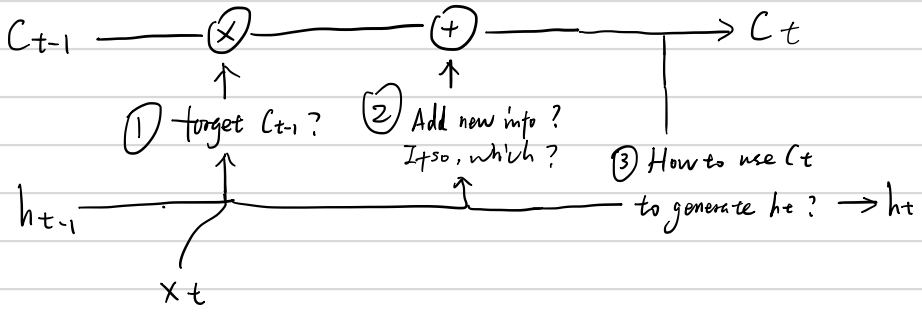
- Gates

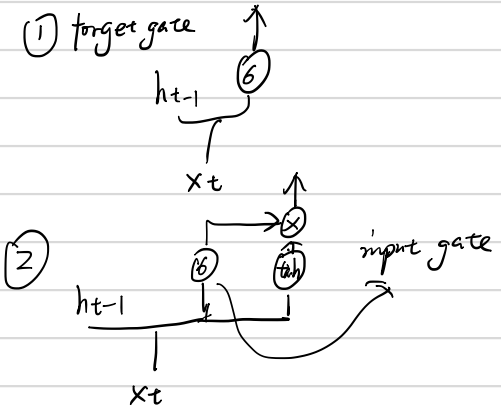


$z = \sigma(Wx+b) \in [0,1]^{m \times n}$

$c' = c \cdot z$ controls which part / how much of c flows through

- Use "Gates" to control RNN when to forget / remember





- Variants
  - GRU → ① combine forget and input gate ② merge ht and Ct ③ output gate
  - Many others.