Backpropagation Through Time in Detail

Mục 9.7.2 – Phân tích chi tiết

Nhóm trình bày

Ngày 21 tháng 12 năm 2025

Mục tiêu

- Trình bày Backpropagation Through Time (BPTT) trên RNN tuyến tính, tập trung vào công thức gradient.
- Làm rõ cách tính $\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{qh}}, \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{hx}}, \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{hh}}$ cùng các bias liên quan.
- Nhấn mạnh ý nghĩa của từng gradient trong cập nhật tham số mô hình.

Mô hình RNN tuyến tính

- Chuỗi đầu vào $\boldsymbol{x}_t \in \mathbb{R}^d$, hidden state $\boldsymbol{h}_t \in \mathbb{R}^h$, đầu ra $\boldsymbol{o}_t \in \mathbb{R}^q$.
- Tham số: $\boldsymbol{W}_{hx} \in \mathbb{R}^{h \times d}$, $\boldsymbol{W}_{hh} \in \mathbb{R}^{h \times h}$, $\boldsymbol{W}_{qh} \in \mathbb{R}^{q \times h}$.
- Bias: $\boldsymbol{b}_h \in \mathbb{R}^h$ cho hidden layer, $\boldsymbol{b}_q \in \mathbb{R}^q$ cho output layer.
- computational graph mở dọc thời gian, tất cả bước t dùng chung các trọng số và bias.

Phương trình forward và loss

$$\boldsymbol{h}_t = \boldsymbol{W}_{hx}\boldsymbol{x}_t + \boldsymbol{W}_{hh}\boldsymbol{h}_{t-1} + \boldsymbol{b}_h, \tag{1}$$

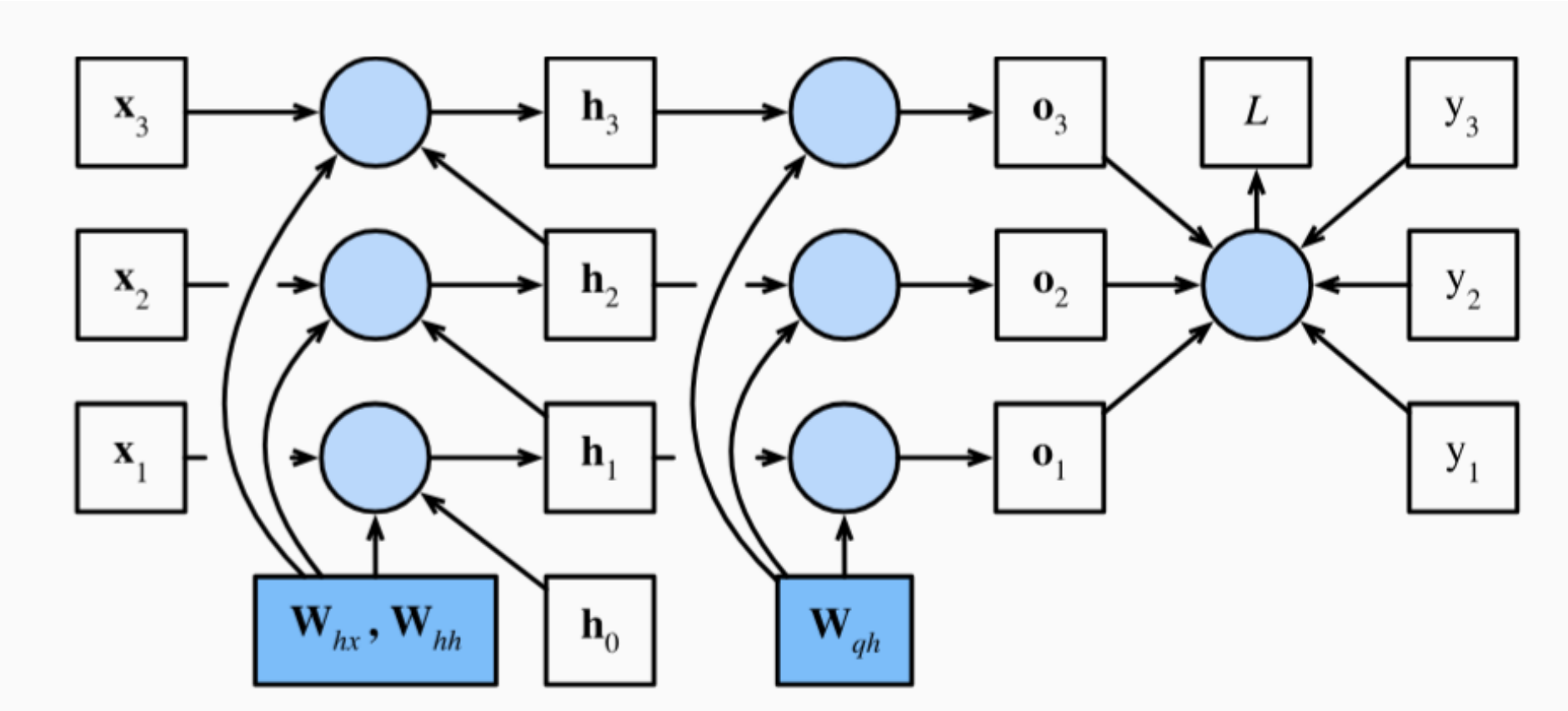
$$\boldsymbol{o}_t = \boldsymbol{W}_{qh}\boldsymbol{h}_t + \boldsymbol{b}_q \tag{2}$$

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T} \ell(\boldsymbol{o}_t, \boldsymbol{y}_t) \tag{3}$$

- Hidden state nhận thông tin từ input hiện tại và hidden state trước đó.
- Loss trung bình giúp gradient không phụ thuộc độ dài chuỗi.

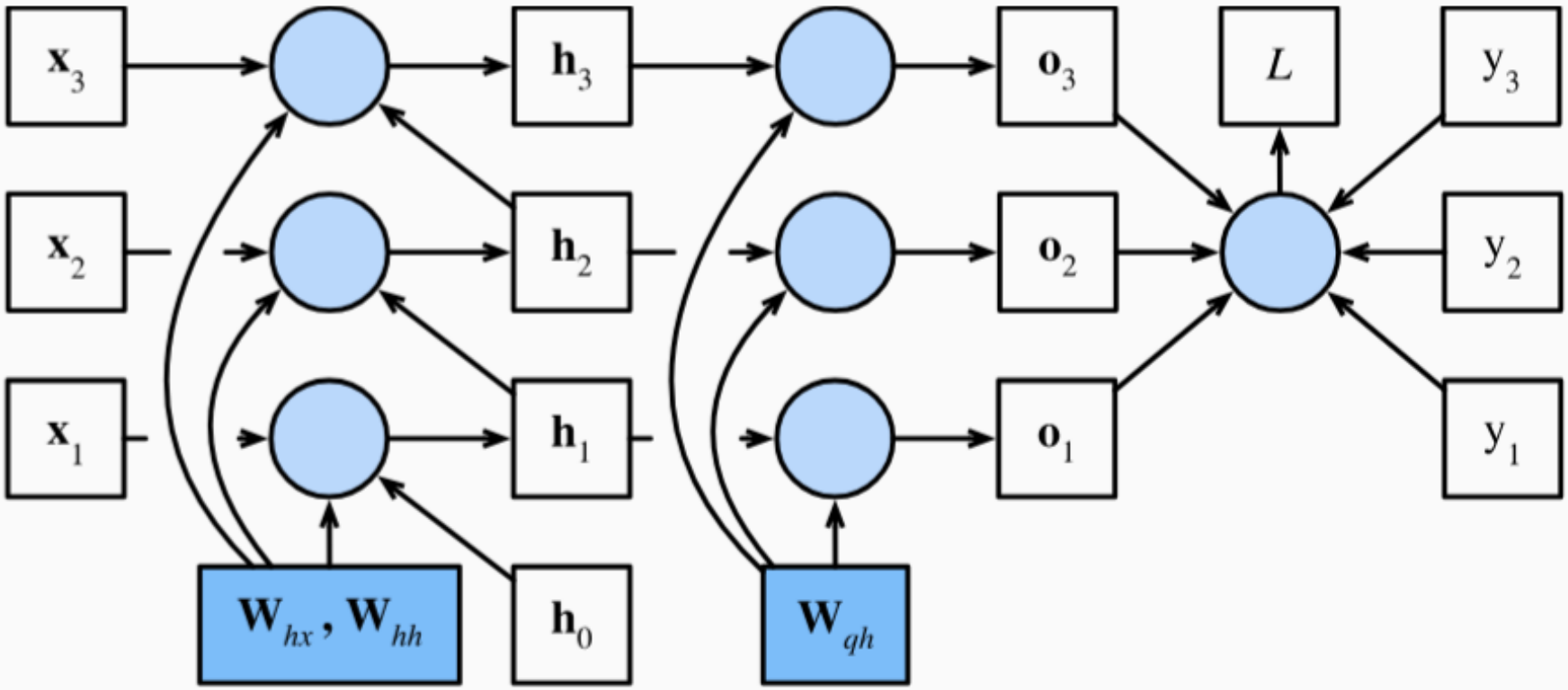
Đồ thị phụ thuộc

Để trực quan hóa phụ thuộc giữa biến và tham số trong quá trình tính toán RNN, ta vẽ một computational graph như Hình 9.7.2. Ví dụ, việc tính hidden state tại bước thời gian 3, $\boldsymbol{h}_3$, phụ thuộc vào các tham số $\boldsymbol{W}_{hx}$ và $\boldsymbol{W}_{hh}$, hidden state của bước trước $\boldsymbol{h}_2$, và input tại bước hiện tại $\boldsymbol{x}_3$.



Đồ thị phụ thuộc (tiếp)

Theo các phụ thuộc trong Hình 9.7.2, ta có thể đi ngược chiều mũi tên để lần lượt tính và lưu các gradient.



Gradient theo output tại time step t

$$L = \frac{1}{T} \sum_{k=1}^{T} \ell(o_k, y_k). \tag{1}$$

Vì o_t chỉ xuất hiện trong một hạng tử của tổng, ta có:

$$\begin{aligned} \frac{\partial L}{\partial o_t} &= \frac{\partial}{\partial o_t} \left(\frac{1}{T} \sum_{k=1}^{T} \ell(o_k, y_k) \right) \\ &= \frac{1}{T} \frac{\partial \ell(o_t, y_t)}{\partial o_t} \in \mathbb{R}^{q \times 1}. \end{aligned} \tag{2}$$

Gradient theo W_{qh} (1/2)

Output layer tuyến tính:

$$o_t = W_{qh} h_t + b_q. \tag{3}$$

Viết theo từng phần tử:

$$(o_t)_j = \sum_{i=1}^{h} (W_{qh})_{ji} (h_t)_i + (b_q)_j. \tag{4}$$

Áp dụng chain rule:

$$\begin{aligned} \frac{\partial L}{\partial (W_{qh})_{ji}} &= \sum_{t=1}^{T} \sum_{k=1}^{q} \frac{\partial L}{\partial (o_t)_k} \frac{\partial (o_t)_k}{\partial (W_{qh})_{ji}} \\ &= \sum_{t=1}^{T} \frac{\partial L}{\partial (o_t)_j} (h_t)_i. \end{aligned} \tag{5}$$

Gradient theo W_{qh} (2/2)

Viết dưới dạng ma trận:

$$\frac{\partial L}{\partial W_{qh}} = \sum_{t=1}^{T} \left(\frac{\partial L}{\partial o_t} \right) h_t^\top \in \mathbb{R}^{q \times h}. \tag{6}$$

- Gradient cộng dồn theo thời gian vì W_{qh} được chia sẻ cho mọi bước.

Gradient theo h_T

Tại thời điểm cuối:

$$o_T = W_{qh} h_T + b_q. \tag{7}$$

Áp dụng chain rule theo chỉ số:

$$\begin{aligned} \frac{\partial L}{\partial (h_T)_i} &= \sum_{j=1}^{q} \frac{\partial L}{\partial (o_T)_j} \frac{\partial (o_T)_j}{\partial (h_T)_i} \\ &= \sum_{j=1}^{q} (W_{qh})_{ji} \frac{\partial L}{\partial (o_T)_j}. \end{aligned} \tag{8}$$

Viết dạng vector:

$$\frac{\partial L}{\partial h_T} = W_{qh}^{\top} \frac{\partial L}{\partial o_T} \in \mathbb{R}^{h \times 1}. \tag{9}$$

Gradient theo h_t với $t < T$

Forward:

$$h_{t+1} = W_{hx}x_{t+1} + W_{hh}h_t + b_h, \qquad o_t = W_{qh}h_t + b_q. \tag{10}$$

Áp dụng chain rule:

$$\frac{\partial L}{\partial h_t} = \left(\frac{\partial h_{t+1}}{\partial h_t}\right)^\top \frac{\partial L}{\partial h_{t+1}} + \left(\frac{\partial o_t}{\partial h_t}\right)^\top \frac{\partial L}{\partial o_t}. \tag{11}$$

Vì các ánh xạ là tuyến tính:

$$\frac{\partial h_{t+1}}{\partial h_t} = W_{hh}, \qquad \frac{\partial o_t}{\partial h_t} = W_{qh}. \tag{12}$$

Suy ra:

$$\frac{\partial L}{\partial h_t} = W_{hh}^\top \frac{\partial L}{\partial h_{t+1}} + W_{qh}^\top \frac{\partial L}{\partial o_t}. \tag{13}$$

Gradient theo W_{hx} (1/2)

Hidden state tại thời điểm t:

$$h_t = W_{hx}x_t + W_{hh}h_{t-1} + b_h. \tag{14}$$

Vì L phụ thuộc vào W_{hx} thông qua toàn bộ các h_t, ta áp dụng chain rule:

$$\frac{\partial L}{\partial W_{hx}} = \sum_{t=1}^{T} \text{prod}\left(\frac{\partial L}{\partial h_t}, \frac{\partial h_t}{\partial W_{hx}}\right). \tag{15}$$

Xét từng phần tử của h_t:

$$(h_t)_i = \sum_{j=1}^{d}(W_{hx})_{ij}(x_t)_j + \sum_{k=1}^{h}(W_{hh})_{ik}(h_{t-1})_k + (b_h)_i. \tag{16}$$

Suy ra:

$$\frac{\partial(h_t)_i}{\partial(W_{hx})_{ij}} = (x_t)_j. \tag{17}$$

Gradient theo W_{hx} (2/2)

Do đó:

$$\frac{\partial L}{\partial (W_{hx})_{ij}} = \sum_{t=1}^{T} \frac{\partial L}{\partial (h_t)_i} \frac{\partial (h_t)_i}{\partial (W_{hx})_{ij}}$$
$$= \sum_{t=1}^{T} \frac{\partial L}{\partial (h_t)_i} (x_t)_j. \tag{18}$$

Viết dưới dạng ma trận:

$$\frac{\partial L}{\partial W_{hx}} = \sum_{t=1}^{T} (\ paramfracLh_t)\, x_t^\top \in \mathbb{R}^{h \times d}. \tag{19}$$

Gradient theo W_{hh} (1/2)

Từ phương trình hidden state:

$$h_t = W_{hx}x_t + W_{hh}h_{t-1} + b_h, \tag{20}$$

W_{hh} ảnh hưởng đến L thông qua toàn bộ chuỗi hidden states $h_1, \ldots, h_T$.
Áp dụng chain rule:

$$\frac{\partial L}{\partial W_{hh}} = \sum_{t=1}^{T} \text{prod}\left(\frac{\partial L}{\partial h_t}, \frac{\partial h_t}{\partial W_{hh}}\right). \tag{21}$$

Viết theo từng phần tử:

$$(h_t)_i = \sum_{k=1}^{h} (W_{hh})_{ik}(h_{t-1})_k + \cdots \tag{22}$$

Suy ra:

$$\frac{\partial (h_t)_i}{\partial (W_{hh})_{ik}} = (h_{t-1})_k. \tag{23}$$

Gradient theo W_{hh} (2/2)

Do đó:

$$\frac{\partial L}{\partial (W_{hh})_{ik}} = \sum_{t=1}^{T} \frac{\partial L}{\partial (h_t)_i} \frac{\partial (h_t)_i}{\partial (W_{hh})_{ik}}$$
$$= \sum_{t=1}^{T} \frac{\partial L}{\partial (h_t)_i} (h_{t-1})_k. \tag{24}$$

Viết dưới dạng ma trận:

$$\frac{\partial L}{\partial W_{hh}} = \sum_{t=1}^{T} \left(\frac{\partial L}{\partial h_t} \right) h_{t-1}^{\top} \in \mathbb{R}^{h \times h}. \tag{25}$$

Gradient theo bias b_h

Hidden state:

$$h_t = W_{hx}x_t + W_{hh}h_{t-1} + b_h. \tag{26}$$

Xét đạo hàm theo b_h:

$$\frac{\partial h_t}{\partial b_h} = I_h. \tag{27}$$

Áp dụng chain rule:

$$\begin{aligned} \frac{\partial L}{\partial b_h} &= \sum_{t=1}^{T} \frac{\partial L}{\partial h_t} \frac{\partial h_t}{\partial b_h} \\ &= \sum_{t=1}^{T} \frac{\partial L}{\partial h_t} \in \mathbb{R}^{h\times 1}. \end{aligned} \tag{28}$$

Gradient theo b_q

Từ biểu thức:

$$o_t = W_{qh} h_t + b_q, \tag{29}$$

ta có:

$$\frac{\partial o_t}{\partial b_q} = I_q. \tag{30}$$

Áp dụng chain rule:

$$\begin{aligned} \frac{\partial L}{\partial b_q} &= \sum_{t=1}^{T} \frac{\partial L}{\partial o_t} \frac{\partial o_t}{\partial b_q} \\ &= \sum_{t=1}^{T} \frac{\partial L}{\partial o_t} \in \mathbb{R}^{q \times 1}. \end{aligned} \tag{31}$$

Kết luận: Gradient theo các tham số của mô hình

Gradient theo output layer:

$$\frac{\partial L}{\partial W_{qh}} = \sum_{t=1}^{T} \left(\frac{\partial L}{\partial o_t} \right) h_t^\top,$$

$$\frac{\partial L}{\partial b_q} = \sum_{t=1}^{T} \frac{\partial L}{\partial o_t}. \tag{32}$$

Gradient theo hidden layer:

$$\frac{\partial L}{\partial W_{hx}} = \sum_{t=1}^{T} \left(\frac{\partial L}{\partial h_t} \right) x_t^\top,$$

$$\frac{\partial L}{\partial W_{hh}} = \sum_{t=1}^{T} \left(\frac{\partial L}{\partial h_t} \right) h_{t-1}^\top,$$

$$\frac{\partial L}{\partial b_h} = \sum_{t=1}^{T} \frac{\partial L}{\partial h_t}. \tag{33}$$