Exploding And Vanishing Gradient Problems

Mục 9.7.2 – Phân tích chi tiết

Nhóm trình bày

Ngày 21 tháng 12 năm 2025

Nguồn gốc của vanishing và exploding gradients

Từ chương trước, gradient theo hidden state tại thời điểm t có dạng:

$$\frac{\partial L}{\partial h_t} = \sum_{k=t}^{T} (W_{hh}^{\top})^{k-t} W_{qh}^{\top} \frac{\partial L}{\partial o_k}. \tag{1}$$

Do đó, độ lớn của gradient phụ thuộc trực tiếp vào các lũy thừa của ma trận $W_{hh}^{\top}$.

Vấn đề cốt lõi:

- Gradient bị nhân lặp lại bởi cùng một ma trận qua nhiều bước thời gian.
- Hành vi của gradient được quyết định bởi tính chất phổ của W_{hh}.

Ước lượng độ lớn gradient

Lấy chuẩn hai vế:

$$\left\| \frac{\partial L}{\partial h_t} \right\| \leq \sum_{k=t}^{T} \left\| (W_{hh}^{\top})^{k-t} \right\| \left\| W_{qh}^{\top} \right\| \left\| \frac{\partial L}{\partial o_k} \right\|. \tag{2}$$

Với chuẩn ma trận dưới chuẩn vector:

$$\left\| (W_{hh}^{\top})^{k-t} \right\| \leq \| W_{hh} \|^{k-t}. \tag{3}$$

Do đó:

$$\left\| \frac{\partial L}{\partial h_t} \right\| \lesssim \sum_{k=t}^{T} \| W_{hh} \|^{k-t}. \tag{4}$$

Điều kiện vanishing và exploding gradients

Xét giới hạn khi độ dài chuỗi tăng ($T \to \infty$):

- Nếu $\|W_{hh}\| < 1$:

$$\|W_{hh}\|^{k-t} \to 0 \quad \Rightarrow \quad \text{gradient vanishing.}$$

- Nếu $\|W_{hh}\| > 1$:

$$\|W_{hh}\|^{k-t} \to \infty \quad \Rightarrow \quad \text{gradient exploding.}$$

- Nếu $\|W_{hh}\| \approx 1$:

gradient được duy trì ổn định.

Trong thực hành, điều kiện này tương đương với bán kính phổ (spectral radius) của W_{hh}.

Tóm tắt vấn đề

Vanishing và exploding gradients là hệ quả trực tiếp của việc:

- Nhân lặp cùng một ma trận W_{hh} qua nhiều bước thời gian.
- Gradient là tổng của các chuỗi lũy thừa ma trận.

Do đó, việc huấn luyện RNN chuỗi dài gặp khó khăn ngay cả trong trường hợp tuyến tính đơn giản.

Giải pháp 1: Gradient Clipping

Gradient clipping giới hạn độ lớn gradient trong quá trình cập nhật.

Cụ thể:

$$g \leftarrow \frac{g}{\max\left(1, \frac{\|g\|}{\tau}\right)}, \tag{5}$$

trong đó g là gradient và τ là ngưỡng.

Đặc điểm:

- Không loại bỏ nguyên nhân gốc rễ.
- Ngăn gradient exploding trong thực tế.
- Dễ cài đặt, được dùng phổ biến.

Giải pháp 2: Long Short-Term Memory (LSTM)

LSTM thay đổi kiến trúc RNN bằng cách:

- Tạo đường truyền gradient gần như tuyến tính theo thời gian.
- Tránh nhân lặp trực tiếp bởi cùng một ma trận.

Trạng thái cell c_t được cập nhật dưới dạng:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \tag{6}$$

trong đó f_t là forget gate.

Gradient có thể đi xuyên qua nhiều bước thời gian mà không bị suy giảm hoặc bùng nổ nghiêm trọng.