UIT - CS431.O12.KHCL
Nguyễn Hoàng Tân. (21521413)

20/11/2023 Task

Luong Attention và Bahdanau Attention là hai mechanisms khác nhau được sử dụng trong context của attention mechanisms trong các mô hình machine translation (NMT). Các attention mechanisms này được thiết kế để giúp mô hình tập trung vào các phần khác nhau của chuỗi đầu vào khi tạo ra từng phần của chuỗi đầu ra. Hãy tìm hiểu về những khác biệt chính giữa Luong Attention và Bahdanau Attention:

# 1 Calculation of Attention Scores

- Bahdanau Attention: Được giới thiệu bởi Dzmitry Bahdanau vào năm 2014, cơ chế attention này tính **attention score** bằng cách sử dụng một feedforward neural network. Các **attention score** được học trong quá trình huấn luyện.
- Luong Attention: Được đề xuất bởi Minh-Thang Luong vào năm 2015, cơ chế attention này tính **attention score** bằng cách sử dụng một tích vô hướng đơn giản hoặc một phép nhân giữa hidden state của decoder và hidden state của encoder.

# 2 Alignment Vector

- Bahdanau Attention: Tính toán một alignment vector dựa trên tổng có trọng số của các hidden state của encoder, trong đó các trọng số chính là các attention scores.
- Alignment vector được tính bằng cách lấy tổng trọng số của các hidden state của encoder, trong đó các trọng số là các attention scores.

# 3 Scoring Mechanism

- Bahdanau Attention: Sử dụng một mô hình liên kết có thể học được (một mạng nơ-ron nhỏ) để tính điểm độ quan trọng của mỗi hidden state của encoder đối với một bước giải mã cụ thể.
- Luong Attention: Sử dụng một cơ chế tính điểm đơn giản hơn, chẳng hạn như tích vô hướng hoặc tương tác nhân tích, mà không cần mạng nơ-ron bổ sung.

# 4 Coverage Mechanism

- Tích hợp một coverage mechanism, giúp theo dõi những phần của chuỗi đầu vào đã được attended.
- Ban đầu không bao gồm coverage mechanism, nhưng các biến thể sau như "Global Attention" đã được giới thiệu để giải quyết hạn chế này.

# 5 Model Complexity

- Bahdanau Attention: Thường được coi là phức tạp về mặt tính toán do có mạng nơ-ron bổ sung được sử dụng để tính điểm.
- Luong Attention: Đơn giản hơn về mặt tính toán vì sử dụng một cơ chế điểm trực tiếp mà không có mạng nơ-ron bổ sung.