XLNet

**Cách hoạt động:**

**1. Permutation Language Modeling (PLM)**

* **Mục tiêu:** XLNet dùng PLM thay vì Masked Language Model (MLM) như BERT. PLM cho phép mô hình dự đoán mỗi từ dựa trên tất cả các hoán vị có thể của các từ trong câu, giúp mô hình học được ngữ cảnh toàn diện hơn.
* **Lợi ích:** Điều này giúp XLNet hiểu được ngữ cảnh theo cả hai hướng (trước và sau từ cần dự đoán) và giảm thiểu vấn đề không nhất quán giữa giai đoạn tiền huấn luyện và tinh chỉnh.

**2. Two-Stream Self-Attention**

* **Cách thức hoạt động:** XLNet sử dụng cơ chế "two-stream self-attention", nơi có hai luồng xử lý thông tin: "content stream" cho các từ đã biết và "query stream" cho vị trí từ cần dự đoán.
* **Lợi ích:** Cơ chế này giúp XLNet dự đoán từ mà không cần phải che dấu chúng trong quá trình tiền huấn luyện, giúp xử lý thông tin một cách hiệu quả và chính xác hơn.

**3. Hiểu Ngữ Cảnh Song Phương và Cải Thiện Hiệu Suất**

* **Hiểu Ngữ Cảnh Song Phương:** Bằng cách sử dụng PLM và "two-stream self-attention", XLNet có khả năng hiểu được ngữ cảnh một cách toàn diện hơn, từ đó cải thiện đáng kể hiệu suất trên nhiều tác vụ xử lý ngôn ngữ tự nhiên.
* **Cải Thiện Hiệu Suất:** Những cải tiến này giúp XLNet vượt trội so với BERT và các mô hình tiền thân khác trong nhiều tác vụ xử lý ngôn ngữ tự nhiên, từ phân loại văn bản đến trả lời câu hỏi và tóm tắt văn bản.

**Các bước tiến hành huấn luyện và đánh giá:**

**1. Đọc và Tiền Xử Lý Dữ Liệu:**

- Đầu tiên, dữ liệu được đọc từ các file JSON. Mỗi file chứa các đoạn văn (context) và câu hỏi liên quan, cùng với vị trí của câu trả lời trong đoạn văn.

- Tiền xử lý bao gồm việc sử dụng tokenizer của XLNet để chuyển đoạn văn và câu hỏi thành dạng token mà mô hình có thể hiểu được. Tokenizer cũng tạo ra "attention masks", giúp mô hình biết được phần nào của dữ liệu cần tập trung vào.

**2. Chuẩn Bị Mô Hình và Cài Đặt Huấn Luyện:**

- Mô hình XLNet được tải và chuẩn bị sẵn sàng cho việc huấn luyện. Bên cạnh đó, một bộ tối ưu hóa (optimizer) và lịch trình điều chỉnh tốc độ học (learning rate scheduler) được thiết lập.

**3. Huấn Luyện Mô Hình:**

- Trong quá trình huấn luyện, dữ liệu được đưa vào mô hình theo từng batch. Mô hình sẽ dự đoán vị trí bắt đầu và kết thúc của câu trả lời dựa trên context và câu hỏi. Sau mỗi lần dự đoán, loss (sự khác biệt giữa dự đoán và thực tế) được tính toán và sử dụng để cập nhật trọng số của mô hình, nhằm mục tiêu giảm thiểu loss.

**4. Đánh Giá Mô Hình:**

- Sau khi huấn luyện, mô hình được đánh giá trên tập test để xem nó hoạt động tốt như thế nào với dữ liệu chưa từng thấy. Quá trình này không cập nhật trọng số mô hình; thay vào đó, nó đo lường độ chính xác của mô hình thông qua loss và tỉ lệ dự đoán đúng.

- Kết quả đánh giá bao gồm độ chính xác (tỉ lệ câu trả lời được mô hình dự đoán chính xác) và loss trung bình trên tập test, cung cấp cái nhìn tổng quan về hiệu suất mô hình.