

Kiến Trúc Transformer – Cổ Máy Xử Lý Ngôn Ngữ Đỉnh Cao

Trong bài này, chúng ta sẽ khám phá thế giới của **Transformer** – không phải những robot biến hình trong phim, mà là một kiến trúc AI cách mạng hóa xử lý ngôn ngữ tự nhiên (NLP). Hãy cùng tìm hiểu một cách đơn giản và thú vị nhé!

1. Giới thiệu: Transformer - "Attention is All You Need"

Năm 2017, nhóm nghiên cứu **Google Brain** công bố bài báo kinh điển "[Attention is All You Need](#)", giới thiệu kiến trúc **Transformer**. Điểm đột phá nằm ở cơ chế "**tập trung**" (**attention**), giúp mô hình xử lý dữ liệu tuần tự (như câu văn) hiệu quả hơn hẳn các mô hình cũ (RNN, LSTM).

2. Kiến Trúc Transformer: Encoder & Decoder

2.1. Giai đoạn Encoder: "Mã hóa" thông tin đầu vào

- **Input Embedding:** Biến mỗi từ (token) thành **vector số học**. Ví dụ câu "*Transformers are awesome*" sẽ được chuyển thành các vector tương ứng.
- **Positional Encoding:** Thêm thông tin **vị trí từ** trong câu (vì Transformer không xử lý tuần tự như RNN).
- **Multi-Head Attention:**
 - Cơ chế "đa đầu" giúp mô hình **tập trung vào nhiều phần khác nhau** của câu cùng lúc. Ví dụ:
 - "*The cat sat on the mat*" → Một "head" tập trung vào quan hệ "*cat - mat*", head khác phân tích "*sat - on*".
 - Giống như bạn vừa nghe podcast, vừa đọc phụ đề, lại vừa ghi chú keywords!
- **Add & Norm:** Kết hợp thông tin cũ/mới (residual connection) và chuẩn hóa dữ liệu để ổn định quá trình học.
- **Feedforward Network:** Tinh chỉnh thông tin qua các phép biến đổi tuyến tính và phi tuyến (ReLU).

2.2. Giai đoạn Decoder: "Giải mã" để tạo kết quả

- **Output Embedding + Positional Encoding:** Tương tự encoder nhưng áp dụng cho chuỗi đầu ra (ví dụ câu dịch từ Anh sang Pháp).
- **Masked Multi-Head Attention:**
 - Khác biệt lớn nhất! Decoder bị "**che**" (**mask**) để không nhìn trước các từ tương lai, đảm bảo khi dịch/dự đoán từ thứ N, nó chỉ dựa vào từ 1 → N-1.
 - Ví dụ: Dịch "*I love AI*" sang tiếng Việt, khi sinh ra từ "*yêu*", mô hình chỉ biết "*Tôi*", không biết trước "*AI*".
- **Multi-Head Attention kết hợp Encoder-Decoder:**
 - Decoder "hỏi" encoder: "*Phần nào của câu gốc liên quan đến từ tôi đang dịch?*" → Cơ chế này giúp dịch chính xác ngữ cảnh.
- **Linear + Softmax:** Biến đổi thành xác suất để chọn từ tiếp theo (ví dụ: sau "*Tôi*" là "*yêu*" với xác suất 80%, "*thích*" 15%...).

3. Tại sao Transformer "xịn"?

- **Ưu điểm vượt trội:**
 - **Song song hóa:** Xử lý cả câu cùng lúc (khác RNN phải tuần tự), tốc độ nhanh hơn.
 - **Hiểu ngữ cảnh sâu:** Nhờ cơ chế attention, nó nắm bắt được quan hệ giữa các từ dù cách xa nhau (ví dụ: "*The cat, which was hungry, sat on the mat*" → hiểu "*cat*" liên quan "*sat*").
 - **Linh hoạt:** Ứng dụng được cho dịch máy (Google Translate), sinh văn bản (ChatGPT), tổng hợp văn bản...
- **Ứng dụng thực tế:**
 - **ChatGPT, Gemini, Claude:** Đều dựa trên biến thể của Transformer.
 - **BERT (Google):** Transformer chỉ dùng encoder, tối ưu cho phân tích ngữ nghĩa.

Kết luận

Transformer là "**trái tim**" của các mô hình ngôn ngữ hiện đại. Hiểu được kiến trúc này, bạn sẽ thấy rõ tại sao AI có thể dịch thuật, trò chuyện, hay viết văn mượt mà đến thế!

“ **Fun fact:** Giọng đọc trong video có thể không chuẩn (như tác giả tự nhận ☹️), nhưng kiến thức thì cực chất!

Tác giả: **Đỗ Ngọc Tú**
Công Ty Phần Mềm **VHTSoft**

Phiên bản #2
Được tạo 23 tháng 4 2025 14:05:25 bởi Đỗ Ngọc Tú
Được cập nhật 26 tháng 4 2025 10:16:37 bởi Đỗ Ngọc Tú