T5

1. T5 là gì?

T5 (Text-to-Text Transfer Transformer) là mô hình Transformer do nhóm của Colin Raffel (Google) giới thiệu trong paper: *"Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer"* (2020).

T5 đại diện cho một bước tiến vượt bậc trong các mô hình Xử lý ngôn ngữ tự nhiên (NLP), khi nó giới thiệu một **khuôn khổ "text-to-text" (văn bản – sang – văn bản)** hoàn toàn mới.
Khác với các mô hình truyền thống được thiết kế riêng cho từng nhiệm vụ cụ thể, **T5 coi tất cả các tác vụ NLP đều là quá trình chuyển đổi văn bản sang văn bản**.

Trong khuôn khổ này, **cả đầu vào và đầu ra của mọi tác vụ NLP đều được coi là chuỗi văn bản**, giúp mô hình trở nên **linh hoạt và đa năng** hơn.

Với cách tiếp cận này, **T5 có thể thực hiện nhiều nhiệm vụ khác nhau** như phân loại văn bản, dịch ngôn ngữ, tóm tắt, hỏi – đáp, v.v... bằng cách **chuyển đổi các bài toán đó thành bài toán sinh văn bản**.
Cách tiếp cận này **đơn giản hóa kiến trúc và quá trình huấn luyện mô hình**, đồng thời giúp **tận dụng hiệu quả tri thức học được trong giai đoạn huấn luyện trước (pre-training)** cho các tác vụ phía sau (downstream tasks).

2. Kiến trúc mô hình

T5 được xây dựng dựa trên **Transformer Encoder–Decoder** giống như mô hình dịch máy (sequence-to-sequence).

Cấu trúc:

- **Encoder**: đọc input text và trích xuất đặc trưng ngữ nghĩa.
- **Decoder**: sinh ra output text từng token một.

3. Pre-Train T5

T5 được **pre-train** trên một tập dữ liệu cực lớn tên là **C4 (Colossal Clean Crawled Corpus)**:

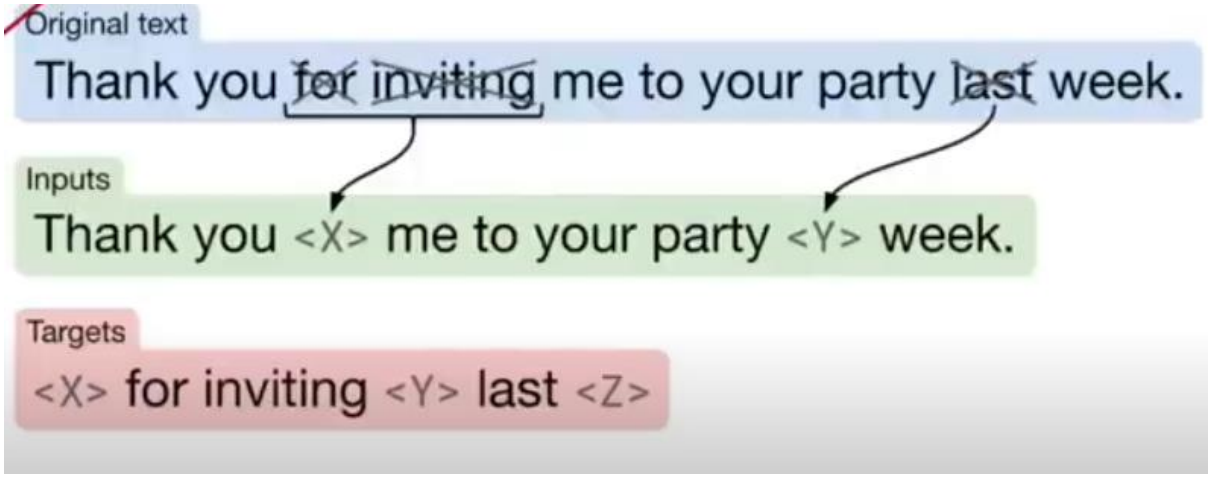
- Gồm ~750GB văn bản sạch được lọc từ Common Crawl.
- Sau khi làm sạch, còn khoảng **3.4 tỷ câu**.

Nhiệm vụ pre-training chính: "Span-Corruption" (một dạng *denoising autoencoder*):

Thay vì che (mask) **một số token riêng lẻ** như BERT, T5 **che (mask) toàn bộ một cụm từ (span)** trong văn bản.

- Các span bị che được **thay thế bằng token đặc biệt** `<X>`, `<Y>`, …
- Mô hình phải **sinh lại toàn bộ nội dung bị che** theo đúng thứ tự.

Ví dụ:



Như vậy, T5 phải **dự đoán lại cả các đoạn văn bị thiếu**, thay vì chỉ từng từ đơn lẻ — điều này buộc mô hình **hiểu sâu về ngữ cảnh và cấu trúc câu**.

Sau khi huấn luyện xong:

- Encoder học được **biểu diễn ngữ cảnh giàu ý nghĩa**.
- Decoder học cách **sinh văn bản mạch lạc, logic**.
- Mô hình **hiểu được cú pháp, ngữ nghĩa, ngữ dụng**, và có khả năng **khái quát hóa tốt** trên các tác vụ khác nhau.

4. So sánh nhanh với BERT & GPT

Mô hình	Kiến trúc	Nhiệm vụ pretrain	Đầu ra	Hướng dữ liệu	Đặc điểm
BERT	Encoder	Masked LM (từng token)	Embedding	Bidirectional	Hiểu ngữ cảnh, không sinh văn bản
GPT	Decoder	Next Token Prediction	Text	Left-to-right	Giỏi sinh text
T5	Encoder–Decoder	Span Corruption	Text	Seq2Seq	Giỏi cả hiểu và sinh text

=> T5 **tổng hợp ưu điểm của cả BERT và GPT**, nhờ đó vừa hiểu sâu ngôn ngữ (giống BERT) vừa sinh tốt (giống GPT).