

# Transformers

**Transformers** là một kiến trúc mạng nơ-ron (neural network architecture) được giới thiệu bởi Google trong bài báo nổi tiếng năm 2017: **"Attention is All You Need"**.

Nó **thay thế hoàn toàn RNN/LSTM** trong việc xử lý chuỗi dữ liệu (như văn bản), và **trở thành nền tảng** cho hầu hết các mô hình ngôn ngữ hiện đại (LLM).

## Ý tưởng cốt lõi: Attention

Điểm mạnh của Transformers là **cơ chế Attention**, cụ thể là **Self-Attention**.

Giả sử bạn có câu:  
**"The cat sat on the mat because it was tired."**

Từ **"it"** cần hiểu là đang nói đến **"the cat"**.  
Cơ chế **attention** giúp mô hình xác định được từ nào trong câu **liên quan nhất** đến từ hiện tại.

## Cấu trúc của Transformer

Có 2 phần chính:

### 1. Encoder (bộ mã hóa)

- Dùng trong BERT, T5 (phần mã hóa).
- Hiểu toàn bộ ngữ cảnh của chuỗi đầu vào.

### 2. Decoder (bộ giải mã)

- Dùng trong GPT, T5 (phần sinh văn bản).
- Dự đoán từ tiếp theo dựa trên các từ trước đó.

## Tóm tắt:

Mô hình	Dùng phần nào?
BERT	Encoder
GPT	Decoder
T5	Encoder + Decoder

## Thành phần chính trong mỗi layer

### 1. Multi-Head Self Attention

- Cho phép mô hình "chú ý" đến nhiều phần khác nhau của câu cùng lúc.

### 2. Feed-Forward Neural Network

- Một MLP đơn giản sau mỗi attention.

### 3. Layer Normalization

- Giúp mô hình ổn định trong quá trình huấn luyện.

### 4. Residual Connections

- Giúp tránh mất thông tin và tăng hiệu quả học.

## Vị trí từ (Positional Encoding)

Transformers **không có khái niệm tuần tự** như RNN.

→ Phải **thêm thông tin vị trí** bằng Positional Encoding để mô hình biết thứ tự từ trong câu.

## Vì sao Transformers lại mạnh?

- **Huấn luyện song song** (không tuần tự như RNN) → nhanh hơn rất nhiều.
- **Tăng khả năng học ngữ cảnh xa** (không bị "quên" từ đầu câu).
- **Học được từ dữ liệu lớn**, dẫn đến khả năng tổng quát mạnh mẽ.

## Hugging Face Transformers là gì?

Đây là **thư viện mã nguồn mở** giúp bạn dễ dàng:

- Sử dụng các mô hình transformer như BERT, GPT, T5, LLaMA...
- Dùng để **fine-tune**, huấn luyện, đánh giá mô hình.
- Tích hợp với **datasets**, **tokenizers**, và **PEFT** (fine-tuning hiệu quả).

## Ví dụ sử dụng nhanh (với Hugging Face):

```
from transformers import pipeline

qa = pipeline("question-answering", model="distilbert-base-cased-distilled-squad")

qa({
    "context": "The cat sat on the mat because it was tired.",
    "question": "Why did the cat sit on the mat?"
})
```

## Tóm tắt dễ nhớ

Thuật ngữ	Ý nghĩa ngắn gọn
Transformer	Kiến trúc xử lý chuỗi mạnh mẽ, thay thế RNN
Attention	Cơ chế "chú ý" đến phần quan trọng của chuỗi

Thuật ngữ	Ý nghĩa ngắn gọn
Self-Attention	Mỗi từ chú ý đến các từ khác trong chuỗi
Encoder / Decoder	Mã hóa / Sinh văn bản
Hugging Face	Thư viện dễ dùng để tận dụng mô hình này

Phiên bản #1  
Được tạo 7 tháng 5 2025 06:02:03 bởi Đỗ Ngọc Tú  
Được cập nhật 7 tháng 5 2025 06:06:29 bởi Đỗ Ngọc Tú