GPT

1.Tổng quan về GPT

GPT (Generative Pre-trained Transformer) là một mô hình ngôn ngữ (Language Model) được OpenAI giới thiệu lần đầu năm 2018. Được xây dựng dựa trên cấu trúc decoder của mô hình transfomer.

Thuật ngữ “Generative Pre-trained Transformer” mô tả chính xác các đặc tính và bản chất cốt lõi của kiến trúc mô hình ngôn ngữ này:

- **Generative (Tạo sinh)**: Mô hình có khả năng sản sinh ra văn bản mới mô phỏng ngôn ngữ của con người.
- **Pre-trained (Đào tạo trước)**: Được huấn luyện trước trên khối lượng dữ liệu văn bản cực lớn để hiểu ngôn ngữ tự nhiên. Xuyên suốt giai đoạn này, mô hình học cách dự đoán từ tiếp theo trong một chuỗi dựa vào các từ đứng trước nó.
- **Transformer (Bộ chuyển đổi)**: “Dựa trên kiến trúc Transformer (do Vaswani et al., 2017) – sử dụng Attention để hiểu mối quan hệ giữa các từ.

2. Sự phát triển của các mô hình GPT

GPT-1 (2018)

- **Paper**: *“Improving Language Understanding by Generative Pre-Training” (Radford et al., 2018).*
- **Đặc điểm**: Mô hình đầu tiên sử dụng kiến trúc Transformer và phương pháp pre-training + fine-tuning.

- **Kích thước mô hình**: 117 triệu tham số, 4,6 GB data, 12 lớp Transformer decoder.

- **Ý nghĩa**:Đặt nền móng cho mô hình ngôn ngữ lớn, khai phá khả năng học từ dữ liệu lớn và áp dụng cho các tác vụ khác nhau mà không cần huấn luyện riêng biệt.

GPT-2 (2019)

- **Paper**: *“Language Models are Unsupervised Multitask Learners”.*

- **Đặc điểm**: Thí nghiệm trên các bộ dữ liệu lớn và kỹ thuật chuẩn bị dữ liệu để huấn luyện mô hình trên nhiều tác vụ khác nhau. Không cần fine-tune vẫn có thể làm các tác vụ khác nhau qua prompt.

- **Kích thước mô hình**: 1,5 tỷ tham số, 40 GB data

- **Ý nghĩa**: Khai sinh khái niệm Zero-Shot Learning.

GPT-3 (2020)

- **Paper**: *“Language Models are Few-Shot Learners” (Brown et al., 2020).*

- **Đặc điểm**: Tiếp tục các đột phá với việc sử dụng học ít ví dụ In-Context Learning(zero-shot/one-shot/few-shot) thay vì tinh chỉnh mô hình
- **Kích thước mô hình**: 175 tỷ tham số, 600 GB data

- **Ý nghĩa**: Mở ra tiềm năng lớn cho các ứng dụng thực tế trong nhiều lĩnh vực.

GPT-4 (2023)

- **Paper**: ‘’*GPT-4 Technical Report*”
- **Đặc điểm**: Hiểu đa ngôn ngữ, xử lý hình ảnh, âm thanh, lập luận phức tạp.
- **Kích thước mô hình**: Không công khai (rất lớn)
- **Ý nghĩa**: Đưa AI đến một tầm cao mới trong việc xử lý các tác vụ phức tạp, như y tế, pháp lý, và khoa học.

3. GPT-1

Mô hình GPT-1 xây dựng nền móng cho các mô hình sau này

Bộ giải mã (decoder) gồm **12 lớp** trong kiến trúc **Transformer**.

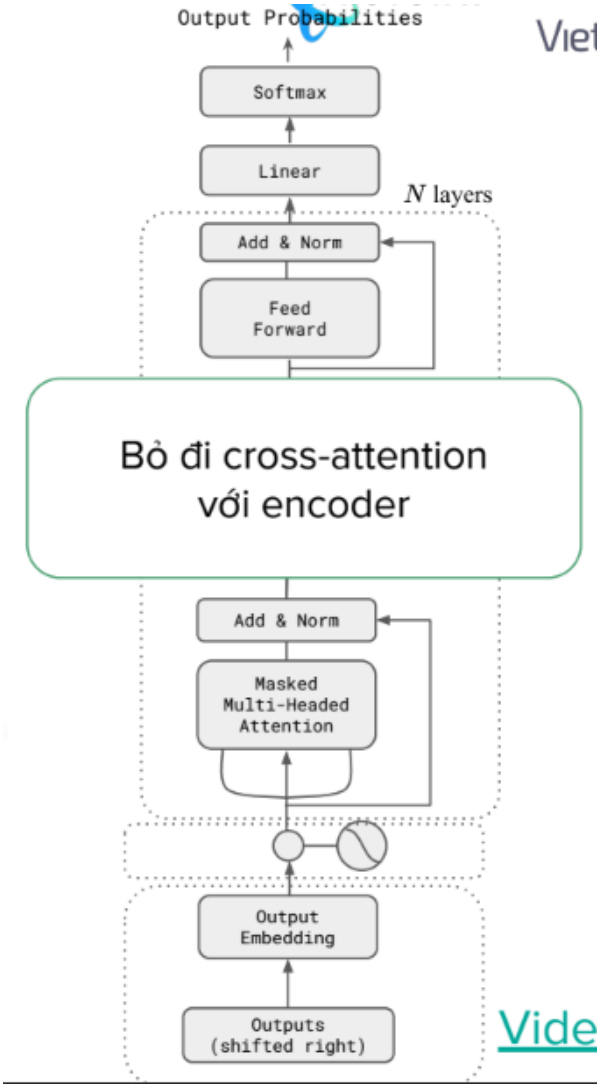
Dữ liệu huấn luyện, bao gồm hơn **7.000 cuốn sách khác nhau** (tổng cộng **4,6 GB văn bản**), được lấy từ **BooksCorpus**.

3.1. Unsupervised pre-training

Input representation:

Input = Token Embedding + Positional Embedding

→ Không có Segment Embedding (vì GPT không dùng cặp câu).



$$h_0 = UW_e + W_p$$

$$h_l = transformer\_block(h_{l-1}) \quad \forall i \in [1, n]$$

$$P(u) = softmax(h_n W_e^T)$$

Trong đó:

- $U = (u_{-k}, \ldots, u_{-1})$ là **vector ngữ cảnh** của các token,
- n là **số lớp**,
- W_e là **ma trận embedding của token**,
- W_p là **ma trận embedding vị trí (position embedding)**.

Hàm Loss:

$$L_1(U) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

Trong đó:

- k: kích thước cửa sổ ngữ cảnh (context window)
- Θ:parameters

3.2. Supervised fine-tune

Sau khi GPT-1 được **pre-train không giám sát**, mô hình được **fine-tune có giám sát** trên các nhiệm vụ cụ thể.

- **Dữ liệu đầu vào:** chuỗi token x^1,…,x^m và **nhãn** y.
- **Cách làm:** lấy đầu ra cuối cùng của Transformer h_l^m, đưa qua **một lớp tuyến tính mới** với tham số W_y

$$P(y|x_1, \ldots, x_m) = softmax(h_l^m W_y)$$

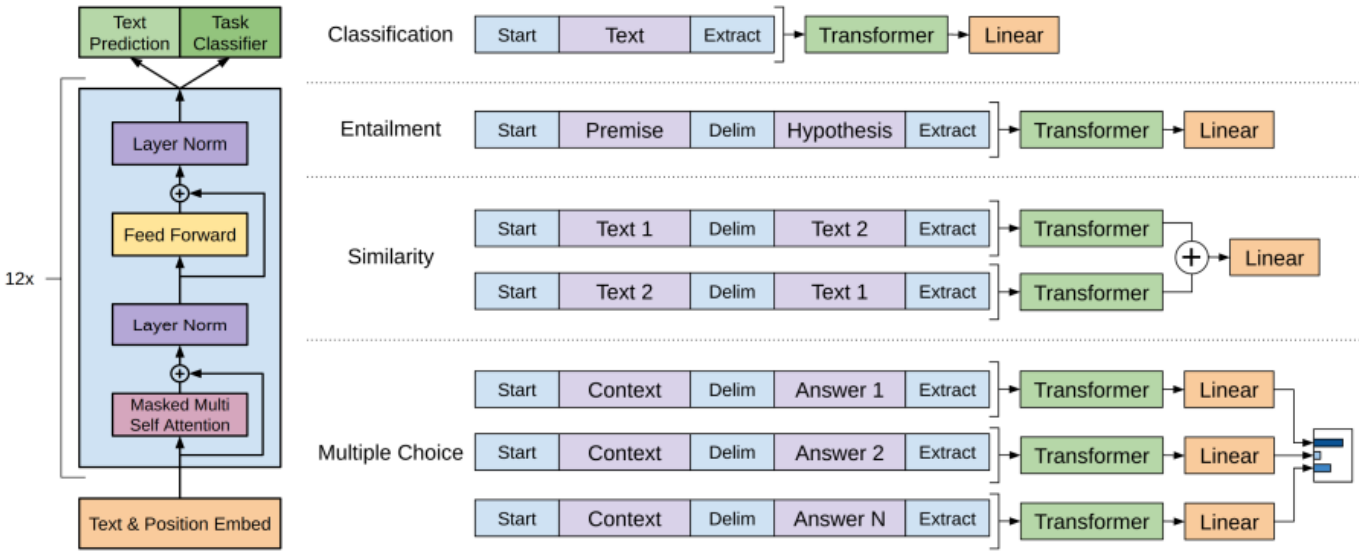
  - **Mục tiêu chính:** tối đa hóa xác suất dự đoán đúng nhãn

$$L_2(C) = \sum \log P(y|x)$$

- **Cải tiến:** thêm mục tiêu phụ là **language modeling loss** (từ giai đoạn pre-train) để giúp:
  - cải thiện khả năng khái quát hóa,
  - tăng tốc độ hội tụ.

→ Tổng loss khi fine-tuning:

$$L_3(C) = L_2(C) + \lambda L_1(C)$$



4. GPT-2

GPT-2 **chỉ được huấn luyện** theo mục tiêu duy nhất:

$$P(w_t | w_1, w_2, ..., w_{t-1})$$

→ Dự đoán từ kế tiếp trong chuỗi.

Tuy nhiên, nhờ:

- Quy mô **1.5B tham số**,
- Dữ liệu **WebText rất đa dạng (dịch, hỏi đáp, tóm tắt, hội thoại, truyện, lập trình, …)**,

nó **vô tình học được quy luật** của nhiều nhiệm vụ ngôn ngữ.
Vì vậy, khi ta cho một prompt mô tả rõ nhiệm vụ, GPT-2 có thể thực hiện được — dù không hề được huấn luyện cụ thể, do đó khai sinh ra khái niệm **zero-shot learning**

Không cần fine-tuning — chuyển sang “zero-shot learning”:

- GPT-1 cần fine-tuning trên từng task (như phân loại, trả lời câu hỏi).

- GPT-2 sử dụng **Zero-Shot Learning (ZSL)** là khả năng của mô hình giải quyết một nhiệm vụ mà nó chưa từng được huấn luyện trực tiếp, chỉ dựa vào mô tả ngôn ngữ trong prompt, nhờ học được mẫu ngữ cảnh (in-context learning) trong quá trình pre-train.

5. GPT-3

OpenAI nhận ra:

> “Nếu mô hình đủ lớn, chỉ cần **cho nó vài ví dụ** trong prompt, nó có thể **hiểu và bắt chước nhiệm vụ**.”

Vì vậy GPT-3 (175B parameters, 2020) được thiết kế để **học trong ngữ cảnh (in-context learning)**:

- Dùng chính prompt làm "tập huấn luyện nhỏ".
- Không cập nhật trọng số (model weights không đổi).
- Mô hình tự nội suy quy luật từ ví dụ → giống như “học tạm thời trong bộ nhớ ngắn hạn”.

GPT-3 được xem là mô hình đầu tiên **thể hiện rõ năng lực in-context learning**.

Kiểu học	Mô tả	Ví dụ
Zero-shot	Không cho ví dụ nào, chỉ đưa hướng dẫn	"Translate English to French: cat → ?"
One-shot	Cho 1 ví dụ mẫu	"Translate: cat → chat. dog → ?"
Few-shot	Cho vài ví dụ (2–10) trong prompt	"Translate: cat→chat, dog→chien, house→maison, tree→?"

5. GPT-4

GPT-4 là mô hình đa phương thức (multimodal), có khả năng xử lý cả văn bản và hình ảnh, đánh dấu bước tiến lớn từ GPT-3 (chỉ văn bản).

GPT-4 được huấn luyện trên dữ liệu khổng lồ và tinh chỉnh với RLHF (Reinforcement Learning from Human Feedback) để cải thiện an toàn và tính hữu ích. Nó là nền tảng cho ChatGPT Plus và các công cụ doanh nghiệp, với quy mô ước tính hơn 1 nghìn tỷ tham số (chi tiết chính thức không công khai).

GPT-4o (o = omni) mở rộng sang âm thanh và video thời gian thực.