Week 8 - LLMs

Ngày 18 tháng 10 năm 2025

1 Khái niệm và các kiến trúc phổ biến

1.1 Khái niệm

Mô hình ngôn ngữ lớn (Large Language Models) là một dạng các mô hình học sâu tiên tiến, được thiết kế để hiểu, xử lý và sinh ngôn ngữ một cách mạch lạc và đúng ngữ cảnh. Khác với các mạng nơ-ron truyền thống, LLM sỡ hữu quy mô vượt trội, thường gồm hàng chục đến hàng trăm lớp và chứa từ vài tỷ tham số trở lên, cho phép chúng học được các cấu trúc và quy luật của ngôn ngữ.

Các mô hình này được huấn luyện trên khối lượng dữ liệu văn bản khổng lồ, sử dụng kiến trúc Transformer làm nền tảng, giúp mô hình nắm bắt được mối quan hệ giữa các từ trong ngữ cảnh dài.

Ngày nay, LLMs được ứng dụng rộng rãi trong nhiều tác vụ NLP như: Text Generation, Text Summarization, Machine Translation, etc. Một nhánh phát triển nâng cao hơn là Multimodal LLMs có khả năng hiểu và sinh ra nội dung từ nhiều dạng dữ liệu khác nhau như văn bản, hình ảnh, âm thanh.

1.2 Các kiến trúc phổ biến

Sự xuất hiện của kiến trúc Transformer vào năm 2017 đã đánh dấu một bước ngoặc quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên, khi thay thế các Sequential Models truyền thống như RNN, LSTM. Các LLMs hiện đại hầu hết đều được xây dựng dựa trên kiến trúc Transformer, và có thể được phân chia thành 3 nhóm chính:

- Encoder-only
- Decoder-only
- Encoder-Decoder

1.2.1 Encoder-only

Kiến trúc tự mã hóa dựa trên bộ mã hóa (encoder) của Transformer. Mục tiêu huấn luyện là giúp mô hình hiểu và biểu diễn ngữ cảnh trong văn bản đầu vào thông qua cơ chế Masked Language Modeling,tức là che khuất ngẫu nhiên một số token trong câu và yêu cầu mô hình dự đoán lại các token bị ẩn. Nhờ cách huấn luyện này, mô hình học được biểu diễn ngữ nghĩa ngữ cảnh giàu thông tin, rất phù hợp cho các tác vụ hiểu ngôn ngữ tự nhiên như phân loại văn bản, trích xuất thực thể (NER), hoặc trả lời câu hỏi (QA).

Các ví dụ tiêu biểu: **BERT, RoBERTa, DistilBERT**

1.2.2 Decoder-only

Kiến trúc tự hồi quy dựa trên bộ giải mã (decoder) của Transformer, trong đó mô hình được huấn luyện theo cơ chế Causal Language Modeling (CLM) — tức là dự đoán token kế tiếp dựa trên chuỗi các token trước đó. Cụ thể, mô hình học xác suất có điều kiện của chuỗi từ:

$$P(w_1, w_2, \ldots, w_n) = \prod_{t=1}^{n} P(w_t \mid w_1, w_2, \ldots, w_{t-1})$$

Nhờ đặc trưng này, mô hình dạng này đặc biệt mạnh trong các tác vụ sinh ngôn ngữ tự nhiên, nơi cần tạo ra văn bản mạch lạc, liên tục và có ngữ cảnh. Đây cũng là loại kiến trúc phổ biến nhất trong các LLM hiện đại, được ứng dụng trong chatbot, sinh văn bản, sinh code.

Các ví dụ tiêu biểu: **GPT, LLaMa** series.

1.2.3 Decoder-Encoder

Kiến trúc này kết hợp cả 2 thành phần Encoder và Decoder trong Transformer, cho phép mô hình hiểu ngữ cảnh đầu vào sinh ra đầu ra có điều kiện dựa trên thông tin đó. Encoder trước tiên xử lý chuỗi đầu vào để trích xuất đặc trưng ngữ nghĩa, sau đó Decoder sử dụng các đặc trưng này để tạo ra chuỗi đầu ra tương ứng. Kiến trúc này đặc biệt thích hợp cho các tác vụ chuyển đổi ngôn ngữ hoặc nội dung như Machine Translation, Text Summerization, etc.

Các ví dụ tiêu biểu: **T5, BART, Pangu series**

1.2.4 Mixture of Experts

Mixture of Experts (MoE) là một xu hướng kiến trúc được ưa chuộng trong các LLMs giai đoạn gần đây, được xem như một giải pháp mở rộng quy mô mô hình, hiệu quả về chi phí. Thay vì kích hoạt tất cả tham số của mô hình trong quá trình suy luận, MoE cho phép tăng tổng số tham số của mô hình, trong khi chỉ kích hoạt một tập con nhỏ các tham số gọi là experts cho mỗi token đầu vào. Cách tiếp cận này được gọi là tính toán có điều kiện, giúp mô hình vừa học được khả năng biểu diễn lớn vừa tiết kiệm tài nguyên tính toán đáng kể.

Về mối quan hệ giữa quy mô và hiệu suất, MoE mang lại hiệu năng cao tương đương hoặc vượt trội so với các mô hình dense cùng kích thước hoạt động, nhưng với chi phí tính toán thấp hơn nhiều. Nhờ chỉ kích hoạt một phần nhỏ các experts trong mỗi lần suy luận, mô hình có thể mở rộng đến hàng trăm tỷ tham số mà vẫn duy trì khả năng triển khai thực tế trên hạ tầng giới hạn. Đây là một cách tiếp cận hiệu quả về chi phí để nâng cao năng lực của LLM mà không cần tăng tuyến tính lượng tài nguyên huấn luyện.

Về mối quan hệ kiến trúc, nhiều LLM tiên tiến gần đây đã áp dụng cơ chế MoE và có cấu trúc nội bộ tương đồng. Chẳng hạn, LLaMA 4 của Meta được cho là đã tích hợp cơ chế MoE tương tự như DeepSeek-V3, một mô hình khổng lồ với 671 tỷ tham số, nhưng chỉ kích hoạt khoảng 37 tỷ tham số cho mỗi token trong quá trình suy luận. Cách thiết kế này thể hiện rõ xu hướng "mô hình cực lớn nhưng tính toán hiệu quả", vốn đang định hình thế hệ LLM mới hiện nay.

1.3 Mối quan hệ giữa các kiến trúc LLM

Ba loại kiến trúc **Encoder-only**, **Decoder-only**, và **Encoder–Decoder** đều bắt nguồn từ mô hình Transformer gốc, nhưng được tinh chỉnh cho các mục tiêu khác nhau trong xử lý ngôn ngữ tự nhiên. Chúng chia sẻ cùng nền tảng kỹ thuật,bao gồm cơ chế *multi-head self-attention*, mạng *feed-forward*, *residual connection*, *layer normalization*, và *positional encoding*,nhưng khác nhau về mục tiêu huấn luyện.

- **Encoder-only:** (ví dụ: BERT) tập trung vào việc hiểu ngữ cảnh hai chiều (*bidirectional context*), phù hợp cho các tác vụ hiểu ngôn ngữ như phân loại văn bản, trích xuất thông tin và tìm kiếm ngữ nghĩa.
- **Decoder-only:** (ví dụ: GPT, LLaMA) dựa trên mô hình tự hồi quy (*auto-regressive*), chỉ sử dụng ngữ cảnh phía trước, tối ưu cho các tác vụ sinh ngôn ngữ như hội thoại, viết văn bản hoặc lập trình.
- **Encoder–Decoder:** (ví dụ: T5, Pangu series) kết hợp cả hai thành phần, trong đó *Encoder* mã hóa đầu vào và *Decoder* sinh đầu ra dựa trên thông tin đã mã hóa, thích hợp cho các tác vụ sinh có điều kiện (*conditional generation*) như dịch máy, tóm tắt hoặc trả lời câu hỏi.

Về mối quan hệ phát triển, ba kiến trúc này không tách biệt hoàn toàn mà bổ sung cho nhau:

- Kiến trúc **Decoder-only** có thể được xem là trường hợp rút gọn của **Encoder–Decoder**, khi bỏ phần *Encoder* và chỉ giữ lại cơ chế tự chú ý trong *Decoder*.
- Ngược lại, **Encoder-only** có thể coi là một nửa của **Encoder–Decoder**, chỉ sử dụng phần mã hóa mà không sinh đầu ra.
- Các mô hình hiện đại (ví dụ: GPT-4V, Gemini 1.5, Claude 3.5) thường kết hợp linh hoạt các đặc điểm của cả ba kiến trúc, cùng với các mở rộng như **Mixture-of-Experts (MoE)** hoặc **Multimodal**, hình thành nên các kiến trúc lai mạnh mẽ hơn.

2 So sánh các nhánh kiến trúc LLM

2.1 So sánh các Đại diện LLM Mã nguồn đóng (Closed-Source)

Các mô hình mã nguồn đóng thường đạt hiệu suất dẫn đầu (*state-of-the-art performance*) nhưng thiếu tính minh bạch về cấu trúc chi tiết và dữ liệu huấn luyện. Dưới đây là ba đại diện tiêu biểu:

Bảng 1: So sánh các mô hình LLM mã nguồn đóng

Mô hình	**Kiến trúc tiêu biểu**	**Ưu điểm (Advantages)**	**Nhược điểm**
GPT series (GPT-3, GPT-4)	Auto-regressive/ Decoder-only. Cấu trúc dựa trên Transformer gốc. GPT-3 có 175B tham số, trong khi GPT-4o là mô hình đa phương thức (*multimodal*).	Tiên phong trong kiến trúc tự hồi quy. GPT-4 và GPT-4o đạt độ chính xác vượt trội trên các điểm chuẩn như *HellaSwag* và *WinoGrande* và nổi bật với khả năng few-shot.	Mã nguồn đóng, chi tiết cấu trúc nội bộ và kỹ thuật tối ưu hóa không được công khai.
Gemini	Phát triển dựa trên kiến trúc *Pathways* của Google. Hỗ trợ kích hoạt thưa thớt (*sparse activations*) cho các tác vụ đa phương thức và đa nhiệm.	Gemini Ultra và GPT-4o là những mô hình có điểm trung bình cao nhất trên các điểm chuẩn như MMLU. Gemini 1.5 được ghi nhận có khả năng hiểu ngữ cảnh đa phương thức với hàng triệu token.	Mã nguồn đóng, thông tin chi tiết không được công khai.
Claude	Auto-regressive	ác mô hình trong dòng Claude 3 đã thể hiện sự tiến bộ đáng kể trong khả năng suy luận thông thường (commonsense inference capabilities) trên các điểm chuẩn như WinoGrande. Khả năng sinh code cũng là 1 điểm đáng chú ý của dòng Claude .	Mã nguồn đóng, hạn chế về khả năng tùy chỉnh và tái huấn luyện.

2.2 So sánh các Đại diện LLM Mã nguồn mở (Open-Source)

Các mô hình mã nguồn mở (Open-weight) đang thúc đẩy mạnh mẽ sự minh bạch và đổi mới trong cộng đồng nghiên cứu. Dưới đây là ba đại diện nổi bật:

Bảng 2: So sánh các mô hình LLM mã nguồn mở

Mô hình	Kiến trúc tiêu biểu	Ưu điểm	Nhược điểm
LLaMA (LLaMA 2 – LLaMA 4)	Decoder-only. Sử dụng *RMSNorm* thay cho *LayerNorm* và *Rotary Positional Embedding (RoPE)*. LLaMA 2 giới thiệu *Grouped-Query Attention (GQA)*; LLaMA 4 tích hợp *Mixture-of-Experts (MoE)* tương tự DeepSeek-V3.	Hiệu quả và dễ triển khai. Các phiên bản nhỏ (7B, 13B) phù hợp cho môi trường cục bộ. LLaMA 2 tăng gấp đôi độ dài ngữ cảnh so với bản đầu.	Hiệu suất ban đầu chưa vượt trội so với GPT-3. Các phiên bản nhỏ hạn chế trong suy luận logic phức tạp.
Mistral (Mistral 7B, Mixtral, Mistral Small 3.1)	Decoder-only. Mistral 7B và Mixtral sử dụng *Mixture-of-Experts (MoE)*; Mistral Small 3.1 dùng *Grouped-Query Attention (GQA)*.	Rất hiệu quả và nhanh: Mistral Small 3.1 (24B) vượt Gemma 3 (27B) trên nhiều điểm chuẩn (trừ toán học). Các mô hình MoE đạt hiệu suất tương đương hoặc cao hơn GPT-3.5 với kích thước nhỏ hơn.	Hiệu suất giảm nhẹ trong các bài kiểm tra toán học và suy luận số học.
BLOOM	Decoder-only. Mô hình quy mô lớn với 176B tham số.	Dự án hợp tác mã nguồn mở đa ngôn ngữ của hơn 1000 nhà nghiên cứu. Huấn luyện trên tập dữ liệu gồm 46 ngôn ngữ tự nhiên và 13 ngôn ngữ lập trình.	Kích thước rất lớn khiến chi phí huấn luyện và triển khai cao; hiệu suất suy luận chậm hơn các mô hình nhỏ hơn như LLaMA.

3 Đánh giá LLM

Các bộ benchmarks đóng vai trò quan trọng trong việc đánh giá năng lực tổng quát hóa, hiểu ngôn ngữ và khả năng suy luận của các mô hình ngôn ngữ lớn (LLM). Dưới đây là bốn bộ điểm chuẩn tiêu biểu được sử dụng rộng rãi trong cộng đồng nghiên cứu.

3.1 MMLU (Massive Multitask Language Understanding)

MMLU là một trong những điểm chuẩn toàn diện và có ảnh hưởng nhất trong việc đánh giá khả năng tổng quát hóa kiến thức của LLM. Bộ dữ liệu này được giới thiệu bởi Hendrycks et al. (2021).

- **Mục tiêu:** Đánh giá khả năng hiểu ngôn ngữ và mức độ khái quát hóa kiến thức chuyên môn của mô hình trên các chủ đề đa lĩnh vực.
- **Phạm vi:** Bao gồm 57 tác vụ (*tasks*) trải rộng trên nhiều lĩnh vực, từ nhân văn (*humanities*) đến khoa học tự nhiên và kỹ thuật (*STEM*), tương đương các kỳ thi bậc trung học và đại học.
- **Định dạng:** Các tác vụ là câu hỏi trắc nghiệm bốn lựa chọn (*multiple-choice*), yêu cầu mô hình vận dụng kiến thức ngôn ngữ và kiến thức chuyên ngành để chọn đáp án đúng.
- **Metric:** Độ chính xác (*Accuracy*) — tỉ lệ câu trả lời đúng trên toàn bộ tập dữ liệu.
- **Thực tế:** Các mô hình tiên tiến như **GPT-4o** và **Gemini Ultra** đã đạt kết quả hàng đầu trên MMLU, thể hiện khả năng hiểu ngôn ngữ và khái quát hóa kiến thức vượt trội.

3.2 SuperGLUE (Super General Language Understanding Evaluation)

SuperGLUE được phát triển bởi Wang et al. (2019) nhằm mở rộng và nâng cao mức độ thách thức so với bộ *GLUE* trước đó, hướng tới đánh giá khả năng suy luận và hiểu ngôn ngữ tự nhiên ở cấp độ cao hơn.

- **Mục tiêu:** Đo lường khả năng suy luận phản biện (*critical reasoning*) và khả năng hiểu ngôn ngữ sâu của các mô hình AI.

- **Đặc điểm:** Bao gồm một tập hợp các tác vụ phức tạp và đa dạng hơn GLUE, được thiết kế để phản ánh những thách thức thực tế trong hiểu ngôn ngữ.
- **Loại tác vụ:** Gồm trả lời câu hỏi (*question answering*), suy luận ngữ nghĩa (*entailment*), phân giải đồng tham chiếu (*coreference resolution*), và phân biệt nghĩa từ (*word sense disambiguation*).
- **Metric:** Điểm tổng hợp (*Composite Score*) tính trung bình từ hiệu suất trên tất cả các tác vụ, khuyến khích mô hình duy trì sự cân bằng giữa các năng lực khác nhau.
- **Thực tế:** Các mô hình như **Claude 3** và **GPT-4** đạt điểm cao nhất trên SuperGLUE, chứng minh khả năng suy luận và hiểu ngôn ngữ nâng cao.

3.3 HellaSwag

HellaSwag (Zellers et al., 2019) là một điểm chuẩn quan trọng để đánh giá khả năng suy luận thông thường (*commonsense reasoning*) của mô hình ngôn ngữ.

- **Mục tiêu:** Kiểm tra khả năng hiểu các tình huống đời thường và chọn phần tiếp theo hợp lý nhất trong một chuỗi mô tả.
- **Định dạng:** Nhiệm vụ chọn câu kết hợp lý nhất trong bốn tùy chọn được cung cấp cho mỗi ngữ cảnh.
- **Metric:** Độ chính xác (*Accuracy*) của dự đoán.
- **Thực tế:** Các mô hình như **GPT-4** và **PaLM 2** đã vượt trội đáng kể so với các mô hình thế hệ trước, chứng minh tiến bộ lớn trong khả năng suy luận thông thường.

3.4 WinoGrande

WinoGrande (Sakaguchi et al., 2020) là bộ dữ liệu mở rộng dựa trên thí nghiệm kinh điển *Winograd Schema Challenge*, được thiết kế nhằm kiểm tra khả năng suy luận ngữ nghĩa và giải quyết mơ hồ ngữ cảnh của mô hình.

- **Mục tiêu:** Đánh giá khả năng suy luận thông thường và xử lý ngữ nghĩa trong các tình huống có yếu tố mơ hồ ngữ pháp hoặc ngữ cảnh.
- **Đặc điểm:** Bao gồm hàng chục nghìn mẫu dữ liệu kiểm tra khả năng hoàn thành câu bằng cách xác định thực thể đúng trong các đại từ mơ hồ (*ambiguous pronouns*).
- **Metric:** Độ chính xác (*Accuracy*) trong việc chọn thực thể đúng.
- **Thực tế:** Các mô hình hiện đại như **GPT-4o** và **Claude 3** đạt hiệu suất gần mức con người, thể hiện năng lực hiểu ngữ cảnh tự nhiên ngày càng sâu sắc.