생성형 AI의 핵심 원리 및 최신 동향 요약

1. 생성형 AI의 기본 원리 (트랜스포머 기반)

\* 다음에 올 단어 예측 학습: 모델은 문장을 한 칸씩 뒤로 미룬 후 Self-Attention을 이용해 다음에 올 단어를 예측하도록 학습합니다. 특히 디코더에서는 Masked Self-Attention을 통해 미래 정보를 보지 못하게 합니다.

\* 모델 표현력 증대: \*\*FFN (Feed-Forward Network)\*\*은 Self-Attention 레이어의 출력값을 비선형적으로 변환하여 모델의 복잡한 패턴 학습 능력을 높여줍니다.

\* 순서 정보 주입: 트랜스포머는 순환 구조가 없어 단어의 순서 정보를 명시적으로 알지 못합니다. 이를 위해 Position Embedding을 단어 임베딩에 더해 문장 내 위치 관계를 파악하게 합니다.

\* 단어 벡터화: 컴퓨터가 텍스트를 처리하기 위해 단어는 Vectorization (임베딩) 과정을 거쳐 숫자로 표현됩니다.

\* 학습된 모델로 추론: 이렇게 학습된 모델은 새로운 입력에 대해 다음에 올 단어를 순차적으로 예측하며 문장을 \*\*생성(추론)\*\*합니다.

2. 대화형 챗봇(예: Gemini)에서의 활용

대화형 챗봇은 위 기본 원리 위에 다음과 같은 기술들이 추가됩니다.

\* Fine-tuning (미세 조정): 사전 학습된 모델을 특정 대화 데이터셋으로 추가 학습시켜 대화 맥락 이해 및 답변 능력을 강화합니다.

\* Turn-taking 및 Context Management: 여러 차례의 대화 턴을 거치며 이전 대화 내용을 기억하고 맥락을 유지하여 자연스러운 대화를 이어갑니다.

\* Retrieval Augmented Generation (RAG): 내부 지식 기반이나 웹 검색에서 관련 정보를 찾아 답변을 생성하여 정확성과 신뢰성을 높입니다.

\* Safety & Alignment: 유해하거나 편향된 답변을 생성하지 않도록 사람의 피드백을 통한 강화 학습(RLHF) 등으로 모델을 조정합니다.

\* Multimodality (멀티모달리티): 텍스트 외에 이미지, 오디오 등 다양한 형태의 정보를 이해하고 생성하는 능력을 포함합니다.

3. 모델 고도화를 위한 최신 기법

모델 성능을 더욱 향상시키기 위한 최신 연구 동향 및 핵심 기술은 다음과 같습니다.

\* Self-Attention의 MLA (Multi Head Latent Attention) 도입:

\* Google의 Perceiver 아키텍처에서 사용되는 개념입니다.

\* 대용량의 입력 데이터를 직접 처리하는 대신, 작은 고정 크기의 \*\*'Latent Array'\*\*를 도입합니다.

\* 입력 데이터와 Latent Array 간의 Cross-Attention 및 Latent Array 내부의 Self-Attention을 통해 효율적으로 정보를 압축하고 처리하여, 입력 시퀀스 길이에 따른 계산 복잡도 문제를 완화합니다.

\* FFN의 MoE (Mixture of Experts) 도입:

\* FFN 레이어에 여러 개의 '전문가 네트워크'를 두고, '게이팅 네트워크'가 입력 토큰에 따라 적절한 전문가들을 선택하도록 하는 방식입니다.

\* 전체 파라미터 수는 크게 늘지만 실제 계산량은 크게 증가하지 않으면서, 모델의 용량(Capacity)을 비약적으로 늘려 성능을 크게 향상시킵니다.

\* Position Embedding의 RoPE (Rotary Position Embedding) 도입:

\* 단어 임베딩에 직접 위치 정보를 더하는 대신, 쿼리(Query)와 키(Key) 벡터에 회전 행렬을 적용하여 상대적인 위치 관계를 인코딩합니다.

\* 이를 통해 Self-Attention이 장거리 의존성을 더 잘 포착하고, 더 긴 시퀀스에서도 안정적인 성능을 유지할 수 있도록 돕습니다.

deepseek r1 논문

<https://arxiv.org/abs/2501.12948>

<https://arxiv.org/pdf/2501.12948>

attention is all you need

<https://arxiv.org/pdf/1706.03762>

unsloth

<https://devocean.sk.com/blog/techBoardDetail.do?ID=166285&boardType=techBlog>

파인튜닝!