중간고사 요약정리

1. 딥러닝이 주목받은 이유

딥러닝은 기존 기계학습을 뛰어넘어 현대 AI 혁신을 이끈 핵심 기술이다. 그 이유는 크게 **표현학습, 스케일링, 범용성** 세 가지로 정리할 수 있다.

---

2. NLP 발전과 LLM의 등장

자연어 처리(NLP)는 크게 네 단계로 발전하였다.

1. **규칙/통계 기반 모델** (n-그램 등)
   → 문맥 길이 한계, 차원의 저주 문제 발생
2. **신경망 기반 모델** (Word2Vec, RNN, LSTM)
   → 단어 의미를 벡터로 표현, 문맥 처리 가능
3. **Transformer 등장 (2017)**
   → Self-Attention 메커니즘
   → 병렬 처리 가능, 장기 의존성 해결
4. **대규모 언어모델(LLM)**
   → GPT, BERT 등
   → 사전학습 + 미세조정 패러다임 확립

---

3. 사전학습과 미세조정

현대 LLM은 **사전학습 → 미세조정** 구조를 따른다.

① 사전학습 (Pre-training)

- 대규모 텍스트로 "다음 단어 예측" 학습
- 언어의 일반 패턴과 상식 습득

② 미세조정 (Fine-tuning)

- 특정 작업에 맞춰 추가 학습
- 적은 데이터로도 높은 성능

최근에는 효율적 튜닝 기법인 **LoRA**가 활용되며, 전체 모델을 재학습하지 않고 일부 가중치만 조정하여 자원 효율을 높인다.

---

4. LLM의 작동 원리

LLM의 본질은 **확률 기반 다음 단어 예측 모델**이다.

수식적으로는

$$P(w_t \mid w_{<t})$$

즉, 이전 단어들을 조건으로 다음 단어의 확률을 계산한다.

Self-Attention은 각 단어 간 관계를 계산하여 문맥을 반영한다.
Query, Key, Value 구조를 통해 중요 단어에 더 높은 가중치를 부여한다.

---

5. GPT 발전과 RLHF

GPT 모델은 다음과 같이 발전하였다.

- GPT-1 → 구조 실험 단계
- GPT-2 → 문맥 유지 능력 향상
- GPT-3 → In-Context Learning 등장
- GPT-3.5/4 → Instruction Tuning + RLHF 적용

RLHF (Reinforcement Learning from Human Feedback)

3단계 구조:

1. SFT (지도 미세조정)
2. 보상모델 학습
3. PPO 강화학습

→ 인간 선호 기반으로 모델을 정렬(alignment)
→ 더 안전하고 공손한 답변 생성

---

6. LLM의 한계: 환각(Hallucination)

환각이란 **사실이 아닌 내용을 그럴듯하게 생성하는 현상**이다.

원인:

- 확률 기반 생성 구조
- 데이터 한계
- 노출 편향

유형:

- 사실 오류
- 논리 비약
- 인용 오류
- 형식 오류

완화 전략:

- 근거 요구
- 자가 검증
- 명확한 프롬프트
- RAG 활용
- 파라미터 조정
- 가드레일 적용

---

7. 자원과 비용 문제

초거대 모델 학습에는 막대한 자원이 필요하다.

- 수백~수천억 파라미터
- 수백만~수억 달러 비용
- 대규모 GPU 인프라
- 높은 전력 소비

따라서 기술적 발전과 함께 경제적·환경적 고려가 필수적이다.

8. LLM의 기본 원리와 한계

대형 언어 모델(LLM)은 **다음 단어 예측 확률 모델**이다.
수식적으로는

$$P(w_t \mid w_{<t})$$

즉, 이전 문맥을 조건으로 다음 단어의 확률을 계산한다.
Transformer의 Self-Attention 구조를 통해 장기 의존성을 처리하며, GPT 계열 모델은 이를 기반으로 발전하였다.

LLM은 사전학습(Pre-training) 후 미세조정(Fine-tuning)을 거치며, 특히 RLHF(SFT → 보상 모델 → PPO)를 통해 인간 선호에 정렬(alignment)된다.

한계

- 환각(Hallucination): 사실이 아닌 내용을 그럴듯하게 생성
- 데이터 편향
- 최신 정보 한계
- 확률 기반 구조로 인한 진실성 보장 불가

---

9. 사전학습과 In-Context Learning

사전학습(Pre-training)

- 대규모 데이터로 언어 패턴 학습
- 일반적 지식과 문맥 처리 능력 확보

In-Context Learning

- 추가 파라미터 업데이트 없이
- 프롬프트 내 예시를 통해 새로운 작업 수행

이는 GPT-3 이후 핵심 능력으로 주목되었으며, 모델은 문맥을 통해 패턴을 추론한다.

---

10. Zero-shot, One-shot, Few-shot

샷(shot)은 프롬프트에 포함된 예시 개수를 의미한다.

방식	특징	장점	한계
Zero-shot	예시 없음	빠르고 간결	정확도·형식 통제 약함
One-shot	예시 1개	형식 안정성 증가	일반화 한계
Few-shot	2~5개 예시	정확도·일관성 향상	토큰 비용 증가

일반적으로 **Few-shot ≥ One-shot ≥ Zero-shot** 순으로 안정성이 높아진다.

---

11. 프롬프트의 본질

프롬프트는 **자연어 기반 프로그래밍 인터페이스**이다.

프롬프트 품질이 곧 출력 품질을 결정한다.

좋은 프롬프트의 핵심 요소는:

- Role (역할)
- Task (과업)
- Context (맥락)
- Format (형식)
- Constraints (제약)
- Examples (예시)

이를 RTF-ICE 구조로 설명할 수 있다.

---

12. 좋은 프롬프트 vs 나쁜 프롬프트

나쁜 프롬프트 특징

- 모호함
- 과도하게 광범위
- 맥락 부족
- 형식 미지정

- 복합 지시 혼합

좋은 프롬프트 원칙

- 명확성(Clarity)
- 구체성(Specificity)
- 맥락 제공(Context)
- 구조화(Structure)
- 제약 조건(Constraints)
- 역할 지정(Role)
- 예시 활용(Few-shot)

---

13. 반복 개선(Iterative Refinement)

프롬프트는 한 번에 완성되지 않는다.

과정:

1. Draft (초안 작성)
2. Run (실행)
3. Observe (결과 평가)
4. Refine (수정)

반복을 통해

- 정확도 향상
- 과잉 설명 제거
- 형식 안정화
- 환각 감소

---

14. 역할 지시 프롬프트(Role Prompt)

Role Prompt는 모델에 특정 페르소나를 부여하여

어조·관점·정보 선택을 변화시키는 기법이다.

예:

- 역사 교사 → 교육적 설명
- 관광 가이드 → 친근한 설명
- 전문가 → 전문적 용어 사용

역할은 톤 제어의 강력한 도구이며,
대상(Audience)과 결합하면 어휘 난도까지 조절 가능하다.

---

15. 톤/말투 제어

톤은 감정적 어조이며 스타일 제어의 핵심 요소이다.

유형:

- 공손한 톤
- 친근한 톤
- 냉정·객관적 톤
- 전문적 톤
- 설득적 톤
- 유머러스한 톤

프롬프트에서 직접 명시하거나
역할을 통해 간접 제어할 수 있다.