3-2 이미지 파운데이션 모델

목차

1. 딥러닝 및 이미지 파운데이션 모델

학습 목표

- 이미지 기반 학습 모델의 구조와 작동 방식을 이해하고, 일반화 성능을 높이는 기법을 학습한다.
- 이미지 생성 및 분류의 차이를 파악하고, 실습을 통해 기본 모델을 적용한다.
- 다양한 모델 구조와 개선 방법을 사례 중심으로 분석한다.

1. 파운데이션 모델

1-1. 파운데이션 모델이란?

AI 모델

- 예시 : 뉴럴네트워크
- 입력 → 뉴럴네트워크 → 출력

이상적인 AI 모델

- 만약 AI 모델이 이 세상의 모든 데이터를 모두 기억한다면?
- 내가 얻고 싶은 답과 유사한 답이 이미 DB에 저장되어 있을 확률이 높음 → 검색엔진

현실적인 기계학습 모델

- 학습 = AI 모델에 데이터를 패턴화하여 압축

파운데이션 모델이란?

- 대규모 데이터를 폭넓게 학습한 후, 다양한 문제에 빠르게 적응할 수 있는 범용 대형 AI 모델
- 2021, 스탠포드 대학

기존 딥러닝 : 아기처럼 기본적인 것들 (시각, 촉각, 청각 등) 부터 배워야 함
파운데이션 모델 : 거대 모델 (뇌) + 대규모 데이터 학습 (많은 지식과 경험) 기반

- 새로운 일을 처음 접해도 금방 배우고 잘함

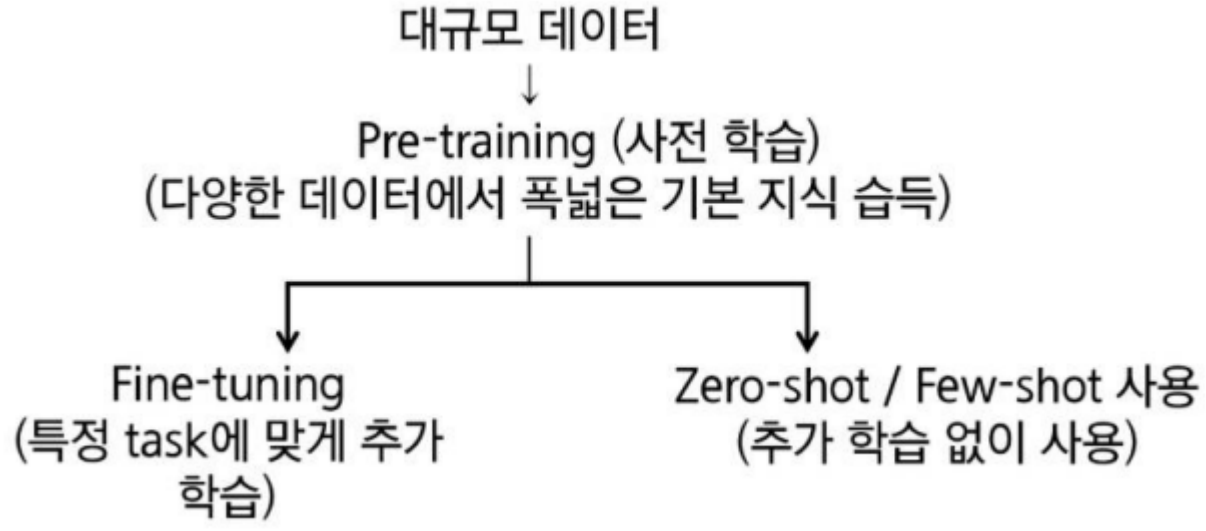
파운데이션 모델 특징

- 트랜스포머 모델 + 대규모 언어 학습이 주요한 범주를 구성

- 모달리티에 재한 없이 비슷한 패턴들이 등장
- 대개 비지도학습으로 훈련된 모델들이 주류
    - 쉬운 데이터 수집 + 대규모 학습
- 높은 Fine-tuning 성능 : 높은 태스크 적응 성능
- 한정되지 않은 출력 지원 : 만 개 이상의 물체 구분, 인식 가능

파운데이션 모델의 활용

- 추가적인 미세조정 (전이학습, 적응(Adaptation) 학습)
    - zero-shot : 예시 없이 처음보는 문제도 바로 적응
    - few-shot : 예제 몇 개 보여주면 바로 적응
    - Fine-tuning : 처음부터 배우지 않아도 조금만 알려주면 금방 적응



문제를 바라보는 방식의 변화

- 과거에는 모델을 새로 학습했지만, 이제 잘 학습된 모델을 얼마나 잘 활용하느냐가 핵심
- 파운데이션 모델 하나 확보하는 데 투여되는 계산 리소스는 대기업 외에 불가

- 내 문제는 이미 학습된 모델의 능력으로 해결할 수 있을까?
- 없다면, 처음부터 학습시켜야 할 정도의 문제일까? → Zero-shot
- 어떤 방식으로 조금만 튜닝해서 해결할 수 없을까? → Fine-tuning

세가지 접근 방식 비교

- Zero-shot
    - 예시 없이 질문만 던져서 문제 해결
    - 사전 학습된 모델이 배경 지식으로 대응
- Few-shot
    - 예시 몇개를 함께 제시하여 문제 해결 유도
    - 모델이 패턴을 스스로 감지하여 다음 입력에 적용
- Fine-tuning
    - 새 task에 맞춰 실제 추가 학습 진행
    - 모델 자체가 바뀜 (weight update)

2. 컴퓨터 비전 파운데이션 모델들

2-1. 영상 파이운데이션 모델이란?

영상 파운데이션 모델 개요

- 컴퓨터 비전(CV) 에서 방대한 데이터를 학습한 모델들
- 분할(Segmentaion), 탐지(Detection), 3D 및 깊이 예측(3D & Depth) 등 다양한 작업 수행 가능

2-2. 이미지와 텍스트 간의 관계 모델

CLIP (2021, OpenAI)

- 가장 유명함
- 언어와 이미지의 유사도 학습
- Vision Language Model의 눈으로 사용

2-3. 멀티모달 언어모델

이미지, 소리, 비디오 등 다양한 모달리티를 함께 이해하고 처리할 수 있는 언어 모델

- 대표적인 모델
  - LLAVA : Language 모델과 Vision 인코더 모델을 결합한 Multi-modal 모델
- 응용사례
  - 텍스트와 이미지를 결합한 대화형 AI, 이미지 설명 생성, 비디오 분석등 다양한 분야에서 사용

2-4. 도메인 특화 모델 - 의료

- 의료 이미지 (X-Ray, MRI, CT 등)을 입력 받아, 병적 진단 및 원인 설명 등의 태스크 수행
- Contrastive learning을 통해 학습
  - BiomedCLIP 모델 구조
- MedCLIP(2022)
  - 의료 텍스트와 이미지 임베딩을 정합시킨 의료용 CLIP 모델
  - 텍스트 입력으로부터 이미지 상의 질병을 탐지하거나 특정 종류의 의료 이미지를 검색하는 방식 등으로 활용 가능
- LLaVA-Med(2023)
  - LLaVA를 의료 데이터에 파인튜닝한 의료 특화 모델
  - 의료 이미지를 포함한 지시문 데이터(visual instruction-following data)를 통해 의료 이미지 기반 챗봇 대화가 가능한 멀티모달 모델

2-4. 도메인 특화 모델 - 제조업

- AnomalyGPT(2023)
  - 제조업 환경에서 발생하는 결함이나 불량을 탐지(anomaly detection)하기 위한 모델
  - 챗봇 형식으로 이미지 상 결함에 대해 텍스트로 질의응답을 주고 받을 수 있음
  - ImageBind의 이미지 인코더와 Vicuna를 언어 모델로 활용하여 제조업 데이터에 파인튜닝
  - 이미지 안에 결합이 있는 지 여부 판단

2-5. 3D 언어 모델

- 3차원 표현(point-cloud) 과 자연어의 관계를 학습한 파운데이션 모델

    - 3D LLM 모델 구조
    - 3차원 지도와 언어 간 관계 학습

2-6. 이미지 세그멘테이션 모델

- Segment Anything(SAM, 2023; SAM2, 2024, Meta)
    - 컴퓨터 비전에서도 방대한 양의 데이터로 파운데이션 모델을 만들 수 있음을 보여준 모델
    - 클릭 등의 유저 입력을 받아, 원하는 영역 마스크를 추출하는 고성능 분할 모델

2-6. 이미지 내 물체 탐지 모델

- Grounding DINO(2023, IDEA Research)
    - 텍스트 입력을 통해 이미지 내 물체를 탐지하는 모델
    - 방대한 데이터를 바탕으로 다양한 종류의 물체에 대해 높은 일반화 성능을 가짐
    - 객체 탐지 분야에서 파운데이션 모델이 높은 성능을 달성할 수 있음을 보여줌
    - 응용 : 이미지 검색, 탐지, 분류

2-6. 이미지 내 인스턴스 탐지 및 세그멘테이션 모델

- Grounded SAM(2024, IDEA Research)

2-6. 비디오 내 인스턴스 탐지 및 세그멘테이션 모델

- SAMURAI(2024, 워싱턴 대학)

2-7. 로봇 작업을 위한 텍스트-행동 변환 모델

- 입력 : 사람의 텍스트 명령 + 로봇 시점 영상
- 출력 : 로봇 행동 = { 위치변화, 관절 움직임 }

3. 영상 파운데이션 모델들

3-1. 영상 생성 파운데이션 모델들

- 이미지 생성 (Image Generation)
    - 대규모 이미지로 학습되어 텍스트 설명을 토대로 새로운 이미지 생성

3-2. Closed 이미지 생성 모델

- DALL E3 (OpenAI , 2023)
- Midjourney v7(Midjourney Inc, 2025)
- FLUX(2024, Black Forest Labs)
- Sora(OpenAI, 2024)

3-3. Closed 비디오 생성 모델

- Veo2 (Google Gemini)

3-4. Open Source 비디오 생성모델

- Hunyuan(2024, Tencent)

4. 3D 파운데이션 모델들

4-1. 3D 깊이(Depth) 추정 모델

- Depth Anything v2(HKU, TikTok 2024)

4-2. 노블-뷰 생성 모델

- Zero123XL(콜롬비아 대학, 2023)

4-3. 이미지 & 3D 동시 생성 모델

- JointDiT(Microsoft, POSTECH, 2025)

4-4. 단일 비디오 기반 Dynamic 3D 복원 모델

- MegaSaM(Google DeepMind, 2024)
- CUT3R(UC Berkeley, 2025)

4-5. 사람 전문 모델

- Sapiens(Meta, 2024)
- CLIP-Actor(POSTECH, 2024)

5. 오디오-비전 파운데이션 모델들

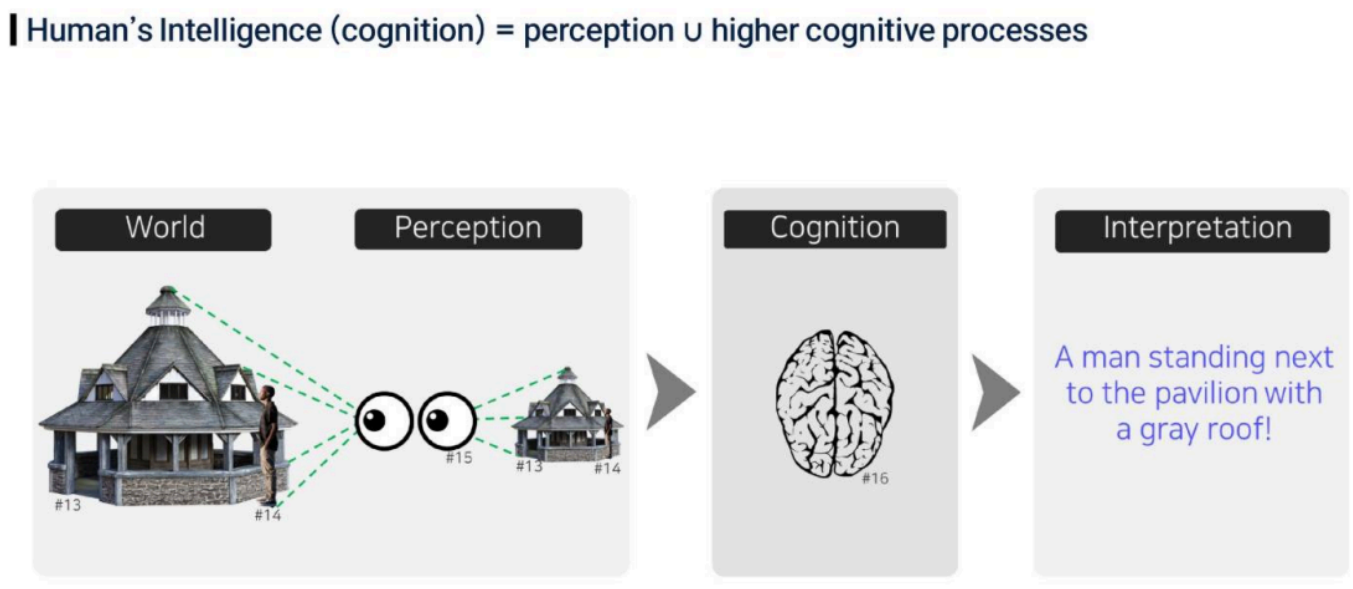
5-1. Audio-Vision 언어모델

- 대규모 언어 모델에 영상, 소리 입력을 확장해 멀티모달 언어모델로 확장 발전중
- ImageBind 기반 비디오 입력 : OneLLM(2024)
- 프레임 단위 비디오 입력 : VideoLLaMA2(2024)

- NExT-GPT : Any-to-Any Multimodal Large Language Model(2023)
- HeyGen’s Avatar IV
- Veo3 (Google)

2.

AGI를 향해서



- 사고 능력과 언어 능력만으로 현실 세계를 살기에 충분할까?

LLM에 눈을 달아볼까? (GPT-4)

- GPT-4(2023)
  - 자연어 입력에 국한된 기존의 거대 언어 모델에서 더 나아가 이미지, 문서, 음성 등 멀티모달 데이터를 처리할 수 있는 모델
  - GPT-4 API를 활용하여 다양한 도메인의 이미지 데이터와 결합한 모델이 개발됨

1. CLIP 모델

1-1. CLIP의 등장 배경

In-domain Generalization

- 기존 딥러닝에서 보편적으로 추구한 일반화 영역 : 학습한 데이터와 유사한 도메인에 대한 일반화 능력
- ex) 사진 데이터로 학습한 이미지 분류 모델이 스케치에서는 잘 동작할 거라 기대하지 않음

Zero-shot Generalization

- 새로운 데이터 처리 : 학습하지 않은 새로운 도메인의 데이터에서도 좋은 성능을 발휘
- 특징
  - 유연성 : 다양한 모멘이나 작업에 쉽게 적용
  - 확장성 : 데이터가 부족한 환경에서도, 학습 없이 성능 발휘

1-2. CLIP(2021, OpenAI)

대조 학습 기반(Contrastive Pre-training) 의 언어-이미지 사전 학습

- 자연어 감독(supervision)을 통해 시각적 개념 학습
- 다양한 이미지-자연어 쌍으로 학습
  - 인터넷에서 수집된 4억 개의 이미지-텍스트 쌍
  - 이미지 인코더 : ViT-B (or ResNet50)
  - 텍스트 인코더 : Transformer

1-3. CLIP 구조 - 텍스트 인코더 (Transformer 기반 Text Encoder)

CLIP 텍스트 인코더 트랜스포머 (2017, Transformer)

- 트랜스포머 구조

- Encoder Only 구조 (BERT 형태) 를 대표적으로 사용

CLIP의 입력 구성

- 토큰이라는 단위의 입력
- 입력된 토큰 간의 관계성을 집중하는 Attention 메커니즘으로 구성
- L 길이의 입력 토큰은 D-차원 특징 벡터(임베딩)의 배열 형태로 입력 (L x D)
- 자연어 데이터 : Sub-word 단위의 임베딩

1-4. CLIP 구조 - 이미지 인코더

CLIP의 입력 구성

- 영상 데이터 : 패치 단위의 임베딩

Self-Attention 만으로 구성한 이미지 인코더

- ViT : 비전 분야에 트랜스포머를 적용한 모델
- 전형적인 트랜스포머 구조를 이미지에 바로 적용
- 이미지를 패치 단위로 나눠서 처리
- 패치에 마스킹 한 후, 예측하는 방식으로 학습 진행

동작 방식

- 이미지를 작은 패치로 나눔
- 각 패치를 1D로 Flatten
- Positional Encoding 수행
  - 이미지 내에서 각 패치의 위치 정보를 추가
- Transformer encoder : 패치 처리
- MLP Head를 통해 분류 작업 수행
  - Head를 수정하여 다른 작업을 위한 transfer learning 활용 가능

CNN vs ViT

- 국소적인 특징을 추출하는 CNN과 달리, ViT는 이미지 전반적인 특징을 추출
- CNN 모델과 비교했을 때, 더좋은 성능을 보임 (e.g. EfficientNet, ResNet)
- Downstream task에 따라 전이 학습 가능

1-5. CLIP 구조 - 대조 학습

대조 학습(contrastive learning)

- 목표 이미지(앵커)를 일치하는 이미지(양성)로 끌어당기기
- 일치하지 않는 여러 이미지(음성)로부터 앵커를 밀어내기

CLIP의 Zero-shot Generation

- CLIP은 zero-shot 성능도 매우 뛰어남

1-6. 멀티 모달 정합

멀티 모달 정합(Multi-modal Matching)

- 서로 다른 두 가지 이상의 모달리티 간의 공통된 임베이딩 벡터 공간을 연결
  - 이미지와 텍스트
- 대표적인 모델
  - CLIP(OpenAI)
  - ImageBind(Meta)

CLIP은 softmax를 이용하여 이미지와 텍스트 임베딩 간의 코사인 유사도를 확인한다.

1-7. SigLIP

SigLIP은 softmax 대신 sigmoid 기반 손실함수 사용

- CLIP과 달리 일치하지 않은 음성 이미지에 제한된 영향만 받기 떄문에 안정적
- 학습 효율성이 뛰어나 적은 메모리 사용량으로도 높은 성능 달성
- CLIP 대비 압도적인 성능, 최신 VLM에 널리 활용

서로 다른 모달리티 간의 변환

- 모달리티 변환을 위한 두 가지 디자인
  - 변환 (Translating)
  - 정렬 (Matching)

1-8. 서로 다른 모달을 하나로 : Cross-modal Translation

정합(Matching)을 통한 크로스모달 변환

- 멀티모달 정합 손실 함수(Multi-modal alignment loss)

1-9. CLIP loss

CLIP의 loss 는 텍스트 - 이미지 간 정렬 정도 측정 (대조학습)이다.

- 정답일수록 가깝고, 오답일 수록 멀다.
- 이 방법을 다양한 멀티모달 모델에서 활용한다.
  - 이미지 캡셔닝
  - Text-to-image 등등

2. LLaVA 모델

2-1. 멀티모달 언어모델

이미지, 소리, 비디오 등 다양한 모달리티를 함께 이해하고 처리하는 언어 모델

대표적인 모델 : LLaVA(2023) : Language 모델과 Vision 인코더 모델을 결합한 비전-언어모델

2-3. LLaVA(2023)

특징

- 이미지 인식과 텍스트 생성을 결합, 이미지 설명 생성 또는 시각적 질문 응답 작업에서 뛰어난 성능
- 이미지 , 명령, 답변이 주어진 데이터셋 구축, Instruction tuning으로 학습

- 효율적인 메모리 사용 : 적은 자원으로 큰 모델을 학습
- 다중 모달 학습 : 텍스트, 시각 데이터 결합하여 응답 생성
- Fine-tuning : 특정 작업에 맞춰 모델 미세조정

Step1 : 사전 학습

- 표현 공유 : 이미지를 텍스트 표현을 변환하는 선형 레이어를 학습
- 효율적인 학습 : 적은 파라미터만 학습

Step2 : Fine-tuning

- 표현 공유 : 특정 작업에 맞춰 선형 레이어와 언어 모델등 필요한 부분 미세조정
- 효율적인 학습 : FP16과 같은 정밀도 최적화, 저비용 학습 기법으로 메모리 사용량 절감

LLaVA 학습 데이터

- 합성 데이터
- GPT를 활용하여 시각 설명 데이터 생성

3. 실습. 배포/서빙 실습

3-1. 허깅 페이스

AI 관련 오픈 소스 모델과 데이터셋을 공유하는 플랫폼

- 특징 : Pretrained Model 가중치 제공, 모델 학습을 위한 다양한 데이터 셋 제공
- 응용 분야 : 자연어 처리(NLP), 컴퓨터비전(CV), 음성인식 등 사용

허깅페이스의 응용

- 파운데이션 모델 실습
- Gradio : 머신러닝 모델을 웹 인터페이스로 쉽게 배포할 수 있게 도와주는 오픈 소스 라이브러리

모델 서빙

- 사용자에게 모델의 예측 결과를 전달하는 절차
- 주요 요소
  - 배포 : 학습된 모델을 서비스 가능한 상태로 변환하여 시스템에 설치 및 실행 유지
  - API 제공 : 모델에게 입력을 전달하고 실행할 수 있는 인터페이스 제공
  - 운영 보조 기능 : 확장 및 라우팅, 모니터링 등의 툴 제공

허깅페이스 inference API

- 별도의 서버 구축 없이 허깅 페이스 플랫폼에서 REST API만으로 모델을 바로 사용할 수 있는 서비스
- 연구 및 테스트 목적 : Request 횟수 제한 있음
- 모든 게시 모델이 지원하지 않음

3. 이미지 파운데이션 모델들

1. sVLM

1-1. OpenVLM

너무 무겁다

1-2. sVLM

다양한 온디바이스 모델

실경량화된 소형 VLM을 만들기 위한 시도들

1-3. SmolVLM

허깅페이스가 개발한 sVLM

1-4. Moondream 0.5B

모바일 기기나 엣지 디바이스에서의 실시간 실행을 염두에 두고 개발

손쉬운 사용법

1-5. Gemini Nano

온디바이스용 경량 Gemini

1-6. 갤럭시 온디바이스AI

모바일 NPU로 이미지, 언어, 오디오, 영상 작업을 기기 내에서 직접 생성형 AI를 구현

1-7. InternVL(2024)

GPT-4o에 맞서는 오픈소스 VLM

- OpenGBLam 에서 개발한 멀티모달 학습 라이브러리

1-8. LMDeploy

LLM의 효율적 압축, 배포, 서빙을 지원하는 오픈소스 툴킷

2. 한국어 sVLM

2-1. 언어별 구조적, 형태적 차이에 따른 토큰화 복잡성

언어별 토큰 길이 격차 (토큰 정보밀도 차이)

- 언어에 따라 동일한 문장이라도 토큰화 후 길이에 큰 차이를 보임
- 영어는 일부 언어보다 최대 2.5배 높은 토큰 정보밀도, 더 많은 내용 담을 수 있음
- 비영어권 언어는 컨텍스트 활용 효율 낮고, 토큰 낭비 발생

토크나이저 언어 편중 이슈

- 토크나이저들은 주로 빈도가 높은 언어에 맞춰 생성

형태소 풍부한 언어의 토큰화 (교착어, 굴절어 어려움)

- 핀란드어, 독일어처럼 복잡한 형태론을 지닌 언어는 서브워드 단위로 쪼개질 토큰 수 증가

2-2. HyperCLOVAX-SEED-Vision-Instruct-3B (Naver)

한국어 특화 멀티모달 모델

텍스트와 이미지를 동시에 이해하고 텍스트 생성

2-3. Open Source 언어모델 - 한국어 지원

Phi (Microsoft)

- 비용 효율 좋음, 온디바이스 AI 활용

HyperCLOVA X SEED (NAVER)

- 한국어에 우수한 성능

ExaONE(LG AI Research)

- 영어-한국어 이중 언어에 능통

SOLAR(Upstage)

- 영어-한국어 번역 능력 좋음

Gemma(Google)

4. 파운데이션 모델 응용

1. 파운데이션 모델 응용

1-1. 파운데이션 모델 + Fine-tuning

방대한 데이터로 학습한 파운데이션 모델을 최신 정보, 특정 작업/도메인 최적화를 위해 파인 튜닝을 함

1-2. AI 리터러시

차별화된 AI 종합 활용 능력

- AI의 작동 원리를 이해하고, AI가 생성한 정보를 비판적으로 분석하며, AI를 도구로서 효과적으로 활용할 수 있는 역량 + AI를 내 입맛 대로 변경해서 사용할 수 있는 능력

1-3. 파인튜닝이란?

미세조정 : 추가 학습을 통해 이미 학습된 모델을 조금만 튜닝하는 것

- 미세 조정을 통해 특정 작업에 특화된 모델을 개발할 수 있다.
- 파운데이션 모델 + 파인튜닝 ⇒ 실용적인 개인화 파운데이션 모델
  - 적은 데이터로 학습 가능
  - 학습 리소스 절약 가능
  - 특정 작업에 대한 우수한 성능
- 프롬프팅 보다 더 좋은 퀄리티와 결과물 생성
- 프롬프트의 길이가 줄면서 토큰 개수 절약

- 응답시간 감소

파인튜닝을 위해 학습률을 작은 비율로 반영해야 한다.

- 미세 조정으로 왜곡 적게

1-4. PEFT (Parameter-Efficient Fine-tuning)

AI 모델의 크기가 너무 커짐

→ 파운데이션 모델에 파인 튜닝을 해야 하는 데, 모델이 너무 커져서 파인 튜닝조차도 힘든 지경

→ PEFT의 필요성 증대

1. 프롬프트 디자인(prompt design)

- 언어 모델에서 주로 활용, 모델이 원하는 레벨의 결과를 출력할 수 있도록 입력 텍스트를 변형하는 방법
- 장점 : 추가 학습 없이 예측 성능 올림
- 단점 : 프롬프트 설계 필요, 성능 향상 제한적

2. 프롬프트 튜닝(prompt tuning, 2021) → 파인튜닝과 프롬프트 디자인의 중간

- 학습 가능한 프롬프트로, 가상 토큰(virtual token)을 입력에 추가
- 역전파를 통해 오직 가상 토큰에 대한 임베딩만 학습하고, 나머지 모델 고정
- 장점
  - 사람의 디자인 없이 모델 스스로 프롬프트 학습
  - 사전학습된 모델을 고정할 수 있음 (지식 손실 X)
  - 적은 비용으로 새로운 데이터셋의 모델을 학습 가능
- 단점
  - 학습된 프롬프트는 해석 X

3. Adaptor 모듈 추가 학습

- Activation을 변경하기 위해 작은 모듈을 추가하여 학습하는 기법

2. 합성 데이터

2-1. 지식증류(Knowledge Distillation)

사전 학습된 고성능 모델의 지식을 작은 모델에 압축해서 빠르고 효율적으로 만들기 위해 도입

높은 성능의 모델(선생님)을 모방하도록 가벼운 모델(학생)을 학습하는 방법

- 선생님 모델이 예측한 soft-label 값과 학생 모델의 예측 값이 가까워지도록 학습 유도
  - soft-label : [0,1] 사이의 모델의 예측을 가짜 라벨(정답)으로 사용

2-2. 텍스트 기반 이미지 편집 모델 학습용 데이터

InstructPix2Pix(2023) : 이미지 편집 지시사항을 수행할 수 있도록 학습

- 이미지와 지시사항이 주어지면, 모델은 적절한 편집을 수행

방법 1.

- 지시사항 기반 이미지 편집을 지도 학습 문제로 전환
- {이미지편집에 대한 지시사항, 편집 전 이미지, 편집 후 이미지} 형식의 학습 데이터셋 생성

방법2.

- 생성된 텍스트 데이터셋을 기반으로 별도의 이미지 편집 생성 모델로 영상 데이터 쌍 생성

방법3.

- 생성된 이미지-명령 쌍 데이터셋 기반으로, 최종 이미지편집 생성 모델을 학습(파인튜닝)

2-4 합성데이터 활용

합성 데이터 : 실제 데이터를 모방하거나 새로 생성한 인공 데이터

- 데이터 수집하거나 사용하기 어려운 경우 대체
- 데이터 부족 문제 해결, 모델 성능 개선 사용