파운데이션 모델 및 멀티모달 AI 총 요약

1. 파운데이션 모델(Foundation Model) 개요

정의 및 핵심 개념

파운데이션 모델이란? 대규모 데이터를 폭넓게 학습한 후, 다양한 문제에 빠르게 적용할 수 있는 범용 대형 AI 모델

등장 배경

- 스탠포드 대학에서 2021년 새로운 AI 범주로 제시
- 전통적 방식(DB 저장 후 검색)의 한계를 극복
- 데이터를 패턴화하여 압축(Neural Networks)하여 새로운 데이터에 일반화 가능

패러다임 전환

기존 방식 vs 파운데이션 모델

- 기존: 작업마다 모델을 처음부터 새롭게 학습
- 현재: 거대 모델 + 대규모 데이터 사전학습 → 새로운 작업에 빠르게 적응

개발 프로세스 Data Creation → Data Curation → Training → Adaptation → Deployment

주요 특징

1. **대규모**: 트랜스포머 모델 + 대규모 언어 데이터, 비지도학습
2. **적응성**: 높은 파인튜닝 성능으로 다양한 태스크에 쉽게 적응
3. **범용성**: 기존 20개 → 수만 개 물체 인식 가능

활용 방법

- **Zero-shot**: 처음 보는 문제를 추가 학습 없이 바로 적용 → 추가 학습 x
- **Few-shot**: 처음 보는 문제를 예제 몇 개만으로 적응 → 추가 학습 x
- **Fine-tuning**: 소량 데이터로 특정 작업에 최적화 → 추가 학습 o

---

2. AGI를 향한 멀티모달 AI

AI 발전 단계

2022년 11월 이전

- 각 분야별로 부분적 능력만을 개별적으로 모델링

2022년 11월 (ChatGPT) 이후

- 대규모 언어모델(LLM)이 높은 사고/추론 성능 달성
- 인간 수준 근접

핵심 질문 사고 능력과 언어 능력만으로 현실 세계를 이해하기에 충분할까? → 시각언어모델의 필요성

GPT-4 - 멀티모달 언어모델

- 이미지, 문서, 음성 등 멀티모달 데이터 처리
- Generic Anomaly Detection 등 다양한 응용

---

3. CLIP - 이미지-언어 파운데이션 모델

CLIP 개념

Contrastive Language-Image Pre-training (2021, OpenAI)

- 언어와 이미지의 유사도 학습
- 인터넷에서 수집된 수억 개의 {텍스트, 이미지} 쌍 데이터 활용

구조

텍스트 인코더 (Transformer)

- Sub-word 단위 토큰 임베딩
- Attention 메커니즘으로 관계성 포착

이미지 인코더 (ViT: Vision Transformer)

- 이미지를 패치(16x16x3)로 분할
- Learnable position embedding
- Transformer encoder로 처리

학습 방법

대조 학습(Contrastive Learning)

- 목표 이미지(앵커)를 대응 텍스트(양성)와 가깝게
- 일치하지 않는 텍스트(음성)와는 멀게
- Softmax 기반 Image↔Text Loss

응용 - 제로샷 이미지 인식

- 텍스트로 카테고리 리스트 준비
- 텍스트 임베딩을 Vector DB로 변환
- 쿼리 이미지와 비교하여 가장 높은 점수의 카테고리 반환

---

4. SigLIP - CLIP의 개선

CLIP의 한계

어느 정도 멀게 배치한 음성 데이터에도 계속 거리를 벌리기 위해 학습 진행

SigLIP의 개선점

- Softmax 대신 Sigmoid 기반 손실함수 사용
- 어느 정도 거리가 멀어지면 고려 안함
- 노이즈에 강건한 성능 제공

성능

학습 데이터에 노이즈가 많을 때 CLIP보다 압도적 성능

---

5. 멀티모달 정합(Multi-modal Alignment)

개념

서로 다른 두 가지 이상의 모달리티 간의 공통된 임베딩 벡터 공간 구성

→ 서로 다른 모달리티 임베딩 간 유사도 비교 가능

주요 모델

CLIP (OpenAI)

- 이미지와 텍스트 간 Multi-modal Alignment

ImageBind (Meta)

- 소리, 텍스트, 이미지, 열화상, 깊이맵 등 6가지 모달리티 통합

활용 사례

1. Cross-Modal Retrieval: 오디오 → 이미지/비디오/텍스트 검색
2. Embedding-Space Arithmetic: 모달리티 간 연산
3. Audio to Image Generation

4. Audio to Segmentation

---

6. 멀티모달 언어 모델 (VLM)

VLM 개념

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

작동 방식 대충 개 사진 → Image Encoder → Linear Projection → Text Decoder → 이 사진은 개가 찍혀 있습니다.

LLaVA (2023)

특징

- Vision과 Language 모델 결합
- 효율적인 2단계 학습

학습 과정

1. Pre-training: 이미지의 텍스트 표현을 변환하는 Projection layer만 학습 (자원 절약)
2. Fine-tuning: 특정 task에 맞게 추가 학습

학습 데이터 ChatGPT를 활용한 합성 데이터 생성 (대화, 상세 설명, 복잡한 추론)

Qwen-VL 시리즈 (Alibaba)

Qwen-VL

- 여러 이미지 입력, 번역, OCR, 객체 인식
- 3단계 학습: Pretraining → Multi-task → Finetuning

Qwen2-VL

- 다국어, 임의 해상도 처리
- Agent, Code/math, Video, Live chat

Qwen2.5-VL

- 강력한 문서 파싱 (OCR, 테이블, 차트, 악보)
- 정밀한 객체 그라운딩
- 장시간 비디오 이해

Qwen2.5-Omni

- 읽고, 쓰고, 보고, 듣고, 말하는 모든 모달리티 통합

InternVL (2024)

특징

- OpenGBLab의 오픈소스 모델
- 6B 파라미터 Vision encoder
- 점진적으로 LLM을 키워가며 학습

---

7. VLM 기반 모델 생태계

CLIP 기반 VLM

- BLIP-2: CLIP + OPT/FlanT5
- InstructBLIP: BLIP-2의 instruction tuning 버전
- LLaVA: CLIP + Vicuna
- MiniGPT-4, mPLUG-Owl

SigLIP 기반 VLM

- PaLI-X: SigLIP + PaLM
- SmolVLM

최근 트렌드

CLIP, SigLIP의 성공적인 레시피를 기반으로 특화 Vision encoder 개발 중

---

8. VLM 성능 향상 기법

Set of Mark (SoM)

개념

- 이미지에 번호를 매겨 객체를 명확히 식별
- 다른 물체 탐지, 세그멘테이션 파운데이션 모델 활용

효과

- VLM의 부족한 시각 능력 보완
- Computer Agent 모델에 매우 유용
- 객체 위치 정확도 대폭 향상

핵심 정리

파운데이션 모델의 혁신

- 대규모 사전학습 → 다양한 작업에 빠른 적응
- Zero-shot, Few-shot, Fine-tuning으로 효율적 활용

멀티모달 AI의 발전

- LLM → VLM: 언어에서 시각까지 확장
- CLIP/SigLIP: 이미지-언어 정합의 기초
- 다양한 모달리티 통합 (ImageBind, Omni 등)

주요 VLM 모델

- LLaVA: 효율적 학습 방법
- Qwen-VL: 문서, 비디오, 다국어 특화
- InternVL: 대규모 오픈소스

미래 방향

- 더 많은 모달리티 통합
- 실시간 처리 능력 향상
- Agent 응용 확대
- 6장 0~76p 문제

파운데이션 모델 & 멀티모달 AI 핵심 문제 10개

문제 1. 파운데이션 모델의 정의

Q: 파운데이션 모델(Foundation Model)이란 무엇이며, 기존 AI 모델 개발 방식과 어떤 차이가 있나요?

정답:

- 정의: 대규모 데이터를 폭넓게 학습한 후, 다양한 문제에 빠르게 적용할 수 있는 범용 대형 AI 모델
- 차이점:
  - 기존: 작업마다 모델을 처음부터 새롭게 학습
  - 파운데이션 모델: 거대 모델 + 대규모 데이터 사전학습 → 새로운 작업에 빠르게 적응

문제 2. 파운데이션 모델의 세 가지 특징

Q: 파운데이션 모델의 3가지 주요 특징(대규모, 적응성, 범용성)을 각각 설명하세요.

정답:

1. **대규모**: 트랜스포머 모델 + 대규모 언어 데이터, 비지도학습으로 훈련
2. **적응성**: 높은 파인튜닝 성능으로 다양한 태스크에 쉽게 적응 가능
3. **범용성**: 다양한 작업 수행 가능 (예: 기존 20개 물체 인식 → 파운데이션 모델은 수만 개 인식)

문제 3. Zero-shot, Few-shot, Fine-tuning 비교

Q: Zero-shot, Few-shot, Fine-tuning의 차이점을 설명하고, 각각 언제 사용하는지 서술하세요.

정답:

- **Zero-shot**: 추가 학습 없이 바로 적용 (학습 데이터가 전혀 없을 때)
- **Few-shot**: 예제 몇 개만으로 적응 (소량의 예시만 있을 때)
- **Fine-tuning**: 소량 데이터로 특정 작업에 최적화 (특정 태스크에 맞춤 성능이 필요할 때)

문제 4. CLIP의 구조와 학습 방법

Q: CLIP(Contrastive Language-Image Pre-training)의 주요 구조(텍스트 인코더, 이미지 인코더)와 대조 학습(Contrastive Learning) 방법을 설명하세요.

정답:구조:

- 텍스트 인코더: Transformer 기반, Sub-word 단위 토큰 임베딩
- 이미지 인코더: ViT(Vision Transformer), 이미지를 패치(16x16x3)로 분할

대조 학습:

- 목표 이미지(앵커)를 대응하는 텍스트(양성)와 가깝게 배치
- 일치하지 않는 여러 텍스트(음성)와는 멀게 배치
- Softmax 기반 Image↔Text Loss 계산

문제 5. SigLIP vs CLIP

Q: SigLIP이 CLIP의 어떤 한계를 개선했으며, 어떤 방식으로 개선했나요?

정답:CLIP의 한계:

- 어느 정도 이미 멀게 배치한 음성 데이터에도 계속 거리를 벌리기 위해 학습이 진행됨

SigLIP의 개선:

- Softmax 대신 Sigmoid 기반 손실함수 사용
- 어느 정도 거리가 멀어지면 고려하지 않음
- 결과: 노이즈에 강건한 성능, 학습 데이터에 노이즈가 많을 때 압도적 성능

문제 6. VLM의 작동 방식

Q: 멀티모달 언어 모델(VLM)의 기본 작동 방식을 단계별로 설명하세요.

정답:작동 방식:

1. **Image Encoder**: 이미지를 특징 벡터로 변환
2. **Linear Projection**: 이미지 특징을 언어 모델이 이해할 수 있는 토큰으로 변환 (Soft prompts)
3. **Text Decoder**: 변환된 이미지 토큰과 텍스트 입력을 결합하여 응답 생성

예: 이미지 + "이 이미지를 설명해줘" → "빨간 스카프를 두르고 있는 고양이 사진이에요"

문제 7. LLaVA의 학습 과정

Q: LLaVA의 2단계 학습 과정(Pre-training, Fine-tuning)을 설명하고, 각 단계에서 무엇을 학습하는지 서술하세요.

정답:Step 1: Pre-training (사전 학습)

- Projection layer만 학습 (Vision Encoder와 Language Model은 Frozen)
- 이미지를 텍스트 표현으로 변환하는 선형 레이어 학습
- 자원과 시간 절약

Step 2: Fine-tuning

- 특정 작업에 맞춰 Projection layer 또는 Language Model을 미세 조정
- 적은 메모리로 큰 모델 학습 가능 (FFP16 등 최적화 기법 활용)
- ChatGPT로 생성한 합성 데이터 사용 (대화, 상세 설명, 복잡한 추론)

문제 8. Qwen-VL 시리즈의 발전

Q: Qwen-VL → Qwen2-VL → Qwen2.5-VL → Qwen2.5-Omni로 발전하면서 추가된 주요 기능들을 설명하세요.

정답:Qwen-VL:

- 여러 이미지 입력, 번역, OCR, 객체 인식
- 3단계 학습 파이프라인

Qwen2-VL:

- 다국어 텍스트 및 이미지 내 텍스트 이해
- 임의의 이미지 해상도 처리
- Agent 응용, Code/math 추론, Video 분석, Live chat

Qwen2.5-VL:

- 강력한 문서 파싱 (OCR, 테이블, 차트, 악보)
- 정밀한 객체 그라운딩
- 장시간 비디오 이해

Qwen2.5-Omni:

- 읽고, 쓰고, 보고, 듣고, 말하는 모든 모달리티 통합 처리
- Vision Encoder + Audio Encoder 통합

문제 9. 멀티모달 정합(Multi-modal Alignment)

Q: 멀티모달 정합(Multi-modal Alignment)이란 무엇이며, CLIP과 ImageBind를 비교 설명하세요.

정답:개념: 서로 다른 두 가지 이상의 모달리티(예: 이미지와 텍스트) 간의 공통된 임베딩 벡터 공간을 구성하는 것

비교:

- **CLIP (OpenAI)**: 이미지와 텍스트 간 Multi-modal Alignment (2개 모달리티)
- **ImageBind (Meta)**: 소리, 텍스트, 이미지, 열화상, 깊이맵 등 6가지 모달리티 통합

활용: Cross-Modal Retrieval, Embedding-Space Arithmetic, Audio to Image Generation 등

문제 10. Set of Mark (SoM)

Q: Set of Mark(SoM)이란 무엇이며, VLM 성능 향상에 어떻게 기여하나요?

정답:개념:

- 이미지에 번호를 매겨 객체를 명확히 식별하는 기법
- 다른 물체 탐지, 세그멘테이션 파운데이션 모델 활용

효과:

- VLM의 부족한 시각 능력 보완
- 객체 위치 파악 정확도 대폭 향상
- Computer Agent 모델에 매우 유용 (예: "12번 램프 옆에 앉으세요" 같은 정확한 지시 가능)

예시:

- 입력: 이미지만 → "컵이나 머그가 있어요" (부정확)
- 입력: 이미지 + SoM → "12번으로 표시된 램프가 있어요" (정확)

[심화] CLIP 응용: 멀티모달 정합 손실함수로 활용

1. 멀티모달 정합 (Multi-modal Alignment)

- 정의: 이미지, 텍스트 등 서로 다른 양식(Modality)의 정보를 공통 임베딩 공간(Embedding Space)에 매핑(mapping)하는 것.
- 목적: 공통 공간 내에서 데이터 간의 의미적 유사도를 측정하기 위함.
- 정합의 2가지 방식
  - 변환 (Translating)
    - 모달리티 A ➡ 모달리티 B로 직접 변환
  - 정합 (Matching) (CLIP 방식)
    - 모달리티 A와 B를 각각 인코딩 ➡ **정렬된 피처 공간 (Aligned Feature Space)**으로 매핑

---

2. 멀티모달 정합 손실 함수 (Multi-modal Alignment Loss)

정의: 멀티모달 데이터(예: 이미지-텍스트 쌍)가 공통의 임베딩 공간에서 얼마나 잘 정렬되었는지(얼마나 가까운지) 그 정도를 측정하는 함수.

- CLIP Loss (Radford21):
  - 방식: 대조 학습 (Contrastive Learning)
  - 원리: 행렬에서, 올바른 (이미지, 텍스트) 쌍의 유사도는 높이고, 잘못된 쌍의 유사도는 낮추도록 학습.
- SDS Loss (Poole23):
  - 방식: Score Distillation Sampling
  - 원리: 사전학습된 Text-to-Image Diffusion Model을 손실 함수처럼 활용.

---

3. CLIP loss 예시

** 해당 부분은 그림을 보면서 이해야 할 것 같음. 해당 교안을 확인하는게 좋을 듯합니다.**

** p83~86 **

- 단일 데이터에 대해서만 학습 = 최적화(학습X) : 일반적인 모델 학습이 아니라 단일 입력에 대한 최적화 과정이다.
- 텍스트-이미지 간 정렬 정도 측정(손실함수로 사용됨) → 유사도 최대화

CLIP의 '역(逆)활용': 손실 함수로 사용하기

사전학습된(pre-trained) CLIP 모델을 고정시키고, 이를 손실 함수로 삼아 입력을 최적화(Optimization)함.

동작 원리 (Backpropagation 활용)

1. 목표 설정: 특정 텍스트 프롬프트 (예: "Pepper the aussie pup")를 Text Encoder에 입력 ➡ 목표 임베딩(T_1) 획득.
2. 입력 정의: 최적화할 대상(예: 3D 모델, 이미지, 텍스트)을 **'학습 가능한 파라미터'**로 설정.
3. 유사도 계산: 이 파라미터에서 생성된 결과물(예: 렌더링된 이미지)을 Image Encoder에 입력 ➡ 결과 임베딩(I_1) 획득.
4. 손실 계산: T_1과 I_1 간의 **유사도(Loss)**를 측정.
5. 역전파 (Backprop): 이 Loss를 '학습 가능한 파라미터' 쪽으로 역전파하여 파라미터를 업데이트.
6. 결과: T_1과 I_1의 유사도가 최대화되는 방향으로 '학습 가능한 파라미터'가 최적화됨.

! 중요: "단일 데이터에 대해서만 학습 = 최적화 (학습X)".

이는 일반적인 모델 '학습(Training)'이 아니라, 단일 입력에 대한 '최적화(Optimization)' 과정임.

---

4. 주요 응용 사례

텍스트 프롬프트 기반 생성 및 편집 예시

1. Image-to-Text (이미지 캡셔닝) - 예: ZeroCLIP

- 유형: 이미지 캡셔닝
- 입력/출력: (입력: 이미지) / (출력: 캡션)
- 특징: 구체적인 텍스트 생성, 제한적인 OCR(글자 인식) 가능. (예: "The Simpsons.", "Stanford University.")

2. Text-to-Image (이미지 스타일 변환) - 예: StyleCLIP

- 유형: 이미지 편집
- 입력/출력: (입력: 텍스트 명령, 원본 이미지) / (출력: 편집된 이미지)
- 원리: "Mohawk hairstyle", "Gothic church", "Without makeup" 같은 텍스트 명령에 맞게 원본 이미지를 편집.

3. Text-to-Motion (3D 애니메이션 생성) - 예: CLIP-Actor

- 유형: 3D 애니메이션
- 입력/출력: (입력: 텍스트 설명) / (출력: 3D 애니메이팅 아바타)
- 원리: "Captain America is jump kicking" 텍스트에 맞춰 3D 아바타의 동작(Motion)을 생성.

Small VLM과 파운데이션 모델들 소개

1-1. OpenVLM 리더보드

- `InternVL`, `Qwen2.5-VL` 등 상위권 VLM 모델들은 파라미터가 수십 B (Billion) 단위로, 성능은 좋지만 너무 무겁다는 한계 있음 (예: 78.4B, 38.4B)

1-2.

- sLLM(소형 언어모델)의 경량화 시도가 VLM(비전 언어모델)에도 이어지며, 다양한 온디바이스 모델 실경량화된 소형 VLM을 만들기 위한 시도가 이뤄지고 있음

주요 소형 VLM 및 온디바이스 AI 모델

- **1-3. SmolVLM**
  - Huggingface가 개발한 소형 VLM
  - `SigLIP` (Vision Encoder)와 `SmolLM2 1.7B` (LLM)를 결합하여 모델 크기를 줄임 (Llama 3.1 8B → SmolLM2 1.7B)
- **1-4. Moondream 0.5B**
  - 모바일/엣지 디바이스에서의 실시간 실행을 목표로 개발
  - 크기: 20억 개 파라미터
    - 4비트 양자화(int4) 시: 다운로드 375MB, 실행 메모리 816MB 수준으로 매우 가벼움
  - 제공 기능: 이미지 캡셔닝, VQA(시각적 질의응답), 객체 탐지, 좌표 지정(Pointing), 시선 감지, OCR 및 문서 이해
  - 사용법: `pip install moondream` 으로 쉽게 설치 및 사용 가능
- **1-5. Gemini Nano**
  - Google의 온디바이스용 경량 Gemini 모델
  - 크기: Nano-1 (18억 개), Nano-2 (32.5억 개) 파라미터
  - 적용: 2024년 픽셀 9 시리즈에 탑재
  - 기능: 녹음 앱에서 실시간 요약, 카메라로 보이는 자료를 AI가 이해하고 관련 텍스트 생성
- **1-6. 갤럭시 온디바이스 AI**
  - 모바일 NPU(신경망 처리 장치)를 활용해 기기 내부에서 생성형 AI 실행
  - 기능: 텍스트 기반 이미지 생성, 인페인팅/아웃페인팅, 어조 변환 및 문법 교정, 자연어 기반 사진 촬영 (예: "스케이트보드 탈 때 찍어줘")

1-7. VLM 배포 및 최적화 도구: LMDeploy

- LMDeploy = LLM/VLM의 효율적인 압축(양자화), 배포, 서빙을 지원하는 오픈소스 툴킷
- 주요 기능:
  - 효율적 추론: 지속적 배치(Persistent batching), 고성능 CUDA 커널 등으로 최대 1.8배 빠른 처리량
  - 효과적 양자화: 4비트 양자화 등을 통해 2.4배 빠른 추론 속도 달성
  - 대화형 추론: KVCache를 통해 대화 이력 재처리를 줄여 효율적인 대화형 추론 지원
  - 높은 호환성: 다양한 VLM (LLaVA, Qwen-VL, DeepSeek-VL 등) 및 LLM (Phi-3, Llama3 등)을 지원

1-8. 기타 sVLM

- 최신 sVLM
  - Qwen 2.5-VL 3B
    - 다양한 크기의 이미지와 장시간 영상 처리 가능
    - 표, 약보, 화학식 등 다양한 형태의 데이터 및 JSON 형식 등 처리 가능
- Phi-3.5 vision instruct 4.2B
  - OCR, 차트 분석, 비디오 용약 등에 특화된 경량 멀티모달 모델
- DeepSeek-VL2 1B
  - 중국 AI 스타트업으로 저비용 오픈소스 LLM 및 VLM 개발
- Gemma 3 1B
  - Google의 멀티모달 오픈 소스 모델
  - IB ~27B의 다양한 크기의 모델 제공

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
  - 영어 중심 토크나이저 • 영어는 일부 언어보다 최대 2.5배 높은 토큰 정보 밀도를 보여, 같은 토큰 길이에 더 많은 내용을 담을 수 있음 • 비영어권 언어는 컨텍스트 활용 효율이 낮고 토큰 낭비가 발생하는 "구조적" 불이익 존재 (주의: 언어 자체의 한계가 아님. 토큰화 방법의 효율성 차이)
- **토크나이저의 언어 편중 이슈**
  - 토크나이저의 언어 편중 이슈
    - 주로 빈도가 높은 표현에 대한 설계되어, 사용 빈도가 적거나 형태가 설계 언어와 다른 언어는 비효율적으로 긴 토큰 시퀀스가 생성
  - 형태소가 복잡한 언어의 토큰화
    - 핀란드어, 독일어의 경우, 하나의 단어가 매우 길거나 여러 의미를 접합해 표현하므로, 서브워드 단위로 쪼개지는 토큰 수가 크게 증가
- **한국어 토큰화 시도 (예시)**
  - **문장:** "이번 방학 때 뭐해?"
  - **Base 토크나이저:** 19개 토큰 (비효율적)
  - **Extended 토크나이저:** 8개 토큰 (효율적)
  - → 한국어에 최적화된 토크나이저가 필요함을 시사함

2-2. 한국어 sVLM (소형 비전언어모델) 모델

- **모델명: HyperCLOVAX-SEED-Vision-Instruct-3B**
- NAVER가 개발한 한국어 특화 멀티모달 모델로, 텍스트와 이미지를 동시에 이해하여 텍스트를 생성
- **사용 방법 (Huggingface):**
  1. Huggingface 로그인 및 모델 다운로드

2. 커스텀 모델 등록 및 불러오기
3. 대화 데이터 구성 및 입력 생성
4. 텍스트 생성

다른 이미지 파운데이션 모델들 소개

2-1. 이미지 파운데이션 모델

- 영상 파운데이션 모델 개요
  - 컴퓨터 비전(CV)에서 방대한 데이터를 학습한 모델들
  - 분할(Segmentation), 탐지(Detection), 3D 및 깊이 예측(3D & Depth) 등 다양한 작업 수행 가능
  - 예시: 이미지-텍스트 유사도 측정, 비전-언어 모델, 멀티 모달, 3D LLM 모델 구조

2-2. 이미지 세그멘테이션 모델

- Segment Anything(SAM, 2023; SAM2, 2024) - Meta
  - 컴퓨터 비전에서도 방대한 양의 데이터로 Foundation model을 만들 수 있음을 보여준 모델
  - 클릭, 박스, 부분 세그먼트, 텍스트 등의 유저의 입력을 받아, 원하는 영역 마스크를 추출하는 고성능 분할 모델
  - 약 1,100만개의 이미지(약 10억 개의 마스크)로 학습

2-2. 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(generalization) 성능을 가짐
  - 객체 탐지 분야에서 Foundation 모델이 높은 성능을 달성할 수 있음을 보여준 모델
  - 응용 : 이미지 검색, 탐지, 분류 작업에 사용

2-2. 이미지 내 인스턴스 탐지 및 세그멘테이션 모델

- Grounded SAM (2024, IDEA Research)
  - Grounding DINO와 SAM 결합 모델
  - 텍스트 입력으로부터 객체 탐지 뿐만 아니라 분할까지 동시에 수행할 수 있는 모델
  - Grounding DINO를 통해 추출된 박스를 SAM의 입력으로 활용하여 개별 물체 분할

2-2. 비디오 내 인스턴스 탐지 및 세그멘테이션 모델

- SAMURAI(2024) - Univ. Washington
  - 비주얼 물체 트랙킹 State-of-the-art (Updated at 2025. 05)
  - SAM 2 기반 응용 모델

    - 응용 사례: 비디오 편집, 물체 지우기와 결합, 이상행동 감지, CCTV 자동 분석, 스포츠 중계

2-3. 영상 생성 파운데이션 모델들

- 이미지 생성 : 대규모 이미지로 학습되어 텍스트 설명을 토대로 새로운 이미지를 생성
- Closed Source(상용)와 Open Source(오픈소스)

A. Closed (상용) 모델

- DALL·E 3 (OpenAI, 2023.10)
    - ChatGPT에 통합되어 대화를 통해 프롬프트를 개선하고 이미지를 생성 가능
    - 복잡한 프롬프트도 정교하게 반영하며, 텍스트 인식 및 생성 능력이 향상
    - 예시: "지브리 스타일", "스티커 이모지" 등 특정 스타일 생성
- Midjourney v7 (Midjourney Inc., 2025.04 알파)
    - 긴 프롬프트 이해도와 세밀한 스타일 조절(예: 색상, 음영) 능력이 뛰어남
    - 손과 신체 표현의 일관성이 크게 향상되었고, 텍스트가 포함된 이미지 생성도 지원함

B. Open Source 모델

- Stable Diffusion 3 / 3.5 (Stability AI, 2024.02 / 2024.10)
    - Diffusion Transformer 아키텍처를 도입하여 이미지 품질과 다중 객체 프롬프트 처리 능력을 개선함
    - 이미지 내 텍스트 및 글자 표현 정확도가 크게 향상됨
    - 다양한 모델 크기(800M ~ 8B 파라미터)를 제공하여 성능과 자원 요구 사항 간의 균형 능
- FLUX (Flux.1) (Black Forest Labs, 2024.08)
    - Rectified Flow Transformer 기반의 120억 파라미터 최신 모델
    - 복잡한 장면 생성에 강점을 보이며, 선명한 글자나 숫자 표현 능력 탁월
    - 세 가지 모델 변형
        - Pro: 최고 성능 (API 유료)
        - Dev: 오픈소스 (연구용/비상업용)
        - Schnell: 경량, 고속 (오픈소스/상업용 허용)

---

p140

ControlNet(2023)

- 컨트롤 조건 입력을 기반으로 사용자가 원하는 이미지를 생성해주는 이미지 생성 모델
- 커뮤니티 중심으로 매우 활발하게 응용되고 있음

노블 -뷰 생성 모델 - Zero123XL(콜롬비아 대학:2023) : 2D에서 3D로 변환

- 2D 이미지를 입력으로 받아 해당 물체를 특정 위치의 카메라 뷰로 바라보았을 때의 모습을 생성하는 모델
- 추가로 2D 이미지 입력만으로 해당 물체의 3D 전체 모습을 재현할 수 있음
- 응용 : 3D 모델링, 가상현실(VR) 및 증강현실(AR) 콘텐츠 생성에 사용

정교환 #D Depth map 추정 - Marigold

- 단안 깊이 추정을 위해 이미지 생성 Diffusion 모델을 합성데이터에 파인튜닝

이미지 & 3D 동시에 생성 모델 - JointDit

- 기능 : 이미지와 3D 깊이 맵 동시 생성, 입력 이미지의 3D 추정, 3D 기반 이미지 생성 등 다양한 기능 지원
- 물리적으로 더욱 그럴듯한 장면 생성

Depth Anythin v2

- SAM 이후 연구된 많은 vision foundation 모델 중 깊이맵 예측을 위한 모델
- SAM과 마찬가지로 약 150만 개의 방대한 데이터로 학습됨
- 이때 약 6,200만개의 라벨링 되지 않은 데이터를 추가로 활영하여 성능 극대화
- 응용 : 자율 주행, 로봇 비전, 3D 복원 등 다양한 작업에서 사용

사람 중심 모델 - Sapiens

- 인간 형태 인식을 위해 3000만 개의 이미지로 학습된 파운데이션 모델
- 사람 중심 태스크들

Sora

- 텍스트 → 비디오 생성 및 이미지 / 동영상 → 비디오 확장 모두 지원
- ChagGPT와 통합된 대화형 편집-사용자 피드백으로 반복 개선 가능
- 물리적인 이해를 보여줌

Veo 2

- 8초 길이, 720p 해상도의 와이드스크린 비디오 생성
- 생성 영상에는 워터마크가 포함되어 합성 비디오 식별 가능
- TikTok, YouTube 등에 바로 공유 가능 편의성 제공

Veo 3

- 자연스럽게 싱크된 소리까지 같이 생성

HeyGen’s Avatar IV

- 입력(텍스트 스크립트, 목소리 샘플, 사진 한 장) → 출력(스피킹 비디오)
- 기술스택

Wan 2.2

- 지원 모드 : text -to -video, text & image-to-video, sound-to-sound, and image-to-video
- 시네마 퀄리티

MegaSaM

- 단안 카메라 동영상에서의 정확한 카메라 포즈 및 깊이 추정

CUT3R

- 가상 시점에서의 미관측 영역 추론
- 비디오 스트림이나 순서가 없는 사진 모음과 같은 다양한 길이의 이미지를 자연스럽게 처리

Audio - Vision Language Models

- 대규모 언어모델에 영상, 소리 입력을 확장해 멀티모달 언어모델로 확장 발전 중

파운데이션 모델 + Fine-tuning

파운데이션 모델과 미세조정이 필요한 이유

- 방대한 데이터로 학습된 초대형 딥러닝 모델. 다양한 작업이나 범용적인 문제에 바로 적용 가능
- 최근에는 텍스트 뿐만 아니라 이미지, 오디오, 비디오 등의 다양한 입력 데이터를 처리할 수 있는 멀티모달로 확장
- 하지만 최신 정보나 특정한 작업/도메인에 최적화 되어 있지 않아, 즉시 활용이 어려운 경우가 반드시 있음

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 입맛대로 변경해서 사용할 수 있는 능력

미세조정 : 추가 학습을 통해 이미 학습된 모델에 조금만 튜닝하는 것

- 미세조정을 통해 특정 작업에 특화된 모델을 개발할 수 있다.
- 파운데이션 모델 + Fine-tuning = 실용적인 개인화 파운데이션 모델
- 적은 데이터로 학습 가능
- 학습 리소스 절약 가능
- 특정 작업에 대한 우수한 성능
- (MLLM 가정) 사전 학습된 모델에 프롬프팅을 통한 작업을 했을 때보다 더 좋은 퀄리티의 결과물을 생성
- 프롬프트에 넣는 예제보다 훨씬 더 많은 예제를 통해 학습 가능
- 프롬프트 길이가 줄어들면서 토큰 개수 절약
- 응답하는데 걸린느 시간(latency)을 단축

하이퍼파라미터 - Learning Rate

- 손실함수가 큰 값일 때 미세하게 조정하기 어려우므로 뉴럴넷 모델에 작은 비율로 반영함
- Learning rate : 반영할 비율
- 적절한 learning rate 값
- 모델과 데이터마다 달라 실험을 통해 구함

너무 낮은 learning rate 값

- local minimum에 빠져서 global minimum에 도달할 가능성이 낮아짐

너무 높은 learning rate 값

- 마구 점프를 뛰다보니 global minimum으로 딱 맞춰 가기 어려워짐

효율적인 모델 학습

- 오픈소스로 공개된 고성능 파운데이션 모델을 출발점으로 미세조정하는 접근이 일반화되었으나, 여전히 높은 비용
- 효율적인 미세조정 방법

프롬프트 디자인

- 언어모델에서 주로 활용. 모델이 원하는 레벨의 결과를 출력할 수 있도록 입력 텍스트를 변형하는 방법
- 장점 : 추가 학습 없이 사전학습된 모델의 예측 성능을 끌어올릴 수 있음
- 단점 : 프롬프트를 사람이 직접 설계해야 한다는 부담감이 있으며 성능 향상이 제한적

프롬프트 튜닝

- 학습 가능한 프롬프트로서, 가상 토근을 입력에 추가
- 역전파를 통해 오직 가상 토큰에 대한 임베딩만 학습하고 나머지 모델은 고정

장점

- 사람의 디자인 없이 스스로 프롬프트를 학습할 수 있음
- 사전 학습된 모델을 고정할 수 있음
- 지식 손실 없음 : 일반 파인튜닝은 지식 손실 발생
- 적은 비용으로 새로운 데이터셋의 모델을 학습할 수 있음
- 학습된 프롬프트는 해석 x (그저 숫자 열 일뿐)

Adaptor 모듈 추가 학습

- Activation을 변경하기 위해 작은 모듈을 추가하여 학습하는 기법

합성데이터 활용법

Knowledge Distillation (Teacher - Student 학습)

- 사전학습된 고성능 모델의 지식을 작은 모델에 압축해서 빠르고 효율적으로 만들 수 없을까? 라는 생각에서 시작
- 지금까지 배운 미세조정과 같은 전이학습은 사전학습된 모델과 새로 학습할 모델의 구조가 동일한 경우를 가정하고 있었음

지식증류 : 높은 성능의 무거운 모델을 모방하도록 가벼운 모델을 학습하는 방법

- 크기가 작은 모델만으로 충분히 학습하기 어려운 데이터 특징을 학습하기 위해, 비교적 무겁고 성능이 높은 모델의 도움을 받는 기법으로 볼 수 있음
- 선생님 모델이 예측한 soft-label 값과 학생 모델의 예측 값이 가까워지도록 학습 유도

2-1. 합성데이터 활용법 1 | 지식증류: 높은 성능의 무거운 모델(선생님)을 모방하도록 가벼운 모델(학생)을 학습하는 방법 • 또는, 크기가 작은 모델(student)만으로 충분히 학습하기 어려운 데이터 특징을 학습하기 위해, 비교적 무겁고 성능이 높은 모델(teacher)의 도움을 받는 기법으로 볼 수 있음

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

| 파운데이션 모델들을 툴로 활용하는 방법 - InstructPix2Pix (2023) • 명령(지시사항; instruction)에 따라 이미지 편집을 수행하는 모델 • 기존 방법: 입력 이미지와 출력 이미지에 대한 상세 설명 필요 • 본 방법: 입/출력 이미지 상세 설명 없이, 명령만으로 편집 수행

InstructPix2Pix 방법 (1) • 기존 범용 데이터셋: {이미지, 이미지 설명 (캡션)} • 지시사항(instruction) 기반 이미지 편집을 지도학습 문제로 전환 (입력 데이터-정답 쌍 필요) • 가장 먼저, {이미지 편집에 대한 지시사항, 편집 전 이미지, 편집 후 이미지} 형식의 학습 데이터셋 생성

InstructPix2Pix 방법 (2) • 기존 범용 데이터셋: {이미지, 이미지 설명 (캡션)} • 지시사항(instruction) 기반 이미지 편집을 지도학습 문제로 전환 (입력 데이터-정답 쌍 필요) • 가장 먼저, {이미지 편집에 대한 지시사항, 편집 전 이미지, 편집 후 이미지} 형식의 학습 데이터셋 생성

InstructPix2Pix 방법 (3) • 그 후, 생성된 텍스트 데이터셋을 기반으로 별도의 이미지 편집 생성 모델로 영상 데이터 쌍 생성

InstructPix2Pix 방법 (4) • 생성된 이미지-명령 쌍 데이터셋을 기반으로, 최종 이미지 편집 생성 모델을 학습 (fine-tuning)

LLaVA 학습 데이터 • GPT를 활용하여 시각 설명 데이터(visual instruction data) 생성 기존에 존재하는 이미지, 캡션, 탐지 데이터셋 정답 데이터 활용

- GPT를 이용하여 문제-정답 데이터 쌍을 생성
-

간단한 시뮬레이션 기반 합성 데이터 : 실제 데이터를 모방하거나 새로 생성한 인공 데이터 • 예시 : 가상의 이미지, 텍스트, 소리 등을 알고리즘을 통해 생성 • 실제 데이터를 수집하거나 사용하기 어려운 경우에 대체 가능 • 데이터 부족 문제를 해결하고 모델 성능을 개선하는 데 사용

합성 데이터는 데이터 취득이 어려운 문제에서 특히 더 유용 • 실제 촬영하기 어려운 움직임이나 환경을 시뮬레이션하여 다양성 높은 데이터를 생성 가능

- 예시) 모션 증폭 : 심박수 변화, 건물 진동 등의 미세한 움직임을 합성 데이터를 통해 학습 가능

AGI를 향해서 | Human's Intelligence (cognition) = perception U higher cognitive processes

언어모델 + 검색증강생성

언어모델(LLM) + 검색증강생성(Retrieval Augmented Generation; RAG) • 유연성 : 추가 학습 없이 최신 정보 제공 • 개인화 : 검색과 언어모델의 합성으로 개인 맞춤형 답변 가능 • 정확성 : Verification을 통한 환각(Hallucination) 현상 감소

Claude Computer Use: 텍스트를 기반으로 컴퓨터를 사람처럼 사용할 수 있는 서비스 • 컴퓨터가 수행하길 바라는 지시사항을 텍스트로 입력하면 자동으로 명령 수행 • 각 명령어를 실행하는 agentic tool로 구성되어 있는 서비스

Agent Laboratory • Agent system을 활용하여 연구를 자동으로 수행하는 시스템 개발 • 연구를 진행하는 단계를 agent system처럼 구축하여 자동으로 연구 주제 탐색 및 논문 작성을 유기적으로 수행

AI 관련 오픈 소스 모델과 데이터셋을 공유하는 플랫폼 • 주요 특징 : 사전학습 모델 가중치 제공, 모델 학습을 위한 다양한 데이터셋 제공 • 응용 분야 : 자연어 처리(NLP), 컴퓨터 비전(CV), 음성 인식(Speech) 등 다양한 분야에서 사용 허깅페이스 활용 • 허깅페이스에서 제공하는 모델을 직접 불러와 다양한 작업에 적용 • 파운데이션 모델을 활용한 손쉬운 실습 가능

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 (예: REST API. gRPC)
- (운영 보조 기능): 확장 및 라우팅, 모니터링 등의 툴 제공

Hugging Face Inference API • 별도 서버 구축 없이 Hugging Face 플랫폼에서 REST API만으로 모델을 바로 사용할 수 있는 서비스 • 연구 및 테스트 목적:Request 수 제한 있음 (무료, 유료 버전)

안드로이드에서 Gemma 모델 실행하기 · 모바일 서빙 프레임웍 MediaPipe, MLC LLM을 통해 쉽게 LLM 모델 실행 가능 · 모바일 안드로이드에서 LLM 모델을 사용할 경우, 개인 정보 보안 문제와 인터넷 연결이 끊긴 상태에서도 사용할 수 있다는 장점이 존재

다양한 온디바이스 모델 실행 • LLM뿐만 아니라 얼굴 인식에 사용되는 모델(MobileFace) 등, 다양한 모델들도 온디바이스 환경에서 실행 가능 • 추가학습: 멀티플렛폼 호환성을 위해 고안된 ONNX를 공부하고, 온디바이스 배포 방법에 대해서 조사해보기

Gradio • 머신러닝 모델을 웹 인터페이스로 쉽게 배포할 수 있게 도와주는 오픈 소스 라이브러리

- 코딩 지식이 없는 사용자도 웹 브라우저를 통해 모델과 상호작용할 수 있는 환경 제공 Hugging Face와 통합되어 사용자가 Gradio 기반의 데모를 쉽게 생성하고 배포할 수 있음. 다양한 입력(이미지, 텍스트, 오디오 등)과 출력을 지원

■ 파인튜닝은 모델을 특정 도메인/사용자에 맞게 효율적으로 적용할 수 있는 매우 실용적 기술이다.

■ 합성데이터 활용법은 학습/평가 데이터 취득이 어려울 때, 매우 실용적인 방법론이다. ■ 생각보다 간단한 데이터가 활용성이 좋을 수 있다.

■ 이미지 파운데이션 모델의 개발 동향: 이미지+언어모델의 성공 사례 ⇒ 멀티모달 언어모델로 확장

■ 실습 ■ HuggingFace는 다양한 모델이 공유되어 활용할 수 있는 플랫폼이다. ■ Gradio를 통해 모델을 손쉽게 서빙할 수 있다.