[객관식 20문항 정답 포함]

1. 파운데이션 모델(Foundation Model)의 주요 특징으로 옳지 않은 것은?
   ① 대규모 데이터 학습
   ② **특정 태스크에만 한정된 모델** ☑
   ③ 범용성 확보
   ④ 빠른 적응성

2. Zero-shot 학습의 정의로 가장 적절한 것은?
   ① 예제 몇 개만 보고 학습
   ② **추가 학습 없이 새로운 문제에 바로 적용** ☑
   ③ 특정 데이터로 미세 조정
   ④ 전혀 학습하지 않은 상태로 예측 불가

3. CLIP 모델의 학습 방식은?
   ① 지도학습
   ② 강화학습
   ③ **대조학습** ☑
   ④ 자가회귀학습

4. SigLIP 모델의 주요 개선점은?
   ① **Softmax 제거, Sigmoid 기반 손실함수 사용** ☑
   ② ViT 대신 CNN 구조 사용
   ③ 텍스트 인코더 제거
   ④ 학습 데이터 크기 감소

5. 멀티모달 정합(Multi-modal Alignment)의 목표는?
   ① 데이터 증강
   ② **공통 임베딩 공간 구성** ☑
   ③ 하이퍼파라미터 최적화
   ④ 손실함수 단순화

6. CLIP의 텍스트 인코더 구조는?
   ① CNN

② **Transformer** ☑
③ RNN
④ LSTM

7. LLaVA의 1단계 학습(Pre-training)에서 학습되는 것은?
① 전체 모델
② Vision Encoder
③ **Projection Layer** ☑
④ Text Decoder

8. Qwen2.5-Omni의 특징은?
① 텍스트만 처리
② 다국어 불가능
③ **모든 모달리티 통합(읽고, 쓰고, 보고, 듣고, 말함)** ☑
④ 시각적 처리 불가능

9. Set of Mark(SoM)의 주요 목적은?
① **객체 인식 정확도 향상** ☑
② 데이터 압축
③ 모델 경량화
④ 오디오 합성

10. CLIP에서 이미지 인코더로 사용되는 기본 구조는?
① CNN
② **ViT(Vision Transformer)** ☑
③ RNN
④ BERT

11. ImageBind 모델의 특징은?
① 이미지-텍스트만 학습
② **6가지 모달리티 통합** ☑
③ 텍스트만 인식
④ 비디오만 처리

12. SmolVLM이 기존 모델 대비 가지는 장점은?

① 대규모 파라미터
② **경량화된 구조** ☑
③ 음성 인식 전용
④ 대규모 데이터셋 미사용

13. Moondream 모델의 주요 목적은?
① 대규모 서버용 모델
② **모바일/엣지 디바이스용 실시간 VLM** ☑
③ 이미지 생성 전용
④ 텍스트 요약 전용

14. Gemini Nano 모델이 사용된 디바이스는?
① 아이폰 16
② **픽셀 9 시리즈** ☑
③ 갤럭시 S25
④ 화웨이 메이트70

15. Grounding DINO의 역할은?
① **이미지 내 객체 탐지** ☑
② 음성 인식
③ 비디오 합성
④ 텍스트 요약

16. Segment Anything(SAM) 모델의 목적은?
① 이미지 분류
② **이미지 세분화(Segmentation)** ☑
③ 음성 합성
④ 3D 모델링

17. Depth Anything v2의 주요 응용은?
① 음악 생성
② **자율주행 및 로봇 비전** ☑
③ 텍스트 분류
④ 기계번역

18. Fine-tuning(미세조정)의 장점으로 옳지 않은 것은?
① 적은 데이터로 학습 가능
② 특정 작업에 최적화
③ 모델 지식 손실 최소화
④ **항상 대규모 학습 필요** ☑

19. 프롬프트 튜닝의 장점은?
① 프롬프트를 사람이 직접 설계해야 함
② **사전학습 모델을 고정한 채 가상 토큰만 학습** ☑
③ 높은 비용이 필요
④ 지식 손실이 큼

20. Knowledge Distillation(지식 증류)의 개념은?
① **무거운 모델이 가벼운 모델을 학습하도록 도움** ☑
② 데이터 증강 기법
③ 학습률 조정
④ 하이퍼파라미터 탐색

[주관식 10문항 정답 포함]

1. **파운데이션 모델의 개발 프로세스를 순서대로 쓰시오.**
👉 Data Creation → Data Curation → Training → Adaptation → Deployment

2. **CLIP의 대조학습(Contrastive Learning) 원리를 간단히 설명하시오.**
👉 이미지와 대응되는 텍스트 쌍의 유사도를 높이고, 일치하지 않는 쌍의 유사도를 낮추는 방식으로 학습한다.

3. **SigLIP이 CLIP보다 노이즈에 강건한 이유를 기술하시오.**
👉 Softmax 대신 Sigmoid 손실을 사용해 일정 거리 이상 떨어진 음성 샘플은 학습에서 제외하여 노이즈 영향을 줄인다.

4. **멀티모달 정합 손실함수(CLIP Loss)의 목적은 무엇인가?**
   👉 서로 다른 모달리티(예: 이미지-텍스트)가 공통 임베딩 공간에서 의미적으로 가깝게 정렬되도록 학습하기 위함이다.

5. **LLaVA 모델의 학습 과정에서 ChatGPT가 어떤 역할을 하는지 서술하시오.**
   👉 ChatGPT를 이용해 이미지 설명, 질의응답 등 합성 데이터(visual instruction data)를 생성하여 모델의 Fine-tuning 데이터로 사용한다.

6. **Qwen2.5-VL이 이전 버전 대비 강화된 기능을 두 가지 이상 쓰시오.**
   👉 문서 파싱(OCR, 표, 차트 등) 강화, 정밀한 객체 그라운딩, 장시간 비디오 이해 기능 추가.

7. **SoM(Set of Mark)이 VLM의 시각 능력 향상에 기여하는 이유를 설명하시오.**
   👉 이미지 내 객체에 번호를 부여하여 객체 위치를 명확히 식별함으로써 시각적 인식 정확도를 높인다.

8. **HyperCLOVAX-SEED-Vision-Instruct-3B 모델의 주요 특징을 요약하시오.**
   👉 NAVER가 개발한 한국어 특화 멀티모달 모델로, 텍스트와 이미지를 동시에 이해하며 HuggingFace에서 실행 가능하다.

9. **Fine-tuning과 Prompt Tuning의 차이를 간략히 비교하시오.**
   👉 Fine-tuning은 모델 가중치를 직접 조정하고, Prompt Tuning은 모델은 고정한 채 학습 가능한 가상 토큰만 조정한다.

10. **Knowledge Distillation(지식 증류)의 학습 구조(Teacher-Student)를 설명하시오.**
    👉 성능이 높은 Teacher 모델의 예측 결과(Soft-label)를 참고하여 Student 모델이 이를 모방하도록 학습한다.