CLIP (Contrastive Language–Image Pre-Training)

작성일시	@2025년 11월 12일 오후 1:53
복습	☐

1) 모델 개발 배경

1. 기존 컴퓨터 비전 모델의 한계

- 전통적인 이미지 분류 모델은 레이블링된 데이터셋 위주로 학습
- 라벨링은 비용과 시간 부담이 크다 + 라벨링되어 있는 클래스 외의 새로운 데이터에 대한 일반화가 약한 문제
- 직접적인 라벨링 없이도, 일반화된 의미를 학습할 수 있는 새로운 접근 필요

2. 이미지에 해당하는 단어/문장은 풍부하게 존재한다

- 이미지에 해당하는 단어/문장(캡션, 태그)는 인터넷상에서 풍부하게 존재함
- 해당 이미지-텍스트쌍을 공유된 표현 공간에서 매핑하여 이미지-캡션을 매칭하도록 한다

---

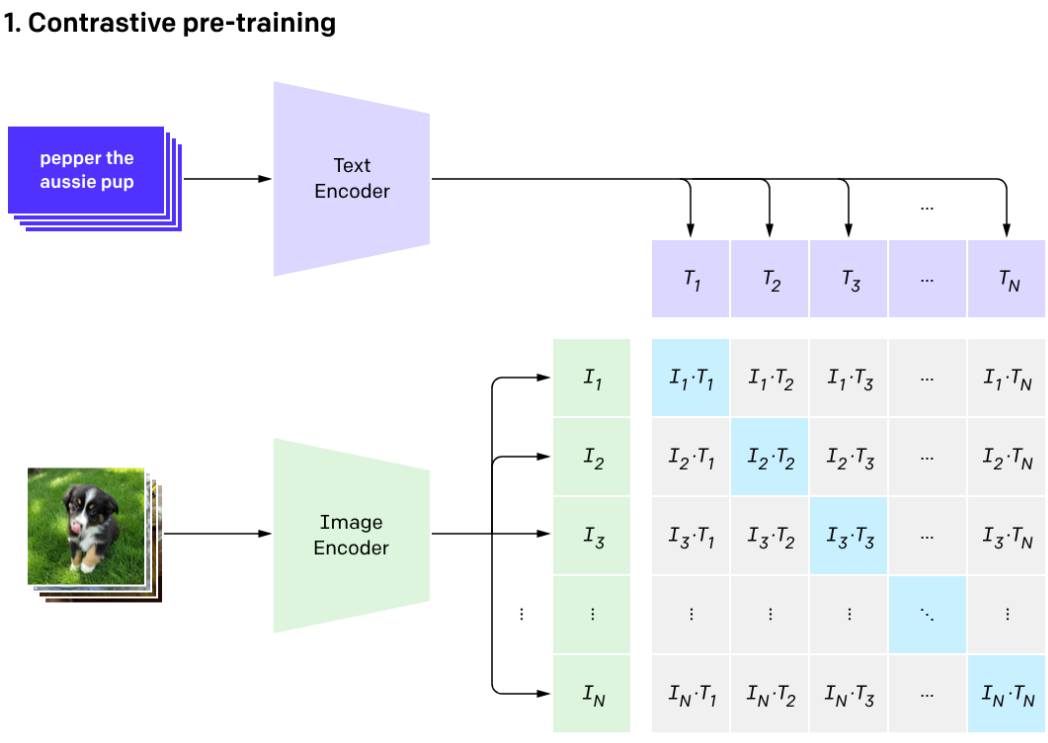
2) 모델의 주요 목적

- 주된 목적은, 인터넷에 존재하는 방대한 (이미지,캡션)을 같은 Representation 공간에서 연결하는 것
- zero-shot 방식으로 다양한 시각 Task에 대응하는 것
  - 기존: 특정 클래스 (”개”, “고양이”, “자동차”)에 대해 미리 지정한 Label에 대해서만 학습
  - CLIP은 추가 라벨링없이 → 자연어 설명만 있어도 어떤 클래스에 속하는지 분류 가능
- 결국, 별도의 라벨링 비용을 줄이고, 이미지에 대해서, 일반화된 이해를 할 수 있는 Vision 모델을 만들고자 함

---

3) CLIP의 학습 과정

1) Contrastive Pre-training

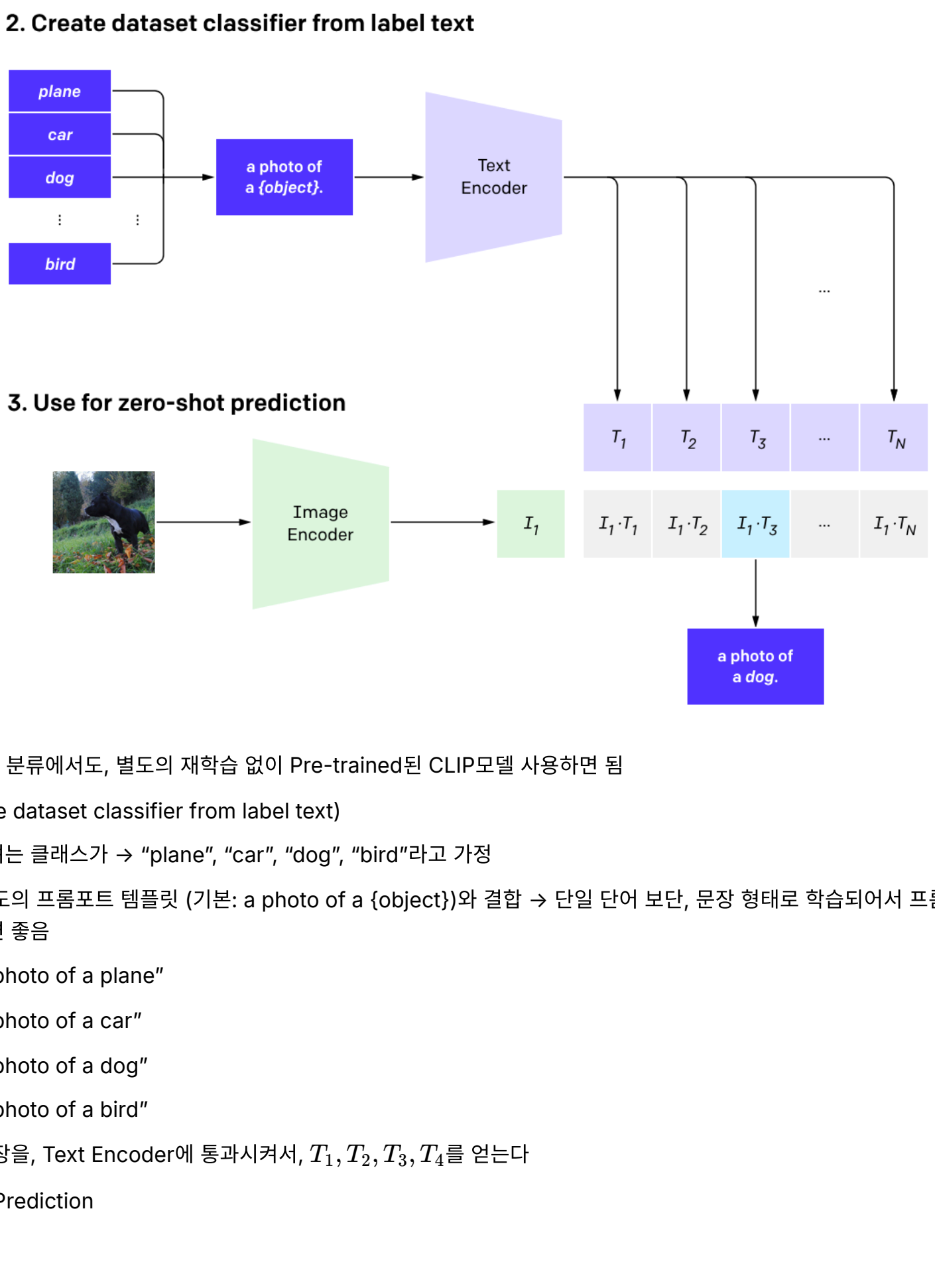


- 이미지와 텍스트를 각각의 Encoder로 변환하여 → Embedding Space에서 서로 짝이 맞는건 가깝게, 짝이 아닌 것은 멀게끔 학습

▼ 예시

- 데이터
  - 이미지: [강아지 사진, 고양이 사진]
  - 텍스트: [강아지가 웃고있다, 고양이가 밥을 먹는다]
- 이미지 임베딩
  - 강아지 사진 → I_1
  - 고양이 사진 → I_2
- 텍스트 임베딩
  - “강아지가 웃고 있다” → T_1
  - “고양이가 밥을 먹는다” → T_2
- 서로 맞는쌍은 가깝게
  - (I_1, T_1)과, (I_2, T_2)는 서로 유사도가 가깝게
  - (I_1, T_2)와, (I_2, T_1)은 서로 유사도가 멀게

2) Create dataset classifier from label text and Zero-shot Prediction



- 새로운 Task 분류에서도, 별도의 재학습 없이 Pre-trained된 CLIP모델 사용하면 됨

▼ 예시 (Create dataset classifier from label text)

- 분류하려는 클래스가 → “plane”, “car”, “dog”, “bird”라고 가정
- 이를 별도의 프롬포트 템플릿 (기본: a photo of a {object})와 결합 → 단일 단어 보단, 문장 형태로 학습되어서 프롬포트 형식으로 변환하면 좋음
  - “a photo of a plane”
  - “a photo of a car”
  - “a photo of a dog”
  - “a photo of a bird”
- 해당 문장을, Text Encoder에 통과시켜서, T_1, T_2, T_3, T_4를 얻는다

▼ Zero-shot Prediction

- 한 번도, 학습하지 않은 데이터셋에도 텍스트 설명만 있다면 분류가 가능해짐

▼ 예시

- 이전 예시에서 이미지를 → “plane”, “car”, “dog”, “bird”로 분류하려고 함
- 현재 첫번째 이미지(강아지 사진)를 Image Encoder에 통과 → I_1
- 이후, 이전 단계에서 만든 Text Embedding Vector와 내적하여, 가장 유사도 높은 T_i를 고른다
- T_i에 해당하는 object로 분류하게 된다 (현재 예시에선 Dog)

---

4) 응용 결과 (Performance)

- zero-shot 방식 이미지 분류에서 큰 폭의 성능 향상
- 일반 객체 분류 데이터셋인 STL10에서 99.3%의 정확도 (Zero-shot으로 매우 높은 성능)
- 다만, 세부 클래스 구분이 까다로운 Task(ex, 비행기 기종 식별, 식물 종 분류)에선 CLIP이 10%가량 성능이 떨어짐

▼ 해석

- CLIP은 많은 라벨링 학습 없이도, 꽤 좋은 성능을 보여준다
- 일반 객체인식에선 매우 좋은 성능
- 하지만 매우 정밀한 구분을 요구하는 Task, 인터넷에 자료가 별로 없던 전문적인 Domain 이미지는 아직 한계