Learning Transferable Visual Models From Natural Language Supervision

https://arxiv.org/abs/2103.00020

0. Introduction

- 기존 컴퓨터 비전 모델은 고정된 범주의 예제로만 학습되어 일반화에 한계가 있음
- 이미지-텍스트 쌍에서 직접 학습하는 언어 기반의 감독 방식은 더 다양한 개념 학습 가능성을 제공함
- CLIP은 이미지와 캡션 매칭을 예측하도록 사전학습하여, 언어와 시각을 동시에 이해하는 범용 모델을 목표로 함

1. Overview

- 이미지와 텍스트를 각각 인코딩하는 dual-encoder 구조
- 4억 개의 이미지-텍스트 쌍(WebImageText)으로 사전학습 수행
- contrastive learning 방식으로 이미지와 해당 캡션의 임베딩 유사도를 최대화하고, 잘못 매칭된 쌍 유사도는 최소화함
- 사전학습된 모델은 텍스트 프롬프트만으로 다양한 시각 과제에서 바로 적용(Zero-shot) 가능

2. Challenges

- 다중 레이블 사전 학습 및 고정 범주 의존 방식은 새로운 개념에 적용하는 데 한계가 있음
- 이미지-텍스트 쌍은 레이블이 아닌 풍부한 자연 언어 정보를 담고 있어, 이를 제대로 활용할 방법이 필요했음

3. Method

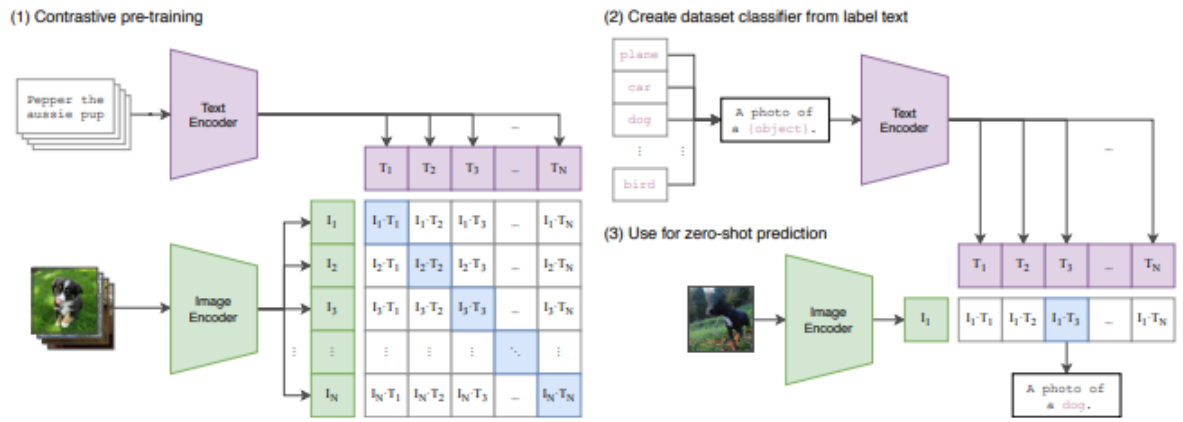


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

- dual-encoder: 이미지 인코더 (ResNet 또는 ViT), 텍스트 인코더 (Transformer 기반)
- 공통 임베딩 공간으로의 선형 투영
- 배치 내 모든 이미지·텍스트 쌍에 대해 쌍별 유사도 측정 → N×N 매칭 행렬 구성
- 양의 쌍(정확히 매칭된 이미지-텍스트)은 유사도를 높이고, 음의 쌍은 낮게 하도록 대조 손실(대칭 교차엔트로피) 최적화
- 온도 매개변수(temperature)를 학습하여 softmax 스케일 조정
- 추론 시, "A photo of a {label}" 형태로 프롬프트를 생성해 텍스트 목록과 이미지 유사도를 계산한 후 최고 유사도 텍스트를 예측 결과로 선택

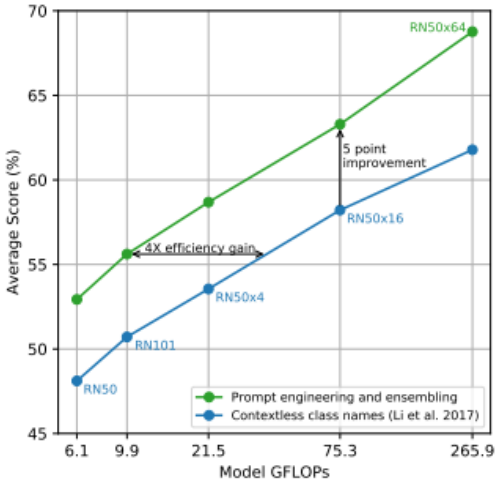
4. Experiments

- 사전학습: WebImageText(약 4억 쌍), ResNet50 및 ViT 모델 사용
- 하이퍼파라미터: Adam optimizer, cosine learning rate decay, 대형 배치 (32768), ResNet50×64는 592 GPU × 18일, ViT-L/14는 256 GPU × 12일 학습

5. Results

	aYahoo	ImageNet	SUN
Visual N-Grams	72.4	11.5	23.0
CLIP	**98.4**	**76.2**	**58.5**

Table 1. Comparing CLIP to prior zero-shot transfer image classification results. CLIP improves performance on all three datasets by a large amount. This improvement reflects many differences in the 4 years since the development of Visual N-Grams (Li et al., 2017).



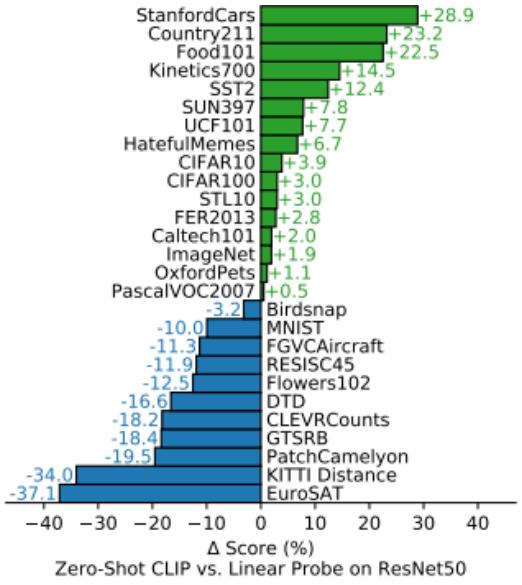


Figure 5. **Zero-shot CLIP is competitive with a fully supervised baseline.** Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.

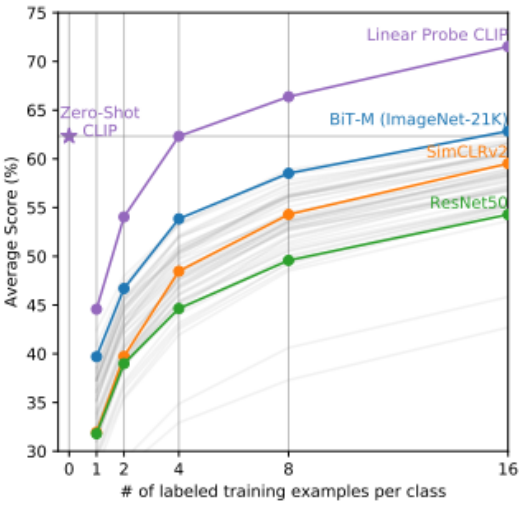


Figure 6. **Zero-shot CLIP outperforms few-shot linear probes.** Zero-shot CLIP matches the average performance of a 4-shot linear classifier trained on the same feature space and nearly matches the best results of a 16-shot linear classifier across publicly available models. For both BiT-M and SimCLRv2, the best performing model is highlighted. Light gray lines are other models in the eval suite. The 20 datasets with at least 16 examples per class were used in this analysis.

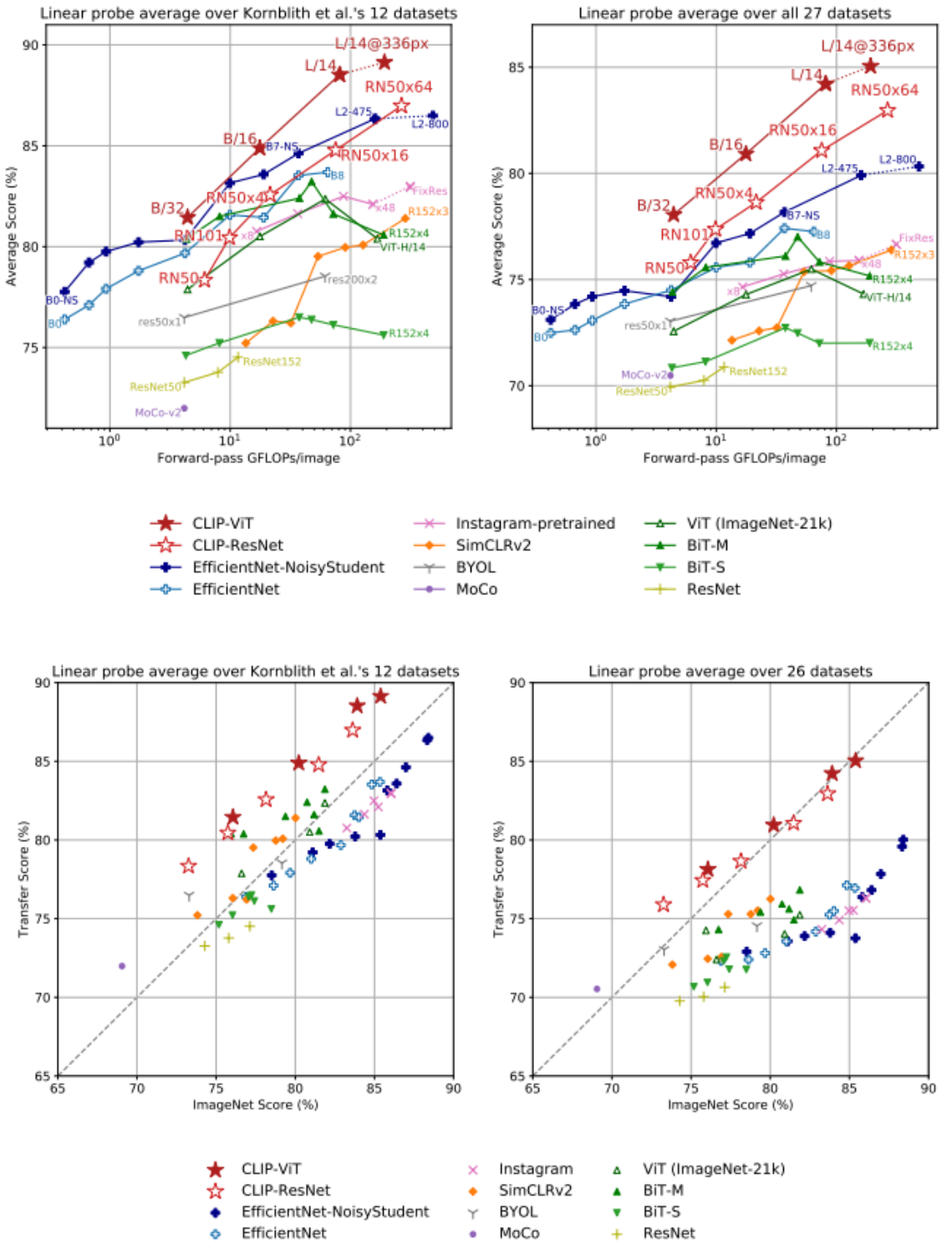


Figure 12. **CLIP's features are more robust to task shift when compared to models pre-trained on ImageNet.** For both dataset splits, the transfer scores of linear probes trained on the representations of CLIP models are higher than other models with similar ImageNet performance. This suggests that the representations of models trained on ImageNet are somewhat overfit to their task.

	Accuracy	Majority Vote on Full Dataset	Accuracy on Guesses	Majority Vote Accuracy on Guesses
Zero-shot human	53.7	57.0	69.7	63.9
Zero-shot CLIP	**93.5**	**93.5**	**93.5**	**93.5**
One-shot human	75.7	80.3	78.5	81.2
Two-shot human	75.7	85.0	79.2	86.1

- ImageNet zero-shot 성능 76.2%, supervised ResNet-50 수준 성능
- 총 30개 이상의 시각 과제(OCR, 영상 행동 인식, 지리 위치 추론, 정밀 분류 등)에서 다양한 베이스라인에 경쟁 또는 우위 성능
- 프롬프트 엔지니어링(예: “A photo of a big {label}”)과 앙상블이 성능 향상에 기여

6. Insight

- 자연어 감독만으로 이미지 표현을 학습하여 광범위한 zero-shot 전이가 가능함을 입증
- 이후 CLIP 기반 모델들이 Stable Diffusion 등의 다양한 분야에서 핵심 역할 수행함
- 세분화된 태스크(예: 종(species)의 구분, 객체 수 세기) 및
  OOD(Out-of-Distribution) 데이터에 취약
- 대형 학습에 높은 계산 비용이 드는 구조임
- 성능 향상을 위해 추가적으로 1000배 이상 계산 필요하다고 추정