Learning Transferable Visual Models From Natural Language Supervision

Radford et al., ICML 2021

2024.7.10

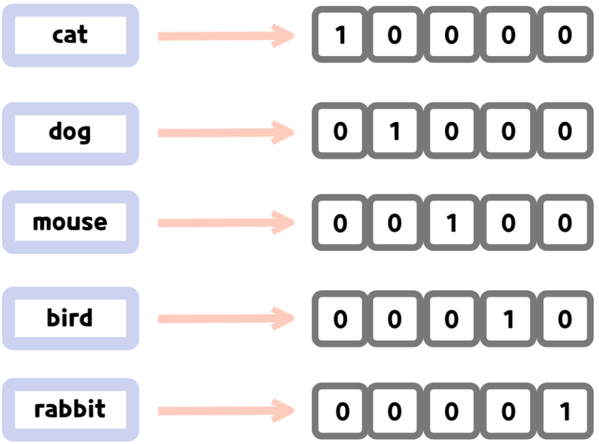
Lab Seminar

Harim Noh

Duksung Women's University
Dept. of Computer Engineering

Introduction

- 기존 비전 모델의 pre-training 문제 = 분류 문제
- ImageNet, JFT(Joint First Task) 데이터셋 등 큰 규모의 데이터셋을 활용해 모델을 미리 pre-training
- 문제점 : **새로운 라벨에 적용할 수 없음 = 모델의 일반화 ↓**
- 학습 당시 없던 Label의 이미지가 들어오면 제대로 분류하지 못함
- 새로운 Label을 추가해서 다시 학습 필요



One-Hot Encoding

Introduction

⇒ 더 나은 일반화 성능을 갖는 pre-trained 모델 필요

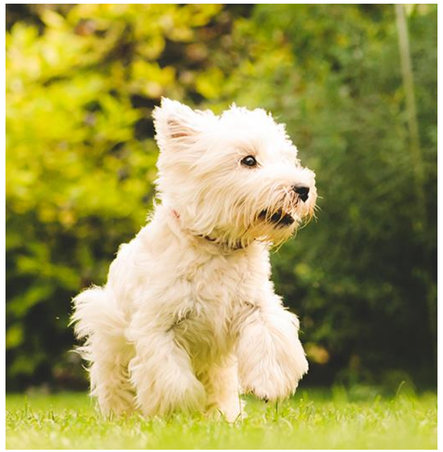
⇒ 자연어처리 분야처럼 웹 상에서 수집된 데이터셋으로부터 pre-training 하는 방식을 적용할 수 있을까?

- CLIP (Contrastive Language-Image Pre-training)
- OpenAI에서 개발
- 이미지와 자연어(텍스트)를 서로 대조하면서 학습시키는 방법론

CLIP

approach

- Natural Language Supervision
- 자연어를 이용한 방법론의 이점 = **scale을 키우기 쉬움**
- 학습할 때 텍스트는 자연어와 유동적으로 관련지어 학습
- 단순히 이미지를 텍스트로 표현하도록 학습하는 것이 아닌 모델이 이미지에 대해 깊은 이해를 가질 수 있도록 함

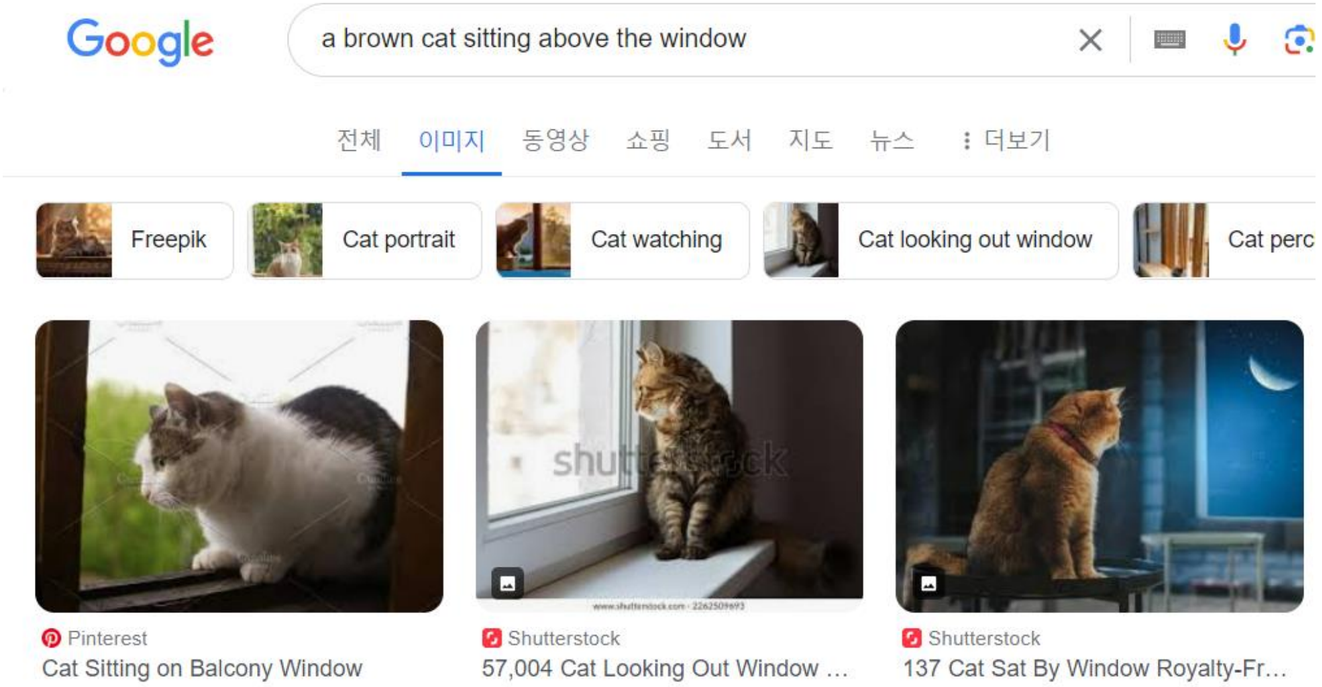
dog ⇦  ⇨ A dog playing in the park

CLIP

approach

- 본 연구에서는 인터넷에서 약 4억 개의 이미지와 텍스트 쌍으로 구성된 데이터셋을 수집
- 다양한 분야의 텍스트 및 이미지를 수집하기 위해 50만 건의 검색을 수행
- 검색어 당 최대 2만개의 이미지를 수집

⇒ WebImageText, WIT

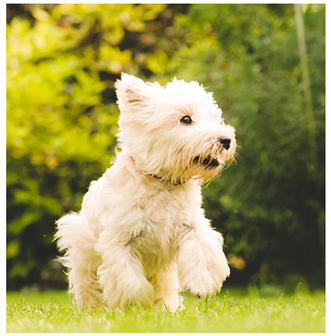


CLIP

approach

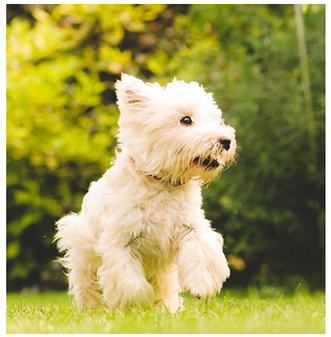
- 본 연구의 초기 접근 방법

1. Transformer-based : CNN / transformer를 이용해 학습 & 예측



“A dog playing in the park”

2. Bag of Words-based : 문장의 문법적 구조나 단어 순서는 무시하고, 단어의 출현 여부만을 고려

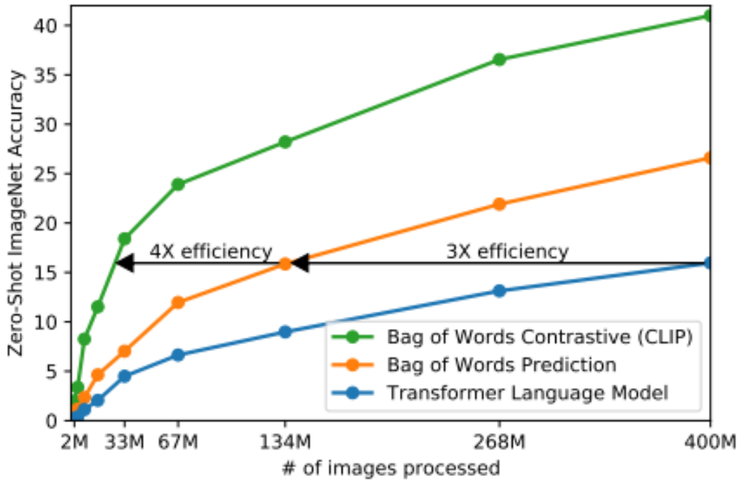


[“dog”, “puppy”, “playing”]

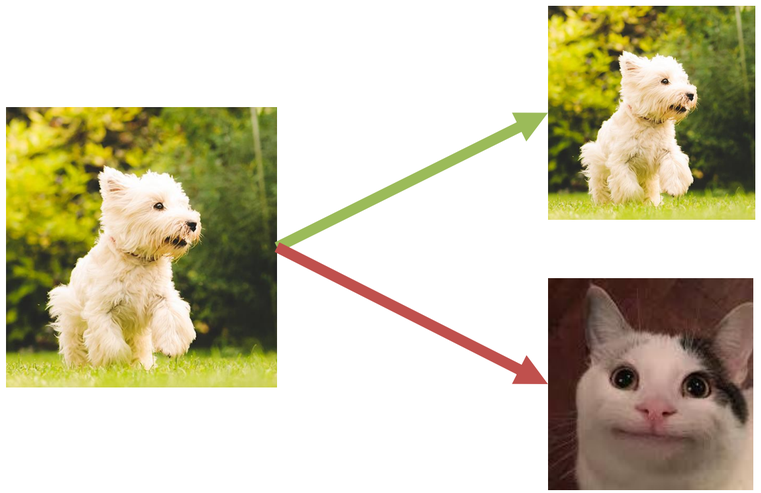
=> 학습 & 예측 시간이 길어 비효율적

CLIP

approach



- **Bag-of-words contrastive** 방식
- Bag-of-words encoding을 사용해 contrastive learning
- Zero-shot 예측에서도 가장 우수한 성능을 보임
- contrastive learning : positive & negative pairs 를 학습하여 데이터의 특성을 파악



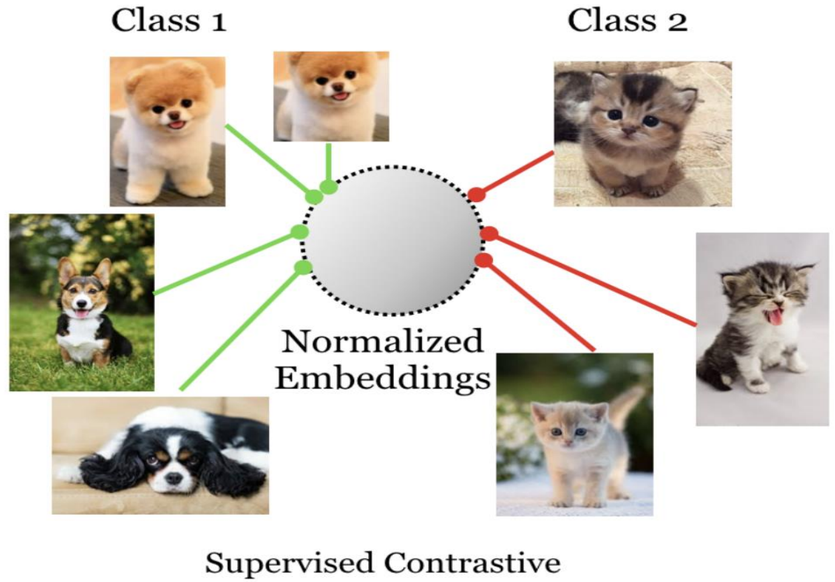
Positive : 관련이 있는 데이터 => 임베딩 벡터가 가까워짐

Negative : 관련이 없는 데이터 => 임베딩 벡터가 멀어짐

=> 양성 쌍의 유사도를 높이고 음성 쌍의 유사도를 낮춤

CLIP

approach



- 유사한 샘플 쌍을 가깝게, 유사하지 않은 샘플 쌍을 멀리 표현하도록 모델을 학습시키는 자기 지도 학습 방법
- 즉, 라벨이 없는 데이터를 활용하여 모델을 학습시키는 방법
- Contrastive learning 과정

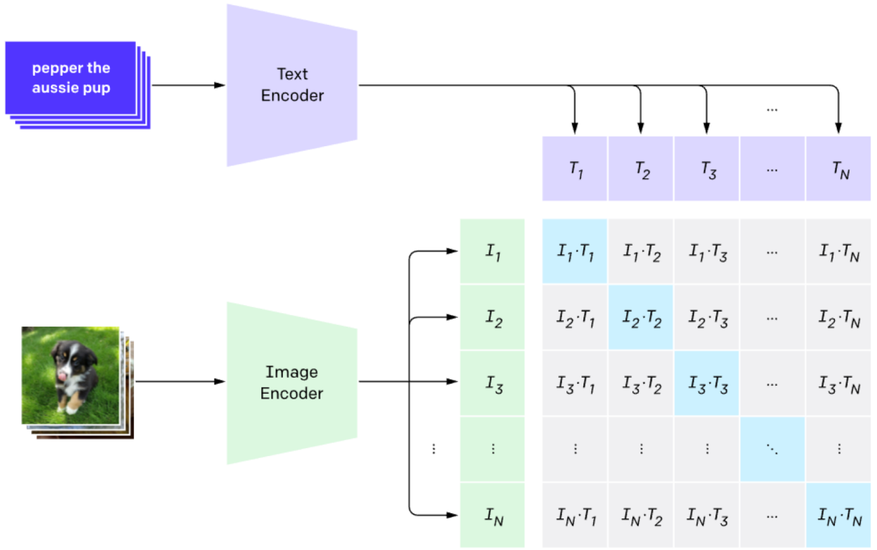
1. 앵커 (Anchor) 샘플 선택: 훈련 데이터에서 임의로 하나의 샘플을 앵커 샘플로 선택
2. 양수 (Positive) 샘플 선택: 앵커 샘플과 유사한 다른 샘플을 양수 샘플로 선택
3. 음수 (Negative) 샘플 선택: 앵커 샘플과 유사하지 않은 다른 샘플을 음수 샘플로 선택
4. 모델 학습: 앵커 샘플, 양수 샘플, 음수 샘플을 모델에 입력하고, 앵커 샘플과 양수 샘플은 가깝게, 앵커 샘플과 음수 샘플은 멀리 표현하도록 모델을 학습

https://paperswithcode.com/method/supervised-contrastive-loss

CLIP

methodology

- **Contrastive pre-training**
- 배치 단위로 이루어진 N개의 이미지와 텍스트를 인코더에 통과해 임베딩 벡터 산출
- L2 정규화를 통해 벡터의 크기를 제거하고 방향 정보를 유지함
- 벡터 간의 내적을 통해 코사인 유사도 구함




```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

CLIP

methodology

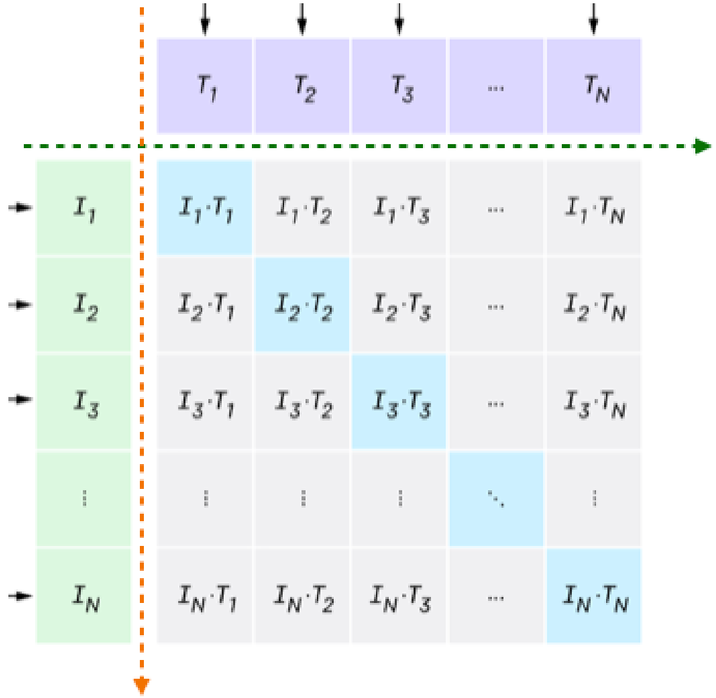
- **Contrastive pre-training**
- N 개 = 실제 (image, text) 페어 / $N^2 - N$ 개 = 서로 다른 (image, text) 페어
- 페어에 해당하지 않으면 다르다는 관계하에 InfoNCE Loss 계산
- t_i 이미지 i에 해당하는 정답 텍스트

이미지 i와 텍스트 j 사이의 코사인 유사도

$$L_{image} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(x_i, y_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(x_i, y_j)/\tau)}$$

유사도 값($\text{sim}(x_i, y_i)$)을 temperature 파라미터 τ로 나눠 줌으로써 유사도 값의 분포를 조정

I번째 이미지와 다른 모든 텍스트 간의 유사도 합



⇒ 이미지 i와 해당하는 텍스트 t_i 의 유사도 최대화
⇒ 다른 텍스트와의 유사도 최소화

CLIP

methodology

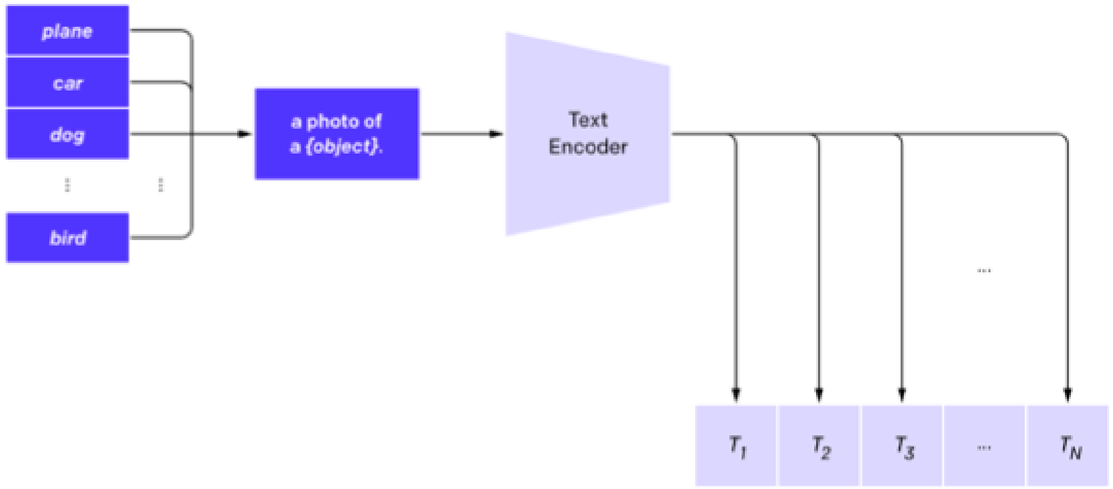
- **Create dataset classifier from label text**
- **데이터셋 분류기 생성**
- 주어진 데이터셋의 레이블(label) 정보를 활용해 해당 데이터를 분류할 수 있는 모델을 구축하는 것

ex) 개와 고양이를 분류하는 데이터셋이 있다면, Fine-Tuning 없이 개와 고양이를 구분하는 모델을 학습

- 적용하고자 하는 하위 문제의 데이터셋의 label을 텍스트로 변환

⇒ 구로 변환하여 입력 시 성능 향상

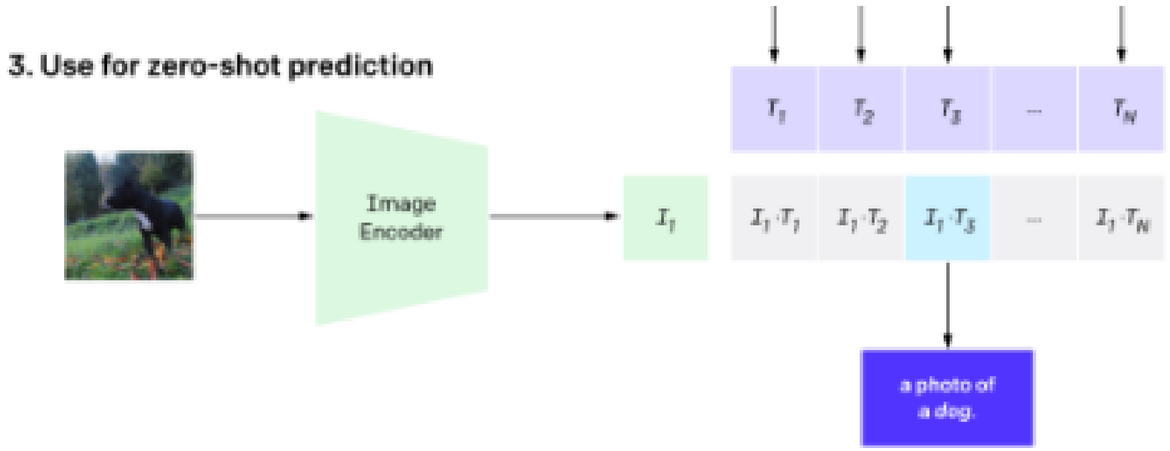
⇒ "a photo of a Dog“



CLIP

methodology

- **Use for zero-shot prediction**
- 사전 학습된 모델을 사용하여, 예측하고자 하는 이미지의 벡터와 각 텍스트 벡터 간의 유사도를 계산
- 계산된 유사도 값 중에서 가장 높은 값을 가지는 텍스트가 모델의 예측 결과로 선택



CLIP

results

- 27개 중 16개의 데이터셋에서 ResNet-50 baseline에 비해 더 좋은 성능을 보임
- 특수하거나 복잡한 경우 CLIP의 성능은 baseline보다 많이 낮음

⇒ EuroSAT, RESISC45은 satellite image를 분류

- 실험에 대한 자세한 내용은 논문을 참고

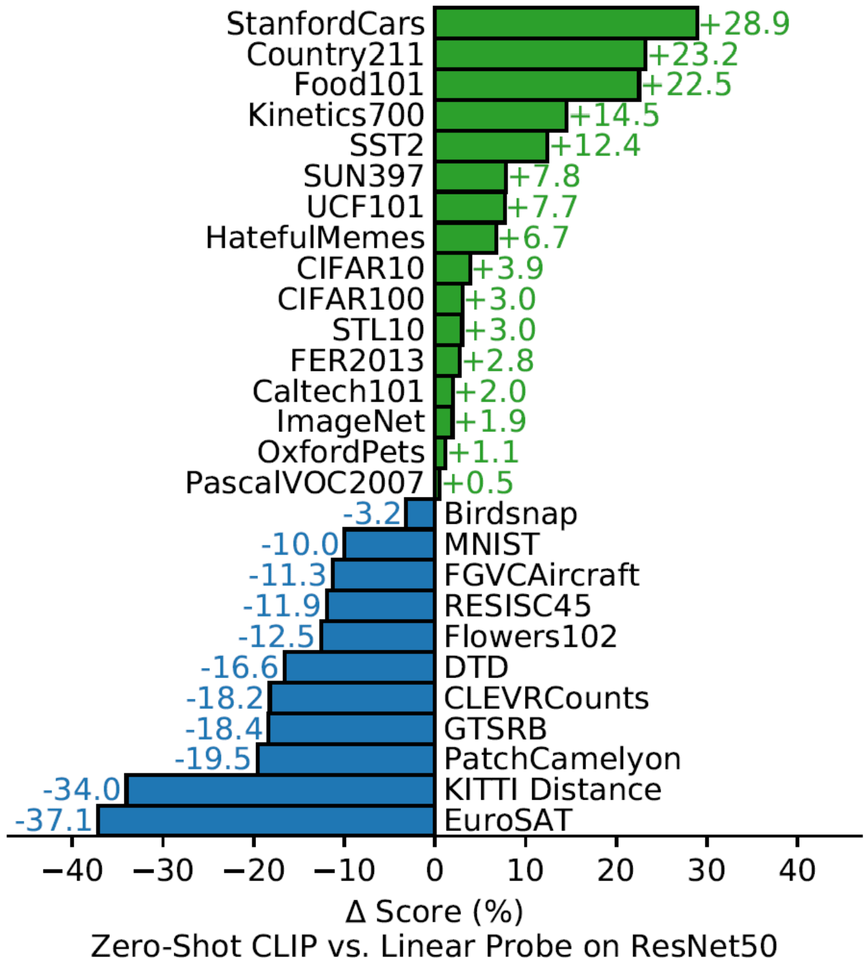


Figure 5. **Zero-shot CLIP is competitive with a fully supervised baseline.** Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.