간단 논세

Learning Transferable Visual Models From Natural Language Supervision

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

25.01.02
유하영

Introduction

NLP에서의 text-to-text는 특정 downstream에 대해 맞춤화 할 필요없이 발전해옴.
하지만, 당시 Computer vision에서는 ImageNet과 같은 label dataset에서 모델을 훈련하는 것이 관행

Natural language를 통한 Image Representation learning은 여전히 힘듦..

1. flexibility와 zero-shot capability 제한
2. 모델 scale 부족

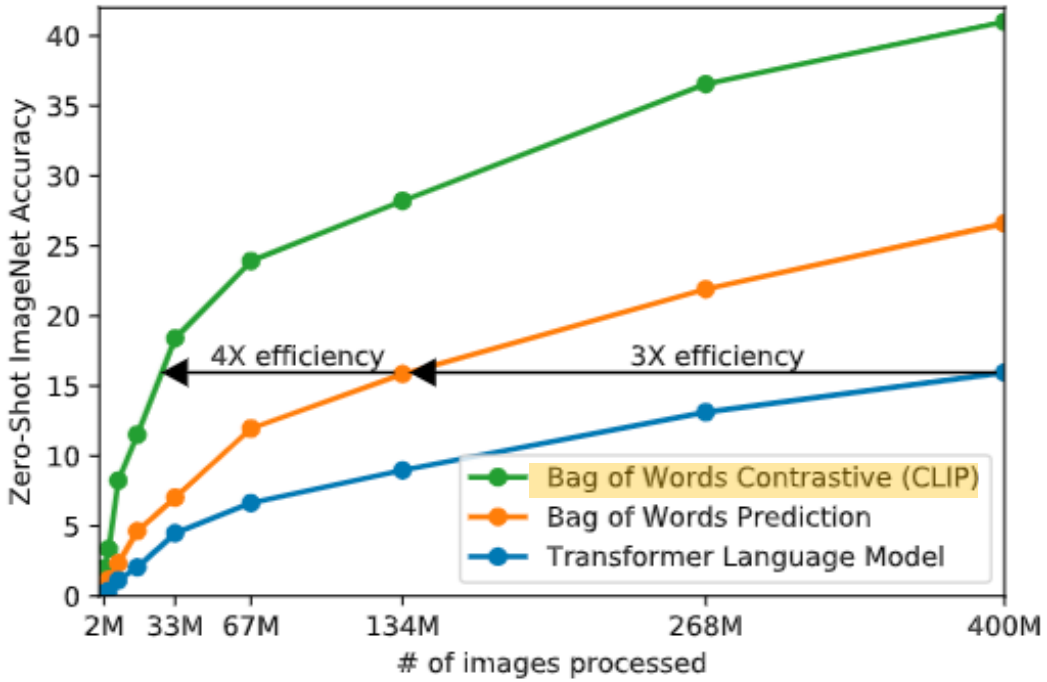
In this paper,

1. image-text를 연결하는 dataset으로 학습
2. 대규모 dataset으로 contrastive learning
3. Zero-shot transfer 강화
4. 모델 scale강화

Method

Initial Approach

VirTex = CNN + Text transformer



이전 연구

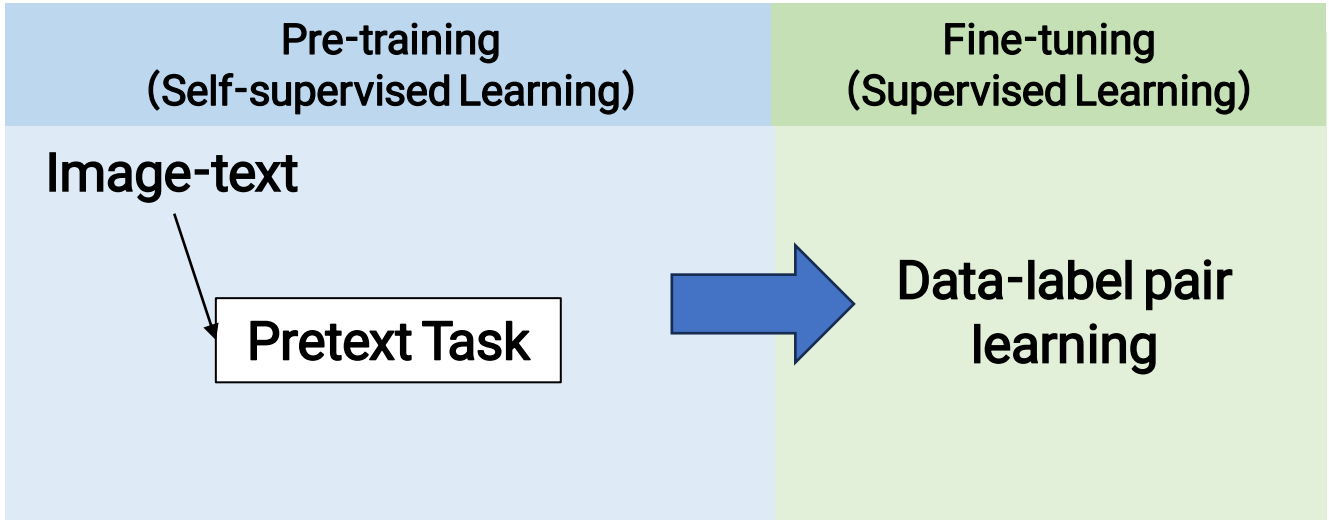
'Contrastive Objectives가 동등한 predictive objective 보다 더 나은 표현을 배울 수 있다. '

➔ 본 논문에서는 BoW 인코딩 기준선을 사용하되, predictive objective 대신 Contrastive Objectives를 사용

Method

Main Idea
: image와 text를 공통 임베딩 공간에 학습시켜, 새로운 작업이나 데이터셋에서도 추가 학습 없이 성능을 발휘하는 것 (zero-shot transfer)

1. Supervised Natural language



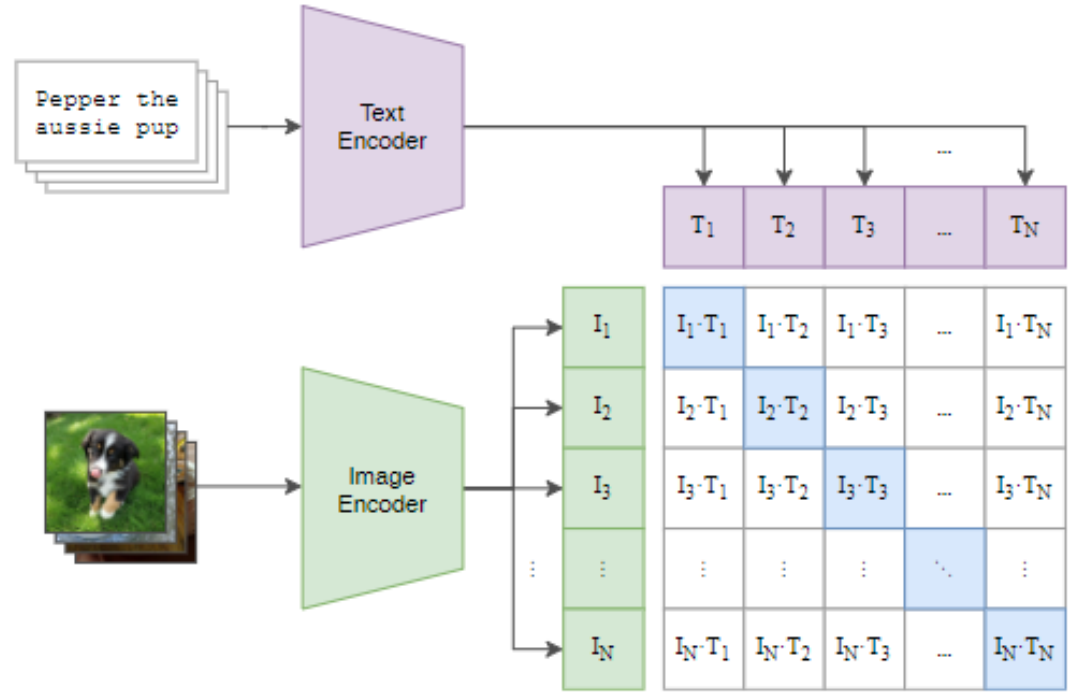
Contrastive Learning

Label X

Method

Pre-training

(1) Contrastive pre-training



Method

```
# 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
```

Figure 3. Numpy-like pseudocode for the core of an implementation of CLIP.

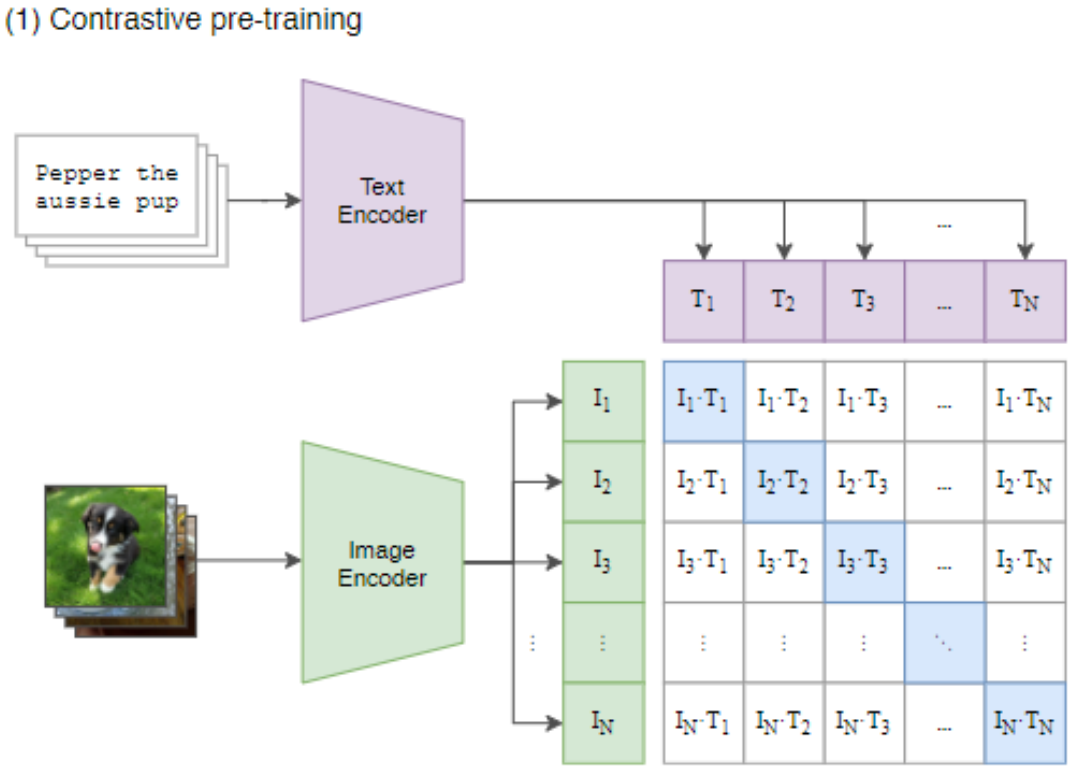
$$L_i = -\frac{1}{N}\sum_{i=1}^{N} \log \text{Softmax}(\text{logits}[i,:])[i]$$

$$L_t = -\frac{1}{N}\sum_{i=1}^{N} \log \text{Softmax}(\text{logits}[:,i])[i]$$

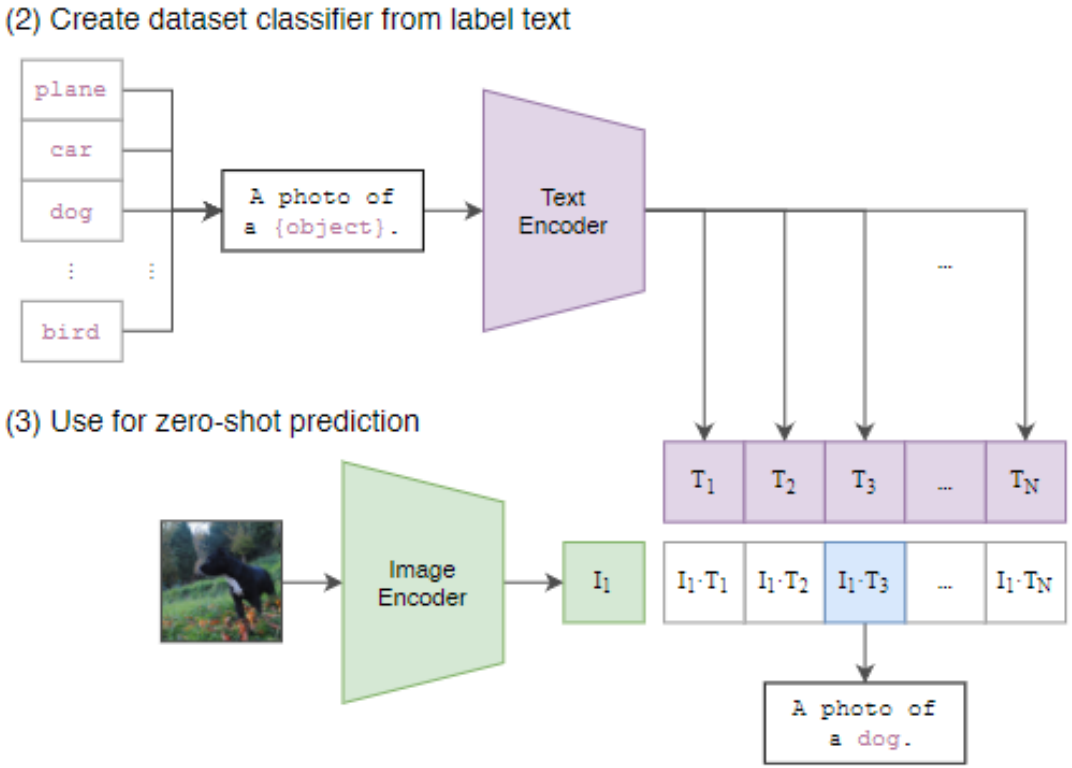
$$L = \frac{L_i + L_t}{2} \quad \text{// 대칭적}$$

Method

Pre-training



Inference



Method

- Creating a sufficiently large dataset

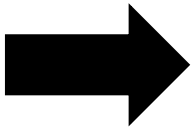
WIT(Web Image Text)

[사전 query]

50만개 query

- Wikipedia 기반,
- Bi-gram으로 구성

Internet 검색



[전체 데이터 개수]

4억개 (image,text)

Method

Train 세부사항

1) Encoder를 사전훈련된 weight로 초기화 X
2) Non-linear projection 사용 X
3) 단일 문장 sampling t_u 사용 X
4) Data Augmentation
5) Temperature Parameter T를 직접 최적화

Train Model

1) ResNet-50
2) ResNet-101
3) EfficientNet Style Scaling 적용
   RN50x4, RN50x16, RN50x64..
4) ViT – B/32, B/16, L/14

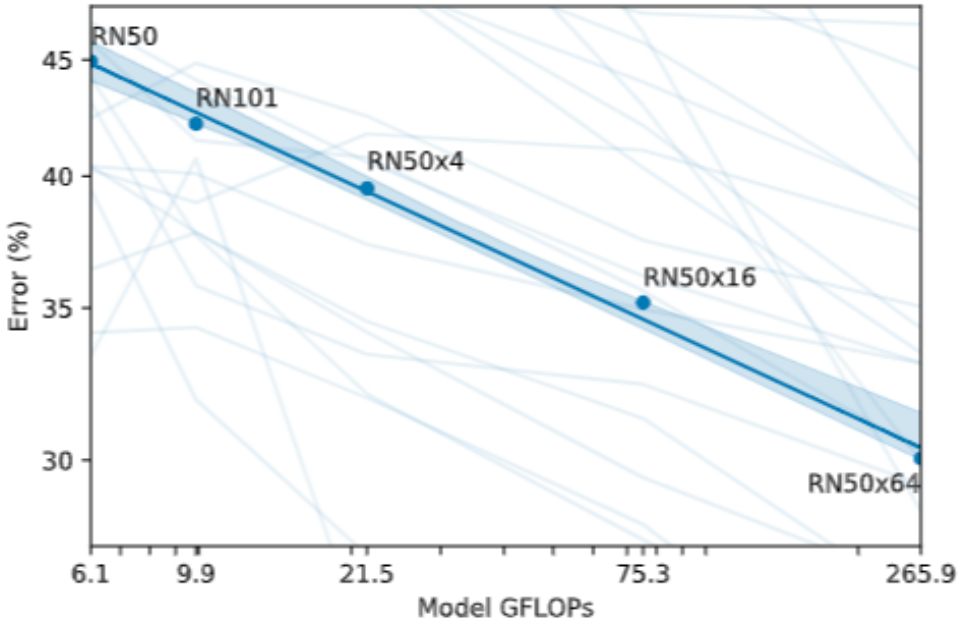


Figure 9. **Zero-shot CLIP performance scales smoothly as a function of model compute.** Across 39 evals on 36 different

Text, Image Encoder Scaling Method

- Image Encoder Scaling
  너비, 깊이, 해상도 모두 동일 비율로 증가시키는 스케일링
- Text Encoder Scaling
  ResNet의 너비 증가에 비례하여 Text Encoder의 width만 증가

Experiments

	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 classi-fication results.

Experiments

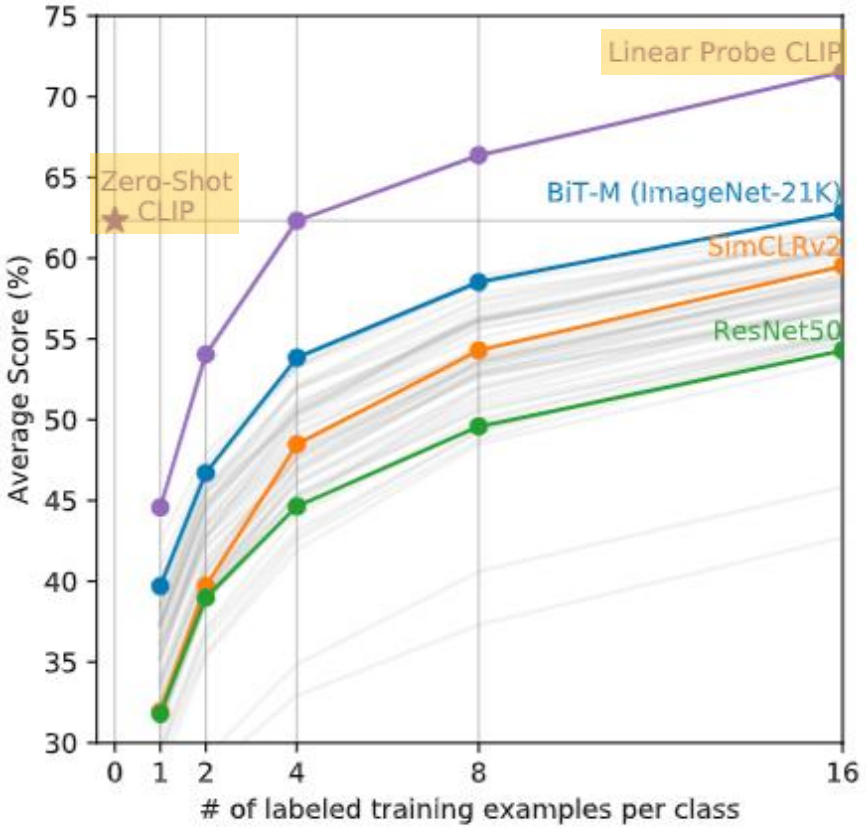


Figure 6. **Zero-shot CLIP outperforms few-shot linear probes.**

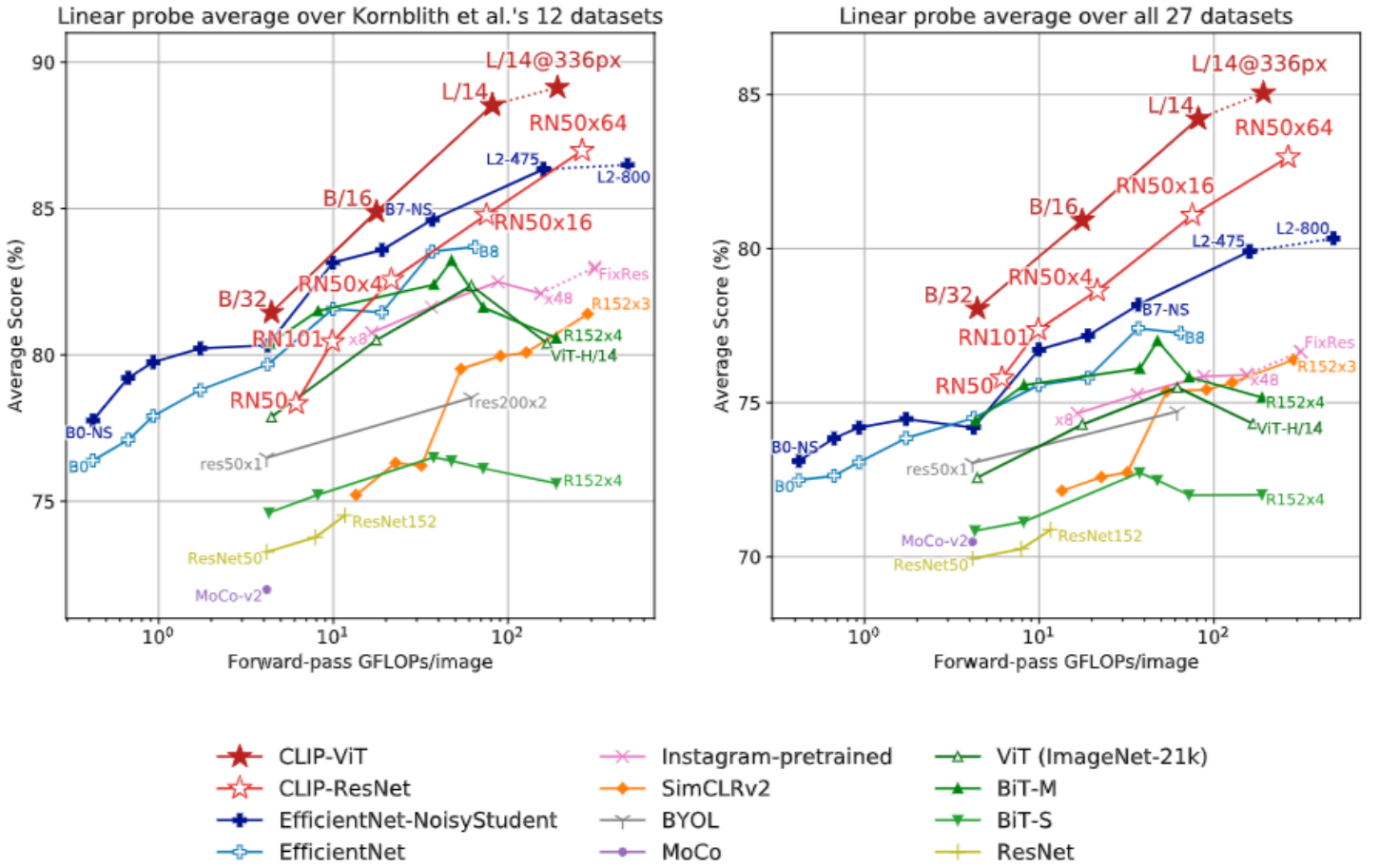


Figure 10. **Linear probe performance of CLIP models in comparison with state-of-the-art computer vision models**, including

Experiments

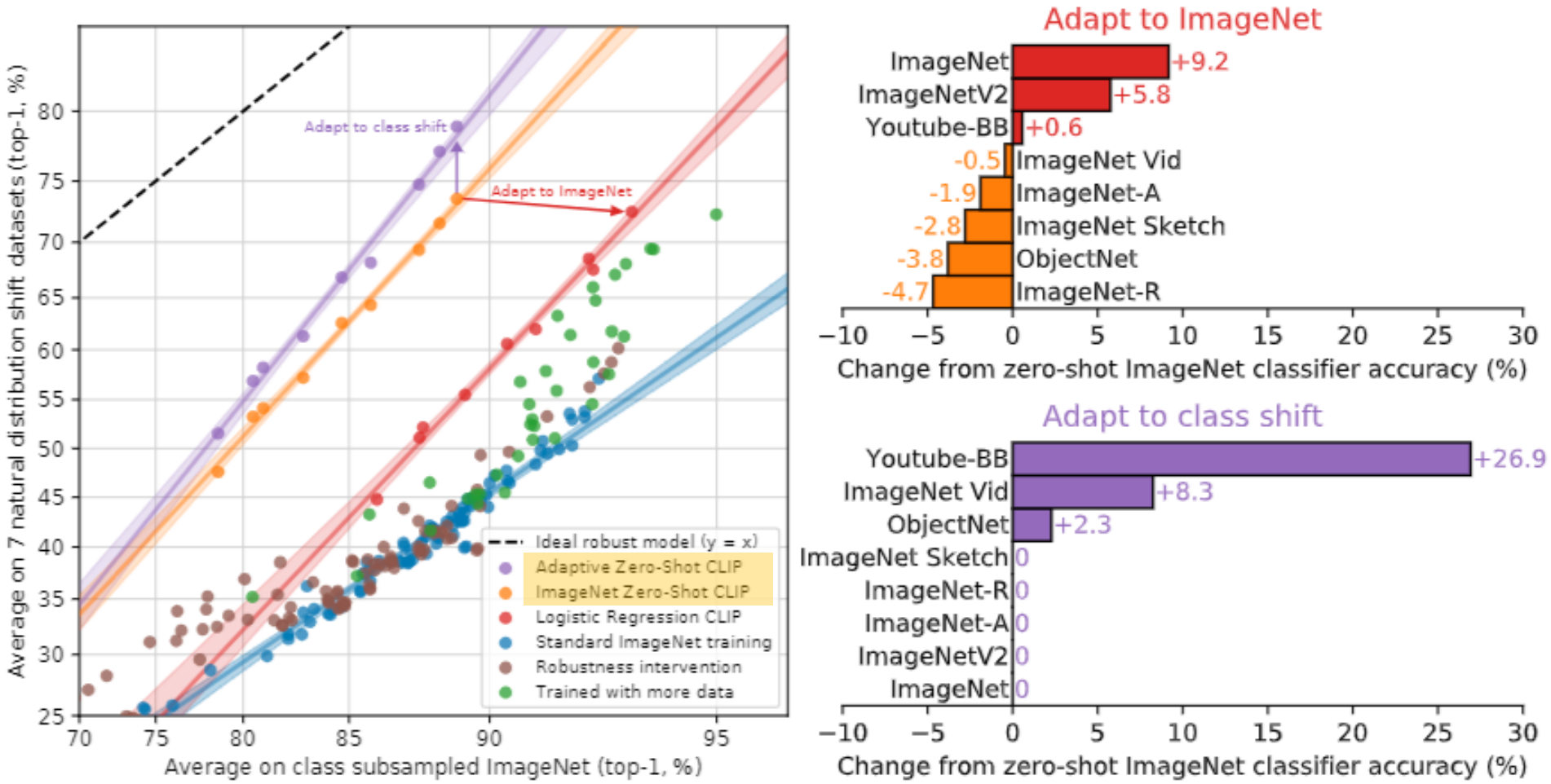


Figure 14. **While supervised adaptation to ImageNet increases ImageNet accuracy by 9.2%, it slightly reduces average robustness.**

Conclusions

Limitations

1. 모델 성능의 한계

- Zero-Shot 성능의 제약
  Zero-shot CLIP의 성능은 단순한 ResNet-50 기반 선형 분류기와 비슷하지만, SOTA 달성은 X
- 세부적인 분류에서 성능이 떨어짐
  물체 수를 세는 것과 같은 추상적이고 체계적인 작업에서 어려움을 겪음.
- Distribution Shift에 대한 한계
  CLIP은 여러 자연 이미지 분포에 대해 강건하지만, 진정한 분포 이동 데이터에는 일반화가 잘 되지 않음

2. 데이터 비효율성

- CLIP은 방대한 이미지-텍스트 페어 데이터를 사용해 훈련되었지만, 데이터 효율성이 낮음.
  예를 들어, CLIP 모델의 12.8억 이미지 학습은 초당 1장의 이미지를 처리한다고 가정했을 때 405년이 걸림.
  더 효율적인 데이터 사용을 위한 연구가 필요.

3. 유연성의 한계

- Zero-shot 에서만 강점, few-shot 최적화 X
  CLIP은 제로샷 학습에는 강점이 있지만, Few-Shot 학습에는 직접적으로 최적화되지 않음.
  제로샷에서 Few-Shot으로 전환할 때 성능이 기대만큼 향상되지 않음.

Conclusions

CLIP 은 사전 학습으로 다양한 task을 학습

-> natural language prompting를 활용하여 **Zero-Shot Transfer** 가능

CLIP 모델이 다양한 기존 데이터셋에 바로 적용 가능.

이 접근법은 많은 태스크에서 task-specific supervised models과 경쟁할 만한 성능을 보임.