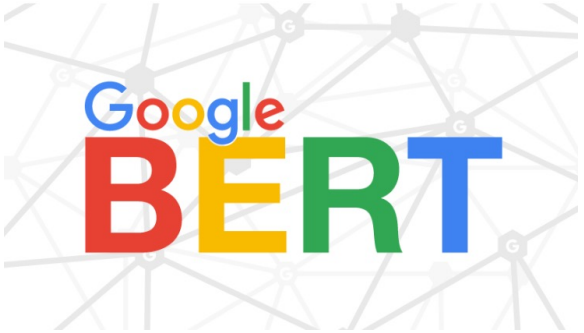
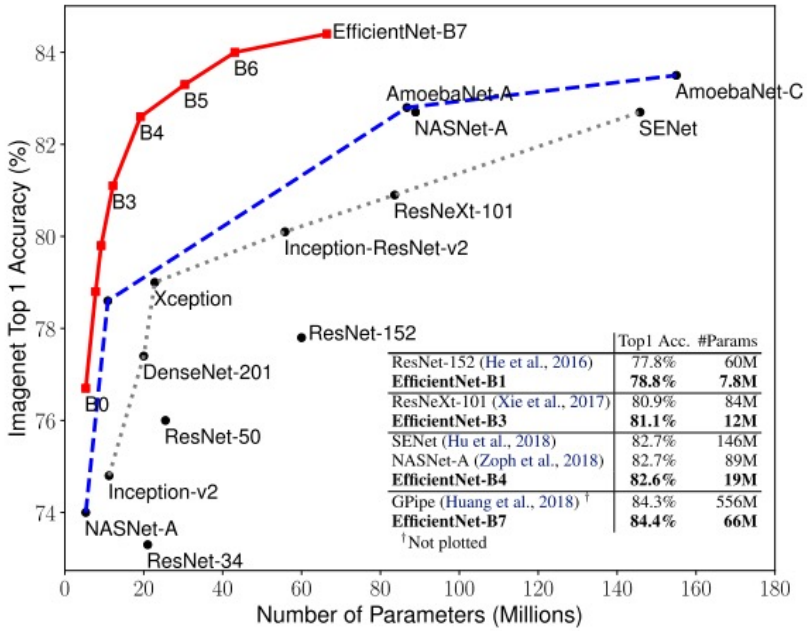
Learning Transferable Visual Models From Natural Language Supervision

Background



✓Pre-trained 방법은 NLP 분야에 있어 혁신적인 변화를 가져왔으며 zero-shot으로도 여러 Task에 적용할 수 있는 기회의 장을 열음

Background



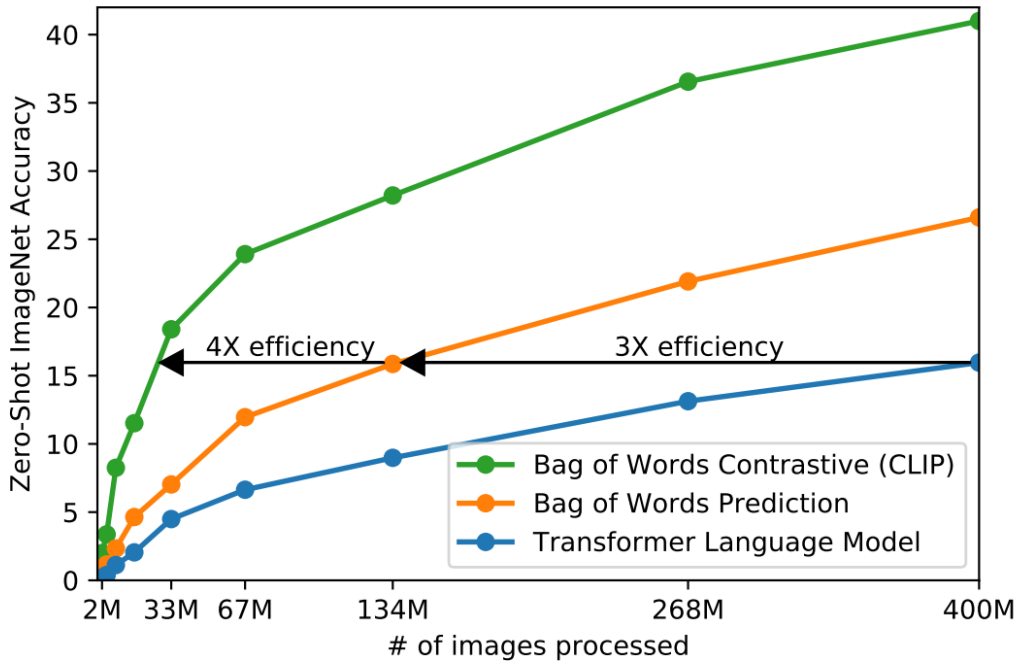
	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**78.8%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.1%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**82.6%**	**19M**
GPipe (Huang et al., 2018)†	84.3%	556M
EfficientNet-B7	**84.4%**	**66M**

†Not plotted

✓ Vision분야에서는 ImageNet과 같은 Labeled data에만 의존(Gold-label)

✓ 이미지에 대한 representation learning에서는 natural language supervision에 대한 시도가 부족하며 이는 아무래도 성능에서 매우 차이가 많이 나기 때문

CLIP results overview



✓그럼에도 불구하고 이를 사용하는 훈련은 사진에 대한 더 많은 정보를 포함할 수 있기에 범용성의 측면에서 이점을 얻을 수 있음, 즉 zero-shot Transfer 가능해짐

✓그러기에 우리는 CLIP이라는 모델을 만들어 범용성을 높이는 것이 목표

Approach

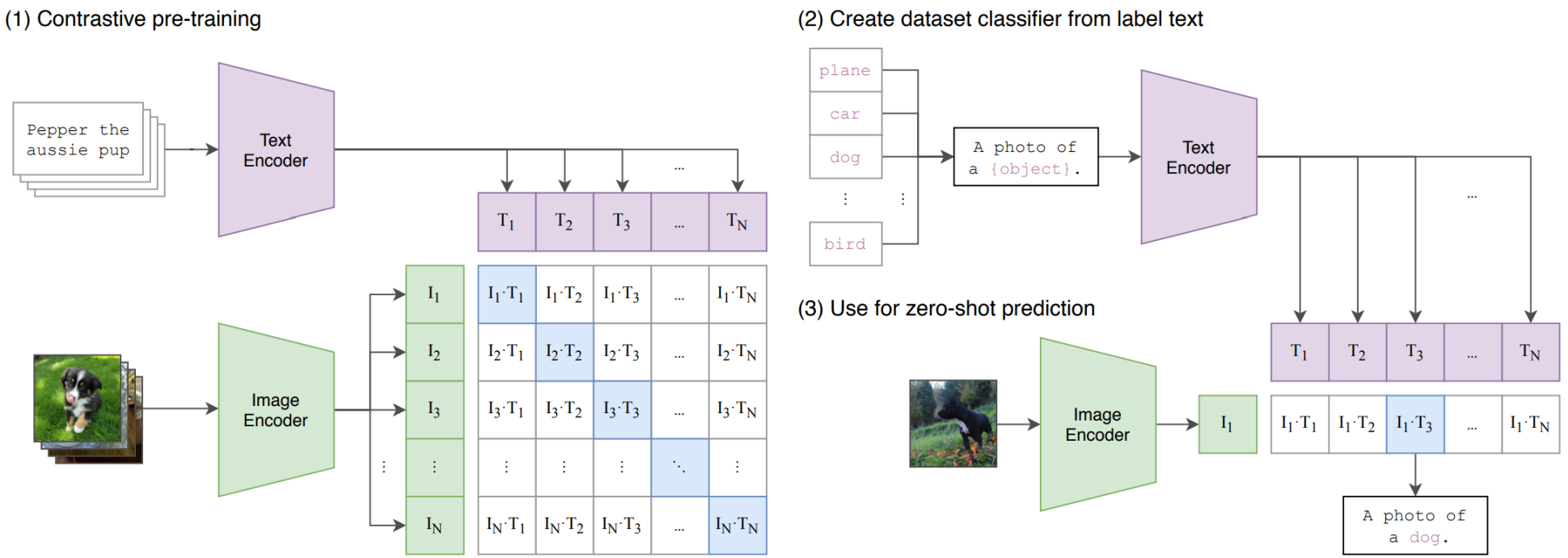
Natural Language Supervision

- ✓ Image-text pair를 통해 visual representation을 학습하는 것이 목표
- ✓ 일명 Gold label이라 불리는 class를 annotation 할 필요 없이 그냥 인터넷에서 크롤링
- ✓ 이러한 방식은 representation을 학습할 뿐만 아니라 언어 능력을 학습할 수 있음

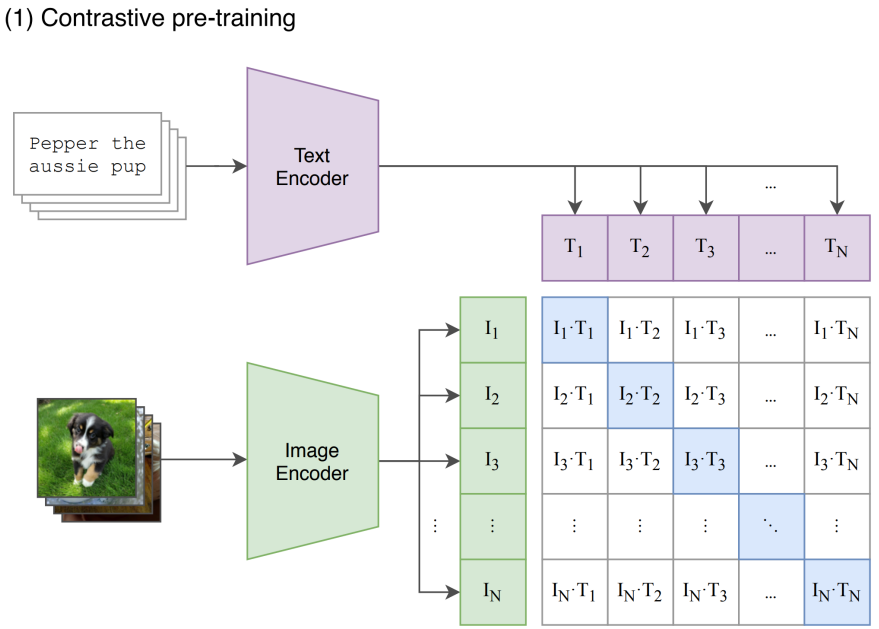
Creating a Sufficiently Large Dataset

- ✓ 인터넷에서 크롤링 한 400만개의 image-text pair
- ✓ 이는 GPT-2를 학습시킨 데이터셋의 양과 유사할 정도로 거대한 데이터셋
- ✓ 이를 WIT 일명 WebImageText Dataset이라 이름 붙힘

Architecture



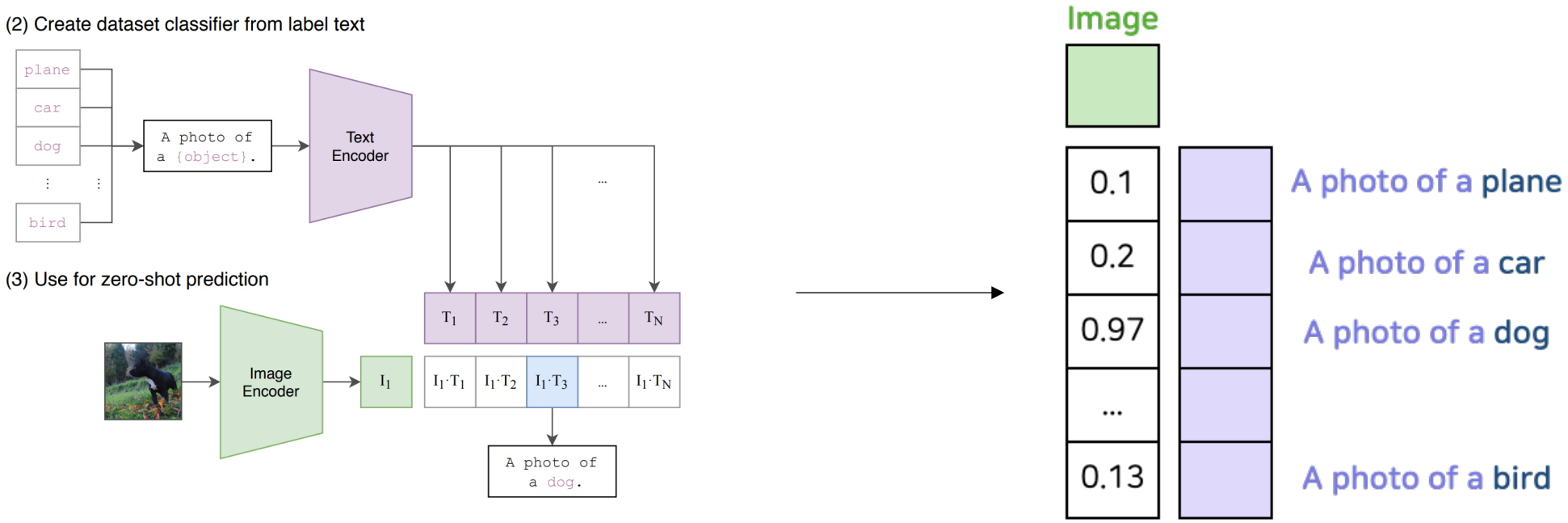
Architecture - Contrastive pre-training




```
# 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
```

- ✓ 매칭하는 데이터와는 가까워지게, 나머지는 멀어지도록 학습하는 방법
- ✓ N개의 Image Feature와 N개의 Text Feature의 조합
- ✓ Text Encoder는 Transformer, Vision Encoder는 5가지 종류의 ResNet, 3가지 종류의 Vit 실험

Architecture - Zero shot prediction



- ✓ 각각의 encoder를 활용해 N개의 Feature들을 추출(여기서 N개는 클래스의 종류)
- ✓ A photo of {object} 프롬프트를 활용해 표현하는 문장을 만들어 줌
- ✓ 마지막으로 코사인 유사도를 측정하고 가장 높은 값을 정답으로 출력

Experiments

Zero-shot Transfer

	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).

✓ 최초로 zero-shot Transfer를 시도한 Visual N-Grams와 비교

✓ 세가지 데이터 셋 모두 훌륭한 성능 개선이 있음을 확인 가능

Experiments

Zero-shot Transfer VS fully super-vised baseline

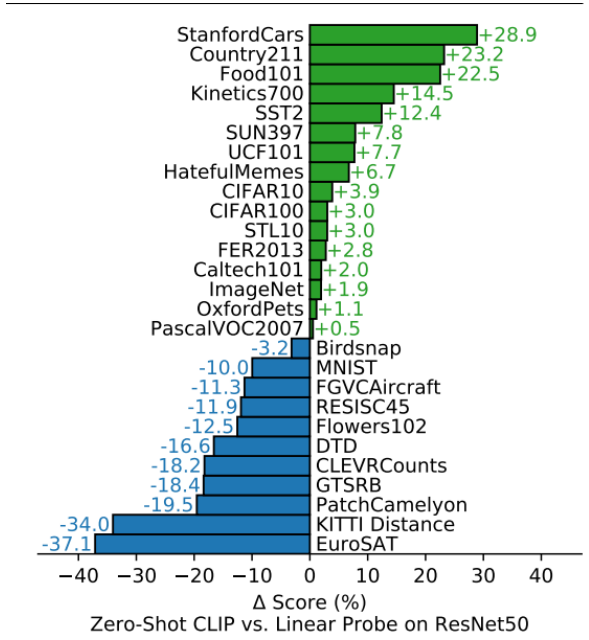


Figure 5. **Zero-shot CLIP is competitive with a fully super-vised baseline.** Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.

✓ 27개 데이터셋에 대한 ResNet-50과의 비교

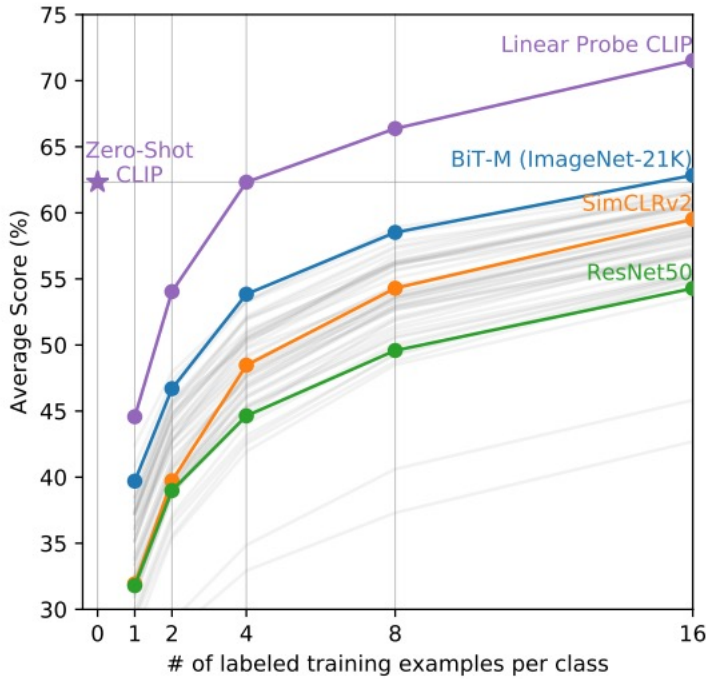
✓ 16개의 데이터 셋에서 베이스라인을 능가함을 확인할 수 있음

✓ 특히 STL10 Dataset은 Zero shot에도 불구하고 SOTA를 달성

✓ EuroSAT, RESISC45같은 매우 특수한 분야(인공위성)에는 약함을 확인

Experiments

Zero-shot Transfer VS Few-shot

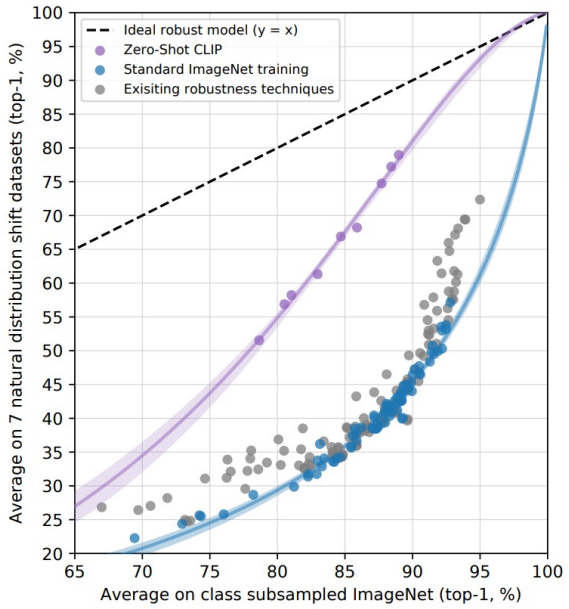


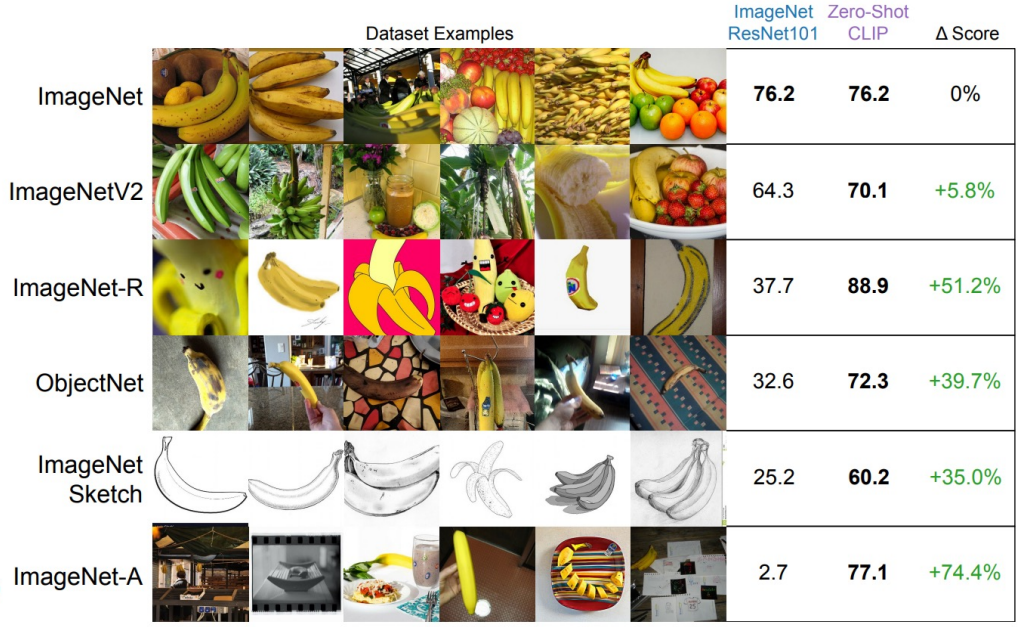
✓ Few-shot에서도 가장 훌륭한 성능을 보임

✓ 1shot에서는 오히려 성능이 떨어짐을 확인할 수 있음

Experiments

Robustness to natural distribution shift





✓ ImageNet과 유사한 데이터셋에서도 CLIP은 훌륭한 성능을 보임

✓ 이는 ResNet은 Generalization이 매우 약한 것을 확인할 수 있음

Experiments

Comparison to Human Performence

	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

✓ Zero-shot에서도 인간의 성능을 훨씬 뛰어넘음

✓ 재밌는 점은 인간은 one-shot과 two-shot의 차이가 나지 않음

Conclusion

limitation

✓인터넷에 표현된 Bias가 그대로 학습될 우려 존재

✓대규모 데이터셋으로 인한 고계산비용과 자원 소모

✓매우 특수된 분야에서는 일반화가 어려움

✓분류에는 효과적일 수 있지만 Caption같은 분야에서는 성능이 떨어짐

Conclusion

✓NLP의 Pre-trained를 다른 Domain으로 훌륭하게 이식한 사례

✓더 다양한 natural language prompting을 학습한다면 더욱 더 개선된 성능을 보일 것으로 기대

감사합니다 ☺