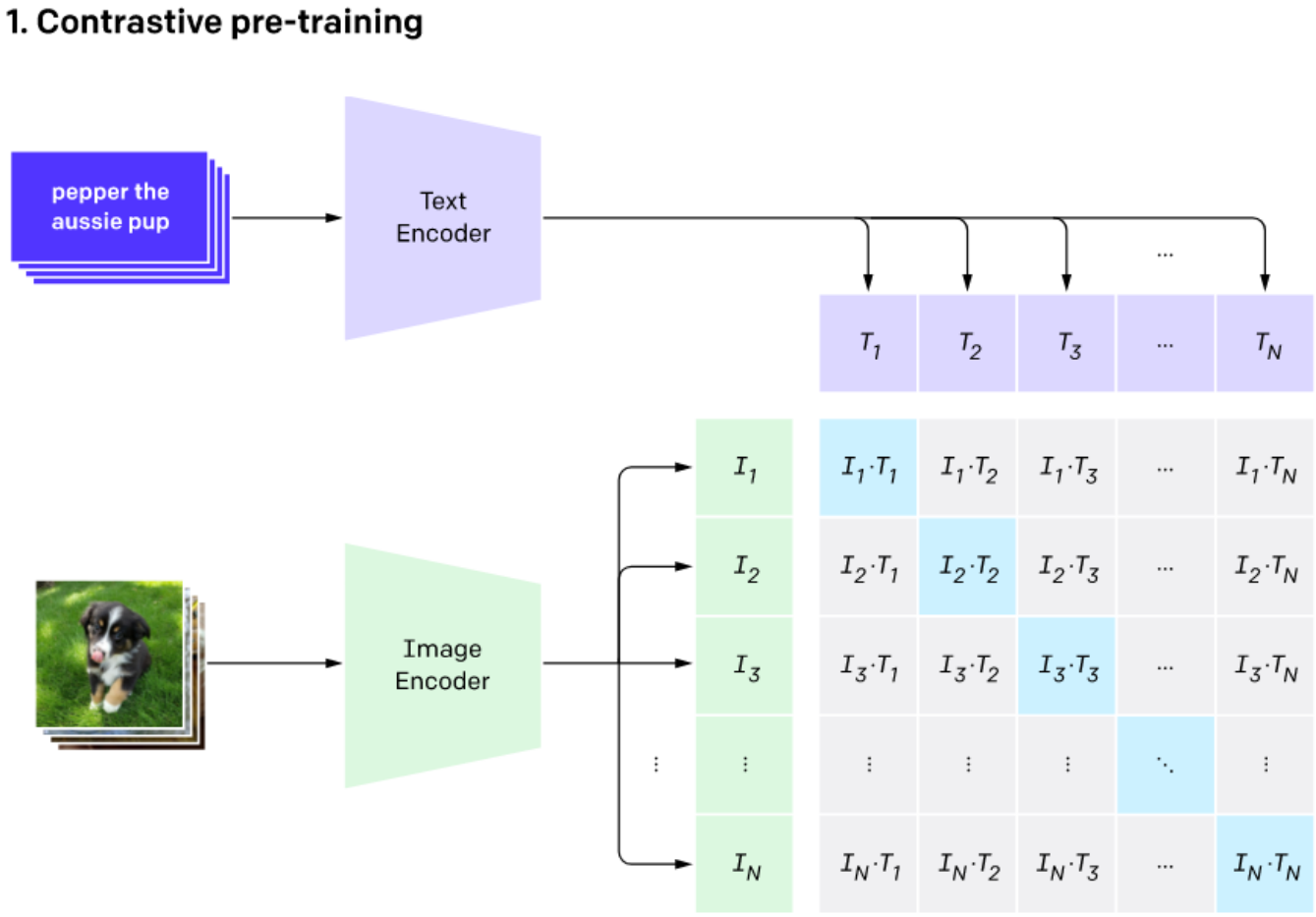
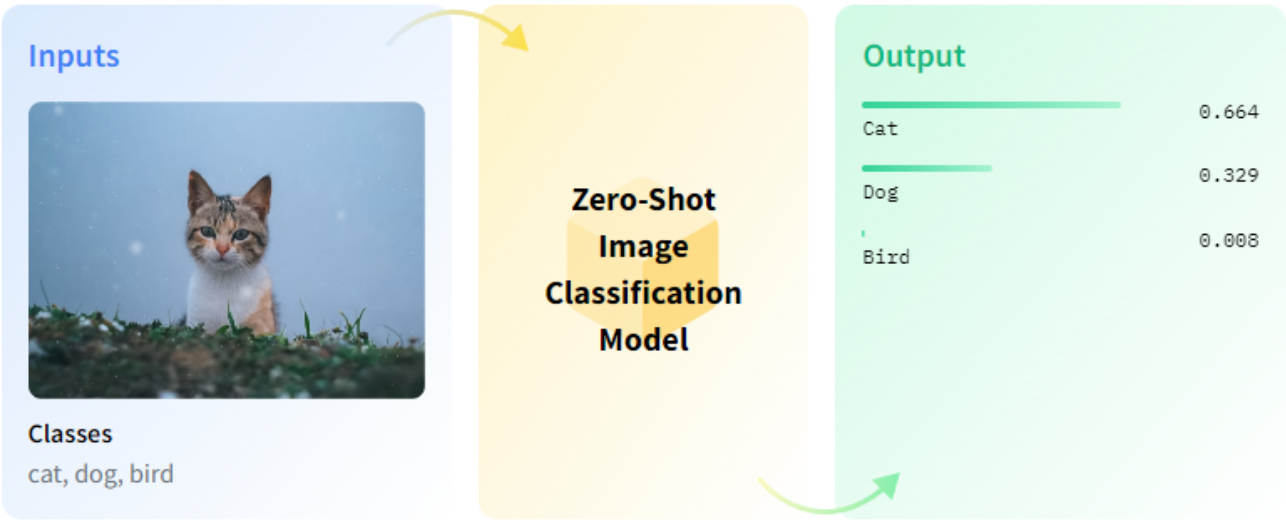


[CLIP] Learning Transferable Visual Models From Natural Language Supervision

*본 템플릿은 DSBA 연구실 이유경 박사과정의 템플릿을 토대로 하고 있습니다.

1. 논문이 다루는 Task



Task: Zero-Shot Image Classification

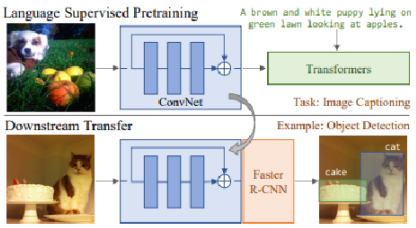
- Input: Image
- Output:Text
- Zero-shot : 어떻게 하면 Training set에 없는 보지 않은 데이터를 예측할까?, 어떻게 하면 데이터에 없는 새로운 클래스를 분류할까? 기존의 방식은 고양이와 개를 분류하는 모델을 훈련시키면 이 모델은 고양이와 개만 분류한다. 하지만 Zero-shot의 경우 고양이와 개를 분류하는 모델을 훈련시켜도 호랑이를 넣었을때 호랑이를 분류해낼 수 있다.

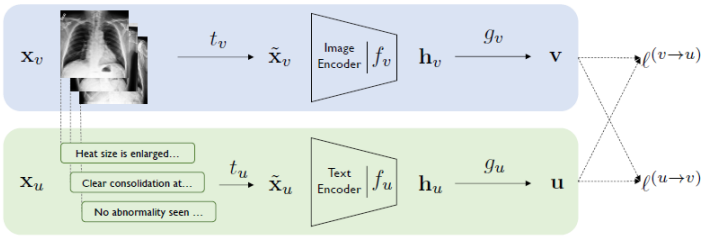
2. 기존 연구 한계

2-1. Vision & NLP의 훈련 방법의 차이

최근 State-of-the-art Vision 모델들은 고정된 카테고리를 통해 예측하도록 훈련이 진행된다. NLP의 경우 GPT-3와 같은 모델은 fine-tuning 데이터가 필요없이도 많은 Task에서 경쟁력을 유지한다. 하지만 Vision에서는 아직까지 관행적으로 ImageNet,JFT-300M 등과 같은 곳에서 훈련시키고 fine-tuning 데이터로 학습시키는게 일반화되어 있다.

2-2. 기존 Natural Language Supervision



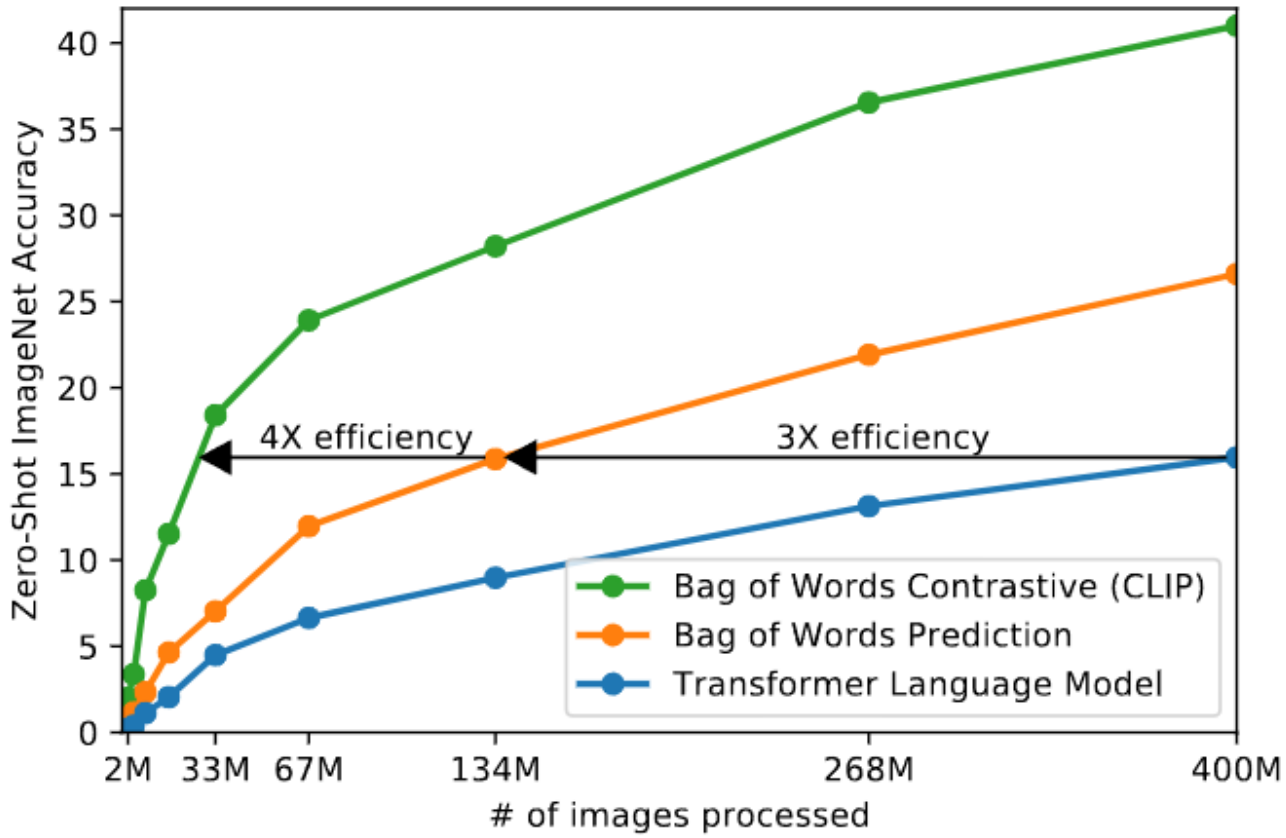


Natural Language Supervision는 이미지 학습을 위해 자연어 감독을 진행하는 것이다. 현재 논문에서는 위와 같은 Natural Language Supervision을 사용하며 지금까지의 연구는 VirTex,Con-VIRT모델이 있었다. 하지만 두 작업 모두 1000, 18291개의 클래스로 학습되어 zero-shot에 제한사항이 존재한다.

2-3. Dummy Dataset

기존 작업에서는 주로 사용되는 데이터 셋은 MS-COCO, Visual Genome, YFCC100M을 주로 사용해왔다. MS-COCO와 Visual Genome은 고품질의 데이터 셋이지만 갯수가 10만개로 적으며 YFCC100M은 1억개의 데이터가 존재하지만 품질이 들쑥날쑥하다. 이를 영어로 제목이 달려있거나 설명이 있는 이미지만 걸렀을 경우 약 6배가 감소하여 1500만장의 사진만 남게되어 ImageNet의 데이터 크기와 동일해지는 아쉬움이 존재한다. 만약 레이블을 지정해서 데이터셋(ImageNet)을 만들경우에는 OpenAI에 따르면 14M개의 이미지에 22000개의 label을 지정하는 경우 25000명의 인원이 소요된다고 한다.

2-4. Pre-Training Method



ImageNet 단 1000개를 훈련시키는데 필요한 컴퓨팅 파워는 ResNeXt101-32x48d을 훈련시키는데 19개의 GPU가 필요하였으며 NoisyStudent EfficientNet-L2를 교육시키는데는 33개의 TPUv3 들어갔다. 이는 단순히 1000개의 클래스만을 예측하는데 필요한 컴퓨팅 파워가 굉장히 비효율적이라고 주장하고 있다. 본 논문에서는 VirTex와 유사하게 6300만 파라미터의 모델을 처음부터 교육하였는데 이미지 캡션을 예측하기 위하여 CNN과 텍스트를 인코딩하는 BoW를 훈련시키는데 효율성이 기존 BoW을 교육하는 것에 비해 약 3배의 효율성이 필요하다. 이때 파란색은 VirTex의 트랜스포머이며 주황색은 BoW, 초록색은 논문에서 주장하는 CLIP모델이다.

3. 제안 방법론

3-1. Natural Language Supervision & Zero-shot Classification

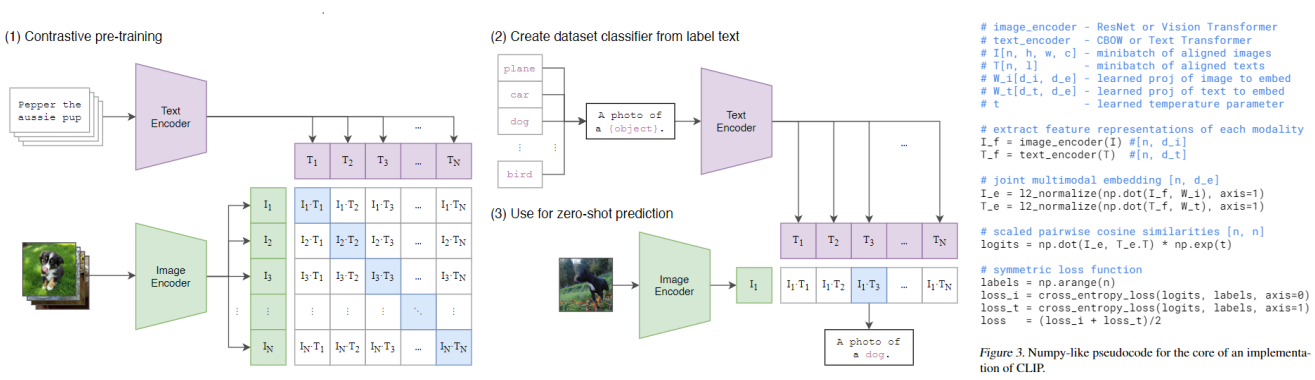
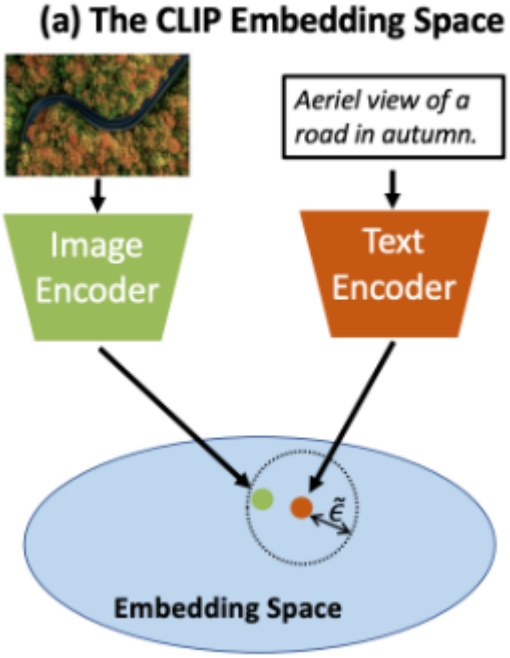


Figure 3. Numpy-like pseudocode for the core of an implementation of CLIP.

논문에서는 위와 같은 방법을 주장하였다. 훈련 시에는 이미지,텍스트 인코더를 함께 훈련시키면서 각 훈련 샘플들을 같은 차원에 맵핑하였다. 위 그림의 예시에서 $$[T_1,T_2,T_3\cdots T_N], [I_1,I_2,I_3\cdots I_N]_{N∈\R^{batchsize}}$$ 일때 $$T$$와 $$I$$는 각각 텍스트와 이미지 Pair를 나타낸다. 이 둘을 행렬곱을 수행하고 코사인 유사도를 계산하게 된다. 이때 대각선의 행렬은 각 이미지,텍스트의 맞는 쌍이므로 **Positive Pair**가 되고 나머지는 **Negative Pair**가 된다. 즉 $$N$$개의 Positive Pair의 코사인 유사도를 최대화 하며 나머지 $$N^2-N$$ Negative Pair의 코사인 유사도는 최소화시키며 훈련을 진행하게 된다.

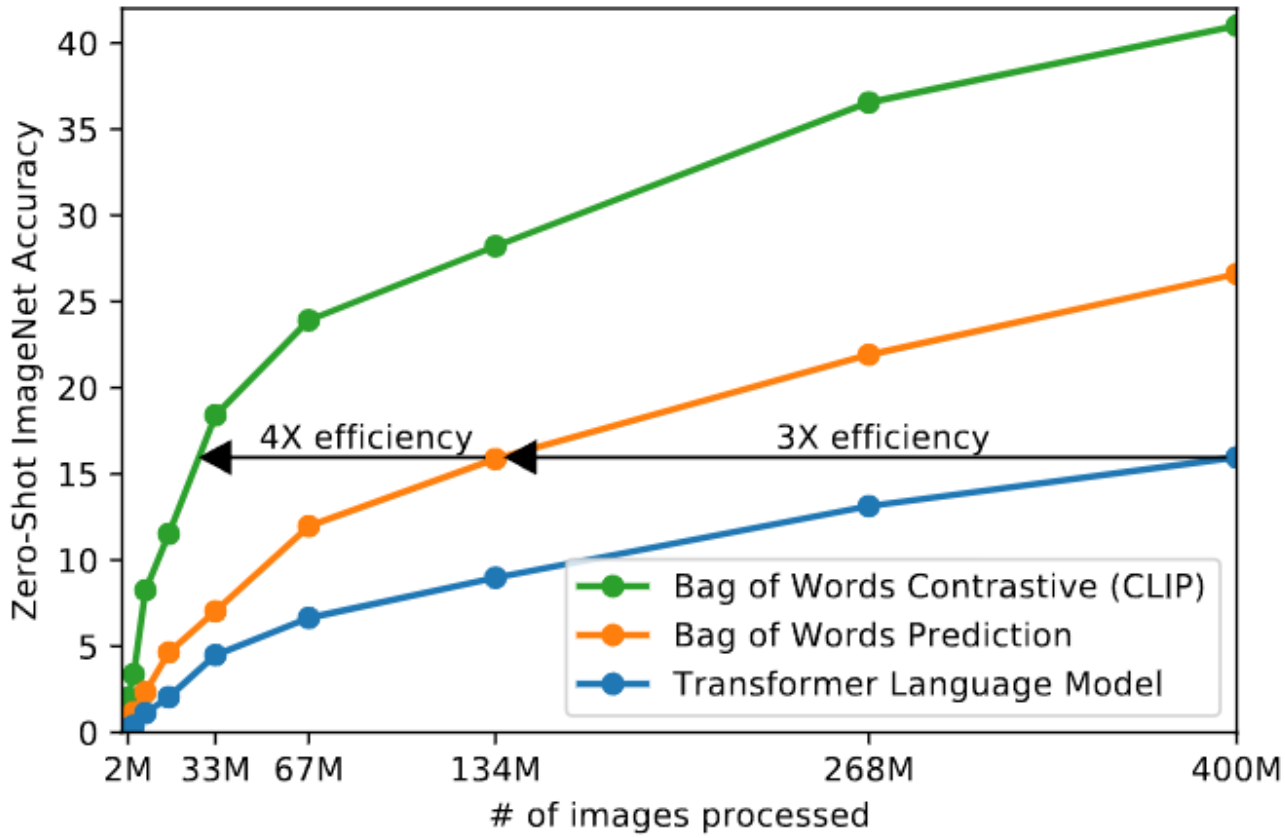
추론 과정에서는 한 개의 이미지가 입력되고 클래스를 직접 여러개를 지정할 수 있게되는데 이미지,텍스트 모두 인코더를 통과하게 되면 $$I_1$$의 차원은 $$I_1∈\R^{1*dim}$$ 이 되며 $$[T_1,T_2,T_3\cdots T_N]$$의 경우는 $$T_N∈\R^{N*dim}$$의 차원을 갖게되어 Transpose를 한 후 행렬곱을 수행하게 되면 $$I_1 · T_N.T ∈\R^{1*N}$$의 차원을 가지게 되며 위 그림처럼 그 중 가장 높은 값을 가지는 텍스트를 뽑아낸다.



이런식으로 Contrastive Learning 수행하게 될경우 같은 차원에 있는 같은 이미지와 텍스트끼리 페어는 가깝게 위치하게 되고 다른 페어끼리는 멀리 위치하게 된다.

이로써 Text embedding을 클래스로 사용하며 사용자가 직접 클래스를 정의하며 기존의 Natural Language Supervision 모델들과는 달리 Zero-shot을 수행할 수 있게 되었다.

3-2. Selecting an Efficient Pre-Training Method



기존의 Pre-Training Method은 많은 컴퓨팅 비용이 소모된다. CLIP같은 경우에 학습에 사용된 이미지가 약 4억개가 소모된다고 논문에서는 나와있다. 위의 초록을 제외한 그래프 처럼 학습 목표를 가져갈 경우 많은 컴퓨팅 자원이 소모됨으로 CLIP은 위의 3-1의 설명처럼 Contrastive Learning을 수행하게 되었다고 한다. 위 그래프에서는 약 기존의 BoW방식모다 4배의 효율성을 자랑한다.

3-3. Choosing and Scaling a Model

논문에서는 이미지 인코더와 텍스트 인코더를 각각 둠으로써 여러 모델 테스트를 진행하였다. 이미지 인코더는 5 ResNets and 3 Vision Transformers를 선택하였다.

이미지 인코더

- Resnet-50, ResNet-101, ResNet-50x4, ResNet-50x16, ResNet-50x64(xN은 EfficientNet-style로 모델 스케일링을 진행)
- ViT-B/32, ViT-B/16, ViT-L/14

텍스트 인코더

- Sparse Transforemr(이미지 인코더 크기에 따른 모델의 너비만 조정)

위와 같은 모델로 실험을 진행하였으며 세부사항은 아래와 같다.

Training detail

- epoch : 32
- optimizer : Adam
- schedule : cosine schedule
- learning rate : 논문에서는 감각적으로 진행하였다고 한다.
- batch size : 32768

- others : weight decay regularization, Mixed-precision

3-4. Creating a Sufficiently Large Dataset

2-3의 설명처럼 데이터셋의 제한사항으로 CLIP에서는 데이터를 새로 수집하였다고 한다. 새로 수집된 데이터는 인터넷에서 공개적으로 수집할 수 있는 이미지,텍스트 쌍 4억개이다. 이때 조금 더 광범위 하게 수집하기 위해 쿼리는 50만개를 사용하였으며 데이터 균형을 위해 각 쿼리당 대략 최대 2만개의 이미지,텍스트 페어를 수집하였다. 사용된 총 단어의 개수는 GPT-2와 비슷하며 WebImageText로 불린다.

✨Contribution

- Natural Language Supervision을 수행하여 보다 심도있는 멀티모달을 구성하였다.
- Contrastive Learning을 목적으로 훈련시켜 학습의 효율성을 기존보다 4~7배 증가 시켰다.
- Contrastive Learning으로 인해 Zero-shot Classification을 효과적으로 적용하였으며 모델 구조 또한 간단하게 적용하였다.
- 크라우드 소싱이나 직접 레이블 보다 인터넷에 있는 데이터로 대규모 데이터 셋을 만들어 학습을 진행하여 대규모 데이터 셋을 운용가능하게끔 하였다..

4. 실험 및 결과(Zero-Shot Transfer)

4-1. INITIAL COMPARISON TO VISUAL N-GRAMS

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

Dataset

- aYahoo, ImageNet, SUN

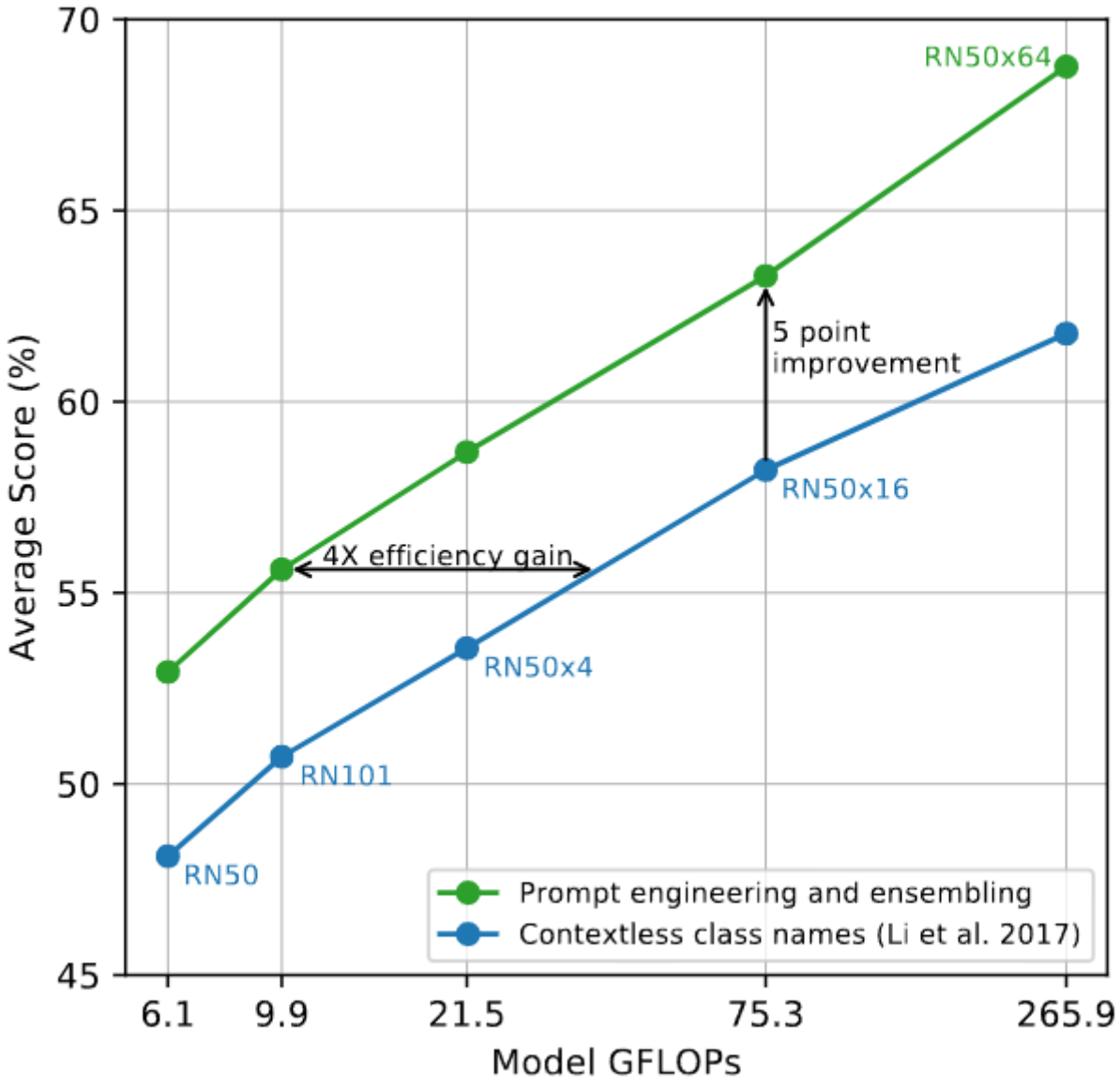
Baseline

- CLIP, Visual N-Grams (논문에서는 가능한 비교대상이 Visual N-Grams밖에 없었으며 성능 비교의 목적은 아니라고 주장하였다. 10배많은 데이터와 예측에는 100배의 컴퓨팅 파워, 훈련에는 1000배의 컴퓨팅 파워가 소모되었으며 기존에는 없던 Transformer같은 아키텍처를 선택하였기 때문이다.)

결과

- 메인 결과 Visual N-Grams에 비해 CLIP 모델은 3가지의 데이터 셋에서 높은 성능 향상을 보여주었다. 특히 ImageNet에서 Visual N-Grams이 11.5%인 반면 CLIP은 76.2%를 기록하였다.
- Analysis CLIP은 ImageNet에서 ResNet-50과 성능이 동일하였으며 Visual N-Grams은 ImageNet 가중치에 초기 훈련된 반면 CLIP은 스크래치부터 훈련되었다. 전체적으로 성능이 올라갔다.

4-2. PROMPT ENGINEERING AND ENSEMBLING



Dataset

- 36개의 이미지 데이터 셋의 평균

Baseline

- ResNet50~50x16, ResNet101

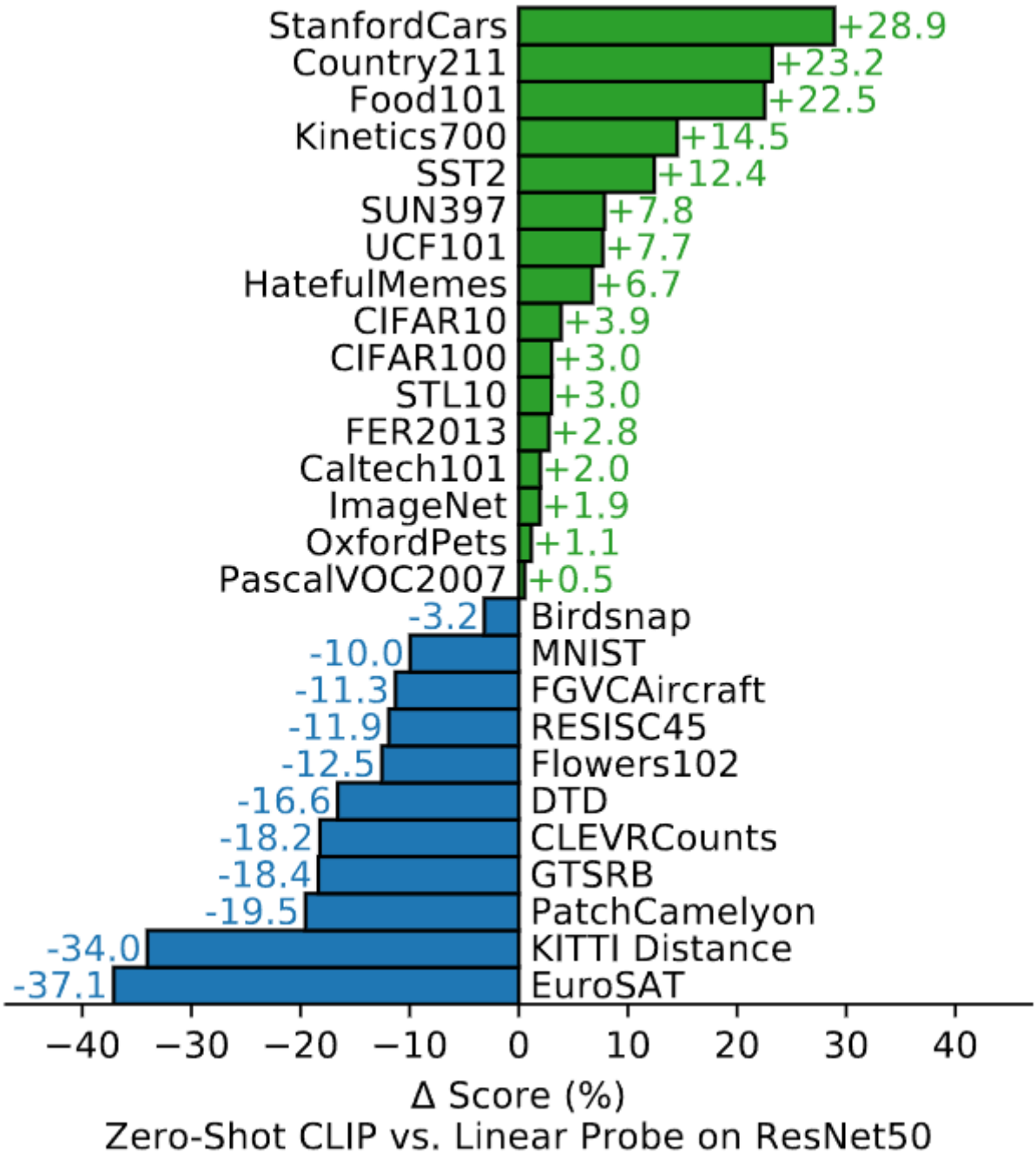
결과

- 메인결과 GPT처럼 프롬프트 엔지니어링을 수행하고 앙상블을 할 경우 약 5점의 점수 향상이 존재하였다. 클래스를 직접 지정하는 Zero-shot인 만큼 프롬프트 엔지니어링이 가능하다.
- Analysis
- PROMPT ENGINEERING ImageNet에서 클래스 cranes인 2가지로 분류된다고 한다. 건설용 cranes과 두루미과를 cranes이라고 지칭한다. 또한 Oxford-IIIT Pet 데이터 셋의 경우도 boxer를 개 품종과 운동 선수를 덜 훈련된 텍스트 인코더의 경우 헷갈린다고 나와있다. 이러한 현상을 해결하기 위해 'A photo of a {label}' 프롬프트를 주어 이것만으로도 ImageNet 정확도가 1.3% 향상되었다.

더욱 범주를 세밀화 하게 될경우 데이터 셋마다 'A photo of a {label}, a type of pet' 이런식으로 프롬프트를 줄 경우 더욱 효과적이라고 주장한다.

- ENSEMBLING 앙상블의 경우 'A photo of a big {label}', 'A photo of a small{label}'로 프롬프트를 준다고 한다. 이때 하드보팅, 소프트보팅처럼 기존의 방식이 아닌 임베딩 스페이스에서의 앙상블을 진행하여 컴퓨팅 파워는 동일하게 소모된다고 주장하여 높은 효율성을 가지고 있다 주장한다.
  위의 그래프에서는 80개의 다른 프롬프트를 사용하여 앙상블을 진행한 결과 단일 프롬프트 모델에 비해 ImageNet에서 3.5%의 성능 향상 효과가 있었다.

4-3. ANALYSIS OF ZERO-SHOT CLIP PERFORMANCE



Dataset

- 27개의 이미지 데이터 셋

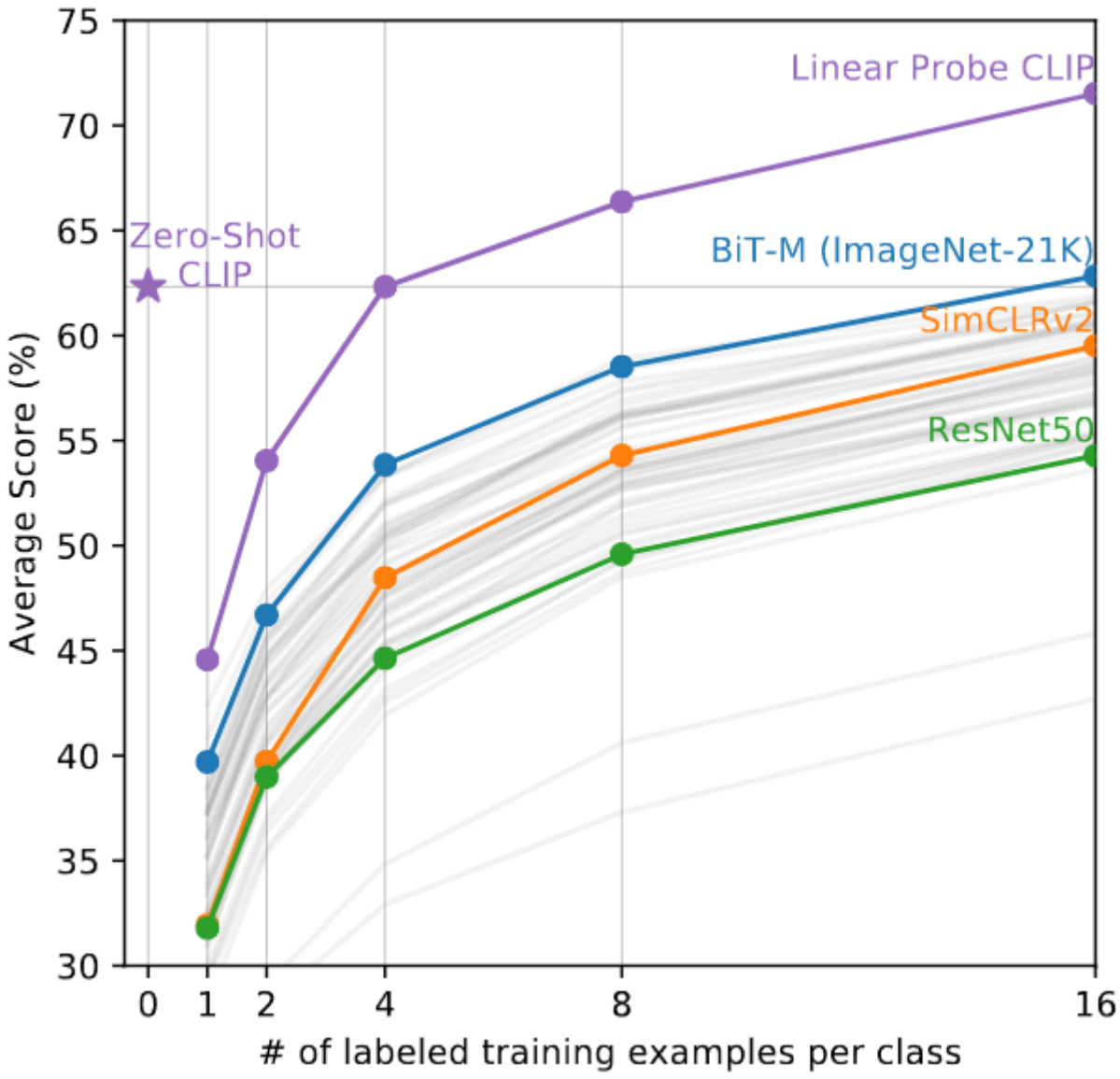
Baseline

- Zero-shot CLIP, ResNet50

결과

- 메인결과 Zero-shot CLIP은 ResNet50과 비교하여 위와 같이 성능의 차이가 존재한다.
- Analysis ImageNet과 CIFAR10 같은 '일반'적인 범주의 데이터 셋의 같은 경우는 점수가 비슷하지만 Kinetics700과 같은 동작을 인식하는 비디오 데이터 셋의 같은 경우는 조금 더 높은 성능을 달성하게 된다. 논문에서는 시각적 개념을 동사와 같이 텍스트로 감독하여 비교적 높은 성능을 달성하였다고 주장한다. STL10 같은 경우에는 기존에 존재하지 않는 SOTA 성능을 달성하였다.
  EuroSAT같은 위성이미지, 독일 교통 표지판 분류 같은 데이터의 경우는 많이 감소하게 되는데 논문에서는 조금 더 세부적이고 복잡한 이미지의 경우 CLIP이 분류하기가 힘들며 사람 또한 이러한 작업에서는 일부분은 잘하지만 대부분 약세인것과 동일하다고 주장하였다. 마치 지금 당장 위성 이미지 분류를 하라면 못하는 것과 같이 말이다.

4-4. Zero-shot vs Few shot



Dataset

- 27개의 이미지 데이터 셋

Baseline

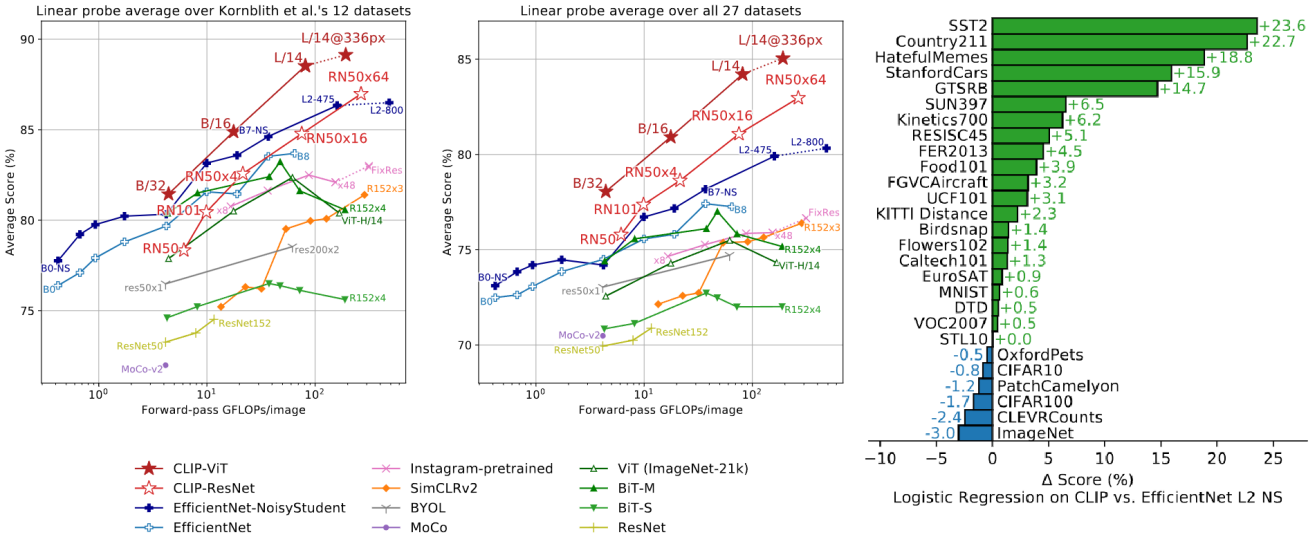
- Zero-Shot CLIP, BiT-M, CimCLRv2, ResNet50

결과

- 메인결과 Zero-Shot에 Linear Layer를 추가하여 학습하였다 이때 X축은 클래스당 학습 이미지의 수인 N-Shot이며 Y축은 Average Score이다. 결론적으로 Zero-Shot은 4-Shot과 성능이 동일하였다.
- Analysis 중요한 점으로는 논문에서는 One-Shot이 Zero-Shot보다 높을 거라고 예상하였지만 오히려 낮은 결과를 나타냈다. 논문에서는 그 이유로 CLIP 모델이 자체만으로 이를 잘 분류할 수 있지만 이러한 임베딩이 존재하여도 Linear Layer를 스크래치부터 학습시키다 보니 각 클래스당 예제가 부족하여 성능이 낮아지는것을 확인 할 수 있다. 하지만 샘플이 늘어날수록 역시 예측을 잘하게 된다.

5. 실험 및 결과(Representation Learning)

5-1. CLIP Representation Learning



Dataset

- 이미지 편향을 최소화 시킨 Kornblith et al.(2019) 12개의 이미지 데이터 셋
- 27개의 이미지 데이터 셋

Baseline

- CLIP-ViT, CLIP-ResNet
- EfficientNet, EfficientNet-NoisyStudent
- BiT-M, BiT-S
- ViT(ImageNet-21k), Resnet, MoCo, BYOL, SimCLRv2, Instagram-pretrained

결과

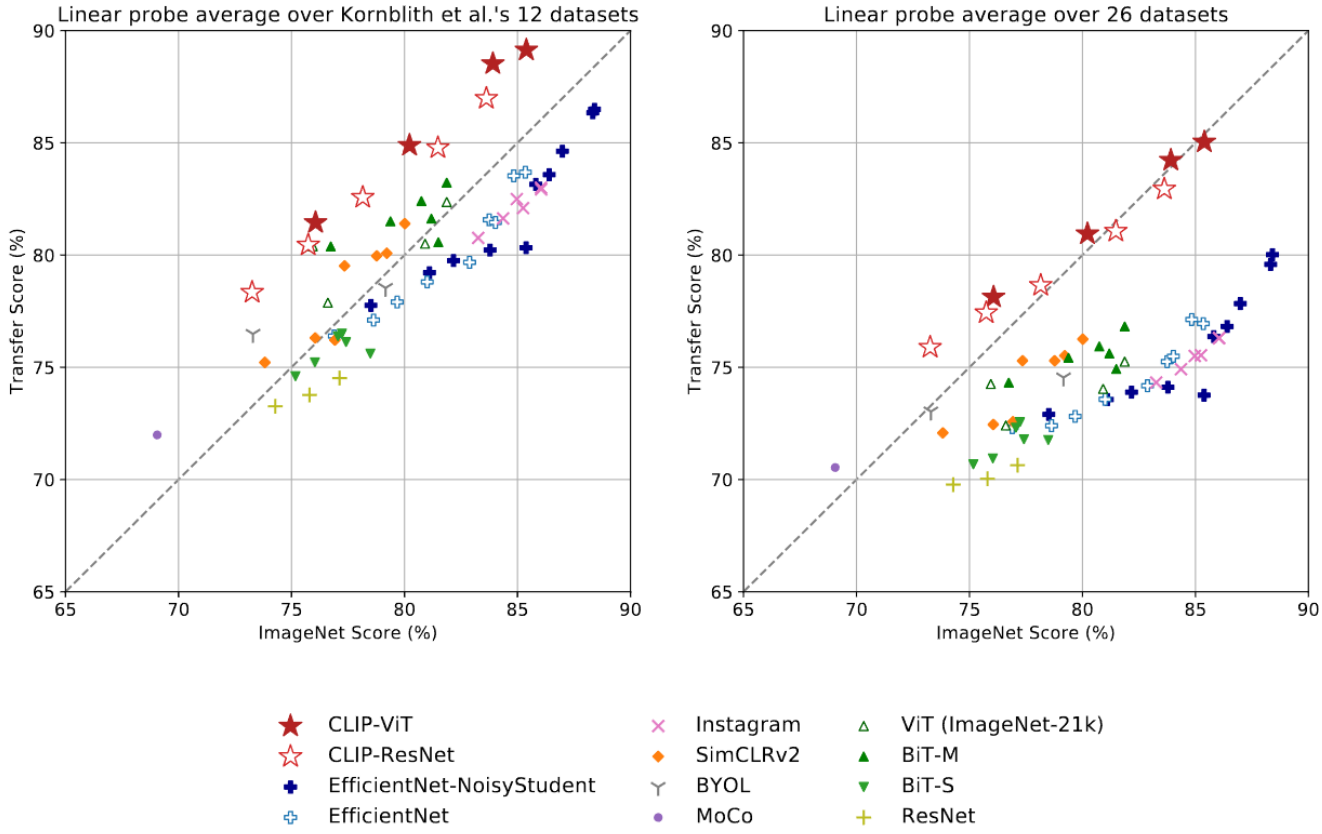
- 메인결과 이 실험에서는 Linear Layer을 통해 CLIP의 이미지 표현 성능을다른 모델들과의 성능을 측정한다. 이미지의 feature을 각 모델 별로 뽑아낸 후 이를 Linear Layer를 통해 비교한다. 그래프를 보면 현재 실험에서는 두 데이테 셋 모두 SOTA를 뽑아내 다른 모델보다 좋은 이미지 feature을 뽑아내는 것을 볼 수 있다.

- Analysis 분석으로는 기존의 SOTA였던 EfficientNet-NoisyStudentL2를 점수와 컴퓨팅 효율성에 대해서 보다 높았다. CLIP의 백본으로 사용된 ResNet과 ViT의 경우 기본적으로 ResNet은 CLIP논문에서 자체 설계한 ResNet이긴 하지만 비교적 ResNet이 ViT보다 컴퓨터 효율성 그리고 Score도 높았다. 역시 ViT의 논문처럼 대규모 데이터셋에선 이런 트랜스포머 구조가 Conv 기반보다 높은 성능을 가진다는 것을 입증하였다. 기존의 SOTA였던 EfficientNet L2 NS과의 비교에서도 비교적 21개의 데이터 셋에서 조금 더 높은 성능을 기록하였다. 전체적인 상세 성능은 아래의 엄청난 실험에 기록되어 있다.

		Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech101	Flowers	MNIST	FER2013	STL10*	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	Kinetics700	CLEVR	HatefulMemes	SST	ImageNet
LM RN50		81.3	82.8	61.7	44.2	69.6	74.9	44.9	85.5	71.5	82.8	85.5	91.1	96.6	60.1	95.3	93.4	84.0	73.8	70.2	19.0	82.9	76.4	51.9	51.2	65.2	76.8	65.2
CLIP-RN	50	86.4	88.7	70.3	56.4	73.3	78.3	49.1	87.1	76.4	88.2	89.6	96.1	98.3	64.2	96.6	95.2	87.5	82.4	70.2	25.3	82.7	81.6	57.2	53.6	65.7	72.6	73.3
	101	88.9	91.1	73.5	58.6	75.1	84.0	50.7	88.0	76.3	91.0	92.0	96.4	98.4	65.2	97.8	95.9	89.3	82.4	**73.6**	26.6	82.8	84.0	60.3	50.3	68.2	73.3	75.7
	50x4	91.3	90.5	73.0	65.7	77.0	85.9	57.3	88.4	79.5	91.9	92.5	97.8	98.5	68.1	97.8	96.4	89.7	85.5	59.4	30.3	83.0	85.7	62.6	52.5	68.0	76.6	78.2
	50x16	93.3	92.2	74.9	72.8	79.2	88.7	62.7	**89.0**	79.1	93.5	93.7	98.3	**98.9**	68.7	98.6	97.0	91.4	89.0	69.2	34.8	83.5	88.0	66.3	53.8	71.1	**80.0**	81.5
	50x64	94.8	94.1	78.6	77.2	81.1	90.5	67.7	**88.9**	**82.0**	94.5	95.4	98.9	**98.9**	**71.3**	99.1	97.1	92.8	90.2	69.2	40.7	83.7	89.5	69.1	55.0	**75.0**	**81.2**	83.6
CLIP-ViT	B/32	88.8	95.1	80.5	58.5	76.6	81.8	52.0	87.7	76.5	90.0	93.0	96.9	**99.0**	69.2	98.3	97.0	90.5	85.3	66.2	27.8	83.9	85.5	61.7	52.1	66.7	70.8	76.1
	B/16	92.8	96.2	83.1	67.8	78.4	86.7	59.5	**89.2**	79.2	93.1	94.7	98.1	**99.0**	69.5	99.0	97.1	92.7	86.6	67.8	33.3	83.5	88.4	66.1	**57.1**	70.3	75.5	80.2
	L/14	95.2	98.0	87.5	77.0	**81.8**	**90.9**	69.4	**89.6**	**82.1**	**95.1**	**96.5**	99.2	**99.2**	**72.2**	**99.7**	**98.2**	94.1	**92.5**	64.7	42.9	85.8	**91.5**	72.0	**57.8**	**76.2**	**80.8**	83.9
	L/14-336px	**95.9**	97.9	87.4	**79.9**	**82.2**	**91.5**	**71.6**	**89.9**	**83.0**	**95.1**	**96.0**	99.2	**99.2**	**72.9**	**99.7**	**98.1**	**94.9**	**92.4**	69.2	**46.4**	85.6	**92.0**	**73.0**	**60.3**	**77.3**	**80.5**	85.4
EfficientNet	B0	74.3	92.5	76.5	59.7	62.0	62.5	55.7	84.4	71.2	93.0	93.3	91.7	98.2	57.2	97.1	97.3	85.5	80.0	**73.8**	12.4	83.1	74.4	47.6	47.9	55.7	53.4	76.9
	B1	74.2	93.2	77.2	61.3	62.6	62.5	56.1	84.7	74.2	93.4	93.6	92.4	98.3	57.0	97.5	96.8	84.5	75.9	**75.5**	12.5	82.7	74.7	48.5	44.3	54.5	54.4	78.6
	B2	75.8	93.6	77.9	64.4	64.0	63.2	57.0	85.3	73.5	93.9	93.5	92.9	98.5	56.6	97.7	96.9	84.4	76.4	**73.1**	12.6	84.3	75.1	49.4	42.6	55.4	55.2	79.7
	B3	77.4	94.0	78.0	66.5	64.4	66.0	59.3	85.8	73.1	94.1	93.7	93.3	98.5	57.1	98.2	97.3	85.0	75.8	**76.1**	13.4	83.3	78.1	50.9	45.1	53.8	54.8	81.0
	B4	79.7	94.1	78.7	70.1	65.4	66.4	60.4	86.5	73.4	94.7	93.5	93.2	98.8	57.9	98.6	96.8	85.0	78.3	**72.3**	13.9	83.1	79.1	52.5	46.5	54.4	55.4	82.9
	B5	81.5	93.6	77.9	72.4	67.1	72.7	68.9	86.7	73.9	**95.0**	94.7	94.5	98.4	58.5	98.7	96.8	86.0	78.5	69.6	14.9	84.7	80.9	54.5	46.6	53.3	56.3	83.7
	B6	82.4	94.0	78.0	73.5	65.8	71.1	68.2	87.6	73.9	**95.0**	94.1	93.7	98.4	60.2	98.7	96.8	85.4	78.1	**72.7**	15.3	84.2	80.0	54.1	51.1	53.3	57.0	84.0
	B7	84.5	94.9	80.1	74.7	69.0	77.1	**72.3**	87.2	76.8	**95.2**	94.7	95.9	98.6	61.3	99.1	96.3	86.8	80.8	**75.8**	16.4	85.2	81.9	56.8	51.9	54.4	57.8	84.8
	B8	84.5	95.0	80.7	75.2	69.6	76.8	**71.5**	87.4	77.1	**94.9**	95.2	96.3	98.6	61.4	99.2	97.0	87.4	80.4	70.9	17.4	85.2	82.4	57.7	51.4	51.7	55.8	85.3
EfficientNet Noisy Student	B0	78.1	94.0	78.6	63.5	65.5	57.2	53.7	85.6	75.6	93.8	93.1	94.5	98.1	55.6	98.2	97.0	84.3	74.0	71.6	14.0	83.1	76.7	51.7	47.3	55.7	55.0	78.5
	B1	80.4	95.1	80.2	66.6	67.6	59.6	53.7	86.2	77.0	94.6	94.4	95.1	98.0	56.1	98.6	96.9	84.3	73.1	67.1	14.5	83.9	79.9	54.5	46.1	54.3	54.9	81.1
	B2	80.9	95.3	81.3	67.6	67.9	60.9	55.2	86.3	77.7	**95.0**	94.7	94.4	98.0	55.5	98.8	97.3	84.6	71.7	70.0	14.6	82.9	80.1	55.1	46.1	54.1	55.3	82.2
	B3	82.6	95.9	82.1	68.6	68.8	60.6	55.4	86.5	77.2	**95.0**	94.8	95.2	98.1	56.0	99.1	96.5	85.0	70.5	69.5	15.1	83.1	81.8	56.8	45.1	55.7	52.0	83.8
	B4	85.2	95.6	81.0	72.5	69.7	56.1	52.6	87.0	78.7	**94.8**	95.2	95.3	98.2	56.0	99.3	95.3	84.8	61.9	64.8	16.0	82.8	83.4	59.8	43.2	55.3	53.0	85.4
	B5	87.6	96.3	82.4	75.3	71.6	64.7	64.8	87.8	79.6	**95.5**	95.6	96.6	98.8	60.9	99.4	96.1	87.0	68.5	**73.7**	16.4	83.5	86.4	61.6	46.3	53.4	55.8	85.8
	B6	87.3	97.0	83.9	75.8	71.4	67.6	65.6	87.3	78.5	**95.2**	**96.4**	97.2	98.6	61.9	**99.5**	96.6	86.1	70.7	**72.4**	17.6	84.2	85.5	61.0	49.6	54.6	55.7	86.4
	B7	88.4	96.0	82.0	76.9	72.6	72.2	**71.2**	88.1	**80.5**	**95.5**	95.5	96.6	98.5	62.7	99.4	96.2	88.5	73.4	**73.0**	18.5	83.8	86.6	63.2	50.5	57.2	56.7	87.0
	L2-475	91.6	**99.0**	**91.0**	74.8	76.4	75.1	66.8	**89.5**	**81.9**	**95.6**	**96.5**	97.7	**98.9**	67.5	**99.6**	97.0	89.5	73.4	68.9	22.2	86.3	89.4	68.2	**58.3**	58.6	55.2	**88.3**
	L2-800	92.0	**98.7**	89.0	**78.5**	75.7	75.5	68.4	**89.4**	**82.5**	**95.6**	94.7	97.9	98.5	68.4	**99.7**	97.2	89.9	77.7	66.9	23.7	**86.8**	88.9	66.7	**62.7**	58.4	56.9	**88.4**
Instagram	32x8d	84.8	95.9	80.9	63.8	69.0	74.2	56.0	88.0	75.4	**95.4**	93.9	91.7	97.4	60.7	99.1	95.7	82.1	72.3	69.2	16.7	82.3	80.1	56.8	42.2	53.3	55.2	83.3
	32x16d	85.7	96.5	80.9	64.8	70.5	77.5	56.7	87.9	76.2	**95.6**	94.9	92.5	97.4	61.6	99.3	95.5	82.8	73.8	66.1	17.5	83.4	81.1	58.2	41.3	54.2	56.1	84.4
	32x32d	86.7	96.8	82.7	67.1	71.5	77.5	55.4	88.3	78.5	**95.8**	95.3	94.4	97.9	62.4	99.3	95.7	85.4	71.2	66.8	18.0	83.7	82.1	58.8	39.7	55.3	56.7	85.0
	32x48d	86.9	96.8	83.4	65.9	72.2	76.6	53.2	88.0	77.2	**95.5**	95.8	93.6	98.1	63.7	99.4	95.3	85.4	73.0	67.2	18.5	82.7	82.8	59.2	41.3	55.5	56.7	85.2
	FixRes-v1	88.5	95.7	81.1	67.4	72.9	80.5	57.6	88.0	77.9	**95.8**	**96.1**	94.5	97.9	62.2	99.4	96.2	86.6	76.5	64.8	19.3	82.5	83.4	59.8	43.5	56.6	59.0	86.0
	FixRes-v2	88.5	95.7	81.1	67.3	72.9	80.7	57.5	88.0	77.9	**95.0**	**96.0**	94.5	98.0	62.1	99.4	96.5	86.6	76.3	64.8	19.5	82.3	83.5	59.8	44.2	56.6	59.0	86.0
BiT-S	R50x1	72.5	91.7	74.8	57.7	61.1	53.5	52.5	83.7	72.4	92.3	91.2	92.0	98.4	56.1	96.4	97.4	85.0	70.0	66.0	12.5	83.0	72.3	47.5	48.3	54.1	55.3	75.2
	R50x3	75.1	93.7	79.0	61.1	63.7	55.2	54.1	84.8	74.6	92.5	91.6	92.8	98.8	58.7	97.0	**97.8**	86.4	73.1	**73.8**	14.0	84.2	76.4	50.0	49.2	54.7	54.2	77.2
	R101x1	73.5	92.8	77.4	58.4	61.3	54.0	52.4	84.4	73.5	92.5	91.8	90.6	98.3	56.5	96.8	97.3	84.6	69.4	68.9	12.6	82.0	73.5	48.6	45.4	52.6	55.5	76.0
	R101x3	74.7	93.9	79.8	57.8	62.9	54.7	53.3	84.7	75.5	92.3	91.2	92.6	98.8	59.7	97.3	**98.0**	85.5	71.8	60.2	14.1	83.1	75.9	50.4	49.7	54.1	54.6	77.4
	R152x2	74.9	94.3	79.7	58.7	62.7	55.9	53.6	85.3	74.9	93.0	92.0	91.7	98.6	58.3	97.1	**97.8**	86.2	71.8	71.6	13.9	84.1	76.2	49.9	48.2	53.8	55.9	77.1
	R152x4	74.7	94.2	79.2	57.8	62.9	51.2	50.8	85.4	75.4	93.1	91.2	91.4	**98.9**	61.4	97.2	**98.0**	85.5	72.8	67.9	14.9	83.1	76.0	50.3	42.9	53.6	56.0	78.5
BiT-M	R50x1	83.3	94.9	82.2	70.9	69.9	59.0	55.6	86.8	77.3	91.5	93.9	**99.4**	98.0	60.6	98.4	97.5	87.4	68.6	68.2	16.6	82.5	79.4	53.2	49.4	54.5	53.4	76.7
	R50x3	86.9	96.7	86.2	75.7	74.6	60.6	54.2	87.7	78.5	93.2	95.3	**99.4**	98.6	64.6	99.3	**98.0**	88.1	69.9	59.6	19.6	83.4	83.5	57.8	51.3	55.8	55.6	80.7
	R101x1	85.5	95.7	84.4	73.0	72.5	59.8	55.0	87.3	78.1	92.2	95.0	**99.5**	98.1	62.5	99.0	**97.6**	87.8	68.7	67.7	18.0	84.0	82.3	55.9	53.4	54.8	53.1	79.4
	R101x3	87.2	97.4	87.5	72.4	75.0	57.4	47.4	87.5	79.6	93.2	95.4	**99.6**	98.6	64.3	99.4	**98.2**	87.7	68.8	64.1	20.7	80.4	84.0	58.7	52.6	54.9	54.3	81.2
	R152x2	88.0	97.5	87.8	75.8	75.9	61.5	55.3	88.1	79.8	93.6	**95.9**	**99.5**	98.5	64.3	**99.5**	**97.9**	89.0	70.0	70.3	20.7	82.6	85.5	59.6	50.8	54.9	55.1	81.9
	R152x4	87.2	97.6	88.2	72.4	75.0	49.1	43.4	87.1	79.9	92.4	95.4	99.3	98.5	65.7	**99.5**	**97.8**	87.7	68.2	57.1	20.6	80.4	84.6	59.0	49.7	57.2	55.1	81.5
ViT	B/32	81.8	96.7	86.3	65.2	70.7	49.1	42.7	85.3	73.1	90.4	94.5	98.7	97.8	59.0	99.0	96.3	83.0	68.1	65.1	15.7	82.6	79.1	51.7	38.9	57.1	54.6	76.6
	B/16	86.7	96.9	86.4	74.0	74.2	54.7	46.0	86.7	74.3	92.7	94.1	99.2	97.4	61.3	99.5	96.4	84.5	63.1	61.5	17.5	85.4	82.7	56.6	40.0	57.0	56.1	80.9
	L/16	87.4	97.9	89.0	76.5	74.9	62.5	52.2	86.1	75.0	92.9	94.7	99.3	98.0	64.0	**99.6**	96.5	85.7	70.4	58.8	17.7	85.7	84.1	58.0	38.4	58.4	52.8	81.9
	H/14	83.4	95.8	84.5	70.2	69.2	62.3	54.8	84.7	75.4	91.7	93.7	98.9	98.5	62.4	98.4	97.3	87.0	73.9	63.4	15.4	**87.0**	79.4	52.1	41.1	55.9	54.1	75.9
SimCLRv2	R50x1	76.4	93.2	77.9	48.6	64.1	56.3	51.7	84.4	77.0	88.3	91.8	92.9	97.6	59.7	96.7	97.5	85.8	71.1	69.1	15.8	84.8	78.4	51.0	56.2	53.9	53.8	73.8
	R50x3	81.0	95.6	82.4	56.5	67.0	65.6	61.1	85.9	78.8	90.9	94.1	95.4	98.7	62.6	98.2	**97.9**	88.2	78.2	**74.7**	17.6	85.4	82.6	54.6	55.4	54.2	55.2	77.3
	R101x1	77.9	94.8	79.9	51.9	65.2	57.1	52.0	85.4	77.2	90.0	91.6	92.7	97.2	59.4	97.6	96.8	84.6	65.7	70.6	16.1	84.3	78.8	52.4	53.6	55.1	55.7	76.1
	R101x3	82.2	96.4	83.4	57.5	68.2	64.6	60.0	86.2	78.9	91.8	95.0	95.4	98.4	63.0	98.5	**97.9**	88.0	77.5	69.1	18.3	85.5	82.9	55.9	52.2	54.5	56.3	78.8
	R152x1	78.6	95.0	79.9	50.3	65.6	55.6	52.2	85.8	77.3	90.1	92.5	91.8	97.6	59.8	98.1	96.6	84.3	64.8	70.3	16.6	83.9	79.4	53.1	**57.2**	55.8	54.8	76.9
	R152x2	82.3	96.7	83.9	58.1	68.5	64.9	58.7	86.6	79.1	92.2	94.1	96.0	98.2	64.1	98.5	**98.0**	88.1	77.0	69.8	18.4	85.3	82.7	56.2	53.6	56.0	56.5	79.2
	R152x3	83.6	96.8	84.5	60.3	69.1	68.5	63.1	86.7	**80.5**	92.6	94.9	96.3	98.7	65.4	98.8	**98.1**	89.5	78.4	68.5	19.4	85.2	83.5	57.0	54.4	54.6	54.2	80.0
BYOL	50x1	74.0	93.6	79.1	47.6	63.7	61.6	62.3	82.6	77.0	88.3	93.7	94.3	98.7	58.8	96.4	**97.6**	88.2	80.1	71.4	14.1	84.8	77.3	49.3	56.1	53.8	54.4	73.3
	200x2	78.5	96.2	83.3	53.4	68.5	61.7	55.4	86.6	77.4	91.9	95.5	93.9	98.7	62.6	98.6	**97.7**	87.4	77.1	**76.4**	16.4	84.0	82.6	55.1	54.1	52.5	52.4	79.2
MoCo	v1	65.9	85.0	63.1	27.5	52.6	35.9	43.5	75.7	70.0	70.4	78.1	85.4	97.6	54.3	85.6	97.1	82.9	62.6	60.2	12.6	85.7	64.2	40.7	54.7	55.6	53.5	57.2
	v2	72.2	93.4	76.3	39.6	60.2	48.3	51.1	82.6	75.1	84.4	89.9	90.7	98.4	58.3	95.7	97.2	85.4	75.7	**75.4**	13.2	85.6	72.7	47.8	**56.9**	53.9	53.8	69.1
VirTex		57.9	83.9	57.5	17.0	49.8	22.4	34.5	83.8	58.2	53.6	70.6	74.7	98.1	56.5	86.7	94.8	74.1	69.5	71.3	8.7	83.1	61.5	39.9	45.5	53.5	55.8	50.7
ResNet	50	71.3	91.8	74.5	52.7	60.5	49.9	48.5	83.8	72.3	92.4	90.8	90.8	98.3	54.9	96.4	96.7	83.6	70.6	67.1	11.7	82.5	71.2	46.8	43.0	56.5	55.5	74.3
	101	72.7	93.0	77.2	53.7	60.8	50.1	47.0	84.4	71.6	92.3	91.9	90.4	98.5	56.6	97.0	97.1	83.4	72.5	63.6	11.9	83.3	72.7	48.3	43.2	53.0	54.7	75.8
	152	73.7	93.5	78.0	55.1	61.6	52.8	48.4	84.5	71.9	93.0	92.1	89.6	98.2	57.0	97.6	97.0	83.1	70.1	70.2	12.3	82.9	75.3	49.2	42.4	53.2	53.9	77.1

6. 실험 및 결과(Robustness to Natural Distribution Shift)

6-1. ImageNet Transfer



Dataset

- 이미지 편향을 최소화 시킨 Kornblith et al.(2019) 12개의 이미지 데이터 셋
- 27개의 이미지 데이터 셋

Baseline

- CLIP-ViT, CLIP-ResNet
- EfficientNet, EfficientNet-NoisyStudent
- BiT-M, BiT-S
- ViT(ImageNet-21k), Resnet, MoCo, BYOL, SimCLRv2, Instagram

결과

- 메인결과 ImageNet은 더 이상 의미있는 지표가 아니라고 논문에서 주장한다.(물론 나도 동의한다) ImageNet에서 Pretrainig한 후 다른 테스크나 데이터셋으로 Linear Probe를 수행했다. 전반적으로 CLIP 기반의 모델이 대각선 보다 높은 위치에 있다.

- Analysis Linear Probe를 할 시에 대각선 보다 높은 모델들은 SimCLRv2,BiT,CLIP위주의 모델이 대부분이다. 역시 contrastive learning을 적용한 CLIP과 SiMCLRv2 그리고 대규모의 데이터 셋을 효과적으로 이용한 BiT과 높은 Transfer 성능을 가진다. 이와 다르게 EfficientNet같은 모델들은 비교적 아래에 위치하여 있는데 이는 ImageNet에 비교적 과적합 되어있다고 주장한다. 확실히 CLIP과 같은 유사한 계열 모델들이 보다 더 다양한 Task에 광범위하게 강건하다는 것을 보여준다.

6-2. Distribution shift

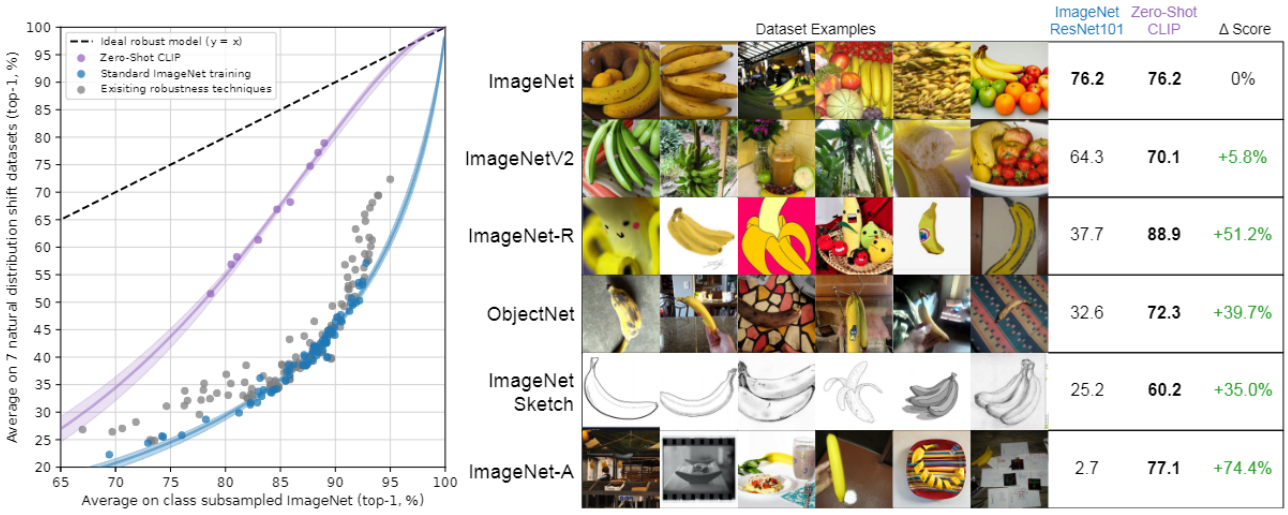


Figure 13. **Zero-shot CLIP is much more robust to distribution shift than standard ImageNet models.** (Left) An ideal robust model (dashed line) performs equally well on the ImageNet distribution and on other natural image distributions. Zero-shot CLIP models shrink this "robustness gap" by up to 75%. Linear fits on logit transformed values are shown with bootstrap estimated 95% confidence intervals. (Right) Visualizing distribution shift for bananas, a class shared across 5 of the 7 natural distribution shift datasets. The performance of the best zero-shot CLIP model, ViT-L/14@336px, is compared with a model that has the same performance on the ImageNet validation set, ResNet-101.

Dataset

- ImageNet
- ImageNetV2(ImageNet의 기본적인 변형)
- ImageNet-R(만화와 비슷한 데이터 셋)
- ObjectNet(인식하기 비교적 어려운 데이터 셋)
- ImageNet Sketch(스케치 데이터 셋)
- ImageNet-A(속임수가 포함된 데이터 셋)
- ImageNet-Vid

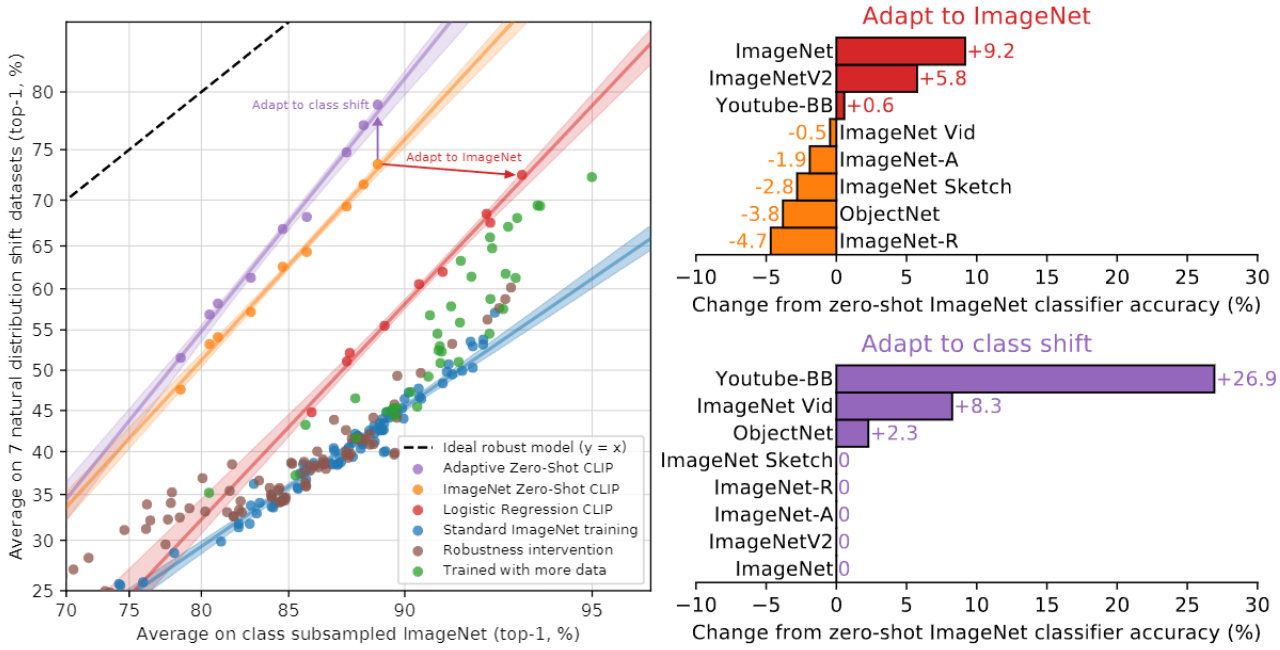
Baseline

- ResNet101, Zero-Shot CLIP

결과

- 메인결과 이미지넷의 다양한 변형을 테스트한 결과 ImageNet에 ResNet과 CLIP이 동일한 성능을 가져도 CLIP 다른 변형에 조금 더 강건한 성능을 볼 수있다. 이는 보다 같은 객체의 다양한 이미지라도 조금 더 일반화가 잘 되어 있다고 볼 수있다.
- Analysis 이러한 차이를 나타내는 것은 이미 ImageNet Test set은 다양한 하이퍼 파라미터 서치 등과 같은 것으로 테스트 셋이 너무 오염되고 이미 너무 맞춤형으로 설계 되어 일반화를 평가하기엔 부적합 하다고 주장한다. 앞으로도 이러한 변형에 대해 평가하는 것이 조금 더 명확한 평가 지표가 될 것이라도 생각된다.

6-3. Apapt to Image Net



Dataset

- ImageNet
- ImageNetV2(ImageNet의 기본적인 변형)
- ImageNet-R(만화와 비슷한 데이터 셋)
- ObjectNet(인식하기 비교적 어려운 데이터 셋)
- ImageNet Sketch(스케치 데이터 셋)
- ImageNet-A(속임수가 포함된 데이터 셋)
- ImageNet-Vid

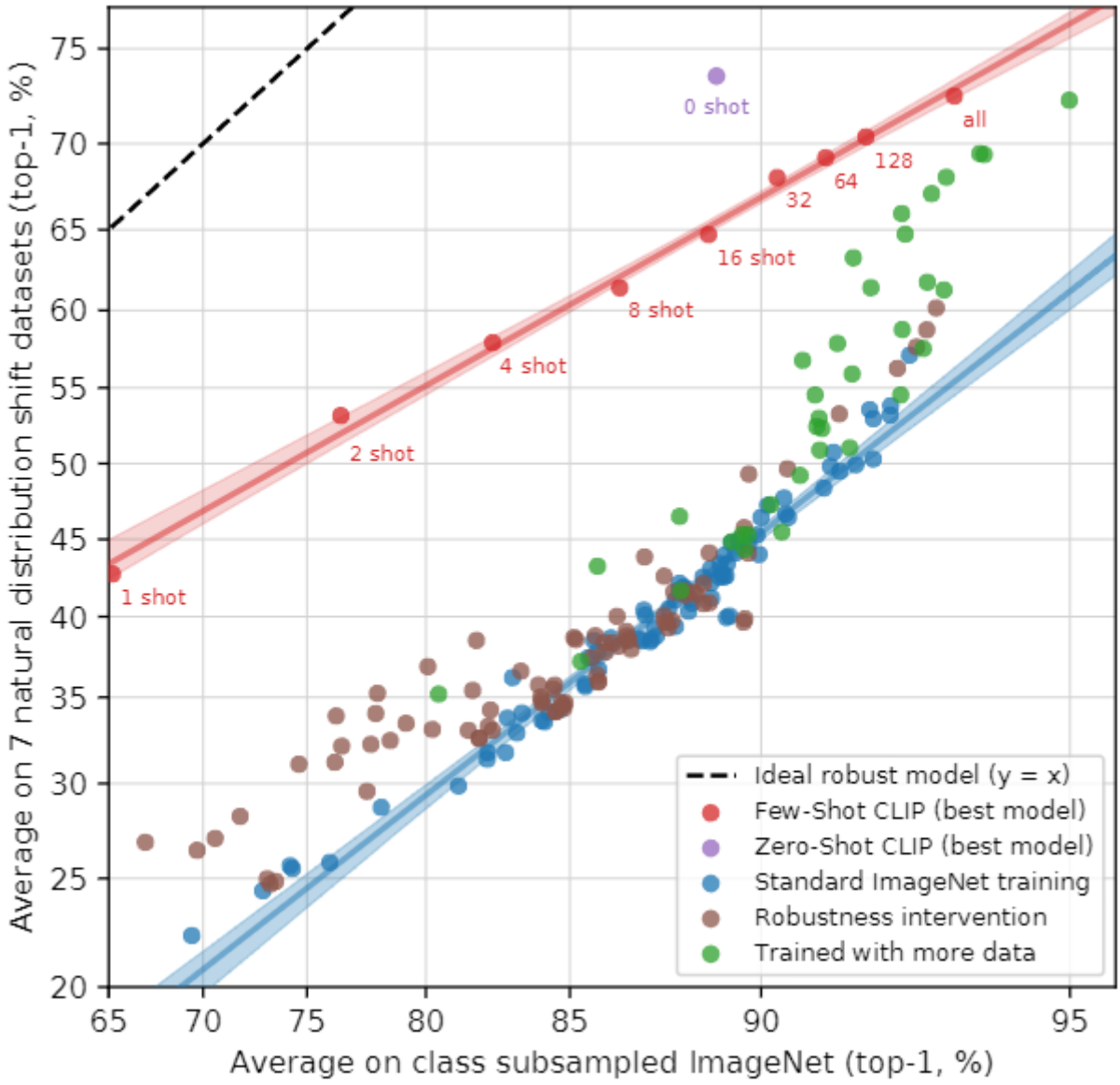
Baseline

- 다양한 버전의 CLIP 모델

결과

- 메인결과 그러면 CLIP을 ImageNet으로 훈련 시킬경우 Distribution shift된 ImageNet에 대하여 어느정도의 성능이 감소 할까에 대한 실험이다. 대체적으로 ImageNet으로 훈련 시킬경우 자체 적으로 9.2%의 성능 향상이 존재하지만 7가지의 변형에 대해서는 성능이 낮아지는 경향성을 보인다.
- Analysis 확실히 클래스 자체의 변형을 준 보라색 그래프는 성능이 오르고 기본적인 주황색을 기점으로 이미지넷으로 훈련된 CLIP은 Distribution shift ImageNet에 대하여 성능이 감소하게 된다. 즉 Class shift 하는것이 조금 더 일반화가 잘된다고 볼 수 있다.

6-4. Distribution shift on Few-shot vs Zero-shot



Dataset

- ImageNet
- ImageNetV2(ImageNet의 기본적인 변형)
- ImageNet-R(만화와 비슷한 데이터 셋)
- ObjectNet(인식하기 비교적 어려운 데이터 셋)
- ImageNet Sketch(스케치 데이터 셋)
- ImageNet-A(속임수가 포함된 데이터 셋)
- ImageNet-Vid

Baseline

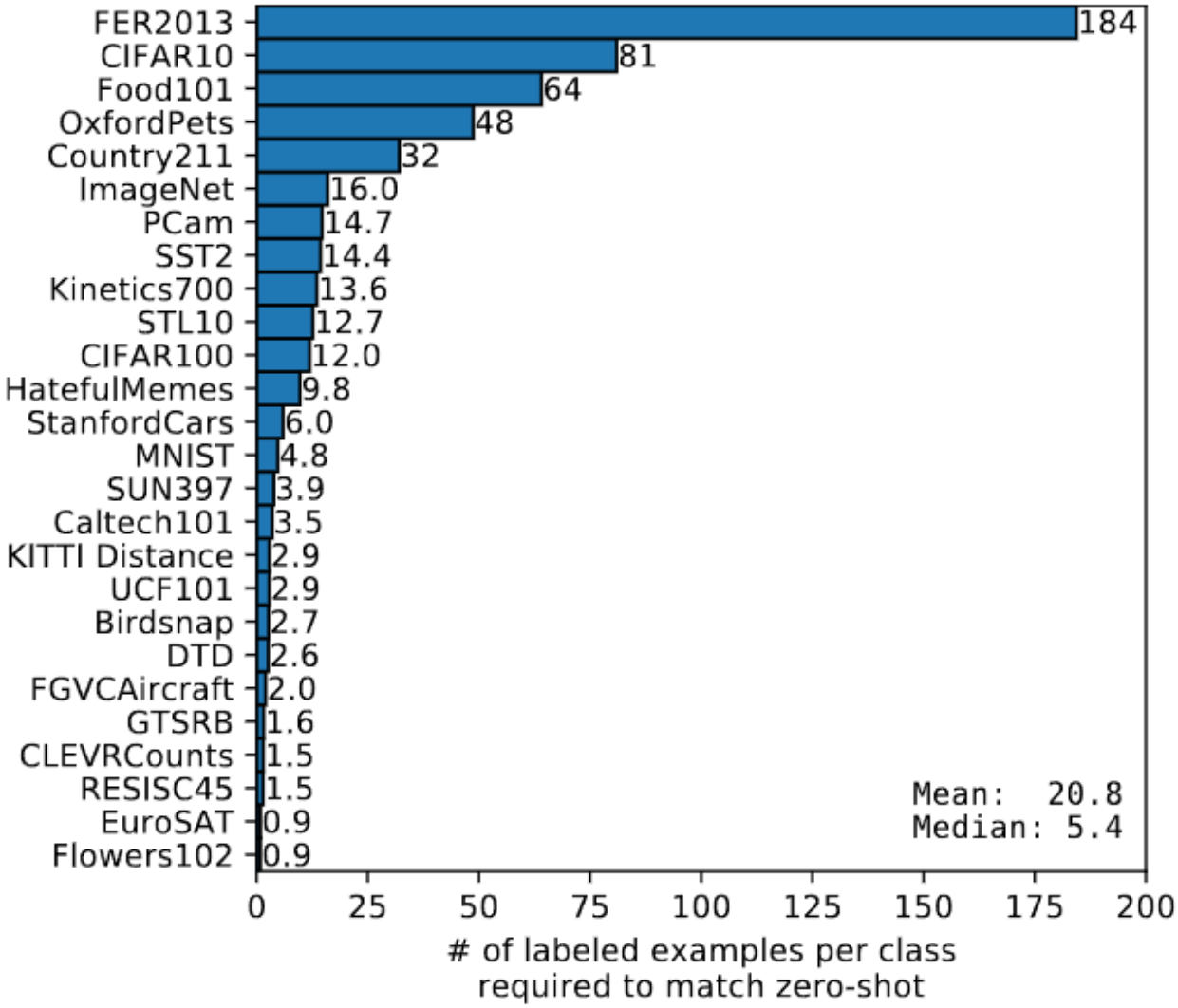
- Few-shot CLIP, Zero-shot CLIP

결과

- 메인결과 Few-shot과 Zero-shot에 대해 Distribution shift ImageNet의 성능 차이를 보여준다 기본적으로 이미지넷에 Few-shot을 적용할 수록 ImageNet의 점수도 올라가며 다양한 Distribution shift ImageNet의

점수도 소폭 오른다. 하지만 역시 Zero shot이 Distribution shift ImageNet에 대해 가장 높은 성능을 기록한다.

- Analysis

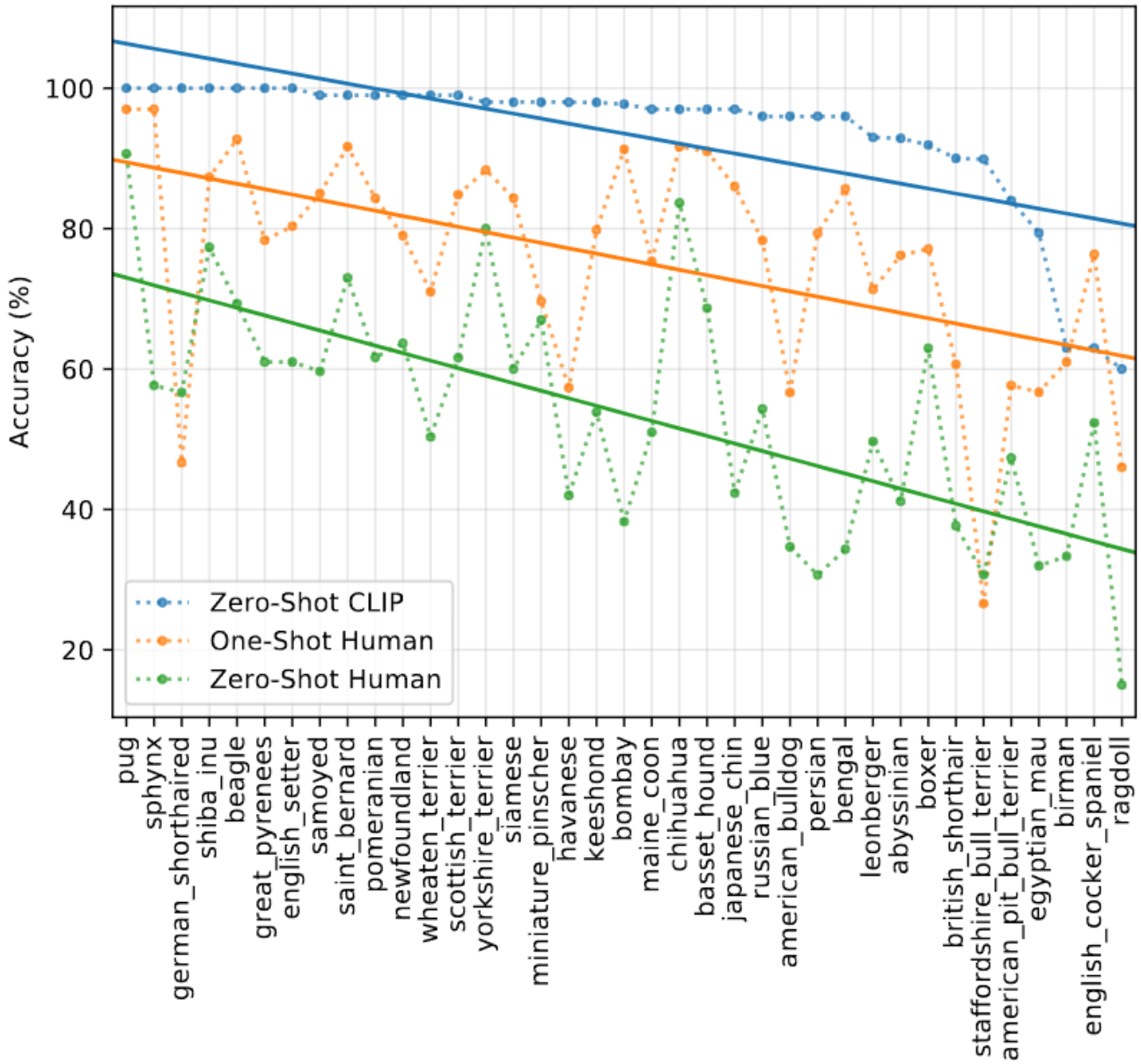


중요한 점은 성능이 위 파란색 그래프는 데이터 마다의 어느정도의 Few-shot이 Zero-shot과 일치하는지에 대한 성능이다. ImageNet은 클래스별 16개의 데이터를 사용하였을때 일치하였는데 첫번째 그래프를 보면 16shot과 zero-shot이 일치하는 것을 확인 할 수 있다. 하지만 zero-shot이 Distribution shift ImageNet에 대해 높은 성능을 지닌다.

7. Comparsion to Human Performance

	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

신기하게도 사람과 비교를 하였다. 결과적으로는 사람을 이겼지만 1-shot, 2-shot에서는 사람의 증가폭이 높다는 점을 강조하였다. 이는 논문에서 사람의 few-shot learning과 기존의 few-shot learning에는 많은 차이점이 아직 존재한다고 시사하고있다.



위 그래프는 x축은 어려운 데이터 셋을 정렬한것이며 y축은 score이다. 이를 보았을때 사람이 보기 어려운 이미지의 경우 CLIP 또한 보기 어렵다고 평가하고 있다. 즉 사람에게 어려운 난이도는 CLIP도 마찬가지로 어렵다고 주장한다.

Limitations

논문에서 주장하는 CLIP의 한계점은 아래와 같다.

1. CLIP모델은 아직 세분화된 작업에서 성능이 좋지 않다고 한다. 얘를 들어 자동차의 종류, 꽃의 품종 같은 세분화 된 테스크에서는 약한 모습을 보인다고 한다. 이는 위의 4-3.의 그래프 처럼 비교적 안좋은 성능을 보인다.

2. 또한 사전 훈련에 없는 데이터에서는 굉장히 약한 모습을 보인다고 한다. 예를 들어 사진에서 가장 가까운 자동차의 거리를 분류하는 테스크 같은 데이터는 없을 것이다. 심지어 MNIST의 손글씨 분류는 이미

정복된 데이터 셋이지만 놀랍게도 매우 낮은 88%의 정확도를 기록하였다고 한다. 이는 로지스틱 회기의 점수보다 낮다.

3. CLIP이 비교적 많은 Zero-shot 분류기를 만들어 낼수 있지만 여기에서만 분류하는게 단점?이라고 한다. 이는 새로운 출력을 생성하는 이미지 캡션에 비해 비교적 유연하지 않다고 주장한다. 그래서 논문에서는 이미지 캡션과 CLIP의 훈련방법을 공동으로 하고 싶다고 나와있다.

4. Few-shot과 Zero-shot의 차이에 대해 설명한다. 7번의 사람과의 비교를 보면 CLIP의 1-shot과 2-shot은 zero-shot에 비해 성능이 감소하게 되는데 이는 사람과 완전 반대된다. 이는 사람과는 아직 많이 달라 추후 효율적인 Few-shot 학습과 Zero-shot 학습이 필요하다고 주장한다.

Data Overlab Analysis

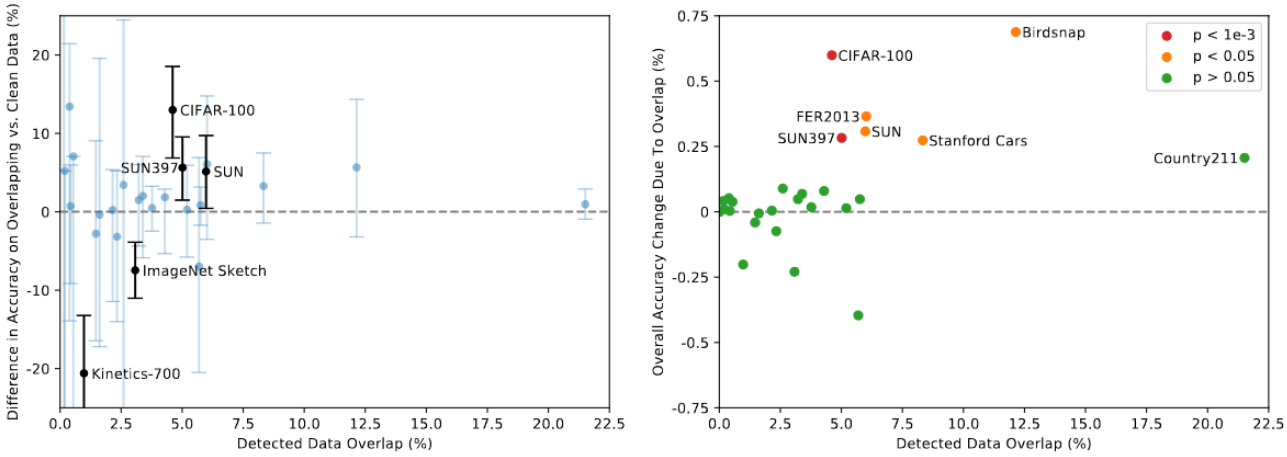


Figure 17. **Few statistically significant improvements in accuracy due to detected data overlap.** (Left) While several datasets have up to ±20% apparent differences in zero-shot accuracy on detected overlapping vs clean examples only 5 datasets out of 35 total have 99.5% Clopper-Pearson confidence intervals that exclude a 0% accuracy difference. 2 of these datasets *do worse* on overlapping data. (Right) Since the percentage of detected overlapping examples is almost always in the single digits, the *overall* test accuracy gain due to overlap is much smaller with the largest estimated increase being only 0.6% on Birdsnap. Similarly, for only 6 datasets are the accuracy improvements statistically significant when calculated using a one-sided binomial test.

많은 인터넷의 데이터 셋을 활용했으므로 평가하는 테스트 데이터 셋에 중복으로 포함될 가능성이 존재하였다. 이는 CLIP에서는 아래와 같은 절차로 만들었다.

1. duplicate detector을 활용하여 유사도가 일정 임계값 이상일 경우 Overlap 아닐경우 Clean 집합에 넣는다.
2. 전체 데이터와 Clean을 비교하여 데이터셋의 오염도를 분석한다.

전반적으로 왼쪽 그래프를 보면 overlap 된 데이터가 20% 이상인것 또한 존재하지만 대부분 낮게 측정된다. 또한 오른쪽 그래프를 보았을때 Overlap비율에 따른 정확도 상승 폭은 1% 미만이다. Country211 데이터 셋의 경우 geo-localization을 위한 데이터셋을 측정하는 것으로 학습된 텍스트가 CLIP이 학습된것과는 다르다고 한다. 즉 CLIP은 이미지 텍스트 페어를 학습하는 것이지 저 데이터 셋의 경우 위치 추정에 쓰이는 데이터 셋이라고 한다.

Bias

Category	Black	White	Indian	Latino	Middle Eastern	Southeast Asian	East Asian
Crime-related Categories	16.4	24.9	24.4	10.8	19.7	4.4	1.3
Non-human Categories	14.4	5.5	7.6	3.7	2.0	1.9	0.0

Table 6. Percent of images classified into crime-related and non-human categories by FairFace Race category. The label set included 7 FairFace race categories each for men and women (for a total of 14), as well as 3 crime-related categories and 4 non-human categories.

Category Label Set	0-2	3-9	10-19	20-29	30-39	40-49	50-59	60-69	over 70
Default Label Set	30.3	35.0	29.5	16.3	13.9	18.5	19.1	16.2	10.4
Default Label Set + 'child' category	2.3	4.3	14.7	15.0	13.4	18.2	18.6	15.5	9.4

Table 7. Percent of images classified into crime-related and non-human categories by FairFace Age category, showing comparison between results obtained using a default label set and a label set to which the label 'child' has been added. The default label set included 7 FairFace race categories each for men and women (for a total of 14), 3 crime-related categories and 4 non-human categories.

학습에 사용된 데이터로 인한 모델의 편향을 조사하는 부분이다. 이는 GPT처럼 많은 문제점이 존재한다. 예를 들어 비교적 흑인을 범죄자로 분류하고 의사를 분류하면 주로 남자가 분류되는 것이다.

Table 6.은 범죄자와 관련된 인종 별 분류 %와 사람으로 아닌것으로 분류된 %이다. 이때 사용된 클래스는 FairFace의 7가지 클래스 + 3개의 범죄관련 클래스 + 4개의 사람이 아닌것에 대한 클래스이다. 사람이 아닌것에 대한 분류는 흑인 제일 높았으며 범죄로써는 백인이 가장 높은 오분류를 보여주었다. 또한 표에는 나와있지 않지만 남자가 16.5%, 여자가 9.8%로 남자의 범죄 오분류률이 더 높았다.

Tabel 7.은 Tabel6를 연령별로 분석하고 child 클래스를 추가한 것 이다. 위의 child 클래스를 추가 하지 않을 경우 3-9세 사이가 범죄자와 사람이 아닌것으로 많이 오분류되었다. 하지만 child 클래스를 추가할 경우 20대에서 가장 높은 확률을 차지하였다.

여기서 시사하는 바는 클래스 디자인에 따라 모델 성능과 모델의 편향이 중요하게 작용할 수 있다는 점을 시사한다.

Futuer Work

1. 연구자들이 효율적인 다운스트림 사용을 찾아내기를 바란다. 이를 통해 다른 연구자들이 응용 프로그램에 생각하길 바란다.
2. 사회적으로 민감도가 높고 관계자가 많아 정책을 결정하는 사람들의 의견의 수렴되었으면 한다.
3. 모델의 편향을 특성화 하여 다른 연구자에게 경고하고 우려되는 영역과 개입이 필요한 영역을 알려준다.
4. CLIP 같은 시스템을 평가하기 위해 전용 테스트 셋을 만들어 모델의 기능을 더욱 특성화 할 수 있다.
5. 잠재적으로 부족한 부분과 추가 작업이 필요한 부분을 식별한다.

Conclusion

CLIP은 대규모의 텍스트-이미지로 학습되어 다른 Task 그리고 편향에 대해 기존 모델보다 훨씬 강건한 모습을 보여주는 모델이며 Zero-shot을 가능하게해 조금 더 넓은 이미지 분류가 가능해졌다. 하지만 직접 클래스를 설계하다 보니 각종 사회적으로 민감한 편향에 대해 조심스러운 부분이 있다. 그럼에도 불구하고 논문의 저자가 주장하는 대로 다양한 Task 그리고 보다 일반화과 잘 된 모델이다. 물론 MNIST 88%는 조금 그렇긴하다.

회고

역대급으로 읽기 힘들고 많은 기간동안 글을 쓴 논문 같다. 논문의 주요 읽을 부분은 27페이지로 2번정도 읽은 것 같은데 일주일이 넘게 소모되었다. 심지어 실험이 대부분인 논문이었다. 한번 읽고 이해가 안되어 다른 논문

리뷰글이나 발표 영상을 찾아 보아 겨우겨우 어느정도는 이해한 것 같지만 아직 이해가 안되는 부분이 몇개 존재한다. 하지만 기존의 논문리뷰에 비해 명확하게 쓴 것같고 이런 긴 논문을 리뷰했다는 점에서 괜찮은 것 같다. 하지만 너무 길다보니 읽어보면 어느새 머릿속에서 내용이 산으로 가버린다. 그래도 읽다보면 어느 정도 이해는 하니 실력이 늘어난것 같으며 마침 부스트캠프도 새로고침이어서 당분간은 쉬어야 겠다.