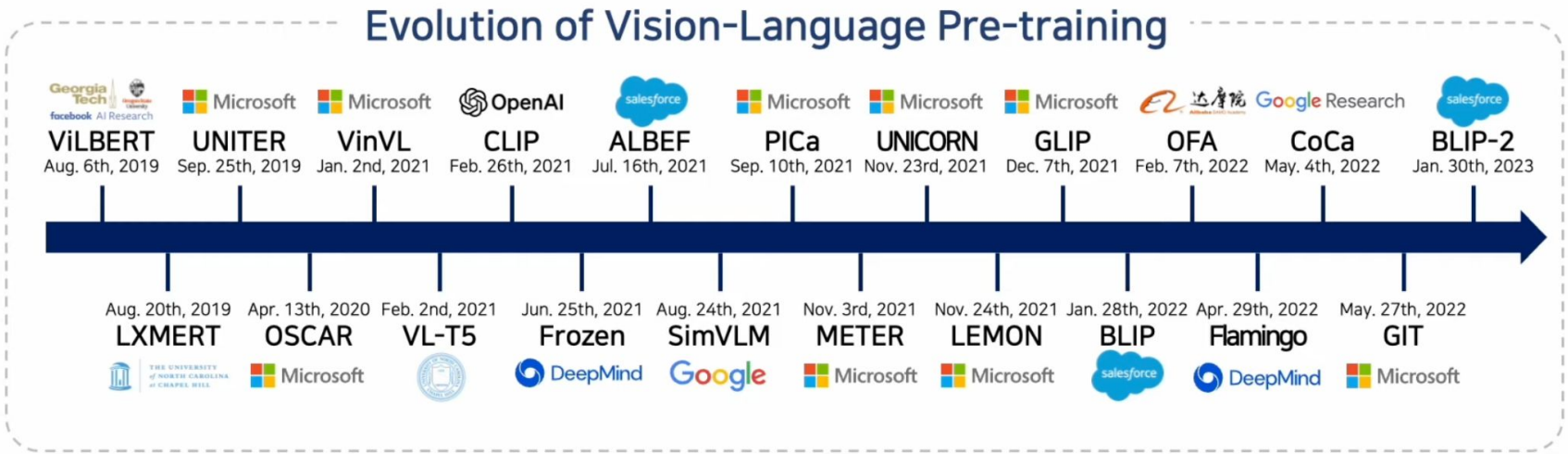
Learning Transferable Visual Models From Natural Language Supervision (CLIP)

VLM의 등장

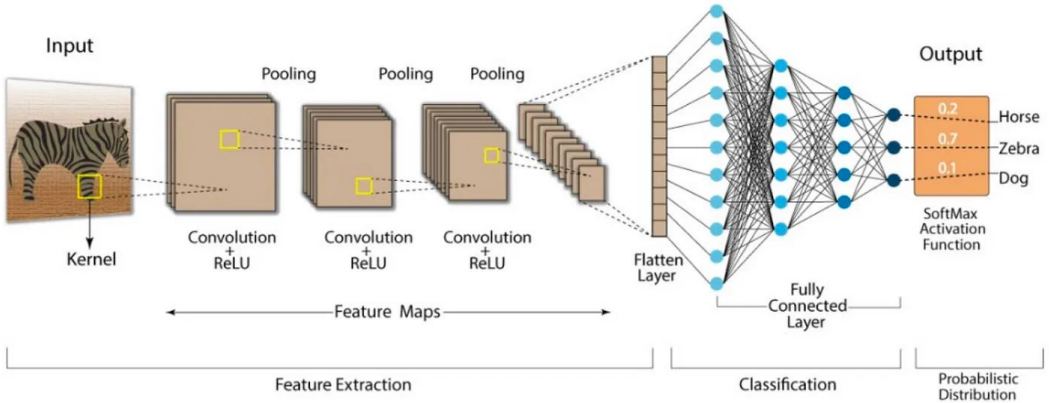
Vision-Language Pre-training

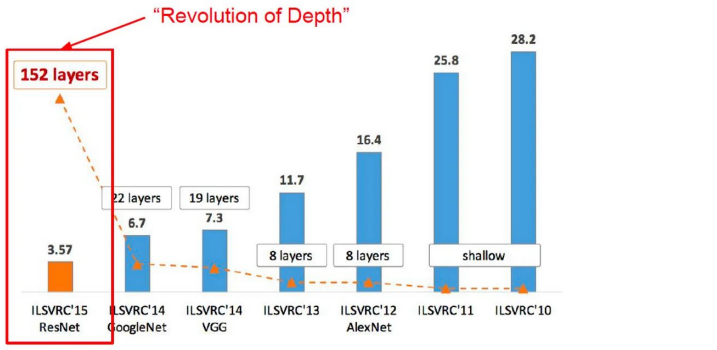


Vision Model

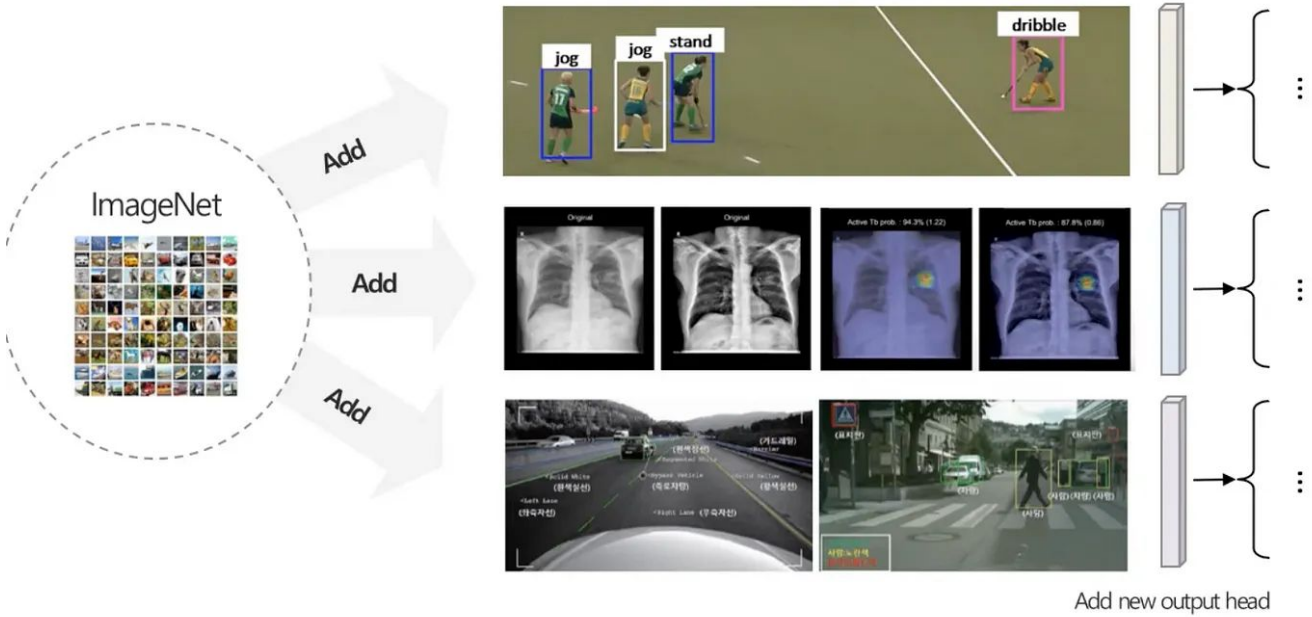
이미지를 입력받아 어떻게 모델을 구성하면 더 좋은 표현을 학습하는지를 고민.

-> 이미지만 학습한 모델은 고질적으로 일반화 능력이 부족하고 작은 노이즈에도 취약한 약점을 보임

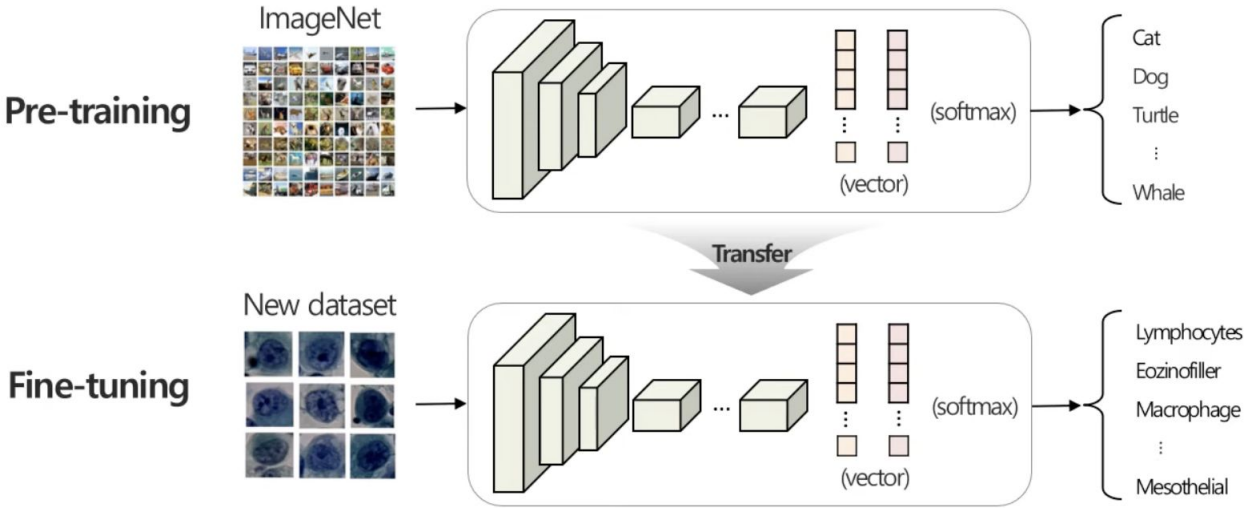




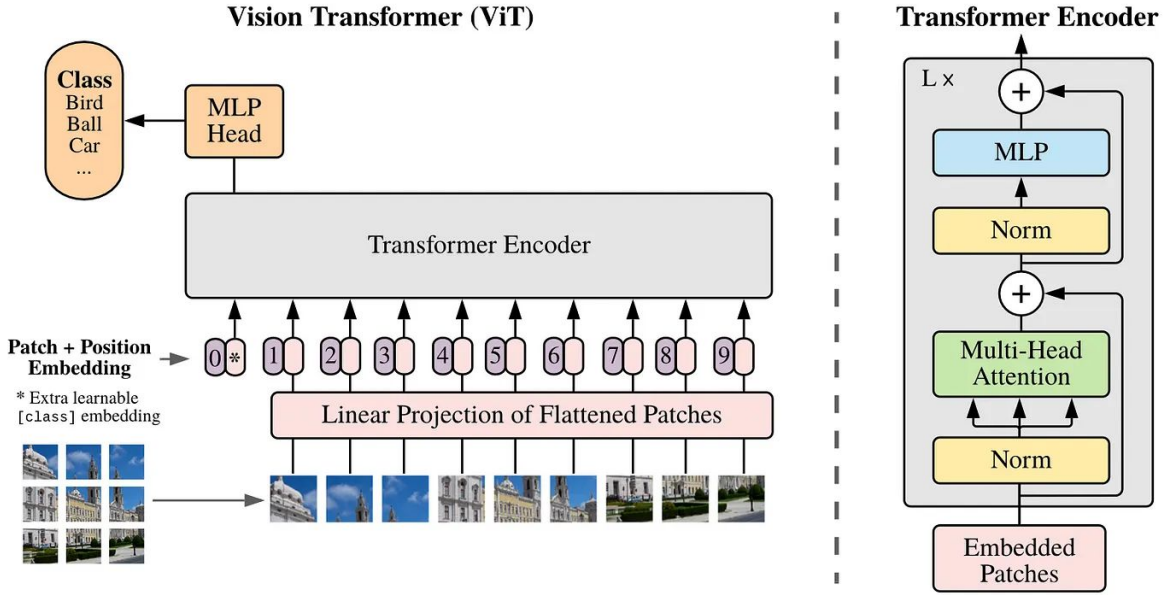
학습되지 않은 데이터에 약함.



Transfer Learning

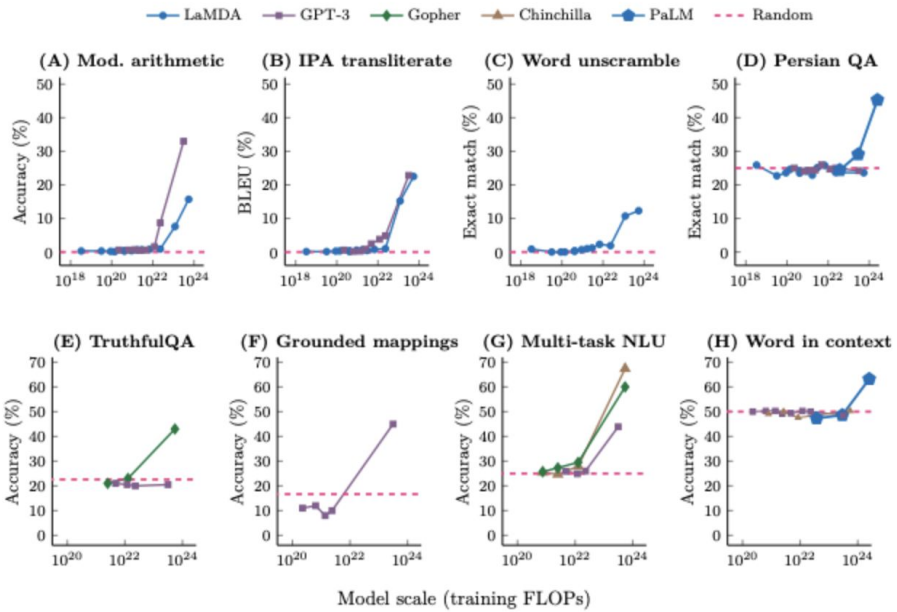


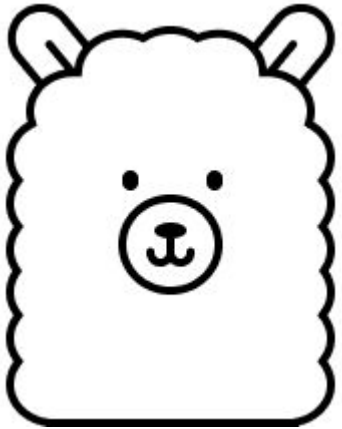
Vision Transformer



Attention 이후의 인공지능

큰 모델, 큰 데이터만 있으면 다 된다. -> ChatGPT

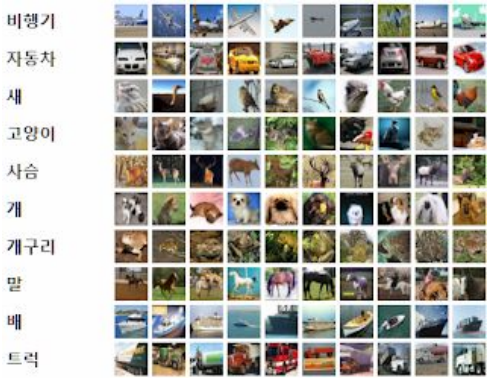




Llama3의 경우 15T(15조개 토큰 학습)

이미지 데이터셋

CIFAR(Canadian Institute For Advanced Research)

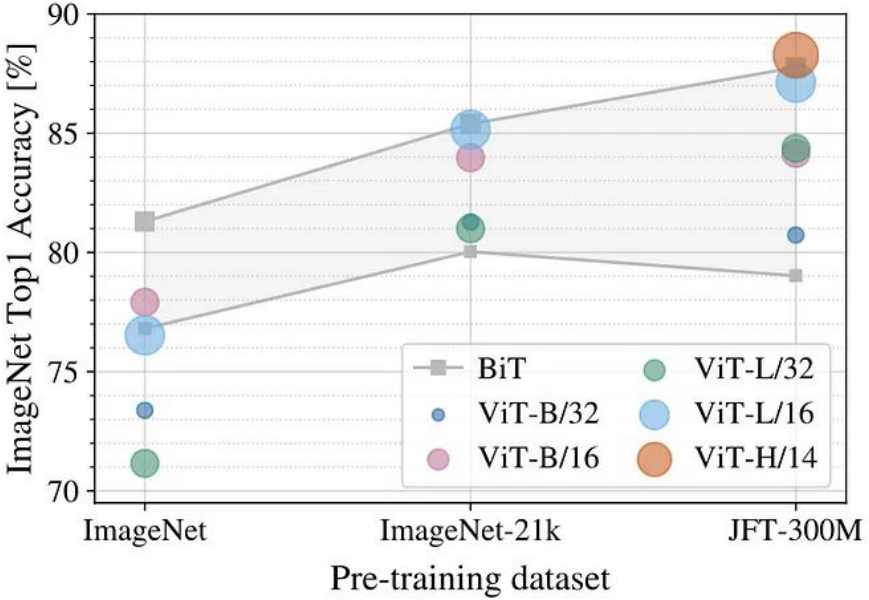


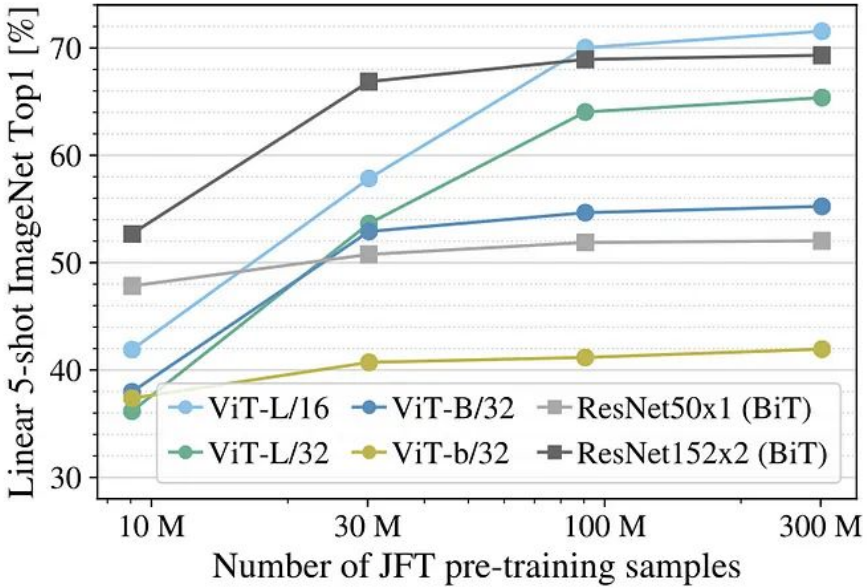
ImageNet(ImageNet Large Scale Visual Recognition Challenge - ILSVRC)

COCO Dataset(Microsoft COCO: Common Objects in Context)

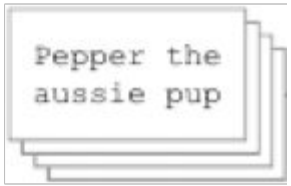
사람이 직접 라벨링하기 때문에 데이터 크기의 한계가 있다.
(위 3개 합쳐도 1500만장 정도.)

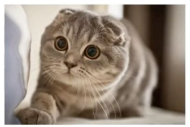
Vision Transformer(data problem)





데이터 부족 문제 해결?





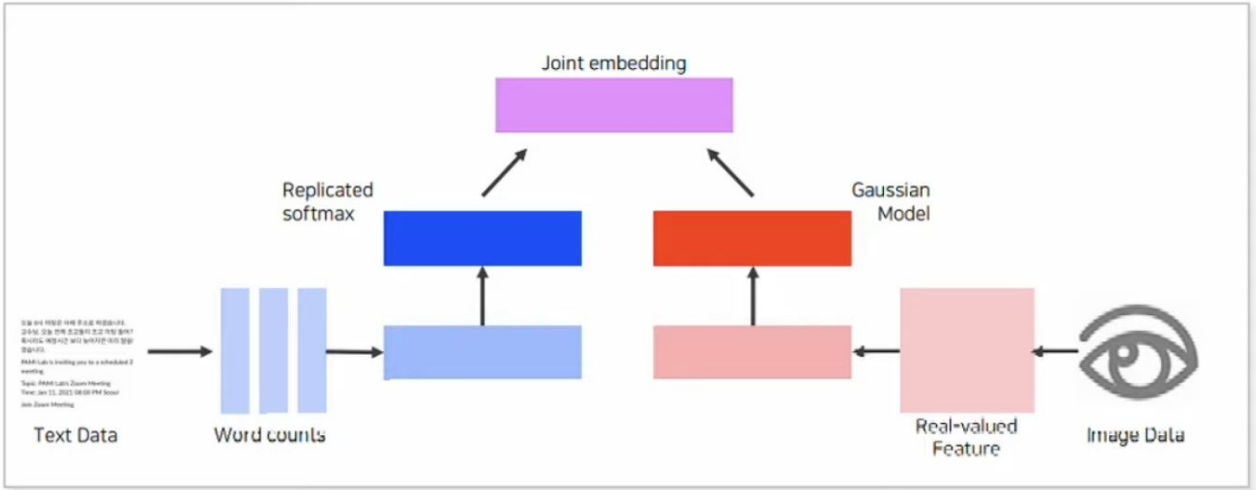
Visual representation

vs.

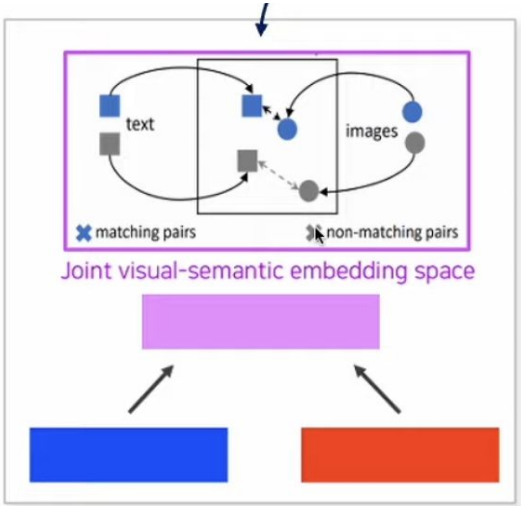
Visual representation + Semantic information

인터넷상에서 이미지마다 달려 있는 자연어 문장을 그대로 Supervision으로 사용하자는 아이디어
-> 4억장의 이미지 - 텍스트 데이터셋 구축

Multimodal Learning

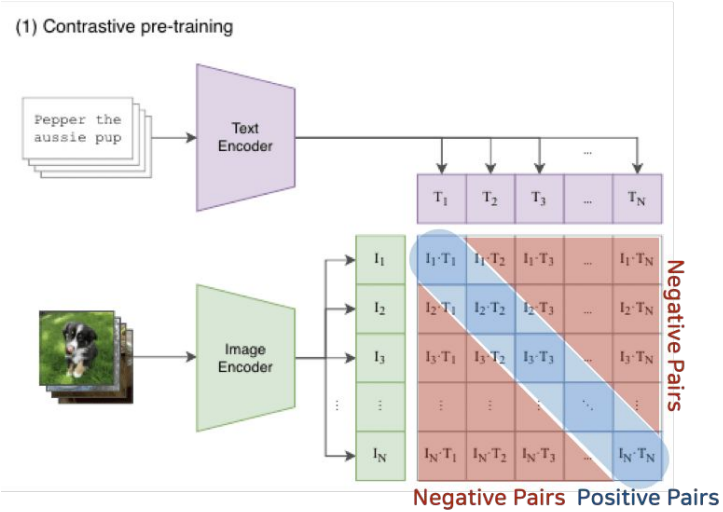


< Joint Embedding Method >



< Joint Embedding Space >

Contrastive Learning

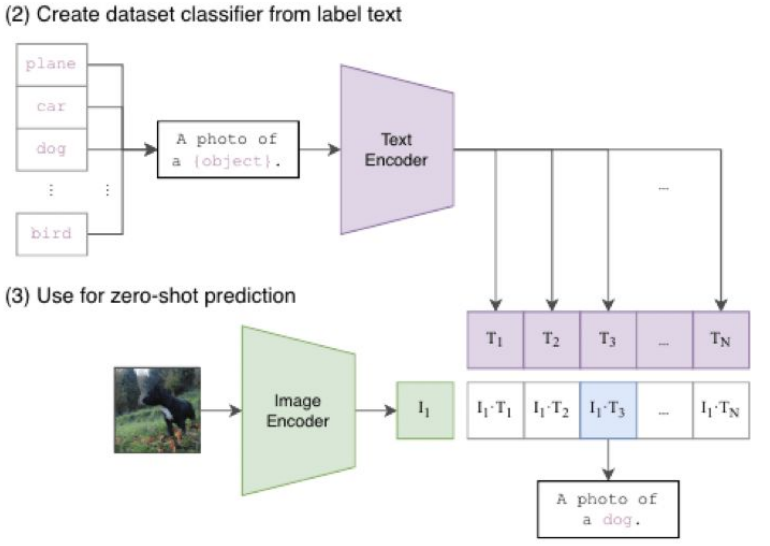


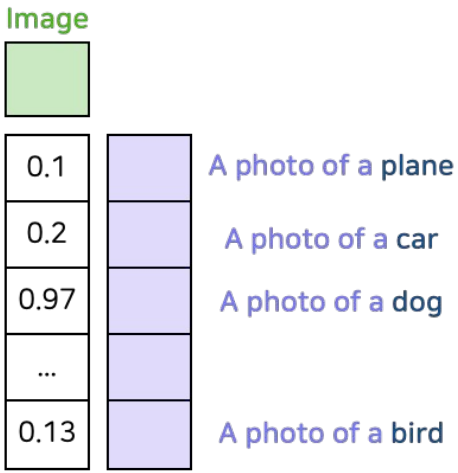

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned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

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

라벨의 갯수가 정해져있는 분류 문제가 아니라, 자연어를 이미지의 감독(Supervision)으로 활용하기 때문에 기존 분류 작업처럼 cross entropy loss로 학습하는 것이 불가능함
-> 대조 학습(Contrastive Learning)

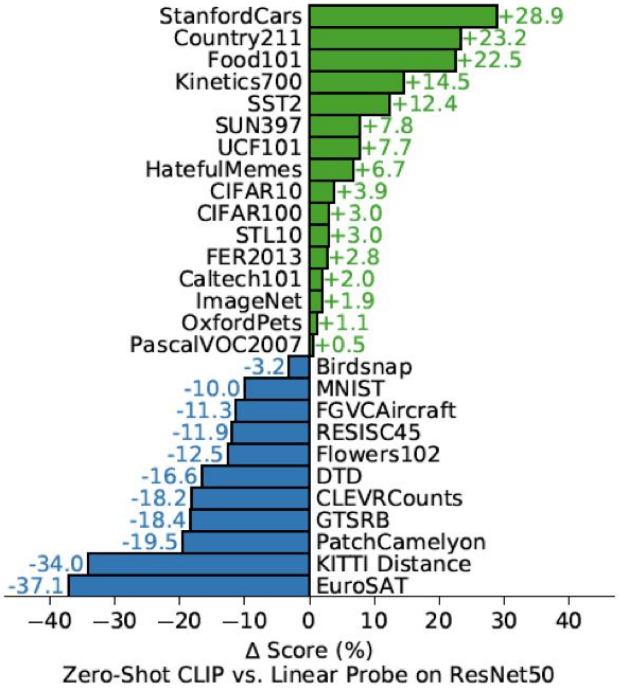
Zero Shot Prediction





이러한 방법을 통해서 CLIP은 고정되지 않은 개수의 클래스에 대해 예측이 가능.
기존의 Label을 사용하여 이미지의 클래스를 구분하는 방식이 아닌, 이미지와 자연어의 정렬 (Align)을 학습했기 때문.

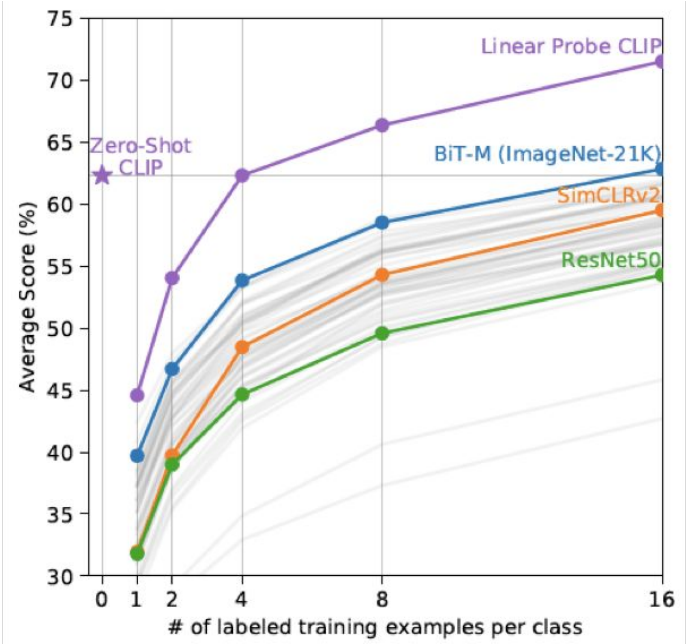
Zero Shot Transfer(1)

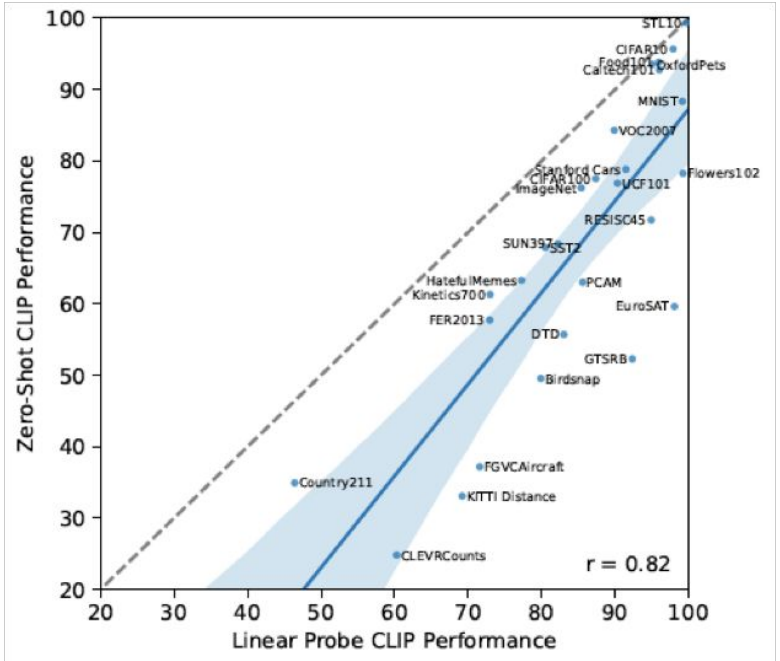


Linear Probe란 학습이 완료된 Encoder를 가져와 Supervised Learning으로 Classifier만 재학습해주는 방법

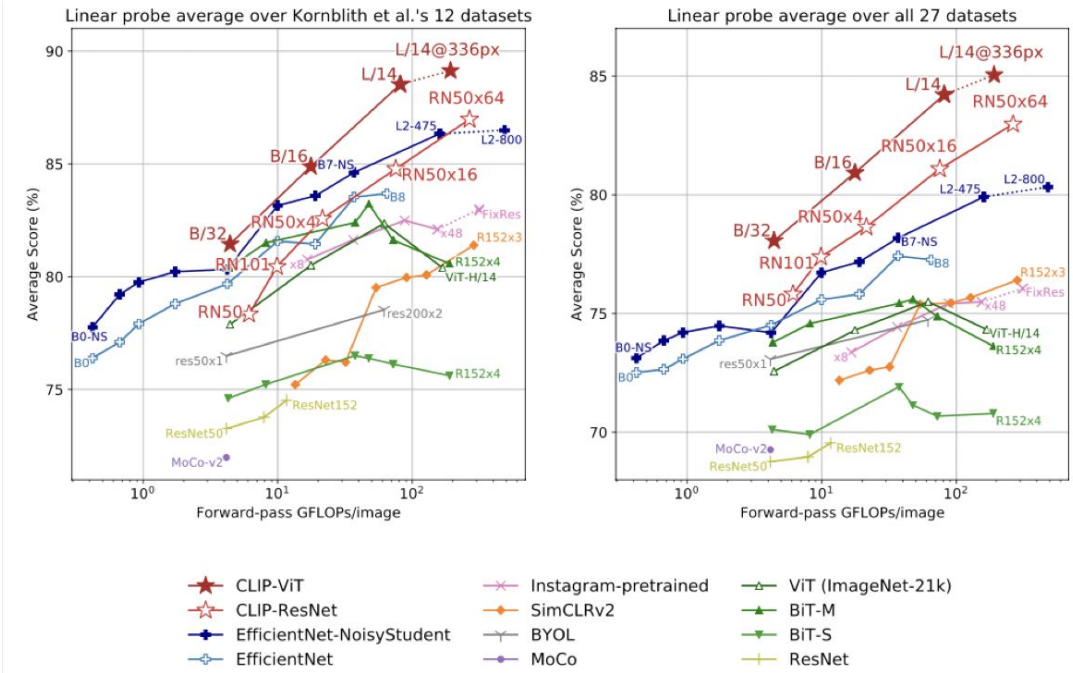
초록색이 CLIP, 즉 한번도 학습하지 문제(label)를 잘 해결할 수 있다는 것을 증명

Zero Shot Transfer(2)





Representation Learning



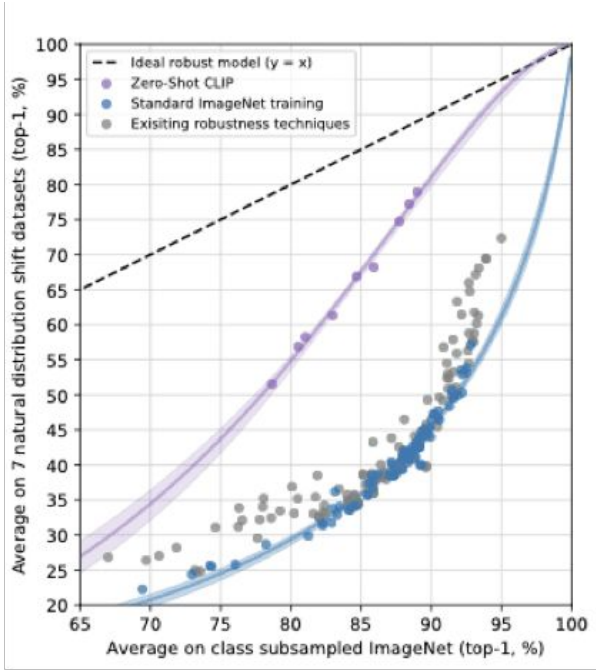
Robustness

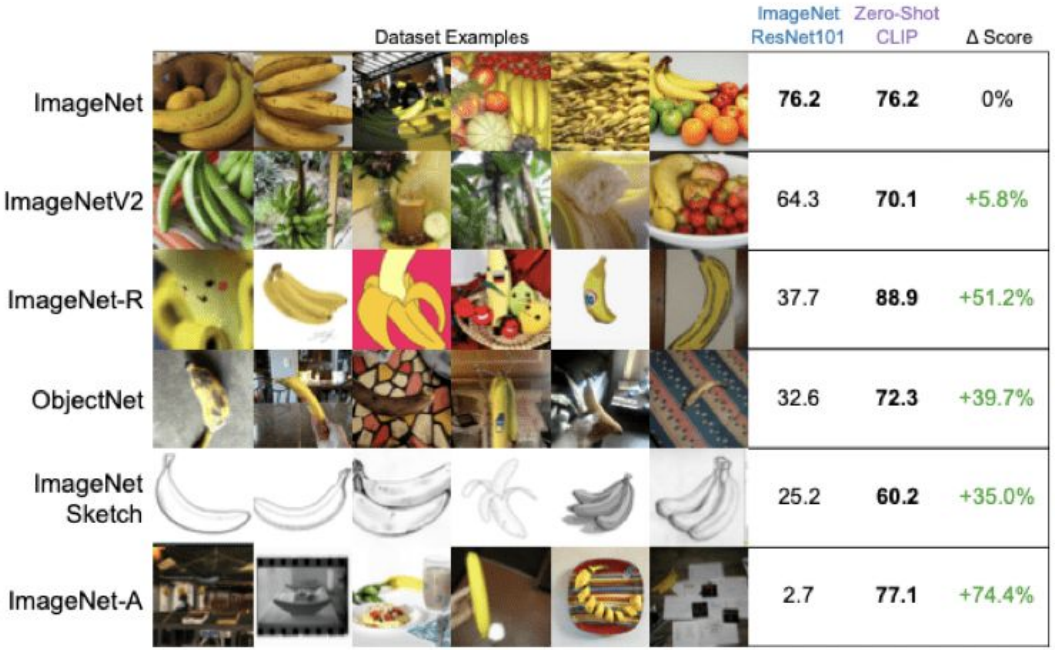
ImageNet

gap

Real-world dataset

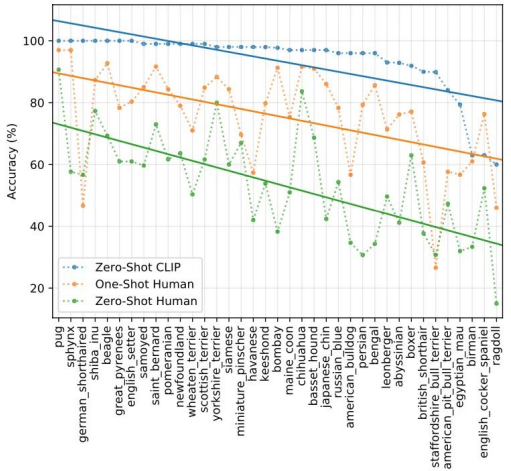
Robustness to natural distribution shift





VS human

	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



사람은 하나의 예시만 주더라도 성능이 엄청나게 좋아진다 -> 메타인지가 있다.

CLIP은 하나의 샘플을 재학습해도 성능이 좋아지지 않는다. 메타인지의 부족.

의의

- 자연어와 이미지의 의미있는 융합 방법을 제안
- Zero Shot 학습 방법의 가능성
- 다양한 데이터셋에 대한 적용성
- 편향에 대한 인식과 대응을 촉발
- 인공지능 연구의 새로운 방향을 제시
- 실용적인 응용 가능성을 제시

출처

- https://ffighting.net/deep-learning-paper-review/multimodal-model/clip/
- https://openai.com/index/clip/
- https://www.youtube.com/watch?v=dELmmuKBUtI
- https://medium.com/@taewan2002/clip-connecting-text-and-images-1c76cc1bae65