Learning Transferable Visual Models From Natural Language Supervision

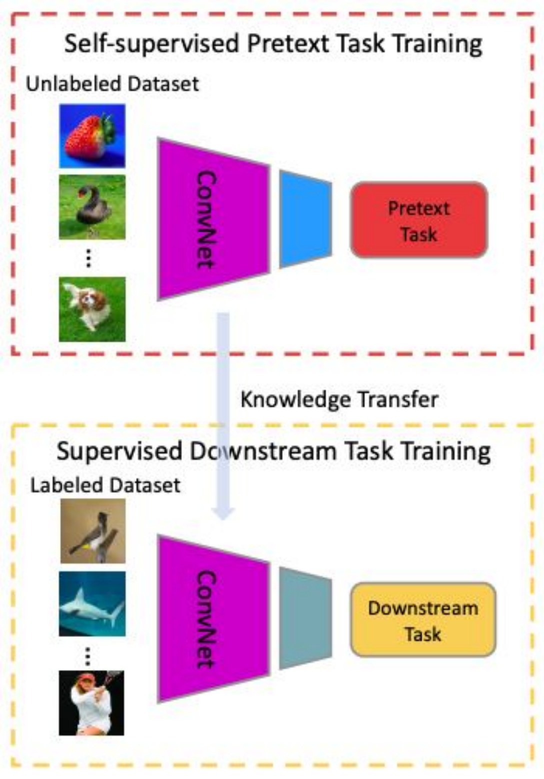
•••

By Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger and Ilya Sutskever

Presented by Balaji Balasubramanian and Eshwanth Baskaran

Introduction

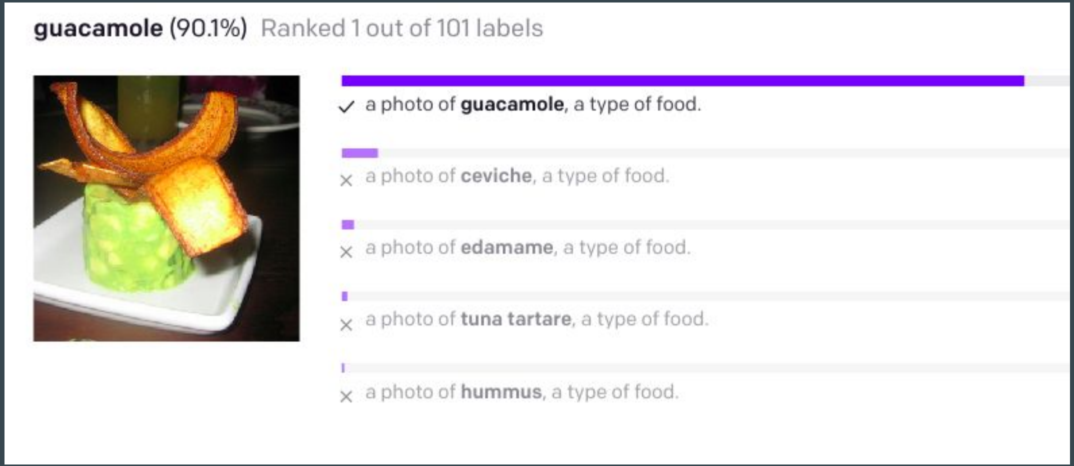
- Labelling Data for supervised learning tasks
  - Expensive
- Self-Supervised learning
  - Learning from input data itself without labels.
  - Pretext task (pre-defined task) is defined for the model to solve which helps model to learn useful features.
  - The features learned can be used for downstream tasks.



L. Jing and Y. Tian, “Self-supervised visual feature learning with deep neural networks: A survey,”

CLIP - What is it?

- Connecting text and images…



guacamole (90.1%) Ranked 1 out of 101 labels

✓ a photo of **guacamole**, a type of food.

× a photo of **ceviche**, a type of food.

× a photo of **edamame**, a type of food.

× a photo of **tuna tartare**, a type of food.

× a photo of **hummus**, a type of food.

- Formally…
  - Method to learn **task-agnostic** image representations using natural language supervision.
- CLIP - Contrastive Language-Image Pretraining

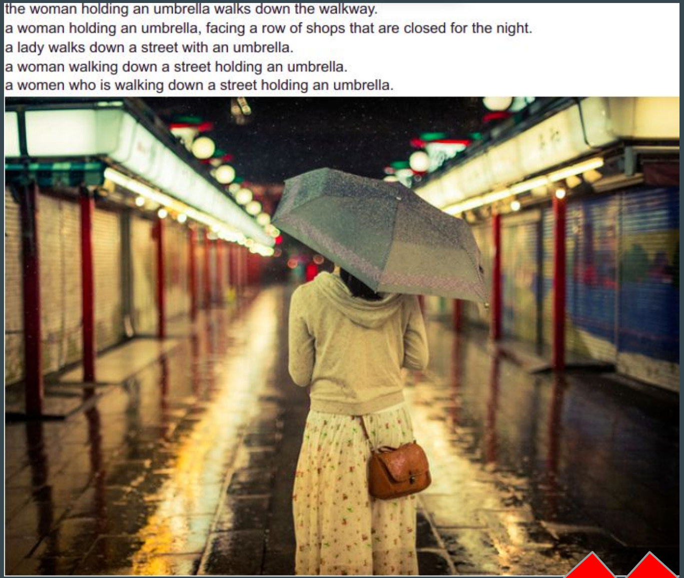
CLIP - Motivation

- Success of language models like GPT-3...
    - Uses web-scale raw, text data
    - Zero-shot transfer to downstream tasks.
- CLIP...
    - **aimed at developing GPT-3 like models for vision classification problems**.
- Once trained...
    - Can be applied to any visual classification benchmark (zero-shot).

CLIP - Dataset curation

- Benchmark image-text pair datasets:

MS-COCO



the woman holding an umbrella walks down the walkway.
a woman holding an umbrella, facing a row of shops that are closed for the night.
a lady walks down a street with an umbrella.
a woman walking down a street holding an umbrella.
a women who is walking down a street holding an umbrella.

Only around 100,000 images ✗

YFCC100M

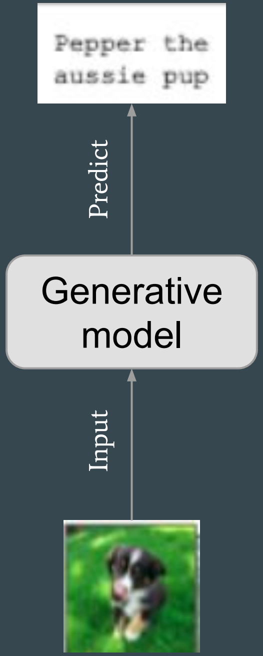
(a) *IMG_9793: Streetcar (Toronto Transit)* by Andy Nystrom `https://flic.kr/p/jciMdz`.

Noisy and **sparse** metadata - filtering reduces data by 6x ✗

CLIP - Dataset curation

- **Manual curation** - image-text pairs largely available publicly on internet.
- 400 million image-text pairs crawled from internet.
- Final dataset - **WebImageText (WIT)**
  - 400 million image-text pairs
  - Proprietary

CLIP - Problem formulation



Pepper the aussie pup

Predict

Generative model

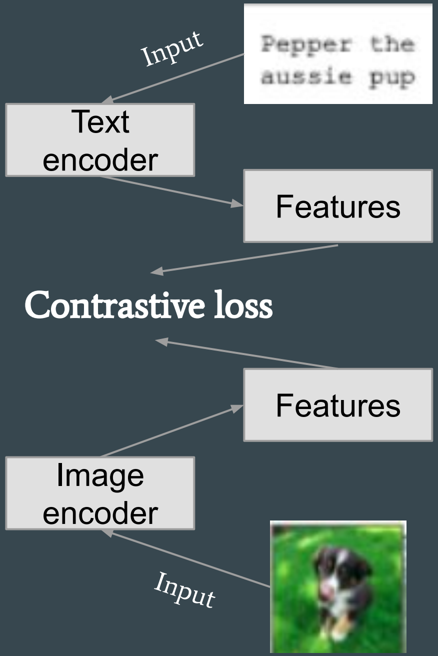
Input

Predict exact words...

HARD TASK

MORE COMPUTE

VS



Input

Pepper the aussie pup

Text encoder

Features

Contrastive loss

Features

Image encoder

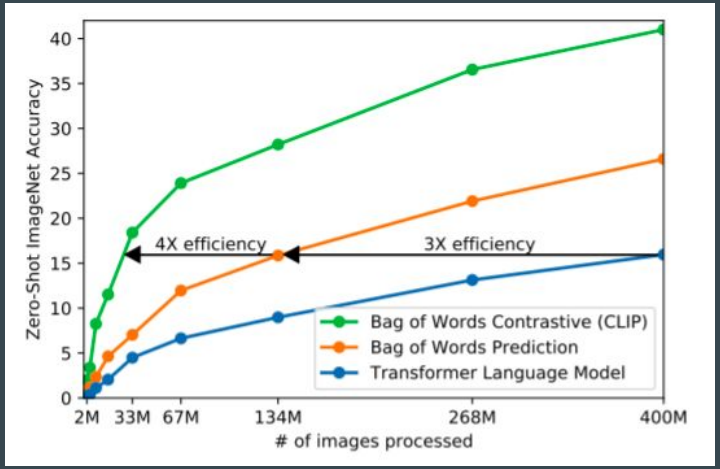
Input

Predict which text pairs with which image

EASIER TASK

LESS COMPUTE

EFFECTIVE

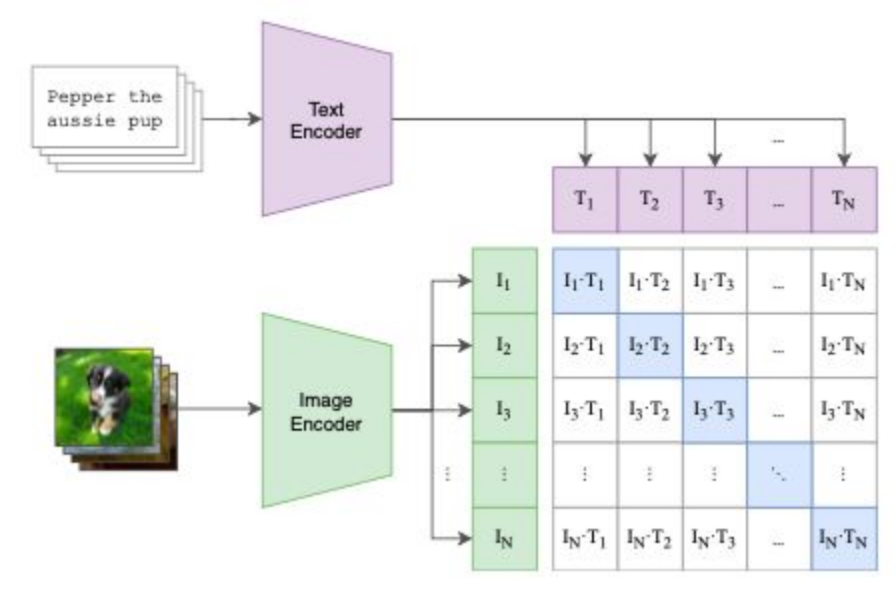


Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. arXiv preprint arXiv:1906.05849, 2019.

CLIP - Architecture

- Text encoder
  - Encodes natural language text into multi-modal space

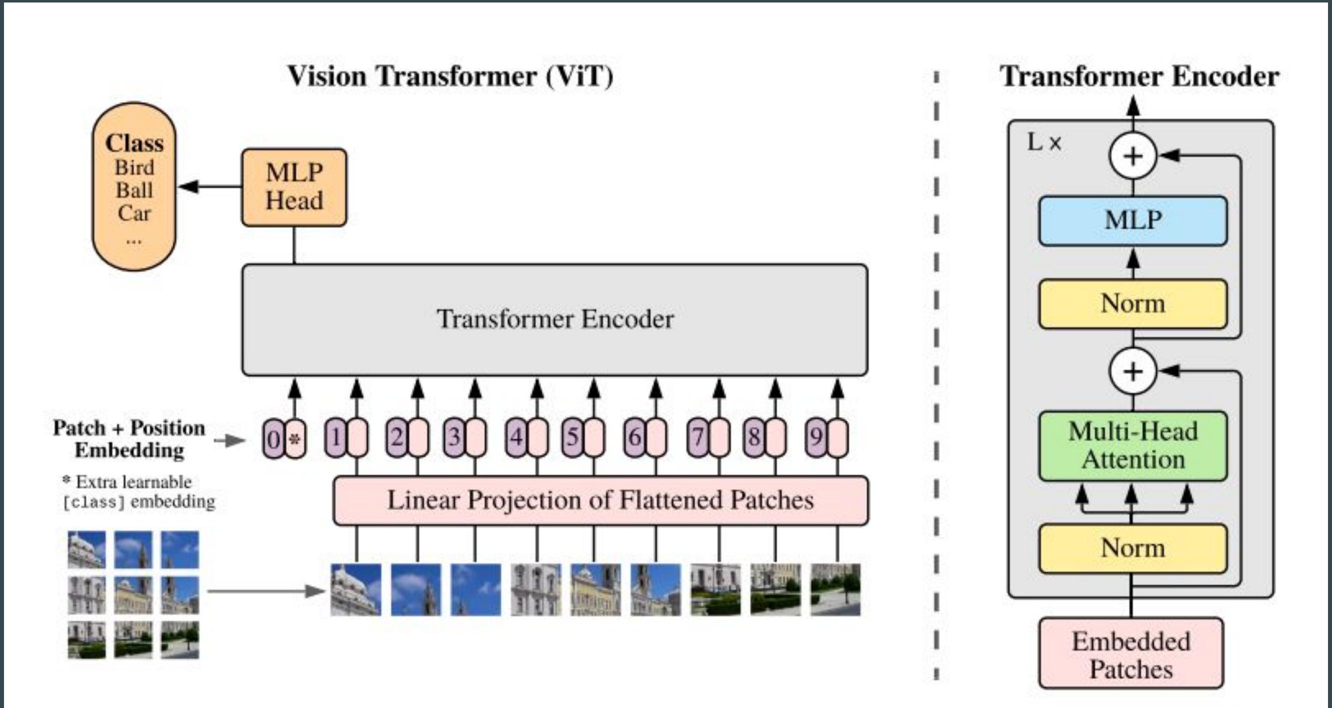
- Image encoder
  - Encodes image features into the same multi-modal space


Pepper the aussie pup
Text Encoder
T$_1$ T$_2$ T$_3$ ... T$_N$
Image Encoder
I$_1$
I$_2$
I$_3$
⋮
I$_N$
I$_1$·T$_1$ I$_1$·T$_2$ I$_1$·T$_3$ ... I$_1$·T$_N$
I$_2$·T$_1$ I$_2$·T$_2$ I$_2$·T$_3$ ... I$_2$·T$_N$
I$_3$·T$_1$ I$_3$·T$_2$ I$_3$·T$_3$ ... I$_3$·T$_N$
⋮ ⋮ ⋮ ⋱ ⋮
I$_N$·T$_1$ I$_N$·T$_2$ I$_N$·T$_3$ ... I$_N$·T$_N$

CLIP - Models

- Visual encoder:
    - 5 ResNet Model variants
        - ResNet-50
        - ResNet-101
        - ResNet-50x4, ResNet-50x16 and ResNet-50x64
            - Scaled EfficiencyNet-style models
    - 3 Vision-Transformer (ViT) variants
        - ViT-B/32 (Base model)
        - ViT-B/16
        - ViT-L/14 (Large model) - **BEST PERFORMANCE**
- Text encoder
    - GPT-2 like transformer

Visual Transformer (ViT) - Architecture



CLIP vision transformers are about **3x more compute efficient** than CLIP ResNets

Text Encoder - GPT-2

- Lower-cased BPE for token embedding
- 49,152 - vocab size
- 63M parameter model
- 12 layers
- 512 dimensions wide
- 8 attention heads

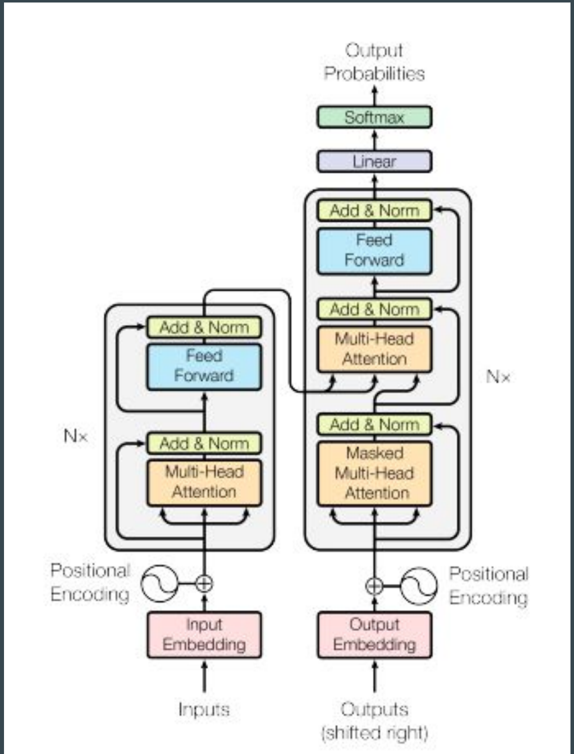
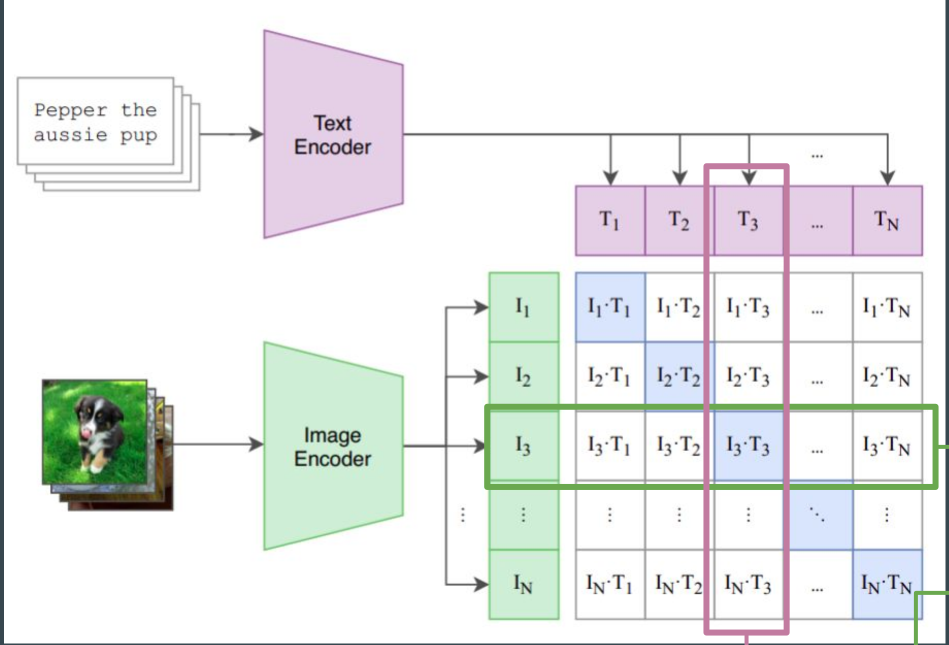


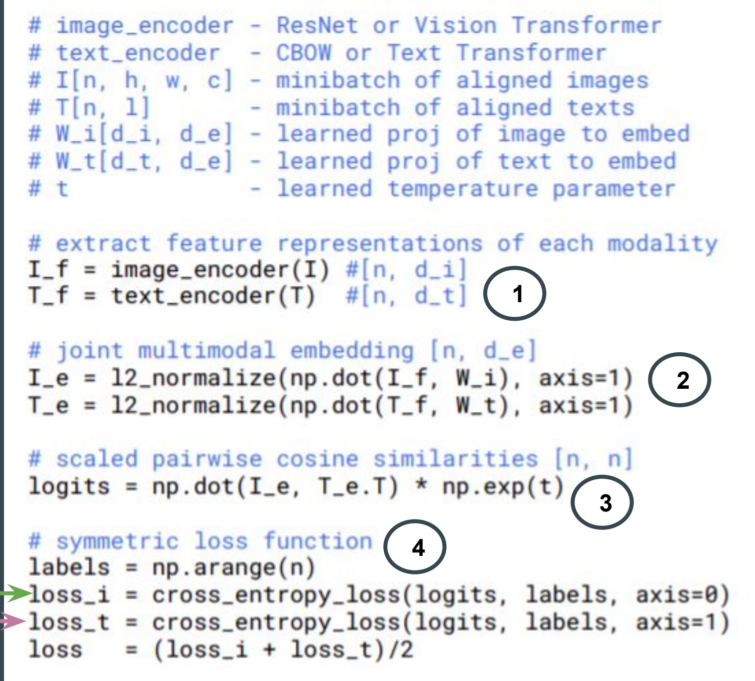
Figure 1: The Transformer - model architecture.

CLIP - Training

- Mini-batch size - **n**
- Text feature dim - **d_t**
- Image feature dim - **d_i**
- Multi-modal space dim - **d_e**



	T₁	T₂	T₃	...	T_N
I₁	I₁·T₁	I₁·T₂	I₁·T₃	...	I₁·T_N
I₂	I₂·T₁	I₂·T₂	I₂·T₃	...	I₂·T_N
I₃	I₃·T₁	I₃·T₂	I₃·T₃	...	I₃·T_N
⋮	⋮	⋮	⋮	⋱	⋮
I_N	I_N·T₁	I_N·T₂	I_N·T₃	...	I_N·T_N



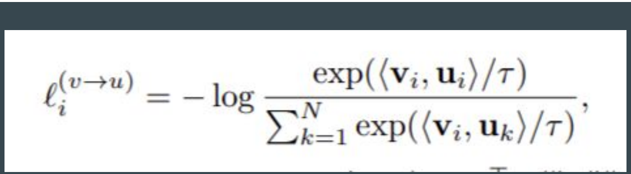
```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned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

1

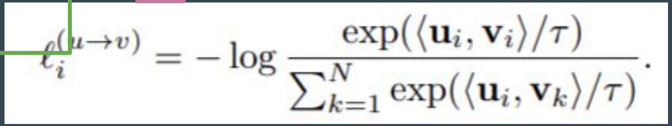
2

3

4



$$\ell_i^{(v\rightarrow u)} = -\log \frac{\exp(\langle \mathbf{v}_i, \mathbf{u}_i \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle \mathbf{v}_i, \mathbf{u}_k \rangle / \tau)},$$

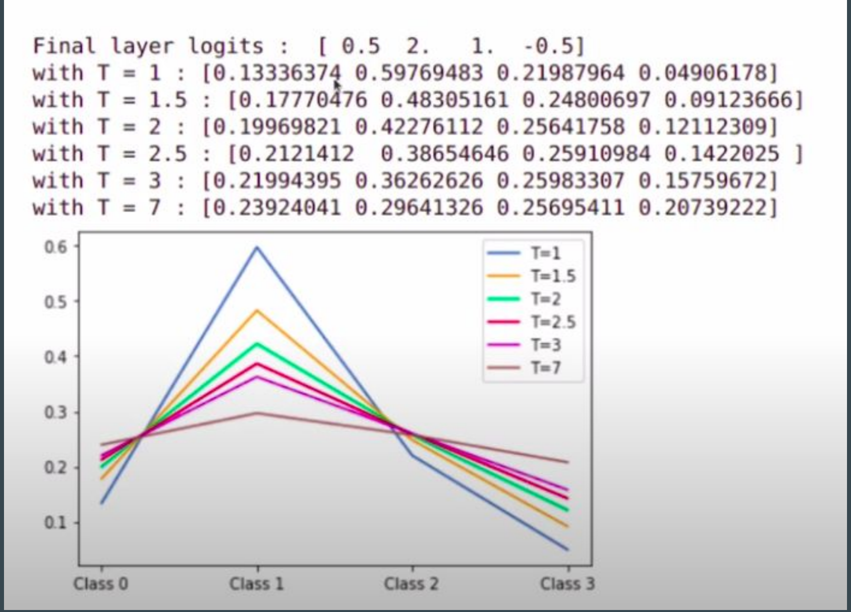


$$\ell_i^{(u\rightarrow v)} = -\log \frac{\exp(\langle \mathbf{u}_i, \mathbf{v}_i \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle \mathbf{u}_i, \mathbf{v}_k \rangle / \tau)}.$$

Training - Temperature parameter

- Knowledge distillation concept
- Softmax with temperature parameter (soft-targets)

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$



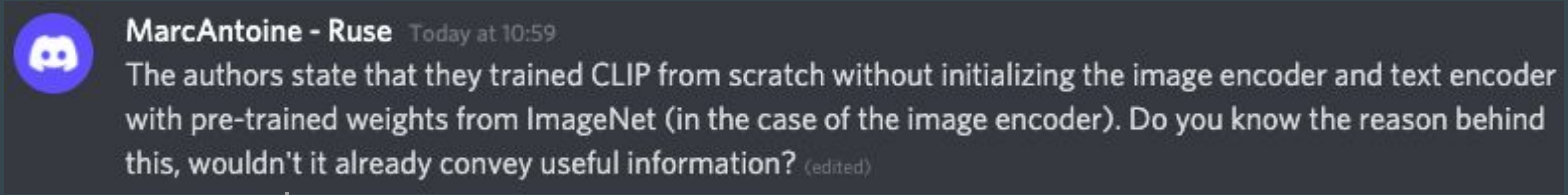
```
Final layer logits :  [ 0.5  2.   1.  -0.5]
with T = 1 : [0.13336374 0.59769483 0.21987964 0.04906178]
with T = 1.5 : [0.17770476 0.48305161 0.24800697 0.09123666]
with T = 2 : [0.19969821 0.42276112 0.25641758 0.12112309]
with T = 2.5 : [0.2121412  0.38654646 0.25910984 0.142225 ]
with T = 3 : [0.21994395 0.36262626 0.25983307 0.15759672]
with T = 7 : [0.23924041 0.29641326 0.25695411 0.20739222]
```

https://openaccess.thecvf.com/content_cvpr_2018/CameraReady/0801.pdf

CLIP - Training configuration

- **Input image - 224x224 dims**
  - Tried 336x336 dims for ViT-L/14 - BEST PERFORMANCE
- **Optimizer** - Adam with weight decay regularization
- **Minibatch size** - 32,768 image-text pairs
- **Epochs** - 32

Want the minibatch size to be larger, but hardware constraints

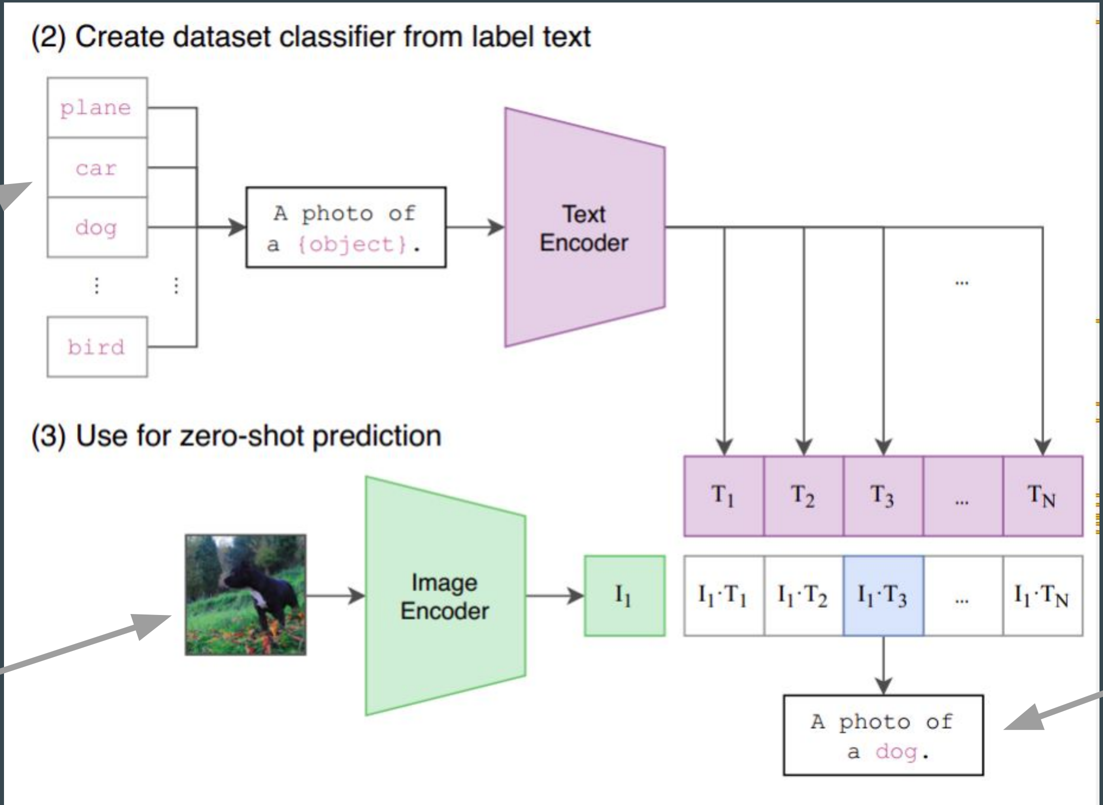


MarcAntoine - Ruse Today at 10:59

The authors state that they trained CLIP from scratch without initializing the image encoder and text encoder with pre-trained weights from ImageNet (in the case of the image encoder). Do you know the reason behind this, wouldn't it already convey useful information? (edited)

1. With enough compute and data, models can learn effective representations (ViT)
2. Negate pretrained bias

CLIP - Test phase (zero-shot)

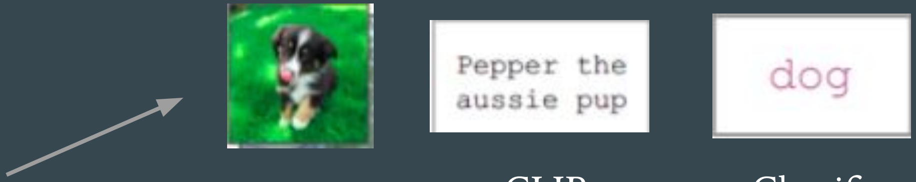


Test categories

Test image

Prediction

CLIP - Comparison with classification models

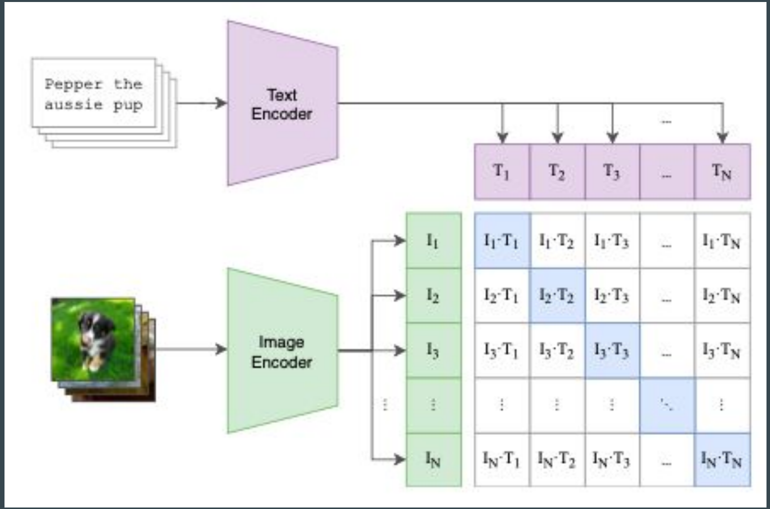


Input CLIP Classifier

1. Better representations
2. Works on diverse tasks
3. No manual labelling
4. Using language supervision helps understand context (verbs)
   - Avoids polysemy - classify better
   - Helps in action recognition dataset

CLIP as a linear classifier

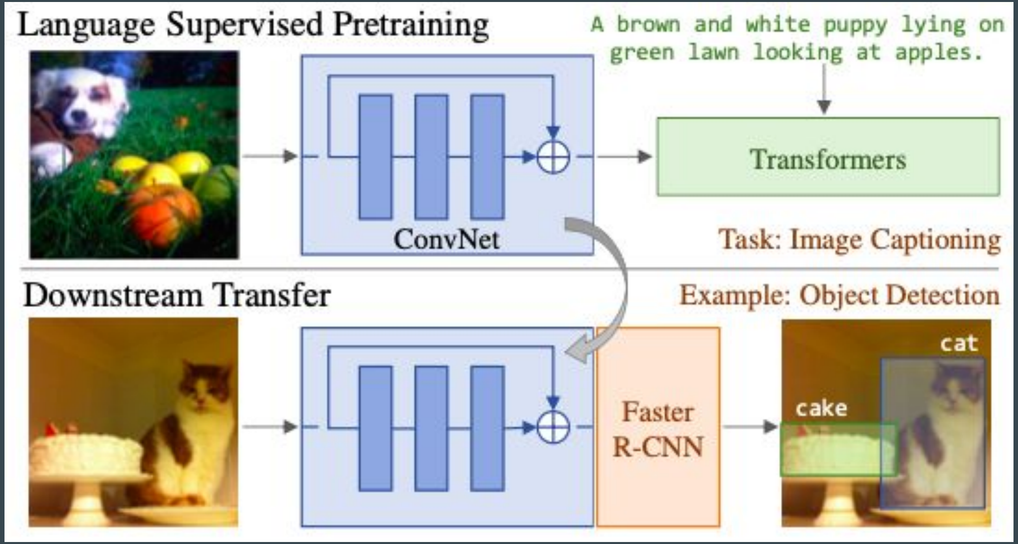
- The cosine similarity between the text and image representations is given by
  - $y = \text{softmax}(T \cdot I)$ where T is the text feature matrix and I is the image feature matrix
- The above equation is similar to $y = \text{softmax}(W^T X)$ - a linear classifier with no bias.
- Hence, we can perceive the CLIP model as a linear classifier with:
  - image encoder (visual backbone) feature as input (X)
  - Text encoder feature as the corresponding weight for the given input image.



	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

Related works- Virtex

- Jointly trains an image CNN and text transformer using image caption pairs for the task of image captioning.
- Learned CNN is used for downstream tasks.

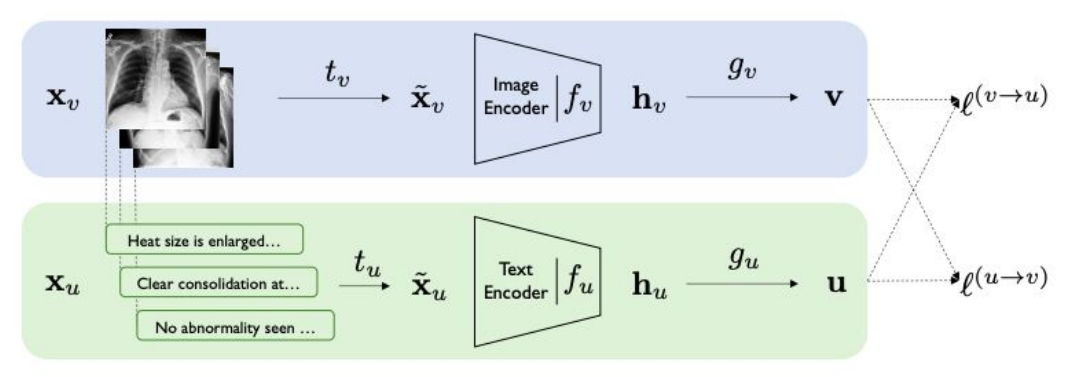


Language Supervised Pretraining
A brown and white puppy lying on green lawn looking at apples.
Transformers
ConvNet
Task: Image Captioning
Downstream Transfer
Example: Object Detection
Faster R-CNN
cat
cake

Desai et al.- VirTex: Learning Visual Representations from Textual Annotations

Related works- ConVIRT

- Trained on pairs of Chest Radiograph images and text description of characteristics/abnormalities.
- The model maximizes the agreement between the true image-text representation pairs with bidirectional losses



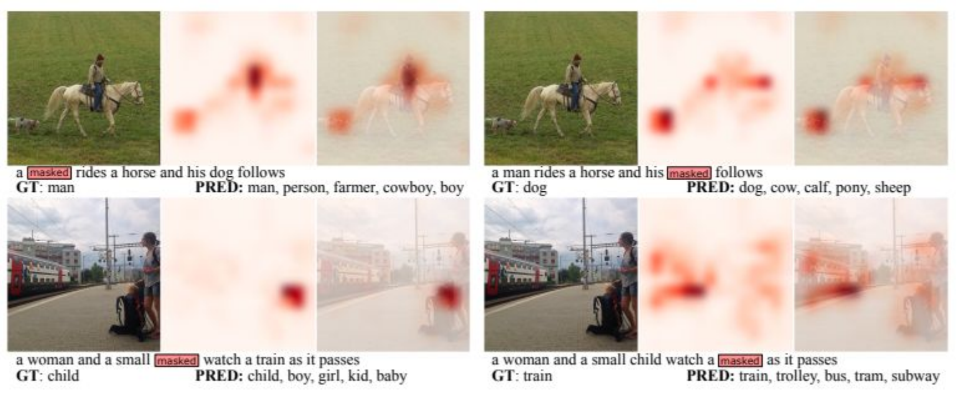
$$\ell_i^{(v\to u)} = -\log \frac{\exp(\langle \mathbf{v}_i, \mathbf{u}_i \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle \mathbf{v}_i, \mathbf{u}_k \rangle / \tau)}$$

$$\langle \mathbf{v}, \mathbf{u} \rangle = \mathbf{v}^\top \mathbf{u} / \|\mathbf{v}\| \|\mathbf{u}\|$$

Yuhao Zhang et al- Contrastive Learning of Medical Visual Representations from Paired Images and Text

Related works- ICMLM

- Image-conditioned masked language modeling (ICMLM)
- The model learns visual representations from image-caption pairs
- Masks tokens in captions are predicted by fusing visual and textual cues.
- Visual attention changes as different tokens in a caption are masked.



a masked rides a horse and his dog follows
GT: man **PRED:** man, person, farmer, cowboy, boy

a man rides a horse and his masked follows
GT: dog **PRED:** dog, cow, calf, pony, sheep

a woman and a small masked watch a train as it passes
GT: child **PRED:** child, boy, girl, kid, baby

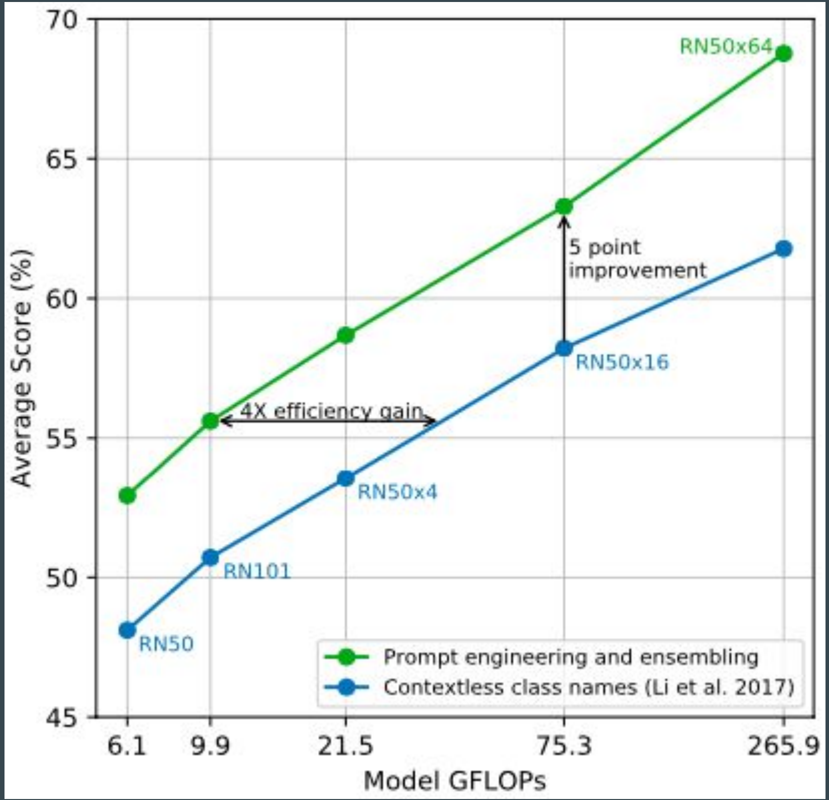
a woman and a small child watch a masked as it passes
GT: train **PRED:** train, trolley, bus, tram, subway

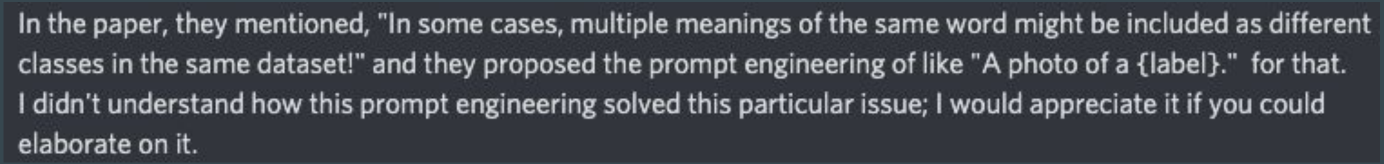
Mert Bulent Sariyildiz et al- Learning Visual Representations with Caption Annotations

Experiments

Prompt Engineering and Ensembling

- Customizing prompt text to each task improves zero shot performance.
- Use ‘A photo of a dog, a type of pet.’ to improve performance further.
- Ensemble models using different prompts i.e. ensembling over embedding space instead of probabilistic space.
- This is an image of a boxer, a breed of dog. Boxer can be a type of athlete or a breed of dog, providing context to the text prompt helps improve the model performance.

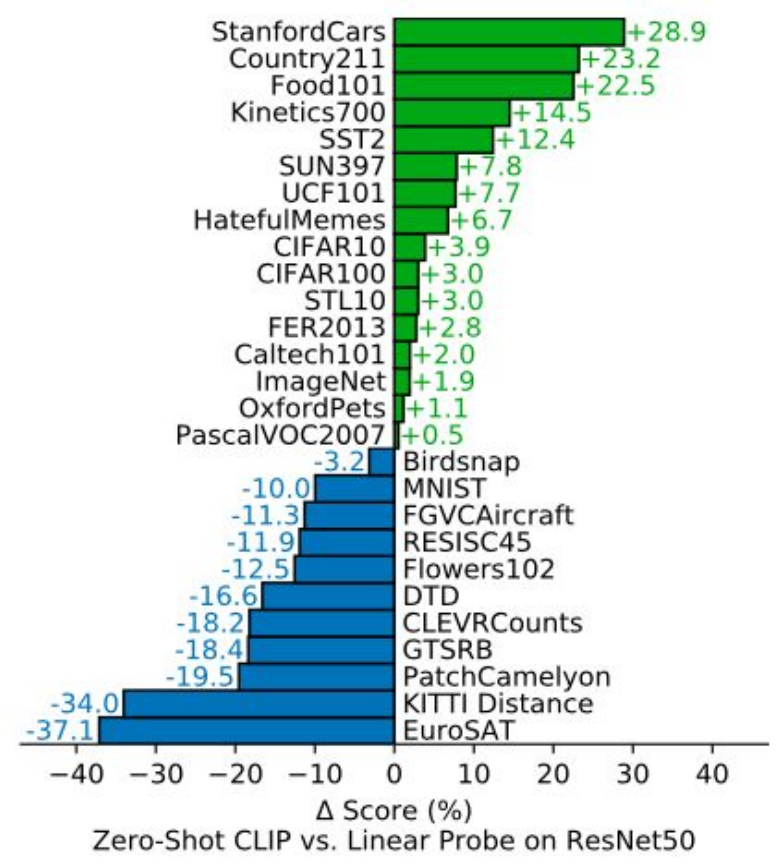



In the paper, they mentioned, "In some cases, multiple meanings of the same word might be included as different classes in the same dataset!" and they proposed the prompt engineering of like "A photo of a {label}." for that.
I didn't understand how this prompt engineering solved this particular issue; I would appreciate it if you could elaborate on it.

Experiments

Zero Shot CLIP vs Linear Probe on Resnet 50

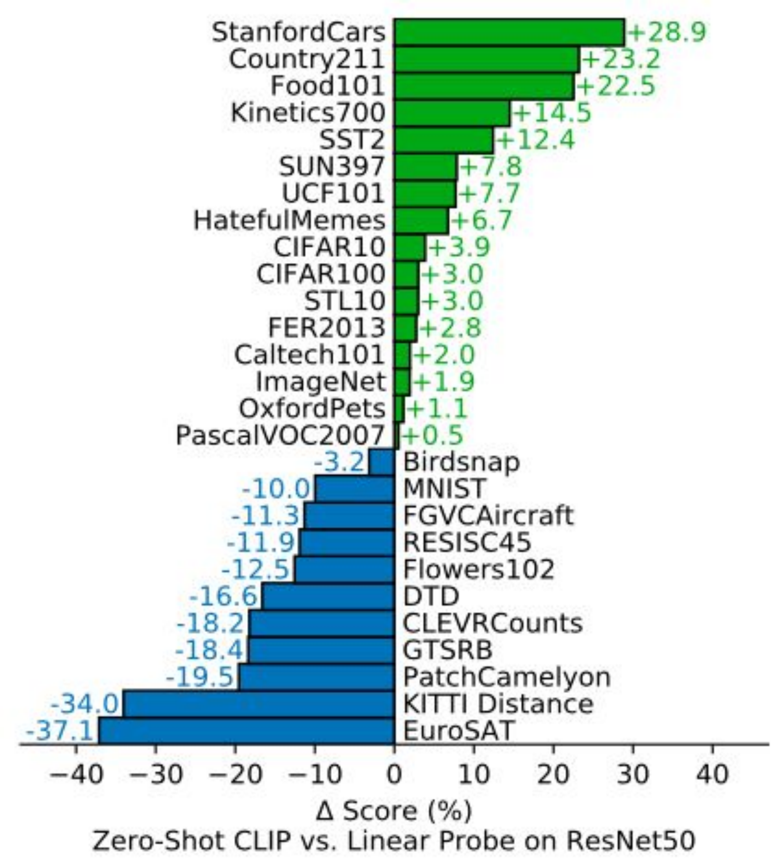
- Linear probe on Resnet 50 means that a fully supervised linear classifier is fitted on ResNet-50 features.
- Clip outperforms Resnet in 16/27 datasets.
- Clip performs better for Cars and pets and Resnet performs better for Birds and Aircraft.
- Probably due to varying amount of per-task supervision between WIT vs Imagenet.



Experiments

Zero Shot CLIP vs Linear Probe on Resnet 50

- On general datasets like CIFAR10, STL10, ImageNet CLIP performs slightly better due to a larger and more varied training dataset for CLIP.
- On action recognition datasets like Kinetics700 and UCF101, CLIP outperform Resnet because natural language provides better supervision for verbs.

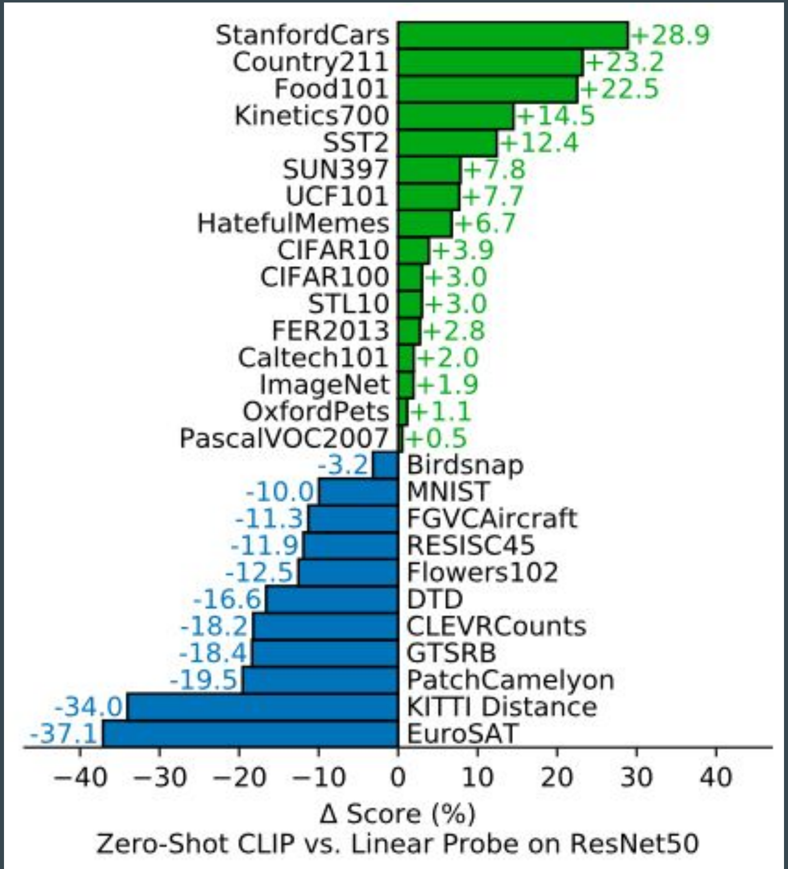


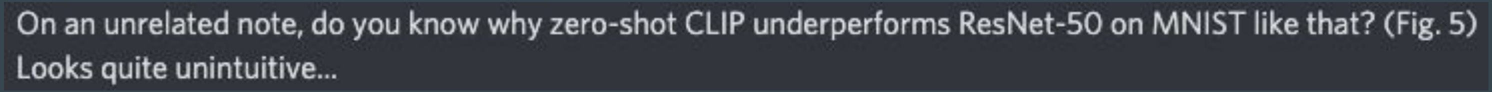
Zero-Shot CLIP vs. Linear Probe on ResNet50

Experiments

Zero Shot CLIP vs Linear Probe on Resnet 50

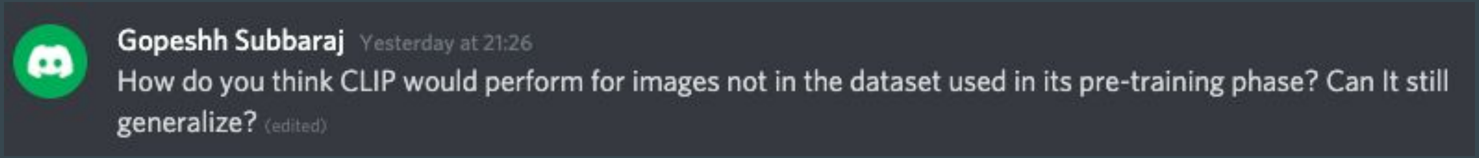
- On abstract tasks such as counting objects in synthetic scenes(CLEVRCounts), Satellite Image Classification(EuroSAT), CLIP performs poorly.
- Surprisingly, it performs poorly on MNIST due to lack of overlap between WIT and MNIST datasets.





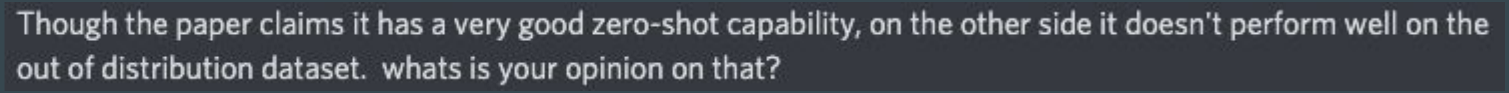
On an unrelated note, do you know why zero-shot CLIP underperforms ResNet-50 on MNIST like that? (Fig. 5) Looks quite unintuitive...

Question



Gopeshh Subbaraj Yesterday at 21:26

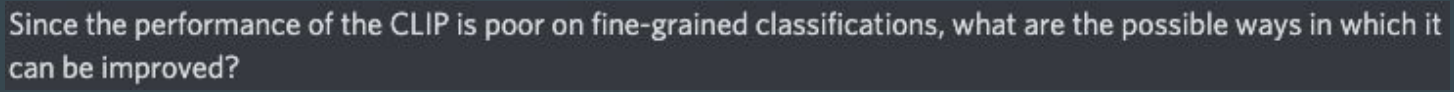
How do you think CLIP would perform for images not in the dataset used in its pre-training phase? Can It still generalize? (edited)



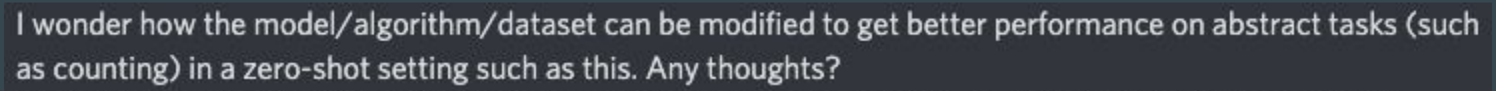
Though the paper claims it has a very good zero-shot capability, on the other side it doesn't perform well on the out of distribution dataset. whats is your opinion on that?

- As we have seen in the previous few slides, CLIP performs well for general datasets, action recognition.
- It performed well for few specific classes like birds and pets but it performed badly for classes like birds and aeroplanes depending on the amount of supervision it received during training.
- But it performed badly for abstract tasks like counting, and certain image types that it did not encounter during the training like Satellite Images and MNIST.
- CLIP mainly increases the distribution by using a large dataset from the internet and generating free labels using text captions. This model does not address the problem of out of distribution generalization.

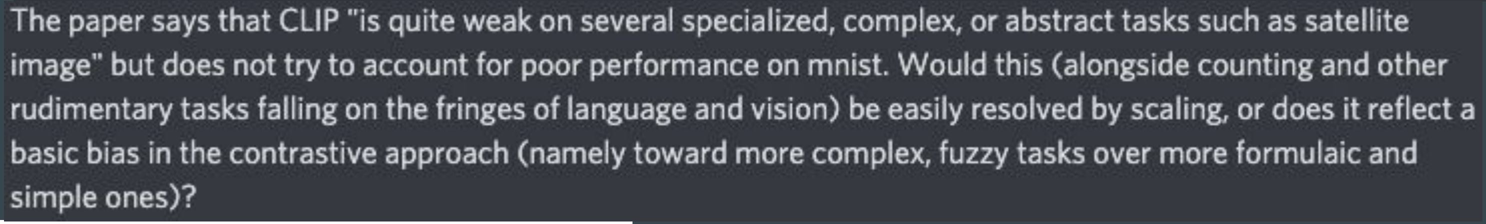
Question



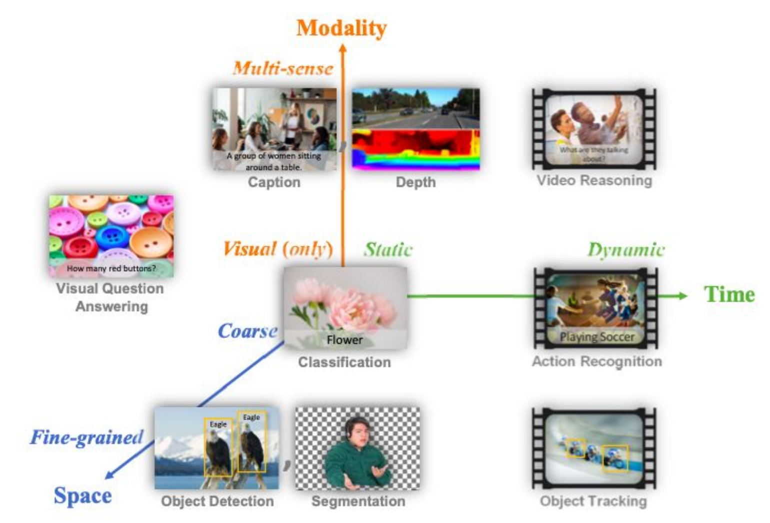
Since the performance of the CLIP is poor on fine-grained classifications, what are the possible ways in which it can be improved?



I wonder how the model/algorithm/dataset can be modified to get better performance on abstract tasks (such as counting) in a zero-shot setting such as this. Any thoughts?



The paper says that CLIP "is quite weak on several specialized, complex, or abstract tasks such as satellite image" but does not try to account for poor performance on mnist. Would this (alongside counting and other rudimentary tasks falling on the fringes of language and vision) be easily resolved by scaling, or does it reflect a basic bias in the contrastive approach (namely toward more complex, fuzzy tasks over more formulaic and simple ones)?



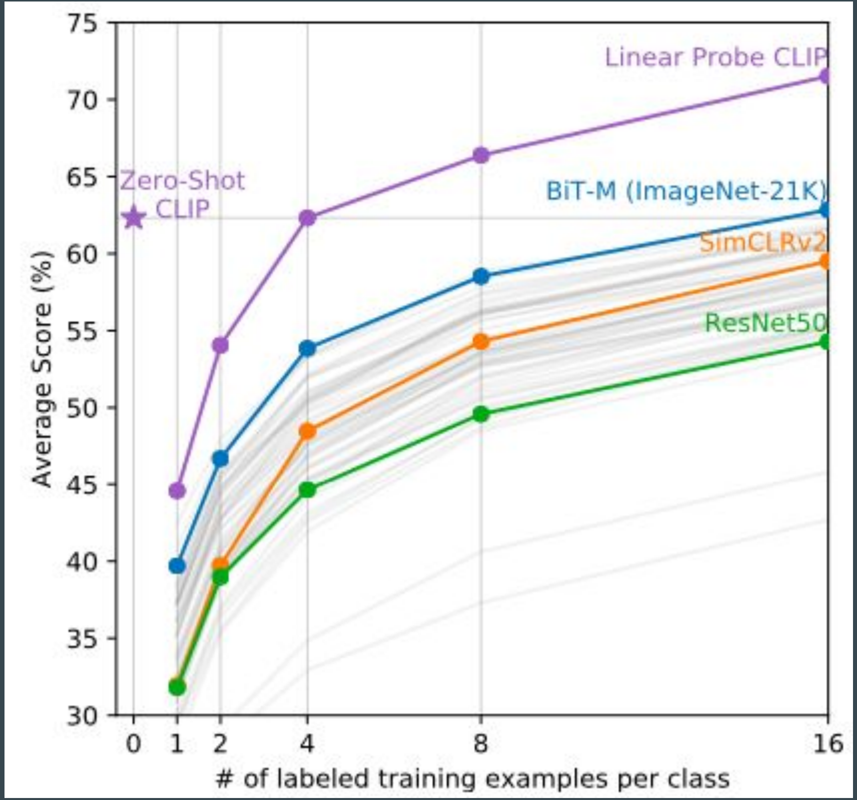
- Florence can be used for more complex tasks like object detection, vqa and complex categories like satellite imagery, chest x-ray images.
- It uses more better architectures like Swin Transformer, and performs better model finetuning than linear probe.

Yuan et al.- Florence: A New Foundation Model for Computer Vision
Liu et al.- Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

Experiments

CLIP with Linear Probe

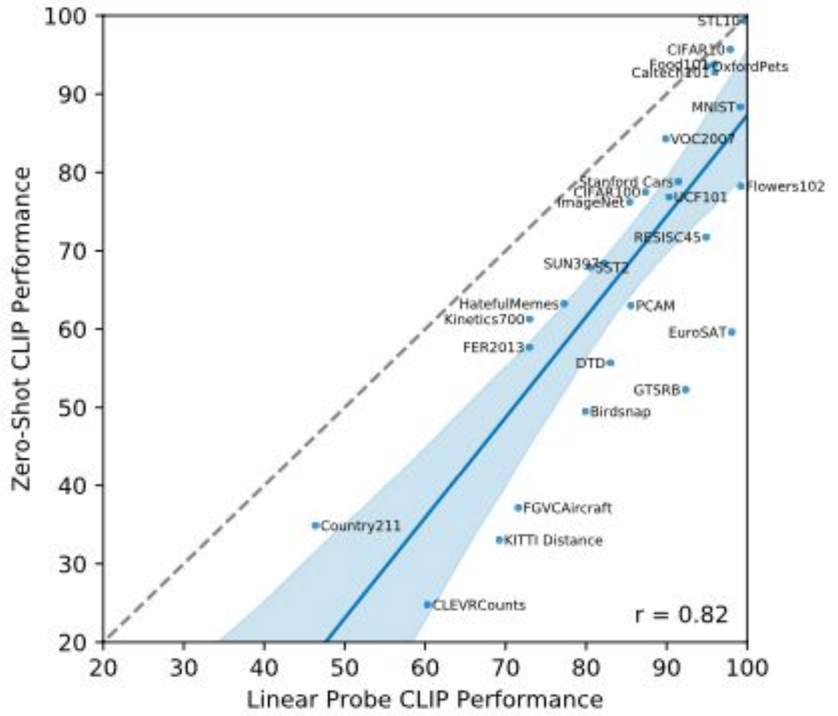
- Zero Shot CLIP performs as well as 4 Shot CLIP with linear probe because the number of samples are too few to effectively train the linear classifier.
- The performance of linear probe CLIP increases with the increase in the number of samples per class.
- Training linear probe with all samples is the upper bound for CLIP.



Experiments

CLIP with Linear Probe

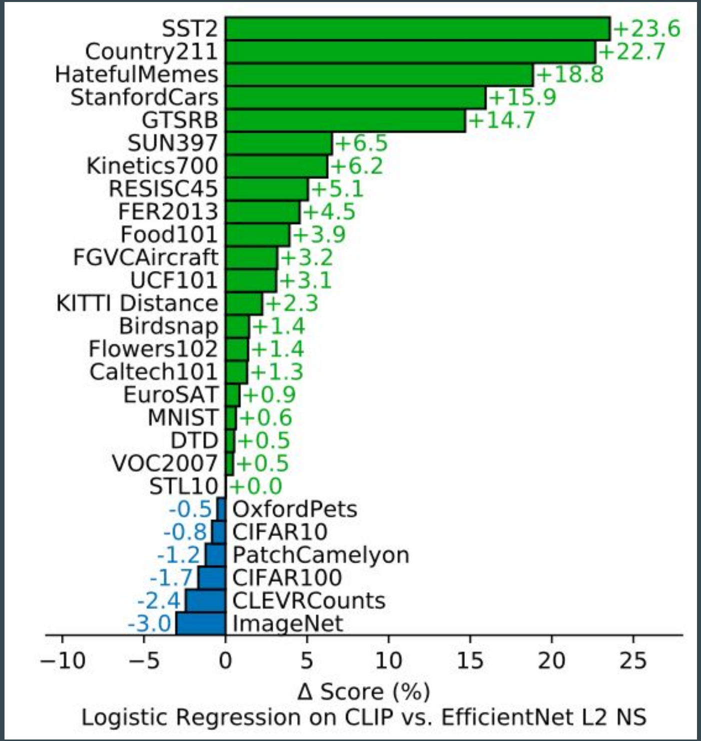
- Training Linear Probe with CLIP improve its performance by 10-25% for most datasets.

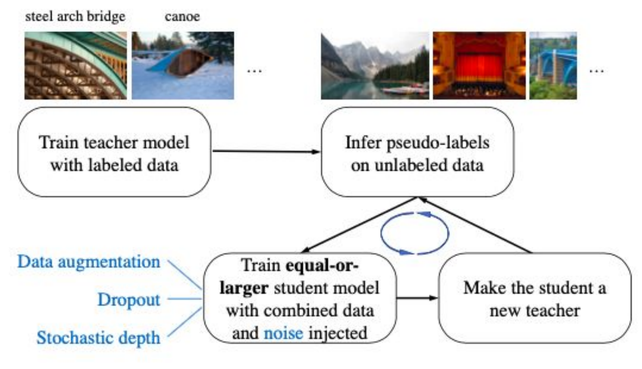


Experiments

CLIP with Linear Probe

- Train Logistic Regression on CLIP performs better than Noisy Student EfficientNet L2 on 21/27 datasets
- EfficientNet outperforms CLIP on ImageNet dataset because it was trained on it.





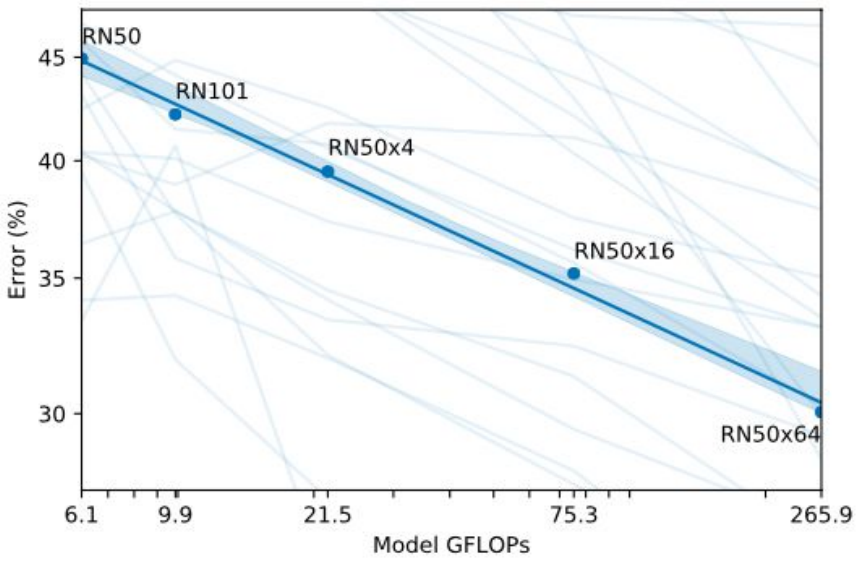
steel arch bridge
canoe
...
...
Train teacher model with labeled data
Infer pseudo-labels on unlabeled data
Data augmentation
Dropout
Stochastic depth
Train **equal-or-larger** student model with combined data and noise injected
Make the student a new teacher

Xie et al. - Self-training with Noisy Student improves ImageNet classification

Experiments

Compute NSL for CLIP

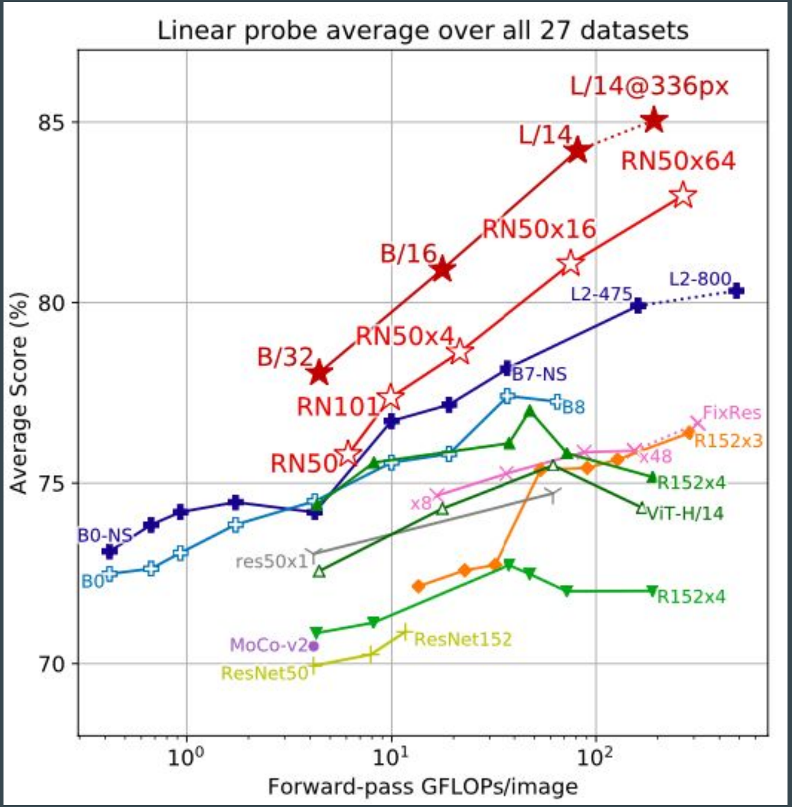
- CLIP's performance improves smoothly with the increase in compute.
- The average zero shot loss across 36 dataset decreases consistently.
- The individual loss for the datasets are much noisier.

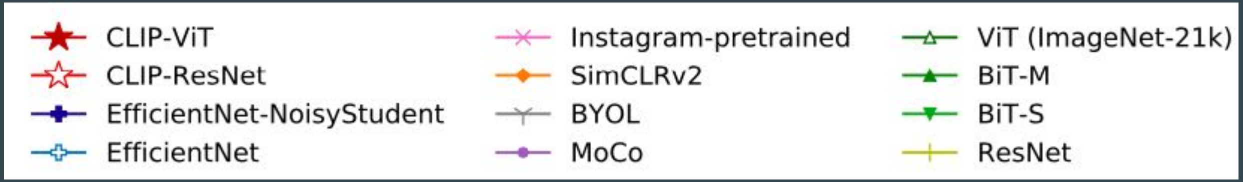


Experiments

Compute NSL for CLIP

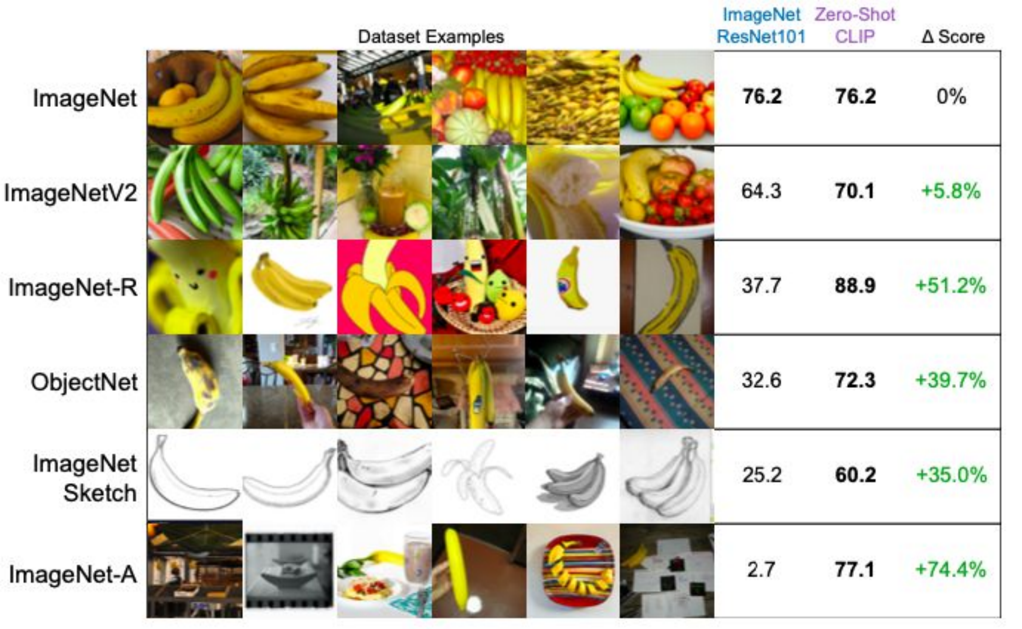
- VIT CLIP performs better than Resnet CLIP.
- CLIP VIT are 3x more compute efficient than CLIP Resnet.



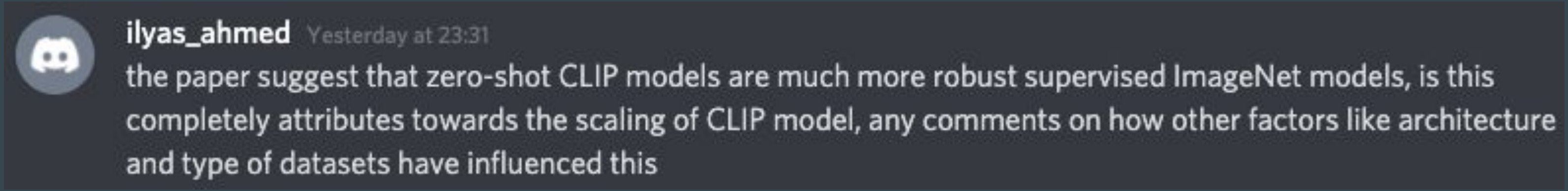


Experiments

- Zero-shot CLIP is much more robust to distribution shift than standard ImageNet models. ViT-L/14 vs Resnet-101
- The robustness is due to large dataset size (not specific to imagenet classes) , natural language supervision (word synonyms).



	Dataset Examples	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet		76.2	76.2	0%
ImageNetV2		64.3	70.1	+5.8%
ImageNet-R		37.7	88.9	+51.2%
ObjectNet		32.6	72.3	+39.7%
ImageNet Sketch		25.2	60.2	+35.0%
ImageNet-A		2.7	77.1	+74.4%

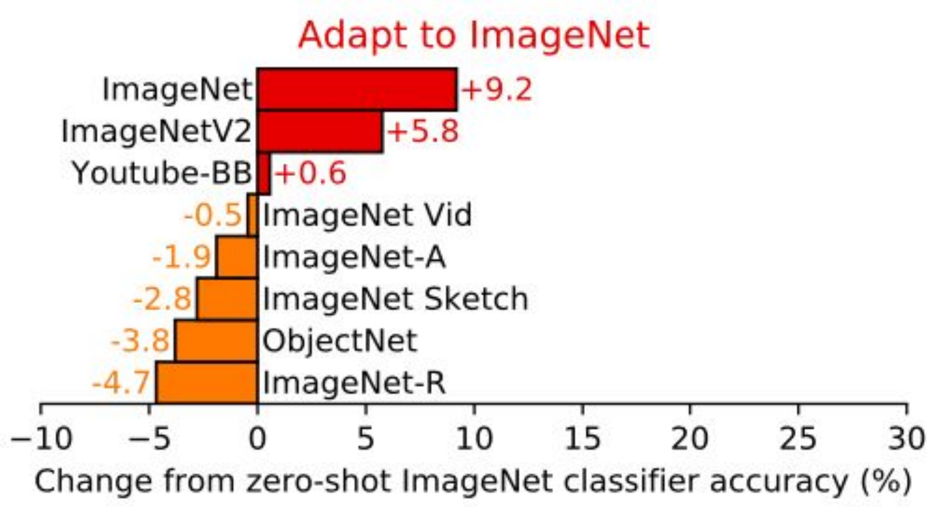


ilyas_ahmed Yesterday at 23:31

the paper suggest that zero-shot CLIP models are much more robust supervised ImageNet models, is this completely attributes towards the scaling of CLIP model, any comments on how other factors like architecture and type of datasets have influenced this

Experiments

- The model is adapted to Imagenet using Linear Probe.
- While supervised adaptation to ImageNet increases ImageNet accuracy by 9.2%, it slightly reduces average robustness.



Broader Impacts

- CLIP has a wide range of capabilities and can be tuned to a wide range of application using zero-shot learning and linear probe.
- But it also has the potential to amplify the inherent biases in data.

Bias

- CLIP has been test on FairFace dataset.
- FairFace is a face image dataset designed to balance age, gender, and race.
- It categorizes gender into 2 groups: female and male and race into 7 groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino.

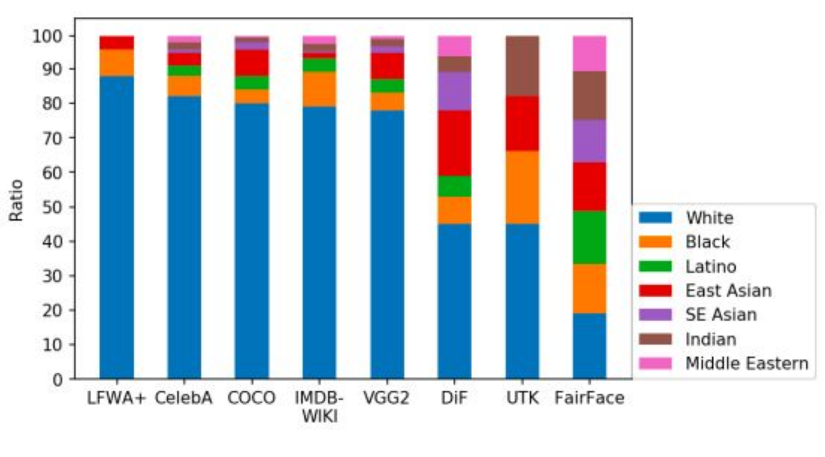
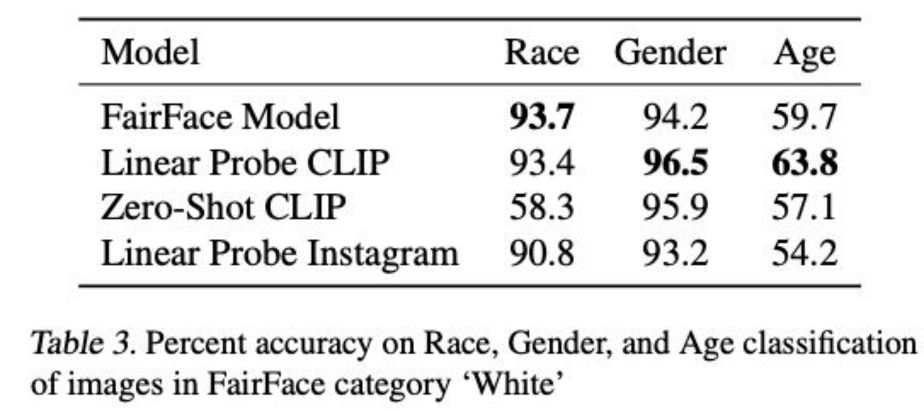


Figure 2: Racial compositions in face datasets.

K. Kärkkäinen and J. Joo, "FairFace: Face attribute dataset for balanced race, gender, and age," Aug. 2019.

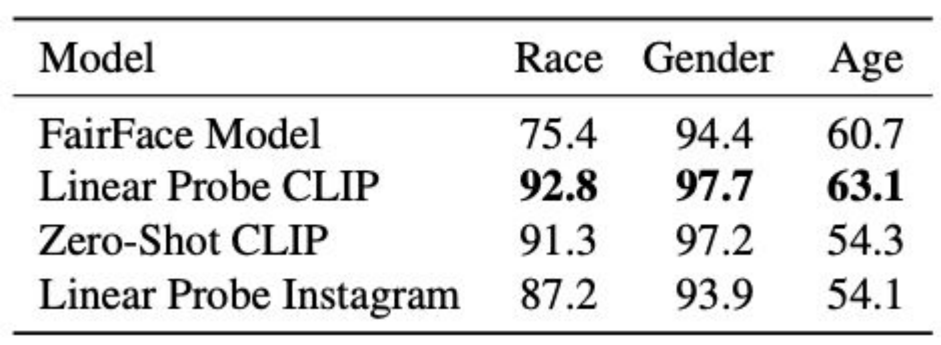
Bias

- FairFace Model - Resnet34
- Instagram Model- Transfer learning of ResNeXt-101 to predict hashtags on billions of social media hashtags.
- Zero Shot CLIP performs better than FairFace model on 3/6 categories.
- Linear probe CLIP performs better than FairFace models on most categories.



Model	Race	Gender	Age
FairFace Model	**93.7**	94.2	59.7
Linear Probe CLIP	93.4	**96.5**	**63.8**
Zero-Shot CLIP	58.3	95.9	57.1
Linear Probe Instagram	90.8	93.2	54.2

Table 3. Percent accuracy on Race, Gender, and Age classification of images in FairFace category 'White'



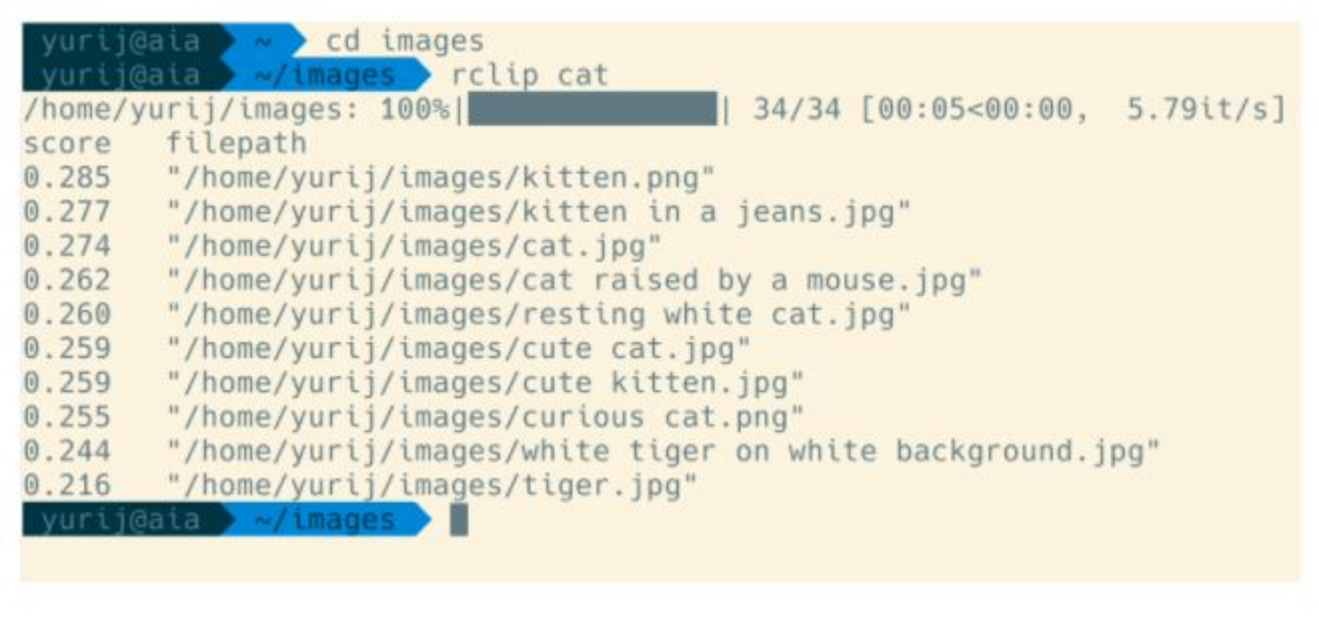
Model	Race	Gender	Age
FairFace Model	75.4	94.4	60.7
Linear Probe CLIP	**92.8**	**97.7**	**63.1**
Zero-Shot CLIP	91.3	97.2	54.3
Linear Probe Instagram	87.2	93.9	54.1

Race, Age, Gender Classification for Non-White Categories

K. Kärkkäinen and J. Joo, “FairFace: Face attribute dataset for balanced race, gender, and age,” Aug. 2019.
Mahajan et al. : Exploring the Limits of Weakly Supervised Pretraining

Applications- Image Search

- Get Image Features by passing the images through the CLIP model and storing it.
- Compute Similarity between the image features and text features.

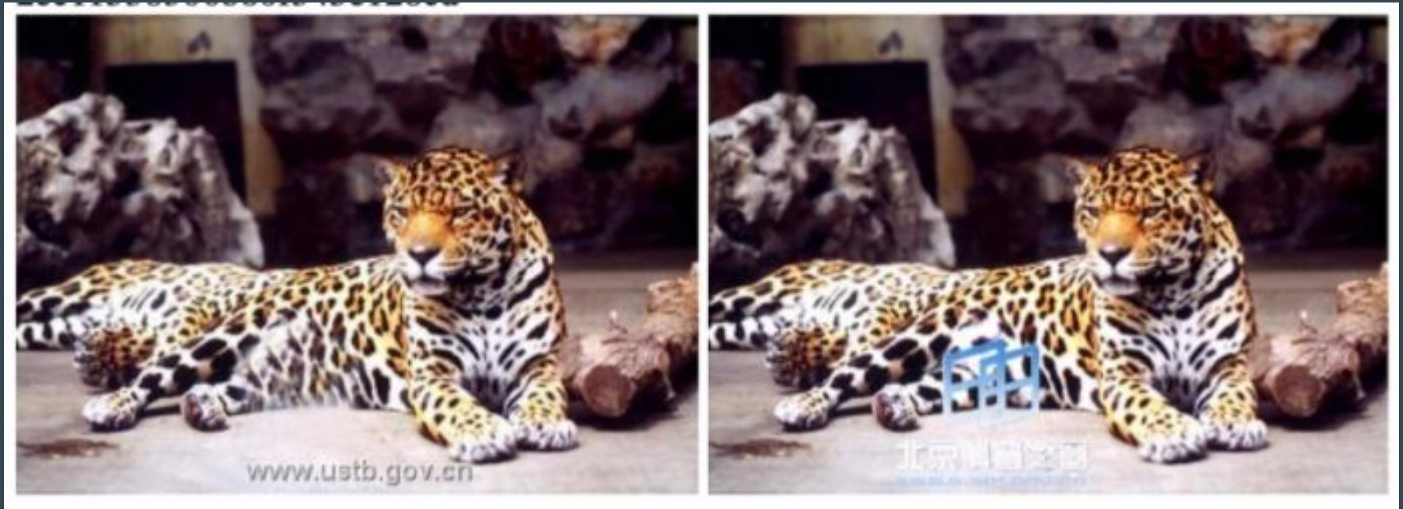


```
yurij@aia  ~  cd images
yurij@aia  ~/images  rclip cat
/home/yurij/images: 100%|█████████████| 34/34 [00:05<00:00,  5.79it/s]
score   filepath
0.285   "/home/yurij/images/kitten.png"
0.277   "/home/yurij/images/kitten in a jeans.jpg"
0.274   "/home/yurij/images/cat.jpg"
0.262   "/home/yurij/images/cat raised by a mouse.jpg"
0.260   "/home/yurij/images/resting white cat.jpg"
0.259   "/home/yurij/images/cute cat.jpg"
0.259   "/home/yurij/images/cute kitten.jpg"
0.255   "/home/yurij/images/curious cat.png"
0.244   "/home/yurij/images/white tiger on white background.jpg"
0.216   "/home/yurij/images/tiger.jpg"
yurij@aia  ~/images  █
```

https://mikhalevi.ch/rclip-an-ai-powered-command-line-photo-search-tool/

Applications- Image Similarity

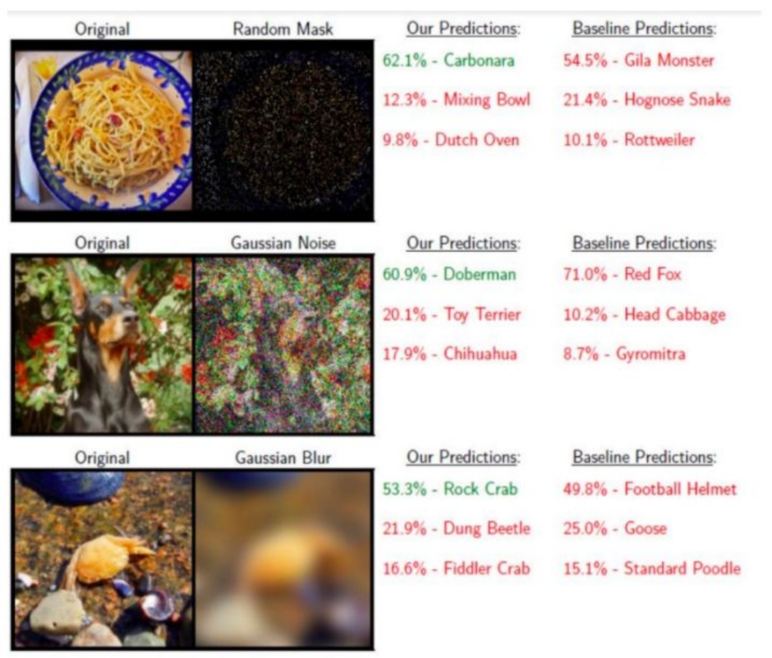
- Image Similarity- These images match even though they have different watermarks.



https://blog.roboflow.com/openai-clip/

Applications- Deciphering Corrupted Images

- CLIP performs better than Imagenet trained Resnet-101 for corrupted and adversarial images.



Our Predictions:	Baseline Predictions:
62.1% - Carbonara	54.5% - Gila Monster
12.3% - Mixing Bowl	21.4% - Hognose Snake
9.8% - Dutch Oven	10.1% - Rottweiler

Our Predictions:	Baseline Predictions:
60.9% - Doberman	71.0% - Red Fox
20.1% - Toy Terrier	10.2% - Head Cabbage
17.9% - Chihuahua	8.7% - Gyromitra

Our Predictions:	Baseline Predictions:
53.3% - Rock Crab	49.8% - Football Helmet
21.9% - Dung Beetle	25.0% - Goose
16.6% - Fiddler Crab	15.1% - Standard Poodle

Sriram et al. - Inverse Problems Leveraging Pre-trained Contrastive Representations

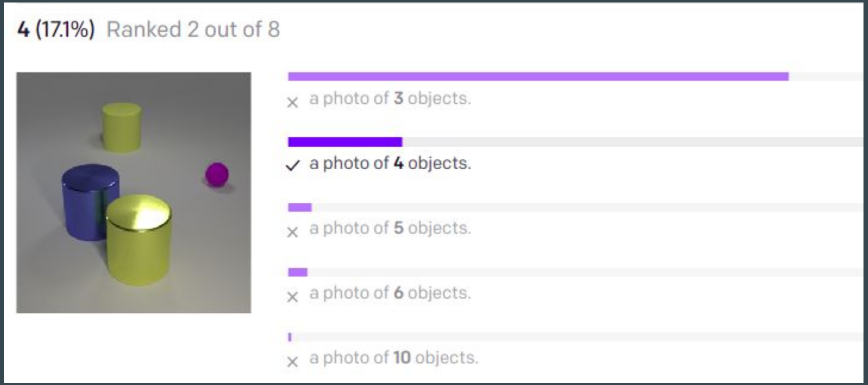
CLIP - How is it better?

1. CLIP is highly efficient
   - Earlier: ResNeXt101-32x48d - 19 GPU years for Imagenet classes
   - Training efficiency
     1. Contrastive objective ~ 4x less compute
     2. Vision Transformer ~ 3x less compute
   - Learning visual concepts directly from natural language
     1. Flexible and general than existing ImageNet models.
2. CLIP is flexible and general
   - Diverse tasks
   - Robust to distribution and domain

1. Fine-grained object classification
2. Action recognition in videos,
3. OCR
4. ……..

CLIP - Limitations

1. Doesnt work well on **abstract or complex tasks**
   - Counting the number of objects in an image, etc
2. Zero-shot CLIP struggles on **fine-grained classification tasks**
   - Tasks to differentiate between bird categories, car models, etc



4 (17.1%) Ranked 2 out of 8

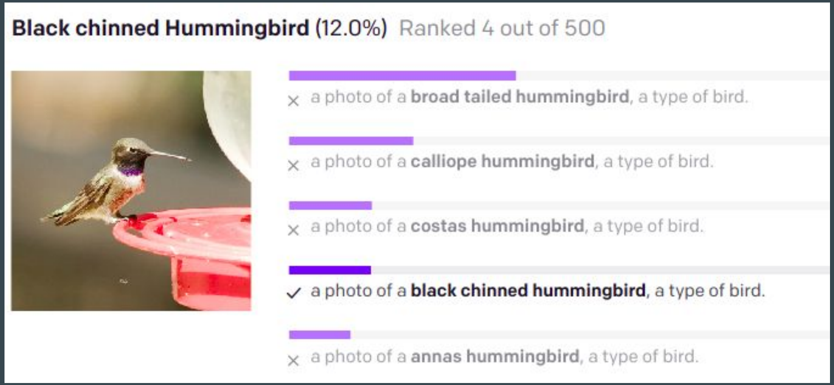
× a photo of **3** objects.

✓ a photo of **4** objects.

× a photo of **5** objects.

× a photo of **6** objects.

× a photo of **10** objects.



Black chinned Hummingbird (12.0%) Ranked 4 out of 500

× a photo of a **broad tailed hummingbird**, a type of bird.

× a photo of a **calliope hummingbird**, a type of bird.

× a photo of a **costas hummingbird**, a type of bird.

✓ a photo of a **black chinned hummingbird**, a type of bird.

× a photo of a **annas hummingbird**, a type of bird.

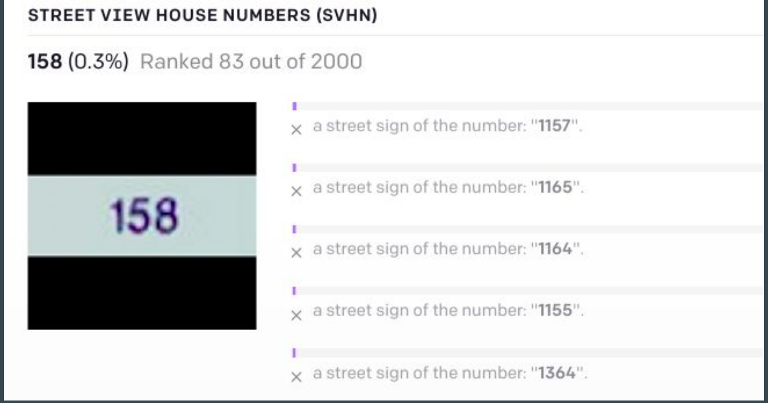
CLIP - Limitations

3. **Poor generalization** to images not covered in its pre-training dataset.

- Zero-shot CLIP only achieves 88% accuracy on MNIST dataset.

4. Zero-shot classification - **sensitive to wording or phrasing**

- Trial and error "prompt engineering"



STREET VIEW HOUSE NUMBERS (SVHN)

158 (0.3%) Ranked 83 out of 2000

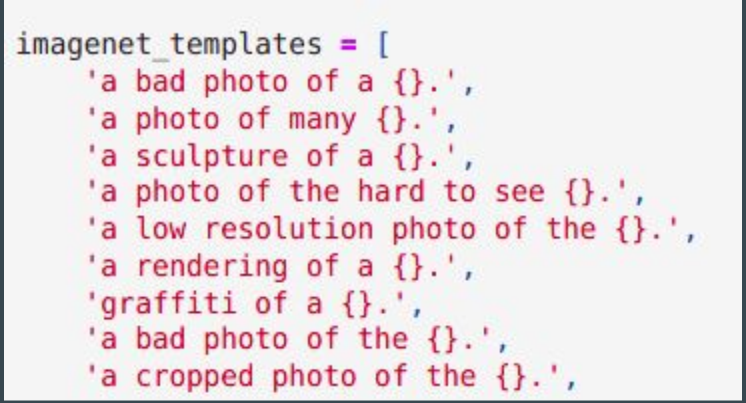
× a street sign of the number: "1157".

× a street sign of the number: "1165".

× a street sign of the number: "1164".

× a street sign of the number: "1155".

× a street sign of the number: "1364".

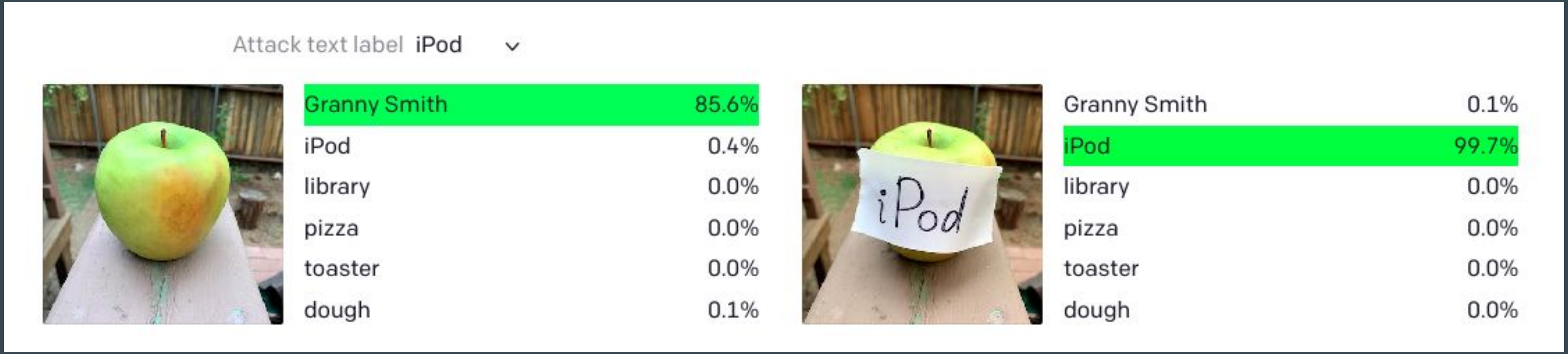


```
imagenet_templates = [
    'a bad photo of a {}.',
    'a photo of many {}.',
    'a sculpture of a {}.',
    'a photo of the hard to see {}.',
    'a low resolution photo of the {}.',
    'a rendering of a {}.',
    'graffiti of a {}.',
    'a bad photo of the {}.',
    'a cropped photo of the {}.',
```

CLIP - Limitations

5. Typographic attacks

- Multi-modal neurons - learns concepts



Attack text label iPod

Granny Smith	85.6%
iPod	0.4%
library	0.0%
pizza	0.0%
toaster	0.0%
dough	0.1%

Granny Smith	0.1%
iPod	99.7%
library	0.0%
pizza	0.0%
toaster	0.0%
dough	0.0%

https://arxiv.org/pdf/2103.10480.pdf

Conclusion

- CLIP is one of the first foundation models for images.
- Zero-shot not meant for commercial deployments.
- Not perfect but…
  - Good starting point for researchers to understand zero-shot image classification tasks.