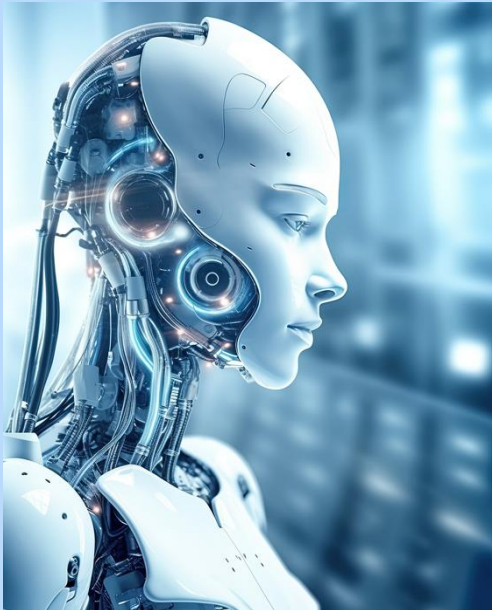


Variants of **Vision** Transformer



OpenAI **CLIP** Model

Contrastive Language-Image Pre-training

<https://arxiv.org/pdf/2103.00020>

Vahid Mirjalili (<https://vmirly.github.io>)

PyML Studio
Vision Transformers Series

CLIP Model Overview

Learning perception from supervision
contained in natural language

Natural Language as a training signal

- Predicting which caption goes with which image
- Collected a dataset of 400M image, text pairs from the internet
- Self-supervised pre-training
- Enabling zero-shot transfer to down-stream tasks

Image Representation Learning using Natural Language Supervision

Limited amount of supervised data (gold labels)
vs.
unlimited amount of raw text

**Advantages of
natural language
supervision**

1. Easier to scale

2. Connecting learned representations to language

WebImageText (WIT) Dataset

- Constructed a dataset of 400M image/text pairs
- Based on 500000 queries collected from high frequent (+100) words in English Wikipedia
- Balancing the dataset with a cap of 20000 (image, text) pair per query

An example of image-text pair



A vibrant street scene in New York City, bustling with pedestrians crossing the street and yellow cabs navigating through the traffic. The image captures the dynamic urban life amid the backdrop of towering skyscrapers under a partly cloudy sky, epitomizing the bustling energy of the city that never sleeps.

Visual Representation Learning

Existing Approaches for Image-Text Pretraining

1. VirTex

2. ConVIRT

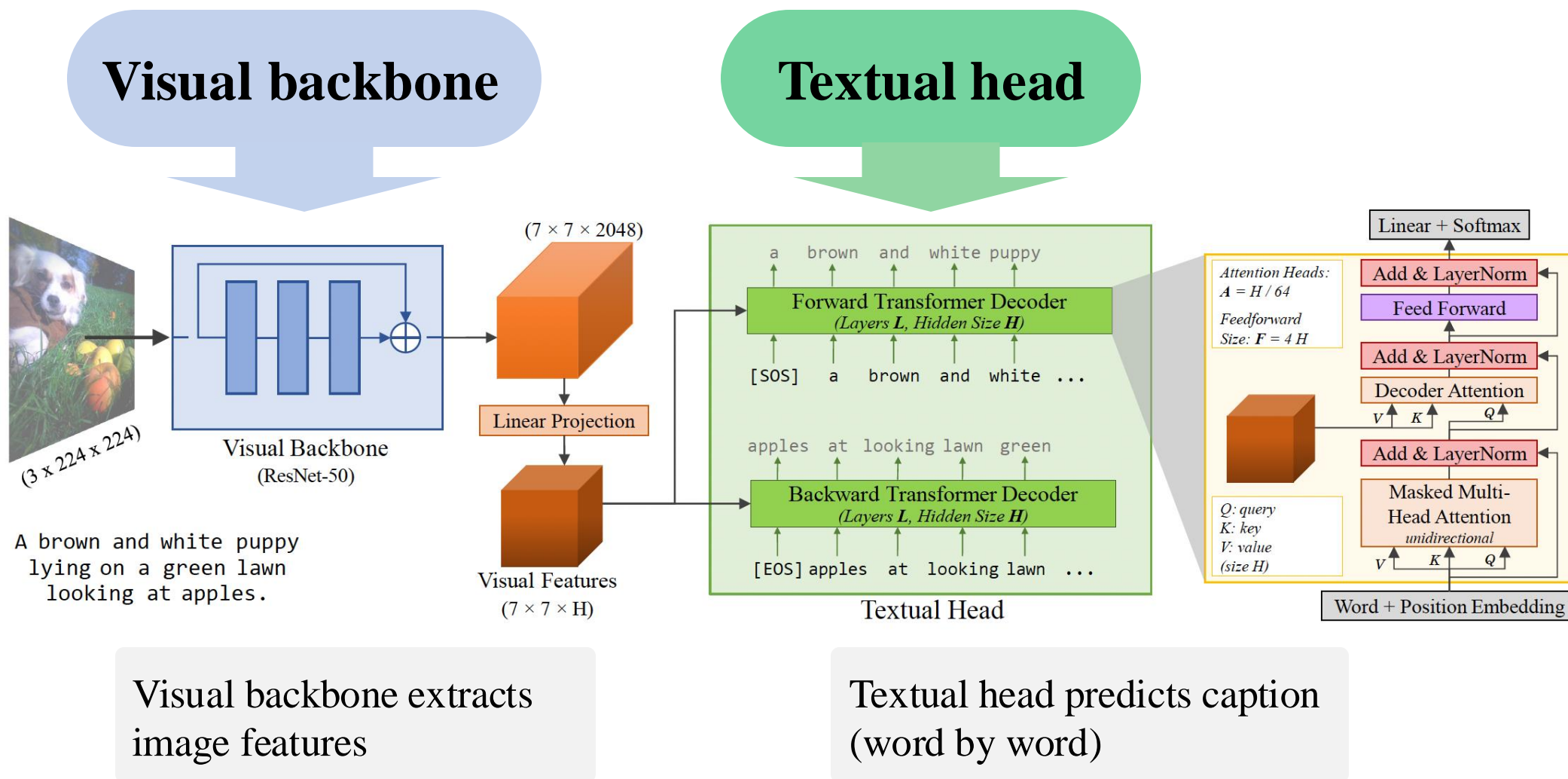
Motivation:

Supervised learning requires high-quality annotations (gold labels)

- Difficult to obtain labels in specialized domains
- High annotation cost
- Limited labeled data

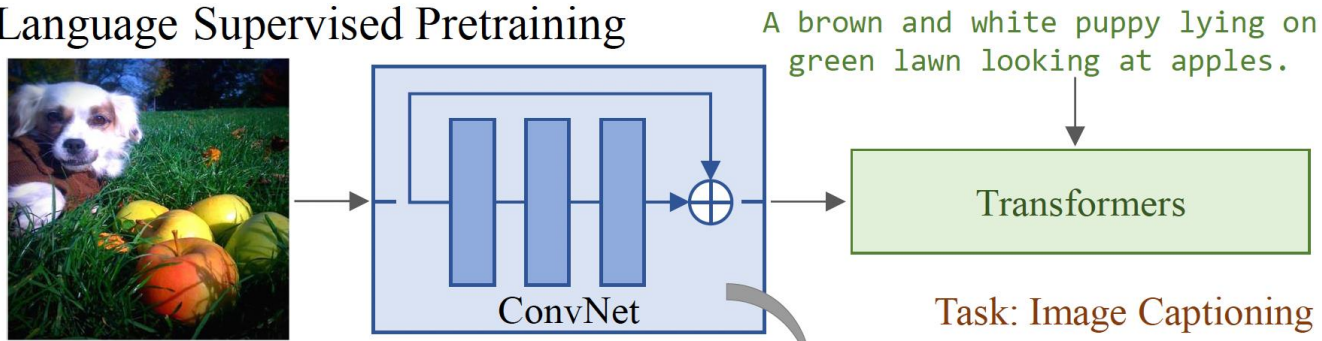
➔ Pre-training using self-supervised learning with abundant textual data

VirTex Model Overview

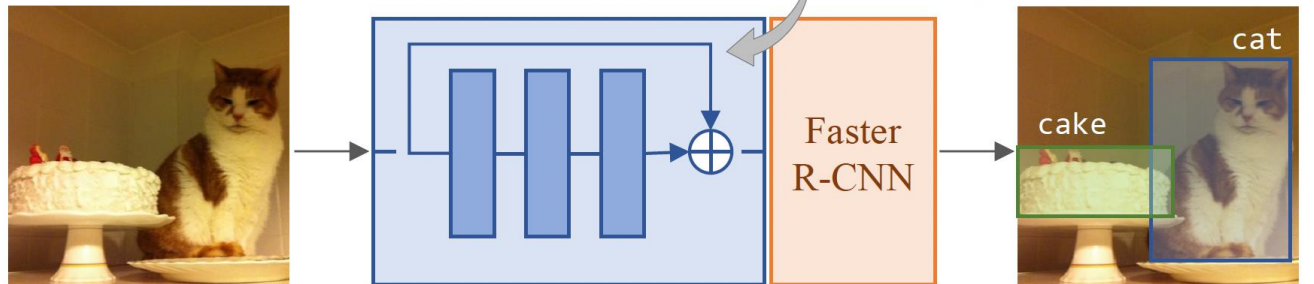


VirTex Overview

Language Supervised Pretraining



Downstream Transfer



Pre-training

Jointly train the visual backbone and the textual head on image captioning

Transfer to downstream tasks

E.g., object detection

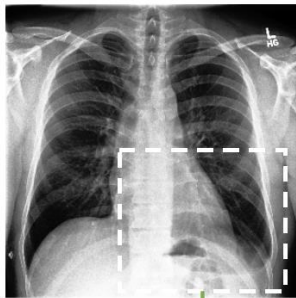
Drawbacks: Predicting exact word is difficult

ConVIRT Model Overview

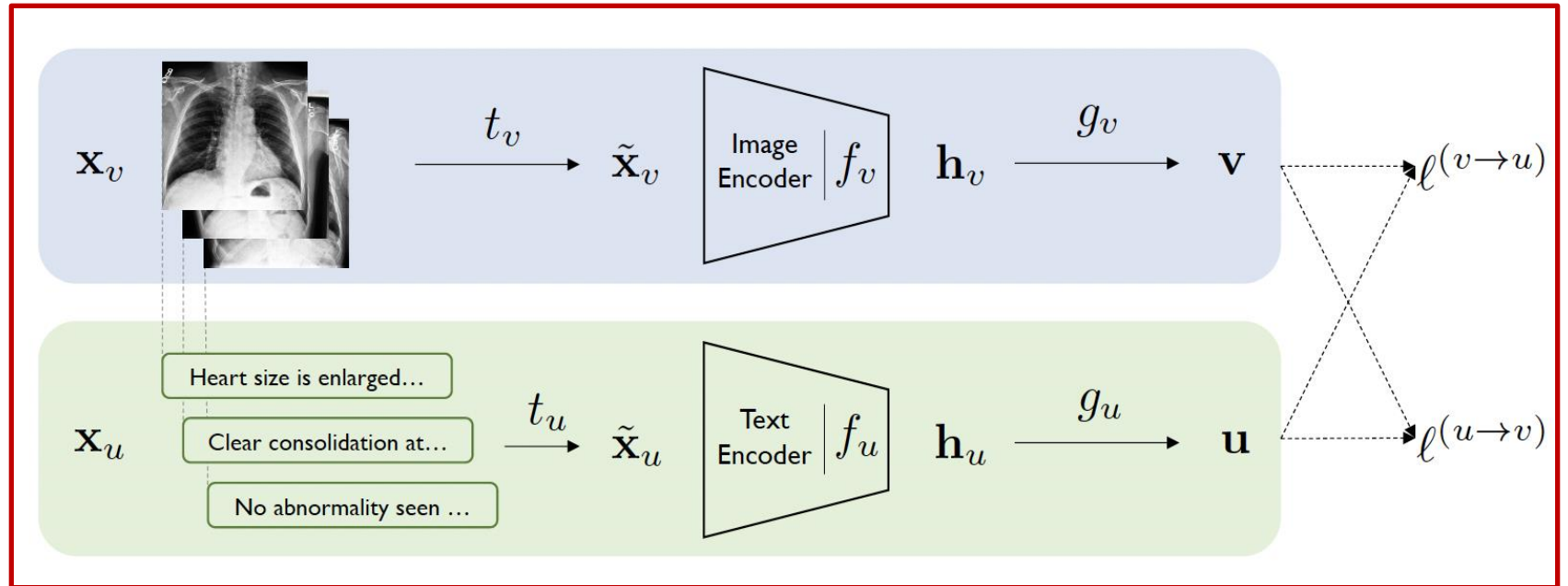
Example medical images
with sample text descriptions



Severe **cardiomegaly**
is noted in the image
with enlarged...



Radiograph shows
pleural effusion in
the right...



- Using contrastive loss to determine which image/caption pair is accurate
(instead of predicting the exact caption)

CLIP Objective Function

Battle of predictive vs. contrastive objectives

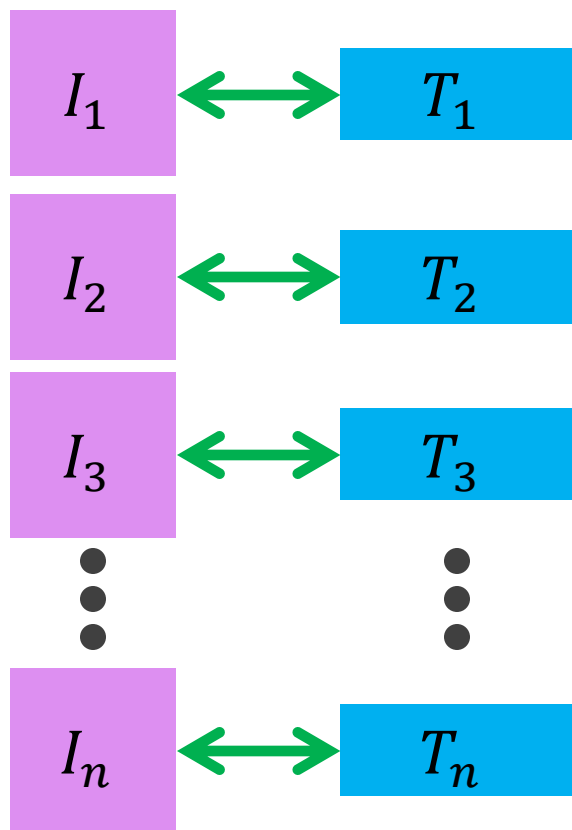
- Predicting exact words in the caption is difficult (due to wide variety of possible captions for an image)
- Contrastive learning has shown better representations

CLIP Objective: Contrastive loss on (image, text) pairs

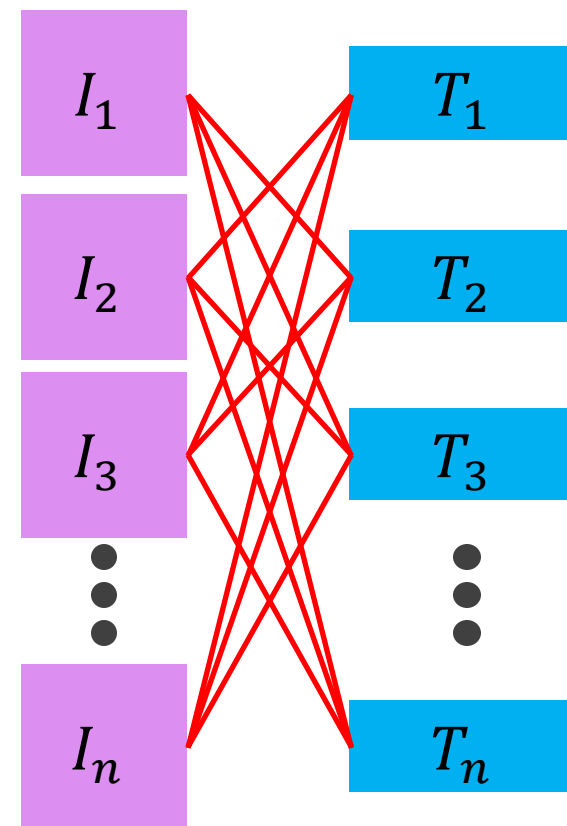
Which text (as a whole, not word-by-word) goes with which image

CLIP Image-Text Pair Matching

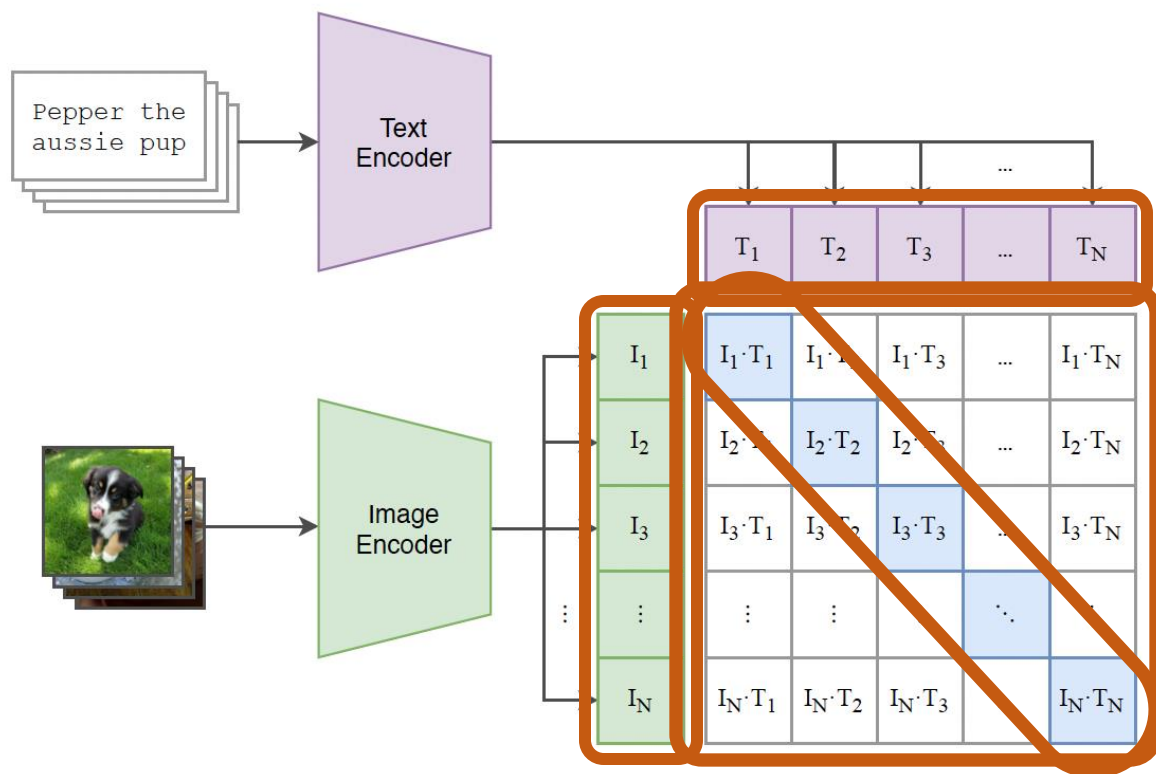
- Increase the cosine similarity of correct pairs in a batch



- Reduce the cosine similarity of $n^2 - n$ incorrect pairings



CLIP 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
```

Image Encoder in CLIP

ResNet

- Replace global average pooling with attention pooling (transformer-style QKV attention)

ResNet-50, ResNet-101
RN50-x4, RN50-x16, RN50-x64

ViT

ViT-B/32, ViT-B/16, ViT-L/14

Additional model pre-trained
at higher resolution:
ViT-L14@336px

Best model: ViT-L/14@336px

Text Encoder

GPT-2 Architecture

- 63M-parameters
- Tokenizer: Lower-cased BPE with a 49k vocab size
- Max Sequence length 76
- Features extracted from token [EOS]

Experiments

➤ Zero-shot Transfer

Generalization to unseen datasets

➤ Representation Learning

- Transfer learning:
- Fitting a linear classifier
 - Fine-tuning (not in this study)

➤ Robustness to natural distribution shift

Out of distribution testing

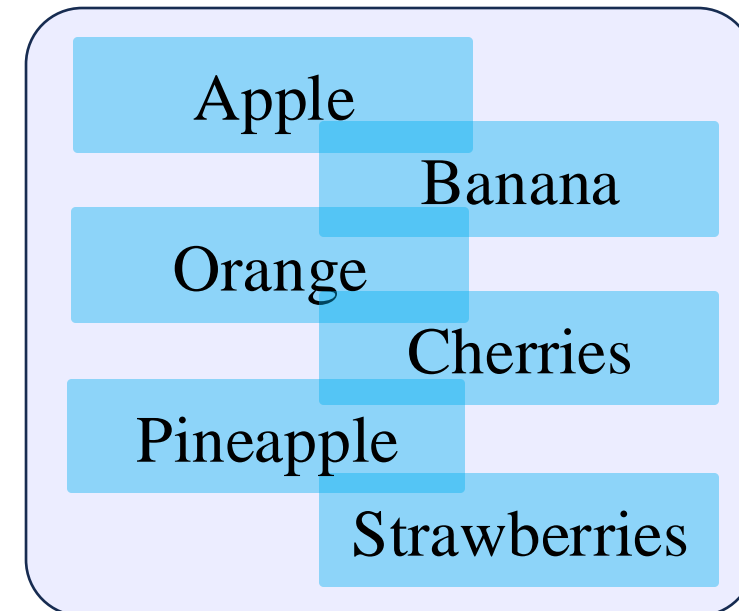
Zero-shot Transfer: Experiment Setup

- Based on the image-text pairing
- Using the class names as the set of all potential text pairings
- Predict the most probable text for each image

Images



Class names



Zero-shot Transfer: Experiment Setup

- Prompt engineering to improve the performance

On a fruit dataset:

A photo of `[label]', which is a type of fruit.

On a food dataset:

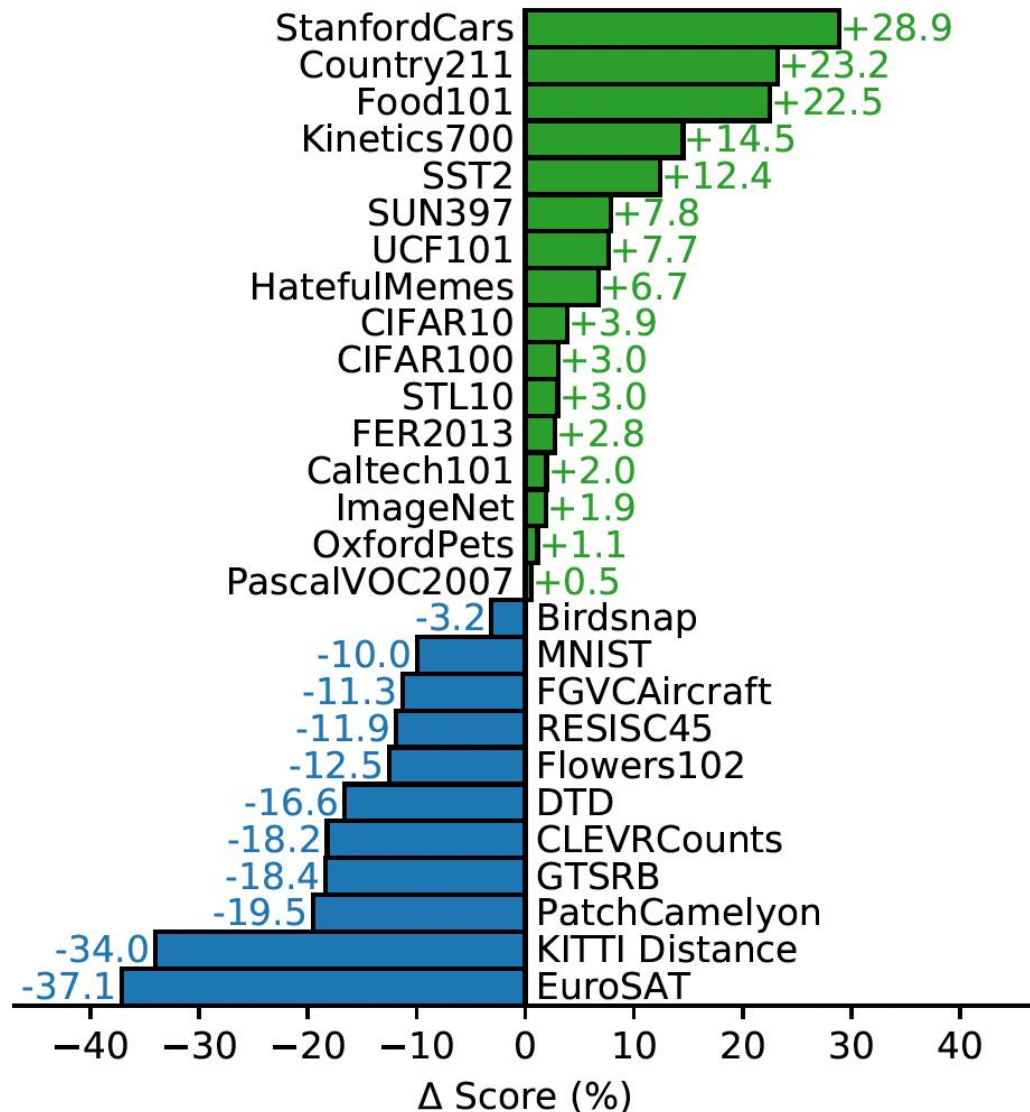
A photo of `[label]', which is a type of food.

- Ensembling

A photo of a big `[apple]', which is a type of fruit.

A photo of a small `[apple]', which is a type of fruit.

Results of Zero-shot Transfer



Zero-Shot CLIP vs. Linear Probe on ResNet50

Baseline

- A logistic regression on the features of ResNet-50 (fully supervised)

Zero-shot CLIP

- Outperforms baseline by +20% on StanfordCars, Country211, Food101
- Lower performance on specialized and complex datasets such as satellite image datasets, lymph node tumor detection, ...

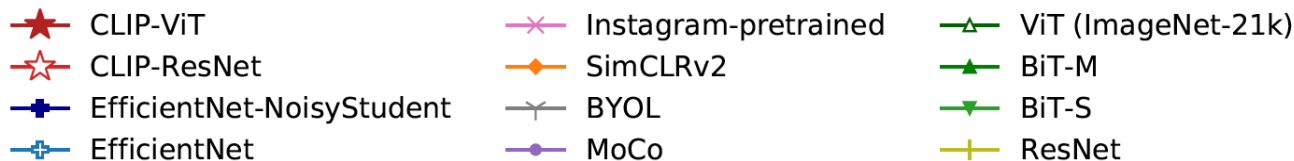
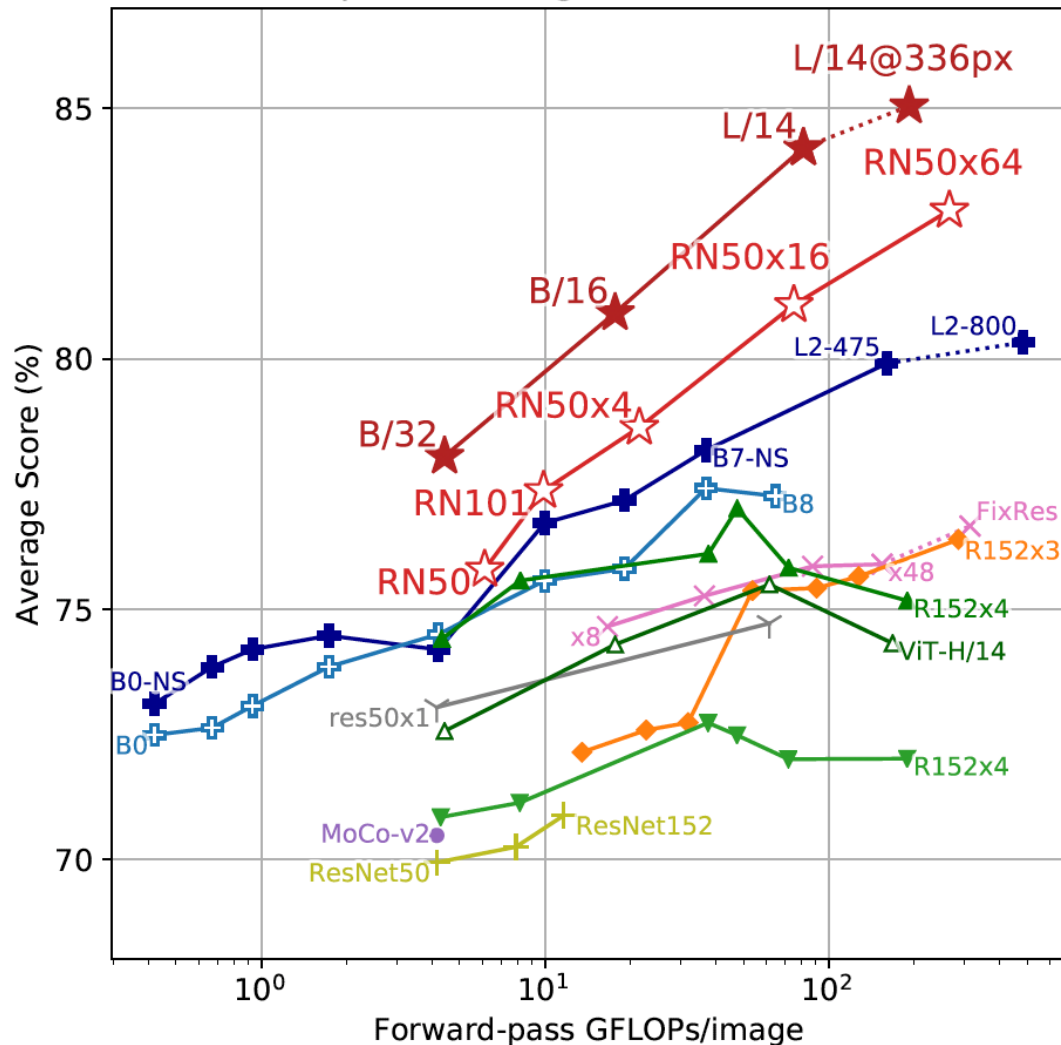
Representation Learning

- Fitting a linear classifier on representations extracted from the model







CLIP Representations:

- CLIP features outperform the best ImageNet model

Linear probe average over all 27 datasets



Robustness to Natural Distribution Shift

| | 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% |

CLIP Representations:

- Zero-shot CLIP significantly outperforms ResNet-101
- ➔ Zero-shot CLIP is more robust to natural distribution shift

Natural distribution shift for the class of bananas

CLIP: Contrastive Language-Image Pre-training

Thanks for watching