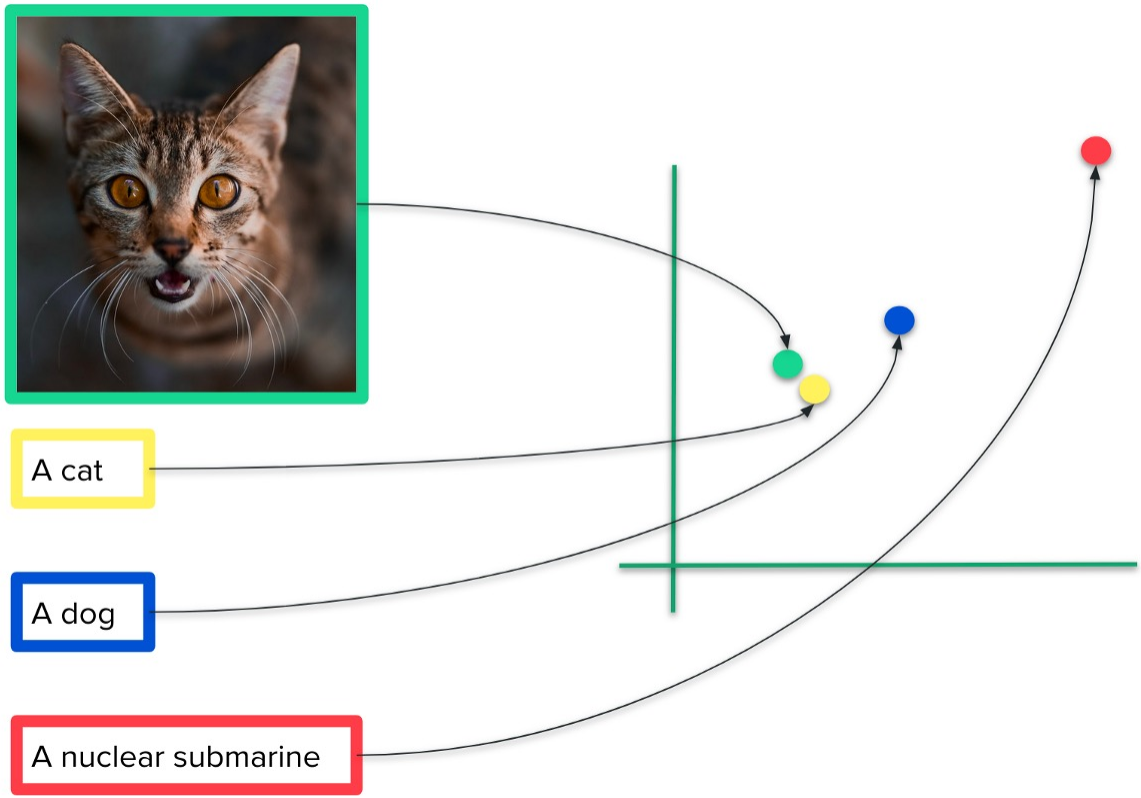
# Session #17:
# Self-Supervised Vision-Lang Models

Tuesday, October 25
CSCI 601.771: Self-supervised Statistical Models

**Goal:**

A joint representation
for vision and language


A cat
A dog
A nuclear submarine

# Learning Transferable Visual Models From Natural Language Supervision

Haoyue Guan and Karan Thakkar

# Motivation

This is a dog
This is not a bird
This is not a squirrel
This is not a horse

This is not a dog
This is a bird
This is not a squirrel
This is not a horse

This is not a dog
This is not a bird
This is a squirrel
This is not a horse

This is not a dog
This is not a bird
This is not a squirrel
This is a horse

MODEL TRAINING

Learn that this examples is a dog but not bird, squirrel and a horse through language

MODEL TESTING

- Like MCQ Questions
- Chose what is correct

A. This is a dog
B. This is a bird
C. This is a squirrel
D. This is a horse

✍: Names

# Introduction

- Vision Models: Restricted!
  - Most state-of-the-art computer vision systems are trained to predict a fixed set of predetermined object categories. *Limit "zero-shot" capabilities
  - This restricts the generality and plagues the model performance when facing unseen data/visual concepts. *Curtail flexibility

Solution: Inspired from the method from NLP

- Autoregressive modeling
- Masked language modeling
- Doesn't require output heads or dataset customization

Incorporate text information with images!

Method: Given a specific caption, find its corresponding matching images.

- Novelty: Using prompt template, the model doesn't have classification head

# Model

## Astonishing Performance

- Same accuracy as ResNet-50 on ImageNet zero-shot without needing to use any of the 1.28 million training examples.
- Easily transfer over 30 different existing computer vision datasets, spanning wide range of vision tasks. (OCR, action recognition, fine-grained object classification, etc.)
- Dataset: 400 million (image,text) pairs

**Without using any dataset specific training!**

# Model

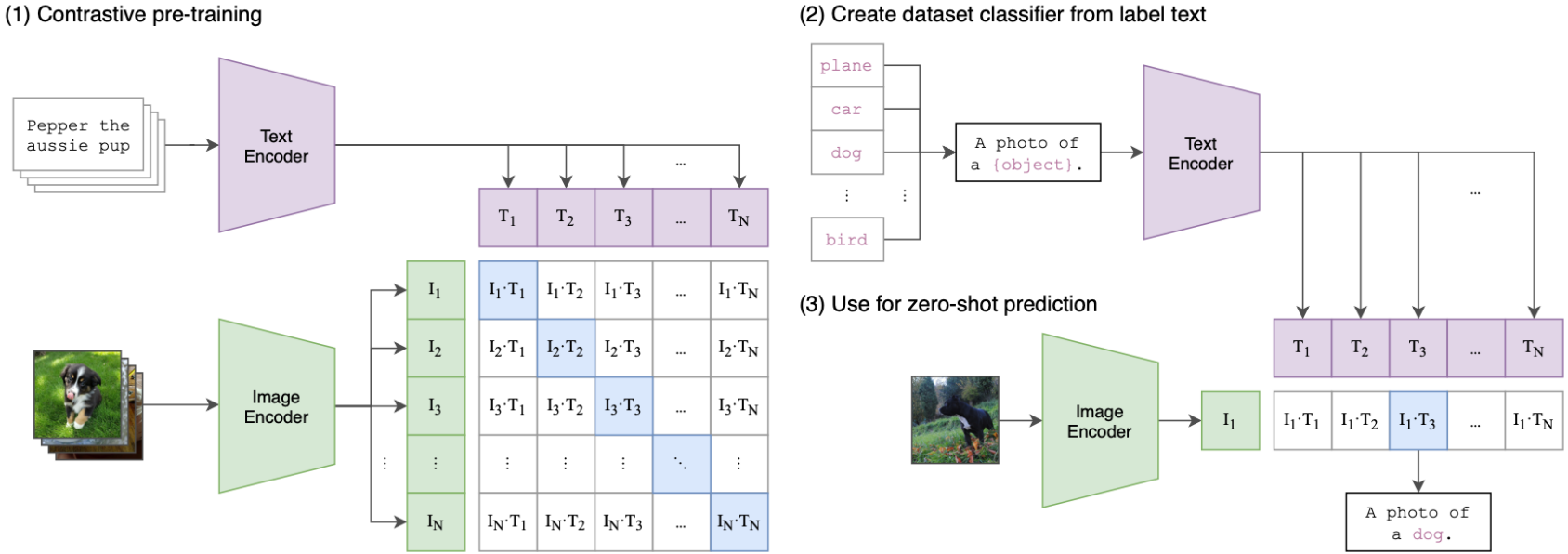
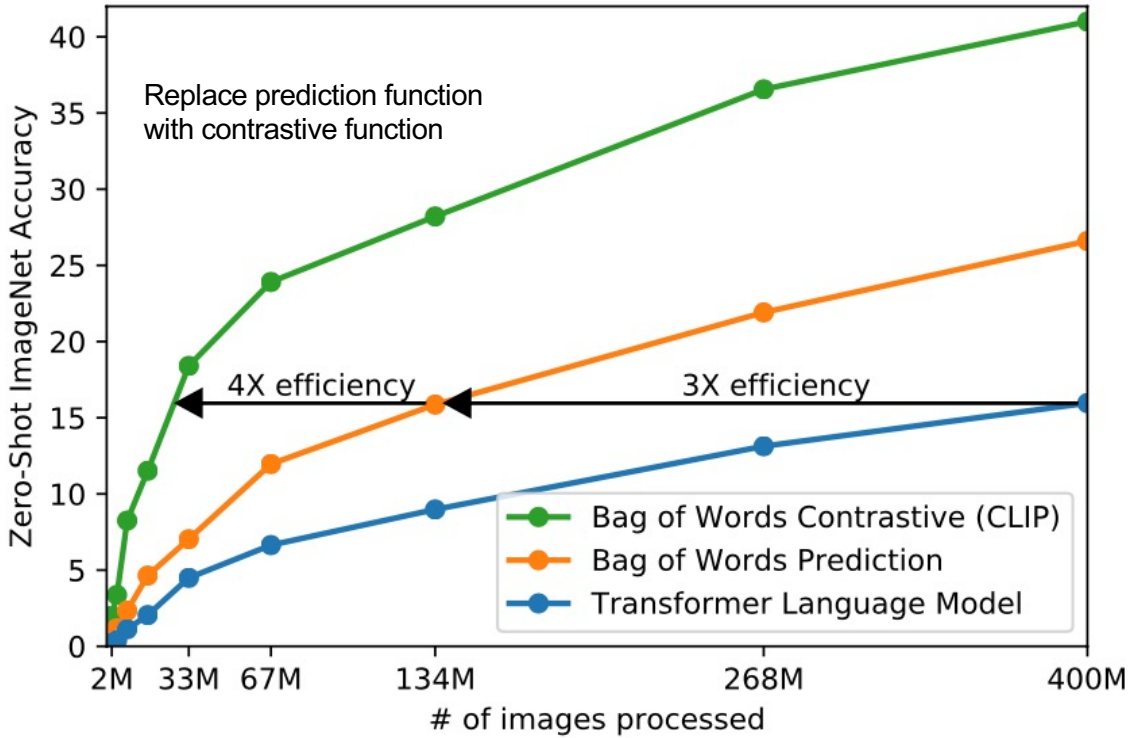



*Figure 1.* Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

✍: Names

# Model




Difficult to train due to variety of description/comments

Contrastive learning alleviate the difficulty of supervision

CLIP is much more efficient at zero shot transfer

✍: Names

# Model

Simple implementation

- No difference with previous contrastive learning work
- N real pairs, N^2-N incorrect pairs
- Symmetric cross entropy loss to minimize similarity scores

Large Dataset

- Avoid overfitting
- No need for pretraining image encoder and text encoder

```
# 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.

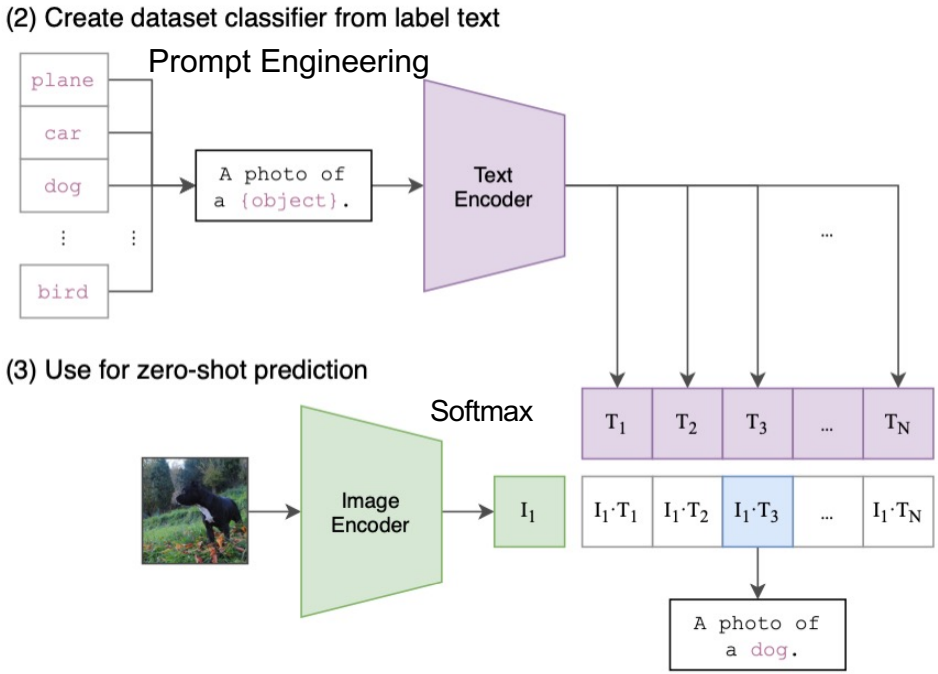
✍: Names

# Experiment

## Why Prompt Engineering?

- Polysemy
- Same words might have different meaning in same dataset!
  - ImageNet: construction crane vs carne
  - Oxford-IIIT Pet: Boxer (pet vs athlete)

  - Use 80 templates in CLIP



```python
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 {}.',
    'a tattoo of a {}.',
    'the embroidered {}.',
    'a photo of a hard to see {}.',
    'a bright photo of a {}.',
    'a photo of a clean {}.',
    'a photo of a dirty {}.',
    'a dark photo of the {}.',
    'a drawing of a {}.',
    'a photo of my {}.',
    'the plastic {}.',
    'a photo of the cool {}.',
    'a close-up photo of a {}.',
    'a black and white photo of the {}.',
    'a painting of the {}.',
    'a painting of a {}.',
    'a pixelated photo of the {}.',
    'a sculpture of the {}.',
    'a bright photo of the {}.',
    'a cropped photo of a {}.',
    'a plastic {}.',
    'a photo of the dirty {}.',
```

✍: Names

# Experiment

Better performance on object classification

- If object exists in the image, the corresponding text should contains it.

Limited on Texture classification and Object counting

- No informative label
- Few-shot might be more appropriate for complex tasks

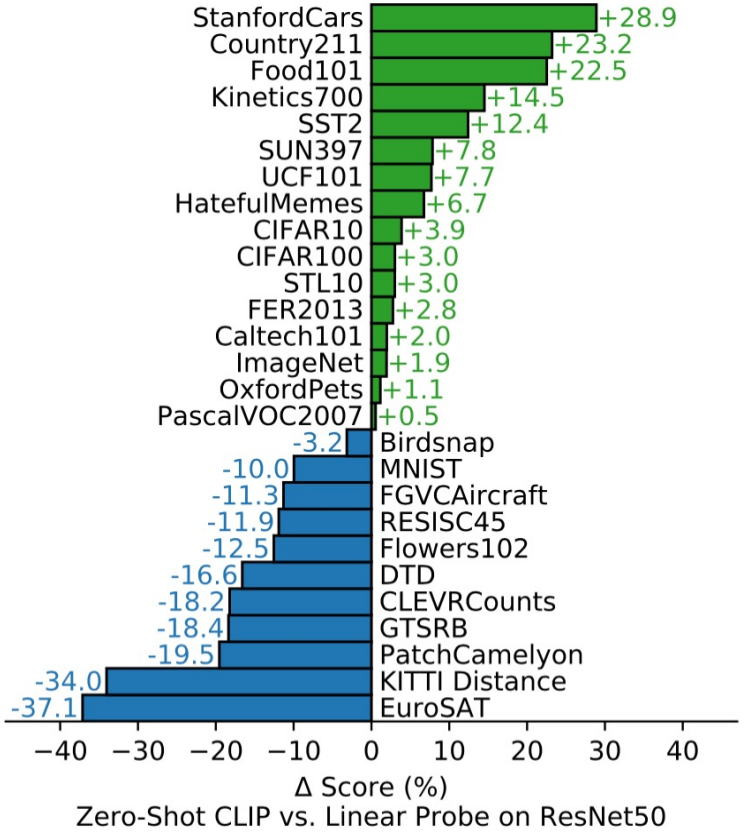



*Figure 5.* **Zero-shot CLIP is competitive with a fully supervised baseline.** Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.

✍: Names

# Experiment

- BiT-M: designed for transfer learning by Google, one of the best model in few-shot transfer learning (Strong Baseline)

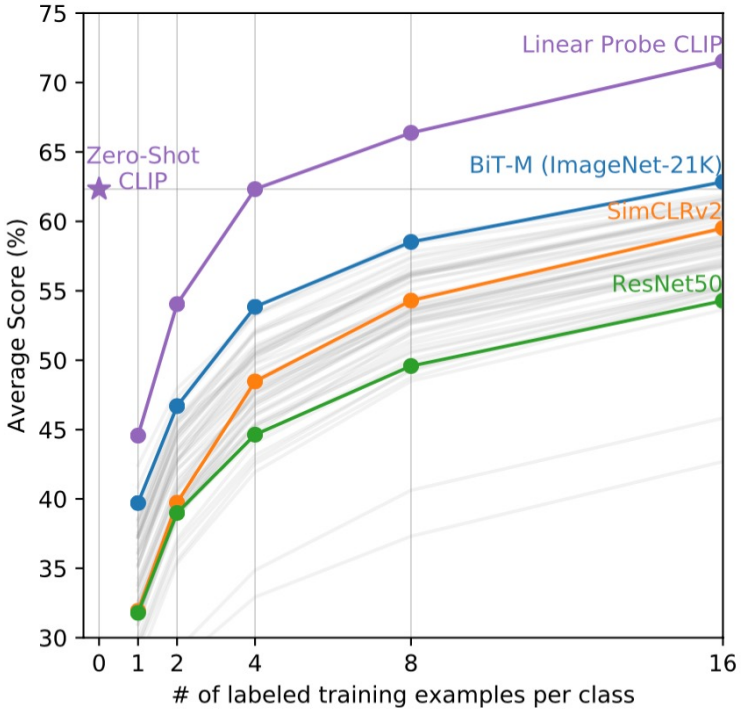
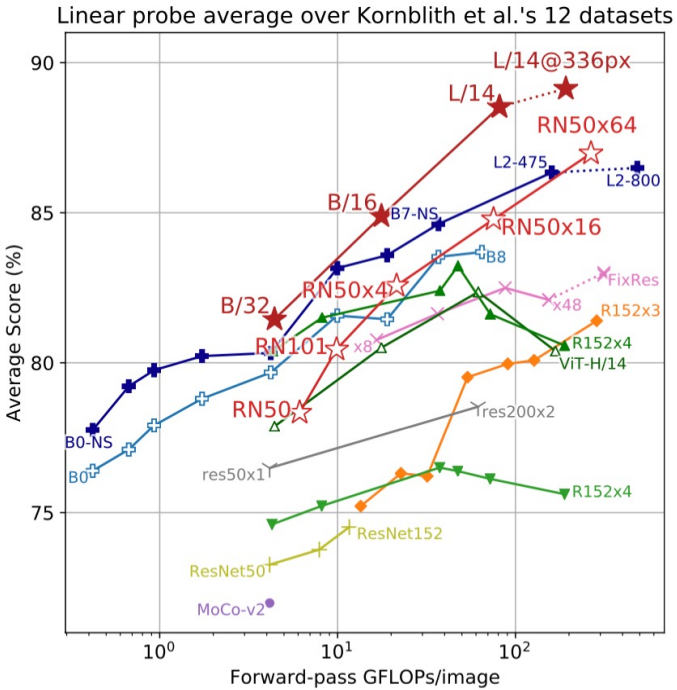
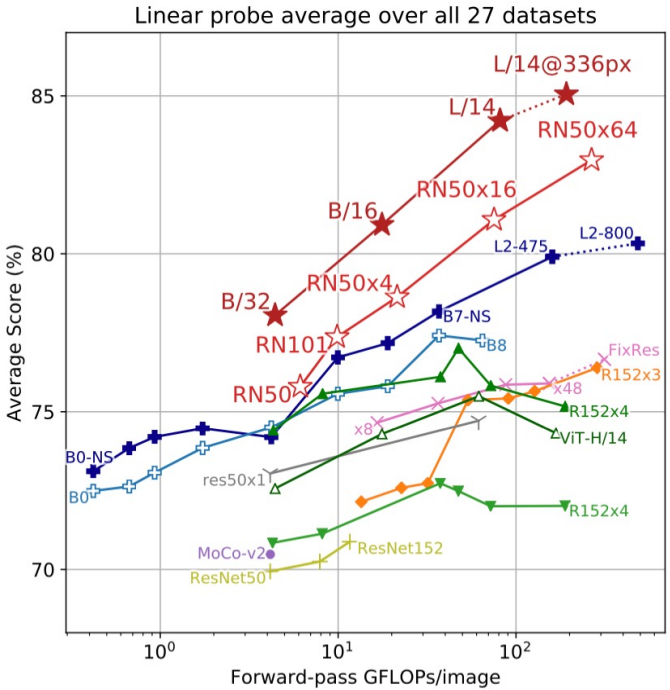



*Figure 6.* **Zero-shot CLIP outperforms few-shot linear probes.** Zero-shot CLIP matches the average performance of a 4-shot linear classifier trained on the same feature space and nearly matches the best results of a 16-shot linear classifier across publicly available models. For both BiT-M and SimCLRv2, the best performing model is highlighted. Light gray lines are other models in the eval suite. The 20 datasets with at least 16 examples per class were used in this analysis.

✍: Names




Supervised EfficientNet

- CLIP-ViT
- CLIP-ResNet
- EfficientNet-NoisyStudent
- EfficientNet

Contrastive learning

- Instagram-pretrained
- SimCLRv2
- BYOL
- MoCo

Classic Supervised model

- ViT (ImageNet-21k)
- BiT-M
- BiT-S
- ResNet

✍: Names

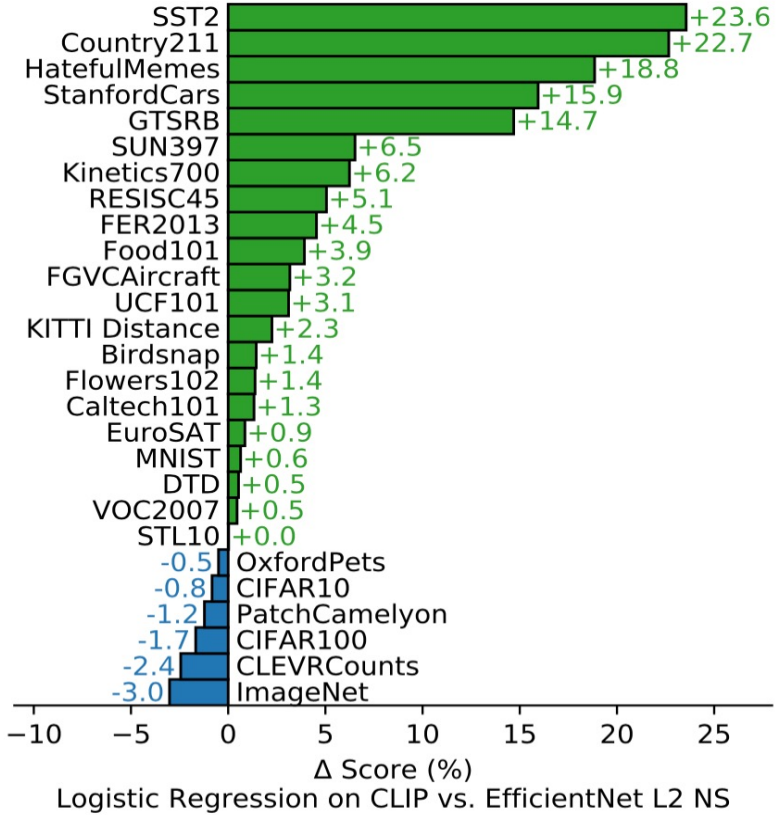



*Figure 11.* **CLIP's features outperform the features of the best ImageNet model on a wide variety of datasets.** Fitting a linear classifier on CLIP's features outperforms using the Noisy Student EfficientNet-L2 on 21 out of 27 datasets.

# Robustness to Natural Distribution Shift

ImageNet Models Overfit

Whereas CLIP is more robust!

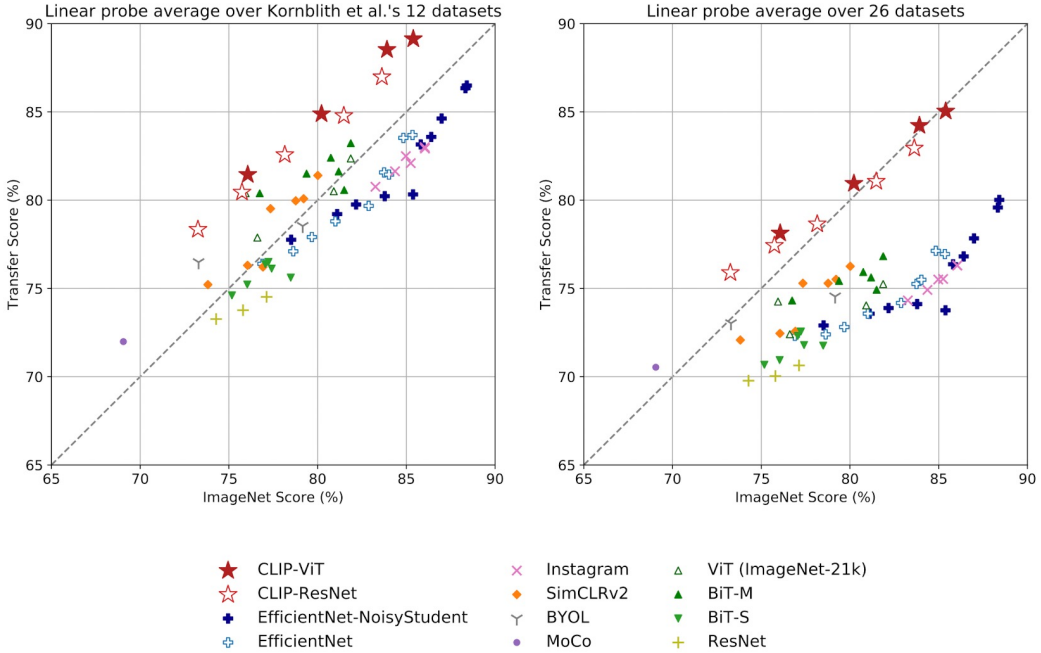



*Figure 12.* **CLIP's features are more robust to task shift when compared to models pre-trained on ImageNet.** For both dataset splits, the transfer scores of linear probes trained on the representations of CLIP models are higher than other models with similar ImageNet performance. This suggests that the representations of models trained on ImageNet are somewhat overfit to their task.

Makes Sense Intuitively cause Zero Shot Models are not trained to work better on a specific task.

✍: Names

# Robustness to Natural Distribution Shift

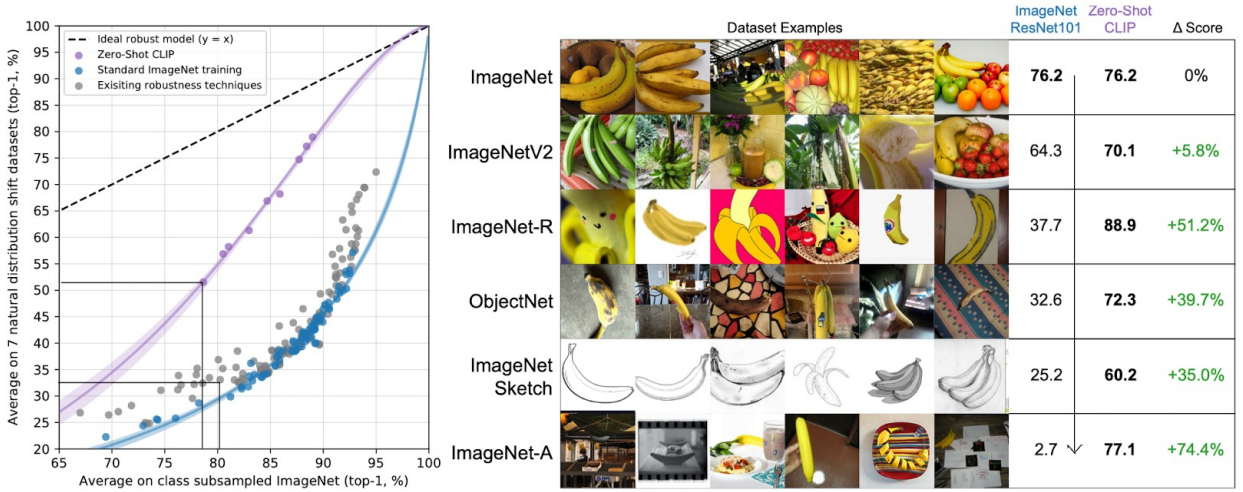
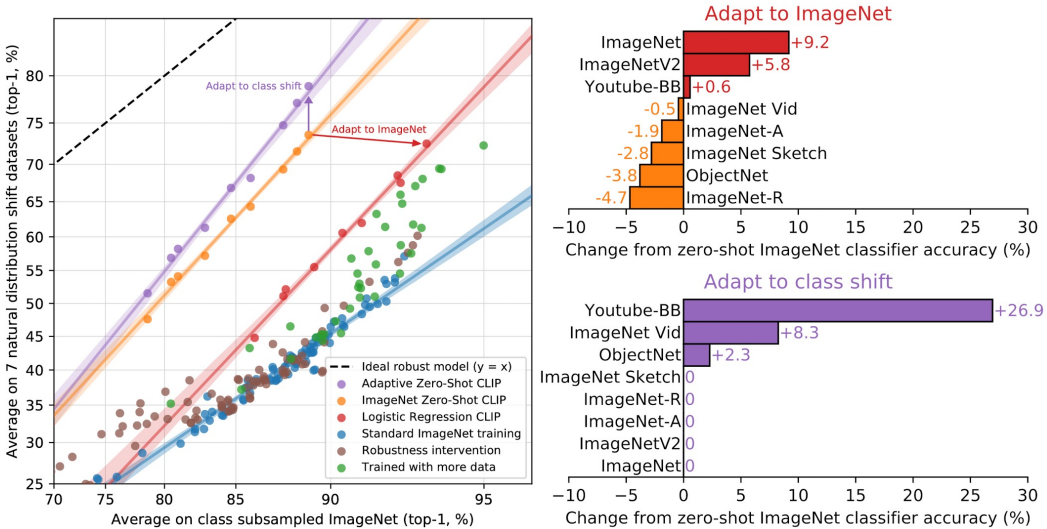



*Figure 13.* **Zero-shot CLIP is much more robust to distribution shift than standard ImageNet models.** (Left) An ideal robust model (dashed line) performs equally well on the ImageNet distribution and on other natural image distributions. Zero-shot CLIP models shrink this "robustness gap" by up to 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.

# Robustness to Natural Distribution Shift

How is it possible to improve accuracy by 9.2% on the ImageNet dataset with little to no increase in accuracy under distribution shift? Is the gain primarily from "exploiting spurious correlations"? Is this behavior unique to some combination of CLIP, the ImageNet datatset, and the distribution shifts studied, or a more general phenomena? Does it hold for end-to-end finetuning as well as linear classifiers? We do not have confident answers to these questions at this time.



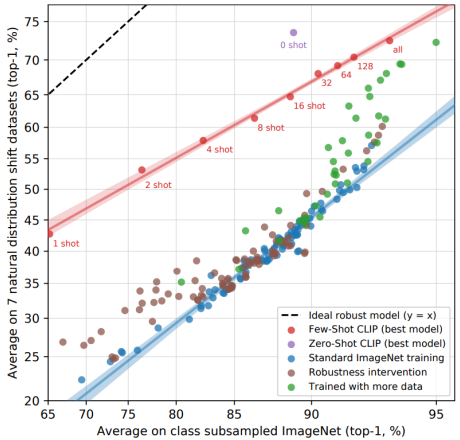




*Figure 15.* **Few-shot CLIP also increases effective robustness compared to existing ImageNet models but is less robust than zero-shot CLIP.** Minimizing the amount of ImageNet training data used for adaption increases effective robustness at the cost of decreasing relative robustness. 16-shot logistic regression CLIP matches zero-shot CLIP on ImageNet, as previously reported in Figure 7, but is less robust.

✍: Names

# CLIP Vs Human

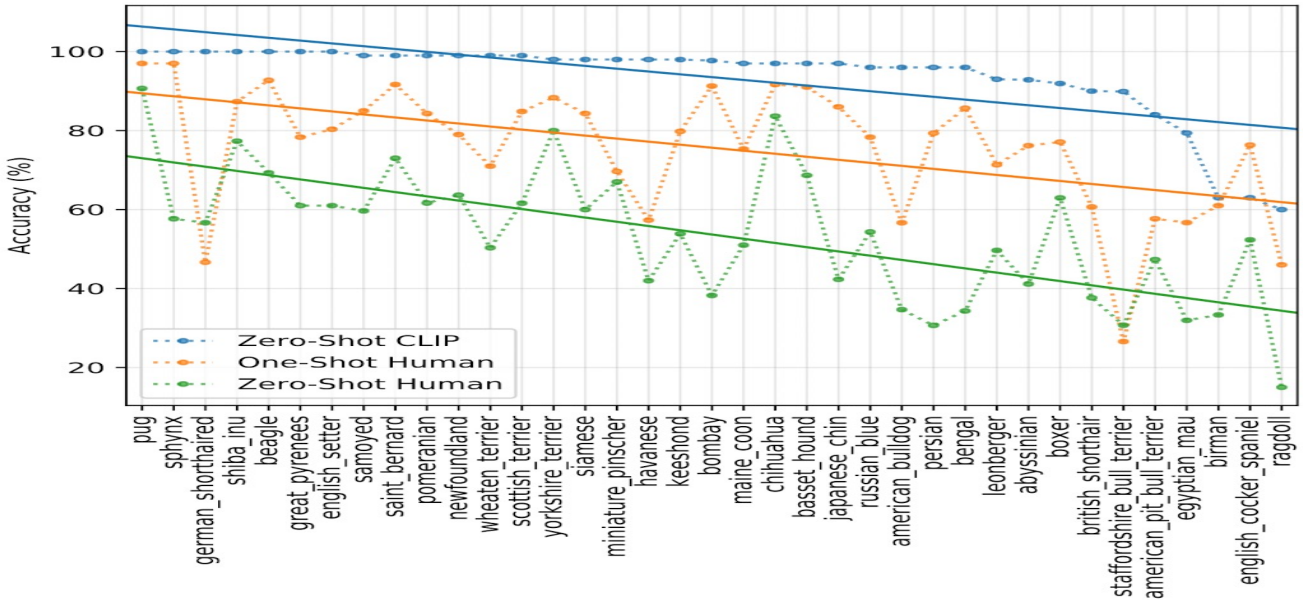



*Figure 16.* The hardest problems for CLIP also tend to be the hardest problems for humans. Here we rank image categories by difficulty for CLIP as measured as probability of the correct label.

✍: Names

# Bias

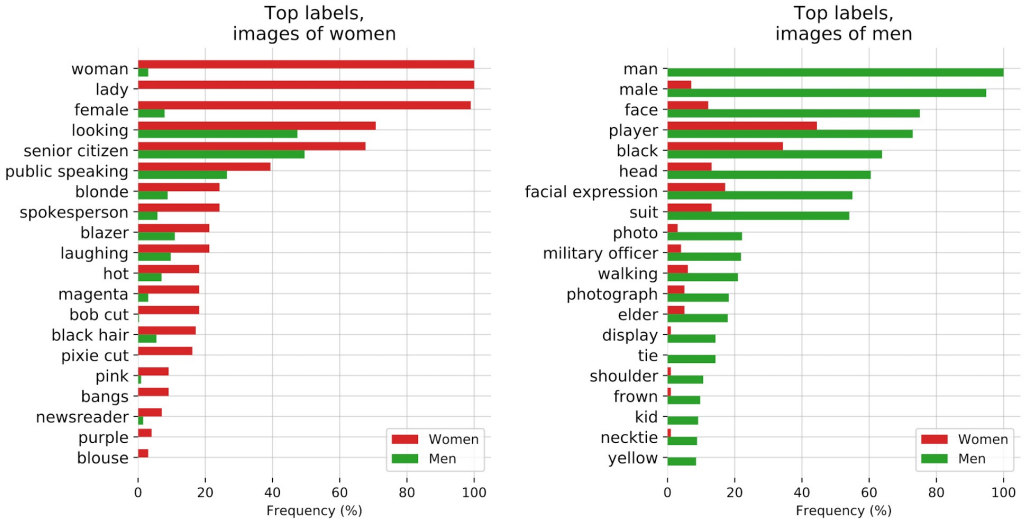



*Figure 18.* CLIP performance on Member of Congress images when given the combined returned label set for the images from Google Cloud Vision, Amazon Rekognition and Microsoft Azure Computer Vision. The 20 most gendered labels for men and women were identified with $\chi^2$ tests with the threshold at 0.5%. Labels are sorted by absolute frequencies. Bars denote the percentage of images for a certain label by gender.

# Surveillance

| Model | 100 Classes | 1k Classes | 2k Classes |
|---|---|---|---|
| CLIP L/14 | 59.2 | 43.3 | 42.2 |
| CLIP RN50x64 | 56.4 | 39.5 | 38.4 |
| CLIP RN50x16 | 52.7 | 37.4 | 36.3 |
| CLIP RN50x4 | 52.8 | 38.1 | 37.3 |

*Table 8.* CelebA Zero-Shot Top-1 Identity Recognition Accuracy

# Empiricists

https://colab.research.google.com/drive/17PUGjMAueIKXoiopvahoQ4LyFsBBn7B7?usp=sharing

# Strengths

- Personally think this is a **impactful** paper
  - Tackle an important problem in zero-shot learning (no need retraining, labeling new dataset)
- Very well written paper
- Provided detailed experiment settings
- Easy access to additional resources: slides, source-code, talk after publication

# Strengths

**Broader Impact of paper:**

- CLIP allows for easy classification for categorization
- Significant promise of varied task handling such as Image search and retrieval
- Addresses Bias
- Addresses acknowledgement of data gathering source, participants and researchers involved

🔎: Ha, Elisée

# Strengths

CLIP significantly exceeds the performance of conventional 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).

Source: https://arxiv.org/abs/2103.00020

🔎: Ha, Elisée

# Weaknesses

- Often require prompt engineering
- Detail paper, but also too many words in arXiv version
- Lack of notation clarification
- Personally expect more theoretical analysis for a ICML paper
  - Contrastive Learning
  - Uncertainty under Distributional-Shift
- Personally expect more investigation on distributional shift experiment

🔎: Ha, Elisée

# Weaknesses - notation

- Misunderstand the number of probability spaces, definition clarification:
  - **N need to be the same as test and train?**

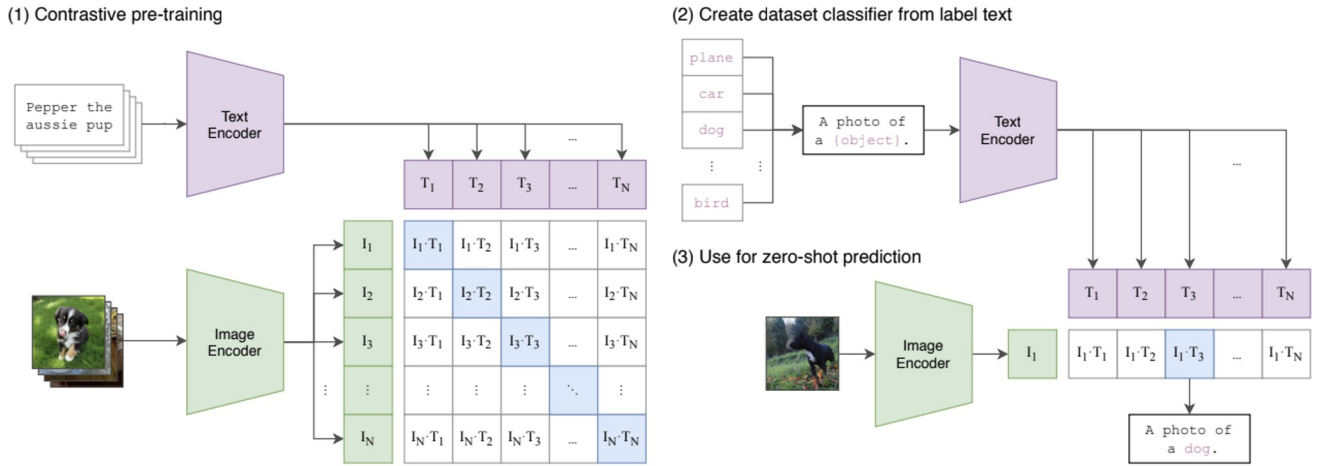
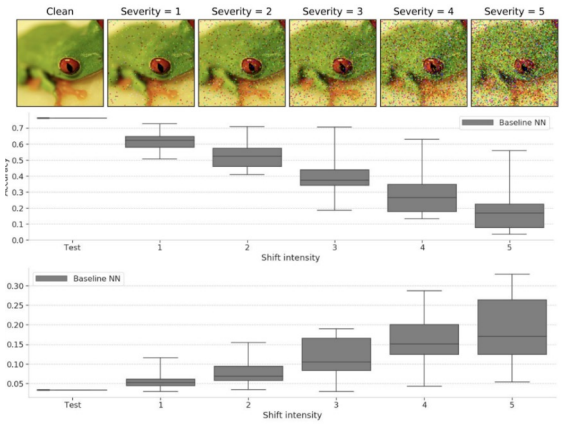
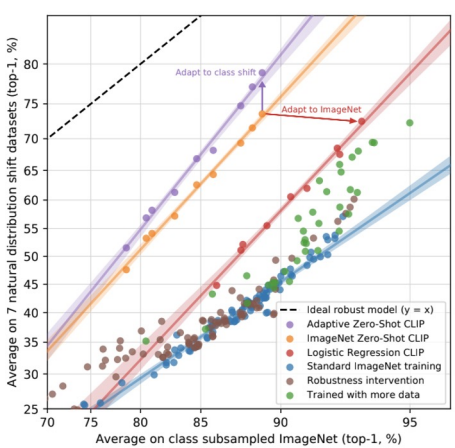


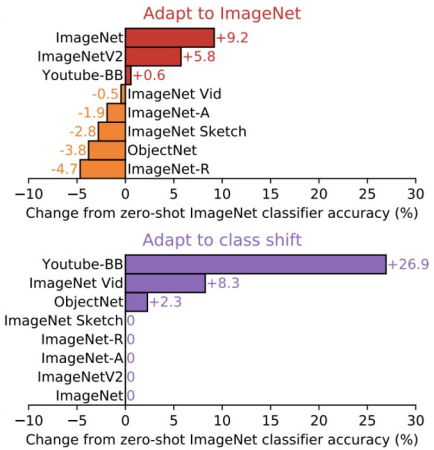

*Figure 1.* Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

# Weaknesses - distributional shift experiment

- Distributional-shift experiment:
  - Expect shift intensity by ImageNet-C
  - Adapt to ImageNet: There is still unsolved question like "How is it possible to improve accuracy by 9.2% on the ImageNet dataset with little to no increase in accuracy under distribution shift?"
  - Adapt to class shift is look obvious



# Conclusion

Personally think:

- An impactful paper: very well written and experiments
- Expect more theoretical analysis and notation elaboration

=> Strongly accept.

🔎: Ha, Elisée

# Problem

- Crafting a good data set for semi supervised learning is hard

- Clips data set picks from captioned images online

- This obviously creates the problem of OOD Data !

# Noisy teacher student model (Xie 2020)

1) Train a teacher model on labeled images

2) Use the teacher to generate pseudo labels on unlabeled images

3) Train a student model on the combination of labeled images and pseudo labeled images

# Reviewer

Google slide:
https://docs.google.com/presentation/d/1Xacs-2FQ_fM9D-XxQoJjA_EMeK5uQKKT/edit?usp=sharing&ouid=106673766508452900823&rtpof=true&sd=true