Learning Transferable Visual Models From Natural Language Supervision (CLIP)

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

Presented by: Kien Tran, Pratik Ramesh, Suyash Kumar

Georgia Tech

Agenda

1. Problem Statement
2. Related works
3. Approach
4. Experiment
5. Bias & Fairness
6. Strengths and Weaknesses

Problem statement - Previous discussions

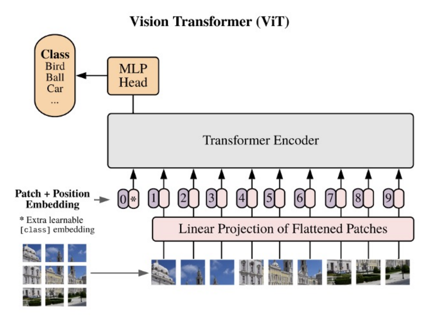
Pre-training methods for image representation:

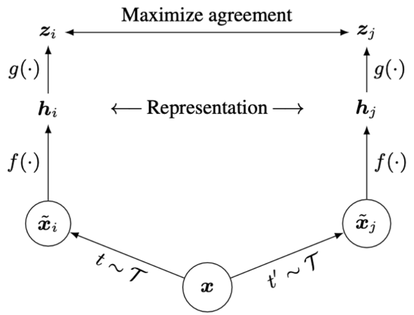
ViT: Supervised by **classification task**

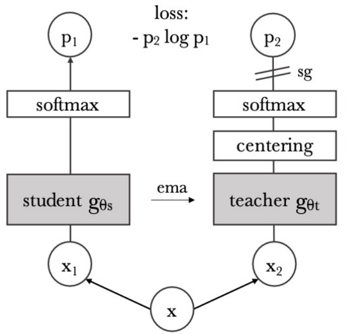
SimCLR: Supervised by **augmented views**

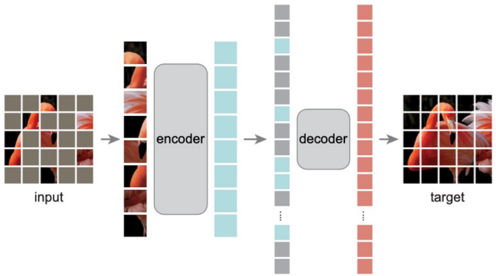
DINO: Supervised by **a teacher net**

MAE: Supervised by **masked patches**









Limitations:

- *Classification: Label collection cost and Limited capacity of expression*
- *Single modality Self-supervise: Semantically sparse supervision and no additional information*

Problem statement - CLIP

Supervise image representation model with **Natural Language**

Screen shot from DALLE 2 website - OpenAI

Advantages:

- *Scalable & Cost effective data collection*
- *Unlimited capacity of expression*
- *Semantically dense supervision*
- *Generalization and Zero-shot learning capability*

Related works - Bag of words approach

Joulin et al. 2016: Bag of words + Multi-class logistic loss

$$\ell(\theta, \mathbf{W}; \mathcal{D}) = \frac{-1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} y_{nk} \log \left[\frac{\exp(\mathbf{w}_k^\top f(\mathbf{x}_n; \theta))}{\sum_{k'=1}^{K} \exp(\mathbf{w}_{k'}^\top f(\mathbf{x}_n; \theta))} \right].$$

Li et al. 2017: Extract n-grams + Smoothed n-grams loss

$$\ell(\mathbf{I}, w; \theta, \mathbf{E}) = -\sum_{i=1}^{K} \log p\left(w_i^i | w_{i-n+1}^{i-1}, \phi(\mathbf{I}; \theta); \mathbf{E}\right)$$

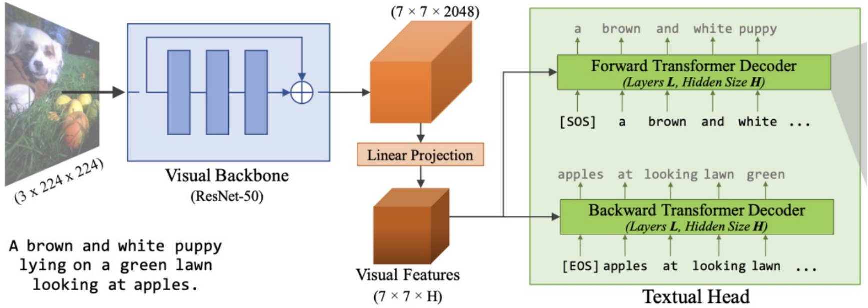
Limitations:

- *Ambiguity (Synonyms and Polysemy)*
- *Mainly model concepts, not semantic relationships*
- *Classification task not suitable for zero-shot transfer*

Related works - VirTex

Desai & Johnson, 2020 - Supervised by an autoregressive decoder:

Visual backbone + Autoregressive decoders (Textual head) + Token-wise NLL losses



$$\mathcal{L}(\theta, \phi) = \sum_{t=1}^{T+1} \log \Big(p(c_t \mid c_{0:t-1}, I; \phi_f, \theta) \Big) + \sum_{t=0}^{T} \log \Big(p(c_t \mid c_{t+1:T+1}, I; \phi_b, \theta) \Big)$$

Limitations:

- *Difficult training task due to arbitrary captions*
- *Large decoder => Computation cost*
- *Small training datasets*

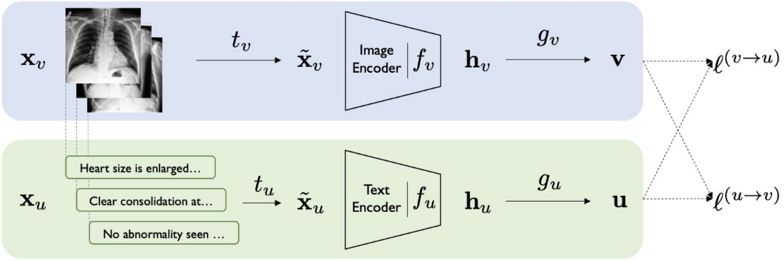
An image of a dog and a human

Both are valid?!

An image of the K-9 training activity

Related works - ConVIRT

Zhang et al., 2020 - Contrastive learning: Image encoder + Image decoder + Contrastive loss



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

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

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left(\lambda \ell_i^{(v\rightarrow u)} + (1-\lambda) \ell_i^{(u\rightarrow v)} \right),$$

Advantages:

- *Light-weight model, easier task compared to VirTex*

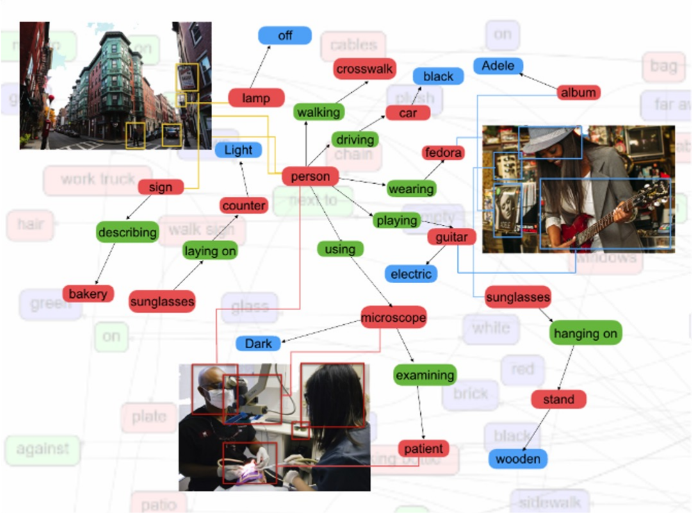
Limitations:

- *Small, domain-specific training datasets*

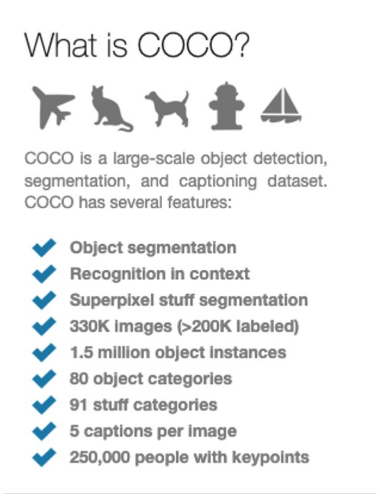
Approach - Dataset

Existing works -

- Coco & Visual Genome - 100,000 images scale
- YFCC100M - 100M scale
    - sparse metadata
    - metadata quality inconsistent



VIsualGenome



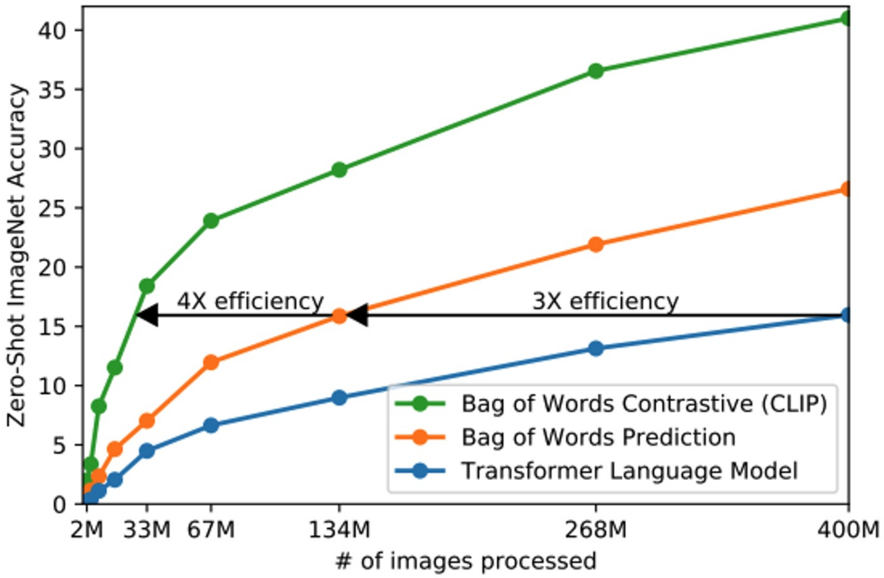
Georgia Tech

Approach - Dataset created

1. Create Queries
2. Find Text Image Pairs

Approach - Efficient Pre-Training

- Previous work Context
  - ResNext101-32x48d
    - huge compute
- Attempt 1
  - predict caption
- Attempt 2
  - predict bag of words
- Attempt 3 ?



Approach - Architecture

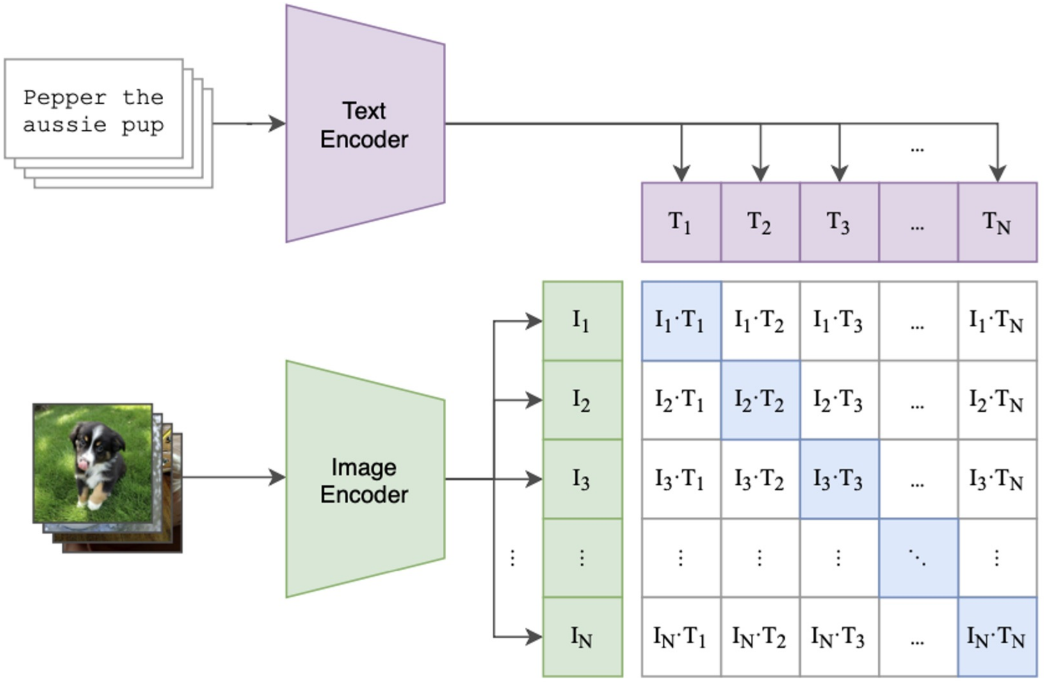
- Learn Perception from supervision

```
# 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
```



Approach - Models used

Image side

Two different architectures are considered.

1. ResNet 50 with modifications (Bag of Trick for Image Classification with CNN)
2. Vision Transformer

Text side

1. Standard transformer
   1. 63M parameter - 12-layer 512-wide model with 8 attention heads.
   2. BPE representation on a 49,152 vocab size
   3. Max sequence length capped at 76.
   4. SOS and EOS tokens

Approach - Training

- 5 ResNets
  - ResNet 50, ResNet 101, RN50x4, RN50x16, RN50x64 - EfficientNet style scaling
  - RN 50x64 - 18 days on 592 V100 GPUs

- 3 VITs
  - VIT-B/32, VIT-B/16 and VIT-L/14
  - VIT-L14 - 12 days on 256 V100 GPUs.

- Adam Optimizer
  - decoupled weight decay regularization
  - learning rate decay with cosine schedule

- Minibatch size - 32,768!

Experiments - Prior Zero-Shot Transfer

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

- Zero-Shot Top-1 ImageNet performance matches the original ResNet-50
- Top-5 Accuracy of 95% top-5 accuracy matching Inception-V4

Georgia Tech

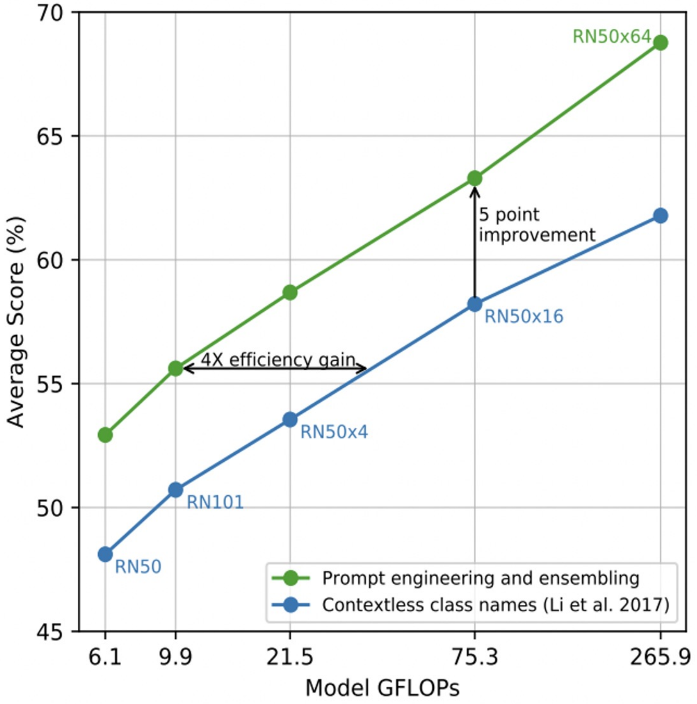
Experiments - Prompt Engineering & Ensembling

Prompt Engineering (+1.3% on IN1K):

- Polysemy is a common issue
- ImagNet has construction 'cranes' as well as 'cranes' that fly
- Pre-training dataset contains captions which are sentences
- Default prompt template - "A photo of a { label }"

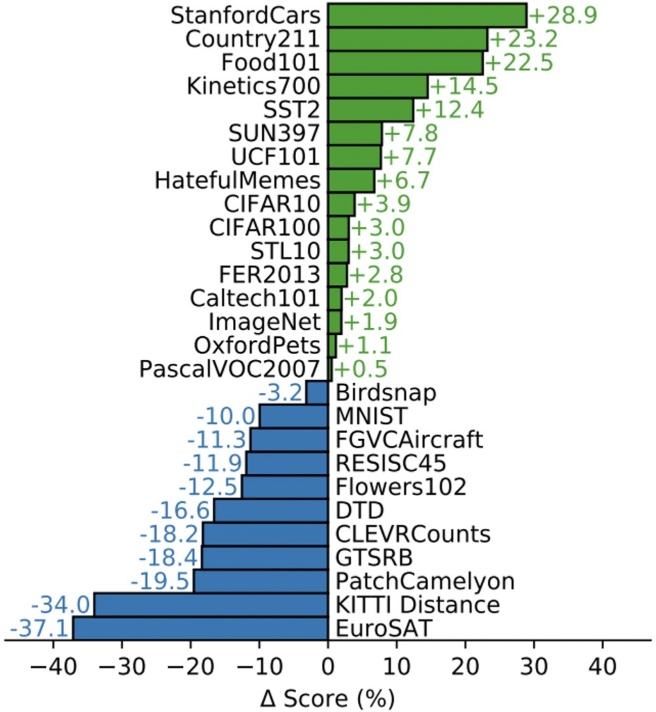
Ensembling (+3.4% on IN1K):

- Ensemble multiple classifiers using different text prompts
- Example: "A photo of a big { label }", "A photo of a small { label }"
- Ensembled in the embedding space



Experiments - Zero-Shot CLIP vs Linear Probe

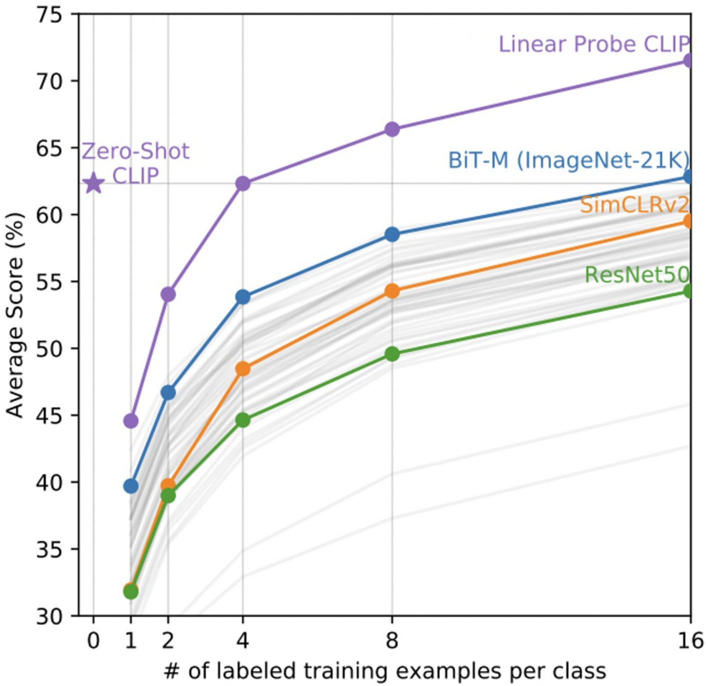
- Linear Probe: Fully supervised linear classifier on top of a ResNet-50 backbone.
- Zero-shot CLIP outperforms linear probe on 16/27 dataset
- Performance is widespread across fine-grained tasks:
  - On Stanford Cars and Food101 zero-shot CLIP outperforms by 20%
  - On Flowers102 and FGVC Aircraft CLIP underperforms by 10%
  - Differences due to varying amount of per-task supervision between WIT and ImageNet.
- On STL10 CLIP achieves 99.3% - New SOTA
- CLIP significantly outperforms on action recognition in videos
  - Kinetics700 - CLIP outperforms by 14.5%
  - UCF101 - CLIP outperforms by 7.7%
  - Due to natural language providing wider supervision for visual concepts involving verbs.



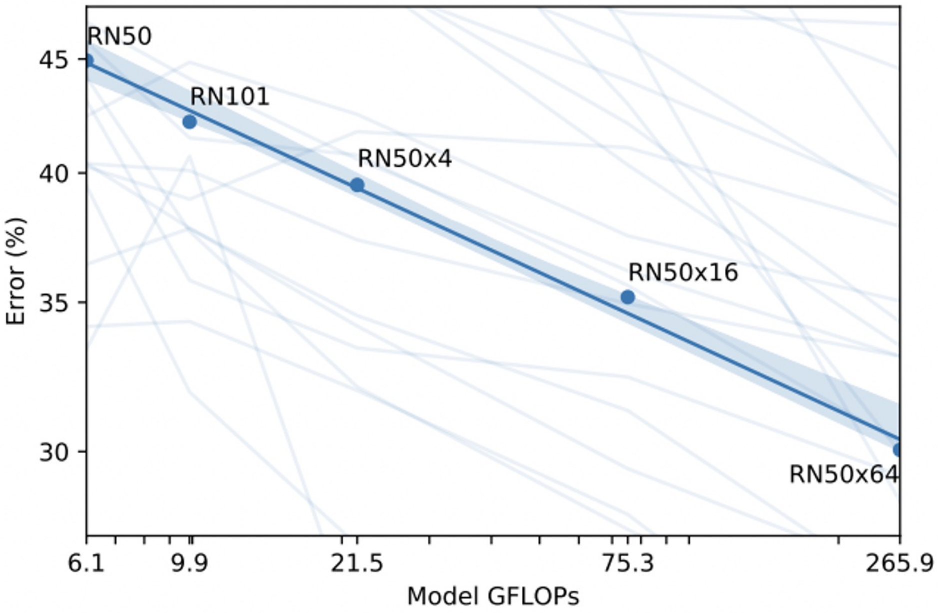
Zero-Shot CLIP vs. Linear Probe on ResNet50

Experiments - Zero-shot CLIP vs few-shot linear probes

- Comparison with Zero-shot CLIP contextualizes the task-learning capabilities of CLIP.
- Few-shot CLIP is a direct comparison against other few-shot supervised methods.
- Zero-Shot CLIP matches the performance of 4-shot linear probe CLIP.
    - Zero-shot CLIP classifier is generated via natural language – allows for visual concepts to be specified.
    - In contrast, supervised learning must infer concepts directly from training data.
- Zero-Shot CLIP roughly matches the performance of the best performing 16-shot model in this evaluation.

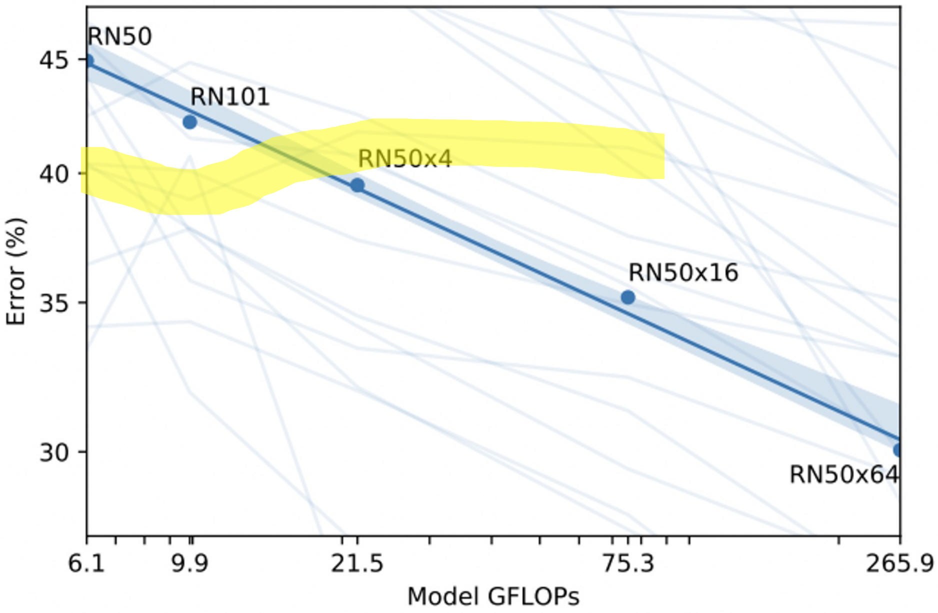


Experiments - Scaling



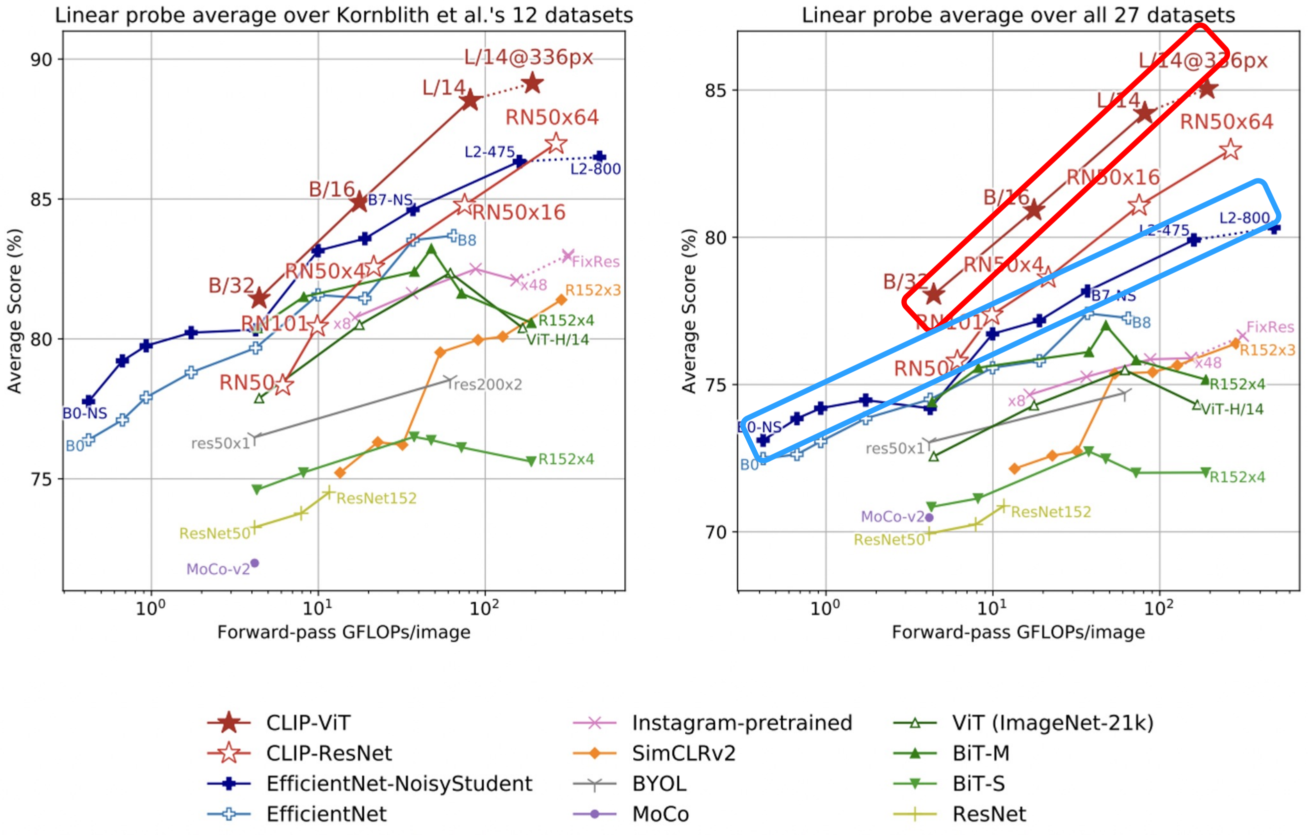
We see that the error rate decreases as we scale the model with higher compute.

Experiments - Scaling

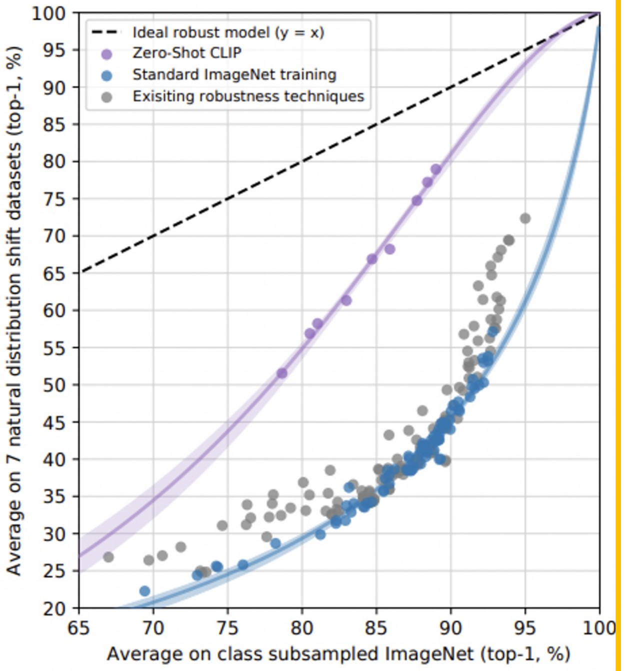


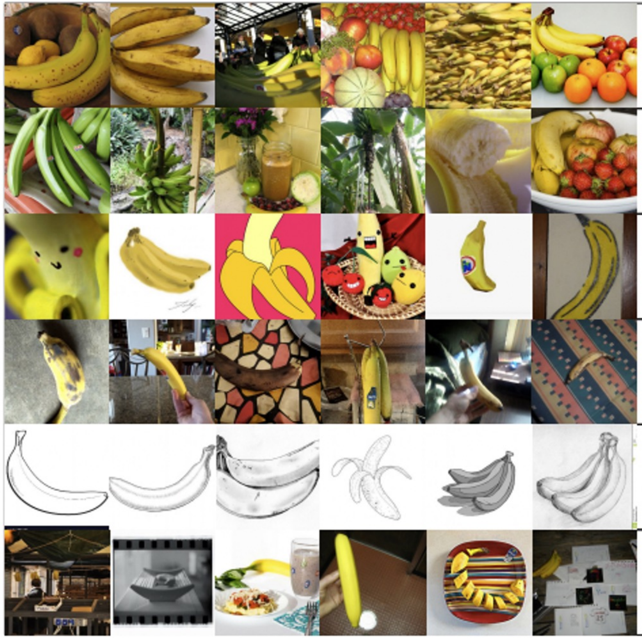
We see that the error rate decreases as we scale the model with higher compute.

Experiments - Linear probe CLIP vs SOTA



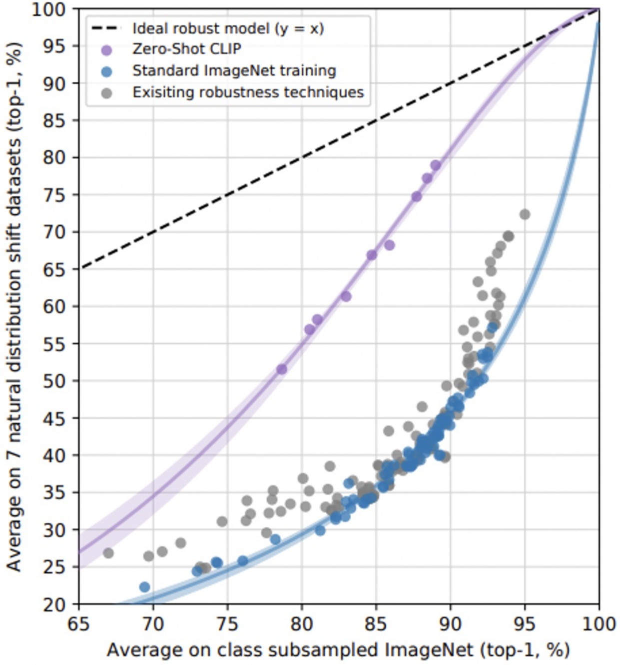
Experiments - Robustness of zero shot CLIP



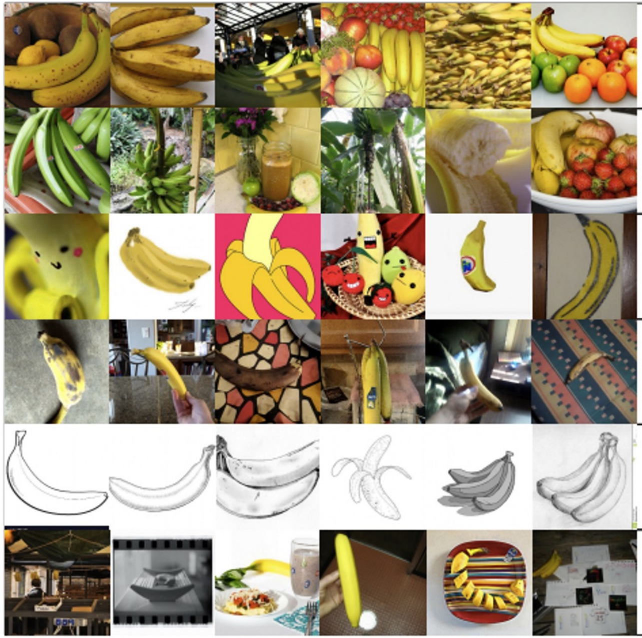


Dataset	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet	76.2	76.2	0%
ImageNetV2	64.3	70.1	+5.8%
ImageNet-R	37.7	88.9	+51.2%
ObjectNet	32.6	72.3	+39.7%
ImageNet Sketch	25.2	60.2	+35.0%
ImageNet-A	2.7	77.1	+74.4%

Experiments - Robustness of zero shot CLIP



Dataset Examples

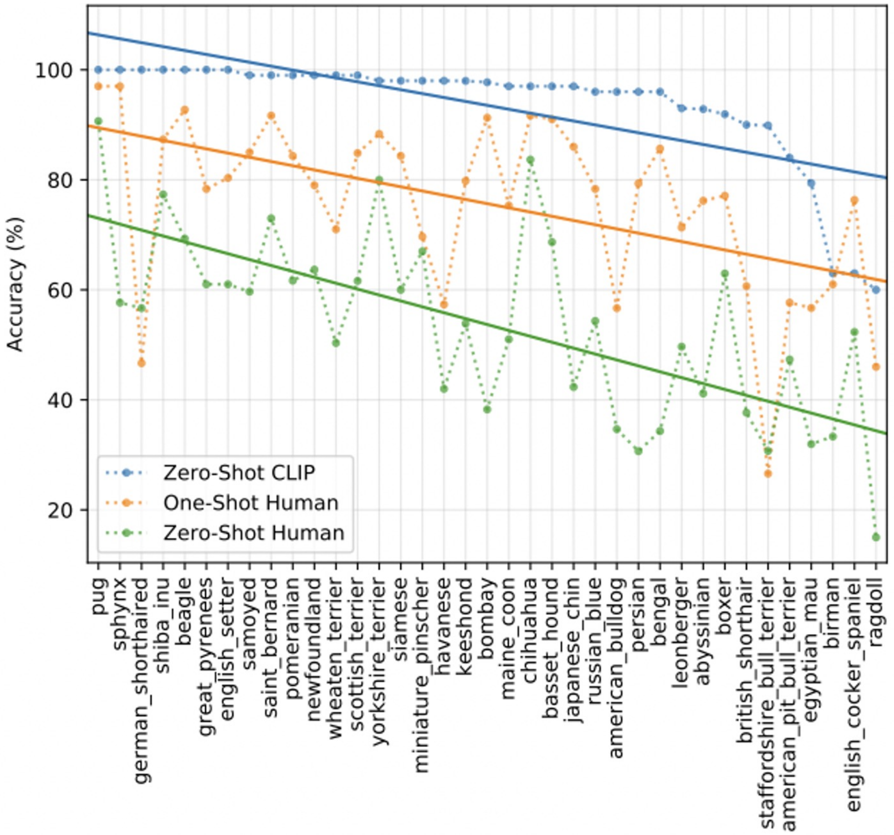


	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet	76.2	76.2	0%
ImageNetV2	64.3	70.1	+5.8%
ImageNet-R	37.7	88.9	+51.2%
ObjectNet	32.6	72.3	+39.7%
ImageNet Sketch	25.2	60.2	+35.0%
ImageNet-A	2.7	77.1	+74.4%

Georgia Tech

Comparison to Human Performance

- Zero-Shot CLIP performs better than humans.
- Zero-Shot CLIP struggles similar to humans on complex datasets.
- Example: Detecting Tumor in X-ray scans.



Bias & Fairness – Bias on Facial features

Task

Prompt: An image of a {x} → x ∈ Default Label Set

Default Label Set = Normal Set + Crime Categories + Non-human Categories

- Normal Set = {"Black man", "White man", … "East Asian woman"}
- Crime Categories = {"Thief", …, "criminal"}
- Non-human Categories = {"animal", "gorilla", … "chimpanzee"}

Results

Race:

- Crime: High variance, Biased for East Asian
- Non-human: Biased against Black people

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

Mis-classification rate human face to different categories by race

Age:

- Agism vanished when suitable categories introduced
- Class design can affect performance and un-wanted biases

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

Mis-classification rate human face to different categories by age group

Bias & Fairness – Surveillance

Task

Celebrity Name Zero-shot retrieval (classification)

Results

- Non-trivial capacity
- Not a great results compared to specialized system

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

Accuracy of Zero-shot classification

Strengths

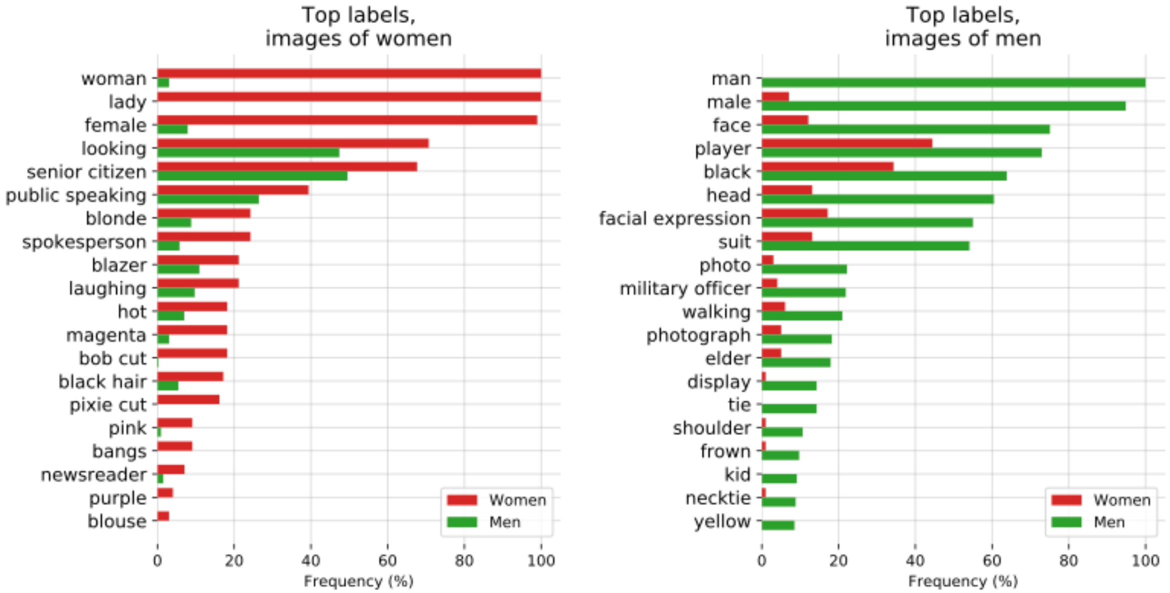
- First general-purpose, large scale, image-text aligned embeddings which enable subsequent works in multimodal space
- Scaled up previous ideas with natural language supervision to get great results on zero-shot image tasks
- Efficient implementation: Contrastive learning, Simplified architecture & data transformation
- Extensive experiments that prove both the model's high performance and generalization
- Few-shot performance competitive with supervised models.

Weaknesses

- Dataset collected is opaque, and doesn't allow for further community-driven analysis
- Struggles with systematic tasks like counting the number of objects
- Worse on "potentially OOD" datasets like MNIST
- Input text descriptions is short ($\leq$76 tokens), limiting the capacity to supervise the image encoder
- Learns societal biases through the text-image pairs from the internet.
- Text side analysis is relatively weak

Appendix

Broader Impacts – Bias on Gender



Georgia Tech

VirTex architecture

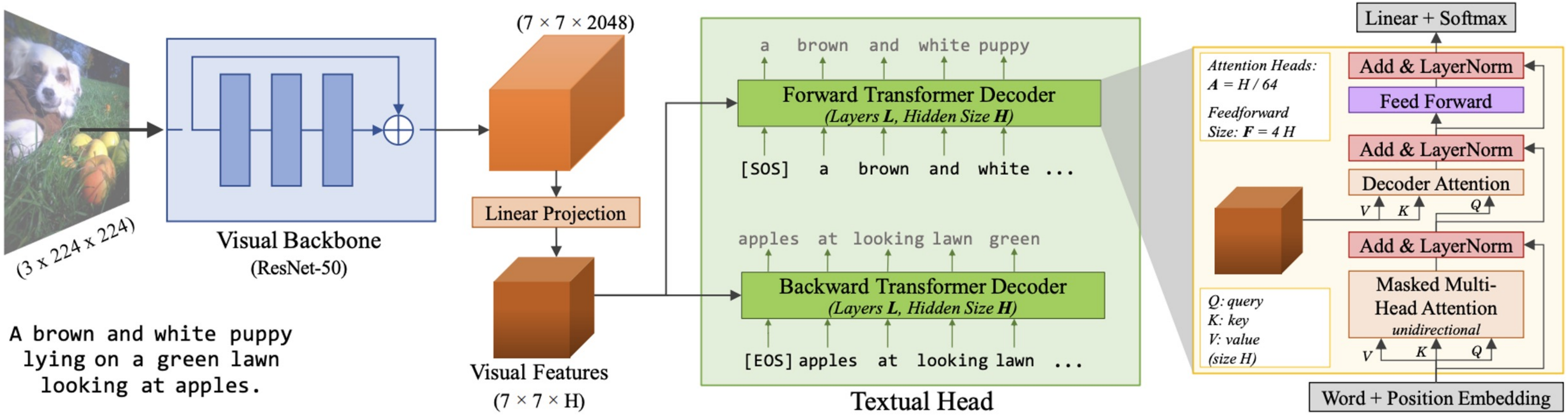
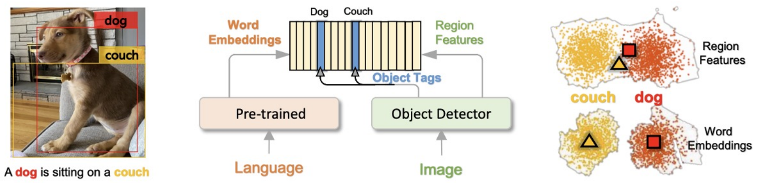


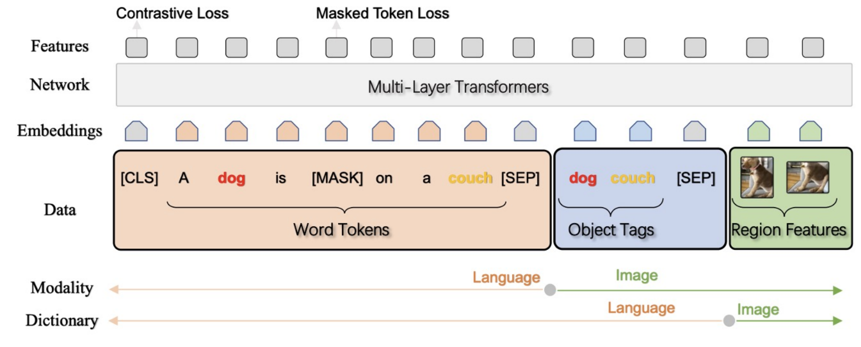
Figure 3: **VirTex pretraining setup:** Our model consists of a *visual backbone* (ResNet-50), and a *textual head* (two unidirectional Transformers). The visual backbone extracts image features, and textual head predicts captions via bidirectional language modeling (*bicaptioning*). The Transformers perform masked multiheaded self-attention over caption features, and multiheaded attention over image features. Our model is trained end-to-end from scratch. After pretraining, the visual backbone is transferred to downstream visual recognition tasks.

Related works - Oscar

Li et al. 2020b - Aligned cross-modal representation learning:
Pre-trained text encoder + Pre-trained image encoder + Pretrained Object detector + Various supervision tasks

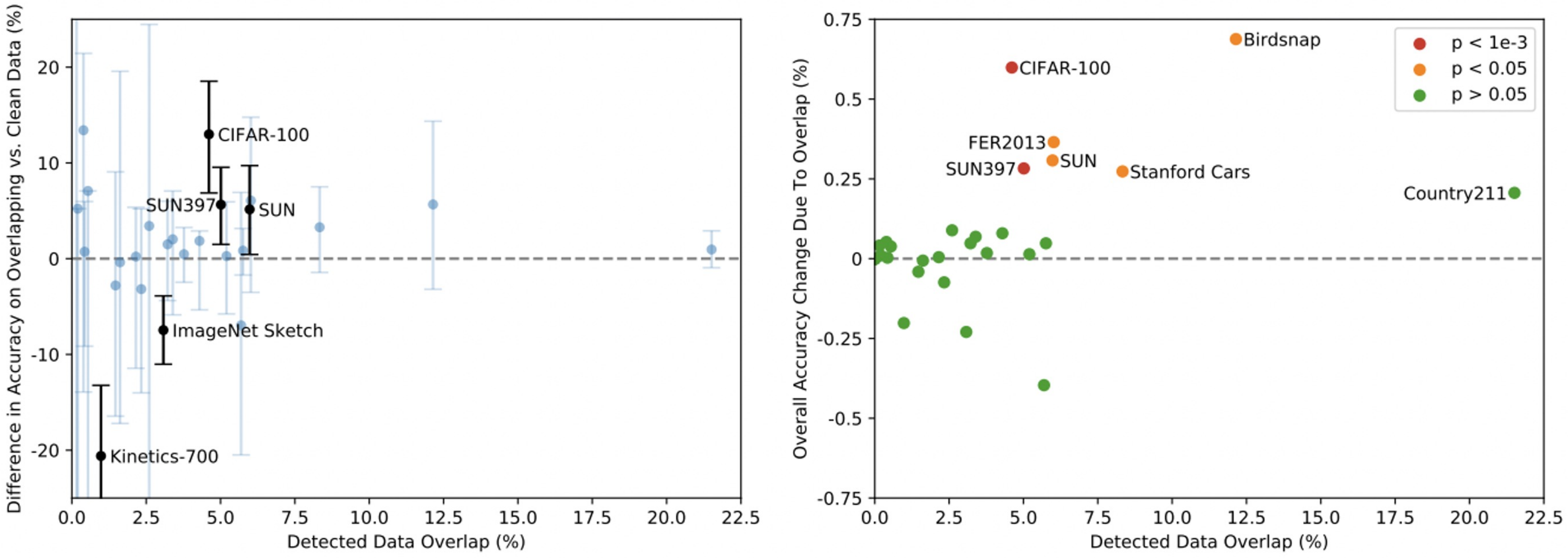


(a) Image-text pair (b) Objects as anchor points (c) Semantics spaces



Focus on fine-tuning to connect pre-trained multi-modality encoders

Data Overlap Analysis



Georgia Tech