

CLIP: The Anchoring Inflection Point

Vineeth N Balasubramanian

Department of Computer Science and Engineering
Indian Institute of Technology, Hyderabad



Contrastive Language Image Pre-training (CLIP)¹

The Paper

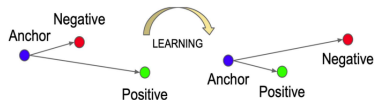
Learning Transferable Visual Models from Natural Language Supervision (ICML '21)

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

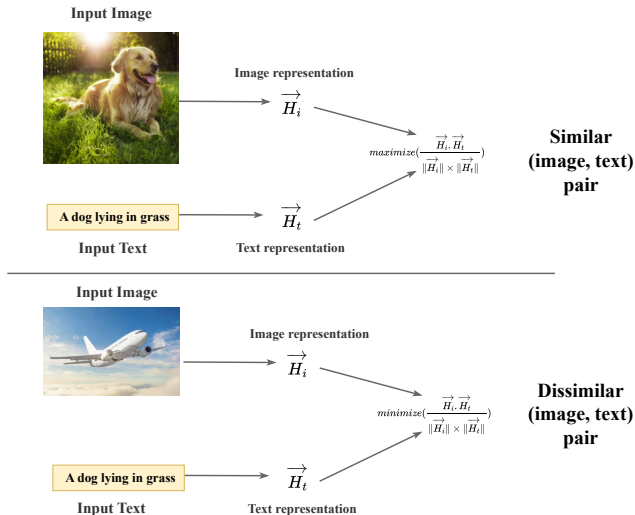
- Mechanism for text supervision in vision models
- Pair an image with its caption using contrastive learning
- Outperforms fully supervised learning baseline on many datasets
- Can be used as a zero-shot classifier!

¹<https://openai.com/research/clip>

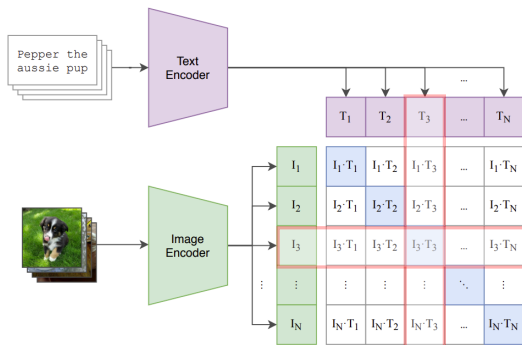
Recall: Contrastive Learning



Recall the **triplet loss** that minimizes the distance between an *anchor* and a *positive*, both of which have the same identity, and maximizes the distance between the *anchor* and a *negative*



CLIP: Training



m_i - one-hot encoded label vector for the i^{th} image sample

y_i^m - cosine similarities vector for i^{th} image sample

t_i - one-hot encoded label for the i^{th} text sample

y_i^t - cosine similarities vector for the i^{th} text sample

ϕ - Cross entropy loss

$$L_m = \frac{\sum_{i=1}^N \phi(y_i^m, m_i)}{N}; \quad L_t = \frac{\sum_{i=1}^N \phi(y_i^t, t_i)}{N}$$

$$L = \frac{L_m + L_t}{2}$$

CLIP: Pseudocode

Simple to implement!

```
# 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

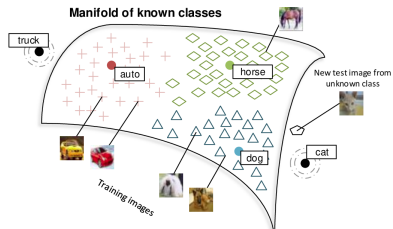
# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T) #[n, d_t]

# 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
```

Zero-Shot Learning²

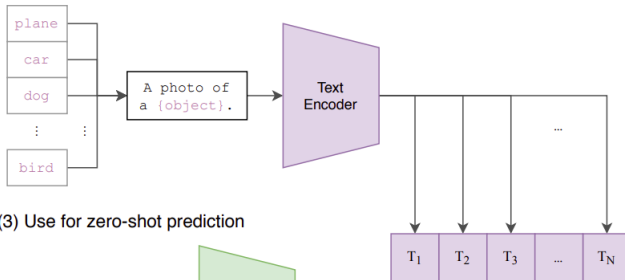


- Unlike supervised learning, model is trained on seen class labeled data – however, it is evaluated on unseen (and seen) classes
- Traditional image classification models are limited to a fixed label space, and may lack generalization to new labels at inference time
- Recall zero-shot learning methods like Relation Networks, f-CLSWGAN, meta-learning methods, etc

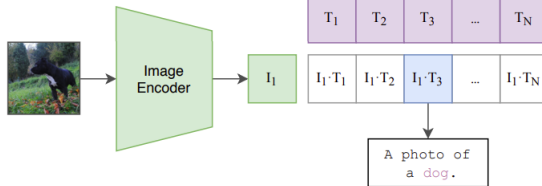
²Socher et al, “Zero-Shot Learning Through Cross-Modal Transfer”, NeurIPS 2013

CLIP for Zero-Shot Classification

(2) Create dataset classifier from label text



(3) Use for zero-shot prediction



Training the CLIP Model

Training Datasets

- Existing crowd-labeled datasets: MS-COCO, Visual Genome, YFCC100M
 - MS-COCO and Visual Genome relatively small with 100,000 images each
 - YFCC100M an alternate dataset with relatively sparse metadata; filtering images to retain ones with natural description titles results in 15M which is similar in magnitude as ImageNet
- WebImageText (WIT)** dataset is hence introduced; consists of 400M (image,text) pairs curated from various publicly available sources on the internet
 - Similar to WebText dataset used to train GPT-2 (same word count)

Training the CLIP Model: Other Implementation Details

Architectures

- **Vision:**
 - 5 Resnet Variants: ResNet-50, ResNet-101 and EfficientNet-Style model scaling upto 4x, 16x and 64x of ResNet-50
 - 3 ViT Variants: ViT-B/16, ViT-B/32, ViT-L/14
- **Text:** Adaptation of standard Transformer, similar to GPT

Training the CLIP Model: Other Implementation Details

Architectures

- **Vision:**
 - 5 Resnet Variants: ResNet-50, ResNet-101 and EfficientNet-Style model scaling upto 4x, 16x and 64x of ResNet-50
 - 3 ViT Variants: ViT-B/16, ViT-B/32, ViT-L/14
- **Text:** Adaptation of standard Transformer, similar to GPT

Optimization Details

- Adam optimizer; Temperature scaling factor of 0.007
- Batch size of 32,768; Models trained for 32 epochs

Training the CLIP Model: Other Implementation Details

Architectures

- **Vision:**
 - 5 Resnet Variants: ResNet-50, ResNet-101 and EfficientNet-Style model scaling upto 4x, 16x and 64x of ResNet-50
 - 3 ViT Variants: ViT-B/16, ViT-B/32, ViT-L/14
- **Text:** Adaptation of standard Transformer, similar to GPT

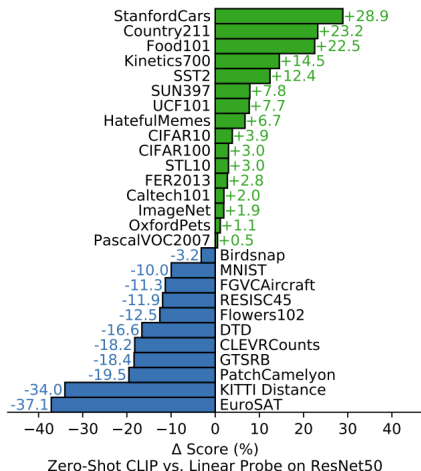
Optimization Details

- Adam optimizer; Temperature scaling factor of 0.007
- Batch size of 32,768; Models trained for 32 epochs

Hardware and Training Time

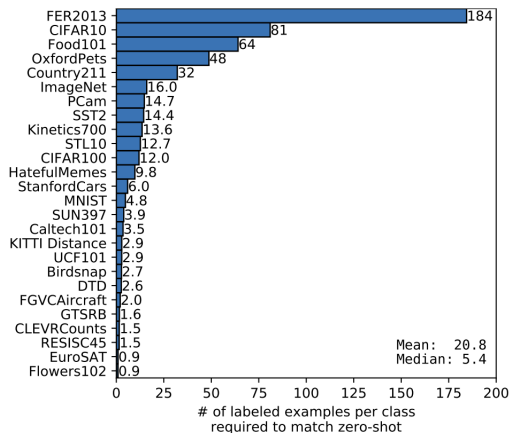
- For the largest ResNet (RN50x64), it took 18 days on 592 V100s
- For the largest ViT (ViT-L/14), it took 12 days on 256 V100s

Experiments: CLIP's Zero-Shot Transfer



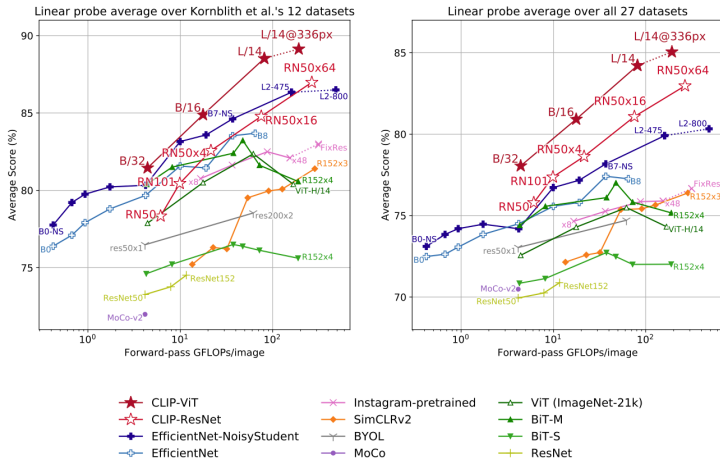
- CLIP outperforms baselines on wide variety of popular datasets
- Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet
- Underperforms on complex datasets like Satellite images, Tumor images
- Not suited for hyper-specific tasks, requires fine-tuning

Experiments: CLIP's Zero Shot Transfer



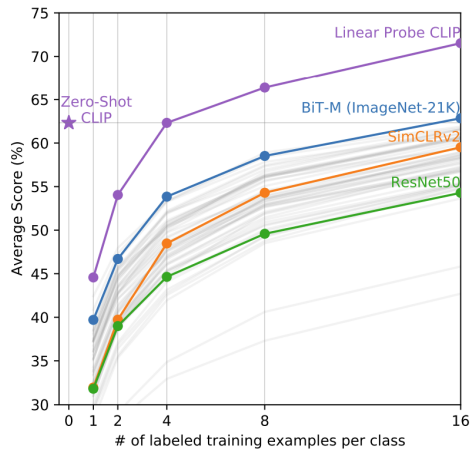
The number of labeled examples per class required to train a linear classifier on the CLIP feature space to match the performance of the zero-shot classifier contextualizes the effectiveness of zero-shot transfer in CLIP

Experiments: Linear Probing Performance of CLIP









ViT based CLIP outperforms all baselines; performance gap increases with GFLOPS

Few-shot performance



- Zero-Shot CLIP outperforms few-shot linear probes upto 16 shots
- Linear probe CLIP outperforms all baselines with a glaring margin

Experiments: Robustness to Distribution Shifts

	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%

- Zero-shot CLIP is much more robust to distribution shift than standard ImageNet models
- Visualizing distribution shift for bananas, a class shared across 5 of the 7 natural distribution shift datasets

More Information

Resources

- <https://openai.com/research/clip>
- HuggingFace Link for CLIP