Learning Transferable Visual Models From Natural Language Supervision

CVPR 2021

JEEON BAE
SAMSUNG SW ACADEMY FOR YOUTH 13th

January 30, 2025

Overview

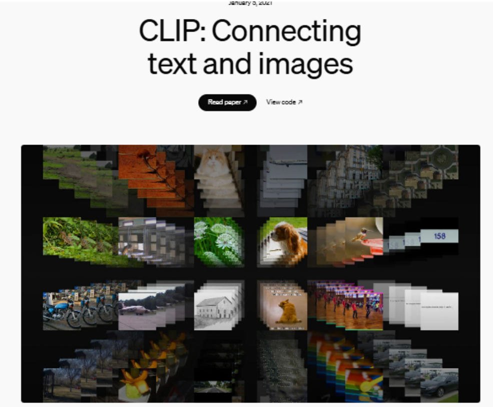
1. Introduction

2. Approach

3. Experiments

4. Conclusion

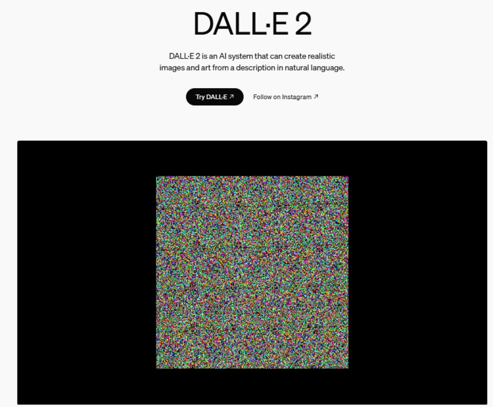
Two image-text models from OpenAI



January 5, 2021

CLIP: Connecting text and images

Figure: CLIP Model

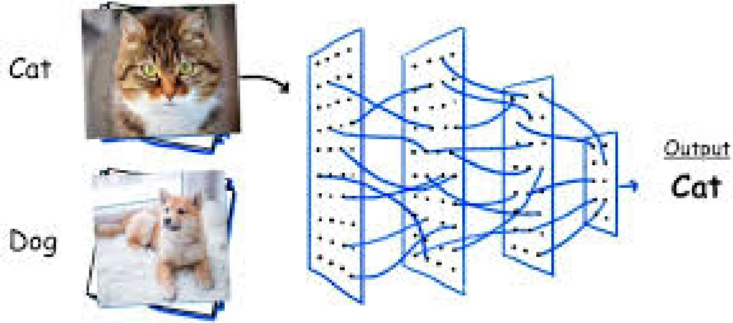


DALL·E 2

DALL·E 2 is an AI system that can create realistic images and art from a description in natural language.

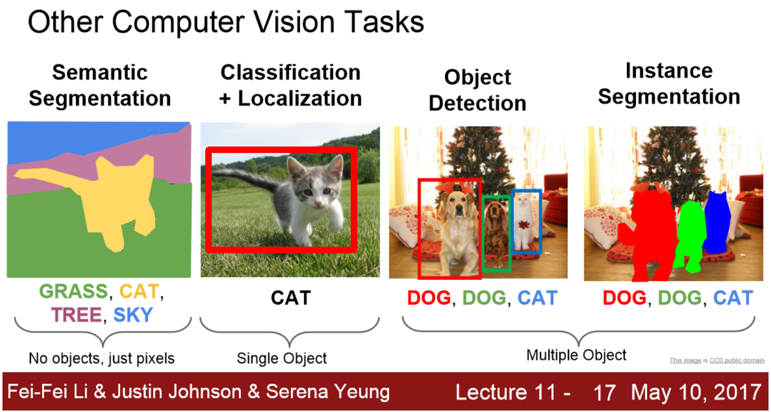
Figure: DALLE Model

Fixed Data Format in SOTA Vision Models



The computer vision models are trained using fixed [*image*, *label*] data formats.

Limitations of Fixed Data Formats



Other Computer Vision Tasks

Semantic Segmentation

Classification + Localization

Object Detection

Instance Segmentation

GRASS, CAT, TREE, SKY

CAT

DOG, DOG, CAT

DOG, DOG, CAT

No objects, just pixels

Single Object

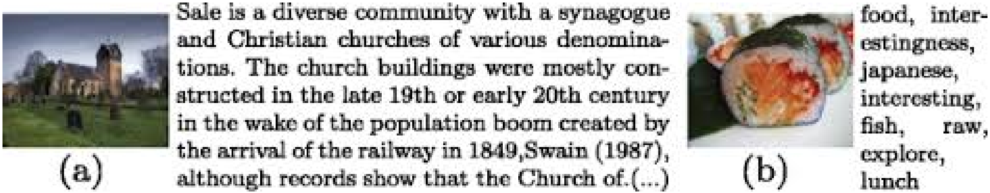
Multiple Object

This image is CC0 public domain

Fei-Fei Li & Justin Johnson & Serena Yeung Lecture 11 - 17 May 10, 2017

Training with such fixed data formats limits the generalizability and usability

Proposed Solution: Using Raw Text Descriptions



Sale is a diverse community with a synagogue and Christian churches of various denominations. The church buildings were mostly constructed in the late 19th or early 20th century in the wake of the population boom created by the arrival of the railway in 1849,Swain (1987), although records show that the Church of.(...)

food, interestingness, japanese, interesting, fish, raw, explore, lunch

- Proposed using raw text descriptions as labels instead of simple categorical labels to enhance generalization and adaptability.
- The pre-training in raw text has led to huge success in NLP.

Methods

Natural Language Supervision

Utilizing natural language to supervise vision models, allowing broader generalization.

Creating a Sufficiently Large Dataset

Scaling up data collection to improve model performance and generalization.

Selecting an Efficient Pre-Training Method

Optimizing training strategies to enhance model efficiency and accuracy.

Natural Language Supervision

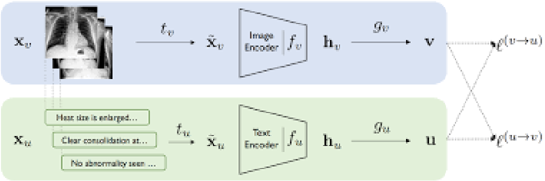


Figure: ConVIRT Model

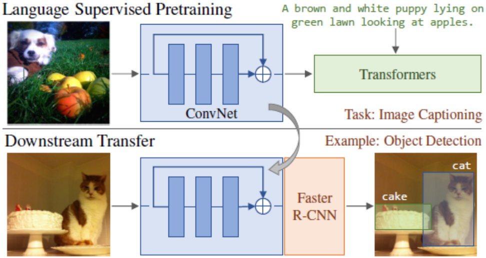

Language Supervised Pretraining
A brown and white puppy lying on green lawn looking at apples.
ConvNet
Transformers
Task: Image Captioning
Downstream Transfer
Example: Object Detection
Faster R-CNN
cat
cake

Figure: VirTEX Model

Natural Language Supervision

Key Idea: Learning perception from supervision contained in natural language.

Motivation:

- Prior works describe similar methods with different terms: **unsupervised**, **self-supervised**, **weakly supervised**, **supervised**.
- Commonality: Using **natural language** as a training signal.

Advantages of Learning from Natural Language:

- **Scalability:** Easier than crowd-sourced labeling since it does not require structured annotations.
- **Passive Learning:** Models can leverage vast textual data available on the internet.
- **Zero-shot Transfer:** Connects learned representations directly to language, enabling flexible adaptation.

Conclusion: Natural language supervision is a promising alternative to traditional training methods.

Creating a Sufficiently Large Dataset

Challenges:

- **MS-COCO, Visual Genome**: High-quality but small (100K images each).
- **YFCC100M**: Larger (100M images), but lacks structured metadata.
- Filtering for meaningful English titles/descriptions shrinks the dataset by **6×** to only 15M images.

New dataset: WebImageText (WIT)

- **400M (image, text) pairs** collected from publicly available web sources.
- Designed to cover a broad range of visual concepts using **500,000 queries**.
- Up to **20,000 pairs per query** to ensure class balance.
- Similar word count to WebText dataset used in GPT-2 training.

Challenges in Pre-Training Large-Scale Vision Models

Computational Cost: - State-of-the-art vision models require massive compute resources. - ResNeXt101-32x48d: **19 GPU-years** (Mahajan et al., 2018). - Noisy Student EfficientNet-L2: **33 TPUv3 core-years** (Xie et al., 2020).

Limitations of Early Approaches:

- Inspired by VirTex, early methods jointly trained an image CNN and text transformer.
- These models struggled to scale efficiently.
- A 63M parameter transformer learned ImageNet classes **3x slower** than a simpler bag-of-words model.

Contrastive Learning as an Efficient Alternative

Problem with Early Methods: - Predicting the exact words in captions is difficult due to the diversity of descriptions.

Key Insight: - Contrastive learning provides better representations than predictive models (Tian et al., 2019). - Generative models can learn high-quality features but require **10x more computationation** (Chen et al., 2020a).

New Approach: - Instead of predicting exact words, learn to identify which text belongs to which image. - Swapping a predictive objective with a contrastive objective improved **zero-shot transfer efficiency by 4x**.

CLIP Training and Optimization

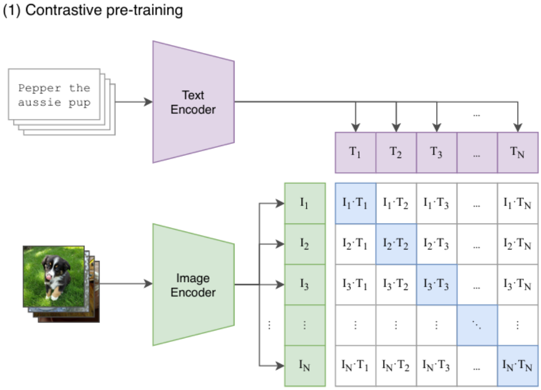
Training Objective: - Given a batch of **N (image, text) pairs**, predict which of the **N²** possible pairs are correct. - Jointly train an image encoder and a text encoder to maximize cosine similarity for correct pairs. - Minimize similarity for incorrect pairs using symmetric cross-entropy loss.

Implementation Simplifications:

- Trained **from scratch** without ImageNet or pre-trained weights.
- Removed non-linear projection in contrastive embedding space.
- Simplified image transformation: only **random square crop** as augmentation.
- Optimized temperature parameter as a **trainable scalar**.

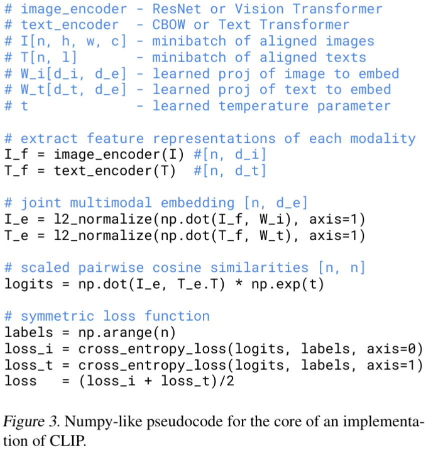
Outcome: - A more efficient model with **4x faster zero-shot transfer** compared to previous approaches.

Contrastive Pre-training



	T$_1$	T$_2$	T$_3$	...	T$_N$
I$_1$	I$_1$·T$_1$	I$_1$·T$_2$	I$_1$·T$_3$	...	I$_1$·T$_N$
I$_2$	I$_2$·T$_1$	I$_2$·T$_2$	I$_2$·T$_3$	...	I$_2$·T$_N$
I$_3$	I$_3$·T$_1$	I$_3$·T$_2$	I$_3$·T$_3$	...	I$_3$·T$_N$
⋮	⋮	⋮	⋮	⋱	⋮
I$_N$	I$_N$·T$_1$	I$_N$·T$_2$	I$_N$·T$_3$	...	I$_N$·T$_N$

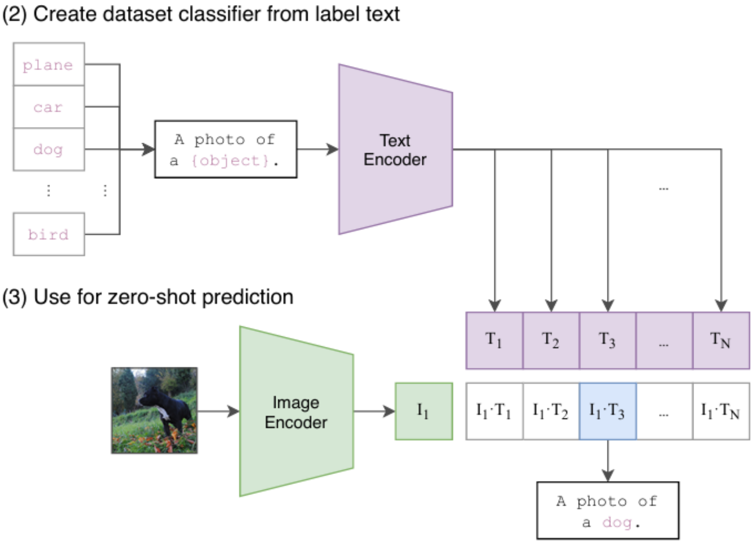
Figure: Overview of CLIP



```
# 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.

Inference


(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
T_1
T_2
T_3
...
T_N
Image Encoder
I_1
$I_1 \cdot T_1$
$I_1 \cdot T_2$
$I_1 \cdot T_3$
...
$I_1 \cdot T_N$
A photo of a dog.

What is Zero-Shot Learning?

Traditional Definition: - Zero-shot learning refers to **generalizing to unseen object categories** in image classification (Lampert et al., 2009).

Our Definition: - We extend the concept to **generalization to unseen datasets**. - This serves as a proxy for learning unseen tasks, inspired by **Zero-Data Learning** (Larochelle et al., 2008).

Why is this Important? Unsupervised learning focuses on **representation learning**. - We study zero-shot transfer to measure **task learning** in machine learning systems.

Prompt Engineering for Zero-Shot Transfer

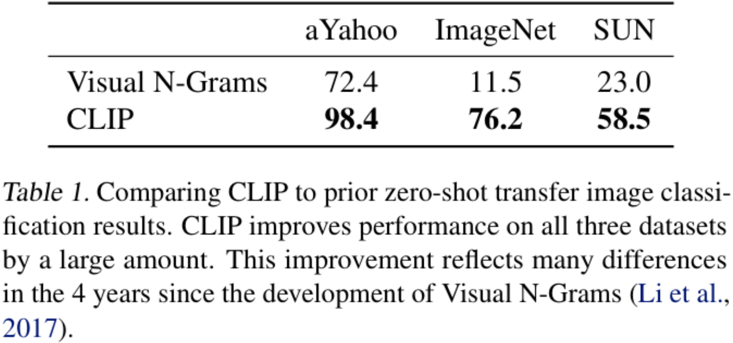
Problem: Challenges in Zero-Shot Transfer

- Many datasets provide only **numeric labels**, limiting natural language supervision.
- **Polysemy issue**: Class names lack context.
- Example: *crane* (bird) vs. *crane* (construction).

Solution: Contextualized Prompts

- Simple template: **"A photo of a [label]."** (+1.3% ImageNet accuracy).
- Task-specific improvements:
  - **Oxford-IIIT Pets:** "A photo of a [label], a type of pet."
  - **OCR datasets:** "A photo of '[label]'."
  - **Satellite images:** "A satellite photo of a [label]."

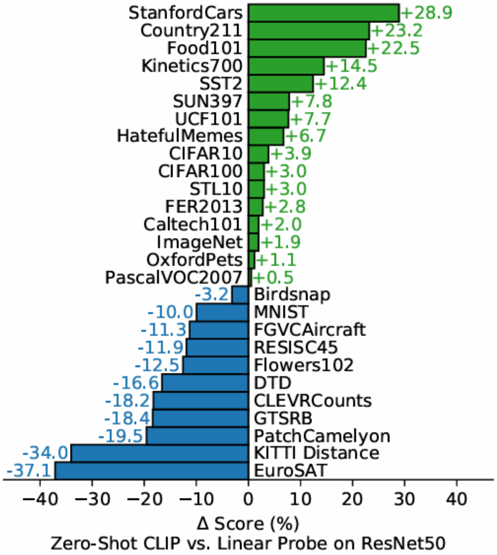
EXP 1: Zero-shot Transfer (compare w/ Visual N-grams)



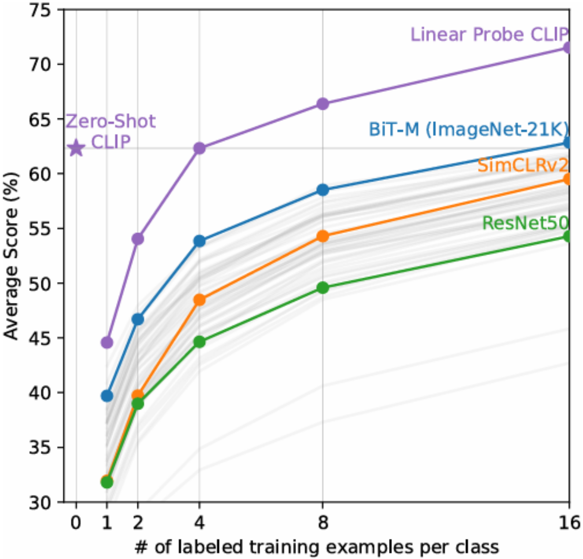
	aYahoo	ImageNet	SUN
Visual N-Grams	72.4	11.5	23.0
CLIP	**98.4**	**76.2**	**58.5**

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).

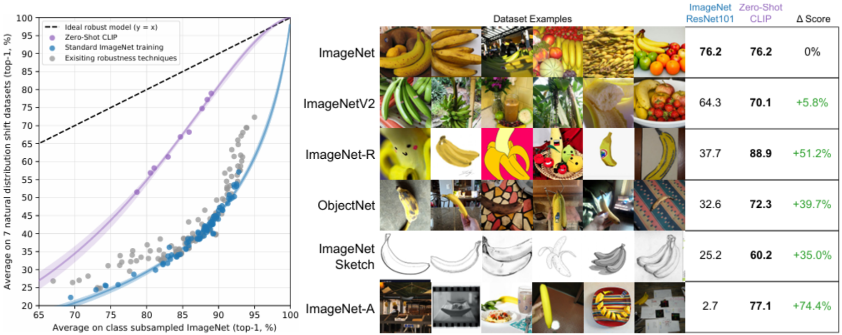
EXP 2: Zero-shot Transfer (compare w/ supervised model)



EXP 3: Zero-shot Transfer vs Few-shot



EXP 4: Robustness to natural distribution shift



	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%

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.

Conclusion

Key Findings

- Task-agnostic **web-scale pre-training**, successful in NLP, also benefits **computer vision**.
- CLIP learns **diverse tasks** during training, enabling **zero-shot transfer** via **natural language prompting**.
- At sufficient scale, this approach **competes with task-specific supervised models** but still has room for improvement.

Social Implications

- Generalized models can impact **AI accessibility, fairness, and bias**.
- Further research is needed to **refine and improve task transfer capabilities**.

References

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

The End