Review of *Learning Transferable Visual Models From Natural Language Supervision*

CLIP

Matthew Evans

May 8 , 2025

Overview

The Contrastive Language-Image Pre-Training (CLIP)[1] model addresses the need for flexible image classification without task-specific labeled data by learning from the vast collection of (image, caption) pairs available online. By jointly training image and text encoders to align their representations, CLIP enables zero-shot recognition of novel categories simply specified in natural language. This design removes the requirement for retraining on each new task, greatly reducing both the annotation burden and deployment time.

Approach

Prior Work

CLIP builds on a rich history of using natural language to guide visual learning and on joint image–text embedding methods. Key prior developments include:

- **Distributional Semantics & Language Models.** Early work in NLP established that word co-occurrence statistics (e.g., word2vec[2]) capture meaningful semantics, laying the groundwork for using language representations as supervision for other domains.
- **Image–Text Retrieval & Joint Embeddings.** From Mori et al. (1999)[3] through kernel Canonical Correlation Analysis (CCA)[4] and ranking losses[5], researchers learned to align image and text modalities in a shared space to enable caption retrieval and cross-modal search.
- **Natural Language Supervision for Vision.** Methods like Ramanathan et al. (2013)[6] and He & Peng (2017)[7] showed that free-form image captions and descriptions can improve tasks such as video event recognition and fine-grained classification without requiring curated class labels.
- **Webly Supervised & Large-Scale Datasets.** Webly supervised approaches [8, 9] used noisy image–query pairs from search engines to train classifiers, while recent automatically constructed caption datasets (e.g., Conceptual Captions) demonstrated the value of scaling to millions of examples.
- **Contrastive & Self-Supervised Vision.** Advances in contrastive learning (InfoNCE[10], MoCo[11], SimCLR[12]) revealed how pulling together augmented views of the same image can yield powerful representations, a principle that CLIP extends to the multi-modal setting.

Novelty

The CLIP approach is novel in its combination of web-scale natural-language supervision with a simple, efficient contrastive learning objective to train joint image–text representations. By harnessing 400 million image–text pairs without manual labeling, CLIP achieves broad zero-shot transfer simply by embedding class names and images into a shared vector space. This stands in contrast to prior methods that relied on fixed label sets, small curated datasets, or complex generative losses, making CLIP both scalable and flexible for new tasks.

Web-Scale Natural-Language Supervision

CLIP leverages the raw captions and surrounding text of 400 million images sourced from the internet, rather than hand-annotated labels. This massive, diverse corpus captures a vast variety of visual concepts and contexts, enabling the model to generalize across domains without task-specific retraining.

Contrastive Multimodal Pre-Training

Rather than predicting discrete tokens, CLIP uses an InfoNCE-style loss to align image and text embeddings:

$$\mathcal{L}_i = -\log \frac{\exp(s_{i,i^+}/\tau)}{\sum_{k=1}^{N} \exp(s_{i,k}/\tau)}, \quad s_{i,j} = \frac{f(x_i) \cdot g(t_j)}{\|f(x_i)\| \|g(t_j)\|}.$$

Here f and g are the image and text encoders, and cosine similarity $s_{i,j}$ drives matching pairs together and mismatches apart within each batch.

Prompt-Based Zero-Shot Classification

At inference, CLIP treats class names or natural-language descriptions as "prompts," embedding them via the text encoder and comparing to image embeddings. This nearest-neighbor softmax over cosine similarities creates a zero-shot classifier that requires no additional gradient updates or task-specific head.

Compound Model Scaling & Architecture

CLIP experiments with both ResNet- and Vision Transformer–based image encoders, applying compound scaling (width, depth, resolution) and attention pooling to improve capacity. The text encoder is a 12-layer Transformer over BPE tokens, scaled in width to match visual model compute, ensuring balanced multi-modal representation power.

Efficient Large-Batch Training

Utilizing huge minibatches (32,768 pairs), mixed-precision, gradient checkpointing, and a learnable temperature τ, CLIP achieves rapid convergence despite its scale. This efficient setup makes training on billions of computed similarities tractable, unlocking high zero-shot performance without impractical compute overhead.

Considerations

Strengths

- **Broad Zero-Shot Transfer.** CLIP matches or exceeds supervised baselines on 16 out of 27 diverse benchmarks without any task-specific training examples, including fine-grained (STL-10) and action recognition (UCF101) datasets.
- **Robustness to Natural Distribution Shift.** In out-of-distribution tests like ImageNetV2 and ObjectNet, CLIP closes up to 75% of the accuracy gap between in- and out-of-distribution performance compared to ImageNet-trained models.
- **Scalable Contrastive Pre-Training.** By aligning image and text embeddings via a simple InfoNCE loss over 400 M pairs, CLIP converges 4× faster than generative captioning objectives and leverages raw web text at scale.
- **Prompt-Based Flexibility.** Users can define new classification tasks at inference with natural-language prompts alone, avoiding any model fine-tuning and greatly reducing deployment time and annotation burden.
- **Balanced Model Scaling.** Compound scaling of ResNets (width, depth, resolution) and proportional width scaling of the text Transformer deliver consistent model capacity increases across both modalities, enabling the model to leverage web-scale data.

Weaknesses

- **High Compute & Data Requirements.** Training CLIP (32,768-sample batches, mixed precision, checkpointing) on 400 M pairs demands hundreds of GPUs for weeks; extrapolating to state-of-the-art zero-shot performance would require $\sim$1000$\times$ more compute, which is currently impractical.
- **Poor Fine-Grained & Systematic Reasoning.** CLIP struggles to differentiate closely related subclasses (e.g., car models, flower species) and tasks requiring counting or measurement, often performing near chance.
- **Brittle Out-of-Distribution Generalization.** Despite robustness to natural shifts, CLIP fails on truly novel domains (e.g., handwritten MNIST), where even basic pixel-based classifiers outperform it.
- **Limited Expressivity.** Unlike generative caption models, CLIP's classification is constrained to predefined prompts and cannot produce novel descriptions or explanations without additional mechanisms.
- **Inefficient Few-Shot Adaptation.** Attaching a linear head on frozen features yields minimal improvement from zero to few-shot compared to humans' large gains after one example, highlighting a sample-efficiency gap.

Measures of Success

CLIP's primary validation rests on its ability to generalize via zero- and few-shot classification across diverse vision benchmarks, briefly summarized below.

Zero-Shot

- On ImageNet, CLIP achieves 76.2% top-1 accuracy without any fine-tuning, matching a supervised ResNet-50 trained on ImageNet.
- Across 27 datasets—including general object recognition (CIFAR-10/100), fine-grained classification (Stanford Cars, Food101), and action recognition (UCF101, Kinetics700)—zero-shot CLIP outperforms a supervised linear probe on ResNet-50 features in 16 tasks, setting new state-of-the-art on STL-10 (99.3%) and improving UCF101 by +7.7 points.

Few-Shot

- By training a simple logistic regression head on frozen CLIP embeddings, CLIP's k-shot performance often rivals or exceeds models trained with extensive labeled data.
- Zero-shot accuracy is equivalent to a 16-shot linear probe on the same features, and the median "effective shots" across tasks is only 5.4 examples per class.

Out-of-Distribution

- On natural shift benchmarks (ImageNetV2, ObjectNet, ImageNet-Sketch), CLIP closes up to 75% of the accuracy gap between in-distribution and shifted test sets compared to an ImageNet-trained model.
- Adapting CLIP via a linear ImageNet head improves in-domain accuracy by +9.2% but reduces out-of-distribution robustness, underscoring the strength of pure zero-shot transfer.
- CLIP performs poorly on truly novel domains—for example, handwritten MNIST, where zero-shot accuracy falls below 60% and even raw-pixel logistic regression outperforms it.
- Tasks requiring precise enumeration (e.g., counting objects) or novel distance estimates often yield near-random performance, highlighting limits of its purely contrastive training.

Impact

CLIP's contrastive, multi-modal pre-training paradigm directly inspired large-scale vision–language models such as Google's ALIGN[13], which scales noisy web text–image pairs to billions and demonstrates similarly strong zero-shot classification. Its joint embedding framework also underpins "CLIP guidance" in text-to-image diffusion models—first seen in OpenAI's GLIDE[14]—where a frozen CLIP model steers generation toward semantically relevant samples. Beyond generation, CLIP's image encoder was adopted (with frozen weights) in DeepMind's Flamingo[15] to provide rich visual features for few-shot multimodal reasoning. Subsequent work such as ALBEF[16] and LiT[17] have refined CLIP's contrastive objective with fine-grained alignment and larger datasets, propelling rapid advances across retrieval, classification, and generative tasks. Overall, CLIP's scalable alignment of vision and language has reshaped the field's approach to zero-shot and few-shot learning.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[2] Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013, 01 2013.

[3] Yasuhide Mori, Hironobu Takahashi, and Ryu ichi Oka. Image-to-word transformation based on dividing and vector quantizing images with words. 1999.

[4] Jason Weston, Samy Bengio, and Nicolas Usunier. Large scale image annotation: Learning to rank with joint word-image embeddings. *Machine Learning*, 81:21–35, 10 2010.

[5] Richard Socher and Fei-Fei Li. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. pages 966–973, 06 2010.

[6] Vignesh Ramanathan, Percy Liang, and Li Fei-Fei. Video event understanding using natural language descriptions. In *Proceedings of the 2013 IEEE International Conference on Computer Vision*, ICCV '13, page 905–912, USA, 2013. IEEE Computer Society.

[7] Xiangteng He and Yuxin Peng. Fine-graind image classification via combining vision and language. *CoRR*, abs/1704.02792, 2017.

[8] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1816–1823 Vol. 2, 2005.

[9] Santosh Kumar Divvala, Ali Farhadi, and Carlos Guestrin. Learning everything about anything: Webly-supervised visual concept learning. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3270–3277, 2014.

[10] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *CoRR*, abs/1911.05722, 2019.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020.

[13] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *CoRR*, abs/2102.05918, 2021.

[14] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. *CoRR*, abs/2112.10741, 2021.

[15] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022.

[16] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *CoRR*, abs/2107.07651, 2021.

[17] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. *CoRR*, abs/2111.07991, 2021.