# CLIP

### **Learning Transferable Visual Models From Natural Language Supervision** (2021) 3.1k

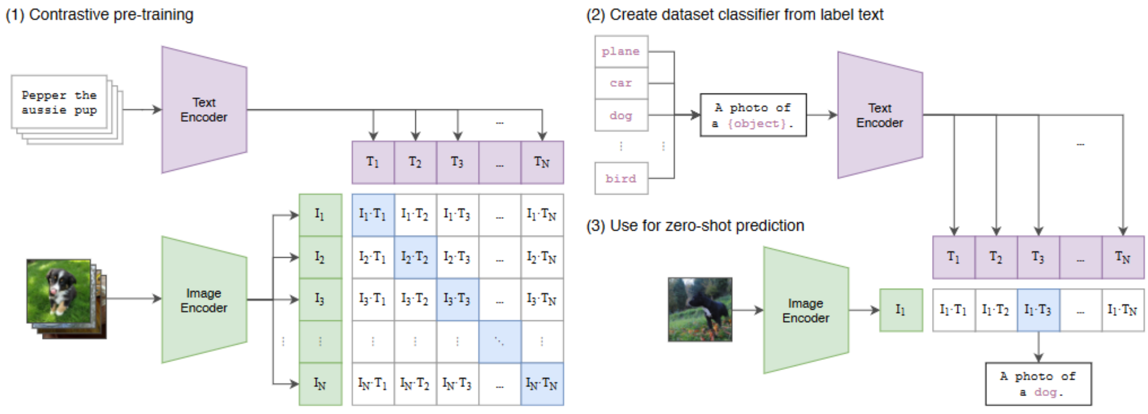


*Figure 1.* Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time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

fixed set of predetermined object categories → limited generality

CLIP: Contrastive Language-Image Pre-training, no classifier or categorical label, transferable to new lables / datasets / tasks: e.g., zero-shot accuracy on ImageNet = ResNet-50, more robust

### **Contrastive pre-training**

larger dataset of (image, text) pairs, larger model

due to training effiency, use contrastive methods instead of gpt; prediction task is too hard, contrastive task is more reasonable

image_encoder(I), text_encoder(T) → cosine similarities, positive samples on diagonal entry

### **Zero-Shot Transfer**

pretrained text_encoder: labels → text features for all label →

pretrained image_encoder: image → image feature → cosine similarity → softmax

### **Prompt Engineering and Ensembling**

PS. a good reseach topic in fine-tuning / reasoning stage that doesn't require too much computational resources)

Issues:

- polysemy: multiple meanings of the same word
- in pre-training dataset, text = full sentence → in classification task, single word

  → A photo of a {label}. accuracy +1.3%

  → A bright photo of a {label}, a type of pet. 80 templates

**Representation Learning (**using all data in downstream tasks)

Common methods:

- linear probe: fix the model, fit a linear classifier on a representation extracted from the model and measure its performance
- fine tune: measure the performance of end-to-end fine-tuning of the model

**Limitation**

Poor on

- fine-grained classification such as differentiating models of cars, species of flowers
- more abstract and systematic tasks such as counting the number of objects in an image.
- novel tasks which are unlikely to be included in CLIP's pre-training dataset, such as classifying

the distance to the nearest car in a photo

- data that is truly out-of-distribution for it, only 88% accuracy on MNIST

Repeatedly query performance on full validation sets to guide the development of CLIP → not truly zero-shot transfer