https://arxiv.org/pdf/2103.00020.pdf

Most CV systems are trained to predict a fixed set of predetermined object categories -> limits their generality and usability bc additional labeled data is needed to specify any other visual concept

Learning from raw text about images (predicting which caption goes with which image)

- Pretrain learn SOTA image representations from scratch on a dataset of 400 million (image, text) pairs

Natural language used to reference learned visual concepts (or describe new ones) enabling zero shot transfer to downstream tasks

Benchmark on 30 existing CV datasets from OCR to action recognition to object classification

- Transfers non trivially to most tasks and competitive with fully supervised baseline without dataset specific training

NLP pretraining (task agnostic objectives): autoregressive and masked language modeling

- Task agnostic architectures to zero shot transfer to downstream datasets, removing need for specialized output heads or dataset specific customization
- Aggregate supervision accessible to modern pre-training methods within web-scale collections of text surpasses that of high quality crowd labeled NLP datasets
    - Computer vision currently doesn't have this (ImageNet)

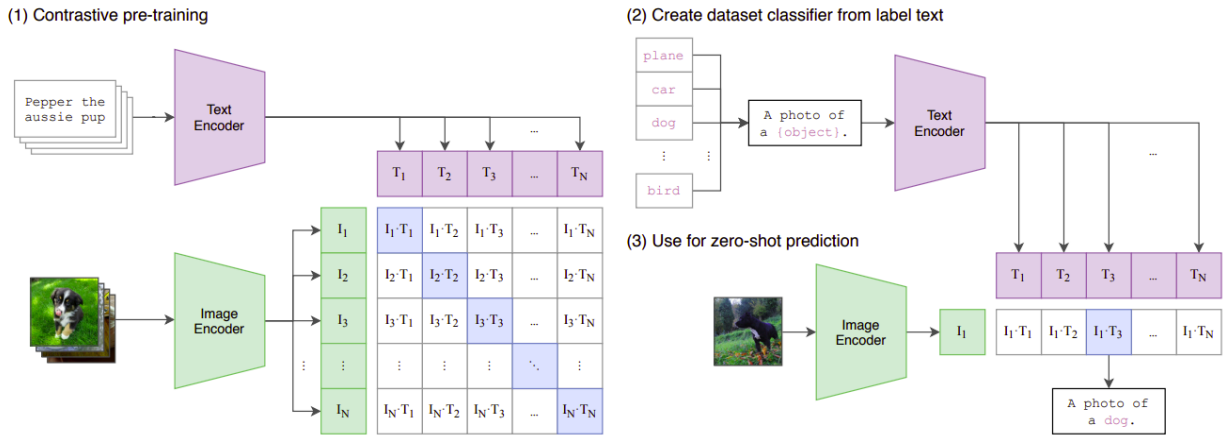

Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the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

Current approaches use a static softmax classifier for prediction and lack a mechanism for dynamic outputs -> limits flexibility and zero shot capabilities

Middle ground between learning from a limited amount of supervised "gold labels" and learning from unlimited amounts of raw text

- Natural language is able to express and supervise much wider set of visual concepts

Scale is important

- Propose simplified version of ConVIRT -> CLIP (contrastive language image pretraining) [language supervision for images]

Transfer performance is smoothly predictable function of compute (eight models spanning 2 orders of magnitude)

Linear probe representation learning analysis: linear classifier taking layer activations as input and measuring discrimination of networks (logistic regression vs a nonlinear layer)

CLIP outperforms best available ImageNet while being more computationally efficient

- Zero shot CLIP are more robust than equivalent accuracy supervised ImageNet
- Zero shot evaluation of task agnostic models is more representative of a model's capability

Much easier to scale natural language supervision compared to standard crowd sourced labeling for image classification (no gold label)

- Connects learned representation to language which enables flexible zero shot transfer

OpenAI is really good at dataset generation

Contrastive objectives can learn better representations than their equivalent predictive objective

- Generative models need a magnitude more compute than contrastive models

Train a system to solve the easier proxy task of predicting only which text as a whole is paired with which image and not the exact words of that text

CLIP predicts which N x N possible (image, text) pairings across a batch actually occur

- Learn multi modal embedding space by jointly training image encoder and text encoder to maximize cosine similarity of image and text embeddings of real pairs (N) while minimizing cosine similarity of N^2 - N incorrect pairings
- "Multi class N pair loss", infoNCE loss

Studying zero shot transfer as a way of measuring task- learning capabilities

- Dataset evaluates performance on a task on a specific distribution, not entire distribution

For each dataset, name of classes in dataset are used as set of potential text pairings and predict most probable pair

- Interpret image encoder as computer vision backbone (compute feature representation for image
- Text encoder hypernetwork which generates weights of linear classifier based on text specifying visual concepts that classes represent

- Every step of CLIP pre training is optimizing performance of randomly created proxy to a CV dataset which contains 1 example per class and 32,768 defined via natural language descriptions
- Zero shot: cache zero shot classifier once it has been computed by text encoder and reuse for all subsequent predictions

Scaling resnet in both width and depth (resnet 50 performed best)
“Prompt engineering” helps

CLIP’s features outperform features of best ImageNet model on a wide variety of datasets

- More robust to task/ distribution shift when compared to models pre-trained on ImageNet