Learning Transferable Visual Models From Natural Language Supervision

Minwoo Jung
mai-lab.net, *Medical Artificial Intelligence Laboratory*
Electrical and Electronic Engineering
Yonsei University

Medical Artificial
Intelligence Laboratory
At Yonsei University

Visual Language Model

CLIP

PMLR

2021

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

Abstract

SOTA computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since additional labeled data is needed to specify any other visual concept. Learning directly from raw text about images is a promising alternative which leverages a much broader source of supervision. We demonstrate that the simple pre-training task of predicting which caption goes with which image is an efficient and scalable way to learn SOTA image representations from scratch on a dataset of 400 million (image, text) pairs collected from the internet. After pre-training, natural language is used to reference learned visual concepts (or describe new ones) enabling zero-shot transfer of the model to downstream tasks. We study performance on over 30 different computer vision datasets, spanning tasks such as OCR, action recognition in videos, geo-localization, and many types of fine-grained object classification. The model transfers non-trivially to most tasks and is often competitive with a fully supervised baseline without the need for any dataset specific training. For instance, we match the accuracy of the original ResNet50 on ImageNet zero-shot without needing to use any of the 1.28 million training examples it was trained on. We release our code and pre-trained model weights at https://github.com/OpenAI/CLIP.

interface (McCann et al., 2018; Radford et al., 2019; Raffel et al., 2019) has enabled task-agnostic architectures to zero-shot transfer to downstream datasets. Flagship systems like GPT-3 (Brown et al., 2020) are now competitive across many tasks with bespoke models while requiring little to no dataset specific training data.

These results suggest that the aggregate supervision accessible to modern pre-training methods within web-scale collections of text surpasses that of high-quality crowd-labeled NLP datasets. However, in other fields such as computer vision it is still standard practice to pre-train models on crowd-labeled datasets such as ImageNet (Deng et al., 2009). Could scalable pre-training methods which learn directly from web text result in a similar breakthrough in computer vision? Prior work is encouraging.

Joulin et al. (2016) demonstrated that CNNs trained to predict words in image captions can learn representations competitive with ImageNet training. Li et al. (2017) then extended this approach to predicting phrase n-grams in addition to individual words and demonstrated the ability of their system to zero-shot transfer to other image classification datasets. Adopting more recent architectures and pre-training approaches, VirTex (Desai & Johnson, 2020), ICMLM (Bulent Sariyildiz et al., 2020), and ConVIRT (Zhang et al., 2020) have recently demonstrated the potential of transformer-based language modeling, masked language modeling, and contrastive objectives to learn image representations from text.

However, the aforementioned models still under-perform

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.

Visual Language Model

CLIP

Background & Goal

Background

- SOTA Computer Vision systems are trained to predict a fixed set of predetermined object categories

→ This restricted form of super-vision limits their generality and usability

Goal

- Aim to bring about groundbreaking progress in the field of computer vision through scalable pre-training methods that learn directly from web text

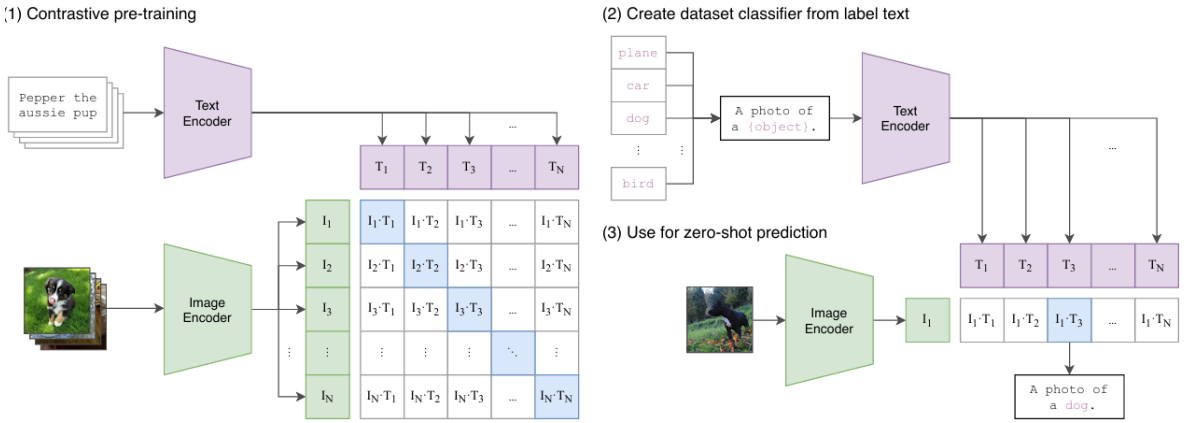

Fig 1. CLIP Overview

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.

Visual Language Model

CLIP

Approach

Natural Language Supervision

- Learning perception from supervision contained in natural language
- It's much easier to scale natural language supervision & also connects that representation to language

Creating a Sufficiently Large Dataset

- A major motivation for natural language supervision is the large quantities of data
- Constructed a new dataset of 400 million (image, text) pairs

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.

Visual Language Model

CLIP

Approach

Selecting an Efficient Pre-training Model

- Learns a multi-modal embedding space to maximize the cosine similarity of image & text embeddings of the N real pairs & minimizing the cosine similarity of the embedding of the $N^2 - N$ incorrect pairings
- Optimize a symmetric cross entropy loss over these similarity scores

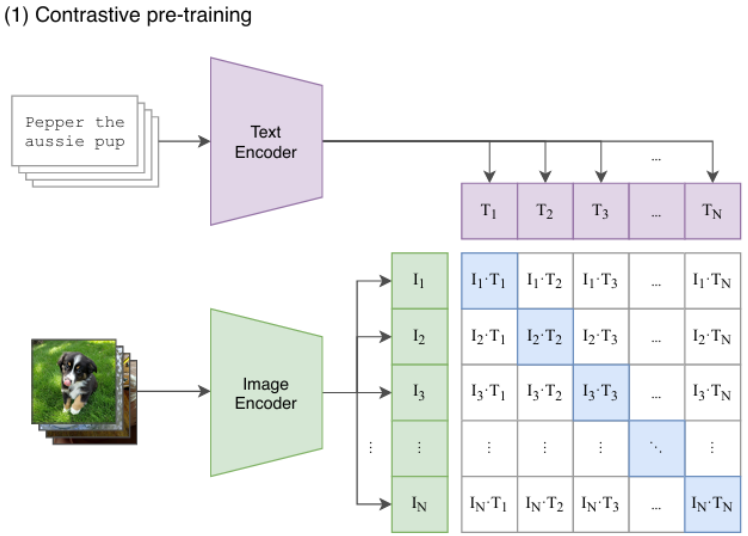

Fig 2. Overview of learning multi-modal embedding space

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.

Visual Language Model

CLIP

Approach

Choosing a Model

- Image Encoder: Use ResNet-D model with rect-2 blur pooling and replace the global average pooling layer with an attention pooling mechanism & ViT with layer normalization
- Text Encoder: Uses a Transformer that operates on the lower-cased BPE representation of the text

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.

Visual Language Model

CLIP

Experiments

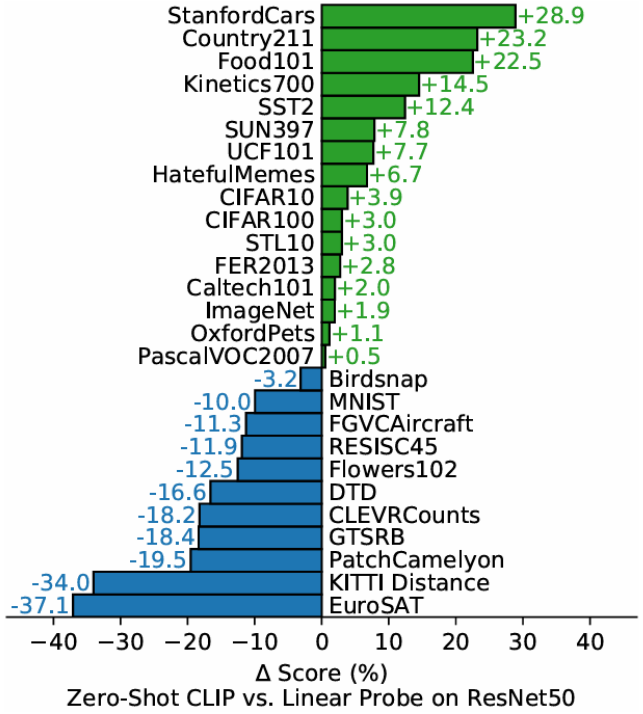

Fig 3. Zero-shot CLIP is competitive with a fully supervised baseline

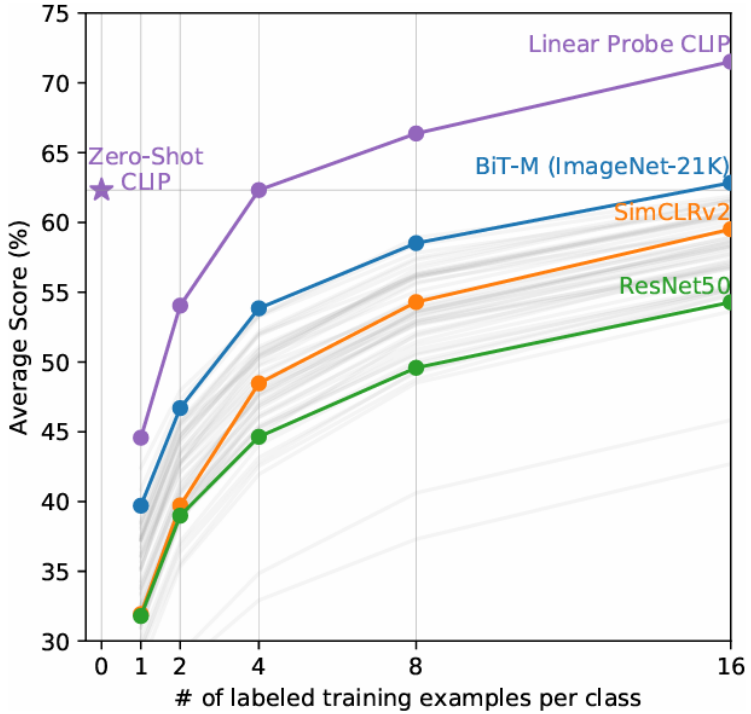

Fig 4. Zero-shot CLIP outperforms few-shot linear probes

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.

Visual Language Model

CLIP

Experiments

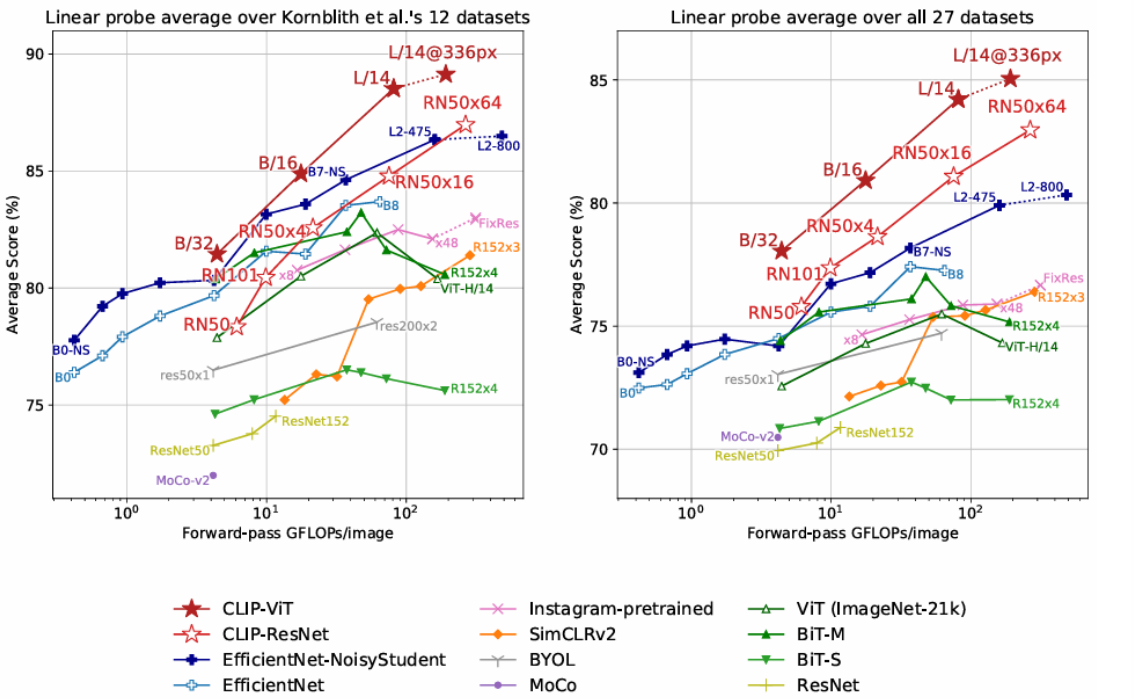

Fig 5. Linear probe performance of CLIP models in comparison with SOTA models

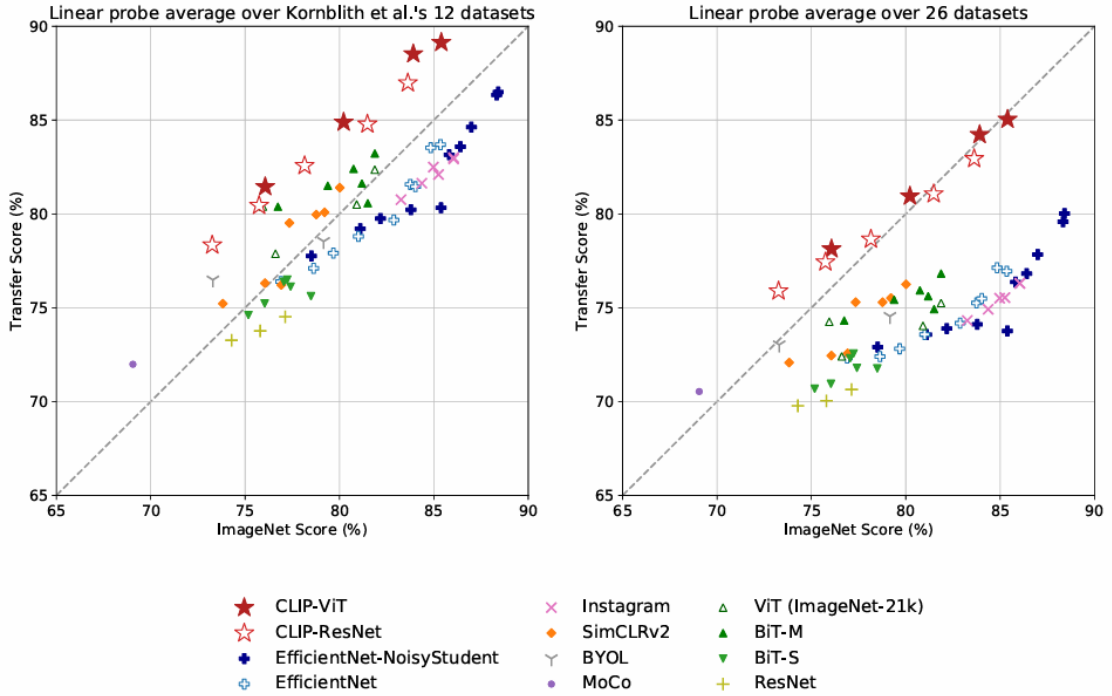

Fig 6. CLIP's features are more robust to task shift when compared to other models

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.

Visual Language Model

CLIP

Implications & Limitations

Implications

- CLIP has a wide range of capabilities due to its ability to carry out arbitrary image classification tasks

Limitations

- CLIP's zero-shot performance is still quite weak on several kinds of tasks
- For tasks which are unlikely to be included in CLIP's pre-training dataset, performance can be near random
- Image-text pairs are unfiltered and uncurated and result in CLIP models learning many social biases

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.