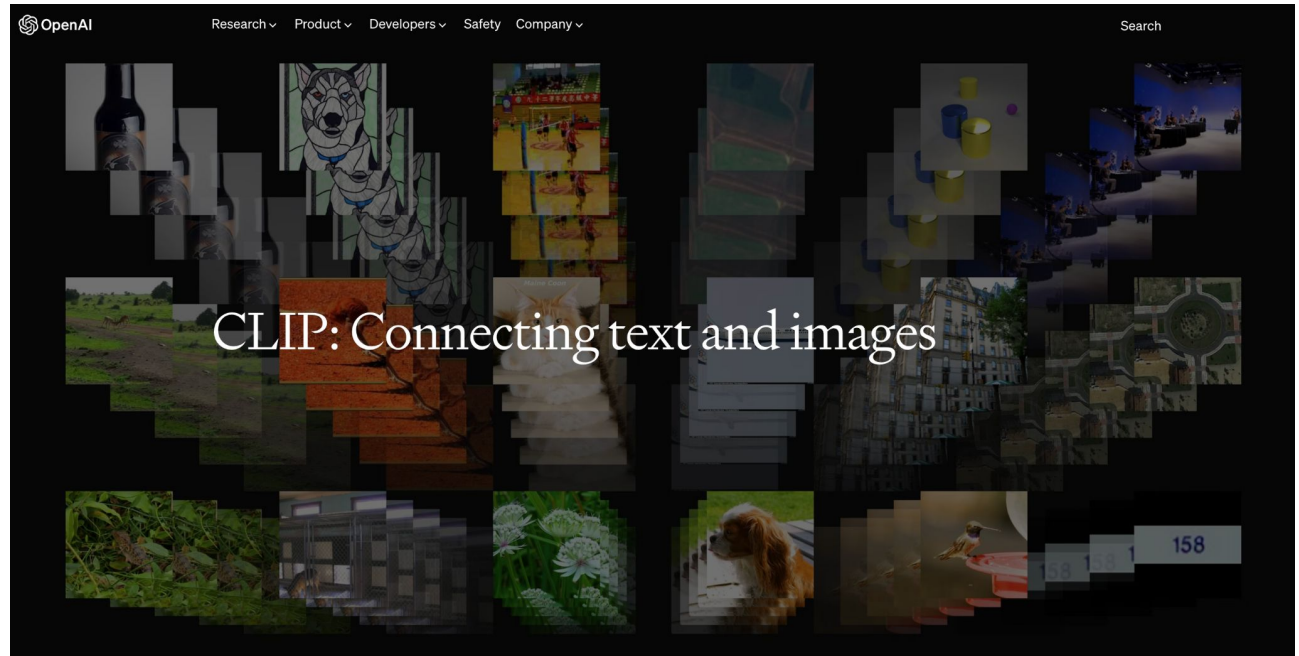




Deep Learning

CLIP

Contrastive Language-Image Pre-Training



Contrastive Language-Image Pre-Training

arXiv:2103.00020v1 [cs.CV] 26 Feb 2021

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford^{*1} Jong Wook Kim^{*1} Chris Hallacy¹ Aditya Ramesh¹ Gabriel Goh¹ Sandhini Agarwal¹
Girish Sastry¹ Amanda Askell¹ Pamela Mishkin¹ Jack Clark¹ Gretchen Krueger¹ Ilya Sutskever¹

Abstract

State-of-the-art 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 the performance of this approach by benchmarking on over 30 different existing 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 ResNet-50 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>.

1. Introduction and Motivating Work

Pre-training methods which learn directly from raw text have revolutionized NLP over the last few years (Dai & Le, 2015; Peters et al., 2018; Howard & Ruder, 2018; Radford et al., 2018; Devlin et al., 2018; Raffel et al., 2019).

^{*}Equal contribution ¹OpenAI, San Francisco, CA 94110, USA.
Correspondence to: <[alec, jongwook]@openai.com>.

Task-agnostic objectives such as autoregressive and masked language modeling have scaled across many orders of magnitude in compute, model capacity, and data, steadily improving capabilities. The development of “text-to-text” as a standardized input-output 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 removing the need for specialized output heads or dataset specific customization. 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.

Over 20 years ago Mori et al. (1999) explored improving content based image retrieval by training a model to predict the nouns and adjectives in text documents paired with images. Quattoni et al. (2007) demonstrated it was possible to learn more data efficient image representations via manifold learning in the weight space of classifiers trained to predict words in captions associated with images. Srivastava & Salakhutdinov (2012) explored deep representation learning by training multimodal Deep Boltzmann Machines on top of low-level image and text tag features. Joulin et al. (2016) modernized this line of work and demonstrated that CNNs trained to predict words in image captions learn useful image representations. They converted the title, description, and hashtag metadata of images in the YFCC100M dataset (Thomee et al., 2016) into a bag-of-words multi-label classification task and showed that pre-training AlexNet (Krizhevsky et al., 2012) to predict these labels learned representations which preformed similarly to ImageNet-based pre-training on transfer tasks. 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

Contributions

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**.



Contributions

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**.



Contributions

We study the performance of this approach by benchmarking on over **30 different existing CV datasets**, spanning tasks such as OCR, action recognition in videos, geo-localization, and many types of fine-grained object classification.

airplane



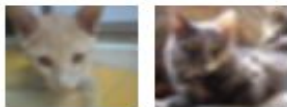
automobile



bird



cat



deer



Traditional learning paradigm in CV

- Create a training set of pairs:
(image, label)
- Retrain the model for new categories
- Learn a different model for each task

what worked

What we also tried...

Webly supervised learning

Learning Object Categories from Google's Image Search

R. Fergus¹

L. Fei-Fei²

P. Perona²

A. Zisserman¹

¹Dept. of Engineering Science
University of Oxford
Parks Road, Oxford
OX1 3PJ, U.K.
{fergus.az}@robots.ox.ac.uk

²Dept. of Electrical Engineering
California Institute of Technology
MC 136-93, Pasadena
CA 91125, U.S.A.
{feifei,perona}@vision.caltech.edu

Abstract

Current approaches to object category recognition require datasets of training images to be manually prepared, with varying degrees of supervision. We present an approach that can learn an object category from just its name, by utilizing the raw output of image search engines available on the Internet. We develop a new model, TSI_pLSA, which extends pLSA (as applied to visual words) to include spatial information in a translation and scale invariant manner. Our approach can handle the high intra-class variability and large proportion of unrelated images returned by search engines. We evaluate the models on standard test sets, showing performance competitive with existing methods trained on hand prepared datasets.

1. Introduction

The recognition of object categories is a challenging problem within computer vision. The current paradigm [1, 2, 5, 10, 14, 15, 21, 22, 24] consists of manually collecting a large training set of good exemplars of the desired object category; training a classifier on them and then evaluating it on novel images, possibly of a more challenging nature. The assumption is that training is a hard task that only needs to be performed once, hence the allocation of human resources to collecting a training set is justifiable. However, a constraint to current progress is the effort in obtaining large enough training sets of all the objects we wish to recognize. This effort varies with the size of the training set required, and the level of supervision required for each image. Examples range from 50 images (with segmentation) [15], through hundreds (with no segmentation) [10], to thousands of images [14, 23].

In this paper we propose a different perspective on the problem. There is a plentiful supply of images available at the typing of a single word using Internet image search engines such as Google, and we propose to learn visual models directly from this source. However, as can be seen in Fig. 1, this is not a source of pure training images: as many

as 85% of the returned images may be visually unrelated to the intended category, perhaps arising from polysemes (e.g. "iris" can be iris-flower, iris-eye, Iris-Murdoch). Even the 15% subset which do correspond to the category are substantially more demanding than images in typical training sets [9] – the number of objects in each image is unknown and variable, and the pose (visual aspect) and scale are uncontrolled. However, if one can succeed in learning from such noisy contaminated data the reward is tremendous: it enables us to automatically learn a classifier for whatever visual category we wish. In our previous work we have considered this source of images for training [11], but only for the purpose of re-ranking the images returned by the Google search (so that the category of interest has a higher rank than the noise) since the classifier models learnt were too weak to be used in a more general setting, away from the dataset collected for a given keyword.

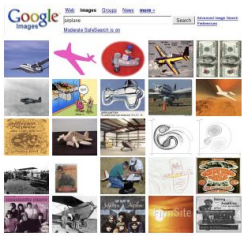
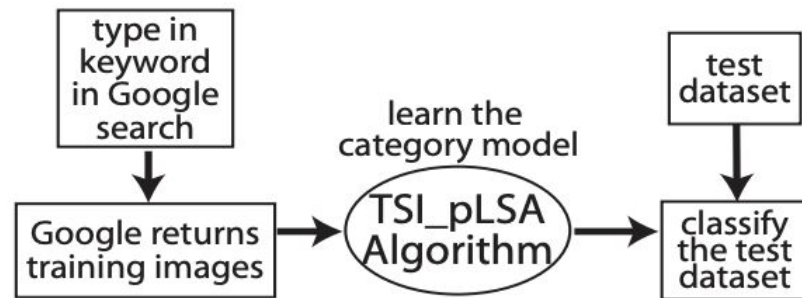


Figure 1: Images returned by Google's image search using the keyword "airplane". This is a representative sample of our training data. Note the large proportion of visually unrelated images and the wide pose variation.



ICCV, 2005

What we also tried...

Webly supervised learning

HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips

Antoine Miech^{1,2*} Dimitri Zhukov^{1,2*} Jean-Baptiste Alayrac^{2*}
Makarand Tapaswi² Ivan Laptev^{1,2} Josef Sivic^{1,2,3}
¹Ecole Normale Supérieure ²Inria ³CHRC, CTU
<https://www.di.ens.fr/willow/research/howto100m>

Abstract

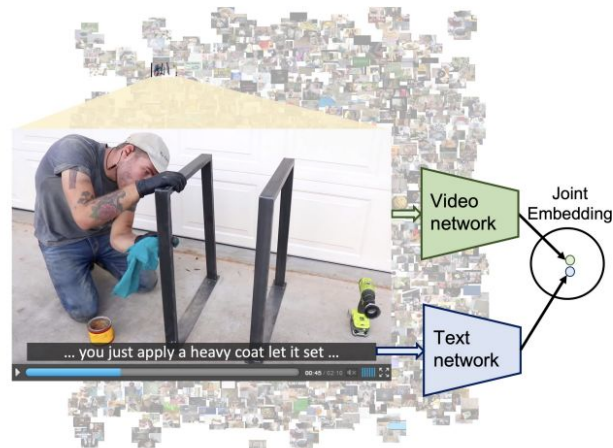
Learning text-video embeddings usually requires a dataset of video clips with manually provided captions. However, such datasets are expensive and time consuming to create and therefore difficult to obtain on a large scale. In this work, we propose instead to learn such embeddings from video data with readily available natural language annotations in the form of automatically transcribed narrations. The contributions of this work are three-fold. First, we introduce HowTo100M: a large-scale dataset of 136 million video clips sourced from 1.22M narrated instructional web videos depicting humans performing and describing over 23k different visual tasks. Our data collection procedure is fast, scalable and does not require any additional manual annotation. Second, we demonstrate that a text-video embedding trained on this data leads to state-of-the-art results for text-to-video retrieval and action localization on instructional video datasets such as YouCook2 or CrossTask. Finally, we show that this embedding transfers well to other domains: fine-tuning on generic Youtube videos (MSR-VTT dataset) and movies (LSMDC dataset) outperforms models trained on these datasets alone. Our dataset, code and models are publicly available [1].



Figure 1: We learn a joint text-video embedding by watching millions of narrated video clips of people performing diverse visual tasks. The learned embedding transfers well to other instructional and non-instructional text-video datasets.

tends naturally to artificial agents that need to understand the visual world and communicate about it with people. Examples of tasks that still represent a significant challenge for current artificial systems include text-to-video retrieval [25, 32, 54, 55, 63], text-based action or event localization [15], video captioning [36, 61], and video question answering [51, 63]. Yet, progress on these problems is important for a host of applications from searching video archives to human-robot communication.

A common approach to model visual concepts described with language is to learn a mapping of text and video into a shared embedding space, where related text fragments and video clips are close to each other [15, 32, 37, 38, 59]. Learning a good representation often requires a large set of paired video clips and text captions. In fact, given the huge variability of video scenes and their textual descriptions, learning a generic embedding space may require millions of paired video clips and text captions. However, existing datasets (e.g. MSR-VTT [58], DiDeMo [15], EPIC-



arXiv:1906.03327v2 [cs.CV] 31 Jul 2019

1. Introduction

Communicating about the visual world using language is a key ability of humans as intelligent beings. A three year old child can manipulate objects, observe its own actions and describe them to others using language; while adults can learn new skills by reading books or watching videos. This interplay between video and language ex-

*Equal contribution.

*Now at DeepMind.

¹Département d'informatique de l'ENS, École normale supérieure, CNRS, PSL Research University, 75005 Paris, France.

²Czech Institute of Informatics, Robotics and Cybernetics at the Czech Technical University in Prague.

ICCV, 2019

What we also tried...

Learning from descriptions

VirTex: Learning Visual Representations from Textual Annotations

Karan Desai Justin Johnson
University of Michigan
(kdesai, justincj)@umich.edu

Abstract

The de-facto approach to many vision tasks is to start from pretrained visual representations, typically learned via supervised training on ImageNet. Recent methods have explored unsupervised pretraining to scale to vast quantities of unlabeled images. In contrast, we aim to learn high-quality visual representations from fewer images. To this end we revisit supervised pretraining, and seek data-efficient alternatives to classification-based pretraining. We propose VirTex – a pretraining approach using semantically dense captions to learn visual representations. We train convolutional networks from scratch on COCO Captions, and transfer them to downstream recognition tasks including image classification, object detection, and instance segmentation. On all tasks, VirTex yields features that match or exceed those learned on ImageNet – supervised or unsupervised – despite using up to ten times fewer images.

1. Introduction

The prevailing paradigm for learning visual representations is first to *pretrain* a convolutional network [1, 2] to perform image classification on ImageNet [3, 4], then *transfer* the learned features to downstream tasks [5, 6]. This approach has been wildly successful, and has led to significant advances on a wide variety of computer vision problems such as object detection [7], semantic [8] and instance [9] segmentation, image captioning [10–12], and visual question answering [13, 14]. Despite its practical success, this approach is expensive to scale since the pretraining step relies on images annotated by human workers.

For this reason, there has been increasing interest in *unsupervised pretraining* methods that use unlabeled images to learn visual representations which are then transferred to downstream tasks [15–21]. Some recent approaches have begun to match or exceed supervised pretraining on ImageNet [22–26], and have been scaled to hundreds of millions [22, 25, 27, 28] or billions [24] of images.

Continuing to scale unsupervised pretraining to ever-larger sets of unlabeled images is an important scientific

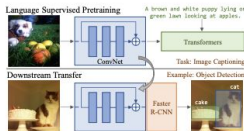


Figure 1. Learning visual features from language: First, we jointly train a ConvNet and Transformers using image-caption pairs, for the task of image captioning (top). Then, we transfer the learned ConvNet to several downstream vision tasks, for example object detection (bottom).

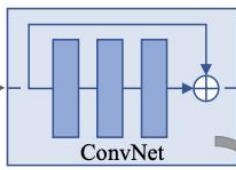
goal. But we may also ask whether there are alternate ways of pretraining that learn high-quality visual representations with fewer images. To do so, we revisit supervised pretraining and seek an alternative to traditional classification pretraining that uses each image more efficiently.

In this paper we present an approach for learning Visual representations from Textual annotations (VirTex). Our approach is straightforward: first, we jointly train a ConvNet and Transformer [29] from scratch to generate natural language captions for images. Then, we transfer the learned features to downstream visual recognition tasks (Figure 1).

We believe that using language supervision is appealing due to its *semantic density*. Figure 2 compares different pretraining tasks for learning visual representations. Captions provide a semantically denser learning signal than unsupervised contrastive methods and supervised classification. Hence, we expect that using textual features to learn visual features may require fewer images than other approaches.

Another benefit of textual annotations is simplified data collection. To collect classification labels, typically human experts first build an ontology of categories [3, 4, 30, 31], then complex crowdsourcing pipelines are used to elicit labels from non-expert users [32, 33]. In contrast, natural language descriptions do not require an explicit ontology and can easily be written by non-expert workers, leading to a

Language Supervised Pretraining

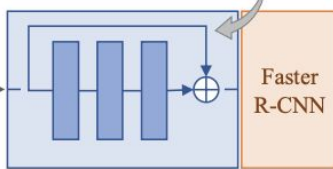


A brown and white puppy lying on green lawn looking at apples.

Transformers

Task: Image Captioning

Downstream Transfer

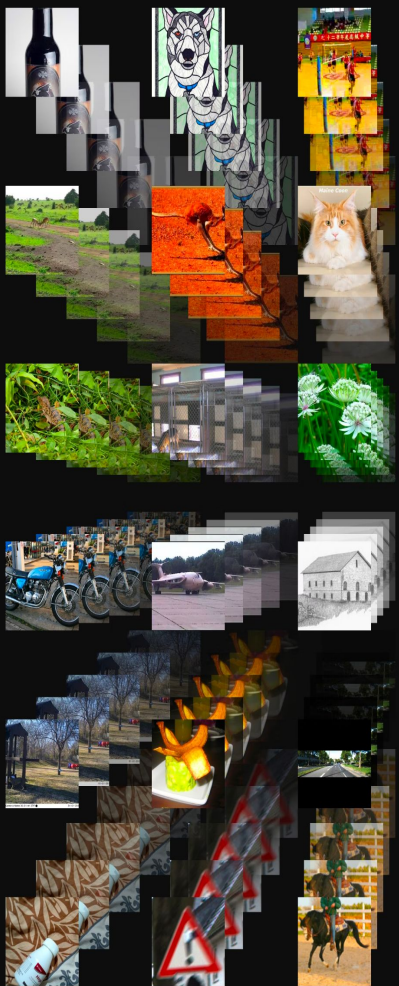


Example: Object Detection





Back to CLIP...

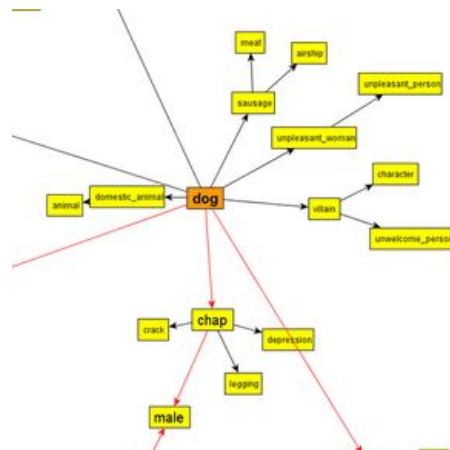


Dataset: 400M image-text pairings

- Minimally-curated web-scraped data
- Quality and appropriateness of the captions not guaranteed
- Manually checking captions is financially prohibitive
- Reflects biases inherent in the data

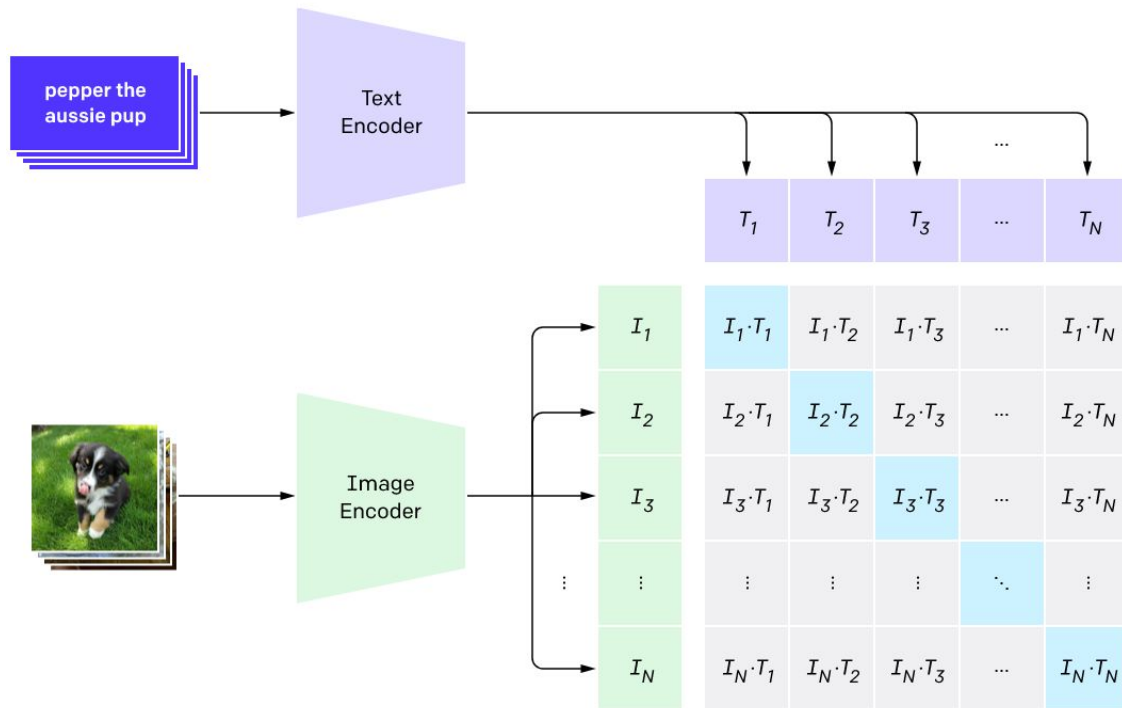
WebImageText (WIT)

- Searching the web with 500K queries.
- Queries:
 - ◆ words occurring at least 100 times in English Wikipedia
 - ◆ bi-grams (with high MI) to augment the initial queries
 - ◆ names of wikipedia articles above a search volume threshold
 - ◆ WordNet synsets
- Approximate class balancing: include up to 20K (image, text pairs) per query.



01. Contrastive Pretraining

Pre-train an image encoder and a text encoder to predict which images were paired with which texts in the dataset.



01. Contrastive Pretraining

CLIP trains an image and text encoders to maximise cosine similarities of the valid pairs within each batch and minimises those of invalid pairings.

```
# 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
```

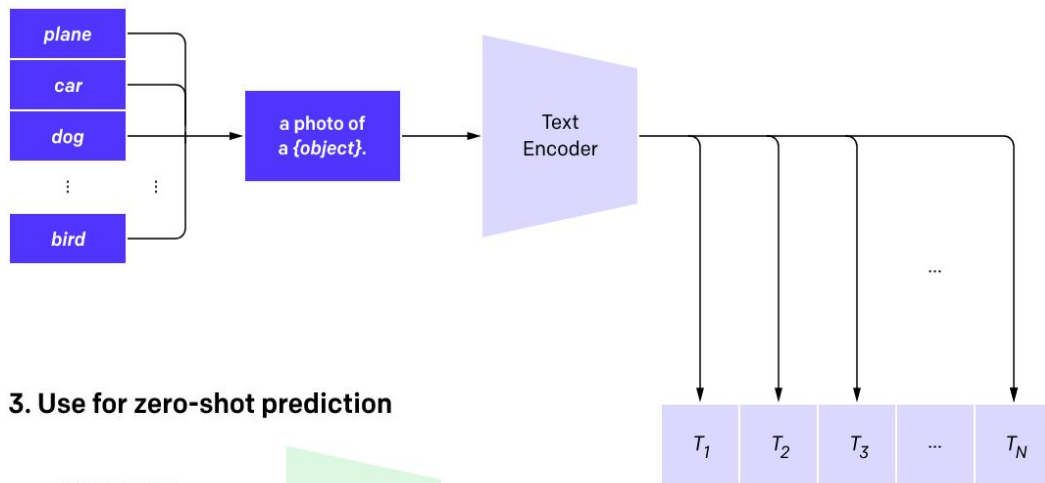
Figure 3. Numpy-like pseudocode for the core of an implementation of CLIP.

知乎 @千佛山彭子晏

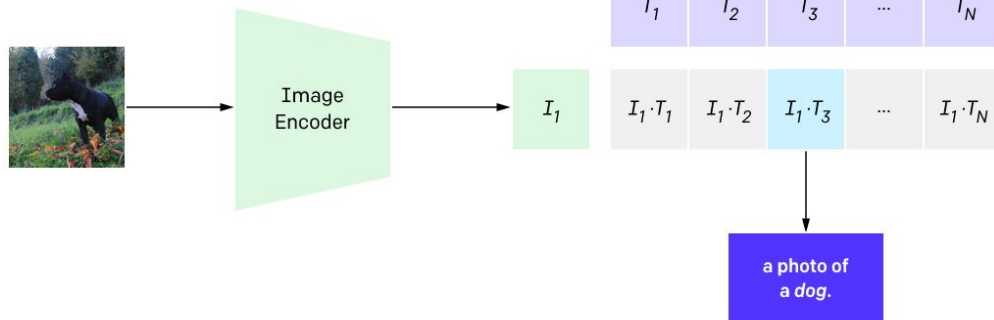
02. Inference

Prompting and zero-shot prediction

2. Create dataset classifier from label text



3. Use for zero-shot prediction



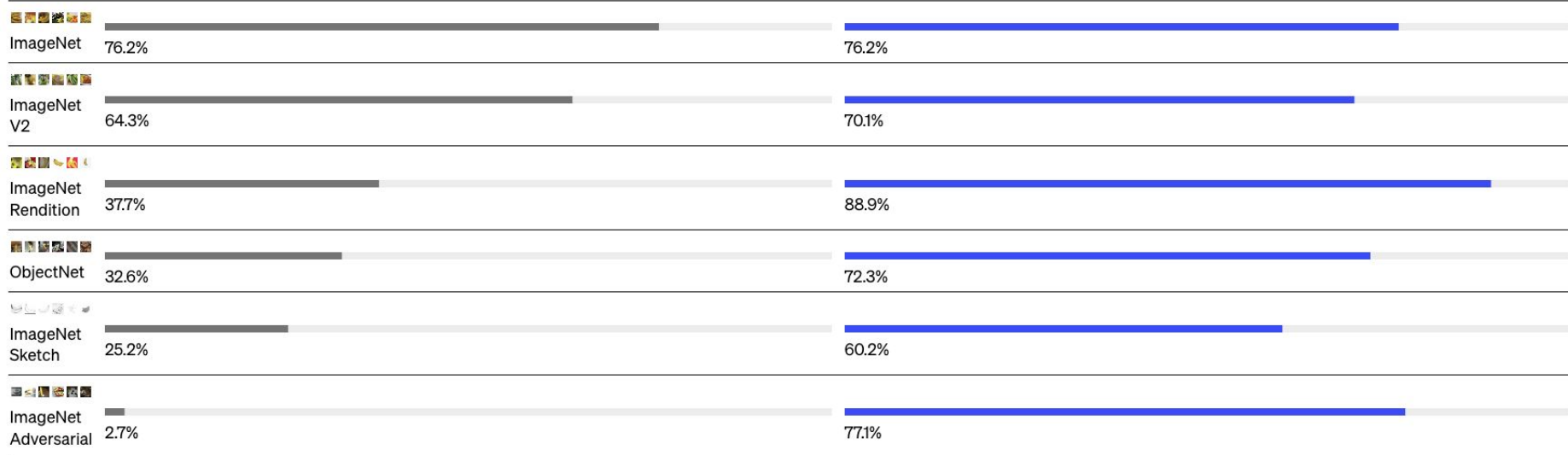


Check the paper for implementation details...

Results

Dataset ImageNet ResNet101

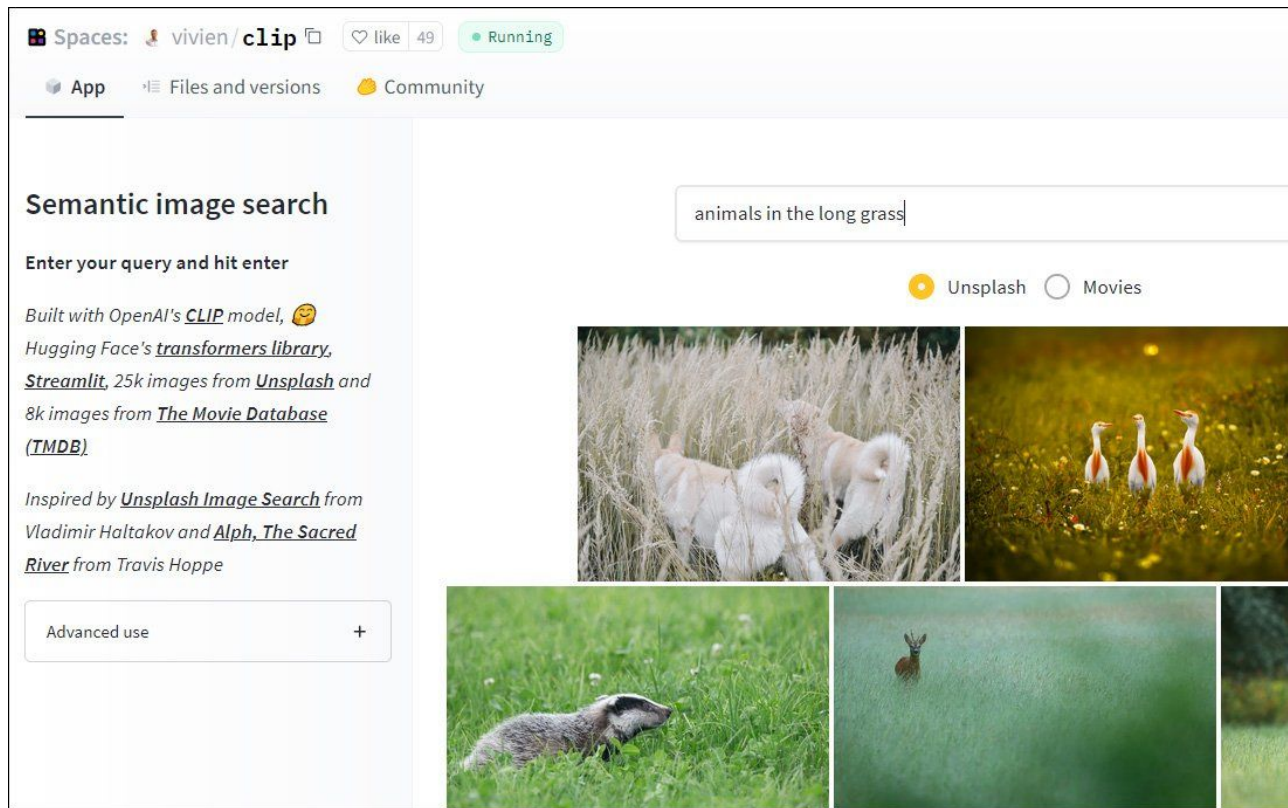
CLIP ViT-L



Although both models have the same accuracy on the ImageNet test set, CLIP's performance is much more representative of how it will fare on datasets that measure accuracy in different, non-ImageNet settings. For instance, ObjectNet checks a model's ability to recognize objects in many different poses and with many different backgrounds inside homes while ImageNet Rendition and ImageNet Sketch check a model's ability to recognize more abstract depictions of objects.

Use of CLIP

HuggingFace



Applications

Non-English languages, e.g. Chinese

ChineseCLIP

[ModelScope](#) | [Demo](#) | [Paper](#) | [Blog](#)

本项目为CLIP模型的中文版本，使用大规模中文数据进行训练（~2亿图文对），旨在帮助用户快速实现中文领域的[图文特征&相似度计算](#)、[跨模态检索](#)、[零样本图片分类](#)等任务。本项目代码基于[open_clip project](#)建设，并针对中文领域数据以及在中文数据上实现更好的效果做了优化。本项目提供了API、训练代码和测试代码，下文中将详细介绍细节。

新闻

- 2023.3.20 新增对比学习的[梯度累积](#)支持，可模拟更大batch size的训练效果
- 2023.2.16 新增[FlashAttention](#)支持，提升训练速度，降低显存占用，详见[flash_attention.md](#)
- 2023.1.15 新增部署[ONNX](#)和[TensorRT](#)模型支持（并提供预训练TensorRT模型），提升特征推理速度，满足部署需求，详见[deployment.md](#)
- 2022.12.12 新增实现[FLIP](#)训练策略，在finetune训练时可[激活使用](#)（感谢@zwkkk同学[贡献代码](#)❤️）
- 2022.12.3 公开[ELEVATER](#)图像分类数据集的中文版本，详见[数据文档](#)
- 2022.12.1 Chinese-CLIP模型代码&特征提取API，同步合入Huggingface transformers📦代码库
- 2022.11.22 新增[零样本图像分类](#)代码，可支持[ELEVATER benchmark](#)零样本分类评测任务
- 2022.11.3 新增RN50，ViT-H-14模型，公开[技术报告](#)
- 2022.9.22 新增ViT-L-14，ViT-L-14-336模型
- 2022.7.13 新增[图文特征提取快速API](#)，并行代码快速调用由中文CLIP模型计算图文特征&相似度

Applications

NeRF-based 3D model generation



night city with vaporwave aesthetic

Applications

Robotics: semantic localization

CLIP-Fields: Weakly Supervised Semantic Fields for Robotic Memory

Nur Muhammad (Mahi) Shafiullah^{1†} Chris Paxton² Lerrel Pinto¹ Soumith Chintala² Arthur Szlam²

Abstract—We propose CLIP-Fields, an implicit scene model that can be trained with no direct human supervision. This model learns a mapping from spatial locations to semantic embedding vectors. The mapping can then be used for a variety of tasks, such as segmentation, instance identification, semantic search over space, and view localization. Most importantly, the mapping can be trained with supervision coming only from web-image and web-text trained models such as CLIP, Detic, and Sentence-BERT. When compared to baselines like Mask-RCNN, our method outperforms on few-shot instance identification or semantic segmentation on the HM3D dataset with only a fraction of the examples. Finally, we show that using CLIP-Fields as a scene memory, robots can perform semantic navigation in real-world environments. Our code and demonstrations are available here: <https://mahis.life/clip-fields>

I. INTRODUCTION

Recently, a class of models for representing 3D scenes implicitly [1] has shown great promise as a tool for computer vision [2], [3]. These neural radiance fields (NeRFs), and implicit neural representations more generally [4], can serve as differentiable databases of spatio-temporal state that can be used by robots for scene understanding, SLAM, and planning [5]–[8].

Another line of recent work has shown that web-scale weakly-supervised vision-language models (e.g. CLIP [9]) capture powerful semantic abstractions. These have proven useful for a range of robotics applications, including object understanding [10] and multi-task learning from demonstration [11]. These applications have been limited, however, by the fact that these trained representations assume a single 2D image as input; it has been an open question how best to use these to enable 3D reasoning with all the advantages these vision-language models have to offer.

In this work, we introduce a method for building weakly supervised semantic neural fields, called CLIP-Fields. The key idea is to build a mapping from locations in space $g(x, y, z) : \mathbb{R}^3 \rightarrow \mathbb{R}^d$ that serves as a generic differentiable spatial database. The database is augmented with “modality” specific heads that interface g to off-the-shelf weakly-supervised language and vision models, which are used to train g and the heads. We assume that we have access to depth images of the scene of interest, and approximately, the corresponding 6D camera poses. From these, we train CLIP-Fields with a contrastive loss that penalizes mismatches



Fig. 1: Our approach, CLIP-Fields, integrates multiple views of a scene and can capture 3D semantics from relatively few examples. This results in a scalable 3D semantic representation that can be used to infer information about the world from relatively few examples and functions as a 3D spatial memory for a mobile robot.

between the vector output of the modality-specific head at the back-projected point in space corresponding to a pixel in an image, and the web-image-trained vectors corresponding to the location in the image; but encourages differences with vector representations of other images and regions of space.

Thus, from the point of view of a robot using CLIP-Fields as a spatial database for scene-understanding, training g itself can be entirely self-supervised – the full pipeline, including training the underlying image models, need not use any explicit supervision. On the other hand, as we will show in our experiments, the spatial database g can naturally incorporate scene-specific labels, if they are available.

We demonstrate our method quantitatively on instance segmentation and identification. Furthermore, we give qual-

¹ New York University

² FAIR Labs

[†] Corresponding author, email: mahi@cs.nyu.edu

Visual Grounding


The **dog** and the **boy**



Locate the most relevant region in an image
based on a natural language query



Mandatory Reading

- ★ Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Aspell, Pamela Mishkin, Jack Clark, et al. *Learning transferable visual models from natural language supervision*. In ICML 2021.
- 



Questions?