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

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). 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

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

从自然语言监督中学习可迁移的视觉模型

亚历克·拉德福德 [*1] 金钟旭 [*1] 克里斯·哈拉西 [1] 阿迪蒂亚·拉梅什 [1] 加布里埃尔·吴 [1] 桑迪尼·阿加瓦尔 [1] 吉里什·萨斯特里 [1] 阿曼达·阿斯克尔 [1] 帕梅拉·米什金 [1] 杰克·克拉克 [1] 格雷琴·克鲁格 [1] 伊利亚·苏茨克弗 [1]

摘要

最先进的计算机视觉系统经过训练，可以预测一组固定的预定对象类别。这种受限的监督形式限制了它们的通用性和可用性，因为需要额外的标注数据来指定任何其他视觉概念。直接从关于图像的原始文本中学习是一种有前景的替代方案，它利用了更广泛的监督来源。我们证明，预测哪个标题与哪个图像对应的简单预训练任务，是一种高效且可扩展的方法，可以从互联网收集的4亿个（图像，文本）对数据集中从头开始学习最先进的图像表示。预训练后，自然语言被用来引用学习到的视觉概念（或描述新的概念），使模型能够零样本迁移到下游任务。我们通过在30多个不同的现有计算机视觉数据集上进行基准测试来研究这种方法的性能，涵盖的任务包括OCR、视频中的动作识别、地理定位以及多种类型的细粒度对象分类。该模型在大多数任务上都有显著的迁移效果，并且通常与完全监督的基线模型竞争，无需任何特定数据集的训练。例如，我们在ImageNet上零样本匹配了原始ResNet-50的准确率，而无需使用其训练所依赖的128万个训练样本中的任何一个。我们在https://github.com/OpenAI/CLIP 发布了代码和预训练模型权重。

1. 引言与动机研究

直接从原始文本中学习的预训练方法在过去几年彻底改变了自然语言处理领域（Dai & Le, 2015; Peters et al., 2018; Howard & Ruder, 2018; Radford et al., 2018; Devlin et al., 2018; Raffel et al., 2019）。诸如自回归和掩码语言建模这类任务无关的目标，在计算量、模型容量和数据规模上跨越了多个数量级的扩展，持续提升了模型能力。“文本到文本”作为一种标准化输入输出接口的发展（McCann等人，2018；Radford等人，2019；Raffel等人，2019），使得任务无关的架构能够零样本迁移到下游数据集，无需专用输出头或针对特定数据集的定制化。像GPT-3（Brown等人，2020）这样的旗舰系统，如今在许多任务上与定制模型表现相当，同时几乎不需要特定数据集的训练数据。

这些结果表明，在现代预训练方法中，通过网络规模文本集合可获得的总体监督超过了高质量众包标注的NLP数据集。然而，在其他领域如计算机视觉中，使用ImageNet（Deng等人，2009）等众包标注数据集进行模型预训练仍是标准做法。直接从网络文本中学习的可扩展预训练方法是否能为计算机视觉带来类似突破？先前的研究令人鼓舞。

20多年前，Mori等人（1999）通过训练模型预测与图像配对的文本文档中的名词和形容词，探索了改进基于内容的图像检索方法。Quattoni等人（2007）证明，通过在用于预测图像相关标题中词汇的分类器权重空间中进行流形学习，可以获得数据效率更高的图像表示。Srivastava与Salakhutdinov（2012）通过在低层图像和文本标签特征上训练多模态深度玻尔兹曼机，探索了深度表示学习。Joulin等人（2016）将这一研究方向现代化，证明了通过训练CNN预测图像标题中的词汇可以学习到有效的图像表示。他们将YFCC100M数据集（Thomee等人，2016）中图像的标题、描述和标签元数据转化为词袋多标签分类任务，并表明通过预训练AlexNet（Krizhevsky等人，2012）来预测这些标签所学习的表示，在迁移任务上的表现与基于ImageNet的预训练效果相当。随后Li等人（2017）将此方法扩展到预测短语n-gram及独立词汇，并展示了其系统向其他图像任务的零样本迁移能力。

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

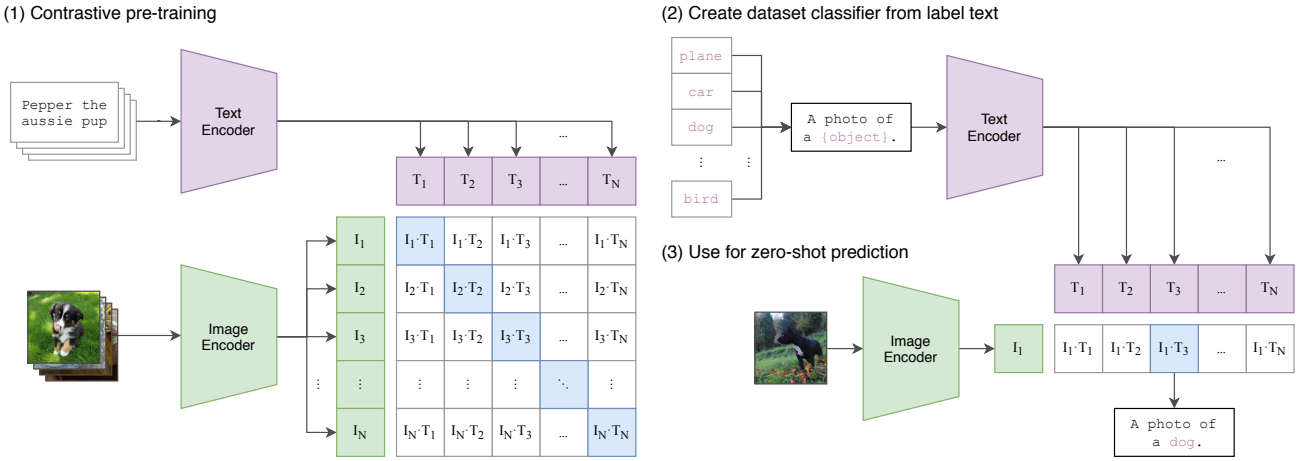

Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the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

classification datasets by scoring target classes based on their dictionary of learned visual n-grams and predicting the one with the highest score. 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.

While exciting as proofs of concept, using natural language supervision for image representation learning is still rare. This is likely because demonstrated performance on common benchmarks is much lower than alternative approaches. For example, Li et al. (2017) reach only 11.5% accuracy on ImageNet in a zero-shot setting. This is well below the 88.4% accuracy of the current state of the art (Xie et al., 2020). It is even below the 50% accuracy of classic computer vision approaches (Deng et al., 2012). Instead, more narrowly scoped but well-targeted uses of weak supervision have improved performance. Mahajan et al. (2018) showed that predicting ImageNet-related hashtags on Instagram images is an effective pre-training task. When fine-tuned to ImageNet these pre-trained models increased accuracy by over 5% and improved the overall state of the art at the time. Kolesnikov et al. (2019) and Dosovitskiy et al. (2020) have also demonstrated large gains on a broader set of transfer benchmarks by pre-training models to predict the classes of the noisily labeled JFT-300M dataset.

This line of work represents the current pragmatic middle ground between learning from a limited amount of supervised "gold-labels" and learning from practically unlimited amounts of raw text. However, it is not without compromises. Both works carefully design, and in the process limit, their supervision to 1000 and 18291 classes respectively. Natural language is able to express, and therefore supervise, a much wider set of visual concepts through its generality. Both approaches also use static softmax classifiers to perform prediction and lack a mechanism for dynamic outputs. This severely curtails their flexibility and limits their "zero-shot" capabilities.

A crucial difference between these weakly supervised models and recent explorations of learning image representations directly from natural language is scale. While Mahajan et al. (2018) and Kolesnikov et al. (2019) trained their models for accelerator years on millions to billions of images, VirTex, ICMLM, and ConVIRT trained for accelerator days on one to two hundred thousand images. In this work, we close this gap and study the behaviors of image classifiers trained with natural language supervision at large scale. Enabled by the large amounts of publicly available data of this form on the internet, we create a new dataset of 400 million (image, text) pairs and demonstrate that a simplified version of ConVIRT trained from scratch, which we call CLIP, for Contrastive Language-Image Pre-training, is an efficient method of learning from natural language supervision. We study the scalability of CLIP by training a series of eight models spanning almost 2 orders of magnitude of compute and observe that transfer performance is a smoothly predictable function of compute (Hestness et al., 2017; Kaplan et al., 2020). We find that CLIP, similar to the GPT family, learns to perform a wide set of tasks during pre-training including OCR, geo-localization, action recognition, and many others. We measure this by benchmarking the zero-shot transfer performance of CLIP on over 30 existing datasets and find

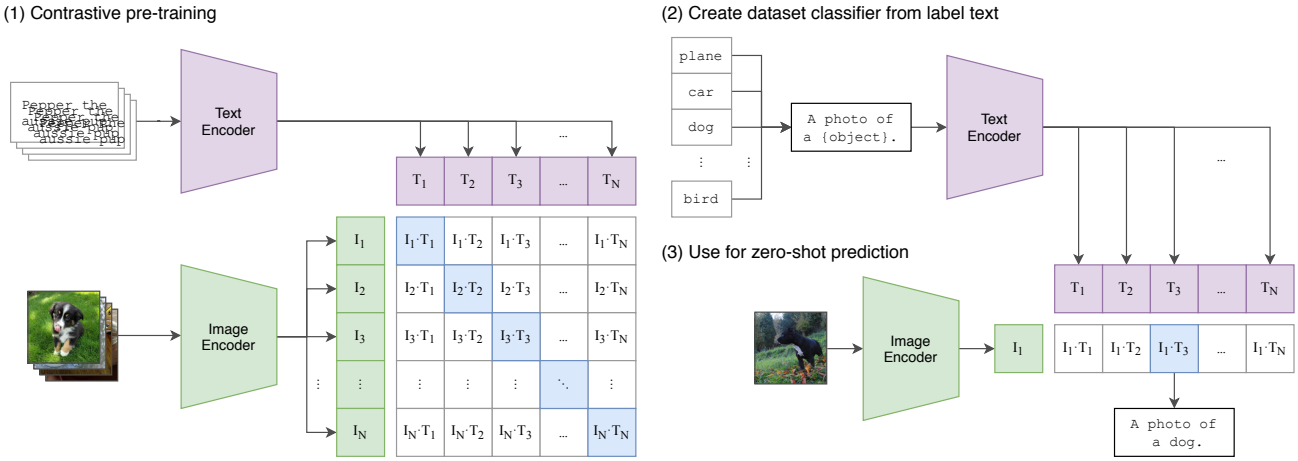

图1. 我们的方法概述。标准图像模型联合训练图像特征提取器和线性分类器以预测某些标签，而CLIP则联合训练图像编码器和文本编码器，以预测一批（图像，文本）训练样本的正确配对。在测试时，学习到的文本编码器通过嵌入目标数据集类别的名称或描述，合成一个零样本线性分类器。

通过基于已学习的视觉n-gram词典对目标类别进行评分，并预测得分最高的类别，我们在分类数据集上取得了进展。采用更近期的架构和预训练方法，VirTex（Desai & Johnson，2020）、ICMLM（Bulent Sariyildiz等人，2020）以及Con-VIRT（Zhang等人，2020）最近展示了基于Transformer的语言建模、掩码语言建模和对比目标在从文本中学习图像表征方面的潜力。

尽管作为概念验证令人兴奋，但利用自然语言监督进行图像表征学习仍属罕见。这很可能是因为在常见基准测试中已展示的性能远低于其他方法。例如，Li等人（2017）在ImageNet的零样本设置中仅达到11.5%的准确率，远低于当前最先进技术88.4%的准确率（Xie等人，2020），甚至不及经典计算机视觉方法50%的准确率（Deng等人，2012）。相比之下，范围更窄但目标明确的弱监督应用反而提升了性能。Mahajan等人（2018）表明，预测Instagram图像上与ImageNet相关的标签是一种有效的预训练任务。当对这些预训练模型进行ImageNet微调时，准确率提升了超过5%，并推动了当时整体技术水平的进步。Kolesnikov等人（2019）和Dosovitskiy等人（2020）也通过预训练模型预测带噪声标签的JFT-300M数据集类别，在更广泛的迁移基准测试中实现了显著提升。

这项工作代表了当前在从有限的有监督“黄金标签”学习和从几乎无限的原始文本学习之间的实用中间地带。然而，这并非没有妥协——

米塞斯。这两项工作都精心设计，并在过程中将监督类别分别限制在1000和18291类。自然语言凭借其通用性，能够表达并监督更广泛的视觉概念。两种方法还都使用静态softmax分类器进行预测，缺乏动态输出的机制。这严重限制了它们的灵活性，并制约了其“零样本”能力。

这些弱监督模型与近期直接从自然语言学习图像表征的探索之间存在一个关键差异：规模。Mahajan等人（2018）和Kolesnikov等人（2019）在数百万至数十亿张图像上进行了加速器年量级的训练，而VirTex、ICMLM和ConVIRT仅在十万至二十万张图像上进行了加速器天量级的训练。在本研究中，我们弥合了这一差距，并系统探究了在大规模自然语言监督下训练的图像分类器的行为特征。借助互联网上公开的海量此类数据，我们构建了一个包含4亿个（图像，文本）对的新数据集，并证明ConVIRT的简化版本——我们称之为CLIP（对比性语言-图像预训练）——是一种高效的自然语言监督学习方法。我们通过训练八个计算量跨越近两个数量级的模型系列，研究了CLIP的可扩展性，发现其迁移性能与计算量呈平滑可预测的函数关系（Hestness等人，2017；Kaplan等人，2020）。与GPT系列类似，我们发现CLIP在预训练过程中能学习执行广泛任务，包括光学字符识别、地理定位、动作识别等。我们通过在30多个现有数据集上对CLIP进行零样本迁移性能基准测试来衡量这一能力，并发现

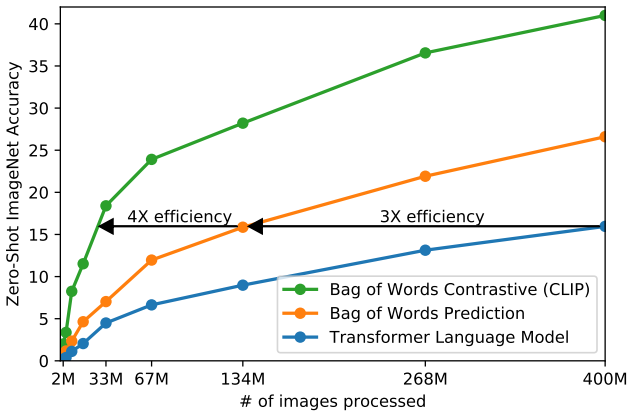

Figure 2. **CLIP is much more efficient at zero-shot transfer than our image caption baseline.** Although highly expressive, we found that transformer-based language models are relatively weak at zero-shot ImageNet classification. Here, we see that it learns 3x slower than a baseline which predicts a bag-of-words (BoW) encoding of the text (Joulin et al., 2016). Swapping the prediction objective for the contrastive objective of CLIP further improves efficiency another 4x.

it can be competitive with prior task-specific supervised models. We also confirm these findings with linear-probe representation learning analysis and show that CLIP outperforms the best publicly available ImageNet model while also being more computationally efficient. We additionally find that zero-shot CLIP models are much more robust than equivalent accuracy supervised ImageNet models which suggests that zero-shot evaluation of task-agnostic models is much more representative of a model's capability. These results have significant policy and ethical implications, which we consider in Section 7.

2. Approach

2.1. Natural Language Supervision

At the core of our approach is the idea of learning perception from supervision contained in natural language. As discussed in the introduction, this is not at all a new idea, however terminology used to describe work in this space is varied, even seemingly contradictory, and stated motivations are diverse. Zhang et al. (2020), Gomez et al. (2017), Joulin et al. (2016), and Desai & Johnson (2020) all introduce methods which learn visual representations from text paired with images but describe their approaches as unsupervised, self-supervised, weakly supervised, and supervised respectively.

We emphasize that what is common across this line of work is not any of the details of the particular methods used but the appreciation of natural language as a training signal. All these approaches are learning from *natural language supervision*. Although early work wrestled with the complexity of natural language when using topic model and n-gram representations, improvements in deep contextual representation learning suggest we now have the tools to effectively leverage this abundant source of supervision (McCann et al., 2017).

Learning from natural language has several potential strengths over other training methods. It's much easier to scale natural language supervision compared to standard crowd-sourced labeling for image classification since it does not require annotations to be in a classic "machine learning compatible format" such as the canonical 1-of-N majority vote "gold label". Instead, methods which work on natural language can learn passively from the supervision contained in the vast amount of text on the internet. Learning from natural language also has an important advantage over most unsupervised or self-supervised learning approaches in that it doesn't "just" learn a representation but also connects that representation to language which enables flexible zero-shot transfer. In the following subsections, we detail the specific approach we settled on.

2.2. Creating a Sufficiently Large Dataset

Existing work has mainly used three datasets, MS-COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017), and YFCC100M (Thomee et al., 2016). While MS-COCO and Visual Genome are high quality crowd-labeled datasets, they are small by modern standards with approximately 100,000 training photos each. By comparison, other computer vision systems are trained on up to 3.5 *billion* Instagram photos (Mahajan et al., 2018). YFCC100M, at 100 million photos, is a possible alternative, but the metadata for each image is sparse and of varying quality. Many images use automatically generated filenames like `20160716_113957.JPG` as "titles" or contain "descriptions" of camera exposure settings. After filtering to keep only images with natural language titles and/or descriptions in English, the dataset shrunk by a factor of 6 to only 15 million photos. This is approximately the same size as ImageNet.

A major motivation for natural language supervision is the large quantities of data of this form available publicly on the internet. Since existing datasets do not adequately reflect this possibility, considering results only on them would underestimate the potential of this line of research. To address this, we constructed a new dataset of 400 million (image, text) pairs collected form a variety of publicly available sources on the Internet. To attempt to cover as broad a set of visual concepts as possible, we search for (image, text) pairs as part of the construction process whose text includes one of a set of 500,000 queries.[1] We approximately class

[1]The base query list is all words occurring at least 100 times in the English version of Wikipedia. This is augmented with bi-grams

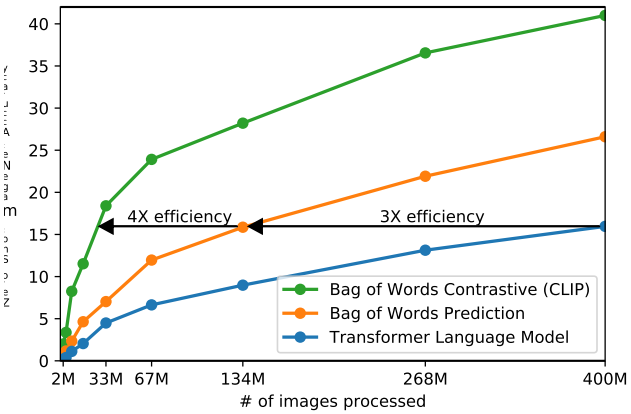

图2. CLIP在零样本迁移方面比我们的图像描述基线高效得多。尽管表达能力很强，但我们发现基于Transformer的语言模型在零样本ImageNet分类方面相对较弱。在此我们看到，其学习速度比预测文本词袋（BoW）编码的基线（Joulin等人，2016）慢3倍。将预测目标替换为CLIP的对比目标后，效率进一步提升至4倍。

它可以与先前针对特定任务的监督模型相竞争。我们还通过线性探针表示学习分析证实了这些发现，并表明CLIP在计算效率更高的同时，其表现也优于目前公开可用的最佳ImageNet模型。此外，我们发现零样本CLIP模型比同等准确度的监督式ImageNet模型具有更强的鲁棒性，这表明对任务无关模型进行零样本评估更能代表模型的实际能力。这些结果具有重要的政策与伦理影响，我们将在第7节对此进行探讨。

2. 方法

2.1. 自然语言监督

我们方法的核心思想是从自然语言所含的监督信息中学习感知能力。正如引言中所讨论的，这并非全新概念，然而描述这一领域研究时使用的术语却多种多样，甚至看似矛盾，且其动机阐述也各不相同。Zhang等人（2020）、Gomez等人（2017）、Joulin等人（2016）以及Desai与Johnson（2020）都提出了从图文配对数据中学习视觉表征的方法，却分别将其方法描述为无监督、自监督、弱监督和有监督学习。

我们强调，这一系列工作的共同点不在于所采用具体方法的细节，而在于对自然语言作为训练信号的重视。所有这些方法都从*natural language super-vision*尽管早期研究在使用主题模型和n-gram表示时需应对自然语言的复杂性，但深度上下文表征学习的进展表明，我们现在已拥有有效利用这一丰富监督源的工具（McCann等人，2017年）。

从自然语言中学习相比其他训练方法具有若干潜在优势。与图像分类的标准众包标注相比，自然语言监督的扩展性要强得多，因为它不需要将标注转换为经典的“机器学习兼容格式”，例如规范的N选一多数投票“黄金标签”。相反，基于自然语言的方法可以从互联网海量文本中包含的监督信息中进行被动学习。与大多数无监督或自监督学习方法相比，从自然语言中学习还有一个重要优势：它不仅学习表征，还将该表征与语言联系起来，从而实现灵活的零样本迁移。在接下来的小节中，我们将详细说明最终采用的具体方法。

2.2. 创建足够大的数据集

现有研究主要使用了三个数据集：MS-COCO（Lin等人，2014）、Visual Genome（Krishna等人，2017）和YFCC100M（Thomee等人，2016）。尽管MS-COCO和Visual Genome是高质量的人工标注数据集，但以现代标准衡量规模较小，各自仅包含约10万张训练照片。相比之下，其他计算机视觉系统训练时使用的照片数量高达35亿张*billion* Instagram照片（Mahajan等人，2018）。拥有1亿张照片的YFCC100M虽是一个潜在替代方案，但其每张图像的元数据稀疏且质量参差不齐。许多图像使用自动生成的文件名（如20160716_113957.JPG）作为“标题”，或仅包含相机曝光参数的“描述”。经过筛选仅保留含英语自然语言标题和/或描述的图像后，数据集规模缩减至原来的1/6，仅剩1500万张照片，这与ImageNet的规模大致相当。

自然语言监督的一个主要动机是互联网上公开可用的大量此类数据。由于现有数据集未能充分反映这种可能性，仅基于它们来评估结果会低估这一研究方向的实际潜力。为此，我们构建了一个包含4亿个（图像，文本）对的新数据集，这些数据来自互联网上各种公开可用的资源。为了尽可能覆盖广泛的视觉概念，我们在构建过程中搜索了文本包含50万个查询词集中至少一个词汇的（图像，文本）对。[1] 我们大致分类

[1]The base query list is all words occurring at least 100 times in the English version of Wikipedia. This is augmented with bi-grams

balance the results by including up to 20,000 (image, text) pairs per query. The resulting dataset has a similar total word count as the WebText dataset used to train GPT-2. We refer to this dataset as WIT for WebImageText.

2.3. Selecting an Efficient Pre-Training Method

State-of-the-art computer vision systems use very large amounts of compute. Mahajan et al. (2018) required 19 GPU years to train their ResNeXt101-32x48d and Xie et al. (2020) required 33 TPUv3 core-years to train their Noisy Student EfficientNet-L2. When considering that both these systems were trained to predict only 1000 ImageNet classes, the task of learning an open set of visual concepts from natural language seems daunting. In the course of our efforts, we found training efficiency was key to successfully scaling natural language supervision and we selected our final pre-training method based on this metric.

Our initial approach, similar to VirTex, jointly trained an image CNN and text transformer from scratch to predict the caption of an image. However, we encountered difficulties efficiently scaling this method. In Figure 2 we show that a 63 million parameter transformer language model, which already uses twice the compute of its ResNet-50 image encoder, learns to recognize ImageNet classes three times slower than a much simpler baseline that predicts a bag-of-words encoding of the same text.

Both these approaches share a key similarity. They try to predict the *exact* words of the text accompanying each image. This is a difficult task due to the wide variety of descriptions, comments, and related text that co-occur with images. Recent work in contrastive representation learning for images has found that contrastive objectives can learn better representations than their equivalent predictive objective (Tian et al., 2019). Other work has found that although generative models of images can learn high quality image representations, they require over an order of magnitude more compute than contrastive models with the same performance (Chen et al., 2020a). Noting these findings, we explored training a system to solve the potentially easier proxy task of predicting only which text *as a whole* is paired with which image and not the exact words of that text. Starting with the same bag-of-words encoding baseline, we swapped the predictive objective for a contrastive objective in Figure 2 and observed a further 4x efficiency improvement in the rate of zero-shot transfer to ImageNet.

Given a batch of N (image, text) pairs, CLIP is trained to predict which of the $N \times N$ possible (image, text) pairings across a batch actually occurred. To do this, CLIP learns a multi-modal embedding space by jointly training an image encoder and text encoder to maximize the cosine similarity of the image and text embeddings of the N real pairs in the batch while minimizing the cosine similarity of the embeddings of the $N^2 - N$ incorrect pairings. We optimize a symmetric cross entropy loss over these similarity scores. In Figure 3 we include pseudocode of the core of an implementation of CLIP. To our knowledge this batch construction technique and objective was first introduced in the area of deep metric learning as the *multi-class N-pair loss* Sohn (2016), was popularized for contrastive representation learning by Oord et al. (2018) as the InfoNCE loss, and was recently adapted for contrastive (text, image) representation learning in the domain of medical imaging by Zhang et al. (2020).

Due to the large size of our pre-training dataset, over-fitting is not a major concern and the details of training CLIP are simplified compared to the implementation of Zhang et al. (2020). We train CLIP from scratch without initializing the image encoder with ImageNet weights or the text encoder with pre-trained weights. We do not use the non-linear projection between the representation and the contrastive embedding space, a change which was introduced by Bachman et al. (2019) and popularized by Chen et al. (2020b). We instead use only a linear projection to map from each encoder's representation to the multi-modal embedding space. We did not notice a difference in training efficiency between the two versions and speculate that non-linear projections may be co-adapted with details of current image only in self-supervised representation learning methods. We also remove the text transformation function t_u from Zhang et al. (2020) which samples a single sentence at uniform from the text since many of the (image, text) pairs in CLIP's pre-training dataset are only a single sentence. We also simplify the image transformation function t_v. A random square crop from resized images is the only data augmentation used during training. Finally, the temperature parameter which controls the range of the logits in the softmax, τ, is directly optimized during training as a log-parameterized multiplicative scalar to avoid turning as a hyper-parameter.

2.4. Choosing and Scaling a Model

We consider two different architectures for the image encoder. For the first, we use ResNet-50 (He et al., 2016a) as the base architecture for the image encoder due to its widespread adoption and proven performance. We make several modifications to the original version using the ResNet-D improvements from He et al. (2019) and the antialiased rect-2 blur pooling from Zhang (2019). We also replace the global average pooling layer with an attention pooling mechanism. The attention pooling is implemented as a single layer of "transformer-style" multi-head QKV attention where the query is conditioned on the global average-pooled

with high pointwise mutual information as well as the names of all Wikipedia articles above a certain search volume. Finally all WordNet synsets not already in the query list are added.

通过为每个查询包含最多20,000个（图像，文本）对来平衡结果。最终数据集的总词数与用于训练GPT-2的WebText数据集相近。我们将此数据集称为WIT（WebImageText）。

2.3. 选择高效的预训练方法

最先进的计算机视觉系统需要大量的计算资源。Mahajan等人（2018）训练其ResNeXt101-32x48d模型耗费了19个GPU年，而Xie等人（2020）训练Noisy Student EfficientNet-L2模型则消耗了33个TPUv3核心年。考虑到这两个系统仅训练用于预测1000个ImageNet类别，从自然语言中学习开放视觉概念的任务显得尤为艰巨。在我们的研究过程中，我们发现训练效率是成功扩展自然语言监督的关键，并基于这一指标选择了最终的预训练方法。

我们最初的方法与VirTex类似，从头开始联合训练图像CNN和文本变换器，以预测图像的标题。然而，我们在高效扩展该方法时遇到了困难。在图2中，我们展示了一个拥有6300万参数的变换器语言模型，其计算量已经是ResNet-50图像编码器的两倍，但学习识别ImageNet类别的速度比一个简单得多的基线模型慢三倍，该基线模型预测的是同一文本的词袋编码。

这两种方法有一个关键的共同点：它们都试图预测每张图片所附文本中的*exact*个单词。由于与图片同时出现的描述、评论及相关文本种类繁多，这是一项艰巨的任务。最近在图像对比表征学习方面的研究发现，对比目标能够比同等的预测目标学习到更好的表征（Tian等人，2019）。另有研究指出，尽管图像的生成模型可以学习到高质量的表征，但达到相同性能时，其所需的计算量比对比模型高出一个数量级以上（Chen等人，2020a）。基于这些发现，我们尝试训练一个系统来解决一个可能更简单的代理任务：仅预测哪段文本*as a whole*与哪张图像配对，而不预测该文本的具体用词。我们从相同的词袋编码基线出发，在图2中将预测目标替换为对比目标，结果观察到在ImageNet上的零样本迁移效率进一步提升了4倍。

给定一批N (图像、文本)对，CLIP的训练目标是预测在一个批次中$N \times N$种可能的（图像，文本）配对里哪些真实存在。为此，CLIP学习一个通过联合训练图像编码器和文本编码器，构建多模态嵌入空间，旨在最大化批次中N真实配对图像与文本嵌入的余弦相似度，同时最小化$N^2 - N$错误配对嵌入的余弦相似度。我们基于这些相似度分数优化对称交叉熵损失。图3提供了CLIP实现核心部分的伪代码。据我们所知，这种批次构建技术与目标函数最早由Sohn（2016）在深度度量学习领域提出，被称为*multi-class N-pair loss*；随后被Oord等人（2018）推广为对比表示学习中的InfoNCE损失；最近又被Zhang等人（2020）应用于医学影像领域的对比（文本、图像）表示学习。

由于我们的预训练数据集规模庞大，过拟合并非主要问题，因此CLIP的训练细节相较于Zhang等人（2020）的实现有所简化。我们从头开始训练CLIP，未使用ImageNet权重初始化图像编码器，也未使用预训练权重初始化文本编码器。我们未采用表示与对比嵌入空间之间的非线性投影（该技术由Bachman等人于2019年提出，并由Chen等人在2020b年推广），而仅使用线性投影将各编码器的表示映射到多模态嵌入空间。我们未观察到两个版本在训练效率上的差异，推测非线性投影可能仅在与当前图像细节共同适应时，在自监督表示学习方法中发挥作用。我们还移除了Zhang等人（2020）中的文本转换函数t_u（该函数从文本中均匀采样单个句子），因为CLIP预训练数据集中的许多（图像，文本）对仅包含单个句子。同时简化了图像转换函数t_v：训练期间使用的唯一数据增强方法是从调整尺寸后的图像中随机裁剪正方形区域。最后，用于控制softmax中逻辑值范围的温度参数τ在训练期间直接作为对数参数化的乘法标量进行优化，以避免将其作为超参数进行调整。

2.4. 模型选择与缩放

我们为图像编码器考虑了两种不同的架构。首先，由于ResNet-50（He等人，2016a）的广泛采用和已验证的性能，我们将其作为图像编码器的基础架构。我们对原始版本进行了几项改进：采用了He等人（2019）的ResNet-D优化方案，以及Zhang（2019）的抗锯齿rect-2模糊池化技术。此外，我们将全局平均池化层替换为注意力池化机制。该注意力池化通过单层“Transformer风格”的多头QKV注意力实现，其中查询向量以全局平均池化的特征为条件。

---

with high pointwise mutual information as well as the names of all Wikipedia articles above a certain search volume. Finally all WordNet synsets not already in the query list are added.

```
# 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.

representation of the image. For the second architecture, we experiment with the recently introduced Vision Transformer (ViT) (Dosovitskiy et al., 2020). We closely follow their implementation with only the minor modification of adding an additional layer normalization to the combined patch and position embeddings before the transformer and use a slightly different initialization scheme.

The text encoder is a Transformer (Vaswani et al., 2017) with the architecture modifications described in Radford et al. (2019). As a base size we use a 63M-parameter 12-layer 512-wide model with 8 attention heads. The transformer operates on a lower-cased byte pair encoding (BPE) representation of the text with a 49,152 vocab size (Sennrich et al., 2015). For computational efficiency, the max sequence length was capped at 76. The text sequence is bracketed with `[SOS]` and `[EOS]` tokens and the activations of the highest layer of the transformer at the `[EOS]` token are treated as the feature representation of the text which is layer normalized and then linearly projected into the multi-modal embedding space. Masked self-attention was used in the text encoder to preserve the ability to initialize with a pre-trained language model or add language modeling as an auxiliary objective, though exploration of this is left as future work.

While previous computer vision research has often scaled models by increasing the width (Mahajan et al., 2018) or depth (He et al., 2016a) in isolation, for the ResNet image encoders we adapt the approach of Tan & Le (2019) which found that allocating additional compute across all of width, depth, and resolution outperforms only allocating it to only one dimension of the model. While Tan & Le (2019) tune the ratio of compute allocated to each dimension for their EfficientNet architecture, we use a simple baseline of allocating additional compute equally to increasing the width, depth, and resolution of the model. For the text encoder, we only scale the width of the model to be proportional to the calculated increase in width of the ResNet and do not scale the depth at all, as we found CLIP's performance to be less sensitive to the capacity of the text encoder.

2.5. Training

We train a series of 5 ResNets and 3 Vision Transformers. For the ResNets we train a ResNet-50, a ResNet-101, and then 3 more which follow EfficientNet-style model scaling and use approximately 4x, 16x, and 64x the compute of a ResNet-50. They are denoted as RN50x4, RN50x16, and RN50x64 respectively. For the Vision Transformers we train a ViT-B/32, a ViT-B/16, and a ViT-L/14. We train all models for 32 epochs. We use the Adam optimizer (Kingma & Ba, 2014) with decoupled weight decay regularization (Loshchilov & Hutter, 2017) applied to all weights that are not gains or biases, and decay the learning rate using a cosine schedule (Loshchilov & Hutter, 2016). Initial hyper-parameters were set using a combination of grid searches, random search, and manual tuning on the baseline ResNet-50 model when trained for 1 epoch. Hyper-parameters were then adapted heuristically for larger models due to computational constraints. The learnable temperature parameter τ was initialized to the equivalent of 0.07 from (Wu et al., 2018) and clipped to prevent scaling the logits by more than 100 which we found necessary to prevent training instability. We use a very large minibatch size of 32,768. Mixed-precision (Micikevicius et al., 2017) was used to accelerate training and save memory. To save additional memory, gradient checkpointing (Griewank & Walther, 2000; Chen et al., 2016), half-precision Adam statistics (Dhariwal et al., 2020), and half-precision stochastically rounded text encoder weights were used. The calculation of embedding similarities was also sharded with individual GPUs computing only the subset of the pairwise similarities necessary for their local batch of embeddings. The largest ResNet model, RN50x64, took 18 days to train on 592 V100 GPUs while the largest Vision Transformer took 12 days on 256 V100 GPUs. For the ViT-L/14 we also pre-train at a higher 336 pixel resolution for one additional epoch to boost performance similar to FixRes (Touvron et al., 2019). We denote this model as ViT-L/14@336px. Unless otherwise specified, all results reported in this paper as "CLIP" use this model which we found to perform best.

```
# 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
```

图3. CLIP实现核心部分的类Numpy伪代码。

图像的表示。对于第二种架构，我们尝试了最近引入的视觉变换器（ViT）（Dosovitskiy等人，2020）。我们严格遵循他们的实现，仅做了微小修改：在变换器之前对合并的补丁和位置嵌入添加了额外的层归一化，并采用了略有不同的初始化方案。

文本编码器是一个Transformer（Vaswani等人，2017），其架构修改遵循Radford等人（2019）的描述。基础规模采用6300万参数、12层、512宽度、8个注意力头的模型。该Transformer处理文本的小写字节对编码（BPE）表示，词汇表规模为49,152（Sennrich等人，2015）。为提升计算效率，最大序列长度限制为76。文本序列以[SOS]和[EOS]标记作为边界，Transformer最高层在[EOS]标记处的激活值被视为文本特征表示，该表示经过层归一化后线性投影到多模态嵌入空间。文本编码器中使用了掩码自注意力机制，以保留初始化预训练语言模型或将语言建模作为辅助目标的可能性，但相关探索留待未来工作。

尽管以往的计算机视觉研究通常通过单独增加宽度（Mahajan等人，2018）或深度（He等人，2016a）来扩展模型，但对于我们采用的ResNet图像编码器，我们借鉴了Tan & Le（2019）的方法，该方法发现将额外计算资源同时分配给宽度、深度和分辨率三个维度，比仅将其分配给单一维度表现更优。模型的一个维度。虽然Tan & Le (2019)为他们的EfficientNet架构调整了每个维度分配的计算比例，但我们采用了一个简单的基线，即均等地将额外计算资源分配给增加模型的宽度、深度和分辨率。对于文本编码器，我们仅按比例缩放模型宽度，使其与ResNet计算出的宽度增加成比例，完全不缩放深度，因为我们发现CLIP的性能对文本编码器容量的敏感度较低。

2.5. 训练

我们训练了5个ResNet系列模型和3个Vision Transformer模型。对于ResNet，我们训练了ResNet-50、ResNet-101，以及另外三个采用EfficientNet风格模型缩放方法、计算量分别约为ResNet-50的4倍、16倍和64倍的模型，它们分别被标记为RN50x4、RN50x16和RN50x64。对于Vision Transformer，我们训练了ViT-B/32、ViT-B/16和ViT-L/14。所有模型均训练32个周期。我们采用Adam优化器（Kingma & Ba, 2014），并对非增益或偏置的所有权重应用解耦权重衰减正则化（Loshchilov & Hutter, 2017），学习率通过余弦调度进行衰减（Loshchilov & Hutter, 2016）。初始超参数通过在基线ResNet-50模型上进行1周期训练时的网格搜索、随机搜索和手动调优组合设定。由于计算资源限制，较大模型的超参数采用启发式方法调整。可学习的温度参数τ初始化为相当于（Wu et al., 2018）中0.07的值，并进行截断处理以防止逻辑值缩放超过100倍——我们发现这对防止训练不稳定是必要的。我们使用了非常大的小批量尺寸32,768。采用混合精度训练（Micikevicius et al., 2017）以加速训练并节省内存。为节省额外内存，我们使用了梯度检查点技术（Griewank & Walther, 2000; Chen et al., 2016）、半精度Adam统计量（Dhariwal et al., 2020）以及半精度随机舍入的文本编码器权重。嵌入相似度的计算也进行了分片处理，每个GPU仅计算其本地嵌入批次所需的成对相似度子集。最大的ResNet模型RN50x64在592块V100 GPU上训练了18天，而最大的Vision Transformer模型在256块V100 GPU上训练了12天。对于ViT-L/14，我们还以更高的336像素分辨率进行了额外1个周期的预训练以提升性能，该方法类似FixRes（Touvron et al., 2019）。我们将该模型标记为ViT-L/14@336px。除非特别说明，本文中所有标注为“CLIP”的结果均使用该性能最优的模型。

3. Experiments

3.1. Zero-Shot Transfer

3.1.1. Motivation

In computer vision, zero-shot learning usually refers to the study of generalizing to unseen object categories in image classification (Lampert et al., 2009). We instead use the term in a broader sense and study generalization to unseen datasets. We motivate this as a proxy for performing unseen tasks, as aspired to in the zero-data learning paper of Larochelle et al. (2008). While much research in the field of unsupervised learning focuses on the *representation learning* capabilities of machine learning systems, we motivate studying zero-shot transfer as a way of measuring the *task-learning* capabilities of machine learning systems. In this view, a dataset evaluates performance on a task on a specific distribution. However, many popular computer vision datasets were created by the research community primarily as benchmarks to guide the development of generic image classification methods rather than measuring performance on a specific task. While it is reasonable to say that the SVHN dataset measures the task of street number transcription on the distribution of Google Street View photos, it is unclear what "real" task the CIFAR-10 dataset measures. It is clear, however, what distribution CIFAR-10 is drawn from - TinyImages (Torralba et al., 2008). On these kinds of datasets, zero-shot transfer is more an evaluation of CLIP's robustness to distribution shift and domain generalization rather than task generalization. Please see Section 3.3 for analysis focused on this.

To our knowledge, Visual N-Grams (Li et al., 2017) first studied zero-shot transfer to existing image classification datasets in the manner described above. It is also the only other work we are aware of that has studied zero-shot transfer to standard image classification datasets using a generically pre-trained model and serves as the best reference point for contextualizing CLIP. Their approach learns the parameters of a dictionary of 142,806 visual n-grams (spanning 1- to 5- grams) and optimizes these n-grams using a differential version of Jelinek-Mercer smoothing to maximize the probability of all text n-grams for a given image. In order to perform zero-shot transfer, they first convert the text of each of the dataset's class names into its n-gram representation and then compute its probability according to their model, predicting the one with the highest score.

Our focus on studying zero-shot transfer as an evaluation of task learning is inspired by work demonstrating task learning in the field of NLP. To our knowledge Liu et al. (2018) first identified task learning as an "unexpected side-effect" when a language model trained to generate Wikipedia articles learned to reliably transliterate names between languages. While GPT-1 (Radford et al., 2018) focused on pre-training as a transfer learning method to improve supervised fine-tuning, it also included an ablation study demonstrating that the performance of four heuristic zero-shot transfer methods improved steadily over the course of pre-training, without any supervised adaption. This analysis served as the basis for GPT-2 (Radford et al., 2019) which focused exclusively on studying the task-learning capabilities of language models via zero-shot transfer.

3.1.2. Using CLIP for Zero-Shot Transfer

CLIP is pre-trained to predict if an image and a text snippet are paired together in its dataset. To perform zero-shot classification, we reuse this capability. For each dataset, we use the names of all the classes in the dataset as the set of potential text pairings and predict the most probable (image, text) pair according to CLIP. In a bit more detail, we first compute the feature embedding of the image and the feature embedding of the set of possible texts by their respective encoders. The cosine similarity of these embeddings is then calculated, scaled by a temperature parameter τ, and normalized into a probability distribution via a softmax. Note that this prediction layer is a multinomial logistic regression classifier with L2-normalized inputs, L2-normalized weights, no bias, and temperature scaling. When interpreted this way, the image encoder is the computer vision backbone which computes a feature representation for the image and the text encoder is a hypernetwork (Ha et al., 2016) which generates the weights of a linear classifier based on the text specifying the visual concepts that the classes represent. Lei Ba et al. (2015) first introduced a zero-shot image classifier of this form while the idea of generating a classifier from natural language dates back to at least Elhoseiny et al. (2013). Continuing with this interpretation, every step of CLIP pre-training can be viewed as optimizing the performance of a randomly created proxy to a computer vision dataset which contains 1 example per class and has 32,768 total classes defined via natural language descriptions. For zero-shot evaluation, we cache the zero-shot classifier once it has been computed by the text encoder and reuse it for all subsequent predictions. This allows the cost of generating it to be amortized across all the predictions in a dataset.

3.1.3. Initial Comparison to Visual N-Grams

In Table 1 we compare Visual N-Grams to CLIP. The best CLIP model improves accuracy on ImageNet from a proof of concept 11.5% to 76.2% and matches the performance of the original ResNet-50 despite using none of the 1.28 million crowd-labeled training examples available for this dataset. Additionally, the top-5 accuracy of CLIP models are noticeably higher than their top-1, and this model has a 95% top-5 accuracy, matching Inception-V4 (Szegedy et al., 2016). The ability to match the performance of a strong, fully supervised baselines in a zero-shot setting suggests

3. 实验

3.1. 零样本迁移

3.1.1. 动机

在计算机视觉领域，零样本学习通常指图像分类中泛化到未见物体类别的研究（Lampert等，2009）。我们则以更广义的方式使用该术语，研究向未见数据集的泛化能力。我们将此视为执行未见任务的代理目标，正如Larochelle等（2008）在零数据学习论文中倡导的那样。虽然无监督学习领域的许多研究聚焦于机器学习系统的*representation learn-ing*能力，但我们主张将零样本迁移研究作为衡量机器学习系统*task-learning*能力的一种方式。从这个视角看，数据集评估的是特定分布上某项任务的性能。然而，许多流行的计算机视觉数据集由研究社区创建时，主要是作为指导通用图像分类方法开发的基准，而非衡量特定任务的性能。虽然可以合理地说SVHN数据集衡量的是谷歌街景照片分布上的街景门牌号转录任务，但CIFAR-10数据集衡量的“真实”任务却并不明确。不过CIFAR-10的分布来源是清晰的——它源自TinyImages图像库（Torralba等，2008）。对于这类数据集，零样本迁移更多是评估CLIP对分布偏移和领域泛化的鲁棒性，而非任务泛化能力。相关分析请参见第3.3节。

据我们所知，Visual N-Grams（Li等人，2017）首次以上述方式研究了向现有图像分类数据集的零样本迁移。这也是我们已知的唯一另一项研究，它使用通用预训练模型探索了向标准图像分类数据集的零样本迁移，并可作为理解CLIP的最佳参考基准。他们的方法学习了包含142,806个视觉n元语法（涵盖1至5元）词典的参数，并利用Jelinek-Mercer平滑的微分形式优化这些n元语法，以最大化给定图像所有文本n元语法的概率。为实现零样本迁移，他们首先将数据集中每个类别名称的文本转换为n元语法表示，随后根据其模型计算概率，最终预测得分最高的类别。

我们专注于研究零样本迁移作为任务学习的评估方法，这一思路受到自然语言处理领域任务学习研究的启发。据我们所知，Liu等人（2018）首次将任务学习界定为一种“意外副作用”——当训练用于生成维基百科文章的语言模型时，它学会了在不同语言间可靠地音译人名。虽然GPT-1（Radford等人，2018）主要侧重于预训练——作为一种提升监督式微调的迁移学习方法，训练过程还包含了一项消融研究，该研究表明四种启发式零样本迁移方法的性能在预训练过程中稳步提升，无需任何监督适应。这一分析为GPT-2（Radford等人，2019年）奠定了基础，该研究专注于通过零样本迁移探索语言模型的任务学习能力。

3.1.2. 使用CLIP进行零样本迁移

CLIP经过预训练，能够预测图像与文本片段是否在其数据集中配对。为实现零样本分类，我们复用这一能力。针对每个数据集，我们使用数据集中所有类别的名称作为潜在文本配对集合，并依据CLIP预测最可能的（图像，文本）配对。具体而言，我们首先通过各自的编码器计算图像的特征嵌入和所有可能文本的特征嵌入。随后计算这些嵌入的余弦相似度，通过温度参数τ进行缩放，并经由softmax归一化为概率分布。需注意，此预测层是一个具有L2归一化输入、L2归一化权重、无偏置项和温度缩放的多项式逻辑回归分类器。按此解读，图像编码器是计算图像特征表示的计算机视觉主干网络，而文本编码器则是一个超网络（Ha等人，2016），它基于描述类别所代表视觉概念的文本来生成线性分类器的权重。Lei Ba等人（2015）首次提出了这种形式的零样本图像分类器，而基于自然语言生成分类器的思想至少可追溯至Elhoseiny等人（2013）。延续这一解读，CLIP预训练的每个步骤都可视为对随机创建的计算机视觉数据集代理的性能优化——该数据集每个类别仅含1个样本，并通过自然语言描述定义了总计32,768个类别。在零样本评估中，我们会在文本编码器计算出零样本分类器后将其缓存，并复用于所有后续预测。这使得生成分类器的成本能够分摊到数据集中所有预测任务上。

3.1.3. 与视觉N元语法的初步比较

在表1中，我们将视觉N-Gram与CLIP进行对比。最佳的CLIP模型将ImageNet上的准确率从概念验证的11.5%提升至76.2%，并且在不使用该数据集128万个人工标注训练样本的情况下，达到了原始ResNet-50的性能水平。此外，CLIP模型的Top-5准确率明显高于其Top-1准确率，该模型的Top-5准确率达到95%，与Inception-V4（Szegedy等人，2016）的表现相当。这种在零样本设定下能与强大的全监督基线模型性能匹配的能力表明

	aYahoo	ImageNet	SUN
Visual N-Grams	72.4	11.5	23.0
CLIP	**98.4**	**76.2**	**58.5**

Table 1. Comparing CLIP to prior zero-shot transfer image classification results. CLIP improves performance on all three datasets by a large amount. This improvement reflects many differences in the 4 years since the development of Visual N-Grams (Li et al., 2017).

CLIP is a significant step towards flexible and practical zero-shot computer vision classifiers. As mentioned above, the comparison to Visual N-Grams is meant for contextualizing the performance of CLIP and should not be interpreted as a direct methods comparison between CLIP and Visual N-Grams as many performance relevant differences between the two systems were not controlled for. For instance, we train on a dataset that is 10x larger, use a vision model that requires nearly 100x more compute per prediction, likely used over 1000x their training compute, and use a transformer-based model which did not exist when Visual N-Grams was published. As a closer comparison, we trained a CLIP ResNet-50 on the same YFCC100M dataset that Visual N-Grams was trained on and found it matched their reported ImageNet performance within a V100 GPU day. This baseline was also trained from scratch instead of being initialized from pre-trained ImageNet weights as in Visual N-Grams.

CLIP also outperforms Visual N-Grams on the other 2 reported datasets. On aYahoo, CLIP achieves a 95% reduction in the number of errors, and on SUN, CLIP more than doubles the accuracy of Visual N-Grams. To conduct a more comprehensive analysis and stress test, we implement a much larger evaluation suite detailed in Appendix A. In total we expand from the 3 datasets reported in Visual N-Grams to include over 30 datasets and compare to over 50 existing computer vision systems to contextualize results.

3.1.4. Prompt Engineering and Ensembling

Most standard image classification datasets treat the information naming or describing classes which enables natural language based zero-shot transfer as an afterthought. The vast majority of datasets annotate images with just a numeric id of the label and contain a file mapping these ids back to their names in English. Some datasets, such as Flowers102 and GTSRB, don't appear to include this mapping at all in their released versions preventing zero-shot transfer entirely.[2] For many datasets, we observed these labels may be

[2]Alec learned much more about flower species and German traffic signs over the course of this project than he originally anticipated.

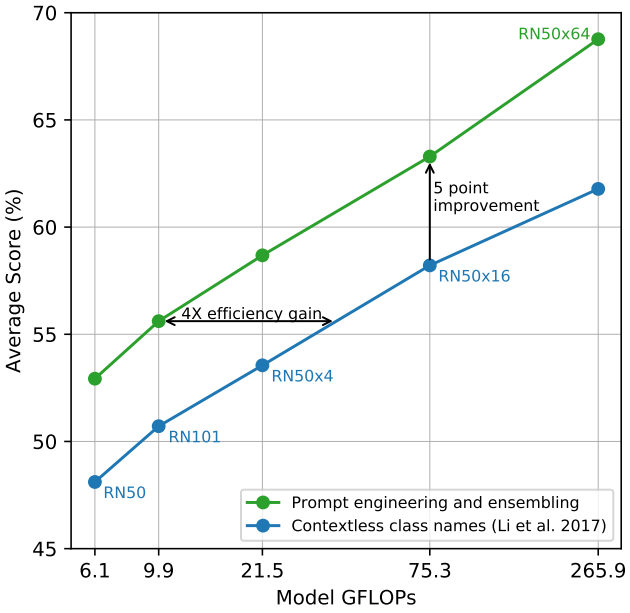

Figure 4. **Prompt engineering and ensembling improve zero-shot performance.** Compared to the baseline of using contextless class names, prompt engineering and ensembling boost zero-shot classification performance by almost 5 points on average across 36 datasets. This improvement is similar to the gain from using 4 times more compute with the baseline zero-shot method but is "free" when amortized over many predictions.

chosen somewhat haphazardly and do not anticipate issues related to zero-shot transfer which relies on task description in order to transfer successfully.

A common issue is polysemy. When the name of a class is the only information provided to CLIP's text encoder it is unable to differentiate which word sense is meant due to the lack of context. In some cases multiple meanings of the same word might be included as different classes in the same dataset! This happens in ImageNet which contains both construction cranes and cranes that fly. Another example is found in classes of the Oxford-IIIT Pet dataset where the word boxer is, from context, clearly referring to a breed of dog, but to a text encoder lacking context could just as likely refer to a type of athlete.

Another issue we encountered is that it's relatively rare in our pre-training dataset for the text paired with the image to be just a single word. Usually the text is a full sentence describing the image in some way. To help bridge this distribution gap, we found that using the prompt template "`A photo of a {label}.`" to be a good default that helps specify the text is about the content of the image. This often improves performance over the baseline of using only the label text. For instance, just using this prompt improves accuracy on ImageNet by 1.3%.

	aYahoo	ImageNet	SUN
Visual N-Grams	72.4	11.5	23.0
CLIP	**98.4**	**76.2**	**58.5**

表1. 对比CLIP与先前零样本迁移图像分类结果。CLIP大幅提升了三个数据集的性能表现。这一进步反映了自视觉N元语法（Li等人，2017）提出四年来的诸多技术差异。

CLIP是朝着灵活实用的零样本计算机视觉分类器迈出的重要一步。如上所述，与Visual N-Grams的比较旨在为CLIP的性能提供背景参照，不应被解读为CLIP与Visual N-Grams之间的直接方法对比，因为两个系统间许多影响性能的差异并未得到控制。例如，我们使用的训练数据集规模扩大了10倍，采用的视觉模型每次预测所需的计算量增加近100倍，训练总计算量可能超过其1000倍，并且使用了Visual N-Grams发布时尚未存在的基于Transformer的模型。为进行更贴近的比较，我们在Visual N-Grams训练所用的相同YFCC100M数据集上训练了CLIP ResNet-50模型，发现其仅需单个V100 GPU运行一天即可达到他们报告的ImageNet性能水平。该基线模型同样是从零开始训练，而非像Visual N-Grams那样使用预训练的ImageNet权重进行初始化。

CLIP在其他两个已报告的数据集上也优于Visual N-Grams。在aYahoo数据集上，CLIP实现了95%的错误率降低；在SUN数据集上，CLIP的准确率比Visual N-Grams提高了一倍以上。为了进行更全面的分析和压力测试，我们实施了附录A中详述的更大规模评估体系。总体而言，我们将评估范围从Visual N-Grams报告的3个数据集扩展到涵盖30多个数据集，并与50多个现有计算机视觉系统进行比较，以便对结果进行背景化分析。

3.1.4. 提示工程与集成

大多数标准图像分类数据集将用于命名或描述类别的信息视为事后补充，这些信息使得基于自然语言的零样本迁移成为可能。绝大多数数据集仅用标签的数字ID来标注图像，并包含一个将这些ID映射回其英文名称的文件。一些数据集，例如Flowers102和GTSRB，在其发布版本中似乎完全未包含此类映射，从而完全阻碍了零样本迁移。[2] 对于许多数据集，我们观察到这些标签可能

[2]Alec learned much more about flower species and German traffic signs over the course of this project than he originally anticipated.

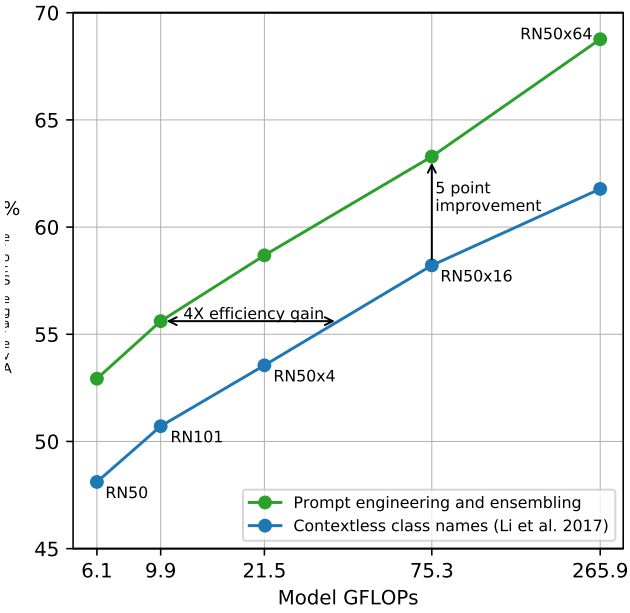

图4. 提示工程与集成方法提升零样本性能。相较于使用无上下文类别名称的基线方法，提示工程与集成策略在36个数据集上将零样本分类性能平均提升近5个百分点。这一改进效果类似于基线零样本方法增加4倍计算量所获得的增益，但在大量预测任务中分摊时可视为“零成本”优化。

选择有些随意，且未预见到与零样本迁移相关的问题，这种迁移依赖于任务描述才能成功进行。

一个常见的问题是歧义性。当类别的名称是提供给CLIP文本编码器的唯一信息时，由于缺乏上下文，它无法区分词语的具体含义。在某些情况下，同一个词的多种含义可能作为不同类别出现在同一数据集中！例如，ImageNet中既包含建筑起重机，也包含飞行的鹤。另一个例子出现在牛津-IIIT宠物数据集的类别中：根据上下文，“boxer”一词明显指代一种犬类，但对于缺乏上下文的文本编码器来说，它同样可能指代一种运动员类型。

我们遇到的另一个问题是，在我们的预训练数据集中，与图像配对的文本仅为单个单词的情况相对较少。通常，文本是一个完整的句子，以某种方式描述图像。为了弥合这种分布差距，我们发现使用提示模板“一张{v*}标签{v*}的照片。”是一个很好的默认设置，有助于明确文本是关于图像内容的。这通常比仅使用标签文本的基线方法提高了性能。例如，仅使用此提示就将ImageNet的准确率提高了1.3%。

Similar to the "prompt engineering" discussion around GPT-3 (Brown et al., 2020; Gao et al., 2020), we have also observed that zero-shot performance can be significantly improved by customizing the prompt text to each task. A few, non exhaustive, examples follow. We found on several fine-grained image classification datasets that it helped to specify the category. For example on Oxford-IIIT Pets, using "`A photo of a {label}, a type of pet.`" to help provide context worked well. Likewise, on Food101 specifying *a type of food* and on FGVC Aircraft *a type of aircraft* helped too. For OCR datasets, we found that putting quotes around the text or number to be recognized improved performance. Finally, we found that on satellite image classification datasets it helped to specify that the images were of this form and we use variants of "`a satellite photo of a {label}.`".

We also experimented with ensembling over multiple zero-shot classifiers as another way of improving performance. These classifiers are computed by using different context prompts such as '`A photo of a big {label}`" and "`A photo of a small {label}`". We construct the ensemble over the embedding space instead of probability space. This allows us to cache a single set of averaged text embeddings so that the compute cost of the ensemble is the same as using a single classifier when amortized over many predictions. We've observed ensembling across many generated zero-shot classifiers to reliably improve performance and use it for the majority of datasets. On ImageNet, we ensemble 80 different context prompts and this improves performance by an additional 3.5% over the single default prompt discussed above. When considered together, prompt engineering and ensembling improve ImageNet accuracy by almost 5%. In Figure 4 we visualize how prompt engineering and ensembling change the performance of a set of CLIP models compared to the contextless baseline approach of directly embedding the class name as done in Li et al. (2017).

3.1.5. Analysis of Zero-Shot CLIP Performance

Since task-agnostic zero-shot classifiers for computer vision have been understudied, CLIP provides a promising opportunity to gain a better understanding of this type of model. In this section, we conduct a study of various properties of CLIP's zero-shot classifiers. As a first question, we look simply at how well zero-shot classifiers perform. To contextualize this, we compare to the performance of a simple off-the-shelf baseline: fitting a fully supervised, regularized, logistic regression classifier on the features of the canonical ResNet-50. In Figure 5 we show this comparison across 27 datasets. Please see Appendix A for details of datasets and setup.

Zero-shot CLIP outperforms this baseline slightly more often than not and wins on 16 of the 27 datasets. Looking at individual datasets reveals some interesting behavior. On fine-grained classification tasks, we observe a wide spread in performance. On two of these datasets, Stanford Cars and Food101, zero-shot CLIP outperforms logistic regression on ResNet-50 features by over 20% while on two others, Flowers102 and FGVCAircraft, zero-shot CLIP underperforms by over 10%. On OxfordPets and Birdsnap, performance is much closer. We suspect these difference are primarily due to varying amounts of per-task supervision between WIT and ImageNet. On "general" object classification datasets such as ImageNet, CIFAR10/100, STL10, and PascalVOC2007 performance is relatively similar with a slight advantage for zero-shot CLIP in all cases. On STL10, CLIP achieves 99.3% overall which appears to be a new state of the art despite not using any training examples. Zero-shot CLIP significantly outperforms a ResNet-50 on two datasets measuring action recognition in videos. On Kinetics700, CLIP outperforms a ResNet-50 by 14.5%. Zero-shot CLIP also outperforms a ResNet-50's features by 7.7% on UCF101. We speculate this is due to natural language providing wider supervision for visual concepts involving verbs, compared to the noun-centric object supervision in ImageNet.

Looking at where zero-shot CLIP notably underperforms,

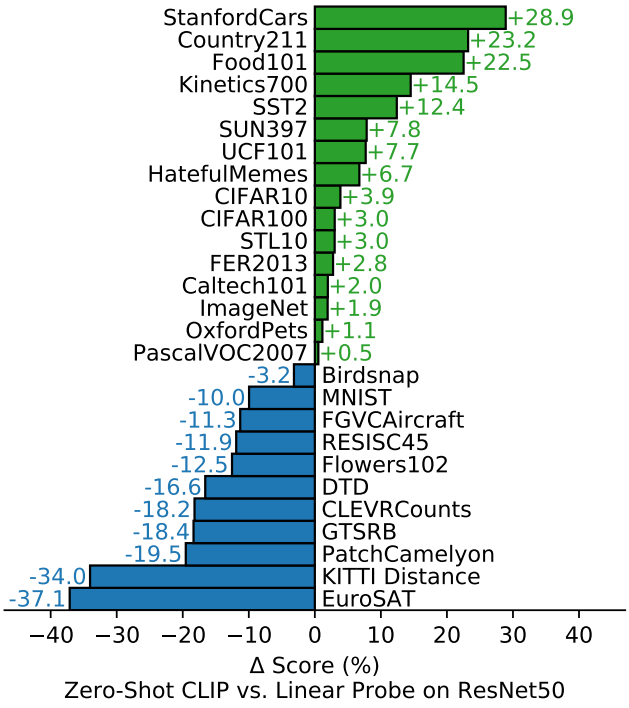

Figure 5. **Zero-shot CLIP is competitive with a fully supervised baseline.** Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.

类似于围绕GPT-3的“提示工程”讨论（Brown等人，2020；Gao等人，2020），我们也观察到，通过为每个任务定制提示文本，零样本性能可以得到显著提升。以下是一些非详尽的示例。我们在多个细粒度图像分类数据集上发现，指定类别有助于提升效果。例如在Oxford-IIIT Pets数据集上，使用“一张{标签}的照片，一种宠物。”来提供上下文效果良好。同样，在Food101数据集上指定*a type of food*，在FGVC Aircraft数据集上指定*a type ofaircraft*也有帮助。对于OCR数据集，我们发现将要识别的文本或数字用引号括起来可以提高性能。最后，我们发现在卫星图像分类数据集上，注明图像属于此类形式并使用“一张{标签}的卫星照片。”的变体有所助益。

我们还尝试了集成多个零样本分类器作为另一种提升性能的方法。这些分类器通过使用不同的上下文提示来计算，例如“一张大{标签}的照片”和“一张小{标签}的照片”。我们在嵌入空间而非概率空间上构建集成。这使得我们能够缓存一组平均文本嵌入，从而在多次预测中分摊计算成本时，集成的计算开销与使用单个分类器相同。我们观察到，通过集成多个生成的零样本分类器能够可靠地提升性能，并将其应用于大多数数据集。在ImageNet上，我们集成了80种不同的上下文提示，这比上述单一默认提示的性能额外提高了3.5%。综合来看，提示工程和集成技术共同将ImageNet的准确率提升了近5%。在图4中，我们可视化了提示工程和集成如何改变一系列CLIP模型的性能，并与Li等人（2017）中直接将类别名称嵌入的无上下文基线方法进行了对比。

3.1.5. 零样本CLIP性能分析

由于计算机视觉中任务无关的零样本分类器尚未得到充分研究，CLIP为深入理解此类模型提供了重要契机。本节我们将系统探究CLIP零样本分类器的各项特性。首要问题是评估零样本分类器的基本性能。为提供参照基准，我们将其与经典ResNet-50特征上训练的完全监督、正则化逻辑回归分类器进行对比，该基线模型可直接调用现有框架实现。图5展示了在27个数据集上的对比结果，具体数据集设置详见附录A。

Zero-shot CLIP 略微超越这一基线更多——

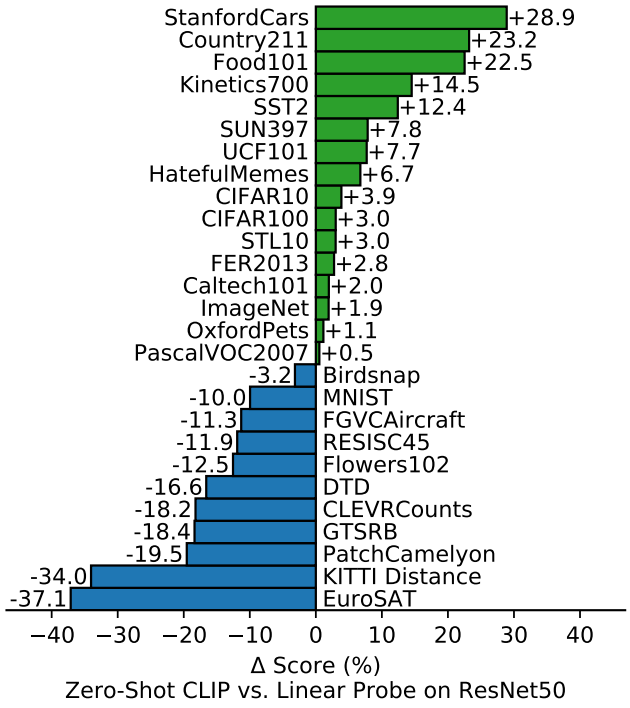

图5. 零样本CLIP与全监督基线模型表现相当。在涵盖27个数据集的评估体系中，零样本CLIP分类器在包括ImageNet在内的16个数据集上超越了基于ResNet-50特征训练的全监督线性分类器。

在27个数据集中的16个上，零样本CLIP的表现优于逻辑回归。观察单个数据集时，我们发现了一些有趣的现象。在细粒度分类任务中，我们观察到性能差异很大。在其中两个数据集——Stanford Cars和Food101上，零样本CLIP的表现比基于ResNet-50特征训练的logistic回归高出20%以上；而在另外两个数据集——Flowers102和FGVCAircraft上，零样本CLIP的表现则低了10%以上。在OxfordPets和Birdsnap上，两者的性能则非常接近。我们怀疑这些差异主要是由于WIT和ImageNet在不同任务上提供的监督信息量不同所致。在“通用”物体分类数据集上，如ImageNet、CIFAR10/100、STL10和PascalVOC2007，两者的性能相对接近，但零样本CLIP在所有情况下都略有优势。在STL10上，CLIP达到了99.3%的整体准确率，这似乎是一个新的最高水平，尽管它没有使用任何训练样本。在两个衡量视频动作识别的数据集上，零样本CLIP显著优于ResNet-50。在Kinetics700上，CLIP比ResNet-50高出14.5%。在UCF101上，零样本CLIP也比ResNet-50的特征高出7.7%。我们推测这是因为自然语言为涉及动词的视觉概念提供了更广泛的监督，而ImageNet的监督则更侧重于名词性的物体。

观察零样本CLIP明显表现不佳的领域，

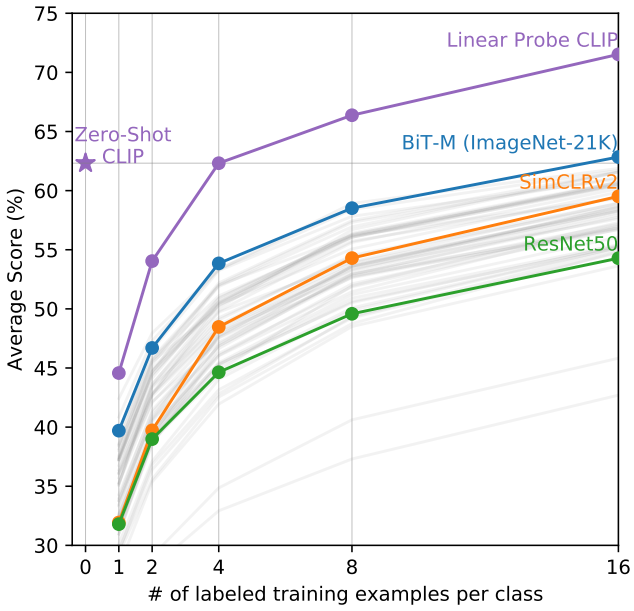

Figure 6. **Zero-shot CLIP outperforms few-shot linear probes.** Zero-shot CLIP matches the average performance of a 4-shot linear classifier trained on the same feature space and nearly matches the best results of a 16-shot linear classifier across publicly available models. For both BiT-M and SimCLRv2, the best performing model is highlighted. Light gray lines are other models in the eval suite. The 20 datasets with at least 16 examples per class were used in this analysis.

we see that zero-shot CLIP is quite weak on several specialized, complex, or abstract tasks such as satellite image classification (EuroSAT and RESISC45), lymph node tumor detection (PatchCamelyon), counting objects in synthetic scenes (CLEVRCounts), self-driving related tasks such as German traffic sign recognition (GTSRB), recognizing distance to the nearest car (KITTI Distance). These results highlight the poor capability of zero-shot CLIP on more complex tasks. By contrast, non-expert humans can robustly perform several of these tasks, such as counting, satellite image classification, and traffic sign recognition, suggesting significant room for improvement. However, we caution that it is unclear whether measuring zero-shot transfer, as opposed to few-shot transfer, is a meaningful evaluation for difficult tasks that a learner has no prior experience with, such as lymph node tumor classification for almost all humans (and possibly CLIP).

While comparing zero-shot performance to fully supervised models contextualizes the task-learning capabilities of CLIP, comparing to few-shot methods is a more direct comparison, since zero-shot is its limit. In Figure 6, we visualize how zero-shot CLIP compares to few-shot logistic regression on the features of many image models including the best publicly available ImageNet models, self-supervised learning methods, and CLIP itself. While it is intuitive to expect zero-shot to underperform one-shot, we instead find that zero-shot CLIP matches the performance of 4-shot logistic regression on the same feature space. This is likely due to an important difference between the zero-shot and few-shot approach. First, CLIP's zero-shot classifier is generated via natural language which allows for visual concepts to be directly specified ("communicated"). By contrast, "normal" supervised learning must infer concepts indirectly from training examples. Context-less example-based learning has the drawback that many different hypotheses can be consistent with the data, especially in the one-shot case. A single image often contains many different visual concepts. Although a capable learner is able to exploit visual cues and heuristics, such as assuming that the concept being demonstrated is the primary object in an image, there is no guarantee.

A potential resolution of this discrepancy between zero-shot and few-shot performance is to use CLIP's zero-shot classifier as a prior for the weights of the few-shot classifier. While adding an L2 penalty towards the generated weights is a straightforward implementation of this idea, we found that hyperparameter optimization would often select for such a large value of this regularizer that the resulting few-shot classifier was "just" the zero-shot classifier. Research into better methods of combining the strength of zero-shot transfer with flexibility of few-shot learning is a promising direction for future work.

When comparing zero-shot CLIP to few-shot logistic regression on the features of other models, zero-shot CLIP roughly matches the performance of the best performing 16-shot classifier in our evaluation suite, which uses the features of a BiT-M ResNet-152x2 trained on ImageNet-21K. We are certain that a BiT-L model trained on JFT-300M would perform even better but these models have not been publicly released. That a BiT-M ResNet-152x2 performs best in a 16-shot setting is somewhat surprising since, as analyzed in Section 3.2, the Noisy Student EfficientNet-L2 outperforms it in a fully supervised setting by almost 5% on average across 27 datasets.

In addition to studying the average performance of zero-shot CLIP and few-shot logistic regression, we also examine performance on individual datasets. In Figure 7, we show estimates for the number of labeled examples per class that a logistic regression classifier on the same feature space requires to match the performance of zero-shot CLIP. Since zero-shot CLIP is also a linear classifier, this estimates the effective data efficiency of zero-shot transfer in this setting. In order to avoid training thousands of linear classifiers, we estimate the effective data efficiency based on a log-linear interpolation of the performance of a 1, 2, 4, 8, 16-shot (when possible), and a fully supervised linear classifier trained on each dataset. We find that zero-shot transfer can

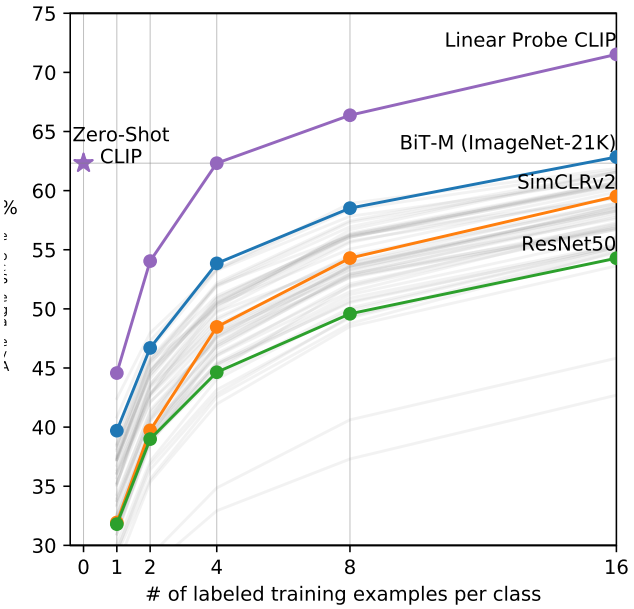

图6. 零样本CLIP超越少样本线性探针。零样本CLIP与在同一特征空间上训练的4样本线性分类器的平均性能相当，并且几乎与公开可用模型中16样本线性分类器的最佳结果持平。对于BiT-M和SimCLRv2，均突出显示了性能最佳的模型。浅灰色线条表示评估套件中的其他模型。本分析使用了每类至少包含16个样本的20个数据集。

我们发现，零样本CLIP在多项专业化、复杂或抽象任务上表现相当薄弱，例如卫星图像分类（EuroSAT和RESISC45）、淋巴结肿瘤检测（PatchCamelyon）、合成场景中的物体计数（CLEVRCounts），以及自动驾驶相关任务如德国交通标志识别（GTSRB）、判断最近车辆距离（KITTI Distance）等。这些结果凸显了零样本CLIP在处理更复杂任务时的能力不足。相比之下，非专业人士却能稳健地完成其中多项任务，例如计数、卫星图像分类和交通标志识别，这表明模型仍有巨大的改进空间。然而，我们需要谨慎指出：对于学习者毫无先验经验的困难任务（例如几乎所有人类——可能也包括CLIP——都未曾接触过的淋巴结肿瘤分类），衡量零样本迁移而非少样本迁移是否构成有意义的评估标准，目前尚不明确。

在将零样本性能与全监督模型进行比较时，CLIP的任务学习能力得以情境化；而与少样本方法对比则更为直接，因为零样本正是其极限。在图6中，我们可视化展示了零样本CLIP与基于多种图像模型特征的少样本逻辑回归的对比结果，这些模型包括当前公开可用的最佳ImageNet模型、自监督学习方法以及CLIP本身。尽管直观上我们预期零样本学习会表现不如单样本学习，但结果却发现零样本CLIP在相同特征空间上的表现与4样本逻辑回归相当。这很可能源于零样本与少样本方法间的一个重要差异。首先，CLIP的零样本分类器是通过自然语言生成的，这使得视觉概念能够被直接“传达”和指定。相比之下，“常规”监督学习必须从训练样本中间接推断概念。缺乏上下文的情境下，基于样本的学习存在一个缺陷：许多不同的假设都可能与数据相符，尤其在单样本情况下更为明显。单张图像通常包含多种不同的视觉概念。尽管强大的学习者能够利用视觉线索和启发式方法（例如假设所演示的概念是图像中的主要对象），但这并不能得到保证。

解决零样本与少样本性能差异的一个潜在方案是，将CLIP的零样本分类器作为少样本分类器权重的先验。虽然向生成权重添加L2惩罚是这一想法的直接实现方式，但我们发现超参数优化往往会为该正则项选择过大的值，导致最终的少样本分类器“仅仅”等同于零样本分类器。未来工作的一个前景广阔的方向是研究如何更好地结合零样本迁移的优势与少样本学习的灵活性。

在将零样本CLIP与其他模型特征的少样本逻辑回归进行比较时，零样本CLIP大致匹配了我们评估套件中表现最佳的16样本分类器的性能，该分类器使用了在ImageNet-21K上训练的BiT-M ResNet-152x2的特征。我们确信，在JFT-300M上训练的BiT-L模型表现会更优，但这些模型尚未公开发布。BiT-M ResNet-152x2在16样本设置中表现最佳有些令人意外，因为如第3.2节所分析，在完全监督设置下，Noisy Student EfficientNet-L2在27个数据集上的平均表现比它高出近5%。

除了研究零样本CLIP和少样本逻辑回归的平均性能外，我们还考察了在单个数据集上的表现。在图7中，我们展示了在同一特征空间上，逻辑回归分类器需要每个类别多少标注样本才能达到零样本CLIP的性能水平。由于零样本CLIP本身也是一个线性分类器，这估算了在此设定下零样本迁移的有效数据效率。为了避免训练数千个线性分类器，我们基于对数线性插值来估算有效数据效率，插值数据点包括1、2、4、8、16样本（在可能的情况下）以及每个数据集上训练的全监督线性分类器的性能。我们发现零样本迁移能够

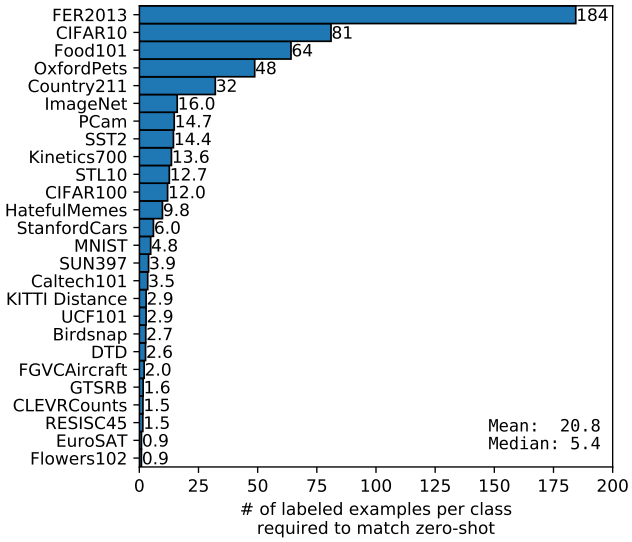

Figure 7. **The data efficiency of zero-shot transfer varies widely.** Calculating the number of labeled examples per class a linear classifier on the same CLIP feature space requires to match the performance of the zero-shot classifier contextualizes the effectiveness of zero-shot transfer. Values are estimated based on log-linear interpolation of 1, 2, 4, 8, 16-shot and fully supervised results. Performance varies widely from still underperforming a one-shot classifier on two datasets to matching an estimated 184 labeled examples per class.

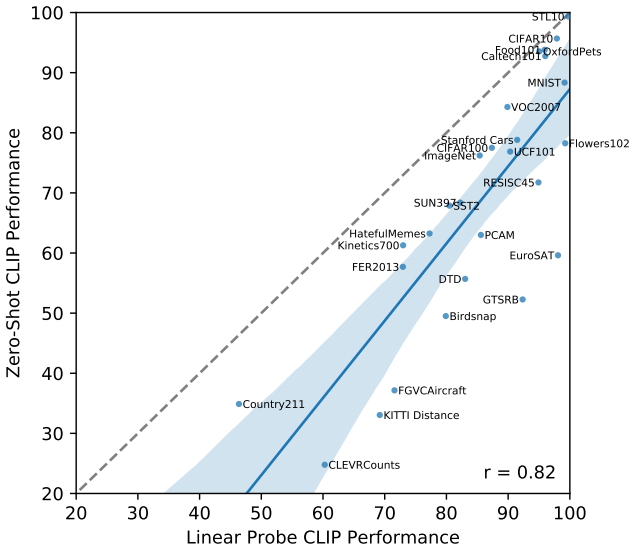

Figure 8. **Zero-shot performance is correlated with linear probe performance but still mostly sub-optimal.** Comparing zero-shot and linear probe performance across datasets shows a strong correlation with zero-shot performance mostly shifted 10 to 25 points lower. On only 5 datasets does zero-shot performance approach linear probe performance ($\leq$3 point difference).

have widely varying efficiency per dataset from less than 1 labeled example per class to 184. Two datasets, Flowers102 and EuroSAT underperform one-shot models. Half of the datasets require less than 5 examples per class with a median of 5.4. However, the mean estimated data efficiency is 20.8 examples per class. This is due to the 20% of datasets where supervised classifiers require many labeled examples per class in order to match performance. On ImageNet, zero-shot CLIP matches the performance of a 16-shot linear classifier trained on the same feature space.

If we assume that evaluation datasets are large enough that the parameters of linear classifiers trained on them are well estimated, then, because CLIP's zero-shot classifier is also a linear classifier, the performance of the fully supervised classifiers roughly sets an upper bound for what zero-shot transfer can achieve. In Figure 8 we compare CLIP's zero-shot performance with fully supervised linear classifiers across datasets. The dashed, $y = x$ line represents an "optimal" zero-shot classifier that matches the performance of its fully supervised equivalent. For most datasets, the performance of zero-shot classifiers still underperform fully supervised classifiers by 10% to 25%, suggesting that there is still plenty of headroom for improving CLIP's task-learning and zero-shot transfer capabilities.

There is a positive correlation of 0.82 (p-value $< 10^{-6}$) between zero-shot performance and fully supervised performance, suggesting that CLIP is relatively consistent at connecting underlying representation and task learning to zero-shot transfer. However, zero-shot CLIP only approaches fully supervised performance on 5 datasets: STL10, CIFAR10, Food101, OxfordPets, and Caltech101. On all 5 datasets, both zero-shot accuracy and fully supervised accuracy are over 90%. This suggests that CLIP may be more effective at zero-shot transfer for tasks where its underlying representations are also high quality. The slope of a linear regression model predicting zero-shot performance as a function of fully supervised performance estimates that for every 1% improvement in fully supervised performance, zero-shot performance improves by 1.28%. However, the 95th-percentile confidence intervals still include values of less than 1 (0.93-1.79).

Over the past few years, empirical studies of deep learning systems have documented that performance is predictable as a function of important quantities such as training compute and dataset size (Hestness et al., 2017; Kaplan et al., 2020). The GPT family of models has so far demonstrated consistent improvements in zero-shot performance across a 1000x increase in training compute. In Figure 9, we check whether the zero-shot performance of CLIP follows a similar scaling pattern. We plot the average error rate of the 5 ResNet CLIP models across 39 evaluations on 36 different datasets and find that a similar log-log linear scaling trend holds for CLIP across a 44x increase in model compute. While the overall trend is smooth, we found that performance on individual evaluations can be much noisier. We are unsure whether

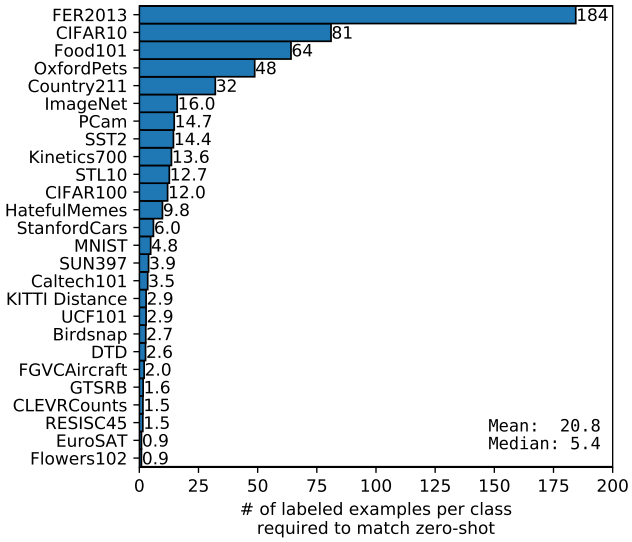

图7. 零样本迁移的数据效率差异显著。通过计算在同一CLIP特征空间上训练的线性分类器为达到零样本分类器性能所需每类标注样本数量，可以量化零样本迁移的实际效能。该数值基于1、2、4、8、16样本及全监督结果的半对数线性插值估算。不同数据集表现差异巨大：在两个数据集上仍低于单样本分类器性能，而在最佳情况下可达到相当于每类184个标注样本的预估效果。

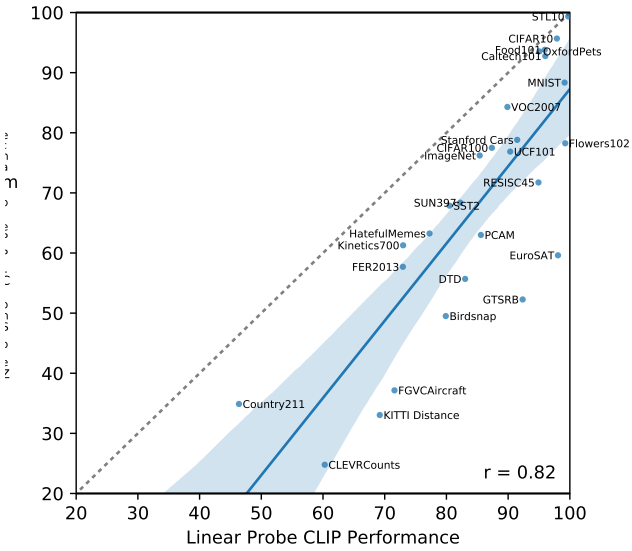

图8. 零样本性能与线性探针性能相关，但多数情况下仍非最优。跨数据集比较零样本与线性探针性能显示二者存在强相关性，零样本性能普遍低10至25个百分点。仅在5个数据集中，零样本性能接近线性探针性能（{v*}3个百分点的差异）。

不同数据集间的效率差异巨大，每个类别所需的标注样本量从不足1个到184个不等。Flowers102和EuroSAT两个数据集的表现甚至低于单样本模型。半数数据集每个类别只需不到5个样本，中位数为5.4个。但平均数据效率估算值为每个类别20.8个样本，这是因为有20%的数据集需要大量标注样本才能使监督分类器达到同等性能。在ImageNet上，零样本CLIP的性能与在同一特征空间训练的16样本线性分类器相当。

如果我们假设评估数据集足够大，使得在其上训练的线性分类器参数能得到良好估计，那么由于CLIP的零样本分类器同样属于线性分类器，全监督分类器的性能大致设定了零样本迁移所能达到的上限。在图8中，我们对比了CLIP的零样本性能与各数据集上的全监督线性分类器。虚线$y = x$代表与全监督版本性能匹配的“最优”零样本分类器。对于大多数数据集，零样本分类器的性能仍低于全监督分类器10%至25%，这表明提升CLIP的任务学习与零样本迁移能力仍有充足空间。

零样本性能与全监督性能之间存在0.82的正相关性（p值$< 10^{-6}$）。

表现，这表明CLIP在将底层表示和任务学习与零样本迁移联系起来方面相对一致。然而，零样本CLIP仅在5个数据集上接近完全监督性能：STL10、CIFAR10、Food101、OxfordPets和Caltech101。在这5个数据集中，零样本准确率和完全监督准确率均超过90%。这表明，对于其底层表示质量也较高的任务，CLIP在零样本迁移方面可能更有效。通过线性回归模型预测零样本性能作为完全监督性能的函数，其斜率估计显示：完全监督性能每提高1%，零样本性能提高1.28%。然而，95%置信区间仍包含小于1的值（0.93-1.79）。

过去几年，深度学习系统的实证研究表明，性能可依据训练计算量和数据集规模等重要变量进行预测（Hestness等人，2017；Kaplan等人，2020）。GPT系列模型迄今已证明，在训练计算量增长1000倍的过程中，其零样本性能始终保持稳定提升。在图9中，我们检验了CLIP的零样本性能是否遵循相似的缩放规律。通过绘制5个ResNet CLIP模型在36个不同数据集上39项评估的平均错误率，我们发现CLIP在模型计算量增长44倍的过程中，同样呈现出对数-对数线性缩放趋势。虽然整体趋势平稳，但我们发现单项评估的性能可能存在较大波动。目前尚不确定

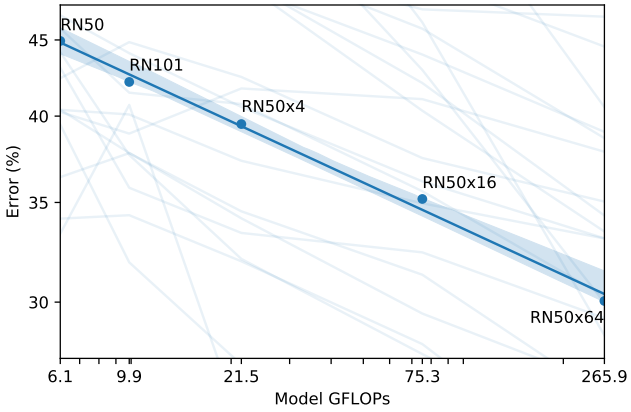

Figure 9. **Zero-shot CLIP performance scales smoothly as a function of model compute.** Across 39 evals on 36 different datasets, average zero-shot error is well modeled by a log-log linear trend across a 44x range of compute spanning 5 different CLIP models. Lightly shaded lines are performance on individual evals, showing that performance is much more varied despite the smooth overall trend.

this is caused by high variance between individual training runs on sub-tasks (as documented in D'Amour et al. (2020)) masking a steadily improving trend or whether performance is actually non-monotonic as a function of compute on some tasks.

3.2. Representation Learning

While we have extensively analyzed the task-learning capabilities of CLIP through zero-shot transfer in the previous section, it is more common to study the representation learning capabilities of a model. There exist many ways to evaluate the quality of representations as well as disagreements over what properties an "ideal" representation should have (Locatello et al., 2020). Fitting a linear classifier on a representation extracted from the model and measuring its performance on various datasets is a common approach. An alternative is measuring the performance of end-to-end fine-tuning of the model. This increases flexibility, and prior work has convincingly demonstrated that fine-tuning outperforms linear classification on most image classification datasets (Kornblith et al., 2019; Zhai et al., 2019). While the high performance of fine-tuning motivates its study for practical reasons, we still opt for linear classifier based evaluation for several reasons. Our work is focused on developing a high-performing task and dataset-agnostic pre-training approach. Fine-tuning, because it adapts representations to each dataset during the fine-tuning phase, can compensate for and potentially mask failures to learn general and robust representations during the pre-training phase. Linear classifiers, because of their limited flexibility, instead highlight these failures and provide clear feedback during development. For CLIP, training supervised linear classifiers has the added benefit of being very similar to the approach used for its zero-shot classifiers which enables extensive comparisons and analysis in Section 3.1. Finally, we aim to compare CLIP to a comprehensive set of existing models across many tasks. Studying 66 different models on 27 different datasets requires tuning 1782 different evaluations. Fine-tuning opens up a much larger design and hyper-parameter space, which makes it difficult to fairly evaluate and computationally expensive to compare a diverse set of techniques as discussed in other large scale empirical studies (Lucic et al., 2018; Choi et al., 2019). By comparison, linear classifiers require minimal hyper-parameter tuning and have standardized implementations and evaluation procedures. Please see Appendix A for further details on evaluation.

Figure 10 summarizes our findings. To minimize selection effects that could raise concerns of confirmation or reporting bias, we first study performance on the 12 dataset evaluation suite from Kornblith et al. (2019). While small CLIP models such as a ResNet-50 and ResNet-101 outperform other ResNets trained on ImageNet-1K (BiT-S and the originals), they underperform ResNets trained on ImageNet-21K (BiT-M). These small CLIP models also underperform models in the EfficientNet family with similar compute requirements. However, models trained with CLIP scale very well and the largest model we trained (ResNet-50x64) slightly outperforms the best performing existing model (a Noisy Student EfficientNet-L2) on both overall score and compute efficiency. We also find that CLIP vision transformers are about 3x more compute efficient than CLIP ResNets, which allows us to reach higher overall performance within our compute budget. These results qualitatively replicate the findings of Dosovitskiy et al. (2020) which reported that vision transformers are more compute efficient than convnets when trained on sufficiently large datasets. Our best overall model is a ViT-L/14 that is fine-tuned at a higher resolution of 336 pixels on our dataset for 1 additional epoch. This model outperforms the best existing model across this evaluation suite by an average of 2.6%.

As Figure 21 qualitatively shows, CLIP models learn a wider set of tasks than has previously been demonstrated in a single computer vision model trained end-to-end from random initialization. These tasks include geo-localization, optical character recognition, facial emotion recognition, and action recognition. None of these tasks are measured in the evaluation suite of Kornblith et al. (2019). This could be argued to be a form of selection bias in Kornblith et al. (2019)'s study towards tasks that overlap with ImageNet. To address this, we also measure performance on a broader 27 dataset evaluation suite. This evaluation suite, detailed in Appendix A includes datasets representing the aforementioned tasks, German Traffic Signs Recognition Benchmark (Stallkamp et al., 2011), as well as several other datasets adapted from VTAB (Zhai et al., 2019).

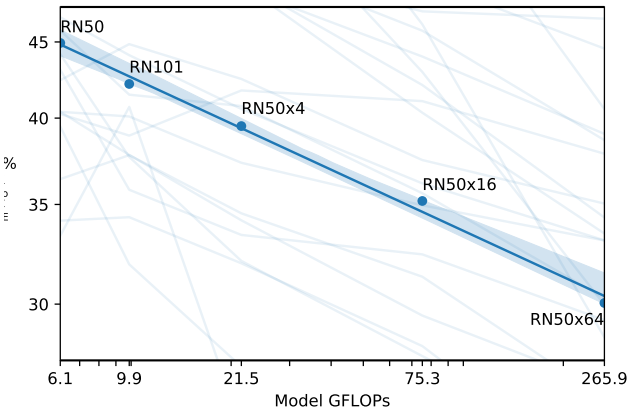

图9. 零样本CLIP性能随模型计算量呈平滑扩展趋势。在涵盖5个不同CLIP模型、计算量跨度达44倍的39次评估（涉及36个数据集）中，平均零样本误差可通过对数-对数线性趋势精准建模。浅色细线代表各单项评估的表现，表明尽管整体趋势平滑，但具体性能仍存在显著波动。

这是由于在子任务上个别训练运行之间的高方差（如D’Amour等人（2020）所记载）掩盖了稳步改善的趋势，还是某些任务上的性能实际上随着计算量的增加而非单调变化。

3.2. 表征学习

尽管我们在上一节中通过零样本迁移广泛分析了CLIP的任务学习能力，但更常见的研究方向是评估模型的表征学习能力。衡量表征质量的方法多种多样，且关于“理想”表征应具备何种特性也存在争议（Locatello等人，2020）。在模型提取的表征上拟合线性分类器并测量其在各数据集上的性能是一种常用方法。另一种方法是测量模型端到端微调的性能。这种方式灵活性更高，先前研究已令人信服地证明在大多数图像分类数据集上，微调优于线性分类（Kornblith等人，2019；翟等人，2019）。虽然微调的高性能出于实用考量值得研究，但我们仍选择基于线性分类器的评估方法，原因如下：我们的工作聚焦于开发高性能、任务与数据集无关的预训练方法。微调在调整阶段会使表征适配每个数据集，这可能补偿并掩盖预训练阶段未能学习通用鲁棒表征的缺陷。而线性分类器因其有限的灵活性，反而能凸显这些缺陷，为开发过程提供清晰反馈。对于CLIP，训练监督线性分类器的另一个好处是，其方法与其零样本分类器所采用的方法非常相似，这使得在第3.1节中能够进行广泛的比较和分析。最后，我们的目标是将CLIP与一系列全面的现有模型在多项任务中进行比较。在27个不同数据集上研究66个不同模型需要调整1782次不同的评估。微调会开启更大的设计和超参数空间，这使得公平评估变得困难，并且在计算上比较多种技术也成本高昂，正如其他大规模实证研究中所讨论的那样（Lucic等人，2018；Choi等人，2019）。相比之下，线性分类器只需最少的超参数调整，并具有标准化的实现和评估流程。有关评估的更多详细信息，请参阅附录A。

图10总结了我们的发现。为最小化可能引发确认偏误或报告偏误担忧的选择效应，我们首先研究了Kornblith等人（2019）提出的12个数据集评估套件上的性能。虽然小型CLIP模型（如ResNet-50和ResNet-101）在ImageNet-1K上训练的ResNet（BiT-S及原始版本）表现更优，但它们逊于在ImageNet-21K上训练的ResNet（BiT-M）。这些小型CLIP模型的表现也不及计算需求相似的EfficientNet系列模型。然而，采用CLIP训练的模型展现出极佳的扩展性——我们训练的最大模型（ResNet-50x64）在综合得分和计算效率上均略优于现有最佳模型（Noisy Student EfficientNet-L2）。我们还发现CLIP视觉变换器的计算效率约为CLIP ResNets的3倍，这使我们在既定计算预算内达到了更高的综合性能。这些结果在定性上复现了Dosovitskiy等人（2020）的发现：当在足够大规模数据集上训练时，视觉变换器比卷积网络具有更高计算效率。我们最佳的综合模型是ViT-L/14，该模型在我们的数据集上以336像素更高分辨率微调了1个额外周期，在此评估套件中平均超越现有最佳模型2.6%。

如图21定性所示，CLIP模型学习到的任务范围比以往任何从随机初始化端到端训练的单一计算机视觉模型都更广泛。这些任务包括地理定位、光学字符识别、面部情绪识别和动作识别。Kornblith等人（2019）的评估体系中并未涵盖这些任务。这或许可被视为Kornblith等人（2019）研究中存在的一种选择偏差，即偏向于与ImageNet重叠的任务。为解决这一问题，我们还采用了一个更广泛的27个数据集评估体系进行性能测量。该评估体系（详见附录A）包含了代表上述任务的数据集、德国交通标志识别基准（Stallkamp等人，2011），以及多个改编自VTAB（Zhai等人，2019）的其他数据集。

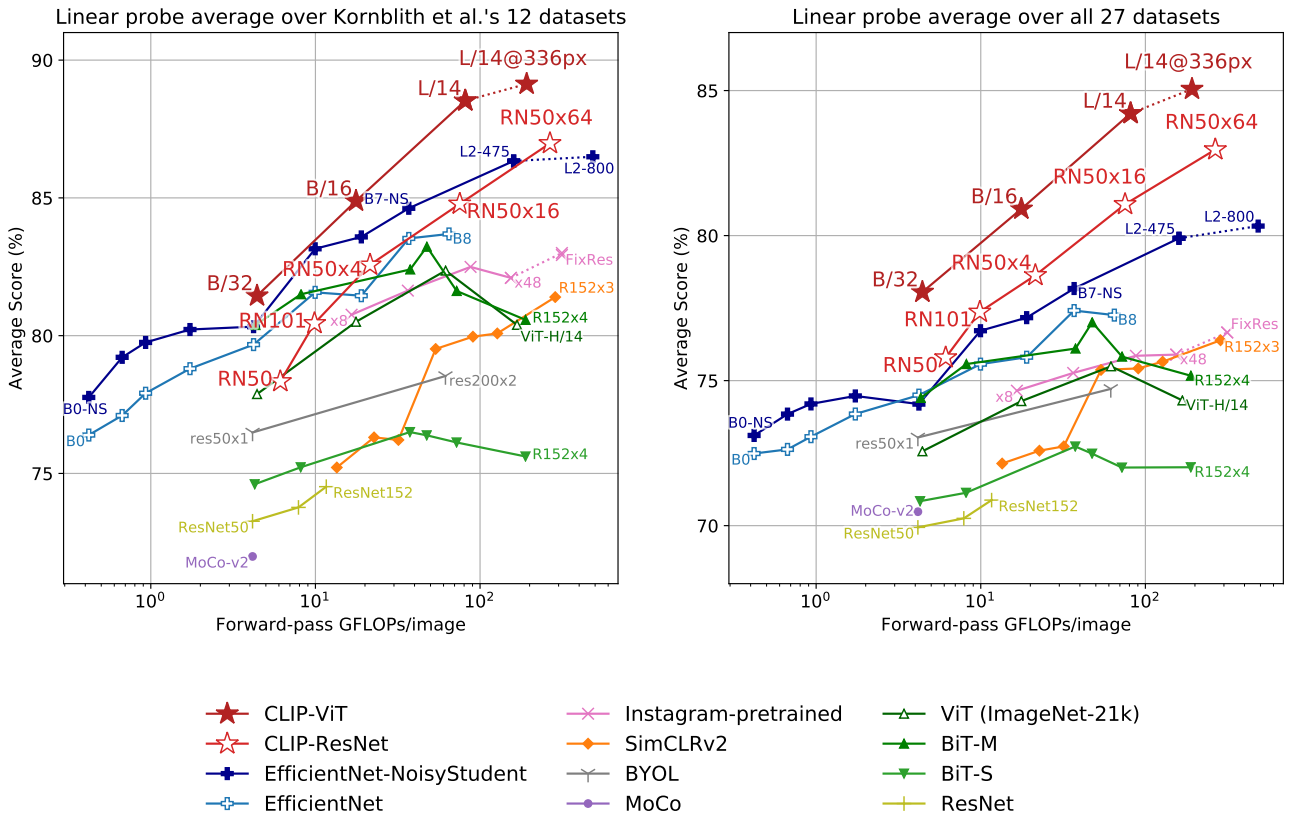

Figure 10. **Linear probe performance of CLIP models in comparison with state-of-the-art computer vision models**, including EfficientNet (Tan & Le, 2019; Xie et al., 2020), MoCo (Chen et al., 2020d), Instagram-pretrained ResNeXt models (Mahajan et al., 2018; Touvron et al., 2019), BiT (Kolesnikov et al., 2019), ViT (Dosovitskiy et al., 2020), SimCLRv2 (Chen et al., 2020c), BYOL (Grill et al., 2020), and the original ResNet models (He et al., 2016b). (Left) Scores are averaged over 12 datasets studied by Kornblith et al. (2019). (Right) Scores are averaged over 27 datasets that contain a wider variety of distributions. Dotted lines indicate models fine-tuned or evaluated on images at a higher-resolution than pre-training. See Table 10 for individual scores and Figure 20 for plots for each dataset.

On this broader evaluation suite, the benefits of CLIP are more clear. All CLIP models, regardless of scale, outperform all evaluated systems in terms of compute efficiency. The improvement in average score of the best model over previous systems increases from 2.6% to 5%. We also find that self-supervised systems do noticeably better on our broader evaluation suite. For instance, while SimCLRv2 still underperforms BiT-M on average on the 12 datasets of Kornblith et al. (2019), SimCLRv2 outperforms BiT-M on our 27 dataset evaluation suite. These findings suggest continuing to expand task diversity and coverage in order to better understand the "general" performance of systems. We suspect additional evaluation efforts along the lines of VTAB to be valuable.

In addition to the aggregate analysis above, we visualize per-dataset differences in the performance of the best CLIP model and the best model in our evaluation suite across all 27 datasets in Figure 11. CLIP outperforms the Noisy Student EfficientNet-L2 on 21 of the 27 datasets. CLIP improves the most on tasks which require OCR (SST2 and HatefulMemes), geo-localization and scene recognition (Country211, SUN397), and activity recognition in videos (Kinetics700 and UCF101). In addition CLIP also does much better on fine-grained car and traffic sign recognition (Stanford Cars and GTSRB). This may reflect a problem with overly narrow supervision in ImageNet. A result such as the 14.7% improvement on GTSRB could be indicative of an issue with ImageNet-1K, which has only a single label for all traffic and street signs. This could encourage a supervised representation to collapse intra-class details and hurt accuracy on a fine-grained downstream task. As mentioned, CLIP still underperforms the EfficientNet on several datasets. Unsurprisingly, the dataset that the EfficientNet does best relative to CLIP on is the one it was trained on: ImageNet. The EffcientNet also slightly outperforms CLIP on low-resolution datasets such as CIFAR10 and CIFAR100. We suspect this is at least partly due to the lack of scale-based data augmentation in CLIP. The EfficientNet also does slightly better on PatchCamelyon and CLEVRCounts, datasets where overall performance is still

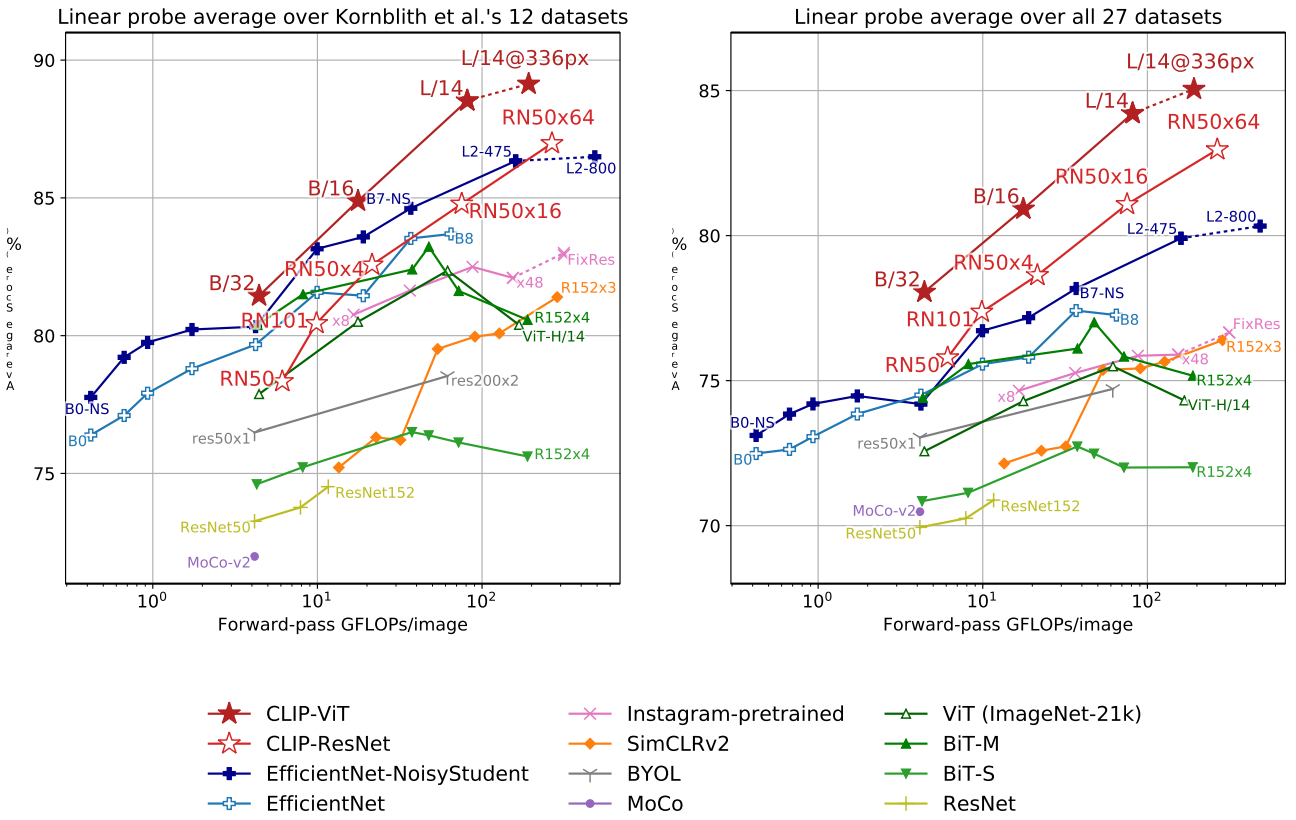

图10. CLIP模型与前沿计算机视觉模型的线性探针性能对比，包括EfficientNet（Tan & Le, 2019; Xie et al., 2020）、MoCo（Chen et al., 2020d）、Instagram预训练的ResNeXt模型（Mahajan et al., 2018; Touvron et al., 2019）、BiT（Kolesnikov et al., 2019）、ViT（Dosovitskiy et al., 2020）、SimCLRv2（Chen et al., 2020c）、BYOL（Grill et al., 2020）以及原始ResNet模型（He et al., 2016b）。（左）分数为Kornblith等人（2019）研究的12个数据集上的平均值。（右）分数为涵盖更广泛分布类型的27个数据集上的平均值。虚线表示在高于预训练分辨率图像上进行微调或评估的模型。各数据集详细分数见表10，各数据集性能曲线见图20。

在这个更广泛的评估套件中，CLIP的优势更为明显。所有CLIP模型，无论规模大小，在计算效率方面都优于所有评估系统。最佳模型的平均得分相较于先前系统的提升从2.6%增加到了5%。我们还发现，自监督系统在我们更广泛的评估套件中表现明显更好。例如，尽管SimCLRv2在Kornblith等人（2019）的12个数据集上平均表现仍不及BiT-M，但在我们的27个数据集评估套件中，SimCLRv2超越了BiT-M。这些发现表明，继续扩展任务多样性和覆盖范围有助于更好地理解系统的“通用”性能。我们认为，沿着VTAB方向的进一步评估工作将具有重要价值。

除了上述的总体分析外，我们在图11中可视化了所有27个数据集上最佳CLIP模型与我们评估套件中最佳模型之间的性能差异。在27个数据集中，CLIP在21个数据集上表现优于Noisy Student EfficientNet-L2。CLIP在需要OCR的任务（如SST2）上提升最为显著。

以及HatefulMemes）、地理定位与场景识别（Country211、SUN397），以及视频中的活动识别（Kinetics700和UCF101）。此外，CLIP在细粒度汽车和交通标志识别（Stanford Cars和GTSRB）上也表现更佳。这可能反映了ImageNet中监督过于狭窄的问题。例如，在GTSRB上14.7%的改进可能暗示了ImageNet-1K的一个问题，该数据集对所有交通和街道标志仅使用单一标签。这可能导致监督式表征压缩类内细节，从而损害细粒度下游任务的准确性。如前所述，CLIP在多个数据集上仍落后于EfficientNet。不出所料，EfficientNet相对于CLIP表现最佳的数据集正是其训练所用的数据集：ImageNet。在低分辨率数据集如CIFAR10和CIFAR100上，EfficientNet也略微优于CLIP。我们怀疑这至少部分归因于CLIP缺乏基于尺度的数据增强。EfficientNet在PatchCamelyon和CLEVRCounts上也略胜一筹，这些数据集的整体性能仍然

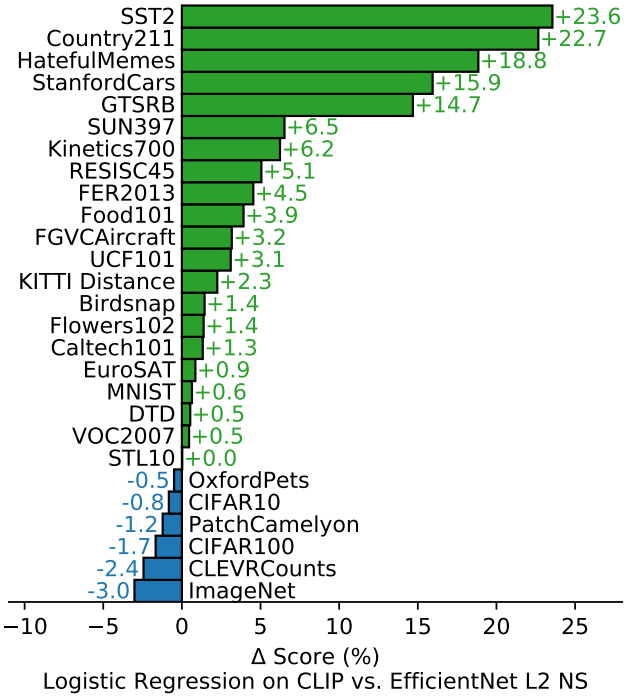

Figure 11. **CLIP's features outperform the features of the best ImageNet model on a wide variety of datasets.** Fitting a linear classifier on CLIP's features outperforms using the Noisy Student EfficientNet-L2 on 21 out of 27 datasets.

low for both approaches.

3.3. Robustness to Natural Distribution Shift

In 2015, it was announced that a deep learning model exceeded human performance on the ImageNet test set (He et al., 2015). However, research in the subsequent years has repeatedly found that these models still make many simple mistakes (Dodge & Karam, 2017; Geirhos et al., 2018; Alcorn et al., 2019), and new benchmarks testing these systems has often found their performance to be much lower than both their ImageNet accuracy and human accuracy (Recht et al., 2019; Barbu et al., 2019). What explains this discrepancy? Various ideas have been suggested and studied (Ilyas et al., 2019; Geirhos et al., 2020). A common theme of proposed explanations is that deep learning models are exceedingly adept at finding correlations and patterns which hold across their training dataset and thus improve in-distribution performance. However many of these correlations and patterns are actually spurious and do not hold for other distributions and result in large drops in performance on other datasets.

We caution that, to date, most of these studies limit their evaluation to models trained on ImageNet. Recalling the topic of discussion, it may be a mistake to generalize too far from these initial findings. To what degree are these failures attributable to deep learning, ImageNet, or some combination of the two? CLIP models, which are trained via natural language supervision on a very large dataset and are capable of high zero-shot performance, are an opportunity to investigate this question from a different angle.

Taori et al. (2020) is a recent comprehensive study moving towards quantifying and understanding these behaviors for ImageNet models. Taori et al. (2020) study how the performance of ImageNet models change when evaluated on *natural distribution shifts*. They measure performance on a set of 7 distribution shifts: ImageNetV2 (Recht et al., 2019), ImageNet Sketch (Wang et al., 2019), Youtube-BB and ImageNet-Vid (Shankar et al., 2019), ObjectNet (Barbu et al., 2019), ImageNet Adversarial (Hendrycks et al., 2019), and ImageNet Rendition (Hendrycks et al., 2020a). They distinguish these datasets, which all consist of novel images collected from a variety of sources, from *synthetic distribution shifts* such as ImageNet-C (Hendrycks & Dietterich, 2019), Stylized ImageNet (Geirhos et al., 2018), or adversarial attacks (Goodfellow et al., 2014) which are created by perturbing existing images in various ways. They propose this distinction because in part because they find that while several techniques have been demonstrated to improve performance on synthetic distribution shifts, they often fail to yield consistent improvements on natural distributions.[3]

Across these collected datasets, the accuracy of ImageNet models drop well below the expectation set by the ImageNet validation set. For the following summary discussion we report average accuracy across all 7 natural distribution shift datasets and average accuracy across the corresponding class subsets of ImageNet unless otherwise specified. Additionally, for Youtube-BB and ImageNet-Vid, which have two different evaluation settings, we use the average of pm-0 and pm-10 accuracy.

A ResNet-101 makes 5 times as many mistakes when evaluated on these natural distribution shifts compared to the ImageNet validation set. Encouragingly however, Taori et al. (2020) find that accuracy under distribution shift increases predictably with ImageNet accuracy and is well modeled as a linear function of logit-transformed accuracy. Taori et al. (2020) use this finding to propose that robustness analysis should distinguish between *effective* and *relative* robustness. Effective robustness measures improvements in accuracy under distribution shift above what is predicted by the documented relationship between in-distribution and out-of-distribution accuracy. Relative robustness captures any improvement in out-of-distribution accuracy. Taori et al. (2020) argue that robustness techniques should aim to improve both effective robustness and relative robustness.

Almost all models studied in Taori et al. (2020) are trained

[3]We refer readers to Hendrycks et al. (2020a) for additional experiments and discussion on this claim.

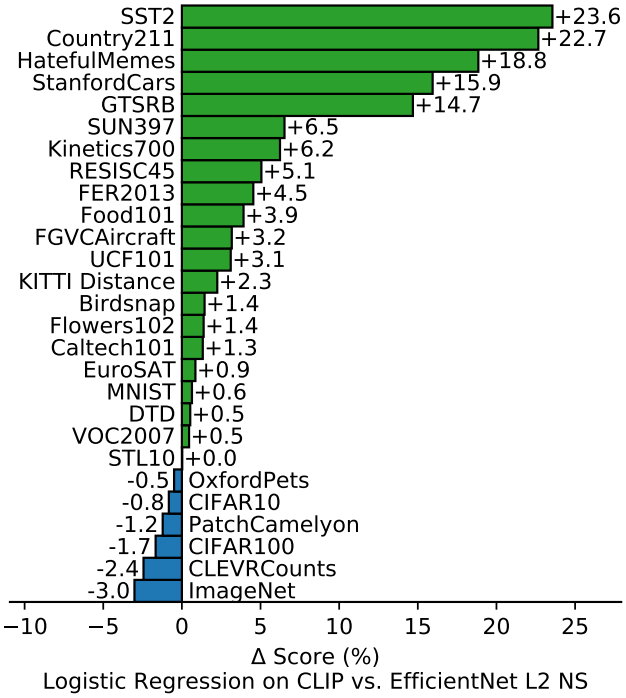

图11. CLIP的特征在多种数据集上表现优于最佳ImageNet模型的特征。在27个数据集中的21个上，基于CLIP特征拟合的线性分类器性能优于使用Noisy Student EfficientNet-L2模型。

两种方法的效果都不理想。

3.3. 对自然分布偏移的鲁棒性

2015年，有研究宣布深度学习模型在ImageNet测试集上的表现超越了人类（He等人，2015）。然而，随后几年的研究反复发现，这些模型仍会犯许多简单的错误（Dodge & Karam, 2017; Geirhos等人，2018; Alcorn等人，2019），而针对这些系统设计的新基准测试往往显示其性能远低于它们在ImageNet上的准确率及人类准确率（Recht等人，2019; Barbu等人，2019）。如何解释这种差异？学界已提出并研究了多种观点（Ilyas等人，2019; Geirhos等人，2020）。这些解释的一个共同主题是：深度学习模型极其擅长发现训练数据集中普遍存在的相关性和模式，从而提升在分布内数据上的性能。然而，这些相关性和模式中有许多实际上只是虚假关联，并不适用于其他数据分布，导致模型在其他数据集上出现性能大幅下降。

我们提醒，迄今为止，这些研究大多将评估局限于在ImageNet上训练的模型。回顾讨论的主题，从这些初步发现中过度推广可能是一个错误。这些失败在多大程度上可归因于深度学习、ImageNet，或是某些两者的结合？CLIP模型通过在海量数据集上进行自然语言监督训练，具备出色的零样本性能，这为我们从不同角度探究这一问题提供了契机。

Taori等人（2020）近期开展了一项综合性研究，致力于量化和理解ImageNet模型在这些行为上的表现。Taori等人（2020）研究了ImageNet模型在*natural distribution shifts*上评估时性能如何变化。他们测量了模型在7种分布偏移数据集上的性能：ImageNet V2（Recht等人，2019）、ImageNet Sketch（Wang等人，2019）、Youtube-BB与ImageNet-Vid（Shankar等人，2019）、ObjectNet（Barbu等人，2019）、ImageNet Adversarial（Hendrycks等人，2019）以及ImageNet Rendition（Hendrycks等人，2020a）。他们将这些从多种来源收集的新图像构成的数据集，与*synthetic distribution shifts*（例如ImageNet-C（Hendrycks & Dietterich, 2019）、风格化ImageNet（Geirhos等人，2018）或对抗攻击（Goodfellow等人，2014））区分开来——后者是通过以各种方式扰动现有图像生成的。他们提出这种区分，部分原因是发现尽管已有多种技术被证明能提升模型在合成分布偏移上的性能，但这些技术往往无法在自然分布偏移上带来一致的改进。[3]

在这些收集的数据集中，ImageNet模型的准确率远低于ImageNet验证集设定的预期。在接下来的总结讨论中，除非另有说明，我们将报告所有7个自然分布偏移数据集的平均准确率，以及ImageNet对应类别子集的平均准确率。此外，对于Youtube-BB和ImageNet-Vid这两种具有不同评估设置的数据集，我们采用pm-0和pm-10准确率的平均值。

ResNet-101 在这些自然分布偏移上的评估错误率是 ImageNet 验证集上的 5 倍。然而令人鼓舞的是，Taori 等人（2020）发现，分布偏移下的准确率会随 ImageNet 准确率可预测地提升，并且在对数几率变换后的准确率上能很好地用线性函数建模。Taori 等人（2020）利用这一发现提出，鲁棒性分析应区分 *effective* 鲁棒性与 *relative* 鲁棒性。有效鲁棒性衡量的是分布偏移下准确率的提升幅度超出原始分布与分布外准确率之间已知关系所预测的部分；相对鲁棒性则捕捉分布外准确率的任何改进。Taori 等人（2020）主张，鲁棒性技术应致力于同时提升有效鲁棒性与相对鲁棒性。

Taori等人（2020）研究中几乎所有模型都经过训练

[3]We refer readers to Hendrycks et al. (2020a) for additional experiments and discussion on this claim.

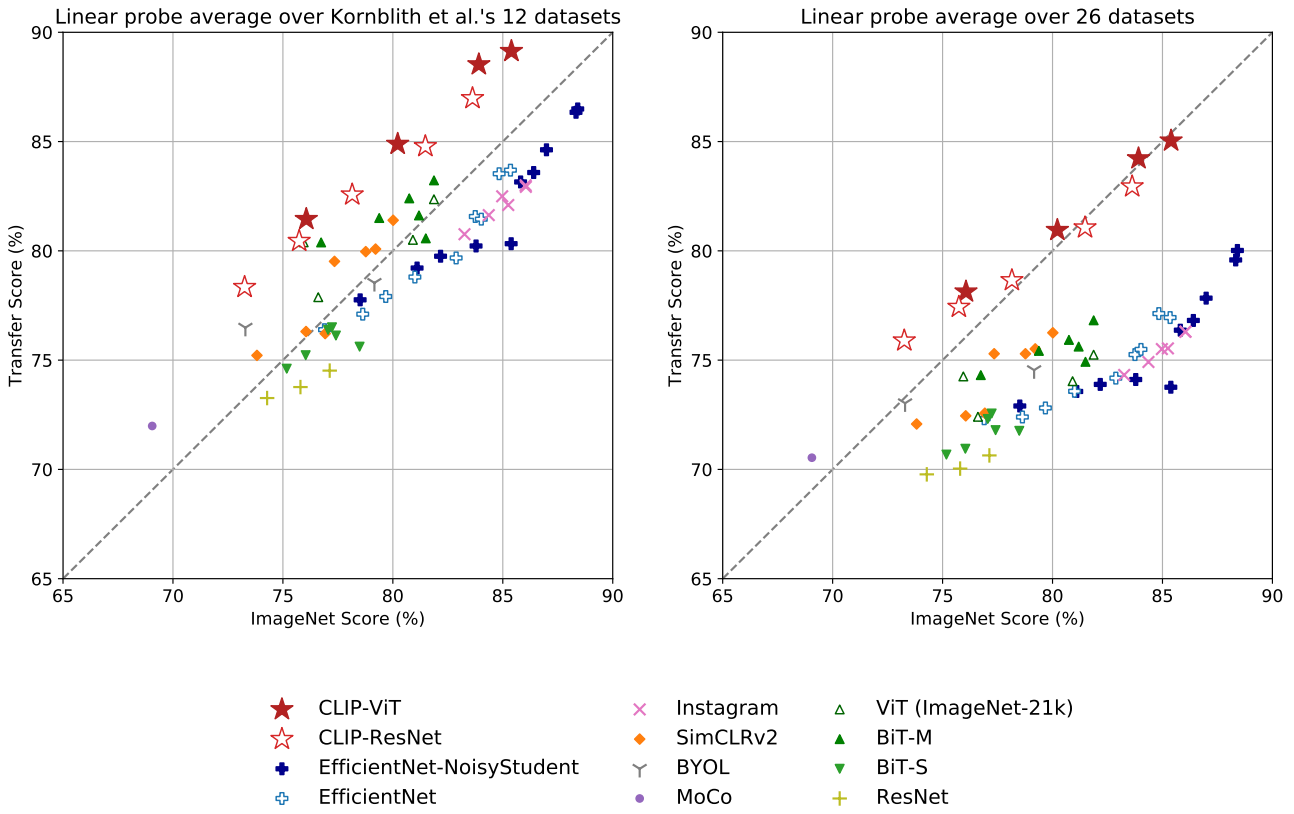

Figure 12. **CLIP's features are more robust to task shift when compared to models pre-trained on ImageNet.** For both dataset splits, the transfer scores of linear probes trained on the representations of CLIP models are higher than other models with similar ImageNet performance. This suggests that the representations of models trained on ImageNet are somewhat overfit to their task.

or fine-tuned on the ImageNet dataset. Returning to the discussion in the introduction to this section - is training or adapting to the ImageNet dataset distribution the cause of the observed robustness gap? Intuitively, a zero-shot model should not be able to exploit spurious correlations or patterns that hold only on a specific distribution, since it is not trained on that distribution. [4] Thus it is reasonable to expect zero-shot models to have much higher effective robustness. In Figure 13, we compare the performance of zero-shot CLIP with existing ImageNet models on natural distribution shifts. All zero-shot CLIP models improve effective robustness by a large amount and reduce the size of the gap between ImageNet accuracy and accuracy under distribution shift by up to 75%.

While these results show that zero-shot models can be much more robust, they do not necessarily mean that supervised learning on ImageNet causes a robustness gap. Other details of CLIP, such as its large and diverse pre-training dataset or use of natural language supervision could also result in much more robust models regardless of whether they are zero-shot or fine-tuned. As an initial experiment to potentially begin narrowing this down, we also measure how the performance of CLIP models change after adapting to the ImageNet distribution via a L2 regularized logistic regression classifier fit to CLIP features on the ImageNet training set. We visualize how performance changes from the zero-shot classifier in Figure 14. Although adapting CLIP to the ImageNet distribution increases its ImageNet accuracy by 9.2% to 85.4% overall, and ties the accuracy of the 2018 SOTA from Mahajan et al. (2018), *average accuracy under distribution shift slightly decreases*.

It is surprising to see a 9.2% increase in accuracy, which corresponds to roughly 3 years of improvement in SOTA, fail to translate into any improvement in average performance under distribution shift. We also break down the differences between zero-shot accuracy and linear classifier accuracy per dataset in Figure 14 and find performance still increases significantly on one dataset, ImageNetV2. ImageNetV2 closely followed the creation process of the original ImageNet dataset which suggests that gains in accuracy from supervised adaptation are closely concentrated around the ImageNet distribution. Performance decreases by 4.7% on

[4] We caution that a zero-shot model can still exploit spurious correlations that are shared between the pre-training and evaluation distributions.

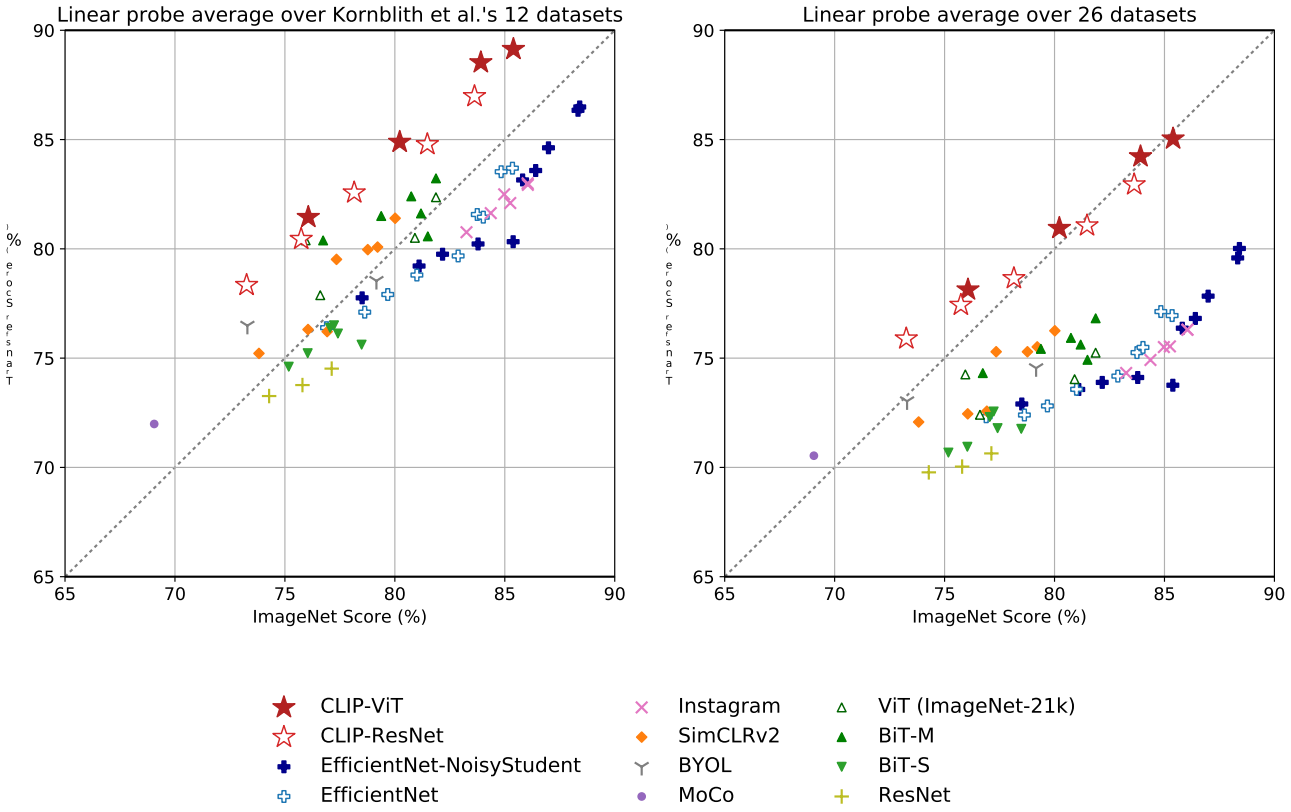

图12. 与在ImageNet上预训练的模型相比，CLIP的特征对任务迁移具有更强的鲁棒性。在两种数据集划分下，基于CLIP模型表征训练的线性探测器的迁移分数均高于ImageNet性能相近的其他模型。这表明在ImageNet上训练的模型表征存在一定程度的任务过拟合。

或是在ImageNet数据集上进行微调。回到本节引言中的讨论——训练或适应ImageNet数据集分布是否是观察到鲁棒性差距的原因？直观来看，零样本模型应当无法利用仅在特定分布中存在的虚假关联或模式，因为它并未在该分布上接受训练。{v*} 因此我们有理由期待零样本模型具备显著更高的有效鲁棒性。在图13中，我们将零样本CLIP与现有ImageNet模型在自然分布偏移下的性能进行了对比。所有零样本CLIP模型均大幅提升了有效鲁棒性，并将ImageNet准确率与分布偏移下准确率之间的差距缩小了最高达75%。

尽管这些结果表明零样本模型可以更加稳健，但这并不一定意味着在ImageNet上的监督学习会导致稳健性差距。CLIP的其他细节，例如其庞大且多样化的预训练数据集或使用自然语言监督，也可能导致这一结果。

[4]We caution that a zero-shot model can still exploit spurious correlations that are shared between the pre-training and evaluation distributions.

在更为稳健的模型中，无论它们是零样本还是经过微调的，情况都是如此。作为一项可能开始缩小这一差距的初步实验，我们还测量了CLIP模型在通过L2正则化逻辑回归分类器适应ImageNet分布后的性能变化，该分类器是基于ImageNet训练集上的CLIP特征进行拟合的。我们在图14中展示了性能从零样本分类器开始的变化情况。尽管将CLIP适应到ImageNet分布使其ImageNet整体准确率提高了9.2%，达到85.4%，并与Mahajan等人（2018）2018年的SOTA准确率持平，*average accuracy under distribution shift slightly decreases*。

令人惊讶的是，准确率提升了9.2%——这大致相当于SOTA（当前最优技术）三年的进步幅度——却未能转化为分布偏移下平均性能的任何改善。我们还在图14中按数据集细分了零样本准确率与线性分类器准确率之间的差异，发现仅在一个数据集ImageNetV2上性能仍有显著提升。ImageNetV2严格遵循了原始ImageNet数据集的创建流程，这表明监督式适应带来的准确率增益高度集中在ImageNet分布范围内。在

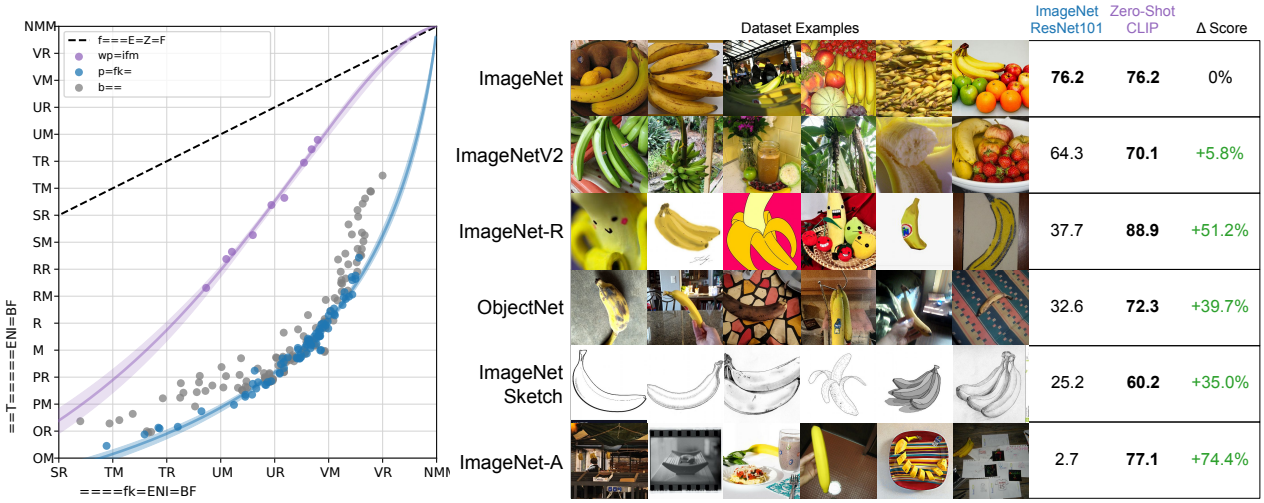

Figure 13. **Zero-shot CLIP is much more robust to distribution shift than standard ImageNet models.** (Left) An ideal robust model (dashed line) performs equally well on the ImageNet distribution and on other natural image distributions. Zero-shot CLIP models shrink this "robustness gap" by up to 75%. Linear fits on logit transformed values are shown with bootstrap estimated 95% confidence intervals. (Right) Visualizing distribution shift for bananas, a class shared across 5 of the 7 natural distribution shift datasets. The performance of the best zero-shot CLIP model, ViT-L/14@336px, is compared with a model that has the same performance on the ImageNet validation set, ResNet-101.

ImageNet-R, 3.8% on ObjectNet, 2.8% on ImageNet Sketch, and 1.9% on ImageNet-A. The change in accuracy on the two other datasets, Youtube-BB and ImageNet Vid, is insignificant.

How is it possible to improve accuracy by 9.2% on the ImageNet dataset with little to no increase in accuracy under distribution shift? Is the gain primarily from "exploiting spurious correlations"? Is this behavior unique to some combination of CLIP, the ImageNet datatset, and the distribution shifts studied, or a more general phenomena? Does it hold for end-to-end finetuning as well as linear classifiers? We do not have confident answers to these questions at this time. Prior work has also pre-trained models on distributions other than ImageNet, but it is common to study and release models only after they have been fine-tuned to ImageNet. As a step towards understanding whether pre-trained zero-shot models consistently have higher effective robustness than fine-tuned models, we encourage the authors of Mahajan et al. (2018), Kolesnikov et al. (2019), and Dosovitskiy et al. (2020) to, if possible, study these questions on their models as well.

We also investigate another robustness intervention enabled by flexible zero-shot natural-language-based image classifiers. The target classes across the 7 transfer datasets are not always perfectly aligned with those of ImageNet. Two datasets, Youtube-BB and ImageNet-Vid, consist of superclasses of ImageNet. This presents a problem when trying to use the fixed 1000-way classifier of an ImageNet model to make predictions. Taori et al. (2020) handle this by max-pooling predictions across all sub-classes according to the ImageNet class hierarchy. Sometimes this mapping is much less than perfect. For the person class in Youtube-BB, predictions are made by pooling over the ImageNet classes for a baseball player, a bridegroom, and a scuba diver. With CLIP we can instead generate a custom zero-shot classifier for each dataset directly based on its class names. In Figure 14 we see that this improves average effective robustness by 5% but is concentrated in large improvements on only a few datasets. Curiously, accuracy on ObjectNet also increases by 2.3%. Although the dataset was designed to closely overlap with ImageNet classes, using the names provided for each class by ObjectNet's creators still helps a small amount compared to using ImageNet class names and pooling predictions when necessary.

While zero-shot CLIP improves effective robustness, Figure 14 shows that the benefit is almost entirely gone in a fully supervised setting. To better understand this difference, we investigate how effective robustness changes on the continuum from zero-shot to fully supervised. In Figure 15 we visualize the performance of 0-shot, 1-shot, 2-shot, 4-shot ..., 128-shot, and fully supervised logistic regression classifiers on the best CLIP model's features. We see that while few-shot models also show higher effective robustness than existing models, this benefit fades as in-distribution performance increases with more training data and is mostly, though not entirely, gone for the fully supervised model. Additionally, zero-shot CLIP is notably more robust than a few-shot model with equivalent ImageNet performance.

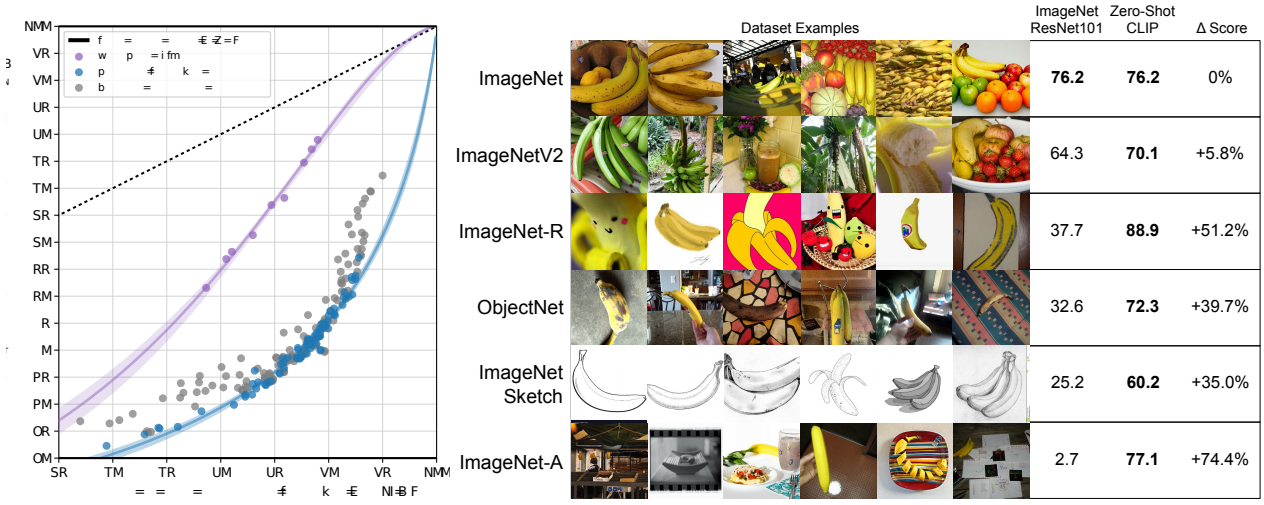

图13. 零样本CLIP模型比标准ImageNet模型对分布偏移具有更强的鲁棒性。（左）理想的鲁棒模型（虚线）在ImageNet分布和其他自然图像分布上表现同样出色。零样本CLIP模型将这种“鲁棒性差距”缩小了高达75%。图中展示了逻辑值转换后的线性拟合结果，并附有自助法估计的95%置信区间。（右）以香蕉类别为例可视化分布偏移情况，该类别在7个自然分布偏移数据集中的5个里均存在。性能最佳的零样本CLIP模型ViT-L/14@336px与在ImageNet验证集上表现相同的ResNet-101模型进行了对比。

ImageNet-R上为3.8%，ObjectNet上为3.8%，ImageNet Sketch上为2.8%，ImageNet-A上为1.9%。在另外两个数据集Youtube-BB和ImageNet Vid上的准确率变化不显著。

在ImageNet数据集上准确率提升9.2%，但在分布偏移下准确率几乎未增长，这如何实现？这种提升主要源于“利用虚假相关性”吗？这种现象是CLIP、ImageNet数据集与所研究分布偏移的特有组合产物，还是更普遍的现象？该结论是否同时适用于端到端微调与线性分类器？目前我们对此尚无明确答案。先前研究虽已在非ImageNet分布上预训练模型，但通常仅在模型经ImageNet微调后才进行研究与发布。为探究预训练零样本模型是否始终比微调模型具有更高的有效鲁棒性，我们鼓励Mahajan等人（2018）、Kolesnikov等人（2019）及Dosovitskiy等人（2020）的作者在可能的情况下，基于各自模型对这些问题展开进一步研究。

我们还研究了另一种鲁棒性干预措施，该措施通过灵活的零样本自然语言图像分类器实现。七个迁移数据集中的目标类别并不总是与ImageNet的类别完全一致。其中两个数据集，Youtube-BB和ImageNet-Vid，包含ImageNet的超类。当尝试使用ImageNet模型的固定1000路分类器进行预测时，这会带来问题。Taori等人（2020）通过最大-

根据ImageNet类别层次结构，对所有子类别的预测结果进行汇总。有时这种映射远非完美。以Youtube-BB数据集中的人物类别为例，其预测是通过汇总ImageNet中棒球运动员、新郎和潜水员等类别的结果得出的。而借助CLIP，我们可以直接基于每个数据集的类别名称生成定制化的零样本分类器。图14显示，这种方法将平均有效鲁棒性提升了5%，但改进主要集中在少数数据集上，且幅度较大。有趣的是，ObjectNet的准确率也提高了2.3%。尽管该数据集的设计初衷是与ImageNet类别高度重合，但使用ObjectNet创建者提供的类别名称，相比必要时使用ImageNet类别名称并汇总预测，仍能带来小幅提升。

尽管零样本CLIP提升了有效鲁棒性，但图14显示在完全监督设置下这一优势几乎完全消失。为了更好地理解这种差异，我们研究了从零样本到完全监督的连续过程中有效鲁棒性的变化。在图15中，我们可视化了基于最佳CLIP模型特征的0样本、1样本、2样本、4样本……128样本及完全监督逻辑回归分类器的性能。我们发现，虽然少样本模型也展现出比现有模型更高的有效鲁棒性，但随着训练数据增加带来的分布内性能提升，这种优势逐渐减弱；对于完全监督模型，该优势虽未完全消失但已大幅衰减。此外，在ImageNet性能相当的情况下，零样本CLIP明显比少样本模型更具鲁棒性。

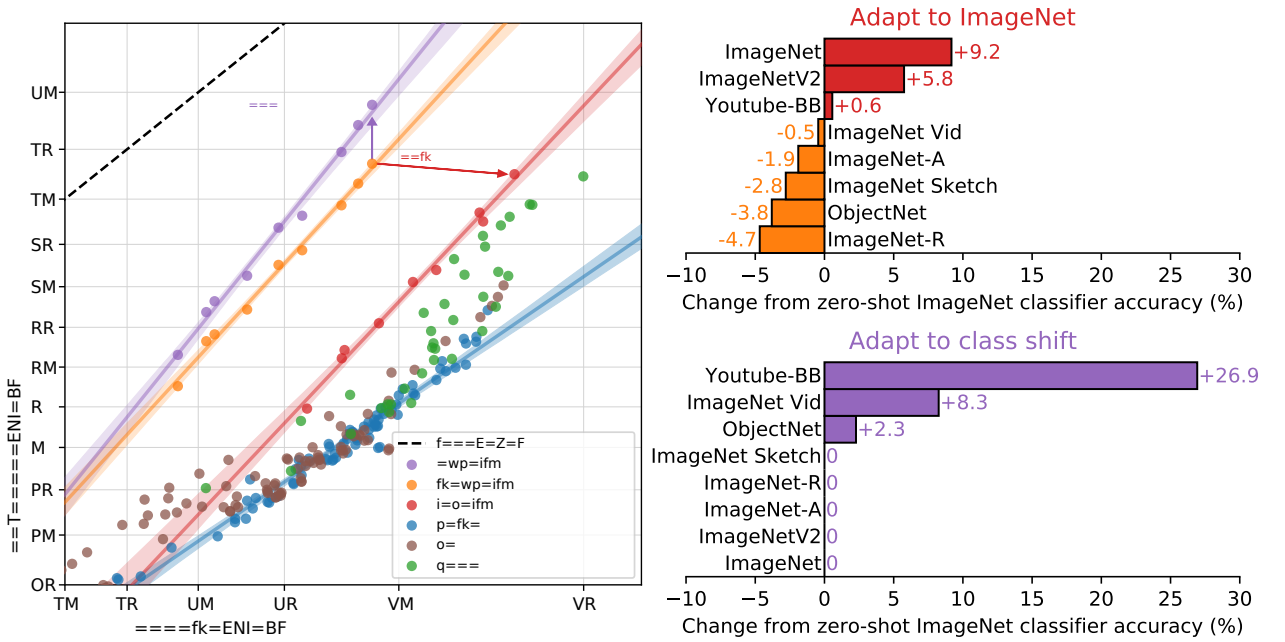

Figure 14. **While supervised adaptation to ImageNet increases ImageNet accuracy by 9.2%, it slightly reduces average robustness.** (Left) Customizing zero-shot CLIP to each dataset improves robustness compared to using a single static zero-shot ImageNet classifier and pooling predictions across similar classes as in Taori et al. (2020). CLIP models adapted to ImageNet have similar effective robustness as the best prior ImageNet models. (Right) Details of per dataset changes in accuracy for the two robustness interventions. Adapting to ImageNet increases accuracy on ImageNetV2 noticeably but trades off accuracy on several other distributions. Dataset specific zero-shot classifiers can improve accuracy by a large amount but are limited to only a few datasets that include classes which don't perfectly align with ImageNet categories.

Across our experiments, high effective robustness seems to result from minimizing the amount of distribution specific training data a model has access to, but this comes at a cost of reducing dataset-specific performance.

Taken together, these results suggest that the recent shift towards large-scale task and dataset agnostic pre-training combined with a reorientation towards zero-shot and few-shot benchmarking on broad evaluation suites (as advocated by Yogatama et al. (2019) and Linzen (2020)) promotes the development of more robust systems and provides a more accurate assessment of performance. We are curious to see if the same results hold for zero-shot models in the field of NLP such as the GPT family. While Hendrycks et al. (2020b) has reported that pre-training improves relative robustness on sentiment analysis, Miller et al. (2020)'s study of the robustness of question answering models under natural distribution shift finds, similar to Taori et al. (2020), little evidence of effective robustness improvements to date.

4. Comparison to Human Performance

How does CLIP compare to human performance and human learning? To get a better understanding of how well humans perform in similar evaluation settings to CLIP, we evaluated humans on one of our tasks. We wanted to get a sense of how strong human zero-shot performance is at these tasks, and how much human performance is improved if they are shown one or two image samples. This can help us to compare task difficulty for humans and CLIP, and identify correlations and differences between them.

We had five different humans look at each of 3669 images in the test split of the Oxford IIT Pets dataset (Parkhi et al., 2012) and select which of the 37 cat or dog breeds best matched the image (or 'I don't know' if they were completely uncertain). In the zero-shot case the humans were given no examples of the breeds and asked to label them to the best of their ability without an internet search. In the one-shot experiment the humans were given one sample image of each breed and in the two-shot experiment they were given two sample images of each breed.[5]

One possible concern was that the human workers were not sufficiently motivated in the zero-shot task. High human accuracy of 94% on the STL-10 dataset (Coates et al., 2011)

[5]There is not a perfect correspondence between the human few-shot tasks and the model's few-shot performance since the model cannot refer to sample images in the way that the humans can.

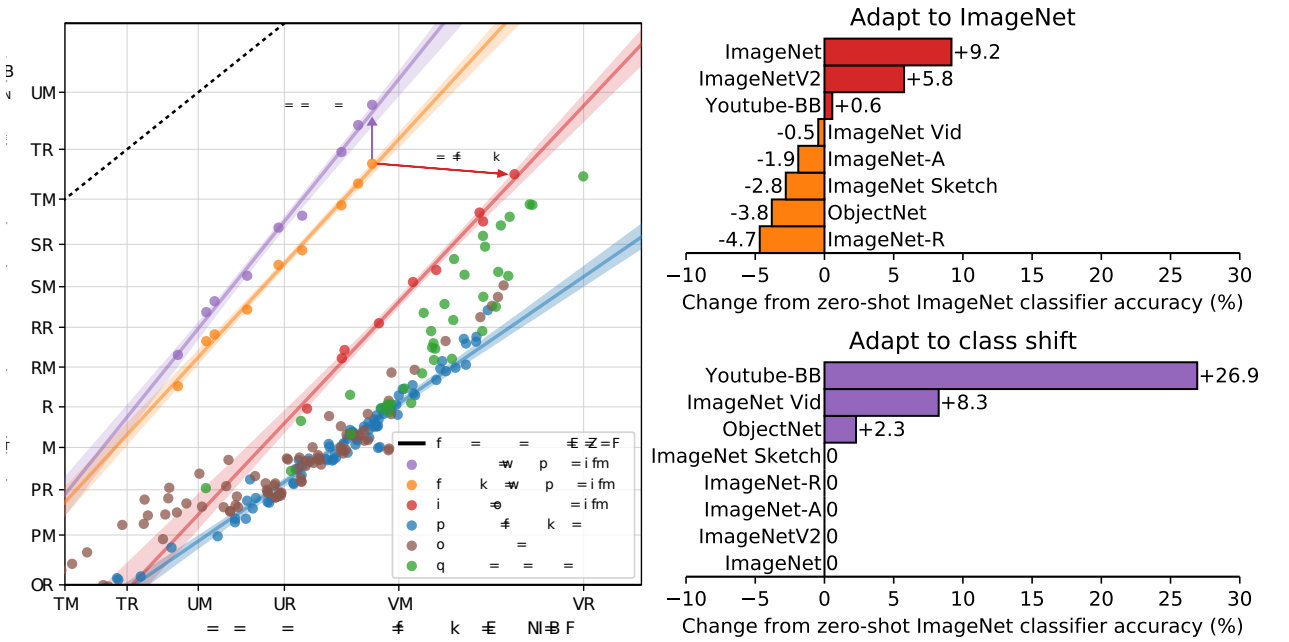

图14. 尽管通过监督式适应ImageNet将ImageNet准确率提升了9.2%，但这略微降低了平均鲁棒性。（左）相较于使用单一静态零样本ImageNet分类器并像Taori等人（2020）那样对相似类别预测进行池化，针对每个数据集定制零样本CLIP能提升鲁棒性。适应ImageNet的CLIP模型与先前最佳的ImageNet模型具有相近的有效鲁棒性。（右）两种鲁棒性干预措施在各数据集上准确率变化的具体细节。适应ImageNet显著提高了ImageNetV2的准确率，但牺牲了在其他若干分布上的准确率。针对特定数据集的零样本分类器可大幅提升准确率，但仅适用于少数类别与ImageNet分类不完全对齐的数据集。

在我们的实验中，高有效鲁棒性似乎源于最小化模型可获取的特定分布训练数据量，但这会以降低数据集特定性能为代价。

综上所述，这些结果表明，近期向大规模任务和数据集无关的预训练转变，结合对广泛评估套件上零样本和少样本基准测试的重新定位（如Yogatama等人（2019）和Linzen（2020）所倡导的），促进了更鲁棒系统的开发，并提供了更准确的性能评估。我们好奇地想知道，在NLP领域（如GPT系列）的零样本模型是否也会出现相同的结果。尽管Hendrycks等人（2020b）报告称预训练提高了情感分析的相关鲁棒性，但Miller等人（2020）对自然分布变化下问答模型鲁棒性的研究发现，与Taori等人（2020）类似，迄今为止几乎没有证据表明有效的鲁棒性改进。

4. 与人类表现对比

CLIP与人类表现及人类学习相比如何？为了更好地理解人类在与CLIP类似的评估环境中的表现，我们评估了

在我们的一项任务中，我们想了解人类在这些任务上的零样本表现有多强，以及如果向他们展示一两个图像样本，人类的表现会提高多少。这有助于我们比较人类和CLIP的任务难度，并识别它们之间的相关性和差异。

我们邀请了五位不同的人员，查看了牛津IIT宠物数据集测试集中（Parkhi等人，2012）的3669张图像，并为每张图像从37个猫或狗品种中选择最匹配的品种（若完全不确定则选择“我不知道”）。在零样本情况下，参与者未获得任何品种示例，仅凭自身能力进行标注，且未进行网络搜索。在单样本实验中，参与者获得了每个品种的一张示例图像；在双样本实验中，则获得了每个品种的两张示例图像。{v*}

一个可能的担忧是，在零样本任务中，人类工作者没有得到足够的激励。在STL-10数据集上，人类达到了94%的高准确率（Coates等人，2011年）

[5]There is not a perfect correspondence between the human few-shot tasks and the model's few-shot performance since the model cannot refer to sample images in the way that the humans can.

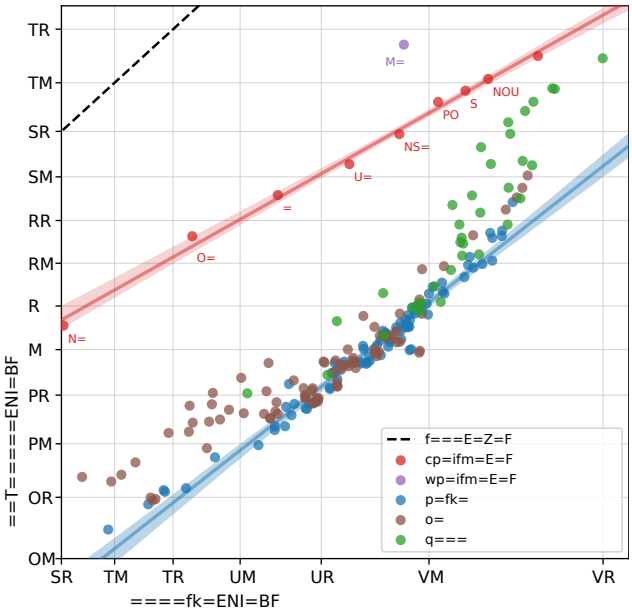

Figure 15. **Few-shot CLIP also increases effective robustness compared to existing ImageNet models but is less robust than zero-shot CLIP.** Minimizing the amount of ImageNet training data used for adaption increases effective robustness at the cost of decreasing relative robustness. 16-shot logistic regression CLIP matches zero-shot CLIP on ImageNet, as previously reported in Figure 7, but is less robust.

and 97-100% accuracy on the subset of attention check images increased our trust in the human workers.

Interestingly, humans went from a performance average of 54% to 76% with just one training example per class, and the marginal gain from an additional training example is minimal. The gain in accuracy going from zero to one shot is almost entirely on images that humans were uncertain about. This suggests that humans "know what they don't know" and are able to update their priors on the images they are most uncertain in based on a single example. Given this, it seems that while CLIP is a promising training strategy for zero-shot performance (Figure 5) and does well on tests of natural distribution shift (Figure 13), there is a large difference between how humans learn from a few examples and the few-shot methods in this paper.

This suggests that there are still algorithmic improvements waiting to be made to decrease the gap between machine and human sample efficiency, as noted by Lake et al. (2016) and others. Because these few-shot evaluations of CLIP don't make effective use of prior knowledge and the humans do, we speculate that finding a method to properly integrate prior knowledge into few-shot learning is an important step in algorithmic improvements to CLIP. To our knowledge, using a linear classifier on top of the features of a high-quality pre-trained model is near state-of-the-art for few shot learning (Tian et al., 2020), which suggests that there is a gap between the best few-shot machine learning methods and human few-shot learning.

	Accuracy	Majority Vote on Full Dataset	Accuracy on Guesses	Majority Vote Accuracy on Guesses
Zero-shot human	53.7	57.0	69.7	63.9
Zero-shot CLIP	**93.5**	**93.5**	**93.5**	**93.5**
One-shot human	75.7	80.3	78.5	81.2
Two-shot human	75.7	85.0	79.2	86.1

Table 2. Comparison of human performance on Oxford IIT Pets. As in Parkhi et al. (2012), the metric is average per-class classification accuracy. Most of the gain in performance when going from the human zero shot case to the human one shot case is on images that participants were highly uncertain on. "Guesses" refers to restricting the dataset to where participants selected an answer other than "I don't know", the "majority vote" is taking the most frequent (exclusive of ties) answer per image.

If we plot human accuracy vs CLIP's zero shot accuracy (Figure 16), we see that the hardest problems for CLIP are also hard for humans. To the extent that errors are consistent, our hypothesis is that this is due to at least a two factors: noise in the dataset (including mislabeled images) and out of distribution images being hard for both humans and models.

5. Data Overlap Analysis

A concern with pre-training on a very large internet dataset is unintentional overlap with downstream evals. This is important to investigate since, in a worst-case scenario, a complete copy of an evaluation dataset could leak into the pre-training dataset and invalidate the evaluation as a meaningful test of generalization. One option to prevent this is to identify and remove all duplicates before training a model. While this guarantees reporting true hold-out performance, it requires knowing all possible data which a model might be evaluated on ahead of time. This has the downside of limiting the scope of benchmarking and analysis. Adding a new evaluation would require an expensive re-train or risk reporting an un-quantified benefit due to overlap.

Instead, we document how much overlap occurs and how performance changes due to these overlaps. To do this, we use the following procedure:

1) For each evaluation dataset, we run a duplicate detector (see Appendix C) on its examples. We then manually inspect the found nearest neighbors and set a per dataset threshold to keep high precision while maximizing recall. Using this threshold, we then create two new subsets, `Overlap`, which contains all examples which have a similarity to a training example above the threshold, and `Clean`, which

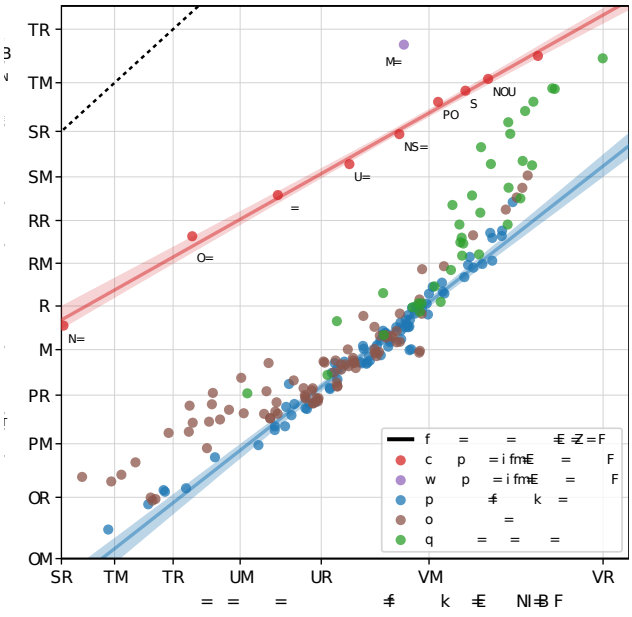

图15。与现有的ImageNet模型相比，少样本CLIP也提升了有效鲁棒性，但其鲁棒性仍低于零样本CLIP。减少用于适配的ImageNet训练数据量会以降低相对鲁棒性为代价来提升有效鲁棒性。如图7先前所示，16样本逻辑回归CLIP在ImageNet上与零样本CLIP表现相当，但鲁棒性较弱。

并且在注意力检查图像子集上达到97-100%的准确率，增强了我们对人工标注者的信任。

有趣的是，人类仅通过每个类别一个训练示例，就从平均54%的表现提升到了76%，而额外训练示例带来的边际增益微乎其微。从零样本到单样本的准确率提升，几乎完全体现在人类原本不确定的图像上。这表明人类“知道自己不知道什么”，并且能够基于单个示例，更新他们最不确定图像的先验认知。鉴于此，尽管CLIP是一种前景广阔的零样本性能训练策略（图5），并且在自然分布偏移测试中表现良好（图13），但人类从少量示例中学习的方式与本文中的少样本方法之间仍存在巨大差异。

这表明，正如Lake等人（2016年）及其他研究者所指出的，仍有算法改进的空间，以缩小机器与人类在样本效率上的差距。由于CLIP的这些少样本评估未能有效利用先验知识，而人类却能做到，我们推测，找到一种将先验知识恰当整合到少样本学习中的方法，是CLIP算法改进的重要一步。据我们所知，在高质量特征之上使用线性分类器——

	Accuracy	Majority Vote on Full Dataset	Accuracy on Guesses	Majority Vote Accuracy on Guesses
Zero-shot human	53.7	57.0	69.7	63.9
Zero-shot CLIP	**93.5**	**93.5**	**93.5**	**93.5**
One-shot human	75.7	80.3	78.5	81.2
Two-shot human	75.7	85.0	79.2	86.1

表2. 牛津IIT宠物数据集上人类表现的比较。如Parkhi等人（2012）所述，评估指标为平均每类分类准确率。从人类零样本情况到单样本情况时，性能提升主要来自参与者高度不确定的图像。“猜测”指将数据集限制在参与者选择了“我不知道”以外答案的情况，“多数投票”指每张图像取最频繁（排除平票）的答案。

高质量预训练模型在小样本学习方面接近最先进水平（Tian等人，2020年），这表明最佳的小样本机器学习方法与人类小样本学习之间仍存在差距。

如果我们绘制人类准确率与CLIP零样本准确率的对比图（图16），会发现CLIP最棘手的难题对人类而言同样困难。在错误具有一致性的范围内，我们的假设是这至少源于两个因素：数据集中的噪声（包括错误标注的图像）以及分布外图像——这两者对人类和模型都构成挑战。

5. 数据重叠分析

在非常庞大的互联网数据集上进行预训练时，一个值得关注的问题是与下游评估数据的无意重叠。这一点至关重要，因为最坏情况下，评估数据集的完整副本可能泄露到预训练数据集中，从而使评估失去作为泛化能力有效测试的意义。防止这种情况的一种方法是在训练模型前识别并移除所有重复数据。虽然这能保证报告真实的留出性能，但需要提前知晓模型可能评估的所有数据。这种做法的缺点在于限制了基准测试和分析的范围。新增评估任务将需要昂贵的重新训练，否则可能因数据重叠而报告无法量化的收益。

相反，我们记录了重叠发生的程度以及这些重叠如何影响性能。为此，我们采用以下步骤：

1) 针对每个评估数据集，我们对其样本运行重复检测器（详见附录C）。随后人工核查发现的最近邻样本，并为每个数据集设定阈值，在保持高精度的同时最大化召回率。基于该阈值，我们创建两个新子集：重叠集（包含所有与训练样本相似度超过阈值的样本）和洁净集（包含

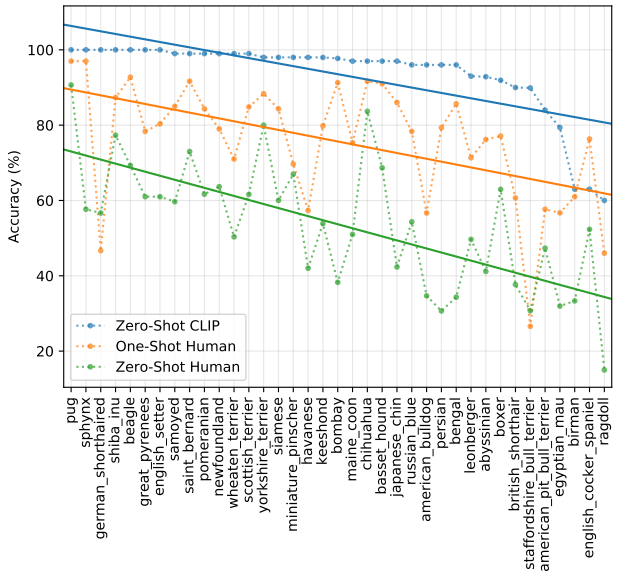

Figure 16. The hardest problems for CLIP also tend to be the hardest problems for humans. Here we rank image categories by difficulty for CLIP as measured as probability of the correct label.

contains all examples that are below this threshold. We denote the unaltered full dataset `All` for reference. From this we first record the degree of data contamination as the ratio of the number of examples in `Overlap` to the size of `All`.

2) We then compute the zero-shot accuracy of CLIP RN50x64 on the three splits and report `All - Clean` as our main metric. This is the difference in accuracy due to contamination. When positive it is our estimate of how much the overall reported accuracy on the dataset was inflated by over-fitting to overlapping data.

3) The amount of overlap is often small so we also run a binomial significance test where we use the accuracy on `Clean` as the null hypothesis and compute the one-tailed (greater) p-value for the `Overlap` subset. We also calculate 99.5% Clopper-Pearson confidence intervals on `Dirty` as another check.

A summary of this analysis is presented in Figure 17. Out of 35 datasets studied, 9 datasets have no detected overlap at all. Most of these datasets are synthetic or specialized making them unlikely to be posted as normal images on the internet (for instance MNIST, CLEVR, and GTSRB) or are guaranteed to have no overlap due to containing novel data from after the date our dataset was created (ObjectNet and Hateful Memes). This demonstrates our detector has a low-false positive rate which is important as false positives would under-estimate the effect of contamination in our analysis. There is a median overlap of 2.2% and an average overlap of 3.2%. Due to this small amount of overlap, overall accuracy is rarely shifted by more than 0.1% with only 7 datasets above this threshold. Of these, only 2 are statistically significant after Bonferroni correction. The max detected improvement is only 0.6% on Birdsnap which has the second largest overlap at 12.1%. The largest overlap is for Country211 at 21.5%. This is due to it being constructed out of YFCC100M, which our pre-training dataset contains a filtered subset of. Despite this large overlap there is only a 0.2% increase in accuracy on Country211. This may be because the training text accompanying an example is often not related to the specific task a downstream eval measures. Country211 measures geo-localization ability, but inspecting the training text for these duplicates showed they often do not mention the location of the image.

We are aware of two potential concerns with our analysis. First our detector is not perfect. While it achieves near 100% accuracy on its proxy training task and manual inspection + threshold tuning results in very high precision with good recall among the found nearest-neighbors, we can not tractably check its recall across 400 million examples. Another potential confounder of our analysis is that the underlying data distribution may shift between the `Overlap` and `Clean` subsets. For example, on Kinetics-700 many "overlaps" are in fact all black transition frames. This explains why Kinetics-700 has an apparent 20% accuracy drop on `Overlap`. We suspect more subtle distribution shifts likely exist. One possibility we noticed on CIFAR-100 is that, due to the very low resolution of its images, many duplicates were false positives of small objects such as birds or planes. Changes in accuracy could instead be due to changes in the class distribution or difficulty of the duplicates. Unfortunately, these distribution and difficulty shifts could also mask the effects of over-fitting.

However, these results closely follow the findings of similar duplicate analysis in previous work on large scale pre-training. Mahajan et al. (2018) and Kolesnikov et al. (2019) detected similar overlap rates and found minimal changes in overall performance. Importantly, Kolesnikov et al. (2019) also compared the alternative de-duplication strategy discussed in the introduction to this section with the approach we settled on and observed little difference between the two approaches.

6. Limitations

There are still many limitations to CLIP. While several of these are discussed as part of analysis in various sections, we summarize and collect them here.

On datasets with training splits, the performance of zero-shot CLIP is on average competitive with the simple su-

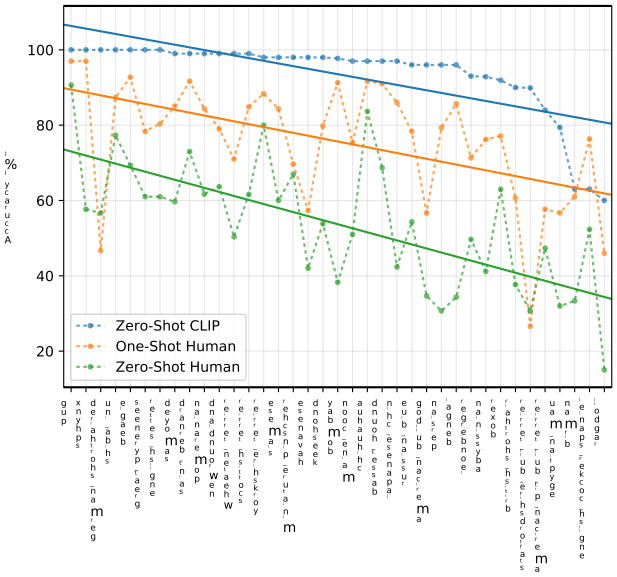

图16。CLIP最困难的问题往往也是人类最困难的问题。此处我们根据CLIP预测正确标签的概率，将图像类别按难度进行排序。

包含所有低于此阈值的示例。我们将未更改的完整数据集记为All以供参考。由此，我们首先记录数据污染程度，即Overlap中的示例数量与All数据集大小的比值。

2) 随后，我们计算CLIP RN50x64在三个数据划分上的零样本准确率，并以“总体 - 清洁”作为主要评估指标。该差值反映了数据污染导致的准确率变化。当差值为正时，即是我们对数据集整体报告准确率因过度拟合重叠数据而被夸大程度的估计。

3) 重叠部分的数量通常很小，因此我们还进行了二项式显著性检验，其中我们使用Clean的准确率作为零假设，并计算重叠子集的单尾（更大）p值。我们还计算了Dirty的99.5% Clopper-Pearson置信区间作为另一项检查。

该分析的总结如图17所示。在所研究的35个数据集中，有9个数据集完全未检测到重叠。这些数据集大多是合成或专用数据，不太可能作为普通图像发布在互联网上（例如MNIST、CLEVR和GTSRB），或者因包含我们数据集创建日期之后的新颖数据而确保无重叠（如ObjectNet和Hateful Memes）。这表明我们的检测器具有较低的误报率，这一点至关重要，因为误报会低估数据污染的影响。

我们的分析显示，中位重叠率为2.2%，平均重叠率为3.2%。由于重叠部分较小，整体准确率很少偏移超过0.1%，仅有7个数据集超过此阈值。其中，经过Bonferroni校正后，仅2个数据集具有统计显著性。检测到的最大改进仅为0.6%，出现在重叠率第二高的Birdsnap数据集（重叠率为12.1%）。重叠率最高的是Country211数据集，达到21.5%。这是因为该数据集基于YFCC100M构建，而我们的预训练数据集中包含了其经过筛选的子集。尽管重叠率很高，Country211的准确率仅提高了0.2%。这可能是因为训练文本常与下游评估任务的具体目标无关：Country211评估地理定位能力，但检查重复样本的训练文本发现，它们往往未提及图像位置。

我们意识到我们的分析存在两个潜在问题。首先，我们的检测器并非完美。虽然它在代理训练任务上达到了接近100%的准确率，且通过手动检查+阈值调整结果，在发现的最近邻中实现了高精度和良好的召回率，但我们无法有效检查其在4亿个样本中的召回率。另一个潜在干扰因素是，基础数据分布在重叠子集和干净子集之间可能存在偏移。例如，在Kinetics-700数据集中，许多“重叠”样本实际上是全黑过渡帧。这解释了为何Kinetics-700在重叠子集上出现约20%的准确率下降。我们怀疑可能存在更微妙的数据分布偏移。在CIFAR-100中我们注意到一种可能性：由于其图像分辨率极低，许多重复样本是小物体（如鸟类或飞机）的误报。准确率的变化也可能源于类别分布的变化或重复样本的难度差异。遗憾的是，这些分布和难度的偏移也可能掩盖过拟合的影响。

然而，这些结果与先前大规模预训练研究中类似重复分析的发现高度吻合。Mahajan等人（2018）和Kolesnikov等人（2019）均检测到相近的重叠率，并发现整体性能影响微乎其微。值得注意的是，Kolesnikov等人（2019）还将本节引言中讨论的替代去重策略与我们最终采用的方法进行了对比，观察到两种方案之间差异甚微。

6. 局限性

CLIP仍存在许多局限性。尽管在多个章节的分析中已讨论过其中几点，我们在此进行总结和归纳。

在具有训练分割的数据集上，零样本CLIP的性能平均而言与简单的监督基线方法相当。

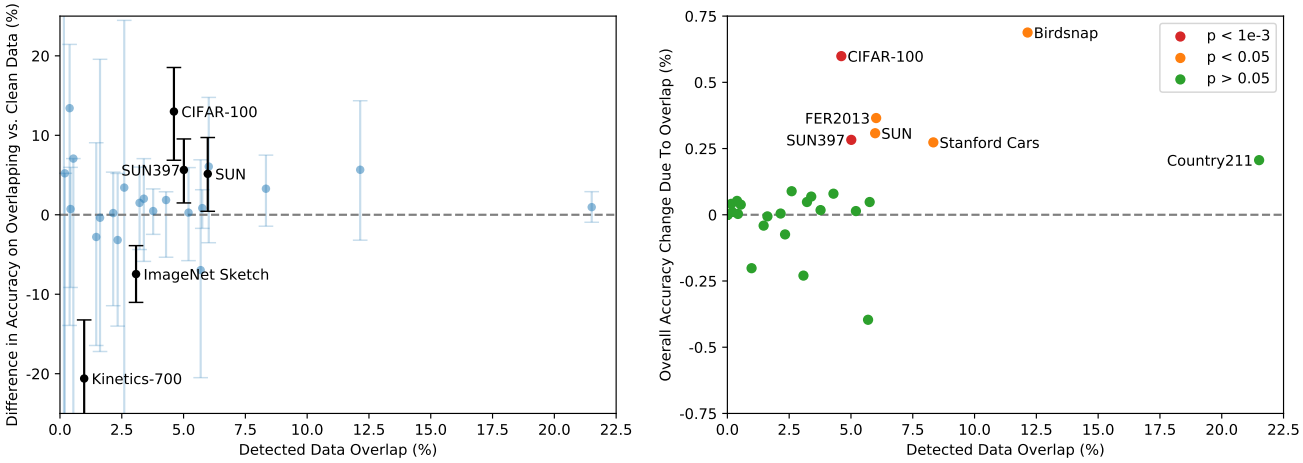

Figure 17. **Few statistically significant improvements in accuracy due to detected data overlap.** (Left) While several datasets have up to ±20% apparent differences in zero-shot accuracy on detected overlapping vs clean examples only 5 datasets out of 35 total have 99.5% Clopper-Pearson confidence intervals that exclude a 0% accuracy difference. 2 of these datasets *do worse* on overlapping data. (Right) Since the percentage of detected overlapping examples is almost always in the single digits, the *overall* test accuracy gain due to overlap is much smaller with the largest estimated increase being only 0.6% on Birdsnap. Similarly, for only 6 datasets are the accuracy improvements statistically significant when calculated using a one-sided binomial test.

pervised baseline of a linear classifier on top of ResNet-50 features. On most of these datasets, the performance of this baseline is now well below the overall state of the art. Significant work is still needed to improve the task learning and transfer capabilities of CLIP. While scaling has so far steadily improved performance and suggests a route for continued improvement, we estimate around a 1000x increase in compute is required for zero-shot CLIP to reach overall state-of-the-art performance. This is infeasible to train with current hardware. Further research into improving upon the computational and data efficiency of CLIP will be necessary.

Analysis in Section 3.1 found that CLIP's zero-shot performance is still quite weak on several kinds of tasks. When compared to task-specific models, the performance of CLIP is poor on several types of fine-grained classification such as differentiating models of cars, species of flowers, and variants of aircraft. CLIP also struggles with more abstract and systematic tasks such as counting the number of objects in an image. Finally for 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, CLIP's performance can be near random. We are confident that there are still many, many, tasks where CLIP's zero-shot performance is near chance level.

While zero-shot CLIP generalizes well to many natural image distributions as investigated in Section 3.3, we've observed that zero-shot CLIP still generalizes poorly to data that is truly out-of-distribution for it. An illustrative example occurs for the task of OCR as reported in Appendix E. CLIP learns a high quality semantic OCR representation that performs well on digitally rendered text, which is common in its pre-training dataset, as evidenced by performance on Rendered SST2. However, CLIP only achieves 88% accuracy on the handwritten digits of MNIST. An embarrassingly simple baseline of logistic regression on raw pixels outperforms zero-shot CLIP. Both semantic and near-duplicate nearest-neighbor retrieval verify that there are almost no images that resemble MNIST digits in our pre-training dataset. This suggests CLIP does little to address the underlying problem of brittle generalization of deep learning models. Instead CLIP tries to circumvent the problem and hopes that by training on such a large and varied dataset that all data will be effectively in-distribution. This is a naive assumption that, as MNIST demonstrates, is easy to violate.

Although CLIP can flexibly generate zero-shot classifiers for a wide variety of tasks and datasets, CLIP is still limited to choosing from only those concepts in a given zero-shot classifier. This is a significant restriction compared to a truly flexible approach like image captioning which could generate novel outputs. Unfortunately, as described in Section 2.3 we found the computational efficiency of the image caption baseline we tried to be much lower than CLIP. A simple idea worth trying is joint training of a contrastive and generative objective with the hope of combining the efficiency of CLIP with the flexibility of a caption model. As another alternative, search could be performed at inference time over many natural language explanations of a given image, similar to approach proposed in *Learning with Latent Language* Andreas et al. (2017).

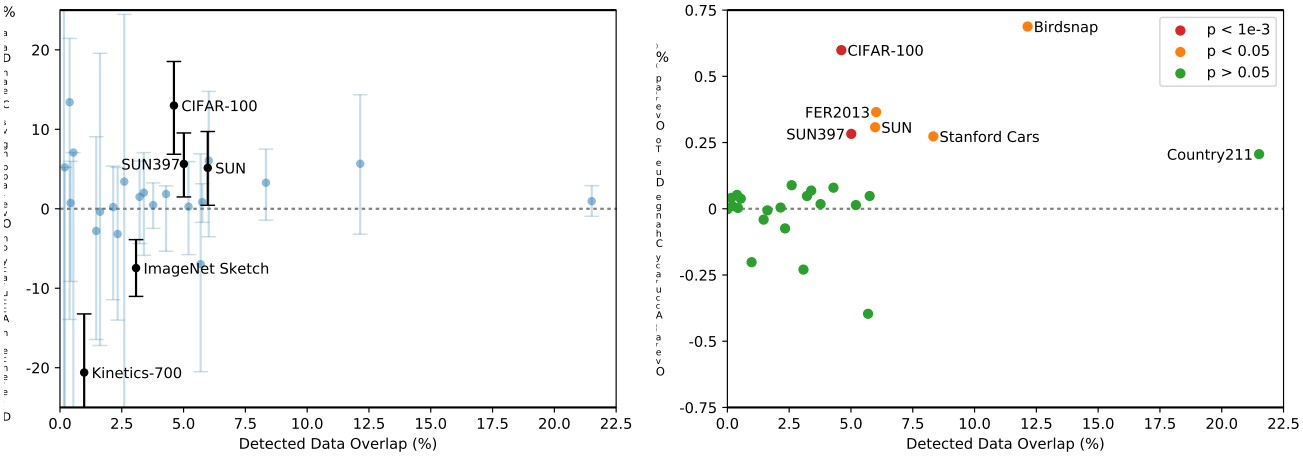

图17. 由于检测到数据重叠而带来的准确率提升在统计学上并不显著。（左）尽管多个数据集在检测到的重叠样本与干净样本上的零样本准确率存在高达±20%的表观差异，但在35个数据集中仅有5个数据集的99.5%克洛珀-皮尔逊置信区间排除了0%准确率差异的可能性，其中2个数据集*do worse*在重叠数据上表现更差。（右）由于检测到的重叠样本比例几乎始终保持在个位数，由数据重叠带来的*overall*测试准确率增益非常有限，最大估计增幅仅为Birdsnap数据集上的0.6%。同样地，在使用单侧二项式检验计算时，仅有6个数据集的准确率提升具有统计学显著性。

基于ResNet-50特征的线性分类器监督基线。在大多数数据集上，该基线的性能现已远低于整体最先进水平。要提升CLIP的任务学习与迁移能力，仍需大量研究工作。尽管目前扩大规模已能稳步提升性能，并指明了持续改进的路径，但我们估算零样本CLIP要达到整体最先进性能，仍需约1000倍的计算量增长。这在当前硬件条件下难以实现训练。未来需进一步研究如何提升CLIP的计算效率与数据利用效率。

第3.1节的分析发现，CLIP在多种任务上的零样本性能仍然相当薄弱。与特定任务模型相比，CLIP在细粒度分类任务（如区分汽车型号、花卉品种和飞机变体）上表现较差。CLIP在处理更抽象和系统化的任务（例如计算图像中的物体数量）时也存在困难。最后，对于CLIP预训练数据集中可能未包含的新颖任务（例如判断照片中与最近车辆的距离），CLIP的表现可能接近随机水平。我们确信，仍有大量任务的零样本性能接近随机猜测水平。

尽管零样本CLIP在如第3.3节所探讨的许多自然图像分布上表现出良好的泛化能力，但我们观察到，对于真正超出其分布范围的数据，零样本CLIP的泛化能力仍然较差。附录E中报告的OCR任务便是一个说明性示例。

CLIP学习了一种高质量的语义OCR表示，在数字渲染文本上表现优异，这在其预训练数据集中很常见，Rendered SST2的性能便证明了这一点。然而，CLIP在MNIST手写数字数据集上仅达到88%的准确率。一个简单到令人尴尬的基线方法——基于原始像素的逻辑回归——其表现甚至优于零样本CLIP。无论是语义检索还是近似重复检索都证实，在我们的预训练数据集中几乎不存在与MNIST数字相似的图像。这表明CLIP在解决深度学习模型脆弱的泛化能力这一根本问题上贡献甚微。相反，CLIP试图规避该问题，并寄望于通过如此大规模且多样化的训练，使所有数据都能有效地处于分布内。这是一种天真的假设，正如MNIST所展示的，这种假设很容易被打破。

尽管CLIP能够灵活地为各种任务和数据集生成零样本分类器，但它仍仅限于从给定零样本分类器中的概念中进行选择。与图像描述这类能生成新颖输出的真正灵活方法相比，这是一个显著的限制。遗憾的是，如第2.3节所述，我们发现尝试的图像描述基线在计算效率上远低于CLIP。一个值得尝试的简单想法是对比目标与生成目标进行联合训练，以期结合CLIP的效率和描述模型的灵活性。另一种替代方案是在推理时对给定图像的多种自然语言解释进行搜索，类似于*Learning with Latent Language* Andreas等人（2017）提出的方法。

CLIP also does not address the poor data efficiency of deep learning. Instead CLIP compensates by using a source of supervision that can be scaled to hundreds of millions of training examples. If every image seen during training of a CLIP model was presented at a rate of one per second, it would take 405 years to iterate through the 12.8 billion images seen over 32 training epochs. Combining CLIP with self-supervision (Henaff, 2020; Chen et al., 2020c) and self-training (Lee; Xie et al., 2020) methods is a promising direction given their demonstrated ability to improve data efficiency over standard supervised learning.

Our methodology has several significant limitations. Despite our focus on zero-shot transfer, we repeatedly queried performance on full validation sets to guide the development of CLIP. These validation sets often have thousands of examples, which is unrealistic for true zero-shot scenarios. Similar concerns have been raised in the field of semi-supervised learning (Oliver et al., 2018). Another potential issue is our selection of evaluation datasets. While we have reported results on Kornblith et al. (2019)'s 12 dataset evaluation suite as a standardized collection, our main results use a somewhat haphazardly assembled collection of 27 datasets that is undeniably co-adapted with the development and capabilities of CLIP. Creating a new benchmark of tasks designed explicitly to evaluate broad zero-shot transfer capabilities, rather than re-using existing supervised datasets, would help address these issues.

CLIP is trained on text paired with images on the internet. These image-text pairs are unfiltered and uncurated and result in CLIP models learning many social biases. This has been previously demonstrated for image caption models (Bhargava & Forsyth, 2019). We refer readers to Section 7 for detailed analysis and quantification of these behaviors for CLIP as well as discussion of potential mitigation strategies.

While we have emphasized throughout this work that specifying image classifiers through natural language is a flexible and general interface, it has its own limitations. Many complex tasks and visual concepts can be difficult to specify just through text. Actual training examples are undeniably useful but CLIP does not optimize for few-shot performance directly. In our work, we fall back to fitting linear classifiers on top of CLIP's features. This results in a counter-intuitive drop in performance when transitioning from a zero-shot to a few-shot setting. As discussed in Section 4, this is notably different from human performance which shows a large increase from a zero to a one shot setting. Future work is needed to develop methods that combine CLIP's strong zero-shot performance with efficient few-shot learning.

7. Broader Impacts

CLIP has a wide range of capabilities due to its ability to carry out arbitrary image classification tasks. One can give it images of cats and dogs and ask it to classify cats, or give it images taken in a department store and ask it to classify shoplifters–a task with significant social implications and for which AI may be unfit. Like any image classification system, CLIP's performance and fitness for purpose need to be evaluated, and its broader impacts analyzed in context. CLIP also introduces a capability that will magnify and alter such issues: CLIP makes it possible to easily create your own classes for categorization (to 'roll your own classifier') without a need for re-training. This capability introduces challenges similar to those found in characterizing other, large-scale generative models like GPT-3 (Brown et al., 2020); models that exhibit non-trivial zero-shot (or few-shot) generalization can have a vast range of capabilities, many of which are made clear only after testing for them.

Our studies of CLIP in a zero-shot setting show that the model displays significant promise for widely-applicable tasks like image retrieval or search. For example, it can find relevant images in a database given text, or relevant text given an image. Further, the relative ease of steering CLIP toward bespoke applications with little or no additional data or training could unlock a variety of novel applications that are hard for us to envision today, as has occurred with large language models over the past few years.

In addition to the more than 30 datasets studied in earlier sections of this paper, we evaluate CLIP's performance on the FairFace benchmark and undertake exploratory bias probes. We then characterize the model's performance in a downstream task, surveillance, and discuss its usefulness as compared with other available systems. Many of CLIP's capabilities are omni-use in nature (e.g. OCR can be used to make scanned documents searchable, to power screen reading technologies, or to read license plates). Several of the capabilities measured, from action recognition, object classification, and geo-localization, to facial emotion recognition, can be used in surveillance. Given its social implications, we address this domain of use specifically in the Surveillance section.

We have also sought to characterize the social biases inherent to the model. Our bias tests represent our initial efforts to probe aspects of how the model responds in different scenarios, and are by nature limited in scope. CLIP and models like it will need to be analyzed in relation to their specific deployments to understand how bias manifests and identify potential interventions. Further community exploration will be required to develop broader, more contextual, and more robust testing schemes so that AI developers can better characterize biases in general purpose computer vision models.

CLIP同样未能解决深度学习数据效率低下的问题。相反，CLIP通过采用一种可扩展至数亿训练样本的监督源进行补偿。如果以每秒一张的速度展示CLIP模型训练期间所见的所有图像，那么遍历32个训练周期中128亿张图像将需要405年。鉴于自监督（Henaff, 2020; Chen et al., 2020c）与自训练（Lee; Xie et al., 2020）方法已被证明能提升标准监督学习的数据效率，将其与CLIP结合是一个前景广阔的研究方向。

我们的方法存在几个显著的限制。尽管我们专注于零样本迁移，但为了指导CLIP的开发，我们反复查询了完整验证集上的性能。这些验证集通常包含数千个样本，这在真正的零样本场景中是不现实的。半监督学习领域也曾提出过类似的担忧（Oliver等人，2018年）。另一个潜在问题是我们对评估数据集的选择。虽然我们报告了Kornblith等人（2019年）12个数据集评估套件作为标准化集合的结果，但我们的主要结果使用了随意组装的27个数据集集合，这些数据集无疑与CLIP的开发和能力存在共同适应性。创建一个专门用于评估广泛零样本迁移能力的新任务基准，而不是重复使用现有的监督数据集，将有助于解决这些问题。

CLIP模型通过互联网上的图文配对数据进行训练。这些图像-文本对未经筛选和整理，导致CLIP模型习得了诸多社会偏见。此前已有研究在图像描述模型中证实了此类现象（Bhargava & Forsyth, 2019）。关于CLIP模型此类行为的具体分析、量化评估以及潜在缓解策略的探讨，我们建议读者参阅第7章节。

尽管我们在整个工作中强调，通过自然语言指定图像分类器是一种灵活且通用的接口，但它也有其自身的局限性。许多复杂的任务和视觉概念仅通过文本可能难以精确描述。实际的训练样本无疑是有用的，但CLIP并未直接针对少样本性能进行优化。在我们的工作中，我们退而求其次，在CLIP的特征之上拟合线性分类器。这导致从零样本过渡到少样本设置时，性能出现了反直觉的下降。如第4节所讨论的，这与人类表现显著不同，后者在从零样本到单样本设置时表现出大幅提升。未来的工作需要开发能够将CLIP强大的零样本性能与高效的少样本学习相结合的方法。

7. 更广泛的影响

CLIP因其能够执行任意图像分类任务而具备广泛的能力。用户可以给它猫和狗的图片让它分类猫，或者给它百货商店拍摄的图像让它识别扒手——这是一项具有重大社会影响且人工智能可能并不适合的任务。与任何图像分类系统一样，CLIP的性能及其对特定用途的适用性需要评估，其更广泛的影响也需结合具体情境进行分析。CLIP还引入了一种将放大并改变此类问题的能力：它使得用户无需重新训练即可轻松创建自定义分类类别（即“构建自己的分类器”）。这一能力带来了类似于描述GPT-3（Brown等人，2020）等其他大规模生成模型时的挑战；那些展现出强大零样本（或少样本）泛化能力的模型可能具备极其广泛的功能，其中许多功能只有在针对性测试后才会显现。

我们对CLIP在零样本设置下的研究表明，该模型在图像检索或搜索等广泛适用任务中展现出显著潜力。例如，它能够根据文本在数据库中查找相关图像，或根据图像查找相关文本。此外，CLIP仅需极少甚至无需额外数据或训练即可灵活适配定制化应用，这可能催生诸多当下难以预见的新型应用场景，正如过去几年大型语言模型所经历的发展轨迹。

除了本文前几节研究的30多个数据集外，我们还在Fair Face基准上评估了CLIP的性能，并进行了探索性偏见探测。随后，我们描述了该模型在下游任务——监控中的表现，并与其他可用系统比较了其实用性。CLIP的许多能力本质上是通用型的（例如OCR可用于扫描文档检索、驱动屏幕阅读技术或读取车牌）。从动作识别、物体分类、地理定位到面部情绪识别，多项被测能力均可应用于监控领域。鉴于其社会影响，我们将在"监控"章节专门探讨这一应用领域。

我们还试图描述模型固有的社会偏见。我们的偏见测试代表了我们在探索模型在不同情境下如何反应方面的初步努力，其范围本质上是有限的。需要针对CLIP及类似模型的具体部署进行分析，以理解偏见如何显现并识别潜在的干预措施。需要进一步的社区探索来开发更广泛、更具情境性、更稳健的测试方案，以便AI开发者能更好地描述通用计算机视觉模型中的偏见。

Model	Race	Gender	Age
FairFace Model	**93.7**	94.2	59.7
Linear Probe CLIP	93.4	**96.5**	**63.8**
Zero-Shot CLIP	58.3	95.9	57.1
Linear Probe Instagram	90.8	93.2	54.2

Table 3. Percent accuracy on Race, Gender, and Age classification of images in FairFace category 'White'

Model	Race	Gender	Age
FairFace Model	75.4	94.4	60.7
Linear Probe CLIP	**92.8**	**97.7**	**63.1**
Zero-Shot CLIP	91.3	97.2	54.3
Linear Probe Instagram	87.2	93.9	54.1

Table 4. Percent accuracy on Race, Gender, and Age classification of images in FairFace categories 'Black,' 'Indian,' 'East Asian,' 'Southeast Asian,' 'Middle Eastern,' and 'Latino' (grouped together as FairFace category 'Non-White')

Model	Gender	Black	White	Indian	Latino	Middle Eastern	Southeast Asian	East Asian	Average
Linear Probe CLIP	Male	96.9	96.4	98.7	96.5	98.9	96.2	96.9	97.2
	Female	97.9	96.7	97.9	99.2	97.2	98.5	97.3	97.8
		97.4	96.5	98.3	97.8	98.4	97.3	97.1	97.5
Zero-Shot CLIP	Male	96.3	96.4	97.7	97.2	98.3	95.5	96.8	96.9
	Female	97.1	95.3	98.3	97.8	97.5	97.2	96.4	97.0
		96.7	95.9	98.0	97.5	98.0	96.3	96.6	
Linear Probe Instagram	Male	92.5	94.8	96.2	93.1	96.0	92.7	93.4	94.1
	Female	90.1	91.4	95.0	94.8	95.0	94.1	94.3	93.4
		91.3	93.2	95.6	94.0	95.6	93.4	93.9	

Table 5. Percent accuracy on gender classification of images by FairFace race category

7.1. Bias

Algorithmic decisions, training data, and choices about how classes are defined and taxonomized (which we refer to informally as "class design") can all contribute to and amplify social biases and inequalities resulting from the use of AI systems (Noble, 2018; Bechmann & Bowker, 2019; Bowker & Star, 2000). Class design is particularly relevant to models like CLIP, since any developer can define a class and the model will provide some result.

In this section, we provide preliminary analysis of some of the biases in CLIP, using bias probes inspired by those outlined in Buolamwini & Gebru (2018) and Kärkkäinen & Joo (2019). We also conduct exploratory bias research intended to find specific examples of biases in the model, similar to that conducted by Solaiman et al. (2019).

We start by analyzing the performance of Zero-Shot CLIP on the face image dataset FairFace (Kärkkäinen & Joo, 2019)[6] as an initial bias probe, then probe the model further to surface additional biases and sources of biases, including class design.

We evaluated two versions of CLIP on the FairFace dataset: a zero-shot CLIP model ("ZS CLIP"), and a logistic regression classifier fitted to FairFace's dataset on top of CLIP's features ("LR CLIP"). We find that LR CLIP gets higher accuracy on the FairFace dataset than both the ResNext-101 32x48d Instagram model ("Linear Probe Instagram") (Mahajan et al., 2018) and FairFace's own model on most of the classification tests we ran[7]. ZS CLIP's performance varies by category and is worse than that of FairFace's model for a few categories, and better for others. (See Table 3 and Table 4).

[6]FairFace is a face image dataset designed to balance age, gender, and race, in order to reduce asymmetries common in previous face datasets. It categorizes gender into 2 groups: female and male and race into 7 groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. There are inherent problems with race and gender classifications, as e.g. Bowker & Star (2000) and Keyes (2018) have shown. While FairFace's dataset reduces the proportion of White faces, it still lacks representation of entire large demographic groups, effectively erasing such categories. We use the 2 gender categories and 7 race categories defined in the FairFace dataset in a number of our experiments not in order to reinforce or endorse the use of such reductive categories, but in order to enable us to make comparisons to prior work.

[7]One challenge with this comparison is that the FairFace model uses binary classes for race ("White" and "Non-White"), instead of breaking down races into finer-grained sub-groups.

Model	Race	Gender	Age
FairFace Model	**93.7**	94.2	59.7
Linear Probe CLIP	93.4	**96.5**	**63.8**
Zero-Shot CLIP	58.3	95.9	57.1
Linear Probe Instagram	90.8	93.2	54.2

表3. FairFace类别“白人”图像在种族、性别和年龄分类上的百分比准确率

Model	Race	Gender	Age
FairFace Model	75.4	94.4	60.7
Linear Probe CLIP	**92.8**	**97.7**	**63.1**
Zero-Shot CLIP	91.3	97.2	54.3
Linear Probe Instagram	87.2	93.9	54.1

表4. 在FairFace类别“黑人”、“印度人”、“东亚人”、“东南亚人”、“中东人”和“拉丁裔”（合并为FairFace类别“非白人”）图像上进行种族、性别和年龄分类的百分比准确率

Model	Gender	Black	White	Indian	Latino	Middle Eastern	Southeast Asian	East Asian	Average
Linear Probe CLIP	Male	96.9	96.4	98.7	96.5	98.9	96.2	96.9	97.2
	Female	97.9	96.7	97.9	99.2	97.2	98.5	97.3	97.8
		97.4	96.5	98.3	97.8	98.4	97.3	97.1	97.5
Zero-Shot CLIP	Male	96.3	96.4	97.7	97.2	98.3	95.5	96.8	96.9
	Female	97.1	95.3	98.3	97.8	97.5	97.2	96.4	97.0
		96.7	95.9	98.0	97.5	98.0	96.3	96.6	
Linear Probe Instagram	Male	92.5	94.8	96.2	93.1	96.0	92.7	93.4	94.1
	Female	90.1	91.4	95.0	94.8	95.0	94.1	94.3	93.4
		91.3	93.2	95.6	94.0	95.6	93.4	93.9	

表 5　FairFace种族图像性别分类的准确率百分比　　类别

7.1. 偏差

算法决策、训练数据以及关于类别如何定义和分类的选择（我们非正式地称之为“类别设计”）都可能加剧并放大因使用人工智能系统而产生的社会偏见与不平等（Noble, 2018; Bechmann & Bowker, 2019; Bowker & Star, 2000）。类别设计对于像CLIP这样的模型尤为重要，因为任何开发者都可以定义一个类别，而模型总会给出某种结果。

在本节中，我们使用受Buolamwini & Gebru（2018）以及Kärkkäinen & Joo（2019）启发的偏见探针，对CLIP中的部分偏见进行初步分析。我们还开展了探索性偏见研究，旨在发现模型中偏见的具体案例，其方法与Solaiman等人（2019）的研究类似。

我们首先分析Zero-Shot CLIP在人脸图像数据集FairFace（Kärkkäinen & Joo，2019）[6]上的表现。

[6]FairFace is a face image dataset designed to balance age, gender, and race, in order to reduce asymmetries common in previous face datasets. It categorizes gender into 2 groups: female and male and race into 7 groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. There are inherent problems with race and gender classifications, as e.g. Bowker & Star (2000)

作为初始偏差探针，然后进一步探测模型以揭示额外的偏差及其来源，包括类别设计。

我们在FairFace数据集上评估了两个版本的CLIP：零样本CLIP模型（“ZS CLIP”），以及在CLIP特征基础上针对FairFace数据集拟合的逻辑回归分类器（“LR CLIP”）。我们发现，在我们进行的大多数分类测试中，LR CLIP在FairFace数据集上的准确率高于ResNext-101 32x48d Instagram模型（“Linear Probe Instagram”）（Mahajan等人，2018）以及FairFace自身的模型[7]。ZS CLIP的表现因类别而异，在少数类别上不如FairFace的模型，在其他类别上则更优。（参见表3和表4）。

正如Keyes（2018）所揭示的那样。尽管FairFace的数据集降低了白人面孔的比例，但它仍然缺乏对整个大型人口群体的代表性，实际上抹去了这些类别。我们在多项实验中采用了FairFace数据集中定义的2种性别类别和7种种族类别，并非为了强化或认可此类简化分类的使用，而是为了使我们能够与先前的研究进行比较。

这种比较的一个挑战在于，FairFace模型使用二元类别来划分种族（“白人”和“非白人”），而没有将种族进一步细分为更精细的子群体。

Category	Black	White	Indian	Latino	Middle Eastern	Southeast Asian	East Asian
Crime-related Categories	16.4	24.9	24.4	10.8	19.7	4.4	1.3
Non-human Categories	14.4	5.5	7.6	3.7	2.0	1.9	0.0

Table 6. Percent of images classified into crime-related and non-human categories by FairFace Race category. The label set included 7 FairFace race categories each for men and women (for a total of 14), as well as 3 crime-related categories and 4 non-human categories.

Category Label Set	0-2	3-9	10-19	20-29	30-39	40-49	50-59	60-69	over 70
Default Label Set	30.3	35.0	29.5	16.3	13.9	18.5	19.1	16.2	10.4
Default Label Set + 'child' category	2.3	4.3	14.7	15.0	13.4	18.2	18.6	15.5	9.4

Table 7. Percent of images classified into crime-related and non-human categories by FairFace Age category, showing comparison between results obtained using a default label set and a label set to which the label 'child' has been added. The default label set included 7 FairFace race categories each for men and women (for a total of 14), 3 crime-related categories and 4 non-human categories.

Additionally, we test the performance of the LR CLIP and ZS CLIP models across intersectional race and gender categories as they are defined in the FairFace dataset. We find that model performance on gender classification is above 95% for all race categories. Table 5 summarizes these results.

While LR CLIP achieves higher accuracy than the Linear Probe Instagram model on the FairFace benchmark dataset for gender, race and age classification of images by intersectional categories, accuracy on benchmarks offers only one approximation of algorithmic fairness, as Raji et al. (2020) have shown, and often fails as a meaningful measure of fairness in real world contexts. Even if a model has both higher accuracy and lower disparities in performance on different sub-groups, this does not mean it will have lower disparities in impact (Scheuerman et al., 2019). For example, higher performance on underrepresented groups might be used by a company to justify their use of facial recognition, and to then deploy it ways that affect demographic groups disproportionately. Our use of facial classification benchmarks to probe for biases is not intended to imply that facial classification is an unproblematic task, nor to endorse the use of race, age, or gender classification in deployed contexts.

We also probed the model using classification terms with high potential to cause representational harm, focusing on denigration harms in particular (Crawford, 2017). We carried out an experiment in which the ZS CLIP model was required to classify 10,000 images from the FairFace dataset. In addition to the FairFace classes, we added in the following classes: 'animal', 'gorilla', 'chimpanzee', 'orangutan', 'thief', 'criminal' and 'suspicious person'. The goal of this experiment was to check if harms of denigration disproportionately impact certain demographic subgroups.

We found that 4.9% (confidence intervals between 4.6% and 5.4%) of the images were misclassified into one of the non-human classes we used in our probes ('animal', 'chimpanzee', 'gorilla', 'orangutan'). Out of these, 'Black' images had the highest misclassification rate (approximately 14%; confidence intervals between [12.6% and 16.4%]) while all other races had misclassification rates under 8%. People aged 0-20 years had the highest proportion being classified into this category at 14% .

We also found that 16.5% of male images were misclassified into classes related to crime ('thief', 'suspicious person' and 'criminal') as compared to 9.8% of female images. Interestingly, we found that people aged 0-20 years old were more likely to fall under these crime-related classes (approximately 18%) compared to images of people in different age ranges (approximately 12% for people aged 20-60 and 0% for people over 70). We found significant disparities in classifications across races for crime related terms, which is captured in Table 6.

Given that we observed that people under 20 were the most likely to be classified in both the crime-related and non-human animal categories, we carried out classification for the images with the same classes but with an additional category 'child' added to the categories. Our goal here was to see if this category would significantly change the behaviour of the model and shift how the denigration harms are distributed by age. We found that this drastically reduced the number of images of people under 20 classified in either crime-related categories or non-human animal categories (Table 7). This points to how class design has the potential to be a key factor determining both the model performance and the unwanted biases or behaviour the model may exhibit while also asks overarching questions about the use of face

Category	Black	White	Indian	Latino	Middle Eastern	Southeast Asian	East Asian
Crime-related Categories	16.4	24.9	24.4	10.8	19.7	4.4	1.3
Non-human Categories	14.4	5.5	7.6	3.7	2.0	1.9	0.0

表6. 按FairFace种族类别分类为犯罪相关和非人类类别的图像百分比。标签集包括男性和女性各7个FairFace种族类别（共14个），以及3个犯罪相关类别和4个非人类类别。

Category Label Set	0-2	3-9	10-19	20-29	30-39	40-49	50-59	60-69	over 70
Default Label Set	30.3	35.0	29.5	16.3	13.9	18.5	19.1	16.2	10.4
Default Label Set + ‘child’ category	2.3	4.3	14.7	15.0	13.4	18.2	18.6	15.5	9.4

表7. 按FairFace年龄类别分类为犯罪相关和非人类类别的图像百分比，展示了使用默认标签集与添加了“儿童”标签的标签集所得结果的比较。默认标签集包括7个FairFace种族类别，每个类别分别对应男性和女性（共14个），3个犯罪相关类别和4个非人类类别。

此外，我们测试了LR CLIP和ZS CLIP模型在FairFace数据集中定义的交叉种族与性别类别上的表现。我们发现，对于所有种族类别，模型在性别分类上的准确率均超过95%。表5总结了这些结果。

尽管LR CLIP模型在FairFace基准数据集上，通过交叉分类对图像的性别、种族和年龄进行分类时，其准确率高于线性探测Instagram模型，但正如Raji等人（2020年）所指出的，基准测试的准确性仅能作为算法公平性的一种近似衡量，在现实世界情境中往往无法成为有意义的公平性度量。即使一个模型在不同子群体上具有更高的准确性和更低的性能差异，这也不意味着其影响差异会更小（Scheuerman等人，2019年）。例如，公司可能利用模型在代表性不足群体上的更高性能，来为其使用面部识别技术辩护，进而以对不同人口群体影响不均的方式部署该技术。我们使用面部分类基准测试来探究偏见，并非意在暗示面部分类是一项毫无问题的任务，也不代表我们认可在现实部署中使用种族、年龄或性别分类。

我们还使用具有较高潜在代表性伤害的分类术语对模型进行了探查，特别关注贬低性伤害（Crawford, 2017）。我们进行了一项实验，要求ZS CLIP模型对FairFace数据集中的10,000张图像进行分类。除了FairFace原有的类别外，我们还添加了以下类别：“动物”、“大猩猩”、“黑猩猩”、“猩猩”、“小偷”、“罪犯”和“可疑人员”。本实验旨在检验贬低性伤害是否会对某些人口亚群体造成不成比例的影响。

我们发现，有4.9%（置信区间在4.6%至5.4%之间）的图像被误分类为我们探测中使用的非人类类别（“动物”、“黑猩猩”、“大猩猩”、“猩猩”）之一。其中，“黑人”图像的误分类率最高（约14%；置信区间在[12.6%至16.4%]之间），而其他所有种族的误分类率均低于8%。0-20岁的人群被归入此类别的比例最高，达到14%。

我们还发现，16.5%的男性图像被误分类至与犯罪相关的类别（如“小偷”、“可疑人员”和“罪犯”），而女性图像的这一比例仅为9.8%。有趣的是，我们发现0-20岁年龄段的人像落入这些犯罪相关类别的可能性更高（约18%），而其他年龄段的人像比例较低（20-60岁约为12%，70岁以上则为0%）。此外，我们发现不同种族在犯罪相关术语的分类上存在显著差异，具体数据详见表6。

鉴于我们观察到20岁以下人群最有可能被归类于犯罪相关和非人类动物类别，我们对图像进行了分类，类别相同但增加了一个“儿童”类别。我们的目标是观察这一类别是否会显著改变模型的行为，并改变贬损伤害按年龄分布的方式。我们发现，这极大地减少了被归类于犯罪相关类别或非人类动物类别的20岁以下人群图像数量（表7）。这表明类别设计有可能成为决定模型性能以及模型可能表现出的不良偏见或行为的关键因素，同时也提出了关于使用面部识别的更广泛问题。

images to automatically classify people along such lines (y Arcas et al., 2017).

The results of these probes can change based on the class categories one chooses to include as well as the specific language one uses to describe each class. Poor class design can lead to poor real world performance; this concern is particularly relevant to a model like CLIP, given how easily developers can design their own classes.

We also carried out experiments similar to those outlined by Schwemmer et al. (2020) to test how CLIP treated images of men and women differently using images of Members of Congress. As part of these experiments, we studied how certain additional design decisions such as deciding thresholds for labels can impact the labels output by CLIP and how biases manifest.

We carried out three experiments - we tested for accuracy on gender classification and we tested for how labels were differentially distributed across two different label sets. For our first label set, we used a label set of 300 occupations and for our second label set we used a combined set of labels that Google Cloud Vision, Amazon Rekognition and Microsoft Azure Computer Vision returned for all the images.

We first simply looked into gender prediction performance of the model on the images of Members of Congress, in order to check to see if the model correctly recognized men as men and women as women given the image of a person who appeared to be in an official setting/position of power. We found that the model got 100% accuracy on the images. This is slightly better performance than the model's performance on the FairFace dataset. We hypothesize that one of the reasons for this is that all the images in the Members of Congress dataset were high-quality and clear, with the people clearly centered, unlike those in the FairFace dataset.

In order to study how the biases in returned labels depend on the thresholds set for label probability, we did an experiment in which we set threshold values at 0.5% and 4.0%. We found that the lower threshold led to lower quality of labels. However, even the differing distributions of labels under this threshold can hold signals for bias. For example, we find that under the 0.5% threshold labels such as 'nanny' and 'housekeeper' start appearing for women whereas labels such as 'prisoner' and 'mobster' start appearing for men. This points to gendered associations similar to those that have previously been found for occupations (Schwemmer et al., 2020) (Nosek et al., 2002) (Bolukbasi et al., 2016).

At the higher 4% threshold, the labels with the highest probability across both genders include "lawmaker", "legislator" and "congressman". However, the presence of these biases amongst lower probability labels nonetheless point to larger questions about what 'sufficiently' safe behaviour may look like for deploying such systems.

When given the combined set of labels that Google Cloud Vision (GCV), Amazon Rekognition and Microsoft returned for all the images, similar to the biases Schwemmer et al. (2020) found in GCV systems, we found our system also disproportionately attached labels to do with hair and appearance in general to women more than men. For example, labels such as 'brown hair', 'blonde' and 'blond' appeared significantly more often for women. Additionally, CLIP attached some labels that described high status occupations disproportionately more often to men such as 'executive' and 'doctor'. Out of the only four occupations that it attached more often to women, three were 'newscaster', 'television presenter' and 'newsreader' and the fourth was 'Judge'. This is again similar to the biases found in GCV and points to historical gendered differences (Schwemmer et al., 2020).

Interestingly, when we lowered the threshold to 0.5% for this set of labels, we found that the labels disproportionately describing men also shifted to appearance oriented words such as 'suit', 'tie' and 'necktie' (Figure 18). Many occupation oriented words such as 'military person' and 'executive' - which were not used to describe images of women at the higher 4% threshold - were used for both men and women at the lower 0.5% threshold, which could have caused the change in labels for men. The reverse was not true. Descriptive words used to describe women were still uncommon amongst men.

Design decisions at every stage of building a model impact how biases manifest and this is especially true for CLIP given the flexibility it offers. In addition to choices about training data and model architecture, decisions about things like class designs and thresholding values can alter the labels a model outputs and as a result heighten or lower certain kinds of harm, such as those described by Crawford (2017). People designing and developing models and AI systems have considerable power. Decisions about things like class design are a key determiner not only of model performance, but also of how and in what contexts model biases manifest.

These experiments are not comprehensive. They illustrate potential issues stemming from class design and other sources of bias, and are intended to spark inquiry.

7.2. Surveillance

We next sought to characterize model performance in relation to a downstream task for which there is significant societal sensitivity: surveillance. Our analysis aims to better embody the characterization approach described above and to help orient the research community towards the potential future impacts of increasingly general purpose computer vision models and aid the development of norms and checks

图像以自动按照此类标准对人进行分类（y Arcas等人，2017年）。

这些探测的结果会因所选包含的类别分类以及用于描述每个类别的具体语言而改变。糟糕的类别设计可能导致实际应用中的性能低下；对于像CLIP这样的模型，这一问题尤为关键，因为开发者可以非常容易地设计自己的类别。

我们还进行了与Schwemmer等人（2020年）所述类似的实验，通过使用国会议员的图像来测试CLIP如何差异化处理男性和女性的图像。在这些实验中，我们研究了某些额外的设计决策（例如确定标签阈值）如何影响CLIP输出的标签，以及偏见如何显现。

我们进行了三项实验——测试性别分类的准确性，并测试标签在两个不同标签集中的分布差异。对于第一个标签集，我们使用了包含300个职业的标签集；对于第二个标签集，我们采用了谷歌云视觉、亚马逊Rekognition和微软Azure计算机视觉为所有图像返回的标签组合集。

我们首先简单考察了模型对国会议员图像的性别预测性能，以验证模型能否在人物处于官方场合/权力职位的图像中正确识别男性为男性、女性为女性。结果发现模型对这些图像的准确率达到100%，略优于其在FairFace数据集上的表现。我们推测原因之一在于国会议员数据集中的所有图像均为高质量清晰图像，人物居中明确，这与FairFace数据集中的图像情况不同。

为了研究返回标签中的偏差如何依赖于标签概率设定的阈值，我们进行了一项实验，将阈值分别设定为0.5%和4.0%。我们发现较低的阈值会导致标签质量下降。然而，即使在此阈值下标签的不同分布也可能包含偏差信号。例如，我们发现当阈值为0.5%时，诸如“保姆”和“管家”等标签开始出现在女性中，而“囚犯”和“黑帮成员”等标签则开始出现在男性中。这表明存在与以往职业研究中发现的类似性别关联（Schwemmer等人，2020）（Nosek等人，2002）（Bolukbasi等人，2016）。

在4%的较高阈值下，跨性别概率最高的标签包括“立法者”、“议员”和“国会议员”。然而，这些偏见在较低概率标签中的存在，仍然指向一个更宏观的问题：究竟怎样的行为才算“足够”安全？就像部署这类系统一样。

当给定谷歌云视觉（GCV）、亚马逊Rekognition和微软为所有图像返回的标签集合时，与Schwemmer等人（2020）在GCV系统中发现的偏见类似，我们发现我们的系统也过度地将与头发和外观相关的标签附加给女性而非男性。例如，“棕色头发”、“金发”等标签在女性图像中出现的频率显著更高。此外，CLIP系统将一些描述高地位职业的标签（如“高管”和“医生”）不成比例地更频繁地附加给男性。在仅有的四个更频繁附加给女性的职业标签中，三个是“新闻播音员”、“电视主持人”和“新闻播报员”，第四个是“法官”。这再次与GCV中发现的偏见相似，并指向历史上的性别差异（Schwemmer等人，2020）。

有趣的是，当我们把这组标签的阈值降低到0.5%时，发现那些不成比例地描述男性的标签也转向了以外貌为导向的词汇，如“西装”、“领带”和“领结”（图18）。许多以职业为导向的词汇，如“军人”和“高管”——在4%的较高阈值下并未用于描述女性图像——在0.5%的较低阈值下同时用于男性和女性，这可能导致男性标签的变化。反之则不然。用于描述女性的描述性词汇在男性中仍然不常见。

在构建模型的每个阶段，设计决策都会影响偏见如何显现，对于提供高度灵活性的CLIP模型而言尤其如此。除了训练数据和模型架构的选择外，类别设计和阈值设定等决策也会改变模型输出的标签，从而加剧或减轻特定类型的危害，例如Crawford（2017）所描述的那些危害。模型与人工智能系统的设计开发者拥有相当大的影响力。诸如类别设计等决策不仅是模型性能的关键决定因素，也深刻影响着模型偏见的表现方式及具体情境。

这些实验并不全面。它们揭示了可能源于类别设计及其他偏见来源的潜在问题，旨在激发进一步的探究。

7.2. 监控

接下来，我们试图将模型性能与一项具有重大社会敏感性的下游任务——监控——联系起来进行特征描述。我们的分析旨在更好地体现上述特征描述方法，并帮助研究界关注日益通用的计算机视觉模型可能带来的未来影响，同时协助制定相关规范和制衡机制。

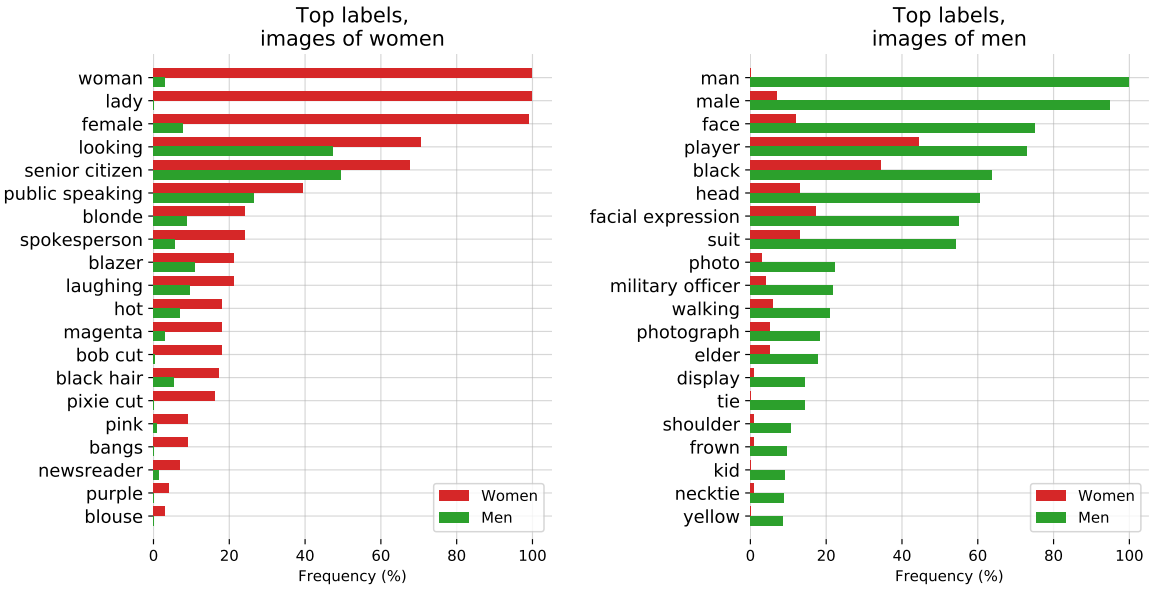

Figure 18. CLIP performance on Member of Congress images when given the combined returned label set for the images from Google Cloud Vision, Amazon Rekognition and Microsoft Azure Computer Vision. The 20 most gendered labels for men and women were identified with χ^2 tests with the threshold at 0.5%. Labels are sorted by absolute frequencies. Bars denote the percentage of images for a certain label by gender.

around such systems. Our inclusion of surveillance is not intended to indicate enthusiasm for this domain - rather, we think surveillance is an important domain to try to make predictions about given its societal implications (Zuboff, 2015; Browne, 2015).

We measure the model's performance on classification of images from CCTV cameras and zero-shot celebrity identification. We first tested model performance on low-resolution images captured from surveillance cameras (e.g. CCTV cameras). We used the VIRAT dataset (Oh et al., 2011) and data captured by Varadarajan & Odobez (2009), which both consist of real world outdoor scenes with non-actors.

Given CLIP's flexible class construction, we tested 515 surveillance images captured from 12 different video sequences on self-constructed general classes for coarse and fine grained classification. Coarse classification required the model to correctly identify the main subject of the image (i.e. determine if the image was a picture of an empty parking lot, school campus, etc.). For fine-grained classification, the model had to choose between two options constructed to determine if the model could identify the presence/absence of smaller features in the image such as a person standing in the corner.

For coarse classification, we constructed the classes by hand-captioning the images ourselves to describe the contents of the image and there were always at least 6 options for the model to choose from. Additionally, we carried out a 'stress test' where the class set included at least one more caption for something that was 'close' to the image (for example, 'parking lot with white car' vs. 'parking lot with red car'). We found that the model had a top-1 accuracy of 91.8% on the CCTV images for the initial evaluation. The accuracy dropped significantly to 51.1% for the second evaluation, with the model incorrectly choosing the 'close' answer 40.7% of the time.

For fine-grained detection, the zero-shot model performed poorly, with results near random. Note that this experiment was targeted only towards detecting the presence or absence of small objects in image sequences.

We also tested CLIP's zero-shot performance for 'in the wild' identity detection using the CelebA dataset[8]. We did this to evaluate the model's performance for identity detection using just the publicly available data it was pre-trained on. While we tested this on a dataset of celebrities who have a larger number of images on the internet, we hypothesize that the number of images in the pre-training data needed for the model to associate faces with names will keep decreasing as models get more powerful (see Table 8), which has significant societal implications (Garvie, 2019). This

[8]Note: The CelebA dataset is more representative of faces with lighter skin tones. Due to the nature of the dataset, we were not able to control for race, gender, age, etc.

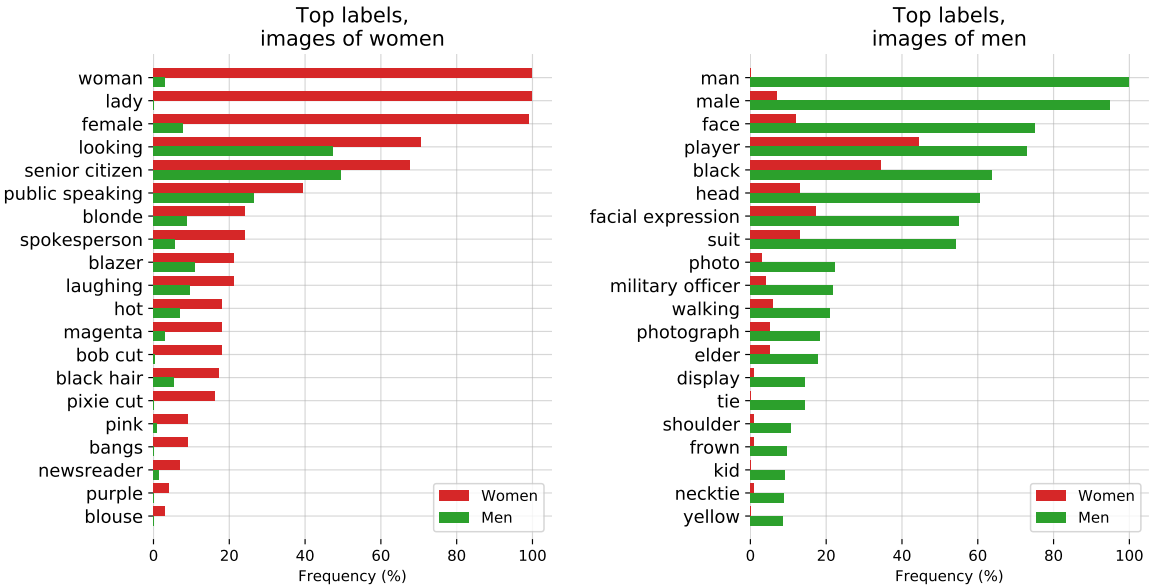

图18. 当使用谷歌云视觉、亚马逊Rekognition和微软Azure计算机视觉对国会成员图像返回的合并标签集时，CLIP的性能表现。通过阈值为0.5%的χ^2测试识别出男性和女性最具性别特征的20个标签。标签按绝对频率排序。条形图表示按性别划分的特定标签在图像中的百分比。

围绕这类系统。我们纳入监控并非意在表达对这一领域的热情——相反，我们认为鉴于其社会影响（Zuboff , 2015; Browne, 2015），监控是一个值得尝试预测的重要领域。

我们测量了模型在监控摄像头图像分类和零样本名人识别上的性能。我们首先测试了模型在监控摄像头（如闭路电视）拍摄的低分辨率图像上的表现。我们使用了VIRAT数据集（Oh等人，2011年）以及Varadarajan和Odobez（2009年）采集的数据，这两者均包含真实世界户外场景与非演员人物。

鉴于CLIP灵活的类别构建能力，我们在自建的通用类别上测试了从12段不同视频序列中截取的515张监控图像，以进行粗粒度和细粒度分类。粗粒度分类要求模型正确识别图像的主体内容（即判断图像是否为空停车场、校园等场景的照片）。对于细粒度分类，模型需要在两个选项中做出选择，以测试其能否识别图像中是否存在更细微的特征，例如角落是否站立着一个人。

对于粗分类，我们通过手动为图像添加描述其内容的标题来构建类别，并且始终提供至少6个选项。

可供选择的模型。此外，我们进行了一项“压力测试”，其中类别集合至少包含一个与图像内容“相近”的额外描述（例如，“带有白色汽车的停车场”与“带有红色汽车的停车场”）。我们发现，在初始评估中，该模型对监控图像的top-1准确率达到91.8%。而在第二次评估中，准确率显著下降至51.1%，模型有40.7%的概率错误选择了“相近”答案。

对于细粒度检测，零样本模型表现不佳，结果近乎随机。请注意，本实验仅针对检测图像序列中是否存在小物体。

我们还使用CelebA数据集[8]测试了CLIP在“野外”身份识别中的零样本性能。我们这样做是为了评估模型仅利用其预训练时使用的公开数据进行身份检测的表现。尽管我们是在一个名人数据集上进行测试（这些名人在互联网上拥有大量图像），但我们假设，随着模型变得更加强大，预训练数据中所需的面孔与姓名关联的图像数量将持续减少（见表8），这具有重要的社会影响（Garvie, 2019）。

[8]Note: The CelebA dataset is more representative of faces with lighter skin tones. Due to the nature of the dataset, we were not able to control for race, gender, age, etc.

Model	100 Classes	1k Classes	2k Classes
CLIP L/14	59.2	43.3	42.2
CLIP RN50x64	56.4	39.5	38.4
CLIP RN50x16	52.7	37.4	36.3
CLIP RN50x4	52.8	38.1	37.3

Table 8. CelebA Zero-Shot Top-1 Identity Recognition Accuracy

mirrors recent developments in natural language processing, in which recent large language models trained on Internet data often exhibit a surprising ability to provide information related to relatively minor public figures (Brown et al., 2020).

We found that the model had 59.2% top-1 accuracy out of 100 possible classes for 'in the wild' 8k celebrity images. However, this performance dropped to 43.3% when we increased our class sizes to 1k celebrity names. This performance is not competitive when compared to production level models such as Google's Celebrity Recognition (Google). However, what makes these results noteworthy is that this analysis was done using only zero-shot identification capabilities based on names inferred from pre-training data - we didn't use any additional task-specific dataset, and so the (relatively) strong results further indicate that before deploying multimodal models, people will need to carefully study them for behaviors in a given context and domain.

CLIP offers significant benefit for tasks that have relatively little data given its zero-shot capabilities. However, large datasets and high performing supervised models exist for many in-demand surveillance tasks such as facial recognition. As a result, CLIP's comparative appeal for such uses is low. Additionally, CLIP is not designed for common surveillance-relevant tasks like object detection and semantic segmentation. This means it has limited use for certain surveillance tasks when models that are designed with these uses in mind such as Detectron2 (Wu et al., 2019) are widely available.

However, CLIP does unlock a certain aspect of usability given how it removes the need for training data. Thus, CLIP and similar models could enable bespoke, niche surveillance use cases for which no well-tailored models or datasets exist, and could lower the skill requirements to build such applications. As our experiments show, ZS CLIP displays non-trivial, but not exceptional, performance on a few surveillance relevant tasks today.

7.3. Future Work

This preliminary analysis is intended to illustrate some of the challenges that general purpose computer vision models pose and to give a glimpse into their biases and impacts. We hope that this work motivates future research on the characterization of the capabilities, shortcomings, and biases of such models, and we are excited to engage with the research community on such questions.

We believe one good step forward is community exploration to further characterize the capabilities of models like CLIP and - crucially - identify application areas where they have promising performance and areas where they may have reduced performance[9]. This process of characterization can help researchers increase the likelihood models are used beneficially by:

- Identifying potentially beneficial downstream uses of models early in the research process, enabling other researchers to think about applications.
- Surfacing tasks with significant sensitivity and a large set of societal stakeholders, which may call for intervention by policymakers.
- Better characterizing biases in models, alerting other researchers to areas of concern and areas for interventions.
- Creating suites of tests to evaluate systems like CLIP on, so we can better characterize model capabilities earlier in the development cycle.
- Identifying potential failure modes and areas for further work.

We plan to contribute to this work, and hope this analysis provides some motivating examples for subsequent research.

8. Related Work

Any model that leverages written, spoken, signed or any other form of human language as part of its training signal is arguably using natural language as a source of supervision. This is an admittedly extremely broad area and covers most work in the field of distributional semantics including topic models (Blei et al., 2003), word, sentence, and paragraph vectors (Mikolov et al., 2013; Kiros et al., 2015; Le & Mikolov, 2014), and language models (Bengio et al., 2003). It also includes much of the broader field of NLP that deals with predicting or modeling sequences of natural language in some way. Work in NLP intentionally leveraging natural language supervision in the form of explanations, feedback, instructions, and advice for tasks such as classification (as opposed to the commonly used representation of supervision as a set of arbitrarily encoded discrete category labels) has

[9] A model could be unfit for use due to inadequate performance or due to the inappropriateness of AI use in the application area itself.

Model	100 Classes	1k Classes	2k Classes
CLIP L/14	59.2	43.3	42.2
CLIP RN50x64	56.4	39.5	38.4
CLIP RN50x16	52.7	37.4	36.3
CLIP RN50x4	52.8	38.1	37.3

表8. CelebA零样本Top-1身份识别准确率

反映了自然语言处理领域的最新进展，其中基于互联网数据训练的大型语言模型常展现出令人惊讶的能力，能够提供与相对次要的公众人物相关的信息（Brown等人，2020年）。

我们发现，在“野生”的8k名人图像中，该模型在100个可能类别中的top-1准确率达到59.2%。然而，当我们将类别数量增加到1k个名人姓名时，其性能下降至43.3%。与谷歌名人识别（Google）等生产级模型相比，这一表现并不具备竞争力。但值得关注的是，这项分析仅基于预训练数据推断出的姓名进行零样本识别——我们没有使用任何额外的任务特定数据集，因此（相对）强劲的结果进一步表明，在部署多模态模型之前，人们需要仔细研究其在特定上下文和领域中的行为。

CLIP凭借其零样本能力，在数据相对稀缺的任务中展现出显著优势。然而，对于人脸识别等许多高需求的监控任务，目前已存在大规模数据集和性能优越的监督模型，因此CLIP在此类应用中的相对吸引力较低。此外，CLIP并非针对目标检测和语义分割等常见监控相关任务而设计。这意味着当Detectron2（Wu等人，2019）这类专为监控场景设计的模型已被广泛使用时，CLIP在某些监控任务中的应用范围会受到限制。

然而，CLIP确实解锁了某种可用性，因为它消除了对训练数据的需求。因此，CLIP及类似模型能够实现定制化、小众的监控应用场景，这些场景目前缺乏量身定制的模型或数据集，同时还能降低构建此类应用的技术门槛。正如我们的实验所示，当前在少数监控相关任务上，零样本CLIP表现出尚可但并非卓越的性能。

7.3. 未来工作

这一初步分析旨在说明通用计算机视觉模型带来的一些挑战，并揭示其偏见与影响。

我们希望这项工作能推动未来研究，深入探讨此类模型的能力、局限性与偏见特征，并期待与研究界就这些问题展开交流。

我们认为，社区探索是向前迈出的重要一步，它能进一步描述像CLIP这样的模型的能力，并关键地识别出它们具有良好性能的应用领域以及可能性能较弱的领域[9]。这种特性描述的过程可以帮助研究人员提高模型被有益使用的可能性，具体通过：

- 在研究过程的早期识别模型潜在有益的后续用途，使其他研究人员能够思考应用。
- 浮现出具有显著敏感性和广泛社会利益相关者的任务，可能需要政策制定者介入干预。
- 更好地描述模型中的偏差，提醒其他研究人员关注需要干预的领域和问题区域。
- 创建测试套件来评估像CLIP这样的系统，以便我们能在开发周期早期更好地描述模型能力。
- 识别潜在故障模式及需要进一步工作的领域。

我们计划为这项工作做出贡献，并希望此分析能为后续研究提供一些具有启发性的示例。

8. 相关工作

任何模型，只要其利用书面、口头、手语或其他任何形式的人类语言作为训练信号的一部分，都可以说是在使用自然语言作为监督来源。这无疑是一个极其广泛的领域，涵盖了分布语义学领域的大部分工作，包括主题模型（Blei等人，2003年）、词向量、句子向量和段落向量（Mikolov等人，2013年；Kiros等人，2015年；Le和Mikolov，2014年）以及语言模型（Bengio等人，2003年）。它还包括更广泛的自然语言处理领域，该领域以某种方式处理自然语言序列的预测或建模。在自然语言处理中，有意利用自然语言监督（以解释、反馈、指令和建议等形式）来完成分类等任务（与通常将监督表示为一组任意编码的离散类别标签相反）的工作

[9] A model could be unfit for use due to inadequate performance or due to the inappropriateness of AI use in the application area itself.

been explored in many creative and advanced ways. Dialog based learning (Weston, 2016; Li et al., 2016; Hancock et al., 2019) develops techniques to learn from interactive natural language feedback in dialog. Several papers have leveraged semantic parsing to convert natural language explanations into features (Srivastava et al., 2017) or additional training labels (Hancock et al., 2018). More recently, ExpBERT (Murty et al., 2020) uses feature representations produced by conditioning a deep contextual language model on natural language explanations and descriptions of relations to improve performance on the task of relation extraction.

CLIP is an example of using natural language as a training signal for learning about a domain other than language. In this context, the earliest use of the term *natural language supervision* that we are aware of is the work of Ramanathan et al. (2013) which showed that natural language descriptions could be used along side other sources of supervision to improve performance on the task of video event understanding. However, as mentioned in the introduction and approach section, methods of leveraging natural language descriptions in computer vision well predate the use of this specific term, especially for image retrieval (Mori et al., 1999) and object classification (Wang et al., 2009). Other early work leveraged tags (but not natural language) associated with images for the task of semantic segmentation (Barnard et al., 2003). More recently, He & Peng (2017) and Liang et al. (2020) demonstrated using natural language descriptions and explanations to improve fine-grained visual classification of birds. Others have investigated how grounded language can be used to improve visual representations and classifiers on the ShapeWorld dataset (Kuhnle & Copestake, 2017; Andreas et al., 2017; Mu et al., 2019). Finally, techniques which combine natural language with reinforcement learning environments (Narasimhan et al., 2015) have demonstrated exciting emergent behaviors such as systematically accomplishing zero-shot tasks (Hill et al., 2019).

CLIP's pre-training task optimizes for text-image retrieval. This areas of research dates back to the mid-90s with the previously mentioned Mori et al. (1999) as representative of early work. While initial efforts focused primarily on predictive objectives over time research shifted towards learning joint multi-modal embedding spaces with techniques like kernel Canonical Correlation Analysis and various ranking objectives (Weston et al., 2010; Socher & Fei-Fei, 2010; Hodosh et al., 2013). Over time work explored many combinations of training objective, transfer, and more expressive models and steadily improved performance (Frome et al., 2013; Socher et al., 2014; Karpathy et al., 2014; Kiros et al., 2014; Faghri et al., 2017).

Other work has leveraged natural language supervision for domains other than images. Stroud et al. (2020) explores large scale representation learning by training a system to pair descriptive text with videos instead of images. Several works have explored using dense spoken natural language supervision for videos (Miech et al., 2019; 2020b). When considered together with CLIP, these works suggest that large scale natural language supervision is a promising way to learn high quality perceptual systems for many domains. Alayrac et al. (2020) extended this line of work to an additional modality by adding raw audio as an additional supervision source and demonstrated benefits from combining all three sources of supervision.

As part of our work on CLIP we also construct a new dataset of image-text pairs. Modern work on image-text retrieval has relied on a set of crowd-sourced sentence level image caption evaluation datasets like Pascal1K (Rashtchian et al., 2010), Flickr8K (Hodosh et al., 2013), and Flickr30K (Young et al., 2014). However, these datasets are still relatively small and limit achievable performance. Several methods have been proposed to create larger datasets automatically with Ordonez et al. (2011) as a notable early example. In the deep learning era, Mithun et al. (2018) demonstrated an additional set of (image, text) pairs collected from the internet could improve retrieval performance and several new automatically constructed datasets such as Conceptual Captions (Sharma et al., 2018), LAIT (Qi et al., 2020), and OCR-CC (Yang et al., 2020) have been created. However, these datasets still use significantly more aggressive filtering or are designed for a specific task such as OCR and as a result are still much smaller than WIT with between 1 and 10 million training examples.

A related idea to CLIP is webly supervised learning. This line of work queries image search engines to build image datasets by querying for terms and uses the queries as the labels for the returned images (Fergus et al., 2005). Classifiers trained on these large but noisily labeled datasets can be competitive with those trained on smaller carefully labeled datasets. These image-query pairs are also often used to improve performance on standard datasets as additional training data (Chen & Gupta, 2015). CLIP also uses search queries as part of its dataset creation process. However CLIP only uses full text sequences co-occuring with images as supervision rather than just the queries, which are often only a single word or short n-gram. We also restrict this step in CLIP to text only querying for sub-string matches while most webly supervised work uses standard image search engines which have their own complex retrieval and filtering pipelines that often involve computer vision systems. Of this line of work, *Learning Everything about Anything: Webly-Supervised Visual Concept Learning* (Divvala et al., 2014) has a notably similar ambition and goal as CLIP.

Finally, CLIP is related to a recent burst of activity on learning joint models of vision and language (Lu et al., 2019; Tan

已在许多创新和先进的方式中得到探索。基于对话的学习（Weston, 2016; Li et al., 2016; Hancock et al., 2019）开发了从对话中的交互式自然语言反馈中学习的技术。多篇论文利用语义解析将自然语言解释转化为特征（Srivastava et al., 2017）或额外的训练标签（Hancock et al., 2018）。最近，ExpBERT（Murty et al., 2020）通过将深度上下文语言模型基于自然语言解释和关系描述来生成特征表示，从而提升关系抽取任务的性能。

CLIP是利用自然语言作为训练信号来学习语言以外领域的一个范例。在此背景下，我们所知最早使用术语*natural languagesupervision*的研究是Ramanathan等人（2013）的工作，该研究表明自然语言描述可以与其他监督源结合使用，以提升视频事件理解任务的性能。然而，如引言和方法部分所述，在计算机视觉中利用自然语言描述的方法远早于这一特定术语的使用，尤其是在图像检索（Mori等人，1999）和物体分类（Wang等人，2009）领域。其他早期研究利用与图像关联的标签（但非自然语言）进行语义分割任务（Barnard等人，2003）。近年来，He与Peng（2017）以及Liang等人（2020）证明了使用自然语言描述和解释可以提升鸟类细粒度视觉分类的效果。另有研究探索了如何在ShapeWorld数据集上运用具身语言改进视觉表征和分类器（Kuhnle & Copestake，2017；Andreas等人，2017；Mu等人，2019）。最后，将自然语言与强化学习环境相结合的技术（Narasimhan等人，2015）已展现出令人兴奋的涌现行为，例如系统性地完成零样本任务（Hill等人，2019）。

CLIP的预训练任务以优化图文检索为目标。这一研究领域可追溯至90年代中期，之前提到的Mori等人（1999年）的研究即是早期工作的代表。尽管最初的研究主要聚焦于预测性目标，但随着时间的推移，研究重心逐渐转向通过核典型相关分析及各类排序目标等技术学习联合多模态嵌入空间（Weston等人，2010年；Socher与Fei-Fei，2010年；Hodosh等人，2013年）。后续研究不断探索训练目标、迁移策略与更具表现力的模型之间的多种组合，逐步提升了性能表现（Frome等人，2013年；Socher等人，2014年；Karpathy等人，2014年；Kiros等人，2014年；Faghri等人，2017年）。

其他研究已利用自然语言监督应用于图像以外的领域。Stroud等人（2020年）探索了通过训练系统将描述性文本与视频而非图像配对，实现大规模表征学习。多项研究探索了利用密集口语自然语言监督处理视频的方法（Miech等人，2019；2020b）。结合CLIP来看，这些研究表明大规模自然语言监督是跨多个领域构建高质量感知系统的有效途径。Alayrac等人（2020）通过引入原始音频作为额外监督源，将这一研究方向扩展至新模态，并论证了融合三种监督源带来的显著优势。

作为CLIP研究的一部分，我们还构建了一个新的图文对数据集。现代图文检索工作依赖于一系列众包句子级图像描述评估数据集，如Pascal1K（Rashtchian等人，2010）、Flickr8K（Hodosh等人，2013）和Flickr30K（Young等人，2014）。然而，这些数据集规模仍相对有限，制约了性能的进一步提升。已有多种方法被提出用于自动构建更大规模的数据集，其中Ordonez等人（2011）的研究是早期重要范例。在深度学习时代，Mithun等人（2018）证明从互联网收集的额外图文对能提升检索性能，随后出现了多个新型自动构建数据集，如Conceptual Captions（Sharma等人，2018）、LAIT（Qi等人，2020）和OCR-CC（Yang等人，2020）。但这些数据集仍采用明显更严格的过滤策略，或专为OCR等特定任务设计，其规模（约100万至1000万训练样本）仍远小于WIT数据集。

与CLIP相关的一个概念是网络监督学习。这一研究方向通过查询图像搜索引擎来构建图像数据集，即使用查询词作为检索图像的标签（Fergus等人，2005）。在这些规模庞大但标注噪声较多的数据集上训练的分类器，其性能可与在较小规模但精细标注的数据集上训练的模型相媲美。这类图像-查询对也常被用作额外训练数据，以提升在标准数据集上的性能表现（Chen & Gupta，2015）。CLIP在其数据集构建过程中同样采用了搜索查询，但CLIP仅使用与图像共现的完整文本序列作为监督信号，而非通常仅为单个词或短n-gram的查询词。此外，CLIP在此步骤中限制为仅通过文本进行子字符串匹配查询，而多数网络监督研究采用标准的图像搜索引擎——这些引擎自带复杂的检索过滤流程，且常涉及计算机视觉系统。在该研究方向中，*Learning Everything about Anything: Webly-Supervised Visual Concept Learning* (Divvala等人（2014）)的工作目标与CLIP有着显著的相似性。

最后，CLIP与近期在视觉与语言联合模型学习方面涌现的大量研究相关（Lu等人，2019；Tan

& Bansal, 2019; Chen et al., 2019; Li et al., 2020b; Yu et al., 2020). This line of work focuses on richly connecting vision and language in order to solve complex downstream tasks such as visual question answering, visual commonsense reasoning, or multimodal entailment. These approaches leverage impressively engineered models which combine 3 (or more) pre-trained subsystems, typically an image feature model, a region proposal / object detection model, and a pre-trained masked language model such as BERT. These systems are then jointly fine-tuned via various training objectives on image-text pairs and applied to the aforementioned tasks and achieve impressive results. CLIP is instead focused on learning visual models from scratch via natural language supervision and does not densely connect the two domains with a joint attention model. The only interaction in a CLIP model between the image and text domain is a single dot product in a learned joint embedding space. We are excited to see CLIP hybridized with this line of work.

9. Conclusion

We have investigated whether it is possible to transfer the success of task-agnostic web-scale pre-training in NLP to another domain. We find that adopting this formula results in similar behaviors emerging in the field of computer vision and discuss the social implications of this line of research. In order to optimize their training objective, CLIP models learn to perform a wide variety of tasks during pre-training. This task learning can then be leveraged via natural language prompting to enable zero-shot transfer to many existing datasets. At sufficient scale, the performance of this approach can be competitive with task-specific supervised models although there is still room for much improvement.

ACKNOWLEDGMENTS

We'd like to thank the millions of people involved in creating the data CLIP is trained on. We'd also like to thank Susan Zhang for her work on image conditional language models while at OpenAI, Ishaan Gulrajani for catching an error in the pseudocode, and Irene Solaiman, Miles Brundage, and Gillian Hadfield for their thoughtful feedback on the broader impacts section of the paper. We are also grateful to the Acceleration and Supercomputing teams at OpenAI for their critical work on software and hardware infrastructure this project used. Finally, we'd also like to thank the developers of the many software packages used throughout this project including, but not limited, to Numpy (Harris et al., 2020), SciPy (Virtanen et al., 2020), ftfy (Speer, 2019), TensorFlow (Abadi et al., 2016), PyTorch (Paszke et al., 2019), pandas (pandas development team, 2020), and scikit-learn (Pedregosa et al., 2011).

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pp. 265–283, 2016.

Alayrac, J.-B., Recasens, A., Schneider, R., Arandjelović, R., Ramapuram, J., De Fauw, J., Smaira, L., Dieleman, S., and Zisserman, A. Self-supervised multimodal versatile networks. *arXiv preprint arXiv:2006.16228*, 2020.

Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S., and Nguyen, A. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4845–4854, 2019.

Andreas, J., Klein, D., and Levine, S. Learning with latent language. *arXiv preprint arXiv:1711.00482*, 2017.

Assiri, Y. Stochastic optimization of plain convolutional neural networks with simple methods. *arXiv preprint arXiv:2001.08856*, 2020.

Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pp. 15535–15545, 2019.

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, pp. 9453–9463, 2019.

Barnard, K., Duygulu, P., Forsyth, D., Freitas, N. d., Blei, D. M., and Jordan, M. I. Matching words and pictures. *Journal of machine learning research*, 3(Feb):1107–1135, 2003.

Bechmann, A. and Bowker, G. C. Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media. *Big Data & Society*, 6(1):205395171881956, January 2019. doi: 10.1177/2053951718819569. URL https://doi.org/10.1177/2053951718819569.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

Bhargava, S. and Forsyth, D. Exposing and correcting the gender bias in image captioning datasets and models. *arXiv preprint arXiv:1912.00578*, 2019.

& Bansal, 2019; Chen et al., 2019; Li et al., 2020b; Yu et al., 2020）。这一系列工作的重点在于密集连接视觉与语言，以解决复杂的下游任务，如视觉问答、视觉常识推理或多模态蕴含。这些方法利用了设计精良的模型，这些模型结合了三个（或更多）预训练子系统，通常包括一个图像特征模型、一个区域提议/目标检测模型，以及一个预训练的掩码语言模型（如BERT）。然后，这些系统通过图像-文本对上的各种训练目标进行联合微调，并应用于上述任务，取得了令人瞩目的成果。而CLIP则侧重于通过自然语言监督从头开始学习视觉模型，并未通过联合注意力模型密集连接这两个领域。在CLIP模型中，图像与文本领域之间唯一的交互是在一个习得的联合嵌入空间中进行一次点积运算。我们期待看到CLIP与这一系列工作相结合。

9. 结论

我们研究了是否有可能将NLP中任务无关的网络规模预训练的成功经验迁移到另一个领域。我们发现，采用这一方法{v*}在计算机视觉领域催生了类似的行为，并探讨了这类研究的社会影响。为了优化训练目标，CLIP模型在预训练期间学会了执行多种任务。这种任务学习能力随后可以通过自然语言提示来利用，实现对许多现有数据集的零样本迁移。在足够规模下，该方法的性能可与针对特定任务的监督模型相媲美，尽管仍有很大的改进空间。

致谢

我们要感谢参与创建CLIP训练数据的数百万人。我们还要感谢Susan Zhang在OpenAI期间对图像条件语言模型的研究，Ishaan Gulrajani对伪代码中错误的指正，以及Irene Solaiman、Miles Brundage和Gillian Hadfield对论文中更广泛影响部分提出的深刻反馈。同时，我们由衷感谢OpenAI的加速与超级计算团队，他们为该项目所依赖的软硬件基础设施做出了关键贡献。最后，我们也要感谢本项目使用的众多软件包的开发者，包括但不限于Numpy（Harris等人，2020）、SciPy（Virtanen等人，2020）、ftfy（Speer，2019）、TensorFlow（Abadi等人，2016）、PyTorch（Paszke等人，2019）、pandas（pandas开发团队，2020）以及scikit-learn（Pedregosa等人，2011）。

参考文献

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., 等. Tensorflow: 一种用于大规模机器学习的系统。载于 *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 第 265–283 页, 2016. Alayrac, J.-B., Recasens, A., Schneider, R., Arandjelović, R., Ramapuram, J., De Fauw, J., Smaira, L., Dieleman, S., 与 Zisserman, A. 自监督多模态通用网络。*arXiv preprint arXiv:2006.16228*, 2020. Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.- S., 与 Nguyen, A. 摆个姿势（就）攻击：神经网络容易被熟悉物体的奇怪姿势欺骗。载于 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 第 4845–4854 页, 2019. Andreas, J., Klein, D., 与 Levine, S. 用潜在语言学习。*arXiv preprint arXiv:1711.00482*, 2017. Assiri, Y. 使用简单方法对普通卷积神经网络进行随机优化。*arXiv preprint arXiv:2001.08856*, 2020. Bachman, P., Hjelm, R. D., 与 Buchwalter, W. 通过跨视图最大化互信息来学习表示。载于 *Advances in Neural Information Processing Systems*, 第 15535–15545 页, 2019. Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gut- freund, D., Tenenbaum, J., 与 Katz, B. Objectnet: 一个用于推动物体识别模型极限的大规模偏置控制数据集。载于 *Advances in Neural Information Processing Systems*, 第 9453–9463 页, 2019. Barnard, K., Duygulu, P., Forsyth, D., Freitas, N. d., Blei, D. M., 与 Jordan, M. I. 匹配词语与图片。*Journal of machine learning research*, 第3卷(2月):1107–1135, 2003. Bechmann, A. 与 Bowker, G. C. 任何其他名称下的无监督：社交媒体人工智能中知识生产的隐藏层面。*Big Data & Society*, 第6卷(1):205395171881956, 2019年1月. doi: 10.1177/ 2053951718819569. URL https://doi.org/10. 1177/2053951718819569. Bengio, Y., Ducharme, R., Vincent, P., 与 Jauvin, C. 一种神经概率语言模型。*Journal of machine learning research*, 第3卷(2月):1137–1155, 2003. Bhargava, S. 与 Forsyth, D. 揭示并修正图像描述数据集和模型中的性别偏见。*arXiv preprint arXiv:1912.00578*, 2019.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022, 2003.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357, 2016.

Bowker, G. C. and Star, S. L. *Sorting things out: Classification and its consequences*. MIT press, 2000.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Browne, S. *Dark Matters: Surveillance of Blackness*. Duke University Press, 2015.

Bulent Sariyildiz, M., Perez, J., and Larlus, D. Learning visual representations with caption annotations. *arXiv e-prints*, pp. arXiv–2008, 2020.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91, 2018.

Carreira, J., Noland, E., Hillier, C., and Zisserman, A. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I. Generative pretraining from pixels. In *International Conference on Machine Learning*, pp. 1691–1703. PMLR, 2020a.

Chen, T., Xu, B., Zhang, C., and Guestrin, C. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020b.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020c.

Chen, X. and Gupta, A. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1431–1439, 2015.

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020d.

Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019.

Cheng, G., Han, J., and Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Choi, D., Shallue, C. J., Nado, Z., Lee, J., Maddison, C. J., and Dahl, G. E. On empirical comparisons of optimizers for deep learning. *arXiv preprint arXiv:1910.05446*, 2019.

Coates, A., Ng, A., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223, 2011.

Crawford, K. The trouble with bias. *NIPS 2017 Keynote*, 2017. URL https://www.youtube.com/watch?v=fMym_BKWQzk.

Dai, A. M. and Le, Q. V. Semi-supervised sequence learning. In *Advances in neural information processing systems*, pp. 3079–3087, 2015.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Deng, J., Berg, A. C., Satheesh, S., Su, H., Khosla, A., and Fei-Fei, L. Ilsvrc 2012, 2012. URL http://www.image-net.org/challenges/LSVRC/2012/.

Desai, K. and Johnson, J. Virtex: Learning visual representations from textual annotations. *arXiv preprint arXiv:2006.06666*, 2020.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., and Sutskever, I. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 潜在狄利克雷分配。*Journal of machine Learning research*, 3(1月): 993−1022, 2003.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. 男人之于计算机程序员，如同女人之于家庭主妇？消除词嵌入中的偏见。*Advances in neural information processing systems*, 29:4349−4357, 2016.

鲍克，G. C. 和斯塔尔，S. L. *Sorting things out: Classifica- tion and its consequences*。麻省理工学院出版社，2000年。

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., 等。语言模型是小样本学习者。*arXiv preprint arXiv:2005.14165*，2020。

布朗，S. *Dark Matters: Surveillance of Blackness*。杜克大学出版社，2015年。

Bulent Sariyildiz, M., Perez, J., and Larlus, D. 使用标题注释学习视觉表示。*arXiv e-prints*, pp. arXiv−2008, 2020.

Buolamwini, J. 和 Gebru, T. 性别阴影：商业性别分类中的交叉准确性差异。于 *Conference on fairness, accountability and transparency*，第77−91页，2018年。

Carreira, J., Noland, E., Hillier, C., 和 Zisserman, A. 关于 Kinetics-700 人类动作数据集的简短说明。*arXiv preprint arXiv:1907.06987*, 2019.

陈明、拉德福德、柴尔德、吴军、俊、栾德和苏茨克维尔。从像素生成预训练。于 *International Conference on Machine Learning*，第1691−1703页。PMLR，2020a。

陈天奇、徐波、张弛和Guestrin, C. 训练具有次线性内存成本的深度网络。*arXiv preprint arXiv:1604.06174*，2016年。

陈天奇、西蒙·科恩布利斯、穆罕默德·诺鲁齐和杰弗里·辛顿，视觉表征对比学习的简单框架。*arXiv preprint arXiv:2002.05709*，2020b。

陈天奇、科恩布利斯、斯沃斯基、诺鲁齐和辛顿，《大型自监督模型是强大的半监督学习器》。*arXiv preprint arXiv:2006.10029*，2020c。

陈、X. 与 Gupta, A. 卷积网络的网络监督学习。收录于 *Proceedings of the IEEE International Conference on Computer Vision*，第 1431−1439 页，2015 年。

陈、范、Girshick、何。基于动量对比学习的改进基线。*arXiv preprint arXiv:2003.04297*，2020d。

陈，Y.-C.，李，L.，余，L.，Kholy，A. E.，Ahmed，F.，甘，Z.，程，Y.，和刘，J. Uniter：学习通用图像文本表示。*arXiv preprint arXiv:1909.11740*，2019。

程、韩、卢。遥感图像场景分类：基准与最新进展。*Proceed- ings of the IEEE*, 105(10):1865−1883, 2017。

Choi, D., Shallue, C. J., Nado, Z., Lee, J., Maddison, C. J., and Dahl, G. E. 关于深度学习优化器的实证比较。*arXiv preprint arXiv:1910.05446*, 2019。

Coates, A., Ng, A., and Lee, H. 无监督特征学习中单层网络的分析。于 *Proceedings of the fourteenth international conference on artificial intelligence and statistics*，第215−223页，2011年。

Crawford, K. 偏见的困扰。*NIPS 2017 Keynote*，2017。网址 https://www.youtube.com/ watch?v=fMym_BKWQzk。

戴, A. M. 和 乐, Q. V. 半监督序列学习。于 *Advances in neural information processing systems*, 页 3079−3087, 2015。

D’Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., 等。在现代机器学习中，欠明确性对可信度提出了挑战。*arXiv preprint arXiv:2011.03395*，2020。

邓嘉、董伟、Socher, R.、李飞飞、李凯、李飞飞。ImageNet：一个大规模分层图像数据库。于 *CVPR09*，2009年。

邓杰、伯格、萨提什、苏、科斯拉和费飞。ILSVRC 2012，2012年。网址 http://www.image-net.org/challenges/LSVRC/2012/。

Desai, K. 和 Johnson, J. Virtex：从文本注释中学习视觉表示。*arXiv preprint arXiv:2006.06666*，2020。

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT：用于语言理解的深度双向Transformer预训练。*arXiv preprint arXiv:1810.04805*，2018。

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A. 与 Sutskever, I. Jukebox：一种音乐生成模型。*arXiv preprint arXiv:2005.00341*，2020。

Divvala, S. K., Farhadi, A., and Guestrin, C. Learning everything about anything: Webly-supervised visual concept learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3270–3277, 2014.

Dodge, S. and Karam, L. A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th international conference on computer communication and networks (ICCCN)*, pp. 1–7. IEEE, 2017.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Elhoseiny, M., Saleh, B., and Elgammal, A. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2584–2591, 2013.

Faghri, F., Fleet, D. J., Kiros, J. R., and Fidler, S. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

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

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pp. 2121–2129, 2013.

Gan, Z., Chen, Y.-C., Li, L., Zhu, C., Cheng, Y., and Liu, J. Large-scale adversarial training for vision-and-language representation learning. *arXiv preprint arXiv:2006.06195*, 2020.

Gao, T., Fisch, A., and Chen, D. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020.

Garvie, C., May 2019. URL https://www.flawedfacedata.com/.

Geiger, A., Lenz, P., and Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.

Gomez, L., Patel, Y., Rusiñol, M., Karatzas, D., and Jawahar, C. Self-supervised learning of visual features through embedding images into text topic spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4230–4239, 2017.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., et al. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63, 2015.

Google. Google cloud api: Celebrity recognition. URL https://cloud.google.com/vision/docs/celebrity-recognition.

Griewank, A. and Walther, A. Algorithm 799: revolve: an implementation of checkpointing for the reverse or adjoint mode of computational differentiation. *ACM Transactions on Mathematical Software (TOMS)*, 26(1):19–45, 2000.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Ha, D., Dai, A., and Le, Q. V. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

Hancock, B., Bringmann, M., Varma, P., Liang, P., Wang, S., and Ré, C. Training classifiers with natural language explanations. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, pp. 1884. NIH Public Access, 2018.

Hancock, B., Bordes, A., Mazare, P.-E., and Weston, J. Learning from dialogue after deployment: Feed yourself, chatbot! *arXiv preprint arXiv:1901.05415*, 2019.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Fernández del

Divvala, S. K., Farhadi, A., and Guestrin, C. 学习万物：基于网络监督的视觉概念学习。发表于 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*，第3270–3277页，2014年。

Dodge, S. 和 Karam, L. 一项关于人类与深度学习在视觉失真下识别性能的研究与比较。载于 *2017 26th international conference on computer communication and networks (ICCCN)*，第 1–7 页。IEEE，2017。

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., 等。一幅图像相当于16x16个词汇：大规模图像识别的Transformer模型。*arXiv preprint arXiv:2010.11929*, 2020。

Elhoseiny, M., Saleh, B., and Elgammal, A. 编写一个分类器：仅使用文本描述的零样本学习。在 *Proceedings of the IEEE International Conference on Computer Vision*，第2584–2591页，2013年。

Faghri, F., Fleet, D. J., Kiros, J. R., 与 Fidler, S. Vse++: 通过困难负样本改进视觉-语义嵌入。*arXiv preprint arXiv:1707.05612*, 2017.

Fergus, R., Fei-Fei, L., Perona, P., 和 Zisserman, A. 从谷歌图像搜索中学习对象类别。载于 *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*，第2卷，第1816–1823页。IEEE，2005年。

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T. Devise: 一种深度视觉语义嵌入模型。发表于 *Advances in neural information processing systems*，第2121–2129页，2013年。

Gan, Z., Chen, Y.-C., Li, L., Zhu, C., Cheng, Y., and Liu, J. 面向视觉与语言表示学习的大规模对抗训练。*arXiv preprint arXiv:2006.06195*, 2020。

Gao, T., Fisch, A., and Chen, D. 让预训练语言模型成为更好的少样本学习器。*arXiv preprint arXiv:2012.15723*, 2020.

加维，C.，2019年5月。网址 https://www.flawedfacedata.com/。

Geiger, A., Lenz, P., 与 Urtasun, R. 我们是否已为自动驾驶做好准备？KITTI视觉基准测试集。收录于 *Conference on Computer Vision and Pattern Recognition (CVPR)*，2012年。

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. 经ImageNet训练的CNNs是偏向纹理；增加形状偏置能提高准确性和鲁棒性。*arXiv preprint arXiv:1811.12231*, 2018。

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. 深度神经网络中的捷径学习。*arXiv preprint arXiv:2004.07780*, 2020.

Gomez, L., Patel, Y., Rusiñol, M., Karatzas, D., and Jawahar, C. 通过将图像嵌入文本主题空间进行视觉特征的自监督学习。载于 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*，第4230–4239页，2017年。

Goodfellow, I. J., Shlens, J., and Szegedy, C. 对抗样本的解释与利用。*arXiv preprint arXiv:1412.6572*, 2014.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., 等。表示学习中的挑战：三项机器学习竞赛报告。*Neural Networks*, 64:59–63, 2015。

谷歌。谷歌云API：名人识别。网址 https://cloud.google.com/vision/docs/celebrity-recognition。

Griewank, A. 和 Walther, A. 算法 799：revolve：计算微分反向或伴随模式中检查点机制的一种实现。*ACM Trans- actions on Mathematical Software (TOMS)*，26(1):19–45，2000。

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., 等。自展潜在表示：自监督学习的新方法。*arXiv preprint arXiv:2006.07733*，2020。

哈，D.，戴，A.，和乐，Q. V. 超网络。*arXiv preprint arXiv:1609.09106*，2016年。

Hancock, B., Bringmann, M., Varma, P., Liang, P., Wang, S., and Ré, C. 使用自然语言解释训练分类器。载于 *Proceedings of the conference. Association for Computational Linguistics. Meeting*，第2018卷，第1884页。NIH公共访问，2018年。

Hancock, B., Bordes, A., Mazare, P.-E., and Weston, J. 部署后从对话中学习：自我反馈，聊天机器人！*arXiv preprint arXiv:1901.05415*，2019年。

哈里斯，C. R.，米尔曼，K. J.，范德沃尔特，S. J.，戈默斯，R.，维尔塔宁，P.，库尔纳波，D.，维泽，E.，泰勒，J.，伯格，S.，史密斯，N. J.，克恩，R.，皮库斯，M.，霍耶，S.，范克尔克韦克，M. H.，布雷特，M.，霍尔丹，A.，费尔南德斯·德尔

Río, J., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. Array programming with NumPy. *Nature*, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2.

Hays, J. and Efros, A. A. Im2gps: estimating geographic information from a single image. In *2008 ieee conference on computer vision and pattern recognition*, pp. 1–8. IEEE, 2008.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016a.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016b.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., and Li, M. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 558–567, 2019.

He, X. and Peng, Y. Fine-grained image classification via combining vision and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5994–6002, 2017.

Helber, P., Bischke, B., Dengel, A., and Borth, D. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Henaff, O. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pp. 4182–4192. PMLR, 2020.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Hendrycks, D. and Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020a.

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., and Song, D. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020b.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M., Ali, M., Yang, Y., and Zhou, Y. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Hill, F., Lampinen, A., Schneider, R., Clark, S., Botvinick, M., McClelland, J. L., and Santoro, A. Environmental drivers of systematicity and generalization in a situated agent. In *International Conference on Learning Representations*, 2019.

Hodosh, M., Young, P., and Hockenmaier, J. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47: 853–899, 2013.

Hongsuck Seo, P., Weyand, T., Sim, J., and Han, B. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 536–551, 2018.

Howard, J. and Ruder, S. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pp. 125–136, 2019.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Jaderberg, M., Simonyan, K., Vedaldi, A., and Zisserman, A. Deep structured output learning for unconstrained text recognition. *arXiv preprint arXiv:1412.5903*, 2014.

Jaderberg, M., Simonyan, K., Zisserman, A., et al. Spatial transformer networks. *Advances in neural information processing systems*, 28:2017–2025, 2015.

R´o, J.、Wiebe, M.、Peterson, P.、Gérard-Marchant, P.、Sheppard, K.、Reddy, T.、Weckesser, W.、Abbasi, H.、Gohlke, C. 与 Oliphant, T. E. 使用 NumPy 进行数组编程。*Nature*，585:357–362，2020。doi: 10.1038/s41586-020-2649-2。Hays, J. 与 Efros, A. A. Im2gps：从单张图像估计地理信息。收录于 *2008 ieee conference on computer vision and pattern recognition*，第 1–8 页。IEEE，2008。He, K.、Zhang, X.、Ren, S. 与 Sun, J. 深入研究整流器：在 ImageNet 分类上超越人类水平性能。收录于 *Proceedings of the IEEE international conference on computer vision*，第 1026–1034 页，2015。He, K.、Zhang, X.、Ren, S. 与 Sun, J. 用于图像识别的深度残差学习。收录于 *Proceedings of the IEEE conference on computer vision and pattern recognition*，第 770–778 页，2016a。He, K.、Zhang, X.、Ren, S. 与 Sun, J. 用于图像识别的深度残差学习。收录于 *Proceedings of the IEEE conference on computer vision and pattern recognition*，第 770–778 页，2016b。He, K.、Fan, H.、Wu, Y.、Xie, S. 与 Girshick, R. 用于无监督视觉表示学习的动量对比。收录于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第 9729–9738 页，2020。He, T.、Zhang, Z.、Zhang, H.、Zhang, Z.、Xie, J. 与 Li, M. 卷积神经网络图像分类技巧集。收录于 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*，第 558–567 页，2019。He, X. 与 Peng, Y. 通过结合视觉与语言进行细粒度图像分类。收录于 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*，第 5994–6002 页，2017。Helber, P.、Bischke, B.、Dengel, A. 与 Borth, D. Eurosat：用于土地利用和土地覆盖分类的新型数据集和深度学习基准。*IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*，12(7):2217–2226，2019。Henaff, O. 使用对比预测编码进行数据高效图像识别。收录于 *International Conference on Ma- chine Learning*，第 4182–4192 页。PMLR，2020。

Hendrycks, D. 和 Dietterich, T. 神经网络对常见损坏与扰动的鲁棒性基准测试。*arXiv preprint arXiv:1903.12261*，2019。

Hendrycks, D. 与 Gimpel, K. 高斯误差线性单元（gelus）。*arXiv preprint arXiv:1606.08415*，2016。

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. 自然对抗样本。*arXiv preprint arXiv:1907.07174*, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., 等。稳健性的多面性：对分布外泛化的批判性分析。*arXiv preprint arXiv:2006.16241*，2020a。

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., and Song, D. 预训练变换器提升分布外鲁棒性。*arXiv preprint arXiv:2004.06100*，2020b。

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M., Ali, M., Yang, Y., 与 Zhou, Y. 深度学习缩放是可经验预测的。*arXiv preprint arXiv:1712.00409*，2017。

Hill, F., Lampinen, A., Schneider, R., Clark, S., Botvinick, M., McClelland, J. L., and Santoro, A. 环境驱动因素对具身智能体系统性与泛化能力的影响。发表于 *International Conference on Learning Repre- sentations*，2019年。

Hodosh, M., Young, P. 与 Hockenmaier, J. 将图像描述构建为排序任务：数据、模型与评估指标。*Journal of Artificial Intelligence Research*, 47: 853–899, 2013。

Hongsuck Seo, P., Weyand, T., Sim, J., and Han, B. Cplanet：通过地图的组合分区增强图像地理定位。于 *Proceedings of the European Conference on Computer Vision (ECCV)*，第 536–551 页，2018。

Howard, J. 和 Ruder, S. 面向文本分类的通用语言模型微调。*arXiv preprint arXiv:1801.06146*, 2018.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. 对抗样本不是缺陷，而是特征。于 *Advances in Neural Information Processing Systems*，第 125–136 页，2019。

Ioffe, S. 与 Szegedy, C. 批归一化：通过减少内部协变量偏移加速深度网络训练。*arXiv preprint arXiv:1502.03167*，2015。

Jaderberg, M., Simonyan, K., Vedaldi, A., 与 Zisserman, A. 面向无约束文本识别的深度结构化输出学习。*arXiv preprint arXiv:1412.5903*, 2014.

Jaderberg, M., Simonyan, K., Zisserman, A. 等人。空间变换网络。*Advances in neural information processing systems*, 28:2017–2025, 2015。

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017.

Joulin, A., Van Der Maaten, L., Jabri, A., and Vasilache, N. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pp. 67–84. Springer, 2016.

Kalfaoglu, M., Kalkan, S., and Alatan, A. A. Late temporal modeling in 3d cnn architectures with bert for action recognition. *arXiv preprint arXiv:2008.01232*, 2020.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Karpathy, A., Joulin, A., and Fei-Fei, L. F. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pp. 1889–1897, 2014.

Keyes, O. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–22, 2018.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., and Testuggine, D. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790*, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kiros, R., Salakhutdinov, R., and Zemel, R. S. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. Skip-thought vectors. *Advances in neural information processing systems*, 28: 3294–3302, 2015.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. Large scale learning of general visual representations for transfer. *arXiv preprint arXiv:1912.11370*, 2019.

Kornblith, S., Shlens, J., and Le, Q. V. Do better imagenet models transfer better? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2661–2671, 2019.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Kuhnle, A. and Copestake, A. Shapeworld-a new test methodology for multimodal language understanding. *arXiv preprint arXiv:1704.04517*, 2017.

Kärkkäinen, K. and Joo, J. Fairface: Face attribute dataset for balanced race, gender, and age, 2019.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people, 2016.

Lampert, C. H., Nickisch, H., and Harmeling, S. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 951–958. IEEE, 2009.

Larochelle, H., Erhan, D., and Bengio, Y. Zero-data learning of new tasks. 2008.

Le, Q. and Mikolov, T. Distributed representations of sentences and documents. In *International conference on machine learning*, pp. 1188–1196, 2014.

LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

Lee, D.-H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks.

Lei Ba, J., Swersky, K., Fidler, S., et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4247–4255, 2015.

Li, A., Jabri, A., Joulin, A., and van der Maaten, L. Learning visual n-grams from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4183–4192, 2017.

Li, G., Duan, N., Fang, Y., Gong, M., and Jiang, D. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. 2020a.

Li, J., Miller, A. H., Chopra, S., Ranzato, M., and Weston, J. Learning through dialogue interactions by asking questions. *arXiv preprint arXiv:1612.04936*, 2016.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. Clevr: 一个用于组合语言与基础视觉推理的诊断性数据集。于 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 页 2901–2910, 2017。

Joulin, A., Van Der Maaten, L., Jabri, A., 与 Vasilache, N. 从大规模弱监督数据中学习视觉特征。载于 *European Conference on Computer Vision*，第67–84页。Springer，2016。

Kalfaoglu, M., Kalkan, S., 和 Alatan, A. A. 用于动作识别的3D CNN架构中结合BERT的后期时序建模。*arXiv preprint arXiv:2008.01232*, 2020。

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. 神经语言模型的缩放定律。*arXiv preprint arXiv:2001.08361*, 2020.

Karpathy, A., Joulin, A., and Fei-Fei, L. F. 深度片段嵌入用于双向图像句子映射。于 *Advances in neural information processing systems*，第1889–1897页，2014年。

Keyes, O. 性别误判机器：自动性别识别的跨性别/人机交互启示。*Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–22, 2018.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., 与 Testuggine, D. 仇恨性迷因挑战：在多模态迷因中检测仇恨言论。*arXiv preprint arXiv:2005.04790*, 2020。

Kingma, D. P. 和 Ba, J. Adam：一种随机优化方法。*arXiv preprint arXiv:1412.6980*，2014。

Kiros, R., Salakhutdinov, R., 和 Zemel, R. S. 使用多模态神经语言模型统一视觉-语义嵌入。*arXiv preprint arXiv:1411.2539*, 2014.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., 与 Fidler, S. Skip-thought 向量。*Advances in neural information processing systems*, 28: 3294–3302, 2015.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. 用于迁移的大规模通用视觉表征学习。*arXiv preprint arXiv:1912.11370*, 2019.

Kornblith, S., Shlens, J., and Le, Q. V. 更好的ImageNet模型迁移效果更好吗？发表于 *Proceedings of the IEEE conference on computer vision and pattern recognition*，第2661–2671页，2019年。

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., 等. 视觉基因组：利用众包密集图像标注连接语言与视觉. *In- ternational journal of computer vision*, 123(1):32–73, 2017.

Krizhevsky, A., Sutskever, I., 和 Hinton, G. E. 使用深度卷积神经网络进行ImageNet分类。于 *Advances in neural information processing systems*，第1097–1105页，2012年。

Kuhnle, A. 与 Copestake, A. 的《Shapeworld——一种多模态语言理解的新测试方法》。*arXiv preprint arXiv:1704.04517*，2017年。

Kärkkäinen, K. 与 Joo, J. Fairface：面向平衡种族、性别和年龄的人脸属性数据集，2019。

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. 构建像人类一样学习和思考的机器，2016。

Lampert, C. H., Nickisch, H. 与 Harmeling, S. 通过类间属性迁移学习检测未见过的物体类别。发表于 *2009 IEEE Conference on Computer Vision and Pattern Recognition*，第951–958页。IEEE，2009年。

Larochelle, H., Erhan, D., 和 Bengio, Y. 新任务的零数据学习。2008。

Le, Q. 和 Mikolov, T. 句子和文档的分布式表示。于 *International conference on machine learning*，第 1188–1196 页，2014 年。

LeCun, Y. 手写数字的mnist数据库。*http://yann. lecun. com/exdb/mnist/*。

Lee, D.-H. 伪标签：一种用于深度神经网络的简单高效半监督学习方法。

雷巴, J., 斯沃斯基, K., 菲德勒, S., 等. 使用文本描述预测深度零样本卷积神经网络. 于 *Proceedings of the IEEE International Conference on Computer Vision*, 页 4247–4255, 2015.

Li, A., Jabri, A., Joulin, A., and van der Maaten, L. 从网络数据中学习视觉 n-gram。于 *Proceedings of the IEEE International Conference on Computer Vision*，第 4183–4192 页，2017。

Li, G., Duan, N., Fang, Y., Gong, M., and Jiang, D. Unicoder-vl：一种通过跨模态预训练实现的视觉与语言通用编码器。2020a。

Li, J., Miller, A. H., Chopra, S., Ranzato, M., and Weston, J. 通过提问进行对话交互学习。*arXiv preprint arXiv:1612.04936*, 2016.

Li, X., Yin, X., Li, C., Hu, X., Zhang, P., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. *arXiv preprint arXiv:2004.06165*, 2020b.

Liang, W., Zou, J., and Yu, Z. Alice: Active learning with contrastive natural language explanations. *arXiv preprint arXiv:2009.10259*, 2020.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Linzen, T. How can we accelerate progress towards human-like linguistic generalization? *arXiv preprint arXiv:2005.00955*, 2020.

Lippe, P., Holla, N., Chandra, S., Rajamanickam, S., Antoniou, G., Shutova, E., and Yannakoudakis, H. A multimodal framework for the detection of hateful memes. *arXiv preprint arXiv:2012.12871*, 2020.

Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., and Shazeer, N. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*, 2018.

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. A sober look at the unsupervised learning of disentangled representations and their evaluation. *arXiv preprint arXiv:2010.14766*, 2020.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pp. 13–23, 2019.

Lu, Z., Xiong, X., Li, Y., Stroud, J., and Ross, D. Leveraging weakly supervised data and pose representation for action recognition, 2020. URL https://www.youtube.com/watch?v=KOQFxbPPLOE&t=1390s.

Lucic, M., Kurach, K., Michalski, M., Gelly, S., and Bousquet, O. Are gans created equal? a large-scale study. *Advances in neural information processing systems*, 31: 700–709, 2018.

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and van der Maaten, L. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 181–196, 2018.

McCann, B., Bradbury, J., Xiong, C., and Socher, R. Learned in translation: Contextualized word vectors. In *Advances in neural information processing systems*, pp. 6294–6305, 2017.

McCann, B., Keskar, N. S., Xiong, C., and Socher, R. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*, 2018.

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017.

Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., and Sivic, J. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE international conference on computer vision*, pp. 2630–2640, 2019.

Miech, A., Alayrac, J.-B., Laptev, I., Sivic, J., and Zisserman, A. Rareact: A video dataset of unusual interactions. *arXiv preprint arXiv:2008.01018*, 2020a.

Miech, A., Alayrac, J.-B., Smaira, L., Laptev, I., Sivic, J., and Zisserman, A. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9879–9889, 2020b.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26:3111–3119, 2013.

Miller, J., Krauth, K., Recht, B., and Schmidt, L. The effect of natural distribution shift on question answering models. *arXiv preprint arXiv:2004.14444*, 2020.

Mishra, A., Alahari, K., and Jawahar, C. Scene text recognition using higher order language priors. 2012.

Mithun, N. C., Panda, R., Papalexakis, E. E., and Roy-Chowdhury, A. K. Webly supervised joint embedding for cross-modal image-text retrieval. In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 1856–1864, 2018.

Mori, Y., Takahashi, H., and Oka, R. Image-to-word transformation based on dividing and vector quantizing images with words. Citeseer, 1999.

Mu, J., Liang, P., and Goodman, N. Shaping visual representations with language for few-shot classification. *arXiv preprint arXiv:1911.02683*, 2019.

Li, X., Yin, X., Li, C., Hu, X., Zhang, P., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., 等. Oscar：面向视觉-语言任务的物体语义对齐预训练模型。*arXiv preprint arXiv:2004.06165*, 2020b.

梁、邹、于。Alice：基于对比自然语言解释的主动学习。*arXiv preprint arXiv:2009.10259*，2020年。

林，T.-Y.，Maire，M.，Belongie，S.，Hays，J.，Perona，P.，Ramanan，D.，Dollár，P.，和 Zitnick，C. L. Microsoft COCO：上下文中的常见物体。载于 *European conference on computer vision*，第740–755页。Springer，2014年。

Linzen, T. 我们如何加速实现类人语言泛化的进展? *arXiv preprint arXiv:2005.00955*, 2020.

Lippe, P., Holla, N., Chandra, S., Rajamanickam, S., Antonious, G., Shutova, E., and Yannakoudakis, H. 一种用于检测仇恨性表情包的多模态框架。*arXiv preprint arXiv:2012.12871*, 2020。

刘，P. J.，萨利赫，M.，波特，E.，古德里奇，B.，塞帕西，R.，凯泽，L.，和沙泽尔，N.。通过总结长序列生成维基百科。*arXiv preprint arXiv:1801.10198*，2018年。

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. 对无监督解耦表示学习及其评估的冷静审视。*arXiv preprint arXiv:2010.14766*, 2020。

Loshchilov, I. 和 Hutter, F. Sgdr：带热重启的随机梯度下降。*arXiv preprint arXiv:1608.03983*，2016。

Loshchilov, I. 和 Hutter, F. 解耦权重衰减正则化。*arXiv preprint arXiv:1711.05101*，2017。

Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert：面向视觉与语言任务的与任务无关的视觉语言表示预训练。于 *Advances in Neural Information Processing Systems*，第13–23页，2019年。

Lu, Z., Xiong, X., Li, Y., Stroud, J., and Ross, D. 利用弱监督数据和姿态表示进行动作识别，2020。URL https://www.youtube.com/watch?v=KOQFxbPPLOE&t=1390s。

Lucic, M., Kurach, K., Michalski, M., Gelly, S., and Bousquet, O. GAN的生成效果是否等同？一项大规模研究。*Advances in neural information processing systems*, 31: 700–709, 2018.

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and van der Maaten, L. 探索弱监督预训练的极限。载于 *Proceedings of the European Conference on Computer Vision (ECCV)*，第181–196页，2018年。

McCann, B., Bradbury, J., Xiong, C., 与 Socher, R. 翻译中的学习：上下文相关的词向量。于 *Advances in neural information processing systems*，第 6294–6305 页，2017。

McCann, B., Keskar, N. S., Xiong, C., 与 Socher, R. 自然语言十项全能：多任务学习作为问答。*arXiv preprint arXiv:1806.08730*，2018。

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., 等。混合精度训练。*arXiv preprint arXiv:1710.03740*，2017。

Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., and Sivic, J. Howto100m：通过观看上亿条旁白视频片段学习文本-视频嵌入。于 *Proceedings of the IEEE international conference on computer vision*，第 2630–2640 页，2019年。

Miech, A., Alayrac, J.-B., Laptev, I., Sivic, J., 与 Zisserman, A. Rareact：一个记录异常交互的视频数据集。*arXiv preprint arXiv:2008.01018*，2020a。

Miech, A., Alayrac, J.-B., Smaira, L., Laptev, I., Sivic, J., 和 Zisserman, A. 从未经筛选的教学视频中端到端学习视觉表示。于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第9879–9889页，2020b。

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. 词与短语的分布式表示及其组合性。*Advances in neural informa- tion processing systems*, 26:3111–3119, 2013.

Miller, J., Krauth, K., Recht, B., and Schmidt, L. 自然分布偏移对问答模型的影响。*arXiv preprint arXiv:2004.14444*, 2020.

Mishra, A., Alahari, K., 与 Jawahar, C. 使用高阶语言先验的场景文本识别。2012。

Mithun, N. C., Panda, R., Papalexakis, E. E., 与 Roy-Chowdhury, A. K. 基于网络监督的跨模态图像-文本检索联合嵌入方法。发表于 *Proceedings of the 26th ACM international conference on Multimedia*，第1856–1864页，2018年。

森, 康之, 高桥, 宏和, 冈, 良一. 基于图像分割与词汇向量量化的图像到文字转换. Citeseer, 1999.

Mu, J., Liang, P., and Goodman, N. 利用语言塑造视觉表征进行少样本分类。*arXiv preprint arXiv:1911.02683*, 2019.

Muller-Budack, E., Pustu-Iren, K., and Ewerth, R. Geolocation estimation of photos using a hierarchical model and scene classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 563–579, 2018.

Murty, S., Koh, P. W., and Liang, P. Expbert: Representation engineering with natural language explanations. *arXiv preprint arXiv:2005.01932*, 2020.

Narasimhan, K., Kulkarni, T., and Barzilay, R. Language understanding for text-based games using deep reinforcement learning. *arXiv preprint arXiv:1506.08941*, 2015.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Noble, S. U. Algorithms of oppression: How search engines reinforce racism. 2018.

Nosek, B. A., Banaji, M. R., and Greenwald, A. G. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, 6(1):101, 2002.

Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.-C., Lee, J. T., Mukherjee, S., Aggarwal, J., Lee, H., Davis, L., et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, pp. 3153–3160. IEEE, 2011.

Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., and Goodfellow, I. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31:3235–3246, 2018.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ordonez, V., Kulkarni, G., and Berg, T. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24:1143–1151, 2011.

pandas development team, T. pandas-dev/pandas: Pandas, February 2020. URL https://doi.org/10.5281/zenodo.3509134.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. V. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

Qi, D., Su, L., Song, J., Cui, E., Bharti, T., and Sacheti, A. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020.

Quattoni, A., Collins, M., and Darrell, T. Learning visual representations using images with captions. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. IEEE, 2007.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training, 2018.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., and Denton, E. Saving face: Investigating the ethical concerns of facial recognition auditing, 2020.

Ramanathan, V., Liang, P., and Fei-Fei, L. Video event understanding using natural language descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 905–912, 2013.

Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 139–147, 2010.

Muller-Budack, E., Pustu-Iren, K., 和 Ewerth, R. 使用层次模型和场景分类进行照片的地理位置估计。于 *Proceedings of the European Conference on Computer Vision (ECCV)*，第 563–579 页，2018 年。

Murty, S., Koh, P. W., and Liang, P. Expbert：使用自然语言解释的表征工程。*arXiv preprint arXiv:2005.01932*，2020。

Narasimhan, K., Kulkarni, T., 与 Barzilay, R. 使用深度强化学习进行基于文本游戏的语言理解。*arXiv preprint arXiv:1506.08941*, 2015。

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. 使用无监督特征学习读取自然图像中的数字。2011。

诺布尔，S. U.《压迫的算法：搜索引擎如何强化种族主义》。2018年。

Nosek, B. A., Banaji, M. R., 和 Greenwald, A. G. 从一个演示网站中获取内隐的群体态度和信念。*Group Dynamics: Theory, Research, and Practice*, 6(1):101, 2002.

哦，S.，胡格斯，A.，佩雷拉，A.，昆图尔，N.，陈，C.-C.，李，J. T.，穆克吉，S.，阿加瓦尔，J.，李，H.，戴维斯，L.，等。用于监控视频事件识别的大规模基准数据集。载于*CVPR 2011*，第3153–3160页。IEEE，2011年。

奥利弗，A.，奥德纳，A.，拉弗尔，C. A.，丘布克，E. D.，与古德费洛，I. 深度半监督学习算法的现实评估。*Advances in neural information pro- cessing systems*，31卷：3235–3246页，2018年。

Oord, A. v. d., Li, Y., and Vinyals, O. 基于对比预测编码的表征学习。*arXiv preprint arXiv:1807.03748*, 2018.

Ordonez, V., Kulkarni, G., 和 Berg, T. Im2text: 使用一百万张带标题的照片描述图像。*Advances in neural information processing systems*, 24:1143–1151, 2011.

pandas开发团队，T. pandas-dev/pandas：Pan-das，2020年2月。URL https://doi.org/10.5281/zenodo.3509134。

Parkhi, O. M., Vedaldi, A., Zisserman, A., 和 Jawahar, C. V. 猫与狗。收录于 *IEEE Conference on Computer Vision and Pattern Recognition*, 2012。

帕斯克, A., 格罗斯, S., 马萨, F., 勒雷, A., 布拉德伯里, J., 钱南, G., 基林, T., 林, Z., 吉梅尔辛, N., 安蒂加, L., 德斯梅森, A., 科普夫, A., 杨, E., 德维托, Z., 雷森, M., 特贾尼, A., 奇拉姆库尔蒂, S., 斯坦纳, B., 方, L., Bai, J., 与 Chintala, S. Pytorch：一种命令式风格的高性能深度学习库。收录于 *Advances in Neural Information Processing Systems 32*，第 8024–8035 页，2019。

佩德雷戈萨，F.，瓦罗夸，G.，格拉姆福特，A.，米歇尔，V.，蒂里翁，B.，格里塞尔，O.，布隆德尔，M.，普雷滕霍费尔，P.，韦斯，R.，杜堡，V.，范德普拉斯，J.，帕索斯，A.，库尔纳波，D.，布鲁彻，M.，佩罗，M.，和杜谢奈，E. Scikit-learn：Python中的机器学习。*Journal of Machine Learning Research*，12卷：2825–2830页，2011年。

Pennington, J., Socher, R., and Manning, C. D. Glove: 用于词表示的全局向量。于 *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*，第1532–1543页，2014年。

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. 深度上下文词表示。*arXiv preprint arXiv:1802.05365*, 2018。

齐、D.，苏、L.，宋、J.，崔、E.，巴蒂、T.，和萨切蒂、A.。Imagebert：基于大规模弱监督图像-文本数据的跨模态预训练。*arXiv preprint arXiv:2001.07966*，2020年。

Quattoni, A., Collins, M., and Darrell, T. 使用带标题的图像学习视觉表示。于 *2007 IEEE Conference on Computer Vision and Pattern Recognition*，第 1–8 页。IEEE，2007。

Radford, A., Narasimhan, K., Salimans, T., 与 Sutskever, I. 通过生成式预训练提升语言理解能力，2018。

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. 语言模型是无监督多任务学习者。2019。

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. 探索统一文本到文本转换器的迁移学习极限。*arXiv preprint arXiv:1910.10683*，2019。

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., and Denton, E. 拯救面子：调查面部识别审计中的伦理问题，2020。

Ramanathan, V., Liang, P., and Fei-Fei, L. 使用自然语言描述的视频事件理解。In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 905–912, 2013.

Rashtchian, C., Young, P., Hodosh, M. 和 Hockenmaier, J. 使用亚马逊的Mechanical Turk收集图像标注。载于 *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*，第139–147页，2010年。

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? *arXiv preprint arXiv:1902.10811*, 2019.

Salimans, T. and Kingma, D. P. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in neural information processing systems*, pp. 901–909, 2016.

Scheuerman, M. K., Paul, J. M., and Brubaker, J. R. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–33, 2019.

Schwemmer, C., Knight, C., Bello-Pardo, E. D., Oklobdzija, S., Schoonvelde, M., and Lockhart, J. W. Diagnosing gender bias in image recognition systems. *Socius*, 6: 2378023120967171, 2020.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

Shankar, V., Dave, A., Roelofs, R., Ramanan, D., Recht, B., and Schmidt, L. Do image classifiers generalize across time? *arXiv preprint arXiv:1906.02168*, 2019.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and Rohrbach, M. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019.

Socher, R. and Fei-Fei, L. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 966–973. IEEE, 2010.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., and Ng, A. Y. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.

Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in neural information processing systems*, pp. 1857–1865, 2016.

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., and Wang, J. Release strategies and the social impacts of language models, 2019.

Soomro, K., Zamir, A. R., and Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Speer, R. ftfy. Zenodo, 2019. URL https://doi.org/10.5281/zenodo.2591652. Version 5.5.

Srivastava, N. and Salakhutdinov, R. Multimodal learning with deep boltzmann machines. In *NIPS*, 2012.

Srivastava, S., Labutov, I., and Mitchell, T. Joint concept learning and semantic parsing from natural language explanations. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 1527–1536, 2017.

Stallkamp, J., Schlipsing, M., Salmen, J., and Igel, C. The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In *IEEE International Joint Conference on Neural Networks*, pp. 1453–1460, 2011.

Stroud, J. C., Ross, D. A., Sun, C., Deng, J., Sukthankar, R., and Schmid, C. Learning video representations from textual web supervision. *arXiv preprint arXiv:2007.14937*, 2020.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*, 2016.

Tan, H. and Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

Tan, M. and Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. *arXiv preprint arXiv:2007.00644*, 2020.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Image Net分类器能否泛化至ImageNet？ *arXiv preprint arXiv:1902.10811*, 2019.

Salimans, T. 和 Kingma, D. P. 权重归一化：一种加速深度神经网络训练的简单重参数化方法。载于 *Advances in neural information pro- cessing systems*，第901–909页，2016年。

Scheuerman, M. K., Paul, J. M., 与 Brubaker, J. R. 计算机如何识别性别：商业面部分析服务中的性别分类评估。*Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–33, 2019.

Schwemmer, C., Knight, C., Bello-Pardo, E. D., Oklobdzija, S., Schoonvelde, M., 与 Lockhart, J. W. 诊断图像识别系统中的性别偏见。*Socius*, 6: 2378023120967171, 2020。

Sennrich, R., Haddow, B., 和 Birch, A. 使用子词单元进行稀有词的神经机器翻译。*arXiv preprint arXiv:1508.07909*, 2015.

Shankar, V., Dave, A., Roelofs, R., Ramanan, D., Recht, B., and Schmidt, L. 图像分类器是否具有跨时间泛化能力？ *arXiv preprint arXiv:1906.02168*, 2019.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. 概念性标注：一个用于自动图像描述、经过清洗和上位词化的图像替代文本数据集。载于 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*，第2556–2565页，2018年。

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and Rohrbach, M. 迈向能够阅读的VQA模型。于 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 页 8317–8326, 2019。

Socher, R. 与 Fei-Fei, L. 连接多模态：使用未对齐文本语料库进行半监督图像分割与标注。收录于 *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*，第 966–973 页。IEEE，2010。

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. 基于情感树库语义组合性的递归深度模型。发表于 *Proceedings of the 2013 conference on empirical methods in natural language processing*，第1631–1642页，2013年。

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., 与 Ng, A. Y. 基于组合语义的图像句子检索与描述。*Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.

Sohn, K. 改进的深度度量学习与多类别n对损失目标。于 *Advances in neural information processing systems*，第1857–1865页，2016年。

索莱曼，I.，布伦戴奇，M.，克拉克，J.，阿斯克尔，A.，赫伯特-沃斯，A.，吴，J.，拉德福德，A.，克鲁格，G.，金，J. W.，克雷普斯，S.，麦凯恩，M.，纽豪斯，A.，布拉扎基斯，J.，麦格菲，K.，和王，J.。语言模型的发布策略与社会影响，2019。

Soomro, K., Zamir, A. R., 和 Shah, M. Ucf101：一个包含 101 个野外视频人类动作类别的数据集。*arXiv preprint arXiv:1212.0402*, 2012。

Speer, R. ftfy. Zenodo, 2019. 网址 https://doi.org/10.5281/zenodo.2591652. 版本 5.5。

Srivastava, N. 和 Salakhutdinov, R. 使用深度玻尔兹曼机进行多模态学习。收录于 *NIPS*，2012年。

Srivastava, S., Labutov, I., and Mitchell, T. 从自然语言解释中联合学习概念与语义解析。载于 *Proceedings of the 2017 conference on empirical methods in natural language processing*, 第1527–1536页, 2017年。

Stallkamp, J., Schlipsing, M., Salmen, J., 与 Igel, C. 德国交通标志识别基准：一个多类别分类竞赛。载于 *IEEE International Joint Conference on Neural Networks*，第1453–1460页，2011年。

Stroud, J. C., Ross, D. A., Sun, C., Deng, J., Sukthankar, R., and Schmid, C. 从文本网络监督中学习视频表示。*arXiv preprint arXiv:2007.14937*, 2020。

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. Inception-v4、Inception-ResNet与残差连接对学习的影响。*arXiv preprint arXiv:1602.07261*，2016。

Tan, H. 和 Bansal, M. Lxmert：从Transformers学习跨模态编码器表示。*arXiv preprint arXiv:1908.07490*，2019。

Tan, M. 与 Le, Q. V. EfficientNet：重新思考卷积神经网络的模型缩放。*arXiv preprint arXiv:1905.11946*，2019。

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. 测量图像分类中对自然分布偏移的鲁棒性。*arXiv preprint arXiv:2007.00644*, 2020.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. Yfcc100m：多媒体研究中的新数据。*Communications of the ACM*, 59(2): 64–73, 2016.

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

Tian, Y., Wang, Y., Krishnan, D., Tenenbaum, J. B., and Isola, P. Rethinking few-shot image classification: a good embedding is all you need? *arXiv preprint arXiv:2003.11539*, 2020.

Torralba, A., Fergus, R., and Freeman, W. T. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

Touvron, H., Vedaldi, A., Douze, M., and Jégou, H. Fixing the train-test resolution discrepancy. In *Advances in neural information processing systems*, pp. 8252–8262, 2019.

Varadarajan, J. and Odobez, J.-M. Topic models for scene analysis and abnormality detection. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pp. 1338–1345. IEEE, 2009.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., and Welling, M. Rotation equivariant CNNs for digital pathology. June 2018.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Vo, N., Jacobs, N., and Hays, J. Revisiting im2gps in the deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2621–2630, 2017.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pp. 10506–10518, 2019.

Wang, H., Lu, P., Zhang, H., Yang, M., Bai, X., Xu, Y., He, M., Wang, Y., and Liu, W. All you need is boundary: Toward arbitrary-shaped text spotting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12160–12167, 2020.

Wang, J., Markert, K., and Everingham, M. Learning models for object recognition from natural language descriptions. In *BMVC*, volume 1, pp. 2, 2009.

Weston, J., Bengio, S., and Usunier, N. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81(1):21–35, 2010.

Weston, J. E. Dialog-based language learning. In *Advances in Neural Information Processing Systems*, pp. 829–837, 2016.

Weyand, T., Kostrikov, I., and Philbin, J. Planet-photo geolocation with convolutional neural networks. In *European Conference on Computer Vision*, pp. 37–55. Springer, 2016.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

Wu, Z., Xiong, Y., Yu, S., and Lin, D. Unsupervised feature learning via non-parametric instance-level discrimination. *arXiv preprint arXiv:1805.01978*, 2018.

Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10687–10698, 2020.

y Arcas, B. A., Mitchell, M., and Todorov, A. Physiognomy's new clothes. 2017. URL https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a.

Yang, Z., Lu, Y., Wang, J., Yin, X., Florencio, D., Wang, L., Zhang, C., Zhang, L., and Luo, J. Tap: Text-aware pre-training for text-vqa and text-caption. *arXiv preprint arXiv:2012.04638*, 2020.

Yogatama, D., d'Autume, C. d. M., Connor, J., Kocisky, T., Chrzanowski, M., Kong, L., Lazaridou, A., Ling, W., Yu, L., Dyer, C., et al. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*, 2019.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Tian, Y., Krishnan, D., and Isola, P. 对比多视图编码。*arXiv preprint arXiv:1906.05849*, 2019.

Tian, Y., Wang, Y., Krishnan, D., Tenenbaum, J. B., and I sola, P. 重新思考小样本图像分类：一个好的嵌入向量就是全部所需吗？*arXiv preprint arXiv:2003.11539*, 20 20.

Torralba, A., Fergus, R., and Freeman, W. T. 八千万微小图像：用于非参数化物体与场景识别的大规模数据集。*IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

Touvron, H., Vedaldi, A., Douze, M., and Jégou, H. 修复训练与测试分辨率不一致的问题。于 *Advances in neural information processing systems*，第 8252–8262 页，2019。

Varadarajan, J. 与 Odobez, J.-M. 面向场景分析与异常检测的主题模型。收录于 *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*，第1338–1345页。IEEE，2009年。

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, ., and Polosukhin, I. Attention is all you need. 发表于 *Advances in neural information processing systems*, 第 5998–6008 页, 2017.

Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., and W elling, M. 用于数字病理学的旋转等变卷积神经网络。2 018年6月。

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., We ckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilso n, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., M oore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimr man, R., Henriksen, I., Quintero, E. A., Harris, C. R., Arch ibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., 以及 SciPy 1.0 贡献者。SciPy 1.0：Python 科学计算基础算法。*Nature Methods*, 17:261–272, 2020. doi: 10.1 038/s41592-019-0686-2.

Vo, N., Jacobs, N., 和 Hays, J. 在深度学习时代重新审视 im2gps。收录于 *Proceedings of the IEEE International Conference on Computer Vision*，第 2621–2630 页，2017 年。

王，A.，辛格，A.，迈克尔，J.，希尔，F.，利维，O. ，和鲍曼，S. R. GLUE：一个用于自然语言理解的多任务基准与分析平台。*arXiv preprint arXiv:1804.07461* ，2018年。

王, H., Ge, S., Lipton, Z., 和 Xing, E. P. 通过惩罚局部预测能力学习鲁棒的全局表示。于 *Advances in Neural Information Processing Systems*, 页 1 0506–10518, 2019。

王, H., 卢, P., 张, H., 杨, M., 白, X., 徐, Y., 何, M., 王, Y., 刘, W. 你所需要的只是边界：面向任意形状的文本定位。于 *Proceedings of the AAAI Conference on Artificial Intelligence*，第34卷，第 12160–12167页，2020年。

王，J.，Markert, K. 和 Everingham, M. 从自然语言描述中学习物体识别模型。于 *BMVC*，第1卷，第2页，200 9年。

韦斯顿，J.，本吉奥，S.，和乌苏尼尔，N. 大规模图像标注：利用联合词-图像嵌入学习排序。*Machine learning*，81(1):21–35，2010。

韦斯顿, J. E. 基于对话的语言学习. 于 *Advances in Neural Information Processing Systems*, 页 829–837, 20 16.

Weyand, T., Kostrikov, I., 和 Philbin, J. 使用卷积神经网络进行行星照片地理定位。载于 *European Conference on Computer Vision*，第37–55页。Springer ，2016。

吴，Y.，基里洛夫，A.，马萨，F.，罗，W.-Y.，和吉尔希克，R. Detectron2。https://github.com/ facebookrese arch/detectron2，2019。

吴，Z.，熊，Y.，余，S.，和林，D. 通过非参数实例级判别进行无监督特征学习。*arXiv preprint arXiv:1805.01978*，2018。

谢, Q., Luong, M.-T., Hovy, E., 和 Le, Q. V. 使用带噪声学生的自训练改进 ImageNet 分类。于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 页 10687–10698, 2020。

y Arcas, B. A., Mitchell, M., 和 Todorov, A. 相面术的新装。2017. URL https://medium.com/@blaisea/ physiogno mys-new-clothes-f2d4b59fdd6a。

杨, Z., 卢, Y., 王, J., 尹, X., Florencio, D., 王, L., 张, C., 张, L., 和 罗, J. Tap: 面向文本VQA与文本描述的文本感知预训练。*arXiv preprint arXiv:2012.04638*, 2020.

Yogatama, D., d’ Autume, C. d. M., Connor, J., Kocisky, T., Chrzanowski, M., Kong, L., Lazaridou, A., Ling, W., Y u, L., Dyer, C., 等人。学习与评估通用语言智能。*arXiv preprint arXiv:1901.11373*，2019。

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. 从图像描述到视觉指称：事件描述语义推理的新相似性度量。*Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Yu, F., Tang, J., Yin, W., Sun, Y., Tian, H., Wu, H., and Wang, H. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 2020.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A. S., Neumann, M., Dosovitskiy, A., et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

Zhang, R. Making convolutional networks shift-invariant again. *arXiv preprint arXiv:1904.11486*, 2019.

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

Zuboff, S. Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, 30(1):75–89, 2015.

Yu, F., Tang, J., Yin, W., Sun, Y., Tian, H., Wu, H., and Wang, H. Ernie-vil: 通过场景图实现知识增强的视觉-语言表示。*arXiv preprint arXiv:2006.16934*, 2020.

Zeiler, M. D. 和 Fergus, R. 可视化和理解卷积网络。于 *European conference on computer vision*，第 818–833 页。Springer，2014。

翟旭、普伊格塞尔韦、科列斯尼科夫、鲁伊森、里克尔梅、卢西奇、乔隆加、平托、纽曼、多索维茨基等人在2019年进行了一项大规模研究，利用视觉任务适应基准探讨表征学习。*arXiv preprint arXiv:1910.04867*

张, R. 让卷积网络再次具有平移不变性。*arXiv preprint arXiv:1904.11486*, 2019.

张，Y.，江，H.，三浦，Y.，曼宁，C. D.，和兰格洛茨，C. P. 从配对图像和文本中对比学习医学视觉表示。*arXiv preprint arXiv:2010.00747*，2020。

祖博夫，S. 大他者：监控资本主义与信息文明的前景。*Journal of Infor- mation Technology*, 30(1):75–89, 2015.

A. Linear-probe evaluation

We provide additional details for linear probe experiments presented in this paper, including the list of the datasets and models used for evaluation.

A.1. Datasets

We use the 12 datasets from the well-studied evaluation suite introduced by (Kornblith et al., 2019) and add 15 additional datasets in order to assess the performance of models on a wider variety of distributions and tasks. These datasets include MNIST, the Facial Expression Recognition 2013 dataset (Goodfellow et al., 2015), STL-10 (Coates et al., 2011), EuroSAT (Helber et al., 2019), the NWPU-RESISC45 dataset (Cheng et al., 2017), the German Traffic Sign Recognition Benchmark (GTSRB) dataset (Stallkamp et al., 2011), the KITTI dataset (Geiger et al., 2012), PatchCamelyon (Veeling et al., 2018), the UCF101 action recognition dataset (Soomro et al., 2012), Kinetics 700 (Carreira et al., 2019), 2,500 random samples of the CLEVR dataset (Johnson et al., 2017), the Hateful Memes dataset (Kiela et al., 2020), and the ImageNet-1k dataset (Deng et al., 2012). For the two video datasets (UCF101 and Kinetics700), we use the middle frame of each video clip as the input image. STL-10 and UCF101 have multiple predefined train/validation/test splits, 10 and 3 respectively, and we report the average over all splits. Details on each dataset and the corresponding evaluation metrics are provided in Table 9.

Additionally, we created two datasets that we call Country211 and Rendered SST2. The Country211 dataset is designed to assess the geolocation capability of visual representations. We filtered the YFCC100m dataset (Thomee et al., 2016) to find 211 countries (defined as having an ISO-3166 country code) that have at least 300 photos with GPS coordinates, and we built a balanced dataset with 211 categories, by sampling 200 photos for training and 100 photos for testing, for each country.

The Rendered SST2 dataset is designed to measure the optical character recognition capability of visual representations. To do so, we used the sentences from the Stanford Sentiment Treebank dataset (Socher et al., 2013) and rendered them into images, with black texts on a white background, in a 448×448 resolution. Two example images from this dataset are shown in Figure 19.

A.2. Models

In combination with the datasets listed above, we evaluate the following series of models using linear probes.

LM RN50 This is a multimodal model that uses an autoregressive loss instead of a contrastive loss, while using the ResNet-50 architecture as in the smallest contrastive model. To do so, the output from the CNN is projected into four tokens, which are then fed as a prefix to a language model autoregressively predicting the text tokens. Apart from the training objective, the model was trained on the same dataset for the same number of epochs as other CLIP models.

CLIP-RN Five ResNet-based contrastive CLIP models are included. As discussed in the paper, the first two models follow ResNet-50 and ResNet-101, and we use EfficientNet-style (Tan & Le, 2019) scaling for the next three models which simultaneously scale the model width, the number of layers, and the input resolution to obtain models with roughly 4x, 16x, and 64x computation.

CLIP-ViT We include four CLIP models that use the Vision Transformer (Dosovitskiy et al., 2020) architecture as the image encoder. We include three models trained on 224-by-224 pixel images: ViT-B/32, ViT-B/16, ViT-L/14, and the ViT-L/14 model fine-tuned on 336-by-336 pixel input images.

EfficietNet We use the nine models (B0-B8) from the original EfficientNet paper (Tan & Le, 2019), as well as the noisy-student variants (B0-B7, L2-475, and L2-800) (Tan & Le, 2019). The largest models (L2-475 and L2-800) take the input resolutions of 475x475 and 800x800 pixels, respectively.

Instagram-pretrained ResNeXt We use the four models (32x8d, 32x16d, 32x32d, 32x48d) released by (Mahajan et al., 2018), as well as their two FixRes variants which use higher input resolutions (Touvron et al., 2019).

Big Transfer (BiT) We use BiT-S and BiT-M models (Kolesnikov et al., 2019), trained on the ImageNet-1k and ImageNet-21k datasets. The model weights for BiT-L is not publicly available.

Vision Transformer (ViT) We also include four ViT (Dosovitskiy et al., 2020) checkpoints pretrained on the ImageNet-21k dataset, namely ViT-B/32, ViT-B/16, ViT-L/16, and ViT-H/14. We note that their best-performing models, trained on the JFT-300M dataset, are not available publicly.

SimCLRv2 The SimCLRv2 (Chen et al., 2020c) project released pre-trained and fine-tuned models in various settings. We use the seven pretrain-only checkpoints with selective kernels.

BYOL We use the recently released model weights of BYOL (Grill et al., 2020), specifically their 50x1 and 200x2

A. 线性探针评估

我们提供了本文中线性探测实验的额外细节，包括用于评估的数据集和模型列表。

A.1. 数据集

我们采用了(Kornblith等人，2019)提出的经过充分研究的评估套件中的12个数据集，并额外添加了15个数据集，以评估模型在更广泛分布和任务上的性能。这些数据集包括MNIST、面部表情识别2013数据集(Goodfellow等人，2015)、STL-10(Coates等人，2011)、EuroSAT(Helber等人，2019)、NWPU-RESISC45数据集(Cheng等人，2017)、德国交通标志识别基准(GTSRB)数据集(Stallkamp等人，2011)、KITTI数据集(Geiger等人，2012)、PatchCamelyon(Veeling等人，2018)、UCF101动作识别数据集(Soomro等人，2012)、Kinetics 700(Carreira等人，2019)、CLEVR数据集(Johnson等人，2017)的2500个随机样本、恶意表情包数据集(Kiela等人，2020)以及ImageNet-1k数据集(Deng等人，2012)。对于两个视频数据集(UCF101和Kinetics700)，我们使用每个视频片段的中间帧作为输入图像。STL-10和UCF101分别具有10个和3个预定义的训练/验证/测试划分，我们报告所有划分的平均结果。各数据集详情及对应评估指标见表9。

此外，我们创建了两个数据集，分别命名为Country211和Rendered SST2。Country211数据集旨在评估视觉表征的地理定位能力。我们对YFCC100m数据集（Thomee等人，2016）进行筛选，找出拥有至少300张带GPS坐标照片的211个国家（定义为具有ISO-3166国家代码），并通过为每个国家抽取200张训练照片和100张测试照片，构建了一个包含211个类别的平衡数据集。

渲染后的SST2数据集旨在衡量视觉表征的光学字符识别能力。为此，我们采用斯坦福情感树库数据集（Socher等人，2013）中的句子，将其渲染为448{v*}448分辨率、白底黑字的图像。图19展示了该数据集的两个示例图像。

A.2. 模型

结合上述列出的数据集，我们使用线性探针评估以下系列模型。

LM RN50 这是一个多模态模型，它采用自回归损失而非对比损失，同时使用采用与最小对比模型相同的ResNet-50架构。具体而言，将CNN的输出投影为四个标记，随后将其作为前缀输入语言模型，以自回归方式预测文本标记。除训练目标外，该模型使用的训练数据集和训练周期数均与其他CLIP模型保持一致。

CLIP-RN 包含五个基于ResNet的对比CLIP模型。如论文所述，前两个模型遵循ResNet-50和ResNet-101架构，而后三个模型采用EfficientNet风格（Tan & Le, 2019）的缩放方法，同步扩展模型宽度、层数和输入分辨率，从而获得计算量分别约为4倍、16倍和64倍的模型。

CLIP-ViT 我们纳入了四个采用视觉变换器（Dosovitskiy等人，2020）架构作为图像编码器的CLIP模型。其中包括三个基于224×224像素图像训练的模型：ViT-B/32、ViT-B/16、ViT-L/14，以及一个在336×336像素输入图像上微调的ViT-L/14模型。

EfficientNet 我们采用了原始 EfficientNet 论文（Tan & Le，2019）中的九个模型（B0-B8），以及噪声学生变体（B0-B7、L2-475 和 L2-800）（Tan & Le，2019）。其中最大的模型（L2-475 和 L2-800）分别采用 475x475 和 800x800 像素的输入分辨率。

Instagram预训练的ResNeXt 我们采用了(Mahajan等人，2018年)发布的四种模型（32x8d、32x16d、32x32d、32x48d），以及它们两个使用更高输入分辨率的FixRes变体（Touvron等人，2019年）。

大迁移（BiT）我们使用了在ImageNet-1k和ImageNet-21k数据集上训练的BiT-S和BiT-M模型（Kolesnikov等人，2019年）。BiT-L的模型权重尚未公开。

视觉变换器（ViT）我们还包含了四个在ImageNet-21k数据集上预训练的ViT（Dosovitskiy等人，2020）检查点，即ViT-B/32、ViT-B/16、ViT-L/16和ViT-H/14。我们注意到，它们在JFT-300M数据集上训练的最佳性能模型并未公开提供。

SimCLRv2项目（Chen等人，2020c）发布了多种设置下的预训练和微调模型。我们使用了七个仅预训练的检查点，并采用了选择性核技术。

BYOL 我们使用了最近发布的BYOL模型权重（Grill等人，2020年），具体采用了其50x1和200x2版本。

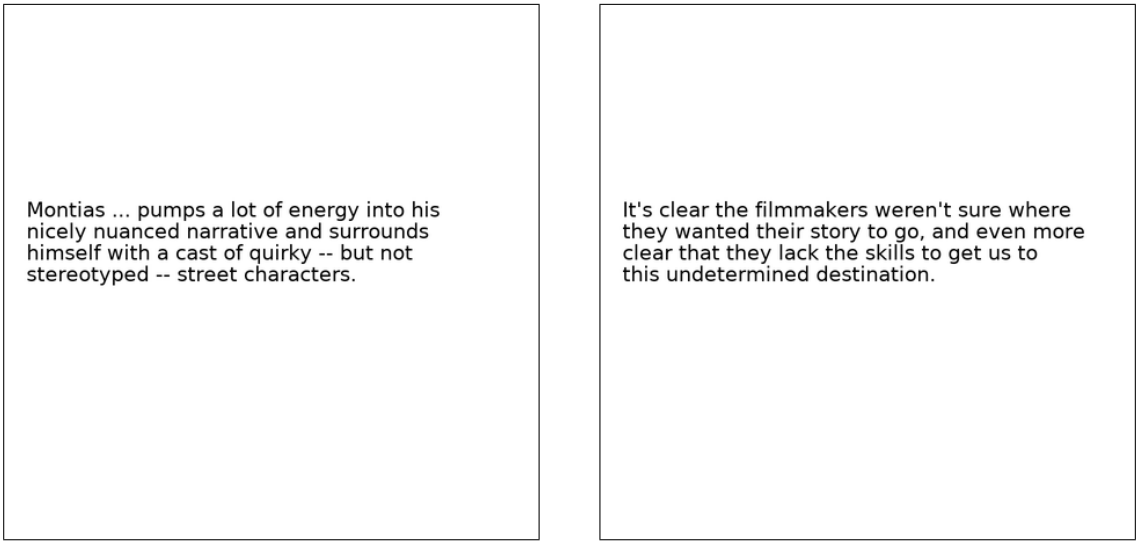

Figure 19. Two example images from the Rendered SST2 dataset

checkpoints.

Momentum Contrast (MoCo) We include the MoCo-v1 (He et al., 2020) and the MoCo-v2 (Chen et al., 2020d) checkpoints.

VirTex We use the pretrained model of VirTex (Desai & Johnson, 2020). We note that VirTex has a similar model design to CLIP-AR but is trained on a 1000x smaller dataset of high-quality captions from MSCOCO.

ResNet We add the original ResNet checkpoints released by (He et al., 2016b), namely ResNet-50, ResNet-101, and ResNet152.

A.3. Evaluation

We use image features taken from the penultimate layer of each model, ignoring any classification layer provided. For CLIP-ViT models, we used the features before the linear projection to the embedding space, which corresponds to `I_f` in Figure 3. We train a logistic regression classifier using scikit-learn's L-BFGS implementation, with maximum 1,000 iterations, and report the corresponding metric for each dataset. We determine the L2 regularization strength λ using a hyperparameter sweep on the validation sets over the range between 10^{-6} and 10^{6}, with 96 logarithmically spaced steps. To save compute required for the sweeps, we perform a parametric binary search that starts with $\lambda = [10^{-6}, 10^{-4}, 10^{-2}, 1, 10^{2}, 10^{4}, 10^{6}]$ and iteratively halves the interval around the peak until it reaches a resolution of 8 steps per decade. The hyperparameter sweeps are performed on a validation split of each dataset. For the datasets that contain a validation split in addition to a test split, we use the provided validation set to perform the hyperparameter search, and for the datasets that do not provide a validation split or have not published labels for the test data, we split the training dataset to perform the hyperparameter search. For the final result, we combine the validation split back with the training split and report the performance on the unused split.

A.4. Results

The individual linear probe scores are provided in Table 10 and plotted in Figure 20. The best-performing CLIP model, using ViT-L/14 archiecture and 336-by-336 pixel images, achieved the state of the art in 21 of the 27 datasets, i.e. included in the Clopper-Pearson 99.5% confidence interval around each dataset's top score. For many datasets, CLIP performs significantly better than other models, demonstrating the advantage of natural language supervision over traditional pre-training approaches based on image classification. See Section 3.2 for more discussions on the linear probe results.

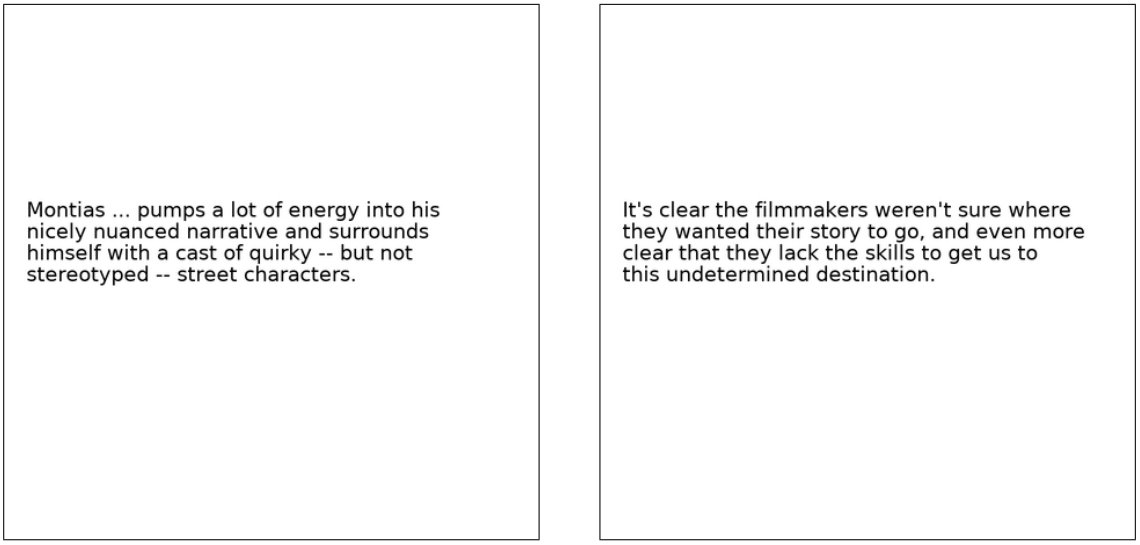

图19. Rendered SST2数据集中的两个示例图像

检查点。

动量对比（MoCo）我们包含了MoCo-v1（He等人，2020年）和MoCo-v2（Chen等人，2020年d）的检查点。

VirTex 我们采用VirTex（Desai & Johnson, 2020）的预训练模型。我们注意到VirTex的模型设计与CLIP-AR相似，但其训练数据集规模小1000倍，使用的是来自MSCOCO的高质量标注数据。

ResNet 我们添加了由（He等人，2016b）发布的原始ResNet检查点，即ResNet-50、ResNet-101和ResNet152。

A.3. 评估

我们采用从每个模型倒数第二层提取的图像特征，忽略任何提供的分类层。对于CLIP-ViT模型，我们使用线性投影到嵌入空间之前的特征，即图3中的I_f。我们使用scikit-learn的L-BFGS实现训练逻辑回归分类器，最大迭代次数为1000次，并报告每个数据集的相应指标。我们通过在验证集上进行超参数扫描来确定L2正则化强度λ，扫描范围从10^$^{-6}$到10^^{6}，按对数间隔分为96步。为减少扫描计算量，我们采用参数化二分搜索：初始搜索点为$\lambda = [10^{-6}, 10^{-4}, 10^{-2}, 1^{}, 10^{2}, 10^{4}, 10^{6}]$，随后围绕峰值逐次将区间减半，直至达到每十倍距8步的分辨率。超参数扫描在各数据集的验证集分割上进行。对于除训练集外还包含验证分割的数据集，在测试分割中，我们使用提供的验证集进行超参数搜索；对于未提供验证分割或未发布测试数据标签的数据集，我们则分割训练数据集以执行超参数搜索。最终，我们将验证分割重新合并至训练分割中，并在未使用的分割上报告性能表现。

A.4. 结果

各个线性探测得分详见表10，并绘制于图20中。表现最佳的CLIP模型采用ViT-L/14架构和336×336像素图像，在27个数据集中的21个达到了当前最优水平，即其成绩均落入各数据集最高分数的Clopper-Pearson 99.5%置信区间内。在多数数据集上，CLIP模型显著优于其他模型，这证明了基于自然语言监督的方法相较于传统图像分类预训练方式的优势。关于线性探测结果的更多讨论，请参见第3.2节。

Dataset	Classes	Train size	Test size	Evaluation metric
Food-101	102	75,750	25,250	accuracy
CIFAR-10	10	50,000	10,000	accuracy
CIFAR-100	100	50,000	10,000	accuracy
Birdsnap	500	42,283	2,149	accuracy
SUN397	397	19,850	19,850	accuracy
Stanford Cars	196	8,144	8,041	accuracy
FGVC Aircraft	100	6,667	3,333	mean per class
Pascal VOC 2007 Classification	20	5,011	4,952	11-point mAP
Describable Textures	47	3,760	1,880	accuracy
Oxford-IIIT Pets	37	3,680	3,669	mean per class
Caltech-101	102	3,060	6,085	mean-per-class
Oxford Flowers 102	102	2,040	6,149	mean per class
MNIST	10	60,000	10,000	accuracy
Facial Emotion Recognition 2013	8	32,140	3,574	accuracy
STL-10	10	1000	8000	accuracy
EuroSAT	10	10,000	5,000	accuracy
RESISC45	45	3,150	25,200	accuracy
GTSRB	43	26,640	12,630	accuracy
KITTI	4	6,770	711	accuracy
Country211	211	43,200	21,100	accuracy
PatchCamelyon	2	294,912	32,768	accuracy
UCF101	101	9,537	1,794	accuracy
Kinetics700	700	494,801	31,669	mean(top1, top5)
CLEVR Counts	8	2,000	500	accuracy
Hateful Memes	2	8,500	500	ROC AUC
Rendered SST2	2	7,792	1,821	accuracy
ImageNet	1000	1,281,167	50,000	accuracy

Table 9. Datasets examined for linear probes. We note that, for the Birdsnap and Kinetics700 datasets, we used the resources that are available online at the time of this writing.

Dataset	Classes	Train size	Test size	Evaluation metric
Food-101	102	75,750	25,250	accuracy
CIFAR-10	10	50,000	10,000	accuracy
CIFAR-100	100	50,000	10,000	accuracy
Birdsnap	500	42,283	2,149	accuracy
SUN397	397	19,850	19,850	accuracy
Stanford Cars	196	8,144	8,041	accuracy
FGVC Aircraft	100	6,667	3,333	mean per class
Pascal VOC 2007 Classification	20	5,011	4,952	11-point mAP
Describable Textures	47	3,760	1,880	accuracy
Oxford-IIIT Pets	37	3,680	3,669	mean per class
Caltech-101	102	3,060	6,085	mean-per-class
Oxford Flowers 102	102	2,040	6,149	mean per class
MNIST	10	60,000	10,000	accuracy
Facial Emotion Recognition 2013	8	32,140	3,574	accuracy
STL-10	10	1000	8000	accuracy
EuroSAT	10	10,000	5,000	accuracy
RESISC45	45	3,150	25,200	accuracy
GTSRB	43	26,640	12,630	accuracy
KITTI	4	6,770	711	accuracy
Country211	211	43,200	21,100	accuracy
PatchCamelyon	2	294,912	32,768	accuracy
UCF101	101	9,537	1,794	accuracy
Kinetics700	700	494,801	31,669	mean(top1, top5)
CLEVR Counts	8	2,000	500	accuracy
Hateful Memes	2	8,500	500	ROC AUC
Rendered SST2	2	7,792	1,821	accuracy
ImageNet	1000	1,281,167	50,000	accuracy

表9. 线性探针研究的数据集。我们注意到，对于Birdsnap和Kinetics700数据集，我们使用了撰写本文时在线可用的资源。

		Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech101	Flowers	MNIST	FER2013	STL10*	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	Kinetics700	CLEVR	HatefulMemes	SST	ImageNet
LM RN50		81.3	82.8	61.7	44.2	69.6	74.9	44.9	85.5	71.5	82.8	85.5	91.1	96.6	60.1	95.3	93.4	84.0	73.8	70.2	19.0	82.9	76.4	51.9	51.2	65.2	76.8	65.2
CLIP-RN	50	86.4	88.7	70.3	56.4	73.3	78.3	49.1	87.1	76.4	88.2	89.6	96.1	98.3	64.2	96.6	95.2	87.5	82.4	70.2	25.3	82.7	81.6	57.2	53.6	65.7	72.6	73.3
	101	88.9	91.1	73.5	58.6	75.1	84.0	50.7	88.0	76.3	91.0	92.0	96.4	98.4	65.2	97.8	95.9	89.3	82.4	**73.6**	26.6	82.8	84.0	60.3	50.3	68.2	73.3	75.7
	50x4	91.3	90.5	73.0	65.7	77.0	85.9	57.3	88.4	79.5	91.9	92.5	97.8	98.5	68.1	97.8	96.4	89.7	85.5	59.4	30.3	83.0	85.7	62.6	52.5	68.0	76.6	78.2
	50x16	93.3	92.2	74.9	72.8	79.2	88.7	62.7	**89.0**	79.1	93.5	93.7	98.3	**98.9**	68.7	98.6	97.0	91.4	89.0	69.2	34.8	83.5	88.0	66.3	53.8	71.1	**80.0**	81.5
	50x64	94.8	94.1	78.6	77.2	81.1	90.5	67.7	**88.9**	**82.0**	94.5	95.4	98.9	**98.9**	**71.3**	99.1	97.1	92.8	90.2	69.2	40.7	83.7	89.5	69.1	55.0	**75.0**	**81.2**	83.6
CLIP-ViT	B/32	88.8	95.1	80.5	58.5	76.6	81.8	52.0	87.7	76.5	90.0	93.0	96.9	**99.0**	69.2	98.3	97.0	90.5	85.3	66.2	27.8	83.9	85.5	61.7	52.1	66.7	70.8	76.1
	B/16	92.8	96.2	83.1	67.8	78.4	86.7	59.5	**89.2**	79.2	93.1	94.7	98.1	**99.0**	69.5	99.0	97.1	92.7	86.6	67.8	33.3	83.5	88.4	66.1	**57.1**	70.3	75.5	80.2
	L/14	95.2	98.0	87.5	77.0	**81.8**	**90.9**	69.4	**89.6**	**82.1**	**95.1**	**96.5**	99.2	**99.2**	**72.2**	**99.7**	**98.2**	94.1	**92.5**	64.7	42.9	85.8	**91.5**	72.0	**57.8**	**76.2**	**80.8**	83.9
	L/14-336px	**95.9**	97.9	87.4	**79.9**	**82.2**	**91.5**	**71.6**	**89.9**	**83.0**	**95.1**	**96.0**	99.2	**99.2**	**72.9**	**99.7**	**98.1**	**94.9**	**92.4**	69.2	**46.4**	85.6	**92.0**	**73.0**	**60.3**	**77.3**	**80.5**	85.4
EfficientNet	B0	74.3	92.5	76.5	59.7	62.0	62.5	55.7	84.4	71.2	93.0	93.3	91.7	98.2	57.2	97.1	97.3	85.5	80.0	**73.8**	12.4	83.1	74.4	47.6	47.9	55.7	53.4	76.9
	B1	74.2	93.2	77.2	61.3	62.6	62.5	56.1	84.7	74.2	93.4	93.6	92.4	98.3	57.0	97.5	96.8	84.5	75.9	**75.5**	12.5	82.7	74.7	48.5	44.3	54.5	54.4	78.6
	B2	75.8	93.6	77.9	64.4	64.0	63.2	57.0	85.3	73.5	93.9	93.5	92.9	98.5	56.6	97.7	96.9	84.4	76.4	**73.1**	12.6	84.3	75.1	49.4	42.6	55.4	55.2	79.7
	B3	77.4	94.0	78.0	66.5	64.4	66.0	59.3	85.8	73.1	94.1	93.7	93.3	98.5	57.1	98.2	97.3	85.0	75.8	**76.1**	13.4	83.3	78.1	50.9	45.1	53.8	54.8	81.0
	B4	79.7	94.1	78.7	70.1	65.4	66.4	60.4	86.5	73.4	94.7	93.5	93.2	98.8	57.9	98.6	96.8	85.0	78.3	**72.3**	13.9	83.1	79.1	52.5	46.5	54.4	55.4	82.9
	B5	81.5	93.6	77.9	72.4	67.1	72.7	68.9	86.7	73.9	**95.0**	94.7	94.5	98.4	58.5	98.7	96.8	86.0	78.5	69.6	14.9	84.7	80.9	54.5	46.6	53.3	56.3	83.7
	B6	82.4	94.0	78.0	73.5	65.8	71.1	68.2	87.6	73.9	**95.0**	94.1	93.7	98.4	60.2	98.7	96.8	85.4	78.1	**72.7**	15.3	84.2	80.0	54.1	51.1	53.3	57.0	84.0
	B7	84.5	94.9	80.1	74.7	69.0	77.1	**72.3**	87.2	76.8	**95.2**	94.7	95.9	98.6	61.3	99.1	96.3	86.8	80.8	**75.8**	16.4	85.2	81.9	56.8	51.9	54.4	57.8	84.8
	B8	84.5	95.0	80.7	75.2	69.6	76.8	**71.5**	87.4	77.1	**94.9**	95.2	96.3	98.6	61.4	99.2	97.0	87.4	80.4	70.9	17.4	85.2	82.4	57.7	51.4	51.7	55.8	85.3
EfficientNet Noisy Student	B0	78.1	94.0	78.6	63.5	65.5	57.2	53.7	85.6	75.6	93.8	93.1	94.5	98.1	55.6	98.2	97.0	84.3	74.0	71.6	14.0	83.1	76.7	51.7	47.3	55.7	55.0	78.5
	B1	80.4	95.1	80.2	66.6	67.6	59.6	53.7	86.2	77.0	94.6	94.4	95.1	98.0	56.1	98.6	96.9	84.3	73.1	67.1	14.5	83.9	79.9	54.5	46.1	54.3	54.9	81.1
	B2	80.9	95.3	81.3	67.6	67.9	60.9	55.2	86.3	77.7	**95.0**	94.7	94.4	98.0	55.5	98.8	97.3	84.6	71.7	70.0	14.6	82.9	80.1	55.1	46.1	54.1	55.3	82.2
	B3	82.6	95.9	82.1	68.6	68.8	60.6	55.4	86.5	77.2	**95.0**	94.8	95.2	98.1	56.0	99.1	96.5	85.0	70.5	69.5	15.1	83.1	81.8	56.8	45.1	55.7	52.0	83.8
	B4	85.2	95.6	81.0	72.5	69.7	56.1	52.6	87.0	78.7	**94.8**	95.2	95.3	98.2	56.0	99.3	95.3	84.8	61.9	64.8	16.0	82.8	83.4	59.8	43.2	55.3	53.0	85.4
	B5	87.6	96.3	82.4	75.3	71.6	64.7	64.8	87.8	79.6	**95.5**	95.6	96.6	98.8	60.9	99.4	96.1	87.0	68.5	**73.7**	16.4	83.5	86.4	61.6	46.3	53.4	55.8	85.8
	B6	87.3	97.0	83.9	75.8	71.4	67.6	65.6	87.3	78.5	**95.2**	**96.4**	97.2	98.6	61.9	**99.5**	96.6	86.1	70.7	**72.4**	17.6	84.2	85.5	61.0	49.6	54.6	55.7	86.4
	B7	88.4	96.0	82.0	76.9	72.6	72.2	**71.2**	88.1	**80.5**	**95.5**	95.5	96.6	98.5	62.7	99.4	96.2	88.5	73.4	**73.0**	18.5	83.8	86.6	63.2	50.5	57.2	56.7	87.0
	L2-475	91.6	**99.0**	**91.0**	74.8	76.4	75.1	66.8	**89.5**	**81.9**	**95.6**	**96.5**	97.7	**98.9**	67.5	**99.6**	97.0	89.5	73.4	68.9	22.2	86.3	89.4	68.2	**58.3**	58.6	55.2	**88.3**
	L2-800	92.0	**98.7**	89.0	**78.5**	75.7	75.5	68.4	**89.4**	**82.5**	**95.6**	94.7	97.9	98.5	68.4	**99.7**	97.2	89.9	77.7	66.9	23.7	**86.8**	88.9	66.7	**62.7**	58.4	56.9	**88.4**
Instagram	32x8d	84.8	95.9	80.9	63.8	69.0	74.2	56.0	88.0	75.4	**95.4**	93.9	91.7	97.4	60.7	99.1	95.7	82.1	72.3	69.2	16.7	82.3	80.1	56.8	42.2	53.3	55.2	83.3
	32x16d	85.7	96.5	80.9	64.8	70.5	77.5	56.7	87.9	76.2	**95.6**	94.9	92.5	97.4	61.6	99.3	95.5	82.8	73.8	66.1	17.5	83.4	81.1	58.2	41.3	54.2	56.1	84.4
	32x32d	86.7	96.8	82.7	67.1	71.5	77.5	55.4	88.3	78.5	**95.8**	95.3	94.4	97.9	62.4	99.3	95.7	85.4	71.2	66.8	18.0	83.7	82.1	58.8	39.7	55.3	56.7	85.0
	32x48d	86.9	96.8	83.4	65.9	72.2	76.6	53.2	88.0	77.2	**95.5**	95.8	93.6	98.1	63.7	99.4	95.3	85.4	73.0	67.2	18.5	82.7	82.8	59.2	41.3	55.5	56.7	85.2
	FixRes-v1	88.5	95.7	81.1	67.4	72.9	80.5	57.6	88.0	77.9	**95.8**	**96.1**	94.5	97.9	62.2	99.4	96.2	86.6	76.5	64.8	19.3	82.5	83.4	59.8	43.5	56.6	59.0	86.0
	FixRes-v2	88.5	95.7	81.1	67.3	72.9	80.7	57.5	88.0	77.9	**95.0**	**96.0**	94.5	98.0	62.1	99.4	96.5	86.6	76.3	64.8	19.5	82.3	83.5	59.8	44.2	56.6	59.0	86.0
BiT-S	R50x1	72.5	91.7	74.8	57.7	61.1	53.5	52.5	83.7	72.4	92.3	91.2	92.0	98.4	56.1	96.4	97.4	85.0	70.0	66.0	12.5	83.0	72.3	47.5	48.3	54.1	55.3	75.2
	R50x3	75.1	93.7	79.0	61.1	63.7	55.2	54.1	84.8	74.6	92.5	91.6	92.8	98.8	58.7	97.0	**97.8**	86.4	73.1	**73.8**	14.0	84.2	76.4	50.0	49.2	54.7	54.2	77.2
	R101x1	73.5	92.8	77.4	58.4	61.3	54.0	52.4	84.4	73.5	92.5	91.8	90.6	98.3	56.5	96.8	97.3	84.6	69.4	68.9	12.6	82.0	73.5	48.6	45.4	52.6	55.5	76.0
	R101x3	74.7	93.9	79.8	57.8	62.9	54.7	53.3	84.7	75.5	92.3	91.2	92.6	98.8	59.7	97.3	**98.0**	85.5	71.8	60.2	14.1	83.1	75.9	50.4	49.7	54.1	54.6	77.4
	R152x2	74.9	94.3	79.7	58.7	62.7	55.9	53.6	85.3	74.9	93.0	92.0	91.7	98.6	58.3	97.1	**97.8**	86.2	71.8	71.6	13.9	84.1	76.2	49.9	48.2	53.8	55.9	77.1
	R152x4	74.7	94.2	79.2	57.8	62.9	51.2	50.8	85.4	75.4	93.1	91.2	91.4	**98.9**	61.4	97.2	**98.0**	85.5	72.8	67.9	14.9	83.1	76.0	50.3	42.9	53.6	56.0	78.5
BiT-M	R50x1	83.3	94.9	82.2	70.9	69.9	59.0	55.6	86.8	77.3	91.5	93.9	**99.4**	98.0	60.6	98.4	97.5	87.4	68.6	68.2	16.6	82.5	79.4	53.2	49.4	54.5	53.4	76.7
	R50x3	86.9	96.7	86.2	75.7	74.6	60.6	54.2	87.7	78.5	93.2	95.3	**99.4**	98.6	64.6	99.3	**98.0**	88.1	69.9	59.6	19.6	83.4	83.5	57.8	51.3	55.8	55.6	80.7
	R101x1	85.5	95.7	84.4	73.0	72.5	59.8	55.0	87.3	78.1	92.2	95.0	**99.5**	98.1	62.5	99.0	**97.6**	87.8	68.7	67.7	18.0	84.0	82.3	55.9	53.4	54.8	53.1	79.4
	R101x3	87.2	97.4	87.5	72.4	75.0	57.4	47.4	87.5	79.6	93.2	95.4	**99.6**	98.6	64.3	99.4	**98.2**	87.7	68.8	64.1	20.7	80.4	84.0	58.7	52.6	54.9	54.3	81.2
	R152x2	88.0	97.5	87.8	75.8	75.9	61.5	55.3	88.1	79.8	93.6	**95.9**	**99.5**	98.5	64.3	**99.5**	**97.9**	89.0	70.0	70.3	20.7	82.6	85.5	59.6	50.8	54.9	55.1	81.9
	R152x4	87.2	97.6	88.2	72.4	75.0	49.1	43.4	87.1	79.9	92.4	95.4	99.3	98.5	65.7	**99.5**	**97.8**	87.7	68.2	57.1	20.6	80.4	84.6	59.0	49.7	57.2	55.1	81.5
ViT	B/32	81.8	96.7	86.3	65.2	70.7	49.1	42.7	85.3	73.1	90.4	94.5	98.7	97.8	59.0	99.0	96.3	83.0	68.1	65.1	15.7	82.6	79.1	51.7	38.9	57.1	54.6	76.6
	B/16	86.7	96.9	86.4	74.0	74.2	54.7	46.0	86.7	74.3	92.7	94.1	99.2	97.4	61.3	99.5	96.4	84.5	63.1	61.5	17.5	85.4	82.7	56.6	40.0	57.0	56.1	80.9
	L/16	87.4	97.9	89.0	76.5	74.9	62.5	52.2	86.1	75.0	92.9	94.7	99.3	98.0	64.0	**99.6**	96.5	85.7	70.4	58.8	17.7	85.7	84.1	58.0	38.4	58.4	52.8	81.9
	H/14	83.4	95.8	84.5	70.2	69.2	62.3	54.8	84.7	75.4	91.7	93.7	98.9	98.5	62.4	98.4	97.3	87.0	73.9	63.4	15.4	**87.0**	79.4	52.1	41.1	55.9	54.1	75.9
SimCLRv2	R50x1	76.4	93.2	77.9	48.6	64.1	56.3	51.7	84.4	77.0	88.3	91.8	92.9	97.6	59.7	96.7	97.5	85.8	71.1	69.1	15.8	84.8	78.4	51.0	56.2	53.9	53.8	73.8
	R50x3	81.0	95.6	82.4	56.5	67.0	65.6	61.1	85.9	78.8	90.9	94.1	95.4	98.7	62.6	98.2	**97.9**	88.2	78.2	**74.7**	17.6	85.4	82.6	54.6	55.4	54.2	55.2	77.3
	R101x1	77.9	94.8	79.9	51.9	65.2	57.1	52.0	85.4	77.2	90.0	91.6	92.7	97.2	59.4	97.6	96.8	84.6	65.7	70.6	16.1	84.3	78.8	52.4	53.6	55.1	55.7	76.1
	R101x3	82.2	96.4	83.4	57.5	68.2	64.6	60.0	86.2	78.9	91.8	95.0	95.4	98.4	63.0	98.5	**97.9**	88.0	77.5	69.1	18.3	85.5	82.9	55.9	52.2	54.5	56.3	78.8
	R152x1	78.6	95.0	79.9	50.3	65.6	55.6	52.2	85.8	77.3	90.1	92.5	91.8	97.6	59.8	98.1	96.6	84.3	64.8	70.3	16.6	83.9	79.4	53.1	**57.2**	55.8	54.8	76.9
	R152x2	82.3	96.7	83.9	58.1	68.5	64.9	58.7	86.6	79.1	92.2	94.1	96.0	98.2	64.1	98.5	**98.0**	88.1	77.0	69.8	18.4	85.3	82.7	56.2	53.6	56.0	56.5	79.2
	R152x3	83.6	96.8	84.5	60.3	69.1	68.5	63.1	86.7	**80.5**	92.6	94.9	96.3	98.7	65.4	98.8	**98.1**	89.5	78.4	68.5	19.4	85.2	83.5	57.0	54.4	54.6	54.2	80.0
BYOL	50x1	74.0	93.6	79.1	47.6	63.7	61.6	62.3	82.6	77.0	88.3	93.7	94.3	98.7	58.8	96.4	**97.6**	88.2	80.1	71.4	14.1	84.8	77.3	49.3	56.1	53.8	54.4	73.3
	200x2	78.5	96.2	83.3	53.4	68.5	61.7	55.4	86.6	77.4	91.9	95.5	93.9	98.7	62.6	98.6	**97.7**	87.4	77.1	**76.4**	16.4	84.0	82.6	55.1	54.1	52.5	52.4	79.2
MoCo	v1	65.9	85.0	63.1	27.5	52.6	35.9	43.5	75.7	70.0	70.4	78.1	85.4	97.6	54.3	85.6	97.1	82.9	62.6	60.2	12.6	85.7	64.2	40.7	54.7	55.6	53.5	57.2
	v2	72.2	93.4	76.3	39.6	60.2	48.3	51.1	82.6	75.1	84.4	89.9	90.7	98.4	58.3	95.7	97.2	85.4	75.7	**75.4**	13.2	85.6	72.7	47.8	**56.9**	53.9	53.8	69.1
VirTex		57.9	83.9	57.5	17.0	49.8	22.4	34.5	83.8	58.2	53.6	70.6	74.7	98.1	56.5	86.7	94.8	74.1	69.5	71.3	8.7	83.1	61.5	39.9	45.5	53.5	55.8	50.7
ResNet	50	71.3	91.8	74.5	52.7	60.5	49.9	48.5	83.8	72.3	92.4	90.8	90.8	98.3	54.9	96.4	96.7	83.6	70.6	67.1	11.7	82.5	71.2	46.8	43.0	56.5	55.5	74.3
	101	72.7	93.0	77.2	53.7	60.8	50.1	47.0	84.4	71.6	92.3	91.9	90.4	98.5	56.6	97.0	97.1	83.4	72.5	63.6	11.9	83.3	72.7	48.3	43.2	53.0	54.7	75.8
	152	73.7	93.5	78.0	55.1	61.6	52.8	48.4	84.5	71.9	93.0	92.1	89.6	98.2	57.0	97.6	97.0	83.1	70.1	70.2	12.3	82.9	75.3	49.2	42.4	53.2	53.9	77.1

Table 10. Linear probe performance of various pre-trained models over 27 datasets. Scores within the 99.5% Clopper-Pearson confidence interval of each dataset's top score are shown in bold.

*We updated the STL10 scores from the previous version of this paper after fixing a CUDA-related bug.

		Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech101	Flowers	MNIST	FER2013	STL10	CIFAR100	ImageNet	HatefulMemes	SST2	KITTI	Country211	Mean
	LM RN50	81.3	82.8	61.7	44.2	69.6	74.9	44.9	85.5	71.5	82.8	85.5	91.1	96.6	60.1	95.3	93.4	84.0	73.8	70.2	19.0	82.9	76.4	51.9	51.2	65.2	76.8	65.2
NR PLC	50	86.4	88.7	70.3	56.4	73.3	78.3	49.1	87.1	76.4	88.2	89.6	96.1	98.3	64.2	96.6	95.2	87.5	82.4	70.2	25.3	82.7	81.6	57.2	53.6	65.7	72.6	73.3
	101	88.9	91.1	73.5	58.6	75.1	84.0	50.7	88.0	76.3	91.0	92.0	96.4	98.4	65.2	97.8	95.9	89.3	82.4	**73.6**	26.6	82.8	84.0	60.3	50.3	68.2	73.3	75.7
	50x4	91.3	90.5	73.0	65.7	77.0	85.9	57.3	88.4	79.5	91.9	92.5	97.8	98.5	68.1	97.8	96.4	89.7	85.5	59.4	30.3	83.0	85.7	62.6	52.5	68.0	76.6	78.2
	50x16	93.3	92.2	74.9	72.8	79.2	88.7	62.7	**89.0**	79.1	93.5	93.7	98.3	**98.9**	68.7	98.6	97.0	91.4	89.0	69.2	34.8	83.5	88.0	66.3	53.8	71.1	**80.0**	81.5
	50x64	94.8	94.1	78.6	77.2	81.1	90.5	67.7	**88.9**	**82.0**	94.5	95.4	98.9	**98.9**	**71.3**	99.1	97.1	92.8	90.2	69.2	40.7	83.7	89.5	69.1	55.0	**75.0**	**81.2**	83.6
TV PLC	B/32	88.8	95.1	80.5	58.5	76.6	81.8	52.0	87.7	76.5	90.0	93.0	96.9	**99.0**	69.2	98.3	97.0	90.5	85.3	66.2	27.8	83.9	85.5	61.7	52.1	66.7	70.8	76.1
	B/16	92.8	96.2	83.1	67.8	78.4	86.7	59.5	**89.2**	79.2	93.1	94.7	98.1	**99.0**	69.5	99.0	97.1	92.7	86.6	67.8	33.3	83.5	88.4	66.1	**57.1**	70.3	75.5	80.2
	L/14	95.2	98.0	87.5	77.0	**81.8**	**90.9**	69.4	**89.6**	**82.1**	**95.1**	**96.5**	99.2	**99.2**	**72.2**	**99.7**	**98.2**	94.1	**92.5**	64.7	42.9	85.8	**91.5**	72.0	**57.8**	**76.2**	**80.8**	83.9
	L/14-336px	**95.9**	97.9	87.4	**79.9**	**82.2**	**91.5**	**71.6**	**89.9**	**83.0**	**95.1**	**96.0**	99.2	**99.2**	**72.9**	**99.7**	**98.1**	**94.9**	**92.4**	69.2	**46.4**	85.6	**92.0**	**73.0**	**60.3**	**77.3**	**80.5**	85.4
EfficientNet	B0	74.3	92.5	76.5	59.7	62.0	62.5	55.7	84.4	71.2	93.0	93.3	91.7	98.2	57.2	97.1	97.3	85.5	80.0	**73.8**	12.4	83.1	74.4	47.6	47.9	55.7	53.4	76.9
	B1	74.2	93.2	77.2	61.3	62.6	62.5	56.1	84.7	74.2	93.4	93.6	92.4	98.3	57.0	97.5	96.8	84.5	75.9	**75.5**	12.5	82.7	74.7	48.5	44.3	54.5	54.4	78.6
	B2	75.8	93.6	77.9	64.4	64.0	63.2	57.0	85.3	73.5	93.9	93.5	92.9	98.5	56.6	97.7	96.9	84.4	76.4	**73.1**	12.6	84.3	75.1	49.4	42.6	55.4	55.2	79.7
	B3	77.4	94.0	78.0	66.5	64.4	66.0	59.3	85.8	73.1	94.1	93.7	93.3	98.5	57.1	98.2	97.3	85.0	75.8	**76.1**	13.4	83.3	78.1	50.9	45.1	53.8	54.8	81.0
	B4	79.7	94.1	78.7	70.1	65.4	66.4	60.4	86.5	73.4	94.7	93.5	93.2	98.8	57.9	98.6	96.8	85.0	78.3	**72.3**	13.9	83.1	79.1	52.5	46.5	54.4	55.4	82.9
	B5	81.5	93.6	77.9	72.4	67.1	72.7	68.9	86.7	73.9	**95.0**	94.7	94.5	98.4	58.5	98.7	96.8	86.0	78.5	69.6	14.9	84.7	80.9	54.5	46.6	53.3	56.3	83.7
	B6	82.4	94.0	78.0	73.5	65.8	71.1	68.2	87.6	73.9	**95.0**	94.1	93.7	98.4	60.2	98.7	96.8	85.4	78.1	**72.7**	15.3	84.2	80.0	54.1	51.1	53.3	57.0	84.0
	B7	84.5	94.9	80.1	74.7	69.0	77.1	**72.3**	87.2	76.8	**95.2**	94.7	95.9	98.6	61.3	99.1	96.3	86.8	80.8	**75.8**	16.4	85.2	81.9	56.8	51.9	54.4	57.8	84.8
	B8	84.5	95.0	80.7	75.2	69.6	76.8	**71.5**	87.4	77.1	**94.9**	95.2	96.3	98.6	61.4	99.2	97.0	87.4	80.4	70.9	17.4	85.2	82.4	57.7	51.4	51.7	55.8	85.3
EfficientNet NoisyStudent	B0	78.1	94.0	78.6	63.5	65.5	57.2	53.7	85.6	75.6	93.8	93.1	94.5	98.1	55.6	98.2	97.0	84.3	74.0	71.6	14.0	83.1	76.7	51.7	47.3	55.7	55.0	78.5
	B1	80.4	95.1	80.2	66.6	67.6	59.6	53.7	86.2	77.0	94.6	94.4	95.1	98.0	56.1	98.6	96.9	84.3	73.1	67.1	14.5	83.9	79.9	54.5	46.1	54.3	54.9	81.1
	B2	80.9	95.3	81.3	67.6	67.9	60.9	55.2	86.3	77.7	**95.0**	94.7	94.4	98.0	55.5	98.8	97.3	84.6	71.7	70.0	14.6	82.9	80.1	55.1	46.1	54.1	55.3	82.2
	B3	82.6	95.9	82.1	68.6	68.8	60.6	55.4	86.5	77.2	**95.0**	94.8	95.2	98.1	56.0	99.1	96.5	85.0	70.5	69.5	15.1	83.1	81.8	56.8	45.1	55.7	52.0	83.8
	B4	85.2	95.6	81.0	72.5	69.7	56.1	52.6	87.0	78.7	**94.8**	95.2	95.3	98.2	56.0	99.3	95.3	84.8	61.9	64.8	16.0	82.8	83.4	59.8	43.2	55.3	53.0	85.4
	B5	87.6	96.3	82.4	75.3	71.6	64.7	64.8	87.8	79.6	**95.5**	95.6	96.6	98.8	60.9	99.4	96.1	87.0	68.5	**73.7**	16.4	83.5	86.4	61.6	46.3	53.4	55.8	85.8
	B6	87.3	97.0	83.9	75.8	71.4	67.6	65.6	87.3	78.5	**95.2**	**96.4**	97.2	98.6	61.9	**99.5**	96.6	86.1	70.7	**72.4**	17.6	84.2	85.5	61.0	49.6	54.6	55.7	86.4
	B7	88.4	96.0	82.0	76.9	72.6	72.2	**71.2**	88.1	**80.5**	**95.5**	95.5	96.6	98.5	62.7	99.4	96.2	88.5	73.4	**73.0**	18.5	83.8	86.6	63.2	50.5	57.2	56.7	87.0
	L2-475	91.6	**99.0**	**91.0**	74.8	76.4	75.1	66.8	**89.5**	**81.9**	**95.6**	**96.5**	97.7	**98.9**	67.5	**99.6**	97.0	89.5	73.4	68.9	22.2	86.3	89.4	68.2	**58.3**	58.6	55.2	**88.3**
	L2-800	92.0	**98.7**	89.0	**78.5**	75.7	75.5	68.4	**89.4**	**82.5**	**95.6**	94.7	97.9	98.5	68.4	**99.7**	97.2	89.9	77.7	66.9	23.7	**86.8**	88.9	66.7	**62.7**	58.4	56.9	**88.4**
ResNeXt WSL	32x8d	84.8	95.9	80.9	63.8	69.0	74.2	56.0	88.0	75.4	**95.4**	93.9	91.7	97.4	60.7	99.1	95.7	82.1	72.3	69.2	16.7	82.3	80.1	56.8	42.2	53.3	55.2	83.3
	32x16d	85.7	96.5	80.9	64.8	70.5	77.5	56.7	87.9	76.2	**95.6**	94.9	92.5	97.4	61.6	99.3	95.5	82.8	73.8	66.1	17.5	83.4	81.1	58.2	41.3	54.2	56.1	84.4
	32x32d	86.7	96.8	82.7	67.1	71.5	77.5	55.4	88.3	78.5	**95.8**	95.3	94.4	97.9	62.4	99.3	95.7	85.4	71.2	66.8	18.0	83.7	82.1	58.8	39.7	55.3	56.7	85.0
	32x48d	86.9	96.8	83.4	65.9	72.2	76.6	53.2	88.0	77.2	**95.5**	95.8	93.6	98.1	63.7	99.4	95.3	85.4	73.0	67.2	18.5	82.7	82.8	59.2	41.3	55.5	56.7	85.2
	FixRes-v1	88.5	95.7	81.1	67.4	72.9	80.5	57.6	88.0	77.9	**95.8**	**96.1**	94.5	97.9	62.2	99.4	96.2	86.6	76.5	64.8	19.3	82.5	83.4	59.8	43.5	56.6	59.0	86.0
	FixRes-v2	88.5	95.7	81.1	67.3	72.9	80.7	57.5	88.0	77.9	**95.0**	**96.0**	94.5	98.0	62.1	99.4	96.5	86.6	76.3	64.8	19.5	82.3	83.5	59.8	44.2	56.6	59.0	86.0
SIMCLR	R50x1	72.5	91.7	74.8	57.7	61.1	53.5	52.5	83.7	72.4	92.3	91.2	92.0	98.4	56.1	96.4	97.4	85.0	70.0	66.0	12.5	83.0	72.3	47.5	48.3	54.1	55.3	75.2
	R50x3	75.1	93.7	79.0	61.1	63.7	55.2	54.1	84.8	74.6	92.5	91.6	92.8	98.8	58.7	97.0	**97.8**	86.4	73.1	**73.8**	14.0	84.2	76.4	50.0	49.2	54.7	54.2	77.2
	R101x1	73.5	92.8	77.4	58.4	61.3	54.0	52.4	84.4	73.5	92.5	91.8	90.6	98.3	56.5	96.8	97.3	84.6	69.4	68.9	12.6	82.0	73.5	48.6	45.4	52.6	55.5	76.0
	R101x3	74.7	93.9	79.8	57.8	62.9	54.7	53.3	84.7	75.5	92.3	91.2	92.6	98.8	59.7	97.3	**98.0**	85.5	71.8	60.2	14.1	83.1	75.9	50.4	49.7	54.1	54.6	77.4
	R152x2	74.9	94.3	79.7	58.7	62.7	55.9	53.6	85.3	74.9	93.0	92.0	91.7	98.6	58.3	97.1	**97.8**	86.2	71.8	71.6	13.9	84.1	76.2	49.9	48.2	53.8	55.9	77.1
	R152x4	74.7	94.2	79.2	57.8	62.9	51.2	50.8	85.4	75.4	93.1	91.2	91.4	**98.9**	61.4	97.2	**98.0**	85.5	72.8	67.9	14.9	83.1	76.0	50.3	42.9	53.6	56.0	78.5
MOCO	R50x1	83.3	94.9	82.2	70.9	69.9	59.0	55.6	86.8	77.3	91.5	93.9	**99.4**	98.0	60.6	98.4	97.5	87.4	68.6	68.2	16.6	82.5	79.4	53.2	49.4	54.5	53.4	76.7
	R50x3	86.9	96.7	86.2	75.7	74.6	60.6	54.2	87.7	78.5	93.2	95.3	**99.4**	98.6	64.6	99.3	**98.0**	88.1	69.9	59.6	19.6	83.4	83.5	57.8	51.3	55.8	55.6	80.7
	R101x1	85.5	95.7	84.4	73.0	72.5	59.8	55.0	87.3	78.1	92.2	95.0	**99.5**	98.1	62.5	99.0	**97.6**	87.8	68.7	67.7	18.0	84.0	82.3	55.9	53.4	54.8	53.1	79.4
	R101x3	87.2	97.4	87.5	72.4	75.0	57.4	47.4	87.5	79.6	93.2	95.4	**99.6**	98.6	64.3	99.4	**98.2**	87.7	68.8	64.1	20.7	80.4	84.0	58.7	52.6	54.9	54.3	81.2
	R152x2	88.0	97.5	87.8	75.8	75.9	61.5	55.3	88.1	79.8	93.6	**95.9**	**99.5**	98.5	64.3	**99.5**	**97.9**	89.0	70.0	70.3	20.7	82.6	85.5	59.6	50.8	54.9	55.1	81.9
	R152x4	87.2	97.6	88.2	72.4	75.0	49.1	43.4	87.1	79.9	92.4	95.4	99.3	98.5	65.7	**99.5**	**97.8**	87.7	68.2	57.1	20.6	80.4	84.6	59.0	49.7	57.2	55.1	81.5
ViT	B/32	81.8	96.7	86.3	65.2	70.7	49.1	42.7	85.3	73.1	90.4	94.5	98.7	97.8	59.0	99.0	96.3	83.0	68.1	65.1	15.7	82.6	79.1	51.7	38.9	57.1	54.6	76.6
	B/16	86.7	96.9	86.4	74.0	74.2	54.7	46.0	86.7	74.3	92.7	94.1	99.2	97.4	61.3	99.5	96.4	84.5	63.1	61.5	17.5	85.4	82.7	56.6	40.0	57.0	56.1	80.9
	L/16	87.4	97.9	89.0	76.5	74.9	62.5	52.2	86.1	75.0	92.9	94.7	99.3	98.0	64.0	**99.6**	96.5	85.7	70.4	58.8	17.7	85.7	84.1	58.0	38.4	58.4	52.8	81.9
	H/14	83.4	95.8	84.5	70.2	69.2	62.3	54.8	84.7	75.4	91.7	93.7	98.9	98.5	62.4	98.4	97.3	87.0	73.9	63.4	15.4	**87.0**	79.4	52.1	41.1	55.9	54.1	75.9
2vRLCms	R50x1	76.4	93.2	77.9	48.6	64.1	56.3	51.7	84.4	77.0	88.3	91.8	92.9	97.6	59.7	96.7	97.5	85.8	71.1	69.1	15.8	84.8	78.4	51.0	56.2	53.9	53.8	73.8
	R50x3	81.0	95.6	82.4	56.5	67.0	65.6	61.1	85.9	78.8	90.9	94.1	95.4	98.7	62.6	98.2	**97.9**	88.2	78.2	**74.7**	17.6	85.4	82.6	54.6	55.4	54.2	55.2	77.3
	R101x1	77.9	94.8	79.9	51.9	65.2	57.1	52.0	85.4	77.2	90.0	91.6	92.7	97.2	59.4	97.6	96.8	84.6	65.7	70.6	16.1	84.3	78.8	52.4	53.6	55.1	55.7	76.1
	R101x3	82.2	96.4	83.4	57.5	68.2	64.6	60.0	86.2	78.9	91.8	95.0	95.4	98.4	63.0	98.5	**97.9**	88.0	77.5	69.1	18.3	85.5	82.9	55.9	52.2	54.5	56.3	78.8
	R152x1	78.6	95.0	79.9	50.3	65.6	55.6	52.2	85.8	77.3	90.1	92.5	91.8	97.6	59.8	98.1	96.6	84.3	64.8	70.3	16.6	83.9	79.4	53.1	**57.2**	55.8	54.8	76.9
	R152x2	82.3	96.7	83.9	58.1	68.5	64.9	58.7	86.6	79.1	92.2	94.1	96.0	98.2	64.1	98.5	**98.0**	88.1	77.0	69.8	18.4	85.3	82.7	56.2	53.6	56.0	56.5	79.2
	R152x3	83.6	96.8	84.5	60.3	69.1	68.5	63.1	86.7	**80.5**	92.6	94.9	96.3	98.7	65.4	98.8	**98.1**	89.5	78.4	68.5	19.4	85.2	83.5	57.0	54.4	54.6	54.2	80.0
LOYO	50x1	74.0	93.6	79.1	47.6	63.7	61.6	62.3	82.6	77.0	88.3	93.7	94.3	98.7	58.8	96.4	**97.6**	88.2	80.1	71.4	14.1	84.8	77.3	49.3	56.1	53.8	54.4	73.3
	200x2	78.5	96.2	83.3	53.4	68.5	61.7	55.4	86.6	77.4	91.9	95.5	93.9	98.7	62.6	98.6	**97.7**	87.4	77.1	**76.4**	16.4	84.0	82.6	55.1	54.1	52.5	52.4	79.2
BoCoM	v1	65.9	85.0	63.1	27.5	52.6	35.9	43.5	75.7	70.0	70.4	78.1	85.4	97.6	54.3	85.6	97.1	82.9	62.6	60.2	12.6	85.7	64.2	40.7	54.7	55.6	53.5	57.2
	v2	72.2	93.4	76.3	39.6	60.2	48.3	51.1	82.6	75.1	84.4	89.9	90.7	98.4	58.3	95.7	97.2	85.4	75.7	**75.4**	13.2	85.6	72.7	47.8	**56.9**	53.9	53.8	69.1
	VirTex	57.9	83.9	57.5	17.0	49.8	22.4	34.5	83.8	58.2	53.6	70.6	74.7	98.1	56.5	86.7	94.8	74.1	69.5	71.3	8.7	83.1	61.5	39.9	45.5	53.5	55.8	50.7
ResNet	50	71.3	91.8	74.5	52.7	60.5	49.9	48.5	83.8	72.3	92.4	90.8	90.8	98.3	54.9	96.4	96.7	83.6	70.6	67.1	11.7	82.5	71.2	46.8	43.0	56.5	55.5	74.3
	101	72.7	93.0	77.2	53.7	60.8	50.1	47.0	84.4	71.6	92.3	91.9	90.4	98.5	56.6	97.0	97.1	83.4	72.5	63.6	11.9	83.3	72.7	48.3	43.2	53.0	54.7	75.8
	152	73.7	93.5	78.0	55.1	61.6	52.8	48.4	84.5	71.9	93.0	92.1	89.6	98.2	57.0	97.6	97.0	83.1	70.1	70.2	12.3	82.9	75.3	49.2	42.4	53.2	53.9	77.1

表10. 各种预训练模型在27个数据集上的线性探针性能。得分位于各数据集最高分99.5% Clopper-Pearson置信区间内的结果以粗体显示。

*We updated the STL10 scores from the previous version of this paper after fixing a CUDA-related bug.

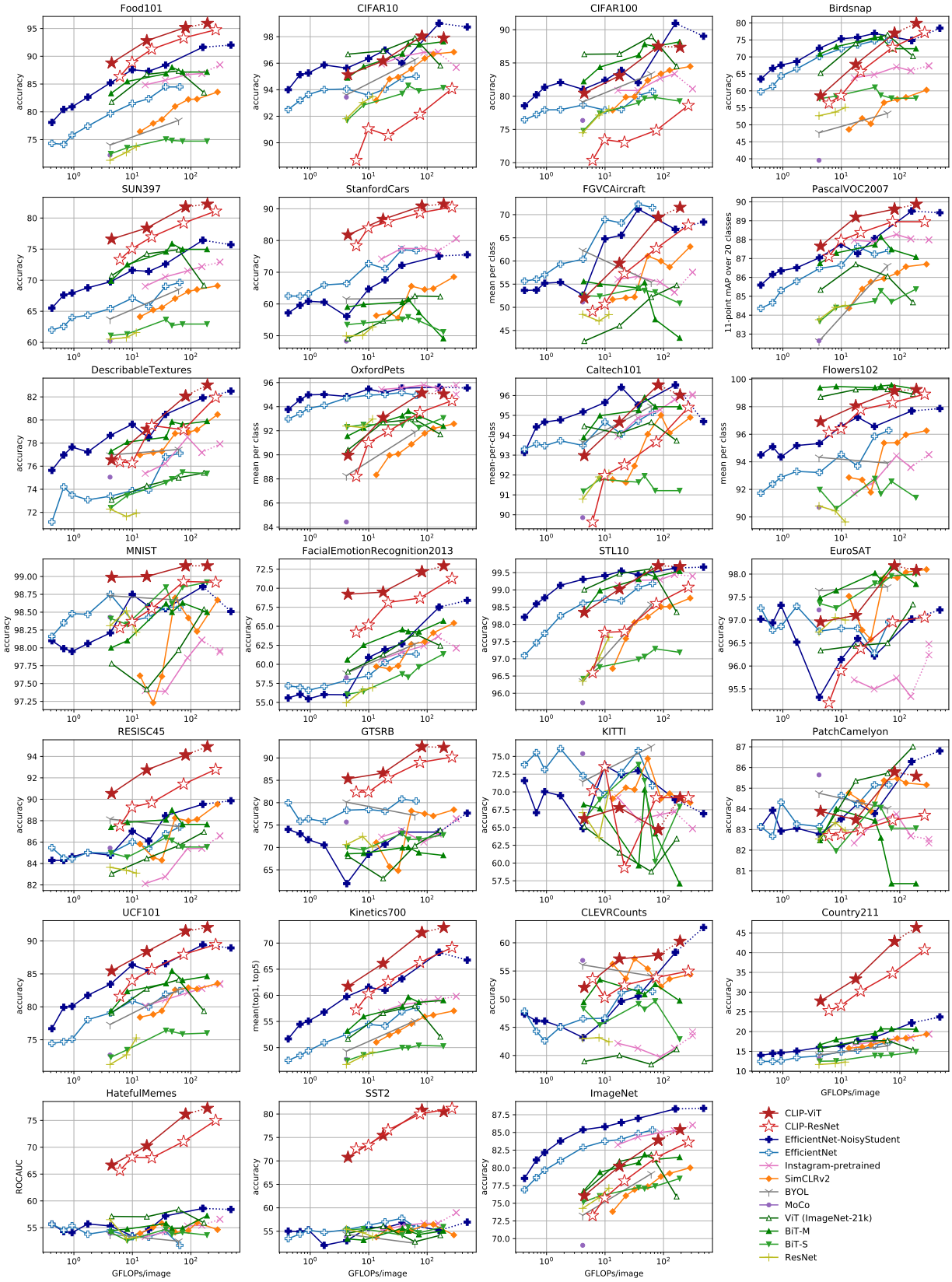

Figure 20. Linear probe performance plotted for each of the 27 datasets, using the data from Table 10.

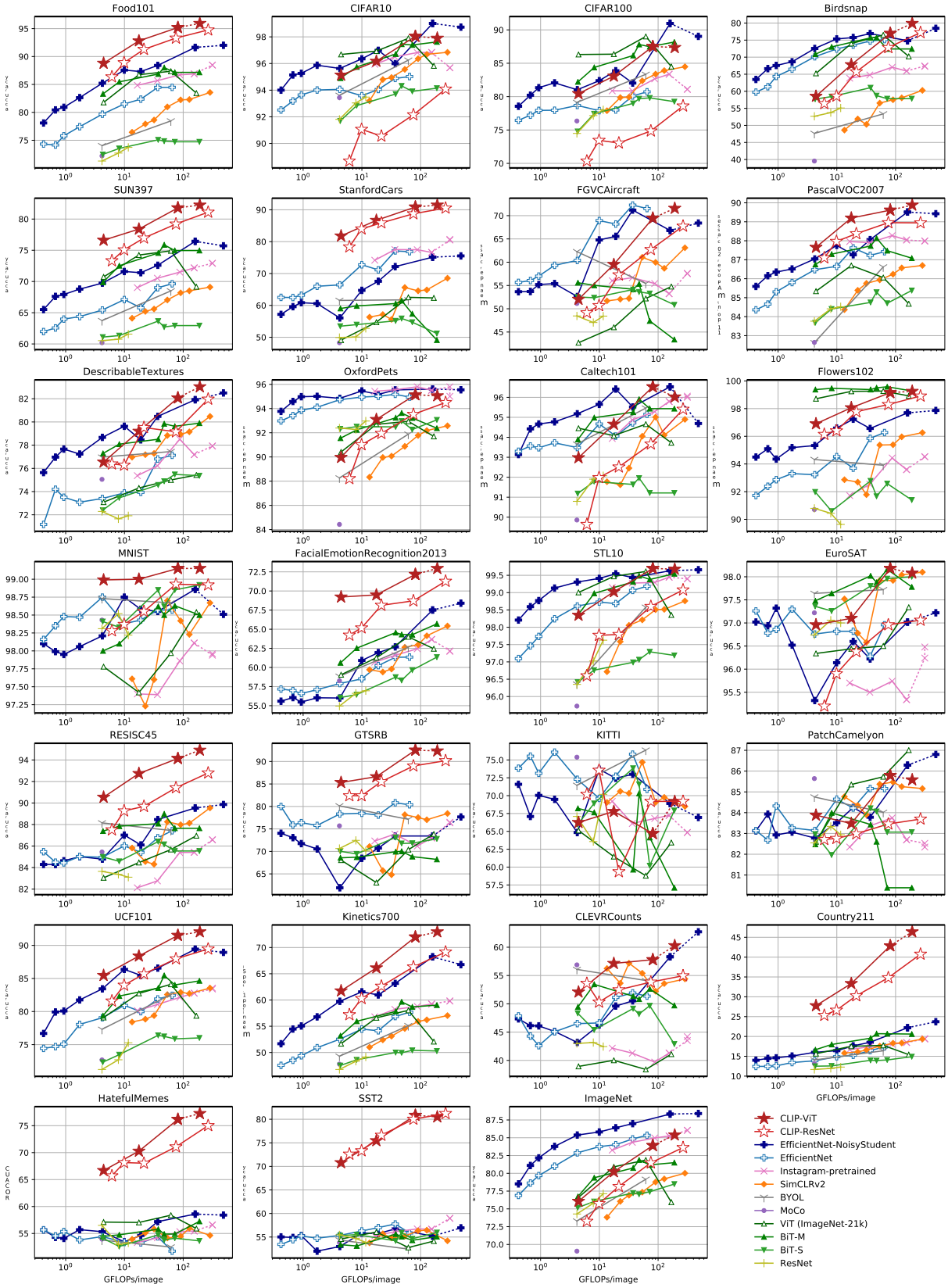

图20。使用表10中的数据，为27个数据集中的每一个绘制线性探测性能图。

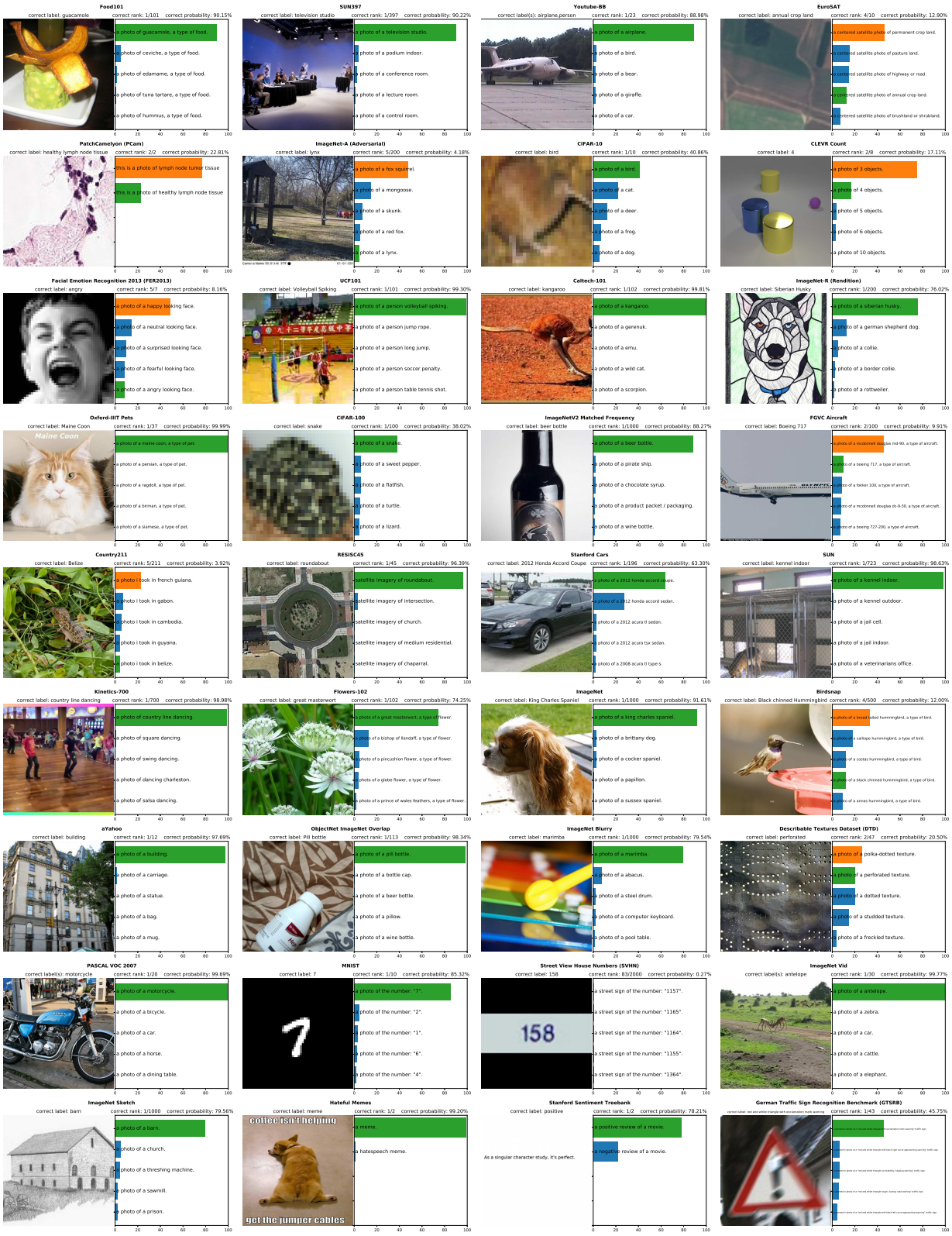

Figure 21. Visualization of predictions from 36 CLIP zero-shot classifiers. All examples are random with the exception of reselecting Hateful Memes to avoid offensive content. The predicted probability of the top 5 classes is shown along with the text used to represent the class. When more than one template is used, the first template is shown. The ground truth label is colored green while an incorrect prediction is colored orange.

图21. 36个CLIP零样本分类器的预测可视化。除为避免冒犯性内容而重新选择"恶意梗图"外，所有示例均为随机选取。图中展示前5个类别的预测概率及其对应文本表示。当使用多个模板时，仅显示首个模板。真实标签以绿色标注，错误预测以橙色标注。

		Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Stanford Cars	FGVC Aircraft	VOC2007	DTD	Oxford Pets	Caltech101	Flowers102	MNIST	FER2013	STL10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCam	UCF101	Kinetics700	CLEVR	HatefulMemes	Rendered SST2	ImageNet
CLIP-ResNet	RN50	81.1	75.6	41.6	32.6	59.6	55.8	19.3	82.1	41.7	85.4	82.1	65.9	66.6	42.2	94.3	41.1	54.2	35.2	42.2	16.1	57.6	63.6	43.5	20.3	59.7	56.9	59.6
	RN101	83.9	81.0	49.0	37.2	59.9	62.3	19.5	82.4	43.9	86.2	85.1	65.7	59.3	45.6	96.7	33.1	58.5	38.3	33.3	16.9	55.2	62.2	46.7	28.1	61.1	64.2	62.2
	RN50x4	86.8	79.2	48.9	41.6	62.7	67.9	24.6	83.0	49.3	88.1	86.0	68.0	75.2	51.1	96.4	35.0	59.2	35.7	26.0	20.2	57.5	65.5	49.0	17.0	58.3	66.6	65.8
	RN50x16	90.5	82.2	54.2	45.9	65.0	72.3	30.3	82.9	52.8	89.7	87.6	71.9	80.0	56.0	97.8	40.3	64.4	39.6	33.9	24.0	62.5	68.7	53.4	17.6	58.9	67.6	70.5
	RN50x64	91.8	86.8	61.3	48.9	66.9	76.0	35.6	83.8	53.4	93.4	90.6	77.3	90.8	61.0	98.3	59.4	69.7	47.9	33.2	29.6	65.0	74.1	56.8	27.5	62.1	70.7	73.6
CLIP-ViT	B/32	84.4	91.3	65.1	37.8	63.2	59.4	21.2	83.1	44.5	87.0	87.9	66.7	51.9	47.3	97.2	49.4	60.3	32.2	39.4	17.8	58.4	64.5	47.8	24.8	57.6	59.6	63.2
	B/16	89.2	91.6	68.7	39.1	65.2	65.6	27.1	83.9	46.0	88.9	89.3	70.4	56.0	52.7	98.2	54.1	65.5	43.3	44.0	23.3	48.1	69.8	52.4	23.4	61.7	59.8	68.6
	L/14	92.9	96.2	77.9	48.3	67.7	77.3	36.1	84.1	55.3	93.5	92.6	78.7	87.2	57.5	99.3	59.9	71.6	50.3	23.1	32.7	58.8	76.2	60.3	24.3	63.3	64.0	75.3
	L/14-336px	93.8	95.7	77.5	49.5	68.4	78.8	37.2	84.3	55.7	93.5	92.8	78.3	88.3	57.7	99.4	59.6	71.7	52.3	21.9	34.9	63.0	76.9	61.3	24.8	63.3	67.9	76.2

Table 11. Zero-shot performance of CLIP models over 27 datasets.

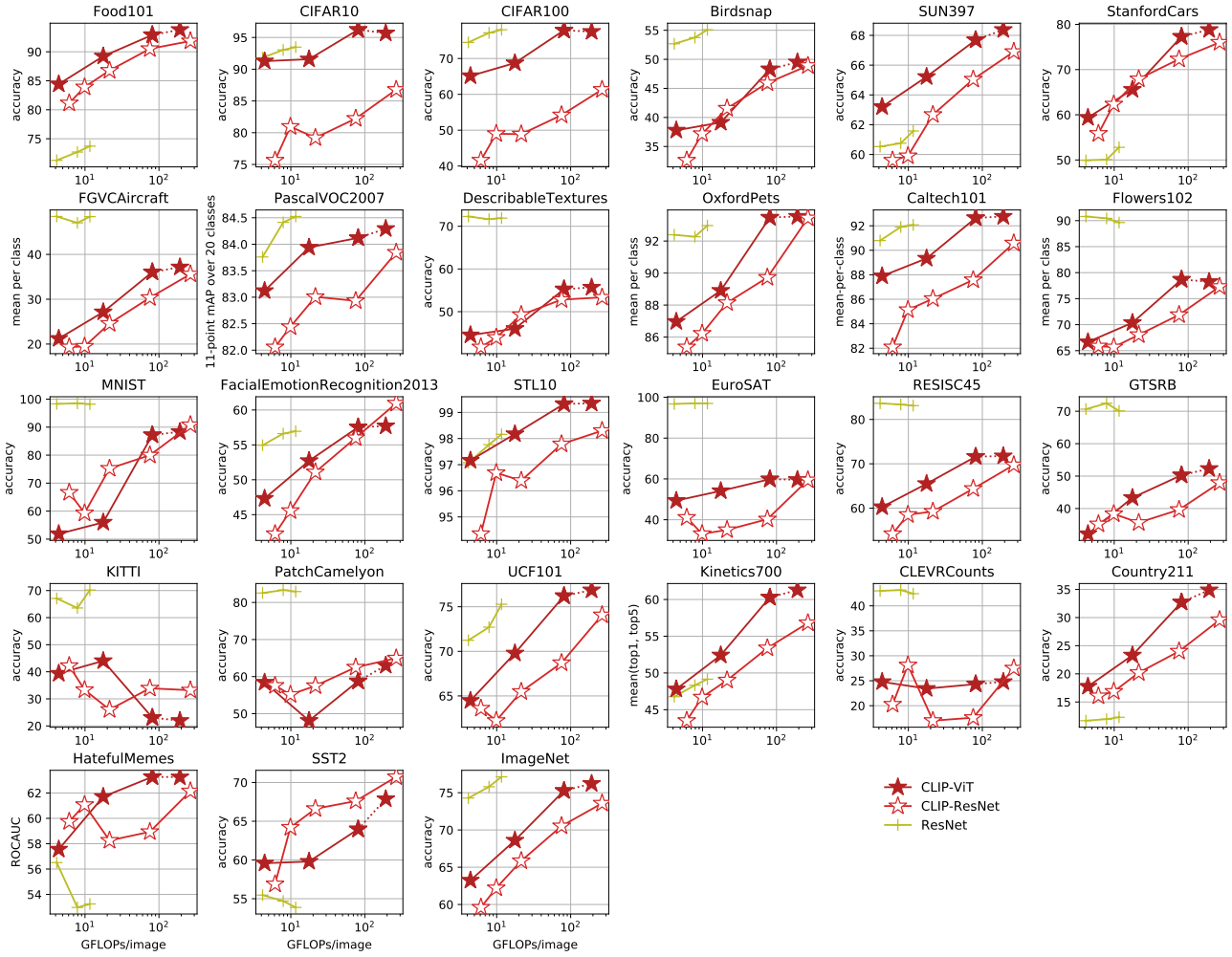

Figure 22. CLIP's zero-shot performance compared to linear-probe ResNet performance

		Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Stanford Cars	FGVC Aircraft	VOC2007	DTD	Oxford Pets	Caltech101	Flowers102	MNIST	FER2013	STL10	EuroSAT	RESISC45	GTSRB	KITTI	Pcam	UCF101	Kinetics700	CLEVR	Country211	HatefulMemes	SST2	ImageNet	ObjectNet	average
RN	RN50	81.1	75.6	41.6	32.6	59.6	55.8	19.3	82.1	41.7	85.4	82.1	65.9	66.6	42.2	94.3	41.1	54.2	35.2	42.2	16.1	57.6	63.6	43.5	20.3	59.7	56.9	59.6	40.4	56.4
	RN101	83.9	81.0	49.0	37.2	59.9	62.3	19.5	82.4	43.9	86.2	85.1	65.7	59.3	45.6	96.7	33.1	58.5	38.3	33.3	16.9	55.2	62.2	46.7	28.1	61.1	64.2	62.2	45.9	58.8
	RN50x4	86.8	79.2	48.9	41.6	62.7	67.9	24.6	83.0	49.3	88.1	86.0	68.0	75.2	51.1	96.4	35.0	59.2	35.7	26.0	20.2	57.5	65.5	49.0	17.0	58.3	66.6	65.8	48.3	60.7
	RN50x16	90.5	82.2	54.2	45.9	65.0	72.3	30.3	82.9	52.8	89.7	87.6	71.9	80.0	56.0	97.8	40.3	64.4	39.6	33.9	24.0	62.5	68.7	53.4	17.6	58.9	67.6	70.5	52.8	64.1
	RN50x64	91.8	86.8	61.3	48.9	66.9	76.0	35.6	83.8	53.4	93.4	90.6	77.3	90.8	61.0	98.3	59.4	69.7	47.9	33.2	29.6	65.0	74.1	56.8	27.5	62.1	70.7	73.6	56.1	68.3
ViT	B/32	84.4	91.3	65.1	37.8	63.2	59.4	21.2	83.1	44.5	87.0	87.9	66.7	51.9	47.3	97.2	49.4	60.3	32.2	39.4	17.8	58.4	64.5	47.8	24.8	57.6	59.6	63.2	45.4	59.6
	B/16	89.2	91.6	68.7	39.1	65.2	65.6	27.1	83.9	46.0	88.9	89.3	70.4	56.0	52.7	98.2	54.1	65.5	43.3	44.0	23.3	48.1	69.8	52.4	23.4	61.7	59.8	68.6	48.6	62.1
	L/14	92.9	96.2	77.9	48.3	67.7	77.3	36.1	84.1	55.3	93.5	92.6	78.7	87.2	57.5	99.3	59.9	71.6	50.3	23.1	32.7	58.8	76.2	60.3	24.3	63.3	64.0	75.3	57.9	67.7
	L/14-336px	93.8	95.7	77.5	49.5	68.4	78.8	37.2	84.3	55.7	93.5	92.8	78.3	88.3	57.7	99.4	59.6	71.7	52.3	21.9	34.9	63.0	76.9	61.3	24.8	63.3	67.9	76.2	58.9	68.7

表 11. CLIP 模型在 27 个数据集上的零样本性能。

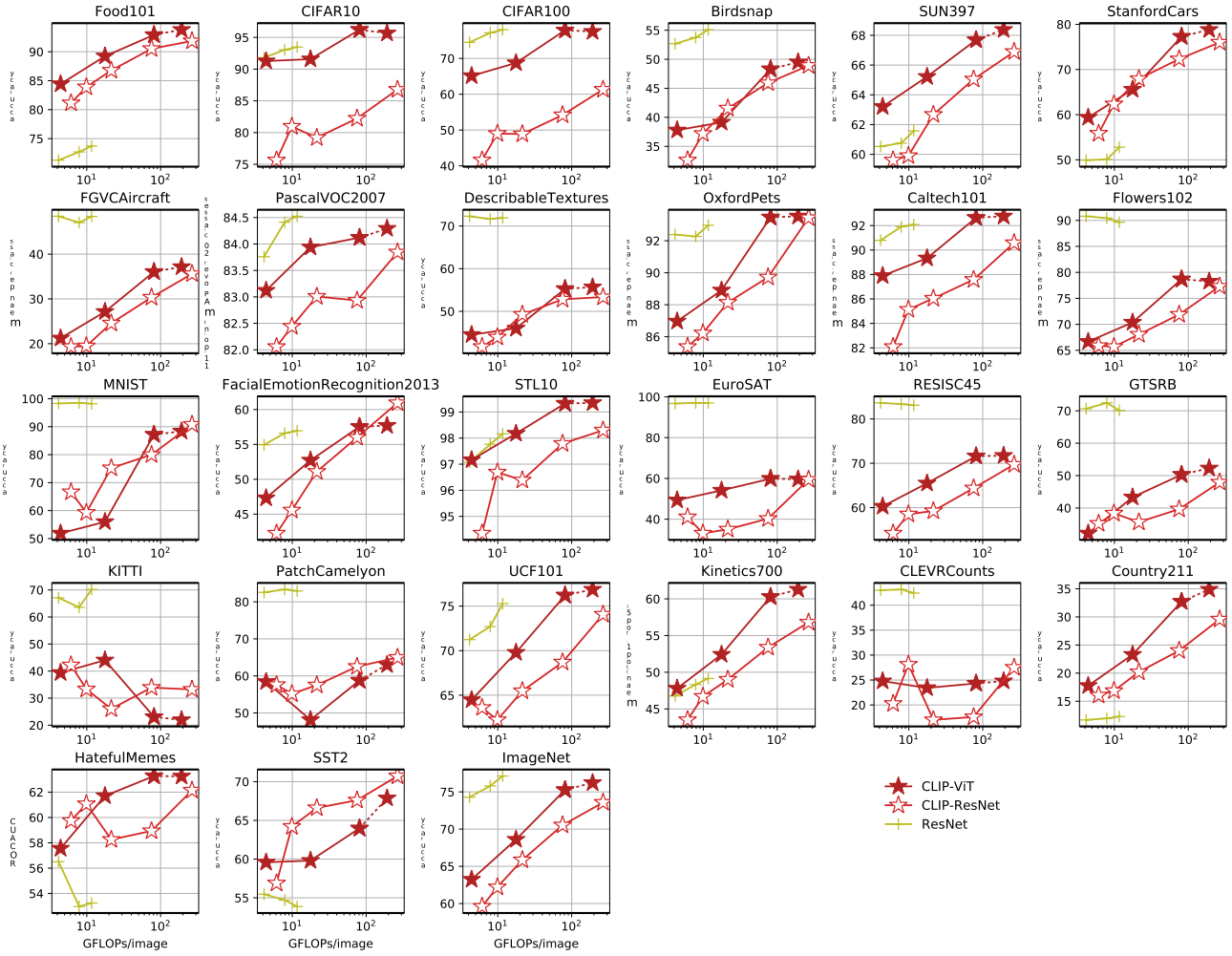

图22. CLIP的零样本性能与线性探针ResNet性能对比

B. Zero-Shot Prediction

To provide a qualitative summary / overview of CLIP's zero-shot performance we visualize a randomly selected prediction for 36 different zero-shot CLIP classifiers in Figure 21. In addition, Table 11 and Figure 22 show the individual zero-shot performance scores for each dataset.

C. Duplicate Detector

Our early attempts at duplicate detection and analysis used nearest neighbors in the model's learned embedding space. While it is intuitive to use a model's own notion of similarity, we encountered issues. We found the model's feature space is weighted very heavily towards semantic similarity. Many false positives occurred due to distinct objects that would be described similarly (soccer balls, flowers of the same species, etc...) having almost perfect similarity. We also observed the model was quite poor at assigning certain kinds of near-duplicates high similarity scores. We noticed repeatedly that images with high-frequency textures (such as fur or stripe patterns) pre-processed by different resizing algorithms (nearest neighbor vs bi-linear) could have surprisingly low similarity. This resulted in many false negatives.

We built our own near-duplicate detector to fix this issue. We created a synthetic data augmentation pipeline that combined a variety of common image manipulations. The augmentation pipeline combines random cropping and zooming, aspect ratio distortion, downsizing and upscaling to different resolutions, minor rotations, jpeg compression, and HSV color jitter. The pipeline also randomly selects from different interpolation algorithms for all relevant steps. We then trained a model to maximize the similarity of an image and its transformed variant while minimizing similarity to all other images in a training batch. We used the same n-pair / InfoNCE loss as CLIP but with a fixed temperature of 0.07.

We selected a ResNet-50 as the model architecture. We modified the base ResNet-50 with the anti-alias improvements from (Zhang, 2019) and used weight norm (Salimans & Kingma, 2016) instead of batch norm (Ioffe & Szegedy, 2015) to avoid leaking information about duplicates via batch statistics - a problem previously noted in (Henaff, 2020). We also found the GELU activation function (Hendrycks & Gimpel, 2016) to perform better for this task. We trained the model with a total batch size of 1,712 for approximately 30 million images sampled from our pre-training dataset. At the end of training it achieves nearly 100% accuracy on its proxy training task.

	Linear Classifier			*Zero Shot*		
Dataset	YFCC	WIT	Δ	YFCC	WIT	Δ
Birdsnap	47.4	35.3	+12.1	19.9	4.5	+15.4
Country211	23.1	17.3	+5.8	5.2	5.3	+0.1
Flowers102	94.4	89.8	+4.6	48.6	21.7	+26.9
GTSRB	66.8	72.5	−5.7	6.9	7.0	−0.1
UCF101	69.2	74.9	−5.7	22.9	32.0	−9.1
Stanford Cars	31.4	50.3	−18.9	3.8	10.9	−7.1
ImageNet	**62.0**	60.8	+1.2	**31.3**	27.6	+3.7
Dataset Average	65.5	**66.6**	−1.1	29.6	**30.0**	−0.4
Dataset "Wins"	10	**15**	−5	**19**	18	+1

Table 12. **CLIP performs similarly when trained on only YFCC100M.** Comparing a ResNet-50 trained on only YFCC100M with a same sized subset of WIT shows similar average performance and number of wins on zero shot and linear classifier evals. However, large differences in dataset specific performance occur. We include performance on the 3 datasets where YFCC does best and worst compared to WIT according to a linear probe in order to highlight this as well as aggregate performance across all linear and zero-shot evals and the canonical ImageNet dataset.

D. Dataset Ablation on YFCC100M

To study whether our custom dataset is critical to the performance of CLIP, we trained a model on a filtered subset of the YFCC100M dataset (details described in Section 2.2) and compared its performance to the same model trained on an equally sized subset of WIT. We train each model for 32 epochs at which point transfer performance begins to plateau due to overfitting. Results are shown in Table 12. Across our whole eval suite, YFCC and WIT perform similarly on average for both zero-shot and linear probe settings. However, performance on specific fine-grained classification datasets can vary widely - sometimes by over 10%. Our speculation is that these differences in performance reflect the relative density of relevant data in each pre-training dataset. For instance, pre-training on YFCC100M, which might contain many photos of birds and flowers (common subjects for photographers), results in better performance on Birdsnap and Flowers102, while pre-training on WIT results in better car and pet classifiers (which appear common in our dataset).

Overall, these results are encouraging as they suggest our approach can use any reasonably filtered collection of paired (text, image) data. This mirrors recent work which reported positive results using the same contrastive pre-training objective on the relatively different domain of medical imaging (Zhang et al., 2020). It also is similar to the findings of noisy student self-training which reported only slight improvements when using their JFT300M dataset over YFCC100M (Xie et al., 2020). We suspect the major advantage of our dataset over the already existing YFCC100M is its much larger size.

B. 零样本预测

为了对CLIP的零样本性能进行定性总结/概述，我们在图21中可视化了36个不同零样本CLIP分类器的随机预测结果。此外，表11和图22展示了每个数据集的独立零样本性能得分。

C. 重复检测器

我们早期在重复检测和分析方面的尝试使用了模型学习到的嵌入空间中的最近邻方法。虽然利用模型自身的相似性概念很直观，但我们遇到了一些问题。我们发现模型的特征空间非常侧重于语义相似性。由于不同物体（如足球、同一品种的花等）可能被描述得相似，导致许多误报情况，它们的相似度几乎完美。我们还观察到模型在给某些类型的近似重复项分配高相似度分数方面表现相当差。我们反复注意到，经过不同缩放算法（最近邻与双线性）预处理的高频纹理图像（如毛皮或条纹图案）可能具有出乎意料的低相似度。这导致了许多漏报情况。

我们构建了自己的近重复检测器来解决这个问题。我们创建了一个合成数据增强流程，该流程结合了多种常见的图像处理技术。增强流程包括随机裁剪和缩放、纵横比扭曲、降采样和上采样至不同分辨率、轻微旋转、JPEG压缩以及HSV颜色抖动。该流程还会在所有相关步骤中随机选择不同的插值算法。随后，我们训练了一个模型，旨在最大化图像与其变换版本之间的相似度，同时最小化与训练批次中所有其他图像的相似度。我们采用了与CLIP相同的n-pair/InfoNCE损失函数，但将温度参数固定为0.07。

我们选择了ResNet-50作为模型架构。我们在基础ResNet-50上采用了（Zhang, 2019）提出的抗锯齿改进，并使用权重归一化（Salimans & Kingma, 2016）替代批量归一化（Ioffe & Szegedy, 2015），以避免通过批量统计信息泄露重复样本的信息——这一问题先前已在（Henaff, 2020）中指出。我们还发现GELU激活函数（Hendrycks & Gimpel, 2016）在此任务中表现更佳。我们使用总批大小为1,712对模型进行训练，训练数据约包含从预训练数据集中采样的3,000万张图像。训练结束时，模型在其代理训练任务上达到了接近100%的准确率。

	Linear Classifier			*Zero Shot*		
Dataset	YFCC	WIT	Δ	YFCC	WIT	Δ
Birdsnap	47.4	35.3	+12.1	19.9	4.5	+15.4
Country211	23.1	17.3	+5.8	5.2	5.3	+0.1
Flowers102	94.4	89.8	+4.6	48.6	21.7	+26.9
GTSRB	66.8	72.5	−5.7	6.9	7.0	−0.1
UCF101	69.2	74.9	−5.7	22.9	32.0	−9.1
Stanford Cars	31.4	50.3	−18.9	3.8	10.9	−7.1
ImageNet	**62.0**	60.8	+1.2	**31.3**	27.6	+3.7
Dataset Average	65.5	**66.6**	−1.1	29.6	**30.0**	−0.4
Dataset “Wins”	10	**15**	−5	**19**	18	+1

表12. 仅使用YFCC100M训练时，CLIP表现相似。将仅在YFCC100M上训练的ResNet-50与WIT中相同规模的子集进行比较，结果显示在零样本和线性分类器评估中，两者的平均性能和获胜次数相近。然而，在特定数据集上的性能存在显著差异。我们列出了YFCC相对于WIT表现最佳和最差的3个数据集（基于线性探针评估）的性能，以突出这一点，同时汇总了所有线性与零样本评估以及经典ImageNet数据集上的整体表现。

D. YFCC100M数据集消融实验

为了研究我们的定制数据集是否对CLIP的性能至关重要，我们在YFCC100M数据集的过滤子集上训练了一个模型（细节见第2.2节），并将其性能与在同等规模的WIT子集上训练的相同模型进行了比较。每个模型均训练32个周期，此时由于过拟合，迁移性能开始趋于稳定。结果如表12所示。在整个评估体系中，YFCC和WIT在零样本和线性探针设置下的平均表现相似。然而，在特定的细粒度分类数据集上，性能可能存在显著差异——有时超过10%。我们推测这些性能差异反映了每个预训练数据集中相关数据的相对密度。例如，在可能包含大量鸟类和花卉照片（摄影师的常见主题）的YFCC100M上进行预训练，会导致在Birdsnap和Flowers102上表现更好；而在WIT上进行预训练，则能产生更优的汽车和宠物分类器（这些类别在我们的数据集中较为常见）。

总体而言，这些结果令人鼓舞，因为它们表明我们的方法可以利用任何经过合理筛选的（文本、图像）配对数据集合。这与近期一项研究相呼应，该研究在医学影像这一相对不同的领域使用相同的对比预训练目标报告了积极成果（Zhang et al., 2020）。这也与噪声学生自训练的发现相似，该研究在使用其JFT300M数据集相较于YFCC100M时仅报告了轻微改进（Xie et al., 2020）。我们推测，我们的数据集相较于现有YFCC100M的主要优势在于其规模要大得多。

Finally, we caution that WIT includes this filtered subset of YFCC100M. This could result in our ablation underestimating the size of performance differences between YFCC100M and the rest of WIT. We do not think this is likely as YFCC100M is only 3.7% of the overall WIT data blend and it did not noticeably change the performance of models when it was added to the existing data blend during the creation of WIT.

E. Selected Task and Dataset Results

Due to the large variety of datasets and experiments considered in this work, the main body focuses on summarizing and analyzing overall results. In the following subsections we report details of performance for specific groups of tasks, datasets, and evaluation settings.

E.1. Image and Text Retrieval

CLIP pre-trains for the task of image-text retrieval on our noisy web-scale dataset. Although the focus of this paper is on representation learning and task learning for the purpose of transfer to a wide variety of downstream datasets, validating that CLIP is able to achieve high transfer performance transfer on exactly what it is pre-trained for is an important sanity check / proof of concept. In Table 13 we check the zero-shot transfer performance of CLIP for both text and image retrieval on the Flickr30k and MSCOCO datsets. Zero-shot CLIP matches or outperforms all prior zero-shot results on these two datasets. Zero-shot CLIP is also competitive with the current overall SOTA for the task of text retrieval on Flickr30k. On image retrieval, CLIP's performance relative to the overall state of the art is noticeably lower. However, zero-shot CLIP is still competitive with a fine-tuned Unicoder-VL. On the larger MS-COCO dataset fine-tuning improves performance significantly and zero-shot CLIP is not competitive with the most recent work. For both these datasets we prepend the prompt "`a photo of`" to the description of each image which we found boosts CLIP's zero-shot R@1 performance between 1 and 2 points.

E.2. Optical Character Recognition

Although visualizations have shown that ImageNet models contain features that respond to the presence of text in an image (Zeiler & Fergus, 2014), these representations are not sufficiently fine-grained to use for the task of optical character recognition (OCR). To compensate, models are augmented with the outputs of custom OCR engines and features to boost performance on tasks where this capability is required (Singh et al., 2019; Yang et al., 2020). Early during the development of CLIP, we noticed that CLIP began to learn primitive OCR capabilities which appeared to steadily improve over the course of the project. To evaluate this qualitatively noticed behavior, we measured performance on 5 datasets requiring the direct and indirect use of OCR. Three of these datasets MNIST (LeCun), SVHN (Netzer et al., 2011), and IIIT5K (Mishra et al., 2012) directly check the ability of a model to perform low-level character and word recognition, while Hateful Memes (Kiela et al., 2020) and SST-2 (Socher et al., 2013) check the ability of a model to use OCR to perform a semantic task. Results are reported in Table 14.

CLIP's performance is still highly variable and appears to be sensitive to some combination of the domain (rendered or natural images) and the type of text to be recognized (numbers or words). CLIP's OCR performance is strongest Hateful Memes and SST-2 - datasets where the text is digitally rendered and consists mostly of words. On IIIT5K, which is natural images of individually cropped words, zero-shot CLIP performs a bit more respectively and its performance is similar to Jaderberg et al. (2014) early work combining deep learning and structured prediction to perform open-vocabulary OCR. However, performance is noticeably lower on two datasets involving recognition of hand written and street view numbers. CLIP's 51% accuracy on full number SVHN is well below any published results. Inspection suggests CLIP struggles with repeated characters as well as the low resolution and blurry images of SVHN. CLIP's zero-shot MNIST performance is also poor and is outperformed by supervised logistic regression on raw pixels, one of the simplest possible machine learning baselines.

SST-2 is a sentence level NLP dataset which we render into images. We include SST-2 in order to check whether CLIP is able to convert low level OCR capability into a higher level representation. Fitting a linear classifier on CLIP's representation of rendered sentences achives 80.5% accuracy. This is on par with the 80% accuracy of a continuous bag of words baseline using GloVe word vectors pre-trained on 840 billion tokens (Pennington et al., 2014). While this is a simple NLP baseline by today's standard, and well below the 97.5% of the current SOTA, it is encouraging to see that CLIP is able to turn an image of rendered text into a non-trivial sentence level representation. Fully supervised CLIP is also surprisingly strong on Hateful Meme detection, where CLIP is only 0.7 points behind the current single model SOTA and several points above the best baseline from the original paper. Similar to SST-2, these other results on Hateful Memes use the ground truth text which CLIP does not have access to. Finally, we note that zero-shot CLIP outperforms the best results using fully supervised linear probes across all other 56 models included in our evaluation suite. This suggests CLIP's OCR capability is at least somewhat unique compared to existing work on self-supervised and supervised representation learning.

最后，我们提醒注意，WIT包含了YFCC100M的这个过滤子集。这可能导致我们的消融实验低估了YFCC100M与WIT其余部分之间的性能差异规模。我们认为这种情况不太可能发生，因为YFCC100M仅占整个WIT数据混合的3.7%，且在创建WIT时将其添加到现有数据混合中并未显著改变模型的性能。

E. 选定任务与数据集结果

鉴于本研究中考虑的数据集和实验种类繁多，正文部分主要侧重于总结和分析整体结果。在接下来的小节中，我们将针对特定任务组、数据集和评估设置报告详细的性能表现。

E.1. 图像与文本检索

CLIP在我们的嘈杂网络规模数据集上进行了图像-文本检索任务的预训练。尽管本文的重点在于表征学习和任务学习，旨在迁移至广泛的下游数据集，但验证CLIP能否在其预训练任务上实现高迁移性能，是一项重要的概念验证与合理性检验。在表13中，我们评估了CLIP在Flickr30k和MSCOCO数据集上文本与图像检索的零样本迁移性能。零样本CLIP在这两个数据集上达到或超越了所有先前的零样本结果。在Flickr30k的文本检索任务中，零样本CLIP也与当前整体最优方法具有竞争力。在图像检索方面，CLIP相对于整体最优技术的性能明显较低。然而，零样本CLIP仍可与经过微调的Unicoder-VL相媲美。在更大的MS-COCO数据集上，微调显著提升了性能，零样本CLIP则无法与最新研究工作竞争。针对这两个数据集，我们在每张图像的描述前添加了提示语“一张关于”，这一操作使CLIP的零样本R@1性能提升了1到2个百分点。

E.2. 光学字符识别

尽管可视化显示ImageNet模型包含对图像中文本存在做出响应的特征（Zeiler & Fergus, 2014），但这些表征不够精细，无法用于光学字符识别（OCR）任务。为弥补这一不足，模型通过集成定制OCR引擎的输出和特征来增强，以提升在需要此能力任务上的性能（Singh等人，2019；Yang等人，2020）。在CLIP开发的早期阶段，我们注意到CLIP开始学习初级的OCR能力，且这种能力在项目推进过程中持续提升。为定性评估这一观察到的行为，我们测量了其性能表现。

在5个需要直接或间接使用OCR的数据集上。其中三个数据集——MNIST（LeCun）、SVHN（Netzer等人，2011）和IIIT5K（Mishra等人，2012）直接检验模型执行低级字符和单词识别的能力，而Hateful Memes（Kiela等人，2020）和SST-2（Socher等人，2013）则检验模型利用OCR执行语义任务的能力。结果详见表14。

CLIP的表现仍然存在很大差异，并且似乎对领域（渲染图像或自然图像）与待识别文本类型（数字或单词）的某种组合较为敏感。CLIP的OCR性能在Hateful Memes和SST-2数据集上表现最强——这些数据集的文本均为数字渲染且主要由单词构成。在IIIT5K数据集（包含单独裁剪单词的自然图像）上，零样本CLIP的表现相对更好，其性能与Jaderberg等人（2014）早期结合深度学习与结构化预测进行开放词汇OCR的研究成果相近。然而，在涉及手写数字和街景门牌号识别的两个数据集上，CLIP的表现明显较差。CLIP在完整SVHN数字数据集上51%的准确率远低于任何已发表的研究结果。分析表明CLIP在处理重复字符以及SVHN的低分辨率模糊图像时存在困难。CLIP在零样本MNIST任务上的表现同样不佳，甚至被基于原始像素的监督逻辑回归（一种最简单的机器学习基线方法）所超越。

SST-2是一个句子级别的自然语言处理数据集，我们将其渲染为图像。我们引入SST-2是为了检验CLIP能否将低层级的OCR能力转化为更高层级的表征。在CLIP对渲染句子的表征上拟合线性分类器，达到了80.5%的准确率。这与基于GloVe词向量（在8400亿词例上预训练）的连续词袋基线模型80%的准确率相当（Pennington等人，2014）。尽管以现今标准看这是一个简单的自然语言处理基线，且远低于当前最优模型97.5%的准确率，但令人鼓舞的是，CLIP能够将渲染文本的图像转化为具有实际意义的句子级表征。完全监督的CLIP在恶意表情包检测任务上也表现出惊人的强大性能，仅比当前单模型最优结果低0.7个百分点，且比原论文中的最佳基线高出数个百分点。与SST-2类似，这些在恶意表情包上的其他结果使用了CLIP无法访问的真实文本标签。最后我们注意到，零样本CLIP在我们评估套件包含的其他56个模型中，均优于使用完全监督线性探针的最佳结果。这表明CLIP的OCR能力相较于当前自监督与监督表征学习的研究成果，至少具备一定独特性。

		Text Retrieval						Image Retrieval					
		Flickr30k			MSCOCO			Flickr30k			MSCOCO		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Finetune	Unicoder-VL[a]	86.2	96.3	99.0	62.3	87.1	92.8	71.5	90.9	94.9	46.7	76.0	85.3
	Uniter[b]	87.3	98.0	99.2	65.7	88.6	93.8	75.6	94.1	**96.8**	52.9	79.9	88.0
	VILLA[c]	87.9	97.5	98.8	-	-	-	76.3	**94.2**	**96.8**	-	-	-
	Oscar[d]	-	-	-	**73.5**	**92.2**	**96.0**	-	-	-	**57.5**	**82.8**	**89.8**
	ERNIE-ViL[e]	**88.7**	98.0	99.2	-	-	-	**76.7**	93.6	96.4	-	-	-
Zero-Shot	Visual N-Grams[f]	15.4	35.7	45.1	8.7	23.1	33.3	8.8	21.2	29.9	5.0	14.5	21.9
	ImageBERT[g]	-	-	-	44.0	71.2	80.4	-	-	-	32.3	59.0	70.2
	Unicoder-VL[a]	64.3	86.8	92.3	-	-	-	48.4	76.0	85.2	-	-	-
	Uniter[b]	83.6	95.7	97.7	-	-	-	68.7	89.2	93.9	-	-	-
	CLIP	88.0	**98.7**	**99.4**	58.4	81.5	88.1	68.7	90.6	95.2	37.8	62.4	72.2

Table 13. **CLIP improves zero-shot retrieval and is competitive with the best fine-tuned result on Flickr30k text retrieval.** Bold indicates best overall performance while an underline indicates best in category performance (zero-shot or fine-tuned). For all other models, best results from the paper are reported regardless of model size / variant. MSCOCO performance is reported on the 5k test set. [a](Li et al., 2020a) [b](Chen et al., 2019) [c](Gan et al., 2020) [d](Li et al., 2020b) [e](Yu et al., 2020) [f](Li et al., 2017) [g](Qi et al., 2020)

		MNIST	SVHN	IIIT5K 1k	Hateful Memes	SST-2
Finetune	SOTA	**99.8**[a]	**96.4**[b]	**98.9**[c]	**78.0**[d]	**97.5**[e]
	JOINT[f]	-	-	89.6	-	-
	CBoW[g]	-	-	-	-	80.0
Linear	Raw Pixels	92.5	-	-	-	-
	ES Best	98.9[h]	-	-	58.6[h]	59.0[i]
	CLIP	99.2	-	-	77.3	80.5
ZS	CLIP	88.4	51.0	90.0	63.3	67.9

Table 14. OCR performance on 5 datasets. All metrics are accuracy on the test set except for Hateful Memes which reports ROC AUC on the dev set. Single model SOTA reported to best of knowledge. *ES Best* reports the best performance across the 56 non-CLIP models in our evaluation suite. [a](Assiri, 2020) [b](Jaderberg et al., 2015) [c](Wang et al., 2020) [d](Lippe et al., 2020) [f](Jaderberg et al., 2014) [g](Wang et al., 2018) [h](Xie et al., 2020) [i](Mahajan et al., 2018)

		UCF101	K700	RareAct	
		Top-1	AVG	mWAP	mWSAP
Finetune	R(2+1)D-BERT[a]	**98.7**	-	-	-
	NS ENet-L2[b]	-	**84.8**	-	-
	HT100M S3D[d]	91.3	-	-	-
	Baseline I3D[e]	-	70.2	-	-
Linear	MMV FAC[f]	91.8	-	-	-
	NS ENet-L2[c]	89.4[c]	68.2[c]	-	-
	CLIP	92.0	73.0	-	-
ZS	HT100M S3D[d]	-	-	30.5	34.8
	CLIP	80.3	69.6	**40.7**	**44.8**

Table 15. Action recognition performance on 3 video datasets. Single model SOTA reported to best of knowledge. Note that *linear CLIP* and *linear NS ENet-L2* are trained and evaluated on a single frame subsampled version of each dataset and not directly comparable to prior work. On Kinetics-700, we report the ActivityNet competition metric which is the average of top-1 and top-5 performance. [a](Kalfaoglu et al., 2020) [b](Lu et al., 2020) [c](Xie et al., 2020) [d](Miech et al., 2020b) [e](Carreira et al., 2019) [f](Alayrac et al., 2020)

E.3. Action Recognition in Videos

For the purpose of learning, a potentially important aspect of natural language is its ability to express, and therefore supervise, an extremely wide set of concepts. A CLIP model, since it is trained to pair semi-arbitrary text with images, is likely to receive supervision for a wide range of visual concepts involving both common and proper nouns, verbs, and adjectives. ImageNet-1K, by contrast, only labels common nouns. Does the lack of broader supervision in ImageNet result in weaker transfer of ImageNet models to tasks involving the recognition of visual concepts that are not nouns?

To investigate this, we measure and compare the performance of CLIP and ImageNet models on several video action classification datasets which measure the ability of a model to recognize verbs. In Table 15 we report results on UCF-101 (Soomro et al., 2012) and Kinetics-700 (Carreira et al., 2019), two common datasets for the task. Unfortunately, our CPU based linear classifier takes a prohibitively long time to evaluate on a video dataset due to the very large number of training frames. To deal with this, we aggressively sub-sample each video to only a single center frame, effectively turning it into an image classification dataset. As a result, our reported performance in a linear evaluation setting likely under estimates performance by a moderate amount.

		Text Retrieval						Image Retrieval					
		Flickr30k			MSCOCO			Flickr30k			MSCOCO		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Fine-tune	Unicoder-VL[a]	86.2	96.3	99.0	62.3	87.1	92.8	71.5	90.9	94.9	46.7	76.0	85.3
	Uniter[b]	87.3	98.0	99.2	65.7	88.6	93.8	75.6	94.1	**96.8**	52.9	79.9	88.0
	VILLA[c]	87.9	97.5	98.8	-	-	-	76.3	**94.2**	**96.8**	-	-	-
	Oscar[d]	-	-	-	**73.5**	**92.2**	**96.0**	-	-	-	**57.5**	**82.8**	**89.8**
	ERNIE-ViL[e]	**88.7**	98.0	99.2	-	-	-	**76.7**	93.6	96.4	-	-	-
Zero-Shot	Visual N-Grams[f]	15.4	35.7	45.1	8.7	23.1	33.3	8.8	21.2	29.9	5.0	14.5	21.9
	ImageBERT[g]	-	-	-	44.0	71.2	80.4	-	-	-	32.3	59.0	70.2
	Unicoder-VL[a]	64.3	86.8	92.3	-	-	-	48.4	76.0	85.2	-	-	-
	Uniter[b]	83.6	95.7	97.7	-	-	-	68.7	89.2	93.9	-	-	-
	CLIP	88.0	**98.7**	**99.4**	58.4	81.5	88.1	68.7	90.6	95.2	37.8	62.4	72.2

表13。CLIP提升了零样本检索性能，并在Flickr30k文本检索任务中与最佳微调结果具有竞争力。粗体表示整体最佳性能，下划线表示类别内最佳性能（零样本或微调）。对于其他所有模型，均报告论文中的最佳结果（不考虑模型规模/变体）。MSCOCO性能基于5k测试集报告。[a](Li等人，2020a) [b](Chen等人，2019) [c](Gan等人，2020) [d](Li等人，2020b) [e](Yu等人，2020) [f](Li等人，2017) [g](Qi等人，2020)

		MNIST	SVHN	IIIT5K 1k	Hateful Memes	SST-2
Fine-tune	SOTA	**99.8**[a]	**96.4**[b]	**98.9**[c]	**78.0**[d]	**97.5**[e]
	JOINT[f]	-	-	89.6	-	-
	CBoW[g]	-	-	-	-	80.0
Linear	Raw Pixels	92.5	-	-	-	-
	ES Best	98.9[h]	-	-	58.6[h]	59.0[i]
	CLIP	99.2	-	-	77.3	80.5
Zero-Shot	CLIP	88.4	51.0	90.0	63.3	67.9

表14. 5个数据集上的OCR性能。除Hateful Memes在开发集上报告ROC AUC外，所有指标均为测试集准确率。单模型SOTA据我们所知的最佳报告。*ES Best*报告了我们评估套件中56个非CLIP模型中的最佳性能。[a](Assiri, 2020) [b](Jaderberg et al., 2015) [c](Wang et al., 2020) [d](Lippe et al., 2020) [f](Jaderberg et al., 2014) [g](Wang et al., 2018) [h](Xie et al., 2020) [i](Mahajan et al., 2018)

		UCF101	K700	RareAct	
		Top-1	AVG	mWAP	mWSAP
Fine-tune	R(2+1)D-BERT[a]	**98.7**	-	-	-
	NS ENet-L2[b]	-	**84.8**	-	-
	HT100M S3D[d]	91.3	-	-	-
	Baseline I3D[e]	-	70.2	-	-
Linear	MMV FAC[f]	91.8	-	-	-
	NS ENet-L2[c]	89.4[c]	68.2[c]	-	-
	CLIP	92.0	73.0	-	-
Zero-Shot	HT100M S3D[d]	-	-	30.5	34.8
	CLIP	80.3	69.6	**40.7**	**44.8**

表15. 在3个视频数据集上的动作识别性能。据我们所知，报告的是单模型SOTA。请注意，*linear*、*CLIP*和*linear NS ENet-L2*是在每个数据集的单帧子采样版本上进行训练和评估的，与先前工作不直接可比。在Kinetics-700上，我们报告的是ActivityNet竞赛指标，即top-1和top-5性能的平均值。[a](Kalfaoglu等人，2020) [b](Lu等人，2020) [c](Xie等人，2020) [d](Miech等人，2020b) [e](Carreira等人，2019) [f](Alayrac等人，2020)

E.3. 视频中的动作识别

为了学习的目的，自然语言一个潜在的重要方面是它能够表达并因此监督极其广泛的概念集合。CLIP模型由于被训练来将半任意文本与图像配对，很可能接收到涉及普通名词、专有名词、动词和形容词的广泛视觉概念的监督。相比之下，ImageNet-1K仅标注普通名词。ImageNet中更广泛监督的缺失是否会导致ImageNet模型在涉及非名词视觉概念识别任务上的迁移能力较弱?

为了研究这一点，我们测量并比较了CLIP与ImageNet模型在多个视频任务上的性能。

动作分类数据集用于衡量模型识别动词的能力。在表15中，我们报告了在UCF-101（Soomro等人，2012年）和Kinetics-700（Carreira等人，2019年）这两个该任务常用数据集上的结果。遗憾的是，由于训练帧数量极大，我们基于CPU的线性分类器在视频数据集上的评估耗时过长。为解决这一问题，我们对每个视频进行激进下采样，仅保留单个中心帧，从而将其转化为图像分类数据集。因此，我们在线性评估设置中报告的性能可能被适度低估。

	IN Top-1	IN-V2 Top-1	IN-A Top-1	IN-R Top-1	ObjectNet Top-1	IN-Sketch Top-1	IN-Vid PM0	IN-Vid PM10	YTBB PM0	YTBB PM10
NS EfficientNet-L2[a]	**88.3**	**80.2**	**84.9**	74.7	68.5	47.6	88.0	82.1	67.7	63.5
FixResNeXt101-32x48d V2[b]	86.4	78.0	68.4	80.0	57.8	59.1	85.8	72.2	68.9	57.7
Linear Probe CLIP	85.4	75.9	75.3	84.2	66.2	57.4	89.1	77.2	68.7	63.1
Zero-Shot CLIP	76.2	70.1	77.2	**88.9**	**72.3**	**60.2**	**95.3**	**89.2**	**95.2**	**88.5**

Table 16. Detailed ImageNet robustness performance. IN is used to abbreviate for ImageNet. [a](Xie et al., 2020) [b](Touvron et al., 2019)

Despite this handicap, CLIP features transfer surprisingly well to this task. CLIP matches the best prior result on UCF-101 in a linear probe evaluation setting and also outperforms all other models in our evaluation suite. On Kinetics-700, CLIP also outperforms the fine-tuned I3D baseline from the original paper. Since it does not require a training stage, we report CLIP's zero-shot performance when averaging predictions across all frames. CLIP also performs well in this setting and on Kinetics-700 its performance is within 1% of the fully supervised I3D baseline which is trained on 545000 labeled videos. Encouraged by these results, we also measure CLIP's performance on the recently introduced RareAct dataset (Miech et al., 2020a) which was designed to measure zero-shot recognition of unusual actions like "hammering a phone" and "drilling an egg". CLIP improves over the prior state of the art, a S3D model trained on automatically extracted captions from 100 million instructional videos, by 10 points.

While CLIP has encouragingly strong performance on the task of action recognition, we note that there are many differences between the models being compared beyond just their form of supervision such as model architecture, training data distribution, dataset size, and compute used. Further work is needed to more precisely determine what specific design decisions contribute to achieving high performance on this task.

	1km	25km	200km	750km	2500km
ISNs[a]	**16.9**	**43.0**	**51.9**	**66.7**	**80.2**
CPlaNet[b]	16.5	37.1	46.4	62.0	78.5
CLIP	13.9	32.9	43.0	62.0	79.3
Deep-Ret+[c]	14.4	33.3	47.7	61.6	73.4
PlaNet[d]	8.4	24.5	37.6	53.6	71.3

Table 17. Geolocalization performance on the IM2GPS test set. Metric is percent of images localized within a given radius. Models are ordered by average performance. [a](Muller-Budack et al., 2018) [b](Hongsuck Seo et al., 2018) [c](Vo et al., 2017) [c](Weyand et al., 2016)

E.4. Geolocalization

Another behavior we noticed during the development of CLIP was its ability to recognize many places and locations. To quantify this we created the Country211 dataset as described in Appendix A and report results on it throughout the paper. However it is a new benchmark so to compare with prior work on geolocalization we also report results on the IM2GPS test set from Hays & Efros (2008) in Table 17. Since IM2GPS is a regression benchmark, we guess the GPS coordinates of the nearest image in a set of reference images using CLIP's embedding space. This is not a zero-shot result since it uses nearest-neighbor regression. Despite querying only 1 million images, which is much less than prior work, CLIP performs similarly to several task specific models. It is not, however, competitive with the current state of the art.

E.5. Robustness to Distribution Shift

Section 3.3 provides a high level summary and analysis of ImageNet-related robustness results. We briefly provide some additional numerical details in this appendix. Performance results per dataset are provided in Table 16 and compared with the current state of the art results reported in Taori et al. (2020)'s evaluation suite. Zero-shot CLIP improves the state of the art on 5 of the 7 datasets, ImageNet-R, ObjectNet, ImageNet-Sketch, ImageNet-Vid, and Youtube-BB. CLIP's improvements are largest on ImageNet-Vid and Youtube-BB due to its flexible zero-shot capability and on ImageNet-R, which likely reflects CLIP's pre-training distribution including significant amounts of creative content. A similar behavior has been documented for the Instagram pre-trained ResNeXt models as discussed in Taori et al. (2020).

	IN Top-1	IN-V2 Top-1	IN-A Top-1	IN-R Top-1	ObjectNet Top-1	IN-Sketch Top-1	IN-Vid PM0	IN-Vid PM10	YTBB PM0	YTBB PM10
NS EfficientNet-L2[a]	**88.3**	**80.2**	**84.9**	74.7	68.5	47.6	88.0	82.1	67.7	63.5
FixResNeXt101-32x48d V2[b]	86.4	78.0	68.4	80.0	57.8	59.1	85.8	72.2	68.9	57.7
Linear Probe CLIP	85.4	75.9	75.3	84.2	66.2	57.4	89.1	77.2	68.7	63.1
Zero-Shot CLIP	76.2	70.1	77.2	**88.9**	**72.3**	**60.2**	**95.3**	**89.2**	**95.2**	**88.5**

表16. 详细的ImageNet鲁棒性性能。IN用于缩写ImageNet。[a](Xie等人，2020年) [b](Touvron等人，2019年)

尽管存在这一局限，CLIP特征在此任务上的迁移表现却出人意料地出色。在线性探测评估设置中，CLIP在UCF-101数据集上达到了先前最佳结果，并在我们的评估体系中超越了所有其他模型。在Kinetics-700数据集上，CLIP同样超越了原论文中经过微调的I3D基线模型。由于无需训练阶段，我们通过平均所有帧的预测结果来报告CLIP的零样本性能。在此设置下CLIP表现优异，其在Kinetics-700上的性能与经过54.5万条标注视频全监督训练的I3D基线仅相差1%。受这些结果的鼓舞，我们还测量了CLIP在近期推出的RareAct数据集（Miech等人，2020a）上的表现，该数据集专为评估非常规动作（如“锤击手机”“钻鸡蛋”）的零样本识别能力而设计。CLIP较先前基于1亿条教学视频自动提取字幕训练的S3D模型提升了10个百分点，刷新了该领域的最佳性能纪录。

尽管CLIP在动作识别任务上表现出令人鼓舞的强大性能，但我们注意到，除了监督形式之外，被比较的模型之间还存在许多差异，例如模型架构、训练数据分布、数据集大小和使用的计算资源。需要进一步的工作来更精确地确定哪些具体的设计决策有助于在此任务上实现高性能。

	1km	25km	200km	750km	2500km
ISNs[a]	**16.9**	**43.0**	**51.9**	**66.7**	**80.2**
CPlaNet[b]	16.5	37.1	46.4	62.0	78.5
CLIP	13.9	32.9	43.0	62.0	79.3
Deep-Ret+[c]	14.4	33.3	47.7	61.6	73.4
PlaNet[d]	8.4	24.5	37.6	53.6	71.3

表 17. IM2GPS 测试集上的地理定位性能。指标为定位在给定半径内的图像百分比。模型按平均性能排序。[a](Muller-Budack 等人，2018) [b](Hongsuck Seo 等人，2018) [c](Vo 等人，2017) [c](Weyand 等人，2016)

E.4. 地理定位

在开发CLIP过程中，我们注意到的另一个现象是它能够识别许多地点与位置。为量化这一能力，我们按照附录A所述创建了Country211数据集，并在全文中报告其评估结果。由于这是一个新基准，为与先前地理定位研究进行对比，我们同时在表17中报告了Hays & Efros（2008）提出的IM2GPS测试集结果。鉴于IM2GPS属于回归任务基准，我们利用CLIP的嵌入空间在参考图像集中通过最近邻匹配推测目标图像的GPS坐标。该方法采用最近邻回归策略，因此不属于零样本学习范畴。尽管仅查询了100万张图像（远少于先前研究），CLIP的表现仍与多个专用模型相当，但尚未达到当前最优技术水平。

E.5. 对分布偏移的鲁棒性

第3.3节对ImageNet相关的鲁棒性结果进行了高层级的总结与分析。本附录将简要补充部分数值细节。各数据集的性能结果见表16，并与Taori等人（2020）评估套件中报告的最新前沿成果进行了对比。零样本CLIP在7个数据集中的5个上实现了性能突破，包括ImageNet-R、ObjectNet、ImageNet-Sketch、ImageNet-Vid和Youtube-BB。CLIP在ImageNet-Vid和Youtube-BB上的提升最为显著，这得益于其灵活的零样本能力；而在ImageNet-R上的进步，可能源于CLIP预训练数据集中包含了大量创意内容。类似现象在Taori等人（2020）讨论的Instagram预训练ResNeXt模型中也曾有记载。

F. Model Hyperparameters

Hyperparameter	Value
Batch size	32768
Vocabulary size	49408
Training epochs	32
Maximum temperature	100.0
Weight decay	0.2
Warm-up iterations	2000
Adam β_1	0.9
Adam β_2	0.999 (ResNet), 0.98 (ViT)
Adam ϵ	10^{-8} (ResNet), 10^{-6} (ViT)

Table 18. Common CLIP hyperparameters

Model	Learning rate	Embedding dimension	Input resolution	ResNet blocks	ResNet width	Text Transformer layers	Text Transformer width	Text Transformer heads
RN50	5×10^{-4}	1024	224	(3, 4, 6, 3)	2048	12	512	8
RN101	5×10^{-4}	512	224	(3, 4, 23, 3)	2048	12	512	8
RN50x4	5×10^{-4}	640	288	(4, 6, 10, 6)	2560	12	640	10
RN50x16	4×10^{-4}	768	384	(6, 8, 18, 8)	3072	12	768	12
RN50x64	3.6×10^{-4}	1024	448	(3, 15, 36, 10)	4096	12	1024	16

Table 19. CLIP-ResNet hyperparameters

Model	Learning rate	Embedding dimension	Input resolution	Vision Transformer layers	Vision Transformer width	Vision Transformer heads	Text Transformer layers	Text Transformer width	Text Transformer heads
ViT-B/32	5×10^{-4}	512	224	12	768	12	12	512	8
ViT-B/16	5×10^{-4}	512	224	12	768	12	12	512	8
ViT-L/14	4×10^{-4}	768	224	24	1024	16	12	768	12
ViT-L/14-336px	2×10^{-5}	768	336	24	1024	16	12	768	12

Table 20. CLIP-ViT hyperparameters

F. 模型超参数

Hyperparameter	Value
Batch size	32768
Vocabulary size	49408
Training epochs	32
Maximum temperature	100.0
Weight decay	0.2
Warm-up iterations	2000
Adam β_1	0.9
Adam β_2	0.999 (ResNet), 0.98 (ViT)
Adam ϵ	10^{-8} (ResNet), 10^{-6} (ViT)

表 18. 常见 CLIP 超参数

Model	Learning rate	Embedding dimension	Input resolution	ResNet blocks	ResNet width	Text Transformer layers	Text Transformer width	Text Transformer heads
RN50	5×10^{-4}	1024	224	(3, 4, 6, 3)	2048	12	512	8
RN101	5×10^{-4}	512	224	(3, 4, 23, 3)	2048	12	512	8
RN50x4	5×10^{-4}	640	288	(4, 6, 10, 6)	2560	12	640	10
RN50x16	4×10^{-4}	768	384	(6, 8, 18, 8)	3072	12	768	12
RN50x64	3.6×10^{-4}	1024	448	(3, 15, 36, 10)	4096	12	1024	16

表 19. CLIP-ResNet 超参数

Model	Learning rate	Embedding dimension	Input resolution	Vision Transformer layers	Vision Transformer width	Vision Transformer heads	Text Transformer layers	Text Transformer width	Text Transformer heads
ViT-B/32	5×10^{-4}	512	224	12	768	12	12	512	8
ViT-B/16	5×10^{-4}	512	224	12	768	12	12	512	8
ViT-L/14	4×10^{-4}	768	224	24	1024	16	12	768	12
ViT-L/14-336px	2×10^{-5}	768	336	24	1024	16	12	768	12

表 20. CLIP-ViT 超参数