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]

摘要

当前的计算机视觉（CV）模型通常被训练用于预测有限的物体类别。这种严格的监督训练方式限制了模型的泛化性和实用性，因为这样的模型通常还需要额外的标注数据来完成训练时未曾见过的视觉“概念”。直接从图片的描述文本中学习是一个有潜力的选择，因为这样我们可以获取更多的监督信号。这篇文章中，我们证明了利用一个简单的预训练任务（即预测哪个文本描述对应当前图像）在一个从互联网上搜集的 4 亿个（图像，文本）对的数据集上可以取得 SOTA 的图像表征。预训练完之后，在下游任务上，我们可以通过用自然语言（文本）匹配视觉概念（图像）从而实现零次学习转移。我们在 30 个不同类型的下游 CV 任务上进行了基准测试，并展示了我们模型强大的迁移能力，其在很多下游任务上不需要任何额外的数据也能比拟完全监督的模型。比如，我们的模型在 ImageNet 上的零次学习准确率能达到在 ImageNet 上全监督训练的 ResNet-50 的性能。

关键词：多模态；无监督

1 引言

在 NLP 中，预训练的方法目前其实已经被验证很成功了，像 BERT 和 GPT 系列之类的。其中，GPT-3 从网上搜集了 4 亿对数据进行预训练然后可以在很多下游任务上实现 SOTA 性能和无监督学习。这其实说明从网络上的的数据中学习是可以超过高质量的人工标注的 NLP 数据集的。然而，对于 CV 领域，目前预训练模型基本都是基于人工标注的 ImageNet 数据集（含有 1400 多万张图像），那么借鉴 NLP 领域的 GPT-3 从网上搜集大量数据的思路，我们能不能也从网上搜集大量图像数据用于训练视觉表征模型呢？作者先是回顾了并总结了和上述相关的两条表征学习路线：（1）构建图像和文本的联系，比如利用已有的图像文本对数据集，从文本中学习图像的表征；（2）获取更多的数据（不要求高质量，也不要求全部有标签）然后做弱监督预训练，就像谷歌使用的 JFT-300M 数据集进行预训练一样（在 JFT 数据集中，类别标签是有噪声的）。具体来说，JFT 中一共有 18291 个类别，这能教模型的概念比 ImageNet 的 1000 类要多得多，但尽管已经有上万类了，其最后的分类器其实还是静态的、有限的，因为你最后还是得固定到 18291 个类别上进行分类，那么这样的类别限制还是限制了模型的零次学习能力。这两条路线其实都展现了相当的潜力，前者证明图像文本对可以用来训练视觉表征，后者证明扩充数据能极大提升性能，即使数据有噪声。于是从上层出发，作者考虑从网上爬取大量的图像文本对以扩充数据，同时这样的数据可以用来训练视觉表征的。作者随即在互联网上采集了 4 亿个图像文本对，准备开始训练模型。

2 相关工作

在过去的几年里，直接从原始文本中学习的预训练方法已经彻底改变了 NLP[1]。任务不可知的目标，如自回归和屏蔽语言建模，已经在计算、模型容量和数据方面扩展了许多数量级，能力稳步提

高。“文本到文本”作为标准化输入输出接口的发展使任务不可知架构能够零次学习迁移运用到到下游数据集，从而无需专门的输出头或特定于数据集的定制[2]。像 GPT-3 这样的系统现在在许多定制模型任务中具有竞争力，同时几乎不需要特定于数据集的训练数据[3]。

这些结果表明，在网络规模的文本集合中，现代预训练方法, 可获得的聚合监督超过了高质量的人群标记 NLP 数据集。然而，在计算机视觉等其他领域，在 ImageNet 等人群标记的数据集上预训练模型仍然是标准做法[4]。直接从网络文本中学习的可扩展预训练方法能否在计算机视觉领域取得类似的突破? 之前的工作令人鼓舞。

3 本文方法

3.1 本文方法概述

图 1 所示为 CLIP 模型的概述图。传统的图像分类模型同过共同训练特征提取器以及线性分类器来执行图像分类的预测任务。CLIP 模型通过对两个编码器（文本编码器以及图像编码器）来实现预测一个图像和一个文本的匹配性。在进行预测的时候，通过将目标数据集的描述词嵌入到句子中，再将这个句子通过文本编码器来构成零次预测的一部分结构。

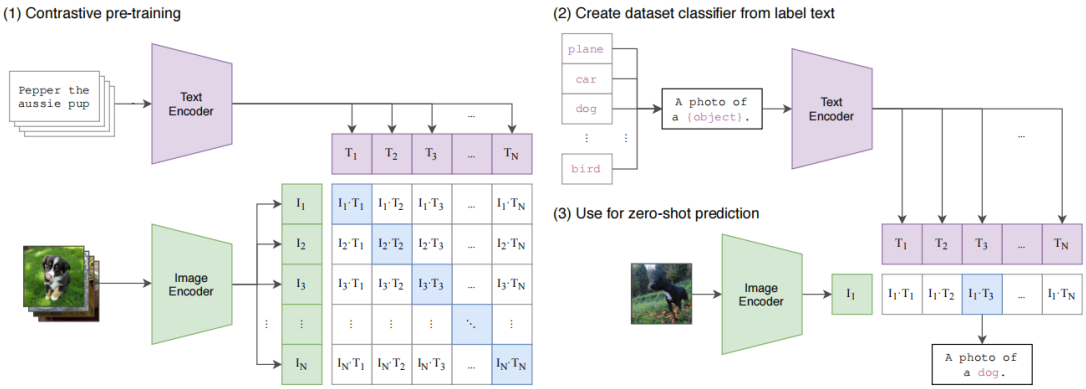

图 1: CLIP 模型示意图

3.2 自然语言监督

本模型的方法核心是从自然语言中包含的监督学习感知的思想。与其他训练方法相比，从自然语言中学习有几个潜在的优势。与用于图像分类的标准众包标签相比，扩展自然语言监督要容易得多，因为它不需要注释是经典的“机器学习兼容格式”。相反，处理自然语言的方法可以被动地从互联网上大量文本中的监督中学习。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的优势，那就是它不仅“只是”学习一种表示，而且还将该表示与语言联系起来，从而实现灵活的零次学习。在下面的小节中，我们将详细介绍我们确定的具体方法。

3.3 数据集建立

自然语言监管的一个主要动机是大量文本数据可以在互联网上公开获得。我们构建了一个新的数据集，从互联网上各种公开可用的资源中收集了 4 亿 (图像，文本) 对。为了尝试覆盖尽可能广泛的视觉概念集，我们搜索 (图像，文本) 对作为构建过程的一部分，其文本包含 500,000 个查询集中的一个。通过每个查询包含多达 20,000 对 (图像，文本) 对来平衡结果。结果数据集的总字数与用于训练 GPT-2 的 WebText 数据集相似。

3.4 选择模型及训练

对于图像编码器，作者准备了两种选择，ResNet-50 和 Vision Transformer（Vit）。对于文本编码器，作者选择了 Transformer。作者训练了 5 个 ResNets 和 3 个 Vision transformer。对于 Resnet，训练了一个 ResNet-50，一个 ResNet-101，然后还有 3 个遵循高效网络风格的模型缩放，并使用大约 4 倍，16 倍和 64 倍的 ResNet-50 计算。它们分别表示为 RN50x4、RN50x16 和 RN50x64。对于 Tranformer，训练了一个 vitb /32，一个 vitb /16 和一个 vitl /14。

如图 2 所示，作者给出 CLIP 实现核心的伪代码。给定一批 N(图像，文本) 对，CLIP 被训练来预测一批中 N × N 个可能的 (图像，文本) 对中哪一个实际发生。为了做到这一点，CLIP 学习了具有高点互信息的内容，通过联合训练图像编码器和文本编码器来最大化批处理中 N 个实对的图像和文本嵌入的余弦相似度，同时最小化 N 个 2−N 个错误对的嵌入的余弦相似度。作者优化了这些相似度分数上的对称交叉熵损失。

```
# 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 2: CLIP 核心伪代码

4 复现细节

本人复现了 CLIP 模型的零次学习预测以及线性预测这两部分功能。零次学习预测是指直接将图片输入预训练好的 CLIP 模型。线性探测是指，保持 CLIP 模型的权重不变，将它当作一个特征提取器，再在后面加 softmax 层，我们只训练 softmax 层。

4.1 实验环境搭建

执行下列指令：

$ conda install –yes -c pytorch pytorch=1.7.1 torchvision cudatoolkit=11.0

$ pip install ftfy regex tqdm

$ pip install git+https://github.com/openai/CLIP.git

4.2 界面分析与使用说明

Zero-shot.py 和 Linear.py 分别是零次学习和线性探测的实现代码，默认当前目录下 1.jpeg 时预测的图像，CIFAR100 为线性探测的数据集。直接执行 Zero-shot.py 和 Linear.py 即可运行。

5 实验结果分析

如图 3 所示为零次学习实验结果。我们保证模型未经训练同类图像，直接测试该图片。与传统图像分类模型不同，该模型和的输入有两种变量组成：图像和长句，所以，当一张图片确定时，概率分布会随着长句的变化而变化。我们可以看到，CLIP 的泛化能力很强，他可以准确识别出图形的种类、颜色、相关节日与情感，当一个长句中匹配的标签越多时，模型也认为这句话更加匹配。这也证明了自然语言监督可以学习多种表示，而且将图像示与语言联系起来。

图 3: 零次学习实验结果示意

如图 4 所示为线性探测程序结果。数据集选择的是 CIFAR100。我们可以看到，其准确率为 80%，这高于 CLIP 零次学习的预测结果，同时也高于 ResNet-50 有监督学习下的结果。这证明了，在某些数据集下，线性探测的方法可以进一步提升预测准确性，弥补零次学习细粒度不足的问题。

```
At iterate 1000    f=  2.80827D+04    |proj g|=  1.33646D+00

           * * *

Tit   = total number of iterations
Tnf   = total number of function evaluations
Tnint = total number of segments explored during Cauchy searches
Skip  = number of BFGS updates skipped
Nact  = number of active bounds at final generalized Cauchy point
Projg = norm of the final projected gradient
F     = final function value

           * * *

   N    Tit     Tnf  Tnint  Skip  Nact     Projg        F
51300   1000   1034      1     0     0   1.336D+00   2.808D+04
  F =   28082.714728929379

STOP: TOTAL NO. of ITERATIONS REACHED LIMIT
Accuracy = 80.000

进程已结束,退出代码0
```

图 4: 线性探测实验结果示意

6 总结与展望

本文提出了一个自然语言监督的、泛化能力极强的预训练模型，在零次学习模式下，便可在大部分数据集中变现的比有监督训练的 ResNet-50 模型效果更好。

然而，我们看到，零次学习 CLIP 在一些专业、复杂或抽象的任务上相当弱，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)、与自动驾驶相关的任务，如德国交通标志识别 (GTSRB)、识别到最近汽车的距离 (KITTI distance)。这些结果说明了零次学习 CLIP 在处理复杂任务时能力较差。CLIP 模型仍有很大的改进空间。

参考文献

[1] A.M. D, Q.V. L. Advances in neural information processing systems[J]. Semi-supervised sequence learning, 2015: 3079-3087.

[2] B. M, N.S. K, C. X, et al. The natural language decathlon: Multitask learning as question answering[J]. arXiv preprint arXiv, 2018, 1806.08730.

[3] B. B T, B. M, N. R, et al. Language models are few-shot learners[J]. arXiv preprint arXiv, 2020, 2005.14165.

[4] J. D, W. D, R. S, et al. Language models are few-shot learners[J]. CVPR09, 2009.