**Title:** Learning Transferable Visual Models From Natural Language Supervision

**Abstract:**

当前CV模型训练用来预测一组预定的物体类别。这种严格的监督训练方式限制了模型的泛化性和可用性，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有前途的替代方案，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zero-shot transfer。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supervised的模型。比如，我们的模型在ImageNet上的zero-shot accuracy能达到在ImageNet上全监督训练的ResNet-50的性能。

**Introduction and Motivating Work：**

作者介绍了NLP领域的原始数据预训练的进展：像GPT-3这样的旗舰模型在许多定制数据集上都有很强的竞争力，GPT-3从网上搜集了400 billion byte-pair-encoded tokens进行预训练然后可以在很多下游任务上实现SOTA性能和zero-shot learning，这其实说明从web-scale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。

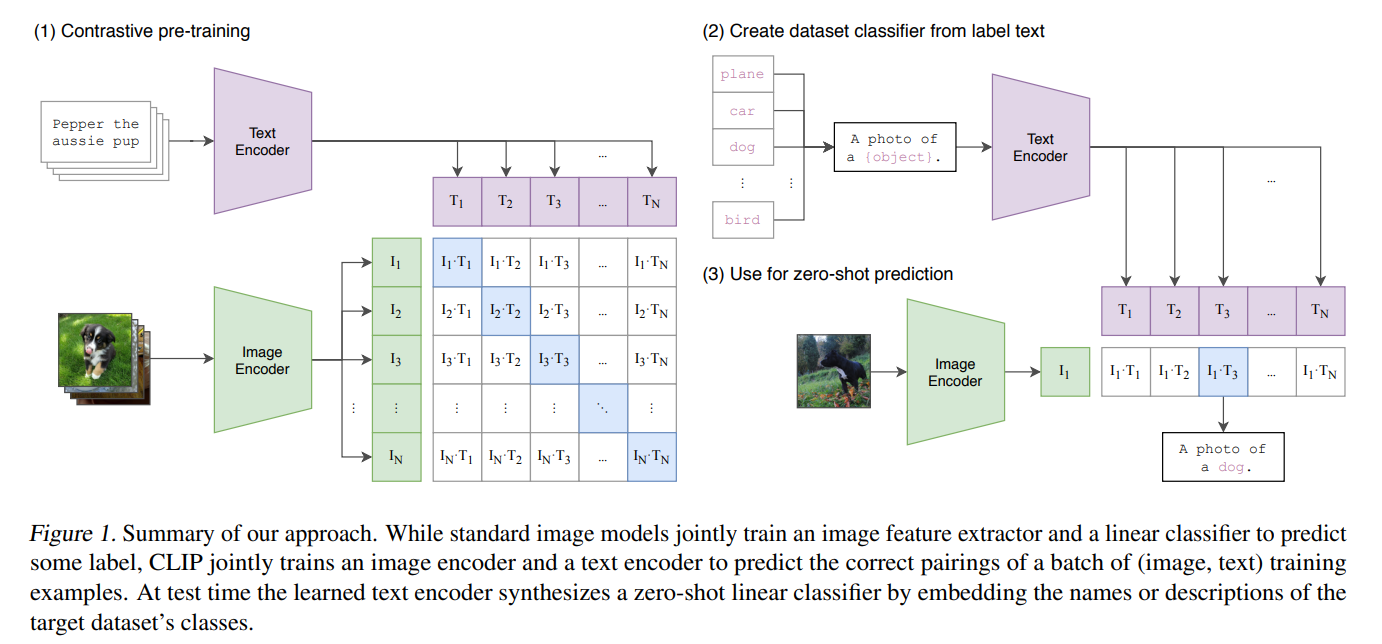
然而，在CV领域仍然还是基于人工标注的ImageNet数据集，从NLP中直接学习的可扩展预训练方法能否在计算机视觉领域带来类似的突破？

作者列举了一系列text-image，构建了image和text的联系，说明了一些模型证明具有迁移学习的能力

用自然语言监督进行图像表征学习的工作的证明很少，但是很有潜力，如Mahajan等人用Instagram上具有ImageNet标签的图片做预训练，再去ImageNet上fine-tune，比基准准确度高5%，或用JFT-300M做预训练，它有18291个类别, 这两条路线其实都展现了相当的潜力，前者证明paired text-image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。

现在实用的立场是在全监督训练和无限文本数据训练中取妥协，但他们都使用静态的softmax做预测输出，这大大降低了他们的灵活性和zero-shot能力。

作者大量研究了使用自然语言模型训练图像分类器的行为，并扩充了数据集（从网上的公开数据）。创建了4亿个图像-文本对，用来训练CLIP模型（ConVIRT的简化版本）。我们学习了CLIP的可扩充性，分为8个模型，相差2个计算数量级，可迁移性基本是一条平滑的线。CLIP和GPT家族类似，作者在超过30个数据集上测量了它的zero-shot迁移表现，发现其对比定制模型有竞争力。在ImageNet上表现为最好的公开模型且计算效率更高，在等精度下鲁棒性更好。

**Model-Training：**

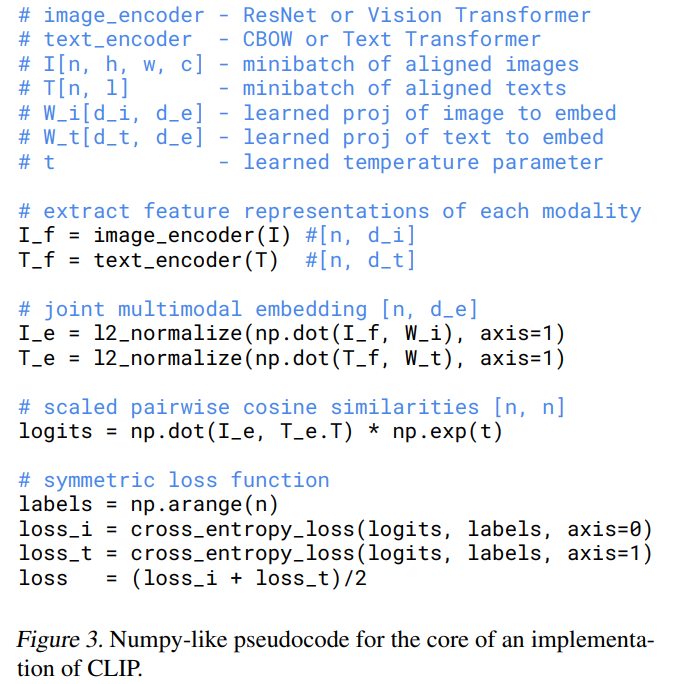
本文方法的总结。

1）预训练过程中，将text encoder 和image encoder的输出结果进行点乘，然后算loss进行训练

2）从Sample得到N个输入，都经过text encoder编码得到对应的N个class text embedding

3）要预测的图片经过Image encoder得到image embedding，然后与N个class text embedding做点乘，相似度Top-1的为预测类

伪代码如下：



编码器输出

L2正则

拿到输出

算loss

关于模型：

基于ConVIRT，CLIP主要做出了以下简化：

ConVIRT中的image encoder的参数是ImageNet初始化的，而CLIP直接用random初始化；

ConVIRT的projection head是non-linear的，而CLIP采用linear的projection；

CLIP去掉了ConVIRT中text transformation（指均匀从text中采样句子），因为CLIP数据集中有很多只出现过一次的（image，text）；

CLIP的image transformation只用了resize和squared crop；

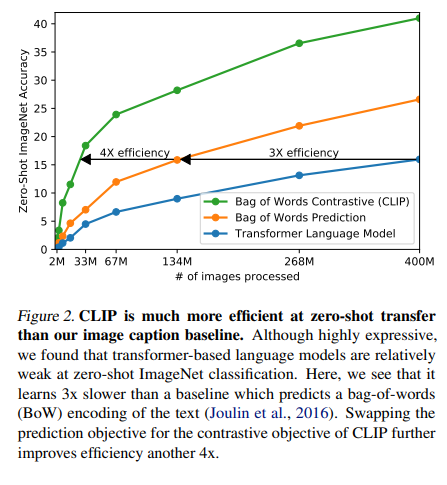
CLIP loss中的temperature参数τ是可学的。

Creating a Sufficiently Large Dataset：

We constructed a new dataset of 400 million (image, text) pairs collected form a variety of publicly available sources on the Internet.

构造了足够大的数据集，和GPT-2使用的数据集有相似的总字数，把它称为WIT（WebImage Text）

Selecting an Efficient Pre-Training Method：

训练效率是成功扩展自然语言监督的关键

最初的方法类似VirTex，使用CNN处理图像，Transformer处理语言，效率如Figure2所示

预测每个图片的确切文本是艰巨的任务

改变策略，只预测文本与哪个图像配对，而不是预测图像对应的确切单词

训练效率对比如图，模型复杂度降低，训练效率提升

Choosing and Scaling a Model – Training：

1）text encoder

作者统一采用GPT-2里的Transformer结构；对于base size model，使用63M-parameter 12-layer 512-width model with 8 attention heads；model width则随着image encoder的size增加而增加。输入句子的最大长度为76。

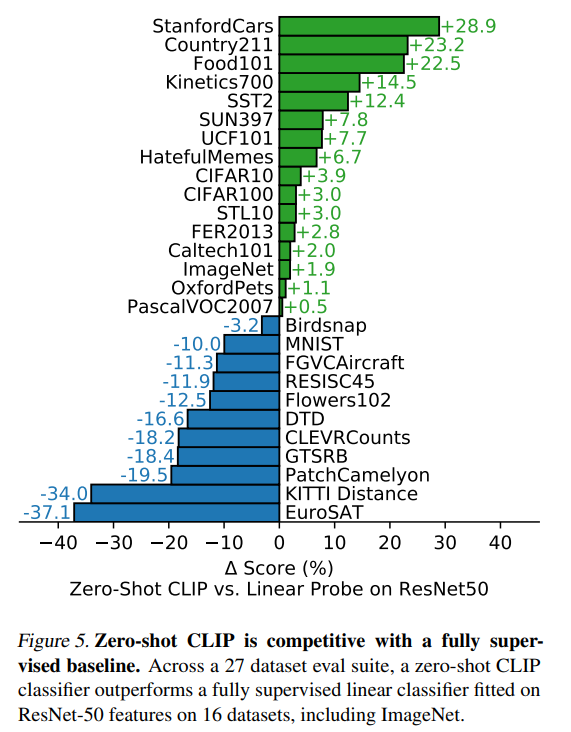
2）image encoder

这里作者一共训练了8个不同的image encoder（5 ResNets & 3 ViTs）

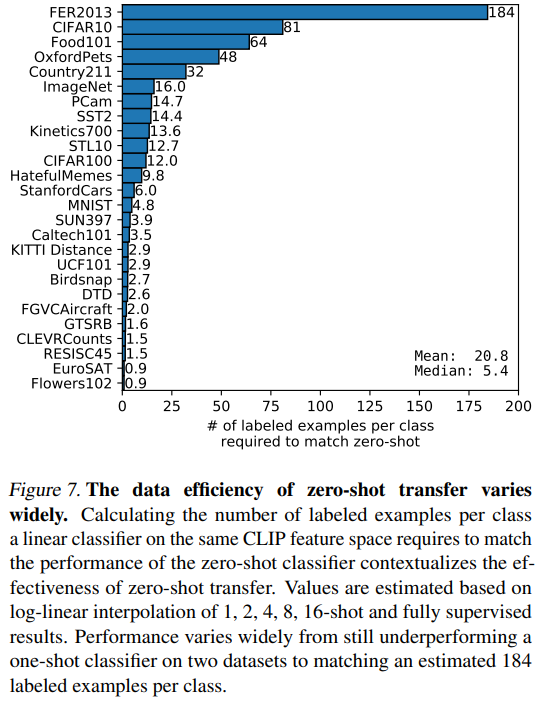
ResNet50、ResNet101、ResNet50x4、ResNet50x16、ResNet50x64、ViT-B/32、ViT-B/16、ViT-L/14

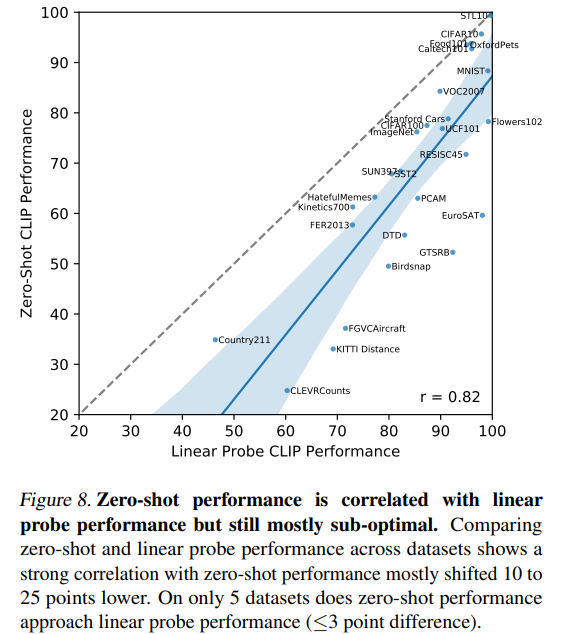
其中ResNets做了一个小的修改：将ResNet编码出来的结果再经过一个attention pooling；对于ViTs也做了一个小的修改：在tokens（patch tokens和pos tokens相加）被送到Transformer之前，让tokens先经过一个layer norm层，此外参数的初始化和原来的ViTs也有微小的不同。

**Experiment：**

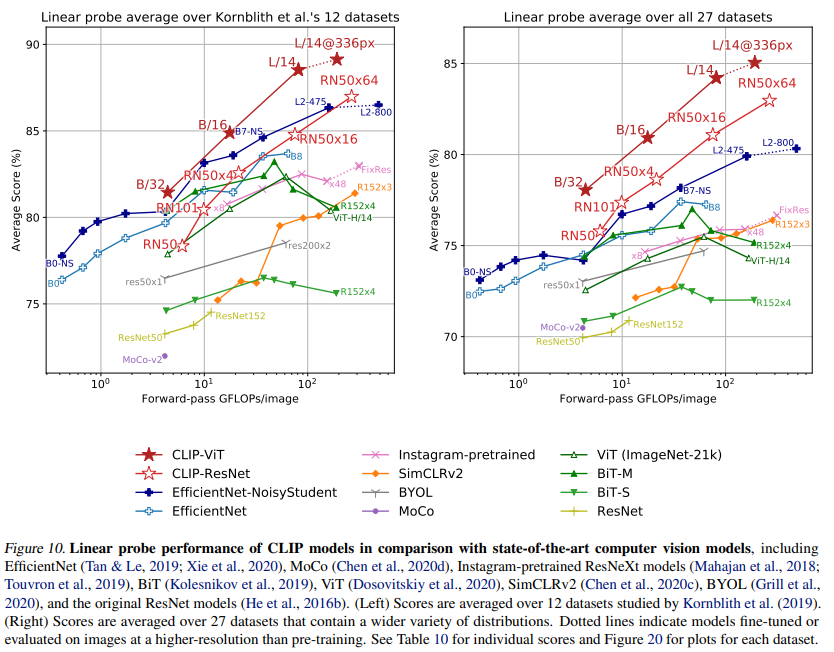


零样本CLIP与完全监督的基线相比具有竞争力。在27个数据集评估套件中，零样本CLIP分类器在16个数据集上（包括ImageNet）的性能优于在ResNet-50特征上完全监督的线性分类器。

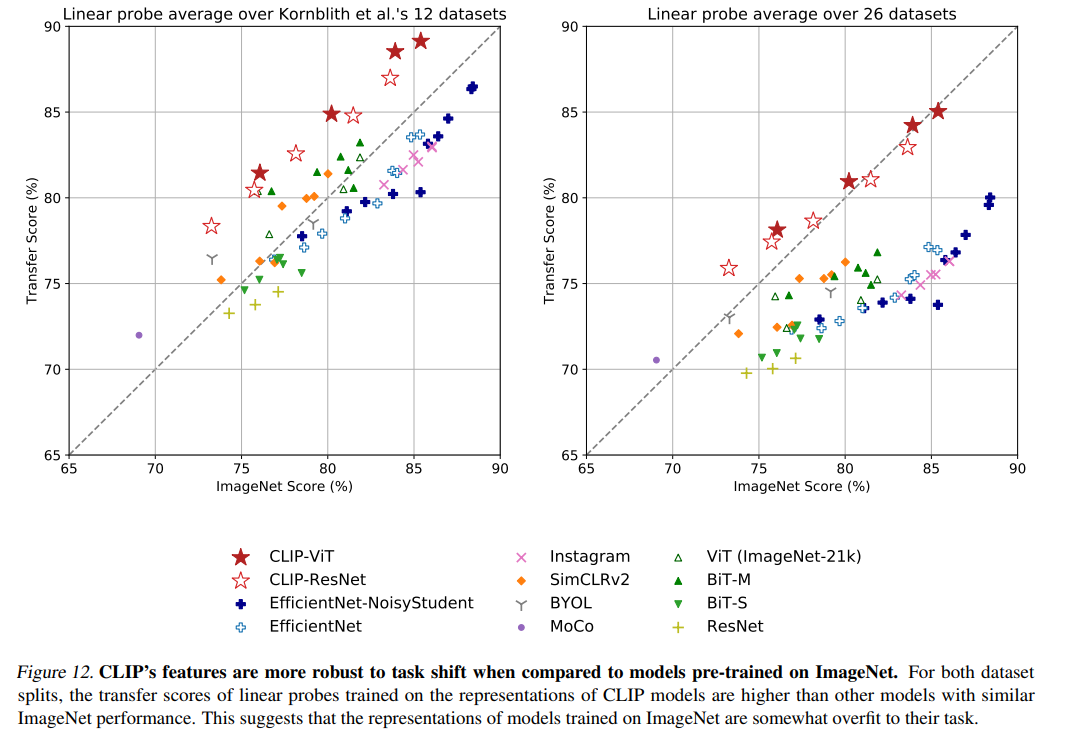
Few-shot (linear probing) CLIP （保持CLIP encoder 参数fixed，加一层逻辑回归分类器微调）平均需要20.8-shots才能match zero-shot CLIP性能。这里相当于保持了the same CLIP feature space上，观察few-shot finetuning和zero-shot的性能差异。这里其实说明通过自然语言学到的视觉概念比少量样本finetune学到的好。

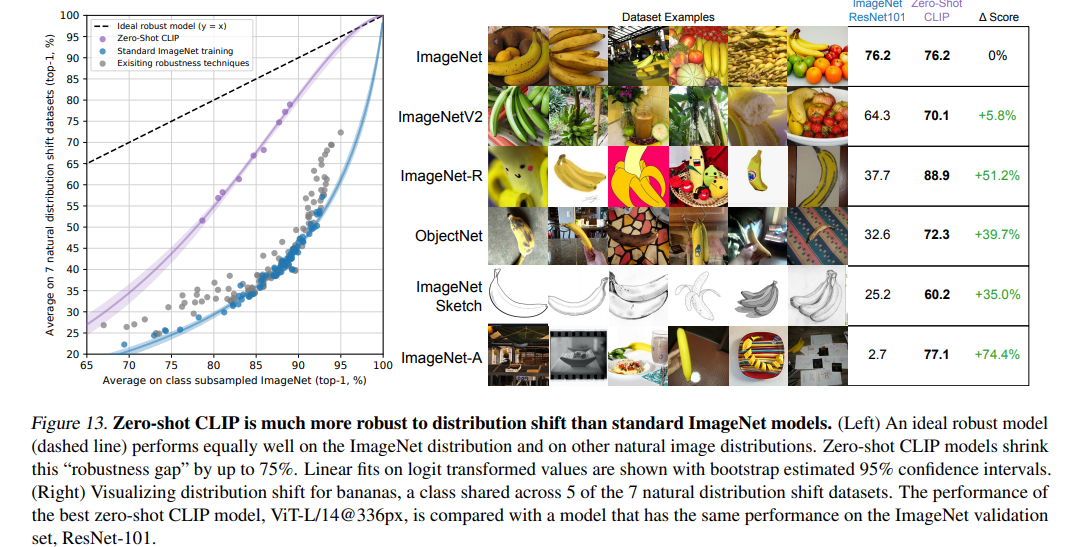


各样本zero-shot的对比



上图是模型训练过程中Linear probe的值，可以看到CLIP表现很好，说明CLIP训练效率很高

上图说明CLIP比其他在ImageNet上训练的训练集更具有鲁棒性，其余模型多少有些过拟合

本图也是说明CLIP比标准ImageNet模型更具有鲁棒性，根据左图，CLIP更接近理想鲁棒模型，在右图实例预测，CLIP比大多数模型得分更高

Limitation:

CLIP的zero-shot性能虽然总体上比supervised baseline ResNet-50要好，但其实在很多任务上比不过SOTA methods，因此CLIP的transfer learning有待挖掘；

CLIP在这几种task上zero-shot性能不好：fine-grained分类（花的分类、车的分类之类的）、抽象的任务（如计算图中object的个数）以及预训练时没见过的task（如分出相邻车辆的距离）。BTW，在这些任务上zero-shot性能不好，不代表CLIP pretrained encoders就没用了，CLIP encoders还是能提供很强的视觉先验的；

Zero-shot CLIP在真正意义上的out-of-distribution data上性能不好，比如在OCR中；

尽管CLIP zero-shot classifier能在很广泛的任务上work，但究其本质CLIP还是在有限的类别中进行对比、推理，而不能像image caption那样完全的flexible地生成新的概念（如：词），这是CLIP功能上的缺陷，CLIP终究不是生成模型；

CLIP仍然没有解决深度学习poor data efficiency的问题，结合CLIP和self-training可能是一个能提高data efficiency的方向；

CLIP的方法论上也存在几个缺陷：在训练和挑选CLIP模型时，作者采用在几个数据的validation performance来做指导，这其实是不准确的，因为它不能完全代表CLIP的zero-shot性能。如果，设计一套框架来evaluate zero-shot performance对于之后的研究是很重要的；

CLIP的训练数据是从网上采集的，这些image-text pairs没有做data clear和de-bias，这可能会使模型有一些social biases；

很多视觉任务很难用text来表达，如何用更高效的few-shot learning方法优化CLIP也很重要。

**Conclusion：**

我们研究了是否可以将NLP领域中无关任务的网络预训练的成功应用到其他领域。我们发现，采用这种方法会导致计算机视觉领域出现类似的行为，并讨论了这一研究方向的社会影响。为了优化其训练目标，CLIP模型在预训练期间学习执行多种任务。然后，可以利用这种任务学习能力，通过自然语言提示实现向许多现有数据集的零样本迁移。在足够规模的情况下，这种方法的性能可以与特定任务的监督模型竞争，但仍有很多改进的空间。

-what did authors try to accomplish?

Transferability of CV models. NLP field models have good ablity to handle downstream tasks because of its pre-training in big dataset.

-what the key elements of the approach?

400 million text-image pairs be used to pre-train CLIP and effective training approach.

-what can you use yourself?

I know CV models also can use big dataset for pre-train to improve its robustness.

-what other reference do you want to follow?