惊呆！不用一张图片，却训出个图像识别SOTA?

原创 橙橙子 夕小瑶的卖萌屋 2021-04-14 17:55

文 | 橙橙子

如果老板派给你一个任务，不使用一张图片，让你训练一个视觉预训练模型，你会不会觉得老板疯了。最近有一篇论文，不仅没用一张真实图片和标注，还训练出个媲美SOTA的效果，甚至超过了MoCov2和SimCLRv2，你敢信么？今天，就让我们来看一下这篇神作！

论文题目:

Can Vision Transformers Learn without Natural Images?

论文链接:

https://arxiv.org/pdf/2103.13023.pdf

项目地址:

https://hirokatsukataoka16.github.io/Vision-Transformers-without-Natural-Images/

也可以在 **【夕小瑶的卖萌屋】**订阅号后台回复关键词 **【0414】** 下载论文PDF~

什么是不需要真实图像的ViT?

视觉Transformer（ViT）最近强势进军CV界，并取得绝佳效果，大有要取代曾经的王者卷积神经网络（CNNs）的趋势。不过，ViT也有诸多问题，在预训练阶段经常需要上亿级别的真实图像数据和标注预训练才能和CNNs一较高低，这直接带来了诸如隐私保护、标注成本、AI伦理等问题。随着自监督学习方法（Self-Supervised Learning, SSL）如Moco、SimCLR的成功，标注问题被极大地解决，但是在真实图像上进行训练仍然会触发诸如侵犯隐私和公平性保护的问题。譬如，正因为图像版权相关的问题，著名的ImageNet数据集只能用于非商业用途。

如果能不使用任何真实图像数据和人工标注情况下训练ViT，还能达到甚至超过真实图像训练的最优模型，数据问题荡然无存，模型轻松放心大胆用，这简直完美，岂不快哉！

公式驱动的监督学习

重点来了！通过什么方式达成这一目标呢？本文提出了一种基于公式驱动的监督学习方法（Formula-Driven Supervised Learning，FDSL）。这种方法依赖于没有自然图像的数据库，即分形数据库（FractalDB）。通过分配分形来自动生成图像模式及其类别标签，这些分形基于现实世界背景知识中存在的自然规律。FractalDB最早被提出于[1]，值得一提的是，这篇文章同样也是本文作者所写，并获得了ACCV 2020最佳论文提名奖。

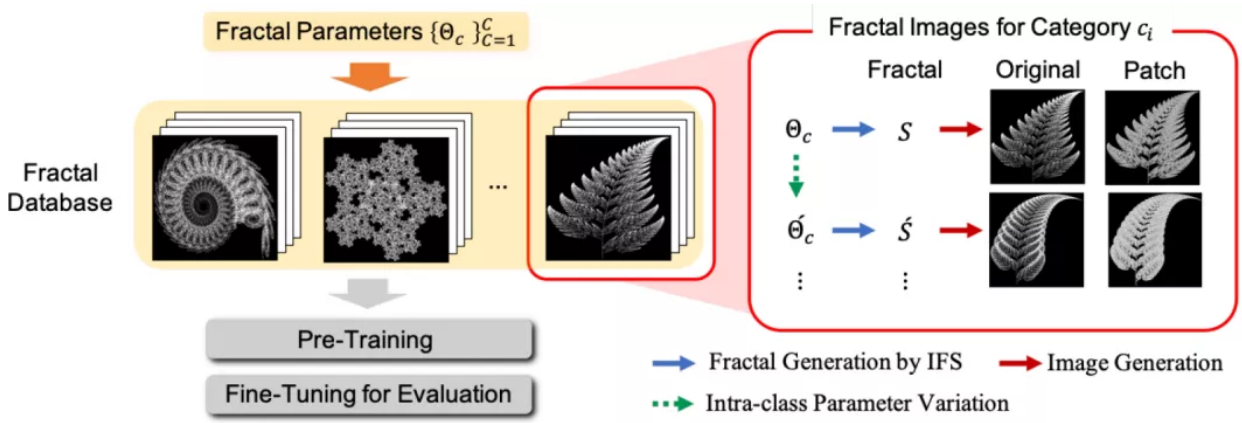
FractalDB的构造过程可以分为两步：

（1）使用迭代函数系统（iterated function system，IFS）自动生成基础分形和对应的类别。熟悉计算机图形学的同学们会比较熟悉，使用IFS生成分形分为选定随机起始像素点x_i、随机生成N种仿射变换$\Theta_i = \{(\theta_j, p_j)\}_{j=1}^{N}$（$\theta_j$包含6个参数：4个旋转参数和2个平移参数；$p_j$表示采样概率）、依据概率分布对当前点采样变换函数生成新的描绘点、重复迭代这个过程直至达到设定像素点阈值这几个过程。最终的分形由这些像素点绘制而成，由于它由Θ_i确定，所以对应的类别就是Θ_i。这种方法能够保证只使用简单的公式就可以生成接近自然物体的复杂模式。

（2）对基础分形做扩展，得到当前类下的不同样例（intra-category instances）。这个步骤的目的是为了扩充数据。类别内扩充的基本原则是在保持分形基本形状不变的情况下，尽可能增加多样性。论文提出了三种方式：a. 对IFS的6个参数进行一定weight缩放：预设了4种weight，可以产生25种（$1 + 6 * 4$）不同的变种。b.旋转：包括不旋转、水平旋转、垂直旋转、水平-垂直共4种。c. 块渲染：基础分形使用了$1 * 1$的像素渲染，为了制造差异性，块渲染使用10种$3 * 3$的像素块。这样，对于每一种类别，我们可以构造出1000（$25 * 4 * 10$）个样例。

最终，FractalDB含有两种不同的规模。FractalDB-1K含有1k类别，共计1M样例。FractalDB-10k含有10k类别，共计10M样例。

下图展示了分形数据库的构造过程：



分形数据库联合ViT

FractalDB可以直接应用在ViT上么？答案是肯定的，不过本文也针对ViT的特点做了一些使用方式上的修改。首先，真实图像是彩色图，而分形没有背景，是灰度图。为了让模型学到一些色彩的分布，论文

对FractalDB进行了色彩增强，即在渲染时随机使用颜色像素。进一步，参考自监督学习的成功经验，论文进行了更长时间的充分训练。

好了，数据已ready，剩下的就交给强大的ViT了！这里，论文使用了DeiT (Data-Efficient Image Transformers)[2]. 在FractalDB上训练ViT和在真实图像上训练方法一样，将2D图像$x \in \mathbb{R}^{H \times W \times C}$拆分成$P * P$大小的多个patch，并平铺在一起组成多个visual token的1D输入$x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$，然后开心快乐的feed到Transormer中训练就好啦～

呼唤实验效果

实验是检验真理的唯一标准，效果好不好，结果看一下。论文使用了经典的pretrain-finetune方法，首先在FractalDB上预训练的DeiT，接着在各个视觉下游任务数据集上微调。

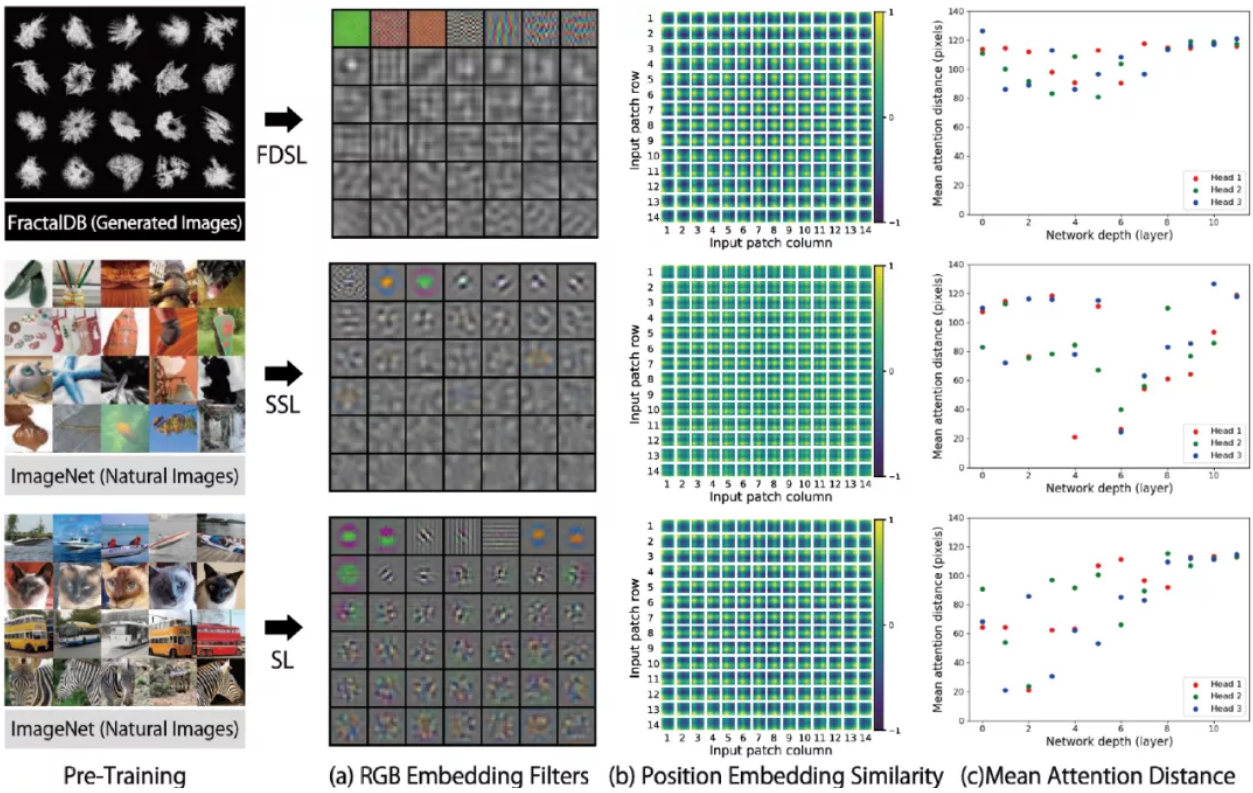
首先和多种有监督方法进行了效果对比。尽管论文方法没有完全超过在在Imagenet-1k（1.28M）上训练的效果，但是已经非常接近了。这可是完全一张真实图像都没有用啊喂! 另外我们可以看到，使用预训练和不使用，效果差距是非常明显的。

PT	PT Img	PT Type	C10	C100	Cars	Flowers	VOC12	P30	IN100
Scratch	–	–	78.3	57.7	11.6	77.1	64.8	75.7	73.2
Places-30	Natural	Supervision	95.2	78.5	69.4	96.7	77.6	–	86.5
Places-365	Natural	Supervision	**97.6**	**83.9**	**89.2**	**99.3**	84.6	–	**89.4**
ImageNet-100	Natural	Supervision	94.7	77.8	67.4	97.2	78.8	78.1	–
ImageNet-1k	Natural	Supervision	**98.0**	**85.5**	**89.9**	**99.4**	**88.7**	**80.0**	–
FractalDB-1k	Formula	Formula-supervision	96.8	81.6	86.0	98.3	84.5	78.0	87.3
FractalDB-10k	Formula	Formula-supervision	**97.6**	83.5	87.7	98.8	**86.9**	**78.5**	**88.1**

另一方面，论文和流行的自监督学习方法进行了实力对比。论文方法的平均表现亮眼，超过了MoCov2、SimCLRv2等方法。

Method	Use Natural Images?	C10	C100	Cars	Flowers	VOC12	P30	Average
Jigsaw	YES	96.4	82.3	55.7	98.2	82.1	**80.6**	82.5
Rotation	YES	95.8	81.2	70.0	96.8	81.1	79.8	84.1
MoCov2	YES	96.9	83.2	78.0	98.5	85.3	**80.8**	87.1
SimCLRv2	YES	**97.4**	**84.1**	**84.9**	**98.9**	**86.2**	80.0	**88.5**
FractalDB-10k	NO	**97.6**	**83.5**	**87.7**	**98.8**	**86.9**	78.5	**88.8**

最后，论文也做了一些可视化分析。使用分形数据库训练的模型相对于有监督模型和自监督模型而言，过滤器的范围要更广，可以在更大的范围内获取特征。



总结一下

论文另辟蹊径的在不使用任何真实图像和标注的条件下，成功训练了一个强大的ViT模型，虽然距离现在的有监督方法还有微弱差距，但是已经超过了目前最优秀的自监督模型MoCov2和SimCLRv2，是一项非常有趣的工作，相信它在AI伦理和版权保护方面有重要意义。

不过，笔者私以为，抛开数据使用问题，研究角度还是很期望看到自动构造的分形数据和真实图像数据的融合训练，说不定会有意想不到的效果呢。

参考文献

[1] Pre-training without Natural Images
(https://arxiv.org/pdf/2101.08515.pdf)

[2] Training data-efficient image transformers & distillation through attention
(https://arxiv.org/pdf/2012.12877.pdf)

喜欢此内容的人还喜欢

我不看好data2vec这类多模态融合的研究

夕小瑶的卖萌屋