A Simple Framework for Contrastive Learning of Visual Representations（*SimCLR*）



1. 摘要

本文介绍了SimCLR：一个简单的**视觉表征对比学习**的框架。我们简化了最近提出的对比性自我监督学习算法，而不需要专门的架构或内存库。为了了解是什么使对比性预测任务能够学习有用的表征，我们系统地研究了我们框架的主要组成部分。我们表明：（1）数据扩增在定义有效的预测任务中起着关键作用；（2）在表征和对比性损失之间引入可学习的非线性转换，极大地提高了学习表征的质量；（3）与监督学习相比，对比学习得益于更大的批次规模和更多的训练步骤。通过结合这些发现，我们能够在ImageNet上的自我监督和半监督学习方面大大超过以前的方法。通过SimCLR学习的自监督表征训练的线性分类器达到了76.5%的最高准确率，这比以前的技术水平提高了7%，与有监督的ResNet-50的性能相当。当只对1%的标签进行微调时，我们达到了85.8%的最高5级精度，超过了标签数量少100倍的AlexNet。

1. 问题重述

在没有人类监督的情况下学习有效的视觉表征是一个长期存在的问题。大多数主流的方法都属于两类之一：生成性或判别性。

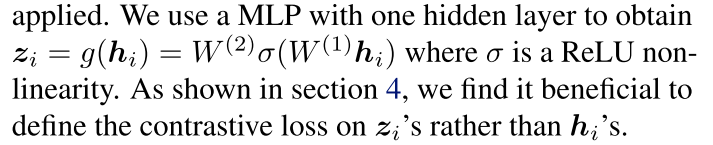
1. 生成式方法学习生成或以其他方式模拟输入空间中的像素。然而，像素级别的生成在计算上是昂贵的，对于表征学习来说可能不是必需的。
2. 鉴别性方法使用与监督学习类似的目标函数来学习表征，但训练网络来执行辅助任务，其中的输入和标签都来自于未标记的数据集。

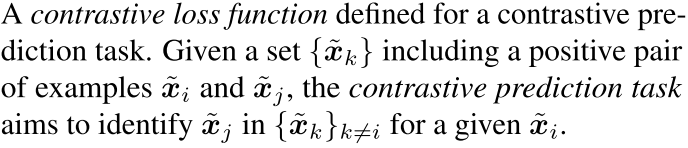
许多这样的方法依靠启发式方法来设计借口任务（Doersch等人，2015；Zhang等人，2016；Noroozi & Favaro，2016；Gidaris等人，2018），这可能会限制所学表征的通用性。最近，基于潜伏空间的对比学习的判别方法显示出巨大的前景，取得了最先进的结果（Hadsell等人，2006；Dosovitskiy等人，2014；Oord等人，2018；Bachman等人，2019）。

为了了解是什么促成了良好的对比性表达学习，我们系统地研究了我们框架的主要组成部分，并表明：

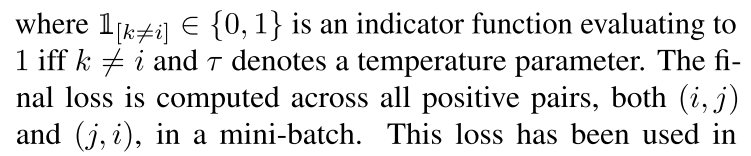
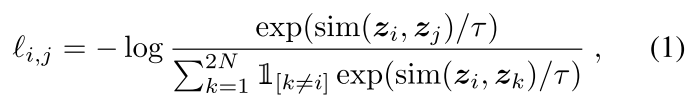
1. 在定义产生有效表征的对比性预测任务时，多种数据增强操作的组成是至关重要的。此外，无监督的对比性学习比有监督的学习更受益于强大的数据增强。
2. 在表征和对比性损失之间引入可学习的非线性转换，大大改善了学习表征的质量。
3. 具有对比性交叉熵损失的表征学习得益于规范化的嵌入和适当调整的温度参数。
4. 与有监督的学习相比，对比性学习受益于更大的批次规模和更长的训练。与监督学习一样，对比学习也得益于更深更广的网络。
5. 详细阐述
   1. 对比学习框架

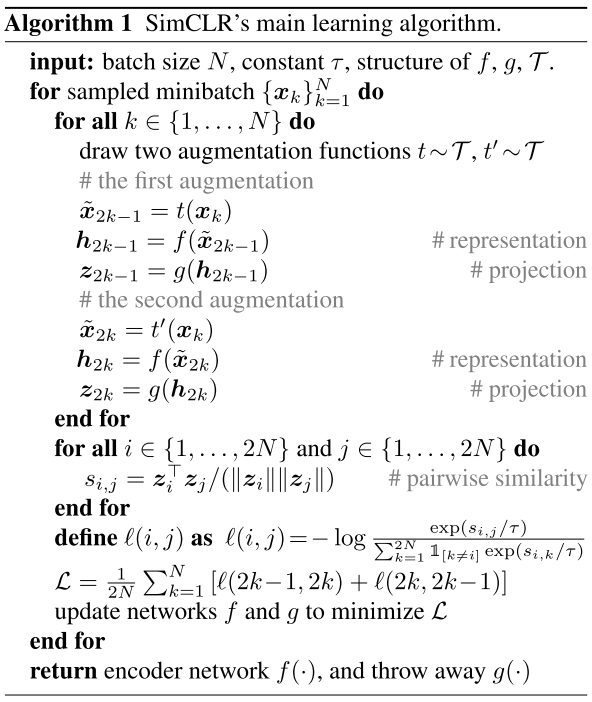
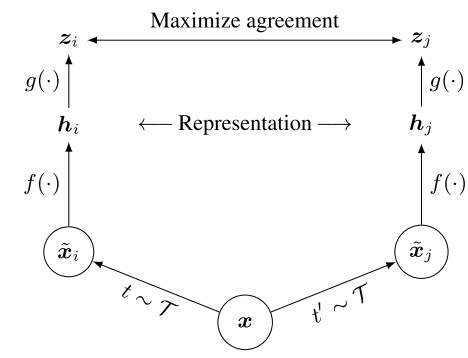
受最近的对比学习算法的启发（见第7节的概述），SimCLR通过潜在空间中的对比损失，使同一图像的不同增强视图之间的一致性最大化，从而学习表征。如图2所示，这个框架包括以下四个主要部分。

1. 一个随机的数据增强模块，对任何给定的数据例子进行随机转换，导致同一例子的两个相关视图，表示为˜x***i***和˜x***j***，我们认为这是一个正数对。在这项工作中，我们依次应用了三个简单的增强：*随机裁剪*，然后调整到原始尺寸，*随机颜色扭曲*，以及*随机高斯模糊*。如第3节所示，**随机裁剪和颜色扭曲的组合对于实现良好的性能至关重要**。
2. 一个神经网络基础编码器*f*(·)，从增强的数据实例中提取代表向量。我们的框架允许对网络结构进行各种选择，没有任何限制。我们采用常用的ResNet来获得，是平均池化层的输出。
3. 一个小的神经网络投影头*g*(·)，将表征映射到应用对比性损失的空间里。



我们随机抽取N个样本构成minibatch，并将对比性预测任务定义在来自于该样本的一对增强样本上，从而得到2N个数据点。我们不对负面的例子进行显式抽样。相反，与（Chen等人，2017）类似，我们将minibatch中的其他2（N-1）个增强的例子视为负面例子。让sim(u, v)= uTv / ||u|| ||v||去代表归一化的u和v之间的点积（即余弦相似度）。那么，一对正的例子（i，j）的损失函数定义为：





* 1. 大batch训练

为了保持简单，我们不使用记忆库来训练模型（Wu等人，2018；He等人，2019）。相反，我们将训练批次大小N从256变化到8192。8192的批次大小给了我们每个来自两个增强视图的阳性对16382个负面例子。在使用线性学习率缩放的标准SGD/Momentum时，大批量的训练可能是不稳定的（Goyal等人，2017）。为了稳定训练，我们对所有批次大小使用[LARS优化器](https://www.jianshu.com/p/e430620d3acf)（You等人，2017）。我们用云计算TPU训练我们的模型，根据批次大小，使用32到128个核心。

**全局BN**。标准的ResNets使用批量归一化（Ioffe & Szegedy，2015）。在具有数据并行性的分布式训练中，BN的平均值和方差通常在每个设备上进行本地汇总。在我们的对比学习中，由于正数对是在同一个设备中计算的，模型可以利用本地信息泄露来提高预判的准确性，而不需要改善表征。我们通过在训练期间将BN的平均值和方差汇总到所有设备上来解决这个问题。其他的方法包括在不同的设备上洗数据例子（He等人，2019年），或用层规范代替BN（Hénaff等人，2019年）。

* 1. 评估

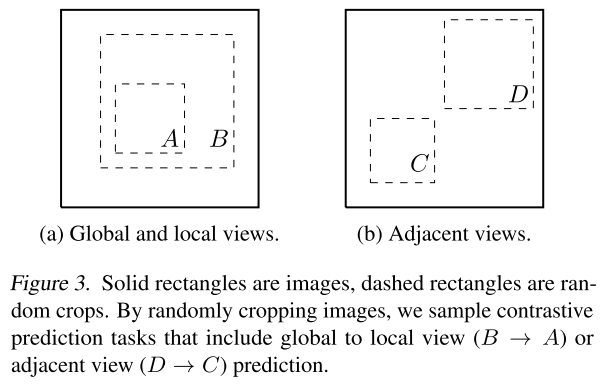
在这里，我们为我们的实证研究制定了协议，其目的是了解我们框架中的不同设计选择。

**数据集和度量**。我们对无监督的预训练（在没有标签的情况下学习编码器网络f）的大部分研究是使用ImageNet ILSVRC-2012数据集（Rus-sakovsky等，2015）完成的。

预训练（在没有标签的情况下学习编码器网络f）的大部分研究是使用ImageNet ILSVRC-2012数据集（Rus-sakovsky等人，2015）。在CIFAR-10（Krizhevsky & Hinton, 2009）上的一些额外预训练经验可以在附录B.9中找到。我们还在广泛的数据集上测试了预训练的结果，以进行迁移学习。为了评估学到的表征，我们遵循广泛使用的线性评估协议（Zhang等人，2016；Oord等人，2018；Bachman等人，2019；Kolesnikov等人，2019），在冻结的基础网之上训练一个线性分类器，并将测试准确性作为表征质量的代理。除了线性评估，我们还与半监督和转移学习的最先进技术进行比较。

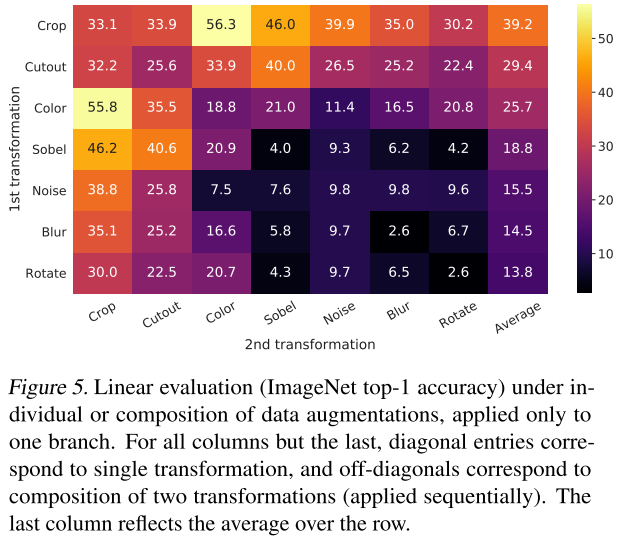
* 1. 对比性表征学习的数据增强

数据增强定义了预测性任务。虽然数据增强已被广泛用于有监督和无监督的表示学习，它还没有被认为是一种系统的方法来定义对比预测任务。许多现有的方法通过改变结构来定义对比性预测任务。例如，Hjelm等人（2018）；Bachman等人（2019）通过约束网络架构中的感受野实现了全局到局部的视图预测，而Oord等人（2018）；Hénaff等人（2019）通过固定的图像分割程序和上下文聚合网络实现了邻近视图的预测。我们表明，这种复杂性可以通过对目标图像进行简单的随机裁剪（调整大小）来避免，这创造了一个包含上述两个预测任务的系列，如图3所示。这种简单的设计选择方便地将预测任务与其他组件（如神经网络结构）相分离。更广泛的对比性预测任务可以通过扩展增量系列和随机组合来定义。



* + 1. 数据增强操作的组成对于学习好的表征至关重要

为了系统地研究数据增强的影响，我们在此考虑几种常见的增强方式。一种增强涉及数据的空间/几何变换，如裁剪和调整大小（水平翻转）、旋转（Gidaris等人，2018）和剪切（De- Vries & Taylor，2017）。另一类增强涉及外观转换，如颜色失真（包括颜色下降、亮度、对比度、饱和度、色调）（Howard，2013；Szegedy等人，2015），高斯模糊和Sobel过滤。图4直观地显示了我们在这项工作中研究的增强方法。

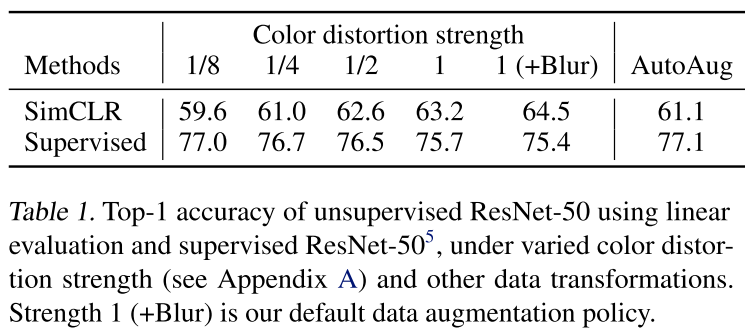


为了了解单个数据增强的效果和增强组成的重要性，我们研究了单独或成对应用增强时我们框架的性能。由于ImageNet图像的尺寸不同，我们总是应用裁剪和调整图像的大小（Krizhevsky等人，2012；Szegedy等人，2015），这使得我们很难在没有裁剪的情况下研究其他增强。为了消除这种干扰，我们考虑对这种消融进行不对称的数据转换设置。具体来说，我们总是先随机裁剪图像并将其调整到相同的分辨率，然后我们只对图2中框架的一个分支应用目标转换（s），而将另一个分支作为身份（即t（xi）=xi）。

图5显示了单个和组合变换下的线性评价结果。我们观察到，尽管模型在对比性任务中几乎可以完美地识别正数对，但没有任何一个转化足以学习到好的表示。当组成增量时，对比性预测任务变得更难，但表征的质量却大大改善。附录B.2提供了一个关于组成更广泛的增量的进一步研究。

有一种增强的组成方式很突出：随机裁剪和随机颜色失真。我们猜想，当只使用随机裁剪作为数据增强时，一个严重的问题是，一个图像中的大多数斑块都有类似的颜色分布。图6显示，仅靠颜色图谱就足以区分图像。神经网络可以利用这个捷径来解决预测任务。因此，将剪裁与颜色失真结合起来，以学习可归纳的特征是非常关键的。

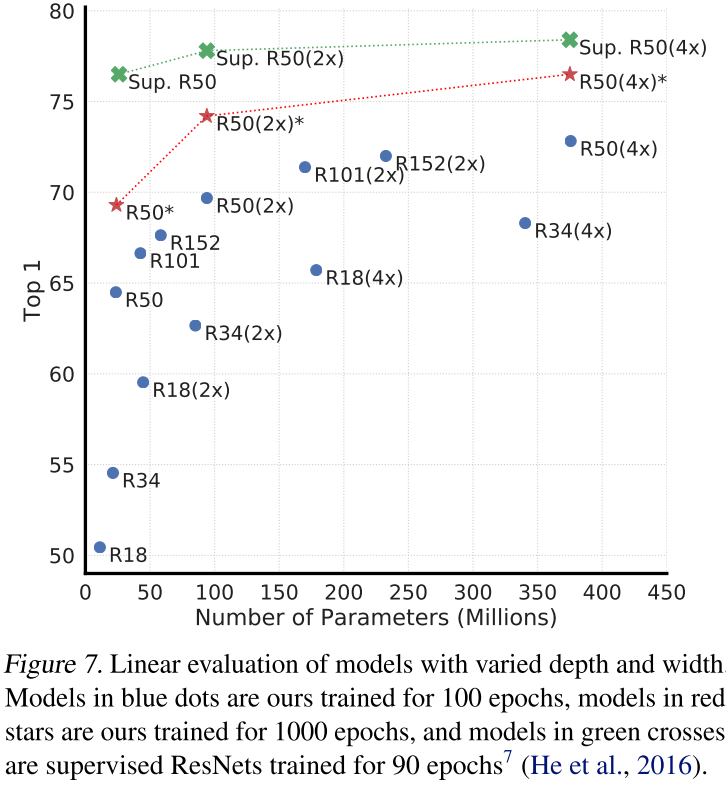
* + 1. 对比性学习需要比监督性学习更强的数据增强



为了进一步证明颜色增强的重要性，我们调整了颜色增强的强度，如表1所示。较强的颜色增强大大改善了对所学无监督模型的线性评价。在这种情况下，AutoAugment（Cubuk等人，2019年），一个使用监督学习发现的复杂的增强策略，并不比简单的裁剪+（更强的）颜色失真效果好。当用相同的增强集训练有监督的模型时，我们观察到更强的颜色增强并没有改善，甚至损害了它们的性能。因此，我们的实验表明，与监督学习相比，无监督的对比学习从更强的（颜色）数据增强中受益。尽管之前的工作报告称，数据增强对自我监督学习很有用（Doersch等人，2015；Bachman等人，2019；Hénaff等人，2019；Asano等人，2019），但我们表明，对监督学习没有产生准确性好处的数据增强仍然可以对对比学习有很大帮助

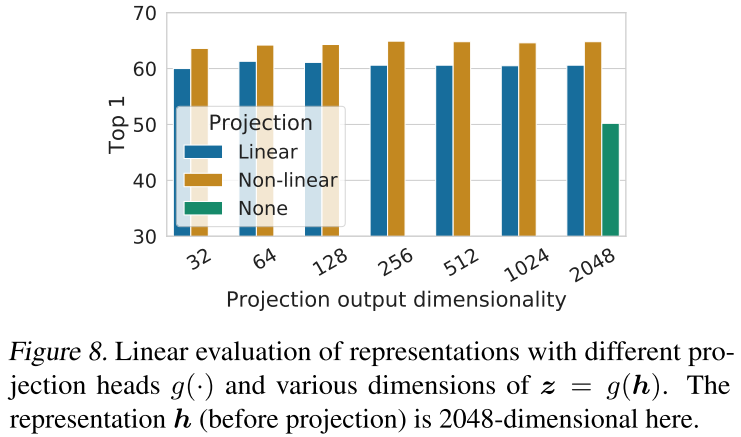
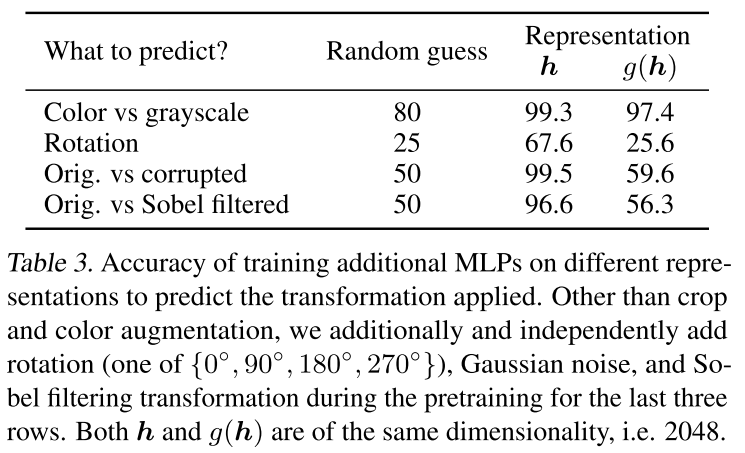
1. Encoder和Head的结构
   1. Unsupervised contrastive learning benefits (more) from bigger models

图7显示，也许毫不奇怪，增加深度和宽度都能提高性能。虽然类似的发现也适用于监督学习（He等人，2016），但我们发现监督模型和在无监督模型上训练的线性分类器之间的差距随着模型大小的增加而缩小，这表明无监督学习比监督学习更受益于更大的模型。



* 1. 非线性投影头提高了前一层的表示质量

然后我们研究包括投影头，即g(h)的重要性。图8显示了使用三种不同架构的头部的线性评估结果：（1）身份映射；（2）线性投影，如以前的几种方法（Wu等人，2018）；（3）默认的非线性投影，有一个额外的隐藏层（和ReLU acti- vation），类似于Bachman等人（2019）。我们观察到，非线性投射比线性投射好（+3%），比没有投射好得多（>10%）。当使用projection head时，无论输出维度如何，都能观察到类似的结果。此外，即使使用非线性投影，投影头之前的层h仍然比之后的层z=g(h)好很多（>10%），这表明投影头之前的隐藏层比之后的层有更好的表现。

我们猜想，在非线性投影之前使用表征的重要性是由于对比性损失所引起的信息损失。特别是，z = g(h)被训练成对数据转换不变的。因此，g可以去除可能对下游任务有用的信息，如物体的颜色或方向。通过利用非线性变换g(-)，可以在h中形成和保持更多的信息。为了验证这一假设，我们进行了实验，使用h或g(h)来学习预测预训练期间应用的变换。这里我们设定g(h)=*W*(2)*σ*(*W*(1)*h*)，输入和输出维度相同（即2048）。表3显示h包含更多关于所应用的变换的信息，而g(h)则失去了信息。

1. 损失函数和批量大小
   1. 可调节温度的归一化交叉熵损失比替代方案效果更好

我们将NT-Xent损失与其他常用的对比性损失函数进行比较，如Logistic损失（Mikolov等人，2013）和margin损失（Schroff等人，2015）。表2显示了目标函数以及损失函数输入的梯度。看一下梯度，我们观察到：1）L2归一化（即余弦相似性）与温度一起有效地权衡了不同的例子，适当的温度可以帮助模型从hard negative example中学习；2）与交叉熵不同，其他目标函数并不按其相对硬度权衡负数。因此，对于这些损失函数，我们必须采用半硬性负面挖掘（Schroff等人，2015）：在不计算所有损失项的梯度的情况下，我们可以使用半硬性负面项（即那些在损失边际内且距离最近，但比正面例子更远的）来计算梯度。

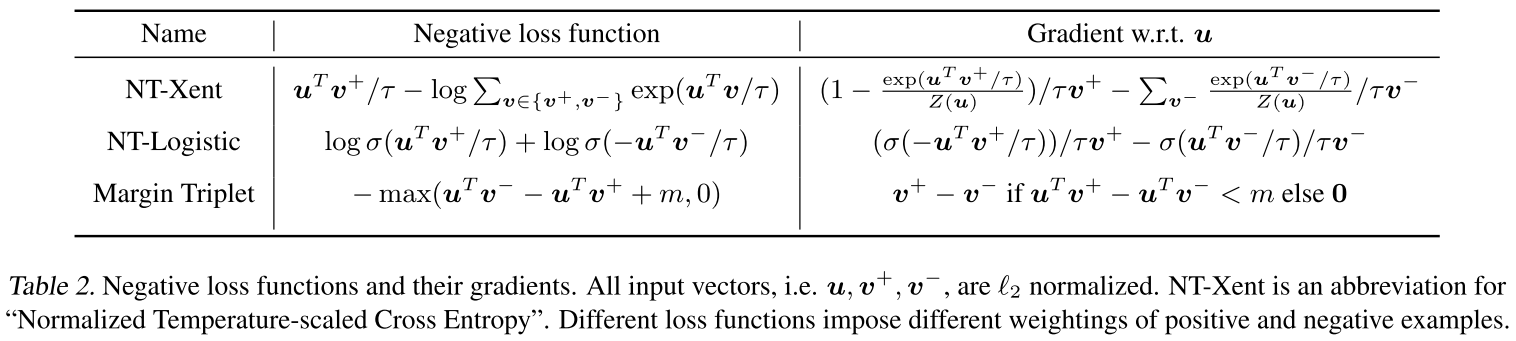
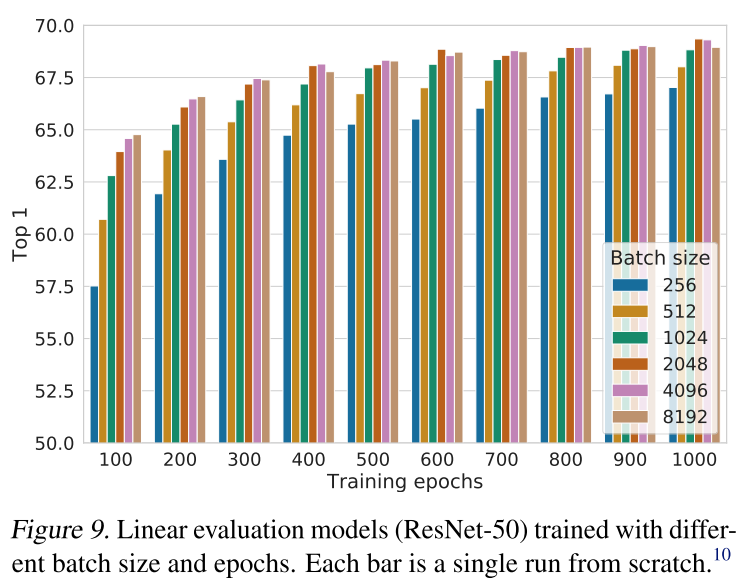


表5显示，如果没有归一化和适当的温度比例，性能会明显下降。如果不进行归一化，对比性任务的准确性会更高。任务的准确性较高，但在线性评估下，所得到的表征是在线性评估下更差。

* 1. Contrastive learning benefits (more) from larger batch sizes and longer training



在对比学习中，较大的批次规模提供了更多的负面例子，促进了收敛（即在给定的准确度下需要较少的历时和步骤）。更长的训练时间也能提供更多的负面例子，改善结果。在附录B.1中，我们提供了更长的训练步骤的结果。

1. Comparison with State-of-the-art

在本小节中，与Kolesnikov等人（2019）；He等人（2019）类似，我们使用ResNet-50的3种不同隐藏层宽度（宽度乘数为1×、2×和4×）。为了更好地收敛，我们的模型在这里被训练了1000个epochs。

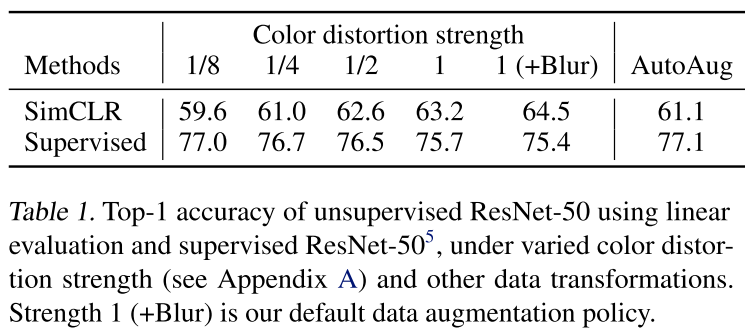


表1显示了不同方法之间的更多数字比较。与之前需要专门设计架构的方法相比，我们能够使用标准的网络来获得明显更好的结果。用我们的ResNet-50获得的最佳结果（4倍）可以与有监督的预训练的ResNet-50相匹配。

**半监督学习**。我们遵循Zhai等人（2019）的做法，以类平衡的方式对已标记的ILSVRC-12训练数据集进行1%或10%的抽样（每类分别为12.8张∼128张）。11 我们只是在没有正则化的标记数据上对整个基础网络进行微调（见附录B.5）。表7显示了我们的结果与最近的方法的比较（Zhai等人，2019；Xie等人，2019；Sohn等人，2020；Wu等人，2018；Donahue & Simonyan，2019；Misra & van der Maaten，2019；Hénaff等人，2019）。来自（Zhai等人，2019）的监督基线由于密集搜索超参数（包括增强）而很强大。同样，我们的方法在使用1%和10%的标签时都比最先进的方法有明显改善。有趣的是，在完整的ImageNet上对我们预训练的ResNet-50（2×，4×）进行微调，也明显优于从头开始训练（达2%，见附录B.2）。

**转移学习**。我们在线性评估（固定特征提取器）和微调设置中评估了12个自然图像数据集的转移学习性能。按照Kornblith等人（2019）的做法，我们对每个模型-数据集组合进行超参数调整，并在验证集上选择最佳超参数。表8显示了ResNet-50（4×）模型的结果。当进行微调时，我们的自监督模型在5个数据集上明显优于自监督基线，而监督基线只在2个数据集（即宠物和花）上有优势。在其余5个数据集上，这两个模型在统计学上是并列的。完整的实验细节以及标准ResNet-50架构的结果在附录B中提供。

1. 相关工作

使图像的表征在小的变换下相互一致的想法可以追溯到Becker和Hinton（1992）。我们通过利用最近在数据增强、网络结构和连贯性损失方面的进展来扩展它。一个类似的一致性想法，但用于类别标签预测，已经在其他情况下被探索过，如半监督学习。

手工制作的辅助任务。最近自我监督学习的复兴始于人工设计的借口任务，如相对补丁预测（Doersch等人，2015），解决拼图（Noroozi和Favaro，2016），着色（Zhang等人，2016）和旋转预测（Gidaris等人，2018；Chen等人，2019）。虽然通过更大的网络和更长的训练可以获得良好的结果（Kolesnikov等人，2019年），但这些借口任务在某种程度上依赖于临时启发式的方法，这限制了所学表征的通用性。

**对比性视觉表征学习**。追溯到Hadsell等人（2006），这些方法通过对比正向对和反向对来学习表征。沿着这些思路，Dosovitskiy等人（2014）提出将每个实例视为一个由特征向量（参数化形式）表示的类。Wu等人（2018）提出使用记忆库来存储实例类表示向量，这种方法在最近的几篇论文中被采用和扩展（Zhuang等人，2019；Tian等人，2019；He等人，2019；Misra & van der Maaten，2019）。其他工作探讨了使用批内样本来代替记忆库进行负向采样（Doersch & Zisserman, 2017; Ye et al., 2019; Ji et al., 2019）。

最近的文献试图将其方法的成功与潜在表征之间的相互信息最大化联系起来（Oord等人，2018；Hénaff等人，2019；Hjelm等人，2018；Bachman等人，2019）。然而，目前还不清楚对比性方法的成功是由相互信息决定的，还是由对比性损失的具体形式决定的（Tschannen等人，2019年）。

我们注意到，我们的框架工作中几乎所有的单个组件都出现在以前的工作中，尽管具体的实例可能有所不同。我们的框架相对于以前的工作的优越性不是由任何单一的设计选择来解释的，而是由它们的组成来解释的。我们在附录C中对我们的设计选择和以前的工作进行了全面的比较。

1. 总结

在这项工作中，我们提出了一个简单的框架和它的实例，用于对比性的视觉表征学习。我们仔细研究了其组成部分，并展示了不同设计选择的效果。通过结合我们的研究结果，我们比以前的自我监督、半监督和转移学习方法有了很大的改进。

我们的方法与ImageNet上的标准监督学习的不同之处仅在于对数据增强的选择、在网络末端使用非线性头以及损失函数。这个简单框架的优势表明，尽管最近兴趣大增，但自我监督学习的价值仍然被低估。

1. 评析