A Simple Framework for Contrastive Learning of Visual Representations

Ting Chen Simon Kornblith Mohammad Norouzi Geoffrey Hinton

Abstract

This paper presents SimCLR: a simple framework for contrastive learning of visual representations. We simplify recently proposed contrastive self-supervised learning algorithms without requiring specialized architectures or a memory bank. In order to understand what enables the contrastive prediction tasks to learn useful representations, we systematically study the major components of our framework. We show that (1) composition of data augmentations plays a critical role in defining effective predictive tasks, (2) introducing a learnable nonlinear transformation between the representation and the contrastive loss substantially improves the quality of the learned representations, and (3) contrastive learning benefits from larger batch sizes and more training steps compared to supervised learning. By combining these findings, we are able to considerably outperform previous methods for self-supervised and semi-supervised learning on ImageNet. A linear classifier trained on self-supervised representations learned by SimCLR achieves 76.5% top-1 accuracy, which is a 7% relative improvement over previous state-of-the-art, matching the performance of a supervised ResNet-50. When fine-tuned on only 1% of the labels, we achieve 85.8% top-5 accuracy, outperforming AlexNet with 100× fewer labels.

【摘要】

本文提出了一个简单的视觉表征对比学习框架 SimCLR。我们简化了最近提出的对比自监督学习算法，而不需要专门的架构或内存库。为了了解是什么使得对比预测任务能够学习有用的表征，我们系统地研究了框架的主要组成部分。（1）数据扩充部分对定义有效的预测任务起着至关重要的作用，（2）在表示和对比损失之间引入一种可学习的非线性变换，大大提高了学习表征的质量，（3）与监督学习相比，对比学习需要更大的批量和更多的训练步骤。通过结合这些发现，我们的方法能够在 ImageNet 上大大优于以前的自监督和半监督学习方法。利用 SimCLR 学习的自监督表示训练的线性分类器达到了 76.5%的 top-1 准确率，比先前的技术水平提高了 7%，与有监督 ResNet-50 的性能相当。当仅对 1%的标签进行微调时，我们的前 5 名准确率达到 85.8%，比 AlexNet 少 100 倍的标签。

1. Introduction

Learning effective visual representations without human supervision is a long-standing problem. Most mainstream approaches fall into one of two classes: generative or discriminative. Generative approaches learn to generate or otherwise model pixels in the input space (Hinton et al., 2006; Kingma & Welling, 2013; Goodfellow et al., 2014).However, pixel-level generation is computationally expensive and may not be necessary for representation learning. Discriminative

approaches learn representations using objective functions similar to those used for supervised learning, but train networks to perform pretext tasks where both the inputs and labels are derived from an unlabeled dataset. Many such approaches have relied on heuristics to design pretext tasks (Doersch et al., 2015; Zhang et al., 2016; Noroozi & Favaro, 2016; Gidaris et al., 2018), which could limit the generality of the learned representations. Discriminative approaches based on contrastive learning in the latent space have recently shown great promise, achieving state-of-the-art results (Hadsell et al., 2006; Dosovitskiy et al., 2014; Oord et al., 2018; Bachman et al., 2019).

【1.介绍】

在无人为监督下学习有效的视觉表征是一个长期存在的问题。大多数主流的方法分为两类：生成式方法和判别式方法。生成方法学习在输入空间中生成或以其他方式建模像素（Hinton 等人，2006 年；Kingma&Welling，2013 年；Goodfellon 等人，2014 年）。然而，像素级的生成在计算上是昂贵的，并且可能不是表示学习所必需的。判别式的方法使用与监督学习类似的目标函数来学习表示，但是训练网络执行代理任务，其中输入和标签都来自未标记的数据集。许多这样的方法依赖于启发式来设计代理任务（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泛化性。基于潜在空间中的对比学习的判别式方法最近显示出巨大的前景，取得了最先进的结果（Hadsell 等人，2006 年；Dosovitskiy 等人，2014 年；Oord 等人，2018 年；Bachman 等人，2019 年）。

In this work, we introduce a simple framework for contrastive learning of visual representations, which we call SimCLR. Not only does SimCLR outperform previous work (Figure 1), but it is also simpler, requiring neither specialized architectures (Bachman et al., 2019; Hénaff et al., 2019) nor a memory bank (Wu et al., 2018; Tian et al., 2019; He et al., 2019; Misra & van der Maaten, 2019).

在这项工作中，我们介绍了一个简单的视觉表征对比学习框架，我们称之为 SimCLR。SimCLR 不仅优于以前的工作（图 1），而且更简单，既不需要专门的架构（Bachman et al.，2019；Hénaff et al.，2019），也不需要内存库（Wu et al.，2018；Tian et al.，2019；He et al.，2019；Misra&van der Maaten，2019）。

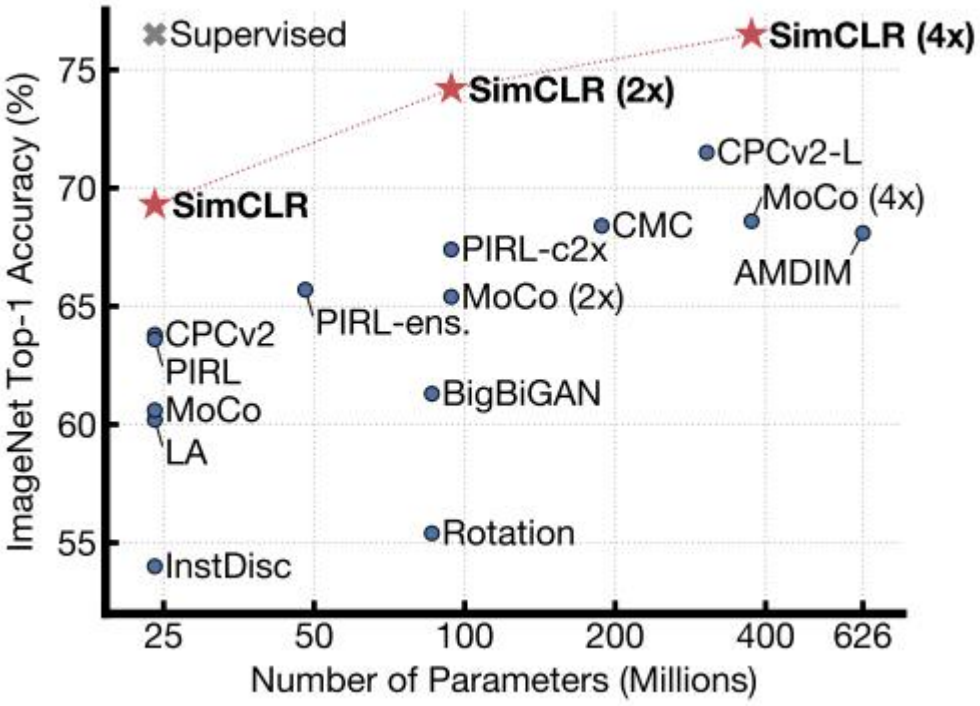


Figure 1. ImageNet Top-1 accuracy of linear classifiers trained on representations learned with different self-supervised methods (pretrained on ImageNet). Gray cross indicates supervised ResNet-50. Our method, SimCLR, is shown in bold.
图 1　ImageNet 线性分类器在不同自监督方法（在 ImageNet 上预先训练）下学习的表征训练的最高精度。灰色十字表示有监督的 ResNet-50。我们的方法 SimCLR 以粗体显示。

In order to understand what enables good contrastive representation learning, we systematically study the major components of our framework and show that:

- Composition of multiple data augmentation operationsis crucial in defining the contrastive prediction tasks thatyield effective representations. In addition, unsupervised contrastive learning benefits from stronger data augmentation than supervised learning.
- Introducing a learnable nonlinear transformation between the representation and the contrastive loss substantially improves the quality of the learned representations.
- Representation learning with contrastive cross entropy loss benefits from normalized embeddings and an appropriately adjusted temperature parameter.
- Contrastive learning benefits from larger batch sizes and longer training compared to its supervised counterpart. Like supervised learning, contrastive learning benefits from deeper and wider networks.

为了了解什么能使对比表征学习变得更好，我们系统地研究了框架的主要组成部分，并表明：

- 多个数据扩充操作的组合对于定义的、能都产生有效表示的对比预测任务至

关重要。此外，与有监督学习相比，无监督对比学习的数据增强能力更强。

- 在表征和对比损失之间引入一种可学习的非线性变换，大大提高了学习表征的质量。
- 具有对比交叉熵损失的表征学习得益于标准化嵌入和适当调整的温标（temperature ）参数。
- 对比学习与监督学习相比，更大的批量和更长的培训时间有利于对比学习。与监督学习一样，对比学习得益于更深更广的网络。

We combine these findings to achieve a new state-of-the-art in self-supervised and semi-supervised learning on ImageNet ILSVRC-2012 (Russakovsky et al., 2015). Under the linear evaluation protocol, SimCLR achieves 76.5% top-1 accuracy, which is a 7% relative improvement over previous state-of-the-art (Hénaff et al., 2019). When fine-tuned with only 1% of the ImageNet labels, SimCLR achieves 85.8% top-5 accuracy, a relative improvement of 10% (Hénaff et al., 2019). When fine-tuned on other natural image classification datasets, SimCLR performs on par with or better than a strong supervised baseline (Kornblith et al., 2019) on 10 out of 12 datasets.

我们结合这些发现，在 ImageNet ILSVRC-2012 上实现了自我监督和半监督学习的新技术（Russakovsky 等人，2015）。在线性评估协议下，SimCLR 达到了 76.5%的 top-1 准确率，这比之前的最先进水平相对提高了 7%（Hénaff et al.，2019）。当仅使用 1%的 ImageNet 标签进行微调时，SimCLR 达到了 85.8%的前 5 名准确率，相对提高了 10%（Hénaff et al.，2019）。当对其他自然图像分类数据集进行微调时，SimCLR 在 12 个数据集中的 10 个数据集上的性能与强监督基线（Kornblith 等人，2019）相当或更好。

2. Method

2.1. The Contrastive Learning Framework

Inspired by recent contrastive learning algorithms (see Section 7 for an overview), SimCLR learns representations by maximizing agreement between differently augmented views of the same data example via a contrastive loss in the latent space. As illustrated in Figure 2, this framework comprises the following four major components.

- A stochastic data augmentation module that transforms any given data example randomly resulting in two correlated views of the same example, denoted $\tilde{x}_i$ and $\tilde{x}_j$, which we consider as a positive pair. In this work, we sequentially apply three simple augmentations: random cropping followed by resize back to the original size, random color distortions, and random Gaussian blur. As shown in Section 3, the combination of random crop and color distortion is crucial to achieve a good performance.
- A neural network base encoder $f(\cdot)$ that extracts representation vectors from augmented data examples. Our framework allows various choices of the network architecture without any constraints. We opt for simplicity and adopt the commonly used ResNet (He et al., 2016)to obtain $h_i = f(\tilde{x}_i) = \text{ResNet}(\tilde{x}_i)$ where $h_i \in \mathbb{R}^d$ is the output after the average pooling layer.

- A small neural network projection head g(•) that maps representations to the space where contrastive loss is applied. We use a MLP with one hidden layer to obtain zi = g(hi) = W(2) σ (W(1)hi) where σ is a ReLU non-linearity. As shown in section 4, we find it beneficial to define the contrastive loss on zi’s rather than hi’s.
- A contrastive loss function defined for a contrastive prediction task. Given a set {x ̃ k} including a positive pair of examples x ̃ i and x ̃ j , the contrastive prediction task aims to identify x̃j in {x̃k}k=i for a given x̃i.

【2 方法】

2.1 对比学习框架

受最近的对比学习算法（见第 7 节概述）的启发，SimCLR 通过在潜在空间中的对比损失最大化同一数据示例的不同增强视图之间的一致性来学习表示。如图 2 所示，该框架包括以下四个主要组件：

- 一个随机数据扩充模块，它随机转换任何给定的数据示例，产生同一示例的两个相关视图，表示为 x̃i 和 x̃j，我们将其视为正对。在这项工作中，我们依次应用了三个简单的扩展：随机裁剪，然后将大小调整回原始大小；随机颜色失真；随机高斯模糊。如第 3 节所示，随机裁剪和颜色失真的结合是获得良好性能的关键。
- 一种基于神经网络的编码器 f(•)，从增强数据示例中提取表示向量。我们的框架允许在没有任何约束的情况下选择各种网络架构。我们选择简单，采用常用的 ResNet（He 等人，2016），得到 h=f（x̃i）=ResNet（x̃i），其中 hi∈Rd 是平均池化层后的输出。
- 一个小的神经网络映射头 g（•），将“表示”映射到应用对比损失的空间。我们使用带有一个隐藏层的 MLP 得到 zi=g（hi）=W（2） σ （W（1）hi），其 σ 中是一个 ReLU 非线性。如第 4 节所示，我们发现定义的对比损失在 zi 上比 hi 上更有效。
- 为对比预测任务定义的对比损失函数。在给定一组{x̃k}包括一对正例 x̃i 和 x̃j，对比预测任务旨在从{x̃k}k/=i 中确定给定 x̃i 的 x̃j。

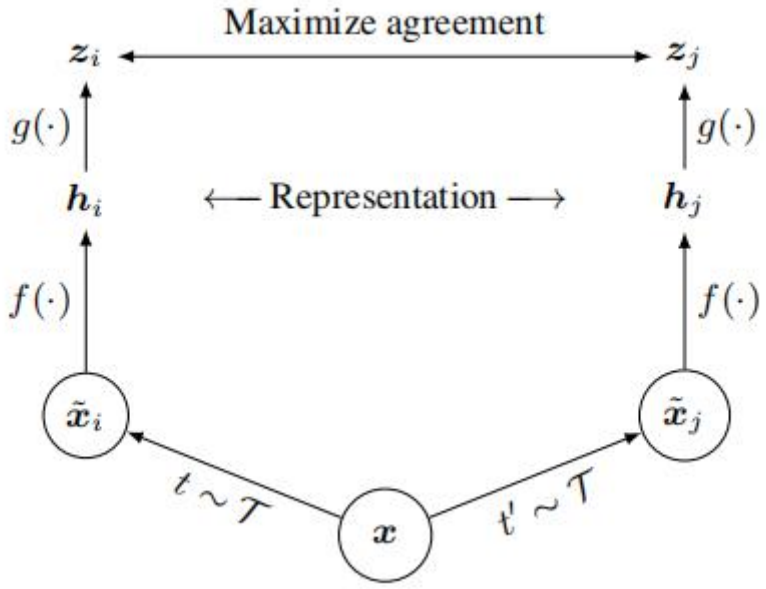


Figure 2. A simple framework for contrastive learning of visual representations. Two separate data augmentation operators are sampled from the same family of augmentations (t ~ T and t' ~ T) and applied to each data example to obtain two correlated views. A base encoder network f(•) and a projection head g(•) are trained to maximize agreement using a contrastive loss. After training is completed, we throw away the projection head g(•) and use encoder f(•) and representation h for downstream tasks.
图 2　视觉表征对比学习的一个简单框架。从同一个扩充族（t~ T 和 t'~ T）中抽取两个独立的数据扩充操作，并将其应用于每个数据示例，以获得两个相关视图。一个基本编码器网络 f（ • ）和一个映射头 g（ • ）被训练成使用对比损失最大化一致性。训练结束后，我们扔掉映射头 g（ • ），使用编码器 f（ • ）和表示 h 来完成下游任务。

We randomly sample a minibatch of N examples and define the contrastive prediction task on pairs of augmented examples derived from the minibatch, resulting in 2N data points. We do not sample negative examples explicitly. Instead, given a positive pair, similar to (Chen et al., 2017), we treat the other 2(N-1) augmented examples within a minibatch as negative examples. Let sim(u, v) = $\text{sim}(u, v) = u^T v / \|u\| \|v\|$ denote the dot product between l2 normalized u and v (i.e. cosine similarity). Then the loss function for a positive pair of examples (i, j) is defined as

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}, \qquad (1)$$

where $\mathbb{1}_{[k \neq i]} \in \{0, 1\}$ is an indicator function evaluating to 1 iff k / = i and τ denotes a temperature parameter. The final loss is computed across all positive pairs, both (i, j) and (j, i), in a mini-batch. This loss has been used in previous work (Sohn, 2016; Wu et al., 2018; Oord et al., 2018); for convenience, we term it NT-Xent (the

normalized temperature-scaled cross entropy loss).

我们随机抽取一小批样本(minibatch)，并对从中衍生出成对的扩充样本，同时定义对比预测任务，因此得到 2N 个数据点。我们不显式地抽取负例。相反，给定一对正样本，类似于（Chen et al.，2017），我们将一个小批次中的另外 2*（N−1）个扩充示例视为负例。设 $\mathrm{sim}(u,v)=u^Tv/\|u\|\|v\|$ 表示 L2 正则化的 u 和 v 之间的点积（即余弦相似性）。然后将一对正例（i，j）的损失函数定义为：

$$\ell_{i,j}=-\log\frac{\exp(\mathrm{sim}(\boldsymbol{z}_i,\boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N}\mathbb{1}_{[k\neq i]}\exp(\mathrm{sim}(\boldsymbol{z}_i,\boldsymbol{z}_k)/\tau)}, \qquad (1)$$

其中 $\mathbb{1}_{[k\neq i]}\in\{0,1\}$ 是指示函数，表示是否 k == i， τ 表示温标超参数。最后的损失是在一个小批量（minibatch）中计算所有正对（i，j）和（j，i）的。此损失已在以前的工作中使用过（Sohn，2016；Wu et al.，2018；Oord et al.，2018）；为方便起见，我们将其称为 NT-Xent（归一化温标尺度交叉熵损失）。

Algorithm 1 SimCLR's main learning algorithm.

input: batch size N, constant τ, structure of $f, g, \mathcal{T}$.
for sampled minibatch $\{\boldsymbol{x}_k\}_{k=1}^{N}$ **do**
  for all $k\in\{1,\dots,N\}$ **do**
    draw two augmentation functions $t\sim\mathcal{T}, t'\sim\mathcal{T}$
    # the first augmentation
    $\tilde{\boldsymbol{x}}_{2k-1}=t(\boldsymbol{x}_k)$
    $\boldsymbol{h}_{2k-1}=f(\tilde{\boldsymbol{x}}_{2k-1})$ # representation
    $\boldsymbol{z}_{2k-1}=g(\boldsymbol{h}_{2k-1})$ # projection
    # the second augmentation
    $\tilde{\boldsymbol{x}}_{2k}=t'(\boldsymbol{x}_k)$
    $\boldsymbol{h}_{2k}=f(\tilde{\boldsymbol{x}}_{2k})$ # representation
    $\boldsymbol{z}_{2k}=g(\boldsymbol{h}_{2k})$ # projection
  end for
  for all $i\in\{1,\dots,2N\}$ and $j\in\{1,\dots,2N\}$ **do**
    $s_{i,j}=\boldsymbol{z}_i^\top\boldsymbol{z}_j/(\|\boldsymbol{z}_i\|\|\boldsymbol{z}_j\|)$ # pairwise similarity
  end for
  define $\ell(i,j)$ **as** $\ell(i,j)=-\log\frac{\exp(s_{i,j}/\tau)}{\sum_{k=1}^{2N}\mathbb{1}_{[k\neq i]}\exp(s_{i,k}/\tau)}$
  $\mathcal{L}=\frac{1}{2N}\sum_{k=1}^{N}[\ell(2k-1,2k)+\ell(2k,2k-1)]$
  update networks f and g to minimize $\mathcal{L}$
end for
return encoder network $f(\cdot)$, and throw away $g(\cdot)$

Algorithm 1 summarizes the proposed method.

2.2. Training with Large Batch Size

To keep it simple, we do not train the model with a memory bank (Wu et al., 2018; He et al., 2019). Instead, we vary the training batch size N from 256 to 8192. A batch size of 8192 gives us 16382 negative examples per positive pair from both augmentation views. Training with large batch size may be unstable when using standard SGD/Momentum with linear learning rate scaling (Goyal et al., 2017). To stabilize the training, we use the LARS optimizer (You et al., 2017) for all batch sizes. We train our model with Cloud TPUs, using 32 to 128 cores depending on the batch size.

2.2 大的批量训练

为了简单起见，我们不使用存储库训练模型（Wu et al.，2018；He et al.，2019）。相反，我们将训练批大小从 256 变为 8192。批处理大小为 8192，从两个增强视图来看，为每对正样本都有 16382 个负样本。当使用带有线性学习率缩放的标准 SGD/Momentum 时，大批量培训可能不稳定（Goyal 等人，2017 年）。为了稳定训练，我们使用 LARS 优化器（You 等人。，N 2017)适用于所有批量。我们用云 TPU 来训练我们的模型，根据批量大小使用 32 到 128 个核心。

Global BN. Standard ResNets use batch normalization (Ioffe & Szegedy, 2015). In distributed training with data parallelism, the BN mean and variance are typically aggregated locally per device. In our contrastive learning, as positive pairs are computed in the same device, the model can exploit the local information leakage to improve prediction accuracy without improving representations. We address this issue by aggregating BN mean and variance over all devices during the training. Other approaches include shuffling data examples across devices (He et al., 2019), or replacing BN with layer norm (Hénaff et al., 2019).

全局 BN 标准 Resnet 使用批归一化（Ioffe&Szegedy，2015）。在具有数据并行性的分布式训练中，BN 均值和方差通常在每个设备上局部聚集。在我们的对比学习中，由于正样本对是在同一个设备上计算的，该模型可以在不改进表示的情况下利用局部信息泄漏来提高预测精度。我们通过在训练期间汇总所有设备的 BN 均值和方差来解决这个问题。其他的方法包括在设备间对数据示例进行洗牌（He et al.，2019），或者用层规范代替 BN（Hénaff et al.，2019）。

2.3. Evaluation Protocol

Here we lay out the protocol for our empirical studies, which aim to understand different design choices in our framework.

Dataset and Metrics. Most of our study for unsupervised pretraining (learning encoder network f without labels) is done using the ImageNet ILSVRC-2012 dataset (Russakovsky et al., 2015). Some additional pretraining experiments on CIFAR-10 (Krizhevsky & Hinton, 2009) can be found in Appendix B.9. We also test the pretrained results on a wide range of datasets for transfer learning. To evaluate the learned representations, we follow the widely used linear evaluation protocol (Zhang et al., 2016; Oord et al., 2018; Bachman et al., 2019; Kolesnikov et al., 2019), where a linear classifier is trained on top of the frozen base network, and test accuracy is used as a proxy for representation quality. Beyond linear evaluation, we also compare against state-of-the-art on semi-supervised and transfer learning.

Default setting. Unless otherwise specified, for data augmentation we use random crop and resize (with random flip), color distortions, and Gaussian blur (for details, see Appendix A). We use ResNet-50 as the base encoder network, and a 2-layer MLP projection head to project the representation to a 128-dimensional latent space. As the loss, we use NT-Xent, optimized using LARS with learning rate of 4.8 (= 0.3 × BatchSize/256) and weight decay of 10-6 . We train at batch size 4096 for 100 epochs. Furthermore, we use linear warmup for the first 10 epochs, and decay the learning rate with the cosine decay schedule without restarts (Loshchilov & Hutter, 2016).

2.3 评估方案

在这里，我们列出了我们的实验研究的方案，目的是了解我们框架中不同的设计选择。

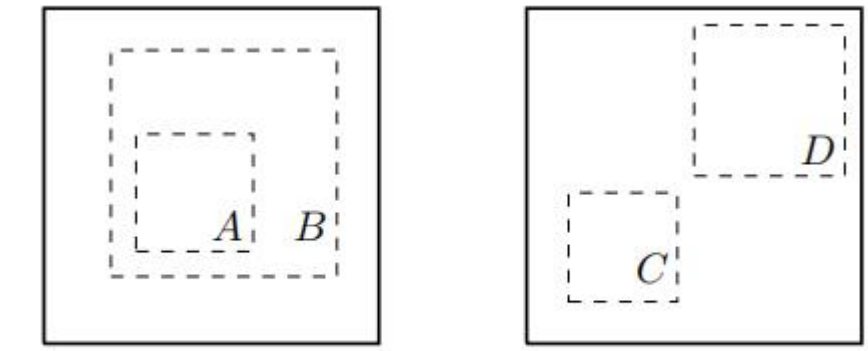
数据集和指标 我们对无监督预训练（无标签的学习编码器网络）的大部分研究是使用 ImageNet ILSVRC-2012 数据集完成的（Russakovsky 等人，2015）。关于 CIFAR-10 的一些额外的预训练实验（Krizhevsky&Hinton，2009）见附录 B.9。我们也在大量的数据集上测试了预训练的迁移学习结果。为了评估学习到的表征，我们遵循广泛使用的线性评估协议（Zhang 等人，2016；Oord 等人，2018；Bachman 等人，2019；Kolesnikov 等人，2019），其中线性分类器在冻结基础网络上进行训练，测试精度用作表征质量的代理。除了线性评估之外，我们还比较了半监督和转移学习的最新进展。

默认设置 除非另有规定，对于数据增强，我们使用随机裁剪和调整大小（随机翻转）、颜色失真和高斯模糊（有关详细信息，请参阅附录 A）。我们使用 ResNet-50 作为基本编码网络，并使用一个 2 层 MLP 映射头将表示映射到 128 维的潜在空间。对于损失函数，我们使用 NT-Xent，并使用 LARS 优化，学习率为 4.8（=0.3×BatchSize/256），权重衰减为 10-6。我们训练的批量（batchsize）大小 4096，有 100 个 epoch。而且，我们对前 10 个阶段使用线性预热，并在不重新启动的情况下使用余弦衰减时间表衰减学习率（Loshchilov&Hutter，2016）。

3. Data Augmentation for Contrastive Representation Learning

Data augmentation defines predictive tasks. While data augmentation has been widely used in both supervised and unsupervised representation learning (Krizhevsky et al.,2012; Hénaff et al., 2019; Bachman et al., 2019), it has not been considered as a systematic way to define the contrastive prediction task. Many existing approaches define contrastive prediction tasks by changing the architecture. For example, Hjelm et al. (2018); Bachman et al. (2019) achieve global-to-local view prediction via constraining the receptive field in the network architecture, whereas Oord et al. (2018); Hénaff et al. (2019) achieve neighboring view prediction via a fixed image splitting procedure and a context aggregation network. We show that this complexity can be avoided by performing simple random cropping (with resizing) of target images, which creates a family of predictive tasks subsuming the above mentioned two, as shown in Figure 3. This simple design choice conveniently decouples the predictive task from other components such as the neural network architecture. Broader contrastive prediction tasks can be defined by extending the family of augmentations and composing them stochastically.

数据扩充明确了预测任务 虽然数据扩充在有监督和无监督表征学习中得到了广泛的应用（Krizhevsky 等人，2012；Hénaff et al.，2019；Bachman et al.，2019），但它并没有被视为明确对比预测任务的系统方法。许多现有的方法通过改变结构来定义对比预测任务。例如，Hjelm 等人。（2018 年）；Bachman 等人（2019）通过限制网络架构中的接收场（receptive field）实现全局到局部视图预测，而 Oord 等人。（2018 年）；Hénaff 等人。（2019）通过固定图像划分过程和上下文聚合网络实现邻域视图预测。我们表明，通过对目标图像执行简单的随机裁剪（调整大小）可以避免这种复杂性，这将创建一系列包含上述两个方法的预测任务，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（如神经网络体系结构）解耦。更广泛的对比预测任务可以通过一系列扩充和随机组合来明确。



(a) Global and local views. (b) Adjacent views.

Figure 3. Solid rectangles are images, dashed rectangles are random crops. By randomly cropping images, we sample contrastive prediction tasks that include global to local view (B → A) or adjacent view (D → C) prediction.
图 3 实心矩形是图像，虚线矩形是随机裁剪。通过随机裁剪图像，我们采样的对比预测任务，包括全局到局部视图（B→A）或相邻视图（D→C）的预测。

3.1. Composition of data augmentation operations is crucial for learning good representations

To systematically study the impact of data augmentation, we consider several common augmentations here. One type of augmentation involves spatial/geometric transformation of data, such as cropping and resizing (with horizontal flipping), rotation (Gidaris et al., 2018) and cutout (DeVries & Taylor, 2017). The other type of augmentation involves appearance transformation, such as color distortion (including color dropping, brightness, contrast, saturation, hue) (Howard, 2013; Szegedy et al., 2015), Gaussian blur, and Sobel filtering. Figure 4 visualizes the augmentations that we study in this work.

To understand the effects of individual data augmentations and the importance of augmentation composition, we investigate the performance of our framework when applying augmentations individually or in pairs. Since ImageNet images are of different sizes, we always apply crop and resize images (Krizhevsky et al., 2012; Szegedy et al., 2015), which makes it difficult to study other augmentations in the absence of cropping. To eliminate this confound, we consider an asymmetric data

transformation setting for this ablation. Specifically, we always first randomly crop images and resize them to the same resolution, and we then apply the targeted transformation(s) only to one branch of the framework in Figure 2, while leaving the other branch as the identity (i.e. t(xi) = xi). Note that this asymmetric data augmentation hurts the performance. Nonetheless, this setup should not substantively change the impact of individual data augmentations or their compositions.

3.1 数据扩充操作的组合是学习良好表征的关键

为了系统地研究数据扩充的影响，我们在这里考虑几种常见的扩充。一种类型的增强涉及数据的空间/几何转换，例如裁剪和调整大小（水平翻转）、旋转（Gidaris 等人，2018）和剪切（DeVries&Taylor，2017）。另一种类型的增强涉及外观转换，例如颜色失真（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 滤波。图 4 显示了我们在这项工作中研究的扩展。

为了了解单个数据扩充的影响和扩充组合的重要性，我们研究了我们的框架在单独或成对应用扩充时的性能。由于 ImageNet 图像的大小不同，我们总是使用裁剪和调整大小的图像（Krizhevsky 等人，2012；Szegedy 等人，2015），这使得在没有裁剪的情况下很难研究其他增强效果。为了消除这种混淆，我们考虑了一个不对称的数据转换设置。具体地说，我们总是首先随机裁剪图像并将其调整到相同的分辨率，然后将目标转换仅应用于图 2 中框架的一个分支，而将另一个分支作为标识（即 i.e. t(xi) = xi）。请注意，这种不对称的数据扩充会损害性能。尽管如此，这些数据的扩充或扩充的组成不应发生实质性的变化。

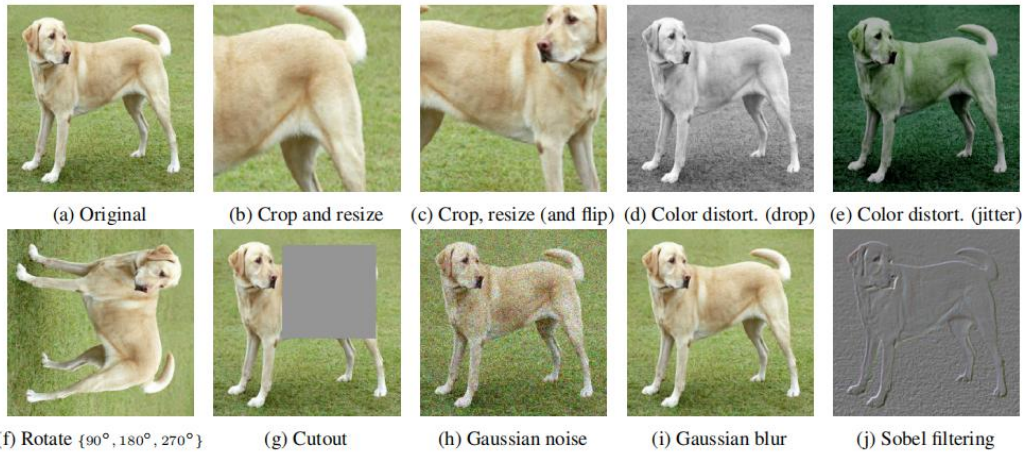


Figure 4. Illustrations of the studied data augmentation operators. Each augmentation can transform data stochastically with some internal parameters (e.g. rotation degree, noise level). Note that we only test these operators in ablation, the augmentation policy used to train our models only includes random crop (with flip and resize), color distortion, and Gaussian blur. (Original image cc-by: Von.grzanka)

图 4 所研究的数据扩充算子图解。每一种增广方法都可以随机地对数据进行一些内部参数（如旋转度、噪声级）的变换。注意，我们只在消融实验中测试这些操作，用于训练模型的增强策略只包括随机裁剪（带翻转和调整大小）、颜色失真和高斯模糊。

Figure 5 shows linear evaluation results under individual and composition of

transformations. We observe that no single transformation suffices to learn good representations, even though the model can almost perfectly identify the positive pairs in the contrastive task. When composing augmentations, the contrastive prediction task becomes harder, but the quality of representation improves dramatically. Appendix B.2 provides a further study on composing broader set of augmentations.
One composition of augmentations stands out: random cropping and random color distortion. We conjecture that one serious issue when using only random cropping as data augmentation is that most patches from an image share a similar color distribution. Figure 6 shows that color histograms alone suffice to distinguish images. Neural nets may exploit this shortcut to solve the predictive task. Therefore, it is critical to compose cropping with color distortion in order to learn generalizable features.

图 5 显示了单个和组合变换下的线性评估结果。我们观察到，没有一个单一的转换就足以学习好的表征，即使这个模型几乎可以完美地识别对比任务中的正对。在组合增强项时，对比预测任务变得更加困难，但表示质量却显著提高。附录 B.2 提供了关于组成更广泛的扩充集的进一步研究。
其中一个突出的组成部分是：随机裁剪和随机颜色失真。我们推测，当仅使用随机裁剪作为数据增强时，一个严重的问题是来自图像的大多数面片共享相似的颜色分布。图 6 显示，仅颜色直方图就足以区分图像。神经网络可以利用这一捷径来解决预测任务。因此，为了学习可归纳的特征，用颜色失真与裁剪组合是非常重要的。

Figure 5. Linear evaluation (ImageNet top-1 accuracy) under individual or composition of data augmentations, applied only to one branch. For all columns but the last, diagonal entries correspond to single transformation, and off-diagonals correspond to composition of two transformations (applied sequentially). The last column reflects the average over the row.
图 5 单独或组合数据扩充下的线性评估（ImageNet top-1 精度），仅适用于一个分支。对于除最后一列之外的所有列，对角线条目对应于单个转换，而非对角线对应于两个变换的组合（顺序应用）。最后一列反映行的平均值。

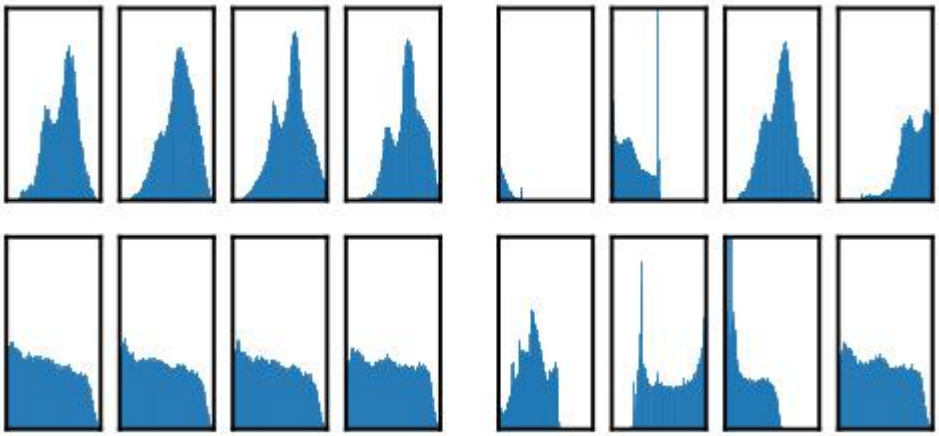


Figure 6. Histograms of pixel intensities (over all channels) for different crops of two different images (i.e. two rows). The image for the first row is from Figure 4. All axes have the same range.
图 6 两个不同图像（即两行）的不同 crop 的像素强度（在所有通道上）的直方图。第一行的图像来自图 4。所有轴的范围相同。

3.2. Contrastive learning needs stronger data augmentation than supervised learning

To further demonstrate the importance of the color augmentation, we adjust the strength of color augmentation as shown in Table 1. Stronger color augmentation substantially improves the linear evaluation of the learned unsupervised models. In this context, AutoAugment (Cubuk et al., 2019), a sophisticated augmentation policy found using supervised learning, does not work better than simple cropping+ (stronger) color distortion. When training supervised models with the same set of augmentations, we observe that stronger color augmentation does not improve or even hurts their performance. Thus, our experiments show that unsupervised contrastive learning benefits from stronger (color) data augmentation than supervised learning. Although previous work has reported that data augmentation is useful for self-supervised learning (Doersch et al., 2015; Bachman et al., 2019; Hénaff et al., 2019; Asano et al., 2019), we show that data augmentation that does not yield accuracy benefits for supervised learning can still help considerably with contrastive learning.

3.2 对比学习比监督学习需要更强的数据扩充能力

为了进一步证明颜色增强的重要性，我们将颜色增强的强度调整为如表 1 所示。更强的颜色增强大大提高了学习的无监督模型的线性评估。在这种情况下，AutoAugment（Cubuk et al.，2019）是一种使用监督学习发现复杂增强策略的方法，其效果并不比简单裁剪+（更强）颜色失真更好。当使用相同的增强集训练监督模型时，我们观察到更强的颜色增强并没有改善甚至损害它们的性能。因此，我们的实验表明，与有监督学习相比，无监督对比学习得益于更强的（彩色）数据增强。尽管先前的研究报告数据扩充对自监督学习是有用的（Doersch 等人，2015；Bachman 等人，2019；Hénaff et al.，2019；Asano et al.，2019），但我们表明，数据扩充对于监督学习不产生准确性益处，仍然可以在很大程度上有助于对比学习。

Methods	Color distortion strength 1/8	1/4	1/2	1	1 (+Blur)	AutoAug
SimCLR	59.6	61.0	62.6	63.2	64.5	61.1
Supervised	77.0	76.7	76.5	75.7	75.4	77.1

Table 1. Top-1 accuracy of unsupervised ResNet-50 using linear evaluation and supervised ResNet-50, under varied color distortion strength (see Appendix A) and other data transformations. Strength 1 (+Blur) is our default data augmentation policy.
表 1。在不同的颜色失真强度（见附录 A）和其他数据转换下，使用线性评估和有监督 ResNet-50 的无监督 ResNet-50 的最高精度。强度 1（+Blur）是我们默认的数据扩充策略。

4. Architectures for Encoder and Head

4.1. Unsupervised contrastive learning benefits (more) from bigger models

Figure 7 shows, perhaps unsurprisingly, that increasing depth and width both improve performance. While similar findings hold for supervised learning (He et al., 2016), we find the gap between supervised models and linear classifiers trained on unsupervised models shrinks as the model size increases, suggesting that unsupervised learning benefits more from bigger models than its supervised counterpart.
图 7 显示，也许并不奇怪，增加深度和宽度都可以提高性能。虽然类似的研究结果也适用于监督学习（He et al.，2016），但我们发现，随着模型尺寸的增加，有监督模型和训练在无监督模型上的线性分类器之间的差距缩小，这表明无监督学习从更大的模型中获益比在有监督的模型中获益更多。

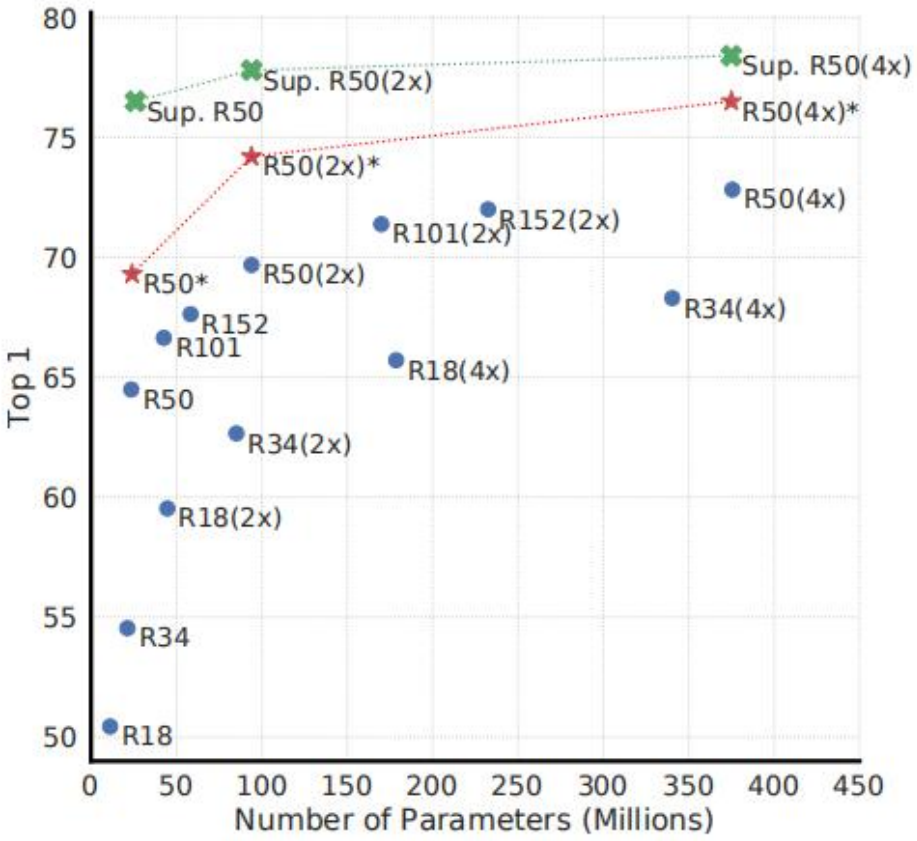


Figure 7. Linear evaluation of models with varied depth and width. Models in blue dots are ours trained for 100 epochs, models in red stars are ours trained for 1000 epochs, and models in green crosses are supervised ResNets trained for 90 epochs7 (He et al., 2016).
图 7 不同深度和宽度模型的线性评估。蓝色圆点的模型是我们训练的 100 个 epochs，红星上的模型是我们训练的 1000 个 epochs，绿色十字的模型是 90 个 epochs 的监督 ResNet 网络（He et al.，2016）。

4.2. A nonlinear projection head improves the representation quality of the layer before it

We then study the importance of including a projection head, i.e. g(h). Figure 8 shows linear evaluation results using three different architecture for the head: (1) identity mapping; (2) linear projection, as used by several previous approaches (Wu et al., 2018); and (3) the default nonlinear projection with one additional hidden layer (and ReLU activation), similar to Bachman et al. (2019). We observe that a nonlinear projection is better than a linear projection (+3%), and much better than no projection (>10%). When a projection head is used, similar results are observed regardless of output dimension. Furthermore, even when nonlinear projection is used, the layer before the projection head, h, is still much better (>10%) than the layer after, z = g(h), which shows that the hidden layer before the projection head is a better representation than the layer after.

We conjecture that the importance of using the representation before the nonlinear projection is due to loss of information induced by the contrastive loss. In particular, z = g(h) is trained to be invariant to data transformation. Thus, g can remove information that may be useful for the downstream task, such as the color or orientation of objects. By leveraging the nonlinear transformation g(·), more information can be formed and maintained in h. To verify this hypothesis, we conduct experiments that use either h or g(h) to learn to predict the transformation applied during the pretraining. Here we set g(h) = W(2)σ(W(1)h), with the same input and output dimensionality (i.e. 2048). Table 3 shows h contains much more information about the transformation applied, while g(h) loses information. Further analysis can be found in Appendix B.4.

4.2 非线性投影头提高了前面层的表征质量

然后，我们研究了包含投影头的重要性，即：g（h）。图 8 显示了对头部使用三种不同架构的线性评估结果：（1）身份映射；（2）线性投影，如之前几种方法所使用的（Wu 等人，2018 年）；（3）默认的非线性投影，带有一个额外的隐藏层（和 ReLU 激活），类似于 Bachman 等人（2019 年）。我们观察到非线性投影比线性投影好（+3%），也比没有投影好得多（>10%）。当使用投影头时，无论输出尺寸如何，都可以观察到类似的结果。此外，即使使用非线性投影，投影头之前的层 h 仍然比后面的层 z=g(h)更好（>10%），这表明投影头之前的隐藏层比后面的层更好。

我们推测，在非线性投影之前使用表征的重要性是由于对比损失导致的信息丢失。特别是，z=g（h）被训练成对数据转换不变性。因此，可以删除可能对下游任务有用的信息，例如对象的颜色或方向。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）

来学习预测在预训练期间应用的变换。这里我们设置（h）=W（2）σ（W（1）h），具有相同的输入和输出维数（即 2048）。表 3 显示 h 包含更多关于应用的转换的信息，而 g（h）则丢失了信息。进一步分析可以附录 B.4 查看。

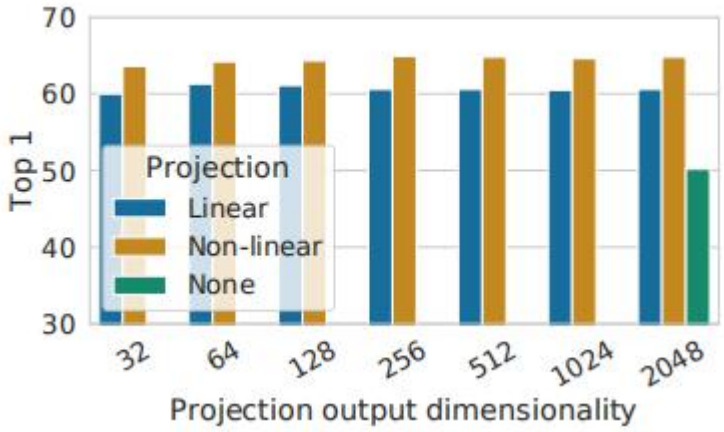


Figure 8. Linear evaluation of representations with different projection heads g(·) and various dimensions of z = g(h). The representation h (before projection) is 2048-dimensional here.
图 8 不同投影头 g（·）和 z=g（h）的不同维数表示的线性估计。在这里，表征 h（投影前）是 2048 维的。

What to predict?	Random guess	Representation h	Representation $g(h)$
Color vs grayscale	80	99.3	97.4
Rotation	25	67.6	25.6
Orig. vs corrupted	50	99.5	59.6
Orig. vs Sobel filtered	50	96.6	56.3

Table 3. Accuracy of training additional MLPs on different representations to predict the transformation applied. Other than crop and color augmentation, we additionally and independently add rotation (one of {0◦, 90◦, 180◦, 270◦}), Gaussian noise, and Sobel filtering transformation during the pretraining for the last three rows. Both h and g(h) are of the same dimensionality, i.e. 2048.
表 3 在不同的表征上训练额外的 mlp 以预测所应用的变换的准确性。除了裁剪和颜色增强，我们还额外独立地添加了旋转（其中之一{0◦, 90◦, 180◦, 270◦}）高斯噪声和 Sobel 滤波变换在最后三行预训练时。h 和 g（h）的维数相同，即 2048。

5. Loss Functions and Batch Size

5.1. Normalized cross entropy loss with adjustable temperature works better than alternatives

We compare the NT-Xent loss against other commonly used contrastive loss functions, such as logistic loss (Mikolov et al., 2013), and margin loss (Schroff et al., 2015). Table 2 shows the objective function as well as the gradient to the input of the loss function. Looking at the gradient, we observe 1) L2 normalization (i.e. cosine similarity) along with temperature effectively weights different examples, and an appropriate temperature can help the model learn from hard negatives; and 2) unlike

cross-entropy, other objective functions do not weigh the negatives by their relative hardness. As a result, one must apply semi-hard negative mining (Schroff et al., 2015) for these loss functions: instead of computing the gradient over all loss terms, one can compute the gradient using semi-hard negative terms (i.e., those that are within the loss margin and closest in distance, but farther than positive examples).

To make the comparisons fair, we use the same L2 normalization for all loss functions, and we tune the hyperparameters, and report their best results. Table 4 shows that, while (semi-hard) negative mining helps, the best result is still much worse than our default NT-Xent loss.

We next test the importance of the L2 normalization (i.e. cosine similarity vs dot product) and temperature τ in our default NT-Xent loss. Table 5 shows that without normalization and proper temperature scaling, performance is significantly worse. Without L2 normalization, the contrastive task accuracy is higher, but the resulting representation is worse under linear evaluation.

【5 损失函数与批量大小】

5.1 可调温标下的归一化交叉熵损失比其他方案更好

我们将 NT-Xent 损失与其他常用的对比损失函数进行比较，如 logistic 损失（Mikolov et al.，2013）和边际损失（Schroff et al.，2015）。表 2 显示了目标函数以及损失函数的输入梯度。从梯度上看，我们观察到 1）L2 归一化（即余弦相似性）与温标一起有效地加权不同的样本，适当的温标可以帮助模型从硬负样本中学习；2）与交叉熵不同，其他目标函数不通过负例的相对难度来衡量负例的权重。因此，我们必须对这些损失函数应用半硬负挖掘（Schroff et al.，2015）：与其计算所有损失项的梯度，不如使用半硬负项计算梯度（即那些在损失范围内、距离最近但比正例更远的项）。

为了使比较公平，我们对所有损失函数使用相同的 L2 标准化，我们调整超参数，并报告它们的最佳值结果。表 4 表明，虽然（半硬）负挖掘有帮助，但最好的结果仍然比我们默认的 NT-Xent 损失要糟糕得多。

接下来，我们测试 L2 归一化（即余弦相似性与点积）和温标在默认 NT-Xent 损耗中的重要性。表 5 显示，如果不进行标准化和适当的温标缩放，性能会显著降低。在没有 L2 标准化的情况下，对比任务的准确率较高，但在线性评估下得到的表现更差。

Name	Negative loss function	Gradient w.r.t. $\boldsymbol{u}$
NT-Xent	$\boldsymbol{u}^T\boldsymbol{v}^+/\tau - \log \sum_{v\in\{v^+,v^-\}} \exp(\boldsymbol{u}^T\boldsymbol{v}/\tau)$	$(1 - \frac{\exp(\boldsymbol{u}^T\boldsymbol{v}^+/\tau)}{Z(\boldsymbol{u})})/\tau\boldsymbol{v}^+ - \sum_{v^-} \frac{\exp(\boldsymbol{u}^T\boldsymbol{v}^-/\tau)}{Z(\boldsymbol{u})}/\tau\boldsymbol{v}^-$
NT-Logistic	$\log \sigma(\boldsymbol{u}^T\boldsymbol{v}^+/\tau) + \log \sigma(-\boldsymbol{u}^T\boldsymbol{v}^-/\tau)$	$(\sigma(-\boldsymbol{u}^T\boldsymbol{v}^+/\tau))/\tau\boldsymbol{v}^+ - \sigma(\boldsymbol{u}^T\boldsymbol{v}^-/\tau)/\tau\boldsymbol{v}^-$
Margin Triplet	$-\max(\boldsymbol{u}^T\boldsymbol{v}^- - \boldsymbol{u}^T\boldsymbol{v}^+ + m, 0)$	$\boldsymbol{v}^+ - \boldsymbol{v}^-$ if $\boldsymbol{u}^T\boldsymbol{v}^+ - \boldsymbol{u}^T\boldsymbol{v}^- < m$ else $\boldsymbol{0}$

Table 2. Negative loss functions and their gradients. All input vectors, i.e. u, v+, v-, are L2 normalized. NT-Xent is an abbreviation for “Normalized Temperature-scaled Cross Entropy”. Different loss functions impose different weightings of positive and negative examples.

表 2 负损失函数及其梯度。所有输入向量，即 u, v+, v-，都是 L2 标准化的。NT-Xent 是“归一化温标交叉熵”的缩写。不同的损失函数会对正反两个例子施加不同的权

重。

Margin	NT-Logi.	Margin (sh)	NT-Logi.(sh)	NT-Xent
50.9	51.6	57.5	57.9	63.9

Table 4. Linear evaluation (top-1) for models trained with different loss functions. "sh" means using semi-hard negative mining.
表 4 使用不同损失函数训练的模型的线性评估（top-1）。"sh"是指使用半硬阴性挖掘法。

ℓ_2 norm?	τ	Entropy	Contrastive acc.	Top 1
Yes	0.05	1.0	90.5	59.7
	0.1	4.5	87.8	64.4
	0.5	8.2	68.2	60.7
	1	8.3	59.1	58.0
No	10	0.5	91.7	57.2
	100	0.5	92.1	57.0

Table 5. Linear evaluation for models trained with different choices of L2 norm and temperature τ for NT-Xent loss. The contrastive distribution is over 4096 examples.
表 5。是否使用 L2 范数和不同的温标τ 选择训练的模型对 NT-Xent 损耗的线性评估。对比分布超过 4096 例。

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

Figure 9 shows the impact of batch size when models are trained for different numbers of epochs. We find that, when the number of training epochs is small (e.g. 100 epochs), larger batch sizes have a significant advantage over the smaller ones. With more training steps/epochs, the gaps between different batch sizes decrease or disappear, provided the batches are randomly resampled. In contrast to supervised learning (Goyal et al., 2017), in contrastive learning, larger batch sizes provide more negative examples, facilitating convergence (i.e. taking fewer epochs and steps for a given accuracy). Training longer also provides more negative examples, improving the results. In Appendix B.1, results with even longer training steps are provided.

5.2 对比学习从更大的批量 batchsize 和更长的训练中获益

图 9 显示了当模型为不同的时期（epochs）数训练时批(batchsize)大小的影响。我们发现，当训练周期(epochs)数较少（例如 100 个训练周期 epochs）时，较大的批量比较小的批量具有显著的优势。随着训练步骤/epochs 的增加，不同批次之间的差距减小或消失，前提是这些批次被随机重新取样。与监督学习（Goyal 等人，2017）比较，在对比学习中，较大的批量提供了更多的负面示例，促进了收敛（即在给定精度下使用更少的时间和步骤）。训练时间越长，负面例子越多，效果越好。附录 B.1 提供了更长培训步骤的结果。

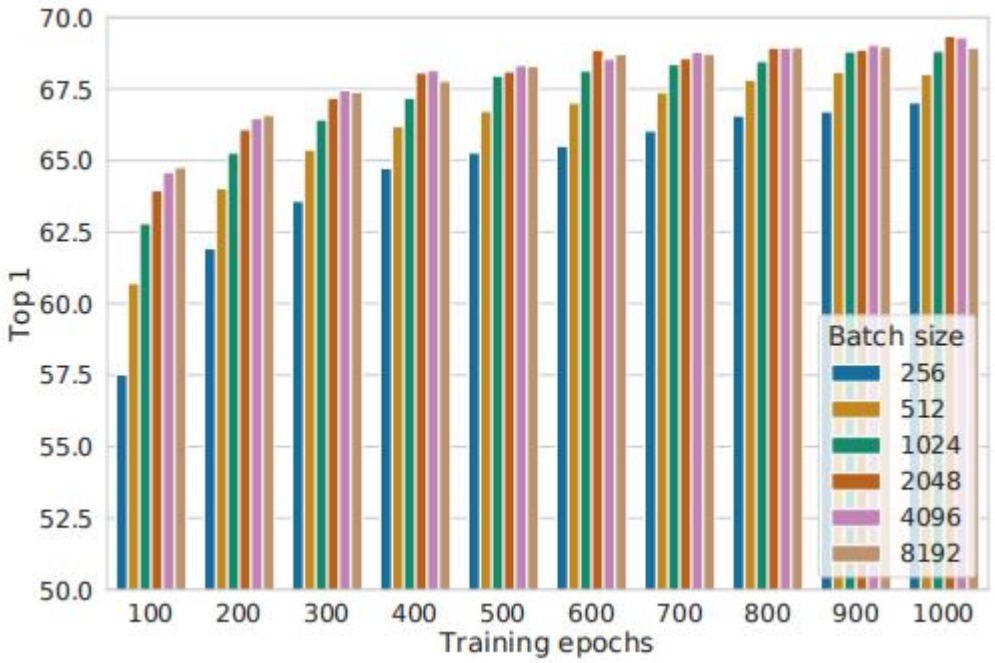


Figure 9. Linear evaluation models (ResNet-50) trained with different batch size and epochs. Each bar is a single run from scratch.
图 9 线性评估模型（ResNet-50）在不同的批次大小和时间下进行训练。每个柱子都是从零开始的运行。

6. Comparison with State-of-the-art
In this subsection, similar to Kolesnikov et al. (2019); He et al. (2019), we use ResNet-50 in 3 different hidden layer widths (width multipliers of 1×, 2×, and 4×). For better convergence, our models here are trained for 1000 epochs.

【6. 与最新技术的比较】
在本小节中，类似于 Kolesnikov 等人。（2019 年）；He 等人。（2019 年），我们在 3 种不同的隐藏层宽度（1×、2×、4×）中使用 ResNet-50。为了更好地融合，我们这里的模型训练了 1000 个 epochs。

Linear evaluation. Table 6 compares our results with previous approaches (Zhuang et al., 2019; He et al., 2019; Misra & van der Maaten, 2019; Hénaff et al., 2019; Kolesnikov et al., 2019; Donahue & Simonyan, 2019; Bachman et al.,2019; Tian et al., 2019) in the linear evaluation setting (see Appendix B.6). Table 1 shows more numerical comparisons among different methods. We are able to use standard networks to obtain substantially better results compared to previous methods that require specifically designed architectures. The best result obtained with our ResNet-50 (4×) can match the supervised pretrained ResNet-50.

线性评估 表 6，在线性评估设置中（见附录 B.6），将我们的结果与以前的方法进行了比较（Zhang 等人，2019 年；He 等人，2019 年；Misra 和 van der Maaten，2019 年；Hénaff 等人，2019 年；Kolesnikov 等人，2019 年；Donahue 和 Simonyan，2019 年；Bachman 等人。，2019；Tian 等人，2019 年）。表 1 显示了不同方法之间的更多数值比较。我们能够使用标准网络来获得比以前需要专门设计的体系

结构更好的结果。用我们的 ResNet-50（4×）得到的最佳结果可以与监督预训练 ResNet-50 相匹配。

Semi-supervised learning. We follow Zhai et al. (2019) and sample 1% or 10% of the labeled ILSVRC-12 training datasets in a class-balanced way (~ 12.8 and ~ 128 images per class respectively). We simply fine-tune the whole base network on the labeled data without regularization (see Appendix B.5). Table 7 shows the comparisons of our results against recent methods (Zhai et al., 2019; Xie et al., 2019; Sohn et al., 2020; Wu et al., 2018; Donahue & Simonyan, 2019; Misra & van der Maaten, 2019; Hénaff et al., 2019). The supervised baseline from (Zhai et al., 2019) is strong due to intensive search of hyper-parameters (including augmentation). Again, our approach significantly improves over state-of-the-art with both 1% and 10% of the labels. Interestingly, fine-tuning our pretrained ResNet-50 (2×, 4×) on full ImageNet are also significantly better then training from scratch (up to 2%, see Appendix B.2).

半监督学习 我们跟随 Zhai et al.（2019 年）并以类别平衡的方式抽取 1%或 10%的 ILSVRC-12 训练数据集（每类分别为 12.8 和 128 幅图像）。我们只需根据标记的数据对整个基础网络进行微调，而无需进行正则化（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 et al.，2019；Xie et al.，2019；Sohn et al.，2020；Wu et al.，2018；Donahue&Simonyan，2019；Misra&van der Maaten，2019；Hénaff et al.，2019）。监督基线来自 Zhai et al.（2019 年）由于对超参数（包括增强）的密集搜索而变得强大。同样，我们的方法明显改进了现有的 1%和 10%的标签。有趣的是，在全 ImageNet 上微调我们预先训练过的 ResNet-50（2×，4×）也明显优于从头开始训练（高达 2%，见附录 B.2）。

Transfer learning. We evaluate transfer learning performance across 12 natural image datasets in both linear evaluation (fixed feature extractor) and fine-tuning settings. Following Kornblith et al. (2019), we perform hyperparameter tuning for each model-dataset combination and select the best hyperparameters on a validation set. Table 8 shows results with the ResNet-50 (4×) model. When fine-tuned, our self-supervised model significantly outperforms the supervised baseline on 5 datasets, whereas the supervised baseline is superior on only 2 (i.e. Pets and Flowers). On the remaining 5 datasets, the models are statistically tied. Full experimental details as well as results with the standard ResNet-50 architecture are provided in Appendix B.8.

迁移学习 我们评估了 12 个自然图像数据集在线性评估（固定特征提取）和微调设置下的迁移学习性能。遵循 Kornblith 等人（2019 年），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。经过微调，我们的自监督模型在 5 个数据集上显著优于监督基线，而监督基线仅在 2 个数据集（即宠物和鲜花）上优于我们的自监督模型。在剩下的 5 个数据集中，模型在统计学上是一致的。完整的实验细节以及标准 ResNet-50 架构的结果在附录 B.8 中提供。

7. Related Work

The idea of making representations of an image agree with each other under small transformations dates back to Becker & Hinton (1992). We extend it by leveraging recent advances in data augmentation, network architecture and contrastive loss. A

similar consistency idea, but for class label prediction, has been explored in other contexts such as semi-supervised learning (Xie et al., 2019; Berthelot et al., 2019).

【7 相关工作】

在小变换下使图像的表示相互一致的想法可以追溯到 Becker&Hinton（1992）。我们利用数据扩充、网络体系结构和对比丢失方面的最新进展来扩展它。类似的一致性思想，但对于类标签预测，已经在诸如半监督学习（Xie et al.，2019；Berthelot et al.，2019）等其他环境中进行了探索。

Handcrafted pretext tasks. The recent renaissance of self-supervised learning began with artificially designed pretext tasks, such as relative patch prediction (Doersch et al., 2015), solving jigsaw puzzles (Noroozi & Favaro, 2016), colorization (Zhang et al., 2016) and rotation prediction (Gidaris et al., 2018; Chen et al., 2019). Although good results can be obtained with bigger networks and longer training (Kolesnikov et al., 2019), these pretext tasks rely on somewhat ad-hoc heuristics, which limits the generality of learned representations.

手工制作的代理任务 最近，自我监督学习的复兴始于人为设计的代理任务，如相对补丁预测（Doersch et al.，2015）、解决拼图难题（Noroozi&Favaro，2016）、色彩化（Zhang et al.，2016）和旋转预测（Gidaris et al.，2018；Chen et al.，2019）。尽管通过更大的网络和更长的训练时间可以获得良好的结果（Kolesnikov 等人，2019 年），这些代理任务依赖于某种程度上的临时启发式，这限制了学习表征的普遍性。

Contrastive visual representation learning. Dating back to Hadsell et al. (2006), these approaches learn representations by contrasting positive pairs against negative pairs. Along these lines, Dosovitskiy et al. (2014) proposes to treat each instance as a class represented by a feature vector (in a parametric form). Wu et al. (2018) proposes to use a memory bank to store the instance class representation vector, an approach adopted and extended in several recent papers (Zhuang et al., 2019; Tian et al., 2019; He et al., 2019; Misra & van der Maaten, 2019). Other work explores the use of in-batch samples for negative sampling instead of a memory bank (Doersch & Zisserman, 2017; Ye et al., 2019; Ji et al., 2019).

Recent literature has attempted to relate the success of their methods to maximization of mutual information between latent representations (Oord et al., 2018; Hénaff et al., 2019; Hjelm et al., 2018; Bachman et al., 2019). However, it is not clear if the success of contrastive approaches is determined by the mutual information, or by the specific form of the contrastive loss (Tschannen et al., 2019).

We note that almost all individual components of our framework have appeared in previous work, although the specific instantiations may be different. The superiority of our framework relative to previous work is not explained by any single design choice, but by their composition. We provide a comprehensive comparison of our design choices with those of previous work in Appendix C.

对比视觉表征学习 可以追溯到 Hadsell 等人。（2006），这些方法通过对比正对和负对来学习表征。沿着这些思路，Dosovitskiy 等人。（2014）建议将每个实例

视为一个由特征向量（以参数形式）表示的类。Wu 等人。（2018）建议使用内存库来存储实例类表示向量，这是最近几篇论文中采用和扩展的方法（庄等人，2019 年；田等，2019 年；他等，2019 年；米斯拉和范德马滕，2019）。其他研究探索了使用批内样本而不是记忆库进行负采样（Doersch&Zisserman，2017；Ye 等人，2019；Ji 等人，2019）。

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

我们注意到，我们的框架中几乎所有单独的组件都出现在以前的工作中，尽管具体的实例化可能不同。与以前的工作相比，我们的框架的优越性不是由任何单一的设计选择来解释的，而是由它们组成的。我们在附录 C 中提供了我们的设计选择与先前工作的综合比较。

8. Conclusion

In this work, we present a simple framework and its instantiation for contrastive visual representation learning. We carefully study its components, and show the effects of different design choices. By combining our findings, we improve considerably over previous methods for self-supervised, semi-supervised, and transfer learning.

Our approach differs from standard supervised learning on ImageNet only in the choice of data augmentation, the use of a nonlinear head at the end of the network, and the loss function. The strength of this simple framework suggests that, despite a recent surge in interest, self-supervised learning remains undervalued.

【8 结论】

在这项工作中，我们提出了一个简单的框架和它的实例对比视觉表征学习。我们仔细研究了它的组成部分，并展示了不同设计选择的效果。通过结合我们的发现，我们在自我监督、半监督和迁移学习方面比以前的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只是在数据扩充的选择，网络末端使用非线性映射头，以及损失函数上。这个简单框架的力量表明，尽管最近人们对自我监督学习的兴趣激增，但它仍然被低估了。

附录还没翻译