A Simple Framework for Contrastive Learning of Visual Representations

一种简单的视觉表征对比学习框架

Ting Chen [1] Simon Kornblith [1] Mohammad Norouzi [1] Geoffrey Hinton [1]
Ting Chen [1] Simon Kornblith [1] Mohammad Norouzi [1] Geoffrey Hinton [1]

Abstract

摘要

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. [1]

本文提出了 SimCLR: 一种简单的视觉表征对比学习框架。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。为了理解是什么使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。我们展示了 (1) 数据增强的组合在定义有效的预测任务中起着关键作用，(2) 在表征和对比损失之间引入可学习的非线性变换显著提高了学习到的表征的质量，以及 (3) 与监督学习相比，对比学习受益于更大的批量大小和更多的训练步骤。通过结合这些发现，我们能够在 ImageNet 上显著超越之前的自监督和半监督学习方法。一个在 SimCLR 学习的自监督表征上训练的线性分类器达到了 76.5% 的 top-1 准确率，相较于之前的最先进技术有 7% 的相对提升，匹配了监督 ResNet-50 的性能。当仅在 1% 的标签上进行微调时，我们达到了 85.8% 的 top-5 准确率，超越了 AlexNet，且使用了 100× 更少的标签。[1]

1. Introduction

1. 引言

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).

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。大多数主流方法可以归入两类: 生成式或判别式。生成式方法学习生成或以其他方式建模输入空间中的像素 (Hinton et al., 2006; Kingma & Welling, 2013; Goodfellow et al., 2014)。

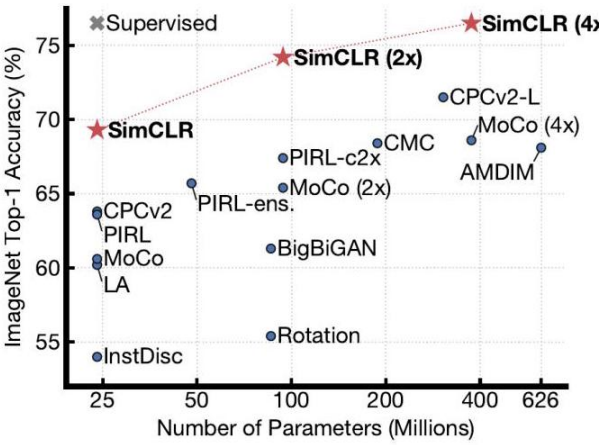


Figure 1. ImageNet Top-1 accuracy of linear classifiers trained on representations learned with different self-supervised methods (pretrained on ImageNet). Gray cross indicates supervised ResNet-50. Our method, SimCLR, is shown in bold.

图 1. 基于不同自监督方法 (在 ImageNet 上预训练) 学习的表示的线性分类器的 ImageNet Top-1 准确率。灰色十字表示监督学习的 ResNet-50。我们的方法 SimCLR 以粗体显示。

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).

然而，像素级生成计算成本高，并且在表示学习中可能并不是必要的。判别方法使用与监督学习相似的目标函数来学习表示，但训练网络执行预文本任务，其中输入和标签均来自未标记的数据集。许多此类方法依赖于启发式设计预文本任务 (Doersch et al., 2015; Zhang et al., 2016; Noroozi & Favaro, 2016; Gidaris et al., 2018)，这可能限制了学习表示的普遍性。基于潜在空间对比学习的判别方法最近显示出巨大的潜力，取得了最先进的结果 (Hadsell et al., 2006; Dosovitskiy et al., 2014; Oord et al., 2018; Bachman et al., 2019)。

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示对比学习框架，我们称之为 SimCLR。SimCLR 不仅超越了之前的工作 (图 1)，而且更简单，不需要专门的架构 (Bachman et al., 2019; Hénaff et al., 2019) 或内存库 (Wu et al., 2018; Tian et al., 2019; He et al., 2019; Misra & van der Maaten, 2019)。

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

为了理解什么使得良好的对比表示学习成为可能，我们系统地研究了我们框架的主要组成部分，并展示了:

- Composition of multiple data augmentation operations is crucial in defining the contrastive prediction tasks that yield effective representations. In addition, unsupervised contrastive learning benefits from stronger data augmentation than supervised learning.

[1] Google Research, Brain Team. Correspondence to: Ting Chen <iamtingchen@google.com>.
[1] Google Research, Brain Team. 通信地址:Ting Chen <iamtingchen@google.com>。
Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.
Proceedings of the 37^{th} 国际机器学习会议，维也纳，奥地利，PMLR 119，2020。
[1] Code available at https://github.com/google-research/simclr.
[1] 代码可在 https://github.com/google-research/simclr 获取。

- 多种数据增强操作的组合在定义有效表示的对比预测任务中至关重要。此外，无监督对比学习受益于比监督学习更强的数据增强。

- Introducing a learnable nonlinear transformation between the representation and the contrastive loss substantially improves the quality of the learned representations.

- 在表示与对比损失之间引入可学习的非线性变换显著提高了学习表示的质量。

- Representation learning with contrastive cross entropy loss benefits from normalized embeddings and an appropriately adjusted temperature parameter.

- 使用对比交叉熵损失的表示学习受益于归一化的嵌入和适当调整的温度参数。

- 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.

- 与其监督学习对应物相比，对比学习受益于更大的批量大小和更长的训练时间。与监督学习一样，对比学习也受益于更深和更宽的网络。

We combine these findings to achieve a new state-of-the-art in self-supervised and semi-supervised learning on Ima-geNet 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，在自监督和半监督学习上实现了 ImageNet ILSVRC-2012(Russakovsky 等，2015) 的新状态-of-the-art。在线性评估协议下，SimCLR 达到了 76.5% 的 top-1 准确率，相较于之前的状态-of-the-art(Hénaff 等，2019) 有 7% 的相对提升。当仅使用 1% 的 ImageNet 标签进行微调时，SimCLR 达到了 85.8% 的 top-5 准确率，相对提升 10%(Hénaff 等，2019)。在其他自然图像分类数据集上进行微调时，SimCLR 在 12 个数据集中有 10 个表现与强监督基线 (Kornblith 等，2019) 相当或更好。

2. Method

2. 方法

2.1.The Contrastive Learning Framework

2.1. 对比学习框架

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.

受到近期对比学习算法的启发 (见第 7 节的概述)，SimCLR 通过在潜在空间中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。正如图 2 所示，该框架包括以下四个主要组件。

- A stochastic data augmentation module that transforms any given data example randomly resulting in two correlated views of the same example, denoted $\widetilde{\mathbf{x}}_i$ and $\widetilde{\mathbf{x}}_j$, 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.

- 一个随机数据增强模块，它随机转换任何给定的数据示例，从而生成同一示例的两个相关视图，记作 $\widetilde{\mathbf{x}}_i$ 和 $\widetilde{\mathbf{x}}_j$ ，我们将其视为一个正对。在本研究中，我们依次应用三种简单的增强: 随机裁剪后调整回原始大小、随机颜色失真和随机高斯模糊。如第 3 节所示，随机裁剪和颜色失真的组合对于实现良好的性能至关重要。

- 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)

- 一个神经网络基础编码器 $f(\cdot)$ ，用于从增强的数据示例中提取表示向量。我们的框架允许在网络架构上进行各种选择，而没有任何限制。我们选择简单性，采用常用的 ResNet(He et al., 2016)。

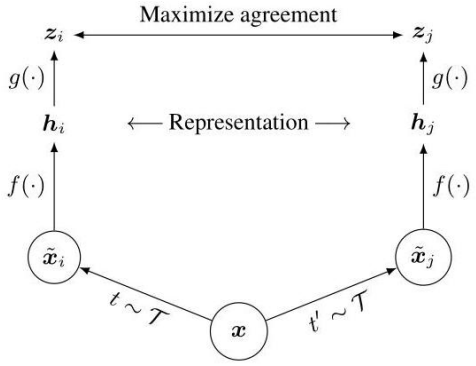


Figure 2. A simple framework for contrastive learning of visual representations. Two separate data augmentation operators are sampled from the same family of augmentations ($t \sim \mathcal{T}$ and $t' \sim \mathcal{T}$) and applied to each data example to obtain two correlated views. A base encoder network $f(\cdot)$ and a projection head $g(\cdot)$ are trained to maximize agreement using a contrastive loss. After training is completed, we throw away the projection head $g(\cdot)$ and use encoder $f(\cdot)$ and representation $\mathbf{h}$ for downstream tasks.

图 2. 一个简单的视觉表示对比学习框架。两个独立的数据增强操作从同一增强家族中采样 ($t \sim \mathcal{T}$ 和 $t' \sim \mathcal{T}$ ，并应用于每个数据示例，以获得两个相关视图。一个基础编码器网络 $f(\cdot)$ 和一个投影头 $g(\cdot)$ 被训练以使用对比损失最大化一致性。训练完成后，我们丢弃投影头 $g(\cdot)$ ，并使用编码器 $f(\cdot)$ 和表示 $\mathbf{h}$ 进行下游任务。

to obtain $\mathbf{h}_i = f(\widetilde{\mathbf{x}}_i) = \text{ResNet}(\widetilde{\mathbf{x}}_i)$ where $\mathbf{h}_i \in \mathbb{R}^d$ is the output after the average pooling layer.

以获得 $\mathbf{h}_i = f(\widetilde{\mathbf{x}}_i) = \text{ResNet}(\widetilde{\mathbf{x}}_i)$ ，其中 $\mathbf{h}_i \in \mathbb{R}^d$ 是平均池化层后的输出。

- A small neural network projection head $g(\cdot)$ that maps representations to the space where contrastive loss is applied. We use a MLP with one hidden layer to obtain $\mathbf{z}_i = g(\mathbf{h}_i) = W^{(2)}\sigma\left(W^{(1)}\mathbf{h}_i\right)$ where σ is a ReLU nonlinearity. As shown in section 4, we find it beneficial to define the contrastive loss on $\mathbf{z}_i$ ' s rather than $\mathbf{h}_i$ ' s.

- 一个小型神经网络投影头 $g(\cdot)$ ，将表示映射到应用对比损失的空间。我们使用一个具有一个隐藏层的多层感知机 (MLP) 来获得 $\mathbf{z}_i = g(\mathbf{h}_i) = W^{(2)}\sigma\left(W^{(1)}\mathbf{h}_i\right)$ ，其中 σ 是 ReLU 非线性激活函数。如第 4 节所示，我们发现将对比损失定义在 $\mathbf{z}_i$ 上而不是 $\mathbf{h}_i$ 上是有益的。

- A contrastive loss function defined for a contrastive prediction task. Given a set $\{\widetilde{\mathbf{x}}_k\}$ including a positive pair of examples $\widetilde{\mathbf{x}}_i$ and $\widetilde{\mathbf{x}}_j$, the contrastive prediction task aims to identify $\widetilde{\mathbf{x}}_j$ in $\{\widetilde{\mathbf{x}}_k\}_{k\neq i}$ for a given $\widetilde{\mathbf{x}}_i$.

- 为对比预测任务定义的对比损失函数。给定一个集合 $\{\widetilde{\mathbf{x}}_k\}$ ，包括一对正例 $\widetilde{\mathbf{x}}_i$ 和 $\widetilde{\mathbf{x}}_j$ ，对比预测任务旨在识别给定 $\widetilde{\mathbf{x}}_i$ 的 $\widetilde{\mathbf{x}}_j$ 在 $\{\widetilde{\mathbf{x}}_k\}_{k\neq i}$ 中的位置。

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 $\text{sim}(\mathbf{u},\mathbf{v}) = \mathbf{u}^\top\mathbf{v}/\parallel\mathbf{u}\parallel\parallel\mathbf{v}\parallel$ denote the dot product between ℓ_2 normalized $\mathbf{u}$ and $\mathbf{v}$ (i.e. cosine similarity). Then the loss function for a positive pair of examples(i, j)is defined as

我们随机抽取一个小批量的 N 示例，并在从小批量派生的增强示例对上定义对比预测任务，从而产生 $2N$ 数据点。我们不显式抽取负示例。相反，给定一对正例，类似于 (Chen et al., 2017)，我们将小批量内的其他 $2(N-1)$ 增强示例视为负示例。设 $\text{sim}(\mathbf{u},\mathbf{v}) = \mathbf{u}^\top\mathbf{v}/\parallel\mathbf{u}\parallel\parallel\mathbf{v}\parallel$ 表示 ℓ_2 归一化 $\mathbf{u}$ 和 $\mathbf{v}$ 之间的点积 (即余弦相似度)。那么一对正例 (i, j) 的损失函数定义为

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

where $\mathbb{1}_{[k\neq i]} \in \{0,1\}$ is an indicator function evaluating to 1 iff $k \neq i$ 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). Algorithm 1 SimCLR's main learning algorithm.

其中 $\mathbb{1}_{[k\neq i]} \in \{0,1\}$ 是一个指示函数，当且仅当 $k \neq i$ 和 τ 时评估为 1，后者表示温度参数。最终损失是在小批量内所有正对 (i, j) 和 (j, i) 上计算的。这个损失在之前的工作中被使用过 (Sohn, 2016; Wu et al., 2018; Oord et al., 2018)；为了方便，我们称之为 NT-Xent(归一化温度缩放交叉熵损失)。算法 1 SimCLR 的主要学习算法。

---

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

Algorithm 1 summarizes the proposed method.
算法 1 总结了所提出的方法。

2.2. Training with Large Batch Size

2.2. 使用大批量训练

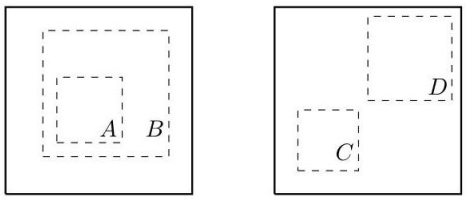
To keep it simple, we do not train the model with a memory bank (Wu et al., 2018; He et al., 2019). Instead, we vary the training batch size N from 256 to 8192. A batch size of 8192 gives us 16382 negative examples per positive pair from both augmentation views. Training with large batch size may be 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]

为了简单起见，我们不使用记忆库来训练模型 (Wu et al., 2018; He et al., 2019)。相反，我们将训练批量大小 N 从 256 变为 8192。批量大小为 8192 时，我们从两个增强视图中为每对正样本提供 16382 个负样本。使用标准 SGD/Momentum 和线性学习率缩放进行大批量训练可能不稳定 (Goyal et al., 2017)。

为了稳定训练，我们对所有批量大小使用 LARS 优化器 (You et al., 2017)。我们使用云 TPU 训练模型，具体使用 32 到 128 个核心，具体取决于批量大小。[2]

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 ResNets 使用批量归一化 (Ioffe & Szegedy, 2015)。在数据并行的分布式训练中，BN 的均值和方差通常在每个设备上局部聚合。在我们的对比学习中，由于正样本对是在同一设备上计算的，模型可以利用局部信息泄漏来提高预测准确性，而不改善表示。我们通过在训练期间对所有设备的 BN 均值和方差进行聚合来解决这个问题。其他方法包括在设备之间打乱数据示例 (He et al., 2019)，或用层归一化替代 BN(Hénaff et al., 2019)。



(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 ($B \rightarrow A$) or adjacent view ($D \rightarrow C$) prediction.

图 3. 实心矩形为图像，虚线矩形为随机裁剪。通过随机裁剪图像，我们采样包括从全局到局部视图 ($B \rightarrow A$) 或相邻视图 ($D \rightarrow C$) 的对比预测任务。

2.3. Evaluation Protocol

2.3. 评估协议

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

在这里，我们制定了我们的实证研究协议，旨在理解我们框架中的不同设计选择。

Dataset and Metrics. Most of our study for unsupervised pretraining (learning encoder network f without labels) is done using the ImageNet ILSVRC-2012 dataset (Rus-sakovsky 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.

数据集和指标。我们对无监督预训练 (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) 进行的。附录 B.9 中可以找到一些关于 CIFAR-10(Krizhevsky & Hinton，2009) 的额外预训练实验。我们还在广泛的数据集上测试预训练结果，以进行迁移学习。为了评估学习到的表示，我们遵循广泛使用的线性评估协议 (Zhang 等，2016；Oord 等，2018；Bachman 等，2019；Kolesnikov 等，2019)，在冻结的基础网络上训练线性分类器，并使用测试准确率作为表示质量的代理。除了线性评估外，我们还与半监督和迁移学习的最新技术进行比较。

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 $4.8\,(= 0.3 \times \text{BatchSize}/256)$ and weight decay of 10^{-6} . We train at batch size 4096 for 100 epochs. [3]

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).

默认设置。除非另有说明，对于数据增强，我们使用随机裁剪和调整大小 (带随机翻转)、颜色失真和高斯模糊 (详细信息见附录 A)。我们使用 ResNet-50 作为基础编码器网络，并使用 2 层 MLP 投影头将表示投影到 128 维潜在空间。作为损失函数，我们使用 NT-Xent，通过 LARS 优化，学习率为 $4.8\,(=0.3\times \text{BatchSize}/256)$，权重衰减为 10^{-6}。我们以批量大小 4096 训练 100 个周期。[3] 此外，我们在前 10 个周期使用线性预热，并使用余弦衰减计划 (不重启) 来衰减学习率 (Loshchilov & Hutter，2016)。

3. Data Augmentation for Contrastive Representation Learning

3. 对比表示学习的数据增强

Data augmentation defines predictive tasks. While data augmentation has been widely used in both supervised and unsupervised representation learning (Krizhevsky et al.,

数据增强定义预测任务。虽然数据增强在监督和无监督表示学习中得到了广泛应用 (Krizhevsky 等，2010)，

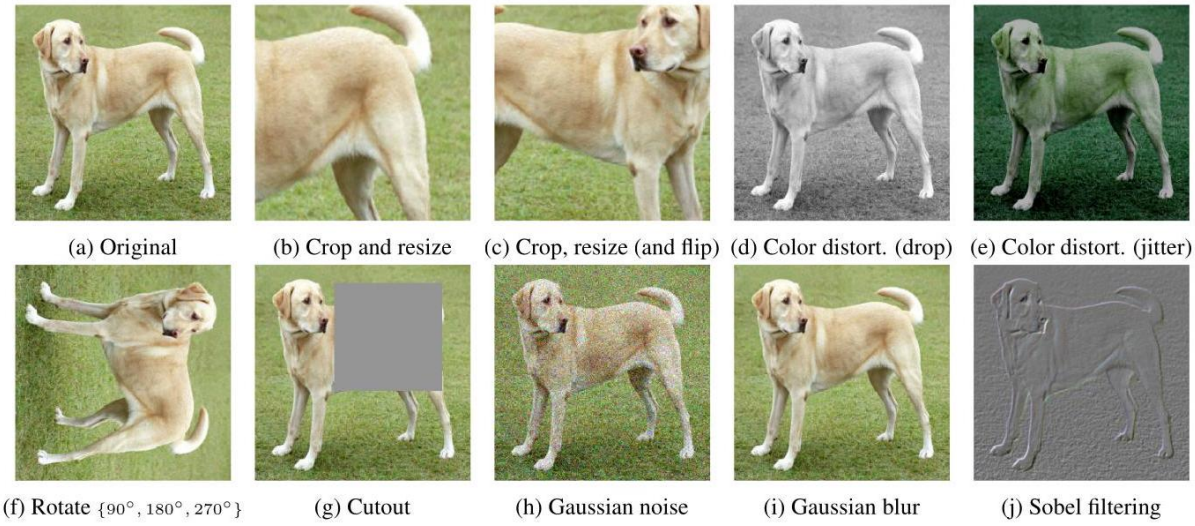


Figure 4. Illustrations of the studied data augmentation operators. Each augmentation can transform data stochastically with some internal parameters (e.g. rotation degree, noise level). Note that we only test these operators in ablation, the augmentation policy used to train our models only includes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

图 4. 研究的数据增强操作的示例。每种增强可以通过一些内部参数 (例如旋转角度、噪声水平) 随机转换数据。请注意，我们仅在消融实验中测试这些操作，用于训练我们模型的增强策略仅包括随机裁剪 (带翻转和调整大小)、颜色失真和高斯模糊。(原始图像 cc-by: Von.grzanka)

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

[2] With 128 TPU v3 cores, it takes $\sim$ 1.5 hours to train our ResNet-50 with a batch size of 4096 for 100 epochs.

[2] 使用 128 个 TPU v3 核心，训练我们的 ResNet-50 需要 $\sim$ 1.5 小时，批量大小为 4096，训练 100 个周期。

[3] Although max performance is not reached in 100 epochs, reasonable results are achieved, allowing fair and efficient ablations.

[3] 尽管在 100 个周期内未达到最大性能，但取得了合理的结果，从而允许公平和高效的消融实验。

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.

尽管在一些研究中 (2012；Hénaff 等，2019；Bachman 等，2019) 没有将其视为定义对比预测任务的系统方法，但许多现有方法通过改变架构来定义对比预测任务。例如，Hjelm 等 (2018)；Bachman 等 (2019) 通过限制网络架构中的感受野实现全局到局部视图预测，而 Oord 等 (2018)；Hénaff 等 (2019) 则通过固定的图像分割程序和上下文聚合网络实现邻近视图预测。我们表明，通过对目标图像进行简单的随机裁剪 (并调整大小)，可以避免这种复杂性，这样可以创建一个包含上述两种任务的预测任务家族，如图 3 所示。这一简单的设计选择方便地将预测任务与神经网络架构等其他组件解耦。通过扩展增强家族并随机组合它们，可以定义更广泛的对比预测任务。

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

3.1. 数据增强操作的组合对学习良好表示至关重要

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 (De-Vries & 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.

为了系统地研究数据增强的影响，我们在这里考虑几种常见的增强方法。一种增强类型涉及数据的空间/几何变换，例如裁剪和调整大小 (包括水平翻转)、旋转 (Gidaris 等，2018) 和 cutout(De-Vries & Taylor，2017)。另一种增强类型涉及外观变换，例如颜色失真 (包括颜色丢失、亮度、对比度、饱和度、色调)(Howard，2013；Szegedy 等，2015)、高斯模糊和 Sobel 滤波。图 4 可视化了我们在本研究中研究的增强方法。

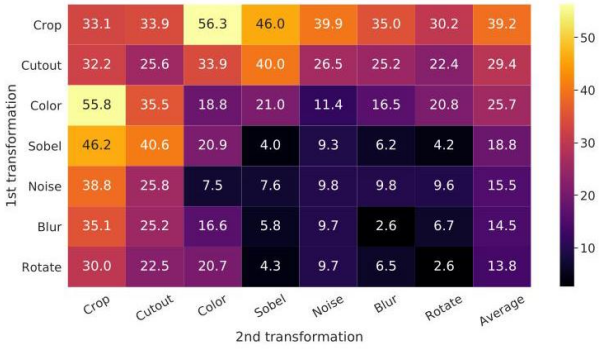


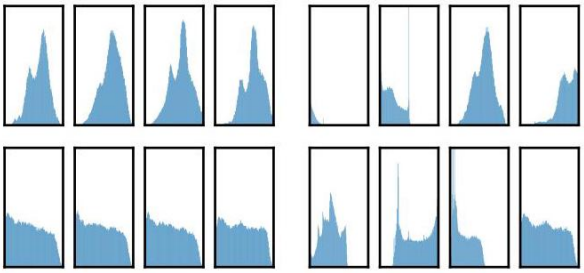
Figure 5. Linear evaluation (ImageNet top-1 accuracy) under individual or composition of data augmentations, applied only to one branch. For all columns but the last, diagonal entries correspond to single transformation, and off-diagonals correspond to composition of two transformations (applied sequentially). The last column reflects the average over the row.

图 5. 在单个或组合数据增强下的线性评估 (ImageNet top-1 准确率)，仅应用于一个分支。除了最后一列外，所有列的对角线条目对应于单一变换，而非对角线条目对应于两个变换的组合 (顺序应用)。最后一列反映了该行的平均值。

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. $t(\mathbf{x}_i) = \mathbf{x}_i$). Note that this asymmet-

为了理解单个数据增强的效果以及增强组合的重要性，我们研究了在单独或成对应用增强时我们框架的性能。由于 ImageNet 图像的大小不同，我们始终应用裁剪和调整大小 (Krizhevsky 等，2012；Szegedy

等，2015)，这使得在没有裁剪的情况下研究其他增强变得困难。为了消除这种混淆，我们考虑了一种不对称的数据变换设置用于此消融实验。具体而言，我们始终首先随机裁剪图像并将其调整为相同的分辨率，然后仅将目标变换应用于图 2 中框架的一个分支，而将另一个分支保持为恒等 (即 $t(\mathbf{x}_i) = \mathbf{x}_i$)。请注意，这种不对称性



(a) Without color distortion. (b) With color distortion.

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. 两幅不同图像的不同裁剪 (即两行) 的像素强度直方图 (在所有通道上)。第一行的图像来自图 4。所有坐标轴的范围相同。

Methods	Color distortion strength					AutoAug
	1/8	1/4	1/2	1	1 (+Blur)	
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

方法	颜色失真强度					AutoAug
	1/8	1/4	1/2	1	1 (+ 模糊)	
SimCLR	59.6	61.0	62.6	63.2	64.5	61.1
监督	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[5] , under varied color distortion strength (see Appendix A) and other data transformations. Strength 1 (+Blur) is our default data augmentation policy.

表 1. 在不同颜色失真强度 (见附录 A) 和其他数据变换下，使用线性评估的无监督 ResNet-50 的 Top-1 准确率，以及监督 ResNet- 50[5] 。强度 1(+ 模糊) 是我们的默认数据增强策略。

ric data augmentation hurts the performance. Nonetheless, this setup should not substantively change the impact of individual data augmentations or their compositions.

过度的数据增强会损害性能。然而，这种设置不应实质性改变单个数据增强或其组合的影响。

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.

图 5 显示了在单独和组合变换下的线性评估结果。我们观察到，没有单一的变换足以学习良好的表示，尽管模型几乎可以完美识别对比任务中的正对。当组合增强时，对比预测任务变得更加困难，但表示的质量显著提高。附录 B.2 提供了对组合更广泛增强集的进一步研究。

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.

一种增强组合脱颖而出: 随机裁剪和随机颜色失真。我们推测，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补丁共享相似的颜色分布。图 6 表明，仅颜色直方图就足以区分图像。神经网络可能利用这一捷径来解决预测任务。因此，将裁剪与颜色失真组合在一起以学习可泛化的特征是至关重要的。

3.2. Contrastive learning needs stronger data augmentation than supervised learning

3.2. 对比学习需要比监督学习更强的数据增强

To further demonstrate the importance of the color augmentation, we adjust the strength of color augmentation as

为进一步证明颜色增强的重要性，我们调整颜色增强的强度为

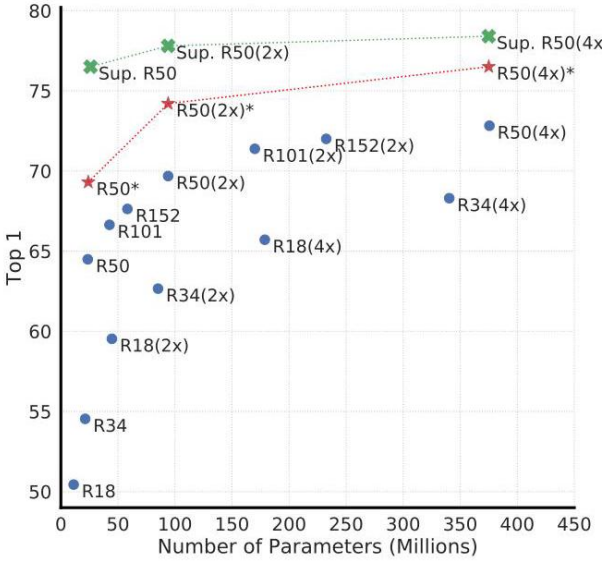


Figure 7. Linear evaluation of models with varied depth and width. Models in blue dots are ours trained for 100 epochs, models in red stars are ours trained for 1000 epochs, and models in green crosses are supervised ResNets trained for 90 epochs [7] (He et al.,2016).

图 7. 不同深度和宽度模型的线性评估。蓝点表示我们训练了 100 个周期的模型，红星表示我们训练了 1000 个周期的模型，绿色交叉表示监督 ResNet 训练了 90 个周期的模型 [7] (He et al., 2016)。

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.

如表 1 所示。更强的颜色增强显著改善了学习到的无监督模型的线性评估。在这种情况下，使用监督学习找到的复杂增强策略 AutoAugment (Cubuk et al., 2019) 的效果不如简单的裁剪 + (更强的) 颜色扭曲。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，甚至损害了它们的性能。因此，我们的实验表明，无监督对比学习比监督学习更能从更强的 (颜色) 数据增强中受益。尽管之前的研究报告称数据增强对自监督学习是有用的 (Doersch et al., 2015; Bachman et al., 2019; Hénaff et al., 2019; Asano et al., 2019)，我们表明，对于监督学习没有带来准确性收益的数据增强，仍然可以显著帮助对比学习。

4. Architectures for Encoder and Head

4. 编码器和头部的架构

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

4.1. 无监督对比学习更能从更大的模型中受益

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，增加深度和宽度都能提高性能，这并不令人惊讶。虽然监督学习也有类似的发现 (He et al., 2016)，但我们发现，随着模型规模的增加，监督模型与训练于无监督模型的线性分类器之间的差距缩小，这表明无监督学习比其监督对应物更能从更大的模型中受益。

Name	Negative loss function	Gradient w.r.t. $\mathbf{u}$
NT-Xent NT-Logistic Margin Triplet	$\mathbf{u}^T\mathbf{v}^+/\tau - \log \sum_{\mathbf{v}\in\{\mathbf{v}^+,\mathbf{v}^-\}} \exp\left(\mathbf{u}^T\mathbf{v}/\tau\right)$ $\log\sigma\left(\mathbf{u}^T\mathbf{v}^+/\tau\right) + \log\sigma\left(-\mathbf{u}^T\mathbf{v}^-/\tau\right)$ $-\max\left(\mathbf{u}^T\mathbf{v}^- - \mathbf{u}^T\mathbf{v}^+ + m, 0\right)$	$\left(1 - \frac{\exp\left(\mathbf{u}^T\mathbf{v}^+/\tau\right)}{Z(\mathbf{u})}\right)/\tau\mathbf{v}^+ - \sum_{\mathbf{v}^-} \frac{\exp\left(\mathbf{u}^T\mathbf{v}^-/\tau\right)}{Z(\mathbf{u})}/\tau\mathbf{v}^-$ $\left(\sigma\left(-\mathbf{u}^T\mathbf{v}^+/\tau\right)\right)/\tau\mathbf{v}^+ - \sigma\left(\mathbf{u}^T\mathbf{v}^-/\tau\right)/\tau\mathbf{v}^- \; \mathbf{v}^+ - \mathbf{v}^-$ if $\mathbf{u}^T\mathbf{v}^+ - \mathbf{u}^T\mathbf{v}^- < m$ else0

名称	负损失函数	关于 $\mathbf{u}$ 的梯度
NT-Xent NT-Logistic 边际三元组	$\mathbf{u}^T\mathbf{v}^+/\tau - \log \sum_{\mathbf{v}\in\{\mathbf{v}^+,\mathbf{v}^-\}} \exp\left(\mathbf{u}^T\mathbf{v}/\tau\right)$ $\log\sigma\left(\mathbf{u}^T\mathbf{v}^+/\tau\right) + \log\sigma\left(-\mathbf{u}^T\mathbf{v}^-/\tau\right)$ $-\max\left(\mathbf{u}^T\mathbf{v}^- - \mathbf{u}^T\mathbf{v}^+ + m, 0\right)$	$\left(1 - \frac{\exp\left(\mathbf{u}^T\mathbf{v}^+/\tau\right)}{Z(\mathbf{u})}\right)/\tau\mathbf{v}^+ - \sum_{\mathbf{v}^-} \frac{\exp\left(\mathbf{u}^T\mathbf{v}^-/\tau\right)}{Z(\mathbf{u})}/\tau\mathbf{v}^-$ $\left(\sigma\left(-\mathbf{u}^T\mathbf{v}^+/\tau\right)\right)/\tau\mathbf{v}^+ - \sigma\left(\mathbf{u}^T\mathbf{v}^-/\tau\right)/\tau\mathbf{v}^- \; \mathbf{v}^+ - \mathbf{v}^-$ 如果 $\mathbf{u}^T\mathbf{v}^+ - \mathbf{u}^T\mathbf{v}^- < m$ 否则 0

Table 2. Negative loss functions and their gradients. All input vectors, i.e. $\mathbf{u}, \mathbf{v}^+, \mathbf{v}^-$, are ℓ_2 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 $\mathbf{u}, \mathbf{v}^+, \mathbf{v}^-$ ，都是 ℓ_2 归一化的。NT-Xent 是 "归一化温度缩放交叉熵" 的缩写。不同的损失函数对正负样本施加不同的权重。

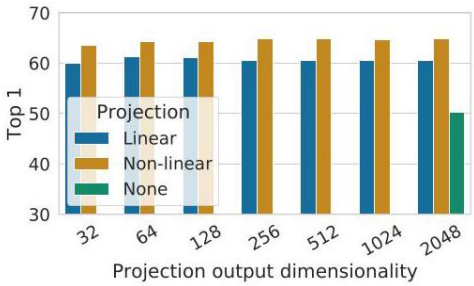


Figure 8. Linear evaluation of representations with different projection heads $g(\cdot)$ and various dimensions of $\mathbf{z} = g(\mathbf{h})$. The representation $\mathbf{h}$ (before projection) is 2048-dimensional here.

图 8. 使用不同投影头 $g(\cdot)$ 和各种维度的 $\mathbf{z} = g(\mathbf{h})$ 进行的表示线性评估。表示 $\mathbf{h}$ (投影前) 在这里是 2048 维的。

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

4.2. 非线性投影头提高了其之前层的表示质量

We then study the importance of including a projection head, i.e. $g(\mathbf{h})$. Figure 8 shows linear evaluation results using three different architecture for the head: (1) identity mapping; (2) linear projection, as

[5] Supervised models are trained for 90 epochs; longer training improves performance of stronger augmentation by $\sim 0.5\%$.

[5] 监督模型训练了 90 个周期；更长的训练通过 $\sim 0.5\%$ 提高了更强增强的性能。

[7] Training longer does not improve supervised ResNets (see Appendix B.3).

[7] 更长时间的训练并不会改善监督的 ResNets(见附录 B.3)。

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, $\mathbf{h}$, is still much better ($> 10\%$) than the layer after, $\mathbf{z} = g(\mathbf{h})$, which shows that the hidden layer before the projection head is a better representation than the layer after.

我们接着研究包含投影头的重要性，即 $g(\mathbf{h})$ 。图 8 显示了使用三种不同架构的头的线性评估结果:(1) 恒等映射；(2) 线性投影，如若干先前方法所用 (Wu et al., 2018)；(3) 默认的非线性投影，具有一个额外的隐藏层 (和 ReLU 激活)，类似于 Bachman et al. (2019)。我们观察到，非线性投影优于线性投影 (+3%) ，并且远远优于没有投影 ($> 10\%$) 。当使用投影头时，无论输出维度如何，观察到的结果都是相似的。此外，即使使用非线性投影，投影头之前的层 $\mathbf{h}$ 仍然比之后的层 $\mathbf{z} = g(\mathbf{h})$ 明显更好，这表明投影头之前的隐藏层是比之后层更好的表示。

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, $\mathbf{z} = g(\mathbf{h})$ 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 $g(\cdot)$, more information can be formed and maintained in $\mathbf{h}$. To verify this hypothesis, we conduct experiments that use either $\mathbf{h}$ or $g(\mathbf{h})$ to learn to predict the transformation applied during the pretraining. Here we set $g(h) = W^{(2)}\sigma\left(W^{(1)}h\right)$, with the same input and output dimensionality (i.e. 2048). Table 3 shows $\mathbf{h}$ contains much more information about the transformation applied, while $g(\mathbf{h})$ loses information. Further analysis can

我们推测，在非线性投影之前使用表示的重要性是由于对比损失引起的信息丢失。特别地，$\mathbf{z} = g(\mathbf{h})$ 被训练为对数据变换不变。因此，g 可能会去除对下游任务有用的信息，例如物体的颜色或方向。通过利用非线性变换 $g(\cdot)$ ，可以在 $\mathbf{h}$ 中形成并保持更多的信息。为了验证这一假设，我们进行实验，使用 $\mathbf{h}$ 或 $g(\mathbf{h})$ 来学习预测在预训练期间施加的变换。这里我们设置 $g(h) = W^{(2)}\sigma\left(W^{(1)}h\right)$ ，具有相同的输入和输出维度 (即 2048)。表 3 显示 $\mathbf{h}$ 包含了关于施加的变换的更多信息，而 $g(\mathbf{h})$ 则丢失了信息。进一步的分析可以

What to predict?	Random guess	Representation	
		h	$g(\mathbf{h})$
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

预测什么?	随机猜测	表示	
		h	$g(\mathbf{h})$
彩色与灰度	80	99.3	97.4
旋转	25	67.6	25.6
原始与损坏	50	99.5	59.6
原始与 Sobel 过滤	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 $\mathbf{h}$ and $g(\mathbf{h})$ are of the same dimensionality, i.e. 2048.

表 3. 在不同表示上训练额外的 MLP 以预测施加的变换的准确性。除了裁剪和颜色增强外，我们在预训练期间额外且独立地添加了旋转 ($\{0°, 90°, 180°, 270°\}$ 的一种)、高斯噪声和 Sobel 滤波变换，最后三行均如此。$\mathbf{h}$ 和 $g(\mathbf{h})$ 的维度相同，即 2048。

be found in Appendix B.4.

可以在附录 B.4 中找到。

5. Loss Functions and Batch Size

5. 损失函数和批量大小

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

5.1. 可调温度的归一化交叉熵损失优于其他替代方案

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) ℓ_2 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).

我们将 NT-Xent 损失与其他常用的对比损失函数进行比较，例如逻辑损失 (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输入的梯度。观察梯度，我们发现 1) ℓ_2 归一化 (即余弦相似度) 以及温度有效地加权不同的示例，适当的温度可以帮助模型从困难的负样本中学习；2) 与交叉熵不同，其他目标函数并不根据相对难度对负样本进行加权。因此，必须对这些损失函数应用半难负样本挖掘 (Schroff et al., 2015): 可以使用半难负样本 (即那些在损失边际内且距离最近，但比正样本更远的样本) 来计算梯度，而不是对所有损失项计算梯度。

To make the comparisons fair, we use the same ℓ_2 normalization for all loss functions, and we tune the hyperparameters, and report their best results. [8] Table 4 shows that, while (semi-hard) negative mining helps, the best result is still much worse than our default NT-Xent loss.

为了使比较公平，我们对所有损失函数使用相同的 ℓ_2 归一化，并调整超参数，报告其最佳结果。[8] 表 4 显示，虽然 (半难) 负样本挖掘有帮助，但最佳结果仍然远不如我们默认的 NT-Xent 损失。

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

边缘	NT-逻辑	边缘 (sh)	NT-逻辑 (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" 表示使用半难负样本挖掘。

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

ℓ_2 范数?	τ	熵	对比准确率	前 1
是	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
否	10	0.5	91.7	57.2
	100	0.5	92.1	57.0

[8] Details can be found in Appendix B.10. For simplicity, we only consider the negatives from one augmentation view.

[8] 详细信息见附录 B.10。为简化起见，我们仅考虑来自一个增强视图的负样本。

Table 5. Linear evaluation for models trained with different choices of ℓ_2 norm and temperature τ for NT-Xent loss. The contrastive distribution is over 4096 examples.

表 5. 使用不同选择的 ℓ_2 规范和温度 τ 对 NT-Xent 损失进行线性评估。对比分布覆盖 4096 个示例。

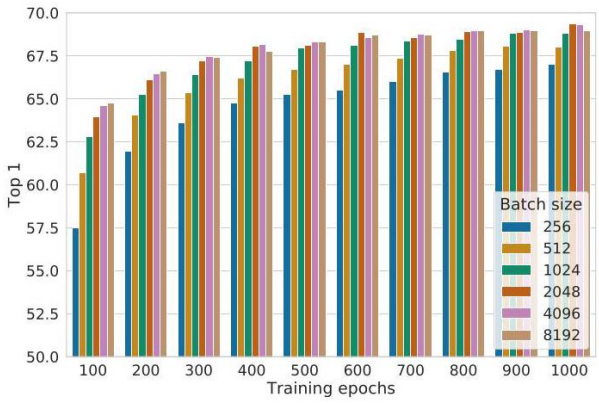


Figure 9. Linear evaluation models (ResNet-50) trained with different batch size and epochs. Each bar is a single run from scratch. [10]

图 9. 使用不同批量大小和训练轮次的线性评估模型 (ResNet-50)。每个柱状图代表从头开始的单次运行。[10]

We next test the importance of the ℓ_2 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 ℓ_2 normalization, the contrastive task accuracy is higher, but the resulting representation is worse under linear evaluation.

我们接下来测试 ℓ_2 归一化 (即余弦相似度与点积) 和温度 τ 在我们默认的 NT-Xent 损失中的重要性。表 5 显示，在没有归一化和适当温度缩放的情况下，性能显著下降。没有 ℓ_2 归一化时，对比任务的准确性更高，但在线性评估下得到的表示效果更差。

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

5.2. 对比学习从更大的批量大小和更长的训练中获益 (更多)

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.

图 9 显示了在不同训练轮数下批量大小的影响。我们发现，当训练轮数较少 (例如 100 轮) 时，较大的批量大小相较于较小的批量大小具有显著优势。随着训练步骤/轮数的增加，不同批量大小之间的差距减少或消失，前提是批次被随机重采样。与监督学习 (Goyal 等，2017) 相比，在对比学习中，较大的批量大小提供了更多的负例，促进了收敛 (即在给定准确性下需要更少的轮数和步骤)。更长的训练也提供了更多的负例，从而改善了结果。在附录 B.1 中提供了更长训练步骤的结果。

Method	Architecture	Param (M)	Top 1	Top 5
Methods using ResNet-50:				
Local Agg.	ResNet-50	24	60.2	-
MoCo	ResNet-50	24	60.6	-
PIRL	ResNet-50	24	63.6	-
CPC v2	ResNet-50	24	63.8	85.3
SimCLR (ours)	ResNet-50	24	69.3	89.0
Methods using other architectures:				
Rotation	RevNet-50 (4×)	86	55.4	-
BigBiGAN	RevNet-50 (4×)	86	61.3	81.9
AMDIM	Custom-ResNet	626	68.1	-
CMC	ResNet-50 (2×)	188	68.4	88.2
MoCo	ResNet-50 (4×)	375	68.6	-
CPC v2	ResNet-161 (*)	305	71.5	90.1
SimCLR (ours)	ResNet-50 (2×)	94	74.2	92.0
SimCLR (ours)	ResNet-50 (4×)	375	76.5	93.2

方法	架构	参数 (M)	前 1	前 5
使用 ResNet-50 的方法:				
局部聚合	ResNet-50	24	60.2	-
MoCo	ResNet-50	24	60.6	-
PIRL	ResNet-50	24	63.6	-
CPC v2	ResNet-50	24	63.8	85.3
SimCLR(我们的)	ResNet-50	24	69.3	89.0
使用其他架构的方法:				
旋转	RevNet-50 (4×)	86	55.4	-
BigBiGAN	RevNet-50 (4×)	86	61.3	81.9
AMDIM	自定义 ResNet	626	68.1	-
CMC	ResNet-50 (2×)	188	68.4	88.2
MoCo	ResNet-50 (4×)	375	68.6	-
CPC v2	ResNet-161 (*)	305	71.5	90.1
SimCLR(我们的)	ResNet-50 (2×)	94	74.2	92.0
SimCLR(我们的)	ResNet-50 (4×)	375	76.5	93.2

Table 6. ImageNet accuracies of linear classifiers trained on representations learned with different self-supervised methods.

表 6. 使用不同自监督方法学习的表示上训练的线性分类器在 ImageNet 上的准确性。

Method	Architecture	Label fraction	
		1%	10%
		Top 5	
Supervised baseline	ResNet-50	48.4	80.4
Methods using other label-propagation:			
Pseudo-label	ResNet-50	51.6	82.4
VAT+Entropy Min.	ResNet-50	47.0	83.4
UDA (w. RandAug)	ResNet-50	-	88.5
FixMatch (w. RandAug)	ResNet-50	-	89.1
S4L (Rot+VAT+En. M.)	ResNet-50 (4×)	-	91.2
Methods using representation learning only:			
InstDisc	ResNet-50	39.2	77.4
BigBiGAN	RevNet-50 (4×)	55.2	78.8
PIRL	ResNet-50	57.2	83.8
CPC v2	ResNet-161(*)	77.9	91.2
SimCLR (ours)	ResNet-50	75.5	87.8
SimCLR (ours)	ResNet-50 (2×)	83.0	91.2
SimCLR (ours)	ResNet-50 (4×)	85.8	92.6

方法	架构	标签比例	
		1%	10%
		前五名	
监督基线	ResNet-50	48.4	80.4
使用其他标签传播的方法:			
伪标签	ResNet-50	51.6	82.4
VAT+ 熵最小化	ResNet-50	47.0	83.4
UDA(带随机增强)	ResNet-50	-	88.5
FixMatch (带随机增强)	ResNet-50	-	89.1
S4L (旋转 +VAT+ 增强学习)	ResNet-50 (4×)	-	91.2
仅使用表示学习的方法:			
InstDisc	ResNet-50	39.2	77.4
BigBiGAN	RevNet-50 (4×)	55.2	78.8
PIRL	ResNet-50	57.2	83.8
CPC v2	ResNet-161(*)	77.9	91.2
SimCLR (我们的)	ResNet-50	75.5	87.8
SimCLR (我们的)	ResNet-50 (2×)	83.0	91.2
SimCLR (我们的)	ResNet-50 (4×)	85.8	92.6

Table 7. ImageNet accuracy of models trained with few labels.
表 7. 使用少量标签训练的模型在 ImageNet 上的准确性。

6. Comparison with State-of-the-art

6. 与最先进技术的比较

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 $1\times, 2\times$, and $4\times$). For better convergence, our models here are trained for 1000 epochs.

在本小节中，类似于 Kolesnikov 等 (2019)；He 等 (2019)，我们在 3 种不同的隐藏层宽度 (宽度乘数为 $1\times, 2\times$ 和 $4\times$) 下使用 ResNet-50。为了更好的收敛，我们在这里将模型训练了 1000 轮。

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.,

线性评估。表 6 将我们的结果与之前的方法进行了比较 (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) 在线性评估设置中 (见附录 B.6)。表 1 显示了不同方法之间更多的数值比较。我们能够使用标准网络获得显著优于需要特定设计架构的先前方法的结果。我们使用的 ResNet-50 [latex0] 所获得的最佳结果可以与监督预训练的 ResNet-50 相匹配。

	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
SimCLR (ours)	76.9	95.3	80.2	48.4	65.9	60.0	61.2	84.2	78.9	89.2	93.9	95.0
Supervised	75.2	95.7	81.2	56.4	64.9	68.8	63.8	83.8	78.7	92.3	94.1	94.2
Fine-tuned:												
SimCLR (ours)	89.4	98.6	89.0	78.2	68.1	92.1	87.0	86.6	77.8	92.1	94.1	97.6
Supervised	88.7	98.3	88.7	77.8	67.0	91.4	88.0	86.5	78.8	93.2	94.2	98.0
Random init	88.3	96.0	81.9	77.0	53.7	91.3	84.8	69.4	64.1	82.7	72.5	92.5

	食物	CIFAR10	CIFAR100	鸟类快照	SUN397	汽车	飞机	VOC2007	DTD	宠物	Caltech-101	花卉
线性评估:												
SimCLR(我们的)	76.9	95.3	80.2	48.4	65.9	60.0	61.2	84.2	78.9	89.2	93.9	95.0
监督学习	75.2	95.7	81.2	56.4	64.9	68.8	63.8	83.8	78.7	92.3	94.1	94.2
微调:												
SimCLR(我们的)	89.4	98.6	89.0	78.2	68.1	92.1	87.0	86.6	77.8	92.1	94.1	97.6
监督学习	88.7	98.3	88.7	77.8	67.0	91.4	88.0	86.5	78.8	93.2	94.2	98.0
随机初始化	88.3	96.0	81.9	77.0	53.7	91.3	84.8	69.4	64.1	82.7	72.5	92.5

[10] A linear learning rate scaling is used here. Figure B. 1 shows using a square root learning rate scaling can improve performance of ones with small batch sizes.
[10] 这里使用了线性学习率缩放。图 B.1 显示，使用平方根学习率缩放可以提高小批量大小模型的性能。

Table 8. Comparison of transfer learning performance of our self-supervised approach with supervised baselines across 12 natural image classification datasets, for ResNet-50 (4×) models pretrained on ImageNet. Results not significantly worse than the best ($p > 0.05$, permutation test) are shown in bold. See Appendix B. 8 for experimental details and results with standard ResNet-50.

表 8。我们自监督方法与监督基线在 12 个自然图像分类数据集上的迁移学习性能比较，针对在 ImageNet 上预训练的 ResNet-50 (4×) 模型。结果显示在粗体中，表示与最佳结果 (($p > 0.05$ ，置换检验) 没有显著差异。有关实验细节和标准 ResNet-50 的结果，请参见附录 B.8。

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.

2019; Tian et al., 2019) 在线性评估设置中 (见附录 B.6)。表 1 显示了不同方法之间更多的数值比较。我们能够使用标准网络获得显著优于需要特定设计架构的先前方法的结果。我们使用的 ResNet-50 (4×) 所获得的最佳结果可以与监督预训练的 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). [11] 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 hyperparameters (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 ($2\times, 4\times$) on full ImageNet are also significantly better then training from scratch (up to 2% , see Appendix B.2).

半监督学习。我们遵循 Zhai 等人 (2019) 的研究，以类别平衡的方式对标记的 ILSVRC-12 训练数据集进行采样 1% 或 10% (每个类别分别为 ~ 12.8 和 ~ 128 图像)。[11] 我们仅在标记数据上对整个基础网络进行微调，而不进行正则化 (见附录 B.5)。表 7 显示了我们结果与近期方法的比较 (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) 的监督基线由于对超参数 (包括增强) 的密集搜索而表现强劲。再次强调，我们的方法在使用 1% 和 10% 标签时显著优于最先进的技术。有趣的是，在完整的 ImageNet 上微调我们预训练的 ResNet-50 ($2\times, 4\times$) 的效果也显著优于从头开始训练 (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。我们在 12 个自然图像数据集上评估迁移学习的性能，包括线性评估 (固定特征提取器) 和微调设置。遵循 Kornblith 等人 (2019) 的研究，我们对每个模型-数据集组合进行超参数调优，并在验证集上选择最佳超参数。表 8 展示了使用 ResNet-50 (4×) 模型的结果。当进行微调时，我们的自监督模型在 5 个数据集上显著优于监督基线，而监督基线仅在 2 个数据集 (即宠物和花卉) 上表现更佳。在其余 5 个数据集中，模型的表现统计上是平局的。完整的实验细节以及使用标准 ResNet-50 架构的结果在附录 B.8 中提供。

7. Related Work

7. 相关工作

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).

使图像的表示在小变换下相互一致的想法可以追溯到 Becker 和 Hinton(1992)。我们通过利用数据增强、网络架构和对比损失的最新进展来扩展这一思想。类似的一致性思想，但用于类别标签预测，已在其他上下文中探讨，例如半监督学习 (Xie 等，2019；Berthelot 等，2019)。

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.

手工设计的前置任务。自监督学习的最近复兴始于人工设计的前置任务，例如相对补丁预测 (Doersch 等，2015)、拼图解决 (Noroozi 和 Favaro，2016)、上色 (Zhang 等，2016) 和旋转预测 (Gidaris 等，2018；Chen 等，2019)。尽管使用更大的网络和更长的训练可以获得良好的结果 (Kolesnikov 等，2019)，但这些前置任务依赖于某种程度的临时启发式方法，这限制了学习表示的普遍性。

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).

对比视觉表征学习。早在 Hadsell 等人 (2006) 的研究中，这些方法通过将正对比对与负对比对进行对比来学习表征。在此基础上，Dosovitskiy 等人 (2014) 提出将每个实例视为由特征向量 (以参数形式表示) 表示的一个类。Wu 等人 (2018) 建议使用内存库来存储实例类表征向量，这一方法在几篇近期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)。其他研究则探索了使用批内样本进行负采样，而不是使用内存库 (Doersch & Zisserman，2017；Ye 等人，2019；Ji 等人，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.

最近的文献试图将其方法的成功与潜在表征之间互信息的最大化联系起来 (Oord 等人，2018；Hénaff 等人，2019；Hjelm 等人，2018；Bachman 等人，2019)。然而，目前尚不清楚对比方法的成功是由互信息决定，还是由对比损失的具体形式决定 (Tschannen 等人，2019)。我们注意到，我们框架的几乎所有单独组件都出现在之前的工作中，尽管具体的实例化可能有所不同。我们框架相对于之前工作的优越性并不是由任何单一设计选择解释的，而是由它们的组合所决定。我们在附录 C 中提供了我们设计选择与之前工作的全面比较。

8. Conclusion

8. 结论

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.

在本研究中，我们提出了一个简单的框架及其实例，用于对比视觉表示学习。我们仔细研究了其组成部分，并展示了不同设计选择的影响。通过结合我们的发现，我们在自监督、半监督和迁移学习的先前方法上有了显著的改进。

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.

[11] The details of sampling and exact subsets can be found in

[11] 采样和确切子集的细节可以在 https://www.tensorflow.org/datasets/catalog/imagenet2012_subset.

我们的方法与在 ImageNet 上的标准监督学习的不同之处仅在于数据增强的选择、网络末端使用非线性头以及损失函数。这个简单框架的强大表明，尽管最近对自监督学习的兴趣激增，但它仍然被低估。

Acknowledgements

致谢

We would like to thank Xiaohua Zhai, Rafael Müller and Yani Ioannou for their feedback on the draft. We are also grateful for general support from Google Research teams in Toronto and elsewhere.

我们要感谢 Xiaohua Zhai、Rafael Müller 和 Yani Ioannou 对草稿的反馈。我们也感谢 Google Research 在多伦多及其他地方的团队提供的支持。

References

参考文献

Asano, Y. M., Rupprecht, C., and Vedaldi, A. A critical analysis of self-supervision, or what we can learn from a single image. arXiv preprint arXiv:1904.13132, 2019.

Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning representations by maximizing mutual information across views. In Advances in Neural Information Processing Systems, pp. 15509–15519, 2019.

Becker, S. and Hinton, G. E. Self-organizing neural network that discovers surfaces in random-dot stereograms. Nature, 355 (6356):161-163, 1992.

Berg, T., Liu, J., Lee, S. W., Alexander, M. L., Jacobs, D. W., and Belhumeur, P. N. Birdsnap: Large-scale fine-grained visual categorization of birds. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2019-2026. IEEE, 2014.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. Mixmatch: A holistic approach to semi-supervised learning. In Advances in Neural Information Processing Systems, pp. 5050-5060, 2019.

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101-mining discriminative components with random forests. In European conference on computer vision, pp. 446-461. Springer, 2014.

Chen, T., Sun, Y., Shi, Y., and Hong, L. On sampling strategies for neural network-based collaborative filtering. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 767-776, 2017.

Chen, T., Zhai, X., Ritter, M., Lucic, M., and Houlsby, N. Self-supervised gans via auxiliary rotation loss. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12154-12163, 2019.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3606- 3613. IEEE, 2014.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation strategies from data. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 113-123, 2019.

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552, 2017.

Doersch, C. and Zisserman, A. Multi-task self-supervised visual learning. In Proceedings of the IEEE International Conference on Computer Vision, pp. 2051-2060, 2017.

Doersch, C., Gupta, A., and Efros, A. A. Unsupervised visual representation learning by context prediction. In Proceedings of the IEEE International Conference on Computer Vision, pp. 1422-1430, 2015.

Donahue, J. and Simonyan, K. Large scale adversarial representation learning. In Advances in Neural Information Processing Systems, pp. 10541-10551, 2019.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. Decaf: A deep convolutional activation feature for generic visual recognition. In International Conference on Machine Learning, pp. 647-655, 2014.

Dosovitskiy, A., Springenberg, J. T., Riedmiller, M., and Brox, T. Discriminative unsupervised feature learning with convolutional neural networks. In Advances in neural information processing systems, pp. 766-774, 2014.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. International Journal of Computer Vision, 88(2):303-338, 2010.

Fei-Fei, L., Fergus, R., and Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop on Generative-Model Based Vision, 2004.

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728, 2018.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In Advances in neural information processing systems, pp. 2672-2680, 2014.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677, 2017.

Hadsell, R., Chopra, S., and LeCun, Y. Dimensionality reduction by learning an invariant mapping. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR' 06), volume 2, pp. 1735-1742. IEEE, 2006.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. arXiv preprint arXiv:1911.05722, 2019.

Hénaff, O. J., Razavi, A., Doersch, C., Eslami, S., and Oord, A. v. d. Data-efficient image recognition with contrastive predictive coding. arXiv preprint arXiv:1905.09272, 2019.

Hinton, G. E., Osindero, S., and Teh, Y.-W. A fast learning algorithm for deep belief nets. Neural computation, 18(7):1527- 1554, 2006.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670, 2018.

Howard, A. G. Some improvements on deep convolutional neural network based image classification. arXiv preprint arXiv:1312.5402, 2013.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 2015.

Ji, X., Henriques, J. F., and Vedaldi, A. Invariant information clustering for unsupervised image classification and segmentation. In Proceedings of the IEEE International Conference on Computer Vision, pp. 9865-9874, 2019.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

Kolesnikov, A., Zhai, X., and Beyer, L. Revisiting self-supervised visual representation learning. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1920-1929, 2019.

Kornblith, S., Shlens, J., and Le, Q. V. Do better ImageNet models transfer better? In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2661-2671, 2019.

Krause, J., Deng, J., Stark, M., and Fei-Fei, L. Collecting a large-scale dataset of fine-grained cars. In Second Workshop on Fine-Grained Visual Categorization, 2013.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/kriz/ learning-features-2009-TR.pdf.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pp. 1097-1105, 2012.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983, 2016.

Maaten, L. v. d. and Hinton, G. Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579-2605, 2008.

Maji, S., Kannala, J., Rahtu, E., Blaschko, M., and Vedaldi, A. Fine-grained visual classification of aircraft. Technical report, 2013.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

Misra, I. and van der Maaten, L. Self-supervised learning of pretext-invariant representations. arXiv preprint arXiv:1912.01991, 2019.

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In Computer Vision, Graphics & Image Processing, 2008. ICVGIP'08. Sixth Indian Conference on, pp.

722-729. IEEE, 2008.

Noroozi, M. and Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In European Conference on Computer Vision, pp. 69-84. Springer, 2016.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. Cats and dogs. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3498-3505. IEEE, 2012.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3):211-252, 2015.

Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815-823, 2015.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. In Advances in neural information processing systems, pp. 1857-1865, 2016.

Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., and Raffel, C. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. arXiv preprint arXiv:2001.07685, 2020.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9, 2015.

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. arXiv preprint arXiv:1906.05849, 2019.

Tschannen, M., Djolonga, J., Rubenstein, P. K., Gelly, S., and Lucic, M. On mutual information maximization for representation learning. arXiv preprint arXiv:1907.13625, 2019.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3733-3742, 2018.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3485-3492. IEEE, 2010.

Xie, Q., Dai, Z., Hovy, E., Luong, M.-T., and Le, Q. V. Unsupervised data augmentation. arXiv preprint arXiv:1904.12848, 2019.

Ye, M., Zhang, X., Yuen, P. C., and Chang, S.-F. Unsupervised embedding learning via invariant and spreading instance feature. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6210-6219, 2019.

You, Y., Gitman, I., and Ginsburg, B. Large batch training of convolutional networks. arXiv preprint arXiv:1708.03888, 2017.

Zhai, X., Oliver, A., Kolesnikov, A., and Beyer, L. S41: Self-supervised semi-supervised learning. In The IEEE International Conference on Computer Vision (ICCV), October 2019.

Zhang, R., Isola, P., and Efros, A. A. Colorful image colorization. In European conference on computer vision, pp. 649-666. Springer, 2016.

Zhuang, C., Zhai, A. L., and Yamins, D. Local aggregation for unsupervised learning of visual embeddings. In Proceedings of the IEEE International Conference on Computer Vision, pp. 6002-6012, 2019.