Small Data Challenges in Big Data Era: A Survey of Recent Progress on Unsupervised and Semi-Supervised Methods

**摘要**

小数据挑战已经出现在许多学习问题中，因为深层神经网络的成功往往依赖于大量的标记数据的可用性，而这些标记数据的收集成本很高。为了解决这一问题，人们在无监督和半监督的方式下，对具有小数据的复杂模型进行了大量的训练。本文将回顾这两大类方法的最新进展。大量的小数据模型将被归类到一个大的图景中，在这里我们将展示它们如何相互作用以激发新思想的探索。我们将回顾学习变换等变、分离、自监督和半监督表示的标准，这是最近发展的基础。许多无监督和半监督生成模型的实例都是基于这些标准开发的，通过探索未标记数据的分布以获得更强大的表示，极大地扩展了现有自动编码器、生成对抗网（GANs）和其他深层网络的领域。在我们关注无监督和半监督方法的同时，我们还将对其他新出现的课题进行更广泛的回顾，从无监督和半监督域自适应到变换等价性和不变性在训练广谱深网络中的基本作用。我们不可能写一本独家的百科全书来收录所有相关的作品。相反，我们的目标是探索这一领域的主要思想、原则和方法，以揭示我们在应对大数据时代的小数据挑战的道路上的前进方向。

**索引项**--小数据方法，无监督方法，半监督方法，域自适应，变换等变表示，分离表示，生成模型，自动编码器，生成对抗网络，自动回归模型，基于流的生成模型，变压器，自监督方法，教授student模型。

**1.概述**

本文旨在全面综述无监督和半监督方法的最新进展，以应对在大量无标记数据可用的情况下，使用少量标记数据的训练模型所面临的挑战。深度学习的成功往往取决于大量标记数据的可用性，在这些数据中，数百万幅图像被标记为训练深度神经网络[1]，[2]，以使这些模型与人类的性能相当甚至超过人类的性能。

然而，在许多情况下，要收集足够多的标记数据是一个挑战，这激发了许多研究工作去探索标记数据之外的无监督信息，以训练各种学习任务的鲁棒模型。

无标签数据。虽然标记数据的数量非常少，但未标记数据可能非常大。这些未标记数据的分布提供了学习健壮表示的重要线索，这些表示可推广到新的学习任务。未标记的数据可以以无监督和半监督的方式利用，取决于是否利用附加的标记示例来训练模型。无标记数据还可以帮助模型缩小不同任务之间的域差距，这导致了一大类无监督和半监督的域自适应方法。

辅助任务。辅助任务也可以作为辅助信息的一个重要来源来缓解小数据问题。例如，相关任务可以是与目标任务相关的不相交概念集上的学习问题。这属于零样本学习（ZSL）和小样本学习（FSL）问题的范畴。广义上，ZSL问题可以看作是一个无监督学习问题，目标任务上没有标记实例，而FSL是半监督的，标记数据很少。两者都旨在传递语义知识或学习知识(例如从源任务到目标任务的元学习。

本研究的重点是探讨无监督与半监督的方法，藉由探讨未标记的范例来解决小资料问题。虽然我们不会回顾利用辅助任务信息的ZSL和FSL方法，但是从一个大的角度来看所有这些方法对我们都是有益的。这将使我们更好地了解我们所处的位置，征服小数据挑战的旅程。利用各种信息源的不同方法导致了广泛的学习方法，从不同的角度解决小数据挑战，如图1所示。

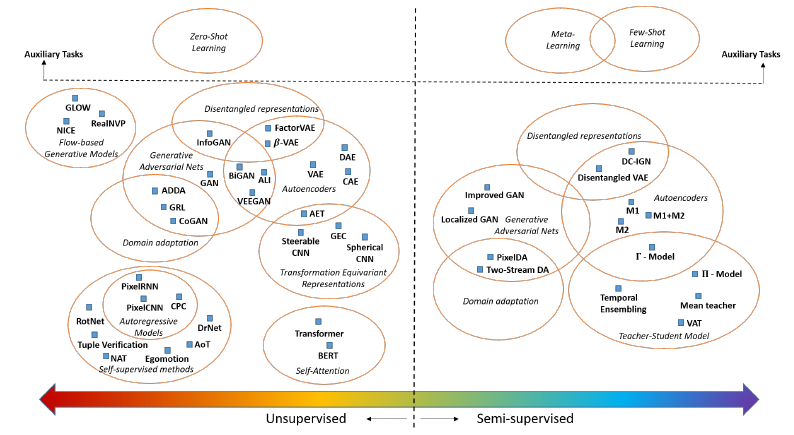


图1：小数据方法的概况。此图显示了不同方法之间的关系，以及它们相互交叉的位置。请参考图2了解本次调查中这些方法的分类

**1.1无监督方法**

在频谱的最左端是无监督的方法，训练时没有标记数据。这些无监督的方法寻求学习表征，这些表征具有足够的泛化性，以适应未来的各种学习任务。在这种情况下，从无监督方法学习到的表示通常是基于这些表示之上的下游分类任务的性能来评估的。各种原理和模型都致力于训练无监督的表示。如图2所示，我们将从几个不同的角度审查它们。

首先，我们将回顾Hinton开创性的工作[60]中提出的变换等变表示（TER）的新原理，以及这种表示的无监督训练的最新公式[9]。接下来回顾了许多最新模型中具有代表性的生成网络，包括自动编码器、生成对抗网络（GANs）、基于流的生成网络和Transformers的变体（见图2）。学习从这些生成模型中分离表示的原则也是许多无监督方法的核心，我们将回顾如何从无标记数据中提取可解释的生成因素。最后，自监督方法构成了一大类无监督模型，我们将回顾自回归模型以及图像和视频表示的自监督训练。

零样本学习（Zero-Shot Learning，ZSL）也位于频谱的左端，它通过挖掘通常不相交的概念集上的辅助任务来实现。与单纯的无监督方法相比，它经常通过嵌入单词和视觉属性来探索概念之间的语义关联，并利用它们将知识从源概念转移到目标概念。给定一个新的样本，零样本学习可以将其分配给一个不可见的概念，其语义嵌入最接近样本的表示。我们请感兴趣的读者参考ZSL方法的更详细的评论[61]。

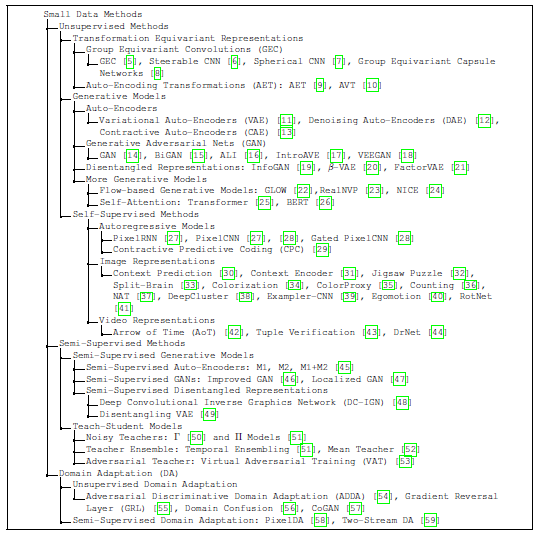


图2：图表展示了在无监督和半监督方法的情况下小数据方法的分类。虽然许多方法位于多个类别的交叉点（如图1所示），但它们是按在本次调查中查看它们的位置分组的。这不是所有小数据方法的百科全书。

**1.2半监督方法**

沿着谱的右边是半监督方法，它探索未标记和标记的例子来训练模型。其思想在于，未标记的示例提供了关于数据在空间中一般分布的重要线索，并且可以通过探索这种分布来训练鲁棒模型。例如，稳健模型应在沿数据流形方向的随机变换（如GAN的平移、旋转、翻转甚至随机扰动[47]）下进行稳定和平滑的预测，或避免将其决策边界置于数据分布的高密度区域。

沿着这个方向，如图2所示，我们将回顾半监督生成模型，扩展它们的非监督对应模型，如半监督自动编码器和GANs，以及它们的分离表示。还将通过鼓励teacher和student模型在有标签和无标签数据上的一致性来审查各种师生模型，以训练半监督模型。它们可以按照获得teacher模型的不同方式进行分类——通过应用随机扰动和敌对扰动[51]，[53]student模型集合的平均值[51]。

在半监督方法中，当利用不相交概念集上的辅助任务来改进模型训练时，也存在小样本学习（FSL）。一方面，当概念关联可以通过嵌入的表示在不同的概念之间共享信息时，这就像零样本学习。另一方面，可以从一个基本概念库中抽取一组辅助任务，学习一个元模型，用少量实例（如初始点、模型参数更新规则）提取训练和更新模型的知识。FSL可以看作是一个半监督问题，因为很少有标记的例子可以通过更新模型来解决目标任务。关于FSL的详细回顾，我们希望读者参考[62]。

请注意，我们不可能审查一份关于如此广泛的无监督和半监督方法的所有工作的独家清单。在本文中，我们只选择回顾该领域在理念、原则和模式方面的最新进展，以展示我们取得的成就，以及我们将如何应对小数据挑战。

本文的其余部分安排如下。无监督方法将在第2节中进行审查，然后在第3节中对半监督方法进行调查。第4节将讨论相关和新兴的主题，包括无监督和半监督域自适应，以及变换等变和不变性在模型训练中的基本作用。最后，我们将在第5节结束调查。

**2 无监督方法**

在本节中，我们将调查关于学习无监督表示的文献。从未标记的示例中训练无监督表示的目的是确保它可以在将来推广到新的任务。

我们将从新出现的学习转换等变表示原理、各种有代表性的生成模型及其对可解释生成因子的分离表示、以及各种用于训练图像和视频表示的自监督方法开始进行回顾。

**2.1无监督表示学习**

无监督表示的训练方法大致分为以下三组研究。

**变换等变表示**。近年来，从无标记数据中学习变换等变表示（TERs）在无监督和有监督两种方法中都引起了广泛的关注。特别是，一个好的三元等价物，它具有不同类型的变换，这样图像中的场景结构就可以被压缩编码成它的表示形式。然后，在训练的TER的基础上，对未见的视觉概念进行连续的识别。TER的概念最初是由Hinton等人提出的[60]在介绍胶囊网的过程中，我们将在第2.2节中对其进行回顾。

**生成模型**。在无监督学习问题中，自动编码、生成对抗网等生成模型得到了广泛的研究，从这些模型中可以学习到描述无标记数据生成过程的紧致表示。我们将回顾这些模型的学习和推理问题，并讨论如何将结果表示分离为能够解释内在和外在数据变化的生成因素。除了自动编码器和GANs之外，更多的生成模型也将在第2.3节中讨论.

**自监督的方法**。此外，还存在大量的自监控信号，用于训练不访问任何标记数据的模型，包括自监控重建数据的自回归模型。我们将在第2.4节中回顾学习无监督表示的不同类型的自我监督信号。

**2.2 变换等变表示**

在我们开始回顾无监督表示学习的方法之前，思考一个好的表示应该具有哪些性质是有益的，特别是从卷积神经网络（CNNs）的巨大成功中。这应该为学习无监督表示的实践奠定基础。

尽管目前还缺乏一个可靠的理论，但人们认为，图像翻译的等价性和不变性对CNNs的成功起着至关重要的作用，特别是对于有监督的分类任务[1]，[60]。典型的卷积神经网络（CNN）由两部分组成：通过多个卷积层的输入图像特征映射和将特征映射到目标标签的全连通层分类器。

当得到的特征映射与输入图像的平移等价时，完全连接的分类器应该是对任何变换都不变的预测标签。在Hinton等人提出学习变换等变表示（TER）概念之前[5] 文献[60]、[63]、[64]主要关注的是利用变换不变性准则来训练有监督模型，通过最小化各种变换增强的标记图像的分类误差[1]。不幸的是，不可能直接应用变换不变性来训练一个无监督的表示-没有标签监督的指导，这将导致一个平凡的表示不变性的所有例子。

因此，采用变换等价性作为训练无监督表示的标准是一种自然的选择，希望它可以推广到不知道其标签的未见任务。这与转换不变性的标准相反，转换不变性的标准倾向于将学习到的表示更专门化为给定任务的标签。事实上，很容易看出，通过卷积层生成的特征图与翻译的特征图是等价的，翻译后的图像的特征图也会以同样的方式移动，从而受到边缘填充效应的影响[1]。这激发了许多工作来概括这一思想，以考虑翻译以外的更多类型的转换（例如，一般的图像扭曲和投影转换）[5]。通过对图像的内在视觉结构进行编码，这可以学习图像的良好表示，这些结构与许多变换是等价的。

在此基础上，通过直接训练特征映射作为不同变换群的函数，提出了群等变卷积（GEC）[5]。证明了所得到的特征映射在指定变换下是等价的。然而，群等变卷积的形式是严格定义的，这限制了它在许多应用中表示的灵活性。或者，通过最大化结果表示和所选转换之间的依赖性，探索一种更灵活的强制转换等价性的方法，从而产生自动编码转换（AET）[9]。与GEC相比，AET不完全符合变换等价性准则，在追求无监督表示形式的灵活性的方面。

**2.2.1 群等变卷积**

考虑一个组G，它可以由各种变换组成，如旋转、平移和镜像反射。群等变卷积（GEC）的目标是生成与群中所有变换等价的特征映射。

为了正式引入变换等价性的概念，我们可以将输入图像和特征映射f看作图像网格上的函数，



其中f（p）给出像素位置p处的特征。为了简单起见，我们只考虑单通道特征映射，但它可以直接扩展到多通道场景而没有任何困难。当将变换应用于f时，其结果是变换图像或特征映射。然后我们说一个带核滤波器的卷积是g的等价变换，如果即与转换输入的卷积等于与原始输入的卷积的转换。

为了使变换等价，在GEC中，特征映射被认为是组的一个函数，定义为。

然后，在上与输入图像f的群卷积被定义为



生成群卷积特征映射因此，输入图像之后的所有特征映射都是的函数，这样的特征映射f与滤波器的组卷积被定义为



其中滤波器也定义在G上。如果我们将G组限制为平移，则不难证明该组卷积减少为传统卷积。

Cohen和Welling[5]证明了上述群卷积的变换等价性，



式中，定义了将变换u应用于输入f的算符Lu。这表明变换输入的卷积等于卷积输入的变换，即变换等变。

群卷积通常以有监督的方式训练，以在神经网络中与一些分类层（例如，完全连接层和软最大层）一起表示图像[5]。原则上，也可以通过将它们视为自动编码器体系结构中的编码器来执行它们的无监督训练。此外，通过将群卷积分解为滤波器变换和平面卷积，存在一种有效的实现方法[5]。

值得注意的是，训练群等变表示的思想被扩展到了更多场景中的变换等变表示。例如，群等变胶囊网络将群等变卷积与动态路由机制结合起来训练胶囊网络[8]；球形CNNs[7]在SO（3）群中分析球形图像，而可转向CNNs[6]在SO（2）群中研究可转向表示的等变性质。有关更多实现细节，请参阅[5]、[6]、[8]。

**2.2.2自动编码转换**

虽然群卷积在数学上保证了变换的等价性，但它们作为所考虑的变换群的函数有一种非常有限的特征映射形式。在许多应用中，我们通常更喜欢更灵活的表示形式，可以通过探索未标记数据的分布以无监督的方式进行训练。在这一节中，我们将回顾最近提出的自动编码变换（AET）[9]和变分方法-自动编码变分变换（AVT）[10]。

***自动编码转换***

与传统的自动编码数据（AED）范式通过重建数据来学习表示不同，AET试图通过解码原始图像和转换图像的表示来训练无监督模型。它假设如果变换可以重构，则表示应该包含变换前后图像的视觉结构的所有必要信息，以便表示是变换等价的。此外，对表示的形式没有限制，这使得为将来的任务选择合适的表示形式变得更加灵活。

形式上，考虑从分布p（t）采样的变换t，以及从数据分布p（x）绘制的图像x。通过对x应用t，可以将x转换为t（x）。然后，AET旨在学习带参数的编码器，提取给定样本x的表示。同时，转换解码器还学习了t，它通过解码原始图像和转换图像的表示来估计输入变换t的。

自动编码变换（AET）的学习问题归结为联合学习表示编码器和变换解码器。为此，可以通过最小化变换t与其估计值之间的以下重建误差来训练AET，



其中，变换的估计值是编码器和解码器的函数，使得，并且期望值被接管采样的变换和图像。然后，通过反向传播损耗梯度，在小批量上联合更新和的网络参数

在[9]中，AET模型考虑了三种类型的变换：参数变换、组织变换和非参数变换。这表明可以将一系列广泛的转换集成到AET模型中。

**自编码变分变换**

从信息论的角度来看，Qi等人。[10] 提出了一种可供选择的自动编码变分变换（AVT）模型，该模型通过最大化变换与表示之间的相互信息来揭示它们之间的联系。它假设一个好的TER应该最大化它对变换的概率依赖性，这样当图像的视觉结构被外部变换时，表示包含了解码变换的内在信息。

形式上，转换图像t（x）的表示z由AVT中的平均和方差指定，使得



式中由正态分布N（0；I）得出，表示按元素划分的积，是模型参数。

通过这种概率表示，可以最大化相互信息来学习，即



直接最大化互信息可能是困难的，并且是一个变分的下界



通过引入一个代理变换解码器，它是变换t在表示z和图像x上的条件概率。

这使得我们能够通过最大化上述互信息的下界，有效地联合训练表示编码器和转换解码器。我们请感兴趣的读者参阅[10]了解更多详情。

**2.3 生成表示**

生成模型，如生成对抗网[14]，自动编码器及其变体已经成为以无监督方式从未标记数据中提取表达性表示的强大工具。在本小节中，我们将回顾基于无监督模型的表示学习的几个方向，特别是GANs和自动编码器，以及它们的表示分离对应方，用于建模独立和可解释的生成因子，这些生成因子对许多下游任务很有用。

我们将展示这些生成模型在很大程度上是相关的。例如，GANs依赖于学习编码器来从数据推断表示[15]、[16]和减少模式崩溃[18]，而自动编码器可以通过对抗性训练得到增强，以从整个潜在代码空间生成更清晰的数据重构[65]。在这些生成模型的基础上，我们还学习了各种形式的分离表示，为从表示中提取、分离和解释生成因素开辟了一个积极的研究方向。

**2.3.1自动编码器**

自动编码器和许多变体[11]、[12]、[13]、[67]是寻求通过联合训练一对编码器（推断组件）和解码器（重构组件）来重构输入数据的生成模型。这里我们将回顾变分自动编码器（VAE）[11]以及去噪自动编码器（DAE）[12]、[67]和压缩自动编码器（CAE）[13]，它们与第2.3.3节中的解纠缠表示的正则化机制和第3.2节中的半监督方法密切相关。

***变分自动编码器***

变分自动编码器（VAE）[11]通过使参数化模型的边缘数据似然的变分下界最大化来训练自动编码器模型。为此，使用变分编码器来近似难以处理的后验，从而导致以下不等式的下界边缘似然：



其中p（z）是表示的优先项，是解码器。

重新参数化技巧也被引入到来自的样本中



式中是从平均值和单位偏差为零的简单高斯分布中随机抽取的，且是按元素划分的乘积。该方法将模型参数从随机噪声中分离出来，通过神经网络对误差信号进行反向传播训练。

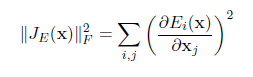
稍后，在回顾第2.3.3节中的分离表示时，我们将看到VAE提供了一个强大的工具来研究和实现表示分离，以提供可解释的生成因素。

***走向鲁棒的自动编码器***

去噪自动编码器（DAE）[12]和压缩自动编码器（CAE）[13]都旨在学习对输入数据噪声不敏感的鲁棒表示。

与典型的自动编码器不同，DAE[12]将噪声破坏的样本作为输入，并尝试重建原始数据。这迫使神经网络学习可用于恢复未损坏干净数据的鲁棒表示。有很多方法可以破坏数据。例如，可以随机删除输入数据的某些部分，DAE尝试恢复丢失的部分；也可以通过旋转、平移和镜像翻转随机转换图像，DAE旨在学习健壮的表示，以便在转换之前从中恢复原始图像。

CAE[13]以不同的方式学习健壮表示。CAE不依赖解码器来重建DAE中的原始数据，而是直接惩罚编码器E在输入数据存在小扰动的情况下学习到的表示的变化。结果对输入样本x周围Jacobi矩阵的Frobenius范数进行以下惩罚以训练CAE



其中Ei表示x的编码表示的第i个元素。

通过在模型输入甚至模型本身加入噪声来对模型训练进行正则化的思想，导致了许多训练鲁棒有监督和无监督模型的正则化方法。对抗性噪声比随机噪声更能训练鲁棒分类器，其方法是鼓励对受对抗性影响的有标记和无标记样本进行平滑预测。我们将在第3.2节中的半监督方法的背景下仔细研究它们。

**2.3.2 基于GAN的表示**

在GAN模型中，数据是由输入到发生器的噪声产生的，因此这些噪声可以看作是发生器产生的数据的自然表示。

考虑到文献[68]，[69]证明了许多GAN变体具有生成分布与实例不可区分的数据的广义能力，所有真实数据的GAN表示也都是完整的。

然而，当给定一个真实的样本时，我们需要反转生成器以获得与样本相对应的噪声表示。因此，需要一个编码器，它可以直接输出噪声，从中产生相应的样本，并由此表示。

为此，在双向生成对抗网络（BiGAN）[15]和对抗学习推理（ALI）[16]中分别独立地发展了对抗训练生成器及其相应编码器的思想。该思想后来被集成到一个规则化的损耗敏感GAN模型中，该模型具有已证明的分布一致性和可推广性，以生成真实数据[70]。

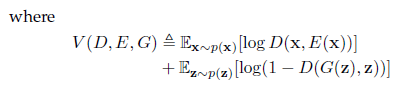
***BiGAN and ALI: Adversarial Representation Learning***

形式上，这些方法旨在从GAN模型中学习三个元素：1）生成器从输入噪声的分布p（z）到生成样本X的分布的映射；2）编码器将样本映射回噪声，使得理想的G（Z）等于x，即，是G的逆；3）鉴别器指定区分实对和伪对的概率。

与经典的GAN相比，有两大区别。首先，编码器是表示学习的额外元素。第二，鉴别器具有联合样本噪声对而不是单个样本作为输入，以区分真假对。

以上三个元素可以用极大极小目标联合训练





p（x）是真实的数据分布。这个极大极小问题可以通过交替梯度的方法来解决，比如训练经典的GAN[14]。

Donahue等人[15] 已证明在理想情况下，生成的编码器E几乎处处反转生成器G，即（见[15]中的定理2）。此外，还证明了E和G在（2）的联合训练是通过最小化自动编码器的“损失”（见[15]的定理3）来完成的，这使得E成为其输入样本的期望表示模型。

***更多相关工作***

除了BiGAN和ALI之外，还有其他的混合方法联合训练自动编码和GANs在一个集成的框架中进行对抗性表征学习和推理[17]、[18]、[65]、[66]、[71]

例如，Larsen等人[65]使用来自GAN鉴别器的中间表示来测量重建图像和输入图像之间的相似性作为重建误差来训练VAE。或者，对抗性自编码器[66]被提议通过匹配来自数据分布p（x）和先验分布p（z）的噪声的聚集后验来训练VAE。通过训练鉴别器来区分p（z）和q（z），并引导编码器产生与前一个不可区分的聚集后验，实现分布之间的匹配。这样，VAE的训练被正则化，以确保解码器产生一个生成模型，该生成模型在期望的数据分布之前映射给定的数据。

为缓解模式崩溃问题，文中还探讨了VAE与GAN的联姻问题。例如，Srivastava等人[18] 训练编码器（在那篇论文中称为重构器）来反转生成器，并通过使编码数据的分布与输入高斯噪声相匹配来减少生成样本的模式崩溃。假设如果发生模式崩溃，重构器不太可能将所有生成的样本映射回原始高斯噪声的分布，这将导致一个强大的学习信号来训练生成器和重构器。

Huang等人[17] 进一步介绍了一种以后验为判别函数的内凹模型，用以区分真实数据和伪数据。特别地，在实际样本x上的z的后验被鼓励与先前的p（z）相匹配，而在生成的样本上的z的后验被假定偏离p（z）。然后，通过将后验样本与先验样本相匹配，训练生成器生成样本。结果表明，IntroAVE能够生成与真实样本不可区分的数据。

**2.3.3分离表示**

提出了分离表示法[19]，通过提供可解释和显著的属性来描述数据，从而促进下游任务。Bengio等人[72]提出在一个分离的表示应该随着数据的变化而变化，以响应真实世界的事件和转换。

例如，一组有意义的属性，例如面部表情、姿势、眼睛颜色、发型、性别甚至身份，可以单独分配来分离面部图像，它们对于解决未来的识别问题非常有用，而不必暴露在某些有监督的数据中。这就意味着，可以推广到自然监督任务的良好表示应该尽可能地分离，以提供一组丰富的因子化属性来描述数据。

***InfoGAN：分离基于GAN的表示***

分离表示的努力导致了文献中的InfoGAN[19]及其变体[20]，[73]来训练生成模型，该模型可以从分离的表示中创建数据。具体地说，InfoGAN假设有两种类型的噪声变量输入到它的生成器中：1）不可压缩噪声的矢量z，它不分解为任何语义表示，并且可以像在传统GAN中一样被生成器以纠缠的方式使用；2）潜在码的矢量c，它表示显著分离的关于生成的样本x的信息在生成过程中不会丢失。

因此，InfoGAN的假设是通过组合这两种噪声，使潜在码c和生成的样本G（z，c）之间的互信息最大化。它应该防止生成器忽略对包含有关生成样本的显著知识的潜在代码的依赖性。在生成器G上最大化互信息以训练InfoGAN和传统GAN的minimax目标。利用代理分布导出了的一个可处理的变分下界，以逼近真后验：



其中p（c）是潜在码的先验分布，H（c）是其熵。关于InfoGAN的更多细节可以在[19]中找到。

***-VAE：分离VAE表示***

分离表示的思想也扩展到了其他无监督模型。其中有 -VAE[20]，其目的是通过施加约束使其与各向同性高斯匹配来分离推断的后验。它通过限制其容量在推断的后验上造成潜在的信息瓶颈。这样的正则化不仅鼓励了数据更有效的表示，而且由于各向同性的先验，使得表示分离为独立的因子。

将下列目标最大化以训练VAE模型



其中正拉格朗日乘子来自约束。

当=1时，公式简化为传统的VAE模型。随着的增加，对潜在信息瓶颈的一个更强的约束是强制控制表示的能力和条件独立性。较高的将在-VAE模型的重建保真度和学习表示的分离度之间进行权衡。

***解纠缠度***

衡量学习表示的解纠缠度，解纠缠度评分[20]是在假设解纠缠表示即使使用一个简单的分类器，也可以根据其表示对数据进行稳健分类的基础上设计的。通过固定表示中的一个生成因子，同时随机抽取所有其他生成因子，生成大量图像。然后用一个低性能的线性分类器对该因子进行识别，得到的准确度报告为解纠缠度得分。显然，如果分离表示的独立性和可解释性保持不变，则固定因子的方差应该很小，因此分类器应该具有较高的识别精度，并给出较高的分离分数。

然而，有人认为，线性分类器仍然可能对超参数和优化器敏感，并且如果仅将K个因子中的K-1分离，其分离度量将遭受失效模式。为了解决这一问题，本文提出了一个替代指标[20]，直接使用结果表示中每个维度的方差作为固定因子的指标，并应用多数票分类器来预测所选因子。这避免了优化超参数的调整，也避免了其他度量的故障模式。

***更多的分离表示***

除InfoGAN和-VAE之外，许多其他生成模型都在寻求分离表示。FactorVAE[21]提出最小化聚集后验q（z）与其因子化形式之间的总相关（TC），该形式测量多个随机因素的相关性。根据密度比技巧，训练一个鉴别器来区分两个后验的样本，并输出样本z来自真实聚合后验q（z）的概率。然后通过最小化VAE下界和得到的TC来训练分解后的VAE。与-VAE相比，FactorVAE避免了对互信息项的不必要的惩罚，从而在充分分离生成因子表示的同时，获得更好的数据重建。

此外，在半监督方法[48]、[74]的背景下也研究了分离表示，这将在下一节中进行回顾。

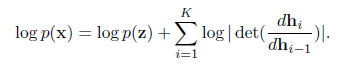
**2.3.4更多生成模型**

***基于流的生成模型***

基于流的生成模型[22]、[23]、[24]通过一系列双射函数将从简单分布（例如，多变量高斯分布）提取的随机噪声z映射到数据样本x



这种可逆函数序列称为流。它允许我们通过变量公式的变化来计算x的对数似然



采用Actnorm、可逆卷积和仿射耦合层三种不同的可逆流函数[22]构造了深生成流（GLOW）模型中的一步流。压缩运算符还定义了GLOW中具有不同数据抽象级别的多尺度结构[23]。GLOW中的每一步都有一个对数行列式，由于它有一个三角Jacobian矩阵，所以可以很容易地计算出来，从而可以有效地最大化得到的数据对数似然来训练模型。

***Self-Attention and Transformer***

Transformer [25]被提出作为递归神经网络的替代品，它堆叠了自关注层，以及点全连接层和位置编码，以捕获其编码器和解码器组件中输入和输出序列之间的依赖关系。

自我关注是关键。序列中的每个嵌入都映射到查询、键和值的元组。然后，每个位置的输出是由当前查询和序列键之间的相似性加权的值之和。通常采用多头注意的方法，用不同的投影权重对查询、键和值进行多次线性投影，并将这些线性投影的结果连接起来投影到最终结果。

除了自我注意之外，编码器和解码器中的每一层还包含一个完全连接的前馈网络，该前馈网络分别且相同地应用于每个位置。通过正弦和余弦函数的位置编码也被添加到每个嵌入中，它提供关于序列中位置的信息。Transformer已经成为自然语言任务中单词嵌入的一种强大的无监督表示，关于Transformer及其应用的更多细节可以在[25]，[26]中找到。

**2.4自我监督方法**

大量的自监控信号也被用来训练无监控的表示。我们将从回顾自回归模型作为一大类自监督模型开始，并继续学习图像和视频的自监督表示。

**2.4.1自回归模型**

其中一类自监督模型是通过预测上下文、缺失或未来数据来训练的，通常称为自回归模型。其中包括PixelRNN[27]、PixelCNN[28]、[75]和Transformer[25]。它们可以生成有用的无监督表示，因为预测数据中不可见部分的上下文通常依赖于相同的共享潜在表示。

***PixelRNN***

具体地说，在PixelRNN[27]中，图像被划分成规则的小斑块网格，并构建一个递归结构来根据当前斑块的上下文预测其特征。提出了三种不同的RNNPixels，分别用行LSTM、对角BiLSTM和多尺度PixelRNN来生成图像块序列。

对于ROW LSTM[27]，从上到下逐行生成图像，并且面片的上下文大致是面片上方的三角形。相反，对角线LSTM从顶部的一个角对角扫描图像，并到达底部的另一个角，因此它具有对角线上下文。多尺度PixelRNN[27]由无条件PixelRNN和一个或多个PixelRNN层组成。首先应用无条件PixelRNN生成原始图像的子采样，然后条件PixelRNN层以较小的图像作为输入生成原始较大图像。多层条件像素层可以叠加，从低分辨率到高分辨率逐步生成原始图像。

***PixelCNN***

行和对角线LSTM的一个缺点是计算量大，因为每个面片的特征必须按顺序计算。这可以通过使用卷积结构一次计算所有面片的特征来避免。蒙蔽卷积用于避免仅依赖于前一个上下文而不是未来上下文的条件依赖冲突。与具有潜在无限依赖范围的PixelRNN相比，PixelCNN[28]的代价是将每个贴片的上下文限制为一个有界的感受野。因此，可以堆叠多个卷积层以增加上下文大小。

另一方面，门控激活被引入到PixelCNN中[28]。这就产生了一个门控像素CNN，它能够模拟不同补丁之间更复杂的相互依赖关系。此外，门控PixelCNN还增加了一个水平堆栈，该堆栈的条件是目前为止的行，以及一个垂直堆栈依赖于所有以前的行。通过合并两个堆栈的输出，可以避免接收场中的盲点。

***对比预测编码***

自回归模型可以作为自动编码体系结构中的解码器，在该体系结构中，它们被迫输出对预测未来补丁有用的强大表示。这使我们能够以自回归的方式训练表示，而无需访问任何标记的数据。

对比预测编码（CPC）[29]在训练这种自回归模型方面做出了显著的努力。它的目的是最大化上下文c的潜在表示和未来样本x之间的互信息I（c；x），从而通过序列最大化地共享信息可以做出更准确的未来预测。有关CPC及其在训练自回归模型和学习无监督特征中的应用的更多细节，请参见[29]。

**2.4.2图像表示**

除了自回归模型外，自监督方法还探索了其他形式的自监督信号来训练深层神经网络。这些自监督信号可以直接从数据本身导出，而无需手动标记它们。

上下文。例如，Doersch等人[30]使用从图像中随机抽取的两个面片的相对位置作为自监督信息来训练模型。Pathak等人[31]训练上下文编码器，通过最小化像素级重建误差和对抗性损失的组合，从周围环境生成缺失部分的内容。Mehdi和Favaro[36]提出通过解拼图来训练卷积神经网络。

着色。在文献[34]，[35]中，图像彩色化也被用于自监督任务中训练卷积网络。Zhang等人[33]通过从另一子集重建数据信道的子集来呈现跨信道自动编码器，其中跨信道特征被连接为数据表示。

代理类、目标和群集。多索维茨基等人[39]通过对一组代理类进行分类来训练cnn，每个类都是通过对单个图像应用各种转换而形成的。相反，Bojanowski等人[37]使用噪声作为目标（NAT），方法是联合学习表示，并将每个样本分配给一组固定的目标值中的一个。相反，Caron[38]等人。通过迭代聚类特征并使用结果表示更新网络来训练DeepCluster模型。

数数、运动和旋转。诺鲁齐等人[32]学习满足下采样图像和平铺图像之间等价关系的计数特征。Egomotion[40]也被用作一种自我监控信号，用于对连续图像中的视觉元素的表示进行建模，以便在代理在环境中移动时找到它们的对应关系。Gidaris等人[41]通过对离散集合中的图像旋转进行分类来训练神经网络。它学习变换等变表示的特殊情况，因为学习的表示应该通过与应用的旋转等变来编码关于它们的信息。

**2.4.3视频表示**

自我监督的思想也被用来通过探索潜在的时间信息来训练视频的特征表示。例如，时间箭头（AoT）[42]被用作监控信号，以学习视频的高级语义和低级物理表示，同时避免来自视频制作而不是物理世界的人工提示。

帧序列的顺序还可以监督视频表示的训练以捕获时空信息[43]。为此，提出了一种元组验证方法，通过提取单个帧的表示并确定随机采样的帧元组是否在正确的顺序来训练CNN模型，以消除视频片段中的方向混淆。

利用视频帧间的时间一致性，提出了一种图像的分离表示方法。训练DrNet模型[44]将每个帧分解为一个固定内容表示和一个具有对抗性损失的时变姿态表示。它假设姿态表示不包含视频身份的信息，并且对抗性损失防止姿态特征从一个视频区别到另一个视频。DrNet可以学习强大的内容和姿势表示，这些表示可以结合起来生成比现有方法更远的帧[44]。

**2.5 评价**

无监督方法通常根据其在下游任务上的性能进行评估。特别是，学习到的无监督表示可用于对基准数据集（如CIFAR-10、ImageNet、Places和Pascal VOC）执行分类任务。

为了公平比较，在这些数据集上制定了标准的评估协议。在本小节中，我们将回顾这样一个被许多无监督方法广泛采用的协议。尽管由于遗留原因，并非所有方法都与本议定书进行了比较，但文献中出现了允许在许多最近的方法之间进行公平和直接的比较的方法[9]、[37]、[39]、[41]、[76]、[77]、[78]。

**2.5.1 评估方案**

无监督模型的评估通常包括两个阶段。第一阶段是在无监督的情况下只训练未标记的例子。在第二阶段中，在学习到的表示的基础上训练一个有监督的分类器，以评估其对新分类任务的泛化性能。

以ImageNet数据集上的评估协议为例。AlexNet被广泛用作学习无监督表示的骨干，它由五个卷积层和三个完全连接层（包括一个具有1000路输出的softmax层）组成。下面有几个设置用于测试分类任务的无监督模型。

非线性分类器在这种情况下，直到Conv4和Conv5的卷积层在第一阶段进行无监督训练后被冻结。在评估阶段，对Conv4和Conv5以上的所有层进行有标记数据的监督训练。也就是说，为了评估的目的，非线性分类器被训练在无监督的表示上。例如，在Conv4设置中，使用标记的示例训练Conv5和三个完全连接的层，包括最后一个1000way输出层。表1显示了几个无监督模型之间的比较。表中还包括全监督模型（ImageNet标签）和随机模型，分别显示分类性能的上界和下界。

线性分类器也可以在无监督表示的基础上增加一个完全连接的层来训练弱线性分类器。如表2所示，线性层在特征映射的不同卷积层上训练。结果表明，线性分类器能够很好地训练出线性分类器，在经过适当训练的无监督表示之上的分类器训练效率和测试精度之间取得很好的折衷。

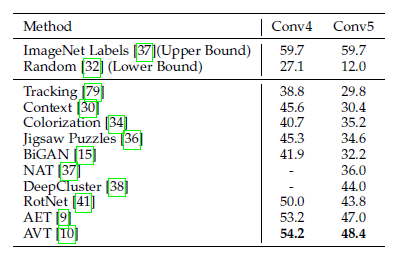


表1:ImageNet上非线性层的Top-1精度。使用AlexNet作为骨干来训练无监督模型。学习无监督特征后，在Conv4层和Conv5层上训练非线性分类器，并用标注的例子比较它们的性能。全监督模型和随机模型具有上下有界性。

跨数据集任务还执行跨数据集任务，以比较无监督表示与新数据集上的任务的通用性。如表3所示，也通过对ImageNet数据集进行预训练来评估无监督模型。然后在具有位置标签的特征地图的不同卷积层上训练单层logistic回归分类器。表4显示了PASCAL VOC上的分类、对象检测和语义分割，其中模型仍然基于AlexNet变体，并以无监督的方式在ImageNet上预训练。将这两个结果与用位置标签和ImageNet标签训练的完全监督模型以及随机网络进行了比较。

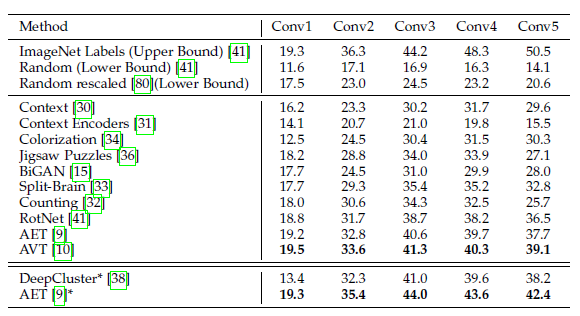


表2:ImageNet上线性层的Top-1精度。以AlexNet为骨干，对比较中的无监督模型进行训练。一个1000路的线性分类器被训练在各种卷积的特征映射层上，这些特征映射的空间大小被调整为有大约9000个元素。文中还报道了全监督模型和随机模型，以显示无监督模型性能的上下界。仅使用一种crop，在AET测试期间不使用dropout或局部响应标准化，但使用\*表示的模型除外，其中10种作物用于比较结果。

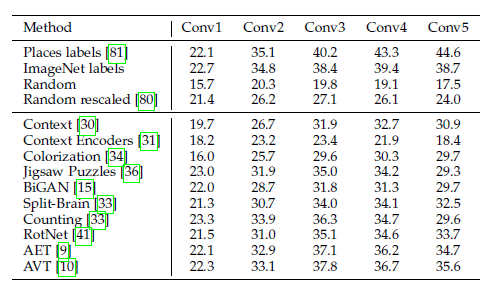


表3:Places数据集具有线性层的Top-1精度。一个205路的逻辑回归分类器被训练在不同层次的特征图上，这些特征图的空间大小被调整为有大约9000个元素。所有无监督特征都是在ImageNet数据集上预先训练的，这些数据集在训练带有位置标签的logistic回归层时被冻结。将无监督模型与用位置标签和ImageNet标签训练的完全监督网络以及随机模型进行了比较。

**3.半监督方法**

在本节中，我们将从两个不同的角度回顾半监督方法

半监督生成模型。在第3.1节中，将回顾半监督自动编码器、GANs和非监督表示。我们将展示如何从相应的无监督生成模型中导出这些半监督生成模型，从而揭示无监督和半监督方法之间的内在联系。

教授student模型。这是一大类半监督模型，在文献中已经达到了当前最佳的水平，其中一个或一个teacher模型集合被训练来预测未标记的例子，并且预测的标签被用来监督student模型的训练。我们将在第3.2节中回顾各种类型的teacher模型-噪音teacher、teacher合奏和对抗性teacher，并展示如何针对各种噪音和/或对抗性机制对其进行培训，以建立更稳健的半监督模型。

**3.1半监督生成模型**

在这一节中，我们将回顾大量的半监督生成模型。

**3.1.1半监督自动编码器**

Kingma等人[45]将无监督变分自编码扩展到两种半监督模型。

第一个潜在特征判别模型（M1）相当简单。在VAE模型对样本x的潜在表示z的基础上，训练分类器来预测其标签。在训练集的有标签部分和无标签部分训练VAE的同时，基于有标签的例子训练分类器。

第二代半监督模型（M2）更为复杂。除了潜在表示z之外，样本x由另一个类变量y生成，该类变量y对于未标记的x是潜在的，或者对于已标记的x是可见的。考虑到额外的类变量，数据由生成过程解释：



其中p（y）是类先验的多项式分布。

与VAE不同，M2引入了一对变分后验来推断z和y：



然后通过可以推断出z和y联合后验。

其中，可以作为分类器来预测测试样本的标签。为了训练M2，考虑了两种情况[45]，分别导出了标记对（x,y）的边缘分布和未标记样本x的边缘分布的变分下界。将这两个边界结合起来会产生一个最大对数似然问题。

然而，在最终的目标函数中还需要增加一个额外的分类代价，这样分类器就可以同时训练有标签和无标签的例子。与VAE类似，重新参数化技巧用于执行反向传播[11]。

最后，M1和M2可以通过使用M1模型的嵌入表示z1学习M2来组合。M2模型有自己的潜在表示z2以及每个样本的标签变量y。这就产生了一个两层的深层生成模型，分别从生成z1和从z1生成x：。

文献[49]、[82]、[83]中除了M1和M2模型和混合模型外，还以不同的方式将监督信息引入到变分自动编码器中。稍后，我们将回顾如何通过部分指定随机变量子集之间的图形依赖性来从半监督vae中分离表示，以便分解和解释数据变化。

**3.1.2半监督GANs**

还采用了GANs从两个不同的角度实现了半监督学习。其中一个考虑训练一个K+1分类器，用K个给定的标签进行分类，用一个伪类来表示生成的样本。它通过将未标记的示例视为属于第一个K实类来探索其分布，并使用特征匹配技巧来释放竞争性性能[46]。

相反，另一种范式将学习GAN模型的生成器看作是数据流形的（局部）参数化，这样就可以在流形上沿着其切线刻画标签不变性。这与经典的基于图的半监督模型中仅由图拉普拉斯逼近的拉普拉斯-贝尔特拉米算子密切相关。

我们将在下面回顾这两种半监督的GANs范式。

**用特征匹配训练K+1分类器**

Salimans等人。[46]提出了改进的半监督GANs训练技术。通过将真实样本和生成的样本放在一起，它训练分类器将每个样本标记为K个真实类或一个伪类中的一个。对于前K个类中的一个，所有未标记的数据都被分类为真实示例，而生成的示例被分类为假示例。传统的分类代价是定义在标签数据上的，并结合无监督的GAN损耗来训练模型。

此外，还必须使用一种称为特征匹配的技巧来训练生成器。它不是通过最大化生成的样本分类为K个实类的可能性来训练生成器，而是通过最小化从分类器中间层提取的实样本和生成的样本之间的差异来训练生成器。

这个技巧在训练半监督GANs的竞技表现[46]过程中起到了至关重要的作用

**利用局部GANs追求标签不变性**

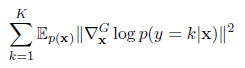
图拉普拉斯（graph-Laplacian）被广泛用于描述图中连接的样本上标签的变化。最小化拉普拉斯图可以对连接的相似样本之间的标签进行平滑预测。当使用图来逼近未知数据流形时，图Laplacian实际上是对底层数据流形上的Laplace-Beltrami算子的逼近。

在[47]中，我们做了一个显著的工作来学习局部GAN，它定义了一个局部生成器G（x,z）围绕每个带有z的样本x。这就产生了数据流形上每个样本x的局部坐标，其中x是原点，即。这样，整个数据流形可以被一系列局部坐标覆盖。它允许我们定义分类函数f（x）在流形上的梯度为



因此，可以直接计算Laplace-Beltrami算子，而不必使用经典半监督方法中常用的基于近似图的Laplacian。

然后通过最小化数据流形上的标签不变性来训练半监督分类器



以及半监督GANs的loss[46]。

此外，局域GAN允许我们从几何的角度来解释流形局部塌缩成低维时的模式塌缩。然后对雅可比矩阵施加一个正交约束来训练生成器，防止它在流形上崩溃。

**3.1.3半监督分离表示**

逆向图形网络

深卷积逆向图形网络（DCIGN）[48]通过将视觉模型作为逆向图形来实现半监督变分自动编码器模型。换句话说，它的目标是学习一组“图形代码”，通过这些代码，图像可以像在图形程序中一样进行转换和渲染。这些图形代码被视为图像的分离表示。

DC-IGN建立在VAE模型的基础上，但采用半监督的方式进行训练。所学习的表示被分解为几个外部变量，如方位角、仰角和光源方位角，以及一些描述身份、形状、表达式和表面纹理的内部变量。在一个小批量中，只有一个因子是固定不变的，其他因子都是固定不变的，生成的图像只有一个主动变换对应于选择的因子，并通过网络转发。与非活动变换相对应的其他变量被钳制为其平均值。误差信号的梯度通过网络反向传播，而与非活动变换相对应的梯度被强制使其与小批量上的平均值不同，这可以训练编码器，使所有关于活动变换的信息都集中在所选择的变量上。

DC-IGN是半监督的逆向图形工程，因为可以从三维人脸和椅子数据集中获得各种变换的训练图像。我们还注意到一些逆图形模型[85]、[86]、[87]被提出用于训练分离表示。

其中包括采用lambertian反射模型并隐式构造3D表示的深lambertian网络[88]和使用域特定解码器重建图像的变换自动编码器[60]、[89]，以及[90]用一个近似可微的渲染器来显式地捕捉模型参数变化和图像观察之间的关系。

分离半监督VAEs

在文献[49]中，提出了一种半监督VAEs的广义形式，用于将可解释变量从潜在表示中分离出来。它编制了用神经网络建立对观测和未观测潜在变量的一般依赖关系的图形模型，并使用随机计算图[91]来推断和训练生成模型。

为此，使用重要抽样估计来最大化有监督和半有监督似然性的下界。通过将每个随机节点展开成一个子图，建立随机计算图来训练生成的模型。具体来说，在生成模型和推理模型中，为每个节点指定了一个分布类型和一个参数函数神经网络。采用重参数化方法对无监督和半监督变量进行抽样，并根据所有半监督变量的联合概率计算重要性抽样的权重。

该模型使我们能够灵活地指定对分离表示的依赖关系，以解释数据变化，并使其余未指定的依赖关系以纠缠的方式学习。

**3.2 师生模式**

半监督学习的师生模型背后的思想是获得一个或多个teacher，并以对未标记示例的预测为目标来监督student模型的训练。最大化师生一致性，提高student对未标记样本分类的绩效和稳定性。

培训teacher的各种方法以及最大限度地提高teacher和student之间的一致性，导致了这一类的各种半监督模型。具体来说，将随机噪声应用于模型的输入层和隐藏层可以追溯到[92]、[93]、[94]，这已经被证明相当于在目标函数中添加额外的正则项。在teacher-student方法中，通过将噪声样本馈送到损坏的模型中来获得噪声teacher，并且将预测偏差最小化，以训练teacher和student之间的模型（“模型50”）或模型的两个损坏副本之间的模型（“模型51”）。

这一思想被扩展到召集一个跨epoch的teacher团队来指导student的培训。他们预测的指数移动平均被用来提高teacher集合对未标记示例（时间集合[51]）的预测标签的准确性。或者，可以对模型参数进行指数加权平均，以形成teacher集合（Mean teacher[52]）的预测。这两种方法分别依赖于在输入样本和模型参数中加入随机噪声来提高teacher和student模型一致性时挖掘未标记数据的稳健性。

与添加随机噪声不同，对抗性的例子是通过student模型计算出来的，可以最大限度地改变预测的标签。这就产生了一个对抗性的teacher，通过最小化teacher对对抗性例子的偏离来训练和更新student。

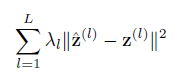
这就产生了虚拟对抗训练（VAT），它在半监督学习方面取得了最新的成绩。

下面，我们将详细介绍不同的教学方法。

**3.2.1 Noisy Teachers：和模型**

和模型的开发都基于这样的信念，即在数据的任意随机变换和对模型的扰动下，稳健模型应具有稳定的预测[95]。这样可以将决策边界推离训练样本，使模型对数据和模型参数上的噪声不敏感。因此，在student模型的输入和参数中加入随机噪声和扰动，形成一个有噪声的teacher，并将teacher对预测的偏差最小化，以训练student模型。

具体地说，模型[50]具有每层l的多层潜在表示，并且使用自动编码器通过从损坏的进行去噪来获得估计的。然后，层上的batch normalized估计和净潜在表示之间的平方误差之和



通过最小化训练不同层次上正超参数系数加权的清晰student模型。

相反，通过最小化噪声输出之间的差异，简化了-模型[51]。在半监督学习问题中，给定一个有标签或无标签的样本x，它被一些噪声污染，并被随机退出和合并方案扰动进入模型[95]。此进程运行两次，生成其输出和的两个版本。然后，将它们之间的平方误差最小化，以鼓励噪声输出之间的一致性，并结合标记示例的分类成本来训练模型。不同于匹配干净和损坏表示的-模型，-模型运行损坏分支两次以匹配噪声输出。

然而，这两个模型都依赖于随机噪声来探索它们对噪声输入和扰动模型的弹性，这对于寻找一个合格的teacher来训练鲁棒模型是无效的。因此，在各个时期对teacher集合进行跟踪，以形成一个更有能力的teacher模型，从而产生以下时间集合[51]和平均teacher[52]方法。

**3.2.2 Teacher Ensemble: Temporal Ensembling and Mean Teacher**

时间集合[51]和mean teacher[52]在随时间跟踪模型集合以获得更好的teacher模型方面彼此相似。然而，它们在通过被跟踪模型的参数（mean teacher）保持指数移动平均值超过预测（时间整合）方面有所不同。

在形式上，考虑一个由参数化的模型在模型参数和/或输入中添加了一些噪声的情况下输出输入x的预测y。

对于时间序列，在每个epoch，给定样本x上的目标预测以指数移动平均（EMA）方式在线更新为



具有正平滑系数。将得到的EMA预测进一步规范化，构造目标y，通过最小化来训练模型



再次，该目标与小批量的分类成本相结合，以训练被噪声污染的模型。由于每次迭代都要更新单个实例上的预测值，因此每个历元只更新一次预测值，从而使早期模型中的信息以较慢的速度融入到模型的训练中。

与时间整合相反，mean teacher在模型参数上保持一个EMA，而不是个人预测



使用当前student模型的参数。然后通过最小化过来更新student模型



当时间加扰和平均teacher跟踪一组先前的模型来预测teacher的目标以监督训练过程时，他们仍然依赖于添加随机噪声来训练预测一致的稳定模型。研究表明，在不知道模型易受对抗性噪声影响的情况下，通过训练模型对抗随机抽取的噪声，无法获得样本周围的局部各向同性输出分布[96]。这激发了另一种方法，即使用敌对的teacher来监督培训过程。

**3.2.3 对抗性teacher：虚拟对抗性培训**

对抗训练已经被用来规范化一个模型，并使其在对抗示例中具有健壮性[96]，[97]。具体来说，该模型被训练成沿着输入示例的敌对方向进行平滑预测。这种方法已经扩展到虚拟对抗训练（VAT）[53]，在这种训练中，可以围绕未标记的数据寻找对抗方向，沿着这些数据，模型的变化最大。这允许以半监督的方式训练模型。

形式上，考虑一个有标签或无标签的示例x，以及一个输出标签的条件分布的参数化模型。VAT通过以下公式来找到最具对抗性的方向



在两个分布之间有一个发散测度D，其中在样本周围的半径内寻找对抗方向。

然后将对抗损失最小化训练模型



在标记和为标记的样本上，加上最小的分类成本。

对抗方向是的Hessian矩阵的第一主特征向量，它是r=0时r的函数，这反过来允许快速幂迭代算法求解。这可以很容易地集成到随机梯度方法中，以便在小批量上迭代更新。

**3.3 评价**

首先介绍了半监督方法实验中广泛使用的两个数据集。

**3.3.1数据集**

CIFAR-10数据库集。数据集[98]包含50000个训练图像和10个图像类别的10000个测试图像。我们在实验中训练半监督LGAN模型，其中每个类标记100个和400个标记示例，其余示例不标记。在这个数据集上的实验结果是通过10次以上的平均运行来报告的。

SVHN数据集。数据集[99]包含32X32街景房号，大致以图像为中心。训练集和测试集分别包含73257和26032个门牌号。在一个实验中，每个数字使用50个和100个标记的例子来训练模型，其余的未标记的例子作为辅助数据，以半监督的方式训练模型。

**3.3.2 结果**

CIFAR-10和SVHN常被用来评价半监督模型的性能，方法是用所有未标记的训练图像和不同数量的标记样本对其进行训练。然后在单独的测试集中报告错误率。

为了便于比较，通常采用13层卷积神经网络对模型进行训练（见[51]表5）。对于大多数模型，随机平移和水平翻转被用作输入图像的数据增强。在许多模型中也使用了两种形式的噪声（例如，模型、mean teacher、时间集成）：高斯噪声应用于输入层，而随机漏失应用于网络中。

表6和表5分别对CIFAR-10和SVHN数据集的结果进行了比较，从中我们可以看出，在这两个数据集上，teach-student模型类的性能都优于其他方法。特别是VAT在这些比较模型中取得了最为突出的性能。

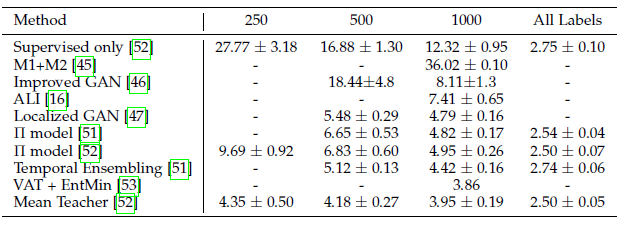


表5:SVHN上的错误率，以及用于训练不同模型的各种标记示例。注：在模型上分别报告了两个版本的结果。

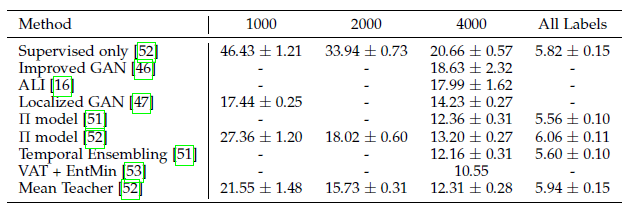


表6:CIFAR-10上的错误率，以及用于训练不同型号的各种标记示例。注：在模型上分别报告了两个版本的结果。

**4．相关和新兴话题**

在本节中，我们将回顾无监督和半监督方法的相关和新兴话题。

**4.1领域适应**

我们将以无监督和半监督的方式审查领域适应问题。

**4.1.1无监督域自适应**

GANs的一个有趣的应用是将学习到的表示和模型从源域自适应到目标域。具体地，对于无监督域自适应，一组标记源实例从源域的分布中采样，而另一组未标记的实例从目标域的分布中得到。然后，无监督域自适应的目标是学习在目标分布上具有低风险的分类器f。我们将无监督域自适应分类为无监督方法，因为目标域不包含监督信息，尽管源域是受监督的。

有几种不同的方法来解决无监督域自适应问题。在这里，我们将重点回顾与GAN模型密切相关的对抗性学习方法，利用它可以生成与目标样本分布不可区分的样本的特性。

如[54]所述，在开发无监督域自适应算法时有三种设计选择：1）绑定权重：是否在源域和目标域的表示模型之间共享权重；2）基本模型：是否从源域到目标域自适应区分或生成模型；3） 用于训练模型的对抗性目标。

不同的选择导致了不同的模式。

**对抗性区分域适应**

对抗性区分域自适应（ADDA）[54]对源域和目标域的表示模型的权重进行了求解。相反，它学习两个独立的模型MS和MT，将源和目标样本映射到它们各自的表示。首先，基于源域中标记的示例，在表示Ms的顶部训练分类器f：



其中，是标记示例上的分类错误。

然后固定MS，训练目标表示模型MT，使两个模型输出的分布一致，相互匹配。基于GAN的目标通过学习区分源表示和目标表示的域鉴别器来实现这一点，



然后通过混淆域鉴别器由MT生成的表示来自源域来最小化对抗性损失以训练目标表示MT：

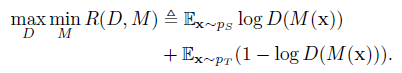


对判别器D和目标表示MT进行迭代优化，使其收敛。然后，根据训练好的分类器f和目标模型MT，用f（MT（x））对样本x进行分类。

**梯度反转层**

与ADDA不同，梯度反转层（GRL）模型[55]选择绑定源和目标表示的权重（即MS=MT=M）。分类器f、共享表示M和域鉴别器D将被联合训练。

它在共享M上引入了以下正则化器和域鉴别器D



换句话说，共享表示学习将样本映射到同一个分布，无论是源域还是目标域，这样D就无法区分它们。

该正则化器与分类损失相结合，产生联合优化问题



与ADDA相比，分类器与表示联合训练，优化了易受消失梯度影响的真实minimax目标[54]。

**4.1.2半监督域自适应**

当目标域上有附加的标记示例时，无监督域和半监督域自适应之间的边界变得模糊。例如，在GRL中，分类损失（3）不仅可以用标记的源实例，也可以用标记的目标实例最小化。

或者，像素级域适配（PixelDA）[58]选择使用用于采样噪声z的GAN生成器G（x，z）将源图像直接适配到其目标对应图像，以与目标分布pT匹配。然后，通过以半监督方式将标记的适应图像和标记的目标图像组合起来训练分类器。为了利用图像自适应过程的先验知识，还可以最小化额外的内容相似性损失。

此外，还提出了一种双流结构来同时训练源域和目标域的两个网络。它不试图直接强制域不变性，因为域不变性特征可能会削弱所学习分类器的鉴别能力。相反，它通过建模源数据和目标数据之间的相似性和差异来显式地建模域移动。

具体来说，它在来自两个域的标记数据上分别训练两个网络流。通过将两个网络流的权值之差最小化到线性变换，引入了权值正则化器。这鼓励两个相关流建模域不变性，同时承认域之间存在差异。然后，通过源样本和目标样本的表示，将域差异最小化。这可以通过最小化内核空间中的最大平均差异（MMD）[100]、[101]、[102]、[103]来实现。同时，GRL[55]的思想也可以用于训练当两个域的表示变得不可区分时应该表现不佳的域分类器。

**4.1.3更多相关工作**

基于对抗性或非对抗性训练的无监督领域适应方法还有其他变体。例如，域混淆[56]提出了一个目标，在该目标下，通过以相同的方式查看两个域，训练两个不带符号的表示映射到均匀分布上。CoGAN[57]训练两个分别生成源图像和目标图像的gan。通过将两个GANs的高级参数连接起来，实现了区域不变性，并根据判别器的输出对分类器进行训练。

**4.2 变换等变与不变性**

一个更具理论意义的课题是揭示学习表征中变换等价性与不变性之间的内在联系。一方面，对变换等变表示（TER）的追求被认为是实现无监督学习的最新性能的关键标准之一[60]。然而，在有监督的图像和目标识别任务中，应用变换不变性来训练带有标记数据的判别网络也是非常重要和必要的。

乍一看，同时执行两个标准似乎是个两难的选择，但它们实际上在支持深度学习巨大成功的著名卷积神经网络中很好地共存——卷积特征映射等同于翻译，而输出预测在各种转换下应该是不变的[1]。最近在将翻译等价性推广到泛型转换方面的工作[5]也展示了在上面训练更强大的表示和判别模型以解决小数据挑战[5]、[9]、[10]的巨大潜力。

然而，对于变换等价性和变换不变性之间的关系的深入理解，对于如何弥合训练无监督模型和有监督模型之间的差距，仍然缺乏。在无监督和有监督的模型中，变换等变和不变性的基本作用是毋庸置疑的，但我们仍然不知道如何将它们以一致的方式集成。

事实上，无监督表示学习更关注新任务的泛化性，而有监督任务更关注给定任务的区分能力。如何将变换等变性和不变性的追求恰当地结合起来，在泛化和判别之间达到更好的平衡？我们是否还应该将变换等价表示的无监督学习与变换不变分类器的有监督训练分离开来？我们相信，对这些问题的有见地的回答可以带来更具变革性和效率的方式，将这两个原则结合起来，以应对每天出现的新任务的小数据挑战。这是我们今后要回答的一个基本问题。

**5．结论**

在本文中，我们回顾了两大类小数据方法-无监督和半监督方法。特别地，在有监督和半有监督的情况下，我们回顾了大量的生成模型，包括自动编码器、GANs、基于流的模型和自回归模型类别。我们还比较了培训这些模型时出现的一些标准和原则，例如训练无监督和有监督表示的变换等价性和不变性，以及分解和可解释深网络的无监督和半监督表示的分离。本文还综述了无监督域和半监督域的自适应，以揭示在无标记和有标记数据存在的情况下，如何弥合不同域分布之间的差距的最新进展。讨论了变换等变性和不变性在小数据模型训练中的基本作用。