**MixMatch: A Holistic Approach to Semi-Supervised Learning**

代码：https://github.com/google-research/mixmatch

**Abstract**

已证明半监督学习是利用未标记数据来减轻对大型标记数据集的依赖的强大范例。 在这项工作中，我们统一了当前用于半监督学习的主流方法，以产生一种新算法MixMatch，该算法猜测数据增强的未标记示例的低熵标记，并使用MixUp混合标记和未标记的数据。 MixMatch可以在许多数据集和标记的数据量上大幅度获取最新的结果。 例如，在具有250个标签的CIFAR-10上，我们将错误率降低4倍（从38％降至11％），而在STL-10上降低2倍。 我们还演示了MixMatch如何帮助实现差异性隐私的显着更好的准确性-隐私权衡。 最后，我们进行了消融研究，以弄清楚MixMatch的哪些成分对其成功最重要。

**1 Introduction**

最近在训练大型深度神经网络方面取得的成功，在一定程度上要归功于大型标记数据集的存在。 然而，对于许多学习任务而言，收集标记数据非常昂贵，因为它必然涉及专家知识。 最好用医疗任务来说明这一点，在这种情况下，测量需要昂贵的设备，而标签则是耗时的分析的成果，该分析是从多位人类专家那里得出的。 此外，数据标签可能包含私人信息。 相比之下，在许多任务中，获取未标记的数据要容易得多或便宜得多。

半监督学习[6]（SSL）试图通过允许模型利用未标记的数据来大大减轻对标记数据的需求。 半监督学习的许多最新方法都增加了一个损失项，该损失项是根据未标记的数据计算的，并鼓励该模型更好地推广到未见的数据。 在最近的工作中，该损失项属于以下三类之一（在第2节中进一步讨论）：熵最小化[18，28]-鼓励模型对未标记的数据输出可信的预测； 一致性正则化-鼓励模型在其输入受到干扰时产生相同的输出分布； 和通用正则化-鼓励模型很好地概括并避免过度拟合训练数据。

在本文中，我们介绍了混合算法（MixMatch），这是一种SSL算法，它引入了一种单一损失，可以很好地统一这些主要方法用于半监督学习。 与以前的方法不同，MixMatch一次性针对所有属性，我们发现这些属性带来以下好处：

通过实验，我们证明MixMatch在所有标准图像基准（第4.2节）上均获得了最新的结果，并将CIFAR-10的错误率降低了4倍； 我们在消融研究中进一步表明，MixMatch大于其各部分的总和； 我们在第4.3节中演示了MixMatch对于差异化的私人学习非常有用，它使PATE框架中的学生[36]获得了最新的最新结果，同时增强了隐私保障和准确性。

简而言之，MixMatch为未标记的数据引入了统一的损失项，可以无缝地减少熵，同时保持一致性并保持与传统正则化技术的兼容性。

**2 Related Work**

为了为MixMatch做好准备，我们首先介绍SSL的现有方法。 我们主要关注那些基于MixMatch的最新技术。 关于SSL技术的文献很多，我们在这里不予讨论（例如，“转换”模型[14、22、21]，基于图的方法[49、4、29]，生成模型[3、27、41， 9、17、23、38、34、42]等）。 在[49，6]中提供了更全面的概述。 在下文中，我们将引用通用模型pmodel（y j x;），该模型将为具有参数的输入x在类标签y上产生分布。

**2.1一致性正则化**

监督学习中一种常见的正则化技术是数据增强，它应用了假定不影响类语义的输入转换。 例如，在图像分类中，通常会弹性变形或向输入图像添加噪声，这可以在不更改其标签的情况下显着改变图像的像素内容[7、43、10]。 粗略地说，这可以通过生成几乎无限的新数据，经过修改的数据流来人为地扩大训练集的大小。一致性正则化通过利用分类器即使在扩展后也应为未标记的示例输出相同的类分布的思想，将数据扩展应用于半监督学习。 更正式地讲，一致性正则化要求将未标记的示例x分类为Augment（x），这是对自身的扩充。

在最简单的情况下，对于未标记的点x，先前的工作[25，40]加上损失项



请注意，Augment（x）是随机变换，因此等式（1）中的两个项不一样。 “平均教师” [44]替换了等式（1）中的一个术语，使用模型参数值的指数移动平均值进行模型输出。 这提供了一个更稳定的目标，并凭经验发现可显着改善结果。 这些方法的缺点是它们使用特定于域的数据增强策略。“虚拟对抗训练” [31]（VAT）通过替代地计算附加扰动来解决此问题，以将其应用于输入，从而最大程度地改变输出类别的分布。 MixMatch通过对图像（随机水平翻转和裁剪）使用标准数据增强来利用一致性正则化的形式。

**2.2 Entropy Minimization**

在许多半监督学习方法中，一个常见的基本假设是分类器的决策边界不应通过边缘数据分布的高密度区域。

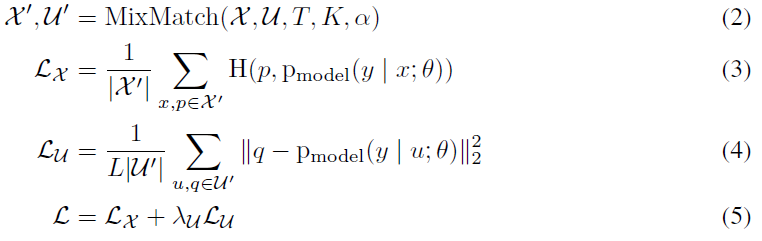
强制执行此操作的一种方法是要求分类器对未标记的数据输出低熵预测。 这在[18]中用损失项明确完成，该项使未标记数据x的的熵最小。 在[31]中，这种形式的熵最小化与VAT相结合以获得更强大的结果。 “ Pseudo-Label” [28]通过根据未标记数据的高可信度预测构建硬（1-hot）标记并将其用作标准交叉熵损失的训练目标，来隐式地最小化熵。 MixMatch还通过在目标分布上针对未标记数据使用“锐化”功能来隐式实现熵最小化，如3.2节所述。

**2.3传统正则化**

正则化是指对模型施加约束的一般方法，以使其难以记忆训练数据，因此希望使其更好地泛化到未见的数据[19]。 我们使用权重衰减惩罚模型参数的L2范数[30，46]。 我们还使用MixMatch中的MixUp [47]来鼓励示例之间的凸出行为。 我们将MixUp既用作正则化器（应用于标记的数据点），又将其作为半监督学习方法（应用于未标记的数据点）。 MixUp先前已应用于半监督学习； 特别是，[45]的并行工作使用了MixMatch中使用的方法的子集。 我们澄清了我们的消融研究中的差异（第4.2.3节）。

**3 MixMatch**

在本节中，我们将介绍我们提出的半监督学习方法MixMatch。 MixMatch是一种“整体”方法，结合了第2节中讨论的SSL主导范式的思想和组件。给定批处理的个标记示例具有一个热门目标（代表L个可能的标签之一）和大小相等的批 对于未标记的示例，MixMatch会生成一批经过处理的增强标记的示例X’和一批带有“猜测”标记的增强的未标记示例。 然后，将和X’用于计算单独的标记和未标记损耗项。 更正式地说，半监督学习的组合损失L定义为



其中H（p; q）是分布p和q之间的交叉熵，T，K，和U是下面描述的超参数。 算法1中提供了完整的MixMatch算法，图7中显示了标签猜测过程的示意图。接下来，我们描述MixMatch的每个部分。

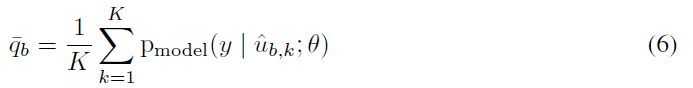
**3.1 Data Augmentation**

正如许多SSL方法中的典型做法一样，我们在标记和未标记的数据上都使用数据增强。对于这批标记数据X中的每个xb，我们生成一个转换后的版本（算法1，第3行）。 对于这批未标记数据U中的每个ub，我们生成K个增量

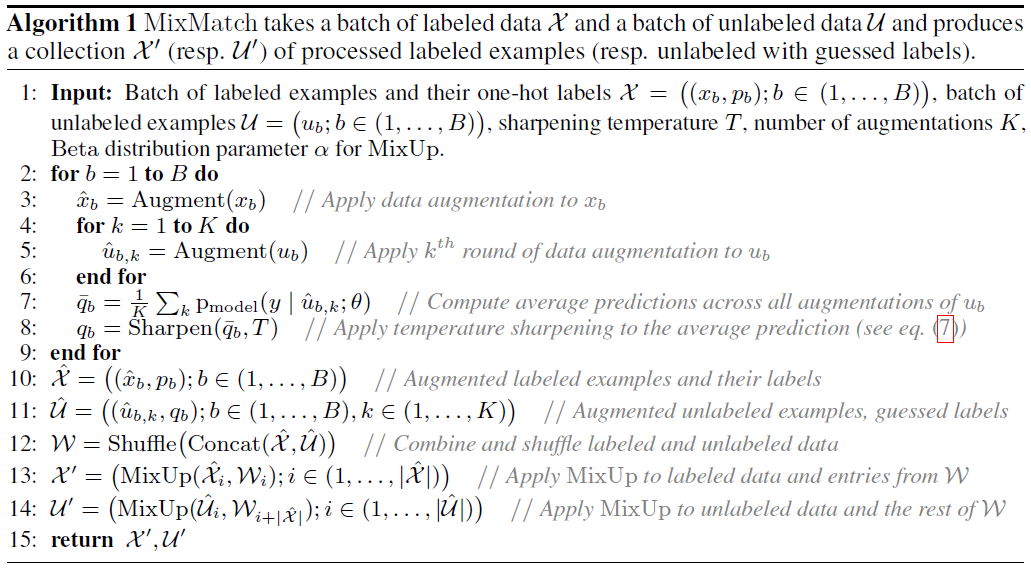
（算法1，第5行）。 通过以下小节中描述的过程，我们使用这些单独的扩充为每个ub生成一个“猜测标签” qb。

**3.2 Label Guessing**

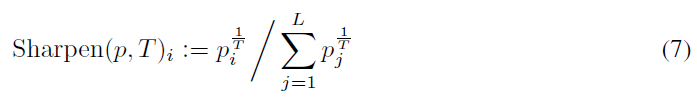
对于U中每个未标记的示例，MixMatch使用模型的预测为该示例的标签生成一个“猜测”。 稍后将这种猜测用于无监督损失项中。 为此，我们通过ub的所有K个增量计算模型的预测类分布的平均值



在算法1，第7行中。在一致性正则化方法中，通常使用数据扩充来获得未标记示例的人工目标[25、40、44]。



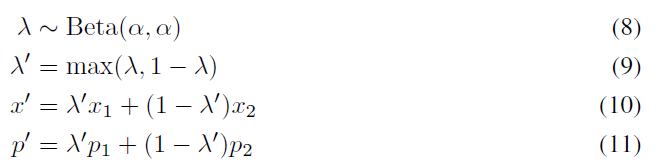
**锐化**。在产生标签猜测时，我们从半监督学习中成功实现了熵最小化（在2.2节中进行了讨论）的启发，又执行了另一步。 给定对扩增qb的平均预测，我们应用了锐化函数来减少标签分布的熵。 在实践中，对于锐化功能，我们使用调整此分类分布[16]的“温度”的通用方法，这被定义为操作



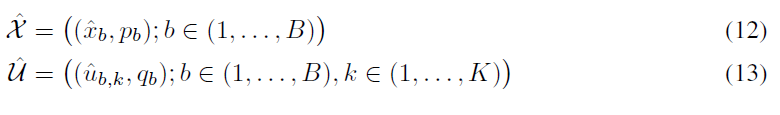
其中p是一些输入类别分布（特别是在MixMatch中，p是在增强qb上的平均类别预测，如算法1，第8行所示），T是超参数。 作为时，Sharpen（p; T）的输出将接近Dirac（“one-hot”）分布。 由于我们稍后将使用作为模型针对ub增大的预测的目标，因此降低温度会鼓励模型产生较低熵的预测。

**3.3 MixUp**

我们使用MixUp进行半监督学习，与SSL的以往工作不同，我们将带标签的示例和未带标签的示例与标签猜测（如3.2节所述生成）混合在一起。 为了与我们的单独损失条款兼容，我们定义了MixUp的稍作修改的版本。 对于带有相应标签概率的两个示例对我们计算：



为超参数，Vanilla MixUp省略了（9）（即设置为）。 鉴于已标记和未标记的示例在同一批次中串联在一起，我们需要保留批次的顺序以适当地计算各个损失成分。 这是通过等式（9）实现的，确保比x2更接近x1。 要应用MixUp，我们首先将收集带有标签的所有增强标签的示例和带有其猜测标签的所有未标签的示例



（算法1，第10-11行）。 然后，我们将这些集合合并，并将结果混洗以形成W，它将用作MixUp的数据源（算法1，第12行）。 对于中的第ith个示例标签对，我们计算并将结果添加到集合中（算法1，第13行）。 我们针对计算 ，有意地使用W的其余部分，而的构造中没有使用过W的剩余部分（算法1， 第14行）。 总而言之，MixMatch将X转换为，这是带有数据增强和MixUp（可能与未标记的示例混合）的标记的示例的集合。 类似地，U转换为，即每个未标记示例的多个增强的集合以及相应的标记猜测。

**3.4 Loss Function**

给定我们已处理的批次和，我们使用等式（3）至（5）中所示的标准半监督损耗。 公式（5）将标签和的模型预测之间的典型交叉熵损失与预测和的猜测标签的平方L2损失相结合。 我们在等式（4）中使用此L2损失（多类Brier得分[5]），因为与交叉熵不同，它有界且对错误的预测较不敏感。 由于这个原因，它经常被用作SSL中的未标记数据丢失[25，44]以及预测不确定性的度量[26]。 我们不会像标准[25、44、31、35]那样通过计算猜测的标签来传播梯度

**3.5 Hyperparameters**

由于MixMatch结合了多种机制来利用未标记的数据，因此引入了各种超参数-特别是锐化温度T，未标记的扩增数K，MixUp中Beta的参数以及无监督的损失权重。 在实践中，具有许多超参数的半监督学习方法可能会出现问题，因为使用较小的验证集很难进行交叉验证[35、39、35]。 但是，实际上，我们发现大多数MixMatch的超参数都是固定的，不需要根据每个实验或每个数据集进行调整。具体而言，对于所有实验，我们将T设为0.5，将K设为2。此外，我们仅基于每个数据集更改和； 我们发现= 0.75和 = 100是调整的良好起点。 在所有实验中，按照惯例，我们会在训练的前16,000步中将U线性增加至最大值。[44]

**4 Experiments**

我们在标准SSL基准测试（第4.2节）上测试MixMatch的有效性。 我们的消融研究将MixMatch每种成分的贡献分开（第4.2.3节）。 作为附加应用程序，我们将在第4.3节中考虑保护隐私的学习。

**4.1 Implementation details**

除非另有说明，否则在所有实验中，我们均使用[35]中的“ Wide ResNet-28”模型。 除了以下差异之外，我们对模型和训练过程的实现与[35]的实现非常匹配（包括使用5000个示例选择超参数）：首先，我们使用指数移动平均值为，而不是降低学习率，而是对模型进行了评估。 其参数的衰减率为0.999。 其次，对于Wide ResNet-28模型，我们在每次更新时都应用0.0004的权重衰减。 最后，我们每216个训练样本检查一个点，并报告最后20个检查点的中值错误率。 例如，通过平均检查点[2]或选择具有最小验证误差的检查点，可以简化分析过程，但会降低准确性。

**4.2 Semi-Supervised Learning**

首先，我们评估MixMatch在四个标准基准数据集上的有效性：CIFAR-10和CIFAR-100 [24]，SVHN [32]和STL-10 [8]。 在前三个数据集上评估半监督学习的标准做法是将大多数数据集视为未标记，而将一小部分用作标记数据。 STL-10是专门为SSL设计的数据集，具有5,000个标记的图像和100,000个未标记的图像，这些图像是从与标记的数据略有不同的分布中得出的。

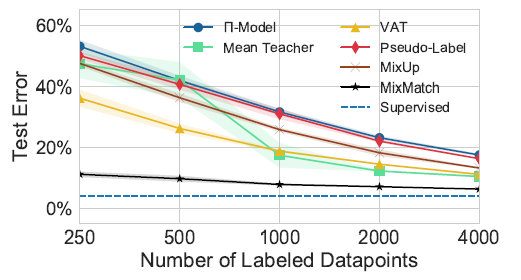


图2：对于不同数量的标签，MixMatch与CIFAR-10上基线方法的错误率比较。 确切数字见表5（附录）。 “监督”是指使用所有50000个培训示例进行培训，并且没有未标记的数据。 具有250个标签的MixMatch达到的错误率可与拥有4000个标签的次佳方法的性能相媲美。

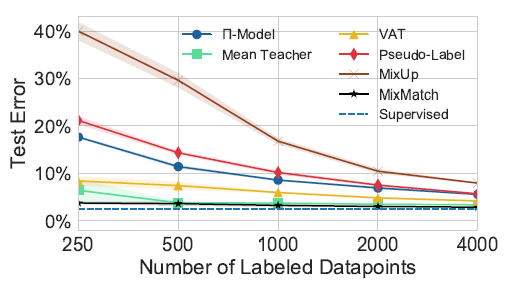


图3：对于不同数量的标签，MixMatch与SVHN上基线方法的错误率比较。 确切数字在表6（附录）中提供。 “监督”是指使用所有73257个培训示例进行培训，并且没有未标记的数据。 通过250个示例，MixMatch几乎达到了该模型监督训练的准确性。

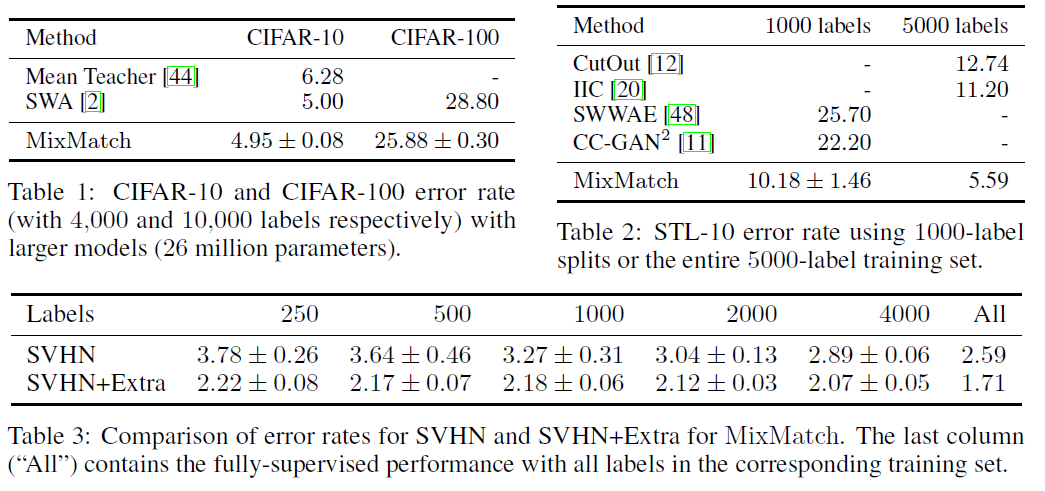
**4.2.1 Baseline Methods**

作为基线，我们考虑在第2节中介绍的[35]中考虑的四种方法（-模型[25，40]，平均教师[44]，虚拟对抗训练[31]和伪标签[28]）。 我们还单独使用MixUp [47]作为基准。 MixUp被设计为用于监督学习的正则化器，因此我们将其应用于SSL进行修改，方法是将其应用于具有扩展标记的示例和具有其相应预测的扩展非标记示例。 根据MixUp的标准用法，我们在MixUp生成的猜测标签和模型的预测之间使用了交叉熵损失。 正如[35]所提倡的，我们在相同的代码库中重新实现了每种方法，并将它们应用于相同的模型（在第4.1节中进行了描述），以确保公平地进行比较。 我们重新调整了每种基线方法的超参数，与[35]中的方法相比，通常会导致边际精度提高，从而为测试MixMatch提供了更具竞争力的实验设置.

**4.2.2 Results**

CIFAR-10 对于CIFAR-10，我们使用从250到4000的不同数量的带标记示例（按照标准惯例）来评估每种方法的准确性。 结果可以在图5中看到。 2.对于CIFAR-10，我们使用= 75。 我们为每个标记点数创建了5个分割，每个分割点具有不同的随机种子。 每个模型都在每个分割上进行训练，错误率由各个分割的均值和方差来报告。 我们发现MixMatch的性能大大优于其他所有方法，例如4000个标签的错误率达到6：24％。 作为参考，在同一模型上，对所有50000个样本进行完全监督的训练可获得4：17％的错误率。此外，MixMatch仅使用250个标签即可获得11：08％的错误率。 为了进行比较，在250个标签上，次佳的方法（VAT [31]）的错误率达到36:03，比MixMatch高出4：5，这是因为我们的模型获得的错误限制为4：17％ 在完全监督的学习中 另外，在4000个标签上，性能次佳的方法（Mean Teacher [44]）获得10：36％的错误率，这表明MixMatch可以在只有1 / 16个标签的情况下达到类似的性能。 我们相信，最有趣的比较是带有很少标记的数据点，因为它揭示了方法的采样效率，这对于SSL至关重要。

具有更大模型的CIFAR-10和CIFAR-100一些先前的工作[44，2]还考虑了使用更大的2600万参数模型。 我们在[35]中使用的基本模型只有1：5百万个参数，这与这些结果的比较令人困惑。 为了与这些结果进行更合理的比较，我们测量了增加基本ResNet模型的宽度的效果，并评估了MixMatch在28层Wide Resnet模型上的性能，该模型每层有135个滤镜，产生了2600万个参数。 我们还将在具有10000个标签的CIFAR-100上对该更大的模型进行MixMatch评估，以与[2]中的相应结果进行比较。 结果示于表1。 通常，MixMatch匹配或优于[2]的最佳结果，尽管我们注意到由于[44，2]的模型还使用了更多的结果，因此比较仍然存在问题复杂的“摇动”正则化[15]。 对于此模型，我们使用了0，0008的权重衰减。 对于CIFAR-10，我们使用U = 75；对于CIFAR-100，我们使用 = 150。



SVHN和SVHN + Extra与CIFAR-10一样，我们评估SVHN上每种SSL方法的性能，标签数量从250到4000不等。按照惯例，我们首先考虑设置73257示例训练集为 分为标记和未标记的数据。 结果如图3所示。我们使用 =250。在这里，针对每个标记点数（每个都有不同的随机种子）对5个分割进行模型评估。 我们发现，在所有标记数据中，MixMatch的性能相对稳定（并且优于所有其他方法）。 令人惊讶的是，经过额外的调整，我们能够从Mean Teacher [44]中获得出色的性能，尽管其错误率始终比MixMatch的错误率略高。

请注意，SVHN有两个训练集：train和extra。 在完全监督的学习中，将这两个集合连接起来以形成完整的训练集合（604388个样本）。 在SSL中，由于历史原因，多余的设置被搁置，仅使用火车（73257个样本）。 我们认为，利用未标记数据的训练和额外优势会更有趣，因为它显示出未标记样本比已标记样本更高的比例。 我们在表3中报告SVHN和SVHN + Extra的错误率。对于SVHN + Extra，我们使用= 0:25；  = 250，并且由于可用数据量较大，重量衰减较小，为0：000002。 我们发现，在两个训练集上，MixMatch几乎都几乎立即在同一训练集上达到了完全监督的性能–例如，与SVHN + Extra相比，MixMatch在SVHN + Extra上只有250个标签，实现了2：22％的错误率。 监督绩效为1：71％。 有趣的是，在SVHN + Extra MixMatch上，对于每个考虑的标记数据量，在SVHN上的表现均优于完全监督的训练，而没有额外的（2：59％错误）。 为了强调这一点的重要性，请考虑以下情形：您可以从SVHN中获得73257个示例，其中有250个示例被标记，并且可以选择：您可以再获取8个未标记的数据并使用MixMatch，也可以获取293个更多的标记数据并充分使用 监督学习。 我们的结果表明，获得更多的未标记数据并使用MixMatch更为有效，这可能比获得293个以上的标签便宜得多。

STL-10 STL-10包含5000个训练示例，旨在与10个预定义的折叠一起使用（我们仅使用前5个折叠），每个折叠有1000个示例。 但是，所有5000个示例都需要进行一些先前的培训。 因此，我们在两种实验设置中进行比较。 使用1000个示例，MixMatch不仅超越了1000个示例的最新技术水平，而且超越了所有带有5000个标记示例的最新技术水平。请注意，表2中的基线均未使用相同的实验设置（即模型），因此很难直接比较结果； 但是，由于MixMatch获得的最低误差是原来的两倍，因此我们将其视为对我们的方法的信任。我们使用 = 50。

**4.2.3 Ablation Study**

由于MixMatch结合了各种半监督学习机制，因此与文献中的现有方法有很多共通之处。 结果，我们研究了去除或添加组件，以便进一步了解MixMatch的性能。具体来说，我们衡量的是

使用K个扩充的均值类别分布或针对单个扩充使用类别分布（即，设置K = 1）

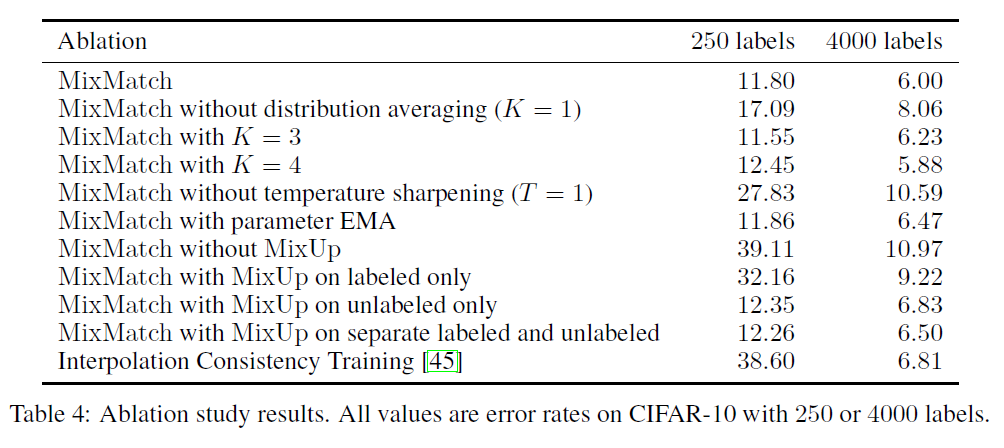
消除温度锐化（即设置T = 1）

产生猜测标签时，使用模型参数的指数移动平均值（EMA），如Mean Teacher所做的[44]

仅在标记的示例之间，仅在未标记的示例之间执行MixUp，并且不会在标记和未标记的示例之间进行混合

使用插值一致性训练[45]，可以将其视为该消融研究的特例，其中仅使用未标记的混合，不应用锐化，而EMA参数用于标签猜测。

我们在具有250和4000个标签的CIFAR-10上进行了消融； 结果示于表4。我们发现每种成分都对MixMatch的性能有所贡献，其中250标签设置的差异最为明显。 尽管Mean Teacher在SVHN上很有效（图3），但我们发现使用类似的EMA参数值会稍微损害MixMatch的性能。



**4.3 Privacy-Preserving Learning and Generalization**

私下学习可以让我们评估方法的概括能力。 确实，保护训练数据的隐私等于证明该模型不会过拟合：如果添加，修改或删除其训练样本中的任何一个，则学习算法被认为是差分私有的（最广泛接受的隐私技术定义）。 保证不会导致学习到的模型参数有统计学上的显着差异[13]。 因此，在实践中，具有差异性隐私的学习是一种正规化形式[33]。 每次培训数据访问都会构成潜在的隐私泄漏，将其编码为一对输入及其标签。 因此，从私人训练数据进行深度学习的方法（例如DP-SGD [1]和PATE [36]）受益于在计算模型参数更新时访问尽可能少的带标签私人训练点。 半监督学习非常适合这种情况。

我们使用PATE框架进行隐私学习。 从公开的，未标记的数据中对学生进行半监督的培训，部分数据由一群可以访问私有的，已标记数据的教师标记。 学生达到固定准确度所需的标签越少，它提供的隐私保证就越强。 教师使用嘈杂的投票机制来回答学生提出的标签问题，当他们无法达成足够强烈的共识时，他们可以选择不提供标签。 因此，如果MixMatch改善了PATE的性能，那么这也将说明MixMatch从每个类的几个典范范例中得到的改进的概括性。

我们将通过MixMatch实现的准确性-隐私权衡与SVHN的增值税[31]基线进行比较。VAT达到了以前最先进的91：6％的测试准确度，隐私丢失率为“ = 4:96 [37]。由于MixMatch在标记点很少的情况下表现良好，因此可以实现95:21+ 0：17％的测试准确率，隐私损失“ = 0:97”要小得多。因为使用e来测量隐私度，所以改进约为=55，这是一项重大改进。 低于1的隐私损失表示更强大的隐私保证。请注意，在私人培训设置中，学生模型仅使用10,000个示例。

**5 Conclusion**

我们引入了MixMatch，这是一种半监督学习方法，它结合了当前SSL主流范式的思想和组成部分。 通过对半监督和隐私保护学习的大量实验，我们发现，在我们研究的所有设置中，MixMatch与其他方法相比，其表现出显着改善的性能，通常降低了两个或两个以上的错误率。 在未来的工作中，我们有兴趣将半监督学习文献中的其他思想纳入混合方法，并继续探索哪些成分会产生有效的算法。 另外，大多数现代半监督学习算法的工作都是根据图像基准进行评估的。 我们有兴趣探索MixMatch在其他领域的有效性。