经噪声标签训练的深度神经网络的理解与应用

摘要

噪声标签在现实世界的数据集中是无处不在的，这提出了要强健地训练深度神经网络（DNN）的挑战，因为DNN通常具有很高的能力来记住噪声标签。 在本文中，我们发现可以根据数据集中的噪声比对测试精度进行定量表征。特别是在对称噪声的情况下，测试精度是噪声比的二次函数，这解释了先前发表的实验结果。根据我们的分析，我们将交叉验证应用于随机分裂的嘈杂数据集，该数据集可识别大多数具有正确标签的样本。然后，我们采用共同教学策略，该策略充分利用已识别的样本来针对噪声标签对DNN进行鲁棒训练。与广泛的最新方法相比，我们的策略在合成和实际训练噪声下均不断提高DNN的泛化性能。

1引言

DNN在监督学习任务上的显著成功很大程度上取决于大量带有准确标签的训练样本。 正确地标记大量数据的成本太高，而诸如众包（Yan等，2014； Chen等，2017）和在线查询（Schroff等，2011； Divvala等，2014）等替代方法可以廉价地获取数据， 但不可避免地会产生嘈杂的标签。带有太多噪声标签的训练会降低DNN的泛化性能，因为网络很容易在损坏的标签上过度拟合（Zhang等人，2017； Arpit等人，2017）。 要利用大量的噪声数据，了解噪声标签如何影响DNN的训练和泛化是第一步。在此基础上，我们可以设计特定的方法来在实际应用中稳健地训练DNN。

已经提出了许多处理噪声标签的方法。有几种方法专注于估计噪声转换矩阵并相应地校正目标函数，例如前向或后向校正（Patrini等人，2017），S模型（Goldberger＆Ben-Reuven，2017）。然而，准确估计噪声转换矩阵是一个挑战。另一种方法是对选定或加权的样本进行训练，例如解耦（Malach＆Shalev-Shwartz，2017），MentorNet（Jiang等，2018），基于梯度的加权（Ren等，2018）和联合教学（Han et al。，2018）。剩下的问题是设计一个可靠且令人信服的选择或加权样本的标准。另一种方法建议使用DNN的预测来校正标签，例如Bootstrap（Reed等人，2015），Joint Optimization（Tanaka等人，2018）和D2L（Ma等人，2018），它们都容易受到过度拟合。为了提高鲁棒性，联合优化引入了正则化项，要求先验知识了解实际类别在所有训练样本之间的分布方式。但是，在实践中通常没有先验知识。

嘈杂的标签如何影响DNN的训练和泛化还没有得到很好的理解，这值得引起更多关注，因为它可能会促进针对噪声进行鲁棒训练DNN的基本方法。在没有标签损坏的情况下，泛化错误会受到复杂性度量的限制，例如VC维数（Vapnik，1998），Rademacher复杂度（Bartlett＆Mendelson，2002）和统一稳定性（Mukherjee et al。，2002; Bousquet＆Elisseeff，2002; Poggio等人，2004）。但是，在存在嘈杂标签的情况下，边界变得微不足道。张等。 （2017）证明了DNN具有很高的适应随机标签的能力，但泛化误差较大。张等。 （2017）也显示了泛化误差与噪声比之间的正相关，这意味着DNN确实从噪声数据中捕获了一些有用的信息。 Arpit等。 （2017）表明，在训练过程中，DNN倾向于先学习简单的模式，然后逐渐记住所有样本，这证明了广泛使用的小损失标准：将训练损失小的样本视为干净的样本（Han等，2018; Jiang等人，2018）。 Ma等。（2018）定性地将DNN的泛化性能差归因于潜在特征子空间的维数增加。通过广泛的实验，这些工作获得了对经过噪声标签训练的DNN有趣行为的经验见解，而理论和定量的解释仍未出现。

在本文中，我们可以定量地阐明通常使用噪声标签训练的DNN的泛化性能。 为了验证我们的理论分析，我们使用交叉验证对一组收集的被某些噪音污染的样本进行随机拆分。 DNN可以在一个子集上进行训练，然后在其余数据集上进行评估，以比较理论和经验结果在泛化性能上的表现。 我们发现，DNN可以精确地拟合有噪声的训练集，并在分布上进行泛化(详见权利要求1)。 因此，我们可以根据数据集中的噪声比定量地描述测试准确性。 特别是在对称噪声的情况下，测试精度是噪声比的二次函数。 在张等人。 （2017），根据经验发现DNN的泛化性能高度依赖于噪声比。 我们的贡献之一是对他们的经验发现进行彻底的解释。

根据我们的分析，我们进一步开发了一种针对噪声标签训练DNN的特定方法。我们的方法是在共同教学策略的基础上开发的，该策略首先在Blum＆Mitchell（1998）中提出，然后在（Han et al。，2018）中进行了修改以处理具有令人印象深刻性能的嘈杂标签。在协同教学策略中，一个同时训练两个网络:从整个有噪声的训练集中抽取小批量样本，然后每个网络选择一定数量的小损失样本，并将其反馈给它的对等网络。但是，当训练集的噪声比增加时，联合教学的性能会严重下降。而且，根据训练集的噪声比来设置每个小批量中选择的小损失样本的数量，这在实践中是不可用的。幸运的是，我们可以基于对DNN泛化性能的理论分析来解决这些问题。特别地，我们提出了迭代噪声交叉验证（INCV）方法来选择样本子集，该子集的噪声比比原始数据集小得多，从而使DNN的训练过程更加稳定。此外，我们可以自动估计所选集合的噪声比，这使我们的方法更适合工业应用。简而言之，我们的主要贡献是

•理论上将DNN的泛化性能与标签噪声相关联，

•选择干净标签和训练抗噪DNN的实用算法。

合成和真实噪声标签的实验表明，与最新方法相比（Patrini等人，2017; Malach＆Shalev-Shwartz，2017; Han等人，2018; Jiang等人， 2018; Ma et al。，2018），使用我们的策略训练的DNN在干净测试集上达到最佳测试准确性。 特别是，我们的方法在（i）CIFAR-10数据集（Krizhevsky＆Hinton，2009）上得到了验证，该数据具有通过随机翻转原始标签而生成的合成噪声标签，以及（ii）WebVision数据集（Li等人，2017） ，这是一个大型基准，包含从网站抓取的240万张图片，其中包含真实的嘈杂标签。

2.预备

对于一个c类分类，我们收集数据集，其中是第t个样本，其观察到的标签为。 如前所述，观察到的标签y可能会损坏，因为示例x经常通过在线查询或众包系统进行标签。令y表示真实标签，我们可以通过引入噪声转换矩阵来描述集合D的破坏过程，其中表示概率D。 将第i个类示例标记为j。 在交叉验证中，我们将收集的样本D随机分为两半和。 这样，和共享相同的噪声转换矩阵T。 令表示由参数化的神经网络，而表示由网络给出的x的预测标签。

3. 理解经过噪音标签训练的DNNs

（Zhang et al。，2017）中的大量实验表明，DNN可以拟合训练集中包含的嘈杂甚至随机的标签，但是即使在具有相同噪声的测试集中，泛化误差也很大。 在本节中，我们使用先前介绍的噪声过渡矩阵T对通常使用噪声标签训练的DNN的泛化性能进行理论上的量化，这完美地解释了（Zhang等人，2017年）报道的经验发现。

在经典的可能近似正确框架（Valiant，1984）中，良好的泛化性能意味着预测和观察到的测试标签y与随机变量近似相同，即对于每个测试样本x，它们应该相等。 如果没有标签损坏，泛化错误可以受到VC维数（Vapnik，1998年），Rademacher复杂度（Bartlett和Mendelson，2002年）等的限制。但是，在处理带有噪声标签的DNN时， = y可能不会 当在每个测试示例x上进行评估时保持不变，会导致较大的泛化误差（Zhang等人，2017）。 幸运的是，我们发现泛化仍然发生在分布的意义上，即分布的泛化，如以下claim1要求所示，在交叉验证中，我们将有噪声的数据集D随机分为两半和。

Claim 1.（分配中的泛化）。 令为在上训练并在上测试的网络。 如果我们假设

1. 观察到的输入示例x是集合D中的i.i.d.(独立同分布)
2. f具有足够高的容量，

则在上，预测真正的第i类测试样本的概率为j为

（1）

其中表示和共享的噪声转换矩阵

Claim 1揭示了这样的事实，即预测和测试标签y具有相同的分布。实际上，如果在上训练的模型在另一个带有真实标签的干净测试集上进行测试，则公式（1）仍然成立，但在这种情况下，这意味着将第i类测试样本预测为j的概率等于训练集的。我们将通过5.1中的实验证明Claim1的合理性。

测试准确度是一种广泛使用的度量标准，测试准确度被定义为预测观察到的标记y的测试示例的比例。在下面的命题1中，我们在测试集上制定测试精度

命题1。令和为两个具有相同噪声过渡矩阵T的数据集，为在上训练并在上进行测试的网络。 根据Claim1中的假设，对于任何类别i∈[c]的测试精度为

（2）

证明。 基于Claim1，和y具有以T为特征的相同分布。假设标签损坏过程是独立的，那么在测试集上，我们有

(3)

因此公式（2）遵从于

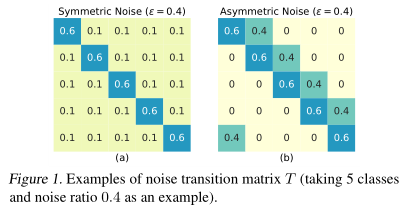
3.1 对称和非对称噪声

根据先前的文献（Ren等人，2018; Han等人，2018; Jiang等人，2018; Ma等人，2018），在本小节中，我们重点研究两种典型的噪声类型：对称噪声和非对称噪声可以定义如下（示例见图1），

定义1.

在比率为，的对称噪声的情况下，我们定义，并且，。

对于比率为，的非对称噪声，我们定义，对于某些定义，否则定义。



在对称和非对称噪声的情况下，我们可以使用噪声比 量化DNN的测试精度，这些DNN分别在前面提到的噪声数据集和上进行训练和测试。

**推论1.1** 对于比率为的对称噪声，测试精度为

(4)

对于比率为的非对称噪声，测试精度为

(5)

证明。根据命题1，我们有

注意，对于对称和非对称噪声，∀*i*∈[c]，是由给出的常数。因此，通过将插入等式，可以得出所需的结果。

有趣的是，等式（4）非常适合图1（c）所示的泛化精度的实验结果，并使我们能够根据实验测试的准确性估算数据集的噪声比

**4.针对噪声标签训练DNN**

在本节中，我们提出一种基于协同教学策略的方法，以针对噪声标签对DNN进行鲁棒训练。 正如前面介绍的那样，当训练集的噪声比增加时，协同教学的性能会严重下降并变得不稳定，这在我们的实验中得到了进一步证明。 为了解决这个问题，我们建议首先选择一个样本子集，该子集的噪声比比原始数据集小得多。

如果样本（x，y）的观测标签y等于其潜在真实类别y，则该样本是干净的。 但是，ˆ y实际上是不可用的。如果样本（x，y）的观察标记y等于网络给出的预测标记，我们建议将其识别为干净的。 如果我们要确定样本（x，y）是否干净，则应将该样本排除在训练集中。 一个直观的方法可以在算法1中找到，即噪声交叉验证（NCV）方法，其有效性将通过下面的理论分析和下一部分的大量实验证明。

|  |
| --- |
| 算法1：嘈杂的交叉验证（NCV）：从嘈杂的样本中选择干净的样本 |
| 输入：噪声数据集D，训练周期E  1：S=，初始化网络  2：将D随机分为两等分和  3：在数据集上训练，周期为E  4：选择样本，满足  5：重新初始化网络  6: 在数据集上训练，迭代周期为E  7：选择样本，满足  8:S=  输出：选定的集合S |

按照标准指标（Powers，2011），我们根据标签精度Label Precision（LP）（Han等人，2018）和标签召回Label Recall（LR）来衡量识别性能，

(6)

其中是如算法1中给出的选定子集，|| 表示一组中的样本数。 代表中纯净样品的比例，而代表S的纯净样品的占中所有纯净样品的比例。根据以上定义，所选集合S的噪声比。 对于任何i∈[c]类，我们也有LP和LR：

(7)

基于第3节中介绍的分析，我们在下面的命题2中量化算法1的性能。

**命题2** 使用算法1去选择干净的样本，我们有，

(8)

证明 根据算法1，我们可以重新表示公式（7）为

将式(2)和式(3)代入上述方程即可得到期望的结果。

4.1 对称和非对称噪声

由于,公式(8)可以推出

**推论2.1**

（9）

有趣的是，我们可以看到，对于对称噪声的，我们得到了不等式(9)的上界，对于非对称噪声，我们得到了不等式(9)的下界。 在对称和非对称噪声的情况下，我们还有，因此我们可以在随后的推论2.2中重新构造LP和LR。

**推论2.2**

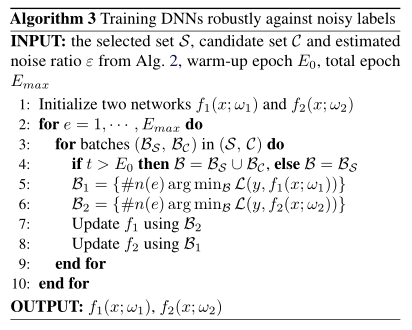
对于对称噪声比 ，我们有

, （10）

对于非对称噪声比 ，我们有

, （11）

给定由公式（4）或者（5）预估的数据集D的噪声比，给上述推论2.2进一步使我们能够估计指标LP和LR。 回想一下，根据LP的定义，所选子集S的噪声比为。 在实际情况下（，在中最大，j∈[c]），算法1始终会产生一个噪声比 <较小的子集。了解更多详情参见附录D。



4.2 使用INCV方法改善协同教学

虽然子集由Alg选择。 1通常具有比原始集合小的噪声比，DNN的鲁棒训练可能需要更多的训练样本。 为了解决此问题，我们提出了迭代噪声交叉验证（INCV）方法，以通过应用Alg增加选择的样本数。 1迭代。 INCV的更多详细信息可以在Alg中找到。 2.除了选择干净的样本外，INCC还会删除每次迭代中分类交叉熵损失较大的样本。 去除率r确定将去除多少样品。

在详细剖析了嘈杂的数据集D后，Alg。 如图2所示，我们可以进一步改善协同教学，以充分利用所选集S和候选集C的优势。 具体来说，我们让这两个网络在第一个E 0时期集中于所选集合S，然后合并候选集合C。 因此，训练稳定性和测试准确性均得到改善。 我们的方法的更多细节可以在Alg3中找到.

|  |
| --- |
| 算法2：迭代噪声交叉验证(INCV):从噪声样本中选出干净的样本 |
| 输入：噪声数据集D，迭代次数N，训练周期E，去除率r  1：选择集，候选集C = D  2：for i=1,…,N do  3： 初始化网络  4： 将C随机分为两等分和  5： 在数据集上训练，周期为E  6: 选择样本，满足  7： 确定数量的样本将会被去除    8: if i=1,使用公式(4)估算噪声比  9: 重新初始化网络  10: 在数据集上训练，周期为E  11: 选择样本，满足  12: 确定数量的样本将会被去除    13:  14:end for  输出：选定的集合S,剩余候选集合C和估计噪声比 |

5.实验

本节包括三个部分。 首先，我们通过实验验证了Sec中提出的理论结果。 3＆4。 然后，我们演示了Alg中显示的INCV方法。 2可以识别更多具有正确标签的样品。 最后，我们表明在Alg中概述了我们提出的方法。 3可以针对噪声标签强大地训练DNN，并且优于最新方法（Patrini等人，2017; Malach＆Shalev-Shwartz，2017; Han等人，2018; Jiang等人，2018; Ma 等人，2018）。 我们的代码位于https://github.com/chenpf1025/noisy\_label\_understanding\_utilizing。

实验装置。 为了验证我们的理论并测试算法，我们首先对通过随机破坏CIFAR-10中的原始la-bels产生的合成噪声标签进行实验（Krizhevsky＆Hinton，2009）。 我们专注于两种代表性的噪声类型：对称噪声和非对称噪声，如Def中所定义。 图1和图1所示。要验证我们在真实噪声标签上的方法，我们使用WebVision数据集（Li等，2017），该数据集包含使用ImageNet ILSVRC12中的1,000个概念从网站抓取的240万幅图像（邓等人）。 （2009年）。 WebVision的训练集包含许多没有人为注释的真实噪声标签。 Supp中提供了更多的实现细节。 答：在以下小节中，我们将重点放在实验结果和讨论上。

5.1 用噪声标签训练的DNN的行为

对于通常使用噪声标签训练的DNN，我们在理论上通过以下指标来表征其行为：（i）等式中给出的测试准确性。 （4）＆（5），（ii）等式中给出的LP。 （10）和（11）; （iii）等式中给出的LR （10）和（11）。 在本小节中，我们在广泛的实验中评估了这三个指标，并表明实验结果证实了我们的理论分析。 给定嘈杂的数据集D，我们实施交叉验证将其随机分为两半D 1，D 2，然后在D 1上训练ResNet-110（He et al。，2016b）在D 2上进行测试

实验结果证实了理论分析。 如图2所示，实验结果与理论估计一致。 特别是，图2（a）再现了（Zhang et al。，2017）中的观察结果，即测试精度高度依赖于噪声比。 （Zhang et al。，2017）并没有提出任何理论上的解释，但我们明确地用公式Eq。 （4）测试精度是噪声比的二次函数。 在图2（b）和（e）中，实验LP由我们的公式精确给出。 可以看出，对于某些数据点，实验测试精度和LR略小于我们的理论值。 这是合理的，因为D 2的分布与D 1不完全相同，并且即使没有噪声，泛化误差也不会变为0。

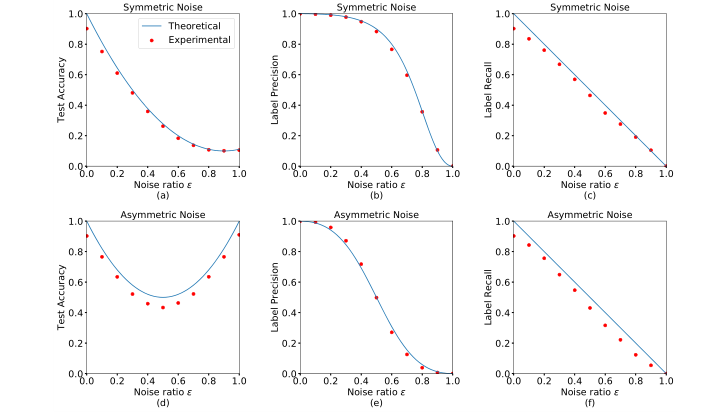


图2.手动损坏的CIFAR-10的测试精度，标签精度（LP）和标签召回率（LR），带噪声比。 第一行对应于对称噪声，第二行对应于非对称噪声。 交叉验证后，我们在嘈杂的数据集的一半上训练ResNet-110，并在其余一半上进行测试。 实验结果与理论曲线吻合。

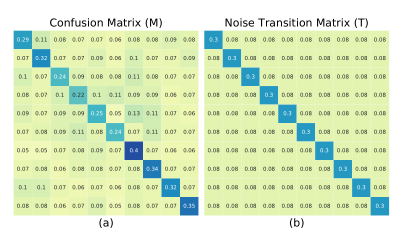


图3. RseNet-110，其通常是上训练的混淆矩阵手动损坏CIFAR-10与噪声转换矩阵T.中号≈T满足权利要求1中提出的语句。

为了进一步研究使用噪声标签训练的DNN的预测行为，我们定义了一个混淆矩阵M，其第ij个条目表示预测第i类测试样本的概率为j，s.t.，



图3展示了在人工破坏的CIFAR-10上以比率0.7的对称噪声训练的DNN的混淆矩阵，我们发现M≈T满足权利要求1中的陈述。更多的结果可以在Supp中找到。 B，我们证明M≈T仍然成立。

训练精度收敛到极低的值并不与我们的发现相矛盾。我们发现在较大的对称噪声下，模型的训练精度始终会收敛到极低的值。在实验中，当使用比率为0.7，0.8，0.9和1.0的对称噪声进行训练时，训练精度分别仅为0.58，0.40，0.24和0.36。但是，我们在图2和3中显示，我们的理论结果始终与实验结果一致。这种现象进一步提出了一个基本问题：高训练精度是学习和概括的必要条件吗？在没有数据扩充的情况下，有限样本表达性定理（Zhang等人，2017）表明DNN在有限数量的训练样本上始终可以实现0个训练误差。然而，在我们的实现中使用了标准数据增强（He等，2016a），这使得难以实现高训练精度，尤其是在大的对称噪声下。直观地，由于存在噪声标签，来自同一类别的附近样本可能具有不同的标签，从而需要对许多小区域进行不同的分类。增强容易产生违反先前学习的分类器区域的随机样本，因此增加了训练误差。即使在这种情况下，我们先前提出的理论公式仍然成立，如图2和3所示。在这里，我们得出结论，只要对足够丰富的深度神经网络进行足够多的步骤训练直至收敛，该网络就可以适合训练设置并归纳分布，即使标签上有嘈杂并且训练精度也很低。我们要求在将来对这种有趣的现象进行更多的理论解释。

5.2。 INCV鉴定更多干净的示例

图2（b）和（e）验证了Alg选择的子集。 1通常具有比原始设置小的噪声比。 有时，训练DNN需要大量训练样本。 在这里，我们演示了Alg。 2（INCV）可以通过迭代识别更多干净的样本。 为了提高效率，我们使用ResNet-32并设置N = 4，E = 50而不进行微调。 ε是使用公式自动估算的。 （4）在所有实验中。

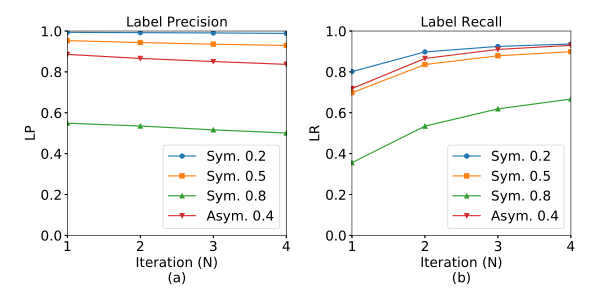


图4.手动损坏的CIFAR-10上INCV的LP和LR。 在每个图中，四个曲线分别对应于比率为0.2、0.5、0.8的对称噪声和比率为0.4的非对称噪声。

INCV可以准确识别大多数清洁样品。 图4示出了Alg的平均LP和LR值。 2，通过重复所有实验5次计算。 如图所示，即使经过一次迭代，LP和LR也比理论下限更好。 与Sec中使用的ResNet-110相比。 5.1，在本小节中，我们每次迭代仅训练50个时期的ResNet-32。 一个简单得多的模型自然会消除过拟合问题，从而产生更好的LP和LR。 此外，图4还表明LR随着迭代的增加而大大增加，而LP则略有下降。 经过四次迭代，INCV可以准确识别出大多数干净的样品。 例如，在比率为0.5的对称噪声下，它将选择大约90％（= LR）的干净样本，并且所选集合的噪声比率会降低到大约10％（= 1-LP）。

甚至在原始的CIFAR-10中也存在嘈杂的标签。 我们还对原始CIFAR-10运行INCV仅进行了1次迭代，并检查了被识别为损坏的样本。 有趣的是，有几个令人困惑的样本，如图5所示。这表明即使在原始CIFAR-10中也存在嘈杂的标签。 尽管CIFAR-10中包含的损坏样本很少见，对训练的影响可忽略不计，但能够识别它们意味着INCV是清除噪声标签的强大算法。

5.3。 针对噪声标签对DNN进行强大的培训

概述Alg.3，制定了联合教学以充分利用我们的INCB方法。 以下内容阐明了一些对Alg3的实际实现有用的问题。

* 问：如何设置从C和S提取的迷你批次B C和B S的尺寸？

答：通常，绘制小批量以便| BC | / | BS |合理。 = | C | / | S | 。 但是，当C很大时，会导致从C提取太多样本，这会损害训练过程，因为C通常包含许多损坏的样本。 因此，我们通过设置| B C | / | BS | |来稍微调整策略。 = min（0.5，| C | / | S |）。 在实验中，我们设置批次大小| BS | 到128，然后计算| BC | 相应地。

* 问：每个小批量样品中应保留多少样品？

答：在每个小批量生产中，我们使用训练损失较小的#n（e）个样本来更新网络，其中e是当前纪元。 在协同教学之后（Han et al。，2018），我们设置n（e）= | B S |（1-εS min（e / 10,1）），这意味着我们从| B S |中减去n（e）。 在前10个周期线性地| B S |（1-εS）固定，然后固定。 回想一下，εS = 1-LP表示S的噪声比。



图5. CIFAR-10中包含并由INCV识别的嘈杂标签。 原始标签标注在图像下方。 （a）被标记为卡车的人。 （b）标为卡车，实际上是汽车吗？ （c）玩具车上的一只鸟。 （d）标为飞机。 （e）卡车旁边的汽车。 （f）标记为猫。 （g）标为狗，实际上是一匹马？ （h）标为船舶。

可比的方法。 我们比较Alg。 3具有以下基准（1）F校正（Patrini等，2017）。 它首先训练网络以估计T，然后相应地校正损耗函数。 （2）去耦（Malach＆Shalev-Shwartz，2017）。 它在样本上训练了两个网络，而这两个网络的预测是不同的。 （3）共同教学（Han et al。，2018）。 它维护两个网络。 每个网络从迷你批次中选择训练损失较小的样本，并将其馈送到另一个网络。 （4）MentorNet（Jiang et al。，2018）。 教师网络是经过预训练的，它提供了一个样本加权方案来训练学生网络。 （5）D2L（Ma et al。，2018）。 对于每个样本，它将线性组合原始标签和网络预测作为新标签。 合并权重取决于潜在特征子空间的维数（Amsaleg 等人，2017）。

实验已手动损坏CIFAR-10。 我们首先通过手动破坏带有不同类型噪声的标签来评估CIFAR-10上的所有方法。 对于对称噪声，我们测试噪声比0.2，0.5和0.8。 对于非对称噪声，我们选择非平凡且具有挑战性的噪声比0.4，因为大于0.5的非对称噪声是微不足道的。 尽管如此，我们仍使用ResNet-32并将所有实验重复五次。

表1.在不同噪声类型和噪声比下的平均测试准确度（％，5次运行），带有标准偏差。 我们在手动损坏的CIFAR-10上训练RseNet-32，并在干净的测试集上进行测试。 最好的结果是用黑体标记。

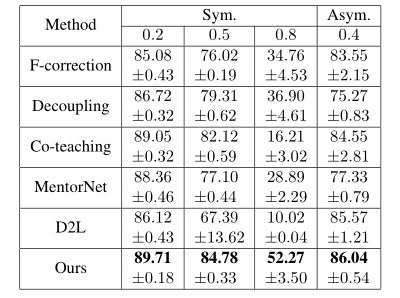
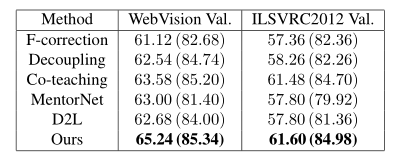


表2. WebVision验证集和ImageNet ILSVRC12验证集的验证准确性（％）。 括号外（内）的数字表示Top-1（Top-5）分类的准确性。 我们在WebVision训练集的前50个课程中训练了inception-resnet v2，其中包含真实的嘈杂标签。 最好的结果是用黑体标记。



如表1所示，在所有情况下，我们的方法始终可以达到最佳测试精度（以黑体字标出）。 即使对于大多数方法都具有挑战性的比率为0.8的对称噪声，我们也可以达到良好的测试精度。 图6展示了在每个训练时期之后，在干净测试集上所有方法的测试准确性。 可以发现，我们的方法在所有设置下都能令人印象深刻地达到最佳测试精度，而某些基线方法在训练的后期阶段会过度拟合，例如图6（b）和（d）中所示的F校正，去耦和MentorNet。 ），并在所有四个子图中显示D2L。 特别是，与联合教学相比（Han等人，2018），我们的方法进一步享有更稳定的训练过程，并且首先通过在干净的子集上进行训练获得了更好的测试准确性。

在实际的嘈杂标签上进行实验。 为了验证我们的方法在实际嘈杂标签上的实际用法，我们使用了WebVision数据集1.0（Li等人，2017），其训练集包含许多真实世界的嘈杂标签。 由于数据集很大，为了进行快速实验，我们使用Inception-Resnet v2比较了Google图像子集的前50个类的所有方法（Szegedy等人，2017）。 我们在人工注释的WebVision验证集和ILSVRC12验证集上测试训练后的模型。 如表2所示，在测试准确性方面，我们的方法始终优于其他最新技术。 而且，Supp。 C包含通过我们的INCV方法（图2）从WebVision数据集自动识别的一些嘈杂示例，这意味着INCV在包含实际噪声标签的数据集上是可靠的。

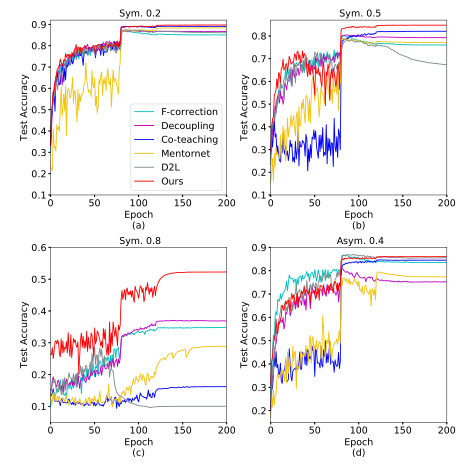


图6。在不同的噪音类型和噪音比率下，训练期间的平均测试准确度(5次)。我们在手动损坏的CIFAR-10上训练RseNet-32，并在干净的测试集上进行测试。

6.结论

在这项工作中，我们开始对噪音标签进行正式研究。 我们首先针对通过噪声标签训练的DNN的泛化提出了一些发现。 理论分析和广泛的实验被提出来证明我们的陈述是正确的。 根据我们的发现，我们然后提出了INCV方法，该方法将噪声数据集随机划分，然后利用交叉验证来识别干净的样本。 我们为INCV提供了理论上的保证，然后通过实验证明它能够准确识别大多数清洁样品。 最后，我们采用共同教学策略，该策略充分利用已识别的样本来针对噪声标签对DNN进行稳健的训练。 通过与广泛的基准进行比较，我们证明了我们的方法在干净的测试装置上达到了最新的测试精度。 将来，我们关于使用噪声标签训练的DNN的泛化性能的表述可能会促进处理标签损坏的更基本方法。