**Meta-Weight-Net: Learning an Explicit Mapping For Sample Weighting**

会议：NIPS2019

代码：<https://github.com/xjtushujun/meta-weight-net>

**摘要**

当前的深度神经网络（DNN）会轻易地过度拟合带有损坏的标签或类不平衡的偏向训练数据。 通常，通过设计从训练损失到样本权重的权重函数映射，然后在权重重新计算和分类器更新之间进行迭代，样本重加权策略可缓解此问题。 但是，当前方法需要手动预先指定加权函数及其附加的超参数。 由于依赖于研究的问题和训练数据的适当加权方案的显着变化，这使得它们很难在实践中普遍应用。 为了解决这个问题，我们提出了一种能够直接从数据中自适应学习显式加权函数的方法。 加权函数是具有一个隐藏层的MLP，构成了几乎任何连续函数的通用逼近器，这使得该方法能够适应广泛的加权函数，包括传统研究中假设的那些。 在少量的无偏元数据的指导下，加权函数的参数可以与分类器的学习过程同时精确地更新。 综合和真实的实验证实了我们的方法在类不平衡和嘈杂标签情况下实现适当加权功能的能力，完全符合传统方法的常规设置，并且比常规情况更复杂。 这自然会导致其准确性比其他最新技术更好。

**1简介**

由于DNN具有对复杂输入模式进行建模的强大功能，最近它们在各种应用程序上均取得了令人印象深刻的良好性能。 但是，DNN可以轻松地过度适应有偏差的训练数据，例如包含损坏的标签[2]或类别不平衡[3]的数据，在这种情况下，它们的综合表现不佳。 这个健壮的深度学习问题已在多个文献中得到了理论证明[4、5、6、7、8、9]。

然而，实际上，通常会遇到这种有偏差的训练数据。 例如，实际收集的训练样本始终包含损坏的标签[10、11、12、13、14、15、16、17]。 一个典型的示例是从众包系统[18]或搜索引擎[19、20]大致收集的数据集，这可能会产生大量的嘈杂标签。 偏差训练数据的另一种流行类型是类不平衡的那些。 现实世界的数据集通常被描述为具有长尾配置的偏态分布。 少数类别占了大多数数据，而大多数类别的代表性不足。 因此，利用这些有偏差的训练数据进行有效的学习是很重要的，同时又是挑战机器学习中的重要问题[1，21]，这些偏差的训练数据被认为偏离了评估/测试数据。

样本重加权方法是解决此强大学习问题的常用策略。 主要方法是设计一个从训练损失到样本权重（具有超参数）的加权函数映射，然后在从当前训练损失值计算权重和最小化加权训练损失以进行分类器更新之间进行迭代。 构建这样的损耗映射有两个完全矛盾的想法。 如图1（a）所示，函数使函数单调增加，即，强制学习以更加强调具有较大损失值的样本，因为它们更像是位于分类边界上的不确定硬样本。 此类的典型方法包括AdaBoost [22，23]，硬负挖矿[24]和焦点损失[25]。已知这种样本加权方式对于班级不平衡问题是必需的，因为它可以优先考虑具有相对较高训练损失的少数班级。

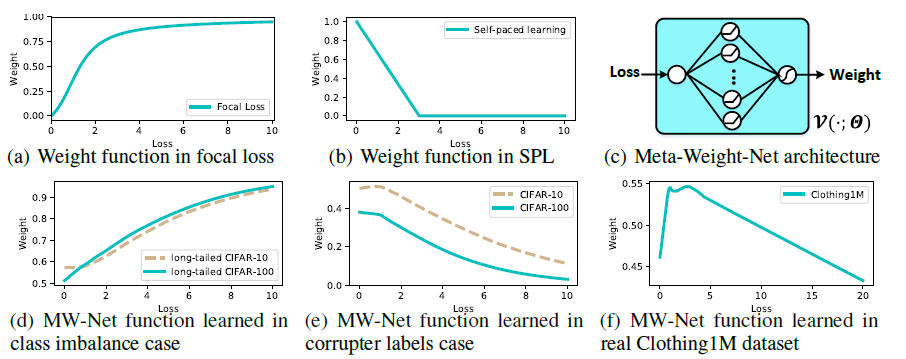


图1：（a）-（b）在focal loss和自定进度学习（SPL）中设置的权重函数。 （c）Meta-Weighting Net体系结构。 （d）-（f）通过我们的方法分别在类不平衡（不平衡因子100），嘈杂的标签（40％均匀噪声）和实际数据集中学习到了元加权网络功能。

相反，另一种方法将加权函数设置为单调递减，如图1（b）所示，以将具有较小损失值的样本作为更重要的样本。 合理性在于，这些样本更有可能是带有清晰标签的高可信度样本。 典型的方法包括自定进度学习（SPL）[26]，迭代重加权[27、17]和多个变体[28、29、30]。 此加权策略已特别用于嘈杂的标签情况，因为它倾向于抑制损耗值非常大（可能带有错误的不正确标签）的样本的影响。

尽管这些样本重加权方法有助于提高学习算法对有偏差训练样本的鲁棒性，但它们在实践中仍然存在明显的缺陷。 一方面，当前的方法需要基于对训练数据的某些假设来手动设置加权函数的特定形式。 但是，当我们几乎不了解基础数据知识或标签条件过于复杂时（例如训练集既不平衡又嘈杂的情况），这往往是不可行的。 另一方面，即使我们指定了某些加权方案（例如焦点损失[25]或SPL [26]），它们也不可避免地会涉及到超参数，例如前者中的聚焦参数和后者中的年龄参数，这些参数可以手动预设或通过交叉验证进行调整。 这往往会进一步增加它们的应用难度，并降低它们在实际问题中的性能稳定性。

为了缓解上述问题，本文提出了一种自适应样本加权策略，可以从数据中自动学习显式加权函数。 主要思想是参数化加权函数为只有一个隐藏层的MLP（多层感知器）网络（如图1（c）所示），称为Meta-Weight-Net，理论上是几乎所有连续函数的通用逼近器[31]，并且 然后使用一个小的无偏验证集（元数据）来指导其所有参数的训练。 加权函数的显式形式最终可以最终获得，特别适合于学习任务。

总而言之，本文做出了以下三方面的贡献：

1）我们建议通过元学习从数据中自动学习由MLP参数化的显式损失加权函数。 由于该权重网具有通用的逼近能力，因此它可以很好地适应各种加权函数，包括传统研究中使用的那些。

2）实验证明，我们的方法学习的加权函数高度符合传统的手动预设加权方式，适用于不同的训练数据偏差，例如图1（d）和1（e）所示的类不平衡和带有噪声的标签情况）。 这表明，所提方法学习的加权方案倾向于帮助揭示对数据偏差见解的更深入理解，尤其是在复杂的偏差情况下，其中提取的加权函数具有复杂的趋势（如图1（f）所示）。

3）可以很好地解释为什么所提出的方法有效的见解。 特别地，可以通过以下方式来解释用于Meta-Weight-Net参数的更新方程：改进那些更好地符合元数据知识的样本的样本权重，而抑制那些违反这种元知识的样本。 这符合我们对这个问题的常识：我们应该减少那些高度偏差的影响，同时强调那些没有偏差的。

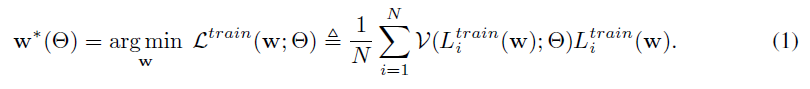
本文的结构如下。 第2节介绍了提出的元学习方法，以及详细的算法及其收敛性分析。 第三部分讨论相关工作。第四部分演示了实验结果，并最终得出结论。

**2拟议的元权重网络学习方法**

**2.1元学习目标**

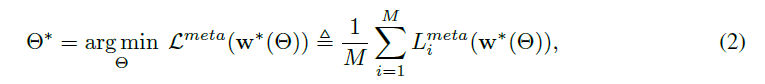
考虑训练集的分类问题，其中xi表示第i个样本，是c类的标签向量，N是整个训练数据的数量。表示分类器，w表示其参数。 在当前应用中，始终设置为DNN。 因此，为了方便起见，我们也采用DNN，并将其称为分类器网络。

通常，可以通过最小化在训练集合上计算的损失来提取最佳分类器参数。为了表示方便，我们表示。在存在偏差训练数据的情况下，样本重加权方法通过对第i个样本损失施加权重来增强训练的鲁棒性，其中表示权重 ，表示其中包含的参数。通过最小化以下加权损失来计算最佳参数**w**：



**Meta-Weight-Net**：我们的方法旨在以元学习的方式自动学习超参数。 为此，我们将表示为一个MLP网络，该网络只有一个包含100个节点的隐藏层，如图1（c）所示。为了方便参考，我们称此权重网为Meta-Weight-Net或MW-Net。 每个隐藏节点具有ReLU激活功能，输出具有Sigmoid激活功能，以确保输出位于[0，1]。 尽管简单，但该网络被称为几乎所有连续函数的通用逼近器[31]，因此可以拟合包括传统研究中使用的那些加权函数。

**元学习过程**。可以通过使用元学习思想[32、33、34、35]来优化MW-Net中包含的参数。具体来说，假设我们有少量的无偏元数据集（即具有清晰的标签和平衡的数据分布），表示真实样本标签分布的元知识，其中M是元样本的数量，。可以通过最小化以下元损失来获得最佳参数：

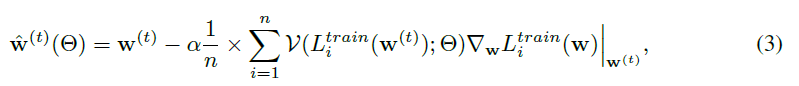


其中在元数据上计算。

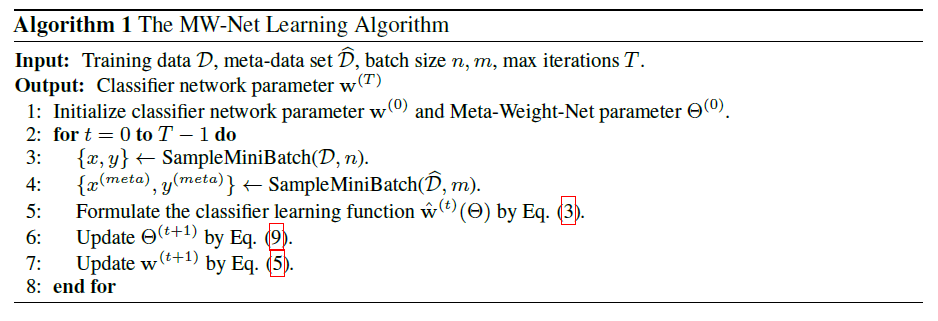
**2.2 Meta-Weight-Net学习方法**

计算最优需要两个嵌套的优化循环。 在这里，我们采用在线策略分别通过单个优化循环进行更新，以保证算法的效率。

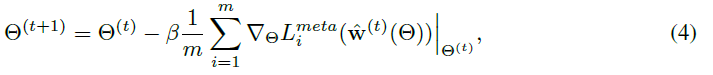
**制定分类器网络的学习方式**。 作为一般的网络训练技巧，我们采用SGD来优化训练损失（1）。具体而言，在每次训练迭代中，训练样本的小批量被采样，其中n是最小批大小。然后，可以通过沿着等式（1）在小批量训练的数据中目标损失的下降方向移动当前来制定分类器网络参数的更新方程：



为步长。

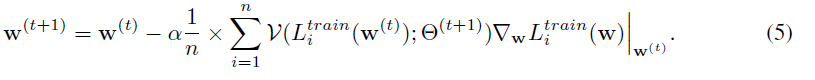


**更新Meta-Weight-Net的参数**：从等式（3）收到分类器网络参数更新公式的反馈后，然后，可以根据公式（2）轻松地更新Meta-Weight-Net的参数。即沿着等式（2）的目标梯度移动当前参数在元数据上计算：

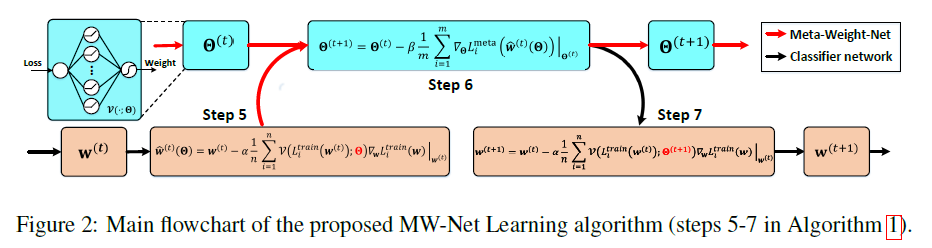


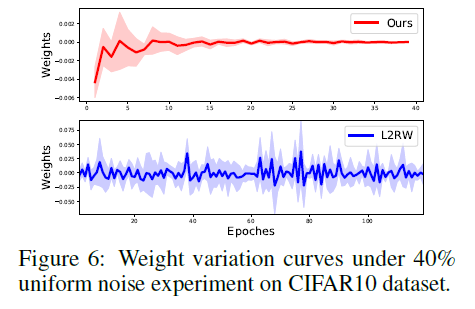
其中为步长。

更新分类器网络的参数：然后，使用更新的改善分类器网络的参数w，即



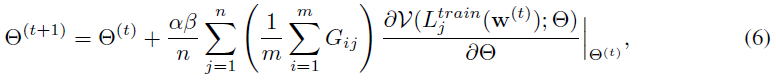
然后可以在算法1中总结MW-Net学习算法，图2说明其主要实现过程（步骤5-7）。 梯度的所有计算都可以通过自动微分技术有效地实现，并且可以推广到分类器网络的任何深度学习架构。 使用流行的深度学习框架（如PyTorch [36]）可以轻松实现该算法。 不难看出，分类器网络和MW-Net在学习过程中会根据最后一步中计算出的值逐渐改善其参数，从而可以稳定地更新权重，如图6所示。

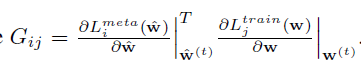




2.3元权重网加权方案分析

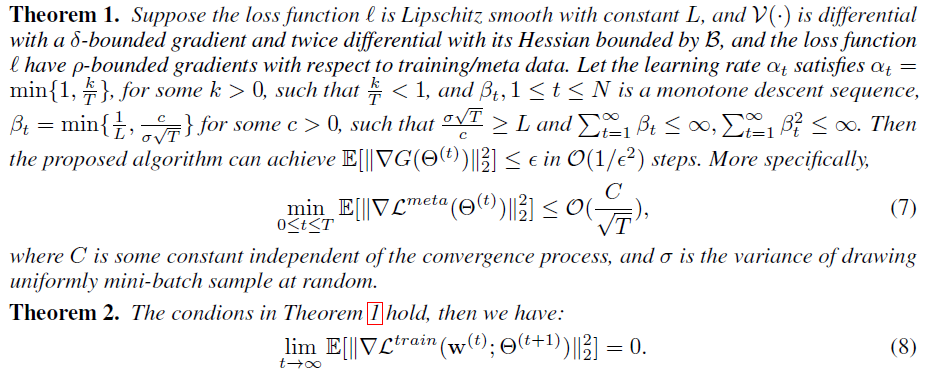
等式（9）通过反向传播可以重写为：



其中，忽略系数，很容易看出总和中的每个项都指向权重函数的上升梯度。，第j个梯度项的系数，表示基于训练损失计算的第j个训练样本的梯度与基于元损失计算的小批量元数据的平均梯度之间的相似性 。 这意味着，如果训练样本的学习梯度与元样本的学习梯度相似，那么它将被视为有利于获得正确的结果，并且其权重往往会增加。 相反，样品的权重倾向于被抑制。 这种理解与为什么著名的MAML起作用的原因是一致的[37，38，39]。

**2.4 MW-Net学习算法的收敛性**

我们的算法涉及两级目标的优化，因此我们从理论上证明了在定理1和定理2中的某些温和条件下，我们的方法收敛到了元损失和训练损失函数的临界点。 分别。 证明列在补充材料中。



**3相关工作**

**样本加权方法**。 重新加权示例的想法可以追溯到数据集重新采样[40、41]或实例重新加权[42]，后者通过使用对任务或数据的某些先验知识来将样本权重作为预处理步骤进行预先评估。 使样品权重更灵活的合适数据，最近的研究者专注于预先设计从训练损失到样本权重的加权函数映射，并在训练过程中动态地改善权重[43，44]。 设计加权函数主要有两种方式。 一种是使其单调增加，在类不平衡的情况下特别有效。 典型的方法包括增强算法（如AdaBoost [22]）及其多种变化形式[45]，困难的示例挖掘[24]和焦点损失[25]，这会给具有较大损失值的方法施加更大的权重。 相反，另一系列方法将加权函数指定为单调递减，特别是在嘈杂的标签情况下使用。 例如，SPL [26]及其扩展[28，29]，迭代重加权[27，17]和其他近期工作[46，30]，将更多的精力放在具有较小损失的简单样本上。 这些方法的局限性在于，它们都需要手动预先指定加权函数的形式以及它们的超参数，这增加了它们在实际应用中易于使用的难度。

**元学习方法**。 受元学习发展的启发[47、48、49、37、50]，最近提出了一些方法来从数据中学习自适应加权方案，以使学习更加自动化和可靠。 这方面的典型方法包括FWL [51]，学习教学[52、32]和MentorNet [21]方法，它们的权重函数分别设计为贝叶斯函数逼近器，具有注意力机制的DNN，双向LSTM网络。 然而，它们不仅使用损耗值作为经典方法的输入，而且使用的加权函数（即元学习器）具有更为复杂的形式，并且需要输入复杂的信息（例如样本特征）。 这使得它们不仅难以继承传统方法所具有的良好性能，而且还容易被普通用户复制。

与我们密切相关的一种称为L2RW [1]的方法采用了类似的元学习机制。 主要区别在于，权重是在此处隐式学习的，而没有显式的加权函数。 但是，这可能会导致训练期间的加权行为不稳定以及无法进行概括。 相比之下，通过显式但简单的Meta-Weight-Net，我们的方法可以以更稳定的方式学习权重，如图6所示，并且可以很容易地从特定任务推广到相关的其他任务（请参见参考资料）。 补充材料）。

**类不平衡的其他方法**。 处理数据不平衡的其他方法包括：[53，54]尝试将从主要课程中学到的知识转移到次要课程。 还开发了基于度量学习的方法来有效利用尾部数据来提高泛化能力，例如三头损失[55]和范围损失[56]。

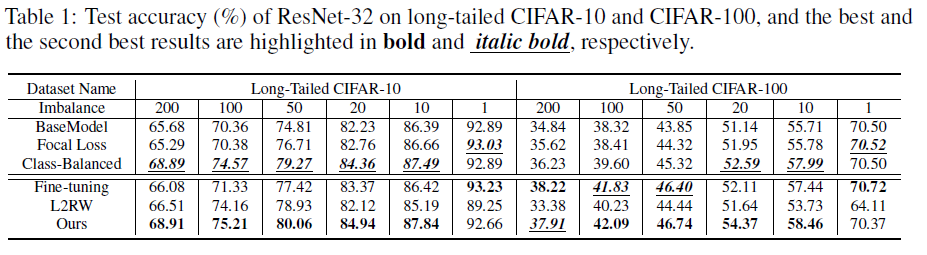
**标签损坏的其他方法**。 为了处理嘈杂的标签问题，已经设计了多种方法，通过补充清洁标签推理步骤[11、14、57、13、21、1、15]将嘈杂的标签校正为真实标签。 例如，GLC [15]提出了一种损耗校正方法，以减轻标签噪声对DNN分类器的影响。 其他方法包括Reed [58]，协同训练[16]，D2L [59]和S模型[12]。

**4实验结果**

为了评估该算法的性能，我们对具有类不平衡和嘈杂标签问题的数据集以及具有更复杂数据偏差的真实数据集进行了实验。

**4.1类不平衡实验**

我们使用长尾CIFAR数据集[60]，它根据指数函数减少每个类的训练样本数量，其中i是类索引，ni是训练图像的原始数量 和。数据集的不平衡因子定义为最大类别中的训练样本数除以最小样本数。 我们对ResNet-32 [61]进行了SGD的softmax交叉熵损失训练，其动量为0.9，权重衰减为，初始学习率为0.1。ResNet-32的学习率在80和90个时代之后（总共100个时代）除以10，WN-Net的学习率固定为。我们在验证集中每类随机选择10张图像作为元数据集。 比较的方法包括：1）BaseModel，它使用softmax交叉熵损失在训练集上训练ResNet-32； 2）焦点损失[25]和类平衡[60]代表了预定义样本重加权技术的最新水平； 3）微调，微调BaseModel对元数据集的结果； 4）L2RW [1]，它利用附加的元数据集来自适应地分配训练样本的权重。



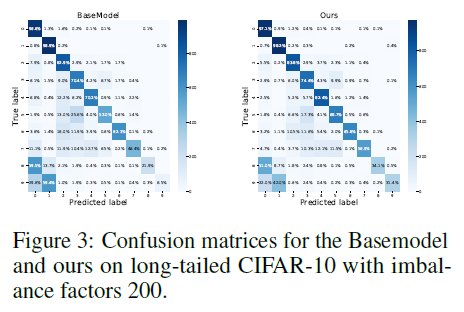
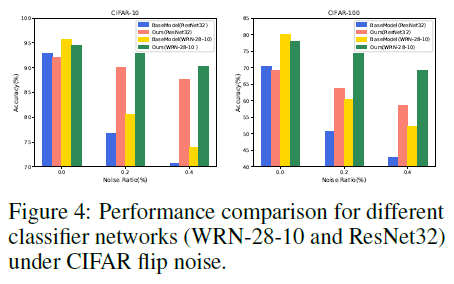


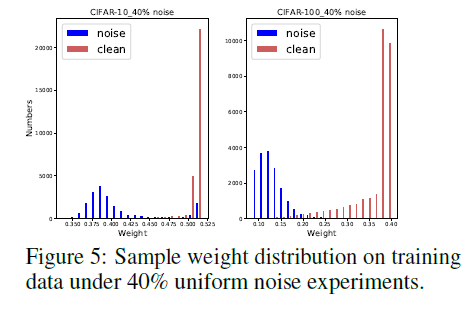
表1显示了ResNet-32在测试集上的分类准确性，混淆矩阵如图3所示（更多详细信息在补充材料中列出）。 可以观察到：1）我们的算法在具有类不平衡的数据集上明显优于其他竞争方法，显示了在这种数据偏差情况下的鲁棒性； 2）当不平衡因子为1时，即所有类别的样本数均相同，微调效果最佳，并且我们的方法仍能达到可比的性能； 3）当长尾CIFAR-100上的不平衡因子为200时，最小的类别只有两个样本。 进行额外的微调可提高性能，而我们的方法在这种极端的数据偏差下仍能保持良好的性能。

为了了解MW-Net的权重方案，我们在图1（d）中描述了所学MW-Net的权重相对于损失的趋势曲线，该曲线与经典的最佳加权方式相符。 即，应该对损失较大的样本施加较大的权重，这些样本更有可能是少数类别的样本。

**4.2标签损坏的实验**

我们在训练集上研究了两个损坏的标签设置：1）均匀噪声。 遵循[2]中的相同设置，每个样本的标签以概率p独立更改为随机类别。 2）翻转噪音。 每个样本的标签以总概率p独立地翻转到相似类别。 在我们的实验中，我们以相等的概率随机选择两个类别作为相似类别。 使用了两个基准数据集：CIFAR-10和CIFAR-100 [62]。 两者都普遍用于评估嘈杂的标签[59，16]。随机选择验证集中具有干净标签的1000张图像作为元数据集。 我们采用了广泛的ResNet-28-10（WRN-28-10）[63]来处理均匀噪声，将ResNet-32 [61]来处理翻转噪声作为我们的分类器网络模型。我们尝试了在每个噪声设置下将不同的分类器网络体系结构用作分类器网络，以表明我们的算法适用于不同的深度学习体系结构。 我们在图4中显示了这种效果，验证了我们的方法在两个分类器网络设置中始终如一的良好性能。

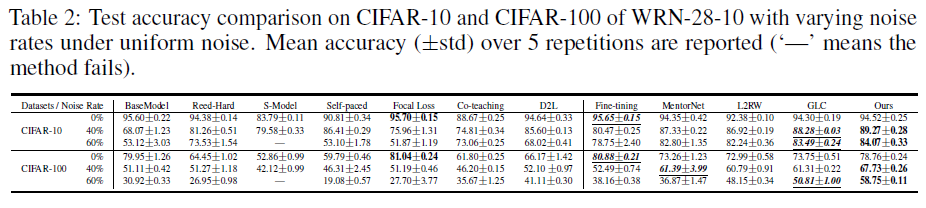




比较方法包括：BaseModel，指的是在我们的方法中使用的相似分类器网络，同时直接在有偏差的训练数据上进行训练； 强大的学习方法Reed [58]，S-Model [12]，SPL [26]，焦点损失[25]，协同教学[16]，D2L [59]； 微调，使用干净的标签微调BaseModel在元数据上的结果，以进一步提高其性能； 典型的元学习方法MentorNet [21]，L2RW [1]，GLC [15]。 我们还仅在1000个元图像上训练了基准网络。 由于忽略了大量训练样本背后的知识，因此性能明显比所提出的方法差。 因此，我们没有将其结果进行比较。

所有基线网络均使用SGD进行了训练，其动量为0.9，权重衰减为，初始学习率为0.1。 分类器网络的学习率在均匀噪声下的36个时期和38个时期（总共40个时期）除以10，而在40个时代和50个时期之后（平均总共60个epoch）在反转的噪音中。WN-Net的学习率固定为。 我们使用不同的随机种子重复了5次实验，以进行网络初始化和生成标签噪声。

我们在表2和表3中报告了每个实验系列和每个竞争方法在5次重复测试中得到的平均准确度。可以观察到，除了40％的Flip，第二种方法在几乎所有数据集和所有噪声率上均具有最佳性能 噪声。 在噪声情况为0％（无偏差）的情况下，我们的方法仅比BaseModel稍差一点。 对于其他损坏的标签案件，我们方法的优势显而易见。 此外，可以看出，在均匀噪声下，当噪声率从40％增加到60％时，我们与所有其他竞争方法之间的性能差距会增大。 与CIFAR100数据集的第二好的结果相比，即使标签噪声为60％，我们的方法仍可以获得相对较高的分类精度，并且获得了超过15％的准确度增益，这表明我们的方法在这种情况下的鲁棒性。



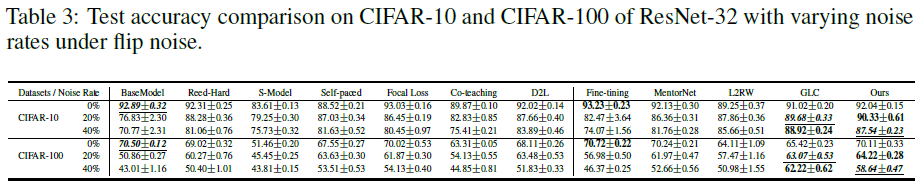


图4显示了WRN-28-10与ResNet32在固定翻转噪声设置下的性能比较。 我们可以观察到，两个网络之间的方法和BaseModel的性能增益几乎是相同的。 这意味着我们方法的性能改进不取决于分类器网络体系结构的选择。

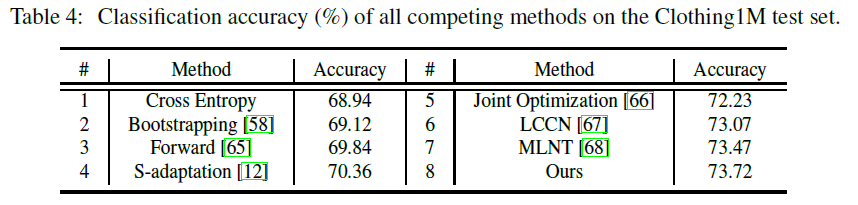
如图1（e）所示，学习权重函数的形状描述为单调递减，符合针对此偏差条件的传统最佳设置，即，对具有相对较大损失的样本施加较小的权重以抑制损坏的影响 标签。 此外，我们在图5中绘制了干净和有噪声训练样本的权重分布。可以看出，几乎所有大的权重都属于干净样本，并且有噪声样本的权重小于干净样本的权重，这意味着训练有素的样本 Meta-Weight-Net可以区分干净和嘈杂的图像。

图6在我们的方法和L2RW的CIFAR10数据集上绘制了在40％噪声下的权重变化以及训练训练。 y轴表示在相邻的历元之间计算的权重之差，x轴表示历元的数量。 随机选择十个有噪声的样本以计算它们的平均曲线，并在相应区域中用包围该样本的标准偏差的区域进行说明。 可以看出，我们的方法的权重不断变化，随着迭代逐渐稳定，最后收敛。 相比之下，L2RW学习过程中的权重波动相对较大。 这可以解释我们的方法与竞争方法相比始终具有更好的性能。

**4.3在Clothing1M上进行的实验**

为了验证所提出方法在现实世界数据上的有效性，我们在Clothing1M数据集[64]上进行了实验，其中包含从网上购物获得的100万幅服装图像具有14个类别的网站，例如T恤，衬衫，针织品。 标签是通过使用卖方提供的图像的周围文本生成的，因此包含许多错误。 我们使用7k干净数据作为元数据集。 遵循先前的工作[65，66]，我们使用了在ImageNet上经过预训练的ResNet-50。 对于预处理，我们将图像调整为256X256，裁剪中间的224X224作为输入，然后执行归一化。 我们使用动量为0.9，权重衰减为10 weight3，初始学习率为0.01，批处理大小为32的SGD。ResNet-50的学习率在5个历元（总共10个历元）后除以10。 WN-Net的学习率固定为。

结果总结在表4中。这表明所提出的方法达到了最佳性能。 图1（f）绘制了学习到的MW-Net函数的趋势曲线，揭示了丰富的数据见解。 具体来说，当损失的值较小时，加权函数会随着损失的增加而增加，这意味着它倾向于更加强调具有丰富分类知识的硬边际样本。 而当损耗逐渐变化较大时，加权函数开始单调减小，这意味着它倾向于抑制损耗值相对较大的噪声标签样本。 这种复杂的本质不能通过常规的权重函数很好地传递。



**5结论**

我们提出了一种新的元学习方法，用于自适应地提取样本权重，以在存在训练数据偏差的情况下保证强大的深度学习。 与当前需要手动设置权重函数形式的重新加权方法相比，该新方法能够直接从数据中得出合理的加权方法。 我们的算法的工作原理可以很好地解释，并且我们的方法的过程可以轻松地重现（附录A提供了我们算法的Pytorch实现（少于30行代码）），并且完整的培训代码可在https//github.com/xjtushujun/meta-weight-ne上获得）。 我们的经验结果表明，在一般的数据偏差情况下，例如类不平衡，标签损坏和更复杂的实际情况下，建议的方法都可以执行得更好。 此外，这种自适应权重学习方法有望应用于机器学习中的其他权重设置问题，例如集成方法和多视图学习。