EvidentialMix: 混合开闭集标签噪声的学习

Ragav Sachdeva,Filipe R. Cordeiro,Vasileios Belagiannis,
Ian Reid and Gustavo Carneiro

摘要

深度学习的有效性依赖于大规模的数据集，这些数据集是通过可靠而精心策划的数据获取和标注过程得到的。然而，获取具有精确标注的大规模数据集是非常昂贵和耗时的，而廉价的替代品通常会产生带有噪声标签的数据集。本文研究的领域通过关注下面两种标签噪声的训练模型来解决这个问题: 1) 闭集噪声，即一些训练样本被错误地标注为非真实类别的标签，但该标签在类别集合中; 2) 开集噪声，即训练集包含的样本具有 (严格地) 不包含在已知训练标签集中的真实类别。在本工作中，我们研究了一个标签噪声问题的变体，它结合了开集和闭集的标签噪声，并引入了一个基准评价来评价在这个设定下训练算法的性能。我们认为这样的问题更普遍，更好地体现了现实中噪声标签的场景。此外，我们提出了一种新的算法，称为 EvidentialMix，它解决了这个问题。我们将其性能与最先进的方法进行比较，设定为我们提出的在闭集和开集噪声上的基准。结果表明我们的方法比最先进的模型产生了更好的分类结果和更好的特征表示。我们的代码可以在如下的网址获取: https://github.com/ragavsachdeva/EvidentialMix。

1 介绍

深度学习利用大量且精心收集的训练数据，在几个重要的分类问题上取得了显著的结果 [12,3]。然而，社会中大多数感兴趣的数据集可用的数量级大而不容易收集，这意味着数据可能包含获取和标记的错误，可以导致较差的泛化能力 [26]。所以本邻域的一个重要的挑战就是能够处理这种噪声标签数据集的方法的研发。最近，研究人员极大地促进了这一领域的发展，他们研究了受控的合成标签噪声，发现了可以应用于现实世界噪声数据集的理论或方法。

目前所研究的标签噪声类型可分为两类: 闭集噪声和开集噪声。虽然这些术语 ("闭集"和"开集") 是最近由 wang 等人在 [24] 中创造的，他们在论文中引入了有开集噪声的标签学习问题，但闭集标签噪声问题已经被广泛地研究了很久。在处理闭集标签噪声时，大多数的学习算法假定有一组固定的训练标签 [14,22]。在这个设定中，一些训练样本被标注到一个不正确的标签上，而它们的真实类出现在训练标签集中。这些错误可以是完全随机的，即标签被任意地翻转到一个不正确的类别上; 也可以是一致的，如标注者确实对一个特定示例的标注感到困惑。一个较少研究的标签噪声问题是开集噪声标签问题 [24]，其中我们错误地采样了一些数据，导致它们的真实标注不包含在已知的训练标签集内。这种设定的一个夸张的例子可能是：在用于建模猫和狗的二分类问题中，训练集中出现一个马的图像。从他们的定义可以明显看出，这两种标签噪声是互斥的，即一个给定的噪声标签不可能同时是闭集和开集。

很容易证明开放集和闭集噪声可能同时出现在真实世界的数据集中。例如，最近的大规模数据收集方法提出使用查询商业搜索引擎获取数据 (例如，谷歌图像检索)。其中搜索关键字作为查询图像的标签。从图1中可以看出，使用这种方法采集图像会导致同时存在开集噪声和闭集噪声。然而到目前为止还没有关于混合标签噪声的系统研究，尽管已经有论文对他们提出的方法进行了评估 [13,24]，但训练数据集被闭集噪声或开集噪声独立地破坏了，但从未被联合地破坏过。

在本文中，我们提出了一种新的基准评价来解决标签噪声学习问题，该问题由闭集噪声和开集噪声的组合组成。这个提出的基准评估由三个变量定义: 1) 标签噪声在训练集中所占的总比例，用 $\rho \in [0,1]$ 表示; 2) 标签噪声样本集中闭集标签噪声的比率，使用 $\omega \in [0,1]$ 表示;（这表示了整个训练集中有 $\rho \times \omega$ 比例的样本含有闭集标签噪声，并且有 $\rho \times (1-\omega)$ 比例的样本含有开集标签噪声）3) 开集噪声数据的来源。注意这种设定同时是两种标签噪声的泛化。因为如果 $\omega \in \{0,1\}$，则样本可以被破坏而具有其中的一种噪声。

最先进的 (SOTA) 方法旨在解决闭集噪声标签问题，其重点是识别不正确注释的样本，并使用半监督学习 (SSL) 方法，更新它们的标签的 [14] 用于下一次训练迭代。这种策略很可能在开集问题中失败，因为它假设每个训练样本的真实标签都在研究类别中存在，但事实并非如此。另一方面，解决开集噪声问题的主要方法 [24] 重点在学习过程中，识别噪声样本并降低其权值。这种策略在闭集问题上效率很低，因为闭集的噪声样本在 SSL 阶段仍然很有意义。因此，为了在存在闭集和开集噪声样本的情况下具有鲁棒性，学习算法必须能够识别影响每个训练样本的标签噪声类型。根据类型，如果是闭集标签噪声样本，则更新标签，如果是开集标签噪声样本则减少其权重。为了实现这一目

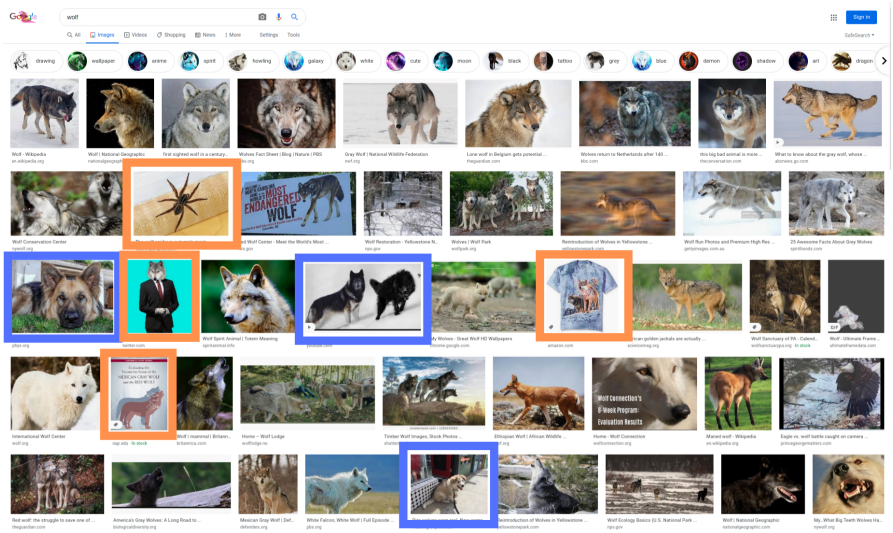

图 1: wolf-vs-dog 二元分类器数据的搜索引擎查询结果。这里使用的搜索关键字是“狼”。由一个橙色边框包围的图像是开集噪声样本 (即既不是狼也不是狗)，被蓝色边框包围的是闭集噪声样本 (即标签是狼，但实际上是狗)。

标，我们提出了一种新的学习算法，名为 EvidentialMix (EDM)，见图2。我们提出的算法的关键贡献如下:

- EDM 能够准确区分干净的、开集和闭集噪声样本，从而允许它根据类型应用不同的学习机制。相比之下，以前的方法 [14,24] 只能将干净样本与噪声样本分离，而不能将闭集噪声与开集噪声样本分离。
- 我们发现我们的方法可以学习到比以前的研究方法更好的特征表示，如图4 中的 t-SNE 图所示，其中我们的方法对每个已知的标签/类都有一个唯一的聚类，而对开集样本给出另一个单独的聚类。相比之下，之前的方法显示出在开集样本上很大程度的过拟合，并错误地将它们聚到一个已知的类中。
- 我们的实验表明，EDM 得到的分类精度，在不同的标签噪声率设定下（包括 $\omega \in \{0, 1\}$ 这样的极端情况）能够与之前的方法相媲美或能超过它们。

2 先前的研究

人们对标签噪声设定下深度学习分类建模问题的研究越来越感兴趣。对于闭集噪声，Reed 等人 [19] 提出了最早的方法之一，使用转移矩阵来表示标签如何在不同类别之间切换。转移矩阵在不同的方法中被进一步探索 [18,5]，但它们都没有显示出有竞争力的结果，可能是因为它们不包括识别和处理噪声样本的机制。数据增广方法 [27] 相继被闭集标签噪声方法探索过，其思想是数据增广可以自然增加对噪声标签的训练鲁棒性。元学习是另一种在闭集噪声标签问题中探索的技术 [15,20]，但由于需要干净的验证数据集或人工的新训练任务，这项技术相对来说尚未被探索。研究者也探索了利用课程学习 (CL)[7] 解决闭集噪声问题的方法，在训练过程中根据训练样本的损失值，动态地对训练

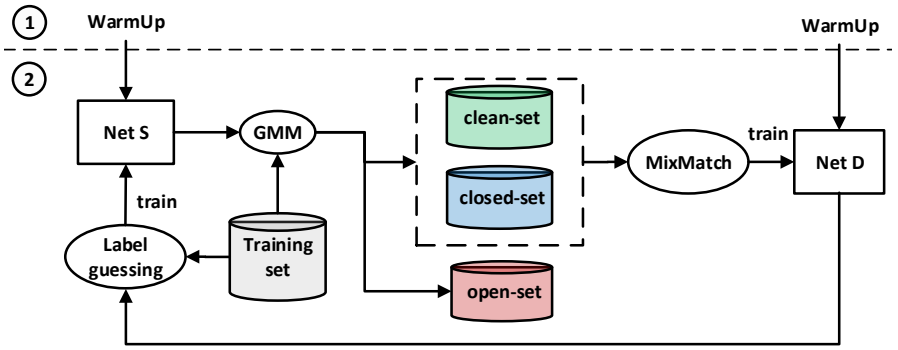

图 2: 我们提出的方法 EvidentialMix 依赖于两个网络，NetD 和 NetS。这两种模型最初都是使用一个简单的训练机制进行几个 epoch(见 (1)WarmUp) 的训练，该机制不包含任何处理标签噪声的方法。接下来在 (2) 中，我们在 NetS 的样本损失上拟合一个 ψ-分量高斯混合模型 (GMM)，以便将训练集分离为干净的、闭集和开集样本。在这之后，使用 SSL 学习机制 [14] 训练 NetD，但只使用预测的干净和闭集样本。最后，NetS 在整个训练集上进行训练，以使用 NetD 估计的标签最小化主观逻辑损失 [21]，并且 (2) 中的过程重复。

样本进行重新标记。该方法已经扩展成多模型训练方法 [17,25]，目的是聚焦到损失小的样本的训练上，这些样本被多个模型不一致地分类。最近，Kim 等人 [9] 研究了使用消极学习显式识别噪声样本，获得了有竞争力的结果。另一种处理标签噪声的重要方法是 Tarvainen 等人 [23] 提出的模型集成。利用具有鲁棒性的生成分类器 (RoG) 来提高判别分类器的性能的方法在 Lee et al.[13] 已经进行了研究，他们从训练好的判别模型的几个层中提取特征，构建鲁棒的线性判别集成模型。实际上，这种方法有可能提高任何方法的性能，并且已经接连在闭集和开集噪声场景中进行了测试。

开集标签噪声学习直到最近才被 wang 等人 [24] 研究，其思想是识别出含有噪声标签的样本，并在训练过程中减少它们的权重，因为它们几乎可以肯定属于一个不在训练标签集合中的类别。鉴于他们的方法是唯一明确处理开集噪声的方法，他们的方法是解决开集标签噪声学习的主要基准。

目前用于闭集标签噪声学习的 SOTA 方法是 SELF[22] 和 DivideMix[14]——它们都结合了上述的多种方法。SELF[22] 结合了模型集成、重标记、噪声样本识别和数据增强; 而 DivideMix[14] 使用了多模型训练、噪声样本识别和数据增强 [1]。这两种方法很容易受到开集噪声的影响，因为它们假设训练样本必须属于某个训练类别——这个假设对于开集噪声来说是不正确的。

3 方法

3.1 问题定义

我们定义训练集为 $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{|\mathcal{D}|}$，其中 RGB 图像为 $\mathbf{x} : \Omega \to \mathbb{R}^3$ (Ω 表示图像晶格)。训练标签集合用 $\mathcal{Y}$ 表示，形成了 $|\mathcal{Y}|$ 维度下的标准基，$\mathbf{y} \in \{0,1\}^{|\mathcal{Y}|}$ ($\sum_{c=1}^{|\mathcal{Y}|} \mathbf{y}(c) = 1$, 表示是一个多分类问题)。注意 $\mathbf{y}_i$ 表示的是 $\mathbf{x}_i$ 的噪声标签，潜在的真实标签用 $\mathbf{y}_i^*$ 表示。

对于**闭合标签噪声问题，噪声率**为 $\zeta \in [0,1]$，我们假定 $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}$ 以概率 $1-\zeta$ 被标记为 $\mathbf{y}_i = \mathbf{y}_i^*$，并以 ζ 概率被标记为 $\mathbf{y}_i \sim r(\mathcal{Y})$，$r(\mathcal{Y}, \theta_r)$ 表示一个随机函数，从 $\mathcal{Y}$ 选择一个标记，由一个特定分布以及其参数 θ_r 控制。

对于**开集标签噪声问题，噪声率**为 $\eta \in [0,1]$，我们需要定义一个新的训练集 $\mathcal{D}'$（满足 $\mathcal{D}' \cap \mathcal{D} = \emptyset$）。其中 $\mathcal{D}'$ 的标签集合用 $\mathcal{Y}'$ 表示（满足 $\mathcal{Y}' \cap \mathcal{Y} = \emptyset$）——这意味着在 $\mathcal{D}'$ 中的图像不再具有 $\mathcal{Y}$ 中的标签。在这样的开集标签噪声学习问题中，有 $1-\eta$ 比例的样本从 $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}$ 且 $\mathbf{y}_i = \mathbf{y}_i^*$ 的样本中抽取。另外 η 比例的样本从 $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}'$ 且 $\mathbf{y}_i \sim r(\mathcal{Y}, \theta_r)$ 的样本中抽取。

对于**混合的开集和闭集标签噪声问题，噪声率**为 $\rho, \omega \in [0,1]$，混合了上述两种噪声定义。更具体地，$1-\rho$ 比例的训练集包含图像 $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}$，标记为 $\mathbf{y}_i = \mathbf{y}_i^*$，然而 $\omega \times \rho$ 比例的样本采样自 $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}$，且标签为 $\mathbf{y}_i \sim r(\mathcal{Y}, \theta_r)$。$(1-\omega) \times \rho$ 比例的样本来自集合 $\mathcal{D}'$，且被标记为 $\mathbf{y}_i \sim r(\mathcal{Y}, \theta_r)$。

3.2 噪声分类

处理这个问题时的主要障碍是需要识别闭集和开集噪声样本，因为它们必须用不同的方法处理。一种可能的方法是将闭集噪声样本与置信度高而分类错误时 [14] 计算出的高损失相关联，将开集样本与不确定的分类相关联。为了实现这一点，我们使用了主观逻辑 (SL) 损失函数 [21]，依赖于证据推理理论和 SL 来量化分类不确定性。SL 损失利用 Dirichlet 分布来代表主观意见，对信念和不确定性进行编码。使用 SL 损失进行训练的网络试图将预测后验的参数构成 Dirichlet 密度函数进而对训练样本进行分类。对于给定样本的输出结果被认为是在一组类别标签上对该样本进行分类的证据。图3显示了使用 SL 损失训练的网络中样本的损失分布。很容易观察到干净、闭集和开集噪声样本之间的区别。

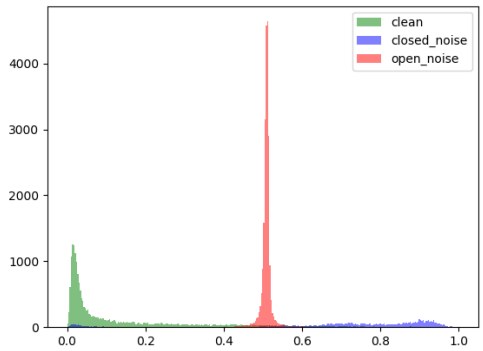

图 3: 预热阶段过后，使用 SL 损失函数时逐样本的损失分布。其中全部噪声比率 $\rho = 0.6$，闭合噪声比率 $\omega = 0.25$（也就是开集噪声比率为 $1-\omega = 0.75$）。$\mathcal{D}$ 为 CIFAR-10，$\mathcal{D}'$ 为 ImageNet32。

3.3 EvidentialMix

我们提出的 EvidentialMix 方法同时训练两个网络: 使用 SL 损失 [21] 的 NetS、使用 SSL 训练机制和 DivideMix(DM) 损失 [14] 的 NetD。广义地说，SL 损失估计分类不确定性的能力允许 NetS 将训练集划分为干净样本、开集和闭集噪声样本。预测的干净样本和闭集样本随后被用来配合 MixMatch（如 [14] 所述）训练 NetD，而预测的开集噪声样本在当前周期被丢弃。在此之后，NetD 重新标记整个训练数据集 (包括预测的开集样本)，然后用来训练 NetS。

随着 NetS 不断地从 NetD 预测的标签中学习，它可以更好地将数据分成三个集合。这是因为 NetD 预测的标签在训练过程中变得更加准确，因为它只在干净样本和闭集样本上进行训练，而从不训练被预测为开集的样本。这两个网络相互补充，从而对组合的闭集和开集噪声问题产生准确的分类结果。下面列出了详细的解释，而算法1详细地描述了整个训练过程。

算法1训练 NetD，通过 $f_{\theta^{(D)}}(c|\mathbf{x})$ 表示，并且训练 NetS，通过 $f_{\theta^{(S)}}(c|\mathbf{x})$ 表示（其中 $c \in \{1, ..., |\mathcal{Y}|\}$）。——两个网络都可以得到 $\mathbb{R}^{|\mathcal{Y}|}$ 空间的（logit）。在预热阶段（见 WarmUp($\mathcal{D}$) 行），我们在有限数量的周期内同时训练两个网络，其中对于 $f_{\theta^{(D)}}(c|\mathbf{x})$ 的训练使用交叉熵损失，此时概率通过 Softmax 激活函数 $\sigma(.)$ 获得，形成 $p_{\theta^{(D)}}(c|\mathbf{x}) = \sigma(f_{\theta^{(D)}}(c|\mathbf{x}))$，对于 $f_{\theta^{(S)}}(c|\mathbf{x})$ 使用如下的 SL 损失:

$$\mathcal{L}^{(S)} = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \ell^{(S)}(\mathbf{x}_i, \mathbf{y}_i, \theta^{(S)}), \tag{1}$$

根据 [21]:

$$\ell^{(S)}(\mathbf{x}_i, \mathbf{y}_i, \theta^{(S)}) = \sum_{c=1}^{|\mathcal{Y}|} (\mathbf{y}_i(c) - \alpha_{ic}/S_i)^2 + \frac{\alpha_{ic}(S_i - \alpha_{ic})}{S_i^2(S_i + 1)}, \tag{2}$$

其中，对于每个类别 $c \in \{1, ..., |\mathcal{Y}|\}$，有 $\alpha_{ic} = \varphi(f_{\theta^{(S)}}(c|\mathbf{x}_i)) + 1$ 。$\varphi(.)$ 表示 ReLU 激活函数，且 $S_i = \sum_{c=1}^{|\mathcal{Y}|} \alpha_{ic}$。

对样本分类成干净、开集、闭集三个类别是通过公式(2)对于整个训练集 $\mathcal{D}$ 的 SL 损失实现的。更具体地，我们使用了损失集合 $\{\ell^{(S)}(\mathbf{x}_i, \mathbf{y}_i, \theta^{(S)})\}_{i=1}^{|\mathcal{D}|}$ 并使用期望极大化算法（EM）拟合了一个 ψ-分量的高斯混合模型（GMM）。我们在本文中探索的想法在于，模型对**干净样本**的输出趋于自信，同时与原始标签一致，产生一个小的损失。模型对 **闭集噪声样本**的输出趋于自信但是同时和原始标签不一致，产生一个大的损失值。然而，对于**开集噪声样本**，模型将不会产生可信输出，导致损失值既不大也不小。因此，多分量 GMM 模型将会捕获每个这些集合，得到样本分类为干净的概率 w_i，即后验概率 $p(\mathcal{G}|\ell^{(S)}(\mathbf{x}_i, \mathbf{y}_i, \theta^{(S)}))$。其中 $\mathcal{G}$ 表示均值 $\leq \mu_{min}$ 的高斯分量的集合（也就是说小损失的分量集合）。分类为闭集噪声样本的后验概率为 $w_i^{\text{cl}} = p(\mathcal{G}^{\text{cl}}|\ell^{(S)}(\mathbf{x}_i, \mathbf{y}_i, \theta^{(S)}))$ ，通过均值 $\geq \mu_{max}$ 的高斯分量集合进行计算。

Algorithm 1: EvidentialMix(EDM)

Input: $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{|\mathcal{D}|}$, number of augmentations M, temperature sharpening T, loss weights $\lambda^{(\mathcal{U})}$ and $\lambda^{(reg)}$, MixMatch parameter α, number of epochs E.

Output: output result

$f_{\theta^{(D)}}(c|\mathbf{x}), f_{\theta^{(S)}}(c|\mathbf{x}) = \text{WarmUp}(\mathcal{D})$;

while $e < E$ **do**

  $\mathcal{W}, \mathcal{W}^{\text{op}}, \mathcal{W}^{\text{cl}} = \text{GMM}(\mathcal{D}, f_{\theta^{(S)}}(c|\mathbf{x}))$

  `// Train NetD`

  $\mathcal{X} = \{(\mathbf{x}_i, \mathbf{y}_i, w_i) | (\mathbf{x}_i, \mathbf{y}_i, w_i) \in (\mathcal{D}, \mathcal{W}), w_i > \max(w_i^{\text{op}}, w_i^{\text{cl}})\}$

  $\mathcal{U} = \{\mathbf{x}_i | (\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}, w_i^{\text{cl}} > \max(w_i, w_i^{\text{op}})\}$

  for *iter=1* ***to*** *num_iters* **do**

    $\{(\mathbf{x}_b, \mathbf{y}_b, w_b)\}_{b=1}^{B} \subset \mathcal{X}$ `// randomly pick` B `samples from` $\mathcal{X}$

    $\{\mathbf{u}_b\}_{b=1}^{B} \subset \mathcal{U}$ `// randomly pick` B `samples from` $\mathcal{U}$

    for *b=1* ***to*** B **do**

      for *m=1* ***to*** M **do**

        $\hat{\mathbf{x}}_{b,m} = \text{DataAugment}(\mathbf{x}_b)$

        $\hat{\mathbf{u}}_{b,m} = \text{DataAugment}(\mathbf{u}_b)$

      end

      for *c=1* ***to*** $|\mathcal{Y}|$ **do**

        $\mathbf{p}_b(c) = \frac{1}{M} \sum_m p_{\theta^{(D)}}(c|\hat{\mathbf{x}}_{b,m})$

        $\mathbf{q}_b(c) = \frac{1}{M} \sum_m p_{\theta^{(D)}}(c|\hat{\mathbf{u}}_{b,m})$

      end

      $\hat{\mathbf{y}}_b = \text{TempSharpen}_T(w_b \mathbf{y}_b + (1 - w_b)\mathbf{p}_b)$

      $\hat{\mathbf{q}}_b = \text{TempSharpen}_T(\mathbf{q}_b)$

    end

    $\hat{\mathcal{X}} = \{(\hat{\mathbf{x}}_{b,m}, \hat{\mathbf{y}}_b)\}_{b \in (1,\ldots,B), m \in (1,\ldots,M)}$

    $\hat{\mathcal{U}} = \{(\hat{\mathbf{u}}_{b,m}, \hat{\mathbf{q}}_b)\}_{b \in (1,\ldots,B), m \in (1,\ldots,M)}$

    $\mathcal{X}', \mathcal{U}' = \text{MixMatch}_\alpha(\hat{\mathcal{X}}, \hat{\mathcal{U}})$

    $\theta^{(D)} = \text{SGD}(\mathcal{L}^{(D)}, \theta^{(D)}, \mathcal{X}', \mathcal{U}')$

  end

  `// Train NetS`

  for *i=1* ***to*** $|\mathcal{D}|$ **do**

    $\hat{c}_i = \arg\max_{c \in \mathcal{Y}} \left[(w_i^{\text{cl}}) p_{\theta^{(D)}}(c|\mathbf{x}_i) + (1 - w_i^{\text{cl}}) \mathbf{y}_i(c) \right]$

    $\hat{\mathbf{y}}_i = \text{onehot}(\hat{c}_i)$

  end

  $\theta^{(S)} = SGD(\mathcal{L}^{(S)}, \theta^{(S)}, \{(\mathbf{x}_i, \hat{\mathbf{y}}_i)\}_{i=1}^{|\mathcal{D}|})$

end

分类为开集噪声样本的概率则使用剩余的高斯分量进行计算，分量的均值 $\in (\mu_{min}, \mu_{max})$ ——这些后验概率形成了三个集合 $\mathcal{W} = \{w_i\}_{i=1}^{|\mathcal{D}|}$，$\mathcal{W}^{\text{cl}} = \{w_i^{\text{cl}}\}_{i=1}^{|\mathcal{D}|}$ 和 $\mathcal{W}^{\text{op}} = \{w_i^{\text{op}}\}_{i=1}^{|\mathcal{D}|}$。使用这些后验概率, 我们可以建立干净样本集，表示为 $\mathcal{X}$，包含被预测为干净样本的概率 w_i 大于其他两种类型的样本。同时建立闭合噪声样本集，表示为 $\mathcal{U}$，包含被预测为闭集标签噪声样本的概率 w_i^{cl} 大于其他两种类型的样本。

接下来，我们通过干净样本集 $\mathcal{X}$ 和闭合噪声样本集 $\mathcal{U}$ 训练 NetD，样本集在上文中已定义。一个小批次从 $\mathcal{X}$ 和 $\mathcal{U}$ 中抽取。此外，我们增广了每个集合中的样本 M 次 [14]。对于干净样本和闭集噪声样本的平均预测概率通过计算 M 次增广的平均值得到，计算是在“温度尖锐化”之后进行的，通过 $\text{TempSharpen}_T(.)$ 表示，其中 T 表示温度。从而对干净和闭集噪声样本形成“新”的样本和标签，分别为 $\hat{\mathcal{X}} = \{(\hat{\mathbf{x}}_{b,m}, \hat{\mathbf{y}}_b)\}_{b,m=1}^{B,M}$ 和 $\hat{\mathcal{U}} = \{(\hat{\mathbf{u}}_{b,m}, \hat{\mathbf{q}}_b)\}_{b,m=1}^{B,M}$ 在随机梯度下降（SGD）的上一个阶段是 MixMatch 过程 [1]，其中 $\hat{\mathcal{X}}$ 和 $\hat{\mathcal{U}}$ 的样本是 $\mathcal{X}'$ 和 $\mathcal{U}'$ 中样本的线性组合。SGD 最小化 DM 损失，该损失融合了如下的多种损失 [14]：

$$\mathcal{L}^{(D)} = \mathcal{L}^{(\mathcal{X})} + \lambda^{(\mathcal{U})}\mathcal{L}^{(\mathcal{U})} + \lambda^{(reg)}\mathcal{L}^{(reg)}, \tag{3}$$

其中 $\lambda^{(\mathcal{U})}$ 表示无标签样本集合的损失的权重，$\lambda^{(reg)}$ 表示正则化项的权重。公式(3)中损失项的定义如下所示：

$$\mathcal{L}^{(reg)} = \sum_{c=1}^{|\mathcal{Y}|} \frac{1}{|\mathcal{Y}|} \log\left(\frac{1}{|\mathcal{X}'| + |\mathcal{U}'|} \sum_{\mathbf{x}\in(\mathcal{X}'\bigcup\mathcal{U}')} p_{\theta^{(D)}}(c|\mathbf{x})\right), \tag{4}$$

其中

$$\begin{aligned}
\mathcal{L}^{(\mathcal{X})} &= -\frac{1}{|\mathcal{X}'|} \sum_{(\hat{\mathbf{x}},\hat{\mathbf{y}})\in\mathcal{X}'} \sum_{c=1}^{|\mathcal{Y}|} \hat{\mathbf{y}}(c) \log(p_{\theta^{(D)}}(c|\hat{\mathbf{x}})), \\
\mathcal{L}^{(\mathcal{U})} &= \frac{1}{|\mathcal{U}'|} \sum_{(\hat{\mathbf{u}},\hat{\mathbf{q}})\in\mathcal{U}'} \|\hat{\mathbf{q}} - p_{\theta^{(D)}}(:|\hat{\mathbf{u}})\|_2^2,
\end{aligned} \tag{5}$$

公式中 $p_{\theta^{(D)}}(:|\hat{\mathbf{u}}) \in [0,1]^{|\mathcal{Y}|}$ 表示模型在输入为 $\hat{\mathbf{u}}$ 时对所有 $|\mathcal{Y}|$ 个标签的输出。在训练了 NetD 之后，我们使用更新后的训练集通过最小化 SL 损失(1)训练 NetS 模型 $f_{\theta^{(S)}}(c|\mathbf{x})$。更新过后的训练集表示为 $\{(\mathbf{x}_i, \hat{\mathbf{y}}_i)\}_{i=1}^{|\mathcal{D}|}$，其中

$$\hat{c}_i = \arg\max_c \left[(w_i^{\text{cl}}) p_{\theta^{(D)}}(c|\mathbf{x}_i) + (1 - w_i^{\text{cl}})\mathbf{y}_i(c))\right], i \in \{1, ..., |\mathcal{D}|\}$$

用于产生新的标签 $\hat{\mathbf{y}}_i = \text{onehot}(\hat{c}_i)$。

对于测试样本 $\mathbf{x}$ 的测试完全依赖于 NetD 分类器，通过如下的公式表示：$c^* = \arg\max_{c\in\mathcal{Y}} p_{\theta^{(D)}}(c|\mathbf{x})$。

3.4 实现

我们使用随机梯度下降 (SGD) 训练 18 层 PreAct Resnet[6](对于 NetS 和 NetD)，动量为 0.8，权重衰减系数为 0.0005，一个批次大小为 64。WarmUp 阶段的学习速率为 0.02，主要训练过程中前 100 个周期的学习速率为 0.02，然后衰减到 0.002。对于 NetD

和 NetS, WarmUp 阶段分别持续 10 和 30 个周期，其中 NetD 使用交叉熵损失进行训练而 NetS 使用主观逻辑损失 SL 进行训练，两者都使用训练集 $\mathcal{D}$。在 WarmUp 阶段之后，两个模型都接受 $E = 200$ 周期的训练。与 [14] 相似，增加的数据样本数为 $M = 2$，锐化温度为 $T = 0.5$，MixMatch 参数为 $\alpha = 4$，DM 损失(3)的正则化项权重为 $\lambda^{(reg)} = 1$。然而，不像 [14] 中人工基于 ρ 的取值选择权重 $\lambda^{(\mathcal{U})}$，我们对于所有的实验都设置 $\lambda^{(\mathcal{U})} = 25$。对于 GMM，我们使用 $\psi = 20$ 个分量，$\mu_{min} = 0.3$ 和 $\mu_{max} = 0.7$，因为这些值得到了稳定的结果。

4 实验

在前人研究闭集和开集噪声问题 [14,24,13] 的基础上，我们对闭集噪声的 CIFAR-10 数据集 [11] 进行了实验 [14,13]，使用开集噪声场景的 CIFAR-100(小规模)[11] 和 ImageNet32(大规模)[2] 数据集 [24,13]。CIFAR-10 拥有 10 个类，每类 5000 张 32×32 像素的训练图像 (共 50000 张训练图像)，以及 10000 张 32×32 像素的测试集，每类 1000 张图像。CIFAR-100 拥有 100 个类，每个类拥有 5000 张 32×32 像素的图像，而 ImageNet32 是 ImageNet[3] 的下采样变体，它拥有 1281149 张图像，1000 个类别，每幅图像拥有 32 ×32 个像素点。以上所有的数据集都是用提供的标签建立的，所以下面我们引入一个新的噪声标签基准评估，它结合了合成的开集和闭集标签噪声。

4.1 混合的开集和闭集标签噪声基准

本文提出的基准评估由实验中的标签噪声率定义，由 $\rho \in \{0.3, 0.6\}$ 表示，闭集标签噪声的比例由 $\omega \in \{0, 0.25, 0.5, 0.75, 1\}$ 表示，闭集标签噪声通过随机从 CIFAR-10 数据集中选择 $\rho \times \omega \times 100\%$ 比例的样本并且对称地打乱它们的标签，和 [14] 中使用的人工标签噪声相似。开集标签噪声通过随机选择 $\rho \times (1 - \omega) \times 100\%$ 比例的 CIFAR-10 训练样本然后使用从 CIFAR-100 或 ImageNet32 中任意选择的图像替换它们。此时一个随机的 CIFAR-10 标签被赋予每一张开集噪声图像。结果是基于 CIFAR-10 的干净测试集的分类精度，通过上面提出的基准设定进行评估。我们同时利用 t-SNE[16] 在特征空间中比较了 EDM 和其他相关方法的样本分布，以及 EDM 分离干净、闭集和开集噪声样本的有效性。

4.2 相关的对比方法

我们将提出的方法与下面列出的三种方法进行比较:

DivideMix[14] 是目前的 SOTA 方法，它将封闭集噪声标签学习问题转换成半监督学习问题。它遵循一个多模型方法，通过对每个 epoch 的训练样本的损失值拟合一个 2 分量高斯混合模型 (GMM) 来将训练数据分割成干净的和有噪声的子集。接下来，框架丢弃预测为噪声样本的标签，并使用 MixMatch[1] 来训练模型。

ILON[24] 介绍了开集标签噪声学习问题。文章中提出的方法是基于迭代的方法，基于异常因子算法（LOF）对样本重新设置权重。

RoG[13] 构建了一个生成分类器集合，该分类器由从 ResNet 模型的多层特征提取而成。RoG 的作者分别在闭集噪声和开集噪声上测试了他们的方法，这使得在我们的组合噪声设定中这个模型成为一个重要的基准。

4.3 结果和讨论

分类精度: 表1显示了本文提出的 EDM 模型在基准下的计算结果，与 RoG [13], ILON[24], 和 DivideMix[14] 的结果进行比较。评估依赖于不同的标签噪声率（$\rho \in 0.3, 0.6$ 和闭合噪声率 $\omega \in \{0, 0.25, 0.5, 0.75, 1\}$），并使用 CIFAR-100 和 ImageNet32 作为开集噪声数据集。结果表明，我们的方法 EDM 在 20 个噪声设定中的 17 个中优于所有竞争的方法，在其余的 3 个设定中接近第二。对 $\rho = 0.3$ 的情况，不管 ω 选择何值，也无论选择哪一个开集噪声数据集，EDM 都优于所有其他模型。在某些设定中 EDM 比第二好的模型精度高 3%。另一方面，RoG 和 ILON 的表现明显不如 EDM 和 DivideMix，特别是 $\rho = 0.6$ 的场景，精度的差异在某些情况下超过 15%。一般来说，当闭集噪声的比例增加时，RoG 和 ILON 的表现更差，而 DivideMix 和 EDM 则相反。同样明显的是，EDM 对于开集噪声比 DivideMix 更鲁棒，从分类结果中可以看出，当 ω 很小时非常明显。

特征表示: 在总噪声率 $\rho = 0.6$ 的情况下，我们在图4中显示了所有方法的 t-SNE 图 [16]，设定中闭合标签噪声率为 $\omega = 0.5$，使用 CIFAR-100 和 ImageNet32 作为开集噪声数据集。特别地，所有方法的特征都是从模型的最后一层中提取出来的 (在我们的例子中，我们使用来自 NetD 的特征，它是用于分类的，正如在第 3.3 节中解释的那样)。在图中，棕色样本来自开集噪声数据集，而所有其他颜色代表真正的 CIFAR-10 类。这清楚地表明，我们提出的 EDM 在将开集样本与其他干净的和闭集噪声样本的分离方面是相当有效的，而 DivideMix 和 ILON 对这些样本产生了过拟合，从 CIFAR-10 类周围开集噪声样本的分布就可以看出这一点。有趣的是，RoG 也显示有很好的分离性，但显然获得的分布比 EDM 更复杂。

噪声分类: 图5给出了在噪声率 $\rho = 0.6$，闭集噪声率为 $\omega \in \{0, 0.5, 1\}$ 的情况下，干净样本、开集样本和闭集样本的损失值分布，其中使用来自 CIFAR-100 和 Imagenet32 的样本作为开集噪声数据集。从这些图中可见，显然 SL 损失能够成功地区分上述三种样本，即使是只有一种噪声类型存在时也是如此，例如当 $\omega \in \{0, 1\}$。这表明 SL 损失在不确定性的识别方面对开集噪声样本是有效的。本文测试的方法中 DivideMix[14] 也试图将样本分离成干净样本和噪声样本。然而, 结果分布似乎不足以让三组之间明确分离，因为开集和闭集噪声样本的标签基本上是没有区别的。因此，DivideMix 能够将干净的样本从有噪声的样本中分离出来，而不是将闭集噪声从开集噪声中分离出来，从而迫使它在训练过程中对这两种噪声类型进行相似的处理 (即，两种类型都被视为闭集噪声)。这是不理想的，因为开放集样本将被分配一个不正确的训练标签，这最终会导致网络过拟合这些样本。

		ρ	0.3					0.6				
		ω	0	0.25	0.5	0.75	1	0	0.25	0.5	0.75	1
ImageNet32	RoG	Best	91.9	90.7	90.2	89.6	89.5	87.8	85.7	84.5	83.1	82.9
		Last	91.0	88.7	86.6	86.2	83.9	85.9	78.1	70.3	64.7	59.8
	ILON	Best	91.8	90.7	88.0	86.5	85.8	87.7	83.4	81.2	78.7	77.3
		Last	90.6	86.9	82.0	77.3	72.7	85.5	72.6	58.9	54.4	46.5
	DivideMix	Best	92.4	92.5	93.4	93.9	94.3	**92.5**	92.8	93.2	93.9	**94.7**
		Last	92.0	92.5	93.0	93.7	94.1	**92.5**	92.2	92.8	93.2	**94.6**
	EDM (Ours)	Best	**93.2**	**94.4**	**94.7**	**95.1**	**95.2**	91.2	**93.7**	**94.0**	**94.1**	94.1
		Last	**92.5**	**93.7**	**94.5**	**94.7**	**94.8**	90.9	**93.1**	**93.4**	**93.9**	94.1
CIFAR-100	RoG	Best	91.4	90.9	89.8	90.4	89.9	88.2	85.2	84.1	83.7	83.1
		Last	89.8	87.4	85.9	84.9	84.5	82.1	72.9	66.3	62.0	59.5
	ILON	Best	90.4	88.7	87.4	87.2	86.3	83.4	82.6	80.5	78.4	77.1
		Last	87.4	84.3	80.0	74.6	73.8	78.0	67.9	55.2	48.7	45.6
	DivideMix	Best	89.3	90.5	91.5	93.0	94.3	89.0	90.6	91.8	93.4	**94.4**
		Last	88.7	90.1	90.9	92.8	94.0	88.7	89.8	91.5	93.0	**94.3**
	EDM (Ours)	Best	**92.9**	**93.8**	**94.5**	**94.8**	**95.3**	**90.6**	**92.9**	**93.4**	**93.7**	94.3
		Last	**91.9**	**93.1**	**94.0**	**94.5**	**95.1**	**89.4**	**91.4**	**92.8**	**93.4**	94.0

表 1: 所有竞争方法和我们提出的 EDM 方法的基准评估结果。干净数据是从 CIFAR-10 中采样的，开集噪声样本是从 ImageNet32 和 CIFAR-100 中获得的。训练集中全部的噪声是通过 $\rho \in \{0.3, 0.6\}$ 表示的，闭集标签噪声比率为 $\omega \in \{0, 0.25, 0.5, 0.75, 1\}$，开集标签噪声比率为 $1 - \omega$。

5 结论

在本文中，我们研究了一种结合开集 [24,13] 和闭集噪声 [14,13] 的标签噪声学习问题的变体。为了在这个新的问题设定下测试各种方法，我们提出了一个新的基准，系统地改变总噪声率和闭集与开集噪声的比例。开集样本来自小规模数据集 (CIFAR-100) 或大规模数据集 (ImageNet32)，这样这些样本的真实标签不包含在主要数据集 (CIFAR-10) 中。我们认为，这样的问题设置更一般，类似现实生活中的标签噪声场景。然后，我们提出了 EvidentialMix 算法，成功地解决了这种新的噪声类型学习问题。我们使用主观逻辑损失 [21] 产生低损失的干净样本，高损失的闭集噪声样本，和中等损失的开集样本。训练数据中明确的分工让我们 **(1)** 识别和移除开集噪声样本以避免过度拟合，考虑到他们不属于任何已知的类；**(2)** 用半监督学习方法从预测的闭集样本 [14] 中学习。评估表明，我们提出的 EDM 方法比目前最先进的闭集问题研究方法 [14,13] 和开集问题研究方法 [24,13] 更有效地解决了这种新的组合开闭集标签噪声学习问题。

未来的工作: 引入这个问题的动机是发起研究社区的对话，以研究混合的开集和闭集标签噪声。接下来，我们将探索更具挑战性的噪声设定，例如将不对称的噪声 [19] 和语义噪声 [13] 合并到提出的组合标签噪声问题中。由于我们是第一个在受控的设定下解决

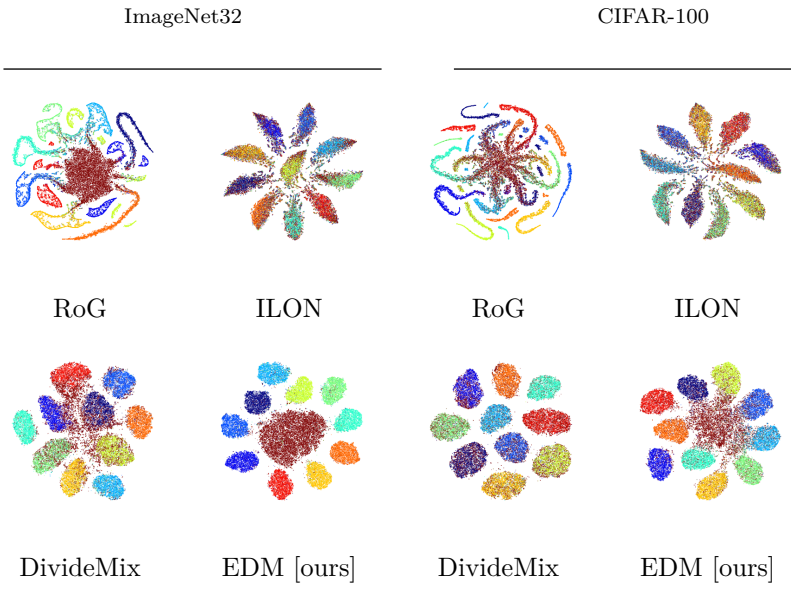

图 4: 相关方法和我们提出的 EDM 方法的 TSNE 图，其中整体的噪声率是 $\rho = 0.6$，开集噪声的比率是 $\omega = 0.5$，CIFAR-100 和 ImageNet32 表示开集数据集。其中**棕色**样本表示开集噪声样本，其他颜色表示真实的 CIFAR-10 类别。

这个问题的，没有先例说明如何将这些具有挑战性的噪声设定进行有意义的融合。例如，尽管不对称闭集噪声之前已经在文献 [19] 中进行了研究，但是它的对等物——不对称开集标签噪声需要什么是不清楚的; 例如，目前还不清楚如何在 CIFAR-10 和 ImageNet 类之间建立一个噪声转换矩阵。此外，我们认为研究其他类型的不确定性来识别开集噪声是有价值的，例如使用贝叶斯学习 [4] 来识别。我们也旨在探索这样的方法。

6 参考文献

[1] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. MixMatch: A Holistic Approach to Semi-Supervised Learning. arXiv eprints, page arXiv:1905.02249, May 2019.

[2] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. arXiv preprint arXiv:1707.08819, 2017.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.

[4] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In international conference on machine learning, pages 1050–1059, 2016.

[5] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. 2017.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings

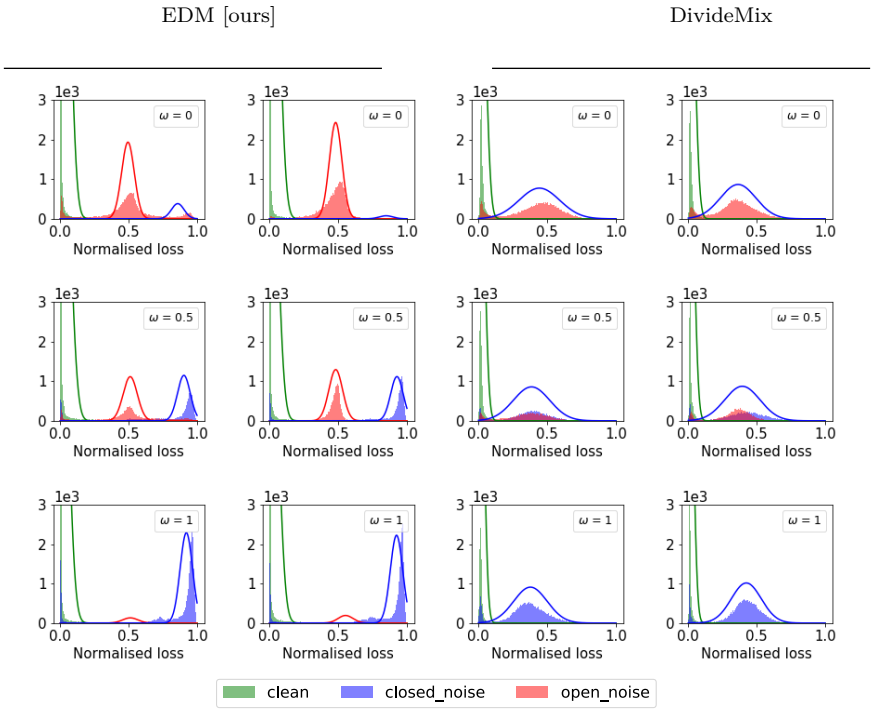

图 5: 对开集样本（橙色）、闭集样本（蓝色）、干净样本（绿色）的逐样本的 loss 分布。左边由 EDM 产生，右边由 DivideMix 产生。周期数为 $e = 100$，并且 $\rho = 0.6$ 和 $\omega = 0.3$。开集噪声样本 $\in$ {CIFAR-100,ImageNet32}。我们同时展示了估计的对干净样本，开集样本、闭集样本的 GMM 后验概率（对于 EDM），并展示了使用 DivideMix 模型时干净样本和噪声样本的后验概率分布。

in deep residual networks. In European conference on computer vision, pages 630–645. Springer, 2016.

[7] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In International Conference on Machine Learning, pages 2304–2313, 2018.

[8] Lance Kaplan, Federico Cerutti, Murat Sensoy, Alun Preece, and Paul Sullivan. Uncertainty Aware AI ML: Why and How. arXiv e-prints, page arXiv:1809.07882, Sept. 2018.

[9] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In Proceedings of the IEEE International Conference on Computer Vision, pages 101–110, 2019.

[10] Hans-Peter Kriegel, Peer Kr ́oger, Erich Schubert, and Arthur Zimek. Loop: local outlier probabilities. In Proceedings of the 18th ACM conference on Information and knowledge management, pages 1649–1652, 2009.

[11] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[12] Yann LeCun, Y. Bengio, and Geoffrey Hinton. Deep learning. Nature, 521:436–44, 05 2015.

[13] Kimin Lee, Sukmin Yun, Kibok Lee, Honglak Lee, Bo Li, and Jinwoo Shin. Robust inference via generative classifiers for handling noisy labels. arXiv preprint arXiv:1901.11300, 2019.

[14] Junnan Li, Richard Socher, and Steven C. H. Hoi. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. arXiv e-prints, page arXiv:2002.07394, Feb. 2020.

[15] Junnan Li, YongkangWong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5051–5059, 2019.

[16] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579–2605, 2008.

[17] Eran Malach and Shai Shalev-Shwartz. Decoupling"when to update"from"how to update". In Advances in Neural Information Processing Systems, pages 960 – 970, 2017.

[18] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1944–1952, 2017.

[19] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training Deep Neural Networks on Noisy Labels with Bootstrapping. arXiv e-prints, page arXiv:1412.6596, Dec. 2014.

[20] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. arXiv preprint arXiv:1803.09050, 2018.

[21] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty, 2018.

[22] Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. SELF: Learning to Filter Noisy Labels with Self-Ensembling. arXiv e-prints, page arXiv:1910.01842, Oct. 2019.

[23] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Advances in neural information processing systems, pages 1195–1204, 2017.

[24] Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative Learning with Open-set Noisy Labels. arXiv e-prints, page arXiv:1804.00092, Mar. 2018.

[25] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? arXiv preprint arXiv:1901.04215, 2019.

[26] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. arXiv preprint

arXiv:1611.03530, 2016.

[27] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412, 2017.