计 算 机 研 究 与 发 展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202330382
61(12): 3121−3133, 2024

一种基于在线蒸馏的轻量化噪声标签学习方法

黄贻望[1,2,3]　黄雨鑫[2]　刘　声[1,3]

[1]（铜仁学院大数据学院　贵州铜仁　554300）

[2]（福建理工大学计算机科学与数学学院　福州　350001）

[3]（贵州省公共大数据重点实验室（贵州大学）　贵阳　550025）

（hyxhh0226@163.com）

A Lightweight Noise Label Learning Method Based on Online Distillation

Huang Yiwang[1,2,3], Huang Yuxin[2], and Liu Sheng[1,3]

[1]（*School of Data Science* , *Tongren University*, *Tongren* , *Guizhou* 554300）

[2]（*School of Computer Science and Mathematics*, *Fujian University of Technology*, *Fuzhou* 350001）

[3]（*Guizhou Provincial Key Laboratory of Public Big Data* (*Guizhou University*), *Guiyang* 550025）

Abstract　Training deep learning models with noisy data containing lossy labels is a hot research topic in machine learning. Studies have shown that deep learning model training is susceptible to overfitting due to noisy data. Recently, a method combining meta-learning and label correction can make the model better adapt to the noisy data to mitigate the overfitting phenomenon. However, this meta-label correction method relies on the model's performance, and the lightweight model does not have good generalization performance under noisy data. To address this problem, we propose a knowledge distillation-based meta-label correction learning method (KDMLC), which treats the meta label correction model (MLC) composed of a deep neural network and a multilayer perceptron as a teacher model to correct the noise labels and guide the training of the lightweight model, and at the same time, KDMLC adopts a two-layer optimization strategy to train and enhance the generalization ability of the teacher model, so as to generate a higher-quality pseudo-labels for training the lightweight model. The experiments show that KDMLC improves the test accuracy by 5.50% compared with MLC method at high noise level; meanwhile, using Cutout data enhancement on the CIFAR10 dataset, KDMLC improves the test accuracy by 9.11% compared with MLC at high noise level, and the experiments on the real noisy dataset, Clothing1M, also show that KDMLC outperforms the other methods, verifying that KDMLC is better than the other methods, which verifies the feasibility and validity of KDMLC method.

Key words　pseudo labels; label correction; meta learning; knowledge distillation; noise data

摘　要　利用含有有损标签的噪声数据来训练深度学习模型是机器学习中的研究热点.研究表明深度学习模型训练易受噪声数据的影响而产生过拟合现象.最近,一种将元学习与标签校正相结合的方法能够使模型更好地适应噪声数据以减缓过拟合现象,然而这种元标签校正方法依赖于模型的性能,同时轻量化模型在噪声数据下不具备良好的泛化性能.针对这一问题,本文结合元学习提出一种基于在线蒸馏的轻量化噪声标签学习方法KDMLC（knowledge distillation-based meta-label correction learning),该方法将深度

收稿日期: 2023-05-15; **修回日期**: 2023-12-04

基金项目: 国家自然科学基金项目（62066040, 62261047）; 贵州省公共大数据重点实验室开放基金项目（2018BDKFJJ011）; 铜仁市科技局项目（铜仁市科研 [2022]5 号）

This work was supported by the National Natural Science Foundation of China (62066040, 62261047), the Foundation of Guizhou Provincial Key Laboratory of Public Big Data (2018BDKFJJ011), and the project of Tongren Science and Technology Bureau (Tongren City Scientific Research [2022]5).

通信作者: 黄贻望（yjsyhyw@gztrc.edu.cn）

神经网络与多层感知机构成的元标签校正（meta label correction, MLC）模型视为教师模型，对噪声标签进行校正并指导轻量化模型进行训练，同时采用双层优化策略训练并增强教师模型的泛化能力，从而生成更高质量的伪标签用于训练轻量化模型. 实验表明，KDMLC 在高噪声水平下对比 MLC 方法准确率提高了 5.50 个百分点；同时对 CIFAR10 数据集使用 Cutout 数据增强，KDMLC 在高噪声水平下对比 MLC 准确率提升了 9.11 个百分点，而在真实噪声数据集 Clothing1M 上的实验，KDMLC 也优于其他方法，验证了 KDMLC 的可行性和有效性.

关键词 伪标签；标签校正；元学习；知识蒸馏；噪声数据

中图法分类号 TP181

在现实生活中可以通过搜索引擎与自动标签软件等方式轻易地获取大量的非专业化标注的数据集，但这些数据集往往会含有一定程度的噪声标签，导致深度神经网络出现过拟合现象，从而降低模型的泛化能力[1]. 同时 Nakkiran 等人[2]的工作表明轻量化模型更难以在噪声数据下进行有效的训练. 因此，使不同规模的深度神经网络能够有效地在噪声标签数据下进行鲁棒性训练成为当下一个重要的研究热点.

为解决这一热点问题，学术界涌现出多种噪声标签学习方法，Jindal 等人[3]设计一个适用于在噪声数据下学习的鲁棒性架构，通过估计的噪声转移矩阵作为辅助信息能够使深度神经网络在噪声数据具有良好的泛化能力；同时研究人员也将正则化技术[4]广泛运用到噪声标签学习中，但单纯的使用正则化方法并不能很好地解决模型对噪声数据的过拟合问题. 欧阳宵等人[5]利用标签相关性作为先验知识有效地处理了多标签数据中的标签噪声. 苗壮等人[6]设计了一种双教师共识去噪策略的伪标签去噪方法，利用蒸馏算法有效地去除了初始伪标签中的噪声. Jiang 等人[7]提出一个弱监督学习方法，训练集由一组较小的干净数据与一组较大的噪声数据构成来训练深度神经网络. 基于这种弱监督学习方法，Zheng 等人[8]将元学习[9]应用于标签校正策略中，并通过分类网络与多层感知机之间的元学习过程生成更高质量的伪标签，增强了模型的自适应能力. 当前的标签校正网络架构生成的伪标签质量过于依赖主模型的性能，且大容量模型难以部署在边缘设备上[10]，而且尽管轻量级模型适用于边缘设备，但在处理噪声数据时识别准确度有限.

针对轻量化模型难以在噪声数据中稳定训练的问题，可以采用知识蒸馏[11]技术在不显著影响模型性能[12]的前提下解决. 元伪标签（meta pseudo labels, MPL）方法[13]利用知识蒸馏，使无标签数据通过教师模型生成伪标签用于训练学生模型，再通过元学习框架使学生模型对教师模型进行反馈，从而起到一个辅助作用使教师模型能够生成更高质量的伪标签. 而元标签校正[8]（meta label correction, MLC）方法生成的伪标签质量过于依赖分类网络性能（主模型），在分类任务中对于一些样本较少的类别并不具备良好的鲁棒性，甚至对比于普通的深度神经网络会更快地适应噪声数据.

本文提出一种基于在线蒸馏的轻量化噪声标签学习方法 KDMLC（knowledge distillation-based meta-label correction learning）. KDMLC 在采用 MLC 作为教师模型的前提下，引入知识蒸馏技术，利用轻量化模型并结合元学习（反馈机制）辅助 MLC 模型中的主模型进行训练. 具体而言，轻量化模型利用 MLC 中的主模型生成的软标签与多层感知机（multi-layer perceptron, MLP）网络生成的伪标签进行训练，并将轻量化模型在干净数据上的损失反馈给主模型，从而有效地缓解主模型过拟合噪声数据，增强主模型的泛化性能. 同时 KDMLC 方法使轻量化模型也能够在噪声数据下拥有接近甚至超越主模型的性能.

本文的主要贡献包括 3 个方面：

1）设计了基于知识蒸馏的标签校正模型，利用轻量化模型的训练反馈对主模型进行更新，提升了主模型性能，缓解了模型的过拟合现象.

2）增强了轻量化模型（学生模型）在噪声数据中的泛化能力，使其在噪声数据训练下达到与在干净数据集训练出的轻量化模型相近的性能.

3）所提模型在不同噪声系数下的数据集上的一系列对比实验取得了较好的效果，教师模型与学生模型性能均优于 MLC.

1 相关工作

1.1 标签噪声学习

在噪声数据集中，深度神经网络需要尽可能减

少噪声标签带来的负面影响，才能在噪声数据中得到有效的训练. 丁家满等人[14]提出了一种基于正则化的半监督标签学习方法，从实例相似性与标签相关性角度构建正则化项以提高模型分类效果. Van Rooyen 等人[15]通过噪声转移矩阵可以将干净数据集的干净标签转化为噪声标签.

一些工作利用噪声转移矩阵作为辅助信息，例如 Goldberger 等人[16]将其作为可学习参数内嵌至神经网络，以端到端的形式与网络模型参数一起学习，但这种方法在高噪声水平下难以稳定发挥；Patrini 等人[17]提出前向损失校正与后向损失校正，其仅需估计每个类被破坏成另一个类的概率而不受网络体系结构影响. 基于前向损失校正，Hendrycks 等人[18]提出了 GLC（gold loss correction），利用一小部分干净数据来更加准确地估计噪声转移矩阵.

基于 GLC 的工作，Li 等人[19]从小型干净数据集中提炼出所学到的知识，再通过知识图谱来指导蒸馏过程，以促进模型更好地在噪声数据集下进行训练. 除此之外，文献 [20] 使用元学习技术辅助生成软标签，并以端到端的方式学习深度神经网络参数. Wang 等人[21]将元学习应用于标签校正学习，通过一个 MLP 网络对噪声标签进行校正后生成伪标签用于主模型的训练，并采用双层优化策略使其形成了一种端到端的训练模式.

KDMLC 方法不同于其他利用知识蒸馏进行噪声标签学习方法，其更侧重于模型整体性能的稳定性. 而 MPL 侧重于将学生模型视为辅助模型帮助教师模型训练；文献 [19] 则是使用小型干净数据集训练教师模型，并以离线蒸馏的方式指导学生模型训练. 具体提升如表 1 所示.

Table 1 Comparison of Label Correction Methods Based on Knowledge Distillation

表 1 基于知识蒸馏的标签校正方法对比

	MPL[13]	Li 等人[19]所提方法	KDMLC（本文）
在线蒸馏策略	√	×	√
轻量化模型性能提升	×	√	√
教师模型性能提升	√	×	√
教师模型的训练数据类型	噪声数据	干净数据	噪声数据

1.2 知识蒸馏

知识蒸馏是指通过教师模型提炼出不同类型的“知识”并有效地传递给学生模型学习的一种模型压缩技术. 基于响应的知识采用教师模型最后输出的数据对学生模型进行训练. 而基于特征的知识[22]是对基于响应的知识的一个扩展，它采用教师模型中间层输出的特征知识对学生模型相应的卷积层进行指导. 基于关系的知识[23]则是通过教师模型不同层与不同样本之间的关系知识指导学生模型学习.

一般来说，教师模型的训练过程是相对独立，Hinton 等人[11]通过预先训练教师模型，再采用离线蒸馏的方式将基于响应的知识从教师模型转移到学生模型中. Heo 等人[24]从决策边界的角度出发，认为决策边界应当与真实的分类边界接近，从而利用决策边界对学生网络进行蒸馏. 这是一种单方向和 2 阶段的蒸馏方法，但训练复杂的教师模型无法避免，且在离线蒸馏中学生模型过于依赖教师模型和过于侧重教师模型迁移知识的内容，它们之间的性能始终存在差距.

在线蒸馏与自蒸馏能够有效缓解离线蒸馏的局限性. Fang 等人[25]提出的 FastDFKD（faster data-free knowledge distillation）将元学习与生成对抗网络相结合应用于知识蒸馏，通过元生成器生成高质量的数据在线训练学生网络. 魏秀参等人[26]将知识蒸馏与多示例学习相结合，使得模型能够在多示例学习下以极低的存储很好地保留模型的旧知识. 同时 Chen 等人[27]提出一种知识回顾提炼的知识蒸馏方法 ReviewKD（distilling knowledge via knowledge review），该方法将基于响应的知识与基于特征的知识相结合，且教师模型通过将不同卷积层的特征知识迁移给学生模型同级卷积层或浅层卷积层，使学生模型达到更优的效果.

2 KDMLC 方法

本文所提方法 KDMLC 主要将一个元标签校正模型作为教师模型，基于元标签校正模型生成的伪标签通过知识蒸馏框架训练轻量化学生模型，并通过元学习算法将学生模型的损失反馈给教师模型，形成一个端到端的模型结构.

2.1 基本元标签校正模型（MLC）

$D=\{x_l,y_l\}^m$, $l=1,2,\cdots,m$, 代表一组干净的标签数据示例，$D_u=\{x_u,y_u\}^n$, $u=1,2,\cdots,n$, 代表含噪声标签的数据示例，干净数据示例的大小为 m，噪声数据集示例的大小为 n, $m<n$.

MLC 是由一个主模型 f_{θ_t} 与标签校正网络（label correction network, LCN）g_ω 共同构成一个元学习模块如图 1 所示.

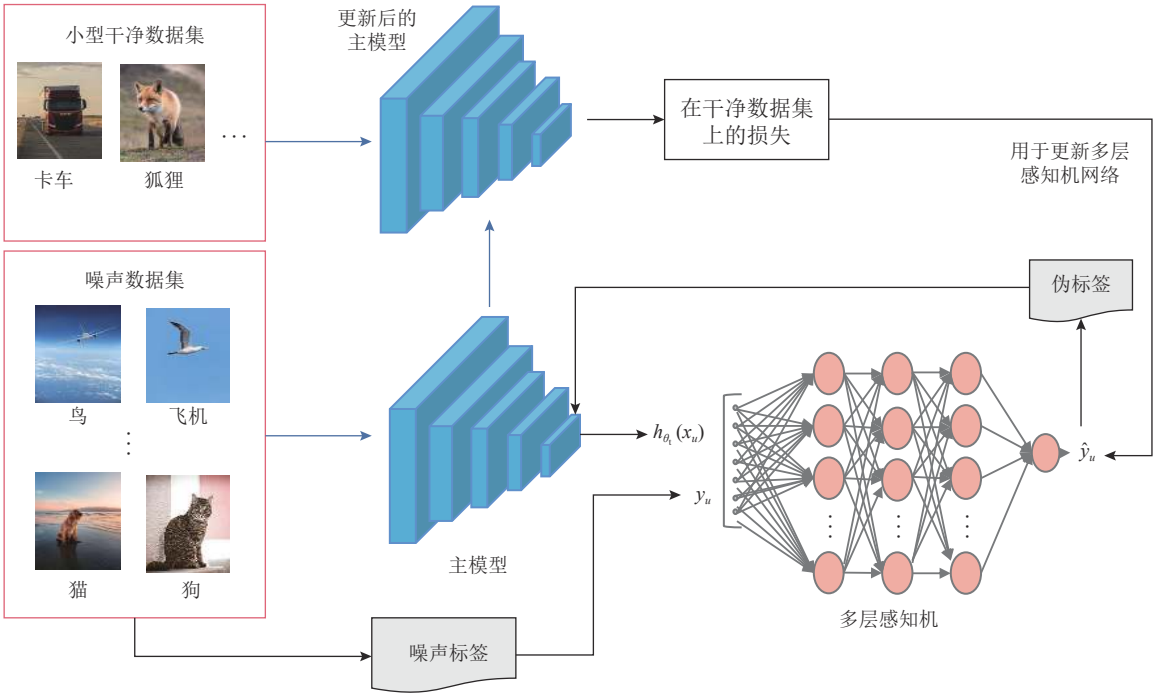


Fig. 1 MLC framework

图 1 MLC 框架

主模型是一个参数为θ_t的 Resnet 网络，它可以表示为$f_{\theta_t}(x_u)$，y_t是$f_{\theta_t}(x_u)$经过 softmax 层所输出的伪标签；LCN 是一个参数为ω的多层感知机网络，函数形式可以表示为$\hat{y}_u=g_\omega\left(h_{\theta_t}(x_u),y_u\right)$，其中$h_{\theta_t}(x_u)$表示噪声数据$x_u$在主模型上的特征输出. 主模型与 LCN 之间相互依赖，主模型通过 LCN 生成的元伪标签进行训练并更新模型参数，主模型参数更新为

$$\theta_t'=\theta_t'(\omega)=\theta_t-\eta_t\nabla_{\theta_t}L_{D_u}(\omega,\theta_t)\,, \tag{1}$$

其中η_t为超参数代表主模型的学习率，且$L_{D_u}(\omega,\theta_t)\triangleq E_{D_u}L\left(g_\omega\left(h_{\theta_t}(x_u),y_u\right),f_{\theta_t}(x_u)\right)$，若 LCN 生成了高质量的伪标签，则更新后的主模型在干净数据集上应实现较低的损失，MLC 的总体形式可以表现为函数：

$$\begin{aligned}&\min_{\omega}E_DL\left(y_l,f_{\theta_t'}(x_l)\right),\\&\text{s.t. }\theta_t'(\omega)=\arg\min_{\theta_t}E_{D_u}L\left(g_\omega\left(h_{\theta_t}(x_u),y_u\right),f_{\theta_t}(x_u)\right).\end{aligned} \tag{2}$$

2.2 元伪标签学习（MPL）

MPL 的模型框架由教师模型与学生模型共同构成. 未标记的数据可以通过教师模型生成伪标签从而指导学生模型训练，同时将更新后的学生模型在标记的数据集上的表现反馈给教师模型，使得教师模型能够生成更好的伪标签训练学生模型.

MPL 方法采用 2 个相同的模型作为教师模型f_{θ_t}与学生模型f_{θ_s}. $\hat{y}_u=f_{\theta_t}(x_u;\theta_t)$代表教师网络对无标签数据$x_u$的预测伪标签，同时$f_{\theta_s}(x_u;\theta_s)$与$f_{\theta_s}(x_l;\theta_s)$分别代表学生网络对干净标签数据与噪声标签数据的软预测. 而学生网络的参数更新依赖于教师模型预测的伪标签$\hat{y}_u$，即θ_s的更新依赖于θ_t，可以显式地将其依赖性表现为$\theta_s^{\mathrm{PL}}(\theta_t)$，即

$$\theta_s^{\mathrm{PL}}(\theta_t)=\theta_s-\eta_s\nabla_{\theta_s}L_u(\theta_t,\theta_s)\,, \tag{3}$$

其中$L_u(\theta_t,\theta_s)=E_{x_u}\left[CE\left(f_{\theta_t}(x_u;\theta_t),f_{\theta_s}(x_u;\theta_s)\right)\right]$，$\eta_s$为学生网络的学习率. 而教师网络是依据学生网络在标签数据上的损失$L_l\left(\theta_s^{\mathrm{PL}}(\theta_t)\right)$进行更新：

$$\theta_t'=\theta_t-\eta_t\nabla_{\theta_t}L_l\left(\theta_s-\nabla_{\theta_s}L_u(\theta_t,\theta_s)\right). \tag{4}$$

MPL 的总体形式可以表现为：

$$\begin{aligned}&\min_{\theta_t}L_l\left(\theta_s^{\mathrm{PL}}(\theta_t)\right),\\&\text{s.t. }\theta_s^{\mathrm{PL}}(\theta_t)=\arg\min_{\theta_s}L_u(\theta_t,\theta_s).\end{aligned} \tag{5}$$

2.3 基于响应的知识蒸馏

传统的知识蒸馏利用性能更好的教师模型所提炼出的监督信息来训练学生模型，使学生模型拥有更好的性能. Hinton 等人[11]通过引入温度系数 *temp* 来改造 softmax 函数

$$q_i=\frac{\exp(z_i/temp)}{\sum_j\exp\left(z_j/temp\right)}. \tag{6}$$

原始 softmax 函数的 *temp*=1，当温度系数 *temp* 越高时，softmax 函数输出每个值的概率分布越均匀，即增加了对负标签的关注程度，而负标签中都包含一定的信息，尤其是一些值显著高于平均值的负标签对提升模型性能具有巨大价值. 如图 2 所示，知识蒸

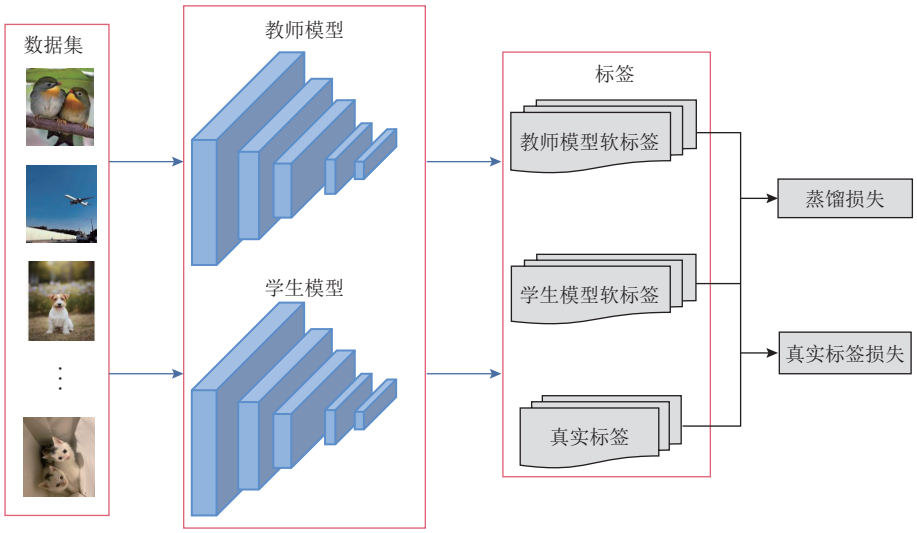


Fig. 2 Response based knowledge distillation framework

图 2 基于响应的知识蒸馏框架

馏通过真实标签与教师模型生成的软标签对学生模型进行训练：

$$L_{\text{soft}} = -\sum_{j}^{N} y_j^{\text{soft}} log\left(q_j\right), \tag{7}$$

$$L_{\text{KD}} = \alpha L_{\text{soft}} + \beta L_{\text{hard}}, \tag{8}$$

其中L_{soft}是指教师模型生成的软标签与学生模型输出之间的损失，而L_{hard}代表真实标签与学生模型输出之间的交叉熵损失，y_j^{soft}是学生模型经过 softmax 后的输出. 使用真实标签能够有效地避免教师模型将错误的信息传给学生模型.

2.4 基于在线蒸馏的元伪标签学习

本文采用了 MLC 作为教师模型，学生模型为普通的 Resnet 网络. 学生模型的参数为θ_{s}，其函数表达式为$y_{\text{s}} = f_{\theta_{\text{s}}}(x_u)$. MLC 的优点在于它可以通过 LCN 生成的元伪标签来训练主模型，并利用参数更新后的主模型在干净数据上的训练损失反馈给 LCN，使 LCN 能生成质量更高的伪标签.

本文模型的主体框架如图 3 所示，由 MLC 模块与知识蒸馏模块构成. 其中学生模型通过主模型与 LCN 生成的伪标签进行训练. 由于知识蒸馏对真实标签有较高的依赖，所以将 LCN 生成的元伪标签视为真实标签，具体流程如式(9)(10)(11)所示：

$$L_{\text{real}} = L_{D_u}(\theta_{\text{s}}, \omega) = CrossEntropy\left(f_{\theta_{\text{s}}}(x_u), \hat{y}_u\right), \tag{9}$$

$$L_{\text{pseudo}} = L_{D_u}(\theta_{\text{s}}, \theta_{\text{t}}) = L_{\text{KL}}\left(y_{\text{s}}^{temp}, y_{\text{t}}^{temp}\right), \tag{10}$$

$$L_{\text{KD}}(\theta_{\text{s}}, \theta_{\text{t}}, \omega) = \alpha L_{\text{real}} + \beta L_{\text{pseudo}}, \tag{11}$$

其中，α与β为超参数，L_{real}是指在噪声数据集下 LCN 生成的元伪标签与学生模型输出之间的交叉熵损失，L_{pseudo}代表主模型生成的伪标签与学生模型输出之间的 KL(Kullback-Leibler divergence)散度，y_{s}^{temp}与y_{t}^{temp}分别代表主模型与学生模型蒸馏输出.

从而能够通过蒸馏损失L_{KD}来更新学生模型参数：

$$\theta'_{\text{s}} = \theta_{\text{s}} - \eta_{\text{s}} \nabla_{\theta_{\text{s}}} L_{\text{KD}}(\theta_{\text{s}}, \theta_{\text{t}}, \omega), \tag{12}$$

其中η_{s}代表学生模型的学习率.

学生模型依赖于教师模型生成的伪标签进行训练更新，即$\theta'_{\text{s}} = \theta_{\text{s}}^{\text{PL}}(\theta_{\text{t}}, \omega)$，所以教师模型生成伪标签的质量决定了学生模型的性能. 为了使教师模型生成更高质量的伪标签，可以通过借鉴 MPL 模型的学生模型反馈机制(元学习双层优化策略)，将学生模型在小型干净数据集$D = \{x_l, y_l\}^m$上的训练损失反馈给主模型，从而提升主模型性能并生成更高质量的伪标签，具体公式为：

$$\begin{aligned} &\min_{\theta_{\text{t}}} L_{\text{MPL}}\left(\theta_{\text{s}}^{\text{PL}}(\theta_{\text{t}}, \omega)\right), \\ &\text{s.t. } \theta'_{\text{s}} = \theta_{\text{s}}^{\text{PL}}(\theta_{\text{t}}, \omega) = \theta_{\text{s}} - \eta_{\text{s}} \nabla_{\theta_{\text{s}}} L_{\text{KD}}, \end{aligned} \tag{13}$$

其中$L_{\text{MPL}} = E_D\left(CE\left(y_l, f_{\theta_{\text{s}}}\left(x_l, \theta'_{\text{s}}\right)\right)\right)$. 同时主模型的参数更新也依赖于 LCN 所生成的元伪标签$\hat{y}_u$，即

$$\theta'_{\text{t}} = \theta_{\text{t}} - \eta_{\text{t}} \nabla_{\theta_{\text{t}}}\left(L_{\text{MPL}}\left(\theta'_{\text{s}}\right) + L_{D_u}(\omega, \theta_{\text{t}})\right), \tag{14}$$

再通过计算更新后的主模型在干净数据集上的损失，反馈给 LCN，从而更新 LCN 参数：

$$\omega' = \omega - \eta_{\omega} \nabla CE\left(y_l, f_{\theta'_{\text{t}}}(x_l)\right), \tag{15}$$

其中η_{ω}代表 LCN 的学习率.

2.5 反馈机制的理论推导与分析

针对式(14)中的$\nabla_{\theta_{\text{t}}} L_{\text{MPL}}\left(\theta'_{\text{s}}\right)$，在一批噪声数据中主模型生成的伪标签为$y_{\text{t}} \sim D_u(x_u, y_u)$，LCN 生成的元

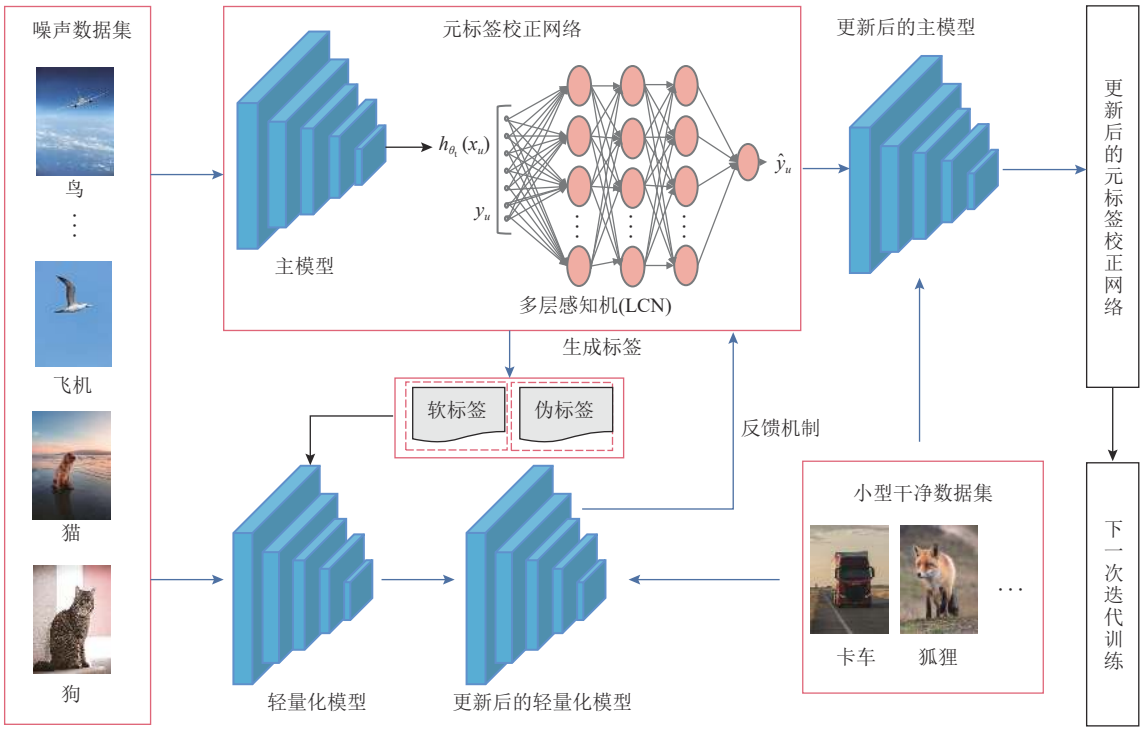


Fig. 3 Framework of KDMLC

图 3 KDMLC 框架

伪标签为$\hat{y}_u \sim D_u(x_u, y_u)$，为了简化公式，将学生模型与教师模型分别表示为 S 与 T，且 LCN 用 Y 表示. 通过更新主模型参数，以最小化学生模型在小型干净数据集 $D(x_l, y_l)$ 上的期望.

$$\frac{\partial R}{\partial \theta_t} = \frac{\partial}{\partial \theta_t} CE\left(y_l, S\left(x_l; E_{(y_t,\hat{y}_u)}\left[\theta_s - \eta_s \nabla_{\theta_s} \times \left[L_{KL}(y_t, S(x_u, \theta_s)) + CE(\hat{y}_u, S(x_u; \theta_s))\right]\right]\right)\right). \tag{16}$$

为了简化符号，作如下定义：

$$\theta'_s = E_{(y_t,\hat{y}_u)}\left[\theta_s - \eta_s \nabla_{\theta_s}\left[L_{KL}(y_t, S(x_u; \theta_s)) + CE(\hat{y}_u, S(x_u; \theta_s))\right]\right], \tag{17}$$

从而依据链式法则，可将式(16)转化为

$$\frac{\partial R}{\partial \theta_t} = \frac{\partial}{\partial \theta_t} CE\left(y_l, S\left(x_l; \theta'_s\right)\right) = \frac{\partial CE\left(y_l, S\left(x_l; \theta'_s\right)\right)}{\partial \theta'_s} \frac{\partial \theta'_s}{\partial \theta_t}. \tag{18}$$

式(18)中的第 2 个等号右侧的第 1 项可由反向传播直接计算得出，故主要关注于第 2 项的推导过程，即

$$\begin{aligned} \frac{\partial \theta'_s}{\partial \theta_t} = {} & \frac{\partial}{\partial \theta_t} E_{(y_t,\hat{y}_u)}\left[\theta_s - \eta_s \nabla_{\theta_s}\left[L_{KL}(y_t, S(x_u; \theta_s)) + CE(\hat{y}_u, S(x_u; \theta_s))\right]\right] = \\ & \frac{\partial}{\partial \theta_t} E_{(y_t,\hat{y}_u)}\left[\theta_s - \eta_s\left[\frac{\partial L_{KL}(y_t, S(x_u; \theta_s))}{\partial \theta_s} + \frac{\partial CE(\hat{y}_u, S(x_u; \theta_s))}{\partial \theta_s}\right]\right], \end{aligned} \tag{19}$$

故为了简化符号，作如下定义：

$$g_t(y_t) = \frac{\partial L_{KL}(y_t, S(x_u; \theta_s))}{\partial \theta_s}, \tag{20}$$

$$g_u(\hat{y}_u) = \frac{\partial CE(\hat{y}_u, S(x_u; \theta_s))}{\partial \theta_s}, \tag{21}$$

从而式(19)可转化为

$$\frac{\partial \theta'_s}{\partial \theta_t} = -\eta_s \frac{\partial}{\partial \theta_t} E_{(y_t,\hat{y}_u)}\left[g_t(y_t) + g_u(\hat{y}_u)\right], \tag{22}$$

y_t 与 $\hat{y}_u$ 通过教师模型生成，所以与 θ_t 存在依赖关系. 但由式(20)(21)可知，$g_t(y_t)$ 和 $g_u(\hat{y}_u)$ 与 θ_t 无关，其中 $E_{(y_t,\hat{y}_u)} \times \left[g_t(y_t) + g_u(\hat{y}_u)\right] = E_{y_t} \times g_t(y_t) + E_{\hat{y}_u} \times g_u(\hat{y}_u)$. 所以可以根据贝尔曼期望与策略梯度可做如下转化：

根据策略梯度中的状态价值函数将式(20)假设为

$$V_t(x_u; \theta_t) = E_{y_t}\left[g_t(y_t)\right] = \sum_{y_t} P(y_t|x_u; \theta_t) g_t(y_t), \tag{23}$$

其中期望 E_{y_t} 可以写为概率密度函数的等价连加的形式，即 $E_{y_t} = \sum_{y_t} P(y_t|x_u; \theta_t) g_t(y_t)$. 从而可通过策略梯度将公式转化为

$$\begin{aligned} \frac{\partial V_t}{\partial \theta_t} = {} & \sum_{y_t} \frac{\partial P(y_t|x_u; \theta_t)}{\partial \theta_t} g_t(y_t) = \\ & \sum_{y_t} P(y_t|x_u; \theta_t) \frac{\partial \log P(y_t|x_u; \theta_t)}{\partial \theta_t} g_t(y_t) = \\ & E_{y_t}\left[\frac{\partial \log P(y_t|x_u; \theta_t)}{\partial \theta_t} g_t(y_t)\right] = \\ & E_{y_t}\left[\frac{\partial CE(y_t, T(x_u; \theta_t))}{\partial \theta_t} g_t(y_t)\right]. \end{aligned} \tag{24}$$

式(24)中可以根据交叉熵损失函数的定义对$\log P(y_t|x_u;\theta_t)$进行替换，同理式(21)可假设为

$$V_u(x_u;\theta_t)=E_{\hat{y}_u}\left[g_u(\hat{y}_u)\right]=\sum_{\hat{y}_u}P(\hat{y}_u|x_u;\omega)g_u(\hat{y}_u),\quad(25)$$

即可推导出

$$\frac{\partial V_u}{\partial\theta_t}=E_{\hat{y}_u}\left[\frac{\partial logP(\hat{y}_u|x_u;\omega)}{\partial\theta_t}g_u(\hat{y}_u)\right]=E_{\hat{y}_u}\left[\frac{\partial CE\left(\hat{y}_u;Y\left(h_{\theta_t}(x_u);\omega\right)\right)}{\partial\theta_t}g_u(\hat{y}_u)\right],\quad(26)$$

故式(18)可转化为

$$\frac{\partial R}{\partial\theta_t}=\frac{\partial CE\left(y_l,S\left(x_l,\theta_s'\right)\right)}{\partial\theta_s'}\frac{\partial\theta_s'}{\partial\theta_t}=\eta_s\frac{\partial CE\left(y_l,S\left(x_l,\theta_s'\right)\right)}{\partial\theta_s'}\times E_{(y_t,\hat{y}_u)}\left\{\left[\frac{\partial CE(y_t,T(x_u;\theta_t))}{\partial\theta_t}g_t(y_t)\right]+\left[\frac{\partial CE\left(\hat{y}_u;Y\left(h_{\theta_t}(x_u);\omega\right)\right)}{\partial\theta_t}g_u(\hat{y}_u)\right]\right\}.\quad(27)$$

最终可根据式(14)更新主模型参数，KDMLC方法的具体流程为：

算法 1. 基于在线蒸馏的轻量化噪声标签学习方法.

输入：超参数α, β，噪声水平系数γ，蒸馏温度*temp*，学习率*lr*；

输出：更新后的模型参数θ_t', θ_s'.

① for i=0 to $N-1$

② 教师模型对噪声数据生成相应的软标签y_t与伪标签$\hat{y}_u$；

③ 使用生成的标签训练学生模型，并根据式(9)更新学生模型参数；

④ 计算主模型对数输出与元伪标签$\hat{y}_u$之间的交叉熵损失并计算其梯度$\nabla_{\theta_t}L_{D_u}(\omega,\theta_t)=\nabla_{\theta_t}CE\left(f_{\theta_t}(x_u),\hat{y}_u\right)$；

⑤ 计算更新后的学生模型在干净数据集上的损失，并根据式(24)计算损失梯度；

⑥ 根据式(14)更新主模型参数；

⑦ 计算更新后的主模型在小型干净数据集上的损失来计算其梯度，并根据式(15)更新LCN参数；

⑧ end for

3 实验与分析

3.1 实验数据

为了评估模型的性能，本节使用3种图像数据集来进行实验，数据集分别是CIFAR10, CIFAR100, Clothing1M[28].

CIFAR10与CIFAR100分别拥有10个类别与100个类别，CIFAR10和CIFAR100都包含50 000个训练数据与10 000个测试数据，每个数据图像的大小为32×32. 我们从训练数据中分割出1 000个干净的训练样本作为元数据集，用于评估参数更新后的模型性能. 然后通过人工合成噪声数据的方式将剩余的49 000个训练数据转变为噪声数据，本节主要通过以下2种方法合成噪声数据集：

1)统一标签噪声(UNIF). 在具有n个类别的干净数据集中，设定一个标签损坏概率为γ，根据标签损坏概率将真实标签y均匀损坏为任何其他类，同时有$(1-\gamma)$的概率保持原有的真实标签.

2)翻转标签噪声(FLIP). 在具有n个类别的干净数据集中，一个具有真实标签y的干净例子被随机地翻转到其余类中的概率为γ，并以$(1-\gamma)$的概率保持原有的真实标签.

Clothing1M是一个大规模的带有弱标签的真实噪声数据集，拥有14个类别，共100万张图像. Clothing1M中的大部分图像受现实噪声的影响而被随机分配伪标签，共7.24万张图像被挑选到干净数据集. Clothing1M中带有强标签的图像被用于训练、验证、测试，分别包含5万、1.4万、1万张图像. 详细信息如表2所示.

Table 2 Settings of Sample Numbers in the Dataset

表2 数据集样本个数设置

数据集类型	数据集(类别数)		
	CIFAR10(10)	CIFAR100(100)	Clothing1M(14)
训练集	50×10^3	50×10^3	$1\,050\times10^3$
测试集	10×10^3	10×10^3	10×10^3
干净数据集	1×10^3	1×10^3	50×10^3
噪声数据集	49×10^3	49×10^3	$1\,000\times10^3$

3.2 基准模型及实验设置

本节将所提出的方法与Fine-tuning, GEC[29], GLC[18], MW-Net[30], MLC[8], Co-teaching[31], L2RW[32]等方法进行对比，来验证本文方法的性能.

本节主要将CIFAR10和CIFAR100数据集人工合成为对称噪声数据和非对称噪声数据用于模型的训练，同时为了确保实验的可对比性，mini-batch大小均设为100，迭代次数设为120，采用随机梯度下降(SGD)更新模型参数，同时采用标准的数据增强方式使实验结果更加公正.

在CIFAR10数据集下，采用Resnet34网络作为教师模型的主模型，在CIFAR100数据集下采用Resnet32作为教师模型的主模型，学生模型采用Resnet18网络. 教师模型与学生模型采用不同的学习率优化方法，教师模型使用MultiStepLR方法分别在80次与100次迭代中改变学习率大小，学生模型将warmup与余弦法相结合作为学习率优化策略. 由于蒸馏温度对实验的影响较大，所以在CIFAR10数据集上教师模型与学生模型之间蒸馏温度 *temp* 设置为3，在CIFAR100数据集上主模型与学生模型之间蒸馏温度 *temp* 设置为1.

3.3 实验结果分析

3.3.1 CIFAR10实验结果分析

表3显示了上述方法对CIFAR10数据集在2种噪声类型下与不同方法的测试结果对比. 对于UNIF噪声类型我们对比了4种不同的噪声水平，分别为20%, 40%, 60%, 80%. 同时在FLIP噪声类型下对20%和40%共2个噪声水平下进行测试.

从表3可以观察到，在CIFAR10中所有的噪声水平上，KDMLC方法与其他方法对比均取得了较好的结果. 在低噪声水平下，学生模型性能均优于MLC；在较高的噪声水平下，尤其是在噪声水平为80%时，学生模型的准确率比MLC方法的准确率高出5.50个百分点，同时教师模型的准确率也达到了77.34%，均高于目前方法的性能.

图4显示了MLC与KDMLC在CIFAR10下不同噪声水平的测试精度. 当噪声水平大于50%时KDMLC与基础方法MLC相比有明显的提升，证明KDMLC在高噪声水平下拥有更优的泛化性能. 而在FLIP噪声类型下，由于一个类中的任何实例只能以固定概率翻转到另一个类中，在这种噪声下，噪声水平应低于50%，否则错误标记数据占大多数将给模型的泛化性能带来较大的影响. 在高噪声水平时，KDMLC的鲁棒性远优于MLC.

Table 3 Performance Comparison of KDMLC and Different Methods on CIFAR10 Dataset

表3 KDMLC与不同方法在CIFAR10数据集的性能对比 %

方法	UNIF噪声的4种噪声水平				FLIP噪声的2种噪声水平	
	20%	40%	60%	80%	20%	40%
Fine-tuning	91.17	87.34	83.75	69.28	93.11	91.04
GEC[29]	90.27	88.50	83.70	57.27	90.11	85.24
Co-teaching[31]	86.05	75.15	73.31		83.68	75.62
L2RW[32]		87.11	82.60		87.86	85.66
GLC[18]	91.47	88.52	84.08	64.21	92.46	91.74
MW-Net[30]	91.48	87.77	81.98	65.88	91.47	91.64
MLC[8]	89.83	87.32	83.92	74.43	91.81	91.35
$KDMLC_{NT}$	89.97	88.16	85.13	77.34	92.08	91.62
$KDMLC_{NS}$	**90.25**	**89.78**	**86.76**	**79.93**	**92.87**	**92.48**

注：NT与NS分别代表教师模型与学生模型，黑体数值表示最优分类准确率.

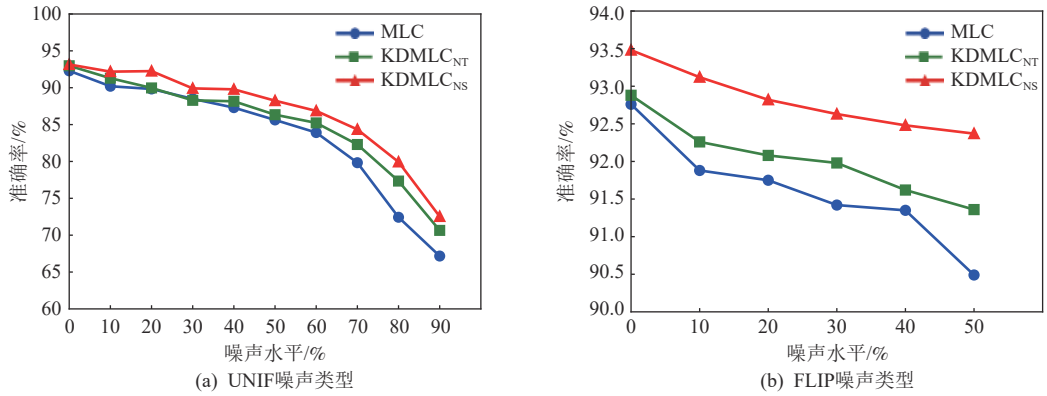
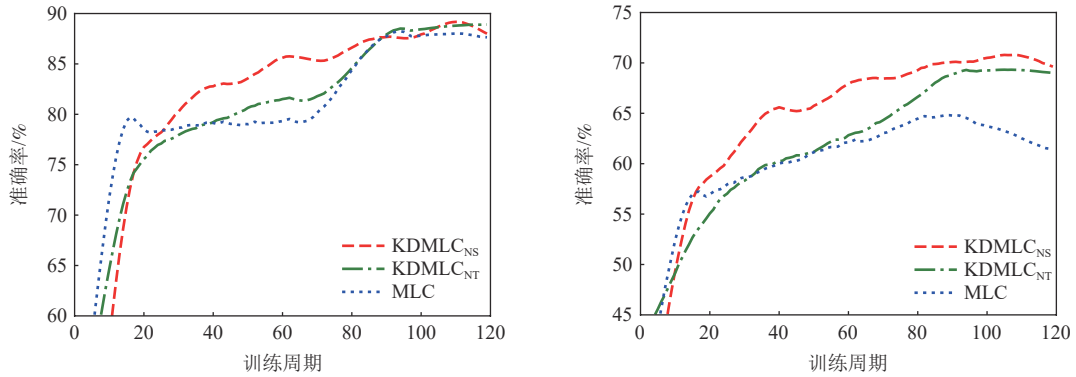


Fig. 4 Accuracy comparison of model on CIFAR10 dataset

图4 模型在CIFAR10数据集下的准确率对比

图 5(a)和图 5(b)分别代表噪声水平为 40% 和 90% 时在 120 次迭代中的准确率对比. 在高噪声水平时, MLC 模型在 80 次迭代后明显出现了过拟合现象, 而采用 KDMLC 方法的模型仍然呈现平稳上升的趋势并未发生过拟合现象, 证明 KDMLC 方法在高噪声水平下拥有更优的鲁棒性与正则化效果.

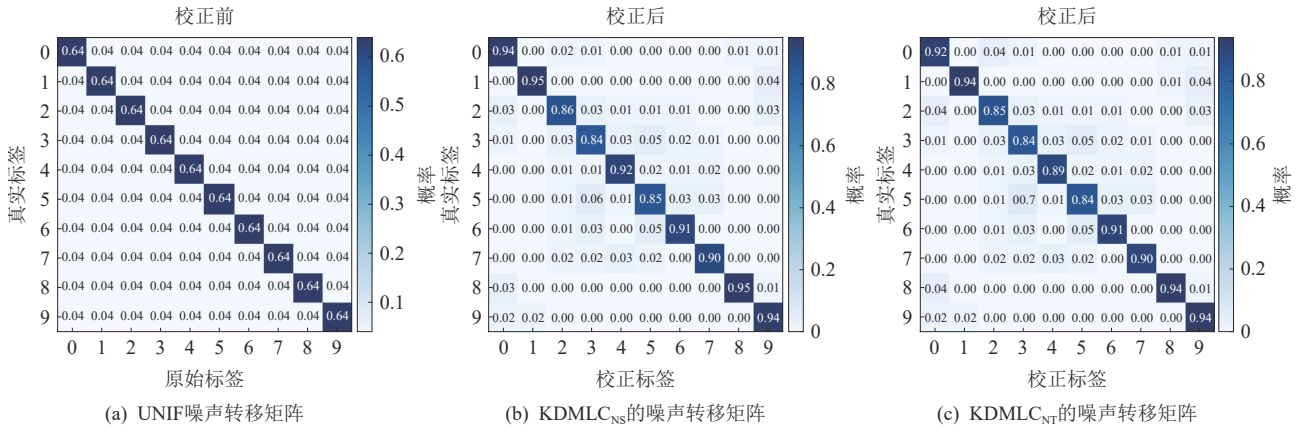
图 6(a)为 CIFAR10 数据集经过 UNIF 噪声处理后, 噪声水平为 40% 的噪声转移矩阵, 图 6(b)(c)为使用 KDMLC 方法修正后的噪声转移矩阵. 通过比较图 6(a)和图 6(b)可以发现, 经过校正后的大多数对角线项的概率超过 90%, 这表明 KDMLC 方法具有较高的校正准确率.



(a) 40%的噪声水平 (b) 90%的噪声水平

Fig. 5 Accuracy comparison of CIFAR10 dataset with UNIF noise type

图 5 在 UNIF 噪声类型下 CIFAR10 数据集的准确率对比



(a) UNIF噪声转移矩阵 (b) KDMLC$_{NS}$的噪声转移矩阵 (c) KDMLC$_{NT}$的噪声转移矩阵

Fig. 6 Noise transition matrix for CIFAR10 dataset under UNIF noise type with a 40% noise level

图 6 在 40% 噪声水平的 UNIF 噪声类型下 CIFAR10 数据集的噪声转移矩阵

3.3.2 CIFAR100 实验结果分析

表 4 显示了 CIFAR100 在 UNIF 与 FLIP 噪声类型的不同噪声水平下, MLC 模型与 KDMLC 模型的实验结果. 可以观察出 MLC 并不能很好地适应 CIFAR100, 尤其是在 UNIF 噪声类型下, 噪声数据被均匀分布到不同类别中, 且在数据量较少的情况下容易增加模型过拟合的概率. 而采用 KDMLC 的学生模型的性能对比 MLC 均有较大的提升, 同时教师模型的性能也得到了优化, 验证了学生模型反馈机制的有效性.在 CIFAR100 中采用学习率 lr=0.1, 蒸馏温度 $temp$=1 时获得的效果较优, 若采用与 CIFAR10 数据集相同的参数设置, 所获得的结果就截然相反. 从图 7(a)(b)中可以发现在 CIFAR100 中蒸馏温度对实验结果具有较大影响, 图 7(a)中 KDMLC 方法明显低于原 MLC

Table 4 Accuracy Comparison of KDMLC and MLC on CIFAR100 Dataset with Two Types of Noise

表 4 KDMLC 与 MLC 在 2 种噪声类型下 CIFAR100 数据集的准确率对比 %

方法	UNIF 噪声的 4 种噪声水平				FLIP 噪声的 2 种噪声水平	
	20%	40%	60%	80%	20%	40%
MLC	58.51	36.43	23.59	16.50	60.19	55.69
KDMLC$_{NT}$	63.26	44.92	28.74	19.65	64.84	63.88
KDMLC$_{NS}$	**67.79**	**50.61**	**31.33**	**19.83**	**68.73**	**67.59**

注：黑体数值表示最优值.

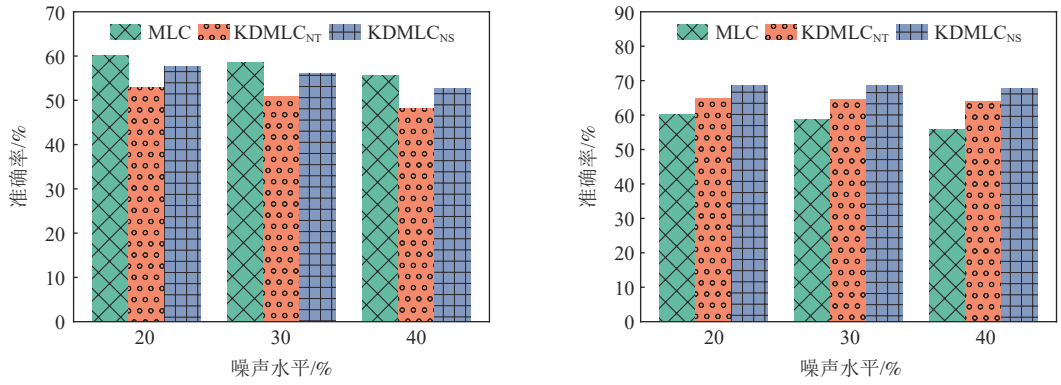


Fig. 7 Comparison of testing accuracy on CIFAR100 dataset under different parameter settings

图 7 在 CIFAR100 数据集的不同参数设置下测试准确率对比

方法的效果，相差近 10 个百分点的准确率. 而在改变学习率与蒸馏温度后，教师模型与学生模型的性能均优于 MLC.

从图 8 可以发现在 80 次迭代时，噪声水平越高，MLC 模型与教师模型之间的差距便越大，证明 KDMLC 方法的确增强了教师模型的性能并提高了其生成伪标签的质量，同时使其在高噪声水平数据集下拥有更强的鲁棒性.

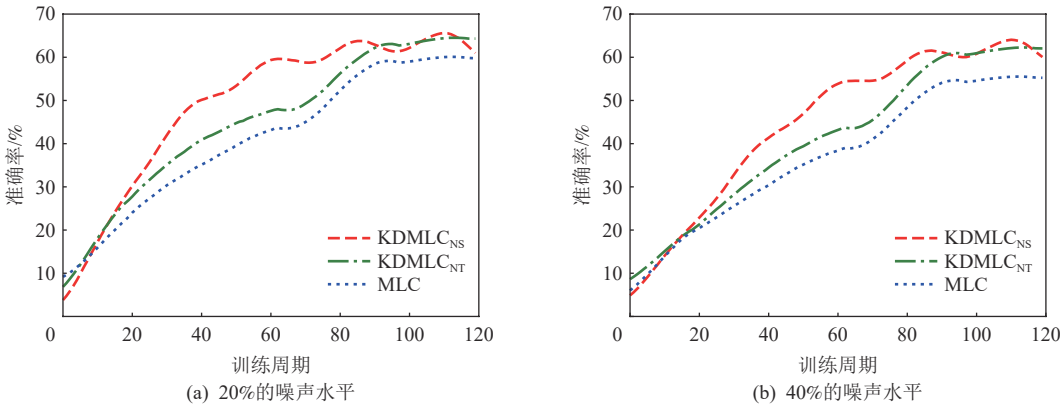


Fig. 8 Accuracy comparison of CIFAR100 dataset with FLIP noise type

图 8 在 FLIP 噪声类型下 CIFAR100 数据集的准确率对比

3.3.3 消融实验与数据增强

在相同的实验环境下，我们将 KDMLC 与 MLC，MW-Net 所需的训练时间进行对比，如图 9 所示，KDMLC 采用在线蒸馏的训练方式，对比其他方法将消耗更多的训练时间.

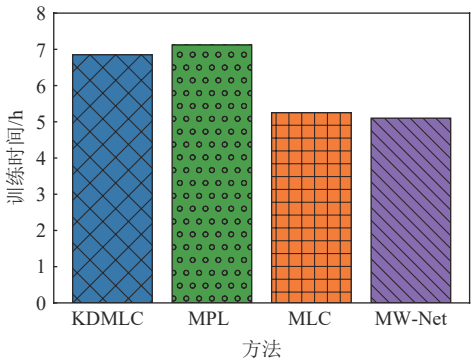


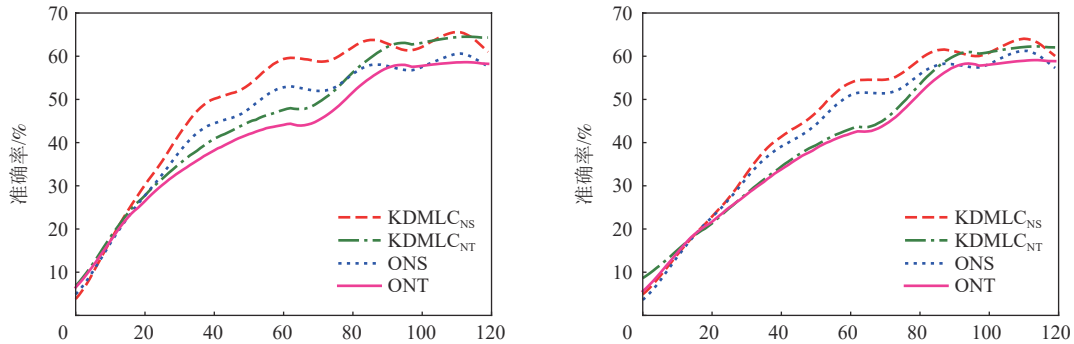
Fig. 9 Comparison of training time

图 9 训练时间对比

图 10 显示了 KDMLC 在软标签(soft labels)与 One-Hot 类型标签下的训练对比，可以观察出，采用 soft labels 训练出的模型性能更优. 其中，ONS 表示采用 One-Hot 标签训练的学生模型，ONT 表示采用 One-Hot 标签训练的教师模型.

表 5 对 KDMLC 进行了系统的探索，分别在数据集 CIFAR10(UNIF)与 CIFAR100(FLIP)下进行实验对比，其中考虑了 2 种策略：1)无学生反馈训练；2)KDMLC 训练.

在此基础下，为了进一步确认 KDMLC 在 CIFAR100 数据集下能否通过更多数据使其具有更强的泛化性能，引入 Cutout 数据增强来解决 CIFAR100 数据量较



(a) 20%的噪声水平 (b) 40%的噪声水平

Fig. 10 Comparison of training accuracy for different types of labels on CIFAR100 dataset with FLIP noise type

图 10 在 FLIP 噪声类型下 CIFAR100 数据集上对不同类型标签训练准确率对比

Table 5 Comparison of Accuracy in KDMLC Ablation Experiments

表 5 KDMLC 消融实验准确率对比 %

训练策略	模型	数据集/噪声类型/噪声水平	
		CIFAR10/UNIF/80%	CIFAR100/FLIP/40%
无学生反馈	KDMLC$_{NT}$	74.51	55.73
	KDMLC$_{NS}$	76.48	59.61
KDMLC 方法	KDMLC$_{NT}$	77.34	63.88
	KDMLC$_{NS}$	**79.93**	**67.59**

注：黑体数值表示最优值.

少的问题，实验设置的裁剪块数为 1，裁剪大小为 8×8. 表 6 中 CNT 与 CNS 分别代表在 Cutout 数据增强下的教师模型与学生模型.

Table 6 Accuracy Comparison of KDMLC and MLC with Two Types of Noise

表 6 KDMLC 与 MLC 在 2 种噪声类型下的准确率对比 %

本文方法	UNIF 噪声的 3 种噪声水平			FLIP 噪声的 3 种噪声水平		
	40%	60%	80%	20%	30%	40%
KDMLC$_{NT}$	88.16	85.13	77.34	64.84	64.37	63.88
KDMLC$_{NS}$	89.78	86.76	79.93	**68.73**	**68.23**	67.59
CNT	89.05	86.37	82.28	64.67	64.18	63.91
CNS	**90.26**	**87.44**	**83.54**	68.59	67.98	**67.76**

注：黑体数值表示最优分类准确率.

从表 6 中可以发现在 UNIF 噪声类型下，CIFAR10 通过 Cutout 数据增强后使模型性能都得到了明显的提升，且随着噪声水平的提高模型性能提升更为明显. 尤其在 80% 噪声水平上，教师模型与学生模型的准确率分别提升了 4.94 个百分点与 3.61 个百分点，表明在数据充足的环境下，KDMLC 方法在高噪声水平下还具备较高的可提升空间. 而在经过 FLIP 噪声类型预处理的 CIFAR100 数据集上，KDMLC 方法在低噪声水平下并未取得较为理想的结果.

在 CIFAR100 数据集下，图 11 显示了在 UNIF 噪声类型下噪声水平 80% 时，可以发现在数据增强后，师生模型的测试精度远高于平均水平，且对比无数据增强的模型本文方法也有较高的提升.

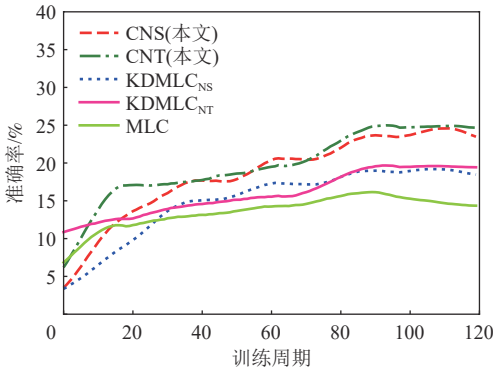


Fig. 11 Comparison of model testing accuracy under Cutout data enhancement

图 11 在 Cutout 数据增强下模型测试准确率对比

3.3.4 Clothing1M 实验结果分析

表 7 显示了真实噪声数据集 Clothing1M 的实验结果，我们采用 Resnet50 网络和 Resnet34 网络分别作为主模型与学生模型，可以看出学生模型在带有真实噪声标签的 Clothing1M 数据集上的准确率均高于

Table 7 Testing Accuracy Comparison of Clothing1M

表 7 Clothing1M 测试准确率对比 %

方法	准确率	方法	准确率
Joint Optimization[33]	72.23	Bootstrap[34]	69.12
U-correction[35]	71.00	Forward[17]	69.84
MW-Net[30]	73.72	MSLC[36]	74.02
GLC[18]	73.69	KDMLC$_{NS}$（本文）	**74.23**

注：黑体数值表示最优值.

其他方法,表明 KDMLC 方法能够在元学习框架下生成质量更好的伪标签.

4 总 结

本文结合元学提出了一种基于在线蒸馏的轻量化噪声标签学习方法.具体来说,通过使用元标签校正模型 MLC 作为教师模型生成的伪标签对学生模型进行训练,缓解了轻量化模型在噪声数据下难以训练的问题,同时我们采用元学习的双层优化框架联系教师模型与学生模型进行联合优化,使得教师模型能够生成更高质量的伪标签.在 CIFAR10 和 CIFAR100 数据集上,通过采用不同的噪声类别与噪声水平进行实验,验证了本文所提方法 KDMLC 的有效性,增强了教师模型的性能,使其生成更高质量的伪标签从而使学生模型得到更好的训练,同时学生模型的性能也超越了教师模型.在未来的工作中,将重点研究 KDMLC 方法的鲁棒性问题,为后续研究提供更好的理论基础.

作者贡献声明: 黄贻望提出了方法思路和实验方案;黄雨鑫负责完成实验并撰写论文;刘声提出指导意见并修改论文.

参 考 文 献

[1] Zhang Chiyuan, Bengio S, Hardt M, et al. Understanding deep learning (still) requires rethinking generalization[J]. Communications of the ACM, 2021, 64(3): 107−115

[2] Nakkiran P, Kaplun G, Bansal Y, et al. Deep double descent: Where bigger models and more data hurt[J]. Journal of Statistical Mechanics: Theory and Experiment, 2021, 2021(12): 124003

[3] Jindal I, Nokleby M, Chen Xuewen. Learning deep networks from noisy labels with dropout regularization[C]//Proc of the 16th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2016: 967−972

[4] Pereyra G, Tucker G, Chorowski J, et al. Regularizing neural networks by penalizing confident output distributions[J]. arXiv preprint, arXiv: 1701.06548, 2017

[5] Ouyang Xiao, Tao Hong, Fan Ruidong, et al. Weakly supervised multi-label learning using prior label correlation information[J]. Journal of Software, 2023, 34(4): 1732−1748 (in Chinese)
(欧阳宵, 陶红, 范瑞东, 等. 利用标签相关性先验的弱监督多标签学习方法[J]. 软件学报, 2023, 34(4): 1732−1748)

[6] Miao Zhuang, Wang Yapeng, Li Yang, et al. Robust Hash learning method based on dual-teacher self-supervised distillation[J]. Computer Science, 2022, 49(10): 159−168 (in Chinese)
(苗壮, 王亚鹏, 李阳, 等. 一种鲁棒的双教师自监督蒸馏哈希学习方法[J]. 计算机科学, 2022, 49(10): 159−168)

[7] Jiang Lu, Zhou Zhengyuan, Leung T, et al. Mentornet: Learning data-driven curriculum for very deep neural networks on labels[C]//Proc of the 35th Int Conf on Machine Learning. New York: ACM, 2018: 2304−2313

[8] Zheng Guoqing, Awadallah A H, Dumais S. Meta label correction for noisy label learning[C]//Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 11053−11061

[9] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//Proc of the 34th Int Conf on Machine Learning. New York: ACM, 2017: 1126−1135

[10] Ji Rongrong, Lin Shaohui, Chao Fei, et al. Deep neural network compression and acceleration: A review[J]. Journal of Computer Research and Development, 2018, 55(9): 1871−1888 (in Chinese)
(纪荣嵘, 林绍辉, 晁飞, 等. 深度神经网络压缩与加速综述[J]. 计算机研究与发展, 2018, 55(9): 1871−1888)

[11] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint, arXiv: 1503.02531, 2015

[12] Yang Zhendong, Li Zhe, Gong Yuan, et al. Rethinking knowledge distillation via cross-entropy[J]. arXiv preprint, arXiv: 2208.10139, 2022

[13] Pham H, Dai Zihang, Xie Qizhe, et al. Meta pseudo labels[C]//Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 11557−11568

[14] Ding Jiaman, Liu Nan, Zhou Shujie, et al. Semi-supervised weak-label classification method by regularization[J]. Chinese Journal of Computer, 2022, 45(1): 69−81 (in Chinese)
(丁家满, 刘楠, 周蜀杰, 等. 基于正则化的半监督弱标签分类方法[J]. 计算机学报, 2022, 45(1): 69−81)

[15] Van Rooyen B, Williamson R C. A Theory of learning with corrupted labels[J]. Journal of Machine Learning Research, 2017, 18(1): 8501−8550

[16] Goldberger J, Ben-Reuven E. Training deep neural-networks using a noise adaptation layer[C/OL]//Proc of the 5th Int Conf on Learning Representations. 2017[2023-01-13].https://openreview.net/forum?id=H12GRgcxg

[17] Patrini G, Rozza A, Krishna Menon A, et al. Making deep neural networks robust to label noise: A loss correction approach[C]//Proc of the 16th IEEE Int Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 1944−1952

[18] Hendrycks D, Mazeika M, Wilson D, et al. Using trusted data to train deep networks on labels corrupted by severe noise[C]//Proc of the 31st Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2018: 10456−10465

[19] Li Yuncheng, Yang Jianchao, Song Yale, et al. Learning from noisy labels with distillation[C]//Proc of the 16th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 1910−1918

[20] Algan G, Ulusoy I. Meta soft label generation for noisy labels[C]//Proc of the 25th IEEE Int Conf on Pattern Recognition. Piscataway, NJ: IEEE, 2021: 7142−7148

[21] Wang Zhen, Hu Guosheng, Hu Qinghua. Training noise-robust deep neural networks via meta-learning[C]//Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ:

IEEE, 2020: 4524−4533

[22] Passban P, Wu Yimeng, Rezagholizadeh M, et al. Alp-KD: Attention-based layer projection for knowledge distillation[C]//Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 13657−13665

[23] Passalis N, Tzelepi M, Tefas A. Heterogeneous knowledge distillation using information flow modeling[C]//Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 2339−2348

[24] Heo B, Lee M, Yun S, et al. Knowledge distillation with adversarial samples supporting decision boundary[C]//Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 3771−3778

[25] Fang Gongfan, Mo Kanya, Wang Xinchao, et al. Up to 100x faster data-free knowledge distillation[C]//Proc of the 36th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022: 6597−6604

[26] Wei Xiushen, Xu Shulin, An Peng, et al. Multi-instance learning with incremental classes[J]. Journal of Computer Research and Development, 2022, 59(8): 1723−1731 (in Chinese)
(魏秀参, 徐书林, 安鹏, 等. 面向增量分类的多示例学习[J]. 计算机研究与发展, 2022, 59(8): 1723−1731)

[27] Chen Pengguang, Liu Shu, Zhao Hengshuang, et al. Distilling knowledge via knowledge review[C]//Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 5008−5017

[28] Xiao Tong, Xia Tian, Yang Yi, et al. Learning from massive noisy labeled data for image classification[C]//Proc of the 28th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 2691−2699

[29] Zhang Zhilu, Sabuncu M. Generalized cross entropy loss for training deep neural networks with noisy labels[C]//Proc of the 31st Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2018: 8778-8788

[30] Shu Jun, Xie Qi, Yi Lixuan, et al. Meta-weight-Net: Learning an explicit mapping for sample weighting[C]//Proc of the 32nd Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2019: 1917−1928

[31] Han Bo, Yao Quanming, Yu Xingrui, et al. Co-teaching: Robust training of deep neural networks with extremely noisy labels[C]//Proc of the 31st Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2018: 8527−8537

[32] Ren Mengye, Zeng Wenyuan, Yang Bin, et al. Learning to reweight examples for robust deep learning[C]//Proc of the 35th Int Conf on Machine Learning. New York: ACM, 2018: 4334−4343

[33] Tanaka D, Ikami D, Yamasaki T, et al. Joint optimization framework for learning with noisy labels[C]//Proc of the 31st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 5552−5560

[34] Reed S, Lee H, Anguelov D, et al. Training deep neural networks on noisy labels with bootstrapping[J]. arXiv preprint, arXiv: 1412.6596, 2014

[35] Arazo E, Ortego D, Albert P, et al. Unsupervised label noise modeling and loss correction[C]//Proc of the 36th Int Conf on Machine Learning. New York: ACM, 2019: 312−321

[36] Wu Yichen, Shu Jun, Xie Qi, et al. Learning to purify noisy labels via meta soft label corrector[C]//Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 10388−10396

Huang Yiwang, born in 1978. PhD, professor, Master supervisor. Member of CCF. His main research interests include artificial intelligence, business process management, and software formalization method.

黄贻望, 1978 年生. 博士, 教授, 硕士生导师, CCF 会员. 主要研究方向为人工智能、业务过程管理、软件形式化方法.

Huang Yuxin, born in 1998. Master candidate. Member of CCF. His research interests include computer vision and label noise learning.

黄雨鑫, 1998 年生. 硕士研究生. CCF 会员. 主要研究方向为计算机视觉、标签噪声处理.

Liu Sheng, born in 1984. PhD, professor. His research interest includes information processing.

刘　声, 1984 年生. 博士, 教授. 主要研究方向为信息处理.