APAM：面向噪声标签与长尾学习的语言模型自适应预训练与自适应元学习

Sunyi Chi
sunyi.chi@uth.tmc.edu

Bo Dong
dongbd@amazon.com

Yiming Xuy
mxu@亚马逊.com

Zhenyu Shi
szhenyu@amazon.com

Zheng Du
zhengdu@amazon.com

arXiv:2302.03488v2 [cs.CL] 2 May 2023

摘要

实际应用中的自然语言处理（NLP）任务通常呈现长尾分布，并伴有噪声标签。这些问题对深度神经网络（DNNs）等复杂模型的泛化能力和鲁棒性构成挑战。一些常用的重采样技术，如过采样或欠采样，很容易导致过拟合。利用少量元数据来学习数据权重正日益受到关注。此外，近期研究表明，自监督预训练具有优势，尤其有助于代表性不足的数据。在这项工作中，我们提出了一个通用框架，用于处理长尾和噪声标签问题。该模型通过对比学习的方式适应问题领域。重新加权模块是一个前馈网络，它能够学习显式的加权函数，并根据元数据调整权重。该框架还通过结合交叉熵损失的多项式展开与焦点损失，进一步调整损失函数中各项的权重。大量实验表明，所提出的框架始终优于基线方法。最后，我们的敏感性分析强调了所提出框架处理长尾问题并减轻噪声标签负面影响的能力。

1 引言

深度神经网络（DNNs）已成为利用大规模标注数据解决复杂问题的默认建模选择。它们在自然语言处理和计算机视觉等各种领域的监督学习中取得了显著成功。这些成功依赖于大量高质量的标注数据。在实践中，大规模获取干净标签通常代价高昂。这要么需要多轮盲审轮次和裁决者的决定，要么需要审计员进行质量保证。两者都耗费人力且耗时。近年来，微调（Devlin 等人，2019年；Cui 等人，2018年）、领域自适应（Tzeng 等人，2017年；Xu 等人，2020年；Ganin 和 Lempitsky，2015年；Xu 等人，2021年）和少样本学习（Brown 等人，2020年）的进展缓解了对大量标注数据的需求。尽管取得了这些进展，深度学习模型仍然依赖准确的标注数据。

另一个会损害模型泛化能力的因素是数据分布偏移。数据偏移可能由多种来源造成。长尾问题是一个常见的例子。例如，为了训练一个对购物商品进行分类的模型，很难获得稀有商品的充足图像（He 和 McAuley，2016年）。训练一个能够接触到足够多低频话题或用户意图的对话模型同样是一项挑战。在不平衡数据集上训练模型时，模型训练会偏向多数类。由于可供学习的样本数量更多，模型能够学会在多数类上表现良好；但由于缺乏足够的样本，模型无法学到有意义的模式，而这些模式本可以帮助其学习少数类。模型在这些尾部类别上的表现成为深度神经网络在实际应用中的瓶颈，因此，改善这类情况至关重要。

许多研究提出了缓解噪声标签或长尾类别问题的方法。为减轻噪声标签的影响，研究者探讨了样本选择（Jiang等人，2018年）、标签校正（Patrini等人，2017年；Sanchez等人，2019年）以及噪声感知损失（Liu等人，2022年；Castells等人，2020年）。数据集重采样（如SMOTE（Chawla等人，2002））是一种常见方法：选取一定比例的数据来训练网络，或为每个样本学习一个权重。这些权重通过最小化训练损失进行优化。该方法应用于多种知名算法，例如AdaBoost （ Freund 和 Schapire，1997 ）、自步学习（Kumar等人，2010年），

以及强调高方差样本的算法（Jiang等人，2018年；Chang等人，2017年）。为解决长尾分布问题，一些研究修改了采样算法，以确保所有类别都得到均衡表示（Kubát和Matwin，1997；Chawla等人，2002）。其他常见方法包括调整损失函数（Menon等人，2021年），使其偏向少数类，以及事后校正（Kang等人，2020年）。

然而，这些方法的假设彼此矛盾。一方面，为了缓解噪声标签问题，我们会提高干净标注数据的权重。另一方面，针对长尾问题的算法会强调少数类别，因为这些类别更有可能具有较高的训练损失。因此，这些方法无法处理噪声标签与长尾类别并存的问题。为了同时处理噪声标签和长尾问题，一些研究（Shu 等人，2021年；Ren 等人，2018年）提出了一种元学习范式，其遵循一种更为自然的假设：最佳样本加权应使干净数据的损失最小化。这些方法从少量干净数据中学习实例权重，并取得了令人鼓舞的结果。尽管如此，我们认为元学习范式的通用性不足，而且由于需要平衡的元验证数据集，成本仍然很高。

在我们的工作中，我们提出了自适应预训练与自适应元学习方法（APAM），这是一个用于共同处理噪声标签和长尾类别并存问题的通用框架。该方法自然地将上述元学习范式（Shu 等人，2021年；Ren 等人，2018年）纳为特例。我们的方法不需要平衡的元数据来指导重新加权。这一特点减少了元学习所需的干净数据量，因此更加可行，成本也更低。

此外，我们通过对比学习Gururangan 等人，2020年引入了一个领域自适应预训练阶段，以提高模型的鲁棒性。所提出的自适应过程会降低噪声样本的权重；另一方面，APAM 会提高权重，以同时应对噪声和长尾问题。Gao 等人，2021年

我们进一步在两个数据集上评估 APAM，结果表明它优于其他方法（Shu 等人，2021年；Lin 等人，2017年；Cui 等人，2019年；Devlin 等人，2019年；Gao 等人，2021年）。此外，我们还进行了全面的消融研究和敏感性分析。在这些实验中，我们观察到 APAM 始终优于其他方法。

总而言之，我们工作的主要贡献如下：

•为应对文本分类中长尾和噪声标签并存的问题，我们提出了一种通用的两阶段深度学习框架，包括领域自适应预训练阶段和通过自适应重新加权进行监督微调的阶段。该框架在评估数据集上的表现优于最先进的方法。

• 我们证明，所提出的自适应加权方法不需要平衡的元数据，因此能够减轻长尾问题对大量元数据的依赖。

•实验结果表明，领域自适应对比学习能够持续提升 APAM框架的性能。

• 通过全面的消融研究和敏感性分析，我们展示了所提出方法中各组成部分的贡献，以及该方法在带有噪声标签的长尾问题上的有效性。

2 相关工作

2.1 噪声标签学习

深度学习近期的成功在很大程度上依赖于大量带有高质量标签的数据。然而，获取准确标签既昂贵又费力。一些众包平台（如亚马逊 Mechanical Turk）已被广泛用于降低标注成本。不过，这些方案往往会导致标签损坏（Paolacci 等人，2010年；Mason 和 Suri，2012年）。Song 等人（Song 等人，2022年）总结指出，现实世界中标签的噪声比例通常在 8.0% 至 38.5% 之间（Song 等人，2019 年；Xiao 等人，2015年；Lee 等人，2017年）。存在损坏标签时，深度学习模型往往会对噪声标签过拟合。Zhang 等人（Zhang 等人，2021a）通过实验表明，深度神经网络很容易拟合噪声比例各异的训练数据集，从而导致模型泛化能力较差。

许多研究探索了噪声标签学习，可将相关方法分为四类：正则化、设计损失函数、采样方法，以及修改模型架构以抵御噪声数据。

具体而言，正则化方法得到了广泛采用，包括 Dropout（Srivastava 等人，2014年）、权重衰减（Krogh 和 Hertz，1991年）和批归一化（Ioffe 和 Szegedy，2015年），以及标签平滑（Pereyra 等人，2017年；Lukasik 等人，2022年）、mixup（Zhang 等人，2018年）和对抗训练（Goodfellow 等人，2015年）等隐式正则化方法。

许多研究提出了鲁棒损失（Zhang 和 Sabuncu，2018年；Ghosh 等人，2017年；Wang 等人，2019年）或损失调整（Reed 等人，2015年；Song 等人，2019年；Patrini 等，2017；Wu 等人，2023a；Sanchez 等，2019）方法，以缓解噪声标签问题。元学习（Ren 等人，2018年；Shu 等，2021；Finn 等人，2017年；Dong 等人，2022年）是一类自动化的权重调整或标签校正方法，通过学习如何重新加权或校正标签来实现这一目标。采样方法（Song 等人，2019年；Han 等人，2018年；Yu 等人，2019年）侧重于从含噪训练数据集中选取真实标签样本。此外，能够抵抗噪声标签的模型架构（Sukhbaatar 等人，2014年；Goldberger 和 Ben-Reuven，2017年）也得到了广泛应用，例如噪声适应层。

2.2 长尾学习

现实世界的应用通常属于长尾问题，其中一些类别关联的数据点很少（Cui 等人，2019年；Menon 等人，2021年；Liu 等人，2019年）。长尾问题与类别不平衡问题有两点不同。首先，尽管少数类的规模相对多数类较小，但仍可能拥有大量样本。其次，长尾问题包含大量类别，尾部类别样本通常非常稀少。相比之下，类别不平衡问题中的类别数量可能很少。因此，长尾问题可视为一种更具体且更具挑战性的类别不平衡问题（Zhang 等人，2021b）。

近年来，人们提出了大量深度长尾学习方法（Cui 等人，2019年；Liu 等人，2019年）以解决长尾问题。这些方法可分为三类（Zhang 等人，2021b），分别是信息增强、类别再平衡和模型改进。基于信息增强的方法旨在通过向模型训练中引入额外信息来改善长尾学习。常见方法包括迁移学习（Wang 等，2017年；Tan 等，2018年）和数据增强（Shorten 和 Khoshgoftaar，2019年）。类别再平衡方法包括重采样（Chawla 等，2002；Kang 等，2020年；Huang 等，2021年）和损失调整方法（Leng 等人，2022年；Lin 等，2017；Sun 等，2007）。最后，改进模型也被证明对许多长尾问题很有帮助。常见方法包括集成学习（Li 等，2020年）和表征学习（Wang 等，2021年；Wu 等人，2023b；Dong 等，2020年；Wang 等，2020年）。

2.3 领域自适应预训练

近年来，来自各种语料库的预训练语言模型重塑了自然语言处理。Devlin 等人（Devlin 等人，2019年）提出了自监督学习（SSL），这是一种无监督学习方法，让模型通过解决辅助任务来学习数据表示，而无需标签。这些预训练任务可以提升模型的泛化能力，并防止模型在数量有限的类别标签上过拟合（Dong 等人，2022年）。

此外，许多研究（Gururangan 等人，2020年；Zhou 等人，2021年）已证明，通过第二阶段的领域内预训练，使预训练模型适应目标任务所在领域是有益的。Gururangan 等人（Gururangan 等人，2020年）研究了四个不同出版领域中的八项分类任务，这些领域包括生物医学、计算机科学和新闻，并表明领域自适应预训练（DAPT）能够提升性能。Zhang 等人（Zhang 等人，2022年）则致力于使预训练语言模型适用于长尾问题。

对比学习（Hadsell 等人，2006年）已广泛应用于预训练任务。其核心思想是通过拉近语义相似的邻居并推远非邻居来学习表征。研究者（Gao 等人，2021年；Wu 等人，2020年）提出通过对比学习来学习句子嵌入。近期研究表明，在无监督任务上大规模进行对比预训练，能够有效地学习文本和代码的表征（Neelakantan 等人，2022年）。

3 方法论

受近期对比学习和元学习研究的启发，我们提出了 APAM，这是一种通用的两阶段框架，用于解决噪声数据环境下 NLP 长尾问题。我们假设该问题中只有少量干净数据和大量带噪标注数据。

APAM 首先通过自监督式对比学习进行领域自适应预训练。该自监督学习任务通过对丢弃层随机应用掩码，学习句子对的语义相近特征表示，其中正样本对由同一源句子生成。

在第二阶段，我们提出自适应权重策略，利用少量干净数据来指导编码器参数的自适应调整。在迭代过程中，带噪训练样本将在前向传播中获得权重，并在后续迭代中利用这些权重。此过程可以缓解长尾性和噪声标签问题。所提出的框架如图1所示。下面我们将详细介绍上述两个主要组成部分。

3.1 领域自适应预训练

我们框架的第一阶段使用自监督学习对基础模型进行预训练，从而舍弃噪声且不平衡的标签。在预训练阶段，我们采用简单的句子嵌入对比学习方法（simCSE）（Gao 等，2021）来学习领域特定表示。我们通过对整组未标注句子 $\{x_i\}$进行自适应训练，将基础编码器微调至特定任务领域。在此，我们使用相同的句子对（x_i, x_i^+）作为正样本对，其中 $x_i^+ = x_i$通过独立采样的丢弃掩码生成。具体而言，我们将相同的输入两次馈入编码器 F（x, z），并使用不同的丢弃掩码 z 和 z'获得两个嵌入，其中 z 表示用于 Dropout 的随机掩码。我们将相同句子对（x_i, x_i^+）的特征表示记为（s_i, s_i^+），其中 $s_i = F$（x, z）且 $s_i^+ = F$（x^+, z'）。批次内其他不同句子被视为负样本对。对比损失如公式 1 所示。

$$l_i = -log\frac{e^{sim(s_i,s_i^+)/\tau}}{\sum_{j=1}^{N} e^{sim(s_i,s_j^+)/\tau}} \quad (1)$$

令 N 表示一个小批次中的句子数量。 τ 是一个固定的温度参数至 $\tau = 0.05$ 遵循先前的研究（Gao 等人，2021年）； sim（s_1， s_2）是余弦相似度$\frac{s_1^T s_2}{||s_1||||s_2||}$。在此阶段，我们利用特定领域的大量未标注文本，获得一个经过领域自适应预训练的编码器。

3.2 自适应元学习

在第二阶段，我们利用少量干净数据集进行自适应元学习，迭代微调基础模型。受 PolyLoss（Leng 等人，2022年）的启发，我们在交叉熵损失上使用自适应多项式项，如公式2

$$l(W) = -log(P_c) + \epsilon(1 - P_c) \quad (2)$$

其中， ϵ 是多项式项的权重，在我们的实验中取值为 1； P_c 表示目标类别的预测概率； W 表示由模型参数和损失函数中的自适应项组成的参数集。

算法1自适应元学习算法

要求： 训练数据 D_t， 元数据 D_m。
输入： 批大小 n、m，最大迭代次数 T。
输出： W_T

初始化 θ_0, W_0
对于 $t = 0, \ldots, T\text{-}1$ **执行**
  $\{x, y\} \leftarrow$ 采样小批次(D_t, n)
  $\{x^{meta}, y^{meta}\} \leftarrow$采样小批次$(D_m, m)$
  $\hat{}y \leftarrow$前向传播(x, y, W_t)
  $\nabla W_t \leftarrow$反向传播$(W_t(\theta))$
  $\hat{}W_t(\theta) \leftarrow W_t - \alpha\nabla W_t$
  $\hat{}y^{meta} \leftarrow$前向传播$(x^{meta}, y^{meta}, \hat{}W_t(\theta))$
  $\nabla\theta \leftarrow$反向传播(θ)
  更新 $\hat{\theta}$
  $\hat{}\nabla W_t \leftarrow$反向传播$(W_t(\theta))$
  $W_{t+1} \leftarrow$ 优化器步骤$(W_t, \nabla W_t)$
</style}结束循环

受一些现有工作的启发（Ren 等人，2018年；Shu 等人，2021年），我们使用一个仅含一个隐藏层的多层感知机网络，该隐藏层有 100 个节点，以损失 L^{train}（p）为输入，自动学习实例权重。该模型需要进行两层嵌套优化循环，如算法1所示。具体来说，在每次训练迭代中，我们通过沿目标损失的下降方向移动当前 W，并使用初始化的元模型参数 W，来更新训练模型参数 θ，其中 α 是步长。获得更新后的 W（θ）后，元模型的参数 θ

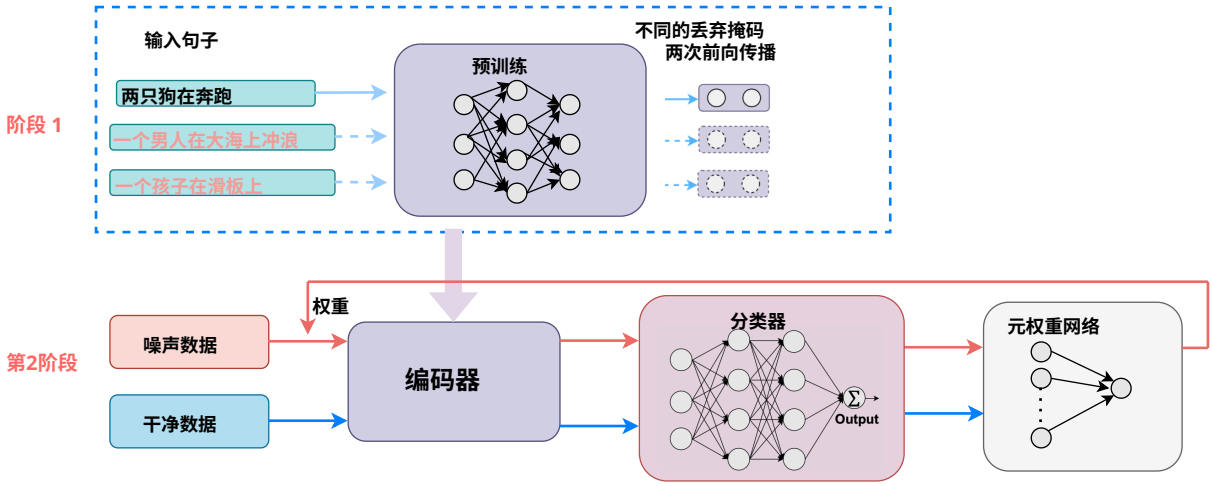

图1：APAM框架概览，其中 θ 表示元模型的参数， W 表示主模型的参数。

随后，沿着在元数据上计算得到的目标梯度移动当前参数，从而更新这些参数。然后，利用更新后的 θ 通过传播来更新参数 W（θ）。

4 实验

为评估 APAM 的性能，我们在两个真实世界数据集上的多项分类任务中开展了实验。我们还将 APAM 与长尾学习和噪声标签学习领域的最先进方法进行了比较。

4.1 数据集

Amazon注释数据集 该 Amazon注释数据集取样自 Alexa 实时流量，是一个企业规模的口语语言数据集，记录了人与对话式人工智能代理之间的对话交互。请注意，我们已对数据进行处理，确保用户无法被识别。该数据集包含话语的自动语音识别 (ASR) 文本，并由人工标注员进行标注。标注员聆听音频，并从预定义的类别分类体系中选择标签。此任务有 30 多个不同类别。更严格地说，我们将数据集的不平衡因子定义为最大类别的训练样本数量除以最小类别的训练样本数量。该数据集的不平衡因子为 110。

此外，数据中包含 6% 的噪声标签，这些标签来自一个标注流程，其中只有一名标注员对话语所属领域进行标注。相比之下，我们从另一条数据流程中获得干净标签，该流程由三名标注员根据多数投票结果提供标签；如果标注员无法达成一致，则由额外的裁决员作出裁定。

亚马逊评论数据集 除了企业级的 Amazon 注释数据集外，我们还在数据集 Amazon 评论数据集上评估 APAM[1]，，该数据集是规模最大的产品评论数据集之一。它广泛用于不平衡分类基准方法。我们将 APAM 与多种基线方法进行比较，例如 BERT（Devlin 等人，2019年）、MWNet（Shu 等人，2021年）等，并通过改变噪声率进行敏感性分析。该数据集包含跨越 19 年的 Amazon产品评论及元数据，共有 29 个产品类别中的 1.428 亿条评论，例如书籍、电子产品、美容等。评论包括文本评论、评分和投票；产品元数据则包括描述、类别信息、价格和品牌。最大类别的评论数量超过少数类的 200 倍。我们随机抽取了一个数据子集，其中所有用户和商品都至少有 5 条评论。不平衡因子为 51.3。表 1 展示了亚马逊评论数据集的 2 个头部类别和 2 个尾部类别样本及其对应文本。

4.2 实验设置

合成噪声比例在 Amazon 评论数据集中，我们从整个训练集中抽取 1% 作为干净数据集。噪声集通过按照均匀标签噪声破坏其余数据点的标签生成。我们生成了噪声比率 $\rho = (10\%, 20\%, 30\%, 40\%, 50\%, 60\%)$。对于一个

[1]https://jmcauley.ucsd.edu/data/amazon/

表 1：亚马逊评论数据集的样本类别和文本

样本类别	样本文本
书籍	喜欢这个故事
服装、鞋履和珠宝	舒适又华丽
庭院、草坪和花园	看起来很漂亮，而且和我们的后院完美契合
乐器	做工非常精良，是买来送礼的

表 2：模型在 Amazon Annotation 数据集上的性能

(%)	准确率	Δ准确率	精确率	召回率	F1 分数
BERT + 交叉熵损失	83.33	+ 0	83.13	83.33	82.79
BERT + 焦点损失	83.66	+ 0.33	83.38	83.66	82.90
BERT + 类别平衡损失	83.97	+ 0.64	83.45	83.97	83.05
BERT + MWN	84.07	+ 0.74	83.93	84.07	83.44
BERT + simCSE	84.08	+ 0.75	83.90	84.08	83.25
BERT + AML	84.22	+ 0.89	83.90	84.22	83.50
BERT + DAPT-SSL	84.60	+ 1.27	84.39	84.60	83.78
BERT + APAM	**85.38**	+ 2.05	**84.77**	**85.38**	**84.56**

包含 C 个类别的数据集，真实标签为 y 的干净样本会以概率 ρ/C 被随机改为所有可能类别中的任意一个。请注意，标签也有 ρ/C 的概率保持真实，因此被破坏后的标签也可能与原始标签相同。

模型架构与训练 在整个实验过程中，我们均使用未区分大小写的 Bert-base 作为编码器骨干网络。在预训练 BERT-base 的顶部添加了两个全连接层，分别包含 256 和 128 个隐藏节点。训练批大小设为 128，初始学习率为 $1e^{-5}$。我们使用 PyTorch 实现所有模型和实验。所有模型的训练轮次相同。

基线方法 我们将 APAM 与 BERT 和元权重网络（MWN）进行比较，将其作为基线方法。MWN 是一种元学习范式，其遵循一个自然假设：最佳的样本加权方式应使干净数据的损失最小化。该方法从一个小型干净数据集中学习实例权重，并在实践中表现出色。此外，我们还进行了全面的消融研究，以证明领域自适应预训练和自适应元学习的积极作用。

4.3 主要结果

我们使用准确率、精确率、召回率和 F1 分数作为性能评估指标。表格 2 展示了以预训练 BERT-base 作为主分类器的 APAM 的准确率、精确率、召回率和 F1 分数，并将其与 Amazon Annotation 数据集上其他最先进的方法进行比较。该数据集具有真实世界中类别不平衡和噪声标签的特点。采用交叉熵损失的基线模型 BERT-base 达到 83.33% 的准确率、83.13% 的精确率、83.33% 的召回率和 82.79% 的 F1 分数。APAM 达到 85.38% 的准确率、84.77% 的精确率、85.38% 的召回率和 84.56% 的 F1 分数。与基线模型相比，APAM 利用噪声且不平衡的 Amazon Annotation 数据集中少量干净数据集进行自适应实例权重学习，使准确率提高了 2.05%。

此外，表 3 展示了 APAM 相对于人工标注员标注结果的预测性能。我们观察到，APAM 的表现与人工标注员相当。值得注意的是，APAM 在某些尾部类别上优于人工标注，例如播客和导航类别。这些观察结果表明，标签自动化具有潜力，尤其是在代表性不足的类别上。

4.4 消融研究

为了评估 APAM框架各组件的贡献，我们进行了全面的消融研究，以了解每个组件的实际效果，并量化 APAM框架带来的收益。结果见表 2。在自监督学习中进行领域自适应预训练（DAPT-SSL），即在基线模型的预训练基础上，额外使用未标注文本进行预训练，使准确率较基线提高了 1.27%。自适应元学习利用小型干净数据集对每个实例进行加权，

表3：APAM 与人工标注员的比较。

		APAM			人工		
		精确率	召回率	F1	精确率	召回率	F1
主要类别	实体购物	91.32	98.34	94.70	93.54	99.24	96.30
	名人声音	90.88	82.52	86.50	97.36	99.01	98.18
	命令与控制	88.73	93.91	91.25	94.59	96.19	95.39
	音乐	93.27	96.52	94.87	95.87	97.60	96.73
未成年人类别	播客	**90.24**	**97.37**	**93.67**	89.47	89.47	89.47
	食谱	76.67	**95.83**	85.19	100.00	91.67	95.65
	图库	**94.44**	**94.44**	**94.44**	94.44	94.44	94.44
	导航	85.71	**100.00**	**92.31**	100.00	83.33	90.91

表4：亚马逊评论数据集上的敏感性分析结果

（噪声比率 %）	10	20	30	40	50	60	平均值
BERT + 交叉熵损失	61.37	59.57	57.09	55.04	52.77	22.58	51.40
BERT + MWN	61.53	61.13	59.91	57.60	55.70	53.18	58.18
BERT + APAM	**65.70**	**65.15**	**64.15**	**63.18**	**62.70**	**61.63**	**63.75**

准确率达到 84.22%。APAM结合了 DAPT-SSL 和 AML 的优势，在 Amazon Annotation 数据集上使准确率较基线模型提高了 2.05%，这证明了 APAM 的各个组件在具有噪声标签的真实世界长尾数据集中均有效。

4.5 敏感性分析

表 4 展示了在不平衡因子为 51、UNIF 噪声水平分别为 10%、20%、30%、40%、50% 和 60% 时，使用基准数据集 Amazon 评论进行的敏感性分析结果。显然，由于基线方法无法处理噪声和不平衡标签，随着噪声水平升高，其性能显著下降；当噪声水平达到 60% 时，性能更是极差。相比之下，基于重新加权的方法（MWN 和 APAM）在长尾数据上能够抵御严重的标签噪声。这与（Shu 等人，2021 年）报告的图像数据结果一致：重新加权方法在噪声率变化或真实世界的不平衡率下都表现良好。此外，我们观察到，对于存在噪声和不平衡问题的文本数据，APAM 比以往的重新加权方法更加有效。噪声问题越严重，APAM 相较于基线方法和其他重新加权方法的优势就越明显。

5 结论

本文从自适应权重学习和元学习的角度，探讨如何利用噪声标签和长尾文本数据进行学习。我们提出了一种新颖的框架，可对未标注的领域特定文本进行自适应预训练，并提取样本权重，从而确保在训练数据偏差存在的情况下实现稳健的深度学习。实证结果表明，在真实世界数据集和合成数据集上，所提出的框架 APAM 在长尾学习和噪声学习领域均优于其他最先进的方法。所提出的自适应框架 APAM 能够抵御不同程度的噪声和不平衡。此外，它还兼容其他基线架构。

参考文献

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, JaredDKaplan, PrafullaDhariwal, Arvind Neelakantan 等。2020年。语言模型是少样本学习者。载于《神经信息处理系统进展》第 33 卷，第 1877–1901 页。Curran Associates, Inc.

Thibault Castells、Philippe Weinzaepfel 和 Jerome Revaud。2020年。Superloss：一种适用于鲁棒课程学习的通用损失函数。载于《神经信息处理系统进展》第 33 卷，第 4308–4319 页。Curran Associates, Inc.

Haw-Shiuan Chang、Erik Learned-Miller 和 Andrew McCallum。2017年。 Active bias：训练更准确的——

通过强调高方差样本来衡量神经网络。

N. V. Chawla、K. W. Bowyer、L. O. Hall 和 W. P. Kegelmeyer。2002年。SMOTE：合成少数类过采样技术。《人工智能研究期刊》，16：321–357。

Yin Cui、Menglin Jia、Tsung-Yi Lin、Yang Song 和 Serge Belongie。2019年。基于样本有效数量的类别平衡损失。载于《2019年 IEEE/CVF 计算机视觉与模式识别会议》，第 9260–9269 页。

Yin Cui、Yang Song、Chen Sun、Andrew Howard 和 Serge J. Belongie。2018年。大规模细粒度分类与领域特定迁移学习。载于《2018年 IEEE 计算机视觉与模式识别会议》，2018年，地点：美国犹他州盐湖城，2018年6月18日至22日，第 4109–4118 页。计算机视觉基金会 / IEEE计算机学会。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：用于语言理解的深度双向Transformer预训练。载于《2019年北美计算语言学协会分会人类语言技术会议论文集》第1卷，第 4171–4186 页，美国明尼苏达州明尼阿波利斯。计算语言学协会。

Bo Dong、Cristian Lumezanu、Yuncong Chen、Dongjin Song、Takehiko Mizoguchi、Haifeng Chen 和 Lati- fur Khan。2020年。音速：高效音频场景分类。载于《2020年多媒体检索国际会议论文集》（ICMR '20），第301–305页，美国纽约州纽约市。计算机协会。Bo Dong、Yiyi Wang、Hanbo Sun、Yunji Wang、Alireza Hashemi 和 Zheng Du。2022年。CML：一种用于估计人工标签置信度分数并降低数据收集成本的对比元学习方法。载于第五届电子商务与自然语言处理研讨会论文集，第35–43页，爱尔兰都柏林。计算语言学协会。Chelsea Finn、Pieter Abbeel 和 Sergey Levine。2017年。用于深度网络快速适应的模型无关元学习。CoRR，abs/1703.03400。Yoav Freund 和 Robert E Schapire。1997年。在线学习的决策理论泛化及其在提升中的应用。《计算机与系统科学杂志》，55(1):119–139。Yaroslav Ganin 和 Victor Lempitsky。2015年。通过反向传播实现无监督领域自适应。载于《第32届国际会议机器学习国际会议论文集》第37卷，第1180–1189页。JMLR.org。Tianyu Gao、Xingcheng Yao 和 Danqi Chen。2021年。SimCSE：句子嵌入的简单对比学习。载于2021年自然语言处理实证方法会议，页码6894–6910，线上举办，地点为多米尼加共和国蓬塔卡纳。计算语言学协会。

Aritra Ghosh、Himanshu Kumar 和 P. S. Sastry。2017年。针对深度神经网络中标签噪声的鲁棒损失函数。载于《第三十一届 AAAI 人工智能会议论文集》，第 1919–1925 页。AAAI出版社。

Jacob Goldberger 和 Ehud Ben-Reuven。2017年。使用噪声适应层训练深度神经网络。载于国际学习表征会议。

Ian J. Goodfellow、Jonathon Shlens 和 Christian Szegedy。2015年。解释并利用对抗样本。载于第三届国际学习表征会议，2015年，美国加利福尼亚州圣迭戈，2015年5月7日至9日，会议论文集。

Suchin Gururangan、Ana Marasovi´c、Swabha Swayamdipta、Kyle Lo、Iz Beltagy、Doug Downey 和 Noah A. Smith。2020年。预训练不应止步：使语言模型适应领域和任务。载于《第58届计算语言学协会年会论文集》，第8342–8360 页，线上举行。计算语言学协会。

R. Hadsell、S. Chopra 和 Y. LeCun。2006年。通过学习不变映射进行降维。载于 2006年 IEEE计算机学会计算机视觉与模式识别会议，第2卷，第1735–1742页。

Bo Han、Quanming Yao、Xingrui Yu、Gang Niu、Miao Xu、Weihua Hu、Ivor Wai-Hung Tsang 和 Masashi Sugiyama。2018年。《协同教学：利用极度噪声的标签训练鲁棒的深度神经网络》。载于神经信息处理系统大会。Ruining He 和 Julian McAuley。2016年。起伏变迁：利用单类协同过滤对时尚趋势的视觉演变进行建模。载于第25届万维网国际会议论文集，第507–517页，瑞士日内瓦共和国及州。国际万维网会议指导委员会。Yi Huang、Buse Giledereli、Abdullatif Köksal、Arzu- can Özgür 和 Elif Ozkirimli。2021年。针对长尾类别分布的多标签文本分类平衡方法。载于2021年自然语言处理实证方法会议论文集，第8153–8161页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。Sergey Ioffe 和 Christian Szegedy。2015年。批量归一化：通过减少内部协变量偏移加速深度网络训练。载于第32届国际会议

机器学习会议——第37卷，第448–456页。JMLR.org。

Lu Jiang、Zhengyuan Zhou、Thomas Leung、Li-Jia Li 和 Li Fei-Fei。2018年。MentorNet：针对损坏标签，为极深神经网络学习数据驱动课程。载于第35届国际机器学习会议论文集，《机器学习研究论文集》第80卷，第2304–2313页。PMLR。

Bingyi Kang、Saining Xie、Marcus Rohrbach、Zhicheng Yan、Albert Gordo、Jiashi Feng 和 Yannis Kalantidis。2020年。解耦长尾识别中的表征与分类器。ArXiv，abs/1910.09217。

Anders Krogh 和 John Hertz。1991年。简单的权重衰减即可改善泛化。载于《神经信息处理系统进展》第4卷。Morgan-Kaufmann。

Miroslav Kubát 和 Stan Matwin。1997年。应对不平衡训练集的诅咒：单边选择。载于 ICML。

M. Pawan Kumar、Benjamin Packer 和 Daphne Koller。2010年。面向潜变量模型的自步学习。载于第23届神经信息处理系统国际会议论文集——第1卷，NIPS'10，第1189–1197页，美国纽约州雷德胡克。Curran Associates Inc.

Kuang-Huei Lee、Xiaodong He、Lei Zhang 和 Linjun Yang。2017年。CleanNet：面向标签噪声下可扩展图像分类器训练的迁移学习。

Zhaoqi Leng、Mingxing Tan、Chenxi Liu、Ekin Do-gus Cubuk、JayShi、Shuyang Cheng 和 Dragomir Anguelov。2022年。PolyLoss：从多项式展开视角看分类损失函数。载于国际学习表征会议。

Yu Li、Tao Wang、Bingyi Kang、Sheng Tang、Chunfeng Wang、Jintao Li 和 Jiashi Feng。2020年。利用平衡组 softmax 克服长尾目标检测中的分类器不平衡。2020年 IEEE/CVF计算机视觉与模式识别会议，第10988–10997页。

Tsung-Yi Lin、Priya Goyal、Ross Girshick、Kaim- ing He 和 Piotr Dollár。2017年。用于密集目标检测的焦点损失。载于《2017年IEEE国际计算机视觉会议》，第 2999–3007 页。

Sheng Liu、Jonathan Niles-Weed、Narges Razavian 和 Carlos Fernandez-Granda。2022年。早期学习正则化可防止对噪声标签的记忆。载于第34届神经信息处理系统国际会议论文集，NIPS'20，美国纽约州雷德胡克。Curran Associates Inc.

Ziwei Liu、Zhongqi Miao、Xiaohang Zhan、Jiayun Wang、Boqing Gong 和 Stella X. Yu。2019年。开放世界中的大规模长尾识别。载于2019 IEEE/CVF计算机视觉与模式识别会议，页码2532–2541。Michal Lukasik、Srinadh Bhojanapalli、Aditya Krishna Menon 和 Sanjiv Kumar。2022年。标签平滑能缓解标签噪声吗？载于第37届国际机器学习会议论文集。JMLR.org。Winter A. Mason 和 Siddharth Suri。2012年。在亚马逊 Mechanical Turk 上开展行为研究。《行为研究方法》，44:1–23。Aditya Krishna Menon、Sadeep Jayasumana、Ankit Singh Rawat、Himanshu Jain、Andreas Veit 和 Sanjiv Kumar。2021年。通过 logit调整实现长尾学习。ArXiv，abs/2007.07314。

Arvind Neelakantan、Tao Xu、Raul Puri、Alec Radford、Jesse Michael Han、Jerry Tworek、Qiming Yuan、Nikolas Tezak、Jong Wook Kim、Chris Hallacy、Johannes Heidecke、Pranav Shyam、Boris Power、Tyna Eloundou Nekoul、Girish Sastry、Gretchen Krueger、David Schnurr、Felipe Petroski Such、Kenny Hsu、Madeleine Thompson、Tabarak Khan、Toki Sherbakov、Joanne Jang、Peter Welinder 和 Lilian Weng。2022年。通过对比预训练获得文本和代码嵌入。

Gabriele Paolacci、Jesse Chandler 和 Panos Ipeirotis。2010年。使用亚马逊 Mechanical Turk 开展实验。《判断与决策》，5:411–419。

Giorgio Patrini、Alessandro Rozza、Aditya Krishna Menon、Richard Nock 和 Lizhen Qu。2017年。使深度神经网络对标签噪声具有鲁棒性：一种损失修正方法。载于2017 IEEE计算机视觉与模式识别会议，第2233–2241页。

Gabriel Pereyra、G. Tucker、Jan Chorowski、Lukasz Kaiser 和 Geoffrey E. Hinton。2017年。通过惩罚高置信度输出分布来正则化神经网络。ArXiv，abs/1701.06548。

Scott E. Reed、Honglak Lee、Dragomir Anguelov、Christian Szegedy、D. Erhan 和 Andrew Rabinovich。2015年。使用自举在噪声标签上训练深度神经网络。CoRR，abs/1412.6596。

Mengye Ren、Wenyuan Zeng、Binh Yang 和 Raquel Urtasun。2018年。学习重新加权样本，以实现鲁棒深度学习。ArXiv，abs/1803.09050。

Eric Arazo Sanchez、Diego Ortego、Paul Albert、NoelE. O'Connor 和 Kevin McGuinness。2019年。无监督标签噪声建模与损失修正。CoRR，abs/1904.11238。

Connor Shorten 和 Taghi Khoshgoftaar。2019年。关于深度学习图像数据增强的综述。《Journal of BigData》，6。

Jun Shu、Qi Xie、Lixuan Yi、Qian Zhao、Sanping Zhou、Zongben Xu 和 Deyu Meng。2021年。Meta-Weight-Net：学习用于样本加权的显式映射。Curran Associates Inc.，美国纽约州雷德胡克。

Hwanjun Song、Minseok Kim 和 Jae-Gil Lee。2019年。SELFIE：为鲁棒深度学习修复不洁样本。载于第 36 届国际机器学习会议论文集，《机器学习研究论文集》第 97 卷，第 5907–5915 页。PMLR。

Hwanjun Song、Minseok Kim、Dongmin Park、Yooju Shin 和 Jae-Gil Lee。2022年。使用深度神经网络从噪声标签中学习：综述。IEEE神经网络与学习系统汇刊，第 1–19 页。

Nitish Srivastava、Geoffrey Hinton、AlexKrizhevsky、Ilya Sutskever 和 Ruslan Salakhutdinov。2014年。Dropout：一种防止神经网络过拟合的简单方法。《机器学习研究期刊》，15(1)：1929–1958。

Sainbayar Sukhbaatar、Joan Bruna、Manohar Paluri、Lubomir D. Bourdev 和 Rob Fergus。2014年。使用噪声标签训练卷积网络。arXiv：计算机视觉与模式识别。Yanmin Sun、Mohamed S. Kamel、Andrew K.C. Wong 和 Yang Wang。2007年。用于不平衡数据分类的代价敏感提升。模式识别，40(12):3358–3378。Chuanqi Tan、Fuchun Sun、Tao Kong、Wenchang Zhang、Chao Yang 和 Chunfang Liu。2018年。深度迁移学习综述。载于《人工神经网络与机器学习——2018》，第270–279页，沙姆。施普林格国际出版公司。Eric Tzeng、Judy Hoffman、Kate Saenko 和 Trevor Darrell。2017年。对抗性判别域适应。载于2017 IEEE计算机视觉与模式识别会议，第2962–2971页。Peng Wang、Kai Han、Xiu-Shen Wei、Lei Zhang 和 Lei Wang。2021年。基于对比学习的长尾图像分类混合网络。载于2021 IEEE/CVF计算机视觉与模式识别会议，第943–952页。Yisen Wang、Xingjun Ma、Zaiyi Chen、Yuan Luo、Jinfeng Yi 和 James Bailey。2019年。用于噪声标签鲁棒学习的对称交叉熵。载于2019 IEEE/CVF国际计算机视觉会议，第322–330页。Yu-Xiong Wang、Deva Ramanan 和 Martial Hebert。2017年。学习建模长尾。载于第31届神经信息处理系统国际会议论文集，NIPS'17，第7032–7042页，美国纽约州雷德胡克。Curran Associates Inc.

Zhuoyi Wang、Yigong Wang、Bo Dong、Sahoo Pracheta、Kevin Hamlen 和 Latifur Khan。2020年。基于自适应边界的深度对抗度量学习。载于2020年IEEE第六届云上大数据安全国际会议（BigDataSecurity）、IEEE高性能与智能计算国际会议（HPSC）和IEEE智能数据与安全国际会议（IDS），第100–108页。

Jun Wu、Xuesong Ye 和 Yanyuet Man。2023a。Bot-trinet：通过度量学习实现社交机器人检测的统一高效嵌入方法。

Jun Wu、Xuesong Ye、Chengjie Mou 和 Weinan Dai。2023b。FineEHR：精炼临床笔记表征以提升死亡率预测。

Zhuofeng Wu、Sinong Wang、Jiatao Gu、Madian Khabsa、Fei Sun 和 Hao Ma。2020。CLEAR：用于句子表示的对比学习。ArXiv，abs/2012.15466。

Tong Xiao、Tian Xia、Yi Yang、Chang Huang 和 Xiaogang Wang。2015。利用海量带噪标注数据进行图像分类学习。载于2015年IEEE计算机视觉与模式识别会议，第2691–2699页。

Yiming Xu、Lin Chen、Zhongwei Cheng、Lixin Duan 和 Jiebo Luo。2020年。通过多模态领域自适应实现开放式视觉问答。自然语言处理经验方法会议。Yiming Xu、Lin Chen、Lixin Duan、Ivor W. Tsang 和 Jiebo Luo。2021年。采用软未知类别拒绝的开放集领域自适应。IEEE神经网络与学习系统汇刊，第1–12页。Xingrui Yu、Bo Han、Jiangchao Yao、Gang Niu、Ivor Wai-Hung Tsang 和 Masashi Sugiyama。2019年。分歧如何帮助提升应对标签污染的泛化能力？载于 ICML。Chen Zhang、Lei Ren、Jingang Wang、Wei Wu 和 Dawei Song。2022年。让预训练语言模型成为优秀的长尾学习器。Chiyuan Zhang、Samy Bengio、Moritz Hardt、Benjamin Recht 和 Oriol Vinyals。2021a。理解深度学习（仍然）需要重新思考泛化。Commun. ACM，64(3):107–115。Hongyi Zhang、Moustapha Cisse、Yann N. Dauphin 和 David Lopez-Paz。2018年。mixup：超越经验风险最小化。载于国际学习表征会议。Yifan Zhang、Bingyi Kang、Bryan Hooi、Shuicheng Yan 和 Jiashi Feng。2021b。深度长尾学习：综述。ArXiv，abs/2110.04596。Zhilu Zhang 和 Mert R. Sabuncu。2018年。用于训练带噪声标签的深度神经网络的广义交叉熵损失。载于论文集

第32届神经信息处理系统国际会议，第8792–8802页，纽约州雷德胡克：Curran Associates Inc。

MengZhou、Zechen Li 和 Pengtao Xie。2021 年。用于文本分类的自监督正则化。《计算语言学协会会刊》，9：641–656。