杭州电子科技大学

硕 士 学 位 论 文

题　　目:基于 BERT 的带噪半监督文本分类方法研究

研 究 生　任子扬

专　　业　电子信息

指导教师　姚英彪　教授

完成日期 2023 年 4 月

中国知网　https://www.cnki.net

杭州电子科技大学硕士学位论文

基于 BERT 的带噪半监督文本分类方法研究

研 究 生：任子扬

指导教师：姚英彪　教授

2023 年 4 月

中国知网　https://www.cnki.net

Dissertation Submitted to Hangzhou Dianzi University
for the Degree of Master

Research on noisy semi-supervised text classification method based on BERT

Candidate: Ziyang Ren
Supervisor: Prof. Yingbiao Yao

April, 2023

中国知网 https://www.cnki.net

摘 要

文本分类是自然语言处理领域（Natural Language Process，NLP）的一个基本问题，现在最流行的方法即利用大数据训练深度神经网络。但是，大量数据的获取可能需要耗费很高的人力时间等成本，数据标签正确率可能也无法保证。故半监督学习、带噪学习等更为廉价的深度学习方法应运而生。基于 BERT（Bidirectional Encoder Representations from Transformers）模型结构，本文分别在半监督学习和带噪学习两方面做出研究和探索。

（1）针对半监督学习中文本数据增强过程中语义的过度变化以及噪声标签的问题，本文提出了一种对噪声标签鲁棒的对抗半监督学习方法 ASSL（Adversarial Semi-Supervised Learning）用于文本分类。ASSL 的主要贡献包括改进的对抗数据增强方法和损失函数。针对对抗数据增强，本文提出了最大隐藏梯度下降（Maximum Hidden Gradient Descent，MHGD）方法，让最强对抗扰动作用于 BERT 模型的 Transformer 隐藏层的特定语义表示张量，在提升了模型一致性的同时也降低了语义剧烈变化的风险。对于损失函数，本文将反向交叉熵作为噪声容忍项，与交叉熵相结合，提出了 Flex-Symmetric Cross Entropy（Flex-SCE）。Flex-SCE 在训练过程中动态地降低了标记数据的影响。因此，ASSL 在有限的标记数据（尤其是噪声标签）上最小化了模型的过拟合。实验表明，与几种先进的半监督学习方法相比，ASSL 在多个数据集上取得了优异的训练效果，并显著降低了噪声标签的影响。

（2）对于小样本带噪的文本分类任务，本文提出了一种基于置信学习（Confident Learning，CL）改进的小样本带噪学习方法（FewCL）。置信学习适用于大规模数据集合的数据筛选降噪处理，但在样本较少时，其依靠每个类别的模型预测概率均值作为阈值做出正误判断的方法仍存在改进空间。一方面，欠拟合的模型可能会过于武断地去除数据集中的困难样本，甚至导致数据量过少或者种类的不平衡；另一方面，受训练集中错误标签影响的模型对标签正误的判断不一定可靠。不同于置信学习利用模型预测概率来判断标签正误，本方法利用 BERT 编码语句语义，将每条语句映射到高维空间得到语义表达张量和每个类别的“原型”张量。以聚类的思想，本方法计算每个语义表达张量到该类别“原型”的“距离”，然后以该“距离”及其均值、标准差为基准，将样本标签细分为错误标签、正确困难标签、正确简单标签三类，并且在滤除错误标签样本之后按照“距离”来衡量标签置信度，给剩余样本赋予权重并参与后续模型训练。实验证明，相比传统全监督训练、对称交叉熵和置信学习，本方法在多个数据集和多种标签错误率下都表现出了最好的模型训练效果。

关键词：文本分类、半监督学习、带噪学习、深度神经网络

中国知网 https://www.cnki.net

ABSTRACT

Text classification is a basic problem in the field of Natural Language Process (NLP). The most popular method is to train deep neural network models with a large amount of data. However, the acquisition of large amounts of data may require a lot of labor time and other costs, and the accuracy of data labels may not be guaranteed. Therefore, cheaper deep learning methods such as semi-supervised learning and noisy learning have emerged. Based on the BERT (Bidirectional Encoder Representations from Transformers) model structure, this paper makes research and exploration in both semi-supervised learning and noisy learning.

(1) Aiming at the problem of excessive semantic changes and noise labels in the process of text data enhancement in semi-supervised learning, this paper proposes an Adversarial Semi-Supervised Learning (ASSL) robust to noise labels for text classification. The main contributions of ASSL include improved adversarial data augmentation methods and loss functions. For adversarial data enhancement, this paper proposes a Maximum Hidden Gradient Descent (MHGD) method, which allows the strongest adversarial perturbation to act on the specific semantic representation tensor of the Transformer hidden layer of the BERT model, which improves the consistency of the model and reduces the risk of dramatic semantic changes. For the loss function, this paper proposes Flex-Symmetric Cross Entropy (Flex-SCE) by combining reverse cross entropy with cross entropy as a noise tolerance term. Flex-SCE dynamically reduces the influence of labeled data during training. Therefore, ASSL minimizes the overfitting of the model on limited labeled data (especially noise labels). Experiments show that compared with several advanced semi-supervised learning methods, ASSL achieves excellent training performance on multiple datasets and significantly reduces the impact of noise labels.

(2) For small sample noisy text classification tasks, this paper proposes an improved small sample noisy learning method (FewCL) based on Confident Learning (CL). Confident learning is suitable for data filtering and noise reduction processing of large-scale data sets. However, when the number of samples is small, there is still room for improvement in the method of relying on the mean value of the model prediction probability of each category as the threshold to make a judgment. On the one hand, under-fitting models may remove difficult samples in the data set too arbitrarily, and even lead to too little data or imbalance of types ; on the other hand, the model affected by the wrong label in the training set is not necessarily reliable in judging the correctness of the label. Different from belief learning, which uses model prediction probability to judge whether the label is correct or not, this

中国知网 https://www.cnki.net

method uses BERT to encode the semantics of sentences, and maps each sentence to a high-dimensional space to obtain a semantic expression tensor and a 'prototype' tensor for each category. Based on the idea of clustering, this method calculates the 'distance' of each semantic expression tensor to the 'prototype' of the category. Then, based on the ' distance ' and its mean and standard deviation, the sample labels are subdivided into three categories: error labels, correct difficult labels and correct simple labels. After filtering out the wrong label samples, the label confidence is measured according to the 'distance', and the remaining samples are given weights and participate in the subsequent model training. Experiments show that compared with traditional supervised training, Symmetric Cross Entropy and Confident learning, this method shows the best model training effect under multiple data sets and multiple label error rates.

Keywords: Text classification, Semi-supervised learning, Noisy learning, Deep neural network

中国知网 https://www.cnki.net

目 录

第 1 章 绪论.....1
1.1 研究背景和意义.....1
1.2 国内外研究现状.....2
1.2.1 传统的文本分类方法.....2
1.2.2 基于深度学习的文本分类方法.....2
1.2.3 基于半监督学习的文本分类方法.....3
1.2.4 文本分类数据集.....4
1.2.5 本文涉及的相关概念.....4
1.3 研究难点与挑战.....6
1.4 主要研究内容.....7
1.5 论文章节安排.....8
第 2 章 文本分类算法理论基础.....9
2.1 深度神经网络.....9
2.1.1 卷积神经网络.....9
2.1.2 循环神经网络.....9
2.1.3 Transformer 与注意力机制.....10
2.1.4 BERT 深度神经网络.....10
2.2 算法原理及发展.....11
2.2.1 损失函数以及反向传播算法.....11
2.2.2 对抗学习方法.....12
2.2.3 半监督学习.....13
2.2.4 数据增强.....14
2.2.5 带噪学习.....14
2.2.6 小样本学习.....15
2.3 评价指标.....15
2.4 本章小结.....16
第 3 章 ASSL:一种对标签噪声鲁棒的对抗半监督学习方法.....17
3.0 ASSL 系统框架.....17
3.1 最大隐藏梯度下降.....18
3.2 损失函数.....20
3.3 ASSL 半监督学习过程.....23
3.4 实验及结果分析.....23
3.4.1 实验模型结构与数据集.....23
3.4.2 训练方法比较.....24

中国知网 https://www.cnki.net

3.4.3 在不同隐藏层添加扰动的效果比较.....24
3.4.4 消融与对比实验.....25
3.4.5 噪声标签测试.....26
3.5 本章小结.....26
第 4 章 基于置信学习改进的小样本带噪学习方法.....28
4.1 任务场景及其符号.....28
4.2 相关工作及其改进.....29
4.2.1 置信学习方法.....29
4.2.2 小样本学习方法.....31
4.2.3 待改进之处的分析.....32
4.3 基于置信学习的小样本带噪学习文本分类方法设计.....32
4.4 实验及结果分析.....35
4.4.1 模型结构与数据集.....35
4.4.2 BERT 模型各隐藏层训练结果比较.....36
4.4.3 不同 K 值效果对比.....37
4.4.4 带噪学习效果.....38
4.5 本章总结.....40
第 5 章 总结与展望.....42
5.1 研究工作总结.....42
5.2 研究工作展望.....42
参考文献.....44

中国知网 https://www.cnki.net

第 1 章 绪论

1.1 研究背景和意义

近几年 IT 与互联网发展迅速，人类已经步入第三次科学技术革命——大数据时代。计算机应用的发展使得很多分析和决策可以更多的基于非结构化数据[1,2]。与此同时，各种数据信息生产与传播方式也经历着一场深刻变革。杂志报刊渐渐退出了历史舞台，手机、平板电脑等移动终端早已随处可见，各类新兴媒体不断涌现，渗透到社会的各个角落。万事万物的互联互通使得数据信息成倍增长，并以文字、图片、视频、音频等方式给人带来更加直观有效的体验。其中，文本文字是信息密度最高、资料最多、成本最低、最易为网民获得的信息传播媒介。因此，文本信息处理技术成为当前热门的研究领域。

在目前信息爆炸大环境中，我们享受着科技给生活品质带来的提升，也痛苦于资讯过量带来的烦恼。在目前大量数据信息源源不断的大环境中，对于文本信息进行分析与管理是非常必要且重要的，其中之一便是文本分类技术。

人工智能时代的到来使得文本分类技术得到了迅速发展。文本分类研究开始于 50 年代后期，有关技术已迭代更新数年，目前已被各种搜索引擎、互联网论坛等平台广泛使用。从最初简单地依靠人工完成到后来利用计算机算法来完成分类任务，再发展成为今天的智能化系统，这一系列过程都体现出了人工智能与文本分类之间紧密关系。最初的文本分类主要依靠人工操作，文本的分类文员要吃透文本，要在了解文本意义的前提下，将文本归类。这种方式准确率不高且耗时耗力，于是人们便尝试将机器学习引入到文本分类中。伴随着机器学习[3]的产生与更替，机器自动分类已逐步代替手工工作，并作为自然语言处理领域[4,5]（Natural Language Process，NLP）的一个课题发展至今。机器学习技术作为一种基于计算机模拟人类思维的新型智能技术，具有高度自适应能力，可以从大量无结构化的文本数据中提取出有用的特征信息。伴随着计算机科学技术向高性能大数据领域的发展以及深度学习理论的出现，某些传统机器学习算法很难解决的问题实现了更进一步地突破。通过有效的文本分类技术，媒体平台可以对关键信息进行有效的定位和分类筛选。在对复杂信息的梳理、优化搜索效率等等方面，现代文本分类技术都起着其他手段无法代替的作用。

在文本分类技术的实际应用场景中，文本数据需要经过标注后才能用于深度学习模型训练。但是，大量高质量训练数据的获取却始终是一个问题，数据的大量标注需要耗费大量的人力、时间成本。与此同时，线上每时每刻都在产生大量的无标签数据样本。因此，同时使用了带标签和无标签[16]数据的半监督学习方法应运而生。

现有的半监督模型训练方法同时需要带标签数据和无标签数据[6,7,8]，训练效果十分依赖

中国知网 https://www.cnki.net

于带标签数据，但很少有人考虑其中带标签数据的标签正确率问题。实际上，人工标注结果可能并不十分理想，获取的标注数据可能含有一定比例的错误标签，即标签噪声。研究表明，神经网络模型在训练过程中，总是会学到错误标签中的信息，而半监督的训练过程甚至可能会加剧这种错误印象，当标签错误率超出一定范围，多用了很多无标签数据的半监督训练可能还不如使用少量标签数据的全监督训练得到的模型精度高。

因此，半监督学习、带噪学习方法的研究很有必要。它们可以降低深度学习方法投入实际应用的门槛，在数据不充分、且标签质量得不到保证的情况下，尽量提升模型训练效果，增强半监督学习过程的鲁棒性。越来越多的学者也在开展相关的研究。

1.2 国内外研究现状

1.2.1 传统的文本分类方法

上世纪 90 年初期，特殊工程与浅分类模式的有机融合逐渐受到学术领域的关注，这种分类方法相对传统[17]，通常由文本表示、文本预处理、机器学习[18]几种方法组成[9]。最被广泛应用的文本表示方法是词袋模型（Bag of Word, BOW）。N-gram 模型[15]是一种类似于词袋模型的文本表示方法[20]，其主要将文本语序纳入考量范围。文本预处理[23]通常涵盖对标点符号、停用词全面删除等符合实际业务的操作。机器学习的方法可以理解为分类器，一般情况下其主要由支持向量机、逻辑回归等算法构成。此外，还可通过文本特征过滤的方式进行处理。

上述方法大部分都需要利用分散的词汇来表达文本特征，并通过其他线性模型对其展开分类处理，虽然在分类效果上有所提升，但传统的机器学习方法在面对一些复杂的分类任务时，分类的精度不高。

1.2.2 基于深度学习的文本分类方法

深度学习方法多年来受到国内外学者的关注，在语音识别，计算机视觉等领域都取得了较为理想的发展，也被广泛地应用于各种自然语音场景。以深度学习方法基础上对文本进行分类，主要有以下两个优势：

（1）灵活度高。深度学习主要利用深层神经网络进行文本中各种特征的提取，并对语义语法等信息做出相应表示，对人工设计规则和特征选择的需求不大，能够有效减少人工成本。

（2）精度较高。近年来，卷积神经网络等等、循环神经网络等等、注意机制等模型和算法正在演化和应用。上述模型和算法能够更大地提高文本信息进行识别和分类的精准度。因此，基于深度学习的方法是文本分类研究的优先选择。

卷积神经网络（Convolutional Neural Network，CNN）[24]可以在固定区域内对文本特征进行抓取。卷积神经网络中的卷积层可以较好的提取局部特征，所以卷积神经网络在计算机视觉（Computer Vision，CV）领域取得很大的成就，后来有学者将其引入自然语言处理领域。在 2014 年，Yoon Kim[35]针对 CNN 的输入层做了一些变形，提出了文本分类模型 TextCNN。

中国知网 https://www.cnki.net

与传统图像的 CNN 网络相比，TextCNN 在网络结构上没有任何变化，甚至更加简单。TextCNN 只有一层卷积，一层池化层，最后将输出外接 Softmax 来多分类。文本信息可以转化为字/词向量来表示，网络卷积层中，可通过文本矩形卷积核进行计算，卷积核从文章的开头移动到文章的末尾。有时候可以利用单个卷积核来完成文档特征提取。Conneau[25]等设计了 VDCNN（Very deep convolutional neural network），通过增加卷积和池化层的个数，从而得到多个窗口的具体特征信息。VDCNN 更适用于大规模数据集，且其性能的提升并不依赖其他的数据增强技术。

除 CNN 之外，还有能获得上下文信息的循环神经网络（Recurrent Neural Network，RNN）[34]，它在文本分类方面起到了至关重要的影响与作用。RNN 的主导思想是利用组合前一时间所具有的单元信息与当前单元信息之间所产生的关联，并借此引入“递归”的功能，然而若使用这种结构处理较长的文字，梯度消失与梯度爆炸问题极易伴随着错误向后扩散。Hochereiter[26]等人在他们的研究中分析了随时间反向传播（BPTT）带来的梯度爆炸和梯度消失问题，第一次提及了“长-短期记忆”（Long-Short Term Memory, LSTM）这一概念，实现“忘却”与“记忆”效应，解决 RNN 长期依赖的问题。

1.2.3 基于半监督学习的文本分类方法

现有半监督学习多用于图像分类中，近年来主流方法大多数采用一致性正则[27]，熵最小化[28]等思想。一方面如果在数据输入中加入噪声，鲁棒模型输出应该是类似的；另一方面要尽量最小化无标签数据中模型的熵。

“一致性正则”意味着无标签数据经过一次或者多次的增强之后，被输入到模型中，以该模型“较低熵”输出作为其“伪标签”。同时，通过另一种增强方式将无标签数据输入到模型中，这时，模型输出要与刚获得的“伪标签”有相似性。刘文豪[10]等学者采用协同训练[7]的方式来避免错误伪标签不断被使用。张晓龙等[19]学者采用有监督学习和无监督学习的一致性训练方式，针对金融领域文本引入无监督数据增强方法，即对特定任务使用特定目标的数据增强方法，以产生更有效的数据。仝鑫[21]等学者使用繁体、拼音替换等多种攻击策略生成与原句语义一致的对抗样本并以此解决中文文本分类模型的鲁棒性问题。

“低熵”的度量方法通常是观察模型输出概率分布的最大可能性是否大于某阈值。如何度量“相似性”，则依赖于选用什么样的损失函数。通常的选择有 KL 散度，交叉熵和 L2 正则等。闫云飞[22]等人针对数据集的类别数量不均衡和分类难易不均衡问题，在传统焦点损失函数的基础上提出了一种可以根据样本不均衡性特点动态调整的变焦损失函数。

文本半监督分类中存在着两个问题。一方面是如何把图像半监督分类一致性正则与最小化熵的概念运用到文本处理中。图像的旋转，镜像、转换灰度等这些简单的操作，便可在不改变样本标签的情况下，达到数据增强的目的。但是，NLP 下，信息在文本中的分布存在着一定的离散特性，很难用简单的变换产生大量语义恒定的扩充样本。特别在短文本里，修改

中国知网 https://www.cnki.net

某些词，极有可能会直接改变整个语义，故直接在词向量上添加扰动并非最佳选择。有些学者利用生成时对抗网络

另一方面，通过设置概率阈值对模型输出置信度进行评判，此法更简便，更直接，但是实际运行过程中存在着一些问题：由于阈值的存在，无标签数据的添加可能需要到模型训练的中期，甚至是末期，而这个时候那些早已经参加了训练的少量带标签的数据，可能已经过拟合。使用过拟合后的模型来对数据进行标签预测很可能包含错误标签，即噪声标签，并且该误差不是均匀分散随机的，而是该模型“死读书”后输出的一种系统性误差。所以，如何处理带标签和无标签数据的协同训练，如何应对模型预测过程中产生的噪声标签，都是需要克服的难题。

1.2.4 文本分类数据集

本文基于三个公共英文文本分类数据集和一个私有中文文本分类数据集开展了实验。这三个公共数据集是 IMDB[29]、AG news[30]和 DBpedia[31]。其中，部分训练集被用于验证。原始测试集被用来测试训练性能。私有数据集 Best 来自浙江百世科技有限公司的智能客服，并随机选择一部分经过检查的样本作为测试集。几个未检查的样本被用来训练和验证模型。

IMDB 数据集包含来自互联网电影数据库（IMDB）的 50000 条评论。训练集和测试集都包含 50%的正面评论和 50%的负面评论。

AG news 包含 AG news 语料库中 2000 多个新闻来源的文章。这些数据包括四个类别：国际、体育、商业和科学/技术。

DBpedia 分类数据集是通过从 DBpedia2014 中选择 14 个类别的数据构建的。

Best 意图分类数据集包括“订单交付”、“提醒”、“物流信息查询”和“投诉”四个意图类别。这些标签来自于百世公司内部员工对在线客服数据的注释。由于员工在工作中可能存在疏忽，这些标签可能存在一定的错误。

1.2.5 本文涉及的相关概念

（1）预训练语言模型

预训练语言模型已经证明对改进许多自然语言处理[3]任务是有效的[8]。这些任务包括句子级任务，如自然语言推理[10]和释义[11]，旨在通过整体分析来预测句子之间的关系，以及实体级任务，如命名实体识别[12]，其中模型需要在实体级别生成细粒度输出。

将预训练语义表示向量应用于下游任务有两种现有策略：基于特征和微调。基于特征的方法，使用特定于任务的体系结构，其包括预训练的产生的语义向量作为附加特征。微调方法，例如 Generative Pre-trained Transformer（GPT）[14]，引入了最小的任务特定参数，并通过简单地微调预训练参数来训练下游任务。这两种方法在预训练期间可以共享相同的目标函数，且使用单向语言模型来学习一般语言表示。

从语言模型（Language Model, LM）迁移学习的最新趋势是在任务目标上预先训练一些

中国知网 https://www.cnki.net

模型架构，然后对监督下游任务的相同模型进行微调。这些方法的优点是需要从头开始学习很少的参数。至少部分任务可以借助这一优势，比如 GPT 在 GLUE 基准测试中获得了许多句子级任务的最好结果。

有学者认为单向语言模型严重限制了预训练向量表示语义的能力，特别是对于微调方法。语言模型的单向限制了在预训练期间可以使用的体系结构的选择。例如，GPT 使用了从左到右的体系结构，其中每个实体只能处理 Transformer 自我注意层中的前一个实体[36]。这些限制对于句子级别任务来说是次优的，在这些任务中，从两个方向合并上下文至关重要。于是有学者提出 BERT（Bidirectional Encoder Representations from Transformers）[37]来改进基于微调的方法。

（2）张量

张量是一种数学对象，它可以用来表示多维数据，比如矩阵、向量、标量等等。张量在物理学、工程学、计算机科学等领域中都有广泛的应用，比如在机器学习中，张量被用来表示数据集和模型的参数。在数学中，张量是向量和矩阵的推广，它可以包含任意维度的数组。张量具有特定的变换规则，可以进行加法、乘法等基本运算。

回顾对“量”的认知历程，代数手法中的“数”（标量）和“数组”（列矩阵、行矩阵）都能有效地表达数量和向量。在很多数学研究和应用中若不使用这种方法则很难发现其规律和特征。若运用几何学的技巧，实数在标量中可表示为一维坐标系（例如数轴）上的一点，而复数则可表示为二维坐标系（例如复平面）上的一点。这样就可以将矢量与量纲联系起来。在二维、三维或更高维坐标系中，向量可以被描述为一个空间点或有向线段，这是一种常见的表示方式。这样就使向量既具有了一定程度的直观性，又具备了一定的抽象思维能力。

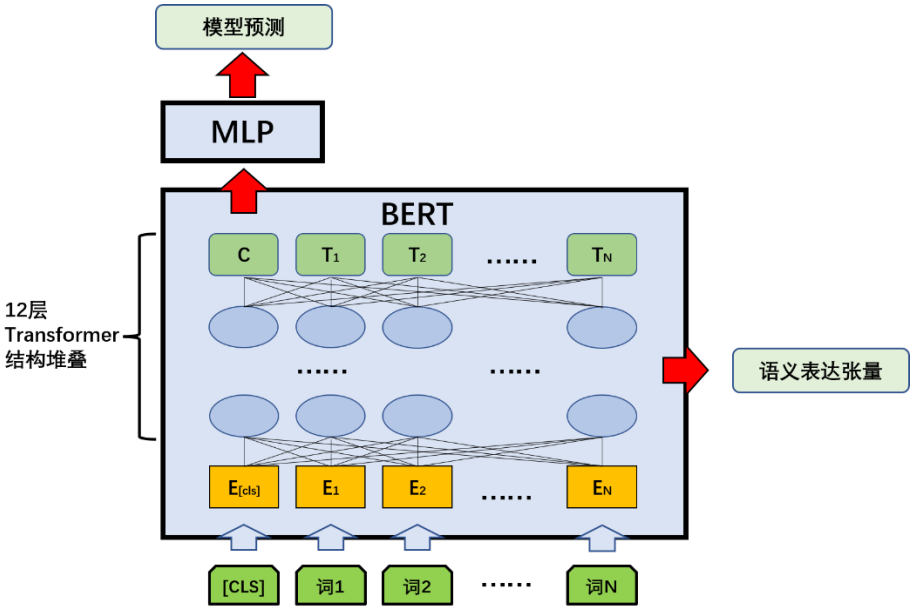

图 1.1 BERT 模型大致结构图

当使用深度学习模型处理 NLP 问题时，文本、字符串无法用数学工具处理，所以一般文

中国知网 https://www.cnki.net

本被转化为张量后输入模型。深度学习模型一般具有多层结构，每层的输入输出都是张量数据，每层的计算过程都包含大量线性代数运算。在对输入张量层级递进的计算处理过程中，原始文本的语义、语法等特征被逐渐提取出来。以 BERT[39]模型为例，首先将一句话转化为一个一维张量（也可理解为数组或向量），长度为语句截断长度。然后，将该一维张量输入 BERT 模型。BERT 模型（如图 1.1 所示）包含一个词向量嵌入层和一个 12 层 Transformer 结构的 Encoder 编码器。一维张量数据输入词向量嵌入层后会匹配每个词的词向量，然后将该一维张量转化为二维张量（也可理解为矩阵或者向量组），大小为语句截断长度×词向量维度。Encoder 编码器中的每层 Transformer 结构的输出张量都可看作是“语义表达张量”，都包含了不同程度上提取的语义特征。这个过程也可看作为 BERT 模型将语句语义编码映射到高维空间得到高维语义表达张量。

（3）词向量

词向量是自然语言处理中的一种技术，它将每个单词表示为一个向量，使得单词之间的相似度可以通过向量之间的距离来计算。常见的词向量模型有 Word2Vec[32]、GloVe[33]等。这些模型通过对大量语料进行学习，将每个单词映射到一个高维向量空间中，使得相似的单词在向量空间中距离较近，不相似的单词距离较远。这种词向量表示可以应用于文本分类、情感分析、机器翻译等自然语言处理任务中，提高了模型的准确性和效率。

（4）语料库

语料库是指用于语言学研究的文本数据集合，通常包括大量的现实语言使用情况，可以是文本、口语、书籍、新闻、社交媒体帖子等等。语料库用于描述、研究和了解某种语言的语法、词汇、语用等方面的特点和规律。它也是自然语言处理技术的重要数据源之一，通过对大规模语料库的分析和学习，可以提高文本分类、情感分析、机器翻译、文本生成等任务的准确性和效率。常见的语料库包括 Brown 语料库、Penn Treebank 语料库、Gutenberg 语料库等。

1.3 研究难点与挑战

文本分类任务的目标是通过一定数量的样本训练得到模型，然后利用该模型准确预测未知文本所属的类别。但是面对实际丰富多样的语料样本，达到目标存在诸多难点和挑战：

（1）数据集可能存在错误。在数据集的产生过程中可能存在标签错误，这些标签错误会对模型训练过程造成很大影响，如何使模型对于标签错误有更多的鲁棒性是一个有挑战性的问题。

（2）部分场景下，文本样本的数量可能不充分。传统深度学习方法的训练过程需要大量带标签样本，然而标签的产生需要耗费大量人力和时间。

（3）文本分类样本数据可能不均衡。有些数据集中，不同类别的样本数量呈现长尾分布，少数几个类别中包含大量样本，而相当一部分类别只包含少数样本。这样训练出的模型对于

中国知网 https://www.cnki.net

不同类别的预测精准度会有很大差距。

（4）样本不同类别之间可能存在相关性。部分文本在语义上可能处于多个类别的“边界区域”，模型在判断时也会在这些模糊区域出现“误判”，甚至有时候机器的“误判”结果，经检查后会发现其实算不上“误判”。

（5）文本有长短区别，部分信息较少的短文本，或者信息过于分散的长文本都给模型的训练和预测带来困难。

（6）文本预处理阶段需要经过分词处理，而分词的结果会受到文本极大的影响，分词的结果也会影响训练预测的效果。

1.4 主要研究内容

在大量文本数据应用过程中，分类是基础工作。目前常用于文本分类的算法有很多，包括机器学习，神经网络等。传统的文本分类通常是建立在带标签的数据上进行监督学习的，但是在现实具体的工业场景下，数据多数是无标签原始数据，对数据进行标注会消耗大量人力资源，且人工标注的标签质量如何也值得怀疑，在现实中，甚至会发生“训练数据愈丰富训练效果愈差”现象。针对这一问题，本文提出一种新的基于半监督的文本分类算法，通过对有标记数据进行特征提取，并将其用于构建分类器，从而达到提高准确率的目的。故只使用少量带带标签数据，而使用大量没带标签数据的半监督学习方法，便成为一种较为实用，低成本的选择。

（1）对于标签的质量无法保证的文本半监督分类任务，本文提出了一种对抗性半监督学习（Adversarial Semi-Supervised Learning，ASSL）方法。为了使模型对噪声标签更加稳健，本文提出了用于标签训练的 Flex-Symmetric Cross Entropy（Flex-SCE）损失函数。通过 Flex-SCE 损失，交叉熵项的动态权重将逐渐减少标签的影响，特别是避免了在有限的噪声标签数据上的过拟合。无标签训练过程中采用了一致性正则化的思想，其中提出了 MHGD 对抗性学习方法来产生最强的扰动，用于文本数据的增强。为了避免语义信息的过度变化，最强扰动作用于语义表示张量。

（2）对于小样本[38]带噪的文本分类任务，本文提出了一种基于置信学习改进的小样本带噪学习方法（FewCL）。为了让模型的纠错不会因为预先训练过程中的过拟合现象影响，本文采取类似 K 折验证的思路，将数据集分成多块，分块验证筛选错误标签的同时逐步对模型做训练和微调。为了适用于小样本学习场景，本文利用 BERT 模型提取语义并且映射到高维空间后形成语义表达张量，通过比较每条语句的高维语义表达张量和该类别的“原型”张量之间的“距离”，判断数据标签的正误。最后再给每条判断为“正确”的样本根据其“难易程度”赋予不同的权重，然后将其投入后续训练过程。

中国知网 https://www.cnki.net

1.5 论文章节安排

根据研究课题，本文章节安排如下：

第一章：绪论。阐述了文本分类问题的研究背景和意义、国内外研究现状、研究难点及挑战，最后概述了全文内容及章节安排。

第二章：文本分类相关理论基础。首先介绍了不同神经网络结构及其发展，包括了 CNN 卷积神经网络、RNN 循环神经网络、Transformer 注意力机制以及 BERT 模型结构等；然后阐述对抗学习等相关深度学习模型训练方法；最后叙述了文本分类问题的相关参数和评价指标。

第三章：一种基于对抗学习的带噪半监督文本分类方法。首先提出了 MHGD 对抗性学习方法用于文本数据增强，以保持整体语义信息，同时提高正则化性能。然后构建了 Flex-SCE 损失函数，其中交叉熵项被赋予一个按照一定规则衰减的权重系数，以提高对噪声标签的鲁棒性。最后搭建了一种对抗性半监督学习方法 ASSL，用于标记数据不足和有噪声的文本分类任务。

第四章：基于置信学习改进的带噪学习文本分类方法。首先介绍了一下置信学习的算法流程及其思路，然后分析了其方法可改进的地方，然后提出了一种适用于小样本场景的带噪学习文本分类方法。

第五章：总结与展望。对本论文的研究内容进行系统总结，对目前工作中存在的实际价值和不足进行了剖析，并对以后可优化之处进行了展望。

中国知网 https://www.cnki.net

第2章 文本分类算法理论基础

2.1 深度神经网络

2.1.1 卷积神经网络

卷积神经网络（Convolutional Neural Network, CNN）[24]是一种深度学习模型，主要应用于图像识别，语音识别、自然语言处理等领域。卷积神经网络以多层卷积操作与池化操作为核心思想，从图像、序列或其他数据抽取特征，把这些特征转移到全连接层，以完成分类，回归和其他工作。

卷积神经网络主要部件为卷积层，它利用卷积核（Kernel）对输入数据进行特征提取。传统的卷积核都采用了固定大小的卷积核，但随着网络规模增大，这种方法已经无法满足大规模处理的需要。卷积核可看作是权重的集合，通过滑动输入数据，对卷积操作进行计算，所得结果即为特征图。为减少计算量，卷积神经网络一般在卷积层的后面加上池化层，在保持重要特征的前提下使特征图减小。

除卷积层，池化层外，卷积神经网络进一步包含激活函数、全连接层和其他部件。通过对激活函数进行修改，可将其应用于不同的网络环境。激活函数能够提高模型非线性能力，并利用全连接层输出最后分类结果。

整体上卷积神经网络作为一个高效的模型能够在图像，语音和文本等方面的技术应用有较为出色的表现。

2.1.2 循环神经网络

循环神经网络（Recurrent Neural Network，RNN）是一种深度学习模型，主要用于处理序列数据。与传统的前馈神经网络（Feedforward Neural Network）不同，循环神经网络可以利用过去的状态信息来处理当前的输入，使得模型能够记忆之前的信息并将其应用于当前的任务中。

在循环神经网络中，每个时刻的输入会被传递给一个隐藏层，同时这个隐藏层的输出也会被传递到下一个时刻。这个过程可以被视为对序列中状态的递归处理。为了增强模型的记忆能力，循环神经网络通常会使用一种称为“长短时记忆网络”（Long Short-Term Memory，LSTM）的结构，它可以针对长序列中的梯度消失问题进行优化。

与卷积神经网络不同，循环神经网络中的权重是在时刻上共享的，这意味着模型可以对任意长度的序列进行处理。另外，由于循环神经网络可以自然地处理序列之间的依赖关系，因此它在语音识别、自然语言处理、时间序列预测等任务中表现良好。

总的来说，循环神经网络可以处理序列数据的各种任务，在许多领域都有广泛的应用。

中国知网 https://www.cnki.net

2.1.3 Transformer 与注意力机制

Transformer[36]是一种基于注意力机制的神经网络模型，由 Google 在 2017 年提出，用于自然语言处理领域中的机器翻译任务。与传统的循环神经网络（RNN）和卷积神经网络（CNN）不同，Transformer 不依赖于序列顺序，可以一次性处理整个序列，大大加速了模型训练和推理的速度。

Transformer 广泛使用了注意力机制，用于对输入序列中的不同位置进行加权，以便更好地捕捉序列中的关键信息。Transformer 中的注意力机制可以分为自注意力机制和多头注意力机制两种。

自注意力机制（Self-Attention）是指在一个序列中，每个位置都可以与其他位置计算相似度，通过计算相似度得到每个位置的权重，从而确定每个位置的重要程度，然后将这些位置的特征加权求和作为输出。自注意力机制使得模型可以自动关注输入序列中的不同部分，并且可以处理变长的输入序列。

多头注意力机制（Multi-Head Attention）是指将输入序列分别进行不同的线性变换，然后进行自注意力机制，最后将得到的结果拼接在一起。通过多头注意力机制，模型可以将不同的特征组合在一起，从而更好地捕捉输入序列中的关键信息。

总的来说，Transformer 模型通过注意力机制的设计，使得模型可以快速、高效地处理序列数据，在自然语言处理等领域中取得了非常优秀的成果。

2.1.4 BERT 深度神经网络

BERT（Bidirectional Encoder Representations from Transformers）[37]是由 Google 在 2018 年提出的一种预训练语言模型，采用基于 Transformer 的架构。BERT 采用双向 Transformer 编码器，通过预训练模型来获取通用的语义表示，进而可以在各种下游任务中进行微调，例如文本分类[39]、命名实体识别[25]等。BERT 模型是一种非常强大的预训练语言模型，在刚提出时甚至在 11 项 NLP 任务中取得了当时的最佳结果，令人吃惊。很多学者围绕 BERT 也开展了研究。

BERT 模型的预训练过程采用了两个任务。第一个任务是 Masked Language Model（MLM），即遮盖语言模型。该任务将输入序列中的一些 token（词符）随机遮盖，然后让模型预测遮盖 token 的原始内容。这个任务可以使模型学习到对语言中单词的上下文和共现关系进行建模的能力。第二个任务是 Next Sentence Prediction（NSP），即下一句预测。该任务旨在检测模型是否能够正确地理解两个句子之间的关系。模型输入两个句子，然后预测这两个句子是否是连续的。

BERT 模型的优点在于它可以通过预训练来学习通用的语言表示，从而可以在各种下游任务中进行微调，而无需从头开始训练。此外，由于采用了双向 Transformer 编码器，BERT 可以有效地处理长文本，同时遮盖语言模型的设计也可以增强模型对上下文的理解。

中国知网 https://www.cnki.net

本文中使用 BERT-base-uncased 模型。模型结构包括一个词向量嵌入层和一个 12 层 Transformer 隐层的 Encoder 编码器。每个隐层的输入输出都是 768 维张量，12 个自注意力头，共 110M 参数量。

2.2 算法原理及发展

2.2.1 损失函数以及反向传播算法

损失函数（Loss Function）是深度学习模型中的一个重要组成部分，用于衡量模型预测结果与真实值之间的差异程度。训练的过程会尽量使模型的预测结果逐渐接近真实值，因此需要定义一个损失函数来量化预测结果的误差。

常见的损失函数包括均方误差（Mean Squared Error，MSE）、交叉熵损失（Cross-entropy Loss）、对数损失（Log Loss）等。不同的损失函数适用于不同的场景和任务。

均方误差（MSE）是一种用于回归问题的损失函数，它测量预测值与真实值之间的差距的平方和。交叉熵损失（SCE）是一种用于分类问题的损失函数，它测量分类器预测的概率分布与实际标签的概率分布之间的差异。对数损失（LL）也是一种常用的分类问题损失函数，它测量分类器的预测值与实际标签之间的差距。

在深度学习模型中，损失函数的选择和设计是非常重要的，不同的损失函数可以影响模型的学习效果和性能。在训练过程中通常使用反向传播算法来计算损失函数的梯度，并使用优化算法来调整模型参数，使得损失函数的值最小化，以达到提高模型泛化能力的目的。

反向传播算法（Back propagation）是一种深度学习模型中常用的优化算法，用于计算损失函数关于模型参数的梯度。反向传播算法的基本思想是通过链式法则（Chain Rule）将损失函数的梯度向后传递，从而计算神经网络中每个参数对损失函数的贡献程度。

反向传播算法分为两个阶段：前向传播和反向传播。在前向传播阶段，输入数据通过网络中的各个层，计算出网络的输出结果。在反向传播阶段，首先计算损失函数关于输出结果的梯度，然后将梯度向前传递，计算每一层的梯度，最终计算出损失函数关于每个参数的梯度。具体来说，在反向传播算法中，需要计算每个参数对损失函数的偏导数，然后使用优化算法（如梯度下降）来更新参数。由于反向传播算法可以自动计算梯度，因此可以大大减少手动计算梯度的工作量。

总的来说，反向传播算法是深度学习模型中非常重要的优化算法，它可以通过计算梯度来更新模型参数，使得模型能够更好地拟合数据，提高模型的泛化能力和性能。图 2.1 描绘了模型一般的参数更新步骤。

中国知网 https://www.cnki.net

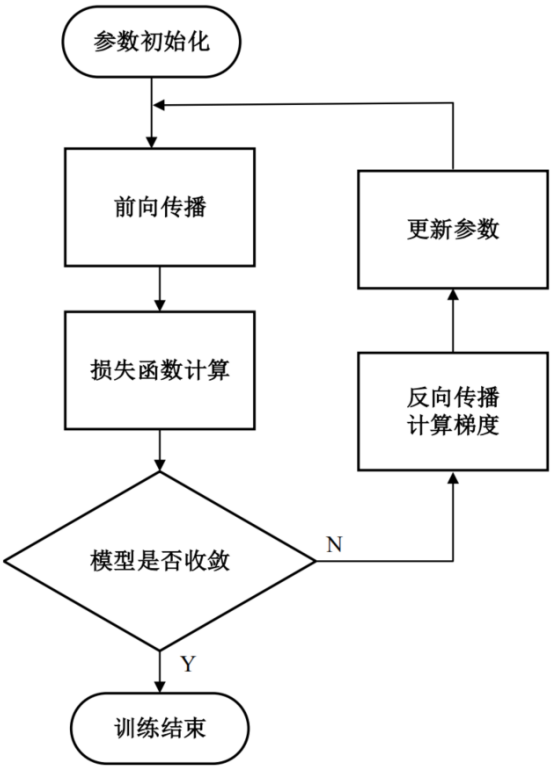

图 2.1 一般模型训练过程

2.2.2 对抗学习方法

Goodfellow[40]在 2015 年提出了对抗性训练的概念，即对原始输入样本加入扰动，构建对抗性样本进行训练。

对于扰动，Goodfellow 认为，由于具有线性特性，神经网络容易受到线性扰动的影响。因此，他提出了 FGSM 方法（Fast Gradient Sign Method）[40]，利用损失相对于输入的梯度来计算输入数据的扰动，从而构建对抗性样本。通过对对抗性样本的训练，可以在一定程度上提高模型的鲁棒性、预测精度和泛化能力。

Madry[41]总结了以前的工作，从优化的角度重新定义了向模型添加扰动的过程为寻找鞍点的问题。与 Goodfellow 的“一步到位”不同，Madry 认为“小步走，多走几步”可以产生更强的扰动，并提出了 PGD 方法（Project Gradient Descent, 梯度投影下降）[41]，两者对比如图 2.2 所示。

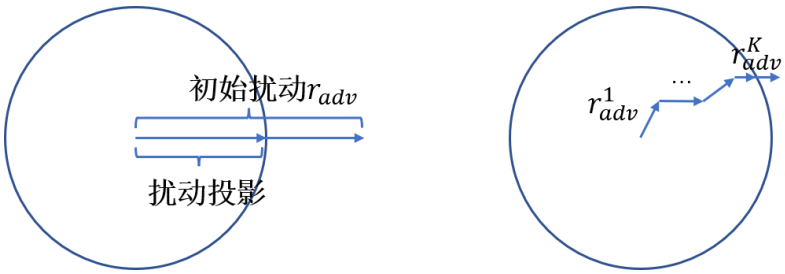

图 2.2 FGSM（左）与 PGD（右）对比示意图

中国知网 https://www.cnki.net

其他一些学者也对 PGD 做了很多修改，比如 FreeAT 方法（Free Advertising Training）[42]。FreeAT 的想法是同步更新扰动和模型的参数，这样总的训练时间与传统网络的训练时间几乎相同。然而，这种情况下的训练数据集都是对抗性样本，没有干净的样本。

上述方法主要用于 CV，其输入数据是连续的 RGB 值。在 NLP 中，人们一般会在嵌入词向量中加入扰动来构建对抗性样本，这样可以提高模型的泛化性能。

Miyato 等人[43]在 2017 年将 FGSM 的思想应用于文本分类领域，在文本输入数据中加入虚拟对抗性扰动。

Zhu 等[44]学者指出了 FreeAT 的问题，即每个扰动对当前参数都是次优的，因此提出了类似 PGD 的方法，Freelb（Free Large-Batch）。与 PGD 不同的是，PGD 采取最后一次迭代的梯度来计算扰动，Freelb 采取多次迭代的平均梯度。

上述对抗性学习方法中，基本上都是从梯度上升的思路出发，一般有两个方向可改进：一是简化梯度过程中的计算；二是优化扰动的效果，提升模型泛化性能。本文在第三章中的方法旨在提高对抗性扰动的效果，而不关注模型的训练速度。

2.2.3 半监督学习

半监督学习的研究是建立在缺乏标注数据的基础上的。一致性正则化和熵最小化是半监督学习方法中最常使用的。

2.2.3.1 一致性正则

一致性正则化是指对于未标记的数据，无论有无扰动，模型都会输出相同的输出。π 模型[45]充分运用了一致性正则化的思想。在无标签数据中加入高斯噪声两次后，通过前向推理分别得到两个预测概率分布。两个分布之间的差距用平均平方误差（MSE）损失来衡量，通过减少差距来提高模型在扰动下的一致性。Temporal Ensembling[45]使用当前模型预测值和历史预测值的平均值来计算 MSE 损失。这种方法有效地保留了历史信息并稳定了当前值。

上面提到的对无标签数据添加扰动的方式[46,47]进行数据扩增，主要是使用简单的随机噪声。相关研究发现，数据扩增方式对模型的性能提升有重要影响。无监督数据增强（UDA）[48]是针对无标签数据提出的采用更多样化、更真实的数据增强方法，如回译、TF-IDF（term frequency–inverse document frequency）词汇替换等。

虚拟对抗训练（Virtual Adversarial Training，VAT）[49]与 π 模型方法类似，仍采用一致性正则化的方法。作者认为 π 模型中的随机噪声不能模拟复杂情况的输入。VAT 在类似于 FGSM 的对抗训练方式中加入扰动，并选择 KL 散度作为损失函数。

2.2.3.2 熵最小化

熵最小化意味着决策边界应尽可能地通过数据稀疏的地方，以避免将密集的样本点分割到决策边界的两侧。

伪标签的应用遵循了这一思路。伪标签使用模型来生成标签，与带标签的数据混合，以提

中国知网 https://www.cnki.net

供额外的信息。Self-training[50]使用用少量标注样本训练的模型为未标注的数据分配伪标签，并使用伪标签进行监督学习。然而，有学者指出，模型对自己预测结果的盲目自信可能会放大预测偏差。

伪标签的产生往往伴随着大量的错误预测。许多算法，如 UDA[48]和 Fixmatch[51]，为无监督损失设置了高的、固定的阈值，以选择具有较高信心的伪标签。每次只选择置信度超过一定阈值的预测。然而，当批次太小或有许多困难的数据时，这种方法可能导致数据的浪费。因此，Flexmatch[52]构建了动态置信度阈值，该阈值随着每个类别的学习效果而动态变化。

2.2.4 数据增强

考虑到文本数据中信息表达的复杂性和离散性，NLP 中的数据增强操作具有一定的挑战性。数据增强的目的是获得具有类似含义但表达方式不同的文本。有些研究直接针对句子本身的变化。Wei 等人[53]使用随机交换、同义词替换、随机删除、随机插入等方法。尤丛丛等人[54]和 Xie 等人[48]以词汇替换和回译[55]的形式对数据进行一致性训练。一些学者基于词的嵌入对文本数据进行增强。VAT 和其他一些方法[56,57]以对抗的方式产生扰动，用于数据增强。π 模型中的随机高斯噪声在 NLP 中也有一定影响。

目前，大多数 NLP 数据增强方法直接作用于文本本身或多层模型的嵌入层。因为文本本身的信息表达是离散的，替换句子中的个别单词可能真的会改变句子的含义。因此，传统的词语替换和顺序变化的方法可能仍有改进的余地。对于更广泛使用的回译方法，它在很大程度上依赖于额外的机器翻译工具，但其成本甚至可能超过数据标注本身。

研究[58]表明，通过对隐藏层的两个语义表征张量进行插值解码，可以生成一个新的句子，而新的句子包含了两个原始句子的混合语义信息。这也意味着，BERT[37]等多层语言模型可以对文本的语义信息进行连续编码。这时，图像中的一些数据增强方法可以应用于这些语义表示张量。

多层语言模型可以用来编码句子语义信息。本文中的数据增强方法试图作用于隐藏层的语义表示张量。

2.2.5 带噪学习

由于实际场景中的数据质量有时无法保证，因此噪声学习受到了广泛关注。深度神经网络（Deep Neural Network, DNN）在训练过程中总是学习噪声标签中的信息，这干扰了神经网络的预测精度。考虑到经验风险最小化方案的稳健性，一些研究人员试图设计稳健的损失函数。在有噪声标签的情况下，未知的干净数据的损失仍然非常小，从而模型的训练和参数更新不会受到太大的影响。

在分类任务中，交叉熵（CE）具有很强的泛化能力和收敛性。但在面对噪声标签时，平均绝对误差（MAE）[59]由于其泛化能力，表现得比 CE 更好。但是 MAE 在处理复杂的数据时并不擅长。同时基于 CE 和 MAE，广义交叉熵（GCE）[60]被提出。为了使 CE 更加对称，

中国知网 https://www.cnki.net

Wang 等人[61]将反向交叉熵（RCE）作为噪声容忍项与交叉熵相结合，构建了对称交叉熵（SCE）。该损失函数被用来提高模型对噪声标签的鲁棒性。在 SCE 中，CE 和 RCE 的权重系数都是固定的超参数，它们对 SCE 的性能有很大影响。然而，在模型训练的过程中，不同时期对数据的过拟合程度和对噪声标签的敏感性是不同的。更为灵活的方式是动态改变权重系数，随着模型的变化而调整，从而进一步提高整个训练过程的鲁棒性。

2.2.6 小样本学习

小样本学习（Few-shot learning，FSL）[62,63,64]任务中，一个分类器在只有少量样本的情况下，被训练适用于新样本的分类。这种情况下传统的机器学习算法往往难以取得好的效果。小样本学习是一种比较新的机器学习领域，旨在解决这个问题。

小样本学习的主要方法是利用已有的知识来辅助学习。例如，利用迁移学习的思想，将已经学习好的模型应用到新的任务中，或者利用元学习的思想，从已有的任务中学习出一些通用的知识，用于新任务的学习。Zhang 等人[65]利用匹配网络（matching networks）在标记的数据集（支持集）上训练使用注意力机制来预测未标记点（查询集）的类。汪雨竹等人[66]基于元学习的思路，并且结合卷积神经网络和 Transformer 编码器，构建了一个三阶段表示学习模型。

小样本学习方法在进行训练时，通常假设训练集是干净的，具有准确标记的样本。但这种假设往往并不现实：训练集无论多么大小，仍然可以包含错误标记的样本。因此，对于小样本学习方法的实际应用来说，模型训练过程对标签噪声的鲁棒性至关重要。

一些小样本学习算法，例如基于“原型”的算法、基于“关系”的算法等等，都是为了更好地利用有限的训练样本来提高模型的泛化能力。为了解决小样本学习中的错误标记样本，Liang[67]等学者提供了简单有效的特征聚合方法，改进了“原型”网络（ProtoNet）这种流行的小样本学习技术的“原型”，并且提出了一种新颖的少样本学习模型 TraNFS。

2.3 评价指标

对于文本分类任务的衡量，有多种指标。一方面是对于文本分类任务本身的评价，包括通过标签噪声比例体现的分类任务的难度等；另一方面是对任务完成结果的评价，包括训练所得的模型的预测精度。

对于文本分类任务，本文的评价指标主要针对训练数据集，包括数据集中带标签样本数量和无标签样本数量，带标签样本中的标签错误率。为了测试不同训练方法在不同标签错误率下的表现，本文实验中会在带标签样本中按照一定比例添加随机错误标签。

本文中的实验会在不同标签错误率的条件下，测量不同方法训练出的模型在测试集上的准确率（Accuracy），并以准确率作为衡量模型训练效果的指标。由于模型训练过程中模型参数初始值、训练数据随机顺序等偶然因素都会对最后模型在测试集上的准确率一定影响，所

中国知网 https://www.cnki.net

以本文在实验中固定这些偶然因素的随机种子，分别取多次随机种子实验后平均的结果来衡量模型训练的效果。

2.4 本章小结

本章首先对不同神经网络的结构和发展情况进行了描述，其中包括 CNN 卷积神经网络、RNN 循环神经网络、Transformer 注意机制和 BERT 模型的构造等等；接着对损失函数和反向传播算法、对抗学习、带噪学习、半监督训练、数据增强等算法以及各自领域的相关成果做了一定阐述；最后，对文本分类问题中的有关参数及评价指标进行描述，包括文本分类实验中的不同类别样本数量、标签错误率、模型测试准确率等评价标准。

中国知网 https://www.cnki.net

第 3 章 ASSL：一种对标签噪声鲁棒的对抗半监督学习方法

本章阐述了一种对标签噪声鲁棒的对抗半监督学习方法（Adversarial Semi-Supervised Learning，ASSL），用于标记数据不足和有噪声的文本分类任务。本章一方面提出了 MHGD 对抗性学习方法用于文本数据增强，以保持整体语义信息的同时提高正则化性能；另一方面构建了 Flex-SCE 损失函数，其中交叉熵项被赋予一个按照一定规则衰减的权重系数，以提高对噪声标签的鲁棒性。

3.0 ASSL 系统框架

ASSL 方法分为带标签学习和无标签学习两部分。无标签学习利用一致性正则的思路，在 BERT 模型的某一 Transformer 隐层中添加最大梯度扰动后得到预测结果，并鼓励该预测结果靠近无扰动结果。带标签学习过程利用 Flex-SCE 损失函数显著降低标签噪声对模型参数更新的影响。ASSL 的整体框如图 3.1 所示，本文的主要符号见表 3.1。

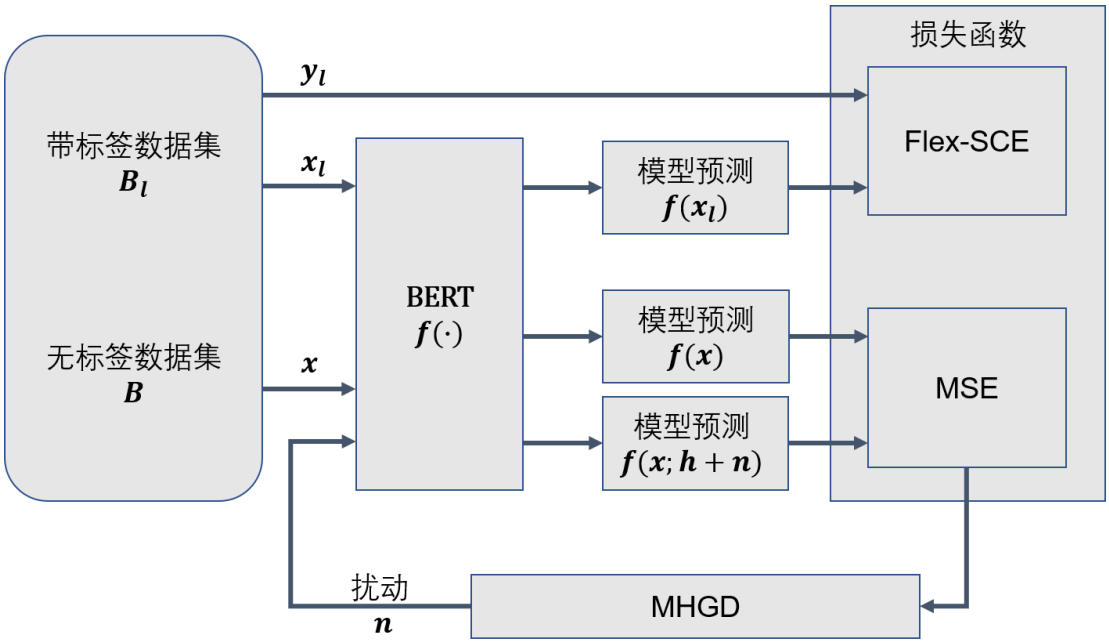

图 3.1 ASSL 方法整体框架

对于一个 c 类文本分类任务，指定 $B=\{(x_i,y_i)|i=1,2,3,...,b\}$ 作为一个批次训练数据 (x_i,y_i)，其中 b 表示批次中数据数量。标签 $y_i\in\{0,1,2,\cdots,c-1,NU\}$，其中 NU 是无标签数据的标记。指定 $B_l=\{(x_{li},y_{li})|(x_{li},y_{li})\in B\ \&\ y_{li}\neq NU\}$ 为带标签数据。ASSL 整体框架来自于一致性正则和鲁棒性损失的思想。主要过程可以分为两部分：在 B_l 上的标记训练和在 B 上的无标签训练。

中国知网 https://www.cnki.net

表 3.1 符号及意义总结

符号	意义
c	样本的种类数量
B	一批次样本
x	输入数据
y	样本标签
h	模型某隐藏层输出的语义表达张量
r	高斯随机扰动张量
n	MHGD 扰动张量
$f(x_i)$	不加扰动的模型预测概率分布
$f(x_i; h_i + n_i)$	在隐藏层添加 MHGD 扰动的模型预测概率分布
$f(x_i; h_i + r_i)$	在隐藏层添加高斯随机扰动的模型预测概率分布
K	MHGD 方法中的反向传播次数
$H(\cdot)$	损失函数
θ	模型参数集合
g	语义表达张量对输入数据求梯度
η	MHGD 扰动的反射参数
ε	MHGD 扰动的改变量
N	模型训练轮次数量
t	模型训练轮次索引

3.1 最大隐藏梯度下降

对抗学习的一个重要问题是关于在一定范围内扰动能否更强更有效。Madry[41]将添加扰动的过程重新定义为一个MIN-MAX问题。如公式(3.1)所示，内部最大化代表找到最强的扰动，使损失最大化，也就是“攻击”；外部最小化代表经验风险最小化，也就是“防御”。其中 S 是扰动的范围， D 是输入样本的分布。

$$\min_{\theta} E_{(x,y)\sim D}\left[\max_{r_{adv}\in S} H\left(f(x+r_{adv}), y\right)\right] \tag{3.1}$$

在语义不发生过度变化的前提下，使损失变大的扰动可以带来更好的训练效果，这也符合对抗学习的MIN-MAX问题的准则。

对于文本数据的增强，本文提出了“最大隐藏梯度下降”方法（Maximum Hidden Gradient Descent, MHGD），它遵循PGD的梯度上升思想。在每一轮的梯度上升中，前一轮的对抗性损失被用来生成当前轮次的扰动。然后，由于当前轮次前向推理过程中增加的扰动，当前轮次的损失很可能高于前一轮的损失。最终目标是根据MIN-MAX问题，在梯度上升的迭代过程中使损失最大化。然而，如图3.2所示，在梯度上升的整个过程中，损失呈现出波动上升的趋势。PGD采用的最后一次扰动不一定是最强的扰动，也不一定表现出最佳的泛化性能。所以MHGD实时更新最大损失。最大损失对应于整个过程中最强的扰动，这样MHGD就相对完成了损失最大化的任务。

中国知网 https://www.cnki.net

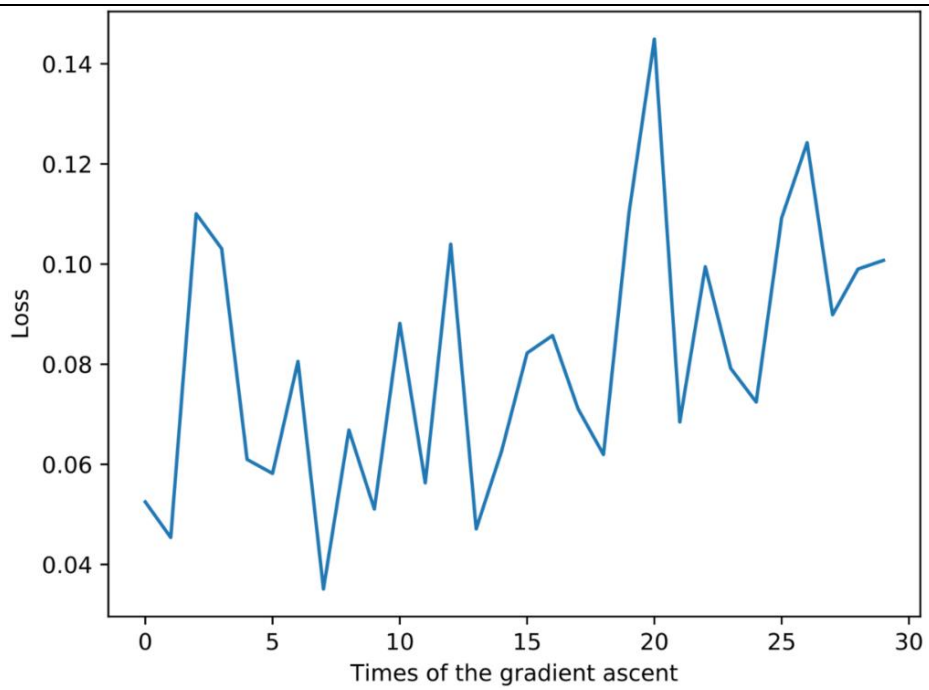

图3.2 PGD过程中的损失变化

然而，最强扰动的最佳效果有一个前提，即有扰动的句子仍然具有与原句大致相同的语义信息。目前，大多数的文本数据增强都是直接对句子中的词语进行变换或对词向量进行扰动。在训练的早期阶段，词向量一般是随机的或在多维空间中具有聚类属性的向量。如果词向量是由随机值组成的，直接添加到词向量中的扰动可能会使一些词的含义或整体语义信息发生巨大变化。如果初始的词向量具有聚类的性质，添加扰动的操作就类似于在词向量空间中寻找近义词。由于语言本身的特点，一些词的替换可能会导致语义的过度变化。因此，这种类似于近义词替换的操作仍然存在着语义剧变的风险。

为了保证最强的扰动不会导致对句子表达信息的过度改变，本文将MHGD应用于多层模型中隐藏层的语义表示张量。

BERT-base-uncased模型的结构包括一个词向量嵌入层，一个12层的Transformer结构，以及一个全连接结构的Pooler层。

对于一个句子，我们可以用多层神经网络编码语义信息，并获得语义表达张量，在此基础上进行最终预测。一些学者研究了BERT不同隐藏层的作用，发现一些隐藏层具有更强的表达能力，这些层掌握了从语法到语义层面的不同信息表达形式。

在本文中，MHGD被应用于BERT的特定隐藏层的语义表示张量。这可以提高扰动强度并尽可能地保证语义的相似性。

MHGD的伪代码如表3.2中的**算法1**。

中国知网 https://www.cnki.net

表3.2 MHDG方法伪代码

算法 1 MHGD 方法

输入：一批带标签和无标签的样本 $B=\{(x_i,y_i)\,|\,i=1,2,3,\cdots,b\}$。反向传播的数量 K。损失函数 $H(\cdot)$。初始噪声 n^0。

1： $l_m \leftarrow 0, n \leftarrow n^0$

2： $l_{adv} \leftarrow \frac{1}{b}\sum_{i=1}^{b} H\left(f(x_i), f\left(x_i; h_i+n_i\right)\right)$

3： **for** k **in** $[1,K]$ **do**

4： $g=\nabla_h l_{adv}, n^k=\varepsilon\frac{g}{\|g\|_2}$

5： $n \leftarrow n+n^k$

6： **if** $\|n\|_2>\eta$ **then**

7： $n \leftarrow \eta\frac{n}{\|n\|_2}$

8： $l_{adv} \leftarrow \frac{1}{b}\sum_{i=1}^{b} H\left(f(x_i), f\left(x_i; h_i+n_i\right)\right)$

9： **if** $l_{adv}>l_m$ **then**

10： $l_m \leftarrow l_{adv}$

输出：最终损失 l_m

算法 1 介绍了在训练一个批次时的 MHGD 过程。

首先在模型进行一次初始化前向传播，初始化损失 l_{adv}。然后，MHGD 重复以下过程 K 次：计算损失 l_{adv} 相对于语义表达张量 h 的梯度 g，然后计算得到扰动 n_k，将其加入总扰动张量 n。在判断是否进行如在第 6-7 行所示的投影操作后，进行本轮此前向传播，前向传播的过程中将总扰动加到特定隐藏层的输出张量 h，前向传播后更新本轮此的损失 l_{adv}。最后更新最大损失 l_m，最大损失 l_m 则产生于目前最强的扰动。

在随后的实验中，我们发现如果将 K 设置为一个固定值，那么从一开始就最强的扰动可能会对训练初期的参数优化产生负面影响。因此，可以将 K 设置为一个动态增加的值，使扰动的强度随着训练过程的进展而增加。

3.2 损失函数

对于带标签的数据，考虑到标签的准确性可能并不理想，在半监督学习的过程中，模型更可能在少量的带标签的数据上过度拟合。本文提出了 Flex-SCE 作为标注数据的损失函数。Wang 等人[61]将作为噪声容忍度项的反向交叉熵与交叉熵相结合，构建了对称交叉熵（SCE），如公式(3.1)所示。这两个项的权重参数分别为 α 和 β。

$$H_{SCE}\left(f\left(x_{li}\right),y_{li}\right)=\alpha H_{CE}\left(f\left(x_{li}\right),y_{li}\right)+\beta H_{RCE}\left(f\left(x_{li}\right),y_{li}\right). \tag{3.1}$$

中国知网 https://www.cnki.net

然而，SCE 的实际性能对参数α很敏感。根据 Wang 等人[61]的实验，参数α对收敛和正则化性能之间的平衡有很大影响。因此，在训练过程中，本文构建了$\alpha(t)$来做出动态的调整，其中t在$[0, N-1]$。在本文中，$\alpha(t)$是一个逐渐递减的权重。Flex-SCE 如下：

$$\begin{aligned} & H_{Flex-SCE}\left(f\left(x_{li}\right), y_{li}\right) \\ & =\alpha(t) H_{CE}\left(f\left(x_{li}\right), y_{li}\right)+\beta H_{RCE}\left(f\left(x_{li}\right), y_{li}\right) \\ & =\alpha(t) y_{li} \cdot \log\left(f\left(x_{li}\right)\right)+\beta f\left(x_{li}\right) \cdot \log\left(y_{li}\right). \end{aligned} \tag{3.2}$$

当标签y_{li}被扩展为 one-hot 形式时，它必须包含元素 0，所以负的无限项$\log 0$必定会出现在反向交叉熵的计算中。设超参数$A=\beta log 0, A<0$。设$f\left(x_{li}\right)_y$为元素 1 在 one-hot 标签y_{li}对应位置的概率。Flex-SCE 损失函数可以表示为：

$$\begin{aligned} & H_{Flex-SCE}\left(f\left(x_{li}\right), y_{li}\right) \\ & =\alpha(t) y_{li} \cdot \log\left(f\left(x_{li}\right)\right)+A\left(1-f\left(x_{li}\right)_y\right). \end{aligned} \tag{3.3}$$

本文构造了交叉熵的动态权重参数$\alpha(t)$，表示为：

$$\alpha(t)=\alpha_0\left[1+\lambda\left(\sin(\gamma)-\sin\left(\frac{t}{N}\left(\frac{\pi}{2}-\gamma\right)+\gamma\right)\right)\right]. \tag{3.4}$$

超参数$\gamma \in(0, \pi / 2)$，　$\lambda \in(0,1]$，　$\alpha_0 \in(0,+\infty)$。

$$\frac{\partial \alpha}{\partial t}=-\left(\frac{\pi}{2}-\gamma\right)\frac{\alpha_0 \lambda}{N}\cos\left(\frac{t}{N}\left(\frac{\pi}{2}-\gamma\right)+\gamma\right)<0. \tag{3.5}$$

如公式(3.5)所示，随着训练的进行，$\alpha(t)$从α_0到$\alpha_0[1+\lambda(\sin(\gamma)-1)]$逐渐减小。在训练的后期，模型更容易受到有噪声标签的影响，模型的预测精度更高，所以这里逐渐降低带标签数据的影响，避免对带标签数据的过拟合，对模型的预测投以更高的信任。

超参数γ控制了$\alpha(t)$减小变化的程度。如图 3.3 所示，在 0~$\pi/2$范围内，γ越小，$\alpha(t)$下降的程度越深。对于不同难度和不同应用场景的数据集，可以通过调整γ来调整$\alpha(t)$的变化程度。对于质量较高、训练效果较好的数据集，我们可能会对γ较大的交叉熵赋予较高的信任度；当数据质量无法保证或同质化程度较高时，可尝试适当降低γ。

中国知网　https://www.cnki.net

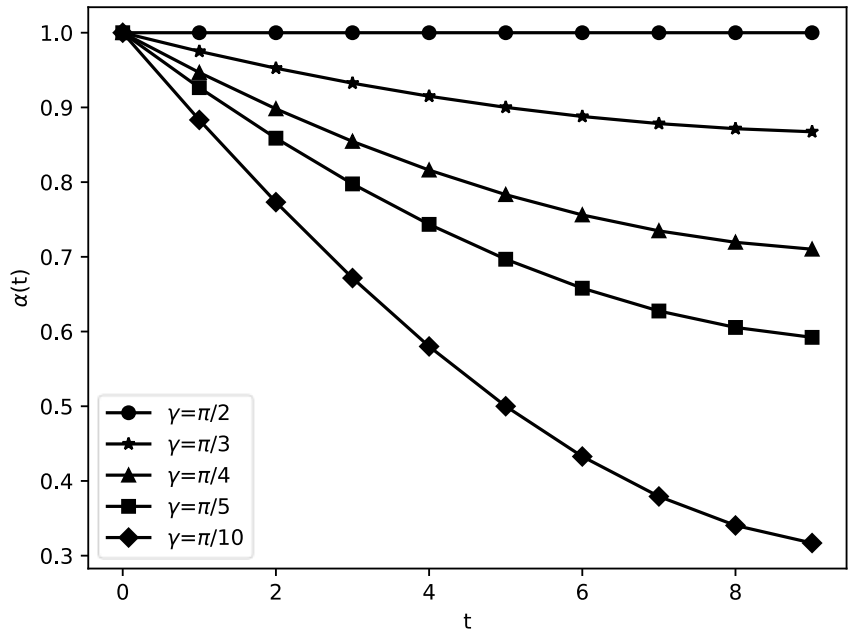

图 3.3 随着 γ 的变化，$\alpha(t)$ 的不同下降趋势（其中 $\alpha_0=1.0,\lambda=1.0,N=10$ ）

在训练初期，模型对数据的分布情况了解较少。数据中含有未知信息，模型参数会快速变化。在训练末期时，可以学到的知识较少，模型已经有了较为准确的先验预测。参数相对稳定，参数变化率逐渐减小。如公式(3.4)、(3.5)和图 3.3 所示，根据模型参数在训练过程中的变化规律，$\alpha(t)$ 的变化率也呈现出随参数的增大而减小的趋势

$$\frac{\partial^2\alpha}{\partial t^2}=\left(\frac{\pi}{2}-\gamma\right)^2\frac{\alpha_0\lambda}{N}\sin\left(\frac{t}{N}\left(\frac{\pi}{2}-\gamma\right)+\gamma\right)>0. \tag{3.6}$$

对于无标签数据，本文选择 MSE 作为损失函数。许多半监督学习方法，如 UDA 和 Fix-match，通过设置置信度阈值对样本进行筛选，将其转换为 one-hot 标签或锐化的概率分布，然后结合另一个预测计算损失。一方面，这样的操作可能会加剧错误预测的误导性。另一方面，可能导致对困难样本的忽略和欠拟合，也可能容易形成对简单样本的模型过拟合。因此，本文放弃筛选过程，选择使用高斯 ramp-up 曲线构造无标签损失项并赋予一个逐渐增大的权重系数 $w(t)$ 。损失函数计算如下：

$$w(t)H_{\mathrm{MSE}}\left(f\left(x_i\right),f\left(x_i;h_i+n_i\right)\right)=w(t)\frac{1}{c}\left\|f\left(x_i\right)-f\left(x_i;h_i+n_i\right)\right\|_2^2. \tag{3.7}$$

$$w(t)=\exp\left[-\mu\left(1-\frac{t}{N}\right)^2\right]. \tag{3.8}$$

$t\in[0,N-1]$，超参数 $\mu>0$ 。$\|\cdot\|_2^2$ 表示张量中各个元素的平方和。$w(t)$ 从 $\exp(-\mu)$ 逐渐增加到 $\exp(-\mu/N^2)$ 。因此，对模型预测的信任度逐渐增加。

中国知网 https://www.cnki.net

3.3 ASSL 半监督学习过程

基于 MHGD 和上述损失函数，本节给出 ASSL 的详细过程。

对于 B_l 上的带标签训练，预测概率分布 $f(x_{li})$ 和对应标签 y_{li} 之间的 Flex-SCE 损失为 $H_{Flex-SCE}(f(x_{li}),y_{li})$，该损失对噪声标签具有鲁棒性。Flex-SCE 的计算公式如下：

$$\frac{1}{|B_l|}\sum_{i=1}^{|B_l|}H_{Flex-SCE}\left(f(x_{li}),y_{li}\right). \tag{3.9}$$

对于 B 上的无标签训练，利用一致性正则的思想，有扰动和无扰动的概率分布应尽可能相同。两个预测概率分布 $f(x_i)$ 和 $f(x_i;h_i+n_i)$ 之间的 MSE 损失为 $H_{MSE}(f(x_i),f(x_i;h_i+n_i))$。在语义表示张量中加入来自 MHGD 的对抗扰动 n_i。MSE 的计算公式如下：

$$\frac{1}{b}w(t)\sum_{i=1}^{b}H_{MSE}\left(f(x_i),f\left(x_i;h_i+n_i\right)\right). \tag{3.10}$$

具体过程的伪代码如表 3.3 所示。

表 3.3 ASSL 伪代码

算法 2：ASSL 方法流程

输入： 一批次带标签和无标签的输入数据 $B=\{(x_i,y_i)\,|\,i=1,2,3,\cdots,b\}$。一批次带标签数据 $B_l=\{(x_{li},y_{li})\,|\,(x_{li},y_{li})\in B\,\&\,y_{li}\neq NU\}$。训练轮次数量 N。高斯随机噪声张量 r。

输出： 模型参数集合 θ。

1： for t in $[0,N-1]$ do

2： for each B do

3： 使用算法 1($B\leftarrow B, K\leftarrow t, H\leftarrow H_{MSE}, n^0\leftarrow r$)，计算得到 l_m

4： $l_s\leftarrow\frac{1}{|B_l|}\sum_{i=1}^{|B_l|}H_{Flex-SCE}\left(f(x_{li}),y_{li}\right)$

5： $FinalLoss=l_s+l_m$

6： 以 $FinalLoss$ 更新模型参数 θ

3.4 实验及结果分析

3.4.1 实验模型结构与数据集

本文的实验采用了流行的 BERT 模型。本实验使用 BERT-base-uncased tokenizer 来对文本进行分词标记。句子的截断长度被设置为 32。BERT-base-uncased 模型被用作编码器。整个模型由词向量嵌入层、具有 12 层 Transformer 结构的 Encoder 编码器和两层感知器（Multilayer Perceptron，MLP）组成。嵌入层和编码器的第一个隐藏层的学习率被设定为 2e-5，从编码器的第 2 层到第 11 层隐藏层的衰减系数为 0.95，逐渐递减。对于 MLP，学习率为 1e-3。

本次实验在 AG news，IMDB，DBpedia，Best 四个数据集上取出部分数据开展文本分类半监督训练实验及其效果比较。数据构成如表 3.4 所示。

中国知网 https://www.cnki.net

表 3.4 实验数据集相关信息

数据集	种类数量	带标签样本数量	无标签样本数量
AG news	4	400	1000
IMDB	2	500	1000
DBpedia	14	280	1000
Best	4	400	1000

3.4.2 训练方法比较

- 全监督训练方法[39]。 在 BERT 的基础上，用标注的数据对 BERT-base-uncased 中的参数进行进一步微调。
- π-model 半监督训练方法[45]。通过缩小高斯噪声和 dropout 导致的两个前向运算结果之间的差距，提高了模型的一致性和鲁棒性。
- VAT 半监督学习方法[49]。对于数据输入的扰动，不再使用随机噪声，而是使用梯度上升来提高模型的鲁棒性。

3.4.3 在不同隐藏层添加扰动的效果比较

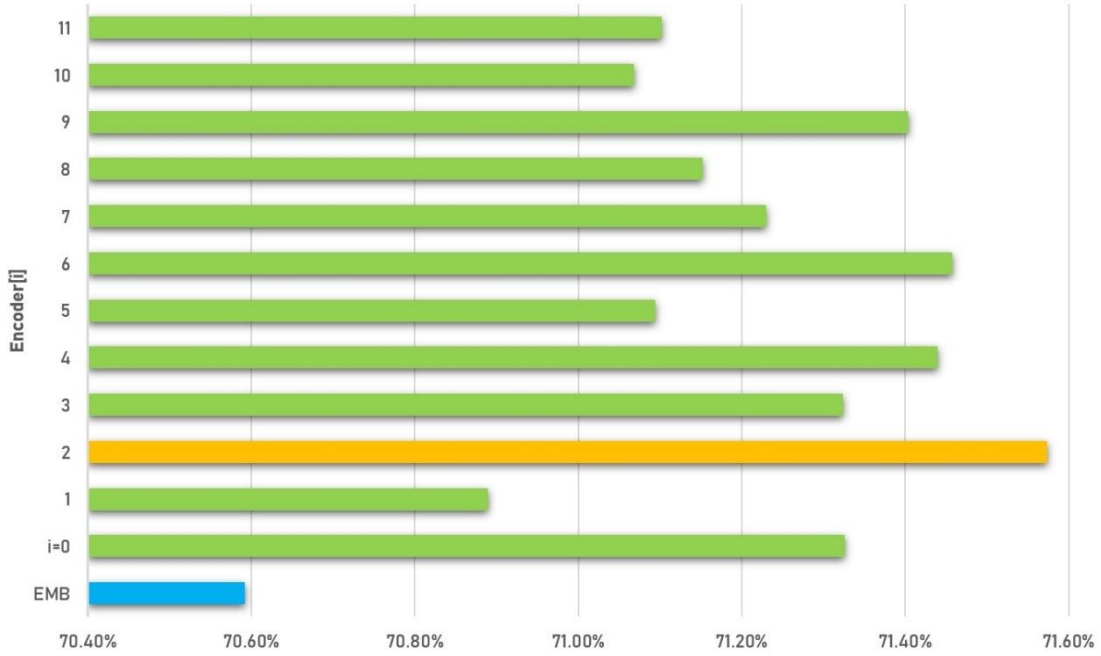

图 3.4 在不同隐藏层添加扰动的效果对比。Encoder[i]表示第“i+1”隐藏层，EMB 表示词向量嵌入层。

实验3.4.3使用AG news数据集，在编码器或嵌入层的不同隐藏层中加入MHGD扰动，探究ASSL的性能。

从图 3.4 可以看出，直接在嵌入层添加扰动效果最差，这是因为 MHGD 产生的最强扰动导致词嵌入剧烈变化，甚至影响原有的语义信息，对模型训练产生负面影响。编码器的隐藏层之间的整体效果似乎相似，但在 $i \in \{2,4,6,9\}$ 的情况下，性能更好。其中，在 Encoder[2]中

中国知网 https://www.cnki.net

加入扰动后训练的模型，即第 3 隐藏层，在测试集中的准确率最高。这也证明了对隐藏层添加扰动的效果要好于直接对词向量进行扰动。

3.4.4 消融与对比实验

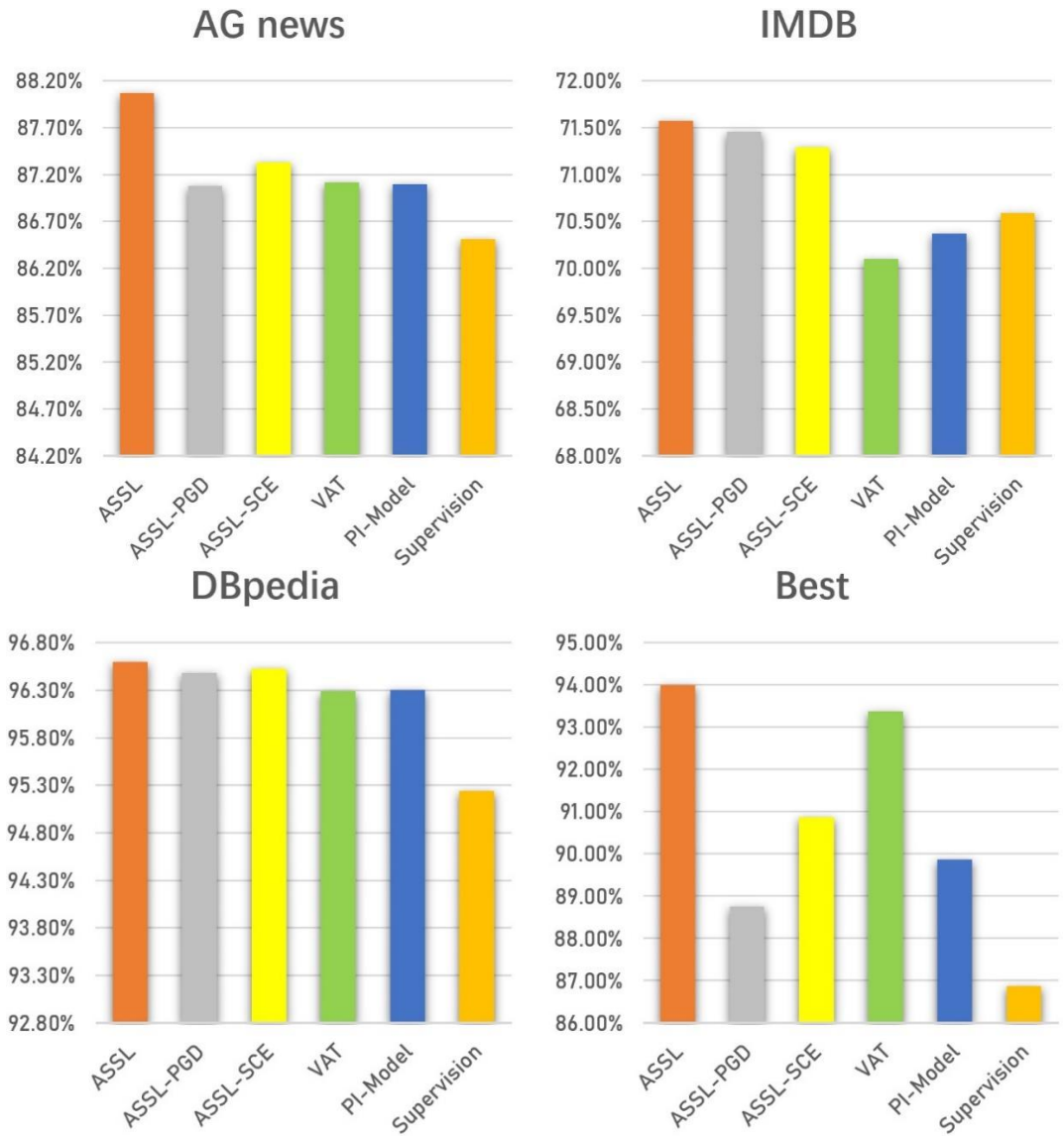

图 3.5 ASSL、ASSL 去除各个组件、其他训练方法在四个数据集上的表现

消融对比实验旨在验证 ASSL 中各组件的实际效果。实验结果如图 3.5 所示。

ASSL 代表本文提出的方法。ASSL-PGD 方法是指使用传统的 PGD 对抗方法进行数据增强，但作用于隐藏层的语义表示张量。ASSL-SCE 采用传统的 SCE 损失函数。

实验表明：

ASSL-PGD，对抗学习方法遵循 PGD 思路。由于其扰动仍然作用于隐藏层的语义表示张量，因此仍然保持句子的一般语义信息。在 IMDB 和 DBpedia 上，ASSL-PGD 取得了比 VAT、π-model 和全监督更好的效果。然而，ASSL-PGD 在 AG news 和 Best 上表现不佳，在 4 个数据集上都不如 ASSL，因为前者没有达到最强扰动。

中国知网 https://www.cnki.net

对于 ASSL-SCE，ASSL 使用 SCE 损失，其中交叉熵的权重设置为固定值。在 AG news、IMDB 和 DBpedia 上的表现优于 VAT、π-model 和全监督方法，在 Best 上表现不佳。由于对称交叉熵对交叉熵和反向交叉熵之间的平衡过于敏感，模型在训练时容易在少量带标签数据上过拟合，因此在四个数据集上其效果均低于 ASSL。

此外，在 Best 数据集上，由于标签中存在噪声，全监督方法的性能并不理想。在不增加数据标注工作的情况下，ASSL 获得了最佳的训练性能，显著提高了模型在该意图识别数据集上的预测精度。

综上所述，如果 ASSL 的某些组成部分沿用传统的方法，如 PGD 对抗方法和 SCE 损失函数，性能会立刻有所降低。与 VAT 和 π-model 等半监督学习方法相比，ASSL 在这些数据集上也取得了最好的性能，改进确实有效。

3.4.5 噪声标签测试

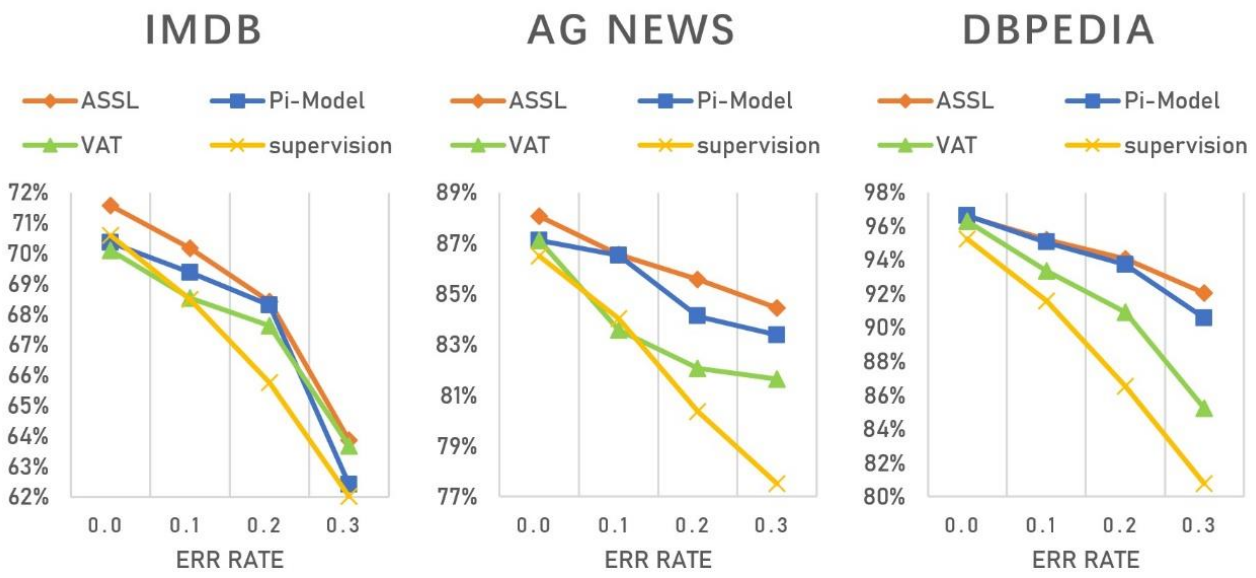

图 3.6 在不同标签噪声率下，四种方法的训练效果比较

实验 3.4.5 的目的是验证 ASSL 在数据集中面对噪声标签的性能。

本实验在带标签数据中按一定比例随机选取数据，随机改变标签，从而在数据集中添加噪声标签。错误率范围为 0%到 30%。在不同错误率的情况下，实验对比了 ASSL、VAT、π-model 和全监督训练的性能。从训练结果来看，由于噪声比例的增加，四种方法的性能都有所降低。在此过程中，ASSL 在这些方法中始终具有最高的分类精度。如图 3.6 所示，噪声标签的加入确实会对模型的训练过程和结果产生影响。然而，无论错误率如何，ASSL 总是表现出最好的性能和对噪声标签足够的鲁棒性。

3.5 本章小结

ASSL 充分利用了隐藏层中最强有效的扰动，并将其与 Flex-SCE 结合应用于半监督学习。

中国知网 https://www.cnki.net

通过使用不同的数据集进行实验，本章验证了各个组成部分对最终效果的影响。与其他方法相比，ASSL 虽然在训练速度上逊色，但在存在噪声标签的情况下具有最好的鲁棒性能，在多个数据集上取得了最高的预测精度。

中国知网 https://www.cnki.net

第 4 章 基于置信学习改进的小样本带噪学习方法

很多数据集中的标签可能存在错误，即“噪声标签”。据报道，真实数据集中标签损坏的比例从 8.0 %到 38.5 %不等。在有噪声标签的情况下，已知训练深度神经网络容易受到噪声标签的影响，因为大量的模型参数使得深度神经网络在错误标签上过度拟合，并且能够学习任何复杂的函数。深度神经网络可以很容易地用任意比例的噪声标签来拟合整个训练数据集，这最终导致了在测试数据集上的可泛化性差。

对于小样本带噪的文本分类任务，本章基于提出了一种基于置信学习改进的小样本带噪学习方法（FewCL）。为了让模型的纠错不会因为预先训练过程中的过拟合现象影响，本文采取类似 K 折验证的思路，将数据集分成多块，分块验证筛选错误标签的同时逐步对模型做训练和微调。为了适用于小样本学习场景，本文引入“原型”的思路，利用 BERT 模型提取语义并且映射到高维空间后形成语义表达张量，通过比较每条语句的高维语义表达张量和该类别的“原型”张量之间的“距离”，判断数据标签的正误。最后再给每条判断为“正确”的样本根据其“难易程度”赋予不同的权重，然后将其投入后续训练过程。

4.1 任务场景及其符号

考虑一个以 X 为特征空间，$Y=\{0,1\}^C$ 为标记空间的 C 分类问题。假设所有的标签都是 one-hot 向量，用 e_c 表示 c 类对应的 one-hot 向量。令 $\tilde{S}=\{(x_i,\tilde{y}_i),i=1,2,...,b\}$ 为在 $X\times Y$ 上按照分布 D 抽取的独立同分布样本。这里的任务是学习一个在按照分布 D 抽取的测试集上表现良好的分类器。然而，我们得到的是一个符合分布 D_η 的训练集 $S=\{(x_i,y_i),i=1,2,...,b\}$。这里的 y_i 表示噪声标签（即可能存在错误的标签），$\tilde{y}_i$ 表示正确标签，二者通过以下等式产生关联：

$$P[y_i=e_{c'}|\tilde{y}_i=e_c]=\eta_{cc'} \tag{4.1}$$

这里 $\eta_{cc'}$ 称为噪声率。这种错误标签一般将其称为类条件噪声，因为在这里标签损坏的概率依赖于原始标签。这种情况的一个特例是对称噪声，这里假设 $\eta_{cc}=(1-\eta)$ 且 $\eta_{cc'}=\frac{\eta}{C-1}$，$\forall c'\neq c$。这里，$\eta$ 表示错误标签出现的概率。在对称噪声下，被破坏的标签是其他任何标签的概率相等。

将 $\eta_{cc'}$ 表示为一个矩阵，并假设它是对角占优的（即 $\eta_{cc'}<\eta_{cc}$，$\forall c'\neq c$）。（注意，当 $\eta<\frac{C-1}{C}$ 时，对称噪声也是如此）。在这种情况下，如果取标签损坏的训练集中由特定类标记的所有数据，那么真正属于该类别的数据在该集合中仍然占多数。现在标签噪声下的鲁棒学习问题可以表述为：我们希望在一个符合分布 D_η 的训练集中，学习一个分布为 D 的分类器。相关符号及其意义如表 4.1 所示：

中国知网 https://www.cnki.net

表 4.1 本章符号及其意义

符号	意义
C	样本的种类数量
$[C]$	样本的种类集合
b	训练集样本数量
$S,\tilde{S}$	带噪训练集，无噪训练集
S_k,S_k^c	带噪验证集，带噪验证集中类别为c的样本集合
x_i	输入数据
$y_i,\tilde{y}_i$	噪声标签，正确标签
h_i	模型某隐藏层输出的语义表达张量
$f\left(x_i\right)$	模型对样本x_i的预测概率分布向量
$f\left(x_i;y_i=j\right)$	模型对样本x_i的预测概率分布中，x_i属于类别j的概率
t_j	样本的正确标签为j的置信度阈值
$H\left(\cdot\right)$	损失函数
θ	模型参数集合

4.2 相关工作及其改进

4.2.1 置信学习方法

置信学习（Confident Learning，CL）是通过对数据集中的标签错误进行表征和识别，基于概率阈值估计、置信度排序、去除噪声数据的原则，来提升标签质量。在这里，置信学习在类条件噪声过程的假设上，估计噪声标签（给定）和未损坏标签（未知）之间的联合分布。

数据集中噪声标签的存在引入了两个问题。如何识别带带标签错误的数据，以及如何在有噪声标签的情况下进行良好的学习。在这里，置信学习遵循以数据为中心的方法，从理论和实验上研究了带噪标签学习的关键在于准确和直接地表征数据中标签噪声的不确定性这一前提。

置信学习的具体流程分三个步骤：估计正确标签和噪声标签的联合分布、滤除噪声样本、重新调整各类样本权重并重新训练。具体操作如下：

步骤一：估计正确标签和噪声标签的联合分布。

首先用已有的带噪声训练集$S=\{(x_i,y_i),i=1,2,...,b\}$训练初步训练得到一个模型$f_0(\cdot)$及其对于训练集的预测$f_0(x)$，其中$f_0(x_i)$表示模型一开始对数据$x_i$的预测概率分布，$f\left(x_i;y_i=j\right)$表示模型预测数据$x_i$属于类别$j$的概率。计算每个类别$j$的置信度阈值$t_j$。以预测$f_0(x)$为基准构造混淆矩阵$C_{confusion}$和置信矩阵$C_{y\tilde{y}}$。$t_j$和$C_{y\tilde{y}}$的计算过程如表 4.2 和表 4.3 所示：

中国知网 https://www.cnki.net

表 4.2 置信度阈值计算

计算置信度阈值 t_j
1：for $j \leftarrow 1, C$ **do**
2： **for** $i \leftarrow 1, b$ **do**
3： $l \leftarrow$ **空列表**[]
4： **if** $\tilde{y}_i = j$ **then**
5： **将** $f_0(x_i; y_i = j)$ **这一概率值放入列表** l
6： $t_j \leftarrow average(l)$
输出：类别 j **的置信度阈值** t_j

表 4.3 置信矩阵构造

构造置信矩阵 $C_{y\tilde{y}}$
初始化， $C_{y\tilde{y}} \leftarrow C \times C$ **的全零矩阵**
1：for $i \leftarrow 1, b$ **do**
2： $cnt \leftarrow cnt + 1$
3： **for** $j \leftarrow 1, C$ **do**
4： **if** $f_0(x_i; y_i = j) \geq t_j$ **then**
5： $cnt \leftarrow cnt + 1$
6： $\tilde{y}_i \leftarrow j$
7： **if** $cnt > 1$ **then**
8： $\tilde{y}_i \leftarrow \arg\max_j f(x_i; y_i = j)$
9： **if** $cnt > 0$ **then**
10： $C_{y\tilde{y}}[y_i][\tilde{y}_i] \leftarrow C_{y\tilde{y}}[y_i][\tilde{y}_i] + 1$
输出：置信矩阵 $C_{y\tilde{y}}$

构造置信矩阵$C_{y\tilde{y}}$后，将其内部各个元素做归一化处理后形成噪声标签和真实标签的联合估计分布$Q_{y\tilde{y}}$，具体公式如下：

$$Q_{y\tilde{y}}[i][j] = \frac{\frac{C_{y\tilde{y}}[i][j]}{\sum_{j \in [C]} C_{y\tilde{y}}[i][j]} \cdot \left| X_{y=i} \right|}{\sum_{i \in [C], j \in [C]} \left(\frac{C_{y\tilde{y}}[i][j]}{\sum_{j \in [C]} C_{y\tilde{y}}[i][j]} \cdot \left| X_{y=i} \right| \right)} \tag{4.2}$$

混淆矩阵$C_{confusion}$是带噪标签y_i和模型预测$f(x_i)$的数量统计，带噪标签可能存在标签错误。主对角线上为各个类别中模型预测和带噪标签相符合的样本数量，主对角线之外则为模型预测和样本标签不同。

置信矩阵$C_{y\tilde{y}}$描绘了训练集中样本的带噪标签和“真实标签”的数量统计情况。对某一样本x_i，若其模型预测$f(x_i)$在类别j的概率分布高于阈值t_j，则标签j就是该样本的“真实标签”。主对角线上的是各个类别中带噪标签和“真实标签”相符合的样本数量，主对角线之

中国知网 https://www.cnki.net

外则分布着带噪标签不符合“真实标签”的样本。

步骤二：找出并过滤错误标签。

这里有五种方法过滤错误标签，如下：

策略 1：根据混淆矩阵 $C_{confusion}$ 过滤。这是一种基本方法，滤除所有混淆矩阵 $C_{confusion}$ 主对角线之外的样本，即将所有不符合模型预测的样本标签均视作噪声标签，然后予以滤除。

策略 2：根据置信矩阵 $C_{y\tilde{y}}$ 过滤。滤除所有置信矩阵 $C_{y\tilde{y}}$ 主对角线之外的样本，即若一个样本的模型预测的概率分布峰值并未高于其对应阈值，则认为该标签是错误标签，予以滤除。

策略 3：根据 PBC（Prune By Class）原则过滤。对于类别 j，选择 $b\cdot\sum_{c\in[C]:c\neq j}\left(Q_{y=j,\tilde{y}=c}[j]\right)$ 个置信度 $f\left(x_i;y=j\right)$ 最低的样本，予以滤除。该策略的思路为对于每个类别而言，以噪声标签和真实标签的联合估计分布 $Q_{y\tilde{y}}$ 估计每个类别可能存在的噪声标签数量，然后将置信度最低的那些样本予以滤除。

策略 4：根据 PBNR（Prune By Noise Rate）原则过滤。对于所有分布在联合估计分布 $Q_{y\tilde{y}}$ 的非主对角线 $Q_{y=i,\tilde{y}=j},i\neq j$ 上的样本，选择 $b\cdot Q_{y=i,\tilde{y}=j}$ 个模型预测概率分布在带噪标签和“真实标签”两项上差距最大的样本。该策略设想，若一个样本在联合估计分布 $Q_{y\tilde{y}}$ 的非主对角线上，则其模型预测和带噪标签之间的差距越大，则越其标签是错误标签的概率越高。

策略 5：C+NR 策略。将以上四种策略中选取几种相结合。例如，若一个样本标签在策略 3 和策略 4 下均被判定为错误标签，则将其滤除。

步骤三：模型微调

滤除在步骤二中被判定为错误标签的样本，重新调整各类别样本的权重，重新训练。对于类别 j，将其类别的样本的权重系数修正为 $\frac{1}{p(y=j|\tilde{y}=j)}=\frac{Q_{\tilde{y}}[j]}{Q_{y\tilde{y}}[j][j]}$。

4.2.2 小样本学习方法

近年来，深度神经网络作为深度学习方法的代表，在各种机器学习任务中都取得了卓越的效果，许多任务甚至已经超过人工的表现。但使用这些深度学习方法中一个关键就是要有大量标注数据作为支撑。但在实际工业场景中，在许多情况下，获取大量训练数据是很困难的一件事，后期的数据标注更需要耗费大量的人力和时间成本。所以，一种仅需少量样品的训练便可得到一个优质模型的方法则很有必要，小样本学习由此走上历史舞台。

原型网络（Prototypical Networks），如图 4.1，是基于这样一种思想，即存在一个对映函数（实际上可以是深度学习模型），它将每一条输入数据都映射为一个多维向量或者多维嵌入空间的一个点，这个多维向量可理解为多维嵌入空间中一点的坐标，多维向量的元素则包含了输入数据中的信息，可以是一幅图片的内容、或者是一条文本的语义。这里仍然是在解决分类问题，而每个类的点聚集在单个“原型”表示向量（Prototypical Representation）周围。“原型”表示向量即可理解为包含该类别的所有信息的向量，或者高维嵌入空间对应的点。

中国知网 https://www.cnki.net

为了做到这一点，可以使用神经网络学习输入数据到多维嵌入空间的非线性映射，并将类的“原型”表示向量作为其训练集中该类向量在嵌入空间中的均值。然后，通过简单地找到最近的类“原型”对嵌入的查询点进行分类。

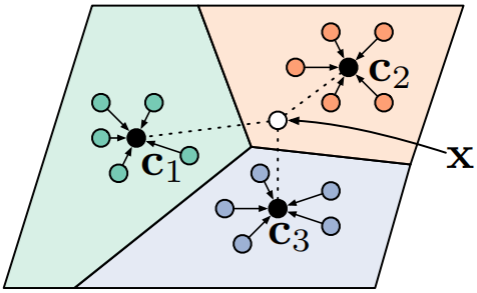

图 4.1 小样本学习分类图示

4.2.3 待改进之处的分析

置信学习在带噪学习领域取得了一定突破性的进展，该方法实用性很广，可作用于计算机视觉、自然语言处理等多个领域，也适用于多种深度学习模型结构和数据集，没有引入多余的超参数，即插即用，十分方便。但在具体使用场景中，置信学习方法却仍然存在一定的提升空间：

（1）置信学习方法适用于大规模情景，否则仅以预测概率不高为由直接削减则会进一步降低数据量。该方法倾向于去除损失较大的样本，可能会误删一些真正具有价值的难样本。而且在数据量较小的情况下，过于武断的删减可能会造成各个类别数据量的不平衡，并进而造成不同类别的数据相互影响，加重模型对不同种类之间的混淆。

（2）在小样本场景下，在训练集上学习得到的模型，去预测训练集的标签。如果模型在训练集上过拟合，甚至于已经被训练集的错误标签影响而学习到错误的信息，则以这样的模型为基准去删减甚至纠正数据标签，可能会加剧后续训练过程中的过拟合现象。

基于置信学习的改进思路：

（1）将训练集分成多个部分，采取类似 K 折交叉验证的方式。

（2）引入小样本带噪学习的方法，不是采用概率分布为基准，而是以各个类别的数据和“原型”数据之间的“距离”为基准，将训练数据按信任度分为错误样本、正确困难样本、正确简单样本，然后分别做赋予权值并且筛选。

（3）考虑到困难样本和错误样本的区别，将 PBC 和 PBNR 策略相结合，对每个类别，若虽然判断于标签不符合，但其“距离”并未近于阈值，则将其判断为困难数据，赋予相对较小的权值而并非直接删除。

4.3 基于置信学习的小样本带噪学习文本分类方法设计

这里对置信学习进行改进，将其适用于小样本且带噪的训练集环境，并结合文本分类的

中国知网 https://www.cnki.net

一些特征，提出了一种小样本的带噪文本分类方法（FewCL）。如图 4.2 所示，本方法将模型在训练集上的训练采用类似 K 折交叉验证的思路，然后利用 BERT 模型对输入语句进行语义编码，其某一隐层输出张量即作为语义表达张量。每一个类别的语义表达张量的平均值作为该类别的“原型”张量。计算每个语义表达张量到“原型”张量的“距离”，以该“距离”计算并赋予不同的数据权值，且按照一定策略将不同样本划分为错误样本、正确困难样本、正确简单样本三类然后对其分别处理。在所有数据赋予了权值并做筛选的过程中，模型本身也在不断训练和微调。当最后全部训练数据的权值赋予和筛选过程结束后，对模型做最终微调，后得到最终训练结果模型。

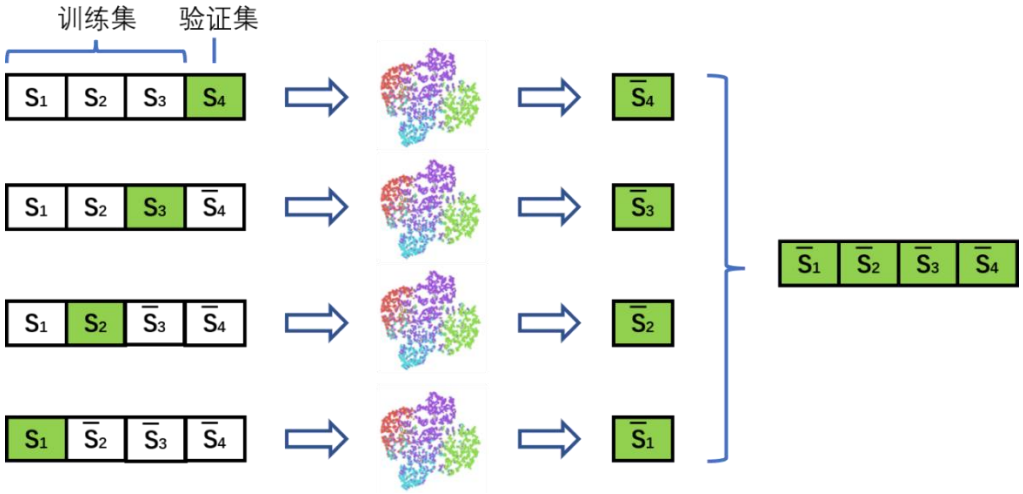

图 4.2 基于置信学习的小样本带噪学习方法示意图

FewCL 方法的具体过程伪代码如表 4.4 中**算法 3** 所示。具体方法如下：

步骤一：初始模型训练及验证。

该步骤对应算法 3 第 3 行。采用类似 K 折交叉验证的方式。将训练集 S 分为 K 份，即 $S=\{S_k \,|\, k=1,2,3,...,K\}$，选取其中一个训练子集 S_k 作为验证集，剩下的 $S-S_k$ 作为本轮的训练集，在此训练集上做初始训练后得到分类器模型 $f_k(\cdot)$。

步骤二：赋予权值及筛选

该步骤对应算法 3 第 4-17 行。取子集 S_k 中的标签（带噪）为类别 c 的集合 S_v^c 中的样本 $x_i^c, x_i^c \in S_k^c, c \in \{C\}$，输入模型 $f_k(\cdot)$。由于模型采用 BERT 神经网络，所以每个 x_i^c 会被 BERT 做语义编码后，在某一隐藏层输出得到语义表达张量 h_i^c。计算每个类别的语义表达张量的平均张量，并将其作为该类别的“原型”语义张量 p^c：

$$p^c = \frac{1}{|S_k^c|}\sum_{i=1}^{|S_k^c|} h_i^c \tag{4.3}$$

计算 h_i^c 距离同属于类别 c 的“原型”语义张量 p^c 的平均“距离” d_i^c。

在这里语义表达张量之间的“距离”可以考虑用三种方式描述：

（1）均方欧式距离：

中国知网 https://www.cnki.net

$$d_i^c = \left\| h_i^c - p^c \right\|_2^2 \tag{4.4}$$

（2）绝对值距离：

$$d_i^c = \left| h_i^c - p^c \right| \tag{4.5}$$

（3）余弦相似度：

$$d_i^c = \frac{h_i^c \cdot p^c}{\left\| h_i^c \right\| \cdot \left\| p^c \right\|} \tag{4.6}$$

三种描述“距离”的方式均有利弊，实际操作时可根据张量的分布情况调整。

根据每个语义表达张量 h_i^c 距离各自类别的“原型”语义张量 p^c 的“距离” d_i^c，计算集合 S_k^c 中，类别 c 的各个语义表达张量 h_i^c 到“原型”语义张量 p^c 的平均距离 t^c 以及标准差 σ^c 。

$$t^c = \frac{1}{\left| S_k^c \right|} \sum_{i=1}^{\left| S_k^c \right|} d_i^c \tag{4.7}$$

$$(\sigma^c)^2 = \frac{1}{\left| S_k^c \right|} \sum_{i=1}^{\left| S_k^c \right|} \left\| d_i^c - t^c \right\|_2^2 \tag{4.8}$$

根据 d_i^c 和 t^c 的值以及样本 x_i^c 的标签 c，对样本 x_i^c 的标签正确性做出判定。这里将标签划分为错误标签，正确困难标签，正确简单标签。错误标签指样本的标签标注错误；正确困难标签指数据集中较难短期学习完全的样本；正确简单标签指模型已经能够以较高的正确率和置信度做出判断的样本。

故有三种情况：

（1）若 $d_i^c < t^c$，则样本 x_i^c 的标签 c 为正确简单标签，保留该样本；

（2）若 $t^c < d_i^c < t^c + \sigma^c$，则样本 x_i^c 的标签 c 为正确困难标签，保留该样本；

（3）若 $d_i^c > t^c + \sigma^c$，则样本 x_i^c 的标签 c 为错误标签，去除该样本。

按上述策略对验证集 S_k^c 做样本筛选之后得到 $\bar{S}_k^c$，分别计算并赋予 $\bar{S}_k^c$ 中各个样本 $\bar{x}_i^c$ 权值 ω_i^c：

$$\omega_i^c = \frac{\exp(-d_i^c)}{\sum_{j \in \bar{S}_k^c} \exp(-d_j^c)} \tag{4.9}$$

此时可以看到，ω_i^c 随着 d_i^c 单调递减。通过这种方式，我们可以自适应调节各个样本的权重。d_i^c 越小，则意味着该样本映射到多维语义空间中的点距离该类别的“原型”更近，故我们对其投以更高的权重；d_i^c 越大，则意味着该样本映射到多维语义空间中的点距离该类别的“原型”更远，虽然我们仍然保留该样本，但仍然谨慎地投以相对更低的权重。

此时得到了子集 S_k^c 筛选后的“干净”数据集 $\bar{S}_k^c$，已在带噪数据集上初步训练后的模型 $f_k(\cdot)$。

步骤三：模型的微调。

此时，S_k 已经经过清理，得到了干净的数据集 $\bar{S}_k$，而步骤一和二中筛选并赋予权值的行为同样也在训练集 S 中生效得到 S'。在训练集 S' 的 K 块数据子集中重新选择一个非 $\bar{S}_k$ 的子集做为新的验证集 $S_{k'}$，在集合 $S' - S_{k'}$ 中重复上述步骤一和步骤二的行为。因为模型已经过初步

中国知网 https://www.cnki.net

训练得到 $f_k(\cdot)$，故后续的训练过程可将学习率逐渐调低，对模型的参数不做过于剧烈的修改。在集合 $S'-S_{k'}$ 中，由于包含了被赋予权值的 $\bar{S}_k$ 样本，这些样本在模型训练时将各自权值乘以其损失后，求梯度更新模型参数。$f_k(\cdot)$ 经过小学习率的微调后得到 $f_{k'}(\cdot)$。然后将验证集 $S_{k'}$ 中的样本输入 $f_{k'}(\cdot)$，得到语义表达张量，并计算 $S_{k'}$ 中各类的“原型”表达张量，各个样本到“原型”表达张量的“距离”及其均值和标准差。然后筛选集合 $S_{k'}$ 中样本并赋予权值。重复这个步骤直至训练集 $S=\{S_k \mid k=1,2,3,...,K\}$ 中全部 K 块子集均得到了筛选和权值赋予后，数据清理及模型训练过程结束。

表 4.4 FewCL 方法流程

算法 3：FewCL 方法

1：对每个样本 $x_i \in S$，初始化样本权重 $\omega_i \leftarrow 1$，模型为 $f(\cdot)$

2：for $k \leftarrow 1, K$ **do**

3： **在 $f(\cdot)$ 基础上，以 $S-S_k$ 为训练集训练得到模型 $f_k(\cdot)$**

4： **for** $c \leftarrow 1, C$ **do**

5： **for** $i \leftarrow 1, \left|S_k^c\right|$ **do**

6： **将标签为 c 的样本 x_i^c 输入模型计算 h_i^c**

7： **类别 c 的原型张量** $p^c \leftarrow \frac{1}{\left|S_k^c\right|}\sum_{i=1}^{\left|S_k^c\right|} h_i^c$

8： **for** $i \leftarrow 1, \left|S_k^c\right|$ **do**

9： **利用(4.4)-(4.6)计算 h_i^c 到 p^c 的距离 d_i^c**

10： **类别 c 的样本-原型平均距离** $t^c \leftarrow \frac{1}{\left|S_k^c\right|}\sum_{i=1}^{\left|S_k^c\right|} d_i^c$

11： **类别 c 的样本-原型距离标准差** $\sigma^c \leftarrow \sqrt{\frac{1}{\left|S_k^c\right|}\sum_{i=1}^{\left|S_k^c\right|}\left\|d_i^c - t^c\right\|_2^2}$

12： **for** $i \leftarrow 1, \left|S_k^c\right|$ **do**

13： **if** $d_i^c > t^c + \sigma^c$ **then**

14： $\omega_i^c \leftarrow 0$

15： **else**

16： $\omega_i^c \leftarrow \frac{\exp(-d_i^c)}{\sum_{j \in \bar{S}_k^c}\exp(-d_j^c)}$

17： $f(\cdot) \leftarrow f_k(\cdot)$

输出：净化后的样本集合 S，模型 $f(\cdot)$

4.4 实验及结果分析

4.4.1 模型结构与数据集

中国知网 https://www.cnki.net

本章实验利用 BERT 深度神经网络提取编码文本中的语义。由于 BERT 神经网络模型由一个词向量嵌入层和一个编码器 Encoder 构成，该编码器包括了一个 12 隐藏层的 Transformer 结构，有研究表明，对于一句输入语句，编码器的各个隐藏层都对一句话中的语义有不同程度的特征提取效果。

这里使用 BERT-base-uncased tokenizer 来对文本进行标记。句子的截断长度设置为 64。BERT-base-uncased 模型被用作编码器。嵌入层和编码器的第一个隐藏层的学习率被设定为 2e-5，从编码器的第 2 层到第 11 层隐藏层的衰减系数为 0.95，逐渐递减。对于 MLP，学习率为 1e-3。在使用本章所述的小样本带噪学习方法进行训练时，所有学习率设定一个衰减系数为 0.9。

模型训练效果采用训练好的模型在测试集上的总分类准确率评判，为了降低偶然情况的影响，本实验固定了模型初始参数和数据混合的随机种子，且通过不同的随机种子分别进行模型的训练过程，训练过程中每隔一定数量的批次训练后在验证集上验证模型的预测准确率，时时保存验证集上预测准确率最高的模型参数作为最佳模型，最后以该最佳模型在测试集上的准确率的平均值作为最终测试效果。

数据集采用 DBpedia、AG news、IMDB 三个数据集，分别取其中部分数据做带噪学习相关实验，为了模拟小样本场景，实验中尽量让训练集中每种类别的样本数量保持在 30 个左右。具体数据信息如表 4.5 所示。

表 4.5 实验数据集相关信息

数据集	种类数量	带标签样本数量
AG news	4	120
IMDB	2	60
DBpedia	14	420

4.4.2 BERT 模型各隐藏层训练结果比较

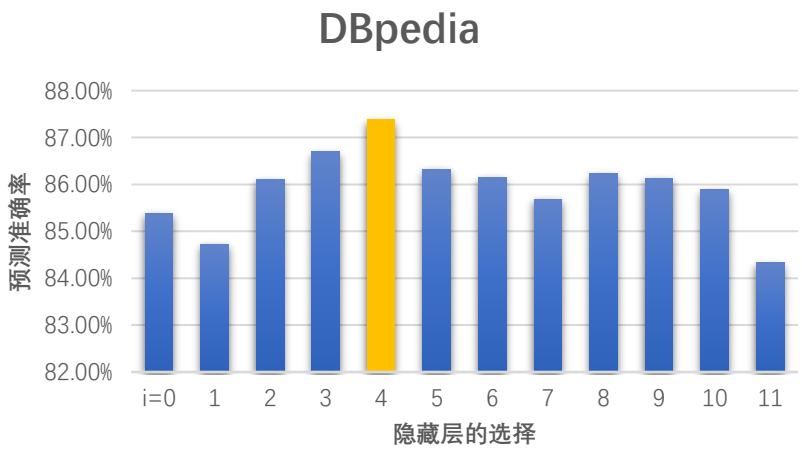

图 4.3 选择 BERT 中各个隐藏层使用 FewCL 的效果对比

中国知网 https://www.cnki.net

此实验在 DBpedia14 分类数据集上开展，且并不添加噪声，采用本章所述的基于置信学习改进的小样本带噪学习方法。目标为探明本方法在 BERT 编码器中各隐藏层的效果对比。实验结果如图 4.3 所示，在第五隐藏层效果最好，故后面的实验在采用本章所属方法时，均选择第五隐藏层（i=4）输出语义表达向量。某次训练过程中的模型预测准确率和损失值的变化情况如图 4.4 所示。

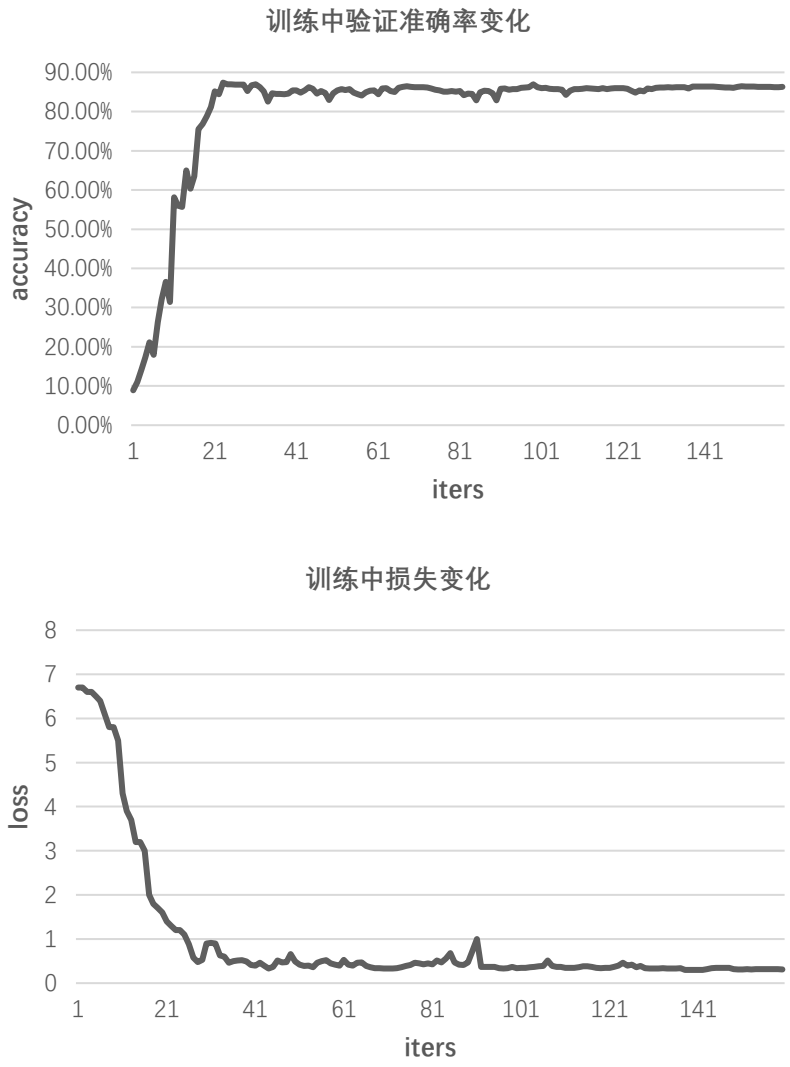

图 4.4 模型训练过程中的预测准确率和损失变化

4.4.3 不同 K 值效果对比

FewCL 一开始需要先将训练集分为 K 个子集，然后对 K 个子集分别做模型训练以及样本标签的清理验证。本实验的目的是验证训练集划分的子集个数 K 的值对模型训练效果的影响。实验在 DBpedia14 分类数据集上开展，FewCL 过程中的语义表达张量产生于第 5 隐藏层（根据实验 4.4.2），数据集中分别随机加入不同比例的错误标签。K=1 是一种消融实验，意味着初始阶段放弃对训练集的分割，但仍保留后续的模型训练及对带噪样本标签的筛选过程。

中国知网 https://www.cnki.net

当 K 值较小的时候，每一轮筛选过程中训练模型所用的数据量也较小；随着 K 值的逐渐加大，每一轮筛选过程中模型遇到的数据重复度也逐渐加大。实验结果如表 4.6 所示。

表 4.6 不同 K 值、不同标签错误率情况下训练最后得到的模型预测准确度

K 值		1	2	3	4	5	6
标签错误率	0%	87.31%	**87.41%**	87.35%	87.38%	87.25%	87.11%
	20%	82.15%	82.86%	82.94%	**83.02%**	82.88%	82.53%
	40%	74.31%	75.97%	**76.25%**	76.20%	75.33%	75.54%

实验结果表明，随着训练集中错误标签比例的上升，FewCL 的训练效果都在降低，这意味着错误标签确实对深度学习模型的训练效果产生了影响。

当训练集中标签错误率为 0%，K=1,2,3,4 时模型训练效果基本一致，当 K≥5 后训练出的模型精度开始下降。当 K≥5 时，模型则更有可能在过于重复的数据上过拟合，所以训练效果有所降低。因为此时标签错误率为 0%，所以虽然过程中依然会对数据有所筛选，但因为 FewCL 的筛选策略相对较为保守，会相对更多地保留难以判断的困难样本，所以筛选后的结果并不会造成太大影响。

当训练集中标签错误率为 20%~40%，此时数据集中包含相当数量的错误样本标签。

若 K=1，即不将训练集分开，而是将训练集上训练出的模型，用于预测和判断训练集上的标签正误，因为模型本身已经在该训练集上拟合，故对于训练集本身存在的错误，可能存在误判的风险，所以 K=1 时模型训练效果最低。

若 K≥2，则采用 FewCL 方法训练模型，从结果可看出，K=3,4 时的模型训练效果最佳。每轮筛选时，模型训练所用的样本数据量为$(1-1/K)|S|$。当 K 较小时（K=2），因为本文都在小样本情景中，样本数据量较少，则每轮训练时的样本数量更少，这对模型的训练不利。当 K 较大时（K=5,6）,每轮训练时模型遇到的数据重复度过高，也可能导致模型后期因在部分数据上过拟合导致对标签的筛选判断失误。但只要 K≥2，FewCL 方法的效果都在 K=1（即不讲训练集分开）的方式上有所增益。

总体来说，当标签错误率很低时，在 1~4 范围内 K 值的设置影响不明显。当标签错误率开始提升，设置 K 值大于 1，即划分数据集的意义开始体现，此时即使过大的 K 值可能导致模型的过拟合，但依然比不划分数据集效果更好。K 值在 3~4 范围中效果最好，且其相对于不划分数据集（K=1）的差距随着标签错误率的提升而愈加明显。

4.4.4 带噪学习效果

此实验目的是探究比较在 IMDB 影评数据集、AG news 新闻数据集、DBpedia 新闻分类数据集三个文本分类数据集上，本章所述的方法 FewCL 的效果。

其他训练方法比较：

（1）一般全监督训练方法[39]，损失函数选择传统的 CE 交叉熵。在 BERT 的基础上，用

中国知网 https://www.cnki.net

标注的数据对 BERT-base-uncased 中的参数进行进一步训练。作为基准方法比较。

（2）一般全监督训练方法，损失函数选择 SCE 对称交叉熵[61]。对称交叉熵时优化损失函数类型的带噪学习方法。该方法时带噪学习的基本方法之一，通过将交叉熵对称化的方式，可在模型训练时自适应调节近似错误标签的损失值大小。

（3）置信学习方法（Confident Learning，CL）[38]。样本筛选类型的带噪学习方法。在 BERT 模型上，利用置信学习方法对数据集做筛选后，再对模型重新训练。筛选数据时，滤除的策略选择 PBC、PBNR。

由于 DBpedia、IMDB、AG news 这些数据集质量较高，故这里通过随机修改一定比例的标签的方式添加标签噪声。标签噪声的比例从 0%~40%。采用噪声概率均匀的对称噪声。实验结果如图 4.5~4.7 所示。

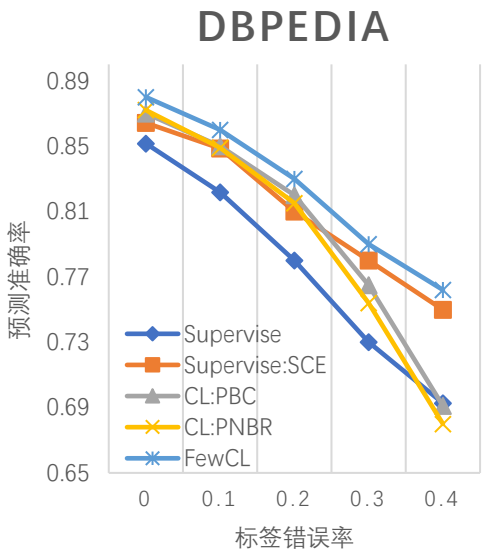

图 4.5 五种深度学习方法在不同比例的噪声标签的 DBpedia 数据集的训练效果

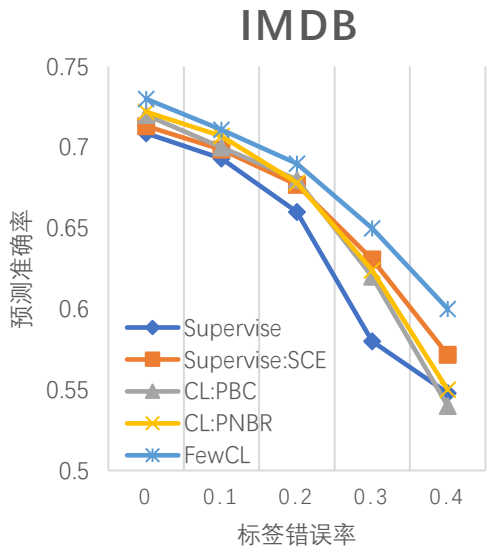

图 4.6 五种深度学习方法在不同比例的噪声标签的 IMDB 数据集的训练效果

中国知网 https://www.cnki.net

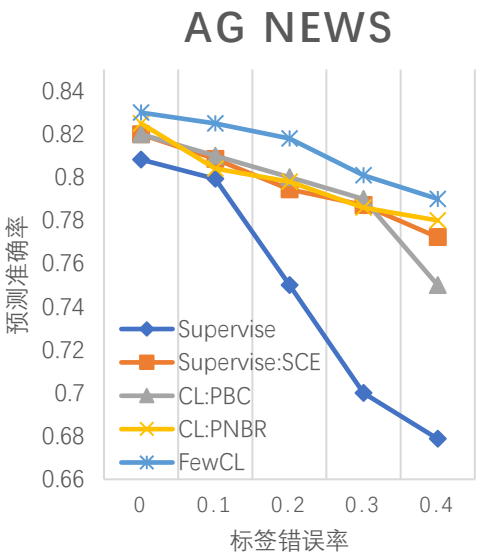

图 4.7 五种深度学习方法在不同比例的噪声标签的 AG news 数据集的训练效果

实验结果可看出，在标签噪声为 0%时，本文所述的方法训练所得模型的预测精度虽然略微高于另外三种方法，但差距并不明显。

当噪声标签比例逐渐增加的时候，四种方法的训练所得的模型分类精度都在降低，这说明标签噪声的添加确实对模型的训练效果造成了负面影响。

在噪声比例处在 0%~30%时，基础的一般全监督训练方法的训练效果都比不上另外三种方法，这说明另外三种方法采用的改进都确实起了积极作用，提高了模型训练过程的抗噪性能。当噪声比例处于 40%时，在 DBpedia 和 IMDB 数据集上，置信学习方法效果甚至均低于一般全监督训练方法。这是因为本实验中数据集数据量并不大，在噪声比例过大时，置信学习滤除了过多的样本，训练数据量过少，最终导致训练效果并不理想。

无论标签噪声比例是多少，对称交叉熵的应用都提高了一般全监督训练的效果，但在噪声比例在 0%~20%区间内，其效果略逊于置信学习，因为此时置信学习滤除的数据量并不大，数据的筛选确实取得了良好的效果；但当噪声比例高于 20%时，置信学习在小样本学习上的劣势便展现了出来。

但从始至终，本章所述的基于置信学习改进的小样本带噪学习方法都取得了最好的效果。当标签噪声率较低时，本方法很好地保留了置信学习思路的优势，滤除了少数带噪样本；当标签噪声率较高时，本方法却并没有过于武断地滤除过多样本，保留了数据中部分困难样本，即使在部分错误样本没有滤除，本方法也通过赋予权重的方式尽量降低了其影响力，取得了最佳的小样本带噪模型训练效果。

4.5 本章总结

本章首先介绍了置信学习（Confident Learning，CL）和相关方法的思路和主要流程。之后本章分析了置信学习方法的可改进之处及其改进思路，即模型在带噪数据集上可能存在的

中国知网 https://www.cnki.net

过拟合现象会影响对标签正误判断的可靠性；置信学习在小样本学习的场景下效果可能效果并不好。继而本章将小样本学习的“原型”的相关思路引入，提出了基于置信学习改进的小样本带噪文本分类方法（FewCL）。该方法利用 BERT 模型将文本语句映射到一个高维语义空间，每条语句都有一个高维的语义表达张量，每个种类的样本都有一个“原型”张量，通过计算每个计算高维语义表达张量到“原型”张量之间的“距离”，将样本标签分为错误标签、正确困难标签、正确简单标签，滤除错误标签然后给正确标签赋予不同的权重。为了尽量降低模型过拟合造成的影响，这里采取类似 K 折验证的思路，在训练集的 K-1 个子集中训练或微调模型后，判断和验证剩余 1 个子集的标签正误并作后续筛选操作。后续实验表明，（1）利用 BERT 编码器的第 5 层输出语义表达张量可达到最好的效果；（2）对数据集合理的划分有助于提升模型在带噪数据集上的训练效果；（3）随着标签错误率的提升，所有方法训练的模型预测精度都在下降，说明错误标签确实对模型训练有很大影响；（4）在小样本学习场景中，由于置信学习倾向于滤除错误和困难标签，置信学习在标签噪声较高时效果突然变差；（5）本章提出的 FewCL 在多个数据集，多种方式中都达到了最佳的训练效果。

中国知网 https://www.cnki.net

第5章 总结与展望

5.1 研究工作总结

随着深度学习在图像识别、语音识别领域的大放异彩，人们对深度学习在 NLP 领域的价值寄予厚望。如今 ChatGPT 的成功使得相关的研究和应用变得炙手可热。自然语言处理作为人工智能领域的认知智能，成为目前大家关注的焦点。分类和预测（回归）是深度学习模型的两大任务，文本分类旨在将特定领域中的语句判断所属标签。这是自然语言处理领域的一个基础性课题。但深度学习模型的训练十分依靠高质量大规模数据集，大规模高质量数据集的搭建维护本身需要成本，并且数据的大量标注也需要很多的人力和时间，故更低成本的半监督学习和带噪学习有着很广阔的应用空间和实际价值。

本文首先介绍了不同神经网络结构以及相关理论进展，然后选择了目前先进的 BERT 模型结构做文本语义特征的提取，并且基于 BERT 模型构建文本分类方法。

（1）针对带标签样本的数量和质量都无法保证的苛刻条件下，本文在第三章提出了一种对抗性的带噪半监督文本分类方法（ASSL）。我们提出了 MHGD 对抗学习方法用于文本数据增强，在保留整体语义信息的同时提升正则化性能。基于对称交叉熵 SCE，我们提出 Flex-SCE 损失函数，其中交叉熵项被赋予一个权重系数，该权重系数按照一定的规则衰减，以提高对噪声标签的鲁棒性。为了证明 ASSL 的有效性，实验在不同的数据集上开展，ASSL 与现有的深度学习方法，即全监督训练、VAT、π-model 和去除特定组件的 ASSL 进行比较，均取得了最高的预测精度。

（2）由于置信学习（Confident Learning, CL）在样本数量不多时候的表现存在提升空间，本文在第四章引入小样本领域的学习和 K 折验证的思路，在 BERT 模型基础上改进了置信学习在小样本学习情景下的表现，提出了 FewCL 小样本带噪文本分类方法。为了减轻过拟合对模型的影响，FewCL 将带噪的训练集分为 K 个子集，选择其中 K-1 个子集训练或微调模型，然后去剩下 1 个子集做标签的正误判断和筛选。相比置信学习，FewCL 对样本的筛选更加细化，通过计算样本语义表达张量和类别“原型”张量的“距离”，将样本标签分为错误标签、正确困难标签、正确简单标签。然后筛选并给样本赋权值后尽兴下一轮模型的训练和微调，直至最后完成。实验表明，FewCL 在多个数据集，多种标签错误率下都取得了最佳的训练效果。

5.2 研究工作展望

本文中主要对文本分类问题分别提出了半监督和带噪学习方法，两种思路都取得了良好

中国知网 https://www.cnki.net

的效果。但仍有可以提升的空间。

（1）运行速率的提升。两个算法的整个流程都并不简单，ASSL 方法需要多次梯度变化，这对运行机器的运算速率以及缓存空间都提出了挑战。FewCL 方法中一个模型需要经历不断的多轮次的微调，实验中也存在一种情况：数据集还未清理结束，但模型的预测精度已经达到了最高。之后我们会考虑能否更加轻量化地解决问题。

（2）数据集合的扩充。这次的实验更多地在三个英文数据集上开展，因为它们的质量更高。后续可以考虑扩充中文高质量数据集，并尝试在中文场景下是否有新的想法。

中国知网 https://www.cnki.net

参考文献

[1] 沈艳,陈赟,黄卓.文本大数据分析在经济学和金融学中的应用:一个文献综述[J].经济学(季刊),2019,18(4):1153-1186.

[2] 李一昊,滕伊洋,张亚群等.毒性病理学中人工智能和机器学习的应用研究进展[J].中国新药杂志,2023,32(06):598-604.

[3] 秦璐,李易,林仙铖等.基于机器学习的比特币实体分类方法研究综述[J].海南师范大学学报(自然科学版),2023,36(01):38-45+52.

[4] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning[C].Proceedings of the 25th international conference on Machine learning. 2008: 160-167.

[5] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. nature, 2015, 521(7553): 436-444.

[6] Mourad R. Semi-supervised learning improves regulatory sequence prediction with unlabeled sequences[J]. BMC bioinformatics, 2023, 24(1): 1-15.

[7] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[C]. Proceedings of NAACL-HLT. 2018: 2227-2237.

[8] 孙凯丽,罗旭东,罗有容.预训练语言模型的应用综述[J].计算机科学,2023,50(01):176-184.

[9] Bowman S, Angeli G, Potts C, et al. A large annotated corpus for learning natural language inference[C]. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 632-642.

[10] 刘文豪,姜胜明.基于无标签半监督学习的商品识别方法[J].计算机应用与软件,2022,39(07):167-173.

[11] Howard J, Ruder S. Universal Language Model Fine-tuning for Text Classification[C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 328-339.

[12] Tjong Kim Sang E F, De Meulder F. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition[C]. Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. 2003: 142-147.

[13] Rajpurkar P, Zhang J, Lopyrev K, et al. SQuAD: 100,000+ Questions for Machine Comprehension of Text[C]. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 2383-2392.

[14] Alec R, Karthik N, Tim S, et al. Improving language understanding with unsupervised learning[J].

中国知网 https://www.cnki.net

Citado, 2018, 17: 1-12.

[15] Brown P F, Della Pietra V J, Desouza P V, et al. Class-based n-gram models of natural language[J]. Computational linguistics, 1992, 18(4): 467-480.

[16] Ando R K , Tong Z . A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data[J]. Journal of Machine Learning Research, 2005, 6:1817-1853.

[17] 吴子玥.基于自然语言处理和机器学习的文本分类及其运用[J].电子技术与软件工程,2023,No.249(07):216-219.

[18] 蒋玉茹,张禹尧,毛腾等.汉语零形回指消解研究综述[J].中文信息学报,2020,34(03):1-12.

[19] 张晓龙,支龙,高剑等.一种半监督学习的金融新闻文本分类算法[J].大数据,2022,8(02):134-144.

[20] 阎亚亚.词袋模型和 TF-IDF 在文本分类中的比较研究[J].电脑知识与技术,2021,17(28):138-140.

[21] 仝鑫,王罗娜,王润正等.面向中文文本分类的词级对抗样本生成方法[J].信息网络安全, 2020, 20(09): 12-16.

[22] 闫云飞,孙鹏,张杰勇等.基于领域 BERT 模型的服务文本分类方法[J].空军工程大学学报,2023,24(01):103-111.

[23] Le Q, Mikolov T. Distributed representations of sentences and documents[C]. International conference on machine learning. PMLR, 2014: 1188-1196.

[24] Zhang Y, Wallace B C. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification[C]. Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2017: 253-263.

[25] CONNEAU A, SCHWENK H, BARRAULT L, et al. Very Deep Convolutional Networks for Text Classification[C/OL]. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Valencia, Spain. 2017.

[26] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[27] Abuduweili A, Li X, Shi H, et al. Adaptive consistency regularization for semi-supervised transfer learning[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 6923-6932.

[28] Saito K, Kim D, Sclaroff S, et al. Semi-supervised domain adaptation via minimax entropy[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2019: 8050-8058.

[29] Maas A, Daly R E, Pham P T, et al. Learning word vectors for sentiment analysis[C]. Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies. 2011: 142-150.

中国知网 https://www.cnki.net

[30] Maas A, Daly R E, Pham P T, et al. Learning word vectors for sentiment analysis[C]. Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies. 2011: 142-150.

[31] Lehmann J, Isele R, Jakob M, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia[J]. Semantic web, 2015, 6(2): 167-195.

[32] MIKOLOV T, CHEN K, CORRADO GregS, et al. Efficient Estimation of Word Representations in Vector Space[Z]. International Conference on Learning Representations. 2013.

[33] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C]. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.

[34] 刘颖,杨鹏飞,张立军等.前馈神经网络和循环神经网络的鲁棒性验证综述[J/OL].软件学报:1-33[2023-05-27].https://doi.org/10.13328/j.cnki.jos.006863.

[35] KIM Y. Convolutional Neural Networks for Sentence Classification[C/OL]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar. 2015. http://dx.doi.org/10.3115/v1/d14-1181. DOI:10.3115/v1/d14-1181.

[36] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[J]. Neural Information Processing Systems. 2017.

[37] Kenton J D M W C, Toutanova L K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]. Proceedings of NAACL-HLT. 2019: 4171-4186.

[38] Northcutt C, Jiang L, Chuang I. Confident learning: Estimating uncertainty in dataset labels[J]. Journal of Artificial Intelligence Research, 2021, 70: 1373-1411.

[39] Sun C, Qiu X, Xu Y, et al. How to fine-tune bert for text classification?[C]. Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18. Springer International Publishing, 2019: 194-206.

[40] Goodfellow I J, Shlens J, Szegedy C. EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES[J]. stat, 2015, 1050: 20.

[41] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[J]. stat, 2017, 1050: 9.

[42] Shafahi A, Najibi M, Ghiasi M A, et al. Adversarial training for free![J]. Advances in Neural Information Processing Systems, 2019, 32: 3358–3369.

[43] Miyato T, Maeda S, Koyama M, et al. Virtual adversarial training: a regularization method for supervised and semi-supervised learning[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(8): 1979-1993.

[44] Zhu C, Cheng Y, Gan Z, et al. Freelb: Enhanced adversarial training for language

中国知网 https://www.cnki.net

understanding[J]. 2019.

[45] Laine S, Aila T. Temporal Ensembling for Semi-Supervised Learning[C]. International Conference on Learning Representations.

[46] Wang J, Perez L. The effectiveness of data augmentation in image classification using deep learning[J]. Convolutional Neural Networks Vis. Recognit, 2017, 11: 1-8.

[47] Miyato T, Dai A M, Goodfellow I. ADVERSARIAL TRAINING METHODS FOR SEMI-SUPERVISED TEXT CLASSIFICATION[J]. stat, 2017, 1050: 6.

[48] Xie Q, Dai Z, Hovy E, et al. Unsupervised data augmentation for consistency training[J]. Advances in neural information processing systems, 2020, 33: 6256-6268.

[49] Miyato T, Maeda S, Koyama M, et al. Virtual adversarial training: a regularization method for supervised and semi-supervised learning[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(8): 1979-1993.

[50] Zou Y, Yu Z, Liu X, et al. Confidence regularized self-training[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 5982-5991.

[51] Sohn K, Berthelot D, Carlini N, et al. Fixmatch: Simplifying semi-supervised learning with consistency and confidence[J]. Advances in neural information processing systems, 2020, 33: 596-608.

[52] Zhang B, Wang Y, Hou W, et al. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling[J]. Advances in Neural Information Processing Systems, 2021, 34: 18408-18419.

[53] Wei J, Zou K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks[C]. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 6382-6388.

[54] 尤丛丛,高盛祥,余正涛等.基于同义词数据增强的汉越神经机器翻译方法[J].计算机工程与科学,2021,43(08):1497-1502.

[55] Sennrich R, Haddow B, Birch A. Improving Neural Machine Translation Models with Monolingual Data[C]. 54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics (ACL), 2016: 86-96.

[56] Nagesh A, Surdeanu M. An exploration of three lightly-supervised representation learning approaches for named entity classification[C]. Proceedings of the 27th International Conference on Computational Linguistics. 2018: 2312-2324.

[57] Narayan P L, Nagesh A, Surdeanu M. Exploration of Noise Strategies in Semi-Supervised Named Entity Classification[C]. Joint Conference on Lexical and Computational Semantics. Association for Computational Linguistics, 2019: 186-190.

中国知网 https://www.cnki.net

[58] Chen J, Yang Z, Yang D. MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification[C]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 2147-2157.

[59] Ghosh A, Kumar H, Sastry P S. Robust loss functions under label noise for deep neural networks[C]. Proceedings of the AAAI conference on artificial intelligence. 2017, 31(1): 16-20.

[60] Zhang Z, Sabuncu M R. Generalized cross entropy loss for training deep neural networks with noisy labels[C]. Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018: 8792-8802.

[61] Wang Y, Ma X, Chen Z, et al. Symmetric cross entropy for robust learning with noisy labels[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 322-330.

[62] Miller E G, Matsakis N E, Viola P A. Learning from one example through shared densities on transforms[C]. Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662). IEEE, 2000, 1: 464-471.

[63] Lake B M, Salakhutdinov R, Gross J, et al. One shot learning of simple visual concepts[C]. 33rd Annual Meeting of the Cognitive Science Society: Expanding the Space of Cognitive Science, CogSci 2011. The Cognitive Science Society, 2011: 2568-2573.

[64] 潘雪玲,李国和,郑艺峰.面向深度网络的小样本学习综述[J/OL].计算机应用研究:1-10[2023-05-28].https://doi.org/10.19734/j.issn.1001-3695.2023.02.0074.

[65] Zhang L, Liu J, Luo M, et al. Scheduled sampling for one-shot learning via matching network[J]. Pattern Recognition, 2019, 96: 1-11.

[66] 汪雨竹,彭涛,朱蓓蓓等.基于元学习的小样本知识图谱补全[J].吉林大学学报(理学版),2023,61(03):623-630.

[67] Liang K J, Rangrej S B, Petrovic V, et al. Few-shot learning with noisy labels[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 9089-9098.

中国知网 https://www.cnki.net