Computer Science and Application 计算机科学与应用, 2023, 13(10), 1938-1947
Published Online October 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.1310192

Hans汉斯

基于BERT知识蒸馏的情感分析模型

孙杨杰，常青玲

五邑大学智能制造学部，广东 江门

收稿日期：2023年9月16日；录用日期：2023年10月16日；发布日期：2023年10月24日

摘　要

目前，BERT预训练语言模型在文本分析领域得到普遍应用，在情感分析任务上更是取得了SOTA级别的表现，但是在边缘设上部署BERT模型仍具有挑战性。而一般用于解决情感分析的传统的机器学习模型(SVM，NB和LR)较易部署，但精度不如BERT模型。本文旨在实现对两种不同方法的优势进行融合，训练出一个精度高且易部署的模型并用于解决情感分析任务。先前的工作大多是将BERT模型蒸馏进一个浅层的神经网络结构，这种方法能够减少BERT模型的参数，但依然保留了上百万的参数，难以在边缘设备上部署。本文提出将已经训练好的BERT模型设定为教师模型，将传统机器学习模型(SVM，NB和LR)设定为学生模型，并在输出层面完成知识转移。训练学生模型使用教师模输出的软标签和logits，并证明了学生模型在软标签上进行训练可以简化学生模型的学习过程，同时强调了从教师模型获得的所有文本特征之间的关系是平等的，并用蒸馏后的学生模型在IMDB数据集上进行验证。实验结果表明，蒸馏了BERT预训练语言模型知识的传统机器学习模型在测试数据上的性能得到明显提升，相比较于基线模型精度都提高了1%以上，且参数量维持在BERT模型的1/10水平。

关键词

BERT，知识蒸馏，机器学习，情感分析，文本分类

BERT-Based Knowledge Distillation for Sentiment Analysis Model

Yangjie Sun, Qingling Chang

Faculty of Intelligent Manufacturing, Wuyi University, Jiangmen Guangdong

Received: Sep. 16th, 2023; accepted: Oct. 16th, 2023; published: Oct. 24th, 2023

Abstract

In recent years, pretrained language models have achieved remarkable performance in various

文章引用: 孙杨杰, 常青玲. 基于 BERT 知识蒸馏的情感分析模型[J]. 计算机科学与应用, 2023, 13(10): 1938-1947.
DOI: 10.12677/csa.2023.1310192

fields, including computer vision, natural language processing, and multimodal tasks. However, while pretrained language models excel in accuracy across various tasks, they come with high computational costs and long inference times. On the other hand, traditional machine learning models are more easily deployable but often lag behind in accuracy compared to pretrained language models. This paper aims to combine the strengths of both approaches to create a highly accurate and deployable model. Inspired by knowledge distillation techniques (teacher-student models), this study sets a pretrained BERT language model as the teacher model and traditional machine learning models as the student models. It extracts additional soft-label knowledge from the large, high-weight teacher model to train lightweight student models. The goal is to use the distilled student models to tackle sentiment analysis tasks in textual data, with a focus on the benchmark movie review dataset. The main steps include data preprocessing, feature extraction, and knowledge distillation. The research results demonstrate that traditional machine learning models, distilled with knowledge from large pretrained language models, significantly improve performance on test data, with accuracy gains of more than 1% compared to baseline models, and the number of parameters is maintained at 1/10 level of the BERT model.

Keywords

BERT, Knowledge Distillation, Machine Learning, Sentiment Analysis, Text Classification

1. 引言

在现代自然语言处理(NLP)领域，大型预训练语言模型被得到广泛应用，例如源自 Transformer 架构的 BERT 和 GPT-2 等双向预训练语言模型，尤其是文本分类任务。然而，这些大型预训练模型的计算成本相当高，使得它们难以在运行时得以有效应用。以 BERT 模型为例，它包含着 3.44 亿个参数和 24 个 Transformer 层[1]。然而这些庞大的模型参数导致它们无法轻松部署在边缘设备上，目前可部署模型的精度又不如大型预训练语言模型。为了解决这个问题，已经提出了多种压缩方法，包括蒸馏、量化和权值修剪等技术[2]。蒸馏方法是当前备受研究人员关注的一种主动模型压缩方法，用以应对模型体积庞大、推理计算时间过长等问题[3]。知识蒸馏最早是由[4]提出的，该方法有助于弥合了大型模型和轻量级模型在可学习性和表达能力之间的差距，其主要思路是通过训练一个小型、轻量级的机器学习或深度学习模型，该模型借鉴了从一个较大、高权重的已经训练好的模型中提取的额外知识，从而弥合教师模型和学生模型之间的精度差距，模型结构图如下图 1 所示。本文的目标是通过将与 BERT 预训练语言模型的知识提取到这些机器学习模型中，用来提升这些易部署模型的精度。利用蒸馏方法来增强支持向量机(Support Vector Machine, SVM)、朴素贝叶斯(Naive Bayes, NB)和逻辑回归(Logistic Regression, LR)等轻量级可部署模型在基准电影评论数据集上进行情感分析任务的性能。最后，我们将在准确率方面对这些机器模型性能进行比较。

2. 相关工作

根据特定任务对 BERT 预训练模型进行微调，以在特定的自然语言任务中使用，会得到精度更高的表现[1]。下图 2 显示我们可以对 BERT 进行微调，以用于问题回答(Question Answering, QS)、命名实体识别(Name Entity Recognition, NER)和分类任务等[5]。

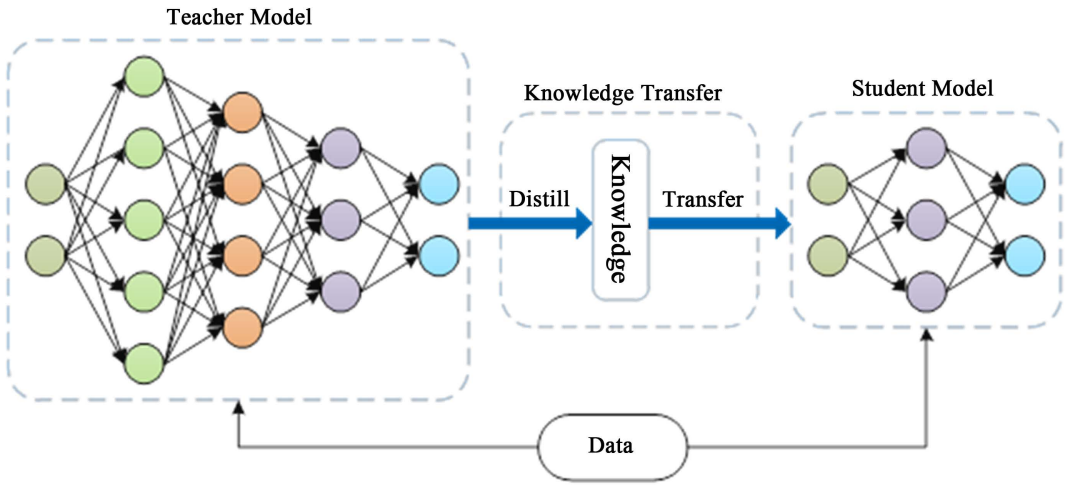

Figure 1. Original knowledge distillation teacher-student framework [3]
图 1. 原始知识蒸馏教师－学生模型框架[3]

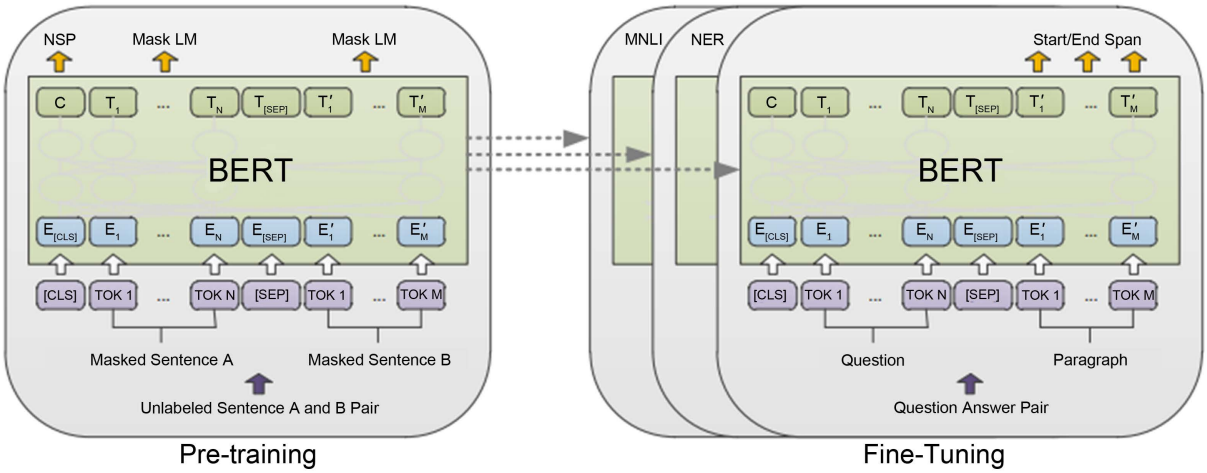

Figure 2. BERT fine-tuning procedures [5]
图 2. BERT 模型微调过程[5]

但是微调后，BERT 模型仍然难以部署到低算力资源设备上，其庞大复杂的结构在预处理过程中会产生巨量的模型参数，让算力资源有限的研究者无法需要解决的特定任务来微调 BERT 预训练语言模型[6]。据此，本文提出要解决的问题：基于 BERT 知识蒸馏提升机器学习模型在文本任务上的性能研究。其中情感分析(Sentiment Analysis, SA)是文本任务中的一个热门的研究领域，被认为是自然语言处理中的一个复杂任务，其中可能包含多种编码和否定等因素[7]。因此，应对这些挑战的能力是提高情感分析分类器准确性的一个重要方面[8] [9]。

在情感分析领域，已经提出了一些知识蒸馏方法。[10] [11] [12]表明通过合适的蒸馏方法可以提高情感分析性能、减小模型体积和计算时间。而对 BERT 模型进行知识蒸馏的研究已经在 tinyBERT 模型[13]和 DistilBert 模型[14]中被引入。有研究者已经提出了 BERT 模型的蒸馏版本，并用于解决情感分析问题[6]，它利用了大型的 BERT 模型作为教师模型，而学生模型使用了较少编码层的 BERT 模型，蒸馏模型同时学习了学习目标、最终隐藏状态的总和以及大模型的软标签。另外一项工作提出了一种应对情感分析任务的方法，其中包括一组在捷克电影评论数据集上训练的 BERT 模型，采用增强的池化层以实现更好的情感预测，然后将这些知识蒸馏到一个简单的 BERT 模型中。在[15]的作者旨在通过提出一种领域

感知方法来增强 BERT，用于跨领域情感分类任务，数据集是亚马逊评论基准数据集。[16]解决了[15]中的问题，它提出了一种无监督的领域自适应策略，将知识蒸馏与对抗性领域自适应相结合。这种方法已在跨领域情感分类任务中进行了评估，并显示出在选择适当的温度值时，它可以提高性能。而[17]开发了一种患者知识蒸馏策略，将 BERT 模型压缩成一个轻量级的小型网络。这种蒸馏方法不使用原始教师的最后一层输出，而是从中间教师层学习。[18]提出了一种知识蒸馏学习策略，将用户特征、评论和产品信息作为输入集成在一起。这种方法将教师模型中的用户特征传递到学生神经网络的权重中，即教师模型将其训练数据的预测分布传递给学生模型。作者提出了一个名为 L2PG 的深度学习模型，用于处理在创建的包含十个类别的亚马逊数据集上的终身情感分析任务[19]，该方法通过使用知识蒸馏来保留从早期任务中学到的知识，并仅传递有用的子集特征以理解新任务。在基于特征级别的情感分析方向，[20]提出了一种结合了门控卷积神经网络和注意机制的蒸馏模型，该模型在各种情感分析数据集上进行了评估，在模型大小和推理时间方面表现出比其他模型更高的性能。[21]提出的一项工作实现了一种基于双门策略的方法来提取相关的方面情绪特征在这方面也有所突破。另外，知识蒸馏也被应用于数据集，它被称为数据蒸馏，其目的是将大型数据集的知识封装到合成的小型数据集中，如[22]。

3. 模型架构

3.1. 学生模型

与过往的蒸馏模型不同的是，我们采用的学生模型是传统机器模型，本文针对于情感分析任务，所选用具体的学生模型是 SVM、NB 和 LR。传统的机器学习模型通常使用词袋模型和 TF-IDP 来表示文本数据，将文本数据转化为离散的特征向量，如下图 3 所示。而一般作用在 BERT 预训练语言模型的词嵌入，通常不直接用于传统的机器学习模型。选用机器学习模型作为学生模型，也是因为其没有池化层，残差连接，注意力机制等等结构，所以模型参数相对预训练语言模型少了很多。

Figure 3. Student model workflow diagram
图 3. 学生模型工作流程图

3.2. 教师模型

本文提出的教师模型就是谷歌团队在 2019 年提出的 BERT-Large 模型[1]。对教师模型最重要的步骤是根据对应的下游任务来对 BERT 模型进行微调。在情感分析任务中，通常使“下个句子预测”任务(Next Sentence Prediction, NSP)任务来微调 BERT 预训练语言模型，微调流程如下图 4 所示。

但本文在对数据进行预处理的时候，删掉了长度为 1 的字母和标点符号，所以据此考虑，也需要用“掩码语言模型”任务(Maske Language Model, MLM)来对 BERT 模型进行微调。选用的损失函数为两个任务联合学习的损失函数，公式(1)所示。其中，θ：BERT 中编码器分的参数；θ_1：是 Mask-LM 任务中在编码器上所接的输出层中的参数；θ_2：是句子预测任务中在编码器接上的分类器参数；

$$L(\theta,\theta_1,\theta_2)=L_1(\theta,\theta_1)+L_2(\theta,\theta_2) \tag{1}$$

在 MLM 的函数中，如果被遮掩的词的集合为 M，因为它是一个词典大小$|V|$上的多分类问题，所用的损失函数叫做负对数似然函数，如公式(2)所示：

$$L_1(\theta,\theta_1)=-\sum_{i=1}^{M}\log p(m=m_i\mid\theta,\theta_1),m_i\in[1,2,\cdots,|V|] \tag{2}$$

在 NSP 的函数中，也属于是一个分类任务的损失函数，如公式(3)所示：

$$L_2(\theta,\theta_2)=-\sum_{j=1}^{N}\log p(n=n_i\mid\theta,\theta_2),n_i\in[\text{IsNext},\text{NotNext}] \tag{3}$$

两个任务联合学习的损失函数如公式(4)所示：

$$L(\theta,\theta_1,\theta_2)=-\sum_{i=1}^{M}\log p(m=m_i\mid\theta,\theta_1)-\sum_{j=1}^{N}\log p(n=n_i\mid\theta,\theta_2) \tag{4}$$

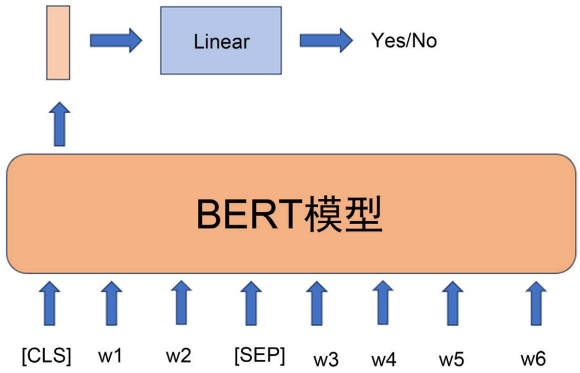

Figure 4. Flowchart for fine-tuning teacher model with NSP
图 4. 用 NSP 微调教师模型的流程图

3.3. 特征提取

特征提取是为了将评论文本转换为数值特征向量。为此，我们进行了两种不同的特征提取，一种用于学生模型，另一种用于教师模型。对于学生模型，我们使用 CountVectorizer 模块创建了一个(1, 3) n-gram 范围的词袋模型，以训练机器学习算法。而对于教师模型，我们使用了一种特殊的标记化工具——BertTokenizer 进行标记化。然后，从每个文本评论中仅使用 510 个标记，同时在每个文本的开头和结尾添加了两个 BERT 模型特定的标记，即[CLS]和[SEP]，如下图 5 所示。接下来，将每个文本填充到最大的 512 个标记，因为 BERT 模型期待这样大小的输入。最后，将每个标记转换为其向量以训练教师模型(BERT)。

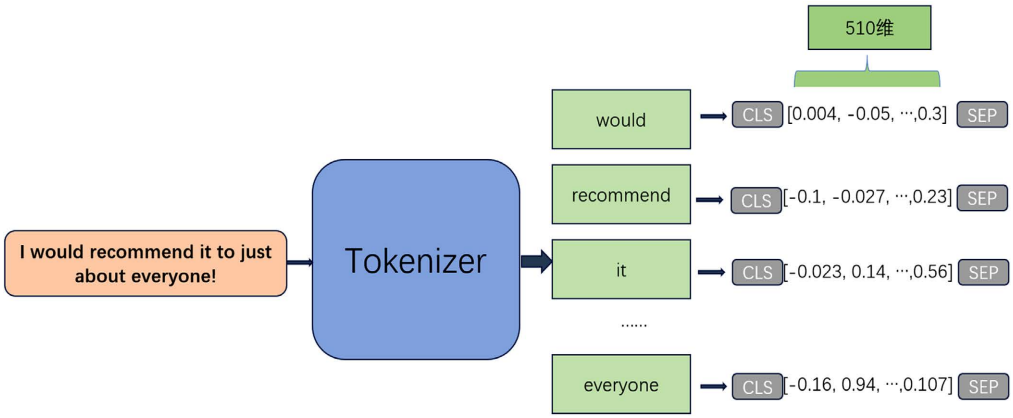

Figure 5. Feature extraction process of teacher model
图 5. 教师模型特征提取图

3.4. 蒸馏架构

如图 6，在这项工作中，知识蒸馏方法实现了基于教师模型输出的知识传递，学生模型学习模仿教师模型在任何给定输入下的行为。与在原始数据集的硬标签上训练学生模型不同，我们将训练学生模型使用教师模型的输出。例如，一些文本具有强烈的情感极性，而其他文本似乎是中性的。但如果我们只使用教师模型预测的独热标签(one-hot)，我们可能会错过与预测不确定性相关的重要知识。因此，我们将训练学生模型使用教师模型的 logits，即在进入 Sigmoid 函数之前教师模型的输出层。在这些输出或 logits 上进行训练可以简化学生的学习过程，另外，学生模型从教师模型获得的所有标签的关系是平等强调的。

Figure 6. Knowledge distillation process
图 6. 蒸馏模型流程图

神经网络的离散概率输出如公式(5)给出，其中， w_i 表示 i 行的 softmax 权值，z 则相当于 $w^{\mathrm{T}}h$ 。

$$\tilde{y}_i = \mathrm{softmax}(z) = \frac{\exp\{w_i^{\mathrm{T}}h\}}{\sum_j \exp\{w_j^{\mathrm{T}}h\}} \tag{5}$$

Softmax 函数的参数被称为“logits”，在 logits 上进行训练使得学生模型更容易学习。蒸馏的目标是对学生网络的 logits 与教师网络的 logits 之间的均方误差(MSE)损失进行弥合，如公式(6)，其中 $z^{(B)}$和 $z^{(S)}$ 分别是教师和学生的 logits。

$$\mathcal{L}_{\text{distill}} = z^{(B)} - z^{(S)}{}_2^2 \tag{6}$$

其他度量方式，比如使用软目标的交叉熵也是可行的。然而，在我们的初步实验中，我们发现均方误差(MSE)表现稍微更好。

在训练时，可以将蒸馏目标与传统的交叉熵损失结合使用，交叉熵损失针对一个独热标签 t 给定，具体如下：

$$\begin{aligned}\mathcal{L} &= \alpha \cdot \mathcal{L}_{\text{CE}} + (1-\alpha) \cdot \mathcal{L}_{\text{distill}} \\ &= -\alpha \sum_i t_i \log y_i^{(S)} - (1-\alpha) \left\| z^{(B)} - z^{(S)} \right\|_2^2\end{aligned} \tag{7}$$

其中，在使用标记数据集进行蒸馏时，独热标签 t 是真实标签。在使用未标记数据集进行蒸馏时，我们使用教师模型的预测标签，即如果 $i = \arg\max y^{(B)}$ 则 $t_i = 1$，否则为 0。

4. 实验

图 7 展示了我们在这项工作中遵循的方法论工作流程。该工作流包括六个步骤，首先是收集 IMDB 数据集、预处理和准备数据。接下来是特征提取、模型构建和所提出的知识蒸馏方法的描述。最后，我们使用准确度度量来评估模型的性能。以下是每个步骤的详细信息。

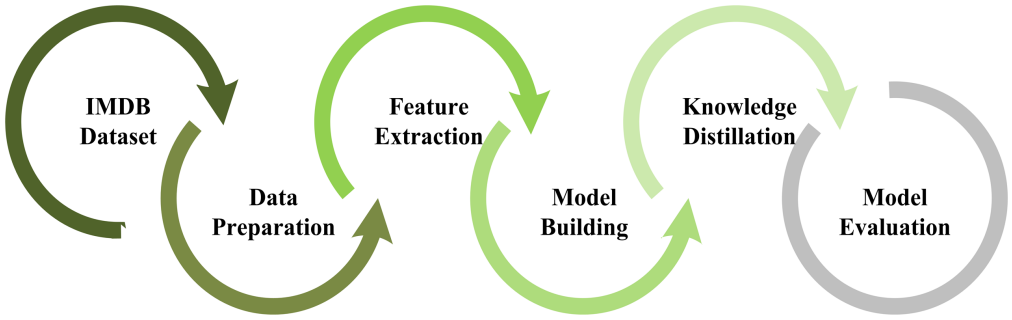

Figure 7. Methodology workflow process
图 7. 实验步骤

4.1. 数据集

这项工作的重点是情感二元分类任务，即将给定的文本分类为积极或消极。为此，我们使用了 IMDB 电影评论——情感分析基准数据集之一。它包含 50,000 条电影评论(25,000 条负面评论和 25,000 条正面评论) [23] [24]。每条记录由两部分组成，即一个评论和一个二进制标签，该标签是用来表明该条评论文本是积极的还是消极的。

4.2. 实验设置

对于 BERT 模型，我们使用大型变种 BERT-Large 作为教师模型，从预训练的权重开始，按照原始的、任务特定的微调过程[1]进行微调。我们使用 AdamW 算法进行微调，学习率为{2, 3, 4, 5} * 10^{-5}，在验证集上选择最佳模型。对于我们的模型，我们将原始数据集与合成的示例一起提供给任务特定的、经过微调的 BERT 模型，以获得预测的 logits。

预处理是执行任何机器学习算法的关键步骤之一，应用这一步骤有助于提高算法的精度。为此，我们将文本根据空白分割为多个标记，并删除了所有标点符号、字母数字单词、停止词和长度为 1 的所有单词。然后，使用自然语言处理工具包(NLTK)将评论中的每个单词转换成小写字母。

4.3. 模型评估标准

模型评估步骤旨在估计基线模型(学生模型)、教师模型和蒸馏模型在未见样本(即测试数据)上的泛化准确性。我们使用 Accuracy 函数来计算准确率，该函数可以在度量模块中找到。计算方式如下方公式(8)所示，其中 TP 表示积极样本，TN 表示消极样本：

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total of observations}} \tag{8}$$

4.4. 实验结果

NB、SVM 和 LR 传统机器学习模型对测试数据使用(1, 3) n-gram 的 BOW 文本表示方法，并采用了从 BERT 模型中获得了不同的数据标签(硬标签和软标签)，从而得到的实验结果如表 1 和图 8 所示，本节对此进行讨论分析。

Table 1. Comparison table of results data
表 1. 结果数据对比表

Metrics	Model Type	Accuracy	Parameters (Millions)
NB	Baseline	78.7	13 M
NB	Distilled	**80.3**	16 M
LR	Baseline	79.5	22 M
LR	Distilled	**80.1**	26 M
SVM	Baseline	79	18 M
SVM	Distilled	**80.1**	29 M
BERT	Baseline	90	**334 M**

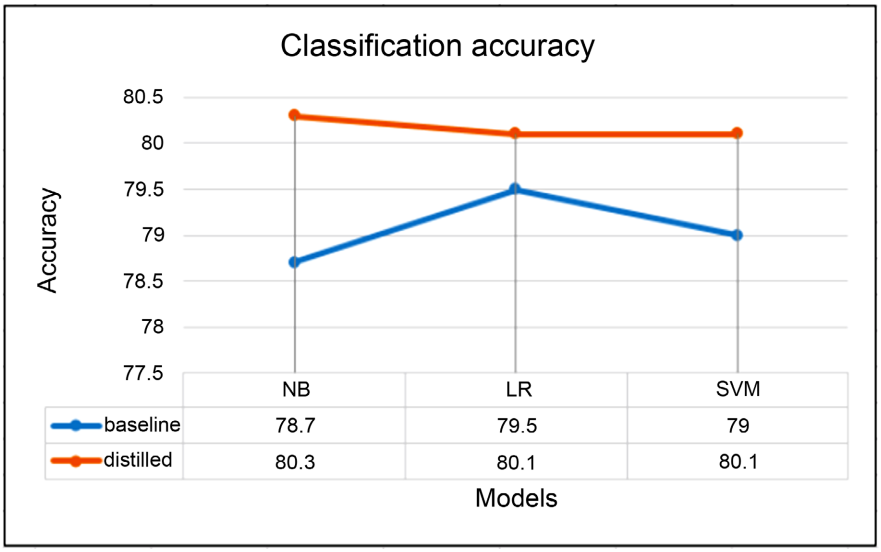

Figure 8. Classifiers accuracy using BOW
图 8. 使用 BOW 的机器学习模型准确度

4.5. 实验结果分析

由表 1 我们得知，使用 BOW 和硬标签的学生模型(NB、SVM 和 LR)产生了几乎相同的准确率结果，但是 SVM 模型相对有一点优势。当应用本文提出的蒸馏方法后，即传统机器学习模型在软标签上训练(BERT 的 logits 输出)时，传统机器学习模型的准确度有所提高，这证明了这种蒸馏方法，即在软标签上训练学生模型而不是硬标签上进行训练是可以提高学生模型在情感分析任务上的性能的。

对于本次实验中，蒸馏后的传统机器学习模型的准确率几乎相同，分别为 80.3，80.1 和 80.1，相比较于之前的表现，准确率普遍提高了约 1%至 2%。其中，蒸馏后的 NB 模型准确率提高了 2%以上，而 LR 和 SVM 相比较于之前准确率提高了约 1%。因此，NB 在所有机器学习模型中取得了最好的成绩。

观察学生模型和基线模型的参数，蒸馏后的学生模型参数相比较基线模型参数略有增加，例如 NB 的参数量相较于基线模型增加了 3 M 个参数，参数量增加最多的是 SVM，增加了大约 11 M 个参数。虽然蒸馏后学生模型参数量会增加，但是相比较于 BERT 预训练语言模型的参数量仍有很大差距，蒸馏后的 NB，LR，SVM 的参数量比 BERT 模型少了大约 19 倍，11 倍和 10 倍，保留住了传统机器学习模型可部署的特点。

4.6. 模型部署

为了验证了我们的学生模型是否可以应用在边缘设备上，通过构建一个用于情感分析的移动应用程序来进行比较，该应用程序在最新的智能手机(iPhone 14)上的平均推理时间与我们之前基于 BERT 模型训练的情感分析模型相对比，在不考虑分词步骤的情况下，我们的蒸馏后的学生模型(NB，LR，SVM)分别比 BERT 模型快了 97.3%，98.3%和 96.5%，具体的推理时间如下表 2 所示。

Table 2. Comparison table of inference latency
表 2. 推理时间比较表

Metrics	NB	NB(Distilled)	LR	LR (Distilled)	SVM	SVM (Distilled)	BERT
Infer.time (second)	2.3	3.4	2.6	2.8	3.1	4.9	158

5. 总结

预训练语言模型在 NLP 领域有着不可或缺的地位，BERT 预训练语言模型就是一个著名的例子，同时 BERT 模型也是 NLP 领域的卓越解决方案之一，尤其在文本分类任务(如情感分析)方面。然而，这些优点常常伴随着一些缺点，比如模型庞大，拥有近百万(甚至数十亿)参数，并且计算成本过高，这成为了该模型在 Web 应用程序、边缘设备或嵌入式设备的客户端部署上的障碍。

本文旨在利用蒸馏方法，通过提取 BERT 模型的知识，增强轻量级传统机器学习模型(可部署模型)在基准电影评论数据集上的情感分析任务性能。研究结果表明，这种策略可以提高 LR、SVM 和 NB 等轻量级机器学习模型在测试数据上的性能，并将准确率提高了超过 1%。

6. 未来工作展望

在未来的工作中，我们计划探索其他蒸馏技术，并尝试将将这些蒸馏技术与传统的机器学习模型相结合，并研究这些方法的能力，以提高具有更轻层和参数的小型神经网络的整体性能。

参考文献

[1] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*, Vol. 1, 4171-4186.

[2] 刘欢, 张智雄, 王宇飞. BERT 模型的主要优化改进方法研究综述[J]. 数据分析与知识发现, 2021, 5(1): 3-15.

[3] 邵仁荣, 刘宇昂, 张伟, 等. 深度学习中知识蒸馏研究综述[J]. 计算机学报, 2022, 45(8): 1638-1673.

[4] Hinton, G., Vinyals, O. and Dean, J. (2015) Distilling the Knowledge in a Neural Network. 1-9. http://arxiv.org/abs/1503.02531

[5] Li, H., Ma, Y., Ma, Z. and Zhu, H. (2021) Weibo Text Sentiment Analysis Based on Bert and Deep Learning. *Applied Sciences*, **11**, Article No. 10774. https://doi.org/10.3390/app112210774

[6] Wei, S., Yu, D. and Lv, C. (2020) A Distilled BERT with Hidden State and Soft Label Learning for Sentiment Classification. *Journal of Physics*: *Conference Series*, **1693**, Article ID: 012076. https://doi.org/10.1088/1742-6596/1693/1/012076

[7] Vashisht, G. and Sinha, Y.N. (2021) Sentimental Study of CAA by Location-Based Tweets. *International Journal of Information Technology*, **13**, 1555-1567. https://doi.org/10.1007/s41870-020-00604-8

[8] Ali Salmony, M.Y. and Rasool Faridi, A. (2021) Supervised Sentiment Analysis on Amazon Product Reviews: A Survey. 2*nd International Conference on Intelligent Engineering and Management* (*ICIEM*), London, 28-30 April 2021, 132-138. https://doi.org/10.1109/ICIEM51511.2021.9445303

[9] Wiegand, M., Balahur, A., Roth, B., Klakow, D. and Montoyo, A. (2010) A Survey on the Role of Negation in Sentiment Analysis. *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, Uppsala, 10 July 2010, 60-68. http://dl.acm.org/citation.cfm?id=1858959.1858970

[10] Ruffy, F. and Chahal, K. (2019) The State of Knowledge Distillation for Classification. 1-8. http://arxiv.org/abs/1912.10850

[11] Gou, J., Yu, B., Maybank, S.J. and Tao, D. (2021) Knowledge Distillation: A Survey. *International Journal of Computer Vision*, **129**, 1789-1819. https://doi.org/10.1007/s11263-021-01453-z

[12] 曾桢, 王擎宇. 融合 BERT 中间隐藏层的方面级情感分析模型[J]. 科学技术与工程, 2023, 23(12): 5161-5169.

[13] Jiao, X., *et al*. (2020) TinyBERT: Distilling BERT for Natural Language Understanding. *Findings of the Association for Computational Linguistics*: *EMNLP*, 16-20 November 2020, 4163-4174. https://doi.org/10.18653/v1/2020.findings-emnlp.372

[14] Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2019) DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. 2-6. http://arxiv.org/abs/1910.01108

[15] Du, C., Sun, H., Wang, J., Qi, Q. and Liao, J. (2020) Adversarial and Domain-Aware BERT for Cross-Domain Sentiment Analysis. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, 5-10 July 2020, 4019-4028. https://doi.org/10.18653/v1/2020.acl-main.370

[16] Ryu, M. and Lee, K. (2020) Knowledge Distillation for BERT Unsupervised Domain Adaptation. 1-11. http://arxiv.org/abs/2010.11478

[17] Sun, S., Cheng, Y., Gan, Z. and Liu, J. (2020) Patient Knowledge Distillation for BERT Model Compression. *Proceedings of the* 2019 *Conference on Empirical Methods in Natural Language Processing and the* 9*th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*), Hong Kong, November 2019, 4323-4332. https://doi.org/10.18653/v1/D19-1441

[18] Song, J. (2019) Distilling Knowledge from User Information for Document Level Sentiment Classification. 2019 *IEEE* 35*th International Conference on Data Engineering Workshops* (*ICDEW*), Macao, 8-12 April 2019, 169-176. https://doi.org/10.1109/ICDEW.2019.00-15

[19] Qin, Q., Hu, W. and Liu, B. (2020) Using the Past Knowledge to Improve Sentiment Classification. *Findings of the Association for Computational Linguistics*: *EMNLP* 2020, November 2020, 1124-1133. https://doi.org/10.18653/v1/2020.findings-emnlp.101

[20] Ren, F., Feng, L., Xiao, D., Cai, M. and Cheng, S. (2020) DNet: A Lightweight and Efficient Model for Aspect Based Sentiment Analysis. *Expert Systems with Applications*, **151**, Article ID: 113393. https://doi.org/10.1016/j.eswa.2020.113393

[21] Shuang, K., Yang, Q., Loo, J., Li, R. and Gu, M. (2020) Feature Distillation Network for Aspect-Based Sentiment Analysis. *Information Fusion*, **61**, 13-23. https://doi.org/10.1016/j.inffus.2020.03.003

[22] Li, Y. and Li, W. (2021) Data Distillation for Text Classification. *Association for Computing Machinery*, **1**.

[23] Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y. and Potts, C. (2011) Learning Word Vectors for Sentiment Analysis. 49*th Annual Meeting of the Association for Computational Linguistics*: *Human Language Technologies*, Vol. 1, 142-150.

[24] Kumar, K., Harish, B.S. and Darshan, H.K. (2018) Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method. *International Journal of Interactive Multimedia and Artificial Intelligence*, **5**, 109-114. https://doi.org/10.9781/ijimai.2018.12.005