分类号 TP391.1 学号 18023019

UDC 004 密级 公开

工学硕士学位论文

基于模型不确定性的细粒度文本情感分析研究

硕士生姓名 郭晓亭

学科专业 计算机科学与技术

研究方向 人工智能

指导教师 王晓东 研究员

国防科技大学研究生院

二〇二一年六月

Model Uncertainty Aware Fine Grained Text Sentiment Analysis

Candidate: Guo Xiaoting

Supervisor: Prof. Wang Xiaodong

A dissertation

Submitted in partial fulfillment of the requirements

for the degree of Master of Engineering

in Computer Science and Technology

Graduate School of National University of Defense Technology

Changsha, Hunan, P. R. China

June, 2021

目 录

摘 要 i

ABSTRACT iii

第一章 绪论 1

1.1 研究背景和意义 1

1.1.1 细粒度情感分析的研究背景和意义 1

1.1.2 模型不确定性的研究背景和意义 4

1.2 研究内容和主要创新点 5

1.2.1 研究内容 5

1.2.2 主要创新点 6

1.3 论文组织结构 7

第二章 相关工作 9

2.1 词嵌入表示 9

2.1.1 Word2Vec 9

2.1.2 GloVe 11

2.1.3 BERT 11

2.2 细粒度情感分析的国内外研究现状 12

2.2.1 细粒度情感分析的研究现状 12

2.2.2 细粒度情感分类的基本方法 14

2.2.3 局限性 19

2.3 模型不确定性的国内外研究现状 20

2.3.1 发展历史 21

2.3.2 研究成果 22

2.4 本章小结 23

第三章 模型不确定性感知的损失函数设计 25

3.1 基于贝叶斯神经网络的不确定性度量 25

3.1.1 研究方法 25

3.1.2 形式化表述 27

3.2 不确定性感知的损失函数设计 29

3.2.1 研究思路 30

3.2.2 不确定性表示 30

3.2.3 方法设计 31

3.3 实验及结果分析 32
3.3.1 数据集 33
3.3.2 baseline 模型 33
3.3.3 实验结果 34
3.3.4 结果分析与可视化 36
3.3.5 测试样本预测结果变化举例 42
3.4 本章小结 42
第四章 基于不确定性的细粒度情感分类模型 43
4.1 研究内容及整体框架 43
4.1.1 研究内容 43
4.1.2 方法设计与整体框架 43
4.2 数据处理 45
4.2.1 数据准备 45
4.2.2 词的嵌入表示 46
4.3 不确定性感知的两阶段修正预测模型 UATR 46
4.3.1 UAL 指导的不确定性估计模型 47
4.3.2 不确定性阈值的选择 47
4.3.3 基于图神经网络的修正预测模型 48
4.4 实验及结果分析 50
4.4.1 实验设置 50
4.4.2 结果分析 50
4.4.3 消融研究 54
4.4.4 讨论 55
4.5 本章小结 56
第五章 总结与展望 57
5.1 全文总结 57
5.2 未来展望 58
致谢 59
参考文献 61
作者在学期间取得的学术成果 67

表 目 录

表 1.1 细粒度情感分析任务总览 2
表 1.2 baseline 模型在 restaurant 上预测结果统计 2
表 1.3 极性预测相反的样本示例 3
表 2.1 baseline 模型在 SemEval14 数据集上极性预测相反样本占总比统计 20
表 2.2 标签值与预测值差值跨度 20
表 3.1 不确定大小与预测结果的指示关系表 30
表 3.2 SemEval14 数据集样本统计 33
表 3.3 各基线模型及其变体在 SemEval14 数据集上的实验结果 35
表 3.4 RGAT 模型在 Restaurant 数据集上极性预测相反的样本修正示例 42
表 4.1 restaurant 数据集采样的数据样例 45
表 4.2 不同模型在 laptop 数据集性能比较 50
表 4.3 不同模型在 restaurant 数据集性能比较 52
表 4.4 UATR 方法实例说明 53
表 4.5 关于 UAL 和 UATR 两阶段修正方法的消融研究 54

图 目 录

图 2.1 lstm 结构示意图 15
图 2.2 rgat 模型示意图[24] 16
图 2.3 attention 机制示意图 17
图 2.4 ian 模型结构示意图[60] 19
图 2.5 普通神经网络的回归输出 21
图 2.6 使用高斯过程的回归输出 22
图 3.1 基于 MC Dropout 的不确定性模型操作示意图 28
图 3.2 UAL 指导模型训练示意图 32
图 3.3 基线模型与 UAL 指导变体模型 mix 的极性预测相反样本占比 opp 对比图. 36
图 3.4 RAM 在 laptop 上误分类样本在不确定性区间的分布图（不同颜色的面积大小代表该不确定性区间的样本在所有误分类样本中所占的比例，下同） 37
图 3.5 RAM 在 restaurant 上误分类样本在不确定性区间的分布图 37
图 3.6 RGAT 在 laptop 上误分类样本在不确定性区间的分布图 38
图 3.7 RGAT 在 restaurant 上误分类样本在不确定性区间的分布图 38
图 3.8 无 UAL 指导的分类模型 RAM_unc 在 laptop 上误分样本累积分布图（横坐标代表不确定性从 0 到 1 间隔为 0.05 的区间，下同） 39
图 3.9 有 UAL 指导的分类模型 RAM_mix 在 laptop 上误分样本累积分布图 . 39
图 3.10 无 UAL 的指导分类模型 RGAT_unc 在 restaurant 上误分类样本累积分布图 40
图 3.11 有 UAL 指导的分类模型 RGAT_mix 在 restaurant 上误分类样本累积分布图 40
图 3.12 无 UAL 指导的分类模型 RAM_unc 在 laptop 上误分样本累积分布图 .. 41
图 3.13 有 UAL 指导的分类模型 RGAT_unc 在 laptop 上误分样本累积分布图 41
图 4.1 基于不确定性的两阶段修正预测模型 UATR 结构图 44
图 4.2 依存关系树示例 45
图 4.3 bert 分类任务应用结构示意图 46
图 4.4 正误分类样本在不确定性区间分布的差异值 48
图 4.5 GAT 模型结构图 49
图 4.6 RGAT 在 laptop 上误分类样本累积分布图（柱状面积的大小代表着小于该区间上界的不确定性区间内误分类样本的数量，下同） 51

图 4.7　UATR 在 laptop 上误分类样本累积分布图 52
图 4.8　RGAT 和 UATR 在 restaurant 上误分类样本累积分布对比图 52
图 4.9　RGAT 和 UATR 在 laptop 上误分类样本累积分布对比图 53

摘　要

随着自然语言处理技术的兴起，情感分析得到了广泛的研究。细粒度情感分析由于其分析对象更具体、更明确，成为近年来的研究热点。主流方法首先获取表征特定目标的句子嵌入层，然后将其馈送到各种经典神经网络以学习特征，最后通过分类器获得给定目标的情感极性。从 CNN，LSTM，RAM，IAN 到 GNN 的各种变体，研究人员挖掘各模型优势尝试对给定目标与其上下文之间的关系进行更好的建模。尽管已经取得了一些成果，但细粒度的多分类情感技术由于其准确率不够高、受数据集规模和类别样本是否均衡影响较大等特点，仍然无法完全应用于正式的分析场合，甚至出现相当比例极性完全预测相反的情况，同时模型不能对预测的结果进行解释。相比之下，通过贝叶斯神经网络建模模型的不确定性，对模型权重的分布进行推断，相当于集成权重分布上的多组神经网络模型，对小数据的学习有更好的鲁棒性，为深度学习模型提供了概率解释。因此，我们从模型的预测结果与模型置信度（不确定性）的关系入手，利用贝叶斯神经网络可以将概率建模与神经网络相结合的特性，对预测结果给出相应的模型不确定性，设计不确定性感知的损失函数 UAL(Uncertainty Aware Loss) 指导模型训练，为预测结果提供更合理的不确定性解释，并且提出不确定性感知的两阶段修正预测模型 UATR(Uncertainty Aware Two-stage Refinement Model)，设计动态获得各批次不确定性阈值的策略，当不确定性大于一定阈值时，样本将会被筛选进行二次预测，建模句子成分间的依赖关系以获得更好的分类效果。

本文的主要内容包括：

1）针对规模较小、类别样本不均衡的数据集中出现极性预测相反比例较大的问题，探究模型不确定性应用于细粒度情感分析任务的有效性。

2）在基于贝叶斯神经网络获取模型不确定性的基础上，设计并实现不确定性感知的损失函数 UAL，使模型预测结果趋于正确分类样本伴随较小不确定性、错误分类样本伴随较大不确定性的区间分布，为预测结果提供可靠的不确定性解释，降低极性预测相反样本的占比。

3）在得到误分类样本在不确定性区间合理分布的前提下，设计得到二次修正不确定性阈值的策略，提出基于不确定性的细粒度情感分析模型 —— 不确定性阈值指导的两阶段修正预测模型 UATR，进一步提高细粒度情感多分类准确率和宏观 F1 值。

通过在标准基准测试集 SemEval14 上设计实验，与目前的 SOTA 模型相比，

本文提出 03 的模型 UATR 在 laptop 数据集上的准确率和宏观 F1 值分别提高了 2.66%，4.19%，restaurant 数据集上的准确率和宏观 F1 值分别提高了 1.7%，2.2%。

关键词: 细粒度情感分析；注意力机制；图神经网络；贝叶斯神经网络；模型不确定性

ABSTRACT

With the rise of natural language processing, sentiment analysis has been extensively studied. Fine-grained sentiment analysis has become a research hotspot in recent years since its analysis objects are more specific and clear. The mainstream method first obtains the sentence embedding layer that characterizes a specific target, and then feeds it to various classical neural networks to learn features, and finally obtains the sentiment polarity of a given target through a classifier. From various variants of CNN, LSTM, RAM, IAN to GNN, researchers tap the advantages of each model to try to model the relationship between a given target and its context better. Although some results have been achieved, the fine-grained multi-class sentiment classification technology cannot be fully applied to formal analysis situations due to its insufficient accuracy and the size of datasets, and even give a considerable proportion of polarities completely predict the opposite. In addition, model cannot give an explanation for predicted results. In contrast, Bayesian neural network can model uncertainties by inferring the distribution of model weights, which is more robust to small data learning, and provides a probabilistic explanation for deep learning models. Therefore, we start with the relationship between the model's prediction results and model confidence, using Bayesian neural networks to combine probabilistic modeling with neural networks, design and implement the loss function UAL of uncertainty perception to guide model training phase and give the corresponding model uncertainty to the prediction results. More important, we propose uncertainty aware two-stage refinement model to correct samples with high possibility to be wrong. Specifically, when uncertainty is greater than a certain threshold, the sample will be selected for secondary prediction. The dependency relationship between sentence components is modeled to get better classification effect.

The main contents and innovations of this paper include:

1) Aiming at the problem of a relatively large proportion of opposite polarity predictions that are prone to appear in small-scale data sets with unbalanced category samples, this paper explores the effectiveness of model uncertainty applied to fine-grained sentiment analysis tasks.

2) On the basis of obtaining the uncertainty of the model, design and implement the uncertainty perception loss function UAL, so that the model prediction results tend

to be the interval distribution of correctly classified samples with less uncertainty, and incorrectly classified samples with greater uncertainty , Provide a reliable explanation of uncertainty for the prediction results, and reduce the proportion of samples with opposite polarity predictions.

3) On the premise of obtaining reasonable misclassified samples in the uncertainty interval, design strategies to determine the uncertainty threshold for the second correction, and propose a fine-grained sentiment analysis model based on uncertainty——Uncertainty Aware Two-stage Refinement model UATR to further improve the accuracy of multi-classification and the macro F1 value .

By designing experiments on the standard benchmark dataset SemEval14, compared with the current SOTA model, the accuracy and macro F1 value of our model UATR on the laptop dataset are increased by 2.66% and 4.19%, respectively. On the restaurant dataset, the accuracy and macro F1 value of the model have been increased by 1.7% and 2.2% respectively.

Key Words: Aspect-based Sentiment Analysis; Attention Machinism; Graph Neural Network; Bayesian Neural Network; Model Uncertainty

符号使用说明

SOTA	最优表现 (the state-of-the-art)
baseline	基准模型
ABSA	方面词级情感分析（aspect-based sentiment analysis）
AOWE	方面词级情感词抽取（aspect-oriented Opinion Wprds Extraction）
E2E-ABSA	端到端的方面词级情感分析（End-to-End Aspect-based Sentiment Analysis）

第一章　绪论

随着人工智能时代的到来，数字化信息已呈现出指数级增长的趋势。这些看似孤立和分散的信息隐藏着复杂而多样的关系。在数据挖掘技术日趋成熟的背景下，自然语言处理（NLP）技术应运而生。人们需要帮助机器理解自然语言，学习语义关系，甚至生成自然语言文本。文本情感分析，也称为意见挖掘，是指使用自然语言处理，文本挖掘和计算机语言学来识别和提取原始资料中的主观信息[1]。

1.1　研究背景和意义

1.1.1　细粒度情感分析的研究背景和意义

1.1.1.1　细粒度情感分析的研究背景

情感分析是一种重要的信息组织方式，研究的目标是自动挖掘和分析文本中的立场、观点、看法、情绪和喜恶等主观信息[2][3]。在社会管理、商业决策和信息预测等各个方面有着广泛而重要的应用价值。纵观情感分析[4][5]的研究历史，其研究方法大致可以概括为三个阶段，2002 年 -2008 年，基于词典的规则方法[6]：早期的情感分析工作主要围绕主客观判定和情感极性判定两方面展开，人们根据给定情感词典获得观点词的极性，通过语法分析器和词典进行语义分析，并定义规则来计算关于否定词、程度副词等的情感分数；2008 年 -2011 年，基于统计的机器学习方法：利用统计机器学习方法，如朴素贝叶斯（NB），最大值熵模型（ME），支持向量机（SVM）等[7]，在有标注的训练数据上训练情感分类器模型，总是依靠大量有标记语料；2011 年至今，基于深度学习的方法：深度学习由于其更强大的特征提取和学习能力而在自然语言处理的许多任务中表现出色，长短时期记忆网络以及注意力机制凭借能够捕获远距离信息和重要信息的优势而变得流行[8][9][10][11]。

情感分析任务的研究对象从最初的篇章级[12][13]、句子级[14][15]，发展为更细粒度的方面词级[16]，即对于一个分析文本中涉及到的所有方面属性，分别预测对应的情感类别，例如在 restaurant 数据集中，句子”the staff is arrogant, the prices are way high for Brooklyn” 中包含两个方面词 staff 和 prices，其中 staff 对应的情感极性是消极的，prices 对应的情感极性也是消极的。细粒度情感分析针对具体的描述对象给出情感类别的预测，其分析对象更明确、更具体，可以更好地满足实际应用的需求，已经成为新的研究趋势。

细粒度情感分析根据不同的需求有多种不同的任务定义形式[17]，具体可以概括为以下三种：第一种是经典的 ABSA (Aspect-based sentiment analysis) 问题，旨在

对句子中给定方面词做情感判断；第二类是 AOWE (Aspect-oriented Opinion Words Extraction)，提取句子中方面词所对应的情感词、观点词等；第三类是 E2E-ABSA (End-to-End Aspect-based Sentiment Analysis)，端到端的方面词情感分析，即首先从句子中提取出现的方面词，然后针对各方面词给出相应的情感判断，是实际应用中比较理想的任务，但因其中还涉及到序列标注问题，实际效果会受到一定的限制。各任务的输入输出如表1.1所示。

表 1.1　细粒度情感分析任务总览

任务	输入	输出
ABSA	句子，方面词	方面词的情感极性
AOWE	句子，方面词	方面词对应的情感词
E2E-ABSA	句子	方面词，方面词对应的情感极性

细粒度情感分析的研究中，衍生了一批用于学术研究的基准数据集。主要包括 SemEval14[18] 发布的 Restaurant 和 Laptop 数据集，Twitter 数据集[19]，斯坦福发布的 SST 数据集[20]，Hu 和 Liu 发布的 digital product 数据集[21]，MPQA 语料库[22]，康奈尔大学发布的 Movie Review 数据集[23]。其中 SemEval14 的 Restaurant 和 Laptop 数据集上的三分类研究最为广泛，在该任务上的最新 SOTA 模型是 2020 年提出的 RGAT 模型[24]，在不引入额外训练数据的情况下，以 Restaurant 数据集为例，最高可以达到 86.59（原论文）的准确率，通过对经典的 baseline 模型以及最新 SOTA 模型的预测结果进行观察分析，发现在错误分类的样本中，极性预测相反的样本几乎占据所有错误分类样本的四分之一到三分之一，统计信息如表1.2所示：

表 1.2　baseline 模型在 restaurant 上预测结果统计

model	accuracy	macro-F1	wrong as opposite(%)
td-lstm	79.11	68.51	32.05
atae-lstm	77.41	66.55	34.78
ram	80.71	71.18	27.31
aen-bert	80.98	70.03	23.94
bert-spc	84.82	77.95	25.29

极性预测相反的样本是指针对该样本中的方面词，通过模型预测得到的情感

极性与标签极性恰恰相反，标签为正向的预测为负向，标签为负向的预测为正向，即预测误差跨度为 2，表1.3给出了极性预测相反的示例 (0 代表负向情感 negative，2 代表正向情感 positive)。在实际的情感分析场景中，相较于中立情感和正负向情感之间的误分，客户更难以容忍地是将正负向情感完全预测相反，尤其在情感极性敏感的应用中，极性预测完全相反将会给后续任务带来很大的误导。

表 1.3　极性预测相反的样本示例

content	aspect	label	pred
... and the **place** is very **clean, sterile**.	place	2	0
The **staff should** be a bit **more friendly**.	staff	0	2
The main **draw** of this place is the **price**.	price	2	0

通过观察，极性预测相反的样本其语句类型主要可以概括为以下两类：

1、有明显的正负向情感词，模型依然判断错误；

2、对方面词的情感描述使用了比较级，针对方面词所表达的情感与句子中出现的情感词相反，无法识别语句真实情感，被情感词所误导。

因此，在细粒度情感分析任务中，不仅要以提高样本整体的预测准确率为目标，将对句子、目标词更深入的嵌入表示和句子成分之间的准确依存性分析作为研究的重点和难点，而且要尽量保证模型预测结果尽可能小地偏离标签值、减小极性预测相反样本占总预测样本的比例。

1.1.1.2　细粒度情感分析的研究意义

随着社交平台的普及和流行，人们可以更及时地获取第一手资讯，更自由地发表自己的观点、看法和评价。自媒体和社交媒体较传统媒体更具时效性和普遍性而成为信息发布和传递的主要媒介，情感分析就是对海量信息进行分析梳理的有效手段。其主要目的是从一段主观文本表述中识别出用户对事件、商品或人物的态度、情感、意见等信息。这样的分析可以应用到对用户的喜好判断、商品推荐、人物画像等任务中，有很高的研究价值，在商业决策、信息预测和社会管理方面都有着重要的意义[2][3]。

商业决策方面：消费者对消费产品进行主观评价，商家对不同消费者针对各商品的评价进行分析，有助于改进产品，也可以为其它消费者提供参考，帮其确定购买意向。

信息预测方面：可以根据社交媒体上的信息进行选情预测，制定对应的宣传策略。

社会管理方面：可以进行舆情监控，网民通过对热点事件的评论表达自己的观点、态度，机构可以对内容进行分析，进而发现民众意见倾向，客观反映舆情状态。亦可对个人进行定点分析，观测其对于某一事件的立场态度是否有随时间推移的动态变化。

细粒度情感分析可以更具体地捕捉文本中涉及到的所有评论对象，更有针对性地挖掘到用户关于事件或商品某个方面更精准的情感倾向和态度立场，并且可以通过收集某一方面的海量信息得到网民关于该事件、该商品某一属性的整体评价；甚至按照时间维度随时间变化进行情感动态监测，对于人们的生产生活有着重要的指导意义。此外，关注预测样本中极性预测相反的样本占比，使模型预测结果尽可能小地偏离标签值、更接近真实值，有助于模型对其预测结果有更好的解释性。

1.1.2 模型不确定性的研究背景和意义

1.1.2.1 模型不确定性的研究背景

深度学习因其强大的非线性拟合能力在很多任务中达到了惊人的预测精度，伴随着深度神经网络越来越强大，模型的复杂度也在增加。复杂度的增加带来了一系列新的挑战，模型解释性就是其中之一。尤其是在医疗诊断或自动驾驶等敏感应用中，必须保证模型对正确特征的依赖，人们希望能够解释机器学习、深度学习模型所学到的知识，以便识别偏差和故障并相应地修改模型。为了构建更具鲁棒性、更能抵抗对抗攻击的模型，在过去的几年中，研究人员尝试了多种方法来探究模型可解释性，具体地包括[25]：

- LIME[26]：通过局部线性逼近来解释模型预测的方法。
- Activation Maximization[26]：探究输入模式对模型预测结果影响最大的方法。
- 特征可视化：通过可视化每层特征，捕捉网络所关注的特征的方法。
- 不确定性估计：利用模型多轮预测结果关联模型不确定性的方法。

在上面的方法中，不确定性估计不仅可以通过模型预测结果和对应的不确定性为模型提供一定的解释性，而且可以设置不确定性阈值来进行后续修正预测，辅助人类决策。建立模型置信度是非常重要的一个指标，但是大多数深度模型都不提供此类信息。回归模型输出一个回归到数据均值上的单个向量；分类模型中，最终经 softmax 输出的概率向量常常被误认为模型置信度。但实际上，即使 softmax 概率输出值很高，模型仍然可能不确定。

深度学习中，不确定性主要分为偶然不确定性和认知不确定性[27]。偶然不确定性是数据集本身存在的误差，与模型无关，是不可避免的；认知不确定性是模型训练中产生的不确定性，其度量的是，测试数据是否存在于训练数据的分布之

中。为了捕捉神经网络的认知不确定性，我们假设模型权重满足一定的先验分布，比如高斯分布。这样的模型称为贝叶斯神经网络，贝叶斯神经网络[28][29]将确定性网络的权重参数替换为这些参数的分布，网络中每个参数的权重都不再是确定的数字，而是权重的先验分布。训练目标不再是直接优化网络权重，而是对所有可能的权重进行平均（称为边际化）。贝叶斯推断用来计算参数的后验分布，在给定数据的情况下，该后验分布捕捉到一组合理的模型参数用于最终的模型预测。

1.1.2.2 模型不确定性的意义

神经网络虽然在多数情况下会给出一个很好的结果，但它们通常被视为黑匣子，无法为其结果提供解释性依据。偶尔会给出一个特别糟糕的结果，而这个特别糟糕的结果在不同领域的容忍度是不一样的。当特别糟糕的结果不被容忍时，我们希望模型在输出这个结果的同时，给出一个非常大的不确定性，以此为指导进行二次干预。基于贝叶斯的深度学习网络模型，与标准的模型相比，能够实现不确定性估计，通过为神经网络的权重引入不确定性进行正则化，相当于集成多组权重分布上的神经网络进行预测，因此可以较容易地从小数据集中学习，对小数据的过度拟合具有更好的鲁棒性[30]。

1.2 研究内容和主要创新点

根据上述对细粒度情感分析任务的背景调研和有关模型不确定性的应用研究，在实际的应用当中，对情感分析预测结果的准确率有很高的要求，模型训练完成后，我们不仅希望在整体预测结果上达到高的准确率，而且希望每个预测值都尽量小地偏离真实值，尤其不希望出现样本极性预测结果与真实值完全相反的情况。虽然深度学习在很多任务中都展现出很强的非线性拟合能力，但是在数据集较少的情况下，模型效果就会大打折扣，出现严重的过拟合现象，存在一定的局限性。为模型引入不确定性度量不仅可以有效地解决过拟合问题，对结果进行预测，而且还可以对预测结果给出相应的不确定性解释。本文探究如何利用模型不确定性提高细粒度情感分析的性能。

1.2.1 研究内容

细粒度情感分析问题虽然已经被广泛研究，研究人员几乎探索了所有的神经网络模型在该任务上的表现，但是仍然面临着多分类准确率不够高、受数据集规模大小以及类别样本是否均衡的影响的问题。主流的深度学习模型是对数据敏感的，模型一定的情况下，数据规模较大的往往可以有更好的实验效果，数据规模小、类别样本不均衡是实际应用中数据集普遍存在的两个特点，这通常会影响模型的性能。为模型引入不确定性在一定程度上相当于集成某权重分布上的无穷多

组神经网络进行预测，可以有效缓解过拟合的问题，对小数据也有更好的鲁棒性，可以削减数据规模太小带来的问题。建模模型不确定性的有效方法是将模型修改为基于贝叶斯神经网络的模型，与标准神经网络不同的是，基于贝叶斯神经网络的模型不再训练固定的权重参数值，而是去拟合权重参数满足的分布，这样模型通过对权重的多次采样和预测结果的反馈尽可能地逼近真实的参数分布。因此，本文探究引入模型不确定性对细粒度情感分类任务的影响，在建模不确定性的基础上设计方法改善数据分布不均衡情况下极性预测相反样本占较大的现象，并且进一步提高细粒度多分类结果的准确率。

本文聚焦数据规模较小、研究较为广泛的基准数据集 SemEval14，基于模型不确定性对细粒度文本情感分析展开研究，任务定义为在基准数据集 SemEval14 (Restaurant 和 Laptop) 上，对给定目标词做情感极性三分类的研究。

首先对数据集进行分析，针对其数据规模较小、类别不均衡的特点，为经典 baseline 模型引入不确定性，观察将其改为基于贝叶斯神经网络的不确定性模型是否可以改善整体的分类效果。在此基础上，探究引入不确定性对模型进行改进的可行性，考虑模型预测结果和相应不确定性的一致性关系，提出不确定性感知的损失函数，将不确定性反馈到模型对预测结果的损失函数上，用来指导模型训练，使模型尽可能以较小不确定性将样本分对、以较大不确定性将样本分错，在为模型预测结果提供更可靠不确定性解释的同时降低情感极性预测相反的样本占所有预测样本的比例。

其次，在上述为模型预测结果提供更可靠不确定性解释的基础上，提出不确定性感知的两阶段修正预测模型。以不确定性损失函数为指导训练模型，得到最终预测样本在不确定性区间的分布，根据不确定性区间上正确预测样本与错误预测样本的分布差异确定进行二次修正的最佳不确定性阈值，我们希望超过该不确定性阈值的错误分类样本尽可能地多，正确分类样本尽可能地少。接下来，以不确定性阈值为界，对大于不确定性阈值的样本进行二次预测，第二阶段的修正模型对第一阶段的基础预测模型形成互补的优势，以此提高模型在细粒度情感分析问题上的准确率。

1.2.2 主要创新点

1、针对经典 baseline 模型在小规模数据集上出现相当比例极性预测相反的情况，利用贝叶斯神经网络建模权重分布的特性，为模型引入不确定性度量，减小数据规模较小带来的问题，达到泛化模型性能的目的。使模型能够意识到对哪些输入样本的预测是不确定的，从而将模型预测结果与模型不确定性关联起来，方便二次干预。

2、为了更好地利用模型不确定性、将其用于指导模型训练，进而降低极性预测相反样本的占比，本文设计并实现了不确定性感知的损失函数，以较小不确定性分错的样本给予较高的惩罚，以较小不确定性分对的样本给予较高的奖励，促使模型参数分布趋于增多小不确定性分对的样本，减少小不确定性分错的样本，在保证多分类准确率的前提下，使模型对其预测结果有更大的把握，为模型预测结果提供更可靠的不确定性解释。

3、在获取模型不确定性的基础上，利用不确定性感知的损失函数使模型趋向于以较小不确定性将样本分对、以较大不确定性将样本分错，进而根据误分类样本在不确定性区间上的分布得到适合二次干预的不确定性阈值，设计两阶段修正预测模型，对不确定性超过阈值的样本筛选，送入基于注意力机制的图神经网络进行二次预测，利用句子成分之间的依存关系进一步丰富句子表示，可以更好地融合不同网络的优势，避免将大量分对的模型修正错误，进一步提高模型的预测精度。

1.3 论文组织结构

本文一共有五章，其内容组织如下：

第一章，从细粒度情感分析和模型不确定性的研究背景和研究意义出发，宏观地介绍了细粒度情感分析所涉及到的子任务、国内外的研究现状和技术手段的发展历程，模型不确定性在国内外的应用现状。随后，总结了当前模型中存在的问题并给出了本文的研究内容和创新点。

第二章，主要从技术原理方面介绍本文相关的技术，首先是词嵌入表示 Word2vec 和 Bert，其次是细粒度情感分类的基础模型，分别从循环神经网络、图神经网络（以 SOTA 模型为例）和注意力机制展开介绍。最后是贝叶斯神经网络与模型不确定性的发展历史和研究成果。

第三章，介绍将标准神经网络模型改进为基于贝叶斯神经网络模型的原理，实现引入不确定性度量的 baseline 模型，提出不确定性感知的损失函数。通过在 SemEval14 两大经典数据集上进行实验，从模型准确率、宏观 F1 值等方面进行对比分析，降低极性预测相反的样本占比，验证引入不确定性方法的有效性，为模型预测结果提供不确定性解释。

第四章，基于不确定性的细粒度多分类情感分析，提出了不确定性感知的两阶段修正预测模型。首先阐述了模型的研究动机和整体框架，然后分别就变分 Dense 草稿标签模型和基于注意力机制的图神经网络修正模型的原理和实现展开叙述，通过与 2020 年的 SOTA 模型进行实验对比，进行模型性能的验证。

第五章，总结与展望。主要包括对本文工作内容的总结，以及对未来工作的展望。

第二章　相关工作

本章从解决细粒度情感分析问题的关键步骤出发，详细介绍了主流的词嵌入方法的原理，细粒度情感分析任务的研究现状、分析了当前方法的局限性，最后介绍了模型不确定性的发展历史和最新研究成果。

2.1　词嵌入表示

词嵌入是将词汇映射到实数向量的方法总称。经过词嵌入，可以把自然语言转化为计算机能够识别的数字编码。词嵌入的提出旨在捕捉文档中词语的上下文关系，语义、语法相似性等。最简单的词嵌入方法是 one-hot 编码，该编码方式是建立一个词汇表长度的全零向量，将每个单词在词汇表对应的索引维度置为 1，其它元素保持不变，就得到了每个词的 one-hot 向量。这种编码方式虽然易于实现，但有着很明显的缺陷，在词汇表较大的时候很容易陷入维度爆炸的困境，而且向量编码很稀疏，存储和计算代价很大，词与词之间彼此独立，完全不能反映词语之间的语义远近关系。考虑到这些因素，在实际的应用中，one-hot 作为词向量的场景并不常见。接下来将主要介绍目前主流的分布表示 Word2vec[31]、GloVe[32]、Bert[33]。

2.1.1　Word2Vec

word2vec[31] 是 Mikolov 提出的分布式词表示方法，是使用浅层神经网络学习单词嵌入的最流行的技术之一。分布式表示的思想是使具有相似上下文的单词占据相对紧密的空间位置，在数学上，矢量之间的角度的余弦值接近 1，即角度接近 0，越表示两个向量之间的距离小，对应到词，则两个词之间的语义环境较为近似。word2vec 是神经语言模型的产物，神经语言模型用来判断一个句子出现的概率，输入词序列，经过两层非线性变换，输出该词序列成句的概率。在训练神经语言模型时，作为中间参数的词向量就是我们需要的。

word2vec 舍弃了神经语言模型的两个非线性层，降低了复杂度，提出了两种高效的模型 CBOW(Continuous Bag-of-Words) 和 Skip-Gram。CBOW 模型的出发点在于给定被预测单词的固定窗口大小的上下文，来预测该单词，即我们的预测目标如下（式2.1所示）：

$$p\left(w_t \mid w_{t-k}, w_{t-(k-1)}, \cdots, w_{t-1}, w_{t+1}, \cdots, w_{t+(k-1)}, w_{t+k}\right) \tag{2.1}$$

具体做法是，输入层采用上下文单词的 one-hot 向量 $x_{t-k} \cdots x_{t+k}$，加和后经权重矩阵 W 映射，取平均值作为隐藏层的向量 h（如公式2.2所示），隐藏层到输出

层再经权重矩阵 U 变换，最后通过 softmax 操作得到每个单词的概率值 y（如公式2.3所示）。

$$h = \frac{W^T \cdot x_{t-k} + \cdots + W^T \cdot x_{t-1} + W^T \cdot x_{t+1} + \cdots + W^T \cdot x_{t+k}}{2 \cdot k} \tag{2.2}$$

$$y = \text{softmax}\left(U^T \cdot h\right) \tag{2.3}$$

最终的优化目标就是使被预测单词所对应的概率值最大 (V 是词汇表的大小)，记为 L（如公式2.4所示）:

$$L = \prod_{t=1}^{V} p\left(w_t \mid w_{t-k}, w_{t-(k-1)}, \cdots, w_{t-1}, w_{t+1}, \cdots, w_{t+(k-1)}, w_{t+k}\right) \tag{2.4}$$

Skip-Gram 的目标和 CBOW 相反，它是根据当前词 W_t，预测窗口大小为 k 的上下文单词，其预测目标如下（如式2.5所示）:

$$p\left(w_{t+p} \mid w_t\right)(-k \leq p \leq k, k \neq 0) \tag{2.5}$$

类似地，Skip-Gram 的输入层也是词的 one-hot 表示，记为 x_t，经权重矩阵 W 映射到隐藏层得到 h（如公式2.6所示），再经权重矩阵 U 映射到输出层。与 CBOW 不同的是，损失函数变成了 2k 个词损失函数的总和（如公式2.8）。

$$h = W^T x_t \tag{2.6}$$

$$y = \text{softmax}\left(U^T h\right) \tag{2.7}$$

$$L = \prod_{t=1}^{V} \prod_{p=-k, t \neq 0}^{k} p\left(w_{t+p} \mid w_t\right) \tag{2.8}$$

word2vec 中的 CBOW 和 Skip-Gram 模型在最后一层都有 softmax 的操作，每次都是在词汇表大小维度 V 上的映射，计算代价较大。因此，采用了两种加速训练的方式，分别是层级 softmax 和负采样，来降低复杂度。层级 softmax 采用的数据结构是哈夫曼树，根据词频大小来建树，遵循高近低远的原则来构建与根节点的距离，词频低、距离远的节点参数就越多，这样在训练的过程中，词频较低的路径上的参数就可以被更多的训练和更新，因此效果会更好，同时复杂度也从原来的 $O(V)$ 降为 $O(\log V)$。负采样，提高那些相对来说词频不高的词的词向量精度，将词以一定的概率舍弃。重采样高频词去掉哪些经常出现而没有实际意义的词，降低他们出现在训练样本中的频率。在增大正样本的概率的同时降低所有负样本的概率，根据某种负采样策略挑选负例，使其预测概率尽可能小，既可以减

少计算量，又可以达到近似的效果。

综上，word2vec 省掉了 NNLM 中非线性的隐含层，实现了在大文本数据集上训练出维度在 50-100 之间的词向量，不仅在传统词相似度任务上效果好，而且在词类比任务上也取得了惊人的效果。

2.1.2 GloVe

word2vec 只考虑了上下文信息，没有考虑到全局信息。GloVe[32] 的全称是 Global Vectors for Word Representation，即全局的词向量表示，其初衷是基于共现矩阵学习 word2vec 中词类比任务上的线性关系。本质上，GloVe 是具有加权最小二乘目标的对数双线性模型，是一种基于统计矩阵使用机器学习的混合方法。两个单词之间的欧几里得距离为测量单词的语义相似度提供了一种有效的方法，但是两个单词往往容易显示出可以比单个数字捕获更为丰富的关系的能力，比如“国王”与“王后”相似，因为两个词都描述了一种头衔，代表着人类，但是另一方面，这两个词通常又被认为是相反的，因为它们突出了性别的不同主体。如果想以定量的方式来衡量区分语义相近词义相反的单词之间的细微差别，那么需要将多个数字与单词对相关联，两个词向量之间的向量差就是一个自然的选择，GloVe 设计的目的就是使这样的向量差尽可能多地捕获两个单词并列所指定的含义。

GloVe 模型在全局的单词 - 单词共现矩阵的非零实体上训练，该矩阵列出了单词在给定语料库中彼此共现的频率，填充该矩阵虽然需要遍历整个语料库来收集统计信息，但是这个过程只需进行一次，随后的训练迭代要快得多，因为非零矩阵实体的数量远小于语料库中单词的总数。GloVe 的训练目标是学习单词向量，使其点积等于单词共现概率的对数。由于比值的对数等于相应对数之差，因此该目标将共现概率的比值对数与词向量空间中的向量差关联起来。这些比值可以编码某种形式的含义，所以此类信息也被编码为矢量差[34]，通过 GloVe 生成的词向量在衡量 word2vec 的词类比任务上也取得了非常好的效果。

2.1.3 BERT

BERT[33] (Bidirectional Encoder Representations from Transformers) 通过在海量语料的基础上运行自监督学习方法为单词学习好的特征表示，是一个多任务模型，分别是掩码语言模型 MLM(Masked Language Model) 和 NSP(Next Sentence Prediction)，采用了多层 Transformer 架构，通过多种 attention 机制将任意位置的两个单词的距离转化为 1，有效地解决了 NLP 任务中的长期依赖问题，而且规避了 CNN 无法很好利用序列顺序的问题以及 RNN 不能大规模并行化的问题。MLM 的思想和 CBOW 有相通之处，在训练的时候随机从训练语料中掩码屏蔽掉一些单

词，通过上下文预测该单词，在 BERT 的实验中，采取的策略是，15% 的单词被随机屏蔽，模型训练时，一个句子会被多次送到模型中学习，其中有 80% 的几率会被替换为 [Mask]，10% 的几率随机替换为任意词，还有 10% 的几率保持原来的单词不变。通过预测被屏蔽掉的内容，让网络来学习通用的词义、句法和语义信息。BERT 的输出结果，其一是训练后每个 token 的词向量，另一个池化输出，是每个句子开头填充的 [CLS] 所代表的句子向量的输出，常用于分类任务。

BERT 训练的词向量可以有效缓解 Word2vec 静态词向量无法表示一词多义的问题，针对这个问题，ELMO[35] 和 GPT[36] 也采取了不同的举措，ELMO 除了嵌入词语特征外，还通过前后向语言模型联合学习单词的表示，根据不同的上下文语境动态地调整词语的嵌入表示。ELMO 确实在一定程度上解决了一词多义的表示问题，但是其拼接式的双向融合特征能力以及 LSTM 抽取特征的能力都较后来的 GPT 和 BERT 更弱。GPT 就采用了 transformer 结构，与 BERT 不同的是，GPT 采用的是 transformer 的 Decoder 部分，decoder 端的多头自注意力需要做 mask，需要将当前预测的单词及其之后的单词全部 mask 掉，因此 GPT 是单向的，学习效果较 BERT 逊色。关于 BERT 的网络深度与学习的内容，Ganesh 通过实验佐证[37]，在 BERT 预训练的过程中，低层网络结构学习短语句法信息，中层网络学习丰富的语言学特征，高层网络则学习丰富的语义信息特征。

2.2 细粒度情感分析的国内外研究现状

细粒度情感分类是细粒度情感分析的任务之一，即只需要对给定的目标词作情感极性的判断即可，具体地，可分为二分类和多分类。本节首先介绍细粒度情感分析总任务的研究现状，包括一般化的处理流程和近年来在模型上的一些新的尝试；然后，围绕本文的研究目标细粒度情感分类，详细介绍了应用于该任务的主流方法的原理。最后，分析了当前的方法普遍存在的问题。

2.2.1 细粒度情感分析的研究现状

本节从解决细粒度情感分析问题的一般化流程和主流方法两个方面展开介绍。

2.2.1.1 一般化流程

细粒度情感分析任务涉及到的所有子任务可以概括为：词嵌入表示，术语检测，情感判断三个方面。

词的嵌入表示是建模语言和捕获特征，测量术语之间的相似性以及成为深度学习模型学习语义关系的向量输入的重要组成部分。Word2vec，GloVe，Bert 已被用于现有研究中的情感分析，其具体原理如上一节描述。概括来说，word2vec 和 Gloves 是无上下文的分布式表示形式，它们编码潜在的语义信息，而 Bert 作

为预训练模型的杰出代表，提供了上下文相关的动态词向量，人们可以根据需要适应域的单词向量对其进行微调。陶[38] 设计了一个针对 textCNN，BiLSTM，AT-BiLSTM 的实验，以比较不同的单词矢量表示将如何影响模型的性能。实验结果表明，以上三种词嵌入方法均取得了较好的效果，其中 Bert 效果最佳，glove 运行时间最短，word2vec 效果和运行时间均处于中间。因此，可以选择 word2vec 来在性能和时间之间进行权衡。由于 Bert 具有强大的特征提取功能，因此它在泛化方面具有更好的性能，如果追求最终效果并拥有足够的计算资源，则可以选择 Bert。

方面和观点方面的术语检测是基本部分。这部分的方法可以分为基于无监督词典的方法[39]，统计机器学习方法和基于深度学习的模型[40][41][42]。基于无监督词典（WordNet[43]，NTUSD[44]，MPQA[22]，Hownet[45]，中国情感词汇本体[46]）的方法主要利用给定的词典和句法分析器来检测意见项。同时，将通过对不同依赖解析器的依赖关系来检测方面项，否定词，程度副词等。统计机器学习和基于深度学习的方法主要分为有监督方法和无监督方法。无监督的方法包括主题建模，频繁的模式挖掘和标签传播，而有监督的方法总是可以归因于解决序列标签问题。He 等[47] 提出了一种无监督的神经注意力模型，用于方面词提取，并在餐厅语料库和啤酒语料库上均取得了不错的效果。

情感判断是最后的步骤，旨在用先前的信息推导相应的情感极性或结果。在基于经典无监督词典的传统方法中，人们根据给定词典获得见语的极性，并定义用于计算否定词，程度副词等的情感分数的规则。统计机器学习方法，如朴素贝叶斯（NB），最大熵模型 (ME)，支持向量机（SVM）[7] 始终依靠大标记语料库来训练分类器模型。深度学习由于其更强大的特征提取和学习能力而在自然语言处理的许多任务中均胜过它们，因此人们已经开始尝试将深度学习应用于情感分析。长时记忆网络和短时记忆网络以及注意力机制的基本结合是当前基于深度学习方法的重要组成模块[8][9][10][11]。Zhang 等人[5] 和 Zhou 等人[16] 详细总结了基于 CNN，RNN，记忆和注意力的经典深度学习模型。

2.2.1.2 最新研究趋势

胶囊网络和迁移学习是细粒度情感分析任务上的两个新的尝试方向，近年来，学者们开始逐步探索其分别在捕获分层信息和领域迁移方面带来的优势。

Hinton 提出的胶囊网络[48] 通过提取矢量形式的特征来补偿卷积神经网络的表示缺陷。其出发点是在神经网络中构建更多的结构，然后希望这些新结构可以帮助模型更好地进行泛化。Wang[49] 首次尝试通过基于 RNN 的胶囊模型执行情感分析，他设计了一个具有属性，状态，三种模型的简单胶囊结构，并使每个胶囊专注于一个特定的情感类别。他们的工作在 Movie Review 和 SST 数据集上实现了 2018

年的 SOTA 效果。他们于 2019 年提出了纵横比情感胶囊模型（AS-Capsules）[50]，它可以将方面词识别和在方面词级的情感分类有效地联合起来，通过共享组件利用了方面词和相应情感之间的关联。实验结果表明，SemEval14 餐厅语料库具有良好的性能。此外，他们通过将模型直接应用于 Yelp 上的其他餐厅评论来验证模型的鲁棒性。Du 等在文献[51] 提出了一种具有交互作用的胶囊网络（IACapsNet），通过 EM 路由算法将特征特征表示为聚类特征，从而在基准 SemEval14 和 twitter 上产生了 SOTA 结果。最近，Su 等人将 XLNet 和胶囊网络的组合，称为 XLNetCN[52]。他们通过使用具有动态路由算法的胶囊网络来构造辅助和局部特征表示，从而生成与全局句子方面相关的表示。他们的工作解决了特定于方面和本地特征的问题，并在 SemEval14 上取得了最新的研究成果（引入了辅助语句和领域知识）。可见，胶囊模型是一个很有前途的结构，可以在基本模型中使用以捕获分层信息。

现有的基于评估对象的端到端情感分析的研究仅考虑单个域的性能，而忽略了跨域迁移和归纳的能力。高质量的数据标注昂贵，将迁移学习应用于各种领域的应用成为可行的。Fang[53] 使用了基于 Bert 的迁移学习模型来解决预定义知识结构缺失的问题，Li 等人提出了一种双重记忆交互（DMI）机制[54]，以自动捕获评估者和意见词之间的隐藏关系。DMI 通过与本地内存（神经网络隐藏状态）和全局评估程序内存（点单词内存）进行多次交互来推断每个单词的关系表达式，以便可以对在该任务下迁移的知识进行排序。此外，他们提出了选择性对抗学习（SAL），以使句子中的重要方面保持一致。所有这些设置都极大地影响了从相对域到目标域的知识利用，这对于跨域迁移和泛化能力是一个非常有意义的方向。

2.2.2 细粒度情感分类的基本方法

基于深度学习的细粒度情感分析方法在二十世纪十年代以来得到了广泛的研究，以 RNN、GNN、Attention 为基础模型的研究交替达到了各数据集的 SOTA 水平，接下来将主要围绕循环神经网络、图神经网络和注意力机制的原理及其变体模型展开介绍。

2.2.2.1 基于循环神经网络

基于循环神经网络 RNN(Recurrent neural network) 的模型由于序列建模的优势在许多自然语言处理任务中都显示出了强大的功能。LSTM(Long Short Term Memory)[55] 是一种特定形式的循环神经网络，是为了解决 RNN 在处理长期依赖时面临梯度消失或梯度爆炸的问题而提出的一种门限 RNN(Gated RNN)[56]。Tang 等人[57] 为了更灵活地捕获方面词及其上下文之间的语义关系，提出了 ABSA 任务中经典 baseline 模型：目标依赖的 LSTM(TD-LSTM) 和目标连接的 LSTM(TC-LSTM)，显式地将目标词和上下文以及目标词表示和上下文词语表示结合起来。

在 transformer 提出之前，基于循环神经网络的模型占据了情感分析模型的主体，即使其特征抽取能力与 transformer、Bert 相比较弱，但是由于其模型轻量化、而且在处理序列任务时保留了顺序特征，依然是主流模型的重要组成部分。结构示意图如图2.1所示：

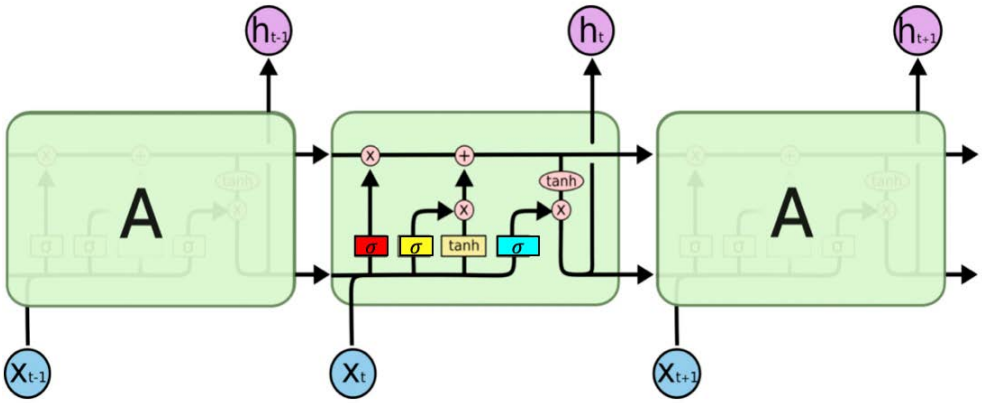


图 2.1　lstm 结构示意图

LSTM 通过引入遗忘门、输入门和输出门使得自循环的权重是变化的，当 t 时刻来临时，长时记忆单元 C_{t-1} 通过遗忘门遗忘一些过去的信息，输入门则将当前时刻的部分新信息 $\tilde{C}_t$ 写入长时记忆单元，激活当前时刻的长时记忆单元准备输出。输出门会把目前积累下来的长时记忆 C_t 利用 $tanh$ 激活函数选出部分相关的记忆生成 t 时刻的短时记忆 h_t，并把这部分记忆输出。其中，遗忘门、输入门、输出门受当前时刻的外部输入 x_t 和上一时刻的短时记忆 h_{t-1} 控制，具体的计算公式如下：

$$\begin{aligned} f_t &= \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \\ i_t &= \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \\ \tilde{C}_t &= \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \\ C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \\ o_t &= \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \\ h_t &= o_t * \tanh\left(C_t\right) \end{aligned} \tag{2.9}$$

2.2.2.2　基于图神经网络

尽管传统的欧氏空间图片声音等数据有很好的平移不变性，但是对于实际应用场景中的网络数据而言，大部分都是非欧氏空间生成的，没有平移不变性。图神经网络 GNN(Graph Neural Network) 是在图结构上进行操作的一种神经网络模型，用来建模图中节点之间的相互依赖关系。在细粒度情感分析任务中，图神经

网络近年来被用来编码句子的语法结构，获得了不错的结果。具体地，研究人员利用图卷积神经网络 GCN[58] 和图注意力网络 GAT[59] 来编码句法特征以便目标词和上下文可以交换信息、根据依存树捕捉词语之间的相互依赖关系。

Bai 等人[24] 提出的 RGAT 模型在 SemEval14 上取得了最新的 SOTA 效果（不引入额外训练数据），该模型以图注意力网络为基础，融合了关系扩展。与已有工作不同的是，他们利用了依存关系信息并证明了它的有效性；使用了独立的编码器建模结构化信息，而不是用语法结构去修正上下文表示。论文中指出，RGAT 模型双编码器结构有以下优势：结构编码器是非常灵活的，而且可以相对容易地应用于新的序列编码器模型（尤其是当句子长度和图中节点数不一致时）；可以减少自动解析的依存树带来的误差传播，因为损失不会从树的表示反向传播到序列编码器上。实现上，以句子的依存树建立句法图，$G = (V, A, R)$，其中，V 代表图中的节点，即句子中的单词。A 是邻接矩阵，当两个词之间有依存关系时，其值为 1，单词与其本身的邻接矩阵值为 1，否则为 0。除此之外，R 是标签矩阵，记录邻接矩阵对应位置值为 1 的标签，否则 $R_{ij} = None$。对应的模型结构图如图2.2所示：

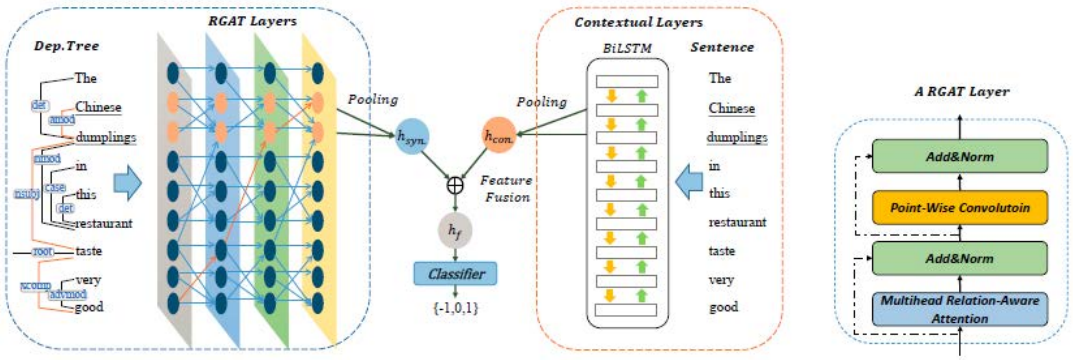


图 2.2　rgat 模型示意图[24]

RGAT 将关系特征整合到注意力计算和融合过程中，以获得具有更丰富信息的表示。RGAT 计算两种注意力分布，分别是节点感知的注意力 e_{ij}^N 和关系感知的注意力 e_{ij}^R，将两者的组合一起作为最后特征融合的注意力权重，其中 f() 是缩放点积注意力函数，r_{ij} 是关系标签矩阵映射为关系嵌入维度的向量。

$$e_{ij}^N = \begin{cases} f\left(h_i^{l-1}, h_j^{l-1}\right), & j \in \mathcal{N}(i) \\ -\inf, & \text{otherwise} \end{cases} \tag{2.10}$$

$$e_{ij}^R = \begin{cases} f\left(h_i^{l-1}, r_{ij}\right), & j \in \mathcal{N}(i) \\ -\inf, & \text{otherwise} \end{cases} \tag{2.11}$$

在残差链接和标准化之后，又添加了一层点对点卷积变换层，使每个节点可以容纳更多信息。

2.2.2.3 基于注意力机制

注意力机制从计算机视觉领域迁移而来，因其能摆脱距离依赖、较好地捕捉句子中所有单词之间的关系而在众多的自然语言处理任务中崭露头角，主要应用于机器翻译、机器问答和机器阅读理解等任务中。

注意力机制[9]是计算某个时刻的输出在输入 x 的各个部分上的相关性，也就是权重，在注意力机制的考量中，有三个重要的组成部分：Query，Key，Value，通过 Query 和 Key 中的元素得到最终要求解的注意力系数，当 Query 和 Value 相同时，计算句子中各个成分与其它成分之间的注意力权值，此时称为 self-attention；当 Query 和 Value 不一样时，可以理解为 encoder-decoder attention，计算 encoder 序列中各个成分对 decoder 序列中各个成分的注意力权值。self-attention 的特点在于无视词之间的距离直接计算依赖关系，从而能够学习到序列的内部结构，实现起来也比较简单，较 RNN 和 LSTM 更容易捕获句子中长距离的相互依赖的特征。而且还可以轻松并行。Transformer 中提出的多头注意力机制，将模型分成多个头，形成多个子空间，让模型去关注不同方面的信息，类似于 CNN 中多个卷积核的作用，多头的注意力有助于网络捕捉到更丰富的特征信息。注意力机制主要有三个步骤（如图2.3所示）：

score-function：度量环境向量和输入向量的相似性，有多种计算方式（公式2.12），其中 transformer 结构中的注意力分数计算对点积计算方式进行了进一步的缩放，$\frac{Q^T K}{\sqrt{d_k}}$，减小数值，使梯度变大，加速了神经网络的计算。

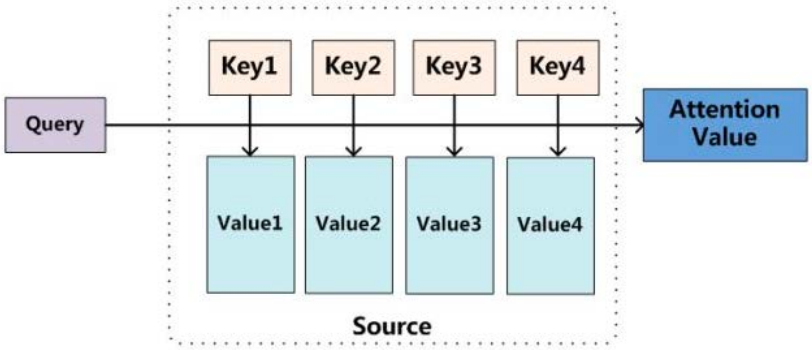


图 2.3 attention 机制示意图

$$f\left(Q,K_i\right)=\begin{cases}Q^TK_i & \text{dot}\\ Q^\tau W_\alpha K_i & \text{general}\\ W_d\left[Q;K_i\right] & \text{concat}\\ v_a^T\tanh\left(W_aQ+U_aK_i\right) & \text{perceptron}\end{cases} \tag{2.12}$$

alignment function：对上述权值使用 softmax 进行归一化（公式2.13）。

$$a_i=\text{soft}\max\left(f\left(Q,K_i\right)\right)=\frac{\exp\left(f\left(Q,K_i\right)\right)}{\sum_j\exp\left(f\left(Q,K_j\right)\right)} \tag{2.13}$$

generate weighted value：根据上一步计算出来的归一化注意力权值，对 value 进行加权求和，即可得到注意力分布后的向量（公式2.14）。

$$\text{Attention}\left(Q,K,V\right)=\sum_t a_iV_i \tag{2.14}$$

Ma 等人[60] 在 2017 年首次提出了交互式注意力建模机制应用于细粒度情感分析中，交互式地学习特定方面词和上下文之间的权重关系。现在主流的注意力机制主要是基于基础 attention 和交互式 attention 进行改进，下面将主要介绍交互式 attention 机制的原理。

交互式注意力提出之前，研究人员通过将目标词表示和上下文单词表示结合起来等方法生成特定于目标词的表示来建模上下文，往往没有对目标词采用单独的建模。IAN 的出发点是利用目标词和上下文之间的相互影响对其协同建模，目标词和上下文都需要运用注意力机制得到不同词的注意力权重，论文中的原始模型结构图如图2.4所示。

以 LSTM 为基础模型，得到目标词和上下文的隐藏状态表示，对各个时间步的隐藏状态值进行加权平均得到初始的向量表示，c_{avg} 和 t_{avg}，分别用 c_{avg} 对目标词的各个时间步隐藏状态值做注意力计算（公式2.15），t_{avg} 对上下文各个时间步隐藏状态值做注意力机制计算（公式2.16），加权求和后得到目标词和上下文的最终表示，将其拼接，得到整体的句子向量表示。

$$\alpha_i=\frac{\exp\left(\gamma\left(h_c^i,t_{avg}\right)\right)}{\sum_{j=1}^n\exp\left(\gamma\left(h_c^j,t_{avg}\right)\right)} \tag{2.15}$$

$$\beta_i=\frac{\exp\left(\gamma\left(h_t^i,c_{avg}\right)\right)}{\sum_{j=1}^m\exp\left(\gamma\left(h_t^j,c_{avg}\right)\right)} \tag{2.16}$$

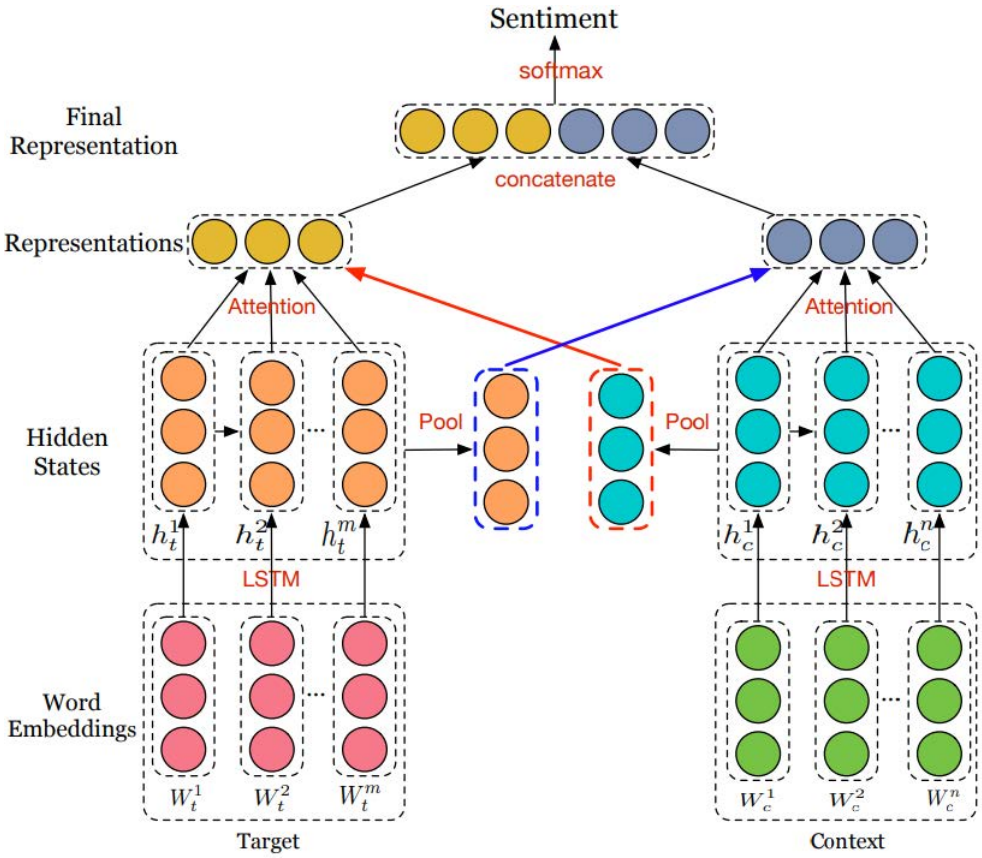


图 2.4　ian 模型结构示意图[60]

$$\begin{aligned} c_r &= \sum_{i=1}^{n} \alpha_i h_c^i \\ t_r &= \sum_{i=1}^{m} \beta_i h_t^i \end{aligned} \tag{2.17}$$

2.2.3　局限性

在细粒度情感分析被广泛研究的今天，研究人员分别从是否基于句法、是否采用动态词嵌入表示等方面入手，提出了基于各种基础神经网络模型的变体，这些模型的提出都是基于更完全地挖掘句子中各成分之间的相互依存关系、更丰富地表达词语的嵌入表示的初衷，经过我们有限的实验分析，在当前的研究环境下，无论是基于 Bert 的、还是基于语法结构的模型，极性预测相反的样本都占据了一定的比例，高达四分之一到三分之一（如表2.1所示），与理论上预测标签误差跨度为 2 的样本比例一致 (如表2.2所示，错误分类的所有可能情况中，跨度为 2 的占 1/3，跨度为 1 的占 2/3)，而不是惯性思维下模型更容易将正负向的样本和中性样本混淆，因此，我们针对标准神经网络模型应用于细粒度情感分析任务中存在的极性预测相反样本占比较大的问题，基于贝叶斯神经网络对主流模型进行改进，让模型在参数分布上多次采样，避免标准神经网络模型静态地以单次高 softmax 值输出不确定的预测结果，将模型预测结果与不确定性结合起来，使模型对其预

测结果更有把握，在保证多分类情感分析准确率的前提下，降低极性预测相反样本的占比。

表 2.1　baseline 模型在 SemEval14 数据集上极性预测相反样本占总比统计

model	laptop	restaurant
td-lstm	26.88	32.05
atae-lstm	22.53	34.78
ram	25.58	27.31
aen-bert	19.18	26.70
bert-spc	18.06	25.29

表 2.2　标签值与预测值差值跨度

预测值 / 标签值	0	1	2
0	0	-1	-2
1	1	0	-1
2	2	1	0

2.3　模型不确定性的国内外研究现状

不确定性最初被用在生命科学相关的领域，在这些领域中，研究人员需要模型返回对预测结果的置信度，不仅仅是整体的预测准确率。例如医生在根据病人的病历建议使用某些药物时，病人希望医生能够对他的建议有足够的把握。而此类信息在人工智能领域也有很大的应用需求，强化学习中的 agent 不断探索环境以使得奖励最大化，也是不确定性发挥作用的例子[28]，agent 会尽量减少其对不同动作的不确定性，在新动作的探索与对已知知识的利用进行权衡。不确定性信息对于从业人员也很重要，在使用深度学习模型解决视觉或自然语言方面的不同任务时，了解模型是否不太确定或错误地过分自信可以帮助获得更好的性能。不确定性量化主要针对回归任务和分类任务。传统的可以建模模型不确定性的方法包括计算机视觉应用中的粒子滤波和条件随机场。但是，现在许多应用要求使用深度学习以达到最先进的性能，而用于分类和回归的标准深度学习工具不能捕捉模型不确定性，大部分不确定性估计算法都是基于贝叶斯神经网络的。

2.3.1 发展历史

贝叶斯神经网络与深度神经网络的不同处在于，其权重参数均为随机变量，而非确定的值，通过将概率建模和神经网络结合起来，得到预测结果的不确定性。其先验用来描述关键参数，并作为神经网络的输入，神经网络的输出则用来描述特定的概率分布的似然，通过采样或变分推断来计算参数的后验分布。因此，贝叶斯神经网络具有不确定性量化能力，具有非常强的鲁棒性，深度神经网络利用可用的数据、定义好的网络结构和损失函数，优化获得点估计的预测，拟合的图像确定唯一（如图2.5所示）；贝叶斯神经网络是一种概率估计，拟合出来的图像存在一个范围（如图2.6所示）。此外，贝叶斯是算法层面的概念，而非模型层面，任何良定义的模型都有其贝叶斯形式。

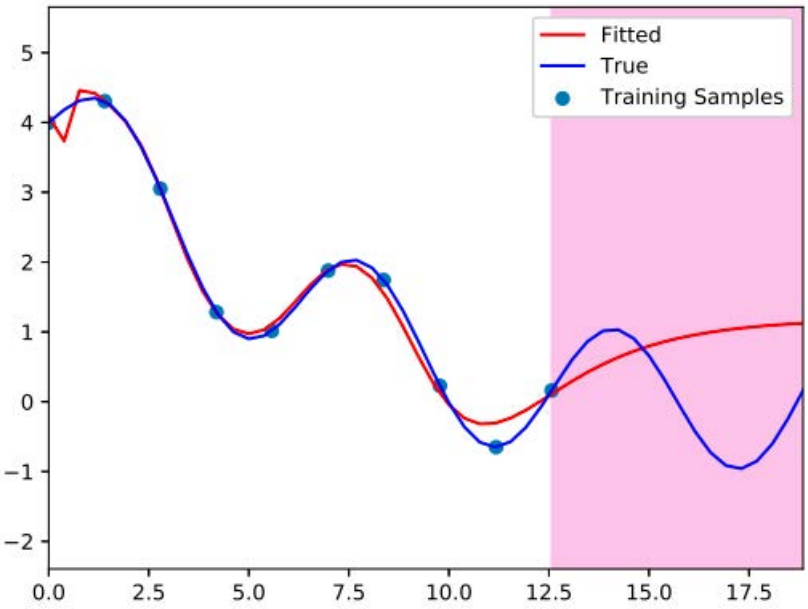


图 2.5　普通神经网络的回归输出

模型不确定性估计的关键在于对模型权重分布的建模，最早可以追溯到 1987 年，Denker 等人[61] 首先提出为权重参数设置分布的想法，对于有限数量的权重，MacKay[62] 和 Neal[63] 先后通过将分布放置在权重上来获得不确定性。从某种意义上讲，不确定性估计可以用高斯过程的估计来模拟，以深度学习为例，可以将有限模型视为高斯过程的近似，利用贝叶斯神经网络获得不确定性。其原理就是求参数关于训练数据的后验分布，而后验分布的计算涉及真实的数据分布，理论上很难获得。BNN 为了解决后验不好计算的问题，一般使用变分推理来近似，找出和真实后验分布 KL 散度最小的简单分布。21 世纪以来，这些想法以技术变体的

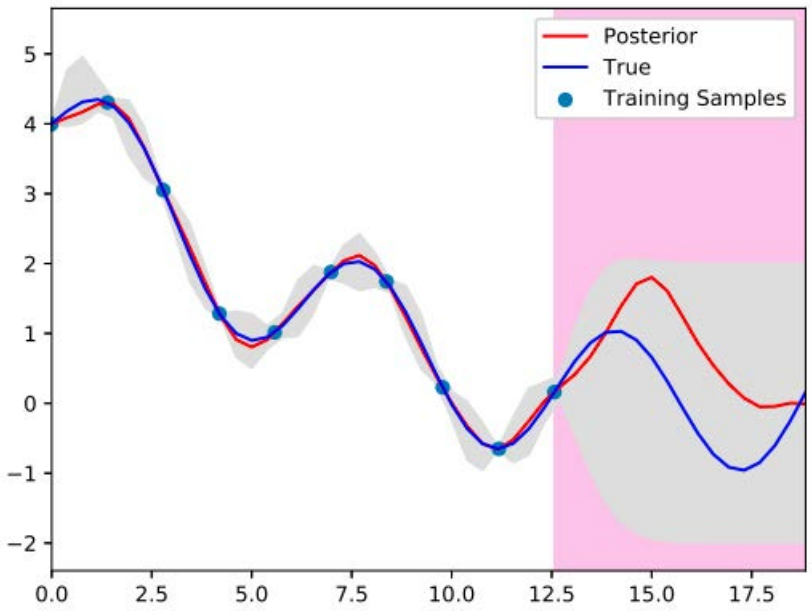


图 2.6　使用高斯过程的回归输出

形式赋予不同的名字重新被 Graves 等人应用，但是以高斯分布近似贝叶斯神经网络中后验分布的变分方法会使参数量翻倍，消耗大量的计算资源，没有带来足够高的性能提升，因此没有得到广泛的应用。

Gal 针对模型不确定性展开了非常丰富的研究，首先证明神经网络中的 dropout 可以解释为著名的贝叶斯模型——高斯过程的近似[30]，并且在 2016 年提出并证明了 dropout 网络可以映射为贝叶斯神经网络中伯努利变分推断的近似[64]，进而可以通过 dropout 实现模型不确定性估计，该想法的核心是用伯努利近似变分分布，通过在每个权重层之前训练带有 dropout 的模型来完成贝叶斯推断，与标准神经网络训练不同的是，在测试阶段也要开启 dropout 来模拟从参数分布中多次采样的操作，这种方法是在等效地执行近似变分推断，伯努利变量的使用不要求引入额外的参数，与标准的神经网络模型相比不会带来大的计算代价。由此，贝叶斯神经网络的实现被简化为训练中的每个权重层之后执行 dropout 操作。

2.3.2　研究成果

最初的 dropout 是用于输入层或全连接层，其目的是为了防止由于数据量或者模型过大导致的过拟合问题。但是 dropout 一般不作用于卷积层和循环神经网络的循环部分，Gal 分别针对卷积神经网络和循环神经网络做了相应的基于贝叶斯的模型研究。

卷积神经网络需要大量的数据来实现正则化，很容易在小数据集上出现过拟

合的情况，dropout 一般不使用在卷积层之后，这是因为测试错误会受到显著影响，放大卷积神经网络的负面结果：卷积会导致 dropout 近似失效。使少量数据建模成为艰巨的任务。为了解决卷积神经网络在小数据上极易过拟合的问题，Gal 提出了基于贝叶斯的高效卷积神经网络[30]，通过将概率分布放在卷积内核上，用伯努利变分分布近似模型不可解的后验分布，缓解了卷积神经网络不适合用 dropout 的问题，并且在 CIFAR-10 分类数据集上的准确率指标上得到了一定的提高。

对于循环神经网络而言，如果在循环部分使用 dropout 的话，经验结果使人们相信噪声会随着循环步的进行被放大而导致信息逐渐丢失。因此，当时存在的研究大多是仅仅将 dropout 技术应用在循环神经网络的输入和输出层。Gal 的实验显示，这样的操作仍然会导致过拟合，随后他们在 RNN 中执行近似变分推断概率贝叶斯模型（他们称其为变分循环神经网络）[65]，新提出的 dropout 变体，与当时的技术中仅对输入层和输出层每个时间步执行不同的 dropout 掩码、不在循环层使用 dropout 掩码的做法不同，分别对输入层、输出层、循环层在每个时间步执行相同的 dropout 掩码，并且在语言建模和情感分析任务上进行了有效性的验证。

在已有的关于深度学习方法不确定性的研究中，主要体现出以下两个优势：

可解释性，偶然不确定性可以反映任务本身的一些需要关注的难点，认知不确定性可以捕捉脱离训练数据的样本点。例如在逐像素深度回归任务中，论文[66] 的实验结果表明，不确定性能够捕捉该任务的许多方面，这些方面本质上是困难的，在对图像区域不确定性结果可视化之后，发现偶然不确定性在图像有较大深度、反射面和遮挡边界的区域更大。

小数据的鲁棒性，引入模型不确定性之后，可以更有效地防止过拟合，论文[30] 在 MNIST 数据集上进行了实验，通过对数据集随机抽取 1/4，1/32，结果表明在 1/4 数据集上引入不确定性的模型有更好的鲁棒性，但是当数据集足够小时，模型均表现出了过拟合的现象。

2.4 本章小结

本章主要介绍课题的相关工作。主要包括以下几个部分：

第一部分，简要介绍了主流的词嵌入方法，主要是分布式词嵌入，包括 Word2vec，Glove 和 BERT 的原理和优劣。

第二部分，介绍细粒度情感分析任务涉及的子任务和相应的方法。然后着重介绍了细粒度情感分类任务常用的三种基础模型，循环神经网络、图神经网络和注意力机制，最后根据对多个模型在 SemEval14 数据集上的预测结果进行分析归纳了现有方法的局限性。

第三部分，从模型不确定性的发展历史出发，描述了基于贝叶斯神经网络建模不确定性的技术沿革，最后简要介绍了现有的研究所反映出来的建模不确定性

带来的优势。

第三章　模型不确定性感知的损失函数设计

标准的神经网络模型在多分类任务上极易受数据集规模大小以及不同类别样本是否均衡的影响，出现相当比例极性预测相反的情况。针对上述问题，本章通过建模模型不确定性 —— 集成权重分布上的多组神经网络模型，来弥补数据规模较小带来的不足，在保证多分类准确率的前提下，探索如何利用模型不确定性降低多分类问题中极性预测相反的样本占总预测样本的比例，并且为预测结果提供可靠的不确定性解释。

不确定性估计的算法通常是基于贝叶斯神经网络的，在基于贝叶斯的神经网络模型的相关研究中，大多只是简单地应用贝叶斯神经网络得到基于不确定性的模型，把不确定性作为观测指标，以此表示模型对预测结果的置信度，忽略了不确定性大小对模型预测结果的反馈作用。本章从不确定性大小和预测正误的关系出发，利用模型不确定性对小规模数据有更好鲁棒性的优势，使不确定性反馈到模型训练过程中，作用于损失函数，以不确定性大小和相应预测结果的正误同时指导模型训练，提出不确定性感知的损失函数 UAL(Uncertainty-Aware Loss)，在保证多分类准确率的基础上，降低极性预测相反的占比。

接下来，首先介绍基于贝叶斯神经网络的不确定性度量方法（即求解参数后验分布的近似解法），获得模型不确定性之后，介绍本文提出的模型不确定性感知的损失函数 UAL 的设计思路，将不确定性反馈到模型训练中，最后在数据集上实验验证方法的有效性。

3.1　基于贝叶斯神经网络的不确定性度量

3.1.1　研究方法

贝叶斯神经网络 BNN(Bayesian Neural Network) 将确定性网络的权重参数替换为这些参数的分布，网络中每个参数的权重都不再是确定的数字，而是权重的先验分布[29]。通过为神经网络的权重引入不确定性进行正则化，相当于集成多组权重分布上的神经网络进行预测。对参数分布进行建模而非求解固定的参数值可以从分布中进行多次采样，并观察多次采样对模型预测结果的影响，如果每次采样都给出一致的预测，那么我们认为模型对此预测结果的置信度很高。贝叶斯神经网络包括两个过程，学习过程和推断过程，学习过程是给定数据和模型，学习参数的分布（公式3.1）；推断过程则是在此基础上对数据进行预测（公式3.2）。因此，问题求解的关键在于推出给定数据和模型下，参数的后验分布，m 代表模型，

D 代表给定的数据，θ 表示参数。

$$p(\theta \mid \boldsymbol{D}, m)=\frac{p(m \mid \boldsymbol{D}, \theta) p(\theta)}{\int p(m \mid \boldsymbol{D}, \theta) p(\theta) d \theta} \tag{3.1}$$

$$p(x \mid \boldsymbol{D}, m)=\int p(x \mid \theta, \boldsymbol{D}, m) p(\theta \mid \boldsymbol{D}, m) d \theta \tag{3.2}$$

贝叶斯推断通常可以分为以下四步（假设模型一定）:

a. 假设先验 $p(\theta)$，建模 $p(y \mid x, \theta)$

b. 计算后验 $p(\theta \mid D)=\frac{p(D|\theta) p(\theta)}{p(D)}$

c. 计算预测值的分布 $p(y \mid x, D)=\int p(y \mid x, \theta) p(\theta \mid D) d \theta$

d. 计算预测值的期望 $E[y \mid x, D]=\int y p(y \mid x, D) d y$

但参数的后验分布需要首先得到数据的先验分布，其解析解的求解非常困难，变分推断和蒙特卡洛采样是求解贝叶斯神经网络后验分布的近似解法，下面将分别展开介绍。

3.1.1.1 变分推断

变分推断是通过选择一种已知分布（比如高斯分布），并不断修改其参数，直到接近要计算的后验分布。最佳参数的求解可以看作一个优化过程，正在优化的分布称为变分分布，数学公式表明，如何为变分分布选择最佳参数等价于最大化下限（示例如下）。

变分贝叶斯的目标就是找到一个和该后验分布 KL 散度最小的简单分布 $q(\theta)$，这个分布是参数化的，通过学习参数 θ 的分布，让简单分布 $q(\theta)$ 和参数后验分布 $p(\theta \mid \mathcal{D})$ 尽可能地接近，将其作为后验分布的近似解。即变分推断的优化目标是最小化简单分布与参数后验分布的 KL 散度（公式3.3）:

$$\mathbb{KL}(q(\theta) \| p(\theta \mid \mathcal{D}))=\mathbb{E}_q\left(\log \left(\frac{q(\theta)}{p(\theta \mid \mathcal{D})}\right)\right) \tag{3.3}$$

经过贝叶斯公式的转换，也可表示如下（公式3.4）.

$$\mathbb{KL}(q(\theta) \| p(\theta \mid \mathcal{D}))=-\left(\mathbb{E}_q(\log p(\mathcal{D}, \theta))-\mathbb{E}_q(\log q(\theta))\right)+\log p(\mathcal{D}) \tag{3.4}$$

其中，$\mathbb{E}_q(\log p(\mathcal{D}, \theta))-\mathbb{E}_q(\log q(\theta))$ 称为 Evidence Lower BOund(ELBO)，同时由于 KL 散度不小于 0，所以 ELBO 还提供了 $\log p(\mathcal{D})$ 的下界，最终的优化目标等价于最大化 ELBO。

3.1.1.2 蒙特卡洛采样

蒙特卡洛方法是一类通过随机采样来求解问题的算法的统称，要求解的问题是某随机事件的概率或某随机变量的期望。蒙特卡洛采样法是指从一个分布采样

很多样本，然后用这些样本来代替这个分布进行相关计算。当积分的解析解不好计算时，采用蒙特卡洛的方法，从 $p(x)$ 中采样很多样本，将样本带入被积函数，把得到的结果求均值作为最后的解。根据大数定律，采样的样本越多，得到的解也就越接近解析解。

将前述贝叶斯推断的第三步和第四步结合起来，可以得到 $E[y \mid x, D] = \iint yp(y \mid x, \theta)p(\theta \mid D)d\theta dy$，调整积分顺序，可以进一步化简（公式3.5），显然，$\int yp(y \mid x, \theta)dy$ 就是 $p(y \mid x, \theta)$ 的均值。

$$E[y \mid x, D] = \int \left(\int yp(y \mid x, \theta)dy \right) p(\theta \mid D)d\theta \tag{3.5}$$

在多数情况下，在对似然率建模的时候，会间接通过对均值建模实现。在回归问题中，神经网络的输出作为高斯分布的均值建模似然率；分类问题中，神经网络的输出作为伯努利分布中不为 0 的概率，也是伯努利分布的均值。最终，蒙特卡洛的核心在于对不同的分布进行采样，复杂分布的采样可以采用拒绝采样、重要性采样、马尔科夫链蒙特卡洛方法采样，一般的研究中对伯努利分布、高斯分布采样即可。

3.1.1.3 MC Dropout

MC Dropout 可以理解为对 dropout 正则化技术的贝叶斯解释。根据上一节的介绍，变分推断是一种贝叶斯方法，它使用变分分布来估计后验分布；dropout 是神经网络的一种正则化形式，在训练过程中，神经元会随机打开或关闭，以防止网络依赖任何特定的神经元。而 MC Dropout 的核心思想在于：dropout 可以用来模拟变分推断，其中变分分布来自伯努利分布（对应 dropout 的激活或失活），以蒙特卡洛的形式对 dropout 进行采样。在具体的实现中，通过 mc dropout 将传统网络变成贝叶斯网络，只需要在测试阶段打开 dropout，保证预测时对模型参数进行多次采样即可，等效于从伯努利分布中进行抽样，并提供了模型确定性的度量，操作示意如图3.1所示。

3.1.2 形式化表述

如前所述，深度学习中的不确定性包括认知不确定性（模型不确定性）和偶然不确定性，这两种不确定性分别建模参数上的概率分布、模型输出上的概率分布。认知不确定性通过对模型的参数施加概率分布达到衡量模型参数随不同数据的变化情况的效果，进而判断所给数据与训练数据分布的远近。偶然不确定性则是通过对模型输出施加概率分布来建模。例如在回归任务中，模型输出可以被高斯随机噪声干扰，在这种情况下，计算噪声和不同输入下的输出之间的方差。

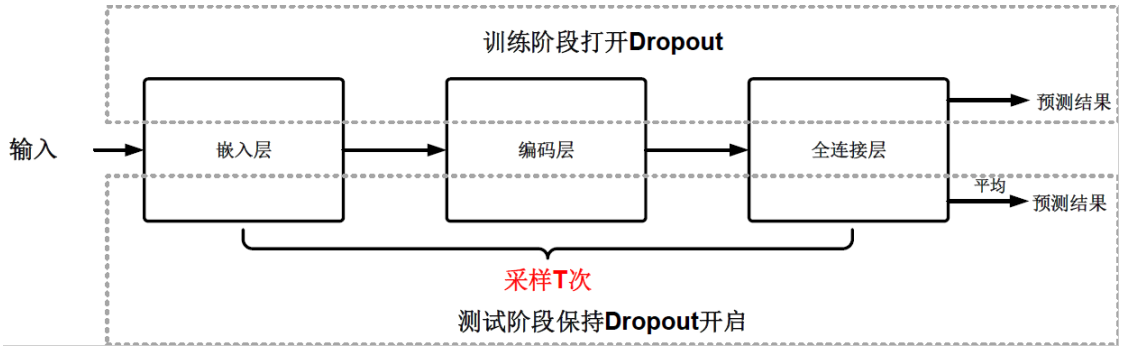


图 3.1　基于 MC Dropout 的不确定性模型操作示意图

记贝叶斯神经网络采样的随机权重为 $\mathrm{f}^{\mathrm{W}}(\mathrm{x})$，模型的似然估计为 $p\left(\mathbf{y} \mid \mathbf{f}^{\mathbf{W}}(\mathbf{x})\right)=\mathcal{N}\left(\mathbf{f}^{\mathbf{W}}(\mathbf{x}), \sigma^{2}\right)$，给定数据集 $D(X, Y)$ 和观测噪音标量 σ，贝叶斯推断用来计算 $p(\mathbf{W} \mid \mathbf{X}, \mathbf{Y})$。

在回归任务中，定义似然为高斯分布，其中均值是模型输出确定，方差是观测噪音 σ。

$$p\left(\mathbf{y} \mid \mathbf{f}^{\mathbf{W}}(\mathbf{x})\right)=\mathcal{N}\left(\mathbf{f}^{\mathbf{W}}(\mathbf{x}), \sigma^{2}\right) \tag{3.6}$$

分类任务中，利用 softmax 对模型输出进行归一化压缩，从相应的概率向量中采样。

$$p\left(\mathbf{y} \mid \mathbf{f}^{\mathbf{W}}(\mathbf{x})\right)=\operatorname{Softmax}\left(\mathbf{f}^{\mathbf{W}}(\mathbf{x})\right) \tag{3.7}$$

论文[64] 中给出在基于不确定性的模型中，优化目标可以表示为 $\mathcal{L}(\theta, p)$（公式3.8），其中 N 代表样本总数，p 是 dropout 概率，θ 是要优化的简单分布的参数集合。具体地，在回归任务中，负的对数似然可以进一步简化为高斯似然（公式3.9），σ 作为观测噪声参数，用来捕捉输出中的噪声。分类任务中，似然估计可以由蒙特卡洛积分近似如下（公式3.10）。

$$\mathcal{L}(\theta, p)=-\frac{1}{N} \sum_{i=1}^{N} \log p\left(\mathbf{y}_{i} \mid \mathbf{f}^{\widehat{\mathbf{W}}_{i}}\left(\mathbf{x}_{i}\right)\right)+\frac{1-p}{2 N}\|\theta\|^{2} \tag{3.8}$$

$$-\log p\left(\mathbf{y}_{i} \mid \mathbf{f}^{\widehat{\mathbf{W}}_{i}}\left(\mathbf{x}_{i}\right)\right) \propto \frac{1}{2 \sigma^{2}}\left\|\mathbf{y}_{i}-\mathbf{f}^{\widehat{\mathbf{W}}_{i}}\left(\mathbf{x}_{i}\right)\right\|^{2}+\frac{1}{2} \log \sigma^{2} \tag{3.9}$$

$$p(y=c \mid \mathbf{x}, \mathbf{X}, \mathbf{Y}) \approx \frac{1}{T} \sum_{t=1}^{T} \operatorname{Softmax}\left(\mathbf{f}^{\widehat{\mathbf{W}}_{t}}(\mathbf{x})\right) \tag{3.10}$$

mc dropout 学习和推断阶段的算法伪代码如下（算法3.1，算法3.2），其中 η_1 表示认知不确定性，η_2 表示偶然不确定性。

算法 3.1 mc dropout 学习阶段算法

已知: 数据 x^*，编码器 $g(\cdot)$，预测网络 $h(\cdot)$，dropout 概率 p，采样次数 B
求: 预测值 $\hat{y}^*_{mc}$，不确定性 η_1

for $b = 1$ to B **do**
  $e^*_{(b)} \leftarrow$ VariationalDropout $\left(g\left(x^*\right), p\right)$
  $z^*_{(b)} \leftarrow$ Concatenate $\left(e^*_{(b)},\ \text{extFeatures}\right)$
  $\hat{y}^*_{(b)} \leftarrow$ Dropout $\left(h\left(z^*_{(b)}\right), p\right)$
end for
// 预测
$\hat{y}^*_{mc} \leftarrow \frac{1}{B}\sum_{b=1}^{B} \hat{y}^*_{(b)}$
// 模型不确定性
$\eta_1^2 \leftarrow \frac{1}{B}\sum_{b=1}^{B}\left(\hat{y}^*_{(b)} - \hat{y}^*_{mc}\right)^2$
return $\hat{y}^*_{mc}, \eta_1$

算法 3.2 mc dropout 推断阶段算法

已知: 数据 x^*，编码器 $g(\cdot)$，预测网络 $h(\cdot)$，dropout 概率 p，采样次数 B
求: 预测值 $\hat{y}^*$，不确定性 η

// 预测，模型不确定性
$\hat{y}^*, \eta_1 \leftarrow$ MCdropout $\left(x^*, g, h, p, B\right)$
// 内在噪音
for x'_v in validation set $\{x'_1, \ldots, x'_V\}$ **do**
  $\hat{y'}_v \leftarrow h\left(g\left(x'_v\right)\right)$
end for
$\eta_2^2 \leftarrow \frac{1}{V}\sum_{v=1}^{V}\left(\hat{y'}_v - y'_v\right)^2$
// 模型整体预测不确定性
$\eta \leftarrow \sqrt{\eta_1^2 + \eta_2^2}$
return $\hat{y}^*, \eta$

为了捕获模型的偶然不确定性 $\mathcal{L}_{\mathrm{NN}}(\theta)$（公式3.11），需要调整观测噪音参数 σ，在非贝叶斯神经网络中，观测噪音参数是作为权重衰减的一部分保持不变的。

$$\mathcal{L}_{\mathrm{NN}}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2\sigma\left(\mathrm{x}_i\right)^2}\left\|\mathbf{y}_i - \mathbf{f}\left(\mathrm{x}_i\right)\right\|^2 + \frac{1}{2}\log\sigma\left(\mathrm{x}_i\right)^2 \tag{3.11}$$

3.2 不确定性感知的损失函数设计

通过将标准神经网络与贝叶斯神经网络结合起来，利用 MC dropout 的近似方法，获得预测结果与相应的模型认知不确定性，针对小规模数据集上的三分类问题，提出不确定性感知的损失函数 UAL，将模型的认知不确定性反馈到预测结果

的正误中，在保证模型准确率的情况下，为模型预测结果提供更可靠不确定性解释的同时降低极性预测相反的样本占比。

3.2.1 研究思路

在 UAL 中，训练目标不单单是将样本类别分对，而且要以较小不确定性分对，我们的出发点是在原来损失函数的基础上，对分对的样本施加奖励，降低其损失函数（以较小不确定性分对的其奖励系数更大）；分错的样本施加惩罚，增大其损失函数（以较小不确定性分错的样本其惩罚系数更大）。最终的目标是使模型尽量以小不确定性分对，或者以大不确定性分错，即让模型尽量对自己的预测结果有较大置信度（如表3.1所示）：

表 3.1 不确定大小与预测结果的指示关系表

预测结果 / 不确定性	正确	错误
大	减少	增多
小	增多	减少

3.2.2 不确定性表示

方差 σ^2 表示随机变量的变化程度，由随机变量的取值和随机变量的均值差的平方取平均得到（公式3.12）。信息熵 $H(\mathbf{p})$ 表示随机变量的不确定性（公式3.13），可以用来描述随机变量所包含的信息量，信息的信息量与不确定性有着直接的关系，它们往往有一致的变化趋势。方差关注的是随机变量本身的取值，而信息熵更多地是关注随机变量的分布，而非具体取值，当随机变量可取的类别数一定时，其取每种类别的概率分布越平均，熵值也就越大。根据任务的目标不同，分类任务和回归任务分别使用信息熵和方差作为其不确定性的度量标准。

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} \tag{3.12}$$

分类任务中，使用概率向量的信息熵 $H(\mathbf{p})$（公式3.13）表示模型不确定性，其中 $\widehat{\mathbf{W}}_t$ 从 dropout 权重分布 $q_\theta^*(\mathbf{W})$ 中采样 $\widehat{\mathbf{W}}_t \sim q_\theta^*(\mathbf{W})$。

$$H(\mathbf{p}) = -\sum_{c=1}^{C} p_c \log p_c \tag{3.13}$$

回归任务中，使用预测方差 $\mathrm{Var}(\mathbf{y})$（公式3.14）表示模型不确定性：

$$\mathrm{Var}(\mathbf{y}) \approx \sigma^2 + \frac{1}{T}\sum_{t=1}^{T} \mathbf{f}^{\widehat{\mathbf{W}}_t}(\mathbf{x})^T \mathbf{f}^{\widehat{\mathbf{W}}_t}(\mathbf{x}) - E(\mathbf{y})^T E(\mathbf{y}) \tag{3.14}$$

其中，$E(\mathbf{y})$ 是从模型采样 T 次得到的预测均值，$E(\mathbf{y}) \approx \frac{1}{T}\sum_{t=1}^{T} \mathbf{f}^{\widehat{\mathbf{W}}_t}(\mathbf{x})$，$\sigma^2$ 代表的是偶然不确定性，内在噪音，后面的预测方差就是需要求解的认知不确定性，当模型 T 次采样结果都一样时降为 0。

3.2.3 方法设计

根据上一节的分析，我们需要根据模型分对分错的结果对损失函数施加不同的奖惩系数，而这个系数与模型对该样本预测结果的不确定性有一致的变化关系，在分类问题中，用来度量模型认知不确定性的信息熵 $H(p)$ 的取值范围如下（不等式3.15），其中 k 是分类问题中的类别数：

$$0 \leqslant H(p) \leqslant \log k \tag{3.15}$$

对其归一化处理之后，将其作为施加在原损失函数上的奖惩系数，即分对的样本乘以不确定性，分错的样本除以不确定性，表示为 $loss_{new}$（公式3.17），loss_ori 由公式3.8给出，代表无 UAL 指导的损失函数（公式3.16）。

$$loss_{ori} = -\frac{1}{N}\sum_{i=1}^{N} \log p\left(\mathbf{y}_i \mid \mathbf{f}^{\widehat{\mathbf{W}}_i}(\mathbf{x}_i)\right) + \frac{1-p}{2N}\|\theta\|^2 \tag{3.16}$$

$$\mathrm{loss}_{\mathrm{new}} = \frac{loss_{ori}}{\mathrm{uncertainty}^{\,(-1)^{y_{true} \Leftrightarrow y_{pred}}}} \tag{3.17}$$

考虑到不确定性值可能非常小，当其作用于分错的样本时导致损失无穷大，因此，需要设置阈值 t 来加以控制，最终的 loss 设计如下：

$$\mathrm{Loss} = \begin{cases} \mathrm{loss}_{\mathrm{ori}} & H(p) \leqslant t \\ \mathrm{loss}_{\mathrm{new}} & H(p) > t \end{cases} \tag{3.18}$$

原理示意如下，样本经过 MC Dropout 模型得到预测值及相应的不确定性估值，根据预测值与标签值是否一致来指导不确定性估值增大或减小的方向，与标准的损失函数不同，UAL 损失函数使模型有两个优化方向，一个是尽量使正确分类的样本数增多，另一个是使分对的样本伴随小的不确定性，使分错的样本伴随大的不确定性，如图3.2所示：

在标准的交叉熵损失中，损失函数的主要影响因素是标签类别所对应的概率值，减小损失的方向就是不断增大预测结果中标签类别所对应的概率值，不同预

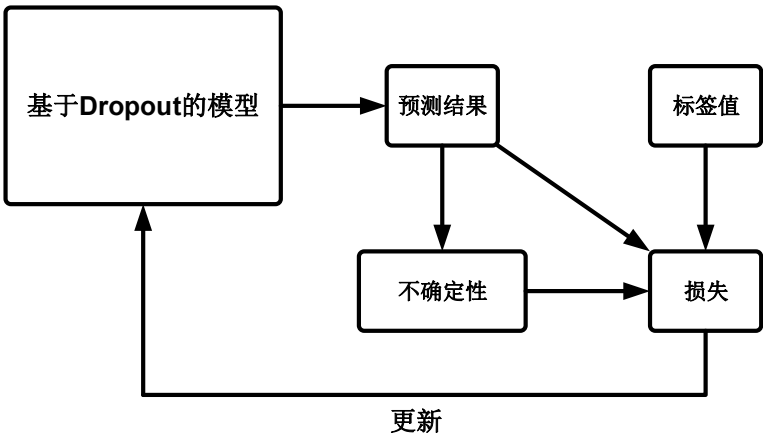


图 3.2　UAL 指导模型训练示意图

测结果导致的损失函数值的大小关系如下（式3.19）。

$$L(\text{预测错误}) > L(\text{预测正确}) \tag{3.19}$$

本章在不确定性指导的损失函数 UAL 下，不同预测结果带来的损失函数的大小关系由如下不等式给出（式3.20）：

$$\begin{aligned} &L(\text{以较小不确定性预测错误}) > L(\text{以较大不确定性预测错误}) > \\ &L(\text{以较大不确定性预测正确}) > L(\text{以较小不确定性预测正确}) \end{aligned} \tag{3.20}$$

3.3　实验及结果分析

本节是实验及结果分析部分，首先介绍了实验使用的数据集，然后对细粒度多分类情感分析任务中的 baseline（包括最新 SOTA）模型进行了介绍，接着围绕不确定性为基线模型设置了三组对照实验，进行实验结果分析与可视化，最后展示了测试样本预测结果的变化作为方法有效性的例证。

情感分析在一定程度上可以视为分类问题，因此其评估标准可以是准确率，精确率，召回率和 F 评分，这些指标实际上已广泛用于各种分类任务的评估中，其中 F 评分是调和平均值，是对准确率和召回率的总体衡量。对于情感分析，方面词与方面类别之间的映射以及情感极性的确定都涉及多分类问题，为了更公正、客观地衡量分类器的总体效果，尽量减小类别不均衡带来的影响，引入了宏观平均。以所有类别的 F1 值（公式3.21）的算术平均值作为宏观平均值，近年来，宏观平均值指标已越来越多地用于情感分析评估中。

本实验中，我们采用准确率 acc、宏观 F1 值、极性预测相反的样本占总测试样本的比例 opp 作为模型的评价指标，其中 opp 是验证本实验提出的方法有效性

的指标，用来衡量不同模型应用到数据集之后，预测结果中极性预测相反的样本占总预测样本的比例的变化情况。

$$F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{3.21}$$

3.3.1 数据集

实验使用的数据集是 SemEval14 中的 Laptop 和 Restaurant 三分类数据集，分别是对电脑和餐厅的评论数据，情感分类的对象是评论语句中出现的方面词。实验中所用到的 baseline 模型都已经根据数据集做了对齐实验（RGAT 模型被提出时虽然实验数据集也选用了 SemEval14，但是对样本进行了删减，具体见[24]，本文以较完整且被大多数论文使用的 SemEval14 数据集作为基准），具体的数据集样本信息统计如表3.2所示。laptop 中，正负中性样本比例约为 2：2：1，restaurant 数据集中，正负中性样本比例约为 3：1：1，即 restaurant 数据集中样本不均衡的现象尤为明显。

表 3.2 SemEval14 数据集样本统计

field / class	restaurant		laptop	
	train	test	train	test
positive	2164	728	994	341
negative	807	196	870	128
neutral	637	196	464	169

3.3.2 baseline 模型

实验中，选取了细粒度多分类情感分析任务中经典的 baseline 模型，以及 2020 年最新的 SOTA 模型（在不引入外部训练数据的情况下），对模型原理进行了介绍，通过将 baseline 模型改为基于贝叶斯神经网络的不确定性模型，应用不确定性感知的损失函数训练，每个模型进行四组对比实验，最终以极性预测相反样本占比作为方法有效性的评估。在数据集 laptop 和 restaurant 上展开实验，各模型的核心思想介绍如下：

TD-LSTM[57]：分别对给定的 target 词上下文（包含目标词）进行建模，然后拼接上下文的隐藏表示，经 softmax 输出判断。

ATAE-LSTM[9]：学习方面词的嵌入表示，利用这个表示和上下文词语之间做注意力计算，得到最后的加权和用作情感分类；在模型输入部分，把方面词嵌入和上下文单词嵌入结合起来。

RAM[11]：多重注意机制的方法合成难句结构中的重要特征，使得较远的信息也能理解（多重关注的结果与 GRU 网络相结合，从 RNN 中继承了不同的行为，即用非线性的方法把注意力的结果结合起来）；在输入和注意力层之间加入了 memory 记忆模块。

BERT-SPC[67]：是将 BERT 应用到情感分析问题上的基础模型，其流程比较简洁，首先利用 Bert 预训练模型对句子和目标词进行编码，然后利用 Bert 输出的表示句子特征的向量进行分类预测。

RGAT[24]：在图注意力网络的基础上，引入关系标签，进一步对句子成分之间的依赖关系建模，融合关系感知的注意力和单词感知的注意力，最后对图编码向量和句子编码向量进行特征融合，得到最后的类别标签。

3.3.3 实验结果

我们基于上述 baseline 基线模型，分别设计了三组基于不确定性的模型变体，与基线模型一起构成 4 组对比实验：其中 ori 表示基线模型本身，unc 表示无 UAL 指导的基于不确定性的基线模型变体，new 表示有 UAL 指导但不设置阈值 t 的基于不确定性的基线模型变体，mix 表示有 UAL 指导且设置阈值 t 的基于不确定性的基线模型变体，对损失函数进行掩码操作（实验中的阈值设置 t 为 0.01），各基线模型的变体模型中，实验参数设置与基线模型保持一致。其中 RGAT 基线模型 ori 的实验结果是利用论文[24] 公开的源代码在本文选择的完整 SemEval14 数据集上运行的结果。

实验的观测指标有准确率，宏观 F1 值，极性预测相反的样本占总测试样本的比例 opp（为了使模型的不同对照组之间具有可比性），单位均为百分比。表格中的加粗字体表示每个评价指标在该模型该数据集上取得的最好效果，opp 一栏中的斜体分别对应基线模型和 UAL 指导的基于不确定性的模型变体相应的极性预测相反占总预测样本的比例。

观察对 baseline 模型进行四组对照实验得到的结果（如表3.3所示），可以得到以下分析：

1、基于贝叶斯神经网络的模型在小规模数据集上的表现往往有一定的提升效果，具体体现在准确率 acc 和宏观 F1 值上，而且对特征提取能力有限的模型效果提升较明显（除 SOTA 模型之外，其余模型在两个数据集上的最大 acc 和 F1 值都在基于贝叶斯神经网络的模型中得到）。

2、从整体上看，本文提出的不确定性感知的优化目标 UAL 与贝叶斯神经网络的结合（对应表格中的 mix 栏），在多分类准确率和宏观 F1 值与基线模型相当的情况下，极性预测相反的比例 opp 的变化趋势在进行实验的所有 baseline 模型

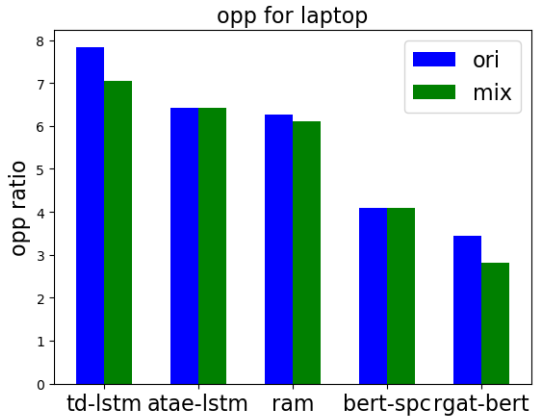
表 3.3　各基线模型及其变体在 SemEval14 数据集上的实验结果

	laptop				restaurant			
baseline	变体	准确率	宏观 F1	opp	变体	准确率	宏观 F1	opp
TD-LSTM	ori	70.85	65.43	*7.84*	ori	79.11	68.51	*6.70*
	unc	70.85	64.52	7.68	unc	79.55	68.68	**5.80**
	new	71.63	65.98	8.77	new	79.20	67.16	6.52
	mix	**72.73**	**67.47**	***7.05***	mix	**79.64**	**69.78**	*5.89*
ATAE-LSTM	ori	**71.47**	66.28	***6.43***	ori	77.41	66.55	*7.86*
	unc	68.81	62.55	7.05	unc	**78.66**	**66.82**	**6.25**
	new	**71.47**	65.69	7.52	new	78.04	65.50	6.34
	mix	71.32	**66.42**	***6.43***	mix	**78.66**	**66.82**	*6.79*
RAM	ori	72.41	67.32	*6.27*	ori	80.71	71.18	*5.27*
	unc	72.57	66.59	8.31	unc	**80.98**	**71.32**	4.91
	new	72.26	**67.88**	7.52	new	79.82	69.16	**4.73**
	mix	**72.73**	67.73	***6.11***	mix	80.54	71.18	*4.91*
BERT-SPC	ori	77.43	74.03	*4.08*	ori	84.82	77.95	*3.84*
	unc	77.90	73.15	4.86	unc	84.73	77.21	2.86
	new	77.74	74.01	4.39	new	84.55	75.92	4.64
	mix	**78.95**	**75.02**	***4.08***	mix	**85.54**	**78.66**	***2.68***
RGAT	ori	**79.17**	**74.55**	*3.45*	ori	**85.89**	**79.83**	*2.95*
	unc	78.55	73.47	3.29	unc	85.45	79.63	2.77
	new	77.77	73.87	4.23	new	85.54	79.73	3.39
	mix	78.06	74.49	***2.82***	mix	85.09	79.33	***2.41***

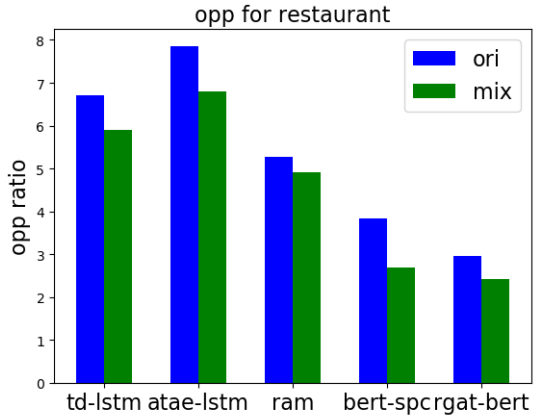
上都有着一致的效果，其 opp 值均小于等于基线模型的 opp 值，而未经 UAL 指导的基于贝叶斯神经网络的变体模型 unc 和未设置阈值的不确定性感知损失函数指导的模型 new 的 opp 值则没有这样的性质。

3、根据对 laptop 和 restaurant 数据集的统计分析，restaurant 数据集类别样本不均衡现象更为明显，比较两大数据集上各模型的准确率和宏观 F1 值，基于贝叶斯神经网络的模型 unc 的优势在 restaurant 数据集上更能得到体现（表现在普遍提高的准确率上），从侧面反映了模型不确定性可以弥补模型在数据规模较小以及类别样本不均衡时的不足。

柱状图3.3直观地展示了将不确定性损失函数 UAL 用于指导模型训练的方法的效果，与基线模型相比，UAL 指导的基于不确定性的变体模型在两个数据集上都取到了最小的 opp 值，即极性预测相反的样本占总预测样本的比例达到了最低。



(a) 基线模型与 mix 变体模型在 laptop 上 opp 对比图



(b) 基线模型与 mix 变体模型在 restaurant 上 opp 对比图

图 3.3　基线模型与 UAL 指导变体模型 mix 的极性预测相反样本占比 opp 对比图.

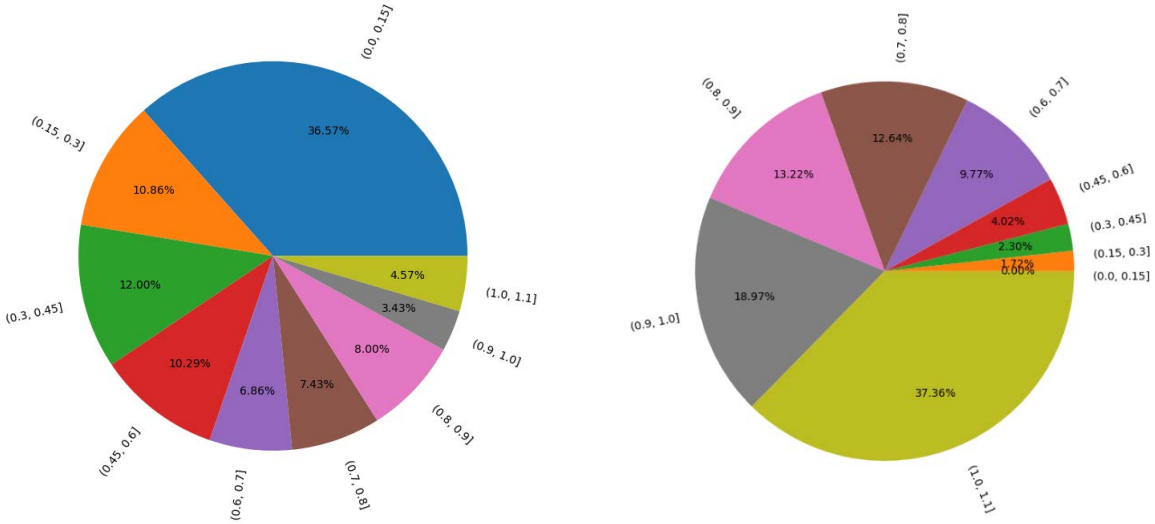
3.3.4　结果分析与可视化

通过对误分类样本数随不确定性变化的分布进行可视化，对非 Bert 模型中分类效果最好的模型 RAM 和最新 SOTA 模型 RGAT 的 unc 实验结果和 mix 实验结果对比分析，以验证 UAL 的有效性。

3.3.4.1　误分类样本在各不确定性区间的分布

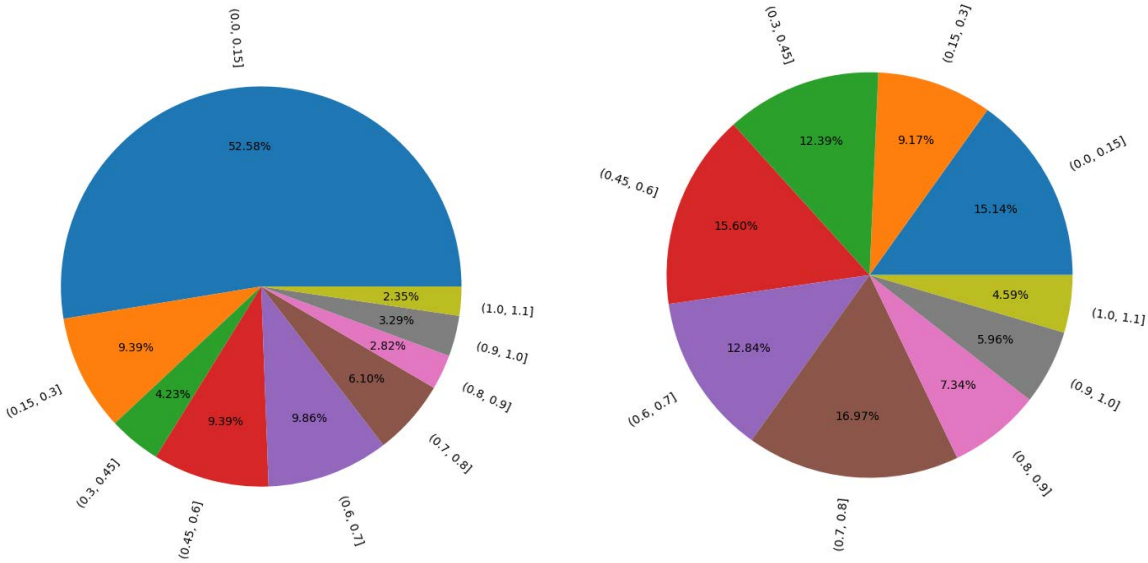
本节对 RAM 模型和 RGAT 模型五分类样本在各不确定性区间的分布进行可视化，选取了无 UAL 指导的不确定性模型 unc 以及在 UAL 指导下的不确定性模型 mix 作为分析对象，以 [0.0,0.15,0.3,0.45,0.6,0.7,0.8,0.9,1.0] 为不确定性区间，分别统计两个模型在各区间上的误分类样本数，观察不确定性感知的损失函数 UAL

带来的模型预测错误样本在不确定性区间的分布变化，验证是否满足让模型对分错的样本伴随较大的不确定性。



(a) ram_lap_unc　　(b) ram_lap_mix

图 3.4　RAM 在 laptop 上误分类样本在不确定性区间的分布图（不同颜色的面积大小代表该不确定性区间的样本在所有误分类样本中所占的比例，下同）

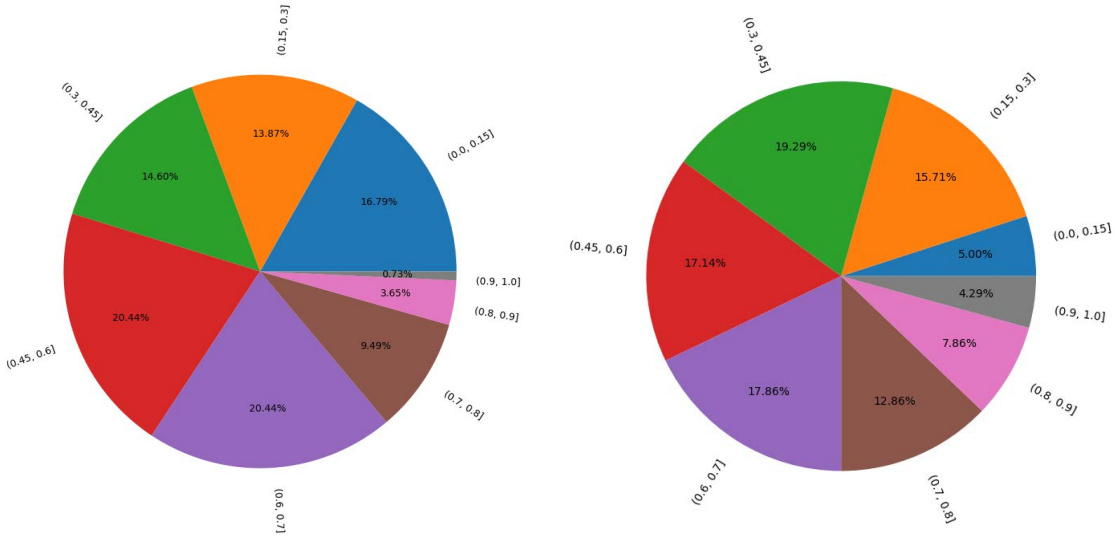


(a) ram_res_unc　　(b) ram_res_mix

图 3.5　RAM 在 restaurant 上误分类样本在不确定性区间的分布图

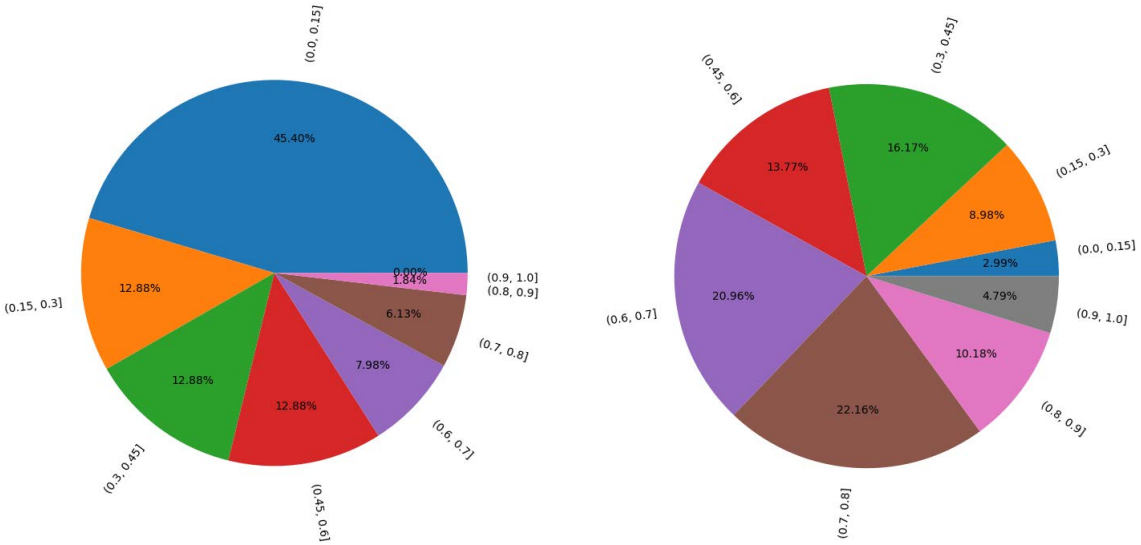
图3.4，图3.5分别显示了在数据集 laptop 和 restaurant 上，未经 UAL 指导的 RAM_unc 和经 UAL 指导的 RAM_mix 的误分类样本中，样本数随不确定性变化

的分布关系，可以直观地观察到，经 UAL 指导的 RAM_mix 误分类样本不确定性整体增大。例如在 laptop 数据集上的实验结果中，以不确定性值 0.45 为界，不确定性不大于 0.45（对应饼状图中，蓝色、橙色、绿色的集合）的误分类样本比例由 59.43% 降到了 4.02%。



(a) rgat_lap_unc　　(b) rgat_lap_mix

图 3.6　RGAT 在 laptop 上误分类样本在不确定性区间的分布图



(a) rgat_res_unc　　(b) rgat_res_mix

图 3.7　RGAT 在 restaurant 上误分类样本在不确定性区间的分布图

SOTA 模型 RGAT 的实验结果如图3.6，图3.7所示，同样地，与未经 UAL 指导

的不确定性模型 RGAT_unc 相比，RGAT_mix 在两个数据集上的误分类样本的不确定性也呈现整体增大的趋势，经 UAL 指导的模型 RGAT_mix 的误分类样本主要集中在不确定性大于 0.45 的区间。

3.3.4.2 误分类样本随不确定性阈值变化的分布

图3.8，图3.9展示了使用 UAL 前后，基于贝叶斯神经网络的 RAM 模型在

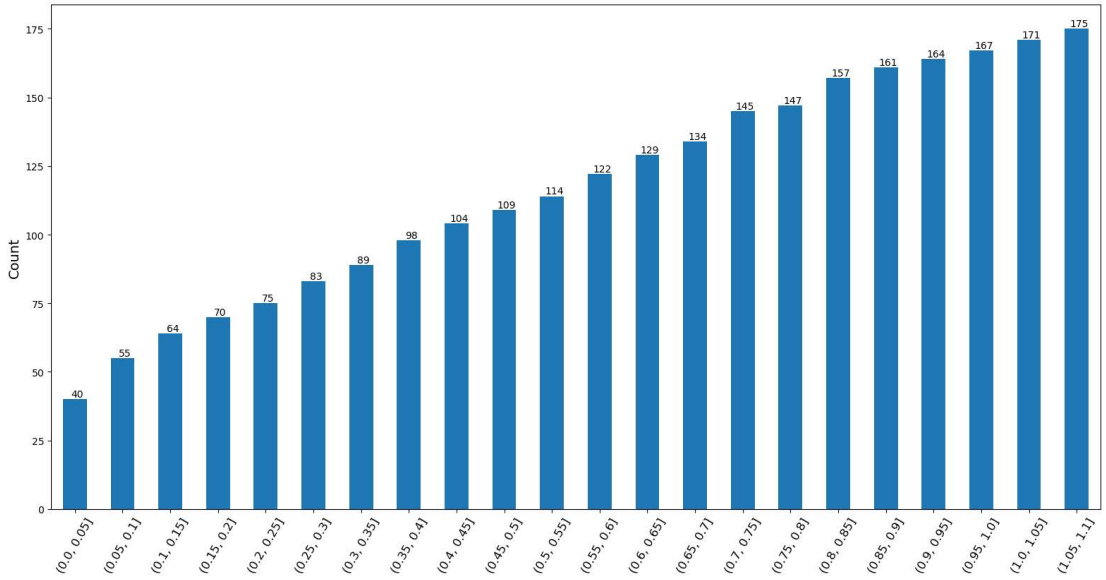


图 3.8　无 UAL 指导的分类模型 RAM_unc 在 laptop 上误分样本累积分布图（横坐标代表不确定性从 0 到 1 间隔为 0.05 的区间，下同）

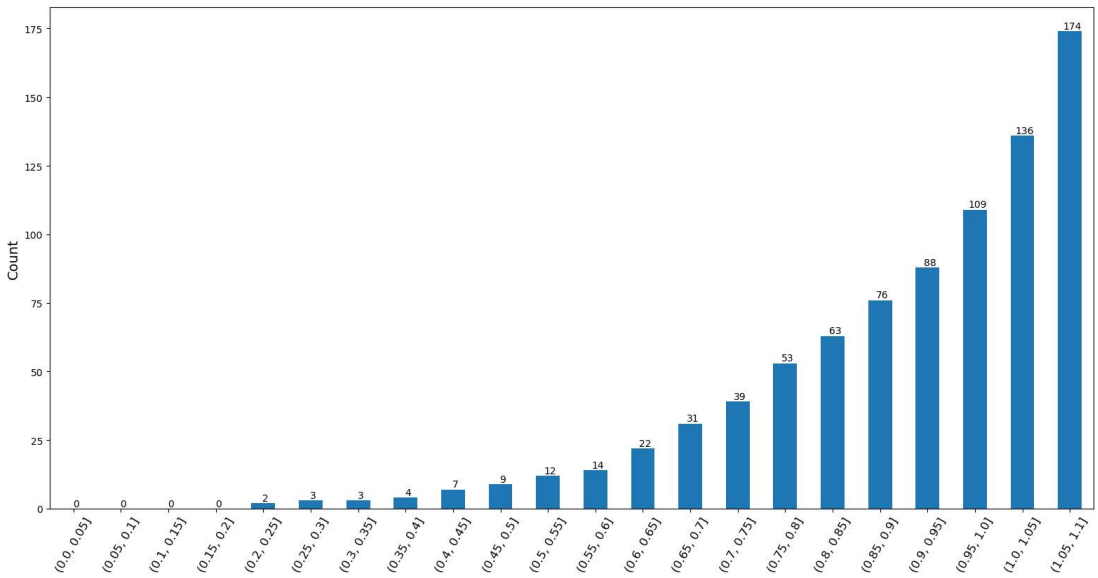


图 3.9　有 UAL 指导的分类模型 RAM_mix 在 laptop 上误分样本累积分布图

laptop 数据集上，误分类样本以各不确定性阈值为上界的累积分布情况；图3.10、

图3.11展示了使用 UAL 前后，基于贝叶斯神经网络的 RGAT 模型在 restaurant 数据集上，误分类样本以各不确定性阈值为界的累积分布情况。很显然，在使用 UAL 之后，不同的模型在不确定性区间上表现出一致的变化趋势，随着不确定性的增加，错误分类的样本数也在显著增加。此外，可以根据使用 UAL 实验前后样本数量的转折点判断需要修正的不确定阈值。

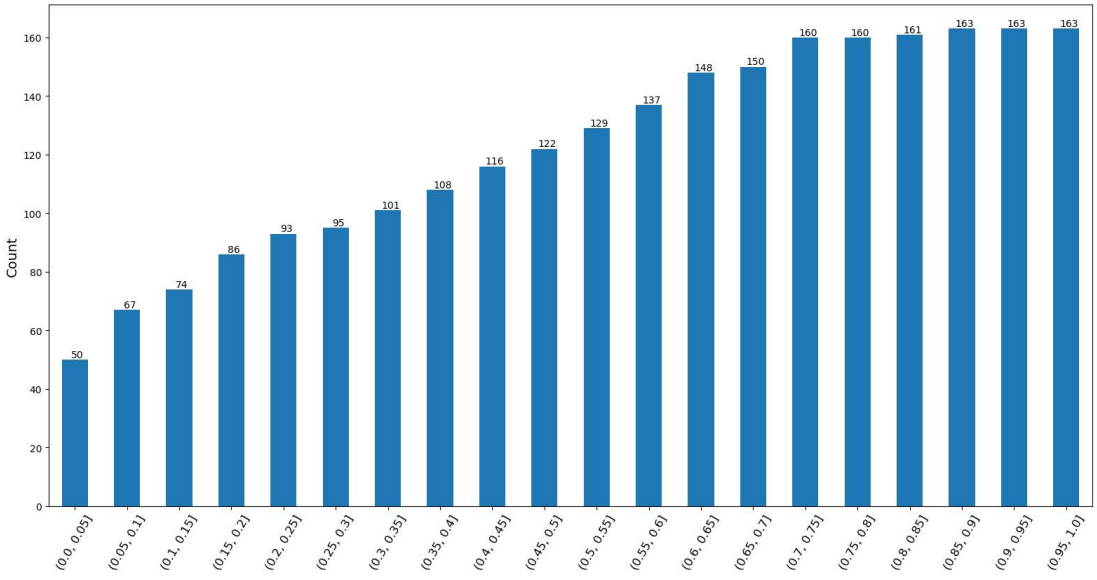


图 3.10　无 UAL 的指导分类模型 RGAT_unc 在 restaurant 上误分类样本累积分布图

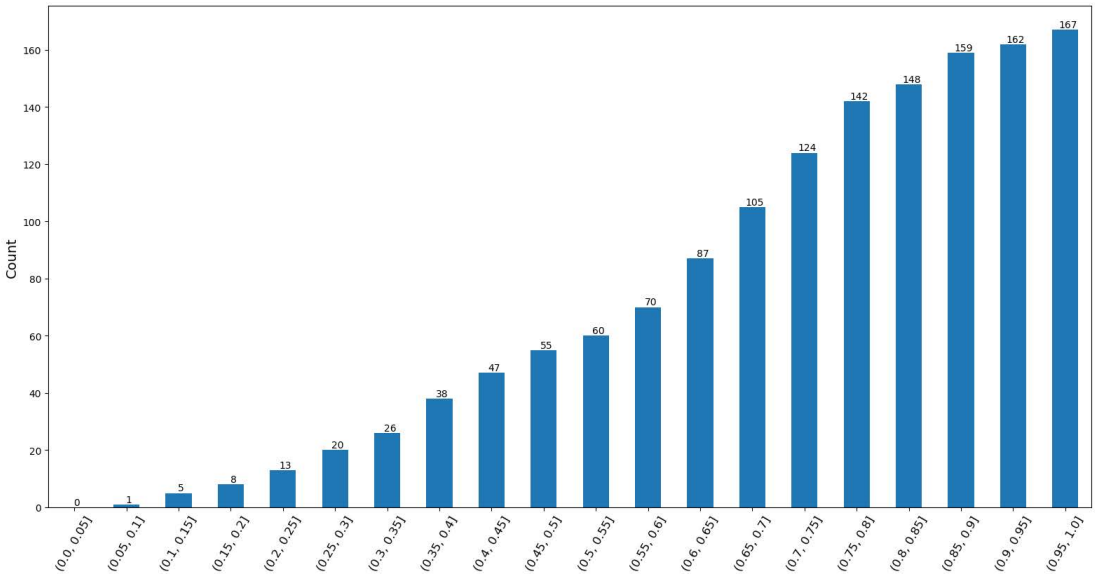


图 3.11　有 UAL 指导的分类模型 RGAT_mix 在 restaurant 上误分类样本累积分布图

图3.12，图3.13分别是基于贝叶斯的 RAM 和 RGAT 模型在 laptop 数据集上未使用 UAL 的情况下，误分类样本数随不确定性阈值变化的分布图，可以看到，

SOTA 模型 RGAT 较 RAM 而言，其误分类样本的主体分布在不确定性较大的区间，以不确定性值 0.3 为阈值，RAM_unc 不确定性大于 0.3 的误分类样本占比 52.57%，RGAT_unc 不确定性大于 0.3 的误分类样本占比 69.34%，即 RGAT 模型的不确定性和分类结果有着更一致的关系，分错的样本大部分分布在不确定性较大的区间，对预测结果有更好的不确定性解释，从侧面反映了 RGAT 模型的特征提取能力更强。

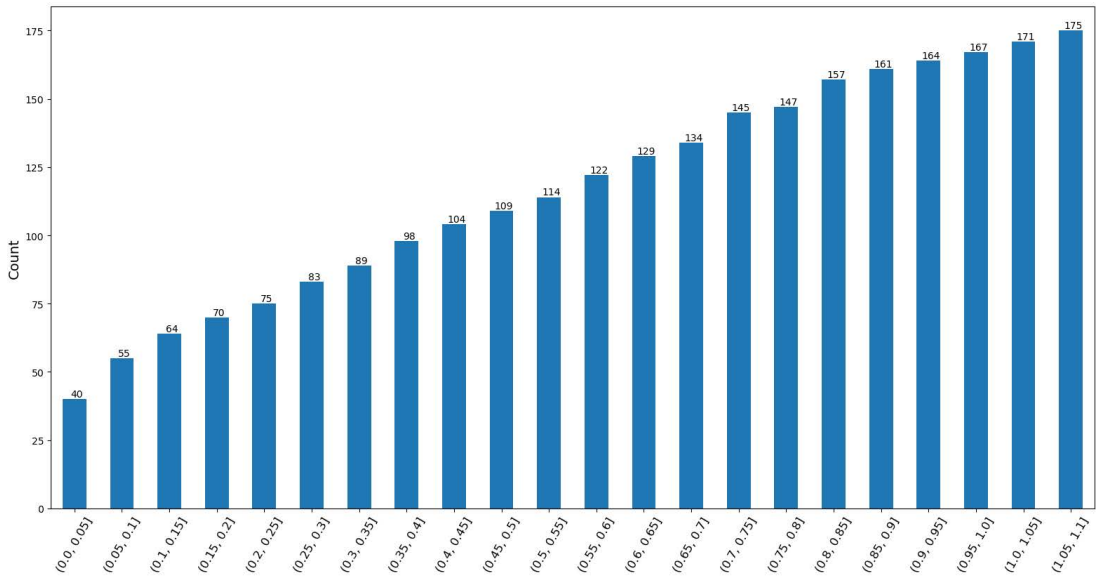


图 3.12　无 UAL 指导的分类模型 RAM_unc 在 laptop 上误分样本累积分布图

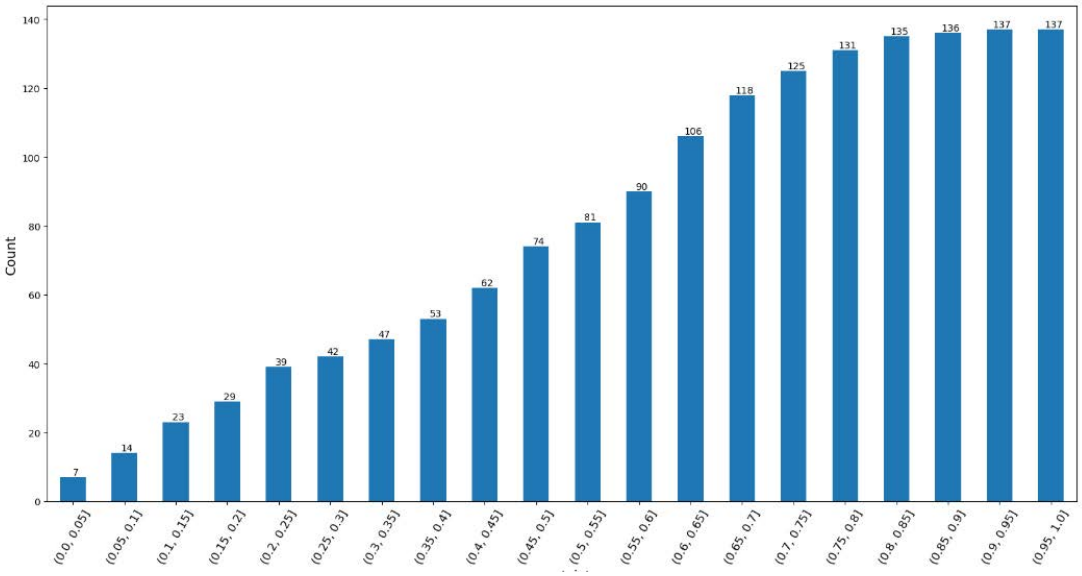


图 3.13　有 UAL 指导的分类模型 RGAT_unc 在 laptop 上误分样本累积分布图

3.3.5 测试样本预测结果变化举例

以 SOTA 模型在 restaurant 数据集上的表现为例进行分析，基线 SOTA 模型 RGAT 的预测样本中，有 33 个极性预测完全相反的样本，而使用 UAL 指导的基于不确定性的变体模型 RGAT_mix 预测之后，有 8 个样本被预测正确（部分示例样本如表3.4所示），5 个样本预测值与标签值之间的误差跨度变为 1，总的极性预测相反的样本降为 27 个。

表 3.4 RGAT 模型在 Restaurant 数据集上极性预测相反的样本修正示例

content	label	RGAT_ori	RGAT_mix
the main draw of this place is the **price**.	**2**	0	**2**
the **decor** is what initially got me in the door.	**2**	0	**2**
Save room for **deserts** - they ’re to die for.	**2**	0	**2**

3.4 本章小结

本章针对标准深度学习方法在小规模、类别样本不均衡数据集上极性预测相反的样本占比较大的问题，引入模型不确定性度量，提出了不确定性感知的损失函数 UAL，通过在多个模型、两个经典数据集上的实验，达到了降低极性预测相反样本占比的效果，最后通过可视化的方法验证了 UAL 的有效性，使模型对分错的样本有较大的不确定性，从而让模型以不确定值大小对预测结果提供一定的解释性。具体地，可分为以下几部分：

第一部分，介绍基于贝叶斯神经网络的不确定性度量方法，包括变分推断、蒙特卡洛采样和 MC Dropout 三种近似解法，详细介绍了基于 MC Dropout 方法的原理和实现框架。

第二部分，在获得模型不确定性之后，提出不确定性感知的损失函数 UAL，指导模型更好地利用不确定性反馈，参与到模型训练中去，使模型尽可能地以较小的不确定性将样本分对，并且以较大的不确定性将样本分错，方便后续对预测结果进行二次干预。

第三部分，将上述方法应用到一些经典模型和最新的 SOTA 模型中，在 laptop 和 restaurant 数据集进行实验，对实验结果进行可视化分析，验证方法的有效性。

第四章　基于不确定性的细粒度情感分类模型

在上一章中，我们基于贝叶斯神经网络估计模型不确定性，在此基础上提出不确定性感知的损失函数 UAL，将模型对预测结果的不确定性反馈到模型参数更新过程，指导模型向着以较小不确定性分对、以较大不确定性分错的方向训练，使得模型更好地建模不确定性和预测结果之间的一致性关系，有效降低了极性预测相反样本占总预测样本的比例，同时为预测结果提供了可靠的不确定性解释；此外，引入 UAL 之后，通过对误分类样本在不确定性区间的分布进行可视化分析，可以更直观地得到适合进行二次干预的不确定性阈值。但是经过对实验结果观察发现，经 UAL 指导的模型并没有在所有指标上达到最好的效果，其准确率和 F1 值均有待提高。

因此，在 UAL 指导模型训练的基础上，本章利用正误分类样本数在不确定性区间的分布差异得到适合进行二次干预的不确定性阈值，并以此为指导提出基于不确定性的细粒度情感分类模型，降低极性预测相反样本所占比例的同时，达到提高分类预测准确率和 F1 值的目的。

4.1　研究内容及整体框架

4.1.1　研究内容

为了更好地将模型不确定性应用于情感极性分类，本文提出基于不确定性的细粒度情感分类模型 —— 不确定性感知的两阶段修正模型 UATR(Uncertainty Aware Two-stage Refinement Model)，首先是 UAL 指导的变分 Dense 层，利用 MC Dropout 方法模拟伯努利分布，基于贝叶斯神经网络估计第一阶段模型预测结果的不确定性，根据该阶段正误分类样本在不同不确定性区间的分布差异得到适合进行二次干预的不确定性阈值，那么有较大概率预测错误的样本就会以高于阈值的不确定性被筛选出来，同时，大部分正确预测的样本以低于阈值的不确定性得以保留，有效避免大量正确预测样本被错误修正的情况。在修正预测的第二阶段，我们使用基于注意力的图神经网络，对第一阶段不确定性较大的样本，额外利用句子中各成分之间的依赖关系，弥补第一阶段简单网络不能深度建模目标词与其上下文之间关系的不足。

4.1.2　方法设计与整体框架

第一阶段将 MC Dropout 方法应用到全连接层获得的样本的预测值和相应的不确定性，在每一个训练轮次，单个批次训练完成后，遵循错误分类样本尽量多

地分布在大于不确定性阈值的区间内、正确分类样本尽量少地分布在大于不确定性阈值区间的原则，确定该轮次该批次最佳的不确定性阈值，接着不确定性大于该阈值的就会送入第二阶段的模型进行再预测。这样让模型在每个训练阶段动态地获得该阶段的不确定性阈值，可以更准确、更科学地指导第二阶段模型进行修正。模型示意图如4.1所示：

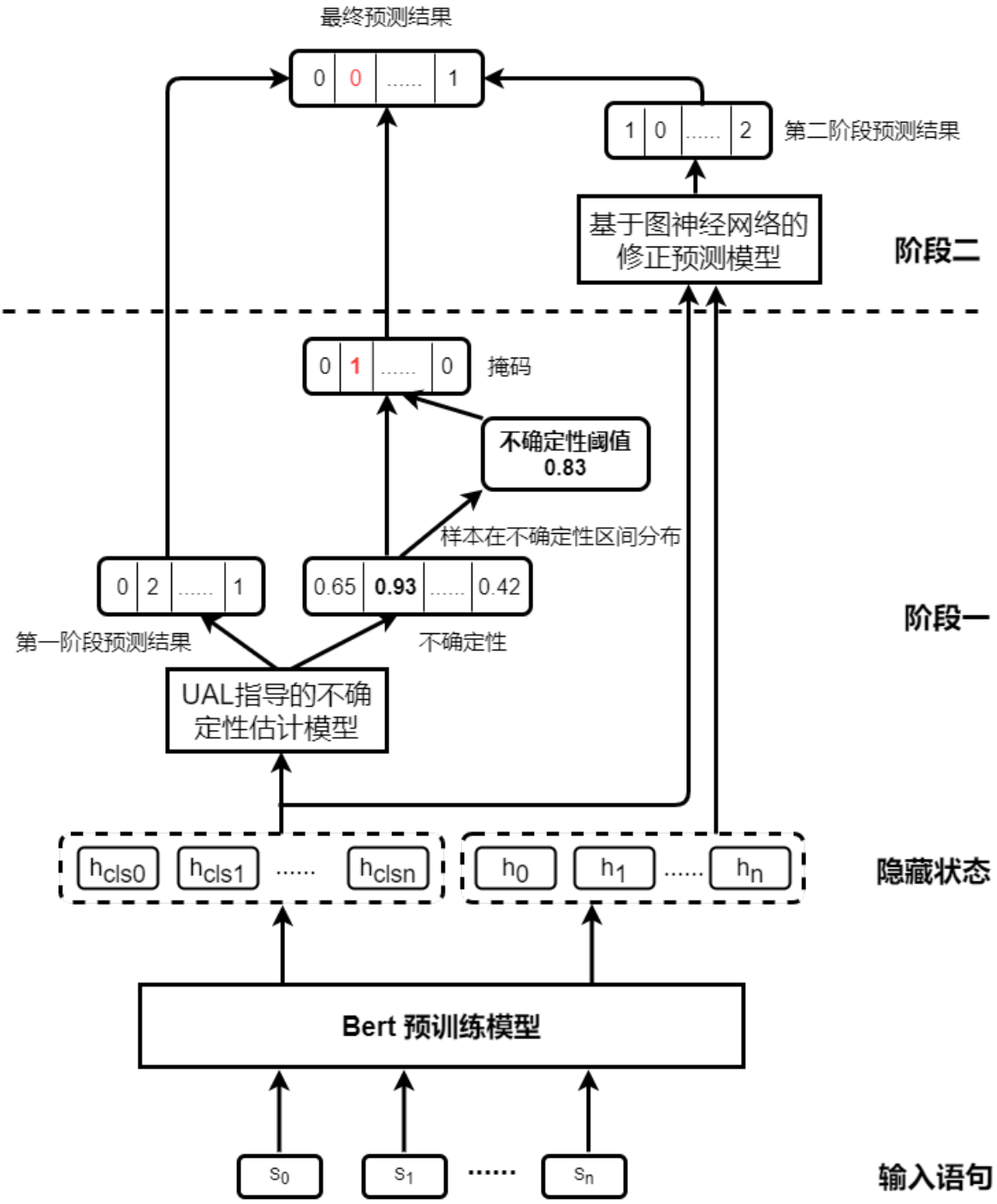


图 4.1　基于不确定性的两阶段修正预测模型 UATR 结构图

4.2 数据处理

本节简要介绍模型需要的数据信息和实验采用的词嵌入方式，并给出数据样例进行展示。

4.2.1 数据准备

实验依然使用 SemEval14 的 Laptop 和 Restaurant 数据集，模型所需的数据内容包括评论语句、方面词，方面词在句子中的位置、词性标注信息以及针对方面词的情感分类标签，具体的数据格式示例如下：

表 4.1 restaurant 数据集采样的数据样例

item	content
tokens	”The”,”portions”,”of”,”the”,”food”,”that”,”came”,”out”,”were”,”mediocre”,”.”
pos_tag	”DT”,”NNS”,”IN”,”DT”,”NN”,”WDT”,”VBD”,”RP”,”VBD”,”JJ”,”.”
head	”2”,”10”,”5”,”5”,”2”,”7”,”5”,”7”,”10”,”0”,”10”
aspect	”portions”,”of”,”the”,”food”
from	1
to	5
polarity	”neutral”

其中，Bai 等人[24] 设计实验，对比了不同依存分析工具在 Penn Treebank 标注集上的表现，Biaffine Parser 和 Bert Biaffine Parser 取得了更好的效果，Biaffine Depedency parser 使用双仿射注意力机制代替双线性或传统的基于 MLP 的注意力机制，使用 Biaffine 依存标签分类器。在双仿射变换之前，将降维 MLP 应用于每个循环输出。因此，本实验中输入到模型的评论语句的句法依存关系沿用 RGAT-BERT 实验中所使用的工具，即根据 biaffine dependency parser 得到句法依存关系。下图给出了例句的句法依存关系分析图：

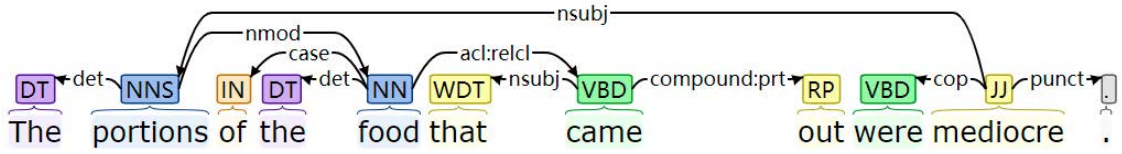


图 4.2 依存关系树示例

4.2.2 词的嵌入表示

细粒度情感分析任务上基于 Bert 的模型效果总是比没有引入 Bert 的模型效果更好，而且最近的 SOTA 模型都是基于 Bert 取得，Bert 强大的特征提取能力使其包含隐含的语法信息，对下游任务大有裨益。因此，本文使用 Bert 预训练模型作为嵌入层，表征句子和其中的目标词。Bert 采用了多层 Transformer 架构，通过多种 attention 机制将任意位置的两个单词的距离转化为 1，有效地解决了 NLP 任务中的长期依赖问题，而且可以很好地并行化。实验中，Bert 的输入采用”[CLS]” + sentence + ”[SEP]” + target + ”[SEP]” 的形式，经 Bert 模型输出后，既可以得到”[CLS]” 位置对应的表征整个句子特征的 bert_pool_output 向量，也可以得到句子中每个单词对应的融合上下文丰富语义信息的向量表示 bert_out。

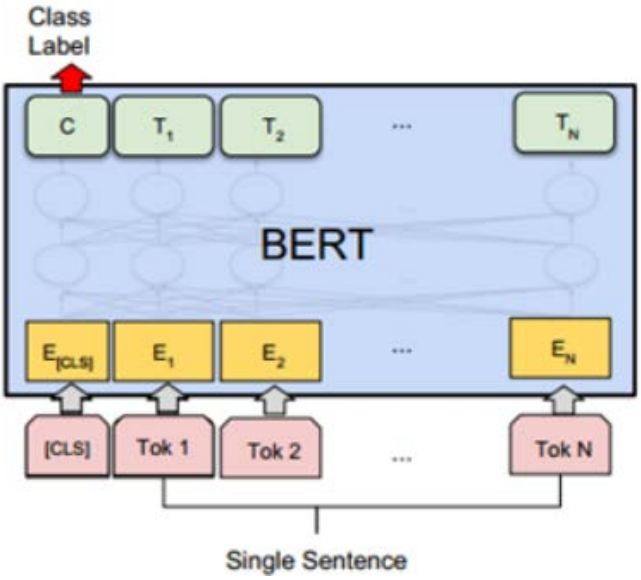


图 4.3 bert 分类任务应用结构示意图

4.3 不确定性感知的两阶段修正预测模型 UATR

本节将对不确定性感知的两阶段修正模型 UATR 展开介绍，首先是第一阶段由 UAL 指导的不确定性估计模型，运用 MC Dropout 方法实现的变分全连接层，完成第一阶段的标签预测和对应样本模型不确定性的估计，本阶段的目标是为第二阶段的修正预测确定合适的筛选阈值，使第一阶段分错的样本有更大的概率被送去二次修正，而第一阶段分对的样本尽量避免进行二次修正。权衡分对样本和分错样本在不确定性区间的分布差异，选取合适的阈值。第二阶段的修正模型选用基于图神经网络的模型，相比第一阶段的模型，更侧重对句子各成分之间的依赖关系的建模。

4.3.1 UAL 指导的不确定性估计模型

不确定性感知的两阶段修正模型 UATR 的第一阶段是由 UAL 指导的不确定性估计模型，由基于 MC Ddropout 的变分全连接层实现，变分全连接层的目的是将全连接层转变为基于贝叶斯神经网络的全连接层，给定数据集 D，通过找到与参数后验分布 $p(\mathbf{W} \mid \mathcal{D})$ KL 散度最小的简单分布 $q_\theta^*(\mathbf{W})$，在参数分布上多次采样得到多个预测值取平均作为最终模型的输出，同时对最终的预测值计算信息熵（公式3.13），以此来建模模型输出的不确定性并确定有较大不确定性的样本。

变分全连接层的实现中，采用 MC Dropout 的方法来模拟变分推断，变分分布来自伯努利分布，概率 p 和矩阵 M 都是变分参数，θ_j 为 0 表示该单元的参数失活。

$$\mathbf{W} = \mathbf{M} \cdot \operatorname{diag}\left(\left[\theta_j\right]_{j=1}^{K}\right) \tag{4.1}$$

$$\theta_j \sim \operatorname{Bernoulli}(p) \tag{4.2}$$

输入语句经过 Bert 预训练模型得到的句子向量表示 bert_pool_output，作为变分 Dense 层的输入，保持训练阶段与测试阶段 dropout 的设置完全一样，每一个训练批次得到该批次样本的不确定性后，根据 UAL（公式3.17）计算损失并指导模型训练，使模型尽可能以较小的不确定性分对，以较大的不确定性分错。训练完成后，得到最接近参数后验分布的伯努利分布。在测试阶段，对训练得到的参数服从的伯努利分布进行蒙特卡洛采样，集成分布上多次采样结果作为最终的预测结果和相应的不确定性（公式4.3）。

$$\text{loss}_{\text{new}} = \frac{loss_{ori}}{\text{uncertainty}^{\,(-1)^{y_{true} \Leftrightarrow y_{pred}}}}$$

$$p\left(\mathbf{y}^* \mid \mathbf{x}^*, \mathcal{D}\right) \approx \frac{1}{T} \sum_{j=1}^{T} p\left(\mathbf{y}^* \mid \mathbf{W}_j, \mathbf{x}^*\right) q_\theta^*\left(\mathbf{W}_j\right) \tag{4.3}$$

4.3.2 不确定性阈值的选择

在某一训练批次内，根据预测结果得到相应的不确定值，结合情感分类标签得到该批次内预测正确样本的不确定性分布区间和预测错误样本的不确定性分布区间，计算以每一个备选不确定性阈值为界，超过不确定性阈值的区间分布上预测错误样本与预测正确样本的差值，取最大差值对应的不确定值作为指导该批次样本进行二次修正预测的不确定性阈值。这里的阈值不是固定不变的，而是在每个训练或测试批次中动态变化的，由当前模型预测正误样本所对应的不确定性分布差异动态决定的，这种自适应的调整不确定性阈值的方式可以使阈值在不同批

次更准确、更科学、更有针对性地指导二次修正。

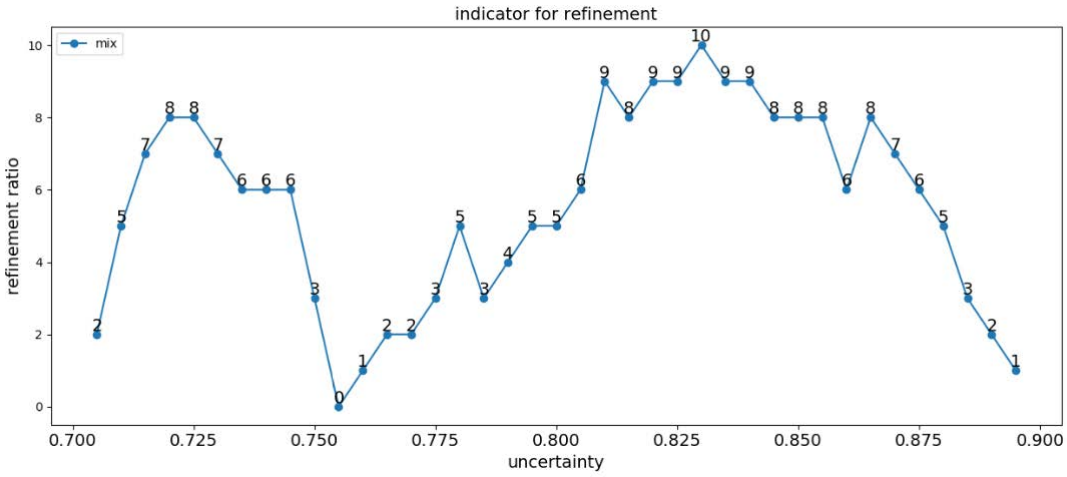


图 4.4　正误分类样本在不确定性区间分布的差异值

图4.4展示了某个批次结束之后，预测正误样本以 0.7 到 0.9 之间的不确定性值作为阈值的分布差异，具体地，当不确定性阈值选择为 0.83 时，在不确定性区间 [0.83,1) 内，预测错误的样本比预测正确的样本差值最大，意味着以不确定性阈值为界，尽量少地将分对的样本送去二次预测的同时，尽量多地将分错的样本送去二次修正预测，这样在二次修正预测的辅助下模型可以取得更好的效果。

文献[68] 于 2020 年提出了不确定性感知的两阶段修正模型解决序列标注问题，在利用模型不确定性上与该工作的关键不同之处在于，本文中不确定性阈值的选择是模型在训练过程中自适应确定的，不是根据历史经验来决定的；而且在整个训练过程中，不同训练阶段不同训练批次的不确定性阈值都是自适应变化的，不是固定不变的，这样有利于在每个阶段更有针对性地利用不确定性阈值作为样本筛选的标准。

4.3.3　基于图神经网络的修正预测模型

本文的第二阶段修正模型选择基于注意力机制的图神经网络模型 (GAT)，GAT 是图神经网络的变体，通过注意力机制对邻居节点作聚合操作，实现对不同节点权重的自适应分配，从而大大提高图神经网络模型的表达能力。

GAT 网络构造一个无标签的语法图（比如句法依存树）$\mathcal{G} = (\mathcal{V}, \mathcal{A})$，$V$ 代表所有的单词节点，A 是反映句子中单词之间依赖关系的邻接矩阵，利用掩码自注意层编码图结构。接下来将介绍邻接矩阵的构建原则和 GAT 层各个模块组成。

4.3.3.1 邻接矩阵构建的原则

1、构成方面词的所有单词之间存在相互依存关系，对应位置邻接矩阵的标签为 1。

2、根据依存分析，存在依赖关系的两个句子成分指向位置的邻接矩阵的标签为 1，无论其指向性如何，即我们构建的是单词节点之间的无向图。

3、自循环，我们认为单词指向自身的依赖关系存在，即邻接矩阵的对角线的标签值为 1。

4.3.3.2 GAT 层的模块介绍

GAT 层是以 Transformer 中的编码器为基本框架的，主要由多头注意力机制、前向网络和残差连接组成。在 transformer encoder 的基础上，根据句子成分之间的依存关系得到包含每个单词节点和其语义邻居节点信息的邻接矩阵，使其作用于多头注意力机制，显式地表征单词与句子中其它单词之间的依赖关系。

图4.5展示了图注意网络 GAT 的模型结构图：

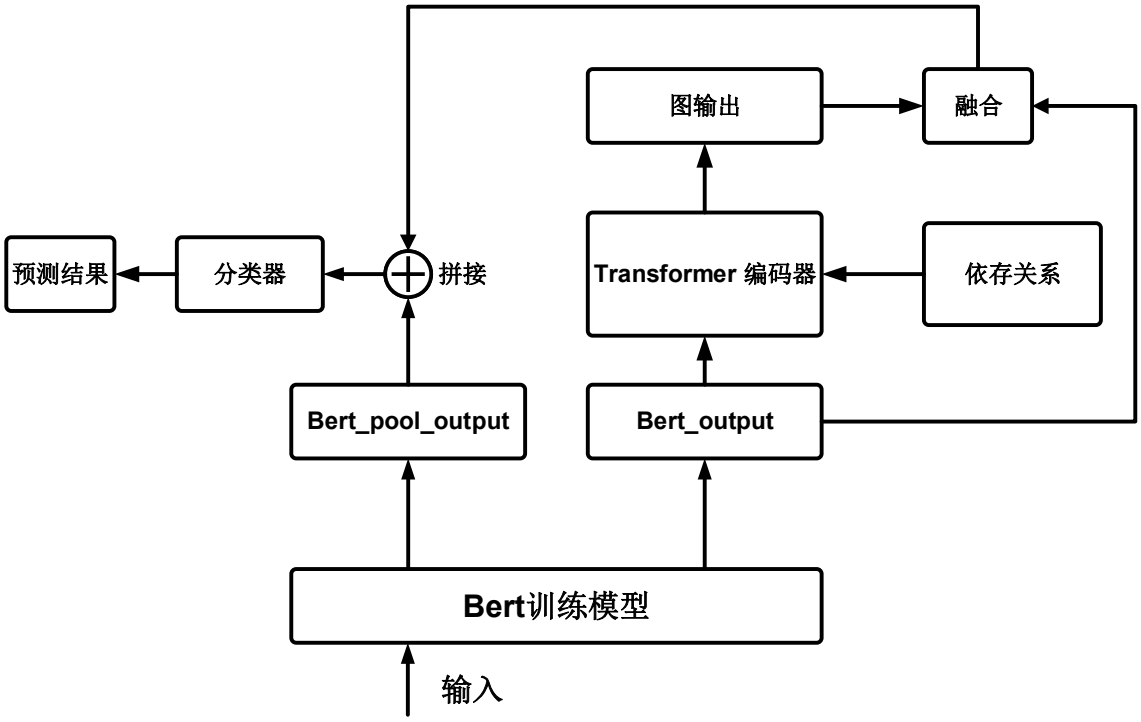


图 4.5 GAT 模型结构图

多头注意力机制：多头的注意力有助于网络捕捉到更丰富的特征和信息。将模型分成多个头，形成多个子空间，让模型去关注不同方面的信息，类似于 CNN 中多个卷积核的作用。图的邻接矩阵也将作用于此模块，用来产生与当前单词节

点没有依赖关系的其它单词的掩码表示，以屏蔽不相关单词，结合自注意力更准确地刻画单词之间的关系。

残差连接：通过将输入和输入的非线性变换叠加，更有效地保留了特征信息，又可以在一定程度上解决梯度消失问题。

特征融合：受 Bai 等人[24] 的启发，我们在 GAT 模型的最后，也选择了将图注意力模型得到的图编码输出和 Bert 模型的 Bert_output 输出进行特征融合，以得到更丰富的特征表示。

4.4 实验及结果分析

在这一节中，我们将介绍不确定性感知的两阶段修正模型的参数设置，对比不同 baseline 模型在 SemEval14 数据集（如表3.2所示）的表现，以多分类准确率 accuracy、宏观 F1 值以及极性预测相反的样本在总预测样本上的占比 opp 为评价指标，对实验结果进行分析并展开讨论。

4.4.1 实验设置

实验中使用的 Bert 预训练模型选择的是 12 层的基础模型，嵌入维度为 768，Bert 嵌入层之后变分 Dense 层的 dropout 率为 0.1，训练阶段与测试阶段保持一致，蒙特卡洛采样数设为 30，采用 Adam 优化器，初始学习率为 $2*10^{-5}$，正则项为 $1*10^{-5}$。图注意力网络 GAT 部分，堆叠两层 transformer 编码层，多头注意力的个数设置为 4。实验所依赖的硬件环境中的 GPU 为 GTX 1080Ti，操作系统为 Ubuntu。

4.4.2 结果分析

表4.2展示了各模型在 laptop 数据集上的性能比较。

表 4.2 不同模型在 laptop 数据集性能比较

model	accuracy	macro-F1	opp
TD-LSTM	70.85	65.43	7.84
ATAE-LSTM	71.47	66.28	6.43
RAM	72.41	67.32	6.27
BERT-SPC	77.43	74.03	4.08
RGAT-BERT	79.17	74.55	3.45
OURS	**81.83**	**78.74**	**2.82**

显然，我们的方法 UATR 在准确率、宏观 F1 值上都取得了最佳效果，极性预测相反的比例也达到了最低（和 UAL 指导的 RGAT-BERT 一样取到了 2.82）。与 SOTA 模型 RGAT-BERT 相比，我们的方法在准确率上有 2.66% 的提升，在宏观 F1 值上更是得到了 4.19% 的提升。宏观 F1 值可以更好地反映模型在类别不均衡数据集上的多分类效果，宏观 F1 值上更显著的提升验证了 UATR 模型在解决类别不均衡问题上的有效性。

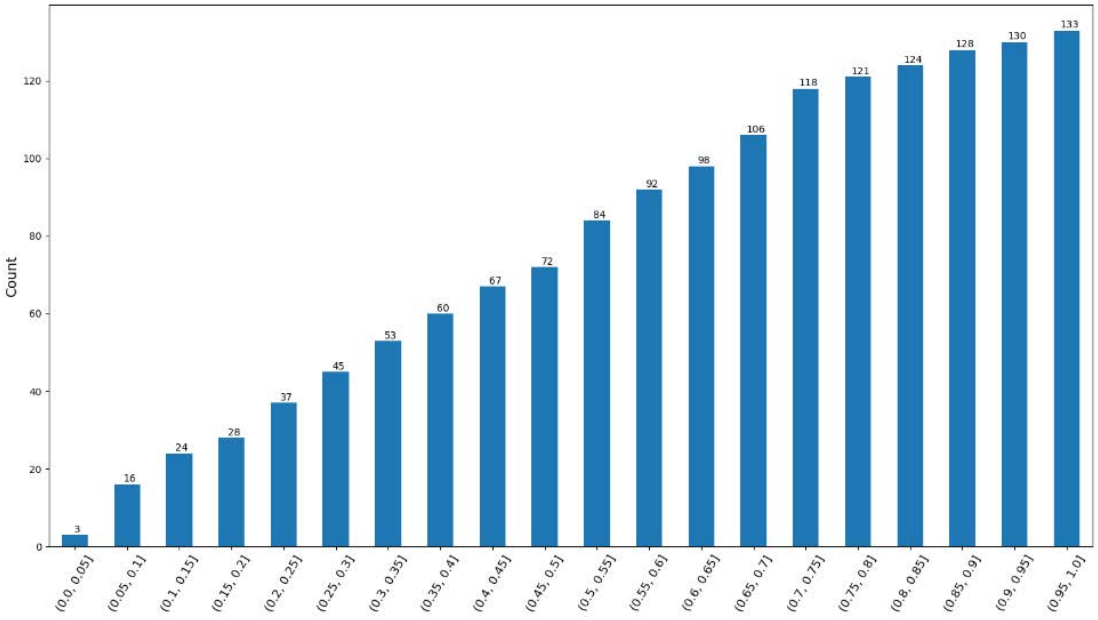


图 4.6　RGAT 在 laptop 上误分类样本累积分布图（柱状面积的大小代表着小于该区间上界的不确定性区间内误分类样本的数量，下同）

图4.6，图4.7对比了 RGAT-BERT 模型和 UATR 模型对测试数据集进行预测后得到的错误预测结果在不确定区间的累积样本分布图，可以直观地看到，UATR 模型作用下的不确定性区间更符合我们的预期，83.6% 的误分类样本集中在不确定性大于 0.5 的区间内，为模型预测结果提供了更可靠的不确定性解释。

同样地，在 restaurant 数据集上，我们的模型也在各项指标中达到了最好的效果，超过了不引入额外训练数据的 SOTA 模型 RGAT-BERT 的表现（如表4.3所示）。具体地，在准确率指标上提高了 1.7 个百分点，在宏观 F1 值上提高了 2.2 个百分点，UATR 模型带来的效果提升更显著地表现在宏观 F1 值上，同样验证了 UATR 模型在处理类别不均衡任务上的优势。除此之外，restaurant 数据集上的极性预测相反比例 opp 创造了新的最低值 2.05，比 UAL 指导的 RGAT-BERT 模型取得的 2.41 更低，可以推测，restaurant 数据集正负类别样本的不均衡程度更大，因此 UATR 用于修正极性预测相反样本占比的空间也大。

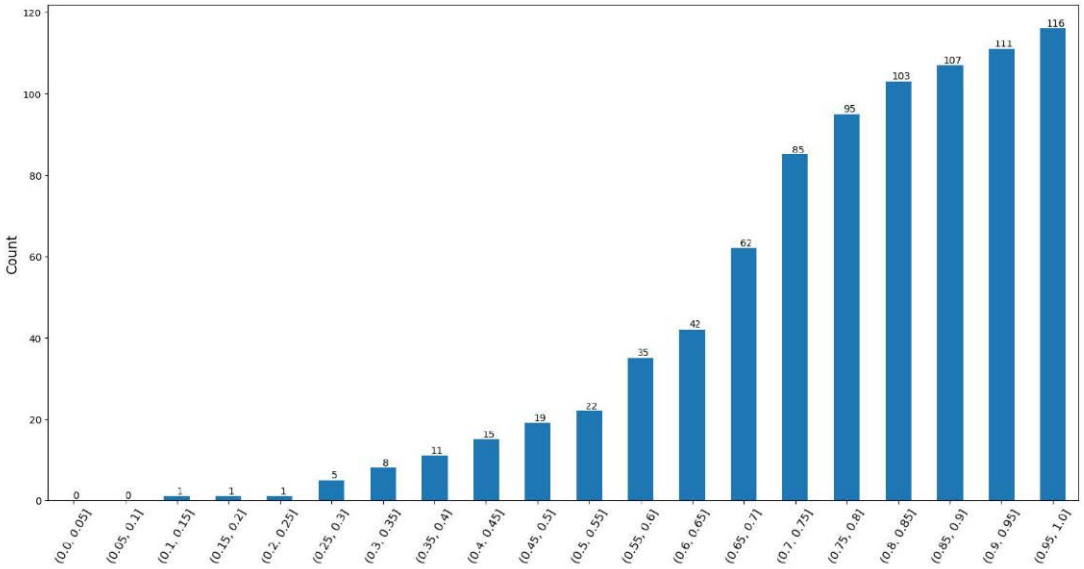


图 4.7　UATR 在 laptop 上误分类样本累积分布图

表 4.3　不同模型在 restaurant 数据集性能比较

model	accuracy	macro-F1	opp
TD-LSTM	79.11	68.51	6.70
ATAE-LSTM	77.41	66.55	7.86
RAM	80.71	71.18	5.27
BERT-SPC	84.82	77.95	3.84
RGAT-BERT	85.89	79.83	2.95
OURS	**87.59**	**82.03**	**2.05**

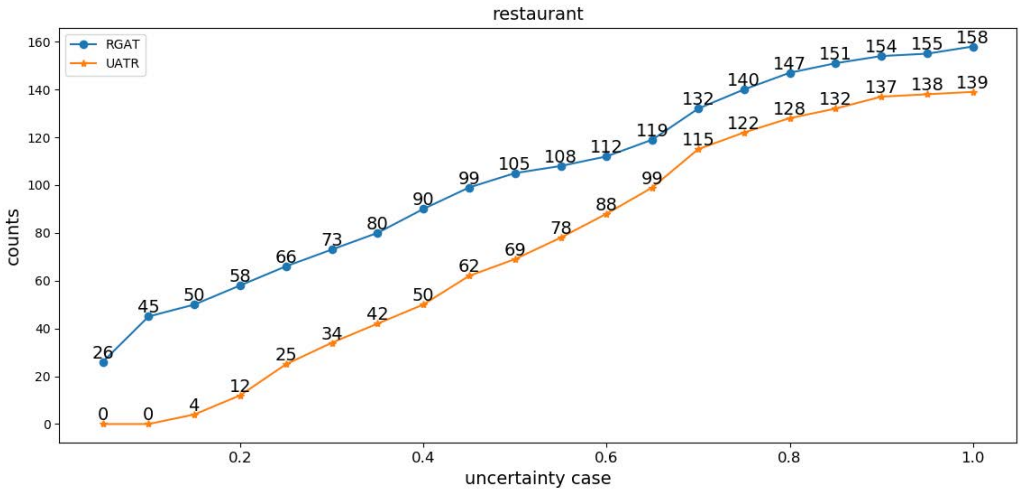


图 4.8　RGAT 和 UATR 在 restaurant 上误分类样本累积分布对比图

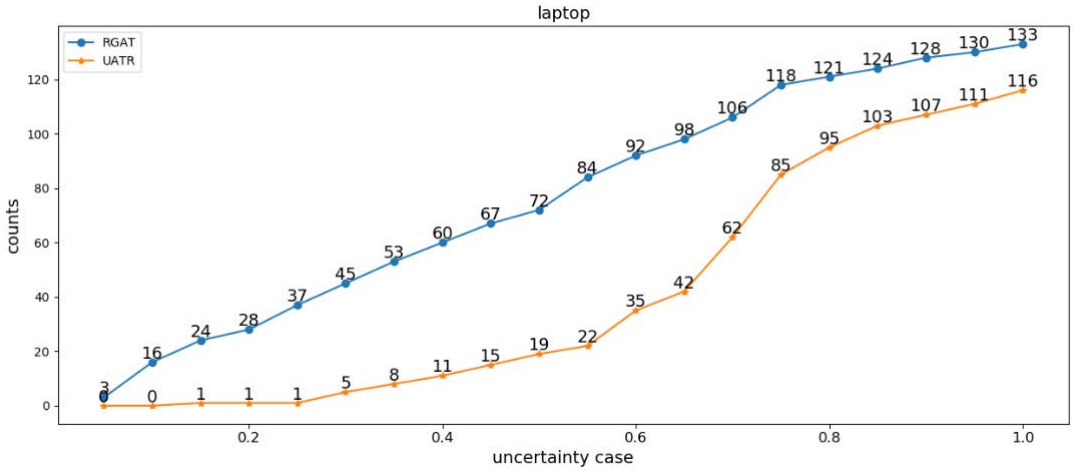


图 4.9　RGAT 和 UATR 在 laptop 上误分类样本累积分布对比图

图4.8，图4.9以折线图的形式对比了 restaurant、laptop 数据集上，RGAT-BERT 模型和 UATR 模型在误分类预测样本在各不确定性区间的累积分布变化，与 RGAT-BERT 相比，UATR 模型的折线图走势一致呈现出起点低、起步平缓，后段陡增的趋势，而且在整个不确定性区间 [0,1) 的各分区段内，UART 模型的误分类样本数都比 RGAT-BERT 模型要少，在为模型提供不确定性解释的同时印证了两阶段修正模型在细粒度多分类任务上的有效性。

表4.4以实例的形式展示了基于不确定性的两阶段修正预测模型 UATR 的修

表 4.4　UATR 方法实例说明

sample 194: the wait ***staff*** was loud and inconsiderate.							
label	pred1	uncertainty	threshold	whether to refine	pred2	final pred	changes
0	**0**	0.229	0.4	no	2	**0**	\
sample 264: my favs here are the ***tacos pastor*** and the tostada de tinga...							
label	pred1	uncertainty	threshold	whether to refine	pred2	final pred	changes
2	**2**	0.433	0.5	no	1	**2**	\
sample 215: ***Service*** is known for bending over backwards to make everyone happy.							
label	pred1	uncertainty	threshold	whether to refine	pred2	final pred	changes
2	0	**0.633**	0.35	**yes**	**2**	**2**	✗→ ✓
sample 262: ***food portion*** was small and below average.							
label	pred1	uncertainty	threshold	whether to refine	pred2	final pred	changes
0	1	**0.845**	0.5	**yes**	**0**	**0**	✗→ ✓

正过程，其中表头部分的 label 表示例句中方面词的真实情感，pred1 表示第

一节阶段 UAL 指导的不确定性估计模型预测的情感极性，uncertainty 是对应结果的不确定性，threshold 表示该批次的不确定性阈值，whether to refinement 表示根据当前不确定性阈值是否需要第二阶段修正，pred2 表示第二阶段修正预测模型的预测结果，final pred 表示综合两阶段得到的最终预测结果，changes 代表了从第一阶段预测结果到最终预测结果的正误变化。（表中加粗单词对应句子中的方面词）

4.4.3 消融研究

这部分我们分别探究 UAL 方法和 UATR 两阶段方法的有效性，设置五组对比实验，分别是用于不确定性感知修正模型的第一阶段预测模型（记为 STAGE1）、用于不确定性感知修正模型的二次修正预测模型 ——图注意力网络（记为 STAGE2）、目前的 SOTA 模型 RGAT-BERT、没有自定义损失函数 UAL 指导的 UATR 模型（UATR$_{-UAL}$）以及在 UAL 指导下的 UATR 模型。表4.5记录了五个模型在 laptop 和 restaurant 数据集上的表现。

表 4.5 关于 UAL 和 UATR 两阶段修正方法的消融研究

	laptop		**restaurant**	
model	accuracy	macro-F1	accuracy	macro-F1
STAGE1	78.95	75.02	85.54	78.66
STAGE2	78.75	74.45	84.91	77.74
RGAT-BERT	79.17	74.55	85.89	79.83
UATR$_{-UAL}$	81.39	78.34	86.25	79.77
UATR	**81.83**	**78.74**	**87.59**	**82.03**

4.4.3.1 UATR 的消融分析

分析表中呈现的结果，分别对比 STAGE1、STAGE2、RGAT-BERT 和 UATR$_{-UAL}$ 与 UATR 模型上的表现，没有 UAL 指导的 UATR$_{-UAL}$ 模型取得了仅次于 UATR 模型的效果，同样超过了 SOTA 模型。不难得出结论，本文提出的不确定性指导的两阶段修正方法是合理的，融合了不确定性解释和图神经网络捕捉句子中成分之间依赖关系的特性，在保证模型整体复杂度的前提下，进一步提高了分类准确率和宏观 F1 值。

4.4.3.2 UAL 的消融分析

观察 UATR$_{-UAL}$ 和 UATR 的实验结果，在两个数据集上，UAL 指导的 UATR 效果均更胜一筹，以此证明了 UAL 方法的有效性。而且与 laptop 数据集相比，在

restaurant 数据集上，UAL 指导的 UATR 模型较没有 UAL 指导的 UATR$_{-UAL}$ 模型效果提升更显著，即不确定性感知的自定义损失函数 UAL 对类别不均衡的数据集改善更好一些。

4.4.4 讨论

本节就 UATR 模型与经典基线模型之间的优劣进行分析，并对 UATR 取得较好效果的原因进行了讨论。

4.4.4.1 优劣比较

UATR 模型与基线模型相比，利用了贝叶斯神经网络整合多组神经网络的预测结果建模不确定性，削弱了数据集规模太小、类别样本不均衡带来的影响。UAL 的设计可以使模型预测结果分布在合理的不确定性区间内，为模型预测结果提供较为可靠的不确定性解释。动态不确定性阈值作用下的两阶段修正预测模型可以很好地发挥两阶段的优势，达到互补的效果。有效降低了预测样本中极性预测相反样本所占的比例，并且进一步提高了准确率和宏观 F1 值。但是仅仅由 UAL 指导的不确定性模型在准确率和宏观 F1 值等指标上没有稳定的提高，不确定性会带来结果的波动性，因此需要把不确定性网络与确定性网络结合起来。

而经典的基线模型都是标准的神经网络模型，是确定性网络，对训练好的模型进行多次预测得到的结果是一样的，不能反映模型对预测结果的置信度，从而限制了模型的性能。

4.4.4.2 优势分析

首先 UATR 模型是基于不确定性设计的模型，第一阶段 UAL 指导的不确定性估计模型中，不确定性的建模需要整合权重分布上的多组神经网络，其预测结果较确定性神经网络的预测结果更具一般性。

其次，不确定性损失函数 UAL 的设计使预测结果的不确定性伴随预测结果的正误一起反馈到模型训练中去，从而最终得到的不确定性可以更好地反映模型对相应预测结果的置信度，为第二阶段根据不确定值筛选样本再预测提供可行性支持。

最后，根据第一阶段预测正误样本在不确定性区间的分布差异得到该批次的不确定性阈值，使不确定性高于阈值的样本有机会进行二次预测。二次修正预测模型，使用图注意网络建模句子中各成分之间的依赖关系得到新的预测结果。第一阶段不确定性模型和第二阶段确定性模型的结合，既发挥了不确定性的优势，又减小了不确定性带来的波动性。

4.5 本章小结

不确定性感知的损失函数 UAL 虽然为模型预测结果提供了可靠的不确定性解释，而且降低了预测样本中极性预测相反的样本所占的比例，但是不能有效地提高模型的分类准确率和宏观 F1 值，依然不能很好地服务于生产生活。在此基础上，本章提出不确定性感知的两阶段修正预测模型 UATR，不仅利用模型不确定性指导训练，而且设计策略动态得到不确定性阈值，从而筛选样本进行二次修正预测，我们的模型在 SemEval14 数据集上取得了比 SOTA 模型更好的效果。具体可分为以下几个部分：

第一部分，对本章的研究内容、方法设计作了整体的介绍，并且给出了模型的整体框架图。

第二部分，介绍了模型所需的数据处理格式和拟采用的词嵌入方式。

第三部分，从第一阶段的预测和第二阶段的修正预测两方面出发，详细介绍了模型设计部分。包括作为第一阶段预测的 UAL 指导的不确定性估计模型的 MC Dropout 实现，作为修正预测阶段的图注意力网络的原理，以及关于不确定性阈值的动态选择策略。

第四部分，实验验证部分，对 UATR 和包括 SOTA 在内的 baseline 进行对比分析，UATR 较 SOTA 模型在 laptop 和 restaurant 数据集上都有更好的效果，其中分类准确率分别提高了 2.66%、1.7%，宏观 F1 值分别提高了 4.19%、2.2%。最后就 UAL 和 UATR 的有效性进行了消融分析。

第五章　总结与展望

本章主要对全文的研究内容和创新点作简要总结，并分别从技术角度和应用角度对课题未来可能的研究方向进行展望。

5.1　全文总结

自 2011 年以来，基于深度学习的方法在细粒度情感分析任务的研究中被广泛使用，较基于词典和基于统计的机器学习方法而言，深度学习凭借其强大的特征拟合能力往往可以取得更好的效果，也更容易得到研究人员的青睐。尽管如此，标准的神经网络对数据集敏感，确切地说，神经网络模型的效果受数据集规模以及类别样本是否均衡的影响。针对实际应用场景中普遍存在的数据集规模较小、类别样本不均衡的特点，本文引入模型不确定性，建模不确定性在一定程度上相当于集成某权重分布上的无穷多组神经网络进行预测，可以弥补数据规模太小、容易过拟合的不足。本文探究了如何利用模型不确定性提高细粒度情感分类任务的性能。

首先，通过将标准的神经网络模型改进为基于贝叶斯神经网络的模型，建模模型的不确定性，考虑模型预测结果和对应不确定性的一致性关系，提出不确定性感知的损失函数 UAL，让不确定性参与到模型的训练当中。使模型尽量以较小的不确定性将样本分对、以较大的不确定性将样本分错，为模型的预测结果提供更可靠的不确定性解释，同时实验结果表明 UAL 指导训练的方法有效降低了模型极性预测相反样本占总测试样本的比例。

其次，在有效降低极性预测相反样本占比、为模型预测结果提供更可靠不确定性解释的前提下，本文进一步提出不确定性感知的两阶段修正模型进行分类预测 UATR，设计策略动态地得到不确定性阈值，以此选择需要二次修正的样本，提高细粒度情感分类的准确率。经实验验证，我们的方法较目前的 SOTA 模型有显著的提高。

本文针对细粒度情感三分类任务中存在相当比例极性预测相反样本的现象，引入模型不确定性考量，为细粒度情感分析任务提供了一个新的研究思路，基于贝叶斯神经网络整合多组神经网络的预测结果来建模不确定性，削弱数据规模太小以及类别样本不均衡带来的影响，并设计了合理有效的策略和模型来对不确定性加以利用，为模型预测结果提供可靠的不确定性解释的同时提高整体的分类准确率，经实验验证，取得了比当前 SOTA 模型更好的效果。

5.2 未来展望

从简单的二分类到多分类，从单一的针对方面词的情感分类任务到端到端的情感词抽取和分类任务，关于细粒度情感分析研究的广度和深度都在增长。尽管本文基于不确定性展开的研究取得了一定成果，但是仍然有很大的研究空间，距离直接应用到实际场景中还有一定的距离。接下来将从技术和应用方面进行展望。

技术方面，可以设计更具针对性的不确定性指导方案，并且探索其它模型的修正预测组合，如胶囊网络等等。

应用方面，可以探究是否在其它任务上有普适性，比如五分类问题，以及其它的自然语言理解和生成任务。

致　谢

三年的时间转瞬即逝，毕业论文的完成意味着我们即将奔赴下一个港口。三年来在科大感受到了别样的人文环境，有期待、有惊喜，有彷徨、有失落，很开心能够加入 UPCOM 大家庭，和大家一起学习进步，迈出应用计算机科学技术学习人工智能从 0 到 1 最重要的一步。

衷心感谢我的导师王晓东研究员，王老师凭借着本领域多年的积淀为我的课题把握着整体的方向，看待问题往往有着开阔的思维和独特的视角，在我的课题研究中提出了很多建设性的意见。王老师不仅在学术上严谨认真，工作中以身作则，而且生活中也是乐观积极，善于开导同学，为同学们排忧解难。在防范疫情就地过年期间，主动关心大家的生活状况并提供帮助。同时感谢吕绍和老师提供的算力支持。两位老师的言传身教将使我终生受益。

感谢所有的任课老师，尽管只有短短一个学期的学习时间，但是你们对学科的专精和热情感染着我，不同知识之间的交叉关联帮助我构建了更完整的知识体系，给我的课题带来了很大的启发；感谢杜冰瑜老师俄语课精彩地讲授。

感谢五队宋继有队长、谭培峰队长、杨红运政委，三位队干部幽默风趣，为我们的日常生活保驾护航，努力为我们提供良好的学习生活环境。

感谢我的家人，家永远是一个人最强大的后盾，在压力较大的日子里，每周的家庭视频是我最轻松、最解压、最快乐的时光，尤其感谢爸爸妈妈和姐姐一直以来的支持，谢谢你们一致鼓励我不断追寻自己喜欢的事情，感谢小宝的胡言胡语为我的生活提亮增色。感谢童志国，你求实求真的态度很可贵，在成长的路上我们亦师亦友，希望未来可以携手并进，一起熠熠发光。

感谢 UPCOM 的所有成员。感谢祁小乐、汪东、涂宏魁、王芝辉、蒋苏师兄，黄韵欣、武玫含师姐营造的浓厚的科研氛围，以及提供的科研建议。感谢余伟、常滔、陈飞师兄，谢谢你们在我的课题研究中提出的中肯的建议，让我少走了很多弯路。感谢同级的常城扬、罗翰文，谢谢你们的头脑风暴和减压神曲。感谢师弟师妹积极组织团建，让实验室充满活力。

感谢在科大遇到的各位好友，很幸运能够遇到温柔可爱的你们。感谢我的室友马锶霞，王元，张文静，我们的性格不尽相同，从你们的身上我学到了很多，也很开心这三年有你们的陪伴。感谢洛洛、瑶瑶、神奇四侠，你们是我在科大收获的独一无二的友情；感谢武协，可以让我在这个专业的地方得到更专业的训练，你们百折不挠的精神永远值得我学习。

感谢学校图书馆提供的丰富的数字资源和方便的检索工具。

最后，也感谢一直努力的自己，希望自己可以继续认真、努力、幸运，成为一个有趣的人。

参考文献

[1] 李超雄. 基于主题模型的网络短文本情感分析研究[D]. 福建师范大学, 2016.

[2] 王伟平. 社交媒体情感分析[R]. 中国科学院信息工程研究所, 2019.

[3] 冯多，林政，付鹏，王伟平. 基于卷积神经网络的中文微博情感分类[J]. 计算机应用与软件, 2017, 34(04): 157~164.

[4] Pang B, Lee L. Opinion Mining and Sentiment Analysis[J/OL]. Foundations and Trends® in Information Retrieval, 2008, 2(1–2): 1~135. http://dx.doi.org/10.1561/1500000011. DOI: 10.1561/1500000011.

[5] Zhang L, Wang S, Liu B. Deep learning for sentiment analysis: A survey[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018, 8(4): e1253.

[6] Sista S P, Srinivasan S H. Polarized Lexicon for Review Classification[C]// MLMTA'04: Proceedings of the International Conference on Machine Learning; Models, Technologies & Applications, CSREA. Press, 2004.

[7] Fersini E, Messina E, Pozzi F A. Sentiment analysis: Bayesian ensemble learning[J]. Decision support systems, 2014, 68: 26~38.

[8] Zhu X, Sobihani P, Guo H. Long short-term memory over recursive structures[C]//International Conference on Machine Learning. 2015: 1604~1612.

[9] Wang Y, Huang M, Zhu X, et al. Attention-based LSTM for aspect-level sentiment classification[C]//Proceedings of the 2016 conference on empirical methods in natural language processing. 2016: 606~615.

[10] Baziotis C, Pelekis N, Doulkeridis C. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis[C]// Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017). 2017: 747~754.

[11] Chen P, Sun Z, Bing L, et al. Recurrent attention network on memory for aspect sentiment analysis[C]//Proceedings of the 2017 conference on empirical methods in natural language processing. 2017: 452~461.

[12] Tang D, Qin B, Liu T. Learning semantic representations of users and products for document level sentiment classification[C]//Proceedings of the 53rd Annual Meeting

of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015: 1014~1023.

[13] Xu J, Chen D, Qiu X, et al. Cached long short-term memory neural networks for document-level sentiment classification[J]. ArXiv preprint arXiv:1610.04989, 2016.

[14] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques[J]. ArXiv preprint cs/0205070, 2002.

[15] Kim Y. Convolutional Neural Networks for Sentence Classification[Z]. 2014. arXiv: 1408.5882 [cs.CL].

[16] Zhou J, Huang J X, Chen Q, et al. Deep learning for aspect-level sentiment classification: Survey, vision, and challenges[J]. IEEE access, 2019, 7: 78454~78483.

[17] Li X, Bing L, Zhang W, et al. Exploiting BERT for end-to-end aspect-based sentiment analysis[J]. ArXiv preprint arXiv:1910.00883, 2019.

[18] Pontiki M, Galanis D, Pavlopoulos J, et al. SemEval-2014 Task 4: Aspect Based Sentiment Analysis[C/OL]//Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: Association for Computational Linguistics, 2014: 27~35. https://www.aclweb.org/anthology/S14-2004. DOI: 10.3115/v1/S14-2004.

[19] Dong L, Wei F, Tan C, et al. Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification[C/OL]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Baltimore, Maryland: Association for Computational Linguistics, 2014: 49~54. https://www.aclweb.org/anthology/P14-2009. DOI: 10.3115/v1/P14-2009.

[20] Socher R, Perelygin A, Wu J, et al. Parsing With Compositional Vector Grammars[G]//EMNLP. 2013.

[21] Hu M, Liu B. Mining and summarizing customer reviews[C]//Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 2004: 168~177.

[22] Deng L, Wiebe J. Mpqa 3.0: An entity/event-level sentiment corpus[C]// Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies. 2015: 1323~1328.

[23] Pang B, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts[J]. ArXiv preprint cs/0409058, 2004.

[24] Bai X, Liu P, Zhang Y. Investigating Typed Syntactic Dependencies for Targeted Sentiment Classification Using Graph Attention Neural Network[J]. IEEE/ACM

Transactions on Audio, Speech, and Language Processing, 2021, 29: 503~514. DOI: 10.1109/TASLP.2020.3042009.

[25] Zeldes Y, Naor I. Using Uncertainty to Interpret your Model[Z]. 2018.

[26] Ribeiro M T, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier[Z]. 2016. arXiv: 1602.04938 [cs.LG].

[27] Der Kiureghian A, Ditlevsen O. Aleatory or epistemic? Does it matter?[J]. Structural safety, 2009, 31(2): 105~112.

[28] Gal Y. Uncertainty in Deep Learning[D]. University of Cambridge, 2016.

[29] Goan E, Fookes C. Bayesian Neural Networks: An Introduction and Survey[G]// Case Studies in Applied Bayesian Data Science. Springer, 2020: 45~87.

[30] Gal Y, Ghahramani Z. Bayesian convolutional neural networks with Bernoulli approximate variational inference[J]. ArXiv preprint arXiv:1506.02158, 2015.

[31] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. ArXiv preprint arXiv:1301.3781, 2013.

[32] 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.

[33] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. ArXiv preprint arXiv:1810.04805, 2018.

[34] 刘瑾莱. 基于深层神经网络推理的图像问答技术研究和应用[D]. 北京邮电大学, 2019.

[35] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[J]. ArXiv preprint arXiv:1802.05365, 2018.

[36] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]., 2018.

[37] Jawahar G, Sagot B, Seddah D. What Does BERT Learn about the Structure of Language?[C/OL]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 3651~3657. https://www.aclweb.org/anthology/P19-1356. DOI: 10.18653/v1/P19-1356.

[38] 陶杨明. 基于深度学习的细粒度文本情感分析研究[D]. 浙江工商大学, 2020.

[39] Kundi F M, Khan A, Ahmad S, et al. Lexicon-based sentiment analysis in the social web[J]. Journal of Basic and Applied Scientific Research, 2014, 4(6): 238~48.

[40] Huang M, Qian Q, Zhu X. Encoding syntactic knowledge in neural networks for sentiment classification[J]. ACM Transactions on Information Systems (TOIS), 2017, 35(3): 1~27.

[41] Tang D, Qin B, Liu T. Document modeling with gated recurrent neural network for sentiment classification[C]//Proceedings of the 2015 conference on empirical methods in natural language processing. 2015: 1422~1432.

[42] Hu X, Tang J, Gao H, et al. Unsupervised sentiment analysis with emotional signals[C]//Proceedings of the 22nd international conference on World Wide Web. 2013: 607~618.

[43] Miller G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39~41.

[44] Ku L W, Chen H H. Mining opinions from the Web: Beyond relevance retrieval[J]. Journal of the American Society for Information Science and Technology, 2007, 58(12): 1838~1850.

[45] Dong Z, Dong Q. HowNet-a hybrid language and knowledge resource[C]// International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003. 2003: 820~824.

[46] 陈建美. 中文情感词汇本体的构建及其应用[D]. 大连理工大学, 2009.

[47] He R, Lee W S, Ng H T, et al. An unsupervised neural attention model for aspect extraction[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 388~397.

[48] Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules[J]. ArXiv preprint arXiv:1710.09829, 2017.

[49] Wang Y, Sun A, Han J, et al. Sentiment analysis by capsules[C]//Proceedings of the 2018 world wide web conference. 2018: 1165~1174.

[50] Wang Y, Sun A, Huang M, et al. Aspect-level sentiment analysis using as-capsules[C]//The World Wide Web Conference. 2019: 2033~2044.

[51] Du C, Sun H, Wang J, et al. Capsule network with interactive attention for aspect-level sentiment classification[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: 5492~5501.

[52] Su J, Yu S, Luo D. Enhancing Aspect-Based Sentiment Analysis With Capsule Network[J]. IEEE Access, 2020, 8: 100551~100561.

[53] Fang X, Tao J. A transfer learning based approach for aspect based sentiment analysis[C]//2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). 2019: 478~483.

[54] Li Z, Li X, Wei Y, et al. Transferable end-to-end aspect-based sentiment analysis with selective adversarial learning[J]. ArXiv preprint arXiv:1910.14192, 2019.

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

[56] Zhang M, Zhang Y, Vo D T. Gated neural networks for targeted sentiment analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 30: 1. 2016.

[57] Tang D, Qin B, Feng X, et al. Effective LSTMs for target-dependent sentiment classification[J]. ArXiv preprint arXiv:1512.01100, 2015.

[58] Sun K, Zhang R, Mensah S, et al. Aspect-level sentiment analysis via convolution over dependency tree[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: 5683~5692.

[59] Huang B, Carley K M. Syntax-aware aspect level sentiment classification with graph attention networks[J]. ArXiv preprint arXiv:1909.02606, 2019.

[60] Ma D, Li S, Zhang X, et al. Interactive attention networks for aspect-level sentiment classification[J]. ArXiv preprint arXiv:1709.00893, 2017.

[61] Denker J S, leCun Y. Transforming Neural-Net Output Levels to Probability Distributions[C]//NIPS-3: Proceedings of the 1990 Conference on Advances in Neural Information Processing Systems 3. Denver, Colorado, USA: Morgan Kaufmann Publishers Inc., 1990: 853~859.

[62] MacKay D J. A practical Bayesian framework for backpropagation networks[J]. Neural computation, 1992, 4(3): 448~472.

[63] Neal R M. Bayesian learning for neural networks.[D]. University of Toronto, 1995.

[64] Gal Y, Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning[C]//International conference on machine learning. 2016: 1050~1059.

[65] Gal Y, Ghahramani Z. A theoretically grounded application of dropout in recurrent neural networks[J]. Advances in neural information processing systems, 2016, 29: 1019~1027.

[66] Kendall A, Gal Y. What uncertainties do we need in bayesian deep learning for computer vision?[J]. ArXiv preprint arXiv:1703.04977, 2017.

[67] Song Y, Wang J, Jiang T, et al. Attentional encoder network for targeted sentiment classification[J]. ArXiv preprint arXiv:1902.09314, 2019.

[68] Gui T, Ye J, Zhang Q, et al. Uncertainty-Aware Label Refinement for Sequence Labeling[J]. ArXiv preprint arXiv:2012.10608, 2020.

作者在学期间取得的学术成果

发表的学术论文

[1] 第一作者. An Overview on Fine-grained Text Sentiment Analysis: Survey and Challenges[C]//Journal of Physics: Conference Series. IOP Publishing, 2021, 1757(1): 012038. (EI 收录, 检索号:20211110059831.)

[2] 第三作者. Ugan: unified generative adversarial networks for multidirectional text style transfer[J]. IEEE Access, 2020, 8: 55170-55180. (SCI 收录, 检索号:20201508385590.)

公开评阅信息

序号	评阅人	职称	导师类型	工作单位	总分	结论	答辩建议	熟悉程度	备注
1	姜晶菲	研究员	硕导	国防科技大学	91	达到	无需修改直接答辩	比较熟悉	
2	李荣春	副研究员	硕导	国防科技大学	91.05	达到	修改后答辩	比较熟悉	

说明：

1. 结论选项包括 2 个：“达到硕士学位论文要求”、“尚未达到硕士学位论文要求”。
2. 答辩建议选项包括 4 个：“无需修改直接答辩”、“修改后答辩”、“修改后复评”、“不予答辩”。
3. 熟悉程度选项包括 3 个：“有深入了解”、“比较熟悉”、“一般了解”。