****

**本科生毕业论文（设计）**



**题 目 基于自然语言处理的鲁棒性问答系统**

**学 院 计算机学院**

**专 业 计算机科学与技术**

**学生姓名 李哲玮**

**学 号 2018141461190 年级 2018**

**指导教师 段 磊**

**教务处制表**

**二〇二二年五月二十日**

**基于自然语言处理的鲁棒性问答系统**

专业：计算机科学与技术

学生：李哲玮 指导老师：段磊

**[摘要]** 问答系统是一个处理自然语言形式问题的系统。它在人类与计算机之间提供了一种便捷的交互方式，近年来已被广泛运用到多个日常生活的领域当中。但是随着网络大数据与人工智能时代的到来，单一领域数据来源的自动问答系统已经无法满足人们对各类信息处理的需求。因此设计一种跨领域间的，具有较强泛化能力的鲁棒性问答系统逐渐成为自然语言处理中的研究热点。为构建一个鲁棒性问答系统，主要利用了自然语言处理技术，在大规模语言预训练模型 BERT 的基础架构上进行改进，提出了针对跨领域问答系统的 RQA­BERT 模型。与此同时，该模型引入了迁移学习中的对抗训练以及数据增强策略。首先对少量域外训练数据进行数据增强，再通过构建一个额外的领域鉴别器模型，让问答系统中的编码器得以混淆两个不同领域数据分布，从而实现跨领域间的鲁棒性问答系统。通过在跨领域的抽取式问答系统数据集上开展实验，最终实验结果证明了该模型在处理跨领域问答上的有效性。

**[关键词]** 自然语言处理；问答系统；预训练语言模型；小样本学习；迁移学习

**An Robust Question Answering System based on Natural**

**Language Processing**

Computer Science and Technology

Student: LI Zhewei Tutor: DUAN Lei

**[Abstract]** A Question Answering (QA) system deals with natural language forms of questions. It provides a convenient way of interaction between humans and computers and has been widely used in many areas of life in recent years. However, with the advent of big data and artificial intelligence, QA from a single data source are no longer able to meet people's needs for various types of information processing. Therefore, the design of a robust question answering system with strong generalization capability across domains has become a hot research topic in natural language processing. In order to build a robust QA system, the RQA-BERT model for cross-domain question answering is proposed by constructing on the pre-trained language model, BERT, using mainly NLP techniques. At the same time, the model introduces adversarial learning and data augmentation strategies in transfer learning. A small amount of out-of-domain training data is augmented with data, and an additional domain discriminator model is employed to encourage the encoder to confuse two invariant domain data distributions, thus completing a robust QA system across domains. The evaluation results demonstrate the effectiveness of the proposed model by gaining an increase on the cross-domain extractive QA benchmark experiment.

**[Key words]**: Natural Language Process; Question Answering System; Pretrained Language Model; Few-shot

Learning; Transfer Learning

目 录

[1 绪论 1](#_Toc103718610)

[1.1 选题背景 1](#_Toc103718611)

[1.2 国内外研究现状 2](#_Toc103718612)

[1.2.1 问答系统 2](#_Toc103718613)

[1.2.2 小样本学习 3](#_Toc103718614)

[1.3 研究问题与挑战 4](#_Toc103718615)

[1.4 论文主要工作 5](#_Toc103718616)

[1.5 论文组织结构 5](#_Toc103718617)

[2 相关基础知识 7](#_Toc103718618)

[2.1 深度神经网络 7](#_Toc103718619)

[2.1.1 全连接层 7](#_Toc103718620)

[2.1.2 层规范化层 8](#_Toc103718621)

[2.2 激活函数 8](#_Toc103718622)

[2.3 损失函数 9](#_Toc103718623)

[2.4 小结 10](#_Toc103718624)

[3 基于 BERT 的鲁棒性问答系统 11](#_Toc103718625)

[3.1 实验设置 11](#_Toc103718626)

[3.1.1 问答系统 11](#_Toc103718627)

[3.1.2 鲁棒性问答系统 11](#_Toc103718628)

[3.2 实验过程 12](#_Toc103718629)

[3.2.1 基于预训练语言模型的问答系统 12](#_Toc103718630)

[3.2.2 对抗训练 15](#_Toc103718631)

[3.2.3 数据增强 16](#_Toc103718632)

[3.2.4 长度惩罚损失 17](#_Toc103718633)

[3.3 小结 18](#_Toc103718634)

[4 实验结果以及分析 19](#_Toc103718635)

[4.1 实验准备 19](#_Toc103718636)

[4.1.1 实验数据集 19](#_Toc103718637)

[4.1.2 评估指标 19](#_Toc103718638)

[4.1.3 对比实验 20](#_Toc103718639)

[4.2 实验结果 20](#_Toc103718640)

[4.2.1 预训练语言模型对实验结果影响 22](#_Toc103718641)

[4.3 消融实验 22](#_Toc103718642)

[4.3.1 对抗训练对实验结果影响 22](#_Toc103718643)

[4.3.2 数据增强对实验结果影响 23](#_Toc103718644)

[4.3.3 长度惩罚对实验结果影响 23](#_Toc103718645)

[4.4 参数敏感性实验 23](#_Toc103718646)

[4.5 小结 24](#_Toc103718647)

[5 结论与展望 25](#_Toc103718648)

[5.1 工作总结 25](#_Toc103718649)

[5.2 工作展望 25](#_Toc103718650)

**[参考文献](#_Toc103718651)** [27](#_Toc103718651)

**[声 明](#_Toc103718652)** [30](#_Toc103718652)

**[致 谢](#_Toc103718653)** [31](#_Toc103718653)

# 绪论

## 选题背景

自然语言处理是实现人工智能技术中至关重要的一个环节。它赋予计算机“理解”人类语言的能力，使得计算机能够通过接受自然语言文本形式的输入，经过处理转换为机器语言，并在计算机上根据目的执行任务，最终面向人类指定任务输出。在如今网络信息以指数爆炸形式增长的时代，如何高效地，精确地检索和处理信息成为了当今世界的一个难题。而自然语言处理技术的应运而生，提供了一个便捷的人机交互方式，在计算机和人类之间搭建起了一座沟通交流的桥梁，让人类和信息之间的交互方式变得更加方便，更加简单。

近年来，随着深度学习研究在自然语言处理领域的不断发展，利用神经网络进行语义分析，信息检索，机器翻译等自然语言处理任务已成为主流研究方法。相较于二十世纪提出的基于规则匹配和复杂逻辑推理的数理统计模型，深度学习的神经网络模型能够更加精准地捕捉到语句中深层次的语义信息，实现端对端的模型训练，避免依赖统计模型中各个独立子模块之间繁琐的信息交互处理。

问答系统作为自然语言处理中的一项信息检索技术的重要应用。它通过接受用户以自然语言文本（非结构化文本) 形式输入问题，从大量数据中自动处理、检索、查找并返回对应问题的答案。问答系统覆盖范围十分广泛，根据不同的划分标准，对问答系统种类的分类也不尽相同。根据问答知识信息来源分类，大致可分为基于知识库问答、基于文档问答、基于网络文件问答。根据问答领域进行分类，又可分为检索类问答、交互性问答、阅读理解性问答、开放性领域问答。在开放性领域进行自动问答则既是重点也是难点。本文研究的方向则是关于基于文档信息的抽取式阅读理解问答系统。作为自然语言处理领域的明日之星，问答系统在人们的日常生活的多个领域中也拥有重大的实际应用价值。比如聊天机器人，知识图谱问答，智能客服，搜索引擎等等。

随着不断有新颖的深度学习模型的提出，自然语言处理各项子任务中的最先进模型的评价指标也逐渐逼近，甚至超越人类的水平[[1](#_bookmark51)]。但是令人遗憾的是，这些模型并不能像人类一样拥有真正阅读理解自然语言的能力。而目前提出的大多数 NLP 模型仅仅考虑训练源领域数据与测试目标领域数据均属于同一领域分布的情况，但是在现实生活应用场景中，训练数据与测试数据属于不同领域分布等情况往往更加普遍。当一个单一领域的最先进模型在面临多领域，少样本的情况下，实验结果的评价指标数值却相对原先结果均有巨大的下降[[2](#_bookmark52)]。例如在基于文档的问答系统中，同一模型在 SQuAD 数据集上 F1 score 评价指标相对混合多领域数据集上测试结果下降超过50%[[2](#_bookmark52)]。

所以，解决不同数据集间分布转移 (Distribution Shift)问题，提高模型鲁棒性已经成为自然语言处理的一项重要研究趋势[[2](#_bookmark52),[3](#_bookmark53)]。这也是把 NLP 技术应用到更加广泛的日常生活中的关键之处，也是实现强人工智能的必经之路。

## 国内外研究现状

### 问答系统

问答系统主要包括三个子模块：自然语言理解(Natural Language Understanding)，信息检索(Information Retrieval)，阅读理解(Reading Comprehension)。在接受到用户的自然语言输入后，经过自然语言理解，获取语句深层次语义，然后在信息资源库中进行信息检索，最终在检索结果上通过阅读理解，进行答案抽取或者文本生成技术输出对应问题的答案。其中的信息检索和答案抽取模块则是其中最为关键的两个部分。

信息检索技术是一种广泛应用于搜索引擎的技术，其主要任务是从海量的数据源或庞大的数据库中通过一系列匹配操作，对用户信息进行分析，查询，并找到与问题匹配的内容。统计学家提出多种数理统计模型对问题和文本进行匹配，比如 Bi­gram 模型，以及TF­IDF[[4](#_bookmark54)] 来衡量检索结果和目标文本的相关性。

在通过信息检索模块后，此时已经从大量数据中选取出了少量的候选段落，语句。此时还需要从候选语句中抽取出相关的句子或词语作为最终答案。虽然某种意义上信息检索和答案抽取都是在数据中对问题进行文本匹配，但是需要注意的是，信息检索和答案抽取的关系更像是宏观阅读和微观阅读的区别。答案抽取需要对候选文本进行粒度更高的划分，因此当一个机器能够根据给出的一段文本，回答一个问题的这个过程，此时需要更加精确的语义层面分析理解，也需要对整个文本段落的上下文有深刻理解。这一步骤又被称为阅读理解(Reading Comprehension)。Burges[[5](#_bookmark55)]指出，当一个机器答对绝大多数人类提出的自然语言问题时，才表明机器真正具有与人类相近的理解自然语言的能力。由此可见，机器阅读理解才是实现通用的，开放领域的问答系统的关键技术。

下文将着重介绍机器阅读理解的方法。目前主流的方法有两种，分别是基于规则匹配的方法，以及基于深度学习方法。

#### 深度学习方法

深度学习自 2012 年成为计算机科学中红极一时的研究方向，研究人员也开始把神经网络应用于 NLP 领域当中，机器阅读理解从此进入神经网络时代。而经验性的结果也表明，深层网络结构确实能捕获到语句之间的深层次语义信息。如今深度学习已成为机器阅读理解中最为主流的研究方法。

机器阅读理解模型总体架构如图1.1所示。从图中可看出模型主要具备三个网络层：编码层，交互层，输出层。编码层主要用于提取上下文段落以及问题的特征向量表示。交互层用于进行文章上下文与问题之间的关联性分析。而经过交互层后，将交互层的得到的关联信息，语义联系输入到输出层中，就可以预测问题的答案了。机器阅读理解模型，在总体架构上面大同小异，而主要的区别则是编码层和交互层使用的神经网络的不同 (RNN，CNN 等….)。

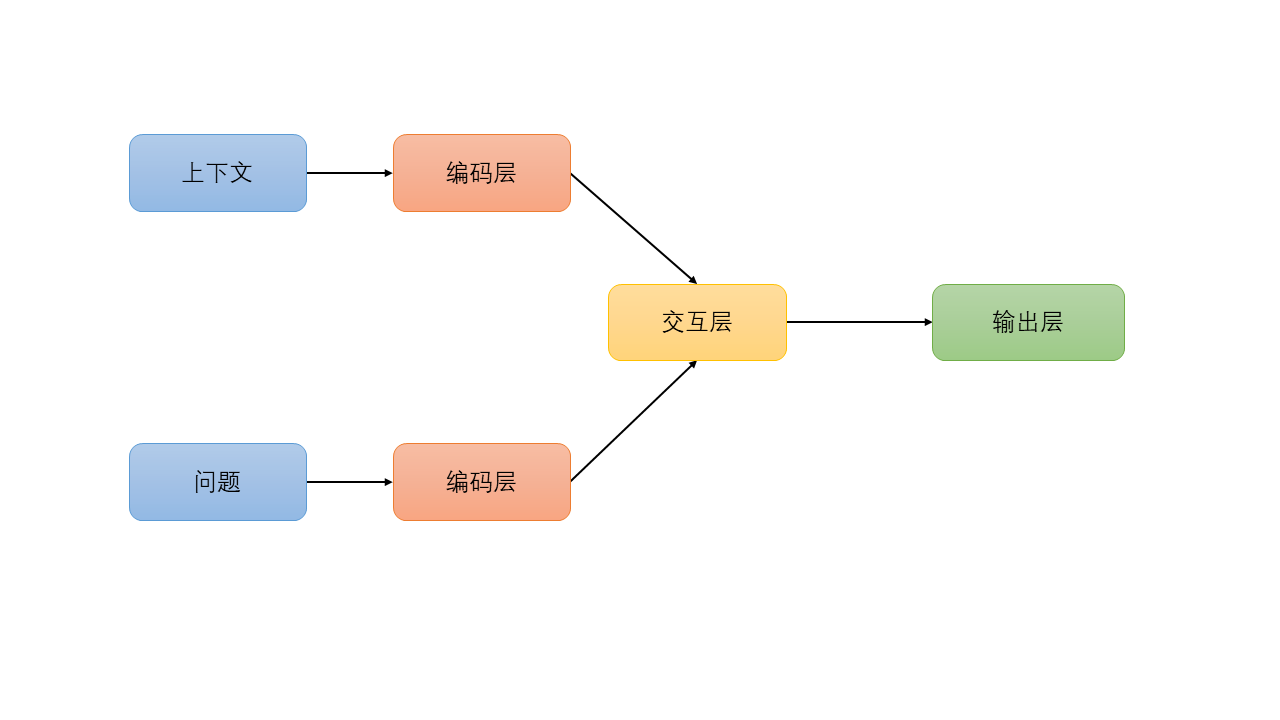


图 1.1 机器阅读模型

最早 Yu[[8](#_bookmark58)] 提出使用卷积神经网络计算统计特征向量的相似性。Chen[[9](#_bookmark59)] 使用多层循环神经网络以及远程监督在单一数据源上训练开放性领域问答系统。Yang[[10](#_bookmark60)] 提出共享值权重的 Attention­based 的网络模型，在不使用额外的统计特征，获得了更好的效果。Seo[[11](#_bookmark61)]在不同粒度 (词向量以及字向量) 的水平上使用多阶段分层双向注意力网络用于提取上下文变量编码。

2018 年，BERT[[12](#_bookmark62)] 模型的横空出世，使其迅速成为 NLP 相关任务中炙手可热的娇子。由 Google Brain 团队提出的 BERT 模型作为预训练语言模型的佼佼者，在 33 亿文本的语料库上，使用了 64 块 TPU，进行了长达 4 天时间的多个无监督训练，最终得到庞大的模型参数。BERT 模型的能够学习到自然语句之间的深层次双向表达，而且 BERT 获得的强大预训练词向量可以万能地在各个任务上进行微调。于是 BERT 模型一经问世，就在自然语言处理的各个下游任务的基准实验上获得了巨大提升，其中包括命名实体识别 (Named Entity Recognition)，情感分析(Sentiment Classification)，词性标注(Part­of­Speech Tagging)。因此，BERT 模型也被用于问答系统[[13](#_bookmark63),[14](#_bookmark64)]中。

### 小样本学习

训练深度学习模型通常需要极为大量的训练数据以支持模型收敛，但当面临训练数据量少或大量数据未标注的情况时，如何更好的训练模型便成为当下一个重要的研究议题，而小样本学习正旨在解决此类问题。

面临少样本学习中最本质的困难，即数据样本少，通过生成更多的数据进行数据增强是一项直觉的解决方法，利用数据分布的不变性即可从少量样本构建新数据。生成对抗网络[[15](#_bookmark65)] 通过对生成器和判别器进行对抗训练后，生成器可被用于构造相似于训练数据分布的数据。但 GAN 模型收敛困难和计算量大的弊端很难满足需求。Wei[[16](#_bookmark66)] 提出使用同义词替换，随机插入，随机交换，随机删除四种简单操作完成数据增强的目的。元学习(Meta Learning) 旨在让模型从大量先验任务中学会元知识，以实现对新任务更好的训练，获得更好的效果。Finn[[17](#_bookmark67)] 提出使用 MAML 模型通过元知识来进行更好的初始化模型参数。Jamal[[18](#_bookmark68)] 提出了算法避免元学习模型对训练任务过拟合的 TAML 模型。

#### 迁移学习

迁移学习是小样本学习的子领域之一。在自然语言处理中，不同的语料数据库可能拥有着不同的主题、语言、字母表、创作流派、文风等等因素。例如问答系统数据集： SQuAD[19] 源于维基百科文章数据，而 NewsQA[20] 来自CNN 社评新闻数据。通过两个语料库的数据分布不一致，从而可以得到它们来源于两个不同的领域的结论。为了客观的定义自然语言处理中的领域概念，Plank[14] 提出每一语料都属于一个潜在未知高维的多样性空间的一个样本，即任何两个不同的语料库均属于不同的领域。而在进行跨领域问答时，由于源领域和目标领域数据分布之间存在一定关联性，但测试样本分布又不完全等同于训练样本分布。为了使模型对测试样本领域数据更具鲁棒性，从训练样本分布情况迁移到新的测试样本数据分布中便成为了难点。于是迁移学习中的领域自适应技术(Domain Adaption)便可以应用于此。

为了使得不同领域分布一致化，Hal[21] 提出使用源领域与目标领域增强后的共享通用特征空间，John[22] 等人提出手动构建结构化共享特征方法。而深度学习中取得巨大成功的自动编码器也被应用来进行特征提取，即建立一个通用领域的去噪自编码器，通过无监督地学习输入数据的潜在表征，然后把提取的特征输入如 SVM 或神经网络的分类器[23,24]。

另一方面，对抗学习(Adversarial Learning)也在领域自适应中取得了不错的效果。Ganin[25] 通过增加额外的领域鉴别器与输入编码器进行对抗，减小源领域与目标领域的分布差异，此时训练得到的编码器便是一个强大的通用特征提取器。Eric[26] 等人提出使用两个不同的特征提取器，分别对抗训练，最终获得一个较好的源领域分类器和目标领域特征提取器的过程。但是利用对抗训练思路的特征提取器可能难以收敛，使得模型效果差。

## 研究问题与挑战

论文以构建一个跨领域的鲁棒性问答系统为目的，探究如何提升问答系统的泛化能力和鲁棒性。由于各问答系统数据集之间拥有不同的体裁、主题、流派等特点，导致两个语料库之间属于不同的数据分布，而跨领域训练过程已经成为了一个棘手的问题。所以构建鲁棒性问答系统主要面临以下挑战:

1. 数据领域分布不同。训练数据和测试数据样本分别来自不同语料库，在训练集上学习到的问答系统模型很难迁移到不同的测试数据分布上，导致模型效果差。
2. 微调样本数据量少。在给定少量有限，甚至缺失微调训练样本时，如何通过迁移学习策略提升模型在测试集上的效果。

目前实现的问答系统工作大多局限于同一领域、同一数据分布、充足训练数据的情况。而当此类模型面临多领域数据，少量训练样本时，却常常表现不佳。于是本文将探索如何在跨领域数据集中，构建一个泛化能力更强，更加鲁棒的问答系统。

## 论文主要工作

本论文为搭建一个更具鲁棒性的问答系统，考虑把小样本学习与问答系统两个领域进行结合。在问答系统中，利用预训练语言模型强大的表征学习能力，训练基于 BERT 模型的问答系统。另一方面，在小样本学习中，利用领域自适应策略，实现模型从训练样本数据分布迁移至测试样本数据分布。同时，考虑到可能面临缺乏微调训练样本的情况，利用数据增强对训练数据进行扩充也是尤为重要。

此论文主要完成四项工作：

1. 构建一个基于短文文本的预训练语言模型 BERT 的问答系统，命名为 RQA­BERT 模型。在此问答系统中，利用  BERT 模型提取问句和上下文中的深层语义特征，并进行编码，然后将此编码输入特征交互层中进行关联性分析，最终输入到输出层，获得预测答案。
2. 构建一个领域鉴别器，利用对抗训练的方式来对跨领域数据进行领域迁移。受到领域自适应的启发，通过引入额外的领域鉴别器，使得训练域和测试域数据分布混淆，从而能够把在训练域数据上学习得到的问答系统运用到测试域数据中。
3. 针对少量训练样本情况，利用数据增强等方法对训练样本进行扩充。数据增强后拥有更多的等价训练，微调样本，为鲁棒性问答系统的提供了更好的效果和更高的收敛性，从而降低了模型过拟合的可能。
4. 通过实验的方式探究不同模型对问答系统泛化能力以及鲁棒性的影响。在提取式问答系统的基准实验上，将 RQA­BERT 模型与传统问答系统 BiDAF 模型进行对比分析；分别针对 RQA­BERT 模型的不同子模块进行消融实验，并且分析，解释其对实验结果的影响。

## 论文组织结构

本论文共分为五个章节，每个章节内容如下:

**第1章:** 绪论。本章主要介绍论文的研究背景，然后分别从问答系统和小样本学习两个方面阐述了国内外相关研究，最终介绍本文的在构建问答系统中遇到的挑战，以及本文的主要工作。

**第2章:** 相关基础知识。对文章所使用的相关技术进行了详细介绍。

**第3章:** 基于 BERT 的鲁棒性问答系统。对鲁棒性问答系统实验进行形式化定义，并详细介绍了提出的 RQA­BERT 模型及其各个变种模块。

**第4章:** 实验结果以及分析。介绍实验所用数据集，并在提取式问答系统基准实验上对比 RQA­BERT 与其他方法的结果，通过消融实验分析各模块影响结果因素。

**第5章:** 结论与展望。最后一章对文章主要工作做出总结，并针对目前实验存在的缺陷，指出未来可能的改进工作。

# 相关基础知识

本文构建鲁棒性问答系统的过程本质上是对深度神经网络模型进行构建的过程。在本章节中，我们将从搭建神经网络的层面出发，介绍实验相关神经网络的基础原理，以及影响深度学习模型的训练过程中的不同损失函数与激活函数。

## 深度神经网络

神经网络模型是一种受到人类大脑神经元之间连接方式的启发，将多个独立的计算单元节点连接在一起组成网状结构的仿生计算模型。其最早起源于二十世纪 四十 年代，但是受限于当时计算机算力低下的原因，深层神经网络模型难以得到充分训练，导致模型效果一直不太好。而随着计算机算力的提升、反向传播算法的提出以及并行计算的运用，让训练深度网络模型逐渐变得可能。而深度学习模型强大的拟合泛化能力也在计算机视觉、自然语言处理领域中大方异彩，效果远超传统机器学习算法，一时间让深度学习成为炙手可热的研究方向。

下面介绍两种常见的神经网络层: 全连接(Dense)层以及本文实验中使用的层规范化(LayerNorm)层。

### 全连接层

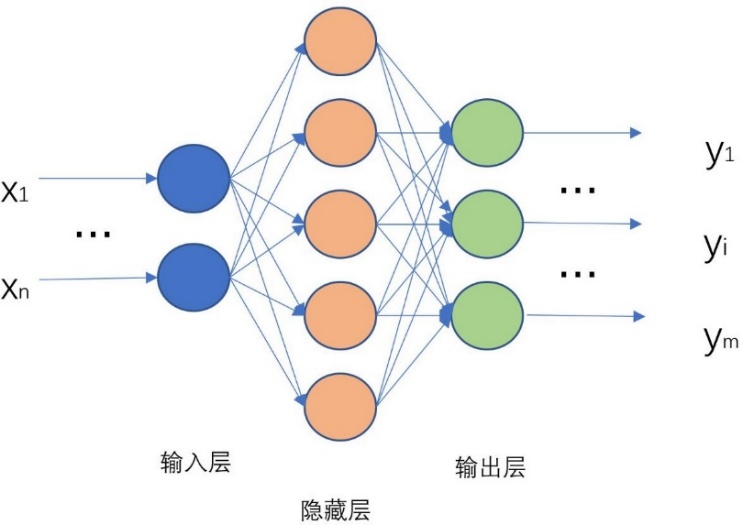
全连接网络是最简单，也是最常用的神经网络之一。其具体结构如下图[2.1](#_bookmark12)所示:

图 2.1 全连接层

顾名思义，全连接网络中每一层的所有神经元均与其前一层的所有神经元相连。每一条相连的有向边代表信息传递的方向，即前层神经元的输出，经过连接边的前馈传递，到达后层神经元且作为其输入之一。由于后层神经元与前层每一神经元均有连接关系，所以后层神经元输入为前一层所有神经元传递信息之和。

全连接层输入与输出数学关系如下:

其中 X, A 分别为前一层网络的输出向量和后一层网络的输入向量，W 是一个可学习连接权重矩阵，b 是可学习偏移向量，*σ*是激活函数，激活函数将在第[2.2](#_bookmark14)中详细介绍。

全连接的特点让相邻两层神经元之间的特征交互变得容易，足够的可学习的参数 W 和 b 也让全连接层具有强大的特征空间变换能力，所以全连接层经常充当分类器的角色。在迁移学习中，全连接网络也常常作为领域鉴别器的基础模型。

### 层规范化层

在训练数据经过层层神经网络后，可能会导致数据分布变化剧烈的情况，此时激活函数的导数在反向传播时无法稳定，可能会出现梯度消失和梯度爆炸现象，让模型训练难以继续进行。为解决上述问题，可以使用规范化层对数据分布进行归一化调整。

主流的规范化层有: 批规范化，层规范化，实例规范化等等。以下介绍实验中所使用的层规范化。

为了解决数据分布偏离，同时考虑到批规范化中存在样本数量过少的情况。层规范化的提出独立于批训练数据，针对每一批样本，计算某一网络层特征的平均值和方差，并对输入样本进行规范化。

其中 N 是网络层节点个数，*µ*和 *σ*分别是此层的特征期望和方差。最终对输入进行规范化，如下：

其中 A 是层规范化的输入向量， 是层规范化的输出，*ϵ*为一个很小的数值。

## 激活函数

激活函数给深度网络模型提供非线性映射特点。若取消激活函数，则过度的堆叠网络层也只等价于一次线性空间映射。在激活函数的支持下，理论上使用三层神经网络即可拟合任意非线性函数，由此可见激活函数的重要性。

常见的激活函数有三种: Sigmoid 函数，Tanh 函数，ReLU 函数。三者各有各自的应用点。

**Sigmoid函数:** 常用于隐藏神经元的输出。由于 Sigmoid 函数平滑和易于求导的特点，因此常用在于 01 分类任务中。其公式如下:

**Tanh函数：**与 Sigmoid 函数相似，因其取值空间在(-1, 1)之间，所以常用于样本标签为-1和1的二分类任务中。

**ReLU函数：**为了进一步缓解模型的计算复杂度，以及避免梯度消失现象，通常使用ReLU函数作为激活函数。

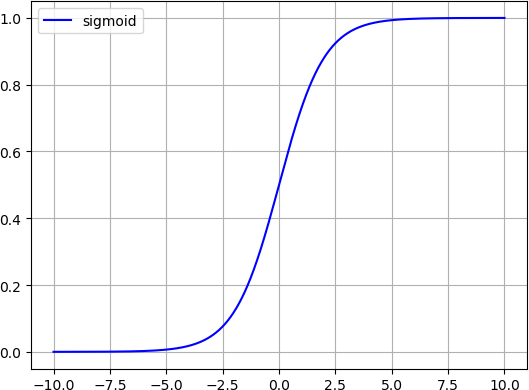
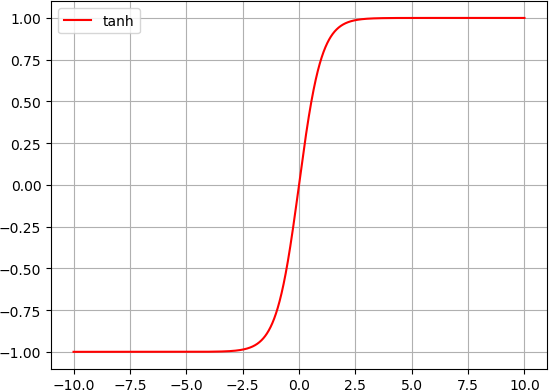


图 2.2 Sigmoid 函数 图 2.3 Tanh 函数

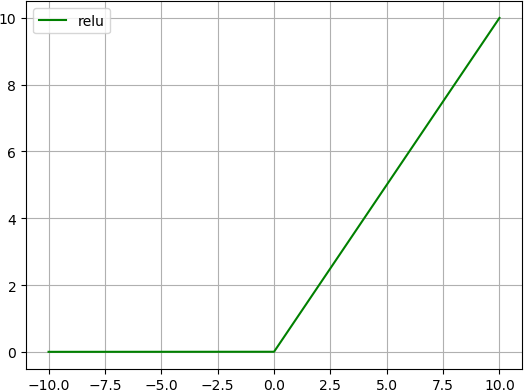


图 2.4 ReLU 函数

## 损失函数

在训练深度学习模型的过程中，损失函数被用来衡量此时模型性能的优劣。同时训练的过程也是不断降低模型预测值与训练样例真值输出之间误差的过程，模型中可训练参数的优化更新所需要的梯度也是通过反向传播损失函数的梯度获取。损失函数的选择对于模型的训练来说至关重要。一个恰当的损失函数不仅能够缩短模型的训练时间，同时能避免模型陷入局部最优点，使模型获得更好的效果。

根据任务类型分类，损失函数主要可分为回归损失和分类损失。其中常见的回归损失有均方误差损失 (Mean Square Error)，平均绝对损失 (Mean Absolute Error)，常见的分类损失有交叉熵损失 (Cross Entropy Loss)，二元交叉熵损失 (Binary Cross Entropy Loss)。

**均方误差损失函数:** 均方损失用于计算模型对输入样本的预测输出值与真值输出的平方差。

其中 N 为训练样本数量，为样本真值，为模型预测值。

**平均绝对损失函数:** 平均绝对损失用于计算模型对输入样本的预测输出值与真值输出的绝对值差。

其中 |x| 表示对 x 取绝对值操作。

**交叉熵损失函数:** 不同于回归任务需要对预测连续空间的数值，分类任务则是针对离散空间进行预测。熵的概念来源于信息论，最早被用来衡量一个系统的确定性。交叉熵则被用来衡量两个分布之间的相似性。其计算公式如下：

**二元交叉熵损失函数:** 二元交叉熵损失是交叉熵损失函数的一种特例，当且仅当进行二分类任务时，由式子[2.9](#_bookmark16)简化得到如下公式:

## 小结

本章节中主要介绍了深度神经网络的相关基础知识。从神经网络层角度出发，对实验中使用的全连接层和层规范化层进行详细介绍。以及在训练神经网络模型过程中，描述了常见的激活函数以及损失函数。上述公式将会在第[3](#_bookmark18)章构建基于 BERT 的鲁棒性问答系统中使用。

# 基于 BERT 的鲁棒性问答系统

## 实验设置

在本小节中，我们将形式化地定义鲁棒性问答系统的实验设置。

### 问答系统

实验研究对象为提取式阅读理解问答系统。首先定义训练测试数据领域相同的一般性问答系统。以 SQuAD 数据集为例，每一个样例输入具有两部分，分别是问题(Question)， 上下文(Context paragraph)。上下文和问题分别由 N，M 个来源于词汇表中的词语组成。而样例输出的答案集合是文章中一段连续的词语组成的短语或句子。

现对问答系统进行具体数学定义：

由于问答系统的答案为短文中连续短语，因此只需要对答案边界进行预测即可。问答系统模型需要学习一个特征提取函数 (Feature Extractor, 定义其函数为 F) 和一个分类器 Classifier(定义其函数为 CLF)，以预测答案边界, 即开始位置与结束位置：

其中为模型在上下文中的提取出的隐藏特征向量。

### 鲁棒性问答系统

在上述问答系统中，假定了训练数据集  与测试数据集  数据样本分布一致。而实现一个鲁棒性的问答系统，即  和  来源于不同领域时的问答系统，正是本文所研究的内容。我们在源数据领域上对训练样本的验证集  进行测试，其中  与  属于同一数据分布。其次，为探究小样本学习对问答系统的影响，我们分别针对模型有无域外数据微调分别进行了测试。

**有域外微调数据:** 为了测试模型的泛化性能，实验假设拥有一部分与测试数据集  数据领域分布相同的微调数据集 ，但微调数据集样本个数远远小于训练数据集样本个数，即 。实验通过小样本学习的方法对微调数据集进行一系列操作，比如：数据增强，领域对抗，元学习等方法，使得源领域数据模型能够从  中获取更多于  相关数据分布的情况，从而达到提升实验最终在测试数据集  上的效果。

**无域外微调数据:** 为了直观感受不同模型之间的鲁棒性能的差异，实验设置了另外一组无域外微调数据的对照实验，旨在验证问答系统在一个从未见过的数据集上的表现。此时并没有微调数据集 ，在训练数据集  上训练得到的问答系统直接在测试数据集  进行测试。

## 实验过程

在此章节中，我们搭建了一个完整的鲁棒性问答系统 RQA-BERT，并且分别构建了三个变种问答系统模块，分别是：领域对抗(domain adversarial) 模块，数据增强  (data augmentation) 模块，以及长度惩罚损失 (length penalty loss) 模块。

图3.1为 RQA-BERT 模型框架，下文将详细阐述模型实现细节。

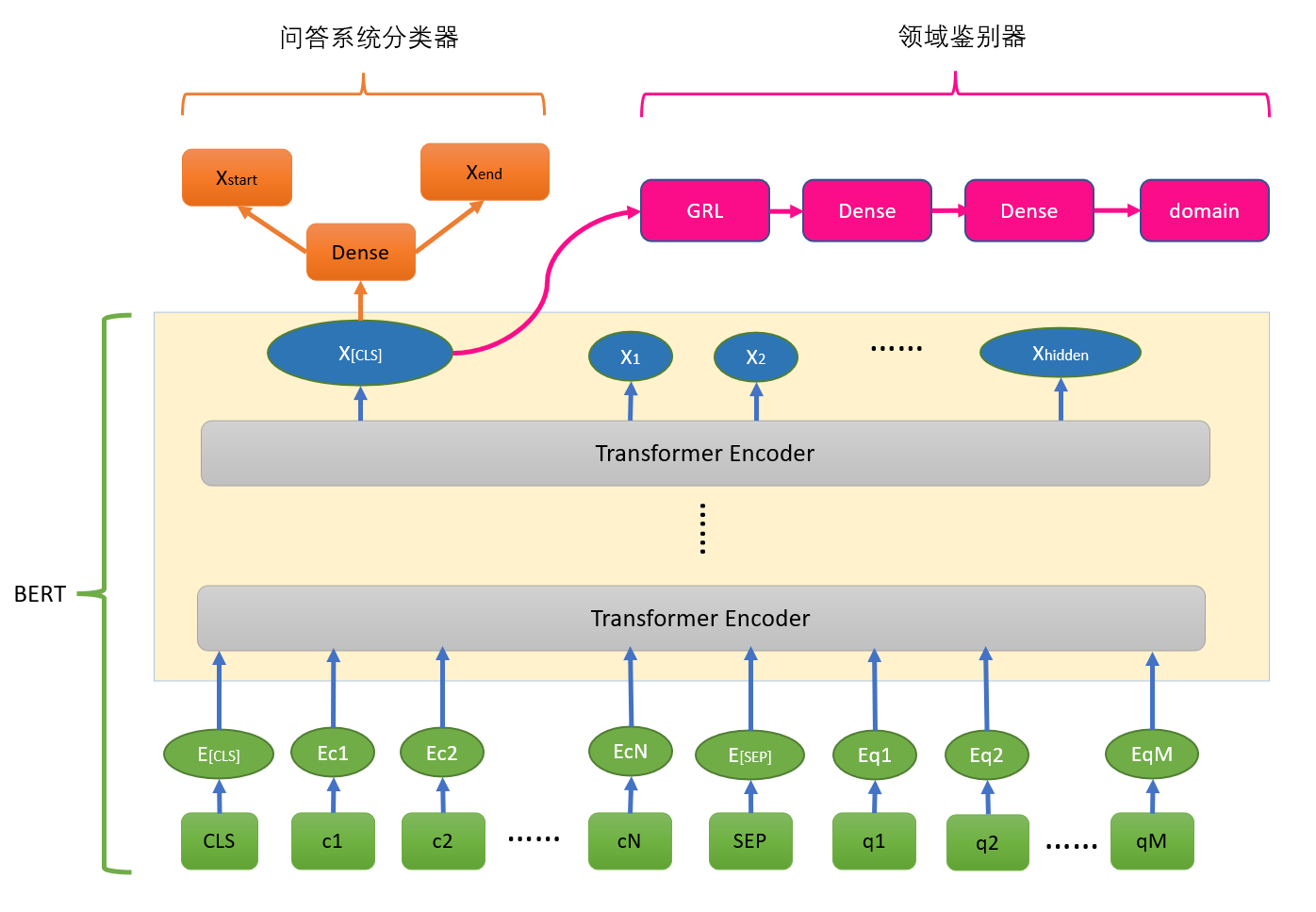


图 3.1 RQA-BERT 模型

### 基于预训练语言模型的问答系统

在大规模语料库上训练的预训练语言模型已经在NLP的多个领域取得了较为出色的结果。而 BERT 模型利用 Transformer 的双向编码表示更是能够精准的捕获到语句中的深层次语义。同时得益于 Transformer 其架构的优势，相较于RNN 网络层训练困难和不可并行性，每一层 Transformer 模型使用多头注意力机制和残差全连接层，极大程度提升了 BERT 模型的性能和可行性。

在 BERT 模型之后，研究人员又相继推出如 ALBERT[[27](#_bookmark77)]，Roberta[[28](#_bookmark78)]，GPT-3[[29](#_bookmark79)] 等等新的预训练变种语言模型。在此实验中，为了同时考虑到时间效率和计算代价的原因，最终使用了经过知识蒸馏的 BERT 模型，即 DistilBERT[[30](#_bookmark80)]。

DistilBERT 利用知识蒸馏技术，对 BERT-base 模型进行模型压缩，在几乎持平 BERT 模型效果的情况下，最终得到一个推理效果更快，模型参数更小，计算开销更少的预训练模型。关于不同预训练模型对实验效果的影响，将在第五章中展开。

#### 将BERT模型用于下游任务问答系统中

BERT[12]模型在大规模未标记的文本上进行MLM(Masked Language Model) 和NSP(Next Sentence Prediction) 预训练任务，已经获得大量自然语言文本深度的双向表达。所以，对于一项新颖的任务，只需要在 BERT 的输出层后加入一层额外的网络，以调整输出结构，便可以广泛应用到 NLP 的下游任务当中。如问答系统，命名实体识别，语言推理等。

#### 输入标记化以及阶段

对于一个问答系统输入有二元对(上下文Context, 问题Question)其中分别由 N 个和 M个词语组成。为了符合 BERT 模型的输入，分别将上下文和问题标记化，并构建标记化的词索引，把连接后的上下文与问题词索引作为 BERT 模型的输入。需要注意的是，这里合并后的词标记可能会超出 BERT 模型输入维度上限，所以需要对超出长度的词索引进行截短。

#### 嵌入层

在嵌入层中，需要进一步把词索引转化为可以被操作的词向量。BERT 模型的向量嵌入层由三个部分组成：词向量嵌入(Token Embedding)，段嵌入(Segment Embedding)， 位置嵌入(Position Embedding)。

* + - * 1. 词向量嵌入：将词索引转换为固定维度的向量。
        2. 段嵌入：由 tokenizer 自动生成，用于识别句子间界限。
        3. 位置嵌入：为 Transformer 架构提供词嵌入向量在原语句中的位置编码信息。

最终 BERT 嵌入层输出为上述三者求和:

#### Transformer架构

BERT 中用来提取特征信息的核心模块来自于 Transformer 架构中的编码器。每一层网络结构中都含有完全相同的 Transformer 块并联组成，且同一层 Transformer 块之间信息通过多头注意力机制双向流通。其主要结构如下图所示：

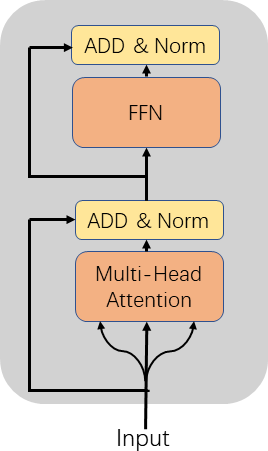


图 3.2 Transformer 层

为了提升模型并行性，Transformer 模型使用注意力机制取代了传统文本信息提取使用的循环神经网络。为了更好的对语句中深度多层语义进行提取和保留，Transformer 使用了多头注意力机制(Multi-head Attention)。

约定 是来自 层的 Transformer 的输出，是经过多头注意力机制后的中间层输出。则有：

其中有 *,* ，代表 Transformer 每层输入以及输出维度。, 是可训练二维权重矩阵参数，用于获取 的注意力投影 Query，Key，Value。代表注意力头数个数。是可训练二维权重矩阵参数，用于把拼接后的各子注意力头投影到与 Transformer 输入维度相同。

上述公式等价于：

所以整个 Transformer 架构如下:

其中 是可训练二维权重矩阵参数。

在本文的实验设置中，DistilBERT 模型使用了6 层Transformer 编码器，= 12*,* = 768*,* = = 64。

### 对抗训练

当训练数据与测试数据属于不同数据分布时，可视为训练数据与测试数据分别来自于两个不同领域，其中训练数据拥有大量带标签数据，而测试数据则拥有少量标签数据或拥有大量未标签数据。在迁移学习中，领域对抗训练启发了我们的工作。

最早将域对抗训练引入到迁移学习的工作是 DaNN[[25](#_bookmark75)]。此论文的中心思想是当源领域数据和目标领域数据分布相同时，即可把源领域数据训练得到的任务分类器用于目标领域数据分类上。所以 Ganin 提出同时训练两个分类器：1. 利用训练域数据训练任务分类器；对源领域任务进行分类。2. 利用训练域数据和测试域数据联合训练一个领域鉴别器 (Domain Discriminator)，对数据的所属领域进行分类。此做法通过让编码器与额外的领域鉴别器对抗训练，以消除训练和测试数据的分布差异，使得编码器 (在本实验中即是 BERT 模型) 把不同领域数据分布映射到同一个空间分布中，从而能够把在训练数据的上得到的源领域任务分类器用于目标领域的任务分类中。

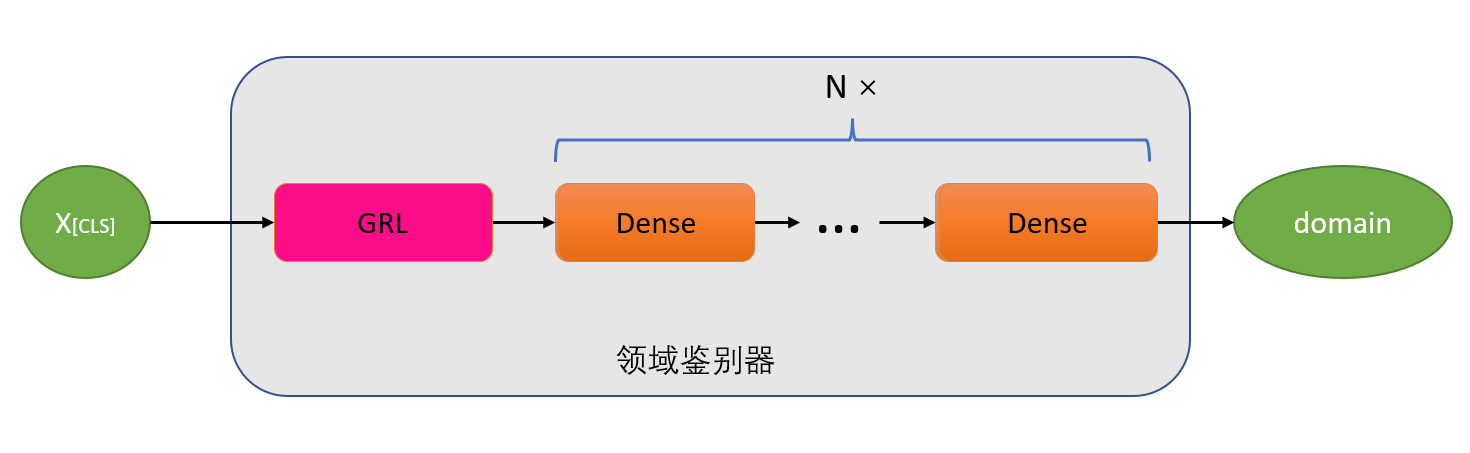


图 3.3 领域鉴别器

考虑到端对端训练的方便简洁性，实验中最终选取使用神经网络作为领域鉴别器架 构。同时，类似于对文本进行情感分类任务，利用了 BERT 模型输出的第一个 CLS 隐藏向量作为领域鉴别器的输入。领域鉴别器具体架构如图3.3所示：

其中为 BERT 模型输出的第一维向量。GRL，Dense分别代表梯度反转层(Gradient Reversal Layer) 以及全连接层。domain 表示此领域鉴别器预测结果。

#### 梯度反转层

为了使得 BERT 模型混淆两个领域数据分布特征，需要使 BERT 模型与领域鉴别器进行对抗训练，即让两者训练目标对抗。梯度反转层的作用在于让梯度在反向传播的过程中通过取反操作，使得梯度反转层后面网络梯度反向传播时混淆梯度反转层前面的网络梯度， 从而同时达到对抗训练两个模块的目的。具体实现如下:

其中 为梯度反转层接受后序层反向传播时提供的输入梯度，为梯度反转层输出梯度，为梯度反转超参数，实验中设置为 -1。

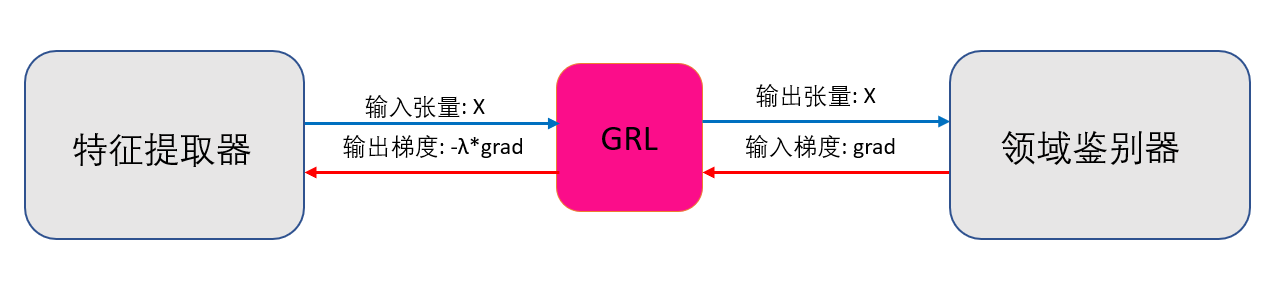


图 3.4 梯度反转层

#### 全连接层

为了简洁性的考虑，领域鉴别器使用全连接神经网络对输入数据的领域进行预测，以及把最终输出向量映射到 2 维。在具体实验设定中，考虑到 输入维度为 768，使用了两层隐藏全连接层的神经网络，分别为设置为 256 维度和 128 维度，并把输出层维度设置为 2。关于全连接层数对实验结果的影响将在第4.4节进行分析。

此处领域鉴别器的训练目标函数如下：

其中 为对抗训练损失比例系数。分别代表领域鉴别器预测输出结果与输入数据领域真值结果。

### 数据增强

深度学习已经广泛应用到 NLP 领域当中，但是在缺乏充足的训练数据时却难以取得令人满意的结果。当面临数据收集困难、数据质量不高、数据缺乏标签的时候，数据增强即可利用现有的数据，人为地生成更多带标签数据。需要注意的是，在问答系统任务中，我们只能对上下文和问题进行数据增强，而不能对答案进行。也正因如此，数据增强技术同样可以用于无标签的域外数据。

实验采用了 Wei[[16](#_bookmark66)] 提出了针对文本的四种增强技术。分别是：

1. 同义词替换 **(SR)**：从句子中随机选取非结束词的n个词 ，并用 的同义词字典随机词在原句对应位置中替换 。
2. 随机插入 **(RI)**：随机选取一个非结束词 ，并将其同义词插入原句任意位置。
3. 随机交换 **(RS)**：在句子中随机选取两个词，并交换两者的位置。重复 n 次。
4. 随机删除 **(RD)**：以 p (超参数，默认设为 0.1) 的概率随机删除句子中的词语。

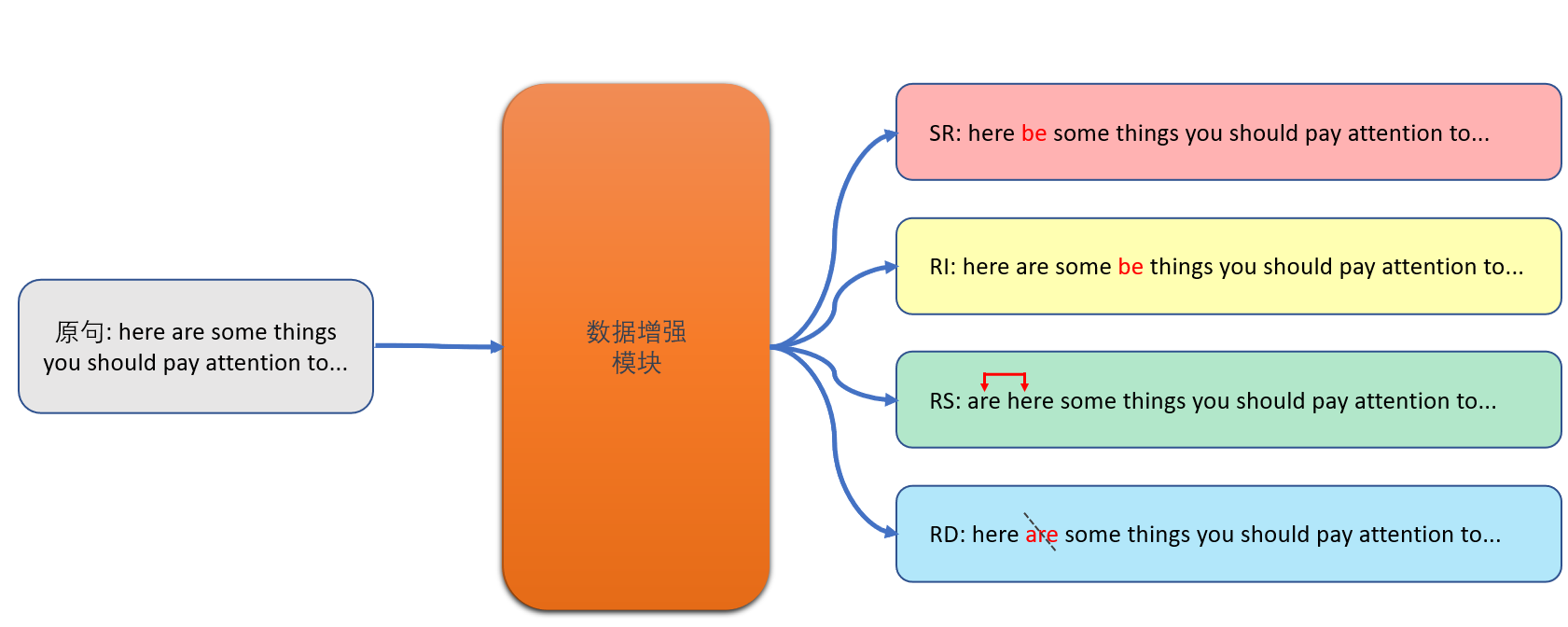


图 3.5 预处理中的数据增强

### 长度惩罚损失

除上述模型架构外，为了尝试进一步提升模型在测试数据集上的评估指标，实验对当前模型在域外数据验证集上进行了错误分析。通过在 RACE-S 验证集上的结果分析，对比了模型预测结果与样例真值之间的差异，最终选择了以下两个样例作为典型错误样例。

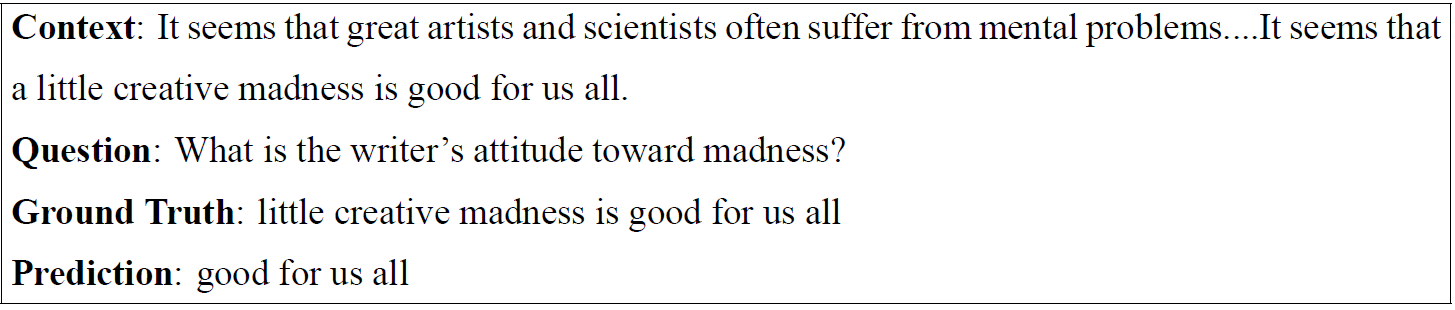


图 3.6.1 错误样例1

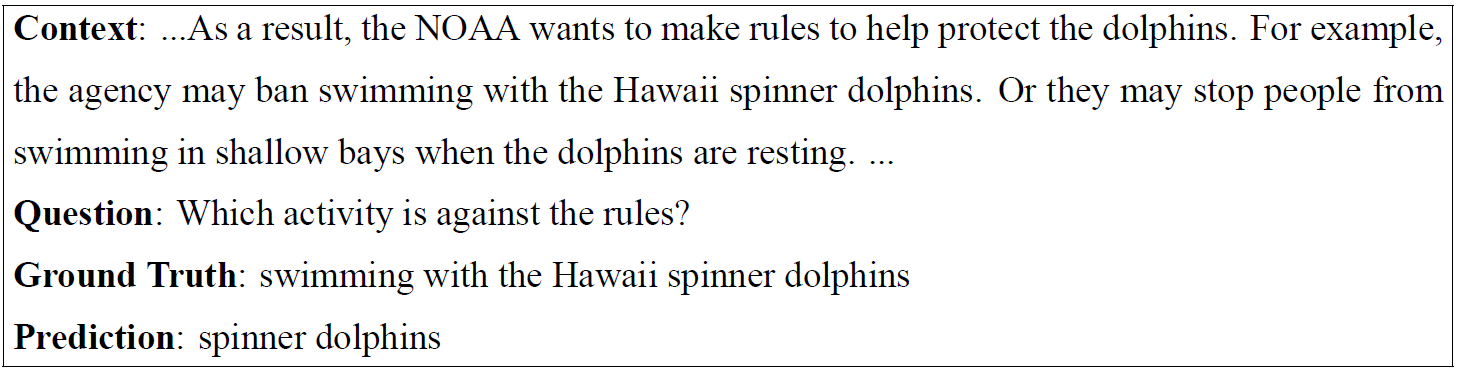


图 3.6.2 错误样例2

通过观察以上样例，可以看出由模型预测出的结果不完整，即是答案中的开始位置和结束位置出现一定距离的偏差。同时，模型输出答案的预测结果相对真值更短。

因此，为了进一步减少由于答案长度偏差以及部分样本答案过短的情况，提出了使用额外的长度损失项对模型过短输出加以惩罚，以鼓励更长的答案输出。

1. 控制答案长度偏差的损失

其中 代表模型预测开始位置与结束位置, 代表测试数据的真实开始位置与结束位置。为了避免同时对起始位置，结束位置以及长度记录损失，此处简化为对起始位置，结束位置以及中间位置的损失。

2. 控制答案长度过短的损失

其中 是长度损失比例系数， 表示若中间项小于零，则取零；若中间项大于等于零，则取原值。

## 小结

在本章中，首先对鲁棒性问答系统的实验具体输入输出进行形式化定义，以及提出针对有微调数据的小样本学习和无微调数据的零样本学习的两种不同的实验设置。然后文章提出基于预训练语言模型的问答系统 RQA-BERT 模型，并构建额外的三个变种问答系统模块: 领域对抗训练模块，数据增强模块，长度惩罚损失模块。文中对各个子模块的实现细节进行了具体的描述。

# 实验结果以及分析

## 实验准备

### 实验数据集

在本实验中需要两个不同领域的问答系统数据集。选择了使用 SQuAD[19] 作为训练领域数据集，RACE[20] 作为测试领域数据集。SQuAD 和 RACE 都是提取式问答系统常用的基准数据集，但是两者在文章来源 (分别来源于维基百科与中国中考、高考阅读理解)， 文章体裁 (记叙文或是议论文)，问答类型 (检索文章或是知识推理) 等方面具有较大不同。与此同时，考虑到实际应用中领域外数据缺失的缘故，我们从RACE中随机选取了128篇文章中的128个问题构造了一个小型RACE的子数据集，称为RACE-S(RACE-SMALL 的缩写)。

各个数据集具体信息如下：

表 4.1 各数据集中样本划分数量

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | #train | #valid | #test |
| SQuAD | 50000 | 10507 | N/A |
| RACE | 87866 | 4887 | 4934 |
| RACE-S | 128 | 128 | 419 |

### 评估指标

本实验选择沿用Rajpurkar[19] 提出的两种提取式问答系统基准实验的评估指标： F1 分数 (F1 score) 以及完全匹配值 (Exact Match Score)。此两个评估指标得分越高表明模型效果更好。

**F1 分数**：将预测结果与真值结果切分为词，统计相同的词语数量，通过精准率和召回率计算 F1 分数。

其中， 分别是指预测词语数量和真值词语数量， 是预测与真值结果均拥有的词语数量。

**完全匹配值**：要求预测结果与真值结果逐词逐句匹配，并统计整个测试样本集完全匹配比例。

其中 是完全匹配答案样本数量， 是预测样本总数量。

### 对比实验

为了体现 RQA-BERT 模型更强的鲁棒性，我们使用 BiDAF[[31](#_bookmark81)] 作为基线模型与 RQA- BERT 模型的结果进行对比。BiDAF 是一个同时计算字符层面，单词层面的上下文嵌入，输入双向 LSTM 模型进行编码，来获得一个上下文-答案相互感知的问答系统模型。

4.1.4 实验设置

实验使用三张 12G 显存的 Tesla P100 显卡对问答系统模型进行训练。预训练 DistilBERT 模型[30] 使用 6 层 Transformer 层用于提取特征。领域分类器使用 3 层全连接网络层， 并且每层全连接层有p=0.1 的概率被设置 Dropout[32]。实验过程中，使用 AdamW[33] 优化器对神经网络参数进行更新。训练过程中设置批大小为 16，学习率为 3e-5。对于 SQuAD[19] 与 RACE[20] 数据集中过长的文本需要在输入到 BERT 模型时进行截断，其中设置最大文本长度为 384。

## 实验结果

实验对比了 RQA-BERT 及其变种模型与基准问答系统模型 BiDAF 的实验结果，以实证的方式展示了 RQA-BERT 模型在跨领域问答系统拥有更强的鲁棒性。

表[4.2](#_bookmark35)为各个模型在 SQuAD 源领域数据集上训练，在域内数据集 SQuAD 和域外数据集 RACE-S 的验证集上的评估指标得分。

表[4.3](#_bookmark36)为各个模型在 SQuAD 源领域数据集上训练，并在域外数据集 RACE-S 上微调后， 在域外数据集 RACE-S 的验证集上的评估指标得分。

在表[4.2](#_bookmark35)中陈列出基准模型 BiDAF 和5个 RQA-BERT 及其变种模型的实验效果。更多关于 RQA-BERT 各个子模块对整个模型效果的具体分析，将会在第[4.3](#_bookmark38)节的消融实验中详细描述。单一 RQA-BERT 模型能够在 RACE-S 上取得 14.06 的完全匹配值以及29.07 的 F1 分数。除开在源领域 SQuAD 验证集上的完全匹配值，所有变种模型均在其他评价指标上均超过了 BiDAF 模型。同时，从表 [4.3](#_bookmark36) 中可以看出，在经过少量测试领域样本的微调情况下，RQA-BERT 及其变种模型在 F1 分数以及完全匹配值上显著超过了 BiDAF 模型。在同时使用对抗训练、数据增强、长度惩罚的集合 RQA-BERT 模型中，取得了 22.05 的完全匹配值以及 32.31 的 F1 分数，在 BiDAF 的基准上分别提升了 8.77 与 7.57。

表 4.2 各项模型在 SQuAD 和 RACE-S 数据集上的验证结果。adv 表示加入对抗训练; DA(SQuAD) 表示使用数据增强后 SQuAD 作为训练集; LP 表示增加长度惩罚项

|  |  |  |
| --- | --- | --- |
| F1 / EM | SQuAD | RACE-S |
| **Train on SQuAD** |  |  |
| BiDAF | 75.17/**64.54** | 21.21/10.94 |
| RQA-BERT | 76.44/61.35 | 29.07/14.06 |
| RQA-BERT + adv | 75.25/59.64 | 29.72/**15.63** |
| RQA-BERT + DA(SQuAD) | **76.70**/62.05 | 26.92/13.28 |
| RQA-BERT + LP | 76.59/61.59 | 28.03/13.28 |
| RQA-BERT + adv + LP | 75.97/61.49 | **30.33**/13.28 |

表 4.3 各项模型在 RACE-S 数据集上的验证结果。† 表示模型在 RACE-S 上经过微调; adv 表示加入对抗训练; DA (SQuAD) 表示使用数据增强后 SQuAD 作为训练集; LP 表示增加长度惩罚项

|  |  |
| --- | --- |
| F1 / EM | RACE-S |
| **Train on SQuAD & Finetune on RACE­S** |  |
| BiDAF | 24.74/13.28 |
| RQA-BERT† | 29.25/16.41 |
| RQA-BERT† + adv | 30.05/17.97 |
| RQA-BERT† + DA(SQuAD) | 27.11/14.06 |
| RQA-BERT† + DA(RACE-S) | 32.29/17.97 |
| RQA-BERT† + LP | 28.75/15.62 |
| RQA-BERT† + LP | 28.75/15.62 |
| RQA-BERT† + adv + LP | **34.08**/17.19 |
| RQA-BERT† + adv + LP + DA(RACE-S) | 32.31/**22.05** |

### 预训练语言模型对实验结果影响

在表 [4.2](#_bookmark35)、表 [4.3](#_bookmark36)中，以实验结果的方式证实预训练语言模型能够提供问答系统更强的鲁棒性。值得注意的是，即使单一 RQA-BERT 模型在目标领域的两个评价指标上均超过了 BiDAF 模型，但是在源训练数据集上的完全匹配值上却有稍微的下降 (64.54 降低到 61.35)。为了解释此反常现象，我们分析了预训练 RQA-BERT 模型与传统问答系统 BiDAF 模型架构的差异，最终得到了以下结论：由于能够对问题–上下文之间进行双向注意力流感知，BiDAF 模型能够对同一数据领域的答案边界更为敏感，因此能够取得更高的完全匹配值。而 BERT 模型由于进行了大规模的文本预训练，使得模型能够更容易的获得文本中的深层次语义信息，让 BERT 模型具备更强大的鲁棒性(体现在域内数据的 F1 指标上) 和泛化能力 (体现在域外数据的 F1 和 EM 指标上)。

## 消融实验

为了进一步验证第[3.2](#_bookmark22)节中介绍的各个子模块在集合 RQA-BERT 中的所作贡献，本小节将分别从三个方面探究不同的 RQA-BERT 变种模块对集合模型最终实验结果造成的影响。

表 4.4 未经过微调模型的消融实验

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| RQA-BERT | adv | LP | DA(SQuAD) | SQuAD  (∆F1/∆EM) | RACE-S  (∆F1/∆EM) |
| ✓ |  |  |  | ±0/±0 | ±0/±0 |
| ✓ | ✓ |  |  | -1.19/-1.71 | +0.70/+1.57 |
| ✓ | ✓ | ✓ |  | -0.74/+0.14 | +1.26/-0.78 |
|  |  |  | ✓ | +0.36/+0.70 | -1.04/-0.78 |

表 4.5 经过微调后模型的消融实验

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| RQA-BERT† | adv | LP | DA(RACE-S) | RACE-S(∆F1/∆EM) |
| ✓ |  |  |  | ±0/±0 |
| ✓ | ✓ |  |  | +6.51/+3.49 |
| ✓ | ✓ | ✓ |  | +10.64/+2.51 |
| ✓ | ✓ | ✓ | ✓ | +8.77/+7.57 |

### 对抗训练对实验结果影响

根据在表 [4.4](#_bookmark39)中结果显示，在无域外微调数据情况下，对 RQA-BERT 模型中加入对抗训练策略使得最终结果在 RACE-S 数据集上的 F1、EM 指标上分别提升了 0.70 和 1.57。但是在 SQuAD 验证集上 F1、EM 指标却下降了 1.19 和 1.71。导致此现象的一个重要原因来自于领域分类器。一个良好的领域分类器使得两个不同的数据领域进行融合的同时，将不可避免地降低特征提取器(也就是这里的BERT 模型) 对源数据领域的特征敏感性。也就是说，特征提取器对单一领域的特征敏感性于对多领域的特征泛化性不能同时具备，因为对抗训练过程本质就是让这两者进行对抗。

### 数据增强对实验结果影响

在本实验设置中，分别对 SQuAD 和 RACE-S 数据集进行了数据增强。其中在无域外微调数据的情况下，对在源数据上对 SQuAD 训练数据增强可得以下结论：F1、EM 指标在SQuAD 数据集上的分别提升了 0.34 和 0.70。但是在 RACE-S 验证集却下降了 2.95 和 0.76。此结果可以证明，对源领域进行数据增强只能增强源领域数据的评估指标，却导致在跨领域数据上的评估指标下降。另一方面，在有域外微调数据集的情况下，对 RACE-S 微调集进行数据增强后，RQA-BERT 模型在扩充后的 RACE-S 训练集上进行微调训练，让 F1 和 EM 分别在 RACE-S 验证集上获得了 3.04 和 1.56 较为明显的提升。由此可见，在只拥有少量域外数据样本时，使用数据增强能够显著地提升跨领域模型在域外数据上的实验结果。

### 长度惩罚对实验结果影响

为了使模型输出更长的答案，可以通过额外给模型输出过短的答案以长度惩罚。从 表[4.4](#_bookmark39)、表[4.5](#_bookmark40)结果得出，单独增加额外的长度惩罚项并不一定会让模型的结果变得更好 (F1 分数与 EM 值均可能上升或下降)。分析其原因可能是 SQuAD 数据集与 RACE-S 数据集答案长度存在不同程度的差异导致的，所以在 SQuAD 数据集上训练得到的长度惩罚损失不能在少样本的情况下迁移到RACE-S 数据集上；其次原因可能来自于为了控制答案偏差而提出的很难控制，因为答案边界中的起始位置与结束位置并没有一个强相关特征可供学习。所以上述两个原因导致长度惩罚对实现结果影响不大。但当长度损失与对抗训练和数据增强模块一起使用的时候，此集合模型能够相比所有其他模型在 RACE-S 数据集上获得至少 4.08 的完全匹配值的显著提升。

## 参数敏感性实验

为进一步探索模型的性能上限，本节对实验中所使用的超参数空间进行搜索。在对抗训练和长度惩罚项中，主要影响结果的超参数有全连接层数量和惩罚长度下限。

全连接层数量**N:** 领域鉴别器本质是一个分类器模型，任何机器学习分类算法均可充当它的角色，但是使用全连接层可以保证神经网络模型端对端地训练，而领域鉴别器中全连接层的数量将决定领域分类器的优劣，从而进一步影响训练与测试领域数据混淆的好坏。实验使用了 的六种设置，其中每层全连接隐藏层为 768 维度，并探究其对模型结果的影响。图[4.1](#_bookmark45)为在测试集上的表现。

从图[4.1](#_bookmark45)可以看出，当全连接层数量从 3 层增加到 6 层时，领域鉴别器分类能力逐渐强大，模型的性能不断提升。但当全连接层数超过 6 层时，模型性能却逐渐下降。

领域鉴别器随着全连接层数量的增加不断拥有更多的参数以及更强大的拟合能力，但是在缺少微调训练样本的情况下，很容易出现了过拟合现象。过少的层数让领域鉴别器性能不佳，而过多的参数却导致不能得到充分的训练，于是最终全连接层数设置为 6 层的时候模型效果最优。

惩罚长度**length K:**实验中通过增加长度惩罚项以限制过短答案的输出。对于答案长度下限的阈值，则使用超参数 K 进行设置。通过五种不同 的设置对模型进行训练并微调。图[4.2](#_bookmark45)展示不同超参数在测试集上的表现。

当惩罚长度过低时，模型在 F1 和 EM 上评估指标表现较差。当设置惩罚长度 6 和 8 时，模型效果达到峰值。随着惩罚长度的继续增加，模型更倾向于输出更长的答案，导致模型的评价指标开始下降。在对惩罚长度参数进行搜索的时候，我们发现无论如何调节 K 值大小，都很难获得巨大的提升，其具体原因也在第[4.3.3](#_bookmark43)节中进行了阐述。

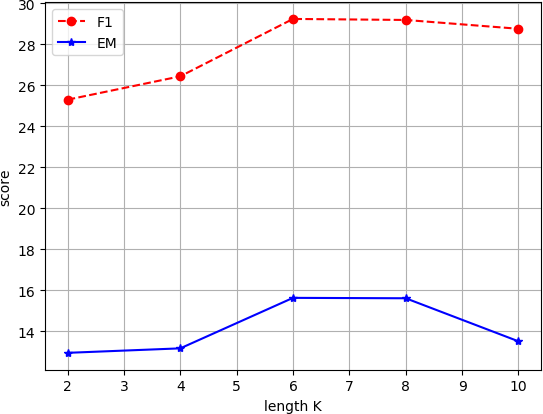
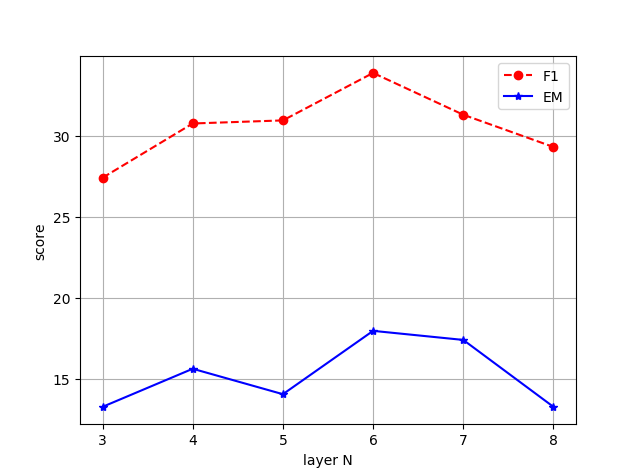


图 4.1 全连接层数 N 对实验结果影响 图 4.2 惩罚长度 K 对实验结果影响

## 小结

本章首先介绍了实验使用的两个数据集，并对数据集进行划分，采样。同时从 RACE 数据集中随机选取样本，构造一个 RACE 的子数据集称为 RACE-S。然后介绍实验中所使用的衡量矩阵、超参数设置以及对比方法中的基准实验。实验分别在有无微调数据的两种设置下进行，并对各自结果进行分析。最终从多方面对 RQA-BERT 集合模型进行了研究，分别对对抗训练、数据增强、长度惩罚三个子模块进行了消融实验，更为深入的探索 RQA-BERT 模型各模块对实验结果的影响。

# 结论与展望

在本论文中，我们以构建一个鲁棒性的问答系统作为研究目标，分别从问答系统和小样本学习两个研究方向出发，在使用预训练语言模型 DistilBERT 的基础上，针对问答系统的下游任务进行了改进，提出了 RQA-BERT 模型。同时，受到迁移学习中领域自适应策略的启发，提出了使用对抗训练技术对跨领域数据进行领域迁移学习。此外，为了应对域外测试样本数量少，模型微调效果差的情况，实验额外使用了针对自然语言的文本数据增强技术以及长度惩罚项。最终以实验的方式验证了 RQA-BERT 及其变种模型的可行性，在以 SQuAD 数据集作为训练领域数据集，RACE 数据集作为测试领域数据集的实验设置下，论文提出的 RQA-BERT 集合模型在 F1 分数以及完全匹配值两个评价指标上分别超过基于传统问答系统基线模型BiDAF 以7.57和8.77。

## 工作总结

以下为对本文主要工作的总结：

1. 通过构建一个鲁棒性问答系统为目的，把问答系统与小样本学习两个研究方向结合。提出一个相对传统单一领域问答系统拥有更强的鲁棒性的跨领域问答系统。
2. 提出 RQA-BERT 模型以及针对跨领域数据训练的三个策略 (对抗训练、数据增强、长度惩罚项)。详细地介绍构造鲁棒性问答系统中的相关技术，利用语言预训练模型构建问答系统；增加领域分类器，并与 BERT 模型进行对抗训练；分别对源领域数据与目标领域数据进行数据增强；增加额外的长度惩罚项以进一步提升模型效果。最终以实证的方式表明，本文提出的集合模型在跨领域基准实验上超过传统问答模型。
3. 进行消融实验分析 RQA-BERT 各个模块对实验最终结果的影响。实验结果显示，对抗训练策略和数据增强能够明显提升模型的鲁棒性，而长度惩罚项对模型的性能的影响并不显著。

## 工作展望

本文完成了对鲁棒性问答系统的基本构建，碍于显卡计算资源缺乏的缘故，所设计的鲁棒性问答系统还存在不足之处，仍需要在未来的工作中不断进行完善。以下为本文将来仍需改进之处：

1. 实验中超参数的设置缺乏一定的理论依据。受到硬件条件的影响，实验过程对模型训练批大小、学习率的设定过于绝对，可能无法找到模型最优解。因此，未来可以考虑使用网格搜索策略对更高维度超参数空间进行进一步搜索。
2. 元学习技术应用的可能性。元学习通过设计出一种拥有强大泛化性的模型，使得即使在少量学习样本的情况下，依然能够保证模型快速在不同数据分布之间进行领域迁移，从而更好的适应、更快的泛化到一个新的数据分布当中。
3. 训练、测试数据集来源单一，跨领域数据特征分散。在对数据进行预处理时，由于服务器性能限制，无法对更大批次的数据进行同时处理，只能在域内、域外选取出一个数据集作为训练、测试领域。如果能够把多个数据集融合到一起，同时作为 RQA-BERT 模型的输入，让模型学习更多来源的数据分布，最终结果可能在测试集上表现会更好。
4. 使用推理能力更强的预训练语言模型。本实验选取 DistilBERT 模型作为 RQA-BERT 模型的骨架网络，但碍于网络模型参数数量限制，使得 DistilBERT 模型推理能力有限，模型效果在域外数据集上表现仍然不佳。未来将考虑使用更为强大的预训练语言模型，以更新颖的预训练方式，更庞大的网络参数，更强大的推理能力让跨领域学习的模型具备更强的鲁棒性。

# **参考文献**

1. HASSAN H, AUE A, CHEN C, et al. Achieving human parity on automatic chinese to English news translation[J]. arXiv preprint arXiv:1803.05567, 2018.
2. JIA R, LIANG P. Adversarial examples for evaluating reading comprehension systems[J]. arXiv preprint arXiv:1707.07328, 2017.
3. TALMOR A, BERANT J. MultiQA: An empirical investigation of generalization and transfer in reading comprehension[J]. arXiv preprint arXiv:1905.13453, 2019.
4. CHOWDHURY G G. Introduction to modern information retrieval[M]. [S.l.] : Facet pub-lishing, 2010.
5. BURGES C J. Towards the machine comprehension of text: An essay[J]. TechReport: MSR- TR-2013-125, 2013.
6. SIMMONS R F, KLEIN S, MCCONLOGUE K. Indexing and dependency logic for answer ing English questions[J]. American Documentation, 1964, 15(3) : 196 – 204.
7. KUPIEC J. MURAX: A robust linguistic approach for question answering using an online encyclopedia[A]. Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval. 1993 : 181 – 190.
8. YU L, HERMANN K M, BLUNSOM P, et al. Deep learning for answer sentence selection[J]. arXiv preprint arXiv:1412.1632, 2014.
9. CHEN D, FISCH A, WESTON J, et al. Reading wikipedia to answer open-domain questions[J]. arXiv preprint arXiv:1704.00051, 2017.
10. YANG L, AI Q, GUO J, et al. anmm: Ranking short answer texts with attention-based neural matching model[A]. Proceedings of the 25th ACM international on conference on information and knowledge management. 2016 : 287 – 296.
11. SEO M, KEMBHAVI A, FARHADI A, et al. Bidirectional attention flow for machine com- prehension[J]. arXiv preprint arXiv:1611.01603, 2016.
12. 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.
13. LEE K, CHANG M-W, TOUTANOVA K. Latent retrieval for weakly supervised open do- main question answering[J]. arXiv preprint arXiv:1906.00300, 2019.
14. PLANK B. What to do about non-standard (or non-canonical) language in NLP[J]. arXiv preprint arXiv:1608.07836, 2016.
15. GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.
16. WEI J, ZOU K. Eda: Easy data augmentation techniques for boosting performance on text

classification tasks[J]. arXiv preprint arXiv:1901.11196, 2019.

1. FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[A]. International conference on machine learning. 2017 : 1126 – 1135.
2. JAMAL M A, QI G-J. Task agnostic meta-learning for few-shot learning[A]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019 : 11719 – 11727.
3. RAJPURKAR P, ZHANG J, LOPYREV K, et al. Squad: 100,000+ questions for machine comprehension of text[J]. arXiv preprint arXiv:1606.05250, 2016.
4. LAI G, XIE Q, LIU H, et al. Race: Large-scale reading comprehension dataset from examinations[J]. arXiv preprint arXiv:1704.04683, 2017.
5. DAUMÉ III H. Frustratingly easy domain adaptation[J]. arXiv preprint arXiv:0907.1815, 2009.
6. BLITZER J, MCDONALD R, PEREIRA F. Domain adaptation with structural correspondence learning[A]. Proceedings of the 2006 conference on empirical methods in natural language processing. 2006 : 120 – 128.
7. ANON. Extracting and composing robust features with denoising autoencoders[C]. 2008 : 1096 – 1103.
8. CHEN M, XU Z, WEINBERGER K, et al. Marginalized denoising autoencoders for domain adaptation[J]. arXiv preprint arXiv:1206.4683, 2012.
9. GANIN Y, USTINOVA E, AJAKAN H, et al. Domain-adversarial training of neural net-works[J]. The journal of machine learning research, 2016, 17(1) : 2096 – 2030.
10. TZENG E, HOFFMAN J, SAENKO K, et al. Adversarial discriminative domain adapta- tion[A]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017 : 7167 – 7176.
11. LAN Z, CHEN M, GOODMAN S, et al. Albert: A lite bert for self-supervised learning of language representations[J]. arXiv preprint arXiv:1909.11942, 2019.
12. LIU Y, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.
13. BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33 : 1877 – 1901.
14. SANH V, DEBUT L, CHAUMOND J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter[J]. arXiv preprint arXiv:1910.01108, 2019.
15. SEO M, KEMBHAVI A, FARHADI A, et al. Bidirectional attention flow for machine com- prehension[J]. arXiv preprint arXiv:1611.01603, 2016.
16. SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent

neural networks from overfitting[J]. The journal of machine learning research, 2014, 15(1) : 1929 – 1958.

1. LOSHCHILOV I, HUTTER F. Fixing weight decay regularization in adam[J], 2018.

# **声 明**

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得四川大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

本学位论文成果是本人在四川大学读书期间在导师指导下取得的，论文成果归四川大学所有，特此声明。

学位论文作者（签名）

论文指导教师（签名）

年 月 日

# **致 谢**

执笔写下致谢，标志着毕业论文工作接近尾声，同时也给大学四年生活画上圆满的句号。四度春秋更替之中，我遇见了无数的人，也发生过太多太多的事，每一段曾经看似平凡的往昔现在都显得弥足珍贵。一路走来，一张张熟悉的脸庞，一件件刻骨铭心的回忆，都让现在的我百感交集。我明白，我有太多要感谢的人和事了。

首先是对父母之恩的感激！作为成长路上最坚实的后盾，一直关爱和督促着我的学业，在精神上给了我最大的支持，给予我鼓励和关心。于是我由衷的感谢父母永远站在我的身边，以及他们那无条件的爱。

感谢我的导师段磊老师，一直以来对我的关心和帮助，让我不断成长；感谢我的同学林俊卿，李涛，申城铭和杨佳智，在我学习生活上长久的支持；以及感谢我的室友一直以来对我的包容和陪伴，能够遇见你们是我的幸运。以及感谢我那些远方的朋友们，虽然我们身处山南海北，但属于我们之间那份最真挚的友谊却毫不褪色。

最后值得感谢的是在大学里坚持和努力的自己。

往者不可谏，来者犹可追。相信在不久的将来，贫瘠之地上也终究会长出最后一朵玫瑰！