上海交通大学学位论文

基于可解释性和鲁棒性的常识性推理研究

姓　　名：黄姗姗
学　　号：017033910040
导　　师：
学　　院：电子信息与电气工程学院
学科/专业名称：计算机科学与技术专业
申请学位层次：博士

2023 年 12 月

A Dissertation Submitted to
Shanghai Jiao Tong University for Master/Doctoral Degree

Research on Commonsense Reasoning Based on Explainability and Robustness

Author: Shanshan Huang
Supervisor: Prof. Mengyue Wu & Prof. Kenny Qili Zhu

School of Shanghai
Shanghai Jiao Tong University
Shanghai, P.R.China
December 23th, 2023

上海交通大学
学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全知晓本声明的法律后果由本人承担。

学位论文作者签名：
日期：　　年　　月　　日

上海交通大学
学位论文使用授权书

本人同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。

本学位论文属于：

公开论文

内部论文，保密 1 年/ 2 年/ 3 年，过保密期后适用本授权书。

秘密论文，保密 _____ 年（不超过 10 年），过保密期后适用本授权书。

机密论文，保密 _____ 年（不超过 20 年），过保密期后适用本授权书。

（请在以上方框内选择打 “√”）

学位论文作者签名：　　指导教师签名：

日期：　　年　　月　　日　　日期：　　年　　月　　日

摘 要

在人工智能领域，常识性推理始终是一个关键且极具挑战性的研究方向。核心问题在于赋予机器类似人类的能力，以处理和理解日常生活中的知识，并据此进行逻辑推理。这种能力对构建更高级、更自然的人机交互系统至关重要。尽管当前的 AI 技术在分析复杂数据、识别模式和执行特定任务方面已取得显著成就，但在进行深入的常识性推理方面，即使最先进的 AI 系统也显示出其局限。这种局限性不仅限制了 AI 系统在复杂环境中的应用效果，而且成为推动人工智能向更高层次发展的主要障碍。

本研究针对这一挑战，探讨了三个核心方向：首先是提升模型的常识性推理能力；其次是深入分析推理模型在鲁棒性方面的不足的原因，尤其关注模型在应对未知或对抗性数据时的反应和表现；最后则是在此基础上增强模型的整体鲁棒性，以应对现实世界中多变和不可预测的环境。

首先，为提升模型的常识性推理能力，我们提出了针对 AI 模型在理解复杂情境中常识性知识不足问题的具体解决策略。为了检验这些策略的有效性，我们选择了一个复杂且富有挑战性的任务——预测叙事故事的结尾。尽管现有模型在分析大规模数据集方面表现出色，但它们在处理复杂的情境和故事理解方面往往表现不佳。我们通过创新的方法简化故事内容，构建更为丰富和细致的故事表征，并结合 ConceptNet 预训练的概念图编码，显著提升了模型在此类任务中的常识性推理能力。

而后，我们聚焦于解析推理模型鲁棒性不足的根本原因。我们开发了两种测试框架，分别从宏观和微观两个层面对模型偏见进行了全面评估和分析。宏观层面的分析通过评估不同数据集上的简单分类模型性能，揭示了模型对某些统计规律的过度依赖，这有助于我们理解模型在泛化能力方面的局限性。而在微观层面，我们设计了 ICQ（“I-see-cue”）框架，该框架结合了特征的分布不平衡性和训练集与测试集分布的相似性，从而能够发现和评估可能影响模型决策的关键特征。通过多维特征划分和细致的性能分析，我们进一步探究了模型在不同特征上的准确性和分布表现。此外，我们还开发了直观的可视化工具，以更有效地识别和理解模型性能差异的根源。

最后，基于对鲁棒性不足的深入理解，我们进一步探讨了如何增强 AI

模型的鲁棒性。这一部分工作不仅关注于改善模型对当前数据的处理能力，更重要的是提高其适应新情境、抵御未知挑战的能力。为此，我们引入了创新的数据增强技术，包括生物学启发的“交叉”和“变异”操作，促进模型在训练过程中更多关注内在逻辑和结构，而非仅依赖表面的数据规律。这不仅提高了模型在复杂环境中的鲁棒性，也为其在现实世界应用中的可靠性和稳定性奠定了基础。

综合来看，本论文通过对常识性推理能力的提升、鲁棒性不足的深入分析，以及鲁棒性的增强，全面推进了人工智能在理解和处理复杂情境下的能力。我们的研究不仅在理论上拓展了 AI 领域的边界，也为构建更高效、更可靠的人工智能系统提供了实际的指导和基础。

关键词：常识性推理，模型鲁棒性，可解释性

ABSTRACT

In the field of artificial intelligence, commonsense reasoning has always been a key and highly challenging research direction. The core issue lies in endowing machines with human-like abilities to process and understand knowledge from daily life and use it for logical reasoning. This capability is crucial for building more advanced and natural human-computer interaction systems. Although current AI technology has made significant achievements in analyzing complex data, identifying patterns, and executing specific tasks, even the most advanced AI systems show their limitations when it comes to in-depth commonsense reasoning. These limitations not only restrict the effectiveness of AI systems in complex environments but also become a major obstacle in driving the development of artificial intelligence to higher levels.

This study addresses these challenges by exploring three core directions: firstly, enhancing the model's commonsense reasoning abilities; secondly, deeply analyzing the causes of deficiencies in the robustness of reasoning models, especially focusing on their responses and performances when dealing with unknown or adversarial data; and finally, based on this, enhancing the overall robustness of the model to cope with the variable and unpredictable environments of the real world.

First, to enhance the model's commonsense reasoning ability, we proposed specific strategies to address the shortcomings of AI models in understanding commonsense knowledge in complex situations. To test the effectiveness of these strategies, we chose a complex and challenging task – predicting the ending of narrative stories. Although current models perform well in analyzing large datasets, they often fall short in dealing with complex situations and understanding stories. We improved the model's commonsense reasoning ability in such tasks significantly by innovating ways to simplify story content, build richer and more detailed story representations, and combine these with the ConceptNet pre-trained conceptual graph encoding.

Then, we focused on analyzing the fundamental reasons behind the lack of robustness in reasoning models. We developed two testing frameworks that comprehensively assess and analyze model biases from both macro and micro perspectives. The macro-level analysis, evaluating the performance of simple classification models on different datasets, revealed the models' over-reliance on certain statistical patterns, helping us understand their limitations in generalization capabilities. At the micro-level, we designed the ICQ ("I-see-cue") framework, which combines the imbalanced distribution of features and the similarity between training and testing set distributions, enabling us to identify and assess key features that might influence model decisions. Through multi-dimensional feature segmentation and detailed performance analysis, we further explored the accuracy and distribution performance of models on different features. Additionally, we developed intuitive visualization tools to more effectively identify and understand the root causes of differences in model performance.

Finally, based on a deep understanding of insufficient robustness, we explored how to enhance the robustness of AI models. This part of the work focuses not only on improving the model's ability to handle current data but more importantly, on enhancing its ability to adapt to new situations and resist unknown challenges. For this purpose, we introduced innovative data augmentation techniques, including biologically inspired "crossover" and "mutation" operations, to encourage models to focus more on internal logic and structure during training, rather than relying solely on superficial data patterns. This not only improved the model's robustness in complex environments but also laid the foundation for its reliability and stability in real-world applications.

In summary, this paper comprehensively advances artificial intelligence's ability to understand and handle complex situations through the enhancement of commonsense reasoning abilities, in-depth analysis of deficiencies in robustness, and further enhancement of robustness. Our research not only theoretically expands the boundaries of the AI field but also provides

practical guidance and a foundation for building more efficient and reliable artificial intelligence systems.

Key words: commonsense reasoning, model robustness, interpretability

目 录

第一章 绪论

1.1 背景

在人类历史上，技术的每一次飞跃都预示着新时代的到来。而在 21 世纪，这一飞跃无疑是人工智能（AI）的崛起。作为数字化浪潮的先锋，AI 正改写着工作、生活甚至思考的方式。它的核心，在于对人类最自然的交流方式——语言的深刻理解与模仿。这正是自然语言处理（NLP）技术的魅力所在，它不仅是 AI 技术发展的重要里程碑，更是将机器智能与人类智慧紧密相连的关键。

随着 AI 技术的发展，我们见证了从简单的语言模型到复杂的 NLP 系统的演变，这些系统不仅能理解文字，还能把握语境和情感，甚至在某些领域超越人类的能力。然而，AI 在追求更深层次理解的路上，面临了一个新的挑战：常识性推理。这种看似平常的知识推理对于 AI 来说却是一座高山，它的攀登过程涉及了从基础知识到复杂推理的全方位挑战。

在本节中，我们的目标是梳理 AI 从其早期发展到自然语言处理技术的演进，特别关注它在理解和应用常识性知识方面的旅程。我们将详细探讨 AI 在常识性推理领域所取得的进展和所面临的挑战，并尝试揭示这个领域的主要发展趋势。

1.1.1 人工智能的全新世界

当我们踏入 21 世纪的第三个十年，我们同样步入了一个划时代的新纪元——人工智能（AI）的时代。在这个由算法塑造的新世界中，机器的能力已经不仅仅局限于模仿人类思维；它们在许多领域已经超越了我们的能力。以自动驾驶汽车为例，它们在繁忙的街道上的灵巧驾驭，展示了机器处理复杂、动态环境的卓越能力。而智能助手如 Siri 和 Alexa，已经成为我们家庭生活的亲密伙伴，它们不只是响应我们的命令，更能预测并适应我们的日常需求。

在医疗领域，AI 的影响同样深远。通过分析庞大的医学图像数据库，AI 辅助的疾病诊断在准确度上有时甚至超过了最资深的医生。而在金融界，AI 的应用从风险评估到欺诈检测，再到个性化投资策略的制定，无一

不在提高效率和精准度。在教育领域，个性化学习系统正根据每个学生的学习速度和风格，提供量身定制的教育内容，彻底改变了传统的教育模式。

这些例子仅仅是人工智能变革力量的一部分展示。随着技术的不断演进，我们预见 AI 将在更广泛的领域——如创意艺术、娱乐甚至环境保护中发挥其独特的影响力。人工智能不仅是技术发展的里程碑，更是推动社会前进、丰富人类生活的强大动力。

1.1.2 自然语言处理的发展进程

自然语言处理（NLP），作为人工智能的关键分支，致力于赋予计算机理解和生成人类语言的能力。过去十几年中，NLP 实现了从基础文本处理到复杂语言结构和情感分析的巨大跨越，这不仅是技术上的突破，更深刻地影响了我们的社会、商业和日常生活。

初期，NLP 系统主要依赖规则和语法分析来执行如词性标注和句法解析等任务。但随着计算能力和数据量的增长，深度学习技术的引入标志着 NLP 领域的一次革命。这使得 NLP 系统能够学习语言的微妙差异和复杂模式，提高了理解和生成自然语言的精准度。

在这一发展过程中，Google 的 BERT[35] 模型成为了里程碑，通过其双向训练方法深入理解文本上下文，极大提高了语言处理能力，并推动了文本分类、命名实体识别和问题回答等多个 NLP 应用的发展。

紧随其后，OpenAI 推出的 GPT 系列 [111] 在文本生成方面取得了显著进展。从 GPT-2[112] 到 GPT-3[13]，这些模型通过大量预训练生成连贯、自然且信息丰富的文本，极大丰富了写作辅助和创意内容生成等领域。

特别值得一提的是，OpenAI 推出的 GPT-4 模型及其基于此模型的应用 ChatGPT[1]，代表了 NLP 技术的最新进展。GPT-4 是一个先进的语言模型，具有巨大的数据集和复杂的算法，它在理解和生成文本方面的能力尤为出色。基于 GPT-4，ChatGPT 则专注于生成对话式文本，展现了在多轮对话中保持话题一致性和相关性的卓越能力。这种专门针对对话场景优化的模型，使得 ChatGPT 在客户服务、虚拟助手和互动式学习等应用中表现卓越。

NLP 技术可应用的领域也是相当广泛的并取得了瞩目的成绩。在机器

[1]https://chat.openai.com/

翻译领域，现代的基于神经网络的翻译系统在流畅度和准确性方面远超早期模型，能够处理更复杂的语言结构并获得更高精准程度的翻译效果。

在商业领域，NLP 的应用正在改变企业与客户的互动方式。智能聊天机器人现在不仅能提供全天候服务，还能处理大量咨询并提供个性化建议。同时，NLP 技术在市场营销中通过分析消费者评论，可以帮助企业深入理解市场需求和消费者情绪。

教育领域也从 NLP 的进步中获益颇丰。语言学习应用程序利用这一技术提供个性化的学习体验，根据学习者的进度和兴趣高效地教授新语言。此外，自动化的学生写作和语言能力评估为教师提供了宝贵的反馈，优化了教学方法。

总的来说，NLP 的发展不仅是技术层面的革命，更深刻地改变了我们的工作、学习和生活方式。随着技术的不断演进和新模型的推出，NLP 未来有望继续推动人工智能技术的边界，创造一个更智能、互动和个性化的数字世界。

1.1.3 常识性知识的重要性和发展进程

在探索人工智能（AI）的复杂领域中，常识性知识扮演着不可或缺的角色。这种从婴儿时期开始积累的基础知识和理解，虽然在日常生活中看似简单直观，对 AI 的发展却至关重要。例如，对于 AI 来说，理解诸如 “动物不会开车” 或 “我的母亲比我年长” 这样的基本事实，实际上是一个巨大挑战。

John McCarthy 的 “Advice Taker” 概念（1959 年）： John McCarthy，作为 AI 领域的先驱之一，在 1959 年提出了具有开创性的 “Advice Taker” 理念。这个假设性的程序旨在让机器利用人类的常识进行决策。McCarthy 认识到，虽然 AI 系统在处理特定任务上表现出色，但面对非典型或复杂问题时常显得无能为力。因此，他提出通过整合常识性知识，AI 系统可以更全面地理解现实世界，做出更准确合理的决策。尽管当时的技术条件限制了 “Advice Taker” 的实际发展，但 McCarthy 的这一构想明确了 AI 向集成常识性推理发展的方向。

专家系统的兴起（1974-1980 年）： 1974 至 1980 年间，随着专家系统的兴起，AI 开始尝试整合特定领域的知识。这些系统虽未完全融合人类的

常识性知识，但它们的出现代表了 AI 技术向更复杂知识集成方向的重要一步。专家系统的发展展现了 AI 在模仿人类专业判断和决策过程中的潜力，预示了未来 AI 在集成常识性知识方面的可能性。

CYC 项目的启动（1984 年）： Douglas Lenat 于 1984 年发起了 CYC 项目，这是一个雄心勃勃的尝试，目标是创建一个具备广泛常识的 AI 系统。CYC 项目致力于编码大量的常识性事实和规则，以模仿人类的常识推理过程。此项目的启动标志着 AI 领域对常识性知识重要性的认识和实践尝试，尽管面临着巨大的挑战和复杂性。

自动化知识获取的探索（1990 年代）：在 1990 年代，AI 开始探索从文本等资料中自动提取常识性知识的方法。通过分析互联网和书籍中的大量文本，AI 系统试图识别普遍存在的模式和关系。这一时期的研究强调了从现实世界的数据中获取和理解常识知识的重要性，为 AI 系统提供了一个更为丰富和动态的知识基础。

语义网络和本体论的发展（2000 年代）：进入 2000 年代，AI 研究者开始利用语义网络和本体论来更好地结构化和组织常识性知识。这些技术的发展使 AI 系统更有效地理解和处理常识性信息，提高了 AI 的知识管理和应用能力。这一时期的进展在于如何让 AI 系统更精确地识别和理解人类语言和行为中的常识性元素。

深度学习带来的突破（2010 年代）： 2010 年代见证了深度学习技术在 AI 领域的突破。这种技术的进步使得 AI 能够更有效地处理和利用大量数据，包括常识性知识。深度学习的引入提高了 AI 系统在理解语境、隐含意义和复杂语言模式方面的能力，这对于实现更加高级的常识性推理至关重要。

集成常识性推理的高级 AI 模型（2020 年代）：到了 2020 年代，高级 AI 模型如 OpenAI 的 GPT-3 开始在集成和应用常识性知识方面取得显著进步。这些模型展示了在理解语言、生成文本和解答复杂问题等方面的高级能力。这一时期的发展突出了 AI 在处理更复杂、更接近人类水平的常识性问题上的能力，预示着 AI 在理解和适应人类世界方面的巨大潜力。

通过这一系列的发展阶段，我们可以看到 AI 如何逐步获得处理和应用人类常识的能力，从最初的理念到现代高级模型的实际应用，这一历程体现了常识性知识在促进 AI 理解和适应人类世界方面的核心作用。

1.1.4 常识性推理的进展和挑战

在之前的讨论中，我们揭示了常识性知识在人工智能发展中的重要性，特别是在其历史演变和对现代 AI 系统的影响方面。这种知识，虽然在人类日常生活中似乎简单直观，却对 AI 理解和适应人类世界提出了复杂的挑战。接下来我们将深入探讨 AI 领域中的一个关键分支——常识性推理及其面临的挑战。

常识性推理的核心在于赋予机器不仅解读文字本身，而是洞察其中隐含的、对人类而言显而易见的知识和逻辑。这项能力的提升在自然语言处理（NLP）领域尤为突出。近年来，AI 在多个推理任务上取得了显著成就，包括但不限于指代消解（Reference Resolution）、问题回答（Question Answering）、文本蕴涵（Textual Entailment）、直觉心理学（Intuitive Psychology）和合理推（Plausible Inference），其中某些模型在特定推理任务上甚至已达到或超越人类的水平。例如，Google 的 BERT 模型在处理 SQuAD（斯坦福问答数据库）1.0[115] 和 2.0[114]，以及 GLUE[146] 数据集时的卓越表现，标志着 AI 在理解常识性知识方面迈出的重要一步。

下面我举几个例子来说明。

示例分析

SQuAD 示例：在 SQuAD 示例中，通常有一个给定的文章（Passage），一个相关问题（Question），以及该问题的正确答案（Answer）。任务的目标是让模型从文段中正确地抽取答案以回答问题。

如图 1-1 在这个例子中，文段提供了关于教师角色和职责的信息，其中包括教师通常使用的工具或方法来帮助学生学习。问题是：“What can a teacher use to help students learn?”（教师可以使用什么来帮助学生学习?）正确的答案是：“lesson plan”（课程计划），因为文段中提到教师可以使用课程计划来促进学生的学习。这种推理涉及到抽取文段中的关键信息，并将其与问题相联系，展示了模型在理解文本和提取相关信息方面的能力。

SNLI 示例：当进行自然语言推理任务并使用 SNLI[11] 数据集时，通常会有一个前提（Premise）句子和一个假设（Hypothesis）句子，以及一个选择（Choice）来描述这两个句子之间的关系。这个任务的目标是让模型判断前提和假设之间的关系是 “蕴含”（Entailment），“矛盾”（Contradiction），

文章（Passage）： The role of teacher is often formal and ongoing, carried out at a school or other place of formal education. In many countries, a person who wishes to become a teacher must first obtain specified professional qualifications or credentials from a university or college. These professional qualifications may include the study of pedagogy, the science of teaching. Teachers, like other professionals, may have to continue their education after they qualify, a process known as continuing professional development. Teachers may use a lesson plan to facilitate student learning, providing a course of study which is called the curriculum.

问题（Question）： What can a teacher use to help students learn?

答案（Answer）： lesson plan

图 1-1 SQuAD 数据集中的一个例子。

还是“中性”（Neutral）。

前提 (Premise) : A man pulling items on a cart.

假设 (Hypothesis) : A man is pushing a baby carriage.

选择 (Choice) : Entailment, Contradiction, Neutral

图 1-2 SNLI 数据集中的一个例子。

在示例图 1-2 中，前提是：“A man pulling items on a cart.”（一个男人在推车上拉着物品。）而假设是：“A man is pushing a baby carriage.”（一个男人正在推婴儿车。）模型的任务是判断前提和假设之间的关系。在这个例子中，前提和假设之间的关系是矛盾的，因为前提描述一个男人在拉车上的物品，而假设描述一个男人在推婴儿车。这两个句子之间的信息是相互矛盾的，因此模型可能会选择“矛盾”（Contradiction）作为答案，表明这两个句子之间存在逻辑上的不一致性。

模型能回答正确这些例子并在相关任务或者数据及上有出色的表现，表明现代模型已经在一定程度上掌握了推理能力，这标志着它们在理解和

应用常识知识方面已迈出了重要的一步。然而，正如我们将在接下来的讨论中深入探讨的，尽管取得了这些显著成就，人工智能在常识推理领域仍面临若干关键性挑战。

这些挑战主要分为三个方面。首先是提升模型在常识推理能力方面的挑战，关键在于如何使人工智能系统不仅理解语言的表层结构，而是深入洞察语言背后的逻辑和常识。其次，针对推理模型在鲁棒性方面的不足的表现，我们需要探究为何在面对未知或异常情境时，AI 系统可能表现出不稳定或错误的推理的原因。最后，这自然引出了如何增强常识推理模型的鲁棒性，即如何确保 AI 系统在面对多变和复杂的现实世界情境时，仍能维持其准确性和可靠性。下面是对这三个挑战的详细介绍。

挑战一：提升模型常识性推理能力的挑战

目前，AI 领域的研究虽然通过分析庞大的序列化数据，在常识性推理方面取得了一定的进展，但研究显示这些模型在深层次理解和高效运用常识性知识上还存在显著不足 [72, 101]。尤其是那些接受了大规模数据训练的预训练模型，在全面理解和精确运用常识性知识方面表现得并不理想。

这一现象指出了 AI 研究的一个核心课题：如何更有效地整合常识性知识到现有的 AI 模型中，从而显著提高其概念性知识推理的能力。当前 AI 模型在处理复杂逻辑推理和深度理解任务时的限制，很大程度上是由于对常识性知识的掌握和运用不足所致。为了应对这一挑战，我们需要从数据集的构建、模型架构的优化、训练方法的创新等多个维度进行深入研究和改进。这种努力不仅对于提升模型性能至关重要，而且是使 AI 在更广泛应用场景中更加贴近人类思维的关键一步。通过这些研究，我们有望在未来实现 AI 对现实世界中各种复杂情境的更好理解和应对。

挑战二：解析模型鲁棒性不足的原因

AI 模型在鲁棒性方面的缺陷，其根源的理解成为了一个重要的研究课题。模型的不鲁棒性可能源于多种因素，如数据的不足、模型架构的局限性、算法的内在缺陷，以及训练过程中的偏差。这些原因主要可以归结为数据问题和模型结构问题，因为最终的决策模型是由训练数据和模型结构共同作用的结果。然而，无论是从数据还是模型结构的角度出发，目前的研究都显得不够充分。未来的研究重点应该是开发出透明、可解释的架构，这对于理解模型在特定情境下的行为及其改进措施至关重要。

在医疗、刑事司法等关键领域，模型的决策透明度和可解释性显得尤为关键。例如，在医疗领域，医生需要清楚地理解 AI 是如何分析 CT 扫描图像来做出病情判断的，以确保诊断的准确性和可靠性。在刑事司法领域，对于释放犯人或批准保释等决策，AI 模型的决策基础的透明度和可解释性对于获得公众信任极为重要。因此，研究人员和开发者正在努力提升 AI 模型的可解释性，以保证这些关键应用的透明度和可靠性。能够解释模型鲁棒性不足的原因是可解释性研究开始阶段重要的一步。

挑战三：增强常识性推理模型的鲁棒性

AI 模型在处理熟悉的、分布内的数据时表现出色，但在面对分布外或对抗性样本时则表现出脆弱性。以自然语言推理任务 SNLI 为例（图 1-2 ），微小的假设变化，例如将假设 “A man is pushing a baby carriage” 更改一个单词，变为 “A man is carrying a baby carriage”，就可能导致如 BERT 等模型作出完全不同的推理结果。这展示了模型在泛化能力上的不足以及对训练数据依赖性的问题。这种不稳定性不仅存在于文本处理领域，图像识别等其他 AI 应用领域也面临着类似的问题。因此，迫切需要开发出能够适应新情境的模型，而不仅仅是依赖于已有数据的模式匹配。

1.2 研究内容和贡献

我的研究重点围绕常识性推理领域中的三个关键挑战展开，如第一节（见 1.1.4节）所述。首先是提高模型在常识性推理方面的能力，其次是解析模型在鲁棒性方面的不足的原因，以及最后一个挑战，即增强常识性推理模型的鲁棒性。

在下文中，我将介绍我在应对这三大挑战方面的主要研究内容及研究的贡献。我的研究的整体结构和逻辑框架在图 图 1-3 中有清晰的展示。

1.2.1 基于知识增强的常识性推理研究

1. 研究内容

在应对常识推理任务的首个挑战，即增强模型在常识性推理方面的能力时，我们的研究聚焦于目前领先的神经网络方法，比如 BERT。我们深入分析了这些方法的应用。虽然神经网络在常识推理任务中展现出强大的

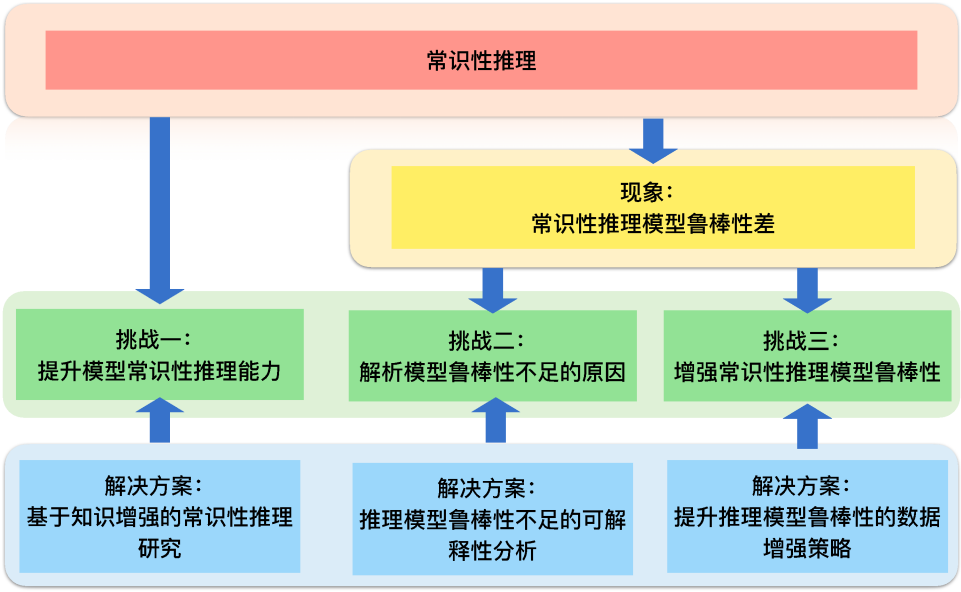


图 1-3 研究内容的整体结构。

性能，但根据 Lin 等人的研究 [72, 101]，在处理常识性知识时，这些方法仍有提升的空间，特别是在精确把握和表达这些知识时。

为了在常识性推理领域实现进一步的突破，我们选择了一个极具挑战性的任务：预测叙事故事的结尾。这一任务的设计灵感来源于早期关于故事理解的研究 [83]，并随后演化为预测故事中可能发生事件的任务 [16]。我们的研究在 ROC 故事填空测试数据集 [92] 上进行了实验，该数据集特别设计用于评估这种预测能力，它要求参与者从两个备选结局中选择一个最符合四句话故事情境的结局。通过这些实验，我们旨在深入探究并提升模型在理解和预测故事结尾方面的能力。

相关研究的初步探索集中在如何计算候选结局与故事情境句子之间的语义相似度。这通常包括对故事句子进行多维度表征，例如通过句子中词向量的聚合 [87]，应用 sentence2vec[64]，DSSM[52] 或采用其他创新特征 [127]。最新的方法趋向于采用双层结构：一是利用 LSTM[49]，GPT[111] 或 BERT[35] 等预训练语言模型构建句子的深层表征；二是在 ROC 数据集上对分类模型进行微调，形成一个全面的故事情境表征。

在此基础上，我们提出了一种新颖的方法，专注于识别和优化故事中的关键概念。我们发现，在故事推理过程中，将注意力集中在特定关键概

念上，而非全部词汇，可以更有效地指引至正确的结局。因此，我们设计了一种策略，在利用语言模型之前，通过突出重要概念并降低不重要概念的影响力，来简化句子。这种方法有效地降低了模型处理的噪音，并使其更专注于捕捉故事的核心信息。这些经过优化的句子被视为事件和概念的集合，而这些概念在 ConceptNet[134]——一个全面覆盖常识推理所需知识的开放领域知识图谱中有所定义。

此外，我们不仅注重于句子内概念的序列化处理，而且还将结构化的常识知识整合进句子表征中。通过融入预训练的概念嵌入到故事句子表征中，我们能够更准确地捕捉故事情境与结局之间的关键概念联系。例如，通过 CapableOf 和 MotivatedBy 这样的预定义关系，我们能够揭示 “long for break” 与 “have a rest” 之间的联系，从而在故事中实现更精准的推断。

值得强调的是，我们的方法在 ROC 故事填空测试数据集上展示了卓越的性能。尽管现有研究已在故事结局预测任务上取得了一定的进展，但这些成果主要集中在原始数据集的验证部分。为了确保我们的方法的有效性和泛化能力，我们避免了使用原始数据集的验证集，而是特意构建了一个新的训练集，该训练集与测试集不共享统计特征。这种严谨的实验设计使我们的评估结果更加可靠，同时也证明了我们所提出方法在提高故事理解和预测能力方面的显著效果。

2. 主要贡献

本研究在常识性推理领域引入了一种创新且有效的方法论。我们实施了一种精心设计的句子简化策略，专注于故事中关键概念的提炼。这种方法不仅降低了模型处理的复杂性，而且通过集中处理核心信息，极大地提高了模型捕捉故事情节的精准度。通过将这些精化的概念与 ConceptNet 数据库中的丰富结构化知识结合，我们在故事推理任务的准确性上实现了显著的提升。

在实验方法上，我们通过精心筛选和优化的训练数据集来确保了评估的公正性和准确性。这种方法学的应用不仅提升了故事结局预测的性能，还为常识性推理中结构化知识的有效运用提供了新的视角。本研究的贡献不仅在于推动了常识性推理领域的发展，也为未来相关领域的研究提供了宝贵的方法和见解。

1.2.2 推理模型鲁棒性不足的可解释性分析

1. 研究内容

从图 1-3 中可见，挑战二和挑战三均源于同一现象：推理模型的鲁棒性不足。尽管多种推理模型在常识性推理任务上取得了显著的成就，但它们在处理对抗性数据或与训练环境不同的场景时表现出了鲁棒性不足的缺陷。

在探究推理模型鲁棒性不足的原因时，关键在于分析模型处理信息的方法和重点。研究表明，模型可能过度依赖数据的某些特定结构元素，例如，在处理前提和假设的关系时，它们可能只关注假设部分，而忽略了前提与假设之间的逻辑联系 [95, 82, 126, 97]。这揭示出模型可能主要学习数据中的统计偏差，也就是所谓的偏见线索（bias cues）或者虚假线索（spurious cues）。

对这一现象出现的原因一种假设是数据集的构建方式存在偏见。模型往往学习特定于数据集的简化规则或者偏见，而非解决问题的通用策略。例如，在自然语言处理中，如果一个数据集中的大多数“正确”样本都包含某个特定词汇，模型可能仅依赖这个词汇做出判断，而未深入理解句子间的逻辑关系。这种简化学习方式导致模型在面对结构或上下文有差异的新数据时变得脆弱，无法有效适应或正确解释。如果这种线索刚好在假设部分，那可能就会造成模型只关注假设部分。

为验证模型可能过度关注数据特定结构元素的假设，研究者们开发了“仅假设”测试 [45] 和注意力图 [145] 这两种方法来分析和量化这一现象。尽管这些工具在解析和评估模型行为方面发挥着关键作用，它们在深入探讨数据偏见如何影响模型核心机制方面存在局限性。因此，深入研究数据偏见对模型机制的具体影响已成为一个紧迫且重要的研究课题。

对此问题，我们设计了两种测试框架来应对这一挑战。第一种是宏观层面的测试框架，基于统计特征来进行评估；第二种则是微观层面的详细统计分析框架。

宏观层面的统计特征测试框架

在宏观层面，我们的研究主要集中于分析自然语言推理任务中的虚假线索问题。我们注意到，尽管人类在处理这类问题时通常会分析问题前提与假设之间的逻辑关系，许多 NLP 模型却倾向于主要关注假设部分，忽略了深入的逻辑分析。这种现象往往是由于数据集中的人为偏差所引起。

在评测和分析数据集方面，我们提出了一种基于统计学线索发现的方法。这个方法首先应用统计学公式来拟合 “仅假设” 模型，以便识别数据集中最具代表性的统计特征。这种方法不仅揭示了数据集中的偏见和虚假线索，还为我们提供了一种衡量数据集质量的有效工具。

进一步地，我们设计了一种多次采样测试方法，将测试数据划分为简单和困难两类。这种划分旨在衡量模型在处理复杂推理任务时的真实能力，同时量化比较简单和困难子集之间的性能差异。通过这种方法，我们能够评估模型是否过度依赖于数据中的简单线索，从而揭示模型在不同情境下的潜在弱点。

此外，这种宏观评估方法不仅揭示了数据集的质量，还为我们提供了关于模型在逻辑推理能力上的深刻见解。这种方法的应用有助于我们理解模型在解决自然语言推理任务时的潜在限制，并指导我们如何改进模型以提高其泛化能力和鲁棒性。

微观层面的统计分析框架

在更细致的层面，我们采用了微观统计分析框架–ICQ（“I-see-cue”，即“我发现线索”）框架，这一工具旨在揭示和分析自然语言推理任务中数据集和模型的偏见性线索。ICQ 框架的应用分为两个主要领域：一是探索数据集中潜藏的偏见线索；二是评估偏见线索对模型性能的影响。

在数据集分析方面，ICQ 框架旨在识别可能在数据集制作过程中不经意间引入的偏见性线索。这些线索可能表现为特定词汇的使用、独特的短语结构或特别的语境模式。通过分析这些线索的词汇频率、共现模式以及它们与目标标签的关联性，ICQ 揭示了潜在的误导源，增强了我们对数据集如何影响模型预测的理解。

在模型分析方面，ICQ 重点分析自然语言推理模型如何在训练过程中学习和依赖这些线索。我们采用了准确性测试和分布测试两种方法，以全

面了解模型对数据集中的偏见性线索的敏感度。准确性测试揭示了模型在处理包含或缺失特定线索的数据样本时的表现，从而展示了模型对这些偏见性线索的依赖程度。而分布测试则通过可视化分析，探究了特定特征分布的变化如何影响模型的预测性能，为我们提供了更直观的理解。

将这两种测试方法结合，ICQ 框架为我们提供了一个全面的视角，以分析和评估模型在学习和反应偏见性线索方面的行为。这种综合性的分析方法不仅增强了我们对模型行为的认识，还促使我们在发现偏见时采取相应措施，以提高模型在处理具有潜在偏见的复杂数据时的鲁棒性和推理能力。这对于构建更加公正、高效的自然语言推理系统具有重要意义。

2. 主要贡献

本研究的主要贡献在于开发了这两种创新的分析框架，它们为深入理解并提升常识性推理模型的鲁棒性提供了新的视角。通过宏观和微观的方法，我们能够全面评估模型在面对复杂数据时的行为，特别是在识别和处理潜在的虚假特征方面。这不仅有助于揭示模型可能的薄弱环节，也促进了对模型学习机制的深入理解，为未来设计更为健壮和可解释的人工智能系统奠定了基础。

本研究的主要贡献在于开发了这两种创新的分析框架，它们为深入理解常识性推理模型的鲁棒性不足的现象提供了新的视角。

在宏观层面，我们分析了自然语言推理数据集的结构和内容，探究了如何在数据集创建过程中可能不经意间引入的偏见性线索。通过识别这些数据集中的特定词汇使用、独特的短语结构，我们深入理解了这些偏见是如何形成的，以及它们可能如何导向模型的预测。

微观层面的分析则集中在应用 ICQ 框架，一个专为分析模型对数据集中偏见性线索的响应而设计的工具。通过 ICQ，我们进行了准确性测试和分布测试，深入探讨了模型如何在训练过程中对这些线索产生依赖，以及这种依赖如何影响模型的推理能力和整体性能。ICQ 的应用使我们能够更细致地观察模型对特定特征分布变化的敏感性，为理解模型行为提供了更加丰富和直观的视角。

这两个层面的分析相辅相成，它们共同揭示了自然语言推理任务中数据集与模型之间的复杂关系，尤其是在理解和处理偏见性问题时。本研究不

仅为理解自然语言推理系统中的数据集和模型交互提供了深入的见解，也为构建更加公平、有效且鲁棒的自然语言推理系统提供了实践上的指导。这些贡献对于未来在该领域的研究和应用具有重要的意义，为改进数据集质量和提升模型处理能力提供了支持。

1.2.3 提升常识性推理模型鲁棒性的数据增强策略

1. 研究内容

本研究着眼于提升常识性推理模型在自然语言推理任务中的鲁棒性的挑战，尤其是在处理多项选择题（MCQs）时面临的挑战。多项选择题是评价常识性推理任务的常用格式，涵盖因果推理 [120]、故事结尾预测 [92, 53]、论证理解 [46] 以及阅读理解 [157] 等任务等多个领域。虽然这些格式在理论上被认为是有效的评估工具，但最新的研究发现，许多先进的神经网络模型在处理这类问题时，并没有真正理解问题的逻辑和语义，而是依靠数据中的统计特征或偏见进行判断。这一发现揭示了一个重要的问题：在自然语言推理任务中，这些模型展现出了一种我们称之为“短路行为”的现象。

这种“短路行为”指的是在 MCQs 问题中，模型在没有充分理解问题内容的情况下，依靠数据中的表面规律或偏见来作出判断。这一行为不仅限制了模型的泛化能力，也挑战了我们评估其真实理解能力的方法。因此，在本研究中，我们将探索和开发新的方法和技术，旨在更精确地识别和克服这种行为，从而提高模型在各类自然语言推理任务中的鲁棒性和有效性。

为了深入探究和量化这种“短路行为”，我们超越了传统的“仅选项测试”方法。我们发展了一种新颖的“代理测试”方法，这种方法不仅从宏观层面评估模型的行为，而且通过构建专门针对“短路行为”设计的新题目，远离了传统的“仅假设”框架。这种代理测试方法能够更精准地揭示模型在面对复杂任务时的真实处理机制。

为了进一步增强数据处理能力，我们采用了“交叉”和“变异”两种数据增强策略。这些策略旨在创造具有挑战性的新训练实例，推动模型超越对数据中表面统计规律的依赖，深入理解问题的本质。

“交叉”操作灵感来自生物遗传中的染色体交叉过程。具体操作是，我们将两个不同问题的选项进行互换，以创造全新的问题实例。例如，从两个不同的 MCQs 中各取一个选项，交换它们的位置，从而生成具有新组合

的问题。这种方法在模型的训练中引入了额外的复杂性和多样性，迫使模型去更深层次地理解问题和选项之间的关系。

而“变异”操作则是对原始问题的某些元素进行细微的调整或修改，以增加问题的复杂性和多样性。这类似于在生物进化中的基因突变，通过对问题或选项中的关键词或短语进行轻微的变化，产生新的问题版本。这些创新性操作能够产生全新的、具有挑战性的训练实例，促使模型超越表面的统计规律，深入学习问题的内在逻辑和结构。

将这些策略应用于当前领先的神经网络模型，如 BERT[35]、XLNet[156]和 RoBERTa[78]，并在 COPA[120]、ROC[92]、ARCT[46] 和 RECLOR[157]等多个基准数据集上进行测试。我们的实验结果显示，这些数据增强策略显著提高了模型在对抗性数据环境中的准确性，并在原始测试集上也实现了性能的提升。

2. 主要贡献

在本研究中，我们主要贡献在于开发了两种高级的数据增强方法，旨在提高常识性推理模型在多样化数据环境下的鲁棒性。灵感来源于生物学的“交叉”和“变异”的数据生成方法的应用经过了在多个基准数据集上的全面测试，证明了它们不仅能够提升模型在“短路问题”代理测试中的表现，而且也能在标准评估环境下维持或提高性能。通过这些创新，我们的研究不只是强化了模型对多样化和对抗性数据的处理能力，也为提升 AI 系统在真实世界应用中的泛化能力和可靠性提供了新的视角和工具。

主要贡献可以细分为下面三点：

1）**深入揭示和评估“短路行为”**：我们首次系统性地揭示了模型处理 MCQs 任务时的一种关键弱点——“短路行为”，即模型依赖于数据中的偏见或统计特征而非深入理解上下文和逻辑关系。通过开发“代理测试”，我们不仅扩展了传统的测试范围，而且专门针对这种行为进行了精准的评估。

2）**开发创新的数据增强策略**：“交叉”和“变异”：我们提出并实施了两种创新的数据增强策略，即“交叉”和“变异”。这些策略有效地引入了更高的问题复杂性和多样性，迫使模型超越对表面统计规律的依赖，深入理解问题的本质。这一策略在提高模型对复杂任务的理解和处理能力方面展现了显著效果。

3）显著提升模型鲁棒性和性能：应用这些策略于领先的神经网络模型并在多个基准数据集上进行测试后，我们的实验结果证实了这些方法不仅在压力测试中显著提升了模型的准确性，而且在常规测试环境中也增强了模型的整体性能。这一成果标志着我们的方法不仅成功地解决了“短路行为”的问题，而且在提高模型在多样化和对抗性数据环境下的整体鲁棒性方面取得了重要进展。

总体而言，本研究为自然语言推理领域提供了关于模型行为深度理解的新视角，并为提升模型在复杂数据环境中的泛化能力和可靠性提供了创新且有效的工具和方法，对构建更加鲁棒和可靠的自然语言推理系统具有重要的实际意义。

1.3 章节安排

在本研究的绪论中，我们对常识性推理的研究背景进行了深入剖析，全面探讨了人工智能和自然语言处理领域的发展历程，以及在常识性推理领域所面临的挑战。针对这些挑战，我们提供了一个概括性的概述，介绍了为应对每项挑战而开展的研究内容及其主要贡献，特别强调了在提升常识性推理能力、增强模型鲁棒性和深化模型可解释性方面所取得的进展。接下来的第三至五章将对每项研究进行更加具体的介绍。

第二章中，我们对与常识性推理紧密相关的关键研究领域进行了详细的梳理，包括任务的界定、基准数据集的构建、推理模型的方法论，以及这些模型在鲁棒性和可解释性方面的最新研究进展。我们对数据集和推理模型的方法论进行了系统性的介绍，这些内容将在后续章节中被频繁引用。

在第三章中，我们着重探讨了如何有效提升模型在常识性推理方面的能力，这是应对常识性推理领域的首个核心挑战。本章详细介绍了我们提出的一种新颖方法论，这包括故事中关键概念的识别与处理，以及通过精简关键概念来减轻模型对次要概念的依赖。此外，本章还探讨了如何通过融合结构化常识知识来深化对故事情境的理解。我们在 ROC 数据集上进行的实验展示了通过概念简化和知识图谱融合技术，我们的方法如何显著提升故事结局预测的准确性。

第四章深入探讨了解析常识性推理模型鲁棒性不足的原因，并提出了两种用于深化理解模型性能的分析框架。我们首先开发了一个宏观层面的

统计特征测试框架，用以评估模型在识别和学习虚假特征的能力。随后，我们引入了微观层面的详细统计分析框架（ICQ），对模型在不同特征上的表现进行了深入的分析。这些框架展示了如何为设计更健壮和可解释的 AI 系统提供新的方法和视角。

第五章则聚焦于提高常识性推理模型鲁棒性的关键策略，特别是在面对多样化和对抗性数据环境时的性能提升。本章介绍了两种创新的数据增强方法：通过在训练数据中注入噪声以减少统计偏差，以及灵感来源于生物学中的 “交叉” 和 “变异” 机制的方法，以创造新的训练实例。这些方法在多个基准数据集上的测试结果表明，它们能够显著提升模型的鲁棒性，减少 “短路” 问题，同时在标准测试集上保持或提升性能。

第六章主要对本论文的工作进行总结，并展望未来可能的研究方向。在这一章节中，我们将回顾研究的主要成果，并讨论未来研究可能探索的新领域。

第二章 研究现状

在上一节的内容中，我们回顾了过去十年间人工智能（AI）和自然语言处理（NLP）领域所取得的显著进展，并探讨了常识性知识领域的进步如何为常识性推理研究提供重要支撑。此外，我们还分析了在这常识性推理领域所面临的挑战，概述了我们在应对这些挑战时主要的研究内容和贡献。为本章内容的深入探讨奠定了基础。本章旨在全面介绍常识性推理的研究现状，涵盖从主要任务和评估基准到近年来流行的方法和模型，以及这些模型在可解释性和鲁棒性评估方面的研究进展。

本章首先聚焦于常识性推理领域的核心任务及其相关评估基准（benchmarks），以此为出发点，探讨了任务的具体内容、目标和挑战。随后，我们将深入讨论推理模型的方法论，包括符号方法、早期统计方法以及神经网络方法，并详细分析这些方法的优势和局限性。这些数据集和推理方法将在后续章节中将被频繁引用。

接下来，本章将重点探讨推理模型鲁棒性的研究。这部分内容包括鲁棒性的定义、测试方法，以及用于增强模型鲁棒性的各种策略。我们将深入分析模型所面对的多样化和对抗性数据类型，并探索提升这些模型鲁棒性的有效途径。

最后，本章将探讨推理模型的可解释性研究，包括模型鲁棒性不足现象分析、模型鲁棒性不足的原因的假设和验证方法。这一部分的内容对于理解模型在复杂环境下的行为模式及其背后的原因至关重要。

综上所述，本章将为读者提供一个全面而深入的视角，以理解常识性推理领域的当前研究现状，包括其核心任务、评估基准、方法论以及鲁棒性和可解释性的研究进展。

2.1 任务

在常识性推理领域，存在多个关键任务 [137]，每个任务都在理解和推动人工智能发展中发挥着至关重要的作用。以下为这些任务的详细介绍：

1. Reference Resolution **（指代消解）**：指代消解任务涉及识别文本中特定表达式（如代词或短语）所指代的对象。这一任务对于自然语言处

理至关重要，它不仅要求理解语言的上下文，还要求捕捉到语言的隐含意义。复杂句子中的多个名词和代词之间的准确指代关系识别，不仅需要深入理解句中实体间的关系和上下文环境，有时更需借助外部常识性知识 [91, 31]。指代消解是理解复杂文本和对话的基础，对于提升机器的语言理解能力至关重要。

2. Question Answering **（问题回答）**：这个任务要求对特定文本片段进行深入理解，以回答有关该文本的问题。这不仅考验了机器在处理语法和词汇基础上的能力，也考验了其在理解文本、提取关键信息、逻辑推理乃至应用外部知识等方面的能力。机器必须能够理解文本的深层含义，并在必要时利用广泛的背景知识来进行回答。

3. Textual Entailment **（文本蕴涵）**：文本蕴涵是由 Dagan 等人（2005）[28] 定义的，指的是文本与假设之间的方向性关系，其中如果一个典型人根据文本会推断假设为真，则可以说文本蕴含假设。一些基准测试通过要求识别矛盾来扩展这项任务，例如第四和第五届 RTE 挑战 [39, 7]。与问题回答相似，文本蕴涵任务需要利用多种简单的语言处理技能，例如命名实体识别和共指解析。不同于问题回答，文本蕴涵还需要理解一个典型人可能做出的推断，因此常识性知识对于这个任务至关重要。

4. Plausible Inference **（合理推理）**：Davis 和 Marcus（2015）[30] 定义的合理推理任务要求系统在有限上下文中做出合逻辑的、中间性的或不确定性的结论。这涉及到在故事中断的关键时刻，选择或生成最合理的后续事件。系统不仅需理解给定信息，还需运用常识性知识和逻辑推理预测最可能的结果。

5. Intuitive Psychology **（直觉心理学）**：直觉心理学是合理推理任务中的一个重要领域，涉及通过行为推断情感和意图，这是人类的一项基本能力。有些基准测试在某些示例中涉及这个主题。直觉心理学任务集中于理解和推断人类行为背后的动机、情感和意图。系统需要理解文本中的事实信息，并对人物的心理状态和社会互动进行深刻洞察。

6. Multiple Tasks **（多任务）**：一些基准测试由几个单独的语言处理或

推理子任务组成，以便在统一的格式下学习和测试不同的阅读理解技能，比如 bAbI[149] 和 GLUE[146]。这些基准可以用作诊断工具，以确定模型在不同领域的表现。子任务通常是从各种已存在的基准测试中重新框定的 [150]。这些基准对于全面评估和提高人工智能系统的语言处理能力至关重要。

2.2 基准数据集

在探索常识性推理领域时，我们特别关注了 2.1节中提及的若干关键任务，每个任务都在理解和推动人工智能发展方面扮演了至关重要的角色。本小节重点介绍了与这些任务紧密相关的主要数据集。这些数据集不仅为研究提供了丰富的实验材料和标准化的评估基准，还直接反映了当前自然语言处理技术的发展水平及其面临的挑战。下面的表格（表 2-1 ）详细展示了各种任务类型相关的关键数据集，包括 SNLI[11]、MNLI[151]、QNLI[146]、COPA[120]、ROC[92]、SWAG[158]、RACE[67]、RECLOR[157]、FEVER[143]、STS[126]、ARCT[46]、CQA（CommonsenseQA）[140] 和 Ubuntu[79]。

接下来，我们将对表 2-1 中提到的每个数据集的格式以及其主要特点进行详细介绍，以更深入地理解它们在评估不同推理任务中的作用和重要性。

1. SNLI（Stanford Natural Language Inference）：如表 2-1 所示，Stanford Natural Language Inference（SNLI），由 Bowman 等人于 2015 年提出（参见文献 [11]），是一项基准测试。它包含近 60 万个句子对，旨在执行类似于第四和第五届 RTE 挑战（参见文献 [39, 7]）的三种推理任务。除了标准的蕴涵、矛盾和中立标签，SNLI 数据集还特别包含了五种群体判断标签，这些标签反映了对每个判断的信心程度和一致性水平。

2. MNLI（Multi-Genre Natural Language Inference）：由 Williams 等人提出 [151]，包含 433,000 个示例，是目前最大的自然语言推理语料库之一。它涵盖了十种不同体裁的书面和口语英语，旨在涵盖现代标准美国英语使用的全部多样性。所有体裁都出现在测试和开发集中，但只有五种体裁包含在训练集中。这种设计允许研究者评估模型在已知来源（匹配）和未知来源（不匹配）的测试示例上的表现。MNLI 数据集的目的是推动自然语言理解（NLU）领域的研究，特别是在领域适应和跨领域转移学习方面。

数据集	任务类型	任务格式	特点
SNLI	文本蕴涵	分类	成对句子判断蕴涵关系
MNLI	文本蕴涵	分类	多体裁文本蕴涵
QNLI	文本蕴含，问题回答	分类	问题与回答的判断
ROC	合情推理	多项选择	选择故事合理结尾
COPA	合理推理	多项选择	因果或效果的选择
SWAG	合情推理	多项选择	预测情境后续
RACE	合情推理	多项选择	中高中水平阅读理解
RECLOR	合情推理	多项选择	逻辑推理阅读理解
FEVER	合理推理	分类	用于评估模型在事实验证方面的能力
STS	合情推理	分类	测试声明和相应证据的准确性
ARCT	问题问答	分类	论证有效性评估
CQA	问题问答	多项选择	评估常识性知识理解
Ubuntu	问题问答	多项选择	对对话模型的评估

表 2-1 常识性推理数据集概览。

3. QNLI（Question Natural Language Inference）数据集 [146]：是基于 SQuAD（Stanford Question Answering Dataset）[115] 构造的，SQuAD 由 Rajpurkar 等人（2016）提出，专注于自然语言理解和问题回答任务。SQuAD 包含超过 10 万个问题，旨在从 Wikipedia 文章段落中找到问题的答案。QNLI 从 SQuAD 中提取段落和问题，转换成自然语言推理的形式，即给定一个声明（问题）和一个文本片段（段落），要求模型确定文本是否包含该声明的答案。这种转换允许 QNLI 评估模型在理解复杂文本及其隐含含义的能力，尤其是在深入分析和推理的背景下。

4. COPA（Choice of Plausible Alternatives）[120]：由 Roemmele 等人在 2011 年提出，是一个专注于评估事件之间因果推理的任务。这个任务需要常识知识来判断通常在世界上发生的事件。COPA 提供一个前提和两个选择，要求从中选择一个作为最合理的原因或效果，测试模型的向前或向后因果推理能力。数据集包含 1,000 个这样的实例，是评估模型在理解和推理因果关系方面的能力的重要工具。

5. ROC[92]：由 Mostafazadeh 等人在 2016 年提出，是一个专注于日常生活故事的语料库，包含大约 50,000 个五句话故事。这些故事涵盖了丰富的因果和时间关系，非常适合于学习和评估常识性知识。其中大约 3,700 个

故事被指定为测试用例，每个测试案例包含一个合理和一个不合理的备选故事结尾，供模型在故事闭幕测试中进行选择。这个测试是 Chambers 和 Jurafsky（2008）提出的叙事任务 [16] 的一个更具挑战性的替代方案，旨在评估模型理解故事情节和进行逻辑推理的能力。

6. SWAG（Situations With Adversarial Generations）[158]: 由 Zellers 等人（2018）提出，是一个包含大约 113,000 个文本开头的基准数据集，每个文本开头有四个可能的结尾。这个基准测试旨在评估模型在情境推理方面的能力。

7. RACE 数据集 [67]: 由 Lai 等人在 2017 年开发，是一个专为评估模型在阅读理解任务上的能力而设计的挑战性数据集。它包含了来自中国中学和高中英语考试的 28,000 篇文章，共计约 98,000 个多项选择题。这些问题不仅覆盖了广泛的主题，而且往往设计得非常巧妙和具有挑战性，经常要求对文章中的多个句子或段落进行深入推理。与一般的阅读理解数据集不同，RACE 中的问题和候选答案通常无法通过简单的文本匹配直接找到答案，而是需要模型进行高级的推理和理解，这使得 RACE 成为评估和提升自然语言理解系统的一个重要工具。

8. RECLOR 数据集 [157]: 由 Yu 等人于 2020 年提出，旨在评估逻辑推理能力在阅读理解中的应用。该数据集从标准化的研究生入学考试（如 GMAT 和 LSAT）中提取了 6,138 个逻辑推理问题，每个问题都包括一个上下文段落、一个问题和四个选择答案，其中只有一个是正确的。RECLOR 的设计特点是它要求模型不仅理解给定文本，还需要进行深入的逻辑推理，以便在多个选择中找到正确答案。这种设计使 RECLOR 成为一个具有挑战性的数据集，用于测试和提高自然语言处理模型在复杂逻辑推理方面的能力。

9. FEVER[143]: 由 Thorne 等人于 2018 年提出，是一个大规模的事实提取和验证数据集。该数据集包含 185,445 个声明，这些声明是通过修改从 Wikipedia 提取的句子生成的，并在没有句子原文的情况下进行了验证。声明被分为三类：支持（Supported）、反驳（Refuted）或信息不足（NotEnoughInfo），由标注者进行分类，其一致性达到 0.6841 Fleiss kappa。对于前两类，标注者还记录了形成其判断所需的证据句子。FEVER 数据集的设计旨在提供一个挑战性的测试平台，帮助推动针对文本来源的声明验

证领域的进步。该数据集通过对声明和正确证据的标注实现了 31.87 % 的最高准确率，而如果忽略证据，准确率达到 50.91%

10. STS（Symmetric Test Set）[126]: 旨在解决在流行的 FEVER[143] 数据集中出现的偏见问题。由 Schuster 等人提出，这个新的 “Symmetric Test Set” 包含 956 个声明-证据对。每个原始声明-证据对都被人工生成了一个具有相同关系（支持或反驳）但表达不同、相反事实的合成对。这个测试集的构造完全消除了模型仅依赖于声明中线索的能力。

11. ARCT（Argument Reasoning Comprehension Task）[46]: 由 Habernal 等人于 2018 年提出，专注于评估模型在理解和分析论证性文本方面的能力。该数据集提供了在线新闻文章评论中的论证结构，包括约 2,500 个例子。每个例子包含一个观点、一个支持或反对该观点的论据，以及两个备选的保证（warrant），其中只有一个能正确地支持论证。任务是识别出正确的保证。ARCT 的挑战在于，许多论证的保证并非直接表达，而是隐含在论证中，需要模型通过外部知识进行推断。这种设计使 ARCT 成为一个有价值的工具，用于推动自然语言处理模型在理解、分析和推理论证性文本方面的进步。

12. CQA 是 CommonsenseQA[140] 数据集的缩写，由 Talmor 等人于 2019 年提出，是一个针对常识性知识的问答（QA）基准测试。它包含 9,500 个三项选择问题，旨在测试模型在解决涉及常识性推理的问题上的能力。每个问题都要求从 ConceptNet 这一常识性知识图中的三个相连概念中消除一个目标概念的歧义。CommonsenseQA 的设计确保了问题不仅直接针对常识性关系，而且所需的常识性知识领域对日常使用来说相当全面，从而在自然语言处理领域中提供了一个具有挑战性的评估基准。

13. Ubuntu 对话语料库 [79] 由 Lowe 等人创建，是一个大规模的多轮对话数据集，用于研究非结构化对话系统。该数据集包含近 100 万个对话，超过 700 万次发言，以及 1 亿个词汇。这些对话来源于 2004 年至 2015 年间的 Ubuntu 聊天日志，主要用于技术支持和问题解答。Ubuntu 对话语料库结合了对话状态跟踪挑战数据集中的多轮对话特性以及类似 Twitter 等微博服务的非结构化交互特性，为基于神经网络的对话系统研究提供了丰富的实验资源。

2.3 推理模型和方法

为了有效地解决 2.1节和 2.2节中所述的基准任务和数据集挑战，研究者们已经开发出了多样化的方法。这些方法的发展覆盖了从初期的符号逻辑与统计方法，到近年来深度学习和神经网络技术的广泛应用。本节旨在简要概述早期的符号逻辑和统计方法，同时将更加深入地讨论在当前和历史基准任务中广泛采用的代表性神经网络方法，展现它们在现代智能系统推理能力形成中的核心作用。

2.3.1 符号方法

符号方法在自然语言推理的应用源起于古典逻辑和演绎推理理论，如亚里士多德的逻辑理论（亚里士多德，1989）[133]，并逐渐发展至现代数学逻辑，例如摩根（1847）[32] 和布尔（1854）[9] 提出的形式逻辑框架。这些方法依靠逻辑形式和推理过程，对人类智力和推理能力的理解产生了深刻影响。在人工智能和语言学领域，例如麦卡锡（1968）[81] 和 Lakoff（1970）[68] 的研究，符号方法为机器的常识推理和语言的语义表示奠定了基础。此外，符号方法也在处理语言问题上显示出独特价值，如 Peirce（1883）[100] 提出的逻辑归纳方法和贝叶斯网络 [33]。

符号方法在初期 RTE 挑战中的应用取得了显著成就。例如，Raina、Ng 和 Manning（2005）[113] 的研究通过将句子转化为逻辑形式，并使用逻辑规则与手工编写的映射进行推理，实现了高准确率。然而，这种方法在大规模数据集的应用上存在局限性，因为手动编写的逻辑规则和映射难以涵盖语言和语义现象的广泛多样性 [63]。

2.3.2 早期统计方法

自 20 世纪 90 年代中期至 2010 年代初，统计方法在自然语言处理领域占据了主导地位。这些方法依赖于精心设计的工程特征和传统统计模型，如决策树 [96] 和朴素贝叶斯分类器 [40]。它们被广泛应用于从词汇特征到更复杂的语言特征分析，在 RTE 挑战中所展现的是这些统计方法在理解和解析语言现象方面的潜力和效果 [28, 47]。

早期统计方法的重大贡献在于它们为理解语言现象提供了基础框架。

然而，它们在处理大规模和高复杂度数据集时表现欠佳。这些方法的性能通常仅略高于随机猜测，其效率受限于特征工程的复杂度和数据的多样性[47, 66]。尽管早期统计方法为深度学习方法的发展奠定了基础，但它们在理解语言的深层结构和复杂性方面逐渐显得不足。

2.3.3 神经网络方法

神经网络方法在自然语言处理领域，特别是在自然语言推理任务中，代表了从早期的基于统计方法到复杂神经网络架构的显著进步。这一转变得益于大数据的可用性，它促使研究人员能够开发并训练更大、更深的神经网络模型，这些模型在多个自然语言推理基准测试中取得了卓越的成绩。图 2-1 展示了自然语言推理任务中一些典型神经网络模型的关键组成部分，主要包括预训练的语言表征、推理模型的核心架构，以及针对特定任务设计的输出层。

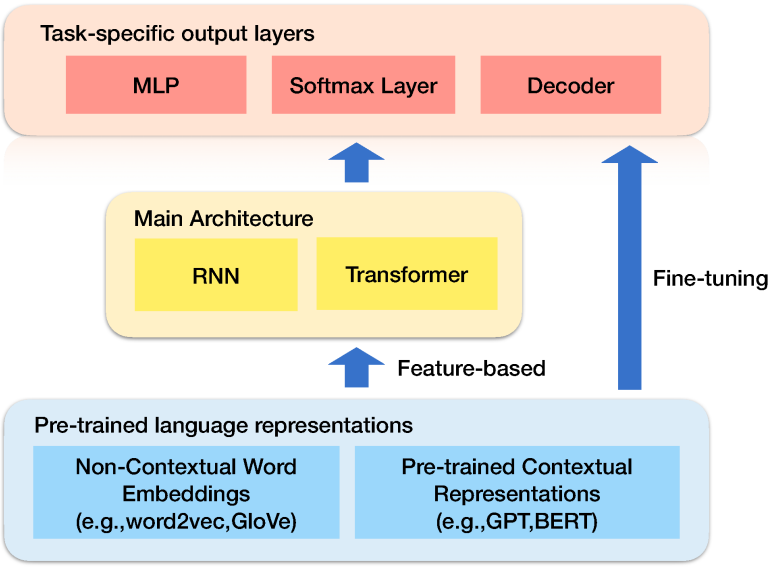


图 2-1 自然语言推理任务中神经网络模型的关键组成部分。

神经网络方法的基石是词或语言的分布式表征，即词向量或（embedding），这些通常通过在大规模文本语料库上训练神经网络得到。早期的词嵌入模型，如 word2vec[87] 和 GloVe[104]，提供的是上下文无关的静态表示。这意味着无论目标词出现在何种语境中，其嵌入向量都保持固定不变，这在处理词义多样性和上下文相关性方面存在局限性。

为了克服这些限制，研究者们发展了基于上下文的词表示模型，如 GPT[111] 和 BERT[35]。这些模型为同一词汇在不同上下文中提供不同的嵌入向量，使得对语言复杂性的捕捉更为精确。这些预训练模型可以直接用作下游任务的特征，也可针对特定应用场景进行微调。例如，GPT 和 BERT 模型采用了一种设计，其中只有很少的部分是特定于特定任务的。这意味着，当这些模型被用于不同的下游任务（如文本分类、问答系统等）时，我们只需要对模型的最后几层进行少量的调整，比如改变输出层的结构或者修改损失函数，就能够使模型适应新的任务。这种设计提供了高度的灵活性，允许同一个预训练模型在多种不同的任务中有效工作，而无需对模型本身进行大规模的重构。

在词嵌入层之上，为了满足不同下游应用的具体需求，研究者们开发了多种网络架构。其中包括循环神经网络（RNN），如长短期记忆网络（LSTM[49]）和门控循环单元（GRU[23]），以及 Transformer 架构 [144]。这些网络架构的选择依赖于特定任务的性质和要求。

针对不同的任务，这些网络架构的输出层也会有所不同。例如，在分类任务中，常用的输出层包括线性层或多层感知器（MLP），并结合 softmax 函数来生成概率分布，从而实现类别的预测。而在语言生成任务中，通常采用语言解码器（Decoder）来生成连贯的文本序列。

神经网络方法的发展不仅增强了系统处理语言多样性和复杂性的能力，还推动了深层次语言理解和推理能力的提升。接下来，我们将详细探讨几种在神经模型领域的经典方法，

- 词向量或嵌入的方法：如 FastText[61]，
- 在词嵌入层之上的特定网络架构：如 ESIM[22]，
- 预训练的上下文表示模型：如 GPT、BERT、XLNet[156] 和 RoBERTa[78]。

1. **词向量或嵌入的方法：**FastText

FastText 结合了简单的线性模型和高效的词嵌入技术，例如层次化 Softmax 和 n-gram 特征，以有效地处理文本分类任务。

1）模型架构

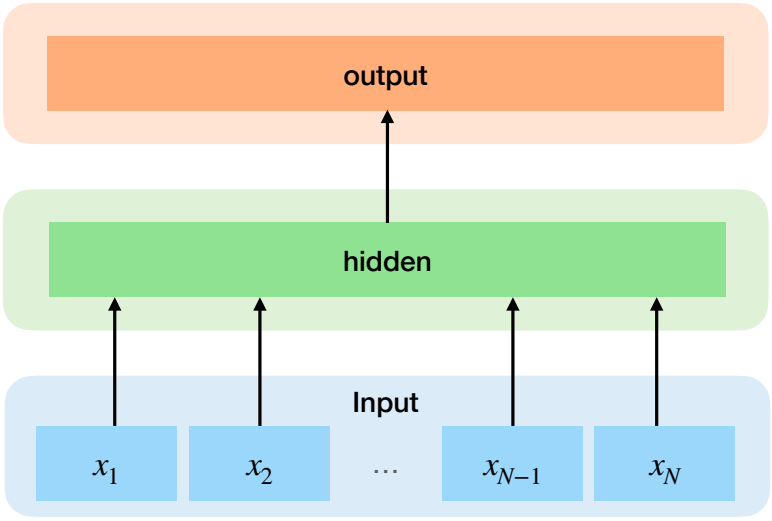


图 2-2 FastText 模型架构。

如图 2-2 所示, FastText 模型架构主要有三部分构成：输入层（input）、隐藏层（hidden）和输出层（output），这个结构跟 word2vec 的 CBOW[86] 的模型架构非常相似。

CBOW（Continuous Bag of Words）模型是自然语言处理领域的一种关键技术，主要用于词嵌入（word embeddings）。这个模型是 Word2Vec 的一部分，由 Google 的研究团队开发。CBOW 的核心思想是使用一个词的上下文（即它周围的词）来预测这个词本身。它通过将上下文中的词转换为向量（通常是 one-hot 编码），然后在模型的隐藏层中计算这些向量的平均值或总和，从而得到一个特征表示。最后，在输出层，CBOW 模型会预测目标词，输出是词汇表上所有单词的概率分布。

在训练过程中，CBOW 模型通过调整网络的权重来减少预测词和实际词之间的差距，从而学习到词汇的向量表示。这些词向量能够捕获词汇之间的复杂语义和语法关系，非常适用于各种自然语言处理任务，如文本分类、情感分析等。

相较于 CBOW，FastText 模型在某些方面采取了不同的方法。我们可以根据 FastText 模型的三个主要层级（输入层、隐藏层和输出层）来介绍其架构，并在此过程中对比其与 CBOW 模型的相似之处和不同之处。

输入层

- FastText：在 FastText 中，输入层接收的是单个文档中的多个单词及其 n-gram 特征。这些特征不仅包括单词本身，还包括其字符级别的 n-gram 表示，这增加了对文本的深层次理解。
- CBOW：而在 CBOW 中，输入层接收的是目标单词的上下文单词，通常仅限于单词本身，没有额外的字符级特征。
- **异同点**：两者都处理单词的向量表示，但 FastText 在输入数据的丰富度和维度上更为先进，包含了字符级的 n-gram 特征。

隐藏层

- FastText **和** CBOW：在这两个模型中，隐藏层的作用都是对输入层的多个词向量进行叠加和平均处理。这一过程形成了一个隐藏的特征表示，用于后续的预测或分类。比如在图 2-2 中，针对一个包含 N 个 n-gram 特征（x_1,x_2 ...,x_{N-1},x_N）的句子。这些特征被编码并平均处理，以形成隐藏向量。
- **异同点**：两者在隐藏层的处理方式上十分相似，主要是对输入的词向量进行集合和平均。

输出层

- FastText：FastText 的输出层用于文档分类，其输出是文档对应的类别标签。FastText 采用分层 Softmax，有效减少了计算复杂度，加快了训练速度。
- CBOW：CBOW 模型的输出则是预测目标单词，基于上下文单词来预测中心单词。
- **异同点**：这是两个模型最显著的不同之处。FastText 专注于文档级别的分类，而 CBOW 则关注于单词级别的预测。

总结来说，FastText 与 CBOW 在输入层的数据类型和输出层的目标上有着显著的不同，而在隐藏层的处理方式上则较为相似。FastText 的设

计使其在处理大规模文本分类任务时更加高效，同时其输入层的丰富特征表示也提高了模型对文本的理解能力。

2）概率模型

模型使用 Softmax 函数来计算预定义类别的概率分布。在训练过程中，模型的目标是最小化数据集中所有文档的负对数似然，公式如下：

$$-\frac{1}{M}\sum_{M=1}^{M}\log(f(BAx_m)) \tag{2.1}$$

其中 M 表示文档的总数，x_m 代表第 m 个文档的标准化词特征包，y_m 是对应的类别标签，A 和 B 是模型的权重矩阵，f 是 Softmax 函数。

3）FastText 的关键特点

- **层次化** Softmax：当类别数量庞大时，为降低计算成本，模型采用基于 Huffman 编码树的层次化 Softmax，将训练时的计算复杂度从 $O(kh)$ 降至 $O(h\log_2(k))$。

- N-gram **特征**：为了部分考虑词序，FastText 使用 n-gram 作为附加特征。通过散列技巧和大量的 bins（对 bigram 使用 10M 个 bins，其他情况下使用 100M 个 bins），实现了 n-gram 的高效映射。

4）模型优势和限制：

优势: FastText 通过结合线性模型的简单性和词嵌入技术的优势，有效地处理了文本分类任务。模型不仅利用了 BoW 来把握单词的分布信息，还通过 n-gram 特征来部分考虑词序。在面对大规模类别时，层次化 Softmax 确保了其高效性，使其特别适用于大规模文本分类任务。

限制：尽管 FastText 在处理罕见词方面取得了一定的进步，但它在捕捉词在不同上下文中的语义变化方面仍有所不足。此外，作为一种静态词嵌入方法，它无法有效解决自然语言推理任务中的复杂推理任务。

2. 在词嵌入层之上的特定网络架构：ESIM

在自然语言推理领域，采用词嵌入技术的网络架构已经取得了显著的发展，特别是在长短时记忆网络（LSTM）的运用上。Bowman 等研究者

[12] 利用基本的 LSTM 架构，在推理任务中实现了重要的突破，为随后的研究工作提供了坚实的基础。继此之后，Munkhdalai 和 Yu 在 2016 年 [94] 提出了一种创新的网络模型。这个模型不仅结合了基于序列的 LSTM 编码和递归网络，而且还融入了复杂的注意力机制。

在这一系列进展中，ESIM（增强序列推理模型）因其在 LSTM 基础上的创新设计和性能提升而引人注目。ESIM 不仅整合了双向长短时记忆网络（BiLSTM）的高效处理能力，还引入了树型长短时记忆网络（Tree-LSTM），以优化对结构化数据的处理。这种独特的结合让 ESIM 在处理自然语言中的序列和结构信息方面表现卓越，其性能超越了先前的模型。

ESIM 模型的整体架构可以在图 2-3 中查看。下面，我将详细介绍 ESIM 模型的各个组成部分，包括它们的结构和功能。

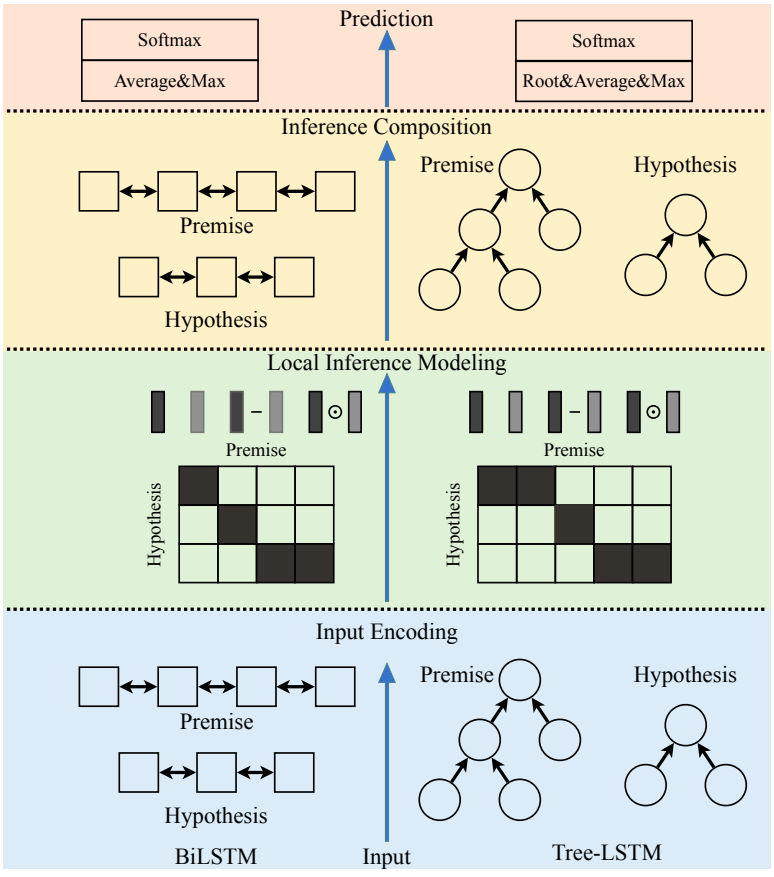


图 2-3 ESIM 模型架构。

1）输入编码（Input Encoding）

在 ESIM 模型的输入编码阶段，我们采用两种编码方式来处理不同特征的输入数据：BiLSTM 和 Tree-LSTM。这一阶段的目标是将输入数据的前提和假设转换为丰富的向量表示，为后续的推理过程打下坚实基础。

BiLSTM **编码**：

对于前提 p 中的每个词 p_i 和假设 h 中的每个词 h_j，BiLSTM 网络编码它们为隐藏状态 $\bar{p}_i$ 和 $\bar{h}_j$。这里 l_p 和 l_h 分别表示前提和假设中的词数。BiLSTM 通过考虑每个词的上下文信息，生成了更加全面的词表示。

$$\bar{p}_i = \text{BiLSTM}(p_i), \quad \forall i \in [1, \ldots, l_p] \tag{2.2}$$

$$\bar{h}_j = \text{BiLSTM}(h_j), \quad \forall j \in [1, \ldots, l_h] \tag{2.3}$$

Tree-LSTM **编码**：

当输入数据呈现树状特征（如解析树）时，可以采用 Tree-LSTM 进行编码。Tree-LSTM 是传统 LSTM 的扩展，特别适用于处理具有层次结构的数据。

传统 LSTM：传统的 LSTM 通过输入门、遗忘门和输出门控制信息的流入、保存和输出，从而有效地处理序列数据。以下公式详细描述了这些门的功能[2]：

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad \text{(遗忘门)} \tag{2.4}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad \text{(输入门)} \tag{2.5}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad \text{(输出门)} \tag{2.6}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad \text{(候选记忆单元)} \tag{2.7}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad \text{(记忆单元更新)} \tag{2.8}$$

$$h_t = o_t \odot \tanh(C_t) \quad \text{(隐藏状态)} \tag{2.9}$$

Tree-LSTM: Tree-LSTM 在每个节点上引入了左右子节点的遗忘门，以更好地处理树形结构中的信息流。以下公式展示了 Tree-LSTM 如何更新每

[2]这里的 h_t 是隐藏状态，跟假设 h 并无关系

个树节点的隐藏状态：

$$h_t = \text{TrLSTM}(x_t, h_{Lt-1}, h_{Rt-1}) \quad \text{（节点更新）} \tag{2.10}$$

$$o_t = \sigma(W_{ox}x_t + U_{Lo}h_{Lt-1} + U_{Ro}h_{Rt-1}) \quad \text{（输出门）} \tag{2.11}$$

$$c_t = f_{Lt} \odot c_{Lt-1} + f_{Rt} \odot c_{Rt-1} + i_t \odot u_t \quad \text{（记忆单元更新）} \tag{2.12}$$

$$f_{Lt} = \sigma(W_f x_t + U_{LLf}h_{Lt-1} + U_{LRf}h_{Rt-1}) \quad \text{（左子节点遗忘门）} \tag{2.13}$$

$$f_{Rt} = \sigma(W_f x_t + U_{RLf}h_{Lt-1} + U_{RRf}h_{Rt-1}) \quad \text{（右子节点遗忘门）} \tag{2.14}$$

$$i_t = \sigma(W_{ix}x_t + U_{Li}h_{Lt-1} + U_{Ri}h_{Rt-1}) \quad \text{（输入门）} \tag{2.15}$$

$$u_t = \tanh(W_{cx}x_t + U_{Lc}h_{Lt-1} + U_{Rc}h_{Rt-1}) \quad \text{（候选记忆单元）} \tag{2.16}$$

2）局部推理建模（Local Inference Modeling）

在局部推理建模阶段，ESIM 模型通过注意力机制计算前提 p 和假设 h 之间每对隐藏状态的相似度 e_{ij}，以建立局部推理关系。

$$e_{ij} = \bar{p}_i^T \bar{h}_j \tag{2.17}$$

使用这些权重 e_{ij} 来计算前提中每个词与假设中每个词之间的加权表示。

$$\tilde{p}_i = \sum_{j=1}^{l_h} \frac{\exp(e_{ij})}{\sum_{k=1}^{l_h}\exp(e_{ik})}\bar{h}_j, \quad \forall i \in [1,\ldots,l_p] \tag{2.18}$$

$$\tilde{h}_j = \sum_{i=1}^{l_p} \frac{\exp(e_{ij})}{\sum_{k=1}^{l_p}\exp(e_{kj})}\bar{p}_i, \quad \forall j \in [1,\ldots,l_h] \tag{2.19}$$

3）推理合成（Inference Composition）

在推理合成阶段，ESIM 模型结合局部推理信息 mp 和 mh，通过将原始向量、它们的差异和逐元素乘积组合在一起，捕捉前提 p 和假设 h 之间的细微差异。

$$mp = [\bar{p}; \tilde{p}; \bar{p} - \tilde{p}; \bar{p} \odot \tilde{p}] \tag{2.20}$$

$$mh = [\bar{h}; \tilde{h}; \bar{h} - \tilde{h}; \bar{h} \odot \tilde{h}] \tag{2.21}$$

4）池化（Pooling）

最后，ESIM 模型通过池化操作将这些信息转换为固定长度的向量，以便于最终的分类任务。这一步骤通过平均池化和最大池化操作，减少了输入序列长度对结果的影响。

$$v = [\mathrm{vp}_{\mathrm{ave}}; \mathrm{vp}_{\mathrm{max}}; \mathrm{vh}_{\mathrm{ave}}; \mathrm{vh}_{\mathrm{max}}] \tag{2.22}$$

通过结合这些池化结果，ESIM 模型得到一个全面的表示，能够用于下游的分类任务。

3. 预训练的上下文表示模型

近年来，自然语言处理领域经历了一场由预训练模型和嵌入向量发展所引领的革命。这些技术不仅能直接作为特征使用，还可针对特定下游任务进行微调。它们主要基于大量无监督文本数据训练，实现了从语义理解到具体应用的重大飞跃。

早期的预训练词嵌入模型，如 word2vec[87] 和 GloVe[104]，在多个领域被广泛应用。但这些模型有一个局限：它们是上下文无关的，即在不同上下文中使用相同嵌入向量，无法捕捉词义多样性。最近的研究通过引入基于上下文的词嵌入模型解决这一问题，代表模型包括 GPT、BERT 及其变体 XLNet 和 RoBERTa。

在深入介绍这些模型之前，了解它们共同依赖的核心架构–Transformer–是关键。2017 年，Google 在其开创性论文 [144] 中首次提出了 Transformer 结构，这在序列处理和翻译任务上是一大进步，超越了传统的循环神经网络 (RNN)。Transformer 的关键创新是自注意力机制，使模型能同时处理序列中的每个元素，并有效捕捉长距离依赖。

接下来，我将详细介绍基础的 transformer 模型基本框架和之后的衍生出的模型结构。

1) Transformer 的基本架构

Transformer 是一个基于注意力机制，特别是自注意力机制的模型，从根本上摒弃了传统的循环神经网络（RNN）和卷积神经网络（CNN）。如图 2-4 , Transformer 模型包含两个主要部分：编码器（图左侧）和解码器（图

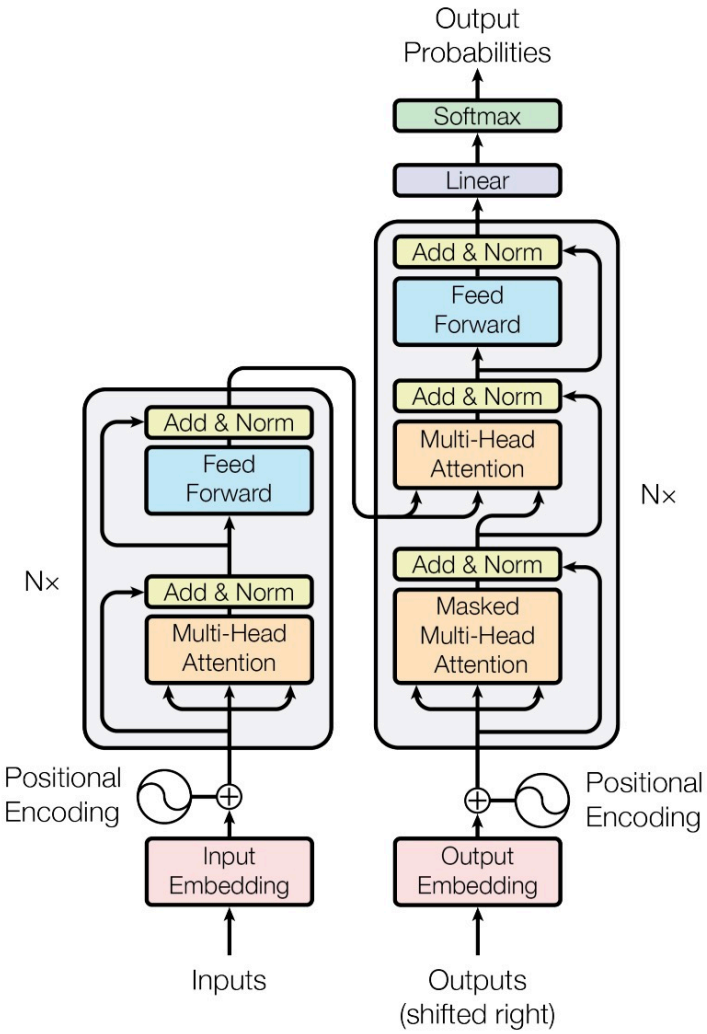


图 2-4 Transformer 结构框架 [144]。

右侧)。这两部分均由数个结构一致的层组成，其中每一层都融合了自注意力机制和前馈神经网络。

编码器的介绍

Transformer 的编码器部分由六个相似的层构成，每层包括两个主要子结构：

- 多头自注意力机制：作为编码器的核心，这一机制使得模型在处理序列中的每个元素时能同时关注其他元素，从而深刻理解整个序列的上下文关系。

- 逐位置的前馈网络：这个全连接网络层对序列中每个位置的数据进行独立的线性变换，进一步加强了模型对位置信息的处理能力。

此外，编码器的每个子层均采用残差连接和层归一化策略，有效防止了深层网络中梯度消失问题的发生，保障了信息的流畅传递。

解码器的介绍

解码器部分的结构与编码器相似，也是由六个层组成，但在子结构上有所不同，主要包括：

- 掩码多头自注意力机制：与编码器类似，但增加了掩码功能，防止解码器在生成序列时提前获取未来的信息，从而确保了信息生成的时序性。
- 编码器-解码器注意力：这一机制使解码器能够关注编码器的所有输出，为序列的生成提供了完整的输入序列上下文。
- 逐位置的前馈网络：与编码器的前馈网络相同，它独立处理序列中每个位置的信息。

解码器的每个子层同样使用残差连接和层归一化，增强了模型的稳定性和训练效率。

缩放点积注意力（Scaled Dot-Product Attention）

在 Transformer 模型中，缩放点积注意力（Scaled Dot-Product Attention）起着核心作用。它通过计算查询（Q）和键（K）之间的点积，来量化序列元素间的关联程度。为了在高维空间中保持梯度的稳定，缩放点积注意力机制将点积结果除以键的维度 d_k 的平方根进行缩放。其数学表达式为：

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2.23}$$

其中 V 代表值（Value），而 softmax 函数确保输出的归一化。这种设计使得每个元素的输出成为输入值的加权和，其中权重反映了元素间的相关性。

这个机制不仅提升了 Transformer 在处理长距离依赖问题上的能力，其并行计算特性也大大提高了模型的效率。

多头注意力（Multi-Head Attention）

多头注意力机制是 Transformer 架构中的一个关键创新，它的设计旨在使模型能够同时从不同的表示子空间捕获信息。这一机制的核心在于，它不是单一地应用注意力机制，而是将其分解为多个并行的“头”，每个头独立地关注输入数据的不同方面，从而提高了模型整体的处理能力和灵活性。

在多头注意力机制中，每个头都独立地执行缩放点积注意力的操作。这一过程可以分为以下几个步骤：

头的分割与独立运算：输入的查询（Q）、键（K）和值（V）首先被分割成多个头，每个头对应一组不同的权重矩阵 W_i^Q、W_i^K 和 W_i^V。这些权重矩阵是模型中的可学习参数，用于将输入映射到相应的子空间中。其中每个头的计算公式为：

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{2.24}$$

缩放点积注意力：已经介绍过。在每个头中，我们对映射后的查询、键和值执行缩放点积注意力计算。这包括计算查询和键的点积，根据键的维度进行缩放，应用 softmax 函数获取注意力权重，最后用这些权重加权求和值。

拼接与输出映射：计算完所有头的缩放点积注意力后，将它们的输出拼接起来，再通过一个额外的权重矩阵 W^O 进行映射，得到最终的输出：

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O \tag{2.25}$$

多头注意力机制的引入，使 Transformer 模型能够同时捕获序列的不同特征（如语法、语义等），并处理各种复杂的依赖关系。这种机制的优势在于其并行性和灵活性，使得模型在处理复杂的序列任务时更加高效和准确。

通过这种细致的解释，我们可以更深刻地理解多头注意力机制的工作原理及其在 Transformer 模型中的重要作用。这种机制不仅是 Transformer 的核心创新之一，也是推动当前自然语言处理领域发展的关键技术。

2) Transformer **的衍生模型**

随着 NLP 技术的不断进步，Transformer 的衍生模型，如 BERT、GPT、XLNet 和 RoBERTa，将 Transformer 结构与无监督学习结合，可以在不同 NLP 任务上实现前所未有的性能，无需每个任务重新训练。这些模型可根据使用的 Transformer 结构分类：

$$\text{衍生模型}\begin{cases}\text{Encoder 模型（如 BERT、RoBERTa），称为自编码模型。}\\ \text{Decoder 模型（如 GPT、XLNet），称为自回归模型。}\end{cases}$$

Encoder **模型**

Encoder 模型只使用 Transformer 模型中的 Encoder 模块，也被称为自编码（auto-encoding）模型。在每个阶段，注意力层都可以访问到原始输入句子中的所有词语，即具有“双向”注意力。Encoder 模型通常通过破坏给定的句子（例如随机遮盖其中的词语），然后让模型进行重构来进行预训练，最适合处理那些需要理解整个句子语义的任务，例如句子分类、命名实体识别（词语分类）、抽取式问答。BERT 是第一个基于 Transformer 结构的纯 Encoder 模型，它在提出时横扫了整个 NLP 界，在流行的 GLUE 基准上超过了当时所有的最强模型。随后的一系列工作对 BERT 的预训练目标和架构进行调整以进一步提高性能。目前，纯 Encoder 模型依然在 NLP 行业中占据主导地位。下面简略介绍一下 BERT 模型及它的变体 RoBERTa。

BERT **模型**

BERT 是基于 Transformer 架构的自然语言处理模型。它采用 Transformer 的编码器部分来学习文本中单词或子单词之间的上下文关系。与传统的单向模型不同，BERT 的编码器一次性读取整个单词序列，因此，虽然通常被描述为“双向”模型，但更准确地说，它是非定向的。

BERT 的训练涉及两种主要策略：Masked LM (MLM) 和 Next Sentence Prediction (NSP)。Masked LM (MLM) 是在输入序列中随机选择大约 15% 的单词并用特殊的 [MASK] 标记替换，然后模型尝试预测这些被遮盖的单词。这需要在编码器的输出顶部添加一个分类层，并且通过词嵌入矩阵将输

出向量转换为词汇量，最后使用 softmax 函数计算每个单词的概率。BERT 的损失函数仅考虑遮盖值的预测。

Next Sentence Prediction (NSP) 策略是 BERT 接受成对的句子作为输入，并学习预测第二个句子是否逻辑上紧跟在第一个句子之后。在输入中，50% 是连续的句子对，另一半则是随机的句子对。在输入模型之前，每个输入序列的开始位置会添加一个特殊的 [CLS] 标记，每个句子的末尾添加 [SEP] 标记，同时引入句子嵌入和位置嵌入。为了预测第二个句子是否跟随第一个句子，模型将 [CLS] 标记的输出转换为一个二维向量，并使用 softmax 计算其为连续句子的概率。

BERT 在多个基准测试上取得了显著的成绩，包括 GLUE、SQuAD、SWAG[158]、CLOTH[154]、DREAM[138] 和 SQuAD 2.0。BERT 通过这种方式，在理解语言的多个方面，尤其是上下文理解方面，取得了新的突破。然而，由于其复杂的结构和依赖于大量数据的特点，BERT 在一些情况下难以解释，并且在处理特定类型的语言数据时可能存在泛化问题。尽管如此，BERT 仍然是自然语言处理领域的一个重要里程碑，并为后续的模型如 XLNet、RoBERTa 等奠定了基础。

RoBERTa 模型

RoBERTa 模型代表了在深度学习和自然语言处理领域中对 BERT 模型的一个重要的进化步骤。这种进化不仅基于对 BERT 潜力的深入挖掘，而且体现了对预训练语言模型的深刻理解。在这一过程中，RoBERTa 的开发者们采用了四个创新和优化措施，旨在全面提升模型的性能和适应性。

首先，关于数据集的扩展和训练时长的增加，这些改变源于一个核心理念：更多的数据和更长的训练周期能够使模型捕捉到更为复杂和微妙的语言模式。通过将数据量从 BERT 的 16GB 增加到 160GB，并将训练迭代次数从 100K 提升至 500K，RoBERTa 能够在更广泛和多样化的文本上学习，从而提升其对于复杂语言结构的理解能力。

动态掩码的引入是 RoBERTa 另一个创新之处。与 BERT 的静态掩码不同，RoBERTa 在每个 epoch 中对文本采用不同的掩码模式。这种方法有效防止了模型对固定掩码模式的过度适应，从而推动模型学习更加泛化的语言表示。具体而言，这种动态掩码策略是在训练过程中实时应用的，确

保即使同一文本在不同的训练阶段也会呈现出不同的掩码模式，增加了训练的多样性和复杂性。

此外，去除 BERT 中的下一个句子预测（NSP）任务也是一个关键决策。尽管 NSP 被设计来提高模型对句子间关系的理解，但研究发现，在去除这一任务后，RoBERTa 在多项下游任务中的表现有了显著提升。这表明对于模型的优化，有时候减法也是一种有效的策略。

再者，RoBERTa 在训练中对样本长度的处理也展示了其对复杂语言模式理解的重视。通过在训练中始终使用全长（512 个令牌）的序列，RoBERTa 强化了对长距离依赖关系的学习，这对于处理更为复杂的语言结构至关重要。

这些改进使得 RoBERTa 在多个基准测试（如 GLUE、RACE[67]、SQuAD 等）中取得了领先性能，甚至超过了人类的表现。综合来看，RoBERTa 的这些创新和优化措施共同构成了其卓越性能的基础。通过在大规模数据集上进行深入训练，实施动态掩码机制，精简训练任务，以及优化样本长度处理，RoBERTa 不仅巩固了 BERT 作为一种强大的语言模型的地位，也为后续的 NLP 研究提供了宝贵的参考和启示。

Decoder 模型

Decoder 模型只使用 Transformer 模型中的 Decoder 模块。在每个阶段，对于给定的词语，注意力层只能访问句子中位于它之前的词语，即只能迭代地基于已经生成的词语来逐个预测后面的词语，因此也被称为自回归（auto-regressive）模型。下面就简要介绍一些常见的 Decoder 模型。

GPT 模型

GPT 模型是第一个使用了 Transformer 架构进行预训练的大型语言生成模型。这个模型标志着自然语言处理中一个重要的转变点。GPT 结合了 Transformer Decoder 架构和迁移学习方法，在大量开放在线数据上进行无监督预训练，特别是在 BookCorpus 数据集上。其预训练任务是根据上文预测下一个单词，从而使模型能够在没有限制的情况下学习语言特征。此外，GPT 还通过微调适应各种下游任务，如文本蕴含、语义相似性、情感分析等，在包括 SNLI、MNLI、ROC、COPA 和 GLUE 等多个 NLP 基准

测试中取得了显著效果。

然而，GPT 也存在一些局限性，比如它主要使用单向（向前）自注意力机制，这意味着在处理文本时，每个词只能看到它前面的词，这限制它对上下文的理解。

XLNet 模型

XLNet 模型也基于 Transformer Decoder 架构，但采用了双向（或更准确地说是排列敏感的）上下文处理。XLNet 通过排列语言建模（Permutation LM）来同时学习文本的前向和后向依赖。XLNet 的出现在自然语言处理领域引发了显著关注，特别是它在特定任务中相较于 BERT 的显著性能提升，标志着 NLP 模型发展的一个新阶段。这种模型不仅继承了自回归语言模型的强大特性，同时也融合了自编码语言模型（如 BERT）的优势，创造了一种新的、更为复杂和有效的预训练机制。

XLNet 的核心在于它的创新性预训练目标——排列语言模型。在 BERT 中，模型预训练是通过随机选择一些单词并将其替换为 [MASK] 标记来进行的，然后模型被训练来预测这些被掩盖的单词。然而，这种方法导致了一个问题：在实际应用中，即微调（fine-tuning）阶段，输入数据中并不存在 [MASK] 标记，这可能导致预训练和微调之间的不一致性。为了解决这个问题，XLNet 引入了排列语言模型。在这种模型中，对于一个给定长度的句子，模型会考虑所有可能的单词排列组合。在每次训练迭代中，模型会随机选择一种排列，并基于这种排列来预测单词。具体来说，对于一个句子中的每个位置，模型尝试预测这个位置的单词，给定在这个特定排列中它之前的所有单词。这种方法有效地模拟了自回归（autoregressive）语言模型的特性，同时也使模型能够在预训练中学习到上下文的全面信息。

在与 BERT 的比较中，XLNet 的预训练机制展现出三个独特的优势：首先，BERT 通过在输入中随机遮盖某些词并预测这些词来进行训练，而 XLNet 则采用了无需明显 [MASK] 标记的内部随机 “遮盖” 单词方法。这种差异使得 XLNet 在预训练和微调阶段保持了更高的一致性，解决了 BERT 预训练与微调不一致的问题。

此外，XLNet 在处理生成型任务上也表现出明显的优势。得益于其自回归特性，XLNet 能够在维持自然的从左到右的生成过程的同时，内部隐

含上下文信息，这对于生成连贯、流畅的文本尤为重要。

针对长文本的处理能力也是 XLNet 的一个亮点。它整合了 Transformer XL[29] 的技术，使得在处理长文档类型的 NLP 任务时，相比 BERT 有着更为显著的优势。

总的来说，XLNet 通过结合自回归和自编码语言模型的优势，通过其排列语言模型和注意力掩码机制，有效地利用上下文信息。这些特性使得 XLNet 在多项 NLP 任务中，尤其是在生成型任务和长文档处理方面，相较于 BERT 展现出更加卓越的性能。这种创新的预训练方法不仅解决了先前模型的某些局限性，也为未来 NLP 模型的发展提供了新的方向。

2.4 推理模型鲁棒性研究

上一小节是我对我所研究的相关的推理方法的介绍，尽管大型预训练语言模型在 NLP 领域取得了显著进展并在现实世界中得到了广泛应用，但它们在处理领域外数据、抵御对抗性攻击 [82, 59] 以及应对输入的微小扰动 [36, 3] 方面仍显脆弱。这些局限性可能会阻碍这些模型在实际环境中的安全部署，并影响用户对 NLP 模型的信任度。为了应对这些挑战，语言技术领域的研究者们正加大力度，旨在深入理解并解决这些鲁棒性问题。在这一节中，我们将进一步讨论自然语言推理任务中模型鲁棒性的相关研究。包括模型的鲁棒性的测试和提高模型鲁棒性的相关方法。

2.4.1 鲁棒性的定义和评估方法

Wang 等人 (2022)[148] 对模型的鲁棒性进行了定义：

定义 1 (鲁棒性). *在自然语言处理模型中，**鲁棒性**指的是模型在面对不同分布的测试数据时保持性能的能力。具体来说，对于一个输入 x 和它的正确标签 y，对于一个在 $(x,y) \sim D$ 上训练的模型 f 及其对 x 的预测 $f(x)$，鲁棒性是通过模型在测试数据 $(x',y') \sim D'$（可能与 D 分布不同）上的表现来衡量的。通常使用诸如鲁棒准确率这样的指标来量化鲁棒性，定义为 $E_{(x',y')\sim D'}[f(x') = y']$。*

在探讨模型鲁棒性的文献中，可以粗略地将他们分为两类：对抗性攻击下的鲁棒性和分布偏移下的鲁棒性。对抗性鲁棒性关注在输入 x 周围创

建微小扰动 δ 形成 x' 时模型的表现，而分布偏移鲁棒性关注在自然发生的不同分布上的模型表现。这些分类共同探讨 D' 是合成分布偏移（如对抗性攻击）还是自然分布偏移。

对抗性攻击下的鲁棒性：对抗性鲁棒性的研究主要集中在模型对输入的微小扰动（δ）的反应上，即在原始输入 x 周围引入扰动，形成 x'，并观察模型的响应。这类攻击通常通过添加对人类几乎不可察觉的噪声，以欺骗模型做出错误判断。这种策略最初在计算机视觉领域被广泛研究 [139, 42]，随后扩展到自然语言处理领域 [119, 141, 128, 10]，凸显了对 NLP 模型进行全面鲁棒性评估的重要性。

分布偏移下的鲁棒性：与之相对，分布偏移下的鲁棒性关注模型在不同自然分布的数据上的表现。这类研究反映了现实世界数据多样性和复杂性的影响，如语法错误、方言、语言差异等因素 [8, 34]，以及同一任务在不同领域的新数据集上的表现。

CheckList[119] 是一个结合了对抗性攻击和分布偏移下鲁棒性测试的方法论。它借鉴软件工程中的“黑盒测试”方法，重点评估 NLP 模型在不同输入变化下的表现。这一方法论包括三个核心部分：功能评估、测试方法和测试用例生成。

1）功能评估：此部分更多关注模型在处理各种 NLP 任务（如语义角色标注、命名实体识别等）时的能力，为后续的鲁棒性测试奠定基础。

2）测试方法和测试用例：这里包括对抗性攻击下的鲁棒性和分布偏移下的鲁棒性两个方面。对于前者，CheckList 通过模拟对抗攻击中的微小变化或扰动（如情感分析中的句子轻微修改），评估模型的反应 [42]。它通过不变性测试（INT）和定向期望测试（DIR）等方法，检验模型在处理对抗性攻击下的鲁棒性（如语法结构变化、情感强度变化）时的稳定性。同时在分布偏移的鲁棒性的测试中，CheckList 通过模板方法简化了大量有效测试用例的创建。例如，可以创建一个模板“我 {NEGATION}{POS_VERB} 这个 {OBJECT}”，并通过不同的词语组合生成多个测试句子，如“我不喜欢这个电影”，来评估模型对否定构造的处理能力。这些例子的主题和分布与原来的测试集已经完全不同。

总结来说，CheckList 为 NLP 领域提供了一个全面的框架，用于评估和提升模型在面对对抗性攻击和分布偏移时的鲁棒性。这不仅有助于揭示

模型的潜在薄弱环节，也为提高模型的整体鲁棒性提供了实践指导。

2.4.2 提升推理模型鲁棒性的方法

在自然语言推理的领域中，关键的目标之一是增强模型的鲁棒性。为此，我们通常采用两种主要策略：一是数据增强技术，用于丰富训练材料；二是优化模型架构和训练流程，以提升模型的内在强度和适应能力。

1）数据增强，作为优化大规模神经网络模型性能的核心策略，在自然语言处理领域的应用尤为关键。我们概述了几种主要的文本数据增强技术及其在学术研究中的应用，展示了它们如何为自然语言推理任务带来显著的性能提升。

回译（Back-translation**）**[129, 37, 153]: 通过将文本从一种语言翻译到另一种语言，再翻译回原语言，创造出文本的新变体，从而增加数据的多样性。

c-BERT **词替换** [152]：运用 BERT 模型来识别并替换文本中的关键词汇，生成文本的新版本，以此来丰富语料库。

混合（Mixup**）**[44, 21]：通过在不同文本样本之间进行线性组合，创造全新的文本实例。

截断（Cutoff**）**[132]：分为两类，第一类叫做 Token 截断，即随机选取 Token，将对应 Token 的嵌入整行置零；第二类是特征截断，即随机选取嵌入的特征，将所选特征维度整列置零。

2）优化模型架构和训练流程的方法，一般指的是调整模型结构来增强模型在自然语言推理任务当中的鲁棒性。最常用的方法是**对抗性训练** [162, 58]，这种方法通过在词嵌入层引入微小的扰动，合成附加样本，提高模型对异常输入的鲁棒性。该方法通过梯度反传生成对抗性扰动，并加到原始嵌入矩阵上，产生增强样本，通常应用于有监督训练场景。

在 Qu(2020 年)[110] 的研究中，通过比较这些模型增强技术在 RoBERTa 模型上处理自然语言推理任务（MNLI）的效果，发现回译和对抗性训练在提升模型性能方面效果显著，明显优于 c-BERT、混合和截断技术。因此，将回译作为自然语言推理任务中的一种强有力的基准进行数据增强方法比较，是一种合理的策略。这些研究成果不仅强调了数据增强在提升模型性能方面的重要性，还为选择适当的数据增强策略提供了宝贵的指导。

2.5 推理模型可解释性研究

2.5.1 推理模型鲁棒性不足的现象

在本节中，我们聚焦于推理模型在处理常识性推理任务时所表现出的鲁棒性不足问题。虽然这些模型在标准测试环境中取得了显著成就，但它们在遭遇对抗性数据或与训练环境显著不同的场景时，却显露出不足 [95, 82, 126, 97, 119]。

以 CheckList[119] 为例，该研究通过对模型进行重新测试，涵盖了多种不同能力，结果显示模型性能与原始测试结果相比有显著差异。这一发现揭示了模型在鲁棒性方面的严重不足。然而，尽管 CheckList 在揭露模型学习过程中的一些缺陷上取得了进展，比如在“语义角色标注”等任务下的表现，它却未能深入分析导致模型鲁棒性不足的根本原因。

2.5.2 推理模型鲁棒性不足的原因的探究

在深入探究推理模型鲁棒性不足的问题时，关键在于分析模型处理信息的方式和焦点。一种常见的假设是，这些模型可能过于专注于数据的特定结构元素，例如，它们可能只关注前提-假设关系对中的假设，而忽视了前提与假设之间的逻辑联系。这种假设揭示了一个更深层次的问题：模型可能主要学习了数据中的统计偏差 [95, 82, 126, 97]，也就是所谓的偏见线索（bias cues）或虚假线索（spurious cues）。

这种现象的根源在于数据集的构建方式。由于数据集的特定特征或模式，模型往往不是学习解决问题的通用策略，而是学习特定于该数据集的简化规则。举个例子，在自然语言处理任务中，如果某个数据集中绝大多数标记为“正确”的样本都包含某个特定的词汇，模型可能就会简单地依赖这个词汇来做出判断，而没有真正理解句子之间的逻辑关系。这种简化的学习方式使模型在面对结构或上下文有所不同的新数据时显得脆弱，无法有效适应或正确解释。如果这种线索刚好在假设部分，那可能就会造成模型只关注假设部分。

因此，探究推理模型的鲁棒性不足，我们不仅要关注模型本身的处理机制，也需要深入理解数据集构建的影响。

2.5.3 推理模型鲁棒性不足的原因的验证

在前面探讨推理模型鲁棒性不足的原因时，研究人员提出了一个关键假设：模型可能过分专注于数据的特定结构元素，如仅关注推理任务中假设部分而忽略了前提。为了验证这个假设，研究者们开发了“仅假设”测试[45]和注意力图[145]这两种方法，旨在深入分析验证和量化此现象。

仅假设测试

“仅假设”测试的核心思想是检验模型在仅凭假设（即问题或选项本身）的情况下的表现。这种测试通常用于评估自然语言理解模型。在这种测试中，模型只被提供假设部分，而没有相应的前提或上下文信息，比如在 1.1.4节中提到的 SNLI 的例子，只提供假设“A man is pushing a baby carriage”而不提供前提，让模型猜测前提跟假设之间的关系。从人的角度来看，解决这个问题是不可能的，因为问题本身有缺陷，很难做出选择，但是如果模型依然做出跟有前提时一致的结果，研究者就可以判断模型可能过度依赖于假设中的某些关键词或短语，而忽略了理解整体意义所必需的上下文。

这种方法的主要优势在于能够揭示模型潜在的缺陷：模型可能仅依赖假设中的关键信息来作出判断，而不是实现全面的理解。然而，其局限性在于它不能完全展现模型在处理真实场景时的推理模式。这是因为模型在训练阶段所依赖的数据包含了前提，而在测试时使用的数据结构却与之不同，仅包含假设而无前提。这种结构变化导致的差异和信息损失是该方法无法克服的。

注意力图

注意力机制是深度学习中的一种重要技术，它帮助模型聚焦于输入数据的关键部分。注意力图是一种可视化工具，它展示了模型在做出决策时对不同输入部分的关注程度。通过观察这些图，研究者和开发者可以更直观地理解模型的决策过程，识别模型关注的是哪些信息。

然而，注意力图的直观性并不总是代表其可靠性或准确性。研究[56]指出，即使模型通过注意力机制突出显示的部分，也可能并不准确地代表

模型决策的真实依据。这表明，尽管注意力图提供了一个理解模型决策过程的窗口，但它不应被视为模型内部工作机制的完整映射。

虽然“仅假设”测试和注意力图作为工具在解析和评估模型行为方面扮演着关键角色，但它们的分析深度有限。这些方法虽然能在一定程度上揭示模型中的偏见，但在深入探讨数据中的偏见如何影响模型的核心机制方面，仍显不足。因此，对数据偏见如何具体作用于模型的机制进行深入研究，已成为当前一个亟待解决的重要课题。

2.6 本章小结

在本章中，我们全面而深入地探讨了常识性推理领域的核心要素。章节始于对该领域的基本任务和评估标准的深度分析，揭示了其核心内容、目标和所面临的挑战。随后，我们对推理模型的发展历程进行了细致的梳理，从符号方法、早期统计方法，一直到神经网络方法，并对这些方法的优势和局限性进行了详尽的讨论。特别地，对于常识性推理中的神经网络方法，我们进行了深入的阐述，这些方法和模型将在后续章节中反复出现和引用。

章节接下来转向推理模型鲁棒性的研究，详细探讨了鲁棒性的定义、测试方法，以及提升模型鲁棒性的多种策略。我们对模型在处理多样化和对抗性数据时的挑战进行了深入的剖析，并探索了增强模型适应性和鲁棒性的有效方法。最终，本章还深入讨论了推理模型的可解释性研究，特别是模型鲁棒性不足的表现、原因的假设以及这些假设的验证方法。

综合来看，本章旨在为读者提供一个全景式的视角，深刻理解常识性推理领域的最新研究趋势和发展，包括其核心任务、评估基准、方法论以及鲁棒性和可解释性研究的最新进展。

第三章　基于知识增强的常识性推理研究

在本研究中，我们致力于解决常识推理领域的关键挑战：增强模型的常识性推理能力。在 第二章节中，我们已经深入探讨了当前领域内的领先技术，特别是 BERT 等先进的神经网络方法。尽管神经网络在常识推理任务中已显示出卓越的性能，但根据 Lin 等人的研究 [72, 101]，在精确处理和表达常识性知识方面，这些方法仍有进一步提升的空间。

针对这一挑战，我们选择了一个具有挑战性的任务作为研究重点：预测叙事故事的结局。这一任务的灵感源自早期的故事理解研究 [83]，并随时间演进，扩展为预测故事中可能事件的任务 [16]。我们的实验是在 ROC 故事填空测试数据集 [92] 上进行的，这一数据集为我们提供了评估模型性能的有效平台。

为了在这一领域实现突破，我们提出了一种注重故事中关键概念识别与优化的新方法。这种方法通过专注于故事的关键元素，而非全体词汇，旨在简化句子结构，减少干扰信息，从而提升模型在捕捉故事核心内容方面的效能。此外，我们采用了一种将结构化的常识知识融入句子表达的策略，通过结合结构化预训练的概念嵌入（embedding），增强了模型在理解故事情境与结局之间关键概念联系的能力。

通过将这些精细化的概念与丰富结构化知识相结合，我们不仅深化了对故事情节的理解，也在故事推理任务的准确度上取得了显著的提升。这一成果标志着我们在常识推理任务处理方面迈出了重要的一步。

3.1　概述

在叙事故事中预测“接下来会发生什么”是人工智能中常识推理的一个重要并且富有挑战性的任务。故事理解最初在规划和目标搜索的背景下被研究 [83]，这是人工智能中最重要的问题之一。随着研究的深入，这一任务逐渐演变为预测故事中可能接踵而至的事件 [16]。众多研究聚焦于一个标准数据集——故事填空测试（ROC） [92]。这个挑战要求从两个备选结局中选出一个与四句话故事情境最为吻合的结局，如图 3-1 (a) 所示。

先前的研究成果表明，结构化的常识知识对于深化故事理解颇具帮助 [71]。例如，观察图 3-1 (b) 和 (c)，我们不难发现，结构化知识有助

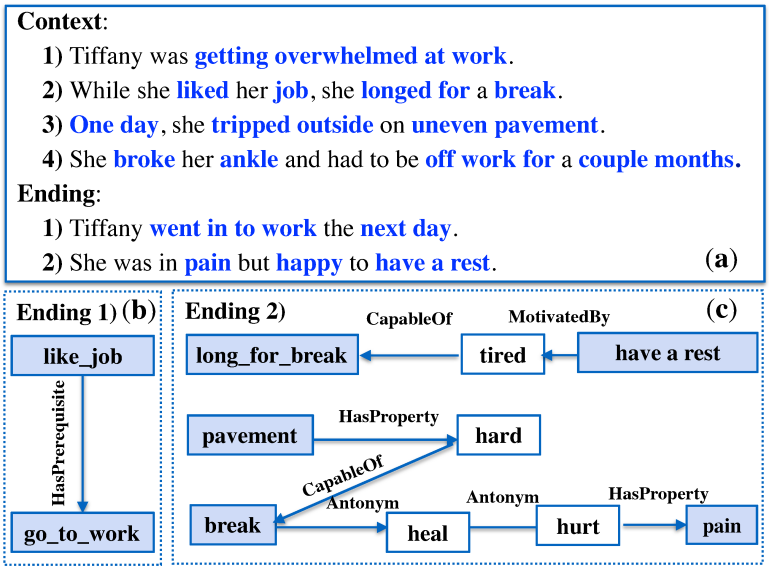


图 3-1　故事填空任务中的一个例子。在 (b) 和 (c) 中，蓝色框内的单词代表故事中的关键概念；白色框内的单词虽非故事概念，但作为桥接节点发挥作用。

于通过分析词汇间的逻辑联系来推断故事的可能结局。更有趣的是，仅凭借关键词（在图中以蓝色高亮显示）就能做出推理，这些词汇在推理过程中提供了丰富信息，而无需解析全文。相对而言，其他未突出显示的词汇不仅信息量较少，甚至可能由于其模糊的语义而对分类器造成干扰。例如，“Tiffany” 这一名字常与珠宝相关联，将这层含义引入故事情境中，往往会产生负面效果。

受以上观察的启发，我们从两个方面运用常识知识来优化故事的表达方式。首先，我们通过从 ConceptNet [134] 中提炼关键概念，对句子进行简化处理。ConceptNet 作为一个社区策划的开放领域知识图谱，涵盖了绝大多数常识推理所需的知识，从而实现了句内概念的有效表示。其次，我们通过融入 ConceptNet 知识图谱中的预训练概念嵌入，将结构化的常识知识引入故事句子的表达中。例如，在图 3-1 (c) 中，“long for break” 与 “have a rest” 之间通过 CapableOf 和 MotivatedBy 关系边相连。这些连接为我们在故事中 “串联起关键点”，帮助我们进行更加深入和有意义的推理分析。

尽管已有的方法在故事结局预测任务上的研究取得了显著进展，但这些成果多是基于原始 ROC 数据集的验证分割实现的。已有研究表明，验证集可能受到注释偏差的影响 [45, 131]，这意味着其中包含的统计特征可能

会被学习，而不需要真正理解故事。因此，在本研究中，我们未使用验证集，而是创建了一个新的训练集，与测试集之间不共享统计线索，从而最大限度地减少训练与测试之间的信息泄露。

本研究的主要贡献是在常识性推理领域引入了一种创新且有效的方法论。我们实施了一种精心设计的句子简化策略，专注于故事中关键概念的提炼。这种方法不仅降低了模型处理的复杂性，而且通过集中处理核心信息，极大地提高了模型捕捉故事情节的精准度。通过将这些精化的概念与 ConceptNet 数据库中的丰富结构化知识结合，我们在故事结局预测的准确率上实现了显著的提升。同时，在实验方法上，我们通过精心筛选和优化的训练数据集来确保了评估的公正性和准确性。

3.2　相关工作

我们的研究是在故事填空任务的背景下进行的，受到了三个研究领域的启发：脚本学习、常识性知识，以及文本推理中的标注偏差。接下来，我们将对这些领域进行详细的回顾。

故事填空任务

故事填空任务 [92] 的目的是为了评估故事理解和常识性推理的能力。众多基准方法 [84, 93] 通过测量情境句子与结局在向量空间中的语义相似度，来评定候选结局的合理性。可以采用不同的句子表征方法，比如对 word2vec 的平均 [87]、entence2vec[64] 和 DSSM[52]。另外，句子长度和字符 n-gram 等风格特征，也在区分正确与错误结局方面发挥作用 [127]。

近年来，多种深度学习方法被用于解决 ROC 任务。这些方法中的大多数遵循两层架构：首先构建每个句子的表征，然后将其聚合为整个故事情境的表征。句子表征可以通过 LSTM、GRU 和注意力层 [147, 161] 从词嵌入中得到，或者通过预训练模型如 Sentence2vec[121, 136] 和 GPT[111, 20] 生成。同样地，情境表征可以通过递归层 [15]、注意力层 [70] 或简单串联 [14] 对所有情境句子的语义进行编码。通过结合浅层特征和 DNN 模型，还可以进一步提高任务的准确性。

统计脚本学习

“脚本” 指的是一系列预定、固定模式的事件序列，用以定义特定的活动，它对文本理解十分重要 [17, 117, 124]。早期研究 [125, 90] 通过从文本中构建知识库来学习脚本。近期，研究人员运用统计模型从大量数据中提取不同类型的表示，包括无监督地学习叙事模式和脚本 [17, 117]，以及事件模式和框架 [18, 2, 130, 51]。为了推理这些知识，研究人员试图将事件预测问题转化为语言模型范式 [106, 122, 50]。对于故事填空任务，语义语言模型（SemLM）[103] 作为框架级别的语言模型，能够有效表示事件的顺序语义 [70, 19]。此外，基于规则的方法 [73] 为给定情境中的显式事件制定了匹配规则。其他一些工作 [89, 116, 88] 则将事件映射到语义嵌入表示中，这可能导致数据稀疏问题。

文本推理中的常识性知识

在许多推理任务中，常识性知识被证明是非常有效的，如阅读理解 [85] 和对话生成 [77] 等领域。最近，ConceptNet 已被纳入到 ROC 模型中。例如，[20] 提出了一种基于概念嵌入相似度的简单而有效的常识性特征。文献 [43] 则通过聚合句子中每个概念标记相邻的概念嵌入，来扩展每个概念的语义。在我们的研究中，常识性知识的运用主要表现在两个方面：一是作为简化过程的指导原则，二是作为增强每个句子语义的额外资源。

文本推理中的标注偏差

在多个文本推理数据集中，如 SNLI[11] 和 MNLI[45]，已经发现存在广泛的标注偏差。一些研究致力于通过人工 [131] 或采用对抗性方法 [158] 来自动生成具有额外标注的新数据集。文献 [121] 提供了一种生成错误样本的简单但有效的方法。其他研究则关注于避免偏见的模型 [24, 160]。尽管这些方法可能减少收集新数据集的工作量，但它们难以迁移到其他模型上，这是其主要局限性。

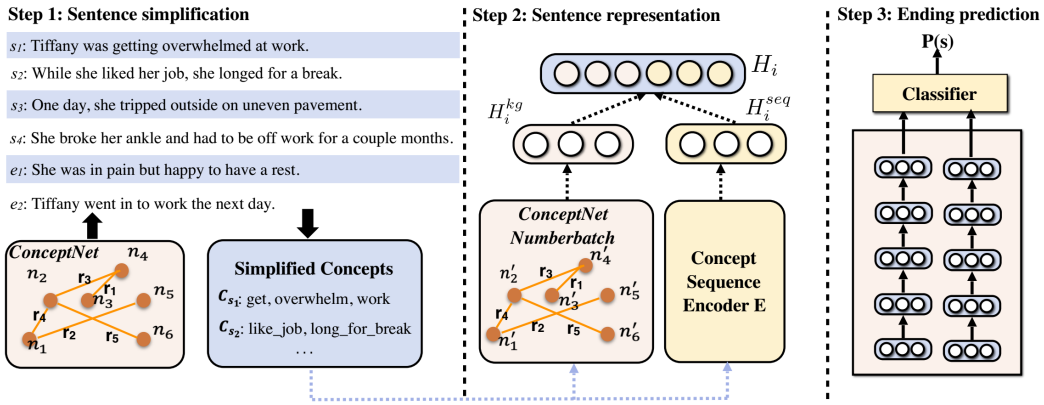


图 3-2　框架概览：我们的框架分为三个主要步骤：句子简化、句子表征和结局预测。$n_1,...n_6$ 表示 ConceptNet 中的标记节点，$r_1,...r_6$ 代表节点间的常识关系。$n_1^{'},...n_6^{'}$ 是在 ConceptNet 上训练得到的对应向量。

3.3　基于常识知识增强的故事结局预测框架

考虑到一个包含 L 个句子 $\mathrm{s}=(s_1,s_2,...,s_L)$ 的故事情境，我们的任务是从两个候选结局句子 e_1 和 e_2 中预测出正确的结局。我们提出的方法旨在通过引入 ConceptNet 中的常识知识，来改进和扩展故事句子的表示，以便更准确地预测故事结局。如图 3-2 所示，该框架包括以下三个主要步骤：句子简化（Sentence simplification）、句子表征（Sentence representation）和故事结局预测（Ending prediction）。

3.3.1　概念提取与句子简化

我们的目标是从包含 N 个单词的输入句子 $s=w_1,...,w_N$ 中提取一系列关键概念和事件 C_s。我们选择 ConceptNet [134] 作为这些概念和事件的来源，因为它覆盖了广泛的常识知识。ConceptNet 中的概念和事件通常由一到两个单词的短语表达，例如 “break ankle”。但在实际语境中，人们可能会使用更多样化的表达方式，如 “break her ankle”。为了弥补这种差异，我们开发了一种模糊匹配启发式方法，允许在 ConceptNet 中的概念短语中加入最多 λ 个额外单词以便在输入句子中实现模糊匹配。

面临的另一个挑战是，提取的概念可能在输入句子中相互重叠。例如，从句子 “She hope it would come back for more later” 中，我们能提取出 “hope”、“come back”、“come for” 和 “more” 等概念。在这种情况下，我们

Algorithm 1 句子简化算法

Input: ConceptNet C, sentence $s = \{w_1, ..., w_N\}$
Output: Concept sequence C_s

```
procedure Simplify(C, s)
    C_s ← {}
    for c ∈ C do
        for w_i ∈ s do
            t ← {w_i, w_{i+1}, ..., w_{i+|c|+λ}}
            if c is a subsequence of t then
                C_s ← C_s + {c}
            end if
        end for
    end for
    for c_k ∈ C_s do
        if ∃c ∈ C_s and c_k is contained by c then
            C_s ← remove c_k from C_s
        end if
    end for
    return C_s
end procedure
```

将保留所有有意义的概念。随后，我们会从 C_s 中移除那些被其他概念完全覆盖的重复概念。例如，“come” 会被 “come back” 覆盖，因此从列表中删除。

完整的简化算法展示在 算法 1 中，其中 $|c|$ 表示概念 c 包含的单词数量。

3.3.2 句子表征构建

经过简化后，原始句子 s 被转换为在 s 中相同顺序的概念序列 C_s。这些概念通常通过 ConceptNet 中的关系连接边相互关联，实际上形成了 ConceptNet 的一个概念子图，代表着重要的结构化知识。接下来，我们将介绍对概念序列和概念子图的编码方法。这两种嵌入的组合成为原始输入句子的完整表示。

概念序列编码

经过简化过程，句子的概念序列通过序列编码器 E 转换为向量表示。在本研究中，我们选用了 DSSM[52]、SKBC[121] 和 BERT[35] 等带有预训练文本表示的文本分类模型。详细信息将在 3.4.1节中描述。为了减少编码器的词汇量，概念序列 C_s 被转换为扁平化的单词序列 s'，即 C_s 中所有概念的单词拼接。与原始句子相比，s' 是一个简化的单词序列，已经丢弃了与常识无关的信息。例如，在上述情况下，简化序列为 “hope come back come for more”。我们认为每个概念中的连续信息至关重要，尽管它们不构成一个连贯的句子，但仍然被保留。

然后，简化序列 s' 被输入到序列编码器中，将 s' 映射成一系列上下文嵌入 H^{seq}：

$$H^{seq} = E\left(s^{'}\right) \tag{3.1}$$

概念图编码

除了简化句子中的扁平化概念序列外，概念之间的关系对于预测故事结局也极为重要。以往的研究 [20, 43] 已经表明，引入 ConceptNet 中的结构化知识能够补充故事中的常识性推理。与现有研究不同的是，我们不是生成手工特征，而是将结构化的常识知识直接纳入句子表征中。Numberbatch[3] 是 ConceptNet 知识图谱的预训练概念嵌入，涵盖超过 2,000,000 个常见概念，并在其他常识表示任务中已被证明有效 [135]。给定从句子中提取的概念序列 C_s，我们将结构化知识表示 H^{kg} 定义为所有概念的向量总和：

$$H^{kg} = \sum_{c \in C_s} Numberbatch(c), \tag{3.2}$$

其中 $Numberbatch(c)$ 表示概念 c 的向量。如果概念不在 Numberbatch 中，我们通过平均 Numberbatch 中所有构成词的向量来近似其概念向量。

最终，句子 s 的完整表示定义为两个组成部分的结合：$H_s = [H_s^{seq}; H_s^{kg}]$。

[3]https://github.com/commonsense/conceptnet-numberbatch

3.3.3 结局预测模型设计

为了从两个候选项中预测正确的结局，我们分别将两个候选结局与上下文句子结合起来。我们应用不同的分类器来判断哪个由 5 句话组成的故事 $\mathrm{s}=(s_1,s_2,s_3,s_4,e), e\in e_1,e_2$ 更可能是正确的：

$$P(y|s)=\begin{cases}\cos\left(H_{s_{[1:4]}},e\right) & \text{e.g., DSSM}\\ \mathrm{softmax}\left(H_{s_{[1:4]}},e\right) & \text{e.g., BERT}\\ \mathrm{softmax}\left(\mathrm{GRU}\left(H_{s_{[1:4]}},e\right)\right) & \text{e.g., SKBC}\end{cases} \tag{3.3}$$

其中 $H_{s_{[1:4]}}$ 是 s_1,s_2,s_3,s_4 的表示向量。分类目标 $y\in 1,2$ 对应于两个候选结局 e1 和 e2。

通过这种方法，我们能够有效地结合常识性知识和深度学习技术，以更准确地预测故事的结局。通过这一框架，我们不仅提高了结局预测的准确性，同时也增加了对故事情境的理解深度，使得预测过程更加符合人类的常识推理方式。

3.4 实验

在本节中，我们将展示我们的实验设计及结果分析，以评估我们提出的方法在故事结局预测中的有效性。首先，我们介绍了与我们方法相比较的几种基线模型，这些模型涵盖了当前故事理解领域的多种典型技术。随后，我们深入讨论了训练和测试数据集的构建，特别强调了新数据集的创建及其在实验中的重要作用。

实验结果和分析部分主要集中于展示不同模型在测试数据集上的表现，以及我们的简化后的概念序列编码和概念图编码方法对提高模型性能的影响。此外，我们还比较了不同的句子简化策略，并探讨了概念嵌入对故事结局预测的作用。最后，实验还包括了对训练数据规模和训练时间影响的分析。

总体而言，这一节旨在通过详尽的实验和细致的分析，全面展示我们方法在提高故事结局预测准确性方面的有效性和潜力。

3.4.1　基线模型与方法

我们的方法与两组基线模型进行了比较。首先，正如 3.1节所述，预训练的故事表征对于选择正确的故事结局非常重要。我们在三种典型的模型上应用了基于概念的故事表征技术：DSSM、SKBC 和 BERT。这些模型通过不同的机制和分类方法进行预训练，并代表了当前流行的预训练模型。

DSSM[92] 计算字符串对在连续语义空间中的相似性。在故事结局预测任务中，DSSM 将四句话的上下文和第五句话映射为语义向量，不考虑字符顺序的原始计数。使用三个 300 维隐藏层对上下文和备选结局进行编码。测试时，DSSM 选择余弦相似性较高的备选结局。

SKBC [121] 基于 Skip-thought[64]，适用于多个语义分类问题。其架构基于 GRU-GRU[49]。我们采用与 SKBC 相同的设置，但在 1000 节点 GRU 隐藏层前增加了 0.4 的 dropout。应用二元交叉熵函数来最大化正确结局的选择概率。所有实验均使用 200 的批量大小和 20 个训练周期。

我们在 BookCorpus 数据集 [163] 上使用 2400 维的语言模型 [64][4] 训练概念序列表示，该数据集包含 11,038 本书的文本。

BERT[35]，基于 Transformer[144] 开发，这个模型的结构我们在第一章已经介绍了。有几种可用的预训练 BERT 模型，它们在模型中使用的层数和参数数量上有所不同（基本版本有 12 层变压器块、768 隐藏大小和 12 个自注意头，共计 110M 参数；大型版本有 24 层变压器块、1024 隐藏大小和 16 个自注意头，共计 340M 参数）。我们选择了基本版本，它是在 BookCorpus 上预训练的。

我们的方法应用于这些模型上。对于简化方法，我们根据经验将额外间隔 λ 固定为 1，因为更大的间隔虽然有助于发现更多概念，但可能引入噪声。例如，在 “Sally went home and wondered about her parents’ marriage” 中，当 λ 等于 2 和 3 时，我们会错误地获得 “go wonder” 和 “go about”。此外，结构化知识表示采用来自 Numberbatch 的 300 维向量。

第二组基准方法涵盖了多种类型的模型：这包括基于特定特征构建的方法、使用生成模型的方法，以及一些与前述 DSSM、SKBC 和 BERT 等模型在方法论上相近的先进技术。

FES-JOINT[102] 结合了框架、实体和情感的特征。这个无监督的联合

[4] https://github.com/ryankiros/skip-thoughts

模型通过计算给定上下文的条件概率来选择适当的结局。

SeqMANN[70] 考虑了多个浅层特征，包括 POS 标签、词嵌入、字符特征、情感否定和 SemLM[103] 特征。

GMSA[43] 通过多源注意力生成故事结局，并引入 ConceptNet 邻域信息。我们比较生成结局和两个备选结局之间的相似性。

CGAN[147] 使用生成对抗网络 (GAN)，应用 GRU 生成训练数据增强的错误结局。

SIMP[136] 使用 Skip-thought 句嵌入，并使用多层密集网络分类器编码整个故事，以确定正确的结局。

GPT[111] 通过训练文本表示的语言模型并进行线性微调，取得了巨大的进步。GPT 也像 BERT 一样使用 Transformer 单元。

ISCK[20] 将情感和上下文与结局之间的常识特征纳入 [111] 的文本表示中，可以获得一些改进。

TransBERT[71] 不仅利用了大规模未标记数据中的一般语言知识，还利用了三个语义相关的转移任务，包括自然语言推理、情感分类和下一步动作预测，对 BERT 进行预训练和初始化。

3.4.2 训练和测试数据集

1. 数据集的选择与挑战

在先前的研究中，为了训练故事结局预测分类模型，研究者普遍使用了 ROC 验证分割，这个数据集包含 3742 个带有正确和错误结局标注的项目。这些项目被进一步分割为两部分，每部分含有 1871 个案例，分别作为训练集和测试集。这种方法在训练集上进行模型训练后，在测试集上取得了良好的表现。然而，这种方法存在明显的局限性：首先，用于训练的案例数量相对较少，可能限制了模型的学习能力；其次，由于训练集和测试集之间的偏见一致性，可能导致评估结果不够公平或准确。

此外，[131] 的研究指出，这些由错误结局标注的验证集由于存在人为创作偏见，不应作为训练资源。类似地，[98] 的研究发现，BERT 在论证推理理解任务 [46] 中的表现完全依赖于数据集中的虚假统计线索。这些发现表明，仅使用 ROC 验证分割进行训练和测试可能无法准确反映模型在实际应用中的效果。

数据集	正确结局	错误结局	总数
验证集	3	70	73
测试集	4	69	73

表 3-1　ROC 验证集 (ROC(V)) 和测试集 (ROC(T)) 中单词 "hate" 出现在正确和错误结局的频率。

2. **新数据集的构建**

基于以上考虑，我们决定采用不同的方法来构建训练和测试数据集。我们的目标是创建一个新的、规模更大且更公正的数据集，以更准确地评估故事结局预测模型的性能。这一决定得到了我们在 Amazon Mechanical Turk (AMT) 上征集的 ROCStories 及其错误结局中的初步实验的支持。如表格表 3-1 所示，我们发现特定词汇（例如 "hate"）在错误结局中出现的频率远高于正确结局，暗示了潜在的信息泄露问题。

针对这一问题，我们重新评估了几种表现优异的算法以及我们的方法。这些算法和方法仅在验证集（ROC(V)）的结局部分进行训练。ROC 测试集的结果展示在表 3-2 中。作为基线，我们还包括了人类表现，即 5 名未经过验证集训练、仅凭常识判断的人类标注者的平均准确率。人类得分远低于部分 "优秀" 算法的情况表明，这些算法并非真正运用 "常识"，而是依赖于训练数据中的模式。

最近发布的修订版本 $ROC_v1.5$ [131] [5] 旨在减少 ROC 中的人为偏见。然而，即使在 $ROC_v1.5$ 中，仅以结局为依据的结果仍然高于人类表现 [131]。此外，它只包含了规模更小的验证集和测试集。因此，这个数据集并不一定解决了在故事闭环测试中提供合适的训练和测试资源的问题。尽管如此，我们还是在 $ROC_v1.5$ 上展示了我们模型的结果（详见 3.4.3节）。

为了解决这些问题，我们选择自动为 ROCStories 语料库添加错误结局，创建一个新的训练数据集。我们遵循了 Roemmele 等人 [121] 的方法，通过随机和向后方法为 ROC 训练集中的 98161 个故事生成错误示例。随机方法将每个故事的结局替换为训练集中另一个故事的随机选定结局，而

[5] $ROC_v1.5$ 发布在https://competitions.codalab.org/competitions/15333#participate-submit_results，但目前已关闭，我们无法获取含有正确标签的完整数据集。我们呈现的结果是在测试阶段结束之前获取的。

模型	ROC(V) (%)	ROCS*(Tr) (%)
SIMP	72.60	59.86
SKBC	72.76	58.18
GPT	77.77	57.93
$\text{TransBERT}_{\text{BASE}}$	79.0	54.52
$\text{TransBERT}_{\text{LARGE}}$	75.84	54.30
人类	62.40	62.40

表 3-2　在 ROC(V) 和 ROCS*(Tr) 上仅使用结局训练的各种模型的测试准确率。

向后方法则通过替换故事的最后一句话来生成错误示例。从这六个备选结局中（4 个来自随机，2 个来自向后），我们随机选择一个以确保正确和错误数据的平衡。这样生成的数据集被称为 ROCS*。我们将 ROCS 分为训练集（ROCS(Tr)）和验证集（ROCS*(V)），比例为 4:1。

在表格表 3-2 中，我们发现仅使用结局进行机器学习的模型表现比人类在我们重构的训练数据集上的表现更差，这反映出了我们数据集构建的有效性。这个数据集在一定程度上弥补了偏见信息泄露的问题。值得注意的是，在以下所有提及的训练数据集、验证数据集和测试数据集中，我们指的是我们新创建的数据集：ROCS*(Tr)、ROCS*(V) 和 ROC(T)。

3. ConceptNet **的应用**

在构建新数据集的过程中，我们面临概念上的词汇表外（OOV）问题。我们选择 ConceptNet 作为我们简化资源的主要原因是它覆盖了大量由多种来源（如 WordNet、DBPedia 和 OpenCyc）构建的常识性知识。如表格表 3-3 所示，在训练集和验证集中，不到 0.020% 和 0.017% 的结局句子完全不包含概念，这表明 ConceptNet 在覆盖广泛的概念领域方面的有效性。

综上所述，我们通过创新地构建训练和测试数据集，并有效利用 ConceptNet，旨在提高故事结局预测模型的准确性和公正性。我们的方法强调了在减少训练数据中的偏见和信息泄露方面的重要性，为模型的实际应用提供了更可靠的基础。在之前的研究中，为了训练故事结局预测分类模型，

数据集	总故事数	上下文零概念数	结局零概念数
训练集	157058	0	26
验证集	39264	0	8
测试集	3742	0	0

表 3-3　训练集、验证集和测试集中故事（上下文和正确或错误结局）的总数和上下文或结局中不含概念的故事数。

模型	原始 (%)	简化 (%)	概念图编码 (%)	简化 + 概念图编码 (%)
DSSM	54.04	58.79	54.0	58.2
SKBC	64.70	68.13	65.12	69.7
$\text{BERT}_{\text{BASE(Ours)}}$	56.54	57.34	59.43	60.24

表 3-4　在 ROC 测试集上应用简化后的概念序列编码和概念图编码方法的端到端准确率。原始 = 基线模型，简化 = 简化后的概念序列编码方法，概念图编码 = 概念图编码方法。

3.4.3　实验结果和分析

1. 端到端结果

首先，我们展示了三种基线模型结合简化后的概念序列编码方法和概念图编码方法的端到端结果。然后，我们评估了在 ROC 上使用新数据集训练的其他模型。

表 3-4 和表 3-5 显示了所有三种典型的预训练故事表征模型均从我们的简化后的概念序列编码方法和概念图编码方法中受益。在表 3-4 中，SKBC 和 DSSM 分别通过简化后的概念序列编码方法实现了 3.43% 和 4.75% 的显著提升。$\text{BERT}_{\text{BASE}}$ 通过简化后的概念序列编码方法获得了 0.8% 的提升，通过概念图编码方法获得了 2.89% 的提升（与原始模型相比）。这

模型	原始 (%)	简化 (%)	概念图编码 (%)	简化 + 概念图编码 (%)
DSSM	54.30	57.83	54.35	58.53
SKBC	64.56	67.30	65.45	67.97
$\text{BERT}_{\text{BASE(Ours)}}$	56.88	58.02	59.79	60.97

表 3-5　在 $ROC_{\text{v}}1.5$ 测试集上应用简化后的概念序列编码和概念图编码方法的端到端准确率。

模型	准确率 (%)
DSSM(实现)	54.04
GMSA	61.20
CGAN	60.90
SeqMANN(实现)	59.74
SIMP(实现)	61.09
FES-LM(实现)	61.60
ISCK(实现)	62.21
GPT(实现)	63.46
SKBC(实现)	64.70
$\text{BERT}_{\text{BASE(Ours)}}$(实现)	56.54
$\text{BERT}_{\text{BASE}}$(实现)	61.46
$\text{BERT}_{\text{LARGE}}$(实现)	64.67
$\text{TransBERT}_{\text{BASE}}$(实现)	61.46
$\text{TransBERT}_{\text{LARGE}}$(实现)	61.89
SKBC+Simp+CE(Ours)	69.7
人类	100

表 3-6　在 ROC 上的故事结局预测实验结果，Simp 是简化后的概念序列编码，CE 是概念图编码。

是因为 BERT 包含了 Transformer 单元，它是一种注意力机制。BERT 可以从预训练中学习信息量大的权重。我们的简化后的概念序列编码方法甚至可以帮助减少 BERT 对信息量小的词的权重。DSSM+CE（CE 是概念图编码）的表现不如 DSSM，主要是因为 DSSM 是一个词袋模型，它将前四个句子作为一个整体进行建模。使用概念图编码时，我们必须将所有四个句子的嵌入求和，然后与 DSSM 的输出向量连接作为最终的表征。这样做不可避免地会丢失顺序信息。从表 3-5 中我们也可以得出相同的结论：简化和概念嵌入可以促进故事结局的预测。

表 3-6 展示了我们新训练和验证数据集上其他先前研究的结果（详见 3.4.2节）。大多数基线模型都是严格按照原论文中的设置，并使用我们提出的新训练数据实现的。$\text{BERT}_{\text{BASE(Ours)}}$ 使用 BookCorpus 重新训练了 $\text{BERT}_{\text{BASE}}$ 的语言模型。我们实现的 $\text{BERT}_{\text{BASE}}$ 和 $\text{BERT}_{\text{LARGE}}$ 使用我们的训练数据进行微调，并且尊重原始语言模型的参数设置。在之前的研究中 SKBC 有最好的报告结果，准确率为 64.7%。$\text{BERT}_{\text{LARGE}}$ 达到了 64.67% 的准确率，在所有基线中排名第二。这表明 BERT 在处理大量文本数据时

事件类型	未应用概念图编码的准确率 (%)	应用概念图编码的准确率 (%)
全词 (SKBC)	64.70	65.12
5-TUPLE	55.12	57.83
FES	60.12	63.66
Ours(Simp)	68.13	69.70

表 3-7　在 ROC 上对 SKBC 应用不同类型事件简化后的概念序列编码的效果（未应用概念图编码）。

事件类型	简化前	事件	简化后
5-TUPLE	10.02	1	3.43
FES	10.02	1.52	8.95
Ours(Simp)	10.02	3.78	4.90

表 3-8　采用不同事件类型进行简化对词序列长度的影响。简化前 = 简化前平均词数，简化后 = 简化后平均词数，事件 = 提取的平均事件数。

具有强大的学习表示能力。我们的 $\text{BERT}_{\text{BASE}}$ 表现不如基础版本，因为我们仅使用 BookCorpus 重新训练了语言模型。尽管更大的语料库，如维基百科，可能带来更好的结果，但我们只是想展示我们的简化后的概念序列编码和概念图编码方法的有效性。BERT 的不佳表现可能是由于训练数据中较少的偏差线索所致。这与 [98] 的工作一致，该工作对抗性地生成测试集，并导致结果急剧下降。采用我们方法的 SKBC 达到了 69.7% 的准确率，在我们的实验中表现最佳。它比我们测试的任何其他常用模型都表现更好。请注意，我们的实验不是为了证明某种特定算法的优越性，而是为了展示我们提出的故事表征方法（即简化后的概念序列编码和概念图编码）适用于多种模型。人类的表现为 100%，可以视为上限 [92]。所有结果都是基于 5 次独立运行的平均值。

2. **不同简化策略的比较**

理解故事需要理解事件序列。为了评估简化后的概念序列编码方法在故事结局预测任务中的有效性，我们比较了两种事件表征方式，即 5-TUPLE[107] 和 FES[102]，它们分别依赖于依存解析或语义角色标注 (SRL)。

5-TUPLE 将事件表示为五元组 (v, e_s, e_o, e_p, p)[6]，其中 v 是动词原形，不能为 *null*，e_s、e_o 和 e_p 分别代表主语、直接宾语和介词宾语的名词参数，p 是连接 v 和 e_p 的介词。另一种事件表征 FES 则联合模型化不同语义知识方面：框架[7]、实体和情感。与原论文不同，我们将这三个方面都以字符串形式而非向量表示（例如，情感标签 POSITIVE 而非情感的独热向量表示）。

表 3-7 显示了在相同 SKBC 模型架构上应用这些简化后的概念序列编码的效果。由于丢失过多信息，5-TUPLE 只达到了 55.12% 的准确率。如表 3-8 所示，简化后的平均词数仅为 3.43。虽然 FES 带来了更多信息，但提取框架的流程容易导致错误传播。

3. 概念嵌入的影响

表 3-7 同样表明，所有事件序列表征都能从结合 ConceptNet 上预训练的概念图编码中受益。Simp 和 Simp+CE 的结果显示，概念之间的关联能引入额外知识，这是无法通过语言模型直接学习到的。

4. 训练数据规模

图 3-3 对比了在增加的训练数据量（ROCS*(Tr)）下，不同方法的模型性能。首先可以看到，我们的两种方法（Simp 和 Simp+CE）即使在较少的训练数据下也能表现出色。事实上，仅使用 10% 的数据，它们就已经超过了 SKBC 在全部数据下的准确率。其次，我们观察到与 SKBC 相比，采用我们方法的模型在训练数据增加时提升更快，这一点从 10% 到 50% 数据量的斜率中尤为明显。最后，比较有无概念嵌入的效果，我们发现，在没有结构化知识支持的情况下，模型难以利用更多训练数据。

5. 训练时间

除了端到端准确率的提升外，简化后的概念序列编码方法还能将训练时间缩短至原来的三分之一。这种效率提升得益于句子中词汇数量的减少和词汇表的缩小。ROC 故事中共有 43,095 个独特词汇，而从 ConceptNet 中提取的简化关键词汇仅有 19,455 个，这大大减少了词汇表的大小。

[6]ROC 故事使用 Stanford CoreNLP 工具进行 5-TUPLE 解析。

[7]语义框架基于 PropBank 框架的语义角色标注注释。

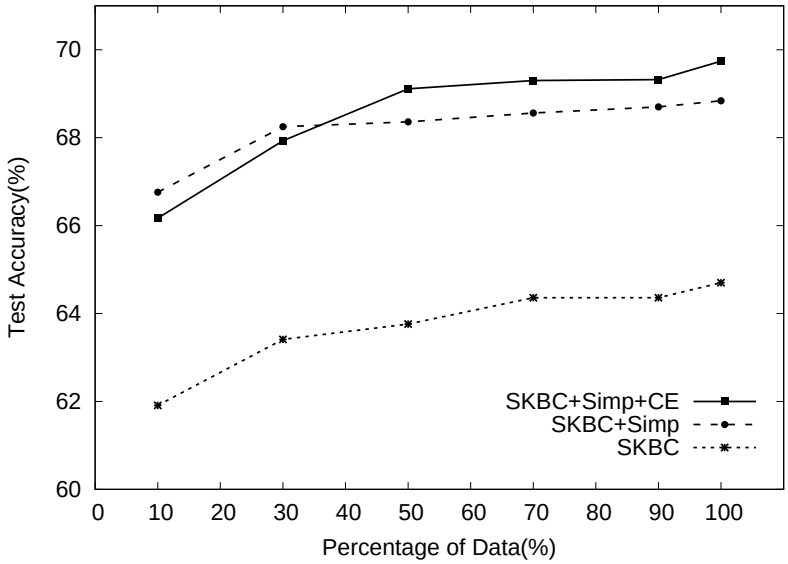


图 3-3　不同训练数据规模下的准确率。

3.5　本章小结

在本章中，我们专注于解决人工智能模型在理解和预测故事结局的任务中面临的一个核心挑战：如何提升模型的常识性推理能力。考虑到故事理解涉及广泛而复杂的常识知识，我们选择了故事结局预测这一特别具有挑战性的任务作为我们研究的焦点。为此，我们提出了一种新颖的方法论，旨在通过精确识别和分析故事中的关键常识概念和事件，从而使模型能够更深入地沿故事线索进行推理。这种方法的创新之处在于，它不仅关注故事的表意，而且深入挖掘故事结构中的隐含逻辑和关系。

我们的方法包括三个关键部分：首先是对故事句子的简化处理，以突出关键概念和事件；其次是基于这些关键元素构建更加丰富和细致的故事表示；最后是运用这种表示来预测故事结局。为了实现这一目标，我们结合了句子的概念序列编码和 ConceptNet 预训练的概念图编码，这一步骤对于捕捉故事中的深层关系至关重要。

在实验部分，我们使用了一个经过精心设计的数据集，旨在减少偏见和提高预测的准确性。我们的模型与多种现有的基线模型（如 DSSM、SKBC 和 BERT）进行了对比。实验结果表明，我们的方法在减少训练数据中的偏见和信息泄露方面表现优异。此外，我们还探讨了不同的简化策略和概

念嵌入的影响，发现结合 ConceptNet 的预训练概念嵌入能够显著提升模型性能。

最后，本章提出了未来研究方向，主要聚焦于如何显式表示故事中的常识关系，以提升模型对预测结果的解释能力。这种方法的发展不仅能增强模型的透明度，还能为未来的研究提供关于故事理解和推理机制的更深刻见解。

第四章　推理模型鲁棒性不足的可解释性分析

在 2.5节中，我们深入探讨了模型在处理常识性推理任务时表现的鲁棒性不足的现象以及这个现象的可解释性上。虽然这些模型在标准测试环境下展现出色，但它们在面对对抗性数据或与训练环境截然不同的情形时，却展现出显著的脆弱性。众多研究提出了一个关键假设，即这种鲁棒性不足可能源自模型对数据中特定结构元素的过度依赖，如偏见线索或虚假线索。这一假设背后的深层次原因在于数据集的构建方式，这导致模型倾向于学习数据特有的简化规则而非广泛适用的解决策略。

为了对这一假设进行验证，研究者们引入了“仅假设”测试和注意力图等工具。这些方法使研究者能够更深入地分析模型可能对数据中某些特定结构元素的过分关注。然而，这些方法在解释模型行为方面存在一定的局限性。例如，“仅假设”测试并不能充分展现模型在处理真实场景时的推理模式，因为它的测试数据结构与模型在训练阶段所接触的数据结构并不一致。同样，尽管注意力图被用来揭示模型的焦点，但相关研究表明，模型通过注意力机制突出显示的内容并不总是准确反映其决策的真实依据。鉴于此，深入探索数据偏见如何具体作用于模型的机制已成为一个迫切且重要的研究课题。

我们将在本章介绍两种创新的测试框架，用以深入分析并理解数据偏见对模型鲁棒性的影响。首先，我们提出的宏观层面测试框架，通过将测试数据划分为简单（easy）和困难（hard）两类，对模型在识别和处理虚假特征方面的能力进行量化评估。此方法揭露了模型对特定统计规律的过度依赖，帮助我们深入理解模型泛化能力的局限。

其次，我们引入了微观统计分析框架——ICQ（“I-see-cue”）框架，它通过多维特征划分和细致的性能分析，探究模型在不同特征上的准确性和分布表现。此外，我们还开发了一种直观的可视化工具，旨在更有效地识别和理解模型性能差异的根源。

总体而言，本研究的主要贡献在于这两种高度创新的分析框架，它们为深入解析常识性推理模型的鲁棒性不足问题提供了新的路径。

4.1　概述

深度神经网络模型在各种自然语言理解（NLU）任务中取得了显著的成就，这些任务包括自然语言推理 [11, 146]、论证分析 [98]、常识推理 [92, 120, 158]、阅读理解 [67]、问题回答 [140] 和对话分析 [79]。这些任务经常采用多项选择框架，正如在斯坦福自然语言推理（SNLI）数据集的例子中所示（见 例子 1 ）。然而，近期研究 [45, 123, 108, 62] 揭示了一些问题，特别是在这些模型对微小变化高度敏感的背景下，我们需要一个更为稳健且精确的评估机制。

Example 1. SNLI *数据集中的自然语言推理示例，正确答案以斜体标出*

Premise: A swimmer playing in the surf watches a low flying airplane headed inland.

Hypothesis: Someone is swimming in the sea.

Label: a）Entailment. b）Contradiction. c）Neutral.

在处理类似 例子 1 这样的任务时，人类通常依赖于前提和假设之间的逻辑关系。与此相反，一些 NLP 模型可能会绕过这种逻辑推理，转而关注数据集中嵌入的偏见，特别是在假设中的偏见 [95, 126]。这些偏见，如情感或表层 n-grams，可能为正确预测提供误导性线索。

当这些偏见在训练和测试数据集中普遍存在，并在预测上保持类似的分布时，我们将其称为“人为虚假线索”（如图 4-1 所示）。例如，在 例子 1 中，某些模型可能过分依赖 “someone” 一词来做出判断。这种情况下，当这些线索缺失或改变时，可能会显著影响模型的性能，凸显了识别这些线索以提高模型鲁棒性的重要性。

本研究旨在探讨如何在各类自然语言理解任务中识别和解决虚假线索的问题。我们特别关注这些线索在数据集制作过程中如何产生，以及模型在训练过程中如何学习并依赖这些线索。我们提出的方法旨在揭示这些线索，并帮助我们深入理解它们如何影响模型的性能和推理能力。为此，我们将从宏观和微观两个角度分析模型是否受到虚假线索的影响。

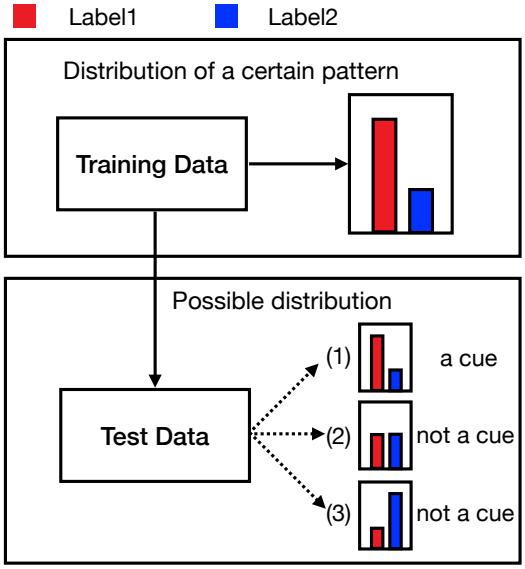


图 4-1　线索的示例。

宏观角度分析

我们提出了一种新方法，用以从宏观角度分析自然语言理解（NLU）任务中虚假线索的问题。在探索人类处理这类问题的方式时，我们发现人们通常会仔细分析问题的前提和假设之间的逻辑关系。然而，之前的研究 [95, 126] 显示，许多自然语言处理（NLP）模型在处理类似问题时，往往只考虑假设部分，忽视了逻辑关系的深入分析。这一现象在很大程度上归咎于数据集中人工制作的假设存在的瑕疵。

虽然“仅假设”测试在理论上是用于识别此类问题的有效方法，但它常依赖于特定的、如 BERT [35] 这样的模型，这就需要进行模型的重新训练。更重要的是，“仅假设”测试并不能充分展现模型在处理真实场景时的推理模式，因为它的测试数据结构与模型在训练阶段所接触的数据结构并不一致。

在我们的研究中，我们开发了一种轻量级框架，专注于识别多项选择自然语言推理数据集中的关键且影响力显著的线索。尽管不是所有多项选择题都包含前提、假设和标签，但我们在第 4.4 节中详细介绍了如何将这些元素标准化。我们的框架以词汇为基础，利用它们作为构建线索的关键特征，因为词汇在大多数现代机器学习方法中是自然语言建模的基石。此外，

即便是复杂的语言特征，如情感、风格和观点，也通常是基于词汇特征。

我们的实验展示了，基于词汇的线索在检测数据集中的统计偏差方面，与资源消耗更大的 “仅假设” 方法同样有效。这一发现对于理解和改进模型的推理能力至关重要。

此外，我们还设计了一种方法，将测试数据基于问题是否能仅通过线索特征被正确回答的能力，划分为简单和困难两个子集。数据集中困难部分的相对大小是衡量其整体质量的重要指标，其中困难部分占比越大，表明数据集质量越高。这种划分同时也是一种 “压力测试”，可用于评估模型的真实推理能力。两个子集间的分类准确率差异揭示了模型在不同情境下的潜在薄弱环节。

综上所述，这个宏观层面的方法的研究贡献有两个。首先，我们开发了一种轻量级但有效的方法，用于评估和识别自然语言理解数据集中的统计偏见和虚假线索（4.4节、4.6.2节）。另外，我们通过将测试数据划分为简单和困难两部分，根据信息泄露程度来评估模型的真实推理能力，从而量化比较这两部分的性能差异（4.6.2节）。

微观角度分析

在微观层面上，我们的研究着重于区分数据集中预先嵌入的线索与模型在训练过程中所学习的线索。传统的偏见检测与缓解工具，如 AI Fairness 360 工具包 [6]，尽管主要聚焦于数据集中的偏见，但在处理模型训练过程中产生的偏见时，它们往往显示出局限性。

现有的方法，例如 “仅假设” 测试和 CheckList，虽然能够揭示模型的脆弱性，却不是专门为识别模型学习的线索而设计的。具体来说，“仅假设” 测试能够突显出数据集中的某些问题，如仅凭假设就能得出正确答案的情况，但这并未涵盖模型在训练和预测过程中使用的完整数据上下文，因此无法全面反映模型的整体能力。

同样地，虽然 CheckList 采用了软件工程中的黑盒测试原则，通过对预定义语言特征的额外压力测试案例来审视模型的弱点，但它对精心设计的模板的依赖限制了其应用范围。这种方法能够揭示模型的脆弱性，但它并不能清楚地揭示模型实际从数据中学到的具体知识或特征。

尽管我们的宏观架构能够在一定程度上分析模型受到数据中虚假线索

的影响程度，但它难以具体指出受到哪些特定线索的影响。

为了克服这些局限，我们引入了 ICQ（“I-see-cue”，意为 “我发现线索”），一个灵活的统计分析框架[8]。ICQ 的设计哲学与传统方法完全不同，它能够在不需要额外测试案例的情况下，有效识别多项选择 NLU 数据集中的偏见。采用黑盒测试方法，ICQ 能够评估模型如何利用这些偏见，从而为理解 NLU 任务中的偏见提供全面且深入的见解。

我们通过在不同的 NLU 数据集上部署 ICQ 来验证其有效性，并探索模型在训练期间可能学习到的线索。ICQ 的应用极大地促进了我们对于诸如 ChatGPT 这类模型如何学习潜在线索的深入理解，并提供了选择合适提示的示例，为模型的优化提供了重要的指导。

综上所述，我们的微观研究主要提供了以下几点贡献：

- 我们推出了 ICQ，这是一种轻量级且功能强大的工具，专门用于识别 NLU 数据集中的统计偏见和线索。我们还提出了简单而有效的测试方法，以量化和直观地评估模型是否在其预测中利用了虚假线索。
- 我们对十个流行的 NLU 数据集和四个模型进行了全面的统计偏见问题评估，从而证实了先前的发现并揭示了新的见解。我们还提供了一个在线演示系统，展示结果并邀请用户评估他们自己的数据集和模型。
- 通过一个案例研究，我们深入探讨了 ChatGPT 如何学习潜在偏见，并为其实际应用提供了宝贵的建议。

4.2　相关工作

我们的工作涉及三个研究方向：虚假特征分析、偏见计算和数据集过滤。

虚假特征分析

近年来，虚假特征分析在自然语言处理（NLP）领域获得了越来越多的关注。研究表明，许多 NLP 模型能够在多项选择题形式的自然语言理解

[8]代码和数据集可在 https://github.com/flora336/icq 获取

问题中取得良好表现，有时甚至无需深入理解问题的核心内容。这类现象在某些研究中被称为“仅假设”测试 [131, 136, 158]。此外，研究发现，这些模型对于假设中的微小但语义重大的变化并不敏感，推测这种“仅假设”表现可能源于假设中的词语与答案标签之间的简单统计关联 [123]。

虚假特征可分为词汇化和非词汇化两类 [11]。词汇化特征主要包括 n-gram 词标和交叉 n-gram 词标，而非词汇化特征则涉及词汇重叠、句子长度以及前提和假设之间的 BLEU 分数。词汇化分类的细化，如否定、数值推理和拼写错误，也在文献中得到了探讨 [95]。另一方面，非词汇化特征如词汇重叠、子序列和成分被进一步细化，同时也考虑了句法结构重叠 [82]。此外，[123] 为未见过的词汇标记提供了额外的词汇化特征。

偏见计算

偏见计算关注于量化线索的严重性。一些研究尝试通过仅假设训练或从嵌入中提取与特定标签相关的特征来隐式编码线索特征 [24, 48, 155]。其他方法则采用统计指标来计算偏见。例如，[157] 使用在特定标签下观察到某个词的概率来对词进行偏见排名。LMI（局部互信息）被用于评估线索并在某些模型中进行重权 [126]。然而，这些研究并没有充分解释为什么选择这些指标而非其他指标。此外，[119] 提供了一种测试数据增强方法，但它没有评估偏见的程度。

数据集过滤

数据集过滤是一种通过减少数据集中人为制作特征来提高数据集质量的方法。事实上，本研究评估的 SWAG 和 RECLOR 等数据集就是采用这种过滤方法的变体生成的。这种方法通过反复扰动数据实例，直到目标模型无法再很好地拟合所得数据集为止。一些方法不是通过预处理数据来去除偏见，而是在训练过程中根据每个阶段的决策排除带有偏见的样本 [155]。[69] 探讨了基于模型的数据集线索减少算法，并设计了一种使用迭代训练的算法。这种方法比人工标注更通用、更高效，但它严重依赖于所使用的模型。不幸的是，不同模型可能会捕捉到不同的线索，因此这些方法可能并不完全。

本节涉及的三个主要研究方向对于当前人工智能领域至关重要。虚假

特征分析揭示了模型可能依赖的非理想特征，偏见计算提供了量化这些特征的方法，而数据集过滤旨在减少数据集中的人为制作特征，以提高其整体质量。这些研究方向为理解和改善 NLP 模型在处理自然语言任务时的表现提供了重要的视角和工具。

4.3 任务表述

在本节中，我们将介绍如何将自然语言推理任务的数据集 X 中的问题实例 x 进行标准化表述。对于数据集 X 中的每个问题实例 x，我们定义其为以下形式：

$$x = (p, h, l) \in X, \tag{4.1}$$

其中 p 代表推理的上下文，即给定的情景或陈述，类似于例子 1中的 “前提”；h 是针对上下文 p 的假设或断言；$l \in \mathcal{L}$ 是一个标签，用于描述上下文 p 和假设 h 之间的关系类型。

在不同的自然语言推理任务中，关系集合 $\mathcal{L}$ 的大小可能会有所不同。通常，这种关系集合能够涵盖任务所需表达的所有可能关系类型。例如，在一个典型的自然语言推理问题中，包含一个前提（p）、一个假设（h）和一个描述前提与假设之间关系的标签（l）。对于描述蕴涵、矛盾和中立三种不同关系的场景，关系集合 $\mathcal{L}$ 的大小为 $|\mathcal{L}| = 3$。

我们认为，这种通用形式能够适用于大多数具有区分性的自然语言推理任务。在第 4.4.3节中，我们将进一步讨论如何将不同类型的自然语言推理任务转换为这种标准化表述，以便于进行更系统和统一的分析。这种标准化的方法不仅有助于简化任务的理解和实现，还为后续的模型训练和评估提供了清晰的指导。

4.4 宏观偏差识别与评估架构

在宏观架构中，我们专注于使用统计特征来评估数据集内的偏差。首先，我们对自然语言推理任务进行通用化的表述。随后，通过计算与每个标签关联的词汇频率，我们设计了衡量词汇和标签相关性的指标，称之为“线索分数”。这些分数揭示了潜在的偏差模式。我们进一步利用简单的统计模型来汇总这些分数，并据此作出预测。最后，我们展示了如何利用这些

快速预测方法，将数据集划分为“简单”和“困难”两部分。

4.4.1 线索度量

对于给定的数据集 X，我们收集其中所有词汇的集合 $\mathcal{H}$。线索度量是衡量特定词在不同标签下出现频率的不平衡性。对于集合中的词 w，我们用以下八种方法中的一种来计算其线索分数 $f_{\mathcal{F}}^{(w,l)}$。我们将这些度量分为两大类：前四种基于简单统计数据，后四种则涉及欧几里得空间中的角度概念。

设 $\mathcal{L}' = \mathcal{L} \setminus \{l\}$，我们定义

$$\#(w, \mathcal{L}') = \sum_{l' \in \mathcal{L}'} \#(w, l'). \tag{4.2}$$

频率（Freq）最基础的度量是词汇和标签的共现频率，其中 #() 表示简单计数。这个度量旨在捕捉词汇在特定标签下的原始频率：

$$f_{Freq}^{(w,l)} = \#(w, l). \tag{4.3}$$

相对频率（RF）相对频率在频率的基础上进行扩展，考虑了词汇在所有标签下的总频率。公式如下：

$$f_{RF}^{(w,l)} = \frac{\#(w, l)}{\#(w)}. \tag{4.4}$$

条件概率（CP）标签 l 在词 w 下的条件概率是另一种重要的度量，用于捕捉特定条件下的概率分布：

$$f_{CP}^{(w,l)} = p(l|w) = \frac{\#(w, l)}{\#(w)}. \tag{4.5}$$

点互信息（PMI）点互信息是信息论中用于衡量词汇和标签关联强度的常用指标。当词汇和标签共现的频率超出独立出现的预期频率时，PMI 值较高。其定义如下，其中 $p(w)$ 和 $p(l)$ 分别是 w 和 l 的概率，$p(w,l)$ 是

它们的联合概率：

$$f_{PMI}^{(w,l)} = \log \frac{p(w,l)}{p(w)p(l)}. \tag{4.6}$$

局部互信息（LMI）局部互信息是 PMI 的变体，它通过词汇和标签的联合概率加权 PMI，给予频繁出现的词汇-标签对更多重视：

$$f_{LMI}^{(w,l)} = p(w,l) \log \frac{p(w,l)}{p(w)p(l)}. \tag{4.7}$$

比率差异（RD）比率差异衡量词汇-标签比率与整体标签比率之间的绝对差异，有助于识别特定标签不成比例相关的词汇：

$$f_{RD}^{(w,l)} = \left| \frac{\#(w,l)}{\#(w,\mathcal{L}')} - \frac{\#(l)}{\#(\mathcal{L}')} \right|. \tag{4.8}$$

角度差异（AD）角度差异类似于比率差异，但通过使用反正切函数，考虑了比率之间的非线性关系，使度量对异常值更加稳健：

$$f_{AD}^{(w,l)} = \left| \arctan \frac{\#(w,l)}{\#(w,\mathcal{L}')} - \arctan \frac{\#(l)}{\#(\mathcal{L}')} \right|. \tag{4.9}$$

余弦（Cos）余弦度量将 $v_w = [\#(w,l), \#(w,\mathcal{L}')]$ 和 $v_l = [(\#(l), \#(\mathcal{L}')]$ 视为二维空间中的两个向量。如果 v_w 和 v_l 共线，则意味着 w 没有提供误导性信息。否则，w 可能是虚假线索，因为它倾向于与某个特定标签 l 更频繁共现。这个度量通过几何方式量化了词汇-标签关系的相似性：

$$f_{Cos}^{(w,l)} = \cos(v_w, v_l). \tag{4.10}$$

加权功率（WP）加权功率结合了余弦度量和基于频率的加权，强调了高频词汇的重要性，帮助优先考虑对模型影响更大的线索：

$$f_{WP}^{(w,l)} = (1 - f_{Cos}^{l}) \#(w)^{f_{Cos}^{l}}. \tag{4.11}$$

总结来说，我们可以将词 w 相对于标签 l 的线索分数表示为 $f^{(w,l)}$，省略方法下标 $\mathcal{F}$。这些度量从多个角度提供了关于词汇和标签之间关联性的洞见，有助于识别潜在的虚假相关性。

4.4.2 聚合方法

我们可以使用简单的方法 $\mathcal{G}$ 来聚合问题实例 x 中的词语线索分数以作出预测。这些方法旨在易于实现和计算效率高，鉴于线索特征的低维性。

平均值和最大值

预测标签最直接的方法是选择具有最高平均或最大线索分数的标签。

$$\mathcal{G}_{\text{average}} = \arg\max_{l} \left(\frac{\sum_{w} f^{w,l}}{|x|} \right), \quad l \in \mathcal{L}, w \in \mathcal{W} \tag{4.12}$$

$$\mathcal{G}_{\max} = \arg\max_{l} \left(\max_{w} (f^{w,l}) \right), \quad l \in \mathcal{L}, w \in \mathcal{W} \tag{4.13}$$

线性模型

为了更好地利用线索分数进行预测，我们采用了两种简单的线性模型：SGDClassifier 和逻辑回归。模型的输入是问题实例 x 中每个标签的线索分数的连接向量：

$$\text{input}(x) = [f^{(w_1,l_1)}, \ldots, f^{(w_d,l_1)}, f^{(w_1,l_2)}, \ldots, f^{(w_d,l_2)}, \ldots, f^{(w_1,l_t)}, \ldots, f^{(w_d,l_t)}] \tag{4.14}$$

在实际应用中，输入向量被填充到相同长度。线性模型的训练损失为：

$$\hat{\phi}_n = \arg\min_{\phi_n} (\text{loss}(\mathcal{G}_{\text{linear}}(\text{input}(x); \phi_n))) \tag{4.15}$$

损失是根据黄金标签 l_g 和预测标签 $\mathcal{G}_{\text{linear}}(\text{input}(x); \phi_n)$ 计算的。ϕ_n 表示在 $\mathcal{G}_{\text{linear}}$ 中最小化 l_g 的损失的最优参数。

4.4.3 多项选择题的标准化转换

到目前为止，我们关注的是具有固定选项集的多项选择题（MCQs）。然而，一些语言推理任务涉及到具有非固定选项的 MCQs，如 ROC 数据集。在这些情况下，我们可以将原始故事分为两个统一实例，$u_1 = (\text{context}, \text{ending1}, \text{false})$ 和 $u_2 = (\text{context}, \text{ending2}, \text{true})$。我们预测每个实例的标签概率，$\mathcal{G}(\text{input}(u_1); \phi)$ 和 $\mathcal{G}(\text{input}(u_2); \phi)$，并选择具有更高概率的结局作为预测。

4.4.4 数据集难度层级分类方法

本小节旨在介绍一种区分数据集中问题难度的方法，通过预测结果将问题分类为简单或困难。这种分类依赖于一个聚合模型，其核心在于利用词频特征进行有效预测。该方法不仅有助于评估模型在处理各种难度问题时的性能，而且能够揭示数据集中的潜在信息泄露。

具体来说，我们对数据集中的每个问题进行预测。如果问题被模型正确预测，我们则将其标记为“简单”；如果预测失败，则标记为“困难”。通过这一流程，我们可以将整个数据集有效地分为两个子集，即简单问题集和困难问题集，进而更深入地理解模型的性能及其在不同情境下的应用潜力。我们可以按照以下步骤将数据集分为简单和困难的问题，参见算法2：

Algorithm 2 将数据集分为简单和困难部分

Require: dataset $\mathcal{D}$, aggregation model $\mathcal{G}$, number of folds n, number of iterations k

Initialize a counter C for each question q in $\mathcal{D}$ as 0
for $i = 1$ to k do
    Perform n-fold random split of $\mathcal{D}$ into $P_1, P_2, \ldots, P_n$
    for $j = 1$ to n do
        Train $\mathcal{G}$ on $P_1, P_2, \ldots, P_{j-1}, P_{j+1}, \ldots, P_n$
        Test $\mathcal{G}$ on P_j
        for each question q in P_j do
            if q is correctly predicted then
                Increment $C[q]$
            end if
        end for
    end for
end for
Label each question q in $\mathcal{D}$ as easy if $C[q] > \frac{k}{2}$, otherwise label as hard
Split $\mathcal{D}$ into easy and hard parts based on the labels

具体来说分成以下六个步骤：

1. 对数据进行 n 折随机划分。
2. 在 $n-1$ 部分上训练聚合模型。
3. 采用循环赛方式，在剩余的部分上测试聚合模型。
4. 每个问题测试一次后，根据模型的预测结果为其分配“简单”或“困难”标签。
5. 重复此过程多次（例如 k 次迭代），并根据 k 次迭代中的多数标签为每个问题标记。
6. 最后，根据每个问题的最终标签，将数据集分为简单和困难两部分。

此算法使我们能够更好地理解模型在不同难度级别上的表现，并仅使用统计特征分析数据集的信息泄露。通过测量词语和标签之间的相关性，聚合线索分数以进行预测，并根据难度将数据集进行分割，我们可以评估模型的真实推理能力，并解决由虚假线索引起的潜在问题。这个过程旨在计算效率高、易于实现，适用于广泛的自然语言推理任务。

4.5 微观偏差识别与评估架构

在微观层面上，偏差识别和评估涉及对数据集的深入分析，以发现和评估潜在的偏见和不平衡。这种分析要求我们关注数据的具体细节，如语言特征和模型对这些特征的响应。本节首先介绍了作为分析基础的关键语言特征，然后描述了我们的 ICQ 框架，它是对这些特征进行系统化检查和评估的工具。

4.5.1 相关语言特征

在微观偏差分析的第一步中，我们关注数据集中的特定语言特征，这些特征可能揭示或促成了偏见的形成。如先前的研究 [95, 62] 所示，我们考虑以下关键语言特征：

词汇：数据集实例中前提或假设中特定词汇的存在。

情感：实例的情感值，计算为单个词语情感极性的总和。

时态：实例的时态特征（过去、现在或未来），由根动词的词性标记决定。

否定：实例中负面词汇（例如 “no”、“not” 或 “never”）的存在，通过依存解析确定。

重叠：至少有一个词（排除停用词）同时出现在前提和假设中。

命名实体识别（NER）：实例中命名实体（例如 PER、ORG、LOC、TIME 或 CARDINAL）的存在，使用 NLTK NER 工具检测。

拼写错误：实例中至少存在一个拼写错误，使用预训练的拼写模型识别。

对于多项选择数据集，除了重叠外的所有特征仅应用于假设。

4.5.2　ICQ 框架

在确定了关键的语言特征后，我们引入 ICQ 框架（如 图 4-2 所示），它是一个三阶段的过程，包括数据提取、线索发现和模型探测。ICQ 框架专注于系统地分析这些语言特征，以揭示潜在的偏差和不一致。

数据提取阶段（Data Extraction**）**：在这一阶段，我们从数据集中提取包含特定语言特征 f 的实例。这一步骤是识别潜在偏见的基础。

线索发现阶段（Cue Discovery**）**：紧接着，我们对提取出的实例进行深入分析，以识别其中的潜在线索。这涉及到对每个预定义特征的详细检查。

模型评估阶段（Model Probing**）**：最后，我们进行两项测试：“准确性测试”（Accuracy Test）和 “分布测试”（Distribution Test）。这些测试旨在评估模型对不同特征的反应，以及这些反应是否揭示了模型的任何偏见或不平衡

我们将在下面更详细地讨论这些阶段：

1. 数据提取阶段

在确定了关键的语言特征之后，我们的系统首先执行的任务是构建数据提取器。对于每个特征值 f，提取器的角色是处理整个数据集并识别与该特定特征值相关联的实例。具体来说，如果数据集中的某个实例包含了这个特征，那么这个实例就会被选中，并归入包含该特征的子集中。这一步是识别特征在数据集中分布和表现的基础。

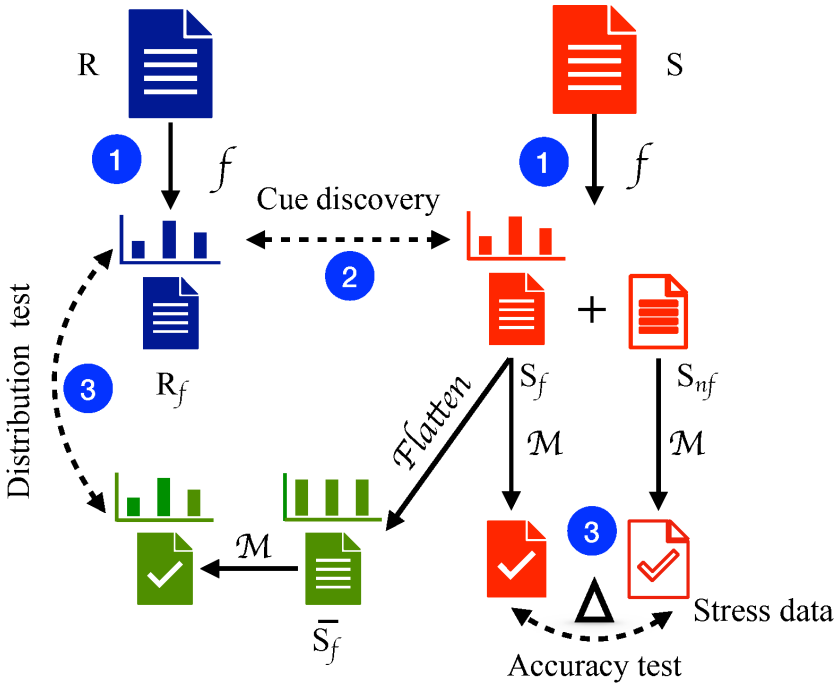


图 4-2　ICQ 工作流程。①：数据提取阶段；②：线索发现阶段；③：模型探测阶段。f= 特定特征，R= 训练数据，S= 测试数据，R_f= 提取的训练数据，S_f= 提取的测试数据，S_{nf}= 不包含特征 f 的剩余测试数据，$\overline{S_f}$= 平衡化测试数据（flatten），m= 特定模型。

2. 线索发现阶段

接下来，我们针对每个特征 f，将其对应的提取器应用于数据集 X 的训练和测试部分，分别标记为 R 和 S，如图 4-2 所示。这导致了基于特征的实例聚类，形成了训练实例子集 R_f 和测试实例子集 S_f。一个特征被视为潜在的数据集线索，当且仅当它同时存在于训练和测试数据中。

为了计算提取集的标签分布偏差，我们使用均方误差（MSE）和 Jensen-Shannon Divergence（JSD）[74]。线索得分反映了数据集 X 针对特征 f 的偏见程度。

均方误差（MSE）计算方法如下：

$$MSE(F) = \frac{1}{|\mathcal{L}|}\sum_i (y_i - \overline{y_i})^2 \tag{4.16}$$

其中，y_i 表示在提取数据集 F 中标签 l_i 的实例数，而 $\overline{y_i}$ 是每个标签的平均

实例数。较大的 $MSE(F)$ 意味着更尖锐的标签分布，指示了更显著的偏见。

如果提取的训练集 R_f 和提取的测试集 S_f 展示出类似的偏见，那么它们分布之间的 JSD 将会很小。JSD 的计算公式为：

$$JSD = \frac{1}{2}\left(Q(R_f) \parallel A\right) + \frac{1}{2}\left(Q(S_f) \parallel A\right), \tag{4.17}$$

其中 $A = \frac{1}{2}\left(Q(R_f) + Q(S_f)\right)$。函数 $Q()$ 表示提取数据集的标签分布。

最后，我们定义线索得分，用以衡量数据集 X 针对特征 f 的偏见程度：

$$cue(f, X) = \frac{MSE(R_f)}{\exp(JSD(R_f, S_f))} \tag{4.18}$$

这个得分结合了标签分布的不均衡（MSE）和在训练与测试集之间分布的一致性（JSD），从而提供了一个全面的偏见指标。

3. **模型测试阶段（**Model Probing**）**

在前两个结算，我们探讨了数据集 X 可能受特定线索 f 影响的可能性。然而，关键的问题在于，基于这个数据集训练出的模型是否真的会利用这一线索。这不仅与数据集的特性有关，而且还取决于模型的结构和训练机制。为了深入探索这一点，在第三阶段介绍了一套框架，用于探测和评估在有偏数据集上训练的模型是否以及在多大程度上利用了特定线索 f。此框架包括两种测试方法：准确性测试和分布测试。

准确性测试

准确性测试旨在评估模型在含有和不含有特定特征 f 的数据子集上的表现差异。这种比较能够揭示模型在面对不同数据情景时的泛化能力。具体来说，如果模型在包含特定特征 f 的数据子集上表现更佳，这通常意味着模型已经学会利用这一特征进行预测。

在准确性测试中，我们专注于测量模型 M 在两个测试集上的表现：一个是包含特征 f 的测试集 S_f，另一个是不包含特征 f 的测试集 S_{nf}。我们通过以下公式计算这两个测试集上准确率的差异：

$$\Delta Acc(f) = acc(S_f) - acc(S_{nf}) \tag{4.19}$$

这里的 ΔAcc 值揭示了模型在处理包含和不包含特定特征的数据子集时性能的变化。ΔAcc 的正值或负值表明模型在比较包含和不包含特定特征的数据子集时性能变化的方向。正值表明模型利用该特征进行预测，而负值则意味着模型在泛化或对该特征的敏感性上存在困难。ΔAcc 的绝对值大小反映了模型性能受到特定特征存在或缺失的影响程度。较大的绝对值表明模型更依赖或对该特征更敏感，而较小的绝对值则表明模型更健壮，对特征的存在或缺失影响较小。

分布测试（Distribution Test）

分布测试着重于通过视觉化的方式检查特定特征的分布变化如何影响模型的预测性能。这一测试的第一步是在提取的测试集 S_f 中创建一个 “平衡化” 测试集 $\overline{S_f}$。我们通过平衡化标签分布来实现这一点，即从每个标签类别中随机移除实例，直到所有类别的实例数量相等。这一过程主要包括以下几个步骤：

- 1）确定最小标签实例数量：首先检查所有标签类别的实例数量，找到其中最少的实例数量，这将成为我们平衡化的目标。
- 2）调整其他标签的实例数量：对于实例数量超过最小类别的标签，我们随机移除多余的实例，使其数量与最小类别相等。这一过程独立地应用于每个标签类别。
- 3）创建平衡化的测试集：经过上述处理，我们得到了一个每个标签实例数量相同的平衡化测试集。这种均衡对于后续测试至关重要，因为它确保了评估模型时对所有标签的公平对待。

通过这种方法，我们能够有效地评估模型对不同标签类别的处理能力，尤其是在面对均衡分布的数据时的表现。

接下来，我们将模型应用于这个平衡化的压力测试集并获取预测结果。然后，我们比较这些预测结果的标签分布与提取的训练数据 R_f 的标签分布。这样做的目的是检查模型是否在预测时偏向于特定标签，尤其是当这

些标签在训练数据中与特定线索相关联时。即使测试集的输入已经被平衡处理，模型可能仍然会放大这种分布上的偏见。我们希望通过这种比较来识别模型的潜在偏差。

综上所述，准确性测试和分布测试在评估模型对特定特征的敏感性和反应上互为补充。分布测试专注于特征分布的变化对模型性能的影响，而准确性测试更关注模型在处理含有和不含有特定特征的数据子集时的表现。通过结合这两种方法，我们能够从不同角度全面评估模型对特定特征的反应。如果两种测试都显示模型对某个特征高度敏感，我们可以更加自信地推断模型确实在利用该特征进行预测。

4.6　实验

在本节中，我们会首先介绍我们探究的任务和相对应的数据集，而后实验的结果和分析我们会根据前面的方法的介绍也分成两部分，就是宏观结果分析和微观结果分析。

4.6.1　实验数据集

本研究的实验覆盖了 12 个相关的数据集，旨在涵盖两大类任务：自然语言推理分类任务和多项选择问题。这些数据集在规模、来源、结构和挑战性方面各具特色，提供了一个全面的研究背景。如表 4-1 所示，我们展示了数据集的分类和一些实例。表 4-2 中详细列出了这 12 个数据集的关键特征。

大部分数据集基于人工编写的假设，但 CQA[140] 和 SWAG[158] 则由基于 LSTM 的语言模型生成。数据集不仅在规模上不同，来源和构成方式也各有特点。部分数据集（如 STS[126] 和 SWAG）引入了对抗性实验，增添了数据多样性和挑战性。

自然语言推理分类任务数据集

在自然语言推理任务中，模型需要判断给定文本对之间的逻辑关系。本研究中包含的自然语言推理数据集有：

<table>
<tr><th>Task Name</th><th>Datasets</th><th colspan="4">Example</th></tr>
<tr><td></td><td></td><td>Original</td><td>“Premise”</td><td>“Hypothesis”</td><td>label</td></tr>
<tr><td>Natural Language Inference</td><td>SNLI, MNLI, QNLI</td><td>(SNLI) Premise: A woman and a child holding on to the railing while on trolley.
Hypothesis: The people are not holding on anything.
Label: contradiction</td><td>A woman and a child holding on to the railing while on trolley .</td><td>The people are not holding on anything.</td><td>contradiction</td></tr>
<tr><td>Argumentation</td><td>ARCT, ARCT_adv</td><td rowspan="2">(ARCT) Reason: Milk isn’t a gateway drug even though most people drink it as children.
Claim: Marijuana is not a gateway drug.
Warrant 1: Milk is similar to marijuana.
Warrant 2: Milk is not marijuana.</td><td>Milk isn’t a gateway drug even though most people drink it as children. Marijuana is not a gateway drug.</td><td>Milk is similar to marijuana.</td><td>true</td></tr>
<tr><td>Reading Comprehension</td><td>RACE, RECLOR</td><td>Milk isn’t a gateway drug even though most people drink it as children. Marijuana is not a gateway drug.</td><td>Milk is not marijuana.</td><td>false</td></tr>
<tr><td>Commonsense Reasoning</td><td>ROCStory, COPA, SWAG</td><td rowspan="3">(COPA) The woman hummed to herself. What was the cause for this?
Alternative1: She was in a good mood.
Alternative2: She was nervous.</td><td>The woman hummed to herself. What was the cause for this?</td><td>She was in a good mood.</td><td>true</td></tr>
<tr><td>Question Answering</td><td>CQA</td><td rowspan="2">The woman hummed to herself. What was the cause for this?</td><td rowspan="2">She was nervous.</td><td rowspan="2">false</td></tr>
<tr><td>Dialogue Analysis</td><td>Ubuntu</td></tr>
</table>

表 4-1　数据集示例和归一化版本。

- SNLI[11]：拥有 570K 个实例，通过人类标注构建而成，旨在测试模型理解自然语言的能力。
- QNLI[146]：含有 11K 个问题-答案对，主要通过众包方式收集，挑战模型在问题回答方面的性能。
- MNLI[151]：具有 413K 个实例，也是通过众包方式构建，用以评估模型在不同文体和话题上的推理能力。

这些数据集的共同点在于它们都要求模型从给定的前提和假设中推断出正确的结论。

多项选择问题数据集

多项选择问题要求参与者（或模型）从多个选项中选择一个正确答案。本研究涉及的相关数据集包括：

- ROC[92] 和 COPA[120]：分别包含 3.9K 和 1K 个实例，这些数据集主要通过众包方式收集，用于评估模型在故事理解和因果推理方面的能力。

数据集	数据大小	数据来源	AE	人类表现 (%)
ROC	3.9k	CD	No	100.0
COPA	1k	CD	No	100.0
SWAG	113k	LM	Yes	88.0
SNLI	570K	CD	No	80.0
QNLI	11k	CD	No	80.0
MNLI	413k	CD	No	80.0
RACE	100k	CD	No	94.5
RECLOR	6k	CD	No	63.0
CQA	12k	CD	No	88.9
ARCT	2k	CD	No	79.8
STS	4k	CD	Yes	-
Ubuntu	100k	随机选择	No	-

表 4-2　数据集中假设收集的方法。AE = **对抗性实验**, LM = **语言模型**, CD = **众包**, **人类表示人类在数据集上的表现。**

- SWAG（含 113K 实例）和 Ubuntu[79]（包含 100K 实例）：分别由基于 LSTM 的语言模型和随机选择方式生成，用于测试模型在生成连贯性和上下文理解方面的性能。
- RACE[67]（含 100K 实例）、RECLOR[157]（含 6K 实例）、CQA（含 12K 实例）、ARCT[46]（含 2K 实例）及 STS（含 4K 实例）：这些数据集通过众包方式收集，每个都提供独特的挑战，如文本理解、逻辑推理和论证分析。

4.6.2　宏观结果分析

我们进行了实验，以展示我们框架在两个方面的有效性：首先，我们应用我们的方法来检测线索，6 种不同任务的 12 个**数据集**（表 4-1 所示）中量化信息泄露量。其次，我们评估了一些流行的 NLP **模型**在被分为简单和困难两部分的原始测试集上的真实推理能力。

1. 数据集线索检测效果评估

我们的线索检测效果评估实验分为以下部分：量化信息泄露的方法论、线索评估技术和度量标准、与数据集中偏差的揭示。

量化信息泄露的方法论

在本研究中，我们专注于量化数据集中信息泄露或偏差的程度。为此，我们开发了一种新的度量标准，命名为 $\mathcal{D}$，其定义为 $\mathcal{D} = \text{Acc} - \text{Majority}$。在这个公式中，Acc 表示模型仅基于虚假线索的准确率，而 Majority 代表通过多数投票机制得到的准确率。

这一度量标准的核心在于评估和比较两种不同情境下的准确率，从而揭示数据集中存在的潜在线索。具体来说：

- $\mathcal{D}$ **的高绝对值**：这意味着数据集中存在大量的线索，这些线索可能被模型用来做出预测。换句话说，如果 $\mathcal{D}$ 的值很高，这通常指示数据集中存在明显的信息泄露问题，可能导致模型偏向于特定的线索，而不是学习到更深层次的、真实的推理模式。

- $\mathcal{D}$ **的低值**：较小的 $\mathcal{D}$ 值并不直接指示偏差较少，而是可能意味着训练和测试数据之间的信息 “泄露” 较少。这可以被解释为模型不过度依赖于数据集中的特定线索，从而可能展现出更加均衡和泛化的学习能力。

- $\mathcal{D}$ **值的正负**：如果 $\mathcal{D}$ 为正，这表明模型在预测时正在利用这些线索。它为研究者提供了一个有力的工具，用于判断模型是否在依赖数据集中的特定信息，而非进行真实的、基于内容的推理。

这种方法论的通用性使其适用于评估任何多项选择类型的数据集。通过应用这一度量标准，我们能够有效地识别和量化不同数据集中的信息泄露问题，从而为进一步的分析和改进提供了坚实的基础。

线索评估技术和度量标准

为了深入探究我们研究中的线索检测效果，我们采用了一种核心技术——“仅假设方法”，将其作为评估虚假线索存在的标准。这种方法独特之处在于，它仅考虑假设部分的信息，而不涉及问题的前提。通过这种方式，如果模型仅凭假设就能准确预测答案，这可能揭示了数据集中的偏见或泄露。

进一步地，为了简化评估过程并寻找类似于“仅假设方法”的有效度量，我们实施了四种更为简单的方法：平均值分类器（Ave）、最大值分类器（Max）、随机梯度下降分类器（SGDC）和逻辑回归（LR）。这些方法基于不同逻辑，旨在从不同角度评估和解释数据集中的线索。这些方法在 4.4中详细阐述

我们的方法与“仅假设方法”最显著的区别在于依赖的线索类型。我们的方法着重于可解释的、基于词汇的线索，相比之下，“仅假设方法”则涉及更复杂、不易解释的线索。这种差异不仅揭示了我们方法的独特性，也强调了我们在方法选择上的多样性。

我们的研究主要关注于**评估和验证我们提出的偏差检测方法**。这种方法的核心在于与传统的基于仅假设模型的表现进行对比分析。我们的目标是证明我们的方法在识别多项选择数据集中的虚假统计线索方面的有效性，进而彰显我们所做工作的贡献。

在这一实验框架下，我们采用了皮尔逊相关系数（Pearson Correlation Coefficient, PCC）来衡量我们的方法与已有的仅假设模型之间的相似度。这里，我们特别关注了 FastText 和 BERT 这两种模型。我们的分析覆盖了 12 个不同的数据集，并运用了八种线索分数度量方法和四种聚合算法。

分析结果，如表4-3所展示的，突出表明了在所有 12 个数据集中，CP 线索分数与逻辑回归模型的结合展示出与仅假设模型（例如 FastText 和 BERT）高度的相关性。与 FastText 和 BERT 模型相比，我们的方法在 PCC 得分上分别达到了 97.17 % 和 97.34 ‰。这些显著的结果使我们得出结论，CP 线索分数和逻辑回归模型的结合为评估所有数据集提供了一种强有力的方法。

基于这些发现，我们坚信基于 CP 的方法是一种强大的工具，能够识别数据集中的关键词特征。这是通过计算我们在4.4章节中描述的“线索度”来实现的。此外，CP 线索和逻辑回归模型的结合为确定多项选择数据集受信息泄露影响的程度提供了一个有效的度量标准，这对于我们的研究领域来说是一个重要的贡献。

此外，我们通过绘制偏差分数 $\mathcal{D}$，来可视化我们的发现，并对我们的 CP+LR 方法与两种仅假设模型（FastText 和 BERT）在 12 个数据集上的表现进行了比较，如图 4-3 所示。图中的线条的一致性显示了我们的方法

	Ave			Max			SGDC			LR		
	FT	BERT	P	FT	BERT	P	FT	BERT	P	FT	BERT	P
PMI	90.87	96.23	93.55	95.37	79.82	87.59	97.81	91.01	94.41	97.14	96.05	96.6
LMI	65.13	49.18	57.16	34.52	30.71	32.62	69.88	79.06	74.47	77.46	81.21	79.33
AD	84.62	72.49	78.56	90.75	73.02	81.89	93.73	76.24	84.98	97.56	86.91	92.24
WP	86.87	73.09	79.98	92.47	79.87	86.17	94.0	22.59	56.53	61.28	75.55	65.86
RD	96.59	93.82	95.21	98.23	91.04	94.63	94.30	93.98	94.14	94.21	95.59	94.90
Cos	94.84	82.94	88.89	92.73	75.40	84.07	98.08	87.86	92.97	87.38	78.44	82.91
Freq	68.00	50.02	59.01	34.45	30.67	32.56	64.08	67.11	65.60	74.58	88.64	81.61
CP	93.09	96.61	94.85	95.29	79.80	87.54	97.19	96.16	96.67	97.17	97.34	97.26

表 4-3　在 12 个数据集上，我们的方法的 $\mathcal{D}$ 与 FastText 和 BERT 的“仅假设”模型的 $\mathcal{D}$ 之间的皮尔逊分数。其中，P 表示 BERT 和 FastText（简称 FT）的平均皮尔逊相关系数。

与仅假设模型之间的强相关性。

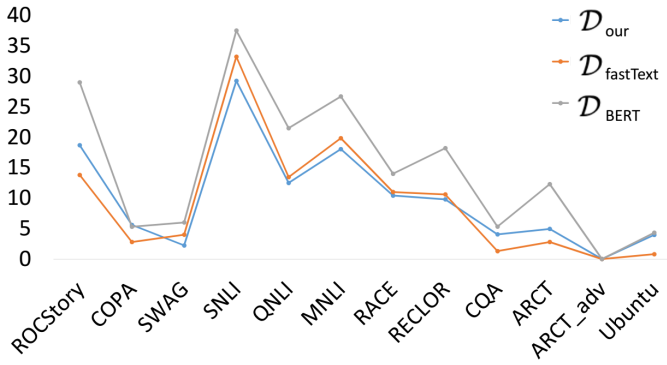


图 4-3　所有 12 个数据集上三种预测模型的偏差分数。

综上所述，我们的方法有效地识别并量化了数据集中的偏差，并且与仅假设模型的高度相关性进一步证明了我们方法的有效性和合理性。

数据集中偏差的揭示

为了更准确地识别那些存在问题的数据集，我们基于实验结果制定了一套详尽的标准。这套标准的核心在于，如果一个模型在任何给定的线索特征上的偏差 $\mathcal{D}$ 超过 10%，我们就认定该数据集存在显著问题。这一简洁的评判标准使我们能迅速辨认出那些含有严重统计线索问题的数据集。

在表4-4中，我们展示了四种简单分类模型在 12 个不同数据集上的最高准确率，以及与多数选择（Majority）的偏差。这一表格详尽记录了使用

词语线索特征时，在各数据集上的选择结果。

数据集	多数选择	Word Cues	
	(%)	Acc.(%)	$\mathcal{D}$(%)
ROC	50.0	68.68	18.68
COPA	50.0	55.60	5.60
SWAG	25.0	27.23	2.23
SNLI	33.33	62.57	29.24
QNLI	50.0	62.49	12.49
MNLI	33.33	51.37	18.04
RACE	25.0	35.42	10.42
RECLOR	25.0	34.80	9.80
CQA	20.0	23.42	3.42
ARCT	50.0	54.95	4.95
STS	50.0	50	0.0
Ubuntu	1.0	4.96	3.96

表 4-4　我们的四种简单分类模型在 12 个数据集上的最高准确率和与多数选择的偏差。

根据这个标准，我们识别出 ROCStories、SNLI、MNLI、QNLI、RACE 和 RECLOR 等数据集存在大量统计线索问题。如表所示，这些数据集中的 $\mathcal{D}$ 值显著高于 10% 的阈值，表明它们包含大量可以被模型利用的虚假统计线索。例如，在 ROCStories 数据集中，我们的方法实现的最高准确率比随机选择高出 20.92%，而在 SNLI 数据集中，这一比例甚至高达 33.59%。这进一步证明了这些数据集中存在的问题。

另外，我们也观察到，在未经过对抗性过滤且受到人为干预的数据集，例如 ARCT，存在更多的虚假统计线索。在对 ARCT 数据集进行的人工调整中，我们发现准确率有显著变化（从 54.95% 降至 50%），这表明人为干预对数据集质量的影响。

综上所述，在表格中我们报告了四种简单分类模型在 12 个数据集上的最高准确率以及与多数选择的偏差。我们的发现表明，偏差 $\mathcal{D}$ 是一个有效的工具，可以用来识别和评估数据集中存在的词语线索问题。

总的来说，我们对问题数据集的深入分析和标准化方法，能够帮助研究人员识别出含有大量统计线索问题的数据集。这一关键洞见将有助于发展更加健壮的模型，使其不那么依赖于表面的统计线索，而是能够更深入

地理解和处理复杂的数据。

2. **模型推理能力的评估**

为了探究模型是否学习了数据集中的线索特征，并因此做出有偏见的选择，本研究提出了一种方法来评估模型是否受到数据集中线索的影响。我们假设，如果一个模型确实受到这些线索的影响，那么它在简单和困难部分的表现将会有所不同。这种表现上的差异被称为**性能差距**，它反映了模型的健壮性。一个健壮的模型，理论上应当对虚假信息不敏感，并在简单和困难两部分的测试中表现出类似的性能。尤其是在困难部分的表现，是衡量模型能力的一个重要指标。

数据集	模型	原始测试	线索分类		
		(%)	简单	困难	**差距**
SNLI	BERT	90.48	94.99	83.02	11.97
	ESIM	87.44	93.27	77.77	15.50
	FastText	54.74	73.16	24.23	48.93
MNLI	BERT	85.10	90.60	79.30	11.30
	ESIM	76.82	85.80	67.34	18.46
	FastText	47.15	66.88	26.31	40.56
QNLI	BERT	88.89	90.92	85.54	5.37
	ESIM	72.17	78.55	61.66	16.89
	FastText	66.33	80.94	42.26	38.67
ROC	BERT	90.54	93.53	84.01	9.52
	ESIM	65.42	72.49	50.00	22.49
	FastText	62.91	71.16	44.90	26.26
RACE	BERT	90.54	93.53	84.01	9.52
	ESIM	65.42	72.49	50.00	22.49
	FastText	62.91	71.16	44.90	26.26
RECLOR	BERT	48.40	61.68	41.74	19.93
	ESIM	40.40	52.69	34.23	18.46
	FastText	31.60	43.11	25.83	17.29

表 4-5　模型在简单和困难测试数据集上的性能差距 (%)。

我们评估了 3 个典型的模型——BERT [35]、ESIM [22] 和 FastText [61]——在 6 个“差”数据集上的表现，分别代表不同的复杂性层级：

FastText: 这个模型将句子视为 n-gram 的集合，并试图独立预测每个答案的正确概率。在这个模型中，单词的表示是基于 300 维的 GloVe 嵌入。在多项选择的设置中，FastText 选择得分最高的答案作为预测结果。由于其简单的结构，FastText 在处理文本时可能更依赖于表面的词汇特征，这可能导致在复杂的推理任务中性能不足。

ESIM: 这个模型通过引入局部推理建模来增强其性能。它在两个片段局部对齐后，对前提和假设之间的推理关系进行建模。我们使用 GloVe 嵌入对自己的 ESIM 模型进行训练，并选择得分最高的答案。ESIM 通过在前提和假设之间建立更复杂的关系来处理任务，这使其在处理需要深层次理解的任务时表现更好。

BERT：BERT 模型基于 Transformer 架构，这是目前深度学习领域非常流行的一种结构。它的训练是在两个无监督的任务上进行的，这两个任务分别是：遮蔽语言模型（MLM）和下一句预测（NSP）。这些训练是在 BooksCorpus 和英文维基百科的文本上完成的。有多种预训练的 BERT 模型可用，它们在层数和参数数量上有所不同。我们选择的是一个基础版本，具有 12 层 Transformer 块、768 个隐藏单元和 12 个自注意力头，共计 110M 个参数。BERT 对前提和假设的连接（带有分隔符）进行了 3 个周期的微调，以预测基于前提和假设的关系。BERT 的这种架构使其能够在处理需要广泛上下文理解和复杂推理的任务时表现出色。

这些模型在不同数据集上的表现及其在简单和困难部分之间的差距如表 4-5 所示。简单和困难部分之间的差距通常在 10-50 % 之间。相比之下，人类在困难集上仍然能够保持良好的表现，这表明这些模型可能在某种程度上都受到了各自数据集中统计线索的影响。“组合” 意味着我们汇总了由词语线索做出的选择。对于 “组合” 而言，简单部分是其他两个线索特征简单部分的并集，而困难部分是两个其他困难部分的交集。我们发现 “组合” 通常会导致更大的差距。

在对比三个模型的性能差距时，我们发现 BERT、ESIM 和 FastText 的差距依次增大。例如，FastText 在 SNLI 数据集上的性能差距高达 48.93%，而 BERT 为 11.97%。这表明 BERT 可能比其他两个模型更具健壮性。然而，所有模型在六个数据集上的性能差距均超过了 10%，这暗示即便是 BERT 也受到了数据集中线索的影响。

值得注意的是，最高分数的模型并不总是最稳定的。例如，在 RACE 和 RECLOR 数据集上，尽管 BERT 的整体准确率更高，但它的性能差距比其他模型更大。ESIM 在 RACE 上的差距相对较小，但考虑到其在困难测试数据上的表现接近随机选择，我们不能简单地认为 ESIM 更优。此外，ESIM 在 RACE 和 RECLOR 数据集上表现出较大的收敛困难。

综合来看，我们提出了一种新的模型评估方法，即通过困难数据测试来考察模型的准确率和健壮性。通过性能差距分析，我们可以深入理解模型的局限性，并努力改进它们。通过减少性能差距，我们可以开发出即使在具有挑战性情境下也能保持高性能的更健壮模型。

4.6.3　微观结果分析

在节中，我们将重点讨论关于线索发现、模型探测和分析的成果以及有关 ChatGPT 模型评估的案例分析。这一整套框架已经在我们的在线演示中得到实现和展示。

1. 数据集中的线索发现

我们利用在 4.5节中描述的方法，对不同数据集中的线索进行了识别和分析。具体来说，我们首先运用本研究定义的各种特征，对每个数据集的训练和测试数据进行筛选。在表 4-6 的左半部分，我们展示了每个数据集中发现的前五个主要线索及其相应的线索得分。

以 STS 数据集为例，这是一个经过精心平衡的对抗性数据集，我们在其中只识别出一个线索——OVERLAP，且该线索的得分非常低。这一发现并不出乎意料，因为 OVERLAP 作为我们列出的特征中唯一涉及前提和假设标记的 “二阶” 特征，很可能已经避开了数据创建者创造的偏见。

在大多数情况下，我们发现的主要线索都是词汇特征。除了 OVERLAP 之外，我们还在列表中发现了如 NEGATION 和 TYPO 等特征。有趣的是，当我们将关注的线索扩展到前十个时，情感（SENTIMENT）和命名实体识别（NER）特征也显现出来。值得注意的是，我们还发现了一些以往其他研究报告中提到的、被认为有偏见的特征，比如 ARCT 中的 “not” 和 NEGATION，MNLI 和 SNLI 中的 “no”，以及 ROC 中的 “like”。尤其在 MNLI 数据集中，我们发现所有五个识别出的线索都与负面词汇相关，这

数据集	Top Cues	Cueness %	FT (Δ)	ES (Δ)	BT (Δ)	RB (Δ)
SNLI	“sleeping”	13.95	30.3	6.81	5.34	4.87
	“no”	13.33	18.09	3.32	2.05	2.6
	“because”	9.24	18.89	4.88	5.61	4.31
	“friend”	8.82	22.96	6.66	3.51	3.05
	“movie”	7.73	16.64	0.06	9.47	-0.19
QNLI	“dioxide”	4.52	9.78	-0.06	4.97	10.56
	“denver”	4.26	13.59	7.14	2.23	3.11
	“kilometre”	4.24	4.85	6.43	4.67	2.55
	“mile”	3.95	7.16	15.64	-1.65	-6.65
	“newcastle”	3.8	3.44	12.0	0.89	-1.23
MNLI	“never”	10.4	29.15	26.41	9.86	10.6
	“no”	8.98	19.49	20.17	1.2	3.32
	“nothing”	8.98	25.5	26.84	5.11	4.32
	“any”	6.79	20.4	19.39	7.76	3.74
	“anything”	5.73	18.43	15.74	3.31	1.14
ROC	“threw”	12.99	1.28	4.69	10.88	0.97
	“now”	8.68	-10.01	14.51	1.75	5.69
	“found”	8.16	-2.31	4.45	5.12	-3.13
	“won”	7.71	2.43	0.74	1.05	5.51
	“like”	7.3	4.77	10.06	8.81	1.67
COPA	“went”	3.61	-10.83	6.46	7.92	1.04
	“got”	2.74	5.45	-9.89	-12.52	-10.3
	“for”	2.14	10.11	-1.89	9.05	11.58
	“with”	1.38	-15.64	-6.98	3.3	13.82
	TYPO	0.84	-12.46	-2.33	3.8	-8.22
SWAG	“football”	7.38	6.13	8.55	1.2	1.55
	“anxious”	6.65	7.55	-4.67	-6.66	-1.67
	“concerned”	6.19	12.6	4.58	8.27	-5.66
	“skull”	5.73	-2.77	0.49	8.43	3.49
	“cop”	5.01	2.79	5.3	-0.92	-0.04
RACE	“above”	13.74	8.73	-8.43	-0.22	-1.92
	“b”	12.84	16.97	-4.8	3.52	-3.45
	“c”	11.83	15.69	-6.94	8.6	-7.6
	“probably”	6.77	9.91	-0.06	-3.8	2.86
	“may”	4.2	7.75	-3.45	-6.67	-1.8
RECLOR	“over”	2.07	1.76	-2.94	-1.35	-4.12
	“result”	1.97	-3.29	-2.69	-1.78	-3.7
	“explanation”	1.81	-6.33	-1.73	-2.76	-7.24
	“proportion”	1.68	-5.64	-4.69	2.37	-2.16
	“produce”	1.4	4.54	-2.98	-14.36	-3.7
ARCT	“not”	3.74	-2.54	7.45	-0.97	-11.96
	NEGATION	2.85	3.49	10.04	6.28	-8.23
	“n’t”	2.52	10.3	5.89	9.49	4.84
	“always”	2.25	-4.66	38.21	-4.35	-8.26
	“doe”	2.06	-0.73	-3.69	-1.15	-7.22
STS	OVERLAP	1.96e-10	1.65	-0.25	2.73	0.57
$\Sigma(\lvert . \rvert)$ (Model weakness)			469.8	361.4	227.7	216.2

表 4-6　数据集、每个数据集的前 5 个线索和每个模型在这些线索上的准确率偏差 Δ。

表明该数据集可能包含了显著的人为痕迹，从而可能导致模型的脆弱性。

此外，我们还注意到某些词汇线索揭示了问题中的特定句法、语义或情感模式。例如，SNLI 数据集中的 “because” 表明了因果关系的存在；ROC 中的 “like” 通常与积极情感相关；RACE 中的 “probably” 和 “may” 表达了不确定性等。这些发现可作为修订数据集时的重要线索。

2. **模型偏见的探测和分析**

为了深入探索模型是否受到数据集中特定线索或特征的影响，我们在原始训练集上对四种模型——FastText、ESIM、BERT 和 RoBERTa——进行了训练。与之前的宏观测试相比，我们新增了 RoBERTa 模型的研究。作为 BERT 的改进版本，RoBERTa 通过在更大的数据集上训练、增加批量大小，并去除了一些原始 BERT 的目标，例如下一句预测（NSP），从而显著提高了性能。这些改进使得 RoBERTa 在多种自然语言处理任务上表现更加出色。我们选用的是 RoBERTa 的基础版（base version）进行测试。接下来，我们通过准确性和分布测试来评估这些模型。

准确性测试

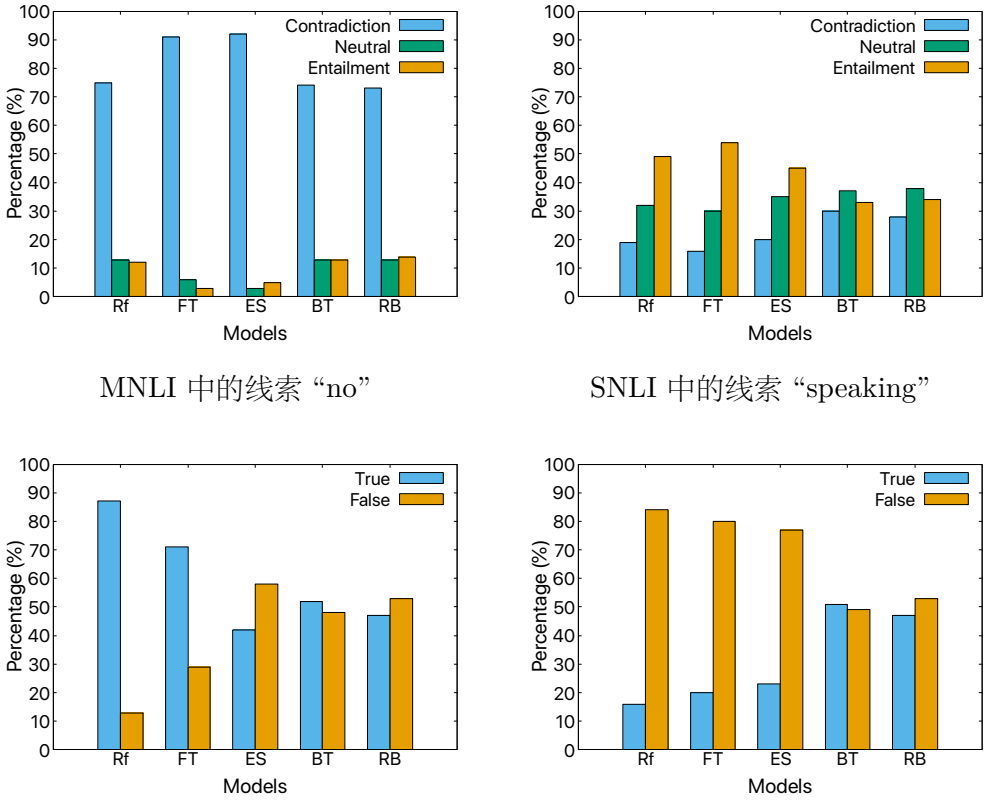
我们在表 4-6 中展示了测试结果。如4.5.2所述，Δ 值的正负号表示模型在含有和不含特定特征的数据子集之间性能变化的方向。Δ 的绝对值大小则反映了模型性能受这些特征影响的程度。较大的 Δ 值表明模型对特定特征有更强的依赖或敏感性，而较小的值则意味着模型更加健壮。

在表 4-6 的底部，我们观察到，跨越所有十个数据集，Δ 绝对值的总和遵循以下顺序：RoBERTA < BERT < ESIM < FastText。这与我们之前的假设和社区对这些模型的普遍看法一致。然而，对单个数据集和特征的细致检查揭示了更复杂的情况。例如，FastText 倾向于捕捉单词级的线索，而不是深层次的语义线索，而像 BERT 和 RoBERTa 这样的更复杂模型则对结构性特征（如 NEGATION 和 SENTIMENT，实际上是词汇类别）更敏感。这种差异可以通过 FastText 的设计理念来解释，它更专注于单词层面，而不是句法或语义结构。

有趣的是，FastText 对 TYPO 特征显示出了强烈的负相关性。我们推测，这可能是因为 FastText 使用了更规范的词汇进行训练，因此对于文本

中的拼写错误表现出了较低的容忍度。

分布测试



MNLI 中的线索 “no”　　SNLI 中的线索 “speaking”

ARCT 中的线索 “above”　　ROC 中的线索 “threw”

图 4-4　四个不同模型的分布测试示例。

在图 4-4 中，我们展示了三个引人关注的发现。这些柱状图代表了四个模型基于每个预测标签的分布百分比。R_f 表示具有特定特征的训练数据分布。我们注意到，在 MNLI 数据集中的 “no” 线索上，所有模型在表 4-6 中均获得了正 Δ 值，尤其是 FastText。与准确性测试一致，我们发现在有 “no” 线索的情况下，FastText 和 ESIM 的预测标签分布偏差在图 4-4 中得到了加强。它们更倾向于预测 “矛盾” 类别，甚至超过了训练数据中的基线。与此相比，BERT 和 RoBERTa 的预测分布更加适度地遵循了训练数据。

尽管 “no” 线索在影响模型方面表现出一定的有效性，但 “above” 这一

线索却并不那么成功。图 4-4 显示，在 ARCT 数据集中，ESIM 的预测结果分布与训练数据完全相反，这解释了表 4-6 中的 Δ 值为-8.43，并暗示即使数据中存在线索，模型也可能不会利用它。同样，BERT 和 RoBERTa 在 “speaking” 线索上的低 Δ 值也解释了它们在表 4-6 中的表现。

特别地，“threw” 线索为 BERT 提供了一个例外情况，因为在分布测试中的结果与准确性测试不一致：尽管 BERT 在准确性偏差上表现很高，但其预测分布却相对均匀。我们并没有遇到很多这样的矛盾情况。然而，当这种情况确实发生时，如在这个例子中，我们认为 BERT 可能没有真正利用 “threw” 这一线索。

通过上述分析，我们不仅揭示了数据集中的线索和特征，还展示了这些线索对不同模型性能的影响。这些发现为深入理解模型行为提供了宝贵的洞见，并为未来的数据集设计和模型开发提供了重要的指导.

3. 案例研究

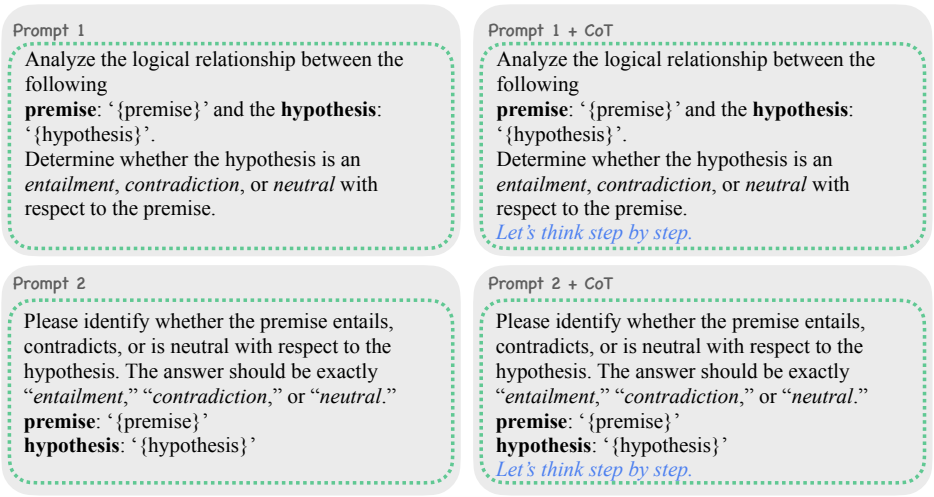


图 4-5　提示语（prompt）。

最近，OpenAI 发布的大型语言模型（LLM）ChatGPT 在自然语言处理（NLP）领域引起了极大关注。ChatGPT，属于 GPT-x 系列模型[9]之一，采用了类似 InstructGPT[99] 的训练方法，即通过强化学习从人类反馈中学习（RLHF）。

[9]目前 x 的值为 3.5 或 4，在我们的实验中，使用的是基于 GPT-3.5 的版本）

在本节中，我们关注的是 ChatGPT 在零样本情况下是否受到偏见特征的影响。具体来说，我们选择了一个聚焦于 MNLI 数据集中 “no” 一词的案例研究。我们的目标是评估不同提示语的有效性，并基于这一单一偏见特征选择最佳提示语，以减轻偏见。

数据集

我们从 MNLI 数据集中挑选了测试实例，专注于 “no” 一词对 ChatGPT 性能的影响。原始测试集包括 3240 个矛盾实例（Contradiction）、3463 个蕴涵实例（Entailment）和 3129 个中立实例（Neutral）。含有 “no” 的实例分布为：矛盾 229 个，蕴涵 38 个，中立 46 个。

为了测试准确性，我们选取了所有含有 “no” 的 313 个实例，并从未包含 “no” 的测试集中随机挑选了相同数量的实例，以确保评估的平衡性。

在分布测试中，我们从每种分类标签中各选取了 38 个含 “no” 的实例，共计 114 个。

提示语

如图4-5所示，我们设计了四种提示语。第一种由 ChatGPT 自行生成，我们询问它关于 MNLI 任务的最佳提示语是什么，并得到了 Prompt 1。第二种提示语受到先前工作 [109] 的启发。第三和第四种提示语在前两种提示语的基础上加入了 “Let’s think step by step”[65] 的元素，采用了 “思维链”（Chain of Thought, CoT）的方法，这种修改方式在提高 InstructGPT 在推理任务上的性能方面已被证明有效，如 Ouyang 等人在 2022 年的研究 [99] 所示。

ICQ 结果

提示语	准确率（含 “no”）	准确率（不含 “no”）	ΔAcc
P1	74.34	77.32	-2.98
P2	75.42	74.18	1.24
P1 + CoT	78.35	77.28	1.07
P2 + CoT	76.67	76.40	0.27

表 4-7　准确性测试结果（%）。P1= **提示语 1**，P2= **提示语 2**。

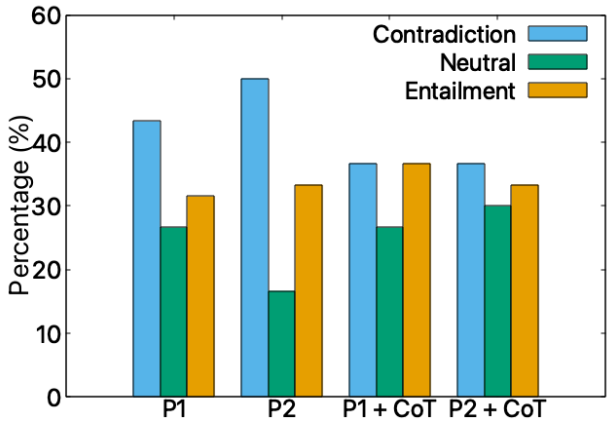


图 4-6　分布测试结果。

我们使用不同的提示语对含有和不含有 “no” 的实例评估了模型的准确率。结果显示在表 4-7 中。P1 展示了负 Δ 准确率，表明在存在 “no” 时难以泛化。P2 显示了正 Δ 准确率，表明更好的泛化能力。将 “CoT” 添加到两个提示语中都减少了偏见风险。P1 + CoT 在准确率（含 “no”）方面显示了最显著的提升，但 P2 + CoT 有最小的绝对 Δ 准确率，表明对 “no” 的敏感性最低，是测试提示语中偏见风险最低的。

此外，我们分析了模型在不同提示语的压力测试集上的预测分布，如图 4-6 ，该测试集包含了具有平衡标签分布的 “no” 一词。分布测试结果揭示了 P1 和 P2 预测分布的不平衡，P1 和 P2 倾向于预测矛盾。添加 “CoT” 减轻了这些不平衡，导致更平衡的分布。P2 + CoT 在标签中呈现了最平衡的分布，支持其最低的偏见风险。

总之，我们的案例研究，特别是当关注 “no” 特征时，表明零样本 ChatGPT 可以受到偏见特征的影响。提示语的选择显著影响其性能。在评估 “no” 特征时，P2 + CoT 在两项测试中显示了最低的偏见风险。“CoT” 策略在减少这一特定特征的偏见风险方面似乎是有效的。然而，重要的是要注意，我们的发现，特别是关于 “no” 特征，表明 ChatGPT 推荐的自我提示语（P1）可能并不总是最优的。这强调了人类干预和持续探索的重要性，以优化性能并最小化偏见风险。未来的研究和结论将受益于更细致和针对特定特征的分析。

4.7　本章小结

在本章节的小结中，我们深入剖析了常识性推理模型鲁棒性不足的核心问题，并探讨了这一现象背后的深层原因。我们着重分析了模型在学习过程中缺乏人类的稳固推理能力，探求了模型究竟学习到了什么。为了深入理解这一问题，我们引入了两种创新的测试框架，目的是对相关问题进行更精细的分析与理解。

鉴于模型性能与数据密切相关，我们从宏观和微观两个角度出发，对数据中的偏见线索及模型的偏见进行了全面的评估和分析。在宏观层面，我们开发的测试框架通过对不同数据集上基于统计特征的简单分类模型的最高准确率进行评估，以及通过与多数选择（Majority）的偏差值来评估数据集中存在的偏差特征程度。我们将测试数据划分为简单（easy）和困难（hard）两类，以量化评估模型在识别和处理虚假特征方面的能力。这一方法揭示了模型对某些统计规律的过度依赖，有助于我们深入理解模型泛化能力的限制。

在微观层面，我们设计了 ICQ（“I-see-cue”）框架。我们首先通过结合特征的分布不平衡性和训练集与测试集分布的相似性来定义 “cueness score”，从而发现可能影响模型的关键特征，并评估数据集中存在的问题。ICQ 通过多维特征划分和细致的性能分析，探究模型在不同特征上的准确性和分布表现。此外，我们还开发了一种直观的可视化工具，以更有效地识别和理解模型性能差异的根源。

值得一提的是，在微观分析中，我们还对 ChatGPT 进行了案例分析，证明了 ICQ 框架的广泛适用性，能够为多种模型提供深入的分析。

总体而言，本研究的主要贡献在于这两种高度创新的分析框架，它们为深入理解和提升常识性推理模型的鲁棒性开辟了新途径。通过结合宏观与微观方法，我们能够全面评估模型在处理复杂数据时的行为模式，特别是在识别潜在虚假特征方面的能力。这些成果不仅揭示了模型的潜在薄弱环节，也为未来设计更健壮、更可解释的人工智能系统奠定了一定的基础。

第五章　提升推理模型鲁棒性的数据增强策略

在人工智能领域，常识性推理作为一项核心任务，其挑战在于赋予机器与人类相似的理解和决策能力。尽管神经网络模型如 BERT[35] 和 RoBERTa[78] 在 ROC[92]、COPA[120]、ARCT[46] 和 RECLOR[157] 等任务上取得了显著成就，它们在处理对抗性数据或不同于训练环境的场景时仍显脆弱。这种局限性通常源于模型对数据集中特定模式的过度依赖，而非深入理解问题本质 [95, 82, 126, 97]。在上一章对模型鲁棒性差的原因的解释中我们也验证了的确受到偏差数据的影响。

针对模型鲁棒性不足的问题，本章提出了一种数据增强方法，旨在提升常识性推理模型的鲁棒性。这种方法通过创新的数据增强手段减少数据的简单统计偏差，从而增强模型的鲁棒性。

本研究的灵感来源于生物学中的“交叉”和“变异”过程。这一策略模仿生物遗传中的染色体交叉和基因变异，创造问题实例之间的新组合和变化。“交叉”操作涉及交换两个不同问题的选项，而“变异”操作涉及对问题的某些部分进行微妙修改。这些操作产生新的、具有挑战性的训练实例，促使模型超越表面的统计规律，深入学习问题的内在逻辑和结构。

这些策略被应用于领先的神经网络模型，并在多个基准数据集上进行测试。实验结果显示，这些策略显著提高了模型在多样化和对抗性数据环境中的鲁棒性，同时在标准测试集上保持或提升了性能。

5.1　概述

选择题（MCQs）是一种常用的格式，用于评估自然语言理解（NLU）能力，涵盖了因果推理 [120]、故事结尾预测 [92, 53]、论证理解 [46] 以及阅读理解 [157] 等任务。这些任务通常由一个情境描述（前提）和几个选择项组成。例如，COPA 数据集 [120] 通过 MCQs 来测试常识性因果推理 [80]，一个典型的例子如下：

Example 2. COPA 选择题示例：

Premise: The man hurt his back.

Choice 1: He stayed in bed for several days. ✓

Choice 2: He went to see a psychiatrist. ✗

近年来的研究开始探究高级神经模型在 NLU 推理问题上的强大表现。特别是，研究发现许多模型可能并不是通过真正理解上下文与选项之间的逻辑和语义联系来取得成功，而是通过利用训练和测试数据中的偏见或统计特征。这一观点得到了“仅选项测试”（也称为“仅假设测试”）的支持 [131, 69]。在这种测试中，模型如 BERT 在没有前提的情况下也能正确回答问题。

我们将这种现象称为自然语言推理中的“短路”。虽然“仅假设测试”为短路行为提供了一些证据，但我们认为它具有一定的局限性。即使模型在没有前提的情况下能正确回答问题，也不一定意味着它在有前提时不会考虑前提。

为了更全面地测试这种“短路”现象，我们尝试了一种新方法：绘制模型在处理完整问题时，最后编码层中单词间的注意力图。如图5-1所示，我们使用了来自 COPA（见例子 2）的示例。注意力图清楚地展示了，在处理完整问题时，模型对于第一个选项与前提之间的联系几乎没有关注，而当仅处理选项，无上下文时，选项内单词之间的注意力保持不变。

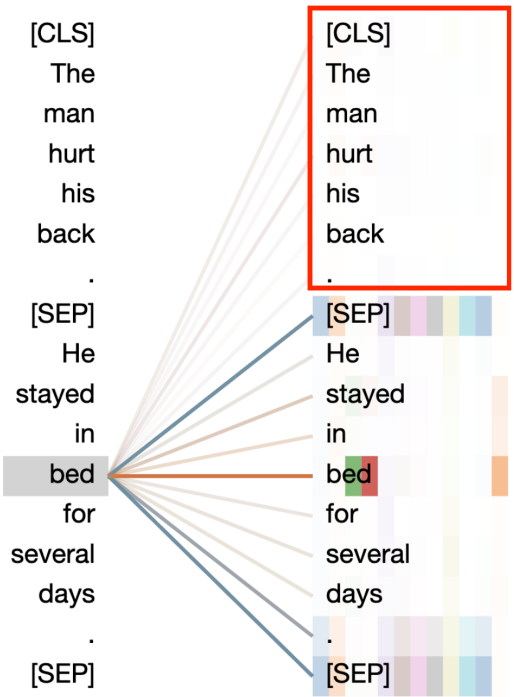


图 5-1　展示 BERT[35] 在 COPA 问题上短路的注意力图。

尽管使用注意力图来手动检查模型的短路行为可以提供洞察，但这一过程既繁琐又成本高昂。因此，我们开发了一种自动化的白盒测试算法，使用阈值来模拟人类的视觉过程。但这种方法也有局限性：它需要访问模型的内部代码，并且只适用于基于注意力的模型。

为了解决这些挑战，我们引入了一种新操作，称为“交叉”，用于 MCQ 问题实例。这种操作通过交换两个 MCQs 的选项来模拟生物繁殖中的染色体交叉过程。这对经常利用短路行为的模型提出了独特挑战，通过构建代理测试案例可以在真实任务中检测这种行为。我们在三个最新的、强大的 NLU 推理模型上进行了这种交叉测试，发现这些模型在实例级别的压力测试中表现出明显的准确率下降，显示出短路行为。

在确认了短路行为的存在后，我们的目标是提高模型的鲁棒性。我们考虑了使用挑战模型难度的压力测试生成更多训练实例的方法。但由于许多压力测试对选项构建有限制，这限制了它们作为通用数据增强方法的有效性。然而，交叉操作及其对应的“突变”操作提供了一个可行的解决方案。这些操作不仅可以用于检测短路行为，还可以作为有效的数据增强技术，减少短路行为的发生，从而提高 NLU 模型的整体鲁棒性。为此，我们应用交叉、突变和反向翻译 [153] 来增强 BERT、XLNet [156] 和 RoBERTa [78] 在 ROC [92]、COPA、ARCT [46] 和 RECLOR [157] 上的表现。我们的实验显示，我们的方法让模型在压力测试上的准确性最多提高了多达 24%，在原始测试数据上最多也提高了 10%。

本研究主要贡献如下：

1. 我们提出了两种检测短路行为的方法：一种基于注意力权重阈值的白盒方法和一种受分子生物学启发的黑盒“交叉”测试。

2. 我们通过实验证实了三个强大、微调过的 NLU 推理模型中存在短路行为。

3. 我们建议使用交叉和突变操作来增强训练数据，鼓励模型考虑问题的上下文。我们的实验确认了这种方法的有效性，不仅在压力测试上，而且在原始测试数据上也显示出模型鲁棒性的显著提升。

5.2　相关工作

我们的研究涉及四个个主要领域：虚假特征的影响、数据增强的有效性，数据集过滤的挑战和好处以及模型分析。

虚假特征的影响

近期研究揭示了虚假特征在自然语言处理模型性能中的显著影响。具体来说，这些模型倾向于依赖表面模式，例如词汇化特征（如 n-gram 及其交叉组合）和非词汇化特征（比如词汇重叠、句子长度和 BLEU 分数）[131, 136, 158, 11, 95, 60]。显著的一点是，某些 NLP 模型能够在多项选择问答任务中取得高准确率，即使它们没有考虑上下文信息，这种现象通过“仅假设测试”暴露了模型对假设中细微但重要语义扰动的不敏感性 [123, 55]。这些发现促使我们直接对模型进行诊断，以减少仅依赖假设中的短路推理方式。

数据增强的应用和挑战

数据增强技术已被广泛应用于增强视觉和语言任务模型的鲁棒性 [105, 4]。然而，研究指出这种方法可能导致模型在增强数据集上过度拟合，从而限制了对新场景的泛化能力 [57, 118, 54, 76, 82]。我们采用一种新策略，通过生成包含多样化“噪声”示例的增强数据，以防止模型过分依赖特定的虚假线索。此外，我们还探索了一系列不依赖特定特征的增强方法，以降低模型依赖于选项短路推理的倾向，这在以往研究中较少涉及 [153, 129, 26, 27, 159, 38]。

数据集过滤的挑战与创新

在提升数据集质量的过程中，尝试通过移除某些人为特征来实现，但这种方法可能过分依赖于模型本身的性能，从而影响其长期的稳定性和可靠性 [155, 69]。因此，我们的研究中着重于开发一种创新策略，而非仅依赖于过滤现有数据集，以此提升模型在处理复杂、真实世界数据时的鲁棒性和准确性。

模型分析

自从大型预训练语言模型的出现，许多研究已经专注于分析这些模型的内部机制，包括语言属性如何被编码在上下文化表示和注意力头中 [41, 25, 75, 142]。与此相比，我们的研究更加关注于模型的高级推理能力。我们通过设计挑战性数据集和执行所谓的短路测试（一种压力测试），补充传统的挑战或基准数据集，从而测试模型是否真正具备推理能力，尤其是在短路行为方面 [5]。

5.3　短路问题的代理测试

在这一节中，我们首先介绍了用于检测模型中短路现象的方法，然后在下一节中我们对这些方法进行了一些修改，以便创建训练数据来解决短路问题并提高模型的鲁棒性。

由于目前没有现成的方法可以确切地证明模型在处理问题时是否发生了“短路”，我们设计了两种作为短路代理测试的方法。这些方法能揭示模型是否倾向于采用短路的方式解决问题，尽管它们不能直接证明短路本身，其作用类似于在天文学中对暗物质的探测。一种方法是在白盒设置下检查模型的注意力图（Attention Weights, AW），另一种方法则是在黑盒设置下，通过对正确选项应用不同操作来生成新的测试案例。

5.3.1　白盒测试

我们可以通过可视化基于注意力的模型的注意力图来直观地检测模型是否采用了短路。考虑一个经过良好训练的模型和一个以 [CLS] 前提 [SEP] 选项 [SEP] 格式正确回答的多项选择题（MCQ），其中 [CLS] 和 [SEP] 是模型使用的分隔符。在这种设置下，选项代表正确的选择。我们首先对输入进行分词处理，将这些令牌序列输入到模型中，并从模型的最后一个编码器层提取所有注意力头的注意力图。

我们使用现成的工具 [145] 将注意力图可视化成用户友好的形式，如图 5-1 所示。然后，我们请人类注释者判断正确选项到前提之间是否存在强的注意力连接。如果超过一半的注释者认为存在强连接，则判断该 MCQ 没有采用短路解决。

虽然手动注释方法准确，但它成本过高，难以应用于大规模的测试。为了克服这个问题，我们提出了一种基于规则的程序，自动检测模型在 MCQ 上的短路行为。具体来说，我们通过对所有注意力头进行最大池化操作，将注意力图聚合成一个单一的图表。然后，我们检查选项中的每个令牌与前提中的每个令牌之间是否存在至少一个高于阈值 t_1 的注意力分数，或至少两个高于阈值 t_2 的分数。特殊令牌如逗号和句号被排除在外。如果这两个条件都不满足，我们认为模型在这个 MCQ 上没有发生短路。实践中，阈值 t_1 和 t_2 需要被调整以最大限度地模拟人类的判断。相关的伪代码如下所示：

Algorithm 3 注意力权重阈值

Input: premise P, correct choice C, model M, threshold t_1 and t_2.
Output: binary 0/1 label L.
initialize counters c_1 and c_2 to 0.
tokenize the formatted input as sequence of tokens S.
feed S into M and extract the last layer's attention maps $Attn_{all}$.
aggregate $Attn_{all}$ into $Attn_{max}$ by max-pooling over all attention heads.
for w_1 in C do
    for w_2 in P do
        if $Attn_{max}(w_1, w_2) > t_1$ then c_1 += 1
        end if
        if $Attn_{max}(w_1, w_2) > t_2$ then c_2 += 1
        end if
    end for
end for
output 1 if $c_1 > 0$ or $c_2 \geq 2$ and 0 otherwise.

这段伪代码是为了自动检测模型在处理多项选择题（MCQ）时是否采用了短路行为设计的。短路行为指的是模型在未充分理解问题内容的情况下快速做出判断。该程序旨在减少人工注释工作的高成本和劳动强度，实现自动化检测。下面是对伪代码的详细解释：

1. **初始化计数器**：设置两个计数器 c_1 和 c_2，用于后续记录满足特定条件的注意力连接数。

2. **分词处理**：将 MCQ 的输入（前提和选项）进行分词处理，转换成模型可以理解的令牌序列 S。

3. **提取注意力图**：将令牌序列 S 输入到模型 M 中，并提取模型最后一层的注意力图 $Attn_{all}$。这个注意力图反映了模型在处理输入时各个令牌之间的关注程度。

4. **注意力图聚合**：通过最大池化方法将所有注意力头的注意力图聚合成一个综合的图表 $Attn_{max}$。最大池化是一种常用的降维方法，旨在保留最重要的特征。

5. **检测注意力连接**：遍历选项中的每个令牌 w_1 和前提中的每个令牌 w_2，检查它们之间的注意力分数是否超过设定的阈值 t_1 或 t_2。如果超过 t_1，则 c_1 加 1；如果超过 t_2，则 c_2 加 1。这一步骤用于评估模型是否在关键词之间建立了足够的关注度。

6. **输出结果**：根据计数器的值来判断模型是否在该 MCQ 上发生了短路。如果 c_1 大于 0 或者 c_2 大于等于 2，则认为没有发生短路，输出 1；否则输出 0，表明模型可能在这个问题上采用了短路策略。

通过这种方法，可以自动化地评估模型是否在理解问题的深度上存在缺陷，从而提高对模型行为分析的效率。

5.3.2 黑盒测试

在多项选择题（MCQ）模型的评估中，基于注意力的测试方法虽然能够检测模型编码器内部的短路现象，但对于评估整个端到端 MCQ 模型的短路行为却显得不足。这主要是因为在基于注意力的预训练语言模型之上，MCQ 模型通常还会包含额外的线性层，这些层在处理和形成最终决策中起着关键作用。仅仅通过分析模型的注意力分配情况，我们无法完全把握模型是如何综合这些层的信息来做出最终选择的。

为了全面评估 MCQ 模型是否存在短路行为，我们提出了一种自动的、端到端的、与模型结构无关的黑盒测试方法。在这种测试中，我们首先观察模型在原始 MCQ 上的表现。如果模型能正确回答，我们随后通过对正确选项实施某种“操作”来轻微修改问题，进而生成一个新的错误选项。这个新生成的问题必须与原始问题共享相同的正确选项。通过分析模型对修

改后问题的响应，我们能推断出模型是否在原始问题上实际理解了问题内容，还是仅仅依赖于短路策略。

在我们的研究中，我们探讨了表 5-1 中列出的多种操作。这些操作中的一些已在先前的研究 [62, 1] 中提出，而其他一些则是我们首次提出的（用 * 标记）。每个操作都有具体的描述和示例，这些示例展示了如何从原始问题的选项中构建假选择。这些操作的目的是保持（p）或改变（c）选项的真实性。基于软件工程中边界测试的理念，我们将这些操作分为三个类别，

Oper.	Description and Example
Neg+	Add negation (c) They called the police to come to my house. ✓ They didn't called the police to come to my house. ✗
Neg-	Remove negation (c) Ben never starts working out. ✓ Ben starts working out. ✗
NER	Randomly replace person names (c) A big wave knocked Mary down . ✓ A big wave knocked Kia down . ✗
PR*	Switch pronoun by gender or quantity (c) She had a great time .✓ He had a great time . ✗
PI*	Instantiate pronoun by randome person (c) They gave Tom a new latte with less ice . ✓ Nathanael gave Tom a new latte with less ice . ✗
Adv	Add adverbs for emphasis (c) The ocean was a calm as a bathtub .✗ In fact the ocean was a calm as a bathtub .✗
CO*	Crossover: Swap the true choices between two questions (p) Josh got sick . ✓ She had a great time .✗
Syn	Replace adj/adv with synonym (p) Dawn felt happy about getting away with it . ✗ Dawn felt glad about getting away with it . ✗
MT*	Mutate: Swap two consecutive words (c) Deb said yes to Tim 's marriage proposal. ✗ Deb said yes Tim to 's marriage proposal .✗
Voice	Swap subject and object (c) Kara asked the neighbors not to litter in their yard . ✓ the neighbors asked Kara not to litter in their yard . ✗

表 5-1　用于代理测试的操作。每个操作的第一行描述了操作本身，接下来的两行则提供了如何从原始问题的选项中构建假选择的示例。操作可能会维持（p）（✓→ ✓, ✗→ ✗）或改变（c）选择的真实性（✓→ ✗）。

这取决于构建的新选项的特性：

1. 第一类包括语法和语义都正确，且新选项与原选项相似的情况；

2. 第二类包括语法和语义都正确，但新选项与原选项明显不同的情况；

3. 第三类则涉及语法或语义错误的选项。

由于模型可能仅因为第三类中选项的错误而做出正确选择，而非基于对前提的深入理解，因此这一类不适合用于测试短路现象。

我们特别关注的操作包括第一类中的否定 [62]、NER [62] 和代词扰动，以及第二类中的副词 [1]、交叉和同义词 [62, 1] 操作。这些操作被设计用来探索模型在不同方面的理解能力，包括语义、语法结构和上下文逻辑。

尽管大多数操作都是直观且易于理解的，但交叉操作独特且值得特别关注。这一操作的灵感来自分子生物学，在数据集中，我们随机抽取一个模型正确回答的 MCQ，并用其真实选项替换原始问题中的错误选项。在代理问题中，新的选项仍然是错误的。这个操作可以通过图 5-2 中的直观展示来解释。

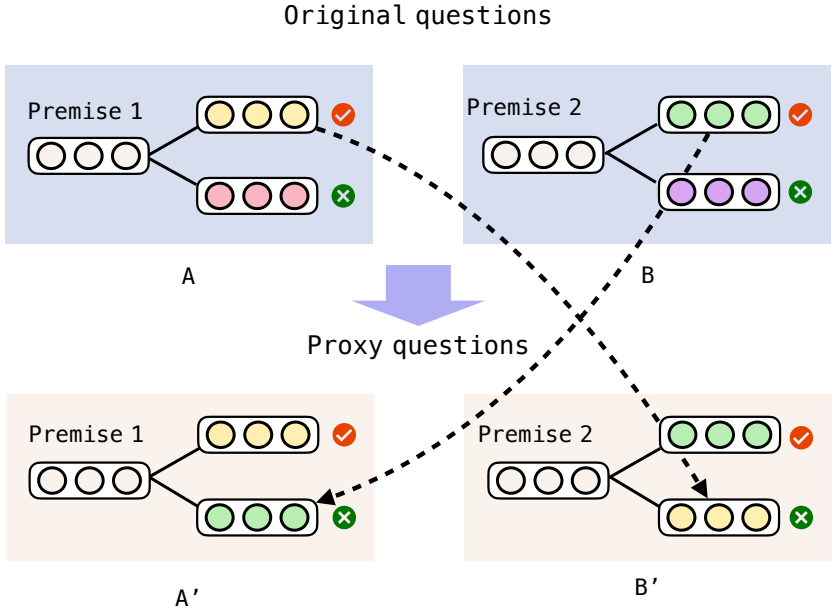


图 5-2　交叉操作示例：用两个问题的正确选项替换这些问题的错误选项，以此创建两个新的代理问题。

与第一和第二类中的其他操作相比，交叉操作生成的代理问题与原始问题最为不同，但从人类角度来看更易理解，因为这两个选项可能完全无关。如果模型处理不当，可能会更明显地暴露出短路行为。因此，交叉操作可能是一种更有效的短路测试方法。

交叉操作的另一个优点是，它允许我们以低成本为每个原始问题生成多个假选择，从而更全面地对每个原始问题进行测试。相比之下，大多数其他操作无法为原始选择生成足够数量的不同变体。

总的来说，所提出的黑盒选项操作提供了一种更具普遍性且与模型无关的方法，用于检测 MCQ 模型中的短路现象。通过应用各种操作来创建代理问题，我们可以更准确地评估模型的性能和鲁棒性，为未来开发更好、更可靠的模型做出贡献。

5.4　数据增强策略

在进行多项选择题（MCQ）模型的代理测试时，如果发现模型倾向于简化处理（即“短路”现象），那么它在面对非训练集中的、更复杂的数据（即域外数据）时，性能很可能会有所下降。针对这一问题，一个有效的解决方法是通过数据增强来扩展训练数据集，从而促进模型在处理问题时更加关注于前提和选项之间的深层次联系。尽管用于构建代理测试的操作也可应用于数据增强，但并非所有操作都具备生成大量有效训练数据的能力。

在众多数据增强技术中，特别值得关注的是交叉和变异操作。这些操作可以被有效地集成到训练数据中，以此提升模型在各类数据上的适应性和稳定性。

5.4.1　交叉操作在数据增强中的应用

作为数据增强的一种有效手段，交叉操作之所以出色，是因为它涉及将两个不同问题中的正确答案相互交换。如果模型依赖于浅层的识别模式（即短路），这些正确答案可能会带有误导性的信号。通过在训练数据中融入交叉操作，模型被迫必须依赖于前提的内容来判断哪个选项是更为合理的。这种策略有效地迫使模型从单纯的关键词匹配，转向对问题内容和上下文的深入理解。

5.4.2　变异操作在数据增强中的应用

变异操作主要有两种形式：一种是仅在正确答案中交换词汇；另一种是同时在正确和错误答案中交换词汇。与交叉操作相比，变异操作可能更

加有效地提升模型的鲁棒性。这是因为变异操作不仅迫使模型依赖于前提内容来区分两个高度相似的选项，而且还让模型对选项中的微小词序变化保持敏感。此外，这种操作还有助于加强模型对于语法和句式结构的理解。

5.4.3 代理测试与数据增强的区别

在使用交叉和变异操作时，理解它们在代理测试和数据增强中的不同用途至关重要。在代理测试的场景中，这些操作被用于修改测试数据集，目的是评估模型是否能在复杂或欺骗性的场景中保持准确性。相反，当这些操作被用于数据增强时，它们被应用于训练数据，目的是增强模型的整体适应性和泛化能力。

综上所述，通过采用交叉和变异操作进行的数据增强不仅能够提升模型对前提和选项间关系的敏感性，还能够加强模型对于细微词序变化的识别能力，从而在实际应用中实现更为精准和可靠的表现。

5.5 实验

数据集	前提	选项	训练集	测试集
ROC	Sarah was home alone. She wanted to stay busy. She turned on the TV. She found a reality show to watch.	Sarah then happily watched the show. ✓ Sarah could not find anything to watch. ✗	1871	1871
ARCT	Reason: Milk isn't a gateway drug even though most people drink it as children. Claim: Marijuana is not a gateway drug.	Warrant 1: Milk is similar to marijuana. ✓ Warrant 2: Milk is not marijuana.✗	1210	444
RECLOR	Context: In a business...to financial prosperity. Question: The reasoning in the argument is flawed because the argument	A: ignores the fact that in... the family's prosperity.✓ B: presumes, without... the family's prosperity.✗ C: ignores the fact... even if they pay high wages.✗ D: presumes, without providing...can succeed.✗	4638	500

表 5-2　数据集示例。

首先，我们将展示实验设置。其次，我们比较了几种测试操作符，用于

发现短路问题。最后，我们评估了不同增强方法的模型在抵御短路方面的鲁棒性。

5.5.1 实验设置

本节将详细阐述我们的实验设置，涵盖所使用的数据集、模型细节，以及我们为发现短路问题设计的测试操作符。

数据集

本研究使用了以下四个数据集，涉及不同的 NLP 任务，以评估模型在多样化场景下的表现，具体示例在5-2中展示：

ROC 是一个故事结尾预测数据集，此数据集要求模型从两个备选故事结尾中选择一个与前四句话的故事前提相符合的。每个案例包含一个短故事和两种可能的结尾，挑战在于理解故事情节并准确预测其合理结尾。

COPA 是因果推理数据集，其示例之前在 5.1节中展示过。给定一个前提情境，COPA 要求选择更合理、因果相关的选项。训练数据和测试数据各有 500 个实例。

ARCT 是论证理解数据集。ARCT 包含一系列论证问题，要求连接原因和主张。每个问题提供一组备选论据，模型需要选择最佳的论证选项。

RECLOR 是逻辑推理阅读理解数据集。RECLOR 要求模型根据给定的文本段落进行逻辑推理，以回答相关问题。这些问题设计来评估模型在理解逻辑结构和推理能力方面的性能。

模型

我们主要研究了三种基于预训练语言模型的流行分类器。预训练模型有多个版本，它们在层数和参数数量上有所不同。我们选择使用每种模型的基础版本。所有模型均在一台配备 GeForce GTX 1080 Ti GPU（11G RAM）和 Intel(R) Xeon(R) CPU E5-2630（128G RAM）的服务器上进行训练和测试。

在本研究中，我们评估了以下三种流行的基于预训练语言模型的分类器：

- BERT（BT）：采用双向 Transformer 架构的模型。它通过大规模语料库的预训练，学习了丰富的语言表示。基础版 BERT 具有 12 个层、隐藏层大小为 768、12 个自注意力头，总共 110M 参数。这种模型在多种 NLP 任务上展示了出色的性能。

- XLNet（XL）：这是一种采用自回归方式训练的语言模型，结合了 BERT 的双向上下文理解能力。XLNet 通过排列语言建模技术，使模型能够更有效地理解和预测文本中的词汇关系。

- RoBERTa（RB）：作为 BERT 的改进版本，RoBERTa 通过在更大的数据集上训练、增加批量大小，并移除某些原始 BERT 目标，如下一句预测（NSP），来提升性能。这些改进使得 RoBERTa 在不同的 NLP 任务上表现更为出色。

所有模型均在配备 GeForce GTX 1080 Ti GPU 和 Intel(R) Xeon(R) CPU 的高性能服务器上进行训练和测试。

压力测试案例

根据 [62] 的研究方法，我们设计了一系列压力测试案例，用于评估不同数据增强方法对模型抵御短路的影响。这些案例是根据表 5-1 中介绍的代理操作创建的，每种操作生成不同数量的测试案例。如 表 5-3 所示。为了评估测试短路的能力，我们将在下一节中使用这些测试案例的子集。

压力测试	ROC	COPA	ARCT	RECLOR
Neg+	1,797	492	297	375
Neg-	94	2	152	119
NER	362	0	5	0
PR	1,073	328	71	72
PI	861	219	56	91
Adv	1,850	496	444	500
CO	1,871	500	444	500
Syn	653	25	303	289
MT	1,871	500	444	500
Voice	1,014	246	174	263
Total	11,446	2,808	2,390	2,709

表 5-3　四个数据集中不同操作产生的压力测试案例数量。

5.5.2 模型短路问题测试

在本节中，我们的目标是选择合适的测试操作符来进行短路测试，并利用这些操作符来评估不同模型在处理多项选择题（MCQ）时短路的程度。

选择短路测试方法

正如在 5.3.2节中所讨论的，我们考虑了基于白盒注意力方法（AW [10]）和黑盒选项操作符中的一些等效类别来评估短路问题。我们现在将研究哪些代理测试更适合短路评估。

根据 5.3.2节的描述，每个测试操作符通过对模型选择正确答案的测试案例进行方向性改变来生成新的测试案例。如果模型在操作后仍然给出正确答案，我们认为它在该测试操作符下没有发生短路。假设人类注意力标注、注意力权重阈值和每个选项操作符都是可行的代理测试，我们可以获得 9 种不同的代理测试。

为了评估不同模型在处理多项选择题（MCQ）时的短路行为，我们从 ROC 测试集中随机抽取了 30 个问题。这些问题已被三种不同的模型——BERT、XLNet 和 RoBERTa——正确地回答过。为了进行短路测试，我们对每种模型应用了一系列的代理测试，每种测试都旨在从不同角度评估模型是否倾向于采取短路策略。

每个代理测试为每种模型生成了一个 30 维的 one-hot 向量，称为“代理向量”。在这个向量中，每个维度代表一个特定的 MCQ，而向量中的值（1 或 0）表示模型在该 MCQ 上是否发生了短路。具体来说，如果模型在某个 MCQ 上发生短路，相应的维度就被标记为 1；否则，就标记为 0[11]。

为了综合评估模型在所有代理测试下的表现，我们对每种模型计算了另一个向量，这个向量代表了所有代理测试的汇总结果。具体操作是，我们对每个 MCQ 的 30 维度进行多数投票。这意味着，对于每个 MCQ，我们检查所有代理测试的结果，并确定哪种结果（即发生短路或未发生短路）在所有测试中出现得最频繁。最终，这种结果被认为是该 MCQ 的综合评估结果。通过这种方式，我们能够从多个角度综合评估模型在特定问题上

[10]在此设置中，t_1 和 t_2 的阈值分别调整为 0.14 和 0.13，基于对 100 个人工标注案例的分析。这些案例从四个数据集的训练数据中随机抽取。

[11]对于某些不适用的代理测试 MCQ，我们随机标记为 1 或 0，以确保分析的一致性。

的短路概率，从而更全面地理解模型的短路行为。

代理测试类型的个体代理向量与集合向量之间较小的欧几里得距离表明了更高的可靠性。完整结果见表 5-4 。我们发现 CO 和 AW 的结果通常更接近集合结果，如较小距离所反映的那样。因此，我们认为 CO 和 AW 是更适合作为短路评估的代理测试。

测试类型	BERT	XLNet	RoBERTa	Ave
Neg+	3.16	3.87	2.45	3.16
Neg-	3.74	3.74	4.12	3.87
NER	3.87	3.87	4.12	3.95
PR	4.0	3.61	3.87	3.83
PI	3.74	3.74	3.74	3.74
CO	2.83	2.63	2.83	2.76
AW	2.45	3.46	2.45	2.79
Choice-only	4.0	3.74	3.87	3.87
Human	3.0	2.55	3.0	2.85

表 5-4　代理向量和集合向量在短路测试中的欧几里得距离（越小越好）。平均值是所有模型的平均分。每个模型的欧几里得距离最小的两个测试被加粗显示。

测试短路问题

我们通过分析注意力权重（AW）和交叉（CO）分数来测试模型是否发生短路。简而言之，较高的 AW/CO 分数意味着模型在解决问题时更加全面，不太可能简单地 “短路”。我们对 BERT、XLNet 和 RoBERTa 等多项选择分类器在四个不同的数据集上进行了微调。在 表 5-5 中，我们发现未经数据增强的原始模型（灰色部分）在 AW 和 CO 分数上普遍较低，暗示它们更容易发生短路。举个例子，XLNet 在 ROC 数据集上的 AW 分数低至 30% 以下，这表明在处理 ROC 数据集时，XLNet 极有可能采取了短路策略。

5.5.3　数据增强模型效果及分析

为了提高模型的整体鲁棒性，我们对 BERT、XLNet 和 RoBERTa 模型在不同数据集上进行了压力测试，并提出了数据增强策略以提升它们的

Model	Short circuit Tests		Robustness Tests	
	AW	CO	Original	Stress
BT(w/o)	98.76	90.80	86.58	81.93
BT+B	99.26	92.54	86.75	82.96
BT+C	99.69	98.47	87.07	84.34
BT+M	99.26	91.47	86.48	86.06
BT+C+M	98.82	97.78	86.75	88.60
XL(w/o)	28.08	83.28	90.81	79.22
XL+B	19.27	84.4	90.43	82.23
XL+C	64.58	98.81	89.47	86.23
XL+M	62.77	86.90	90.17	89.47
XL+C+M	60.25	97.10	90.22	92.64
RB(w/o)	77.41	88.76	92.73	82.33
RB+B	58.15	87.98	92.46	78.50
RB+C	82.71	99.3	91.18	88.92
RB+M	71.73	88.06	92.62	90.29
RB+C+M	93.31	97.44	91.88	93.06

(a) ROC

Model	Short circuit Tests		Robustness Tests	
	AW	CO	Original	Stress
BT(w/o)	89.68	68.71	62.00	57.40
BT+B	96.79	85.42	68.60	68.95
BT+C	98.35	97.25	72.80	78.84
BT+M	95.17	90.62	70.40	79.62
BT+C+M	96.69	96.13	72.40	80.68
XL(w/o)	93.16	60.26	61.40	57.71
XL+B	91.46	65.51	63.20	61.06
XL+C	45.13	94.69	67.80	75.42
XL+M	76.85	57.23	62.20	71.10
XL+C+M	98.51	83.93	67.20	81.32
RB(w/o)	80.89	78.01	76.40	74.85
RB+B	96.36	83.64	77.00	80.26
RB+C	89.62	98.23	79.00	83.31
RB+M	62.26	84.30	72.60	83.53
RB+C+M	61.89	92.70	74.00	87.30

(b) COPA

Model	Short circuit Tests		Robustness Tests	
	AW	CO	Original	Stress
BT(w/o)	99.65	78.52	63.96	58.08
BT+B	99.34	61.18	68.47	56.21
BT+C	98.37	96.08	68.92	65.73
BT+M	98.67	74.42	67.79	69.65
BT+C+M	98.00	90.0	67.57	73.71
XL(w/o)	85.67	59.10	75.45	61.72
XL+B	95.73	60.40	79.05	64.78
XL+C	55.59	92.45	74.55	69.93
XL+M	95.74	59.57	74.10	73.15
XL+C+M	86.26	90.35	77.03	79.11
RB(w/o)	99.14	60.29	78.83	66.16
RB+B	97.78	60.94	81.31	66.02
RB+C	79.19	92.77	77.93	70.64
RB+M	100.00	68.13	77.03	76.64
RB+C+M	71.47	93.39	75.00	78.97

(c) ARCT

Model	Short circuit Tests		Robustness Tests	
	AW	CO	Original	Stress
BT(w/o)	82.46	50.88	45.60	33.91
BT+B	86.01	61.73	48.60	35.99
BT+C	80	96.17	47.00	47.72
BT+M	82.48	58.55	46.80	50.02
BT+C+M	96.79	87.16	43.60	53.79
XL(w/o)	79.64	62.86	56.00	39.77
XL+B	81.40	74.04	57.0	44.6
XL+C	87.87	98.90	54.40	51.66
XL+M	72.76	70.15	53.60	56.99
XL+C+M	48.71	88.56	54.2	58.63
RB(w/o)	85.88	70.2	51.00	36.76
RB+B	15.69	73.73	51.00	38.71
RB+C	89.68	96.83	50.40	50.88
RB+M	100.00	80.38	52.00	59.95
RB+C+M	89.26	88.43	48.40	55.78

(d) RECLOR

表 5-5　短路和鲁棒性测试：对模型上在四个任务进行了带有或不带有(w/o)数据增强的测试。+B = 使用反向翻译进行数据增强，+C = 使用交叉（crossover）进行数据增强，+M = 使用变异（mutation）进行数据增强，CO = 交叉（crossover），AW = 注意力权重评估。压力测试包括 表 5-3 中的所有案例。

性能。我们的分析显示，这些模型普遍缺乏鲁棒性，特别是在面对挑战性强的情景时。为了应对这一问题，我们采用了交叉（+C）、变异（+M）以及这两者的组合（+C+M）等数据增强方法，并将其效果与反向翻译（+B）作为基线进行比较。

模型弱点

正如表 5-5所示，BERT、XLNet 和 RoBERTa 模型在面临压力测试时表现出了明显的性能下降。例如，XLNet 在 ROC 数据集上训练时，其准确率下降了 11.59%，AW 分数仅为 28.8%，这表明该模型在处理 ROC 问题时可能过度依赖于短路策略。同样，在 RECLOR 和 ARCT 数据集上，三个模型的性能也普遍下降约 10%，与较低的 CO 分数相符，暗示短路问题可能是导致性能下降的一个原因。

数据增强

为了缓解识别出的弱点，我们使用两种主要的数据增强方法对模型进行了训练：交叉和变异，这些在前一节中已经讨论过。我们还通过构建同时包含这两种技术的训练数据，组合了这两种方法（+C+M）。我们使用反向翻译 [153] 作为数据增强的基线，因为它在以前的工作中已显示出普遍性和有效性。扩展的数据量与原始数据量保持一致。

表 5-5 展示了“原始测试”的结果。我们观察到，四种数据增强方法不仅没有对模型在原始数据集上的性能产生负面影响，甚至可能帮助模型实现更好的准确性。例如，在 ROC 数据集上，用交叉增强数据训练的 BERT 和 RoBERTa 模型的准确率超过了基础模型，排名第一。交叉方法在 COPA 上也证明是有效的。尽管反向翻译在 A RCT 和 RECLOR 上大多获得更高的分数，+C、+M 和 +C+M 方法与基础模型相比仅略微逊色。

考虑到表 5-5 中的 “Stress” 列，我们发现不同方法显示出不同程度的鲁棒性。总体而言，+C+M 方法在压力测试上表现最好，除了在 RECLOR 数据集上训练 RoBERTa 的情况。这一结果表明，这种类型的数据可以保护模型免受简单扰动的困扰，增强模型的鲁棒性。然而，反向翻译并没有显著提高模型的鲁棒性。虽然单独的交叉方法可以在压力测试下有助于鲁棒性，但它不如 +M 和 +C+M 方法有效。

进一步分析使用短路测试的模型表明，交叉方法始终获得最高的 CO 分数，并且在 AW 分数中通常排名最高。这一发现表明，用交叉数据增强训练的模型更有可能考虑前提，避免短路问题。

结果

总之，我们的研究表明，在开发自然语言理解任务的机器学习模型时，解决模型的鲁棒性和短路问题非常重要。通过研究 BERT、XLNet 和 RoBERTa 模型在不同数据集上的弱点，我们发现这些模型在压力测试下普遍不具备鲁棒性，短路问题是导致它们不稳定的原因之一。

为了克服这些挑战，我们提出并评估了数据增强方法，包括交叉、变异以及两者的组合（+C+M），并将它们与反向翻译基线进行了比较。我们的结果显示，这些数据增强技术不仅保持或提高了模型在原始数据集上的性能，而且在压力测试下显著增强了模型的鲁棒性。特别是，+C+M 方法在大多数情况下表现最佳。

此外，我们从短路测试的发现中得知，交叉方法始终获得最高的 CO 分数，并且在 AW 分数中通常排名最高，表明用交叉数据增强训练的模型更有可能考虑前提，避免短路问题。

未来的工作可以探索额外的数据增强技术及其组合，以进一步增强模型的鲁棒性并减轻短路问题。此外，调查这些增强方法在各种自然语言理解任务和语言中的可转移性，可以为这些方法的普遍性提供有价值的见解。

5.5.4 案例研究

在我们的案例研究中，我们采用了一系列白盒测试，专注于分析注意力模式的变化及其对模型决策的影响。

具体来说，我们选取了一个来自 ROC 数据集的例子进行详细分析，如 表 5-2 所展示的那样。这个案例是围绕一个基于 BERT 模型的注意力图进行的分析，如 图 5-3 所示。在这个特定的例子中，我们观察到，前提中的词 “show” 与正确选项中的短语 “real ity show” 在人类知识中具有显著的关联。

然而，在原始训练集上训练的 BERT 模型未能正确选出与前提中 “show” 相关的选项，可能是因为在选择项和前提之间几乎没有形成有效的注意力连接。这一发现表明，原始模型可能在处理这类问题时忽略了重要的上下文信息。

值得注意的是，经过交叉数据增强训练后的模型显示出了显著的改进。在这种情况下，模型学会了更多地关注前提和前提与选择项之间的联系，例

如，在我们的案例中，它能够识别出 “show” 一词的重要性。类似的趋势也出现在经过变异操作增强训练的模型（即 “BT+M”）中，以及交叉和变异操作的组合（即 “BT+C+M”）中。

这种注意力模式的变化揭示了一个重要的现象:在经过交叉(“BT+C”)、变异（“BT+M”）以及这两者组合（“BT+C+M”）的增强数据训练的模型中，模型能够有效地结合前提中的信息，更准确地从错误选项中区分出正确选项。相比之下，反向翻译（即 “BT+B”）的注意力图显示出较为浅色和稀疏的注意力区块，表明这种增强技术在帮助 BERT 模型建立选择项和前提之间的联系方面并不十分有效。

5.6 本章小结

在本章中，我们全面探讨了自然语言理解（NLU）模型在常识性推理任务中的鲁棒性问题，特别关注了模型的 “短路” 现象。我们发现，尽管神经网络模型在多个任务中表现出色，但它们在面对未知或对抗性数据时的脆弱性揭示了一个关键的缺陷：它们往往依赖于数据中的表面规律而非深入理解，这限制了它们的泛化能力。我们将这种现象定义为 “短路”，即模型依赖简单规律而忽略深层次逻辑推理的倾向。

为了精确探测和深入分析 NLU 模型在处理多项选择题时的这种短路行为，我们采用了两种方法：白盒方法和黑盒方法。在白盒方法中，我们利用模型内部的注意力机制来观察模型如何在不同选项和前提之间分配关注度。这种方法直接反映了模型内部的决策过程，帮助我们理解模型是否在依赖关键词汇的浅层连接，而非深入理解语境。与此同时，黑盒方法通过在原始数据上施加特定的操作（如 NER 改变），创造新的代理测试用例，来考验模型在不同情境下的表现。这些测试用例的设计旨在模拟实际应用中可能遇到的多样化和复杂情境，从而检验模型的泛化能力和逻辑推理能力。

针对模型短路问题，我们提出了一系列创新的数据增强技术，包括受生物学启发的 “ 交叉” 和 “变异” 操作。这些技术通过改变原有数据集的结构，引入新的组合和变体，促使模型在训练过程中不仅关注表面的统计规律，而是更加关注问题的内在逻辑和结构。在实验中，我们将这些数据增强技术应用于包括 BERT、RoBERTa 和 XLNet 在内的先进神经网络模型，并在包括 ROC、COPA、ARCT 和 RECLOR 四个主要的常识性推理基准

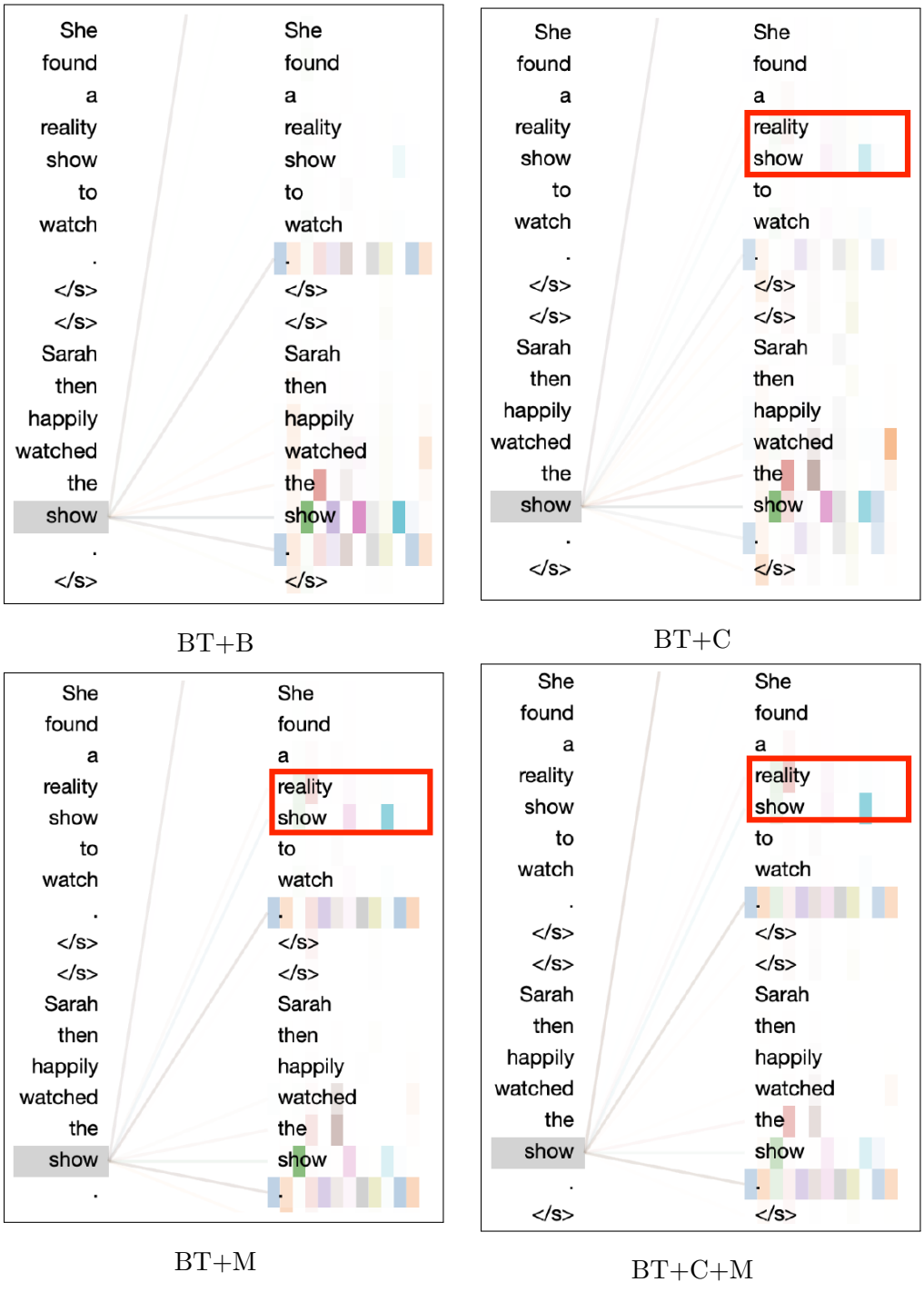


图 5-3　在 ROC 数据集例子上，基于 BERT 模型的注意力图。

数据集上进行了全面测试。结果显示，这些增强技术显著提升了模型在多样化环境中的鲁棒性，同时在标准测试集上保持或提高了性能。

本章的研究不仅揭示了现有 NLU 模型在常识性推理任务中面临的挑战，还提供了解决这些挑战的具体方法。我们的数据增强策略证明了模型的结构本身具有学习更好泛化能力的潜力，但这种潜力常常因为模型过度依赖数据中的“短路”现象而未能充分发挥。未来的研究可以在本研究的基础上，进一步探索和优化数据增强技术，以提升 NLU 模型在更广泛应用场景下的泛化能力和鲁棒性。同时，探索这些方法在其他类型的自然语言处理任务中的应用，将是一个值得追求的方向。

第六章 全文总结

6.1 主要结论

在本论文的收尾章节中，我们将综合回顾针对常识性推理领域所进行的研究工作，这些研究旨在解决人工智能领域提升模型常识性推理能力，鲁棒性和可解释性的关键挑战。通过一系列创新的方法和实验，我们不仅推进了理论研究的边界，还为构建更加高效、鲁棒、和透明的人工智能系统提供了实践基础。以下内容详细梳理了我们在提升常识性推理能力、增强模型鲁棒性和提高模型可解释性这三个核心领域的主要成果。

1. 常识性推理能力的提升

在常识性推理领域，本研究的核心成就是显著提升了模型在故事理解任务中的常识性推理能力。我们针对的主要问题是：如何让模型更好地理解和预测故事的结局，特别是在这些故事涉及到复杂的常识知识时。为了解决这一挑战，我们采用了一种多层次的方法。

首先，我们对故事中的句子进行了简化处理，这有助于模型更清晰地识别和理解故事中的关键概念和事件。其次，我们基于这些关键元素，构建了一个更加丰富和细致的故事表征，这意味着模型不仅仅处理文本的表面意义，而且还能理解故事的深层次结构和含义。最后，我们利用这种复杂的表示来预测故事的结局。

为了进一步加强这一过程，我们结合了概念序列编码和利用 ConceptNet 预训练的概念图编码。这一步骤是关键的，因为它帮助模型捕捉到故事中的深层关系和隐含逻辑。实验结果表明，这种方法大幅减少了训练数据中的偏见和信息泄露，并且当结合了 ConceptNet 的预训练概念编码向量后，模型的性能得到了显著提升。

2. 推理模型的可解释性研究

在常识性推理模型的可解释性方面，本研究开创性地开发了两种测试框架，从宏观和微观两个层面全面评估和分析模型的偏见。我们的目标是使模型的决策过程更加透明和可理解。

宏观层面上，我们通过评估不同数据集上基于统计特征的简单分类模型的最高准确率，以及与多数选择的偏差值，来评估数据集中存在的偏见特征程度。这种方法揭示了模型对某些统计规律的过度依赖，有助于我们理解模型泛化能力的局限性。

微观层面上，我们设计了 ICQ（“I-see-cue”）框架。这个框架通过结合特征的分布不平衡性和训练集与测试集分布的相似性，来发现可能影响模型的关键特征，并评估数据集中存在的问题。通过这种多维特征划分和细致的性能分析，我们能够探究模型在不同特征上的准确性和分布表现。此外，我们还开发了直观的可视化工具，以更有效地识别和理解模型性能差异的根源。

通过这些创新的分析框架，我们不仅深入理解了模型泛化能力的限制，还为设计更健壮、更可解释的人工智能系统奠定了基础。

3. **推理模型的鲁棒性提升**

我们在自然语言推理模型的鲁棒性方面取得了进一步的进展。我们面临的主要问题是：模型在处理未知或对抗性数据时表现出的脆弱性。我们发现，尽管这些模型在特定任务上表现良好，但它们往往依赖于数据中的表面规律，而非深入的逻辑推理。这种现象被我们定义为 “短路”，指的是模型倾向于依赖简单规律而忽略深层次的逻辑推理。

为了说明 “短路” 这一问题，我们采用了两种方法：白盒方法和黑盒方法。白盒方法让我们可以直接观察模型的内部决策过程，特别是模型是如何在不同选项和前提之间分配注意力的。这有助于我们判断模型是否仅依赖于关键词汇的浅层连接。而黑盒方法则通过在原始数据上进行特定的操作，例如改变命名实体，来创造新的测试用例。这些测试用例模拟了实际应用中可能遇到的多样化和复杂情境，用于检验模型的泛化能力和逻辑推理能力。

此外，我们还引入了一系列创新的数据增强技术，包括受生物学启发的 “交叉” 和 “变异” 操作。这些技术通过改变原有数据集的结构，促使模型在训练过程中更加关注问题的内在逻辑和结构，而非表面的统计规律。实验结果显示，这些技术显著提升了模型在多样化环境中的鲁棒性，同时在标准测试集上保持或提高了性能。

6.2 研究展望

在人工智能的迅猛发展中，大型模型如 ChatGPT 的出现开启了一个新的时代，他们不仅在常识性推理领域，同时在更广泛的自然语言处理领域展现了前所未有的能力。它们在处理语言、理解复杂情境、甚至生成创新内容方面的表现，为我们揭示了 AI 技术未来的无限可能。然而，随着这些模型在各个领域的深入应用，我们也逐渐意识到，要充分发挥它们的潜力，同时确保安全和可靠的应用，还需要在几个关键领域进行深入研究。我认为在下一个阶段的研究当中是必然要拥抱大模型的，我的研究展望主要包括下面的三个方面。

首先，尽管大型模型在常识性推理任务上已取得显著进展，但它们在特定场景下的推理能力仍有待加强。未来的研究需要关注如何通过结合更高级的算法和丰富的数据资源，进一步提升这些模型在理解复杂逻辑、因果关系及多维问题上的能力。特别是在一些专业领域，如医疗诊断或法律分析等，这种深化的推理能力将是关键。

接着，大型模型的“黑盒”特性使得它们的决策过程难以解释和理解。这不仅限制了它们在某些场合的应用，也引发了关于透明度和可信赖度的问题。因此，深入探索和提高这些模型的可解释性，将是实现更广泛应用的前提。这涉及到开发新的理论框架，使我们能够更清晰地理解这些模型如何处理信息、做出决策，并向用户展示这一过程。

最后，随着大型模型越来越多地涉及处理敏感数据和执行重要任务，它们的安全性成为了一个不可回避的议题。这不仅关系到个人隐私的保护，更涉及到国家安全层面的考量。因此，构建全面的风险评估和管理体系，确保大型模型即使在处理高风险任务时也能保持稳定和安全，是未来研究中的一项重要任务。

综合来看，未来在这些关键领域的研究不仅将推动大型模型在技术上的进一步发展，也将帮助我们更好地融合人工智能技术与社会需求，确保这些先进技术的发展同时符合伦理标准和安全要求。通过不断探索和创新，我们有望见证一个更智能、更安全、更可靠的人工智能应用时代的到来。

参 考 文 献

[1] Mostafa Abdou, Vinit Ravishankar, Maria Barrett, Yonatan Belinkov, Desmond Elliott, and Anders Søgaard. The sensitivity of language models and humans to winograd schema perturbations. In ACL, pages 7590–7604, 2020.

[2] Niranjan Balasubramanian, Stephen Soderland, Oren Etzioni, et al. Generating coherent event schemas at scale. In EMNLP, 2013.

[3] Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. In International Conference on Learning Representations, 2018.

[4] Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. In ICLR, 2018.

[5] Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. Transactions of the Association for Computational Linguistics, 7:49–72, 2019.

[6] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943, 2018.

[7] Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. The fifth pascal recognizing textual entailment challenge. TAC, 7:8, 2009.

[8] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic dialectal variation in social media: A case study of african-american english. In EMNLP, pages 1119–1130, 2016.

[9] George Boole. An investigation of the laws of thought on which are founded the mathematical theories of logic and probabilities by george boole. Walton and Maberly, 1854.

[10] Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. Bad characters: Imperceptible nlp attacks. In 2022 IEEE Symposium on Security and Privacy (SP), pages 1987–2004. IEEE, 2022.

[11] Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In EMNLP, pages 632–642, 2015.

[12] Samuel Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D Manning, and Christopher Potts. A fast unified model for parsing and sentence understanding. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1466–1477, 2016.

[13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

[14] Michael Bugert, Yevgeniy Puzikov, Andreas Rücklé, Judith Eckle-Kohler, Teresa Martin, Eugenio Martínez-Cámara, Daniil Sorokin, Maxime Peyrard, and Iryna Gurevych. LSDSem 2017: Exploring data generation methods for the story cloze test. In LSDSem, 2017.

[15] Zheng Cai, Lifu Tu, and Kevin Gimpel. Pay attention to the ending: Strong neural baselines for the roc story cloze task. In ACL, 2017.

[16] Nathanael Chambers and Dan Jurafsky. Unsupervised learning of narrative event chains. ACL, 2008.

[17] Nathanael Chambers and Dan Jurafsky. Unsupervised learning of narrative schemas and their participants. In ACL-IJCNLP, 2009.

[18] Nathanael Chambers and Dan Jurafsky. Template-based information extraction without the templates. In HLT, 2011.

[19] Snigdha Chaturvedi, Haoruo Peng, and Dan Roth. Story comprehension for predicting what happens next. In EMNLP, 2017.

[20] Jiaao Chen, Jianshu Chen, and Zhou Yu. Incorporating structured commonsense knowledge in story completion. CoRR, 2018.

[21] Jiaao Chen, Zichao Yang, and Diyi Yang. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In ACL, pages 2147–2157, 2020.

[22] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced lstm for natural language inference. In ACL, pages 1657–1668, 2017.

[23] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In EMNLP, page 1724. Association for Computational Linguistics, 2014.

[24] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In EMNLP-IJCNLP, pages 4060–4073, 2019.

[25] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. In Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 276–286, 2019.

[26] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. arXiv preprint arXiv:1805.09501, 2018.

[27] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 702–703, 2020.

[28] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In Machine learning challenges workshop, pages 177–190. Springer, 2005.

[29] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In ACL, pages 2978–2988, 2019.

[30] Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. Communications of the ACM, 58(9):92–103, 2015.

[31] Ernest Davis, Leora Morgenstern, and Charles L Ortiz Jr. The first winograd schema challenge at ijcai-16. AI Magazine, 38(3):97–98, 2017.

[32] Augustus De Morgan. Formal logic: or, the calculus of inference, necessary and probable. Taylor and Walton, 1847.

[33] Rina Dechter. Reasoning with probabilistic and deterministic graphical models: Exact algorithms, 2013.

[34] Dorottya Demszky, Devyani Sharma, Jonathan H Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. Learning to recognize dialect features. In NAACL-HLT, pages 2315–2338, 2021.

[35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT, pages 4171–4186. Association for Computational Linguistics, 2019.

[36] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification. In ACL, pages 31–36, 2018.

[37] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In EMNLP, pages 489–500, 2018.

[38] Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for nlp. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 968–988, 2021.

[39] Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, Elena Cabrio, and Bill Dolan. The fourth pascal recognizing textual entailment challenge. In TAC, 2008.

[40] Oren Glickman. Applied textual entailment. Citeseer, 2006.

[41] Yoav Goldberg. Assessing bert's syntactic abilities. arXiv preprint arXiv:1901.05287, 2019.

[42] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, ICLR, 2015.

[43] Jian Guan, Yansen Wang, and Minlie Huang. Story ending generation with incremental encoding and commonsense knowledge. arXiv preprint arXiv:1808.10113, 2018.

[44] Hongyu Guo, Yongyi Mao, and Richong Zhang. Augmenting data with mixup for sentence classification: An empirical study. arXiv preprint arXiv:1905.08941, 2019.

[45] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. In NAACL-HLT, pages 107–112, 2018.

[46] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In NAACL-HLT, pages 1930–1940, 2018.

[47] R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second pascal recognising textual entailment challenge. In Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment, volume 7, pages 785–794, 2006.

[48] He He, Sheng Zha, and Haohan Wang. Unlearn dataset bias in natural language inference by fitting the residual. EMNLP-IJCNLP, page 132, 2019.

[49] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 1997.

[50] Linmei Hu, Juanzi Li, Liqiang Nie, Xiao-Li Li, and Chao Shao. What happens next? future subevent prediction using contextual hierarchical lstm. In AAAI, 2017.

[51] Lifu Huang, Taylor Cassidy, Xiaocheng Feng, Heng Ji, Clare R Voss, Jiawei Han, and Avirup Sil. Liberal event extraction and event schema induction. In ACL, 2016.

[52] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. Learning deep structured semantic models for web search using clickthrough data. In CIKM, 2013.

[53] Shanshan Huang, Kenny Q. Zhu, Qianzi Liao, Libin Shen, and Yinggong Zhao. Enhanced story representation by conceptnet for predicting story endings. In CIKM, pages 3277–3280, 2020.

[54] Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. In NAACL-HLT, 2018.

[55] Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. On feature learning in the presence of spurious correlations. NIPS, 35:38516–38532, 2022.

[56] Sarthak Jain and Byron C Wallace. Attention is not explanation. In NAACL-HLT, pages 3543–3556, 2019.

[57] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In EMNLP, pages 2021–2031, 2017.

[58] Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In ACL, pages 2177–2190, 2020.

[59] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In AAAI, volume 34, pages 8018–8025, 2020.

[60] Nitish Joshi, Xiang Pan, and He He. Are all spurious features in natural language alike? an analysis through a causal lens. In EMNLP, pages 9804–9817, 2022.

[61] Armand Joulin, Édouard Grave, Piotr Bojanowski, and Tomáš Mikolov. Bag of tricks for efficient text classification. In ECAL, pages 427–431, 2017.

[62] Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors. ACL, 2020.

[63] Aishwarya Kamath and Rajarshi Das. A survey on semantic parsing. In Automated Knowledge Base Construction (AKBC), 2018.

[64] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In NIPS, 2015.

[65] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. Advances in neural information processing systems, 35:22199–22213, 2022.

[66] Alice Lai and Julia Hockenmaier. Illinois-lh: A denotational and distributional approach to semantics. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 329–334, 2014.

[67] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 785–794, 2017.

[68] George Lakoff. Linguistics and natural logic. Synthese, 22(1-2):151–271, 1970.

[69] Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In ICML, 2020.

[70] Qian Li, Ziwei Li, Jin-Mao Wei, Yanhui Gu, Adam Jatowt, and Zhenglu Yang. A multi-attention based neural network with external knowledge for story ending predicting task. In COLING, 2018.

[71] Zhongyang Li, Xiao Ding, and Ting Liu. Story ending prediction by transferable bert. arXiv preprint arXiv:1905.07504, 2019.

[72] Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. Birds have four legs?! numersense: Probing numerical commonsense knowledge of pre-trained language models. In EMNLP, pages 6862–6868, 2020.

[73] Hongyu Lin, Le Sun, and Xianpei Han. Reasoning with heterogeneous knowledge for commonsense machine comprehension. In EMNLP, 2017.

[74] J. Lin. Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory, 37(1):145–151, 1991.

[75] Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. Linguistic knowledge and transferability of contextual representations. In NAACL-HLT, pages 1073–1094, 2019.

[76] Nelson F Liu, Roy Schwartz, and Noah A Smith. Inoculation by fine-tuning: A method for analyzing challenge datasets. In NAACL-HLT, 2019.

[77] Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. Knowledge diffusion for neural dialogue generation. In ACL, 2018.

[78] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

[79] Ryan Lowe, Nissan Pow, Iulian Vlad Serban, and Joelle Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured

multi-turn dialogue systems. In Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 285–294, 2015.

[80] Zhiyi Luo, Yuchen Sha, Kenny Q. Zhu, Seung-won Hwang, and Zhongyuan Wang. Commonsense causal reasoning between short texts. In Fifteenth international conference on the principles of knowledge representation and reasoning, 2016.

[81] John McCarthy. Programs with common sense, 1959.

[82] Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In ACL, pages 3428–3448, 2019.

[83] James R Meehan. Tale-spin, an interactive program that writes stories. In IJCAI, 1977.

[84] Todor Mihaylov and Anette Frank. Story cloze ending selection baselines and data examination. In LSDSem, 2017.

[85] Todor Mihaylov and Anette Frank. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In ACL, 2018.

[86] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[87] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In NIPS, 2013.

[88] Ashutosh Modi. Event embeddings for semantic script modeling. In CoNLL, 2016.

[89] Ashutosh Modi and Ivan Titov. Inducing neural models of script knowledge. In CoNLL, 2014.

[90] Raymond J. Mooney and Gerald DeJong. Learning schemata for natural language processing. In IJCAI, 1985.

[91] Leora Morgenstern, Ernest Davis, and Charles L Ortiz. Planning, executing, and evaluating the winograd schema challenge. AI Magazine, 37(1):50–54, 2016.

[92] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In NAACL-HLT, pages 839–849, 2016.

[93] Nasrin Mostafazadeh, Lucy Vanderwende, Wen-tau Yih, Pushmeet Kohli, and James Allen. Story cloze evaluator: Vector space representation evaluation by predicting what happens next. ACL, 2016.

[94] Tsendsuren Munkhdalai and Hong Yu. Neural tree indexers for text understanding. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 11–21, 2017.

[95] Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. Stress test evaluation for natural language inference. In COLING, pages 2340–2353, 2018.

[96] Hwee Tou Ng, Leong Hwee Teo, and Jennifer Lai Pheng Kwan. A machine learning approach to answering questions for reading comprehension tests. In 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pages 124–132, 2000.

[97] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In ACL, pages 4885–4901, 2020.

[98] Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. In ACL, pages 4658–4664, 2019.

[99] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35:27730–27744, 2022.

[100] CS Peirce. A theory of probable inference. In Studies in logic by members of the Johns Hopkins University., pages 126–181. Little, Brown and Co, 1883.

[101] Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu, and Qun Liu. Copen: Probing conceptual knowledge in pre-trained language models. In EMNLP, pages 5015–5035, 2022.

[102] Haoruo Peng, Snigdha Chaturvedi, and Dan Roth. A joint model for semantic sequences: Frames, entities, sentiments. In CoNLL, 2017.

[103] Haoruo Peng and Dan Roth. Two discourse driven language models for semantics. In ACL, 2016.

[104] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In EMNLP, pages 1532–1543, 2014.

[105] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621, 2017.

[106] Karl Pichotta and Raymond J Mooney. Statistical script learning with multi-argument events. EACL, 2014.

[107] Karl Pichotta and Raymond J Mooney. Learning statistical scripts with lstm recurrent neural networks. In AAAI, 2016.

[108] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, pages 180–191, 2018.

[109] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is chatgpt a general-purpose natural language processing task solver? arXiv preprint arXiv:2302.06476, 2023.

[110] Yanru Qu, Dinghan Shen, Yelong Shen, Sandra Sajeev, Weizhu Chen, and Jiawei Han. Coda: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding. In ICLR, 2020.

[111] Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. Technical report, Technical report, OpenAI, 2018.

[112] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. 2019.

[113] Rajat Raina, Andrew Y Ng, and Christopher D Manning. Robust textual inference via learning and abductive reasoning. In AAAI, pages 1099–1105, 2005.

[114] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In ACL, pages 784–789, 2018.

[115] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In EMNLP, pages 2383–2392, 2016.

[116] Michaela Regneri, Alexander Koller, and Manfred Pinkal. Learning script knowledge with web experiments. In ACL, 2010.

[117] Michaela Regneri, Alexander Koller, Josef Ruppenhofer, and Manfred Pinkal. Learning script participants from unlabeled data. In RANLP, 2011.

[118] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging nlp models. In ACL, pages 856–865, 2018.

[119] Marco Túlio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with checklist. In ACL, pages 4902–4912, 2020.

[120] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In 2011 AAAI Spring Symposium Series, 2011.

[121] Melissa Roemmele, Sosuke Kobayashi, Naoya Inoue, and Andrew Gordon. An rnn-based binary classifier for the story cloze test. In LSDSem, 2017.

[122] Rachel Rudinger, Pushpendre Rastogi, Francis Ferraro, and Benjamin Van Durme. Script induction as language modeling. In EMNLP, 2015.

[123] Ivan Sanchez, Jeff Mitchell, and Sebastian Riedel. Behavior analysis of nli models: Uncovering the influence of three factors on robustness. In NAACL-HLT, pages 1975–1985, 2018.

[124] Roger C Schank and Robert P Abelson. Scripts, plans, and knowledge. In IJCAI, 1975.

[125] Roger C Schank and Robert P Abelson. Scripts, plans, goals, and understanding: An inquiry into human knowledge structures. Psychology Press, 2013.

[126] Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. Towards debiasing fact verification models. In EMNLP-IJCNLP, pages 3419–3425, 2019.

[127] Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A Smith. Story cloze task: Uw nlp system. In LSDSem, 2017.

[128] Leo Schwinn, René Raab, An Nguyen, Dario Zanca, and Bjoern Eskofier. Exploring misclassifications of robust neural networks to enhance adversarial attacks. Applied Intelligence, pages 1–17, 2023.

[129] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In ACL, pages 86–96, 2016.

[130] Lei Sha, Sujian Li, Baobao Chang, and Zhifang Sui. Joint learning templates and slots for event schema induction. In Proceedings of NAACL-HLT, 2016.

[131] Rishi Sharma, James Allen, Omid Bakhshandeh, and Nasrin Mostafazadeh. Tackling the story ending biases in the story cloze test. In ACL, 2018.

[132] Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. arXiv preprint arXiv:2009.13818, 2020.

[133] Robin Smith et al. Prior analytics. Hackett Publishing, 1989.

[134] Robert Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In AAAI, 2017.

[135] Robert Speer and Joanna Lowry-Duda. Conceptnet at semeval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In SemEval@ACL, 2017.

[136] Siddarth Srinivasan, Richa Arora, and Mark Riedl. A simple and effective approach to the story cloze test. In ACL, 2018.

[137] Shane Storks, Qiaozi Gao, and Joyce Y Chai. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. arXiv preprint arXiv:1904.01172, 2019.

[138] Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. Dream: A challenge data set and models for dialogue-based reading comprehension. TACL, 7:217–231, 2019.

[139] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In ICLR, 2014.

[140] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In NAACL-HLT, pages 4149–4158, 2019.

[141] Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. It's morphin'time! combating linguistic discrimination with inflectional perturbations. In ACL, pages 2920–2935, 2020.

[142] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. What do you learn from context? probing for sentence structure in contextualized word representations. In ICLR, 2019.

[143] James Thorne, Andreas Vlachos, CFEVER: a Large-scale Dataset for Fact Extraction Christodoulopoulos, VERificationhristos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In ACL, 2018.

[144] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[145] Jesse Vig. A multiscale visualization of attention in the transformer model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 37–42, 2019.

[146] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. EMNLP, page 353, 2018.

[147] Bingning Wang, Kang Liu, and Jun Zhao. Conditional generative adversarial networks for commonsense machine comprehension. In IJCAI, 2017.

[148] Xuezhi Wang, Haohan Wang, and Diyi Yang. Measure and improve robustness in nlp models: A survey. In NAACL-HLT, pages 4569–4586, 2022.

[149] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. In ICLR, 2016.

[150] Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. Inference is everything: Recasting semantic resources into

a unified evaluation framework. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 996–1005, 2017.

[151] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In NAACL-HLT, pages 1112–1122, 2018.

[152] Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. Conditional bert contextual augmentation. In International Conference on Computational Science, pages 84–95, 2019.

[153] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. In Proceedings of the 34th International Conference on Neural Information Processing Systems, pages 6256–6268, 2020.

[154] Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. Large-scale cloze test dataset created by teachers. In EMNLP, pages 2344–2356, 2018.

[155] Yadollah Yaghoobzadeh, Remi Tachet, TJ Hazen, and Alessandro Sordoni. Robust natural language inference models with example forgetting. arXiv preprint arXiv:1911.03861, 2019.

[156] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: generalized autoregressive pretraining for language understanding. In NIPS, pages 5753–5763, 2019.

[157] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning. In ICLR, 2020.

[158] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In EMNLP, pages 93–104, 2018.

[159] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. ICLR, 2017.

[160] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In NAACL, 2018.

[161] Mantong Zhou, Minlie Huang, and Xiaoyan Zhu. Story ending selection by finding hints from pairwise candidate endings. IEEE/ACM TASLP, 01 2019.

[162] Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for natural language understanding. In ICLR, 2019.

[163] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In ICCV, pages 19–27, 2015.

攻读学位期间学术论文和科研成果目录

[1] 一作：Shanshan Huang, Siyu Ren, Kenny Q. Zhu: Combating Short Circuit Behavior in Natural Language Reasoning: Crossover and Mutation Operations for Enhanced Robustness. ECAI 2023: 1092-1099（已发表）

[2] 一作：Shanshan Huang, Kenny Q. Zhu: Statistically Profiling Biases in Natural Language Reasoning Datasets and Models. EMNLP 2023 (findings) (已录用)

[3] 一作：Shanshan Huang: Can You Really Reason: A Novel Framework for Assessing Natural Language Reasoning Datasets and Models. ICONIP (15) 2023: 54-66（已发表）

[4] 一作：Shanshan Huang, Kenny Q. Zhu, Qianzi Liao, Libin Shen, Yinggong Zhao: Enhanced Story Representation by ConceptNet for Predicting Story Endings. CIKM 2020: 3277-3280（已发表）

[5] 二作：Zhiyi Luo, Shanshan Huang, Kenny Q. Zhu: Knowledge empowered prominent aspect extraction from product reviews. Inf. Process. Manag. 56(3): 408-423 (2019)（已发表）

[6] 二作：Zhiyi Luo, Shanshan Huang, Frank F. Xu, Bill Y uchen Lin, Hanyuan Shi, Kenny Q. Zhu: ExtRA: Extracting Prominent Review Aspects from Customer Feedback. EMNLP 2018: 3477-3486（已发表）

[7] 一作：Noise-Enhanced Commonsense Reasoning in NLP Models（在投）

致 谢