硕 士 学 位 论 文

基于深度学习的习题理解和应用算法研究

Research on Exercise Understanding and Application Algorithms Based on Deep Learning

作者：冯梦菲

导师：陈一帅

北京交通大学

2020 年 6 月

学位论文版权使用授权书

本学位论文作者完全了解北京交通大学有关保留、使用学位论文的规定。特授权北京交通大学可以将学位论文的全部或部分内容编入有关数据库进行检索，提供阅览服务，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。学校可以为存在馆际合作关系的兄弟高校用户提供文献传递服务和交换服务。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：冯梦菲　　导师签名：陈一帅

签字日期：2020 年 6 月 7 日　　签字日期：2020 年 6 月 7 日

学校代码：10004 密级：公开

北京交通大学

硕士学位论文

基于深度学习的习题理解和应用算法研究

Research on Exercise Understanding and Application Algorithms Based on Deep Learning

作者姓名：冯梦菲 学　　号：17120052

导师姓名：陈一帅 职　　称：副教授

学位类别：工学 学位级别：硕士

学科专业：通信与信息系统 研究方向：信息网络

北京交通大学

2020 年 6 月

致谢

本论文是在我的导师陈一帅老师的悉心指导下完成的。从 17 年本科毕业设计以来，陈老师对我谆谆教诲、认真耐心。每当我遇到困难时，陈老师都会不断鼓励我，与我一起共同面对；每当我懈怠时，陈老师都会及时纠正我，让我步入正轨。陈老师作为我人生路上最为重要的老师，在学术道路上的不断探索、创新指引着我，在生活中对于所喜爱事物的热情、坚持感染着我。除了陈一帅老师，郭宇春老师、赵永祥老师等都对我有着许多的帮助，他们对于学术的专业性让我受益匪浅，对于学术的热情让我颇受感染。

也感谢我生活上的朋友杨晶晶、艾方哲、苏健等人，他们既同我一起在学术上进步，还陪我一起在生活中分享。三年来，我们互相帮助、共享忧乐，度过了我难忘的研究生生涯。

最重要的是我的父母，考研时的忧虑、找工作时的低落，他们都陪我一起面对，不断鼓励我、支持我。他们对于我再三考虑后的决定总是无条件的支持，他们是我人生中最坚实的后盾。

感谢我的老师、父母、朋友，也祝愿我们都有美好的未来。

摘要

将人工智能技术应用于在线教育平台是现代教育领域的迫切需求。目前的线上教育现状显示，优质的老师仍然是稀缺资源，这制约着在线教育的大规模普及；同时，目前的国家政策指出：“为解决社会的教育需求，应当将人工智能应用到在线学习教育平台中去”。

基于自然语言处理技术为学生自动地推荐相似的题目是人工智能在线教育领域中的核心应用，它能够帮助学生掌握知识点。设计合适的习题相似度模型，对于提升习题推荐质量、提高教学效率具有重要的实践意义和应用价值。

传统的相似习题识别模型有两点不足：(1) 仅仅能够得到分开不相似习题表征的超平面，没有着重于增加不相似习题之间的距离，同时减少相似习题之间的距离；(2) 仅仅关注问题文本能带来的信息，忽略了习题解答的作用。这些不足导致目前的模型推荐相似习题的结果不准确。因此，本文的研究目标是：建立综合全面的习题向量空间表征，设计准确的习题相似度模型，提供寻找相似习题的应用工具。

本文针对上述不足创新性地设计能够提升习题表征能力和相似习题推荐效果的习题相似度模型，基于真实在线教育系统的中文数学习题数据集进行实验，具体贡献如下。

1) 为了将习题映射到相似习题距离较近、不相似习题距离较远的习题表征空间中，利用孪生神经网络架构设计出基于 Siamese 架构和 Triplet 架构的习题相似度模型 SBERT-CLS 和 TBERT-CLS。后者的 MAP 分数为 0.61，比表现最好的基线模型 VSM 高 0.23（即相对提升了 60.5%）。

2) 为了更好地捕获习题问题和解答文本之间的关系，本文设计出不仅支持两道习题问题文本和解答文本的各自匹配，还支持一道习题的问题文本和另一道习题的解答文本之间匹配的习题相似度模型 SBERT-QA 和 TBERT-QA。后者的 MAP 分数为 0.65，比仅考虑习题问题文本的 TBERT-CLS 高出 0.04（即相对提升 6.6%）。

3) 为了进一步获取综合全面的习题文本表征，本文利用 Text-CNN 进行池化操作，设计出习题相似度模型 SBERT-CNN、TBERT-CNN、SBERT-QA-CNN 和 TBERT-QA-CNN。其中 TBERT-QA-CNN 的 MAP 分数为 0.66，比未加入 CNN 池化的 TBERT-QA 高出 0.01（即相对提升 1.5%），比表现最好的基线模型 VSM 高出 0.28（即相对提升 73.7%）。

本文对设计的习题相似度模型获得的习题表征进行可视化分析，并用实际推荐案例比较了各模型在推荐相似习题任务上的具体表现，证明了本文设计的模型在改进习题推荐上的效果，验证了模型的实践意义和应用价值。

关键词：寻找相似习题；相似习题推荐；深度学习

ABSTRACT

It’s an urgent need to apply artificial intelligence technology to online education platforms in the field of modern education. The current status of online education shows that high-quality teachers are still scarce resources, which limits the large-scale popularization of online education; Meanwhile, the current national policy states: ‘In order to solve the educational need of society, artificial intelligence should be applied to online learning educational platform.’

Automatically recommending similar exercises to students based on natural language processing technology is a core application in the field of artificial intelligence online education, which can help students master knowledge points. Designing a proper exercise similarity model has important practical significance and application value for improving the quality of exercise recommendation and teaching efficiency.

The traditional similar exercise recognition model has two shortcomings: (1) They only obtain the hyperplane that separates dissimilar exercises, but do not focus on increasing distances between dissimilar exercises while reducing distances between similar exercises; (2) They only focus on the information that the text of exercise’s question can bring, but ignore the role of the exercise’s answer text. These shortcomings lead to inaccurate results of recommending similar exercises by the current models. Therefore, the research goal of this article is to establish a comprehensive vector space representation of exercises, design an accurate exercise similarity model, and provide application tools for finding similar exercises.

This paper innovatively designs exercise similarity models that can improve the representation ability of the exercise and the recommendation effect of similar exercises for the above shortcomings. The experiments are based on the data set of Chinese mathematical exercises in a real online education system. The contributions are as follows.

1) In order to map the exercises to the exercise representation space where the distance between similar exercises is close and the distance between dissimilar exercises is far away, the exercise similarity models SBERT-CLS and TBERT-CLS based on Siamese architecture and Triplet architecture are designed using the Siamese network architecture. The latter has a MAP score of 0.61, which is 0.23 higher than the best-performing baseline model VSM (i.e., a relative increase of 60.5%).

2) In order to better capturing the relationship between the exercise question text and answer text, the exercise similarity models SBERT-QA and TBERT-QA are designed to

support not only the matching of the two question texts and the answer text, but also the question text of one exercise and the answer text of the other exercise. The latter has a MAP score of 0.61, which is 0.04 higher than the TBERT-CLS model that only considers the question text similarity matching (i.e., a relative increase of 6.6%), which proves the importance of the comprehensive consideration of the question text and the answer text for obtaining effective exercise representation.

3) In order to further obtain a comprehensive text representation of the exercises, we use Text-CNN to perform pooling operation and design the exercise similarity models SBERT-CNN, TBERT-CNN, SBERT-QA-CNN and TBERT-QA-CNN. Among them, the MAP score of the TBERT-QA-CNN model is 0.66, which is 0.01 higher than the TBERT-QA model without CNN pooling (i.e., is an increase of 1.5%), and 0.28 higher than the best performing baseline model VSM (i.e., a relative increase of 73.7%).

This article visually analyzes the exercise representations obtained by the designed exercise similarity models, and compares the specific performance of each model in recommending similar exercises with actual recommendation cases, proving the role of the model we designed in improving the recommendation effect of exercises and verifying the practical significance and application value of the model.

KEYWORDS：Finding Similar Exercises; Similar Exercises Recommendation；Deep Learning

目录

1 引言

1.1 研究背景及意义

教育是国家的立足之本。目前，虽然在线教育已取得巨大进步，但在现有的教学系统中，优质的教师资源仍然是在线教育平台的核心，老师的同步跟进仍然是目前最有效的教学方式。然而，优质的老师仍然是稀缺资源，这制约着在线教育的大规模普及。

人工智能是解决上述瓶颈的重要的方法，它在在线教育中的应用已成为我国发展规划战略中最重要的一环。2017 年国务院发布的《新一代人工智能发展规划》[1]指出:“人工智能在教育等应用领域能够发挥积极作用，为解决社会的教育需求，应当将人工智能应用到在线学习教育平台中去”。

目前，虽然在线教育已取得巨大进步，但在现有的教学系统中，优质的教师资源仍然是在线教育平台的核心，老师的同步跟进仍然是目前最有效的教学方式。人工智能发展仍不成熟，无法让教育变得高效、教育水平得到提高。

相似题智能识别和判断是人工智能在线教育领域中的重要问题。从老师的角度，在课程结束时，可以通过习题检索为学生布置习题；从学生的角度，当一个学生做错某类题目或花费的时间过多时，应当推荐类似的习题让学生练习以掌握该知识点。在实际教学中，这些工作都是由教师完成的，耗时耗力，效率不高。如果能够基于人工智能技术准确及时地根据习题相似度推荐习题，不但能够减轻老师的负担，还能让学生得到及时的反馈，提升教学质量。

因此，本论文的研究目标是：建立综合全面的习题向量空间表征，设计准确的习题相似度模型，提供寻找相似习题的应用工具。这为教师的教学、学生的学习带来便利，可以节省老师的精力，将优质教师资源用于更有意义的地方，提升教学效率和学生学习质量。

1.2 国内外研究现状

本节将介绍文本表征和习题文本的相关研究。目前，关于习题深度表征和寻找相似习题任务相关的研究不算太多，但自然语言处理任务相关的研究能够为本文研究带来一定启发。在自然语言处理的研究中，深度学习是最有潜力的一个研究方向。由于深度学习具有强大的表征学习能力，在许多自然语言处理任务中，它们将

文本进行向量表征并进行下游任务，这能够捕捉更多文本隐藏特征，有利于下游任务的进行。这些研究具有对本论文研究的启发作用，对本论文的研究方向有一定的指导意义。

1.2.1 文本表征相关研究

在文本表征方面的研究中，有如下几个方面的研究。

在传统的文本任务研究中，向量空间模型 VSM 通常被用来进行文本的向量表征，具体的，它通过计算文本中每个单词的 TF-IDF 信息进行向量表征，这有利于迅速进行相似性检索[2]。VSM 的这种表征与词袋模型类似，由于它无法捕获实体、单词之间关系，OVSM-TQSM[3]对其进行了改进。OVSM-TQSM 将领域专用术语和 VSM 结合起来以便更精确地进行文本对相似性的计算。

随着自然语言技术的发展，人们逐渐考虑将人类语言习惯引入自然语言模型的设计中，也即语言模型，它也成为了自然语言处理模型的奠基石，启发了早期的相关工作。语言模型通过顺序预测单词将词序信息、单词间关系考虑进来，而不再像 VSM 及其衍生模型一样简单的用 TF-IDF 等人工特征进行建模。此后，利用神经网络进行语言建模的神经概率语言模型 NNLM[4]诞生了，并启发了 Word2vec[5]的相关工作。Word2vec 的出现表示文本表征不再依赖于人工设计的特征（如 TF-IDF），而是在大量语料的训练过程中自发地进行损失计算和参数更新，不断收敛成为向量空间中的某一向量，而向量之间的距离也反映出语义距离。单词向量的出现激励了句子表征，例如用一个句子中所有单词向量的加和将该句子表征成一个向量，但这丢失了大量的单词顺序信息。Doc2vec[6]提出句子向量与单词向量共同训练，利用神经网络自发更新句子向量的特点保留了句子信息，使文本表征质量得到进一步提升。

这些文本表征模型虽然使文本表征能力得到很大的提升，但随着自然语言技术的广泛使用，人们发现简单的通过大量语料更新单词向量无法得到最佳的文本表征，近些年的工作则为得到适当的文本表征做出了大量创新。具体来说，在早期的模型训练时，当遇到一词多义的情况时，每个与该单词相关的训练语料只能改变同一个单词的向量表征，这样的训练方法无法分辨出不同语境下单词的不同语义。为了解决不同语境不同语义的问题，ELMO[7]用双层 LSTM 提取特征，在送入单词向量后，它分别利用下层和上层 LSTM 提取简单的句法特征和高阶的语义特征，并将提取所得的特征进行拼接，在训练双向语言模型后得到相应的文本表征。同时，为了捕捉潜在的单词间关系，已被证明特征提取能力和并行能力远高于 LSTM 的 Transformer[8]也被 GPT[9]所利用，它在预训练时利用 Transformer 捕获文本中每个

单词对之间的关系，将得到的特征进行单向语言模型的训练以得到文本表征。为了减小预训练、微调两阶段间的误差传递，GPT 还在微调阶段根据下游任务的不同对模型结构进行改造，提出了预训练和微调一体化的模型。2018 年，由谷歌提出的 BERT[10]进一步进行改进，它充分利用 Transformer 的优势，将其应用于训练双向语言模型中，并受 GPT 的启发根据不同任务对模型进行改造，刷新了 11 项自然语言处理任务的记录。利用神经网络构建的文本表征模型能够识别模糊的模式，并在一系列自然语言处理任务中保持灵活性。尽管神经网络的复杂性相对于简单方法而言较高，但这些神经网络能够被训练学习足够复杂的生成模型[11]。

孪生神经网络架构的目标是将文本映射到一个语义距离与空间距离一一对应的向量空间中去，它可以将文本映射到相似文本距离较近、不相似文本距离较远的表征空间中。孪生神经网络架构被广泛应用于图像任务中[12-14]，它可以使训练后的空间距离与“语义距离”对应起来。这为自然语言处理任务提供了很好的思路，目前已有研究利用基于 Siamese 网络的 LSTM 应用于句子对相似度识别中[15]。在习题任务的研究中，Liu Qi 等人将 LSTM 应用在孪生神经网络框架上，利用 LSTM 对文本进行表征和相似度计算，将相似题从候选集中挑选出来，实现了在在线教育平台寻找相似数学题的任务，在寻找相似数学习题任务中取得了不错的效果[16]。进一步的，Yu Yin 等人提出了层次预训练网络，同时用语言模型学习低阶的语义而用领域模型学习高阶的逻辑、知识点，他们在双向 LSTM 的基础上加入了自注意力机制，该工作进一步提升了习题的向量表征效果和模型的习题理解能力[17]。

1.2.2 习题相关研究

目前，对习题的相关研究不仅包括习题文本相似的任务，还包括习题文本分类、编程题相似任务等。这些研究通过挖掘与习题相关的特征对具体方案进行设计，对本文的研究有非常重要的指导意义。

在习题任务中，早期研究往往将人工提取的特征与机器学习算法结合起来。例如，人工定义多种特征将数学习题中的每个单词表征成多维向量，并利用 SVM（支持向量机）对单词进行分类[18]。在编程习题任务中也有研究将编程语言中的关键词、嵌套层数等结构特征转换成特征向量，并据此计算编程习题之间的相似度[19]。

传统研究依赖于人工特征的制定，不具备灵活性。随着文本表征在越来越多的自然语言处理任务中展现出的普适性和强大的效果，越来越多的习题研究在文本表征的基础上进行改进。例如，为了实现相似习题检索的任务，借鉴 word2vec，将数学概念利用 skip-gram 方法进行向量表征，并用数学题的相关概念向量进行加权得到数学题的表征，以完成自适应教学系统中的相似题检索任务[20]。由于数学题

中包含有图像、文本、概念等多类信息，也有研究在习题文本表征的基础上加入多项注意力机制（如图像-文本注意力）以计算数学题相似度，提高了寻找相似习题的准确度[16]。

除了习题信息，学生表现数据也可以用于寻找习题之间的关系[21]。研究表明习题知识点、学生模型预测所得的正确概率等特征是计算习题相似度的重要特征，而学生学过的课程、学生回答顺序等特征是预测习题难度的重要特征[22]。在寻找相似习题任务中，还有研究将学生回答的正确率、回答的时间等利用统计的方法计算相似分数[23]。虽然将学生表现数据应用于习题任务是一个大胆的创新，但学生在习题上有相同的表现并不能证明习题的相似[24]，将其应用于习题任务上是有局限的。

尽管习题任务的研究在特征挖掘、文本表征等方面均有所创新，大部分习题任务的研究都忽略了习题解答能够起到的作用，而仅仅关注其问题文本所能带来的信息。在问答相似检索任务中已经发现，不论问题文本是否相似，只要回答文本相似，它们就是相似的[25]。因此，对具有问题和解答的习题文本来说，自然语言处理任务不仅可以通过对习题问题进行文本表征计算相似度，还可以将习题解答文本融入模型中进行文本表征。

1.3 研究内容

寻找相似习题是在线学习教育平台中的一个基本问题。相似习题是指考察学生相同知识点的习题[26]。表 1-1 展示了四道数学习题的问题和解答（在实际的数据集中，数学公式用 latex 公式表示），其中习题 2、3 是习题 1 的相似题，它们都考察了“抢占制胜点”这一数学知识点，而习题 4 考察的是“多位数除法的实际应用”，不是习题 1 的相似题。通过观察可以总结数学习题的特点，也即寻找相似习题任务的两大难点：(1) 数学习题通常以一个故事作为背景来考察数学知识点，由于其文本的复杂性，现有的技术不太容易从故事中确定其知识点；(2) 数学习题通常比较简短，这使得习题中的信息高度密集，这也增加了准确理解问题的困难。

数学习题的特点决定了寻找相似数学习题任务的难度进一步提升，传统的技术已不足以应对。具体来说，传统的技术有两点不足：1) 仅仅能够得到分开不相似习题表征的超平面，没有着重于增加不相似习题之间的距离，同时减少相似习题之间的距离，进而得到不相似习题距离较远、相似习题较近的空间；2) 仅仅关注问题文本所能带来的信息，忽略了习题解答的作用。

为了解决这类问题，本文在已有的深度学习及自然语言处理技术的基础之上进行改进，对习题进行表征并在相似习题任务中取得提升。研究内容主要分为两部分：① 利用孪生神经网络架构，将习题映射到空间距离反映“语义”距离的目标

空间中，提升习题表征能力；② 充分利用习题解答信息，利用文本表征模型捕获习题问题和解答之间的关系提升习题表征能力和下游任务准确度。

表 1-1 习题样例

Table 1-1 Example Exercises

习题 1	问题	两个人从 1 开始按自然数顺序轮流依此报数，每人每次只能报 1~3 个数，不允许不报，谁先报到 25 获胜。你选择先报还是后报？怎样报才能获胜？
	解答	25 / (1 + 3) = 6（组）……1（个），要获胜必须先报，先报 1 个数，然后跟另外一个人凑 4 个数就必胜。
习题 2	问题	艾迪和薇儿两个人轮流在一个凸十六边形中画对角线。规定新画的对角线不能与已经有的相交，画最后一条线者获胜。如果艾迪先画，则谁有必胜的策略？
	解答	艾迪连十六边形相对的两个顶点，将十六边形分成两个九边形。之后不管薇儿怎么连线，艾迪都连与之成轴对称的线段即可。
习题 3	问题	桌子上放着 37 根火柴，聪明昊、神奇涛二人轮流每次取走 1~5 根．规定谁取走最后一根火柴谁获胜．如果双方都采用最佳方法，聪明昊先取，神奇涛后取，你知道会胜吗．
	解答	由 37 /（ 1 + 5 ）= 6 ……1 知聪明昊会胜．
习题 4	问题	小红在计算出发时，把出发 65 写成了 56，结果得到的商是 13 余 52。想一想，正确的商应该是多少？
	解答	先根据错误的除数、商和余数求出正确的被除数：56 * 13 + 52 = 780，再求正确的商：780 / 65 = 12。

1.4 本文的主要贡献

本论文的主要贡献如下：

(1) 为了将习题映射到相似习题距离较近、不相似习题距离较远的习题表征空间中，利用孪生神经网络架构能够将输入映射到用空间距离反映“语义”距离的目标空间中的原理，设计出基于 Siamese 架构和 Triplet 架构的习题相似度模型。本论文所用的中文数学数据集上的实验结果说明：Siamese 架构能对模型效果进行提升，且基于 Triplet 架构的模型效果比基于 Simaese 架构的模型好。例如，在一个习题含有 5 道相似题和 200 道不相似题的候选集中，基于 Siamese 架构的 SBERT-CLS 模型的 MAP 分数为 0.35，比未用 Siamese 架构改进的 BERT 高出 0.33；而基于 Triplet 架构的 TBERT-CLS 模型 MAP 分数为 0.61，比 SBERT-CLS 模型高出 0.26。TBERT-CLS 模型比表现最好的基线模型 VSM 高出 0.23（即相对提升了

60.5%）。

(2) 为了更好地捕获习题问题和解答之间的关系，本论文设计出融合利用习题问题和解答的文本表征模型。具体来说，该模型不仅能够支持两道习题的问题文本-问题文本的匹配和解答文本-解答文本的匹配，还支持一道习题的问题文本和另一道习题的解答文本之间的相似度匹配。实验结果证实了充分利用习题的问题文本和解答文本带来了很大提升。例如，在一个习题含有 5 道相似题和 200 道不相似题的候选集中，TBERT-QA 模型的 MAP 分数为 0.65，比仅考虑习题问题文本相似度匹配的 TBERT-CLS 模型高出 0.04（即相对提升了 6.6%），比表现最好的基线模型 VSM 高出 0.27（即相对提升了 71.1%）。

(3) 本论文还利用 Text-CNN 进行池化操作，进一步获取综合全面的习题文本表征。实验结果说明 CNN 的池化方式能够提升一定的模型效果。例如，在一个习题含有 5 道相似题和 200 道不相似题的候选集中，TBERT-QA-CNN 模型的 MAP 分数为 0.66，比未加入 CNN 池化的 TBERT-QA 模型高出 0.01（即相对提升了 1.5%），比表现最好的基线模型 VSM 的高出 0.28（即相对提升了 73.7%）。

本论文充分利用现有技术的优点和数学习题的特点，设计模型并提升模型在相似习题任务中的效果，进一步加强了人工智能在教育领域的引领作用。

1.5 本文的组织结构

本文的组织结构如下：

第二章介绍本论文相关的技术背景，包括涉及到的各类深度学习模型、自然语言处理模型、孪生神经网络架构、模型评估指标和所采用的工具等。

第三章对本文所使用的中文习题数据集进行统计分析，并在对寻找相似习题任务进行定义后对现有算法进行评估和分析。

第四章介绍本文完成的基于孪生神经网络架构的模型设计及模型构建过程，并从模型训练结果、寻找相似习题任务中的表现、习题表征可视化三个方面对模型进行结果分析。

第五章介绍本文完成的基于习题问题与解答的融合模型设计及模型构建过程，并从模型训练结果、寻找相似习题任务中的表现、习题表征可视化三个方面对模型进行结果分析。

第六章对所有模型结果进行分析，用实际推荐案例比较了各模型在推荐相似习题任务上的具体表现，证明了本文设计的模型在改进习题推荐效果上的作用。

第七章对本文进行总结，并提出了对未来的展望。

2　技术背景

本章将介绍本论文相关的技术背景，包括相关的技术、评估指标和工具。具体的，首先介绍深度学习中常用于自然语言处理任务的几种特征处理器，接着介绍相关的自然语言处理技术概念和模型，并介绍本文相关的模型评估指标，最后介绍模型设计和实验使用的相关工具和平台。

2.1　深度学习

深度学习首次于 1986 年被提出，它是从机器学习领域延伸的，其强大的表征学习能力、自动进行特征挖掘的能力能够得到不错的效果，因此被广泛应用于计算机视觉、自然语言处理等领域，在近十年来迅速发展[11]。

在自然语言处理任务中，越来越多的研究利用深度学习对文本进行表征学习，并将训练结果应用于下游任务中。传统的自然语言处理任务经常用独热编码等方式对文字进行表征，这种高维表征方式在语料库中文字种类众多的情况下极其容易造成维度灾难等问题。由于深度学习对原始数据进行了多重线性、非线性变化等抽象提取，它的表征学习能力远强于传统表征方式，可以大大改善维度灾难的问题。在近年来的自然语言处理研究中，利用深度学习进行分布式表征比比利用传统的独热、词袋表示更有广泛适用性。同时，利用深度学习对维基百科等大量文本进行词向量的预训练，也已被验证能够在自然语言处理任务中加快训练收敛并提高词向量的普适性[5]。

数学习题是一种具有复杂性和简短性的文本，因此寻找相似习题是一种自然语言处理任务。由于深度学习强大的表征学习能力，习题文本中的语义特征等浅层特征和数学逻辑等深层特征能够在深度学习模型的训练中被学习，这使深度学习对数学习题进行适当的表征，并在习题任务中取得效果提升成为可能。

提取习题文本中的浅层、深层特征是习题任务的重点。深度学习中一些神经网络的结构特性决定了可以将语言中的隐藏特征进行抽象提取，因此经常被用作自然语言处理任务中的特征处理器。本节将介绍可以作为自然语言处理任务中可以作为特征处理器的循环神经网络 RNN、长短期记忆网络 LSTM、卷积神经网络 CNN 和 Transformer。

2.1.1 循环神经网络 RNN

循环神经网络是一种将输出状态反馈到自身网络的神经网络[27]，被 Jeffery.L.Elman 提出用于解决时间序列问题[28]。与传统的显示反馈机制，即将输出反馈到输入的结构相比，它将输出前一层的隐藏状态反馈到输入并再次进行下一层的循环，这一改动可以存储每个时刻的状态，保留之前时刻的信息，这样就能够更有效地处理时间序列问题。

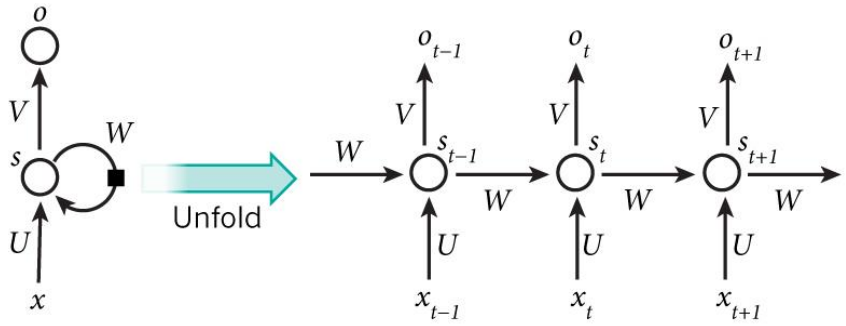


图 2-1 循环神经网络结构

Figure 2-1 Structure of Recurrent Neural Network

循环神经网络的结构如图 2-1 所示。在每个时刻都有一个相应的输入，如 t 时刻的输入是 x_t。由公式（2-1）和（2-2）可以得到 t 时刻的隐藏状态 S_t 和输出状态 O_t：

$$S_t = f\left(Ux_t + WS_{t-1}\right) \tag{2-1}$$

$$O_t = softmax\left(VS_t\right) \tag{2-2}$$

这一隐藏状态可以继续传到下一神经元当中进行运算，并得到每个时刻的隐藏状态和输出，在最后一个时刻可以得到最终的隐藏状态和相应输出。循环神经网络中每个时刻的隐藏状态不仅包含了当前时刻的输入，还存储着之前时刻的“记忆”。循环神经网络的特殊结构不仅能够解决时间序列的表征问题，还能够处理任意长度的时间序列[28]。

循环神经网络也可以为自然语言处理问题带来新的发展方向。本质上，自然语言就是一个时间序列，每当输入一个单词可以视之为一个时刻，并可以得到每个时刻的隐藏状态和输出。循环神经网络已经应用于许多自然语言处理研究工作中，如机器翻译、语言模型等[29]。

2.1.2 长短期记忆网络 LSTM

理论上，循环神经网络所拥有的链式结构可以让它在每个时刻获得上一时刻的状态信息，从而保留所有时刻的信息，但实际上，在输入序列过长时，循环神经网络在很容易“遗忘”距离较远的单词状态。据 Sepp Hochreiter 研究发现，循环神经网络中的误差梯度随距离的增加呈指数性减少[30]；Yoshua Bengio 等人通过理论和实验指出，在误差梯度下降的准则下，很难捕捉长时间的语句依赖关系[31]。

为了解决捕捉长时间依赖关系的问题，Sepp Hochreiter 等人提出了循环神经网络的改进结构——长短期记忆网络[30]。它通过引入“门”的结构在信息传递过程中判断历史信息保留程度，这样的机制能够优先存储更为重要的信息，同时减少不重要信息的存储，能够有效提高循环神经网络的存储效率。如图 2-2 所示即为长短期记忆网络的结构，每时刻所对应的单元细胞（A）中包含了输入门、输出门和遗忘门三个组成单元。

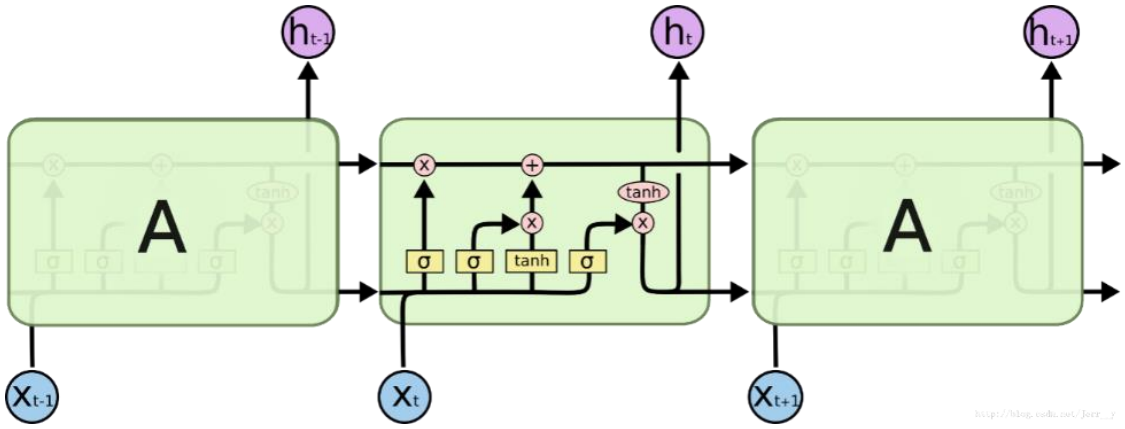


图 2-2 长短期记忆网络结构

Figure 2-2 Structure of Long short-term memory

在长短期记忆网络中，最重要的部分就是遗忘门，它在三类门中对结果的影响最高[32]。在每个时刻都有一个相应的输入，如 t 时刻的输入是 x_t，相应隐藏状态是 h_t。由公式（2-3）计算上一细胞状态 C^{t-1} 能够被存储的比重 f_t：

$$f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right) \tag{2-3}$$

输入门是仅次于遗忘门重要性的门结构，它主要用于计算该时刻细胞状态 C^t 能够被存储的比重 i_t，由公式（2-4）、（2-5）、（2-6）更新 t 时刻细胞状态 C^t：

$$i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right) \tag{2-4}$$

$$C^t = tanh\left(W_c\left[h_{t-1}, x_t\right] + b_c\right) \tag{2-5}$$

$$C^t = f_t * C^{t-1} + i_t * C^t \tag{2-6}$$

输出门是最后一个门结构，在经过遗忘门和输入门的计算后，通过公式（2-7）

和（2-8）计算细胞状态C^t能够被存储的比重o_t和将该时刻的隐藏状态h_t：

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{2-7}$$

$$h_t = o_t * tanh\left(C^t\right) \tag{2-8}$$

在上述计算公式中，W_*、b_*（*代表f、c、o等字符）分别为相应网络结构中的参数，能够在训练中通过反向传播算法进行更新，σ为 sigmoid 激活函数，*tanh* 为 tanh 激活函数。

由于 tanh 函数的函数特性，其导数要么接近 0，要么接近 1，这就使得长短期记忆网络在反向传播更新参数时，不会再出现循环神经网络中由于梯度连乘导致的梯度消失问题[30]。因此，长短期记忆网络能够着重存储更为重要的语句信息，弥补了循环神经网络随时间增长更难捕获序列中依赖关系的缺点，在关系抽取[33]、机器翻译[34]、情感分析[35]等自然语言处理任务中都有广泛的应用。

2.1.3 卷积神经网络 CNN

由于循环神经网络和长短期记忆网络的结构特性能够解决处理语言序列的大部分问题，目前大部分的研究工作都采用了这种链式结构的神经网络。但由于它们每个时刻的状态都依赖于该时刻的输入和上一时刻的状态，因此只能串行处理语言序列，这就大大降低了效率。

卷积神经网络作为一种典型的并行特征处理器，在图像处理任务中被广泛的应用，但在 2014 年才首次被提出用于自然语言处理任务中[36]。卷积神经网络的网络结构包含卷积层、池化层和全连接层，在处理语言序列时对其网络结构进行的改动如图 2-3 所示。

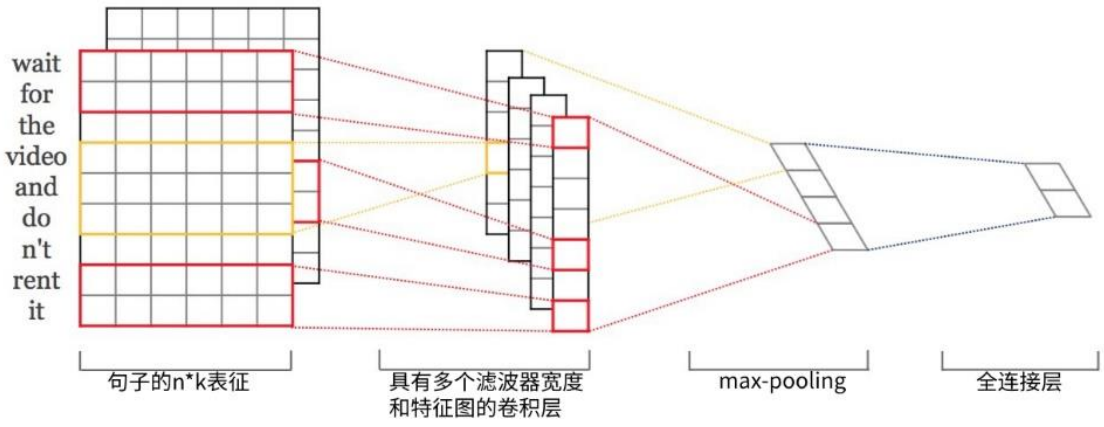


图 2-3 应用于句子分类的卷积神经网络结构[36]

Figure 2-3 Convolutional Neural Network Structure for Sentence Classification[36]

在进入卷积层之前，每个语言序列中的单词都先初始化为一个k维的词向量，当输入的句子长度为n时，将其初始化为$n*k$的词向量矩阵。

卷积层主要用于提取语言序列的深层特征，在将语言序列这样的一维输入经过词向量处理转化为二维输入之后，就可以进行类似图像处理的卷积特征提取。该层用卷积核进行卷积计算，假设卷积核窗口大小为h，卷积核大小就为$h*k$维；第i个卷积核窗口里的单词，即从第i个位置到第$i+h-1$个位置的单词向量$x_{i:i+h-1}$可以通过公式（2-9）得到特征c_i，其中f为激活函数（如 relu 函数）：

$$c_i = f\left(w*x_{i:i+h-1}+b\right) \tag{2-9}$$

通过移动窗口扫描整个句子可以得到$n-h+$个特征，得到特征图：$c=\left[c_1,c_2,...,c_{n-h+1}\right]$。利用$m$个卷积核进行扫描，都可以得到$m$个对应的特征图。

池化层主要用于防止过拟合，比较典型的有 max-pooling 和 mean-pooling。在图 2-3 中进行 max-pooling 操作，经过$c'=\max\{c\}$就可以简单得到该卷积核的特征。得到特征之后再送入全连接层就可以进行简单的文本分类。

比起循环神经网络和长短期记忆网络，卷积神经网络有以下优点：1）由于卷积操作可以并行进行，它的并行能力更高；2）由于滑动窗口的结构特征，它能够捕捉更多的上下文信息；3）由于滑动窗口顺序滑动的特点，它能够记录文本中相对位置的信息。但由于滑动窗口的大小直接决定了捕捉单词依赖关系的距离，超出窗口之外的单词间依赖关系并不能被捕捉。为突破卷积神经网络无法捕捉长距离依赖特征的局限性，相关研究从两个方面对卷积神经网络进行了改进：一是不改变卷积层深度，将连续覆盖的滑动窗口变为不连续覆盖，即空洞卷积[37]；二是加深卷积层深度，使得上下层覆盖窗口叠加[38]。目前，卷积神经网络除了在文本分类上，也在关系抽取[39]、句子重复性识别[40]等自然语言处理任务中有着颇为广泛的应用。

2.1.4 Transformer 模型

在循环神经网络和卷积神经网络提出之后，各类改进方法层出不穷，其中最为普遍的就是利用注意力机制加强对文本中较为重要信息的“注意”。在自然语言处理中，比较典型的是在 seq2seq 这种编码-解码模型上加入注意力机制以进行机器翻译[41]。由于 seq2seq 本质是由循环神经网络组成的，它也具有循环神经网络的局限性，即只能串行处理，效率较低。

2017 年谷歌首次提出 Transformer 结构以进行机器翻译任务[8]。这是一种完全利用注意力机制，与循环神经网络和卷积神经网络都无关的一种新的编码-解码模型，其结构如图 2-4 所示。

注意力机制可以简单转化为键（key）、值（value）和查找（query）的计算。

简单来说，键值为存储的上下文关系，键值对一一对应，当键等于查找时，可以找出相应的值，即注意力分数。一般在自然语言处理任务中，键向量K等于值向量V，查询向量Q与每个句子中的词向量相对应。在 Transformer 中所采用的是自注意力机制，Q和K均为d_k维，V为d_v维。通过公式（2-10）计算得到该句子中每对词之间的注意力分数：

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2-10}$$

Transformer 的另一大创新点是多头注意力机制，即将Q、K、V做h次不同的线性投影，并行分别在每个线性空间上对相应的Q、K、V进行点乘再归一化，最后线性映射输出，输出为$h*d_v$维。需要注意的是，在 Transformer 中应用了残差网络的思想，这样即使网络很深也不会有梯度消失的问题。

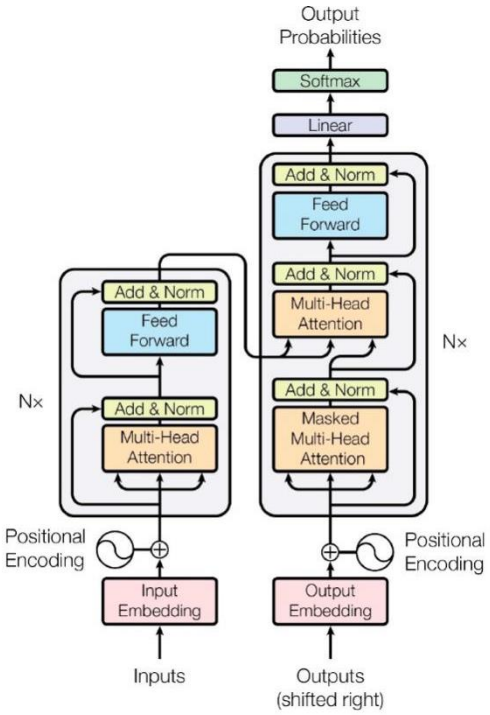


图 2-4 Transformer 结构[8]

Figure 2-4 Structure of Transformer[8]

在编码层–解码层的注意力层中，Q为前一编码层的输出，K和V为该层编码层的输出，这样通过注意力机制能够挖掘解码层的每个位置与编码层的所有位置的联系；在编码层的注意力层中，Q、K、V相等，都为前一编码层的输出，这样通过注意力机制能够挖掘编码层的每个位置与前一编码层所有位置的联系；在

解码的注意力层中，通过注意力机制能够挖掘解码层的每个位置与该位置之前的所有位置的联系[8]。

这种注意力机制虽然并行效率高，但正是因为并行而无法获得句子中每个单词的位置信息，因此在将输入送入编码和解码层前利用正余弦函数进行了位置编码，以获取单词之间的相对位置，如公式（2-11）和（2-12）所示：

$$PE_{(pos,2i)} = sin\left(pos / 10000^{2i/d_{model}}\right) \tag{2-11}$$

$$PE_{(pos,2i+1)} = cos\left(pos / 10000^{2i/d_{model}}\right) \tag{2-12}$$

在并行效率方面，Transformer 和卷积神经网络都比循环神经网络高许多；在提取特征效果方面，Transformer 明显优于循环神经网络和卷积神经网络。目前，对 Transformer 的改进主要有使其不仅在机器翻译任务中，还在问答、语言建模等任务上有更广泛的通用能力[42]，以及延长能够捕获单词间关系的长度[43]。最新的语言模型 BERT[10]也利用 Transformer 进行特征提取，在多项任务中刷新记录。

2.2　自然语言处理技术

自然语言处理是人工智能应用的一个重要分支，包括语音识别、文本分类、机器翻译等多项任务。与图像处理这类感知型任务不同，自然语言处理属于认知类任务，需要进行深层的逻辑分析让机器“理解”自然语言的含义才能完成。

为了让机器更好地“理解”自然语言，需要采取相应的自然语言处理技术和模型。目前，利用大规模的语料进行预训练，并在下游任务中通过自然语言处理模型进行语义分析已经成为主流趋势。最初，许多研究倾向于训练基于词频等人工特征的模型，随着深度学习技术的不断发展，基于自动挖掘语言特征的深度学习模型也越来越多被用来设计自然语言处理模型。

本节将会介绍 TF-IDF、向量空间模型、语言模型、Word2vec 和 BERT。这些自然语言处理模型各有特点，为本文的研究有一定指导意义。

2.2.1　TF-IDF 和向量空间模型

TD-IDF 又名词频-逆文本频率，是一种人工定义的文本特征，主要用来计算单词重要性。它分为两个部分：词频 TF 和逆文本频率 IDF。

词频 TF 即为单词在该文档中出现的频率，可以统计这个单词在该文档中的重要程度。公式（2-13）为 TF 的计算规则，其中 f_d 为单词在该文档中出现的次数，$|D|$ 为该文档的总单词数：

$$TF = f_d / |D| \tag{2-13}$$

逆文本频率 IDF 的计算规则如公式（2-14）所示，可以统计这个单词在所有文档中的重要程度，其中 N_D 为文档总数，N_v 为包含该单词的文档数。如果一个单词在许多文档中都会出现，且频率较高，那它在该文档中的重要程度就较低，而如果这个单词在所有文档中较少出现，且频率较低，那么在出现的文档中的重要程度就较高。

$$IDF = log\left(N_D / \left(N_v + 1\right)\right) \tag{2-14}$$

逆文本频率 IDF 的计算公式利用 log 函数限制了 IDF 的线性变化，并进行了平滑处理。

TF-IDF 值通过公式（2-15）得到，它的值越大说明单词对相应文档的重要程度越高：

$$\text{TF-IDF} = TF * IDF \tag{2-15}$$

利用向量空间模型[44]可以简单的将文本表达成相应的向量，假如所有语料库中的单词数是 10 个，该向量就为 10 维，每一维的值是该文本中相应单词的 TF-IDF 值。具体的，当文本 d_1 可以表示为 $d_1 = \left(w_1, w_2, ...\right)$ 时，向量空间模型可以将其转化为向量 $V_{d_1} = \left(t_1, t_2, ...\right)$，其中 w_i 表示不同的单词，t_i 表示相应单词 w_i 的 TF-IDF 值。当得到文本 d_1 和文本 d_2 的向量时，用余弦相似度可以计算文本相似度，如公式（2-16）所示：

$$\text{sim}\left(d_1, d_2\right) = \frac{V_{d_1} \bullet V_{d_2}}{\left\|V_{d_1}\right\| * \left\|V_{d_2}\right\|} \tag{2-16}$$

由于向量空间模型的训练简单、效率高效，在许多自然语言处理任务中都有应用，如信息检索、文本分类等。但由于它只考虑了单词的频率，而未考虑到单词间的关系[45]，在如今复杂的自然语言处理任务中有待改进。

2.2.2 语言模型

让机器捕获自然语言的规律、“听懂”自然语言的逻辑，进而完成机器翻译、文本分类等认知型任务的基础，是建造一个符合人类语言习惯的数学模型。语言模型在自然语言技术的发展过程中应运而生，它能够计算一句话符合人类语言习惯的概率。

统计语言模型可以根据语料，利用统计模型拟合单词的概率分布函数，在判断一个句子是人类语言的概率时通过单词概率和词组概率进行计算。例如，‘I love NLP’这句话的概率远远大于‘I NLP love’，公式（2-17）和（2-18）分别为这两句话

符合人类语言习惯概率的计算过程：

(2-17)

(2-18)

统计语言模型虽然为自然语言处理任务带来良好的开端，但却不可避免的有两个弊端：1）自由参数过多，当语料库中有$|V|$个单词，待判断是否符合人类语言习惯的句子长度为L时，由于需要计算条件概率，模型的自由参数可达$|V|^L$；2）数据稀疏，句子长度越长，所能构造的词组越多，但实际的训练语料中词组数量有限，不可能包括所有的组合，这样在计算过程中许多词组相应的条件概率都会为 0，最终计算的概率也会为 0。

为了解决这两个问题，可以采用以下措施：1）为了解决自由参数过多的问题，在计算条件概率时利用马尔可夫假设，即一个单词出现的概率仅和前面 n 个单词有关，这种语言模型名为 n-gram 语言模型。例如，当 n=1 时，，此时模型的自由参数仅有$|V|$个，不再随着句子长度而指数级增长。2）为了避免许多词组相应的条件概率为 0 的问题，利用拉普拉斯平滑或其他平滑方法，即在计算概率时，每个单词出现次数都加 1，这样最终计算的概率不会为 0。

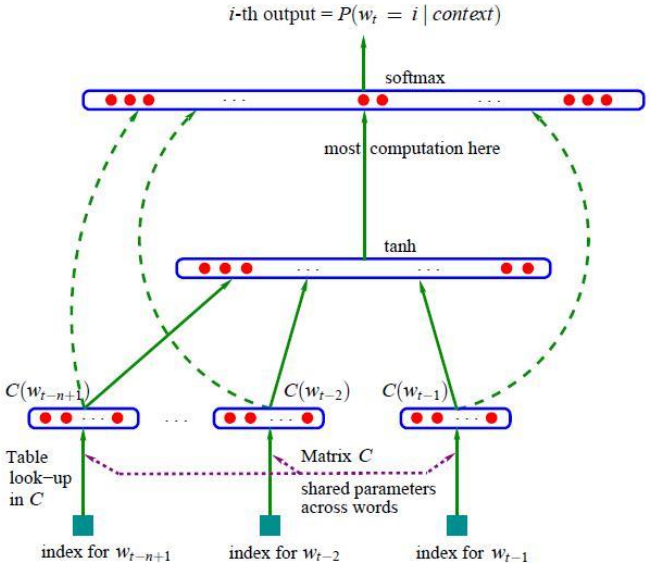


图 2-5 神经网络概率语言模型结构[4]

Figure 2-5 Structure of Nerual Probabilistic Language Model[4]

但改进后的 n-gram 语言存在考虑的上下文单词有限的问题，且没有考虑单词相似性。为解决这些，神经网络概率语言模型 NNLM[4]首次提出词特征向量的概念，结构如图 2-5 所示。它将每个单词表征成一个固定维度的向量，根据词向量计算词序列的概率，并在神经网络的训练过程中同时学习词特征向量和概率函数。神经网络概率语言模型的目标是输入上文，预测下一个单词，在训练过程中得到词特

征向量和概率函数。它包括输入层、投影层、隐藏层和输出层，结构如图所示。输入层上文单词转化成 one-hot 向量，投影层通过一个$|V|*d$（$|V|$为训练语料库单词个数，d 为词向量维度）的参数矩阵和 one-hot 向量相乘得到对应单词的词向量，隐藏层通过 tanh 激活函数将拼接后的上下文向量进行映射，输出层通过 softmax 激活函数计算每个单词是预测单词的概率，所有单词的概率相加为 1。

NNLM 的提出是语言模型领域的一个重大突破，但神经网络架构决定了其不能处理变长语言序列的特性。在后续研究中，Mikolov 等人将 NNLM 中的前馈神经网络改为 RNN，提出了 RNNLM 模型，改善了这一点[46]。

2.2.3 Word2vec 模型

尽管神经网络的引入使得神经网络语言模型比传统的统计语言模型有了很大幅度的优化，神经网络语言模型的计算量仍然很大。Mikolov 等人总结，神经网络语言模型的训练可以分为两步：1、用简单的模型训练出词向量；2、用这些单词的分布式表征训练一个 n-gram NNLM 模型[5]。而 n-gram 模型的训练耗费了大量的时间，使得神经网络语言模型的效率降低。

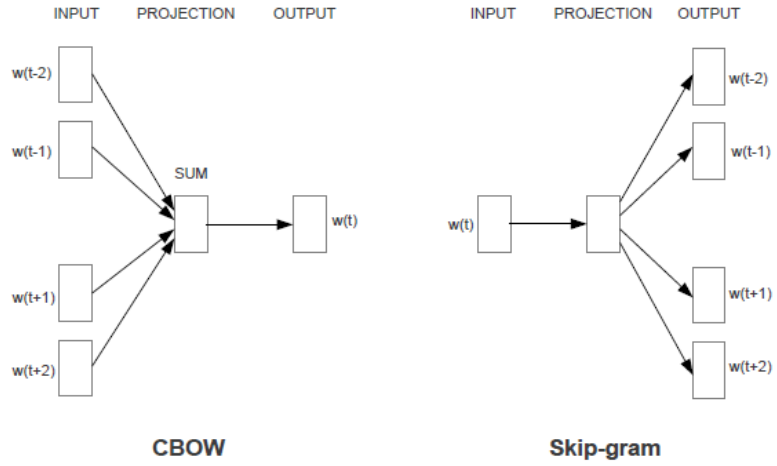


图 2-6 Word2vec 模型结构[5]

Figure 2-6 Structure of Word2vec Model[5]

Word2vec[5]首次提出用大量的数据集训练，以得到单词的连续特征向量为目标。词向量在空间中的表现可以反映它们所代表的语义，如与 Queen 在空间中的表征接近。

通过以往研究的分析，Mikolov 等人发现模型中的非线性隐藏层是计算复杂度过高的主要原因，为了简化计算复杂度，他们提出了 CBOW 和 Skip-gram 两个模型，如图 2-6 所示。

CBOW 的模型架构类似于 NNLM，目标都是通过上下文预测一个单词。但它移除了非线性隐藏层，除此之外，所有单词都可以被映射到相同的位置（它们的向量被平均了），因此单词的顺序并不影响投影。

Skip-gram 模型的目标是将当前单词作为输入送入到带有连续投影层的对数线性分类器中，并根据当前单词预测上下文一定范围内的单词。增加范围可以得到更高质量的单词向量，但也提升了计算复杂度。由于更远的单词相关性不如更近的单词，可以通过采样更少较远的单词以降低其权重。

Word2vec 所提出的训练词向量的方法不仅让自然语言处理问题得到更有效率的处理，还让传统研究跳出了原有的局限，开始关注词向量在自然语言处理问题中的关键作用。在此之后，将句子表征为分布式向量的 Doc2vec[6]、基于共现矩阵学习单词向量的 glove[47]、不再预测单词而是预测标签的 fasttext[48]相继被提出，它们不断扩展了词向量的应用，也改善了模型的效果。

2.2.4 BERT 模型

在 BERT[10]之前，已经有许多利用语言模型进行预训练的算法，如 ELMO[7]、GPT[9]等。BERT 的模型结构是是一个多层双向的 Transformer 编码层，它与 ELMO 和 GPT 的模型结构对比如图 2-7 a)所示。ELMO 利用双向的 LSTM 学习语义特征，并将两个方向的特征拼接起来进行下游任务；GPT 利用单向的 Transformer 学习语义特征，即每个单词只考虑了其在该单词位置之前的表现。只有 BERT 得到的表征在所有层中都与上下文有关。

为了得到高质量的词向量，BERT 对输入和预训练任务提出了改进。输入如图 2-7 b)所示，预训练语料经过处理，将每个句子对的起始位置用[CLS]表示，句子的末位用[SEP]表示，在预训练时将语料处理成句子对进行训练。每个位置的嵌入表征关系如公式（2-19）所示：

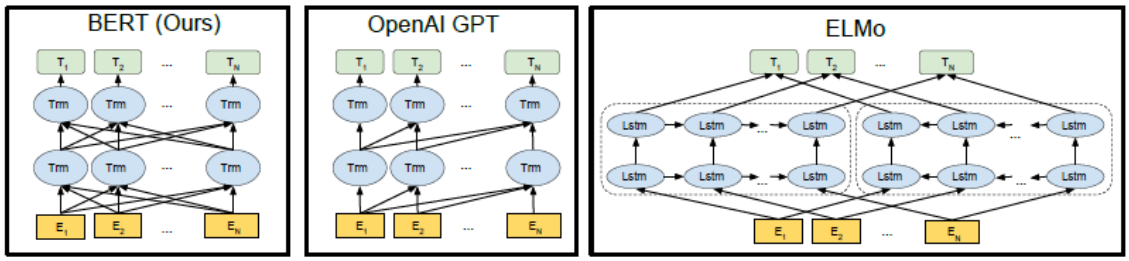
(2-19)

其中，Token Embeddings 为词向量，Segment Embeddings 为句子段向量，表示属于第一句还是第二局，Position Embeddings 与 Transformer 中的向量一样，都表示单词所处的位置。

在预训练时，BERT 不再与传统语言模型一样通过上文预测一个单词，而是首次提出了遮蔽语言模型 MLM，改善了传统语言模型只考虑单向语义信息的不足；除了 MLM，它还同时利用下个句子预测 NSP 训练句子对表征，提升在句子对任

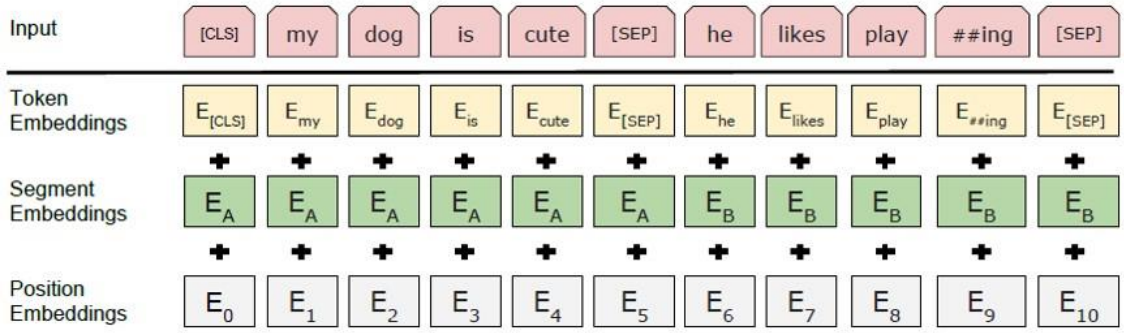
务中的表现。

传统的语言模型至多只能简单的进行单向训练，ELMO[7]将两个方向的特征拼接起来以达到提取双向特征的目的，因为直接双向训练会导致每个单词可以间接“看到”自己。为建立双向训练并克服这一困难，MLM 提出在随机遮蔽一些单词的条件下去预测这些单词，在训练时随机选取了 15%的单词进行遮蔽，将它们用[MASK]标识替换。由于在微调阶段不会存在[MASK]标识，会产生预训练和微调数据的不匹配，这些单词会进行三种变化：1）80%的概率被替换成[MASK]；2）10%的概率被随机替换成一个单词；3）10%的概率不被替换。MLM 利用注意力机制结合了上下文信息，使 BERT 成为一个真正意义上的双向语言模型。



a)　BERT、GPT 与 ELMO 结构

a)　Structures of BERT, GPT and ELMO



b)　BERT 输入

b)　The input of BERT

图 2-7 BERT 模型输入及与其他模型的结构对比[10]

Figure 2-7 Input of BERT Model and Structure Comparison with Other Models[10]

由于传统的语言模型难以捕捉两个句子之间的关系，在如机器问答、自然语言推理等下游任务中难以具有良好的表现。下个句子预测 NSP 的主要目标是预测两个句子是否为上下文关系，在训练时选取了 50%的句子对使得第二句是第一句真实的下一句，50%的句子对第二局是随机选取的。NSP 捕获上下文依赖关系，使 BERT 能够理解句子关系，在需要捕获句子关系的相关任务中的表现有了一定程度

的提升。但后续的研究也发现，在去掉 NSP 任务时，BERT 在某些任务上的表现会更好一些[49-51]。

在 BERT 之后，出现了许多相关的改进研究，如将 Transformer 的进阶版 Transformer-XL 引入到预训练任务的 XLNet[49]，减少大量参数量加快模型训练的 ALBERT[50]，增加大量训练数据并更改训练方法（去掉 NSP 任务、增加训练语料长度等）的 RoBERTa[51]等等。这些模型对 BERT 模型或是更改了训练数据，或是微调了模型的架构，尽管取得了一些效果的提升，但没有进行大幅度的修改，模型体系仍然依赖于 BERT 模型的原始设计。在本论文中，将基于 BERT 模型进行架构上的设计。

2.3 孪生神经网络架构

孪生神经网络架构 Siamese network[12]最初是为了人脸识别问题所提出的。图像处理模型通常利用大量已知类型的数据进行充分训练，但在实际情况中，经常会有类型过多而一些类型下的样本过少的问题，这会导致这些类型的样本在训练时不够充分。

Siamese 网络的主要思想是将输入映射到一个可以用距离反映“语义”距离的目标空间中。它的目标是学习两个输入 X1，X2 的相似度，结构如图 2-8 所示。

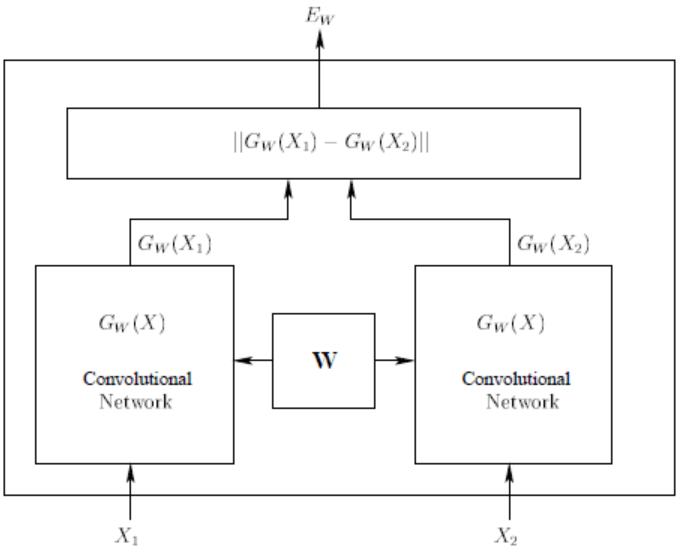


图 2-8 Siamese 网络架构[12]

Figure 2-8 Architecture of Siamese Network[12]

两个输入经过共享权重的模型 $Gw(X)$ 后被映射到目标空间中，并通过公式（2-20）得到两个输入的匹配度：

$$Ew(X1,X2)=\left\|Gw(X1)-Gw(X2)\right\| \tag{2-20}$$

通过公式（2-21）定义目标损失函数为：

$$L_w(X1,X2)=(1-y)L_g\left(Ew(X1,X2)\right)+yL_i\left(Ew(X1,X2)\right) \tag{2-21}$$

其中L_g为相似输入对（y=1）所计算的部分损失函数，L_i为不相似输入对（y=0）所计算的部分损失函数。Raia Hadsell 等人提出了一种对比损失函数[13]，在分类问题中被广泛应用，损失函数如公式（2-22）所示，其中 N 为样本数，d 为两输入的欧式距离，y 为真实标签，$margin$ 为设定的边际距离：

$$L=\frac{1}{2N}\left(\sum_{n=1}^{N}yd^2\right)+(1-y)\max\left(margin-d,0\right)^2 \tag{2-22}$$

在 Siamese 网络之后，研究进一步提出 Triplet 网络[14]，它同时将三个输入送入共享权重的模型之中，其中包括锚样本，一个正例和一个负例，这一神经网络架构可以加快模型的收敛，效果更好，但对训练样本的依赖程度较高。

一系列孪生神经网络架构的提出不仅在图像处理任务中取得了良好的效果，也在自然语言处理任务中被广泛应用，如得到句子表征[52]、句子相似度识别[15]、寻找相似习题[16]等。

2.4 模型评估

在机器学习和深度学习算法中需要对结果进行预测，为了对预测结果进行分析，通常用混淆矩阵对结果进行分析，如表 2-1 所示。当模型将正样本预测为正时，称该样本为 True Positive (TP)；当模型将负样本预测为正时，称该样本为 False Positive (FP)；当模型将正样本预测为负时，称该样本为 False Negative (FN)；当模型将负样本预测为负时，称该样本为 True Negative (TN)。

表 2-1 混淆矩阵

Table 2-1 Confusion Matrix

	预测为正	预测为负
正样本	TP	FN
负样本	FP	TN

常用的模型评估指标有 Precision、Recall、F1 和 AUC 分数。Precision、Recall、F1 分别由公式（2-23）、（2-24）、（2-25）得到：

$$Precision=\frac{TP}{TP+FP} \tag{2-23}$$

$$Recall = \frac{TP}{TP + FN} \tag{2-24}$$

$$F1 = 2\frac{Precision * Recall}{Precision + Recall} \tag{2-25}$$

AUC 分数的计算过程较为复杂。在固定一个阈值时可以通过公式（2-26）和（2-27）分别计算出真阳性率 TPR（等于 Recall）和假阳性率 FPR：

$$TPR = \frac{TP}{TP + FN} \tag{2-26}$$

$$FPR = \frac{FP}{FP + TN} \tag{2-27}$$

设定不同的阈值时，能够得到不同的 TPR 和 FPR，将 FPR 作为横坐标，TPR 作为纵坐标画出一条 ROC 曲线，曲线下方的面积即为 AUC 分数。

2.5 工具

本节主要介绍论文所使用的开发工具，包括用于搭建神经网络模型的 PyTorch 神经网络框架、集成数据分析和数据挖掘功能的机器学习工具 Scikit-Learn 和自然语言处理算法工具 Gensim。

2.5.1 PyTorch 神经网络框架

PyTorch 神经网络框架是基于 Python 语言的神经网络模型开发工具，由 Facebook 人工智能团队开发，并于 2017 年开源。2019 年深度学习论文 PyTorch 的使用率已经达到 80%。PyTorch 的多个优点使得它成为学术界最流行的神经网络框架：PyTorch 和 Tensorflow 都是基于计算图的神经网络框架，不同的是 Tensorflow 必须在运行模型前定义静态图，而 PyTorch 可以在任何时间点灵活更改其动态图；PyTorch 中的张量（Tensor）可以直接转换为 Numpy 格式，便于数据处理；PyTorch 能够利用.cuda()直接转换模型和张量，在 GPU 上进行加速，也能够直接用 cpu()在 CPU 上进行计算。除此之外，PyTorch 还具有调试方便、设计友好等优点。

PyTorch 中 torch.Tensor()是定义张量及其计算的模块，Variable()是定义变量的模块，torch.nn()模块是设计模型并实现反向传播的模块。通过调用各类模块可以构建一个完整的神经网络模型并进行训练。本论文将利用 PyTorch 神经网络框架搭建模型完成训练。

2.5.2 Scikit-Learn 工具包

Scikit-Learn 是基于 python 语言的机器学习工具，依赖于 python 工具库中的 numpy 和 scipy 库。它集成了分类、回归、聚类、降维四类机器学习算法和模型选择、数据预处理两类数据挖掘方法。Scikit-learn 涵盖的机器学习算法包含支持向量机 SVM、K-means、随机森林等模型，在调用时只需将数据处理成对应模型所需的形式就可以直接进行调用。它还涵盖了网格搜索、模型评估指标、数据负采样等功能，可以进行参数选择、模型评估、数据平衡处理等多项功能。本论文将利用 Scikit-learn 计算 AUC、F1 分数等分数进行模型训练结果的评估。

2.5.3 Gensim 工具包

Genism 是基于 python 语言的自然语言处理工具，它集成了主题模型、文档索引和相似检索等功能。Gensim 不仅为开发者提供了较为简单的接口，使得自然语言处理算法的学习更加简便，还由于其流式处理语料数据的特点，耗费内存较小而训练速度快。Gensim 主要包含语料库（Corpus）、向量（Vector）、模型（Model）三大模块：在将语料处理成文档的集合后，每个文档转化为一组特征向量，最后利用模型将文本向量变换成另一个向量空间中的向量。Gensim 中可以调用的模型很多，包括 Word2vec、TF-IDF、文档主题生成模型 LDA、向量空间模型 VSM 等，而且调用简单，训练速度快。本论文将利用 Gensim 搭建部分模型进行相似检索。

2.6 本章总结

本章主要介绍论文研究相关的技术背景，详细介绍了常用于自然语言处理任务的深度学习特征提取器 RNN、LSTM、CNN 和 Transformer，它们能够对文本的深层特征进行提取，接着介绍了自然语言处理常见的模型和孪生神经网络架构，为本文的研究提供一定的指导意义，还介绍了本文中用到的模型评估指标的计算方法，最后介绍了实验开发过程中所使用的主要工具。

3 习题数据统计分析及任务定义

本章将介绍本论文完成的习题统计分析工作，并对寻找相似习题任务进行定义。首先介绍论文的在线教育系统的数学习题数据集，接着利用统计分析、自然语言处理技术等对数据的基本情况进行数据挖掘和文本分析，最后对本论文的寻找相似习题任务进行定义，并对现有模型在该任务上的表现进行分析。

3.1 习题数据集介绍

本论文采用的习题数据集来自于在线教育平台。本论文所采用的是真实的中文数学习题集，知识点覆盖 k12 范围（学前教育至高中教育）。这些数据统一为文本数据，公式部分用 latex 文本表示，每道习题所包含的信息主要有习题 ID、习题问题文本、习题解答文本、习题所属知识点等，具体信息如表 3-1 所示。

表 3-1 习题信息介绍

Table 3-1 The Introduction of Exercise Information

列名	含义
que_id	习题 ID，数据库中习题对应的唯一 ID 号
content	习题问题文本（question），通常利用故事文本阐述数学问题。例如：“两个人从$$1$$开始按自然数顺序轮流依次报数，每人每次只能报$$1\sim 3$$个数，不允许不报$$.$$谁先报到$$25$$谁获胜．你选择先报还是后报？怎样报才能获胜？”
analysis	习题解答文本（answer），通常利用相应知识点的解题思路套用公式。例如：“$$25\div (1+3)=6$$（组）$$\cdots \cdots 1$$（个），要获胜必须先报，先报$$1$$个数，然后跟另外一个人凑$$4$$个数就必胜．”
knowledge	习题知识点，该列存储的知识点最多为 5 级，知识点为树状结构，一个子知识点只能有一个父知识点。在该数据集中存储为 json 格式，用 unicode 文本存储。

本论文首先对习题数据集进行了简单的预处理：将 knowledge 对应的 unicode 文本转化为 utf-8 文本，可以得到该习题对应至多 5 级的知识点。在本论文所用的数据集中，每个习题都至少有一个标签，但不一定每级标签都有。为了简便，本论文将习题对应的最后一级知识点作为该习题的标签（tag），如 knowledge 为“1：

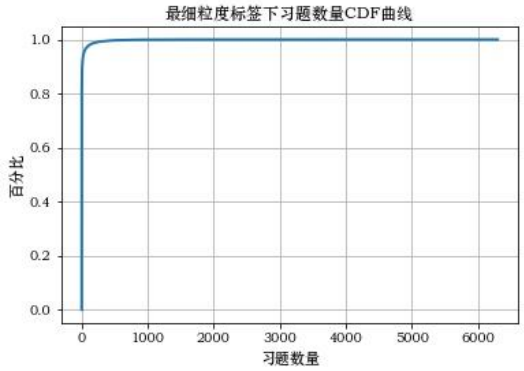
杂题，2：操作与策略，3：游戏策略，4：抢占制胜点”表示一级知识点为“杂题”，二级知识点为“操作与策略”，三级知识点为“游戏策略”，四级知识点为“抢占制胜点”，该类习题可以取“抢占制胜点”作为最细粒度标签（tag）。这可以使每个习题被分类到最细粒度上，保证本论文的模型效果。相似习题是指考察学生相同知识点的习题[26]，因此本文将标签相同的习题认为是相似习题。

3.2 习题数据统计分析

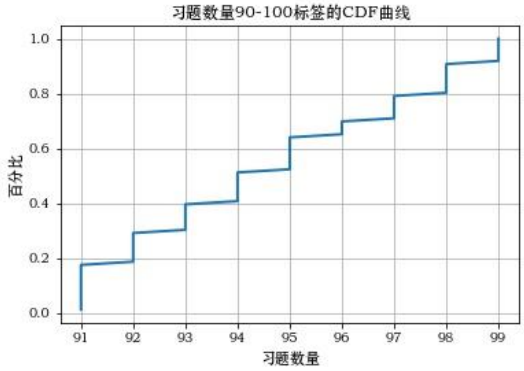
本节对习题数据集进行两个方面的数据挖掘分析：一方面，通过对习题的统计分析进行数据挖掘，另一方面，对习题的文本进行数据挖掘。根据这些分析进行习题数据集预处理和实验参数设置。

首先对习题的标签进行统计分析，进而对习题数据集进行筛选。该中文数学习题在最细粒度标签下的数量分布极度不均衡，它包含 450284 道习题，分布在 40194 个标签下，平均每个标签下约有 11 道题；其中包含习题数量最多的是“四则混合运算”，包含了 6301 道题，而包含 5 道习题及以下的标签共有 34745 个，占总标签数的 86.4%。图 3-1 a) 为对最细粒度标签下习题数量进行统计后所绘制的 CDF 曲线图，也可以看出习题标签分布的不均衡。数据不均衡对深度学习模型的训练影响较大，因此本论文拟首先用习题数量为 90 到 100 之间的标签下的习题集进行模型训练和测试，保证模型的习题数量均衡，即每类习题均进行充分训练的情况下是有效的。这些习题共计 8135 道，属于 86 个标签，其 CDF 曲线图如图 3-1 b) 所示，可以看出习题基本均匀分布在这些标签中。

接着对习题的长度进行统计分析，确定模型的文本输入长度。对习题问题文本和解答文本的长度进行统计发现，在这 450284 道习题中，问题文本长度的平均值为 131.6，中位数为 101；解答文本长度的平均值为 253.6，中位数为 138。图 3-2 为习题问题文本和解答文本长度进行统计后所绘制的 CDF 曲线图（x 轴均经过 log10 对数化），可以看出大部分文本的长度都较小，由于通常深度学习模型处理文本的长度有限，本论文将设置一个定值以截断送入的文本。为了均衡模型性能和模型效率，选取截断长度 L 为 128。



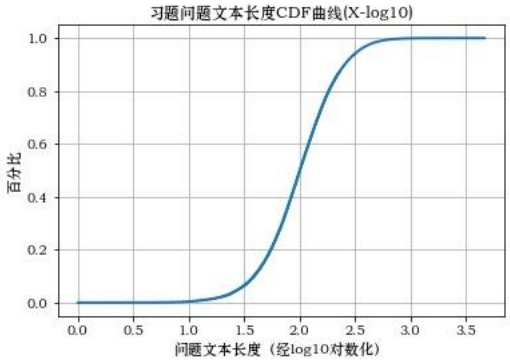
a) 最细粒度标签下习题数量 CDF 曲线
a) CDF Curve of Exercise Number under the Finest Granularity Label



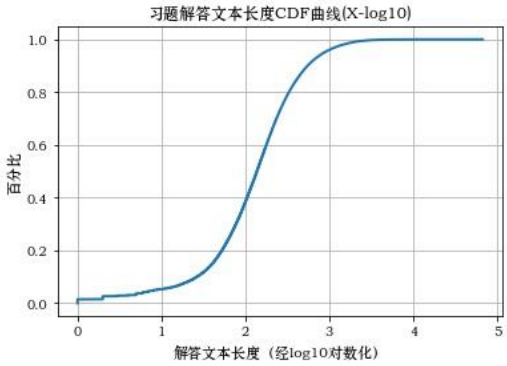
b) 习题数量 90-100 标签的 CDF 曲线
b) CDF Curve of Exercise Number under the Label that has 90-100 Exercsies

图 3-1 标签下习题数量 CDF 曲线

Figure 3-1 CDF Curve of Exercise Number under the Label



a) 习题问题文本长度 CDF 曲线
a) CDF Curve of Exercise Question Text Length



b) 习题问题文本长度 CDF 曲线
b) CDF Curve of Exercise Answer Text Length

图 3-2 习题文本长度 CDF 曲线（X 轴经 log10 对数化）

Figure 3-2 CDF Curve of Exercise Text Length (X axis is log10 logarithmic)

除了对进行数量的统计分析，本论文还利用词云工具对词频排名前 2000 的习题文本关键词进行显性展示，挖掘数学习题的特点。如图 3-3，可以看到数字、运算符、数学术语等均高频出现，这符合常识，也说明数学习题文本的复杂性。可以自然而然想到的是，在寻找相似习题时，只需要搜索出关键词相同的习题，大概率就能找到其真实的相似题。事实上，也确实有研究将关键词等特征从习题文本中提取出来，利用词频矩阵[19]或送入 VSM[45]等传统模型进行相似搜索。这类方法的特点是速度很快，但很大程度上依赖于单词的出现频率和重要程度，这是由于数学习题的数学关系隐藏于大量的故事文本中，依赖于表面的文本难以挖掘其隐藏含义。简单地通过文字匹配、词袋模型等方法进行相似习题的寻找虽然在表述较为简单的习题类型中可能有用，但在故事叙述性强、文字复杂度高的习题类型中是不现实的。

图 3-3 习题文本词云

Figure 3-3 Word Cloud of Exercise Text

3.3 寻找相似习题任务的定义

本节将对寻找相似习题任务进行定义，包括寻找相似习题任务上数据集的构建以及相应的模型评估。

参照 Liu Qi 等人[16]对寻找相似习题任务的数学定义，本论文将寻找相似习题任务定义为排序任务，并利用习题相似度预测模型完成任务。具体的，给定一个习题e，本文的目标是在它的候选习题集E_c中寻找相似习题。为此，首先建立一个模型M去预测两道习题$e1$和$e2$的相似概率$Sim(e1,e2)$，计算过程如公式（3-1）所示：

$$Sim(e1,e2) = M(e1,e2,\theta) \tag{3-1}$$

其中，θ为模型M中的参数。接着计算给定习题e和候选集E_c中所有候选习题e_c^i的相似概率$Sim(e,e_c^i)$，最后对所有相似概率进行排序得到习题e的相似习题。

由于相似习题可以定义为考察学生知识点相同的习题[26]，本论文在构建数据集过程中，将与原题同一标签下的习题认为是该题的相似习题，而不同标签下的习题为该题的不相似习题。对 3.2 节中经过筛选后的 8135 道习题，首先进行 80% / 10% /10%的随机划分，分别得到 6470 的训练集、805 的验证集和 860 的测试集；接着为测试集中的每道习题构建候选集，其中包括相似题和不相似题。本论文在测试集中随机选取 5 道相似题（即在相同标签下随机采样 5 道习题），为了测试模型的稳定性，分别在测试集中随机选取 5、10、50、100、150、200 道不相似题（即在其他标签下随机采样 5 道习题）与这 5 道相似题构建为候选集$E_c = \{e_c^1, e_c^2, ..., e_c^p\}$，其中 p =10、15、55、105、155、205，表 3-2 为构建数据集（ p =10，相似题 5 道，不相似题 5 道）的示例。理想的模型应当能够在候选集中进行相似度概率计算并排序得到所有相似习题。

表 3-2 候选集示例

Table 3-2 Example of Candidate Dataset

原题	相似题（同标签）	不相似题（不同标签）
e1	*e2*、*e4*、*e5*、*e6*、*e7*	*e10*、*e11*、*e15*、*e16*、*e19*
e2	*e1*、*e4*、*e5*、*e7*、*e9*	*e14*、*e15*、*e16*、*e17*、*e18*
e3	*e10*、*e14*、*e15*、*e17*、*e19*	*e21*、*e24*、*e25*、*e28*、*e29*
…	…	…

为了衡量模型的表现，本论文利用平均精度均值（MAP）进行模型在寻找相似习题任务中的评估。首先根据其定义可以得到每道习题的 MAP 分数$p_e@k$，k表示选出前k个相似概率高的习题作为预测所得的相似题：

$$p_e@k = \frac{\sum_{i=1}^{k}(acc(e_c^i)/i)}{k} \tag{3-2}$$

其中，$acc\left(e_c^i\right)$ 表示候选集中的 e_c^i 与原题是否相似：

$$acc\left(e_c^i\right) = \begin{cases} 1 \rightarrow e_c^i \text{与} e \text{相似} \\ 0 \rightarrow \text{其他情况} \end{cases} \tag{3-3}$$

接着，可以计算出第 j 个标签的 MAP 分数 $p_c^j@k$：

$$p_c^j@k = \frac{\sum_{i=1}^{N_c^j} p_e^i@k}{N_c^j} \tag{3-4}$$

其中 N_c^j 为该标签下的测试集习题数量。最后可以得到所有标签的平均 MAP 分数 $p_{total}@k$：

$$p_{total}@k = \frac{\sum_{j=1}^{N_t} p_c^j@k}{N_t} \tag{3-5}$$

其中 N_t 为测试集中标签的数量。由于在构建数据集时选取相似题个数为 5，此时选取 k 为 5 计算上述公式。

3.4 现有模型（VSM）在习题任务上的表现

向量空间模型[2]作为一个典型的传统表征模型，被广泛应用于信息检索、文本分类等任务。本节将介绍本论文应用向量空间模型的具体思路，及其在应用于寻找相似习题任务时的表现。

首先，本论文利用基于 python 的 gensim 工具包实现向量空间模型的训练和测试。Gensim 是一个典型的自然语言处理工具，包含 corpora，models，similarities 等模块。其中：

(1) corpora 模块可以将文档集转化为用单词 id、词频数等代表的表现形式，易于模型训练，具体的，Dictionary 得到训练语料库中的单词与数字 id 对应的字典，而利用该字典可以将语料转化为带有词频信息的词袋模式；

(2) models 模块中包含 word2vec、TfidfModel 等模型，在本论文中可以利用 TfidfModel 得到经过语料库训练的 Tfidf 模型；

(3) similarities 模块能够对文档进行相似度检索，在本论文寻找相似题任务中将用 MatrixSimilarity 对候选集中的题通过计算余弦相似度进行排序，以便找出相似题，其中 num_features 参数表示文档向量对应的长度，也即单词总数，在本论文中取测试题单词数和候选集中单词数较大的数再加 1（平滑化）进行计算。

接着，本文在 3.3 节构建的数据集中进行寻找相似习题任务的实验，固定测试集中每道习题候选集中包含 5 道相似题（相似题数量c=5），分别在包含 5、10、50、100、200 道不相似习题的任务数据集中进行相似度排序得到在该任务中向量空间模型的 MAP 分数（定义见 3.3 节）。

任务结果如表 3-3 所示。从表中可以看到，向量空间模型在相似习题任务上的 MAP 分数随着候选数据集中不相似题数量的增多（相似题数量保持不变）而不断降低，而当不相似题数量大幅增加时，该模型的分数急剧降低。该结果表明，向量空间模型对习题这种复杂文本进行表征的能力有限，难以在相似习题任务获得较好的效果，尤其是习题种类繁多、构成复杂的情况下。

表 3-3 向量空间模型在习题任务上表现

Table 3-3 Vector Space Model Performance on Exercise Task

m（不相似题数量）	5	10	50	100	150	200
MAP 分数	0.80	0.74	0.53	0.45	0.41	0.38

为了进一步分析向量空间模型的文本表征能力，选出两类标签下的习题（图 3-4 中为“抢占制胜点”和“多位数除法的实际应用”），对这两类标签下的习题问题文本利用训练后的模型进行向量表征，并用 T-SNE 进行降维可视化。图 3-4 为这两类标签下习题问题文本表征在二维上的可视化，其中蓝色圆点为“抢占制胜点”标签下的习题表征，0 为它们的中心点，而绿色十字星为“多位数除法的实际应用”标签下的习题表征，1 为它们的中心点。可以发现这两类习题表征有较为明显的分界线，这说明向量空间模型有一定的表征能力，但同时也可以发现，仍有部分习题与非相似习题表征较近，这就会导致向量空间模型在候选集中难以寻找到相似的习题，使该模型在相似习题任务中得不到良好的效果。

经过对传统模型向量空间模型表现的分析，有如下观察：模型在习题任务中的表现与其习题文本表征能力密切相关。因此，在后续的工作中，本文的主要目标就是搭建适合的表征模型，将习题进行文本表征，映射在合适的向量空间中，进而在习题任务中得到一定的提升。

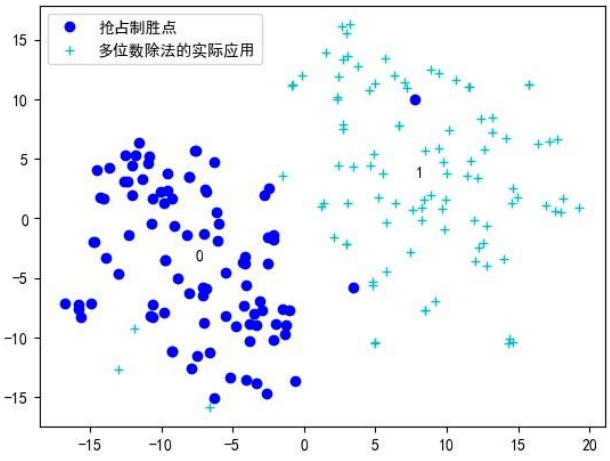


图 3-4 向量空间模型得到的习题表征

Figure 3-4 Exercise Representation Obtained by the Vector Space Model

3.5 本章总结

本章对采用的习题数据集进行了介绍，并对其进行两方面的统计分析：一是对习题标签、文本长度等进行统计展示，二是对习题的文本进行分析和展示；接着本章对本论文的寻找相似习题任务进行具体定义；除此之外，本章还展示了应用向量空间模型的具体思路，分析了该传统自然语言处理模型在习题任务中的效果，并确定后续工作的主要目标。

4 基于孪生神经网络架构的模型设计

本章将介绍本论文完成的基于孪生神经网络架构的习题相似度模型设计，包括其基本思想、模型具体结构设计及模型构建过程，并从模型训练效果、寻找相似习题任务上效果和习题表征可视化三个方面对模型进行分析。这些模型都对习题进行向量表征，并利用设计的模型进行回归预测。

4.1 基本思想

本节利用 BERT[10]对数学习题的表征效果进行可视化分析，发现 BERT 得到的习题表征效果并不理想，在分析原因并引入孪生神经网络架构对其进行相应改进后，用实验结果证明了基于孪生神经网络架构的习题相似度模型能够提升习题表征效果。

4.1.1 BERT 模型的习题表征效果

数学习题表征本质上也是文本的表征，而文本表征目前最先进的被广泛应用的方法是 BERT，本节将通过实验证明其实 BERT 在数学习题集上的表现并不理想。

本文使用 BERT 处理数学习题的复杂文本结构，它是最新的经典自然语言处理模型。如 2.2.4 节中所介绍的，BERT 因加入了 MLM 任务训练而能够捕获传统单向模型所不能捕获的双向语义信息，还因其加入 NSP 任务训练而能够捕获更多的句子关系信息。BERT 的结构优势已在许多自然语言处理任务中被证明有效，本文也将利用 BERT 捕获数学习题中的复杂文本关系。

本文以预测输入的两道习题是否是相似题为目标对 BERT 进行训练，接着用训练好的 BERT 模型得到习题表征，进而完成寻找相似习题等任务。BERT 将两道习题进行拼接送入模型内部，通过自注意力机制捕获起始输入所填充的[CLS]字符位置与输入的两道习题之间的单词联系，再将[CLS]对应的向量送入全连接网络预测两道习题是否相似。

为了分析 BERT 得到的习题表征效果，本文选出两类标签下的习题（图 4-1 中为“抢占制胜点”和“多位数除法的实际应用”），对这两类标签下的习题利用训练后的模型进行向量表征，并用 T-SNE 进行降维可视化（本文中降为二维）。结果如图 4-1 所示。在图 4-1 中，蓝色圆点为“抢占制胜点”标签下的习题表征，0

为它们的中心点，而绿色十字星为“多位数除法的实际应用”标签下的习题表征，1 为它们的中心点。

可以发现，BERT 得到的习题表征并不理想。表示标签为“抢占制胜点”习题的蓝点和表示标签为“多位数除法的实际应用”习题的绿点混杂在一起，两类习题表征没有有效分开。当习题表征效果不理想时，在习题任务中的效果也会大打折扣。

BERT 得到的习题表征效果不理想的原因可能如下：在训练时，BERT 以预测输入的两道习题是否是相似题为目标，在理想情况下 BERT 能够用一个超平面将两类习题分开，但显然的，它没有着重于增加不相似习题之间的距离，同时减少相似习题之间的距离，进而得到不相似习题距离较远、相似习题较近的空间。

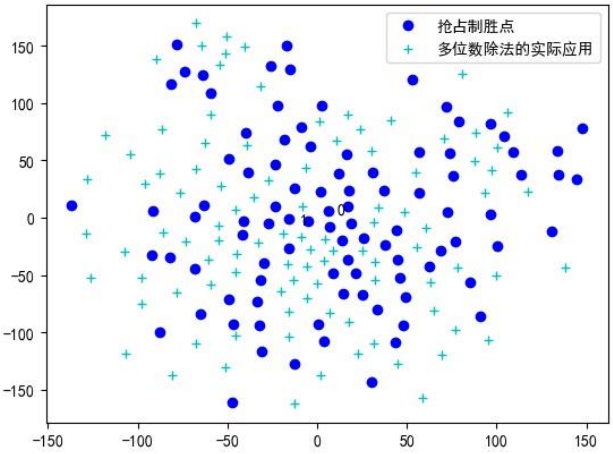


图 4-1 BERT 模型得到的习题表征

Figure 4-1 Exercise Representations Obtained by the BERT Model

4.1.2　孪生神经网络架构的引入

上一节可以发现 BERT 在数学习题上的表征效果并不理想，并分析了原因。孪生神经网络架构最开始用于图像识别任务，近些年的研究中也开始用于自然语言处理任务。

本文提出将孪生神经网络架构用于数学习题的文本表征。这是因为孪生神经网络架构的主要目标为将输入映射到一个可以用距离反映“语义”距离的空间，它能够将习题映射到相似习题距离较近、不相似习题距离较远的习题表征空间。具体来说，本章将利用 Siamese 架构和 Triplet 架构对 BERT 模型进行改进，将通过习题表征得到的表征距离映射到真实的语义距离中，根据这样的模型训练出效果更好的习题表征，以便在下游任务中得到进一步提升。

本文利用基于孪生神经网络架构改进的模型对数学习题进行表征，重复 4.1.1 节关于 BERT 的实验。图 4-2 选取了改进后的其中一个模型 SBERT-CLS。由图 4-2 可见，孪生神经网络可以清晰的把标签为“抢占制胜点”习题的蓝点和表示标签为“多位数除法的实际应用”习题的绿点分开，这证实孪生神经网络架构能够提升模型的习题表征能力。

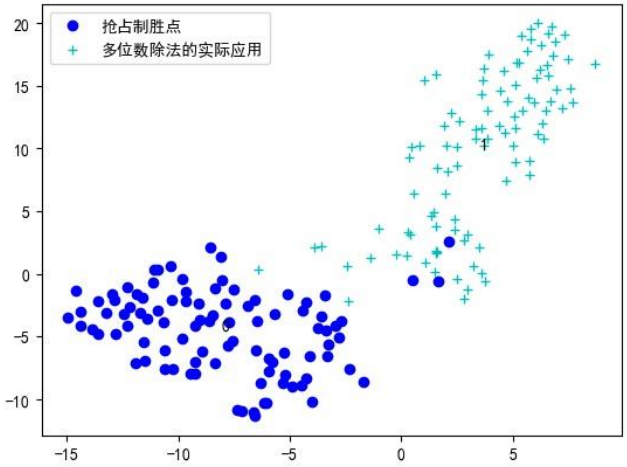


图 4-2 SBERT-CLS 模型得到的习题表征

Figure 4-2 Exercise Representations Obtained by the SBERT-CLS Model

4.2　基于孪生神经网络架构的模型结构

本节将具体介绍本文提出的基于孪生神经网络架构的习题相似度模型结构及其构建过程，分别包括基于 Siamese 架构的 SBERT 模型，基于 Triplet 架构的 TBERT 模型。为了提升输出表征，本文还进一步利用 Text-CNN 进行池化操作。

4.2.1　SBERT 模型

基于 Siamese 架构所设计的 SBERT 模型结构如图 4-3 所示：同时输入一道原题 *e1* 和一道相似/不相似题 *e2*（随机输入相似题或不相似题），两道习题同时通过两个具有相同结构和权重的 BERT 模型得到两个习题表征 R_{e1} 和 R_{e2}，将这两个习题表征相减并取绝对值得到的结果送入全连接网络中，预测它们是否相似（0 表示相似，1 表示不相似）。理论上，SBERT 能够利用 BERT 捕获习题文字中的数学逻辑和关系，并利用 Siamese 架构学习恰当的习题空间向量表征。

具体的，该模型的操作如公式（4-1）和（4-2）所示：

$$R = | R_{e1} - R_{e2} | \tag{4-1}$$

$$S_{\text{softmax}} = \text{softmax}(\mathbf{W}_{d_h \times 2} * R) + \mathbf{b}_{1\times 2} \tag{4-2}$$

其中，$\mathbf{W}_{d_h \times 2}$ 和 $\mathbf{b}_{1\times 2}$ 为全连接网络中的参数，d_h 为习题表征的维度。文本截断长度 L 在实验中设置为 128，由于 Transformer 的结构特性，d_h 必须为 128 的整数倍。在本论文中将 d_h 设置为 768，与 BERT[10]模型一致。

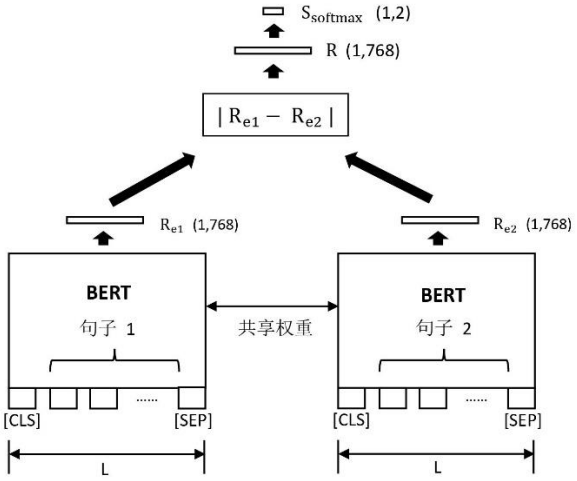


图 4-3 SBERT 模型架构

Figure 4-3 Structure of SBERT Model

为实现该模型的构建，本文利用基于 python 语言的神经网络框架 PyTorch 进行具体实现。首先，对模型进行初始化设置。在分词时，下载谷歌发布的中文词典 vocab.txt 以进行基于字的分词。在初始化词向量、模型矩阵向量等参数时，用继承得到父类 BertModel 的所有结构设置，并通过下载了谷歌中文预训练后的 BERT 模型将模型参数进行载入（从 tensorflow 存储模型类型 ckpt 转化为 PyTorch 存储模型类型 bin）。在超参数设置方面，将该模型分类重要参数 num_label 设置为 2（0 代表两习题输入相似，1 代表两习题输入不相似）。其他超参数通过读入谷歌公布的 bert_config.json 传入模型的 config 中进行定义，其中包括文本表征维度 hidden_size（设置为 768）、隐藏 dropout 概率 hidden_dropout_prob（设置为 0.1）等。

接着本文利用 nn 模块中的 Linear 模块搭建了一个输入维度为 hidden_size，输出维度为 num_label 的全连接层，利用 nn 模块中的 Dropout 模块搭建了 dropout 概率为 hidden_dropout_prob 的 dropout 层，同时，还搭建了 softmax 层和计算损失函数所用到的交叉熵函数。通过调用和串联模块完成了 SBERT 模型的代码搭建。forward 函数在完成模型逻辑后可以用于训练的反向传播，在训练数据集进行输入并通过向量计算得到最终的 loss 后，通过 loss.backward()可以进行模型参数的更新。在计算梯度时，本论文利用 Adam 优化器对参数更新进行调整。

4.2.2 TBERT 模型

基于 Triplet 架构所设计的 TBERT 模型结构如图 4-4 所示：同时输入一道原题 *e1* 和另两道习题 *e2*、*e3*（包括一道相似题和一道不相似题，顺序随机），三道习题同时通过三个具有相同结构和权重的 BERT 模型得到三个习题表征，将 *e1* 和 *e2* 的表征 R_{e1} 和 R_{e2} 相减并取绝对值得到它们在空间中的差别 R_1，将 *e1* 和 *e3* 的表征 R_{e1} 和 R_{e3} 相减并取绝对值得到它们在空间中的差别 R_2，最后模型将 R_1 和 R_2 相减后的结果送入全连接网络中预测原题 *e1* 与哪道题相似（0 表示与第一道题相似，1 表示与第二道题相似）。通常，基于 Triplet 架构的模型会比基于 Siamese 架构的模型取得更明显的效果，同时更依赖于高质量的训练样本。

具体的，该模型的操作如公式（4-3）~公式（4-6）所示：

$$R_1 = | R_{e1} - R_{e2} | \tag{4-3}$$

$$R_2 = | R_{e1} - R_{e3} | \tag{4-4}$$

$$R = R_1 - R_2 \tag{4-5}$$

$$S_{\text{softmax}} = \text{softmax}(\mathbf{W}_{d_h \times 2} * R) + \mathbf{b}_{1\times 2} \tag{4-6}$$

其中，$\mathbf{W}_{d_h \times 2}$ 和 $\mathbf{b}_{1\times 2}$ 为全连接网络中的参数，d_h 为习题表征的维度，在本论文中将其设置为 768。

本文同样利用 PyTorch 神经网络框架进行 TBERT 模型的构建。其参数初始化、模块调用等与 4.2.1 节中 SBERT 相差不大。同样下载谷歌中文词典 vocab.txt、谷歌中文预训练 BERT 模型、超参数设置文件 bert_config.json 对模型和输入进行初始化，并利用 PyTorch 神经网络框架中的 nn 模块搭建全连接层、Dropout 层等。值得一提的是，TBERT 模型中的 num_label 参数虽然也设置为 2，但与 SBERT 模型代码中的含义不同。参数设置和模型载入相同保证了模型的初始化设置相同，其效果完全由模型结构决定。TBERT 和 SBERT 模型的代码实现主要差别在 forward 函数中，在该函数中，将三道习题分别送入三个权重共享、结构相同的 BERT 模型中得到三个习题表征，再通过相减、取绝对值、交叉熵计算等更新损失函数。forward 函数同样可以进行模型参数的更新，在计算 loss 后，利用 loss.backward()进行反向传播更新模型参数，同样在计算梯度时利用 Adam 优化器对参数更新进行调整。

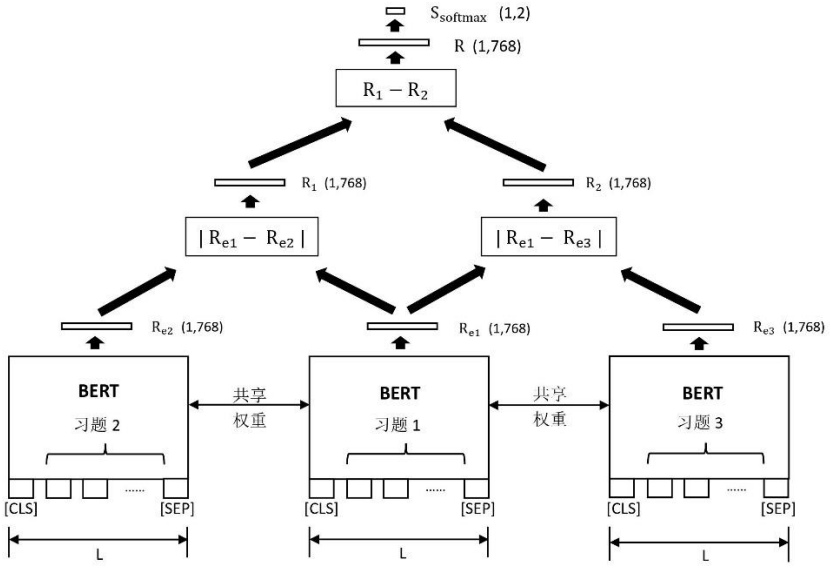
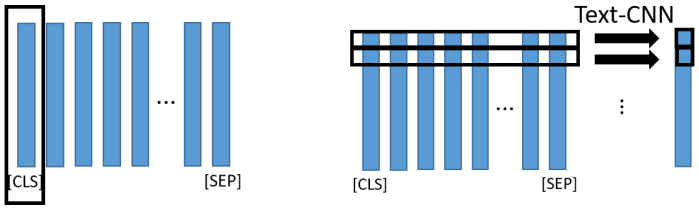


图 4-4 TBERT 模型架构

Figure 4-4 Structure of TBERT Model

4.2.3　CNN 池化操作

传统的 BERT 模型在整个文本的起始位置插入[CLS]标志，结束位置插入[SEP]标志，它将每个位置表征成能够捕获上下文信息的相同维度的向量，并将[CLS]位置对应的向量作为文本表征，这样的操作能够对输入进行表征，但无法充分捕获句子中所有位置上的信息。因此，本文将利用 Text-CNN 对包括[CLS]在内的所有位置上的文本对应的向量进行池化操作，以便获取综合全面的习题文本表征。具体的，设置 Text-CNN 的大小为1*L（L 为 BERT 设置的文本截断长度，在 3.2 节中将其设置为 128），用它对经过 BERT 后的$L*d_h$的向量矩阵进行特征提取，得到$1*d_h$的向量并输出，这就完成了 CNN 池化操作，得到了能够捕获句子中所有位置上信息的向量。



a) 原始池化操作　　b) CNN 池化操作

a) Original Pooling Operation　　b) CNN Pooling Operation

图 4-5 原始池化及 CNN 池化操作实现

Figure 4-5 Implementation of Original Pooling and CNN Pooling Operations

原始的BERT模型代码中（PyTorch版本）包含BertLayerNorm、BertSelfAttention等多个模块，其中，BertPooler 模块将已经过自注意力模块的向量矩阵 hidden_states 中对应于[CLS]的向量取出进行处理作为该句的向量表征，本文在此基础上进行了修改，实现了 CNN 池化操作。原始池化及 CNN 池化操作实现分别如图 4-5 a) 和 b) 所示。

4.3 数据集的构建

本节将介绍用于 SBERT 和 TBERT 模型训练、测试的数据集的构建。由于需要对模型尽量在相同情况下进行对比，在构建数据集时将尽量进行同步构建，使数据集差异最小。

具体的，本文首先对 3.2 节中经过筛选后的 8135 道习题，首先进行 80%/10%/10%的随机划分，分别得到 6470 的训练集、805 的验证集和 860 的测试集（同 3.3 节）。在生成模型需要的训练数据元组时，遍历训练集中的每道习题，在训练集中为其随机选取相似习题（同一标签下但题目 id 不同的习题）和不相似习题（不同标签下的习题），以 50%的概率得到<0，原题，相似习题，不相似习题>的元组，而另 50%的概率得到<1，原题，不相似习题，相似习题>的元组。生成模型需要的验证数据元组和测试数据元组与该过程类似，只不过生成验证数据元组时从验证集中选取习题组成元组，而在生成测试数据元组时从测试集中选取习题组成元组。在生成元组时将重复元组丢弃，遍历 5 次后得到训练数据元组 31849 条，验证数据元组 3551 条，测试数据元组 3786 条。

在训练 SBERT 时，截取格式为<0，原题，相似习题>和<1，原题，不相似习题>的数据元组以进行训练、验证和测试；在训练 TBERT 时，截取格式为<0，原题，相似习题，不相似习题>和<1，原题，不相似习题，相似习题>的数据元组以进行训练、验证和测试。另外，为了方便后续的实验，本文在生成元组时将每个习题的问题文本和解答文本都存储于其中。当利用习题的问题文本训练模型时，仅输入相应习题的问题文本，而当利用习题的解答文本训练模型时，仅输入相应习题的解答文本。

生成的数据元组样例如表 4-1 所示。可以看出，数学习题文本复杂，逻辑仅从文字表面难以挖掘，识别出相似习题也有一定难度。理论上，SBERT 和 TBERT 能够通过送入习题元组不断修正模型参数，对习题进行适当的表征以完成下游任务。由于每次输入都含有三道习题（包括一个正例和一个负例），TBERT 应当比 SBERT 收敛得更快。

表 4-1 数据元组样例

Table 4-1 Example Data Tuple

标志	习题	问题/解答	文本
0	原题	问题	一个箱子内装有$$2016$$颗棋子，两人轮流在其中取棋子，规定每人每次只能提取$$1$$、$$3$$、$$7$$颗棋子，不得不取，也不得多取，取到最后棋子的人取胜．为了确保取胜，你是愿意先手，还是愿意后手？说出你的选择答案和必胜的策略．
		解答	根据规则，虽然不能控制对方每次提取棋子得数量，但可以通过控制自己的数量保证每一轮双方提取棋子总和为$$4$$或$$8$$（对方取$$1$$颗，我方取$$7$$颗或$$3$$颗；对方取$$3$$颗，我方取$$1$$颗；对方取$$7$$颗，我方取$$1$$）．由于$$2016$$是$$8$$的倍数，选择后手提取，可以保证每次自己提取之后，剩余数量都是$$4$$的倍数，直至最后剩下$$8$$颗或$$4$$颗．在$$4$$颗的情况，对方只能取$$3$$或$$1$$颗，我方相应取$$1$$或$$3$$颗，取胜；在$$8$$颗的情况，对方若取$$1$$或$$7$$颗，我方相应取$$7$$或$$1$$颗，取胜；对方若取$$3$$颗，我方取$$1$$颗，转化为$$4$$颗的情况．
	相似题	问题	艾迪和薇儿两个人轮流在一个凸十六边形中画对角线．规定新画的对角线不能与已经有的相交，画最后一条者获胜．如果艾迪先画，则谁有必胜的策略？
		解答	艾迪连十六边形相对的两个顶点，将十六边形分成两个九边形．之后不管薇儿怎么连线，艾迪都连与之轴对称的线段即可，这样艾迪就能保证不败，故艾迪有必胜策略．
	不相似题	问题	已知中心在原点的椭圆$$C$$的右焦点为$$F(1,0)$$，离心率等于$$\frac{1}{2}$$，则$$C$$的方程是（ ）．
		解答	$$\frac{{{x}^{2}}}{3}+\frac{{{y}^{2}}}{4}=1$$$$\frac{{{x}^{2}}}{4}+\frac{{{y}^{2}}}{\sqrt{3}}=1$$$$\frac{{{x}^{2}}}{4}+\frac{{{y}^{2}}}{2}=1$$$$\frac{{{x}^{2}}}{4}+\frac{{{y}^{2}}}{3}=1$$ $$c=1,a=2,b=\sqrt{3}$$

4.4 结果分析

本节对基于孪生神经网络架构设计出的模型分别从模型训练效果、寻找相似习题任务上效果和习题表征可视化三个方面进行结果分析。除了对相同输入下不同的模型结果进行比较，还对同一模型分别输入问题文本和解答文本的结果进行比较。

4.4.1 模型训练效果

首先，本文对进行对照的基线模型以及本章中提出的模型进行简要介绍：

- BERT[10]：参考谷歌提出的 BERT 在句子对分类任务中的应用，本文将两道习题文本拼接起来送入 BERT 模型当中进行训练和测试；
- SLSTM：基于 Siamese 架构的模型，与 SBERT 模型不同的是习题文本送入 LSTM 得到习题表征；
- SBERT-CLS：本章所提出的 SBERT 模型，得到习题表征的方式与原论文相同，以输入文本的头部标志[CLS]对应的表征作为习题表征；
- SBERT-CNN：本章所提出的 SBERT 模型，同时通过 4.2.3 节的 CNN 池化操作得到习题表征；
- TLSTM[16]：基于 Triplet 架构的模型，将习题送入 LSTM 得到习题表征；
- TBERT-CLS：本章所提出的 TBERT 模型，习题表征通过输入文本的头部标志[CLS]经过 BERT 模型对应的表征得到；
- TBERT-CNN：本章所提出的 TBERT 模型，同时通过 4.2.3 节的 CNN 池化操作得到习题表征。

表 4-2 模型训练效果对比

Table 4-2 Model Training Effect Comparison

输入	模型	训练效果指标			
		AUC	准确率	召回率	F1 分数
问题文本	BERT[10]	0.50	0.50	0.48	0.46
	SLSTM	0.93	0.88	0.87	0.87
	SBERT-CLS	0.95	0.91	0.91	0.91
	SBERT-CNN	0.96	0.92	0.92	0.92
	TLSTM[16]	0.91	0.82	0.78	0.78
	TBERT-CLS	**0.94**	0.85	0.81	0.81
	TBERT-CNN	**0.94**	0.86	0.81	0.81
解答文本	SLSTM	0.86	0.80	0.79	0.79
	SBERT-CLS	0.90	0.84	0.83	0.83
	SBERT-CNN	0.92	0.86	0.86	0.86
	TLSTM[16]	0.75	0.82	0.69	0.66
	TBERT-CLS	0.82	0.77	0.73	0.71
	TBERT-CNN	0.82	0.78	0.73	0.71

除了模型的不同，本文还对模型分别输入问题文本和解答文本进行训练效果的对比，当输入文本为问题文本时，意味着训练数据元组、验证数据元祖和测试数

据元组全部由习题问题文本构成，输入为解答文本同理。表 4-2 为这些模型在测试集上的效果指标，包括 AUC 分数、Precision 分数、Recall 分数和 F1 分数。

通过表 4-2 中的结果可以从四个方面进行分析（未说明时，下列结论中所有模型分数均为 AUC 分数），进而得到以下结论：

(1) 从输入类型分析：输入为问题文本的模型比输入为解答文本的模型训练结果更好，这说明习题的解答文本构成比习题问题文本更为复杂，在习题任务中不能完全代替问题文本进行模型的训练。

(2) 从表征模型分析：在输入为习题的问题文本时，SBERT-CLS 比 SLSTM 高出 0.02，TBERT-CLS 比 TLSTM 高出 0.03；在输入为习题的解答文本时，SBERT-CLS 比 SLSTM 高出 0.04，TBERT-CLS 比 TLSTM 高出 0.07。可以看出同样的架构和输入下，BERT 对应的模型效果总是高于利用 LSTM 进行表征的模型，因此，BERT 模型的结构更适合处理习题文本这种逻辑复杂的文本。

(3) 从模型架构分析：基于 Triplet 架构的模型训练效果指标总是低于基于 Siamese 架构的模型，但由于其输入数据元组不同、任务不同，该结果不具备强有力的代表性。

(4) 从池化方式分析：SBERT-CNN 在输入为习题的问题文本和解答文本时分别比 SBERT-CLS 效果提升了 0.01 和 0.02，而 TBERT-CNN 在输入为习题的问题文本和解答文本时效果和 SBERT-CLS 相同。这说明 CNN 的池化方式能够在一些情况下提升一定的模型训练效果（尤其是在未加入 CNN 池化时模型分数较低的情况下），即 CNN 池化能够获取更好的习题文本表征。

最后与基线模型进行对比。当输入问题文本进行训练时，本论文所提出的 TBERT-CLS 和 TBERT-CNN 模型比 TLSTM 模型[16]的 AUC 分数高出 0.03（即相对提升了 3.3%），其余各项分数也均有提升。

4.4.2 寻找相似习题任务

除了对比模型训练的效果，本文还将计算并对比它们在寻找相似习题任务中的 MAP 分数。在 3.4 节中有如下观察：模型在习题任务中的表现与其习题文本表征能力密切相关，也就是说，当设计出的模型能够得到质量较高的习题文本表征时（相似的习题文本表征相距更近，不相似的习题文本表征相距更远），该模型更容易在寻找相似习题这类任务中得到更好的效果。

由表 4-2 可以发现输入习题的解答文本训练效果普遍不如输入习题的问题文本，在后续的实验中也可以发现这一现象，为了简化实验，本文只对 SBERT-CLS、SBERT-CNN、TBERT-CLS、TBERT-CNN 这四类本章提出的模型和 SLSTM、TLSTM

这两个参照模型分别输入习题的问题文本和解答文本，这些实验结果也将在第五章进行进一步的分析。

表 4-3 模型在寻找相似习题任务中的效果对比

Table 4-3 Model Effect Comparison in Finding Similar Exercises Task

输入	模型	寻找相似习题任务中的 MAP 分数					
	m（不相似题数量）	5	10	50	100	150	200
问题文本	VSM[2]	0.80	0.74	0.53	0.45	0.41	0.38
	BERT[10]	0.49	0.32	0.08	0.05	0.03	0.02
	SLSTM	0.92	0.86	0.62	0.48	0.42	0.35
	SBERT-CLS	0.93	0.88	0.65	0.50	0.41	0.35
	SBERT-CNN	0.95	0.91	0.74	0.64	0.58	0.51
	TLSTM[16]	0.88	0.82	0.58	0.46	0.40	0.35
	TBERT-CLS	0.95	0.92	0.78	0.69	0.65	**0.61**
	TBERT-CNN	0.95	0.92	0.77	0.69	0.65	**0.61**
解答文本	SLSTM	0.83	0.73	0.42	0.31	0.24	0.22
	SBERT-CLS	0.86	0.79	0.50	0.36	0.30	0.26
	SBERT-CNN	0.90	0.84	0.59	0.46	0.39	0.34
	TLSTM[16]	0.79	0.70	0.41	0.31	0.26	0.22
	TBERT-CLS	0.87	0.81	0.57	0.48	0.42	0.37
	TBERT-CNN	0.89	0.83	0.62	0.51	0.45	0.41

表 4-3 为不同模型（及不同输入）在固定相似习题为 5 道而不相似习题数分别为 5、10、50、100、200 道不相似习题的随机生成的任务数据集中进行相似度排序得到的 MAP 分数。需要说明的是，本文在实验时所用的数据集构建方式与模型评估和 3.3 节中相同，不同的是，用问题文本进行实验时，每道测试习题的候选集全部为习题的问题文本构成的候选集，而用解答文本进行实验时，每道测试习题的候选集全部为习题的解答文本构成的候选集。同时，尽管基于 Siamese 架构的模型与基于 Triplet 架构的模型训练所需的数据格式不同，但在该任务中，都利用 Siamese 架构计算两道习题相似的概率，只是在载入模型参数时，应当载入不同模型训练后的所存储的参数。

在表 4-3 中可以看到，BERT 在习题任务上的分数很低，近似于“随机”推荐相似习题。而 VSM 是分数最高的基线模型。

通过表 4-3 中的结果本文可以从以下四个方面进行分析（未说明时，下列结论中所有模型 MAP 分数均为候选集里相似题数量为 5、不相似题数量为 200 的寻找相似习题任务中的分数），进而得到以下结论：

(1) 从输入类型分析：输入为问题文本的模型比输入为解答文本的模型在寻找相似习题任务中表现效果更好，这与 4.4.1 节中结果一致，同样说明了习题的解答文本构成较为复杂，不能完全代替问题文本进行模型的训练。

(2) 从表征模型分析：输入为问题文本时，SBERT-CLS 与 SLSTM 相同，TBERT-CLS 比 TLSTM 高出 0.26；输入为解答文本时，SBERT-CLS 比 SLSTM 高出 0.04，TBERT-CLS 比 TLSTM 高出 0.15。这与 4.4.1 节中结果基本一致，即同样的框架和输入下，BERT 对应的模型表现总是优于利用 LSTM 进行表征的模型，进一步说明了 BERT 模型的结构更适合处理习题文本这种逻辑复杂的文本。

(3) 从模型架构分析：输入为问题文本时，TLSTM 与 SLSTM 相同，TBERT-CLS 比 SBERT-CLS 高出 0.26，SBERT-CLS 比 BERT 高出 0.33；输入为解答文本时，TLSTM 与 SLSTM 相同，TBERT-CLS 比 SBERT-CLS 高出 0.11。基于 Siamese 架构的模型分数比未基于孪生网络架构的模型高，而基于 Triplet 架构的模型分数比基于 Siamese 架构的模型高，说明基于 Siamese 架构的模型能够提升模型的文本表征能力，而基于 Triplet 架构的模型文本表征能力能对其进行进一步提升。

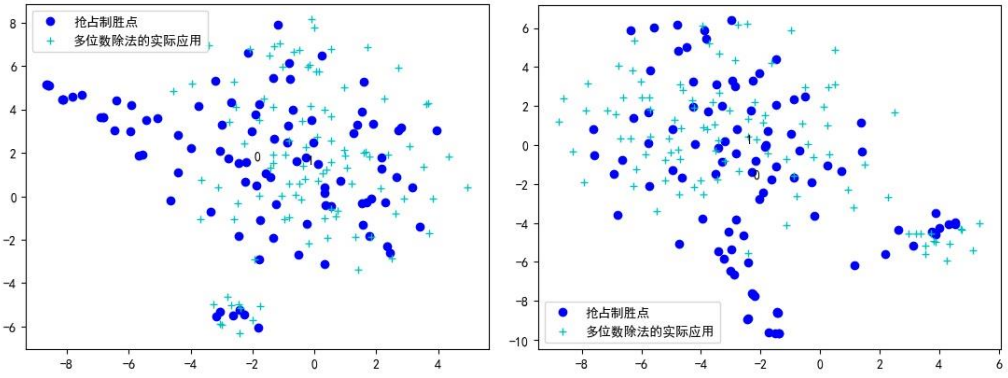
(4) 从池化方式分析：SBERT-CNN 在输入为习题的问题文本和解答文本时分别比 SBERT-CLS 效果提升了 0.16 和 0.08，而 TBERT-CNN 在输入为习题的问题文本和解答文本时分别比 TBERT-CLS 效果提升了 0.0 和 0.04。这与 4.4.1 节中结果一致，即 CNN 的池化方式能够提升一定的模型训练效果（尤其是在未加入 CNN 池化时模型分数较低的情况下），说明了 CNN 池化能够获取更好的习题文本表征。

最后与基线模型进行对比。在用习题问题文本训练和构造候选集的条件下，本论文所提出的 TBERT-CLS 和 TBERT-CNN 模型 MAP 分数为 0.61，表现最好。它们比表现最好的基线模型 VSM 的 MAP 分数高出 0.23（即相对提升了 60.5%）。

4.4.3 习题表征可视化

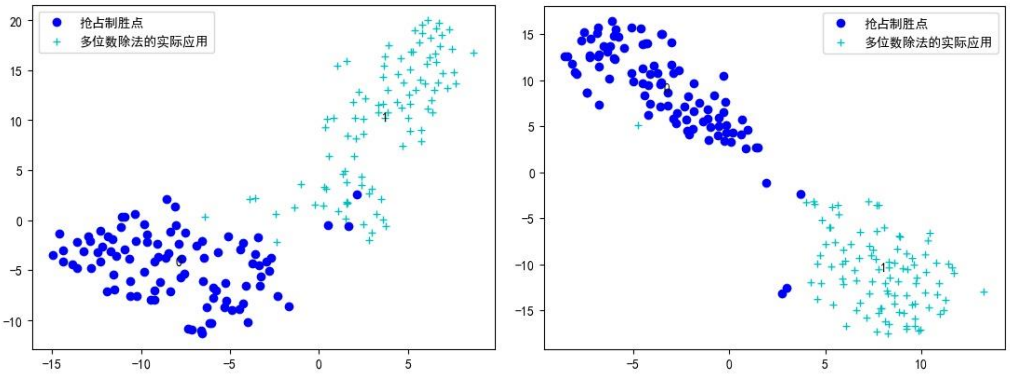
在 3.4 节中，可以得出结论：模型在习题任务中的表现与其习题文本表征能力密切相关。理论上，对 BERT 模型基于孪生神经网络架构的改造可以提升模型对于习题文本的表征能力。

将习题表征进行可视化可以辅助进行实验分析。因此，为了进一步分析得到的习题表征效果，类似于 4.1 节，同样选出两类标签下的习题（图 4-6 中为“抢占制胜点”和“多位数除法的实际应用”），对这些习题利用各类模型进行向量表征及降维可视化。其中，蓝色圆点为“抢占制胜点”标签下的习题表征，0 为它们的中心点，而绿色十字星为“多位数除法的实际应用”标签下的习题表征，1 为它们的中心点。



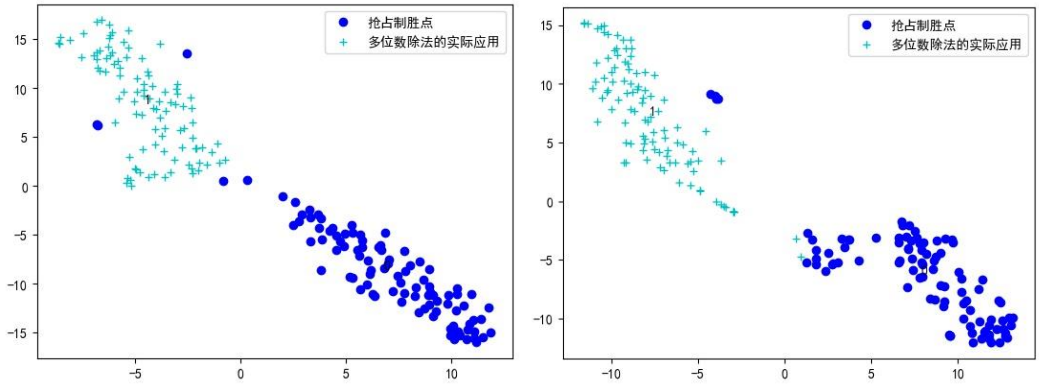
a) SLSTM 得到的习题表征 b) TLSTM 得到的习题表征

a) Exercise Representation of SLSTM b) Exercise Representation of TLSTM



c) SBERT-CLS 得到的习题表征 d) SBERT-CNN 得到的习题表征

c) Exercise Representation of SBERT-CLS d) Exercise Representation of SBERT-CNN



e) TBERT-CLS 得到的习题表征 f) TBERT-CNN 得到的习题表征

e) Exercise Representation of TBERT-CLS f) Exercise Representation of TBERT-CNN

图 4-6 不同模型得到的习题表征

Table 4-6 Exercise Representations Obtained by Different Models

在图 4-6 中，SLSTM 和 TLSTM 模型为参照模型，而 SBERT-CLS、SBERT-CNN、TBERT-CLS、TBERT-CNN 为本章所提出的模型。经过观察可以得到以下结论：

(1) SLSTM 和 TLSTM 模型得到的习题表征在降维后并没有明显的分界线，这说明这两类没有足够好的文本表征能力。

(2) 与 SLSTM（TLSTM）相比，SBERT-CLS 和 SBERT-CNN（TBERT-CLS 和 TBERT-CNN）得到的习题表征有相对明显的分界线，且两类标签下习题表征的中心点距离较远，这说明 BERT 模型能够更好进行习题文本的表征的结论；

(3) 与 BERT 模型对比，SBERT 模型得到的习题表征中，每类习题的表征更靠近相应类别的中心点，而 TBERT 模型得到的习题表征比 SBERT 的类内距离更小，这说明基于 Siamese 架构的模型能够提升文本表征能力，而基于 Triplet 架构的模型具有更强的文本表征能力；

(4) 与 SBERT-CLS（TBERT-CLS）相比，SBERT-CNN（TBERT-CNN）得到的习题表征离群点(与正确类别中心点距离大于与错误类别中心点距离)更少，且每类习题表征更靠近相应类别的中心点，这说明 CNN 池化能提升一定文本表征能力。

综上所述，能够发现 SBERT-CLS、SBERT-CNN、TBERT-CLS、TBERT-CNN 得到的习题表征有明显的效果提升，说明本章中所提出的基于孪生神经网络架构的模型习题文本表征能力较强，进而在习题任务中能够得到良好的效果提升。

4.5 本章总结

本章提出了基于孪生神经网络架构的模型，对其基本思想、模型结构和模型构建过程分别进行了介绍，同时详细介绍了为实现模型训练的数据集的构建过程，并对模型训练的结果进行了分析。最终，本章提出的 TBERT-CLS 和 TBERT-CNN 模型 MAP 分数为 0.61，比表现最好的基线模型 VSM 的 MAP 分数高出 0.23（即相对提升了 60.5%）。

5 基于习题问题与解答的融合模型设计

本章将介绍本论文完成的基于习题问题与解答的融合模型设计，包括其设计思路、模型具体结构设计及模型构建过程，并从模型训练效果、寻找相似习题任务上效果和习题表征可视化三个方面对模型进行分析。

5.1 设计思路

在上一章中，本文基于孪生神经网络架构提升了模型的习题文本表征能力，本章将同时对习题的问题与解答进行表征，并将它们融合利用，实现能够捕获习题问题和解答文本关系的习题相似度模型。

对习题的问题文本和解答文本进行了数据挖掘，可以得到两个重要发现：

(1) 部分习题的问题文本和解答文本会有反复出现的单词，这些单词有时是习题所属类型的重要标志。由于习题文本中的 latex 符号会干扰统计结果，本论文在整个习题数据集上进行了简单的正则处理，接着分别对习题的问题文本和解答文本进行分词。经统计，在问题文本和解答文本中均出现的单词的比例（不计重复）平均为 22.1%，也就是说，平均来看，对于某一道习题来说，问题文本和解答文本中均出现的单词占问题文本和解答文本中所有单词的 22.1%。表 1-1 中习题 1、习题 2 和习题 3 为属于“抢占制胜点”这一标签的习题，可以看出，“胜”、“必胜”等字眼多次重复于习题的问题文本和解答文本中，它们也是该标签下习题的重要标志（但显然的，这 22.1%中的重复单词不都是该类习题的关键词）。

(2) 部分关键词仅在习题的问题文本或解答文本中出现。经统计，有 38.4%的单词仅在问题文本中出现而不在解答文本中出现，而 39.5%的单词仅在解答文本中出现而不在问题文本中出现。举例来说，“先”、“后”等关键词在表 1-1 这 3 道属于“抢占制胜点”标签的习题的问题文本中均有出现，但却仅在习题 1 的解答文本中出现；反过来，习题解答文本中的计算公式也是这一标签的重要计算过程，大部分属于“抢占制胜点”的习题都需计算商和余数以确定策略，但计算公式在很多情况下不会出现在这一类习题的问题文本中（但显然的，不是所有解答文本中有计算商和余数的公式的习题都属于“抢占制胜点”，它也可以属于“多位数除法的实际应用”，如表 1-1 中的习题 4）。

因此，本文要将问题文本和解答文本结合起来，更好地捕获习题问题文本和解答文本之间的关系，提升模型表征的能力。设计出的习题相似度模型不仅应当支持两道习题的问题文本-问题文本的匹配和解答文本-解答文本的匹配，还应当支持一

道习题的问题文本和另一道习题的解答文本之间的相似度匹配。为了实现能够捕获习题问题和解答文本关系的习题相似度模型，本章将同时对习题的问题与解答进行表征，并将它们融合利用。具体来说，本章将在 SBERT 和 TBERT 的基础上将习题问题与解答表征进行拼接设计，进一步提升习题表征质量，以便在下游任务中得到进一步提升。

5.2 基于习题问题与解答的融合模型结构

本节将具体介绍本文提出的基于习题问题与解答的融合模型结构及其构建过程，分别包括在 Siamese 架构上融合习题问题与解答的 SBERT-QA 模型，在 Triplet 架构上融合习题问题与解答的 TBERT-QA 模型。除此之外，CNN 池化操作也能够应用于这些模型上。

5.2.1 SBERT-QA 模型

基于 Siamese 架构融合利用习题问题和习题解答的 SBERT-QA 模型架构如图 5-1 所示：同时输入一道原题*e1*和一道相似/不相似题*e2*（随机输入相似题或不相似题）的问题文本和解答文本，它们同时通过四个权重共享、结构相同的 BERT 模型分别得到四个文本表征 R_{Q1}， R_{A1}， R_{Q2} 和 R_{A2}，将习题对应的问题文本表征和解答文本表征拼接起来得到每道习题的表征 R_{e1} 和 R_{e2}，并将这两个习题表征相减并取绝对值得到的结果送入全连接网络中，预测它们是否相似（0 表示相似，1 表示不相似）。理论上，在 Siamese 架构学习恰当的向量表征的过程中，融合利用习题问题和习题解答能够提升习题表征的质量。

具体的，该模型的操作如公式（5-1）~公式（5-4）所示：

$$R_{e1} = [R_{Q1}, R_{A1}] \tag{5-1}$$

$$R_{e2} = [R_{Q2}, R_{A2}] \tag{5-2}$$

$$R = | R_{e1} - R_{e2} | \tag{5-3}$$

$$S_{\text{softmax}} = \text{softmax}(\mathbf{W}_{d_h \times 2} * R) + \mathbf{b}_{1\times 2} \tag{5-4}$$

其中，$\mathbf{W}_{d_h \times 2}$ 和 $\mathbf{b}_{1\times 2}$ 为全连接网络中的参数，d_h 为习题表征的维度，在本论文中将其设置为 768。

实现该模型的代码基于 PyTorch 完成，基于字分词、参数初始化和超参数设置与 SBERT 模型相同。不同的是，SBERT-QA 模型在利用 BertModel()模块得到习题的问题文本表征和解答文本表征之后，通过 torch.stack()函数将两个表征拼接起来作为相应习题的文本表征。模型搭建完毕之后，在训练过程中利用 Adam 优化器进行反向传播调整参数。

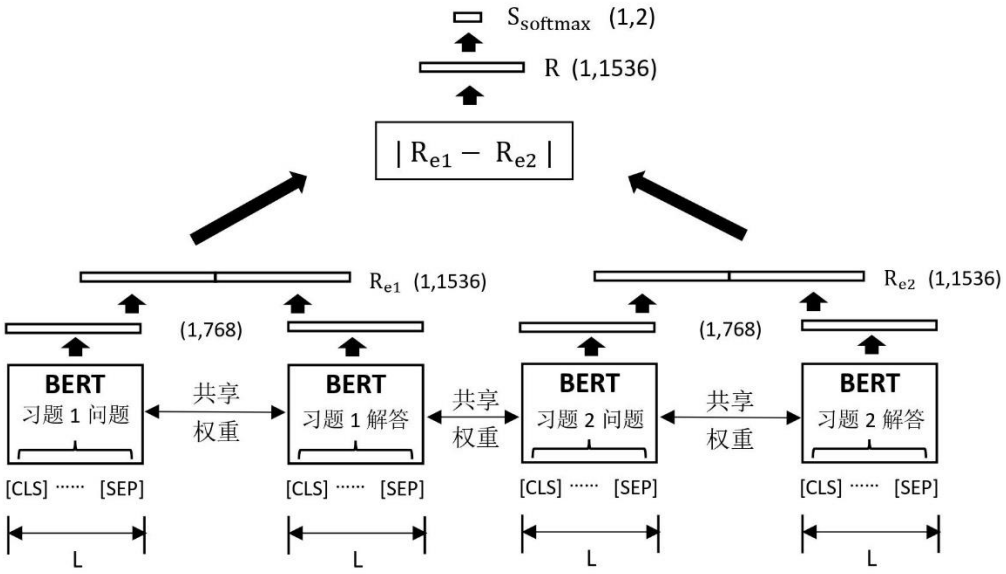


图 5-1 SBERT-QA 模型架构

Figure 5-1 Architecture of SBERT-QA Model

5.2.2　TBERT-QA 模型

基于 Triplet 架构融合利用习题问题和习题解答的 TBERT-QA 模型架构如图 5-2 所示：同时输入一道原题 *e1* 和另两道习题 *e2*、*e3*（包括一道相似题和一道不相似题，顺序随机）的问题文本和解答文本，它们同时通过六个权重共享、结构相同的 BERT 模型得到三个习题问题文本表征 R_{Q1}，R_{Q2}，R_{Q3} 和三个习题解答文本表征 R_{A1}，R_{A2}，R_{A3}，将每道习题的问题文本和解答文本表征拼接起来得到该习题的习题表征。将 *e1* 和 *e2* 的表征 R_{e1} 和 R_{e2} 相减并取绝对值得到它们在空间中的差别 R_1，将 *e1* 和 *e3* 的表征 R_{e1} 和 R_{e3} 相减并取绝对值得到它们在空间中的差别 R_2，最后模型将 R_1 和 R_2 相减后的结果送入全连接网络中预测原题 *e1* 与哪道题相似（0 表示与第一道题相似，1 表示与第二道题相似）。理论上，TBERT-QA 模型能够结合 Triplet 架构的优势和习题文本特征，在习题任务中得到最佳的效果。

具体的，该模型的操作如公式（5-5）~公式（5-11）所示：

$$R_{e1} = [R_{Q1}, R_{A1}] \tag{5-5}$$

$$R_{e2} = [R_{Q2}, R_{A2}] \tag{5-6}$$

$$R_{e3} = [R_{Q3}, R_{A3}] \tag{5-7}$$

$$R_1 = | R_{e1} - R_{e2} | \tag{5-8}$$

$$R_2 = | R_{e1} - R_{e3} | \tag{5-9}$$

$$R = R_1 - R_2 \tag{5-10}$$

$$S_{\text{softmax}} = \text{softmax}(\mathbf{W}_{d_h \times 2} * R) + \mathbf{b}_{1\times 2} \tag{5-11}$$

其中，$\mathbf{W}_{d_h \times 2}$ 和 $\mathbf{b}_{1\times 2}$ 为全连接网络中的参数，d_h 为习题表征的维度，在本论文中将其设置为 768。

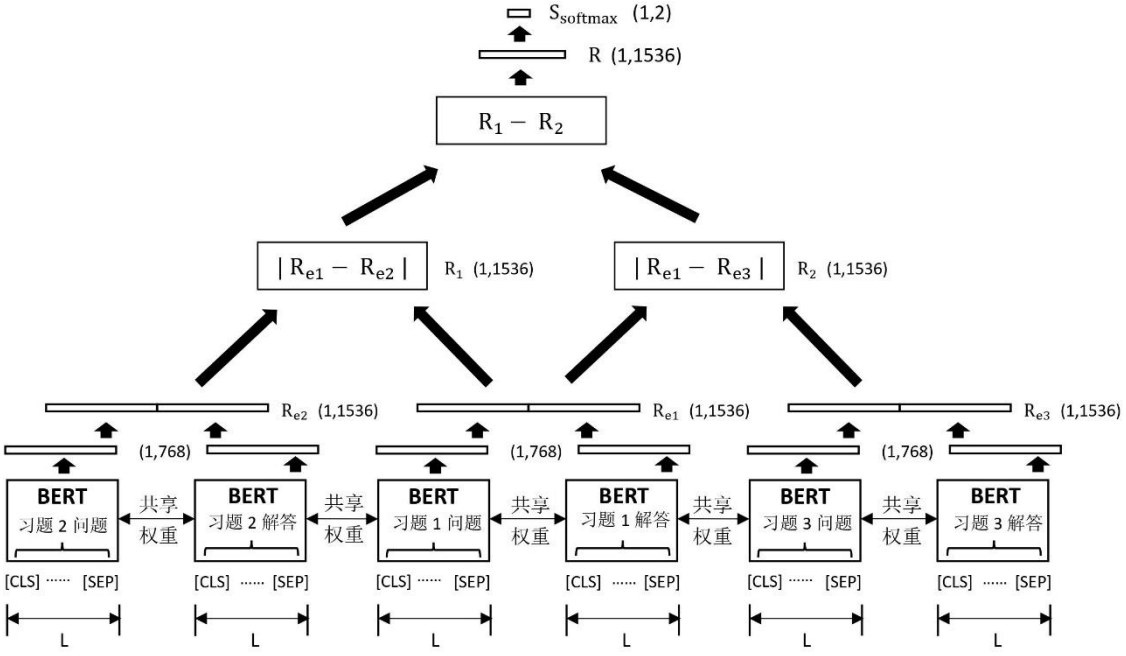


图 5-2 TBERT-QA 模型架构

Figure 5-2 Architecture of TBERT-QA Model

在 TBERT 模型代码的基础上，输入习题问题文本和解答文本，利用 BertModel() 模块和 torch.stack()函数实现习题问题文本表征和解答文本表征的拼接，分别得到原题、相似题和不相似题的习题表征。模型搭建之后的反向传播同样通过 Adam 优化器进行。

类似于 4.2.3 节，同样可以在 SBERT-QA 和 TBERT-QA 模型中利用 CNN 池化操作优化文本表征。具体操作通过 4.2.3 节中提出的 CNN 池化操作实现，这里不再赘述，效果分析将在 5.4 节中介绍。

5.3 数据集的构建

用于 SBERT-QA 和 TBERT-QA 模型训练、测试的数据集的构建与 SBERT 和 TBERT 类似，不同的是，在生成训练数据元组时，将同时输入每道习题的问题文本和解答文本，并在训练时送入模型以生成文本表征。

为了保持训练数据的一致性，SBERT-QA 和 TBERT-QA 模型对应的数据集构建过程与 4.3 节一致，仅在送入模型时选取不同的输入。具体的，在训练 SBERT-QA 时，截取格式为<0，原题问题文本，原题解答文本，相似习题问题文本，相似习题解答文本>和<1，原题问题文本，原题解答文本，不相似习题问题文本，不相似习题解答文本>的数据元组以进行训练、验证和测试；在训练 TBERT-QA 时，用格式为<0，原题问题文本，原题解答文本，相似习题问题文本，相似习题解答文本，不相似习题问题文本，不相似习题解答文本>和<1，原题问题文本，原题解答文本，不相似习题问题文本，不相似习题解答文本，相似习题问题文本，相似习题解答文本>的数据元组以进行训练、验证和测试。与 SBERT 和 TBERT 一致，训练数据元组共 31849 条，验证数据元组共 3551 条，测试数据元组共 3786 条。

5.4 结果分析

本节将对基于习题问题与解答设计出的融合模型进行结果分析。

5.4.1 模型训练效果

首先，本文对进行对照的基线模型以及本章中提出的模型进行简要介绍：

- SBERT-QA：本章所提出的 SBERT-QA 模型，得到习题表征的方式与原论文相同，以输入文本的头部标志[CLS]对应的表征作为习题表征；
- SBERT-QA-CNN：本章所提出的 SBERT-QA 模型，得到习题表征时，其中习题的问题文本表征和解答文本表征分别通过 4.2.3 节的 CNN 池化操作进行池化得到；

- TBERT-QA：本章所提出的 TBERT-QA 模型，得到习题表征的方式与原论文相同，以输入文本的头部标志[CLS]对应的表征作为习题表征；
- TBERT-QA-CON：本章所提出的 TBERT-QA 模型，得到习题表征时，其中习题的问题文本表征和解答文本表征分别通过 4.2.3 节的 CNN 池化操作进行池化得到。

上述模型在训练和测试时均通过输入习题的问题和解答文本进行训练、验证和测试，数据格式在 5.3 节中进行了介绍。表 5-1 为这些模型在测试集上的效果指标，包括包括 AUC 分数、Precision 分数、Recall 分数和 F1 分数。

表 5-1 模型训练效果对比

Table 5-1 Model Training Effect Comparison

输入	模型	训练效果指标			
		AUC	准确率	召回率	F1 分数
问题、解答文本	SBERT-QA	**0.97**	0.93	0.93	0.93
	SBERT-QA-CNN	0.96	0.91	0.91	0.91
	TBERT-QA	0.95	0.89	0.89	0.89
	TBERT-QA-CNN	0.95	0.89	0.88	0.88

通过表 5-1 中的结果本文可以从三个方面进行分析（未说明时，下列结论中所有模型分数均为 AUC 分数），进而得到以下结论：

(1) 从输入类型分析：与表 4-2 对比可以发现，SBERT-QA 模型比 SBERT-CLS 模型输入为问题文本/解答文本时提升了 0.02/0.07；而 TBERT-QA 模型比 TBERT-CLS 模型输入为问题文本/解答文本时提升了 0.01/0.13。这说明习题的问题文本和解答文本中均包含习题的重要信息，充分利用它们能够提升模型的文本表征能力。

(2) 从模型架构分析：基于 Triplet 架构的模型训练效果指标总是低于基于 Siamese 架构的模型，但由于其输入数据元组不同、任务不同，该结果不具备强有力的代表性。

(3) 从池化方式分析：可以看出 SBERT-QA-CNN 模型比 SBERT-QA 模型低了 0.01，而 TBERT-QA-CNN 与 TBERT-QA 一样。从训练结果上看，CNN 池化并不能明显提升模型效果，这一点将在 5.4.2 节中继续分析。

最后与基线模型进行对比。本文提出的 SBERT-QA 模型比输入为问题的 SLSTM 模型高出 0.03（即相对提升了 3.2%），TBERT-QA 模型比输入为问题的 TLSTM 模型高出 0.04（即相对提升了 4.4%）。

5.4.2 寻找相似习题任务

本节将计算并对比上述在寻找相似习题任务中的 MAP 分数。

通过表 5-2 中的结果本文可以从三个方面进行分析（未说明时，下列结论中所有模型 MAP 分数均为候选集里相似题数量为 5、不相似题数量为 200 的寻找相似习题任务中的分数），进而得到以下结论：

(1) 从输入类型分析：与表 4-3 对比可以发现，SBERT-QA 模型比 SBERT-CLS 模型在输入为问题文本和解答文本时分别提升了 0.25 和 0.34；而 TBERT-QA 模型比 TBERT-CLS 模型在输入为问题文本和解答文本时分别提升了 0.04 和 0.28。这说明习题的问题文本和解答文本中均包含习题的重要信息，能够提升模型的文本表征能力。

(2) 从模型架构分析：TBERT-QA 模型比 SBERT-QA 模型高出 0.05，TBERT-QA-CNN 模型比 SBERT-QA-CNN 模型高出 0.11。这说明基于 Triplet 架构的模型比基于 Siamese 架构的模型文本表征能力更强。

(3) 从池化方式分析：SBERT-QA-CNN 模型比 SBERT-QA 模型低 0.05，TBERT-QA-CNN 模型比 TBERT-QA 模型高出 0.01，说明 CNN 池化方式能够在一定情况下提升一定的模型表征能力。

融合习题问题和解答文本的 TBERT-QA 模型 MAP 分数为 0.65，比仅考虑习题问题文本相似度匹配的 TBERT-CLS 模型高出 0.04（即相对提升了 6.6%），且比表现最好的基线模型 VSM（表 4-3）的高出 0.27（即相对提升了 71.1%）。进一步的，加入 CNN 池化方式的 TBERT-QA-CNN 在测试集上的 MAP 分数比 TBERT-QA 高出 0.01（即相对提升了 1.5%），且比表现最好的基线模型 VSM 的高出 0.28（即相对提升了 73.7%）。

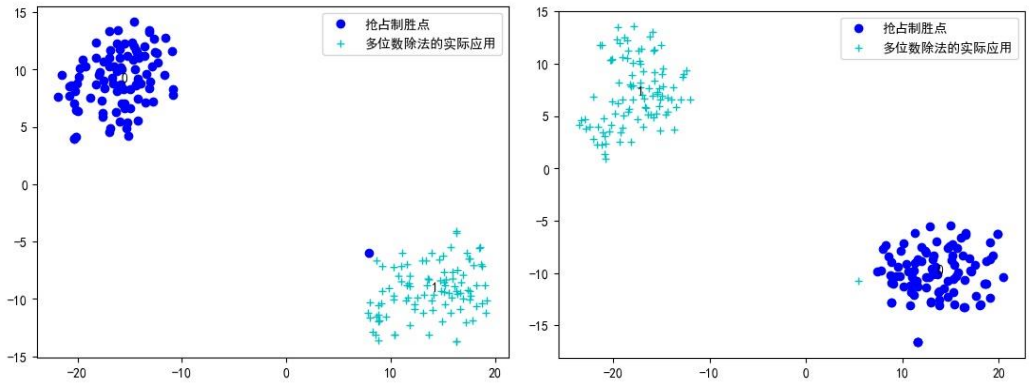
表 5-2 模型在寻找相似习题任务中的效果对比

Table 5-2 Model Effect Comparison in Finding Similar Exercises Task

输入	模型	寻找相似习题任务中的 MAP 分数					
	m（不相似题数量）	5	10	50	100	150	200
问题、解答文本	SBERT-QA	0.96	0.92	0.81	0.71	0.64	0.60
	SBERT-QA-CNN	0.95	0.91	0.77	0.68	0.60	0.55
	TBERT-QA	0.96	0.94	0.82	0.74	0.69	**0.65**
	TBERT-QA-CNN	0.96	0.94	0.81	0.74	0.70	**0.66**

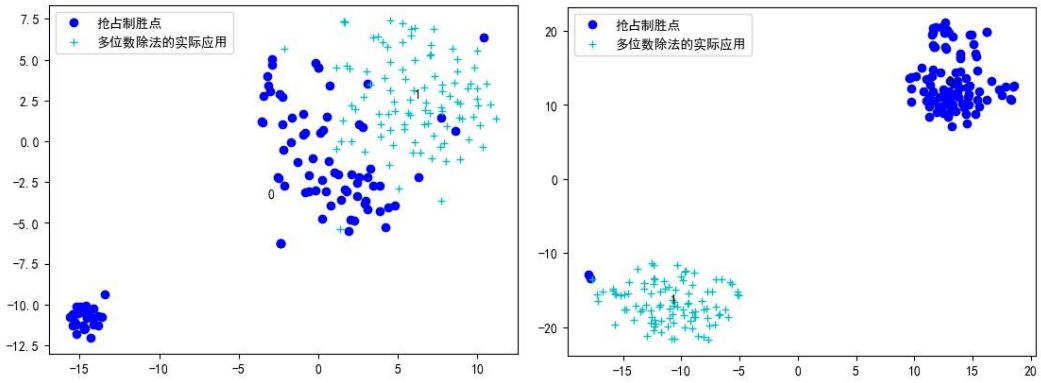
5.4.3　习题表征可视化

习题表征可以辅助本文进行实验分析，因此本节同样选出两类标签下的习题（图 5-3 中为“抢占制胜点”和“多位数除法的实际应用”），对这些习题利用各类模型进行向量表征及降维可视化。其中蓝色圆点为“抢占制胜点”标签下的习题表征，0 为它们的中心点，而绿色十字星为“多位数除法的实际应用”标签下的习题表征，1 为它们的中心点。



a) SBERT-QA 得到的习题表征　　b) SBERT-QA-CNN 得到的习题表征

a) Exercise Representation of SBERT-QA　　b) Exercise Representation of SBERT-QA-CNN



c) TBERT-QA 得到的习题表征　　d) TBERT-QA-CNN 得到的习题表征

c) Exercise Representation of TBERT-QA　　d) Exercise Representation of TBERT-QA-CNN

图 5-3 不同模型得到的习题表征

Table 5-3 Exercise Representations Obtained by Different Models

在图 5-3 中，SBERT-QA、SBERT-QA-CNN、TBERT-QA 和 TBERT-QA-CNN 均为本文本章所提出的模型。经过观察可以得到以下结论：

(1) 与图 4-6 中的 SBERT-CLS 相比，SBERT-QA 得到的习题表征中，每类习题的表征更靠近相应类别的中心点，这说明习题的问题文本和解答文本中均包含习题的重要信息，能够提升模型的文本表征能力。

(2) CNN 池化在 SBERT-QA-CNN 上的作用并不明显，但相比于 TBERT-QA，TBERT-QA-CNN 得到的习题表征质量有很大的提升，说明 CNN 池化能够在一定情况下提升一定的模型表征能力。

综上所述，能够发现这些模型得到的习题表征有一定的效果提升，尤其是 SBERT-QA、SBERT-QA-CNN 和 TBERT-QA-CNN，这说明本章中所提出的基于习题问题与解答的融合模型能进一步提升习题文本表征能力，进而在习题任务中能够得到良好的效果提升。

5.5 本章总结

本章提出了基于习题问题与解答的融合模型，对其设计思路、模型结构和模型构建过程分别进行了介绍，并在简单介绍了数据集构建过程之后，详细对模型结果进行了分析。最终，本文提出的融合习题问题和解答文本的 TBERT-QA 模型 MAP 分数为 0.65，比仅考虑习题问题文本相似度匹配的 TBERT-CLS 模型高出 0.04，（即相对提升了 6.6%），且比表现最好的基线模型 VSM 的高出 0.27（即相对提升了 71.1%）。进一步的，加入 CNN 池化方式的 TBERT-QA-CNN 在测试集上的 MAP 分数比 TBERT-QA 高出 0.01（即相对提升了 1.5%），且比表现最好的基线模型 VSM 的高出 0.28（即相对提升了 73.7%）。

6 性能对比与应用分析

本章介绍本文提出的习题相似度模型在寻找相似习题任务中的表现，进行所有模型表现的对比分析，并通过具体的习题案例说明本文所提出各类模型的应用效果。

6.1 模型整体表现

本节将模型在测试集上的结果展示在表 6-1 中，该分数是在候选集包含 5 道相似题和 200 道不相似题的寻找相似题任务中的 MAP 分数。选取该分数的原因是：候选集中相似题和不相似题的数量比例为 1:20，模型分数能够有一定代表性。

表 6-1 各模型在测试集上寻找相似习题任务中的效果对比

Table 6-1 Model Performance Comparison in Finding Similar Exercises Task on the Testing Sets

输入	模型	MAP 分数
问题、解答	TBERT-QA-CNN	0.66
问题、解答	TBERT-QA	0.65
问题	TBERT-CNN	0.61
问题	TBERT-CLS	0.61
问题、解答	SBERT-QA	0.60
问题、解答	SBERT-QA-CNN	0.55
问题	SBERT-CNN	0.51
解答	TBERT-CNN	0.41
问题	VSM[2]	0.38
解答	TBERT-CLS	0.37
问题	SBERT-CLS	0.35
问题	TLSTM[16]	0.35
问题	SLSTM	0.35
解答	SBERT-CNN	0.34
解答	SBERT-CLS	0.26
解答	TLSTM[16]	0.22
解答	SLSTM	0.22
问题	BERT[10]	0.02

通过分析表 6-1 能够得到以下结论：

(1) 从输入类型分析：TBERT-QA 比输入问题/解答的 TBERT-CLS 高出

0.04/0.28，SBERT-QA 比输入问题/解答的 SBERT-CLS 高出 0.25/0.34。结果说明融合习题问题和解答信息能够提高模型的文本表征能力。

(2) 从表征模型分析：输入为问题/解答时 TBERT-CLS 比 TLSTM 高出 0.26/0.15，输入为问题/解答时 SBERT-CLS 比 SLSTM 高出 0.0/0.04。结果说明 BERT 模型的结构更适合对习题文本这种逻辑复杂的文本进行文本表征。

(3) 从模型架构分析：在输入为问题时 SBERT-CLS 比 BERT 高出 0.33，输入为问题/解答时 TBERT-CLS 比 SBERT-CLS 高出 0.26/0.11，而 TBERT-QA 比 SBERT-QA 高出 0.05。结果说明基于 Siamese 架构的模型能够提升模型的文本表征能力，而基于 Triplet 架构的模型文本表征能力能对其进行进一步提升。

(4) 从池化方式分析：TBERT-QA-CNN 比 TBERT-QA 高出 0.01，SBERT-QA-CNN 比 SBERT-QA 低了 0.05，输入为问题/解答时 TBERT-CNN 比 TBERT-CLS 高出 0.0/0.04，输入为问题/解答时 SBERT-CNN 比 SBERT-CLS 高出 0.16/0.08。结果说明 CNN 的池化方式能够提升一定的模型训练效果，且在不融合习题问题与解答的模型中提升更明显。

(5) 最后进行与基线模型的对比：在表 6-1 中，表现最好的基线模型为 VSM 模型，它的 MAP 分数为 0.38。本文所提出的模型中 TBERT-QA-CNN 的 MAP 为 0.66，比 VSM 高出 0.28（即相对提升了 73.7%）。

6.2 具体案例分析

为了形象说明模型能力，本文在标签为“抢占制胜点”下的测试集中抽取一道习题进行相似习题推荐，根据各习题相似度模型的表现进行具体分析。表 6-2 为所挑选习题的具体信息，表 6-3 为各模型的输入和在该习题上的 MAP 分数（每个模型推荐前 5 道习题的具体情况已附于附录 A 中）。

首先通过表 6-2 分析原题：原题的问题文本长度为 76，解答文本长度为 371。在问题文本中，“轮流”、“获胜”、“先”、“后”等字眼说明该问题是一个典型的对策问题，需要学生制定正确的策略取得胜利；而解答文本利用了“倒推法”（也是这类问题常用的策略）进行解决，通过从最后一次的必胜条件出发，倒推至起始点的必胜条件，其中，“倒推”、“获胜”等字眼与“抢占制胜点”联系较为紧密。为该题推荐相似习题的难点在于：题目中“数”、“自然数”等一类描述过多，如果不进行文本的深层语义提取，很容易与“数”有关的概念混淆。

接着，通过表 6-3 和附录 A 中各模型在该习题上推荐相似习题的具体表现，分析模型表现并得到以下结果：

表 6-2 原题信息

Table 6-2 Exercise Information

原题	问题/解答	习题文本	习题标签
--	问题	甲、乙二人轮流报数，必须报$$1$$~$$6$$中的一个自然数，把两人报出的数依次加起来，谁报数后加起来的数是$$2000$$，谁就获胜．如果甲要取胜，是先报还是后报？报几？以后怎样报？	抢占制胜点
	解答	采用倒推法(倒推法是解决这类问题一种常用的数学方法)．由于每次报的数是$$1$$~$$6$$的自然数，$$2000-1=1999$$，$$2000-6=1994$$，甲要获胜，必须使乙最后一次报数加起来的和的范围是$$1994$$~$$1999$$，由于$$1994-1=1993$$(或$$1999-6=1993$$)，因此，甲倒数第二次报数后加起来的和必须是$$1993$$．同样，由于$$1993-1=1992$$，$$1993-6=1987$$，所以要使乙倒数第二次报数后加起来的和的范围是$$1987$$~$$1992$$，甲倒数第三次报数后加起来的和必须是$$1986$$．同样，由于$$1986-1=1985$$，$$1986-6=1980$$，所以要使乙倒数第三次报数后加起来的和的范围是$$1980$$~$$1985$$，甲倒数第四次报数后加起来的和必须是$$1979$$，…．把甲报完数后加起来必须得到的和从后往前进行排列：$$2000$$、$$1993$$、$$1986$$、$$1979$$、…．观察这一数列，发现这是一等差数列，且公差为$$7$$，这些数被$$7$$除都余$$5$$．因此这一数列的最后三项为：$$19$$、$$12$$、$$5$$．所以甲要获胜，必须先报，报$$5$$．因为$$12-5=7$$，所以以后乙报几，甲就报$$7$$减几，例如乙报$$3$$，甲就接着报$$4\left(=7-3 \right)$$．所以甲要获胜必须先报，甲先报$$5$$；以后，乙报几甲就接着报$$7$$减几．	

表 6-3 各模型在 6-2 习题上寻找相似习题任务中的效果对比

Table 6-3 Model Performance Comparison in Finding Similar Exercises Task on 6-3 Exercise

输入	模型	MAP 分数
问题、解答	TBERT-QA-CNN	1.0
问题、解答	SBERT-QA	0.96
问题	TBERT-CLS	0.91
问题、解答	SBERT-QA-CNN	0.87
问题	SBERT-CLS	0.87
问题、解答	TBERT-QA	0.84
解答	TBERT-CNN	0.70
问题	SBERT-CNN	0.54
问题	TLSTM[16]	0.46
解答	SBERT-CNN	0.45
解答	SBERT-CLS	0.20
问题	SLSTM	0.20
解答	TBERT-CLS	0.13
问题	TBERT-CNN	0.0
解答	TLSTM[16]	0.0
解答	SLSTM	0.0
问题	BERT[10]	0.0
问题	VSM[2]	0.0

一、四个基线模型的表现及原因分析

(1) VSM 模型的 MAP 分数为 0，推荐的前 5 道最相似习题分别属于“加权平均数”、“连续自然数三角形数表之已知位置求数”、“数的整除”、“连续自然数三角形数表之已知位置求数”、“求算式整数部分”这 5 个标签。通过分析推荐题目可以看出，VSM 捕捉了原题中与“数”有关的信息，但却没有捕捉到“获胜”等更重要的信息。这是由于 VSM 依赖于文本中的词频等信息，无法进行文本更深层次的句义捕捉。

(2) BERT 模型的 MAP 分数为 0，推荐的前 5 道最相似习题分别属于“三视图求表面积与体积综合”、“单项式乘多项式”、“定积分的几何意义”、“三角形格点多边形直线的倾斜角与斜率”、“直线的位置关系、直线的方程”这 5 个标签。可以看出 BERT 在该习题上没有捕捉到关键信息，近似于“随机”推荐相似习题。这一结果可能是由于在通过 BERT 能够用一个超平面将两类习题表征分开，但没有着重于增加不相似习题表征之间的距离，同时减少相似习题表征之间的距离。当习题表征效果不理想时，在习题任务中的效果也会大打折扣。（本文提出的 SBERT、TBERT 等均是基于 BERT 的模型改进）。

(3) 输入为问题/解答时 SLSTM 的 MAP 分数为 0.20/0.0。SLSTM 在输入为问题时，推荐的前 5 道最相似习题中有 2 道属于“抢占制胜点”这一概念，其余 3 道属于“复合数字的整除特征应用”和“计数求概率”这 2 个标签；而在输入为解答时，推荐的前 5 道最相似习题分别属于“放缩与估算”、“加权平均数”、“连续自然数三角形数表之已知位置求数”、“调运中的统筹”、“位值原理的完全拆分”这 5 个标签。通过分析推荐题目可以看出，SLSTM 在输入问题文本时能够提升捕获“获胜”等与“抢占制胜点”标签相关的信息，这是由于 SLSTM 不是像 VSM 一样简单地根据词频信息计算文本相似度，而是利用 LSTM 对文本进行深度表征提取出了文本的隐藏语义。但在输入解答文本时仅能捕获与“数”有关的信息，这说明利用 LSTM 的文本表征能力提取习题解答文本隐藏含义还有一定难度。

(4) 输入为问题/解答时 TLSTM 的 MAP 分数为 0.46/0.0，这一分数相比于 SLSTM 有一定程度的提升。TLSTM 在输入为问题文本时，推荐的前 5 道最相似习题中第 1 道属于“抢占制胜点”，其余 4 道分别属于“分数的通分”、“两端植树问题变型”、“数的整除”、“容斥原理”这 4 类标签；而在输入为解答文本时，推荐的前 5 道最相似习题分别属于“计数求概率”、“位值原理的完全拆分”、“智巧趣题”、“分数的通分”、“组合问题”这 5 类标签。通过分析推荐题目同样可以看出，TLSTM 在输入问题文本时能够提升捕获“获胜”等与“抢占制胜点”标签相关的信息。同样能够分析得到：TLSTM 利用 LSTM 模型对文本进行深度表征能够提取出习题问题文本的隐藏语义，但较难提取出习题解答文本中的隐藏含

义。

二、模型架构比较

(1) 输入为问题时，SBERT-CLS 和 TBERT-CLS 的 MAP 分数均比 BERT 高，且比较推荐习题内容可以发现，SBERT-CLS 和 TBERT-CLS 能够捕获习题的关键信息并进行相应相似习题推荐，不再像 BERT 一样“随机”推荐，这说明本文基于 Siamese 架构和基于 Triplet 架构进行的改进有效。具体的，在输入为问题时，TBERT-CLS 的 MAP 分数为 0.91，推荐的前 5 道最相似习题中有 4 道属于“抢占制胜点”标签，仅有 1 道属于“复合数字的整除特征应用”标签；SBERT-CLS 的 MAP 分数为 0.87，推荐的前 5 道最相似习题中有 3 道属于“抢占制胜点”标签，剩下 2 道分别属于“计数求概率”和“组合问题”标签。尽管在少数情况下会发生基于 Triplet 架构的模型效果不如基于 Siamese 架构的模型效果的现象（如 TBERT-CNN 和 SBERT-CNN，这说明某些模型在该题上的作用不够显著，没有足够的能力将该题与其他题目的问题文本区分开），大部分情况下基于 Triplet 架构的模型相对于基于 Siamese 架构的模型效果有一定提升。因此，在输入为问题时，基于 Triplet 架构的模型能够更有效的将属于“抢占制胜点”标签下的习题从与“数”有关的习题类别中剥离出来，这符合本文的预期。

(2) 输入为解答时，TBERT-CNN 的 MAP 分数为 0.70，推荐的前 5 道最相似习题中有 3 道属于“抢占制胜点”，其余 2 道分别属于“描述随机现象发生的可能性大小”和“求算式整数部分”；SBERT-CNN 的 MAP 分数为 0.45 推荐的前 5 道最相似习题中有 3 道属于“抢占制胜点”，其余 2 道分别属于“复合数字的整除特征应用”和“智巧趣题”。但可以发现，TBERT-CLS 和 SBERT-CLS、TLSTM 和 SLSTM 的表现并不是预期的那样。通过分析模型推荐习题可以猜测，尽管解答中含有一定的解题逻辑，但其复杂程度会影响模型进行相似度计算的准确性，如 TLSTM 模型推荐的排序为 Top1 的习题解答中就包含“反向考虑”的字眼，这与“倒推法”是一类性质，但实际上习题解答的公式中计算的是概率相关的信息，该习题的所属类别也是“计数求概率”。因此，在输入为解答时，基于 Triplet 架构的模型能够在一定程度上将属于“抢占制胜点”标签下的习题从与其他习题类别中剥离出来，但由于解答文本的复杂性，无法最大限度地发挥 Triplet 架构的作用。

(3) 输入为问题和解答时，TBERT-QA-CNN 的 MAP 分数为 1.0，推荐的前 5 道最相似习题均属于“抢占制胜点”标签；SBERT-QA-CNN 的 MAP 分数为 0.87，推荐的前 5 道最相似习题中 3 道属于“抢占制胜点”标签，其余 2 道分别属于“多位数除法的实际应用”和“两端植树问题变型”。而 TBERT-QA 的 MAP 分数为 0.84，推荐的前 5 道最相似习题中仅有 1 道属于“2-1-0 赛制”，SBERT-QA 的 MAP 分数为 0.96，仅有 1 道属于“智巧趣题”。可以看出，TBERT-QA 相对于 SBERT-

QA 在该题上效果没有提升，但它们都将该习题与“数”有关的习题分开了，错误推荐的习题尽管不属于“抢占制胜点”标签，但在某些方面与原题更为接近（分析 TBERT-QA 推荐的属于“2-1-0”赛制标签的问题，可以看出 TBERT-QA 能够发现“胜负”等字眼较为重要）。而 TBERT-QA-CNN 的表现说明，Triplet 架构与 CNN 池化叠加能够更有效地将属于“抢占制胜点”标签下的习题从与其他习题类别中剥离出来。

三、模型输入分析

(1) 问题 vs 解答：在大多数情况下，同样的模型输入为问题时的效果比输入为解答时的效果好（如 TBERT-CLS、SBERT-CLS、SBERT-CNN、TLSTM、SLSTM），能够得出如上面分析中提到的结论，尽管解答中含有一定的解题逻辑，但其复杂程度会影响模型进行相似度计算的准确性。

(2) 问题&解答 vs 问题：大多数情况下，同样的架构下输入为问题和解答时的效果比输入仅为问题时的效果好（如 TBERT-QA-CNN 和 TBERT-CNN、SBERT-QA-CNN 和 SBERT-CNN、SBERT-QA 和 SBERT-CLS）。如 SBERT-QA 的 MAP 分数为 0.96，推荐的前 5 道最相似习题中 4 道属于“抢占制胜点”标签，1 道属于“智巧趣题”标签；SBERT-CLS（输入为问题）的 MAP 分数为 0.87，推荐的前 5 道最相似习题中 3 道属于“抢占制胜点”标签，其余 2 道分别属于“计数求概率”和“组合问题”标签。以 SBERT-CLS 推荐的 Top4 习题为例，可以发现该题问题中包含有“获胜”等重要信息，因此在输入为问题时被 SBERT-CLS 判别为相似习题，但事实上它属于“计数求概率”的标签，这在它的解答中也可以看出（解答中包含有“可能”、“可能性”等关键信息）。当将问题和解答输入 SBERT-QA 时，模型能够捕获出这样的差异，因此能够将它们分开。可以得到结论，与输入为问题时相比，输入为问题和解答能够使模型更充分的捕获习题关键信息，进而提升效果。

四、池化类型分析

观察可以看出，CNN 池化能够提升一定的模型效果（尤其是在原模型效果较差时提升较高）。如输入为解答时，TBERT-CNN 的 MAP 分数为 0.70，推荐的前 5 道最相似习题中 3 道属于“抢占制胜点”，其余 2 道分别属于“描述随机现象发生的可能性大小”和“求算式整数部分”；TBERT-CLS 的 MAP 分数为 0.13，推荐的前 5 道最相似习题中 2 道属于“抢占制胜点”，其余 3 道分别属于“容斥原理”、“多位数除法的实际应用”、“计数求概率”。可以看出，相比于 TBERT-CLS，TBERT-CNN 能够提升捕获解答中关键信息的能力，这说明 CNN 池化能够提升一定的模型效果。

6.3 本章总结

本章通过分析习题推荐的具体案例对本论文中已提出的习题相似度模型的效果和原因进行分析，可以得出结论：本论文中提出的模型架构的改进、融合习题问题和解答的改进，以及 CNN 池化操作的应用均能提升一定的模型效果，在习题任务中获得显著的改进。

7 总结和展望

7.1 总结

本文基于真实在线教育系统的中文数学习题数据集进行研究。首先对其进行统计分析，并根据统计分析结果进行习题数据集预处理和实验参数设置。接着，对其进行文本分析，确认中文数学习题任务的困难所在，也为模型设计奠定了基础。在定义寻找相似习题任务的数据集构建过程和算法评估准则后，对现有模型（VSM）在该任务中的表现进行模型构建和实验分析，有如下观察：模型在习题任务中的表现与其习题文本表征能力密切相关。

本文基于最新的自然语言处理模型 BERT，从三个方面进行了模型设计和改进：

(1) 为了将习题映射到相似习题距离较近、不相似习题距离较远的习题表征空间中，利用孪生神经网络架构能够将输入映射到用空间距离反映“语义”距离的目标空间中的原理，设计出两个习题相似度模型：SBERT-CLS 和 TBERT-CLS。其中的 TBERT-CLS 在寻找相似习题任务中，当输入为问题文本时，在测试集上的 MAP 分数为 0.61，比表现最好的基线模型 VSM 高出 0.23（即相对提升了 60.5%）。特别是：

a) 输入为问题文本的模型比输入为解答文本的模型往往表现效果更好，这说明习题的解答文本构成较为复杂，不能完全代替问题文本进行模型的训练。

b) 同样的框架和输入下，BERT 对应的模型表现总是优于利用 LSTM 进行表征的模型。如输入为问题时，TBERT-CLS 在测试集上的 MAP 分数为 0.61，比同样基于 Triplet 架构的模型 TLSTM 高出 0.26。这说明 BERT 的结构更适合处理习题文本这种逻辑复杂的文本。

c) 基于 Siamese 架构的模型可以提升模型效果，如输入为问题时，SBERT-CLS 在测试集上的 MAP 分数为 0.35，比 BERT 高出 0.33；而基于 Triplet 架构的模型比基于 Siamese 架构的模型在习题任务上效果有一定提升，如输入为问题时，TBERT-CLS 在测试集上的 MAP 分数为 0.61，比 SBERT-CLS 高出 0.26。这说明基于 Siamese 架构的模型能够提升模型的文本表征能力，而基于 Triplet 架构的模型文本表征能力能对其进行进一步提升。

(2) 为了更好地捕获习题问题和解答文本之间的关系，同时对习题的问题与解答进行表征，并将它们融合利用，设计融合习题问题和解答的模型。该模型不仅能够支持两道习题的问题文本-问题文本的匹配和解答文本-解答文本的匹配，还支持

一道习题的问题文本和另一道习题的解答文本之间的相似度匹配。设计出两个习题相似度模型：SBERT-QA 和 TBERT-QA。其中的 TBERT-QA 在测试集上的 MAP 分数为 0.65，比仅考虑习题问题文本相似度匹配的 TBERT-CLS 高出 0.04（即相对提升了 6.6%），且比表现最好的基线模型 VSM 的高出 0.27（即相对提升了 71.1%）。特别是：

a) 对习题的问题文本和解答文本进行融合的模型往往比只输入问题文本和解答文本的模型表现效果更好，如 TBERT-QA 在测试集上的 MAP 分数比仅考虑习题问题文本相似度匹配的 TBERT-CLS 提升了 0.04。这说明习题的问题文本和解答文本中均包含习题的重要信息，充分利用它们能够提升模型的文本表征能力。

b) 在融合习题问题和解答的模型中，基于 Triplet 架构的模型相比于基于 Siamese 架构的模型在习题任务上效果也有一定提升，如 TBERT-QA 在测试集上的 MAP 分数为 0.65，比 SBERT-QA 高出 0.05。这说明基于 Triplet 架构的模型比基于 Siamese 架构的模型文本表征能力更强。

(3) 为了进一步获取综合全面的习题文本表征，利用 Text-CNN 进行池化操作提升模型表征能力。设计出四个习题相似度模型：SBERT-CNN、TBERT-CNN、SBERT-QA-CNN 和 TBERT-QA-CNN。其中 TBERT-QA-CNN 在测试集上的 MAP 分数比 TBERT-QA 高出 0.01（即相对提升了 1.5%），且比表现最好的基线模型 VSM 的高出 0.28（即相对提升了 73.7%）。这说明 CNN 池化能够提升一定的习题文本表征效果。

为了证实上述得到的结论，本文还对各模型进行习题表征的可视化分析，通过模型在习题表征上的具体表现加强结论的真实性。最后，为了形象说明模型能力，本文抽取出一道习题进行相似习题推荐，根据各模型的表现进行具体分析，对比各模型的效果并挖掘上述模型改进在其中起到的作用。

7.2 展望

随着在线教育的广泛应用和不断发展，海量习题为教育系统的开发造成一定难度。目前，自然语言处理技术的应用极大提高了工业界处理文本内容的效率，这对提高教育效率、加快教育改革有着重要的作用。未来，在线教育平台将不仅重视数据处理的效率，还将进一步重视个性化教育。把自然语言处理技术、建立学生模型、知识追踪、强化学习等技术应用于个性化题目推荐将成为未来的必然发展趋势。

参考文献

[1] http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm.

[2] Salton G, Buckley C. Term-Weighting Approaches in Automatic Text Retrieval[J]. Information Processing & Management, 1988, 24(5): 513-523.

[3] Yu J, Li D, Hou J, et al. Similarity Measure of Test Questions Based on Ontology and VSM[J]. The Open Automation and Control Systems Journal, 2014, 6(1): 262-267.

[4] Bengio Y, Ducharme R, Vincent P, et al. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research, 2003, 3(6): 1137-1155.

[5] Mikolov T, Chen K, Corrado G S, et al. Efficient Estimation of Word Representations in Vector Space[C]//International Conference on Learning Representations, 2013.

[6] Le Q, Mikolov T. Distributed Representations of Sentences and Documents[C]//International Conference on Machine Learning. 2014: 1188-1196.

[7] Peters M E, Neumann M, Iyyer M, et al. Deep Contextualized Word Representations [C]//North American Chapter of the Association for Computational Linguistics, 2018: 2227-2237.

[8] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. 2017: 5998-6008.

[9] Radford A, Narasimhan K, Salimans T, et al. Improving Language Understanding by Generative Pre-Training[J]. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf, 2018.

[10] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[11] John V. A Survey of Neural Network Techniques for Feature Extraction from Text[J]. arXiv preprint arXiv:1704.08531, 2017.

[12] Chopra S, Hadsell R, LeCun Y. Learning a Similarity Metric Discriminatively, With Application to Face Verification[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, 2005, 1: 539-546.

[13] Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). IEEE, 2006, 2: 1735-1742.

[14] Hoffer E, Ailon N. Deep Metric Learning Using Triplet Network[C]//International Workshop on Similarity-Based Pattern Recognition. Springer, Cham, 2015: 84-92.

[15] Mueller J, Thyagarajan A. Siamese Recurrent Architectures for Learning Sentence Similarity[C]//Thirtieth AAAI Conference on Artificial Intelligence. 2016.

[16] Liu Q, Huang Z, Huang Z, et al. Finding Similar Exercises in Online Education Systems[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 1821-1830.

[17] Yin Y, Liu Q, Huang Z, et al. QuesNet: A Unified Representation for Heterogeneous Test Questions[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 1328-1336.

[18] John R J L, Passonneau R J, McTavish T S. Semantic Similarity Graphs of Mathematics Word Problems: Can Terminology Detection Help?[J]. International Educational Data Mining Society, 2015..

[19] Pelánek R, Effenberger T, Vaněk M, et al. Measuring Item Similarity in Introductory Programming: Python and Robot Programming Case Studies[J]. arXiv preprint arXiv:1806.03240, 2018.

[20] Su D, Yekkehkhany A, Lu Y, et al. Prob2Vec: Mathematical Semantic Embedding for Problem Retrieval in Adaptive Tutoring[J]. arXiv preprint arXiv:2003.10838, 2020.

[21] Stehle S, Spinath B, Kadmon M. Measuring Teaching Effectiveness: Correspondence between Students' Evaluations of Teaching and Different Measures of Student Learning[J]. Research in Higher Education, 2012, 53(8): 888-904.

[22] Chang H S, Hsu H J, Chen K T. Modeling Exercise Relationships in E-Learning: A Unified Approach[C]//EDM. 2015: 532-535.

[23] Rihák J, Pelánek R. Measuring Similarity of Educational Items Using Data on Learners' Performance[J]. International Educational Data Mining Society, 2017..

[24] Huang Z, Liu Q, Chen E, et al. Question Difficulty Prediction for READING Problems in Standard Tests[C]//Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[25] Jeon J, Croft W B, Lee J H. Finding Similar Questions in Large Question and Answer Archives[C]//Proceedings of the 14th ACM International Conference on Information and Knowledge Management. 2005: 84-90.

[26] Williams A E, Aguilarroca N, Tsai M, et al. Assessment of Learning Gains Associated with Independent Exam Analysis in Introductory Biology[J]. CBE- Life Sciences Education, 2011,10(4):346-356.

[27] https://zh.wikipedia.org/zh-hans/循环神经网络[EB/OL].

[28] Elman J L. Finding Structure in Time[J]. Cognitive Science, 1990,14(2):179-211.

[29] Lipton Z C, Berkowitz J, Elkan C. A Critical Review of Recurrent Neural Networks for Sequence Learning[J]. arXiv preprint arXiv:1506.00019, 2015.

[30] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 1997,9(8):1735-1780.

[31] Bengio Y, Simard P, Frasconi P. Learning Long-Term Dependencies with Gradient Descent is Difficult[J]. IEEE Transactions on Neural Networks, 1994, 5(2): 157-166.

[32] Jozefowicz R, Zaremba W, Sutskever I. An Empirical Exploration of Recurrent Network Architectures[C]//International Conference on Machine Learning. 2015: 2342-2350..

[33] Miwa M, Bansal M. End-to-End Relation Extraction Using Lstms on Sequences and Tree Structures[J]. arXiv preprint arXiv:1601.00770, 2016.

[34] Sutskever I, Vinyals O, Le Q V. Sequence to Sequence Learning with Neural Networks[C]//Advances in Neural Information Processing Systems. 2014: 3104-3112.

[35] Tang D, Qin B, Liu T. Document Modeling with Gated Recurrent Neural Network for Sentiment Classification[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 1422-1432..

[36] Kim Y. Convolutional Neural Networks for Sentence Classification[J]. arXiv preprint arXiv:1408.5882, 2014.

[37] Yu F, Koltun V. Multi-scale Context Aggregation by Dilated Convolutions[J]. arXiv preprint arXiv:1511.07122, 2015.

[38] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[39] Nguyen T H, Grishman R. Relation extraction: Perspective from Convolutional Neural Networks[C]//Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. 2015: 39-48.

[40] Yin W, Schütze H. Convolutional Neural Network for Paraphrase Identification[C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015: 901-911.

[41] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. arXiv preprint arXiv:1409.0473, 2014..

[42] Dehghani M, Gouws S, Vinyals O, et al. Universal transformers[J]. arXiv preprint arXiv:1807.03819, 2018.

[43] Dai Z, Yang Z, Yang Y, et al. Transformer-Xl: Attentive Language Models beyond a Fixed-Length Context[J]. arXiv preprint arXiv:1901.02860, 2019.

[44] Salton G, Wong A, Yang C S. A Vector Space Model for Automatic Indexing[J]. Communications of the ACM, 1975, 18(11): 613-620..

[45] Jing Y, Li D, Hou J, et al. Similarity Measure of Test Questions Based on Ontology and VSM[J]. Open Automation & Control Systems Journal, 2014,6(1):262-267.

[46] Mikolov T, Karafiát M, Burget L, et al. Recurrent Neural Network based Language Model[C]//Eleventh Annual Conference of the International Speech Communication Association. 2010.

[47] Pennington J, Socher R, Manning C D. Glove: Global Vectors for Word Representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1532-1543.

[48] Joulin A, Grave E, Bojanowski P, et al. Bag of Tricks for Efficient Text Classification[J]. arXiv preprint arXiv:1607.01759, 2016.

[49] Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized Autoregressive Pretraining for Language Understanding[C]//Advances in Neural Information Processing Systems. 2019: 5754-5764.

[50] Lan Z, Chen M, Goodman S, et al. Albert: A Lite Bert for Self-Supervised Learning of Language Representations[J]. arXiv preprint arXiv:1909.11942, 2019.

[51] Liu Y, Ott M, Goyal N, et al. Roberta: A Robustly Optimized Bert Pretraining Approach[J]. arXiv preprint arXiv:1907.11692, 2019.

[52] Kenter T, Borisov A, De Rijke M. Siamese cbow: Optimizing Word Embeddings for Sentence Representations[J]. arXiv preprint arXiv:1606.04640, 2016..

附录 A

模型（输入类型）及 MAP 分数	相似题排序	习题文本	习题标签
TBERT-QA-CNN（问题、解答） MAP：1.0	TOP1	问题：$$1997$$个空格排成一排，第一格中放有一枚棋子，现有两人做游戏，轮流移动棋子，每人每次可前移$$1$$格、$$2$$格、$$3$$格或$$4$$格；谁选移到最后一格，谁失败．问怎样的移法才能确保获胜？	抢占制胜点
		解答：便于方便，可以把这$$1997$$个空格编成$$1$$号、$$2$$号、$$\cdots\cdots$$、$$1997$$号．要想取胜，应使棋子依次移到号码被$$5$$除余$$1$$的空格处．即$$16$$、$$11$$、$$16$$$$\cdots\cdots$$$$1991$$、$$1996$$	
	TOP2	问题：一个箱子内装有$$2016$$颗棋子，两人轮流在其中取棋子，规定每人每次只能提取$$1$$、$$3$$、$$7$$颗棋子，不得不取，也不得多取，取到最后棋子的人取胜．为了确保取胜，你是愿意先手，还是愿意后手？说出你的选择答案和必胜的策略．	抢占制胜点
		解答：根据规则，虽然不能控制对方每次提取棋子得数量，但可以通过控制自己的数量保证每一轮双方提取棋子总和为$$4$$或$$8$$（对方取$$1$$颗，我方取$$7$$颗或$$3$$颗；对方取$$3$$颗，我方取$$1$$颗；对方取$$7$$颗，我方取$$1$$）．由于$$2016$$是$$8$$的倍数，选择后手提取，可以保证每次自己提取之后，剩余数量都是$$4$$的倍数，直至最后剩下$$8$$颗或$$4$$颗．在$$4$$颗的情况，对方只能取$$3$$或$$1$$颗，我方相应取$$1$$或$$3$$颗，取胜；在$$8$$颗的情况，对方若取$$1$$或$$7$$颗，我方相应取$$7$$或$$1$$颗，取胜；对方若取$$3$$颗，我方取$$1$$颗，转化为$$4$$颗的情况．	
	TOP3	问题：桌子上放着$$37$$根火柴，聪明昊、神奇涛二人轮流每次取走$$1\sim5$$根．规定谁取走最后一根火柴谁获胜．如果双方都采用最佳方法，聪明昊先取，神奇涛后取，你知道会胜吗．	抢占制胜点
		解答：由$$37\div(1+5)=6\cdots\cdots1$$知聪明昊会胜．	
	TOP4	问题：今有两堆火柴，一堆$$6$$根，另一堆$$8$$根．两人轮流在其中任一堆中拿取，取的根数不限，但不能不取．规定取得最后一根者为赢．问：先取者有何策略能获胜？ 如果这两堆火柴数变成一堆$$12$$根，一堆$$12$$根，先取者还是后取者有必胜策略呢？	抢占制胜点
		解答：先取者在$$8$$根一堆火柴中取$$2$$根火柴，使得取后剩下两堆的火柴数相同．以后无论对手在某一堆取几根火柴，你只需在另一堆也取同样多根火柴．只要对手有火柴可取，你也有火柴可取，也就是说，最后一根火柴总会被你拿到．这样先取者总可获胜．如果都变成$$12$$根，后取者只需要在另一堆取跟对手相同多根火柴，则后取者有必胜策略．	
	TOP5	问题：把一枚棋子放在图中左下角的方格内，甲、乙两人玩这样一个游戏：双方轮流移动棋子，只能向上、向右或者向右上方沿$$45{}^\circ $$角移动，一次可以移动任意多格．谁把棋子移到了右上角的方格中即为赢，试问：如果甲先走，是否有必胜的策略，为什么（　　）？有没有不确定	抢占制胜点
		解答：从右上角开始分析哪些位置必胜的，哪些位置是必败的，结果如图所示．因此甲第一步必然走到“√”上，抢占制胜点，故甲必胜．故选$$\text{A}$$.	
SBERT-QA（问题、解答） MAP：0.96	TOP1	问题：桌子上放着$$37$$根火柴，聪明昊、神奇涛二人轮流每次取走$$1\sim5$$根．规定谁取走最后一根火柴谁获胜．如果双方都采用最佳方法，聪明昊先取，神奇涛后取，你知道会胜吗．	抢占制胜点
		解答：由$$37\div(1+5)=6\cdots\cdots1$$知聪明昊会胜．	
	TOP2	问题：$$1997$$个空格排成一排，第一格中放有一枚棋子，现有两人做游戏，轮流移动棋子，每人每次可前移$$1$$格、$$2$$格、$$3$$格或$$4$$格；谁选移到最后一格，谁失败．问怎样的移法才能确保获胜？	抢占制胜点
		解答：便于方便，可以把这$$1997$$个空格编成$$1$$号、$$2$$号、	

		$$\cdots\cdots$$、$$1997$$号．要想取胜，应使棋子依次移到号码被$$5$$除余$$1$$的空格处．即$$16$$、$$11$$、$$16$$$$\cdots\cdots$$$$1991$$、$$1996$$.	
	TOP3	问题：一个箱子内装有$$2016$$颗棋子，两人轮流在其中取棋子，规定每人每次只能提取$$1$$、$$3$$、$$7$$颗棋子，不得不取，也不得多取，取到最后棋子的人取胜．为了确保取胜，你是愿意先手，还是愿意后手？说出你的选择答案和必胜的策略．	抢占制胜点
		解答：根据规则，虽然不能控制对方每次提取棋子得数量，但可以通过控制自己的数量保证每一轮双方提取棋子总和为$$4$$或$$8$$（对方取$$1$$颗，我方取$$7$$颗或$$3$$颗；对方取$$3$$颗，我方取$$1$$颗；对方取$$7$$颗，我方取$$1$$）．由于$$2016$$是$$8$$的倍数，选择后手提取，可以保证每次自己提取之后，剩余数量都是$$4$$的倍数，直至最后剩下$$8$$颗或$$4$$颗．在$$4$$颗的情况，对方只能取$$3$$或$$1$$颗，我方相应取$$1$$或$$3$$颗，取胜；在$$8$$颗的情况，对方若取$$1$$或$$7$$颗，我方相应取$$7$$或$$1$$颗，取胜；对方若取$$3$$颗，我方取$$1$$颗，转化为$$4$$颗的情况．	
	TOP4	问题：把一枚棋子放在图中左下角的方格内，甲、乙两人玩这样一个游戏：双方轮流移动棋子，只能向上、向右或者向右上方沿$$45{}^\circ $$角移动，一次可以移动任意多格．谁把棋子移到了右上角的方格中即为赢，试问：如果甲先走，是否有必胜的策略，为什么（ ）？有没有不确定	抢占制胜点
		解答：从右上角开始分析哪些位置必胜的，哪些位置是必败的，结果如图所示．因此甲第一步必然走到“√”上，抢占制胜点，故甲必胜．故选$$\text{A}$$.	
	TOP5	问题：如果允许砝码放在天平两端，那么能称量出$$1$$～$$121$$克中任何整数克重的物品，至少需要个砝码？	智巧趣题
		解答：$$5$$	
TBERT-CLS（问题） MAP：0.91	TOP1	问题：一个箱子内装有$$2016$$颗棋子，两人轮流在其中取棋子，规定每人每次只能提取$$1$$、$$3$$、$$7$$颗棋子，不得不取，也不得多取，取到最后棋子的人取胜．为了确保取胜，你是愿意先手，还是愿意后手？说出你的选择答案和必胜的策略．	
		解答：根据规则，虽然不能控制对方每次提取棋子得数量，但可以通过控制自己的数量保证每一轮双方提取棋子总和为$$4$$或$$8$$（对方取$$1$$颗，我方取$$7$$颗或$$3$$颗；对方取$$3$$颗，我方取$$1$$颗；对方取$$7$$颗，我方取$$1$$）．由于$$2016$$是$$8$$的倍数，选择后手提取，可以保证每次自己提取之后，剩余数量都是$$4$$的倍数，直至最后剩下$$8$$颗或$$4$$颗．在$$4$$颗的情况，对方只能取$$3$$或$$1$$颗，我方相应取$$1$$或$$3$$颗，取胜；在$$8$$颗的情况，对方若取$$1$$或$$7$$颗，我方相应取$$7$$或$$1$$颗，取胜；对方若取$$3$$颗，我方取$$1$$颗，转化为$$4$$颗的情况．	抢占制胜点
	TOP2	问题：$$1997$$个空格排成一排，第一格中放有一枚棋子，现有两人做游戏，轮流移动棋子，每人每次可前移$$1$$格、$$2$$格、$$3$$格或$$4$$格；谁选移到最后一格，谁失败．问怎样的移法才能确保获胜？	抢占制胜点
		解答：便于方便，可以把这$$1997$$个空格编成$$1$$号、$$2$$号、$$\cdots\cdots$$、$$1997$$号．要想取胜，应使棋子依次移到号码被$$5$$除余$$1$$的空格处．即$$16$$、$$11$$、$$16$$$$\cdots\cdots$$$$1991$$、$$1996$$.	
	TOP3	问题：今有两堆火柴，一堆$$6$$根，另一堆$$8$$根．两人轮流在其中任一堆中拿取，取的根数不限，但不能不取．规定取得最后一根者为赢．问：先取者有何策略能获胜？ 如果这两堆火柴数变成一堆$$12$$根，一堆$$12$$根，先取者还是后取者有必胜策略呢？	抢占制胜点
		解答：先取者在$$8$$根一堆火柴中取$$2$$根火柴，使得取后剩下两堆的火柴数相同．以后无论对手在某一堆取几根火柴，你只需在另一堆也取同样多根火柴．只要对手有火柴可取，你也有火柴可取，也就是说，最后一根火柴总会被你拿到．这样先取者总可获胜．如果都变成$$12$$根，后取者只需要在另一堆取跟对手相同多根火柴，则后取者有必胜策略．	
	TOP4	问题：$$30$$粒珠子依$$8$$粒红色、$$2$$粒黑色、$$8$$粒红色、$$2$$粒黑色$$\cdots \cdots $$的次序串成一圈．一只蚱蜢从第$$2$$粒黑珠子起跳，每次跳过$$6$$粒珠子落在下一粒珠子上．这只蚱蜢至少要跳次才	复合数字的整除特征

		能再次落在黑珠子上.	应用
		解答：这些珠子每$$10$$粒珠子一个周期，我们可以推断出这$$30$$粒珠子数到第$$9$$和$$10$$、$$19$$和$$20$$、$$29$$和$$30\cdots \cdots $$的时候，会是黑珠子．刚才是从第$$10$$粒珠子开始跳，中间隔$$6$$粒，跳到第$$17$$粒，接下来是第$$24$$粒、$$31$$粒、$$38$$粒、$$45$$粒、$$52$$粒、$$59$$粒，一直跳到$$59$$粒的时候会是黑珠子，所以至少要跳$$7$$次．故答案为：$$7$$.	
	TOP5	问题：桌子上放着$$37$$根火柴，聪明昊、神奇涛二人轮流每次取走$$1\sim5$$根．规定谁取走最后一根火柴谁获胜．如果双方都采用最佳方法，聪明昊先取，神奇涛后取，你知道会胜吗.	抢占制胜点
SBERT-QA-CNN（问题、解答） MAP：0.87	TOP1	问题：把一枚棋子放在图中左下角的方格内，甲、乙两人玩这样一个游戏：双方轮流移动棋子，只能向上、向右或者向右上方沿$$45{}^\circ $$角移动，一次可以移动任意多格．谁把棋子移到了右上角的方格中即为赢，试问：如果甲先走，是否有必胜的策略，为什么（　）？有没有不确定	抢占制胜点
		解答：从右上角开始分析哪些位置必胜的，哪些位置是必败的，结果如图所示．因此甲第一步必然走到“√”上，抢占制胜点，故甲必胜．故选$$\text{A}$$.	
	TOP2	问题：一个箱子内装有$$2016$$颗棋子，两人轮流在其中取棋子，规定每人每次只能提取$$1$$、$$3$$、$$7$$颗棋子，不得不取，也不得多取，取到最后棋子的人取胜．为了确保取胜，你是愿意先手，还是愿意后手？说出你的选择答案和必胜的策略.	抢占制胜点
		解答：根据规则，虽然不能控制对方每次提取棋子得数量，但可以通过控制自己的数量保证每一轮双方提取棋子总和为$$4$$或$$8$$（对方取$$1$$颗，我方取$$7$$颗或$$3$$颗；对方取$$3$$颗，我方取$$1$$颗；对方取$$7$$颗，我方取$$1$$）．由于$$2016$$是$$8$$的倍数，选择后手提取，可以保证每次自己提取之后，剩余数量都是$$4$$的倍数，直至最后剩下$$8$$颗或$$4$$颗．在$$4$$颗的情况，对方只能取$$3$$或$$1$$颗，我方相应取$$1$$或$$3$$颗，取胜；在$$8$$颗的情况，对方若取$$1$$或$$7$$颗，我方相应取$$7$$或$$1$$颗，取胜；对方若取$$3$$颗，我方取$$1$$颗，转化为$$4$$颗的情况.	
	TOP3	问题：今有两堆火柴，一堆$$6$$根，另一堆$$8$$根．两人轮流在其中任一堆中拿取，取的根数不限，但不能不取．规定取得最后一根者为赢．问：先取者有何策略能获胜？ 如果这两堆火柴数变成一堆$$12$$根，一堆$$12$$根，先取者还是后取者有必胜策略呢？	抢占制胜点
		解答：先取者在$$8$$根一堆火柴中取$$2$$根火柴，使得取后剩下两堆的火柴数相同．以后无论对手在某一堆取几根火柴，你只需在另一堆也取同样多根火柴．只要对手有火柴可取，你也有火柴可取，也就是说，最后一根火柴总会被你拿到．这样先取者总可获胜．如果都变成$$12$$根，后取者只需要在另一堆取跟对手相同多根火柴，则后取者有必胜策略.	
	TOP4	问题：学校有菊花和月季花共$$118$$盆，菊花比月季花的$$4$$倍少$$12$$盆，学校有菊花和月季花各多少盆？	多位数除法的实际应用
		解答：菊花再增加$$12$$盆，就正好是月季花的$$4$$倍，则这时菊花与月季花的和正好是月季花的$$5$$倍．月季花：$$(118+12)\div (1+4)=26$$（盆）菊花：$$26\times 4-12=92$$（盆）.	
	TOP5	问题：李阿姨在画廊一侧摆放花盆，每隔$$2$$米放一盆，一共放了$$16$$盆花．从第一盆花到最后一盆花的距离有多远？	两端植树问题变型
SBERT-CLS（问题） MAP：0.87	TOP1	问题：把一枚棋子放在图中左下角的方格内，甲、乙两人玩这样一个游戏：双方轮流移动棋子，只能向上、向右或者向右上方沿$$45{}^\circ $$角移动，一次可以移动任意多格．谁把棋子移到了右上角的方格中即为赢，试问：如果甲先走，是否有必胜的策略，为什么（　）？有没有不确定	抢占制胜点
		解答：从右上角开始分析哪些位置必胜的，哪些位置是必败的，结果如图所示．因此甲第一步必然走到“√”上，抢占制胜点，故甲必胜．故选$$\text{A}$$.	
	TOP2	问题：$$1997$$个空格排成一排，第一格中放有一枚棋子，现有两人做游戏，轮流移动棋子，每人每次可前移$$1$$格、$$2$$格、$$3$$格或$$4$$格；谁选移到最后一格，谁失败．问怎样的移法才能确保获胜？	抢占制胜点
		解答：便于方便，可以把这$$1997$$个空格编成$$1$$号、$$2$$号、	

		$$\cdots\cdots$$、$$1997$$号．要想取胜，应使棋子依次移到号码被$$5$$除余$$1$$的空格处．即$$16$$、$$11$$、$$16$$$$\cdots\cdots$$$$1991$$、$$1996$$．	
	TOP3	问题：一个箱子内装有$$2016$$颗棋子，两人轮流在其中取棋子，规定每人每次只能提取$$1$$、$$3$$、$$7$$颗棋子，不得不取，也不得多取，取到最后棋子的人取胜．为了确保取胜，你是愿意先手，还是愿意后手？说出你的选择答案和必胜的策略．	抢占制胜点
		解答：根据规则，虽然不能控制对方每次提取棋子得数量，但可以通过控制自己的数量保证每一轮双方提取棋子总和为$$4$$或$$8$$（对方取$$1$$颗，我方取$$7$$颗或$$3$$颗；对方取$$3$$颗，我方取$$1$$颗；对方取$$7$$颗，我方取$$1$$）．由于$$2016$$是$$8$$的倍数，选择后手提取，可以保证每次自己提取之后，剩余数量都是$$4$$的倍数，直至最后剩下$$8$$颗或$$4$$颗．在$$4$$颗的情况，对方只能取$$3$$或$$1$$颗，我方相应取$$1$$或$$3$$颗，取胜；在$$8$$颗的情况，对方若取$$1$$或$$7$$颗，我方相应取$$7$$或$$1$$颗，取胜；对方若取$$3$$颗，我方取$$1$$颗，转化为$$4$$颗的情况．	
	TOP4	问题：小红、小兰和小明三人玩掷小正方体的游戏，每个小正方体的六个面都分别写着$$1$$、$$2$$、$$3$$、$$4$$、$$5$$、$$6$$．小红说：将两个小正方体一起掷出，看朝上两个数的和是多少．小明说：和是$$6$$，算小红胜；和是$$7$$，算小兰胜；和是$$8$$，算我胜．他们三个人获胜的可能性最大．	计数求概率
		解答：$$6=1+5=2+4=3+3$$，有$$5$$种可能，$$7=1+6=2+5=3+4$$，有$$6$$种可能，$$8=2+6=3+5=4+4$$，有$$5$$种可能，所以，小兰获胜的可能性最大．	
	TOP5	问题：孙一夫在中国好声音的文化测试中，需从$$5$$个试题中任意选答$$3$$题，问：有几种不同的选题方法？若有一道题是必答题，有几种不同的选题方法？	组合问题
		解答：所求不同的选题方法数，就是从$$5$$个不同元素里取出$$3$$个元素的组合数，即$$\text{C}_{5}^{3}=10$$种因为已有一道题必选，所以只要在另外$$4$$道题中选$$2$$道，不同的选题方法有$$\text{C}_{4}^{2}=6$$种	
TBERT-QA（问题、解答） MAP：0.84	TOP1	问题：$$1997$$个空格排成一排，第一格中放有一枚棋子，现有两人做游戏，轮流移动棋子，每人每次可前移$$1$$格、$$2$$格、$$3$$格或$$4$$格；谁选移到最后一格，谁失败．问怎样的移法才能确保获胜？	抢占制胜点
		解答：便于方便，可以把这$$1997$$个空格编成$$1$$号、$$2$$号、$$\cdots\cdots$$、$$1997$$号．要想取胜，应使棋子依次移到号码被$$5$$除余$$1$$的空格处．即$$16$$、$$11$$、$$16$$$$\cdots\cdots$$$$1991$$、$$1996$$	
	TOP2	问题：今有两堆火柴，一堆$$6$$根，另一堆$$8$$根．两人轮流在其中任一堆中拿取，取的根数不限，但不能不取．规定取得最后一根者为赢．问：先取者有何策略能获胜？ 如果这两堆火柴数变成一堆$$12$$根，一堆$$12$$根，先取者还是后取者有必胜策略呢？	抢占制胜点
		解答：先取者在$$8$$根一堆火柴中取$$2$$根火柴，使得取后剩下两堆的火柴数相同．以后无论对手在某一堆取几根火柴，你只需在另一堆也取同样多根火柴．只要对手有火柴可取，你也有火柴可取，也就是说，最后一根火柴总会被你拿到．这样先取者总可获胜．如果都变成$$12$$根，后取者只需要在另一堆取跟对手相同多根火柴，则后取者有必胜策略．	
	TOP3	问题：六个人进行象棋单循环赛，规定胜者得$$2$$分，负者得$$0$$分，和棋双方各得$$1$$分，比赛结束后统计发现，六个人的得分加起来一定是分．	2-1-0 赛制
		解答：六个人单循环赛总共赛$$6\times (6-1)\div 2=15$$（场），每场无论分出胜负还是打平，比赛双方的得分和一定是$$2$$分，因此最终六个人的得分加起来一定是$$2\times 15=30$$（分）．	
	TOP4	问题：把一枚棋子放在图中左下角的方格内，甲、乙两人玩这样一个游戏：双方轮流移动棋子，只能向上、向右或者向右上方沿$$45{}^\circ $$角移动，一次可以移动任意多格．谁把棋子移到了右上角的方格中即为赢，试问：如果甲先走，是否有必胜的策略，为什么（ ）？有没有不确定	抢占制胜点
		解答：从右上角开始分析哪些位置必胜的，哪些位置是必败的，结果如图	

		所示．因此甲第一步必然走到“√”上，抢占制胜点，故甲必胜．故选$$\text{A}$$.	
	TOP5	问题：一个箱子内装有$$2016$$颗棋子，两人轮流在其中取棋子，规定每人每次只能提取$$1$$、$$3$$、$$7$$颗棋子，不得不取，也不得多取，取到最后棋子的人取胜．为了确保取胜，你是愿意先手，还是愿意后手？说出你的选择答案和必胜的策略．	抢占制胜点
		解答：根据规则，虽然不能控制对方每次提取棋子得数量，但可以通过控制自己的数量保证每一轮双方提取棋子总和为$$4$$或$$8$$（对方取$$1$$颗，我方取$$7$$颗或$$3$$颗；对方取$$3$$颗，我方取$$1$$颗；对方取$$7$$颗，我方取$$1$$）．由于$$2016$$是$$8$$的倍数，选择后手提取，可以保证每次自己提取之后，剩余数量都是$$4$$的倍数，直至最后剩下$$8$$颗或$$4$$颗．在$$4$$颗的情况，对方只能取$$3$$或$$1$$颗，我方相应取$$1$$或$$3$$颗，取胜；在$$8$$颗的情况，对方若取$$1$$或$$7$$颗，我方相应取$$7$$或$$1$$颗，取胜；对方若取$$3$$颗，我方取$$1$$颗，转化为$$4$$颗的情况．	
TBERT-CNN（解答） MAP：0.70	TOP1	问题：$$1997$$个空格排成一排，第一格中放有一枚棋子，现有两人做游戏，轮流移动棋子，每人每次可前移$$1$$格、$$2$$格、$$3$$格或$$4$$格；谁选移到最后一格，谁失败．问怎样的移法才能确保获胜？	抢占制胜点
		解答：便于方便，可以把这$$1997$$个空格编成$$1$$号、$$2$$号、$$\cdots\cdots$$、$$1997$$号．要想取胜，应使棋子依次移到号码被$$5$$除余$$1$$的空格处．即$$16$$、$$11$$、$$16$$$$\cdots\cdots$$$$1991$$、$$1996$$	
	TOP2	问题：从一个装有$$4$$个黑球和$$3$$个白球的口袋中任意摸一个球，则摸得白球的可能性是	描述随机现象发生的可能性大小
		解答：总共有$$7$$个球其中白球$$3$$个，占总个数的$$\frac{3}{7}$$.	
	TOP3	问题：桌子上放着$$37$$根火柴，聪明昊、神奇涛二人轮流每次取走$$1\sim5$$根．规定谁取走最后一根火柴谁获胜．如果双方都采用最佳方法，聪明昊先取，神奇涛后取，你知道会胜吗．	抢占制胜点
		解答：由$$37\div(1+5)=6\cdots\cdots1$$知聪明昊会胜．	
	TOP4	问题：一个箱子内装有$$2016$$颗棋子，两人轮流在其中取棋子，规定每人每次只能提取$$1$$、$$3$$、$$7$$颗棋子，不得不取，也不得多取，取到最后棋子的人取胜．为了确保取胜，你是愿意先手，还是愿意后手？说出你的选择答案和必胜的策略．	抢占制胜点
		解答：根据规则，虽然不能控制对方每次提取棋子得数量，但可以通过控制自己的数量保证每一轮双方提取棋子总和为$$4$$或$$8$$（对方取$$1$$颗，我方取$$7$$颗或$$3$$颗；对方取$$3$$颗，我方取$$1$$颗；对方取$$7$$颗，我方取$$1$$）．由于$$2016$$是$$8$$的倍数，选择后手提取，可以保证每次自己提取之后，剩余数量都是$$4$$的倍数，直至最后剩下$$8$$颗或$$4$$颗．在$$4$$颗的情况，对方只能取$$3$$或$$1$$颗，我方相应取$$1$$或$$3$$颗，取胜；在$$8$$颗的情况，对方若取$$1$$或$$7$$颗，我方相应取$$7$$或$$1$$颗，取胜；对方若取$$3$$颗，我方取$$1$$颗，转化为$$4$$颗的情况．	
	TOP5	问题：光明小学评出$$29$$位三好学生，每人奖励一个文具盒和一支钢笔，每个文具盒是$$8$$元，每支钢笔是$$11$$元，王老师大约需要带多少钱去购买这些奖品呢？	求算式整数部分
		解答：三好学生有$$29$$名，每位同学的奖品都是一个铅笔盒和一支钢笔，共$$19$$元，把$$29$$估成$$30$$，$$19$$估成$$20$$，王老师应该带$$29\times 19\approx 600$$（元）．	
SBERT-CNN（问题） MAP：0.54	TOP1	问题：六个人进行象棋单循环赛，规定胜者得$$2$$分，负者得$$0$$分，和棋双方各得$$1$$分，比赛结束后统计发现，六个人的得分加起来一定是分．	2-1-0 赛制
		解答：六个人单循环赛总共赛$$6\times (6-1)\div 2=15$$（场），每场无论分出胜负还是打平，比赛双方的得分和一定是$$2$$分，因此最终六个人的得分加起来一定是$$2\times 15=30$$（分）．	
	TOP2	问题：把一枚棋子放在图中左下角的方格内，甲、乙两人玩这样一个游戏：双方轮流移动棋子，只能向上、向右或者向右上方沿$$45{}^\circ $$角	抢占制胜点

		移动，一次可以移动任意多格．谁把棋子移到了右上角的方格中即为赢，试问：如果甲先走，是否有必胜的策略，为什么（ ）？有没有不确定	
		解答：从右上角开始分析哪些位置必胜的，哪些位置是必败的，结果如图所示．因此甲第一步必然走到“√”上，抢占制胜点，故甲必胜．故选$$\text{A}$$.	
	TOP3	问题：$$1997$$个空格排成一排，第一格中放有一枚棋子，现有两人做游戏，轮流移动棋子，每人每次可前移$$1$$格、$$2$$格、$$3$$格或$$4$$格；谁选移到最后一格，谁失败．问怎样的移法才能确保获胜？	抢占制胜点
		解答：便于方便，可以把这$$1997$$个空格编成$$1$$号、$$2$$号、$$\cdots\cdots$$、$$1997$$号．要想取胜，应使棋子依次移到号码被$$5$$除余$$1$$的空格处．即$$16$$、$$11$$、$$16$$$$\cdots\cdots$$$$1991$$、$$1996$$.	
	TOP4	问题：一个箱子内装有$$2016$$颗棋子，两人轮流在其中取棋子，规定每人每次只能提取$$1$$、$$3$$、$$7$$颗棋子，不得不取，也不得多取，取到最后棋子的人取胜．为了确保取胜，你是愿意先手，还是愿意后手？说出你的选择答案和必胜的策略	抢占制胜点
		解答：根据规则，虽然不能控制对方每次提取棋子得数量，但可以通过控制自己的数量保证每一轮双方提取棋子总和为$$4$$或$$8$$（对方取$$1$$颗，我方取$$7$$颗或$$3$$颗；对方取$$3$$颗，我方取$$1$$颗；对方取$$7$$颗，我方取$$1$$）．由于$$2016$$是$$8$$的倍数，选择后手提取，可以保证每次自己提取之后，剩余数量都是$$4$$的倍数，直至最后剩下$$8$$颗或$$4$$颗．在$$4$$颗的情况，对方只能取$$3$$或$$1$$颗，我方相应取$$1$$或$$3$$颗，取胜；在$$8$$颗的情况，对方若取$$1$$或$$7$$颗，我方相应取$$7$$或$$1$$颗，取胜；对方若取$$3$$颗，我方取$$1$$颗，转化为$$4$$颗的情况．	
	TOP5	问题：桌子上放着$$37$$根火柴，聪明昊、神奇涛二人轮流每次取走$$1\sim5$$根．规定谁取走最后一根火柴谁获胜．如果双方都采用最佳方法，聪明昊先取，神奇涛后取，你知道会胜吗．	抢占制胜点
TLSTM（问题） MAP：0.46	TOP1	问题：桌子上放着$$37$$根火柴，聪明昊、神奇涛二人轮流每次取走$$1\sim5$$根．规定谁取走最后一根火柴谁获胜．如果双方都采用最佳方法，聪明昊先取，神奇涛后取，你知道会胜吗．	抢占制胜点
		解答：由$$37\div(1+5)=6\cdots\cdots1$$知聪明昊会胜．	
	TOP2	问题：请把下面各组分数通分．$$\frac{8}{9}和\frac{5}{6}$$ $$\frac{3}{10}和\frac{1}{4}$$ $$\frac{7}{10}和\frac{4}{15}$$ $$\frac{5}{12}和\frac{3}{8}$$	分数的通分
		解答：$$\frac{8}{9}和\frac{5}{6}=\frac{16}{18}和\frac{15}{18}$$；$$\frac{3}{10}和\frac{1}{4}=\frac{6}{20}和\frac{5}{20}$$；$$\frac{7}{10}和\frac{4}{15}=\frac{21}{30}和\frac{8}{30}$$；$$\frac{5}{12}和\frac{3}{8}=\frac{10}{24}和\frac{9}{24}$$．故答案为：$$\frac{16}{18}和\frac{15}{18}$$；$$\frac{6}{20}和\frac{5}{20}$$；$$\frac{21}{30}和\frac{8}{30}$$；$$\frac{10}{24}和\frac{9}{24}$$.	
	TOP3	问题：长颈鹿召开草原和平大会，邀请了$$28$$只小动物参加会议．最后所有参加会议者併坐成一排拍团体照，如果长颈鹿项坐在正中间，请问它应该坐在从左边算起第几位呢？$$12$$ $$13$$ $$14$$ $$15$$ $$16$$	两端植树问题变型
		解答：若长颈鹿坐在正中间，则其左右两边的动物数应相等．因此由$$28\div 2=14$$可得知长颈鹿的左右两边都是$$14$$只小动物，而$$14+1=15$$，所以长颈鹿坐在左起的第$$15$$位上．故选$$\text{D}$$.	
	TOP4	问题：$$1$$、$$2$$、$$3$$、$$4$$、$$5$$、$$6$$、$$7$$这$$7$$个数中，选出两个互不相同的数，使得这两个数的乘积能被$$3$$整除，不同的选法有几种？$$11$$种 $$12$$种 $$13$$种 $$14$$种	数的整除
		解答：$$2\times 6-1=11$$种.	
	TOP5	问题：有$$2018$$盏亮着的电灯，各有一根拉线开关控制着，现按其顺序编号为$$1$$、$$2$$、$$3$$、……$$2018$$，然后将编号为$$2$$的倍数的灯线拉一下，再将编号为$$3$$的倍数的灯线拉一下，最后将编号是$$5$$的倍数的灯线拉一下，三次拉完后，亮着的灯有盏.	容斥原理
		解答：拉一次，灯变暗，拉两次，灯变亮，所以我们计算的是没有拉过的以及拉过两次的灯：$$2018\div 2=1009$$，$$\left[2018\div 3 \right]=672$$，$$\left[2018\div 5 \right]=403$$，$$\left[2018\div 6	

		\right]=336$$，$$\left[2018\div 10 \right]=201$$，$$\left[2018\div 15 \right]=134$$，$$\left[2018\div 30 \right]=67$$，那么没有拉过的灯有$$2018-\left(1009+672+403-336-201-134+67 \right)=538$$盏，拉过两次的灯有$$336+201+134-67\times 3=470$$盏，共有$$538+470=1008$$盏亮着的.	
SBERT-CNN（解答） MAP：0.45	TOP1	问题：$$30$$粒珠子依$$8$$粒红色、$$2$$粒黑色、$$8$$粒红色、$$2$$粒黑色$$\cdots \cdots $$的次序串成一圈．一只蚱蜢从第$$2$$粒黑珠子起跳，每次跳过$$6$$粒珠子落在下一粒珠子上．这只蚱蜢至少要跳次才能再次落在黑珠子上.	复合数字的整除特征应用
		解答：这些珠子每$$10$$粒珠子一个周期，我们可以推断出这$$30$$粒珠子数到第$$9$$和$$10$$、$$19$$和$$20$$、$$29$$和$$30\cdots \cdots $$的时候，会是黑珠子．刚才是从第$$10$$粒珠子开始跳，中间隔$$6$$粒，跳到第$$17$$粒，接下来是第$$24$$粒、$$31$$粒、$$38$$粒、$$45$$粒、$$52$$粒、$$59$$粒，一直跳到$$59$$粒的时候会是黑珠子，所以至少要跳$$7$$次．故答案为：$$7$$.	
	TOP2	问题：桌子上放着$$37$$根火柴，聪明昊、神奇涛二人轮流每次取走$$1\sim5$$根．规定谁取走最后一根火柴谁获胜．如果双方都采用最佳方法，聪明昊先取，神奇涛后取，你知道会胜吗.	抢占制胜点
		解答：由$$37\div(1+5)=6\cdots\cdots1$$知聪明昊会胜.	
	TOP3	问题：$$1997$$个空格排成一排，第一格中放有一枚棋子，现有两人做游戏，轮流移动棋子，每人每次可前移$$1$$格、$$2$$格、$$3$$格或$$4$$格；谁选移到最后一格，谁失败．问怎样的移法才能确保获胜？	抢占制胜点
		解答：便于方便，可以把这$$1997$$个空格编成$$1$$号、$$2$$号、$$\cdots\cdots$$、$$1997$$号．要想取胜，应使棋子依次移到号码被$$5$$除余$$1$$的空格处．即$$16$$、$$11$$、$$16$$$$\cdots\cdots$$$$1991$$、$$1996$$.	
	TOP4	问题：如果允许砝码放在天平两端，那么能称量出$$1$$～$$121$$克中任何整数克重的物品，至少需要个砝码？	智巧趣题
		解答：$$5$$	
	TOP5	问题：一个箱子内装有$$2016$$颗棋子，两人轮流在其中取棋子，规定每人每次只能提取$$1$$、$$3$$、$$7$$颗棋子，不得不取，也不得多取，取到最后棋子的人取胜．为了确保取胜，你是愿意先手，还是愿意后手？说出你的选择答案和必胜的策略.	抢占制胜点
		解答：根据规则，虽然不能控制对方每次提取棋子得数量，但可以通过控制自己的数量保证每一轮双方提取棋子总和为$$4$$或$$8$$（对方取$$1$$颗，我方取$$7$$颗或$$3$$颗；对方取$$3$$颗，我方取$$1$$颗；对方取$$7$$颗，我方取$$1$$）．由于$$2016$$是$$8$$的倍数，选择后手提取，可以保证每次自己提取之后，剩余数量都是$$4$$的倍数，直至最后剩下$$8$$颗或$$4$$颗．在$$4$$颗的情况，对方只能取$$3$$或$$1$$颗，我方相应取$$1$$或$$3$$颗，取胜；在$$8$$颗的情况，对方若取$$1$$或$$7$$颗，我方相应取$$7$$或$$1$$颗，取胜；对方若取$$3$$颗，我方取$$1$$颗，转化为$$4$$颗的情况.	
SBERT-CLS（解答） MAP：0.20	TOP1	问题：北京、上海分别有$$3$$台和$$14$$台完全相同的机器，准备给武汉$$8$$台，西安$$9$$台，每台机器的运费如下表，如何调运能使总运费最省？	调运中的统筹
		解答：北京应该给西安$$3$$台；上海给武汉$$8$$台，给西安$$6$$台运费最少	
	TOP2	问题：如果允许砝码放在天平两端，那么能称量出$$1$$～$$121$$克中任何整数克重的物品，至少需要个砝码？	智巧趣题
		解答：$$5$$	
	TOP3	问题：$$1997$$个空格排成一排，第一格中放有一枚棋子，现有两人做游戏，轮流移动棋子，每人每次可前移$$1$$格、$$2$$格、$$3$$格或$$4$$格；谁选移到最后一格，谁失败．问怎样的移法才能确保获胜？	抢占制胜点
		解答：便于方便，可以把这$$1997$$个空格编成$$1$$号、$$2$$号、$$\cdots\cdots$$、$$1997$$号．要想取胜，应使棋子依次移到号码被$$5$$除余$$1$$的空格处．即$$16$$、$$11$$、$$16$$$$\cdots\cdots$$$$1991$$、$$1996$$.	
	TOP4	问题：有这样一种数学游戏：在$$3\times 3$$的表格中，要求每个格子中都	分类加

		填上$$1$$、$$2$$、$$3$$三个数字中的某一个数字，且每一行和每一列都不能出现重复的数字，则此游戏共有种不同的填法.	法计数原理、分步乘法计数原理
		解答：根据题意，如图，在$$A$$格中，有$$3$$种填法，在$$B$$、$$C$$格中，都有$$2$$种不同的填法，再分析格$$D$$中，当$$B$$、$$C$$中不同时，则$$D$$中只能与$$A$$的相同，当$$B$$、$$C$$中相同时，考虑每一行和每一列都不能出现重复的数字，则$$D$$中的只能是$$3$$个数，无论$$B$$、$$C$$中相同与否，$$D$$中都只有一种填法，进而可得其他格中，也只有一种填法，则有$$3\times 2\times 2=12$$种不同的填法．故答案为：$$12$$.	
	TOP5	问题：桌子上放着$$37$$根火柴，聪明昊、神奇涛二人轮流每次取走$$1\sim5$$根．规定谁取走最后一根火柴谁获胜．如果双方都采用最佳方法，聪明昊先取，神奇涛后取，你知道会胜吗.	抢占制胜点
		解答：由$$37\div(1+5)=6\cdots\cdots1$$知聪明昊会胜.	
SLSTM（问题） MAP：0.20	TOP1	问题：$$30$$粒珠子依$$8$$粒红色、$$2$$粒黑色、$$8$$粒红色、$$2$$粒黑色$$\cdots \cdots $$的次序串成一圈．一只蚱蜢从第$$2$$粒黑珠子起跳，每次跳过$$6$$粒珠子落在下一粒珠子上．这只蚱蜢至少要跳次才能再次落在黑珠子上.	复合数字的整除特征应用
		解答：这些珠子每$$10$$粒珠子一个周期，我们可以推断出这$$30$$粒珠子数到第$$9$$和$$10$$、$$19$$和$$20$$、$$29$$和$$30\cdots \cdots $$的时候，会是黑珠子．刚才是从第$$10$$粒珠子开始跳，中间隔$$6$$粒，跳到第$$17$$粒，接下来是第$$24$$粒、$$31$$粒、$$38$$粒、$$45$$粒、$$52$$粒、$$59$$粒，一直跳到$$59$$粒的时候会是黑珠子，所以至少要跳$$7$$次．故答案为：$$7$$.	
	TOP2	问题：某个游戏要从袋中摸球；袋中有大小、形状相同的白球$$2$$个，红球$$3$$个；游戏的规则是：每次不放回地摸出$$1$$个球；连续$$2$$次摸出相同颜色的球时游戏结束，没有达到这个条件就把球摸完了，视为游戏失败；请问："某次游戏成功结束了，并且摸出的球中有白球"这件事发生的概率是多少？游戏失败的概率是多少？	计数求概率
		解答：反向考虑：只有在前两个直接摸出"红红"时才没有白球，概率为$$1-\frac {1}{10}-\frac {3}{5}\times \frac {2}{4}=\frac {3}{5}$$．只有"红白红白红"会失败，$$\frac{1}{\text{C}_{5}^{3}}=\frac{1}{10}$$（或$$\frac{(3\times 2\times 1)\times (2\times 1)}{\text{A}_{5}^{5}}=\frac{1}{10}$$）.	
	TOP3	问题：一个箱子内装有$$2016$$颗棋子，两人轮流在其中取棋子，规定每人每次只能提取$$1$$、$$3$$、$$7$$颗棋子，不得不取，也不得多取，取到最后棋子的人取胜．为了确保取胜，你是愿意先手，还是愿意后手？说出你的选择答案和必胜的策略.	抢占制胜点
		解答：根据规则，虽然不能控制对方每次提取棋子得数量，但可以通过控制自己的数量保证每一轮双方提取棋子总和为$$4$$或$$8$$（对方取$$1$$颗，我方取$$7$$颗或$$3$$颗；对方取$$3$$颗，我方取$$1$$颗；对方取$$7$$颗，我方取$$1$$）．由于$$2016$$是$$8$$的倍数，选择后手提取，可以保证每次自己提取之后，剩余数量都是$$4$$的倍数，直至最后剩下$$8$$颗或$$4$$颗．在$$4$$颗的情况，对方只能取$$3$$或$$1$$颗，我方相应取$$1$$或$$3$$颗，取胜；在$$8$$颗的情况，对方若取$$1$$或$$7$$颗，我方相应取$$7$$或$$1$$颗，取胜；对方若取$$3$$颗，我方取$$1$$颗，转化为$$4$$颗的情况.	
	TOP4	问题：用数字$$6$$、$$7$$、$$8$$、$$9$$、$$1$$组三位数，组成无重复数字的三位数的概率是多少？	计数求概率
		解答：用$$6$$、$$7$$、$$8$$、$$9$$、$$1$$组三位数，可以组$$5\times 5\times 5=125$$个；其中无重复数字的三位数有$$5\times 4\times 3=60$$个．概率为$$\frac{60}{125}=\frac{12}{25}$$.	
	TOP5	问题：桌子上放着$$37$$根火柴，聪明昊、神奇涛二人轮流每次取走$$1\sim5$$根．规定谁取走最后一根火柴谁获胜．如果双方都采用最佳方法，聪明昊先取，神奇涛后取，你知道会胜吗.	抢占制胜点
TBERT-CLS（解答）	TOP1	问题：有$$2018$$盏亮着的电灯，各有一根拉线开关控制着，现按其顺序编号为$$1$$、$$2$$、$$3$$、……$$2018$$，然后将编号为$$2$$的倍数的灯线拉一下，再将编号为$$3$$的倍数的灯线拉一下，最后将编号是$$5$$的倍数的灯线拉一下，三次拉完后，亮着的灯有盏.	容斥原理

MAP：0.13		解答：拉一次，灯变暗，拉两次，灯变亮，所以我们计算的是没有拉过的以及拉过两次的灯：$$2018\div 2=1009$$，$$\left[2018\div 3 \right]=672$$，$$\left[2018\div 5 \right]=403$$，$$\left[2018\div 6 \right]=336$$，$$\left[2018\div 10 \right]=201$$，$$\left[2018\div 15 \right]=134$$，$$\left[2018\div 30 \right]=67$$，那么没有拉过的灯有$$2018-\left(1009+672+403-336-201-134+67 \right)=538$$盏，拉过两次的灯有$$336+201+134-67\times 3=470$$盏，共有$$538+470=1008$$盏亮着的.	
	TOP2	问题：学校有菊花和月季花共$$118$$盆，菊花比月季花的$$4$$倍少$$12$$盆，学校有菊花和月季花各多少盆？	多位数除法的实际应用
		解答：菊花再增加$$12$$盆，就正好是月季花的$$4$$倍，则这时菊花与月季花的和正好是月季花的$$5$$倍．月季花：$$(118+12)\div (1+4)=26$$（盆）菊花：$$26\times 4-12=92$$（盆）.	
	TOP3	问题：小红、小兰和小明三人玩掷小正方体的游戏，每个小正方体的六个面都分别写着$$1$$、$$2$$、$$3$$、$$4$$、$$5$$、$$6$$．小红说：将两个小正方体一起掷出，看朝上两个数的和是多少．小明说：和是$$6$$，算小红胜；和是$$7$$，算小兰胜；和是$$8$$，算我胜．他们三个人获胜的可能性最大.	计数求概率
		解答：$$6=1+5=2+4=3+3$$，有$$5$$种可能，$$7=1+6=2+5=3+4$$，有$$6$$种可能，$$8=2+6=3+5=4+4$$，有$$5$$种可能，所以，小兰获胜的可能性最大.	
	TOP4	问题：$$1997$$个空格排成一排，第一格中放有一枚棋子，现有两人做游戏，轮流移动棋子，每人每次可前移$$1$$格、$$2$$格、$$3$$格或$$4$$格；谁选移到最后一格，谁失败．问怎样的移法才能确保获胜？	抢占制胜点
		解答：便于方便，可以把这$$1997$$个空格编成$$1$$号、$$2$$号、$$\cdots\cdots$$、$$1997$$号．要想取胜，应使棋子依次移到号码被$$5$$除余$$1$$的空格处．即$$16$$、$$11$$、$$16$$$$\cdots\cdots$$$$1991$$、$$1996$$	
	TOP5	问题：桌子上放着$$37$$根火柴，聪明昊、神奇涛二人轮流每次取走$$1\sim5$$根．规定谁取走最后一根火柴谁获胜．如果双方都采用最佳方法，聪明昊先取，神奇涛后取，你知道会胜吗.	抢占制胜点
		解答：由$$37\div(1+5)=6\cdots\cdots1$$知聪明昊会胜.	
TBERT-CNN（问题） MAP：0.0	TOP1	问题：有$$2018$$盏亮着的电灯，各有一根拉线开关控制着，现按其顺序编号为$$1$$、$$2$$、$$3$$、……$$2018$$，然后将编号为$$2$$的倍数的灯线拉一下，再将编号为$$3$$的倍数的灯线拉一下，最后将编号是$$5$$的倍数的灯线拉一下，三次拉完后，亮着的灯有盏.	容斥原理
		解答：拉一次，灯变暗，拉两次，灯变亮，所以我们计算的是没有拉过的以及拉过两次的灯：$$2018\div 2=1009$$，$$\left[2018\div 3 \right]=672$$，$$\left[2018\div 5 \right]=403$$，$$\left[2018\div 6 \right]=336$$，$$\left[2018\div 10 \right]=201$$，$$\left[2018\div 15 \right]=134$$，$$\left[2018\div 30 \right]=67$$，那么没有拉过的灯有$$2018-\left(1009+672+403-336-201-134+67 \right)=538$$盏，拉过两次的灯有$$336+201+134-67\times 3=470$$盏，共有$$538+470=1008$$盏亮着的.	
	TOP2	问题：某学生要从物理、化学、生物、政治、历史、地理这六门学科中选三门参加等级考，要求是物理、化学、生物这三门至少要选一门，政治、历史、地理这三门也至少要选一门，则该生的可能选法总数是.	组合问题
		解答：$$C_{6}^{3}-2C_{3}^{3}=18$$.	
	TOP3	问题：有$$13$$个自然数，它们的平均值利用四舍五入精确到小数点后一位是$$26.9$$．那么，精确到小数点后两位数是多少？	放缩与估算
		解答：13 个自然数之和必然是整数．因为$$26.85\leqslant 平均数\leqslant 26.95$$，所以总和在$$13\times 26.85=349.05$$和$$13\times 26.95=350.35$$之间．因此，这些自然数的和未$$350$$，平均数为$$350\div 13\approx 26.923$$，精确到小数点后两位为$$26.92$$.	
	TOP4	问题：对任意一个正整数$$m$$，如果$$m=n(n+1)$$，其中$$n$$是正整数，则称$$m$$为“优数”，$$n$$为$$m$$的最优拆分点，例如：$$72=8\times (8+1)$$，则$$72$$是一个“优数”，$$8$$为$$72$$的最优拆分点．请写出一个“优数”，它的最优拆分点是．把“优数”$$p$$的$$2$$倍与“优数”$$q$$的$$3$$倍的差记为$$D(p,q)$$，例如：$$20=4\times 5$$，$$6=2\times 3$$，则$$D(20,6)=2\times 20-\times 6=22$$．若“优数”$$p$$的	单项式乘多项式

		最优拆分点为$$t+4$$，“优数”$$q$$的最优拆分点为$$t$$，当$$D(p,q)=76$$时，求$$t$$的值并判断它是否为“优数”．求证：若“优数”$$m$$是$$5$$的倍数，则$$m$$一定是$$10$$的倍数．	
		解答：∵$$56=7\times (7+1)$$，∴$$56$$是“优数”，它的最优拆分点是$$7$$，故答案为：$$56$$，$$7$$．由题意知，$$p=(t+4)(t+5)$$，$$q=t(t+1)$$，∵$$D(p,q)=2p-3q=76$$，∴$$2(t+4)(t+5)-3t(t+1)=76$$，∴$$t=3$$或$$t=12$$，∴$$3$$不是“优数”，$$12$$是“优数”．∵“优数”$$m$$是$$5$$的倍数，∴$$n(n+1)$$是$$5$$的倍数，（$$n$$是正整数），当$$n$$为奇数时，$$n+1$$是偶数，∴$$n(n+1)$$是能被$$5$$整除的偶数，故$$n(n+1)$$是$$10$$的倍数，当$$n$$为偶数时，∴$$n(n+1)$$是能被$$5$$整除的偶数，故$$n(n+1)$$是$$10$$的倍数，即：“优数”$$m$$是$$5$$的倍数，则$$m$$一定是$$10$$的倍数．	
	TOP5	问题：如题：两个小数的整数部分分别是$$4$$和$$5$$，那么这两个小数乘积的整数部分共有多少种可能的取值？将两个小数四舍五入到个位后，所得到的数值分别是$$7$$和$$9$$．将这两个小数的乘积四舍五入到个位后共有多少种可能的取值？	求算式整数部分
		解答：设两个小数分别为$$a$$和$$b$$，可得$$4\leqslant a<{}5$$，$$5\leqslant b<{}6$$．因此，我们得到$$a\times b\geqslant 4\times 5=20$$，$$a\times b<{}5\times 6=30$$．所以两个小数乘积的整数部分可取$$20$$到$$29$$之间的任何整数值，一共有$$10$$种可能的取值．设两个小数分别为$$a$$和$$b$$，由于两个小数四舍五入到个位后所得到的数值分别是$$7$$和$$9$$，所以考虑到小数点的情况，可得$$6.5\leqslant a<{}7.5$$，$$8.5\leqslant b<{}9.5$$．因此，我们得到$$a\times b\geqslant 6.5\times 8.5=55.25$$，$$a\times b<{}9.5\times 7.5=71.25$$．所以两个小数乘积的整数部分可取$$55$$到$$71$$之间的任何整数值，一共有$$17$$种可能的取值．	
TLSTM（解答） MAP：0.0	TOP1	问题：某个游戏要从袋中摸球；袋中有大小、形状相同的白球$$2$$个，红球$$3$$个；游戏的规则是：每次不放回地摸出$$1$$个球；连续$$2$$次摸出相同颜色的球时游戏结束，没有达到这个条件就把球摸完了，视为游戏失败；请问：“某次游戏成功结束了，并且摸出的球中有白球”这件事发生的概率是多少？游戏失败的概率是多少？	计数求概率
		解答：反向考虑：只有在前两个直接摸出“红红”时才没有白球，概率为$$1-\frac {1}{10}-\frac {3}{5}\times \frac {2}{4}=\frac {3}{5}$$．只有“红白红白红”会失败，$$\frac{1}{\text{C}_{5}^{3}}=\frac{1}{10}$$（或$$\frac{(3\times 2\times 1)\times (2\times 1)}{\text{A}_{5}^{5}}=\frac{1}{10}$$）．	
	TOP2	问题：$$\overline{abcd}$$，$$\overline{abc}$$，$$\overline{ab}$$，$$a$$依次表示四位数、三位数、两位数及一位数，且满足$$\overline{abcd}$$—$$\overline{abc}$$—$$\overline{ab}$$—$$a$$=$$1787$$，则这个四位数$$\overline{abcd}$$是多少？	位值原理的完全拆分
		解答：原式可表示成：$$889a+89b+9c+d=1787$$，则知$$a$$只能取：$$1$$或$$2$$，当$$a=1$$时，$$b$$无法取，故此值舍去．当$$a=2$$时，$$b=0$$，$$c=0$$或$$1$$，$$d$$相应的取$$9$$或$$0.$$所以这个四位数是：$$2009$$或$$2010$$．	
	TOP3	问题：如果允许砝码放在天平两端，那么能称量出$$1$$～$$121$$克中任何整数克重的物品，至少需要个砝码？	智巧趣题
		解答：$$5$$	
	TOP4	问题：将下列分数的每组分数通分$$\frac{2}{5}$$ $$\frac{3}{7}$$ $$\frac{4}{11}$$ $$\frac{5}{9}$$ $$8\frac{1}{12}$$ $$7\frac{4}{5}$$ $$22\frac{2}{5}$$ $$21\frac{3}{7}$$．$$\frac{5}{6}$$ $$\frac{7}{24}$$ $$\frac{2}{13}$$ $$\frac{2}{39}$$ $$\frac{4}{21}$$ $$4\frac{2}{7}$$ $$9\frac{3}{4}$$ $$5\frac{1}{16}$$．$$\frac{2}{9}$$ $$\frac{5}{12}$$ $$\frac{5}{6}$$ $$\frac{7}{8}$$ $$3\frac{4}{15}$$ $$5\frac{1}{10}$$ $$7\frac{1}{4}$$ $$12\frac{5}{6}$$．	分数的通分
		解答：略略略	
	TOP5	问题：把三张游园票分给$$10$$个人中的$$3$$人，分法有	组合问

		（ ）. $$A_{10}^{3}$$种 $$C_{10}^{3}$$种 $$C_{10}^{3}A_{10}^{3}$$种 $$30$$种	题
		解答：三张票没区别，从$$10$$人中选$$3$$人即可，即$$C_{10}^{3}$$.	
SLSTM（解答） MAP：0.0	TOP1	问题：估算. $$597+299≈$$ $$501+358≈$$ $$723-298≈$$	放缩与估算
		解答：$$597+299≈900$$，$$501+358≈860$$，$$723-298≈420$$.	
	TOP2	问题：某地冬季一周的气温走势如下表所示，那么这一周的平均气温为℃. 温度$$-1$$℃$$1$$℃$$2$$℃$$3$$℃$$4$$℃天数 $$1$$$$2$$$$1$$$$1$$$$2$$	加权平均数
		解答：∵$$\frac{-1\times 1+1\times 2+2\times 1+3\times 1+4\times 2}{7}=2$$，∴一周的平均气温为$$2$$℃.	
	TOP3	问题：如图所示，$$a$$、$$b$$是第$$7$$行的前两个数，$$b$$等于多少？	连续自然数三角形数表之已知位置求数
		解答：数表中，每个数等于它上方两数之和. 第六行第一个数是$$6$$，第六行第二个数$$5+11=16$$，$$b=6+16=22$$.	
	TOP4	问题：北京、上海分别有$$3$$台和$$14$$台完全相同的机器，准备给武汉$$8$$台，西安$$9$$台，每台机器的运费如下表，如何调运能使总运费最省？	调运中的统筹
		解答：北京应该给西安$$3$$台；上海给武汉$$8$$台，给西安$$6$$台运费最少.	
	TOP5	问题：$$\overline{abcd}$$，$$\overline{abc}$$，$$\overline{ab}$$，$$a$$依次表示四位数、三位数、两位数及一位数，且满足$$\overline{abcd}$$—$$\overline{abc}$$—$$\overline{ab}$$—$$a$$=$$1787$$，则这个四位数$$\overline{abcd}$$是多少？	位值原理的完全拆分
		解答：原式可表示成：$$889a+89b+9c+d=1787$$，则知$$a$$只能取：$$1$$或$$2$$，当$$a=1$$时，$$b$$无法取，故此值舍去. 当$$a=2$$时，$$b=0$$，$$c=0$$或$$1$$，$$d$$相应的取$$9$$或$$0.$$所以这个四位数是：$$2009$$或$$2010$$.	
BERT（问题） MAP：0.0	TOP1	问题：如右图所示，由三个正方体木块粘合而成的模型，它们的棱长分别为$$1$$米、$$2$$米、$$4$$米，要在表面涂刷油漆（底面也涂油漆），则模型涂刷油漆的面积是平方米.	三视图求表面积与体积综合
		解答：该图形从前、后、左、右四面观察到的面积都是$${{1}^{2}}+{{2}^{2}}+{{4}^{2}}=21$$平方米，从上面和下面观察到的面积是$${{4}^{2}}=16$$平方米，所以涂刷油漆的面积是$$21\times 4+16\times 2=116$$平方米.	
	TOP2	问题：计算：$$2xy\left(\frac{1}{2}x-3y+1 \right)$$.	单项式乘多项式
		解答：$$2xy\left(\frac{1}{2}x-3y+1 \right)={{x}^{2}}y-6x{{y}^{2}}+2xy$$.	
	TOP3	问题：定积分$$\int_{0}^{3}{\sqrt{9-{{x}^{2}}}dx}$$的值为（ ）. $$9 \pi $$ $$3 \pi $$ $$\frac{9}{4} \pi $$ $$\frac{9}{2} \pi $$	定积分的几何意义
		解答：$$\int_{0}^{3}{\sqrt{9-{{x}^{2}}}dx}$$即为$${{x}^{2}}+{{y}^{2}}=9$$在第一象限图象的面积，即半径为$$3$$的四分之一圆的面积$$\frac{1}{4} \pi \cdot {{3}^{2}}=\frac{9}{4} \pi $$.	
	TOP4	问题：图中相邻三点所形成的等边三角形的面积为$$1$$，那么格点多边形的面积为.	三角形格点多边形
		解答：根据毕克定理，$$(5+4\div 2-1)\times 2=12$$.	
	TOP5	问题：已知直线过点$$A(1,2)$$，且原点到这条直线的距离为$$1$$，则这条直线的方程是（ ）. $$3x-4y+5=0$$和$$x=1$$ $$4x-3y+5=0$$和$$y=1$$ $$3x-4y+5=0$$和$$y=1$$ $$4x-3y+5=0$$和$$x=1$$	直线的倾斜角与斜率、直线的位置关系、直线的方程'
		解答：设直线方程为$$y-2=k(x-1)$$，即$$kx-y+2-k=0$$，∵原点到这条直线的距离为$$1$$，∴$$\frac{\left| -2-k \right|}{\sqrt{{{k}^{2}}+1}}=1$$，解之得$$k=\frac{3}{4}$$. 可得直线方程为$$\frac{3}{4}x-y+2-\frac{3}{4}=0$$，即$$3x-4y+5=0$$. 又∵当直线的斜率不存在时，方程为$$x=1$$，到原点的距离也等于$$1$$. ∴所求直线的方程是$$3x-4y+5=0$$和$$x=1$$. 故选 A.	
VSM（问题）	TOP1	问题：有$$6$$个数，它们的平均数是$$12$$，再添加一个数$$5$$，则这$$7$$个数的平均数是.	加权平均数

MAP：0.0		解答：这$$7$$个数的平均数是$$\frac{12\times 6+5}{7}=11$$.	
	TOP2	问题：如图，从$$1$$开始的自然数按某种方式排列起来，第$$15$$行左起第$$8$$个数是.	连续自然数三角形数表之已知位置求数
		解答：先求项数：一共有$$14$$个完整行，有$$1+2+3+\cdot\cdot\cdot+14=\left(1+14\right)\times14\div2=105$$（项），还剩下$$8$$个数，所以是第$$105+8=113$$（项），因为是从$$1$$开始的连续自然数列，第$$113$$项就是$$113$$.	
	TOP3	问题：$$1$$、$$2$$、$$3$$、$$4$$、$$5$$、$$6$$、$$7$$这$$7$$个数中，选出两个互不相同的数，使得这两个数的乘积能被$$3$$整除，不同的选法有几种？$$11$$种 $$12$$种 $$13$$种 $$14$$种	数的整除
		解答：$$2\times 6-1=11$$种.	
	TOP4	问题：如图所示，$$a$$、$$b$$是第$$7$$行的前两个数，$$b$$等于多少？	连续自然数三角形数表之已知位置求数
		解答：数表中，每个数等于它上方两数之和．第六行第一个数是$$6$$，第六行第二个数$$5+11=16$$，$$b=6+16=22$$.	
	TOP5	问题：如题：两个小数的整数部分分别是$$4$$和$$5$$，那么这两个小数乘积的整数部分共有多少种可能的取值？将两个小数四舍五入到个位后，所得到的数值分别是$$7$$和$$9$$．将这两个小数的乘积四舍五入到个位后共有多少种可能的取值？	求算式整数部分
		解答：设两个小数分别为$$a$$和$$b$$，可得$$4\leqslant a<{}5$$，$$5\leqslant b<{}6$$．因此，我们得到$$a\times b\geqslant 4\times 5=20$$，$$a\times b<{}5\times 6=30$$．所以两个小数乘积的整数部分可取$$20$$到$$29$$之间的任何整数值，一共有$$10$$种可能的取值．设两个小数分别为$$a$$和$$b$$，由于两个小数四舍五入到个位后所得到的数值分别是$$7$$和$$9$$，所以考虑到小数点的情况，可得$$6.5\leqslant a<{}7.5$$，$$8.5\leqslant b<{}9.5$$．因此，我们得到$$a\times b\geqslant 6.5\times 8.5=55.25$$，$$a\times b<{}9.5\times 7.5=71.25$$．所以两个小数乘积的整数部分可取$$55$$到$$71$$之间的任何整数值，一共有$$17$$种可能的取值.	

作者简历及攻读硕士/博士学位期间取得的研究成果

一、作者简历

冯梦菲，女，1995 年 4 月生，2013 年 9 月至 2017 年 7 月就读于北京交通大学通信工程专业，取得工学学士学位。2017 年 9 月至 2020 年 6 月就读于北京交通大学通信与信息系统专业，研究方向是信息网络，取得工学硕士学位。攻读硕士学位期间，主要从事相似数学习题推荐方面的研究工作。

二、发表论文

[1] Feng M, Chen Y, Guo Y, et al. Learning Text Representations for Finding Similar Exercises[C]. international conference on consumer electronics, 2019.

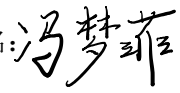
三、参与科研项目

[1] 基于大规模在线视频流系统的会话级 QoE 预测

[2] 基于深度学习的习题理解和应用算法研究

独创性声明

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

学位论文作者签名：  签字日期：2020 年 6 月 7 日

学位论文数据集

表 1.1： 数据集页

关键词*	密级*	中图分类号	UDC	论文资助
寻找相似习题；相似习题推荐；深度学习	公开			
学位授予单位名称*		学位授予单位代码*	学位类别*	学位级别*
北京交通大学		10004	工学	硕士
论文题名*		并列题名		论文语种*
基于深度学习的习题理解和应用算法研究				中文
作者姓名*	冯梦菲		学号*	17120052
培养单位名称*		培养单位代码*	培养单位地址	邮编
北京交通大学		10004	北京市海淀区西直门外上园村 3 号	100044
学科专业*		研究方向*	学制*	学位授予年*
通信与信息系统		信息网络	3	2020
论文提交日期*	2020.6.8.			
导师姓名*	陈一帅		职称*	副教授
评阅人	答辩委员会主席*		答辩委员会成员	
	郭宇春		赵永祥 郑宏云 张立军 孙强	
电子版论文提交格式 文本（ ） 图像（ ） 视频（ ） 音频（ ） 多媒体（ ） 其他（ ） 推荐格式：application/msword；application/pdf				
电子版论文出版（发布）者		电子版论文出版（发布）地		权限声明
论文总页数*	65 页			
共 33 项，其中带*为必填数据，为 21 项。				