规则组合实现大语言模型的数字理解能力

陈钶杰
学号: 22135030

February 27, 2024

1 摘要(Abstract)

以前的研究通常假设大型语言模型在没有计算器工具的情况下无法准确执行算术运算。本文旨在挑战这种误解。通过全新的子规则和复合规则的灵活运用，就如同数学中我们使用最简单的一组基底来表示所有表达式。仅使用一万条训练数据集，语言模型在执行复杂的算术运算时几乎能够达到100%的准确率，而且没有数据泄露，其数学计算能力明显优于现有的语言模型。所使用的规则学习方法不仅在数学问题上适用，对于其他需要严格逻辑推理的问题同样可以借助我们的规则学习方法完成。

2 引言(introduction)

在自然语言处理领域的最新进展中，如GPT-4等大型语言模型已经取得了显著的成功，展示了在各种任务中惊人的语言理解能力。然而，在处理数字问题等专业领域时，这些模型仍然面临挑战。这些模型在理解数字时，主要通过背诵和记忆来理解，而不是深刻理解数字计算本身的规则。本文的研究旨在深入探讨大型语言模型在数字领域中的规则学习，以提高模型在这一领域的性能。

大型语言模型如GPT-4已经展示了在自然语言处理领域各种下游任务中出色的能力[1, 4, 43, 33, 45, 27] 开创性的模型，如GPT-4 [24] 和ChatGPT [23]，已经在大量文本数据上进行了训练，使它们能够生成连贯且上下文相关的响应。它们理解和生成文本的能力使它们在各种自然语言处理任务中非常灵活。此外，LLMs已经被应用于其他领域，涉及数学[5, 17] 和科学[32] 等方面的任务。然而，尽管在各种自然语言处理任务中展现出卓越的能力，GPT-4在数学推理，包括算术任务和中文数学应用问题的处理中可能表现不同水平的熟练度。在算术任务的背景下，一个普遍的假设是LLMs在准确执行复杂算术运算方面存在困难。为了消除这些误解，我们进行了一项调查，评估LLMs的算术能力。具体来说，我们关注LLMs在执行复杂算术运算方面的能力。因此，我们提出了MathRuleGLM，经过精心设计，可以完美执行广泛范围的复杂算术运算，相对于诸如GPT-4等领先的LLMs表现出最佳性能。这些运算包含了单一操作，如加法、减法。重要的是，MathRuleGLM只需要学习一套最基本的规则就有能力灵活处理涉及各种数值形式的算术运算.

(使用图2展示一下我们模型的能力)

为了实现MathRuleGLM在算术任务中展现出的显著性能，我们只用了一套简单的规则构建一个算术数据集，作为MathRuleGLM预训练的基础。该数据集仅包含的一些基本算术以及特定的计算规则，MathRuleGLM学会简单和复杂的算术规则表达式，使其能够准确执行超过8位数字的加、减运算。结果表明，MathRuleGLM的算术性能甚至超过了最强大的LLMs，如GPT-4。具体而言，MathRuleGLM在包含加,减法运算的测试数据集上达到了惊人的(100)%的准确率。相比之下，GPT-4在同一数据集上仅能达到(83)%的准确率。

总体而言，MathRuleGLM可以用最少量规则在算术任务中有最佳的表现。我们的综合实验详细展示了MathRuleGLM相对于GPT-4等语言模型的数学推理的方法的独特性和准确性。这些结果显著挑战了LLMs在复杂算术任务中存在困难的常见误解，揭示了它们在数学推理任务领域卓越潜力。

3 相关工作(relate work)

3.1 背景介绍(back ground)

大语言模型通过大规模数据学习语言规则，但在数学计算中，数字之间的运算是无穷无尽的，但是数字计算是有明确规则的,数字运算对于模型而言是一项具有挑战性的任务。当前的语言模型在这方面仍然存在一些限制，因此有必要进一步研究数字规则学习的方法，以增强其在数字运算中的适用性。

3.2 先前研究综述

早期的已经有不少研究集中在使用大型语言模型在解决数字计算。比如ComputeGPT[3]通过借助python工具来进行数字的计算,这种方法的好处是计算能力强大,且计算的数值不受任何位数的影响,但是这种方法语言模型只是起到了一个文本识别器,将数字,运算符号识别出来,真正文献进行运算的是python,而语言模型本身是不会进行数字计算的.像程序辅助语言模型[1]也是类似的,这是一种新的自然语言推理方法，使用程序作为中间推理步骤。与现有的基于LLM的推理方法不同，主要思想是将求解和计算offload到外部Python 解释器，而不是使用LLM 来理解问题和求解。像MetaMath[5],让语言模型去执行算术和符号推理的任务,通过大量计算题数据的训练,来实现对一些问题的计算,但这种方式更像是背题,而不是深刻理解其中数学规律和算法.MathGPT[4]是一种大型语言模型在不使用计算器工具的情况下无法准确执行算术运算，在足够的训练数据下，20亿参数的语言模型可以准确地进行多位算术运算，此文的核心思路是尽可能的学习5位数以内的数学计算规则,通过这些规则的学习,对高位数运算进行学习.与利用外部工具不同，我们专注于探索如何在不依赖外部工具的情况下增强llm固有的算术能力,但是这个模型并没有让模型真正理解数字运算的基本规律.

3.3 比较和对比

我们的模型同样专注于探索如何在不依赖外部工具的情况下增强llm固有的算术能力,但不同之处在于我们更注重语言模型如何去理解数学计算,而不是通过背诵的方式来解决数学问题.比如对于加法,我们的模型只需要学会99加法表以及加法的进位,对齐规则就可以计算任意位数的加法.这种从加法的数学出发进行计算不仅能够保证计算的准确率,而且更加体现神经网络的泛化能力.

3.4 研究趋势和发展

MLC[2]说明神经网络模型是具有规则学习能力,且具有泛化能力,而目前语言模型在数字计算主流解决方法都是通过采用大量数据训练来提高准确性,其中通过优化训练数据提高最终模型的准确性.

3.5 未解决的问题

通过大规模数据的训练这样虽然能让语言模型掌握一定数学计算能力,但是对于其中计算具有很强的随机性,即使通过使用高质量数据能对模型性能有一定的提升,但是很难有质的改变.因此我们希望能够让语言模型通过学习数学规则来提高数字计算的准确性.只要学会规则,那么模型就能够以极高的准确率去解决各种计算问题.这样是非常符合我们预期的.

3.6 方法和技术的演进

transformer模型在文本处理上依旧是最优秀的,我们使用基本的transformer模型框架,将规则分为基本数学计算规则以及进位,对齐等复合规则,同时进行训练,最终模型在经过训练以后测试其数字运算的能力.

4 研究方法(Methodology)

为了调查语言模型在数学推理中的有效性，我们提出了MathRuleGLM模型，其旨在专注于提升语言模型在数学推理中的性能。首先，MathRuleGLM致力于使用规则来执行准确的算术任务。它通过在其架构中整合数学计算的基本规则并进行规则的复合来实现这一目标。与直接计算复杂算术表达式不同，MathRuleGLM采用这种策略，使用规则的计算方法更接近人类解决数学问题的方式。其次，MathRuleGLM用更少的训练数据来训练模型达到相同的计算结果，以解决复杂的数学问题。通过利用这一策略，MathRuleGLM不仅是在数学问题上可以应用,在其他可以通过子规则,复合规则的问题都可以解决.

4.1 学习算术任务(Learning on Arithmetic Tasks)

算术任务可以广泛分为基本规则调用和复杂规则调用。基本算术运算包括围绕进行涉及两个数字的简单计算的基本数学任务。另一方面，算术任务还包括复杂混合运算的领域，这需要处理各种算术运算和数字格式的组合的技能。MathRuleGLM 涵盖的学习任务的综合类别总结在表1中。

Task	Interger
Addition	x + y
Subtraction	x - y
Multiplication	x * y
Division	x / y

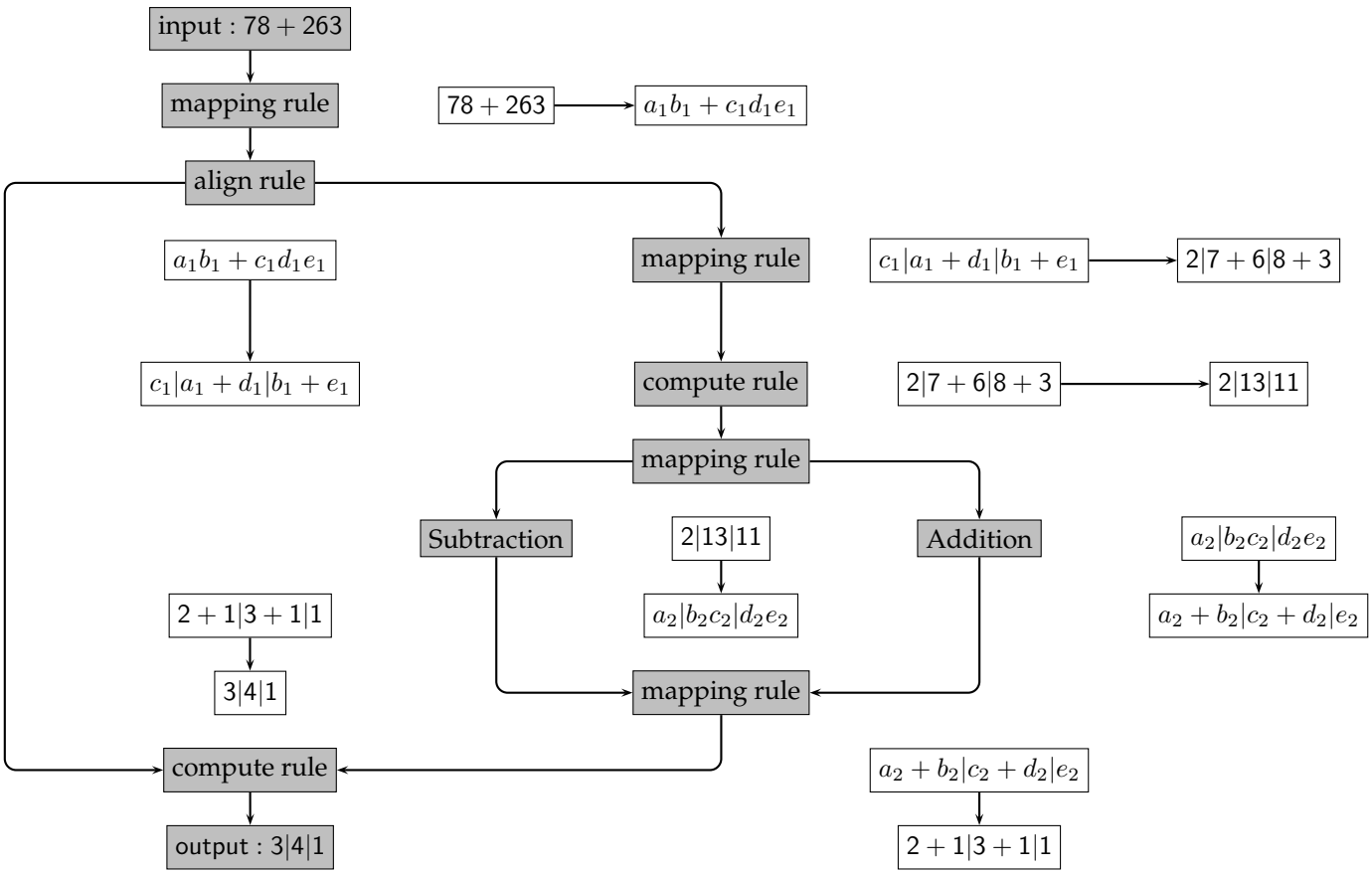

Figure 1: 如何调用规则求解算式

算术训练数据集

用于训练的算术数据集经过精心设计，数据极其精简,但又涵盖全面的算术任务范围。这个数据集经过周到设计，包含各种子规则比如10以内加法表和复合规则如对齐,进位规则。在这些数据集的每一个中，单个算术表达式包括加法（+）、减法（-）数学运算。对于其他算术表达式后续可以添加类似规则来实现.为了与人类的计算习惯相一致，在构建算术数据集时使用的规则均采用了人类进行算术的基本规则。这个策略是将一个复杂算术表达式使用一系列子规则和复合规则来生成最终的答案。这种策略反映了人们在解决复杂算术任务时通常遵循的过程。通过在这样的数据集上进行训练，MathRuleGLM在算术性能上取得了出色的表现，因为它从详细的计算过程中学到了潜在的计算规则。figure 1 提供了一些从算术数据集中提取的训练示例，展示了算术任务的如何利用已学得规则进行计算。

模型和训练过程

表2报告了具有不同模型参数的所有模型的概览。我们的训练工作涵盖了4种不同类型的模型，每种模型具有不同的参数大小。最大的模型具有20亿个参数，使其在容量方面成为最强大的模型。接着，我们训练第二个模型，它有5亿个参数，第三个模型有1亿个参数，最小的模型有1000万个参数。值得注意的是，尽管参数大小有差异，所有模型都使用包含5000万条训练记录的相同数据集规模进行训练。关于MathRuleGLM的分词技术的技术细节见附录.

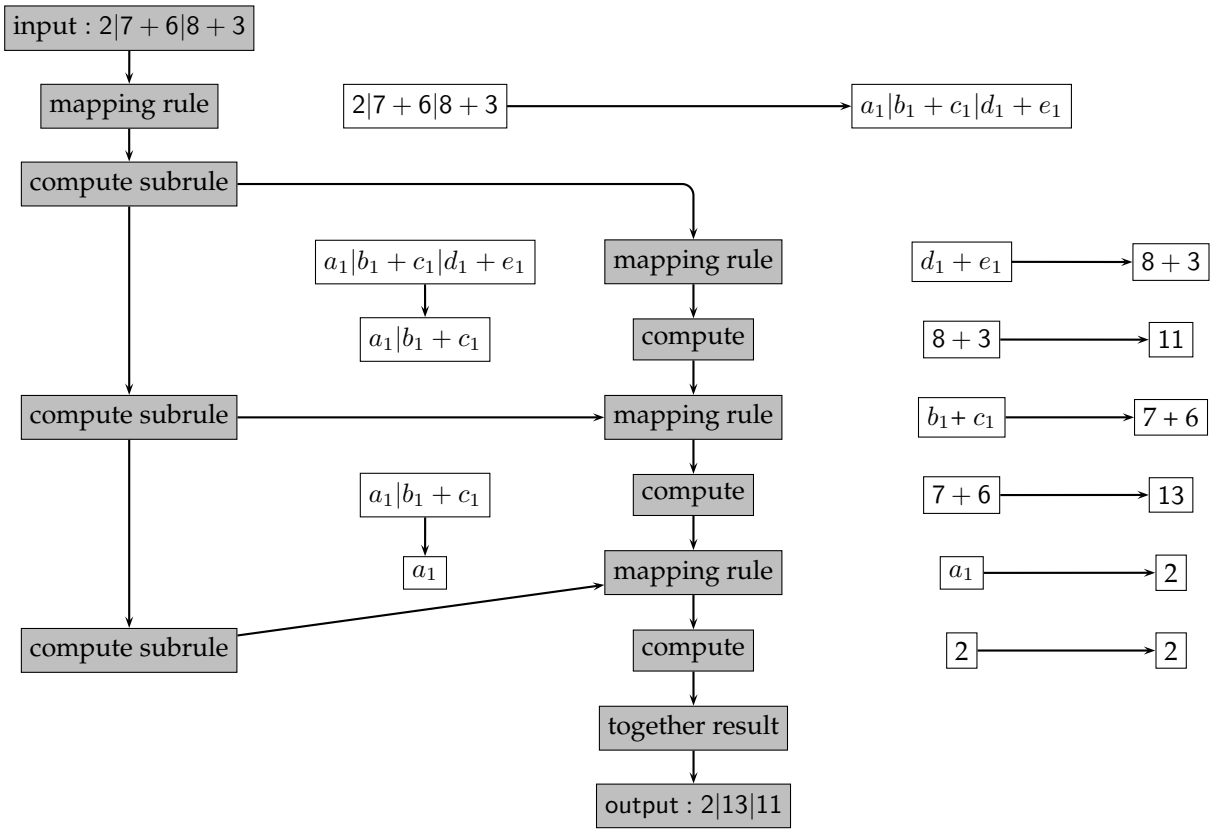

Figure 2: compute rule

Table 1: 不同参数量的模型结果对比

Size	$0-10K$	$10K-1M$	$1M-1B$	$1B-10B$	$10B-70B$	$70B-\infty$
GPT-3.5	-	-	-	-	-	95.29%
GPT-4	-	-	-	-	-	98.32%
llama2-7b	-	-	-	2.3%	-	-
llama2-13b	-	-	-	-	21.89%	-
llama2-70b	-	-	-	-	76.09%	-
Google-PaLM	-	-	-	-	-	95.96%
Qwen-72b-Chat	-	-	-	-	-	93.94%
MathRuleGLM	100%	100%	100%	100%	100%	100%

5 实验(experiments)

MathRuleGLM的总体目标围绕展示语言模型在数学推理领域的能力。为了验证这一点，我们设计了一种复杂算术计算实验，将普通算术问题进行分类,专门针对先有语言模型计算不准确的问题进行测试比较,我们的模型能全面覆盖当前模型在部分计算问题上的不足，为评估模型在数学推理中的熟练程度提供了强有力的评估。

5.1 学习算术

5.1.1 评估指标

为了衡量MathRuleGLM在算术任务上的能力，我们采用以下指标来评估输出结果。准确度通常通过比较MathRuleGLM的输出和实际答案来衡量。相对误差是用来评估MathGLM有效性的另一个重要指标，它量化了MathRuleGLM生成的输出与正确答案之间的差异。

5.1.2 结果与分析

对于算术任务，我们预先训练了一个名为MathRuleGLM的基于Transformer的模型，其在预训练和推理阶段都具有()个模型参数。为了准确评估MathGLM的有效性，我们将其性能与领先的大型语言模型（LLMs）如GPT-4和ChatGPT进行对比。MathRuleGLM在处理算术任务的各个方面上都表现优越，胜过所有其他模型。而且MathRuleGLM是一个非常小的模型变体，即只有()万参数。尽管其紧凑的参数大小，我们的模型在特定算术任务上表现依旧优于GPT-4和ChatGPT。这个惊人的结果展示了规则调用的有效性，该方法涉及将复杂的算术表达式使用复合规则来解决问题，赋予了它识别和理解算术任务中微妙之处的能力。它有效地学习了算术操作的基本规则和原理，使其能够生成准确而精确的解决方案。对于模型在不同参数规模下的表现，我们观察到MathGLM的算术性能与其参数数量的增加直接相关。这一发现表明随着模型大小的增加，其性能呈现相应的增强。总的来说，在复杂的算术任务评估结果中，MathRuleGLM的表现异常出色。通过子规则和复合规则的学习,这些模型很大程度上弥补了GPT-4和ChatGPT的一些数学计算不足之处. 去比对当前的llama模型等模型的计算能力

- 3种特定数据集类型
  - 进位加法:
    1. 48551+1449=
    2. 7223+32777=
  - 负差:
    1. 1140-26787=
    2. 234-15579=
  - 逐位差分
    1. 41085-80976=
    2. 65570-73618=

5.1.3 割除研究

规模分析。为了全面评估模型参数和训练数据规模对性能的影响，我们进行了一系列规模分析实验。(增加一个关于参数量和结果准确度的表格,也可以用图象的方式表示出来)

Table 2: 3种特定的数据集实验结果

Task	进位加法	负差	逐位差分
GPT-3.5	97.2%	99.4	95.29%
GPT-4	100%	97.8%	98.32%
llama2-70b	19.4%	50.2%	76.09%
llama2-13b	5%	20.8%	21.89%
llama2-7b	3.6%	4%	2.3%
Google-PaLM	52.6%	43.6%	95.96%
Qwen-72b-Chat	85.8%	86.4%	93.94%
MathRuleGLM	100%	100%	100%

5.2 针对别人计算不准确的问题进行比较,比如做一些减法问题

用我们的模型以及这些模型进行对比计算:43423-32424

- llama-2-7b:11001
- llama-2-13b:11001
- llama-2-70b:-9001
- Claude-instant-100k:10939
- Google-PaLM:11001
- Qwen-72b-Chat:11009
- Mistral-Medium:77595
- chatgpt3.5:10999
- MathRuleGLM-8k:10999

5.3 使用规则学习去处理其他逻辑推理问题

References

[1] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. *arXiv preprint arXiv:2211.10435*, 2022.

[2] Brenden M Lake and Marco Baroni. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985):115–121, 2023.

[3] Ryan Hardesty Lewis and Junfeng Jiao. Computegpt: A computational chat model for numerical problems. *arXiv preprint arXiv:2305.06223*, 2023.

[4] Zhen Yang, Ming Ding, Qingsong Lv, Zhihuan Jiang, Zehai He, Yuyi Guo, Jinfeng Bai, and Jie Tang. Gpt can solve mathematical problems without a calculator. *arXiv preprint arXiv:2309.03241*, 2023.

[5] Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.