CUT THE CRAP

论文实验复现

马斌

2025. 7. 23

目录

马斌

中国科学技术大学

University of Science and Technology of China

1. 论文概要

1. 论文概要

- 概要：图技术在多智能体协调领域发挥重大作用，本论文针对多智能体的协调拓扑优化，提出了一种经济、简单、健壮的多智能体通信框架 AgentPrune 。
- 背景：拓扑优化主要可分为三大类(根据 *Graphs Meet AI Agents*)：
  - 边缘重要性测量（先全图，再剪枝）
  - 图自动编码器优化（先打分，再连线）
  - 强化学习（多次尝试，逐步优化）

  AgentPrune 即是通过基于低秩原理的图掩码识别关键的多智能体连接

马斌

中国科学技术大学

University of Science and Technology of China

2. 实验设计

2. 实验设计

- 论文提出待解决的四个问题：
  - （RQ1）AgentPrune 在任务完成和令牌效率方面的表现如何?
  - （RQ2）AgentPrune 能否在不影响性能的情况下降低现有多代理系统的经济成本?
  - （RQ3）AgentPrune 是否能有效防御对代理的对抗性攻击?
  - （RQ4）AgentPrune 对其关键组件或参数的敏感度如何?

- 实验分组：*(根据源代码README部分给出的实验步骤)*
  - 对抗攻击实验（RQ3）
  - 性能与框架集成实验（RQ1+RQ2）
  - 现有框架集成实验（RQ2）
  - 性能与成本对比实验（RQ1）

中国科学技术大学

University of Science and Technology of China

3. 复现过程

3. 复现过程

- 尝试-1
  - - github下载到Lab1，按README进行配置
  - - 尝试用本地qwen-0.5B，需修改大量代码的原设计，AI修改后大量报错。

- 尝试-2
  - - 重新github下载到Lab1
  - - 把各个文件大致结构看一遍。发现大多用的是API调用，放弃使用本地模型。
  - - 尝试用API调用glm-4，仅修改URL和API_KEY，并将项目中有关gpt的name全改为glm-4，无法运行，卡死。
  - - 手动结束运行，发现由于在lab1文件夹下，修改了运行代码中关于文件地址的代码，不会卡死，但报错。
  - - AI根据报错反复修改文件，一直无法正常运行，最终又卡死。
  - - 项目文件大量修改后，感觉可能会与原设计有差别，因此选择重开。

3. 复现过程

- 尝试-3
  - 重新github下载到Lab1，经求助使用openai尝试运行实验
  - 发现问题并解决：
    - ❗ 问题：代码地址错误，无法找到.py文件
    - ✓ 解决：将运行代码地址加上 Lab1/
    - ❗ 问题：运行过程dataset无法正常加载
    - ✓ 解决：将代码中parse_args()中加载地址由默认路径修改为绝对路径
    - ❗ 问题：run_gsm8k.py参数错误传入
    - ✓ 解决：删除未定义的 --no_spatial 参数，成功运行
    - ❗ 问题：运行过程MMLU/data下载与解包过程未知原因卡死
    - ✓ 解决：尝试VPN无果后，选择手动下载到本地，再注释掉download()并将加载路径改为项目文件夹中的绝对路径，成功运行。
  - 成功运行了其中有代表性的4个实验代码，并手动叫停以防使用大量 tokens 。

吕斌

中国科学技术大学

University of Science and Technology of China

4. 结果分析

4. 结果分析

Lab1-1

```
(/data0/bma/env/agentprune) amax406% python Lab1/experiments/run_mmlu.py --agent_nums 1 --mod
e DirectAnswer --decision_method FinalMajorVote --agent_names AdverarialAgent --batch_size 4

Number of topics:  57
Total number of questions:  285
Number of topics:  57
Total number of questions:  1531
Evaluating AgentPrune on MMLUDataset split val
  0%|                                                                          | 0/39 [00:00<?, ?it/s]
--------------------------------------------------------------------------------
Batch time 29.656
Raw answer: ['C']
Postprocessed answer: C
Correct answer: C
Accuracy: 100.0% (1/1)
Raw answer: ['D']
Postprocessed answer: D
Correct answer: D
Accuracy: 100.0% (2/2)
Raw answer: ['A']
Postprocessed answer: A
Correct answer: A
Accuracy: 100.0% (3/3)
Raw answer: ['C']
Postprocessed answer: C
Correct answer: C
Accuracy: 100.0% (4/4)
  3%|                                                                        | 1/39 [00:29<18:46, 29.66s/it]
--------------------------------------------------------------------------------
```

运行命令分析:

--agent_nums 1

一个智能体

--mode DirectAnswer

直接给出答案

--decision_method FinalMajorVote

通过多数投票来决定最终答案

--agent_names AdverarialAgent

代理类型为对抗性智能体

--batch_size 4

每次将处理 4 个样本

输出结果分析:

---对抗攻击实验（RQ3）

对应文章结论“单对抗性智能体（AdversarialAgent）会导致系统性能显著下降”，说明其对单一恶意节点的防御效果。

4. 结果分析

Lab1-2

```
(/data0/bma/env/agentprune) amax406% python Lab1/experiments/run_mmlu.py --mode FullConnected
  --batch_size 4 --agent_nums 5 --num_iterations 200 --imp_per_iterations 200 --pruning_rate 0
.5 --num_rounds 1 --optimized_spatial
Number of topics:  57
Total number of questions:  285
Number of topics:  57
Total number of questions:  1531
Iter 0 ---------------------------------------------------------------------------
{'task': 'All other things being equal, which of the following persons is more likely to show
 osteoporosis?\nOption A: An older Hispanic American woman\nOption B: An older African Americ
an woman\nOption C: An older Asian American woman\nOption D: An older Native American woman\n
'}
{'task': 'At which stage in the planning process would a situation analysis be carried out?\n
Option A: Defining the program\nOption B: Planning the program\nOption C: Taking action and i
mplementing ideas\nOption D: Evaluation of the program\n'}
{'task': 'Which of the following is considered an acid anhydride?\nOption A: HCl\nOption B: H
2SO3\nOption C: SO2\nOption D: Al(NO3)3\n'}
{'task': 'In a genetic test of a newborn, a rare genetic disorder is found that has X-linked
recessive transmission. Which of the following statements is likely true regarding the pedigr
ee of this disorder?\nOption A: All descendants on the maternal side will have the disorder.\
nOption B: Females will be approximately twice as affected as males in this family.\nOption C
: All daughters of an affected male will be affected.\nOption D: There will be equal distribu
tion of males and females affected.\n'}
correct answer:C
correct answer:A
correct answer:C
correct answer:C
raw_answers: (['C'], ['B'], ['C'], ['C'])
answers: ['C', 'B', 'C', 'C']
Batch time 82.476
utilities: [1.0, 0.0, 1.0, 1.0]
```

运行命令分析:

--mode FullConnected
配置图为全连接模式（全图）
--batch_size 4
--agent_nums 5
--num_iterations 200
优化迭代总次数为200
--imp_per_iterations 200
每200次迭代进行一次剪枝
--pruning_rate 0.5
剪枝比为0.5
--num_rounds 1
每个查询的优化/推理轮次为1
--optimized_spatial
启用空间图优化

输出结果分析:

--- 性能与成本对比实验（RQ1）
对应文章结论“AgentPrune在全图结构上，以仅$5.6的成本实现了与先进拓扑（成本$43.7）相当的MMLU性能”，证明其在通用推理任务中能大幅降低冗余通信。

4. 结果分析

Lab1-3

```
(/data0/bma/env/agentprune) amax406% python Lab1/experiments_autogen/run_humaneval.py --optim
ized_spatial --optimized_temporal

Batch 0 --------------------------------------------------------------------------------
Batch time 79.367
Accuracy: 1.0
utilities: [True, True, True, True]
Batch 1 --------------------------------------------------------------------------------
Batch time 66.890
Accuracy: 0.875
utilities: [True, True, False, True]
Batch 2 --------------------------------------------------------------------------------
db1111db
db100000db
Batch time 77.593
Accuracy: 0.9166666666666666
utilities: [True, True, True, True]
Batch 3 --------------------------------------------------------------------------------
Batch time 78.940
Accuracy: 0.8125
utilities: [True, True, False, False]
```

运行命令分析:

--optimized_spatial

启用空间图结构优化

--optimized_temporal

启用时间图结构优化

输出结果分析:

--- 现有框架集成实验（RQ2）

对应文章实验设计“将AgentPrune集成到AutoGen等主流框架，验证其经济性”，针对humaneval数据集在AutoGen框架中启用AgentPrune的空间和时间优化，验证其在代码生成任务中对现有框架的成本降低效果。

4. 结果分析

Lab1-4

```
(/data0/bma/env/agentprune) amax406% python Lab1/experiments_autogen/run_gsm8k.py --optimized
_spatial --optimized_temporal
Batch 0 --------------------------------------------------------------------------------
Batch time 101.899
Accuracy: 1.0
utilities: [True, True, True, True]
```

运行命令分析:

--optimized_spatial

启用空间图结构优化

--optimized_temporal

启用时间图结构优化

输出结果分析:

--- 现有框架集成实验（RQ2）

针对GSM8K数据集在AutoGen框架中启用AgentPrune的空间和时间优化，验证其在数学推理任务中对现有框架的成本降低效果。

中国科学技术大学

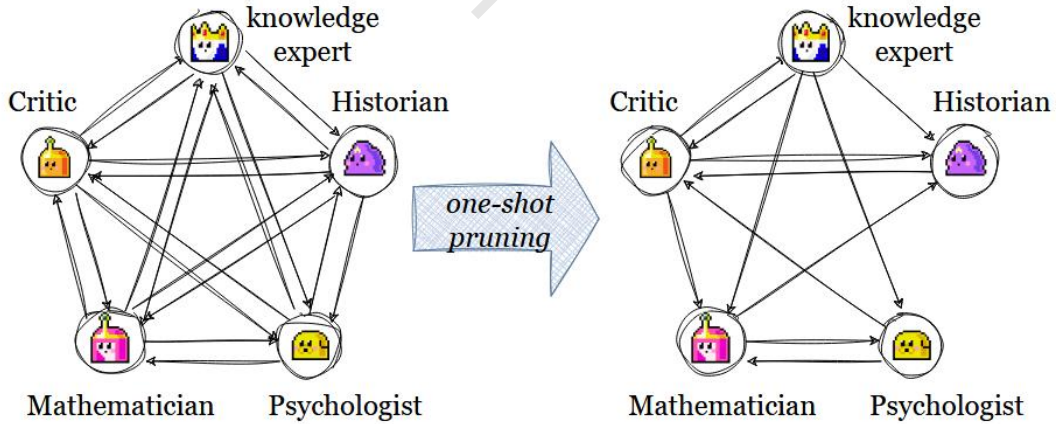
University of Science and Technology of China

5. 总结

5. 总结

文章通过在六个基准数据集上的大量实验（本人显然只复现了其中的极小部分实验），得出结论：AgentPrune 作为一种经济、简单且稳健的多智能体通信框架，能够无缝集成到主流多智能体系统中，有效修剪冗余甚至恶意的通信消息。具体而言，其表现为三个方面：

- 性能上，以仅 5.6 美元的成本实现了与先进拓扑（成本 43.7 美元）相当的结果；
- 经济性上，能与现有多智能体框架无缝集成，实现 28.1%～72.8% 的 token 减少；
- 对抗鲁棒性上，成功防御两种基于智能体的对抗性攻击，带来 3.5%～10.8% 的性能提升。



谢谢！

马斌

2025. 7. 23