MAJOR PROJECT SYNOPSIS

Title: Solve Maths Word Problems using AI

Department of Computer Engineering

Faculty of Engineering and Technology

JAMIA MILLIA ISLAMIA UNIVERSITY

Project By – Debal Hussain Abbas (18BCS046)

Mauwaz Ahmed Farooqui (18BCS048)

ABSTRACT

State-of-the-art language models can match human performance on many tasks, but they still struggle to robustly perform multi-step mathematical reasoning. To diagnose the failures of current models and support research, we use GSM8K, a dataset of 8.5K high quality linguistically diverse grade school maths word problems. We find that even the largest transformer models fail to achieve high test performance, despite the conceptual simplicity of this problem distribution. To increase performance, we propose training verifiers to judge the correctness of model completions. At test time, we generate many candidate solutions and select the one ranked highest by the verifier. We demonstrate that verification significantly improves performance on GSM8K, and we provide strong empirical evidence that verification scales more effectively with increased data than a fine-tuning baseline.

INTRODUCTION

Large language models like GPT-3 have many impressive skills, including their ability to imitate many writing styles, and their extensive factual knowledge. However, they struggle to perform tasks that require accurate multistep reasoning, like solving grade school mathematics word problems. Although the model can mimic the pattern of correct solutions, it regularly produces critical errors in logic.

To match human performance in complex logical domains, the models must learn to recognize their mistakes and to choose their steps carefully. To that end, we train verifiers to evaluate whether or not a proposed solution is correct. To solve a new problem, we use verifiers to select the best among many proposed solutions. We use the new GSM8K dataset to evaluate our methods.

In the examples below, we show a solution generated by our new method

Question

Ali is a dean of a private school where he teaches one class. John is also a dean of a public school. John has two classes in his school. Each class has 1/8 the capacity of Ali's class which has the capacity of 120 students. What is the combined capacity of both schools?

Answer

Method: **175B Verification**

```
Ali's class has a capacity of 120 students.
Each of John's classes has a capacity of
120/8 = 15 students.
The total capacity of John's two classes is
15 students * 2 classes = 30 students.
The combined capacity of the two schools is
120 students + 30 students = 150 students.
```

About the Dataset

GSM8K consists of 8.5K high quality grade school maths word problems. Each problem takes between 2 and 8 steps to solve, and solutions primarily involve performing a sequence of elementary calculations using basic arithmetic operations (+ − × ÷) to reach the final answer. Fine-tuned state-of-the-art language models perform poorly on this dataset, primarily due to the high diversity of problems. At the same time, GSM8K solutions depend only on elementary concepts, so achieving high test performance is an achievable goal.

Solutions in GSM8K are written as natural language rather than as pure maths expressions. By sticking to natural language, model-generated solutions are more readily interpretable by humans.

Methodology : Fine Tuning and Training Verifiers

We investigate two methods to solve problems in GSM8K: finetuning and verification. Fine-tuning, our baseline method, uses the same language modelling objective as the generative pretraining in GPT-3.

One significant challenge in mathematical reasoning is the high sensitivity to individual mistakes. Autoregressive models, which generate each solution token by token, have no mechanism to correct their own errors. Solutions that veer off-course quickly become unrecoverable, as can be seen in the examples provided.

We address this problem by training verifiers to evaluate the correctness of model-generated solutions. Verifiers are given many possible solutions, all written by the model itself, and they are trained to decide which ones, if any, are correct.

To solve a new problem at test time, we generate 100 different probable solutions and then select the solution that is ranked highest by the verifier. Verifiers benefit from this inherent optionality, as well as from the fact that verification is often a simpler task than generation.

We find that we get a strong boost in performance from verification, as long as the dataset is large enough. With datasets that are too small, we believe that the verifiers overfit by memorising the final answers in the training set, rather than learning any more useful properties of mathematical reasoning.

On the full training set, 6B parameter verification slightly outperforms a fine-tuned 175B parameter model, giving a performance boost that is approximately equivalent to a 30x model size increase. Moreover, verification appears to scale more effectively with additional data, if we extrapolate based on current results.

CONCLUSION

Producing correct arguments and recognizing incorrect ones are key challenges in developing more general AI. Grade school maths is an ideal testbed for these capabilities. The problems in GSM8K are conceptually simple, yet a slight mistake is enough to derail an entire solution. Identifying and avoiding such mistakes is a crucial skill for our models to develop. By training verifiers, we teach our models to separate the good solutions from the ones that are not so accurate.

REFERENCES


```
@article{cobbe2021gsm8k,


    title = "Training Verifiers to Solve Math Word Problems”,

    author = "Cobbe, Karl and Kosaraju, Vineet and Bavarian, Mohammad
    and Chen, Mark and Jun, Heewoo and Kaiser, Lukasz and Plappert,
    Matthias and Tworek, Jerry and Hilton, Jacob and Nakano, Reiichiro
    and Hesse, Christopher and Schulman, John”,

    year = "2021",

    publisher = "Association for Computational Linguistics",

}
```