Formatting Instructions For NeurIPS 2023

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Chloe

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

A short description of your goals, task, model, and (for the final report) results. The abstract should make the motivations and the scope of your project clear so that readers can decide whether they are interested in reading your work.

1 Introduction

Gabriel

A description of the motivation behind your work, why the task you chose is interesting/important, and a summary of your (proposed) approach. The problem that you want to solve should be clearly stated in the introduction: especially the input and output of your model and the format of the input and output. This section should also make it clear why your deep learning approach is reasonable for this problem.

2 Background and related work

Chloe

A summary of the background material that students of CSC413 would not already be familiar with. A description of related work done in the area, and how your approach compares with theirs.

If your project builds on previous work, clearly distinguish what they did from what your new contribution is. Also, include a 1-2 sentence summary of other closely related papers. We realize you might not know about all related papers (or have time to carefully read all related papers), and that's OK for this project. Using bibtex is annoying at first, but Google Scholar can give you the bibtex entries.

3 Data

Taha

The dataset used in your model. Include any key exploratory figures that will help readers evaluate the difficulty of your problem and interpret the performance of your model.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

3.1 Datasets

For this project, we decided to use the following datasets:

Math Dataset

The MATH dataset consists of 12,500 (7,500 training and 5,000 test) problems from mathematics competitions including the AMC 10, AMC 12, AIME, and more. Many of these problems can be collected from aops.com/community/c3158_usa_contests. [1]

NuminaMath-CoT

The Numina Math CoT dataset has approximately 860k math problems, where each solution is formatted in a Chain of Thought (CoT) manner. The sources of the dataset range from Chinese high school math exercises to US and international mathematics olympiad competition problems. [2]

NuminaMath-TIR

The Numina Math TIR dataset is a more specific version of the CoT dataset, where 70k problems are selected, with a focus on numerical outputs and integers. Tool-integrated reasoning (TIR) plays a crucial role in this dataset, where the solution to each problem is a sequence of steps that can be executed by a computer program. [3]

3.2 Data Formatting

Problems and solutions are formatted using LATEX. The usage of LATEX ensures that the data is easily readable and can be used to generate math problems and solutions. The data is also formatted in a way that is easy to parse and process for training the model.

The data for the Math Dataset is formatted as a JSON file, with each problem containing the following fields:

- problem: The text of the problem
- level: The difficulty level of the problem (Level 1 up to Level 5)
- type: The type of math problem (e.g. algebra, geometry, etc.)
- solution: The solution to the problem

The data for the Numina Math CoT and TIR datasets are formatted as a JSON file, with each problem containing the following fields:

- An array of objects, where the first object contains
  - content: The text of the problem
  - role: the role assigned to the person who can access this data (user)
- The second object contains
  - content: The solution to the problem
  - role: the role assigned to the person who can access this data (assistant)

We intend to use these datasets to train our model to solve mathematical problems. We will preprocess the data into one format, to ensure that can be used by our model. We will also split the test data into validation and test sets to evaluate the performance of our model.

4 Model architecture

Everyone

A description of your (proposed) model architecture. Please propose an architecture during the proposal phase, but it's okay to change your architecture. In the final report, this section should have enough details to reproduce the work, including all hyperparameters and 3 training settings that you used.

Selected model: PaLM2
Google PaLM2 = transformers + modifications: https://arxiv.org/pdf/2204.02311

Attempt to combine this with PaLM2:
SympyGPT: Transformers for symbolic integration proofs: https://arxiv.org/html/2410.02666v1

Better with word problems? Architecture: PaLM, GPT4: http://research.google/blog/minerva-solving-quantitative-reasoning-problems-with-language-models/

We could also combine the two models (unlikely but look into it):

Standard Transformers Architecture: https://arxiv.org/abs/1706.03762

5 Model architecture figure

Takia

A figure that helps show the overall model or idea. The idea is to make your paper more accessible, especially to readers who are starting by skimming your paper. You must create a new figure, not just use someone else's, even with attribution. Be careful that all figure text are legible, and are approximately the same size as the main text.

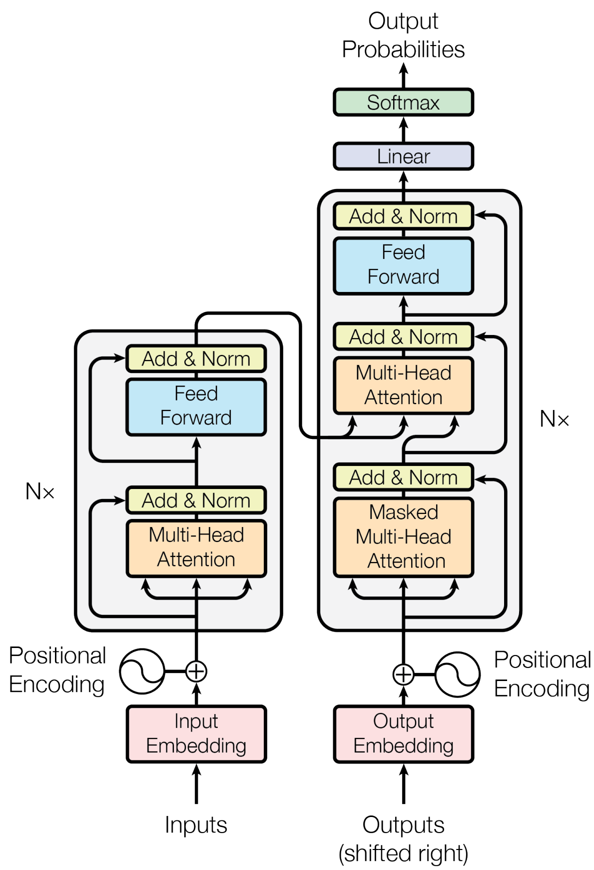

Figure 1: The Transformer - model architecture.

6 Ethical considerations

Taha

Potential ethical issues posed by the use or misuse of your model. Your report should transparently communicate the known or anticipated consequences of building and using machine learning models on this task.

https://neurips.cc/public/EthicsGuidelines

6.1 Ethical Issues

One ethical consideration is the potential for the model to be used to cheat on math assignments/homework. This could encourage students to use the model as a shortcut rather than engaging with the material themselves, which would negatively impact a students capacity for critical thinking and problem-solving.

Another ethical issue is intellectual property rights, as the model is trained on copyrighted data. The model could inadvertently promote plagiarism if it directly provides solutions that students submit as their own, without understanding the learning process. This could lead to academic dishonesty and undermine the integrity of the educational system.

6.2 Mitigation Strategies

To mitigate the first risk, the model should be used as a learning tool rather than a tool for cheating. For example, the model could be used to generate practice problems for students to solve, or to provide explanations for the solutions to problems. This would help students learn and improve their math skills, rather than using the model to cheat. Therefore, while automation can support learning, it is essential that it complements, rather than replaces, active engagement with the educational process.

To mitigate the second risk, the model should be used in a controlled environment where students are guided on how to use the model appropriately. For example, teachers could provide guidelines on how to use the model to check answers or generate practice problems, rather than using it to directly provide solutions. Furthermore, it is crucial to emphasize the value of understanding the material and using tools as aids rather than shortcuts. This would prevent plagiarism and encourage students to engage with the material and develop their problem-solving skills.

7 Work division

A description of how the work will be divided between the team members, and how the team members will be working together (e.g. meet every week Tuesday 4-5 pm).

Chloe - Background and Related Work, Abstract, Model Architecture

Takia - Model Architecture Figure, Model Architecture

Gabriel - Introduction, Model Architecture

Taha - Data, Ethical Considerations, Model Architecture

The team will work on each part, and meet every weekend for additional discussions.

References

[1] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

[2] Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. `[https://huggingface.co/AI-MO/NuminaMath-CoT](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf)`, 2024.

[3] Lewis Tunstall Ben Lipkin Roman Soletskyi Shengyi Costa Huang Kashif Rasul Longhui Yu Albert Jiang Ziju Shen Zihan Qin Bin Dong Li Zhou Yann Fleureau Guillaume Lample Jia LI, Edward Beeching and Stanislas Polu. Numinamath tir. `[https://huggingface.co/AI-MO/NuminaMath-TIR](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf)`, 2024.