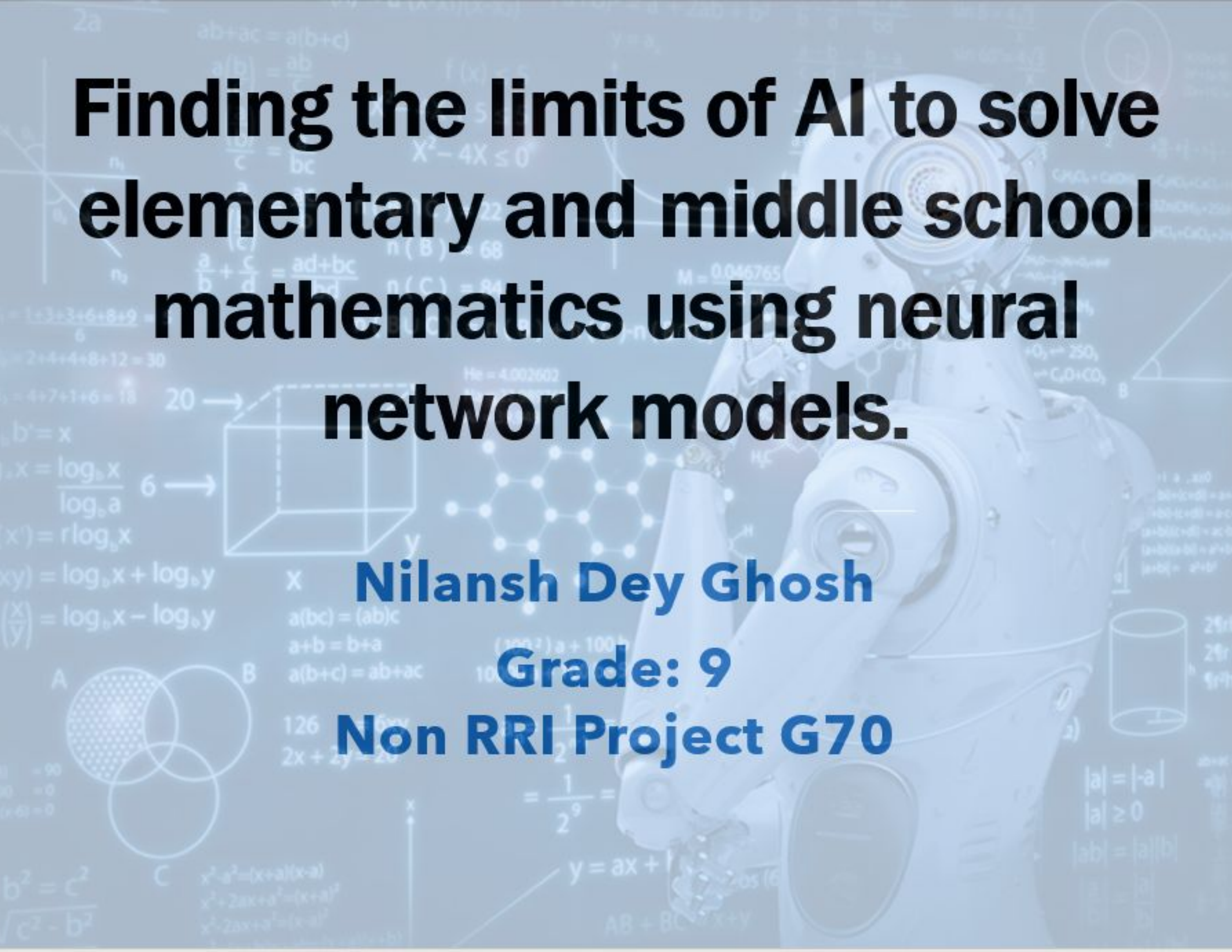

Finding the limits of AI to solve elementary and middle school mathematics using neural network models.

Nilansh Dey Ghosh

Grade: 9

Non RRI Project G70

Introduction

- Goal:
  Design a mechanism to test and find the limits of modern AI (Artificial Intelligence) systems to learn and solve mathematical problems, so that we can understand how far AI can simulate a typical student's brain in elementary and middle school.

- Previous Research:
  - Deepmind, an AI company, previously generated a mathematical dataset to find a machine learning model that would perform the best on school-level math problems.
  - Based on their research, I am recreating the highest performing model (Transformer) and testing its limits to find how accurately it can solve general math problems by experimenting with different subsets in the dataset.
  - In addition to this research, many others have worked on this topic but for more specific problems. For example, in 2014, researchers used a Long Short Term Memory model to predict answers to simple math problems. Another example, is in 2016 when other researchers used an adapted convolutional neural network to solve addition and multiplication .

Methods

- I used Transformer, one of latest ML (Machine-Learning) model used for language translation, to represent an AI system.
- To implement the model in Google Colab, I used the ML libraries Tensorflow and Keras.
- I used Python to generate simple elementary and middle school Mathematics question/answer pairs to form the dataset – based on the implementation described in the paper from Deepmind. This dataset will be used to train the ML model. A separate test dataset (Mathematics questions/answers) will be used to evaluate the accuracy.
- Steps to set up the ML model:
    - Download the source code for generating the mathematical question and answer pairs from GitHub.
    - Generate the mathematical question-answer pairs for question types based on elementary and middle school mathematics e.g., Simple Arithmetic, number comparisons and sorting, measurement, number manipulations.
    - Setup Google Colab and use GPU to train and test the ML model.
    - Setup the Transformer model in Colab.

Methods - Continued

- First, I used the same type of problem to train and test the AI system.
- Then, I used different problem domains for training and testing to evaluate the capability of the AI system to learn from a set of problem domains and transfer that knowledge to solve related problems.

Steps to Train and Test the model:

- Train the model with a specific problem domain or type (e.g. Arithmetic addition) of mathematics dataset.
- Test the model with examples from the same type. Measure accuracy and loss.
- Test the model with examples from different problem types (e.g., Arithmetic multiplication, subtraction, number comparison). Measure accuracy and loss.
- Repeat steps 1-3 with different sets of problem domains (e.g., Train with addition and multiplication, but test with subtraction).

Results

- Original prototype of my Transformer model did not perform as expected
- After making changes to the input text processing and some hyperparamters, my model's accuracy improved

	Train Dataset	Test Dataset	Accuracy of Training (3 epochs)	Accuracy of Testing
1	add/sub	add/sub	98	90
2	add/sub	Add/sub multiple	98	40
3	add/sub	mul/div	98	20
4	Add/sub multiple	Add/sub multiple	98	90
5	mul/div	mul/div	98	90
6	mul/div	add/sub	98	30
7	Num compare	Num compare	98	90
8	Add/sub and mul/div	mixed		

Discussion – Experimental Results

- My model was able to perform well consistently with the same train and test dataset
  - e.g. Train with add/sub - able to solve 2+3;
  - e.g. Train with add/sub_multiple - able to solve 2+3+3
- When the test and train domains were different, there was very poor performance as expected
  - e.g. Train with add/sub but unable to solve 2+3+3 or 2*3
- When attempting to combine knowledge from two domains, my model also performed very poorly
  - e.g. Train with add/sub/mul/div but unable to solve 2*3+4

Discussion – Problems and Issues

- Faced various problems generating the mathematics dataset
    - Fixed issues in the Python code for the dataset generator
    - Fixed issues in loading the dataset onto Colab
- Found proper text pre-processor for the transformer model
    - Experimented with various text processors to fix the skewed results
    - Used the latest text pre-processor library from tf.keras
- Addressed the runtime limit of Google Colab which resulted in less epochs and hence less accuracy during training

Conclusion

- My project mostly turned out as expected. I was able to reach my design criteria and test the model with both same and different problem domains
- I realized ML models are not able to transfer knowledge from one domain to the other as expected
- Even the latest Deep Learning models are not able to combine their knowledge efficiently from different domains and apply them to solve a mixed problem - which an elementary student can do!
- In the future:
  - Understand the inner problem of mathematical knowledge transfer
  - Use that analysis to work with the model's inner structure to improve the accuracy of solving math problems
  - Would also like to test different types of input pre-processing to see which would produce the best results

References

1. Saxton, Grefenstette, Hill, Kohli, 2019, “Analysing Mathematical Reasoning Abilities of Neural Models.” ArXiv.org, 2 Apr. 2019, arxiv.org/abs/1904.01557.

2. Deepmind, 2018, “Analysing Mathematical Reasoning Abilities of Neural Models.” 27 Sept. 2018, https://www.deepmind.com/research/publications/analysing-mathematical-reasoning-abilities-neural-models.

3. Deepmind, 2019, “Mathematics Dataset” https://github.com/deepmind/mathematics_dataset.

. Deepmind/mathematics_dataset in GitHub.

4. “Transformer Model for Language Understanding: TensorFlow Core.” TensorFlow, https://www.tensorflow.org/tutorials/text/transformer, Date Read: July 2020.

5. Ray, Tiernan, 2019, “AI Ain't No A Student: DeepMind Nearly Flunks High School Math.” ZDNet, ZDNet, 4 Apr 2019, www.zdnet.com/article/ai-aint-no-a-student-deepmind-flunks-high-school-math/. Date Read: July 2020. https://towardsdatascience.com/the-limits-of-artificial-intelligence-fdcc78bf263b, "The limits of artificial intelligence.” Marc Botha. Date Read: August 2020.