Mathematical Problem Solver From Natural Language Using Deep Learning

A Thesis submitted to the
Department of Computer Science and Engineering, Jahangirnagar University
in partial fulfillment of the requirements for the degree of
M.Sc. in Computer Science and Engineering

By
Muhammad Rakibul Islam
ID: CSE202203062
Batch: 30

Supervised by
Dr. Morium Akter
Associate Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
JAHANGIRNAGAR UNIVERSITY

OCTOBER 2023

Abstract

We can solve natural language mathematical problems to make life simple. Without manual calculation, we can quickly and easily achieve the outcome of any complex mathematical problem. Every day, we have to measure our daily living costs with manual calculations which are so complicated and time-consuming and difficult for unlearned people that we build a system that can easily and quickly calculate our daily expenses in our natural language. This method uses the natural language equation to be derived.

Declaration

The research work entitled "**Mathematical Problem Solver From Natural Language Using Deep Learning**" has been carried out in the Department of Computer Science and Engineering, Jahangirnagar University is original and conforms the regulations of this University.

I understand the University's policy on plagiarism and declare that no part of this thesis has been copied from other sources or been previously submitted elsewhere for the award of any degree or diploma.

Muhammad Rakibul Islam

Counter Signed by

Dr. Morium Akter

Acknowledgment

I have meticulously completed my thesis, titled "Mathematical Problem Solver From Natural Language Using Deep Learning," and I am deeply grateful to those who made this achievement possible. My foremost thanks go to my parents and our Creator for their unwavering faith and support. I extend my heartfelt appreciation to my esteemed supervisor, Dr. Morium Akter, whose expert guidance and mentorship were indispensable to my success. I also acknowledge the valuable contributions of my friends who assisted me in overcoming challenges, and I express my gratitude to Jahangirnagar University for providing the opportunity and a conducive environment for my research endeavors. My collaborative efforts and the support of these individuals and institutions were integral to my academic journey, and for that, I am truly thankful.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

Symbol	*Description*
$\mathbf{H}$	Random MIMO Channel Matrix
$\boldsymbol{E}\{.\}$	Expectation
$(.)^{+}$	Hermitian Transposition
$\mathbf{E}[\ \mathbf{trace}(\mathbf{H}^{+}\mathbf{H})]$	Expectation of the trace of complex channel matrix ($H^{+}H$)
$\boldsymbol{E}[\boldsymbol{trace}\{(\mathbf{H}^{+}\mathbf{H})^{\mathbf{2}}\}]$	Expectation of the trace of complex channel matrix $(H^{+}H)^{2}$
$\boldsymbol{E}[\boldsymbol{x}^{\boldsymbol{k}}]$	k^{th} Moment of the MIMO Channel
$\dfrac{\boldsymbol{E_b}}{\boldsymbol{N_{0\,min}}}$	Minimum Normalized Transmit Energy per Information Bit
$\dfrac{\boldsymbol{E_b}}{\boldsymbol{N_0}}$	Normalized Transmit Energy per Information Bit
$\boldsymbol{S_0}$	Wideband Slope
$\boldsymbol{R_c}$	Code Rate in bits/s/Hz
$\boldsymbol{C}\left(\dfrac{\boldsymbol{E_b}}{\boldsymbol{N_o}}\right)$	Shannon's capacity function with respect to E_b/N_0
$\boldsymbol{\mu}$	Mean of Gaussian random variable
$\boldsymbol{\sigma}$	Standard deviation
$\boldsymbol{\lambda}$	Scale Parameter
$\mathbf{m}$	Shape Parameter
$\boldsymbol{\Gamma}(\boldsymbol{x})$	Gamma function
$\boldsymbol{\gamma_{av}}$	Ratio of shape and spread parameter
$\mathbf{k}$	Constant

LIST OF MODELS

Chapter 1

Introduction

1.1 Background and Motivation

A computer program can easily calculate equation solutions, but it does not grasp the same problem as the word problem in which the computer has to find out which equation is to be placed, as well as which information is contained in the issue text. In this topic, three models can be used. A baseline is a method used to produce predictions for a data set using heuristic, simple summary statistics, random or master learning. An algorithm to learn a function that models the relationship between the input data and the destination variable (or label). Seq2Seq models are perfect for translating words from one language into sequences of words in another. Long-Short-Term-Memory (LSTM) models are a popular choice for that type of model. The LSTM modules can make sense of the sequence using sequence-dependent data when they remember (or forget) the parts they find important (or unimportant). Transformers solved the problem of transduced sequences or neural machine translation. That means any task that transforms a sequence into a sequence of output. This includes speech recognition, a transformation from text to speech, and so on. LSTM transformers are an architecture for transforming sequences into sequences with two components (encoder and decoder), but they are different from those already described because they do not involve recurrent networks (GRU, LSTM, etc.).

We propose a Mathematical problem-solving approach using a different template. We translate math problem word problems directly into equation templates using the seq2seq, Baseline, and Transformer model instead of previous statistical training approaches.

1.2 Objective

In this paper, I propose an approach to solving math word problems using a different template.

My calculation develops frameworks of conditions while adjusting their factors and numbers to the issue content. MWP solving is believed to be challenging because of the semantic gap between the mathematical expressions and language logic [1]. The requesting within the numerical expression doesn't continuously matter as what is anticipated in straightforward content NLP issues because of commutative laws of expansion and duplication.

1.3 Research Problem

If we want to solve math problems manually then it must take much time. In the current world, People don’t have enough time to solve a mathematical problem on their own when a machine can give the result within a second using the algorithm. In this situation, an automatic mathematical problem solver can solve it easily and faster. On the other hand, we can get an accurate result if we use it.

1.4 Thesis Outline

The first chapter of this dissertation consisted of an explanation of the problem, the purpose, and the background study. It will proceed from here as follows: The proposed model is discussed in Chapter 2 and the key components are briefly discussed. We talked about our methodology and outlined the parts of the proposed model in chapter 3, which contains a detailed description of our work. The fourth chapter describes the whole system as one unit and includes the relevant figures. In the fifth chapter, we discussed, in addition to a conclusion and references, how we plan to extend our work in the future.

Chapter 2

Literature Review

2.1 Evolution of NLP Approaches for Math Word Problem Solving

The purpose of the task of Math Word Problem (MWP) is of course to answer a scientific survey on the content.

The creation of computer frameworks for programmed scientific words (MWP's) has fascinated NLP analysts since 1963. (Feigenbaum et al., 1963; Bobrow, 1964). Measured machine-based learning strategic systems, however, still have to deal with this matter, (Kushman et al., 2014; Amnueypornsakul and Bhat, 2014; Zhou et al., 2015; Mitra and Baral, 2016; Roy and Roth, 2018; Shi et al., 2015; KoncelKedziorski et coll., 2015; Roy and Roth, 2015; HUANG et al., 2017). Endeavors on including planning or formatting. For further literary works that consequently fathom mathematics issues, refer to Zhang et al (2018). The profound neural systems have opened up a contemporary course on computerized MWP understanding (DNNs). Ling et al. (2017) take into account numerous issues of choice input and thus create the contents envisaged and the final choice. At that time Wang et al. are the main attempts to use extensive fortification for the arrangement of mathematical issues (2018). The deep neural solver (DNS) from Wang et al. (2017) is not made of highlights for the natural use of the SEQ2SEQ show to memorize problem-to-equation mapping. Despite having been detailed in a promising way that the demonstration (Wang et al. 2017) still persists in a problem of conditional duplication. Given the implementation, once on November 15, 2018 (Vaswani et al., 2017) of the distinctive models SEQ2seQ for machinery interpretation as demonstrated by repetitive SEQ2SEQ and TransarXiv: 1811.05632v2 [cs.CL], it ensures that they will be adapted for the solution of MWP. In addition, their use is promising.

2.2 Application of Seq2Seq and Transformer Models

The project aims at using Deep Learning to convert mathematical problems automatically such as:

If Andy has 5 apples and Jerry eats 1 then how many are left?

Equation: X = P-Q

Solution: X = 4

In order to accomplish this job, I experimented with seq2seq models and Transformer models. The project is divided into two main components:

1. A baseline evaluation involving a straightforward list of generated problems, like "Two hundred fifty plus Three thousand nine hundred," designed to assess the Seq2Seq model's performance

2. A more extensive phase that involved working with a dataset comprising approximately 38,000 single-variable mathematical word problems, intending to train and compare the performance of both Seq2Seq and Transformer models..

2.3 Challenges and Limitations of Seq2Seq Models in MWP Solving

A standard MWP could be a short account, which would describe a fraction of the planet and a questionable amount. Analysts suggested a number of approaches to fathoming MWPs based on a Seq2Seq show [Sutskever et al., 2014] [Ling et al. 2017, Wang et al. 2017; Huang et al. 2018; Wang et al. 2018a]. These Seq2Seq solvers were found to be strong. Modern expressions are not created in the dataset preparation [Wang et al., 2017]. Seq2Seq models of determination of human discomfort do not fit the goal-driven approach component. When people looked at the content of the MWP problem, they first choose the target amount as the goal and then pay the relevant information to accomplish the goal If the data specifically included fulfilling this purpose, the problem has been resolved; otherwise, the goal should be divided into two sub-targets which are coordinates between an administrator and the important data. The strategy is subsequently revised for and sub-goal until all goals have been achieved. Another problem is that Seq2Seq models can provide invaluable expressions. In spite of the fact that Seq2Seq can understand this issue by using the expression tree post-order cross-section as a target arrangement within the Seq2 Seq, it is difficult for Seq2 Seq to demonstrate the tree-structured expression tree relationship through the intersectional grouping of the tree after order in the interpretation process. The show initializes the root objective vector, which talks about the final goal of the issue, and then the relevant data is approximately combined into the basic vector following the best downward degradation. The vector and its configuration vector are then expected to use a token that indicates if the target is to be helped. The chance that the token will be a numerical appreciation or the goal is to achieve something else straightforward (i.e. Two modern sub-goal vectors, the token anticipated is an administrator One will be created).

Chapter 3

System Model

3.1 Neural Network-Based Approaches for Automatic Problem Solving

3.1.1 Deep Learning

Deep learning is an AI function that I employ to replicate the intricate workings of the human mind in processing data for tasks like object detection, speech recognition, language translation, and decision-making. In my project, deep learning AI serves as a vital tool, capable of learning from both unstructured and unscheduled data without the need for human intervention.

As part of my work, I harness deep learning as a specific form of artificial intelligence, often referred to as deep neural networks or neural learning. The primary aim is to mimic the human brain's functioning in decision-making. It's essentially a facet of machine learning, with the remarkable ability to operate in a nonlinear decision-making process. My project delves into deep learning, where it thrives in making decisions autonomously, particularly when faced with unstructured data. Notably, object recognition, language understanding, and language translation are among the tasks where deep learning excels.

In my project's context, deep learning employs hierarchical neural networks to dissect and comprehend data. These hierarchical neural networks are designed to mirror the structure of the human brain, where neuron codes are interconnected. The multi-layered design of deep learning enables a non-linear approach to data processing across multiple layers, with each layer incorporating additional information, in stark contrast to traditional linear approaches in machine programs.

Furthermore, my project is situated in the digital era, which has evolved hand-in-hand with the explosion of data, often referred to as "big data." This vast pool of data is sourced from diverse origins such as social media, internet search engines, e-commerce platforms, and online streaming services. While this data is readily accessible, the challenge lies in its often unstructured nature, making it a daunting task for individuals to comprehend and extract relevant information. As I work on this project, I recognize the immense potential this wealth of data holds, and I'm keen on adapting automated AI systems to harness this data for various applications.

3.1.2 Natural Language Processing

Natural Language Processing (NLP) is a crucial element of my project, encompassed within the realm of artificial intelligence (AI). Its primary function is to facilitate machines in comprehending and generating human language seamlessly. Within the context of my project, Natural Language Processing (NLP) represents a potent fusion of linguistic principles and computer science, with the goal of deciphering language rules and structures. It empowers intelligent systems not only to comprehend but also to analyze and extract meaning from both text and speech through the utilization of machine learning and NLP algorithms.

NLP plays a central role in unraveling the intricate layers of language, covering aspects such as syntax, semantics, pragmatics, and morphology. In my work within computer science, I take linguistic information and convert it into rule-based, automated learning algorithms. These algorithms possess the remarkable capacity to address specific issues and perform desired tasks with precision and efficiency. For instance, a notable illustration is Gmail's use of NLP through keyword extraction to automatically categorize emails into sections like Promotions, Social, Primary, or Spam. By intelligently associating words in email subject lines with predefined tags, machines learn to categorize emails effectively.

The incorporation of NLP into my project offers numerous advantages, further emphasizing its significance in the field of AI. Here are some high-level benefits that enhance the competitiveness of my work:

- **Conducting Large-Scale Analyses:** NLP equips machines with the capability to automatically comprehend and analyze extensive volumes of unstructured textual data, including social media comments, customer service inquiries, online reviews, and news reports.
- **Real-Time Process Automation:** NLP tools can facilitate the rapid and precise classification and dissemination of data without significant human intervention, operating efficiently around the clock.
- **Customization for Industry-Specific Needs:** NLP algorithms can be tailored to accommodate complex, industry-specific languages, even accounting for nuances such as sarcasm and specialized terminology, aligning with the unique requirements and criteria of my project.

3.1.3 Automatic Problem Solver Baseline Model

A foundational element of my project involves the application of pattern models, which serve as fundamental frameworks capable of delivering practical results for specific tasks with relatively minimal development effort. Pattern models play a pivotal role across various domains and have been extensively employed in my research. These models encompass a variety of methodologies, including linear regression for predicting continuous values, decision trees for classifying structured data, pre-trained neural networks for vision-related tasks, and recurrent neural networks, among others.

While pattern models offer a robust starting point, it's important to recognize that they come with inherent limitations. The simplicity inherent in these models can be viewed as both an advantage and a drawback. For example, in the field of language modeling, basic pattern models often overlook word order within phrases, limiting access to structural information. Additionally, many foundational principles rely heavily on heuristics embedded within the model or make significant assumptions about the underlying data, such as linear modeling.

Moreover, a basic approach may not always yield groundbreaking research insights, which can pose challenges for researchers aiming to publish innovative findings. Nonetheless, these fundamental models play a critical role in establishing the foundation for more complex and sophisticated research within the project.

3.1.4 Seq2Seq

In the realm of machine translation, Google made a significant breakthrough with the introduction of Seq2seq. Prior to this innovation, translation methods were rather rudimentary, involving the direct translation of each word into the target language without considering language structures and sentence patterns. Seq2seq brought about a revolution in translation training by incorporating contextual information. It goes beyond the mere translation of the current word/input and takes into account the context within which it appears. This capability has expanded its application into various domains, including image captions, conversational patterns, and text summarization. As the name suggests, Seq2seq takes a sequence of words as input and generates another sequence of words as output. It employs a recurrent neural network (RNN) to achieve this, although the vanilla RNN is less commonly used due to the vanishing gradient problem. In the Google version, the more robust LSTM (Long Short-Term Memory) architecture is adopted. The context for each word is built by taking two inputs at each time step: the current input word and the previous output word (where the output from the model becomes the input). This Seq2seq model comprises two key components, namely the encoder and decoder, often referred to as the encoder-decoder architecture.

Encoder: The encoder leverages deep neural network layers to transform input words into corresponding hidden vectors. These vectors represent both the current word and its contextual information.

Decoder: The decoder operates in a similar manner to the encoder but in reverse. It utilizes the hidden vector generated by the encoder to produce the subsequent hidden vector, its hidden states, the current word, and predict the next word in the sequence. This sophisticated encoder-decoder framework has significantly advanced the field of machine translation.

3.1.5 Transformer Model

A transformer is a profound learning model, which uses an attention mechanism to assess the impact of various parts of input data. It is mainly used for the processing of natural languages (NLP). It also has applications for video comprehension tasks.

A transformer model handles variable inputs via self-attention layer stacks rather than RNNs or CNNs. There are several advantages to this general architecture:

- The time/space relationships on all the data are not assumed. Ideal for handling a number of objects (for example, StarCraft units).
- Instead of a series like an RNN it is possible to calculate layer outputs parallel.
- Without going through many RNN steps or convolution layers distant items may influence each other's output (see Scene Memory Transformer for example).
- It can learn dependencies for a long time. In many sequence tasks, this is a challenge.

3.1.6 Decoder and Encoder Model

Encoder-decoder models serve as a pivotal component in generating textual descriptions for images or videos through the lens of a learning model. In this process, the model receives an image as input and crafts a sequence of words as output, a functionality that extends to video analysis as well. These models exhibit a remarkable ability to comprehend the input phrases and discern the emotional nuances within them, often rated on a scale ranging from 0 (negative) to 1 (positive). In scenarios like call centers, these models are deployed to analyze customer emotions and reactions concerning specific keywords or company promotions.These models understand the input phrase and the feelings of the input. It is usually rated from 0 to 0 (negative) between 0 and 1 (positive). In call centers, the development of client emotions and reactions to certain keywords or company discounts is analyzed.

The functionality of this model encompasses reading and interpreting an input phrase, translating its content and concepts into a second language. This encoder-decoder architecture is notably integrated into applications like Google Translate. The term "encoding" denotes the process of converting data into the required format, a transformation akin to rendering a word (text) into an image, as illustrated by the game Pictionary. In the domain of machine learning, this process involves converting a sequence of words into a two-dimensional vector, often referred to as a hidden state or condition in Spanish. The encoder is constructed using stacked recurrent neural networks (RNNs), a structure that enables the model to grasp the sequential context and temporal dependencies within the data. The encoder's output, the hidden condition, encapsulates the comprehensive meaning of the entire input sequence, and the length of this vector varies depending on the number of RNN cells used.

Conversely, "decoding" is the process of translating a coded message into a coherent and understandable language. This function is akin to the second person in a Pictionary team translating an image into a word. In the context of machine learning, the decoder's role is to transform the two-dimensional vector into an output sequence, typically an English sentence. This transformation is facilitated through the utilization of RNN layers and a dense layer, which work in concert to predict the English words and assemble the final output sequence. The intricate interplay between the encoder and decoder within these models allows for a rich and nuanced understanding of both textual and visual content.

3.2 Related Work

Past work on programmed math word problem-solving falls into two categories: typical approaches and factual learning approaches. In 1964, Understudy Bobrow (1964) handles logarithmic issues in two steps: to begin with, they change normal dialect sentences into part sentences employing a little set of change designs. At that point, the bit sentences are changed into numerical expressions by design matching. A comparable approach is additionally utilized to unravel English rate issues Charniak (1968, 1969). Liguda and Pfeiffer Liguda and Pfeiffer (2012) propose modeling math word issues with increased semantic systems. In expansion, Addition/subtraction issues are examined most Briars and Larkin (1984); Dellarosa (1986); Bakman (2007); Yuhui et al. (2010); Roy et al. (2015). In 2015, Shi et.al Shi et al. (2015) propose a framework SigmaDolphin which consequently understands math word issues by semantic parsing and thinking. Within the same year, Koncel et.al KoncelKedziorski et al. (2015) moreover formalizes the issue of fathoming multi-sentence arithmetical word issues like that of producing and scoring equation trees. 1We arranged to create a dataset freely accessible when the paper is distributed in 2014, factual learning-based approaches are proposed to fathom the math word issues. Hosseini et al. Hosseini et al. (2014) bargain with the open-domain viewpoint of logarithmic word issues by learning verb categorization from preparing information. Kushman et al. Kushman et al. (2014) proposed a condition format framework to unravel a wide run of variable-based math word issues. Zhou et al. Zhou et al. (2015) encourage expanding this strategy by embracing the max-margin objective, which comes about in higher exactness and lower time fetched. In expansion, Roy and Roth Roy et al. (2015); and Roy and Roth (2016) try to handle math issues with different steps and operations without depending on extra comments or predefined formats. Mitra et al. Mitra and Baral (2016) present a novel strategy to memorize to utilize equations to illuminate straightforward expansion subtraction number-crunching issues. As detailed in 2016 Huang et al. (2016), state-of-the-art approaches have greatly moo execution on an enormous and exceedingly different information set (18,000+ issues). In differentiating between these approaches, we think about the achievability of applying profound learning to the assignment of math word issue tackling.

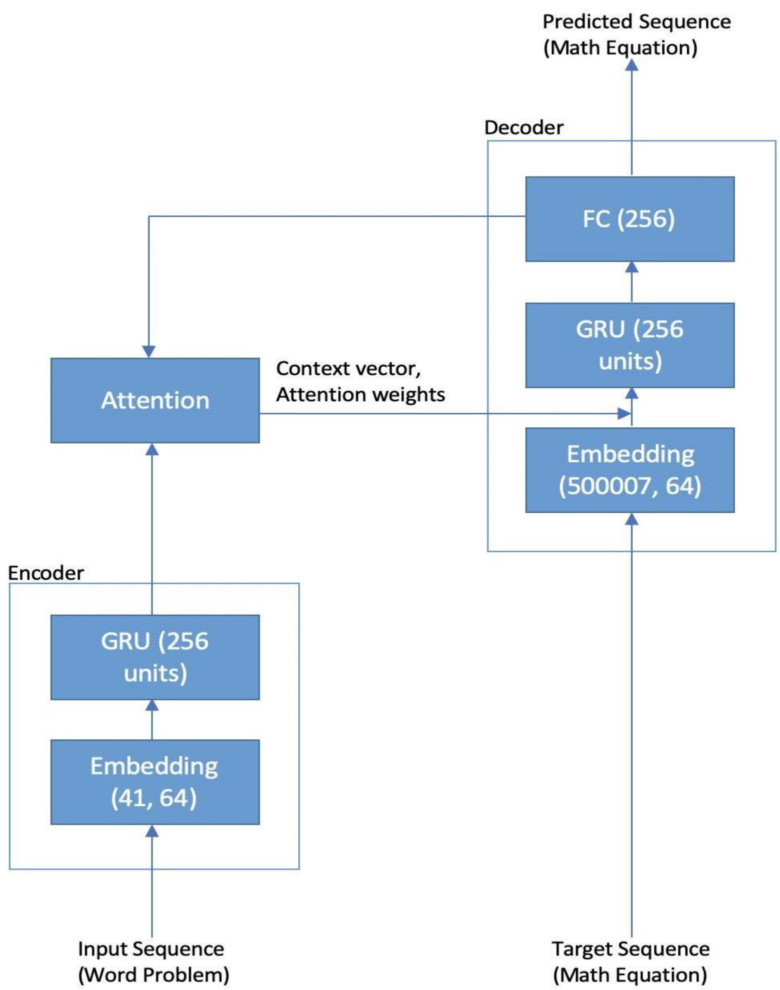
3.2.1 BaseLine Model

Deep learning is an AI function that imitates the functioning of the human mind in processing data for object detection, speech recognition, language translation, and decision-making. Deep learning AI can learn from both unstructured and unscheduled data without human supervision.

The Seq2Seq model used has an encoder, decoder and an attention model.

- The encoder model takes in the word problem as a sequence.
- The Bahdanau Attention computes attention weights for the input sequence along with the most recent prediction.
- The decoder uses the attention weights to predict the next output in the sequence. Teacher forcing is used while training the decoder.

Baseline Model Architecture:



.

Fig 1.1: Baseline Model Architecture

3.2.2 Seq2Seq(Bidir GL) Model:

This model is very similar to the baseline model except of the following:

- The encoder has a bidirectional GRU layer
- A dropout layer is added to reduce overfitting

● LSTM layer is used in the decoder The model does not memorize the dataset, giving us a reasonable BLEU score. However, the attention plots are not as expected- not giving attention to correct parts of the sentence while translating

Seq2Seq(Bidir GL) Model Architecture:

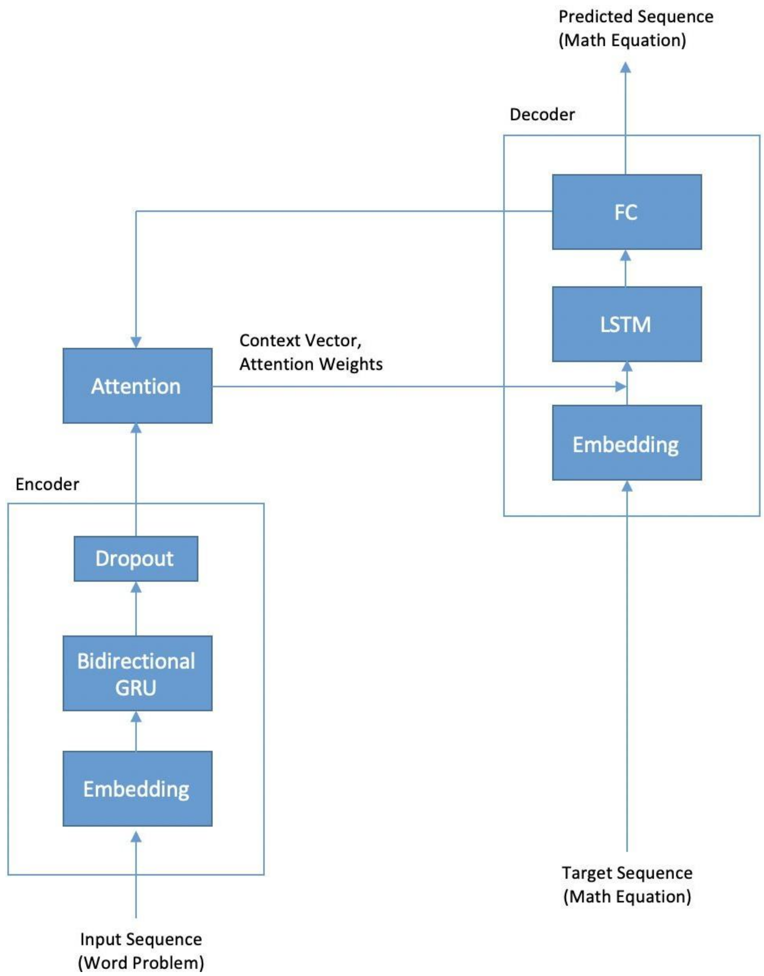


Fig 1.2: Seq2Seq Model Architecture

Chapter 4

Implementation

4.1 Dataset

I take a dataset of approximately 38,000 single variable math word problems and train a seq2seq and Transformer model to compare their performance. I created the dataset using a question generator for math word problems, along with a list of around 2,000 single variable equation questions from the Math Word Problem Repository (MAWPS). In total, I have approximately 38,000 question-equation pairs.

4.2 Proposed Methodology

The transformer model consists of the following blocks:

- An encoder, consisting of N encoder layers.
- Each encoder layer has a multi-head attention block and a feed-forward block
- A decoder, consisting of N decoder layers
- Each decoder layer has a masked multi-head attention block, a multi-head attention block and a feed-forward block
- Positional encoding is added to the input and target sequences since transformers have no recurrent units
- The output of the decoder then goes into a fully-connected layer which gives us our final prediction

Architecture

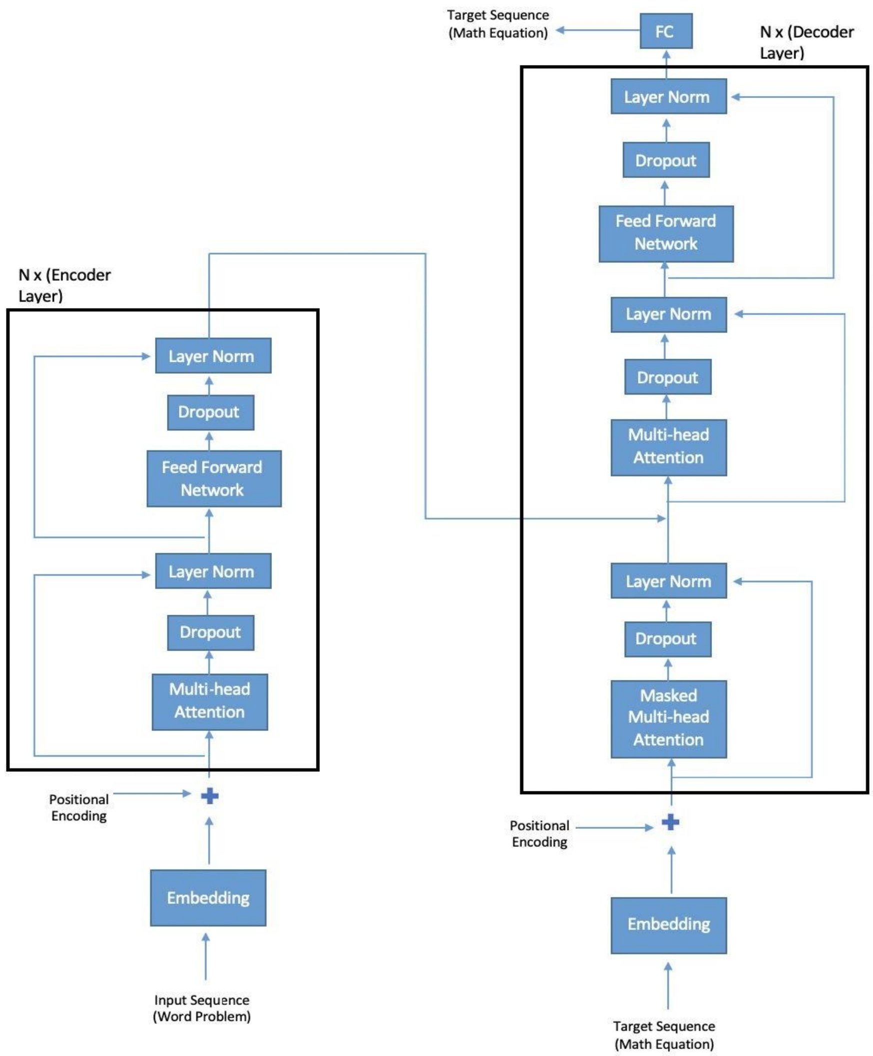


Fig 2.1: Proposed Model Architecture

4.3 Matrices Used

We used two metrics to evaluate the performance of all our models:

1. **Accuracy**: In order to see how exact the predictions of the model are

2. **Corpus BLEU (Bilingual Evaluation Understudy) score**: A metric developed specifically to for auto-translation systems, BLEU score compares n-grams of the candidate and reference translation, so even if the translation is not an exact match, the score is not zero.

4.4 Results

We take a dataset of ~38k single variable math word problems and train a seq2seq and Transformer model and compare their performance.

Comparison of the 3 models discussed above:

Model	**Seq2seq Baseline**	**Seq2seq Bidirectional**	**Transformer**
Corpus BLEU Score	0.9952	0.4478	0.7059
Accuracy	0.9656	0.2568	0.6860

Table 1.1: Comparison of three Models

While the seq2seq model gives a reasonable score on the validation set, when we look at the attention plots that are generated while translating a sentence, we see that attention is not given to the correct tokens when translating. This may indicate that this model is again starting to memorize the dataset rather than learning to actually convert a word problem to an equation. On the other hand, the Transformer model gives a higher score, and the attention plots also indicate that it is performing better than the seq2seq model

4.5 Accuracy

I got Corpus BLEU score of the model: 0.75 and Accuracy of the model: 0.69.

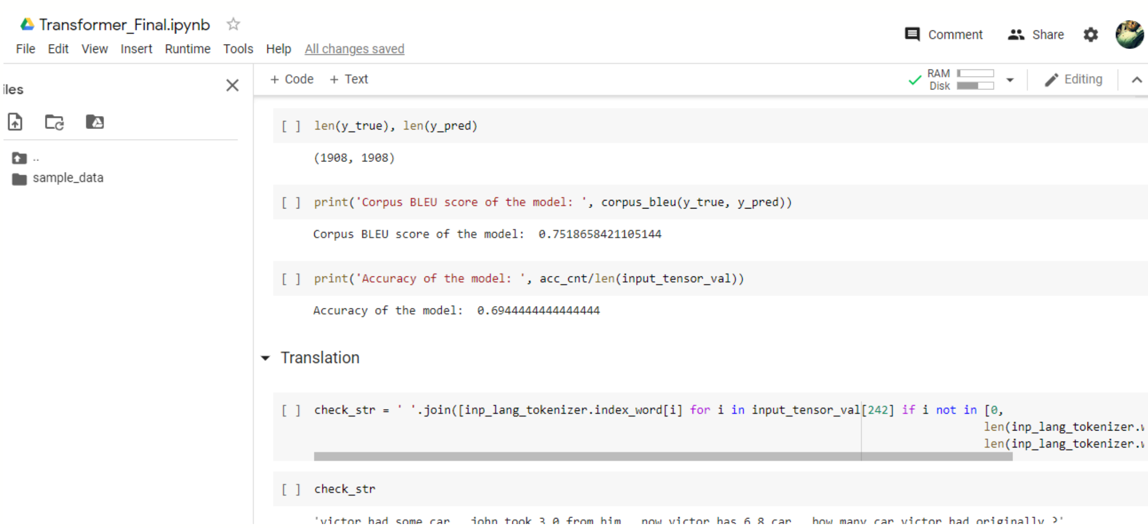


Fig 3.1: Corpus BLEU score

4.6 Translation

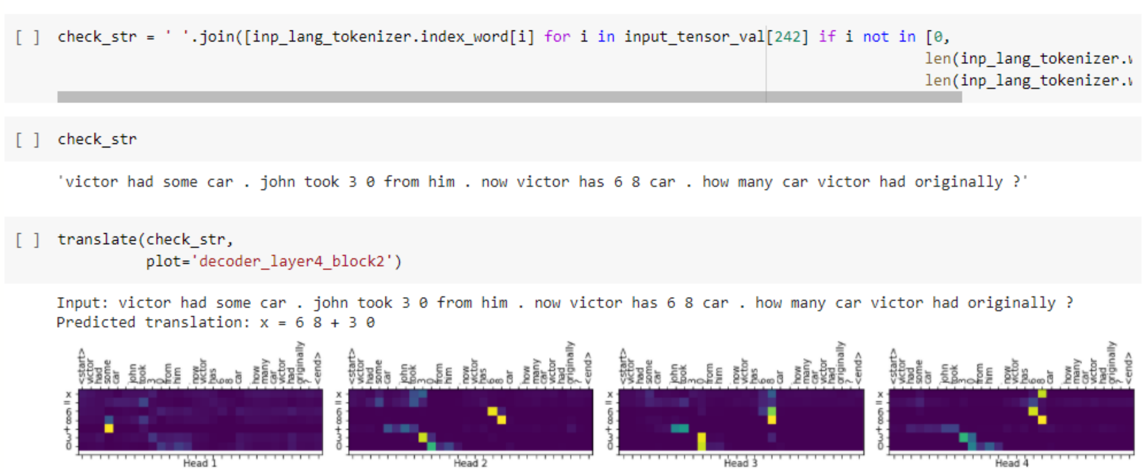


Fig 3.2: Translation

```
[ ] translate("Jerry had 135 pens. John took 19 from him. How many pens Jerry have left?",
              plot='decoder_layer4_block2')
```

```
Input: Jerry had 135 pens. John took 19 from him. How many pens Jerry have left?
Predicted translation: x = 1 3 5 - 1 9
```

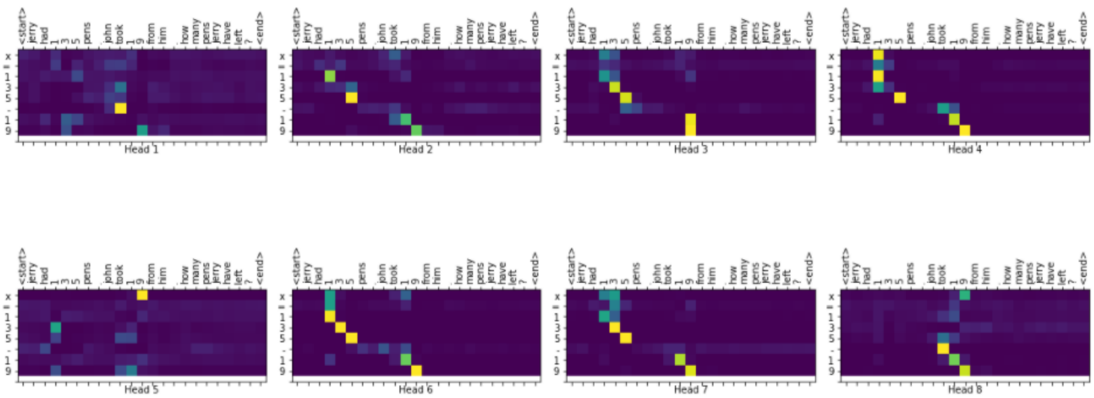


Fig 3.3: Translation

Chapter 5

Conclusion

5.1 Summary

I have proposed a transformer demonstrate to naturally unravel math word issues. This demonstrates specifically changes issue content into a math condition format. This is often the primary work of applying profound learning advances to math word problem-solving. A huge dataset has been developed to demonstrate preparing and experimental assessment. Exploratory comes about to appear that this demonstrates altogether outflanks state-of-the-art factual learning methods in math word problem-solving. The yield of our demonstration may be a single condition containing one obscure variable. In this manner, our approach is as it were appropriate to the issues whose arrangement includes one straight condition of one obscure variable. As future work, we arrange to amplify our demonstration to be able to test with multi-variable conditions and Handle the commutative and affiliated properties of math conditions amid forecast.

5.2 Future Work

Building upon the proposed methodology and the insights gained from its potential advantages and limitations, I see several promising directions for future work in the field of mathematical word problem-solving. These avenues can help enhance and expand the approach, making it more versatile and effective.

- **Larger and Diverse Datasets**: One crucial aspect is the enlargement of the dataset. Collecting a more extensive and diverse set of mathematical word problems can expose the model to a wider range of problem types, complexities, and linguistic variations, contributing to better generalization and adaptability.

- Data Augmentation: The use of data augmentation techniques, such as paraphrasing and variations in problem formulation, can further enrich the dataset. Augmented data can improve the model's ability to handle variations in language and problem structure.
- **Fine-tuning**: Investigating fine-tuning strategies can help customize the model for specific domains or problem types. Fine-tuning the pre-trained model with domain-specific data can enhance its performance in specialized applications.
- **Interpretability and Explainability**: Enhancing the interpretability of the model is a critical area for future work. Developing methods to visualize and explain the model's decision-making process, particularly in complex mathematical word problems, can make it more transparent and trustworthy for users.
- **Human-in-the-loop Systems**: The integration of human-in-the-loop systems can assist in addressing challenging mathematical word problems. Combining automated solutions with human expertise can be a valuable approach, particularly in educational contexts.
- **Multi-Modal Inputs**: Exploring multi-modal inputs, including text, visual, and symbolic representations of problems, can open up new avenues for understanding and solving mathematical word problems. Models that effectively handle such multi-modal inputs may offer superior performance.
- **Real-World Testing**: Transitioning from research settings to real-world applications is a significant step. Testing the methodology in educational platforms, online problem-solving tools, or practical problem-solving applications can provide valuable insights into its real-world performance.
- **Multilingual Support**: Extending the methodology to support multiple languages can make it accessible to a global audience. Adapting the model to different languages and linguistic variations is an essential aspect of future work.
- **Hybrid Models**: Combining the strengths of Transformer models with other architectures, such as graph neural networks or reinforcement learning, may lead to more robust and versatile problem-solving systems.
- **Benchmarking and Evaluation Frameworks**: The development of standardized benchmarks and evaluation frameworks specific to mathematical word problem solving can facilitate fair comparisons between different approaches and foster innovation in the field.

In summary, future work in the domain of mathematical word problem solving should focus on expanding and refining the methodology, improving its interpretability, and ensuring its applicability in practical, real-world scenarios. By addressing these challenges and opportunities, the field can continue to advance and provide valuable solutions for a wide range of users, from students to professionals.

References

[1] W. K. Chen. *Linear Networks and Systems*. Belmont, CA: Wadsworth, 1993, pp. 123-35.

[2] J. E. Bourne. “Synthetic structure of industrial plastics,” in *Plastics*, 2nd ed., vol. 3. J.Peters, Ed. New York: McGraw-Hill, 1964, pp.15-67.

[3] G. Pevere. “Infrared Nation.” *The International Journal of Infrared Design*, vol. 33, pp. 56 - 99, Jan. 1979.

[4] D. B. Payne and H. G. Gunhold. “Digital sundials and broadband technology,” in *Proc. IOOC-ECOC*, 1986, pp. 557-998.

[5] B. Brandli and M. Dick. “Engineering names and concepts,” presented at the 2nd Int.Conf. Engineering Education, Frankfurt, Germany, 1999.

[6] M. Duncan. “Engineering Concepts on Ice. Available online: www.iceengg.edu/staff.html, Accessed on Oct. 25, 2000.

[7] S. Mack. “Desperate Optimism.” M.A. thesis, University of Calgary, Canada, 2000.