Analogical Math Word Problems Solving with Enhanced Problem-Solution Association

Zhenwen Liang[1], **Jipeng Zhang**[2], and **Xiangliang Zhang**[✉1]

[1]University of Notre Dame, {zliang6, xzhang33}@nd.edu
[2]Hong Kong University of Science and Technology, jzhanggr@conect.ust.hk

Abstract

Math word problem (MWP) solving is an important task in question answering which requires human-like reasoning ability. Analogical reasoning has long been used in mathematical education, as it enables students to apply common relational structures of mathematical situations to solve new problems. In this paper, we propose to build a novel MWP solver by leveraging analogical MWPs, which advance the solver's generalization ability across different kinds of MWPs. The key idea, named *analogy identification*, is to associate the analogical MWP pairs in a latent space, i.e., encoding an MWP close to another analogical MWP, while moving away from the non-analogical ones. Moreover, a *solution discriminator* is integrated into the MWP solver to enhance the association between the representations of MWPs and their true solutions. The evaluation results verify that our proposed analogical learning strategy promotes the performance of *MWP-BERT* on Math23k over the state-of-the-art model *Generate2Rank*, with 5 times fewer parameters in the encoder. We also find that our model has a stronger generalization ability in solving difficult MWPs due to the analogical learning from easy MWPs.

1 Introduction

Math word problem (MWP) solving has attracted considerable attention in recent years. Currently, MWP solver design focuses on generating an equation towards an unknown quantity, with an input problem description, as shown in Figure 1. Building such a successful solver is quite challenging, as it requires mathematical understanding and multi-step reasoning abilities to transform the implied logic behind the problem into a mathematical equation composed of operators and quantities. With the emergence of deep learning, MWP has been effectively studied by Seq2Seq models (Xie and Sun, 2019; Zhang et al., 2020b) and pre-trained language models (PLMs) (Tan et al., 2021; Li et al., 2021; Huang et al., 2021; Liang et al., 2022), which treat it as a translation task from the natural language description of a problem to its solution in a mathematical equation.

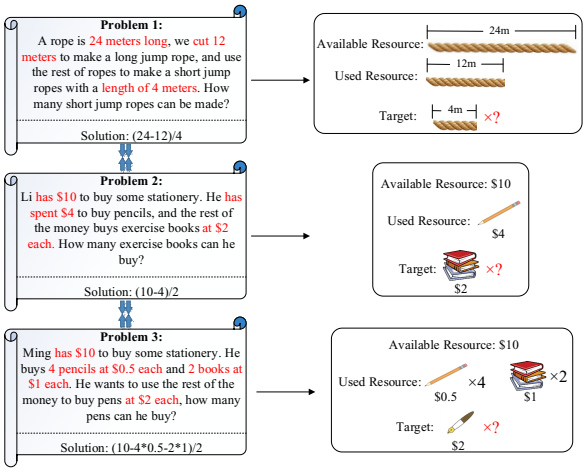

Figure 1: This figure shows three examples of math word problems. We can find mathematical analogy across different topics (P1 and P2) and different difficulty levels (P2 and P3).

In educational science, analogical reasoning (Novick and Holyoak, 1991; Bernardo, 2001) has long been considered as a crucial skill in human problem-solving. In education and training of mathematical reasoning, researchers (Stacey et al., 1982) interpret it as "a dynamics process which, by enabling us to increase the complexity of ideas we can handle, expands our understanding." For example, the MWP P1 and P2 in Figure 1 are in different situations, but both aim to calculate the available quantity of a target object based on known resources, leading to the same solution structure. The MWP P2 and P3 both focus on the knowledge of 'number of units = money / unit price', while the MWP P3 has more reasoning steps. Gaining the skill of generating the solution of P2 to address P1 and P3 is important for students, as well as for automated MWP solvers. A widely accepted defi-

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9454 - 9464
December 7-11, 2022

nition of analogical reasoning is mapping the process of mathematization between different domains (Vosniadou, 1988). We thus propose a strategy of associating the mathematization process of analogical MWP pairs, e.g., P1 and P2, P2 and P3. The key idea is to encode the analogical problems close in the latent representation space, while encoding the non-analogical problems distantly. The introduction of analogical learning in MWP solving can then alleviate the limitations of existing methods such as simple-to-difficult generalization (Zhang et al., 2020b; Jie et al., 2022) and topic-to-topic generalization (Hong et al., 2021).

While the analogical MWPs have close representations, they should be also more associated with their corresponding true solutions than wrong solutions. We thus encode the solution as well, and design a module named solution discrimination to enhance the association of an MWP and its correct solution. The enhancement is powered by generating hard negative samples that are wrong solutions but similar to the correct solution. Negative sampling has been proved effective in training strong MWP solvers (Li et al., 2021; Liang and Zhang, 2021). For example, Li et al. (2021) used negative examples to calculate a contrastive loss and find patterns across MWPs, but the negative examples are retrieved from the given training set, which could be bounded by the dataset size and quality. Liang and Zhang (2021) bridged the problem space and solution space, by building a teacher module to help the solver match problems with their ground truth solutions and distinguish them from randomly generated negative solutions. However, the negative samples in their approach are randomly manipulated variants of the ground truth equation, and are thus limited in their shapes and fail to consider the hardness of negative samples. Therefore, we design a gradient-based manipulation method to find hard negative samples (i.e., wrong solutions that are difficult to be distinguished from a correct solution). In the meantime, to make the form of negative samples less bounded by the ground truth, we introduce more randomness during the generation of negative samples.

Overall, our proposed MWP solver is novel on considering the analogy among MWPs. It contains an *analogy identification* module which aims to transfer mathematization knowledge and skills to generalize the solution to analogical MWPs. Moreover, a *solution discrimination* module is incorporated to enable the solver to bridge problems and their ground truth solutions. Our proposed solver is a flexible framework and can have any MWP encoder-decoder plugged in. The experimental results show that training representative baselines (e.g., GTS (Xie and Sun, 2019) and MWP-BERT (Liang et al., 2022)) by our analogical strategy can further improve their accuracy. Comparing to the current best solver on Math23k, Generate&Rank (Shen et al., 2021), we make MWP-BERT achieve higher accuracy but with 5 times fewer encoder parameters. Ablation studies, qualitative analyses and case studies also demonstrate the effectiveness of our designed solver framework, and verify its generalization ability in solving difficult MWPs due to the analogical learning from easy MWPs.

2 Related Works

2.1 Math Word Problem Solving

The topic of automated math word problem solving was raised in 1960s (Bobrow, 1964). At the beginning, researchers established rules and tried to match every problem with a certain solving rule (Bakman, 2007). Then, statistical and machine learning methods (Shi et al., 2015) were widely applied and sometimes integrated with semantic parsing approaches (Koncel-Kedziorski et al., 2015). Inspired by the great success of deep learning and Seq2Seq models, a Seq2Seq solver with an encoder-decoder structure was proposed and outperformed transitional methods, after which many other Seq2Seq solvers (Wang et al., 2018, 2019; Liu et al., 2019) were proposed. Later, GTS (Xie and Sun, 2019) replaced the sequential decoder with a novel tree-based decoder and achieved great performance. Most following papers (Zhang et al., 2020b; Wu et al., 2020; Lin et al., 2021; Wu et al., 2021; Liang and Zhang, 2021) after GTS focused on the improving of encoder part without touching the tree-based decoder. Recently, due to the development of pre-trained language models (PLMs) (Devlin et al., 2019; Cui et al., 2020), the accuracy of MWP-BERT (Liang et al., 2022) firstly surpasses human performance (Wang et al., 2019). Besides accuracy improvement, there are many works exploring some other aspects of MWP solvers, e.g., teacher-student distillation (Zhang et al., 2020a), auxiliary training tasks (Qin et al., 2021), situation model (Hong et al., 2021). For more comprehensive reviews of MWP solver, please refer to these surveys (Zhang et al., 2019; Faldu et al., 2021).

2.2 Analogical Learning in NLP

The k-nearest neighbor retrieval has been a popular approach deployed in machine translation (He et al., 2021; Jiang et al., 2021) since the born of kNN-MT (Khandelwal et al., 2021). It is simple yet effective in terms of improving the translation performance and explainability. This idea was quickly adopted by other NLP research such as building language models (Liu et al., 2022) and text-generation (Li et al., 2022). There is one recent study in MWP solving that shows the benefits of retrieval-based analogical learning (Huang et al., 2021). The basic idea is to set up a memory module for storing the problems that have been learned. For a new given problem, the solver retrieves similar problems from the memory based on measuring the cosine similarities of the encoded problems and lets the solution of the retrieved problem participate decoding. However, this approach has several limitations. Firstly, the retrieved problems with high cosine similarity may not be analogical to the given problem. They have similar representation vectors perhaps because they have common topic words in the problem description, but their reasoning steps are completely different. These literally similar but non-analogical problems can mislead the decoder give a wrong answer. Secondly, with the increasing size of the training dataset/memory module, the solver will suffer from long search time.

We design our analogical reasoning solver with an *analogy identification* module, which takes advantage of only the problems whose mathematical analogy is confirmed by their solution. Therefore, we have no top-k selection issues. Moreover, unlike (Huang et al., 2021) that introduces the analogical problems in decoding, we use them for improving problem understanding (encoding). Analogical problem pairs are mapped close in a dense region, and non-analogical pairs are separated in different regions (as demonstrated later in Figure 3). This adjusted representation space facilitates the decoding process and leads to more correct results.

3 Approach

3.1 Problem Definition

We aims to train a math word problem solver that receives the problem description X as input and then generates an equation-shaped solution $Y = \{y_1, y_2, ..., y_n\}$ with length n. All equation solutions are transformed into a binary tree representation as proposed in (Xie and Sun, 2019) and sequantialized as their pre-order traversal, thus there exists no parenthesis in Y. The vocabulary of Y contains two parts, operators and numbers. In the solution tree, operator nodes are always parents of numbers, and the number nodes have to be leaf nodes, as shown in Figure 2. Specifically, the vocabulary of operators V_{op} contains 5 operators $\{+, -, *, /, \hat{}\}$ and the vocabulary of numbers V_{num} may vary across different problems. We apply number mapping (Wang et al., 2017) to replace the numbers in Y which can be found in X with placeholders, thus the length of V_{num} depends on how many numbers appear in X. Besides, there are some useful constants like π which are not shown in X but used in Y to get a solution, we also include them in V_{num}.

3.2 Backbone Solvers

Our target is to design a model that can solve MWP by analogical reasoning. Therefore, the model is in fact a flexible framework that can be implemented with any existing MWP solvers. We select GTS (Xie and Sun, 2019) and MWP-BERT (Liang et al., 2022) as backbone solvers and plug them into our novel framework. GTS uses GRU (Cho et al., 2014) as the encoder and MWP-BERT has a BERT encoder that is continually pre-trained on MWP corpus. Both of them apply tree-based decoder and two models can be found at open-sourced[1].

3.3 Analogy Identification

School students are often taught with example problems and tested by analogical problems in exams for evaluating their problem-solving skills. As stated in (Gentner and Holyoak, 1997), how to bridge analogy among problems is one of the key abilities in mathematical problem-solving. To build an automated MWP solver, we first target to also bridge analogy among problems. Thus we design a module named *analogy identification* for addressing two issues, 1) how to find analogical problems; and 2) how to bridge the analogy among problems.

Finding analogical problems Analogical problems usually have associated mathematization processes. In other words, they can be solved by running a similar series of mathematical operators on their own numbers, e.g., the problem 1 and 2 in

[1] https://github.com/ShichaoSun/math_seq2tree and https://github.com/LZhenwen/MWP-BERT

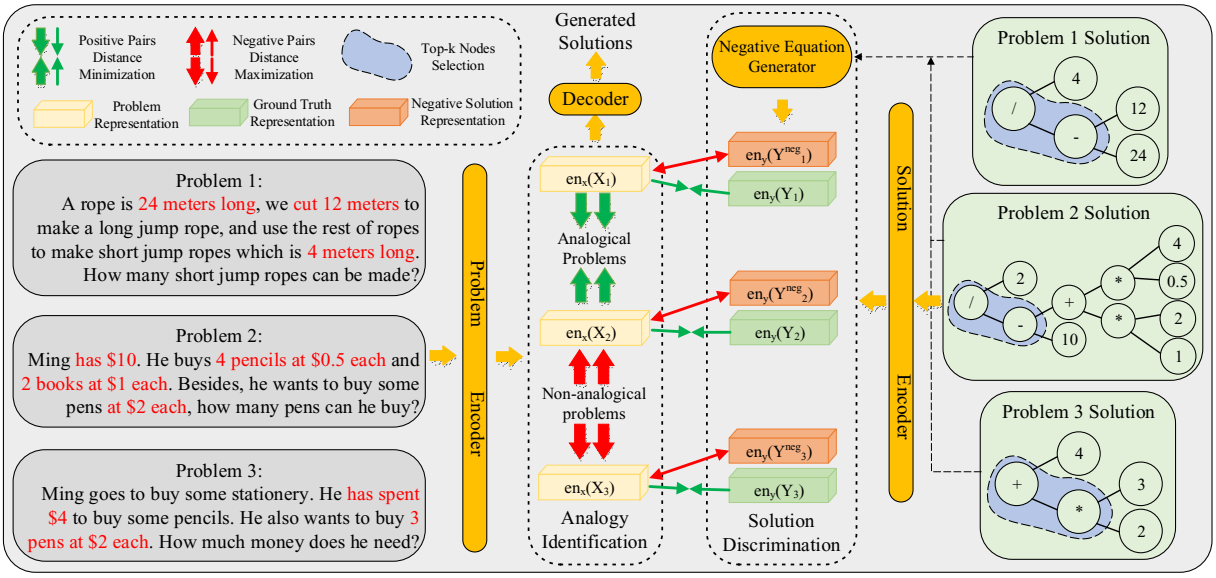

Figure 2: The figure shows the workflow of our proposed framework with 3 example problems and solutions. P1 and P2 have the same top-1 and top-2 nodes, thus they are positive pairs in the analogy identification module, while P2 and P3 is a negative pair. In the solution discrimination part, each problem representation is paired with their ground truth solution to minimize their distance. Then, negative solutions are generated by the generator and the distance between them and the problem representation will be maximized.

Figure 1. For analogical problems with different difficulty levels, e.g., the problem 2 and 3 in Figure 1, the more difficult one can often be simplified into an easier one. For example, '*buys 4 pencils at* $\$0.5$ *each*' in problem 3 can be simplified as '*spends* $4 * 0.5 = \$2$ *on pencils*', which matches well with '*spent* $\$4$ *to buy pencils*' in problem 2. By this simplification, we can see that the problem 2 and 3 both focus on addressing 'number of units = money / unit price'. To find these analogical problems, we first try to understand how a solution tree is generated by a decoder.

Generally, an MWP is solved by a top-down solution tree, where the nodes closer to the top reflect more about the ultimate goal, i.e., the root of the tree. And lower nodes achieve sub-goals by local calculation. Therefore, analogical problems tend to have the same top nodes, which typically manifest similar required knowledge or solving skills. For example, the solution tree of problem 1 and problem 2 in Figure 2 have the same top nodes of running "/" and "-". To find analogical problems, we prune a solution tree in a bottom-up manner, removing the bottom nodes and only keeping the top nodes. Note that we only consider operator nodes because they imply hidden logic (Yang et al., 2022) that different MWPs may share in common, while number nodes represent quantities varying across different MWPs. In this way, all solution trees are simplified to have fewer nodes. In our experimental implementation, we collect analogical problem pairs if they have the same top-1 operator (the same root node), or the same top-1 and top-2 operators (the same root node and its left children following the pre-order transversal). For an instance, problem 2 and problem 3 in Figure 2 are an analogical pair (a positive pair) since they have common root node "/" and its left child "-". We consider our operator matching method as a specific kind of measurement of similarity/analogy, incorporating analogies in (a) different topics and (b) different difficulty levels by matching outermost operators in the solution tree. There are two kinds of existing analogy identification methods - 1. Finding problems with similar semantics. 2. Finding problems with similar solutions (trees). However, the former method neglects the topic-topic analogy, and the latter method neglects the difficulty-level analogy.

Bridging the analogy For one analogical pair of problems (X_1 and X_2), they should be mapped close in a representation space, because they share similar knowledge and skills for being solved. To bridge their analogy, we maximize their analogy score which is defined by:

$$s_a = \sigma(MLP([en_x(X_1) : en_x(X_2)])), \quad (1)$$

where $en_x(X_1)$ yields the problem representation for X_1 by the problem encoder, [:] denotes vector

concatenation, and σ is a sigmod function. The multi-layer perceptrons (MLP) and the encoder $en_x()$ are trained jointly to associate the analogical pair. For two non-analogical problems, their analogy score can be calculated in the same way, and should be minimized.

Then the loss function for bridging the analogy can be defined by cross-entropy:

$$L_a = -(log(s_a) + log(1 - \bar{s}_a)) \quad (2)$$

where s_a is the score for positive pairs and $\bar{s}_a$ is for negative. In each iteration, we randomly sample our positive and negative pairs that are used to update the MLP and encoder.

3.4 Solution Discrimination

Besides the analogical learning part that groups analogical problems together, we also design a solution discrimination module to make the solver associate MWPs stronger to their ground truth solutions. We define a discrimination score that measures the association of a problem X with its ground truth solution Y:

$$s_d = Dis([en_x(X), en_y(Y)]) \quad (3)$$

where $en_y(Y)$ is the representation of equation Y encoded by a gate-recurrent-unit (GRU) (Chung et al., 2014). The discriminator can be empirically performed by a bi-linear similarity function without bias. In other words, the output of $Dis(A, B)$ is $A^T W B$ where W is a learnable matrix. For the problem X, we can also have a wrong solution (negative sample) Y^{neg}. A discrimination score $\bar{s}_d$ can be calculated by Eq.(3) as well. Then, this solution discriminator can be trained to minimize s_d while maximizing $\bar{s}_d$. In other words, the true solution is pulled close to the problem, while the wrong solution is pushed away from the problem. The discriminator is updated once in each mini-batch. Although the discriminator is trained from scratch, it is stable in converging, as it only encodes solutions where the vocabulary only contains placeholders for numbers and operators.

Negative Solution Generator For generating negative solutions, any manipulation of the ground truth solution can lead to a negative one, as implemented in (Liang and Zhang, 2021). However, the random modification neglects the importance of tokens in the solution equation. Although all negative solutions ultimately lead to a wrong answer, the roles they play in minimizing loss functions and serving as contrastive examples to a positive true solution are at different levels of importance.

Our goal is to find variants of the ground truth solution as hard negative samples, which only manipulate the most vulnerable (important) token. In fact, this target is similar to evasion attack (Carlini and Wagner, 2017) on texts, i.e., maximum effect and minimum manipulation. Therefore, borrowing the idea from white-box evasion attack, we regard the token with the largest gradient as the most important and vulnerable one:

$$y_i = \underset{y_i \in Y}{argmax} \ (\nabla Dis([en_x(X), en_y(Y)]). \quad (4)$$

After getting the most important y_i, the next step is to find the replacement token of it. Different from (Liang and Zhang, 2021) which generates it randomly, we notice that the vocabulary for Y is in a small size. Also, the token type has to be consistent after replacement (operator or number). Therefore, it is not costly to find all possible alternatives of token y_i and make them all as negative equations. If y_i is an operator, we have 4 alternatives out of $\{+, -, *, /, \hat{}\}$, leading to 4 negative equations. If y_i is a number, the choices of replacement will depend on the size of V_{num} which is usually less than 5. In this way, we make full use of the result of gradient checking and have considered all possibilities of token replacement to form negative solutions.

Additional Negatives The above-generated negative equations are similar to the ground truth solution, since most part of it has not been changed. To generate diverse negative samples that have loose connections to the ground truth, we also randomly select a solution from the training set as a negative equation, subject to the vocabulary of that solution being a subset of the current solution vocabulary. Because the output vocabulary may vary in different problems, this restriction ensures the generation of reasonable negative samples. Furthermore, we execute the same procedure on this random negative solution, i.e., finding and manipulating the most vulnerable token to obtain more negative samples. To sum up, the negative equation generator of the solution discriminator module will generate two groups of negative samples which are variants of the ground truth solution and a random solution with its variants. After the discriminator Dis gets well-trained by cross-entropy loss, we can construct a new loss term that provides extra supervision to the encoder of the MWP solver, which is

similar to the adversarial loss in the generative adversarial network (GAN) (Goodfellow et al., 2014):

$$L_s = -\log P(s_d = 1|Dis([en_x(X), en_y(Y)])). \quad (5)$$

The solution discriminator thus enhances the association between problem representations and their ground truth equation representations.

3.5 Model Training

There are three components in our proposed framework. (a) An analogy identification module that is jointly trained with the encoder-decoder solver, which plays the role of an auxiliary task. (b) A solution discrimination module that is separately trained with the solver, which ensures the consistency between MWP and its true solution, and provides guidance when the encoder-decoder part is trained. (c) The solver is trained by the summation of Seq2Seq loss, analogy identification loss (Equation 2), and solution discrimination loss (Equation 5), where the Seq2Seq loss is the negative log-probability of generating Y given X:

$$L_{seq} = -\log P(Y|X). \quad (6)$$

The steps of the training pipeline is in Alg. 1. Firstly, we randomly sample some positive pairs and negative pairs according to the introduction of analogical identification in Section 3.3. Secondly, we calculate the gradient as shown in Equation (4), find all vulnerable tokens, and generate all negative equations for solution discrimination. Then we train the discriminator with a cross-entropy loss with ground truth as the positive sample and manipulated solutions as the negative samples (line 3-5). Finally, we train the analogy identification module and the solver with the joint loss (line 6).

4 Experimental Results

In this section, we firstly introduce the used datasets, evaluation metrics and baselines in this paper. Then we conduct an accuracy comparison between our solvers and all baseline methods. Next, our ablation studies show the contribution from different modules. We also analyze the solver accuracy under different solution lengths to evaluate generalization ability. Visualization of different groups of MWPs shows that our solver has better MWP representation ability. Case studies can be found in appendix.

Algorithm 1 Training Pipeline

Input: Problem X, Solution Y, Analogy Identification Module θ, Solution Discrimination Module γ, Encoder-Decoder Module δ
Parameter: Loss Weights λ_1, λ_2
Output: Well-trained θ, γ and δ

1: **for** X, Y in the training set **do**
2: $X_{pos}, X_{neg} \leftarrow$ Sample positive/negative problems from the training set by analogy identification.
3: $y_i \leftarrow$ Get the most important token in Y by gradient as shown in Eq. (4).
4: $Y_{neg} \leftarrow$ Replace y_i with all possible tokens and get the negative solution set.
5: Train the discriminator γ with cross-entropy loss.
6: Train the analogy identification module (i.e., MLP) γ and encoder-decoder module δ with a joint loss: $L = L_{seq} + \lambda_1 L_a + \lambda_2 L_s$.
7: **end for**
8: **return** θ, γ, δ

4.1 Datasets

Math23k Math23k (Wang et al., 2017) is the most commonly used Chinese dataset in MWP solving. It contains 23,162 problems with 21,162 training problems, 1,000 validation problems and 1,000 testing problems. We use the value accuracy as the evaluation metric, which checks whether the equation solution given by the model leads to the ground truth value.

MathQA MathQA (Amini et al., 2019) is an English mathematical problems dataset at GRE level. The original MathQA dataset is annotated in a different way from Math23k with many pre-defined operations. Also, some problems in the dataset suffer from the low-quality issue. Many efforts (Tan et al., 2021; Li et al., 2021; Jie et al., 2022) have been paid to clean and filter the MathQA dataset. In our experiment, we follow the latest version (Jie et al., 2022) of MathQA dataset where 4 arithmetic operators $\{+, -, *, /\}$ are included in this subset. After data filtering, the training set contains 16191 training MWPs and 1605 testing samples. We also use the value accuracy as our evaluation metric for this dataset.

4.2 Baselines

The baselines used in this paper can be divided into RNN-based and PLM-based. For RNN-based

	Math23k	MathQA	#E
RNN-solvers			
GTS	75.6	68.0*	7.2M
GTS-teacher	76.5	68.5*	7.2M
Graph2Tree	77.4	69.5	9.0M
EEH-G2T	**78.5**	–	9.9M
GTS+A&D	78.2	**69.6**	7.2M
PLM-solvers			
REAL	82.3	–	110M
BERT-CL	83.2	73.8	102M
MWP-BERT	84.6	77.2*	102M
Deductive Reasoner	85.1	78.6	102M
Generate&Rank	85.4	–	610M
MWP-BERT+A&D	**85.6**	**79.6**	103M

Table 1: Comparison of answer accuracy (%) and the number of parameters in the encoder (#E). '+A&D' means that the solver is trained with our proposed analogical pipeline containing analogy identification and solution discrimination. Accuracy with a '*' indicates re-produced results by us based on the provided public source code. The other results come from their papers.

solvers, we select GTS (Xie and Sun, 2019) and Graph2Tree (Zhang et al., 2020b) which are the most commonly used baselines in other papers. Besides, we include GTS-teacher (Liang and Zhang, 2021) which pairs up the problems and solutions and checks them with a discriminator, and EEH-G2T (Wu et al., 2021) is the best existing RNN-based solver on Math23k. For PLM-based solvers, we choose the retrieval-based analogical solver REAL (Huang et al., 2021), contrastive learning solver BERT-CL (Li et al., 2021) that determine analogies based on sub-trees in the solution, Deductive Reasoner (Jie et al., 2022) and Generate&Rank (Shen et al., 2021). To our knowledge, Generate&Rank is known as the best solver on Math23k (85.6% accuracy) and Deductive Reasoner performs the best on MathQA (78.6%).

4.3 Implementation Details

Our model is trained and evaluated under Pytorch Framework with a single NVIDIA Tesla V100 GPU. The training iteration follows the steps in Alg. 1 where $\lambda_1 = 0.01$ and $\lambda_2 = 0.001$, which are selected by grid search from [0.0001, 0.001, 0.01, 0.1, 1]. We train the solvers with 160 epochs using a batch size of 32. The learning rate for RNN-based models and PLM-based models is set to 0.001 and 0.00005, which is halved every 30 epochs. One-epoch training takes about 10 minutes and 30 minutes on GTS-based and BERT-based solvers, respectively. AdamW (Kingma and Ba, 2014) optimizer is used with default hyper-parameters in Pytorch. Dropout (Srivastava et al., 2014) probability of 0.5 and 0.1 is used for GTS-encoder and BERT-encoder to avoid potential overfitting. Five-beam search is applied to produce better solutions. For more details, please refer to our code[2].

4.4 Comparative Experiments

In this part, we compare our model with the above baselines in terms of the value accuracy and the number of parameters in the encoder, as shown in Table 1. On Math23k benchmark, the performance of GTS gets considerably raised from 75.4% to 78.2%, which is comparable to the best RNN solver EEH-G2T with fewer parameters in the encoder. For PLM-based solvers, our proposed training method boosts the accuracy of MWP-BERT from 84.6% to 85.6%, achieving state-of-the-art performance with much fewer parameters than Generate&Rank. For MathQA dataset, with our pipeline, the accuracies of GTS and MWP-BERT both get improved and the improved MWP-BERT solver outperforms all other competitors. For some baselines, we failed to re-produce their methods and get a reasonable accuracy on MathQA, thus we leave them empty as many previous papers (Zhang et al., 2020b; Li et al., 2021; Liang et al., 2022) did.

4.5 Ablation Studies

There are two major modules in our framework, i.e., analogy identification and solution discrimination. For the former part, we separately test the influence of top-$k(k \leq 3)$ nodes and find that only top-1 and top-2 have contributions. The reason for the ineffectiveness of top-3 nodes could be that only a small number of MWPs contain more than 3 operators (shown in Table 3), and fewer MWPs would have 3 same operators at the top. Therefore, there will be a considerately limited number of analogical pairs for top-3 MLP. For the solution discrimination module, we evaluate the influence of gradient-guided token selection (we use random token selection in its ablation) and the additional negative solutions that are randomly drawn out of the training set. The results in Table 2 verify the contribution of every individual component to the

[2] https://github.com/LZhenwen/Analogical_MWP

	Analogy Identification (A)			Solution Discrimination (D)		Acc.
	Top-1 MLP	Top-2 MLP	Top-3 MLP	Gradient guided	Extra Negatives	
MWP-BERT	✗	✗	✗	✗	✗	84.7
	✓	✗	✗	✗	✗	84.9
MWP-BERT+A	✓	✓	✗	✗	✗	85.3
	✓	✓	✓	✗	✗	85.1
	✗	✗	✗	✓	✗	85.0
	✗	✗	✗	✗	✓	85.1
MWP-BERT+D	✗	✗	✗	✓	✓	85.3
MWP-BERT+A&D	✓	✓	✗	✓	✓	**85.6**

Table 2: Accuracy of different ablated models on the Math23k dataset. 'A' denotes analogy identification and 'D' denotes solution discrimination. ✓and ✗ indicate without/with that module in the ablated models, respectively.

#OP	Pct.	GTS		MWP-BERT	
		w/o	w	w/o	w
1	35.4%	84.9	86.1	90.1	**90.7**
2	44.0%	80.6	81.7	87.0	**88.0**
3	14.1%	70.7	71.3	62.4	**71.3**
4	3.3%	50.0	60.6	60.6	**78.8**
5*	0.6%	38.2	38.2	50.0	**50.0**

Table 3: The answer of BERT and MWP-BERT on problems with different lengths in Math23k. #op shows the No. of operators in the solution. Pct. is the percentage of problems with different number of operators. 'w/o' means "without our analogy identification and solution discrimination", and 'w' denotes "with our method". * indicates that the result is not statistically meaningful for comparison because there are only 6 samples.

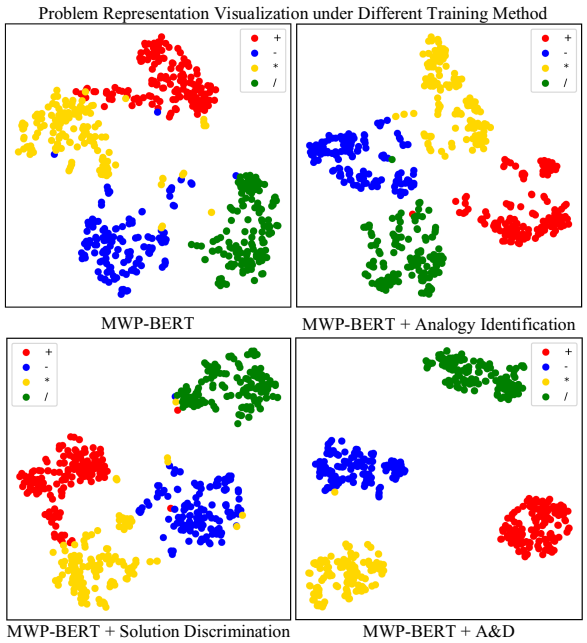

Figure 3: Problem representation visualization by T-SNE. Our model with A&D improves the problem representation learning, which groups analogical problems close and separates non-analogical problems.

proposed training pipeline.

4.6 Analysis of Long-tail Problem Solving

In Table 3, the column Pct. gives the percentage of MWPs with different solution lengths. We can see that the distribution of MWPs follows a long-tail distribution where the simple MWPs with short solutions are the head, and the difficult MWPs with long solutions are the tail. We find that the accuracy on the tails gets more increased than the accuracy on the head with the assistance of our method. Although the overall accuracy improvement on Math23k is about 1% as shown in Table 2, our method makes the solver learn from simple MWPs and generalize to difficult MWPs, bringing a significant accuracy boost on tail problems.

4.7 Visualization

To demonstrate the ability of grouping analogical problems in the representation space, we visualize them by T-SNE (Van der Maaten and Hinton, 2008) and check the clustering phenomenon as did in (Liang and Zhang, 2021; Li et al., 2021). We randomly select 150 problems from Math23k in 4 classes, where each class represents a group of MWPs whose solutions has the same root nodes out of $\{+, -, *, /\}$. Then we visualize them in Figure 3. For vanilla MWP-BERT, many boundary points are not well separated. With the analogy identification, the number of confusing boundary points gets lower. With the solution discrimination module, the inter-class distance is increased and the green cluster is well separated from the others. With both, different clusters are more separated and

Case 1

	Problem Description	Solution
The source problem in testing set	A team is making 2640 cartons, and 240 cartons have been made in 3 days. How many days will it still take to finish the job?	MWP-BERT: (2640)/(240/3) Ours: (2640-240)/(240/3)
Analogical problems in the training set	An engineering team is building a road that is 4.8 km long, and has already built 0.48 km in the first 4 days. How many days will it still take to finish the job?	(4.8-0.48)/(0.48/4)
	A team need to install 480 meters of water pipe, and 120 meters of that pipe have been installed in first 4 days. How many days will it still take to finish the job?	(480-120)/(120/4)

Case 2

	Problem Description	Solution
The source problem in testing set	Fang is reading a 290-page book and plans to finish it in 7 days. She has read 20 pages per day for the first 4 days. How many pages should she read each day to finish it as planned?	MWP-BERT: (290-20*4)/7 Ours: (290-20*4)/(7-4)
Analogical problems in the training set	Ming can finish reading a storybook in 16 days, if he reads 15 pages per day. How many pages should he read each day if he wants to finish it in 10 days?	16*15/10
	Ming reads an a 120-page book, and he reads 18 pages per day for the first 3 days. How many pages should he read each day if he wants to finish it in the next 6 days?	(120-3*18)/6

Figure 4: Case study.

the distribution of each cluster gets denser. This analysis demonstrates that our model effectively improves problem representation learning.

Case Study We show two cases in Figure 4 to demonstrate the potency of our proposed method. We select two problems that MWP-BERT (Liang et al., 2022) failed to solve from the Math23k testing set. We can find that MWP-BERT solver is influenced by analogical problems in the training set and generates an identical solution as the solution of the second analogical problem. The potential reason is that these MWPs are semantically similar and the model fails to distinguish them. In our approach, we design the solution discrimination module to strength the association between the problems and ground truth solutions. Our solver correctly generates the equation through the help of the designed module.

5 Conclusion

We propose a novel analogical training pipeline for math word problem solvers, considering the generalization among analogical MWPs and the association between ground truth solutions and problem descriptions. In this way, problems that need similar solving skills are grouped together to support analogical learning, and solvers are trained to focus on ground truth solutions. The comparative study shows that our method with MWP-BERT outperforms other baselines on Math23k and MathQA, and is much lighter (with fewer parameters). It can also solve more difficult problems due to the analogical reasoning. We believe our work would facilitate the MWP research community and inspire more studies about the analogy in mathematical question answering.

Limitations

Commonsense Knowledge As mentioned in (Lin et al., 2020; Liang et al., 2021), MWP solving in the real-word scenario requires many commonsense knowledge, e.g., 1km = 1000m and one day = 24 hours. When these commonsense constants are not explicitly given in the problem description, our MWP solver has no chance to solve problems that require them. A future direction could be injecting commonsense knowledge into MWP solvers.

Acknowledgement

The research work is partially supported by the Internal Asia Research Collaboration Grant, University of Notre Dame. Thanks for all reviewers for their valuable comments.

References

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *NAACL*, pages 2357–2367.

Yefim Bakman. 2007. Robust understanding of word problems with extraneous information. *arXiv preprint math/0701393*.

Allan BI Bernardo. 2001. Analogical problem construction and transfer in mathematical problem solving. *Educational Psychology*, 21(2):137–150.

DG Bobrow. 1964. Natural language input for a computer problem solving system. *Ph. D. Thesis, Department of Mathematics, MIT*.

Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *EMNLP*, pages 1724–1734.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. In *EMNLP: Findings*, pages 657–668.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Keyur Faldu, Amit Sheth, Prashant Kikani, Manas Gaur, and Aditi Avasthi. 2021. Towards tractable mathematical reasoning: Challenges, strategies, and opportunities for solving math word problems. *arXiv preprint arXiv:2111.05364*.

Dedre Gentner and Keith J Holyoak. 1997. Reasoning and learning by analogy: Introduction. *American psychologist*, 52(1):32.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*, pages 2672–2680.

Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. 2021. Efficient nearest neighbor language models. In *EMNLP*, pages 5703–5714.

Yining Hong, Qing Li, Ran Gong, Daniel Ciao, Siyuan Huang, and Song-Chun Zhu. 2021. Smart: A situation model for algebra story problems via attributed grammar. In *AAAI*.

Shifeng Huang, Jiawei Wang, Jiao Xu, Da Cao, and Ming Yang. 2021. Recall and learn: A memory-augmented solver for math word problems. In *Findings of EMNLP*, pages 786–796.

Qingnan Jiang, Mingxuan Wang, Jun Cao, Shanbo Cheng, Shujian Huang, and Lei Li. 2021. Learning kernel-smoothed machine translation with retrieved examples. In *EMNLP*, pages 7280–7290.

Zhanming Jie, Jierui Li, and Wei Lu. 2022. Learning to reason deductively: Math word problem solving as complex relation extraction. In *ACL*, pages 5944–5955.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. In *ICLR*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015. Parsing algebraic word problems into equations. *Transactions of the Association for Computational Linguistics*, 3:585–597.

Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*.

Zhongli Li, Wenxuan Zhang, Chao Yan, Qingyu Zhou, Chao Li, Hongzhi Liu, and Yunbo Cao. 2021. Seeking patterns, not just memorizing procedures: Contrastive learning for solving math word problems. *arXiv preprint arXiv:2110.08464*.

Zhenwen Liang, Jipeng Zhang, Jie Shao, and Xiangliang Zhang. 2021. MWP-BERT: A strong baseline for math word problems. *CoRR*, abs/2107.13435.

Zhenwen Liang, Jipeng Zhang, Lei Wang, Wei Qin, Yunshi Lan, Jie Shao, and Xiangliang Zhang. 2022. Mwp-bert: Numeracy-augmented pre-training for math word problem solving. In *Findings of NAACL*.

Zhenwen Liang and Xiangliang Zhang. 2021. Solving math word problems with teacher supervision. In *IJCAI*.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! numersense: Probing numerical commonsense knowledge of pre-trained language models. In *EMNLP*.

Xin Lin, Zhenya Huang, Hongke Zhao, Enhong Chen, Qi Liu, Hao Wang, and Shijin Wang. 2021. Hms: A hierarchical solver with dependency-enhanced understanding for math word problem. In *AAAI*, pages 4232–4240.

Qi Liu, Dani Yogatama, and Phil Blunsom. 2022. Relational memory-augmented language models. *Transactions of the Association for Computational Linguistics*, 10:555–572.

Qianying Liu, Wenyv Guan, Sujian Li, and Daisuke Kawahara. 2019. Tree-structured decoding for solving math word problems. In *EMNLP*, pages 2370–2379.

Laura R Novick and Keith J Holyoak. 1991. Mathematical problem solving by analogy. *Journal of experimental psychology: Learning, memory, and cognition*, 17(3):398.

Jinghui Qin, Xiaodan Liang, Yining Hong, Jianheng Tang, and Liang Lin. 2021. Neural-symbolic solver for math word problems with auxiliary tasks. In *ACL*, pages 5870–5881.

Jianhao Shen, Yichun Yin, Lin Li, Lifeng Shang, Xin Jiang, Ming Zhang, and Qun Liu. 2021. Generate & rank: A multi-task framework for math word problems. In *Findings of EMNLP*, pages 2269–2279.

Shuming Shi, Yuehui Wang, Chin-Yew Lin, Xiaojiang Liu, and Yong Rui. 2015. Automatically solving number word problems by semantic parsing and reasoning. In *EMNLP*, pages 1132–1142.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Kaye Stacey, L Burton, and J Mason. 1982. *Thinking mathematically*. Addison-Wesley.

Minghuan Tan, Lei Wang, Lingxiao Jiang, and Jing Jiang. 2021. Investigating math word problems using pretrained multilingual language models. *arXiv preprint arXiv:2105.08928*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Stella Vosniadou. 1988. Analogical reasoning as a mechanism in knowledge acquisition: A developmental perspective. *Center for the Study of Reading Technical Report; no. 438*.

Lei Wang, Yan Wang, Deng Cai, Dongxiang Zhang, and Xiaojiang Liu. 2018. Translating a math word problem to a expression tree. In *EMNLP*, pages 1064–1069.

Lei Wang, Dongxiang Zhang, Jipeng Zhang, Xing Xu, Lianli Gao, Bing Tian Dai, and Heng Tao Shen. 2019. Template-based math word problem solvers with recursive neural networks. In *AAAI*, volume 33, pages 7144–7151.

Yan Wang, Xiaojiang Liu, and Shuming Shi. 2017. Deep neural solver for math word problems. In *EMNLP*, pages 845–854.

Qinzhuo Wu, Qi Zhang, Jinlan Fu, and Xuan-Jing Huang. 2020. A knowledge-aware sequence-to-tree network for math word problem solving. In *EMNLP*, pages 7137–7146.

Qinzhuo Wu, Qi Zhang, and Zhongyu Wei. 2021. An edge-enhanced hierarchical graph-to-tree network for math word problem solving. In *Findings of EMNLP*, pages 1473–1482.

Zhipeng Xie and Shichao Sun. 2019. A goal-driven tree-structured neural model for math word problems. In *IJCAI*, pages 5299–5305.

Zhicheng Yang, Jinghui Qin, Jiaqi Chen, Liang Lin, and Xiaodan Liang. 2022. Logicsolver: Towards interpretable math word problem solving with logical prompt-enhanced learning. *arXiv preprint arXiv:2205.08232*.

Dongxiang Zhang, Lei Wang, Luming Zhang, Bing Tian Dai, and Heng Tao Shen. 2019. The gap of semantic parsing: A survey on automatic math word problem solvers. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2287–2305.

Jipeng Zhang, Roy Ka-Wei Lee, Ee-Peng Lim, Wei Qin, Lei Wang, Jie Shao, and Qianru Sun. 2020a. Teacher-student networks with multiple decoders for solving math word problem. In *IJCAI*, pages 4011–4017.

Jipeng Zhang, Lei Wang, Roy Ka-Wei Lee, Yi Bin, Yan Wang, Jie Shao, and Ee-Peng Lim. 2020b. Graph-to-tree learning for solving math word problems. In *ACL*, pages 3928–3937.