An Augmented Benchmark Dataset for Geometric Question Answering through Dual Parallel Text Encoding

Jie Cao
School of Computer Science
South China Normal University
jiecao@m.scnu.edu.cn

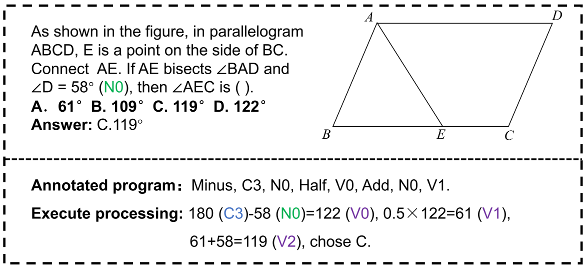
Jing Xiao*
School of Computer Science
South China Normal University
xiaojing@scnu.edu.cn

Abstract

Automatic math problem solving has attracted much attention of NLP researchers recently. However, most of the works focus on the solving of Math Word Problems (MWPs). In this paper, we study on the Geometric Problem Solving based on neural networks. Solving geometric problems requires the integration of text and diagram information as well as the knowledge of the relevant theorems. The lack of high-quality datasets and efficient neural geometric solvers impedes the development of automatic geometric problems solving. Based on GeoQA, we newly annotate 2,518 geometric problems with richer types and greater difficulty to form an augmented benchmark dataset **GeoQA+**[1], containing 6,027 problems in training set and 7,528 totally. We further perform data augmentation method to expand the training set to 12,054. Besides, we design a **D**ual **P**arallel text **E**ncoder (**DPE**) to efficiently encode long and medium-length problem text. The experimental results validate the effectiveness of GeoQA+ and DPE module, and the accuracy of automatic geometric problem solving is improved to 66.09%.

1 Introduction

In recent years, with the continuous development of deep learning technology in NLP, more and more math problem solvers have been developed. However, most of these works focus on solving arithmetic and algebra problems (Xie and Sun, 2019; Lin et al., 2021; Wu et al., 2020). There are few systems for geometric problem solving, especially those based on the method of the neural networks. The solving of geometric problems requires a combination of text and diagram information, and therefore the study of it also helps to promote the development of cross-modal problem-solving.


As shown in the figure, in parallelogram ABCD, E is a point on the side of BC. Connect AE. If AE bisects ∠BAD and ∠D = 58° (N0), then ∠AEC is ( ).
A. 61° B. 109° C. 119° D. 122°
Answer: C.119°

Annotated program: Minus, C3, N0, Half, V0, Add, N0, V1.
Execute processing: 180 (C3)-58 (N0)=122 (V0), 0.5×122=61 (V1), 61+58=119 (V2), chose C.

Figure 1: A typical geometry problem in GeoQA+ dataset and its annotating and solving process.

As shown in Figure 1, a typical geometry problem mainly consists of textual descriptions and geometric diagrams. There are three steps to solve this problem. First, the text and diagram information are encoded separately. Second, the solver needs to understand the semantics of the text and diagram information simultaneously. Third, in order to solve the problem, we may need to combine the information of Text-Diagram with relevant theorem knowledge. For example, the problem in Figure 1 use the theorem of complementary adjacent angles of a parallelogram. Though some previous methods attempt to solve geometric problems automatically, the performance of their solving system is far away from satisfactory (Seo et al., 2014, 2015; Sachan and Xing, 2017). They highly relied on limited handcraft rules and were only validated on small-scale datasets, making it hard to generalize to more complex and real-world cases(Chen et al., 2021). In this case, we mainly focus on building an efficient solving system based on neural networks.

To resolve the mentioned issues, Chen et al. (2021) proposed a geometric problems dataset GeoQA which contains 5,010 geometric problems and the first neural network-based geometric problems solving system NGS. However, we believe that there are some limits to this work. First, we think that the problem type in GeoQA is not rich enough, and it only contains angle and length

[*] Corresponding Author

[1] The source code and benchmark of this paper are available at: https://github.com/SCNU203/GeoQA-Plus

Proceedings of the 29th International Conference on Computational Linguistics, pages 1511–1520
October 12–17, 2022.

problems as well as a very small number of other types. Second, we think that the problems in GeoQA are not difficult enough, and the average solving step of the problems in GeoQA is only 1.96. Third, the geometric problems solver NGS can not effectively solve the problems with long text for the lack of text feature extraction capability.

Inspired by the exiting works (Chen et al., 2021; Seo et al., 2014, 2015), to refresh the research on geometric problem solving and further promote the development of cross-modal numerical reasoning, we newly annotate 2,518 geometric problems containing 636 area-type problems that are not included in GeoQA. The problems we collect are more difficult. The average solving step of our problems is 2.61, which compares to 1.96 of GeoQA. We add our new dataset to the training set of GeoQA to build a new dataset named **GeoQA+** and it contains 7,528 problems in total and 6,027 for training. To the best of our knowledge, GeoQA+ is the largest benchmark dataset for geometry problem solving at present and it improves the overall difficulty and diversity of the original dataset. We further perform data augmentation method on GeoQA+, which expand the data size to 12,054 to obtain more diverse data. As for the model, we design a **D**ual **P**arallel **E**ncoder **DPE** that consists of RoBERTa (Liu et al., 2019) and a Bi-LSTM (Hochreiter and Schmidhuber, 1997) to address the limit of NGS. Our **DPE** module encodes long and medium-length problem text effectively, and we name this new geometric problems solver as **DPE-NGS**. We conduct a series of experiments and the experimental results indicate that the GeoQA+ dataset and our DPE-NGS model show the superiority over the state-of-the-art results.

In summary, our contributions are three-fold:

- To expand GeoQA, we newly annotate 2,518 geometric problems which are more difficult to solve than GeoQA and has more problem types to build a new dataset name GeoQA+, the largest dataset for geometric problem solving at present. In addition, we also perform data augmentation work on GeoQA+ to obtain more diverse data.

- To alleviate the limit of NGS, we design a Dual Parallel Encoder(**DPE**) and propose **DPE-NGS** to effectively solve the geometric problems with long and medium-length text. Experimental results show that our model achieves better accuracy.

- We study the text encoding work of geometric problems. We fine-tune the Pre-training model using a sufficient amount of data for the first time and achieve excellent model performance.

2 Related Work

Geometric Problems Solving Having machine to solve geometric problems has a long history in AI (Wen-Tsun, 1986; Chou et al., 1996). Some researchers proposed methods for geometry theorem proving based on rule-based methods last century(Wenjun, 1984). Wong et al. (2007) designed the first automatic solver LIM-G for geometric problems, but this method was only based on text information to solve the problems. Subsequently, Seo et al. (2014, 2015) constructed the first automatic problems solver that combines text and diagram information with NLP methods and computer vision technology (OCR). However, this method relies too much on handcrafted rules, and it was only verified on the data set with 185 problems. To improve GeoS, Sachan and Xing (2017) replaced these handcraft constraints with geometry axiomatic knowledge in the form of horn-clause rules, but their dataset and code are not released. Lu et al. (2021) proposed Inter-GPS which achieved higher accuracy than all previous geometric problem solvers based on rule-based methods. And their dataset Geometry3k contains 3,002 geometric problems. But Inter-GPS was still designed based on the rule-based method and Geometry3k is not suitable for training neural network-based solvers because of the complex annotating work. Aiming to improve the performance and interpretability of existing models, Chen et al. (2021) proposed a geometric problems dataset GeoQA, and they proposed the first geometric problems solver based on neural networks named NGS. While the GeoQA dataset is not difficult and diverse enough, and the feature extraction ability of NGS is also not good enough. To improve the limits of existing works and promote the development of automatic geometric problem solving, we effectively expand GeoQA dataset and propose DPE-NGS model.

Multimodal Reasoning Visual question answering is a typical multimodal problem. The solving of this kind of problem often requires the model to have a certain reasoning ability (Goyal et al., 2017; Yu et al., 2019). On this basis, some methods propose an implicit reasoning framework to jointly encode multimodal information (Perez et al., 2018; Cohen and Areni, 1991). However, geometric problem solving is more logical and deductive, and the solving process requires additional knowledge of theorems, so these visual problem answering models are not directly applicable to geometric problem solving.

Pre-training Model In NLP Pre-training models have greatly advanced the development of NLP (Song et al., 2021; Zhang et al., 2020). And it has also been applied in the automatic solving of MWPs (Liang et al., 2021). However, it has not been applied to the automatic solving of geometric problems since the lack of dataset. Our experimental results show that the introduction of Pre-training models facilitates the possibility of solving geometric problems based on our newly annotate dataset.

Text Data Augmentation Text data augmentation methods has been widely used in NLP, like EDA(Wei and Zou, 2019) and back translation(Yu et al., 2018) method. We also perform the back-translated method on GeoQA+. We first translate the original data into minor languages and then re-translate the results into the original language. Data back translation enhances the diversity of data.

3 GeoQA+ Dataset

The original GeoQA dataset contains 5,010 geometric problems, 3,509 for training, 746 for validation and 755 for test. We newly annotate 2,518 geometric problems and add them to GeoQA's training set to form a new dataset GeoQA+ which contains 6,027 geometric problems in training set and 7,528 in total.

3.1 Problem and Data Description

Problem Description. Automatic geometry problem solving is defined as solving a geometry problem with diagram and text information. Text-Diagram information are encoded by text and diagram encoder separately, then the encoded results are fused with features from both parts through the Joint Reasoning Module. The decoder module obtains the solving sequence by decoding the output from Joint Reasoning Module, then executes the sequence and gets the answer with additional knowledge of theorems. Figure 1 shows the complete problem definition, and the solving processing uses the knowledge of the properties of parallelograms.

Data Description. Based on the problem definition, we define the data description of the geometry problem, which contains problem text t, diagram d, problem choices c, knowledge points k, problem answer a, solving processing explanation e, and the annotate programs p. Therefore, a geometry problem can be represented as $T(t, d, c, k, a, e, p)$ like Figure 1.

Program Representation. We adapt a domain-specific language(Amini et al., 2019) to represent the geometric problems solving process similar to GeoQA. The program includes the operator OP, operand N, constant operand C, and process variable V. We enrich the representation of the language by synthesizing the data statistics of our newly annotate data. As shown in Table 1, the operators OP are divided into basic and arithmetic operators as well as trigonometric and theorem operators. The constant operands contain various constants such as π, 180°, and 90° that are commonly used. Note that only operators and constant operands are given in Table 1 because both of them are fixed and will not change with different problems. For example, when we solve for the length of the hypotenuse of a right triangle with two right-angled sides known, we will use the Pythagorean theorem to solve the problems, and we need to know the fixed expression of the Pythagoras operation. The operand N is derived from the operands given in the problems and the process variable V is an intermediate variable generated during the operation, both of which vary from problem to problem. The generated sequence expressions of the model show the interpretability of the solving process. We can get a general understanding of the whole problem-solving process from Figure 1.

3.2 Dataset Comparison

The existing geometry problem datasets are generally limited by the size of the data(Seo et al., 2015) and the complex annotating work (Lu et al., 2021), which are not suitable for neural network training. GeoQA is a dataset collected specifically for building a neural network-based geometry problem solver. However, the limits of the

OPR & Const	Programs
Basic	Equal, Double, Half
Arithmetic	Add, Minus, Multiply, Divide, Prescription
Trigonometric	Sin, Cos, Tan, Arc-Sin, Arc-Cos
Theorem & Formula	Pythagorean Add/Minus, Proportion, Circle Area,Circle Perimeter, Cone Area
Constant	30, 60, 90, 180, 360, 540, Π, 0.618

Table 1: An overview of 19 operations of four different types and 8 constants in the defined program set.

GeoQA dataset are the low average difficulty of problem solving and the lack of richness of problem types. Therefore, we newly annotate a dataset with 2,518 problems and add them to the training set of GeoQA to form a new dataset with 7,528 problems, 6,027 problems for training. Compared with GeoQA, our geometry problems are more difficult, and we introduce area-type problems for the dataset. For geometry problems, difficult problems often contain more geometric relationships and geometric attributes than simple problems, and we believe that learning more features of difficult samples help the model to solve difficult problems in the real world. A detailed comparison of the data statistics of the GeoQA's training set with our newly annotate data is shown in Table 2.

As shown in Table 2, our newly annotate dataset introduces 636 problems of area-type that are not available in GeoQA which enhance the data diversity of the dataset. In addition, our dataset are more difficult with 2.61 steps of average solving compare with 1.96 of GeoQA. More solving steps means the problems are more difficult to solve. Besides, our newly annotate problems also add 27 new knowledge points, and there are 77 knowledge points in GeoQA+. The knowledge points of a problem are crucial for solving the question. During the solving process, our model will first apply a Pre-trained module to predict the knowledge points of the problem which helps generating the solving sequence.

As shown in Table 3, the total number of training set in GeoQA+ is 6,027, and the average number of solving step is 2.23, which is nearly 14% higher than the original 1.96, meaning GeoQA+ is much more difficult than GeoQA. More difficult training samples facilitate the model to learn more statistics to improve the ability to solve difficult problems. We name this new training set as **Mix-train**.

3.3 Data Augmentation

We use the back-translation method in this paper to perform data augmentation on our Mix-train training set. We first translate the Mix-train training set data into French and then re-translate the results back to the original Chinese, and finally, we get a back-translated dataset with twice the amount of data, and we name the Back-translated dataset **Backtrans-train** which contains 12,054 problems.

3.4 Data Collection and Annotation

We collect our problems from online education websites. These problems are oriented in grades 6-12, containing various types of problems with corresponding knowledge points and solving explanations. We organized several graduate students to participate in annotating these problems. Each graduate student involved in the data annotation was trained to ensure that the data was annotate consistently with GeoQA. Unlike GeoQA, we allow the existence of problems with up to 8 solving steps while the authors of GeoQA limit the solution steps to 4. We believe that the introduction of difficult problems with long solving steps is beneficial to enhance the inference and generalization ability of the model.

4 Models

To improve the limit of NGS, we redesign the text encoder module. We refer to this improved geometric problems solver as **DPE-NGS**, and the overall structure of DPE-NGS is shown in Figure 2.

4.1 Dual Parallel Text Encoder

Text modeling is commonly used in NLP tasks such as sentiment analysis, topic classification, and problem systems (Li et al., 2020). In previous work, for solving geometric problems, researchers have often encoded the text by rule-based methods (Wong et al., 2007; Seo et al., 2015; Lu et al., 2021). In NGS, an LSTM(Hochreiter and Schmidhuber, 1997) was used to encode the problem text and represented the text as hidden state H in LSTM.

However, by analyzing the statistics of the problems that NGS did not get the result (**No Result** problems), we found that the average problem text length for this category is 68.55, which is much longer than the average problem text length of 52.5

	Properties	Angle	Length	Area\Others	AVG
GeoQA-train	**Number**	1939	1303	267	/
	OP-AVG	1.83	2.10	2.03	1.96
Ours	**Number**	1256	626	636	/
	OP-AVG	**2.78**	**2.27**	**2.60**	**2.61**

Table 2: Comparison of the data statistics of GeoQA-train and our newly annotate data. OP-AVG represents the average solving step of problems.

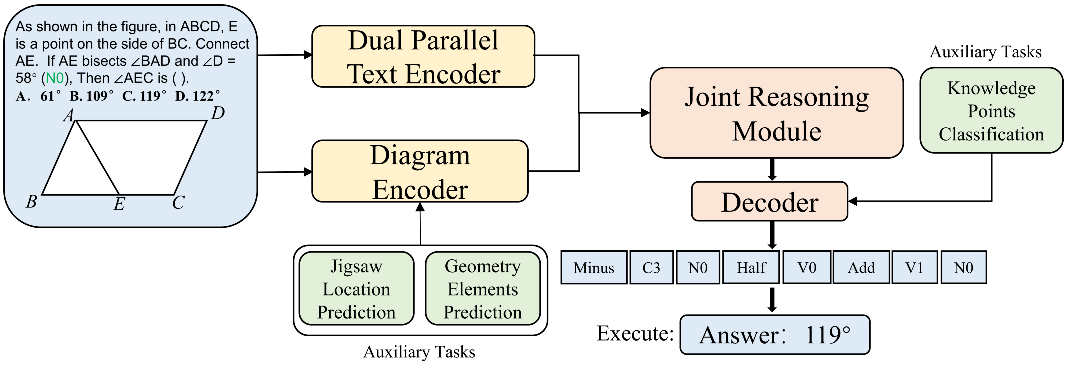

As shown in the figure, in ABCD, E is a point on the side of BC. Connect AE. If AE bisects ∠BAD and ∠D = 58° (N0), Then ∠AEC is ( ).
A. 61° B. 109° C. 119° D. 122°
A
D
B
E
C
Dual Parallel Text Encoder
Diagram Encoder
Jigsaw Location Prediction
Geometry Elements Prediction
Auxiliary Tasks
Joint Reasoning Module
Auxiliary Tasks
Knowledge Points Classification
Decoder
Minus
C3
N0
Half
V0
Add
V1
N0
Execute:
Answer: 119°

Figure 2: Our **DPE-NGS** for geometric problems solving based on Dual Parallel Text Encoder (**DPE**). The model encodes text and image information separately, and then feeds them to the Joint Reasoning Module. The decoder generates the solving sequence based on the output of Joint Reasoning Module, and the executor module finally executes the sequence and gets the answer.

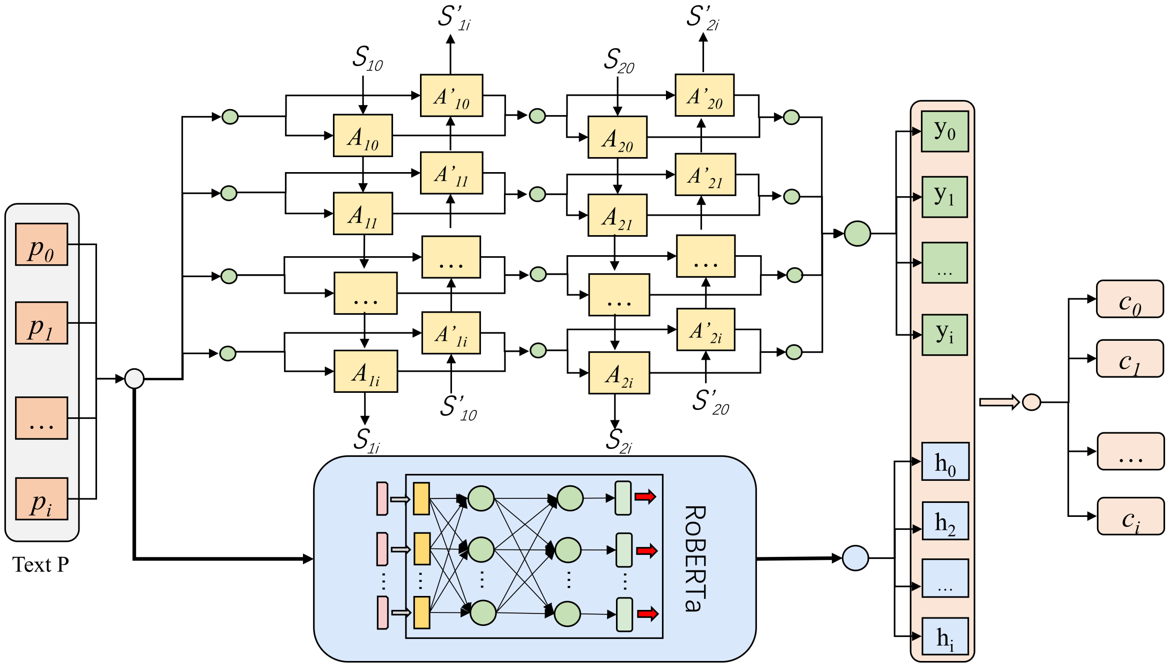


Figure 3: The architecture of our **Dual Parallel Encoder**. We use a two-layer Bi-LSTM and RoBERTa to encode the problem text separately. The encoding results are fed into a fusion layer, and we use the fused encoding information as the final text encoding result.

in GeoQA. This indicates that NGS is still lacking in feature learning for long text problems. In this case, we believe that the problem solving is related to the length of the problem text. In order to distinguish different text lengths, we regard the problems with text lengths between 30∼50 as medium-length problem, and long text problems are those with text lengths more than 50. And there are 2961

	Angle	Length	Others	AVG\Total
Number	3195	1929	903	6027
OP-AVG	2.21	2.15	2.45	2.23

Table 3: Statistics of the new training set in GeoQA+ (Mix-train).

long text problems and 2853 medium-length problems in Mix-train.

The problem text of a geometry problem usually contains many geometric elements expressed, and there are relational dependencies between these geometric elements. As the problem shown in Figure 1, in this example, parallelogram $ABCD$ is the first geometric element that mentions in the text, and $\angle D$ appears later. But in the process of solving the problem, we need to combine the two conditions that $ABCD$ is a parallelogram (parallelogram neighbors are complementary) and $\angle D = 58°$ to derive the next condition $\angle BAD = 122°$ to solve the problem. This is a back-and-forth process in which the key information of a geometry problem interacts with each other. Therefore, we cnsider that we should encode the problem text in a bidirectional way.

Based on the analysis above, we redesign the text encoder and we first introduce the Pre-training model RoBERTa (Liu et al., 2019) as text encoder. RoBERTa is a Bert-based(Devlin et al., 2019) Pre-training model that has been widely used and has greatly advanced various works in NLP. As shown in Table 4, when using RoBERTa as text encoder alone, the model solves 53.10% of long text problems and 69.74% of medium-length problems. Moreover, we also consider encoding problem text with a Bi-LSTM(Hochreiter and Schmidhuber, 1997) with two layers alone. In this case, the model solves 52.49% of long text problems and 70.25% of medium-length text problems. As the results show that the model has a different ability to solve problems with different lengths of text when using RoBERTa or Bi-LSTM as text encoder alone. Specifically, the model solves more long text problems when RoBERTa is used as the encoder, and it performs better in solving medium-length text problems when using Bi-LSTM as encoder.

As the experimental results show, we believe that different encode module have different feature extraction capabilities for various lengths of text during automatic geometry problem solving.

Text Encoder	long(%)	medium-length(%)
LSTM(NGS)	50.37	69.48
RoBERTa	53.10	69.74
Bi-LSTM	52.49	70.26
DPE	**57.48**	**72.56**

Table 4: The ability of the model for different length text problems when using different modules as encoders

To fully extract the features of problems text, we consider combining RoBERTa and Bi-LSTM to form a parallel text encoder. We input the problems text into RoBERTa and Bi-LSTM to encode the text separately. We denote the encoding result of RoBERTa as $H_p = [h_0;...;h_n]$, and we represent the encoding result of Bi-LSTM as $Y_p = [y_0;..;y_n]$. After obtaining the encode outputs H_p and Y_p from RoBERTa and Bi-LSTM, we combine the two sets of features by feeding H_p and Y_p into an Information Convergence layer, and obtain the fusion feature $C_p = [c_0;...;c_n]$. We use C_p to represent the final text encoding result:

$$C_p = [H_p, Y_p].$$

And we name this Dual Parallel Encoder module **DPE**, the structure of our encode module is shown in Figure 3.

Our model solves 57.84% long text problems and 72.56% medium-length problems with DPE as text encoder. Experimental results validate the effectiveness of our DPE encoder. The performance of models for solving problems with long or medium-length text when using different modules as text encoder is shown in Table 4.

4.2 Diagram Encoder

To get the diagram information of problems, we adapt the diagram encoder module based on ResNet (He et al., 2016) from NGS. Two auxiliary tasks (as shown in Figure 2) are applied to pre-train the diagram encoder, which significantly enhance the feature extraction capability of the diagram encoder. We formalize the feature matrix extracted by the diagram encoder as H_d.

4.3 Joint Reasoning Module

After obtaining the text feature C_p and diagram feature H_d, we feed them into the Joint Reasoning Module. In this paper, we use a common attention module named co-attention(Yu et al., 2019) with an attention mechanism for cross-modal data fusing and reasoning. This module consists of 12 self-attention units and 6 guide-attention units. We use the Dual Parallel Encoder output C_p from the text encoder and H_d from the diagram encoder as the input of Joint Reasoning Module. This module fuses and reasons the text-diagram information and outputs F_D, which contains abundant text and diagram information. We further concatenate C_p and F_D to get F_R for decoding program.

4.4 Program Decoder

We use an LSTM(Hochreiter and Schmidhuber, 1997) with attention as the Decoder module, which generates the programs sequentially under the guidance of Reasoning module output F_R. Let $y_t (1 \leq t \leq T)$ be the target program to be generated and P_t as the next program token. In the training process, we use the negative log-likelihood function as the loss function:

$$\mathcal{L}_g(\theta) = \frac{1}{T}\sum_{t=1}^{n} \log P_t(y_t|x, y_1, ...y_{t-1}; \theta),$$

where θ is the parameter of the entire solver model except for Diagram Encoder, and x is the input of the problem text and the diagram feature extracted from the Diagram Encoder.

4.5 Program Executor

The decoder module generates N program sequences [g_1,...,g_n], and the size of N equals to beamsize (beamsize = 10). The executor module selects the first sequence that successfully solves the problem as the prediction sequence. If all the results obtained by computing sequences are not included in the problem options, then the problem will be classified as a No Result problem instead of randomly selecting an answer.

5 Experiments

5.1 Experiment Setup

We conduct experiments on GeoQA and GeoQA+, and we adapt answer accuracy as the evaluation metric. We use the GeoQA-test containing 755 geometric problems as test set. In addition, since most previous work on automatic solving of geometric problems requires additional acceptance of input from OCR, but none of these works has published their associated codes, they are not compared with our methods in this experiment.

Implementation Details. We mention three datasets above: the original training set GeoQA-train with 3,509 problems, Mix-train with 6,027 problems after mixing GeoQA-train with our newly annotate dataset, and the Backtrans-train dataset with 12,054 problems after performing data augmentation on Mix-train. To verify the effectiveness of our datasets, we train our DPE-NGS and NGS with these three datasets separately and test the accuracy of the models on GeoQA-test. In addition, we train two models with GeoQA-train and test the generalization performance on the new test set(the same size as GeoQA-test) randomly extracted from our newly annotate data. Besides, we also compare the performance of a MWPs solver Seq2Prog(Amini et al., 2019), and BERT2Prog: Seq2Prog with BERT as encoder based on GeoQA-train[2]. The learning rate of ResNet is $1e{-}5$, $1e{-}3$ for Bi-LSTM encoder, and $2e{-}5$ for RoBERTa encoder, $1e{-}5$ for the rest. The batch size is 32 and the training epoch is 100.

5.2 Experimental Result

The effectiveness of our dataset. As shown in Table 5, when two models are trained with Mix-train or Backtrans-train, both models show better performance compared to the models train with GeoQA-train. The experiment results prove the effectiveness of our newly annotate dataset. In addition, the dataset after data augmentation is also helpful for accuracy improvement. We believe that because our dataset is more difficult and richer in problem types that expand the training set and makes up for the lack of difficult problems in GeoQA, which helps the models learn more problem features and thus improve the model's performance.

The effectiveness of our model. As shown in Table 5, our DPE-NGS outperforms all models for every training set. DPE-NGS with multimodal reasoning ability becomes the existing best-performing model (66.09%) on GeoQA-test set while train with Back-trains. We further analyze the percentage of No Result type problems generated by the models and found that DPE-NGS produces fewer No Result type problems than the NGS model as shown in Table 6. We believe it is because our DPE-NGS has better feature extraction ability for long text type problems. We also compare the accuracy of the two models for prob-

[2] Results obtained from the paper of Chen et al. (2021).

Traingsets	Model	Total(%)	Angle(%)	Length(%)	Others(%)	No Result(%)
GeoQA-train	**BERT2Prog**	50.3	63.4	33.2	38.9	/
	Seq2Prog	52.6	63.6	39.2	37.0	/
	NGS[3]	60.52	71.53	48.40	40.74	14.94
	DPE-NGS	**62.65**	**74.88**	**47.70**	**50.0**	**12.68**
Mix-train	**NGS**	61.19	72.25	47.70	46.30	12.72
	DPE-NGS	**65.96**	**75.60**	**54.42**	**51.85**	**11.90**
Backtrans-train	**NGS**	63.31	72.97	53.0	42.60	14.03
	DPE-NGS	**66.09**	**76.08**	**55.12**	**46.30**	**10.73**

[3] Results obtained from Chen's open source website: https://github.com/chen-judge/GeoQA

Table 5: Accuracy of the models on GeoQA-test using different training set.

Model	GeoQA-train(%)	Mix-train(%)	Backtrans-train(%)
NGS	14.94	12.72	14.03
DPE-NGS	**12.68**	**11.90**	**10.73**

Table 6: The percentage of **No Result** generated by the two models using different training set.

	OP=1(%)	OP=2(%)	OP=3(%)	OP=4(%)
NGS	76.70	58.42	47.10	38.33
DPE-NGS	**78.95**	**63.57**	**50.0**	**56.67**

Table 7: The accuracy of NGS and DPE-NGS for different difficulty problems using Mix-train. OP=N represents the solving steps of problems. More solving steps means the problem is more difficult.

Model	Total(%)	Angle(%)	Length(%)	Others(%)
NGS	49.14	53.85	45.90	43.07
DPE-NGS	**51.52**	**54.64**	**49.18**	**47.69**

Table 8: Accuracy of two models on our test set when train with GeoQA-train.

lems with different difficulty levels, as shown in Table 7, where our DPE-NGS outperforms NGS on all problems with different solving steps.

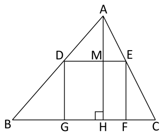
Generalization Performance of Models. We use the GeoQA-train dataset as training set for both models and test the generalization performance on our new test set. As shown in Table 8, since our annotate data are more difficult, neither model achieves a high accuracy, but our DPE-NGS still performs better than NGS, and our model achieves 51.52% compared to 49.14% of NGS indicating that our model shows better generalization performance.

5.3 Ablation Study

To verify the rationality of our model structure design and the validity of the text encoding method. We consider four combinations: 1) NGS, with unidirectional LSTM as text encoder; 2) NGS (RoBERTa), NGS with RoBERTa as encoder; 3) NGS (RoBERTa + LSTM) with an encoder consisting of RoBERTa and a unidirectional LSTM; 4) DPE-NGS, our improved NGS model, with a Dual Parallel Text Encoder consisting of RoBERTa and a Bi-LSTM.

As shown in Table 9, we can see that using only RoBERTa as a text encoder can not improve the performance of the model when train with GeoQA-train, but the accuracy improves considerably when train with our dataset(Mix-train:64.77%, Backtrans-train:65.03%). We believe that the geometry problem text description is far different from the common linguistic description because it contains more geometric expressions, so the RoBERTa module should be fine-tuned with a larger geometric dataset, which also reflects that our new dataset is helpful to apply the Pre-training model to geometric problems solving. In addition, we also see that the model based on RoBERTa and unidirectional LSTM is much less effective than our DPE-NGS which demonstrates the effectiveness of our DPE module structure.

As shown in the figure, to intercept the square DEFG on a piece of paper △ABC. Where G, F in the BC side, D, E, respectively, in the AB, AC side, AH ⊥ BC intersection DE in M, if BC = 12(N0), AH = 8(N1), then the side length of the square DEFG is ().
A. 4.8 B. 4.0 C. 3.4 D. 5.0
Answer: A. 4.8



Annotated program: Mul, N0, N1, Add, N0, N1, Divide, V0, V1

Execute processing: 12(N0)*8(N1)=96(V0), 12+8=20(V1), 96÷20=4.8(V2).

DPE-NGS (No Result): Mul, N0, N1, Divide, V0, N0.

Figure 4: A typical case. A No Result type problem with a complex diagram.

	Trainsets	**GeoQA-train(%)**	**Mix-train(%)**	**Backtrans-train(%)**
Models	**NGS**	60.52	61.19	63.31
	NGS(RoBERTa+LSTM)	59.87	62.12	63.84
	NGS(RoBERTa)	58.28	64.77	65.03
	DPE-NGS	**62.65**	**65.96**	**66.09**

Table 9: Ablation study of different text encoder architecture designs. The content in parentheses indicates the encoder components that the model used.

5.4 Case Analysis

In our best experiment, there are still 10.73% problems for our model that can not get the answer. As shown in Figure 4, it's a typical problem in the No Result category. The diagram of this problem contains nine vertices that can form more than ten line segments and numerous geometric elements. We believe that the diagram is too complex for our Diagram Encoder to extract useful features from it. And this further leads to our inability to select useful diagram information for Joint and Reasoning work with text information, which ultimately affects the model's understanding of the whole problem scenario.

6 Conclusion

In this work, we newly annotate 2,518 geometric problems which are more difficult and with richer problem types to expand GeoQA and form a new benchmark dataset GeoQA+, the largest geometric problem dataset at present. Moreover, we propose a new text-encode method(DPE) to improve the limits of NGS. The experimental results show that both GeoQA+ and DPE-NGS have contributed to the accuracy improvement, and we have improved the baseline accuracy in automatic geometry problem solving from 60.7% to 66.09%. In the future, we will focus on the understanding of problem diagram by enhancing the ability of diagram features extraction as well as the representation of diagram information.

Acknowledgements

This paper is supported by the National Natural Science Foundation of China No. 62177015.

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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. 2021. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523, Online. Association for Computational Linguistics.

Shang-Ching Chou, Xiao-Shan Gao, and Jing-Zhong Zhang. 1996. Automated generation of readable proofs with geometric invariants. *Journal of Automated Reasoning*, 17(3):325–347.

J. B. Cohen and C. S. Areni. 1991. *Affect and Consumer Behavior.* Handbook of Consumer Behavior.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. 2020. A survey on text classification: From shallow to deep learning. *arXiv preprint arXiv:2008.00364*.

Zhenwen Liang, Jipeng Zhang, Jie Shao, and Xiangliang Zhang. 2021. Mwp-bert: A strong baseline for math word problems. *arXiv preprint arXiv:2107.13435*.

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 *Thirty-Fifth AAAI Conference on Artificial 2021*, pages 4232–4240.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6774–6786, Online. Association for Computational Linguistics.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Mrinmaya Sachan and Eric Xing. 2017. Learning to solve geometry problems from natural language demonstrations in textbooks. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 251–261.

Min Joon Seo, Hannaneh Hajishirzi, Ali Farhadi, and Oren Etzioni. 2014. Diagram understanding in geometry questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. 2015. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1466–1476, Lisbon, Portugal. Association for Computational Linguistics.

Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. 2021. BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–177, Online. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Wu Wen-Tsun. 1986. Basic principles of mechanical theorem proving in elementary geometries. *Journal of automated Reasoning*, 2(3):221–252.

Wu Wenjun. 1984. Basic principles of mechanical theorem proving in elementary geometries. *Journal of Systems Science and Mathematical Sciences*, 4(3):207.

Wing-Kwong Wong, Sheng-Cheng Hsu, Shih-Hung Wu, Cheng-Wei Lee, and Wen-Lian Hsu. 2007. Lim-g: Learner-initiating instruction model based on cognitive knowledge for geometry word problem comprehension. *Computers & Education*, 48(4):582–601.

Qinzhuo Wu, Qi Zhang, Jinlan Fu, and Xuan-Jing Huang. 2020. A knowledge-aware sequence-to-tree network for math word problem solving. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7137–7146.

Zhipeng Xie and Shichao Sun. 2019. A goal-driven tree-structured neural model for math word problems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5299–5305. International Joint Conferences on Artificial Intelligence Organization.

Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. 2018. Fast and accurate reading comprehension by combining self-attention and convolution. In *International Conference on Learning Representations*.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290.

Shaohua Zhang, Haoran Huang, Jicong Liu, and Hang Li. 2020. Spelling error correction with soft-masked BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 882–890, Online. Association for Computational Linguistics.