SCIMAT: Science and Mathematics Dataset*

Neeraj Kollepara
CSTAR, IIIT, Hyderabad
neeraj.kollepara@students.iiit.ac.in

Snehith Kumar Chatakonda
CSTAR, IIIT, Hyderabad
snehith.kumar@students.iiit.ac.in

Pawan Kumar
CSTAR, IIIT, Hyderabad
pawan.kumar@iiit.ac.in

Abstract

In this work, we announce a comprehensive well curated and opensource dataset with millions of samples for pre-college and college level problems in mathematics and science. A preliminary set of results using transformer architecture with character to character encoding is shown. The dataset identifies some challenging problems, and invites research on better architecture search for these problems.

1 Introduction and Previous Work

Datasets play an important role in driving research in supervised machine learning research. Some prominant examples being MNIST [13] for hand written digit classification, CIFAR10 [10] and IMAGENET [11] for image classification and generative models, etc. For solving math word problems, semantic rules and various models have been proposed in NLP community since 1963 starting from [1], [2], [6]. Some word problems in [7] are of the form: `Lucy has two dimes. Sarah has six dimes. How many dimes do they have altogether?` or `Dan has six books. Jill has two books. How many books does Dan have more than Jill?` This paper uses Kintsch and Greeno's (1985) theory of comprehension and solution for arithmetic word problems above. These papers used classical approaches with semantic rules. Recently, machine learning models have been used for which large labelled dataset is essential. Hence, there is a dire need of large question-answer dataset for mathematics and science problems; such dataset can have impact on online education, intelligent tutoring and automated grading. For intelligent tutoring, not just the answers, but the step by step hint can be provided; this is explored in [8]. However, tutoring requires some knowledge graph representation. Although, this was shown for simple algebraic and geometric mathematics problems, it remains a challenging task for more advanced problems. No wonder tutoring is a complex task as nicely pointed out in detail in [9]. Given that intelligent tutoring is one of the most challenging task, the datasets and innovative architectures would play a critical role to succeed in this endevour. Recently, question answer dataset[2] for mathematics was proposed in [14] and for word problem sample dataset was proposed in [3], and a comparison of results for character to character encoding for transformer and for LSTM is shown. This dataset has selected problems in mathematics for math exams for British 16 year old school children. Some sample questions are: `Factorise` $x^2 + 7x$ or `Three letters picked without replacement from qqqkkklkqkkk. Give prob of sequence qql.` In [5], a set of 7787 multiple choice questions in high school science questions is proposed as ARC (AI2 Reasoning challenge, 2018). A sample question from this dataset is: `Which property of a mineral can be determined just by looking at it? (A) luster [correct] (B) mass (C) weight (D) hardness.` Moreover, with the ARC challenge a large corpus of 14 million science sentences relevant to the question-answer set is also proposed. A sample sentence from the corpus is: `Random motion of`

[2] `https://github.com/deepmind/mathematics_dataset`

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

the air molecules and turbulence provide upward forces that may counteract the downward force of gravity. Such a corpus allows language understanding and questions with linguistic variations. We remark that any other corpus can be used for training the given architecture for linguistic understandings, which is further trained on the given datasets. For the ARC challenge, several baseline neural models were proposed. There are datasets for logical reasoning and English comprehension. For example, in [18], logical reasoning question answer dataset is proposed. The reasoning is considered to be of various types such as problems involving single supporting fact, two supporting fact, counting, path finding, size reasoning, etc. A sample question for path finding is: The kitchen is north of the hallway. John is hungry. The bathroom is west of the bedroom. John goes to the kitchen. The den is east of the hallway. John grabbed the apple there. The office is south of the bedroom. Daniel is hungry. How do you go from den to kitchen? How do you go from office to bathroom?. The last two sentences are questions with answers west, north and north, west respectively. This dataset is part of bAbI project[3] of facebook research. For algebra word problems, a dataset[4] and code is proposed in [12]. Most of these word problems correspond to solving system of linear equations, their method derives these equations, then solves it. A sample question answer in this dataset taken from [12] is: An amusement park sells 2 kinds of tickets. Tickets for children cost $1.50. Adult tickets cost $4. On a certain day, 278 people entered the park. On that same day the admission fees collected totaled $792. How many children were admitted on that day? How many adults were admitted? with soutions x = 128, y = 150. Continuing along these lines in [16], they propose to translate the math word problem to equation using recurrent neural network (RNN) without doing any complex feature extractions. To the best of our knowledge, a comprehensive **opensource** dataset for mathematics and science for pre-college and college level have been missing. To this end, in the following, we announce a new large dataset named SCIMAT, and we show preliminary results and comparisons [17].

2 SCIMAT: Large Science and Mathematics dataset

We announce a large dataset[5] of hundreds of millions of question-answer for mathematics and science for pre-college and college level, which typically is taught to 15-19 age group around the world. The list of topics covered in science are: Acids And Bases, Atomic Structure, Stoichiometry, Thermodynamics, Units And Dimensions, Kinematics, Laws of Motion, Work Power Energy, Rotatory Motion, Gravitation, Electricity, Moving Charges and Magnetism, Electro Magnetic Induction, Alternating Current, Electro Magnetic Waves, Ray Optics and Optical Instruments, Wave Optics, Dual Nature of Matter, Mechanical Properties of Solids and Liquids, Thermal Properties of Matter, Kinetic theory of Gases, Sound, Waves And Oscillations, SemiConductors, Communication Systems, etc. Each topic contains several subtopics, where each subtopics has hundreds of thousands of question answer dataset.

In any of the chapter, first we identify some of the important problem types, for example in Work Power Energy (WPE) chapter, we have identified problem types such as notion of work, KE, conservation of energy, etc. Then for each of the problem types, we find some questions from basic class 9 to class 12 problems (based on chapter), modify them slightly if needed and create some variants of that problem type.

2.1 Sample Questions in Science

1. **Question:** 33 mL of a solution of HNO3 is found to be completely neutralised by 45 mL of a given solution of NaOH. If we take 12 mL of the same solution of HNO3, the amount of NaOH solution (the same solution as before) required to neutralise it will be. **Answer:** 16.36 ml
2. **Question:** If a diatomic gas of 1 moles at 68 atm and volume 68 lit is adiabatically changed to volume 188 lit, then what will be the pressure. **Answer :** 16.4atm
3. **Question:** A body is dropped from a height of 9578 m with an initial velocity of 42 m/s. With what velocity will it strike the ground ? **Answer:** 435.3 m/s

[3] https://github.com/facebookarchive/bAbI-tasks

[4] http://groups.csail.mit.edu/rbg/code/wordprobs/

[5] https://github.com/misterpawan/scimat2

4. **Question:** A 9062 N force is applied on a body of mass 980 kg placed on a smooth surface, then what is the resulting acceleration obtained ? **Answer:** 9.2 m/s2
5. **Question:** The volume of 549 g of a substance is 116 cm3. If the density of liquid in which substance is placed is 4 g/cm3, will the substance float or sink ? **Answer:** sink
6. **Question:** If a 822 V battery is connected across an unknown resistor, there is 224 A in the circuit, find the value of resistance of the resistor ? **Answer:** 3.7 ohm
7. **Question:** A square coil of side 3 cm consists of 31 turns and carries a current of 5 A. The coil is suspended vertically and the normal to the plane of the coil makes an angle of 53 degress with the direction of a uniform horizontal magnetic field of magnitude 17 tesla. What is the magnitude of the torque experienced by the coil. **Answer:** 1.9 newton-m
8. **Question:** A series LCR circuit is connected to a variable frequency 230 V source with L = 193 H, C = 72 muF, R = 176 ohm. Determine the rms potential drop across resistance? **Answer:** 230 volt
9. **Question:** Suppose that the electric field amplitude of an electromagnetic wave is E0 = 1936 N/C and that its frequency is v = 1512 MHz. Find an expression for B? **Answer:** 6.45e-06sin[3.17e+01x-9.50e+09t]
10. **Question:** During blood transfusion, the needle is inserted in a vein where the gauge pressure is 1720 Pa. If the blood container is placed at 177 mm above the earth level so that blood may just enter the vein, is it safe for the patient?. **Answer:** yes, patient is safe
11. **Question:** A sound wave travels at a speed of 29980.8 m/s, if it's wavelength is 32 m, will the sound wave be audible ? **Answer:** audible
12. **Question:** For an amplitude modulated wave, the maximum amplitude is found to be 18.62 V while the minimum amplitude is found to be 7.91 V. Determine the modulation index. **Answer:** 0.4

Similarly, for mathematics, we append datasets from calculus (differentiation and integration), linear algebra (rank, row reduced echelon form, determinant, trace, etc), set operations, statistics, number theory, probability, etc. Some sample questions in the dataset are as following:

2.2 Sample Question in Mathematics

1. **Question:** Differentiate 293 * x * (sin(x) + sec(x)) with respect to x
   Answer: 293 * x * (cos(x) + tan(x) * sec(x)) + 293 * sin(x) + 293 * sec(x)
2. **Question:** Integrate cot(4*x ^2) + sec(22*x^2) with respect to x
   Answer: 8*x*(-cot(4 * x^2) ^2 - 1) + 44*x* tan(22*x^2) * sec(22*x^2)
3. **Question:** Calculate the Rank of Matrix [[2, 1, 3, 7] , [1, 0, 4, 2] , [3, 1, 7, 9]] **Answer:** 2
4. **Question:** Calculate the Trace of Matrix [[13, 38, 61] , [29, 1, 39] , [92, 16, 45]] **Answer:** 59
5. **Question:** What is the union of { 2, 6, 7, 8, 9 } with { 3, 7, 8 } **Answer:** { 2, 3, 6, 7, 8, 9 }
6. **Question:** What is the median of the sequence (20, 38, 4, 21, 31, 94, 55) **Answer:** 31
7. **Question:** What is 2 (base 3) in base 8? **Answer:** 2
8. **Question:** Three letters picked without replacement from a: 3, c: 1, b: 7, d: 3. Give prob of sequence bdc. **Answer:** 1/104

3 Numerical Experiments

The code for training and testing is written in Python and `PyTorch` framework is used. The models are trained on dual `Intel Xeon E5-2640 v4` processors, providing 40 virtual cores per node, 128 GB of 2400MT/s DDR4 ECC RAM and four `Nvidia GeForce GTX 1080` TiGPUs, providing 14336 CUDA cores. We use the standard transformer described in [15] with our own specifications as follows. We use an encoder which is composed of stack of $N = 4$ identical layers. The embedding size (dmodel) = 128, attention heads (h) = 8. The inner layer size of feed forward network used in each layers of encoder stack (dff) = 512. We minimize the sum of log probabilities of the correct tokens via the Adam optimizer with adaptive learning rate. The model was trained for 100 epochs. For floating point answers, accuracy for two digits after decimal place was matched. In Table 1, 2, we find that there are datasets where it is challenging to obtain high accuracy, and robust architecture or encoding is required. Since lately, many other variants of transformers were proposed, in Table 3, we compare various different transformer with word-to-word and char-to-char encoding. In general, we found that char2char gives best accuracy.

Type of problem	C2C Accuracy
Differentiation of sum	99%
Differentiation of product	100%
Differentiation of composition	100%
Integration of sum	100%
Integration of product	100%
Integration of composition	92.5%
Addition of matrices	49%
Subtraction of matrices	74%
Transpose of matrix	100%
Determinant of matrix	32%
Multiplication of matrices	32%
Trace of a matrix	100%
Product of matrix with a Scalar	100%
Row Reduced echelon	76%
Rank of a matrix	92.5%
Mean of a sequence	95%
Variance of a sequence	39.5%
Median of a sequence	99%
Set Union	100%
Set intersection	97.5%
Set difference	100%
Symmetric difference between sets	100%

Table 1: Comparison of our model trained on new datasets with Char2Char transformer. The Char2Char is denoted by C2C.

Type of problem	C2C Accuracy
Neutralization	82.6%
Adiabatic	76.3%
Refrigrator	82.6%
Estimated value	61.8%
Force, mass, acceleration	45.5%
Momentun conservation	78.7%
Kinetic energy	73.5%
Balancing a metre stick	81.5%
Gravitational field	94.2%
Float or sink?	98%
Ohms Law	89.0%
Torque due to magnetic field	84.5%
LCR circuit	91.3%
Mirror formula for concave	79.6%
Is the sound audible?	76%
Sound wave propogation	31.5%
modulation index	89.6%
Force between wires	75.2%
Conservation of momentum	14.5%
Potential energy	63%
Work, mass, velocity	10%

Table 2: Accuracy for science datasets with Char2Char transformer. The Char2Char is denoted by C2C. See dataset with folder names.

Type of problem	C2C Trans.	W2W Trans.	W2W Perfor.
pH	**99.8%**	97.3%	84.7%
Compare number of atoms	**97.5%**	94.1%	94.4%
Operations with significant digits	**80.4%**	72.9%	67.4%
Equation of motion	8.5%	**12.8%**	12.2%
Kinetic energy	**73.5%**	72.4%	71.8%
Float or sink?	98%	98.7%	**98.8%**
Series/Parallel combination of resistance	**88.9%**	32.5%	45.7%

Table 3: Compare various transformers. Here C2C is char-to-char encoding, W2W is word-to-word encoding, Perfor. stands for performer [4], and Trans. stands for Transformer [15].

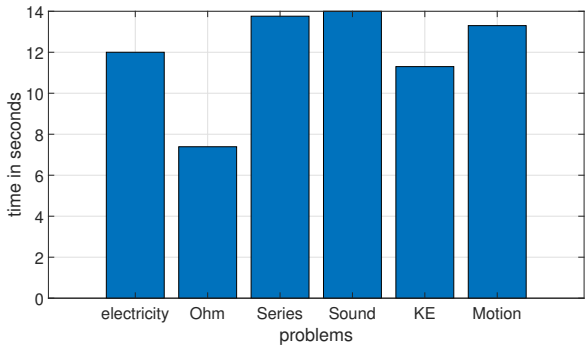

Figure 1: Time for generating some datasets. The generator codes are provided in the data repository.

Acknowledgments and Disclosure of Funding

This work was done at IIIT, Hyderabad. The authors acknowledge all the support of the institute.

References

[1] Daniel G. Bobrow. Natural language input for a computer problem solving system, 1964.

[2] Diane J. Briars and Jill H. Larkin. An integrated model of skill in solving elementary word problems. *Cognition and Instruction*, 1(3):245–296, 1984.

[3] Sizhu Cheng and Nicolas Chung. Simple mathematical word problems solving with deep learning. Technical report, Stanford University, 01 2010.

[4] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers, 2021.

[5] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018.

[6] E.A. Feigenbaum and J. Feldman. *Computers and Thought*. McGraw-Hill, 1963.

[7] Charles R. Fletcher. Understanding and solving arithmetic word problems: A computer simulation. *Behavior Research Methods*, 17(5):565–571, September 1985. Copyright: Copyright 2011 Elsevier B.V., All rights reserved.

[8] Bo Kang, Arun Kulshreshth, and Joseph J. LaViola. Analyticalink: An interactive learning environment for math word problem solving. In *ACM*, IUI '16, page 419–430, New York, NY, USA, 2016. ACM.

[9] Kenneth R. Koedinger, Julie L. Booth, and David Klahr. Instructional complexity and the science to constrain it. *Science*, 342(6161):935–937, 2013.

[10] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research), 2009.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

[12] Nate Kushman, Yoav Artzi, Luke Zettlemoyer, and Regina Barzilay. Learning to automatically solve algebra word problems. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 271–281, Baltimore, Maryland, June 2014. ACM.

[13] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/, 2010.

[14] David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. Analysing mathematical reasoning abilities of neural models. *ArXiv*, abs/1904.01557, 2019.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[16] Yan Wang, Xiaojiang Liu, and Shuming Shi. Deep neural solver for math word problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 845–854, Copenhagen, Denmark, September 2017. ACL.

[17] Snehith Kumar Chatakonda, Neeraj Kollepara, and Pawan Kumar. SCIMAT: Dataset of Problems in Science and Mathematics. In *BDA 2021*, submitted.

[18] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks, 2015.