Machine Number Sense: A Dataset of Visual Arithmetic Problems for Abstract and Relational Reasoning

Wenhe Zhang,[1,3] **Chi Zhang,**[1,2] **Yixin Zhu,**[1,2] **Song-Chun Zhu**[1,2]
[1]UCLA Center for Vision, Cognition, Learning, and Autonomy
[2]International Center for AI and Robot Autonomy (CARA)
[3]Peking University
wenhe@pku.edu.cn, chi.zhang@ucla.edu, yixin.zhu@ucla.edu, sczhu@stat.ucla.edu

Abstract

As a comprehensive indicator of mathematical thinking and intelligence, the *number sense* (Dehaene 2011) bridges the induction of symbolic concepts and the competence of problem-solving. To endow such a crucial cognitive ability to machine intelligence, we propose a dataset, Machine Number Sense (MNS), consisting of *visual* arithmetic problems automatically generated using a grammar model—And-Or Graph (AOG). These visual arithmetic problems are in the form of geometric figures: each problem has a set of geometric shapes as its context and embedded number symbols. Solving such problems is not trivial; the machine not only has to recognize the number, but also to interpret the number with its contexts, shapes, and relations (*e.g.*, symmetry) together with proper operations. We benchmark the MNS dataset using four predominant neural network models as baselines in this visual reasoning task. Comprehensive experiments show that current neural-network-based models still struggle to understand number concepts and relational operations. We show that a simple brute-force search algorithm could work out some of the problems without context information. Crucially, taking geometric context into account by an additional perception module would provide a sharp performance gain with fewer search steps. Altogether, we call for attention in fusing the classic search-based algorithms with modern neural networks to discover the essential number concepts in future research.

1 Introduction

> Number is the ruler of forms and ideas, and the cause of gods and demons.
>
> — Pythagoras, c. 300 (Taylor 1818)

Mathematics is arguably the most elegant and vivid reflection of human intelligence, covering the areas of geometry, arithmetic, algebra, and analysis (Simpson and Weiner 1989). It is the science of logic reasoning, the discipline of abstract forms, and the realm of symbolic languages. Among all the mathematical symbols, numbers are the most familiar and vital elements to us. Although the opinions of Pythagoras that "all is number" are controversial and extreme, the significance of the numbers can never be overestimated: people from all walks of life embrace numbers every day.

Dealing with numbers seems to be a simple task and an innate competence: even newborn infants can discriminate basic numerosities, expressing their surprise when the number of stimuli changes from two to three (Starkey and Cooper 1980). Meanwhile, processing numbers is also a painstaking challenge and a learned skill; it always takes years of efforts for students to practice calculations and more complicated computations. Such a numerical competence is, in fact, very unique; only few other animals possess similar capabilities (and at a much smaller scale compared to human) (Davis and Pérusse 1988; Gallistel 1989; 1990; Brannon and Terrace 1998; Dehaene, Dehaene-Lambertz, and Cohen 1998; Cantlon and Brannon 2007; Jacob and Nieder 2008; Nieder and Dehaene 2009). What is the underlying mechanism of human numerical thinking and the concepts of number? And how to endow a similar capability to machine intelligence?

The *number sense* (Dehaene 2011), a psychological terminology, provides an explanation about the cognitive process of numbers for both human and animals. It refers to the understanding of number concepts, the competence of numerical operations (including counting, comparison, estimation, and calculation), and the ability to flexibly solve mathematical problems (Bobis 1996). People characteristic of good number sense usually possess the abilities of fluent magnitude perception, reasonable result expectation, flexible mental computation, and appropriate presentation formulation (Kalchman, Moss, and Case 2001). Below, we summarize four key observations from the vast body of literature on number sense.

Learned *vs.* Innate Number sense is developed in *acquired* environments in addition to our *innate* capability. Five-month-old infants have already possessed the capacity to represent cardinality and can engage in rudimentary arithmetics—basic addition and subtraction operations on small sets of objects (Wynn 1992). Older children gradually *learn* to establish the abstract connections between the magnitude of the quantities and the symbolic expression of the numbers, which are the foundation of further comparisons and calculations (Temple and Posner 1998). Symbolic numerical processing skills, different from the processing abilities of countable non-symbolic objects, are more closely related to mathematical competence (Schneider et al. 2017). As for

average adults, retrieving the abstract meaning of number symbols has been developed into a highly automated process, thus facilitating more rigorous computations (Dehaene 2011).

Vision *vs.* Language Number sense is in closer relation to vision than language due to a few reasons. First, the definition of number sense emphasizes the estimation of the magnitude of the quantities and the understanding of number symbols based on *visual* input. Second, empirical evidence has suggested the significant relation between vision and number sense. Studies of developmental psychology indicated that people first developed their number sense from vision; babies with limited knowledge of language have expressed the ability to discriminate the numerosity of *visual* objects, and children gradually learn the quantitative meaning of *visual* symbols (Dehaene 2011). Third, in evolutionary psychology, animals, in general, are unable to generate a verbal representation of numbers, but some of them still exhibit the number sense, capable of numerical discrimination and mental operations of *visual* items (Gallistel 2003).

Context and Adaptation Number sense is not only the awareness and manipulation of abstract symbols but also the capacity of conducting flexible mathematical operations in concrete situations. People with good number sense usually display an excellent problem-solving ability (Cobb et al. 1991). To solve mathematical problems effectively, we need to observe the *context* in which the problem is presented, form an *adaptive* representation for problem settings and a proper expectation for possible results, select the most suitable strategy that contains necessary sub-operations, and work with the numbers step by step (Heinze, Star, and Verschaffel 2009).

Quantity *vs.* Rank Two types of neuronal mechanisms were extensively studied in the neuroscience literature (Wiese 2003): (i) *Numerical quantity* refers to the property of cardinality of sets of objects or events (also called numerosity)—"how many?". (ii) *Numerical rank* refers to the property of serial order and pertains to the question—"which position?".

1.1 Overview

In this paper, we hope to use the concept of number sense, an ideal indicator, to evaluate the machine intelligence from the perspective of mathematics; it naturally combines both crystallized intelligence (knowledge and experience of number processing) and fluid intelligence (adaptive problem-solving in a given situation), which comprises the basic structure of human intelligence (Cattell 1963).

Specifically, we propose a new dataset, Machine Number Sense (MNS), in the form of geometric figures. It consists of various types of arithmetic problems, in which integers appear as problem contents and geometric shapes serve as problem contexts; see an example in Figure 1. The task for evaluating machine's number sense is: given training samples as *images*, the algorithms should figure out the underlying latent relations between the numbers in each image panel and fill in the missing number as the answer in the last panel.

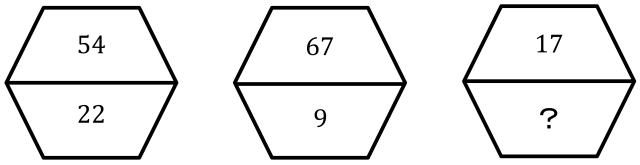


Figure 1: A sample problem in the Machine Number Sense (MNS) dataset using the rule of addition: $54 + 22 = 76$, $67 + 9 = 76$, $17 + ? = 76$. The correct answer is 59.

Here, we are interested in testing a few intriguing questions that correspond to the above four key observations: (i) Given only the visual input, are modern machine learning methods capable of learning and understanding the quantitative meaning of number symbols and the relations between these symbols? (ii) If the spaces of the operations and rules are known, is it possible to work out the problem by symbolic search? What would be the difference between the two streams of methods? (iii) How much does the contextual information contribute to numerical problem-solving? (iv) Could learning-based methods realize the numerical quantity and numerical rank merely from the visual input?

Our experiments show that the predominant neural network models still have a significant cognitive gap between visual symbols and abstract meanings even after extensive training; there must be a missing association between context information and problem-solving skill. In contrast, only taking number symbols as the input, the classic search algorithm manages to solve some problems correctly, but the search is very inefficient. Adding an additional perception module to provide geometric contextual information significantly improves the performance of the algorithm.

1.2 Contributions

This paper makes two major contributions:

- We introduce a new Machine Number Sense (MNS) dataset, composed of various visual arithmetic problems.
- We benchmark the ability of the modern machine learning methods with respect to the quantitative understanding of number symbols, the relational operations between numbers, and the ability of adaptive problem-solving, which together construct a full framework of number sense.

Compared to other mathematical problems in the form of text or language in prior work, the problems presented here are unique in the following aspects:

Token *vs.* Pixel Instead of using tokenized symbols extracted from the texts or languages, testing machine number sense directly from pixel input is much more challenging. By means of language, the quantitative meaning of number symbols and the relations among them could be easily discovered with abundant semantic clues embedded in the sentence. In contrast, it is much harder to establish the connections among different numbers with their visual contexts; the algorithm has to reason and induce from the observed visual pattern using limited examples.

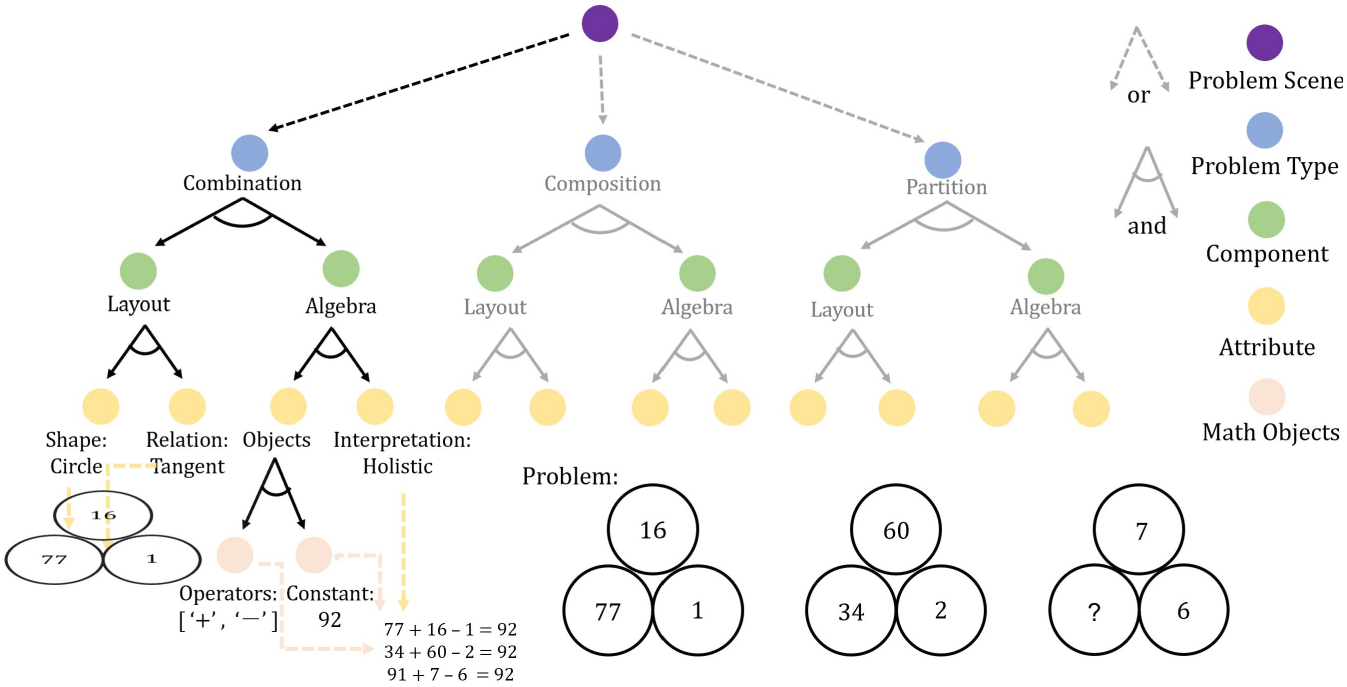


Figure 2: The Machine Number Sense (MNS) dataset creation process. Given grammar production rules together with its attributes, we can generate a test by parsing and sampling an And-Or Graph (AOG).

Sequential *vs.* Hierarchical Using visual inputs also brings up more rigorous requirements for formulating suitable yet flexible representations. The visual pattern is usually hierarchically organized and generated, demanding an algorithm to parse a test into a similar hierarchical representation. This unique property is fundamentally different from the sequential and temporal nature in prior representations in the context of texts or languages. A proper perceptual-grouping (*e.g.*, Gestalt laws (Wertheimer 1923)) for visual elements is necessary. Additionally, we would need a flexible representation for representing a problem based on its context and reconstructing the representation when it is not appropriate; such an adaptation is regarded as a key step for problem-solving (Knoblich et al. 1999).

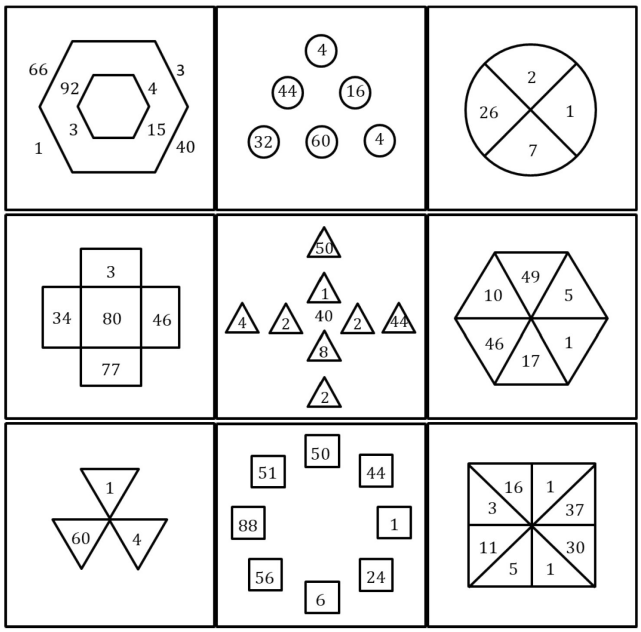
Recognition *vs.* Reasoning The proposed dataset is characterized by both its simplicity and difficulty. In each problem, there are only numbers and geometric shapes, unlike others with various mathematical symbols (Ling et al. 2017; Saxton et al. 2019). However, simple appearance does not indicate trivial problem-solving; in contrast, it enforces the algorithm to reason about the latent structure, relations, and operations within a problem consisting of very “limited” visual information, making the problem-solving process challenging. This nature of the present dataset leads to the focus on reasoning and understanding, rather than the traditional tasks (*e.g.*, recognition) in the field of computer vision.

Human *vs.* Machine There are qualitative differences between the present dataset and previous tests of human number sense. The human tests examine number sense from a clinical perspective, aiming at discriminating children with potential mathematical disabilities, so that the problems in the tests are relatively easy, basic, and eliminative, serving as diagnostic tools. In contrast, our task investigates number sense from a cognitive perspective, measuring machine intelligence from the aspect of number processing; the problems are more comprehensive, flexible, and cognitive-demanding.

2 The Machine Number Sense Dataset

Representation We use And-Or Graph (AOG) as the representation; see an illustration of the structure for the generation process in Figure 2. AOG is a context-free grammar frequently used for hierarchical and compositional data in AI and computer vision (Zhu and Mumford 2007). In MNS dataset, each problem has an internal hierarchical tree structure composed of And-nodes and Or-nodes; an And-node denotes a decomposition of a larger entity in the grammar, and an Or-node denotes an alternative decomposition.

In our design, the root node of the AOG is an Or-node, representing a single test. After the decomposition on the sub-level are three different problem types represented by And-nodes. After selecting the problem type by choosing one of the And-nodes, the problem is divided into layout and algebra components. Sampling the terminal nodes in each component will complete the process of the problem generation. The three image panels within a single problem share common layout and algebraic properties; the only difference among them is the actual integers that appear on the panel.



(a) (b) (c)

Figure 3: Layouts of three different problem types: (a) combination, (b) composition, and (c) partition.

Problem Types We design three types of problems: combination, composition, and partition, each of which has a distinctive layout. Figure 3 shows a few examples using different layouts. In a combination problem, two or three geometric shapes are combined together by a specific spatial relation. In a composition problem, a set of small geometric shapes are composited to outline a larger shape. In a partition problem, one geometric shape is divided into several parts by lines.

Layout Component and Attributes The layout component serves as the problem *context*, consisting of two different geometric attributes, both of which are necessary for the problem generation; see an illustration in Figure 3. The first attribute refers to geometric shape: triangle, square, circle, hexagon, or rectangle. The second attribute varies in different problem types. In combination problems, it indicates the spatial relation by which the geometric shapes group together; in our dataset, two figures could be combined by the relation of overlapping or including, and three figures could be grouped together by tangent relation. In composition problems, the second attribute refers to the format of spatial arrangement of geometric shapes, which can be composed in the forms of line, cross, triangle, square, and circle. In partition problems, this attribute represents the number of parts the geometric shapes are partitioned into.

Algebra Component and Attributes The algebra component serves as the problem *content*; similarly, it is composed of two mathematical attributes. The first attribute indicates the mathematical objects in the problem, including a list of operators and integer constants. The constants range from

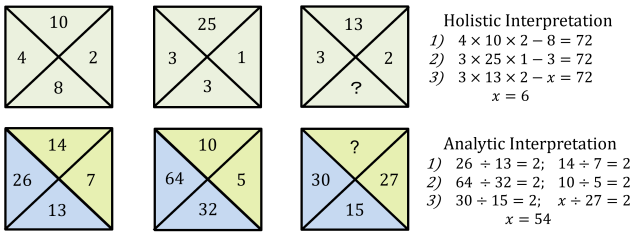


Figure 4: Examples of different algebra components: holistic and analytic interpretation.

1 to 99 and the values of operators are the four elementary operators in arithmetics: "+", "−", "×", "÷".

The second attribute is the styles of interpretation—holistic view and analytic view, which correspond to two basic thinking styles of human cognition (Nisbett et al. 2001); see Figure 4. Holistic cognition emphasizes attending to the entire information input, while analytic cognition focuses on grouping the input into different sub-parts. From a holistic perspective, all the numbers in a panel are involved in the same calculation process together as a whole. From an analytic perspective, the numbers are grouped as several parts, and each part undergoes an individual calculation process. If the interpretation style is analytic, the numbers in a panel can be divided into 2, 3, or 4 parts. The form of grouping is designed on the basis of human perceptual organization laws, especially the law of similarity and the law of proximity (Wertheimer 1923): numbers at neighboring or symmetrical positions tend to be organized as a group.

Sampling Math Objects Once the layout and the interpretation style are determined, the final step is to sample operators and constants to automatically generate a test.

The space of possible operators given the current problem type, component, and attribute is constrained by the problem context. For holistic problems, this space is subject to the number of available integer positions in each *panel*. For analytic problems, the space is subject to the number of available integer positions in each *group*. Parentheses are further randomly inserted to change the operator precedence; this modification dramatically increases the problem space.

The values of integer constants also need to be adjusted to maintain a consistent difficulty among the generated tests. If the center position of the layout has the space to display numbers, we show the integer constant at the center of each panel as a hint for problem-solving. In other situations where the center position is occupied by lines or other shapes, the algorithm needs to reason about what the constant is and how to calculate such a constant. To make a trade-off of difficulty, the values of constants differ in each panel in the former situation, whereas the underlying constants remain the same among different panels in the latter situation.

Instantiation by Calculation Tree The sampled operators and constant slots are fed into an in-order binary tree to sample numbers for instantiation; see Figure 5 for a detailed

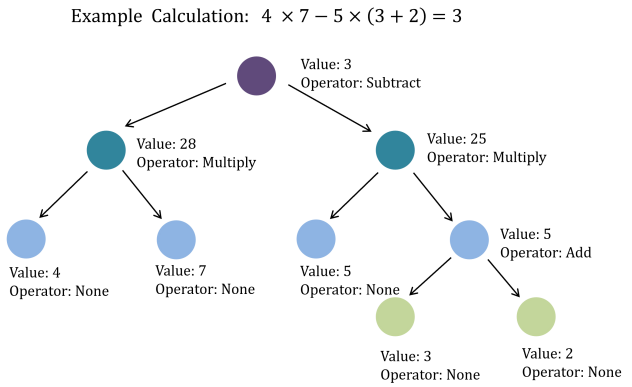


Figure 5: An example of the calculation tree for number generation. The root node is the sampled integer constant.

example of the generation process. We call this binary tree a "calculation tree"; the nodes in it have two properties—numeric value and operator. The value of root node is assigned as the already sampled integer constant, while the values of other nodes need to be sampled. The sampling process follows two constraints: (i) the operation between the left-child value and the right-child value under the parent operator will yield the parent value, and (ii) the value is an integer from 1 to 99. The sampling process terminates when all the leaf nodes have qualified values. If a sampling process cannot generate a problem that satisfies all the constraints, it will be terminated and the entire process will be restarted.

3 Experiment Settings

We benchmark the proposed MNS dataset using both predominant neural network models and classic search-based algorithms. Additionally, human performance on the dataset has also been collected.

3.1 Neural Network Models

We implement state-of-the-art neural-network-based computer vision models for visual problem-solving (Zhang et al. 2019a; Barrett et al. 2018) and examine their competence on the dataset. Specifically, we compare 4 different baselines: (i) a front-end CNN as feature extractor (CNN), (ii) a popular sequential learning model with a CNN backbone combined with an MLP head (LSTM), (iii) an image classifier based on ResNet (He et al. 2016), and (iv) a relational network (RN) (Santoro et al. 2017). We treat the problem as a classification problem and train all models using the cross-entropy loss. All models are optimized using ADAM (Kingma and Ba 2014) and implemented by PyTorch (Paszke et al. 2017); see performance in Table 1 and Figure 6.

CNN In this model, we treat the three panels as a whole and stack them along the channel dimension. Features of the stacked panels are extracted by a CNN model, from which a final answer is predicted.

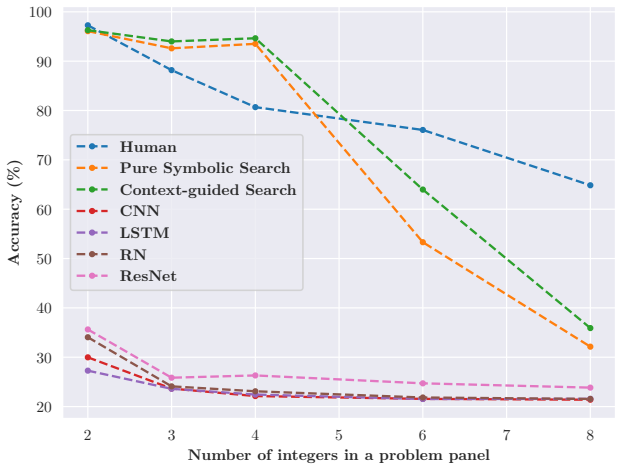


Figure 6: Accuracy w.r.t the number of integers in each panel.

LSTM The sequential nature of calculation and the analogical relations among different panels motivate us to choose the representative LSTM model for sequential learning. Similar to ConvLSTM (Xingjian et al. 2015), we feed each panel independently through a small CNN feature extractor and connect them to the input layer of an LSTM network. Image features are iteratively updated in three steps and finally passed to a multi-layer perceptron for prediction.

ResNet Due to the superior performance in image classification, we also choose to benchmark the dataset using ResNet. The feature extractor used in CNN is now replaced with a ResNet-18 backbone (He et al. 2016). We use the publicly available implementation and train the model from random initialization.

RN Relational network has demonstrated good performance on tasks demanding relational reasoning (Santoro et al. 2017; Barrett et al. 2018). Hence, it is natural to examine whether such a relational structure could be beneficial for number sense. We adopt the relational model and feed image features extracted by a CNN. A multi-layer perceptron is used to predict answers based on the relational representation.

3.2 Symbolic Search-Based Models

We further examine whether the tests can be solved by searching through the problem space. The problem space is fairly large, spanned by various operators, constants, interpretations, shapes, and relations, posing challenges for symbolic search-based models.

We implemented two types of the symbolic search-based models: (i) pure symbolic search, wherein the input is the numbers in each panel, and (ii) context-guided search, taking both the numbers and semantic context information as input. Both the pure symbolic search and context-guided search share similar problem-solving mechanisms: search through the entire problem space until the problem is solved.

Method	Mean	Combination		Composition		Partition	
		Holistic	Analytic	Holistic	Analytic	Holistic	Analytic
Pure Symbolic Search	52.15%	62.98%	56.83%	22.17%	53.73%	51.29%	71.60%
Context-guided Search	56.70%	64.38%	56.08%	29.81%	61.84%	59.70%	67.59%
CNN	22.71%	25.25%	19.65%	22.53%	20.07%	24.44%	23.25%
LSTM	22.16%	24.57%	21.10%	22.21%	20.12%	23.36%	23.83%
RN	22.96%	27.05%	20.47%	22.93%	20.27%	25.81%	23.64%
ResNet	25.29%	27.90%	24.22%	23.42%	23.73%	26.61%	27.78%
Human	77.58%	66.82%	93.64%	61.36%	78.18%	77.27%	88.18%

Table 1: Performance (accuracy) of different models on the machine number sense dataset.

Context-guided search only differs from pure symbolic search in two aspects: (i) additional context information may provide heuristics for solving the problem, and (ii) the relative spatial positions of numbers can be inferred from context information, enabling the model to find the correct order of numbers in calculation more quickly. The performance using these two models are shown in Table 1 and Figure 6.

4 Performance Analysis and Comparison

4.1 Analysis of Neural Network Models

Table 1 shows how models perform on the MNS dataset. As shown in Table 1, neural networks, unlike search algorithms, perform similarly on different interpretations across all problem types. This observation indicates that by purely learning from the paired image and answer, neural network models are not capable of acquiring the essential cognitive process of perception organization for analytic interpretation. Among all the tested models, ResNet achieves the best performance compared to other neural network models. One possible contribution to the better performance of ResNet may come from its considerable depth, which enables the model to extract more distinct features from the problem images (He et al. 2016), helping to discriminate a certain number symbol from others. Although discriminative features on symbols alone may be inadequate for a comprehensive symbolic understanding, it indicates that a strong classifier does help to improve the overall performance.

Figure 6 shows how model performance changes as the number of integers involved increases. One counter-intuitive observation for neural network models is that the accuracy of problem-solving does not significantly decrease as the number of integers increases. Although the accuracy is the highest in 2-integer situation for all models, the performance in cases with more integers remain similar. This observation suggests that neural network models share a common processing mechanism that is invariant to the number of integers, qualitatively different from search algorithms.

4.2 Analysis of Search-based Models

Figure 7 shows that the accuracy of search algorithms improves as the number of search steps increases, in accordance with the intuition that more trials during problem-solving will lead to a higher chance of success. We observe from Table 1 that the performance of search algorithms differs between the two styles of interpretations. In combination problems, the algorithms perform better in holistic interpretation. Conversely, in partition and composition problems, the algorithms perform better in analytic interpretation. This observation follows the design of problem layouts: as there are usually more integers in partition and composition problems, it is more expensive to conduct holistic calculations than grouping the integers into several parts for computation. We also note that four numbers could be a turning point for search-based algorithms as the performance drops significantly when there are more than four integers.

Although pure symbolic search is able to solve some problems, context-guided search has, in general, better performance, especially on problems with higher complexity, *e.g.*, 4-, 6- and 8-integer (see Figure 6). This difference shows the importance of context information in formulating a suitable organization and representation of problem, avoiding invalid trials of low-possibility circumstances, and finding solutions for complicated problems.

4.3 Compare Search *vs.* Neural Network

There are two major differences in performance of search algorithms and neural network models:

- The overall accuracy of neural network models is close to that of pure symbolic search within 100 steps and context-guided search within 50 steps, both of which are relatively small compared to the large problem space.
- The performance of search algorithms varies across different types of problem, different styles of interpretation, and different numbers of integers, in strong contrast to the performance consistency of neural network models.

The underlying reasons for the differences lies in three aspects. First, the representations of number symbols and geometric contexts differ. For search algorithms, the input number symbols are represented as abstract concepts, with clear quantitative meaning and known operational rules, which can be directly fit into each calculation process. Similarly, the context information is given as a high-level semantic concept. In contrast, for neural network models, the input number symbols and geometric contexts are in the form of pixels, so that the models represent the information as a set of extracted features rather than a set of symbolized concepts.

Second, search-based models treat number symbols as independent concepts and process them in a sequential manner, resulting in increased time complexity as the number of integers grows. In contrast, neural network models process

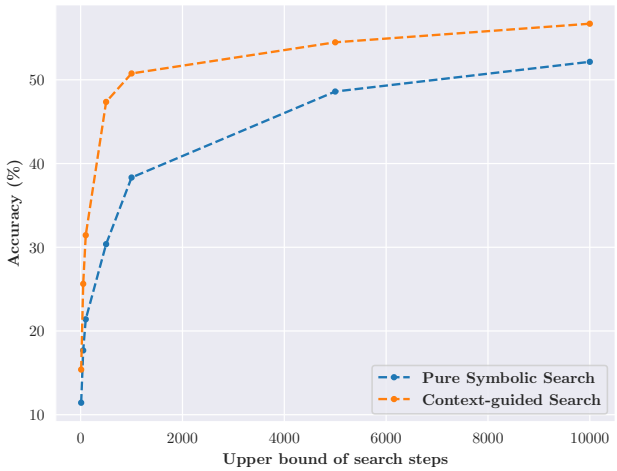


Figure 7: Search performance improves as the number of maximum searching steps increases.

visual features in parallel, so that the model performances are invariant to the number of integers.

Third, since the number symbols and geometric context information are fed into search algorithms separately, the ability of search algorithms to separate problem content from problem context is also advantageous than that of neural network models. We argue that being able to separate contents and contexts based on the pixel input is crucial to achieve high performance for neural network models: geometric figures will lose its meaning if deprived of problem contents, and the interpretation of problem contents will also be harder without problem contexts.

4.4 Compare Human *vs.* Machine Performance

Compared to computational models, human achieves a significantly higher accuracy in *all* types of problems without extensive training. In our experiments, participants have displayed a superb proficiency in comprehending number symbols and an advanced capacity to learn about the operational relations among numbers from just *two* problem panels.

Unlike neural network models and search algorithms, participants consistently perform better in analytic calculation than in holistic calculation. Similar to search algorithms, the accuracy of participants drops as the amount of number symbols in a problem panel increases. A surprising result is that participants only perform better than search algorithms when there are more than four integers in each problem panel; this counter-intuitive result may be due to the fact that the search-based algorithms can almost search the entire problem space within the step limit when the number of integers is small.

5 Related Work

Investigating machine number sense is an important direction that would shed light on many other research topics in the area of artificial intelligence and cognitive science, such as relational reasoning (Waltz et al. 1999; Santoro et al. 2017; Zhang et al. 2019a), visual analogy (Stafford 2001; Davies and Goel 2001; Hill et al. 2019; Zhang et al. 2019a; 2019b), and concept learning (Tenenbaum 1999; Fisher, Pazzani, and Langley 2014; Lake, Salakhutdinov, and Tenenbaum 2015). Below, we briefly review related work in number sense.

5.1 Educational Psychology

To examine the number sense in students' math learning, researchers in the field of educational psychology have been developing and standardizing the diagnostic measurement and intervention training of number sense. A series of tests has been devised to examine different aspects of number sense, which can be classified into two categories—conceptual understanding and procedural operation. For example, the Quantity Discrimination task measures the understanding of the quantitative meaning of number symbols (Chard et al. 2005), and the Number Knowledge Test assesses operational knowledge of numbers (*e.g.*, basic additions and subtractions) with a hierarchy of difficulty (Okamoto and Case 1996).

5.2 Artificial Intelligence

The cognitive ability of machine number sense has not been thoroughly investigated in the field of artificial intelligence. Although mathematical problem-solving has been a research topic with an increasing interest, the focus of previous research work is either on abstract language understanding (Kushman et al. 2014; Upadhyay and Chang 2016; Huang et al. 2016; Wang, Liu, and Shi 2017; Ling et al. 2017) or general mathematical problem-solving (Saxton et al. 2019), leaving out the specific topic of "machine number sense". Crucially, almost all the prior work presented mathematical problems in the form of text, which will lead to a sequential (thus simplified) problem-solving process. This process is different from the flexible nature of human cognition shown in mathematics.

5.3 Machine IQ and Analogy

Another highly related stream of research focuses on relational and analogical reasoning (Barrett et al. 2018; Zhang et al. 2019a; 2019b). Recently, researchers proposed to use deep neural networks to solve Raven's Progressive Matrices (RPM) (Raven 1936; Raven and Court 1998). Unlike the proposed number sense challenge, RPM involves a wider range of object relations, such as figure addition, subtraction, and distribution. However, the requirement of number sense in RPM is less demanding than the proposed dataset: only a limited number of objects are involved in each RPM instance and there is no need for decomposition based on the problem context. A similar setting is studied in (Edmonds et al. 2020; 2019), where an agent needs to reason about the open-lock mechanism by generalizing from mechanistically similar but visually different environments. Our work echoes their conclusion that current methods in training deep neural networks do not help the models acquire a generalizable representation.

To solve RPM problems by computational modeling, previous works also adopted knowledge-based methods, outlining analogical reasoning in a predefined manner. For example, some researchers established structural mappings between RPM image panels, which directly processed the

visual objects as a set of abstract concepts and variation rules (Lovett et al. 2010; Lovett, Forbus, and Usher 2010; Lovett and Forbus 2017). With the provided symbolic concepts and relation rules, the RPM problems can be converted to search problems, leading to an optimal problem-solving accuracy outperforming human. Inspired by these previous investigations, the search-based algorithms in our work were also equipped with prior knowledge of numbers and calculations, as well as analogical mappings between image panels. However, as the arithmetic problems in the MNS dataset possess much larger search space than RPM problems, it is hard and consumptive to solve these problems simply by a pure symbolic search regardless of contextual information. An additional perception module could accelerate the search process with some heuristics from problem contexts, addressing the challenge to some degree; but an apparent gap between search efficiency and human performance still exists. Compared with the previous work, our work reveals the insufficiency of knowledge-based methods in integrating provided knowledge, problem content, and contextual information to conduct human-like adaptive problem-solving.

6 Discussions and Conclusion

In this paper, we propose a dataset generated by And-Or Graph (AOG) to examine the *machine number sense*. Specifically, we evaluate machines' understanding of abstract number symbols and competence of context-based problem-solving. Compared to simple symbolic search-based models, the poor performance of neural network models suggests its insufficiency in symbolic processing and concept understanding, as well as its difficulty in combining content and context to solve problems flexibly.

The dataset and experiments have left room for improvements and brought up inspirations for future work. The critical challenges are how to *emerge* symbolic concepts directly from pixels using minimal supervisions, how to extract *meaningful* relations from the contextual information, and how to reason and make inductions based on these concepts and relations. As the experiments indicated, fusing neural network models' strong capacity of visual feature extraction in large-scale data processing and search-based algorithms' explicit knowledge structure in fit-for-purpose problem-solving may be an effective method for relational and abstract reasoning; the integration of data-driven and knowledge-based methods will complement each other.

Acknowledgement

The authors thank Prof. Hongjing Lu at UCLA Psychology Department for helpful discussions. This work reported herein is supported by MURI ONR N00014-16-1-2007, DARPA XAI N66001-17-2-4029, ONR N00014-19-1-2153, and an NVIDIA GPU donation grant.

References

Barrett, D. G.; Hill, F.; Santoro, A.; Morcos, A. S.; and Lillicrap, T. 2018. Measuring abstract reasoning in neural networks. *arXiv preprint arXiv:1807.04225*.

Bobis, J. 1996. Visualisation and the development of number sense with kindergarten children. *Children's number learning: A research monograph of MERGA/AAMT Adelaide: Australian Association of Mathematics teachers*.

Brannon, E. M., and Terrace, H. S. 1998. Ordering of the numerosities 1 to 9 by monkeys. *Science* 282(5389):746–749.

Cantlon, J. F., and Brannon, E. M. 2007. How much does number matter to a monkey (macaca mulatta)? *Journal of Experimental Psychology: Animal Behavior Processes* 33(1):32.

Cattell, R. B. 1963. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of educational psychology* 54(1):1.

Chard, D. J.; Clarke, B.; Baker, S.; Otterstedt, J.; Braun, D.; and Katz, R. 2005. Using measures of number sense to screen for difficulties in mathematics: Preliminary findings. *Assessment for Effective Intervention* 30(2):3–14.

Cobb, P.; Wood, T.; Yackel, E.; Nicholls, J.; Wheatley, G.; Trigatti, B.; and Perlwitz, M. 1991. Assessment of a problem-centered second-grade mathematics project. *Journal for research in mathematics education* 3–29.

Davies, J., and Goel, A. K. 2001. Visual analogy in problem solving. In *IJCAI*.

Davis, H., and Pérusse, R. 1988. Numerical competence in animals: Definitional issues, current evidence, and a new research agenda. *Behavioral and Brain Sciences* 11(4):561–579.

Dehaene, S.; Dehaene-Lambertz, G.; and Cohen, L. 1998. Abstract representations of numbers in the animal and human brain. *Trends in neurosciences* 21(8):355–361.

Dehaene, S. 2011. *The number sense: How the mind creates mathematics*. OUP USA.

Edmonds, M.; Qi, S.; Zhu, Y.; Kubricht, J.; Zhu, S.-C.; and Lu, H. 2019. Decomposing human causal learning: Bottom-up associative learning and top-down schema reasoning. In *CogSci*.

Edmonds, M.; Ma, X.; Qi, S.; Zhu, Y.; Lu, H.; and Zhu, S.-C. 2020. Theory-based causal transfer: Integrating instance-level induction and abstract-level structure learning. In *AAAI*.

Fisher, D. H.; Pazzani, M. J.; and Langley, P. 2014. *Concept formation: Knowledge and experience in unsupervised learning*. Morgan Kaufmann.

Gallistel, C. R. 1989. Animal cognition: The representation of space, time and number. *Annual review of psychology* 40(1):155–189.

Gallistel, C. R. 1990. *The organization of learning.* The MIT Press.

Gallistel, C. R. 2003. Animal cognition: the representation of space, time and number. *Annual Review of Psychology* 40(40):155–189.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Heinze, A.; Star, J. R.; and Verschaffel, L. 2009. Flexible and adaptive use of strategies and representations in mathematics education. *ZDM* 41(5):535–540.

Hill, F.; Santoro, A.; Barrett, D. G. T.; Morcos, A. S.; and Lillicrap, T. P. 2019. Learning to make analogies by contrasting abstract relational structure. *ArXiv* abs/1902.00120.

Huang, D.; Shi, S.; Lin, C.-Y.; Yin, J.; and Ma, W.-Y. 2016. How well do computers solve math word problems? large-scale dataset construction and evaluation. In *ACL*.

Jacob, S. N., and Nieder, A. 2008. The abc of cardinal and ordinal number representations. *Trends in cognitive sciences* 12(2):41–43.

Kalchman, M.; Moss, J.; and Case, R. 2001. Psychological models for the development of mathematical understanding: Rational numbers and functions. *Cognition and instruction: Twenty-five years of progress* 1–38.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Knoblich, G.; Ohlsson, S.; Haider, H.; and Rhenius, D. 1999. Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, memory, and cognition* 25(6):1534.

Kushman, N.; Artzi, Y.; Zettlemoyer, L.; and Barzilay, R. 2014. Learning to automatically solve algebra word problems. In *ACL*.

Lake, B. M.; Salakhutdinov, R.; and Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. *Science* 350(6266):1332–1338.

Ling, W.; Yogatama, D.; Dyer, C.; and Blunsom, P. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *arXiv preprint arXiv:1705.04146*.

Lovett, A., and Forbus, K. 2017. Modeling visual problem solving as analogical reasoning. *Psychological review* 124(1):60.

Lovett, A.; Tomai, E.; Forbus, K.; and Usher, J. 2010. Solving geometric analogy problems through two-stage analogical mapping. *Cognitive Science* 33(7):1192–1231.

Lovett, A.; Forbus, K.; and Usher, J. 2010. A structure-mapping model of raven's progressive matrices. In *CogSci*.

Nieder, A., and Dehaene, S. 2009. Representation of number in the brain. *Annual review of neuroscience* 32:185–208.

Nisbett, R. E.; Peng, K.; Choi, I.; and Norenzayan, A. 2001. Culture and systems of thought: holistic versus analytic cognition. *Psychological review* 108(2):291.

Okamoto, Y., and Case, R. 1996. Ii. exploring the microstructure of children's central conceptual structures in the domain of number. *Monographs of the Society for research in Child Development* 61(1-2):27–58.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch. In *ICLR*.

Raven, J. C., and Court, J. H. 1998. *Raven's progressive matrices and vocabulary scales*. Oxford pyschologists Press.

Raven, J. C. 1936. Mental tests used in genetic studies: The performance of related individuals on tests mainly educative and mainly reproductive. Master's thesis, University of London.

Santoro, A.; Raposo, D.; Barrett, D. G.; Malinowski, M.; Pascanu, R.; Battaglia, P.; and Lillicrap, T. 2017. A simple neural network module for relational reasoning. In *NeurIPS*.

Saxton, D.; Grefenstette, E.; Hill, F.; and Kohli, P. 2019. Analysing mathematical reasoning abilities of neural models. *arXiv preprint arXiv:1904.01557*.

Schneider, M.; Beeres, K.; Coban, L.; Merz, S.; Susan Schmidt, S.; Stricker, J.; and De Smedt, B. 2017. Associations of non-symbolic and symbolic numerical magnitude processing with mathematical competence: A meta-analysis. *Developmental science* 20(3):e12372.

Simpson, J., and Weiner, E. 1989. Oxford english dictionary. *Dictionary, Oxford English*.

Stafford, B. M. 2001. *Visual analogy: Consciousness as the art of connecting*. MIT press.

Starkey, P., and Cooper, R. G. 1980. Perception of numbers by human infants. *Science* 210(4473):1033–1035.

Taylor, T. 1818. *The Nicomachean ethics*. AJ Valpy.

Temple, E., and Posner, M. I. 1998. Brain mechanisms of quantity are similar in 5-year-old children and adults. *PNAS* 95(13):7836–7841.

Tenenbaum, J. B. 1999. Bayesian modeling of human concept learning. In *NeurIPS*.

Upadhyay, S., and Chang, M.-W. 2016. Annotating derivations: A new evaluation strategy and dataset for algebra word problems. *arXiv preprint arXiv:1609.07197*.

Waltz, J. A.; Knowlton, B. J.; Holyoak, K. J.; Boone, K. B.; Mishkin, F. S.; de Menezes Santos, M.; Thomas, C. R.; and Miller, B. L. 1999. A system for relational reasoning in human prefrontal cortex. *Psychological science* 10(2):119–125.

Wang, Y.; Liu, X.; and Shi, S. 2017. Deep neural solver for math word problems. In *EMNLP*.

Wertheimer, M. 1923. Laws of organization in perceptual forms. *A source book of Gestalt Psychology*.

Wiese, H. 2003. *Numbers, language, and the human mind*. Cambridge University Press.

Wynn, K. 1992. Addition and subtraction by human infants. *Nature* 358(6389):749.

Xingjian, S.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; and Woo, W.-c. 2015. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NeurIPS*.

Zhang, C.; Gao, F.; Jia, B.; Zhu, Y.; and Zhu, S.-C. 2019a. Raven: A dataset for relational and analogical visual reasoning. In *CVPR*.

Zhang, C.; Jia, B.; Gao, F.; Zhu, Y.; Lu, H.; and Zhu, S.-C. 2019b. Learning perceptual inference by contrasting. In *NeurIPS*.

Zhu, S.-C., and Mumford, D. 2007. A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision* 2(4):259–362.