Spatial IQ Test for AI

Erwin Hilton, Qianli Liao and Tomaso Poggio
Center for Brains, Minds and Machines, MIT

***Abstract*—We introduce SITD (Spatial IQ Test Dataset), a dataset used to evaluate the capabilities of computational models for pattern recognition and visual reasoning. SITD is a generator of images in the style of the Raven Progressive Matrices (RPM), a common IQ (Intelligence Quotient) test used to test analytical intelligence. RPMs are purely visual, and require little prior knowledge. RPM tests the users ability to derive abstract rules and patterns from a set of images.**

For the last 100 years, humans have evaluated intelligence using standardized intelligence quotient exams. These tests examine different aspects of intelligence, including verbal, quantitative reasoning, and spatial reasoning ability. In the field of AI, there exists few intelligence established metrics beyond the Turing Test (TT) and the Total Turing Test (TTT). Thus, SITD makes for a useful dataset researchers can use to divide and conquer the task of creating 'intelligent' machines.

***Index Terms*—Dataset, IQ Test, AI, Turing Test**

I. Introduction

What is intelligence? Intelligence is the ability to acquire knowledge and skills. Fundamentally, defining intelligence is a difficult task. Marvin Minsky famously called intelligence a *suitcase* word [5] - a word densely packed with many different definitions. Yet for the last century and a half, psychologists have developed the field of psychometrics to measure humans' mental capacities and processes. Psychometrics use standardized tests such as the IQ (Intelligence Quotient) test to assess cognitive abilities. While not a complete assessment of all types of cognitive ability, IQ tests have shown strong correlations as predictors of human cognitive performance and later success in life. There is also a strong correlation between the performance on IQ tests and performance on other cognitive tasks. This is known as the *g* factor [3], a particular measure of *general* intelligence.

In this paper, we introduce a dataset of spatial intelligence IQ questions based on Raven's Progressive Matrices [6] (**Figure 1**). Raven's Progressive Matrices (RPM) is the most popular IQ test administered to demographic groups ranging from children to the elderly [4].

All of these questions are generated, and therefore not subject to the IQ test designer's copyright.

Currently, psychometric tests are used to evaluate cognitive ability over a wide range of abilities. The definition of intelligence is thus reduced to excelling at a group of established intelligence tests. Evaluating intelligence by a set of intelligence tests also extends to AI agents, as [1] showed. If the various aspects of intelligence are split up and evaluated using different tests, then the development of human-level intelligence AI models can be encouraged by dividing and conquering different intelligence aptitude tests. This goes beyond the Turing Test, which focuses on the *appearance* of intelligence, and instead evaluates each currently known and previously tested aspect of intelligence.

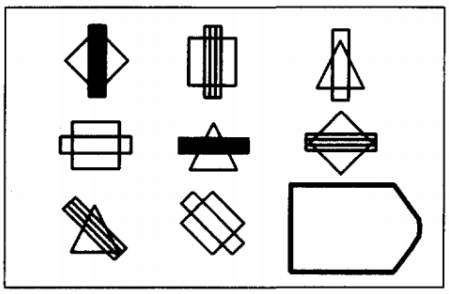

Fig. 1. Example of a Raven's Progressive Matrix, from Carpenter et al.. The task is to identify the best possible candidate figure for the space in the bottom right of the image, out of a set of multiple choice candidate figures. The rules used in the creation of this sample RPM are **distribution of three** and **constant in a row**. The geometric shapes used in each row (diamond, square, triangle) follow a distribution of three. The texture of the bars intersecting the geometric shapes also follow a distribution of three (solid, striped, or clear). The rotation of the bar follows the constant in a row rule (vertical, horizontal, or diagonal).

Thus, by developing freely available psychometric datasets for artificial intelligence researchers, we can use the datasets to evaluate subsets of a machine's cognitive skills.

II. Psychometric Test for AI

III. Related Work

IV. Dataset Generation

The dataset is generated according to a set of five rules previously found [2] to have governed the variation among most of the Raven Progressive Matrices analyzed. These rules are described as follows:

1) **Constant in a row** - the same categorical attribute occurs in the same row, but changes down a column. For example, one row could be made up of squares, the second circles, and the third diamonds
2) **Quantitative Pairwise Progression** - a quantitative increment (size, position, number, etc) occurs between adjacent pairs of images. In this case, using the example for constant in a row would be the shapes gradually scale down as you go along each element in the row.
3) **Distribution of Three** - three values from a categorical attribute, such as the shapes used, are distributed throughout a row. For example, each row could contain three different shapes but the order changes for each row
4) **Figure addition or subtraction** - a figure from one column is added or subtracted from a figure in another column to produce a new figure

5) **Distribution of Two** - three values from a categorical attribute are distributed throughout a row, but for each image only a subset of two values are used while the third value goes unused

The difficulty of a generated RPM is assigned according to a complexity metric $C \in [1, 2, 3]$, which corresponds to the amount of rules used to generate the RPM. For example, if $C = 2$ then any combination of two of the above rules, allowing for duplicates, can be used to generate the RPM.
As mentioned in the rule set description, each of the rules depended on selecting categorical attributes and modifying quantitative values of these attributes. Only a subset of the categorical attributes are able to be quantitatively modified.

The categorical attributes used in the RPM generation are listed as follows:

- **Shape** - a 2d form such as polygons, circles, ellipses, arcs, lines, etc
- **Size** - the shape is set to a certain scale
- **Count** - the amount of similar shapes in the same figure
- **Color** - the color of the shape. The set of colors is denoted by $C = [white, black, blue, red, green, yellow]$
- **Orientation** - the orientation (set rotation) of the shape
- **Position** - where the shape is positioned relative to its background. The background of the shape is constrained to its sub-section of the RPM, where a standard RPM has nine separate sections divided into a 3x3 matrix.

The set of applicable quantitative transformations to the categorical attributes in quantitative pairwise progression:

- **Scaling** - increase or decrease in the size of the shape, where the scale factor $S \in [-1.5, 1.5]$
- **Number count** - the amount of shapes increases or decreases (duplicates or very similar shapes with slight transformations applied). The tally of shapes is a number in $T = [1, 2, 3, 4]$
- **Color change** - the color of a shape changes, to any other color in the set of colors *C*
- **Rotation** - the shape can rotate around its orientation by an angle increment. The possible angle increments range in $R \in [0, 2\pi)$
- **Polygonal side count change** - if the shape is a polygon, the amount of sides can be incremented or decremented. A polygon's side count is constrained to a number in $P_c = [3, 4, 5, ..., 9]$
- **Translation** - the shape moves to a new position

A matrix M of nine different figures are generated using these rules, in the style of Raven's Progressive Matrices.

The matrix M has the form

$$\begin{array}{ccc} \mathrm{M}_{1,1} & \mathrm{M}_{1,2} & \mathrm{M}_{1,3} \\ \mathrm{M}_{2,1} & \mathrm{M}_{2,2} & \mathrm{M}_{2,3} \\ \mathrm{M}_{3,1} & \mathrm{M}_{3,2} & ? \end{array}$$

The figure in the bottom right of the matrix, at indice $M_{3,3} =?$, is the missing component of the RPM left as a task for the examinee to predict. While still generated using the ruleset, $M_{3,3}$ is omitted from the *question* image and replaced with a blank space.

Instead, the correct solution to $M_{3,3}$ is included in a separate set of multiple choice candidate answer images, along with three other erroneous images created using incorrect yet similar variations of the rules and categorical attributes. **Figure 2** shows a sample RPM generated by our code, along with its corresponding answer key.

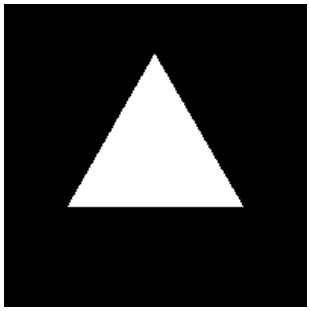

Fig. 2. A generated RPM using our code. This RPM employs the **distribution of three** and **constant in a row**

V. Conclusion

We presented a generative dataset that presents a selection of images based off Raven's Progressive Matrices, a test that is widely used in IQ tests to measure spatial intelligence. The Python dataset generation code will be publicly released.

With this dataset, we hope to open the doors of psychometric IQ testing to artificial intelligence researchers. It is our view that current AI models are unable to solve general reasoning tasks without prior knowledge of sample space. Thus this dataset can be used as a tool for artificial intelligence researchers to use in the development of models that can bridge the gap between the various aspects of machine and human 'intelligence'.

References

[1] Selmer Bringsjord and Bettina Schimanski. What is artificial intelligence? Psychometric AI as an Answer. 2003.

[2] Patricia A. Carpenter, Marcel A. Just, and Peter Shell. What one intelligence test measures: A theoretical account of the processing in the raven progressive matrices test. 1990.

[3] Arthur R. Jensen. The g factor. the science of mental ability. 1998.

[4] R. M. Kaplan and D. P. Saccuzzo. Standardized tests in education, civil service, and the military. psychological testing: Principles, applications, and issues. 2009.

[5] Marvin Minsky. The emotion machine. 2006.

[6] J. Raven, J.C. Raven, and J.H. Court. Manual for raven's progressive matrices and vocabulary scales. 2003, updated 2004.