Similarity-Based Reasoning, Raven's Matrices, and General Intelligence

Can Serif Mekik[a], Ron Sun[a], David Yun Dai[b]
[a]Rensselaer Polytechnic Institute
[b]State University of New York at Albany
mekikc@rpi.edu, rsun@rpi.edu, ydai@albany.edu

Abstract

This paper presents a model tackling a variant of the Raven's Matrices family of human intelligence tests along with computational experiments. Raven's Matrices are thought to challenge human subjects' ability to generalize knowledge and deal with novel situations. We investigate how a generic ability to quickly and accurately generalize knowledge can be succinctly captured by a computational system. This work is distinct from other prominent attempts to deal with the task in terms of adopting a generalized similarity-based approach. Raven's Matrices appear to primarily require similarity-based or analogical reasoning over a set of varied visual stimuli. The similarity-based approach eliminates the need for structure mapping as emphasized in many existing analogical reasoning systems. Instead, it relies on feature-based processing with both relational and non-relational features. Preliminary experimental results suggest that our approach performs comparably to existing symbolic analogy-based models.

1 Introduction

Psychometric intelligence tests can further the understanding of the computational aspects of natural and artificial intelligence [Bringsjord and Schimanski, 2003; Hernández-Orallo *et al.*, 2016]. It has even been suggested that psychometric intelligence tests should form a set of benchmark tasks against which progress in the field of artificial intelligence is assessed [Bringsjord and Schimanski, 2003]. Indeed, psychometric tests present well-defined, validated tasks that place acute demands on key aspects of intelligence. In this paper, a new model of *Raven's Progressive Matrices* (RPM) [Raven *et al.*, 1998a] is presented. This model suggests a subsymbolic approach to similarity-based (including analogical) reasoning and draws a connection between similarity-based reasoning, analogy making, and a form of inductive inference.

Is there a general ability that underlies intelligent behavior, or is intelligent behavior the result of multiple distinct abilities, each addressing a limited domain? If a general ability exists, what purpose does it serve, and how does it operate? These were some key questions Spearman [1927] sought to answer in his influential work on human intelligence. In artificial intelligence, these questions are closely related to the topic of generality, the task of developing systems that can handle a wide variety of tasks, environments, and domains. Traditionally, generality has been a topic in symbolic and logical artificial intelligence, and several domain-general approaches have been proposed [Besold and Schmid, 2016]. These systems often give pride of place to analogical reasoning as it is a paradigm for effectively applying knowledge from unrelated domains to new tasks. Generality is also addressed in subsymbolic and statistical approaches to artificial intelligence, e.g., in transfer learning [Pan and Yang, 2010]. However, these latter efforts focus on techniques for generalizing knowledge between specific tasks to improve learning as opposed to developing flexible and domain-general systems. Notably, subsymbolic approaches to analogical reasoning are relatively rare.

The RPM tests were developed based on Spearman's work on human general intelligence [Raven, 2008; Raven *et al.*, 1998a]. Spearman believed that there exists a set of fundamental cognitive abilities essential for generating novel ideas and dealing with unfamiliar situations. These abilities, collectively termed *eductive ability*, include the ability to draw out and identify relationships between ideas and the ability to form or identify ideas that enter into given relationships with given ideas. RPM tests are designed as direct measures of eductive ability and they are considered to be among the best measures of this ability [Nisbett *et al.*, 2012]. Therefore, RPM tests are ideal material for computational approaches to the investigation of eductive ability [Carpenter *et al.*, 1990].

Although the present study, as the existing computational literature suggests, recognizes the important role of analogical reasoning as a mechanism for generality, it views analogical reasoning as a special case of similarity-based reasoning [Sun, 1995; 1994]. This analysis allows the model to operate using subsymbolic processes, in contrast with existing symbolic models of analogical reasoning. Typically, models of analogical reasoning form analogies by formally aligning structured representations between a *source* and a *target* (i.e., structure mapping) [Gentner and Forbus, 2011]. In the present work, analogy making is viewed in a different way. First, potential analogues are represented using a set of basic or relational features [Sun, 2016]; then, potential targets are evaluated on the basis of their similarity to sources with respect to these features [Turney, 2011]. Similarities are obtained by

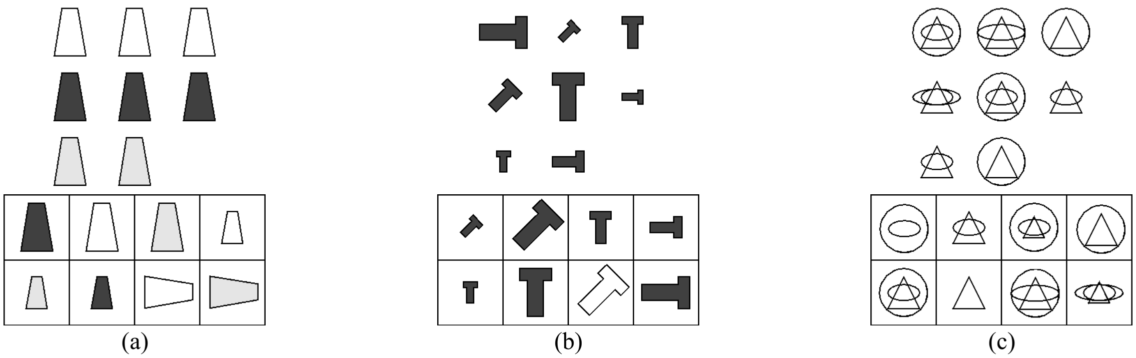

Figure 1: Three matrix problems in the style of Raven's matrices taken from the Matzen *et al.* [2010] dataset.

computing the entropy of the target relative to the source. Thus, the model offers a subsymbolic approach to analogical reasoning where structured representations are replaced by basic and relational features, and structure mapping is replaced by the relative entropy similarity measure.

2 Matrix Problems

All items of the RPM tests are visual pattern completion problems of a specific kind, hereby called *matrix problems*. See Figure 1 for some example matrix problems.

Specifications for matrix problems are presented in Penrose and Raven [1936]. According to Penrose and Raven, matrix problems are modeled on analogical proportions, that is, statements of the form "A is to B as C is to D." Each item features a 2×2 or 3×3 matrix of visual figures where the bottom right corner is left blank. The task is to identify the figure, from a set of six to eight alternatives, that best completes the matrix when inserted into the blank.

Since 3×3 matrices are more difficult, we describe the structure of only these variants in detail. Letting A be a figure and f_1, f_2, ϕ_1, ϕ_2 be figure transformations, 3×3 matrices have the following form:

$$\begin{matrix} A & f_1A & f_2A \\ \phi_1 A & \phi_1 f_1 A & \phi_1 f_2 A \\ \phi_2 A & \phi_2 f_1 A & \phi_2 f_2 A \end{matrix} \tag{1}$$

where transformations are selected so as to ensure that pairs f_i, ϕ_j commute on figure A ($f_i \phi_j A = \phi_j f_i A$) for all $(i,j) \in \{1,2\}^2$.

3 Similarity-Based Approach

Complete matrix rows and columns (those with no blank) are called *matrix sequences*. Inserting an alternative into the blank completes two additional sequences, one each along the row and column axes; we call such sequences *alternative sequences*.

To disambiguate between sequences of the same type (matrix or alternative) and along the same axis (row or column), let n be an index called the *sequence number*. For a matrix sequence, the sequence number refers to the index of the sequence along the relevant axis and, for an alternative sequence, it refers to the alternative array index of the alternative figure used to complete the sequence.

We write $\boldsymbol{s}(m, t, d, n)$ to denote a sequence from item m of type t, along axis d, and of sequence number n (we sometimes use the abbreviated notation $\boldsymbol{s}$). Indexes are assigned according to standard English reading order (left to right, top to bottom). For example, we would denote the top row of Figure 1a as $\boldsymbol{s}(1a, mat, row, 1)$ and the alternative column produced by the bottom right alternative as $\boldsymbol{s}(1a, alt, col, 8)$.

3.1 Selection Procedure

The entropy of a probability distribution q relative to a distribution p, denoted $D(q||p)$, and given, in the discrete case, by

$$D(q||p) = \sum_x q(x) \log \frac{q(x)}{p(x)} \tag{2}$$

is a measure of the similarity of distribution q to distribution p. To the extent that q and p respectively describe characteristics of alternative and matrix sequences for a given matrix and alternative pair, $D(q||p)$ is a measure of the similarity of the alternative sequences to the matrix sequences [cf. Sun, 1995; 1994].

Using a convolutional neural network, our model constructs, for a matrix m, a reference distribution $p(m)$ based on matrix sequence features, as well as alternative distributions $q(m, a)$ for each alternative a (details below). The model then selects an alternative a^* as its response, according to the *Boltzmann distribution* with temperature τ:

$$\Pr[\text{response} = a^*] = \frac{\exp(-D(q(m, a^*)||p(m))/\tau)}{\sum_a \exp(-D(q(m, a)||p(m))/\tau)} \tag{3}$$

The selection procedure thus favors alternatives that minimize $D(q||p)$, that is, those that produce alternative sequences that are more similar to corresponding matrix sequences. For the purposes of the present paper, $\tau = 1$.

3.2 Similarity Computation

Matrix problem specifications suggest that sequences along common axes exhibit common features. Since, for an $N \times N$ matrix, there are $N - 1$ matrix sequences along a given axis d (because one sequence contains the blank item), distributions for a feature F_i (e.g., shape distribution; see 4.1) obtained from these sequences are combined to form a matrix feature distribution, $p_{i,d}(m)$. In the present study, only binary features were used. As such, the combined distributions $p_{i,d}(m)$ are defined as:

Feature	Definition
Constant	All figures are identical.
Shape Distribution	Each figure depicts a distinct shape.
Shading Increments	Figure shadings get progressively lighter or darker by even increments.
Shading Distribution	Each figure features a distinct shading.
Orientation Increments	Figure orientations are progressively incremented/decremented by a set angle.
Orientation Distribution	Each figure features a distinct orientation.
Size Increments	Figure sizes are progressively incremented/decremented by a set factor.
Size Distribution	Each figure features a distinct size.
Numerosity Increments	Figure numerosities are progressively incremented/decremented by a set number.
Numerosity Distribution	Each figure exhibits a distinct numerosity.
Edgewise AND	Third figure corresponds to edgewise intersection of first two figures.
Edgewise XOR	Third figure corresponds to edgewise symmetric difference of first two figures.
Edgewise OR	Third figure corresponds to edgewise union of first two figures.

Table 1: Sequence Feature Definitions. Each definition concerns figures in a given sequence.

$$p_{i,d}(F_i = 1|m) := \left(\prod_{j=1}^{N-1} f_i\big(F_i = 1\big|\boldsymbol{s}(m, \text{mat}, d, j)\big)\right)^{\frac{1}{N-1}} \quad (4)$$

$$p_{i,d}(F_i = 0|m) := 1 - p_{i,d}(F_i = 1|m) \quad (5)$$

where the symbol f_i denotes a probability distribution over possible values of feature F_i. For instance, if F_i is the size increments feature (see 4.1) and $\boldsymbol{s}$ is some figure sequence,

$$f_i(F_i = x|\boldsymbol{s}) := \begin{cases} \Pr[F_i \text{ present}|\boldsymbol{s}] & if\ x = 1 \\ \Pr[F_i \text{ absent}|\boldsymbol{s}] & if\ x = 0 \end{cases} \quad (6)$$

Although all features discussed in the present paper are binary, our model can accommodate non-binary and continuous features.

Note that $p_{i,d}(F_i = 1|m)$ is just the *geometric mean* of $f_i\big(F_i = 1\big|\boldsymbol{s}_j\big)$ obtained for each matrix sequence $\boldsymbol{s}_j$ along d. The distribution $p_{i,d}(m)$ represents whether feature F_i is characteristic of axis d of matrix m. In a similar manner, alternative feature distributions for each alternative a, $q_{i,d}(m, a)$, are constructed for each feature and axis, though no combination process is necessary.

Assuming subjective probabilities for features are mutually independent, the entropy of a joint subjective probability distribution for alternative features, $q(m, a)$, relative to the joint subjective probability distribution for matrix features, $p(m)$, can be expressed as the sum, over features and matrix axes, of relative entropies of individual features. For a matrix m, the overall similarity of alternative sequences constructed using an alternative a to corresponding matrix sequences, denoted as $D(q(m, a)||p(m))$, can thus be calculated as:

$$D(q(m, a)||p(m)) = \sum_i \sum_d D\big(q_{i,d}(m, a)||p_{i,d}(m)\big) \quad (7)$$

where $1 < i < n$ and $d \in \{\text{row}, \text{col}\}$.

3.3 Algorithm

To put everything together, in response to a matrix with 8 alternative answers (e.g., Figure 1a) the model proceeds as follows:

1. Using a *convolutional neural network*, subjective probabilities are obtained for each individual feature in each matrix sequence and each alternative sequence.
2. Feature distributions for matrix sequences are combined, row-wise and column-wise, using *geometric means* (Equations 4 and 5).
3. For each alternative, the similarity of the 2 resulting alternative sequences and the 2 combined matrix sequences is computed, with regard to each feature, row-wise and column-wise, through *relative entropy* (Equation 7).
4. The 8 resulting similarity measures are used to stochastically select an answer out of the 8 alternatives, using the *Boltzmann distribution* (Equation 3).

4 Computational Experiments

The model was tested on a set of matrix problems generated using the Sandia Matrix Generation Tool [Matzen *et al.*, 2010].[1] These generated matrices, henceforth *Sandia matrices*, have the same problem structure as Raven's original matrices. In Matzen *et al.* [2010], tests developed using these matrices were shown to have psychometric characteristics comparable to the *Standard Progressive Matrices* (SPM) [Raven *et al.*, 1998b] variant of the RPM in a norming study.

A deep convolutional neural network was trained to estimate subjective probabilities of features. A total of 1480 input-output pairs (20 from each of 74 matrices) were used to train the model.

To gain some understanding of processing in the hidden layers, further analysis was carried out. For example, a hierarchical cluster analysis on activation patterns in the network's first fully connected layer was used to assess the quality of learned representations.

4.1 Feature Set

For the purposes of the present experiment, the model was implemented using a set of thirteen figure sequence features.

[1] Experimental notes, data and code are available at https://osf.io/7fy2d/.

Feature Type	Pr	Criteria
Constant	1.0	All figures are identical.
	0.5	Exactly two figures are identical.
	0.0	No figures are identical.
X Distribution	1.0	Each figure exhibits a distinct value of feature X (e.g., distribution of different shapes).
	0.5	Only one figure exhibits a distinct value of feature X.
	0.0	All sequence figures exhibit the same value of feature X.
X Increments	1.0	In each figure, feature X is incremented by one unit relative to the previous figure (e.g., progression of sizes).
	0.5	Figures exhibit values of feature X in ascending or descending order, but with variable increment or decrement magnitudes (including zero).
	0.0	Figures exhibit no variation with respect to feature X; or, the values of X are not in descending or ascending order.
Edgewise Operator X	1.0	Edges of the third figure are identical to the result of applying operator X to edges of the first two figures.
	0.5	Edges of the third figure agree with only part of the result of applying operator X to edges of the first two figures; or, the whole result is depicted as well as some unrelated edges.
	0.0	Edges of the third figure depict only part of the result of applying operator X to the edges of the first two figures as well as some unrelated edges; or, edges of the third figure do not depict any discernible non-trivial part of the result of applying X to the edges of the first two figures.

Table 2: Subjective Probability Assignment Criteria. Each criterion concerns figures in a given sequence.

Specific features were defined following feature types present in the literature [e.g., distribution and progression rules in Carpenter *et al.*, 1990]. Detailed feature definitions are given in Table 1. Table 2 defines labeling criteria used to determine subjective probabilities for each feature type. Figure sequences were labeled with subjective probabilities according to these criteria.

4.2 Neural Network

The model's perceptual neural network takes a complete sequence of figures (e.g. a row) as input and outputs subjective probabilities for each sequence feature. The network has a total of 13 outputs whose values range in $[0,1]$. Sequence figures are presented to the network in a 28×28 grayscale format for a total of 2352 input values. (Input values are scaled down by 255 and shifted by -0.5 before being passed to the network in order to avoid saturating nodes in the first layer.)

Architecture

The network has a total of 7 layers and is divided into two major components. The *convolutional stack*, consisting of the first five layers, processes each of the three figures independently. Then, the *fully connected layers* (FC layers) combine the information flowing from the convolutional stack in order to produce subjective probability assignments. Nonlinearities in all but the final layer are given by the tanh function. The final layer makes use of the logistic activation function, $\sigma(x) = \frac{1}{1+e^{-x}}$, to ensure that network outputs lie in the interval $[0,1]$. Remaining network meta-parameters are presented in Table 3; see Figure 2 for a graphical representation of network architecture.

Training Procedure

The network was trained using the ADADELTA algorithm [Zeiler, 2012] for 24000 epochs with learning parameters $\rho = .6$ and $\epsilon = 10^{-10}$. The Xavier method [Glorot and Bengio, 2010] was used for network weight initialization. Training samples were shuffled before being split into 20 mini-batches of 74 input-output pairs. Mini-batch order was shuffled at the start of each epoch, and, every 500 epochs, network weights were recorded. A cross-entropy error measure augmented with L_1 ($\lambda_1 = .003$) and L_2 ($\lambda_2 = .003$) weight penalties was used to compute weight updates.

4.3 Results

The performance metric of interest is the expected percentage of matrices correctly solved, which is equal to the average, over test items, of the probability of a correct response. Maximum performance was found to be 85.01%.

Representation Analysis

A hierarchical clustering technique was used to analyze distributed representations [Servan-Schreiber *et al.*, 1991] in the final convolutional layer (Layer 5) of the neural network.

Layer	Type	Shape	Stride	Padding
1	Conv	(5,1,7,7)	(3,3)	(3,3)
2	Conv	(10,5,2,2)	(2,2)	-
3	Conv	(15,10,3,3)	(1,1)	-
4	Conv	(20,15,2,2)	(1,1)	-
5	Conv	(96,20,2,2)	(1,1)	-
6	Full	(3,96,1,1,96)	-	-
7	Full	(96,13)	-	-

Table 3: Neural Network Meta-Parameters. The Shape column lists weight array dimensions for the corresponding layer. If the layer is convolutional, the shape tuple lists, in respective order, output depth, input depth, height, and width. If the layer is fully connected, the last shape tuple entry corresponds to the number of layer nodes and the remaining entries correspond to dimensions of the input array. Notice that Layer 6 receives a stack of three inputs of shape (96,1,1). These are the final abstract features extracted from each of the three figures in a sequence.

This analysis identifies inputs that are represented in a similar manner by the layer. The analysis was conducted as follows. Pairwise Euclidian distances between node activations representing different figure sequences were first computed, and then figure sequences were incrementally merged into clusters starting with the most similar pairs. Obtained clusters were compared to a reference cluster using the Fowlkes-Mallows index [Fowlkes and Mallows, 1983], with the tree cut so as to produce two clusters. Reference clusters grouped together matrix sequences and answer sequences (alternative sequences containing the answer figure) on the one hand, and distractor sequences (other alternative sequences) on the other. The analysis revealed strong agreement between the reference clusters and clusters obtained from Layer 5 representations ($M = .72, SD = .10$).

Other Experiments

To explore the contributions of the convolutional layers, a version of the neural network was trained with all convolutional nonlinearities removed, thus leaving the network with only a single hidden layer in actuality. Removing nonlinearities reduced model performance significantly, suggesting that the convolutional architecture of the network contributes significantly to the model's feature detection ability.

5 Discussion

This paper presents a new similarity-based model for solving matrix problems. Our approach has some implications for the pursuit of generality in artificial intelligence; our experiments so far seem to suggest so. Our model can capture analogical reasoning, which is a plausible contender for realizing generality, and suggests that feature learning and detection are important to achieving generality (see 5.2).

The relative entropy similarity measure used in this work has a number of interesting characteristics. In particular, some notable formal properties of relative entropy such as non-symmetry and non-subadditivity (possible violations of the triangle inequality) agree with experimental observations about human similarity judgments [Tversky, 1977]. Furthermore, relative entropy minimization is a principle of inductive inference [also known as *cross-entropy minimization*, cf. Shore and Johnson, 1980], which is closely related to Occam's razor [Feldman, 2016].

5.1 Related Work

The present work builds on Mekik *et al.* [2017], where a similar neural network was used for the extraction of features representing relations between two figures. The present model's response selection strategy differs from that of the Mekik *et al.* approach in that it employs a domain-agnostic subsymbolic selection procedure whereas the latter employed a domain-specific rule-based response selection procedure.

A majority of the computational literature on RPM is based on the *rule-induction* approach to matrix problems [e.g., Little *et al.*, 2012; Ragni and Neubert, 2014], first implemented by Carpenter *et al.* [1990]. This approach focuses on identifying sets of rules describing similarities and differences between figures within matrix sequences and using these rules to drive response selection processes. This broad paradigm has encouraged work on analogical mechanisms for rule discovery [Lovett and Forbus, 2017; Rasmussen and Eliasmith, 2011], offering architectural insights for addressing the challenge of generality.

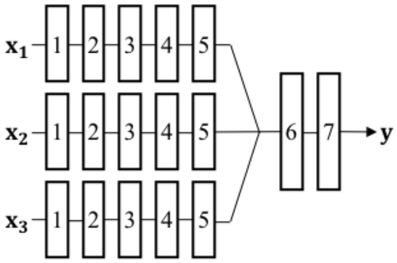

Figure 2: Neural network architecture. Each rectangle represents a layer. Layer parameters are presented in Table 3. The ith sequence figure is represented by $\mathbf{x_i}$; the output vector $\mathbf{y}$ represents feature probabilities determined by the network.

Although explicit rules and symbolic representations dominate the approaches to RPM, there is evidence that simple iconic representations (i.e., pixel arrays) can be used to solve a large proportion of RPM items without extraction of symbolic features, at least at the difficulty level of the SPM [Kunda *et al.*, 2013; McGreggor *et al.*, 2014]. The effectiveness of the iconic approach shows that symbolic representations may not be necessary to tackle matrix problems. If RPM primarily tests a single ability, it may be the case that the mechanisms supporting this ability act at a subsymbolic level. The present study explores one possible architecture for such mechanisms.

Since the experimental items used in the present study are at the difficulty level of the SPM, the model's performance can be roughly situated relative to that of existing models of SPM. The Lovett and Forbus [2017] visual analogical model solves 93.33% of matrices on the SPM, which is currently the highest reported AI performance on that test. The maximal performance achieved by our model compares well to existing models, being second only to the Lovett and Forbus model in performance on SPM-level items. Note that the Lovett and Forbus model makes use of a number of heuristics and operates on visual inputs that have been pre-segmented by an annotator. In contrast, the present model requires no pre-segmentation, and relies on a formally motivated principle of inductive inference. We believe that our model's performance and parsimony are indicative of the promise of the present approach.

5.2 Implications for Artificial Intelligence

The performance of our model depends on the quality of the feature set and the precision with which subjective probability calculations are made. These factors are important for better understanding generality in artificial intelligence, as they appear to be the only non-trivial performance bottlenecks. Indeed, further analyses (omitted due to space limitations) show that our feature set contains features that have some degree of dependency and that these dependencies are likely responsible for several errors. Other errors are attributable to the limited discriminative ability provided by a small feature set or to perceptual errors attributable to the neural network. An enlarged feature set, capable of detecting more subtle differences between figure sequences and free of feature dependencies will likely result in even better performance on the current experimental set. Taken together, these considerations suggest that learning and adaptive detection of features is an important aspect of generality.

Hoshen and Werman [2017] present further evidence of the importance of feature learning in the context of matrix problems. Their work tackles the task of guessing the final figure in a sequence following both a response selection and a response generation approach. Performance of the response generation approach is, reportedly, limited on the test set, and limited performance is attributed to discrepancies in figure complexity between training and test examples.

Instead of focusing on feature learning, the present work explored how, given a (previously acquired) feature set, subjects can carry out similarity-based (including analogical) reasoning at a subsymbolic level. This kind of reasoning is one of the most plausible mechanisms by which generality can be achieved in artificial intelligence. Models of such reasoning are often implemented using symbolic representations [Gentner and Forbus, 2011]. In these models, complex symbolic representations play a crucial role in analogy formation, as analogues are often identified by detecting structural similarities between structured representations. However, structured representations are not always readily available in the wild. Furthermore, analogical reasoning need not be contingent on the presence of structured representations. The present work suggests one way to capture analogical reasoning without relying on structured representations. The relative entropy similarity measure applied to basic and relational features appears to be a viable alternative to structure mapping. We have yet to explore the full potential of our model for explaining all forms of analogy.

Finally, we close with some speculative remarks. Our model suggests some interesting ideas for a computational theory of general intelligence. Our approach suggests integrative links between concepts of inductive inference, simplicity, similarity, and analogy. In particular, the approach suggests that these concepts can all refer to the application of the principle of relative entropy minimization in order to achieve different sets of goals. We believe these links should be further developed, perhaps starting with work on feature learning and on other analogical problems.

6 Conclusion

To conclude, the present work points to interesting possibilities for a broadly scoped and integrative similarity-based computational theory of general intelligence. A limitation of the present model is its performance. Continued development of the model will address this limitation, test the model on original RPM items, and pursue new avenues of research discussed above.

Acknowledgments

This work was supported in part by ARI grant W911NF-17-1-0236. We thank Dr. Laura E. Matzen for providing us with the Sandia Matrices and the Sandia Matrix Generation Tool [Matzen *et al.*, 2010]. We would also like to thank Çiğdem Mekik for input on editing the paper.

References

[Besold and Schmid, 2016] Tarek R. Besold and Ute Schmid. Why Generality Is Key to Human-Level Artificial Intelligence. *Advances in Cognitive Systems*, 4:13–24, 2016.

[Bringsjord and Schimanski, 2003] Selmer Bringsjord and Bettina Schimanski. What is artificial intelligence? Psychometric AI as an answer. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI'03)*, 2003.

[Carpenter *et al.*, 1990] 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. *Psychological Review*, 97(3):404–431, 1990.

[Feldman, 2016] Jacob Feldman. The simplicity principle in perception and cognition. *WIREs Cognitive Science*, 7:330–340, 2016.

[Fowlkes and Mallows, 1983] E. B. Fowlkes and C. L. Mallows. A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383):553–569, 1983.

[Gentner and Forbus, 2011] Dedre Gentner and Kenneth Forbus. Computational models of analogy. *WIREs Cognitive Science*, 2:266–276, 2011.

[Glorot and Bengio, 2010] Xavier Glorot and Yoshua Bengio. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

[Hernández-Orallo *et al.*, 2016] José Hernández-Orallo, Fernando Martínez-Plumed, Ute Schmid, Michael Siebers, and David L. Dowe. Computer Models Solving Intelligence Test Problems: Progress and Implications. *Artificial Intelligence*, 230:74–107, 2016.

[Hoshen and Werman, 2017] Dokhyam Hoshen and Michael Werman. IQ of Neural Networks. *arXiv e-prints*, September 2017.

[Kunda *et al.*, 2013] Maithilee Kunda, Kieth McGreggor, and Ashok K. Goel. A Computational Model for Solving Problems from the Raven's Progressive Matrices Intelligence Test Using Iconic Visual Representations. *Cognitive Systems Research*, 22–23:47–66, 2013.

[Little *et al.*, 2012] Daniel R. Little, Stephan Lewandowsky, and Thomas L. Griffiths. A Bayesian Model of Rule Induction in Raven's Progressive Matrices. In

Proceedings of the 34th Annual Conference of the Cognitive Science Society, 2012.

[Lovett and Forbus, 2017] Andrew Lovett and Kenneth Forbus. Modeling Visual Problem Solving As Analogical Reasoning. *Psychological Review*, 124(1):60–90, 2017.

[Matzen *et al.*, 2010] Laura E. Matzen, Zachary O. Benz, Kevin R. Dixon, Jamie Posey, James K. Kroger, and Ann E. Speed. Recreating Raven's: Software for Systematically Generating Large Numbers of Raven-like Matrix Problems with Normed Properties. *Behavior Research Methods*, 42(2):525–541, 2010.

[McGreggor *et al.*, 2014] Kieth McGreggor, Maithilee Kunda, and Ashok Goel. Fractals and Ravens. *Artificial Intelligence*, 215:1–23, 2014.

[Mekik *et al.*, 2017] Can Serif Mekik, Ron Sun, and David Yun Dai. Deep Learning of Raven's Matrices. In *Fifth Annual Conference on Advances in Cognitive Systems*, 2017.

[Nisbett *et al.*, 2012] Richard E. Nisbett, Joshua Aronson, Clancy Blair, William Dickens, James Flynn, Diane F. Halpern, and Eric Turkheimer. Intelligence: New Findings and Theoretical Developments. *American Psychologist*, 67(2):130–159, 2012.

[Pan and Yang, 2010] Sinno J. Pan and Qiang Yang. A Survey on Transfer Learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[Penrose and Raven, 1936] Lionel S. Penrose and John C. Raven. A New Series of Perceptual Tests: Preliminary Communication. *British Journal of Medical Psychology*, 16(2):97–104, 1936.

[Ragni and Neubert, 2014] Marco Ragni and Stefanie Neubert. Analyzing Raven's Intelligence Test: Cognitive Model, Demand and Complexity, In H. Prade and R. Gilles, Eds., *Computational Approaches to Analogical Reasoning: Current Trends*, Berlin, Heidelberg: Springer-Verlag, 2014.

[Rasmussen and Eliasmith, 2011] Daniel Rasmussen and Chris Eliasmith. A Neural Model of Rule Generation in Inductive Reasoning. *Topics in Cognitive Science*, 3(1):140–153, January 2011.

[Raven, 2008] John Raven. The Raven Progressive Matrices Tests: Their Theoretical Basis and Measurement Model, In J. Raven and C. J. Raven, Eds., *Uses and Abuses of Intelligence: Studies Advancing Spearman and Raven's Quest for Non-arbitrary Metrics*, pages 17–68, 2008.

[Raven *et al.*, 1998a] John Raven, John C. Raven, and J. H. Court. *Manual for Raven's Progressive Matrices and Vocabulary Scales: Section 1, General Overview*, 1998 Ed. Oxford Psychologists Press, 1998.

[Raven *et al.*, 1998b] John Raven, John C. Raven, and J. H. Court. *Manual for Raven's Progressive Matrices and Vocabulary Scales: Section 3, Standard Progressive Matrices*, 1998 Ed. Oxford Psychologists Press, 1998.

[Servan-Schreiber *et al.*, 1991] David Servan-Schreiber, Axel Cleeremans, and James L. McClelland. Graded State Machines: The Representation of Temporal Contingencies in Simple Recurrent Networks. *Machine Learning*, 7:161–193, 1991.

[Shore and Johnson, 1980] John E. Shore and Rodney W. Johnson. Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy. *IEEE Transactions on Information Theory*, 26(1):26–37, 1980.

[Spearman, 1927] Charles Spearman. *The Abilities of Man: Their Nature and Measurement*. New York, NY: The MacMillan Company, 1927.

[Sun, 1994] Ron Sun. *Integrating Rules and Connectionism for Robust Commonsense Reasoning*. New York: John Wiley & Sons, 1994.

[Sun, 1995] Ron Sun. Robust reasoning: Integrating rule-based and similarity-based reasoning. *Artificial Intelligence*, 75:241–295, 1995.

[Sun, 2016] Ron Sun. *Anatomy of the Mind: Exploring Psychological Mechanisms and Processes with the Clarion Cognitive Architecture*. New York: Oxford University Press, 2016.

[Turney, 2011] Peter D. Turney. Analogy perception applied to seven tests of word comprehension. *Journal of Experimental & Theoretical Artificial Intelligence*, 23(3):343–362, 2011.

[Tversky, 1977] Amos Tversky. Features of Similarity. *Psychological Review*, 84(4):327–352, 1977.

[Zeiler, 2012] Matthew D. Zeiler. ADADELTA: An adaptive learning rate method. *arXiv e-prints*, 2012.