Abstract Visual Reasoning

Dhaivat Patel
Rutgers University
dp1081@rutgers.edu

Neel Doshi
Rutgers University
nsd77@rutgers.edu

Shantanu Shinde
Rutgers University
ss3670@rutgers.edu

Shreyas Kulkarni
Rutgers University
spk111@rutgers.edu

Abstract

Abstract reasoning, also known as abstract thinking, is the ability to understand and think with complex concepts that, while real, are not tied to concrete experiences, or memorized situations. This type of reasoning involves higher-order thinking about ideas and principles that are often symbolic or hypothetical. This type of thinking is more complex than thinking that is centered on memorizing and recalling information and facts. Humans do abstract thinking naturally without any effort, But for a Machine learning model this can be a challenging task.

We are working on Raven's Progressive Matrices (RPM) problem. We are using I-RAVEN dataset We are using two models for the project. For our baseline model, We are planning to train our model by using a CNN network for image feature extraction and then feeding those features in LSTM network. We will implement WReN to successfully make a prediction of the target. For our proposed model we are using Scale localized abstract reasoning's slightly modified version. To reduce the computational complexity of the Scale localized reasoning model. We found that proposed model in paper is extremely inefficient compared to other reasoning architectures. We proposed reducing number of levels used in the MRNET in Scale localized abstract reasoning. We also analyzed and experemented with different fusion methods for Scale localized abstract reasoning and implemented different fusion method from proposed in the paper, which was able to generate better accuracy.

1. Introduction

We are training the model in Raven's progressive Matrices as shown in Figure 1. In this problem we are provided with 3×3 figures as shown in 1 consisting of simple elements. Our goal here is to predict correct image structure for the blank space (denoted by '?') which can satisfy the problem best way possible.

The Raven's matrix [14] denoted as $\mathcal{M}$, of size 3×3 contains images at all i,j location except at $\mathcal{M}_{33}$. The aim is to find the best fit image a at $\mathcal{M}_{33}$. We assume the rules are prepared row-wise. To fix the defects of RAVEN dataset, which resulted into model being able to predict output without even looking at the input. We are planning to generate an alternative answer set for each RPM question in RAVEN, forming an improved dataset named I-RAVEN matrix [8]. I-RAVEN using attribute bisection tree avoids such imperfections. On top of that, Extensive experiments conducted on PGM and I-RAVEN datasets, It is shown in the paper of Stratified rule- aware network for abstract visual reasoning [8] that compared with RAVEN, I-RAVEN is more rigorous and fair for evaluating abstract reasoning capability.

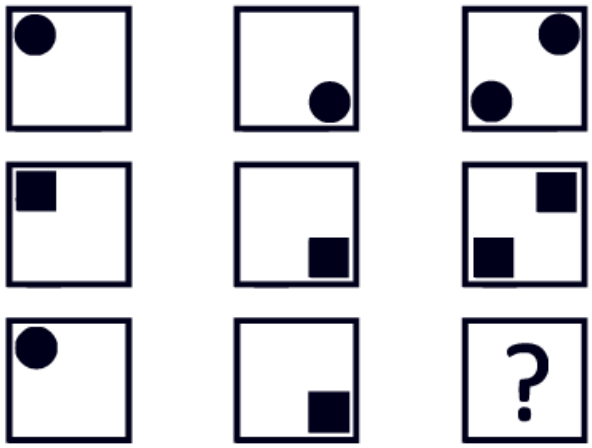

Figure 1. Abstract Reasoning Example

In our baseline model, the model structure can be divided in to 3 parts: CNN [7], LSTM [5] and WReN [13]. Firs in CNN, we used the images generated using I-RAVEN [8] in $\mathcal{M}_{33}$ matrix. This matrix further was fed in to a deep neural network of CNN [7], ResNet [6]. By doing this, we extracted representations of each image. After that we formulated the reasoning based module to learn the row rules. Thus we extracted each image feature from CNN [7]. Then in LSTM, These extracted image features are fed in to an LSTM network [5] sequentially where feedback connec-

tions are used to understand relations. So, the model will become a representation of sequential data instead of just flattend data of all the image representations. In our WReN [13], we used the same architecture as CNN. Each answer feature is paired with each image feature to form a set of ordered pairs. We apply panel embeddings to all input images. This is then fed into a relation module which generates probability score for each answer possibility.

In our propsped modified version of Scale localized abstract reasoning[1], We tried to reduce the number of layers present in MRNET architecture of Scale localized abstract reasoning[1]. The paper also does this but it does this without multihead loss. We do this with multihead loss. We found that multihead loss is vital for model performance and we were able to acheive better results with multihead loss. Because solving Raven's progressive matrices [14] require reasoning across rows and colums. We also augmented rowwise and columnwise features and then evaluated model performance unlike the original paper of Scale localized abstract reasoning. [1]

2. Prior Work

is a non-verbal test typically used to measure general human intelligence and abstract reasoning and is regarded as a non-verbal estimate of fluid intelligence. Its popularity is mainly attributed to its independence from language and writing skills. It has previously been proposed and used by [12], [3], [10]

[8] points out some severe defects in the RAVEN dataset which prevent from the fair evaluation of the abstract reasoning ability. It suggests an answer set generation algorithm called Attribute Bisection Tree (ABT) to form a better dataset which is named Impartial-RAVEN (I-RAVEN) dataset.

[8], [4], [14] all propose a neural network based classifier approach to perform abstract visual reasoning. [8] suggests a Stratified Rule-Aware Network to generate rule embeddings.

[7] makes use of Convolutional Neural Networks (CNN) for solving geometric pattern recognition problems similar to RPM. The CNN receives as input multiple ordered input images and outputs the next image according to the pattern. [9] introduces a reinforcement based model training to overcome limitations of neural network based models on RPM test.

[13] explores the performance of different kinds of models on the RPM test. This analysis is useful to understand the performance and results of Long Short Term Memory and WReN.

[11], [2] makes use of visual reasoning as an inspiration to perform next frame, action and prediction of objects in videos.

3. Theory

To propose a novel model for performing Abstract Visual Reasoning we first explore the prior models that have been implemented for such tasks.

We use the Ravens Progressive Matrices test with changes proposed to the dataset by [8] to generate the I-RAVEN dataset to evaluate the performance of each model being explored. We also use the same tests on our own proposed novel model to perform comparison.

We explored the methods a CNN [7] and ResNet [6] based architecture, an LSTM [5] architecture and the WReN architecture proposed by [13].

All Abstract Visual Reasoning models that attempt to solve the RPM test can be broadly divided into two types based on how the models view the choice images of the RPM test.

The first type of models under this classification is the one that views all choice images one at a time and assigns scores (likelihoods) of the choice image being the correct answer. The image with the highest score is generally chosen as an output. This type of method is categorized as the Single Choice(SC) method.

The second type of models takes all the choice images at once and tries to choose the correct answer choice from these. Here, the model has access to all the choice images at the same time and thus it can make an "informed" decision while choosing the output. This kind of method is categorized as Multiple Choice(MC) method.

Both these methods have their advantages and disadvantages. In the multiple choice method, since the model has access to all the choice images, it can come up with a solution choice image just by looking at the choice images and ignoring the Problem Matrix altogether. This is possible in the RPM test because all the choice images in the RAVEN dataset at generated by making some changes to the correct choice image. So, a smart Abstract Reasoning model can choose the image that has the most common features with all the other choice images. The model will look for patterns in the choice images while ignoring the Problem Matrix. This type of problem however is avoided in the Single Choice method.

The Single Choice method however makes it difficult for the model to perform reasoning as it has access to only one kind of image at a time. This leads to generally bigger models requiring higher compute and can sometimes be inefficient.

To overcome the problems faced in the Multiple Choice type of models, changes have been suggested to the RAVEN dataset. [1] proposes the RAVEN-FAIR dataset which iteratively enlarges the choice image set by making changes to any of the randomly selected image from the current set of choice images.

Similar changes were proposed by [8] in their paper on

the SRAN model to make the I-RAVEN dataset. It is observed that both these dataset have better chances of leading to better reasoning capacities in the Multiple Choice types of models.

3.1. MRNet

[1] proposes the MRNet model which looks at the input images in different resolutions to perform reasoning. MRNet is a Single Choice type of model that assigns scores to each of the choice images after evaluating them independently.

This model searches for relational patterns in different resolutions, where the higher resolution section is able to detect visual relations such as position and patterns while the lower resolution is able to detect semantic relations such as shapes and sizes.

This model is divided into 3 major sections. The first section is the Multi Scale section that encodes the input images into 3 different resolutions. These encoded images are then passed to a pattern module that searches for patterns along the rows and columns of the input Problem Matrix. The output of the pattern module are 6 vectors, one for each row and column. The 3 column vectors are passed through a Fusion function all at once to create a new vector. Similarly, the 3 row vectors are also passed through the same Fusion function. In the end, the 2 output vectors are added with element wise addition before being passed through a Multi Layer Perceptron model which performs the reasoning tasks.

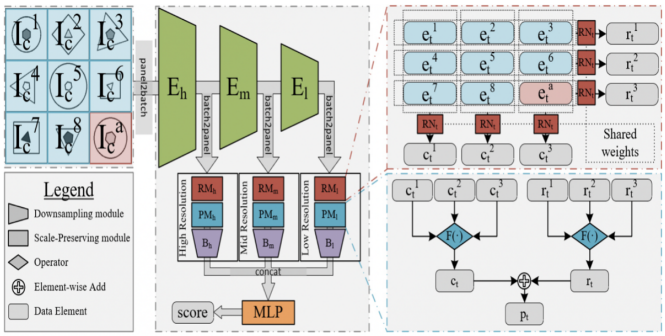

Figure 2. MRNET

3.2. Proposed Model

In order to improve on the performance of the already explored models on the RPM test using the I-RAVEN dataset, namely the ResNet, LSTM, WReN and MRNet, we propose the following changes to the MRNet model which was observed to outperform all the other models.

Firstly, solving RPM's does require reasoning across rows and columns which is missing in current architecture. In the proposed model, there is a change to augment row-wise and columnwise features and then evaluate model performance which means instead of passing all the column vectors and the row vectors of the pattern module in the MRNet directly to the Fusion function, we passed all the combinations of the column vectors, 2 at a time, to the Fusion function. This would mean that instead of the Fusion function being applied only once for the column vectors, it is now used 3 times. A similar computation was performed for the row vectors as well. This will lead to better reasoning and understanding of the patterns between each row and column.

Secondly, it was observed that the outputs of the Fusion function of both the row and the column vectors were added in a element wise addition. This could lead to loss of valuable information before being passed to the MLP module. We instead concatenated the output of each of the 6 Fusion functions, 3 for column and 3 for rows. This concatenated vector was passed as an input to the MLP module to perform better reasoning.

4. Evaluation

4.1. Dataset

To examine the capability of abstract reasoning, we are performing Raven's Progressive Matrices (RPM) test [14]. The Raven's Progressive Matrix [14] looks as shown in 3. In the problem of Raven's Progressive Matrices the model is presented with the first eight images of a 3x3 grid of images called problem matrix, and another eight images are called Answer Set. The objective is to choose the missing ninth image of the grid out of the eight presented choices, by identifying the pattern along the rows and/or columns of the grid. The correct answer is the one that fits the most patterns. For our example shown in Figure 3 the answer will be Image number 7 in the answer set. The reasoning for the example is As you can see in Figure 3 in each row, inner elements are rotated 90 degrees clockwise compared to the elements in the left side in each row. Using that logic We can predict that our answer will be the image labelled '7'.

For both of our proposed model we used to generate the dataset of Raven's Progressive Matrices (RPM) problem [14], we are using I-RAVEN [8]. I-RAVEN [8] Attribute Bisection Tree (ABT) to generate an impartial answer set for any attribute-based Raven's Progressive Matrices question. The Attribute Bisection Tree ensures attribute modifications among the answer set are well balanced. Thus, no clue can be found to guess the correct answer only depending on the answer set, and no distractor can be eliminated without reasoning from the context matrix as well.

Each node in Attribute Bisection Tree indicates a multiple-choice panel, and the root of the tree structure is the correct answer. Different levels of the tree indicate different iterations, where nodes of this level are the candidate

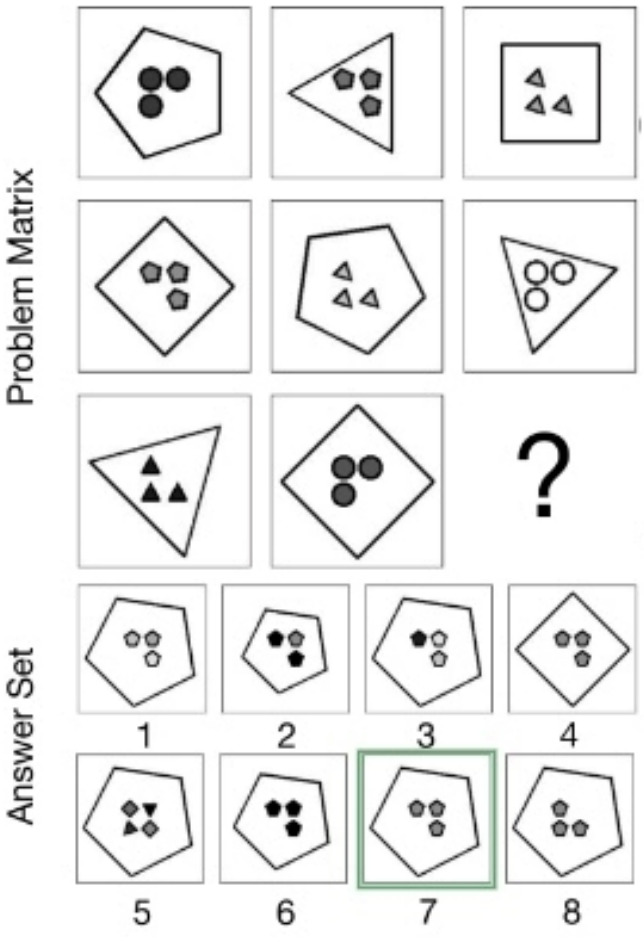

Figure 3. Raven Matrix example

answers of current answer set. The generation process for I-RAVEN's [8] Raven's Progressive Matrices flows in a top-down manner. For each iteration, only one attribute will be modified. At each level, In Attribute Bisection Tree, the a node has two children nodes, where one node remains the same with the father node, the other changes the value of the attribute sampled for this iteration of the father node. Finally, at the bottom level, we could obtain the whole answer set.

To test our model performance, The dataset contains four different configurations (Center Single, 2x2 Grid, Left-Right, Out-In Grid). We generated 10,000 sample matrices of each configuration with 60:20:20 Train:Validation:Test split. We generated these datasets using I-RAVEN proposed in the paper of Stratified rule-aware network for abstract visual reasoning. [8]

4.2. Evaluation Strategy

We evaluated our baseline model by measuring overall accuracy for test dataset on ResNet [6], CNN-MLP [7], LSTM [5] and WReN [13] to determine the overall model performance. We did this for all four configurations (Center Single, 2x2 Grid, Left-Right, Out-In Grid) of I-RAVEN dataset. [8] In addition to that we also measured model accuracy in the interval of 20 epochs for all four configurations. We also plotted the graph of model's accuracy vs epochs for testing dataset.

To evaluate our proposed model, which uses slightly modified version of MRNET architecture proposed in the original paper of Scale localized abstract reasoning. [1]. We plotted a comparison graph of total accuracy against epochs for our proposed model and base MRNET model on I-Raven [8] dataset. We also did an ablation study against our proposed model, original MRNET proposed in the paper [1] and model with reduced resolution and plotted the graph of accuracy against epochs. We also mesured overall accuracy on test dataset for 25 epoches for all four configurations of (Center Single, 2x2 Grid, Left-Right, Out-In Grid) of I-RAVEN dataset and time taken per epoch to mesure model efficiency.

5. Results

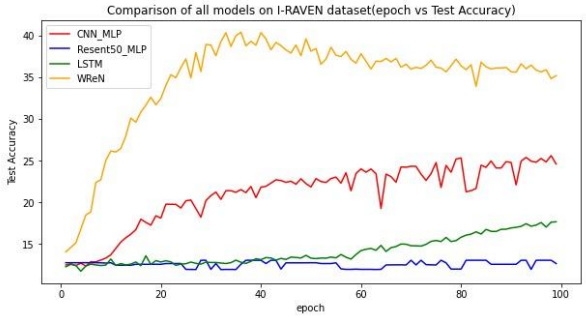

Figure 4. Performance of Baseline Model on I-RAVEN dataset.

In the above Figure 4, the comparison of baseline models on I-RAVEN dataset [8] is shown. WReN [13] performs the best compared to other baselines. Furthermore, accuracies on each configuration is shown in below table.

The configuration wise results, shown in Table 1 which indicate center_single which is considered easier to solve has good accuracy compared to other architectures. One thing about I-RAVEN though is that I-RAVEN still struggles to generate consistent dataset which results into varied performance of model across the configurations. As the Resnet accuracy is low as compared to what author have in the paper, the results were surprising in which data was generated 3 times and still the accuracy was not increasing even when all other models give accuracy stated in the paper.

Configuration	CNNMLP	ResNet	LSTM	WReN
center_single	27.70	13.04	21.29	52.87
distribute_four	17.89	13.24	13.48	20.65
in_distribute_four_out center_single	17.24	13.17	14.96	16.11
left_center _single_right _center_single	24.88	13.05	13.91	17.06
overall	24.61	13.05	17.68	35.23

Table 1. Overall accuracy of baseline model for all four configuration in percentage(%).

Datasize	CNNMLP	ResNet	LSTM	WReN
20	13.47	12.125	13.56	15.0625
40	20.53	12.81	13.15	15.46
60	23.62	12.5	13.54	15.46
80	23.95	13.02	13.85	21.67
100	24.61	13.05	17.68	35.23

The ablation study shows how increase in datasize affects the model performance.

Table 2. Accuracy change for different epochs in percentage(%).

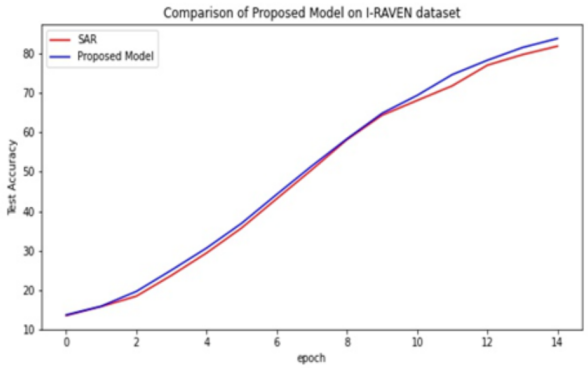

Figure 5. Comparison of Proposed Model on I-RAVEN dataset

In the above Figure 8, the red line shows the test accuracy to epoch plot for the MRNet model [1] which gives about 81.9% on epoch 15 and the blue line shows the proposed model accuracy which is 83.45% at epoch 15. Solving RPMs does require reasoning across rows and columns and this is what changed from the first model.

While doing an analysis of the data, one of the points to keep in mind is that it takes almost 1250 seconds for 1 epoch to run. So, we tried reducing the resolution in the network. After changing the level to '101' instead of '111', we got less accuracy as seen in the plot 6 above with the green line. Each epoch was taking almost 780 seconds to run which is considerably less than the above 2 models. As

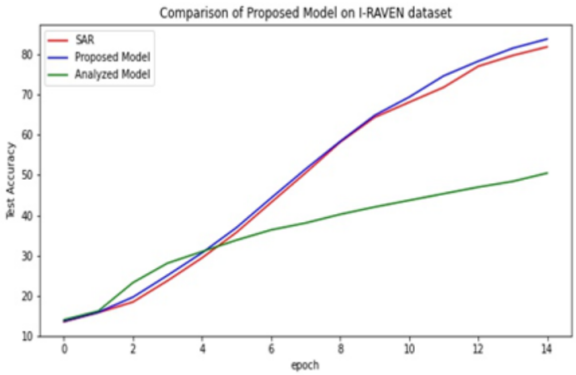

Figure 6. Analysis of Model on I-RAVEN dataset

accuracy is really low as compared to the other 2 models, there is no point in trying this approach.

Epoch(25)	Test Accuracy	Time taken for each epoch
center_single	99.77	312
distribute_four	98.39	312
in_distribute_four _out_center_single	96.36	312
left_center_single _right_center_single	97.60	312
overall	96.65	1250

Table 3. Overall accuracy of proposed model for all four configuration.

The Table 3 shows all the accuracies for the proposed model. For each configuration, each epoch was taking 312 seconds while for the overall dataset each epoch was taking 1250 seconds. As the computational time was more, the model was trained for 25 epochs for each configuration and on the whole dataset. The model is giving 96.65% accuracy on the dataset.

6. Conclusion

As ResNet is giving less accuracy compared to the paper, the I-RAVEN dataset still struggles to generate consistent data which results into varied performance of models. The MRNET is performing better than all the baseline models and with fusion changes in the network model performs even better than the original model. The changes like using two resolutions instead of one reduces the time of implementation but decreases the accuracy considerably.

7. Project Code

Code for baseline models and for newly developed MRNET based model can be found on the following Github repository link: Repository Link

8. Contribution

TASK	DONE BY
Project Step 1	
Code	Shreyas
Step 1 Report	Dhaivat, Shantanu, Neel
Project Step 2	
Code	Shreyas
Step 2 Report	Dhaivat, Shantanu, Neel
Project Step 3	
Code	Shreyas
Project Proposal	Dhaivat, Shantanu, Neel
Final Project Code	Shreyas
Final Report	Dhaivat, Shreyas Shantanu, Neel

References

[1] Y. Benny, N. Pekar, and L. Wolf. Scale-localized abstract reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12557–12565, 2021. 2, 3, 4, 5

[2] A. T. Carl Vondrick, Hamed Pirsiavash. Anticipating visual representations from unlabeled video. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 98–106, 2016. 2

[3] J. M. A. . S. P. Carpenter, P. A. What one intelligence test measures: A theoretical account of the processing in the raven progressive matrices test. *Psychological Review*, 97(3):404–431, 1990. 2

[4] D. G. B. A. S. M. F. Hill, A. Santoro and T. Lillicrap. Learning to make analogies by contrasting abstract relational structure. *arXiv preprint arXiv:1902.00120*, 2019. 2

[5] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2016. 1, 2, 4

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2, 4

[7] D. Hoshen and M. Werman. Iq of neural networks. *arXiv preprint arXiv:1710.01692*, 2017. 1, 2, 4

[8] S. Hu, Y. Ma, X. Liu, Y. Wei, and S. Bai. Stratified rule-aware network for abstract visual reasoning. *arXiv preprint arXiv:2002.06838*, 2020. 1, 2, 3, 4

[9] W. W. Kecheng Zheng, Zheng-Jun Zha. Abstract reasoning with distracting features. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019*, 2019. 2

[10] M.-K. . G. A. K. Kunda, M. A computational model for solving problems from the raven's progressive matrices intelligence test using iconic visual representations. *Cognitive Systems Research*, pages 47–66, 2013. 2

[11] Y. L. Michael Mathieu, Camille Couprie. Deep multi-scale video prediction beyond mean square error. *arXiv:1511.05440*, 2016. 2

[12] J. C. Raven and J. H. C. John Raven. Raven's progressive matrices. *Western Psychological Services.*, 1938. 2

[13] A. Santoro, F. Hill, D. Barrett, A. Morcos, and T. Lillicrap. Measuring abstract reasoning in neural networks. In *International Conference on Machine Learning*, pages 4477–4486, 2018. 1, 2, 4

[14] C. Zhang, F. Gao, B. Jia, Y. Zhu, and S.-C. Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5317–5327, 2019. 1, 2, 3