Residual Attention Network for Abstract Visual Reasoning

Anonymous submission

Abstract

Abstract visual reasoning is a field that identifies and applies patterns between entities, an ability required to build artificial general intelligence (AGI). Existing deep learning methods dealing with Raven's Progressive Matrices (RPM) have mostly struggled to determine how to aggregate and process the information of independently recognized panels, which limits their ability to understand the relationships between the panels. Taking inspiration from the Gestalt principles, which govern what humans attend to when perceiving multiple objects, we enable the model to perceive panels holistically and gain a deeper understanding of the overall context. Specifically, we propose the Residual Attention Network (RANet), which uses residuals between panels to perceive differences among them explicitly. The attention map obtained from residuals augments the panel feature map, aiding the model in understanding and reasoning. Experiments demonstrated the state-of-the-art performance of RANet on various RPM benchmarks and good generalization performance on out-of-distribution settings, showing its ability to effectively capture relationships between panels with a minimal number of parameters.

Introduction

Abstract reasoning capability, the ability to perceive differences between entities and then apply them to new entities, is a hallmark of human intelligence. Raven's Progressive Matrices (RPM) is a test to estimate human abstract visual reasoning capability (Bilker et al. 2012; Raven 1941). Figure 1 shows the examples of RPM problems. Solvers have to perceive and apply patterns between panels. In the computer vision research field, creating Artificial General Intelligence (AGI) requires models to possess abstract visual reasoning capabilities that surpass those of humans (Małkiński and Mańdziuk 2023). Therefore, RPM is also a challenge that needs to be addressed in deep learning models.

Several deep learning models have been proposed to solve the RPM problem by approaching it as a classification task (Mondal, Webb, and Cohen 2023; Zhang et al. 2019b; Zhuo and Kankanhalli 2021), an approach that we follow as well. They mainly focus on recognizing each panel independently using Convolutional Neural Networks (CNNs) (recognition stage) and then on aggregating and processing the information from each panel to effectively find shared rules (reasoning stage), as shown in Figure 2(a). Specifically, they convert the panel feature map into a feature vector. These feature vectors are then concatenated to form rows, columns, or entire matrices, which are fed into the reasoning module of the model to find shared rules and solve the problem. However, finding relationships between panels solely in the form of feature vectors could limit the ability to understand the overall context (Kobayashi 2016).

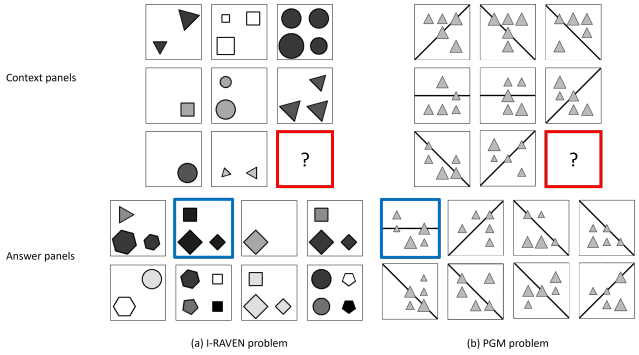

Figure 1: Examples of the I-RAVEN and PGM problems. Each RPM problem has eight context and answer panels each. Context panels comprise a 3×3 matrix. A matrix has several "shared rules" across rows or columns, and the rule exhibiting a compositional nature is defined with the object-attribute-relation form. The solver performs the task of filling each answer panel into the missing panel(red) to find one correct answer panel(blue) that makes the last row or column satisfy the shared rule.

Motivated by the Gestalt principles, we propose a method to identify the relationships between panels not only as feature vectors but also as feature maps. The Gestalt principles in psychology suggest that when humans recognize multiple objects they identify which parts to attend to, making some objects the foreground and others the background and thus inputting information in a much easier-to-understand form (Alawadhi 2010). Our model attends to specific parts of the feature maps from multiple panels, allowing the information from each panel to be processed into a form that is much easier for the reasoning module to understand, as shown in Figure 2(b).

We propose the Residual Attention Network (RANet), which is aware of other panels forming rows or columns when recognizing a panel, enabling it to utilize the overall context among panels. The core idea of RANet is to use residuals between panels, enabling it to explicitly learn the differences between panels, thus making it easier for the model to reason their relationships. RANet has three components: CNN module, residual attention module (RA module), and reasoning module. The CNN module extracts the feature maps for the panels. The RA module calculates the residuals with the preceding and following panels for the feature map of each panel and then creates attention maps to augment feature maps. The reasoning module takes as input a set of augmented feature vectors that form two rows or columns, finds the shared rules of the pairs of rows/columns, and makes total predictions based on the similarity between the shared rules.

RANet shows state-of-the-art performance on the various RPM benchmarks and good generalization performance on out-of-distribution settings, while using fewer parameters than previous methods. The contributions of this study are as follows:

- Instead of recognizing each panel independently, we first propose a method that uses Gestalt principles to find relationships between feature maps during the recognition stage, making reasoning easier.
- By visualizing the residual attention map and evaluating the quality of the augmented feature map, we show that RA module actually aids model in reasoning.
- By demonstrating that the RA module also aids in the reasoning process of other models, we show that the proposed method could be broadly beneficial in the field of abstract visual reasoning.

Related Work

Deep learning methods for solving RPM problems

To measure the abstract visual reasoning capability of a model's, RPM datasets such as PGM (Barrett et al. 2018), RAVEN (Zhang et al. 2019a), I-RAVEN (Hu et al. 2021), and RAVEN-FAIR(Benny, Pekar, and Wolf 2021) have been presented, enabling the model to learn and infer the problems. Various methods using the datasets have been proposed, achieving advancement in this research area. SRAN (Hu et al. 2021), MRNet (Benny, Pekar, and Wolf 2021), and ARII (Zhang et al. 2022) solve RPM problems using the similarity of rows or columns in the matrix. Moreover, WReN (Barrett et al. 2018) SCL (Yuhuai et al. 2020), PredRNet (Yang et al. 2023), and DRNet (Zhao, Xu, and Si 2024) predict the answer panel by evaluating the correctness of the entire matrix. In particular, PredRNet and DRNet are inspired by the human system, similar to our model. However, these methods recognize each panel independently and process information in the form of feature vectors. The proposed method not only uses the similarity of rows or columns with panels, but also considers adjacent panels forming the same row or column when perceiving them.

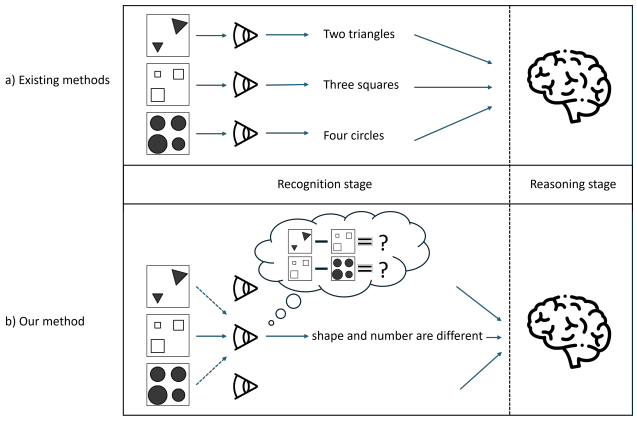

Figure 2: Differences between existing methods and our method in terms of the process of recognizing three panels that form a single row or column.

An approach similar to ours is SRAN, whichsets the input channels to match the number of panels, allowing the CNN encoder to process multiple panels simultaneously. However, it does not explicitly learn the differences between the panels, leading to underwhelming performance. Moreover, it relies on a large number of parameters to learn relationships in the form of feature maps.

Gestalt principles and deep learning

Gestalt principles describe how people tend to perceive the overall structure and pattern of multiple objects rather than individual objects. It suggests that people recognize multiple objects at once; they attend more to specific parts, dividing them into foreground and background. Studies such as (Uhlhaas et al. 2006) and (Zaretskaya and Bartels 2015) have demonstrated that this phenomenon is related to actual brain activity.

There have been attempts to apply Gestalt principle to deep learning models. For example, (Amanatiadis, Kaburlasos, and Kosmatopoulos 2018) demonstrated that the CNN mechanism is similar to the mechanism of the of Gestalt principles. (Song et al. 2022) explicitly incorporated this concept into their architecture in few-shot learning scenario. Moreover, (Hua and Kunda 2019) endeavored to integrate the concept of Gestalt principles with a model for abstract visual reasoning. They employed a VAE-GAN (Larsen et al. 2016) model inspired by the human ability to automatically fill in missing parts of an image, using it to impaint the missing panel of the matrix. Our work is also inspired by the Gestalt principles, ensuring that when recognizing a single panel, the adjacent panels are also considered. Specifically, we used the residuals of the feature maps of panels forming rows or columns, tailored to the RPM task where recognizing differences between patterns of panels is crucial.

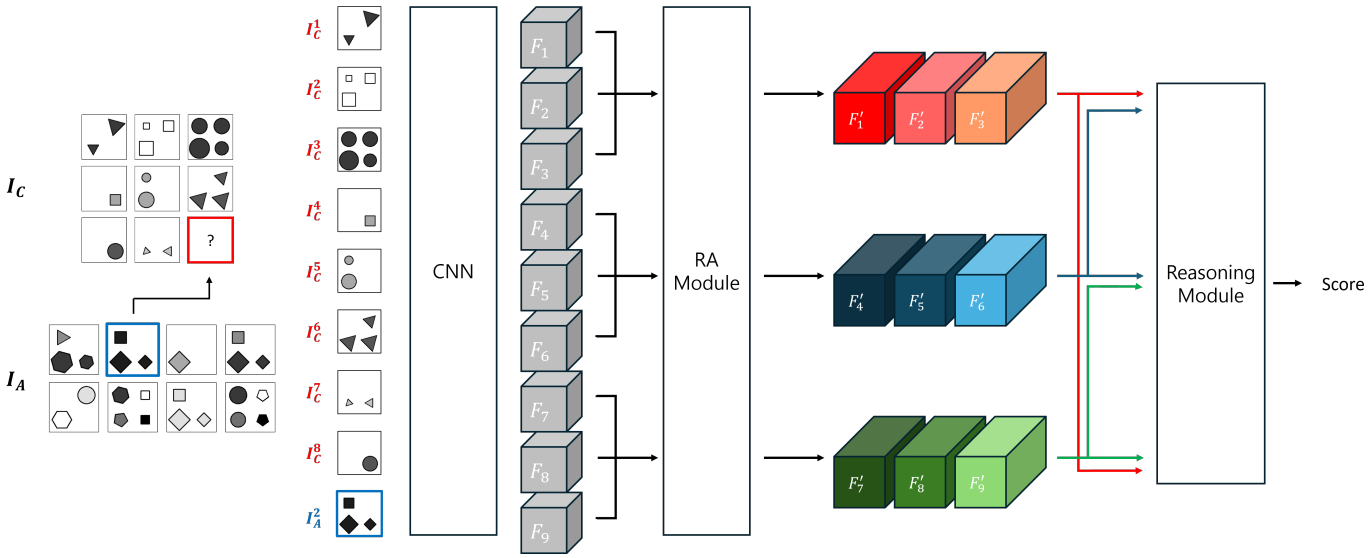

Figure 3: Overall architecture of RANet. The RA module is used to holistically recognize feature maps of panels that form a row or column.

Method

Problem definition

A single RPM problem has eight context and eight answer panels each. The eight context panels form a 3×3 matrix, whose bottom-right corner is a missing place. Solvers have to find one answer panel that satisfies the shared rule by filling the missing place. In particular, let all images of a single RPM problem $I \in R^{16\times1\times H\times W}$. It comprises of context panels $\{I_C^k\}_{k=1}^8$ and answer panels $\{I_A^l\}_{l=1}^8$, $I = I_C \cup I_A$. In the following section, we will consistently use the term "row" instead of "row or column" for clarity.

Overview

RANet is a model that enables the learning of relationships between panels even during the recognition stage. Eight full matrices can be created from input panels, filling each answer panel into missing place iteratively. This operation can be represented by rhe following formula:

$$\{M^l\}_{l=1}^8 = I_C \cup I_A^l, M^l \in R^{9\times1\times H\times W}. \quad (1)$$

RANet performs a classification task where it takes a full matrix M^l, obtained by filling in one answer panel, and outputs the probability p^l that the answer panel is correct.

Figure 3 illustrates the overall architecture of RANet. It comprises three modules: CNN module, RA module, and reasoning module. The CNN module first creates the feature map for each panel. The RA module augments the feature vector of each panel by using residuals between adjacent panels. The reasoning module extracts the common rules between a pair of rows and then calculates probabilities. The probability of a single answer panel is calculated using the similarity between common rules. In the following section, we will only consider cases where the model outputs the probability for a matrix that includes a single answer panel. At the end of the RA module and the reasoning module, a feedforward network following (Vaswani et al. 2017) is added.

CNN module

The CNN module extracts the feature map of each panel. It is based on the ResNet (He et al. 2016) architecture. Specifically, the CNN module consists of 10 layers, with an additional 1×1 convolutional layer added to the first nine layers of ResNet-18. The 1×1 convolutional layer reduces the feature map dimension from 128 to d. Then the total stride of CNN module is 8. Considering the feature map of a single matrix F, this operation can be expressed as follows:

$$F = \text{CNN}(M), F \in R^{9\times d\times \frac{H}{8}\times \frac{W}{8}}. \quad (2)$$

Residual Attention module

The RA module is motivated by the Gestalt principles, augmenting the feature map of a panel by deciding where to attend to from other panels that form the same row. With this module, the model recognizes the feature map in a form that is much easier for reasoning. We will describe the specific operation of the RA module using equations, taking only the first row as shown in top image of Figure 4. The RA module takes the feature maps of the three panels F_1, F_2, F_3 that form a row as input and produces the feature vectors $\boldsymbol{F_1'}$, $\boldsymbol{F_2'}$, and $\boldsymbol{F_3'}$ for the three panels are generated as the output.

The specific operation of the RA module is as follows: First, for each panel, it concatenates the absolute residual between its feature map and the feature maps of the former and later panels. The absolute value of the residual is introduced to make it easier for the model to understand differences regardless of the order of the panels. From the perspective of

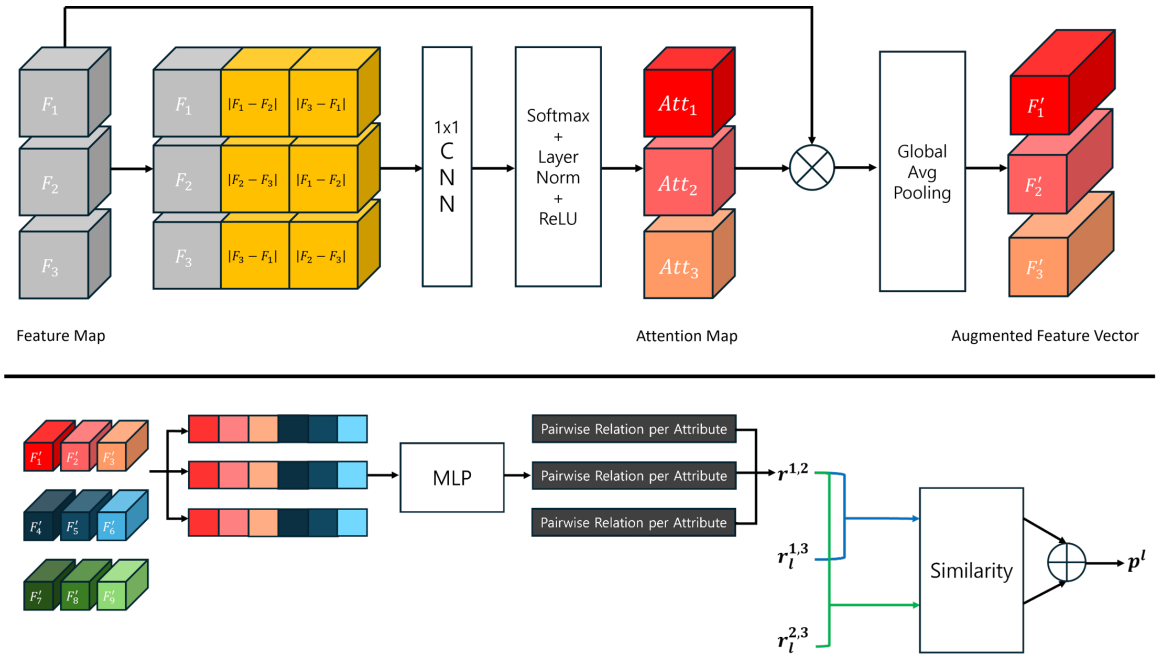

Figure 4: Top: RA module, bottom: Reasoning module.

the first panel, this operation can be represented by the following equation:

$$Cat_1 = \text{Concat}\left[F_1; |F_1 - F_2|; |F_3 - F_1|, dim = 1\right]. \quad (3)$$

Next, the concatenated feature map undergoes dimension reduction by a 1×1 convolutional layer, resulting in the same size as that of the original panel's feature map. Then, softmax is applied along the attribute dimension to emphasize differences between the feature maps of panels forming a row. After applying layer normalization (Ba et al. 2016) for the feature map size and the ReLU activation function, residual attention maps are obtained. These residual attention maps are multiplied element-wise with the feature map of each panel, and global average pooling produces the panel-specific feature vector. From the perspective of the first panel, the overall process can be described using following equations:

$$Embed_1 = \text{Conv}\left(Cat_1\right), \quad (4)$$

$$Embed_1 = \text{Softmax}\left(Embed_1, \text{dim} = 1\right), \quad (5)$$

$$Att_1 = \text{ReLU}\left(\text{LayerNorm}\left(Embed_1\right)\right), \quad (6)$$

$$Aug_1 = F_1 \odot Att_1, \quad (7)$$

$$\boldsymbol{F_1^{'}} = \text{GAP}\left(Aug_1\right), \boldsymbol{F^{'}} \in R^{32}. \quad (8)$$

Reasoning module

The reasoning module first takes the concatenated feature vectors of the panels forming a row as input and outputs the rule embedding that the row satisfies. Then, it calculates the probability that a single answer panel is correct. Instead of finding a rule for a single row, it takes a pair of rows as input to find their shared rule, using the feature vectors of six panels. This approach addresses the distracting features in the RPM problem, where attributes of the objects in the panels that do not satisfy the shared rule can take arbitrary values, making it difficult for the model to identify the shared rule. By finding the common rule between two rows, the influence of distracting features is reduced.

The reasoning module includes several linear layers. However, by using a scattering operation, it can efficiently find the rule embedding from the concatenated vector. This method was devised in SCL and has also been used in ARII, contributing to its powerful performance. The mechanism of the scattering operation is illustrated in Figure 4, where the vector is transposed and fed into a MLP, allowing the module to find common attributes among the panels.

In the case of where the reasoning module takes the first two rows of the RPM problem, the operation can be represented by the following equations:

$$F_{cat} = F^{'}_{k=1\ldots6}, \quad (9)$$

$$F^T_{cat} = \text{Transpose}(F_{cat}), \quad (10)$$

$$\boldsymbol{r^{1,2}} = \text{Concat}(\text{MLP}(F^T_{cat})). \quad (11)$$

By applying the aforementioned process three times, the outputs of the reasoning module for a single RPM problem are $\boldsymbol{r^{1,2}}$, $\boldsymbol{r^{1,3}}$, and $\boldsymbol{r^{2,3}}$. Here, $\boldsymbol{r^{1,2}}$ is derived only from the context panels and can be referred to as the ground truth shared rule. $\boldsymbol{r^{1,3}}$ and $\boldsymbol{r^{2,3}}$ include the answer panel, then their high similarity to the ground truth shared rule indicates that the answer panel is correct. Therefore, the probability for a single answer panel is calculated as follows.

$$p^l = \frac{\text{Sim}\left(\boldsymbol{r^{1,2}}, \boldsymbol{r_l^{1,3}}\right) + \text{Sim}\left(\boldsymbol{r^{1,2}}, \boldsymbol{r_l^{2,3}}\right)}{2}. \quad (12)$$

Method	WReN	SRAN	MRNet	ARII	SCL	PredRNet	DRNet	RANet
PGM-N	62.6	71.3	94.5	88.0	88.9	97.4	**99.0**	98.0
RAVEN	16.8	54.3†	96.6	-	91.6	95.8	**96.9**	96.2
I-RAVEN	23.8	60.8	83.5†	91.1	95.0	96.5	97.6	**98.7**
RAVEN-F	30.3	72.9†	88.4	-	90.1†	97.1	97.6	**97.9**
Average	33.4	64.8	90.8	-	91.4	96.7	**97.8**	97.7
# of param	1.2M	44.0M	19.5M	91.2M	**0.1M**	1.28M	24.6M	0.76M

Table 1: Overall accuracy (%) on i.i.d datasets and the number of parameters of RPM models. Here, PGM-N refers to the neutral regime in the PGM dataset, and RAVEN-F refers to the RAVEN-FAIR dataset. † indicates that the performance was not reported in the original paper; we obtained these results from Table 1 of (Yang et al. 2023).

We used dot product for the similarity function.

Training and inference

The loss function of the training model is as follows:

$$L_{total} = L_{CE} + \alpha \times L_{att}, \tag{13}$$

,where L_{CE} denotes the cross-entropy loss between the probabilities of eight answer panels and one-hot encoded label, and L_{att} represents the sum of residual attention maps. As in (Mascharka et al. 2018), we applied regularization to the attention map to ensure that the residual attention map focused only on the necessary parts. α is a balance term between two losses. When the model infers, it predicts the answer panel having the highest probability among the eight answer panels as correct.

Experiments

Datasets

To evaluate the abstract visual reasoning capability of RANet, we used four different RPM datasets: RAVEN, I-RAVEN, RAVEN-FAIR, and PGM. All datasets are composed of eight context panel and eight answer panels each. In addition to the correct labels, they also provide encoded information about the shared rule, facilitating the model's learning and analysis.

RAVEN-style datasets. In the problem of the RAVEN dataset (Zhang et al. 2019a), rules are shared across rows. Additionally, objects in the panel are arranged differently according to the configuration, enabling the model to develop reasoning abilities under diverse conditions. However, the answer panels in the problem of the RAVEN dataset are biased, which could lead to shortcut learning by the model. To fix this issue, the I-RAVEN (Hu et al. 2021) and RAVEN-FAIR (Benny, Pekar, and Wolf 2021) datasets were proposed. Each dataset includes seven configurations, each containing 6K, 2K, 2K training, validation, and test samples, respectively.

PGM dataset. In the problem of the PGM dataset (Barrett et al. 2018), rules are shared across rows or columns. In addition, the PGM dataset provides various regimes to measure the model's generalization in compositional performance, splitting the training and test sets according to predefined criteria. The PGM dataset comprises eight regimes, each containing 1.2M , 20K , and 200K training, validation, and test samples, respectively.

Experimental settings

We recorded the results for the test set using the hyperparameter settings that achieved the highest accuracy on the validation set of the I-RAVEN dataset. We used a batch size of 32, a learning rate of 0.001, and the Adam optimizer (Kingma and Ba 2015) with a weight decay of 1e-5. The balance term for the loss, α, was set to 6e-5. All panel images were resized to $(H, W) = (160, 160)$ and fed into the model as input. The dimension of feature vector d was set to 32. For experiments on the RAVEN-style dataset, we trained the model on all configurations simultaneously and averaged the accuracy on the test set using three different seeds. However, owing to the large size of the PGM dataset, we reported the accuracy on the test set using a single seed. We used neither data augmentation nor encoded information for shared rules during the training process. In addition, for the PGM dataset, as shared rules could be applied across rows or columns, we calculated the row- and column-wise probability for a single answer panel and averaged them to obtain the model's final output.

Results

Comparison with state of the art. Here, we compared our model with several existing models, namely, WReN (Barrett et al. 2018), SRAN (Hu et al. 2021), MRNet (Benny, Pekar, and Wolf 2021), ARII (Zhang et al. 2022), SCL (Yuhuai et al. 2020), PredRNet (Yang et al. 2023), and DRNet (Zhao, Xu, and Si 2024). Table 1 reveals that the proposed model achieved the state-of-the-art results on the I-RAVEN and RAVEN-FAIR datasets. However, DRNet only used data augmentation for the training process and an ablation study on how data augmentation improved performance revealed an approximately 5% accuracy gain on the I-RAVEN dataset. This suggests that our model achieved the highest average accuracy across four RPM datasets.

Although RANet used fewer parameters than most previous models, it exhibited the best performance because it considered other panels in the recognizing stage. In particular, the results of MRNet and DRNet were comparable with those obtained on the RAVEN dataset and the neutral regime of the PGM dataset. However, they relied on a large number of parameters, with RANet surpassing MRNet on both I-RAVEN and RAVEN-FAIR datasets. Additionally, it provided a comparison between the RANet and SRAN, which are methods in terms of learning relationships

Method	# of param	Neut	Int	Ext	H.O.P	H.O.TP	H.O.T	H.O.LT	H.O.SC	Avg
WReN	1.2M	62.6	62.4	17.2	27.2	41.9	19.0	14.4	12.5	32.4
ARII	91.2M	88.0	72.0*	**29.0**	50.0	64.1	32.1	16.0	12.7	45.5
MRNet	19.5M	93.4	68.1	19.2	38.4	55.3	25.9	**30.1**	**16.9**	43.4
PredRNet	1.28M	97.4	70.5	19.7	63.4	67.8	23.4	27.3	13.1	47.1
DRNet	24.6M	**99.0**	**83.8**	22.2	**93.7**	**78.1**	**48.8**	27.9*	13.1	**58.3**
RANet	**0.76M**	98.0*	71.77	24.41*	77.24*	77.00*	43.72*	22.99	13.32*	53.56*

Table 2: Overall accuracy (%) on all regimes of the PGM dataset, which comprises 1 Neutral and 7 OOD regimes, Neut: Neutral, Int: Interpolation, Ext: Extrapolation, H.O.P: Held-Out Attribute Pairs, H.O.TP: Held-Out Triple Pairs, H.O.T: Held-Out Triples, H.O.LT: Held-Out Line Type, H.O.SC: Held-Out Shape Color. * indicates that this method recorded the second-best performance.

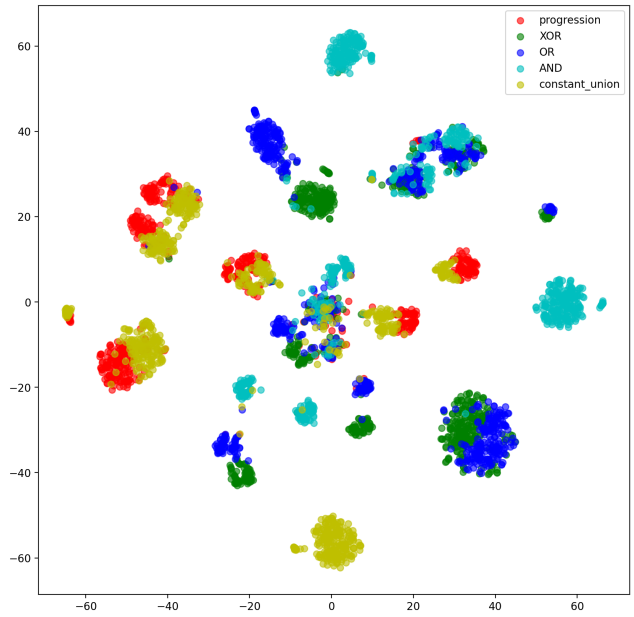

Figure 5: T-SNE visualization of the ground truth shared rule on the neutral regime of the PGM dataset.

between panels during the recognition stage. Our method not only used fewer parameters (44.0M vs. 0.76M), but also significantly outperformed in terms of average accuracy (64.8% vs. 97.7%). These results highlight that RANet explicitly identifies the differences between the feature maps of panels and can effectively learn relationships with fewer parameters.

Out-of-distribution generalization in PGM. We compared the performances of our model and previous models in all regimes of the PGM dataset, including WReN, ARII, MRNet, PredRNet, and DRNet. Table 2 indicates that our method recorded second-best performance among six out of eight regimes and average accuracy. Although RANet used fewer parameters and had good generalization performance, it performed worse in many regimes compared to DRNet. This outcome can be attributed to two reasons. The first is the difference in feature map extractors. In fact, a significant portion of DRNet's performance relies on the vison transformer (Dosovitskiy et al. 2021) encoder, which implies that the model inevitably has a large number of parameters.

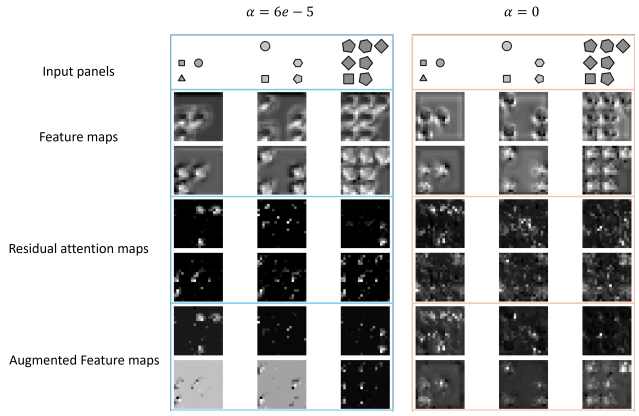

Figure 6: Visualization of feature maps and residual attention maps on the I-RAVEN dataset.

The second reason is the approach of our method toward the PGM dataset. Because it cannot be precisely determined whether the shared rule of the PGM problem applies across rows or columns, we set the model's final output as the average of row- and column-wise similarities. However, in severe out-of-distribution cases, this approach can cause the similarity of the side where the rule is not applied, to act as noise in the model's prediction. We leave the method of learning relationships via operations on feature maps between panels that constitute the entire matrix, rather than in a row- or column-wise manner, as future work.

T-SNE visualization for different relation types. To verify whether RANet could capture the compositional features of the dataset and distinguish relations accordingly, we performed T-SNE visualization (Laurens and Geoffrey 2008) on the ground truth shared rule $\left(r^{1,2}\right)$ of samples from the neutral regime of the PGM dataset. We chose this dataset for this experiment because unlike RAVEN-style datasets, which have multiple shared rules, the PGM dataset generally has only one shared rule per problem, making it more suitable for visualization. Figure 5 shows the T-SNE visualization results, which demonstrates that RANet has the ability to effectively distinguish between different relations, contributing to the model's performance.

Configuration	Similarity	Accuracy
Baseline(CNN + Reasoning)	0.649	97.56
RANet(CNN + RA + Reasoning)	**0.725**	**98.71**

Table 3: Comparison of cosine similarities of two correct row vectors and accuracies on the I-RAVEN dataset between the baseline model and RANet.

Visualization of feature map and residual attention map. To understand how the residual attention map augments the panel feature map and how the regularization on the sum of the residual attention map affects the model, we visualized the feature map and residual attention map. We performed this visualization for the same channel index across three panels forming a single row from the I-RAVEN dataset. Figure 6 shows the visualization results of two models. The images in the three columns on the left are from the model trained by both cross-entropy and regularization on the attention map, while the images in the three columns on the right are from the model trained by only cross-entropy. The first row of Figure 6 displays the input panels, the next two rows show the feature maps for each panel, the next two rows show the residual attention maps for each panel, and the last two rows are augmented feature maps.

In the three left columns, the feature maps are activated according to the positions of objects in each panel. By contrast, the attention maps are either activated similarly to the feature maps or focused on the differences with the former panel; thus, the feature maps of the panels are augmented from various perspectives. However, the attention maps of three right columns are activated for unnecessary parts. These results demonstrate that the RA module and regularization on the residual attention map enable the model to learn the differences with other panels in a more comprehensive manner at the recognition stage.

Effectiveness of the RA module. To verify whether the panel feature maps were augmented into a form that was much easier for the reasoning module to understand after passage through the RA module, we compared two models: one having RA module and the other without it. Specifically, we extracted the feature maps corresponding to the first two rows (six panels) from both models. These feature maps were then flattened to create two row vectors. Given that these row vectors were derived from the context panels and then shared a common rule embedding, we measured the cosine similarity between the two row vectors. A higher cosine similarity indicates that the model understands the features in a way that is much more conducive to inference. Notably, no operations between rows were applied before passage through the reasoning module. Table 3 presents a comparison between two models. The cosine similarity of the baseline was 0.649, and its accuracy was 97.56%, while the cosine similarity and accuracy of RANet were was 0.725 and 98.71%, respectively. Because the configuration of the reasoning module is the same, this result suggests that the RA module enables the model to infer deeper relationships by holistically recognizing the panels that constitute the rows.

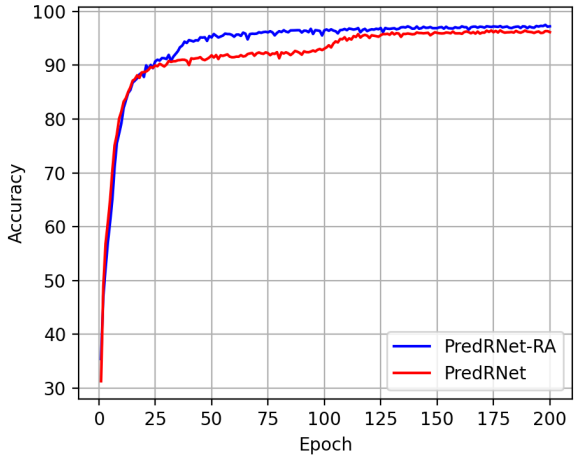

Figure 7: Comparison of accuracy between PredRNet and the model combining PredRNet with the RA module.

Another advantage of the RA module is its compatibility. Although our method outputs a panel's feature vector by applying global average pooling at the end of the RA module, it can also output the feature map by omitting this step. To verify whether considering other panels according to the Gestalt principles improves the model's abstract visual reasoning ability, we included the RA module in a previous state-of-the-art model, PredRNet, and assessed its performance with the I-RAVEN dataset. In particular, we inserted the RA module between the image encoder stage and the PRB stage of PredRNet, calling this model PredRNet-RA. Figure 7 plots the test accuracy of PredRNet and PredRNet-RA across different epochs. It shows that PredRNet-RA not only achieves higher accuracy but also converges faster. This result suggests that the advantage of the RA module can be applied to various abstract visual reasoning models.

Conclusion

We propose RANet to efficiently solve the RPM problem, where each panel is recognized holistically, by incorporating the concept of the Gestalt principles. Specifically, we introduced the RA module to enable the model to explicitly learn the differences among multiple panels using attention from their residuals. Our experiments demonstrated that even in RPM problems consisting of relatively simple shapes, performance improvement could be achieved by processing at the recognition stage. Furthermore, owing to the compatibility of the RA module, we showed that other models could achieve performance enhancement with a minimal number of additional parameters. We expect this method to be widely adopted in the field of abstract visual reasoning in the future.

References

Alawadhi, H. 2010. Form Perception: An Interactive Guide to the Gestalt Principles. *Thesis, Rochester Institute of Tech-*

nology.

Amanatiadis, A.; Kaburlasos, V.; and Kosmatopoulos, E. 2018. Understanding Deep Convolutional Networks through Gestalt Theory. In *Proceedings of the IEEE International Conference on Imaging Systems and Techniques (IST)*, 1–6. Krakow, Poland: IEEE.

Ba, J. L.; Kiros; Ryan, J.; Hinton; and E., G. 2016. Layer Normalization. *arXiv preprint arXiv:1607.06450*.

Barrett, D. G.; Hill, F.; Santoro, A.; Morcos, A. S.; and Lillicrap, T. 2018. Measuring abstract reasoning in neural networks. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 511–520. PMLR.

Benny, Y.; Pekar, N.; and Wolf, L. 2021. Scale-Localized Abstract Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12557–12565.

Bilker, W. B.; Hansen, J. A.; Brensinger, C. M.; Richard, J.; Gur, R. C.; and Gur, R. E. 2012. Development of Abbreviated Nine-item Forms of the Raven's Standard Progressive Matrices Test. *Assessment*, 19(3): 354–369.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hu, S.; Ma, Y.; Liu, X.; Wei, Y.; and Bai, S. 2021. Stratified Rule-Aware Network for Abstract Visual Reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, 1567–1574.

Hua, T.; and Kunda, M. 2019. Modeling Gestalt Visual Reasoning on the Raven's Progressive Matrices Intelligence Test Using Generative Image Inpainting Techniques. arXiv:1911.07736.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Kobayashi, T. 2016. Structured Feature Similarity with Explicit Feature Map. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1211–1219.

Larsen, A. B. L.; Sønderby, S. K.; Larochelle, H.; and Winther, O. 2016. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*.

Laurens, d. M., van; and Geoffrey, H. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9: 2579–2605.

Mascharka, D.; Tran, P.; Soklaski, R.; and Majumdar, A. 2018. Transparency by Design: Closing the Gap Between Performance and Interpretability in Visual Reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4942–4950.

Małkiński, M.; and Mańdziuk, J. 2023. A Review of Emerging Research Directions in Abstract Visual Reasoning. *Information Fusion*, 91: 713–736.

Mondal, S. S.; Webb, T.; and Cohen, J. D. 2023. Learning to Reason Over Visual Objects. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Raven, J. C. 1941. Standardization of Progressive Matrices, 1938. *British Journal of Medical Psychology*, 19: 137–150.

Song, K.; Wu, Y.; Chen, J.; Hu, T.; and Ma, H. 2022. Gestalt-Guided Image Understanding for Few-Shot Learning. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 409–424.

Uhlhaas, P. J.; Linden, D. E. J.; Singer, W.; Haenschel, C.; Lindner, M.; Maurer, K.; and Rodriguez, E. 2006. Dysfunctional Long-Range Coordination of Neural Activity during Gestalt Perception in Schizophrenia. *Journal of Neuroscience*, 26(31): 8168–8175.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems*, volume 30.

Yang, L.; You, H.; Zhen, Z.; Wang, D.-H.; Wan, X.; Xie, X.; and Zhang, R.-Y. 2023. Neural Prediction Errors enable Analogical Visual Reasoning in Human Standard Intelligence Tests. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 39572–39583.

Yuhuai, W.; Honghua, D.; Roger, G.; and Jimmy, B. 2020. The Scattering Compositional Learner: Discovering Objects, Attributes, Relationships in Analogical Reasoning.

Zaretskaya, N.; and Bartels, A. 2015. Gestalt perception is associated with reduced parietal beta oscillations. *NeuroImage*, 112: 61–69.

Zhang, C.; Gao, F.; Jia, B.; Zhu, Y.; and Zhu, S.-C. 2019a. RAVEN: A Dataset for Relational and Analogical Visual rEasoNing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5317–5327.

Zhang, C.; Jia, B.; Gao, F.; Zhu, Y.; Lu, H.; and Zhu, S.-C. 2019b. Learning Perceptual Inference by Contrasting. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Zhang, W.; Mo, S.; Liu, X.; and Song, S. 2022. Learning Robust Rule Representations for Abstract Reasoning via Internal Inferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 33550–33562.

Zhao, K.; Xu, C.; and Si, B. 2024. Learning Visual Abstract Reasoning through Dual-Stream Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 16979–16988.

Zhuo, T.; and Kankanhalli, M. 2021. Effective Abstract Reasoning with Dual-Contrast Network. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Reproducibility checklist

1. This paper:

- Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes)
- Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes)
- Provides well marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes)

2. Does this paper make theoretical contributions? (no)

- All assumptions and restrictions are stated clearly and formally. (yes/partial/no)
- All novel claims are stated formally (e.g., in theorem statements). (yes/partial/no)
- Proofs of all novel claims are included. (yes/partial/no)
- Proof sketches or intuitions are given for complex and/or novel results. (yes/partial/no)
- Appropriate citations to theoretical tools used are given. (yes/no)
- All theoretical claims are demonstrated empirically to hold. (yes/partial/no/NA)
- All experimental code used to eliminate or disprove claims is included. (yes/no/NA)

3. Does this paper rely on one or more datasets? (yes)

- A motivation is given for why the experiments are conducted on the selected datasets (partial)
- All novel datasets introduced in this paper are included in a data appendix. (NA)
- All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (NA)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. (yes)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. (yes)
- All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. (NA)

4. Does this paper include computational experiments? (yes)

- Any code required for pre-processing data is included in the appendix. (yes)
- All source code required for conducting and analyzing the experiments is included in a code appendix. (yes)
- All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes)
- All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes)
- If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. (yes)
- This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. (yes)
- This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. (yes)
- This paper states the number of algorithm runs used to compute each reported result. (yes)
- Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. (no)
- The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). (no)
- This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. (yes)
- This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. (partial)