Visual Perception Inference on Raven's Progressive Matrices by Semi-supervised Contrastive Learning

Aihua Yin[1,2], Weiwen Lu[1,2], Sidong Wang[1,2], Hongzhi You[5], Ruyuan Zhang[4], Dahui Wang[1,3], Zonglei Zhen[1], and Xiaohong Wan[1,2(✉)]

[1] State Key Laboratory of Cognitive Neuroscience and Learning, Beijing, China
[2] IDG/McGovern Institute for Brain Research, Beijing, China
[3] School of Systems Science, Beijing Normal University, Beijing 100875, China
xhwan@bnu.edu.cn
[4] Institute of Psychology and Behavioral Science and Shanghai Mental Health Center, Shanghai Jiao Tong University, Shanghai 200030, China
[5] School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 611731, China

Abstract. The current deep neural networks (DNNs) in mimicking human perception remain challenges for solving visual reasoning tasks. Human perception does not merely involve a passive observer labeling sensory signals, but also contains an active inference about object attributes and their relationships towards an intended output (e.g., an action). In this work, we propose a variational autoencoder (VAE) model to discriminate the ranking relationships between object attribute values by semi-supervised contrastive learning, dubbed as SSCL-VAE. This perception-based model solves the visual reasoning task of Raven's Progressive Matrices (RPM) in three benchmarks (RAVEN, I-RAVEN and RAVEN-Fair), with high accuracy close to humans, as well as many end-to-end supervised models. The current work thus suggests that constructions of general cognitive abilities like human perception may empower the perceptron with DNN to solve high-level cognitive tasks such as abstract visual reasoning in a human-like manner.

Keywords: Perception · Inference · Contrastive learning · Semi-supervised · Autoencoder

1 Introduction

Deep neural networks (DNNs) by end-to-end supervised learning have achieved great success in visual categorizing, but are not versatile for visual reasoning [1–3]. The critical feature in abstract visual reasoning tasks, such as Raven's Progressive Matrices (RPMs) [4], is that the rules governing a sequence of entities are semantically defined by their spatiotemporal relations [5]. Learning these

A. Yin and W. Lu—These authors equally contributed to this study.

L. Fang et al. (Eds.): CICAI 2022, LNAI 13605, pp. 399–412, 2022.
https://doi.org/10.1007/978-3-031-20500-2_33

semantic relationships by supervisions is not straightforward. Even a number of end-to-end supervised DNN models have been developed to achieve high performance in solving RPMs [1–3,6–8], these models lack interpretability and generalizability. Thus far, it remains challenging for DNNs to behave like humans in solving such visual reasoning tasks [9,10].

150 years ago, Dr. von Helmholtz addressed that the properties of the external world are not directly provided by sensory inputs, but are probably inferred through human hierarchic neural processes [11]. Perception is based on higher-order features and their relationships, rather than locally defined features, related to the stimulus. Indeed, these attributes are mixed together and must be disentangled to make explicit percepts [12,13].

Differing from the current DNN methods that need to learn from scratch the associations between the contexts and the supervised labels with a huge number of samples, humans do not rely on such domain-specific knowledge or experiences in RPMs, but their prior general cognitive abilities in recognition of object attributes and their relationships. Although humans at a very early stage of life have no such a concept of semantics and symbols, they can recognize varieties of objects [14], and comprehend simple rules governing the world and apply these rules to new contexts [15]. Importantly, the object attribute representations in the human brain are unique and invariant in different contexts [16,17], and the context-dependent relationships between attributes among the objects are implicitly inferred [11]. For instance, we recognize the same color of 'green' from different objects and further recognize that the 'green' color looks lighter than the 'red' color, but darker than the 'cyan' color.

Inspired by these insights from human perception, we move a step further towards visual reasoning ability of artificial intelligence (AI) on the basis of the general cognitive abilities in object perception as humans do. In this work, we propose a variational autoencoder (VAE) model for visual perception by semi-supervised contrastive learning (SSCL), dubbed as SSCL-VAE. The motivation of the SSCL method is to make embeddings of the same attribute from different objects are close to each other while embeddings of different attributes are separated away from each other [18]. Rule logic execution in this model follows the approach of probabilistic abduction and execution (PrAE) model [19], in which the inference engine aggregates distributed representations of a set of object attributes in the context panels to infer a posterior probabilistic representation of the target panel. Notably, the PrAE model is originally trained by supervisions of both the correct and incorrect answers, and also the ground truths of rules (metadata) contained in each RPM problem as auxiliary annotations. In this work, without these ample supervisions, SSCL-VAE merely enforces human-like perceptual abilities, in particular, the ability of qualitative comparisons between object attribute values from the different objects in the context sets, but not the answer sets. The proposed model obtains accuracies as high as humans, and many of the previous supervised models in solving RPMs in three benchmarks. Importantly, we demonstrate that constructions of human-like perception abilities on DNNs can empower AI such a capability of solving abstract visual reasoning in a human-like manner. To the best of our knowledge, this has not yet been explored in the domain of visual reasoning tasks.

2 Related Work

2.1 Object Representations

Recognition of object attributes is critical to solve visual reasoning tasks, as the latent relations, namely rules, that govern the context of instance are defined by these visual features. DNNs are believed to versatilely fit any desired function with a constraint of the loss function. However, the embeddings of latent attributes are too flexible to comply well with the semantics of object attributes that are used in these tasks, such as types, sizes, and colors in the RAVEN dataset [1–3,6–8]. Instead, the object attributes are often blended in the latent embeddings. Recently, variational autoencoder (VAE) [20–22] and neural-vector [23] models have been proposed to build stable representations and to disentangle the blended representations of object attributes. A straightforward approach to parcel objects into desired attribute representations is to use the metadata of object attributes as auxiliary annotations to train the perception module. However, the prior annotations are needed to label by humans. In image recognitions, SSCL has been used to discover better representations by comparing the relationship representations from the same or different attributes [18,24]. We here leverage this method in cooperation with a VAE model to shape the latent embeddings, in order to enable the model to have simple relational inference capabilities. This proposed model thus constrains the representations of the same attribute to be invariant across different objects, importantly complying with the rankings of attribute values too.

2.2 Visual Reasoning

Most of supervised models designed to solve visual reasoning tasks mainly focus on the visual reasoning process [1–3,6–8], as the baselines of DNNs fails to solve these high-level cognitive tasks. A common motivation for visual reasoning models is to learn relational representations of latent rules by maximizing similarity between analogical relations and minimizing similarity between non-analogical relations [2,3,25–29]. This is achieved by comparing the relational representations with correct and incorrect answers. In striking contrast, the proposed model here embeds the relational representations in the perception module.

2.3 Neuro-symbolic Models

Unlike the monolithic DNN models, the neuro-symbolic models are composed of a perception module at the frontend and an inference module at the backend [3,30–32]. Nonetheless, it remains challenging to train the neuro-symbolic models with the end-to-end supervised training form. Thereby, auxiliary annotations of the latent rules are additionally used to constrain the rule representations in the PrAE model [19]. Although the currently proposed model partially shares the inference engine with the PrAE model, we here use SSCL to train the visual perception module alone. Semi-supervised learning (SSL) has been also used to

solve visual reasoning tasks, such as RPMs [6,33]. However, the conventional SSL combines a large number size of labeled data and a small number size of unlabeled data. In contrast, SSCL used here has no concrete labels, but the pairwise rankings that are partially supervised. For this reason, the current method is also called semi-supervised.

3 Methods

In the RAVEN dataset [35,36], each problem consists of 9 panels in a form of 3×3 matrix with 8 context panels and a missing panel at the last panel. The goal of the task is to find out one from 8 candidate panels that completes the matrix with satisfactions of the row-wise latent rules governing the organization of object attributes Fig. 1. Besides, there are 7 configurations [Center, Left-Right (L-R), Upper-Down (U-D), 2×2 Grid, 3×3 Grid, Out-In Center (O-IC), Out-In Grid (O-IG)] in the RAVEN dataset. In different configurations, objects in panels are organized differently. We train an independent model for each configuration.

Overall, the task requires two independent cognitive abilities of object perception and rule inference. If perception on object attributes is perfect, then the process of identifying the latent rules becomes plain, an exhaustive search within the rule space in a finite set [36]. Differing from the visual perception tasks, the object attributes are also latent in visual reasoning tasks, but are conventionally required to infer from the spatiotemporal relations from the context

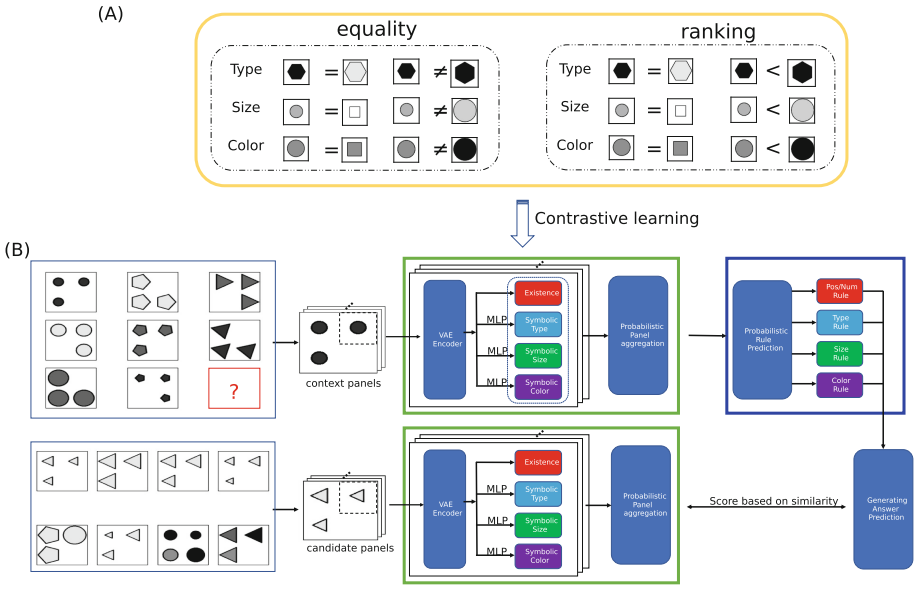

Fig. 1. Description of SSCL-VAE model. (A) The representative examples of semi-supervised contrastive learning in the equality and ranking methods, respectively. (B) A schematic of the model architecture. Please see the main text for details.

in each instance. In other words, both the object attributes and rules remain to be identified. This is hard to implement in DNNs, and also remains challenging for the neuro-symbolic models. The novelty of SSCL-VAE here mainly focuses on the learning approach, rather than the DNN architecture (Fig. 1).

3.1 Object-Based Variational Autoencoder

We use VAE as the backbone of the visual perception module. VAE consists of an encoder that maps the visual inputs to latent representations and a decoder that is required to reconstruct the input images reversely from the latent representations. We first take each object in each panel as the input and pretrain the VAE with the loss as follows,

$$Loss_{\text{VAE}} = \sum_{x} \Big(||x - \hat{x}||_2 + D_{KL}\big(\mathcal{N}(\mu_x, \sigma_x^2)||\mathcal{N}(0,1)\big) \Big) \tag{1}$$

$$\text{with } D_{KL}\big(p||q\big) = \sum_{i} p_i \big(log\big(p_i\big) - log\big(q_i\big)\big), \tag{2}$$

where x is the input (original) image, $\hat{x}$ is the reconstructed image, $||\cdot||$ means L2 norm, D_{KL} means Kullback-Leibler divergence, and $\mathcal{N}(\mu, \sigma^2)$ means Gaussian distribution with the mean μ and standard deviation σ.

The requirement of reconstruction ensures plenty of information in the latent embeddings of objects and also helps to stabilize the representations in later training. In the configuration such as 2×2 Grid, there may exist multiple objects in a panel. We then try to capture all the objects by selecting regions of interest, though the existence of objects in regions is not guaranteed.

3.2 Attribute Discrimination

Differing from the conventional VAE for image reconstruction, the visual perception module is additionally required to discriminate the object attributes, such as type, size and color in RPMs. To do so, we add an additional multi-layer perceptron (MLP) for each attribute to transform the latent embeddings in VAE into the distributed probabilities belonging to the separate attribute values, in which the dimension of type, size and color attribute is 5, 6 and 10, respectively. To train the MLP and finetune the encoder of VAE, we use the information of metadata of object attributes as auxiliary annotations. However, instead of the exact labels, we train the model to acquire two common-sense knowledge on these visual attributes. First, the model is trained to know whether the values of the same attribute from any pair of objects are equivalent or not. In other words, the representations of visual attributes in the model are unique and invariant across different objects. We here denote this ability of visual perception as equality. Obviously, this ability cannot discriminate the relationships between the attribute values. Second, the model is further trained to know the intrinsic orders or rankings of the values of the same attribute from two different objects. In other words, the model acquires the ability to comprehend the relationships

between the attribute values. We here denote this ability of visual perception as ranking. To allow the model to acquire these abilities, we leverage contrastive learning to make pair-wise comparisons between any two objects within each RPM instance. For instance, the model is informed that two objects in an RPM instance share the same color attribute, and further which one has a lighter color, but not the exact color value in the metadata. To be specific, the corresponding loss can be formulated as:

$$Loss_a = \sum_{n=1}^{N} \sum_{i,j=1}^{16} D_{JSD}(s_{ij}^{(n,a)}, d_{ij}^{(n,a)}),\ a \in \{type, size, color\} \tag{3}$$

$$\text{with } s_{ij}^{(n,a)} = \left\{ P(l_i^{(n,a)} > l_j^{(n,a)}), P(l_i^{(n,a)} = l_j^{(n,a)}), P(l_i^{(n,a)} < l_j^{(n,a)}) \right\}, \tag{4}$$

$$d_{ij}^{(n,a)} = \left\{ P(y_i^{(n,a)} > y_j^{(n,a)}), P(y_i^{(n,a)} = y_j^{(n,a)}), P(y_i^{(n,a)} < y_j^{(n,a)}) \right\} \tag{5}$$

$$D_{JSD}(p, q) = \frac{1}{2} D_{KL}(p || \frac{p+q}{2}) + \frac{1}{2} D_{KL}(q || \frac{p+q}{2}), \tag{6}$$

where N is the training batch size, D_{JSD} means Jensen-Shannon divergence, $y_i^{(n,a)}, l_i^{(n,a)}$ are the predicted and ground-truth label of attribute a of the ith object in the nth RPM respectively. Here, only existing objects in each panel are used for training.

On the other hand, the input images may contain no objects, such as in the 2×2 Grid configuration. Hence, the model needs to discern the existence of objects. We add a MLP for the attribute of existence as well. For the sake of simplicity, we use negative log-likelihood loss that counts whether there exists an object. In total, the loss is expressed as follows,

$$Loss = Loss_{exist} + Loss_{type} + Loss_{size} + Loss_{color} \tag{7}$$

3.3 Rule Inference

In the current model, the rule inference module is independent of the visual perception module. Specifically, we implement non-symbolic inference using the rule inference engine as used in the PrAE model [17], in which the probabilities of object attributes are aggregated to obtain the probabilities of panel attributes.

$$p_{panel}^{a} = \sum_{\boldsymbol{exist}} p_{\boldsymbol{exist}} \cdot \exp\left(\frac{\sum_{obj} \log(p_{obj}^{a}) \boldsymbol{exist}_{obj}}{\sum \boldsymbol{exist}_{obj}} \right) \tag{8}$$

where p is a probability vector, $exist$ is a binary vector describing the existence of objects.

For each object attribute, the model calculates the probabilities of the potential rules based on the probabilities of panel attributes, and the rule with maximum normalized probability is chosen to be the predicted rule. Hence, the rule

inference model cannot discover new rules, but discriminates the prior rule candidates, with an assumption that the model has full knowledge of the potential rules.

Specifically, for attribute a, the probabilistic representation of the rule can be obtained by calculating the hadamard-product of the attribute representation and the rule mask,

$$P(r) = \sum_{M \in mask(r)} \sum_{all\ elements} (p_1, p_2, p_3, ..., p_n)^T \odot M \tag{9}$$

where p is a probability vector, M is a rule mask composed of 0, 1, and each column represents the attribute representation of each panel under a certain rule.

While the normalization process of rule probabilities can be formulated as follows,

$$P_{norm}(r) = \frac{P(r)}{\sum_{r' \in E} P(r')} \tag{10}$$

where E denotes the set of potential rules.

3.4 Answer Generation

Finally, the model predicts the potential rules of each attribute containing in RAVEN through the rule inference engine and generates an aggregated probability distribution of attributes in the target panel (Fig. 1). Meanwhile, the aggregated probability distributions of attributes for the candidate panels are also computed by the visual perception module. We then compare the Jensen-Shannon divergence (JSD) between the generated attribute probability distributions with those of the candidates. The candidate with smallest divergence is then selected as the answer. This process is similar to the supervised contrastive learning approach used in the previous studies of CoPINet [3] and PrAE [19].

4 Experiments

4.1 Experimental Setup

We test SSCL-VAE in the RAVEN dataset including three benchmarks of RAVEN [36], I-RAVEN and RAVEN-Fair. The three benchmarks share the same problem contexts, but different candidate sets. We then use the same training model to test its performance in the three benchmarks. We separately train the models for the 7 different RAVEN configurations. We train our model on 6,000 samples in the training dataset and test the model on 2,000 samples in the testing dataset for each configuration, while the validation dataset is not used. The training procedure is separated into two phases. First, we pretrain VAE with 50 neurons in the latent layer to represent the visual images for 100 epochs. In order to better achieve the reconstruction effect, the learning rate of

ADAM optimizer is set to 0.001. Second, we simultaneously train both VAE and MLP to discriminate the attribute values for 100 epochs and the learning rate of the ADAM optimizer is set to 0.01. The batch size is 256 in both phases. The inputs of the object images (160×160) are resized to 32×32. Further, we also test the proposed model on the MNIST benchmark [37]. We train the model on 60,000 samples in the training dataset, and test on 10,000 samples in the testing dataset. The input size of the images is 28×28, and the batch size is also 256. All the models are implemented in PyTorch and runned with Intel(R) Xeon(R) Platinum 8272CL CPUs and NVIDIA Geforce RTX 3090 Founders Edition GPUs.

Table 1. Average accuracy (%) of different models.

	Methods	Avg	Center	2×2	3×3	L-R	U-D	O-IC	O-IG
RAVEN	Equality	39.4	37.4	61.5	33.4	37.4	32.9	28.8	34.4
	Ranking	80.1(+**40.7**)	89.3(+**51.9**)	82.3(+**20.8**)	76.3(+**42.9**)	88.8(+**51.4**)	88.3(+**55.4**)	72.0(+**43.2**)	63.9(+**19.5**)
	Full	91.9	91.8	93.4	91.2	99.8	89.0	97.4	86.6
	PrAE [19]	82.8	100.0	84.4	38.7[a]	95.2	95.9	96.0	69.5
I-RAVEN	Equality	53.4	49.9	67.8	38.8	58.3	55.8	47.3	56.0
	Ranking	85.9(+**32.5**)	92.6(+**42.7**)	85.2(+**17.4**)	81.9(+**43.1**)	92.3(+**34.0**)	91.9(+**36.1**)	84.5(+**37.2**)	72.8(+**16.8**)
	Full	94.4	95.0	95.6	95.1	100.0	92.9	99.0	90.8
	PrAE [19]	87.8	100.0	87.5	55.5[a]	97.6	98.1	98.4	78.0
RAVEN-Fair	Equality	58.5	56.2	72.7	54.3	59.2	57.0	47.4	62.9
	Ranking	88.3(+**29.8**)	94.0(+**37.8**)	87.0(+**14.3**)	84.5(+**30.2**)	94.6(+**35.4**)	93.8(+**36.8**)	83.8(+**36.4**)	80.2(+**17.3**)
	Full	95.6	96.2	95.8	95.7	100.0	94.7	99.0	93.3
	PrAE [19]	90.0	100.0	92.4	58.0[a]	98.0	98.8	98.3	84.8
	Human [36]	84.4	95.4	81.8	79.5	86.4	81.8	86.4	81.8

[a] 3×3 Grid is calculated by the training model of 2×2 Grid.

4.2 Evaluation of General Performance in Three Benchmarks

We first evaluate SSCL-VAE performance in the three benchmarks of RAVEN, I-RAVEN and RAVEN-Fair in comparison with different models that use different sources of metadata. The equality method means that the model has unique and invariant representations of attributes, and the ranking method means that the model can further infer the pair-wise relationships among the same attribute, while the full method means that the model uses the concrete labels of attribute metadata for training the model. We also compare with the PrAE model in I-RAVEN [19], as our model share the rule inference engine with PrAE. Table 1 shows the accuracies of different models. On average, the ranking method achieves accuracies 41%, 32% and 30% higher than the equality method in RAVEN, I-RAVEN and RAVEN-Fair, respectively, although lower than the full method with the detailed labels of metadata of visual attributes. In contrast, the PrAE model that alternatively uses the metadata labels of rules and the correct and incorrect answer panels as supervisions only achieves marginally better performance than the ranking method. Hence, preposition of attribute relationships in the frontend perception module, or rendering inductive inference to the perception module, makes it available for rule inference, close to humans' performance (Table 1). However, simple attribute separation in visual perception (the equality method) are insufficient to acquire such an inference ability. These

results illustrate the reason why end-to-end supervised-learning DNNs are not versatile for visual reasoning tasks.

4.3 Evaluation of Attribute and Rule Representations

To appreciate the benefits of the ranking method in comparison with the equality method, we evaluate their attribute representations and rule representations. As much expected, the representations of entities of each attribute (type, size and color) do not necessarily match the actual entities, although these representations are unique and invariant across objects (Fig. 2A upper). In particular, these representations are not consistent across different configurations. In contrast, the attribute representations in the ranking method considerably align well with the ground truths of the order of attribute values in each category, even across different configurations (Fig. 2A bottom). Accordingly, the rule representations in the equality method are also inconsistent with the true rules (Fig. 2B upper). Please keep in mind that the rules are entirely defined by the attribute values. However, those representations in the ranking method largely align well with the ground truths (Fig. 2B bottom). Hence, the local rankings result in global match with the true order of attribute values, which in turn provides accurate predicted rule simply by aggregations of distributed probabilities of attributes across the context panels.

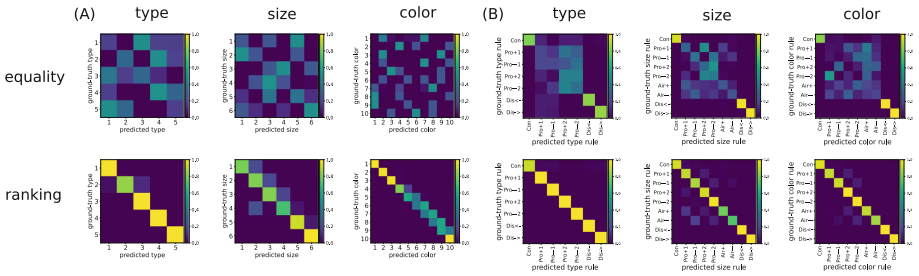

Fig. 2. The representations of object attributes and rules across all of the RAVEN configurations. (A) The confusion matrix between the predicted attribute values and ground truths. (B) The confusion matrix between the predicted rule values and ground truths.

4.4 Evaluations of Generalizability

We further evaluate the generalizability of our proposed model in solving I-RAVEN as examples. First, we evaluate cross-configuration generalizability for our model using the ranking method. The models trained in the configurations of Center, L-R, U-D and O-IC can solve the problems in the other simple configurations, but not on the configurations of the 2×2 Grid, 3×3 Grid, and O-IG. In contrast, the models trained in the configurations of the 2×2 Grid, 3×3 Grid, and O-IG can fairly transfer to solve the problems in other configurations

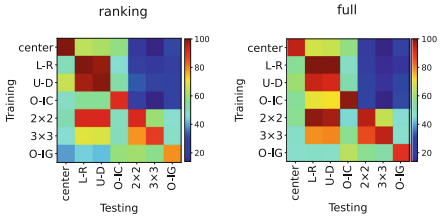

Fig. 3. Cross-configuration generalizability on I-RAVEN.

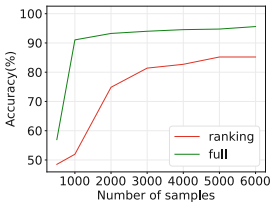

Fig. 4. The accuracy change on I-RAVEN with number of samples.

(Fig. 3). The cross-configuration generalizability using the ranking method is similar as that using the full labels of metadata (Fig. 3). Second, we examine the performance dependent on the training sample size. Figure 4 illustrates that the ranking model is not so much sensitive to the training sample sizes (red line) when the training sample size is larger than 2,000, while the metadata model remains stable until the training sample size is no less than 1,000.

4.5 MNIST

Finally, we test the proposed models also on the MNIST benchmark (Fig. 5A). The performance by the equality method is close to the chance level (10%),

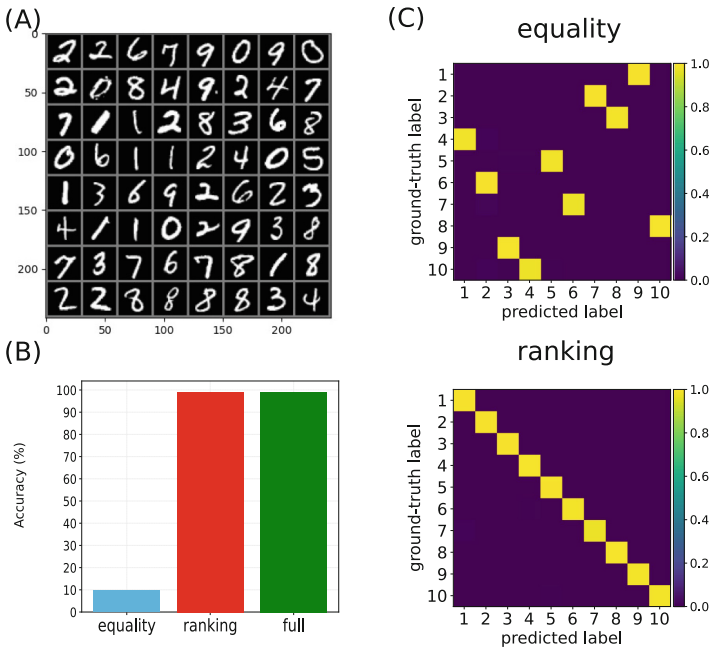

Fig. 5. The performance of the two models on the MNIST dataset.(A) Visualization of the MNIST dataset.(B) Accuracy (%) of different models. (C) The confusion matrix between the predicted attribute values and ground truths.

while that by the ranking method is as high as 99.1%, as same as the full metadata method (Fig. 5B). Again, although the entities are uniquely and invariantly represented in the equality method, the distributed representations are not consistent with the ground truths (Fig. 5C upper). Instead, the representations by the ranking method are aligned well with the ground truths (Fig. 5C bottom). Hence, the SSCL-VAE model may have broad applications in visual perception, including both visual categorizing and visual reasoning.

5 Conclusion

In this paper we present SSCL-VAE, a semi-supervised model that obtains high performance on three RAVEN benchmarks involving abstract visual reasoning. The previous supervised learning on DNNs is dependent on the task-specific knowledge from the answers and auxiliary annotations, and also mainly focus on the rule inference module in the backend. By contrast, SSCL-VAE provides an approach to establish the general cognitive abilities in human perception, but not task-specific knowledge [9,10,38]. Thereby, it has strong robustness even for small sample size for training and generalizability for cross-configuration tests. Importantly, the current model empowers the general perception abilities, in particular, the inference on the relations between visual attributes, and enables non-symbolic inference with interpretability. The simplicity of this approach, we believe, should afford its broad applications in solving other spatiotemporal reasoning tasks [38,39].

The current model of SSCL-VAE also has some important limitations deserved to be improved. First, SSCL needs partial information of metadata, which is sometimes hard to access. It might be improved by self-supervised method, rather than semi-supervised approach. Self-supervised contrastive learning has been broadly applied in computer vision [18,24], natural language processing (NLP) [40,41], and other domains. Second, the model can be trained by independent objects and tasks to construct its general cognitive abilities in object perception including inductive inference. It remains to explore this potential by training independent tasks and testing on other independent tasks. Third, more general cognitive abilities and higher-level cognitive function can be further incorporated into the model to provide more versatile intelligent abilities in solving complex tasks.

Acknowledgments. We thank Dr. Bo Hong's discussions, inspirations and comments. This work was partially supported by grants from the National Science and Technology Innovation 2030 Project of China to Xiaohong Wan (2021ZD0203701).

References

1. Hoshen, D., Werman, M.: IQ of neural networks. arXiv preprint arXiv:1710.01692 (2017)
2. Barrett, D., Hill, F., Santoro, A., Morcos, A., Lillicrap, T.: Measuring abstract reasoning in neural networks. In: International Conference on Machine Learning, pp. 511–520. PMLR (2018)
3. Zhang, C., Jia, B., Gao, F., Zhu, Y., Lu, H., Zhu, S.-C.: Learning perceptual inference by contrasting. In: Advances in Neural Information Processing Systems, pp. 1075–1087 (2019)
4. Raven, J., Court, J., Raven, J.: Raven's Progressive Matrices. Oxford Psychologists Press, Oxford (1938)
5. Lovett, A., Forbus, K., Usher, J.: Analogy with qualitative spatial representations can simulate solving raven's progressive matrices. In: Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 29, no. 29 (2007)
6. Zhuo, T., Kankanhalli, M.: Solving Raven's progressive matrices with neural networks. arXiv preprint arXiv:2002.01646 (2020)
7. Mandziuk, J., Zychowski, A.: DeepIQ: a human-inspired AI system for solving IQ test problems. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2019)
8. Zhuo, T., Kankanhalli, M.: Effective abstract reasoning with dual-contrast network. In: International Conference on Learning Representations (ICLR) (2021)
9. Marcus, G., Davis, E.: Insights for AI from the human mind. Commun. ACM **64**, 38–41 (2020)
10. Fodor, A., WPylyshyn, Z., et al.: Connectionism and cognitive architecture: a critical analysis. Cognition **28**(1–2), 3–71 (1988)
11. von Helmholtz, H.: The aim and progress of physical science. In: Kahl, R. (ed.) Selected Writings of Hermann von Helmholtz, pp. 223–245. Wesleyan University Press, Middletown (Originally Published 1869) (1971)
12. Knill, D.C., Richards, W.: Perception as Bayesian inference. Cambridge University Press, Cambridge (1996)
13. DiCarlo, J.J., Cox, D.D.: Untangling invariant object recognition. Trends Cogn. Sci. **11**(8), 333–341 (2007)
14. Spelke, E.S.: Principles of object perception. Cogn. Sci. **14**(1), 29–56 (1990)
15. Gopnik, A., Glymour, C., Sobel, D.M., Schulz, L.E., Kushnir, T., Danks, D.: A theory of causal learning in children: causal maps and Bayes nets. Psychol. Rev. **111**(1), 3–32 (2004)
16. Li, N., Dicarlo, J.: Unsupervised natural experience rapidly alters invariant object representation in visual cortex. Science, 1502–1507 (2008)
17. Mansouri, F.A., Freedman, D.J., Buckley, M.J.: Emergence of abstract rules in the primate brain. Nat. Rev. Neurosci. **21**, 596–610 (2020)
18. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)
19. Zhang, C., Jia, B., Zhu, S.-C., Zhu, Y.: Abstract spatial-temporal reasoning via probabilistic abduction and execution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9736–9746 (2021)
20. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: International Conference on Learning Representations (ICLR) (2013)

21. Higgins, I., et al.: betavae: Learning basic visual concepts with a constrained variational framework. In: International Conference on Learning Representations (ICLR) (2017)
22. Burgess, C.P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., Lerchner, A.: Understanding disentangling in VAE. arXiv preprint arXiv:1804.03599 (2018)
23. Hersche, M., Zeqiri, M., Benini, L., Sebastian, A., Rahumi, A.: A neuro-vector-symbolic architecture for solving Raven's progressive matrices. https://arxiv.org/abs/2203.04571v1
24. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)
25. Jahrens, M., Martinetz, T.: Solving Raven's progressive matrices with multi-layer relation networks. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–6. IEEE (2020)
26. Malkinski, M., Mandziuk, J.: Multi-label contrastive learning for abstract visual reasoning. arXiv preprint arXiv:2012.01944 (2020)
27. Wu, Y., Dong, H., Grosse, R., Ba, J.: The Scattering Compositional Learner: Discovering Objects, Attributes, Relationships in Analogical Reasoning. arXiv preprint arXiv:2007.04212 (2020)
28. Kiat, N.Q.W., Wang, D., Jamnik, M.: Pairwise relations discriminator for unsupervised Raven's progressive matrices. arXiv preprint arXiv:2011.01306 (2020)
29. Kim, Y., Shin, J., Yang, E., Hwang, S.J.: Few-shot visual reasoning with meta-analogical contrastive learning. In: Advances in Neural Information Processing Systems, vol. 33 (2020)
30. Yi, K., et al.: CLEVRER: collision events for video representation and reasoning. In: International Conference on Learning Representations (2020)
31. Mao, J., Gan, C., Kohli, P., Tenenbaum, J.B., Wu, J.: The neuro-symbolic concept learner: interpreting scenes, words, and sentences from natural supervision. In: International Conference on Learning Representations (ICLR) (2019)
32. Ding, M., Chen, Z., Du, T., Luo, P., Tenenbaum, J.B., Gan, C.: Dynamic visual reasoning by learning differentiable physics models from video and language. In: Advances in Neural Information Processing Systems, vol. 35 (2021)
33. Tarvainen, A., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems (2017)
34. Berthelot, D., Carlini, N., Goodfellow, I., Oliver, A., Papern, N., Raffel, C.: MixMatch: a holistic approach to semi-supervised learning. In: Advances in Neural Information Processing Systems (2019)
35. Matzen, L.E., Benz, Z.O., Dixon, K.R., Posey, J., Kroger, J.K., Speed, A.E.: Recreating Raven's: software for systematically generating large numbers of ravenlike matrix problems with normed properties. Behav. Res. Methods **42**(2), 525–541 (2010)
36. Zhang, C., Gao, F., Jia, B., Zhu, Y., Zhu, S.-C.: Raven: a dataset for relational and analogical visual reasoning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5317–5327 (2019)
37. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
38. Chollet, F.: On the measure of intelligence. arXiv preprint arXiv:1911.01547 (2019)

39. Spratley, S., Ehinger, K., Miller, T.: A closer look at generalisation in RAVEN. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12372, pp. 601–616. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58583-9_36
40. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, vol. 26 (2013)
41. Saunshi, N., Plevrakis, O., Arora, S., Khodak, M., Khandeparkar, H.: A theoretical analysis of contrastive unsupervised representation learning. In: International Conference on Machine Learning, pp. 5628–5637. PMLR (2019)