Hierarchical ConViT with Attention-Based Relational Reasoner for Visual Analogical Reasoning

Wentao He[1], Jialu Zhang[1], *Jianfeng Ren[1,2], Ruibin Bai[1,2], Xudong Jiang[3]

[1]The Digital Port Technologies Lab, School of Computer Science, University of Nottingham Ningbo China
[2]Nottingham Ningbo China Beacons of Excellence Research and Innovation Institute, University of Nottingham Ningbo China
[3]School of Electrical & Electronic Engineering, Nanyang Technological University
{scxwh1,sgxjz1,jianfeng.ren,ruibin.bai}@nottingham.edu.cn, exdjiang@ntu.edu.sg

Abstract

Raven's Progressive Matrices (RPMs) have been widely used to evaluate the visual reasoning ability of humans. To tackle the challenges of visual perception and logic reasoning on RPMs, we propose a Hierarchical ConViT with Attention-based Relational Reasoner (HCV-ARR). Traditional solution methods often apply relatively shallow convolution networks to visually perceive shape patterns in RPM images, which may not fully model the long-range dependencies of complex pattern combinations in RPMs. The proposed ConViT consists of a convolutional block to capture the low-level attributes of visual patterns, and a transformer block to capture the high-level image semantics such as pattern formations. Furthermore, the proposed hierarchical ConViT captures visual features from multiple receptive fields, where the shallow layers focus on the image fine details while the deeper layers focus on the image semantics. To better model the underlying reasoning rules embedded in RPM images, an Attention-based Relational Reasoner (ARR) is proposed to establish the underlying relations among images. The proposed ARR well exploits the hidden relations among question images through the developed element-wise attentive reasoner. Experimental results on three RPM datasets demonstrate that the proposed HCV-ARR achieves a significant performance gain compared with the state-of-the-art models. The source code is available at: https://github.com/wentaoheunnc/HCV-ARR.

Introduction

Research in computer vision has advanced significantly recently (Dosovitskiy et al. 2021; He et al. 2016; Zhang et al. 2022). The research focus is shifting from visual recognition of individual objects to visual understanding of image/video scenes (Zellers et al. 2019). Visual reasoning, as one of the visual understanding tasks, usually consists of two related tasks, "visual perception" and "logical reasoning". The former perceives the image/video through a perception system (Dosovitskiy et al. 2021), and the latter discovers reasoning rules through a cognition system (Crouse et al. 2021). A lot of research efforts have been devoted to developing a system that can not only visually recognize objects from scenes, but also conduct logical reasoning over the perceived visual information (Sekh et al. 2020; Zhang et al. 2019a).

*Corresponding author.

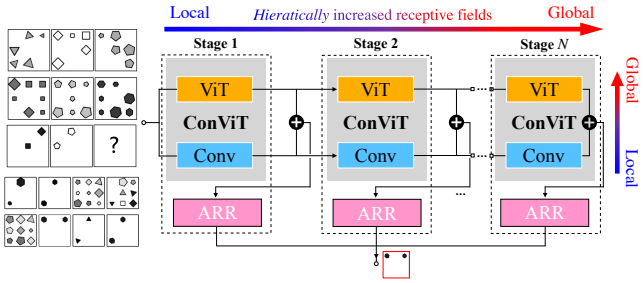

Figure 1: RPM problems are challenging as both local attributes such as `Color`, `Type` and `Size` and global attributes on pattern combinations such as `Number` and `Position` need to be extracted simultaneously and a different rule may be applied on each attribute. The proposed HCV-ARR consists of a set of ConViT blocks that can simultaneously perceive local attributes through convolutional blocks and global attributes through transformer blocks. An element-wise attention-based relational reasoner is then designed to exploit reasoning rules among different attributes.

Raven's Progressive Matrix (RPM) problem is one of the frequently-used tests on human's visual analogical reasoning in cognitive psychology (Raven 2000). An RPM problem is formed by a 3×3 pictorial matrix with the last one left blank, as shown in Fig. 1. The objective is to identify the missing entry from eight candidate answers based on the visual context and inductive rules. To minimize the impact of language barrier and culture bias, the pictorial matrices are often composed of regular polygons. Several RPM databases (Barrett et al. 2018; Benny, Pekar, and Wolf 2021; Hu et al. 2021; Sekh et al. 2020; Teney et al. 2020; Zhang et al. 2019a) have been developed to evaluate the model capability of visual reasoning, *i.e.*, not only visually understand image scenes, but also logically conduct inductive reasoning over pictures (Barrett et al. 2018; Sekh et al. 2020; Teney et al. 2020; Zhang et al. 2019a). Most existing models for visual reasoning (Benny, Pekar, and Wolf 2021; Hu et al. 2021; Spratley, Ehinger, and Miller 2020; Zhuo and Kankanhalli 2021) contain two modules, a perception module that visually perceives the RPM panels to explicitly/implicitly extract the visual attributes, and a logic reasoning module that conducts reasoning over the perceived visual information.

The perception module in many existing models (Spratley, Ehinger, and Miller 2020; Zhang et al. 2019b; Zheng, Zha, and Wei 2019) is built upon shallow convolutional neural networks, as RPM problems are often constructed using simple visual patterns like 2D shapes and lines. But such shape patterns are combined to form complex spatial layouts in RPMs, which introduces global layout attributes such as `Position` and `Number` apart from relatively local attributes such as `Color`, `Size` and `Type`. Applying a shallow network to an image may partially capture local features, but can hardly extract the global panel layouts. Nevertheless, solving an RPM problem requires reasoning over a set of relational rules embedded in both global and local attributes. Lacking of "globality" may lose significant reasoning clues on global patterns. More importantly, real-world images are much more complicated than simple shapes. Existing models built on shallow networks may not be able to handle the tremendous combinations of complex visual recognition tasks and diversified logical reasoning tasks.

Recent approaches (Benny, Pekar, and Wolf 2021; Hu et al. 2021; Spratley, Ehinger, and Miller 2020; Zhuo and Kankanhalli 2021) often tackle the visual reasoning tasks by firstly encoding visual features through convolutions, and then optimizing the "reasoning modules" by computing the row-wise/column-wise similarities of the derived features, to detect the underlying reasoning rules. However, one unique reasoning rule can be applied to every visual attribute. After applying a combination of reasoning rules on various attributes, the resulting features may be different across rows/columns. Requesting the resulting features being similar does not explicitly model the underlying reasoning rules. Hence, it is critical to build a reasoning module that can simulate a wide range of potential reasoning rules and derive the proper rule combinations from a set of RPM images.

To tackle the challenges of RPM problems, we propose an end-to-end solution model, Hierarchical ConViT with Attention-based Relational Reasoner (HCV-ARR). Existing perception models for RPMs (Benny, Pekar, and Wolf 2021) built on convolutional neural networks cannot completely model the global dependencies. In this paper, a ConViT structure is proposed, where a convolutional block is designed to capture the low-level visual attributes, and a transformer block is designed to capture the high-level image semantics. Furthermore, we propose to hierarchically recognize the RPM panel in different levels of receptive fields. The hierarchically designed network structure can capture different aspects of the RPM panels at different scales, from overall global insights of attribute knowledge (*e.g.*, `Number` and `Position`) to specific local understandings of image details (*e.g.*, `Type` and `Size`). To conduct robust analogical reasoning based on the extracted visual features, instead of simply detecting the common recurrent patterns among rows/columns, we design an Attention-based Relational Reasoner (ARR) that dynamically learns the combination of rules applied to attributes across rows/columns. The designed element-wise attention mechanism better models the non-linear relations in each attribute among images. The proposed ARR can uncover a combination of a wide range of relational rules in inductive reasoning.

The proposed method is compared with state-of-the-art models on three benchmark datasets. It outperforms the previous best methods in most of the experimental settings on the RAVEN-FAIR (Benny, Pekar, and Wolf 2021), RAVEN (Zhang et al. 2019a) and I-RAVEN (Hu et al. 2021) datasets, as shown in Tables 3–5. The experimental results demonstrate the effectiveness of the proposed model.

Our contributions can be summarized as: 1) We propose an end-to-end Hierarchical ConViT with an Attention-based Relational Reasoner to solve RPM problems. 2) The proposed ConViT can simultaneously extract the global visual features utilizing the self-attention mechanism and local ones utilizing the shallow convolutional layers. 3) The hierarchically designed ConViTs can better understand the RPM images from different receptive fields. 4) The proposed Attention-based Relational Reasoner can well model the complex relations between rows/columns via the designed element-wise attention-based relational formulation and discover a wide range of reasoning rationales.

Related Work

Visual Reasoning. Visual reasoning visually recognizes attributes from scene images and conducts relational reasoning over the derived attributes. In literature, visual reasoning spans various tasks, *e.g.*, action recognition (Li et al. 2021; Weng et al. 2018, 2020), image captioning (Liu, Ren, and Yuan 2020; Wu et al. 2017), visual question answering (Antol et al. 2015; Johnson et al. 2017; Teney, Wu, and van den Hengel 2017; Wu et al. 2017) and visual IQ tests (Benny, Pekar, and Wolf 2021; Hu et al. 2021; Sekh et al. 2020; Teney et al. 2020; Zhang et al. 2019a; Song et al. 2023).

Activity recognition highly relies on the temporal information, and reasoning the human-object relations over time is challenging. Weng et al. (2020) recognize human activities by reasoning over the discriminative channel-level information through a Progressive Enhancement Module to avoid repeating information extraction from different frames. Image/video captioning from a relation-reasoning perspective has received increasing attention. Liu, Ren, and Yuan (2020) introduced a dual-branch Sibling Convolutional Encoder which combines visual content information with visual-semantic joint embedding using a soft-attention mechanism, and an RNN decoder to generate the captions.

Visual question answering (VQA) is a conventional visual reasoning task for machine understanding of scene-level images. The objective is to derive an accurate natural language answer, given an image and a related natural language question. Early VQAs are based on natural scene images (Antol et al. 2015; Teney, Wu, and van den Hengel 2017). Johnson et al. (2017) developed the CLEVR dataset by replacing natural images with synthetic images to avoid the misleading by background information. Recently, a new form of VQA tasks was developed by Zellers et al. (2019) as Visual Commonsense Reasoning, which aims to answer the question and provide an explanation for why the answer is correct.

Solution Models for RPMs. Visual reasoning on RPMs often consists of two parts: visual perception and logic reasoning (He et al. 2021). Solution models often use neural

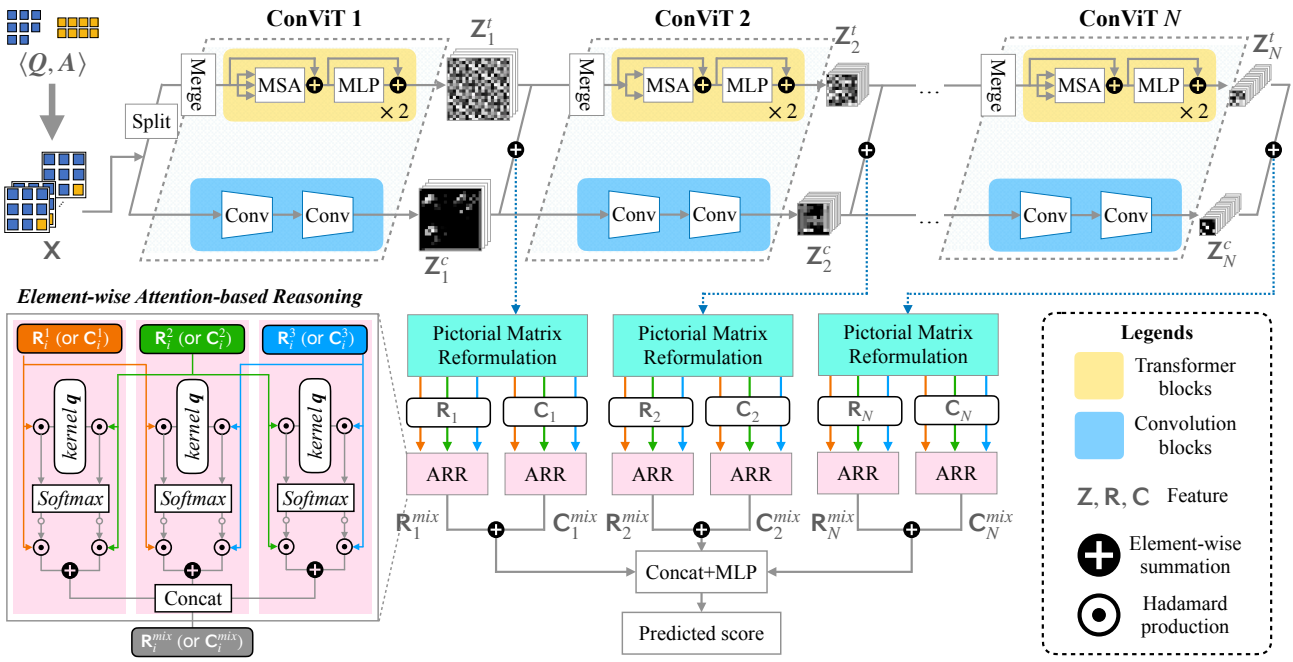

Figure 2: Block diagram of the proposed HCV-ARR, which consists of a Hierarchical ConViT and an Attention-based Relational Reasoner. It extracts the image details locally through the shallow convolutional blocks and the high-level image semantics globally through the transformer blocks. The proposed ARR extracts the element-wise attentional information between two images, uncovers the relations embedded in the image pair and conducts relational reasoning to derive the correct answer.

networks to extract visual features. CoPINet (Zhang et al. 2019b) and Rel-AIR (Spratley, Ehinger, and Miller 2020) both utilize the residual network architecture (He et al. 2016), and MRNet (Benny, Pekar, and Wolf 2021) applies multi-scale convolutional layers to extract features. The neural networks in existing methods (He, Ren, and Bai 2021; Spratley, Ehinger, and Miller 2020; Zhang et al. 2019b; Zheng, Zha, and Wei 2019; Zhuo and Kankanhalli 2021) are often relatively shallow, which may not fully capture complex combinations of visual patterns across RPM panels.

In literature, row-wise and/or column-wise relations are often utilized in solution models to derive the reasoning rules, *e.g.*, CoPINet (Zhang et al. 2019b), LEN (Zheng, Zha, and Wei 2019), MXGNet (Wang, Jamnik, and Lio 2020), Rel-AIR (Spratley, Ehinger, and Miller 2020), DCNet (Zhuo and Kankanhalli 2021) and MRNet (Benny, Pekar, and Wolf 2021). The CoPINet (Zhang et al. 2019b) explicitly contrasts the candidate answers and highlights the difference between options. The LEN (Zheng, Zha, and Wei 2019) utilizes a global encoder that encodes the context and choices to derive the row-/column-wise representations. The MXGNet (Wang, Jamnik, and Lio 2020) and the Rel-AIR (Spratley, Ehinger, and Miller 2020) subtract the common factors from all the option representations. The DCNet (Zhuo and Kankanhalli 2021) implements a dual-contrasting mechanism on both row/column features and choices. The MRNet (Benny, Pekar, and Wolf 2021) applies a multi-scale design and minimizes the squared Euclidean distance between row/column features, to identify recurring patterns. These reasoning models often operate on contrastive information derived from visual features, either within rows/columns or answer candidates, other than seeking for concrete inductive rationales for visual analogical reasoning. A much more sophisticated and comprehensive logic reasoner is needed to uncover the combination of reasoning rules embedded in different attributes of the RPM problems.

Proposed Method

Formally, given a 3×3 RPM pictorial matrix with the last one missing, $\boldsymbol{Q} = \{q_0, q_1, \cdots, q_7\}$, as shown in Fig. 1, the target is to find the missing image $\hat{a}_i$ from the answer set $\boldsymbol{A} = \{a_0, a_1, \cdots, a_7\}$, forming as a complete RPM sample $\langle \boldsymbol{Q}, \boldsymbol{A} \rangle$ with each image of the size of $H \times W$. The same reasoning rule regarding one attribute is shared among three rows or columns. Note that for one RPM sample, the underlying rules for different attributes can be different. In modern RPM solvers, each option image is appended to $\boldsymbol{Q}$, forming 8 groups of image tensors $\mathbf{X}_i \in \mathbb{R}^{9 \times H \times W}$, where $i = 0 \ldots 7$ indicates the option index.

As shown in Fig. 2, the proposed model consists of a Hierarchical ConViT for visual perception and an Attention-based Relational Reasoner for logic reasoning. The former visually perceives the image contents and extracts the visual information, and the latter conducts reasoning on the extracted visual information. The goal of the established network structure is to exploit a discriminative relational mapping that takes the image tensors as the input and outputs the prediction score vector for eight options:

$$\hat{\boldsymbol{y}} = \mathcal{F}_m\{\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_7; \boldsymbol{\Theta}\}, \tag{1}$$

where $\mathcal{F}_m$ is the mapping function, which involves the Hierarchical ConViT for perception and Attention-based Relational Reasoner for reasoning. $\mathbf{\Theta}$ is the network parameters to be optimized. $\hat{\boldsymbol{y}} \in [0,1]^8$ contains the prediction scores of the 8 candidate options. The final output is the option which has the minimum Binary Cross Entropy loss,

$$\mathcal{L} = -\sum_{i=0}^{7} \sigma(y_i) \cdot \log \sigma(\hat{y}_i), \tag{2}$$

where σ represents the sigmoid function and y_i is the ground-truth label concerning the i-th option, and it is an element of the one-hot vector $\boldsymbol{y} \in \{0,1\}^8$.

Hierarchical ConViT for Visual Perception

As shown in Fig. 2, the proposed Hierarchical ConViT encoder $\mathcal{E}$ contains a set of ConViT blocks to extract the visual information from multiple receptive fields, and each ConViT consists of a transformer branch and a convolutional branch. Both low-level image details and high-level image semantics have been shown useful for visual reasoning (Benny, Pekar, and Wolf 2021; Sekh et al. 2020). In Fig. 2, while feature maps of shallow blocks focus more on lower-level semantic relations (Girshick et al. 2014) such as `Color` and `Type`, those of deeper blocks reflect the pattern layout and contain positional relations (Zintgraf et al. 2017; Islam, Jia, and Bruce 2020). The proposed Hierarchical ConViT could concurrently capture the visual information from different receptive fields, which is used as multi-receptive clues for logical reasoning at the later stage.

Existing visual reasoning models (Spratley, Ehinger, and Miller 2020; Zhang et al. 2019b; Zheng, Zha, and Wei 2019; Zhuo and Kankanhalli 2021) often utilize shallow convolutional networks to extract low-level image details, but these shallow networks may not capture the complex visual patterns and spatial correlations across different configurations in RPMs. On the other hand, transformer models (Han et al. 2022) perform well on image classification tasks by partitioning the image into sequences of patches and extracting the attentional information among patches. The question image of RPMs is often formed in patches, *e.g.*, there are 3×3 patches in the `3×3Grid` setting and 2×2 patches in the `2×2Grid` setting in the RAVEN-like dataset. It is hence natural to apply Vision Transformers (ViT) (Dosovitskiy et al. 2021) to better recognize the complex global patterns embedded in RPMs. In the design of ConViT, we encode the image fine details through the shallow convolutional neural networks in one branch, and exploit the spatial correlations and long-range dependencies between complex patterns in another branch through the patch split-and-merge operations and multi-head self-attention.

More specifically, denote $\mathbf{X} \in \mathbb{R}^{8\times 9\times H\times W}$ as the input image tensors, where H and W are the height and width of the image respectively. The output of the convolutional branch $\mathcal{C}_i$ at the i-th stage, $\mathbf{Z}_i^c$, can be derived as:

$$\begin{aligned} \mathbf{Z}_1^c &= \mathcal{C}_1\{\mathbf{X}\}, \\ \mathbf{Z}_i^c &= \mathcal{C}_i\{\mathbf{Z}_{i-1}^c\}. \end{aligned} \tag{3}$$

The output of the transformer branch $\mathcal{T}_i$ at the i-th stage, $\mathbf{Z}_i^t$, can be derived as:

$$\begin{aligned} \mathbf{Z}_0^t &= \mathcal{P}^S\{\mathbf{X}\}, \\ \mathbf{Z}_i^t &= \mathcal{T}_i\{\mathcal{P}_i^M\{\mathbf{Z}_{i-1}^t\}\}, \end{aligned} \tag{4}$$

where $\mathcal{P}^S, \mathcal{P}_i^M$ and $\mathcal{T}_i$ denote the patch splitting operation, the patch merging operation and the transformer at the i-th stage, respectively. $\mathcal{P}_i^M$ contains an unfold layer for down-sampling and a multi-layer perception. The transformer block $\mathcal{T}_i$ propagates a feature map $\mathbf{Z}_{i-1}^t$ and outputs $\mathbf{Z}_i^t$ at stage i as in Fig. 2. The details of model architecture and parameters can be found in Supplementary Materials.

The final output of the encoder $\mathcal{E}$ is a tuple of tensors, recording the feature maps from both convolution blocks and transformer blocks, and concurrently passes it to next reasoning modules for further abstraction.

$$(\mathbf{Z}_1^c + \mathbf{Z}_1^t, \ldots, \mathbf{Z}_N^c + \mathbf{Z}_N^t) = \mathcal{E}\{\mathbf{X}\}. \tag{5}$$

Attention-based Relational Reasoner

Before feeding the perceived visual information to the reasoner, we firstly reformulate the derived features according to rows or columns via Pictorial Matrix Reformulation. Specifically, given $\mathbf{Z}_i = \mathbf{Z}_i^c + \mathbf{Z}_i^t, \mathbf{Z}_i \in \mathbb{R}^{8\times 9\times C_i\times H_i\times W_i}$ for stage i, where C_i, H_i, W_i are shown in Fig. 2, we reformulate the second dimension of the feature tensor into a 3×3 matrix. For notation simplicity, stage index i is omitted. Next, features among rows and columns from the matrix are formed as row features $\mathbf{R} = [\mathbf{R}_1; \mathbf{R}_2; \mathbf{R}_3]$ and column features $\mathbf{C} = [\mathbf{C}_1; \mathbf{C}_2; \mathbf{C}_3]$, respectively, where $\mathbf{R}_1$, $\mathbf{R}_2$, $\mathbf{R}_3$, $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$ are in shape of $(8, 3, C_i, H_i, W_i)$.

Humans often solve visual analogical reasoning tasks by investigating the common rationale spanning over image patterns along rows or columns. The majority of solution models (Benny, Pekar, and Wolf 2021; Spratley, Ehinger, and Miller 2020; Zhang et al. 2019b; Zheng, Zha, and Wei 2019; Zhuo and Kankanhalli 2021) attempt to model this process by minimizing the divergence between row/column feature representations. For example, in (Benny, Pekar, and Wolf 2021), the DIST3 operation within row/column vectors computes the squared Euclidean distance among features. Take row features $\mathbf{R} = [\mathbf{R}_1; \mathbf{R}_2; \mathbf{R}_3]$ as an example,

$$\begin{aligned} &\text{DIST3}(\mathbf{R}_1; \mathbf{R}_2; \mathbf{R}_3) \\ &= |\mathbf{R}_1 - \mathbf{R}_2|^2 + |\mathbf{R}_2 - \mathbf{R}_3|^2 + |\mathbf{R}_3 - \mathbf{R}_1|^2. \end{aligned} \tag{6}$$

Such formulation is optimal only for some special cases. However, a combination of rules applied on different attributes of the problem panel do not necessarily lead to similar image features, *e.g.*, for `Arithmetic` rule, image attributes satisfy some arithmetic relations, and for `Distribute_Three` rule, three distinct attribute values are distributed in images of one row/column. The formulation in Eqn. (6) is not sufficiently expressive to model such relations. To better model the embedded reasoning rules, we propose an Attention-based Relational Reasoner.

Attention mechanism has been widely used in modeling the pairwise relations between two inputs to explore the in-depth inter-relations at the feature level (Chen et al. 2019,

2022; Hori et al. 2017). In RPMs, different inductive rules are applied to different attributes. Therefore, unlike the traditional attention mechanism where all dimensions of the input feature vector share the same learnable weight (Chen et al. 2019), the proposed method weighs each dimension differently to better model the relations across different attributes. As shown in details from Fig. 2, the element-wise attentive relations are learned through a kernel $\boldsymbol{q}$ which has the same dimensionality of the input row/column features. Take row features $\mathbf{R}$ for instance,

$$\mathbf{W}_{\mathsf{i},\mathsf{j}} = \exp(\boldsymbol{q} \odot \mathbf{R}_{\mathsf{i}}) \oslash (\exp(\boldsymbol{q} \odot \mathbf{R}_{\mathsf{i}}) + \exp(\boldsymbol{q} \odot \mathbf{R}_{\mathsf{j}})), \quad (7)$$

where $\odot$ denotes the element-wise (Hadamard) production, $\oslash$ denotes element-wise division, and $\mathsf{i}, \mathsf{j} \in \{1, 2, 3\}$ denotes the row index. The output denoted as $\mathbf{R}^{\mathsf{mix}}$ is obtained as,

$$\begin{aligned} \mathbf{R}^{\mathsf{mix}} = [&\mathbf{W}_{1,2} \odot \mathbf{R}_1 + \mathbf{W}_{2,1} \odot \mathbf{R}_2; \\ &\mathbf{W}_{1,3} \odot \mathbf{R}_1 + \mathbf{W}_{3,1} \odot \mathbf{R}_3; \\ &\mathbf{W}_{2,3} \odot \mathbf{R}_2 + \mathbf{W}_{3,2} \odot \mathbf{R}_3]. \end{aligned} \quad (8)$$

The same operation is applied to the column features $\mathbf{C}$ to derive $\mathbf{C}^{\mathsf{mix}}$. Finally, the predicted scores are aggregated over N stages through a multilayer perceptron $\mathcal{F}_{\mathrm{MLP}}$ as,

$$\hat{\boldsymbol{y}} = \mathcal{F}_{\mathrm{MLP}}\left(\left[\mathbf{R}_1^{\mathsf{mix}} + \mathbf{C}_1^{\mathsf{mix}}; \ldots; \mathbf{R}_N^{\mathsf{mix}} + \mathbf{C}_N^{\mathsf{mix}}\right]\right). \quad (9)$$

Discussion

Attention in Visual Perception. Han et al. (2022) categorizes image representation learning into convolution-based and attention-based. With the development of SENet (Hu, Shen, and Sun 2018), Vision Transformer (ViT) (Dosovitskiy et al. 2021) and its variants, the attention mechanism overwhelms traditional convolutional architectures in substantial CV tasks. ViT is good at capturing long-range dependencies between patches, and its variants such as Swin-T improve the modeling capacity for local information (Han et al. 2022). The proposed HCV simultaneously captures both global and local discriminative information in multiple receptive fields to deeply understand the image contents.

Attention in Relational Reasoning. In analogical reasoning tasks, Relation Network (Santoro et al. 2017) utilizing neural networks has been used to derive the relations of input features for various relational reasoning tasks (Barrett et al. 2018; Santoro et al. 2017). Inspired by the success of the attention mechanism in extracting the attentive information among elements in a sequence (Vaswani et al. 2017), we propose an attention mechanism to deeply explore the relations on feature sequences, and subsequently derive the underlying reasoning rules. Traditional attention-based feature fusion scheme in Two-Stream Convolutional Neural Network (TSCNN) (Chen et al. 2019) has been proven effective in aggregating two sets of features, which uses the standard self-attention to learn a set of weights $\{w_i, i = 1, 2, \ldots, M\}$ corresponding to M sets of features $\{\boldsymbol{f}_i, i = 1, 2, \ldots, M\}$ to generate the aggregated feature $\boldsymbol{f}_a = \sum_i^M \sigma(\boldsymbol{q}^\top \boldsymbol{f}_i)\boldsymbol{f}_i$. To fuse the two feature representations in (Chen et al. 2019), $\boldsymbol{f}_a = w_1\boldsymbol{f}_1 + w_2\boldsymbol{f}_2$, the same weight w_i is assigned to every dimension of the feature vector. Such a mechanism may well fuse multi-modal feature representations, but may be ineffective for modeling high-level complex relations using one learnable weight only, as different rules may be applied on different attributes. In contrast, the proposed element-wise attention-based relational reasoning mechanism assigns a different learnable weight to each dimension, which has a high potential to model the complicated relations embedded in different feature dimensions, and consequently induce a wide range of abstract relations across different attributes.

Experimental Results

The proposed model is evaluated on the RAVEN (Zhang et al. 2019a), I-RAVEN (Hu et al. 2021) and RAVEN-FAIR datasets (Benny, Pekar, and Wolf 2021). Each dataset is randomly split into 10 folds, with 6 folds for training, 2 folds for validation and 2 folds for testing.

Datasets

RAVEN (Zhang et al. 2019a) consists of 70,000 question sets, where each contains 8 question images, arranged as a 3×3 image matrix with the last one missing, and 8 candidate answers. The candidates are generated by permutation from the ground-truth answer image, and each permutated image is derived by randomly shifting one attribute value. The dataset is equally distributed into 7 configurations. Each question contains 6 visual attributes (`Angle`, `Number`, `Position`, `Type`, `Size` and `Color`) and 4 underlying rules (`Constant`, `Progression`, `Arithmetic` and `Distribute_Three`). Extra noise is added to attributes to make the problem more challenging.

I-RAVEN (Hu et al. 2021) is developed to fix the problem in the original RAVEN dataset that the aggregation of the most common values for each attribute could be the correct answer (Benny, Pekar, and Wolf 2021; Hu et al. 2021). In the I-RAVEN dataset, the negative candidate answers are generated by hierarchically permuting one attribute of the ground-truth answer in three iterations. In each iteration, two child nodes are generated, where one node remains the same with the parent node while the other permutes one attribute.

RAVEN-FAIR (Benny, Pekar, and Wolf 2021) iteratively enlarges the answer set starting with the correct answer only and changing one attribute value from either the correct answer or a generated negative answer. Except for the answer generation, the same settings as in original RAVEN are used for the I-RAVEN and RAVEN-FAIR datasets.

Experimental Setup

The proposed method is compared with the following state-of-the-art solutions.

CoPINet (Zhang et al. 2019b) models the probability of each candidate answer by applying a contrasting module.

LEN (Zheng, Zha, and Wei 2019) assembles the possible candidate answer embeddings to the 8 question panel embeddings, calculates scores for all possible combinations (C_9^3 = 84), and predicts the answer with the highest score.

Rel-AIR (Spratley, Ehinger, and Miller 2020) disentangles objects with an initial unsupervised scene decomposition

first, and then encodes it with additional information such as position and scale. A sequence encoder is designed to extract feature relationships and generate the final results.

SRAN (Hu et al. 2021) utilizes a hierarchical rule embedding module and a gated embedding fusion module to output the rule embedding given two-row sequences.

DCNet (Zhuo and Kankanhalli 2021) consists of a rule contrast module and a choice contrast module to exploit the inherent structure of RPMs, which compares the latent rules among rows/columns and increases the choices differences.

MRNet (Benny, Pekar, and Wolf 2021) is established by using multi-resolution convolution layers as the perception module, and the DIST3 row/column operator as the inductive reasoner to conduct relational reasoning.

The number of ConViT blocks is set to $N = 3$. Following the settings in (Benny, Pekar, and Wolf 2021; Zhang et al. 2019a,b), the input images are resized to 80×80 pixels. The maximum number of epochs is 200, and the training is stopped if there is no significant improvement on the validation set over 20 epochs. During training, the learning rate is set to 0.001, and the Adam optimizer is utilized with a weight decay of 1×10^{-5}. The batch size is set to 32.

Ablation Study

To evaluate the performance gains brought by each of the two proposed modules, we conduct an ablation study on the RAVEN-FAIR dataset (Benny, Pekar, and Wolf 2021), by using the most recent state-of-the-art method MRNet (Benny, Pekar, and Wolf 2021) as the baseline model. **HCV** and **ARR** denote methods of replacing the visual perception module or reasoning module of MRNet with the proposed HCV and ARR, respectively. Besides using just convolutions for perception (*i.e.*, Conv + ARR), the usage of ViT only is also evaluated (*i.e.*, ViT + ARR). We also compare the proposed ARR with the traditional attention-based fusion utilized in TSCNN (Chen et al. 2019) on features extracted from the proposed HCV. The results are in Table 1.

Compared with MRNet, all the proposed model components receive consistent performance improvements. For Setting `C`, `L-R`, `U-D` and `O-IC` that do not use high-level image semantics such as `Number` and `Position` (The regular shapes have a fixed `Position` and `Number`) for reasoning, using just convolutions works well, which demonstrates the effectiveness of the convolutions in capturing low-level features for reasoning. Using ViT alone produces very poor results, as most of the attributes, e.g., `Color`, `Type`, `Size` are low-level features, while ViT focuses more on the high-level features such as `Number` and `Position`. When both convolutions and ViTs are used, the proposed ConViT significantly improves the performance on complex configurations such as `2×2G` and `3×3G`, by utilizing the high-level image semantics extracted from ViT blocks.

Methods	Accuracy (%) in Different Configurations							
	Avg.	C	2x2	3x3	L-R	U-D	OIC	OIG
MRNet	86.8	97.0	72.7	69.5	98.7	98.9	97.6	73.3
TSCNN	87.9	96.8	73.5	74.7	94.4	91.8	94.5	88.9
HCV	92.7	99.9	85.7	78.4	99.9	99.8	99.8	85.4
Conv+ARR	93.4	99.9	86.3	79.8	99.8	99.7	99.6	88.7
ViT+ARR	44.9	52.0	39.4	41.0	35.8	35.8	57.8	50.4
Proposed	95.4	99.8	92.9	87.9	99.8	99.6	99.7	88.5

Table 1: Ablation study on different modules of the proposed architecture. Compared with the baseline method, MRNet (Benny, Pekar, and Wolf 2021), both hierarchical ConViT and ARR bring significant performance improvements across all 7 problem configurations.

Stages 1 2 3	Accuracy (%) in Different Configurations							
	Avg.	C	2x2	3x3	L-R	U-D	OIC	OIG
✓✗✗	74.9	83.2	53.1	58.1	90.6	90.3	87.4	61.7
✓✓✗	87.8	98.9	68.1	68.7	99.3	99.2	99.1	78.1
✓✓✓	95.4	99.8	92.9	87.9	99.8	99.6	99.7	88.5

Table 2: Ablation study of the depth of the proposed HCV-ARR on the RAVEN-FAIR dataset.

By introducing the HCV module, the hierarchically increased receptive fields can view the question images from multiple scales, and on each scale both local and global features are richly captured. Hence, the inference performance on all 7 configurations is improved upon the baseline. By using the proposed ARR module, the performance across 7 different settings is also improved upon MRNet, which indicates the benefits brought by the proposed ARR in handling the diverse reasoning rules across both local and global attributes of various scales. The proposed ARR also achieves a significant improvement over the traditional attention mechanism of TSCNN (Chen et al. 2019), which assigns the same weight across different attributes, while the proposed ARR can capture different rules embedded in different attributes.

Additionally, we conduct ablations on the impact of the depth for the proposed HCV-ARR. From Table 2, we can observe the improvements in the reasoning accuracy by considering more receptive fields from deeper ConViT blocks and more reasoning clues from deeper ARRs. The accuracy increases by 12.9% on average when the depth N is set from 1 to 2, and further boosts to 95.4% when $N = 3$. The experimental results show the benefits of utilizing both global and local attributes in visual analogical reasoning.

Comparison Results on RAVEN-FAIR Dataset

The proposed HCV-ARR is compared with state-of-the-art models on the RAVEN-FAIR dataset (Benny, Pekar, and Wolf 2021). We implement and evaluate DCNet (Zhuo and Kankanhalli 2021) and SRAN (Hu et al. 2021) on the RAVEN-FAIR dataset. The results of other compared methods are obtained from (Benny, Pekar, and Wolf 2021). From the results summarized in Table 3, we can see that the proposed method significantly outperforms all the compared methods on all problem configurations. Compared with the previous best model, MRNet (Benny, Pekar, and Wolf 2021), the proposed HCV-ARR achieves a performance gain of 8.6% on average. The proposed method is particularly competitive for challenging settings, *e.g.*, the performance gains

Methods	Accuracy (%) in Different Configurations							
	Avg.	C	2x2	3x3	L-R	U-D	OIC	OIG
CoPINet	36.5	–	–	–	–	–	–	–
LEN	50.9	–	–	–	–	–	–	–
DCNet†	57.0	57.2	48.4	58.2	57.5	59.4	62.0	56.2
SRAN†	76.7	87.4	60.4	62.8	86.5	86.7	77.5	75.9
MRNet	86.8	97.0	72.7	69.5	98.7	98.9	97.6	73.3
Proposed	**95.4**	**99.8**	**92.9**	**87.9**	**99.8**	**99.6**	**99.7**	**88.5**

Table 3: Comparison with state-of-the-art on the RAVEN-FAIR dataset (Benny, Pekar, and Wolf 2021). Results of other methods are obtained from (Benny, Pekar, and Wolf 2021) and † indicates the results by our implementation.

Methods	Accuracy (%) in Different Configurations							
	Avg.	C	2x2	3x3	L-R	U-D	OIC	OIG
CoPINet	46.1	54.4	36.8	31.9	51.9	52.5	52.2	42.8
LEN	41.4	56.4	31.7	29.7	44.2	44.2	52.1	31.7
DCNet†	46.6	56.2	32.7	32.9	54.7	53.9	55.9	39.8
SRAN	60.8	78.2	50.1	42.4	70.1	70.3	68.2	46.3
MRNet†	81.0	99.6	63.4	59.2	98.7	98.3	95.7	51.9
Proposed	**93.9**	**99.9**	**96.2**	**75.5**	**99.4**	**99.6**	**99.5**	**87.3**

Table 4: Comparison with state-of-the-art models on the I-RAVEN dataset. † indicates the results are obtained by us and others are from (Hu et al. 2021).

are 20.2% on 2×2Grid, 18.4% on 3×3Grid, and 15.2% on Out-InGrid settings, respectively. The underlying reasons are two-fold: 1) The proposed HCV module could better perceive the complex patterns in these settings. 2) The proposed ARR could better reason over the combination of rules, with one applied to each attribute.

Comparison Results on I-RAVEN Dataset

We have implemented and evaluated the DCNet (Zhuo and Kankanhalli 2021) and MRNet (Benny, Pekar, and Wolf 2021) on the I-RAVEN dataset, and other results are obtained from (Hu et al. 2021). As shown in Table 4, the proposed model largely outperforms all the compared models. Compared to the previously published best result by SRAN (Hu et al. 2021), the average accuracy improves from 60.8% to 93.9%. The proposed HCV-ARR significantly and consistently outperforms MRNet (Benny, Pekar, and Wolf 2021) on all the settings. Besides the Center, Left-Right, Up-Down and Out-InCenter, HCV-ARR can also effectively solve configurations containing complex spatial relations, such as 2×2Grid and Out-InGrid.

Comparison Results on Original RAVEN Dataset

We also conduct comparison experiments on the original RAVEN dataset (Zhang et al. 2019a). The original RAVEN dataset contains the loophole that the correct answer can be derived by simply aggregating the most common properties from the answer options, without examining the question at all (Hu et al. 2021; Benny, Pekar, and Wolf 2021). In the past, many approaches took advantage of this shortcut to derive a good performance, *e.g.*, by contrasting the candidate answers (Zhang et al. 2019b). We conduct comparison experiments using both settings, with or without the contrast information. The results are summarized in Table 5.

Methods	Accuracy (%) in Different Configurations							
	Avg.	C	2x2	3x3	L-R	U-D	OIC	OIG
CoPINet	18.4	–	–	–	–	–	–	–
SRAN†	46.2	49.0	45.4	52.8	42.4	36.0	49.1	48.8
MRNet	84.0	–	–	–	–	–	–	–
Proposed	**87.3**	**99.8**	**71.4**	**65.9**	**99.9**	**99.8**	**98.0**	**76.2**

(a) Without contrasting on candidate answers.

Methods	Accuracy (%) in Different Configurations							
	Avg.	C	2x2	3x3	L-R	U-D	OIC	OIG
CoPINet	91.4	95.1	77.5	78.9	99.1	99.7	98.5	91.4
LEN	72.9	80.2	57.5	62.1	73.5	81.2	84.4	71.5
Rel-AIR	94.1	99.0	**92.4**	87.1	98.7	97.9	98.0	85.3
DCNet	93.6	97.8	81.7	86.7	99.8	99.8	99.0	91.5
MRNet	**96.6**	–	–	–	–	–	–	–
Proposed	96.0	**99.4**	86.9	**89.1**	**99.9**	**99.9**	**99.8**	**96.8**

(b) With contrasting on candidate answers.

Table 5: Comparison with state-of-the-art on the RAVEN dataset. † indicates the results by our implementation and others are obtained from respective original publications. The proposed HCV-ARR outperforms the previous best method, MRNet, without using the contrast while achieves a comparable performance when using the contrast.

From Table 5b, we can see that many approaches utilizing the contrast information achieve high accuracy. When the proposed HCV-ARR utilizes the contrast information, it achieves slightly poorer than the previous best method, MRNet (Benny, Pekar, and Wolf 2021), but outperforms other methods. When this loophole is eliminated, the proposed method outperforms MRNet (Benny, Pekar, and Wolf 2021) by 3.3%, and significantly outperforms other compared methods as shown in Table 5a. These experimental results validate the effectiveness of the proposed method.

Conclusions

In this paper, a Hierarchical ConViT with Attention-based Relational Reasoner is proposed to solve RPM problems. The proposed Hierarchical ConViT simultaneously extracts the fine image details and global image semantics through shallow convolutional networks and the attention mechanism of vision transformers across multi-level receptive fields. The proposed ARR module effectively models the underlying relations via the designed element-wise attention mechanism, one rule for each attribute, and discovers a wide range of reasoning rationales for better reasoning. The experimental results on three benchmark datasets demonstrate that the proposed HCV-ARR significantly outperforms the state-of-the-art models in almost all the settings.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 72071116, and in part by the Ningbo Municipal Bureau Science and Technology under Grants 2019B10026 and 2022Z173.

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *ICCV*, 2425–2433.

Barrett, D. G.; Hill, F.; Santoro, A.; Morcos, A.; and Lillicrap, T. 2018. Measuring Abstract Reasoning in Neural Networks. In *ICML*, volume 80, 511–520.

Benny, Y.; Pekar, N.; and Wolf, L. 2021. Scale-Localized Abstract Reasoning. In *CVPR*, 12557–12565.

Chen, H.; Hu, G.; Lei, Z.; Chen, Y.; Robertson, N. M.; and Li, S. Z. 2019. Attention-based Two-stream Convolutional Networks for Face Spoofing Detection. *IEEE TIFS*, 15: 578–593.

Chen, S.; He, W.; Ren, J.; and Jiang, X. 2022. Attention-based Dual-stream Visual Transformer for Radar Gait Recognition. In *ICASSP*, 3668–3672.

Crouse, M.; Nakos, C.; Abdelaziz, I.; and Forbus, K. 2021. Neural Analogical Matching. In *AAAI*, volume 35, 809–817.

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 *ICLR*.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *CVPR*, 580–587.

Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; Yang, Z.; Zhang, Y.; and Tao, D. 2022. A Survey on Vision Transformer. *IEEE TPAMI*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778.

He, W.; Ren, J.; and Bai, R. 2021. Data Augmentation by Morphological Mixup for Solving Raven's Progressive Matrices. *arXiv preprint arXiv:2103.05222*.

He, W.; Ren, J.; Bai, R.; and Jiang, X. 2021. Two-stage Rule-induction Visual Reasoning on RPMs with an Application to Video Prediction. *arXiv preprint arXiv:2111.12301*.

Hori, C.; Hori, T.; Lee, T.-Y.; Zhang, Z.; Harsham, B.; Hershey, J. R.; Marks, T. K.; and Sumi, K. 2017. Attention-based Multimodal Fusion for Video Description. In *ICCV*, 4193–4202.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-Excitation Networks. In *CVPR*, 7132–7141.

Hu, S.; Ma, Y.; Liu, X.; Wei, Y.; and Bai, S. 2021. Stratified Rule-Aware Network for Abstract Visual Reasoning. In *AAAI*, volume 35, 1567–1574.

Islam, M. A.; Jia, S.; and Bruce, N. D. 2020. How Much Position Information Do Convolutional Neural Networks Encode? In *ICLR*.

Johnson, J.; Hariharan, B.; Van Der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; and Girshick, R. 2017. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *CVPR*, 2901–2910.

Li, H.; Jiang, X.; Guan, B.; Tan, R. R. M.; Wang, R.; and Thalmann, N. M. 2021. Joint Feature Optimization and Fusion for Compressed Action Recognition. *IEEE TIP*, 30: 7926–7937.

Liu, S.; Ren, Z.; and Yuan, J. 2020. SibNet: Sibling convolutional encoder for video captioning. *IEEE TPAMI*, 43(9): 3259–3272.

Raven, J. 2000. The Raven's Progressive Matrices: Change and Stability over Culture and Time. *Cognitive psychology*, 41(1): 1–48.

Santoro, A.; Raposo, D.; Barrett, D. G.; Malinowski, M.; Pascanu, R.; Battaglia, P.; and Lillicrap, T. 2017. A Simple Neural Network Module for Relational Reasoning. In *NeurIPS*, 4967–4976.

Sekh, A. A.; Dogra, D. P.; Kar, S.; Roy, P. P.; and Prasad, D. K. 2020. Can We Automate Diagrammatic Reasoning? *PR*, 106: 107412.

Song, X.; Jin, J.; Yao, C.; Wang, S.; Ren, J.; and Bai, R. 2023. Siamese-Discriminant Deep Reinforcement Learning for Solving Jigsaw Puzzles with Large Eroded Gaps. In *AAAI*.

Spratley, S.; Ehinger, K.; and Miller, T. 2020. A Closer Look at Generalisation in RAVEN. In *ECCV*, 601–616.

Teney, D.; Wang, P.; Cao, J.; Liu, L.; Shen, C.; and van den Hengel, A. 2020. V-PROM: A Benchmark for Visual Reasoning using Visual Progressive Matrices. In *AAAI*, volume 34, 12071–12078.

Teney, D.; Wu, Q.; and van den Hengel, A. 2017. Visual Question Answering: A Tutorial. *IEEE Sign. Process. Magazine*, 34: 63–75.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is All You Need. In *NeurIPS*, 5998–6008.

Wang, D.; Jamnik, M.; and Lio, P. 2020. Abstract Diagrammatic Reasoning with Multiplex Graph Networks. In *ICLR*.

Weng, J.; Liu, M.; Jiang, X.; and Yuan, J. 2018. Deformable pose traversal convolution for 3D action and gesture recognition. In *ECCV*, 136–152.

Weng, J.; Luo, D.; Wang, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; Jiang, X.; and Yuan, J. 2020. Temporal distinct representation learning for action recognition. In *ECCV*, 363–378.

Wu, Q.; Shen, C.; Wang, P.; Dick, A.; and Van Den Hengel, A. 2017. Image captioning and visual question answering based on attributes and external knowledge. *IEEE TPAMI*, 40(6): 1367–1381.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In *CVPR*, 6720–6731.

Zhang, C.; Gao, F.; Jia, B.; Zhu, Y.; and Zhu, S.-C. 2019a. RAVEN: A Dataset for Relational and Analogical Visual rEasoNing. In *CVPR*, 5317–5327.

Zhang, C.; Jia, B.; Gao, F.; Zhu, Y.; Lu, H.; and Zhu, S.-C. 2019b. Learning Perceptual Inference by Contrasting. In *NeurIPS*, 1075–1087.

Zhang, J.; Zhang, Q.; Ren, J.; Zhao, Y.; and Liu, J. 2022. Spatial-context-aware Deep Neural Network for Multi-class Image Classification. In *ICASSP*, 1960–1964.

Zheng, K.; Zha, Z.-J.; and Wei, W. 2019. Abstract Reasoning with Distracting Features. In *NeurIPS*, 5842–5853.

Zhuo, T.; and Kankanhalli, M. 2021. Effective Abstract Reasoning with Dual-Contrast Network. In *ICLR*.

Zintgraf, L. M.; Cohen, T. S.; Adel, T.; and Welling, M. 2017. Visualizing Deep Neural Network Decisions: Prediction Difference Analysis. In *ICLR*.