COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13347

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13347

Two Computational Approaches to Visual Analogy: Task-Specific Models Versus Domain-General Mapping

Nicholas Ichien,[a,#] Qing Liu,[b,#] Shuhao Fu,[c] Keith J. Holyoak,[c,d] Alan L. Yuille,[e,f] Hongjing Lu[c,g]

[a]*Department of Psychology, University of Pennsylvania*
[b]*Adobe Research*
[c]*Department of Psychology, University of California*
[d]*Brain Research Institute, University of California*
[e]*Department of Computer Science, Johns Hopkins University*
[f]*Department of Cognitive Science, Johns Hopkins University*
[g]*Department of Statistics, University of California*

Received 11 May 2022; received in revised form 21 May 2023; accepted 8 September 2023

Abstract

Advances in artificial intelligence have raised a basic question about human intelligence: Is human reasoning best emulated by applying task-specific knowledge acquired from a wealth of prior experience, or is it based on the domain-general manipulation and comparison of mental representations? We address this question for the case of visual analogical reasoning. Using realistic images of familiar three-dimensional objects (cars and their parts), we systematically manipulated viewpoints, part relations, and entity properties in visual analogy problems. We compared human performance to that of two recent deep learning models (Siamese Network and Relation Network) that were directly trained to solve these problems and to apply their task-specific knowledge to analogical reasoning. We also developed a new model using part-based comparison (PCM) by applying a domain-general mapping procedure to learned representations of cars and their component parts. Across four-term analogies (Experiment 1) and open-ended analogies (Experiment 2), the domain-general PCM model, but not the task-specific deep learning models, generated performance similar in key aspects to that of human reasoners. These findings provide evidence that human-like analogical reasoning is unlikely to be achieved by applying deep learning with big data to a specific type of analogy problem. Rather, humans do (and machines might) achieve analogical reasoning by learning representations that encode structural information useful for multiple tasks, coupled with efficient computation of relational similarity.

Keywords: Analogy; Visual reasoning; Relations; Computational modeling; Deep learning

[#]Joint first authors.

Correspondence should be sent to Nicholas Ichien, Department of Psychology, University of Pennsylvania, 3720 Walnut St., Philadelphia, PA 19104, USA. E-mail: nichien@sas.upenn.edu

1. Introduction: Two computational approaches to analogy

Despite the many recent advances in work on artificial intelligence (AI), capturing core characteristics of human thinking remains an elusive goal. One of the canonical examples of human intelligence is the ability to reason by analogy—recognizing similarities based on *relations* between entities, even when the entities themselves are dissimilar. Computational models of analogical reasoning developed in cognitive science and AI fall into two broad classes (see Fig. 1). An early proposal in cognitive science construed analogy-making as "high-level-perception," a highly context-dependent process in which building representations, and the processes operating over those representations, are in principle inseparable (Chalmers, French, & Hofstadter, 1992). Early advocates never developed a computational model that actually implemented this theoretical position, instead relying on hand-coded representations tailored to toy problems (Hofstadter & Mitchell, 1994). However, contemporary AI work on visual analogy has succeeded in modeling the acquisition of perceptual representations suitable for analogical reasoning by end-to-end manner training from raw inputs of stimuli coded as image pixels (Santoro et al., 2017). Such task-specific models generally assume that training and testing data consist of independent and identically distributed (i.i.d.) samples from an unknown probability distribution. This approach has been applied with some success to solving visual analogy problems inspired by Raven's Progressive Matrices (RPM) (Raven, 1938). After extensive training with RPM-like problems, deep neural networks have achieved human-level performance on test problems with similar basic structure (Hill, Santoro, Barrett, Morcos, & Lillicrap, 2019; Zhang, Gao, Jia, Zhu, & Zhu, 2019).

Early critics of this high-level-perception approach to analogy articulated a number of theoretical reasons to doubt that it accurately characterizes mechanisms underlying human analogical reasoning (Forbus, Gentner, Markman, & Ferguson, 1998). However, contemporary

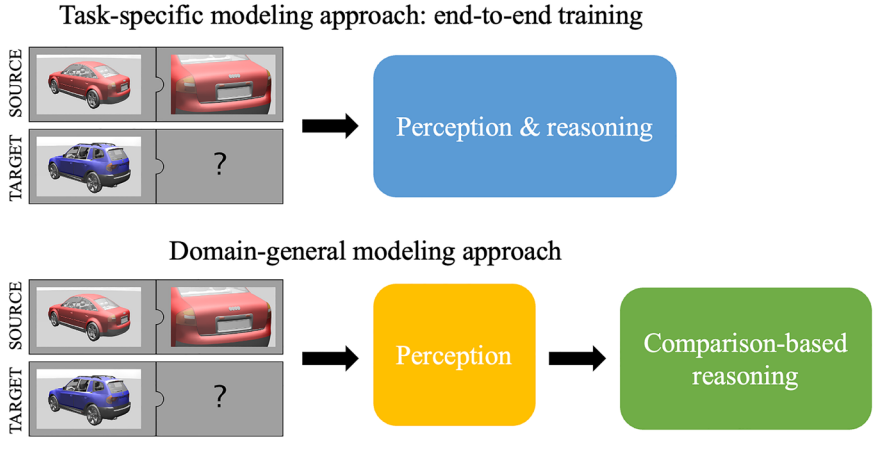

Fig. 1. Overview of the two modeling approaches. The task-specific modeling approach uses end-to-end training with a large number of analogy problems. The domain-general modeling approach replies on perceptual representations learned for visual tasks, coupled with comparisons between representational structures.

AI models based on end-to-end learning offer much more complete and powerful instantiations of the basic approach. In the present paper, we evaluate two leading examples of these models, focusing on a general limitation that such models face: difficulty in generalization. Specifically, the success of these models depends on high similarity between training problems and test problems (due in part to their assumption of i.i.d samples), and on datasets of massive numbers of RPM-like problems (e.g., 1.42 million problems in the PGM dataset (Barrett, Hill, Santoro, Morcos, & Lillicrap, 2018), and 70,000 problems in the RAVEN dataset (Zhang et al., 2019). For example, Zhang and colleagues (Zhang et al., 2019) used 21,000 training problems from the RAVEN dataset, and 300,000 from the PGM dataset.

This dependency on direct training in a reasoning task using big data makes the end-to-end approach qualitatively different from human analogical reasoning. In order for the RPM task to be of interest as a measure of fluid intelligence—the ability to manipulate *novel* information in working memory—extensive pretraining on RPM-like problems must necessarily be avoided (Snow, Kyllonen, & Marshalek, 1984). When the RPM task is administered to a person, "training" is limited to general task instructions. Human analogical reasoning is a prime example of an out-of-task situation involving zero-shot or few-shot learning—the ability to make inferences with minimal prior exposure to structurally similar problems, a core characteristic of human intelligence (Lake, Ullman, Tenenbaum, & Gershman, 2017). There is evidence that humans can deal with both out-of-task and out-of-distribution situations. For example, after observing a handful of examples of a visual concept, people can classify (out-of-distribution) and generate (out-of-task) new instances of the concept by recursive application of a rule (Lake & Piantadosi, 2020).

An alternative approach to visual analogy maintains that the reasoning process is domain-general, and separable from the acquisition of perceptual representations (see Fig. 1 bottom panel). In this domain-general approach, analogical mapping is viewed not as a task to be trained directly, but rather as an inference problem based on comparisons between representational structures. The core of analogical reasoning is assumed to be based on domain-general processes that operate on representations of entities, their attributes, and relations between entities (Hummel & Holyoak, 1997; Lovett & Forbus, 2017). For visual problems, the effective representations are those that support human perception of a three-dimensional world consisting of interrelated objects and their parts, segregated by boundaries between elements (Biederman, Mezzanotte, & Rabinowitz, 1982). These representations provide compositional structures that are not created to solve one particular task; rather, they are used to solve multiple computational problems that arise for intelligent systems. (See Berke, Walter-Terrill, Jara-Ettinger, & Scholl, 2022, for a similar contrast between goal-specific and goal-general representations in human vision.)

Some recent computational models of visual processes have this task-general character. For example, a model trained only on unoccluded objects can classify objects when they are occluded (out-of-distribution) (Zhu, Tang, Park, Park, & Yuille, 2019). Similarly, a model trained to classify objects robustly to occlusion can, as a side effect, estimate the amodal boundary of an object (out-of-task transfer) (Sun, Kortylewski, & Yuille, 2020). Such representations appear to be closely linked to the human capacity to learn from a few examples (sometimes zero-shot or one-shot learning) and to generalize learning to novel situations (Bapst et al., 2019). Once task-general perceptual representations have been acquired,

analogical reasoning can be performed by comparing relations across analogs to assess their similarity. Analogy thus emerges from the ability to represent relational structures and compute their similarity (Lu, Wu, & Holyoak, 2019).

Although theoretically appealing, analogy models based on structural comparison have usually side-stepped the problem of representation learning, often relying on representations hand-coded by the modeler. A practical disadvantage of this reliance is that such models have been unable to extract representations directly from input images or texts, making it difficult to compare their performance to that of end-to-end models using the same training and test data. To compare the task-specific and domain-general computational approaches, the present paper introduces a new database based on pixel-level images of realistic 3D objects—cars—which can be used to generate massive numbers of training data for deep neural networks, and also to create analogy problems suitable for both humans and models. After obtaining benchmark data from human experiments, we compared human performance both to that of task-specific analogy models based on deep neural networks and to that of a new domain-general model that we developed, which extracts and compares representations of cars and their component parts. To create these representations, we use the dataset to train a vision model to identify parts of cars, so that the relations between a whole object and its parts can be identified. Once these representations have been learned, we show that analogy problems can be solved simply by computing relational similarity between analogs, without any additional training on the reasoning task itself (thus instantiating out-of-task testing).

The structure of this paper is as follows: In Experiment 1, we begin with a comparison of task-specific and domain-general models with respect to their ability to reproduce human-like response patterns on traditional, four-term *A:B::C:D* visual analogy problems. In Experiment 2, we go on to compare these modeling approaches using more open-ended mapping problems. In both experiments, all problems required finding analogical correspondences based on realistic images of cars, and they featured part-whole relations, of which knowledge and familiarity are early to develop in humans (Gentner, 1977) and which remain salient to human reasoners into adulthood (Tversky & Hemenway, 1984). We manipulated experimental stimuli to test whether the same qualitative effects found in human performance would also be found in model performance. We elaborate on these manipulations in the next section, but briefly, they, respectively, center on the spatial alignment of analogs, the visual-functional salience of analogous parts, and the visibility of analogous parts. We expected each to impact human analogical reasoning (Matlen, Gentner, & Franconeri, 2020; Tversky & Hemenway, 1984; Zheng, Matlen, & Gentner, 2022). To foreshadow our results: These manipulations did each affect human performance, and across both experiments, we show that our model implementing a domain-general approach to analogical reasoning achieved a better match to human performance both in terms of overall performance and of experimental conditions than did two alternative models implementing a task-specific approach.

2. Experiment 1: Four-term visual analogies with objects and their parts

In previous studies of visual analogy, researchers have often employed simplified visual stimuli consisting of meaningless two-dimensional forms such as those used in RPM

problems (Lovett & Forbus, 2017), Bongard problems (Bongard, 1970), and the Synthetic Visual Reasoning Test (SVRT) (Fleuret et al., 2011). Due to the high complexity of these problems, solving them often involves interleaved processes of representation-building and comparison, in which reasoners iteratively re-represent the same stimulus to support increasingly more informative comparisons (Carpenter, Just, & Shell, 1990). In order to provide a clear comparison between task-specific and domain-general approaches to analogical reasoning, we sought to avoid any ambiguity in our stimuli that might promote iterative re-representation. Such ambiguity might add complexities beyond the scope of the models we are assessing here, and thus blur any performance differences between models instantiating each approach.

In order to focus on a contrast between intermixing the processes of representation-building and comparison (in a task-specific model) or separating them (in a domain- and task-general model), we used visual stimuli intended to evoke unambiguous representations. Specifically, we used visually rich stimuli depicting familiar objects in 3-D, namely, cars (see examples in Fig. 2). Their visual detail enabled variation in the texture, shading, and angle of the depicted objects without compromising experimental control over the analogy problems that they were used to generate. Moreover, the stimulus set is sufficiently extensive to provide a massive amount of data for training both deep learning models of analogy and a model based on the comparison of compositional structures. The three computational models examined in the present paper all avoid hand-coding of representations, instead taking raw pixel-level images as inputs to solve analogy problems.

The analogy problems in this dataset focus on part-whole relations. In general, human perception and thinking show sensitivity to part-whole relations across both visual and semantic domains. Structural description theories of object recognition, which include part relations, can account for viewpoint-invariant recognition in human vision (Biederman, 1987; Marr & Nishihara, 1978). Members of basic-level categories are unified by having a large number of shared parts. For example, Tversky and Hemenway (1984) showed that people converged in assessing the "goodness" of meaningful car components, such as headlights and doors. Critically, part-whole relations also permit early analogical reasoning. For example, children as young as 4 years old can map parts of a human body to their corresponding locations on a tree or mountain (Gentner, 1977).

In the current study, the source analog (*A:B*) in each problem was an image of a car body (Fig. 2A) paired with a subregion of it (Fig. 2B). Each problem was presented as a four-term analogy in *A:B::C:?* format (Fig. 2C). The target was a whole-car image of a different car type than the source, paired with a set of four alternative images of subregions as options to complete the analogy (see Methods). The problems systematically varied the four subregion images: [1] the spatial alignment between images of source and target cars, which were generated by using the same or different orientations of the whole-car images and the subregion images in an analog; [2] whether or not images of analogous subregions depicted unitary car components (part vs. piece conditions); and [3] the visibility of analogous subregions in the whole-car images (visible-component vs. invisible-component conditions). These experimental manipulations enabled us to test whether the same manipulations to experimental stimuli that affect human performance also has the similar impacts on model performance.

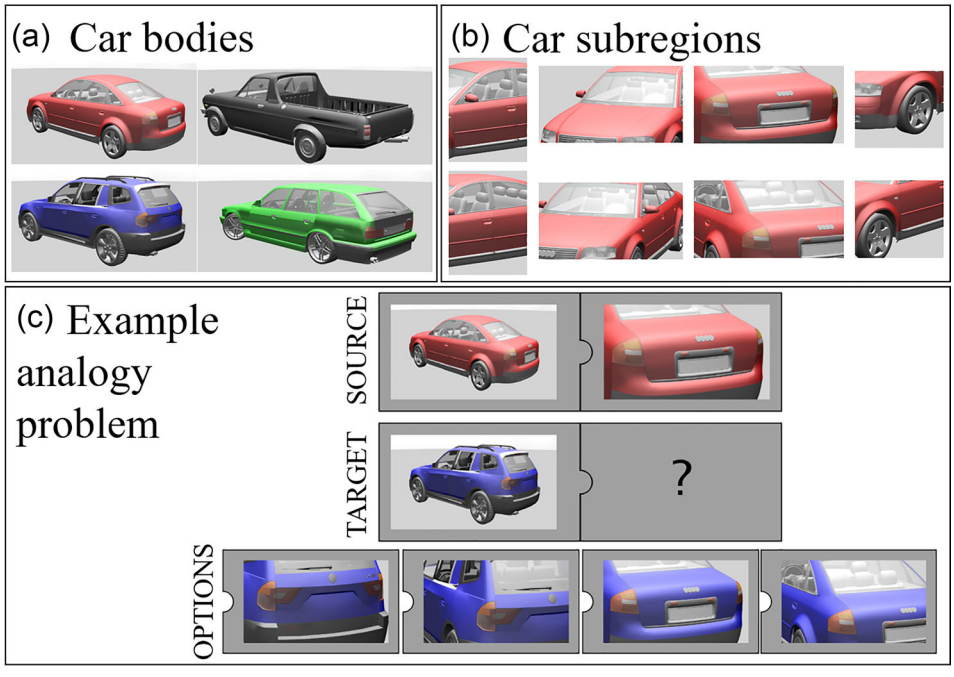

Fig. 2. Experimental stimuli. Panel A: 3D car bodies, a sedan (top left), truck (top right), SUV (bottom left), and station wagon (bottom right). Panel B: Subregions of the sedan car body, a door and window (leftmost column), hood and windshield (second column), trunk and bumper (third column), and headlight and front wheel (rightmost column). These subregions can be entire parts (top row), or pieces that do not correspond to entire parts (bottom row). Panel C: An example analogy problem in *A:B::C:?* format, constructed out of some of the images shown in panels A and B. The row of answer options includes the correct answer (leftmost), a wrong-subregion distractor (second), a wrong-car distractor (third), and a both-wrong distractor (rightmost).

This approach to model evaluation can provide greater insight than merely correlating model performance with human data without systematically manipulating independent variables in experiments.

We predicted that each of these manipulations would affect human performance and thus provide qualitative effects against which to ultimately compare alternative modeling approaches. It is known that spatial alignment between analogs helps children and adults find visual similarities and differences between them (Matlen et al., 2020; Zheng et al., 2022); accordingly, we predicted that people would be more successful in comparing vehicles with the same rather than different orientations. Based on the work of Tversky and Hemenway (1984), we also predicted that mappings based on natural parts on which people agree (e.g., trunk, front door) would be more accurate than mappings based on pieces that cut across part boundaries (e.g., top-back region, side region). Finally, solving analogy problems when the component is not visible in the whole-car image presumably depends on knowledge of 3D structure. Such problems were expected to be more difficult than those in which the analogous components are visible in the whole-car images, because the visible condition is supported

by direct perceptual information, whereas the invisible condition depends on knowledge not provided in the immediate input.

2.1. Four-term analogy task

We tested human participants and computational models on the same visual analogy dataset. Each problem in this dataset was a four-term analogy in *A:B::C:?* format. The source analog (*A:B*) was an image of a car body paired with a subregion of it. The target was a different-style car body, paired with a set of four alternative images of subregions as options to complete the analogy. The problems systematically varied [1] the spatial alignment between images of source and target cars, [2] whether or not images of analogous subregions depicted unitary car components, and [3] the visibility of analogous subregions in the whole-car images.

Car images in analogy problems were generated using 3D car models taken from the ShapeNet 3D Core dataset (Chang et al., 2015). Four subtypes of cars were selected from the dataset: sedan, SUV, station wagon, and truck, each represented by a single 3D car model, and Fig. 2A shows these car bodies. These subtypes yielded a counter-balanced design including the following four pairs of car bodies, in which each body was used twice (once as a source and once as a target): sedan (source)—SUV (target), SUV—station wagon, truck—sedan, station wagon—truck. Car subregion images were generated using the UDA-Part dataset (Liu et al., 2022). This dataset provides detailed part annotations for 3D computer-aided design (CAD) models of vehicles. Each surface mesh on the CAD models was assigned a label from a set of 31 vehicle parts (e.g., back left door, front right door, front left windows, right mirror, and bumper). Eight vehicle parts were selected from the dataset, each of which were spontaneously generated by participants in a part-labeling task: door, window, hood, windshield, trunk, bumper, headlight, and wheel. These eight parts were rated as "good" or characteristic parts of cars (Tversky & Hemenway, 1984). Car subregions used in our visual analogy problems either fully (*parts* condition; top row in Fig. 2B) or partially (*pieces* condition; bottom row in Fig. 2B) depicted one of the mentioned car parts. The images were rendered with Blender software using randomly selected textures. The virtual camera had a resolution of 1024×2048 and a field of view of 90 degrees. Fig. 2B shows examples of different car subregions used to construct the analogy problems.

Each analogy problem (see an example in Fig. 2C) consisted of an incomplete 2×2 image array with the source whole-car image in the top left corner, a subregion of that car body in the top right corner, the target whole-car in the bottom left corner, and a question mark in the bottom right corner. Each problem was based on one of four pairs of car bodies: sedan (source) and SUV (target; pictured in Fig. 2C), SUV and wagon, wagon and truck, or truck and sedan. Below, the array was a set of four answer options presented in a randomized order, and the task was to select the option that best fit the bottom right cell of the 2×2 array. In addition to the correct option (e.g., leftmost in Fig. 2C bottom row), one option depicted a disanalogous subregion of the target car (*wrong subregion*) (e.g., second in Fig. 2C), one option depicted an analogous subregion of the source car (*wrong car*) (e.g., third in Fig. 2C), and a fourth option depicted a disanalogous subregion of the source car (*both wrong*) (e.g., rightmost in

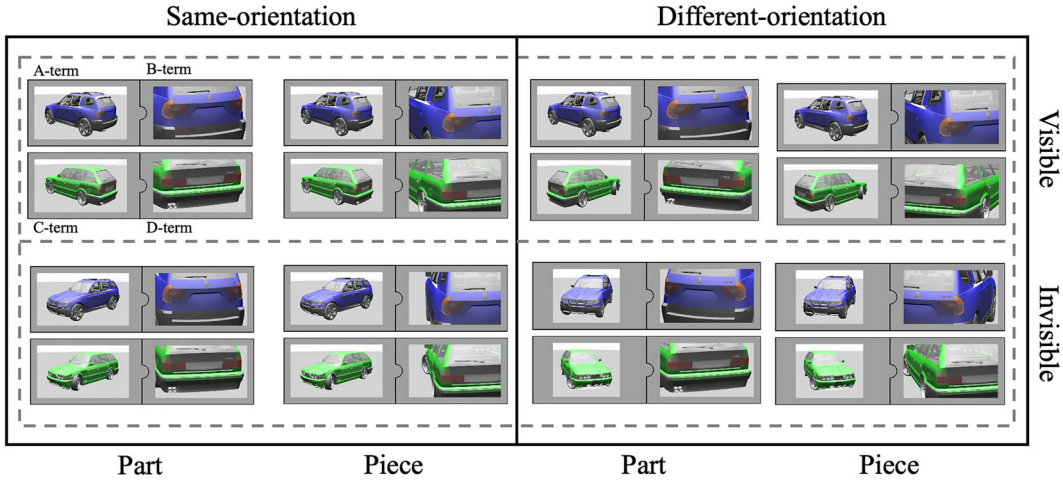

Fig. 3. Examples of each of the eight problem types in the four-term visual analogy dataset tested in Experiment 1. For each problem, the correct completion (*D* term) is shown; *A:B* (source) appears on top, and *C:D* appears on bottom.

Fig. 2C). When the correct answer was a part subregion (e.g., top left image depicting a door and window in Fig. 2B), the wrong-subregion distractor depicted the corresponding piece subregion (e.g., bottom left image depicting a partial view of a door and window in Fig. 2B). These assignments were reversed when the correct answer was a piece subregion.

Our entire test set consisted of 128 analogy problems that varied three factors: orientation of source and target cars (same vs. different), subregion type (part vs. piece), and subregion visibility in the corresponding whole car images (visible vs. invisible). Fig. 3 includes examples of different problem types. Object orientations of cars in source and target images were either same or different in the analogy problems. For part problems, the source and target subregions fully depicted corresponding car components, whereas for piece problems, these subregions included multiple partial components of car parts. For visible problems, the source and target subregions were visible from the images, respectively, depicting the source and target car bodies, whereas for invisible problems, these subregions were occluded in the corresponding images. Crossing each of these factors yielded eight problem types (see examples in Fig. 3). Both humans and models completed these analogy problems; we detail human data collection in the next section.

2.2. Human participants and experimental methods

Seventy-nine participants were recruited from Amazon Mechanical Turk (M_{age} = 42, SD_{age} = 11, age range = [24, 73], 41 female, 38 male). Participants provided online consent in accordance with the University of California, Los Angeles Institutional Review Board and were compensated with a monetary reward of $2.75 (this corresponds to a rate of $11 per hour, assuming a 10-min task, which we determined with some task piloting). In order to avoid excessive fatigue, each participant was asked to complete 32 problems out of the

possible 128. To that end, we created four 32-problem subsets, each consisting of four problems falling into one of the eight problem types, and these subsets were distributed evenly across participants. For a given problem type, each of these four problems were sampled such that they each depicted one of the four car body pairs (sedan-SUV, SUV-wagon, truck-sedan, and wagon-truck) and each of the four car subregions (trunk and bumper, wheel and head light, window and door, and windshield and hood), and the particular problems constituting each of the four 32-problem varied which of the four car body pairs were combined with which of the four car subregions on each of the eight problem types. This experiment was not preregistered; however, experimental stimuli and data are available at the following link: https://osf.io/mkery.

Before starting the task, participants were given a practice analogy problem with line drawings of a gas pump and a lawnmower and a battery and a flashlight (instantiating the common relation *x powers y*). For this practice problem and for the car analogy problems, participants were instructed "to select which option contains a picture that is related to the picture in the left bottom box in the same way as the pictures in the top two boxes." No training or practice on solving the car analogies was provided. No feedback was provided in the experiment.

2.3. Computational models

We implemented models to instantiate task-specific and task-general modeling approaches, respectively. In order to instantiate a task-specific modeling approach to analogical reasoning based on extensive training with highly similar reasoning problems, we implemented two deep learning models, a Siamese Network (Bromley et al., 1993) and a Relation Network (Santoro et al., 2017). A Siamese Network, which has been successfully applied to visual detection tasks (Bromley et al., 1993) and to visual and verbal analogy tasks (Rossiello, Gliozzo, Farrell, Fauceglia, & Glass, 2019; Sadeghi, Zitnick, & Farhadi, 2015), contains two or more identical subnetworks, emphasizing the role of comparisons among multiple inputs in forming comparable feature embeddings as visual representations (see top panel of Fig. 4). The embeddings are compared to assess the similarity between inputs. A Relation Network was developed specifically to learn to solve relational reasoning problems (see bottom panel of Fig. 4). This class of models has been successfully applied to RPM-like analogy problems (Barrett et al., 2018) and query tasks (Santoro et al., 2017).

We compare each of these networks to a new part-based comparison model (PCM), which we introduce to instantiate the task-general representational comparison approach to analogy. This model employs image segmentation algorithms to extract visual features representing part-based structures for 3D objects (see Fig. 5), and then compares representations by computing a generic measure of relational similarity.

2.3.1. Training task-specific models

To train the deep learning models, 30,000 analogy problems were created using part labels to generate images of subregions. We divided the set of 30,000 analogy problems into 27,000 for training and 3000 for test. For each analogy problem, we sampled two whole-car images from the following five different car types of the 3D CAD models (denoted as *A*, *C*): sedan,

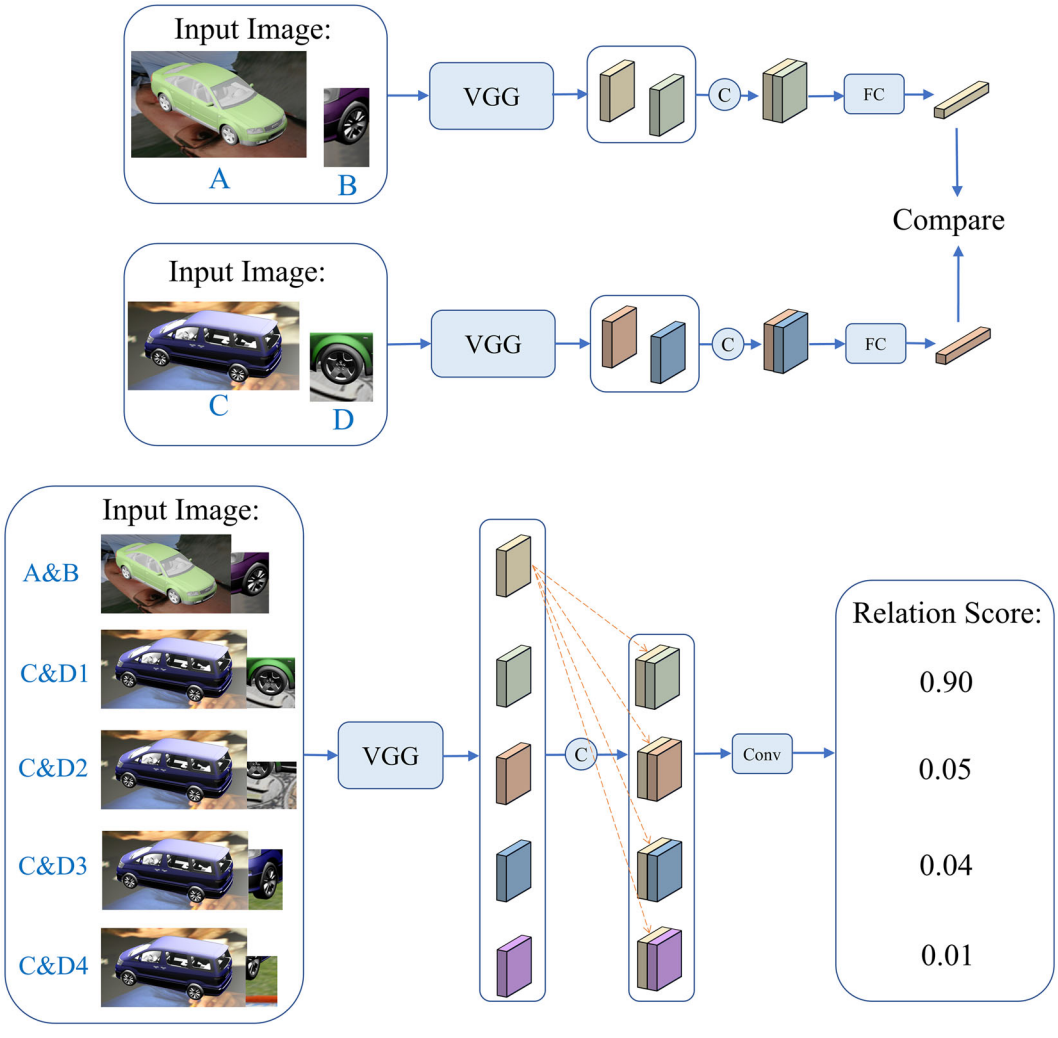

Fig. 4. Siamese Network (top panel) and Relation Network (bottom panel) architectures for answering car analogy problems. "C" in the figure indicates concatenation operation, "FC" indicates fully connected, and "Conv" indicates convolution operation.

SUV, station wagon, truck, and minivan. Then, we randomly selected part or piece subregions of the car in the *A* image to generate the *B* image. The correct option for the *D* image was the unique image that yielded the same whole-part relation for the *C:D* image pair as for the *A:B* pair. For each analogy question, the other three alternative *D* images were one other part/piece images from the car in the *C* image and two subregion images from the car in the *A* image. Whole-car images and car subregion images had random background, lighting, camera position, and object texture/color. None of the images used to train the networks were included in the visual analogy task used to compare model and human performance.

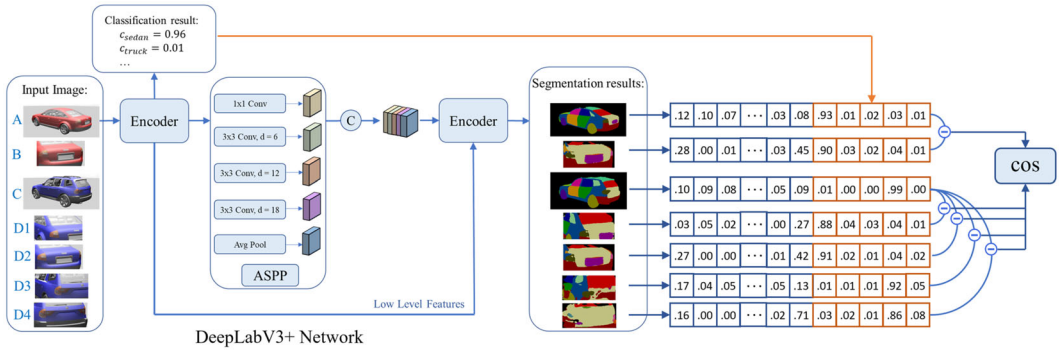

Fig. 5. PCM model architecture, which consists of DeepLabv3+ Network Architecture for identification and segmentation of synthetic car images (left), and a comparison-based reasoning module (right).

2.3.2. *Siamese Network*

We adapted a Siamese Network to learn to solve our car analogy problems. Fig. 3A shows the model architecture for this network, which employs a VGG-16 network to translate pixel-level images to features. Features of whole-car images (*A* and *C* images) are processed by spatial pooling to form embeddings of size 4×4×512; subregion image features (*B* and *D*1-8 images) are pooled to form 1×1×512 embeddings. The feature embedding for image *B* is then expanded to the same size as that for image *A*, and then concatenated with *A* in the channel dimension separately. Feature embeddings for *D* images are similarly expanded and concatenated with the feature embedding for the C image. The concatenated features are passed through another convolutional layer and three fully connected layers to obtain the final embedding of an image pair, which is a vector of size 512. We use contrastive loss with margin 1 to minimize the distance between concatenated feature embeddings from *A:B* images and embeddings from *C:D* images to choose the *D* image that best completes the analogy. More details about training all models are provided in Section 1 of Supplementary Materials.

2.3.3. *Relation Network*

We adopted the implementation by Sung et al. (2018) to set up the Relation Network to learn to solve the car analogy problems; Fig. 3B shows the model architecture for this implementation. The input to the Relation Network is pairs of images including a whole-car image and their corresponding subregions (e.g., *A&B* images, *C&D*1-8 images). To train the network to solve analogy problems, we padded the subregions images B and D images to the same image size as their corresponding whole-car images A and C, and then concatenated the whole-car image with each of the subregion images in the channel dimension, forming 6-channel input image pairs (i.e., the first three channels from color channels of a whole-car image and the second three channels from padded subregion images). The input image-pairs are then passed through the VGG network separately to get a fusion relation feature of *A&B* and *C&D*1-8 images. The extracted features of the *A&B* image pair are then concatenated with the features of the *C&D* image pairs in the channel dimension. The concatenated

features are then passed through two convolutional layers and two fully connected layers to estimate a relation score indicating the probability that a candidate *C&D* pair instantiates the same relation as the question pair of *A&B* images. Cross entropy loss is used to train the Relation Network.

2.3.4. Part-based comparison model

The PCM is our instantiation of the domain- and task-general approach to visual analogy. However, the image representation components in PCM were trained to recognize cars and segment their parts. The training dataset was based on the same UDA-Part dataset used to generate images of car subregions in the four-term visual analogy problems. We rendered 40,000 synthetic images (30,000 for training and 10,000 for test; none of these were used in the analogy task) with automatically generated part segmentation ground-truth. Each training image depicted one of the five car subtypes that were used to train the task-specific models: sedan, SUV, station wagon, truck, and minivan. We rescaled these images to .5 of their original size, which yielded the best performance on part classification and on the analogy problems in Experiment 1. We report the results of control simulations systematically varying this rescaling factor in Section 2 of the Supplementary Materials.

The image representation component in PCM uses a variant of the DeepLabv3+ architecture, a deep neural network that is widely used for semantic segmentation in computer vision (Chen, Zhu, Papandreou, Schroff, & Adam, 2018). The left side of Fig. 5 shows this architecture, which includes encoder layers, an atrous spatial pyramid pooling (ASPP) module, and decoder layers. An input image first is processed through an encoder module composed of a ResNet101 (He, Zhang, Ren, & Sun, 2016), yielding a feature map with 2048 channels and size (height and width) down-sampled by 16. Features extracted from ResNet101 are commonly used for image segmentation. The ASPP module then samples the input features at multiple spatial rates to gather information at different spatial scales. The outputs from different spatial scales are concatenated in the channel dimension and passed through the decoder layers. The output is a mask with the same height and width as the input image. Each pixel of the mask is assigned a label to indicate whether it belongs to the background or each of the 31 parts labeled in the dataset. We added a second output branch after the encoder layers of DeepLabv3+ to predict the car type label, which formed a regular object classifier. Training was conducted using two standard cross-entropy losses: one for segmentation, and one for classification of car types.

Before feeding the images into the network, standard data augmentation was applied (e.g., translation, scaling, and cropping). We controlled the augmentation parameters to ensure the network was only trained with whole car images (i.e., no partial car of subregion images were used during training). Car images used in the human experiment were excluded from the training set. The model achieved high performance for both part segmentation (mean intersect over union [mIoU] = 0.57) and subtype classification (accuracy = 0.99) on the test set. Fig. 6 shows the segmentation results for the trained model.

After training, we applied the network to images used in the analogy problems to obtain segmentation and classification predictions. When applied to the untrained whole-car images used in human experiment (images *A* and *C*), the segmentations were reasonable, and the

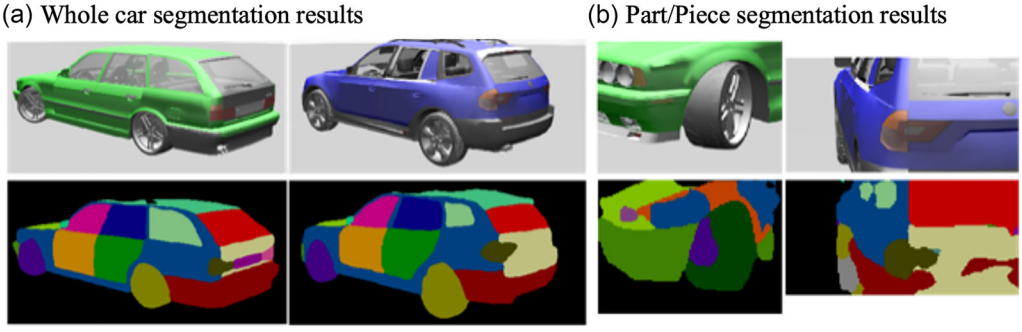

Fig. 6. PCM segmentation results for images used in analogy problems. The model never saw these images during training.

classification accuracy was perfect (Fig. 6A). When applied to the part/piece subregion images used in the experiment (i.e., image *B* and the alternative *D* images), which provide incomplete visual input, the segmentations yielded small errors, and the classification accuracy dropped to .71 (Fig. 6B). These results suggest that the image representation component in PCM based on visual deep learning models of semantic segmentation and object recognition shows a substantial degree of generalization to untrained images for these visual tasks.

Next, we created compositional descriptions of car images based on extracted parts using the segmentation and classification results. We first converted the pixel-level labels in the resulting segmentation map to a low-dimensional feature vector. Specifically, we counted the number of pixels that were parsed into each of the part segments, and computed the proportion for each part (i.e., the number of pixels in a part divided by the total number of pixels in the resulting segmentation map). The dimensions of feature vectors were defined using a taxonomy with 13 parts for cars, adapted from a more generalized segmentation scheme used for parsing complex, real-world scenes in the PASCAL-Part dataset (Chen et al., 2014). The 13 car parts aggregate over subsets of the 31 segments in the UDA-Part dataset on which the part segmentation model was trained. These 13 parts are back part (including back bumper, tail lights, and trunk segments), front part (including front bumper and hood), left frame, right frame, roof, door (including back left door, back bright door, front left door, and front right door), window (including all window in the car), wheel (including all four wheels), back license plate, front license plate, left mirror, right mirror, and head light.

In addition to part-related features, we included the object information of car subtypes by concatenating the part-proportion vector with the car-type classification result, yielding the 18-dimensional feature vector $\mathbf{f} = cat(\mathbb{P}(\mathbf{m}), \mathbf{c})$, where $\mathbf{m}$ is the resulting segmentation map, $\mathbf{c}$ is the car-type prediction, $\Sigma f_{1:13} = 1$ and $\Sigma f_{14:18} = 1$. Thus, for each analogy question, PCM represents images $\mathbf{I}_A, \mathbf{I}_B, \mathbf{I}_C, \mathbf{I}_{D1}, \mathbf{I}_{D2}, \mathbf{I}_{D3}, \mathbf{I}_{D4}$ as feature vectors $\mathbf{f}_A, \mathbf{f}_B, \mathbf{f}_C, \mathbf{f}_{D1}, \mathbf{f}_{D2}, \mathbf{f}_{D3}, \mathbf{f}_{D4}$, respectively.

Finally, to solve the visual analogy problem, a decision is derived by selecting the best *D* image $\in \{D1, D2, D3, D4\}$ such that the relation that holds between image *A* and image *B* is similar to the relation between the two images *C* and *D*. We computed the difference between

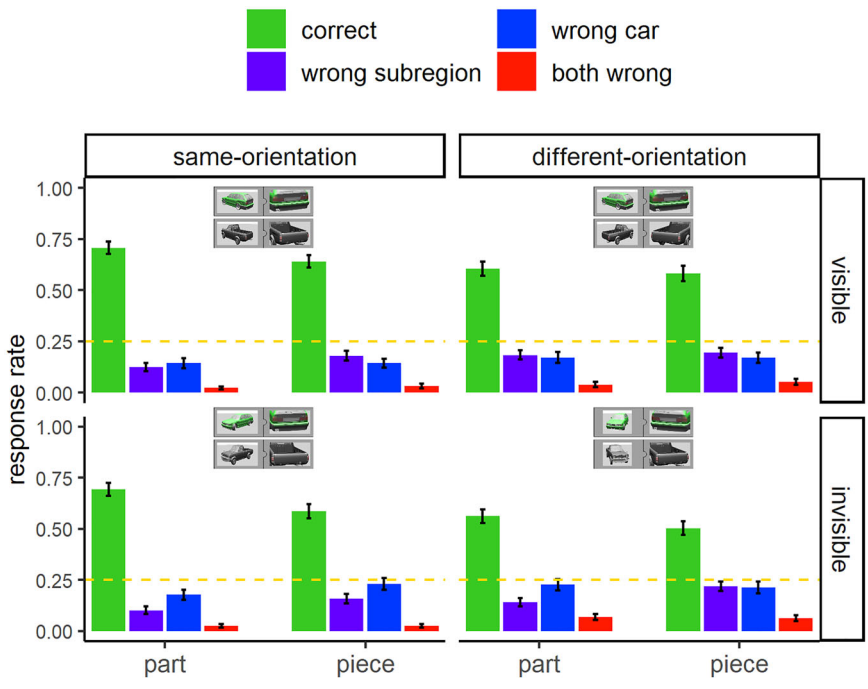

Fig. 7. Human response selections for each problem type. Examples in each quadrant depict an experimental condition (same vs. different orientation between the whole and subregion images for the left and right column panels; subregion images visible vs. invisible in the corresponding whole images for the top and bottom row panels). Dashed line reflects chance performance. Error bars reflect ± 1 standard error of the mean.

feature vectors $\mathbf{f}_A$ and $\mathbf{f}_B$, and between $\mathbf{f}_C$ and $\mathbf{f}_D$, and used cosine distance to measure the dissimilarity of the two difference vectors. The same approach has been used in the Word2vec model (Zhila, Yih, Meek, Zweig, & Mikolov, 2013), and has proved effective in modeling visual analogical reasoning (Lu, Liu, Ichien, Yuille, & Holyoak, 2019). The preferred answer $\hat{D}$ is defined as the D image that generates minimum cosine distance between difference vectors:

$$\hat{D} = \text{argmin}_{D \in \{D_1, D_2, D_3, D_4\}} 1 - \cos\left(\mathbf{f}_B - \mathbf{f}_A,\ \mathbf{f}_D - \mathbf{f}_C\right).$$

2.4. Human performance on four-term visual analogies

Participants achieved an overall mean proportion correct of .61 (SD = .21, range = [.09, .97]), greatly exceeding the chance level of .25 correct. Fig. 7 provides a breakdown of response selections for eight experimental conditions. When participants did not select the correct answer, they more often selected one of the two distractors that included an element of the correct answer (correct subregion or else correct car body). They very seldom selected the option that was completely incorrect (wrong subregion of wrong car body). A three-way repeated-measures ANOVA revealed that accuracy was higher on same-orientation problems (M_{acc} = .65, SD_{acc} = .22) than on different-orientation problems (M_{acc} = .56, SD_{acc} = .23),

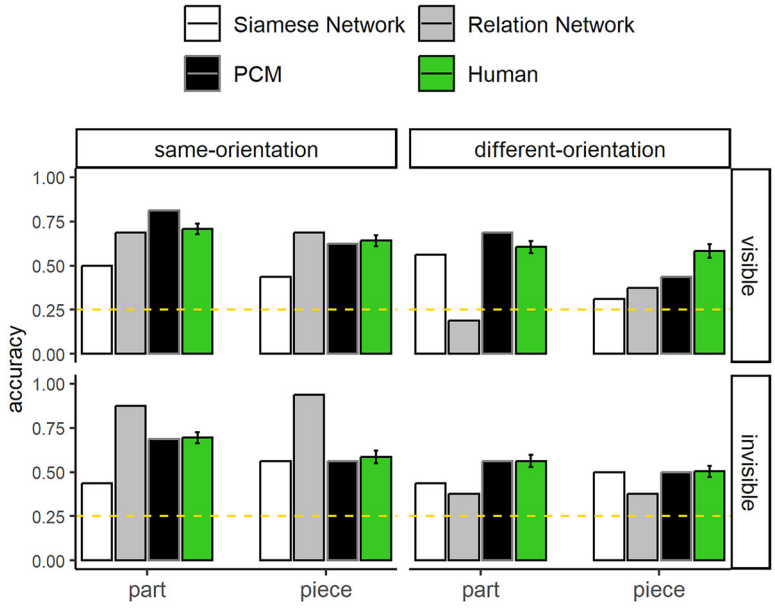

Fig. 8. Model and human performance on the visual analogy task broken down by problem type. Dotted lines indicate chance performance. Error bars reflect ± 1 SEM for human data.

$F(1,75) = 34.20$, $p < .001$, on part problems ($M_{acc} = .64$, $SD_{acc} = .23$) than on piece problems ($M_{acc} = .58$, $SD_{acc} = .22$), $F(1,75) = 14.37$, $p < .001$, and on visible-subregion problems ($M_{acc} = .63$, $SD_{acc} = .23$) than on invisible-subregion problems ($M_{acc} = .59$, $SD_{acc} = .22$), $F(1,75) = 7.95$, $p = .006$. No interaction effects were significant.

2.5. Model performance on four-term visual analogies

All three models (Siamese Network, Relation Network, and PCM) were tested on the same 128 visual analogy problems used in the human experiment. These involve car images never used in training for any of the models. The overall proportion correct was .38 for the Siamese Network, .56 for the Relation Network, and .61 for PCM. Fig. 8 plots model and human performance, broken down according to each experimental condition. The PCM model matched overall human accuracy most closely (.61). More importantly, only PCM captures the qualitative differences in human accuracy across all conditions (Fig. 9). PCM, like humans, shows higher accuracy on same-orientation problems than on different-orientation problems (Fig. 9 top), on part problems than on piece problems (Fig. 9 middle), and on visible-component problems than on invisible-component problems (Fig. 9 bottom). While the two task-specific models also reproduced the effect of orientation, neither was able to capture all three qualitative effects, as PCM did. In order to quantitatively assess the fits of the models to human data, we computed the root mean squared deviation (RMSD) between model accuracy and mean human accuracy for each of the eight types of analogy problems. This measure indicates how much each model deviates from human performance, with a lower RMSD indicating a closer

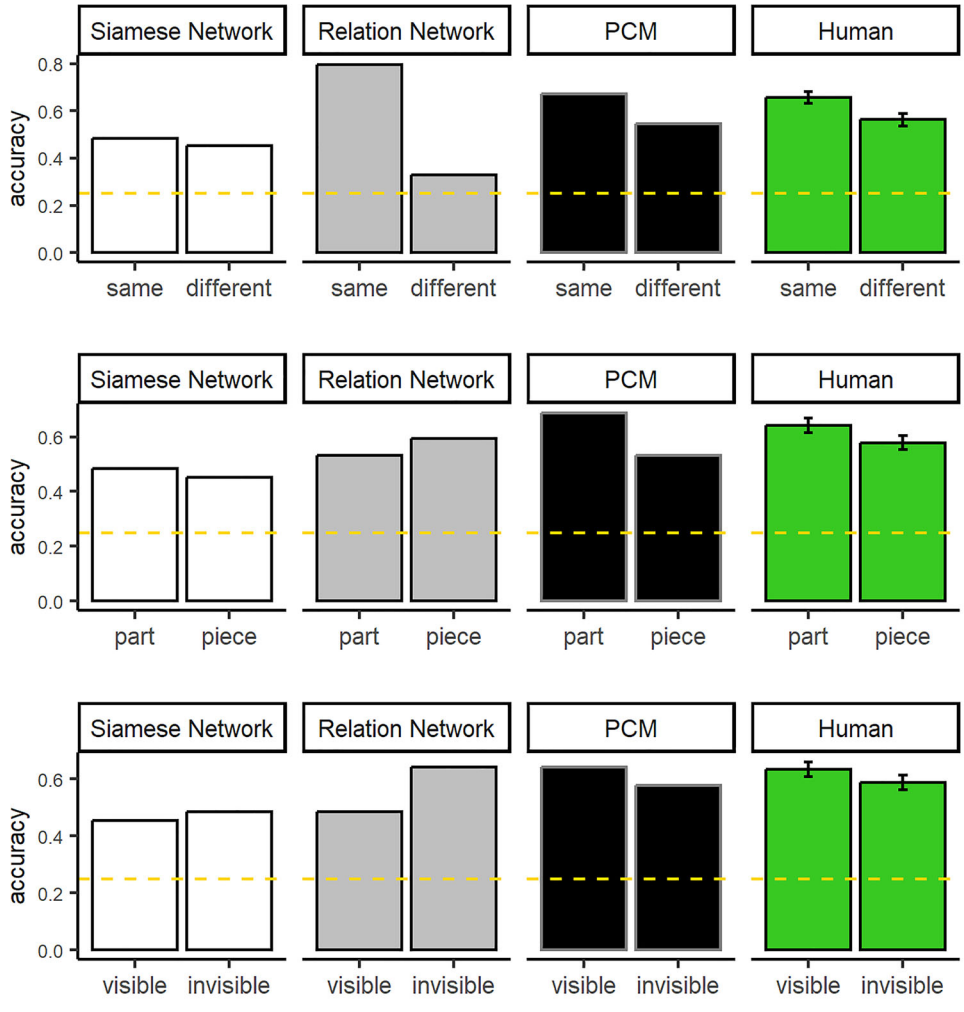

Fig. 9. Model and human performance on the visual analogy task plotted separately for each of the following manipulations: Orientation (top), subregion type (middle), and visibility (bottom). Dotted lines indicate chance performance. Error bars reflect ± 1 SEM for human data.

fit to human data. PCM yielded an RMSD of .07, achieving by far the closest fit to human data. RMSD was .17 for the Siamese Network and .23 for the Relation Network.

Recall that we trained task-specific models using 30,000 analogy problems, and we trained PCM's classifier using 30,000 whole-car images. We examined each model's performance on our analogy problems as a function of size of the training set. We varied the number of analogy problems used to train the Siamese Network and Relation Network: 5000, 10,000, 15,000, and 30,000 problems. Fig. 10 shows the two networks' training (top left) and validation performance (top right), as well as their performance on Experiment 1's four-term analogies (bottom right) as a function of their training set size. Across all three performance

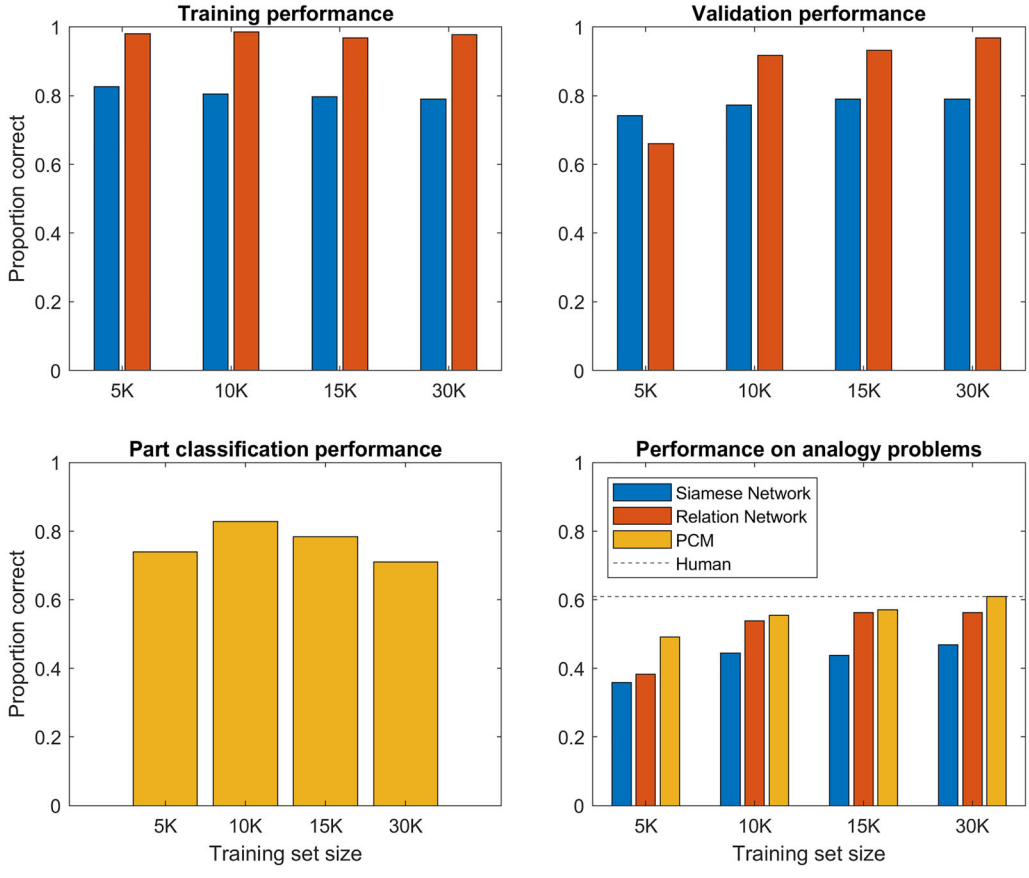

Fig. 10. Siamese Network and Relation Network's performance on training problems (top left) and on held-out validation problems (top right), PCM's part classification performance (bottom left), and all models' overall performance on visual analogy problems (bottom right) all as a function of training set size. Human performance on analogy problems is represented by the dotted line in the bottom right panel.

metrics, the performance of both models plateaus at or before 10,000 training examples. In light of the strong performance of Relation Network (and to a lesser extent, Siamese Network) on validation data, the weak performance of the task-specific models on the visual analogy test problems may seem surprising. A major reason for the poor generalization performance of these models is their lack of any direct pressure to represent part-whole relations between *A* and *B* terms, and between the *C* term and a potential answer option. Instead, these models represent source and target analogs simply as the concatenation of feature vectors, respectively, representing *A* and *B* terms and *C* and answer options. These models are not given any explicit directions to represent any relations between feature vectors.

We also varied the number of images used to train PCM's classifier. PCM's part classification performance on the images used to construct our visual analogy problems showed an inverted U, indicating some overfitting on its training images as the set size increased

beyond 10,000 images (see Fig. 10 bottom left panel). However, increasing the training set size beyond 10,000 images appeared to aid analogy performance, as the model's accuracy on the visual analogy problems steadily increased as a function of its classifier's training set size (Fig. 10 bottom right panel).

To further evaluate the models as accounts of human analogical reasoning, we considered a possible nonanalogical shortcut strategy for answering the problems. Using the problem in Fig. 2C as an example, some problems in the experiment could be answered correctly by selecting the *D* image most visually similar to the *B* image but not identical to it. We tested the possibility that any of the three models might have exploited this shortcut strategy, by providing only the subregion images (*B* image and *D* images) without showing the whole-car images (*A* and *C* images). The Siamese Network and PCM performed at chance on this test, indicating they did not exploit the shortcut strategy. However, Relation Network made the *exact same responses* when the "analogy" problem was reduced to just two of the four terms, implying that the model did not actually learn how to reason by analogy (i.e., by comparing a whole-part relation in source images to a similar relation in target images). This finding further speaks to Relation Network's lack of any explicit relation representation as an explanation for its poor generalization performance from its training data to the visual analogy test set. Judging by its performance breakdown across conditions, Relation Network's clear advantage on same-orientation problems relative to different-orientation problems (Fig. 9 top, gray bars) suggests that the model exploited the visual similarity between spatially aligned car bodies and subregions to select the correct answer among the four response options. In Section 3 of the Supplementary Materials, we report the results of control simulations that replace Relation Network's pixel-level input with the low-dimensional output of PCM's classifier that served as the input to PCM's reasoning module. While providing Relation Network with this alternative input did not improve overall performance on the visual analogy problems in Experiment 1, it did prevent the model's acquisition of this nonanalogical shortcut strategy.

The finding that the Relation Network was able to correctly answer roughly half of the four-term analogy problems using a demonstrably nonanalogical strategy raises concerns about the extent to which this four-term analogy problem set assesses analogical reasoning. Accordingly, we performed Experiment 2 to compare model and human performance on a task that placed stronger demands on *analogical* reasoning, rather than some possible shortcut strategies guided by purely visual features.

3. Experiment 2: Open-ended visual analogies between objects and their parts

In Experiment 2, we asked participants to complete an open-ended visual analogy task, in which they were given an image of a car body with two colored markers on that car and were asked to mark the analogous spots on a different car body. We compared humans and models in their ability to identify and match analogous parts across car types in this open-ended analogy task. None of the models received any additional training beyond that described in connection with Experiment 1. Experiment 2 thus allowed us to compare task-specific and

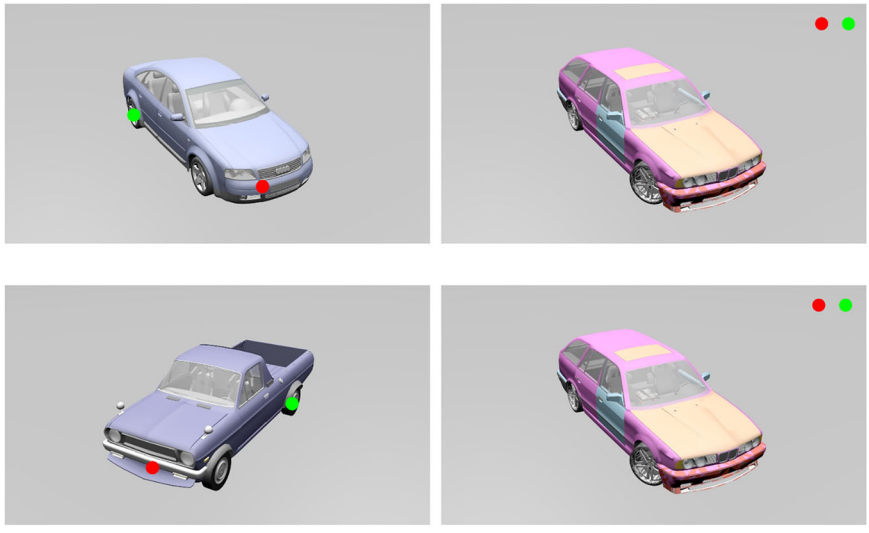

Fig. 11. Two example trials (top and bottom rows) for the open-ended visual analogy task. Given a source image with two colored dots (left), participants were asked to place corresponding dots on the target image (right).

task-general approaches to analogical reasoning on out-of-task generalization. Moreover, examining each model's responses on this more open-ended task will reveal its response strategies, including the nonanalogical strategies learned by the Relation Network.

3.1. Open-ended visual analogy task

In order to produce these open-ended analogy problems, we used the CGPart dataset (Liu et al., 2022) that contains the five car subtypes used to train models: sedans, SUVs, wagons, trucks, and minivans. From these 3D models, we generated 120 unique problems, each consisting of a source and target image showing different car bodies. A few examples of the stimuli are shown in Fig. 11. The CGPart dataset provides detailed pixel-level annotations of vehicles and their component parts, which enabled us to place colored markers on source images so that they were centered at car parts of interest, either on wheels, bumpers, or roofs. Across problems, we manipulated the spatial alignment of the car bodies in the source and target images (the sole manipulation that in Experiment 1 consistently affected the performance of both humans and all three models). The two car bodies in an image pair were shown either in the same orientation or different orientations, 120 degrees apart. Since the visibility and part manipulations did not affect performance in humans and all three models, we did not manipulate them in Experiment 1: All marked parts (e.g., bumpers) remained visible in both images, even when car bodies were not spatially aligned, and all marked parts had a clear semantic label.

Each trial of this task consisted of a pair of images—a source image and a target image—each depicting a different car body. Two different-colored markers, one red and one green, were placed at the center of two parts on the car body in the source image, and participants

were asked to use their mouse to drag a second pair of red and green markers to the analogous locations on the car body depicted in the target image. Specifically, this marker-placement task was to place the green marker in the target image so that it mapped to the green marker in the source image, and similarly for the red marker.

3.2. Human participants and experimental methods

Forty-nine participants ($M_{age} = 19.53$, $SD_{age} = 1.43$, 40 female, 9 male) were recruited from the Psychology Department subject pool at UCLA. Participants provided online consent in accordance with the UCLA Institutional Review Board and were compensated with course credit.

On each trial, participants were presented with one image pair on a computer screen as shown in Fig. 11, and completed a marker-placement task. For each of the two colored markers, they were asked to "move the marker on the top right corner in the target image to the corresponding location that maps to the same-color marker in the source image." If the participant did not think there was an analogy between the two images, they were allowed to move the markers back to the top right corner and leave them there. No time constraint was imposed. On each trial, the exact location of each marker placement was recorded.

Data from 7 of the 49 participants were excluded from analyses either because they indicated they were not serious during the experiment or because they attempted to move the markers fewer than half of the trials. Thus, data from the remaining 42 participants were included in the analyses.

3.3. Model simulations

In order to test models on open-ended mapping problems, we adopted a simple modification to the simulations for the four-term analogy problems in Experiment 1. Because our models could only pinpoint image patches with the same relation to the whole image, we adopted a sliding window method to find the marker locations for the target image that were most analogous to those provided in the source image. From the source image, we used the image to be the whole-car image (image *A*), and we cropped out an image patch centered around the marker location to generate the subregion image (image *B*). We used the target image as the whole-car image for image *C*, and then slid the patch window on the target image with stride of 5 pixels. To generate subregion image *D*, we started by cropping an image patch from the top-left corner of the target image, and then moved 5 pixels rightward to crop the next image patch from the target image. When we reached the rightmost position of the image, we moved the window back to the leftmost position and 5 pixels downward from its previous position, and then repeated the last operation.

This sliding window procedure enabled us to sample hundreds of image patches from the target image to generate the subregion images (image *D*). We then used each model to identify which patch subregion in the target image was most analogous to the patch subregion centered at the marker location of the source image. Thus, the main difference between the testing procedure in Experiment 2 and that used in Experiment 1 is the number of answer options. Whereas the four-term analogy problems in Experiment 1 each had four answer options of

Table 1
The mean distance of model predictions to the human marker center measured in pixels

	PCM		*Relation Network*		*Siamese Network*	
window size	same	different	same	different	same	different
100	17.7	28.0	60.7	84.8	93.0	93.7
110	17.0	30.4	56.1	79.4	98.3	107.5
120	18.0	30.6	45.6	63.2	77.3	91.4
130	19.2	33.1	42.5	65.5	83.2	85.1
140	19.5	38.5	36.6	73.0	79.0	72.6
150	18.1	43.0	40.3	72.0	66.4	76.4

Note. Smaller values indicate more accurate model predictions for human marker locations.

subregion images, the sliding window procedure for the more open-ended analogy problems tested in Experiment 2 yielded hundreds of answer options from patches sampled from the target image. We did not test model predictions on problems where more than 30% of participants failed to find an analogy between the source and target, and, therefore, opted out of responding. This excluded 20 out of 120 problems in the following analyses.

3.4. Human and model performance on open-ended visual analogy task

For each trial, we calculated the mean location of colored marker placements, averaged across participants. We then computed the spatial distance (in pixels) from the marker location provided by each individual participant to the overall mean location. This measure of individual distance to the mean marker location provided a quantitative assessment of human variability in mapping judgments, with smaller distance values indicating higher consistency across participants for their marked locations in the same analogy problem. For the same-orientation image pairs, the average distance to the mean marker location was around 5.7 pixels, whereas different view pairs had an average distance of around 14.9 pixels. Note that relative to the object sizes (average height of 202 pixels and width of 317 pixels), even 15-pixel derivation represents a small spatial distance, indicating strong agreement of people's judgments in analogical mapping. A repeated-measures ANOVA comparing trial-level deviation from the mean marker placements on problems where source and target images depict cars in the same orientation ($M_{same} = 5.93$) with those where cars are in different orientations ($M_{diff} = 15.98$) yielded a reliable main effect of orientation, $F(1,243) = 54.5$, $p < .001$.

When comparing model predictions to human responses, we found that the responses of different model were variably affected by the size of each patch window. Table 1 shows the pixel-wise deviation between model predictions and mean human marker placements for different window sizes on same-orientation and different-orientation problems. Whereas our task-general model PCM was robust to changes in window size, both task-specific models were highly sensitive to such changes. Regardless, PCM's predictions were substantially closer to human marker placements across all window sizes, achieving the best performance using window size of 100.

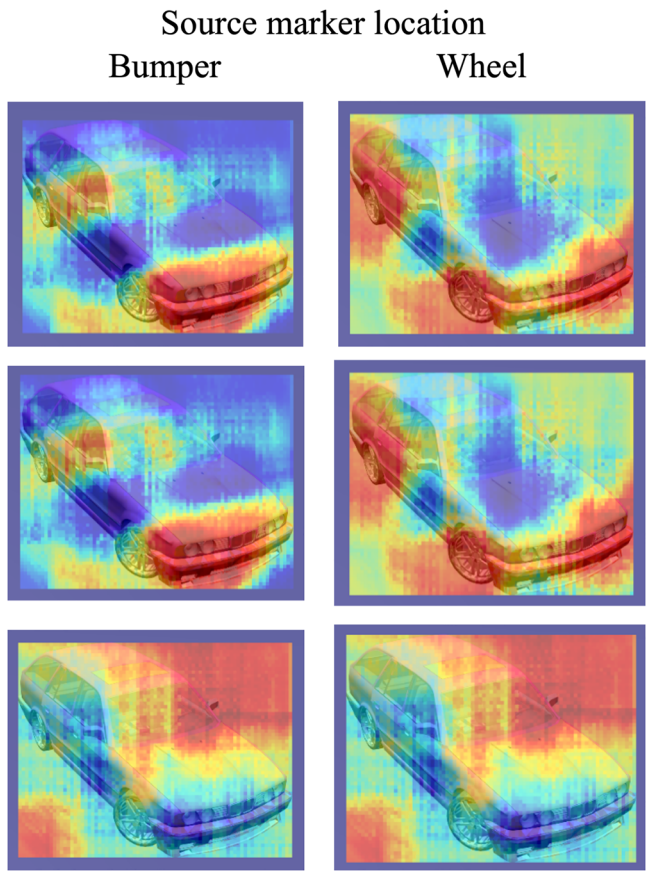

Fig. 12. Prediction scores for Relation Network when the moving patch is centered on each location under various conditions. Left: The source patch is centered on a bump; right: the source patch is centered on a wheel. Top row: All whole-car images and the source marker patch are passed to the network; middle row: only the source marker patch is passed to the network; bottom row: the whole-car source and the target images, along with a blank image for source patch images, are passed to the network. Redness in the heatmap means the network yields a high score, whereas blue indicates the network gives a low score.

We were also able to more precisely characterize the short-cut strategy learned by the Relation Network when it was trained on the type of four-term analogies used in Experiment 1. We depicted the scores of patches at each location on the target image, which are plotted as a heatmap in Fig. 12. These scores were generated using a window size of 140, which showed the best fit of the Relation Network to human performance. Scores reflect performance on same-orientation problems, but the model's prediction for different-orientation problems was similar. When the source image, the source patch subregion centered at a marker location, and the target image were passed in as inputs, the Relation Network generated a reasonable score distribution over the car in the target image. As shown in the top row of Fig. 12, the network focuses on the bumper when the source marker patch is a bumper (left) and focuses

on the wheels when the source marker patch is a wheel (right). However, when given the source marker patch without the whole source and target images, the Relation Network still produced the same score distribution (second row in Fig. 12). When given the whole-car source and target images and a blank image in place of the source marker patch, the network focused on the background, which was most visually similar to the blank image (third row in Fig. 12). Thus, despite being trained to solve four-term visual analogy problems, Relation Network ultimately learned to match subregion images based on visual similarity. In contrast, both the Siamese Network and PCM made use of the whole-car images in their predictions, resulting in very different response distributions when presented with and without whole-car images.

4. Discussion

We compared two computational approaches to visual analogical reasoning. One approach applies task-specific knowledge to reasoning problems, acquired from extensive exposure to highly similar problems. The other approach applies a domain-general comparison procedure to representations of entities and their component structure. The fundamental difference between these two approaches is that the former approach completely merges the formation of perceptual representations and reasoning with them, whereas the latter first learns perceptual representations, to which a domain-general comparison process is then applied.

We introduced the PCM to instantiate the latter approach. We trained this model to segment object parts and identify 3D objects from images in order to ultimately generate representations of object structure in terms of part-whole relations. In Experiment 1, by applying a domain-general comparison procedure to part-based difference vectors, PCM could account for qualitative patterns of human accuracy on four-term visual analogies between images of cars and their subregions. Specifically, PCM's analogy performance, like that of humans, was sensitive to the spatial alignment between analogs, the integrity of analogous subregions as natural parts, and the visibility of analogous subregions in whole-car images.

In Experiment 2, PCM closely approximated human responses on an open-ended analogy task in which participants were asked to freely map parts across car bodies. Since PCM was not trained to solve any analogy problems at all, its performance on the four-term (Experiment 1) and the open-ended task (Experiment 2) both constitute out-of-task testing. In contrast, two popular deep learning models that have been applied to visual analogies—a Siamese Network and a Relation Network—deviated seriously from human performance when confronted with both out-of-sample (but within-task) testing in Experiment 1, and with out-of-task testing in Experiment 2. Indeed, after extensive direct training on four-term analogy problems, the Relation Network did not learn to reason by analogy at all. Instead, the Relation Network acquired a nonanalogical shortcut strategy based on visual similarity. An important methodological implication of these findings is that simply matching (or even exceeding) human overall accuracy on some benchmark task is not a sufficient criterion for inferring that a computational model is simulating human cognitive mechanisms. It is necessary to compare model and human performance in greater detail at the level of controlled manipulations of problem types.

PCM relies on visual processes to learn representations of object parts, offering a parsimonious account of representation learning. Previous approaches to learning similar types of representations have used more computationally intensive (though more efficient) learning processes, such as analogical structure mapping (Chen, Rabkina, McClure, & Forbus, 2019). Unlike such models, PCM does not require that analogical reasoning is already available for use in acquiring representations of entity structure. Rather, we assume analogical reasoning operates on relational representations that can be created by more basic learning mechanisms that operate on nonrelational inputs (Lu et al., 2019). The representations that PCM acquires are expressive enough to provide the basis for a successful approach to visual analogy, based simply on comparisons between relations derived from the model's learned representations.

The failures of the task-specific models in accounting for human performance in the present experiments illustrate a general limitation of treating analogy as the application of task-specific knowledge acquired by simultaneously learning representations of task constituents and also decision procedures for responding on that same task. Such an approach is unlikely to achieve out-of-task generalization. Instead of learning perceptual representations that might be generally useful in multiple tasks, these models acquire representations tailored to the idiosyncrasies of the specific task used in training. A possible way to augment this task-based approach is to incorporate meta-learning, in which a model is trained to solve (and explicitly represent) multiple distinct tasks, and then transfers its acquired knowledge to solve novel tasks, based on their similarity to prior learned tasks (Lampinen & McClelland, 2020). However, it remains to be seen whether meta-learning across a range of analogy tasks can allow a task-based approach to account for the breadth of human analogical reasoning.

The more promising approach, illustrated by PCM, is to first learn componential structure (here, part-whole relations for 3D objects) for visual representations that are generally useful in distinctively different tasks (e.g., object recognition and segmentation) (Berke et al., 2022). By training with varied visual tasks, the learned representations acquire multitask consistency. This approach is broadly consistent with other deep learning models that emphasize the acquisition of representations for use in a range of downstream visual reasoning tasks (Geiger, Carstensen, Frank, & Potts, 2022; Webb et al., 2020; Webb, Sinha, & Cohen, 2020).

We emphasize that we do not propose PCM not as a full-blown computational theory of visual analogy; rather, we use it primarily to instantiate a domain-general approach to visual analogy in order to contrast it with task-specific approaches instantiated by deep learning models. In order to provide a fuller account of the human ability to do analogy, a number of adaptations are necessary. For example, PCM relies heavily on being given representations of all problem stimuli, in order to discriminate among presented answer options. Of course, human reasoners can also *generate* analogy problem solutions. Consider the following four-term problem: $\uparrow : \downarrow :: < : ?$ Without being given any answer options in the problem, human reasoners could easily and reliably generate the correct answer " $>$." In its current form, PCM could not generate any solution in this way. However, one approach to augmenting PCM to ultimately enable it to perform such generative inference is to incorporate a simple pattern completion models to its extant vector representations (e.g., Rumelhart & Abrahamson, 1973), along with an image-rendering processing step. As another example, earlier in this paper, we acknowledged that the difficult nature of many visual reasoning tasks such

as those found in RPM problems, Bongard problems, and SVRT stems from the importance of re-representing problem input until an intuitive rule governing the problem is more easily discernible. In its current form, PCM has no capacity for re-representation. However, an extension might incorporate some probabilistic procedure for representing input, along with some confidence threshold below which the model would sample a new representation. Both of these potential elaborations of the current would help extend its scope to additional aspects of human analogical reasoning.

In the verbal domain, early word-embedding models such as Word2vec acquired lexical representations after having been trained to predict text in sequence within large corpora; these representations were then used to solve simple four-term verbal analogies (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). More recent Large Language Models (LLMs) have been particularly effective in exploiting this approach (Brown et al., 2020), especially when training on text prediction is combined with subsequent human feedback reinforcement learning. These methods yield representations of verbal inputs that support a wide range of capacities, including complex analogical reasoning (Webb, Holyoak, & Lu, 2023). However, text-only LLMs are inapplicable to visual analogy without some external preprocessing that effectively sidesteps the difficult problem of building representations from visual input. Language modeling takes discretized text tokens as inputs, whereas visual perception lacks such well-defined units. Of course, multimodal LLMs that take both text and pixel input data can, in principle, leverage their extensive training to perform visual analogy directly. However, the success of current models on such problems is severely limited by their difficulty with representing relations between visually presented objects (Conwell & Ullman, 2022).

Although PCM requires extensive training to learn its compositional representation for visual inputs, it requires no training at all to solve analogies. Rather, PCM achieves superior performance on our visual analogy tasks simply by comparing its representations of entity structure, using a domain-general similarity measure. Humans can (and machines perhaps may) achieve analogical reasoning by learning representations that encode relational structure, coupled with efficient computation of relational similarity.

Acknowledgments

Preparation of this paper was supported by NSF Grant BCS-1827374 to KJH and HL, and NSF Grant BCS-1827427 to ALY. A preliminary report of part of this research was presented at the 43rd Annual Meeting of Cognitive Science Society. The dataset, model implementation, and experiment data will be released at the website https://osf.io/mkery.

References

Bapst, V., Sanchez-Gonzalez, A., Doersch, C., Stachenfeld, K., Kohli, P., Battaglia, P., & Hamrick, J. (2019). Structured agents for physical construction. International Conference on Machine Learning (pp. 464–474).

Barrett, D. G. T., Hill, F., Santoro, A., Morcos, A. S., & Lillicrap, T. (2018). Measuring abstract reasoning in neural networks. Proceedings of the 35th International Conference on Machine Learning.

Berke, M. D., Walter-Terrill, R., Jara-Ettinger, J., & Scholl, B. J. (2022). Flexible goals require that inflexible perceptual systems produce veridical representations: Implications for realism as revealed by evolutionary simulations. *Cognitive Science*, *46*(10), e13195. https://doi.org/10.1111/cogs.13195

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*, (2), 115–147. https://doi.org/10.1037/0033-295X.94.2.115

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*, (2), 143–177. https://doi.org/10.1016/0010-0285(82)90007-X

Bongard, M. M. (1970). *Pattern recognition*. Spartan Books.

Bromley, J., Bentz, J. W., Bottou, L., Guyon, I., Le Cun, Y., Moore, C., Säckinger, E., & Shah, R. (1993). Signature verification using a "Siamese" time delay neural network. Proceedings of the 7th International Conference of Advances in Neural Information Processing Systems (pp. 737–744).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Ravens Progressive Matrices Test. *Psychological Review*, *97*(3), 404–431.

Chalmers, D. J., French, R. M., & Hofstadter, D. R. (1992). High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental & Theoretical Artificial Intelligence*, *4*, (3), 185–211. https://doi.org/10.1080/09528139208953747

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., & Yu, F. (2015). ShapeNet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*.

Chen, K., Rabkina, I., McClure, M. D., & Forbus, K. (2019). Human-like sketch object recognition via analogical learning. Proceedings of the 33rd AAAI Conference on Artificial Intelligence (pp. 1336–1343).

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the 15th European Conference on Computer Vision (pp. 833–851).

Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., & Yuille, A. (2014). Detect what you can: Detecting and representing objects using holistic models and body parts. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1979–1986).

Conwell, C., & Ullman, T. (2022). Testing relational understanding in text-guided image generation. *arXiv preprint arXiv: 2208.00005*.

Fleuret, F., Li, T., Dubout, C., Wampler, E. K., Yantis, S., & Geman, D. (2011). Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences*, *108*(43), 17621–17625. https://doi.org/10.1073/pnas.1109168108

Forbus, K. D., Gentner, D., Markman, A. B., & Ferguson, R. W. (1998). Analogy just looks like high level perception: Why a domain-general approach to analogical mapping is right. *Journal of Experimental & Theoretical Artificial Intelligence*, *10*(2), 231–257. https://doi.org/10.1080/095281398146842

Geiger, A., Carstensen, A., Frank, M. C., & Potts, C. (2022). Relational reasoning and generalization using non-symbolic neural networks. *Psychological Review*, *130*(2), 308–333.

Gentner, D. (1977). Children's performance on a spatial analogies task. *Child Development*, *48*, (3), 1034. https://doi.org/10.2307/1128356

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. Proceedings of the 13th European Conference on Computer Vision (pp. 630–645).

Hill, F., Santoro, A., Barrett, D., Morcos, A., & Lillicrap, T. (2019). Learning to make analogies by contrasting abstract relational structure. Proceedings of the International Conference on Learning Representations.

Hofstadter, D. R., & Mitchell, M. (1994). The Copycat project: A model of mental fluidity and analogy-making. In K. J. Holyoak & J. A. Barnden (Eds.) *Analogical connections* (pp. 31–112). Ablex Publishing.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*, (3), 427–466. https://doi.org/10.1037/0033-295X.104.3.427

Lake, B. M., & Piantadosi, S. T. (2020). People infer recursive visual concepts from just a few examples. *Computational Brain & Behavior*, *3*, (1), 54–65. https://doi.org/10.1007/s42113-019-00053-y

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253. https://doi.org/10.1017/S0140525×16001837

Lampinen, A. K., & McClelland, J. L. (2020). Transforming task representations to allow deep learning models to perform novel tasks. *Proceedings of the National Academy of Sciences*, *117*, 32970–32981.

Liu, Q., Kortylewski, A., Zhang, Z., Li, Z., Guo, M., Liu, Q., Yuan, X., Mu, J., Qiu, W., & Yuille, A. (2022). Learning part segmentation through unsupervised domain adaptation from synthetic vehicles. Proceedings of the 2022 IEEE Conference on Computer Vision and Pattern Recognition.

Lovett, A., & Forbus, K. (2017). Modeling visual problem solving as analogical reasoning. *Psychological Review*, *124*, (1), 60–90. https://doi.org/10.1037/rev0000039

Lu, H., Liu, Q., Ichien, N., Yuille, A. L., & Holyoak, K. J. (2019). Seeing the meaning: Vision meets semantics in solving pictorial analogy problems. Proceedings of the 41st Annual Meeting of the Cognitive Science Society (pp. 2201–2207).

Lu, H., Wu, Y. N., & Holyoak, K. J. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences*, *116*, (10), 4176–4181. https://doi.org/10.1073/pnas.1814779116

Marr, D. C., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London Series B: Biological Sciences*, *200*, (1140), 269–294. https://doi.org/10.1098/rspb.1978.0020

Matlen, B. J., Gentner, D., & Franconeri, S. L. (2020). Spatial alignment facilitates visual comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *46*, 443–457. https://doi.org/10.1037/xhp0000726

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119.

Raven, J. C. (1938). *Progressive matrices: A perceptual test of intelligence*. Lewis.

Rossiello, G., Gliozzo, A., Farrell, R., Fauceglia, N., & Glass, M. (2019). Learning relational representations by analogy using hierarchical Siamese networks. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3235–3245).

Rumelhart, D. E., & Abrahamson, A. A. (1973). A model for analogical reasoning. *Cognitive Psychology*, *5*(1), 1–28. https://doi.org/10.1016/0010-0285(73)90023-6

Sadeghi, F., Zitnick, C. L., & Farhadi, A. (2015). Visalogy: Answering visual analogy questions. Proceedings of the 29th International Conference on Advances in Neural Information Processing Systems (pp. 1882–1890).

Santoro, A., Raposo, D., Barrett, D. G. T., Malinowski, M., Pascanu, R., Battaglia, P., & Lillicrap, T. (2017). A simple neural network module for relational reasoning. Proceedings of the 31st International Conference on Advances in Neural Information Processing Systems (pp. 4974–4983).

Snow, R. E., Kyllonen, C. P., & Marshalek, B. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (pp. 47–103). Erlbaum.

Sun, Y., Kortylewski, A., & Yuille, A. L. (2020). Amodal segmentation through out-of-task and out-of-distribution generalization with a Bayesian model. Conference on Computer Vision and Pattern Recognition (CVPR).

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H. S., & Hospedales, T. M. (2018). Learning to compare: Relation network for few-shot learning. Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1199–1208).

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General*, *113*(2), 169–193. https://doi.org/10.1037/0096-3445.113.2.169

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, https://doi.org/10.1038/s41562-023-01659-w

Webb, T. W., Dulberg, Z., Frankland, S., Petrov, A., O'Reilly, R., & Cohen, J. (2020). Learning representations that support extrapolation. International Conference on Machine Learning (pp. 10136–10146).

Webb, T. W., Sinha, I., & Cohen, J. D. (2020). Emergent symbols through binding in external memory. *arXiv preprint arXiv:2012.14601*.

Zhang, C., Gao, F., Jia, B., Zhu, Y., & Zhu, S.-C. (2019). RAVEN: A dataset for Relational and Analogical Visual rEasoNing. Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (pp. 5312–5322).

Zheng, Y., Matlen, B., & Gentner, D. (2022). Spatial alignment facilitates visual comparison in children. *Cognitive Science*, *46*(8), e13182. https://doi.org/10.1111/cogs.13182

Zhila, A., Yih, W.-t., Meek, C., Zweig, G., & Mikolov, T. (2013). Combining heterogeneous models for measuring relational similarity. *Proceedings of the 2013 Conference North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1000–1009).

Zhu, H., Tang, P., Park, J., Park, S., & Yuille, A. (2019). Robustness of object recognition under extreme occlusion in humans and computational models. *arXiv preprint arXiv:1905.04598*.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supplementary Material