UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

From Vision to Reasoning: Probabilistic Analogical Mapping Between 3D Objects

Permalink

https://escholarship.org/uc/item/05x3m098

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 44(44)

Authors

Fu, Shuhao
Lu, Hongjing
Holyoak, Keith

Publication Date

2022

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

From Vision to Reasoning: Probabilistic Analogical Mapping Between 3D Objects

Shuhao Fu[1] fushuhao@g.ucla.edu
Keith J. Holyoak[1] holyoak@lifesci.ucla.edu
Hongjing Lu[1,2] hongjing@ucla.edu

[1] Department of Psychology
[2] Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095 USA

Abstract

We see the external world as consisting not only of objects and their parts, but also of relations that hold between them. Visual analogy, which depends on similarities between relations, provides a clear example of how perception supports reasoning. Here we report an experiment in which we quantitatively measured the human ability to find analogical mappings between parts of different objects, where the objects to be compared were drawn either from the same category (e.g., images of two mammals, such as a dog and a horse), or from two dissimilar categories (e.g., a chair image mapped to a cat image). Humans showed systematic mapping patterns, but with greater variability in mapping responses when objects were drawn from dissimilar categories. We simulated the human response of analogical mapping using a computational model of mapping between 3D objects, visiPAM (*visual Probabilistic Analogical Mapping*). VisiPAM takes point-cloud representations of two 3D objects as inputs, and outputs the mapping between analogous parts of the two objects. VisiPAM consists of a visual module that constructs structural representations of individual objects, and a reasoning module that identifies a probabilistic mapping between parts of the two 3D objects. Model simulations not only capture the qualitative pattern of human mapping performance cross conditions, but also approach human-level reliability in solving visual analogy problems.

Keywords: vision; analogy; mapping; graph matching; deep learning

Introduction

Suppose a preschooler is asked questions about pictured objects, such as, “If a tree had a knee, where would it be?” or “Can you point to the eyes of this car?” Children often provide reasonable answers to such questions, providing evidence of the creative nature of human intelligence (Gentner, 1977). Such findings show that the ability to see visual analogies develops early. Visual analogies can also contribute to vivid communication. In a humorous explanation of radio communication, Albert Einstein remarked, "... the wire telegraph is a kind of very, very long cat. You pull his tail in New York and his head is meowing in Los Angeles…. And radio operates exactly the same way: you send signals here, they receive them there. The only difference is that there is no cat." Einstein’s analogy depends on a visual mapping between the imaginary cat and a telegraph line, stripped away from the specific features of either, but linked to the functions of signal and receiver. The analogy is then abstracted further: in the case of radio, there is no solid connector (“no cat”) linking signal and receiver. As an aside, the joke arises from the violation of expectation: Einstein deliberately uses a vivid analogy to *not* convey any insight into the causal mechanism underlying a new technology.

In addition to humor, Einstein’s remark illustrates the operation of analogical reasoning on visuospatial representations. Humans are clearly able to identify systematic relational correspondences between visualizable entities, based on visual imagery (often in response to verbal input), pictures (either realistic or schematic), and/or three-dimensional objects. Whereas analogies stated in language have relations provided as part of the input (via verbs and relational phrases), analogies based on visual inputs more obviously depend on perceptual mechanisms to achieve the *eduction of relations* (Spearman, 1923): the extraction of relations from non-relational inputs, such as pixels in images or spatial mesh regions in 3D objects. Reasoning by analogy from raw visual input (e.g., thousands of pixels or meshes) is clearly a challenging computational problem that demands the integration of perception with reasoning.

Computational and psychological work on visual analogy has largely focused on problems inspired by the Raven’s Progressive Matrices (RPM) (Raven, 2000). After extensive training with RPM-style problems, deep neural networks have achieved human-level performance on test problems with similar basic structures (e.g., Santoro et al., 2017; Zhang et al., 2019). However, the success of these deep learning models depends on datasets of massive numbers (sometimes more than a million) of RPM-style problems, which makes the deep-learning approach fundamentally different from human analogical reasoning. When the RPM task is administered to a person, “training” is limited to general task instructions with at most one practice problem.

A further limitation of empirical and computational work on visual analogy is that efforts have been largely focused on problems based on simple line-drawn geometric forms or line-drawn pictures (e.g., Sternberg, 1977; Krawczyk et al., 2008; Richland, Morrison & Holyoak, 2006). Relatively few studies have used images, such as pictures of cars (Ichien et la., 2021), line drawings (Lu et al., 2019) or photos of human interactions (Green et al., 2017). In addition to the limited range of stimulus types, the form of analogy tasks has also been very constrained. The most common task used in studies of visual analogy is to ask participants to select a valid analogical completion among several invalid foils

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.* ©2022 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

(e.g., Krawczyk et. al., 2008; Green et al., 2017; Lu et al., 2019). Although such forced-choice tasks can be useful to test specific hypotheses, participants' performance heavily depends on what distractors are used.

A different paradigm, first used in the classic study by Gentner (1977), is to let people annotate analogous parts of objects. Gentner asked children and adults to place two dots on a line-drawn object (either a tree or mountain) that would correspond to two human body parts (e.g., mouth and a knee). The advantage of this marker-placement tasks is that it provides a direct measure of analogical mapping that does not require predefining the "correct" response, and which avoids biases that might be triggered by the choice of foils.

Here we propose a flexible computational model of visual analogy for 3D objects, and compare its predictions to findings from an experiment in which people made judgments about correspondences between a range of familiar 3D objects. We focused on analogical mapping between object parts because human perception and thinking show sensitivity to part-whole relations across both visual and semantic domains (e.g., Tversky & Hemenway, 1984; Lee et al., 2021). Our experimental paradigm was adapted from the marker task introduced by Gentner (1977). We refined the paradigm by using analogies based on images of 3D objects (without verbal cues), and obtained more fine-grained quantitative measures of judged correspondences.

Human Experiment

The goal of the experiment was to measure human mapping performance for visual analogy problems in which the two 3D images were drawn either from the same category (e.g., dog and horse images as the analogs), or from two distinctively different categories (e.g., the source image was a chair and the target image was a cat).

Participants Fifty-nine participants (mean age = 20.55 years; 51 female) were recruited from the Psychology Department subject pool at the University of California, Los Angeles. All participants were compensated with course credit.

Stimuli 3D object stimuli were selected from two publicly available datasets used in computer vision: ShapeNetPart (Yi et al., 2016) and a 3D animal dataset ("Animal Pack Ultra 2") from Unreal Engine Marketplace. Nine chairs were selected from the ShapeNetPart dataset (each chair with a different shape), and nine animals from the Animal Pack dataset: horse, buffalo, Cane Corso, sheep, domestic pig, Celtic wolfhound, African elephant, Hellenic hound, and camel. We used the Blender software to render 2D images from the 3D models. The stimulus images were generated using a constant lighting condition, with a gray background. Multiple camera positions were sampled for each object, with 30° separation between camera angles for depth rotation. Two undergraduate research assistants manually annotated the keypoints (i.e., center locations) of predefined parts on the 3D objects. Chair parts included seat, back, and chair legs, while animal parts included spine of torso, head, and legs.

We generated 192 pairs of images. A few examples of the stimuli are shown in Figure 1. Each image pair included a source image (either a chair or an animal), which was annotated with two markers on two different parts of the object. To generate marker locations for source images, we first rendered the images using corresponding 3D object models, and then calculated marker locations on the rendered 2D images using a perspective projection for the predefined camera position. In the within-category condition, the source and target images were from the same general object category (e.g., two images of animals). In the between-category condition, the two images were from different object categories (e.g., a chair image with an animal image). The two objects in an image pair were shown in the same orientation.

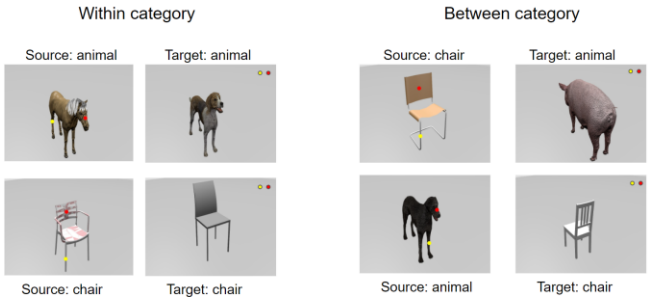


Figure 1. Sample stimuli. Left panel: within-category trials with source and target images from the same object category. Right panel: between-category trials with images from different object categories.

Procedure To measure human mapping judgment, we asked participants to perform a visual analogy task adapted from that used by Gentner (1977). On each trial, participants were presented with one image pair on a computer screen as shown in Figure 1, and completed a marker-placement task. For each of 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. No time constraint was imposed; the entire experiment was completed in about 41 minutes on average. On each trial, the exact location of each marker placement was recorded.

Results Five out of the 59 participants were removed from analysis either because they indicated they were not serious, or because they moved less than 30% of the markers. Thus, data from a total of 54 participants were included in analyses.

Figure 2 shows two representative examples of human responses. The target image was the same for these two comparisons (a horse), while the source image was either a different animal (a Celtic wolfhound) or a chair. The locations marked as analogous by different participants are shown as a heatmap on the target image. These examples

illustrate the general pattern of human performance on the task. Human responses in identifying analogous parts were not idiosyncratic: rather, marker placements were relatively consistent across participants, especially for within-category image pairs (left panel). Variability of responses among participants was greater for between-category comparisons (right panel).

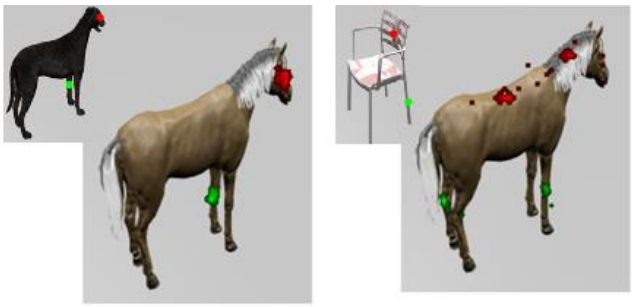


Figure 2. Example heatmaps of human marker placements on target images for two comparisons. The source images have been reduced in size for the purpose of illustration.

For each comparison of an image pair, 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 of marked locations across participants. For the within-category image pairs, the mean distance to the mean marker location was around 8 pixels. Relative to the object sizes (average height of 213 pixels and width of 135 pixels), 8-pixel variability indicates strong agreement of people's judgments in analogical mapping.

Overall similarity between source and target images influenced human response consistency in identifying analogous parts. As shown in Figure 3, variability in mapping judgments was higher when the two analog objects were from distinctively different categories (mean 31.66 pixels) than from the same category (mean 7.66 pixels). A repeated-measures ANOVA with two within-subjects factors (within- vs. between-categories for source and target images, and type of target images) revealed a reliable main effect of category consistency between source and target images, $F(1,192) = 971.92,\ p < .001$. The main effect of target category was not reliable, but a two-way interaction effect was found, $F(1,192) = 15.75,\ p < .001$. This interaction reflects greater within-category distances for comparisons between pairs of chairs than of animals, likely due to greater shape variability among the set of chairs than the set of animals used in the experiment.

Human responses for some between-category problems showed sub-clusters in marked locations. In the example shown in Figure 2 (right), some participants mapped the back of the chair to the head of the horse, as both parts extend out from the main "body" of the object. Other participants instead mapped the back of the chair to the back of the horse, likely based on conceptual knowledge of semantic labels for parts. To ensure that our findings were not solely due to use of a single mean placement for each problem, the KMeans++ algorithm (Arthur & Vassilvitskii, 2007) was applied to human-marked locations for each between-category problem, identifying two clusters for each problem. We then redid the distance analysis for human responses by calculating distances to the closer center of two clusters for between-category image pairs. Using this revised distance measure for between-category placements, mean distance from the closer mean placement was reduced from 31.66 pixels (SD = 13.39) to 13.27 pixels (SD = 4.60). A repeated-measures ANOVA using the new distance measures (based on the closer mean placement) continued to reveal a significant effect of category consistency on distance from the closest mean placement, $F(1,99) = 56.38,\ p < .001$.

visiPAM: From Vision to Analogy

To model human judgments in the visual analogy task, we developed a model, visiPAM, that extends an approach previously applied to verbal analogies (Lu, Ichien, & Holyoak, in press). The overall framework (Figure 4) involves (1) training a visual module to create vector-based representations of visual features for 3D objects, (2) using the learned visual model to form structural representations of individual objects coded as attributed graphs, and then (3) inferring analogical mappings using a probabilistic graph matching algorithm that aims to maximize similarity between mapped analogs subject to a soft isomorphism constraint (preference for one-to-one correspondences).

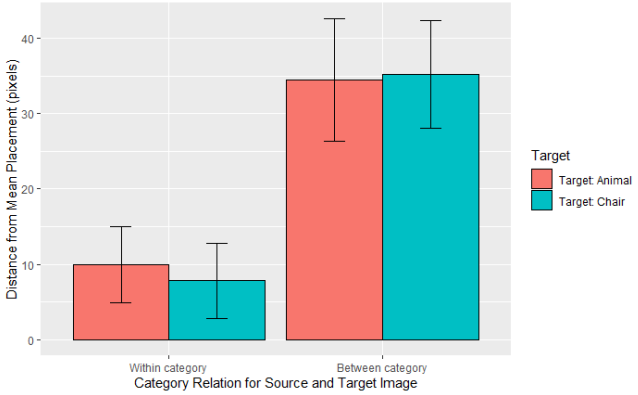


Figure 3. Human judgments in the marker-placement task. Mean distances of marked locations to the mean placements varied as a function of whether the source and target images were drawn from the same or different categories, and which category was used in the target image.

Visual module: Structural representation of objects

The basic aim of the visual module in visiPAM is to form a part-based structural representations of 3D objects that captures both local 3D shape information about parts and the relative spatial relations between parts (see Figure 5). We employed a type of deep neural network, a Dynamic Graph Convolutional Neural Network (DGCNN) (Wang et al., 2019) to capture shape features of 3D objects. The network takes as input a set of 3D points of an object (termed a *point cloud*), which is used to accomplish a wide range of tasks, including object classification and semantic part segmentation. The core component of DGCNN is the EdgeConv operation: for each point, the layer aggregates information from the *K* nearest neighboring points through a nonlinear function to learn features $e_{ij} = h(x_i, x_j)$, where the $h(\cdot)$ function itself is a shared-weight Multilayer Perceptron. The point cloud first passes through three layers of EdgeConv operations. The features created by each EdgeConv layer are max-pooled globally to form a vector, and concatenated with each other to combine geometric properties. These features are then passed to four additional MLP layers to produce a segmentation prediction for each 3D point.

The DGCNN is trained on a supervised part segmentation task (Yi et al., 2016) using 16 types of 3D objects drawn from the ShapeNetPart dataset, which contains about 17,000 models of 3D objects from 16 rigid object categories, including cars, airplanes, and chairs. Each 3D object is annotated with 2-6 parts. After training with a part segmentation task, DGCNN is able to extract local geometric properties from nearby 3D points and encode these as embedding features. Hence, DGCNN transforms the three-dimensional input (x, y, z coordinates) of each 3D point of the object into a 64-dimension embedding vector in the third EdgeConv layer. These embeddings capture critical local geometric properties of 3D shapes, and thus represent informative visual features associated with object parts.

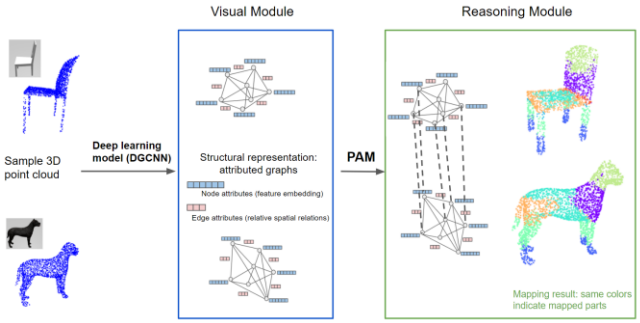


Figure 4. Overview of visiPAM. Left: The visual module takes as input a set of 3D points of an object and forms a structural representation of the object as an attributed graph. Right: The reasoning module then identifies the optimal mappings between nodes in the attributed graphs for two objects. Points on parts predicted to be analogous are coded with the same color in the two objects.

In all the simulation results reported below, the DGCNN was trained on the ShapeNetPart dataset only. Critically, the DGCNN was only trained on man-made objects in the ShapeNetPart dataset and was never trained with 3D animals. About 2000 points were used to represent each 3D object. To reduce the computation cost in the reasoning module, a cluster algorithm (KMeans++ algorithm; Arthur & Vassilvitskii, 2007) is applied to point embeddings to group the points into eight clusters. Each cluster includes 3D points that share similar visual features of geometric shapes. Although this clustering is based entirely on visual embeddings, each cluster typically corresponds to a semantically meaningful part of the object. The clustering algorithm thus approximates the formation of visual representations of object parts.

Using the pipeline described above (see Figure 3), an attributed graph with eight nodes can be constructed to form a structural representation of any 3D object. The part clusters provide the nodes and mean embedding vectors for each cluster constitute node attributes. The eight nodes are fully interconnected to form attributed edges that capture the relative spatial relations among object parts. The relative spatial relations between parts are used to form edges in the attributed graph. For each object part, we calculate the center location by averaging 3D coordinates of points in one cluster. For any pair of parts, we capture spatial relations using three angular distances between cluster centers of the two parts and the object centroid. We denote the 3D coordinates of two cluster centers for two object parts as $\boldsymbol{c_i}$ and $\boldsymbol{c_j}$, and the center location of the whole object as $\boldsymbol{c_0}$. A relation vector that includes three elements is computed using the cosine distances (i.e., angular rotation): $(\cos(\boldsymbol{c_i} - \boldsymbol{c_j}, \boldsymbol{c_i} - \boldsymbol{c_0}), \cos(\boldsymbol{c_i} - \boldsymbol{c_0}, \boldsymbol{c_j} - \boldsymbol{c_0}), \cos(\boldsymbol{c_i} - \boldsymbol{c_j}, \boldsymbol{c_j} - \boldsymbol{c_0}))$. Note that the relation vector is invariant to object rotation.

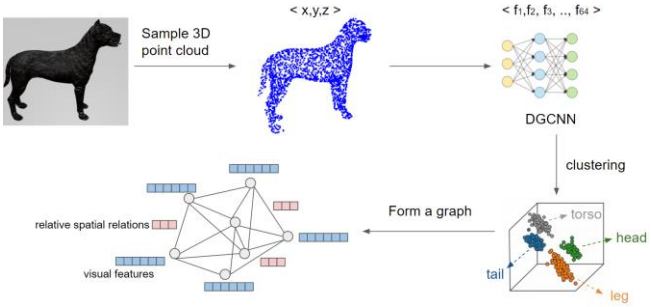


Figure 5. The visual module forms a part-based structural representation of each object organized into an attributed graph.

Reasoning module: Probabilistic Analogical Mapping (PAM)

After forming the structural representation objects in the form of attributed graphs, the reasoning module uses the Probabilistic Analogical Mapping (PAM) model (Lu et al., 2022) to identify correspondences between analogous parts across the two objects. Essentially, PAM is a constrained

graph-matching algorithm that operates on a pair of attributed graphs, G and G', composed of nodes N approximately corresponding to object parts, edges E coding spatial relations between parts, node attributes F based on visual embeddings extracted from the DGCNN model, and edge attributes R based on vectors of relative spatial relations between parts. Let i and j be indices of nodes in the graph. F_i indicates the node attribute (visual features) of the ith node and R_{ij} indicates the spatial relation of the edge linking the ith node to the jth node. We denote the attributed graph for source objects as $G = (N, E, F, R)$ and that for target objects as $G' = (N', E', F', R')$. We use i and j as indices of nodes in the source graph G, and i' and j' as node indices for the target graph G'.

PAM adopts Bayesian inference to estimate the probabilistic mapping matrix ***m,*** consisting of elements denoting the probability that the ith node in the source analog maps to the i'th node in the target analog, $m_{ii'} = P(M_{ii'} = 1)$. $M_{ii'} = 1$ if the ith node in the source object maps to the i'th node in the target object, and $M_{ii'} = 0$ if the two nodes are not mapped. The mapping follows the constraints $\forall i \sum_{i'} M_{ii'} = 1, \forall i' \sum_{i} M_{ii'} = 1$. The optimal mapping identified by PAM is based on maximizing the posterior probability $P(m|G, G')$:

$$P(m|G, G') \propto \& P(G, G'|m)P(m), \quad (1)$$
$$\text{with the constraints } \forall i \sum_{i'} m_{ii'} = 1, \forall i' \sum_{i} m_{ii'} = 1,$$

where the prior term favors isomorphic (one-to-one) mappings, defined as

$$P(m) = e^{\frac{1}{\beta}\sum_i \sum_{i'} m_{ii'} \log m_{ii'}}. \quad (2)$$

The likelihood term $P(G, G'|m)$ is based jointly on visual feature similarity between mapped nodes and relation similarity between mapped parts, defined in the log form as

$$\log(P(G, G'|m)) = (1-\alpha) \sum_i \sum_j \sum_{i'} \sum_{j'} m_{ii'} m_{jj'} S\left(R_{ij}, R'_{i'j'}\right) + \alpha \sum_i \sum_{i'} m_{ii'} S(F_i, F'_{i'}), \quad (3)$$

where $S(\cdot)$ is the cosine similarity function for visual features (node attributes) and for spatial relations (edge attributes). Thus, the first term in Equation 3 corresponds to the weighted sum of edge (relation) similarities multiplied by the corresponding mapping probability, and the second term corresponds to the weighted sum of node (visual feature) similarities multiplied by the corresponding mapping probability. The parameter α is a weight that controls the relative importance of spatial relation similarity (edges) versus visual similarity (nodes), consistent with psychological evidence that a variety of factors can alter human sensitivity to relation versus entity-based similarity. In the simulation, this parameter is set to a constant value of 0.5. PAM is implemented using the graduated assignment algorithm (Gold & Rangarajan, 1996), which iteratively converges on a soft assignment of mapping variables between the source and target analogs.

Using visiPAM to generate placement predictions

The input to visiPAM is the 3D point cloud for each object used in the human experiment. The model is also given the camera orientation for each image used in the experiment, and 3D coordinates of markers for source objects. For each pair of images to be compared, the model takes the point clouds of both objects as the input and employs the DGCNN network to generate structural part representations of the two objects as attributed graphs. The reasoning module then uses the PAM model to identify mappings of analogous parts between the two 3D objects. After the center locations of parts in the two 3D objects are mapped, the marker locations in the target point cloud that are analogous to the markers in the source point cloud are identified by computing the relative locations in the mapped target cluster. The final position of the target markers is determined by the distance of the point to the desired location within the cluster. Since all 3D point clouds are normalized to the same scale, this method works reasonably well for our experiment. After obtaining the mapped location in 3D, the final step is simply to project it onto the 2D image (given camera parameters) in order to compare the model's prediction with human marker placements.

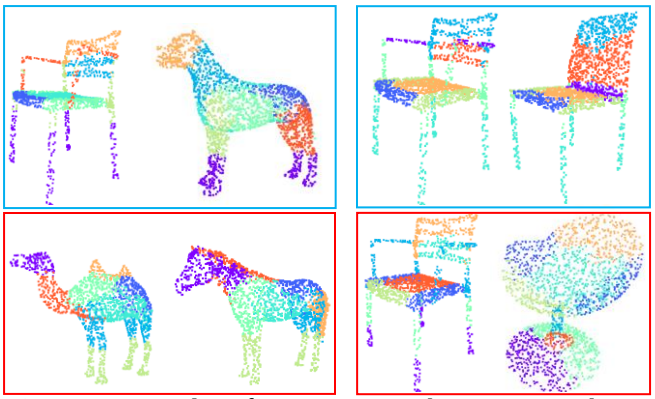


Figure 6. Examples of part mappings between two objects represented as point clouds, generated by visiPAM. The top two examples (in blue boxes) generated sensible mappings (e.g., legs of chair to legs of dog). The bottom two examples (in red boxes) generated some sensible part mappings (e.g., head of camel to head of horse), but also some apparent mismappings (e.g., hump of the camel to tail of the horse; seat of the left chair to stand of the right chair).

Results A few examples of mappings generated by PAM (both successful and less successful) are shown in Figure 6. To compare the model's performance with human responses, we applied the model to all 192 pairs of images used in the experiment, and measured the distance between the marker location predicted by visiPAM to the mean locations of human placements for each pair. We then compared the

distance measure from the model with the mean for human participants. Overall, marker locations of analogous parts predicted by visiPAM were an average of 29 pixels from mean locations of human placements, close to the average human distance to mean locations (20 pixels). Relative to the object sizes, the model and human distances were very similar.

Figure 7 shows a violin plot of human and model distances from mean human placements across the various conditions. As was the case for participants in the experiment, visiPAM's predicted placements were closer to the human mean when the two images were drawn from the same rather than different categories. Model predictions were within 1 std of human distances in the between-category condition, indicating that visiPAM's mapping predictions are near the reliability of human judgments for far analogy problems. Model accuracy was comparable regardless of whether the target was a chair or an animal. This finding suggests that visiPAM is able to generalize its mapping ability to untrained objects.

In addition, we calculated the item-level correlation across the 192 analogy problems between average human distances from mean placement locations and distances of the model predictions from the same mean locations. The model reliably predicted human responses at the item level, $r = 0.58$.

Ablation analysis It is possible to separately evaluate the contributions of visual features (node embeddings) and relations (edge attributes) to visiPAM's mapping performance. Parameter $\boldsymbol{\alpha}$ in Equation 3 controls the relative importance of visual feature similarity and spatial relation similarity in determining mappings. When the contribution of relations is removed (i.e., $\boldsymbol{\alpha} = \mathbf{1}$), the correlation between model and human distance measures was reduced from 0.58 (default model including both visual features and relations) to 0.44. When the contribution of visual feature embeddings is removed (i.e., $\boldsymbol{\alpha} = \mathbf{0}$), the correlation between model and human distances was reduced to 0.12. These ablation results confirm that both visual features and spatial relations contribute to visPAM's ability to identify analogous parts across objects.

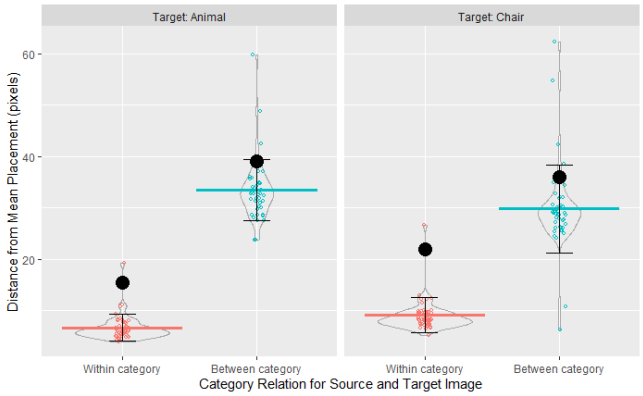


Figure 7. Violin plot of human placements and visiPAM predictions. Each colored dot indicates average distance of marker locations from the human mean for one individual participant. Large black dots indicate visiPAM predictions. Horizontal lines indicate mean human distances, and the error bars indicate one standard deviation.

Discussion

Using a marker-placement task, we found that people can identify consistent mappings between parts of two distinctively different 3D objects, even when the objects being compared are drawn from very different categories (e.g., a chair and a dog). We present a new analogy model, visiPAM, that is able to operate on 3D shapes of visual inputs and compute mappings comparable to those identified by humans. To the best of our knowledge, visiPAM is the first model that can compute analogies between 3D objects.

Most previous machine-learning models that can solve visual analogy problems from pixel-level inputs (e.g., Santoro et al., 2017; Zhang et al., 2019). However, machine-learning models that were originally designed to solve analogy problems based on simple geometric patterns have failed to generalize to analogy problems based on realistic images (Ichien et al., 2021). In contract, visiPAM does not require end-to-end training on massive numbers of analogy problems. Rather, the approach represented by visiPAM assumes that analogical reasoning is a similarity-based cognitive mechanism that naturally operates on representations formed to perform a wide range of perceptual and/or cognitive tasks. Once suitable structural representations have been acquired from raw inputs, analogical reasoning provides a mechanism that promotes generalization and knowledge transfer. The visual module in visiPAM uses a deep learning model that is trained with supervision to perform object classification and part segmentation for a range of 3D objects, coupled with an unsupervised clustering algorithm. The reasoning module, which operates without any training at all, succeeded in finding mappings for images taken from an object category (animals) on which the visual module has not been trained. The integration of structured visual representations coded as visual feature embedding and spatial relations organized into graphs, with a probabilistic mapping algorithm, yields robust analogical mapping that generalizes to novel object categories. Previous models of visual analogical reasoning have also operated on structured visual representations (e.g., Chen et al., 2019; Lovett & Forbus, 2017; Doumas et al., 2022), but visiPAM goes beyond previous models by operating on raw perceptual inputs.

The present results indicate that visiPAM achieves reliability comparable to humans in our marker-placement task, encouraging the possibility of extending the model to other visual analogy tasks. We believe that achieving human-level, generalizable analogical reasoning will require synergy between deep learning with big data (to acquire suitable representations) and similarity-based reasoning over relational structures. Vision and reasoning must be closely coupled in order to "see" the correspondences between distinct objects and scenes.

Acknowledgements

Preparation of this paper was supported by NSF Grant BCS-1827374 awarded to K.J.H and AFRL grant FA8650-19-C-1692/ S00017 to H.L.

References

Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 1027–1035.

Bromley, J., Guyon, I., LeCun, Y., Sackinger, E., & Shah, R. (1993). Signature verification using a "Siamese" time delay neural network. In J. Cowan & G. Tesauro (Eds.), *Advances in Neural Information Processing Systems*.

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

Doumas, L. A., Puebla, G., Martin, A. E., & Hummel, J. E. (2022). A theory of relation learning and cross-domain generalization. *Psychological Review*.

Gentner, D. (1977). Children's performance on a spatial analogies task. *Child Development*, *48*(3), 1034–1039.

Gold, S., & Rangarajan, A. (1996). A graduated assignment algorithm for graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18*(4), 377-388.

Green, A. E., Kenworthy, L., Gallagher, N. M., Antezana, L., Mosner, M. G., Krieg, S., ... & Yerys, B. E. (2017). Social analogical reasoning in school-aged children with autism spectrum disorder and typically developing peers. *Autism*, *21*(4), 403-411.

Ichien, N., Lu, H., & Holyoak, K. J. (2019). Individual differences in judging similarity between semantic relations. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Meeting of the Cognitive Science Society* (pp. 464-470). Austin, TX: Cognitive Science Society.

Ichien, N., Liu, Q., Fu, S., Holyoak, K. J., Yuille, A. L., & Lu, H. (2021). Visual analogy: Deep learning versus compositional models. In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Krawczyk, D. C., Morrison, R. G., Viskontas, I., Holyoak, K. J., Chow, T. W., Mendez, M. F., ... & Knowlton, B. J. (2008). Distraction during relational reasoning: The role of prefrontal cortex in interference control. *Neuropsychologia*, *46*(7), 2020-2032.

Lee, A. L., Liu, Z., & Lu, H. (2021). Parts beget parts: Bootstrapping hierarchical object representations through visual statistical learning. *Cognition, 209,* 104515.

Lovett, A. & Forbus, K. (2017). Modeling visual problem solving as analogical reasoning. *Psychological Review*, *124*, 60-90.

Lu, H., Liu, Q., Ichien, N., Yuille, A. L., & Holyoak, K. J. (2019). Seeing the meaning: Vision meets semantics in solving visual analogy problems. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Meeting of the Cognitive Science Society* (pp. 2201-2207). Austin, TX: Cognitive Science Society.

Lu, H., Ichien, N., & Holyoak, K. J. (2022). Probabilistic analogical mapping with semantic relation networks. *Psychological Review.* DOI: https://doi.org/10.1037/rev0000358

Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology*, *41*(1), 1-48.

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006) Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology*, 94, 249-273.

Santoro, A., Raposo, D., Barrett, D. G., Malinowski, Pascanu, R., Battaglia, P., & Lillicrap, T. (2017). A simple neural network module for relational reasoning. *Advances in Neural Information Processing Systems*, 4967-4976.

Spearman, C. (1923). *The nature of intelligence and the principles of cognition.* London: Macmillan.

Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities.* Hillsdale, NJ: Erlbaum.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General*, *113*(2), 169-193.

Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., & Solomon, J. M. (2019). Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics*, *38*(5), Article No. 146.

Yi, L., Kim, V. G., Ceylan, D., Shen, I.-C., Yan, M., Su, H., Lu, C., Huang, Q., Sheffer, A., & Guibas, L. (2016). A scalable active framework for region annotation in 3D shape collections. *SIGGRAPH Asia*.

Zhang, C., Jia, B., Gao, F., Zhu, Y., Lu, H., & Zhu, S. (2019). Learning perceptual inference by contrasting. In *33rd Conference on Neural Information Processing Systems.*