Modeling Gestalt Visual Reasoning on Raven's Progressive Matrices Using Generative Image Inpainting Techniques

Tianyu Hua TIANYU.HUA@VANDERBILT.EDU
Maithilee Kunda MKUNDA@VANDERBILT.EDU
Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN USA

Abstract

Psychologists recognize Raven's Progressive Matrices as a useful test of general human intelligence. While many computational models investigate various forms of top-down, deliberative reasoning on the test, there has been less research on bottom-up perceptual processes, like Gestalt image completion, that are also critical in human test performance. In this work, we investigate how Gestalt visual reasoning on the Raven's test can be modeled using generative image inpainting techniques from computer vision. We demonstrate that a reasoning agent that has access to an off-the-shelf inpainting model trained only on photorealistic images of objects achieves a score of 27/36 on the Colored Progressive Matrices, which corresponds to average performance for nine-year-old children. We also show that when our agent uses inpainting models trained on other datasets (faces, places, and textures), it does not perform as well. Our results illustrate how learning visual regularities in real-world images can translate into successful reasoning about artificial test stimuli. On the flip side, our results also highlight the limitations of such transfer, which may contribute to explanations for why intelligence tests like the Raven's are often sensitive to people's individual sociocultural backgrounds.

1. Introduction

Consider the matrix reasoning problem in Figure 1; the goal is to select the choice from the bottom that best fits into the portion on top. Such problems are found on many different human intelligence tests (Roid & Miller, 1997; Wechsler, 2008), including on the Raven's Progressive Matrices tests, which are considered to be the most effective single measure of general intelligence across all psychometric tests (Snow et al., 1984).

As you may have guessed, the solution to this problem is answer choice #2. An interesting aspect of this problem is that there are multiple ways to solve it. For example, one might take a top-down, deliberative approach by first deciding that the top two elements are reflected across the horizontal axis, and then reflecting the bottom element to predict an answer–often called an Analytic approach. Alternatively, one might just "see" the answer emerge in the empty space, in a more bottom-up, automatic fashion–often called a Gestalt or figural approach.

Evidence from psychology (Kirby & Lawson, 1983; DeShon et al., 1995; Lynn et al., 2004) and neuroscience (Prabhakaran et al., 1997) supports the existence of these types of diverse strategies in human test performance, as well as surmise about the algorithmic underpinnings of these reasoning

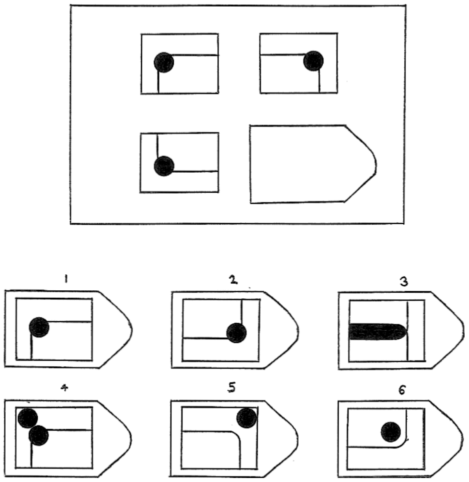

Figure 1. Raven's-like problem eliciting a "Gestalt" strategy.

processes in Hunt's early conceptual analysis of Raven's problem solving (Hunt, 1974). Part of what makes Raven's tests difficult and interesting for humans is that the test allows for many different reasoning strategies, which means that the test is as much a test of metacognitive strategy selection as it is a test of elementary reasoning operations. (This also may contribute to why Raven's test performance is such a good proxy for general intelligence.)

While many artificial reasoning agents explore variations of the Analytic approach, less attention has been paid to the Gestalt approach, though both are critical in human intelligence. In human cognition, Gestalt principles refer to a diverse set of capabilities for detecting and predicting perceptual regularities such as symmetry, closure, similarity, etc. (Wagemans et al., 2012).

Here, we investigate how Gestalt reasoning on the Raven's test can be modeled with generative image inpainting techniques from computer vision. In this paper:

- We describe an agent that solves Raven's problems through Gestalt visual reasoning, using an off-the-shelf image inpainting model as a component.
- We demonstrate that our agent, using an inpainting model trained on object photographs from ImageNet, achieves a score of 27/36 on the Raven's Colored Progressive Matrices test.
- We show that performance is sensitive to the inpainting model's training data. Using models trained on faces, places, and textures yield scores of 11, 17, and 18, respectively, and we offer some potential reasons for these differences.

At minimum, our results demonstrate (i.e. proof by existence) that a computational Gestalt-like strategy is sufficient for solving many Raven's test items, to a level of performance beyond what would likely be expected from such a seemingly simplistic problem-solving strategy. While it is unlikely that people typically use a pure Gestalt strategy like that embodied by our model, it is likely that such a strategy might form part of a person's Raven's problem-solving toolkit. Furthermore, the fact that it is computationally possible, using the relatively parsimonious reasoning mechanisms we propose here, means that we cannot rule out that a particular person might be solving Raven's items in this way (Kunda & Goel, 2011).

Our work also opens up a new avenue of inquiry for visuospatial problem solving mechanisms in cognitive systems research, with implications for visuospatial intelligence tests but also for other types of reasoning tasks that involve "visualizing" the solution from surrounding clues. For example, when we reason about occlusion in the real world, there is likely some aspect of "inpainting" that we use to make inferences about what we cannot see, and these inferences are all the more important when we are transferring knowledge learned from one context to another.

2. Background

2.1 Background: Gestalt Reasoning

In humans, Gestalt phenomena have to do with how we integrate low-level perceptual elements into coherent, higher-level wholes (Wagemans et al., 2012). For example, the left side of Figure 2 contains only scattered line segments, but we inescapably see a circle and rectangle. The right side of Figure 2 contains one whole key and one broken key, but we see two whole keys with occlusion. In psychology, studies of Gestalt phenomena have enumerated a list of principles (or laws, perceptual/reasoning processes, etc.) that cover the kinds of things that human perceptual systems do (Wertheimer, 1923; Kanizsa, 1979).

Likewise, work in image processing and computer vision has attempted to define these principles mathematically or computationally (Desolneux et al., 2007). In more recent models, Gestalt principles are seen as emergent properties that reflect, rather than determine, perceptions of structure in an agent's visual environment.

For example, early approaches to image inpainting—i.e., reconstructing a missing/degraded part of an image—used rule-like principles to determine the structure of missing content, while later, machine-learning-based approaches attempt to learn structural regularities from data and apply them to new images (Schönlieb, 2015).

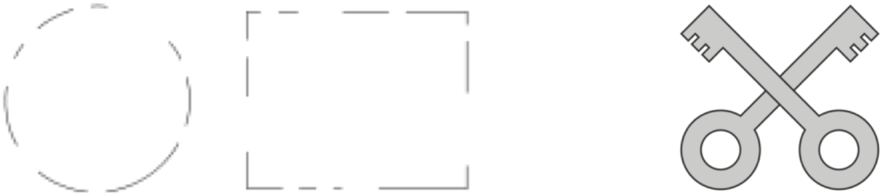

Figure 2. Images eliciting Gestalt "completion" phenomena.

2.2 Background: Image Inpainting in Computer Vision

Machine-learning-based inpainting techniques typically either borrow information from within the occluded image itself (Bertalmio et al., 2000; Barnes et al., 2009; Ulyanov et al., 2018) or from a prior learned from other images (Hays & Efros, 2008; Yu et al., 2018; Zheng et al., 2019). The first approach often uses patch similarities to propagate low-level features, such as the texture of grass, from known background regions to unknown patches. Such approaches can suffer on images with low self-similarity or when the missing part involves semantic-level cognition, e.g., a part of a face.

The second approach aims to generalize regularities in visual content and structure across different images, and several impressive results have recently been achieved with the rise of deep-learning-based generative models. Li et al. (Li et al., 2017) use an encoder-decoder neural network structure, regulated by an adversarial loss function, to recover partly occluded face images. More recently, Yu et al. (Yu et al., 2018) designed an architecture that not only can synthesize missing image parts but also explicitly utilizes surrounding image feature as context to make inpainting more precise. In general, most recent neural-network-based image inpainting algorithms represent some combination of variational autoencoders (VAE) and generative adversarial networks (GAN) and typically contain an encoder, a decoder, and an adversarial discriminator.

2.2.1 Generative Adversarial Networks (GAN)

Generative adversarial networks combine generative and discriminative models to learn very robust image priors (Goodfellow et al., 2014). In a typical formulation, the generator is a transposed convolutional neural network while the discriminator is a regular convolutional neural network. During training, the generator is fed random noise and outputs a generated image. The generated image is sent alongside a real image to the discriminator, which outputs a score to evaluate how real or fake the inputs are. The error between the output score and ground truth score is back-propagated to adjust the weights.

This training scheme forces the generator to produce images that will fool the discriminator into believing they are real images. In the end, training converges at an equilibrium where the generator cannot make the synthesized image more real, while the discriminator fails to tell whether an image is real or generated. Essentially, the training process of GANs forces the generated images to lay within the same distribution (in some latent space) as real images.

2.2.2 Variational autoencoders (VAE)

Autoencoders are deep neural networks, with a narrow bottleneck layer in the middle, that can reconstruct high dimensional data from original inputs. The bottleneck will capture a compressed latent encoding that can then be used for tasks other than reconstruction. Variational autoencoders use a similar encoder-decoder structure but also encourage continuous sampling within the bottleneck layer so that the decoder, once trained, functions as a generator (Kingma & Welling, 2013).

2.2.3 VAE-GAN

While a GAN's generated image outputs are often sharp and clear, the training process can be unstable and prone to problems (Goodfellow et al., 2014; Mao et al., 2016). Even if training problems can be solved, e.g., (Arjovsky et al., 2017), GANs still lack encoders that map real images to latent variables. Compared with GANs, VAE-generated images are often a bit blurrier, but the model structure in general is much more mathematically elegant and more easily trainable. To get the best of both worlds, Larsen et al. (Larsen et al., 2015) proposed an architecture that attaches an adversarial loss to a variational autoencoder, as shown in Figure 3.

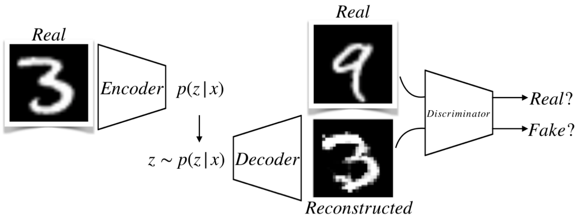

Figure 3. Architecture of VAE-GAN

3. Our Gestalt Reasoning Agent

In this section, we present a general framework for modeling Gestalt visual reasoning on the Raven's test or similar types of problems. Our framework is intended to be agnostic to any type of encoder-decoder-based inpainting model. For our experiments, we adopt a recent VAE-GAN inpainting model (Yu et al., 2018); as we use the identical architecture and training configuration, we refer readers to the original paper for more details about the inpainting model itself.

Our framework makes use of a pre-trained encoder F_θ and corresponding decoder G_ϕ (where θ and ϕ indicate the encoder's and decoder's learned parameters, respectively). The partially visible image to be inpainted, in our case, is a Raven's problem matrix with the fourth cell missing, accompanied with a mask, which is passed as input into the encoder F. Then F outputs an embedded feature representation f, which is sent as input to the generator G. Note that the learned feature representation f could be of any form—a vector, matrix, tensor or any other encoding as long as it represents the latent features of input images.

The generator then outputs a generated image, and we cut out the generated part as the predicted answer. Finally, we choose the most similar candidate answer choice by computing the L_2 distance among feature representations of the various images (the prediction versus each answer choice), computed using the trained encoder F again.

This process is illustrated in Figure 4. More concisely, let x_1, x_2, x_3, be the three elements of the original problem matrix, m be the image mask, and X be the input comprised of these four images. Then, the process of solving the problem to determine the chosen answer y can be written as:

$$y = \operatorname*{argmin}_{k \in \mathbb{S}} \left\| F_\theta\left((G_\phi(F_\theta(X))_{ij})_{\substack{\frac{h}{2}<i\leq h \\ \frac{w}{2}<j\leq w}} \right) - F_\theta(a_k) \right\|$$

where h and w are height and width of the reconstructed image, and $\mathbb{S}$ is the answer choice space.

3.1 Inpainting Models

Our reasoning agent used the same inpainting model architecture (Yu et al., 2018), trained on four different datasets. The first model, which we call Model-Objects, we trained from scratch to evaluate Raven's test performance at multiple checkpoints during training. The latter three models—Model-Faces, Model-Scenes, and Model-Textures—were obtained as pre-trained models (Yu et al., 2018). Details about each dataset are given below.

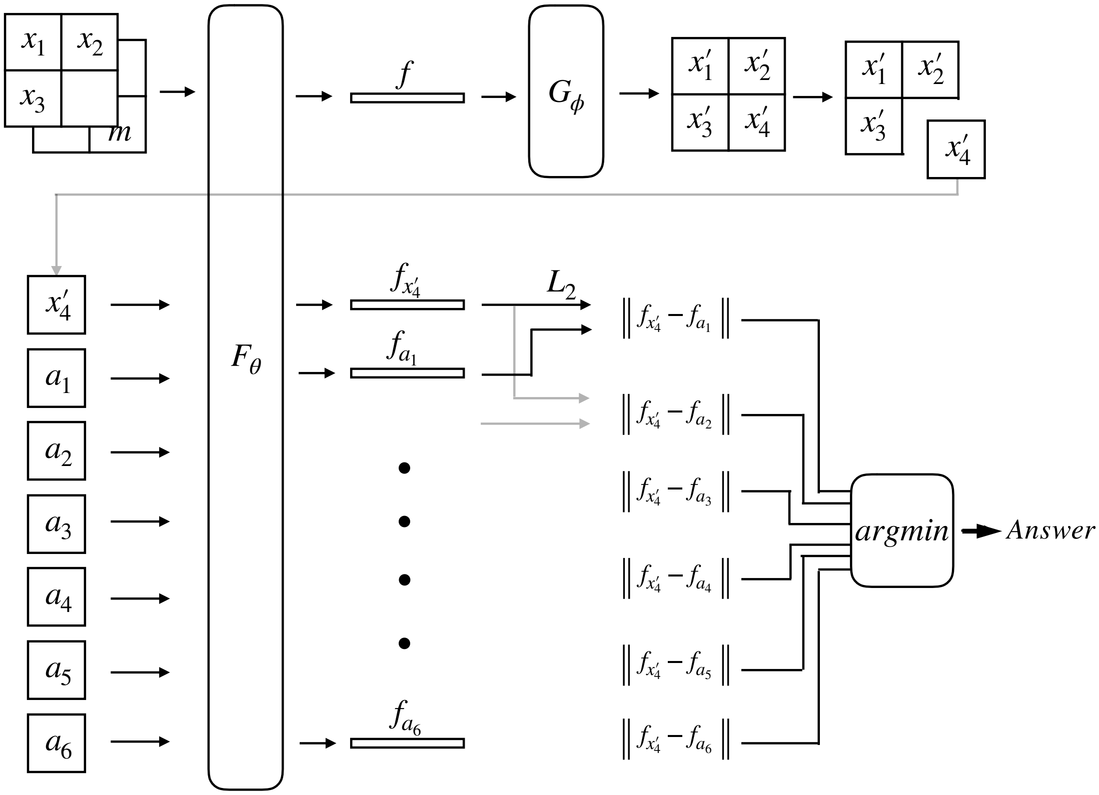
x_1
x_2
x_3
m
f
G_ϕ
x'_1
x'_2
x'_3
x'_4
x'_4
a_1
a_2
a_3
a_4
a_5
a_6
F_θ
$f_{x'_4}$
L_2
f_{a_1}
f_{a_6}
$\| f_{x'_4} - f_{a_1} \|$
$\| f_{x'_4} - f_{a_2} \|$
$\| f_{x'_4} - f_{a_3} \|$
$\| f_{x'_4} - f_{a_4} \|$
$\| f_{x'_4} - f_{a_5} \|$
$\| f_{x'_4} - f_{a_6} \|$
argmin
Answer

Figure 4. Reasoning agent for solving Raven's test problems using Gestalt image completion, using any pre-trained encoder-decoder-based image inpainting model. Elements x_1, x_2, and x_3 from the problem matrix form the initial input, combined into a single image, along with a mask m that indicates the missing portion. These are passed through the encoder F_θ, and the resulting image features f in latent variable space are passed into the decoder G_ϕ. This creates a new complete matrix image X'; the portion x'_4 corresponding to the masked location is the predicted answer to the problem. This predicted answer x'_4, along with all of the answer choices a_i, are again passed through the encoder F_θ to obtain feature representations in latent space, and the answer choice most similar to x'_4 is selected as the final solution.

Note: the reader may wonder why we did not train an inpainting model on Raven's-like images, i.e., black and white illustrations of 2D shapes. Our rationale follows the spirit of human intelligence testing: people are not meant to practice taking Raven's-like problems. If they do, the test is no longer a valid measure of their intelligence (Hayes et al., 2015). Here, our goal was to explore how "test-naive" Gestalt image completion processes would fare.

Model-Objects. The first model was trained from scratch using the full ImageNet (Russakovsky et al., 2015) dataset containing $\sim$14M images non-uniformly spanning 20,000 categories such as "windows," "balloons," and "giraffes." The model converged prior to one full training epoch; we halted training around 300,000 iterations, with a batch size of 36. The best Raven's performance was found at around 80,000 iterations, which means that the final model we used saw only about $\sim$3M images in total during training.

Model-Faces. Our second model was trained on the Large-scale CelebFaces Attributes (CelebA) dataset (Liu et al., 2015), which contains around 200,000 images of celebrity faces, covering around 10,000 individuals.

Model-Scenes. Our third model was trained on the Places dataset (Zhou et al., 2017), which contains around 10M images spanning 434 categories, grouped into three macro-categories: indoor, nature, and urban.

Model-Textures. Our fourth model was trained on the Describable Textures Dataset (DTD) (Cimpoi et al., 2014), which contains 5640 images, divided into 47 categories, of textures taken from real objects, such as knitting patterns, spiderwebs, or an animal's skin.

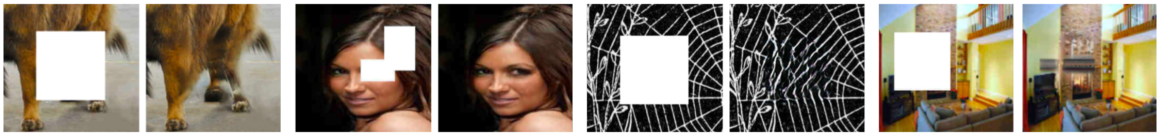

Figure 5. Examples of inpainting produced by same VAE-GAN model (Yu et al., 2018) trained on four different datasets. Left to right: ImageNet (objects), CelebA (faces), DTD (textures), and Places (scenes).

4. Experiment Details and Results

4.1 Raven's Test Materials

Raven's problem images were scanned from paper test booklets. We conducted experiments using two versions: the Standard Progressive Matrices (SPM), intended for the general population, and the Colored Progressive Matrices (CPM), which is an easier test for children and lower-ability adults. In fact, these two tests have substantial overlap: the CPM contains three sets A, AB, and B, with 12 problems each, and the SPM contains five sets A-E also with 12 problems each. Problems increase in difficulty within and across sets.

Initial experiments showed that the inpainting models often failed to work when there was significant white space around the missing element, as in the problem in Figure 1. Thus, when we fed in the matrix images as a combined single image, as in Figure 4, we cropped out this white space. This did change the appearance of problems somewhat, essentially squeezing together the elements in the matrix.

4.2 Model-Objects Results Across Training

Figure 6 shows test results as a function of training for Model-Objects. The top plot shows the loss function being trained on for image inpainting; the model seems to settle into a minimum around 200,000 iterations. The bottom plot shows CPM performance as a function of training, divided into sets A, AB, and B. The model relatively quickly rises above chance performance (1/6 random guessing = total score of 6).

Interestingly, we noticed that the randomly initialized model actually appears to do a bit better than chance; after numerous runs, the average starting score was around 8/36. We believe this can be attributed to intrinsic structure-capturing abilities of the convolutional neural network structure (Ulyanov et al., 2018).

After $\sim$80,000 iterations, CPM performance does not change other than local variations. For the rest of our analyses, we used the model snapshot at the point when it reached peak performance of 27/36 correct. Note that this does yield perhaps a more optimistic estimate of performance than would be obtained by an "average" trained model. Future experiments might be improved by defining stopping criteria based on some function of inpainting loss during training instead.

Figure 7 shows examples of Model-Objects results from various sample problems. (Actual Raven's problems are not shown, in order to protect test security.) Some results are surprisingly good, given that the model was only trained on real-world color photographs.

Interestingly, when we inspected results from the Raven's test, the model generates what look like poor image guesses for certain problems, for example on some of the more difficult problems in set E, but then still chooses the correct answer choice. This could be some form of lucky informed guessing, or, it could be that the image representations in latent space are actually capturing some salient features of the problem and solution.

4.3 Results Across Inpainting Models

Next we compared results across four versions of our reasoning agent: the one using Model-Objects trained as above, and the others using pre-trained versions of Model-Faces, Model-Scenes, and Model-Textures. Figure 8 (top) shows scores achieved by each of the four models on each of the six sets of Raven's problems. As seen in this plot, Model-Objects performs better than any of the other models overall, though Model-Textures does a smidgeon better on Set A (which contains very texture-like problems, so this result makes sense).

None of the models do very well on sets C or D, performing essentially at chance (these problems have 8 answer choices, so chance $\sim$1.5 correct per set). Interestingly, Model-Objects was the only one that consistently generated answers to all problems; the other three models often generated blank images to problems in sets C and D. We are not sure why this occurred. All of the models do rather surprisingly well on set E, which is supposed to be the hardest set of all.

Figure 8 (bottom) shows values called "score discrepancies." When a person takes a Raven's test, the examiner is supposed to check the per-set composition of their score against normative data from other people who got the same total score. So, for example, a score of 27 on the CPM has norms of 10, 10, and 7 for sets A, AB, and B, respectively, which is exactly what Model-Objects scored. This means that Model-Objects was essentially subject to the same difficulty distribution as other people taking the test.

In contrast, if we look at the SPM results, the models do worse than they should have on sets C and D, and better than they should have on set E, in comparison to people getting the same total scores. This means that the difficulty distributions experienced by the models are not the same as what people typically experience.

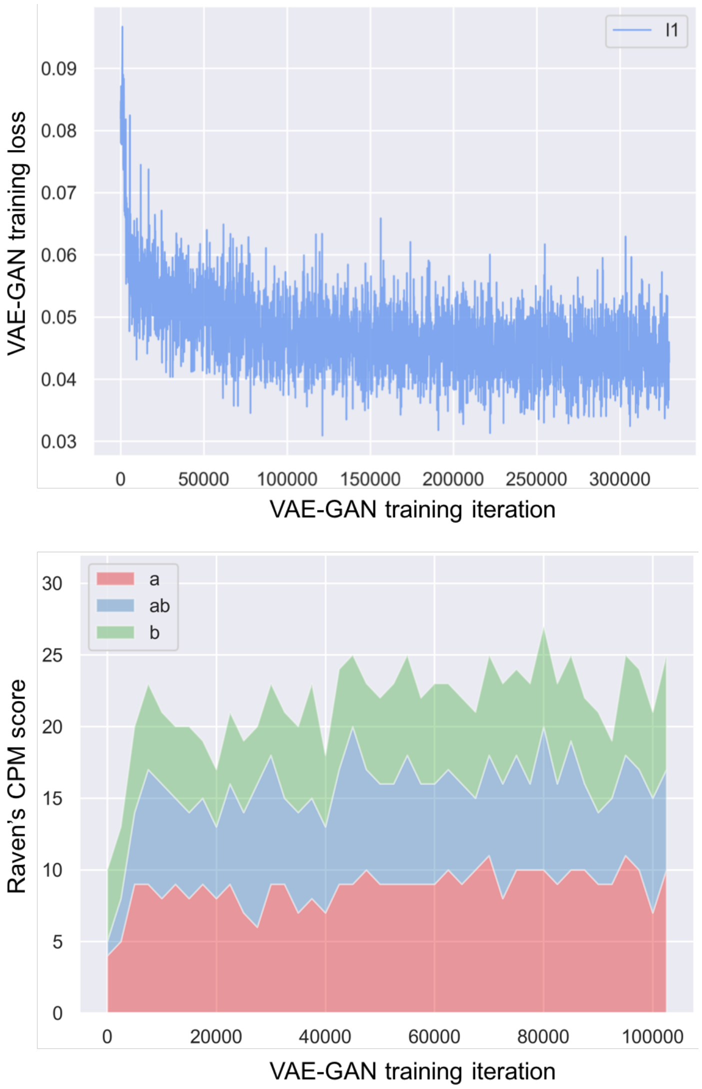

Figure 6. Image inpainting loss (top) and CPM performance (bottom) during training of Model-Objects.

5. Discussion and Future Work

Over the decades, there have been many exciting efforts in AI to computationally model various aspects of problem solving for Raven's matrix reasoning and similar geometric analogy problems, beginning with Evans' classic ANALOGY program (Evans, 1968). In this section, we review some major themes that seem to have emerged across these efforts, situate our current work within this broader context, and point out important gaps that remain unfilled.

Note that our discussion does not focus heavily on absolute test scores. Raven's is not now (and probably never will be) a task that is of practical utility for AI systems in the world to be

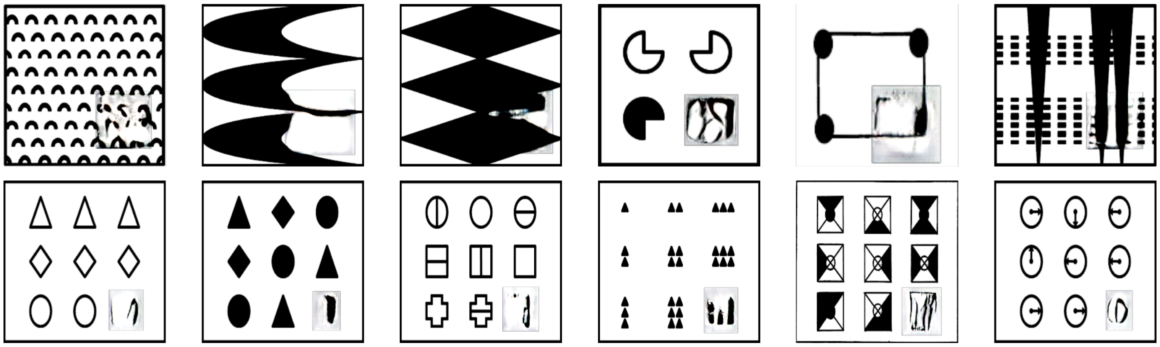

Figure 7. Images generated using Model-Objects for a variety of Raven's-like sample problems.

solving well, and so treating it as a black-box benchmark may be of limited value. However, the test has been and continues to be enormously profitable as a research tool for generating insights into the organization of intelligence, both in humans and in artificial agents. We feel that much valuable scientific knowledge from computational studies of Raven's problem solving has come from systematic, within-model experiments, which is also our aim here.

Knowledge-based versus data-driven. Early models took a knowledge-based approach, meaning that they contained explicit, structured representations of certain key elements of domain knowledge. For example, Carpenter and colleagues (1990) built a system that matched relationships among problem elements according to one of five predefined rules. Knowledge-based models tend to focus on what an agent does with its knowledge during reasoning (Rasmussen & Eliasmith, 2011; Kunda et al., 2013; Strannegård et al., 2013; Lovett & Forbus, 2017); where this knowledge might come from remains an open question.

On the flip side, a recently emerging crop of data-driven models extract domain knowledge from a training set containing example problems that are similar to the test problems the model will eventually solve (Hoshen & Werman, 2017; Barrett et al., 2018; Hill et al., 2019; Steenbrugge et al., 2018; van Steenkiste et al., 2019; Zhang et al., 2019). Data-driven models tend to focus on interactions between training data, learning architectures, and learning outcomes; how knowledge might be represented in a task-general manner and used flexibly during reasoning and decision-making remain open questions.

Our model of Gestalt visual reasoning falls into an interesting grey area between these two camps. On the one hand, the model represents Gestalt principles implicitly, as image priors in some latent space, and these priors are learned in a data-driven fashion. On the other hand, unlike all of the above data-driven models, our model does *not* train on anything resembling Raven's problems. In that sense, it is closer to a knowledge-based model in that we can investigate how knowledge learned in one setting (image inpainting) can be applied to reason about very different inputs.

Constructive matching versus response elimination. Another interesting divide among Raven's models has to do with the overall problem-solving strategy. A study of human problem solving on geometric analogy problems found that people generally use one of two strategies: they come up with a predicted answer first, and then compare it to the answer choices—constructive matching—or

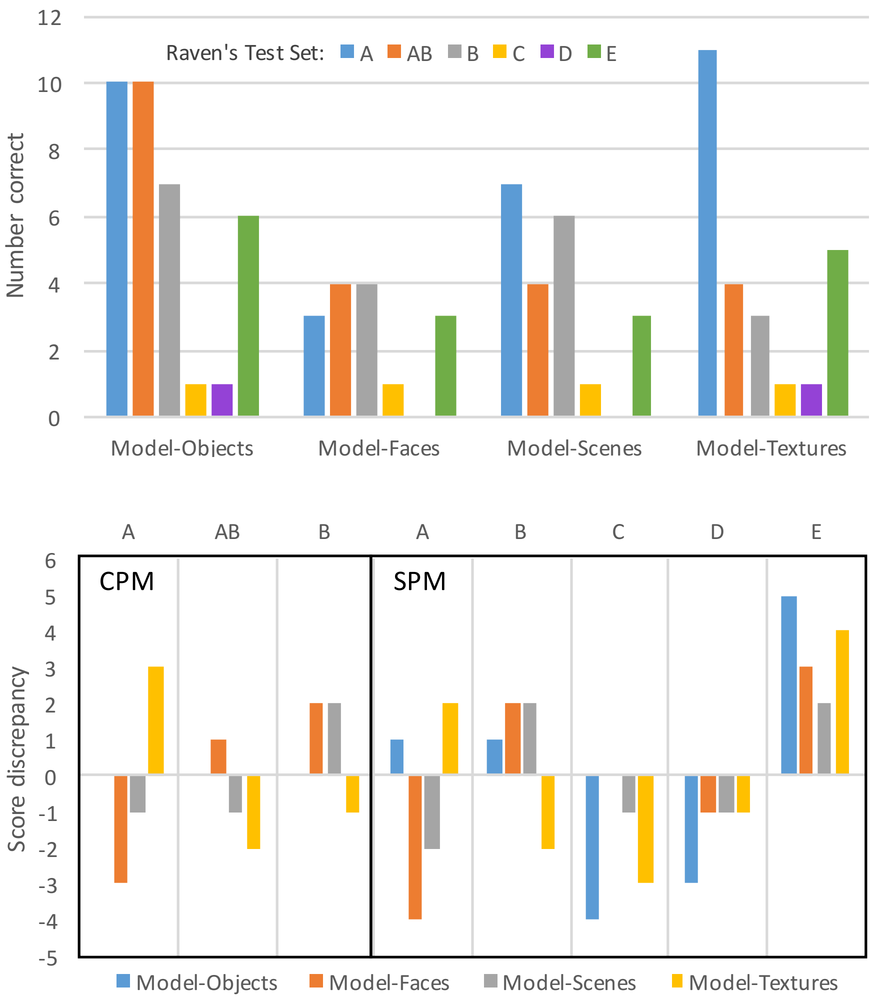

Figure 8. Top: Results for each agent on each set of the Raven's CPM (A, AB, and B) and SPM (A-E). Bottom: Per-set score discrepancies between each model and human norms for same total scores.

they mentally plug each answer choice into the matrix and choose the best one—response elimination (Bethell-Fox et al., 1984). Knowledge-based models have come in both varieties; all of the data-driven models follow the response-elimination approach. Our model uses constructive matching, which we feel is an interesting capability given that the system is not doing any deliberative reasoning (per se) about what should go in the blank space.

Open issues. Our Gestalt model certainly has limitations, as illustrated in the results section. (See Figure 9 for another example.) However, our investigations highlight a form of human reasoning that has not been explored in previous Raven's models. How are Gestalt principles learned, and how do specific types of visual experiences contribute to a person's sensitivity to regularities like symmetry or closure? One fascinating direction for future work will be to explore these relationships in more detail, and perhaps shed light on cultural factors in intelligence testing. For example, would a model trained only on urban scenes (which contain lots of corners, perfect symmetries, and straight lines) do better on Raven's problems than a model trained only on nature scenes?

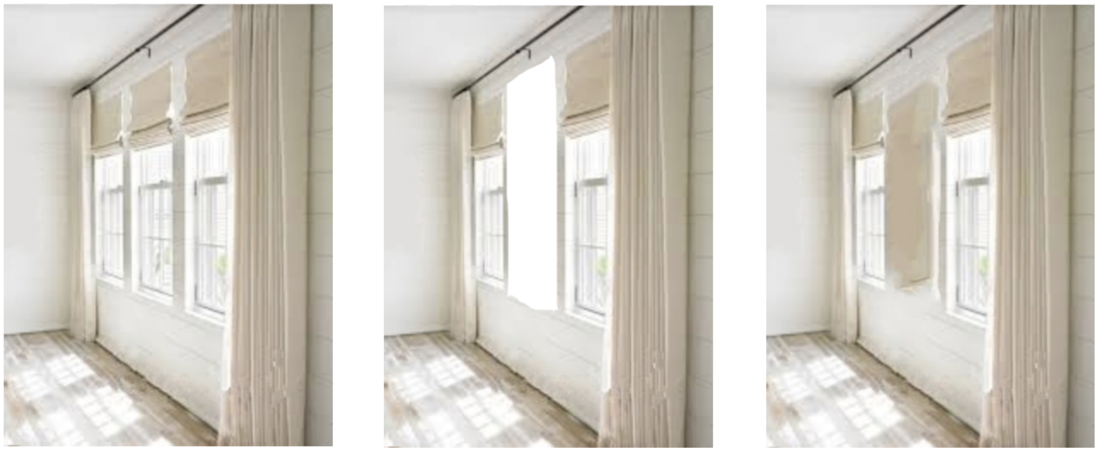

Figure 9. Model-Objects performing inpainting on a row of windows, with the original image on the left, the masked image in the center, and the inpainted image on the right. Note the phantom reflection in the inpainted image. This type of relational, commonsense reasoning requires going beyond a purely Gestalt approach.

Finally, two major open issues for AI models of intelligence tests in general are: metacognitive strategy selection, and task learning. Most AI models tend to adopt a single strategy and see how far its performance can be pushed. However, for humans, a major part of the challenge of intelligence testing is figuring out what strategy to use when, and being able to adapt and switch strategies as needed. In the context of our work, we aim to integrate our Gestalt approach with other, more deliberative reasoning approaches to begin to address this issue. This will introduce many challenges related to having to determine confidence in an answer, planning and decision making, etc.

Relatedly, as with many tasks and systems in AI, previous work on Raven's and other intelligence tests has required the AI system designers to specify the task, its format, goal, etc. for the system. Humans sit down and are given verbal or demonstration-based instructions, and must learn the task, how to represent it internally, and how and what procedures to try. In other words, what is it about a Raven's problem image that leads people (at least some people, some of the time, on some problems) to gaze at the "blank" space in the matrix so that their mind visually fills in the gap? This and many other challenges of task learning (Laird et al., 2017) remains open questions for AI research in intelligence testing (Hernández-Orallo et al., 2016).

Acknowledgements

Many thanks to the anonymous reviewers for their helpful comments. This work was supported in part by National Science Foundation Award #1730044.

References

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

Barnes, C., Shechtman, E., Finkelstein, A., & Goldman, D. B. (2009). Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM TOG* (p. 24). ACM.

Barrett, D. G., Hill, F., Santoro, A., Morcos, A. S., & Lillicrap, T. (2018). Measuring abstract reasoning in neural networks. *arXiv preprint arXiv:1807.04225*.

Bertalmio, M., Sapiro, G., Caselles, V., & Ballester, C. (2000). Image inpainting. *SIGGRAPH* (pp. 417–424).

Bethell-Fox, C., Lohman, D., & Snow, R. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence*, *8*, 205–238.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures. *Psychological review*, *97*, 404.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., & Vedaldi, A. (2014). Describing textures in the wild. *CVPR* (pp. 3606–3613).

DeShon, R. P., Chan, D., & Weissbein, D. A. (1995). Verbal overshadowing effects on raven's advanced progressive matrices: Evidence for multidimensional performance determinants. *Intelligence*, *21*, 135–155.

Desolneux, A., Moisan, L., & Morel, J.-M. (2007). *From gestalt theory to image analysis: a probabilistic approach*, volume 34. Springer Science & Business Media.

Evans, T. G. (1968). A program for the solution of a class of geometric-analogy intelligence-test questions. *Semantic Information Processing*, (pp. 271–353).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *NIPS/NeurIPS* (pp. 2672–2680).

Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2015). Do we really become smarter when our fluid-intelligence test scores improve? *Intelligence*, *48*, 1–14.

Hays, J., & Efros, A. A. (2008). Scene completion using millions of photographs. *CACM*, *51*, 87–94.

Hernández-Orallo, J., Martínez-Plumed, F., Schmid, U., Siebers, M., & Dowe, D. L. (2016). Computer models solving intelligence test problems: Progress and implications. *Artificial Intelligence*, *230*, 74–107.

Hill, F., Santoro, A., Barrett, D. G., Morcos, A. S., & Lillicrap, T. (2019). Learning to make analogies by contrasting abstract relational structure. *arXiv preprint arXiv:1902.00120*.

Hoshen, D., & Werman, M. (2017). Iq of neural networks. *arXiv preprint arXiv:1710.01692*.

Hunt, E. (1974). Quote the raven? nevermore! In L. Gregg (Ed.), *Knowledge and cognition*, 129–158. Erlbaum.

Kanizsa, G. (1979). *Organization in vision: Essays on gestalt perception*. Praeger Publishers.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kirby, J. R., & Lawson, M. J. (1983). Effects of strategy training on progressive matrices performance. *Contemporary Educational Psychology*, *8*, 127–140.

Kunda, M., & Goel, A. K. (2011). Thinking in pictures as a cognitive account of autism. *Journal of autism and developmental disorders*, *41*, 1157–1177.

Kunda, M., McGreggor, K., & Goel, A. K. (2013). A computational model for solving problems from the rpm intelligence test using iconic visual representations. *Cognitive Systems Research*, *22*, 47–66.

Laird, J. E., et al. (2017). Interactive task learning. *IEEE Intelligent Systems*, *32*, 6–21.

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O. (2015). Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*.

Li, Y., Liu, S., Yang, J., & Yang, M.-H. (2017). Generative face completion. *CVPR* (pp. 3911–3919).

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. *ICCV* (pp. 3730–3738).

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

Lynn, R., Allik, J., & Irwing, P. (2004). Sex differences on three factors identified in raven's standard progressive matrices. *Intelligence*, *32*, 411–424.

Mao, X., Li, Q., Xie, H., Lau, R. Y. K., Wang, Z., & Smolley, S. P. (2016). Least squares generative adversarial networks. *ICCV*, (pp. 2813–2821).

Prabhakaran, V., Smith, J. A., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. (1997). Neural substrates of fluid reasoning: an fmri study of neocortical activation during performance of the rpm test. *Cognitive psychology*, *33*, 43–63.

Rasmussen, D., & Eliasmith, C. (2011). A neural model of rule generation in inductive reasoning. *Topics in Cognitive Science*, *3*, 140–153.

Roid, G. H., & Miller, L. J. (1997). Leiter international performance scale-revised (leiter-r). *Wood Dale, IL: Stoelting*.

Russakovsky, O., et al. (2015). Imagenet large scale visual recognition challenge. *Int. journal of computer vision*, *115*, 211–252.

Schönlieb, C.-B. (2015). *Partial differential equation methods for image inpainting*. Cambridge University Press.

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. *Advances in the psychology of human intelligence*, *2*, 103.

Steenbrugge, X., Leroux, S., Verbelen, T., & Dhoedt, B. (2018). Improving generalization for abstract reasoning tasks using disentangled feature representations. *arXiv preprint arXiv:1811.04784*.

van Steenkiste, S., Locatello, F., Schmidhuber, J., & Bachem, O. (2019). Are disentangled representations helpful for abstract visual reasoning? *arXiv preprint arXiv:1905.12506*.

Strannegård, C., Cirillo, S., & Ström, V. (2013). An anthropomorphic method for progressive matrix problems. *Cognitive Systems Research*, *22*, 35–46.

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2018). Deep image prior. *CVPR* (pp. 9446–9454).

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological bulletin*, *138*, 1172.

Wechsler, D. (2008). Wechsler adult intelligence scale–fourth edition (wais–iv). *San Antonio, TX: NCS Pearson*, *22*, 498.

Wertheimer, M. (1923). Untersuchungen zur lehre von der gestalt. ii. *Psychological Research*, *4*, 301–350.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2018). Generative image inpainting with contextual attention. *CVPR* (pp. 5505–5514).

Zhang, C., Gao, F., Jia, B., Zhu, Y., & Zhu, S.-C. (2019). Raven: A dataset for relational and analogical visual reasoning. *CVPR* (pp. 5317–5327).

Zheng, C., Cham, T.-J., & Cai, J. (2019). Pluralistic image completion. *CVPR* (pp. 1438–1447).

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE TPAMI*, *40*, 1452–1464.