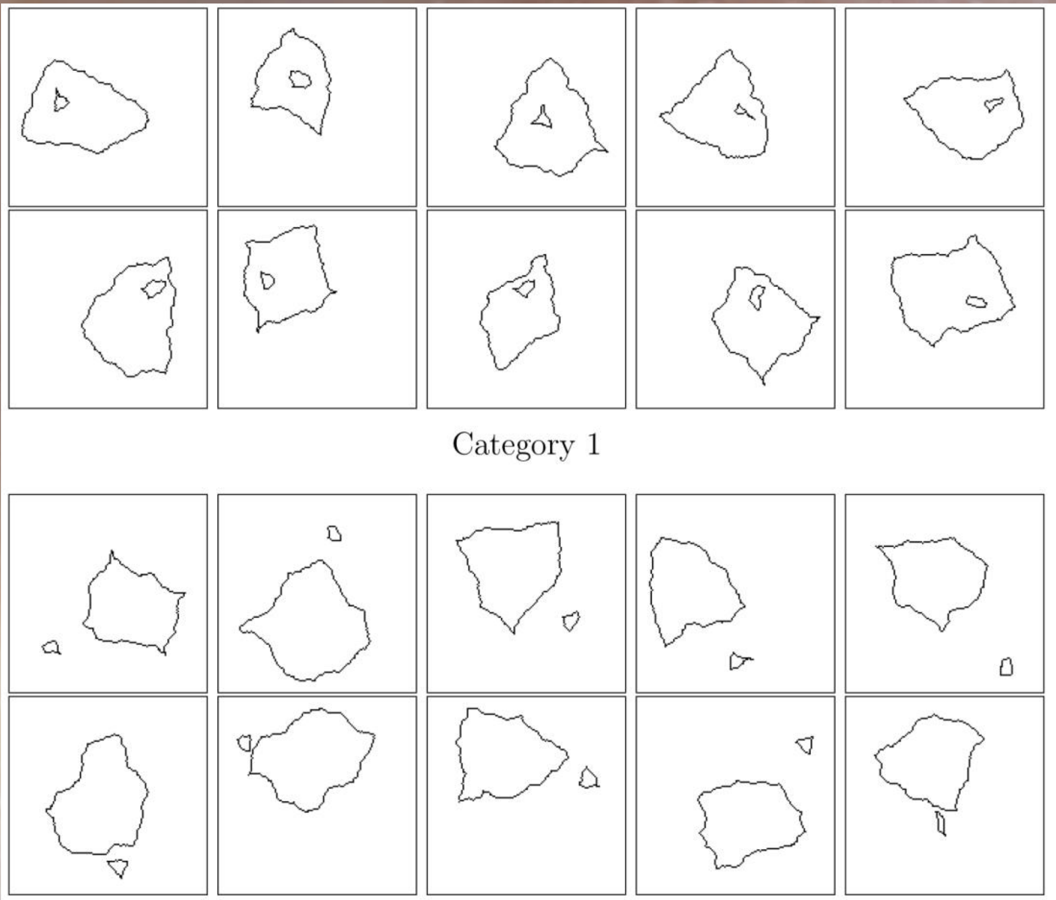
Visual Object Recognition
Computational Models and Neurophysiological Mechanisms
Neuro 130/230. Harvard College/GSAS 78454



Category 1

Category 2

What is different between category 1 and category 2 images?

Visual Object Recognition
Computational Models and Neurophysiological Mechanisms
Neuro 130/230. Harvard College/GSAS 78454

Class 1 [09/11/2023]. Introduction to Vision
Class 2 [09/18/2023]. The Phenomenology of Vision
Class 3 [09/25/2023]. Natural image statistics and the retina
Class 4 [10/02/2023]. Learning from Lesions
Class 5 [10/16/2023]. Primary Visual Cortex
Class 6 [10/23/2023]. Adventures into *terra incognita*
Class 7 [10/30/2023]. From the Highest Echelons of Visual Processing to Cognition
Class 8 [11/06/2023]. First Steps into in silico vision
Class 9 [11/13/2023]. Teaching Computers how to see
Class 10 [11/20/2023]. Computer Vision
Class 11 [11/27/2023]. Connecting Vision to the rest of Cognition
Class 12 [12/04/2023]. Visual Consciousness

FINAL EXAM, PAPER DUE 12/14/2023. No extensions.

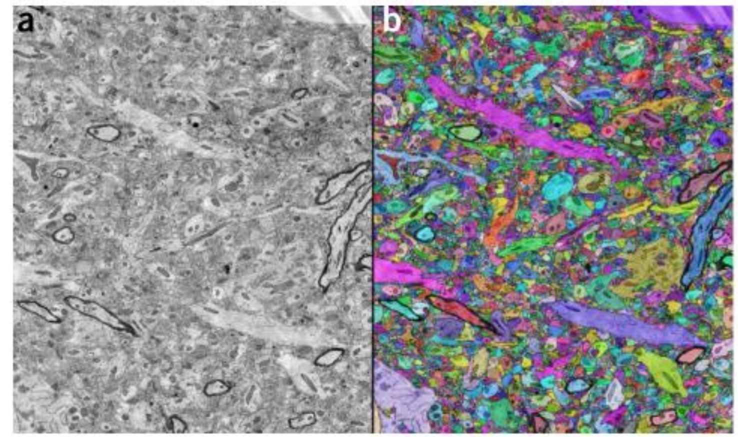
Previously on computer vision…

Species classification and detection

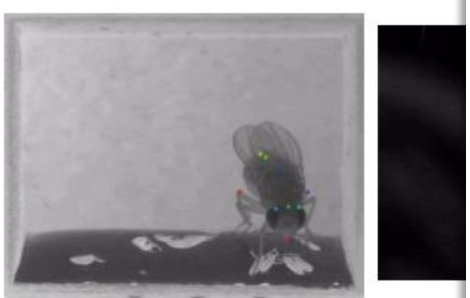
pro

ation+L

Computer vision can help segment biological images



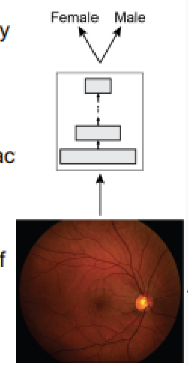
Automatic pose estimation for ethol



Mathis et al 2018

Applications of compute clinical diagnosis

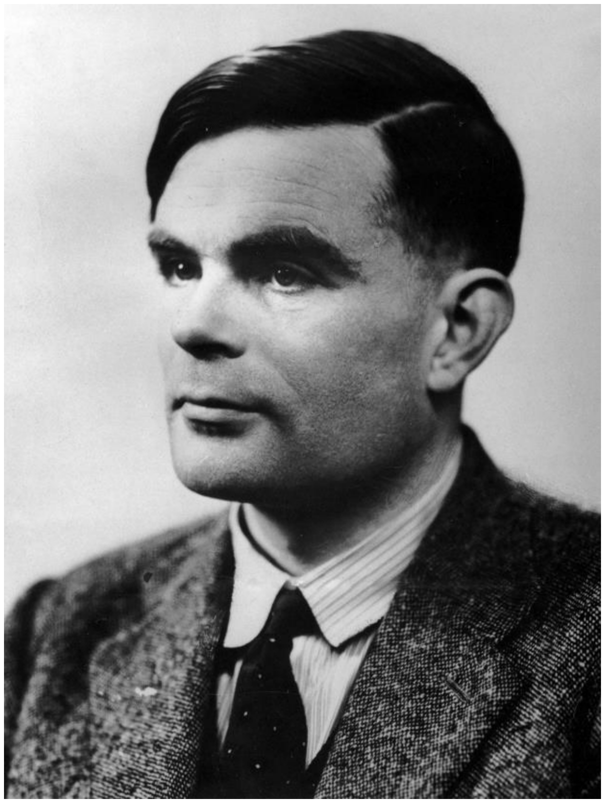
- Excellent performance in many clinical diagnosis tasks
  E.g. breast tumor detection
  E.g. diabetic retinopathy
- Reliability, consistency, accurac
- Machines can discover properties in the data that humans never even thought of before
  E.g. cardiovascular disease risk from fundus photographs
- Beware of incidental findings
- Beware of biases in training data

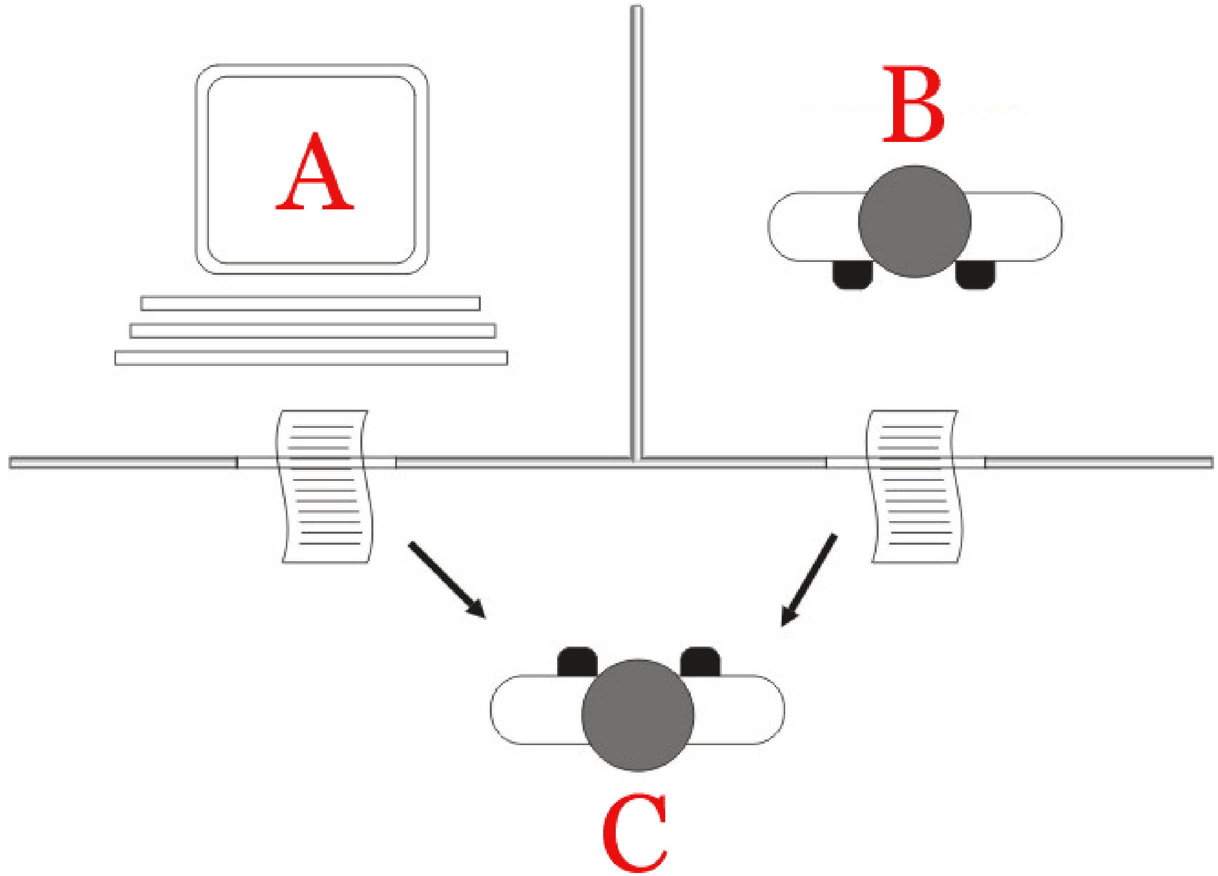


The portrait of Edmond de Belamy

Sold at Christie's auction: $432,500

The Turing test

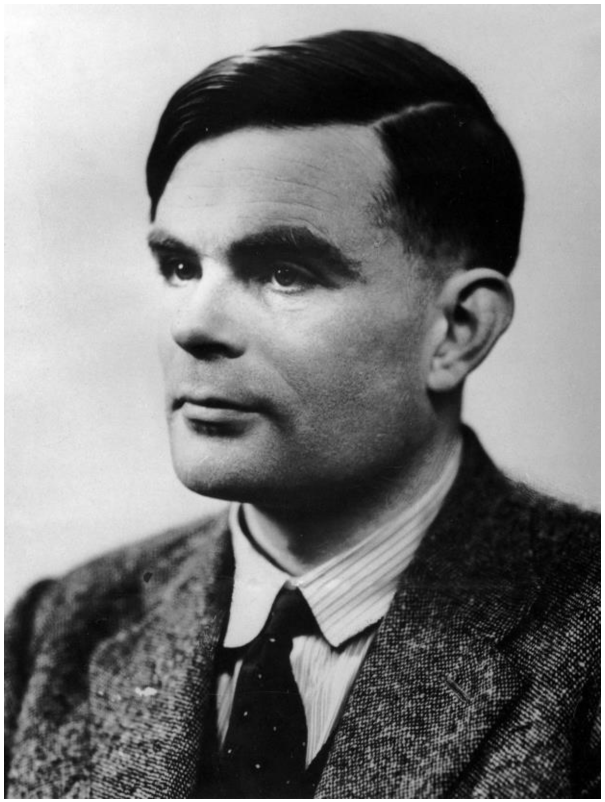


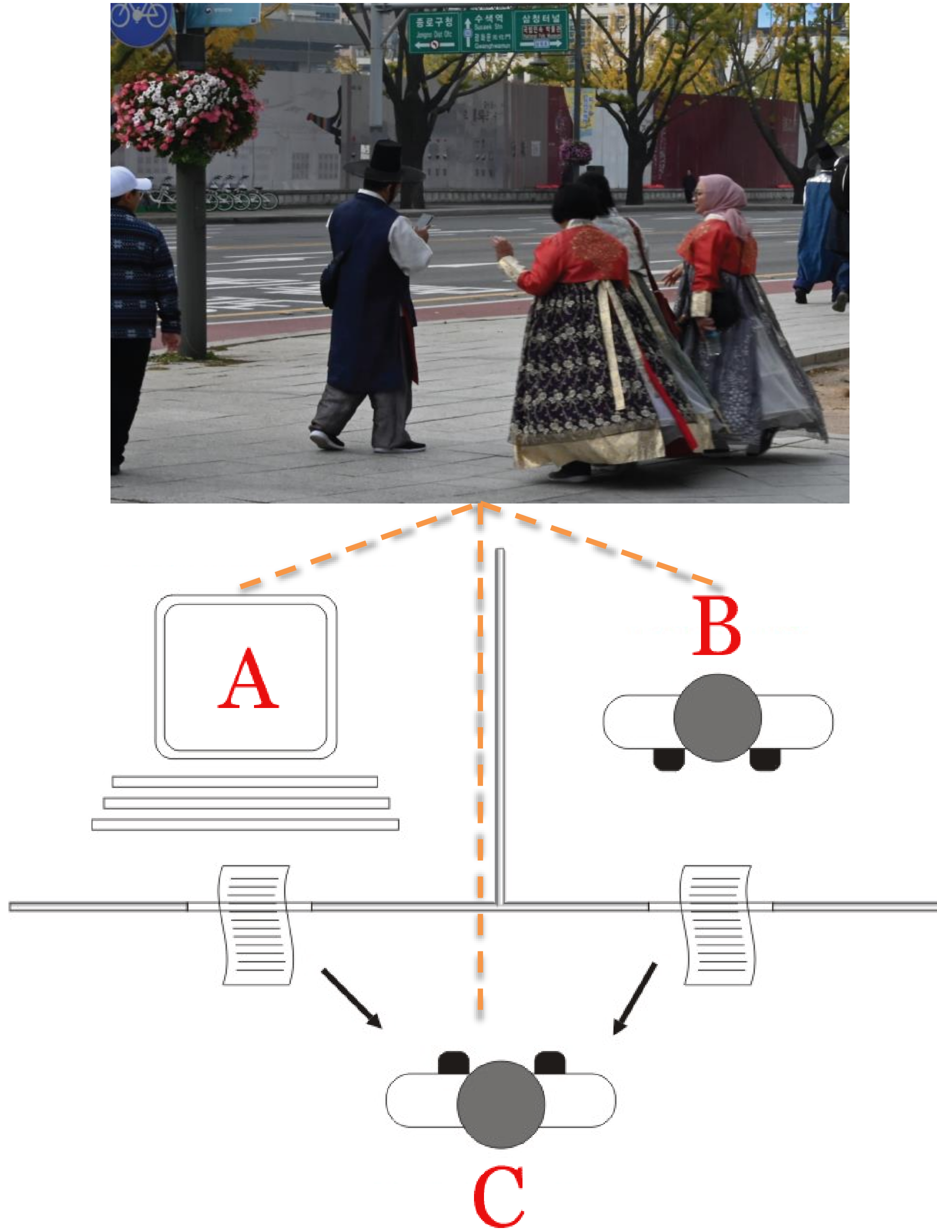


Turing, 1950

Image credit: Encyclopædia Britannica
Juan Alberto Sánchez Margallo, Wikimedia

The Turing test for vision

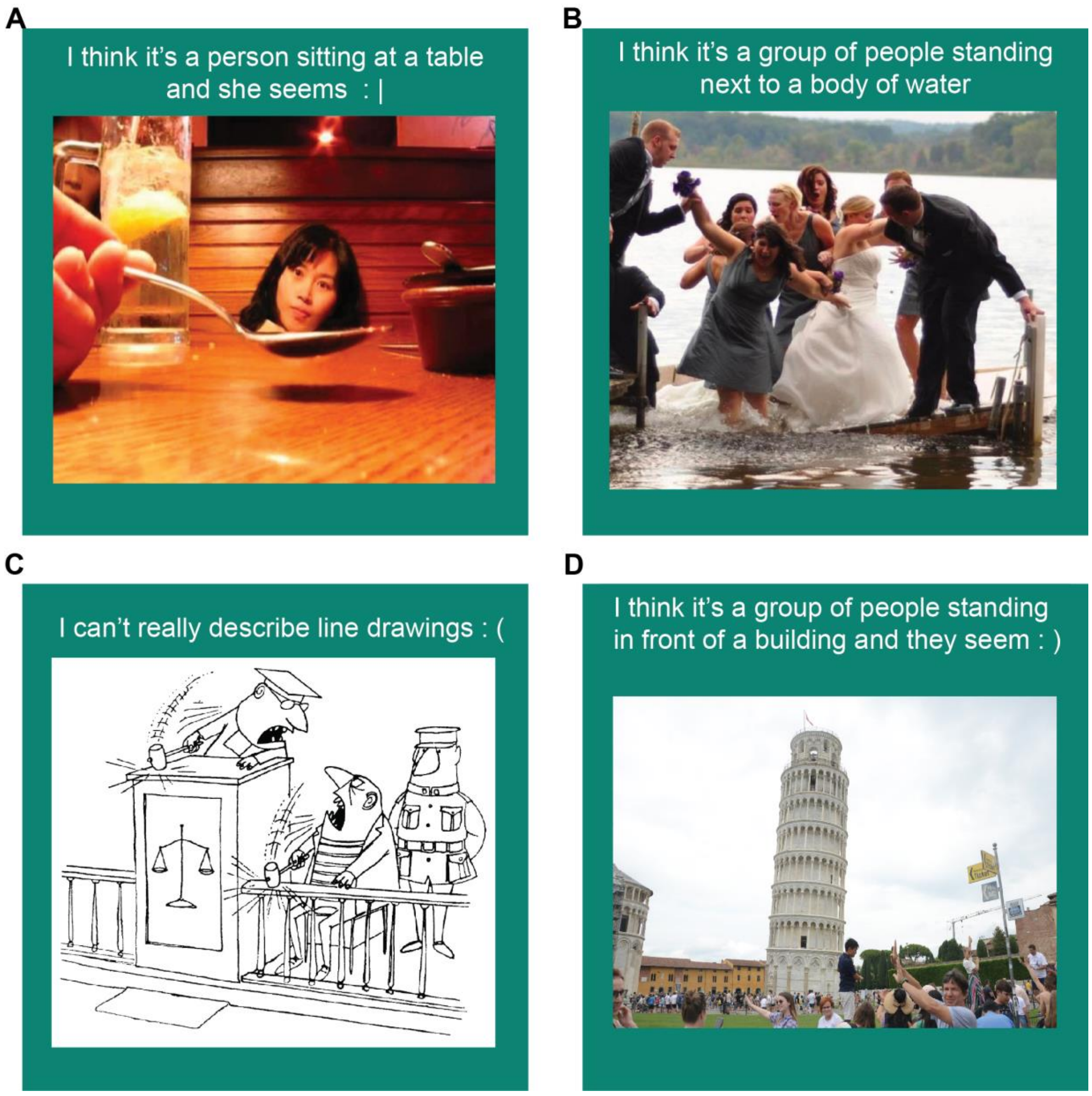




Turing, 1950

Vision-language tasks

Image captioning

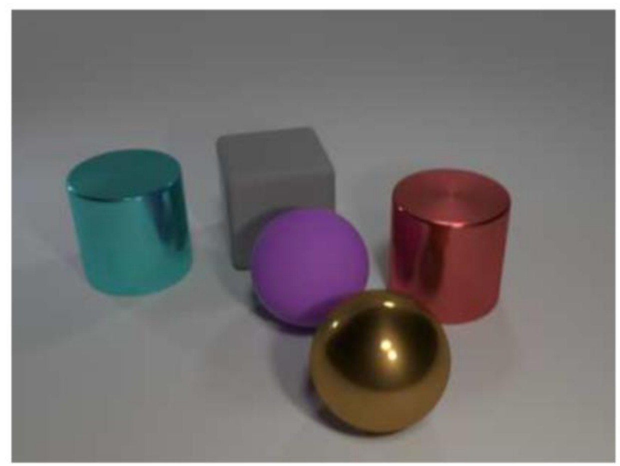


Visual question answering (VQA)

How many chairs are at the table?

Is there a pedestrian in my lane?

Is the person with the blue hat touching the bike in the back?



Is there a matte cube that has the same size as the red metal object?

Johnson et al, ICCV 2017

GPT-4 Vision

Visual inputs: VGA charger

Sample 1 of 7

Next sample

User What is funny about this image? Describe it panel by panel.

Source: hmmm (Reddit)

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

GPT-4 Vision

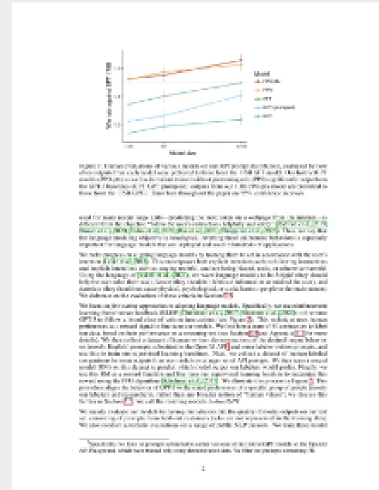
Visual inputs: pixels to paper summaries

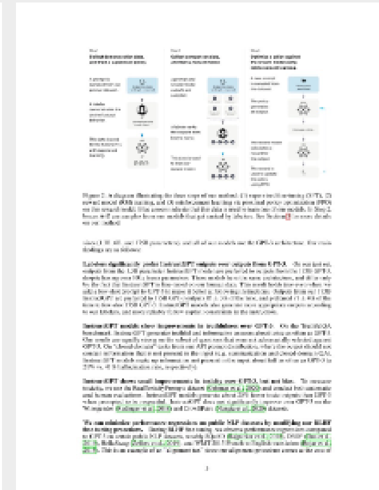
Sample 5 of 7

Next sample

User Below is part of the InstuctGPT paper. Could you read and summarize it to me?


Training language models to follow instructions with human feedback





GPT-4 The InstructGPT paper focuses on training large language models to follow instructions with human feedback. The authors note that making language models larger doesn't inherently make them better at following a user's intent. Large models can generate outputs that are untruthful, toxic, or simply unhelpful.

To address this issue, the authors fine-tune language models on a wide range of tasks using human feedback. They start with a set of labeler-written prompts and responses, then collect a dataset of labeler demonstrations of the desired model behavior. They fine-tune GPT-3 using supervised learning and then use reinforcement learning from human feedback to further fine-tune the model. The resulting model, called InstructGPT, shows improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets.

The authors conclude that fine-tuning with human feedback is a promising direction for aligning language models with human intent.

User Could you explain the process described in Figure 2?

GPT-4 Figure 2 in the InstructGPT paper illustrates the three steps of their method:

1. Supervised fine-tuning (SFT): They collect demonstration data, which consists of labeler-

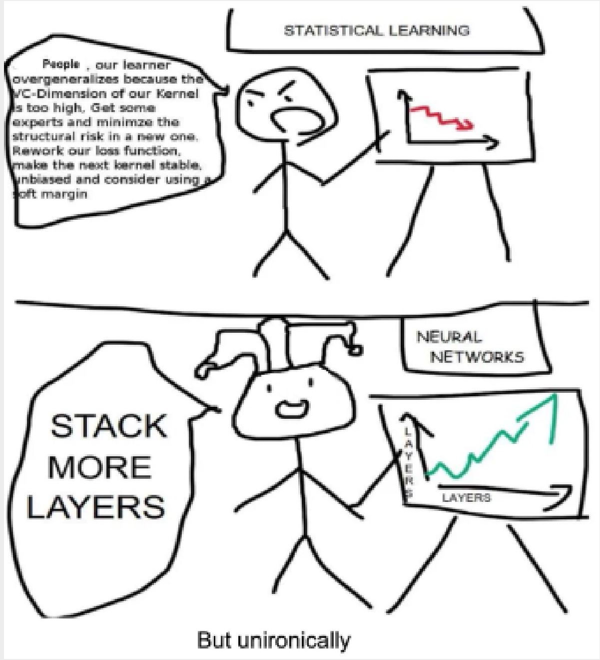
GPT-4 Vision

Visual inputs: moar layers

Sample 7 of 7

Next sample

User Can you explain why this is funny. Think about it step-by-step.



GPT-4 The comic is satirizing the difference in approaches to improving model performance between statistical learning and neural networks.

In statistical learning, the character is shown to be concerned with overfitting and suggests a series of complex and technical solutions, such as minimizing structural risk, reworking the loss function, and using a soft margin.

In contrast, the neural networks character simply suggests adding more layers to the model. This is often seen as a common solution to improving performance in neural networks, but it's also considered a simplistic and brute-force approach.

The humor comes from the contrast between the complexity and specificity of the statistical learning approach and the simplicity and generality of the neural network approach. The "But unironically" comment adds to the humor by implying that, despite being simplistic, the "stack more layers" approach is often effective in practice.

LLaVA: Large Language and Vision Assistant

Visual Instruction Tuning

NeurIPS 2023 (Oral)

Haotian Liu*, Chunyuan Li*, Qingyang Wu, Yong Jae Lee

▶ University of Wisconsin-Madison ▶ Microsoft Research ▶ Columbia University

*Equal Contribution

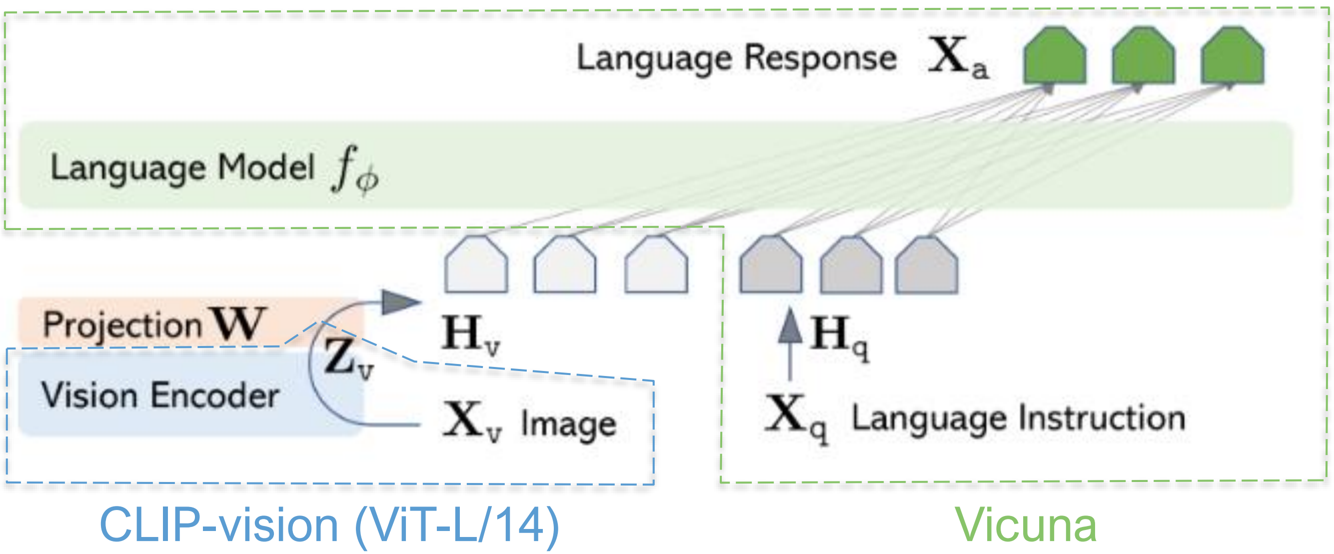
arXiv | arXiv (LLaVA-1.5) | Code | Demo | Dataset | Model

🔥[NEW!] LLaVA-1.5 achieves SoTA on 11 benchmarks, with just simple modifications to the original LLaVA, utilizes all public data, completes training in ~1 day on a single 8-A100 node, and surpasses methods that use billion-scale data.

LLaVA represents a novel end-to-end trained large multimodal model that combines a vision encoder and Vicuna for general-purpose visual and language understanding, achieving impressive chat capabilities mimicking spirits of the multimodal GPT-4 and setting a new state-of-the-art accuracy on Science QA.

LLaVA architecture

LLaVa connects pre-trained CLIP ViT-L/14 visual encoder and large language model Vicuna, using a simple projection matrix.

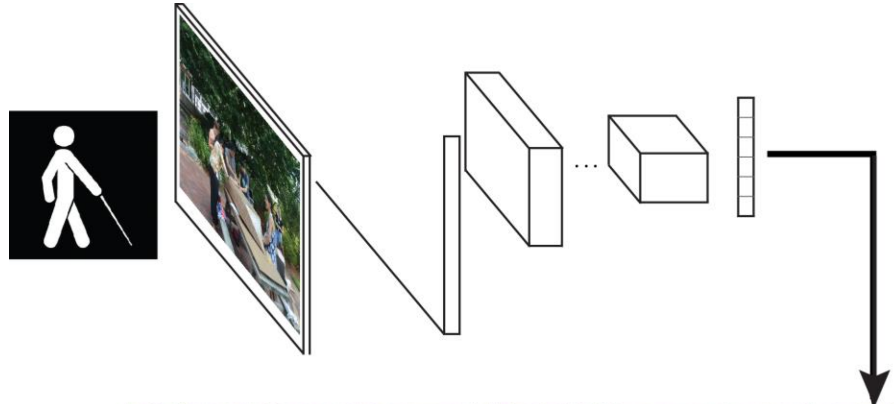


What would LLaVA say?

shorturl.at/esyD6

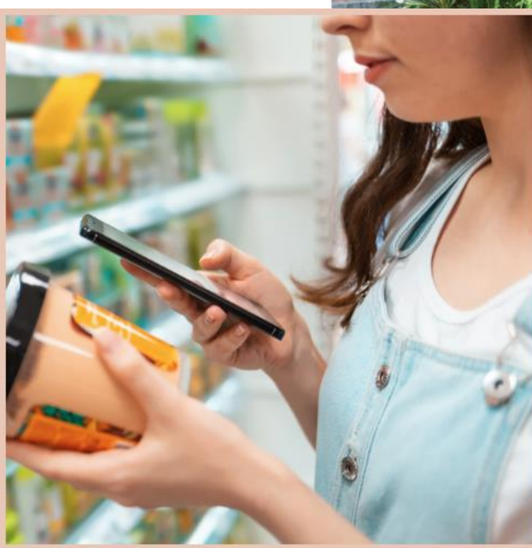
llava.hliu.cc

Computer vision to help people-in-need see

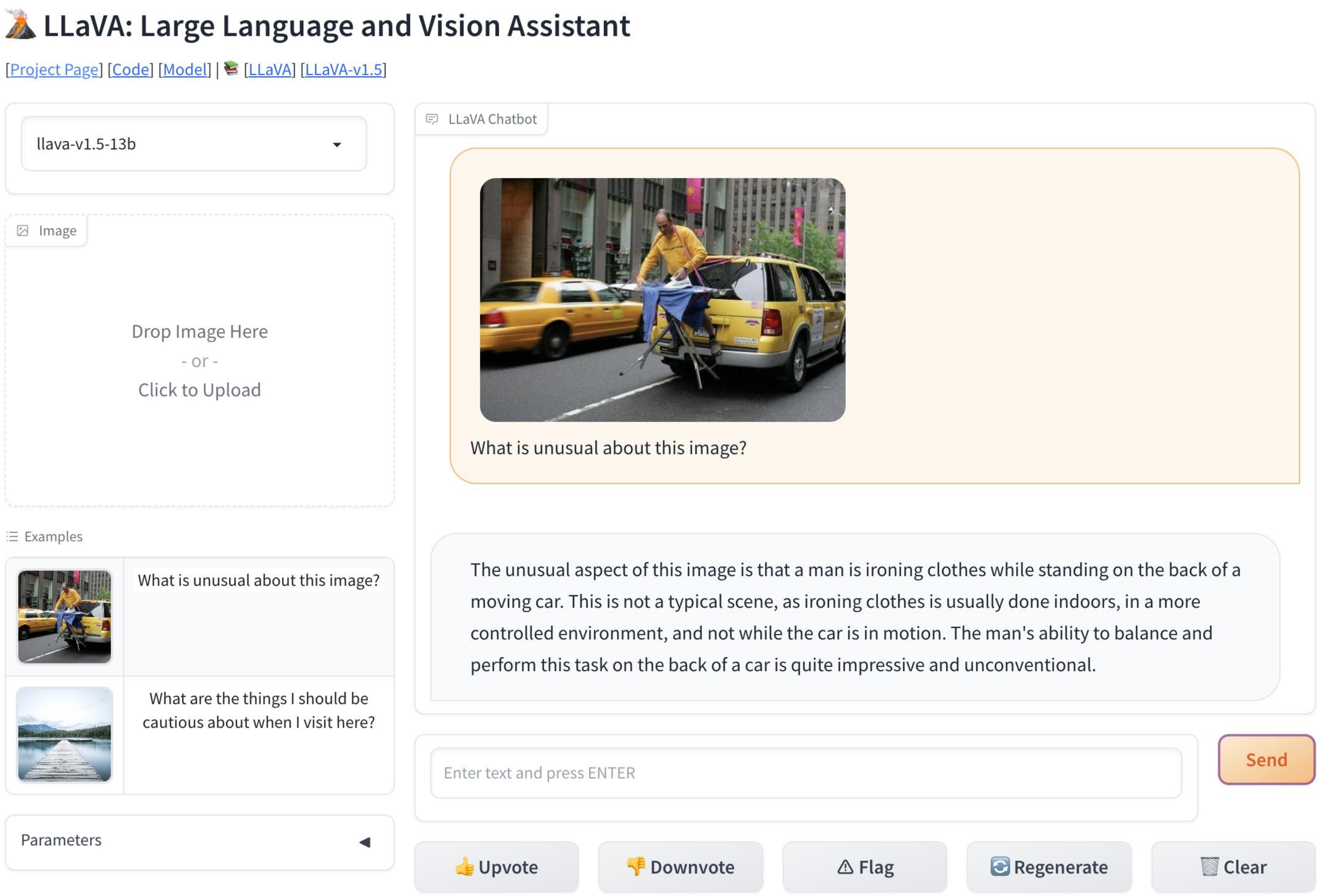


Be My Eyes

Be My Eyes uses GPT-4 to transform visual accessibility.



Can computer vision really reason?



"What is different between categories 1 and 2?"

Hard tests for visual cognition using simplified, controlled datasets

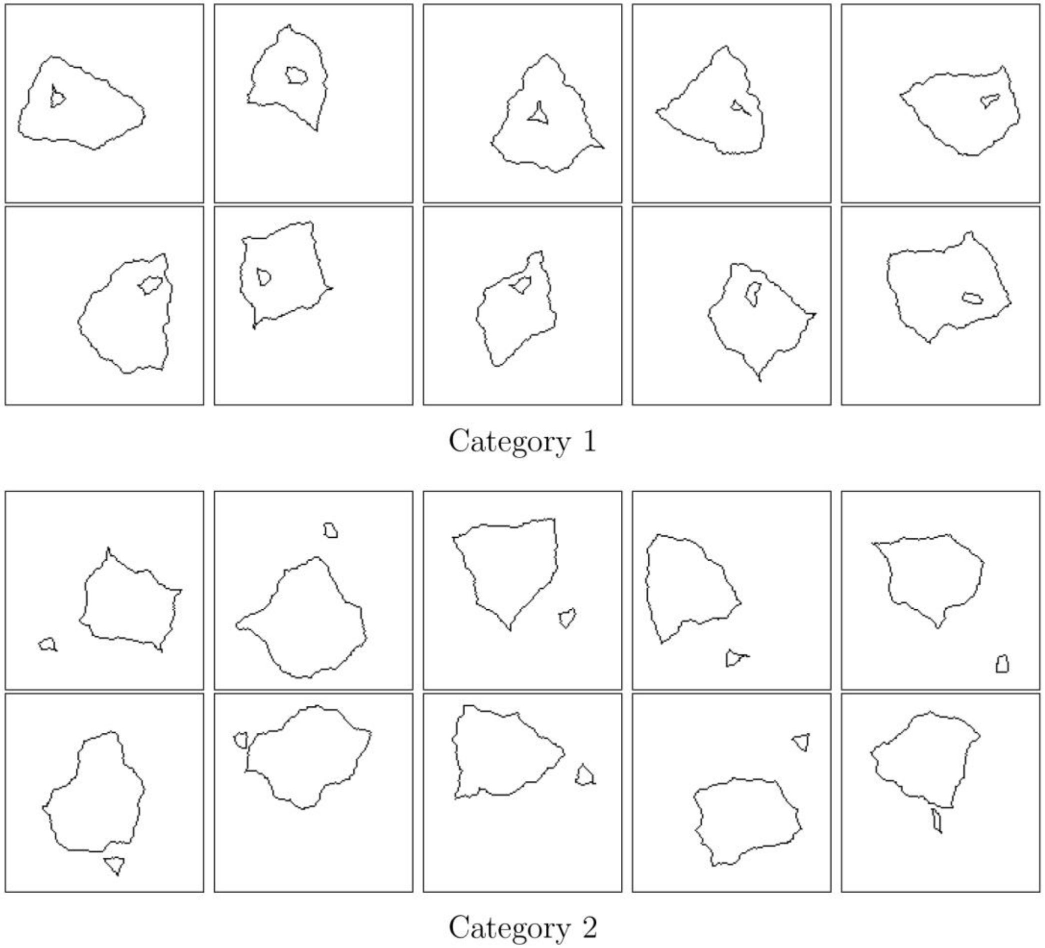
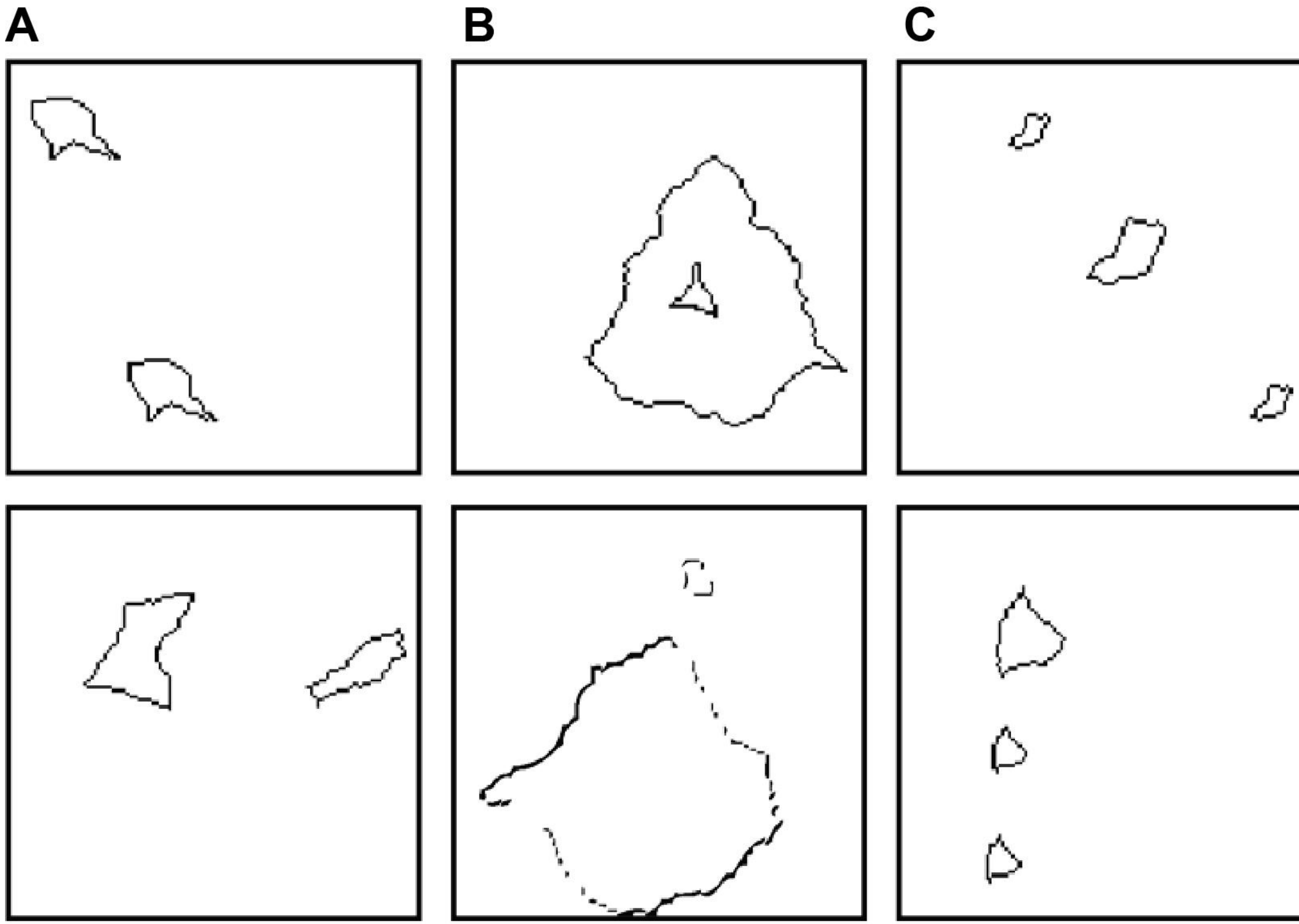


Figure S 10: SVRT Problem #4. Each image contains one big shape and one small shape. In category 1 the small shape is inside the big one and in category 2 the small shape is outside the big one.

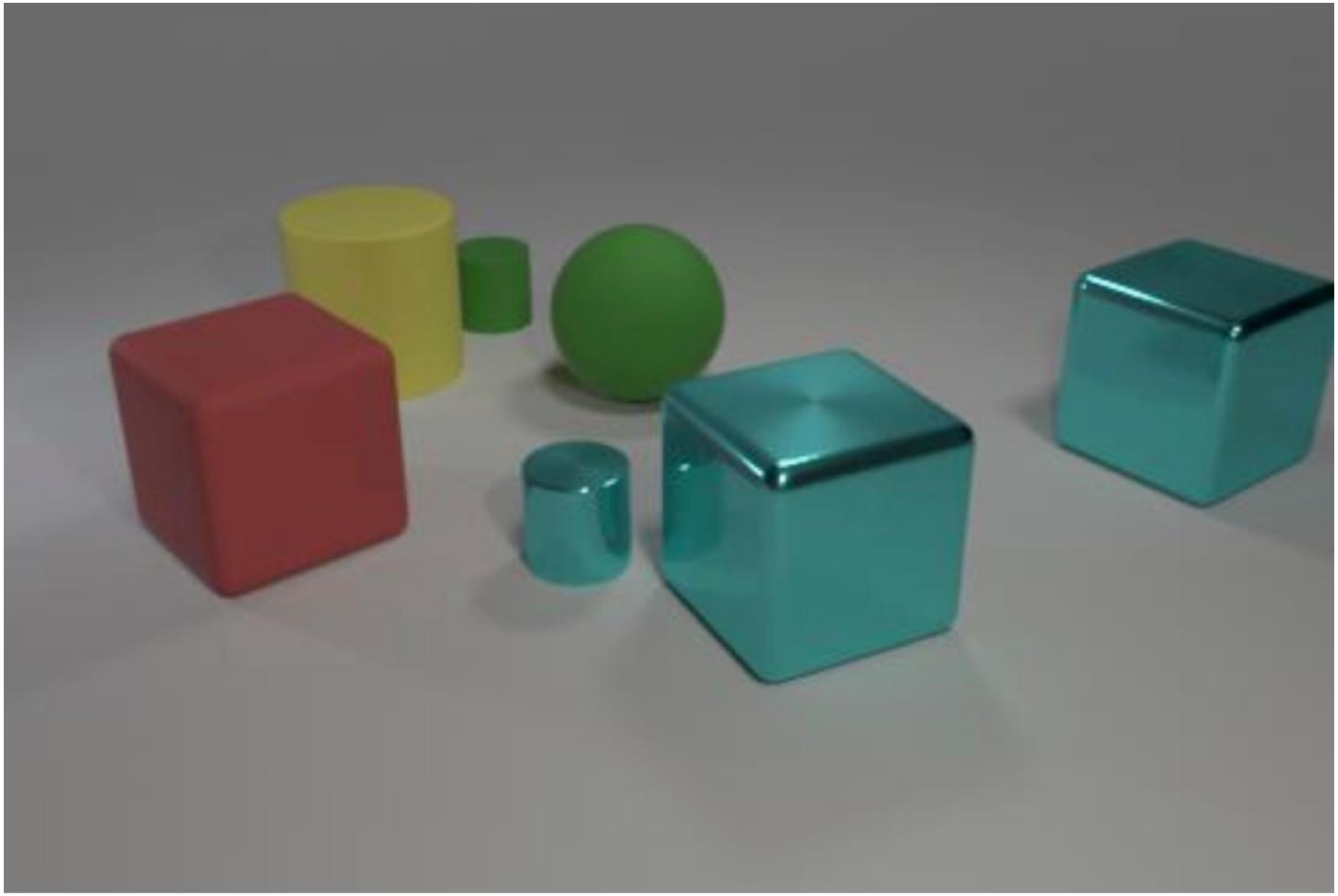
Example visual reasoning tasks



Same/different Inside/outside Large in middle

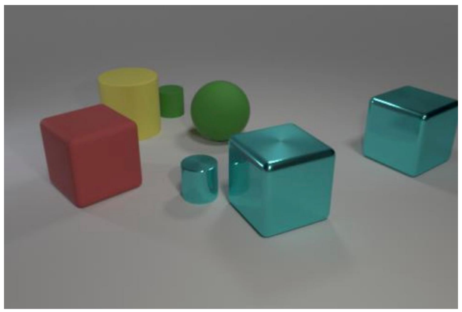
Fleuret et al., 2011
Kim et al., 2018

Example task to test *compositional* visual reasoning—CLEVR dataset

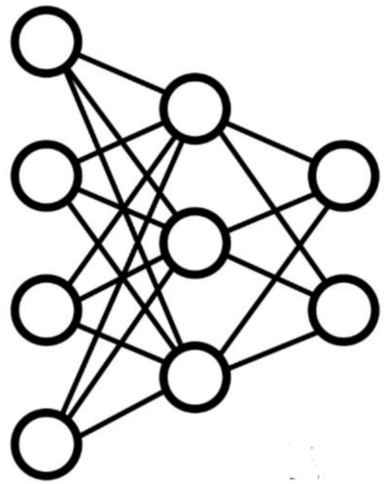


What color is the object with the same size as the blue metallic cylinder?

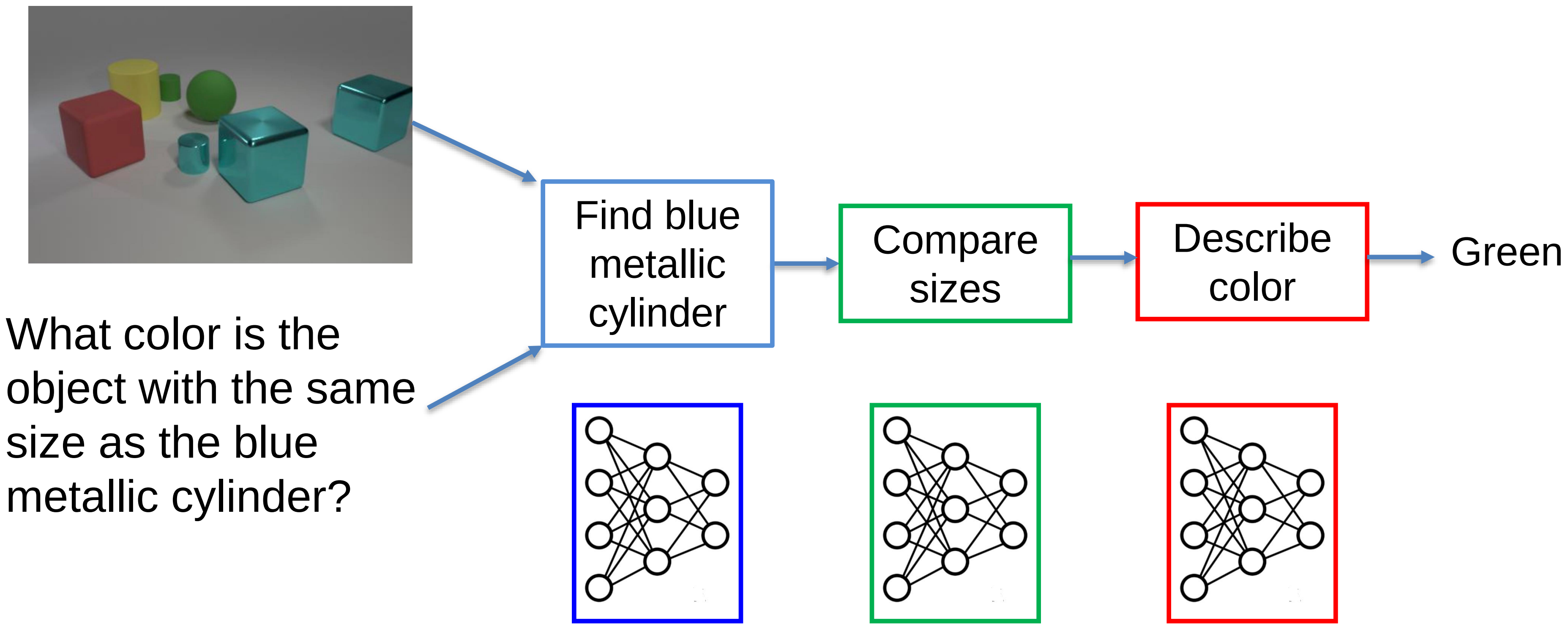
[Johnson et al., 2016](#)



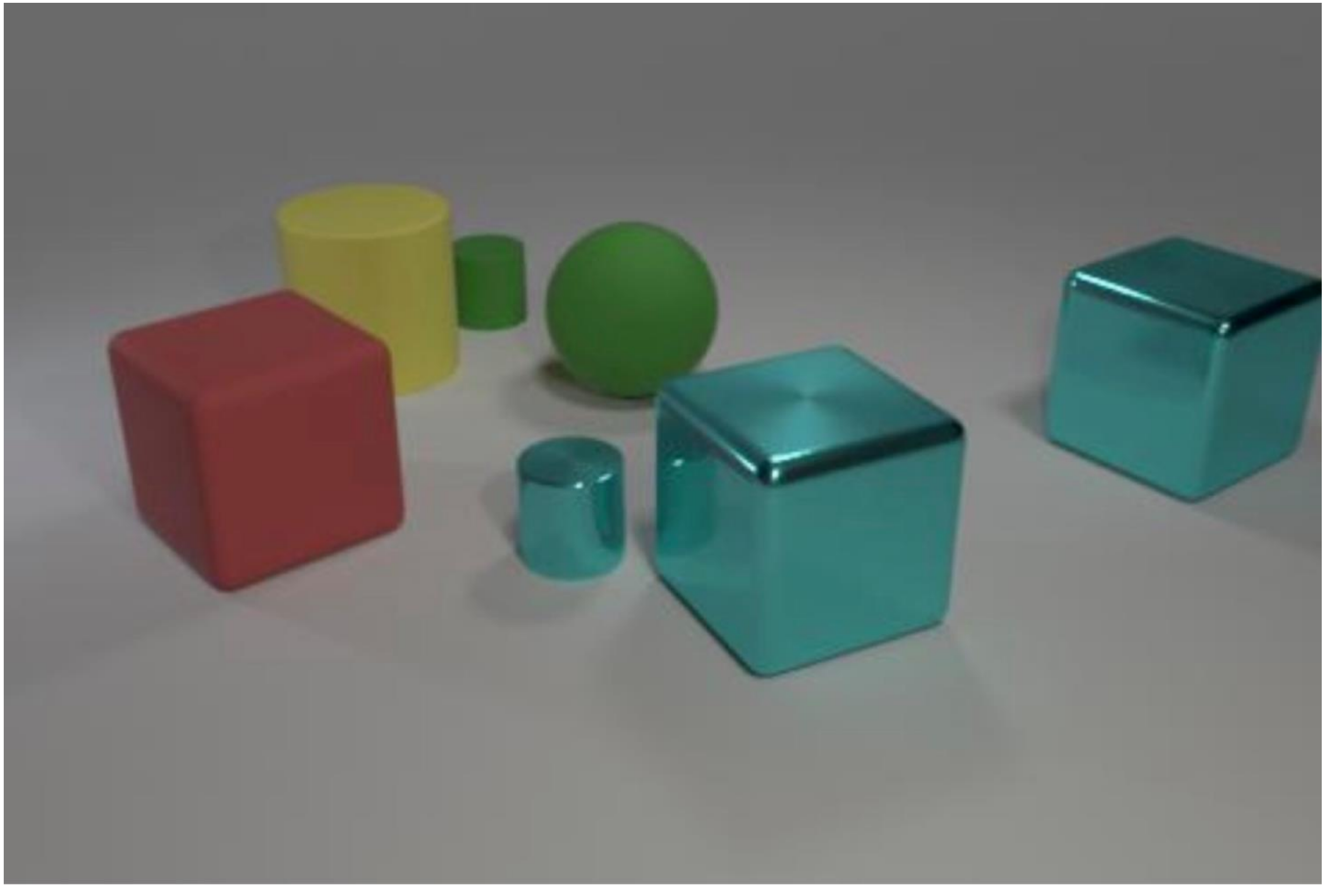
What color is the object with the same size as the blue metallic cylinder?



Green

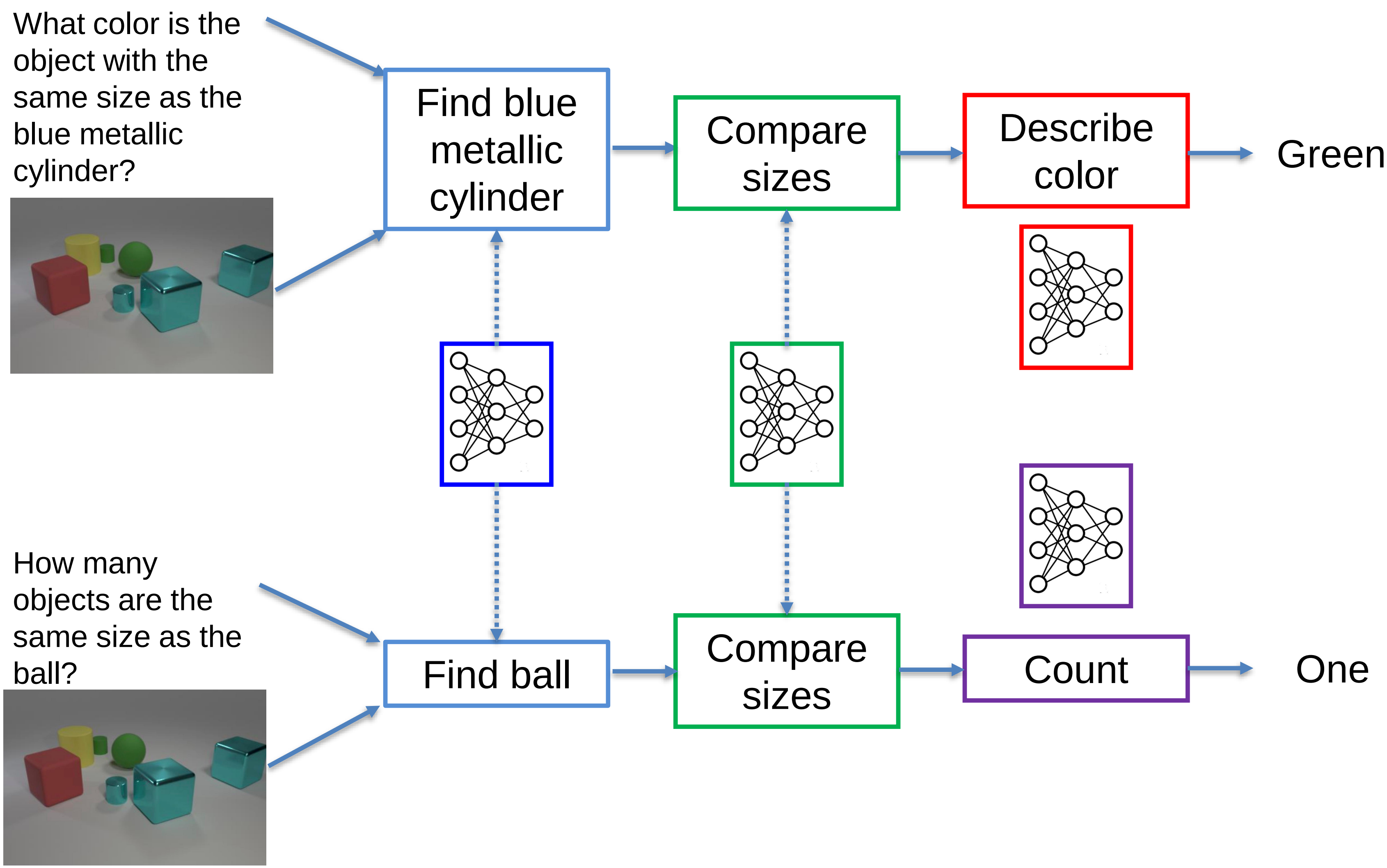

What color is the object with the same size as the blue metallic cylinder?
Find blue metallic cylinder
Compare sizes
Describe color
Green

Visual reasoning



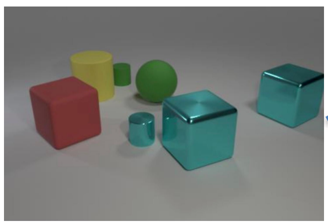
How many objects are the same size as the ball?

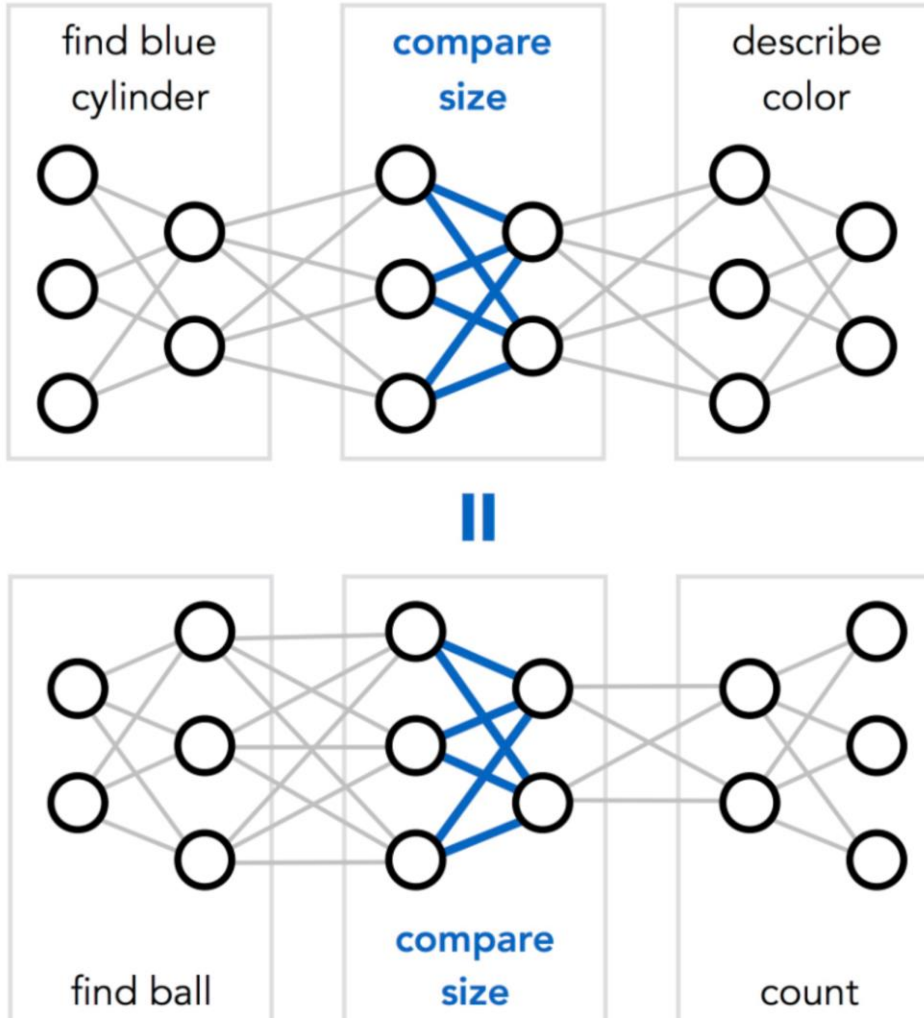
Compositional Visual reasoning
—Reusable subroutines



Compositional Visual reasoning —Reusable subroutines

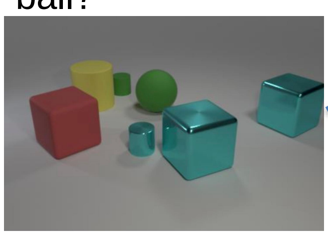
What color is the object with the same size as the blue metallic cylinder?





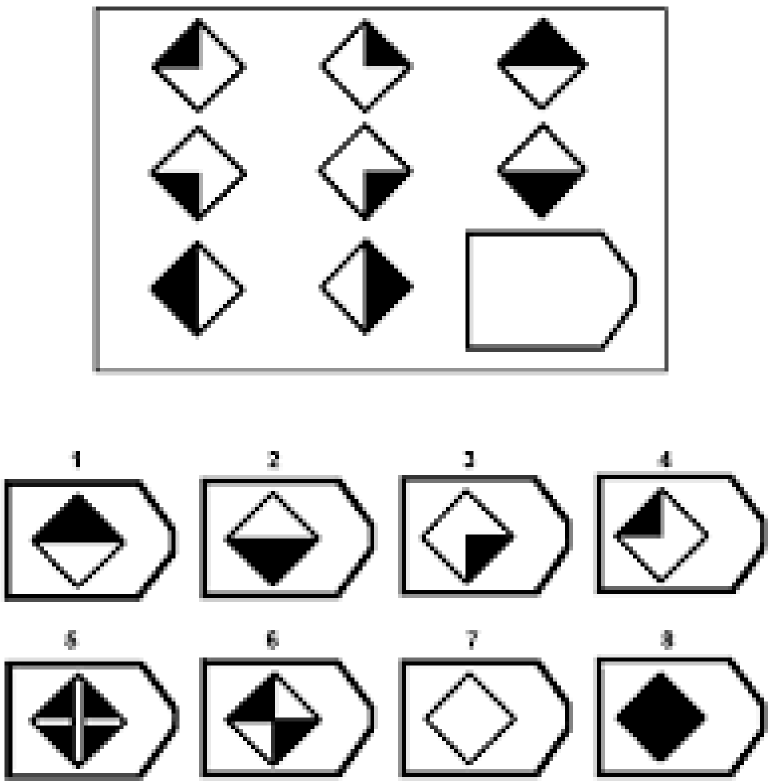
Green

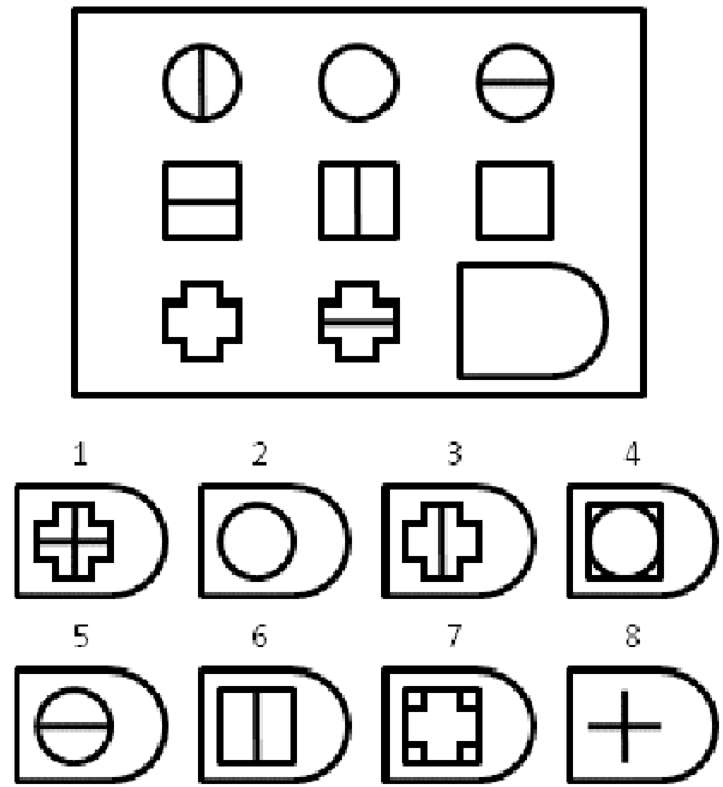
How many objects are the same size as the ball?



One

Raven's progressive matrices (RPMs)





Raven, 1938

RPMs as a testbed for compositional visual reasoning

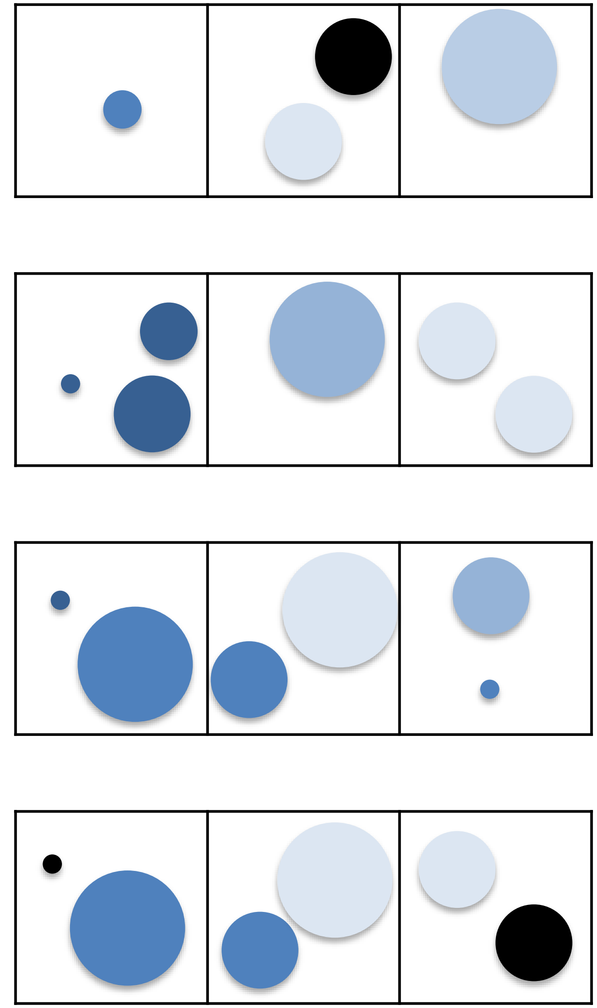
Relation:	Progression	XOR
Attribute:	Size	Color

Standard (i.i.d.) Cross-val.

Train	*Test*
Prog. size	
Prog. color	
XOR size	
XOR color	

Compositional Generalization

Train	Prog. size
	Prog. color
	XOR size
Test	XOR color



Ongoing work by Shane Shang in Gabriel's lab

[Barrett et al., 2018]

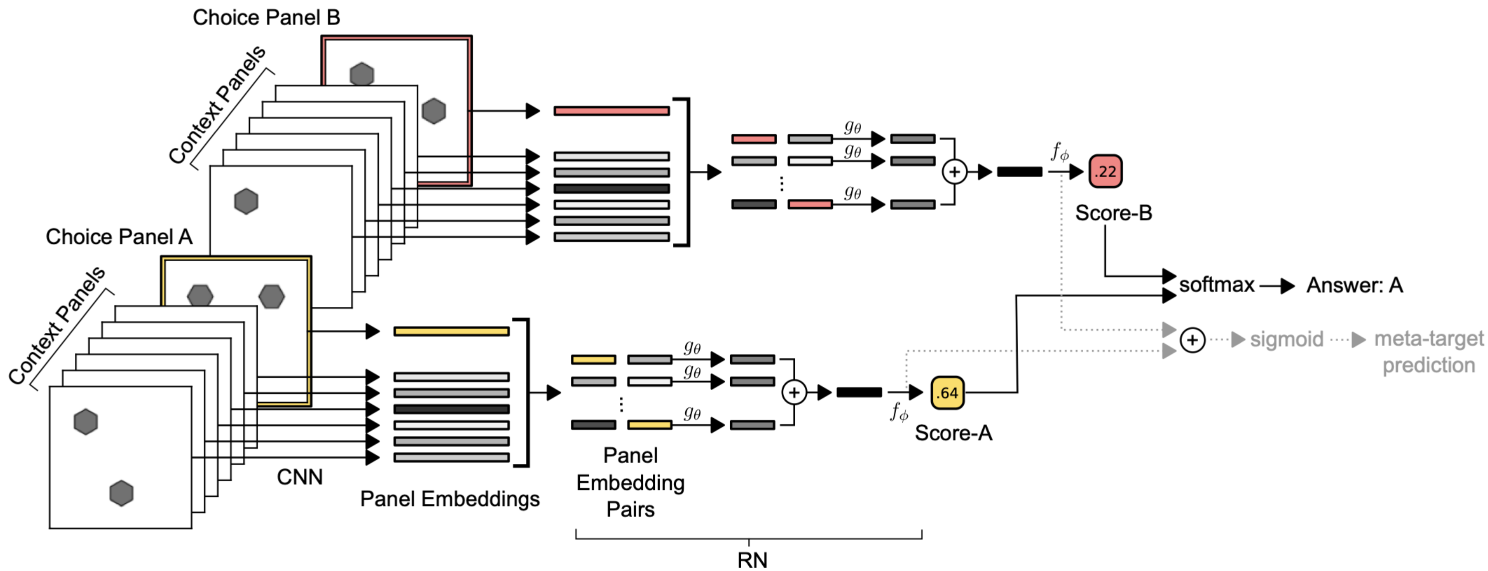


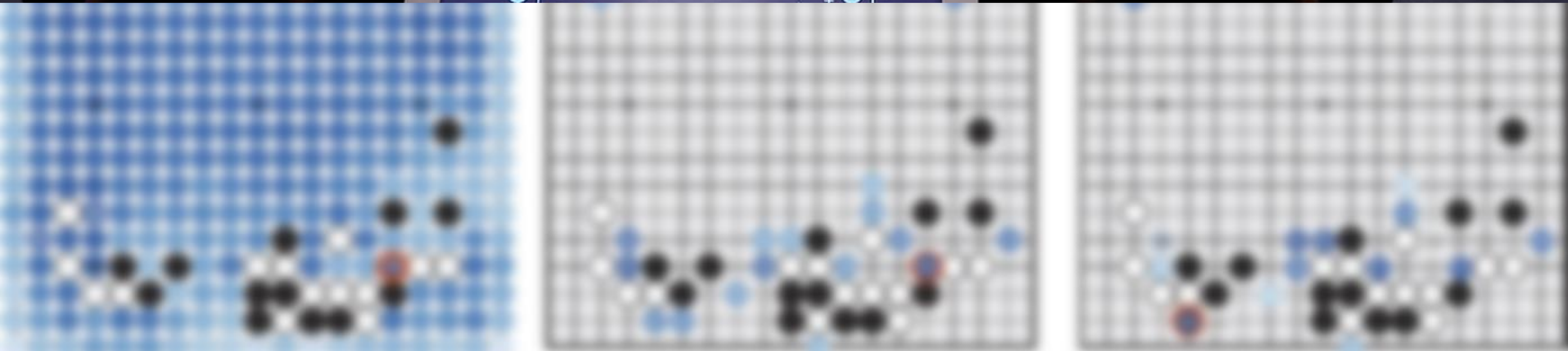
Figure 3. **WReN model** A CNN processes each context panel and an individual answer choice panel independently to produce 9 vector embeddings. This set of embeddings is then passed to an RN, whose output is a single sigmoid unit encoding the "score" for the associated answer choice panel. 8 such passes are made through this network (here we only depict 2 for clarity), one for each answer choice, and the scores are put through a softmax function to determine the model's predicted answer.

Regime	Val. (%)	Test (%)	Diff.
Neutral	63.0	62.6	-0.6
Interpolation	79.0	64.4	-14.6
H.O. Attribute Pairs	46.7	27.2	-19.5
H.O. Triple Pairs	63.9	41.9	-22.0
H.O. Triples	63.4	19.0	-44.4
H.O. `line-type`	59.5	14.4	-45.1
H.O. `shape-colour`	59.1	12.5	-46.6
Extrapolation	69.3	17.2	-52.1

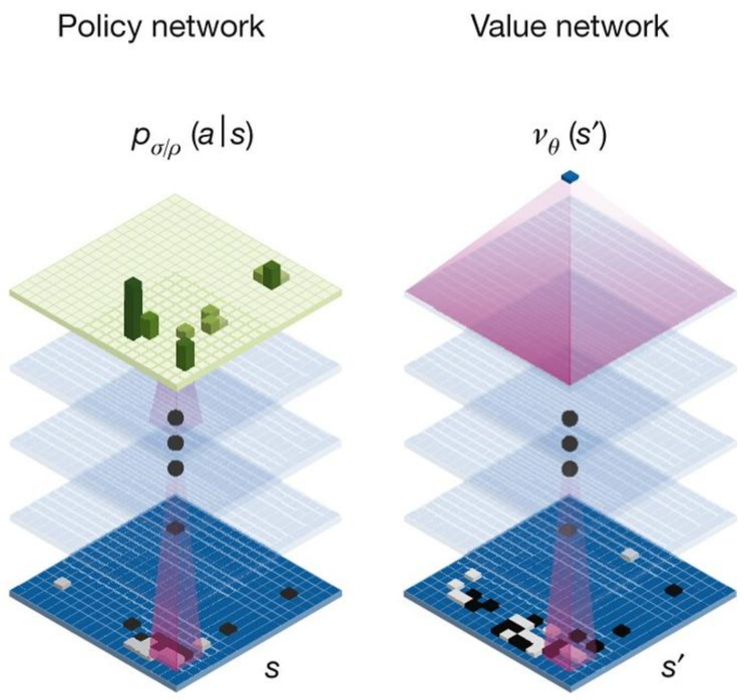
Barrett et al., 2018

Closing the vision-action loop with RL

Closing the vision-action loop with RL—AlphaGo



Closing the vision-action loop with RL—AlphaGo



[Silver et al., 2016]

AlphaGo uses CNNs

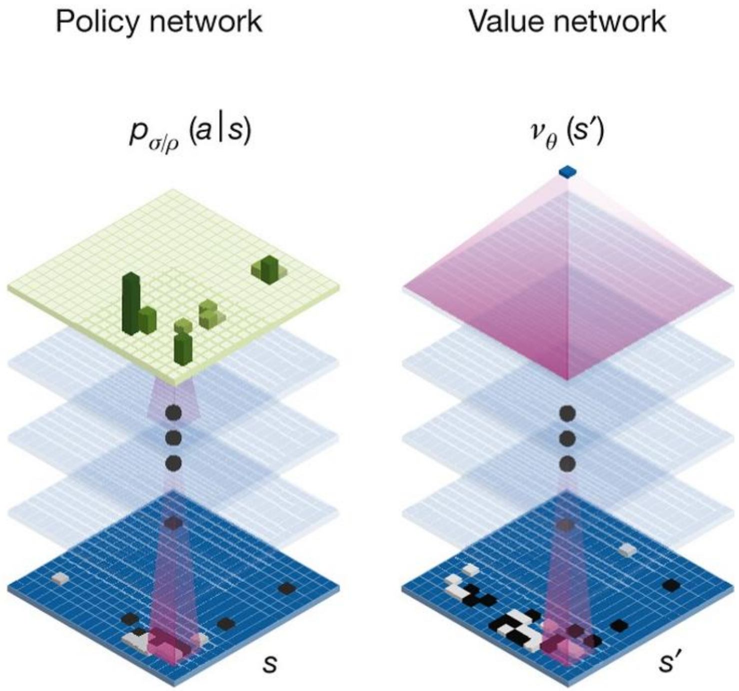
Neural network architecture

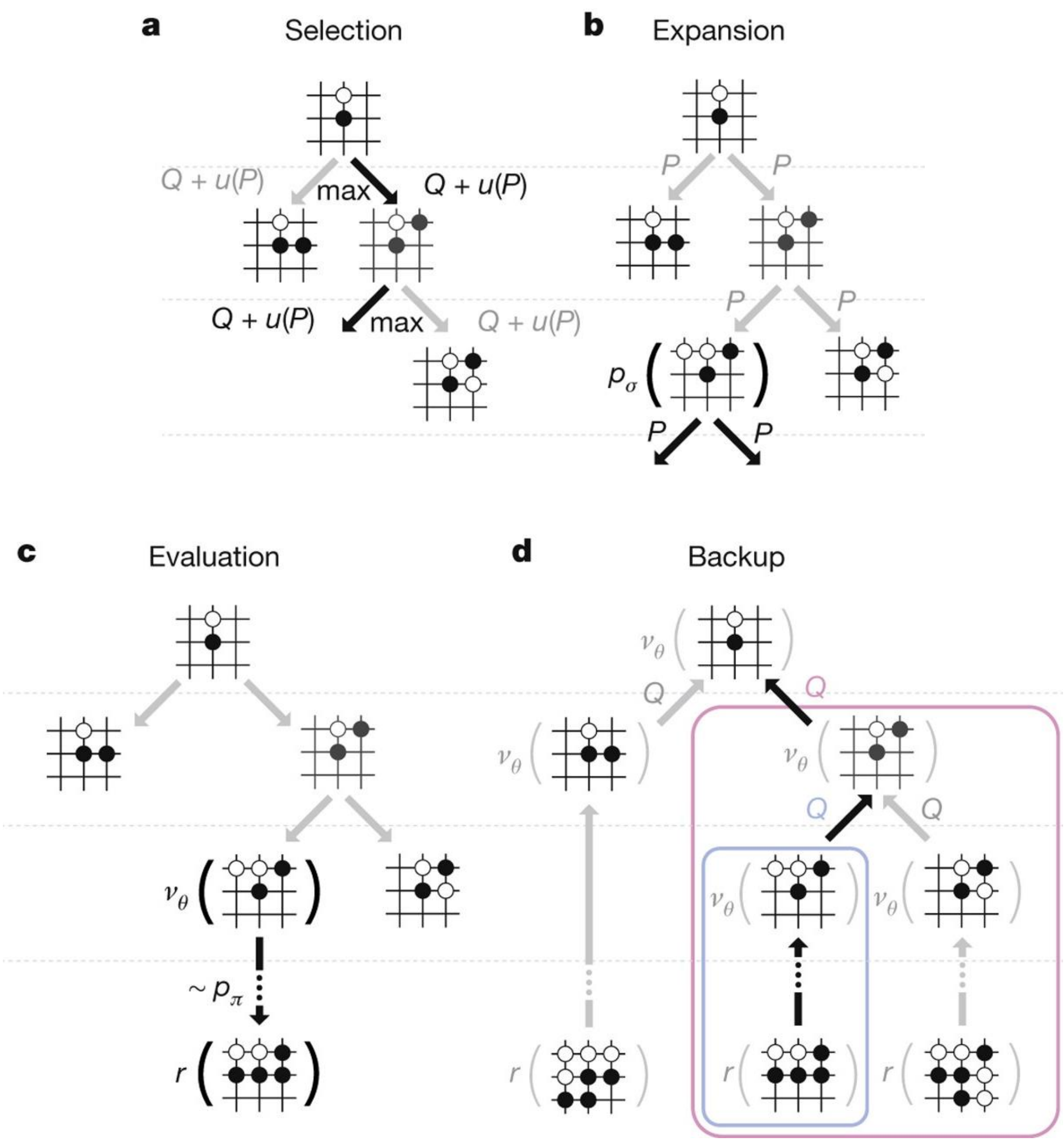
The input to the policy network is a 19 × 19 × 48 image stack consisting of 48 feature planes. The first hidden layer zero pads the input into a 23 × 23 image, then convolves *k* filters of kernel size 5 × 5 with stride 1 with the input image and applies a rectifier nonlinearity. Each of the subsequent hidden layers 2 to 12 zero pads the respective previous hidden layer into a 21 × 21 image, then convolves *k* filters of kernel size 3 × 3 with stride 1, again followed by a rectifier nonlinearity. The final layer convolves 1 filter of kernel size 1 × 1 with stride 1, with a different bias for each position, and applies a softmax function. The match version of AlphaGo used *k* = 192 filters; Fig. 2b and Extended Data Table 3 additionally show the results of training with *k* = 128, 256 and 384 filters.

The input to the value ne
feature plane describing
policy network, hidden la
filter of kernel size 1 × 1 wi
256 rectifier units. The ou

Feature	# of planes	Description
Stone colour	3	Player stone / opponent stone / empty
Ones	1	A constant plane filled with 1
Turns since	8	How many turns since a move was played
Liberties	8	Number of liberties (empty adjacent points)
Capture size	8	How many opponent stones would be captured
Self-atari size	8	How many of own stones would be captured
Liberties after move	8	Number of liberties after this move is played
Ladder capture	1	Whether a move at this point is a successful ladder capture
Ladder escape	1	Whether a move at this point is a successful ladder escape
Sensibleness	1	Whether a move is legal and does not fill its own eyes
Zeros	1	A constant plane filled with 0
Player color	1	Whether current player is black

AlphaGo uses CNNs and tree search





Summary

1. Humans can combine vision with flexible cognitive abilities; algorithms that aspire to do the same can be tested with a “Turing test for vision.”

2. Deep learning models can be built to handle multimodal input.

3. Large language + vision models are making strides toward (casual) human-like ability.

4. An ”intelligent” computer vision system can be of practical help to people who need help seeing.

5. Controlled, compositional tasks are a hard test for visual cognition.

6. Deep learning models can learn vision to support action.

References

- Turing, Alan (October 1950), "Computing Machinery and Intelligence", Mind, **LIX** (236): 433–460, doi:10.1093/mind/LIX.236.433,
- Johnson, Justin, et al. "Inferring and executing programs for visual reasoning." *ICCV*. 2017.
- OpenAI GPT-4. https://openai.com/research/gpt-4
- Liu, Haotian, et al. "Visual instruction tuning." *arXiv preprint arXiv:2304.08485* (2023).
- Fleuret, François, et al. "Comparing machines and humans on a visual categorization test." *PNAS* 108.43 (2011): 17621-17625.
- Kim, Junkyung, Matthew Ricci, and Thomas Serre. "Not-So-CLEVR: learning same–different relations strains feedforward neural networks." *Interface focus* 8.4 (2018): 20180011.
- Johnson, Justin, et al. "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning." *CVPR*. 2017.
- Raven, J. C. (1936). *Mental tests used in genetic studies: The performance of related individuals on tests mainly educative and mainly reproductive.* MSc Thesis, University of London.
- Barrett, David, et al. "Measuring abstract reasoning in neural networks." *International conference on machine learning*. PMLR, 2018.
- Silver, David, et al. "Mastering the game of Go with deep neural networks and tree search." *Nature* 529.7587 (2016): 484-489.
- Also see: Silver, David, et al. "Mastering the game of go without human knowledge." *Nature* 550.7676 (2017): 354-359.

Visual Object Recognition
Computational Models and Neurophysiological Mechanisms
Neuro 130/230. Harvard College/GSAS 78454