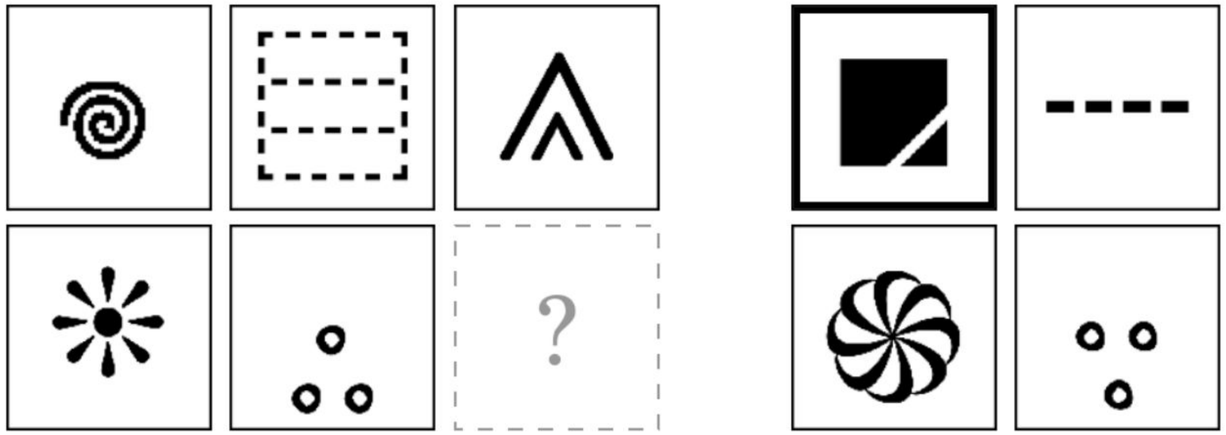


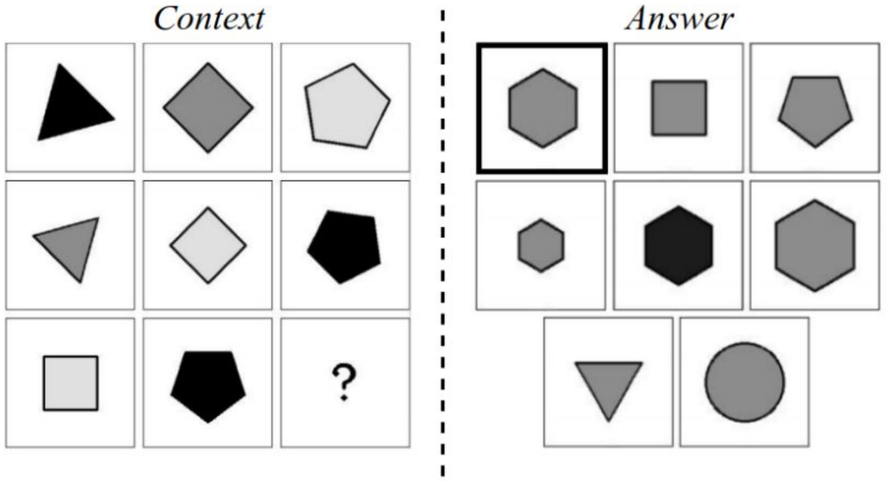
Unicode Analogies:
An Anti-Objectivist Visual Reasoning Challenge

Steven Spratley · Krista A. Ehinger · Tim Miller
School of Computing and Information Systems, The University of Melbourne

THE UNIVERSITY OF MELBOURNE
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

POSTER SESSION: THU-AM-248

Progressive Matrix Problems (PMPs)



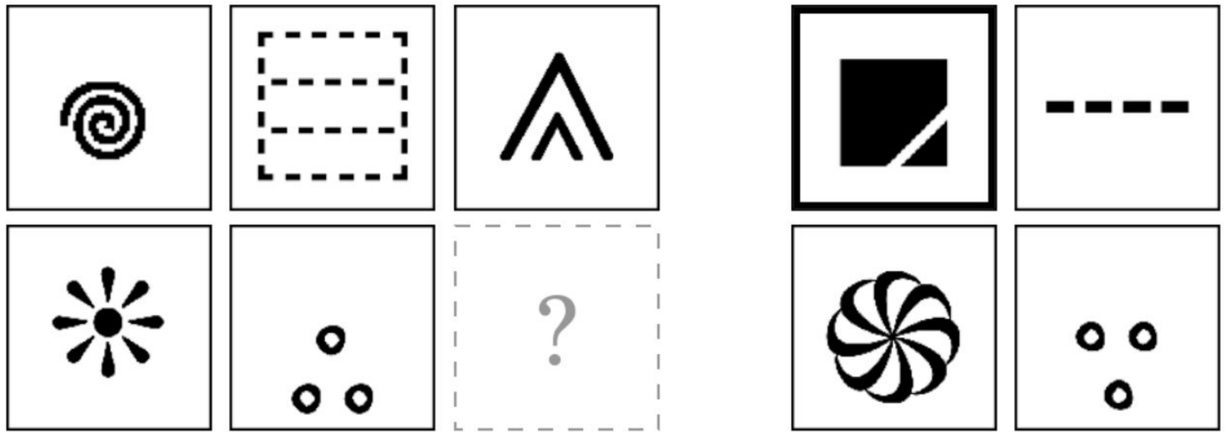
- Started with John Raven's "progressive matrices" in the 1930s
- Current machine learning datasets include PGM and RAVEN
- Traditional three rows of context, with the last frame left blank
- Pick an answer that best completes that row, given the other two

THE UNIVERSITY OF MELBOURNE

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

POSTER SESSION: THU-AM-248

Vision vs. objectivism



- Training systems to learn objectively labelled scenes does not capture the way humans assign concepts ad hoc, in context
- Unicode Analogies is anti-objectivist, polysemic, and gestalt. It challenges solvers to perceive scenes at a level of abstraction suited to the problem, not just the one it was trained on.

THE UNIVERSITY OF MELBOURNE

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

POSTER SESSION: THU-AM-248

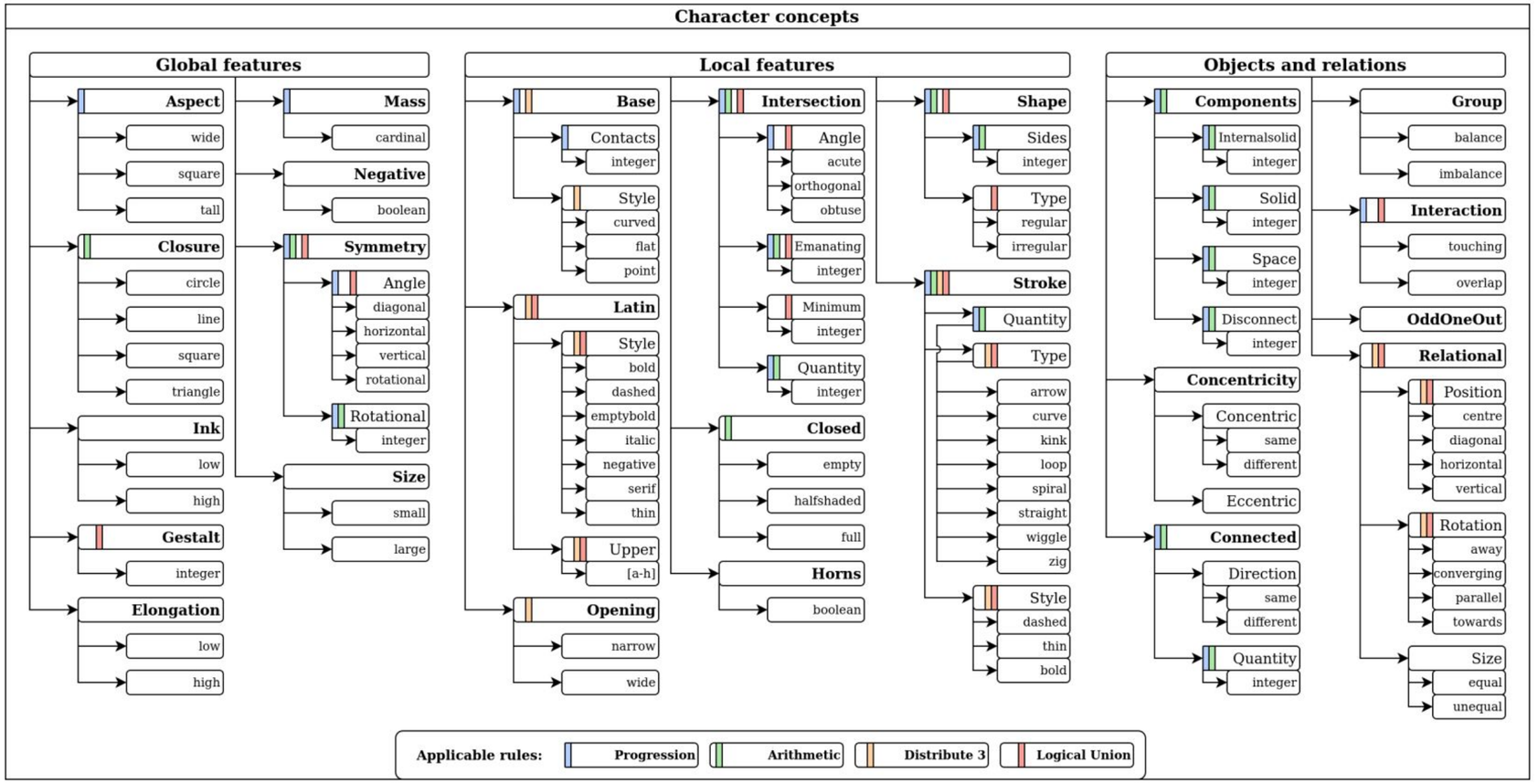
Non-robust features

- How do we perceive the above image?
  - A square, due to gestalt closure
  - An object with high ink level
  - A group with a small triangle
  - A diagonally-arranged pair
  - An arrangement with an equal aspect ratio
  - An arrangement with centred mass
  - An arrangement with two base contacts

- A perceptual take is therefore a question of context and of usefulness, and not of assigning an objective label
- Images in our problems are hand-labelled Unicode characters
- By replacing simple primitives in other datasets with such characters, we create highly diverse problems

THE UNIVERSITY OF MELBOURNE
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA
POSTER SESSION: THU-AM-248

Construction



THE UNIVERSITY OF MELBOURNE

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

POSTER SESSION: THU-AM-248

Experiments

Five key experiments:

1) Rule: Model vs. human performance across all rule types offered by the dataset

2) Schema: Performance explored across three schema subcategories (global, local, and objects & relations)

3) Extrapolation: Models are tested against four extrapolation splits of increasing difficulty

4) Challenge: Concept types are summarised based on a comparison between human and model performance, and used to create a train-test split to probe this disparity

5) Hold-out: The influence of character hold-out strategies is examined, motivating our use of disjoint character sets to exacerbate non-robust feature learning

POSTER SESSION: THU-AM-248

Experiments

Baselines and models:

- Context-blind
- ResNet
- MRNet
- SCL
- Rel-Base
- Human

THE UNIVERSITY OF MELBOURNE

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

POSTER SESSION: THU-AM-248

Results - Rule

Method	Avg	Const	Prog	Arith	Dist3	Union
Blind	27.0	29.5	**29.6**	24.3	28.1	29.7
ResNet	27.4	30.9	26.7	25.7	31.9	30.0
MRNet	**31.1**	33.9	26.8	27.4	34.4	32.9
SCL	28.9	30.1	25.2	25.8	30.7	31.2
RelBase	30.8	**34.5**	28.5	**29.7**	**36.9**	**34.2**
Human	55.5	55.0	65.0	54.0	55.0	42.0

- Model accuracy is almost universally below 35%
- Context-blind solver never achieves more than 5% above random

THE UNIVERSITY OF MELBOURNE
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA
POSTER SESSION: THU-AM-248

Results - Schema and Extrapolation

Method	Global	Local	Obj. & Rel.
Blind	34.0	25.8	25.3
ResNet	35.1	26.5	25.6
MRNet	**39.3**	**30.1**	24.9
SCL	34.1	26.0	24.9
RelBase	39.0	30.0	**26.3**
Human	52.6	58.3	52.2

Method	No Shift	Neutral	Extra	Extra +
Blind	27.0	26.9	26.7	25.6
ResNet	27.4	27.0	27.0	24.9
MRNet	**31.1**	30.2	**28.9**	27.9
SCL	28.9	27.9	27.5	25.7
RelBase	30.8	**31.0**	28.1	**29.5**

THE UNIVERSITY OF MELBOURNE
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA
POSTER SESSION: THU-AM-248

Results - Challenge

Human	Model (RelBase)
Challenge split performance (accuracy and difference)	
71.9%	**31.7%** (-40.2)
Top-5 concepts	
negative	global-size
horns	negative
arrow-quantity	ink
dash-quantity	latin-style
internalsolid	dash-quantity
Bottom-5 concepts	
oddoneout	u-quantity
opening	zig-quantity
base-contacts	arrow-quantity
space	interaction
uniquesolid	uniquesolid

Challenge split performance, other models

	Blind	ResNet	MRNet	SCL
Accuracy	24.8	27.2	28.1	27.7
Difference	-47.1	-44.7	-43.8	-44.2

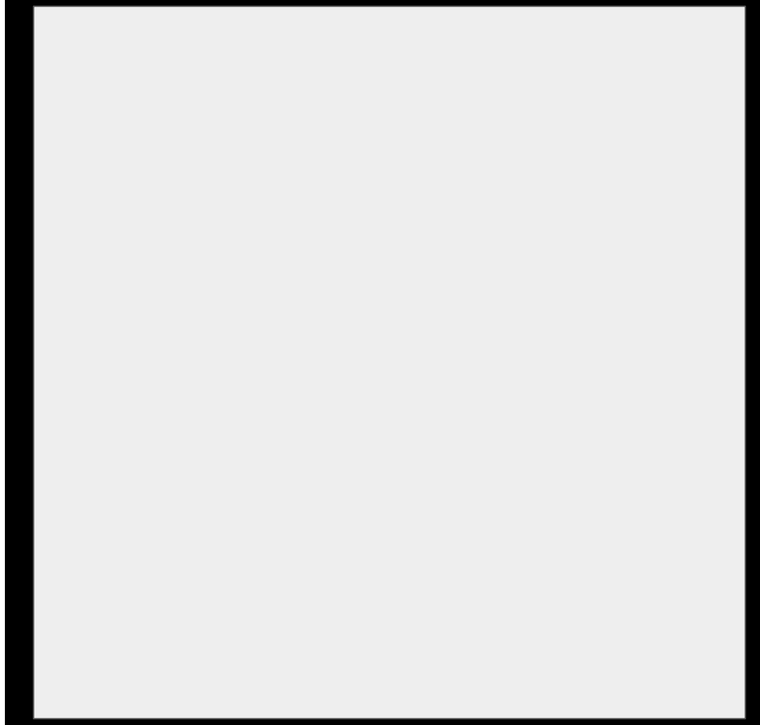
THE UNIVERSITY OF MELBOURNE

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

POSTER SESSION: THU-AM-248

Results - Hold-out

Constant split performance, all models					
H-O	Blind	ResNet	MRNet	SCL	RelBase
None	25.8	31.3	38.5	41.0	**52.2**
Diff.	29.5	30.9	33.9	30.1	34.5



POSTER SESSION: THU-AM-248

Summary: What did we learn?

- Model accuracy is almost universally below 35, often performing close to random
- Context-blind solver never achieves more than 5% above what would be expected of random chance, with more advanced architectures only performing within 10% of it
- Concept-by-concept breakdown of human and model performance
- Performance increases when the same character set is used to assemble both train and test problems, despite rule-class-value tuples being disjoint across splits
- Across all experiments, we notice that despite architectural differences between tested models, similar results were achieved, with the exception of experimentation on different hold-out sets

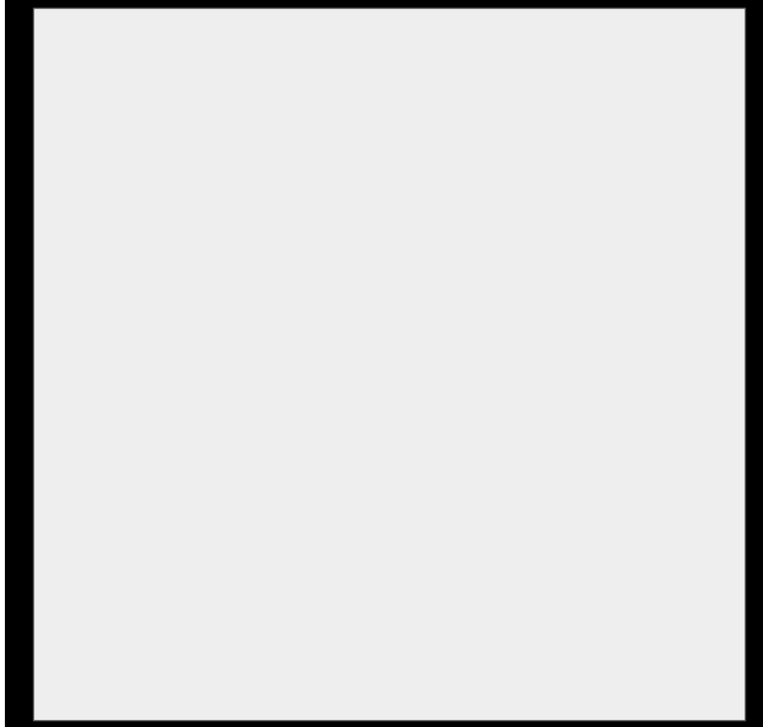
THE UNIVERSITY OF MELBOURNE
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA
POSTER SESSION: THU-AM-248

Impact and future work

- Unicode Analogies is an extensible framework
  - Easily adapted to new concepts and schemas
  - Likely to be of interest in AI and cognitive science
  - Allows for fine-grained analysis on concept acquisition

- We introduced the Unicode Analogies challenge, which assembles novel PMPs from diverse and disjoint sets of character images, and brings fluid perception to the progressive matrix format

- We encourage new solvers, and are excited to see how this framework is adopted by our research community



POSTER SESSION: THU-AM-248