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
https://github.com/SvenShade/UnicodeAnalogies

Abstract

Analogical reasoning enables agents to extract relevant information from scenes, and efficiently navigate them in familiar ways. While progressive-matrix problems (PMPs) are becoming popular for the development and evaluation of analogical reasoning in computer vision, we argue that the dominant methodology in this area struggles to expose the lack of meaningful generalisation in solvers, and reinforces an objectivist stance on perception – that objects can only be seen one way – which we believe to be counterproductive. In this paper, we introduce the Unicode Analogies challenge, consisting of polysemic, character-based PMPs to benchmark fluid conceptualisation ability in vision systems. Writing systems have evolved characters at multiple levels of abstraction, from iconic through to symbolic representations, producing both visually interrelated yet exceptionally diverse images when compared to those exhibited by existing PMP datasets. Our framework has been designed to challenge models by presenting tasks much harder to complete without robust feature extraction, while remaining largely solvable by human participants. We therefore argue that Unicode Analogies elegantly captures and tests for a facet of human visual reasoning that is severely lacking in current-generation AI.

1. Introduction

Traditionally, statistical classification models have been designed to neatly cleave data into categories. Even in tasks such as visual scene decomposition, where data resists full description by any one label, there is an underlying objectivist assumption being made; the expectation of there being an objective number of distinguishable "things" present, themselves belonging to singular classes. Human visual perception makes a departure from this. The symbolic world to which we attend, with firm compositional rules for scenes and their objects, and with their parts and positions, is subsisted by a churning sea of ongoing conceptualisation processes deeply fluid and contextual [15].

In recent years, there has been a proliferation of computer vision architectures built with object-centric inductive biases [21], many of which represent states-of-the-art on popular datasets [7, 35, 41]. This is an important direction, as training models to decompose scenes into objects allows for an explicit abstraction stage promoting feature reuse. However, abstract visual reasoning tasks such as Bongard problems [3] expose philosophically [20] — and in this paper, experimentally — that such an approach might work against the creation of models that possess the ability to abstract and deploy useful concepts. This observation also engages a current debate in the literature regarding the scalability of built-in knowledge and inductive biases [24, 36].

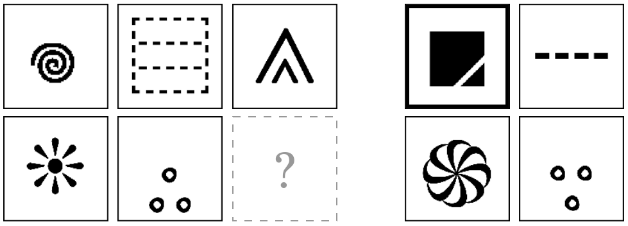

Figure 1. An example problem in UA, instantiating the *Distribute-Three* rule with the *Closure* concept. Five out of six context frames are provided (left), with four answer frames to choose from (right). The correct answer is emboldened.

Humans display flexibility in how they decompose scenes, and perceive such scenes at a level of abstraction informed by past experiences and appropriate to present goals [8, 9]. Scene understanding in humans is therefore undergirded by something other than the perception of static objects [22], and the idea that scene modelling research can separate perception and higher cognition into a pipeline of self-contained modules is strongly critiqued [5].

Noticing other shortcomings of deep-learnt approaches to computer vision, including brittleness to out-of-distribution (OOD) data, a small number of abstraction datasets inspired by Raven's Progressive Matrices have been recently released [23]. Further motivations to this direction include a) the expectation that tasks with such an extended history in general psychometric testing would be useful to import into computer vision research, and b) the opinion that the more broadly applicable a model's abstracted concepts become, the more robust that model will be under OOD conditions [29, 35]. While the applicability of such concepts should ideally be evaluated by these datasets, common approaches to dataset creation feature conceptual schemas consisting of simple objects that can be neatly dropped into scenes, and extracted by scene decomposition stages [35]. This seems to require little in the way of contextual perception, such as Hofstadter's notion of "conceptual slippage" [16].

We observe that the world's writing systems present a diverse resource of characters that are amenable to content analysis, and can assemble novel reasoning problems of their own. We introduce the *Unicode Analogies* (UA) challenge, consisting of character-based progressive matrix problems (PMPs) to benchmark fluid conceptualisation ability in vision systems. The characters in UA are polysemic, and may instantiate any number of concepts, with the salient concept only revealing itself given context (Fig. 1). By generating training and testing problems from disjoint sets of characters, we challenge these systems by presenting tasks much harder to complete without robust feature extraction, while remaining largely solvable by human participants. In doing so, we contribute a dataset that unlike others in this area, operates on a rich conceptual schema that invites fine-grained experimentation, and is easily extensible to new user-defined concepts. Over five key experiments, we explore human and model performance on a number of dataset splits generated by UA, demonstrate that state-of-the-art solvers are still far from achieving the founding goals their datasets were created for, and encourage new solvers to overcome these limitations.

2. Background

2.1. Vision vs. objectivism

For humans, our perceptual world is not populated by firm and unchanging concepts, as if there were some neatly defined mental collection. There is a wealth of psychological research to suggest that cognition – at the levels of conceptualisation [4], reasoning [11], and memory [34] – operates on concepts that are blurred, evolving, fluid, and *ad hoc*. Consider the child who perceives a tree stump surrounded by mushrooms as a dining setting for small creatures. Such analogies are ubiquitious in how we understand scenes not simply as lists of objects, but micro-worlds with physics, rules, structure, intent, and purpose. Via analogy, these worlds, which may not be previously experienced, are "seen as" familiar, in order for us to successfully traverse and manipulate them, efficiently guiding perception and problem solving [12]. We believe that a deep understanding of concepts is demonstrated by the ability to both perceive them in diverse stimuli, and to leverage them for utility. Echoing Odouard and Mitchell [29], the way we assess trained models needs to remain fully aware of this.

2.2. Progressive matrix problems and deep learning

Since their introduction in 1936, Raven's Progressive Matrices (RPMs) have seen extensive use in psychometric testing [32,33], in part due to their abstract, non-verbal, and assumedly culture-agnostic design, as well as their simplicity to administer. RPMs present a visual pattern-matching task requiring solvers to perform analogical reasoning, and such reasoning must depend on the company of context images if a solution is to be found and analogy drawn.

In the field of computer vision, deep learning is ubiquitous in leading models, bringing with it both the remarkable ability to perform rich, automatic feature extraction from large datasets, and a severe brittleness to out-of-distribution data. As analogical reasoning in human and non-human animals is hypothesised to support feats such as tool use and creation, and indeed, general problem solving [15], there has been much interest in creating RPM-inspired datasets amenable to deep learning. In this paper, we refer to the problems presented by all such datasets as belonging to the class of progressive matrix problems (PMPs).

The last five years has seen the release of several abstract reasoning datasets, including two seminal PMP datasets; PGM [1] and RAVEN [39]. PGM is considered the first large-scale dataset of its kind, while RAVEN builds upon it, increasing the diversity of rules and configurations instantiated by problems to discourage memorisation and more accurately assess the generalisation ability of trained models. Since RAVEN's release, there have been a number of research efforts (reviewed by Malkinski and Mandziuk [23]) to benchmark novel architectures on its problems, each analysing model performance primarily informed by overall accuracy. While this presents as a fairly standard methodology in machine learning research, there are more nuanced considerations that this branch of research demands.

2.3. Shortcuts and non-robust features

For any given data, there exists a landscape of "perfect" models, i.e. those that have full explanatory power for those data. Knowing which models will also ultimately capture the knowledge to describe additional data is a contentious question [24]. Humans have evolved many biases, such as the preference for simple explanations [37]. Ironically, the tendency to use analogies has meant expecting broader cognitive abilities of our seemingly mind-like models.

Recent works have exposed the existence of shortcut and non-robust feature learning in neural networks [10, 18]. Geirhos et. al communicate that shortcut learning is a failure to generalise "in the right direction", where a model extracts and depends on features that are not present OOD [10]. Similarly, Langosco et. al explain that learning non-robust features and objectives occurs when a network encapsulates the "wrong" knowledge, i.e. that fulfils optimisation in a way that wasn't intended by researchers [18].

In our research area, such phenomena have resulted in networks failing to learn generalisable features, and exploiting biases in PMP answer sets without the awareness of researchers at the time of publication [35]. Recognising this as an important consideration for dataset creation, we have designed splits in Unicode Analogies to assemble train and test problems from disjoint sets of images, requiring models to learn robust features if they are to perform well.

2.4. Comparisons to other datasets

Datasets such as PGM and RAVEN represent important developments in this field, being the first to automate PMP generation at-scale for deep learning, and with enough diversity to pose a challenge for machine solvers at their times of release. However, they also represent one particular approach to PMP formation, adopting basic object-based schemas, and building complexity by stacking multiple rule instantiations in a given problem. While more recent architectures have become adept at modelling the default splits of these datasets, there is less focus on universally poor extrapolation performance, which has seen relatively little progress [23]. To account for this discrepancy, we hypothesise that this approach to PMP generation and testing is not fully diagnostic of an architecture's analogical reasoning abilities. The familiarity of stimuli invites architectures to separate perception from higher-order cognition, allowing much of the work of the problem to fall to representation learning. If it were not so, rule-stacking would have a greater negative impact on performance than is observed, and introducing modified stimuli would be less detrimental.

Unicode Analogies (UA) is able to broadly express the schemas utilised by these datasets, including a familiar exploration of rules such as progression and arithmetic, objects including shapes and lines, and attributes like size and number, to name a few. However, these are situated within a far richer schema of concepts at multiple levels of abstraction, many of which are inspired by Bongard problems [3] and principles of gestalt perception, including closure, negative space, and grouping. In doing so, it blurs the lines between object and feature, and between perception and cognition, forcing models to incorporate contextual information at all stages of problem solving. This dataset brings PMP research in-line with philosophical criticisms of objectivist approaches to AI [5], prohibiting solvers from relying on scene decomposition stages. It presents just one rule per problem, asking solvers to discover what is salient, instead of learning to represent scenes *a priori*. It also responds to the call for datasets to support concept-based evaluation as voiced by Odouard and Mitchell [29], which is a valid criticism across all other datasets we are aware of.

Most similar to our work is the Bongard-LOGO dataset [28], which also motivates context-dependent perception as a crucial property of human cognition. Bongard-LOGO presents a few-shot benchmark intended for meta-learning, and focuses on capturing the Bongard problem format with frames consisting of generated line drawings. UA instead benchmarks supervised learning approaches such as those built for RAVEN and PGM, importing concepts from Bongard and other formats into progressive matrices. Bongard-LOGO exclusively investigates invariant perception with regards to size, orientation, and position, whereas Unicode Analogies does not limit itself in this way.

Other notable datasets, including KANDINSKY [17], ARC [6], PQA [31], and LABC [14], have related goals. The KANDINSKY set explores spatial and gestalt visual tasks within the context of explainability research, and suggests importing such concepts into progressive matrices as future work. ARC presents a corpus of hand-designed intelligence tasks that require models to generate coloured grids. ARC contains a broader, more complex task base than UA, and is intended to be a very general battery for machine intelligence testing, whereas UA is focused primarily on fluid perception. PQA borrows ARC's gridworld format, and introduces seven tasks related to laws of gestalt perception, already largely solvable by the technique offered in the paper. LABC offers a two-row variant to the problems in PGM, focusing on analogical reasoning across domain shifts, while inheriting the same basic schema as PGM.

3. The *Unicode Analogies* Framework

Unicode Analogies is an extensible framework that allows for the creation of character-based PMP datasets from a conceptual schema. In this section, we introduce this work as a pipeline from schema formation, character annotation, and problem generation, through to defining training splits.

3.1. Conceptual schema

Starting with known concepts with historical usage in PMPs and Bongard problems (e.g. rules involving progression or distribution, and features such as size and shape), the first author performed content analysis on hundreds of characters appearing in the Unicode standard. By following a conventional approach to content analysis resembling Mayring's inductive category development [25], we formalised a broad conceptual schema with which to annotate more characters, allowing software to assemble thousands of novel PMPs. The current conceptual schema is shown in Fig. 2, and depicts many concepts across multiple levels of hierarchy. Generated problems express one of 5 rule types: *constant*, *progression*, *arithmetic*, *distribute-three*, and *union*, with each rule applicable to a subset of concepts. The first 4 of these rules are explored in RAVEN under the same names [39]. In PGM, *distribute-three* is referred to as *consistent-union*, and *union* as *logical-OR* [1]. We refer the reader to these works if such rules are unfamiliar.

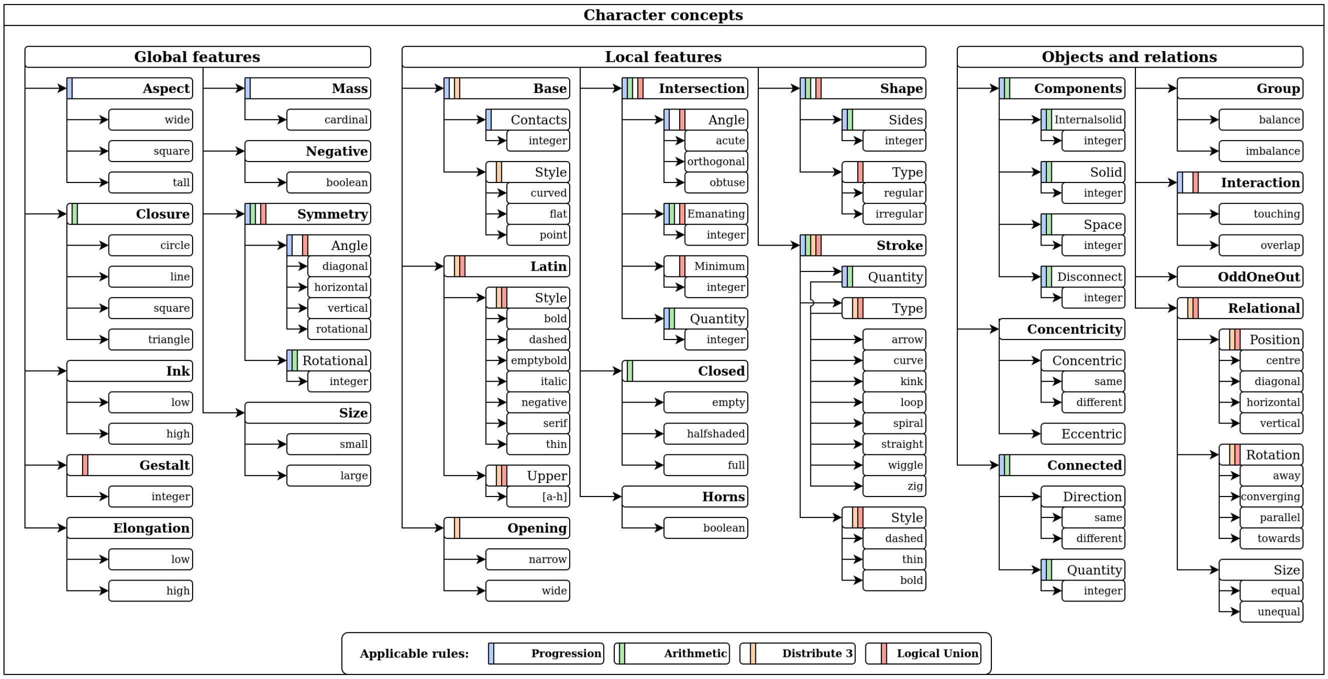

Figure 2. The conceptual schema of *Unicode Analogies 1.0*, at the time of release. *Constant* problems can be generated using any concept, while the other four rules are applicable to subsets of concepts, as labelled by bars. Concepts appear at multiple levels of hierarchy, beginning with grouping global, local, and relational concepts, becoming finer-grained and specifying the values applicable to each.

3.2. Annotation and extensibility

Upon establishing a schema, 4000 characters were preselected from sections of Unicode that feature largely symbolic characters.[1] Qualitative suitability criteria include the simplicity and abstractness of characters as an estimate of their amenability to PMPs. Pragmatic criteria include the availability and copyright of fonts to render chosen characters. Manual annotation was then performed by stepping through concepts in the schema and selecting character images for which these concepts were readily perceived. All selections and annotations were made by the first author. This resulted in a final set of over 2500 annotated characters, each possessing 2.8 annotated features on average, with the most polysemic character featuring 20 annotations.

3.3. Problem structure and generation

To more directly establish the task as analogy-making, while making efficient use of human annotations, the structure of PMPs in UA differ slightly to those found in RAVEN and PGM, consisting of two rows of context, for a total of nine frames per problem (five context, four answers). This resembles the Visual Analogy format introduced in [14]. PMPs do not exhibit multiple rules, instead, each follows a single rule-concept pair, as the goal is to encourage solvers to use context at the perceptual level. Each frame consists of a single Unicode character rendered at 80x80 binary pixels, which is a resolution common to most PMP solvers.

Generating a new dataset split involves random sampling of the problem space. For each problem requested, a tuple is sampled with the structure *rule-concept-shift* (e.g. *constant-shapesides-noshift*). *Context shift* refers to whether or not both context rows will present the same concept values. For example, a problem that instantiates the *progression* rule over the quantity of dots may do so as two rows depicting 'three, two, one' dots. Requesting context shift would mean the second row altering the progression, e.g. 'one, three, five'. Shifted problems are expected to be more difficult as there are less context frames to evidence the rule.

Upon sampling each problem tuple, context frames are selected to instantiate that tuple in two rows. The last context frame is popped from the list and added to the answers, alongside three foils, all annotated as belonging to the parent concept in the problem tuple. Foils do not depict the concept in a way that would complete the intended rule and invalidate the problem. Additionally, foils cannot be drawn from the problem context, nor can they complete an emergent rule (that is, an unintended but valid alternative rule) as far as the system can infer from character annotations. Finally, the pool of potential foils for any given problem is balanced by only accepting a maximum of three instances of each candidate concept, to ensure that over-represented values (i.e. concepts that apply to relatively large numbers of annotated characters) do not dominate answer distributions and introduce an exploitable bias.

[1] The full list of Unicode sections appears in supplementary material.

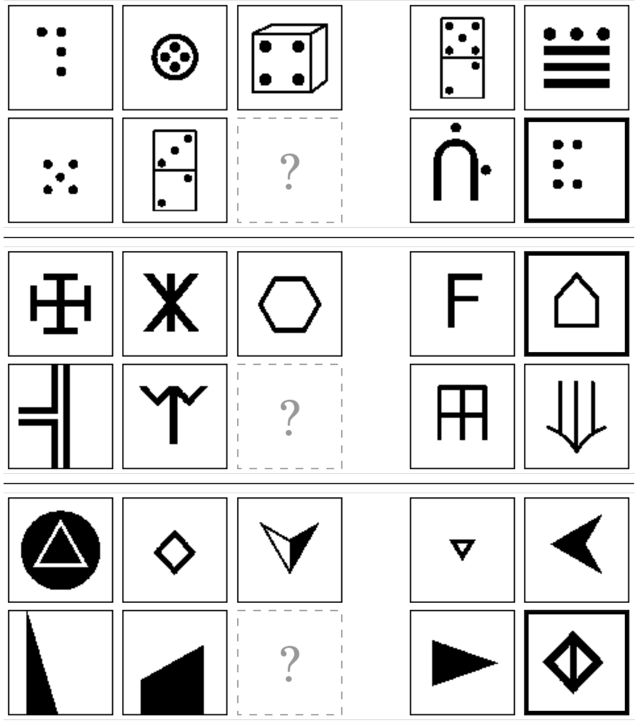

Figure 3. Three example PMPs from the dataset. The top problem demonstrates a constant number of dots in each row. The middle problem demonstrates arithmetic over the meeting points of lines. The lower problem demonstrates a union; the first two frames depict 3-sided and 4-sided shapes, while the final frames depict both. Correct answers are emboldened.

By selecting answers and foils depicting the parent concept, diverse problems can be assembled from orders of magnitude less annotation than naive strategies (i.e. where each image is checked and annotated for each and every feature present in the schema). Answers are also challenging, because the problem rule is guaranteed to be applicable to all candidates, with only one instantiating its concept correctly. Due to the nature of manual annotation, PMPs cannot be guaranteed to always be a) valid (solvable by precisely one candidate answer) or b) human-intuitive in their assembly. Nonetheless, we experimentally confirm that the dataset splits produced by this process remain largely human solvable whilst maintaining a significant challenge for machine solvers, and thereby motivate UA as having utility in exploring this performance gap.

3.4. Parameters for defining splits

A the level of defining a single problem, a tuple specifies the rule, concept, and context shift to be instantiated in the problem frames. At the level of defining a dataset split, there are additional parameters that invite experimentation:

- **Rule sampling**. Defines the subset of rules to be made available when sampling problems.

- **Tuple extrapolation.** Defines which problem tuples to hold out in testing. ***Neutral*** samples all available *rule-class* tuples in both train and test sets. ***Extrapolation*** ensures exclusive tuples across sets. ***Extrapolation-plus*** holds out entire concepts.

- **Context shift.** If true, the generated split will include both shifted and non-shifted problems.

- **Character holdout.** Defines how the character set is split into train and test sets. All problems are generated from these respective sets. If ***None***, all characters are equally available for constructing training and testing problems, but *rule-class-value* tuples are disjoint across splits, to avoid exposing the model to test problems during training. ***Set difference*** holdout ensures train and test character sets are disjoint, requiring models to extract robust features if they are to perform well.

- **Sampling diversity.** ***Balanced*** samples problems uniformly, allowing for equal representation. ***Diverse*** disallows re-instantiating a problem tuple with the same answer to maximise problem diversity and minimise opportunities for memorisation. All experiments in the paper use diverse sampling.

4. Experiments

We ran five key experiments – *Rule*, *Schema*, *Extrapolation*, *Challenge*, and *Hold-out* – designed to deepen an understanding and appreciation of models' visual conceptualisation abilities. Here, we detail the architectures, dataset splits, and evaluation methods used to facilitate this goal. We also describe the acquisition of a human baseline.

4.1. Architectures

We selected three high-performing architectures from the RAVEN literature to fit to each of the dataset splits: the Multi-scale Relation Network [2], the Scattering Compositional Learner [38], and Rel-Base [35]. There exists solvers possessing further structural knowledge for objects and rules [27, 35, 40], obtaining more logical, transparent, and generalisable reasoning at the cost of being bound by strong prior knowledge when asked to correctly parse scenes possessing diverse, overlapping, compound, or gestalt features. We limited experimentation to solvers without such inductive biases, but note that architectures such as the Transformer-based STSN [27] (contemporaneous to this publication) should be investigated further.

To enable modelling the new two-row PMP format by solvers built for three rows, dataset loaders pad problems with empty frames. We noticed no difference in accuracy when comparing this strategy to adjusting architecture input layers. We also selected two baselines as implemented by Spratley et. al [35]: ResNet, and its context-blind variant, used to check for exploitable biases in answer set generation by only viewing answer frames. Instead of treating the blind model as merely a sanity check during development, we subject it to the same tests as all other models to be aware of bias across different splits.

4.2. Method and dataset splits used

1. *Rule.* Model versus human performance is explored across all five rule types offered by the dataset. Models are trained and tested on individual rules, as well as jointly trained on all rules, providing average performance. Parameters for defining these splits are set to defaults: Extrapolation is *neutral*, context shift is disabled, and character holdout is *set difference*.

2. *Schema.* Performance is then explored across the three schema subcategories – *Global*, *Local*, and *Objects & Relations* (Fig. 2) – to provide further understanding of which problem themes were more or less challenging. Parameters for these splits are set to defaults.

3. *Extrapolation.* Models are tested against four extrapolation splits starting with *No Shift* (context-shift disabled), and increasing in difficulty. *Neutral* enables context-shift, while *Extrapolation* and *Extrapolation-plus* also alter the tuple extrapolation parameter to provide finer-grained generalisation results with which to judge the limits of models' extrapolative abilities.

4. *Challenge.* Both easy and challenging concepts are summarised based on a comparison between human and model performance, and the resulting experimentally-informed challenge split is used to further probe the disparity between human and model performance. Parameters are set to defaults.

5. *Hold-out.* The influence of both character hold-out strategies is briefly examined using two splits based on *Constant* rules, in order to exacerbate the effects of non-robust feature learning.

Both model parameter initialisation and dataset seeds needed to be accounted for; the former affects traversal of the optimisation landscape, while the latter affects the distribution of problems across train-test sets, and may prohibit entire problem types from forming across sets in the process of randomly holding out images. To achieve a more robust understanding of model ability, we performed 5-fold cross-validation. Crucially, because dataset splits from this framework aren't amenable to being shuffled and repartitioned (without violating character and tuple holdout), for each dataset split, we generated each fold with random seeds, and trained three randomly-initialised models on each. The size of dataset folds is similar to RAVEN [39], containing 8,000 - 10,000 problems each. All models were trained to a maximum number of epochs given their architecture type, found by preliminary fitting of each model on the *Average* rule set. All materials, including splits, their folds, seeds and other parameters, are made available on our project page.

4.3. Establishing a human baseline

In keeping with RAVEN and Bongard-LOGO, we established a baseline of human performance over a set of representative problems to better direct model development. We employed 30 subjects using the Prolific.co research platform, who were remunerated at a rate consistent with the minimum wage in our country. The two selection criteria required subjects to hold a graduate degree, and to form a gender-balanced sample. Subjects were presented with short instructions as to the structure of problems they would encounter, and familiarised with an initial set of presolved problems, sampled from the train set. They were then instructed to complete a set of 15 problems, sampled randomly from the test set. To obtain these sets, we first generated one fold with a 50-50 train-test split. Across human subjects, at least one problem per potential *rule-concept* tuple was shown. Our experiment was designed using PsychoPy [30] and hosted on Pavlovia.org.

4.4. Experimentally informing a new challenge split

By sorting the list of problem types in the fold used for the human baseline, in order of performance difference (human accuracy minus model accuracy), and retaining the top 50%, we establish a *Challenge* split to help guide the development of perception models towards human-like analogical reasoning. In doing so, there is diagnostic potential to uncover weak spots in vision systems and indicate which inductive biases might be necessary to engineer. This also increases the quality of problems, assuming that human accuracy is representative of problem intuitiveness.

5. Performance analysis

In this section, we present and analyse the outcomes of our five key experiments. Results reported as accuracy (%).

Rule (Tab. 1). Across the different rule sets, we notice that model accuracy is almost universally below 35%, while humans are still above the top solver on the *Average* (joint) set by 24.4%. This difference is increased to 36.5% over the *Progression* set, which is hypothesised to be due to humans excelling at counting objects (a weakness of deep neural networks [13]). This hypothesis is evidenced by the results of the following experiments, performing more fine-grained analysis on concept types. *Union* problems appear unintuitive for humans, but we believe that this performance could be improved with other experimental designs as the different possible rules were not comprehensively described to participants. While this wouldn't invalidate this data (participants would still be required to perform fluid perception over unseen concepts and characters), it was obtained to serve as a baseline, not a goal to beat. The 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, suggesting that overall this dataset succeeds in presenting a significant challenge, while our answer set sampling strategy mitigates exploitable bias.

Schema (Tab. 2). Of the three schema subcategories, the most accurately modelled was *Global*, likely due to concepts such as *ink amount* and *global size* being less abstract and more amenable to feature extraction. Meanwhile, *Object and Relations* saw the human baseline double the leading model's accuracy. Such problems seem to be much harder for machine solvers due to their concepts being abstract and able to be instantiated on a large variety of object types. Unlike many *Global* concepts, which may be partially solvable by pixel counting (e.g. pixels near image borders might be correlated with *global size*), it is unclear how a network might acquire the features to robustly perceive this.

Extrapolation (Tab. 3). Moving from *No Shift* to *Extrapolation-plus*, we observe a general trend of performance loss across all solvers as expected. With stronger future models, we expect this discrepancy to become even more apparent, as these datasets progressively prohibit memorisation and require solvers to extrapolate learned concepts to increasingly OOD problems.

Challenge (Tab. 4). To our knowledge, the *Challenge* split presents the highest discrepancy between human and machine performance of any PMP dataset in this area, with the leading model trailing 40.2% behind, and most models displaying near-random performance. As clued in by the *Schema* experiment, we continue to notice that the most successfully modelled concepts belong to the *Global* category. Humans perform very well in counting local features, while models were largely unable to do so.

To further explore how concepts were perceived in problems, Tab. 5 presents the set unions of concepts deemed relatively 'easy' and 'hard', for both human and machine solvers. To obtain these, we first ordered all concepts in the schema by the average accuracy of problems in which they are featured, and then retained the concepts outside the interquartile range. From this, we can see that *Global* problems are often easier for all solvers, while perceiving empty spaces as objects is unintuitive. Not surprisingly, humans perform very well on problems that explore both global and relational object size, whereas models only succeed at global size, further suggesting that their comprehension of size as an abstract concept is limited.

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

Table 1. Human vs. model performance across rule types. Average (avg) performance is over the combined test set and is therefore weighted to rule types that have more available concepts.

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

Table 2. Human vs. model performance across schema categories.

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**

Table 3. Extrapolation performance on datasets with all rules.

Hold-out (Tab. 6). Comparing models on both *Constant* rule splits – one with character hold-out, and one without – we notice a significant performance increase in some models when the same character set is used to assemble both train and test problems, despite *rule-class-value* tuples being disjoint across splits. This strongly suggests that the use of disjoint character sets is an important design consideration for this framework, and had datasets been constructed without this, we might have critically overestimated model abilities.

A consideration worth mentioning for reproducibility is that models were prone to overfitting, which was partially alleviated by enabling dropout. Given our compute resources, we prioritised *k*-fold cross-validation with maximum epochs to give a useful first pass of contemporary PMP architectures on this dataset. With hyperparameter tweaking and more nuanced regularisation, along with training schemes such as early stopping and best model selection, additional performance might be achieved.

We leave tailoring and developing models to future work. 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. We believe this observation implies that across models, the same kinds of non-robust features are being extracted, and further motivates the UA challenge by inviting a new class of solvers.

6. Broader Impact and Future Work

While this framework is intended to be of primary use for supervised learning techniques in abstract visual reasoning, it is easily extended to new concepts and annotations, inviting future work in artificial intelligence and cognitive science. Investigating the impact of controlling features such as domain shift, distractors and misleading factors, is likely to be of interest in testing models of human concept discovery and category learning. There is also the option to run more targeted generalisation experiments: one could test for model numeracy by generating a split with all numeric concepts, but train and test exclusively on arithmetic and progression problems, respectively. Alternatively, one might want to train on local feature concepts, and test for extrapolation to global features. Or, one might implement and test schemas of their own. The released code performs all experiments automatically, given a user-defined schema.

Since we have chosen the concepts used in problem formation, and possess algorithms that are capable of extracting many of these concepts, there is potential to perform more direct probing of concept acquisition in trained models, using methods such as those introduced by [26]. To our knowledge, this has not yet been performed in this area.

Finally, this framework can be adapted for use across different learning paradigms, including meta and unsupervised learning. For example, the Omniglot dataset [19] presents a challenge for few-shot methods aiming to cluster handwritten characters. The problems in Unicode Analogies are generated from an underlying set of annotated polysemic characters, which might pose its own challenge to such methods.

7. Conclusion

Of the abstraction datasets that aren't focused on gestalt perception (including all based on Raven's Progressive Matrices), the implication for solvers is that there is a singularly correct way to parse a scene. We argue that testing for analogical reasoning needs to incorporate fine-grained and concept-based analysis, over datasets built to expose non-robust feature learning. We introduce 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. In doing so, we demonstrate that state-of-the-art solvers are still far from achieving the founding goals their datasets were created for, and encourage new solvers to overcome these limitations. We are excited to see how this framework is adopted by our research community.

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

Table 4. Breakdown of performance on the *Challenge* split, including a summary of the top and bottom concepts that experimentally informed this split, for human participants and models.

Concepts	Model (RelBase)	
Human	>Q3, 'easy'	<Q1, 'hard'

>Q3, 'easy'	`latin-style,`
	`negative,`
	`global-size,`
	`horns,`
	`dash-quantity`
		`arrow-quantity,`
		`relational-size`

<Q1, 'hard'	`opening,`
	`closure`
		`space,`
		`interaction,`
		`uniquesolid`

Table 5. Two-by-two table depicting set unions of concepts. These concepts feature in problem types outside the interquartile ranges of human and model performance.

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

Table 6. Performance on two *Constant* rule splits, generated with character hold-out set to *None* and *Set difference* (Diff.).

References

[1] David Barrett, Felix Hill, Adam Santoro, Ari Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. In *International conference on machine learning*, pages 511–520. PMLR, 2018. 2, 3

[2] Yaniv Benny, Niv Pekar, and Lior Wolf. Scale-localized abstract reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12557–12565, 2021. 5

[3] M. M. Bongard. *Pattern Recognition*. New York, Spartan Books, 1970. 1, 3

[4] Daniel Casasanto. All concepts are ad hoc concepts. In *The conceptual mind: New directions in the study of the concepts*, pages 543–566. MIT press, 2015. 2

[5] David J Chalmers, Robert M French, and Douglas R Hofstadter. High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental & Theoretical Artificial Intelligence*, 4(3):185–211, 1992. 1, 3

[6] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019. 3

[7] Martin Engelcke, Adam R Kosiorek, Oiwi Parker Jones, and Ingmar Posner. Genesis: Generative scene inference and sampling with object-centric latent representations. *arXiv preprint arXiv:1907.13052*, 2019. 1

[8] Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138, 2010. 1

[9] Karl Friston. The history of the future of the bayesian brain. *NeuroImage*, 62(2):1230–1233, 2012. 1

[10] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 2

[11] Dedre Gentner, Keith J Holyoak, and Boicho N Kokinov. *The analogical mind: Perspectives from cognitive science*. MIT press, 2001. 2

[12] Mary L Gick and Keith J Holyoak. Analogical problem solving. *Cognitive psychology*, 12(3):306–355, 1980. 2

[13] Shuyue Guan and Murray Loew. Understanding the ability of deep neural networks to count connected components in images. In *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–7. IEEE, 2020. 6

[14] Felix Hill, Adam Santoro, David GT Barrett, Ari S Morcos, and Timothy Lillicrap. Learning to make analogies by contrasting abstract relational structure. *arXiv preprint arXiv:1902.00120*, 2019. 3, 4

[15] Douglas R Hofstadter. Analogy as the core of cognition. *The analogical mind: Perspectives from cognitive science*, pages 499–538, 2001. 1, 2

[16] Douglas R Hofstadter, Melanie Mitchell, and Robert Matthew French. *Fluid concepts and creative analogies: A theory and its computer implementation*. University of Michigan, Cognitive Science and Machine Intelligence Laboratory, 1987. 2

[17] Andreas Holzinger, Michael Kickmeier-Rust, and Heimo Müller. Kandinsky patterns as iq-test for machine learning. In *International cross-domain conference for machine learning and knowledge extraction*, pages 1–14. Springer, 2019. 3

[18] Jack Koch, Lauro Langosco, Jacob Pfau, James Le, and Lee Sharkey. Objective robustness in deep reinforcement learning. *arXiv preprint arXiv:2105.14111*, 2021. 2

[19] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. The omniglot challenge: a 3-year progress report. *Current Opinion in Behavioral Sciences*, 29:97–104, 2019. 8

[20] Alexandre Linhares. A glimpse at the metaphysics of bongard problems. *Artificial Intelligence*, 121(1-2):251–270, 2000. 1

[21] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33:11525–11538, 2020. 1

[22] Norman RF Maier. Reasoning in humans. ii. the solution of a problem and its appearance in consciousness. *Journal of comparative Psychology*, 12(2):181, 1931. 1

[23] Mikołaj Małkiński and Jacek Mańdziuk. Deep learning methods for abstract visual reasoning: A survey on raven's progressive matrices. *arXiv preprint arXiv:2201.12382*, 2022. 1, 2, 3

[24] Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018. 1, 2

[25] P Mayring. Qualitative content analysis philipp mayring 3. basic ideas of content analysis. In *Forum Qualitative Sozialforschung*, volume 1, 2000. 3

[26] Thomas McGrath, Andrei Kapishnikov, Nenad Tomašev, Adam Pearce, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik. Acquisition of chess knowledge in alphazero. *arXiv preprint arXiv:2111.09259*, 2021. 8

[27] Shanka Subhra Mondal, Taylor Webb, and Jonathan D Cohen. Learning to reason over visual objects. *arXiv preprint arXiv:2303.02260*, 2023. 5

[28] Weili Nie, Zhiding Yu, Lei Mao, Ankit B Patel, Yuke Zhu, and Anima Anandkumar. Bongard-logo: A new benchmark for human-level concept learning and reasoning. *Advances in Neural Information Processing Systems*, 33:16468–16480, 2020. 3

[29] Victor Vikram Odouard and Melanie Mitchell. Evaluating understanding on conceptual abstraction benchmarks. *arXiv preprint arXiv:2206.14187*, 2022. 2, 3

[30] Jonathan W Peirce. Psychopy—psychophysics software in python. *Journal of neuroscience methods*, 162(1-2):8–13, 2007. 6

[31] Yonggang Qi, Kai Zhang, Aneeshan Sain, and Yi-Zhe Song. Pqa: Perceptual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12056–12064, 2021. 3

[32] John Raven. The raven's progressive matrices: change and stability over culture and time. *Cognitive psychology*, 41(1):1–48, 2000. 2

[33] John C Raven and JH Court. *Raven's progressive matrices*. Western Psychological Services Los Angeles, CA, 1938. 2

[34] Henry L Roediger and Kurt A DeSoto. Psychology of reconstructive memory. *International encyclopedia of the social & behavioral sciences*, 20(2):50–55, 2015. 2

[35] Steven Spratley, Krista Ehinger, and Tim Miller. A closer look at generalisation in raven. In *European Conference on Computer Vision*, pages 601–616. Springer, 2020. 1, 2, 3, 5, 6

[36] Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13:12, 2019. 1

[37] Dorothy Walsh. Occam's razor: A principle of intellectual elegance. *American Philosophical Quarterly*, 16(3):241–244, 1979. 2

[38] Yuhuai Wu, Honghua Dong, Roger Grosse, and Jimmy Ba. The scattering compositional learner: Discovering objects, attributes, relationships in analogical reasoning. *arXiv preprint arXiv:2007.04212*, 2020. 5

[39] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5317–5327, 2019. 2, 3, 6

[40] Chi Zhang, Baoxiong Jia, Song-Chun Zhu, and Yixin Zhu. Abstract spatial-temporal reasoning via probabilistic abduction and execution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9736–9746, 2021. 5

[41] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019. 1