Can you even tell left from right? Presenting a new challenge for VQA

Sai Raam Venkataraman Rishi Sridhar Rao S. Balasubramanian
R. Raghunatha Sarma
Sri Sathya Sai Institute of Higher Learning, Prasanthi Nilayam, Andhra Pradesh
sairaamvenkataraman@gmail.com, rishisridhar96@gmail.com

Chandra Sekhar Vorugunti
Samsung Research India
chandrasekhar.v@iiits.in

Abstract

Visual Question Answering (VQA) needs a means of evaluating the strengths and weaknesses of models. One aspect of such an evaluation is the measurement of compositional generalisation. This relates to the ability of a model to answer well on scenes whose compositions are different from those of scenes in the training dataset. In this work, we present several quantitative measures of compositional separation and find that popular datasets for VQA are not good compositional evaluators. To solve this, we present Uncommon Objects in Unseen Configurations (UOUC), a synthetic dataset for VQA. UOUC is at once fairly complex while also being compositionally well-separated. The object-class of UOUC consists of 380 clasess taken from 528 characters from the Dungeons and Dragons game. The training dataset of UOUC consists of 200,000 scenes; whereas the test set consists of 30,000 scenes. In order to study compositional generalisation, simple reasoning and memorisation, each scene of UOUC is annotated with up to 10 novel questions. These deal with spatial relationships, hypothetical changes to scenes, counting, comparison, memorisation and memory-based reasoning. In total, UOUC presents over 2 million questions. Our evaluation of recent state-of-the-art models for VQA shows that they exhibit poor compositional generalisation, and comparatively lower ability towards simple reasoning. These results suggest that UOUC could lead to advances in research by being a strong benchmark for VQA, especially in the study of compositional generalisation.

1. Introduction

The field of Visual Question Answering (VQA) deals with the development of machine learning models that can understand visual scenes and answer questions about them. These questions can deal with the properties of objects present in the scenes or with the relations among them. Evaluating models involves measuring their ability to answer questions correctly. Two aspects of this are: a model's ability to handle complex scenes, and its ability to generalise well to scenes whose compositions differ from the scenes it was trained on. These, we refer to as *expressivity* and *compositionality*.

Why are these two important? Expressivity is important, as useful data is often complex. Compositionality is important for concept learning. In VQA, concepts are said to be learnt if they can be identified regardless of the scenes or objects they are associated with. To explain this, consider that the concept of 'front of' is independent of objects that are in front other objects. This is true for concepts such as relationships and properties of objects, which are understood across a variety of object instances. This learning, termed compositionality, can be evaluated effectively by having compositionally separated training and testing datasets.

A dataset exhibiting the first property above must have at least a fair diversity in object-complexity. The second property necessitates that the train and test scenes have a minimal overlap in composition, either in terms of object pairs or in terms relationship tuples. This is so that the learning of concepts can be evaluated outside of instances seen in training.

Datasets that are based on natural scenes are significantly complex. Examples of these are VQA-v1 [4], VQA-v2 [9], Visual Genome [17] and GQA [13]. They, however, do not have a mechanism for compositional separation. Moreover, the presence of natural biases in the co-occurrence structures of objects could lead to the presence of compositional similarities in the training and test datasets.

On the contrary, a dataset such as CLEVR [14] offers a structured generation of scenes, where certain aspects of the

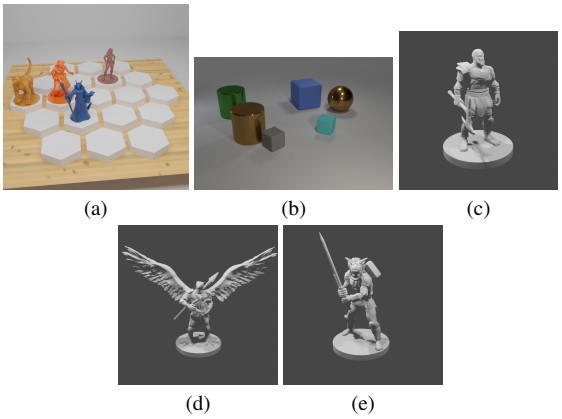


Figure 1. Fig. 1a is an example scene in the test set of UOUC. Every pair of objects in it are never seen in the training dataset - this makes for a strong evaluation of relationships such as spatial ones. Fig. 1b is an example scene in CLEVR. Note that the complexity of objects is low, due to them being simple geometric objects of only three classes. In contrast, consider the objects in Fig. 1a, or in Fig. 1c, 1d and 1e. These possess greater complexity. This, combined with compositional separation, makes for a challenge to models.

train scenes differ from the test scenes. Specifically, the CoGenT dataset of CLEVR presents a colouring-based compositional separation, where certain objects are coloured differently in the training and testing datasets. However, CLEVR suffers from the fact that its object-set is small and simple - consisting of cylinders, cubes and spheres. Each of these objects have only three other properties, namely colour, material and size. Thus, despite the generated compositional difference of colouration, the training and testing sets of CoGenT have several similarities. Another weakness is that due to the simplicity of the object-classes, models may easily learn to recognise objects and thus exhibit high performance.

A major source of this compositional similarity in the training and testing datasets of not only CLEVR, but also other VQA datasets, is the presence of common co-occurrences of object-pairs. In other words, several object-pairs occur together in scenes of both training and testing data. The presence of this common information can make answering questions easier for models, while obscuring the real extent of compositional generalisation.

As a means of solving this problem, we present Uncommon Objects in Unseen Configurations (UOUC). UOUC is significantly more complex than CLEVR. For a rough comparison, the object-class of UOUC is 380 in number with a greater range of appearances as can be seen in contrasting Figure 1b with Figure 1a and the example objects in Figures 1c, 1d and 1e. More scenes are present in the supplementary material of our paper.

UOUC is designed so that no pair of objects co-occur in both the training and testing sets. Thus, compositional separation is based on co-occurrence, a stronger condition than colouration. In order to use this separation condition effectively for VQA, each scene of UOUC is annotated with up to 10 questions. The first four of these deal with the learning of spatial relationships between objects, namely front, back, left, and right. Since no objects co-occur commonly, a model must learn to understand the notion of a spatial relationship independent of the instances it has seen during training. The next two questions deal with simple reasoning such as counting and comparison. The next question queries the presence of an object in a scene. The last three questions deal with memorisation and memory-based reasoning. Memorisation related to non-perceptual attributes is an under-explored area of VQA. Specifically, we wish to make an initial direction in understanding the complexity of memory-based reasoning, in comparision to perception-based reasoning. UOUC thus consists of 200,000 questions in the training dataset, and 30,000 questions in the test set, with over 2 million questions for all the scenes.

Our experiments using recent state-of-the-art models show that UOUC is a strong challenge for compositional reasoning for VQA, indicating its use for future research while also uncovering weaknesses in current VQA.

2. Related Work

There is much work in the field of VQA involving contributions in terms of datasets and models. Several models have been proposed and evaluated on various datasets. MUTAN [6] is a model that combines visual and question features in an efficient manner. Likewise, [16] presented a model that combines features from CNNs, attention, and a sequence-model to again achieve good performance. Both of these achieve good performance on VQA-v1 and VQA-v2. LCGN [11] is another model that uses a language-conditioned graph network with iterative message-passing. MAC is a compositional model for VQA that uses memory and control integrated in a single unit. LCGN and MAC achieve good accuracies on CLEVR and GQA. Another model in the compositional VQA literature is MCAN [27]. MCAN, like MAC and LCGN, continues to be cited as a baseline in the VQA literature, due to its good performance on VQA-v2 and GQA.

More recent models that use newer and more accurate models include AoA [22], TRAR [29], RWSAN [21], and LSAT [23]. AoA presents a attention mechanism that allows for good performance on the VQA-v1 dataset. TRAR uses a routing mechanism within the visual transformer layers to achieve high performance on VQA-v2 and CLEVR. LSAT uses grid features along with attention on inter and intra-windows to capture global and contextual information that allows for high performance on VQA-v2 and CLEVR.

RWSAN presents a light, but high-performing, model for VQA using shared weights between image and text to capture image-question interactions, obtaining state-of-the-art performance on VQA-v2, CLEVR, and GQA.

Datasets for VQA that many models for VQA use are VQA-v1, VQA-v2, GQA, Visual Genome, and CLEVR. VQA-v1 was one of the early datasets for VQA. It uses images from the COCO dataset [7], and provides annotations in the form of questions. VQA-v2 is a balanced version of this dataset that associates each question in VQA-v1 with two images that have a different answer. Visual Genome is a highly-annotated dataset of real images for VQA, with a huge object-class. GQA is a dataset of real scenes that annotates images with questions that involve multiple steps of reasoning. The answer distribution of the questions has been balanced. Moreover, GQA also introduces metrics for evaluating models that move beyond accuracy. CLEVR is a generated dataset for VQA that uses questions that are compositional in nature, and can use multiple steps of reasoning. CLEVR offers a dataset termed CoGenT that uses a separation condition, in terms of object-properties, in its training and testing set. Compositional models that understand properties independent of objects would be able to answer questions that relate to the properties of this separation condition.

3. How does UOUC compare to popular datasets

In this section, we present a comparison of UOUC with several popular datasets for VQA. Our comparison, like our motivation for UOUC, is on the two fronts of complexity and compositional separation.

3.1. Comparing the richness and complexity

We aim to provide a measure of complexity, or richness, by stating the number of questions, objects, and scenes in the datasets. Table 1 provides this information. As can be seen, UOUC has an object set much larger than CLEVR-CoGenT. Moreover, and as mentioned before, based on a comparison between objects in Figure 1a and Figure 1b, we can safely say that UOUC has more complex scenes. UOUC is also quite comparable in terms of the number of scenes, the size of the object-class and number of questions to datasets such as VQA-v1 and VQA-v2. This establishes UOUC as a perceptually harder compositionality evaluator for VQA than CLEVR.

3.2. Comparing compositional separation

As stated before, two measures of compositional separation between the train and the test data are the number of common co-occurring pairs and common co-occuring relations among them. We use these to present four metrics that,

Table 1. The number of images, objects, questions for some datasets (approximate).

Name	Images	Questions	Classes
VQA-v1 (Real)	205K	614K	80
VQA-v2 (Real)	205K	1.1M	80
Visual Genome	108K	1.7M	33,877
GQA	113K	22M	1703
CLEVR-CoGenT	130K	1.3M	3
UOUC	**230K**	**2M**	**380**

Table 2. Measure of compositional separation between the training and testing datasets. Lower indicates better separation. Note that UOUC is the best separated among all the datasets, making it an ideal dataset for evaluating compositional generalisation.

Dataset	VG	GQA	**UOUC**	CGnT	VQA
AvgCoPair	269.93	155.95	**0.00**	20.89	16.88
AvgCoPairOcc	2,270.14	4,023.83	**0.00**	318K	5942.41
AvgCoRel	6.20	45.79	**0.00**	-	11.48
AvgCoRelOcc	21.45	50.01	**0.00**	-	4132.25

when close to zero, indicate a high degree of compositional separation.

3.2.1 Common co-occurrences

The first of these, termed *AvgCoPair*, is the average number of co-occurring pairs of objects in a test scene, that have co-occurred in a train scene. Complementing this is a second score, termed *AvgCoPairOcc*, which gives the average number of times a commonly co-occurring pair is found in the training dataset.

The first measure gives an extent of common information, per test image, present between the training and testing dataset. The second measure gives the extent of the presence of this common information in the training dataset. For example, if two objects - a child and a cake, were to be present in the same scene for some images in both the train the and test datasets, there is some compositional overlap. This compositional overlap can lead to models obtaining higher performance because they have seen these together in the train dataset. Moreover, as mentioned before, this also leads to a difficulty in measuring compositional generalisation for relationships that may have existed between these. If multiple such instances of such co-occurrences are found in the train dataset, then it becomes easier for a model to memorise this information and then answer questions on the test dataset for the common pairs.

3.2.2 Common relationships

The next two measures, termed AvgCoRel and AvgCoRelOcc, extend this concept to relationships between objects instead of only co-occurrences. Thus, AvgCoRel gives the

average number (per test scene) of (subject, object, relation) tuples that occur in the same image in both the training and testing datsets. AvgCoRelOcc gives the average number of the occurrences of such common tuples in the train dataset.

The presence of the same relationships between the same objects in both the training and testing dataset is a stronger case of compositional overlap. Further a larger number of occurrences of any such overlap in the train dataset makes their learning easier for models, and thus impacts a proper evaluation.

Using the same example of a child and a cake, if both the train and the test data have common instances of the relation 'eating' between them, there is a strong compositional overlap. If multiple instances of a child eating cake exist in the train dataset, then a model can find it easy to use this in answering questions on the test dataset.

Thus, we consider both co-occurrence and relational overlap in our metrics of compositionality. Table 2 gives these four metrics for Visual Genome, GQA, UOUC, and CLEVR. We use the mean of 5 random 70-30 train-test splits for Visual Genome to obtain these results, in the manner of [17]. We use the validation set as the test set for GQA, as required information about the test set is not provided publicly. As CLEVR uses no annotated relationships between objects, we compute only AvgCoPair and AvgCoPairOcc. Lastly, we provide an approximate score for VQA-v2 using scene graphs generated by . We include VQA-v2 only as VQA-v1 is hardly used in the recent literature, and VQA-v2 has a high overlap with VQA-v1.

First, we see that Visual Genome, GQA, VQA-v2, and CLEVR have non-zero scores for the computed measures. Thus, there is a high level of compositional overlap between the train and test splits. UOUC has all the scores zero, as no common co-occurring pairs or relationships exist. This is by design. Thus, UOUC offers a stronger evaluation of compositional generalisation for models.

We emphasize a point, here, that has guided our construction of UOUC. If the train and test datasets are separated by co-occurrence then they are naturally separated by relations as well. This is because no common subject-object pairs are possible. UOUC, by its construction, has such a separation.

3.3. Comparison by performance

A final comparison of datasets is based on the level of challenge it offers to recent models. Such a comparison can be based on the accuracy of answering questions. Based on Table 4, we see that models that achieve decent performance on VQA-v2 and GQA, and more importantly high performance on CLEVR, perform poorly on UOUC, especially on compositional generalisation. This allows us to state that UOUC can lead to further advances for compositional generalisation in VQA.

4. Details of UOUC

UOUC, like CLEVR, is synthetically generated so as to allow for a compositional separation of the train and test datasets. We outline the process of construction.

Downloading and pre-processing object models: All the objects were downloaded from https://www.prusaprinters.org/social/39782-mz4250/prints as .STL files. They were pro-processed to make the scales uniform, and the the 3d models of the objects face the viewer. In total, 528 models were used to create 380 classes of objects. The objects are used with the permission of the creator and are licensed under a Creative Commons License 4.0 (Non-Commercial) license.

Categorisation and grouping of objects: The 380 classes were further categorised into 5 categories. These categories are presented in bold in Table 3. The categorisation was done manually. After categorisation, the objects are again grouped randomly into 10 groups. The process is such that each group has roughly similar numbers of objects of each category. This grouping does not introduce any new properties to objects, but is extremely important to the compositional separation of the train and test datasets.

Construction of the base scene: Since the scenes are to be generated using the 3d models of the objects, we use the Blender [8] software to first create 3d scenes which are then rendered to 2d scenes. A requisite for this is a 3d base scene. Figure 2b shows the scene we used. The scene is made so that the objects are placed in the 19 hexagonal structures.

Construction of the scene-structures: Before generating the scenes of the train and test datasets, we first generate text-files that store their structures. The structure of a scene consists of the positions of each of the objects in the scene, the colour of the objects, their rotation about the z-axis, and the camera-rotation. For the training dataset, objects are rotated with an angle chosen randomly between -30 and 30 degrees; the camera is rotated randomly with an angle chosen between -20 and 20 degrees. For the test set, these extents are -45 and 45 degrees for the objects, and -30 and 30 degrees for the camera. These are stored in the text-files, which are then used for the generation of the scenes and the questions and answers. The grouping mentioned earlier is used here. For the training dataset, scenes are generated so that only objects within a group co-occur, while the test set has that objects only from different groups co-occur. The choice of the objects is otherwise random, with a scene having at least 4 objects, and at most 6. Thus, the train and test datasets compositionally-separated. Each group generates 20,000 scenes, and the test set contains 30,000 scenes.

Generation of the scenes: The scenes of UOUC are generated using the scene-structures and the blender software. Based on the scene-structure objects are assigned a colour from one of orange, yellow, green, violet, brown, and light-yellow. The objects are then rotated and positioned ac-

cording to the scene-structure for that scene.

Generation of the questions and answers: The text-files find another use in the generation of the questions and answers for the scenes. Each scene of UOUC has up to 10 questions associated with it. The procedure for the generation of questions and answers is based on using predefined templates for each question and filling-in these templates using the logic for that question, and using certain properties of objects. These properties, apart from colour, are mentioned in Table 3 for each category. The properties are all categorical, and can assume more than 2 values across the objects of a category. The assignment of these properties is based on the role of the object-classes in literature and media. This was done manually. The reader may note that no prior knowledge of Dungeons and Dragons is necessary for using UOUC. Apart from these properties, each object is also assigned a random team, which is one of 'A', 'B', and 'C'. The purpose of this attribute is test the memorisation ability of models.

We describe each of the questions below. Broadly, they can be categorised into 4 categories: Spatial relationship-based (**SRB**), perceptual (**P**), simple reasoning (**SR**), and memory-based (**MB**). Examples of each of the questions are given, in order of introduction, in Figure 2a.

Table 3. Categories and their attributes.

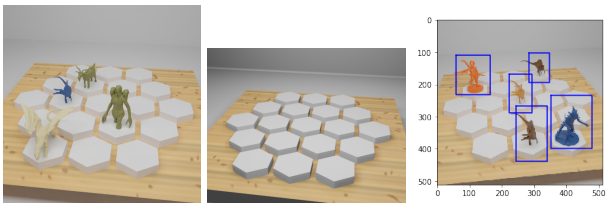
Adventurer	**Dragon**	**Animal**	**Monster**	**Mythical**
Species	Name	Type	Type	Type
Class	Pose	Predator-level	Name	Name
Gender	-	Prey-level	-	Gender
Weapon	-	Name	-	-
Mount	-	Food-habit	-	-
Attackable	-	-	-	-

Q1. Checking for relationships (SRB) This question presents a description of two objects and then asks if a spatial relationship exists between them.

Q2. Checking for chains of relationships (SRB) This question provides the description of three objects, and asks if the first satisfies a given spatial relationship with the second, and the second satisfies a second given spatial relationship with the third.

Q3. Checking for satisfying relationships of two types (SRB) This question provides the descriptions of three objects, and asks if the first satisfies a given spatial relation with the second, and also a given second relationship with the third.

Q4. Text-only swapping and checking for a relationship (SRB) This question provides a description of three objects, suggests a swap of position between two, and asks if the third satisfies a given spatial relationship with the first after swapping. This is new kind of question, similar to some suggested in [5], that changes the scene in text, and asks a question. A model that has learnt to separate an entity and relationships it may be in would find it easier to answer this, than a model that has not.



(a) (b) (c)

Figure 2. Examples of each question-type are given, in order in Figure 2a. **Q1.** Is a green regular-animal front of a blue spinosaurus? **A1.** no. **Q2.** Is there a light-yellow nithe dragon left of a blue spinosaurus that is back of a green goat? **A2.** no. **Q3.** Is there a green goat back of a light-yellow nithe dragon and right of the green coven-horror? **A3.** no. **Q4.** Swapping the position of a green coven-horror with a green goat, is a light-yellow nithe dragon front of it? **A4.** yes. **Q5.** Are there greater, equal or lesser number of dinosaurs than dragons? **A5.** equal. **Q6.** Upon removal of blue dinosaur how many blue dinosaur are present? **A6.** 0. **Q7.** Is there a orange abeloth dragon in the scene? **A7.** no. **Q8.** What is the predation-level of the green goat? **A8.** 1. **Q9.** Will goat attack spinosaurus? **A9.** no. **Q10.** Which category does nithe belong to? **A10.** dragon. Figure 2b shows the base scene. Figure 2c shows a scene with bounding boxes for the objects in the image.

Q5. Comparing based on attributes (SR) This question gives two descriptions of objects, and asks if the number of objects satisfying the first is lesser than, greater than, or equal to the number of objects satisfying the second.

Q6. Count of text-only removal of object (SR) This question suggests a text-only removal of an object, satisfying a description. Then it asks the count of objects satisfying another description. This question is also based on a text-only change of a scene, that necessitates a model understand the concept of removing an object, and counting.

Q7. Checking for an object (P) This question provides a description of an object and asks if that object is there in the scene.

Q8. Stating the properties of an object (MB) This question provides a description of an object and asks a property of that object.

Q9. Checking for non-spatial relationships (MB) Apart from spatial relationships, animals are related to other animals and adventurers by a property-based relationship. Each animal has a predation-level and a prey-level, indicative of some notion of which animal is likely to attack. An adventurer has a property of being attackable, which gives if an animal of sufficient predation-level can attack it. An animal can attack an adventurer if its predation-level is greater

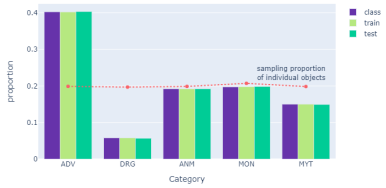


Figure 3. Figure gives the proportion of objects per category (purple) bar, the proportion of object-instances per category for the training dataset (light-green bar), and the proportion of object-instances per category for the test set (cyan bar). The dotted line shows the distribution of the mean number of instances for each object, seen category-wise (The figure can be best viewed in colour).

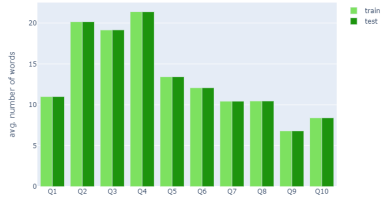


Figure 4. Figure shows the average word-length per question-type, for the train and test dataset (figure can be best viewed in colour).

than 2 and if the adventurer is attackable. An animal can attack another animal if its predation-level exceeds the prey-level of the other animal. This question gives a memory-based relationship that tests memory-based compositional generalisation.

Q10. Stating the category or team of an object (MB) This question provides a description of an object and asks the category or team of the object. While answering this question can be done by only text, it is useful in teaching a model the properties it needs for memory-based compositional generalisation.

A note about memory-based properties and questions: Why are questions that require text-only memorisation included in the VQA tasks of UOUC? For one, we wish to use these as a sanity-check for the models. Secondly, we wish to study if VQA models can also use visual aspects to improve upon text-only processing. Thirdly, we wish to study if memory-based tasks are harder than perceptual tasks. Note that the main challenge of UOUC is given by the first 4 questions, and then the next 3 questions. The memory-based questions are primarily investigative in nature, and can be used as a sanity-check for models.

Apart from these annotations, 2d and 3d bounding boxes for each object in the scenes of UOUC have been provided. An example is provided Figure 2c. The bounding boxes could be used for object-detection features.

4.1. Certain statistics for UOUC

We present the following statistics for UOUC in Figures 3 and 4: the proportion of each category in the object-classes, the proportion of the number of object-instances category-wise, and the proportion of the mean of the instances of each object, seen category-wise.

We see that the category 'Adventurers' has the maximum proportion of objects and object-instances. This is because the original data of the 3d models had a large number of such objects. However, due to the random and unbiased sampling of objects in the scenes, each object, regardless of category, can be seen to be almost as equally likely to be present in a scene. Moreover, the distributions of the object-instances and categories are similar for both the train and test datasets, ensuring that the models are challenged primarily based on question-answering, rather than by other factors relating to the sampling of objects.

We also plot the average word-length per question-type for both the train and test datasets in Figure. We see that the question-lengths are not particularly long, and that they are similar for both the train and test datasets. This again ensures that the challenge of answering questions on the test-set of UOUC is based on the logic of the questions, rather than other extraneous factors. Other statistics are given in the supplementary material.

Availability of UOUC: UOUC can be accessed from https://github.com/sairaamspage/compositionalvqa.

5. Experiments on recent VQA models

We trained and tested several models on UOUC. These are: TRAR [29], RWSAN [21], LSAT [23], AoA [22], LCGN [11], MCAN [27] , MAC [12], SAAA [16], and MUTAN [6]. These models have achieved reported good performance on several existing VQA datasets such as VQA-v1, VQA-v2, GQA, and CLEVR. Their results are given in Table 5.

In order to follow the training strategy as given by the original implementations, we used two kinds of backbone architectures. The first of these is a residual network based backbone. The second of these is an object detector based backbone that uses the NanoDet-Plus object detector `https://github.com/RangiLyu/nanodet`. These features are used so that the VQA models may accurately use class-label or object features for better VQA training.

We trained a ResNet-50 and a ResNet-101 on multi-label classification on UOUC. The models were trained to detect the presence of all possible objects in a scene. The ResNet-50 model achieves a precision of 0.41 and a recall of 0.44, while the ResNet-101 model achieves a precision of 0.51

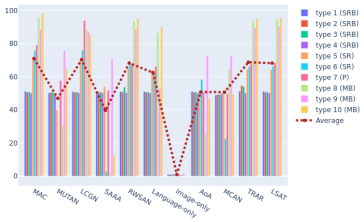


Figure 5. The accuracy per question-type for each model. The average accuracy is given by the dotted line (figure is best viewed in colour).

Table 4. Accuracies (in%) for the models on UOUC. Performance on the various types of questions are presented. l-only refers to language only. i-only refers to image-only.

Model	**SRB**	**SR**	**P**	**MB**	**Global**
MAC	50.72	73.06	79.18	94.12	71.05
LCGN	50.65	72.78	93.88	87.10	70.38
SAAA	50.61	28.46	51.78	28.51	39.67
MUTAN	50.34	20.01	57.7	57.06	47.03
MCAN	49.31	36.90	49.66	62.14	50.71
AoA	50.69	54.81	25.36	59.76	50.76
TRAR	52.55	65.46	67.78	92.87	68.75
RWSAN	51.47	66.25	67.60	92.28	68.29
LSAT	50.67	65.33	68.37	93.50	68.22
l-only	50.69	63.32	66.30	76.38	62.49
i-only	1.03	0.76	1.02	0.92	0.94

and a recall of 0.32 after training for 90 epochs. We use the ResNet models for MAC, LCGN, SAAA, MUTAN, TRAR, and LSAT.

We trained NanoDet-Plus for 100 epochs on UOUC for object detection. NanoDet-Plus achieves a mAP(0.5:0.95) of 0.57 on UOUC. We used the features from the FPN layer as features for AoA and RWSAN.

Table 5. Accuracies of the VQA models on popular VQA datasets. The reader may note that the VQA-v2 and GQA accuracies for RWSAN are on the test-dev datasets , while the others are for the test-std datasets. A comparison with Table 4 shows that the performance is higher for the popular datasets.

Model	VQA	CLEVR	GQA
MAC	-	98.80	54.06
LCGN	-	97.90	56.10
SAAA	64.60 (v1), 59.67 (v2)	-	-
MUTAN	67.36 (v1), 66.41 (v2)	-	-
MCAN	70.90 (v2)	-	57.40
AoA	71.14 (v2)	-	-
TRAR	72.93 (v2)	99.10	-
RWSAN	70.19 (v2)	98.42	57.43
LSAT	71.94 (v2)	98.72	-

We further trained an image-only model and a language-only model for the VQA task. This was done in order to estimate the extent of any biases in the images and questions towards answering the questions. The image-only model is a standard CNN-model, with a fully-connected layer for classification. It uses the ResNet-50 as a pre-trained backbone. The language-only model is a transformer model, with a fully-connected layer for classification. A low performance of these models can be inferred as some evidence for lower bias in the dataset. Since their purpose is only to indicate bias, we have not reported results on the other datasets.

Certain training details, like the source of their implementations and the number of epochs for all these models are given in the supplementary material. The accuracies for the models for each question, along with the average accuracies, is given in Figure 5. The average accuracy for each family of questions (**SRB**, **P**, **SR**, and **MB**) for each model is given in Table 4. We have also reported the global accuracy of each model in Table 4.

6. Discussion and analysis of the results

6.1. Analysis of text and image bias in UOUC

Due to natural correlations or, in our case, co-incidental patterns due to random choices, biases in the data could exist. The measurement of image and text biases allows us to form a lower-bound of sorts to assess the performance of models. That is, a model can be said to perform well if its performance exceeds the performance of the image-only and the language-only models (as well as random chance) by a significant amount. Similar models have been used in the GQA [13] and [14] datasets for the same purpose.

From Figure 5 and Table 4, we see that for all the question types, the image-only model performs very poorly. Its accuracy is below random performance, indicates a low bias in the images of our scenes, and a general lack of spurious correlations between the images and the questions.

From the same figure and table, we also observe that the language-only model performs better. For the SRB questions, this model performs with close to 50% accuracy. This performance indicates a low bias in the questions and answers of UOUC.

We see that the performance for the SR and P questions is not quite random, being around 63% and 66% for these questions. While this is definitely non-zero bias, this is still not very high. There is a great room for performance gains for VQA models that can differentiate them from the language-only model. We interpret the cause for such bias as co-incidence in our random selections of properties and objects for generating questions and answers. In short, we do not see any intentional inclusion of biases in our procedure as it is fully automated and random.

The analysis of the MB questions is more nuanced.

These questions are answerable based on just an analysis of text, with minimal use of the images. Among the three question-types that form the MB questions, only type 9 involves some textual reasoning, while types 8 and 10 are more based on memory. According to intuition, we see that the language-only model performs well on questions 8 and 10. Its performance on question 9 is still low, being only slightly above random chance. We explain this by seeing that questions 8 and 10 do not depend on the compositional separation of UOUC, being used as sanity checks mainly. Question 9, on the other hand, involves compositional generalisation, even if only based on simpler memory, rather than visual perception. We see that even this is quite difficult for the language-only model.

6.2. Analysis of Reasonably-performing models

From Table 4 and Figure 5, we see that the models that perform above the bias of the language-only model are MAC, LCGN, TRAR, RWSAN, and LSAT. These are the models that we consider to have reasonable performance.

SRB questions Each of these models perform close to randonly on the SRB questions. We explain this poor performance by understanding that these are based on visually perceiving spatial relationships and generalising this learnt perception on unseen pairs of objects. In other words, generalising learnt relationships on unseen compositions is hard even for these recent models.

It could be argued that the complexity of the objects could have a role to play. However, all these models use features from pretrained residual networks or NanoDet-Plus, both of which show non-random performance on detecting objects on UOUC. To emphasize this, consider that NanoDet-Plus achieves a mAP(0.5:0.95) score of 0.57 on UOUC, while it achieves a lower score of 0.34 on the COCO [7] dataset. This indicates that the objects of UOUC are easier to detect than those of COCO. Thus, compositional generalisation is hard, as opposed to detecting objects.

SR questions Among these models, only MAC and LCGN show significant gains over the language-only models. The other models show more modest gains. Yet, all of their performances are non-random. We attribute this to the fact that answering the SR questions involves simpler reasoning. However, compositional generalisation still plays a role which makes answering them reasonably hard.

P question Only LCGN is able to accurately detect all the objects with a very high degree of accuracy. Nonetheless, all of these models are able to detect models up to a non-random extent (and above the text bias). This is again proof of the fact that the objects are much easier to detect, when compared to understanding visual relationships.

MB questions All of these models perform very well on the MB questions. This indicates the fact that memorisation is easy for these models, and even compositional generalisation, if only on text, is doable. This leads to believe that compositional generalisation for visual relationships is possible, if an appropriate representation is used.

6.3. Analysis of other models

We see that SAAA, MUTAN, MCAN, and AoA perform rather poorly on all question types, with accuracies below the performance of the language-only model for some question-types. We interpret this performance as the models not being able to handle the change in the distribution of the test data. This change is caused by the images having unseen pairs of objects in a single image, which can lead to a significant change in the distribution of the images. This shift can allow for sanity checks that could lead to the better understanding of the failure points of VQA models.

6.4. Comparison of VQA models on UOUC and other datasets

We see that, in general, the performance on all models on the SRB type of questions is close to random performance. This is not the case for all the models on the other datasets. Even if the performance seem similar, for example MAC obtains 54.06% on GQA and 50.72% on SRB questions of UOUC, the interpretation of them are different. The number of possible answer classes for UOUC is very high as GQA contains open-ended questions. Thus, 54.06% is quite non-random performance. The SRB questions are binary, and 50.72% is close to random chance.

The performance of the models on SR questions for the reasonably performing models is better and quite comparable to their performance on VQA. Yet, given the comparatively lower number of answer-classes, we can still observe that much more performance gains are required before these UOUC questions can be considered solved. The same holds for even the P-type questions.

Only the memory-based questions are easily answered by models. However, we retain these models based on the fact that they can act as sanity checks on models, for example on SAAA, MUTAN, MCAN, and AoA.

One overwhelmingly obvious conclusion is that UOUC is much harder than CLEVR for all visual perception (SRB, SR, and P). Thus, we can use UOUC as a harder test of compositional reasoning.

The main challenge of UOUC is from the SRB questions, while the SR questions are also not very easy to answer. We see that, based on the performance of these recent, high-performing models, UOUC can act as a strong challenge for compositionality in VQA.

References

[1] Learning to compose neural networks for question answering", author = "andreas, jacob and rohrbach, marcus and dar-

rell, trevor and klein, dan. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1545–1554, San Diego, California, June 2016. Association for Computational Linguistics.

[2] Aishwarya Agrawal, Aniruddha Kembhavi, Dhruv Batra, and Devi Parikh. C-vqa: A compositional split of the visual question answering (vqa) v1. 0 dataset. *arXiv preprint arXiv:1704.08243*, 2017.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*, 2015. 1

[5] Christopher Beckham, Martin Weiss, Florian Golemo, Sina Honari, Derek Nowrouzezahrai, and Christopher Pal. Visual question answering from another perspective: Clevr mental rotation tests. 2020. 5

[6] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620, 2017. 2, 6

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 3, 8

[8] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 4

[9] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. Language-conditioned graph networks for relational reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10294–10303, 2019. 2, 6

[12] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations*, 2018. 6

[13] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 7

[14] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 1, 7

[15] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

[16] Vahid Kazemi and Ali Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*, 2017. 2, 6

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 1, 4

[18] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: Asimple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[19] Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022.

[20] David Mascharka, Philip Tran, Ryan Soklaski, and Arjun Majumdar. Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4942–4950, 2018.

[21] Bosheng Qin, Haoji Hu, and Yueting Zhuang. Deep residual weight-sharing attention network with low-rank attention for visual question answering. *IEEE Transactions on Multimedia*, 2022. 2, 6

[22] Tanzila Rahman, Shih-Han Chou, Leonid Sigal, and Giuseppe Carenini. An improved attention for visual question answering, 2020. 2, 6

[23] Xiang Shen, Dezhi Han, Zihan Guo, Chongqing Chen, Jie Hua, and Gaofeng Luo. Local self-attention in transformer for visual question answering. *Applied Intelligence*, pages 1–18, 2022. 2, 6

[24] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, 2019.

[25] Xiaofeng Yang, Guosheng Lin, Fengmao Lv, and Fayao Liu. Trrnet: Tiered relation reasoning for compositional visual question answering. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 414–430. Springer, 2020.

[26] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *arXiv preprint arXiv:1810.02338*, 2018.

[27] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290, 2019. 2, 6

[28] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

[29] Yiyi Zhou, Tianhe Ren, Chaoyang Zhu, Xiaoshuai Sun, Jianzhuang Liu, Xinghao Ding, Mingliang Xu, and Rongrong Ji. Trar: Routing the attention spans in transformer for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2074–2084, 2021. 2, 6

[30] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.