ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230544

Neuro-Symbolic Procedural Semantics for Reasoning-Intensive Visual Dialogue Tasks

Lara Verheyen[a;*], **Jérôme Botoko Ekila**[a], **Jens Nevens**[a], **Paul Van Eecke**[a;**] **and Katrien Beuls**[b;**]

[a]Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Belgium
[b]Faculté d'informatique, Université de Namur, Belgium

Abstract. This paper introduces a novel approach to visual dialogue that is based on neuro-symbolic procedural semantics. The approach builds further on earlier work on procedural semantics for visual question answering and expands it on the one hand with neuro-symbolic reasoning operations, and on the other hand with mechanisms that handle the challenges that are inherent to dialogue, in particular the incremental nature of the information that is conveyed. Concretely, we introduce (i) the use of a conversation memory as a data structure that explicitly and incrementally represents the information that is expressed during the subsequent turns of a dialogue, and (ii) the design of a neuro-symbolic procedural semantic representation that is grounded in both visual input and the conversation memory. We validate the methodology using the reasoning-intensive MNIST Dialog and CLEVR-Dialog benchmark challenges and achieve a question-level accuracy of 99.8% and 99.2% respectively. The methodology presented in this paper responds to the growing interest in the field of artificial intelligence in solving tasks that involve both low-level perception and high-level reasoning using a combination of neural and symbolic techniques.

1 Introduction

Visual dialogue refers to the task in which an artificial agent and a human hold a meaningful and coherent conversation that is grounded in visual input [7]. Typically, an agent needs to answer a sequence of questions about a given image, where the questions can only be understood in relation to previous question-answer pairs.

A schematic depiction of a typical visual dialogue task is shown in Figure 1. In this task, an agent is presented with the image on the left, and needs to answer the sequence of questions Q_1 to Q_4 on the right. The four question-answer pairs constitute a coherent dialogue, in which Q_1 can be answered based on the image alone, but in which Q_2 to Q_4 can only be answered based on the combination of the image and the previous question-answer pairs.

In this paper, we introduce the use of neuro-symbolic procedural semantic representations for solving visual dialogue tasks. We build further on earlier work in the area of visual question answering, in which procedural semantic representations have already been successfully used for representing the meaning of questions in the form of executable queries [2, 16, 26]. Such procedural semantic representations capture the logical structure underlying a question, and can be directly executed on a given image to compute an answer.

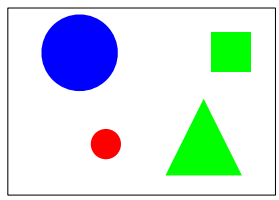

Figure 1. Schematic representation of a typical visual dialogue task, in which an artificial agent needs to answer a sequence of follow-up questions about an image.

An example of a procedural semantic representation for the question '*Are there more squares than circles?*', asked about the image in Figure 1, is shown in Figure 2. The query is composed of six operations, called *primitives*, that need to be performed by an artificial agent in order to compute the answer to the question. First of all, the SEGMENT-SCENE operation segments the image and binds the set of foreground objects to the '*?segmented-scene*' variable. Then, two FILTER operations take this set of objects as input and bind the set of squares and the set of circles to the variables '*?squares*' and '*?circles*' respectively. Then, the set of squares and the set of circles are counted by COUNT operations and the cardinality of each set is computed. Finally, the GREATER-THAN operation checks whether the cardinality of the first set is larger than the cardinality of the second set. The result of this last operation (in this case NO) is at the same time the answer to the question as a whole.

When moving from visual question answering to visual dialogue, the two-step process of first mapping a question to its logical structure and then executing the corresponding query on an image becomes more challenging. For example, in the question '*What is its colour?*', the possessive anaphoric pronoun "its" refers to an object that was introduced by an earlier question-answer pair, and which must be retrieved in order to be able to answer the question. As opposed to visual question answering systems, visual dialogue systems thus need to be able to keep track of the information that has been conveyed during earlier dialogue turns, as well as to use this information for answering questions in later turns.

In order to overcome this challenge, we introduce the use of a *conversation memory* as a data structure that explicitly and incrementally stores the information that is expressed in the subsequent turns of a dialogue. Additionally, we present a procedural semantic representation for visual dialogue tasks, which is able to query both visual input

* Corresponding Author. Email: lara.verheyen@ai.vub.ac.be.
** Joint last authors.

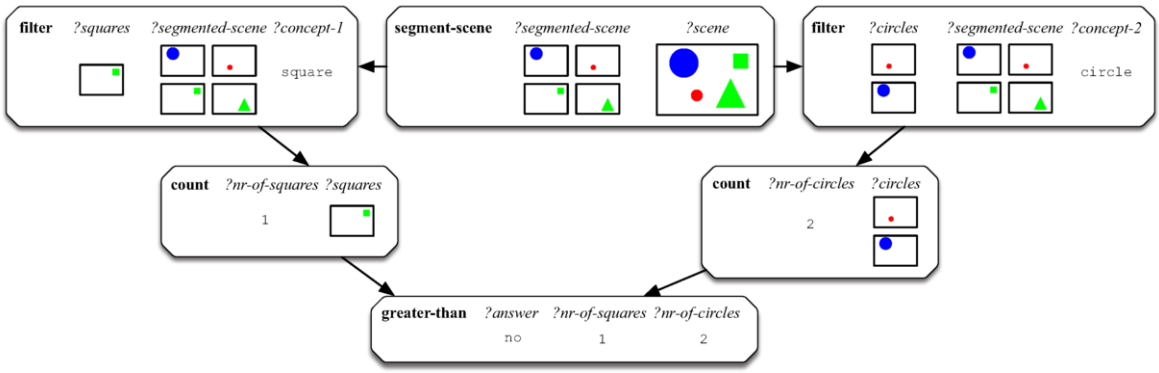

Figure 2. Example of a procedural semantic representation for the question *'Are there more squares than circles?'*, executed on the image in Figure 1. The answer to the question given this image is NO.

and the conversation memory. Due to its neuro-symbolic nature, this semantic representation can exploit both the strengths of subsymbolic systems for interacting with perceptual data, in this case the image, and the strengths of symbolic systems for reasoning based on previously acquired knowledge, in this case by retrieving structured information from the conversation memory.

The evaluation of our novel methodology on the reasoning-intensive MNIST Dialog benchmark [30] and the more challenging CLEVR-Dialog benchmark [20] shows that through the introduction of a conversation memory and the design of a compatible neuro-symbolic procedural semantic representation, we have been able to transfer the success of using procedural semantics in the field of visual question answering to the field of visual dialogue. The methodology presented in this paper contributes to the growing body of research in artificial intelligence that tackles tasks that involve both low-level perception and high-level reasoning using a combination of neural and symbolic techniques.

The rest of this paper is structured as follows. Section 2 presents a brief overview of the state of the art in visual dialogue and procedural semantics. Section 3 introduces our novel methodology for solving reasoning-intensive visual dialogue tasks. Section 4 presents two experiments in which our method is applied to the MNIST Dialog and CLEVR-Dialog benchmark datasets. Section 5 presents the experimental results. Section 6 provides a concluding discussion. A technical appendix is available at https://beehaif.org/docs/verheyen2023neuro-appendix.pdf.

2 Background and related work

The state of the art in visual dialogue is dominated by attention-based neural network approaches, which mainly differ in how they deal with co-references between question-answer pairs. In general, these approaches use an encoder-decoder architecture, which learns to attend to those regions of the image and/or previous question-answer pairs that are most relevant to answering a given question [7, 14, 35, 21]. A next line of research focuses on more explicitly keeping track of the entities that were evoked in earlier dialogue turns and on resolving co-references and ambiguities with respect to these entities. Starting from the observation that the proportion of follow-up questions with non-trivial co-references is limited in existing visual dialogue datasets, in particular VisDial [24, 1], [30] introduce the MNIST Dialog dataset with the specific purpose of evaluating to what extent visual dialogue models are actually capable of reasoning about previously introduced discourse entities. In the same paper, the authors introduce a model that explicitly represents the dialogue history as a combination of previous question-answer pairs and their associated attentions, and is able to retrieve the relevant attention for a given question from this associative memory. Building further on this work, [19] also represent the dialogue history in the form of an associative memory, but the keys are here more fine-grained entity-level descriptions instead of question-answer pairs. The authors introduce a neural module network architecture [3] in which the meaning representation includes two dedicated modules for interacting with the associative memory. [20] introduce the CLEVR-Dialog dataset for studying and benchmarking multi-turn reasoning in visual dialogue. [31] introduce three extensions of memory, attention and composition (MAC) networks [12] that deal with the conversational nature of visual dialogue tasks. A first extension consists in passing information across dialogue turns by initialising the memory state of the first MAC-cell of each turn with the value of the memory state of the last MAC-cell of the previous turn. A second extension concerns a context-aware attention mechanism that implements a transformer-like self-attention mechanism on the previous control states. A final extension consists in appending the entire dialogue history to the current question.

Procedural semantic representations, as pioneered by [37], [36] and [17] capture the meaning of linguistic expressions in the form of programs that can be executed algorithmically. When it comes to the properties of the procedural semantic representations themselves, three different approaches can be distinguished. A first class of models represent the meaning of utterances as queries expressed in a database querying language, such as FunQL [6] or SPARQL [38]. A second class of models represent the meaning of questions using logical forms, often defined in terms of variations on the lambda calculus [29]. The third class of models use formalisms that were especially designed for implementing and processing open-ended procedural semantic languages. Examples of models of this class include meaning representations represented in Incremental Recruitment Language (IRL) [32], as used for example by [28] and [26], or the functional programs used by [2] and [15]. While the primitive operations used in these special-purpose procedural semantics languages need to be implemented or learnt, this approach has the advantage that the languages are open-ended and directly executable.

Primitive operations in procedural semantics can be operationalised symbolically or subsymbolically. Neural module networks have been introduced by [3] as an operationalisation of fully sub-

symbolic procedural semantic representations applied to visual question answering tasks. [19] extend this approach to visual dialogue by adding primitive operations that perform multi-turn co-reference resolution. [39], [23] and [26] present a symbolic approach where the procedural semantic representations are not executed on the image directly, but on a scene graph representation that is generated first. Finally, [22] propose a hybrid procedural semantic engine which integrates neural predicates in probabilistic logic programs.

3 Methodology: high-level overview

Our novel approach to visual dialogue operationalises two main ideas. First, the history of a dialogue is represented explicitly, incrementally and in a structured way. We refer to the data structure holding this information by the term *conversation memory*. Second, the meaning of linguistic utterances is represented using a *neuro-symbolic procedural semantic representation* that combines subsymbolic and symbolic primitive operations.

3.1 Conversation memory

The conversation memory captures all information about the dialogue history that can be relevant for interpreting later dialogue turns, as inspired by the incremental build-up of logical forms in Discourse Representation Theory (DRT) [18]. The conversation memory represents this relevant information in an explicit and human-interpretable way, and is incrementally extended after each dialogue turn. Per turn, the conversation memory stores:

- a timestamp capturing the turn number
- the utterance observed during the turn
- the sentence type of this utterance, indicating for example the question type for questions
- the reply that was produced, if applicable
- the topic of the conversation from an information structure point of view
- a symbolic representation of the set of all entities evoked during the dialogue up to this turn, including all their properties that were mentioned
- for each entity, a pointer to an attention over the image that highlights its grounding in the input

A schematic representation of the conversation memory after processing the dialogue introduced in Figure 1 is shown in Figure 3. At this point, the conversation memory holds information about four subsequent dialogue turns. In the first turn, the question '*Are there any triangles?*' of type QUESTION-EXIST is observed and the answer '*Yes*' is returned. The topic of the conversation at this point is the entity '*object-1*'. Both the grounding of entity '*object-1*' in the input image and its mentioned shape property are stored in the conversation memory. In the second turn, the question '*How many?*' of type QUESTION-COUNT is asked about the current topic of the conversation and the answer '*One*' is returned. The topic of the conversation does not change and no additional information is added. In the third turn, the question '*Is there an object to its left?*' of type QUESTION-EXIST is processed and the answer '*Yes*' is returned. A new entity '*object-2*' is added to the conversation memory with as only information its grounding in the input image. The topic of the conversation shifts to entity '*object-2*'. Finally, at the fourth turn, the question '*What is its colour*?' is processed. The topic of the conversation, namely '*object-2*', is inferred from the previous turn and the answer '*Red*' is returned. The colour property of '*object-2*' is added to the representation of this entity in the conversation memory.

The information that we include in our implementation of the conversation memory reflects the information that is relevant in the visual dialogue tasks that we tackle in Section 4. We do not claim in any way that this information is sufficient to model everyday conversations between human interlocutors, which fall outside the scope of these benchmark challenges. Indeed, further research in pragmatics is needed in order to construct more accurate models of the role that discourse information plays in human conversation.

3.2 Neuro-symbolic procedural semantics

In tandem with the conversation memory, we introduce a neuro-symbolic procedural semantics that is designed to represent the meaning of utterances in their discourse context. The set of primitive operations that is part of our semantics is an extension of the set of operations used in the annotation of the CLEVR VQA dataset [15].

Our neuro-symbolic procedural semantics combines subsymbolic primitives that implement operations over unstructured data, in particular input images or attentions, with symbolic primitives that implement operations over structured data, in particular information contained in the conversation memory. Primitives that can operate on both structured and unstructured input have both a symbolic and a subsymbolic implementation. At runtime, the adequate implementation is then chosen based on the type of the input arguments.

Concretely, the neuro-symbolic procedural semantics consists of 16 primitive operations, which can combine to represent the meaning of statements and questions about objects in an image. The statements and questions can be about the existence and number of objects in the image, their attributes and the spatial relations between the objects. An overview of the different primitive operations as categorised by their symbolic or subsymbolic implementation is shown in Table 1. On the one hand, the set of symbolic primitives consists of primitives that operate on the conversation memory (i.e., GET-TOPIC, GET-PREVIOUS-TOPIC, GET-ATTRIBUTE-CATEGORY and FILTER) and primitives that perform reasoning operations, such as counting or checking the uniqueness of the input (i.e., UNIQUE, COUNT, EXIST, MORE-THAN-ONE, EXIST-OR-COUNT). On the other hand, the set of subsymbolic primitives includes operations related to perception, such as instance segmentation or classification of visual attentions according to their attributes. These primitives build upon a shared inventory of neural modules. The module that underlies the subsymbolic SEGMENT-SCENE primitive is a Mask R-CNN-based network for instance segmentation [10]. The modules underlying the other subsymbolic primitives (i.e. QUERY, FILTER, RELATE, EXTREME-RELATE and IMMEDIATE-RELATE) are implemented by SqueezeNet-based binary classifiers [13] that predict whether an object holds a particular attribute (e.g. BLUE, LARGE or SHINY). The FIND-IN-SCENE and SET-DIFFERENCE primitives bridge between the symbolic and the subsymbolic domains.

3.3 Extending the conversation memory

The conversation memory is extended with new information after each dialogue turn. Concretely, after each turn, a new turn representation is created for the current timestep. The timestep, utterance and reply slots of the turn representation are straightforwardly filled based on the available information. The sentence type is inferred from the final primitive operation executed during the evaluation of

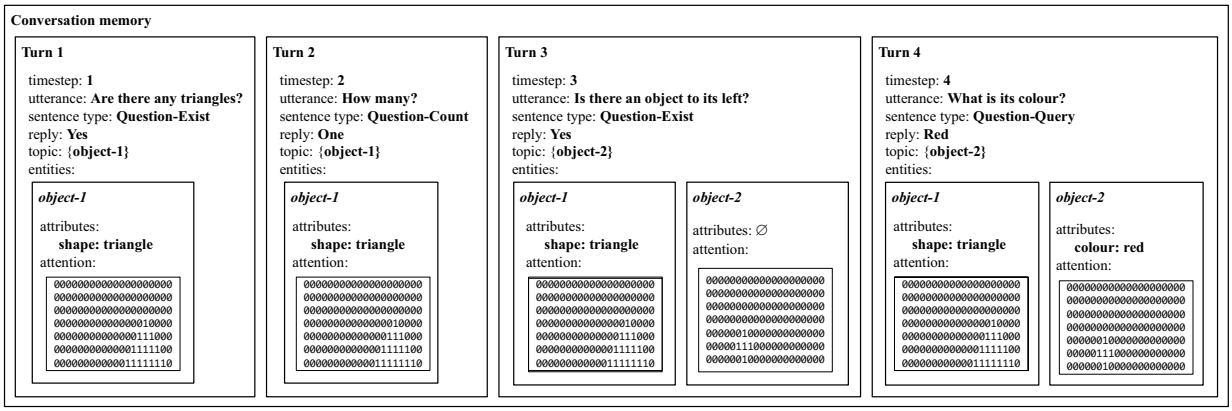

Figure 3. Schematic representation of the conversation memory after the fourth turn of the dialogue sketched in Figure 1. The conversation memory is incrementally updated after each dialogue turn as new information becomes available.

Table 1. Overview of primitive operations categorised by their symbolic or subsymbolic implementation.

symbolic	*subsymbolic*
FILTER	FILTER
UNIQUE	SEGMENT-SCENE
COUNT	RELATE
EXIST	EXTREME-RELATE
MORE-THAN-ONE	IMMEDIATE-RELATE
EXIST-OR-COUNT	QUERY
GET-TOPIC	
GET-PREVIOUS-TOPIC	
GET-ATTRIBUTE-CAT	
FIND-IN-SCENE	
SET-DIFFERENCE	

the semantic network for the current utterance. The topic corresponds to the set of objects that was bound to the input argument of the same primitive operation call. Finally, entities are added or updated based on the properties of the objects that were mentioned during the current turn.

4 Operationalisation of methodology and experimental set-up

We will now operationalise and validate our methodology using two standard benchmark challenges in the field of visual dialogue, in particular MNIST Dialog [30] and CLEVR-Dialog [20]. Both benchmarks were explicitly designed to be bias-free and to include a large proportion of non-trivial co-references across dialogue turns. Due to these two characteristics, answering the questions in the datasets cannot be done based on any statistical properties of the scenes, questions and answers alone, but requires actual reasoning about both the visual content and the discourse context.

4.1 MNIST Dialog

The MNIST Dialog dataset consists of 50,000 images, which are each accompanied by three dialogues. Each dialogue is in turn composed of 10 question-answer pairs. Each image consists in a synthetically generated 4x4 grid of hand-drawn digits with four randomly sampled attributes: colour, background colour, number and style. A symbolic description of the scene is also provided as meta-data, but is not part of the actual benchmark. The questions and answers are automatically generated. The questions can either query attributes of a single digit (e.g. '*What is the color of the digit below it?*') or count digits based on one or more of their attributes (e.g. '*Are there brown digits?*'). They can also include references to the spatial relations between the digits. The answers always take the form of a single word.

There are three main challenges involved in the operationalisation of our methodology for the MNIST Dialog benchmark. First of all, we need a means to map the MNIST Dialog questions to semantic networks that are composed of the primitive operations that we have introduced in Table 1. This is a highly non-trivial task, as the MNIST Dialog dataset does not come with any semantic annotation of the questions. Second, we need to train the neural network modules underlying the subsymbolic primitive operations on the MNIST Dialog images. Finally, we would like to be able to evaluate the process of mapping from questions to semantic networks, the execution of these networks, and the neural modules themselves independently from each other.

In order to operationalise the process of mapping from the MNIST Dialog questions to their semantic representations, we adopt a computational construction grammar approach [33, 34, 4, 5]. Concretely, we extend the computational construction grammar developed by [26] for the CLEVR VQA dataset [15] so that it is able to handle constructions involving co-referential expressions. The meaning predicates contributed by these additional constructions are expressed in terms of the primitive operations defined above. Although interesting in its own right, the details of the grammar itself fall outside the scope of this paper. We refer readers interested in the machine learning of computational construction grammars in the context of visual question answering to [25] and [8]. The execution of the semantic networks is modelled using the Incremental Recruitment Language (IRL) framework [32].

In order to verify the aptness of the semantic representations resulting from the language processing process, we have in a first phase made symbolic implementations of the primitive operations that work on the noise-free meta-data that describe the images rather than on the images themselves. By doing this, we could verify whether the predicted semantic networks would in theory always lead to the correct answer given a question and a scene. We could show that the networks indeed achieved a 100% accuracy when applied to the meta-data of the images. This proves that the primitive operations

presented in Table 1 are indeed sufficient to represent the meaning of the questions in the dataset. It is obviously the temporary noise-free condition of the dataset that makes the 100% figure possible.

The neural modules underlying the primitives were then each trained on the training section of the MNIST Dialog dataset and their accuracy was evaluated on the validation set. All individual modules achieved an accuracy of over 99.8% on the image data. Full technical detail on the training of the modules is provided in the technical appendix.

4.2 CLEVR-Dialog

The CLEVR-Dialog dataset consists of 85,000 images, which are each accompanied by five dialogues. Each dialogue starts with a caption that makes a statement about the contents of the image (e.g. '*There is a gray cube right of a shiny cylinder*'). The caption is then followed by 10 question-answer pairs. The images depict synthetically generated scenes consisting of 3D geometrical objects with randomly sampled attributes: shape, size, colour and material. The questions involve querying an attribute of an object in the scene (e.g. *'What shape is it?'*), counting objects based on one or more of their attributes (e.g. '*How many green spheres are there?*'), and querying whether a set of objects satisfies a given description (e.g. '*Are there any green spheres?*'). The questions can involve reference to different kinds of spatial relations between objects (e.g. 'the left block' and 'the block left of the green cylinder'). In contrast to MNIST Dialog questions, anaphora in CLEVR-Dialog questions can refer back to entities mentioned in any of the previous dialogue turns. Moreover, resolving history-dependent questions can require taking into account the entire dialogue history, as is for example the case in questions such as '*How many other objects are present in the image?*'.

In order to map from utterances to procedural semantic networks, we use the exact same construction grammar as the one used for the MNIST Dialog benchmark. In order to verify the aptness of the programs and language processing system, we create temporary symbolic implementations of the primitives and evaluate the programs that resulted from language processing on the noise-free meta-data that describe the images in the dataset. We achieved an accuracy of 99.99%[1].

The neural modules underlying the primitive operations were trained on the training portion of the CLEVR-Dialog dataset and their accuracy was evaluated on a held-out validation set. All modules achieved an accuracy of over 97.6%. Full technical detail on the training of the modules is provided in the technical appendix.

An operational example of the execution of a semantic network underlying a question from the CLEVR-Dialog dataset on an image is shown in Figure 4. In this example, the question '*What is its colour?*' following the caption '*There is a large sphere.*' is asked. The grammar maps the question to a procedural semantic program consisting of five primitive operations (i.e., SEGMENT-SCENE, GET-TOPIC, FIND-IN-SCENE, UNIQUE, QUERY). After execution of the primitive operations, which includes consulting the conversation memory in order to retrieve the topic of the conversation, the answer '*cyan*' is returned.

[1] The non-perfect accuracy was due to scenes that contained an even number of objects and in which a question relied on reference to the object 'in the middle'.

5 Results

When it comes to evaluating the performance of the overall system on the benchmark challenges, we include two different settings. First of all, in the '*standard*' setting, we evaluate the accuracy of the answers provided by our system as such. In the '*guessing*' setting, the system is allowed to make an educated guess when the execution of a semantic network fails and therefore does not lead to any answer. The educated guess is made based on the question type and the distribution of answers per question type in the training set. The '*guessing*' setting is foreseen in order to be able to compare our results to end-to-end neural approaches which always provide an answer even if its probability is low.

An overview of the evaluation results of our system on the MNIST Dialog and CLEVR-Dialog benchmark datasets is shown at the bottom of Table 2. In the best-performing experimental setting, i.e. the '*guessing*' setting, our system achieves a question-level accuracy of 99.8% on the MNIST Dialog benchmark and of 99.2% on the more challenging CLEVR-Dialog benchmark. In the '*standard*' setting, i.e. without guessing, it achieves a question-level accuracy of 99.8% and 99.0% respectively.

Table 2. Overview of results for MNIST Dialog and CLEVR-Dialog

	MNIST Dial	CLEVR-Dial
Encoder-decoder approaches		
LF [7]	45.1	55.9
HRE [7]	49.1	63.3
MN [7]	48.5	59.6
AMEM [30]	96.4	/
Neural module networks approaches		
N2NMN[2][11]	23.8	56.6
corefNMN [20]	99.3	68.0
MAC network approaches		
MAC-CQ-CAA-MTM [31]	/	98.3
Ours		
standard	**99.8**	99.0
guessing	**99.8**	**99.2**

The table also compares our results against previous approaches, namely the encoder-decoder-based approaches presented by [7] and [30], the neural module networks-based approaches by [11] and [20], and the MAC network-based approach by [31]. We can see that our system outperforms the state of art on both MNIST Dialog and CLEVR-Dialog. While other approaches that tackle both visual dialogue benchmark challenges typically perform much better on the easier MNIST Dialog benchmark as compared to the more challenging CLEVR-Dialog benchmark, our approach obtains consistently good results across both datasets.

6 Discussion and conclusion

In this paper, we have introduced a novel methodology to visual dialogue that is based on neuro-symbolic procedural semantics and have evaluated it on the reasoning-intensive MNIST Dialog and CLEVR-Dialog benchmark challenges. Concretely, our contribution consists in (i) the introduction of a conversation memory as a data structure that explicitly and incrementally represents the information that is

[2] The evaluation of the model on the MNIST Dialog dataset is reported by [19] and of CLEVR-Dialog by [20].

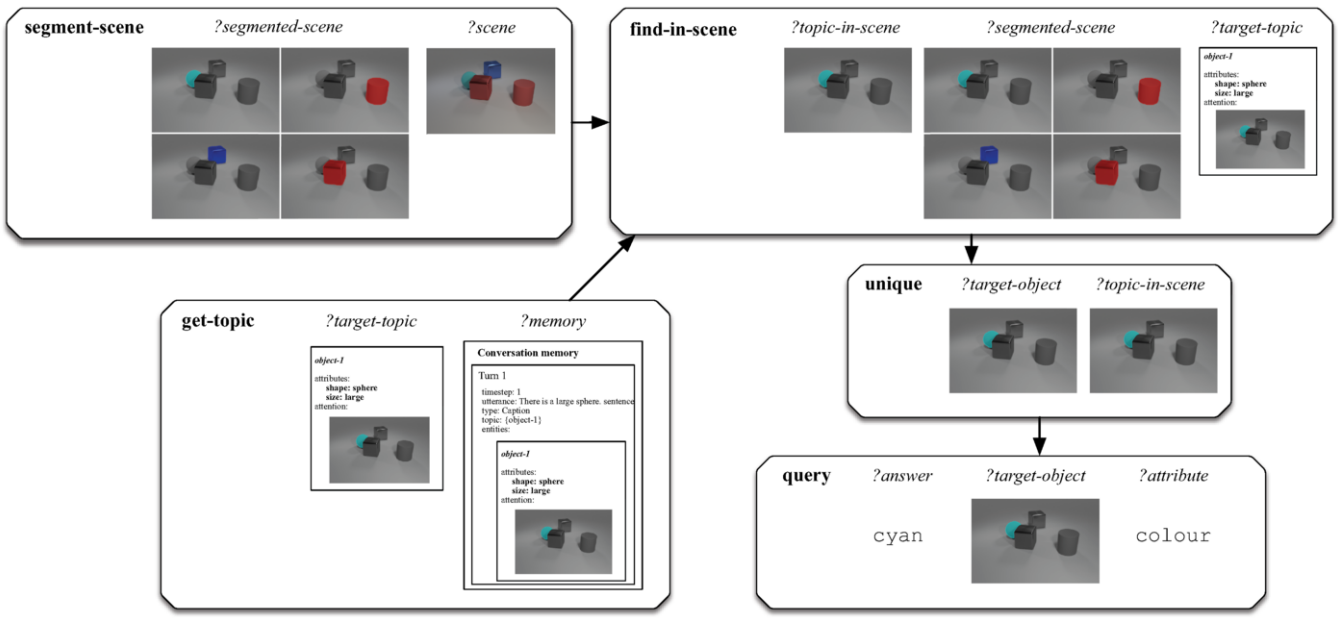

Figure 4. Schematic representation of the execution of the semantic representation for the utterance '*What is its colour?*' following the caption '*There is a large sphere.*' on a scene from the CLEVR-Dialog dataset.

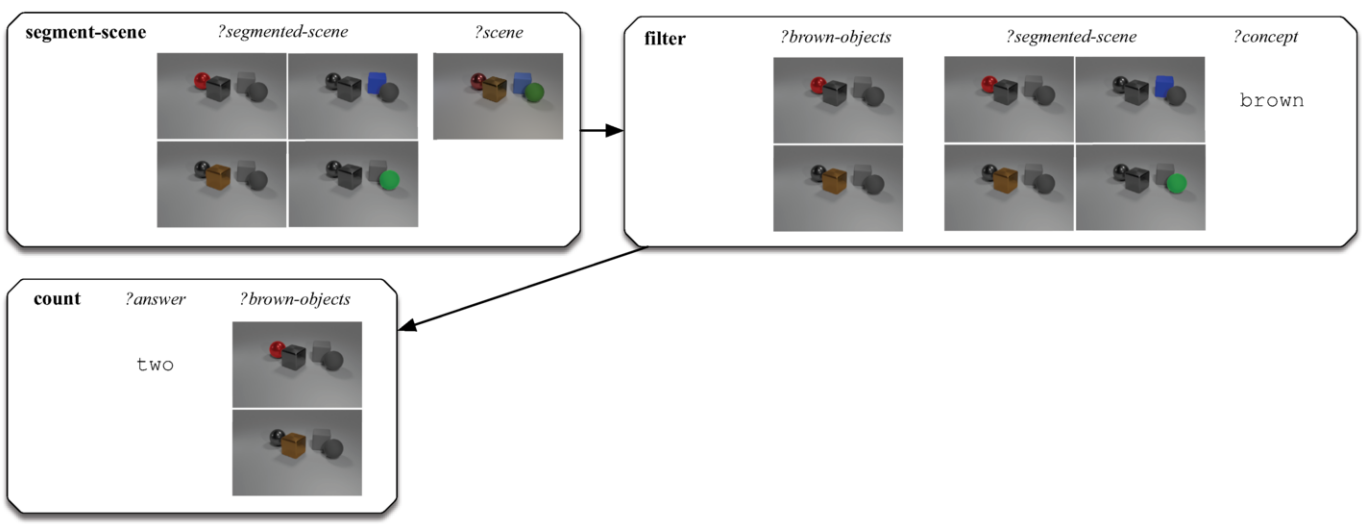

Figure 5. Schematic representation of the execution of the semantic network underlying the utterance '*How many brown objects are there?*' on a scene from the CLEVR-Dialog dataset, illustrating the transparency of the approach. The filter operation wrongly recognises the leftmost object to be brown. As a consequence, two brown objects are counted instead of one.

expressed during the subsequent turns of a dialogue, and (ii) the design of a neuro-symbolic procedural semantics that is grounded in both visual input and the conversation memory.

While the reported benchmark accuracies are definitely important to validate our methodology in comparison to existing approaches, the more prominent contribution of the methodology that we present lies in four main characteristics that distinguish it from the state of the art in visual dialogue. First of all, the methodology is explainable in human-interpretable terms. Input utterances are mapped onto procedural semantic representations, which correspond to logic programs. These programs, which reveal the logical structure underlying an input utterance, are composed of human-interpretable primitive operations, such as COUNT, QUERY and FILTER. This means that the result of the initial language processing step can be inspected and understood by the user. The conversation memory of the system also stores information about the history of a dialogue in a structured and human-interpretable way, thereby being fully transparent about what is remembered by the system. The input and output of each primitive operation can be traced and interpreted, as they consist in either meaningful symbols (human-interpretable categories) or visual attentions over images. Given that these visual attentions are the input and output of human-interpretable operations, humans are able to judge whether an attention corresponds to what is expected or not. As the symbolically implemented primitives can be traced on a meaningful level, the only aspect of the system where the interpretability of the computation is limited is situated in the subsymbolic primitives that deal with perception on the lowest level. By pushing the neuro-symbolic boundary so far down, we ensure that any reasoning capabilities that exceed the perception of basic categories is explainable in human-interpretable terms.

A related advantage of this approach is that it avoids inconsistencies in reasoning by implementing its subsymbolic primitive operations on top of a shared inventory of highly-specialised neural modules. Keeping consistency across reasoning operations is a highly desirable property of intelligent systems, which at the same time leads to a more human-like behaviour. For example, it is obvious that the human capabilities of recognising objects and counting objects rely on the same conceptual distinctions. This is reflected in our system by implementing the COUNT primitive in terms of computing the cardinality of a set of objects returned by a FILTER operation, which is itself implemented based on the same set of binary classifiers as the QUERY operation. The answer to the question '*How many red blocks are there?*' is as a consequence guaranteed to be consistent with the answers to the question '*What is the colour of the block?*' asked for each block in the scene.

A third asset of our approach is that it can effectively monitor its own performance. This has become a topic of high interest in the AI community, since deep neural networks often provide confidence scores of poor quality, especially when it comes to out-of-distribution data [27, 9]. Concretely, in our case, the system knows that it has not been able to answer a question based on sound logic reasoning if the execution of a semantic network fails. While it can still make an educated guess in such cases, the system then indicates that the result should be interpreted with extra care. In fact, the execution of a semantic network fails in 55.0% of the CLEVR-Dialog errors (i.e. errors in the '*standard*' setting) and in 41.7% of the MNIST Dialog errors (in the '*standard*' setting as well). The remaining 45.0% and 58.3% of errors respectively remain undetected by the system. This amounts to only 0.4% of the questions in CLEVR-Dialog and 0.1% of the questions in MNIST Dialog.

A final advantage resides in the modularity of the approach. New primitive operations can be added to the system in order to accommodate new tasks or to model new cognitive capabilities acquired by an artificial agent. These new primitives can add to both the logical and perceptive reasoning capabilities of the agent. Where appropriate, they can reuse neural modules used by existing primitives without needing to retrain them. Neural modules can also dynamically be added, but these might affect the performance of other modules and therefore require retraining some of them. For example, adding a binary classifier for a new colour will likely affect the performance of existing binary classifiers for other colours, as these were trained in the absence of the new colour category.

Figure 5 illustrates the interpretability of our approach by providing an example of a question from the CLEVR-Dialog dataset that was wrongly answered. Concretely, this example shows how the system supports the tracking of the source of errors by providing insight into the logical structure underlying the question, and into the input and output of the different primitive operations that were performed. The example shows the execution of the semantic network underlying the utterance '*How many brown objects are there?*' on a given CLEVR scene. We can see that the question has been analysed into three primitive operations: segmenting the scene (SEGMENT-SCENE), filtering the segmented scene for the colour brown (FILTER) and counting the number of objects in the resulting set (COUNT). The result of the counting operation, which is at the same time returned as the answer to the question, is TWO. However, this answer does not match the gold standard answer from the dataset, which is ONE. Indeed, when scrutinising the execution trace of the semantic network on the scene, it becomes clear that the filter operation has retrieved two brown objects. After a visual inspection of the attentions, the human observer can see that the leftmost object in the scene was wrongly classified as being brown and the source of the error has been found. If we would now query the colour of the leftmost object in the scene, the system is also guaranteed to answer BROWN, as the FILTER and QUERY primitives internally rely on the same neural classifiers. Thus, while the answer to the question is wrong, it is logically consistent with the overall perception and reasoning skills of the system.

In sum, the research reported on in this paper contributes to the growing body of research in artificial intelligence that tackles tasks that involve both low-level perception and high-level reasoning using a combination of neural and symbolic techniques. Neural techniques are used to deal with low-level perception tasks and thereby give rise to meaningful symbols that can then be used as a basis for higher-level reasoning operations. It thereby bears the promise of leading to the development of artificial agents with more explainable, consistent and human-like cognitive capacities.

Acknowledgements

The research reported on in this paper received funding from the EU's H2020 RIA programme under grant agreement no. 951846 (MUHAI), from the Research Foundation Flanders (FWO) through a postdoctoral grant awarded to PVE (grant no. 76929) and from the Flemish Government under the 'Flanders AI Research Program'.

References

[1] Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser, 'History for visual dialog', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8182–8197, Online, (2020). Association for Computational Linguistics.

[2] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein, 'Learning to compose neural networks for question answering', in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, eds., Kevin Knight, Ani Nenkova, and Owen Rambow, pp. 1545–1554. Association for Computational Linguistics, (2016).

[3] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein, 'Neural module networks', in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39–48. IEEE Computer Society, (2016).

[4] Katrien Beuls and Paul Van Eecke, 'Fluid Construction Grammar', in *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, eds., Claire Bonial and Harish Tayyar Madabushi, pp. 41–50, (2023).

[5] Katrien Beuls and Paul Van Eecke, 'Construction grammar and artificial intelligence', in *The Cambridge Handbook of Construction Grammar*, eds., Mirjam Fried and Kiki Nikiforidou, Cambridge University Press, Cambridge, UK, (2024). Forthcoming.

[6] Jianpeng Cheng, Siva Reddy, Vijay Saraswat, and Mirella Lapata, 'Learning an executable neural semantic parser', *Computational Linguistics*, **45**(1), 59–94, (2019).

[7] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra, 'Visual dialog', in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1080–1089. IEEE Computer Society, (2017).

[8] Jonas Doumen, Katrien Beuls, and Paul Van Eecke, 'Modelling language acquisition through syntactico-semantic pattern finding', in *Findings of the Association for Computational Linguistics*, eds., Andreas Vlachos and Isabelle Augenstein, pp. 1317–1327. Association for Computational Linguistics, (2023).

[9] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy, 'Explaining and harnessing adversarial examples', in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, pp. 1–15, (2015).

[10] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick, 'Mask r-cnn', in *2017 IEEE International Conference on Computer Vision (ICCV)*, eds., Rita Cucchiara, Yasuyuki Matsushita, Nicu Sebe, and Stefano Soatto, pp. 2961–2969. IEEE Computer Society, (2017).

[11] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko, 'Learning to reason', in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 804–813. IEEE Computer Society, (2017).

[12] Drew A. Hudson and Christopher D. Manning, 'Compositional attention networks for machine reasoning', in *6th International Conference on Learning Representations, ICLR 2018*. OpenReview.net, (2018).

[13] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size, 2016. arXiv preprint arXiv:1602.07360.

[14] Unnat Jain, Svetlana Lazebnik, and Alexander Schwing, 'Two can play this game', in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5754–5763. IEEE Computer Society, (2018).

[15] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick, 'CLEVR', in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2901–2910. IEEE Computer Society, (2017).

[16] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick, 'Inferring and executing programs for visual reasoning', in *2017 IEEE International Conference on Computer Vision (ICCV)*, eds., Rita Cucchiara, Yasuyuki Matsushita, Nicu Sebe, and Stefano Soatto, pp. 2989–2998. IEEE Computer Society, (2017).

[17] Philip N. Johnson-Laird, 'Procedural semantics', *Cognition*, **5**(3), 189–214, (1977).

[18] Hans Kamp and Uwe Reyle, *From discourse to logic*, volume 42, Springer Science & Business Media, 2013.

[19] Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach, 'Visual coreference resolution in visual dialog using neural module networks', in *Computer Vision – ECCV 2018*, eds., Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, pp. 153–169. Springer, (2018).

[20] Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach, 'Clevr-dialog', in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 582–595, Minneapolis, MN, USA, (2019). Association for Computational Linguistics.

[21] Mingxao Li and Marie-Francine Moens, 'Modeling coreference relations in visual dialog', in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3306—3318, Online, (2021). Association for Computational Linguistics.

[22] Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, and Luc De Raedt, 'Neural probabilistic logic programming in DeepProbLog', *Artificial Intelligence*, **298**, 103504, (2021).

[23] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu, 'The neuro-symbolic concept learner', in *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net, (2019).

[24] Daniela Massiceti, Puneet K. Dokania, N. Siddharth, and Philip Torr, 'Visual dialogue without vision or dialogue', in *Critiquing and Correcting Trends in Machine Learning : NeurIPS 2018 Workshop*, (2018).

[25] Jens Nevens, Jonas Doumen, Paul Van Eecke, and Katrien Beuls, 'Language acquisition through intention reading and pattern finding', in *Proceedings of the 29th International Conference on Computational Linguistics*, eds., Nicoletta Calzolari and Chu-Ren Huang, pp. 15–25. International Committee on Computational Linguistics, (2022).

[26] Jens Nevens, Paul Van Eecke, and Katrien Beuls, 'Computational construction grammar for visual question answering', *Linguistics Vanguard*, **5**(1), 20180070, (2019).

[27] Anh Nguyen, Jason Yosinski, and Jeff Clune, 'Deep neural networks are easily fooled: High confidence predictions for unrecognizable images', in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 427–436. IEEE Computer Society, (2015).

[28] Simon Pauw and Joseph Hilferty, 'Embodied cognitive semantics for quantification', *Belgian Journal of Linguistics*, **30**(1), 251–264, (2016).

[29] Siva Reddy, Mirella Lapata, and Mark Steedman, 'Large-scale semantic parsing without question-answer pairs', *Transactions of the Association for Computational Linguistics*, **2**, 377–392, (2014).

[30] Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal, 'Visual reference resolution using attention memory for visual dialog', in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, eds., Isabelle Guyon, Ulrike Von Luxburg, Samy Bengio, Hanna Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, pp. 3722–3732. Curran Associates Inc., (2017).

[31] Muhammad A. Shah, Shikib Mehri, and Tejas Srinivasan, 'Reasoning over history: Context aware visual dialog', in *Proceedings of the First International Workshop on Natural Language Processing Beyond Text*, pp. 75–83, Online, (2020). Association for Computational Linguistics.

[32] Michael Spranger, Simon Pauw, Martin Loetzsch, and Luc Steels, 'Open-ended procedural semantics', in *Language Grounding in Robots*, eds., Luc Steels and Manfred Hild, 153–172, Springer, New York, NY, USA, (2012).

[33] Paul Van Eecke and Katrien Beuls, 'Exploring the creative potential of computational construction grammar', *Zeitschrift für Anglistik und Amerikanistik*, **66**(3), 341–355, (2018).

[34] Remi van Trijp, Katrien Beuls, and Paul Van Eecke, 'The FCG Editor', *PLOS ONE*, **17**(6), e0269708, (2022).

[35] Yue Wang, Shafiq Joty, Michael R. Lyu, Irwin King, Caiming Xiong, and Steven C. H. Hoi, 'VD-BERT: A unified vision and dialog transformer with BERT', in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3325–3338, Online, (2020). Association for Computational Linguistics.

[36] Terry Winograd, 'Understanding natural language', *Cognitive Psychology*, **3**(1), 1–191, (1972).

[37] William A. Woods, 'Procedural semantics for a question-answering machine', in *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I*, pp. 457–471, New York, NY, USA, (1968).

[38] Mohamed Yahya, Klaus Berberich, Shady Elbassuoni, Maya Ramanath, Volker Tresp, and Gerhard Weikum, 'Natural language questions for the web of data', in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 379–390, (2012).

[39] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum, 'Neural-symbolic VQA', in *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, eds., Samy Bengio, Hanna Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, pp. 1031–1042, (2018).