Question Answering for Visual Navigation in Human-Centered Environments

Daniil E. Kirilenko[1], Alexey K. Kovalev[2,3(✉)], Evgeny Osipov[4], and Aleksandr I. Panov[1,3]

[1] Moscow Institute of Physics and Technology, Moscow, Russia
panov.ai@mipt.ru
[2] HSE University, Moscow, Russia
akkovalev@hse.ru
[3] Artificial Intelligence Research Institute FRC CSC RAS, Moscow, Russia
[4] Lulea University of Technology, Lulea, Sweden
evgeny.osipov@ltu.se

Abstract. In this paper, we propose an HISNav VQA dataset – a challenging dataset for a Visual Question Answering task that is aimed at the needs of Visual Navigation in human-centered environments. The dataset consists of images of various room scenes that were captured using the Habitat virtual environment and of questions important for navigation tasks using only visual information. We also propose a baseline for a HISNav VQA dataset, a Vector Semiotic Architecture, and demonstrate its performance. The Vector Semiotic Architecture is a combination of a Sign-Based World Model and Vector Symbolic Architectures. The Sign-Based World Model allows representing various aspects of an agent's knowledge, and Vector Symbolic Architectures serve on a low computational level. The Vector Semiotic Architecture addresses the symbol grounding problem that plays an important role in the Visual Question Answering Task.

Keywords: Visual question answering · Semiotic approach · Vector symbolic architecture · Habitat · Visual Navigation

1 Introduction

In recent years, models that work with unimodal data have achieved significant results in computer vision (CV) and natural language processing (NLP). And nowadays, researchers have started paying attention to multimodal tasks especially at the intersection of CV and NLP: Image Captioning [19], Visual Question Answering (VQA) [6], Visual Dialog [4], Visual Commonsense Reasoning [32], and Vision-and-Language Navigation (VLN) [2,18].

The advance in virtual assistants requires further development within the framework of embodied Artificial Intelligence (AI). Embodied AI is the study of intelligent systems with a physical or virtual embodiment (robots and egocentric personal assistants). The embodiment hypothesis is the idea that "intelligence

I. Batyrshin et al. (Eds.): MICAI 2021, LNAI 13068, pp. 31–45, 2021.
https://doi.org/10.1007/978-3-030-89820-5_3

emerges in the interaction of an agent with an environment and as a result of sensorimotor activity" [20]. As a testbed, the task of Visual Navigation in human-centered environments might be used as it is possible to supplement it with an ability to interact with humans in the natural language. In VLN, the agent has to navigate the environment following instructions in the natural language, and in VQA, the system has to answer questions about the content of a given image. The combination of VLN and VQA gives an agent that can answer (and, potentially, ask) questions about the environment an opportunity to replenish its scene understanding and clarify instructions.

When an agent faces such tasks, the symbol grounding problem (how symbols get their meanings) [9] arises. We approach this problem from the semiotic point of view by applying Vector Semiotic Architecture, which is a combination of the Sign-Based World Model [22,26] and Vector Symbolic Architecture [12].

The goal of our work is to take one more step toward creating an embodied AI assistant system that will improve human-machine interaction by using a natural language question, which, in the future, can help to set tasks or refine them for AI agents in the simplest way.

The contribution of this paper is twofold: first, a challenging dataset for a Visual Question Answering task that aims to address the needs of Visual Navigation in human-centered environments is proposed. The HISNav VQA dataset is simpler than VQA [6], but more complex than CLEVR [11], and focuses on questions about positions and relations of objects. It also does not suffer from disembodiedness, as images are taken from the robot's camera, and unsituatedness, as scenes resemble environments a robot is supposed to operate in. Second, the Vector Semiotic Architecture baseline is proposed, and its performance is demonstrated on the HISNav VQA dataset. The advantages of Vector Semiotic Architecture are the interpretability of an answering process and the grounding of semiotic representation of objects in robot sensory inputs.

2 Related Works

The research interest in multimodal tasks in the intersection of CV and NLP leads to the emergence of VQA datasets that cover commonsense knowledge and are general-purpose [6], that use synthetic image-question pairs [11], and that are based on questions asked by vision-impaired people [8]. Despite that fact, the area of Visual Navigation in human-centered environments is not covered by VQA datasets, although the ability of an intelligent agent to answer questions about the environment it operates in and, by that, adjust its action is very promising.

These needs are partially satisfied by VLN works [2,18]. But the features of the problem formulation (the agent has to follow linguistic instruction to navigate across the environment based on visual information) do not imply that the agent asks or answers questions to clarify these instructions. That said, this ability is important in situations where the environment can change dynamically and the agent has to re-plan to complete the task.

Most VQA models reduce the task to the classification problem and use neural networks to solve it [1]. Other research directions include the use of external knowledge [31] and neural-symbolic approach [30]. The latter is of particular interest as it combines the advantages of connectionism and symbolism while compensating for their drawbacks. The approach allows obtaining interpretable answering procedures and using rich and efficient tools of deep neural networks. In this work, we use [30] as a starting point. In [30], the scene is represented in the form of the data frame listing objects and their attributes. The process of answering a question is performed by applying a deterministic program that filters the scene data frame and outputs the result as the answer. The disadvantage of representing a scene as a table (data frame) is the loss of structure. We address that problem by applying the scene representation proposed in [17]. In [17], a model that combines a Sign-Based World Model [23,24] and Vector Symbolic Architectures [12] was proposed. Binary Spatter Codes [12] were used to represent causal matrices as vectors. In this work, we applied Spatial Semantic Pointers [15] that provide a convenient way to work with continuous coordinates.

The Sign-Based World Model (SBWM) is a cognitive architecture that allows representing different aspects of agent knowledge. The main information unit in this architecture is a sign – a four-component structure. The meaning component represents the agent's experience of interaction with a concept related to a sign. Knowledge shared across the group of agents is represented by the significance component. The image component serves to distinguish one concept from another. The fourth component is a name that serves a nominative function. SBWM was used for various applications, such as goal setting [25], reasoning [13,16], and hierarchical planning [14]. Each sign component is represented by a binary matrix of a special form, where columns are events and rows are the appearance of a particular feature in the particular event. These matrices are called causal matrices. In [17], causal matrices are encoded using Vector Symbolic Architectures.

Vector Symbolic Architectures (VSA), or hyperdimensional computing [12], is a family of methods that use vectors of high dimensionality to encode concepts and use vector operations to manipulate them. The concept representation is holistic, which means that no particular position in a vector is interpretable. That representation allows reducing symbolic manipulation to vector operations. VSA is instantiated by many variations [5,12,27] that differ in vector space and operations but follow the same computational properties.

3 HISNav VQA Dataset

We present a challenging dataset that consists of images of various room scenes (captured using the Habitat virtual environment [21]) and questions to these images. Our dataset tests the ability of VQA systems to answer questions important for navigation tasks using only visual information. Among the tasks that need to be solved to correctly answer the questions are recognition of objects, the correct determination of their properties, and counting. To answer spatial

questions, it is necessary to determine the position of objects relative both to the observation point and to each other. Thus, a model that can accurately answer such questions can be effectively used to solve navigation problems for robotic platforms, and our dataset is a way to check the necessary visual reasoning abilities. Various samples of image-question pairs from the proposed dataset are demonstrated in Fig. 1. Our HISNav VQA dataset is publicly available.[1]

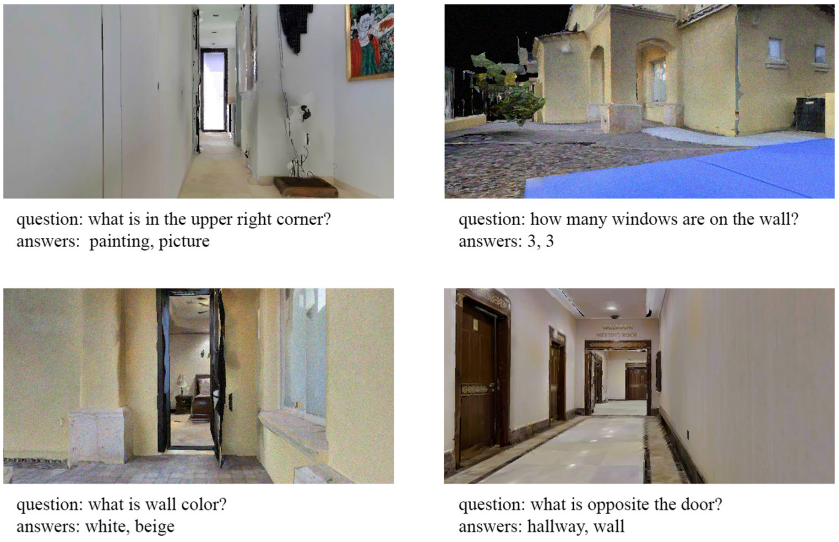

Fig. 1. Examples of image-question pairs from the HISNav VQA dataset

3.1 Images

We used the HISNav dataset [28] as a source of images, which was assembled in the virtual environment Habitat [21]. Habitat is a highly efficient photorealistic 3D simulation for research in embodied AI, that is a great platform for our purposes. For more closeness of synthetic images to images from real cameras, Gaussian noise was added to some of them. All HISNav dataset includes 135,962 images, each RGB image has a resolution 640×320, and ground truth instance labels of 40 classes (wall, floor, chair, door, table, sofa, etc.) correspond to each image. The original dataset includes pictures of 49 unique scenes that present different rooms with various content.

In HISNav, each subsequent image is too close to the previous one, which is undesirable for the dataset purposes. We want the images to be so different from each other that the answers to the same question differ in most cases (or the reason for the answer should be different in the case of the same answer). For this reason, we used only one in thirty image from the initial data. In addition,

[1] https://bit.ly/2XR5OUc.

images with a small amount of content on them (less than five objects on the image) were removed. Images that had strong visual artifacts are also discarded.

3.2 Human-Asked Questions

The crowdsourcing service Yandex.Toloka[2] was used to collect questions and answers to them. The collection of questions and answers had two stages, which were performed by different groups of participants.

First, we asked performers to ask questions about the images (Appendix A). We limited the types of asked questions to the following four main ones: questions about mutual arrangement of objects, quantity, properties of objects, and questions about relative to the observer location. The resulting questions were rigorously assessed following the given instruction. Also, during this stage, performers were asked to give an answer to their question and mark the type of question. This part of the work was not evaluated strictly since the main task of this stage is to collect questions. The total number of workers who participated in this stage was 1,172, with an overall task acceptance rate of 0.63.

The second stage is collecting answers. This time, performers had to answer questions about the images that were collected in the previous step. The instructions for this task were written to bring all answers to the same form and to reduce the number of unique answers in the resulting data. For each image-question pair two answers were collected to reduce the likelihood of erroneous markup. The total number of workers who participated in this stage was 787, with an agreement rate of 0.37.

3.3 Synthetic Questions

For question generation a modular algorithm was written, it generates seven types of questions based on the tabular representation of each scene, which was obtained from the results of the instance segmentation of the corresponding image (Fig. 2). The following question templates were used:

- What color is the nearest object to the [single obj]?
- What color is the [single obj] to the [single obj]?
- What is the nearest object to the [single obj]?
- In which part of the image is the [single obj]?
- How many [multiple obj] are there?
- How many [multiple obj] are to the left/right of the [single obj]?
- Is there a [object] to the left/right of the [single obj]?

Here [single obj] is any object type that is represented in the image in a single instance, [multiple obj] is the type of object that is represented in the image by more than one instance, and [object] is a placeholder for any type of object that may not even be represented.

[2] https://toloka.yandex.com.

Fig. 2. Examples of generated questions

3.4 Dataset Analysis

Our dataset is different from other VQA datasets like CLEVR [11] and VQA v2.0 [6]. VQA v2.0 is a very large and diverse dataset full of common-sense questions. This is good in the case of creating a universal VQA model that can answer a broad variety of questions. CLEVR is much less diverse and, like our dataset, consists of synthetic images. At the same time, it has a weak variety of objects and scenes represented on them, which is why the number of unique words in questions and answers is extremely small. Also, the structure of questions is extremely complex and dissimilar to what people use. Both datasets suffer from disembodiedness, as images are taken from different shoot points (an agent position is not considered), and unsituatedness, as scenes do not resemble the actual environment in which an agent is supposed to operate. That limits the application of these datasets for Visual Navigation in human-centered environments. By the latter, we mean ordinary rooms with furniture and elements of everyday life in which we are all used to living. On the other hand, the proposed dataset is focused on important navigation questions such as the location of various objects, uses images that resemble the operating environment, and are taken from an agent viewpoint. This is what distinguishes it from other VQA datasets.

The HISNav VQA dataset includes 3,500 images with one human-asked question per image and two answers per question. There are 712 unique answers and all questions contain 868 unique words. This is about an order of magnitude more than the CLEVR dataset has and at the same time an order of magnitude less than the VQA dataset (Table 1).

Figure 3(a) shows the distribution of collected human-asked questions by their first four words. The ordering of the words starts toward the center and radiates outward. The arc length is proportional to the number of questions containing the word. White areas are words with contributions too small to show. This figure demonstrates the complexity and variety of human questions,

Table 1. Quantitative comparison of VQA datasets' complexity

Dataset	Unique words	Unique answers
VQA v2.0 [6]	14576	162496
CLEVR [11]	80	28
HISNav VQA (ours)	868	712

synthetic questions are built on seven structural templates of fixed length, while all possible structures of human questions do not fit in this figure, and their length varies from three to 13 words.

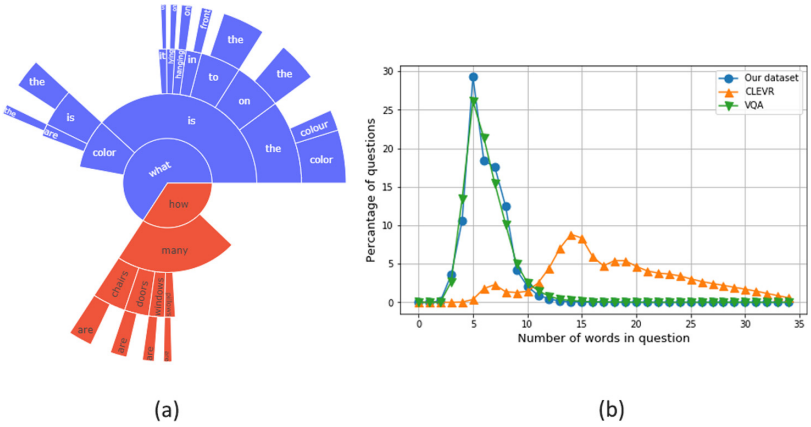

Fig. 3. Comparison of question lengths for different VQA datasets

Figure 3b shows the distribution of question lengths for three datasets: ours, VQA v2.0, and CLEVR. Our dataset and VQA v2.0 turned out to be close since both consist of questions asked by people, and for CLEVR, the distribution turned out to be significantly shifted due to the artificial syntactic complexity of synthetic questions. This fact additionally shows that this dataset is poorly suited for training an assistant who would answer people's questions.

Figure 4 shows various quantitative statistics for our dataset: (a) is the distribution of question types given by Yandex.Toloka workers; (b) is the distribution of answers to questions about color. Due to the specifics of the scenes used, the number of possible colors is also small, but this number is sufficient to test the ability of the model to distinguish the colors of objects, which is important both for the task of navigation and for assistants; (c) is the distribution of answers to count-questions, all the number-answers are in the range from 0 to 9, this is not a large range, but sufficient for room scenes, because there are rare cases when it is needed to count so many objects in rooms; (d) is the distribution of answers

to what-containing questions, which are the most common and important for our purposes.

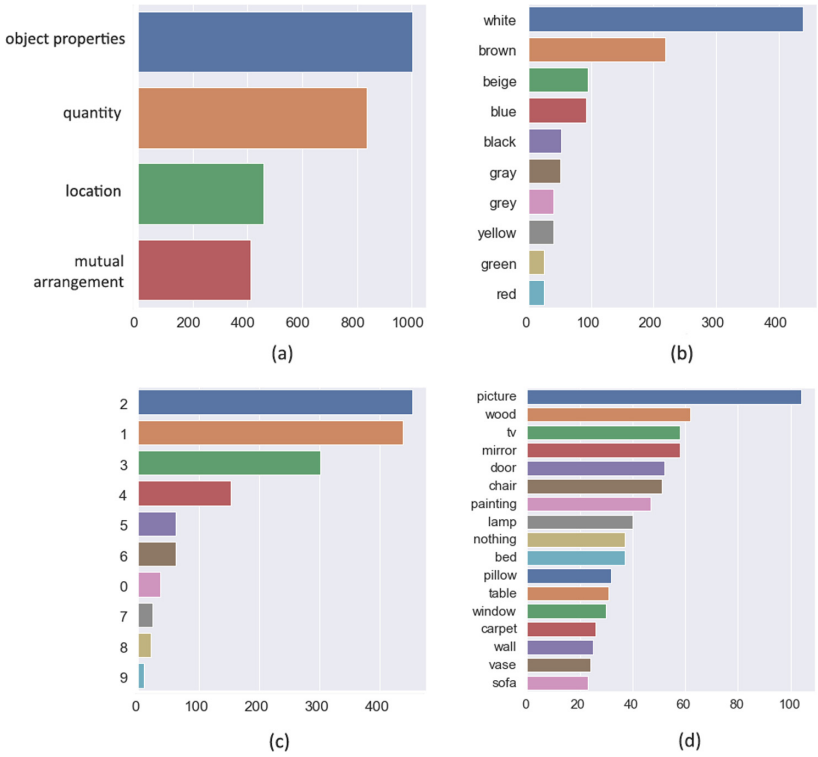

Fig. 4. Quantitative statistics for out dataset

4 Vector Semiotic Architecture Baseline

In this section, we propose a Vector Semiotic Architecture Baseline that addresses the symbol grounding problem and provides an interpretable answering procedure. The Vector Semiotic Architecture model is inspired by NS-VQA [30] and uses the scene representation from [17]. The model consists of three main parts: a scene parser, a question parser, and a program executor. The scene parser uses an instance segmentation model to extract attributes, such as coordinates and color, for each object. The question parser is an attention-based sequence to sequence model [3] that receives a sequence of words in a natural language as input and outputs a sequence of programs for execution. Both the encoder and the decoder have two hidden layers with a 256-dim hidden vector, and the dimension of word vectors is 300. The executor is a collection of functional modules that are sequentially applied to the scene representation to get the answer to a question. The dimension of HD vectors is set to 1,000.

Compared to the work in [17], we use a variance of the Semantic Pointer Architecture (SPA) [5] – Spatial Semantic Pointers [15] – to represent causal matrices and work efficiently with continuous values such as coordinates. Here we use real random vectors with a unit norm. The unique feature of this approach is using special algebraic operations: circular convolution and convolutive power, which is defined as follows

$$\mathbf{u} \otimes \mathbf{w} := IDFT(DFT(\mathbf{u}) \odot DFT(\mathbf{w}))$$

$$\mathbf{u}^p := \Re(IDFT((DFT(\mathbf{u})^p)_{j=0}^{D-1}))$$

where $\mathbf{u}, \mathbf{w}$ are two random vectors, $\Re$ denotes taking the real part of a number, $\odot$ denotes element-wise multiplication, and DFT and IDFT denote the Discrete Fourier Transform and Inverse Discrete Fourier Transform respectively. With these operations, we can encode two numerical values corresponding to the coordinates x, y by generating two random unitary vectors corresponding to the coordinate axes $\mathbf{X}, \mathbf{Y}$ (vector $\mathbf{u}$ is called unitary if $\forall \mathbf{v} : \|\mathbf{v}\| = \|\mathbf{v} \otimes \mathbf{u}\|$):

$$\mathbf{V} = \mathbf{X}^x \otimes \mathbf{Y}^y.$$

Thus, to answer the question "What is the object to the left of the chair?" we have to encode coordinates of a chair into a vector $\mathbf{V}_{chair}$, construct a vector that represents a region left to the chair $\mathbf{V}_{left}$ [15], and compute the similarity between this vector and vectors of other objects coordinates.

5 Experiments

In this section, we test the performance of two baseline models on the HISNav VQA dataset. For human-asked questions, the model's answer is considered correct if it matches at least one of the two ground-truth answers.

5.1 Neural Network Baseline

We implemented a simple neural network baseline for the VQA task. A bag-of-words representation of questions is used as 300-dim vectors. Resnet18 [10] is used to obtain embedding of an image as a vector of dimension 512. Question and image embeddings are concatenated and then passed to a multi-layer perceptron classifier with two layers of 512 hidden units. We used ReLU activation in hidden layers and a softmax for an output layer. The accuracy of this model on synthetic questions is 0.57, and on human questions 0.43. The model was trained using SGD with standard parameters and a batch size of 64. The main problem with this model is that it is sensitive to the bias in our data: the model learns to predict only the most frequent answers to questions.

5.2 Vector Semiotic Architecture

We first trained the VSemA model in a supervised manner on a subsample of synthetic questions (32 questions of each type, 224 in total). This model achieved an accuracy of 0.82 on the validation part of synthetic questions. Further, this model was trained using the REINFORCE [29] algorithm on human questions. The reward was given only for the correct answer, and, as a result, it reached an accuracy of 0.20 on the corresponding validation set.

To obtain the best performance on synthetic data, we used a larger subsample, and, as a result, we got a nearly perfect accuracy of 0.98 (Table 2).

Table 2. Baselines performance (accuracy)

Model	Synthetic questions	Human-asked questions
Our approach	0.98	0.20
Simple NN	0.57	0.43

On human-asked questions, our model demonstrates moderate performance compare to a neural network baseline. Error analysis reveals two main causes. First, our model relies on an instance segmentation mask, and therefore it fails to answer questions about instances that are not segmented. The left half of Fig. 5 depicts a stair with a corresponding question "How many steps are shown on the image?". Our model does not distinguish individual stair steps – it sees them as a whole and, thus, predicts a wrong answer. The second cause also affects the neural network baseline – the ambiguity of questions. In the right half of Fig. 5, both models give wrong answers in terms of ground truth, though predictions are generally correct (both the door and the chair are in the room). This is since there are more than two objects in the room that leads to ambiguity. Other examples of predicted answers are shown in Appendix B.

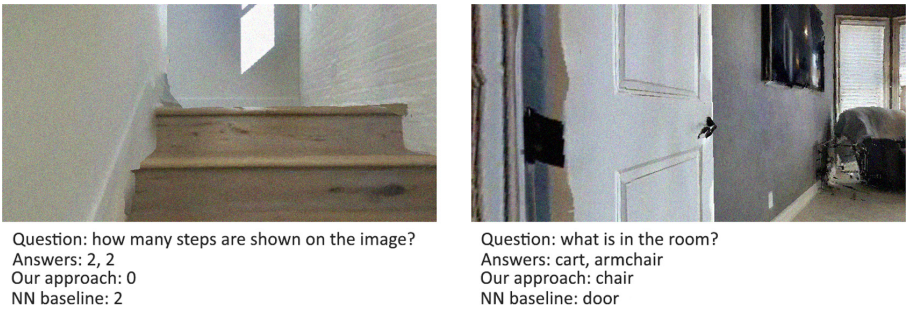

Fig. 5. Examples of Vector Semiotic Architecture model mispredictions

6 Discussion

The proposed HISNav VQA dataset aims to address the needs of Visual Navigation in human-centered environments. The dataset also does not suffer from disembodiedness and unsituatedness and might be used to advance research in the field of assistance robotics or for VLN.

We demonstrated the performance of Vector Semiotic Architecture on a challenging dataset HISNav VQA and achieved nearly perfect performance on synthetic questions. The performance on human-asked questions demonstrates the limitations of our model due to reliance on the instance segmentation mask. This limitation may be addressed by pretraining on a dataset with a large number of classes [7]. On the other hand, our approach provides an interpretable answering procedure compare to a neural network baseline. Also, nearly perfect performance on synthetic questions gives us the ability to construct the domain- and task-specific questions with a high probability of getting the right answer that is crucial for application purposes.

For future work, we plan to increase the performance of Vector Semiotic Architecture on human-asked questions by using datasets with a large number of segmentation classes and exploiting the questions' syntactic structure. We also plan to use HISNav VQA and the proposed model to build a prototype agent that can first operate in a virtual and then in a real-world environment. We hope our work draws researchers' attention to this task, as there are still unresolved problems and challenges.

7 Conclusion

In this work, we propose HISNav VQA – a challenging dataset for a Visual Question Answering task that is based on the Habitat dataset and concentrates on questions about spatial arrangements of the scene objects. The dataset may be used in the scenario of the Vision-and-Language Navigation task in human-centered environments where a robot asks and answers questions to clarify instructions. We also demonstrate the performance of the Vector Semiotic Architecture on the proposed dataset and compare it to a simple neural network baseline.

Acknowledgements. The reported study was supported by RFBR, research Project No. 19-37-90164.

A Appendix: Data Labeling

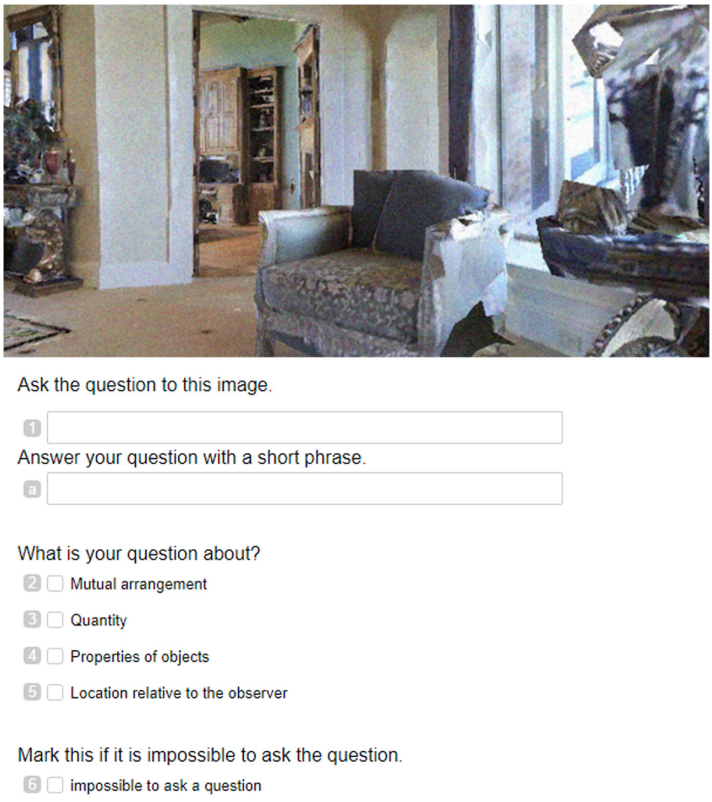

Fig. 6. The user interface for data labeling in Yandex.Toloka

B Appendix: Examples

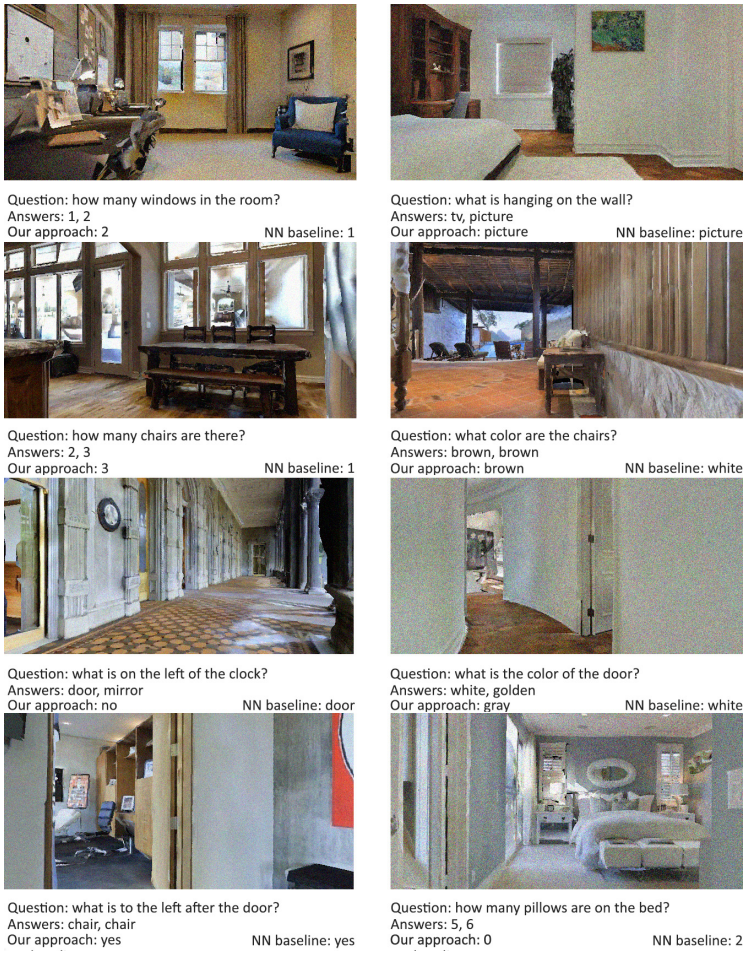

Fig. 7. Examples of predicted answers given by Vector Semiotic Architecture and NN baseline. First row: both models give the right answer. Second row: NN baseline fails. Third row: Vector Semiotic Architecture fails. Last row: both models fail.

References

1. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. arXiv e-prints arXiv:1707.07998, July 2017
2. Anderson, P., et al.: Vision-and-language navigation: interpreting visually-grounded navigation instructions in real environments (2018)
3. Bahdanau, D., Cho, K.H., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, pp. 1–15 (2015)
4. Das, A., et al.: Visual dialog. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

5. Eliasmith, C.: How to Build a Brain: A Neural Architecture for Biological Cognition. Oxford University Press, New York (2013)
6. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: 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)
7. Gupta, A., Dollar, P., Girshick, R.: LVIS: a dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 2019, pp. 5351–5359 (2019)
8. Gurari, D., et al.: VizWiz grand challenge: answering visual questions from blind people. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3608–3617 (2018)
9. Harnad, S.: The symbol grounding problem. Physica D **42**(1), 335–346 (1990)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
11. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.: CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning. In: CVPR (2017)
12. Kanerva, P.: Hyperdimensional computing: an introduction to computing in distributed representation with high-dimensional random vectors. Cogn. Comput. **1**(2), 139–159 (2009)
13. Kiselev, G., Kovalev, A., Panov, A.I.: Spatial reasoning and planning in sign-based world model. In: Kuznetsov, S.O., Osipov, G.S., Stefanuk, V.L. (eds.) RCAI 2018. CCIS, vol. 934, pp. 1–10. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00617-4_1
14. Kiselev, G., Panov, A.: Hierarchical psychologically inspired planning for human-robot interaction tasks. In: Ronzhin, A., Rigoll, G., Meshcheryakov, R. (eds.) ICR 2019. LNCS (LNAI), vol. 11659, pp. 150–160. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-26118-4_15
15. Komer, B., Stewart, T.C., Voelker, A.R., Eliasmith, C.: A neural representation of continuous space using fractional binding. In: 41st Annual Meeting of the Cognitive Science Society. Cognitive Science Society, QC (2019)
16. Kovalev, A.K., Panov, A.I.: Mental actions and modelling of reasoning in semiotic approach to AGI. In: Hammer, P., Agrawal, P., Goertzel, B., Iklé, M. (eds.) AGI 2019. LNCS (LNAI), vol. 11654, pp. 121–131. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-27005-6_12
17. Kovalev, A.K., Panov, A.I., Osipov, E.: Hyperdimensional representations in semiotic approach to AGI. In: Goertzel, B., Panov, A.I., Potapov, A., Yampolskiy, R. (eds.) AGI 2020. LNCS (LNAI), vol. 12177, pp. 231–241. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52152-3_24
18. Ku, A., Anderson, P., Patel, R., Ie, E., Baldridge, J.: Room-across-room: multilingual vision-and-language navigation with dense spatiotemporal grounding. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 4392–4412, November 2020
19. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
20. Linda Smith, M.G.: The development of embodied cognition: six lessons from babies. Artif. Life **11**, 13–29 (2005)
21. Savva, M., et al.: Habitat: a platform for embodied AI research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)

22. Osipov, G.S., Panov, A.I., Chudova, N.V.: Behavior control as a function of consciousness. I. world model and goal setting. J. Comput. Syst. Sci. Int. **53**(4), 517–529 (2014)
23. Osipov, G.S., Panov, A.I.: Relationships and operations in a sign-based world model of the actor. Sci. Tech. Inf. Process. **45**(5), 317–330 (2018). https://doi.org/10.3103/S0147688218050040
24. Osipov, G.S., Panov, A.I.: Rational behaviour planning of cognitive semiotic agent in dynamic environment. Sci. Tech. Inf. Process. **48**(6) (2021)
25. Panov, A.I.: Goal setting and behavior planning for cognitive agents. Sci. Tech. Inf. Process. **46**(6), 404–415 (2019)
26. Panov, A.I.: Behavior planning of intelligent agent with sign world model. Biol. Inspired Cogn. Archit. **19**, 21–31 (2017)
27. Plate, T.A.: Holographic reduced representations. IEEE Trans. Neural Networks **6**(3), 623–641 (1995). https://doi.org/10.1109/72.377968
28. Staroverov, A., Yudin, D.A., Belkin, I., Adeshkin, V., Solomentsev, Y.K., Panov, A.I.: Real-time object navigation with deep neural networks and hierarchical reinforcement learning. IEEE Access **8**, 195608–195621 (2020)
29. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach. Learn. **8**, 229–256 (2004)
30. Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., Tenenbaum, J.B.: Neural-symbolic VQA: disentangling reasoning from vision and language understanding. arXiv e-prints arXiv:1810.02338, October 2018
31. Yu, J., Zhu, Z., Wang, Y., Zhang, W., Hu, Y., Tan, J.: Cross-modal knowledge reasoning for knowledge-based visual question answering. Pattern Recogn. **108**, 107563 (2020)
32. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: visual commonsense reasoning. CoRR abs/1811.10830 (2018)