Applying Vector Symbolic Architecture and Semiotic Approach to Visual Dialog

Alexey K. Kovalev[1,2], Makhmud Shaban[2], Anfisa A. Chuganskaya[1], and Aleksandr I. Panov[1,3]

[1] Artificial Intelligence Research Institute FRC CSC RAS, Moscow, Russia
[2] HSE University, Moscow, Russia
[3] Moscow Institute of Physics and Technology, Moscow, Russia

Abstract. The multi-modal tasks have started to play a significant role in the research on Artificial Intelligence. A particular example of that domain is visual-linguistic tasks, such as Visual Question Answering and its extension, Visual Dialog. In this paper, we concentrate on the Visual Dialog task and dataset. The task involves two agents. The first agent does not see an image and asks questions about the image content. The second agent sees this image and answers questions. The symbol grounding problem, or how symbols obtain their meanings, plays a crucial role in such tasks. We approach that problem from the semiotic point of view and propose the Vector Semiotic Architecture for Visual Dialog. The Vector Semiotic Architecture is a combination of the Sign-Based World Model and Vector Symbolic Architecture. The Sign-Based World Model represents agent knowledge on the high level of abstraction and allows uniform representation of different aspects of knowledge, forming a hierarchical representation of that knowledge in the form of a special kind of semantic network. The Vector Symbolic Architecture represents the computational level and allows to operate with symbols as with numerical vectors using simple element-wise operations. That combination enables grounding object representation from any level of abstraction to the sensory agent input.

Keywords: Visual dialog · Vector symbolic architecture · Sign-Based World Model · Perception

1 Introduction

In recent years, multimodal tasks have attracted increased attention. As an example of such tasks, a Visual Question Answering (VQA) [1] which combines visual and linguistic modalities could be considered. In VQA, a system queried by an image and a question to that image and should output an answer to the given question. This task serves as the starting point for more advanced problems, such as Visual Commonsense Reasoning (VCR) [33] and Visual Dialog [6]. VCR is testing the system's ability to justify its answer by forcing it to choose a rationale for it. Visual Dialog presents an example of a possible scenario of interaction between an intellectual assistant and a user. The crucial feature of the

data that the dataset has collected during the interaction of two agents (Amazon Mechanical Turk workers) with each other. One agent (answerer) is exposed to an image and its caption and its role to answer questions asked by another agent (questioner) who does not see the image but the caption. Thus the questioner implicitly solves the task of refining the representation of a scene depicted on the image. Explicitly, the collected data is used in a situation where the answerer is exchanged with a computer system and asked to answer the last question about the image in the dialog considering dialog history. This peculiarity of the solving problem determines the types of questions asked. They are very clear-cut, as the questioner tries to refine the scene understanding. Most questions ask about external features (color, shape, size, appearance, etc.) and the existence/counting of objects. The symbol grounding problem [10, 4, 23] plays a decisive role in the effectiveness of such systems, as the system should relate the information from the image and the textual data with its internal representation of concepts. However, this problem statement limits the number of question types and does not allow for modeling complex cognitive functions as answers are based directly on the system's sensory input.

In this paper, we approach the symbol grounding problem from the semiotic perspective and apply the Sign-Based World Model [22, 24] cognitive architecture enriched by hyperdimensional computing [13] (vector symbolic architecture) [17]. In the semiotic approach, the information unit is a sign that differs from a symbol in the sense that it possesses an internal structure and a name. Structurally, a sign consists of four components, namely a meaning, significance, an image, and a name. Sign components are represented by causal matrices. Matrices are represented as vectors of high dimensionality via hyperdimensional computing. Signs are not isolated information units but connected into a special kind of hierarchical semantic network – the semiotic network – and on the lowest level are grounded to the agent's sensory input by the image component. The use of Vector Symbolic Architectures allows for reducing operations on signs components to vector operations.

In psychology, the Visual Dialog task relates to perception and the construction of the image of objects based on sensations. The theory of perception can serve as a model for constructing algorithms in artificial intelligence investigations. The contribution of this paper is twofold: first, psychological groundings for such tasks as Visual Dialog are presented, and the Vector Semiotic Architecture is proposed to address the symbol grounding problem in the context of the Visual Dialog task.

2 Related Works

Perception, one of the leading higher mental functions [29] of a person, makes it possible to form some image of an object and subsequently the image of the world. Perception can be understood in two ways: as an image, the result of sensory systems and categorization work, or as a process that is a structure of actions aimed to obtain such an image. Research in the field of activity and the

construction of a system of actions is the most relevant to the development of work algorithms of artificial systems.

The classification of action types and the process of their formation were proposed by Nikolai Bernstein. He identified 5 levels of actions that are characteristic of human activity. Each level was named with a Latin letter: A, B, C, D, E [3]. According to Bernstein, movement levels have their neurophysiological organization. The higher the goal of the action, the higher the level of the corresponding anatomical and physiological organization.

A is the lowest level. This level includes tonic movements, e.g. trembling. They are regulated by simple neurophysiological reactions which are similar both for humans and animals. **B** is the level of coordinated movements. These are coordinated actions without the need for spatial orientation. For example, hand movements while lying on the surface. **C** level needs movement and orientation in space. For example, you need to go around some obstacles. **D** is the subject level that is typical for a person. At this stage, the movements are built according to the logic of the subject. For example, if you need to take a cup without a handle, then a person can find an action consistent with this goal. **E** is the level of the speech muscles movements. This level is carried out when we speak, express our thoughts, or in symbolic movements (dance).

According to the organization of the levels of action, Bernstein proposed the concept of "models of the necessary future". The higher the motive of the activity, the higher the levels connected to its implementation. Levels C and D are significant for the construction of perceptual images.

At the level of perception, actions are associated with the conscious selection of a certain side of a sensory situation and the subsequent categorization of sensory information [32]. Studies of the processes of child perception development show that they are initially included in the external practical actions. The connection of perceptual actions with practical actions (manipulation, movement in space, etc.) is manifested in their expanded motor character, which can be observed externally. In the movements of the hand that feels the object (touch) and in the movements of the eyes that trace the visible contour, there is a continuous comparison of the image with the original, its verification and correction are carried out. In the further development of the activity, there is a reduction in the motor components. This leads to a significant temporary change: the process of perception externally becomes a one-time action. These changes are associated with the formation of a child's system of operations within the framework of perceptual actions and sensory standards. Perceptual actions are implemented using various operations. A similar consideration of the process of constructing a perceptual image is noted in the works of Jean Piaget [25], James Gibson [7], Ulric Neisser [21].

If we consider the process of forming the image of perception in a child of 4-5 years, we can distinguish the following characteristics. The perception is directed for them and carried out in the form of perceptual actions [26, 32]: the shape of the object and the ability to group it according to this attribute; the size of items and the ability to group them; dividing the subject into parts and

vice versa; the ratio of the integral shape of objects and their parts by size; measurements of objects (length, height, width, etc.); color of objects and their parts; the selection of an object from the surrounding environment based on its spatial position relative to other objects, the placement of objects (including the number of parts), based on the knowledge of their position in the space; understanding the similarity-difference relationship; understanding the general-private relationship; understanding of causal relationships-establishing the cause of a particular phenomenon, action, determining the possible consequence of certain actions and place them in the appropriate order; the necessary amount of knowledge about the objects and phenomena of the surrounding world; mathematical representations of the number, geometric shapes, and magnitudes of objects.

In the paper [26], the four stages of perceptual actions of a preschool child were identified in order to build a holistic image of the subject. At the **first stage**, the subject is perceived as a whole. We can say that at this stage there is a comparison of the general characteristics of the object with sensory standards. There is a primary categorization of the object into a certain class. At the **second stage**, the main parts of the object are isolated and their properties (shape, size, color, etc.) are determined. At the same time, the signs that will relate to the main ones will be updated depending on the perceptual attitude, which updates the field of attention and the appropriate signs. At the **third stage**, the spatial relationships of the parts are distinguished relative to each other (above, below, right, left) and to the entire context. At the **fourth stage**, an examination is carried out by repeated holistic perception of the object. All the data obtained about the object properties is analyzed. The results of the performed perceptual actions are synthesized into a single image.

The cultural significance of an object in the social practice of a person or the biological significance of an object in the life of an animal is described in psychology as the objectivity of perception. Experiments in the field of cognitive psychology [28], as well as data from neuropsychology [5] indicate that the recognition of an object occurs not only based on the geometric features of the configuration, but also within the framework of answers to the following questions: “What is it customary to do with this?”and “How can this be used?”. Alexei Leontiev proposed to describe this principle of objectivity through the concept of meaning as the fifth quasi-dimension (existing along with the four dimensions of the space-time continuum), in which the objective world is revealed to a person [19]. The objectivity of perception and mental reflection in general is associated with the use of language for people. Thus, the task that AI specialists set for themselves is completely fundamental to be able to answer questions about the content of the subject scene.

The significant point is the process of analyzing the perceived object and assigning it to a certain class. The question arises about the relationship between the actualization of the perceptual image and semantic information. A number of studies have shown that the selection of semantic features in the pre-adjustment to the process of object perception itself performs a facilitating function if it

meets the principle of semantic expediency. In the research of M. Potter was shown a picture of a hammer. The subject was much faster to name the general semantic category "tool"than when showing the word "hammer"[28].

The process of determining the overall value of the image by narrowing down the field of a diversity of response options occurs simultaneously, and sometimes before the selection of geometric features. In the intermediate phases of the microgenesis of perception, the answer is given to the question: "What does it look like?"[28]. In the perceptual processing of complex realistic images, their general semantic content is highlighted. This is done through the analysis and operation of simple filters that work without feedback. Such images are clustered in the coordinates of the semantic space of the scenario character, for example, "apartment", "forest", "sea coast". The overall meaning of the scene is highlighted before the detailed perception of the individual objects that fill it, providing a quick semantic classification [28].

A significant role of semantic clustering and schematization in the process of image perception is also noted in studies of productive processes of memory and reproduction of observed pictures. In experiments by Frederic Bartlett, the respondents were asked to consider the picture [2] or a story. The first person had the opportunity to see the picture and memorize it. In the future, the second and successive respondents were called. The first person tried to clearly describe its contents to the second, and this description was recorded. It was important what the participants forgot and what the final description was left. All colors except one were immediately forgotten. Random details disappeared. There was a progressive forgetting of unimportant details. However, a few units remain dominant.

Bartlett arrived at the conclusion that the scheme was stored in memory. It is understood as a sequential logical-temporal bundle of images and events. Node components or key events are highlighted in the schema. In the case of memorizing images, this is a spatial scheme. When recreating a scheme, additional details usually refer to the following types: details related to circuit nodes; emotionally charged details; details related to personal experience.

There is a replacement of unfamiliar images with similar ones. When recreating an image, the modifications are based on the schema and script. The actions in the scenario are organized according to a given sequence and are directed at the goal.

One of the directions of modern research of visual perception is the identification of those visual parameters that can act in the construction of a complex (semantic) image. For example, when investigating the problem of increasing the number of informative points used by the visual system in comparison with the number that directly falls on the retina of the eye, in [27] is shown that when an image is perceived, a "field picture"is formed, structured not only by the visible (actually indicated in the image itself) but also by the imaginary, or invisible to the eye axes. Such results determine the prospects of research in the field of the extra-sensory basis of human perceptual activity and the need to focus on it when modeling perception by means of artificial intelligence.

3 Vector Semiotic Architecture for Visual Dialog

In this paper, we apply a Vector Semiotic Architecture to the Visual Dialog task. The Vector Semiotic Architecture is a Sign-Based World Model cognitive architecture enriched with Vector Symbolic Architecture on the low level of representation. Such a combination of approaches attempts to address the symbol grounding problem [10, 4, 23], a fundamental problem of AI that is not solved yet.

The Sign-Based World Model (SBWM) [22, 24] is a framework for modeling cognitive tasks. It bases on principles of the cultural-historical approach of Lev Vygotsky and the activity theory of Alexei Leontiev. SBWM relies on the concept of a **sign** representing the agent's knowledge about the environment it operates in, other agents it interacts with, and itself. The signs form a hierarchical semantic network. Conceptually, the sign is a four-component structure. Four components represent different aspects of the agent's knowledge. The meaning component *(M)* implies the agent's experience. Commonsense knowledge is expressed by the significance component *(S)*. The image component *(I)* is used to distinguish signs. The name *(N)* possesses a nominative function. SBWM is successfully applied to different tasks, e.g., planning [14, 8], role distribution of a group of agents [15], goal setting [24], and reasoning [16].

The *S*, *M*, *I* components of a sign represented by a special structure called a **causal matrix** (CM). A CM z is defined as a tuple of length t of events e_i. Each event e_i represents the appearance of a particular feature f_j at time step i and is a binary vector of length h. Thus a CM is a binary h by t matrix. The 1 in the position z_{ji} in a CM serves as a link to other matrices. That corresponds to the feature f_j and means that the feature f_j is included in the appropriate component of the sign. Thereby CMs are organized into a hierarchical semantic network where the tuples of CMs are the nodes, and the links are the relations between these tuples. The event index t can serve as discrete time to represent dynamic entities.

Vector Symbolic Architectures (VSA) [13] is an umbrella term for bio-inspired methods of representing and manipulating concepts as vectors of high dimensionality (HD vectors). HD vectors use distributed representations, i.e., the information is distributed across vector positions, and only the whole HD vector can be interpreted as a holistic representation of some entity.

For each concept of interest, an atomic HD vector is generated by sampling vector space. Atomic HD vectors are stored in the item memory (IM). With an extremely high probability, all random HD vectors are quasi-orthogonal to each other, which is an important property of high-dimensional spaces. Hyperdimensional computing defines operations and a similarity measure to manipulate atomic HD vectors. Two key operations for computing with HD vectors are bundling and binding. The nature of a vector space could be different for the different realizations of VSA. In this paper we work with bipolar vectors $\mathcal{S} \in \{-1, +1\}^{[d \times 1]}$.

The **binding** binds two HD vectors together and produces another HD vector that is dissimilar to the bounded HD vectors. The semantic interpretation of this

operation is assigning a value to a particular attribute. The Hadamard product is used for the binding operation. The **bundling** is implemented via a position-wise addition. The bundling combines several HD vectors into a single HD vector. The resultant HD vector is similar to all bundled HD vectors. The bundling is used to represent sets.

To map the SBWM structures to VSA operations, we use the approach proposed in [17]. We use **bold font** to denote an HD vector of a corresponding SBWM structure and $\mathbf{H}$ with an appropriate subscript to show that this HD vector is stored in the IM.

As a causal matrix z could be represented as a set of events e_i, then a suitable VSA operation is a bundling. First, we have to map every event e_i to a corresponding HD vector $\mathbf{H}_{e_i}$ and then apply bundling to the collection of vectors $\mathbf{H}_{e_1}, \mathbf{H}_{e_2}, ..., \mathbf{H}_{e_t}$. To represent a link from a causal matrix z_1 to a causal matrix z_2 we, first, transform z_2 to an HD vector $\mathbf{z}_2$, second, split z_1 into events $e_i^{z_1}$, and map them to HD vectors $\mathbf{H}_{e_i^{z_1}}$, and then bind $\mathbf{z}_2$ with a corresponding HD vector $\mathbf{H}_{e_i^{z_1}}$. From the perspective of VSA, each symbol is seen as a high-dimensional vector. Then we can operate on symbols using vector operations and easily switch between representations using item memory $\mathbf{H}$.

The proposed model is in Figure 1. The image and the question are processed separately. Objects with attributes are extracted from the image. SBWM Module uses this information to construct a hierarchical scene representation based on SBWM via causal matrices. Then, the scene representation is encoded into an HD scene vector $\mathbf{z}_{scene}$. An input question, a caption, and a dialog history are fed into the module, where, if necessary, coreference is resolved. Then, the Seq2seq Parser parses the question into a sequence of VSA procedures (program $\mathcal{P}$). Each procedure is a combination of binding, bundling, and similarity operations. Procedures serve a specific purpose, e.g., find an object with a particular attribute value or the value of an object attribute. After that, the VSemA Reasoner executes the program on the scene representation and outputs the answer.

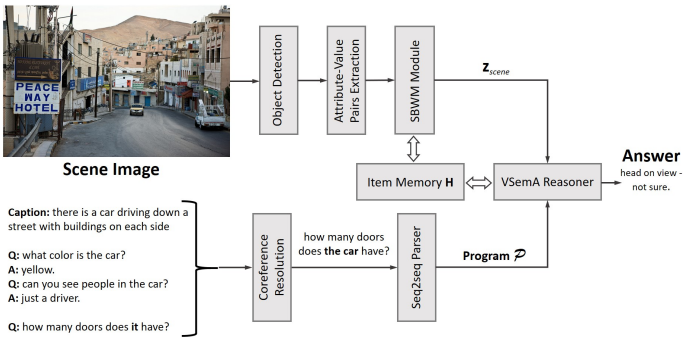

Fig. 1: Vector Semiotic Architecture for Visual Dialog.

4 Experiments

In the Visual Dialog task, the asked question relies on the dialog history. Thus to successfully provide an answer, the model has to consider it. In the proposed approach, we use the coreference resolution to process the dialog history and work with the questions independently. Coreference resolution aims at solving the problem of finding all expressions that correspond to the same entity in the text. We replace the pronouns that refer to the objects mentioned in the previous questions with corresponding nouns. It enables question parsing without relying on the history more than using it for coreference resolution.

To replace pronouns, we use a Huggingface coreference resolution model (NeuralCoref[4]), an AllenNLP library[5], and a simple rule-based replacement. AllenNLP model relies on the approach from [18] with the SpanBERT [12] word embeddings. The rule-based replacement is done by using the spaCy[6] parser to extract part-of-speech tags for each word. The rule was to replace every pronoun in a question with an object from the previous question. We present the quality comparison of used coreference resolution methods in Table 1.

Leaving a pronoun in a sentence makes it impossible for the model to answer correctly. Thus it is better to replace more pronouns at the cost of the percentage of wrong replacements. We chose the AllenNLP model as a final coreference resolution model as it copes well with that task.

Questions	GT	Rule-Based	Huggingface	AllenNLP
what color are bikes? are **they** parked on stock parking?	bikes	they	**bikes**	they
do **they** pose to picture?	people	helmets	their	their
do you see any buildings? are **they** single story?	buildings	**buildings**	any buildings	two
are there other people in the photo? what color are **they** painting with?	person	photo	what color	**person**

Table 1: Results of applying coreference resolution to different dialogues.

Visual Dialog is a real-world dataset, which means that the questions asked in the dialogues are not standardized (e.g., compared to CLEVR [11]). Thus there is no straightforward way to convert a question to a sequence of template procedures, which will produce the answer if executed. Therefore to demonstrate the proposed approach, in this paper, we narrowed down types of questions to existence (Is there an object?) and counting (How many objects are there?).

[4] https://github.com/huggingface/neuralcoref
[5] https://docs.allennlp.org/models
[6] https://spacy.io

The MSCOCO[20] dataset annotation contains 80 classes, which do not cover the entirety of Visual Dialog dataset classes. Therefore we chose a subset of Visual Dialog questions that are about the objects represented in the annotations. There are 20,290 existence questions and 12,472 counting questions. We applied a training pipeline from [31]: the question parser is pretrained on a small subset of question-program pairs in a supervised manner, then REINFORCE [30] is used to fine-tune the parser on the question-answer pairs. There are no program annotations in the Visual Dialog dataset, and we annotated questions manually. It resulted in a total of 39 question-program pairs. Even such a small amount is sufficient to train the model successfully.

We chose the NS-VQA model [31] as a baseline. It achieved an overall accuracy of 50.7%, where 31.7% is the accuracy for counting questions and 71.7% is for existence questions. For Vector Semiotic Architecture, we used HD vectors of size 10000 for scene representation. The proposed model achieved 51.3% accuracy, where the accuracy for counting questions is 31.0%, and the accuracy for existence questions is 73.9%.

5 Discussion

In the experiment, various paintings depicting objects, people, and animals were used. From a psychological point of view, the experiment combines constructing and transmitting a perceptual image. The study design brings it closer, in fact, to the first stage of Bartlett's experiments on productive memory work. The study participant is given a caption as a short description of the image. This semantic framework is the beginning of the construction of a perceptual image. In Example 1 (Figure 2a), a caption will be "a couple of men riding horses on top of a green field". As we noted in the studies of the microgenesis of perception, this will be the initial focus in a certain scenario – "dressage on the field". In Example 2 (Figure 2b), we focus on the scenario field "kitchen, dining room"or "catering establishment: restaurant, cafe". Such a process outlines the range of possible objects that can semantically be in such a field of perception. Within the framework of the human psyche, this process acts as an analog of presetting, which significantly reduces the process of identifying the image.

In the further course of the experiment, a dialogue takes place in the form of an answer to questions that will help reconstruct the perceptual image. From a psychological point of view, such activity carries out the process of recognition, successively passing through separate operations of perceptual actions. In the experiment, a model is created at the C level and partially at the D level, according to Bernstein's typology. So, in Example 1, this is the answer to the question about the number of horses. It is necessary to select an object by identifying it by its contour to implement this perceptual action, i.e., primitive counting based on the visual image. This task implements the principle of sensory standards. The system recognizes an image that corresponds to such generalized examples. A mechanism that models the recognition process based on sensory standards is created due to the training of recognition of a particular image. In Example 2,

color recognition becomes a perceptual action. There is also a search and correlation with color standards to answer that the napkin and tablecloth are white. The design of the question-answer part of the experiment implies the division of the holistic perceptual action of recreation into separate operations. This scheme is most fully classified in the works on child psychology and is described above. In our experiment, we chose three types of operations, i.e., the detection of the object and its name, color, and quantity. They correspond to simple perceptual operations. It is clear that the task of creating an artificial psyche, which imple-

(a) **Caption**: a couple of men riding horses on top of a greenfield.

Q: how many horses are there?
A: 2 horses.

(b) **Caption**: a piece of chocolate dessert on a plate with a napkin and a fork.

Q: what color is the napkin?
A: white.
Q: is it on a table?
A: probably, it's a close up.
Q: is there a stack of napkins?
A: no, just one.
Q: what is the desert?
A: chocolate layer cake.
Q: what color is the plate?
A: white.
Q: how many layers are on the cake?
A: 8 very thin.

Fig. 2: Examples from Visual Dialog [6].

ments the principle of objectivity of mental reflection, has no solution, and it can hardly be posed by anyone seriously. However, the implementation of this principle in problems with a limited set of "names"and associated "images"(spatial-color configurations), "values"(everyday scenarios where the objects presented in the picture play their roles), and "meanings"(evaluative attitude to these objects) seems quite possible. The organization of "recognition"should be carried out "from top to bottom": from the meaning of the entire scene to the properties description of individual objects. The reasons for such a sequence in the algorithm that models human perception are found in two areas of experimental psychology, i.e., in the study of the microgenesis of visual perception and the research of attribution of motives.

One limitation of a proposed approach, that it extracts objects based on annotated classes, and scanty annotations lead to poor performance. As a solution, an instance segmentation model trained on datasets with many categories [9] could be used.

6 Conclusion

In this paper, we propose the Vector Semiotic Architecture for Visual Dialog. The combination of SBWM and VSA in the Vector Semiotic Architecture allows approaching the symbol grounding problem by connecting high-level representations of concepts with sensory inputs of an agent. The proposed architecture achieved 31.0% accuracy for count questions and 73.9% accuracy for existence questions on a subset of the Visual Dialog dataset. Also, we show the limitations of an existing dataset for the Visual Dialog task.

Acknowledgements. The reported study was supported by RFBR, research Projects No. 19-37-90164 and 18-29-22027.

References

1. Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C.L., Batra, D., Parikh, D.: VQA: Visual Question Answering. arXiv e-prints arXiv:1505.00468 (May 2015)
2. Bartlett, F.C.: Remembering: A study in experimental and social psychology. Philosophy **8**(31), 374–376 (1932)
3. Bernstein, A.N.: On dexterity and its development [in Russian]. Publishing House "Physical Culture and Sport", Moscow (1991)
4. Besold, T.R., Kühnberger, K.U.: Towards integrated neural–symbolic systems for human-level ai: Two research programs helping to bridge the gaps. Biologically Inspired Cognitive Architectures **14**, 97 – 110 (2015)
5. Chomskaya, E.D.: Neuropsychology. 4th edition [in Russian]. Peter (2005)
6. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M., Parikh, D., Batra, D.: Visual Dialog. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
7. Gibson, J.: The Perception of the Visual World. : Houghton Mifflin, Boston (1950)
8. Gorodetskiy, A., Shlychkova, A., Panov, A.I.: Delta Schema Network in Model-based Reinforcement Learning. In: Goertzel, B., Panov, A., Potapov, A., Yampolskiy, R. (eds.) Artificial General Intelligence. AGI 2020. Lecture Notes in Computer Science. vol. 12177, pp. 172–182. Springer (2020)
9. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition **2019-June**, 5351–5359 (2019)
10. Harnad, S.: The symbol grounding problem. Physica D: Nonlinear Phenomena **42**(1), 335 – 346 (1990)
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. Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O.: Spanbert: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics **8**, 64–77 (2019)
13. Kanerva, P.: Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors. Cognitive Computation **1**(2), 139–159 (2009)

14. 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.) Artificial Intelligence. pp. 1–10. Springer International Publishing, Cham (2018)
15. Kiselev, G.A., Panov, A.I.: Synthesis of the behavior plan for group of robots with sign based world model. In: Ronzhin, A., Rigoll, G., Meshcheryakov, R. (eds.) Interactive Collaborative Robotics. pp. 83–94. Springer International Publishing, Cham (2017)
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.) Artificial General Intelligence. pp. 121–131. Springer International Publishing, Cham (2019)
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.) Artificial General Intelligence. pp. 231–241. Springer International Publishing, Cham (2020)
18. Lee, K., He, L., Zettlemoyer, L.: Higher-order coreference resolution with coarse-to-fine inference. In: NAACL-HLT (2018)
19. Leontiev, A.N.: Psychology of the image [in russian]. Vestn. Mosk. un-ta. Ser. 14, Psychology. No. 2 pp. 3–13 (1979)
20. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
21. Neisser, U.: Cognition and reality: principles and implications of cognitive psychology. W. H. Freeman and Company (1976)
22. Osipov, G.S., Panov, A.I., Chudova, N.V.: Behavior control as a function of consciousness. i. world model and goal setting. Journal of Computer and Systems Sciences International **53**(4), 517–529 (Jul 2014)
23. Osipov, G.S.: Signs-based vs. symbolic models. In: Sidorov, G., Galicia-Haro, S.N. (eds.) Advances in Artificial Intelligence and Soft Computing. pp. 3–11. Springer International Publishing, Cham (2015)
24. Panov, A.I.: Goal setting and behavior planning for cognitive agents. Scientific and Technical Information Processing **46**(6), 404–415 (Dec 2019)
25. Piaget, J.: Les mécanismes perceptifs [in french]. Presses universitaires de France, Paris (1961)
26. Poddyakov, N.N.: Features of mental development of preschool children [in Russian]. Professional Education Publishing House, Moscow (1996)
27. Shapoval, A.V.: Description of the image structure in modern art criticism analysis [in russian]. Izvestiya Samarskogo nauchnogo tsentra Rossiyskoy akademii nauk **13**(2), 240–246 (2011)
28. Velichkovsky, B.M.: Cognitive science: Fundamentals of the psychology of cognition. In 2 volumes [in Russian]. Smysl/Akademiya, Moscow (2006)
29. Vygotsky, L.: Collected works in 6 volumes. Volume 3 [in Russian]. Pedagogika, Moscow (1983)
30. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning **8**, 229–256 (2004)
31. 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 (Oct 2018)
32. Zaporozhets, A.V., Lisina, M.I.: Development of perception in early and preschool childhood [in Russian]. Prosveshchenie Publishing House, Moscow (1966)
33. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. CoRR **abs/1811.10830** (2018)