National Research University Higher School of Economics

as a manuscript

ALEXEY K. KOVALEV

METHODS AND ALGORITHMS OF NEURO-SYMBOLIC SCENE REPRESENTATION IN MULTIMODAL TASKS

PhD Dissertation Summary

for the purpose of obtaining academic degree
Doctor of Philosophy in Computer Science

Moscow — 2022

The PhD dissertation was prepared at National Research University Higher School of Economics.

Academic Supervisor: Aleksandr I. Panov, Candidate of Science, Docent, Federal Research Center "Computer Science and Control" of Russian Academy of Sciences.

DISSERTATION TOPIC

In recent years artificial intelligence (AI) systems have achieved significant results in many areas, including natural language processing (NLP) tasks, e.g., text translation [68], text generation [62], sentiment analysis [11], computer vision (CV) tasks, e.g., image classification [40], object detection [23], image generation [18], style transfer [15], and others. A common feature of these tasks is the use of one **modality**, where **modality** specifies a representation format in which a particular type of information is stored, i.e. texts for NLP and images for CV.

Having achieved such results, the direction of research began to shift towards the combination of modalities, as a result of which new interesting and challenging problems arise. Such problems are called **multimodal**. The most widespread tasks are those that combine two modalities, e.g., text and image (Image Captioning [69], Visual Question Answering (VQA) [1], Vision-Language Navigation (VLN) [4], Visual Dialog [10], Visual Commonsense Reasoning [75]), video and text (Video Captioning [76], Video Question Answering [70]), audio and image (Audio-Visual Navigation [9]), and others. There are also tasks that involve more than two modalities, e.g., image, semantic mask, and depth map (Place Recognition and Localization [22]).

In this thesis, we concentrated mainly on tasks that combine the modalities of text and image. Such combination is not random if we consider a particular type of an AI system – an intelligent embodied agent that is based on the idea that "intelligence emerges in the interaction of an agent with an environment and as a result of sensorimotor activity" [43]. The main application of such systems is assistance and interaction with the user. These tasks attract the research community, and a lot of attention is paid to the challenges, e.g., within the Embodied AI workshop[1], ALFRED[2] [65] and TEACh[3] [55] challenges are held. To perform these tasks, the system must receive information about the environment (image) and interact with the user using a convenient interface (natural language, text). Possible tasks that may be set for such a system are the clarification of information about the environment (VQA [1]), including help to visually impaired people [20, 19]. This clarification can be expressed by dialogue (Visual Dialog [10]), navigation by instructions (VLN [4]), and others.

In order to successfully perform these tasks, an intelligent agent must present the information received from the environment conveniently. The tasks considered in this dissertation operate in

[1]https://embodied-ai.org/

[2]https://askforalfred.com/EAI22/

[3]https://teachingalfred.github.io/EAI22/

a fixed state of the environment; we will call such a fixed state of the environment a **scene**. The format in which the agent presents the scene will be called the **scene representation**. Depending on the task, different representations of scenes are used, e.g., 3D point clouds, a depth map, a semantic mask, or just an image for the navigation task. Moreover, some tasks do not involve explicit representation of the scenes, e.g., VQA. The most popular solution at the moment is the use of transformer architectures [42, 66] for which it is possible to select the input representation of the scene, but not the internal one.

In the general case, systems based on artificial neural networks represent the scene as a feature vector (embedding) that is transmitted from one layer of the network to another. Such systems currently demonstrate the best results in many tasks. Nevertheless, the embedding representation of the scene has disadvantages. First, it is impossible or difficult to obtain a description of the scene as a set of objects. Second, the process of solving the problem using such a representation is difficult to interpretation, which, third, leads to the complexity of system debugging.

The symbolic representations of the scene do not have these shortcomings. Such representations are well interpreted and allow us to evaluate the process of problem solving. On the other hand, symbolic representations are poorly transferred from one domain to another, requiring a preliminary definition of all the elements included in the scene. Another important drawback is the symbol grounding problem [21]: how the symbolic representation correspond to the elements of the environment, i.e., there is no internal interpretation of symbols within the system.

A possible solution to the symbol grounding problem is neuro-symbolic integration [7], which combines the advantages of the artificial neural networks and symbolic approaches. The main idea of such a combination is to preserve the advantages of both approaches while compensating for the disadvantages. The neuro-symbolic approach is used in many tasks and is gaining popularity among researchers [14, 71, 72, 49, 12, 51]. Thus, the development of methods and algorithms for the neuro-symbolic representation of a scene in multimodal tasks is an urgent problem.

In this thesis, we propose to use a combination of approaches to represent the agent's knowledge at high and low levels of abstraction to achieve neuro-symbolic integration when describing a scene.

At a high level, we propose to employ a semiotic approach [54] that allows structural representation of the agent's knowledge. We use a part of the Sign-Based World Model [53, 52, 58], which is responsible for storing declarative information as a specific implementation of the semiotic approach. The Sign-Based World Model is the representation of the agent about the external environment, its characteristics, goals, motives, other subjects, and operations carried out

based on these representations. The main element of the Sign-Based World Model is the sign, a four-component structure. The components are name, image, meaning, and significance. Each component serves to describe a certain side of the entity. The image component serves to distinguish between entities. The significance represents the knowledge about the entity available to the whole group of agents. Meaning encodes the experience of a particular agent. The name performs a nominative function. In this thesis, we mainly use name and image components of a sign to represent a scene. We also use significance in its rational interpretation, i.e., a category system, as the connections between names are determined by this component.

Such higher cognitive functions as goal-setting [53], planning [56, 58], and role distribution [32] have already been modeled within the framework of the Sign-Based World Model. However, methods for scene representation for multimodal tasks and algorithms to work with have not been addressed yet. The main advantages of the model are that, first, it allows the scene to be represented in a structured hierarchical form, and second, it allows the lower levels of the hierarchy to be grounded to the agent's sensory inputs through causal matrices.

At a low level, we propose to use Vector Symbolic Architectures (VSA) [26] that operate with high-dimensional vectors to represent scenes. High-dimensional vectors are treated as symbols, which makes it possible to reduce operations on symbols to operations on vectors. Thus, VSA facilitates neuro-symbolic integration, provided that high-dimensional vectors are either obtained using neural network approaches, or are themselves fed into artificial neural networks.

As a result of the study, we developed neuro-symbolic methods for scene representation, an apparatus for neuro-symbolic integration based on the Sign-Based World Model and VSA. We also proposed Vector Semiotic Model that uses developed neuro-symbolic scene representation in such tasks as VQA and Visual Dialog. We carried out experiments on the use of neuro-symbolic representation of a scene in multimodal tasks.

The object of the research is the field of artificial intelligence and, more specifically, the knowledge representation. The subject of the research is the neuro-symbolic representation of scenes in multimodal tasks.

Purpose and objectives of the study

The main goal of the thesis research is the development of methods and algorithms for the neuro-symbolic representation of the scene in multimodal tasks. The following tasks are formulated to achieve the goal:

1. to develop methods and algorithms for representing scenes and spatial reasoning in the Sign-Based World Model,

2. to conduct neuro-symbolic integration to represent scenes for multimodal tasks in the Sign-Based World Model using Vector Symbolic Architectures,

3. to develop methods and algorithms for solving multimodal tasks based on the neuro-symbolic scene representation,

4. to conduct experiments to test methods and algorithms for neuro-symbolic scene representation in multimodal tasks.

The degree of development of the research topic

Currently, to solve multimodal problems and represent scenes for them, neural network methods are mainly used. Scenes (images) are represented as feature vectors (embeddings). Such vectors are obtained either from the output of a convolutional neural network [47, 45, 77, 63, 8] or as a result of the work of the attention mechanism [3, 27, 74] or from the output of a transformer [44, 42, 67, 66]. These representations are usually used to predict the response and are not interpretable.

Despite the above, there are works in the neuro-symbolic paradigm [5, 72, 49, 60, 12] that aim not only to solve the problem but also to get an interpretable representation of the scene and the process of question answering. For example, in [72], the scene is represented as a table of objects with a description of their properties. The prediction of the answer comes down to filtering the corresponding cells or comparing the value of the cells.

Of interest are papers on vector symbolic architectures that use high-dimensional vectors to represent scenes. A model for VQA is proposed in [50], and in [73, 33], problems very close to VQA are solved.

In [50], the authors used a simple synthesized dataset with 2D scenes. Each scene contains two geometric shapes in four different shapes and colors. The geometric figures can be in the image in four different positions. The dataset contains all possible combinations of these positions and shapes (3072 images in total). The authors utilized a fully connected artificial neural network with two layers to predict the VSA description of a query image. There were five question templates in

the conducted experiments. The model shows satisfactory results, but the simplicity of the test environment makes it difficult to generalize these results to more complex problems.

In [33], VSA is used to represent stimuli, i.e., groups of patterns and colors, in a maze in which honey bees learn. The proposed episode (scene) encoding is suitable for multimodal tasks since it takes into account objects and their relationships.

In [73], the authors use VSA operations to infer HD representations of images from the CIFAR-10 dataset [39]. HD representations are taken from the hidden layer of the autoencoder, converted to binary form using a threshold function, and passed to the cellular automaton. The evolution of the cellular automaton is calculated according to the chosen rule. The union of several automaton states serves as a high-dimensional vector. Logical reasoning is performed by a series of VSA operations.

Vector symbolic architectures, also known as hyperdimensional computing [26], are a family of methods for representing and manipulating concepts in a high-dimensional space. High fixed-dimensional vectors are the basis for representing information in vector symbolic architectures. These vectors are called high-dimensional vectors or HD vectors. A distributed representation is used to encode information. A distributed representation differs from localist representations, which are commonly utilized for computations, in that no subset of the vector's coordinates can be interpreted. In other words, a specific HD vector coordinate cannot be interpreted. Only the entire high-dimensional vector carries information about a concept as a complete representation.

The encoding of a concept (symbol) using the HD vector is performed by sampling a random vector. Such vectors are called seed HD vectors and are stored in the item memory. Vectors of a more complex structure are constructed from seed vectors using vector operations. For simplicity, the further description will be for the case of bipolar vectors ($\mathbf{H}_{\mathcal{S}} \in \{-1, +1\}^{[d \times 1]}$). An important property of high-dimensional spaces is that, with an extremely high probability, all random HD vectors are dissimilar (quasi-orthogonal).

To operate on seed HD vectors, VSA defines special operations and a measure of similarity for HD vectors. For bipolar vectors, the cosine distance is used. The three key operations for computing with HD vectors are binding, bundling, and permutation.

The binding operation is used to bind two HD vectors together. The result of binding is another HD vector. For bipolar vectors, binding is implemented using a component-wise product (Hadamard product). An important property of the binding operation is that the resulting HD vector is dissimilar to the HD vectors being bound, i.e., the cosine similarity distance is approxi-

mately equal to 0. Binding can be interpreted as assigning a value to a feature.

If it is necessary to obtain a quasi-orthogonal vector without multiplication by another vector, i.e., only one vector should be used in the operation, a permutation (rotation) of the coordinates of the HD vector is applied. It is convenient to use a fixed permutation (e.g., rotation) to relate a character's position in a sequence to an HD vector representing the character at that position.

The bundling operation is implemented by component-wise addition with a threshold. If the result of the summation exceeds the threshold, it is replaced by the threshold value. The bundling operation combines several HD vectors into one HD vector. Unlike the binding and permutation, the resulting HD vector is similar to all bundled HD vectors, i.e., the cosine distance between them is greater than 0. Bundling helps to represent sets of objects. In combination with binding each vector in the set with a corresponding copy of the order vector, shifted (rotated) by a certain number of coordinates, an ordered set of symbols can be obtained.

Bipolar vectors are not the only ones that can be used to build a VSA. Variations of VSA have many different names: Holographic Reduced Representation/HRR [59], Multiply-Add-Permute (MAP) [16], Binary Spatter Codes [25], Sparse Binary Distributed Representations (SBDR) [61], Sparse Block-Codes [41]. All of these models target specific computing hardware but have similar computing properties.

The Sign-Based World Model [53, 57] is a framework for modeling cognitive tasks. It is based on the concept of a sign representing an agent's knowledge of the environment in which it operates, other agents with which it interacts, and itself. Signs are organized into a hierarchical semantic network called the semiotic network. Conceptually, the sign is a four-component structure. The four components are image, meaning, significance, and name. They represent different aspects of the agent's knowledge; the meaning component implies the agent's experience; the significance component represents publicly available information; the image component is used to distinguish signs; the name has a nominative function. The sign components themselves constitute semantic networks based on meanings, meanings, images, and names.

Depending on the problem being solved, one sign component may play a more or less important role than the others. For example, when solving planning problems [2, 30, 31], tasks of distributing roles in a group of agents [32], or goal setting [57], the components of significance and meaning come to the fore, since it is necessary to take into account the experience of agents and the rules of the environment. In the problem of reasoning [30, 35], meaning and image play a major role. In the context of the symbol grounding problem, the crucial role is played by the image component,

which implements the recognition function and allows associating signs with the output signals of the agent's sensors.

KEY RESULTS

1. **A representation of scenes has been developed for the problem of spatial reasoning in the Sign-Based World Model.** In [30], the scene representation and reasoning process is presented in the case of determining the relative position of an agent and an object. In [35], the general case of scene representation as an enumeration of objects and their relations is considered. Also mental actions(generalization, concretization, abstraction, etc.) are proposed with the help of which reasoning is carried out in the Sign-Based World Model.

2. **A method of neuro-symbolic representation of a scene for multimodal tasks was proposed.** [36] proposed the representation of the scene as a high-dimensional vector based on the semantic description of the scene obtained from the Sign-Based World Model and [38] developed it.

3. **Algorithms that implement mental actions by VSA operations have been developed for determining the properties and relations of objects from the neuro-symbolic scene representation.** In [38], software modules that implement mental actions (such as in [35]) have been developed based on operations with high-dimensional vectors, which allow extracting representation information (properties, relations) from the neuro-symbolic representation of the scene that is used to solve multimodal tasks.

4. **A new model for solving multimodal tasks has been developed.** In [38], a new model for the multimodal task based on the neuro-symbolic representation of the scene is developed. The model represents a scene image as an HD vector and performs mental actions in the form of VSA operations on it to obtain an answer. In [38], the model was used for the VQA task; experiments were carried out on the CLEVR dataset [24]. The proposed architecture demonstrates accuracy comparable to the state-of-the-art models [72]. We also demonstrate performance of the model on Human-Centered Environment VQA in [28]. In [37], the model is adapted for the multimodal task of Visual Dialog[10].

5. **The neuro-symbolic scene representation for place recognition was proposed.** In [29], the neuro-symbolic scene representation is used in the two-staged model TSVLoc

for place recognition. The model combines two types of embeddings: 1) embedding from any traditional method (HF-Net, NetVLAD, etc.) 2) semantic embedding constructed from a semantic map of the scene and depth map by the means of Vector Symbolic Architectures. Experiments have shown that the TSVLoc model of semantic place recognition significantly improves previous methods based on the popular neural network models HF-Net and NetVLAD for HPointLoc[4] and Oxford RobotCar [46] datasets.

Personal contribution to the aspects/ideas to be defended

1. **A representation of scenes has been developed for the problem of spatial reasoning in the Sign-Based World Model.** I developed a scene representation method and procedure for spatial reasoning in [30]. In [35], I devised a scene representation and algorithm for reasoning based on mental actions. Besides, I formalized the reasoning process and contributed to the development of the Sign-Based World Model.

2. **A method of neuro-symbolic representation of a scene for multimodal tasks was proposed.** The first modification of the scene representation with a toy example was proposed in [36]. The final variant was proposed in [38].

3. **Algorithms that implement mental actions by VSA were developed for determining the properties and relations of objects from the neuro-symbolic scene representation.** The algorithms were developed by me and implemented as program modules. These modules were used in [38, 28, 37] to predict answers for visual questions.

4. **A new model for solving multimodal tasks has been developed.** The overall architecture was developed by me and proposed in [38]. The variances of architecture were used in [28, 37]. I contributed to experiments conducted in [38] on the CLEVR dataset [24], in [37] conducted on Visual Dialog [10], and in [28] conducted on Visual Question Answering in Human-Centered Environments.

5. **The neuro-symbolic scene representation for place recognition was proposed.** The neuro-symbolic scene representation in [29] based on Vector Symbolic Architectures for place recognition. I also contributed to the development of the overall TSVLoc architecture.

[4]https://github.com/cds-mipt/HPointLoc

PUBLICATIONS AND APPROBATION OF RESEARCH

Below is the list of conferences, workshops, and journals where the main results of the research were presented.

First-tier publications:

1. Kovalev A.K., Shaban M., Osipov E., Panov A.I., Vector Semiotic Model for Visual Question Answering, Cognitive Systems Research, Volume 71, 2022, Pages 52-63, ISSN 1389-0417, https://doi.org/10.1016/j.cogsys.2021.09.001 **(WoS Q1, Scopus Q2)**.

Second-tier publications:

1. Kovalev, A.K., Panov, A.I. (2019). Mental Actions and Modelling of Reasoning in Semiotic Approach to AGI. In: Hammer, P., Agrawal, P., Goertzel, B., Iklé, M. (eds) Artificial General Intelligence. AGI 2019. Lecture Notes in Computer Science, vol 11654. Springer, Cham. https://doi.org/10.1007/978-3-030-27005-6_12 **(Scopus Q2)**.

2. Kovalev, A.K., Shaban, M., Chuganskaya, A.A., Panov, A.I. (2021). Applying Vector Symbolic Architecture and Semiotic Approach to Visual Dialog. In: Sanjurjo González, H., Pastor López, I., García Bringas, P., Quintián, H., Corchado, E. (eds) Hybrid Artificial Intelligent Systems. HAIS 2021. Lecture Notes in Computer Science, vol 12886. Springer, Cham. https://doi.org/10.1007/978-3-030-86271-8_21 **(Scopus Q2)**.

3. Kirilenko, D.E., Kovalev, A.K., Osipov, E., Panov, A.I. (2021). Question Answering for Visual Navigation in Human-Centered Environments. In: Batyrshin, I., Gelbukh, A., Sidorov, G. (eds) Advances in Soft Computing. MICAI 2021. Lecture Notes in Computer Science, vol 13068. Springer, Cham. https://doi.org/10.1007/978-3-030-89820-5_3 **(Scopus Q2)**.

4. Kirilenko D., Kovalev A.K., Solomentsev Y., Melekhin A., Yudin D., Panov A.I., Vector Symbolic Scene Representation for Semantic Place Recognition, 2022 IEEE World Congress on Computational Intelligence, The 2022 International Joint Conference on Neural Networks (IJCNN 2022), Padua, Italy, 18 –23 July 2022 **(Accepted)**.

5. Kovalev, A.K., Panov, A.I., Osipov, E. (2020). Hyperdimensional Representations in Semiotic Approach to AGI. In: Goertzel, B., Panov, A., Potapov, A., Yampolskiy, R. (eds) Artificial General Intelligence. AGI 2020. Lecture Notes in Computer Science, vol 12177. Springer, Cham. https://doi.org/10.1007/978-3-030-52152-3_24

6. Podtikhov, A., Shaban, M., Kovalev, A.K., Panov, A.I. (2021). Error Analysis for Visual Question Answering. In: Kryzhanovsky, B., Dunin-Barkowski, W., Redko, V., Tiumentsev, Y. (eds) Advances in Neural Computation, Machine Learning, and Cognitive Research IV. NEUROINFORMATICS 2020. Studies in Computational Intelligence, vol 925. Springer, Cham. https://doi.org/10.1007/978-3-030-60577-3_34

7. Kiselev, G., Kovalev, A., Panov, A.I. (2018). Spatial Reasoning and Planning in Sign-Based World Model. In: Kuznetsov, S., Osipov, G., Stefanuk, V. (eds) Artificial Intelligence. RCAI 2018. Communications in Computer and Information Science, vol 934. Springer, Cham. https://doi.org/10.1007/978-3-030-00617-4_1

Reports at seminars

1. March 1, 2021, VSAONLINE, Online Speakers' Corner on Vector Symbolic Architectures and Hyperdimensional Computing, Online.

   Combining Vector Symbolic Architecture and Semiotic Approach to Solve Visual Question Answering, Kovalev A.K., Shaban M., Kirilenko D., Osipov E., Panov A.I.,

   (https://sites.google.com/ltu.se/vsaonline/winter-2021)

2. May 15, 2021, Seminar of the Center for Cognitive Modeling of the Moscow Institute of Physics and Technology, Online.

   Visual Question Answering work for 2020, Part 2, Vector Semiotic Architecture for Visual Question Answering, Kovalev A.K., Shaban M., Osipov E., Panov A.I.,

   (https://youtu.be/6RSnEpqDIzg)

3. September 14, 2021, BICA Workshop at IVA'21 (21st ACM International Conference on Intelligent Virtual Agents, 14th - 17th September 2021, Fukuchiyama City, Kyoto, Japan), Online.

Vector Semiotic Model for Visual Question Answering, Kovalev A., Shaban M., Osipov E., Panov A.I.

4. April 28, 2022, Seminar of the Center for Cognitive Modeling of the Moscow Institute of Physics and Technology, Online.

   Methods and Algorithms of Neuro-Symbolic Scene Representation in Multimodal Tasks, Kovalev A.K., Panov A.I.,

   (https://www.youtube.com/watch?v=FAjE8vGkPDY)

5. May 30, 2022, Seminar "Mathematical models of information technologies" at School of Data Analysis and Artificial Intelligence (HSE University), Online.

   Methods and Algorithms of Neuro-Symbolic Scene Representation in Multimodal Tasks, Kovalev A.K., Panov A.I.

Participation in scientific projects

1. Grant of the Ministry of Science and Education of the Russian Federation No. 075-15-2020-799 "Methods for constructing and modeling complex systems based on intelligent and supercomputer technologies aimed at overcoming big challenges".

CONTENTS

In the **Introduction**, we substantiate the relevance of research conducted within the framework of this thesis, present the current state of the research topic, states the purpose and objectives, and claim the main results and personal contributions.

In **Chapter 2**, the background knowledge on Visual Question Answering as an example of multi-modal tasks and the analysis of common source of errors in VQA systems are given. We also provide the background information for the theory of the Sign-Based World Model and Vector Symbolic Architectures information necessary to understand the following chapters.

In **Chapter 3**, we consider two cases of scene representation and spatial reasoning in the Sign-Based World Model. In the first case, the goal is to determine the direction of an agent to an object. For this purpose, we construct focuses of attention of the agent and the object as a set of cells corresponding to the potential directions. We map each cell to a sign in the significance

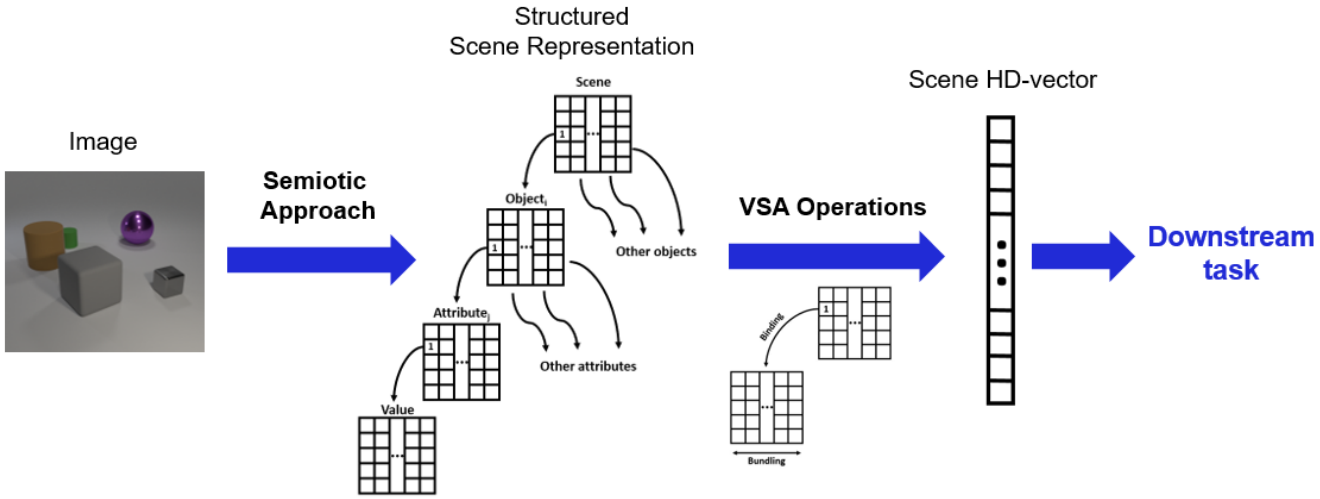

Figure 1: An overview of the neuro-symbolic scene representation pipeline.

causal network, i.e., there is a correspondence between potential directions and signs. Then, we intersect focuses of attention and apply special procedures (mental actions) to exclude cells with narrower significance on the causal network.

In the second case, we consider the more general problem when the situation is described by an enumeration of objects on the scene and relations between these objects. This formulation allows the system to answer a broader class of questions that could be formulated as a conjunction of predicates. To answer a given question, the agent applies a sequence of mental actions, e.g., generalization, concretization, etc., and replenishes the current scene representation by adding new relations as a result of activation of corresponding signs on the semiotic network.

In **Chapter 4**, we propose the neuro-symbolic scene representation in multimodal tasks as a combination of the Sign-Based World Model and Vector Symbolic Architectures (Figure 1). As an example multimodal task, we use VQA.

The neuro-symbolic representation of a scene is obtained in two stages. In the first stage, we use artificial neural networks to detect objects on the scene and extract information about their properties. Then we draw on this information to construct structured scene representation in terms of causal matrices based on the scene representation in SBWM.

In the second stage, we use VSA operations to represent structured scene representation as a scene HD vector for further use in the downstream task. Thus, we have to collapse causal matrices and relations between them into an HD vector. For this, we first represent each causal matrix as a set of corresponding columns by bundling operation and then represent a link from one matrix to another as a binding of the HD vector of matrix to which link is followed by HD vector of a

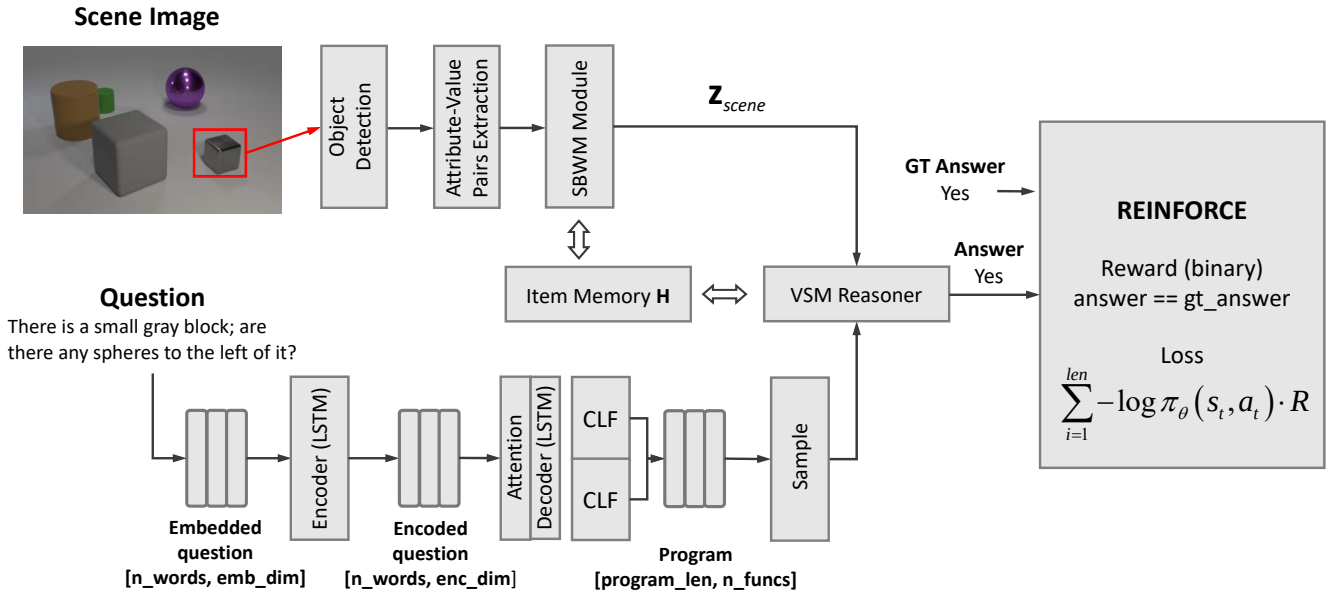

Figure 2: An overview of the proposed Vector Semiotic Model for VQA.

column from which link starts. We used the Multiply-Add-Permute with bipolar vector space [17] as a realization of VSA.

To exploit neuro-symbolic representation in the VQA, task we propose a Vector Semiotic Model (Figure 2) that transform an input image into an HD vector z_scene, represent a corresponding question as a program, i.e. the sequence of mental actions represented by VSA operations, and execute this program on the HD vector z_scene in the VSA Reasoner module. We conducted experiments on the CLEVR dataset [24], results are shown in Table 1. As can be seen, the proposed model achieves results comparable to the state of the art.

Model	Count	Exist	CompN	CompA	Query	Overall
Ours	99.2	99.8	98.9	99.6	99.5	99.4
NS-VQA[72]	**99.7**	**99.9**	**99.9**	**99.8**	**99.8**	**99.8**
NS-CL[49]	98.2	98.8	99.0	99.1	99.3	98.9
RN[63]	90.1	97.8	93.6	97.1	97.9	95.5
multiRN[8]	94.9	99.2	97.2	98.3	98.7	97.7

Table 1: Accuracy metric is used. Our model achieves comparable results to the state-of-the-art results. "CompN"and "CompA"stand for "Compare number"and "Compare attribute", respectively.

In **Chapter 5**, we apply the proposed neuro-symbolic scene representation and Vector Semiotic Model to the VQA task in Human-Centered Environments. We used the HISNav VQA dataset based on the images from the Habitat virtual environment [48]. We also used two types of questions: human-asked, collected from the crowdsourcing service Yandex.Toloka, and synthetic questions with seven templates. The main differences between The HISNav VQA dataset and the CLEVR dataset [24], that it is focused 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.

Compared to the model for the CLEVR dataset [24], as VSA we used a variance of the Semantic Pointer Architecture (SPA)[13] – Spatial Semantic Pointers [34] – to represent causal matrices and work efficiently with continuous values such as coordinates. The proposed model achieves a nearly perfect accuracy of 0.98 on the synthetic questions, but performs more modestly on the human-asked questions, 0.20 (compare to 0.43 of neural network baseline), due to the unstructured nature of questions and high variability of formulations.

In **Chapter 6**, we apply the proposed neuro-symbolic scene representation and Vector Semiotic Model to a Visual Dialog task [10]. The crucial feature of this task is that the dataset has been 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.

As 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 to solve 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.

Visual Dialog is a real-world dataset, which means that the questions asked in the dialogs are not standardized (e.g., compared to CLEVR [24] or synthetic questions of HISNav VQA). Thus, there is no straightforward way to convert a question to a sequence of procedures, which will

produce the answer if executed. Therefore, to demonstrate the proposed approach, we narrowed down types of questions to existence (Is there an object?) and counting (How many objects are there?). The proposed approach achieved 51.3% accuracy, whereas the accuracy for the counting questions is 31.0%, and the accuracy for the existence questions is 73.9%.

In **Chapter 7**, we show how neuro-symbolic scene representation improves results for place recognition. We propose a two-stage model referred as TSVLoc (Figure 3). The proposed model involves two stages. In the first stage, a global embedding is extracted using one of the traditional methods, e.g., HF-Net [64], NetVLAD [6], or Patch-NetVLAD [22]. Then, an image database is queried, and a ranking of images is produced. In the second stage, the semantic HD vector of a query image is constructed from semantic segmentation. Next, we query an image database and get the image ranking for semantic vectors. After that, the current ranking scores, together with the first-stage scores, are passed on to the score fusion module to produce a top-1 ranking image.

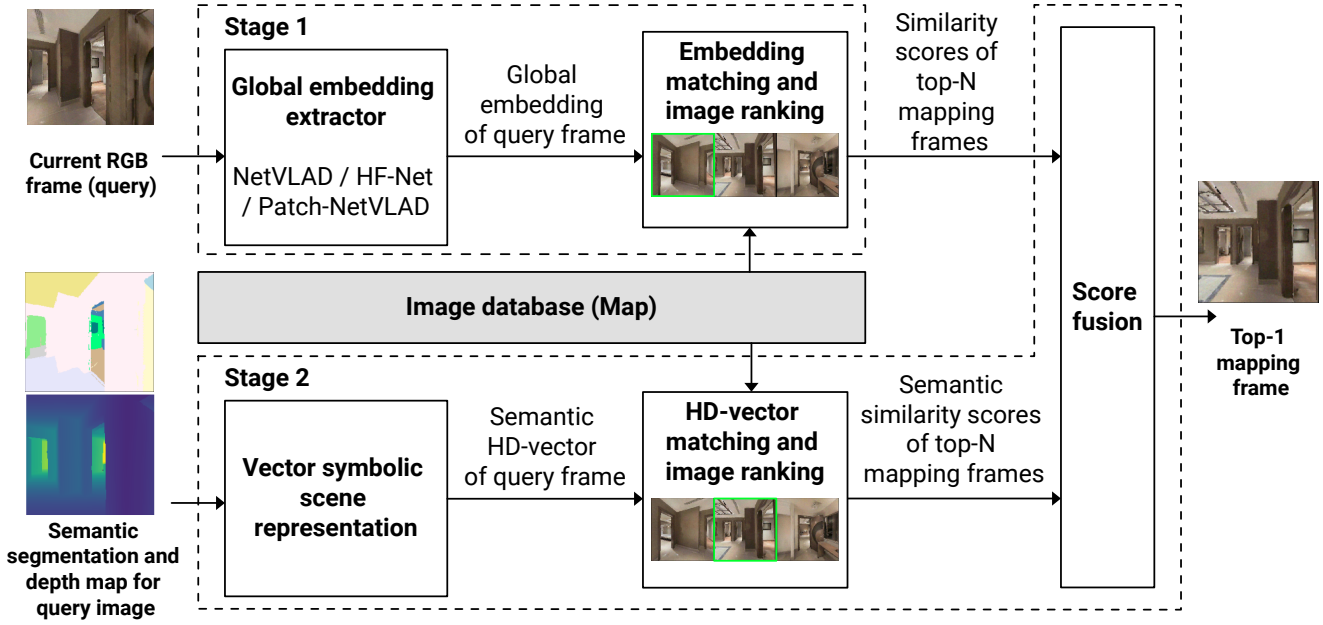

Figure 3: Overview of the proposed model TSVLoc.

To form a semantic HD vector, we use depth and semantic maps of a query frame. For every class C on the image, we generate a high-dimensional vector $\mathbf{c}$, then bind it with a depth vector $\mathbf{d_c}$ of the center of mass of the class instance, and get an object vector $\mathbf{o}$. After that, every pair of vectors that have a common border is bound together through an auxiliary vector $\mathbf{r}$ (relation "near") and summed up.

To evaluate the localization quality of the proposed method, we use the Recall (R) metric

with different thresholds. It is calculated as the fraction of query images whose translation errors do not exceed the specified thresholds $\epsilon_t \in \{0.5m, 1m, 5m, 10m\}$. Such thresholds were chosen to assess the accuracy of solving the global indoor localization problem at various spatial scales. We do not take into account the rotation error because we do not optimize the camera pose after image retrieval.

Applying the semantic representation of scenes for the place recognition task, we started with the simplest representation of scenes and gradually added various improvements to it, achieving an increase in localization metrics with TSVLoc: the base variant is a simple enumeration of the types of objects represented in the image and encoding them into one vector (SE); the first improvement to this method is to encode every pair of objects with common boundaries into one entity through an auxiliary vector (SEB); the next improvement involves the center of mass being calculated for each instance of segmentation, and the depth value at the point of the center of mass being encoded into a vector of this object (SEBD). We validated our approach on the indor HPointLoc dataset (Table 2) and subsamples of the outdoor Oxford RobotCar dataset [46] (Table 3).

As can be seen from the tables, the generation of additional embeddings using Vector Symbolic Architectures based on segmentation and depth maps (SEBD mode of our TSVLoc approach) offers a more accurate solution to the problem of rough global localization.

Method	R(0.5)	R(1)	R(5)	R(10)
HF-Net [64]	0.890	0.892	0.963	0.976
TSVLoc(H+SEB)	0.892	0.893	0.977	0.988
TSVLoc(H+SEBD)	**0.895**	**0.896**	**0.980**	**0.993**
NetVLAD [6]	0.887	0.888	0.962	0.973
TSVLoc(V+SEBD)	**0.892**	**0.893**	**0.976**	**0.987**
Patch-NetVLAD [22]	**0.942**	0.943	0.968	0.978
TSVLoc(P+SEBD)	0.931	**0.946**	**0.978**	**0.982**

Table 2: Localization metrics on all query images from the HPointLoc dataset. R(0.5) means the Recall metric with 0.5m distance threshold.

In **Conclusion**, we summarize the results of the work done during the research, outline possible

Method	R(5)	R(10)	R(25)	R(50)	R(100)
HF-Net [64]	0.485	0.639	0.708	0.737	0.761
TSVLoc(H+SEB)	**0.494**	**0.647**	**0.715**	**0.741**	**0.765**
NetVLAD [6]	0.568	0.725	0.779	0.802	0.822
TSVLoc(V+SEB)	**0.573**	**0.731**	**0.783**	**0.805**	**0.824**
Patch-NetVLAD [22]	0.702	0.842	0.877	0.888	0.898
TSVLoc(P+SEB)	**0.714**	**0.853**	**0.886**	**0.896**	**0.905**

Table 3: Averaged localization metrics on the Oxford RobotCar dataset [46].

ways for the further development of neuro-symbolic representation of scenes for multimodal tasks, and list the results to be defended.

CONCLUSION

In this thesis, we propose a new neural-symbolic scene representation for multimodal tasks. We achieve that by combining the Sign-Based World Model that provides structured hierarchical scene representation and grounding to the sensor inputs of agent with Vector Symbolic Architectures that allows us to operate with high dimensional vectors as with symbols and encode the whole scene into one vector. We also developed a model that use this representation for solving multimodal tasks with two modalities, i.e., text and image. The proposed approach is tested on VQA dataset CLEVR [24], Human-Centered Environment VQA, and Visual Dialog [10].

As a stepping stone to achieving these results, scene representations have been developed for the problem of spatial reasoning in the Sign-Based World Model.

We also propose the two-staged method TSVLoc and neuro-symbolic scene representation for the task of localization which improves results of well-known methods as HF-Net [64], NetVLAD [6], and Ptach-NetVLAD [22].

As a further development of the research topic, we can single out the development of a neuro-symbolic scene representation, which can be used both for the task of Visual Question Answering and for localization with navigation. Such a representation will be useful in solving the more involved tasks of the embodied AI associated with the performance of household tasks, such as in ALFRED [65] and TEACh [55].

Bibliography

[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] Aitygulov, E., Kiselev, G., Panov, A.I.: Task and spatial planning by the cognitive agent with human-like knowledge representation. In: Ronzhin, A., Rigoll, G., Meshcheryakov, R. (eds.) Interactive Collaborative Robotics. pp. 1–12. Springer International Publishing, Cham (2018)

[3] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. arXiv e-prints arXiv:1707.07998 (Jul 2017)

[4] Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I.D., Gould, S., van den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. CoRR **abs/1711.07280** (2017), `http://arxiv.org/abs/1711.07280`

[5] Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Neural module networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 39–48 (2016). https://doi.org/10.1109/CVPR.2016.12

[6] Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition (2016)

[7] 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). https://doi.org/https://doi.org/10.1016/j.bica.2015.09.003, `http://www.sciencedirect.com/science/article/pii/S2212683X15000468`

[8] Chang, S., Yang, J., Park, S., Kwak, N.: Broadcasting convolutional network for visual relational reasoning. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 780–796. Springer International Publishing, Cham (2018)

[9] Chen, C., Majumder, S., Al-Halah, Z., Gao, R., Ramakrishnan, S.K., Grauman, K.: Learning to set waypoints for audio-visual navigation. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=cR91FAodFMe

[10] 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)

[11] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423

[12] Eiter, T., Higuera, N., Oetsch, J., Pritz, M.: A neuro-symbolic asp pipeline for visual question answering. ArXiv **abs/2205.07548** (2022)

[13] Eliasmith, C.: How to build a brain: A neural architecture for biological cognition. Oxford University Press (2013)

[14] d'Avila Garcez, A., Lamb, L.C.: Neurosymbolic AI: the 3rd wave. CoRR **abs/2012.05876** (2020), https://arxiv.org/abs/2012.05876

[15] Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2414–2423 (2016). https://doi.org/10.1109/CVPR.2016.265

[16] Gayler, R.W.: Multiplicative binding, representation operators & analogy. In: Advances in analogy research. pp. 1–4 (1998)

[17] Gayler, R.W.: Multiplicative binding, representation operators, and analogy. In: Advances in analogy research: Integr. oftheory and data from the cogn., comp., and neural sciences, New Bulgarian University (1998)

[18] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 27. Curran Associates, Inc. (2014), https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Pa

[19] Gurari, D., Li, Q., Lin, C., Zhao, Y., Guo, A., Stangl, A., Bigham, J.P.: Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 939–948 (2019). https://doi.org/10.1109/CVPR.2019.00103

[20] Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3608–3617 (2018). https://doi.org/10.1109/CVPR.2018.00380

[21] Harnad, S.: The symbol grounding problem. Physica D: Nonlinear Phenomena **42**(1), 335 – 346 (1990). https://doi.org/https://doi.org/10.1016/0167-2789(90)90087-6, http://www.sciencedirect.com/science/article/pii/0167278990900876

[22] Hausler, S., Garg, S., Xu, M., Milford, M., Fischer, T.: Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14141–14152 (2021)

[23] He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask r-cnn. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 2980–2988 (2017)

[24] 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)

[25] Kanerva, P.: Fully distributed representation. In: Real world computing symposium. pp. 358–365 (1997)

[26] Kanerva, P.: Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. Cognitive Computation **1**(2), 139–159 (Jun 2009). https://doi.org/10.1007/s12559-009-9009-8, https://doi.org/10.1007/s12559-009-9009-8

[27] Kim, J.H., Jun, J., Zhang, B.T.: Bilinear Attention Networks. arXiv e-prints arXiv:1805.07932 (May 2018)

[28] Kirilenko, D.E., Kovalev, A.K., Osipov, E., Panov, A.I.: Question answering for visual navigation in human-centered environments. In: 20th Mexican International Conference on Artificial Intelligence (in print) (2021)

[29] Kirilenko, D.E., Kovalev, A.K., Solomentsev, Y., Melekhin, A., Yudin, D., Panov, A.I.: Vector symbolic scene representation for semantic place recognitio. In: 2022 IEEE World Congress on Computational Intelligence, The 2022 International Joint Conference on Neural Networks (IJCNN 2022), Padua, Italy, 18 –23 July 23 (Accepted (2022)

[30] 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)

[31] Kiselev, G., Panov, A.: Hierarchical psychologically inspired planning for human-robot interaction tasks. In: Ronzhin, A., Rigoll, G., Meshcheryakov, R. (eds.) Interactive Collaborative Robotics. pp. 150–160. Springer International Publishing, Cham (2019)

[32] 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)

[33] Kleyko, D., Osipov, E., Gayler, R.W., Khan, A.I., Dyer, A.G.: Imitation of honey bees' concept learning processes using vector symbolic architectures. Biologically Inspired Cognitive Architectures **14**, 57 – 72 (2015). https://doi.org/https://doi.org/10.1016/j.bica.2015.09.002

[34] 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. QC: Cognitive Science Society (2019)

[35] 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)

[36] 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)

[37] Kovalev, A.K., Shaban, M., Chuganskaya, A.A., Panov, A.I.: Applying vector symbolic architecture and semiotic approach to visual dialog. In: Sanjurjo González, H., Pastor López, I., García Bringas, P., Quintián, H., Corchado, E. (eds.) Hybrid Artificial Intelligent Systems. pp. 243–255. Springer International Publishing, Cham (2021)

[38] Kovalev, A.K., Shaban, M., Osipov, E., Panov, A.I.: Vector semiotic model for visual question answering. Cognitive Systems Research **71**, 52–63 (2022). https://doi.org/https://doi.org/10.1016/j.cogsys.2021.09.001, https://www.sciencedirect.com/science/article/pii/S1389041721000632

[39] Krizhevsky, A.: Learning multiple layers of features from tiny images (2009)

[40] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 25. Curran Associates, Inc. (2012), https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Pa

[41] Laiho, M., Poikonen, J., Kanerva, P., Lehtonen, E.: High-dimensional computing with sparse vectors. In: 2015 IEEE Biomedical Circuits and Systems Conference (BioCAS). pp. 1–4 (2015)

[42] Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. In: Arxiv (2019)

[43] Linda Smith, M.G.: The development of embodied cognition: Six lessons from babies. Artificial Life (2005)

[44] Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: NeurIPS (2019)

[45] Ma, L., Lu, Z., Li, H.: Learning to answer questions from image using convolutional neural network (2015)

[46] Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 Year, 1000km: The Oxford RobotCar Dataset. The International Journal of Robotics Research (IJRR) **36**(1), 3–15 (2017). https://doi.org/10.1177/0278364916679498, http://dx.doi.org/10.1177/0278364916679498

[47] Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: A neural-based approach to answering questions about images. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 1–9 (2015). https://doi.org/10.1109/ICCV.2015.9

[48] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., Batra, D.: Habitat: A Platform for Embodied AI Research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)

[49] Mao, J., Gan, C., Kohli, P., Tenenbaum, J.B., Wu, J.: The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. arXiv e-prints arXiv:1904.12584 (Apr 2019)

[50] Montone, G., O'Regan, J., Terekhov, A.V.: Hyper-dimensional computing for a visual question-answering system that is trainable end-to-end. ArXiv **abs/1711.10185** (2017)

[51] Oltramari, A., Francis, J., Ilievski, F., Ma, K., Mirzaee, R.: Generalizable neuro-symbolic systems for commonsense question answering. CoRR **abs/2201.06230** (2022), https://arxiv.org/abs/2201.06230

[52] Osipov, G.S.: Sign-based representation and word model of actor. In: 2016 IEEE 8th International Conference on Intelligent Systems (IS). pp. 22–26 (2016). https://doi.org/10.1109/IS.2016.7737445

[53] 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). https://doi.org/10.1134/S1064230714040121

[54] 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)

[55] Padmakumar, A., Thomason, J., Shrivastava, A., Lange, P., Narayan-Chen, A., Gella, S., Piramuthu, R., Tür, G., Hakkani-Tür, D.: Teach: Task-driven embodied agents that chat. CoRR **abs/2110.00534** (2021), https://arxiv.org/abs/2110.00534

[56] Panov, A., Yakovlev, K.: Psychologically inspired planning method for smart relocation task. ArXiv **abs/1607.08181** (2016)

[57] Panov, A.I.: Goal setting and behavior planning for cognitive agents. Scientific and Technical Information Processing **46**(6), 404–415 (Dec 2019). https://doi.org/10.3103/S0147688219060066

[58] Panov, A.I.: Behavior Planning of Intelligent Agent with Sign World Model. Biologically Inspired Cognitive Architectures **19**, 21–31 (2017). https://doi.org/10.1016/j.bica.2016.12.001, http://www.sciencedirect.com/science/article/pii/S2212683X16300913

[59] Plate, T.A.: Holographic Reduced Representations: Distributed Representation for Cognitive Structures. Stanford: Center for the Study of Language and Information (CSLI), USA (2003)

[60] Potapov, A., Belikov, A., Bogdanov, V., Scherbatiy, A.: Cognitive module networks for grounded reasoning. In: Hammer, P., Agrawal, P., Goertzel, B., Iklé, M. (eds.) Artificial General Intelligence. pp. 148–158. Springer International Publishing, Cham (2019)

[61] Rachkovskij, D.A.: Representation and Processing of Structures with Binary Sparse Distributed Codes. IEEE Transactions on Knowledge and Data Engineering **3**(2), 261–276 (2001)

[62] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)

[63] Santoro, A., Raposo, D., Barrett, D.G., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrapi, T.: A simple neural network module for relational reasoning. In: Advances in neural information processing systems (2017)

[64] Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From coarse to fine: Robust hierarchical localization at large scale (2019)

[65] Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., Fox, D.: ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020), https://arxiv.org/abs/1912.01734

[66] Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=SygXPaEYvH

[67] Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (2019)

[68] Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR **abs/1609.08144** (2016), http://arxiv.org/abs/1609.08144

[69] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. CoRR **abs/1502.03044** (2015), http://arxiv.org/abs/1502.03044

[70] Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Just ask: Learning to answer questions from millions of narrated videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1686–1697 (October 2021)

[71] Yang, S., Zhang, R., Erfani, S., Lau, J.H.: An interpretable neuro-symbolic reasoning framework for task-oriented dialogue generation (2022). https://doi.org/10.48550/ARXIV.2203.05843, https://arxiv.org/abs/2203.05843

[72] 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)

[73] Yilmaz, O.: Analogy making and logical inference on images using cellular automata based hyperdimensional computing. p. 19–27. COCO'15, CEUR-WS.org, Aachen, DEU (2015)

[74] Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 6274–6283 (2019)

[75] Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. CoRR **abs/1811.10830** (2018), `http://arxiv.org/abs/1811.10830`

[76] Zhang, Z., Shi, Y., Yuan, C., Li, B., Wang, P., Hu, W., Zha, Z.J.: Object relational graph with teacher-recommended learning for video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)

[77] Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., Fergus, R.: Simple baseline for visual question answering (2015)