Neural network based image understanding with ontological approach

Natallia Iskra
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
niskra@bsuir.by

Vitali Iskra
Omnigon Communications LLC
New York, NY, USA
iskra.vitaly@gmail.com

Marina Lukashevich
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
lukashevich@bsuir.by

***Abstract*—In this paper we propose the architecture to perform a task of semantic image analysis. The approach uses the advantages of the state-of-the art deep convolutional neural networks for object detection and builds the semantic graph that represents the scene. Ontological system is used in both graph construction and model verification. The method can be used as a part of a more extensive intelligent system.**

***Keywords*—image understanding, instance segmentation, object detection, ontology, semantic graph, convolutional neural networks, semantic analysis, intelligent system**

I. Introduction

The human interpretation of the image is based on a clear understanding of the meaning of both the scene itself and its individual elements as well as semantic connections between these elements. So, when on an image we see a roadway along with a traffic, we conclude, that the action takes place on the road and obeys the traffic regulations. We easily select the objects of the scene – cars, buses, motorcycles, pedestrians, traffic signs, and road markings. Paying attention to the objects of the scene and their relative position, we understand the situation well. All this happens quickly and naturally. However, for artificial vision systems, such an interpretation is still a challenge today.

In recent years great progress has been made in the field of image classification [1], where the task is to assign a label (or class) to each image. Further development in image analysis went in two directions:

- the improvement of the results in the field of automatic detection of multiple objects in an image (identification of object labels and the locations of the objects) [2];
- semantic description of the image, which, given a set of objects from the image, would allow to obtain a sequence of words describing more complex concepts than simply listing the objects in the image [3], thus creating a text (including the one in simplified natural language) describing relations between the objects in the image.

Solving the problem of understanding and interpreting images today requires the integration of methods from these areas [4]. Thus, in the framework of certain modern approaches a graph model that reflects semantic relations between objects is constructed based on the results of automatic detection [5].

A promising direction for further development in this area is the use of more advanced semantic means, both for describing the results of image analysis (objects and relations), and directly in the analysis process. Such tools currently are knowledge bases and ontologies [6].

Integration of knowledge about the image and the objects represented on it into the knowledge base will allow, on the one hand, to improve the accuracy of understanding through the context and information available in the knowledge base, and on the other hand, to supplement the results of the analysis with new knowledge, that is not clearly presented in the analysis results, but can be generated on the basis of these results and information from the knowledge base (to "discover" the image) [7].

As part of this work, an approach to semantic image analysis based on the integration of a model using convolutional neural networks and information representation and processing tools within the framework of an Open semantic technology for intelligent systems design is considered.

II. Image understanding problem

Currently, the majority of works related to image analysis, including semantic analysis, are devoted to solving image recognition tasks, which involves object detection, classification, and sometimes building semantic links between objects. The result of solving this problem is a description of the depicted objects, which can be both formal and natural-language based. For the formal representation of the identified relations between objects it is convenient to use models based on semantic networks.

However, building complex intelligent system, especially autonomous one, implies the ability of not only processing the images, acquired by the system from the external sources, but also the ability of the system to understand the information that can be obtained by

analyzing the image, i.e. integration of recognition results into the knowledge base of an intelligent system for subsequent analysis, correction, elaboration, decision-making on the basis of the information acquired, and other problems solving.

The task of image understanding can be represented by the following pipeline:

1) Detecting the objects in an image – the estimation of the regions, containing the objects, and their classification.
2) Building an equivalent semantic network, its analysis and verification.
3) Integrating the model into the knowledge base, eliminating synonyms, resolving contradictions.
4) Supplementing the new knowledge with the information already stored in the knowledge base.

The solution to the problem of integrating new fragments into the existing knowledge base, identifying and eliminating synonyms and contradictions, is also carried out in several stages and is discussed in more detail in [8].

The approach to the development of a mechanism for supplementing knowledge obtained by a system from the external sources using information from the knowledge base is discussed in [9].

Thus, this paper will focus on the first two stages, in particular, improving the quality of image recognition through the use of a priori knowledge stored in the knowledge base. Moreover, the approach proposed in this paper can easily be integrated with the approaches considered in the indicated papers and used to solve the problem of image understanding.

III. Existing approaches analysis

To solve such a complex task as understanding and interpreting an image, it is necessary to integrate an artificial neural network with a knowledge database [10]:

- using hybrid neural network architectures, such as convolutional neural networks and recurrent networks [11],
- by application of the semantic text analysis [12].

The solution will naturally include object detection (e.g. by means of convolutional neural network approach) and constructing the semantic structure in the form of graph.

A. Object detection

The first step in image understanding is the detection of the objects in an image – source image processing and feature extraction. Today in the task of object detection the following subtasks can be distinguished:

1) Semantic Segmentation: for each pixel in the input image define its category or class [13]. This problem can be solved by means of e.g. a recurrent network [14], however, due to the large amount of the data processed (each pixel of the image must be processed separately), this approach is very inefficient.

2) Classification and localization: determine the class of the object in the image and its exact location. This task is now considered to be solved [15], however, since the solution is mainly focused on the determining of the class and location of only one object, the existing effective solutions [16] can be applied, for example, to a part of the image (or the scene) with already selected regions.

3) Object Detection: determine the class and the rectangular region for each of the objects in an image. As noted above, the problem can be solved by proposing the number of regions of interest [17] and determining whether there is an object in the selected block and which class it belongs to. It is possible to use already pre-trained model of the convolutional neural network [18].

4) Instance Segmentation: the task is to determine object contours (all visible pixels) and its class on the image with multiple objects [19]. This way it will be possible to analyze the exact relative position of objects in an image, including distortions and occlusions.

For further construction of the semantic network in image analysis the solutions of the two last subtasks based on the class of neural network models with the so-called "region proposal" (R-CNN) are most effective:

- R-CNN [20] is a sequential image processing which generates a set of region proposals using a certain pre-trained convolutional neural network [21] with the final SVM layer [22], and linear regression for more accurate region estimation.
- Fast R-CNN [23] adds the selection of regions and the unification of all neural networks into one model to speed up the performance.
- Faster R-CNN [24] for even greater acceleration selective search of regions is used, and convolutional features are shared between all parts of the network.
- Mask R-CNN [19] in contrast to previous models uses a binary mask to determine not only a rectangular region – a candidate for an object, but also a specific pixels that belong to an object, which, in fact, is the solution to the problem of image segmentation.

In many image analysis systems proposed today precise image segmentation is already a good result [25]. However to fully understand and interpret the image further semantic analysis is necessary.

B. Equivalent semantic network construction

To solve cognitive tasks related to image understanding the designed model should reflect the connections and relations between objects in an image, the properties of these objects (qualitative and quantitative), and other information.

This model can be trained in two ways:

- supervised neural network learning using already labeled graph data [5];

- partially unsupervised learning with pruning of the least likely relations between the objects [26].

Modern methods have a number of general limitations:

- when identifying objects and building links between them one most likely option is chosen;
- only binary relations between objects are considered;
- to cut off the least likely relations only neural network methods, that have a limited set of outputs and high cost of the correction (re-training), are used;
- as a consequence of what is stated above, when cutting off the least likely relations only pairs of objects are considered, while in human image understanding much more complex structures of arbitrary configuration can be analyzed.

In addition to noted above, modern approaches are focused on solving the recognition problem and do not consider the remaining stages of the process of image understanding. This situation is partly due to the fact, that a complete solution of the problem of image understanding implies the possibility of adjusting intermediate results at each of the stages and returning to the previous stages, which in turn implies using a combination of several approaches to information processing, as well as the availability of universal means for representing information of various kinds.

The implementation of such systems on the basis of traditional modern means is a complex task, that involves the combination of heterogeneous components of the system through software interfaces between them. In addition, the system constructed in this way becomes difficult to maintain and develop and making changes to any of the modules is quite costly.

Thus the solution to the problem of image understanding requires a basic technological foundation that would:

- allow to integrate various models of information processing on a unified formal basis, both from the field of machine learning, and, e.g. models of logical inference;
- ensure the unification of the representation of heterogeneous information in the memory of a computer system, including both intermediate recognition results and the previously accumulated knowledge base;
- ensure the ability of adding and adjusting the models of information representation and processing (system re-training).

The standards proposed by the W3C consortium RDF [27] and OWL [28] in particular are currently widely used as the basis for the development of knowledge bases and ontologies. However these standards have a number of significant limitations [29], [30] on the one hand, and regulate only low-level of information representation on the other, almost without regard to approaches to information processing presented in this form.

IV. Proposed approach

In this paper it is proposed to use OSTIS Technology [31] as a formal basis for the implementation of an image understanding system.

The orientation to this technology is due to the presence of the following components:

- a unified version of coding the information of any kind based on semantic networks with set-theoretic interpretation called SC-code;
- a model of an abstract semantic memory storing SC-code constructs (sc-memory) and a model of a basic machine for processing of SC-code structures [31];
- a model of the representation of various types of knowledge and models, methods and tools for developing knowledge bases using SC-code [32];
- models, methods and development tools for hybrid problem solvers in sc-memory based on a multiagent approach [33] that allow to integrate various problem-solving models, including neural networks [34] within one solver.

Thus OSTIS Technology meets the requirements for the technological foundation necessary to implement the proposed approach to image understanding and can serve as the basis for building a hybrid system for semantic image analysis.

Systems developed using this technology are called ostis-systems and it is assumed that the designed system of image understanding can be further integrated as a subsystem into other ostis-system, it is also an ostis-system by itself and follows the same rules.

As mentioned earlier the focus of this paper are the principles of image recognition with regard to the information previously stored in the knowledge base. Let us consider in more detail the stages of solving this problem in the framework of the approach proposed in this paper.

A. Object detection

At this stage selection of objects in the image, classification and building of a topological description of the objects (bounding boxes in the image, see Fig. 1) is performed.

For the object detection the Faster R-CNN (deep convolutional neural network) is used. The output of the network is a set of regions (coordinates of angles) and a vector with class labels probabilities for each region as presented in the table below.

B. Building object relations

In general, building of possible connections between objects can be carried out in several ways:

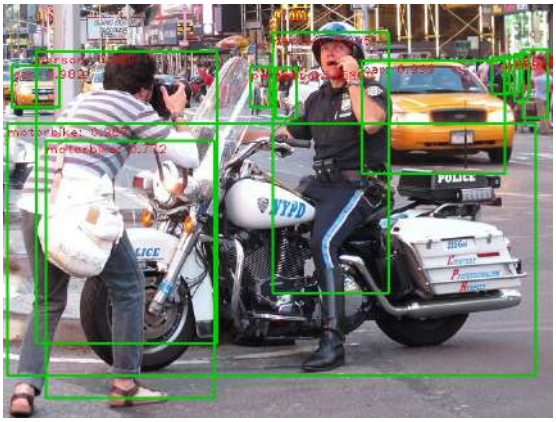

Figure 1. Object detection visualization (Test image from COCO dataset [35]).

Table I
OBJECT DETECTION WITH TOP-ONE CLASSIFICATION SCORES

Class ID	Class Name	Score	y1	x1	y2	x2
7	car	0.9990532	65	417	196	583
7	car	0.98173344	71	9	119	65
14	motorbike	0.98688495	138	4	431	619
14	motorbike	0.7723496	158	49	457	245
15	person	0.99547195	32	313	335	448
15	person	0.9894701	54	38	393	248
15	person	0.9516165	61	568	104	588
15	person	0.8993606	73	316	132	341
15	person	0.8546056	52	604	131	633
15	person	0.76542825	60	582	102	600
15	person	0.66580576	57	594	112	611
15	person	0.52531904	72	287	119	309

- based on machine learning techniques without taking into account the topological connections between objects (e.g. using word2vec approach [12]);
- based on machine learning techniques taking into account the properties of objects and their topology; this option involves the training of the neural network on manually labeled image datasets;
- based on a priori knowledge formalized in the form of ontologies stored in the knowledge base.

It is important to note that for building the knowledge base of big volume containing this kind of information, e.g. probabilistic rules, machine learning methods can also be used.

In the framework of this work the choice was made in favor of the second option, because it allows quickly enough on the basis of available free-access image datasets to train neural network models, which makes possible to quickly create the initial configuration of connections between objects, that can be later adjusted using the information from the knowledge base. At the same time this approach, unlike the first option, allows to take into account the properties and locations of objects, and also requires a relatively small expenditure on the creating of a priori information and has a higher performance compared with the third option.

To build initial version of the semantic model two neural networks are trained:

- the first neural network determines for each pair of objects, whether they can be a subject-object pair, using the probabilities of classes (the idea is, that *person-motorcycle* are probably subject-object, and the *sky-motorcycle* is probably not);
- the second neural network builds a graph for the remaining after the first step subject-object pairs, and marks edges with possible semantic relations (*(person, motorcycle) -> person* ***sits on*** *motorcycle*).

To train a neural network it is necessary for each training image to have a semantic graph, such as provided in Visual Genome dataset [5].

The data in Visual Genome was pre-processed and manually labeled. The dataset consists of seven main components:

- regions
- objects
- attributes
- relations
- region graphs
- scene graphs
- question-answer pairs.

An example of the labeled data is shown in Fig. 2

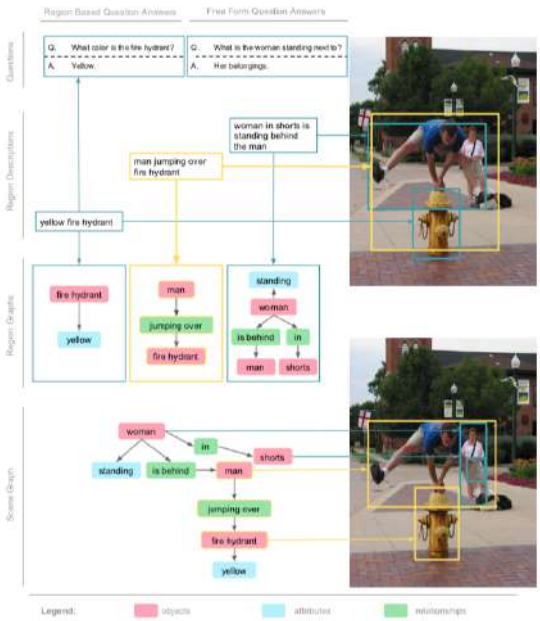

Figure 2. Training data from Visual Genome dataset [5].

The initial configuration of relations at this stage is constructed considering the most probable classes for objects, obtained during the previous stage. For each constructed link there is also a probability of belonging to

certain classes of links (relations), from which the most likely relation is initially selected.

C. Model immersion and verification

At this stage the resulting semantic model is immersed in the knowledge base (at least, the merging of nodes with the same names is performed) and verified using the patterns presented in the corresponding ontologies.

In general, when the number of possible classes for recognizable objects is large, the development of such ontologies is a laborious task and can be simplified by automation, including using neural network models. However, at the current stage, since we are talking about a relatively small number of classes, the development of such an ontology was done manually based on expert knowledge of the subject domain.

The statements presented in the knowledge base are interpreted by the corresponding knowledge base verification agents, that are part of the problem solver of the image understanding subsystem. To implement the solver, the approach developed in the framework of OSTIS Technology proposed in relevant works, e.g. [33], is used. One of the advantages of this approach is the ability to expand the range of agents, that are included in the solver, without significant increase of the labour cost. Thus the tools of knowledge bases verification can be constantly improved.

If in the process of verification using the current version of the ontology and the current set of agents, no contradictions has been detected, the resulting semantic model is accepted as final, and its further immersion into the knowledge base, recognition and merging of synonyms, etc. is performed. Otherwise an adjustment step is performed.

D. Results correction and re-verification

In case verification process detects contradictions, the semantic model is adjusted, which in turn involves several steps:

- the fragment of the model containing contradictions is localized; in the current version of the approach, adjustments are made only within the localized fragment;
- another combination of relations between objects is selected, taking into account the probabilities obtained in step B, the verification is repeated;
- if for the selected classes of objects it is not possible to select a satisfactory combination of relations, then roll back to the beginning of step B is performed and other classes are selected for one or several objects taking into account their probabilities, after which the links between objects are rearranged based on the newly selected classes.

V. System architecture

As previously discussed, the solution is proposed to be an ostis-system with corresponding architecture [36].

To improve the system performance the current implementation uses Pytorch and Tensorflow in the parts related to neural networks. Then the detection results are put into the knowledge base of the ostis-system, where they are verified in accordance with the routine discussed previously.

The proposed system architecture is summarized in the Fig. 3.

A. Object detection unit

The object detection unit is built after the Faster R-CNN [24] architecture.

1) The image is passed to the feature extraction network. The feature extraction is performed by one of the image classification architectures: VGG-16 [37], ResNet-101 [38] or FPN [39], – with few of the outer layers removed.
2) The feature map (e.g. with 256 distinct features for the VGG) is passed to the RPN (Region Proposal Network). The tasks of the RPN is to determine a collection of regions of interest (approx. 2000 - 5000) which have a probability to contain an object.
    a) The first layer of the RPN is the $3 \times 3 \times 1$ convolution, computing the feature vector for each 3×3 window of the feature map. The 3×3 window in the feature map corresponds to sufficiently large local receptive field in the source image (228×228 for the VGG).
    b) Two parallel fully-connected layers produce a region proposition for each of the K anchors situated at the center of each window ($K = 9$ – pre-defined windows with fixed scale and aspect ratios):
        i) The layer that calculates the "objectness" (the probability to contain an object) of each of the K input windows. This layer is basically doing a 2-class anchor classification: for each anchor it decides whether the anchor is having significant (70%) intersection with an object enclosing rectangle.
        ii) Bounding box regression layer: for each anchor with a positive "objectness" the layer calculates a correction to the pre-defined anchor coordinates to match them with the actual object enclosing rectangle (the output is 4 numbers: the correction to the x, y and width/height of the rectangle).
3) Some of the produced regions are going to share a lot of common pixels, such regions are eliminated

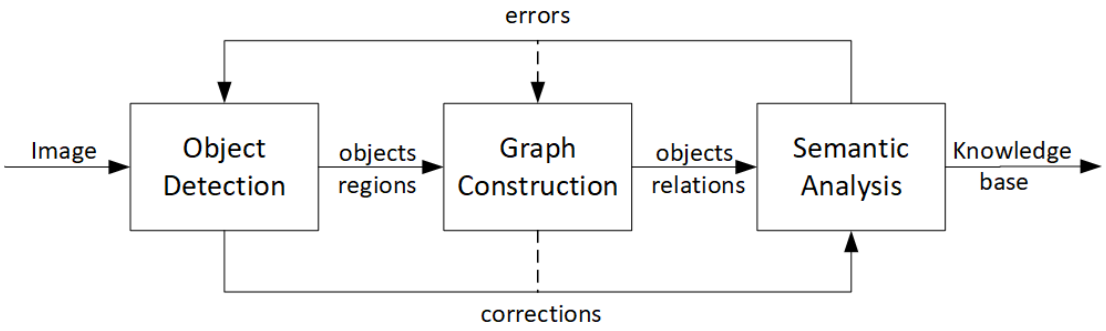

Figure 3. System architecture.

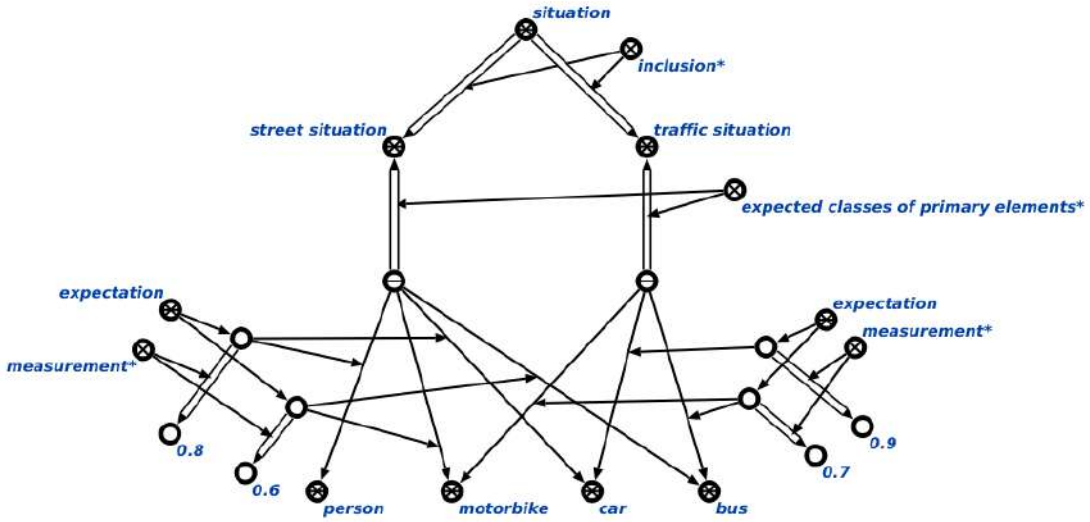

Figure 4. Classes of situations and classes of objects.

using NMS (non-maximum suppression) algorithm to reduce redundancy.

4) Each region of interest is projected to the region of the feature map (the result of step 1).
5) Features extracted by the feature map for the region are adjusted to match dimensions expected by the classification network. One of the following technique is used here:
   a) RoiPool [23] – maximum-pooling layer with fractional stride, ensuring the expected output dimensionality.
   b) RoiCrop [40] – learnable model, that is able to preform scaling with interpolation.
   c) RoiAlign [19] – feature map scaling using bi-linear interpolation.
6) Features are passed to the classifier network to determine an object of the region. Class list is extended with the catch-all "background" class to give the network an opportunity to reject a proposed region. For each class (except for the "background") the network outputs 4 numbers in addition to class probabilities – enclosing window displacement, assuming the object class. It allows the network to correctly detect the window coordinates in case of an RPN error, given that RNP cannot distinguish different classes of objects thus is unable to contain class-specific information to determine correct object region placement.

B. Semantic analysis unit

In general, an ostis-system consists of a system model presented using SC-code (sc-model) and an sc-models interpreting platform. At the same time the sc-model of an ostis-system may be subdivided into an abstract semantic memory model (sc-memory), a knowledge base sc-model, and a problem solver sc-model.

Let us consider the contents of each of these components in more detail from the angle of the image understanding problem.

The approach behind the OSTIS Technology framework is to represent the knowledge base sc-model of

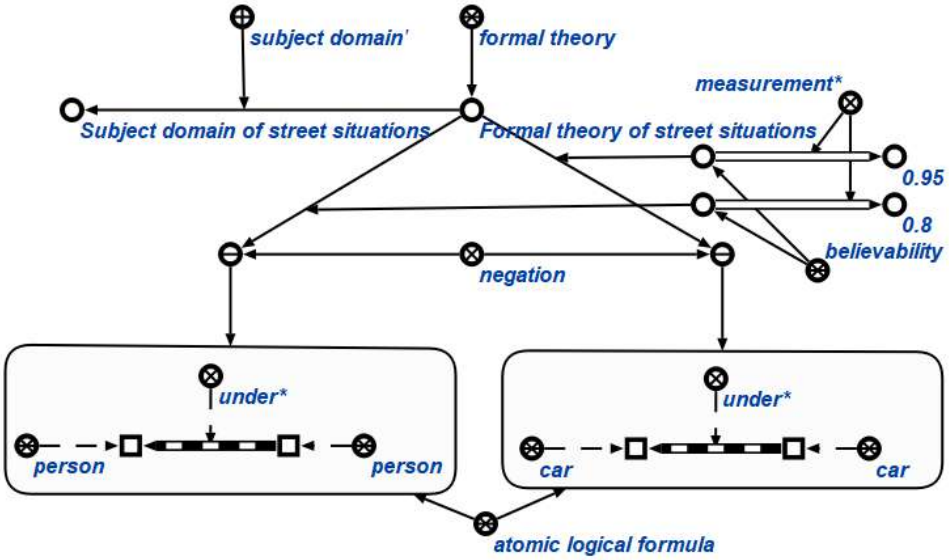

Figure 5. Low-probability situation classes for the example domain.

the ostis-system by hierarchic system of sc-models of the subject domains [32] and corresponding ontologies. As a central reusable knowledge base component, the Knowledge Base Semantic Model Kernel is developed within the technology. The Kernel is included in the each newly-created knowledge base of the system and contains a number of top-level domains and ontologies.

To solve the image understanding problem, the proposed systems uses concepts explored in the following Kernel comprising domains:

- *Subject domain of actions and tasks*
- *Subject domain of situations and events*
- *Subject domain of spatial entities*
- *Subject domain of material entities*
- *Subject domain of temporal entities*
- *Subject domain of parameters and values*

The implemented image understanding system was decided to be oriented towards city traffic images processing. From that followed a development of the model fragments for the following domains describing main object classes occurring on such images, objects relations, and typical situation classes:

- *Subject domain of buildings*
- *Subject domain of living creatures*
- *Subject domain of vehicles*
- *Subject domain of streets and street situations*

The information required on the detection result verification step is specified in the ontologies corresponding to subject domains. In particular, object classes that are expected in the situation of corresponding classes (Fig. 4) and scene classes that are improbable in the context of the given subject domain (Fig. 5) are specified.

In turn, the sc-model of the problem solver is interpreted as a hierarchical system of agents driven by situations and events in shared sc-memory [33]. Such agents are called *"sc-agents"*. Non-atomic sc-agents, that could be decomposed to a simpler sc-agents are considered separately. The structure fragment of a non-atomic knowledge base verification sc-agent in the SCn-code [41] is presented below.

Non-atomic sc-agent for knowledge verification
<= *abstract sc-agent decomposition**:
{
- *Abstract sc-agent for compliance of relations to its domains verification*
- *Abstract sc-agent for compliance of action specification to its class verification*
- *Abstract sc-agent for compliance of class instance to class definition verification*
- *Abstract sc-agent for validation on the base of uniqueness statements*
- *Abstract sc-agent for validation on the base of statements about the impossibility of the given situation existence*

}

VI. Example

Let us consider the following example. The source image is shown in Fig. 6.

The object detection results are regions and objects as presented in Fig. 7 and in the table II.

Based on the detected objects taking into account the coordinates of their locations a closeness graph is constructed (Fig. 8). For clarity the probability values of certain objects of the corresponding classes are omitted.

Figure 6. The source image (from COCO dataset [35]).

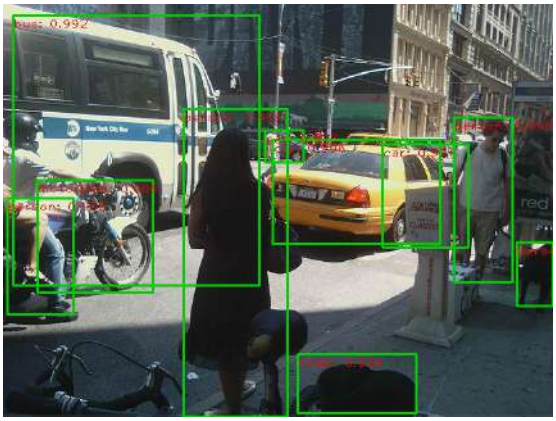

Figure 7. Object detection results.

Next the relations between the neighbouring objects are specified in the graph construction unit (Fig. 9). The semantic analysis unit based on the nature of the selected objects determines the context of the image (*street, traffic*).

Further the unit determines possible errors either in the object detection or in the relations (Fig. 10).

In our example the two fragments of the resulting graph possibly containing errors are localized. The identified fragments correlate with the corresponding classes of incorrect constructions in the knowledge base (in figure the belonging to the indicated classes is omitted).

In this example the first of the situations is considered incorrect, since the presence of the object of the class "chair" is unlikely in a case of the "street situation" class in accordance with the description in the framework of the ontology (Fig. 4).

The second case contradicts the probabilistic statement considered earlier (Fig. 5), that within the scope of the "road situation" there are no situations like "person-under-person" and "car-under-car" (in the current version it is assumed that all participants in the process are on the street level).

Table II
OBJECT DETECTION RESULTS

Class ID	Class Name	Score	y1	x1	y2	x2
6	bus	0.991863	12	11	326	297
7	car	0.9956397	157	313	278	508
7	car	0.92731386	162	441	282	541
7	car	0.9228814	146	304	183	349
9	chair	0.9342198	406	342	475	480
14	motorbike	0.90591204	204	39	334	174
15	person	0.9992173	121	209	479	330
15	person	0.9987311	131	524	322	592
15	person	0.96812516	225	4	360	81
15	person	0.8665556	276	597	350	639

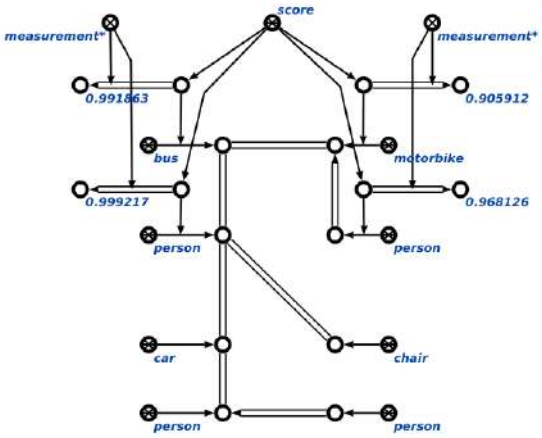

Figure 8. Object detection results in semantic memory.

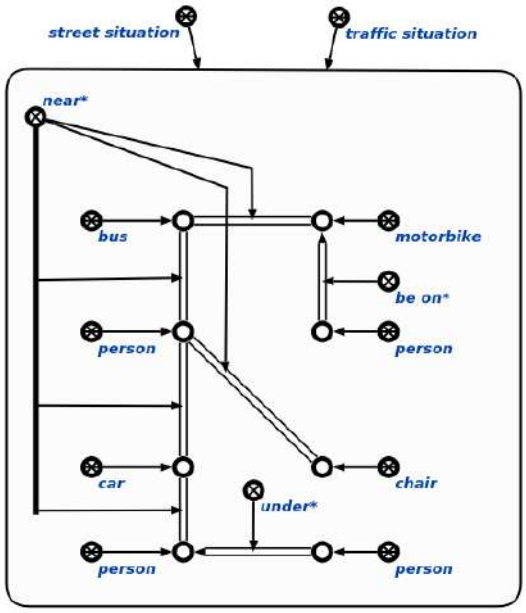

Figure 9. Semantic model with determined relations.

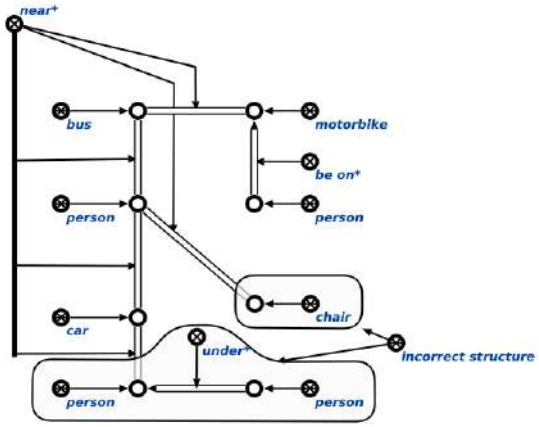

Figure 10. Semantic model with problem fragments localized.

The object detection unit (or the graph construction unit, depending on the nature of the error) re-checks the results for "suspicious" regions (Fig. 11). The model is adjusted (Fig. 12).

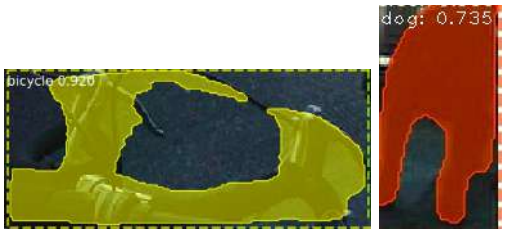

Figure 11. Object detection for "suspicious" regions.

According to the adjusted model, taking into account the subject-object language relationships and replacing the relations designation with appropriate language constructs (in this case, verbs), the following semantic descriptions can be constructed:

"**the** person **rides the** motorbike"

"**the** person **walks the** dog"

VII. Conclusion and further work

The paper considers an approach to improving the quality of image recognition based on the integration of neural network models and the ontological approach. The results obtained will be used further in the context of solving the problem of image understanding.

Furthermore, as one of the ways to develop the proposed approach, it is supposed to use ontologies for training artificial neural networks on the one hand, and to use neural networks and labeled markers for automating ontologies making on the other. The combination of these approaches will reduce the requirements for the volume and quality of a priori information necessary for building recognition and understanding systems, and expand the scope of application of such systems correspondingly.

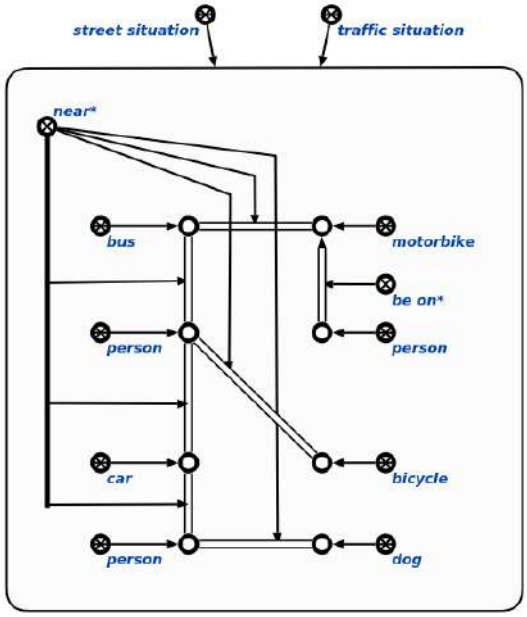

Figure 12. Adjusted semantic model.

Acknowledgments

The research presented in this paper was conducted in close collaboration with the Department of Intelligent Information Technologies of Belarusian State University of Informatics and Radioelectronics. Authors would like to thank the research group of the Department of Intelligent Information Technologies for productive cooperation.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[2] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, and G. Wang, "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018.

[3] M. Z. Hossain, F. A. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *CoRR*, vol. abs/1810.04020, 2018.

[4] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[5] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, pp. 32–73, 2016.

[6] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing," *International journal of human-computer studies*, vol. 43, no. 5-6, pp. 907–928, 1995.

[7] Y. Yang, Y. Li, C. Fermüller, and Y. Aloimonos, "Neural self talk: Image understanding via continuous questioning and answering," *CoRR*, vol. abs/1512.03460, 2015.

[8] V. P. Ivashenko, "Modeli i algoritmy integratsii znanii na osnove odnorodnykh semanticheskikh setei [models and algorithms for the knowledge integration based on homogeneous semantic networks]," *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, vol. 2, pp. 111–132, 2015.

[9] N. Hubarevich, I. Boyko, A. Semenyaka, and A. Hardzei, "Knowledge acquisition based on natural language texts," *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, vol. 2, pp. 199–206, 2018.
[10] V. A. Golovko, V. V. Golenkov, V. P. Ivashenko, V. V. Taberko, D. S. Ivniuk, A. A. Kroshchanka, and M. V. Kovalev, "Integration of artificial neural networks and knowledge bases," *Onthology of Designing*, vol. 8, no. 3(29), pp. 366–386, 2018.
[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.
[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
[14] P. H. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling," in *31st International Conference on Machine Learning (ICML)*, no. EPFL-CONF-199822, 2014.
[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *CoRR*, vol. abs/1709.01507, 2017. [Online]. Available: http://arxiv.org/abs/1709.01507
[17] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European conference on computer vision*. Springer, 2014, pp. 391–405.
[18] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in neural information processing systems*, 2014, pp. 3320–3328.
[19] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.
[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.
[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
[22] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 806–813.
[23] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.
[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.
[25] X. Liu, Z. Deng, and Y. Yang, "Recent progress in semantic image segmentation," *Artificial Intelligence Review*, pp. 1–18, 2018.
[26] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
[27] (2017, Jun.) RDF 1.1 concepts and abstract syntax. [Online]. Available: http://www.w3.org/TR/rdf11-concepts/
[28] (2017, Jun.) OWL 2 Web Ontology Language document overview. [Online]. Available: http://www.w3.org/TR/owl2-overview
[29] (2016, Sep.) Defining N-ary relations on the semantic web. [Online]. Available: https://www.w3.org/TR/swbp-n-aryRelations/#pattern1
[30] (2016, Sep.) Representing specified values in OWL: «value partitions» and «value sets». [Online]. Available: https://www.w3.org/TR/swbp-specified-values
[31] V. Golenkov and N. Gulyakina, "Proekt otkrytoi semanticheskoi tekhnologii komponentnogo proektirovaniya intellektual'nykh sistem. Chast' 2: Unifitsirovannye modeli proektirovaniya [project of open semantic technology of component design of intelligent systems. part 2: Unified design models]," *Ontologiya proektirovaniya [Ontology of design]*, no. 4, pp. 34–53, 2014, (in Russian).
[32] I. Davydenko, "Semantic models, method and tools of knowledge bases coordinated development based on reusable components," *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, vol. 2, pp. 99–118, 2018.
[33] D. Shunkevich, "Agent-oriented models, method and tools of compatible problem solvers development for intelligent systems," *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, vol. 2, pp. 119–132, 2018.
[34] V. Golovko, A. Kroshchanka, V. Golenkov, V. Ivashenko, M. Kovalev, V. Taberko, and D. Ivaniuk, "Integration of artificial neural networks and knowledge bases," *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, vol. 2, pp. 119–132, 2018.
[35] T.-Y. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.
[36] V. Golenkov, N. Guliakina, N. Grakova, I. Davydenko, V. Nikulenka, A. Eremeev, and V. Tarassov, "From training intelligent systems to training their development tools," *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, vol. 2, pp. 81–98, 2018.
[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.
[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
[39] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, 2017.
[40] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.
[41] (2018, Jun.) IMS metasystem. [Online]. Available: http://ims.ostis.net/

НЕЙРОСЕТОВОЕ РАСПОЗНАВАНИЕ ИЗОБРАЖЕНИЙ С ИСПОЛЬЗОВАНИЕМ ОНТОЛОГИЧЕСКОГО ПОДХОДА

Искра Н.А., Искра В.В., Лукашевич М.М.

В настоящей работе предлагается архитектура для выполнения задачи семантического анализа изображений. Подход использует преимущества современных глубоких сверточных нейронных сетей для обнаружения объектов и создает семантический граф, который представляет сцену. Онтологическая система используется как при построении графа, так и при верификации модели. Этот метод можно использовать как часть более сложной интеллектуальной системы.

Received 27.12.18