EasyChair Preprint

№ 7921

Deep Learning Framework for Artificial General Intelligence

Shimon Komarovsky

May 5, 2022

Deep learning framework for Artificial General Intelligence

Shimon Komarovsky[1][0000−0002−9036−0282]

Technion - Israel Institute of Technology, shiman@campus.technion.ac.il

Abstract. This paper proposes two Deep Learning (DL) related models, to serve potentially as parts of an AGI agent. The first one is designed bottom-up, i.e. it is mostly based on DL. The second one is a partial AGI model, specifically concerning the thinking process. It is designed top-down, i.e. it is mainly based on cognition and communication. The latter has not yet been fully designed for implementation. It only describes the representation of the data, and its relevance to DL is by being triggered by some Deep Neural Network (DNN).

Keywords: Deep learning · Neuro-Science · Associative thinking.

1 INTRODUCTION

Since rule-based design tends to be rigid, it is not suitable to construct AGI, but perhaps just act as an inspiration. Instead, it should be based on neuro-sciences, to handle a large variety of scenarios and to have many vital features. Features such as: flexible, fluid, adaptive, and evolving.

We first propose a DL Model (DLM) originated mainly from the neural model in [10]. Then, we propose a model for the important components of an AGI agent: thinking and memory. It models the representation of elements in a memory, and describes how the thinking process accesses them and manipulates them for different tasks. It also encourages flexibility and adaptivity.

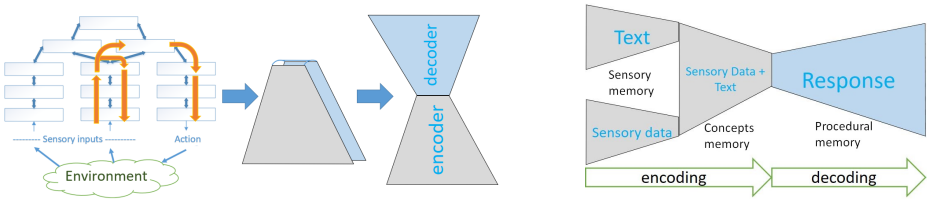
It is evident in neuroscience and DL that knowledge has a hierarchical structure, though there is a controversy about which type is it. In DL and [10] it is a hierarchy of features, while in [11] it is about the compositionality of objects. In our case, our DLM is mainly established on temporal hierarchy. Whereas our model for AGI is based upon associative hierarchy, designated for efficient memory access.

Finally, both of our presented models are based on the System 1 and 2 principle, see [4]. They are both also based on the stimulus-response principle, since we believe that one of AGI's characteristics is that knowledge is operational. In other words, elements that are learned are either objects or their attributes or actions which act upon them. This notion is presented in many papers on associative memory or associative NNs, where an association is a response to a stimulus, which can be either other stimuli [21] or a behavioral response (action) [14]. Associative NNs can also fuse different modalities [12].

2 The proposed DLM

Our DLM is inspired by the neural models and DLMs such as caption generation [26] and Visual-Question-Answering [5]. As shown in Fig. 1(a), the idea is to unwrap the percept-predict structure from the neural model [10] on the left, into a discriminative-generative or an encoder-decoder structure on the right.

The proposed DLM is illustrated in Fig. 1(b). In this structure we encode the data coming from text and sensors. The text includes both information and instructions. Finally, we encode this data into some extracted features representing the whole situation, including what the model is requested to do, and then up-sample it to the actuators (the decoding process).



(a) Percept-predict structure turns into a discriminative-generative structure.

(b) Sensory data and text in encoder-decoder structure.

Fig. 1. Sensory data, text and response in the proposed DLM.

There is evidence of this multi-modal fusion in the literature on image captioning or video recognition tasks. For example, a visual input is encoded into spatio-temporal space, and sentences describing the visual input are encoded into a continuous vector space. Then, the goal is to minimize the distance of the outputs of a deep visual model and a compositional language model in a joint space, and eventually to update these two models jointly [7, 22, 27]. We use this method in our proposed DLM, as discussed in 2.1.

The inner components of the proposed DLM are replaceable, and can be implemented via appropriate DNNs. Sensory input can be handled by e.g. CNN, DBN, and SAE. Text input can be handled by e.g. Transformers, RNNs or their variants: LSTM or GRU. The Sensors-and-Text and the final decoder can also be implemented via sequence-based RNN, as in [1]. Because based on [10], the grasping of a situation is gradual in time. It takes time to figure out the stable situation, and it takes time to follow up on some desired plan. A plan is realized by a sequence of actions, such as in [23], where a robotic-complex-action is transformed, transferred, and then used for the control of base actions. The response, depicted in Fig. 1(b), can be either a physical operation or a sequence of words (e.g. an answer to a question).

Our DLM learns two types of information: objects and actions. We start by teaching basic elements/objects. Then continue with composite objects and actions.

2.1 Proposed DLM function

A more detailed implementation of the proposed DLM is discussed.

The DLM has a hierarchical temporal structure, and it is mainly based on two ideas: the joint learning of multi-modal input, and the learning of intermediate tasks [6,8,13]. The latter is used to implement scene understanding within different time scales (short, mid, and long terms).

The hierarchical temporal structure can be implemented via different clock rates, as suggested in [13]. Another way is via sliding/shifted LSTM blocks as in [6], used to extract different time-scaled features. And another way is via dilated casual convolution, as in [8].

The first idea is about extracting features separately from sensors and text, then learning them together via joint embedding space [22]. Thus, we assume that these inputs are complementary. Since if they are trained together, then if one of them is missing, it is sufficient for recognition as if the second one was there too. These fused features represent spatio-temporal information for the short-term temporal resolution. In the next phase of learning, we extract these joint features further into longer time scales, by freezing first the short-term RNN layers and activating mid-term layers only. The same goes for the long-range layers afterward.

Other types of such gradual learning exist in literature. For example, in [3] gradual learning is proposed from a simple level to a complex level, either manually (expert-guided) or automatically (scoring each sample by its training loss). However, this loss is highly dependent on the models and their hyper-parameters. Hence, different learning takes place: from fewer categories or output tasks (local) to more categories (global).

It is possible also to test adding joint embedding space for spatial information only, before the spatio-temporal short-term joint embedding, as it is done for example with static visual images and simple textual objects [15]. This embedding enables the learning of static compositionality of objects, while later, the inclusion of temporal dimension enables temporal compositionality learning.

The second idea is generally about hierarchical learning of tasks [2, 9, 18], whereby several layers of tasks are learned instead of the usual single output layer of tasks. In temporal hierarchical learning, we learn the current layer of tasks, then later we learn more complex tasks on a new layer, based on the previous tasks.

In our DLM, it is realized by intermediate tasks via RNNs. Using the first idea we simply extract features in different time resolutions as described previously. These features are the hidden and the output layers in RNN. However, to include intermediate tasks for different time resolutions, we use the encoder-decoder structure of RNN, as in translation tasks. In other words, the intermediate tasks are connected to the context signal(s) of the RNN, not to its hidden/output signal(s). A decoder is attached to the context or to the encoder layer in the RNN. Thus, the intermediate tasks are the outputs of each of these decoders. See more in [18].

We can illustrate how training occurs in the DLM via the second idea: visual data and equivalent or/and complementary textual data about objects and actions are merged in a joint space. Then it traverses to one of the operating (short/mid/long-ranged) RNNs, to accomplish the task of predicting the scene, i.e. of the correct relationship between objects and actions in it, for example via graph representation (scene graph).

In conclusion, we have two ways to implement hierarchical temporal learning. Either we can use the first idea, and learn multi-modal data in joint embedding space, at different time scales. Or, we can extract features hierarchically temporarily (via RNN output/hidden layers), and insert intermediate tasks into the temporal structure. Tasks assisting in forming correct and more appropriate (guided) features, as in [2, 18, 20, 24]. Thus, after the recognition of spatio-temporal objects in the features extracted from the two inputs, we should recognize their relationships. Hence the intermediate tasks derive these relationships between objects. Some papers [16, 27] focus on pairwise interactions between perceived objects in an image, e.g. via a 2D graph matrix, whereas [17] models high-order interactions between arbitrary subgroups of objects.

The full sketch of the proposed DLM is shown in Fig. 2. We see that the decoder is also hierarchically-temporarily constructed, as a mirror image of the perceptual encoder, with skip connections, whose function may be: copy, normalization, or addition.

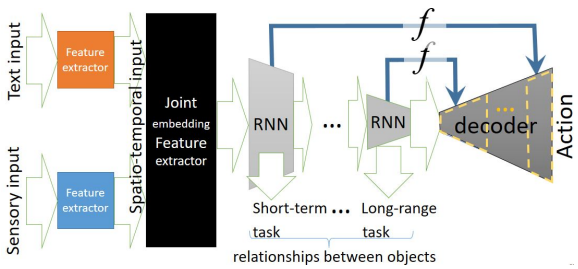


Fig. 2. Hierarchy-temporal DLM.

2.2 Additional suggestions for the proposed DLM

We should include a memory [25], either implicitly in the learning NN components themselves, or explicitly as additional components in the model, with a different type of memories, e.g. sensory, conceptual, and procedural memories as depicted in Fig. 1(b).

A few additional aspects are presented for the proposed model. First, initially we thought to have a single channel for both informative text, describing the current situation, and instructional text, asking the system to perform something. But we figured we should separate these channels, due to several reasons: (i) When our text input is a command, then the NN is trained by using executions

as outputs. However, there is no output to yield from a descriptive text. (ii) Often both channels are needed simultaneously: a descriptive one, such as coming from the user or some online source, and a commanding one. Furthermore, we presume that a commanding input represents our system's objective, and this objective has to be supplied consistently.

The second aspect is about full-model training phase. After training on intermediate tasks to produce more representative features, we train the DLM on its actual output: the response (actions or answers). In this phase, we perform only feature extraction and disregard the intermediate tasks, since their function is needed no more.

Lastly, until now we have discussed gradual learning in the encoder of our DLM, after which we finish with supervised learning for the final response, via the decoder. Nevertheless, we can perform gradual learning also in the decoder. First we teach the encoder-decoder fast tasks with immediate execution. Next, we fix these first layers and teach the mid-term layers, and continue with the same fashion. This resembles the biology-based approach, in which after repeating some task, it becomes automatic for us such that our mind is free from concentrating on it, and now we can deal with other tasks while performing these low-level tasks. Similarly is here: after accomplishing the low-level tasks at a high level of performance, we are free to learn new tasks. This gradual learning can also introduce a working memory or thinking in higher available layers, where the inputs are very slow/stable, and allow the DLM to solve difficult tasks.

3 Associative AGI model

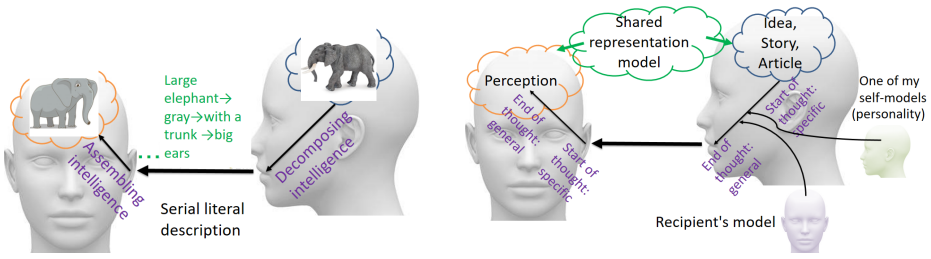
In this section, we describe a model of AGI's most important components. As explained in 3.3, it can utilize our general DLM, 2, as its base memory. This model tries to encapsulate a few cognitive important elements: short-term memory (STM), long-term memory (LTM), working memory (WM), and thinking. As mentioned in the abstract, it is designed in a top-down fashion. Specifically, it originates from our communication model.

3.1 Communication

Our fundamental assumption about human-human communication is that each person is a "black box". Thus, we do not have access to the actual inner interpretation and representation of persons' knowledge. In other words, we communicate externally, via objective tools (the language), but we have hidden subjective perspectives or world models, constructed during a lifetime via different circumstances and experiences. This assumption is illustrated in Fig. 3(a), where the inner representation of the same message varies among people.

Next, our communication model consists of several principles. (i) The sending process is about converting an abstract message, such as a story or technical procedure, into a sequence of words. Hence, this process is generative. It is about decomposing a high-level idea into low-level concepts. Exactly opposite is

the receiving process. In it, the recipient tries to assemble the idea from the low-level concepts, hence it is a discriminative process. These processes are visualized in Fig. 3(a). (ii) These couple of processes can be viewed also temporarily. The sender's thought is materialized fully when he begins his sentence(s). But to fully capture his message, the recipient has to wait till the end of the message. Hence, the end of the thought is the beginning of the message, while its start is the ending of the message. (iii) Additionally, it is about context. Due to the "black-box" assumption, to be maximally understood, the sender must start in the most general context, or common ground, to fit the message to a wide range of different recipients, with a different states of mind. And then gradually lead the recipient to his specific message. Such a chronological process would be optimal for delivering the message as accurately as possible. (iv) Finally, to make the message clearer, both communicators should hold the models of all the relevant participants in the conversation (the recipient, the sender, their shared common knowledge, and their self-models). For principles (ii)-(iv) see Fig. 3(b).



(a) Perception and communication (b) Models in communication

Fig. 3. Communication basics

More generally, principle (iv) reveals that human-AGI communication requires something more than merely a set of models. It requires that the AGI itself hold human-like cognitive properties and capabilities, so that humans and AGI agents would be synchronized during communication and understand each other. Hence, the AGI should have characteristics such as episodic memory, continual learning, abstraction, and generalization.

Furthermore, a more broad interpretation of principle (iv), suggest that humans are actually modeling everything. Although, we model each thing differently - depending on our interaction with it. It applies to both different people (different interactions) and different groups of people. Similarly, it applies to each object/animal or their groups. Interaction with human(s) is unique because it creates a model by conversational interaction. This idea is illustrated in Fig. 4. We probably have also self-modeling, i.e. expectations from us, in the opposite direction of the interaction. In other words, how a person should behave in different groups, with different people, and with different animals and objects.

Moreover, we can model ourselves, while viewing ourselves externally (as if we are another person), to learn and perhaps change our behavior.

Additionally, we perform a passive interaction, i.e. a simple observation. For example, infants mimicking when observing other humans (such as parents or siblings).

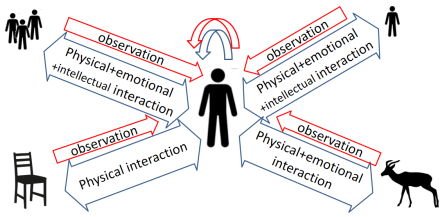


Fig. 4. Human create models from interaction.

3.2 Detailed Associative AGI model

Our AGI model is mainly originated from two aspects: (i) the phenomenon of random bouncing from one thought to another; and (ii) the communicative hypothesis of converting an idea to low-level concepts and vice versa. This model shows how information is represented. It is represented via the dynamic construction of hierarchal structures, similarly to constructing syntactic trees of sentences in NLP. Next, we introduce this model via the illustration of a story.

You can imagine first details about a scene are triggered one by one, and are placed in level 0 of the newly generated hierarchy. Next, another scene is introduced. Each scene is represented by combining all its details in level 1. At the end of chapter 1, we gathered a few scenes. After finishing chapter 2 we connect both chapters to be in level 2. And we can go on and on. See Fig 5(b).

We can see that the lowest level (0) is the most general and the most objective context, since the low-level concepts have so many associations that they lose almost entirely their specificity. However, as you go higher in the levels, the more specific the context becomes, since it is constructed underneath a more specific structure. Hence the highest levels hold the essence of all levels below. Thus, they possess the most accurate message.

The meaning of low-level concepts having the most associations is that they are connected to a huge amount of such hierarchies in the memory we gathered so far. The higher you go in the hierarchy, the fewer associations they have with other hierarchies, until you reach the levels separating this hierarchy from the rest. Furthermore, since it is a story, it has also chronology. Namely, the hierarchy has temporal direction in its levels, to enable us to retrieve it in the right order. See Fig 5(a). But the direction in connections can be extended further. It can represent different types of connections, e.g. between the levels and between the hierarchies; abstraction/generalization; various associative connections, e.g.: comparison, analogy, causality, and correlation.

Regarding the first aspect that we have mentioned earlier, we can conceptualize it as a no-purpose thinking. We can view it as a wandering between existing hierarchies, and randomly jumping from one to another, at random levels within them.

(a) Top view of the trajectory (b) Side view of the trajectory

Fig. 5. Associative thinking via associative trajectories

Associative thinking occurs all the time in our opinion. For example, daily, where the hierarchy is constructed like a long story, with some experience at the top of the story's trajectory, made out of all separate events occurred during this day. But it can also be attached to a previous hierarchy of the previous day, and even the previous week/month/year.

We use associative thinking in most of our cognitive tasks: in generating/perceiving a story/event/message, which is some (non-)linear plot of details; and in planning/simulating/problem solving, which is also a series of possible actions and outcomes.

This thinking model is like a holographic memory, where the triggered neurons are shown in Fig 5(b) on the yellow surface at the bottom. They belong to the concepts memory we have seen in Fig. 1(b). Hence this holographic memory is orthogonal to this base concepts memory. In other words, we can consider triggered neurons in this memory, producing this hierarchical dynamic structure. Of course, procedural memory can participate in this process too.

We propose that the perception operation in our AGI model would be similar to the one in [4]. In it, perception occurs via system 1, a multi-agent system, where agents compete parallelly with each other to decide which pattern is perceived correctly from the senses, and hence also decide which response is suitable for it. A similar idea is presented in [11], where this competition is via triggering all relevant neurons, and then filtering out all irrelevant ones as more clues are coming from the senses. Irrelevant ones predict worse than others, hence we are left eventually with the correct pattern. The process above describes recalling, hence if no pattern is recognized, a new hierarchy/memory is generated.

Both in [10] and in our AGI model this perception idea is expressed by ascending multiple triggered memorized hierarchies, and then descending for prediction or verification. Thus, filtering all the non-relevant memories. When encountered with partial, corrupted, or unorganized information, we can try

to validate it not only by descending, but also by moving in all the different directions in the hierarchies. For example, in recalling a story from a scene, we can move back and forth temporarily in the hierarchies, as we wish.

Associative thinking/approach is much more effective than context alone, since context might consist of many details, while associations can reduce the detail level and emphasize the abstract structure of the thing. Additionally, this allows for minimal communication and minimal resources in cognitive processes, enabling very few items in the WM, e.g. 7 ± 2 items.

It is important to note that this is a data representation model, not yet developed to the actual NN model to construct it. Emerging hierarchies in the WM can be implemented e.g. by some non-parametric method, such as via decision trees, since their structure is dynamic. Moreover, we can store the number of visitations of each node and connection in these hierarchies, to distinguish this way STM from LTM.

Additionally, our AGI model is mature, i.e., it is in the state of adulthood, which is the state reached after there has been some learning stabilization. Hence, this model also lacks the evolution of memory till its mature state. Thus, it is missing all the primary learning and adaptation. It could be fulfilled, for example, via self-supervising learning of predicting the next sensory inputs.

Finally, this model has many implications, similarities with other techniques, examples, and other more thorough considerations, which should be deeply discussed in a much broader paper.

3.3 Memories in the AGI model

Besides having our associative hierarchical structures, as elements in some memory, we also should address the memory structure itself.

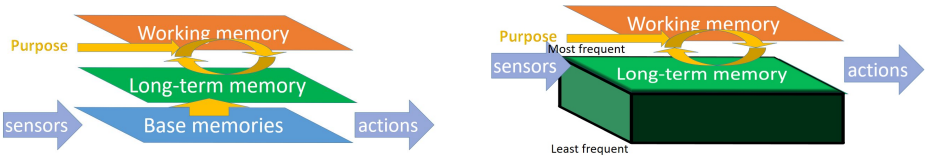
As in humans, systems 0,1 and 2 [19] should be realized here too. System 0 and 1 are expressed when most frequent memory is used, in cases when automatic or no-thinking tasks are performed. Whereas system 2 expressed by thinking, such as in problem solving, and it activates LTM and WM. The AGI model also includes cases where the system is fully utilized, i.e. simultaneously thinking and performing automatic tasks.

We can assume that simple sensory perception is using base memory, similar to system 0 automatic system (no thinking), see Fig. 1(b). Then it provokes LTM concepts or events, "uploading" them to the WM (or STM), see Fig. 6(a). During the sleeping period, the system somehow decides what to consolidate into LTM and what not, due to unimportance or similar memories that already exist there. LTM and WM do not have direct contact with the sensors and executions, perhaps since this is abstract thinking, in which the thinking, depending on some externally-driven task, is moving in purposeful trajectories/hierarchies, mostly regardless to the inputs.

We assume that humans have permanent associative wandering in LTM, producing some final or intermediate results that are updated in WM. Differently, the wandering in AGI must have some purpose. Hence there are some external instructions inserted in this process, guiding it. See Fig. 6(a).

We believe that humans solve any situation/problem this way, i.e. by jumping associatively from element to element with some guiding will, searching for something, meanwhile gathering some intermediate insights, to eventually resolve with some response (good/no/bad solution).

Alternatively, we can regard the base memories, to be simply a part of the LTM. Hence, they represent the most frequent (nearly automatic) part in it. Thus, the least frequently used memory is at the bottom, while the most used memory is at a higher level, while WM serves as the currently used memory, and is located on top of this LTM unit. See Fig. 6(b).



(a) Discrete memories interaction. (b) Continuous memories interaction.

Fig. 6. Memories in the associative thinking model.

4 Generalization in AGI

Generalization interestingly can occur by somehow abstracting out different contexts and grouping the commonalities. For example, when one sees dogs in different circumstances, and for each one of them he is being told that it is a dog, then he connects all these events together, to learn some operational characterization: they have attributes like fur, small bodies, and their unique behavior. Similarly, we learn math by abstracting out the specifics of the many examples we learn, left out eventually with an exact algorithm for doing math. Furthermore, in any skill and action, we can generalize beyond some specific object, to perform the same series of actions over other objects as well. Hence, humans prefer a rule-based approach, since it encapsulates many scenarios, instead of low-level specific examples.

If we combine this AGI characteristic with our need to model everything we interact with (see 3.1), we come up with one possible insight. We need some sort of reorganization of previous data, to turn it into abstract models, on which we can perform predictions. Anything else, which is not modeled, is not assigned for prediction. Models are the most efficient knowledge representation, since beyond prediction they can also simulate different scenarios, e.g. answering questions, understanding different aspects of a concept, and applying counterfactuals.

References

1. Ahmad, W.U., Chang, K.W., Wang, H.: Context attentive document ranking and query suggestion. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 385–394 (2019)
2. Cerri, R., Barros, R.C., De Carvalho, A.C.: Hierarchical multi-label classification using local neural networks. Journal of Computer and System Sciences **80**(1), 39–56 (2014)
3. Cheng, H., Lian, D., Deng, B., Gao, S., Tan, T., Geng, Y.: Local to global learning: Gradually adding classes for training deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4748–4756 (2019)
4. Daniel, K.: Thinking, fast and slow (2017)
5. Desta, M.T., Chen, L., Kornuta, T.: Object-based reasoning in vqa. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1814–1823. IEEE (2018)
6. Diao, X., Li, X., Huang, C.: Multi-term attention networks for skeleton-based action recognition. Applied Sciences **10**(15), 5326 (2020)
7. Dong, J., Li, X., Snoek, C.G.: Predicting visual features from text for image and video caption retrieval. IEEE Transactions on Multimedia **20**(12), 3377–3388 (2018)
8. Ge, L., Li, S., Wang, Y., Chang, F., Wu, K.: Global spatial-temporal graph convolutional network for urban traffic speed prediction. Applied Sciences **10**(4), 1509 (2020)
9. Gu, J., Zhao, H., Lin, Z., Li, S., Cai, J., Ling, M.: Scene graph generation with external knowledge and image reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1969–1978 (2019)
10. Hawkins, J., Blakeslee, S.: On intelligence: How a new understanding of the brain will lead to the creation of truly intelligent machines. Macmillan (2007)
11. Hawkins, J., Lewis, M., Klukas, M., Purdy, S., Ahmad, S.: A framework for intelligence and cortical function based on grid cells in the neocortex. Frontiers in neural circuits **12**, 121 (2019)
12. Huang, K., Ma, X., Song, R., Rong, X., Tian, X., Li, Y.: An autonomous developmental cognitive architecture based on incremental associative neural network with dynamic audiovisual fusion. IEEE Access **7**, 8789–8807 (2019)
13. Hwang, K., Sung, W.: Character-level language modeling with hierarchical recurrent neural networks. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5720–5724. IEEE (2017)
14. Keysermann, M.U., Vargas, P.A.: Towards autonomous robots via an incremental clustering and associative learning architecture. Cognitive Computation **7**(4), 414–433 (2015)
15. Li, A., Lu, Z., Guan, J., Xiang, T., Wang, L., Wen, J.R.: Transferrable feature and projection learning with class hierarchy for zero-shot learning. International Journal of Computer Vision pp. 1–18 (2018)
16. Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.: Scene graph generation from objects, phrases and region captions. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1261–1270 (2017)
17. Ma, C.Y., Kadav, A., Melvin, I., Kira, Z., AlRegib, G., Peter Graf, H.: Attend and interact: Higher-order object interactions for video understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6790–6800 (2018)

18. Nguyen, D.K., Okatani, T.: Multi-task learning of hierarchical vision-language representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10492–10501 (2019)
19. Prokopchuk, Y., Nosov, P., Zinchenko, S., Popovych, I.: New approach to modeling deep intuition. In: Materials of the 13th Scientific and Practical Conference «Modern Information and Innovative Technologies in Transport (MINTT-2021)». Kherson, Ukraine: XSMA. pp. 37–40
20. Sanh, V., Wolf, T., Ruder, S.: A hierarchical multi-task approach for learning embeddings from semantic tasks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 6949–6956 (2019)
21. Shen, F., Ouyang, Q., Kasai, W., Hasegawa, O.: A general associative memory based on self-organizing incremental neural network. Neurocomputing **104**, 57–71 (2013)
22. Socher, R., Karpathy, A., Le, Q.V., Manning, C.D., Ng, A.Y.: Grounded compositional semantics for finding and describing images with sentences. Transactions of the Association for Computational Linguistics **2**, 207–218 (2014)
23. Suzuki, M., Yoshida, Y.: On the development and utility of action control individuality for semi-autonomous intelligent robots. In: 2019 18th European Control Conference (ECC). pp. 3550–3555. IEEE (2019)
24. Wehrmann, J., Cerri, R., Barros, R.: Hierarchical multi-label classification networks. In: International Conference on Machine Learning. pp. 5075–5084 (2018)
25. Weston, J., Chopra, S., Bordes, A.: Memory networks. arXiv preprint arXiv:1410.3916 (2014)
26. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057 (2015)
27. Xu, R., Xiong, C., Chen, W., Corso, J.J.: Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In: AAAI. vol. 5, p. 6. Citeseer (2015)