A Proposal for Common Dataset in Neural-Symbolic Reasoning Studies

Ozgur Yilmaz, Artur d'Avila Garcez, and Daniel Silver

Turgut Ozal University, Computer Science Department, Ankara Turkey
City University London, Department of Computer Science, London UK
Acadia University, Jodrey School of Computer Science, Nova Scotia Canada,
ozyilmaz@turgutozal.edu.tr,a.garcez@city.ac.uk
danny.silver@acadiau.ca

Abstract. We promote and analyze the needs of a common publicly available benchmark dataset to be used for neural-symbolic studies of learning and reasoning. The recently released Visual Genome repository is proposed as a suitable dataset to meet these needs. Along with the original tasks that were suggested by the Visual Genome creators, we propose neural-symbolic tasks that can be used as challenges to promote research in the field and competition between lab groups.

Keywords: Neural-symbolic computing, common dataset, relational learning, reasoning, visual entailment

1 Introduction

Research into neural-symbolic integration seeks to combine learning from sub-symbolic vector representations of data and concepts with symbolic reasoning and knowledge representation. [4–7]. In order to integrate the sub-symbolic neural representations of sensory data with the symbolic knowledge tools developed within AI over the last 60 years of research, a mathematical toolbox has to be designed that has the capability of translating between different levels of knowledge representation. In its infancy, by comparison, neural-symbolic studies are promising ventures towards an AI system which can recognize patterns in sensory data and reason about such commonsense patterns and knowledge.

The existence of a satisfactory dataset has been shown to be fruitful in many computer science fields. It enables a fair comparison of existing approaches and encourages competition. It should be mentioned also that benchmark datasets introduce a potential bias, as problems not covered by the benchmark receive less attention. Due to the growth of the web and abundance of data, ease of annotation by crowd-sourcing and the desire to build accurate applications, many large datasets have been developed within computer vision, such as ImageNet [1], Microsoft COCO [2] and VQA [3]. The size of these datasets is large to accommodate very complex models, specifically deep neural networks, with the promise of use as technological tools in everyday life such as image search and retrieval, or image captioning for the visually impaired.

There are valuable experimental studies in neural-symbolic reasoning, however there is a need for a common publicly-available benchmark dataset to encourage progress and communications in the field. Datasets exist in Statistical Relational Learning (SRL) and Inductive Logic Programming (ILP) which may be suitable for neural-symbolic integration. Recently developed datasets for vision-language tasks such as image caption generation and visual question answering seem attractive for neural-symbolic studies since they require complex pattern recognition over images and symbol manipulation of language. Yet, symbol manipulation and reasoning are limited to image description text that is unstructured, and not amenable to traditional natural language processing (NLP) tools. The ideal dataset for neural-symbolic studies should include a large and complex raw data set for sub-symbolic systems to learn effective and discriminative representations, as well as a formal representation of the raw data (a knowledge-base in first order logic) for symbolic systems to learn general rules and perform logical inference. Existence of both complex sub-symbolic data and its high level symbolic interpretation is essential for developing the above-mentioned translational methods between the two forms of representations which are at the heart of neural-symbolic integration.

In this paper, we propose the use of the Visual Genome dataset [13] as the best challenge benchmark dataset for neural-symbolic integration. The dataset is valuable "as is" towards the goals of neural-symbolic integration, however, we also suggest additional features and challenge tasks for the dataset to meet a wider range of research objectives within neural-symbolic computing.

In Section 2, we recall the goals of neural-symbolic integration (NSI). In Section 3, we describe the visual genome (VG) dataset. In Section 4, we list existing applications of VG to NSI. In Section 5, we propose the new applications and extensions, and in Section 6, we conclude the paper.

2 Neural-Symbolic Reasoning

Neural-symbolic systems [8] integrate logical reasoning and statistical learning by offering sound translation algorithms between network and logic models. They contain three main components: (1) knowledge encoding and reasoning in neural networks, (2) knowledge evolution and network learning, and (3) knowledge extraction from trained networks. In a neural-symbolic system, neural networks provide the machinery for efficient computation and robust learning, while logic provides high-level representations, reasoning and explanation capabilities to the network models, promoting modularity, facilitating validation and maintenance and enabling a better interaction with existing systems.

Neural-symbolic systems have had important applications in diverse areas such as bioinformatics, fraud prevention, assessment and training in simulators, cognitive robotics, general game playing, image, audio and video classification, software verification, and the semantic web. Nevertheless, a major challenge that remains is how to effectively benefit from both (i) robust statistical methods that work well on real-valued vectors and (ii) rich and interpretable represen-

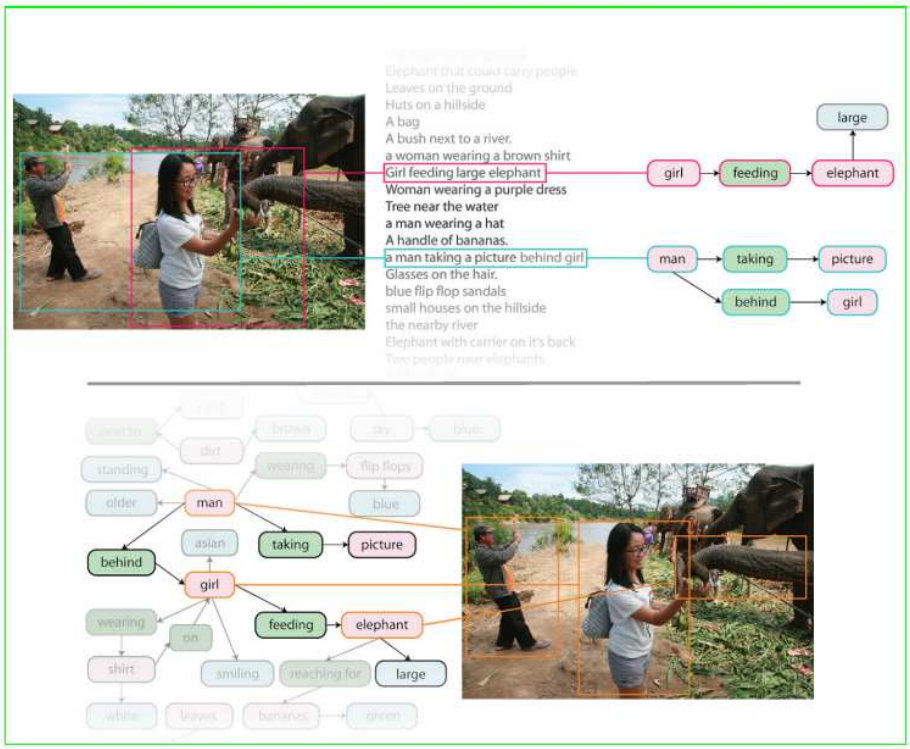

Fig. 1. From perceptual awareness to cognitive understanding of images [13]: images are annotated with numerous region descriptions, objects, attributes, and relationships, e.g.: "girl feeding large elephant" and "a man taking a picture behind girl" (top picture), with objects (e.g. elephant), attributes (e.g. large) and their relationships (e.g. feeding) described in the bottom picture.

tations which enable explanations to be reasoned about and transferred across applications. The above requires the effective translation of relational symbolic knowledge for use by statistical methods which work well with vectors (without the need for grounding all instances of the knowledge-base into the model of choice) and the effective extraction of compact and rich representations from vector-based models following neural network learning.

The emergence of symbolic representations is natural in any complex domain associated with large collections of data. In fact, symbolic representations seem critical to the solution of many interesting challenges involving big data. Consider, for example, the recent AlphaGo experiment[1] or the requirements of life-long learning[9] or intelligent agents who interact with the environment. The above is particularly relevant when neural-symbolic integration meets computer

[1] https://www.technologyreview.com/s/601072/five-lessons-from-alphagos-historic-victory/

vision. As pointed out at a recent Dagstuhl seminar on neural-symbolic computing [2], a serious challenge in the field is the lack of specifically relevant and systematic evaluation mechanisms. The benchmark-based approach, which is useful in some cases, is very limited in others, including the benchmarks used in Statistical Relational Learning (SRL) and Inductive Logic Programming (ILP) [11, 12]. In particular, when the goal is (i) to evaluate how well a system integrates learning and reasoning, or (ii) to evaluate how useful or interpretable the learned descriptions are, existing benchmarks fall short: SRL will tend to ground all representation without a focus on first-order reasoning; ILP tend not to handle real-valued vectors or provide for robust learning. Neural-symbolic systems seek to benefit from the knowledge representation and reasoning capacities of logical symbolic representations, and the robust learning capacities of neural networks, reconciling the logical nature of reasoning and the statistical nature of learning [10]. The provision of a data challenge as proposed here should promote the fair comparative evaluation of: (1) effective learning from noisy data and (2) reasoning about what has been learned.

3 Visual Genome

Visual understanding is suggested to be an AI-complete problem [17], therefore it is a challenging testbed for neural-symbolic studies. A genuine understanding of a visual scene requires detecting objects, recognizing attributes of objects and inferring their interactions and relationships. Understanding images thoroughly requires a grounding of visual concepts onto language and a formalized representation of the components of an image, as stated in [13]: "existing models would be able to detect discrete objects in a photo but would not be able to explain their interactions or the relationships between them. Such explanations tend to be cognitive in nature, integrating perceptual information into conclusions about the relationships between objects in a scene...". Going from **perceptual** to **cognitive**, from image to language, demands a range of operations that must lift the representation from subsymbolic to symbolic, which it is at the core of neural-symbolic computation studies.

Similar to previous attempts on visual knowledge bases [14–16], the Visual Genome provides a large set of images and annotations of image regions which is formalized as a scene graph of objects and their relations. Images in the dataset (see Figure 1) contain multiple image regions each having multiple object instances. The attributes of object instances and their relationship (predicate) with other objects are also recorded. Region graphs are combined to form a scene graph of an image, which can be translated into a knowledge base, as well as plain language using basic NLP tools. The concepts in the dataset can be linked to existing knowledge in other datasets or systems because all objects, attributes and relationships in each image in the Visual Genome can be mapped onto a corresponding WordNet ID, called a synset ID [18]. As described in the Visual

[2] http://www.dagstuhl.de/14381

Genome project main webpage[3], the dataset contains 108,249 images (with an average image size of 500 pixels), 4.2 million region descriptions (with around 75,000 unique image objects), 1.7 million visual question-answers, 2.1 million object instances, 1.8 million attributes (40,500 unique attributes), 1.8 million relationships (40,500 unique relationships), 1.5 million object-object relationship instances, 1.6 million attribute-object instances, 108,249 total scene graphs and 3,788,715 total region graphs.

Therefore, visual genome contains a dense formal knowledge representation of images suitable to be manipulated by symbolic computation approaches, as well as sensory image data ready to be recognized and analyzed by connectionist methods. For vision/language tasks, region descriptions and question-answer pairs related to images are also provided. Overall the dataset enables a wide range of scene understanding applications, which typically require high level symbol manipulation and language processing. Furthermore, the symbolic formalism contained in Visual Genome favors first order logic representations and relational learning. The scale of the dataset means that approaches which perform grounding will probably be less effective than truly relational approaches. In other words, Visual Genome targets a major, arguably the most important, open challenge in neural-symbolic integration: the effective handling of learning from real-valued vectors and reasoning from rich knowledge representations.

4 Existing Applications on the Visual Genome

The developers of the dataset have introduced some interesting tasks, two of which are explained below.

4.1 Attribute and Relationship Prediction

Object class prediction and object detection is at the center of computer vision studies, and successful deep learning algorithms [20, 19] dominate the field. The Visual Genome enables dense and accurate attribute/predicate estimation; bounding boxes that contain an object can be analyzed for predicting attribute/predicate dimensions.

Researchers have found that learning attribute-object class pairs for each bounding box dramatically improves attribute prediction performance possibly due to the unique association of some attributes with specific object classes. Similarly, learning *subject class - predicate - object class* triplets instead of *predicate* only, can improve performance. This is again due to the fact that some relationships occur only among a very small subset of objects classes (e.g. the *drive* predicate accepts the *person* subject exclusively). Such applications can be considered an instantiation of collective classification in relational learning [32].

[3] https://visualgenome.org/

4.2 Caption Generation and Visual Question Answering

The existence of region descriptions and question-answer pairs on images facilitate vision-language processing tasks. The visual representation of images and regions can be used in a generative architecture to produce syntactically and semantically correct text such as automated image caption generation. Recurrent neural network algorithms have been deployed successfully [21] for such vision-language applications. However, a major challenge has been judging performance accuaracy of automated image captioning, e.g. is "A cat is beside a dog under a parked car" the same as "A car is parked over a dog and a cat"?

5 Suggested Applications and Extensions

Visual Genome holds a very rich representation of the visual world, ready to be exploited by cognitive tasks. We envision that the dataset can be used for a wide set of experimental paradigms, or can be extended by additional crowd-sourced annotations as required. We provide a set of novel tasks, which is not meant to be exhaustive. Along with the task definitions, we provide a high level algorithmic description of how to tackle them in order to illustrate how neural-symbolic studies would benefit from the dataset.

Generally, neural-symbolic approaches would ground the sensory data onto symbols and manipulate those, or perform vector algebra on neural representations to form a hierarchy of concepts and rules on the vector space. The main questions are how to accurately and effectively ground the data or how to manipulate the vectors as done with symbols in AI, as well as how to use both mathematical tools simultaneously.

5.1 Visual Entailment

Comprehension of entailment and contradiction in sentences is an important part of language processing. In textual entailment tasks, two sentences need to be understood and the system has to decide whether they contradict each other, they are neutral (unrelated) or they entail each other. The scene graph in Visual Genome is already a valuable asset in the textual entailment task, as utilized in a study in [22], yet there is much more to be done. We propose a new task called visual entailment in which images, relationships and scene graphs are used to detect entailment and contradictions. This is a very natural use of the image representation for neural-symbolic tasks: inference can be performed at the symbolic level if images are grounded onto class and attribute predictions by a classifier, or inference can be partly done at the sub-symbolic level using the neural representations of images. Sub-symbolic computation requires an algebra on semantically meaningful vector representations [33].

We present two image bounding boxes, then ask whether there is entailment/contradiction/neutralism. The decision is very much related to the possible relationships between image boxes. If there is a relationship then the answer

is entailment, if not, it can be neutral or contradiction, depending on the compatibility with commonsense. A *car* and a *tire* imply entailment, a *car* and a *house window* may be neutral but a *car* and a *kitchen sink* is probably a contradiction. The output can be set to a range between -1 (contradiction) and 1 (entailment), at which point the supervised learning may become a regression task instead of classification. It should be noted that visual entailment aims at finding relationships between two scenes thus the proposed task is closely related to link prediction in relational learning, where the goal is to learn the existence of a relationship. Therefore, the idea of contradiction in visual entailment means learning the lack of a relationship, which is not the case in textual entailment task.

The task becomes even more interesting and similar to textual entailment if we allow one or two of the image boxes to be a large region with multiple objects and relationships in it. Then the system needs to analyze the congruence of region graphs, hence knowledge bases. A subsymbolic approach would use neural embeddings of the image boxes to generate rules of entailment on the vector space possibly using a vector symbolic architecture [23, 24] and/or an attention-memory computation framework [25]. A symbolic approach would use the class/attribute/relationship predictors to go up to knowledge base level.

5.2 Scene Graph Estimation

Possibly the hardest task is generating the scene graph of an image because the graph holds the complete high level information regarding the image, we need to go from the sensory to the most complete cognitive level. It requires to focus on specific bounding boxes in the image, estimate object/attribute labels and jump to other image boxes while predicting relationships between them. Thus the graph can be built part by part possibly with multiple passes on the same image region. These multiple passes can possibly be hierarchical in nature, extracting graph structure from coarse to fine details. This workflow resembles the strategy of recurrent architectures with attention-memory mechanisms[26]. Another strategy more in the flavor of neural-symbolic computation would be training the system by encoding regions and scenes in the training dataset with fixed length vector representations and forming a "graph knowledge-base", then matching the test region with the knowledge base to obtain the most representative and similar region description in the training set. After this initial estimation, fine-tuning can optionally be done with the recurrent architectures with attention-memory mechanisms.

The main challenge in this task is related to the variable binding problem: multiple instances of the same object/concept/relationship as it appears in different times and context need to reuse a common function with possibly different values. One possible solution to this problem is transferring learned representation across different contexts [28].

5.3 Visual Rule Extraction and Analogy

Is it possible to mine the scene graphs for extracting logical clauses such as "If **Man** not(**Standing**) Then **Man SitsOn**(**Something**)"? This capability is essential for forming the visual commonsense knowledge mentioned earlier. In a similar flavor, visual analogies can be made such as "**Leg** is to **Man** as **Tire** is to **Car**". These are strictly in the domain of symbolic computation when images are grounded to class/attributes and predicate predictions are processed in the scene graph. However, what if we wanted to retrieve rules and analogies directly using image portions? Then, neural representations of images would need to be processed to harvest conditional and analogical "statements" at the sub-symbolic level [27, 29, 34]. The rules and analogies that form the commonsense knowledge and representations of the images are expected to live on the same space, which is essential for combining connectionist and symbolic capabilities. Visual rule extraction can also be tackled with inductive bias transfer of neural networks across different task domains [30]. More interesting approaches would be again hybrid ones that utilizes the symbolic mechanisms along with vector algebra.

5.4 Collective Classification

Another relevant relational learning task is collective classification: simultaneous prediction of the class of several object bounding boxes in a region given their attributes or relations. This is superficially similar to attribute and relation prediction tasks already examined in [13], yet the proposed task is not bounded by pairwise bounding box queries but all the objects in a region or even in a whole image can be considered for a more challenging collective classification. This is directly related with multiple task learning and inductive bias transfer between many tasks, as studied from a neural-symbolic perspective in [31].

5.5 Unsupervised co-training of a subject class - predicate - object class using images and symbols

Related to prior work discussed in Section 4.1 is the unsupervised co-training of subject class - predicate - object class triples using both image data as well as symbolic logic. The intention is to show that one can learn an unsupervised generative model (e.g. stacked Restricted Boltzmann Machines) that are capable of reconstructing the images given the symbols, and the symbols given the images. Here, symbols could be represented as combinations of textual inputs or as images themselves.

6 Conclusion

We have proposed Visual Genome as a challenge and benchmark dataset for neural-symbolic integration. Along with the original tasks that were suggested by the Visual Genome creators, we also identify tasks specific for neural-symbolic

integration, in particular combining learning from real-valued vectors and reasoning from rich relational knowledge representations, to promote research in the field and competition between lab groups.

Acknowledgments. We would like to thank the reviewers for detailed and very beneficial comments on the paper. Ozgur Yilmaz is supported by The Scientific and Technological Research Council of Turkey (TUBİTAK) Career Grant, No: 114E554.

References

1. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
2. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*, pages 740–755. Springer, 2014.
3. Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.
4. Tarek R Besold and Kai-Uwe Kühnberger. Towards integrated neural–symbolic systems for human-level AI: Two research programs helping to bridge the gaps. *Biologically Inspired Cognitive Architectures*, 14:97–110, 2015.
5. Artur S. d'Avila Garcez, Luis C Lamb, and Dov M Gabbay. *Neural-symbolic cognitive reasoning*. Springer Science & Business Media, 2008.
6. Sebastian Bader and Pascal Hitzler. Dimensions of neural-symbolic integration-a structured survey. *arXiv preprint cs/0511042*, 2005.
7. Artur S. d'Avila Garcez, Tarek R Besold, Luc de Raedt, Peter Földiak, Pascal Hitzler, Thomas Icard, Kai-Uwe Kühnberger, Luis C Lamb, Risto Miikkulainen, and Daniel L Silver. Neural-symbolic learning and reasoning: contributions and challenges. In *Proceedings of the AAAI Spring Symposium on Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches, Stanford*, 2015.
8. Artur S. d'Avila Garcez, Luís C. Lamb, and Dov M. Gabbay. *Neural-Symbolic Cognitive Reasoning*. Cognitive Technologies. Springer, 2009.
9. Daniel L Silver, Qiang Yang, and Lianghao Li. Lifelong machine learning systems: Beyond learning algorithms. In *in AAAI Spring Symposium Series*. Citeseer, 2013.
10. Leslie G. Valiant. Knowledge infusion. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA*, pages 1546–1551, 2006.
11. Jue Wang and Pedro M. Domingos. Hybrid markov logic networks. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008*, pages 1106–1111, 2008.
12. Luc De Raedt, Kristian Kersting, Sriraam Natarajan, and David Poole. *Statistical Relational Artificial Intelligence: Logic, Probability, and Computation*. Synthesis

Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2016.
13. Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalanditis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016.
14. Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. Building a large-scale multimodal knowledge base system for answering visual queries. *arXiv preprint arXiv:1507.05670*, 2015.
15. Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. Neil: Extracting visual knowledge from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1409–1416, 2013.
16. Fereshteh Sadeghi, Santosh K Divvala, and Ali Farhadi. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1456–1464. IEEE, 2015.
17. Dafna Shahaf and Eyal Amir. Towards a theory of ai completeness. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, pages 150–155, 2007.
18. George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
19. Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision andpattern recognition*, pages 580–587, 2014.
20. Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Computer Science Department, University of Toronto, Tech. Rep*, 2009.
21. Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.
22. Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.
23. Simon D Levy and Ross Gayler. Vector symbolic architectures: A new building material for artificial general intelligence. In *Proceedings of the 2008 conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, pages 414–418. IOS Press, 2008.
24. Jeff Mitchell and Mirella Lapata. Vector-based models of semantic composition. In *ACL*, pages 236–244, 2008.
25. Ivo Danihelka, Greg Wayne, Benigno Uria, Nal Kalchbrenner, and Alex Graves. Associative long short-term memory. *arXiv preprint arXiv:1602.03032*, 2016.
26. Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, pages 2431–2439, 2015.
27. Ozgur Yilmaz. Symbolic computation using cellular automata-based hyperdimensional computing. *Neural computation*, 2015.
28. Daniel L Silver. The parallel transfer of task knowledge using dynamic learning rates based on a measure of relatedness. *Connection Science*, 8(2):277–294, 1996.

29. Ozgur Yilmaz. Analogy making and logical inference on images using cellular automata based hyperdimensional computing. In *Advances in Neural Information Processing Systems, Cognitive Computation Workshop*, pages 1–9, 2015.
30. Daniel L Silver. Selective functional transfer: Inductive bias from related tasks. In *IASTED International Conference on Artificial Intelligence and Soft Computing (ASC2001)*. Citeseer, 2001.
31. Daniel L Silver and Liangliang Tu. Image transformation: inductive transfer between multiple tasks having multiple outputs. In *Advances in Artificial Intelligence*, pages 296–307. Springer, 2008.
32. Prithviraj Sen, Galileo Mark Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher and Tina Eliassi-Rad. Collective Classification in Network Data. AI Magazine, 3(29):93–106. 2008.
33. Luciano Serafini and Artur S. d'Avila Garcez. Logic Tensor Networks: Deep Learning and Logical Reasoning from Data and Knowledge. *arXiv preprint arXiv:1606.04422*, 2016.
34. Tarek Richard Besold, Kai-Uwe Kühnberger, Artur S. d'Avila Garcez, Alessandro Saffiotti, Martin H. Fischer and Alan Bundy. Anchoring Knowledge in Interaction: Towards a Harmonic Subsymbolic/Symbolic Framework and Architecture of Computational Cognition. Artificial General Intelligence - 8th International Conference, AGI 2015, AGI 2015, Berlin, Germany, July 22-25, 2015.