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. In this study, we propose a recently released dataset to be used in neural-symbolic studies. We analyze the needs of a common neural-symbolic dataset and show how Visual Genome matches them. Along with the original tasks that were suggested in Visual Genome, we propose neural-symbolic tasks that can be used as challenges to promote competition.

Keywords: Neural-symbolic, Common Dataset, Knowledge Base, Relational Learning, Entailment

1 Introduction

Existence of a satisfactorily large dataset is shown to be very fruitful in many computer science fields. It enables a fair comparison of existing approaches and encourages competition. Due to the growth of web and abundance of data, ease of annotation by crowdsourcing and the existence of will towards building accurate applications, many large datasets have been developed in computer vision such as ImageNet [1], Microsoft COCO [2] or VQA [3]. The size of the datasets are large enough to accommodate very complex models, specifically deep neural networks which can be used in technological tools in everyday life such as image search/retrieval, image captioning for the visually impaired etc.

Neural-symbolic approaches are aiming at utilizing neural representation of data and concepts for symbolic computation of knowledge [4–7]. In order to integrate the subsymbolic neural representation of sensory data and information to the symbolic manipulation tools developed over the last 60 years of AI research, a mathematical toolbox has to be designed that has the capability to transform between levels of knowledge. In its infancy, neural-symbolic studies are promising ventures for a true AI system which has to recognize patterns in the sensory data and reason about it with common sense knowledge.

There are valuable experimental studies in neural-symbolic reasoning, however there seems to be a lack of common dataset which expected to speed up the progress in the field and transfer wisdom among different approaches. Datasets exist in Statistical Relational Learning (SRL), but they are small and almost

all of them have very limited scope and data complexity. Recently developed datasets on vision-language tasks such as image caption generation and visual question answering are certainly attractive for neural-symbolic studies since they require complicated pattern recognition on images and symbol manipulation of language. Yet, symbol manipulation aspect is limited due to the fact that provided text is unstructured, thus forming the knowledge base is limited with NLP performance on the image descriptions. The ideal dataset for neural-symbolic studies has to include complex enough raw data (sensory or textual) for sub-symbolic systems to learn effective and discriminative representations, as well as a formal representation of the data (i.e. knowledge base in first order logic) for symbolic systems to learn general rules and make logical inference. Existence of both complex data and its high level interpretation is essential for developing the necessary transformational methods of representation, which is missing in existing datasets. In this study, we are proposing that Visual Genome dataset

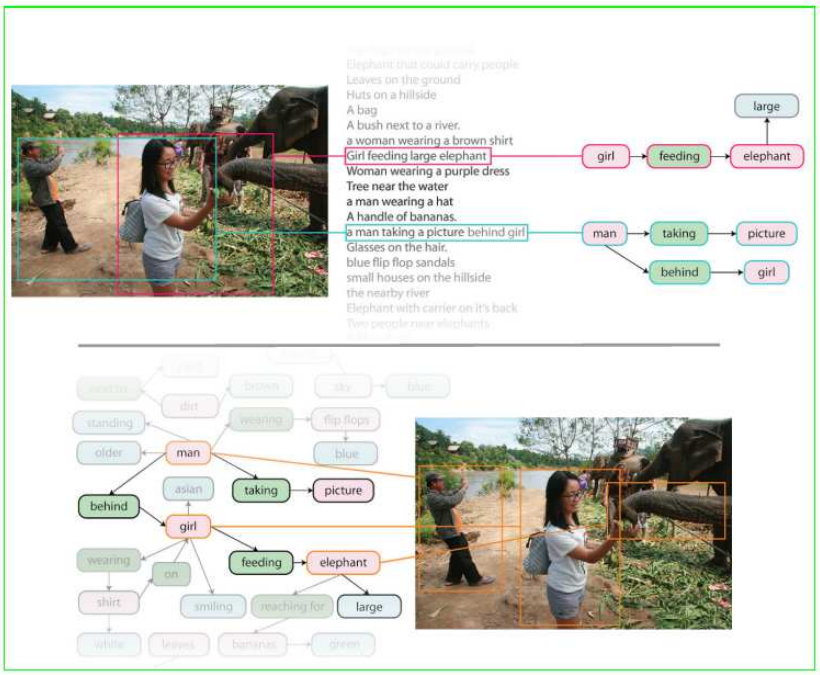

Fig. 1. "An overview of the data needed to move from perceptual awareness to cognitive understanding of images. .... a dataset of images densely annotated with numerous region descriptions, objects, attributes, and relationships. Region descriptions (e.g. girl feeding large elephant and a man taking a picture behind girl) are shown (top). The objects (e.g. elephant), attributes (e.g. large) and relationships (e.g. feeding) are shown (bottom). ... dataset also contains image related question answer pairs (not shown)." Adapted and quoted from [13].

[13] is a good starting point and a very good candidate to fulfill this mission. The dataset is very valuable "as is" for neural-symbolic goals, but we are suggesting to develop additional features in the dataset to include a wider range of experimental approaches.

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 representation 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 representations which enable explanations to be reasoned about and transferred across applications. The above requires the effective translation of relational knowledge for use by such 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 such a model following 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 most 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 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

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

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

ground all representation without a focus on first-order reasoning; ILP will not handle real-valued vectors or focus on robust learning. With this in mind, i.e. given real-valued vectors and rich representations, neural-symbolic systems seek to benefit from the knowledge representation and reasoning capacities of such (logical) representations, and the robust learning capacities of neural networks, reconciling the logical nature of reasoning and the statistical nature of learning [10]. Neural-symbolic integration, therefore, seeks to enable - and the provision of a data challenge as proposed here should promote the fair comparative evaluation of - effective learning from noisy data and reasoning about what has been learned.

3 Visual Genome

Visual understanding is suggested to be an AI-complete problem [14], thus vision 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. As it is stated in the paper [13], understanding images thoroughly requires a grounding of visual concepts to language and a formalized representation of the components of an image. Another way of explaining the unique stance of the dataset is given as: " existing models would be able to detect discreet 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 to lift the representation from subsymbolic to symbolic, which it is at the core of neural-symbolic computation studies.

Visual genome provides a large set of images and annotations of image regions (Figure 1) which is formalized as a scene graph of objects and their relations. Images in the dataset 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 dataset are canonicalized to its corresponding WordNet [15] ID (called a synset ID).

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, most of which require a high level symbol manipulation and language processing.

4 Existing Applications on the Visual Genome

The developers of the dataset introduce some interesting tasks, two sets 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 [17, 16] dominate the field. Visual Genome enables dense and accurate attribute/predicate estimation, and in these set of proposed tasks, bounding boxes that contain an object are analyzed for predicting attribute/predicate dimensions.

It is observed 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 dramatically improves performance. This is again due to the fact that some relationships occur only among a very small subset of objects classes (eg. drive predicate accepts person subject exclusively). These two applications can be considered an instantiation of collective classification task in relational learning literature.

4.2 Caption Generation and Question Answering

Existence of region descriptions and question-answer pairs on images facilitate language processing tasks. The visual representation of images and regions can be used in a generative architecture to produce syntactically and semantically correct text. Recurrent neural network algorithms are successfully deployed [18] for vision-language applications.

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 crowdsourced annotations as required. We provide a set of novel tasks, which is not comprehensive. Along with the task definitions, we provide a high level algorithmic description of tackling them in order to illustrate how neural-symbolic studies would benefit from the dataset.

Generally neural-symbolic approaches would ground the sensory data to symbols and manipulate them, or perform vector algebra on neural representation to form a hierarchy of concepts and rules on this vector space. The main questions are how to accurately and effectively ground the data or how to manipulate vectors as if they are symbols, 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 task, 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 very valuable asset in textual entailment task and it is utilized in a study [19], 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 visual entailment and contradictions. This is a very natural employment of the image representation for neural-symbolic tasks: the inference can be performed in the symbolic domain if images are grounded to class and attribute predictions through a classifier, or inference can be partly done on the subsymbolic space using the neural representation of images. Subsymbolic computation requires a sort of algebra on semantically meaningful vector representations.

We present two image bounding boxes, then ask whether there is entailment/contradiction/neutralism. The decision is very much related to the possible relationship between image boxes. If there is a relationship the answer is entailment, if not it can be neutral or contradiction depending on the compatibility with common sense. A car and a tire is entailment, a car and a house window is neutral but a car and a kitchen sink is contradiction. The ouput can be set to a range between -1 (contradiction) and 1 (entailment), after which the supervised objective becomes regression instead of classification. The proposed task is closely related to link prediction in relational learning literature, for which the existence of a relationship is questioned.

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 [20, 21] and/or an attention-memory computation framework [22]. A symbolic approach would use the class/attribute/relationship predictors to go up to knowledge base level. Hybrid approaches can be utilized that exploit both domains, and exchange information at some 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[23]. 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 [25].

5.3 Visual Rule Extraction and Analogy

Is it possible to mine the scene graphs for extracting horn clauses such as "if **Man** not(**Standing**), then **Man SitsOn**(**Something**)"? This capability is essential for forming the visual common-sense knowledge. 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 if images are grounded to class/attribute and predicate predictions and processed representation is the scene graph. However, what if we want to retrieve rules and analogies directly using image portions? Then, neural representations of images need to be processed to harvest conditional and analogical "statements" in the subsymbolic level [24, 26]. The rules and analogies that forms the common sense knowledge and representation of the image is 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 [27]. 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 and/or relations. This is superficially similar to attribute and relation prediction task 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 have been studied before from neural-symbolic perspective [28].

Acknowledgments. 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 SD'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 dAvila 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. Dafna Shahaf and Eyal Amir. Towards a theory of ai completeness. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, pages 150–155, 2007.
15. George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

16. 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.
17. Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Computer Science Department, University of Toronto, Tech. Rep*, 2009.
18. 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.
19. 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.
20. 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.
21. Jeff Mitchell and Mirella Lapata. Vector-based models of semantic composition. In *ACL*, pages 236–244, 2008.
22. Ivo Danihelka, Greg Wayne, Benigno Uria, Nal Kalchbrenner, and Alex Graves. Associative long short-term memory. *arXiv preprint arXiv:1602.03032*, 2016.
23. Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, pages 2431–2439, 2015.
24. Ozgur Yilmaz. Symbolic computation using cellular automata-based hyperdimensional computing. *Neural computation*, 2015.
25. 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.
26. 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.
27. Daniel L Silver. Selective functional transfer: Inductive bias from related tasks. In *IASTED International Conference on Artificial Intelligence and Soft Computing (ASC2001*. Citeseer, 2001.
28. 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.