Commonsense Reasoning for Identifying and Understanding the Implicit Need of Help and Synthesizing Assistive Actions

Maëlic Neau[1,2,3], Paulo Santos[2], Anne-Gwenn Bosser[1], Nathan Beu[3] and Cédric Buche[1,3]

[1]*Lab-STICC/ENIB, France*

[2]*College of Science and Engineering, Flinders University of South Australia,*
1284 South Rd, Clovelly Park SA 5042, Australia

[3]*CNRS, International Research Lab "CROSSING", Adelaide, Australia*

Abstract

Human-Robot Interaction (HRI) is an emerging subfield of service robotics. While most existing approaches rely on *explicit* signals (i.e. voice, gesture) to engage, current literature is lacking solutions to address *implicit* user needs. In this paper, we present an architecture to (a) detect user implicit need of help and (b) generate a set of assistive actions without prior learning. Task (a) will be performed using state-of-the-art solutions for Scene Graph Generation coupled to the use of commonsense knowledge; whereas, task (b) will be performed using additional commonsense knowledge as well as a sentiment analysis on graph structure. Finally, we propose an evaluation of our solution using established benchmarks (e.g. ActionGenome dataset) along with human experiments. The main motivation of our approach is the embedding of the perception-decision-action loop in a single architecture.

Keywords
Commonsense Reasoning, Knowledge Graph, Vision-to-Language, Cognitive Robotics

1. Introduction

Detecting and understanding user's intentions and needs is the fundamental backbone of service robotics. This question relates to how high-level, abstract, concepts can be inferred from raw sensor data (an issue intimately related to the symbol grounding problem) [1]. Traditional approaches to this problem in robotics use *explicit* signals from the user such as voice [2], gesture [3] or even touch [4]. However, the deployment of service robots in assisting activities of daily life (ADL), especially for impaired or elderly people, is leading the way to more *implicit* interactions with autonomous agents. Previous work has been introduced to understand user's implicit intentions in service robotics. Some use external context to predict user's intentions

[5] while others rely on gaze-based signals [6]. But to the best of our knowledge none of them integrates the use of external commonsense knowledge of the world.

The present paper tackles an important part of this issue, where only non-verbal, visually-observable data are taken into account. We present a system that is able to understand the implicit user needs of help in the realisation of a task and to provide a relevant assistive action, inspired by the way humans act based on commonsense reasoning. In a general sense, the use of commonsense reasoning in the present work can be summarised with the following assumption: a factor of human assistance to one another in the realisation of a task is the perception of danger. For instance, humans are typically able to understand (without explicit prior learning) that anything coming out of an oven is hot, and that a person should protect their hands to avoid hurting themselves. This is connected to the following definition of commonsense reasoning, from [7], p.170:

> *Commonsense causal reasoning is qualitative reasoning about the behavior of a mechanism which can be done without external memory or calculation aids, although it may draw on concepts learned from the advanced study of a particular domain, e.g. automobile mechanics, computer architecture, or medical physiology.*

For instance, in the above example, we can summarise the commonsense reasoning as the following causal relationships:

$$oven \xrightarrow{\text{produce}} heat \tag{1}$$

$$heat \xrightarrow{\text{capable of}} hurt\ skin \tag{2}$$

With a refinement on the visual features, the system is able to ground commonsense knowledge to the scene as follow:

$$heat \xrightarrow{\text{capable of}} hurt\ hand \tag{3}$$

This commonsense reasoning process will also help the robotic agent to build an assistive action. In fact, in some cases, the reasoning about the visual inputs alone is not sufficient to provide accurate help. Recalling the previous example, the system needs external knowledge to come up with the assistive action "bring gloves to the user", as a human would do:

$$glove \xrightarrow{\text{capable of}} protect\ hand \tag{4}$$

We believe that creating such relationships is possible by using commonsense knowledge databases such as the ConceptNet [8] or ATOMIC [9] datasets, and also following some of the ideas for combining logic reasoning with machine learning described in [10]. Another characteristic of our work is the use of Scene Graph (SG) [11] as a tool for knowledge representation. In fact, this type of representation could easily be enriched with external knowledge databases as they share the same data structure: graph. Finally, the use of commonsense reasoning is the part of our architecture, which gives us the possibility of understanding biases and correcting them in an incremental development.

Our position could be summarized as the following statement: from the analysis of human behavior, the use of state-of-the-art solutions from Vision-to-Language combined with Commonsense Reasoning will leverage Cognitive Robotics.

2. Related work

The task of retrieving graph representations from still images or videos is called Scene Graph Generation (SGG), in this section we review current approaches for SGG. To perform efficient reasoning, our solution integrates external commonsense knowledge. Thus, we also review solutions for knowledge-graph enrichment and completion. Finally, as our reasoning system needs to provide a sentiment analysis to retrieve the possibility of risks, we review approaches to connotation lexicon [12] (i.e. lexicon that lists words with connotative polarity).

2.1. Scene Graph Generation

Scene Graph Generation (SGG) [11] is the task of creating a grounded graph of visual entities retrieved from an image with the goal of representing attributes, objects and their relationships in a scene. Such graphs typically contain one or more triplet(s) *(head entity, relation, tail entity)*. Entities present in a scene graph could be person (e.g. *woman*), place (e.g. *street*), object (e.g. *jeans*) or attributes (e.g. *blue, long*). Relations between entities are even spatial positions (e.g. *in front of, behind*), actions (e.g. *walking*) or descriptions (e.g. *wearing*).

While recent approaches for this task may differ, the majority are using object detection and region captioning as baseline [13]. For object detection, the most reported solution is the use of pre-trained Faster-RCNN [14], a highly efficient Convolutional Neural Network (CNN) approach with Region Of Interest (ROI) pooling for object classification. Once objects are detected, the SGG solutions need to pair entities with one another and find the correct predicate to represent this relation. To do so, approaches such as Conditional Random Fields (CRF) [15] or Transitionnal Embeddings (TransE) [16] are used. Lately, Neural Networks have leveraged this task with new RNN/LSTM-based [17] and Graph Convolution Networks (GCN) [18] approaches.

When working with videos, a complex structure is needed to model relationships between images. To this end, [19] use a Temporal Convolution Network (TCN) paired to a GCN for modeling within-image dependencies. In [20] the authors use Target Adaptive Context Aggregation to relate entities to their spatio-temporal context.

2.2. Commonsense Completion

There are multiple ways to use external knowledge to enrich a scene graph. The task of Commonsense Completion is introduced in [21] to define the automatic completion of a knowledge graph using commonsense knowledge, in most cases retrieved from ConceptNet [8]. In this task the knowledge is directly added to the graph, creating new nodes and edges. In COMET [22], the authors describe a model that learns how to generate graph completion based on relationships between events from the ATOMIC [9] and ConceptNet [8] datasets. This method uses a Transformer architecture, the model is trained with a dataset of graphs to predict the next node, given the input previous node and a relation R from the set of relationships of ATOMIC.

As the solution uses a Transformer architecture, the input and output are natural language sentences, with specific tokens to represent relations. We can also see the task of Commonsense Completion of the SG as a Knowledge Graph Fusion. [23] propose a new approach to bridge knowledge and scene graphs using successive message passing on a Graph Neural Network (GNN).

2.3. Word Connotations Lexicon

[12] is the first attempt to build a connotation lexicon — a lexicon that maps words and their intrinsic connotation. The proposed approach learns word connotation using connotative predicates, i.e. predicates that ensure that words often encountered with ones negatively connoted will also be negatively connoted. With this method, the algorithm only needs a small seed of labelled words and a database of texts to learn words' connotations. In [24] the authors extend this method using induction algorithms based on graph structures. The use of Random Walk based on HITS/PageRank, Label/Graph Propagation and Constraint Optimization is reported. With this approach, [24] propose to capture fine-grained inductions, reducing biases from the previous solution (e.g. the world "cure" is often associated with "disease" while not being negatively connoted).

3. Detecting and fulfilling the implicit need of help

From an input video sequence, the system builds its own representation of the task being performed using state-of-the-art approaches of Vision-to-Language. Then, it reasons on this representation using commonsense knowledge to assess risks for a human: if the risks are too high, an assistive action is performed accordingly. This work aims to solve the following questions: (a) Scene Graph enrichment with related Commonsense Knowledge; (b) Scene Graph refinement upon visual features; (c) Sentiment Analysis from Scene Graph; (d) Action Generation from Commonsense Scene Graph.

3.1. Scene Understanding

The efficient scene understanding task from visual inputs, including human intents and objects affordances, is an old an unsolved challenge for Computer Vision. In this task, the use of an appropriate representation of the scene is of utmost importance. In the literature, there are two main approaches for scene representation: graphs and natural language processing; the former has been developed within Scene Graph Generation and the latter within Video Captioning. We choose scene graphs over natural language captions to model our representation of the scene. Scene Graphs provide numerous advantages: each detected entity is clearly represented and grounded, relationships between user and objects can be clearly identified and finally the enrichment of external knowledge is simple as most of knowledge bases also use graph structures [8] [9]. In this section, we will detail our system for understanding the scene and inferring human's risks using SG, breaking down the description in four distinct steps (illustrated in Figure 1):
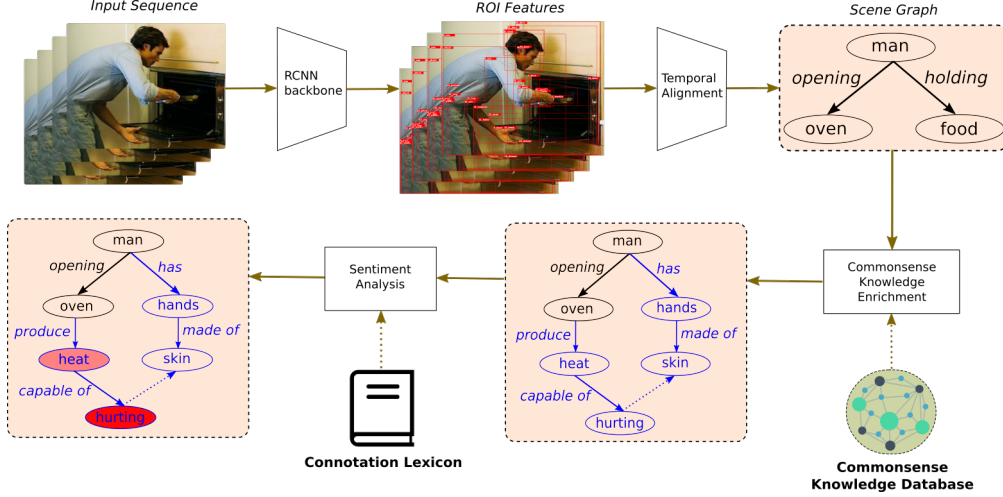
Figure 1: Our proposed architecture for scene representation and sentiment analysis.

Scene Graph Generation First, the representation of the relevant perceived data is critical. As a backbone, we use Faster-RCNN [14] (Figure 1 top left) to retrieve ROI features from the input scene. Given these features, we need to construct a directed graph G (Figure 1 top right) composed of a set of entities E and a set of relations R such that:

$$G = (R, E, \theta) \tag{5}$$

where θ is an incidence function that finds the relation $r \in R$ between the head entity $h \in E$ and the tail entity $t \in E$ such as:

$$\theta : \{h, t\} \in E^2 \rightarrow r \tag{6}$$

To generate such a graph, we follow the approach from [20] that uses Target Adaptive Context Aggregation (TRACE) to embed temporal and spatial information. The approach is as follows: from the visual features, relation candidates are represented as a hierarchical relation tree (HRTree); then, the TRACE module will capture temporal and spatial relationships to model the context with other frames; finally, a classification module will output the best inference.

Scene Graph Enrichment Second, we enrich the graph with relevant commonsense data (Figure 1 bottom right). The challenge here is to bridge the gap between the ontology represented in the scene graph and the ontology represented in the commonsense knowledge graph. ConceptNet is a database of commonsense knowledge, it has been generated using multiple resources such as crowd-sourcing or expert-generated data. For each word in natural English language, ConceptNet relates other words or group of words using commonsense relationships such as "isUsedFor" or "is PartOf". For each relation, ConceptNet gives a set of connected entities. To select the most relevant, we pick the one with the highest confidence match with the data already present in the graph. For example, the relation "knife is used for cutting vegetables" will be selected over "knife is used for stabbing" if instances of vegetables (e.g. "tomato") are already present in the graph. Thus, following [23], we pair similar labeled nodes from both

ontologies with a new edge. Then, all new edges are updated using successive message passing to propagate information across the graph. Grounding is applied by inferring related regions of the image to the corresponding enriched commonsense knowledge [25]. This, for instance, will replace the word "vegetables" in the previous example by the one directly connected with the image, i.e. "tomato". The enrichment of commonsense knowledge is also used to interpret mis-used objects in a task, e.g. from the graph "PersonX is cutting a tomato with an axe" the commonsense knowledge from the word "axe" could be infrequently related to the word "tomato" and thus the task would be declared as *unsafe*.

Sentiment Analysis on Graph Third, sentiment analysis is performed given information from the graph (Figure 1 bottom left). We are evaluating words and their semantic connotation (e.g. the word "heat" is negatively connoted) using connotation lexicon such as [26]. Traditional approaches to build connotation lexicon rely on words prosody in texts, we want to extend this representation to visual features proximity using bounding boxes coordinates associated to every entity. For instance, spatial proximity between "negative" entities and the user in the image features will be highly weighted. We update the graph adding a *sentiment value* $S_n \in [-1; 1]$ to each node that will represent the potential risk of the entity for the human.

Decision Making Fourth, given the sentiment analysis, a pooling is performed with respect to the graph dependencies to retrieve a confidence value. If this value is above a pre-defined threshold, the task is declared as *unsafe* and a decision of assistance is made. This threshold is dynamic and could be adjusted given contextual information such as the presence of a child in the scene.

3.2. Command Generation

Once the autonomous agent understands the immediate necessity of assistance, it needs to provide the appropriate help. To do so, the system needs to generate the best assistive action towards increasing the safety of the task, as in the example introduced in Section 1.

We build what we call a "commonsense response", that means the most probable set of actions to perform for helping the human with the task. The goal here is to complete the weighted graph retrieved from the Sentiment Analysis with the related commonsense knowledge to stabilize the graph. We iterate through the graph using the same process as for Scene Graph Enrichment. The difference here is that for each iteration we also perform Sentiment Analysis and select only the positively weighted nodes. We compute the current sentiment value of the graph G as follows:

$$S(G) = \sum_{i=1}^{n} S_n \tag{7}$$

where n is number of nodes. At the end of the process we obtain a graph similar to the one shown in Figure 2, where the solution will be the highest positively weighted node that represents an object. This object could then be found and provide by the robot to the user. If no satisfying solution is found, one approach could be to warn the user by vocal utterances.
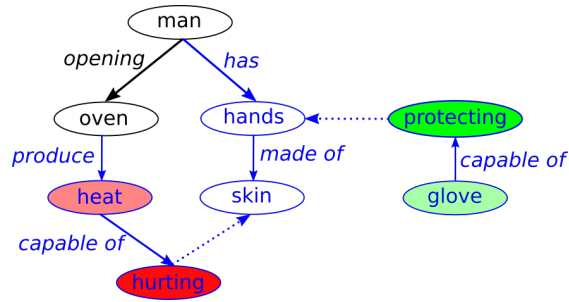
Figure 2: Solution retrieval via Scene Graph Completion.

3.3. Evaluation

We will evaluate our Scene Graph Generation approach using the ActivityNet dataset [27]. This dataset contains 20k videos of 200 everyday life human activities and is used as a benchmark for most SGG approaches. Recently, the ActionGenome dataset was also introduced in [28]. This dataset captures human daily life activities in 265k labeled frames.

To assess user acceptance and perceived performance of the system, we will conduct a series of human experiments. To do so, humans will engage in safe and potentially dangerous scenarios in which the system will detect risk of danger and define its remediating action to minimise detected danger. For practical and ethical reasons, humans will perform the safe task in a real environment and the potentially dangerous task in a Virtual Reality environment in which the task is simulated. We will use the Unified Theory of Acceptance and Use of Technology (UTAUT) [29] and the Technology Acceptance Model 3 (TAM-3) [30] to evaluate acceptance of the system in these scenarios. Additionally, we will supplement these measures with qualitative feedback about the performance and actions of the system. Human-Human interaction in similar scenarios will be used to evaluate the coherence of the assistive action of the system.

4. Conclusion and future work

The proposed work combines a traditional machine learning approach to generate an accurate model of the world with knowledge representation and reasoning. Our solution includes Scene Graph Generation as a high-level representation of the scene, commonsense knowledge enrichment combined with sentiment analysis to asses risk for the user and, finally, a graph completion method to retrieve relevant solutions. As a limitation, our work does not consider other factors of the implicit need of help such as fatigue or stress. Furthermore, this proposal does not take into account the latent context of the captured sequence. For instance, in our running example it would be good to know if the oven is turned on or if it was turned on earlier.

As this proposal is still an on-going work, a number of challenges remain open, such as: the efficient fusion of scene and commonsense knowledge graphs; the sentiment analysis from scene graph and the generation of robot commands from graph entities. All these issues will be considered in future work, along with an investigation of the limitations, and confounding factors, in the automatic interpretation of the implicit need of help.

5. Acknowledgments

References

[1] S. Harnad, The symbol grounding problem, Physica D: Nonlinear Phenomena 42 (1990) 335–346.

[2] Y. Tada, Y. Hagiwara, H. Tanaka, T. Taniguchi, Robust understanding of robot-directed speech commands using sequence to sequence with noise injection, Frontiers in Robotics and AI 6 (2020) 144.

[3] S. Waldherr, R. Romero, S. Thrun, A gesture based interface for human-robot interaction, Autonomous Robots 9 (2000) 151–173.

[4] G. Doisy, Sensorless collision detection and control by physical interaction for wheeled mobile robots, in: Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction, 2012, pp. 121–122.

[5] R. Liu, X. Zhang, S. Li, Use context to understand user's implicit intentions in activities of daily living, in: 2014 IEEE International Conference on Mechatronics and Automation, 2014, p. 1214–1219.

[6] S. Li, X. Zhang, Implicit intention communication in human–robot interaction through visual behavior studies, IEEE Transactions on Human-Machine Systems 47 (2017) 437–448.

[7] B. Kuipers, Commonsense reasoning about causality: deriving behavior from structure, Artificial intelligence 24 (1984) 169–203.

[8] R. Speer, J. Chin, C. Havasi, Conceptnet 5.5: An open multilingual graph of general knowledge, Proceedings of the AAAI Conference on Artificial Intelligence 31 (2017).

[9] M. Sap, R. L. Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, Y. Choi, Atomic: An atlas of machine commonsense for if-then reasoning, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) 3027–3035.

[10] F. van Harmelen, A. t. Teije, A boxology of design patterns for hybrid learning and reasoning systems, Journal of Web Engineering 18 (2019) 97–124. ArXiv: 1905.12389.

[11] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. A. Shamma, M. S. Bernstein, L. Fei-Fei, Image retrieval using scene graphs, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2015, p. 3668–3678.

[12] S. Feng, R. Bose, Y. Choi, Learning general connotation of words using graph-based algorithms, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011, p. 1092–1103.

[13] Y. Li, W. Ouyang, B. Zhou, K. Wang, X. Wang, Scene graph generation from objects, phrases and region captions, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, p. 1270–1279.

[14] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2017) 1137–1149.

[15] B. Dai, Y. Zhang, D. Lin, Detecting visual relationships with deep relational networks, in:

2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, p. 3298–3308.

[16] H. Zhang, Z. Kyaw, S.-F. Chang, T.-S. Chua, Visual translation embedding network for visual relation detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, p. 3107–3115.

[17] R. Zellers, M. Yatskar, S. Thomson, Y. Choi, Neural motifs: Scene graph parsing with global context, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5831–5840.

[18] J. Yang, J. Lu, S. Lee, D. Batra, D. Parikh, Graph r-cnn for scene graph generation, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 670–685.

[19] R. Wang, Z. Wei, P. Li, Q. Zhang, X. Huang, Storytelling from an image stream using scene graphs, Proceedings of the AAAI Conference on Artificial Intelligence 34 (2020) 9185–9192.

[20] Y. Teng, L. Wang, Z. Li, G. Wu, Target adaptive context aggregation for video scene graph generation, Proceedings of the IEEE/CVF International Conference on Computer Vision (2021) 13688–13697.

[21] X. Li, A. Taheri, L. Tu, K. Gimpel, Commonsense knowledge base completion, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2016, p. 1445–1455.

[22] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, Y. Choi, Comet: Commonsense transformers for automatic knowledge graph construction, arXiv:1906.05317 [cs] (2019).

[23] A. Zareian, S. Karaman, S.-F. Chang, Bridging Knowledge Graphs to Generate Scene Graphs, volume 12368 of *Lecture Notes in Computer Science*, Springer International Publishing, 2020, p. 606–623.

[24] S. Feng, J. S. Kang, P. Kuznetsova, Y. Choi, Connotation lexicon: A dash of sentiment beneath the surface meaning, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2013, p. 1774–1784.

[25] A. Zareian, Z. Wang, H. You, S.-F. Chang, Learning visual commonsense for robust scene graph generation, arXiv:2006.09623 [cs] (2020).

[26] S. M. Mohammad, P. D. Turney, Crowdsourcing a word-emotion association lexicon, Computational Intelligence 29 (2013) 436–465.

[27] B. G. Fabian Caba Heilbron, Victor Escorcia, J. C. Niebles, Activitynet: A large-scale video benchmark for human activity understa nding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 961–970.

[28] J. Ji, R. Krishna, L. Fei-Fei, J. C. Niebles, Action genome: Actions as compositions of spatio-temporal scene graphs, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10236–10247.

[29] V. Venkatesh, F. Davis, M. G. Morris, Dead or alive? the development, trajectory and future of technology adoption research., Journal of the association for information systems 8 (2007) 1.

[30] V. Venkatesh, H. Bala, Technology acceptance model 3 and a research agenda on interventions, Decision sciences 39 (2008) 273–315.