Reasoning and Learning in Real-World Systems for Long-Term Autonomy (LTA)
Papers from the AAAI 2018 Fall Symposium

Deep CNN and Probabilistic DL Reasoning for Contextual Affordances

Hazem Abdelkawy*, Sandro Rama Fiorini*, Abdelghani Chibani, Naouel Ayari and Yacine Amirat
LISSI Laboratory
University of Paris-Est Créteil (UPEC),
Vitry-sur-Seine France
{hazem-khaled-mohamed.abdelkawy, sandro.fiorini, abdelghani.chibani, amirat}@u-pec.fr

Abstract

Endowing robots with cognitive capabilities for recognising contextual object affordances is a big challenge, which requires sophisticated and novel approaches. In this paper, we propose a hybrid approach to interpret contextualised object affordances from sensor data. The proposed approach combines both Deep CNN networks for object and indoor place recognition with probabilistic DL reasoning for affordance inference. We argue that our hybrid approach can be an interesting alternative in situations where no specific dataset for contextualised affordances exists.

Introduction

Visual intelligence is one of the most important aspects of human cognition, and the paramount goal of the visual intelligence is the contextual visual reasoning. Take a cup as an example. From a single image, humans can infer its name, texture, colour, and what actions the object affords. In (Gibson 2014), Gibson defined the notion of object affordance as the "properties of an object that determine what actions a human can perform on them." In this paper, we define the notion of *contextual object affordance* as the *relationship between an object and a set of actions this object allows in a given situation*. In other words, objects might afford different actions at different places, times, or situations. In this work, contextual object affordances are proposed as means to filter the possible actions that a companion robot can monitor/do in an ambient environment. Besides, contextual affordances can be used as part of a bigger process to extract an agent intentions, by restricting the possible intentions based on the affordable actions in the environment in a given time.

The previous attempts to recognise object affordances can be divided into two categories: visual features classifications models, knowledge-based inference models. In (Fergus et al. 2005), the proposed approach is able to learn an object category from its name, based on the output of Google Image search results. In (Kjellström, Romero, and Kragić 2011), the inference of the object affordances is based on monitoring humans while they use objects in different actions. In (Yao, Ma, and Fei-Fei 2013), the proposed approach is able to model the affordance of an object based on the majority of human poses while interacting with that object. In (Chu and Thomaz 2017), object affordances are discovered by a guided exploration approach that combines self-learning with supervised learning. In (Do et al. 2017), AffordanceNet deep learning model is proposed to detect ambient objects and their affordances simultaneously from RGB images. In (Zhu, Fathi, and Fei-Fei 2014), the authors propose a Markov Logic Network (MLN) knowledge base to apply zero-shot object affordance prediction besides object recognition given human poses. Despite the previous serious attempts, only the latter model is able to predict the object affordances for unseen novel objects.

From a pure machine learning perspective, while it would be possible to train a model to produce contextual affordances, frequently data sets are not available for this task. To the best of our knowledge, no data set with these characteristics exists. Therefore alternative solutions are needed in order to use this kind of information in autonomous systems.

In this work, we propose our initial findings to extend a previously proposed cognitive architecture (Ayari et al. 2015; 2017) to predict the contextual object affordances based on place information. The extension is based on Deep Convolutional Networks (CNNs) and Probabilistic Description Logics (DL) Reasoning. The role of the probabilistic DL reasoning is to provide the ability to produce contextual affordances based on low-level object and place information. Our contribution is methodological: we demonstrate how the integration of Deep CNNs models and DL reasoning components can produce more valuable output even in a situation where the data of training is missing.

Cognitive Architecture

The overall cognitive architecture proposed in (Ayari et al. 2015; 2017) is depicted in Fig 1. At the low level, a *communication service* is implemented to enable the entities populating the ambient environment to connect and subscribe to cloud services as well as to interchange knowledge. The communication service is based on standard communication technologies such as (XMPP, REST, etc.) In addition to the communication service, emotion recognition, metric maps and topological maps based environment modelling, and multi-modal data sensing services are implemented at the low level.

*These authors contributed equally to this work

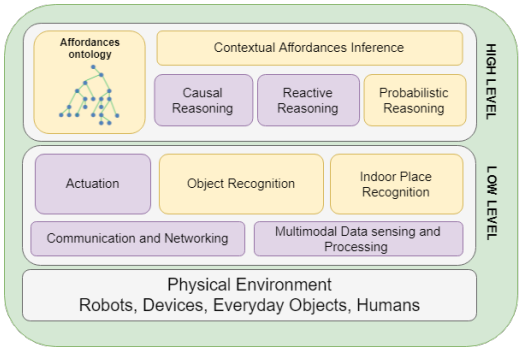

Figure 1: Overview of the architecture.

Here we focus on the *object recognition* and *indoor place recognition* services, as well as the *probabilistic reasoning service*. Their objective is to enable the companion robots to recognise the indoor places and the objects populating the environment. Implemented as Deep CNNs, models, they extract and recognise ambient objects and the indoor places in which these objects are located. This information is fed into the High Level, where the DL reasoning is able to infer the contextual object affordances based on probabilistic object and place data provided by the Low Level, using mainly instance classification and subsumption checking.

Object Recognition

We employ $YOLO$ ("$You\ Only\ Look\ Once$") algorithm (Redmon et al. 2016) in order to recognise the objects populating an environment. $YOLO$ object recognition model predicts the object bounding boxes and the associated probabilities for these boxes. Firstly, the input image is divided into SxS regions, within each region the model outputs a set of N bounding boxes. For each bounding box, the model predicts the object class probability and the location values for the bounding box. Finally, the model filters out bounding boxes with class probabilities below a predefined threshold value.

$YOLO$ model has 24 stacked convolution layers followed by 2 dense fully connected layers. To reduce the output features space, we applied 1X1 convolution reduction layers followed by 3X3 convolution layers.

Indoor Place Recognition

The Deep Residual Network $(ResNet)$ was developed by Microsoft Research Labs and exploited in (He et al. 2016) for image recognition. In this paper, we use a modified ResNet deep architecture to recognise the indoor locations. The model consists of number of stacked convolution layers, combined with residual shortcut connections to train deeper and more sparse networks. The input convolution layer consists of 64 feature maps of size 7x7 with stride of 2. A max-pooling layer of 3x3 kernel and stride of 2 is applied after the input layer to down-sample the feature maps representation. The max-pooling layer is followed by a set of 16 residual blocks, each residual block consisting of 4 convolution layers with 3x3 kernel size. A Global Average Pooling (GAP) layer (Lin, Chen, and Yan 2013) is used to minimise overfitting by reducing the total number of learned parameters. Finally, a dense, fully connected neural network with 2000 neuron is exploited as a classification layer to recognise the indoor location.

Probabilistic Reasoning

The objective of the probabilistic reasoning service is to infer object affordances based on object and place data extracted by the CNN services described above. It is based on a probabilistic DL reasoning model presented by (Riguzzi et al. 2015), supported by an OWL 2 ontology.

The ontology is relatively simple (Fig. 3). It specifies the notion of affordance as a relationship between an object instance and a type of (or class of) action. It defines three main high-level concepts: *object*, *place* and *action type*. Objects are the common objects of daily living and places are the physical places wherein these objects can be found. Objects and places are linked by the relation *placed at*. Action types are reified action concepts; its instances are types of actions which objects may afford. The reification allows one to represent affordances as a first-order relations, without requiring metamodeling subterfuges. The object relationship *affords* captures this relation. Also, representing affordances as relationships instance of class instances simplifies the reasoning, as it avoids the need of creating artificial class instances during DL reasoning, which is not trivial to control. The remainder of the ontology is a taxonomy of objects and places which parametrizes the reasoning algorithm, as well as a set of reified action types represented as instances.

The inference rules are captured by DL subsumption rules with the general formula schema

$$O \sqcap \exists \text{placedAt}.P \sqsubseteq \exists \text{affords}.\{T\},$$

where O is a subclass of the *Object*, P is a subclass of *Place* and T is an instance of *Action Type*. So, if any object instance of a certain class is placed at the right place, it is possible to reason that it affords a given action. Objects and places can be described as specific as necessary.

The probabilistic DL reasoning is based on DISPONTE distribution semantics (Riguzzi et al. 2015) for DL knowledge bases (KBs). In this model, DL axioms are annotated with probabilities, which are assumed to be independent. DISPONTE defines *worlds* that select subsets of axioms of the KB. The probability of a world w is defined as a joint probability over selected and non-selected axioms. Finally, the probability of a query axiom (i.e. a inferred axiom) is ultimately given by a marginalised joint probability over the worlds that entail the query. We use the BUNDLE reasoner (Riguzzi et al. 2015) carry out the inferences. BUNDLE implements DISPONTE by joining traditional DL reasoners (i.e. Pellet[1]) with Binary Decision Diagrams (BDD). In brief, given a probabilistic KB and a query axiom, BUNDLE takes worlds to be possible explanations of a query generated by a standard DL reasoner. The probability of a query then defined by marginalising over the joint probabilities of each of its explanations. This computation is optimised by

[1]https://github.com/stardog-union/pellet

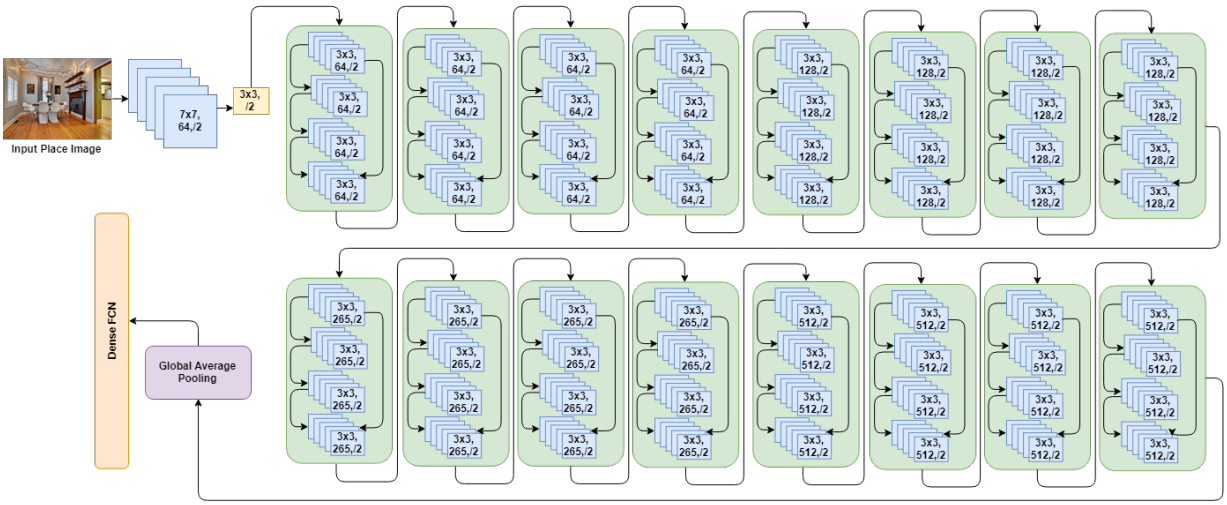

Figure 2: Deep ResNet model for indoor place recognition.

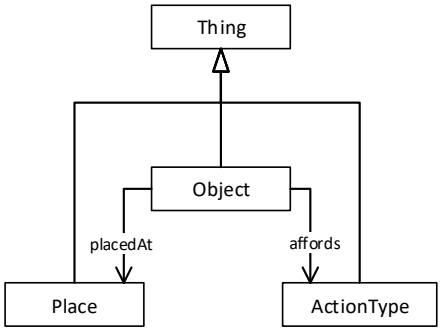

Figure 3: Overview of the ontology being used.

calculating BDDs of a disjunction of the explanations, and subsequently recursively calculating the probability of the query using the BDD.

Our affordances interpretation algorithm takes probability vectors of object and places classes detected by the CNNs as inputs. For a given input frame, the dCNN layer outputs a vector $O = [p_1 : C_1, ..., p_n : C_n]$, where C_i corresponds to the calculated class of the i-th detected object with probability p_i. For places, the dCNN layer outputs a vector $L = [p_1 : C_1, ..., p_n : C_m]$, where C_i corresponds to a possible classification for the *current* place with probability p_i. The affordances reasoner encodes this output as a set of instances of the ontology. An instance l is created in the KB and for each element $p_i : C_i$ of L, an axiom $p_i :: l : C_i$ is added to the ontology. Also, for each element $p_i : C_i$ of O, an instance x_i is created and an axiom $p_i :: x_i : C_i$ is added to the KB. All object instances x_i are defined to be *placed at* l. This KB is loaded into BUNDLE, which infers affordance relationships between instances x_i and actions types present in the model, weighted by probabilities calculated on x_i class and placement. The output of the component is the set of action types afforded by all objects in that frame, weighed by their infered probability. In cases where two objects in the same frame afford the same action, then the action type with maximum probability is taken. The object and place instances are not kept from a frame to the next.

An advantage of this method is that all the ontology axioms are taken into consideration while reasoning, even non probabilistic ones. For example, it is possible to aggregate common object classes in the object dataset under superclasses, to which one can define a single reasoning rule. Such modelling can drastically reduce the amount of inference rules to cover all possible affordances.

Preliminary Evaluation

We carried out a preliminary evaluation of the proposed method through an empirical experiment on real-world datasets. The datasets are as follows:

- Microsoft COCO (Lin et al. 2014) (Common Objects in Context) dataset was exploited to evaluate the performance of $YOLO$ object recognition deep learning model. The dataset consists of 80 different objects with total number of 2.5 million annotated instances. The dataset is divided into 118K images for training, 5K images for validation, 41K images for testing.
- The Place365 standard dataset (Zhou et al. 2017) was used to evaluate the Place recognition deep learning model. The dataset consists of 2 million images of different 365 common places. The dataset is divided into overlapped training, validation, and testing sets (1M, 36K, 300K images, respectively). The training set contains up to 5,000 images per category, while the validation and testing contains 100 and 900 images per category, respectively.
- The Daily Living Activities (ADL) dataset (Pirsiavash and Ramanan 2012) was exploited to evaluate the performance of the DL reasoning. The dataset consists of one million RGB frames of 20 persons while practising un-

scripted 32 daily activities. The dataset annotation consists of objects, activities, hand positions, and environmental events. Compared to the traditional datasets for daily activities, this dataset combining long scale temporal activities for periods up to few minutes and complex object interactions.

- The proposed ontology was populated with objects, places and related action classes from COCO, Place365, and ADL datasets respectively. We defined an initial collection of 16 DL rules to cover a subset of objects, places and action types. These rules have been defined by hand, trying to match activities in ADL to possible combinations of objects and places from COCO and Place365. By aggregating dataset objects and places into superclasses, we were able to define rules with higher reuse potential.

To evaluated the proposed approach, a set of 4577 continuous frames from the ADL dataset were used to obtain some preliminary performance statistics. In the Low Level, the Deep CNNs were able to recognise in average up to three objects ($\bar{x} = 3.43, s = 1.91$) and five indoor places (fixed value). The average processing time of recognising the ambient objects with indoor Places is 260 ms per frame of 288x384 pixels[2]. In the High Level, the reasoner component is able to generate approximately 11 axioms ($\bar{x} = 11.87, s = 3.82$), which were added to the ontology at each frame. Based on this input, the reasoner produced around 3 contextual affordances per frame in average ($\bar{x} = 3.32, s = 2.2$). The average reasoning time for recognition of contextual affordances is approximately 350 ms ($\bar{x} = 354.60, s = 216.29$) for each frame[3].

Conclusion

In this paper, we propose a hybrid approach based on Deep CNNs and DL reasoning to recognise contextual affordances. We evaluated the proposed approach through empirical experiments on real-world datasets. The preliminary evaluation shows that the processing time of the proposed approach is reasonably fitting the constraints of real-time applications.

References

Ayari, N.; Chibani, A.; Amirat, Y.; and Matson, E. T. 2015. A novel approach based on commonsense knowledge representation and reasoning in open world for intelligent ambient assisted living services. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, 6007–6013.

Ayari, N.; Abdelkawy, H.; Chibani, A.; and Amirat, Y. 2017. Towards semantic multimodal emotion recognition for enhancing assistive services in ubiquitous robotics. In *2017 AAAI Fall Symposium Series*.

Chu, V., and Thomaz, A. L. 2017. Analyzing differences between teachers when learning object affordances via guided exploration. *The International Journal of Robotics Research* 36(5-7):739–758.

Do, T.; Nguyen, A.; Reid, I. D.; Caldwell, D. G.; and Tsagarakis, N. G. 2017. Affordancenet: An end-to-end deep learning approach for object affordance detection. *CoRR* abs/1709.07326.

Fergus, R.; Fei-Fei, L.; Perona, P.; and Zisserman, A. 2005. Learning object categories from google's image search. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, 1816–1823. IEEE.

Gibson, J. J. 2014. *The ecological approach to visual perception: classic edition*. Psychology Press.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Kjellström, H.; Romero, J.; and Kragić, D. 2011. Visual object-action recognition: Inferring object affordances from human demonstration. *Computer Vision and Image Understanding* 115(1):81–90.

Lin, T.; Maire, M.; Belongie, S. J.; Bourdev, L. D.; Girshick, R. B.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: common objects in context. *CoRR* abs/1405.0312.

Lin, M.; Chen, Q.; and Yan, S. 2013. Network in network. *arXiv preprint arXiv:1312.4400*.

Pirsiavash, H., and Ramanan, D. 2012. Detecting activities of daily living in first-person camera views. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2847–2854. IEEE.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Riguzzi, F.; Bellodi, E.; Lamma, E.; and Zese, R. 2015. Reasoning with probabilistic ontologies. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, 4310–4316. AAAI Press.

Yao, B.; Ma, J.; and Fei-Fei, L. 2013. Discovering object functionality. In *Proceedings of the IEEE International Conference on Computer Vision*, 2512–2519.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhu, Y.; Fathi, A.; and Fei-Fei, L. 2014. Reasoning about object affordances in a knowledge base representation. In *European conference on computer vision*, 408–424. Springer.

[2]Deep CNNs running on a Intel(R) Core(TM) i7-6820HQ CPU @ 2.70GHz (8 CPUs), 2.7GHz, with 16Gb RAM.

[3]Reasoner running on a Intel(R) Xeon(R) CPU E5-1630 v4 @ 3.70GHz, 4 Core(s), 8 Logical Processor(s), with 8Gb RAM.