Appreciation of Symbolic Attributes in Machine Perception

Mohamadreza Faridghasemnia

Center for Applied Autonomous Sensor System (AASS)

Abstract

In this position paper, we want to attract attention to the importance of symbolic attributes in machine perception. We discuss the benefits of a perception system that not only recognizes the category of objects, but also recognizes many other aspects of objects.

Keywords

Machine perception, Symbolic attribute detection, Context-awareness, Object understanding

1. Introduction

Context-awareness is becoming an important feature of systems that are meant to mimic human intelligence. Specifically, in systems that are designed for Human-Robot Interaction (HRI), context-awareness allows a robot, for example, to perform its action in different situations, or to resolve uncertainty in under-specified tasks. In this paper, we focus on one principal component of context-awareness, which is machine perception. Let's imagine at a time in a day, a person telling her/his assistive robot "Bring me the red one". Although it is obvious for the person which object she/he is referring to, the robot finds more than one candidate that suits the tag "red", depending on the visually perceived context. These candidates can be a red apple, a bottle of red wine, a magazine that is named red, and a plant that the user named "red". Based on contextual information (visual, auditory, time, location, etc.), and past interactions with the user, a robot can decide which "red" fits the user's intention. Notice that we define the visual part of the context as all objects that are perceivable in the environment by the robot.

Let us imagine a smarter system than the previous example. Someone is walking with its assistive robot on a street, the robot analyses its perception and breaks the silence with "Bare trees, white grounds, it sounds like winter", and the user says, "This year winter came on time". Systems in these examples are only some of the simple showcases of what people expect from artificial intelligence before they call them intelligent. While there is much research going on around different aspects of AI and robotics systems to make these examples real, we want to focus on the perception parts of the system, where these days a simple aspect of perception is often underestimated, that is perceiving and understanding the visual attributes. Indeed, the basic information that is needed in these examples is that robots should know the symbolic attributes of objects, e.g. the redness of wine, the red title of a magazine, the bareness of trees,

AIC 2022, 8th International Workshop on Artificial Intelligence and Cognition
mohamadreza.farid@oru.se (M. Faridghasemnia)

and the whiteness of the ground. In other words, robots not only have to recognize apple, wine, magazine, plant, tree, and ground but also have to recognize attributes of each, such as: type, color, labels, states, and other visual attributes.

A while ago, traditional Computer Vision (CV) algorithms were used for recognizing objects in an image, which were limited to recognizing only a very few objects. Other CV algorithms were capable of recognizing very few other features of objects, like color and shape. Nowadays robot perception has extensive improvement, thanks to neural networks which with an acceptable amount of computation, can locate and recognize the category of objects in an image (object detection with neural networks). But research on recognizing other features (attributes) of the objects has been limited. We only found very few works that are addressing this problem with certain limits, which are reviewed in Section 3. We believe the advancements of neural networks (for example YOLO [1] or R-CNN [2] architectures) should be used in recognition of other attributes of objects, in addition to recognition of categories.

2. In defense of symbolic attributes

In this section, we discuss why symbolic attributes are important in HRI. We believe its benefits can be extended to other applications of AI systems that have a visual perception system. We also discuss challenges for realization of a system that can compute symbolic attributes and the current state.

Benefits

By only considering the category of objects, we are missing lots of information that a system can compute from its visual perception module. In HRI, often it is required that a word should be directly grounded to a physical object. In our example of "Bring me the red one", the robot has to find an instance of "red" for the user. In a system with sophisticated intelligence, the robot should find instances of a word ("red" in our example) in whole attributes of all objects, and not limited to only the category of object. For example, when a robot wants to find an object that is "red", it can search in all objects that have any attribute of red, where attributes can be name, labels, colors, etc., and choose the best candidate among objects.

In human conversations, it often happens that a person calls something by an attribute that cannot be grounded directly. For example, "red" can be the type of wine but this may not directly correspond to visually observable features. A robot may find a wine with the label "Merlot" and by querying its ontology find a relation between "Merlot" and "red". The basic requirement for such an approach is that a system should save symbolic attributes of all objects, together with an ontology that it can extract and match information. In other words, since available ontologies and reasoning engines are at symbolic level, having symbolic attributes of objects make it possible for a system to ground expressions indirectly.

As discussed, one major benefit of detecting symbolic attributes is in grounding for HRI, and one might find it similar to end-to-end models that nowadays are being used for grounding referential expressions, but indeed they differ in many aspects that we want to describe here. Many works like the one given in [3] trained an end-to-end neural network which is learned to find a region in an image that fits a given referential expression. Although end-to-end

approaches of this kind are very compelling in grounding one expression to an object in an image, we believe one might find symbolic attribute-based approach better in some criteria, such as:

- `Firstly` : The end-to-end approaches often seem difficult to be explainable. For example, it is not easy to find what words in expressions are grounded.
- `Secondly` : Using symbolic knowledge (either knowledge base or history of observed objects) in neural networks is not as easy as symbolic methods. For example, it is easy to incorporate an existing symbolic ontology and detections, if detections are also symbolic.
- `Thirdly` : Using inductive and deductive engines on the knowledge (knowledge base + history + currently observed objects) is easier on symbolic data, as many state of the art tools for induction and deduction are symbolic.
- `Fourthly` : As a result of of Secondly and Thirdly, understanding an indirect (implicit) reference is possible with symbolic attributes. This is because understanding indirect references requires the deployment of inductive and deductive engines on the combination of ontology and perceivable information, as discussed in Section 2.

3. Challenges

One who wants to make a system capable of recognizing attribute of objects in a scene should expect two main challenges of network architecture and the dataset. Nowadays, neural networks are capable of recognizing categories of objects in a scene. For example, two-stage object detectors (e.g. R-CNN model [2]) take an image of a scene and predict regions of interest in the image at the first stage, and then in the second stage predict the most probable category for each region. In machine learning, while the first stage is a regression problem, the second stage is a multi-class, single-label classification problem. A proper setting for attribute detection seems to be similar to the object detection model but in a multi-class multi-label setting.

But recognition of attributes is not only a change of neural network settings, nor should be considered as an easy task. One problem is that recognition of regions for attributes is not easy, since all regions in an image can have some attributes (think of a sky in an image that is not labeled as an object, but it has the color attribute), but one might simplify the problem from attributes of regions to attributes of objects, which most datasets have labels per object.

An adequate dataset is another requirement for a neural attribute detection system. Adequacy in terms of clean, balanced, categorized, and a large number of labels, etc. should be defined. Current attribute datasets are very noisy, most of them are labeled per-image, and the number of labeled attributes is usually small. For example, VisualGenome [4] come with a huge number of images, with 64000 per-region attributes. But attributes in this dataset are very noisy and extremely imbalanced (many attributes appeared very few), hence no work applied attribute detection on this dataset. On the other hand, imageNet [5] attributes is better in term of balanced labels, and is also a per-region labeled dataset but with only 20 visual attributes. In Pascal attribute dataset [6], labels are balanced, cleaned, but per-image with only 64 attributes. It is worth noting that balanced labels seem almost impossible. Just like other object recognition datasets where humans are labeled more than any other objects, our investigation shows that color attributes are the most appeared attributes.

Current state

We believe recognition of all attributes of objects should get more attention. We found some works that addressed the importance of attributes, and some focused on the classification of attributes. However, we could not find any work on attribute detection. Notice that by classification we mean the task of classifying the most probable label for an image, while by detection we mean the task of locating the region of objects in an image, in addition to classifying the most probable label for each of the recognized regions. For example, the work that is given in [7] is an attribute classification work, where they used a neural network that classifies cropped regions for annotated attributes. At this work, the size of attributes is very limited and does not cover many important aspects of objects, and using the cropped image is too much simplification of the problem. Farhadi et al. in [6] described objects that have to be described with attributes, but images in this dataset have only one object. Such works are similar to image classification, but instead of classifying the category of an image, multiple attributes have to be classified for each image.

In our previous work in [8], we created a framework for capturing multi-modal attributes from language and vision, but our work was limited to few attributes, and attribute acquisition for objects via language was expensive. We also demonstrated the benefits of attributes in [9] and showed the inductions that can be done on symbolic attributes of objects. Other works also demonstrated the benefits of attributes (even embedded attributes) in few-shot learning [10], classification of fashion attributes [11], benefits of passing additional attributes (embedded) for caption generation [12].

4. Conclusion

In this short paper, we discussed the benefits of symbolic attributes for AI systems. As an example, we described that the extraction of symbolic attributes of objects in a scene allows a system to combine this information with available ontologies. We made cases for a perception system that not only recognizes the category of objects but also recognizes other attributes of objects. We also discussed challenges that exist and the current state.

Acknowledgments

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The author would also like to thank Alessandro Saffiotti and Lars Karlsson for their guidance and comments.

References

[1] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[2] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
[3] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, T. L. Berg, Mattnet: Modular attention network for referring expression comprehension, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1307–1315.
[4] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, arXiv preprint arXiv:1602.07332 (2016).
[5] O. Russakovsky, L. Fei-Fei, Attribute learning in large-scale datasets, in: European Conference on Computer Vision, Springer, 2010, pp. 1–14.
[6] A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, in: 2009 IEEE conference on computer vision and pattern recognition, IEEE, 2009, pp. 1778–1785.
[7] S. Banik, M. Lauri, S. Frintrop, Multi-label object attribute classification using a convolutional neural network, arXiv preprint arXiv:1811.04309 (2018).
[8] M. Faridghasemnia, A. Vanzo, D. Nardi, Capturing frame-like object descriptors in human augmented mapping, in: International Conference of the Italian Association for Artificial Intelligence, Springer, 2019, pp. 392–404.
[9] M. Faridghasemnia, D. Nardi, A. Saffiotti, Towards abstract relational learning in human robot interaction, arXiv preprint arXiv:2011.10364 (2020).
[10] B. Zhang, X. Li, Y. Ye, Z. Huang, L. Zhang, Prototype completion with primitive knowledge for few-shot learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3754–3762.
[11] M. Jia, Y. Zhou, M. Shi, B. Hariharan, A deep-learning-based fashion attributes detection model, arXiv preprint arXiv:1810.10148 (2018).
[12] Y. Huang, J. Chen, W. Ouyang, W. Wan, Y. Xue, Image captioning with end-to-end attribute detection and subsequent attributes prediction, IEEE Transactions on Image Processing 29 (2020) 4013–4026.