Training robots to identify object placements by 'hallucinating' scenes

12 February 2020, by Ingrid Fadelli

Oier Mees demonstrating how the new approach works. Credit: Mees et al.

With more robots now making their way into a number of settings, researchers are trying to make their interactions with humans as smooth and natural as possible. Training robots to respond immediately to spoken instructions, such as "pick up the glass, move it to the right," etc., would be ideal in many situations, as it would ultimately enable more direct and intuitive human-robot interactions. However, this is not always easy, as it requires the robot to understand a user's instructions, but also to know how to move objects in accordance with specific spatial relations.

Researchers at the University of Freiburg in Germany have recently devised a new approach for teaching robots how to move objects around as instructed by human users, which works by classifying "hallucinated" scene representations. Their paper, pre-published on arXiv, will be presented at the IEEE International Conference on Robotics and Automation (ICRA) in Paris, this June.

"In our work, we concentrate on relational object placing instructions, such as 'place the mug on the right of the box' or 'put the yellow toy on top of the box,'" Oier Mees, one of the researchers who carried out the study, told TechXplore. "To do so, the robot needs to reason about where to place the mug relative to the box or any other reference object in order to reproduce the spatial relation described by a user."

Training robots to make sense of spatial relations and move objects accordingly can be very difficult, as a user's instructions do not typically delineate a specific location within a larger scene observed by the robot. In other words, if a human user says "place the mug on the left of the watch," how far left from the watch should the robot place the mug and where is the exact boundary between different directions (e.g., right, left, in front of, behind, etc.)?

"Due to this inherent ambiguity, there is also no ground-truth or 'correct' data that can be used to learn to model spatial relations," Mees said. "We address the problem of the unavailability of ground-truth pixelwise annotations of spatial relations from the perspective of auxiliary learning."

The main idea behind the approach devised by Mees and his colleagues is that when given two objects and an image representing the context in which they are found, it is easier to determine the spatial relationship between them. This allows the robots to detect whether one object is to the left of the other, on top of it, in front of it, etc.

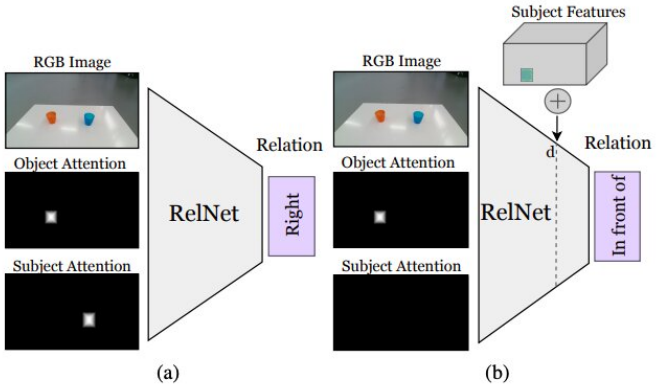

Figure summarizing how the approach devised by the researchers works. An auxiliary CNN, called RelNet, is trained to predict spatial relations given the input image and two attention masks referring to two objects forming a relation. (a) after training, the network can be 'tricked' to classify hallucinated scenes by (b) implementing high-level features of items at different spatial locations. Credit: Mees et al.

While identifying a spatial relationship between two objects does not specify where the objects should be place to reproduce that relation, inserting other objects within the scene could allow the robot to infer a distribution over several spatial relations. Adding these nonexistent (i.e., hallucinated) objects to what the robot is seeing should allow it to evaluate how the scene would look if it performed a given action (i.e., placing one of the objects in a specific location on the table or surface in front of it).

"Most commonly, 'pasting' objects realistically in an image requires either access to 3-D models and silhouettes or carefully designing the optimization procedure of generative adversarial networks (GANs)," Mees said. "Moreover, naively "pasting" object masks in images creates subtle pixel artifacts that lead to noticeably different features and to the training erroneously focusing on these discrepancies. We take a different approach and implant high-level features of objects into feature maps of the scene generated by a convolutional neural network to hallucinate scene representations, which are then classified as an auxiliary task to get the learning signal."

Before training a convolutional neural network (CNN) to learn spatial relations based on hallucinated objects, the researchers had to ensure that it was capable of classifying relations between individual pairs of object based on a single image. Subsequently, they "tricked" their network, dubbed RelNet, into classifying "hallucinated" scenes by implanting high-level features of items at different spatial locations.

"Our approach allows a robot to follow natural-language placing instructions given by human users with minimal data collection or heuristics," Mees said. "Everybody would like to have a service robot at home which can perform tasks by understanding natural-language instructions. This is a first step to enable a robot to better understand the meaning of commonly used spatial prepositions."

Most existing method for training robots to move objects around use information related to the objects" 3-D shapes to model pairwise spatial relationships. A key limitation of these techniques is that they often require additional technological components, such as tracking systems that can trace the movements of different objects. The approach proposed by Mees and his colleagues, on the other hand, does not require any additional tools, since it is not based on 3-D vision techniques.

The researchers evaluated their method in a series of experiments involving real human users and robots. The results of these tests were highly promising, as their method allowed robots to effectively identify the best strategies to place objects on a table in accordance with the spatial relations outlined by a human user's spoken instructions.

"Our novel approach of hallucinating scene representations can also have multiple applications in the robotics and computer vision communities, as often robots often need to be able to estimate how good a future state might be in order to reason over the actions they need to take," Mees said. "It could also be used to improve the performance of many neural networks, such as object detection networks, by using hallucinated scene

representations as a form of data augmentation."

Mees and his colleagues we able to model a set of natural language spatial prepositions (e.g. right, left, on top of, etc.) reliably and without using 3-D vision tools. In the future, the approach presented in their study could be used to enhance the capabilities of existing robots, allowing them to complete simple object shifting tasks more effectively while following a human user's spoken instructions.

Meanwhile, their paper could inform the development of similar techniques to enhance interactions between humans and robots during other object manipulation tasks. If coupled with auxiliary learning methods, the approach developed by Mees and his colleagues may also reduce the costs and efforts associated with compiling datasets for robotics research, as it enables the prediction of pixelwise probabilities without requiring large annotated datasets.

"We feel that this is a promising first step towards enabling a shared understanding between humans and robots," Mees concluded. "In the future, we want to extend our approach to incorporate an understanding of referring expressions, in order to develop a pick-and-place system that follows natural language instructions."

More information: Learning object placements for relational instructions by hallucinating scene representations. arXiv:2001.08481 [cs.RO]. arxiv.org/abs/2001.08481

APA citation: Training robots to identify object placements by 'hallucinating' scenes (2020, February 12) retrieved 14 May 2021 from https://techxplore.com/news/2020-02-robots-placements-hallucinating-scenes.html