Google Research Blog

The latest news from Research at Google

Teaching Robots to Understand Semantic Concepts

Friday, July 21, 2017

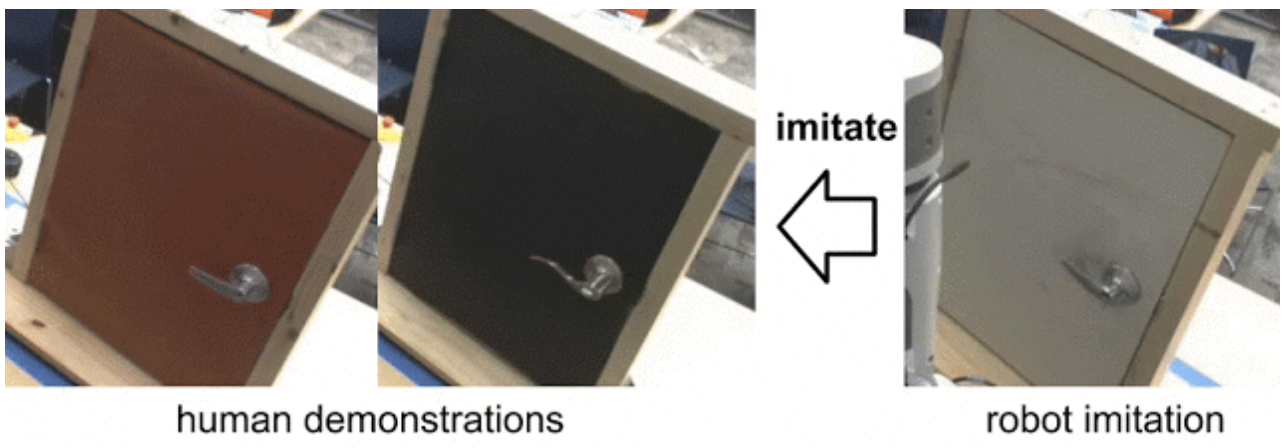
Posted by Sergey Levine, Faculty Advisor and Pierre Sermanet, Research Scientist, Google Brain Team

Machine learning can allow robots to acquire complex skills, such as grasping and opening doors. However, learning these skills requires us to manually program reward functions that the robots then attempt to optimize. In contrast, people can understand the goal of a task just from watching someone else do it, or simply by being told what the goal is. We can do this because we draw on our own prior knowledge about the world: when we see someone cut an apple, we understand that the goal is to produce two slices, regardless of what type of apple it is, or what kind of tool is used to cut it. Similarly, if we are told to pick up the apple, we understand which object we are to grab because we can ground the word “apple” in the environment: we know what it means.

These are semantic concepts: salient events like producing two slices, and object categories denoted by words such as “apple.” Can we teach robots to understand semantic concepts, to get them to follow simple commands specified through categorical labels or user-provided examples? In this post, we discuss some of our recent work on robotic learning that combines experience that is autonomously gathered by the robot, which is plentiful but lacks human-provided labels, with human-labeled data that allows a robot to understand semantics. We will describe how robots can use their experience to understand the salient events in a human-provided demonstration, mimic human movements despite the differences between human robot bodies, and understand semantic categories, like “toy” and “pen”, to pick up objects based on user commands.

Understanding human demonstrations with deep visual features

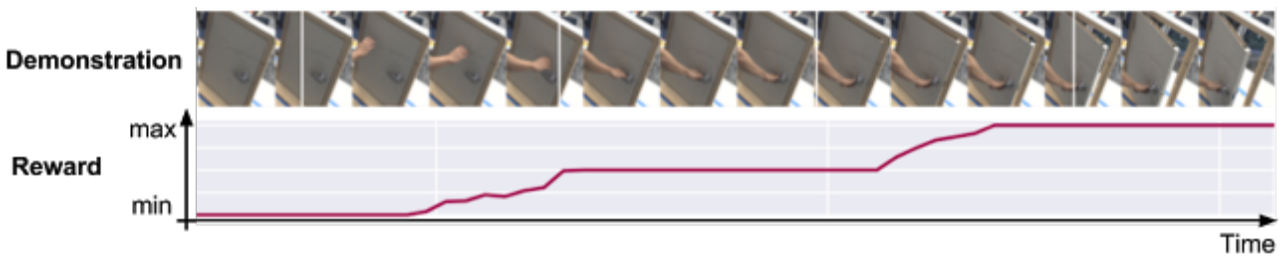
In the first set of experiments, which appear in our paper Unsupervised Perceptual Rewards for Imitation Learning, our is aim is to enable a robot to understand a task, such as opening a door, from seeing only a small number of unlabeled human demonstrations. By analyzing these demonstrations, the robot must understand what is the semantically salient event that constitutes task success, and then use reinforcement learning to perform it.



Examples of human demonstrations (left) and the corresponding robotic imitation (right).

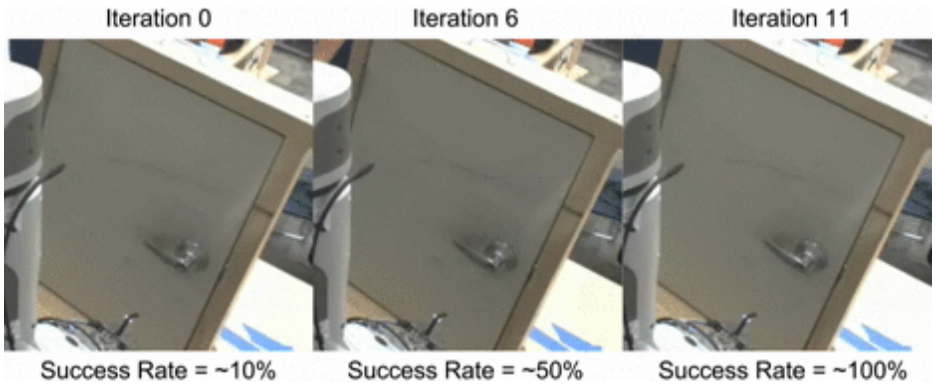
Unsupervised learning on very small datasets is one of the most challenging scenarios in machine learning. To make this feasible, we use deep visual features from a large network trained for image recognition on ImageNet. Such features are known to be sensitive to semantic concepts, while maintaining invariance to nuisance variables such as appearance and lighting. We use these features to interpret user-provided

demonstrations, and show that it is indeed possible to learn reward functions in an unsupervised fashion from a few demonstrations and without retraining.



Example of reward functions learned solely from observation for the door opening tasks. Rewards progressively increase from zero to the maximum reward as a task is completed.

After learning a reward function from observation only, we use it to guide a robot to learn a door opening task, using only the images to evaluate the reward function. With the help of an initial kinesthetic demonstration that succeeds about 10% of the time, the robot learns to improve to 100% accuracy using the learned reward function.

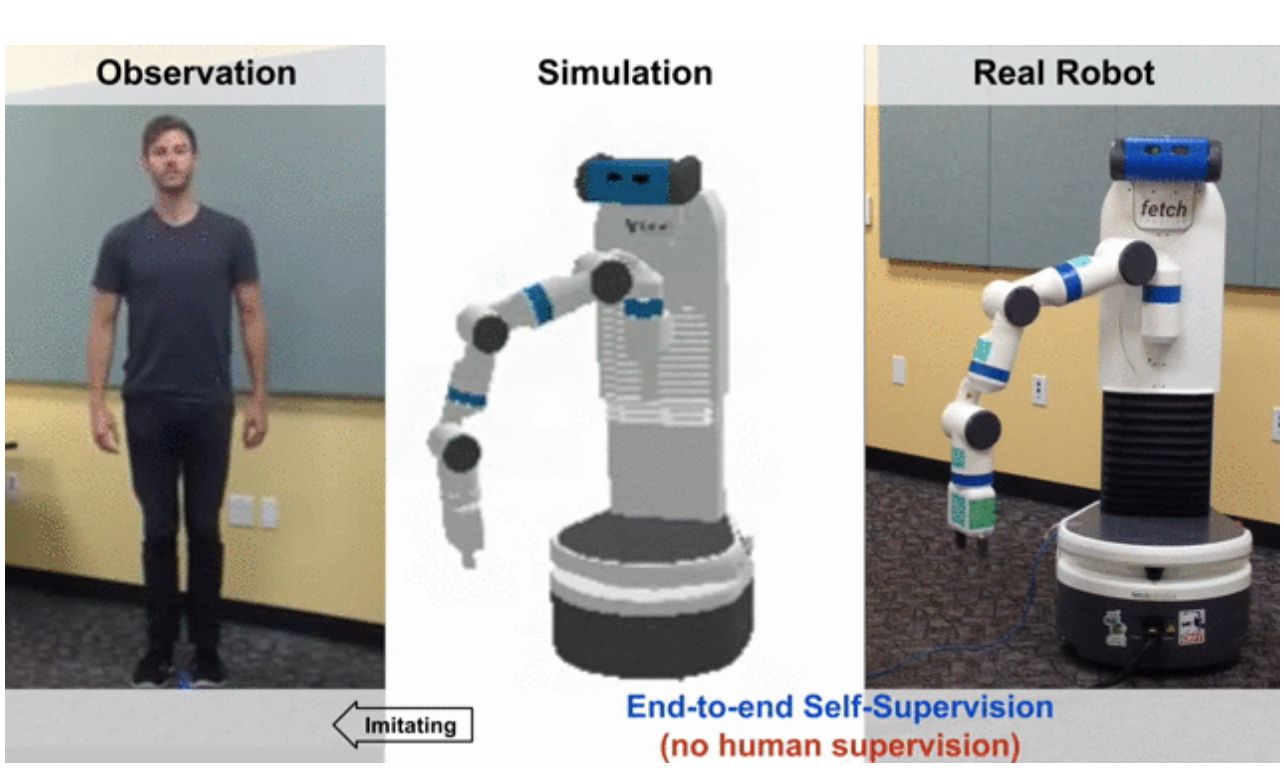


Learning progression.

Emulating human movements with self-supervision and imitation.

In Time-Contrastive Networks: Self-Supervised Learning from Multi-View Observation, we propose a novel approach to learn about the world from observation and demonstrate it through self-supervised pose imitation. Our approach relies primarily on co-occurrence in time and space for supervision: by training to distinguish frames from different times of a video, it learns to disentangle and organize reality into useful abstract representations.

In a pose imitation task for example, different dimensions of the representation may encode for different joints of a human or robotic body. Rather than defining by hand a mapping between human and robot joints (which is ambiguous in the first place because of physiological differences), we let the robot learn to imitate in an end-to-end fashion. When our model is simultaneously trained on human and robot observations, it naturally discovers the correspondence between the two, even though no correspondence is provided. We thus obtain a robot that can imitate human poses without having ever been given a correspondence between humans and robots.



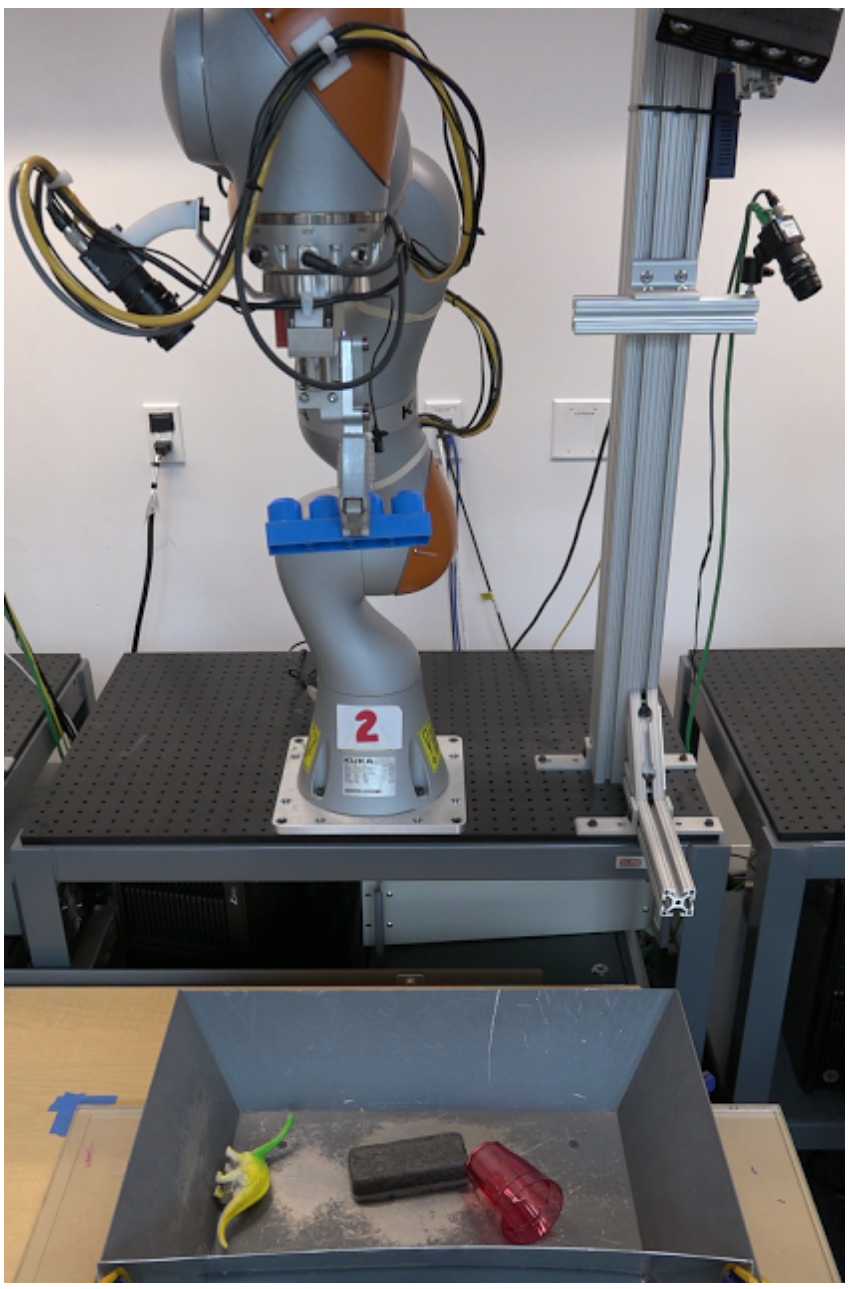
Self-supervised human pose imitation by a robot.

A striking evidence of the benefits of learning end-to-end is the *many-to-one* and *highly*

non-linear joints mapping shown above. In this example, the up-down motion involves many joints for the human while only one joint is needed for the robot. We show that the robot has discovered this highly complex mapping on its own, without any explicit human pose information.

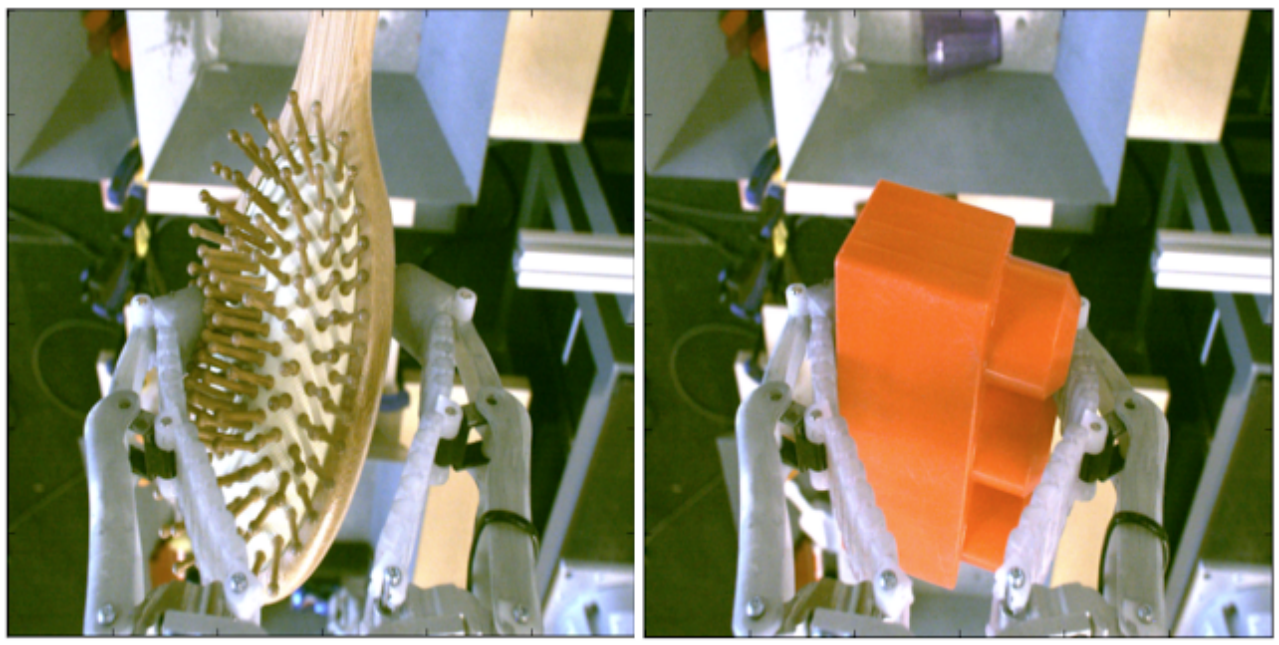
Grasping with semantic object categories

The experiments above illustrate how a person can specify a goal for a robot through an example demonstration, in which case the robots must interpret the semantics of the task -- salient events and relevant features of the pose. What if instead of showing the task, the human simply wants to tell it to what to do? This also requires the robot to understand semantics, in order to identify which objects in the world correspond to the semantic category specified by the user. In End-to-End Learning of Semantic Grasping, we study how a combination of manually labeled and autonomously collected data can be used to perform the task of *semantic grasping*, where the robot must pick up an object from a cluttered bin that matches a user-specified class label, such as “eraser” or “toy.”



In our semantic grasping setup, the robotic arm is tasked with picking up an object corresponding to a user-provided semantic category (e.g. Legos).

To learn how to perform semantic grasping, our robots first gather a large dataset of grasping data by autonomously attempting to pick up a large variety of objects, as detailed in our previous post and prior work. This data by itself can allow a robot to pick up objects, but doesn’t allow it to understand how to associate them with semantic labels. To enable an understanding of semantics, we again enlist a modest amount of human supervision. Each time a robot successfully grasps an object, it presents it to the camera in a canonical pose, as illustrated below.



The robot presents objects to the camera after grasping. These images can be used to label which object category was picked up.

A subset of these images is then labeled by human labelers. Since the presentation images show the object in a canonical pose, it is easy to then propagate these labels to the remaining presentation images by training a classifier on the labeled examples. The labeled presentation images then tell the robot which object was actually picked up, and it can associate this label, in hindsight, with the images that it observed while picking up that object from the bin.

Using this labeled dataset, we can then train a two-stream model that predicts which object will be grasped, conditioned on the current image and the actions that the robot might take. The two-stream model that we employ is inspired by the dorsal-ventral decomposition observed in the human visual cortex, where the ventral stream reasons about the semantic class of objects, while the dorsal stream reasons about the geometry of the grasp. Crucially, the ventral stream can incorporate auxiliary data consisting of labeled images of objects (not necessarily from the robot), while the dorsal stream can incorporate auxiliary data of grasping that does not have semantic labels, allowing the entire system to be trained more effectively using larger amounts of heterogeneously labeled data. In this way, we can combine a limited amount of human labels with a large amount of autonomously collected robotic data to grasp objects based on desired semantic category, as illustrated in the video below:

End-to-End Learning of Semantic Grasping

Future Work

Our experiments show how limited semantically labeled data can be combined with data that is collected and labeled automatically by the robots, in order to enable robots to understand events, object categories, and user demonstrations. In the future, we might imagine that robotic systems could be trained with a combination of user-annotated data and ever-increasing autonomously collected datasets, improving robotic capability and easing the engineering burden of designing autonomous robots. Furthermore, as robotic systems collect more and more automatically annotated data in the real world, this data can be used to improve not just robotic systems, but also systems for computer vision, speech recognition, and natural language processing that can all benefit from such large auxiliary data sources.

Of course, we are not the first to consider the intersection of robotics and semantics. Extensive prior work in natural language understanding, robotic perception, grasping, and imitation learning has considered how semantics and action can be combined in a robotic system. However, the experiments we discussed above might point the way to future work into combining self-supervised and human-labeled data in the context of autonomous robotic systems.

Acknowledgements

The research described in this post was performed by Pierre Sermanet, Kelvin Xu, Corey Lynch, Jasmine Hsu, Eric Jang, Sudheendra Vijayanarasimhan, Peter Pastor, Julian Ibarz, and Sergey Levine. We also thank Mrinal Kalakrishnan, Ali Yahya, and Yevgen Chebotar for developing the policy learning framework used for the door task, and John-Michael Burke for conducting experiments for semantic grasping.

Unsupervised Perceptual Rewards for Imitation Learning was presented at RSS 2017 by Kelvin Xu, and Time-Contrastive Networks: Self-Supervised Learning from Multi-View Observation will be presented this week at the CVPR Workshop on Deep Learning for Robotic Vision.

54 条评论

以“Zhao Haijun”的身份发表评论

热门评论

Research at Google 通过 Google+ 3天前 - 公开分享

Can we teach robots to understand semantic concepts, to get them to follow simple commands specified through categorical labels or user-provided examples? Below on the Google Research blog, we discuss some of our recent work on robotic learning that combines experience that is autonomously gathered by the robot, which is plentiful but lacks human-provided labels, with human-labeled data that allows a robot to understand

翻译

+65 1 · 回复

查看所有 27 条回复

Armada jakenson 2天前

+**Anders Feder** ALL I hear from you is MR Negative! BITTER!

翻译

Anders Feder 2天前

+**Armada jakenson** All I hear from you are phony excuses for mental deficiencies.

翻译

Glenn Gabe 通过 Google+ 12小时前 - 公开分享

Amazing & SCARY. Google explains how its robots learned by *watching human behavior & mimicking human movements*

翻译

+5 1 · 回复

1K Living 12小时前

This is really nice.

翻译

Albert Jones 9小时前

Tap teixe

I.J. Atencio分享了此信息 3天前 - Machine Learning (Supervised Learning)

1

Emre Safak 通过 Google+ 10小时前 - 公开分享

此信息最初是由**Research at Google**分享的

Can we teach robots to understand semantic concepts, to get them to follow simple commands specified through categorical labels or user-provided examples? Below on the Google Research blog, we discuss some of our recent work on robotic learning that combines experience that is autonomously gathered by the robot, which is plentiful but lacks human-provided labels, with human-labeled data that allows a robot to understand semantics.

翻译

1 · 回复

Greg Linden 通过 Google+ 2天前 - 公开分享

Labels: Deep Learning , Google Brain , Research , Robotics