Levels of Abstraction in Machine Learning

Henry Blanchette

1 Introduction

The techniques of machine learning (ML) in the broader field of artificial intellgence (AI) has undeniably made huge headway recently. However there still seem to be some significant inadequacies with the state-of-the-art machine *intelligence* even for relatively narrow tasks. For example, modern ML in the form of a convolutional neural network (CNN) has achieved super-human ability in the classic computer game of Breakout, where success is measured by how quickly the algorithm can beat the game. The CNN achieved this level of skill in Breakout from "the ground up", meaning that it was given no special information about the game other that control over input to the game (moving the paddle) and access to the output from the game (looking at the game screen).

A human reader recognize that, in order to gain skill at Breakout, one needs to recognize or l earn certain patterns about the physics of the game. One learns that the ball bounces off the paddle and this is the only way to keep the ball from going out of bounds, and that the ball bounces off of the bricks causing them to disappear and scores a point. Somewhere along the way, one learns that hollowing out a tunnel through all the brick layers lets the ball bounce around consecutively in the top area scoring lots of points.

But is this how the CNN learned to play? Are there analogous concepts of *ball*, *paddle*, *brick*, and *tunnel*? One way to probe for answers is to change some feature of the game that requires the player to make use of one of these abstractions in order to readjust their strategy. And this is where the CNN exposes the fact that the knowledge that it has in the form of intricate weighted connections between nodes in its layers may not likely resemble anything like we'd expect from a human that is considered knowledgable or skilled at breakout. If the game is modified so that the paddle is slightly higher on the vertical, a human player will have no trouble recognizing how to control this paddle and apply the same skills that they may have learned from the normal version of the game (e.g. tunneling). But the CNN utterly fails at this task. When the paddle is moved, it loses all apparently skill at the game. When it is trained on the new version of the game, starting from its previously-skilled configuration of weights, it basically scrambles its weights randomly before it re-achieves its previous level of skill. The CNN was unable to salvage any of its knowledge when it needed to account for even such a minor change.

The significance of this observation is that it suggest the CNN likely didn't *understand* the game of Breakout in the same way we expect a human to understand the game, even though the CNN was super-human in performance. There are many other examples of this phenomenon, that Melanie brought up during her visit, such as how an image classifier CNN can become highly-proficient at recognizing images of, say, cows, yet when presented with an empty field of grass the CNN still sees cows. Or in some of the more extreme examples, one can engineer a *adversarial example* that is carefully designed to take advantage of an image classifier CNN's trained structure to make it recognize whatever the designer wishes in an image while making the image appear relatively unchanged to a human viewers.

2 Levels of Abstraction

Recalling the Breakout case, the obvious question to ask of the CNN is "if the CNN didn't learn even basic abstractions like *the paddle*, then what *did* the CNN learn that allowed it to play so amazing well?" While it is possible that CNN learned a few human-recognizable abstractions, the fact that it misses many that humans would find extremeley critical (e.g. the abstraction of *the paddle*) suggests that there probably aren't many and even the seeming presence of these patterns might just be a facade for a superficial representation of the abstraction inside the CNN.

A simple and likely explanation for the CNN's behavior that accounts for its pattern-obliviousness is that it discovered a sort of "cheat" in the form of statistical correlations between certain pixels and player moves that were easier to learn that the human-level abstractions that humans would expect. In the previous section when the steps of learning how to play Breakout was touched on, it should have been clear that humans can skip many of the early steps because humans have an intuitive sense of physics and object-ness. Humans know how to expect the ball to bounce around the screen even if they have never looked at a Breakout game. Additionally, humans look at the paddle as an *object* that can possibly move vertically and still remain the same object, as well as the usual left/right. (The question of how or whether humans learn these intuitions, as I describe them, in the first place is an intriguing question that will be addressed later.)

These "cheats" are examples of *first-order patterns*.

3 Experimental Design

4 Results

5 Conclusions

References