Vision as Inverse Graphics

Machine learning techniques towards a program-based model for scene understanding

Vinjai Vale

May 21, 2017
MIT PRIMES CS conference

Scene Understanding

Recognize objects and components

Answer questions about scene

Why is scene understanding hard

Scene understanding is easy for humans.

What is this a picture of?

What is the man doing?

What is the man’s jersey number?

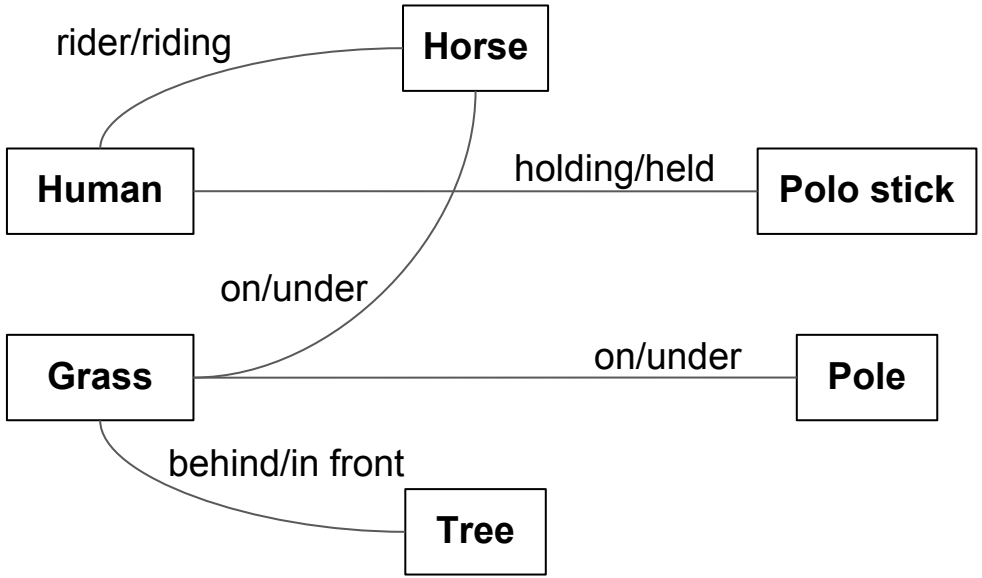
How fast is the horse going?

These questions are second nature for humans to answer, but are really difficult for a computer.

Scene Understanding

Scene: Polo

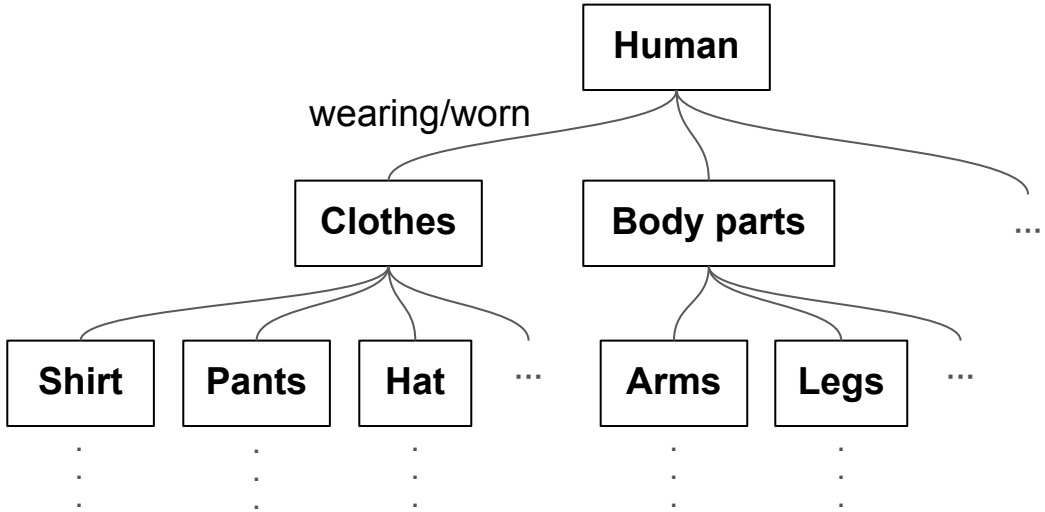
Primary Objects



Scene Understanding

Object: Human

Components (Secondary/Tertiary Objects)



Current approaches

Classify objects and *infer* the scene context based on the types of objects present

Types of objects present: Horse, Human, Stick, Grass, Tree —> Scene is Polo

Top-down approach; gets the gist of the scene

Google image captioner

Image source: Google paper “Show and Tell: A Neural Image Caption Generator,” CVPR 2015

Another example

Computer: A zebra, of course!

Human: A horse in a zebra costume, of course!

Image source: Google Images

Current approaches lack **compositionality**

A *compositional representation* is one where complicated objects or scenes are represented by putting together simpler parts.

Compositionality is second nature to humans, but not to computers!

Alternative approach

Goal: accomplish scene understanding by creating an **abstraction** that includes information about the following:

- The type of each object and each component
- How each component is related to the object that it is a part of, and vice versa (thereby encoding the specifics of each object)

Vision as inverse graphics

Alternate paradigm:

Analyze an image by attempting to synthesize it

Program-based model