CS231n

Thomas Lu

Note: statements that I am unsure about are enclosed in double square brackets, e.g. [[Every even integer greater than 4 can be expressed as the sum of two primes.]]

1 Lecture 1: Introduction

Computer vision is a hugely important and impactful field today. In recent years, the number of cameras in the world has exploded, and so correspondingly has the amount of visual data in existence, and, therefore, the available opportunity from understanding that data. Today (2017) visual data constitutes the vast majority of Internet traffic - for example, 5 hours of video gets uploaded to YouTube every second. There is also biological evidence for the imporance of visual understanding - a "evolutionary big bang" where the number of animal species rapidly shot up around 540M years ago coincided with what we believe to be the first occurrences of eyes in animals.

A brief history of computer vision:

- 1959: An early neurobiological study on vision in cats found that simple cells in the brain respond to edges of certain orientation, and that successive layers of more complex cells depended on the previous ones and responded to higher-level elements.
- 1963: A PhD thesis called Block World that many consider to be foundational to computer vision was introduced. The thesis described an attempt to make a computer visually understand a world made up of geometric shapes.
- 1970s: David Marr suggested that in going from a 2D picture to a 3D representation, human and primate brains first turn the 2D picture into a "primal sketch", then a "2.5D sketch", before finally arriving at a 3D representation of the world.
- $\tilde{1}$980: researchers dreamed up other ways to visually represent objects: Generalized Cylinder and Pictorial Structure were two of the more well-known ones.
- 1997: An image segmentation technique called Normalized Cut was introduced. It would cluster pixels in an image into semantically meaningful groups.
- 1999: SIFT was introduced. SIFT was a method for finding invariant "critical points" in an image that could be used to match similar objects to each other.
- 2001: A group of researchers figured out how to do face detection cheaply enough for the technology to be built into a handheld digital camera.
- 2006: Spatial Pyramid Matching, a technique for image classification, was introduced.
- 2005 (Histogram of Gradients) and 2009 (Deformable Part Model): Human detection techniques.
- 2005: The PASCAL Visual Object Challenge was introduced - an object classification benchmark dataset with around 20k images and 20 classes.

- 2006: Work began on creating ImageNet, a huge image classification dataset with (now, 2021) 14M images and 22k categories.
  - The motivation is that we need a big, complex model to effectively process images, and big models need to be trained on a lot of data.
  - There was also the question of whether it was possible to train a model to "recognize everything."
  - In 2010, the first ImageNet Large Scale Visual Recognition Challenge was held. In the beginning, top models averaged around 75% top-5 recall.
  - A step-function changed in 2012, when a strong CNN model (AlexNet) achieved 84% top-5 recall.
  - By 2015, top-5 recall surpassed human performance (over 95%). Since then, CNNs have gotten even deeper and even more powerful.
  - But CNNs weren't that new! The idea was actually introduced in 1998 by Yann LeCun. But the reason they took a long time show such good performance was because hardware and training data availability needed to catch up.

A lot of the latest advances in computer vision literature have been around image classification, but there are other important fundamental problems too, e.g.

- Semantic segmentation: given an image and a particular pixel, what object is the pixel a part of?
- 3D representation: given a 2D image, construct a 3D model of the scene.
- Activity recognition: given a video, figure out what the subject is doing.
- Image description: given an image, write a detailed natural-language description of the content.
- Humor understanding: given a humorous image, explain why it's funny.

And these problems have a ton of impactful applications: content moderation, medicine, autonomous driving, and more!