HOW WE TEACH COMPUTERS TO UNDERSTAND PICTURES

A WHITEPAPER ON THE HISTORY OF COMPUTER VISION AND STATE OF THE ART ALGORITHMS

DHRUVIT KOTHARI

TABLE OF CONTENTS

Introduction to Computer Vision

For decades now people have been envisioning machines that are equally or more intelligent than humans. One of the main obstacles of this goal is to teach machines how to see like humans. The ability to detect objects and then disentangle information from their images is of paramount importance when making a machine that is smarter than humans.
Computer Vision basically means that the machine/computer is able to extract high level meaningful information from an image. It should be able to understand the surrounding world using that information and then use it to perform more complex tasks like making decisions.
Humans are great at vision. A child by the age of three has mastered vision and in most cases is able to describe what she/he can see in an image.

Figure 1: A man standing besides an elephant
A child would look at the above image and recognise the man and elephant.

Earlier scientists thought that making computers that can beat humans in chess would be harder than making a machine to master human vision. Now years later we can see that humans have not been able to beat the top chess computers for 2 decades, while Computer Vision still remains an unsolved problem. Computer Vision is a very trivial problem at the surface but somehow we still aren't able to teach machines to see and understand things like a human does. The problem seems so trivial to us because we can do it with great accuracy without even thinking about it.For example, a person can summarise a video content after watching it just once.

The state of the art Computer Vision algorithm today can perform a variety of tasks like detecting objects from an image, classifying dogs and cats, and in some cases even identify the breed of the animal. The progress made in the field of Computer Vision is tremendous but it still doesn't compare to human vision.

From the perspective of an engineer, Computer Vision basically means automating machines to do tasks that a human visual system can perform. On the other hand, to a scientist Computer Vision is to replicate the ability of the human visual system.

Why do we want computers to see and make sense of images?

In today's world we are surrounded by images and videos. The constant rate of images and videos being uploaded on social media is so massive that it would take a million years for a human eye to see all of the images/videos that are uploaded in 1 minute. Images and videos have not only transpired from being a token of memories but now are a medium for people to express themselves freely. All these pictures are uploaded and indexed on the internet. We want to create a computer/machine that is able to make sense of the content of these images.

Computer Vision is actively used to perform crucial tasks like surveillance, medical imaging, auto vehicle safety etc. Usage of Computer Vision has increased exponentially in recent years with every company employing Computer Vision techniques to automate jobs and free up company resources. Also, Computer Vision is a key part of the puzzle of developing an artificial general intelligence. The field of Computer Vision is dotted with historical breakthroughs which propelled its progress.

Why is Computer Vision Hard?

On paper, Computer Vision doesn't seem like a hard problem. The abstractness of the idea of human vision is the main hurdle in machine learning and understanding images. A computer can perform any task which is well defined.For example, if we tell a machine to add two numbers it will perform the task very quickly and return the result. The concept of adding two numbers is algorithmically defined with no boundary cases to consider. “2 + 2 =4”, this statement is true for all computers and there is no doubt about what the program is trying to do. On the other hand, Computer Vision is an abstract problem. What does it mean to see? How do humans make relations between different objects? How do humans recognise a face that they have seen before? All these questions are vague and it is very difficult to come up with a definite algorithm that would make a computer learn to see and understand images. Computers only understand the languages of numbers and translating these concepts to computers seems like an impossible task. Many attempts have been made in the past to teach computers to see and detect objects. And we have only recently been able to partially solve this problem of detecting objects. Understanding high level information about the image is still a far fetched proposition.

Figure 2: Obama pulling a prank

Today's state of the art Computer Vision system would see this picture and detect the number of people in the picture. Some other Computer Vision models may even attempt to summarise the image by telling us what is happening in the image. “5 people are standing in the hallway” would be a reasonable estimate of some of these models. But to us this picture has way more content and information. We can recognise that a person is standing on the weight machine to measure his weight while some other person is keeping his foot on the machine to increase the reading. The social context of weight to a person and why the increased reading makes the other people laugh is outside the reach of Computer Vision models’ understanding today. We also recognise the person in the image as Obama and the action by a man of his position makes the situation even more amusing. We hope that one day, a machine is able to grasp all this from a single look at the picture.

The other reason Computer Vision is difficult to solve is because the human visual system is way too advanced when compared to state of the art Computer Vision. A human can recognise faces under all kinds of variation in conditions like illumination, angle of the viewpoint , expression etc. Human visual system is also good at filling up occluded subjects i.e objects which can be only partially seen. A glimpse of the ears of a cat peeking through the box is enough for us to classify it as a cat. There is no practical way as of today to implement this relational knowledge in a Computer Vision system.

History of Computer Vision

Over the years research in Computer Vision has taken great strides. With the advent of big data i.e generation of images and videos at a pace which has never been seen in history, Computer Vision algorithms have gotten better and better. The increase in computing power has also led to quick testing of key ideas in Computer Vision. Today researchers have greater access to data and computing power to develop new Computer Vision algorithms and systems.
But this was not always the case. Here we look into the key milestones without which Computer Vision would still be considered impossible. Today's state of the art algorithms like Convolutional Neural Network are heavily inspired by this research.

a. Hubel and Wiesel Experiment on cats (1959)

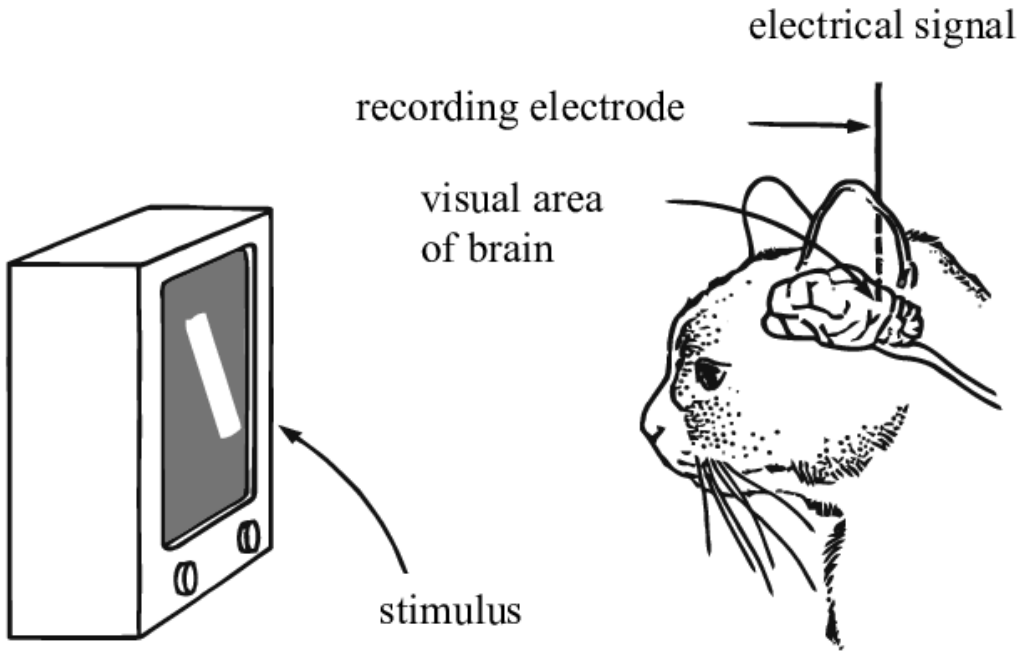


Figure 3: Hubel and Wiesel Experiment on cats

In 1959, neurophysiologist David Hubel and Torsten Wiesel conducted multiple experiments by inserting a microelectrode in the primary visual cortex of an anaesthetised cat. During the experiment they found that when the cat is shown a light and dark pattern of light at a certain angle some of the neurons are fired. When the angle is changed some other neurons in the region are fired. They named these neurons as “simple cells”. They also found neurons which respond to moving patterns of light and dark. These neurons were named as “complex cells”. This led to the important discovery that the visual system makes sense of complex representation by processing multiple simple stimuli operations.
This discovery inspired the deep architecture of convolutional neural networks (state of the art). Many convolutional neural networks today are viewed as a stacked model of cell types inspired by Hubel and Wiesel.

b. Lawrence Roberts Phd. “Machine Perception of Three Dimensional Solids” (1963)

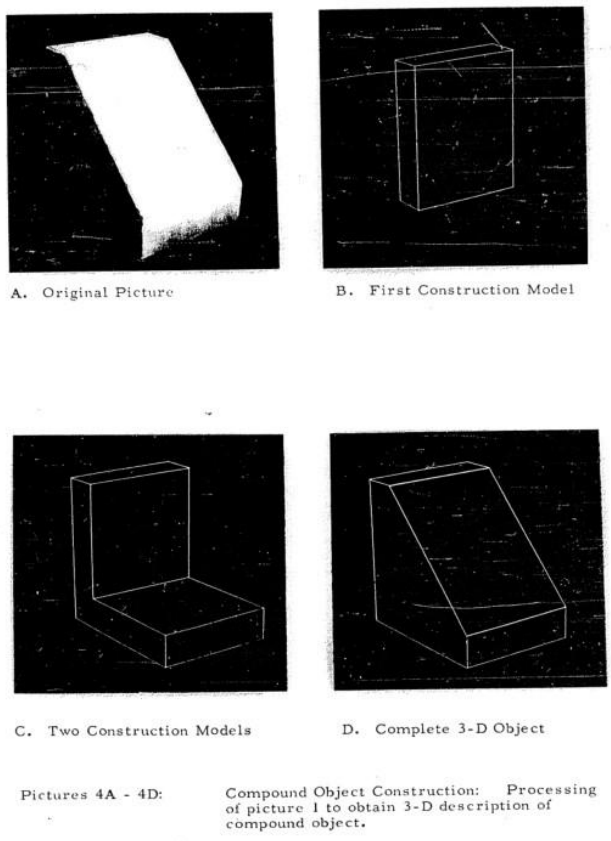
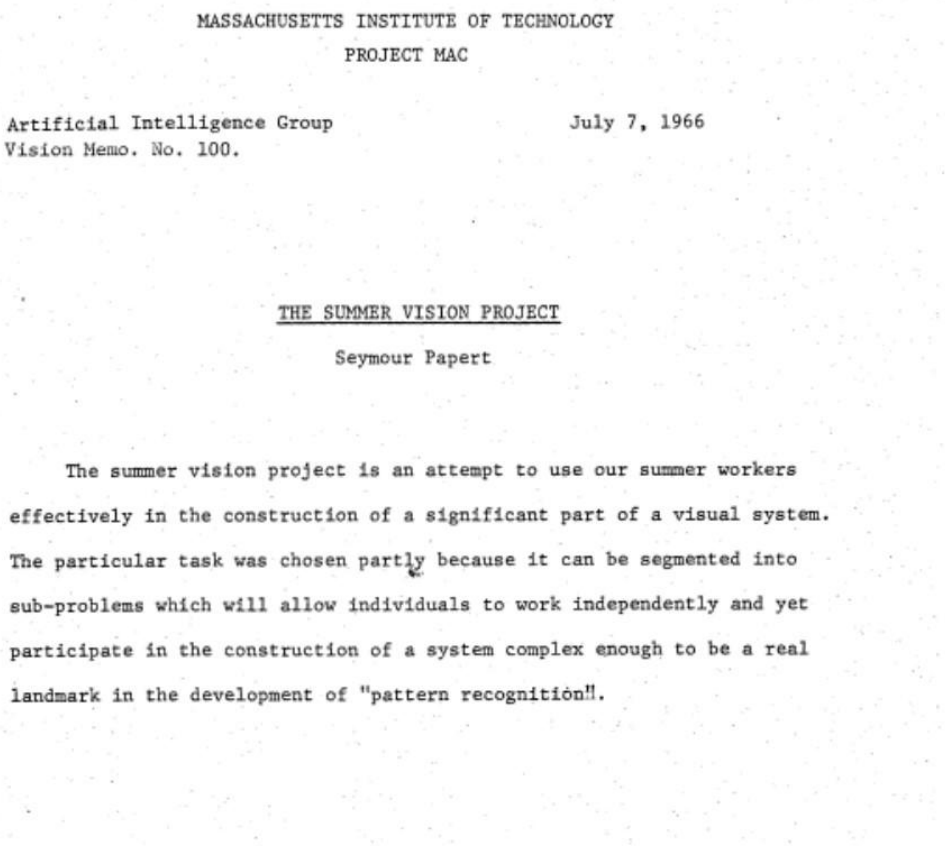


Figure 4: Computer reconstruction of a 3D object

In 1963, Lawrence Roberts published his thesis, “Machine Perception of Three Dimensional Solids”, which is considered by many as the start of the Computer Vision field. He is dubbed as the “Father of Computer Vision” in the Computer Vision community. His thesis that describes the extraction of 3D information of the world from 2D photographs of line drawing is still considered valuable today. This inspired Robert to see that edges and line drawings can be used to construct vision. The work done by him in his thesis led to multiple researchers to follow his work and progress Computer Vision.

c. Marvin Minsky's MIT Summer Project (1966)



MASSACHUSETTS INSTITUTE OF TECHNOLOGY
PROJECT MAC

Artificial Intelligence Group July 7, 1966
Vision Memo. No. 100.

THE SUMMER VISION PROJECT

Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

Figure 5 : MIT Summer Project 1966

In the summer of 1966 , Marvin Minsky gave Gerald Sussan, one of his undergraduate students a project. . The goal of the project was to connect a camera to a computer and make the computer describe what it saw. Thus began the quest to solve the visual input problem of Computer Vision which continues to remain completely unsolved till date. The computer

had to analyse the scene and detect objects present in the scene. This is easier said than done, we have been able to make progress towards the problem but it still remains elusive.

d. Neocognitron (1980)

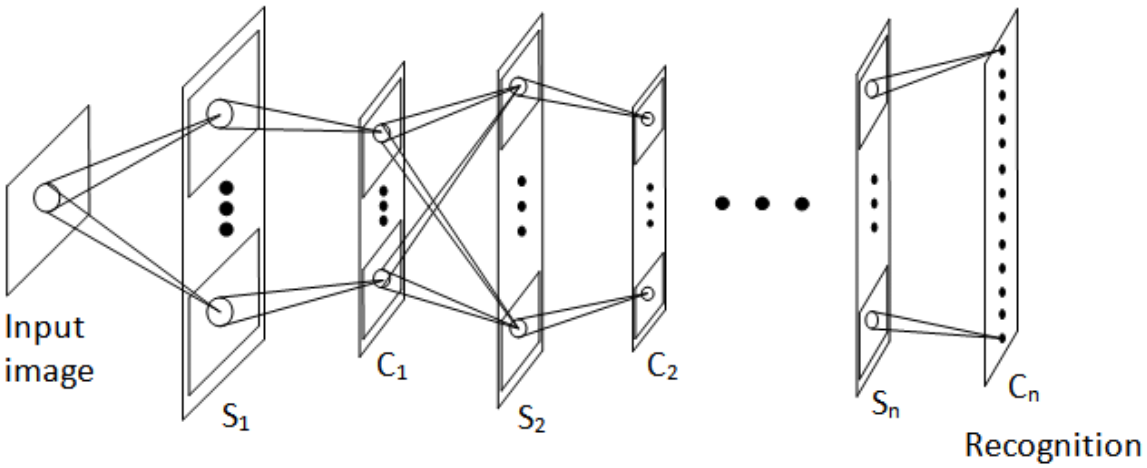


Figure 6: Neocognitron Architecture

In 1980 , Kunihiko Fukushima developed the 'Necognitron' architecture to process complex visual patterns. Fukushima, inspired by the 1959 work of David Hubel and Torsten Wiesel, used two types of cells, 'S-cells' and 'C-cells' in its architecture. The goal of the "C-cells" was to capture complex features from the image like 'a cat' while the other 'S-cells' detected lower level features like lines, edges etc. The neocognitron is considered as the inspiration for all existing convolutional neural networks. It was used for handwritten pattern recognition and other pattern detection tasks. Nine years later Yann LeCun of Bell labs made the first large scale commercially applied Computer Vision system based on the neocognitron architecture.

e. LeNet (1989)

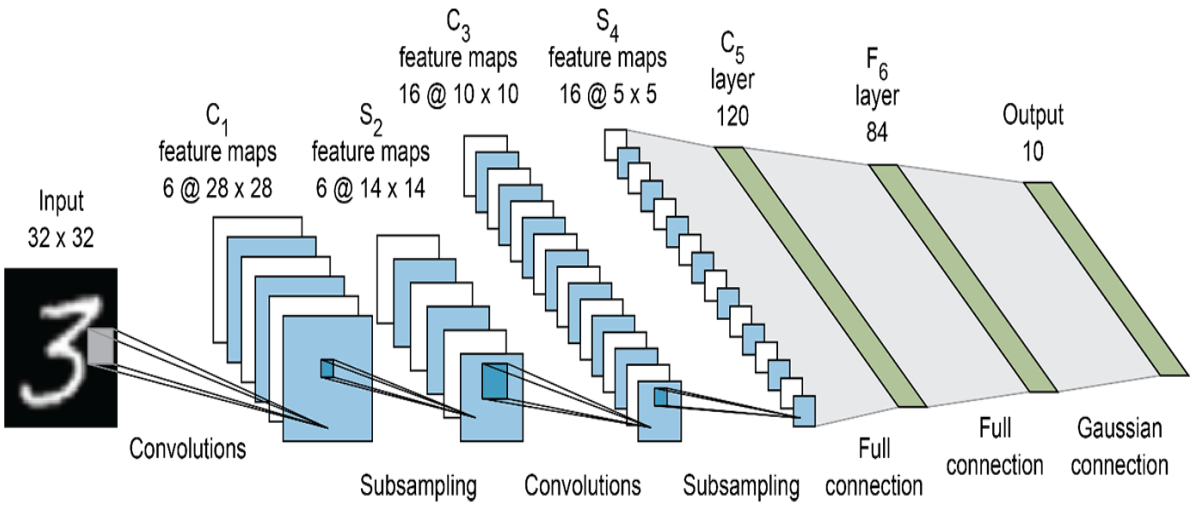


Figure 7: LeNet Architecture

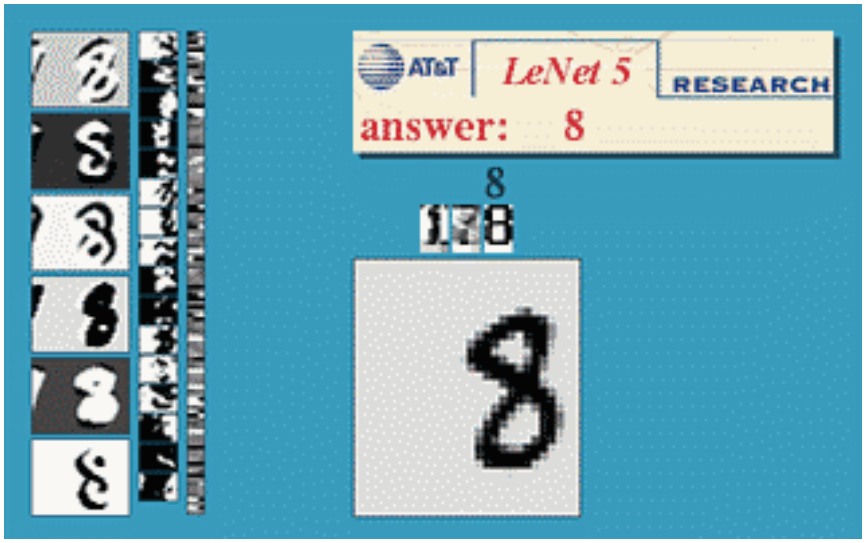


Figure 8: LeNet used to recognise handwritten digits

In 1989, Yann LeCun who worked at the famous Bell Labs invented the first commercial large scale Computer Vision system. He applied the backpropagation algorithm to the convolutional neural architecture. Convolutional Neural Networks, also known as CNN are a subset of general neural network architecture. The main difference between the two is that CNN employs domain information about pictures to increase its performance. Yann LeCun successfully trained CNN to recognise handwritten digit zip code numbers provided by the US Postal Service. This led to the creation of the MNIST dataset which is still used as a benchmark for many Computer Vision algorithms. The MNIST dataset consists of 60,000 training images and 10,000 testing images of handwritten digits. Each digit is a 32 x 32 grayscale image. Yann LeCun's LeNet takes a 32x 32 picture of a handwritten digit and applies different convolution and pooling schemes to give out an output in the end. The training of the neural network using backpropagation made it widely successful.

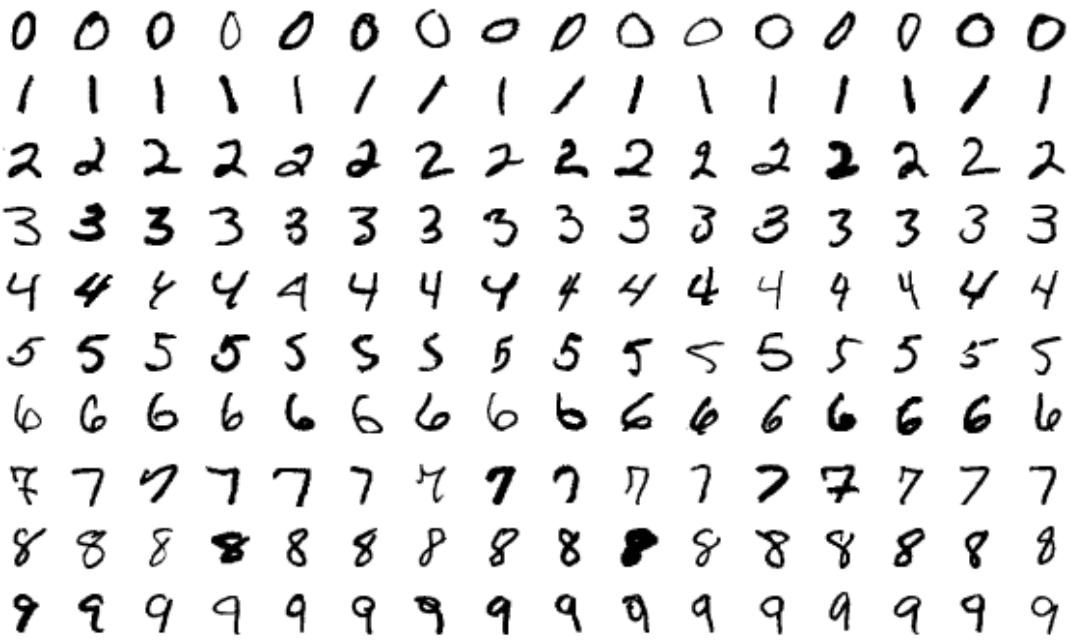


Figure 9: Example of the MNIST Dataset.

f. Eigenface (1991)

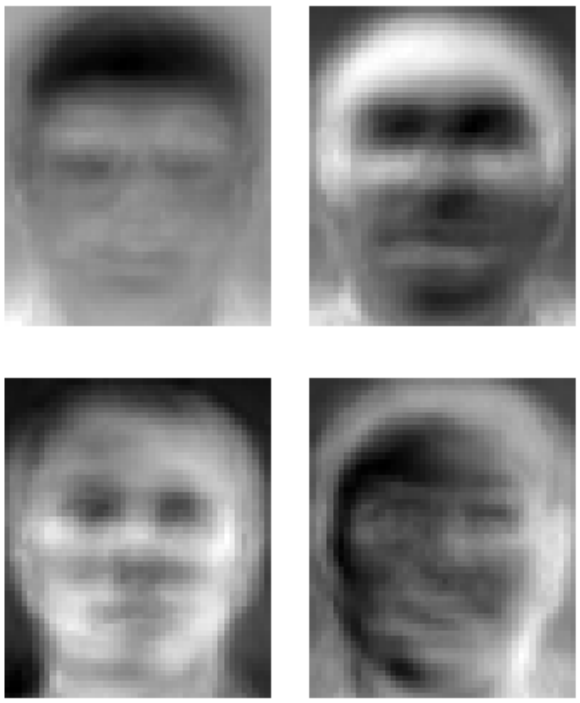


Figure 10: Example of Some Eigenface from the AT&T Laboratories

In 1991, Mathew Turk and Alex Pentland applied principal component analysis on a large set of face pictures to perform facial recognition. These independent vectors extracted by principal component analysis were used to construct eigenfaces of the faces The eigenfaces thus created had patterns to evaluate for symmetry, size of the mouth and nose, position of the hairline, style of facial hair etc. Recognition of the face was achieved by comparing the test images to the constructed eigenfaces. This technique of using principal component analysis can also be used for digit recognition, medical imaging, and sign language/gesture interpretation.

g. Viola-Jones Object Detection Framework (2001)

FIgure 11: Faces detected in real time by the Viola-Jones framework

In 2001 , Computer Vision researchers, Paul Viola and Micheal Jones proposed the Viola-Jones Object Detection Framework. This framework was used to detect objects in images and was particularly good at detecting faces. The Viola-Jones Object Detection Framework combines the concepts of Haar-like Features, Integral Images, the AdaBoost Algorithm, and the Cascade Classifier to create a system for object detection that is fast and accurate. By 2005, almost all mainstream camera companies employed the Viola-Jones algorithm in their cameras. Viola-Jones algorithm works by sliding a window over the picture

and outputting the sliding window which contains the object/face. The classifier trained by using adaBoost outputs fast accurate results during test time.

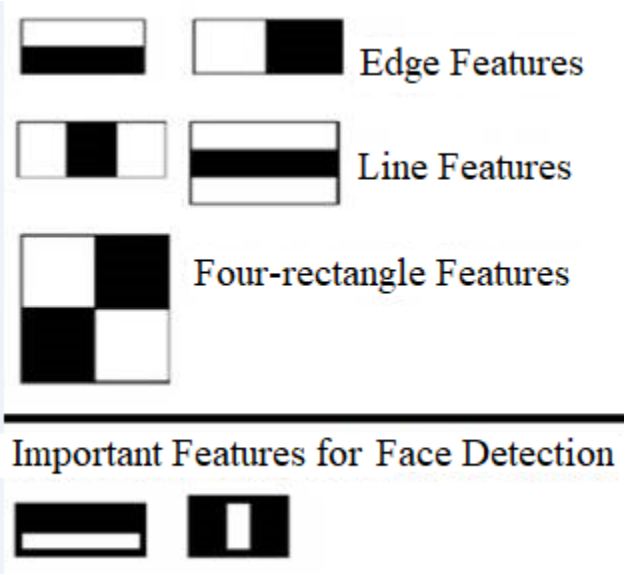


Figure 12: Haar Like features for detecting edges and lines

h. ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (2010)

Figure 13: ImageNet database

In 2010, the first annual ILSVRC competition was held. Its goal was to train an object detection classifier on the ImageNet database. The ImageNet database is a collection of more than 14 million hand-annotated images of objects with their respective bounding box provided. In modern day Computer Vision, the database has become one of the most important benchmarks. The large database provided was presented as a challenge to other Computer Vision practitioners to improve, promote and share ideas about the current vision and object detection algorithms.

I. AlexNet 2012

In 2012, one of the entries submitted to the competition paved the way ahead for the entire Computer Vision community. AlexNet, a convolutional neural network architecture, achieved a top-5 error rate of 16.4% which was 10% lower than the next top entry. This achievement was tremendous for the Computer Vision and neural network community.

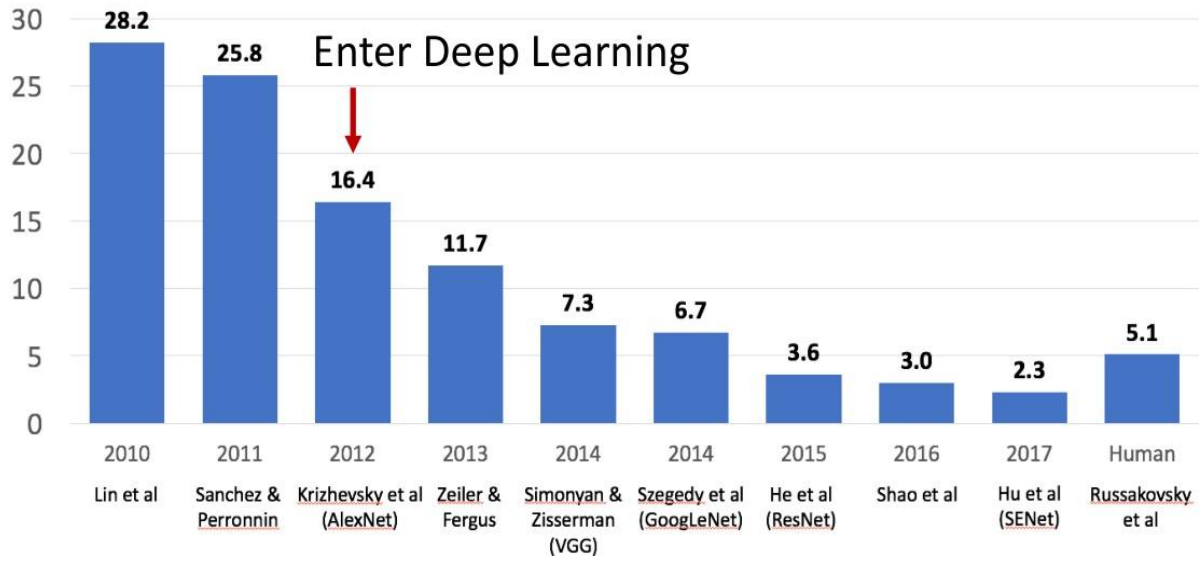


Figure 14: Top-5 Error rate of the ILSVRC competition

AlexNet, designed by Alex Krizhevsky in collaboration with Iliya Sutskever and Geoffrey Hinton, was the first Convolutional neural network implementation to win the ImageNet Challenge. The significant drop in the top-5 error rate from 25.8% to 16.4% attracted a lot of Computer Vision researchers and practitioners to view CNN as a serious tool to solve Computer Vision. So deep was the impact of AlexNet, that all winners of the ImageNet competition in the subsequent years were CNN with wider and deeper architectures.

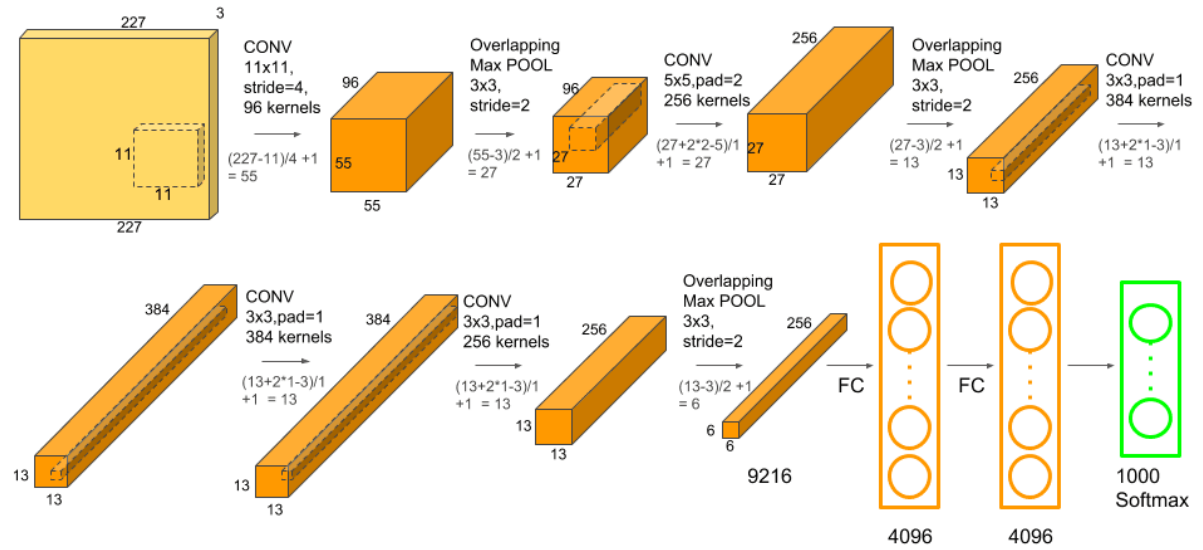


Figure 15: AlexNet Architecture

Convolutional Neural Networks is a deep learning architecture developed employing domain knowledge about Computer Vision. The domain knowledge includes the compositionality of the world i.e the objects are composed of much smaller objects which come together to form the world.For example, the lines and edges in the picture make up a face or an object to detect. The second domain knowledge employed in the architecture is the property of translational invariance in images i.e if an object moved a few pixels to the right it is still the same object. Making a computer understand and employ this knowledge through CNN gave a boost in trivial Computer Vision tasks.
Deep neural network architecture is loosely inspired by the human visual system hence the name 'neural' in them. The connectivity of neurons and firing of different neurons for different stimuli are incorporated in the architectures of many deep neural networks today.

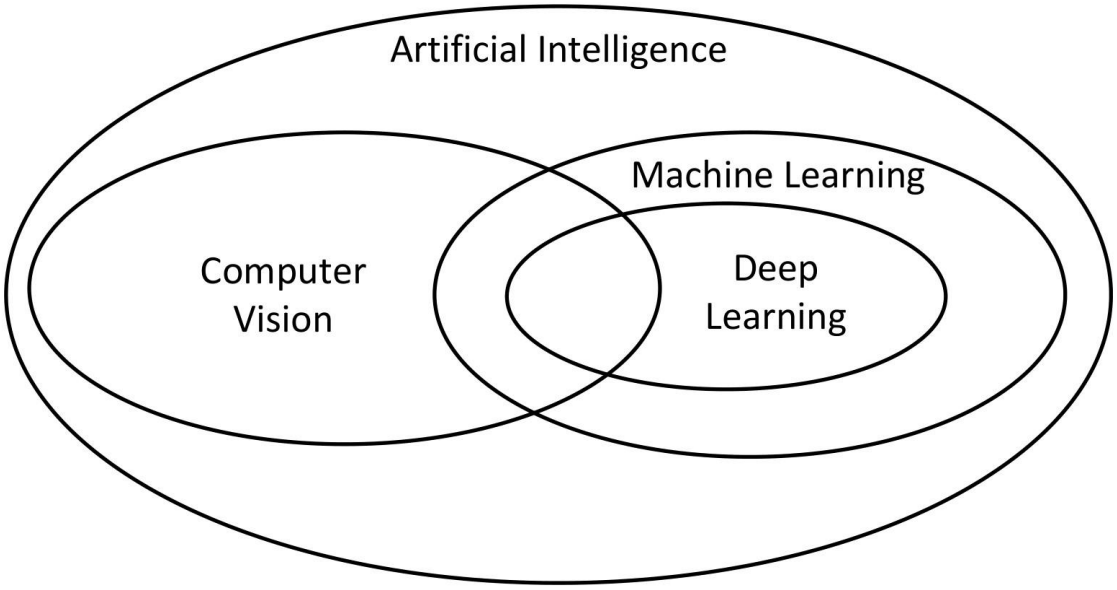


Figure 16: Relation between different fields

Convolutional Neural Networks are composed of two types of layers. The convolutional layer which performs all the computation required to extract meaning from the images and then project them into feature maps (vectors of numbers which make sense to a computer). These feature maps are then later used to further train CNN to classify and detect objects. The second layer is a subsampling layer, also called a pooling layer which down-samples the unnecessary information from the feature map. Deep learning practitioners have found that

stacking these two layers on top of each other creating a deep architecture improves the performance of the neural network significantly.

II. ResNet(2015)

In 2015, ResNet won the ILSVRC competition by surpassing the human error rate. The top-5 error rate achieved by humans was 5.1% while ResNet had an error rate of 3.1%. This was the first time that a Computer Vision algorithm had surpassed human level performance on object detection and classification. The implications of the result were clear. The CNN architecture combined with other deep learning approaches was the future of Computer Vision.

Training deep neural networks is very hard and a lot of trial and error has to be done to train a network efficiently. ResNet further improved the CNN architecture by employing 'residual blocks' which allowed the network to skip connections during training. This made training networks of even 100 and 1000 layers possible. In the case of deep learning, more layers means better training and test results. ResNet is still widely used as of today for object classification and detection.

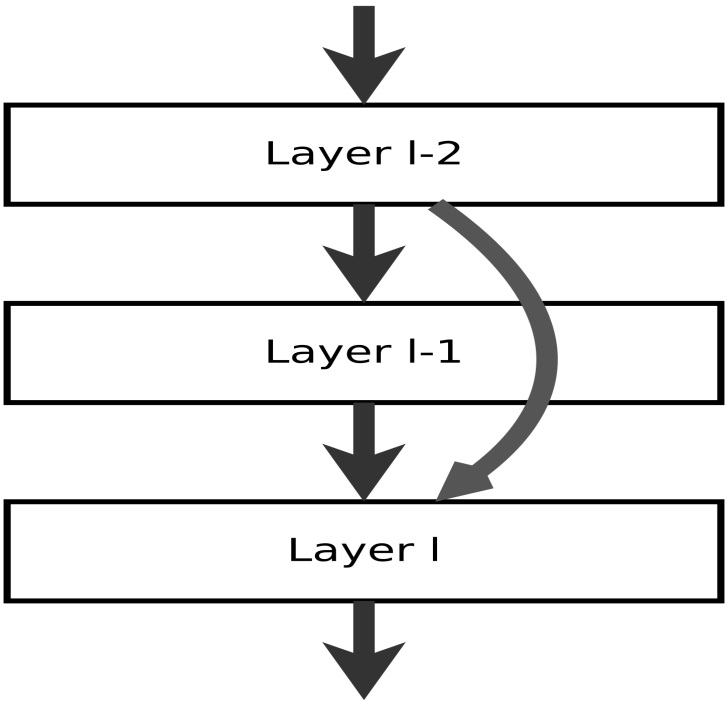


Figure 17: Residual Block employing skip connections.

Present and Future of Computer Vision

The future for Computer Vision looks promising with governments, universities and companies deploying tons of resources to conduct research and improve modern Computer Vision systems. Today, modern Computer Vision systems are able to perform a lot of tasks and in some cases even exceed human level performance.

Self Driving Cars: Computer vision has enabled us to make cars aware of their surroundings and make them smart enough to drive on their own. One of the main goals of self-driving cars is to eliminate the human factor during driving and to reduce accidents on the road. Self driving cars are statistically more safe and are less likely to get into accidents than humans. The technology is still not perfect today but companies like Tesla, Waymo and Amazon are constantly working on perfecting it.

Medical Imaging: Computer vision is used today to detect cancerous moles and tumours from MRI and CT scans. Such systems assist the doctors to give diagnosis and in some cases, even catch something missed by doctors.

Facial Recognition: Computer vision today plays an important role in keeping our workplaces and even our phones safe. Almost all phones use facial recognition of some sort to keep our data safe. Facial recognition is also used by law enforcement around the world to assist apprehend criminals .

The future Computer Vision system will pave the way for artificial intelligence around us. When these systems are used in conjunction with natural language processing models, we will be able to discern much more information about the surroundings from the images. More potent applications to assist humans will be created and the day is not far when Computer Vision will be considered a solved problem.

Summary

- Computer Vision is a difficult task
- Hubel and Wiesel experiment of cats lead to simple cells and complex cells (1959)
- Lawrence Robers thesis was about extracting 3D information from 2D images through lines and edges (1963)
- Marvin Minsky's MIT summer project lead to the field of Computer Vision(1966)
- Necognitrons architecture was inspired by the work of Hubel and Wiesel (1980)
- LeNet was the first commercially deployed CNN architecture trained using backpropagation algorithm. (1989)
- Eigenfaces were created using statistical analysis to perform facial recognition (1991)
- Viola-Jones real time object detection framework implemented in cameras to detect faces and objects.(2001)
- ImageNet competition (ILSVRC) was launched to solve and promote Computer Vision(2010)
- AlexNet won the ImageNet competition, beating the second place model with a 10% difference in top 5 error rate(2012)
- ResNet surpassed the human level performance in the ImageNet competition (2015)
- Modern CNN is used in self-driving cars, medical imaging and facial recognition