IMAGE CAPTIONING AND AUDIO FEEDBACK

Project report submitted in partial fulfillment of the requirement for the Degree of Bachelor of Technology

in

Computer Science and Engineering

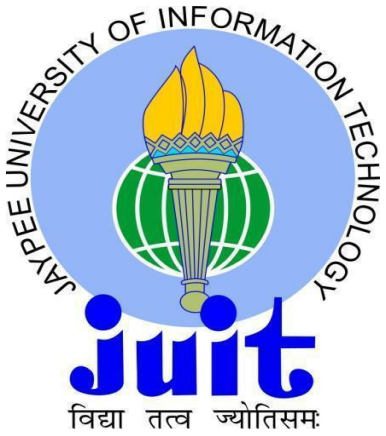
by

Hemant Chauhan
161337

Under the supervision of

Mrs Monika Bharti (**Dept**. **of CSE & IT**)

to



Department of Computer Science & Engineering and Information Technology

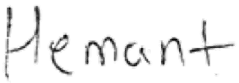
Jaypee University of Information Technology Waknaghat, Solan-173234, Himachal Pradesh

CERTIFICATE

Candidate's Declaration

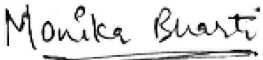
We hereby declare that the work presented in this report entitled **"**Performance Analysis of Deep Learning based approaches for "image captioning and audio feedback**"** in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering submitted in the department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology Waknaghat, Solan is an authentic record of our own work carried out over a period from August 2019 to December 2019 under the supervision of Mrs Monika Bharti **(**Assistant Professor, Department of CSE & IT**)**.

The matter embodied in the report has not been submitted for the award of any other degree or diploma**.**



Hemant Chauhan
161337

This is to certify that the above statement made by the candidate is true to the best of my knowledge**.**



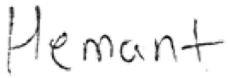
Supervisor Name**:** Mrs Monika Bharti
Designation**:** Assistant Professor
Department Name**:** Department of Computer Science Engineering & Information Technology
Dated**: ..../**Dec**/**2020

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success**.** We are grateful to my project guide Mrs Monika Bharti for the guidance, inspiration and constructive suggestions that helped us in the preparation of the project**.**

We would also like to thank our colleagues who helped us in successful completion of the project by having constructive discussions with us on how to implement the project**.**

Date**: ….**/12**/**2020



Hemant Chauhan
161337

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ML	Machine Learning
PCA	Principal Component Analysis
KNN	K-Nearest Neighbour
IP	Image Processing
AI	Artificial Intelligence
FB	Facebook
SVD	Singular Value Decomposition
APK	Application
INFO	Information
ALGO	Algorithm

LIST OF FIGURES

ABSTRACT

Consequently producing a characteristic language portrayal of a picture is an errand near the core of picture understanding. In this paper, we present a multi-model neural system technique firmly identified with the human visual framework that consequently figures out how to depict the substance of pictures. Our model comprises of two sub-models: an article discovery and limitation model, which separate the data of articles and their spatial relationship in pictures individually; Furthermore, a profound repetitive neural system (RNN) in view of long present moment memory (LSTM) units with consideration instrument for sentences age. Each expression of the portrayal will be consequently adjusted to various items of the info picture when it is produced. This is like the consideration component of the human visual framework. Test results on the Flickr dataset grandstand the value of the proposed technique, which outflank past benchmark models.

CHAPTER-1 INTRODUCTION

1.1 INTRODUCTION

Artificial Intelligence

AI is a wide-ranging aspect of software engineering dealing with the construction of savvy machines prepared to carry out commitments that typically involve human insight. Computer based thinking is a science field with various algorithm, but headways in AI and deep learning shift the outlook in each sector of the tech industry for all practical purposes.

Can machines think?

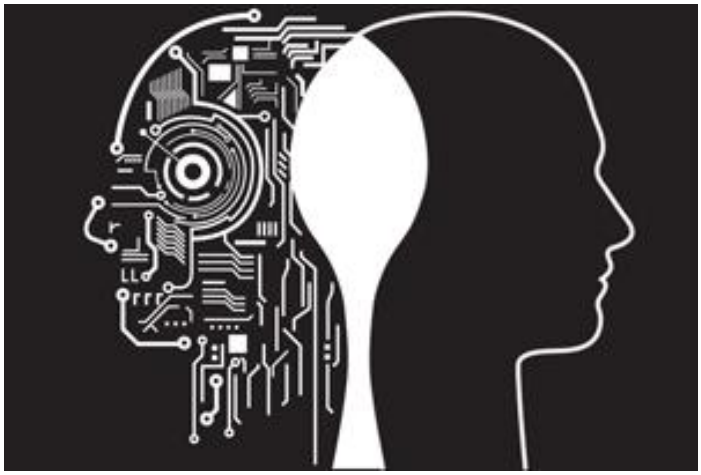


Fig 1.1: Can Machines think

Not even ten years after cracking the Nazi Enigma , the encrypting the system and helping the Army to win World War II and making the edge over everone , the mathematician Alan

Turing changed the history with just a simple inquiry in a second time: "Do really machines think"?

Allen's paper "Processing Machinery and Intelligence" (1950), and Turing Test as a result, established the central goal and perception of man-made brain capacity.

At its heart, AI is the branch of the software engineering that make plans to respond in the confirmed to Turing 's query. It is the responsibility that human intelligence be replicated or recreated in machines.

Man-made reasoning's broad goal has given ascend to various inquiries and debates. To such an degree that no single field value is known all over.

The significant impediment in making AI’s as essentially "building machines that are shrewd" is what that doesn't majorly clarify what man-made consciousness is? What makes a machine smart?

Artificial Intelligence: A Latest Approach in their crucial reading material, Stuart Russell and Peter Norvig encounter the inquiry by bridging their work around the topic of insightful machine operators. In view of this, AI is "the investigation of specialists that get percepts from the earth and perform activities."

How is AI be used?

AI commonly bogus under two general classifications:

Restricted AI: This sort of man-made intelligence, also referred to as "Frail AI," operates within a restricted environment and is a replication of human awareness. Slender AI frequently focuses on performing a solitary task very well and bear in mind that these devices can seem astute, They work within undeniably more demands and confines than even the most basic human insight.

Falsified General Intelligence (AGI): AGI, now and then referred to as "Solid AI," is the kind of logical reasoning by computers we found here in the movies, like to Westworld robots or Star Trek 's Data: The Next Generation. AGI is a particular-knowledge of computers and, like an individual, which can apply that insight to take care of any issue.
Thin Artificial Intelligence

Thin AI is covers us and is efficiently the good acknowledgment of man-made reasoning to date. With its emphasis on performing explicit assignments, Narrow AI has encountered various achievements in the most recent decade that have had "critical cultural advantages and have added to the financial essentialness of the country," as per "Planning for the Future of Artificial Intelligence," a 2016 report by the Obama Administration.

A couple of instances of Narrow AI include:

1.Google search

2.Picture acknowledgment programming

3.Alexa bot ,Siri and other individual collaborators

4.Auto-driving vehicles

5.IBM's Watson

Machine Learning

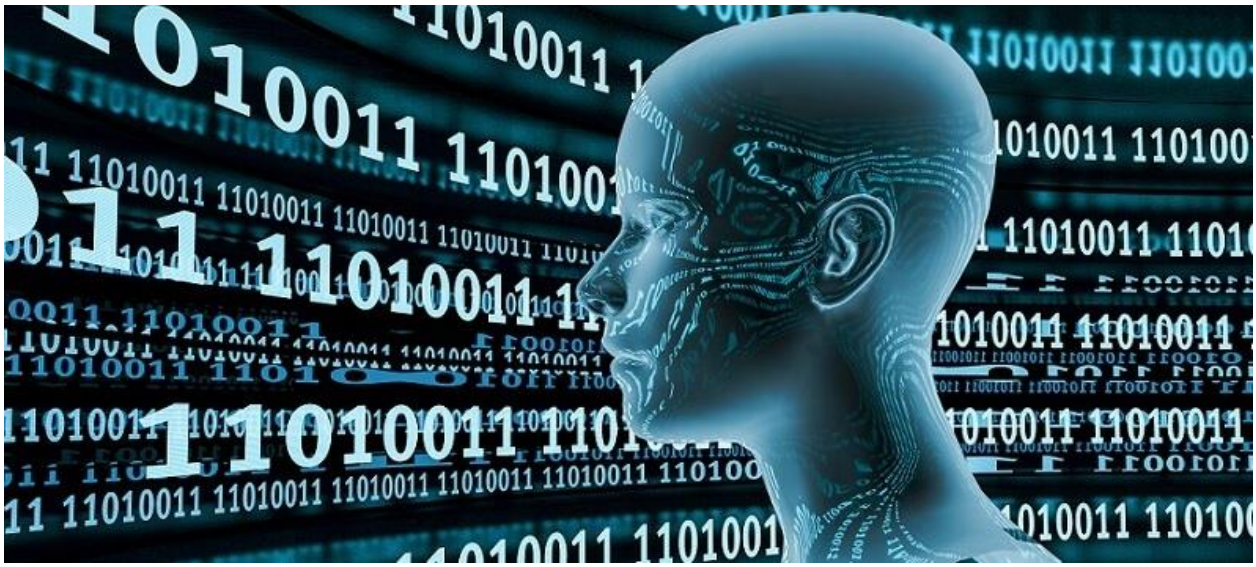


Fig 1.2: Machine learning

Machine learning is an area in artificial in (AI) that gives the ability to the system to understand by itself and enhance from things done earlier without it being programmed by programmer. Machine learning mainly covers on computer programs that navigate and use the information to analyze for themselves.

The learning cycle starts with insights or pattern , like examples, actual experiences or lessons, to search for clues in data and to make informed choices in the futher by relying on the case studies we have. The main objective is to allow machines to automatically learn without humanbeing interferance , and to modify behavior according to previous instances.

Some machine learning methods

Machine learning algorithms are divided into mainly two categories as supervised or unsupervised.

Using named examples to predict future outcomes, machine learning models can be used where past learning is used for new data. Beginning with the train and test dataset, the supervised learning creates a conditional function to make production values predictions. After adequate training from the dataset the system will provide expectations any very new input provided . The algorithm can also make comparison its output with original one , intended output provided and find errors to make further changes in the model accordingly to give a high performance.

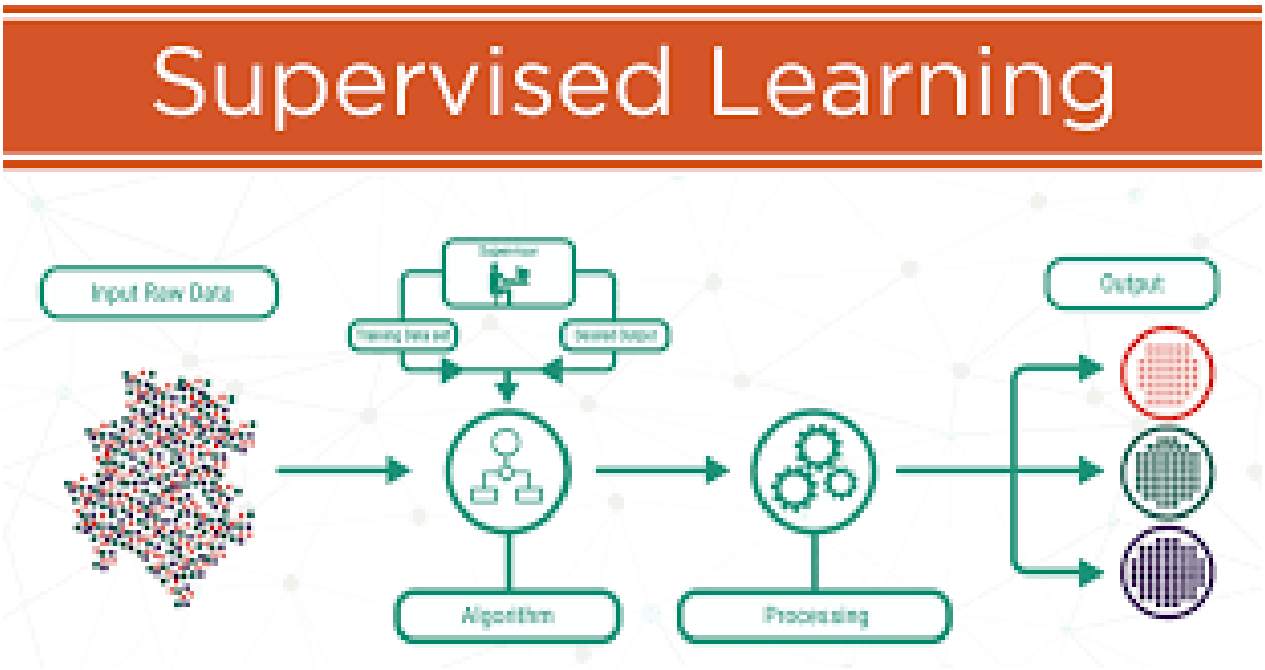


Fig 1.3: Supervised Learning

Unsupervised artificial intelligence algorithms, by contrast, are used where the knowledge used to train can not be identified or labelled. Learning without supervison explores how machines can analyze a function from unidentifiable data to find a secret structure. The machines does not work out the right performance, but it examines the information and can draw matches to explain hidden structures from datasets. structures from unlabeled data.

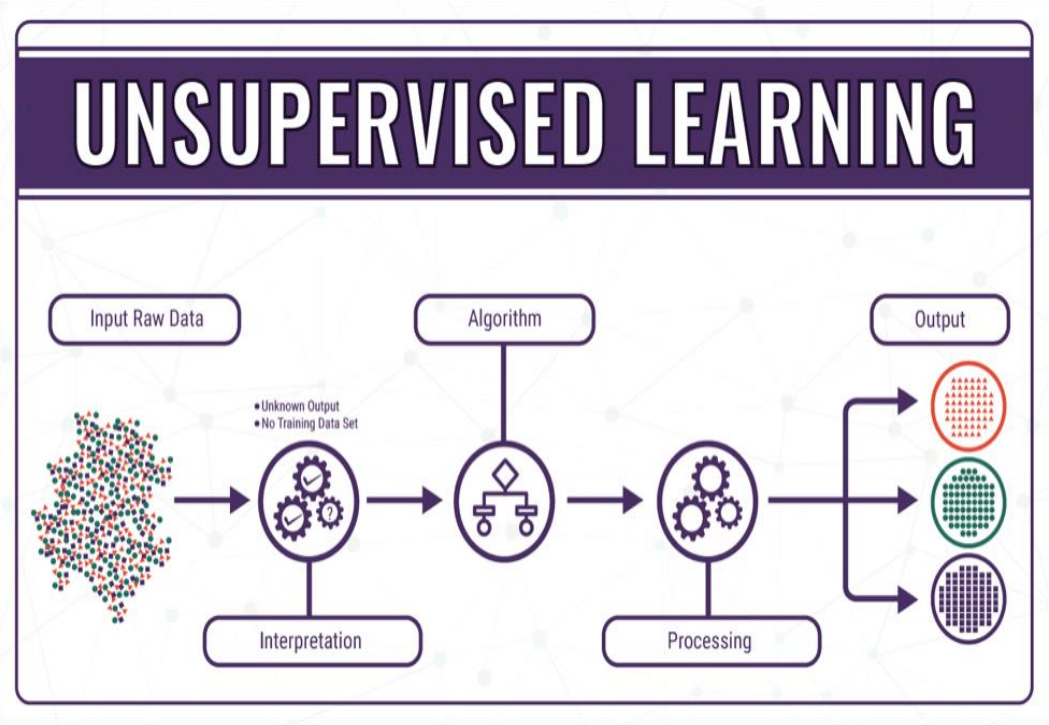


Fig 1.4: Unsupervised Learning

Machine learning helps us to analyze huge amounts of data. While usually providing quicker, more accurate outcomes to see the profitable opportunities or reduce hazardous risks, it can may take some more time and money to train efffectively. The combination of artificial intelligence with Cognitive computing technologies will make this process much faster and more successful

Supervised ML algorithms

1.Linear Regression:

Relapse issues are directed learning issues in which the reaction is ceaseless. Order issues are administered learning issues in which the reaction is straight out. Straight relapse is a method that is valuable for relapse issues.

So, why do we prefer linear regression?

- used largely
- fast responsive time
- easy usability
- It can be interpreted easily
- Acts as base for many other models

Simple Linear Regression Equation

$\hat{y} = b_0 + b_1 x$

Here

- $\hat{Y}$ = response
- x = feature
- b_0 = intercept
- b_1 = coefficient for x

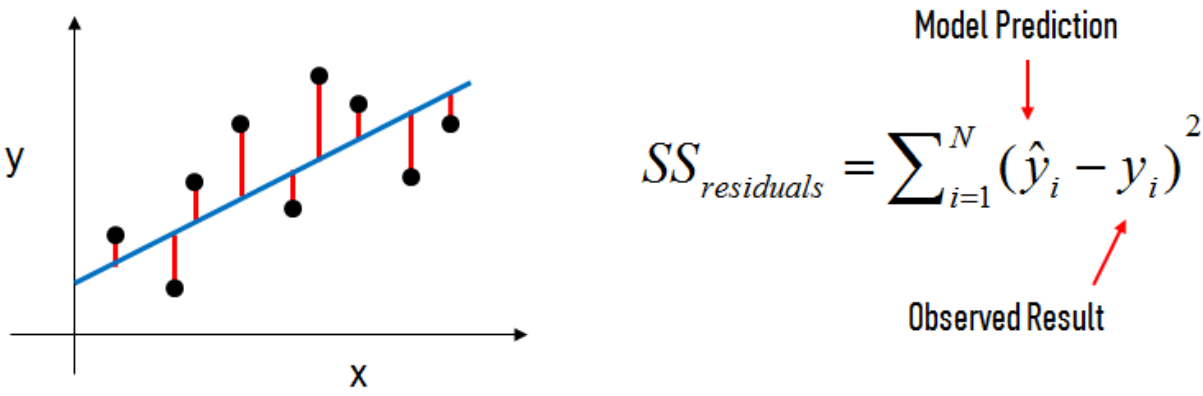


Fig 1.5:Linear Regression

2.Decision Tree

A decision tree is more of a real tree kind of chart with hubs speaking to where we pick a quality and pose an inquiry; edges speak to the appropriate responses the to the inquiry; and the bottom most without child nodes speak to the genuine yield or class mark. They are utilized in non-direct basic leadership with basic straight choice surface.

Decision trees organize the models by organizing them from the roots to a certain leaf center down the tree, with both the leaf hub offering the model a location. -- hub in the tree is conducted as an application for some property, and each edge falling out of that hub is compared to one of the most common responses to it. This process is calling itself again and again in nature and is rehabilitated for each stable subtree.

How about we represent this with assistance of a model. We should expect we need to take on badminton on a specific day — state Saturday — in what capacity will you choose whether should to play or not. Suppose you go and verify if it's sweltering or cold, check the speed of the breeze and mugginess, how the climate is, for example is it bright, shady, or

stormy. You consider every one of these components to choose in the event that you need to play or not.

The baseline methodology that is being in the decision trees is known as the algorithm ID3 (by Quinlan). The ID3 method constructs trees which take decision of it own by using a greedy, from top-down approach. Shortly, the algorithm levels have been:-Pick the best attribute for the NODE — Assign A as the decision attribute (test case). - For every value of node , make a new further of the NODE. – align the order of the training data to the appropriate further node leaf. - If given cases are accurately predicted, then don't do forward and else loop over the new further nodes.

The next major questions, now, is how to select the best characteristic. For ID3, we consider the best attributes in term of which attributes have the most knowledge benefit, a metric which communicates how nicely an attribute divides the data into classes based on predcition.

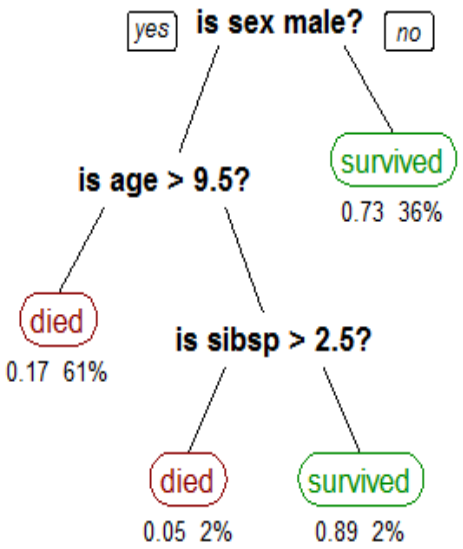


Fig 1.6:Probability

Cost Of Split:

We can reperesent Regresion as sumof(yactual — predictions)^2

We can represent Classifications as : G which is sumof(pk * (1 — pk)

Reasons to stop the splitting

You may request that when quit growing a tree? As an issue for the most part has an enormous arrangement of highlights, it brings about huge number of split, which thus gives a colossal tree. Such trees are mind boggling and can prompt overfitting. All in all, we have to realize when to stop? One method for doing this is to set a base number of preparing contributions to use on each leaf. For instance we can utilize at least 10 travelers to arrive at a decision(died or endure), and disregard any leaf that takes under 10 travelers. Another path is to set greatest profundity of your model. Greatest profundity alludes to the length of the longest way from a root to a leaf.

Pros of CART

- • Easy to understand, to translate and to image.
- Choice trees verifiably perform differentiable screening or highlight determination.
- It Can handle with both numerical and unmitigated information.
- • Option trees typically require little effort on the part of clients to organize information
- Nonlinear parameter relation don’t impact tree execution.

Cons of CART

- It can further create complex trees hence making the algorithm more complex
- They can be unstable as small difference can make the whole result difference
- Greedy algorithms can never assure us to provide the max over all interval best decision tree.

Naïve Bayes

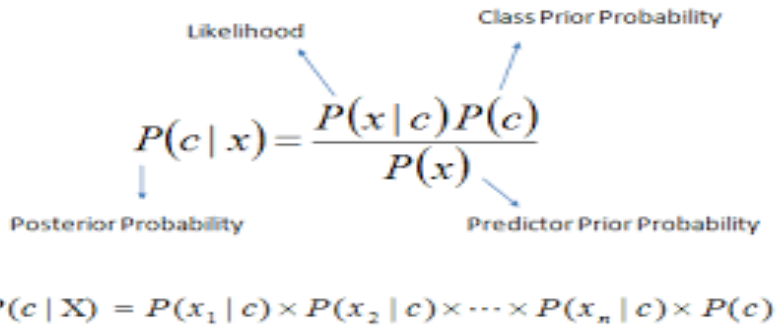


Fig 1.7:Naïve Bayes

It is an order system with a presumption of independence among indicators based on Bayes ' Theorem. In simple terms, a classifier from Naive Bayes assumes proximity of a particular value in a category is random to the proximity of some other element. For eg, an organic item

may be seen as an item on the off probability of being color red , around, and around 3 creeping in length across. If these indicators rely on each other or on the existence of specific highlights, these products freely add to the probability that this natural product is an apple and that's why it's called 'Gullible.'

Guileless Bayes algo is nothing but tough to fabricate and mainly valuable for huge data collections. Alongside smoothly, Naive Bayes is popular to beat even profoundly hard order techniques.

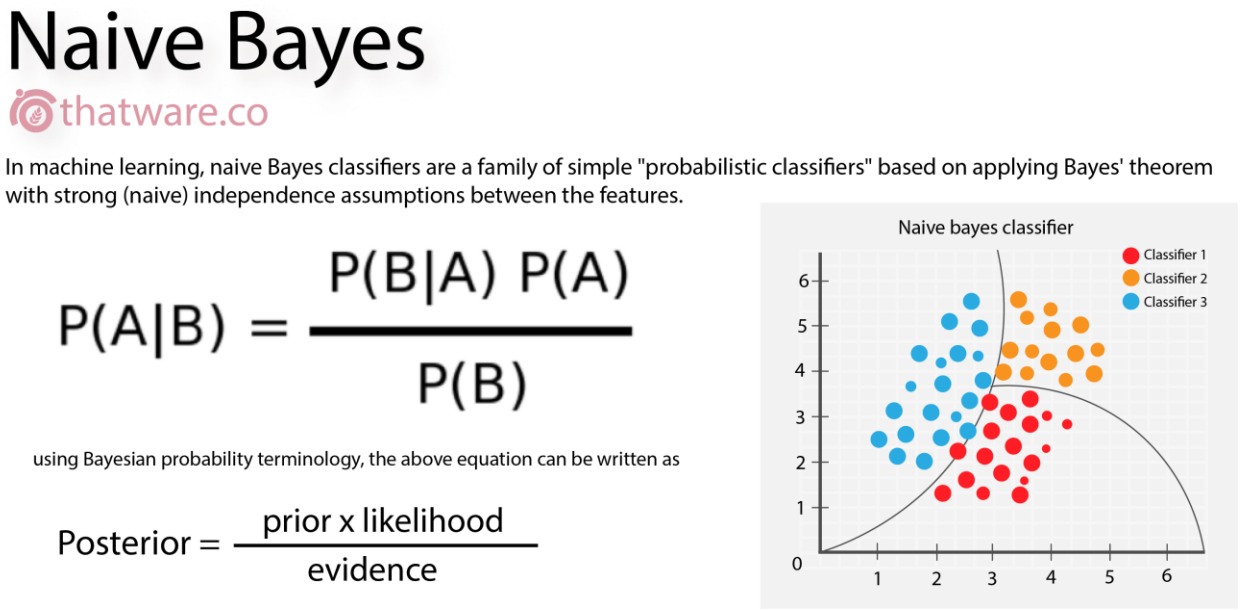


Fig 1.8:Naïve Bayes

Pros

- It is quick and simple to see category of test data collection. It furthermore perform well in multple class predcition They can be unstable as small difference can make the whole result difference

- • At the point where autonomy assumption remains, a Naive Bayes predictor compares good with specific algos such as measured relapse and you need less details to plan

- • It performs well if straight-out data variables contrasting with statistical variable(s) will occur. Typical theft is approved for numerical variable (minus bend, which is a solid suspicion).

Cons

- • If the clear cut parameter has a class (in the test data index) that was not included in the planning of the information index, a 0 (zero) probability will be defined at that point and a prediction will not be made. This is named "zero frequency" occasionally. We may use the smoothing method to understand that. One of the simplest smoothing devices.

- On the contrary , innocent Bayes is otherwise referred to as a bad estimator, and the chance give us output from predicted probability are therefore not to be taken into account.

- Other constraint of Naive Bayes is the presumption of automatic indicators. In actuality, it is in real world outlandish that we get a lot of indicators which are all autonomous.

Disadvantages of Machine Learning:

1. Time and assets-

Machine learning requires lot of assets to work. It might request extra computatuion power. ML requires enough chances to give the calculate a chance to learn and create to make acceptable their proposed reason with a many of accuracy and importance.

2. Data Acquisition

Artificial intelligence needs the training of large data sets, which should be inclusionary / impartial and of good quality. There may also be occasions when they will wait to produce new data.

3. Translation of results-

Precisely deciphering the outcomes produced by the calculations is a difficult errand. One needs to practice alert while picking calculations for their particular reason.

Advantages of Machine Learning:

1. Contiguous improvement-

Learning by the Machine algo improve in correctness and efficiency as they gain by previous attempts. This helps them make good decisions.

2. Distinguishes patterns and examples effectively-

AI is able to audit enormous volumes of data and find explicitly pa tterns and examples that people wouldn't see clearly. For example, for a business based on web site such as Flipkart, it helps to understand their customers' perusing habits and purchase chronicles to help them take into account the right products, plans, and notifications. It uses the possibility to not showing important promotions to them.

3. Automation-

Machine learning doesn't require human mediation. It enables machines to learn. It assists machines with making forecasts and improve the calculations independent from

anyone else. Against infection programming is a typical case of this as they consequently channel new dangers as and when they are perceived

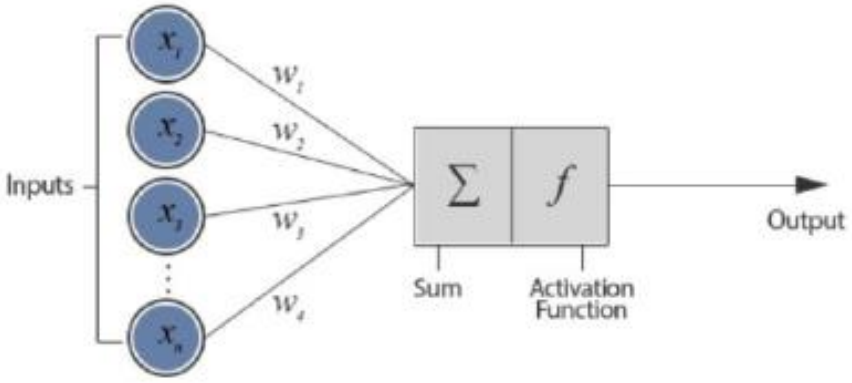


Fig 1.9:Deep Learning

Deep Learning

Deep learning is one of the set of machine learning that function to replicates the way of work of the human mind in handling info and making pattern for use in making the decisions. Deep learning is capable of learning not supervised from data that is not structured or not labeled. Also known as deep neural network.

How Deep Learning Works

In the digital age, deep learning has evolved tremendously, resulting in an blast of data in every forms and with each and every part of the universe. Known as big data, this

information is compiled from outlets such as social networks, search engine and digital cinemas. A enormous qunatity of info is frequently used and can be exchanged via fetch app such as computing on cloud.

The info, though, which is usually highly not structured, is so enormous which it may take years for human beings to grasp it and take out necessary info. Company understand the tremendous more possible that can arise from this abundance of knowledge being scrambled, and are gradually adapting for automated help to Artificial inteligent systems.

Deep Learning Versus Machine Learning

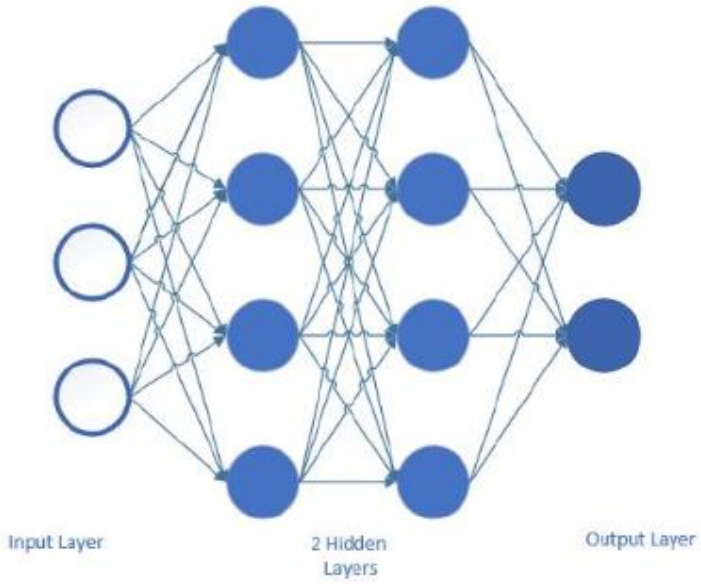


Figure 3.4: An Artificial Neural Network also known as a Multi Layer Perceptron

Fig 1.10: Multi Lever Perceptron

Artificial intelligence is one of the mostly used AI algo used for managing large data, a auto-adaptive algo that continues to evolve with experience or add more data.

To this end, if a online payment companies wanted to find the presence or probability of illegal work in its system, it would use machine learning technique. The network connectivity made into a computer algo will analyze all transfers that exist on the onlines network, identify data set patterns and notice out any deviation that the pattern detects.

Data science, a part of machine learning, follows a orderly level of artificial neural networks which perform machine learning processes. The CNNs are constructed like a human mind, with neurons nodes linked like a spidersWeb. While old-style systems build linear analyzes of the results. the hierarchical way of functioning of deep learning systems allows system to load info with a not so linear approach.

An old-style method to identifying illegal work or money filtering may depend on the volume of transfer that follows, whereas a nonlinear deep learnings idea would have included time, geographical locations, IP addresess, and some other function that would possibly point to fraudulent behavior. The very first layer of the algorithm is used to process a new data given like the transfers quantity and transfer it as output to the next step. The very next layer collects the information from the previous step by adding some extra info such as the IP address of the user and transfers the result on.

The next step includes data output from the second layer and uses the raw data, such as geographic location , making the design of the system even better. This is continuing through all neuron network stages.

CNN

A Convolutional NN is a Deep Learning algorithm that can take in an input data, assign location (learnable weight and bias) to different enitites in the image and distinguish one term

from some other. The pre-treatment required in a simple ConvolutionalNet is much smaller than other classification techniques. Although the filter are arm-engineered in ancient times, with appropriate training scale, ConvNets has the ability to learn these features.

A ConvolutionalNet's architecture is similar to that of Neurons' connectivity pattern working in the Human Mind, and was motivated by the Visuals Cortex's organizations. Sensory cells only react to changes in a restricted portion of the field of vision as the Receptive Field. A collection of such objects joins together to cover the entire area of vision.

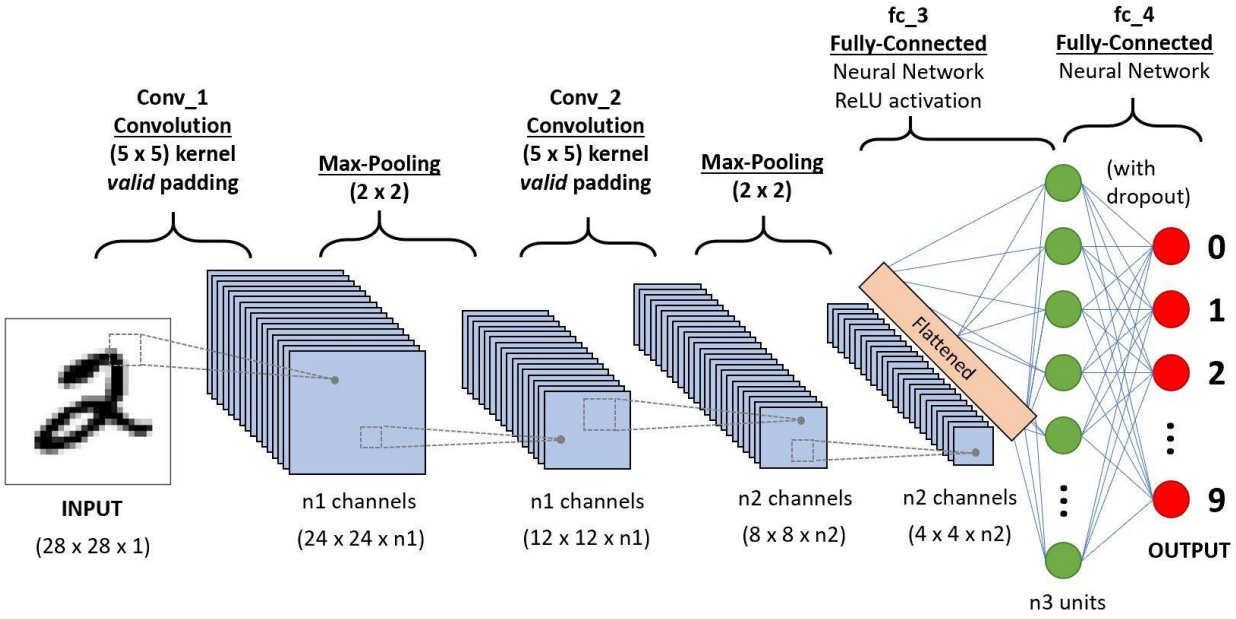


Fig 1.11:CNN

DenseNet

Recent research on Convilutionla Neural Network has showed that training can be significantly dense, more effective and more fastly if they include smaller links middle of structures near to the given data and those near to the output data. We accept this observation in this paper and introduce the Dense Convolutionary Network (DenseNet), which links each

step to each other in a feed-forward like . Although typical CNN with L layers have L links — one across every step and one step thereafter — our link has L(L+1)/2 deep links. The features maps of every previous layer are used as input by each layer and their features maps are used as input in every consequent step DenseNets have many excellent features: they alleviates the problem of the vanishing gradient, reinforce the propagations of features, encourages reuability of features and significantly minimise the amount of data used.

The counterintuitive effects of this complex linking pattern would be that it needs afew criteria than conventional CNN, because terminated feature maps need not be relearned. Current feed-forward systems can be seen as algo in a state transmitted from one layer to the next. Can step read the states from its former layers and write to the step next. This change the environment but still transfer the data that must be processed. ResNets[11] makes this info stored automatically through modification of an additives identity. Latest variation of ResNets shows that several layer makes very less contribution and be dumped at irregular intervals while train the data . This make the states of ResNets the same as (not rolled) RNN, but the number of ResNets parameters is considerably greater since every layers has its veryown weight. Our given architecture for DenseNet specifically varies between information applied to the network and files contained in it. DenseNet layer are very slim (e.g. 12 feature maps per layer), introducing only a limited number of features maps to the channel's "collective information" and making the remaining features maps useless — and the ultimate classifiers make a decision based on all of the channel's feature maps.

Besides improving variable performance, one major advantage of DenseNets is its changed transmission of communication and gradient across the link , which allows us to trains them easily. Every layer has a directs access from the loss function to the gradients and the original control signal, resulting in implicit deep guidance. This help in training of a deeper networks

architecture. Further, we also notice that dense connections have a normalizing effects, which minimizes over- fittings on task with small training set sizes.

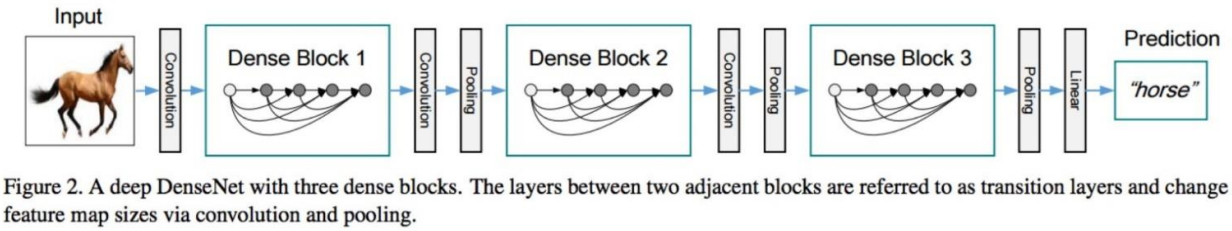


Figure 2. A deep DenseNet with three dense blocks. The layers between two adjacent blocks are referred to as transition layers and change feature map sizes via convolution and pooling.

Fig 1.12: DenseNet

Subsetting extracted features learned from several layer increase variance in the inputs of successive layer and improve their performance. That marks a major gap between DenseNets and ResNets. DenseNets are easier and more powerful compared with Inception channels [35, 36], which often concatenate features from different layers.

ResNet

According to the ultimate approximation principle, we knows that a feedforward link with only one layers is able to prove any function, provided enough power. The layer, however, could be enourmous and the networks would be inclined to overfit the info. There is therefore a shared fashions in the scientific community that our core links need to go deep.

The cutting-edge CNN software is moving deeper and deeper after AlexNet. The VGG networs and GoogleNets (also nam as Inception V1) has nineteen and twenty two layer respectively, while AlexNet had just 5 convolutionary layers.

But through the depth of the channel doesn't work by simply piling layers together. Due to the not so famous disappearing gradient problems, deep neural networks are hard to train — like the gradient is back-propagated to earlier given layers, autorepeated multiplications can made the gradient infinite very little. As the link deep downs, its quality becomes soaked or even starts to degrade rapidly.

Earlier of ResNet, there were many way to deal with the issues of the disappearing gradient, for example, [4] sums an additional losses in the middle layer as additional control, and none of them seems to solve the problem once and for all.

The main premise of using ResNet is to incorporate called "identification shortcut connection" that misses one or more strands, as shown in the figures below

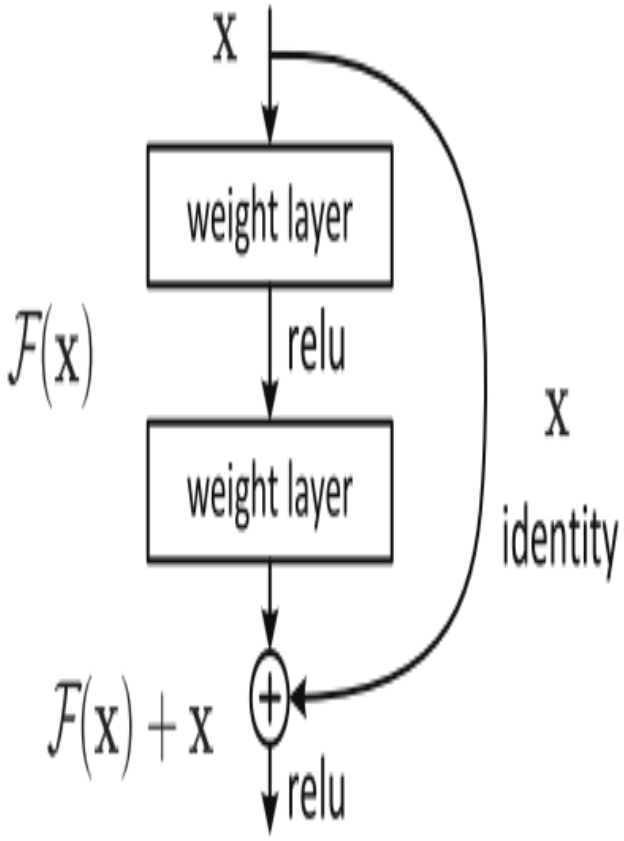


Fig 1.13:ResNet

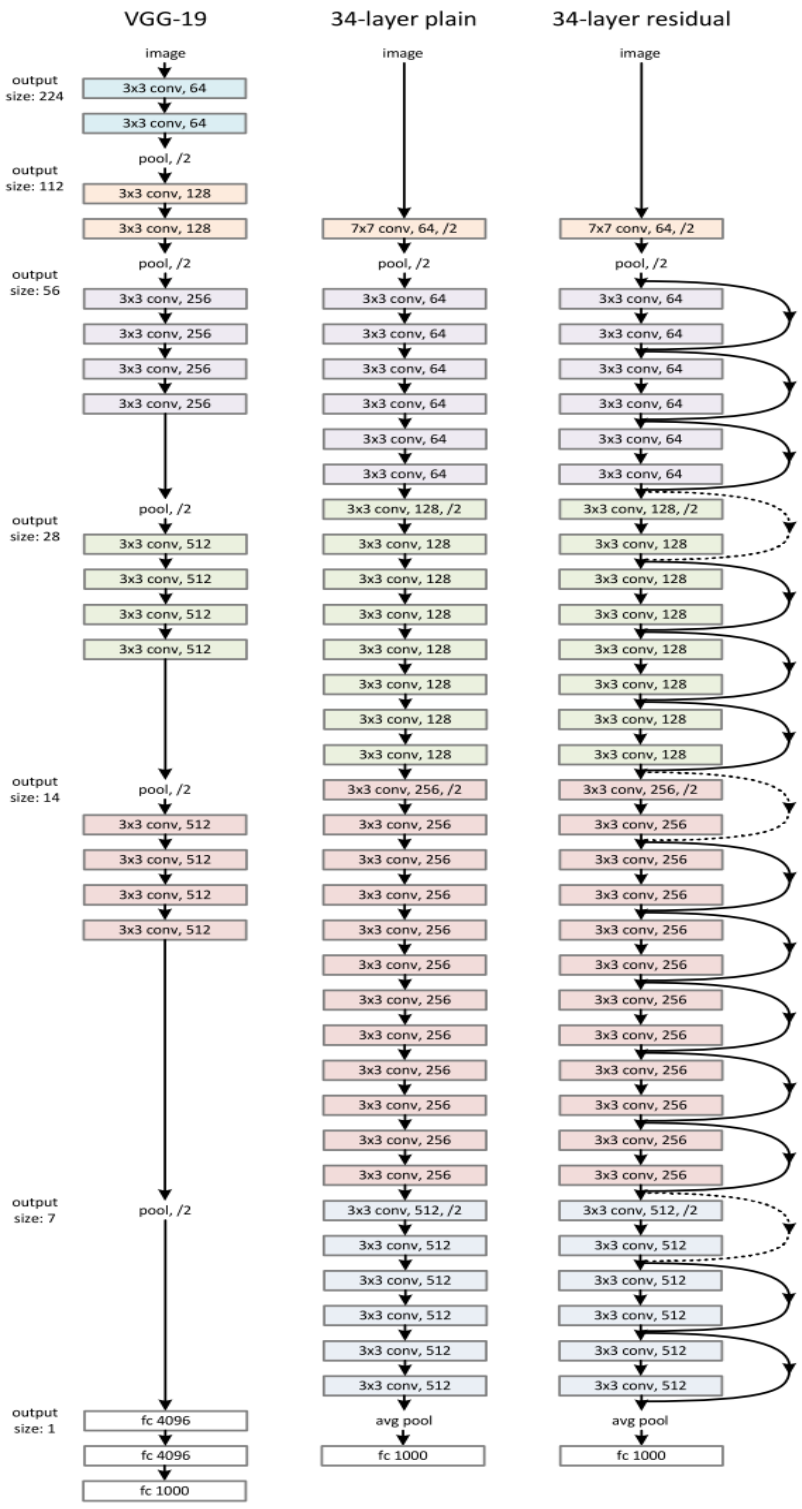


Fig 1.14: ResNet

The writers of [2] argue that piling layers will not reduce the performance of the network since we could actually stack and mapping identities (layer that does nothing) on the existing network and the resultant architecture will do the same thing. It means that a learning error greater than the shallower counterparts will not occur with the deeper model. They hypothesize makes it easier for the pieces to fit a residual mapping than to let them fit the required cognitive process directly. And the face legal above explicitly enables it to do exactly that.

VGG

- While prior AlexNet variants concentrated in the first convolutionary layer on a smaller filter size and strides, VGG tackles a really critical feature of CNNs: width. Let us go over the VGG architectural style:

- VGG takes in an RGB image of 224x224 pixels. The writers clipped out center 224x224 patch for each picture to keep the image size appropriate for the ImageNet contest.

- In VGG, the convolutionary layers use a very small receptive field (3x3, the smallest size still capturing left / right and up / down). There are different 1x1 convolution filters which act as linear awareness workshops, followed by theReLU unit. The convolution stage is set to 1 pixel so that after the convolutio the spatial resolution is conserved

• Completely ConnectedLayers. VGG has 3 final fully connected: the first two have 4096 channels each and third has 1000 layers, 1 channel per class

- CloakedLayers. All extracted features of VGG are using aReLU (a huge innovation from AlexNet that is cutting training time). Generally speaking, VGG donot uses Local Response regularization (LRN), since LRN increases memory requirements and training time without any specific increase in precision.

The Difference

VGG, while premised on AlexNet, has many differences which separate the testable theories from the others:

- Rather than using very large temporal information such as AlexNet (11x11 at 4 step) VGG uses very small temporal information (3x3 at 1 step). Because 3ReLU units are now in place of just one, the way things is more discriminatory. Also there are smaller criteria (27 times the level of channels instead of 49 times the number of channels that AlexNet has).

- VGG contains 1x1 convolution layers in order to make the decision more non-linear without changing the temporal information.

- The small-size convolution filters allow the VGG to get a wide number of layers of weight; more strands of courses lead to better efficiency. That's not a uncommon thing though. The 2014 ImageNet contest also featured GoogLeNet, another platform that uses the deep CNNs and limited convolution filters.

Fig 1.15: VGG

Advantages of Deep Learning:

1. Has top-level execution in matters that ultimately beat different schemes in different fields. This includes discussion, language, sense of direction, messing around like Go etc. However, this is not by a tad but by a vital amount.Diminishes the requirement for include designing, one of the most tedious pieces of AI practice.
2. Is a technology that can be adapted moderately effectively to new issues such as vision, time scheduling, language, etc., using procedures such as convolutionary neural systems, repeated neural systems, lengthy transitory memory, etc..

Disadvantages of Deep Learning:

1. It computationally extremely costly to plan. This takes a long time for the most impredictable models to plan using several computers fitted with expensive GPUs.
2. Choosing the topology / boost / prepare strategy / hyperparameters for deep learning is a dark manner of working without any hypothesis to guide you.
3. Rather computation time costly to plan. It takes a long time for the most impredictable prototypes to prepare using several devices fitted with costly GPUs.

Differences between Machine based Learning and Deep Learning

	Machine Learning	Deep Learning
Dependencies on data	Fantastic exhibitions on a little/medium dataset	Phenomenal execution on a
Hardware dependencies	Work on a low-end machine.	Requires ground-breaking r GPU: DL plays out a duplication
Feature engineering	Features must be understood first then the data	No need to understand th represents the data
Time of Execution	Can take from a few min to hrs	Can take upto weeks but Ne Network requires to process a important no of weight.
Interpretabe	There are a few algo are not so diffficult for interpretation and some of them are really hard to inderstand	hard to impossible

1.2 PROBLEM STATEMENT:

The issue presents a subtitling task, which requires a PC vision framework to both restrict and depict notable areas in pictures in normal language. The picture subtitling task sums up object location when the depictions comprise of a solitary word. Given a lot of pictures furthermore, earlier information about the substance locate the right semantic mark for the whole image(s).

Chapter 1 provides a basic introduction about the project so as to give the basic idea and details regarding what we are going to do and also to familiarize you with the technical and few of the necessary theoretical aspects.

Chapter 2 includes the Literature survey i.e. review from different journals, research papers etc.

Chapter 3 aims at the system design, the techniques and the different tools needed for the project.

Chapter 4 tells us about the project performance analysis

Chapter 5 provides the conclusion, and also tells about scopes in the future for the same

Input: A image

Expected Output: Description of image in the form of subtitles and audio

Dataset: Flickr8k(8000 images)

References:

1.IITM Cse department TPA.

2.Towards Science Blogs

3.Data School

4.Medium Blogs

Libraries Used:

- Keras 1.2.2

- Tensorflow 0.12.1
- tqdm
- numpy
- pandas
- matplotlib
- pickle
- PIL
- Glob
- Google Text to speech API
- Glove

1. 3 OBJECTIVES :

The objectives of the project is to get the output audio of the picture that we have provided.It is basically a image captioning project with a audio feedback for which we have used a Google text to audio API which will give us an output audio mp3 file with the resultant audio. We have used Transfer learning for this as we used a pre –trained model Inception V3 which is a full-convolutional, 48-layer-deep neural network. You can load a pretrained version of the trained network from the ImageNet database on more than one million images. The pre-trained network is capable of classifying images into thousands of classes of objects, such as car , mouse, pen and several creatures.

1. 4 METHODOLOGY:

We will Use Flickr 8K dataset which comprises fo 8000 images and will divide that into 2 parts that are Training Set and Testing Set. Training Set will be dataset on which we gonna

train our model which comprises of 6000 random images while testing set comprises of 2000 images on which we going to test our model . Then we will going to clean out whole training and testing data with the help of some functions. Images are a mere input (X) to our algo. As you might already know, every input to a algo must be vector-shaped. The transformation of every image into a vectors of a Constant size which can then be fed into the NN as input. To this end, we use the InceptionV3 model (Convolutional Neural Network) developed by google Research to opt for transfer learning.

Data Preprocessing Captions

Subtitles are something we wish to predict. So subtitles will be the targeted the parameters (Y) which the algo learns to predicts as during training period.

But the the whole caption 's estimate, given the image, doesn't happen immediately. Word by word, we'll guess the caption. So we need to encode all word into a vector of a constant size. That part should be seen later, even so, when we look at it

Data preprocessing using generators

That is amongst the most significant elements in the study of the whole case. Here we will explain how the info can be configured in a way which is suitable to be provided as input to the deep neural network.

Data generators are a feature that is native to Python. The ImageDataGenerator category given by the Keras API is nothing other than a generator function implemented in Python.

Model

Here comes out model that we have built which will ultimately going to predict output for us.

CHAPTER-2 LITERATURE SURVEY

Convolutionary Neural Networks (CNN) are naturally enhanced Multi Layered Perceptron variations. It consists of at least each convolution layer (regularly with a sub-sampling step) and is then decided to pursue as in a normal multilayered neural system by at least one fully associated layer. A CNN's engineering is designed to exploited the 2D structured of an information image (or, for instance, a discoursed signals, other 2D data). This is achieved with neighborhood groups and bundled heaps pursuing by some sort of pool techniques that brings invariant highlights to analysis. A further benefit of CNNs is that they are easier to plan and have far fewer parameter than fully connected structures with a comparable number of covered units. CNNs have been commonly used and read for image errands and now are cutting-edge systems for object recognition and exploring Repetitive Neural Networks (RNNs) in numerous NLP undertakings. RNNs are called formulaic in view of the fact that for each element of a grouping they perform a similar odd job, with the yield being based on past computations. Then again, RNNs can be considered as "memory" systems that catch up to this point data on what has been defined. Hypothesis is that RNNs can use data in highly subjective long sequences, Nonetheless, they are limited by and by looking back just a few steps. Our model's goal is to construct subtitles or picture representations naturally. Various conferences have done explorations in the past, finding a place in both business and the academic world that looks something like or is in the light of a theme like what we do. From these explorations various parts of our model take reference. The investigation papers we used are referred to separately in the list of sources, similar to those provided by D. Narayanswamy et al[1] which aims to establish marks characterizing the outlines of the film, or that of DElliott, F. Kellers, Photo depiction using portrayals of visual reliance, whereby the creators strive to identify the various constituents of a image. Anyway, the research discussed above is generally concerned and focuses more on using image handling to classify what's more, to differentiate separate books in a image . They never manages different setting

of these articles. Such discovery publications have enables us to understand incredibility the concept of preparing and dividing pictures which let a significant jobs in my framework. Even we seek to achieve the picture's rational representation. In the aforementioned area of knowing the perfecting of picture in a variety of areas, especially in the centre sections of market and the academic world, a lot of present work is also being completed. The new business as usual involves wide-ranging inquiries about Pc programming and NLP based bunches , The ones we've discussed are those at Stanford A. Karpathy, Li Fei and UT, Austin, which are both researching goals to construct photo inscriptions. Cortana is an aspect of artificial intelligence created by Microsoft (Cortana) apart from the scholarly population 's striking modern improvements. , As of their current Microsoft Build meeting, they are organizing the AI in their Bing pursuit management, which will allow the customer to communicate with the User Interface more naturally. The most feasible operation of image preparation and subtitles was by Reddits (the Reverse Image Search), which allowed a customer to move an image and to break down their estimate. and shows images with a compasioned settings, this was trailing by Google’s, despite the facts are that it is to be Seen that both these tasks are still in training stage also, being worked on. Another eminent use of CV by Facebook Team (Image Tag). Through our prototype we aim to provide syntactically correct and outwardly based representation of theoretical artifacts, the given depiction of which will be represented in natural language, e.g. humans viewed. By using systems such as CNN, RNN and data indexes, such as those of Flickr, we are striving to get a human-level scenario of the images in question. Our main knowledge is that we can use these enormous picture sentences datasets by looking at the phrases as fragile names, where synonymous pieces of word refer to some particular yet obscure region of the scene. Our methodology is to infer and use these arrangements to familiarize ourselves with a training algorithm of representations We are building a fundamental model of the neural system which derives the structure between segments of sentences and the district of the image they represent. We present a recurrent neural technology that takes a image of the data and generates its material.

CHAPTER-3:SYSTEM DESIGN

In this part we would discuss the means that we are going to take while develop our framework. Subsequent to experiencing distinctive research papers we intended to go fora non-intrusive method for developing our framework. As we all know that there are lots of different object detection and image captioning strategies are available but we go for the inceptionV3 and beam search technique for designing our system and we have gone for such a strategy because it is increasingly solid and relevant under various circumstances.

We design a system which depends on the following two stages:

1. Object Detection

2. Captioning the image and Audio Feedback

Firstly we will create a dictionary of captions and then we will calculate the unique words. Then we will train our dataset by passing the images and passing the caption. We are using Flicker8K dataset. In Image Captioning, a CNN is utilized to separate the highlights from a picture which is then alongside the inscriptions is nourished into a RNN. To separate the highlights, we utilize a model prepared on Imagenet. I gave a shot VGG-16, Resnet-50 and InceptionV3. Vgg16 has very nearly 134 million parameters and its main 5 mistake on Imagenet is 7.3%. InceptionV3 has 21 million parameters and its main 5 blunder on Imagenet is 3.46%. Human top-5 mistake on Imagenet is 5.1%.

Preprocesing the images using InceptionV3

Beginning V3 is a generally used picture acknowledgment algo appears to be accomplished many more than seventy eight percent exactness on the ImageNet set of Data. The algo is the climax of numerous thoughts created by various specialists throughout the years.

The design itself consists of synchronous and awry building squares like convolutions, regular max pooling ,pooling, concats and fully linked layers. Batch norms is commonly used in the system, and applied to inputs for actuation. Misfortune is handled using Softmax.

The coding which provides 3 core binaries for:

1. Learning an Inception v3 model from begining over different devices and different devices using the training set for the ImageNet 2012 Competition.

2. Building an Inception v3 network with a single across different devices and/or multiple computers that use Training data set for the ImageNet 2012 Competition.

3. Reassign an Inception v3 network on a new assignment and back-propagate the mistakes to fine-tune the weights and biases.

A sequential stochastic gradient descent through several GPUs is used for the preparation. The user can specify how many GPUs they want to utilize. The sequential train of data conducts batch slicing by converting in number of batches over several GPUs

The main knowledge identifies with layer tasks. In a conventional conv net, each layer extricates data from the past layer so as to change the info information into an increasingly helpful portrayal. Be that as it may, each layer type removes an alternate sort of data. The yield of a 5x5 convolutional part discloses to us something other than what's expected from the yield of a 3x3 convolutional bit, which reveals to us something else from the yield of a

maximum pooling piece, etc, etc. At some random layer, how would we know what change gives the most "valuable" data?

An Inception module processes numerous various changes over a similar information map in parallel, linking their outcomes into a solitary yield. As it were, for each layer, Inception does a 5x5 convolutional change, and a 3x3, and a maximum pool. What's more, the following layer of the model gets the chance to choose if (and how) to utilize each snippet of data.

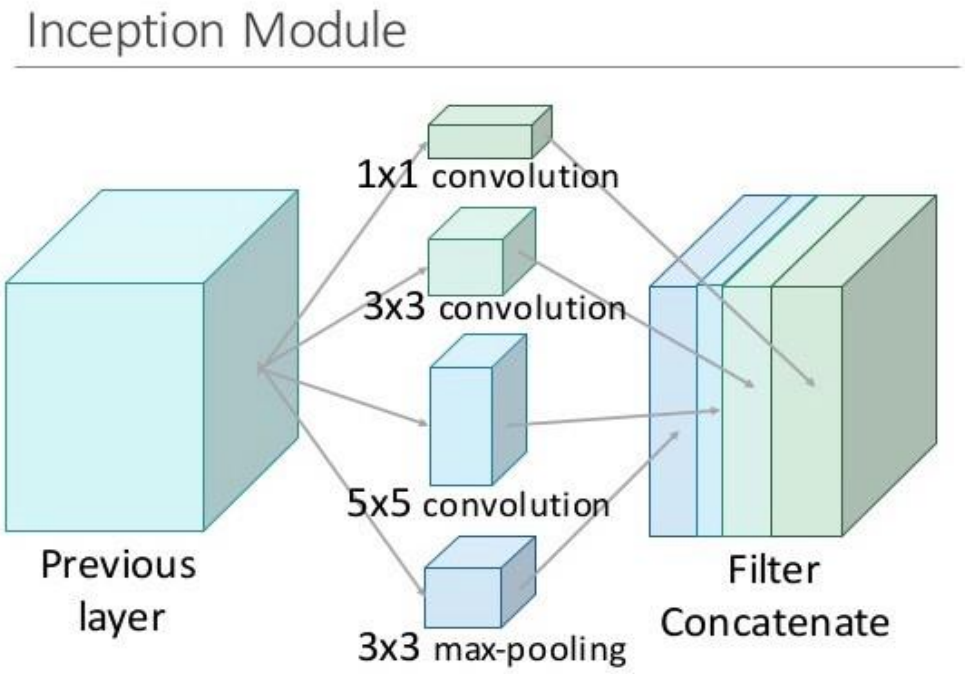


Fig3.1 :Inception Module

The expanded data thickness of this model engineering accompanies one glaring issue: we've radically expanded computational expenses. Not exclusively are enormous (for example 5x5) convolutional channels characteristically costly to process, stacking numerous various channels one next to the other significantly expands the quantity of highlight maps per layer. Furthermore, this expansion turns into a fatal bottleneck in our model. Consider it along these lines. For each extra channel included, we need to convolve over all the info maps to

compute a solitary yield. See the picture underneath: making one yield map from a solitary channel includes figuring over each and every guide from the past layer.

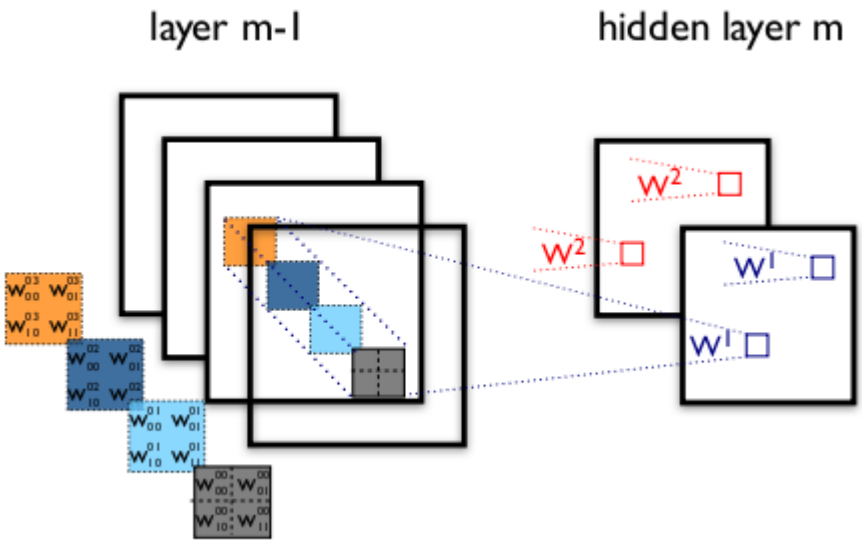
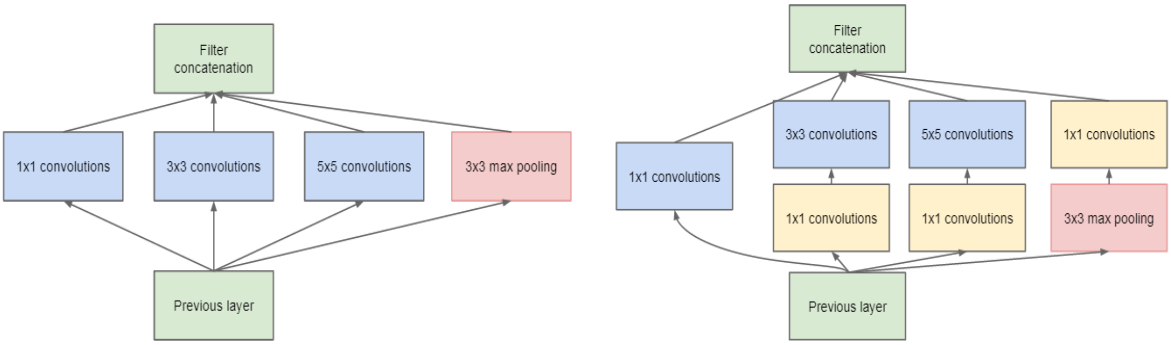


Fig3.2 : Layer structure of Inception Module

suppose there are M input maps. One extra channel implies convolving over M more maps; N extra channels implies convolving over N*M more maps. As it were, as the creators note, "any uniform increment in the quantity of [filters] brings about a quadratic increment of calculation." Our credulous Inception module just significantly increased or quadrupled the quantity of channels. Computationally, this is a Big Bad Thing.

This prompts knowledge #2: utilizing 1x1 convolutions to perform dimensionality decrease. So as to illuminate the computational bottleneck, the creators of Inception utilized 1x1 convolutions to "channel" the profundity of the yields. A 1x1 convolution just sees each an incentive in turn, however over numerous channels, it can remove spatial data and pack it down to a lower measurement. For instance, utilizing 20 1x1 channels, a contribution of size

64x64x100 (with 100 component maps) can be compacted down to 64x64x20. By diminishing the quantity of information maps, the creators of Inception had the option to stack diverse layer changes in parallel, bringing about nets that were at the same time profound (numerous layers) and "wide" (many parallel tasks).



(a) Inception module, naïve version (b) Inception module with dimension reductions

Fig 3.3: Naïve Version and Dimention reductions

Caption generation is a demanding artificially intelligent issue where a textual explanation for a given photograph must be produced.

It requires the two strategies from PC vision to compr/ehend the substance of the picture and a language based algo from the area of characteristic language handling to transform the comprehension of the picture into words organized appropriately. As of late, profound learning strategies have accomplished cutting edge results on instances of this issue.

Profound learning strategies have shown cutting edge results on inscription age issues. What is generally great about these techniques is a solitary start to finish model can be characterized to foresee a subtitle, given a photograph, rather than requiring complex information arrangement or a pipeline of explicitly planned models.

Image Captioning

Image captioning is a famous Artificial Intelligence (AI) area of research that deals with the knowledge of images and a language summary for that image. The perception of images requires to identify objects and recognize them. It also has to recognize the requirement or location of the picture, the properties of the objects and their encounters. Generating much-formed sentences requires a syntactic as well as a linguistic comprehension of the language.. Knowing an image depends primarily on having features of the image. The techniques employed for this reason can be classified into two different classes:

(1) New methods, machine learning algo based and

(2) Deep data science detection algorithm. handmade feature such as Locals Binary Patterns (LBP) in conventional machine learning

Scaled Invariants Feature Transform (SIFT), Oriented Gradient Histogram (HOG), and a variation of these characteristics are commonly used. Attributes are extracted in such techniques From data to entry. To identify an object, they are then moved on to a classifier such as Support Vector Machines (SVM). Because handcrafted features are unique to the mission, they extract characteristics from a large and diverse collection. Also, practical world info such as image and videos are complexed and have many semantical interpretation.While, in deeply machine learning based algo techniques, feature are learned automatically from data training and they can take care a enourmous and diversified set of image and video. For instance, Convolutional Neural Networks (CNN) are commonly used for the learning of features, and identification using a classifiers such as Relu. For producing captions, CNN is usually accompanied by Recurrent Neural Networks (RNN). A huge

number of posts on text categorization have been authored in the last five years, to deep machine teaching which are more popular and used . Deep learning algo can manage the image captioning complexities and problems reasonably well. Only three study papers on this research topic have been published thus far. And although publications provided a nice literature surveys of image annotations, they can cover a some paper on deep learning as many of them were authored after the surveyed paper . Such surveyed paper primarily covered based on model, replication-based, and creating models from really some deep learning-based, novel image captions. Alarge numbers of research on deep learning-based text categorization have been completed, however. The presence of huge and raw dataset has also make the learning based on image transcribing an important topic of study . To include an english translation of the literaturary part , we show surveys mainly on image captioning based on deep learning texts. The main contribution of this study is to show an exhaustive surveys of deep learning for captioning of images. Next, the current captioning of images papers are grouped into 3 main categories:

(1)Image captioning based on Template

(2)Images captioning based on Retrieval

(3)Images caption baased on Novel generations.

The classifications are briefly discussed in Section 2. Many methods of deep learning based images annotations fall under the classes of generations of creative captions. Thus, with machine learning we focus only on the generation of novel captions. Secondly, we combine the approaches of deep learning text categorization into different groups, namely

(1) Based on the space visuals,

(2) Based on multimodals,

(3) Supervised learning

, (4) Other deep learning,

(5) Based on Densed Captioning,

(6) Based on all Scenebase,

(7) Based on Encoding Decoding Architecture

(8) Based on Composition Architecture,

(9)Model based on LSTM(Long Short-Term Memory),

(10) Others language model-based,

(11) Attention-Based,

(12) Based on Semantic,

(13) Stylized captions

CNN

CNN image order took, process and community an data image under different groups (e.g., Dog, Cat, Tiger, Lion). PCs find a picture to be a pixel display and depend on the objectives of the image. It will see height x width x depth(h = Height, w = Width, d = aspect in view of the photo goals). E.g., an pic of 3 x 6 x 6 RGB grid clusters (Three parameter to RGB values) and a pic of 1 x 4 x 4 shows of network of pic of grayscale.

Actually, in-depth learning of CNN algo to prep and check, each knowledge images will go through a continuum of convolution layers with streams (Kernal), fully, pooling, associated layers (FC) and applying Relu capability to characterize an object with stochastic quality somewhere within the range 0 and 1 . The figure below is a completed progression of CNN processing a picture of information and grouping the reports based on values.

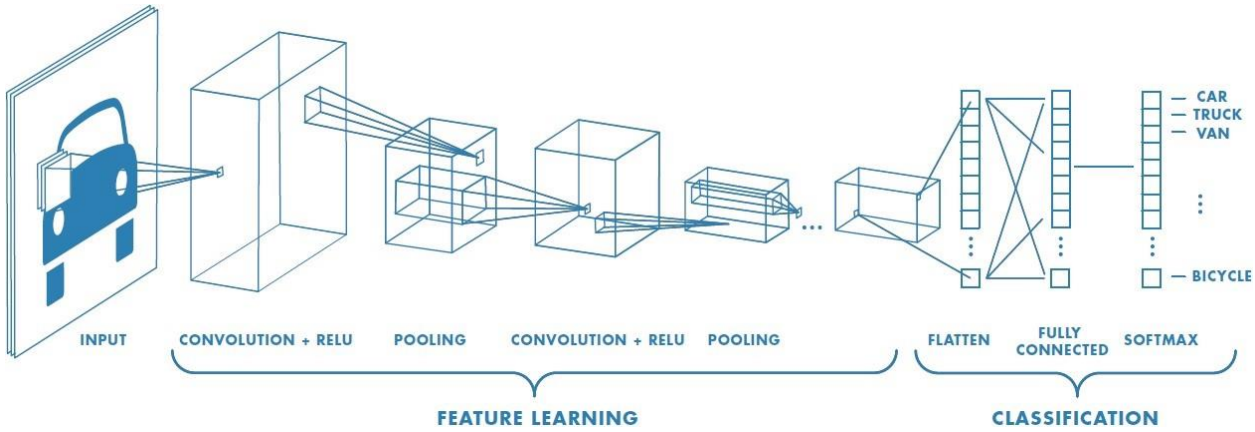


FIG 3.4:Neural network with convolution layer

Glove

GloVe is an unsupervised machine learning algorithm who use words to obtain feature vectors. Implementation is carried out from a corpus on consolidated global phrase-word co-occurrence statistics, and the consequent interpretations show fascinating linear shapes of the word vector space.

The word occurrence statistics in a corpus are A primary sour of info for everyone

Uncontrolled methods of computational methods of verbs, and while many such techniquesnow exist, the question arises how meaningful From such statistics is produced, and how the corresponding word vectors might reflect that meaning.

We shine new light on that question in this section.

Test Plan

This topic displays the execution subtle for the proposed system created for the Image Captioning is discussed here.

The software used in the execution of the project its platform bundles and so on are described as follows

Requirement

Platform:

Widows 7 and above/Max OS 10 or above

Software Used:

Python/Jupyter

Implementation Details

The implementation details are carried out in 3 stages

1.Input Name

We will input the image to the system with various alignments with different illumination conditions and use those image to train the system so that the system is able to recognize different objects

2.Feature Extraction:

Then inceptionV3 will extract features for us of different objects in the picture.

3.Building the model:

Relevant models is being built in order to execute the code

4.Prediction:

Prediction in the form of captions is being displayed and audio is being generated

Conclusion

In this chapter we examined our proposed system in details and how the whole system related with Image Captioning is implemented.We likewise talk about the Algorithm implemented in various periods of image captioning and the means which we will pursue while building our system.

CHAPTER- 4 PERFORMANCE ANALYSIS

While checking out the data points if we take only 2 images it will lead to some number of datapoints let say 15. But in reality dataset we have around 6000 images each having five captions in it , which sums up to 30000 images and caption

If we assume that every caption on a average is 8 . which will ultimately lead to a total of 30000 * 8 = 240000 data points and the size of matrix will going to be so large assuming n as number of data points (240000) , m as size of every data point which is size of the image vector + size of partial caption(x) which combines to be 204b + x. Each term (or index) is mapped to high dimensional space by means of one the techniques. Later, we'll see how every character is being map to a 200 lengthed vector with the use of pre trained mode of GLOVE words embedding while the algo building stage. each series comprises 34 indexes , in which every positon is a 200-length vector. Here x = 34 * 200 = 6800
Consequently m = 2048 + 6800 = 8848.
Last, matrix size= 8848* 240000 = 1857080000 blocks.
even if we presume that one frame takes 2 bytes, then we would need upwards of 3 GB of main memory that stores this data matrix.
That is quite an enormous requirement and even if we able to bring data to the main memory it will somehow make our system laggy .Hence to solve this problem we used generators.
Our model was predicting almost correct caption there are a few images where our model came up with absurd captions.This is because InceptionV3 have trained on a vast variety of objects but there are a few which have still left and its practically impossible to train on every image its just it got better with time as we keep on providing images to it and keep on developing it . Another next reason is som images pixels are so mixed up that it didn’t even differentiate between the two objects hence coming up with different captions.

CHAPTER- 5 CONCLUSION AND FUTURE WORK

5 . 1 OVERVIEW

In this last chapter of our report we might want to finish up our work and talk about the work we are going to execute in the near future. So far we have we have perused various research papers which really talked about the different face image captioning procedures which has implemented up until now. We even did the comparison of various image captioning algorithms with their upsides and downsides which helped us to pick a powerful algorithm which could withstand different drawbacks .We even examined about our framework on which we are going to do the testing procedures. The project aim is basic however a viable method for recognition maintaining a strategic distance from pointless complexities that may hamper to genuine execution. We even talked about the robustness of different algorithms with their memory prerequisites and their handling time. In the whole project work we were centered around extracting highlights from pictures and subsequent to considering different research papers we chose to build up a customary, basic yet an effective algorithm for gender recognition. We can land to this point after the end of this report higher discovery rates are conceivable.

5.2 FURTURE SCOPE

This is a basic simple solution and a lot of modidfications can be made in this like we could have use large data set , architechture could be changed , we could have played with hyper parameters more (like batch size , learning rate ,number of units , dropout rate, etc) , could have used CV set for overfitting , Instead of greedy search we could have used Beam search which traverse through relevant features only , using some other scoring factors It can be further broadened to the real time image captioning which will help people with impairment, self driving cars.For now there are some use cases when the objets of same calss appear there

are chances that captioning will might not be upto the expectation. For example – when passing two buckets In which on contain apple and other contain orange then it captioned both as apple. So this can be solved by applying rigorous algorithms and passing more dataset to the training model.

5. 3 APPLICATIONS

The very first application is the self driving cars In this world of automation we can help car bot to move around as here it will does this all in realtime with greater speed and will check which path to move on. Helping the blind , its really hard for these people to move aroung especially crossing high jammed roads here like we can help them by this this will tell surrounding description and ultimately help in moving . CCTV cameras can also be modified to caption the recording .

REFERENCES

Vinyals O et al, CVPR -Show and Tell: A Neural Image Caption Generator

Karpathy A et al, CVPR 2015- Deep Visual-Semantic Alignments for Generating Image Descriptions -

Chen X et al - Mind's Eye: A Recurrent Visual Representation for Image Caption Generation , 2015.

Donahue J et al, CVPR - Long-term Recurrent Convolutional Networks for Visual Recognition and Description - 2015

Jia X et al - Guiding the Long-Short Term Memory Model for Image Caption Generation , 2015.

Mao J et al - Descriptions of Images -, ICCV 2015

Park C C et al - Expressing an Image Stream with a Sequence of Natural Sentences 2015

Xu K et al -Show, Attend and Tell: Neural Image Caption Generation with Visual Attention , 2015

Vendrov I et al, arXiv preprint - Order-Embeddings of Images and Language - 2015

Mansimov E et al - Generating Images from Captions with Attention - 2015.

APPENDICES

```
In [1]: import numpy as np
from numpy import array
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import string
import os
from PIL import Image
import glob
from pickle import dump, load
from time import time
from keras.preprocessing import sequence
from keras.models import Sequential
from keras.layers import LSTM, Embedding, TimeDistributed, Dense, RepeatVector,\
                         Activation, Flatten, Reshape, concatenate, Dropout, BatchNormalization
from keras.optimizers import Adam, RMSprop
from keras.layers.wrappers import Bidirectional
from keras.layers.merge import add
from keras.applications.inception_v3 import InceptionV3
from keras.preprocessing import image
from keras.models import Model
from keras import Input, layers
from keras import optimizers
from keras.applications.inception_v3 import preprocess_input
from keras.preprocessing.text import Tokenizer
from keras.preprocessing.sequence import pad_sequences
from keras.utils import to_categorical

Using TensorFlow backend.
```

Libraries imported

```
In [14]: # convert the loaded descriptions into a vocabulary of words
def to_vocabulary(descriptions):
    # build a list of all description strings
    all_desc = set()
    for key in descriptions.keys():
        [all_desc.update(d.split()) for d in descriptions[key]]
    return all_desc

# summarize vocabulary
vocabulary = to_vocabulary(descriptions)
print('Original Vocabulary Size: %d' % len(vocabulary))

Original Vocabulary Size: 8763

In [15]: # save descriptions to file, one per line
def save_descriptions(descriptions, filename):
    lines = list()
    for key, desc_list in descriptions.items():
        for desc in desc_list:
            lines.append(key + ' ' + desc)
    data = '\n'.join(lines)
    file = open(filename, 'w')
    file.write(data)
    file.close()

save_descriptions(descriptions, 'descriptions.txt')
```

Code for Data cleaning

```
model_new = Model(model.input, model.layers[-2].output)
```

```
In [29]: # Function to encode a given image into a vector of size (2048, )
def encode(image):
    image = preprocess(image) # preprocess the image
    fea_vec = model_new.predict(image) # Get the encoding vector for the image
    fea_vec = np.reshape(fea_vec, fea_vec.shape[1]) # reshape from (1, 2048) to (2048, )
    return fea_vec
```

```
In [41]: # Call the funtion to encode all the train images
# This will take a while on CPU - Execute this only once
start = time()
encoding_train = {}
for img in train_img:
    encoding_train[img[len(images):]] = encode(img)
print("Time taken in seconds =", time()-start)
```

```
Time taken in seconds = 1650.9092202186584
```

```
In [49]: # Save the bottleneck train features to disk
with open("/Jupyter Sketch/encoded_train_images.pkl", "wb") as encoded_pickle:
    dump(encoding_train, encoded_pickle)
```

Image Encoder

```
# data generator, intended to be used in a call to model.fit_generator()
def data_generator(descriptions, photos, wordtoix, max_length, num_photos_per_batch):
    X1, X2, y = list(), list(), list()
    n=0
    # loop for ever over images
    while 1:
        for key, desc_list in descriptions.items():
            n+=1
            # retrieve the photo feature
            photo = photos[key+'.jpg']
            for desc in desc_list:
                # encode the sequence
                seq = [wordtoix[word] for word in desc.split(' ') if word in wordtoix]
                # split one sequence into multiple X, y pairs
                for i in range(1, len(seq)):
                    # split into input and output pair
                    in_seq, out_seq = seq[:i], seq[i]
                    # pad input sequence
                    in_seq = pad_sequences([in_seq], maxlen=max_length)[0]
                    # encode output sequence
                    out_seq = to_categorical([out_seq], num_classes=vocab_size)[0]
                    # store
                    X1.append(photo)
                    X2.append(in_seq)
                    y.append(out_seq)
            # yield the batch data
            if n==num_photos_per_batch:
                yield [[array(X1), array(X2)], array(y)]
                X1, X2, y = list(), list(), list()
```

Generator function

```
# Load Glove vectors
glove_dir = '/Jupyter Sketch/glove'
embeddings_index = {} # empty dictionary
f = open(os.path.join(glove_dir, 'glove.6B.200d.txt'), encoding="utf-8")

for line in f:
    values = line.split()
    word = values[0]
    coefs = np.asarray(values[1:], dtype='float32')
    embeddings_index[word] = coefs
f.close()
print('Found %s word vectors.' % len(embeddings_index))
```

```
Found 400000 word vectors.
```

```
embedding_dim = 200

# Get 200-dim dense vector for each of the 10000 words in out vocabulary
embedding_matrix = np.zeros((vocab_size, embedding_dim))

for word, i in wordtoix.items():
    #if i < max_words:
```

Glove Embedding

```
inputs1 = Input(shape=(2048,))
fe1 = Dropout(0.5)(inputs1)
fe2 = Dense(256, activation='relu')(fe1)
inputs2 = Input(shape=(max_length,))
se1 = Embedding(vocab_size, embedding_dim, mask_zero=True)(inputs2)
se2 = Dropout(0.5)(se1)
se3 = LSTM(256)(se2)
decoder1 = add([fe2, se3])
decoder2 = Dense(256, activation='relu')(decoder1)
outputs = Dense(vocab_size, activation='softmax')(decoder2)
model = Model(inputs=[inputs1, inputs2], outputs=outputs)
```

```
model.summary()
```

```
Model: "model_3"
```

Layer (type)	Output Shape	Param #	Connected to
input_4 (InputLayer)	(None, 34)	0	
input_3 (InputLayer)	(None, 2048)	0	
embedding_1 (Embedding)	(None, 34, 200)	330400	input_4[0][0]

Model

```
model.layers[2]
```

```
<keras.layers.embeddings.Embedding at 0x7ff80cc3fc18>
```

```
model.layers[2].set_weights([embedding_matrix])
model.layers[2].trainable = False
```

```
model.compile(loss='categorical_crossentropy', optimizer='adam')
```

```
epochs = 10
number_pics_per_bath = 3
steps = len(train_descriptions)//number_pics_per_bath
```

```
## for i in range(epochs):
    generator = data_generator(train_descriptions, train_features, wordtoix, max_length, number_pics_per_bath)
    model.fit_generator(generator, epochs=1, steps_per_epoch=steps, verbose=1)
    model.save('/Jupyter Sketch/model_weights/model_' + str(i) + '.h5')
```

```
for i in range(epochs):
```

Training Batches

```
z=0
z+=1
pic = list(encoding_test.keys())[45]
image = encoding_test[pic].reshape((1,2048))
x=plt.imread(images+pic)
plt.imshow(x)
plt.show()
```


```
print("Greedy:",greedySearch(image))
```

```
Greedy: two children are playing on the bed
```


Final Output

HSR1

by Hemant Sharmar1

Submission date: 27-May-2020 05:33PM (UTC+0530)
Submission ID: 1332756568
File name: HemantR1.docx (2.24M)
Word count: 9530
Character count: 49165

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ML	Machine Learning
PCA	Principal Component Analysis
KNN	K-Nearest Neighbour
IP	Image Processing
AI	Artificial Intelligence
FB	Facebook
SVD	Singular Value Decomposition
APK	Application

LIST OF FIGURES

ABSTRACT

Consequently producing a characteristic language portrayal of a picture is an errand near the core of picture understanding. In this paper, we present a multi-model neural system technique firmly identified with the human visual framework that consequently figures out how to depict the substance of pictures. Our model comprises of two sub-models: an article discovery and limitation model, which separate the data of articles and their spatial relationship in pictures individually; Furthermore, a profound repetitive neural system (RNN) in view of long present moment memory (LSTM) units with consideration instrument for sentences age. Each expression of the portrayal will be consequently adjusted to various items of the info picture when it is produced. This is like the consideration component of the human visual framework. Test results on the COCO dataset grandstand the value of the proposed technique, which outflank past benchmark models.

CHAPTER-1 INTRODUCTION

1.1 INTRODUCTION

Artificial Intelligence

AI is a wide-ranging aspect of software engineering dealing with the construction of savvy machines prepared to carry out commitments that typically involve human insight. Computer based thinking is a science field with various algorithm, but headways in AI and deep learning shift the outlook in each sector of the tech industry for all practical purposes.

Can machines think?

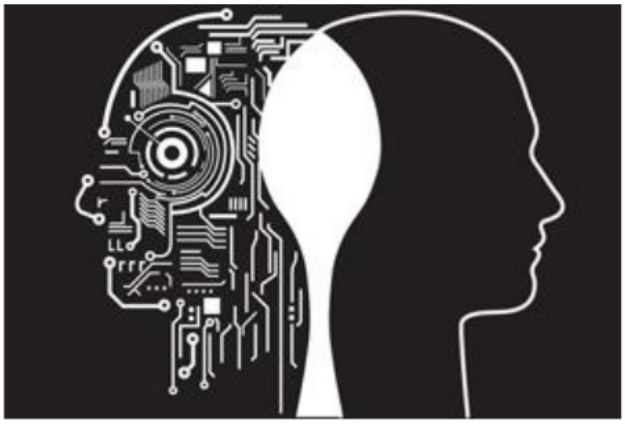


Fig 1.1: Can Machines think

Not even ten years after cracking the Nazi Enigma , the encrypting the system and helping the Army to win World War II and making the edge over everone , the mathematician Alan

Turing changed the history with just a simple inquiry in a second time: "Do really machines think"?

Allen's paper "Processing Machinery and Intelligence" (1950), and Turing Test as a result, established the central goal and perception of man-made brain capacity.

At its heart, AI is the branch of the software engineering that make plans to respond in the confirmed to Turing 's query. It is the responsibility that human intelligence be replicated or recreated in machines.

Man-made reasoning's broad goal has given ascend to various inquiries and debates. To such an degree that no single field value is known all over.

The significant impediment in making AI's as essentially "building machines that are shrewd" is what that doesn't majorly clarify what man-made consciousness is? What makes a machine smart?

Artificial Intelligence: A Latest Approach in their crucial reading material, Stuart Russell and Peter Norvig encounter the inquiry by bridging their work around the topic of insightful machine operators. In view of this, AI is "the investigation of specialists that get percepts from the earth and perform activities."

How is AI be used?

AI commonly bogus under two general classifications:

Restricted AI: This sort of man-made intelligence, also referred to as "Frail AI," operates within a restricted environment and is a replication of human awareness. Slender AI frequently focuses on performing a solitary task very well and bear in mind that these devices can seem astute, They work within undeniably more demands and confines than even the most basic human insight.

Falsified General Intelligence (AGI): AGI, now and then referred to as "Solid AI," is the kind of logical reasoning by computers we found here in the movies, like to Westworld robots or Star Trek 's Data: The Next Generation. AGI is a particular-knowledge of computers and, like an individual, which can apply that insight to take care of any issue.

Thin Artificial Intelligence

Thin AI is covers us and is efficiently the good acknowledgment of man-made reasoning to date. With its emphasis on performing explicit assignments, Narrow AI has encountered various achievements in the most recent decade that have had "critical cultural advantages and have added to the financial essentialness of the country," as per "Planning for the Future of Artificial Intelligence," a 2016 report by the Obama Administration.

A couple of instances of Narrow AI include:

1.Google search

2.Picture acknowledgment programming

3.Alexa bot ,Siri and other individual collaborators

4.Auto-driving vehicles

5.IBM's Watson

Machine Learning

Fig 1.2: Machine learning

Machine learning is an area in artificial in (AI) that gives the ability to the system to understand by itself and enhance from things done earlier without it being programmed by programmer. Machine learning mainly covers on computer programs that navigate and use the information to analyze for themselves.

The learning cycle starts with insights or pattern , like examples, actual experiences or lessons, to search for clues in data and to make informed choices in the futher by relying on the case studies we have. The main objective is to allow machines to automatically learn without humanbeing interferance , and to modify behavior according to previous instances.

Some machine learning methods

Machine learning algorithms are divided into mainly two categories as supervised or unsupervised.

Using named examples to predict future outcomes, machine learning models can be used where past learning is used for new data. Beginning with the train and test dataset, the supervised learning creates a conditional function to make production values predictions. After adequate training from the dataset the system will provide expectations any very new input provided . The algorithm can also make comparison its output with original one , intended output provided and find errors to make further changes in the model accordingly to give a high performance.

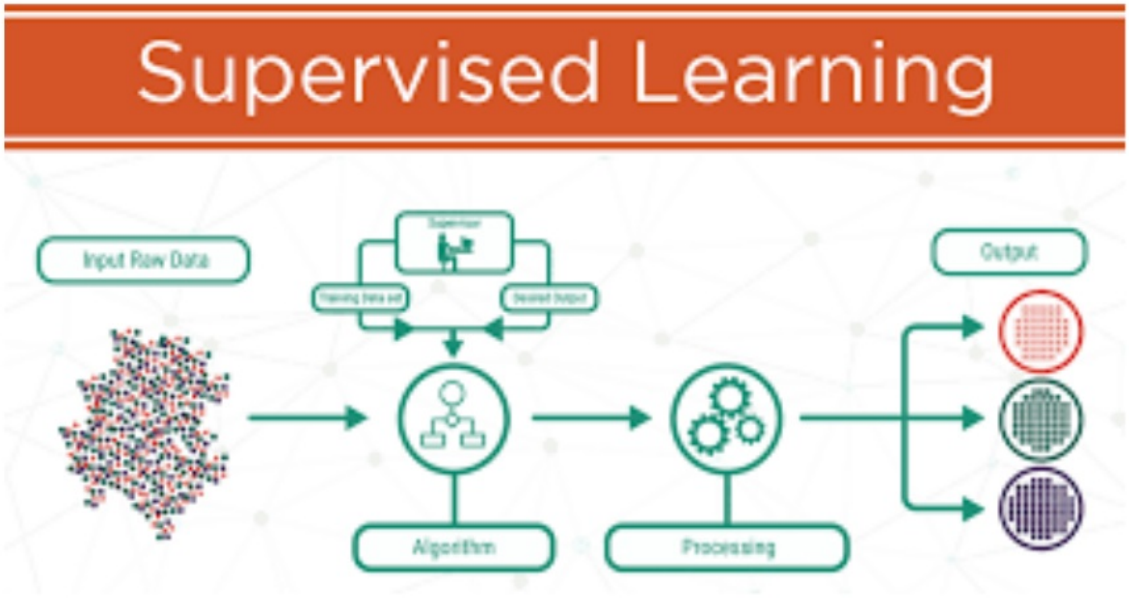


Fig 1.3: Supervised Learning

Unsupervised artificial intelligence algorithms, by contrast, are used where the knowledge used to train can not be identified or labelled. Learning without supervison explores how machines can analyze a function from unidentifiable data to find a secret structure. The machines does not work out the right performance, but it examines the information and can draw matches to explain hidden structures from datasets. structures from unlabeled data.

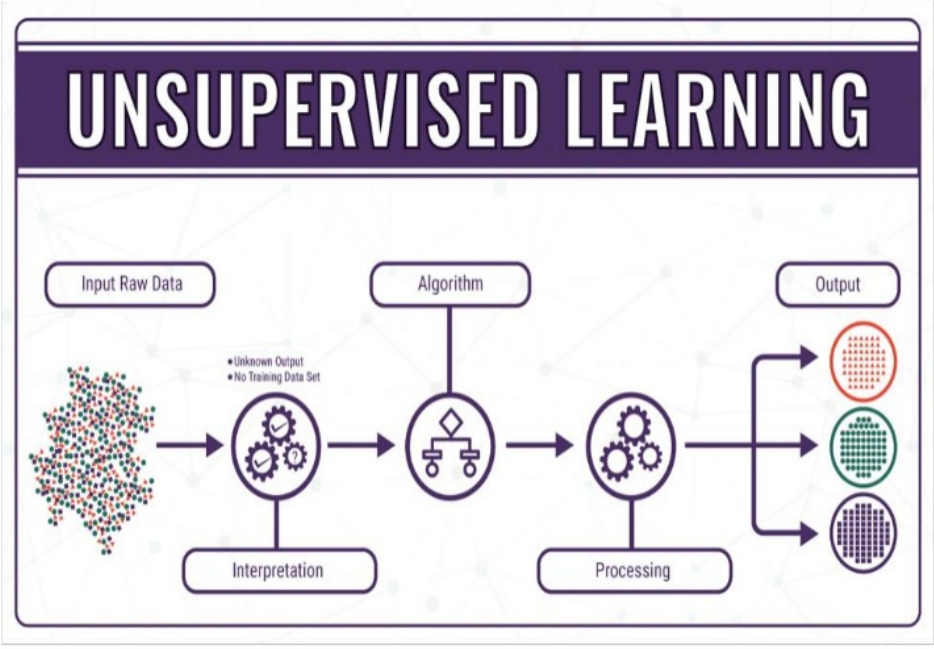


Fig 1.4: Unsupervised Learning

Machine learning helps us to analyze huge amounts of data. While usually providing quicker, more accurate outcomes to see the profitable opportunities or reduce hazardous risks, it can may take some more time and money to train efffectively. The combination of artificial intelligence with Cognitive computing technologies will make this process much faster and more successful

Supervised ML algorithms

1.Linear Regression:

Relapse issues are directed learning issues in which the reaction is ceaseless. Order issues are administered learning issues in which the reaction is straight out. Straight relapse is a method that is valuable for relapse issues.

So, why do we prefer linear regression?

- used largely
- fast responsive time
- easy usability
- It can be interpreted easily
- Acts as base for many other models

Simple Linear Regression Equation

$\hat{y} = b_0 + b_1 x$

Here

- $\hat{Y}$ = response
- x = feature
- b_0 = intercept
- b_1 = coefficient for x

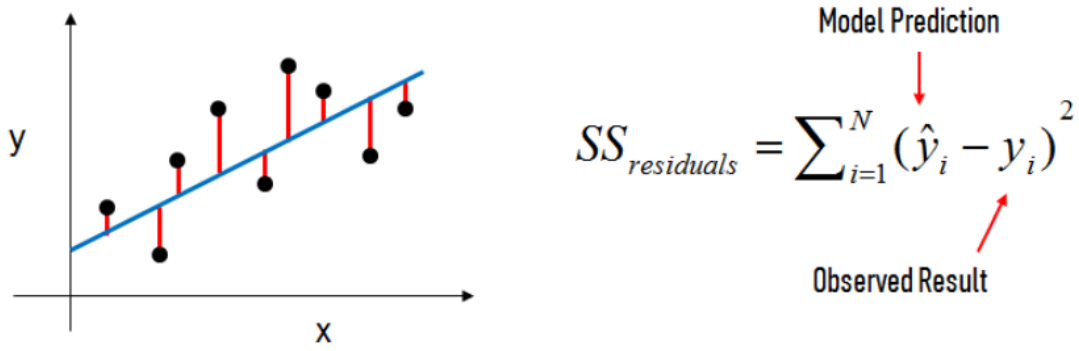


Fig 1.5:Linear Regression

2.Decision Tree

A decision tree is more of a real tree kind of chart with hubs speaking to where we pick a quality and pose an inquiry; edges speak to the appropriate responses the to the inquiry; and the bottom most without child nodes speak to the genuine yield or class mark. They are utilized in non-direct basic leadership with basic straight choice surface.

Decision trees organize the models by organizing them from the roots to a certain leaf center down the tree, with both the leaf hub offering the model a location. -- hub in the tree is conducted as an application for some property, and each edge falling out of that hub is compared to one of the most common responses to it. This process is calling itself again and again in nature and is rehabilitated for each stable subtree.

How about we represent this with assistance of a model. We should expect we need to take on badminton on a specific day — state Saturday — in what capacity will you choose whether should to play or not. Suppose you go and verify if it's sweltering or cold, check the speed of the breeze and mugginess, how the climate is, for example is it bright, shady, or

stormy. You consider every one of these components to choose in the event that you need to play or not.

The baseline methodology that is being in the decision trees is known as the algorithm ID3 (by Quinlan). The ID3 method constructs trees which take decision of it own by using a greedy, from top-down approach. Shortly, the algorithm levels have been:-Pick the best attribute for the NODE — Assign A as the decision attribute (test case). - For every value of node , make a new further of the NODE. – align the order of the training data to the appropriate further node leaf. - If given cases are accurately predicted, then don't do forward and else loop over the new further nodes.

The next major questions, now, is how to select the best characteristic. For ID3, we consider the best attributes in term of which attributes have the most knowledge benefit, a metric which communicates how nicely an attribute divides the data into classes based on predcition.

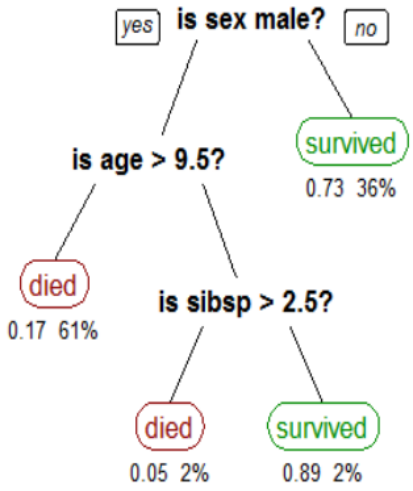


Fig 1.6:Probability

Cost Of Split:

We can reperesent Regresion as sumof(yactual — predictions)^2

We can represent Classifications as : G which is sumof(pk * (1 — pk)

Reasons to stop the splitting

You may request that when quit growing a tree? As an issue for the most part has an enormous arrangement of highlights, it brings about huge number of split, which thus gives a colossal tree. Such trees are mind boggling and can prompt overfitting. All in all, we have to realize when to stop? One method for doing this is to set a base number of preparing contributions to use on each leaf. For instance we can utilize at least 10 travelers to arrive at a decision(died or endure), and disregard any leaf that takes under 10 travelers. Another path is to set greatest profundity of your model. Greatest profundity alludes to the length of the longest way from a root to a leaf.

Pros of CART

- • Easy to understand, to translate and to image.
- Choice trees verifiably perform differentiable screening or highlight determination.
- It Can handle with both numerical and unmitigated information.
- • Option trees typically require little effort on the part of clients to organize information
- Nonlinear parameter relation don't impact tree execution.

Cons of CART

- It can further create complex trees hence making the algorithm more complex
- They can be unstable as small difference can make the whole result difference
- Greedy algorithms can never assure us to provide the max over all interval best decision tree.

Naïve Bayes

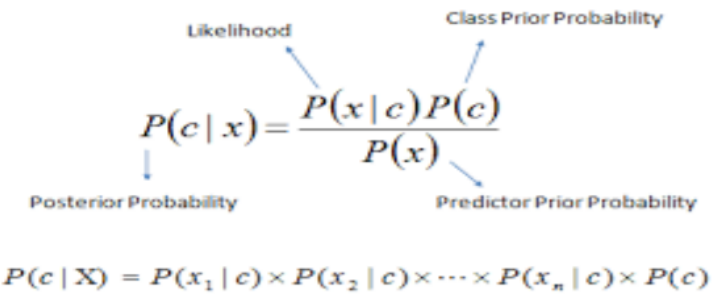


Fig 1.7:Naïve Bayes

It is an order system with a presumption of independence among indicators based on Bayes ' Theorem. In simple terms, a classifier from Naive Bayes assumes proximity of a particular value in a category is random to the proximity of some other element. For eg, an organic item

may be seen as an item on the off probability of being color red , around, and around 3 creeping in length across. If these indicators rely on each other or on the existence of specific highlights, these products freely add to the probability that this natural product is an apple and that's why it's called 'Gullible.'

Guileless Bayes algo is nothing but tough to fabricate and mainly valuable for huge data collections. Alongside smoothly, Naive Bayes is popular to beat even profoundly hard order techniques.

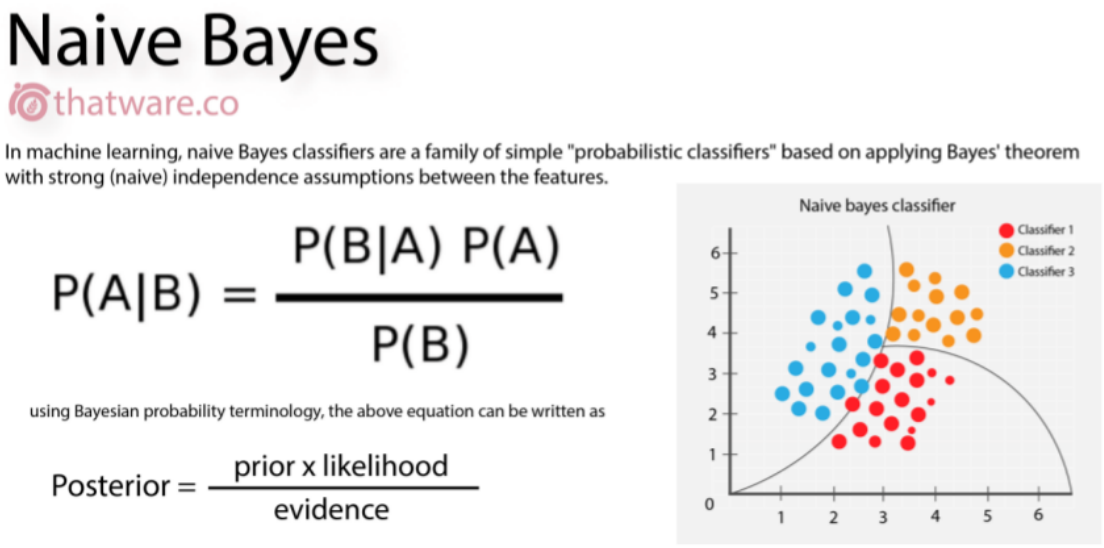


Fig 1.8:Naïve Bayes

Pros

- It is quick and simple to see category of test data collection. It furthermore perform well in multple class predcition They can be unstable as small difference can make the whole result difference
- • At the point where autonomy assumption remains, a Naive Bayes predictor compares good with specific algos such as measured relapse and you need less details to plan
- • It performs well if straight-out data variables contrasting with statistical variable(s) will occur. Typical theft is approved for numerical variable (minus bend, which is a solid suspicion).

Cons

- • If the clear cut parameter has a class (in the test data index) that was not included in the planning of the information index, a 0 (zero) probability will be defined at that point and a prediction will not be made. This is named "zero frequency" occasionally. We may use the smoothing method to understand that. One of the simplest smoothing devices.
- On the contrary , innocent Bayes is otherwise referred to as a bad estimator, and the chance give us output from predicted probability are therefore not to be taken into account.
- Other constraint of Naive Bayes is the presumption of automatic indicators. In actuality, it is in real world outlandish that we get a lot of indicators which are all autonomous.

Disadvantages of Machine Learning:

1. Time and assets-

Machine learning requires lot of assets to work. It might request extra computatuion power. ML requires enough chances to give the calculate a chance to learn and create to make acceptable their proposed reason with a many of accuracy and importance.

2. Data Acquisition

Artificial intelligence needs the training of large data sets, which should be inclusionary / impartial and of good quality. There may also be occasions when they will wait to produce new data.

3. Translation of results-

Precisely deciphering the outcomes produced by the calculations is a difficult errand. One needs to practice alert while picking calculations for their particular reason.

Advantages of Machine Learning:

1. Contiguous improvement-

Learning by the Machine algo improve in correctness and efficiency as they gain by previous attempts. This helps them make good decisions.

2. Distinguishes patterns and examples effectively-

AI is able to audit enormous volumes of data and find explicitly pa tterns and examples that people wouldn't see clearly. For example, for a business based on web site such as Flipkart, it helps to understand their customers' perusing habits and purchase chronicles to help them take into account the right products, plans, and notifications. It uses the possibility to not showing important promotions to them.

3. Automation-

Machine learning doesn't require human mediation. It enables machines to learn. It assists machines with making forecasts and improve the calculations independent from

anyone else. Against infection programming is a typical case of this as they consequently channel new dangers as and when they are perceived

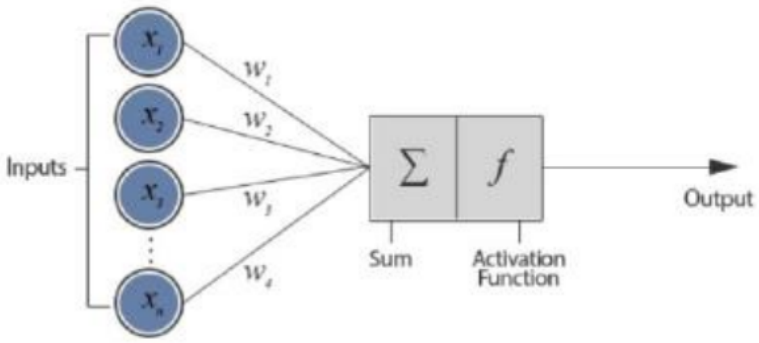


Fig 1.9:Deep Learning

Deep Learning

Deep learning is one of the set of machine learning that function to replicates the way of work of the human mind in handling info and making pattern for use in making the decisions. Deep learning is capable of learning not supervised from data that is not structured or not labeled. Also known as deep neural network.

How Deep Learning Works

In the digital age, deep learning has evolved tremendously, resulting in an blast of data in every forms and with each and every part of the universe. Known as big data, this

information is compiled from outlets such as social networks, search engine and digital cinemas. A enormous qunatity of info is frequently used and can be exchanged via fetch app such as computing on cloud.

The info, though, which is usually highly not structured, is so enormous which it may take years for human beings to grasp it and take out necessary info. Company understand the tremendous more possible that can arise from this abundance of knowledge being scrambled, and are gradually adapting for automated help to Artificial inteligent systems.

Deep Learning Versus Machine Learning

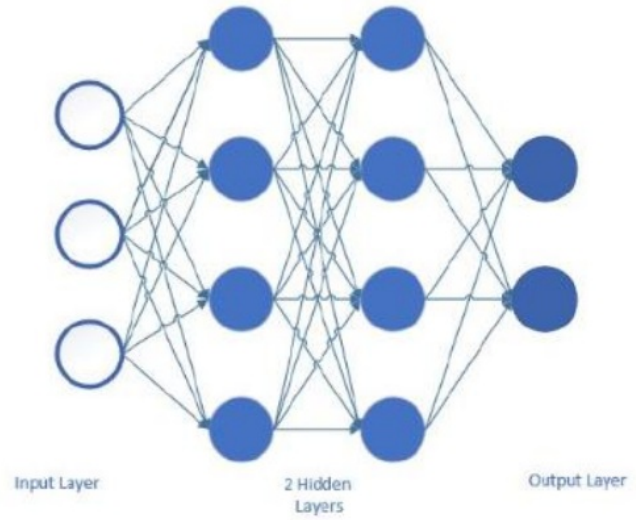


Figure 3.4: An Artificial Neural Network also known as a Multi Layer Perceptron

Fig 1.10: Multi Lever Perceptron

Artificial intelligence is one of the mostly used AI algo used for managing large data, a auto-adaptive algo that continues to evolve with experience or add more data.

To this end, if a online payment companies wanted to find the presence or probability of illegal work in its system, it would use machine learning technique. The network connectivity made into a computer algo will analyze all transfers that exist on the onlines network, identify data set patterns and notice out any deviation that the pattern detects.

Data science, a part of machine learning, follows a orderly level of artificial neural networks which perform machine learning processes. The CNNs are constructed like a human mind, with neurons nodes linked like a spidersWeb. While old-style systems build linear analyzes of the results. the hierarchical way of functioning of deep learning systems allows system to load info with a not so linear approach.

An old-style method to identifying illegal work or money filtering may depend on the volume of transfer that follows, whereas a nonlinear deep learnings idea would have included time, geographical locations, IP addresess, and some other function that would possibly point to fraudulent behavior. The very first layer of the algorithm is used to process a new data given like the transfers quantity and transfer it as output to the next step. The very next layer collects the information from the previous step by adding some extra info such as the IP address of the user and transfers the result on.

The next step includes data output from the second layer and uses the raw data, such as geographic location , making the design of the system even better. This is continuing through all neuron network stages.

CNN

A Convolutional NN is a Deep Learning algorithm that can take in an input data, assign location (learnable weight and bias) to different enitites in the image and distinguish one term

from some other. The pre-treatment required in a simple ConvolutionalNet is much smaller than other classification techniques. Although the filter are arm-engineered in ancient times, with appropriate training scale, ConvNets has the ability to learn these features.

A ConvolutionalNet's architecture is similar to that of Neurons' connectivity pattern working in the Human Mind, and was motivated by the Visuals Cortex's organizations. Sensory cells only react to changes in a restricted portion of the field of vision as the Receptive Field. A collection of such objects joins together to cover the entire area of vision.

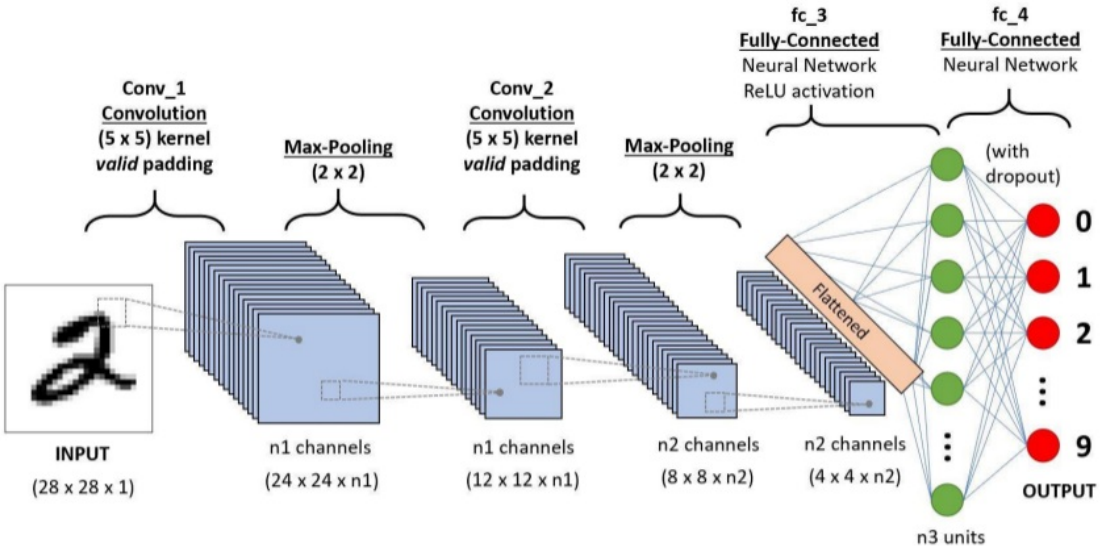


Fig 1.11:CNN

DenseNet

Recent research on Convilutionla Neural Network has showed that training can be significantly dense, more effective and more fastly if they include smaller links middle of structures near to the given data and those near to the output data. We accept this observation in this paper and introduce the Dense Convolutionary Network (DenseNet), which links each

step to each other in a feed-forward like . Although typical CNN with L layers have L links — one across every step and one step thereafter — our link has L(L+1)/2 deep links. The features maps of every previous layer are used as input by each layer and their features maps are used as input in every consequent step DenseNets have many excellent features: they alleviates the problem of the vanishing gradient, reinforce the propagations of features, encourages reuability of features and significantly minimise the amount of data used.

The counterintuitive effects of this complex linking pattern would be that it needs afew criteria than conventional CNN, because terminated feature maps need not be relearned. Current feed-forward systems can be seen as algo in a state transmitted from one layer to the next. Can step read the states from its former layers and write to the step next. This change the environment but still transfer the data that must be processed. ResNets[11] makes this info stored automatically through modification of an additives identity. Latest variation of ResNets shows that several layer makes very less contribution and be dumped at irregular intervals while train the data . This make the states of ResNets the same as (not rolled) RNN, but the number of ResNets parameters is considerably greater since every layers has its veryown weight. Our given architecture for DenseNet specifically varies between information applied to the network and files contained in it. DenseNet layer are very slim (e.g. 12 feature maps per layer), introducing only a limited number of features maps to the channel's "collective information" and making the remaining features maps useless — and the ultimate classifiers make a decision based on all of the channel's feature maps.

Besides improving variable performance, one major advantage of DenseNets is its changed transmission of communication and gradient across the link , which allows us to trains them easily. Every layer has a directs access from the loss function to the gradients and the original control signal, resulting in implicit deep guidance. This help in training of a deeper networks

architecture. Further, we also notice that dense connections have a normalizing effects, which minimizes over- fittings on task with small training set sizes.

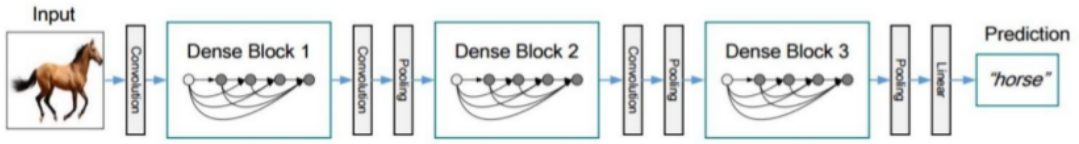


Figure 2. A deep DenseNet with three dense blocks. The layers between two adjacent blocks are referred to as transition layers and change feature map sizes via convolution and pooling.

Fig 1.12: DenseNet

Subsetting extracted features learned from several layer increase variance in the inputs of successive layer and improve their performance. That marks a major gap between DenseNets and ResNets. DenseNets are easier and more powerful compared with Inception channels [35, 36], which often concatenate features from different layers.

ResNet

According to the ultimate approximation principle, we knows that a feedforward link with only one layers is able to prove any function, provided enough power. The layer, however, could be enourmous and the networks would be inclined to overfit the info. There is therefore a shared fashions in the scientific community that our core links need to go deep.

The cutting-edge CNN software is moving deeper and deeper after AlexNet. The VGG networs and GoogleNets (also nam as Inception V1) has nineteen and twenty two layer respectively, while AlexNet had just 5 convolutionary layers.

But through the depth of the channel doesn't work by simply piling layers together. Due to the not so famous disappearing gradient problems, deep neural networks are hard to train — like the gradient is back-propagated to earlier given layers, autorepeated multiplications can made the gradient infinite very little. As the link deep downs, its quality becomes soaked or even starts to degrade rapidly.

Earlier of ResNet, there were many way to deal with the issues of the disappearing gradient, for example, [4] sums an additional losses in the middle layer as additional control, and none of them seems to solve the problem once and for all.

The main premise of using ResNet is to incorporate called "identification shortcut connection" that misses one or more strands, as shown in the figures below

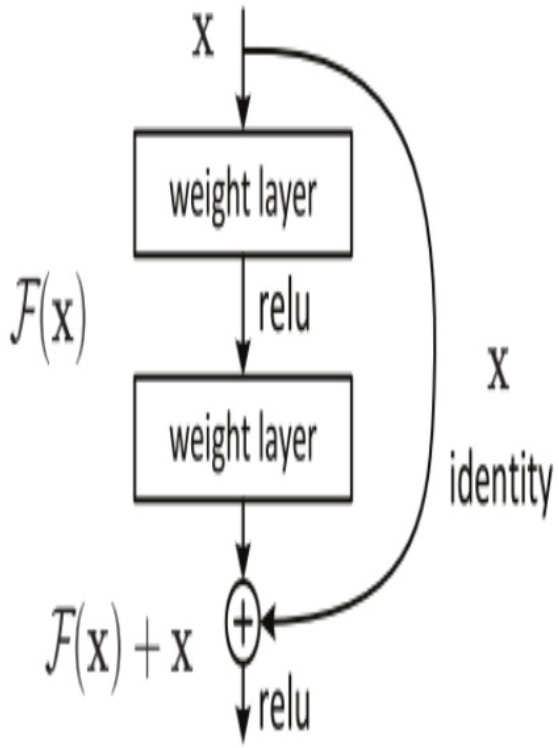


Fig 1.13:ResNet

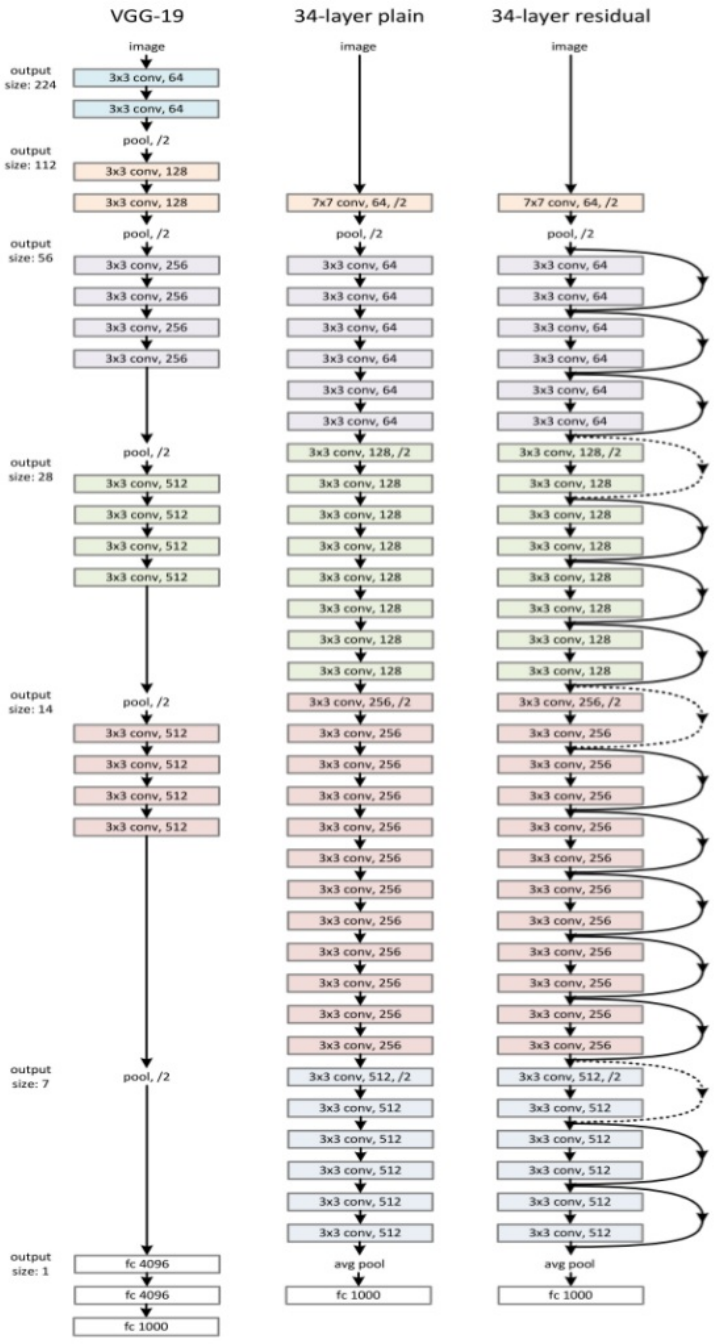


Fig 1.14: ResNet

The writers of [2] argue that piling layers will not reduce the performance of the network since we could actually stack and mapping identities (layer that does nothing) on the existing network and the resultant architecture will do the same thing. It means that a learning error greater than the shallower counterparts will not occur with the deeper model. They hypothesize makes it easier for the pieces to fit a residual mapping than to let them fit the required cognitive process directly. And the face legal above explicitly enables it to do exactly that.

VGG

- While prior AlexNet variants concentrated in the first convolutionary layer on a smaller filter size and strides, VGG tackles a really critical feature of CNNs: width. Let us go over the VGG architectural style:

- VGG takes in an RGB image of 224x224 pixels. The writers clipped out center 224x224 patch for each picture to keep the image size appropriate for the ImageNet contest.

- In VGG, the convolutionary layers use a very small receptive field (3x3, the smallest size still capturing left / right and up / down). There are different 1x1 convolution filters which act as linear awareness workshops, followed by theReLU unit. The convolution stage is set to 1 pixel so that after the convolutio the spatial resolution is conserved

• Completely ConnectedLayers. VGG has 3 final fully connected: the first two have 4096 channels each and third has 1000 layers, 1 channel per class

- CloakedLayers. All extracted features of VGG are using aReLU (a huge innovation from AlexNet that is cutting training time). Generally speaking, VGG donot uses Local Response regularization (LRN), since LRN increases memory requirements and training time without any specific increase in precision.

The Difference

VGG, while premised on AlexNet, has many differences which separate the testable theories from the others:

- Rather than using very large temporal information such as AlexNet (11x11 at 4 step) VGG uses very small temporal information (3x3 at 1 step). Because 3ReLU units are now in place of just one, the way things is more discriminatory. Also there are smaller criteria (27 times the level of channels instead of 49 times the number of channels that AlexNet has).

- VGG contains 1x1 convolution layers in order to make the decision more non-linear without changing the temporal information.

- The small-size convolution filters allow the VGG to get a wide number of layers of weight; more strands of courses lead to better efficiency. That's not a uncommon thing though. The 2014 ImageNet contest also featured GoogLeNet, another platform that uses the deep CNNs and limited convolution filters.

Fig 1.15: VGG

Advantages of Deep Learning:

1. Has top-level execution in matters that ultimately beat different schemes in different fields. This includes discussion, language, sense of direction, messing around like Go etc. However, this is not by a tad but by a vital amount.Diminishes the requirement for include designing, one of the most tedious pieces of AI practice.
2. Is a technology that can be adapted moderately effectively to new issues such as vision, time scheduling, language, etc., using procedures such as convolutionary neural systems, repeated neural systems, lengthy transitory memory, etc..

Disadvantages of Deep Learning:

1. It computationally extremely costly to plan. This takes a long time for the most impredictable models to plan using several computers fitted with expensive GPUs.
2. Choosing the topology / boost / prepare strategy / hyperparameters for deep learning is a dark manner of working without any hypothesis to guide you.
3. Rather computation time costly to plan. It takes a long time for the most impredictable prototypes to prepare using several devices fitted with costly GPUs.

Differences between Machine based Learning and Deep Learning

	Machine Learning	Deep Learning
Dependencies on data	Fantastic exhibitions on a little/medium dataset	Phenomenal execution on a
Hardware dependencies	Work on a low-end machine.	Requires ground-breaking r GPU: DL plays out a duplication
Feature engineering	Features must be understood first then the data	No need to understand th represents the data
Time of Execution	Can take from a few min to hrs	Can take upto weeks but Ne Network requires to process a important no of weight.
Interpretabe	There are a few algo are not so diffficult for interpretation and some of them are really hard to inderstand	hard to impossible

1.2 PROBLEM STATEMENT:

The issue presents a subtitling task, which requires a PC vision framework to both restrict and depict notable areas in pictures in normal language. The picture subtitling task sums up object location when the depictions comprise of a solitary word. Given a lot of pictures furthermore, earlier information about the substance locate the right semantic mark for the whole image(s).

Chapter 1 provides a basic introduction about the project so as to give the basic idea and details regarding what we are going to do and also to familiarize you with the technical and few of the necessary theoretical aspects.

Chapter 2 includes the Literature survey i.e. review from different journals, research papers etc.

Chapter 3 aims at the system design, the techniques and the different tools needed for the project.

Chapter 4 tells us about the project performance analysis

Chapter 5 provides the conclusion, and also tells about scopes in the future for the same

Input: A image

Expected Output: Description of image in the form of subtitles and audio

Dataset: Flickr8k(8000 images)

References:

1.IITM Cse department TPA.

2.Towards Science Blogs

3.Data School

4.Medium Blogs

Libraries Used:

- Keras 1.2.2

- Tensorflow 0.12.1
- tqdm
- numpy
- pandas
- matplotlib
- pickle
- PIL
- Glob
- Google Text to speech API
- Glove

1. 3 OBJECTIVES :

The objectives of the project is to get the output audio of the picture that we have provided.It is basically a image captioning project with a audio feedback for which we have used a Google text to audio API which will give us an output audio mp3 file with the resultant audio. We have used Transfer learning for this as we used a pre –trained model Inception V3 which is a full-convolutional, 48-layer-deep neural network. You can load a pretrained version of the trained network from the ImageNet database on more than one million images. The pre-trained network is capable of classifying images into thousands of classes of objects, such as car , mouse, pen and several creatures.

1. 4 METHODOLOGY:

We will Use Flickr 8K dataset which comprises fo 8000 images and will divide that into 2 parts that are Training Set and Testing Set. Training Set will be dataset on which we gonna

train our model which comprises of 6000 random images while testing set comprises of 2000 images on which we going to test our model .

Then we will going to clean out whole training and testing data with the help of some functions. Images are a mere input (X) to our algo. As you might already know, every input to a algo must be vector-shaped.

The transformation of every image into a vectors of a Constant size which can then be fed into the NN as input. To this end, we use the InceptionV3 model (Convolutional Neural Network) developed by google Research to opt for transfer learning.

Data Preprocessing Captions

Subtitles are something we wish to predict. So subtitles will be the targeted the parameters (Y) which the algo learns to predicts as during training period.

But the the whole caption 's estimate, given the image, doesn't happen immediately. Word by word, we'll guess the caption. So we need to encode all word into a vector of a constant size. That part should be seen later, even so, when we look at it

Data preprocessing using generators

That is amongst the most significant elements in the study of the whole case. Here we will explain how the info can be configured in a way which is suitable to be provided as input to the deep neural network.

Data generators are a feature that is native to Python. The ImageDataGenerator category given by the Keras API is nothing other than a generator function implemented in Python.

Model

Here comes out model that we have built which will ultimately going to predict output for us.

CHAPTER-2 LITERATURE SURVEY

Convolutionary Neural Networks (CNN) are naturally enhanced Multi Layered Perceptron variations. It consists of at least each convolution layer (regularly with a sub-sampling step) and is then decided to pursue as in a normal multilayered neural system by at least one fully associated layer. A CNN's engineering is designed to exploited the 2D structured of an information image (or, for instance, a discoursed signals, other 2D data). This is achieved with neighborhood groups and bundled heaps pursuing by some sort of pool techniques that brings invariant highlights to analysis. A further benefit of CNNs is that they are easier to plan and have far fewer parameter than fully connected structures with a comparable number of covered units. CNNs have been commonly used and read for image errands and now are cutting-edge systems for object recognition and exploring Repetitive Neural Networks (RNNs) in numerous NLP undertakings. RNNs are called formulaic in view of the fact that for each element of a grouping they perform a similar odd job, with the yield being based on past computations. Then again, RNNs can be considered as "memory" systems that catch up to this point data on what has been defined. Hypothesis is that RNNs can use data in highly subjective long sequences, Nonetheless, they are limited by and by looking back just a few steps. Our model's goal is to construct subtitles or picture representations naturally. Various conferences have done explorations in the past, finding a place in both business and the academic world that looks something like or is in the light of a theme like what we do. From these explorations various parts of our model take reference. The investigation papers we used are referred to separately in the list of sources, similar to those provided by D. Narayanswamy et al[1] which aims to establish marks characterizing the outlines of the film, or that of DElliott, F. Kellers, Photo depiction using portrayals of visual reliance, whereby the creators strive to identify the various constituents of a image. Anyway, the research

discussed above is generally concerned and focuses more on using image handling to classify what's more, to differentiate separate books in a image . They never manages different setting of these articles. Such discovery publications have enables us to understand incredibility the concept of preparing and dividing pictures which let a significant jobs in my framework. Even we seek to achieve the picture's rational representation. In the aforementioned area of knowing the perfecting of picture in a variety of areas, especially in the centre sections of market and the academic world, a lot of present work is also being completed. The new business as usual involves wide-ranging inquiries about Pc programming and NLP based bunches ,The ones we've discussed are those at Stanford A. Karpathy, Li Fei and UT, Austin, which are both researching goals to construct photo inscriptions. Cortana is an aspect of artificial intelligence created by Microsoft (Cortana) apart from the scholarly population 's striking modern improvements. , As of their current Microsoft Build meeting, they are organizing the AI in their Bing pursuit management, which will allow the customer to communicate with the User Interface more naturally. The most feasible operation of image preparation and subtitles was by Reddits (the Reverse Image Search), which allowed a customer to move an image and to break down their estimate. and shows images with a compasioned settings, this was trailing by Google's, despite the facts are that it is to be Seen that both these tasks are still in training stage also, being worked on. Another eminent use of CV by Facebook Team (Image Tag). Through our prototype we aim to provide syntactically correct and outwardly based representation of theoretical artifacts, the given depiction of which will be represented in natural language, e.g. humans viewed. By using systems such as CNN, RNN and data indexes, such as those of Flickr, we are striving to get a human-level scenario of the images in question. Our main knowledge is that we can use these enormous picture sentences datasets by looking at the phrases as fragile names, where synonymous pieces of word refer to some particular yet obscure region of the scene. Our methodology is to infer and use these arrangements to familiarize ourselves with a training algorithm of representations We are building a fundamental model of the neural system which derives the structure between segments of sentences and the district of the image they represent. We

present a recurrent neural technology that takes a image of the data and generates its material representation. Our exams show that the statements generated yield fair subjective expectations.

CHAPTER-3:SYSTEM DESIGN

In this part we would discuss the means that we are going to take while develop our framework. Subsequent to experiencing distinctive research papers we intended to go fora non-intrusive method for developing our framework. As we all know that there are lots of different object detection and image captioning strategies are available but we go for the inceptionV3 and beam search technique for designing our system and we have gone for such a strategy because it is increasingly solid and relevant under various circumstances.

We design a system which depends on the following two stages:

1. Object Detection

2. Captioning the image and Audio Feedback

Firstly we will create a dictionary of captions and then we will calculate the unique words. Then we will train our dataset by passing the images and passing the caption. We are using Flicker8K dataset. In Image Captioning, a CNN is utilized to separate the highlights from a picture which is then alongside the inscriptions is nourished into a RNN. To separate the highlights, we utilize a model prepared on Imagenet. I gave a shot VGG-16, Resnet-50 and InceptionV3. Vgg16 has very nearly 134 million parameters and its main 5 mistake on Imagenet is 7.3%. InceptionV3 has 21 million parameters and its main 5 blunder on Imagenet is 3.46%. Human top-5 mistake on Imagenet is 5.1%.

Preprocesing the images using InceptionV3

Beginning V3 is a generally used picture acknowledgment algo appears to be accomplished many more than seventy eight percent exactness on the ImageNet set of Data. The algo is the climax of numerous thoughts created by various specialists throughout the years.

The design itself consists of synchronous and awry building squares like convolutions, regular max pooling ,pooling, concats and fully linked layers. Batch norms is commonly used in the system, and applied to inputs for actuation. Misfortune is handled using Softmax.

The coding which provides 3 core binaries for:

1. Learning an Inception v3 model from begining over different devices and different devices using the training set for the ImageNet 2012 Competition.

2. Building an Inception v3 network with a single across different devices and/or multiple computers that use Training data set for the ImageNet 2012 Competition.

3. Reassign an Inception v3 network on a new assignment and back-propagate the mistakes to fine-tune the weights and biases.

A sequential stochastic gradient descent through several GPUs is used for the preparation. The user can specify how many GPUs they want to utilize. The sequential train of data conducts batch slicing by converting in number of batches over several GPUs

The main knowledge identifies with layer tasks. In a conventional conv net, each layer extricates data from the past layer so as to change the info information into an increasingly helpful portrayal. Be that as it may, each layer type removes an alternate sort of data. The yield of a 5x5 convolutional part discloses to us something other than what's expected from

the yield of a 3x3 convolutional bit, which reveals to us something else from the yield of a maximum pooling piece, etc, etc. At some random layer, how would we know what change gives the most "valuable" data?

An Inception module processes numerous various changes over a similar information map in parallel, linking their outcomes into a solitary yield. As it were, for each layer, Inception does a 5x5 convolutional change, and a 3x3, and a maximum pool. What's more, the following layer of the model gets the chance to choose if (and how) to utilize each snippet of data.

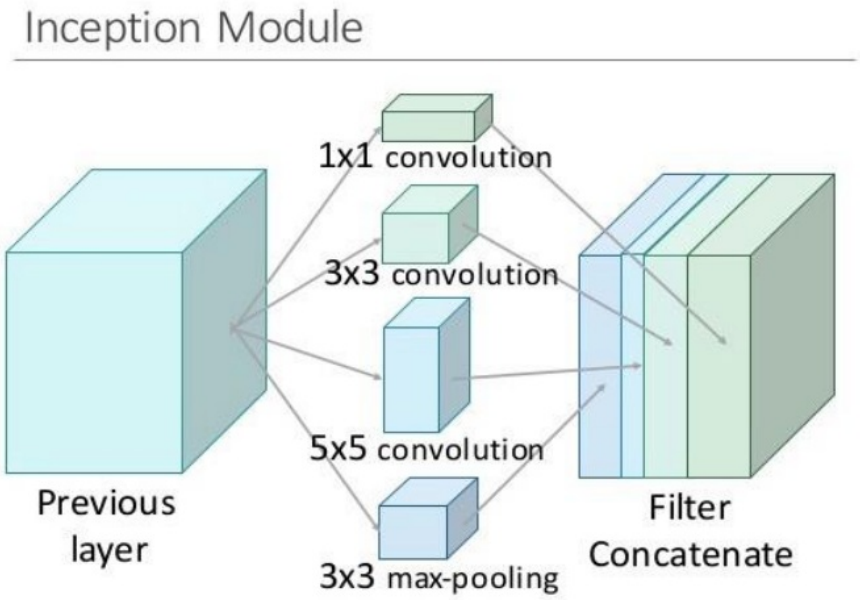


Fig3.1 :Inception Module

The expanded data thickness of this model engineering accompanies one glaring issue: we've radically expanded computational expenses. Not exclusively are enormous (for example 5x5) convolutional channels characteristically costly to process, stacking numerous various channels one next to the other significantly expands the quantity of highlight maps per layer. Furthermore, this expansion turns into a fatal bottleneck in our model. Consider it along these

lines. For each extra channel included, we need to convolve over all the info maps to compute a solitary yield. See the picture underneath: making one yield map from a solitary channel includes figuring over each and every guide from the past layer.

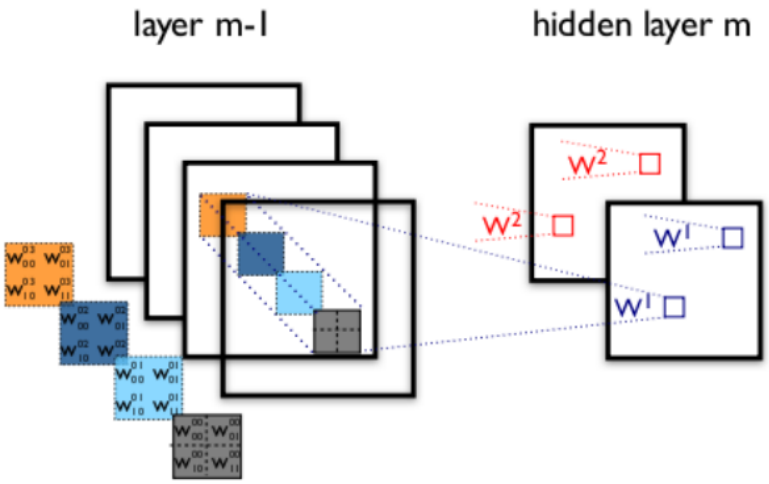
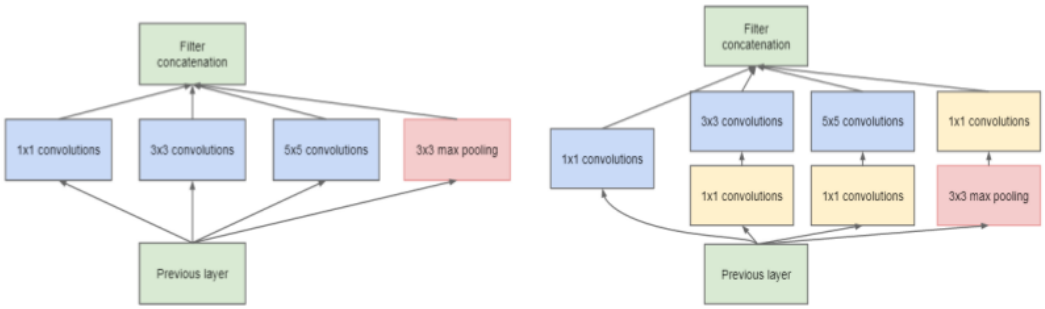


Fig3.2 : Layer structure of Inception Module

suppose there are M input maps. One extra channel implies convolving over M more maps; N extra channels implies convolving over N*M more maps. As it were, as the creators note, "any uniform increment in the quantity of [filters] brings about a quadratic increment of calculation." Our credulous Inception module just significantly increased or quadrupled the quantity of channels. Computationally, this is a Big Bad Thing.

This prompts knowledge #2: utilizing 1x1 convolutions to perform dimensionality decrease. So as to illuminate the computational bottleneck, the creators of Inception utilized 1x1 convolutions to "channel" the profundity of the yields. A 1x1 convolution just sees each an incentive in turn, however over numerous channels, it can remove spatial data and pack it

down to a lower measurement. For instance, utilizing 20 1x1 channels, a contribution of size 64x64x100 (with 100 component maps) can be compacted down to 64x64x20. By diminishing the quantity of information maps, the creators of Inception had the option to stack diverse layer changes in parallel, bringing about nets that were at the same time profound (numerous layers) and "wide" (many parallel tasks).



(a) Inception module, naïve version (b) Inception module with dimension reductions

Fig 3.3: Naïve Version and Dimention reductions

Caption generation is a demanding artificially intelligent issue where a textual explanation for a given photograph must be produced.

It requires the two strategies from PC vision to compr/ehend the substance of the picture and a language based algo from the area of characteristic language handling to transform the comprehension of the picture into words organized appropriately. As of late, profound learning strategies have accomplished cutting edge results on instances of this issue.

Profound learning strategies have shown cutting edge results on inscription age issues. What is generally great about these techniques is a solitary start to finish model can be characterized to foresee a subtitle, given a photograph, rather than requiring complex information arrangement or a pipeline of explicitly planned models.

Image Captioning

Image captioning is a famous Artificial Intelligence (AI) area of research that deals with the knowledge of images and a language summary for that image. The perception of images requires to identify objects and recognize them. It also has to recognize the requirement or location of the picture, the properties of the objects and their encounters. Generating much-formed sentences requires a syntactic as well as a linguistic comprehension of the language.. Knowing an image depends primarily on having features of the image. The techniques employed for this reason can be classified into two different classes:

(1) New methods, machine learning algo based and

(2) Deep data science detection algorithm. handmade feature such as Locals Binary Patterns (LBP) in conventional machine learning

Scaled Invariants Feature Transform (SIFT), Oriented Gradient Histogram (HOG), and a variation of these characteristics are commonly used. Attributes are extracted in such techniques From data to entry. To identify an object, they are then moved on to a classifier such as Support Vector Machines (SVM). Because handcrafted features are unique to the mission, they extract characteristics from a large and diverse collection. Also, practical world info such as image and videos are complexed and have many semantical interpretation.While, in deeply machine learning based algo techniques, feature are learned automatically from data training and they can take care a enourmous and diversified set of image and video. For instance, Convolutional Neural Networks (CNN) are commonly used for the learning of features, and identification using a classifiers such as Relu. For producing

captions, CNN is usually accompanied by Recurrent Neural Networks (RNN). A huge number of posts on text categorization have been authored in the last five years, to deep machine teaching which are more popular and used . Deep learning algo can manage the image captioning complexities and problems reasonably well. Only three study papers on this research topic have been published thus far. And although publications provided a nice literature surveys of image annotations, they can cover a some paper on deep learning as many of them were authored after the surveyed paper . Such surveyed paper primarily covered based on model, replication-based, and creating models from really some deep learning-based, novel image captions. Alarge numbers of research on deep learning-based text categorization have been completed, however. The presence of huge and raw dataset has also make the learning based on image transcribing an important topic of study . To include an english translation of the literaturary part , we show surveys mainly on image captioning based on deep learning texts. The main contribution of this study is to show an exhaustive surveys of deep learning for captioning of images. Next, the current captioning of images papers are grouped into 3 main categories:

(1)Image captioning based on Template

(2)Images captioning based on Retrieval

(3)Images caption baased on Novel generations.

The classifications are briefly discussed in Section 2. Many methods of deep learning based images annotations fall under the classes of generations of creative captions. Thus, with machine learning we focus only on the generation of novel captions. Secondly, we combine the approaches of deep learning text categorization into different groups, namely

(1) Based on the space visuals,

(2) Based on multimodals,

(3) Supervised learning

, (4) Other deep learning,

(5) Based on Densed Captioning,

(6) Based on all Scenebase,

(7) Based on Encoding Decoding Architecture

(8) Based on Composition Architecture,

(9)Model based on LSTM(Long Short-Term Memory),

(10) Others language model-based,

(11) Attention-Based,

(12) Based on Semantic,

(13) Stylized captions

CNN

CNN image order took, process and community an data image under different groups (e.g., Dog, Cat, Tiger, Lion). PCs find a picture to be a pixel display and depend on the objectives of the image. It will see height x width x depth(h = Height, w = Width, d = aspect in view of the photo goals). E.g., an pic of 3 x 6 x 6 RGB grid clusters (Three parameter to RGB values) and a pic of 1 x 4 x 4 shows of network of pic of grayscale.

Actually, in-depth learning of CNN algo to prep and check, each knowledge images will go through a continuum of convolution layers with streams (Kernal), fully, pooling, associated layers (FC) and applying Relu capability to characterize an object with stochastic quality somewhere within the range 0 and 1 . The figure below is a completed progression of CNN processing a picture of information and grouping the reports based on values.

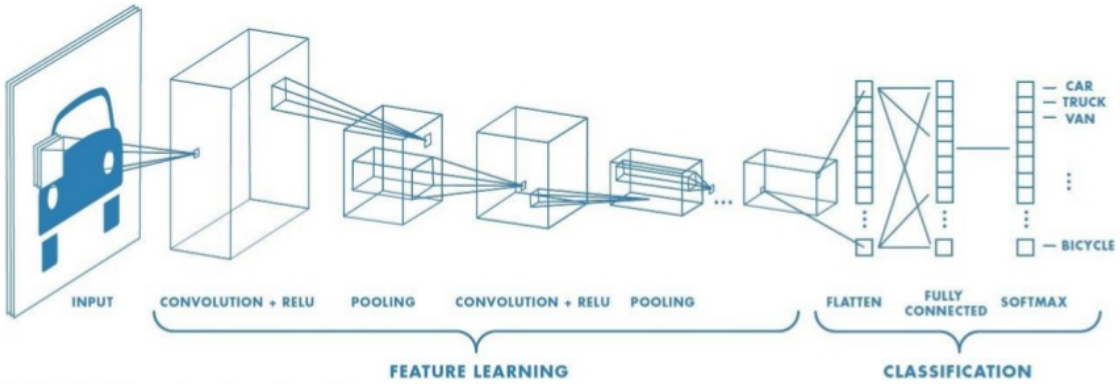


FIG 3.4:Neural network with convolution layer

Glove

GloVe is an unsupervised machine learning algorithm who use words to obtain feature vectors. Implementation is carried out from a corpus on consolidated global phrase-word co-occurrence statistics, and the consequent interpretations show fascinating linear shapes of the word vector space.

The word occurrence statistics in a corpus are A primary sour of info for everyone

Uncontrolled methods of computational methods of verbs, and while many such techniquesnow exist, the question arises how meaningful From such statistics is produced, and how the corresponding word vectors might reflect that meaning.

We shine new light on that question in this section.

Test Plan

This topic displays the execution subtle for the proposed system created for the Image Captioning is discussed here.

The software used in the execution of the project its platform bundles and so on are described as follows

Requirement

Platform:
Widows 7 and above/Max OS 10 or above
Software Used:
Python/Jupyter

Implementation Details

The implementation details are carried out in 3 stages

1.Input Name

We will input the image to the system with various alignments with different illumination conditions and use those image to train the system so that the system is able to recognize different objects

2.Feature Extraction:

Then inceptionV3 will extract features for us of different objects in the picture.

3.Building the model:

Relevant models is being built in order to execute the code

4.Prediction:

Prediction in the form of captions is being displayed and audio is being generated

Conclusion

In this chapter we examined our proposed system in details and how the whole system related with Image Captioning is implemented.We likewise talk about the Algorithm implemented in various periods of image captioning and the means which we will pursue while building our system.

CHAPTER- 4 PERFORMANCE ANALYSIS

While checking out the data points if we take only 2 images it will lead to some number of datapoints let say 15. But in reality dataset we have around 6000 images each having five captions in it , which sums up to 30000 images and caption

If we assume that every caption on a average is 8 . which will ultimately lead to a total of 30000 * 8 = 240000 data points and the size of matrix will going to be so large assuming n as number of data points (240000) , m as size of every data point which is size of the image vector + size of partial caption(x) which combines to be 204b + x. Each term (or index) is mapped to high dimensional space by means of one the techniques. Later, we'll see how every character is being map to a 200 lengthed vector with the use of pre trained mode of GLOVE words embedding while the algo building stage. each series comprises 34 indexes , in which every positon is a 200-length vector. Here x = 34 * 200 = 6800
Consequently m = 2048 + 6800 = 8848.
Last, matrix size= 8848* 240000 = 1857080000 blocks.

even if we presume that one frame takes 2 bytes, then we would need upwards of 3 GB of main memory that stores this data matrix.

That is quite an enormous requirement and even if we able to bring data to the main memory it will somehow make our system laggy .Hence to solve this problem we used generators.

Our model was predicting almost correct caption there are a few images where our model came up with absurd captions.This is because InceptionV3 have trained on a vast variety of objects but there are a few which have still left and its practically impossible to train on every image its just it got better with time as we keep on providing images to it and keep on developing it . Another next reason is som images pixels are so mixed up that it didn't even differentiate between the two objects hence coming up with different captions.

CHAPTER- 5 CONCLUSION AND FUTURE WORK

5 . 1 OVERVIEW

In this last chapter of our report we might want to finish up our work and talk about the work we are going to execute in the near future. So far we have we have perused various research papers which really talked about the different face image captioning procedures which has implemented up until now. We even did the comparison of various image captioning algorithms with their upsides and downsides which helped us to pick a powerful algorithm which could withstand different drawbacks .We even examined about our framework on which we are going to do the testing procedures. The project aim is basic however a viable method for recognition maintaining a strategic distance from pointless complexities that may hamper to genuine execution. We even talked about the robustness of different algorithms with their memory prerequisites and their handling time. In the whole project work we were centered around extracting highlights from pictures and subsequent to considering different research papers we chose to build up a customary, basic yet an effective algorithm for gender

recognition. We can land to this point after the end of this report higher discovery rates are conceivable.

5.2 FURTURE SCOPE

This is a basic simple solution and a lot of modidfications can be made in this like we could have use large data set , architechture could be changed , we could have played with hyper parameters more (like batch size , learning rate ,number of units , dropout rate, etc) , could have used CV set for overfitting , Instead of greedy search we could have used Beam search which traverse through relevant features only , using some other scoring factors It can be further broadened to the real time image captioning which will help people with impairment, self driving cars.For now there are some use cases when the objets of same calss appear there are chances that captioning will might not be upto the expectation. For example – when passing two buckets In which on contain apple and other contain orange then it captioned both as apple. So this can be solved by applying rigorous algorithms and passing more dataset to the training model.

5. 3 APPLICATIONS

The very first application is the self driving cars In this world of automation we can help car bot to move around as here it will does this all in realtime with greater speed and will check which path to move on. Helping the blind , its really hard for these people to move aroung especially crossing high jammed roads here like we can help them by this this will tell surrounding description and ultimately help in moving . CCTV cameras can also be modified to caption the recording .

HSR1

ORIGINALITY REPORT

5%	1%	0%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Antonine University Student Paper	2%
2	Submitted to La Trobe University Student Paper	1%
3	Submitted to London School of Commerce Student Paper	<1%
4	Submitted to University of Auckland Student Paper	<1%
5	vtucs.com Internet Source	<1%
6	Submitted to University of Northumbria at Newcastle Student Paper	<1%
7	s-space.snu.ac.kr Internet Source	<1%
8	Submitted to K. J. Somaiya College of Engineering Vidyavihar, Mumbai Student Paper	<1%

9	Submitted to The University of Buckingham Student Paper	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

HSR1

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

PAGE 28

PAGE 29

PAGE 30

PAGE 31

PAGE 32

PAGE 33

PAGE 34

PAGE 35

PAGE 36

PAGE 37

PAGE 38

PAGE 39

PAGE 40

PAGE 41

PAGE 42

PAGE 43

PAGE 44

PAGE 45

PAGE 46

PAGE 47

PAGE 48

PAGE 49

PAGE 50

PAGE 51

PAGE 52

PAGE 53

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

PLAGIARISM VERIFICATION REPORT

Date: 25-9-20 **Yes**

Type of Document (Tick): | **PhD** | **M.Tech Dissertation/** | **B.Tech Project** | **Pape**

Name: Hemant **Department: CSE** **Enrolment No 161337**

Contact No. 9805088466

E-mail. Chemant077@gmail.com

Name of the Supervisor: Mrs Monika Bharti

Title of the Thesis/Dissertation/Project

Report/Paper (In Capital letters): Object detection and audio feedback

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

Complete Thesis/Report Pages Detail:

- Total No. of Pages = 59
- Total No. of Preliminary pages = 9
- Total No. of pages accommodate bibliography/references = 5

Hemant

(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found **Similarity Index** at …5….. (%). Therefore, we
are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

Monika Bharti

(Signature of Guide/Supervisor) **Signature of HOD**

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

Copy Received on	**Excluded**	**Similarity Index (%)**	**Generated Plagiarism Report Details** (Title, Abstract & Chapters)	
	• All Preliminary Pages • Bibliography/Images/Quotes • 14 Words String		Word Counts	
Report Generated on			Character Counts	
		Submission ID	Total Pages Scanned	
			File Size	

Checked by

Name & Signature **Librarian**

.....

Please send your complete thesis/report in (PDF) with Title Page, Abstract and Chapters in (Word File) through the supervisor at plagcheck.juit@gmail.com