|  |  |
| --- | --- |
| Index | Summary |
| 1 | **Introduction to Deep Learning**  \*ML or DL has one common purpose that is *to introduce perception or mimic the perception abilities of humans.*  \*Perception: the ability to become aware of something through the senses.  \*Vision: How human understands what he sees.  \*Speech: Communication using vocalization, deliver message by speaking.  \*Ear (Speech Recognition): Receive and understand the spoken message.  \*Text: Understanding the written message.  \*Deep Learning have improved the above-mentioned perceptions for a computing device. |
| 2 | **Importance of Data for Deep Learning Algorithms**  \*In modern world, data is considered as oil.  \*Algorithms of AI such as face identification, recognition, smile identification, action identification and so on, is done well with the help of data.  \*If we have 5000 instance or rows of data of a particular category then we can prepare a classifier using this data.  \*When instances of labeled data such as images approaches to 10million then computer gain quality extent of perception approaching the human perception.  Annotated data: Labeled data |
| 3 | **Interoperability of Deep Learning Algorithm**  \*All the big commercial enterprise such as google, Facebook etc. are investing in Deep Learning because these companies are oriented about the behavior of user. They apply AI algorithms on the user data to make user experience better. E.g. Facebook 10yr challenge, recommendations, friend suggestion etc.  \*Deep learning performs well on small as well as complex problems. All we need is lots of data.  \*Interoperability: Same DL algorithms, tools or techniques are applicable on various problem domains with a minor change algorithm. E.g. the same algorithm can be useful for speech recognition, Bioinformatics, audio-processing or weather data. |
| 4 | **Deep Learning Frameworks**  In deep learning framework, there are 4 blocks   1. Classification (to give classification label to data)    1. Choose a classifier such as Logistic Classifier    2. Stochastic optimization: reduce errors and improve accuracy    3. Data and parameter tuning: Fine tune the data features 2. Deep Networks (multiple layers performing individual task such as eye detection, lips detection is combined to make a face identifier)    1. DNs: Collaborate layers that perform task of similar level    2. Regularization: Understand how sub-problems will solve the parent problem 3. Computer Vision    1. Convolutions    2. Neural Networks 4. Text Processing (or any time series data)    1. Recurrent Neural Networks or RNNs    2. Embeddings    3. Longest Short-term memory or LSTM |
| 5 | **Linear Classification**  \*Classification: Assignment of specified or categorized labels to data E.g. digit recognition.  \*Logistic classifier is an S shaped binary function that classifies the given data between 0 and 1.  \*W stands for **weights.** B stands for **Bias** and it is our prior knowledge**.** X is **input vector** or **input class.** Y is **output vector.**  \*Wx + b yield scores as output vector.  \*W and B come from training data.  \*Softmax function converts scores of output vector into proper probabilities i.e. sum of all entities in output vector is 1. |
| 6  2:40 | **One Hot Encoding**  \*After applying softmax function to assign proper probabilities to output vector entities, the maximum probability is set to ‘1’ and rest to ‘0’. It is called one hot encoding.  \*It is binary classification.  \* Scaling is the technique of standardizing the features of data within a defined range. It tells us that how a particular feature is comparatively significant as compared to other features. |
| 12  2:20 | **Linear Model and Its Advantages**  There are two models used in machine learning   * Linear Model:   + Differentiate between classes using straight line.   + Can’t handle real life problems.   + Let’s break down WX + B = Y into discrete blocks.   + X is input vector that contains N inputs.   + Y is output vector that contain K outputs.   + Total number of possible parameters is (N+1)K where 1 signifies the corresponding missing value. Benefits of Linear model are:     - They have Controlled number of parameters. Due to this limit, linear model is not quite useful in real life problems. Also due to this limit, linear model is simpler and easier to handle.     - They are additive. Linear model should qualify the straight line. It will fail if we multiply two variables e.g. x\*x parabola not a linear model.     - They are stable i.e. they show stability on even low value of X so Y. Linear models are comparatively more stable. i.e there is one-to-one correspondence among X and Y.     - Their derivates are stable and constant i.e. are defined in fixed range.     - They are less complex. So, computation is easy. * Non-Linear Model |
| 13 | **Deep Learning Architecture**  \*To make a non-linear function, it should be the combination of linear functions.  \*we use linear functions because it is less computationally expensive, less complex, easy to handle.  \*Activation function is applied on linear data in order to combine linear functions.  \*The activation function in this module is RELU (Rectified Linear Unit). |
| 14 | **Deep Learning Algorithm | Chain Rule** |
| 15 | **Forward and Backward Propagation**  \*There are two types of propagation in ML   * Forward Propagation (Directed from left to right i.e. from X towards Y or Assignment of labels to input data) * Backward Propagation (Directed from right to left i.e. from Y towards X or To update weights in order to make ANN perform better. It is derivative based)   \*Propagation is stack of simple mathematical operations.  \* Backward propagation is more complex as it involves derivatives. It is 2 time more complex than forward propagation. |
| 16 | **Over Fitting and Under Fitting**  \*Overfitting: Model works on known data and can’t work on incoming unknown data.  \*Underfitting: Model can’t perform well on even known data. We need more data to overcome this problem.  \*Training time: Time taken by the model while training.  \*Validation data: Subset of training data that is used to check the model being trained  \*Cut off the training process when **Validation data** to **Training time** graph begin to decrease. This process is called early termination.  \*In order to know what model will fit on our data, there is a term called regularization.  \*Skinny jeans problem: Some of the ML models perform very good once they fit on data but the fit size is the problem.  \*Regularization: It is the process which regularizes or shrinks the coefficients towards zero. In simple words, regularization discourages learning a more complex or flexible model, to prevent overfitting  \*In regularization, we control the weight in order get the best output y.  \*Common technique is L2 regularization.  \*L’ is updated loss. L is original loss. Beta is controlling parameter.L2 norm is distance between two parameters. |
| 17 | **Deep Learning Neural Networks**  \*Andrew Ng is a renowned scholar in ML. He is co-founder of coursera. He is CS professor at Stanford. He is working in Baidu research lab as well.  \*Google Brain is a deep learning artificial intelligence research team at Google.  \*In a talk in 2013, he explained how we can use to mimic human brain to get high performance easily.  \*ML has 3 basic phases.   * Data pre-processing * Feature engineering * Classification   \*The basic difference that deep learning made was **Automation of Feature Engineering phase**. So, we don’t have to extract features manually. System automatically learns what features to extract and how to use them in a particular application. So that they concluded, It is our best shot at progress towards real AI.  \*Real AI: The more the AI closer to human brain, the more successful the system is.  \*In order to make a normal face detector, we need .5 million faces as data.  \*Loss plateau: Saddle points and local minima where system doesn’t learn any more. |
| 19 | **Examples of Deep Learning**  \*Deep learning is a supervised learning.  \*Deep learning is being used in   * Voice recognition * Biological sequencing * Text Processing * Corona virus understanding   \*There are three basic types in the scope of DL   * Classification * Regression * Anomaly detection   \*Drug repurposing: It is a drug development strategy predicated on the reuse of existing licensed drugs for new medical indications  \*Classification problem is the most important problem amongst.  \*Multinomial Classification: Giving an instance more than one labels e.g. this is a phone, this is an iPhone.  \*Regression Problem: Output is number e.g. What will be the price of house at a particular place?  \*Anomaly detection: e.g. we have many images of patient lungs samples out of which a few are covid-19 patient lungs.  \*DL performs equal to human brain in many cases and even better in a few cases. E.g. sometimes human cant differentiates between two things while DL can.  \*DL is popular particularly in now-a-days because now we have labeled data and high-performance computers i.e. GPUs. Without label, data is nearly useless.  \*IoT: Being able to update things according to the state of the art within a room. |
| 20 | **How Deep Learning Works || Deep Neural Networks**  \*DL is different from ANN because in ANN there are 2 to 3 hidden layers.  \*DL uses deep neural networks in which hidden layers range up to 150.  \*In DL we don’t need to tell or extract features explicitly. System do it automatically.  \*You should choose ML when   * Amount of data is low * Problem is simple * Hardware is not high-end i.e. no GPU   \*In human brain, neurons are attached to each other by means of synapses |
| 21 | **How to Make Application Using Deep Learning Architecture**  \*DL is probability based that is, it doesn’t answer as 0 or 1. Instead, it gives the probabilities.  \*In image processing, we have images. Images can be grayscale or color. Images have pixels.  \*Convolution is the process in which we perform some operations on a digital image.  \*RELU (Rectified Linear Unit): Normalizes values because convolution is a lossy function. It normalizes the values and put them in a range.  \*Pooling: Choosing one best option from many available options.  \*Max-Pooling: Chooses the maximum value amongst.  \*Human perceives things on high-level while computer starts from low level e.g. lighter and darker pixel values of an image.  \*In DL layers   * At first layer, image processing on for color identification is performed. * At second layer, simple shapes such as dots, lines and edges are being identified. * At third layer, complex shapes such as triangles, rectangle or polygon that are formed from simple shapes are being identified. * At fourth layer, the specific shape that is combination of complex shapes and will describe the object e.g. flower, is being identified.   \*DL moves towards output as simple shape, mediocre shape and specific shape in neural network layers while identifying an object.  \*Feature Map: It is the output of one filter applied to the previous layer.  \*DL is a sequential process and feature maps are passed from previous to next layer.  \* Activation of the Network at a Particular Level: Result of a specific layer used to activate next layer.  \*RELU: Normalizes and put the values in a range.  \*Softmax Function: Converts input values into probabilities i.e. 0 and 1.  \*Low-level features at first layer, Mid-level features at middle layers and high-level features at last layers. |
| 22  3:30 | **CNN Structure in Deep Learning || Convolution Neural Network**  \*Larger the architecture, more the repetition of convolution and pooling layers.  \*Loss: The process of classification.  \* RGB image has dimension = 3, Grayscale image has dimension = 1.  \*Filters used in CV are mostly 3x3, so that   * all components contribute evenly in the calculation because of symmetry * it is the smallest possible symmetric filter in which we get a pixel in center   \*Convolution is aka SOP (sum of products).  \*Big size filter has   * large global effect * will miss small local details * require less processing power |
| 23 | **CNN Structure in Deep Learning || Phases in Deep Learning Architecture**  \*Padding: we place 0 or 1 in the part of filter that falls outside of the image.  \*Replication: replicate the pixel values present at the border of the image.  \*Discarding: discard the border pixel values of the image while applying filter.  \*Pooling: After convolution, we have to retain the important information or pixels only in order to fast computation. Spatial pooling can be of three types   * Max pooling * Average pooling * Sum pooling   \*Flattening: Pooled feature map matrix is converted into linear vector.  \*Fully connected layer: Flattened vector is passed to fully connected layer of conventional neural network.  \*Output layer: scores yielded by fully connected layer.  \*Softmax function: converts input scores into proper probabilities.  \*Loss function: it is the difference between predicted class and actual class.  \*Conventional neural network: has input layer, output layer and hidden layers. |
| 24 | **When to Use Convolutional Neural Network**  \*There are 3 flavors of deep learning architecture   * Multi-layer perceptron (when we know what features to extract) * Convolutional neural networks (most popular on image data) * Recurrent neural network (on textual data, have memory)   \*If hidden layers are formed through convolution, the resulting network is called CNN.  \*CNN is most popular because it deals well matrix data e.g. pixel values, and converts input data into another format.  \*Use RNN in textual data. |
| 25 | **Long Short- Term Memory in RNN || Uses of RNN and LSTM || Sequence Prediction Problem**  \*CNN and Multilayer perceptron do not have memory.  \*RNN is used in text, natural language and other time bound (temporal) data processing.  \*Caption of an image is an application of RNN.  \*In RNN, layers are dependent. |
| 26 | **Artificial Neural Networks in Deep Learning || ANN || CNN || RNN**  \*MLP is the simplest one.  \*Mostly used when we have simple low level features.  \*Should not be used on image data. |
| 27 | **Convolutional Layer and Activation Maps in Deep Learning**  \*CNN is aka ConvNet. In CNN, we have an image and a filter. We convolute filter on an image.  \*Low level features: simple features such as edges, color, shape and texture.  \*High level features: more complex features such as eyes, nose.  \*Convolution is aka SOP (sum of products). This means the result will be a single number.  \*In image, we have h x w x d. (height x width x depth). Depth is number of channels of pixels e.g. RGB image has 3 channels so depth = 3.  \*Depth of filter will always be equal to depth of the image. Although, height and width may vary.  \*Steps involved in CNN   * Convolute the filter or multiple filters on the image to get the activation map(s) or feature map(s). It will be a matrix of smaller in size as compared to input image because convolution is a lossy operator. When we convolute filters in sequence, the size of corresponding activation map decreases. Overall depth increases. * Activation functions * Pooling (optional): After convolution, we have to retain the important information or pixels only in order to fast computation. Spatial pooling can be of three types   + Max pooling   + Average pooling   + Sum pooling * Flattening: Pooled feature map matrix is converted into linear vector. * Fully connected layer: Flattened vector is passed to fully connected layer of conventional neural network. * Output layer: scores yielded by fully connected layer. * Softmax function: converts input scores into proper probabilities. * Loss function: it is the difference between predicted class and actual class. |
| 29 | **Convolutional Layer and Activation Maps in Deep Learning (1)**  \*Output size after applying filter: (N-F) / S + 1   * N: input size * F: filter size * Stride: It is the number of pixels shifts over the input matrix. When the stride is 1 then we move the filters to 1 pixel at a time. When the stride is 2 then we move the filters to 2 pixels at a time and so on |
| 30 | **Activation Functions in Deep Learning || Linear and Non-Linear Functions**  \*To work with realistic data, Activation functions are used to convert linear data into non-linear. |
| 31  4:05 | **Activation Functions in Deep Learning | Sigmoid Activation Function | Linear and Non-Linear Function**  \*Activation functions are used for conversion of linear data/operations to non-linear data/operations.  \*Sigmoid:   * It is an S-shaped function that bounds the input values between 0 and 1.   \*We don’t use Sigmoid function in DL because   * It converts all higher and lower values into nearly single constant. * We cannot take derivative of constant values. * Values cannot be optimized because derivative of constant is zero. * It has an exponential component in the formula that is **expensive to compute.** * It is not zero-centric, so it is not useful for data pre-processing, data cleaning and PCA   \*We essentially need optimization in order to reduce loss in final output.  \*RELU is most used function in DL.  \*A straight line has two things   1. Slope 2. Intercept   \*Zero-centric: A line or curve centered (or balanced) on origin.  \*For a line or curve, it is better for it to be zero-centric. So that,   * We don’t need to do an extra calculation i.e. finding the intercept. * Used in data cleaning e.g. Principal component analysis (PCA). * Models operating on normalized data enjoy faster convergence. |
| 32 | **Activation Functions in Deep Learning || Tanh Function || Linear and Non Linear Function**  \*Tanh   * Produces output between +1 and -1. * Produces zero-centric output.   \*Disadvantages of Tanh include   * Derivative of higher values is zero. So, can’t be optimized. * Include four exponential components. So, it is **very expensive to compute.** * It has almost all disadvantages as sigmoid.   \*Linear data: The data that is organized in a linear order in which elements are linked one after the other.   * Price of Gallon with respect to its weight (Linear) * Weight of a person with respect to time (Non-Linear) |
| 33  3:58  5:30 | **Activation Functions in Deep Learning || RELU Function || Rectified Linear Unit**  \*RELU (Rectified Linear Unit) is the most used activation function because   * It is 6 times faster than sigmoid and Tanh. * Can be optimized because derivative is not zero for x > 0.   \*Rectified means corrected.  \*In negative region it is 0 so, it is called **Dead Relu.**  \*It is not zero-centric.  \*Leaky Relu is the solution of Dead Relu  \*Parametric Relu: Value of alpha is learnt while training. |
| 34 | **CNN Architecture in Deep Learning || 1\*1 Convolution Filters || LeNet-5**  \*In a typical CNN architecture, there are two parts.   * Part one   + Convolution operations (Application of filters)   + Activation functions (Convert linear data to non-linear data)   + (optional) Pooling (Choosing one best option from many available options) * Part two (Optional)   + Fully connected Layers that converts the architecture into neural networks.   \*Without thinking a lot and working hard, we can use the existing architecture for our problems.  \*Output size is reduced after conventional convolution.  \*1x1 convolution is important when we want to   * Keep the size of input and output same. * Reduce the depth. So, less computation power is required in future calculations |
| 35 | **AlexNet CNN Architecture in Deep Learning || Convolution Filters**  \*ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is an object recognition challenge in which different classes of objects are given.  \*AlexNet was first DL architecture in which concept of layers was evolved. |
| 36 | **AlexNet CNN Architecture in Deep Learning (1) || Convolution Filters**  \*Loss: Difference between actual and predicted value.  \*To reduce loss, weights are adjusted while training. Also called parameter training.  \*Learning is done in convolution, not in pooling. |
| 37 | **ZF Net Architecture in Deep Learning || Deeper Networks || CNN Architectures**  \*There are two types of DL frameworks   * Wider networks * Deeper networks   \*By reducing filter size, no. of parameters are reduced.  \*In ZFnet,   * Smaller size filters where used for low-level features. * Larger size filters where used for high-level features.   \*In VGG,   * Layers were increased. * Size of filters were reduced. |
| 38 | **GoogLeNet Architecture || CNN Architectures** |
| 39  0:00 | **ResNet Architecture || CNN Architectures**  \*There are two types of errors   1. Training error 2. Testing error   \*Vanishing gradient: Results become minute while moving from one layer to next layer after application of derivative.  \*ResNet gave the solution of Vanishing Gradient. In ResNet, x was added as identity to each layer.  \*Cifar (Canadian Institute for Advanced Research): A dataset for image recognition.  \*Top 5 error: Output will be considered correct even if it fall any of top 5 related classes. |