**Interview Questions**

**(https://chandrasekhargundluru.medium.com/)**

[**https://pub.towardsai.net/16-interview-questions-that-test-your-machine-learning-skills-part-1-52cd58b64fbb**](https://pub.towardsai.net/16-interview-questions-that-test-your-machine-learning-skills-part-1-52cd58b64fbb)

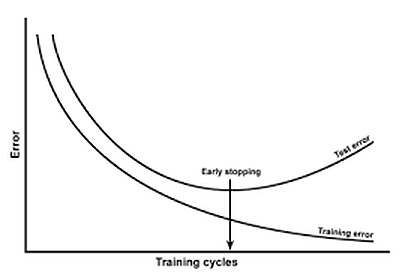
[**https://pub.towardsai.net/16-interview-questions-that-test-your-machine-learning-skills-part-2-386bf3ca0caf**](https://pub.towardsai.net/16-interview-questions-that-test-your-machine-learning-skills-part-2-386bf3ca0caf)

[**https://pub.towardsai.net/best-and-worst-cases-of-machine-learning-models-part-1-36cdb9296611**](https://pub.towardsai.net/best-and-worst-cases-of-machine-learning-models-part-1-36cdb9296611) **--- Diff between ML models**

[**https://pub.towardsai.net/16-interview-questions-every-machine-learning-enthusiast-should-know-a4142d5e00cc**](https://pub.towardsai.net/16-interview-questions-every-machine-learning-enthusiast-should-know-a4142d5e00cc)

1) What is early stopping in deep learning?

* Early stopping is a regularization technique. Overtraining a model on a dataset will cause overfitting.
* Therefore it is required to stop the training when the model starts to overfit. This process of stopping the training early is called early stopping. In early stopping, hyperparameters could be no. of epochs, generalization error, cross-validation, model performance.



## 2) Why does regularization help reduce overfitting?

* Regularization focuses on restricting the model weight. Due to the weights being restricted whenever there is a slight change in input data, the output doesn’t change drastically.
* Though this procedure results in adding bias terms, the model’s variance is reduced, thus generalizing the overall model.

## 3) Which is the most interpretable ML/DL model in your opinion, and why?

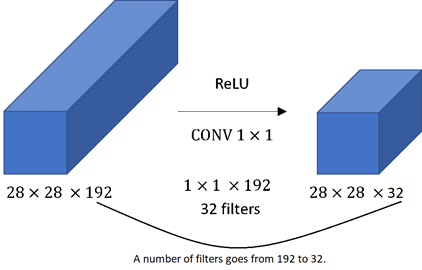
* decision tree algorithms as the most interpretable in AI. It enjoys the ‘most white box’ status from the majority of data scientists too. Because decision trees are order collection of values (data in this case), it’s visualizable ability is the highest as it performs true or false operations.
* Further, entropy used for splitting represents the disorders in the real world.

## 4) Why is bag-of-words not applied to deep learning?

* Bag-Of-Words (BoW) is a tokenization method that does not preserve the order. It generates a token that is understood as text and not as a sequence. Further, the structure of the sequence also won’t be preserved.
* It gives problems while doing feature engineering. So, BoW is typically used in traditional machine learning and not used in deep learning.

## 5) What is the need of using 1x1 convolution filters?

* While pooling reduces height and width, there was no efficient technique available to reduce the depth.
* The depth could be reduced by using filters with padding but using 3x3 or 5x5 filters is quite expensive in terms of training and thus in order to reduce the depth efficiently 1x1 convolution was introduced.



## 6) How to handle low-quality data?

* Data science is driven by data. Low-quality data is a common scenario we can tackle by the following means
* 1) Data cleaning: missing value imputations, outlier handling, balancing data through sampling
* 2) Data augmentation: Generating more data from existing and available data.
* 3) Data risk analysis: we should establish data reliability and risk involved

## 7) Why does SGD with momentum outperform the traditional SGD optimizer?

* The problem with SGD is that it highly oscillates on its path towards global minima. Thus it might happen it gets stuck on some local minima or might take some time to converge.
* In SGD with momentum, we have added a momentum term so that the overall gradient depends on both the current and previous gradients. Thus causing faster convergence and reduction of oscillation.

## 8) What is a noise injection?

* Noise injection is a method of regularization. It involves indirectly penalizing the complex models by adding noise to the training data.
* The added noise is selected based on the chosen kernel and acts as an artificial noise, making it difficult for the model to train. Generally, noise added is selected from a Gaussian distribution, and mean and standard deviation acts as a hyperparameter.

## 9) Why are ResNets not affected by the depth of the network?

* As the depth of the network increases, vanishing/exploding gradients become significant and affect the accuracy of the model.
* To curb this issue, ResNets introduced the concept of the residual network. Here, we use skip connections.
* The advantage of skip connection is that if there is any layer in the network which is hurting our model performance then it will be skipped by regularization. Thus allowing us to increase the depth of the model without worrying about accuracy.

## 10) What is transpose convolution?

* Transpose convolution is used for upsampling in the neural network when we require to increase the height and width of input.
* This operation is generally required in the variational autoencoders and generative adversarial networks.
* Transposed convolution layers require weights to be learned first before performing convolution operation. Therefore traditional interpolation techniques are faster, but their results are not as efficient.

## 11) What is BLEU Score?

* BLEU is the acronym for Bilingual Evaluation Understudy. It is used to evaluate the sentence generated using DNN with respect to a reference sentence.
* It ranges from 0 to 1, where 1 signifies a perfect match and 0 signifies a perfect mismatch.
* BLEU score is easy to compute and understand, is language-independent, has a high correlation with human evaluation, and is widely adopted. BLEU score use case extends to language generation, image caption generation, text summarization, and speech recognition.

## 12) What are the conditions for proper initialization of the weights in deep neural networks?

* Proper weight initialization is essential because it decides the course of training our model takes. A wrong initialization may take a much longer time for convergence or even not converge at all. One of the leading causes of slow convergence is exploding/vanishing gradients. To avoid this, any weight initialization must have the following rules:
* Activations mean should be zero across every layer of the network.
* The variance of the activations should stay constant across every layer of the network.

## ****1. Explain the****68–95–99 rule in normal distribution****?****

* As shown in the image below. nearly 68% of the data is within 1 standard deviation (σ) from the mean (μ), nearly 95% of the data is within 2 standard deviations (σ) from the mean (μ), and nearly 99.7% of the data is within 3 standard deviations (σ) from the mean (μ).

## ****2) What are the non-parametric models in machine learning?****

* **Nonparametric models** are those that do not make strong assumptions about the mapping function from independent features to dependent. Ex: KNN.
* These models assume that the data distribution cannot be defined in terms of a finite set of features.
* Nonparametric learning does not have a training phase. They utilize a similarity metric to make predictions.

## ****4) How do ResNets solve the vanishing gradient problem?****

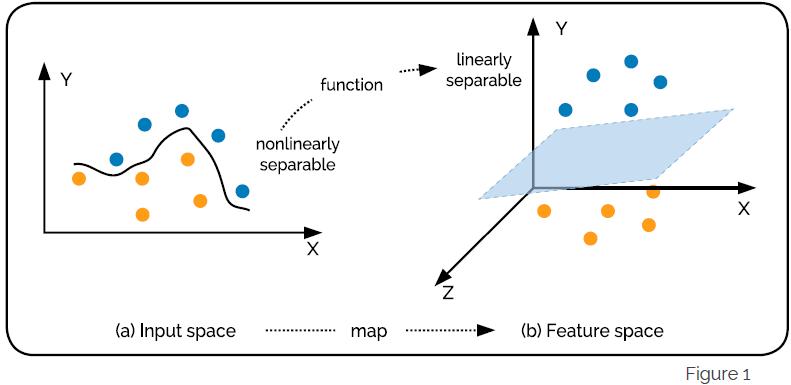
* A ResNet solves the vanishing gradient problem by using skip functions. By only using vector addition, ResNet backpropagates through the identity function.
* Weights will be altered till the output becomes equivalent to the identity function. The gradient is multiplied by one, and the result is maintained in previous layers. Here, the identity function preserves the gradients from vanishing.

## ****5) Why is Mean Absolute Error preferred over Root Mean Squared error loss functions?****

* In MAE each difference will be given equal weightage, which deals with unnecessary biasing in computation. On the other hand, RMSE is strict on larger values and penalizes them.
* In MAE, we indirectly find the median, but in RMSE, it is the median we find. So, if we want a prediction that is robust to outliers, we should use RMSE.

## ****6) What is a Kernel trick in SVM?****

* SVM became the most popular algorithm because of its kernel trick. Kernel trick is simply a dot product of two vectors x and y in high dimensional feature space. Therefore, they are also known as **“generalized dot product**”.
* The kernel trick avoids explicit mapping needed to get linear learning algorithms to learn a nonlinear decision boundary. It can be thought of applying feature transformations internally.



## 7) Why has Adam considered the best optimizer in training a Deep Neural Network?

* An optimizer is used to change the weights and learning rate in a neural network to reduce the losses. There are several types of optimizing algorithms like Gradient Descent, AdaGrad, AdaDelta, Stochastic Gradient Descent, Adam…..
* Although Adam is computationally expensive, it keeps track of the exponentially decaying average of past gradients. Besides that, it converges fast and rectifies the vanishing learning rate and high variance.

## 8) What are anchors in Faster RCNN?

* Anchors are fixed-size rectangles defined over the last convolutional feature map and fed to the Region Proposal Network. For each anchor, RPN predicts the probability of the object being present in it, and the dimensions of anchors have adjusted accordingly.
* The parameters of anchors are also included in the loss function and updated during the training. For positive anchors, all anchors’ parameters are updated, and for negative anchors, only classification loss is considered.

## ****9) How is a Convolution Neural Network different from a standard Multi-Layer Perceptron?****

* The difference between a standard Multi-Layer Perceptron (MLP) and Convolution Neural Network (CNN) is the Convolutional Layer as a hidden layer. CNN can have a non-convolutional layer as well, but the basis of CNN is the convolutional layer.
* Like other layers, a convolutional layer receives input, transforms it, and passes the transformed tensor into the next layer. Here this transformation is the convolution operation. Convolution layers detect patterns in the images. Convolution layers have filters that detect the patterns.

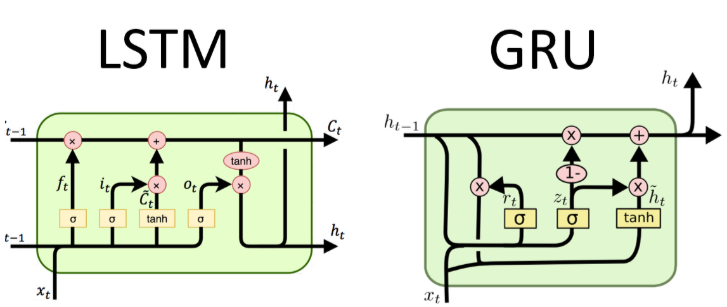
## 10) What are the disadvantages of YOLO?

Some of the disadvantages of YOLO are as follows:-

* They have low accuracy in detecting small objects.
* Since each grid only proposes a limited number of bounding boxes, they struggle in detecting close objects.
* High localization error.
* Low precision for medium and large size objects.

## 11) Which is faster between LSTM and GRU? Explain.

* Typically, Gated Recurrent Units (GRU) are quicker than Long Short-Term Memory (LSTM).
* Because LSTM architecture is a bit complex compared to the GRUs because of its additional forget gate as a memory unit. It retains the information across the time-steps.
* However, GRU combines input gates and forget gates to form an update gate. Hence, GRU is less complicated and consequently faster.



## 12) How do you calculate threshold reconstruction error in autoencoder for anomaly detection?

Fixing a value of autoencoder’s threshold in anomaly is challenging and tricky. It is highly dependent on the business problem and the required KPIs. There are broadly two approaches for finding the thresholds.

* When we have labeled data, based on the percentile of anomalies, we fix the threshold.
* When we don’t have labeled data, we find it by optimizing against the loss function.

## 13) Why is feature engineering less important in Deep learning, unlike traditional machine learning?

* Data scientists perform feature engineering to increase the quality of data representation.
* Features are transformed into a new representation. Traditional ML algorithms’ hypothesis space is not rich enough to learn hypothesis space on its own.
* This necessitates a data scientist to perform feature engineering on his own before feeding data to an algorithm. However, deep learning models learn intricate patterns without any assistance. This property solves a practitioner’s problem to a great extent.

## 14) How do deep learning models outperform traditional models in time series analysis?

Deep learning models such as DeepAR outperform traditional time series models hands down by incorporating the following factors.

1. Considers related variables that impact prediction accuracy. For example price

2. Takes metadata into account to handle seasonality

3. Handles new items with no historical data. For example: If Adidas shoes are new, it will consider Nike shoe data (similar data available)

## ****16)What are the advantages of MobileNet?****

* MobileNet has taken over the mobile deep learning industry by storm since its release. The main advantages of using MobileNet are:-
* It has very few numbers of parameters.
* They are lightweight DNNs and can be used with embedded devices
* They are faster, small in size, and have low latency.
* They have a wide range of applications including object detection, classification, recognition, and localization.