1. What is padding
2. Sigmoid Vs Softmax
3. What is PoS Tagging
4. What is tokenization
5. What is topic modeling
6. What is back propagation
7. What is the idea behind GANs
8. What is the Computational Graph
9. What is sigmoid What does it do
10. What is Named-Entity Recognition
11. Explain the masked language model
12. How do you preprocess text in NLP
13. How do you extract features in NLP
14. How is wordvec different from Glove
15. What Are the Different Layers on CNN
16. What makes CNNs translation invariant
17. How is fastText different from wordvec
18. Explain Generative Adversarial Network
19. What is backward and forward propagation
20. What are Syntactic and Semantic Analysis
21. What is a local optimumWhat is a local optimum
22. Explain gates used in LSTM with their functions
23. What is ReLU How is it better than sigmoid or tanh
24. What is transfer learning have you used it before
25. What is multi-task learning When should it be used
26. Difference between convex and non-convex cost function
27. Why do we remove stop words When do we not remove them
28. Explain the difference between an epoch a batch and an iteration
29. What is the difference between NLP and NLU
30. For online learning which one would you prefer SGD or Adagrad and why
31. What Is a Multi-layer Perceptron MLPWhat Is a Multi-layer Perceptron MLP
32. Is it always bad to have local optimaIs it always bad to have local optima

33. In node2vec, what does embedding represent topological similarity or nearness
34. What do you understand by Boltzmann Machine and Restricted Boltzmann Machines
35. How to compute an inverse matrix faster by playing around with some computational tricks
36. For infrequent/rare words which among CBOW and SkipGram should be used for wordvec training
37. What is pooling in CNN Why do we need it
38. Describe the structure of Artificial Neural Networks & RNN(recurrent neural network)
39. How to Select a Batch Size Will selecting a batch size produce better or worse results?
40. What are N-grams How can we use them
41. How large should be N for our bag of words when using N-grams
42. How can you use neural nets for text classification and computer vision
43. Do gradient descent methods always converge at the same point
44. What is gradient descent How does it work
45. What are autoencoders Explain the different layers of autoencoders and mention three practical usages of them
46. What is vanishing gradient descent
47. difference between Vanishing gradient Vs Exploding gradient
48. How to handle dying node problems in case of ReLU activation function
49. What is the use of the leaky ReLU function
50. What are the different Deep Learning Frameworks
51. What is the difference between machine learning and deep learning
52. What is a dropout layer and how does it help a neural network
53. Explain why dropout in a neural network acts as a regularizer
54. How to know whether your model is suffering from the problem of Exploding Gradients
55. How to handle exploding gradient problem
56. How Does an LSTM Network Work
57. What problem does Bi-LSTM solve instead of only LSTM
58. What is the difference between LSTM and GRU
59. What happens to the predictions of a CNN if an image is rotated
60. How does CNN help in translation and rotation invariance of images
61. Define Term Freuency & Inverse Document Freuency Tf-idf and how to use it for converting text to vector
62. What are three primary convolutional neural network layers How are they commonly put together

63. Describe the architecture of a typical Convolutional Neural Network
64. What do you mean by Dropout and Batch Normalization, When and why use
65. What is the difference between online and batch learning
66. Is dropout used on the test set
67. What is an activation function and discuss the use of an activation function
68. Explain three different types of activation functions
69. What is the range of activation functions
70. Why is Rectified Linear Unit a good activation function
71. Why don't we use the Relu activation function in the output layer
72. What can go wrong if we use a linear activation instead of ReLU
73. Give examples in which a many-to-one RNN architecture is appropriate, Give examples in which a many-to-one RNN architecture is appropriate
74. What is RNN and How does an RNN work
75. Why Sigmoid or Tanh is not preferred to be used as the activation function in the hidden layer of the neural network
76. difference between various Activation functions such as Sigmoid , tanh, Softmax, ReLU, Leaky ReLU
77. Why Tanh activation function preferred over sigmoid
78. What are word embeddings Why are they useful
79. what is WordVec
80. What are some advantages of using character embeddings instead of word embeddings
81. How do you get sentence meanings from word embeddings, considering the position of words in the sentence
82. Would you prefer gradient boosting trees model or logistic regression when doing text classification with bag of words
83. What is bag of words How we can use it for text vectorization
84. What are the advantages and disadvantages of bag of words
85. What is the main difference between Adam and SGD
86. What are the advantages and disadvantages of SGD over gradient descent
87. What is the difference between stochastic gradient descent SGD and gradient descent GD, Batch gradient descent, Stochastic gradient descent, Mini-batch gradient descent , what are the pros and cons for each of them
88. When would you use GD over SDG and vice-versa

89. How would you choose the number of filters and the filter size at each CNN layer
90. How can we use CNN for text classification
91. What are some advantages in using a CNN (convolutional neural network rather than a DNN (dense neural network in an image classification task
92. Describe two ways to visualize features of a CNN in an image classification task
93. Why do segmentation CNNs typically have an encoder-decoder style / structure
94. What is a convolutional layer & Why do we actually need convolutions Can we use fully-connected layers for that
95. What are the advantages of parameter sharing in case of convolution
96. Why do we use convolutions for images rather than just Fully Connected layers
97. Why would you use many small convolutional kernels such as x rather than a few large onesWhy would you use many small convolutional kernels such as x rather than a few large ones
98. Why we generally use Softmax non-linearity function as the last operation in-network
99. How does BatchNormalization differ in training and inferencing
100. How does batch size affect training of neural networks
101. When using mini batch gradient descent, why is it important to shuffle the data
102. Give a simple mathematical argument why a mini-batch version of such ML algorithm might be computationally more efficient than a training with full data set
103. On a simplified and fundamental scale what makes the newly developed BERT model better than traditional NLP models
104. How would you initialize weights in a neural network
105. Why weights are initialized with small random numbers in a neural network What happens when weights are all or constant values
106. Suppose you have a NN with layers and ReLU activations What will happen if we initialize all the weights with the same value
107. What is backpropagation How does it work Why do we need it
108. Why large filter sizes in early layers can be a bad choice How to choose filter size
109. which one is more powerful a layer decision tree or a -layer neural network without any activation function --> Hint non-linearity
110. Both decision trees and deep neural networks are non-linear classifier ie they separates the space by complicated decision boundary Why then it is so much easier for us to intuitively follow a decision tree model vs a deep neural network

111. If you could take advantage of multiple CPU cores would you prefer a boosted-tree algorithm over a random forest Why