SNS COLLEGE OF ENGINEERING
Kurumbapalayam (Po), Coimbatore – 641 107
AN AUTONOMOUS INSTITUTION
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai
DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE
UNIT-2

2-Mark Questions:

1. What is the history of Deep Learning?

Answer: Deep Learning traces its origins to the 1940s and 1950s with early neural network models like the Perceptron. However, it gained significant traction in the 2000s with the rise of powerful computational resources, large datasets, and breakthrough architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs).

2. Define Backpropagation in neural networks.

Answer: Backpropagation is an optimization technique used to minimize the loss function by adjusting the weights of the neural network based on the gradient of the error with respect to each weight. It involves a forward pass followed by a backward pass to compute gradients using the chain rule of differentiation.

3. What is regularization in machine learning?

Answer: Regularization refers to techniques used to prevent overfitting by adding a penalty term to the loss function, such as L1 (Lasso) or L2 (Ridge) regularization, which penalize large weights in the model.

4. What is the purpose of batch normalization in neural networks?

Answer: Batch normalization is a technique used to normalize the activations of each layer in the network to have zero mean and unit variance. It speeds up training and helps to stabilize the learning process.

5. What does the VC dimension measure in neural networks?

Answer: The VC (Vapnik-Chervonenkis) dimension measures the capacity of a model to fit any possible set of data. It is used to assess the model's complexity and generalization ability.

6. How do deep networks differ from shallow networks?

Answer: Deep networks consist of many hidden layers, allowing them to model complex relationships in data. Shallow networks have fewer hidden layers and are less capable of capturing intricate patterns, making deep networks more suitable for complex tasks.

7. What is the key feature of Convolutional Neural Networks (CNNs)?

Answer: CNNs are specialized for processing grid-like data, such as images, and use convolutional layers to automatically learn spatial hierarchies of features, significantly improving image processing tasks.

8. What are Generative Adversarial Networks (GANs)?

Answer: GANs consist of two neural networks, a generator and a discriminator, that compete with each other. The generator creates fake data, and the discriminator attempts to distinguish between real and fake data. GANs are used for generating realistic synthetic data.

9. What is semi-supervised learning?

Answer: Semi-supervised learning is a machine learning approach that utilizes both labeled and unlabeled data for training. It leverages the small amount of labeled data along with a large amount of unlabeled data to improve learning efficiency and performance.

10. What is a probabilistic theory of Deep Learning?

Answer: A probabilistic theory of Deep Learning suggests that neural networks can be interpreted as probabilistic models that aim to learn the underlying probability distributions of the data. This approach integrates uncertainty into the model's predictions.

11. What is the role of non-linearity in deep neural networks?

Answer: Non-linearity, introduced by activation functions, allows neural networks to model complex patterns and relationships in data. Without non-linearity, the network would be limited to performing linear transformations, making it unable to handle complex tasks.

12. Explain the concept of overfitting in machine learning.

Answer: Overfitting occurs when a model learns the training data too well, including noise and outliers, leading to poor performance on unseen data. Regularization techniques and cross-validation help prevent overfitting.

13. What are hidden layers in neural networks?

Answer: Hidden layers are intermediate layers between the input and output layers of a neural network. These layers process the input data, extracting relevant features before the final output layer produces predictions.

14. What is the function of an activation function in neural networks?

Answer: The activation function introduces non-linearity into the network, enabling it to learn and approximate complex functions. Without it, the network would be a linear model, limiting its ability to solve intricate problems.

15. What is dropout regularization?

Answer: Dropout regularization involves randomly "dropping out" (setting to zero) a fraction of the neurons during training to prevent overfitting. This forces the network to learn more robust and generalized features.

16-Mark Questions:

1. Discuss the history of Deep Learning, highlighting key milestones in its development.

Answer:
Deep learning, an area of machine learning inspired by the structure of the brain, has evolved over several decades:

- 1940s-1950s: Early neural network models like the Perceptron laid the groundwork for artificial neurons.
- 1980s: The development of backpropagation and multi-layer networks revived interest in neural networks, despite limitations in computational power and data.
- 2000s: The increase in computational power, availability of large datasets, and breakthroughs in algorithms, such as deep belief networks, made deep learning a powerful tool.

- 2012: The landmark success of AlexNet in the ImageNet competition demonstrated the power of deep convolutional networks, leading to a surge in deep learning research.
- Present: Ongoing advancements in network architectures (e.g., GANs, Transformers), training methods, and large-scale pre-trained models (e.g., GPT, BERT) have pushed deep learning to new heights.

2. Explain the process of Backpropagation and its role in training neural networks.

Answer:
Backpropagation is an algorithm used to minimize the loss function by adjusting the weights of a neural network. It involves:

1. Forward Pass: Input data is passed through the network to compute the output.
2. Loss Calculation: The difference between the predicted and actual output is computed using a loss function.
3. Backward Pass: The error is propagated backward through the network using the chain rule of calculus to compute the gradients of the loss function with respect to the weights.
4. Weight Update: The weights are updated using an optimization algorithm like Gradient Descent. The process is repeated iteratively, allowing the network to learn from data and improve its performance.

3. Discuss the concept of regularization in neural networks. How do techniques like L2 regularization, dropout, and batch normalization help prevent overfitting?

Regularization

A fundamental problem in machine learning is how to make an algorithm that will perform well not just on the training data, but also on new inputs. Many strategies used in machine learning are explicitly designed to reduce the test error, possibly at the expense of increased training error. These strategies are known collectively as regularization.

Definition: - "any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error."

❖ In the context of deep learning, most regularization strategies are based on regularizing estimators.

❖ Regularization of an estimator works by trading increased bias for reduced variance.

An effective regularizer is one that makes a profitable trade, reducing variance significantly while not overly increasing the bias.

- Many regularization approaches are based on limiting the capacity of models, such as neural networks, linear regression, or logistic regression, by adding a parameter norm penalty $\Omega(\theta)$ to the objective function J. We denote the regularized objective function by $\tilde{J}$

$$\tilde{J}(\theta; X, y) = J(\theta; X, y) + \alpha\Omega(\theta)$$

where $\alpha \in [0, \infty)$ is a hyperparameter that weights the relative contribution of the norm penalty term, Ω, relative to the standard objective function J. Setting α to 0 results in no regularization. Larger values of α correspond to more regularization.

- The parameter norm penalty Ω that penalizes only the weights of the affine transformation at each layer and leaves the biases unregularized.

L2 Regularization

One of the simplest and most common kind of parameter norm penalty is L2 parameter & it's also called commonly as weight decay. This regularization strategy drives the weights closer to the origin by adding a regularization term . L2 regularization is also known as ridge regression or Tikhonov regularization. To simplify, we assume no bias parameter, so θ is just w. Such a model has the following total objective function.

- We can see that the addition of the weight decay term has modified the learning rule to multiplicatively shrink the weight vector by a constant factor on each step, just before performing the usual gradient update. This describes what happens in a single step. The minimum of $\hat{J}$ occurs where its gradient $\nabla_w \hat{J}(w) = H(w - w*)$ is

equal to '0' To study the eff ect of weight decay,

- As α approaches 0, the regularized solution $\tilde{w}$ approaches w*. But what happens as α grows? Because H is real and symmetric, we can decompose it into a diagonal matrix Λ and an orthonormal basis of eigenvectors, Q, such that $H = Q\Lambda Q^T$. Applying Decomposition to the above equation, We Obtain

The solid ellipses represent contours of equal value of the unregularized objective. The dotted circles represent contours of equal value of the L 2 regularizer. At the point $\tilde{w}$, these competing objectives reach an equilibrium. In the first dimension, the eigenvalue of the

Hessian of J is small. The objective function does not increase much when moving horizontally away from w∗ . Because the objective function does not express a strong preference along this direction, the regularizer has a strong effect on this axis. The regularizer pulls w1 close to zero. In the second dimension, the objective function is very sensitive to movements away from w∗ . The corresponding eigenvalue is large, indicating high curvature. As a result, weight decay affects the position of w2 relatively little.

L1 Regularization

While L2 weight decay is the most common form of weight decay, there are other ways to penalize the size of the model parameters. Another option is to use L1 regularization.

21

➢ L1 regularization on the model parameter w is defined as the sum of absolute values of the individual parameters.

L1 weight decay controls the strength of the regularization by scaling the penalty Ω using a positive hyperparameter α. Thus, the regularized objective function J˜(w; X, y) is given by

By inspecting equation 1, we can see immediately that the effect of L 1 regularization is quite different from that of L 2 regularization. Specifically, we can see that the regularization contribution to the gradient no longer scales linearly with each wi ; instead it is a constant factor with a sign equal to sign(wi).

2Difference between L1 & L2 Parameter Regularization

➢ L1 regularization attempts to estimate the median of data, L2 regularization makes estimation for the mean of the data in order to evade overfitting.

22

➢ L1 regularization can add the penalty term in cost function. But L2 regularization appends the squared value of weights in the cost function.

➢ L1 regularization can be helpful in features selection by eradicating the unimportant features, whereas, L2 regularization is not recommended for feature selection

➢ L1 doesn't have a closed form solution since it includes an absolute value and it is a non differentiable function, while L2 has a solution in closed form as it's a square of a weight

4. Explain the concept of Generative Adversarial Networks (GANs) and their applications.

Answer:
Generative Adversarial Networks (GANs) consist of two networks: a generator and a discriminator. The generator creates fake data samples, while the discriminator attempts to distinguish between real and fake data. The two networks are trained simultaneously in a game-like setting where the generator aims to fool the discriminator, and the discriminator tries to correctly classify the data.

- Applications: GANs have a wide range of applications, including:
    - Image generation: Creating realistic images, artwork, or even deepfake videos.
    - Data augmentation: Generating synthetic data to augment real-world datasets.
    - Super-resolution: Enhancing the resolution of images or videos.
    - Style transfer: Generating new images in the style of another artist or domain.

5. Compare Deep Networks and Shallow Networks. Discuss the advantages and disadvantages of deep networks.

Answer:

- Shallow Networks: Have a small number of layers, typically just one or two hidden layers. While simpler and faster to train, they are less capable of modeling complex patterns in data.
- Deep Networks: Have multiple hidden layers, allowing them to model hierarchical representations of the data. Deep networks are highly flexible and capable of learning complex patterns from large datasets.
    - Advantages of Deep Networks:
        - Ability to model complex, hierarchical representations.
        - Effective for tasks like image recognition, language processing, and reinforcement learning.
        - Can achieve state-of-the-art performance in many domains.
    - Disadvantages:
        - Require large amounts of data and computational resources.
        - Training can be slow and prone to overfitting if not properly regularized.
        - Susceptible to issues like vanishing gradients and the need for careful initialization and optimization strategies.

6.Explain Batch Normalization.

Batch Normalization:

It is a method of adaptive reparameterization, motivated by the difficulty of training very deep models.In Deep networks, the weights are updated for each layer.

So the output will no longer be on the same scale as the input (even though input is normalized).Normalization - is a data pre-processing tool used to bring the numerical data to a common scale without distorting its shape.when we input the data to a machine or deep learning algorithm we tend to change the values to a balanced scale because, we ensure that our model can generalize appropriately.(Normalization is used to bring the input into a balanced scale/ Range).

Even though the input X was normalized but the output is no longer on the same scale. The data passes through multiple layers of network with multiple times(sigmoidal) activation functions are applied, which leads to an internal co-variate shift in the data.

This motivates us to move towards Batch Normalization

Normalization is the process of altering the input data to have mean as zero and standard deviation value as one.

Procedure to do Batch Normalization:

(1) Consider the batch input from layer h, for this layer we need to calculate the mean of this hidden activation.
(2) After calculating the mean the next step is to calculate the standard deviation of the hidden activations.
(3) Now we normalize the hidden activations using these Mean & Standard Deviation values. To do this, we subtract the mean from each input and divide the whole value with the sum of standard deviation and the smoothing term (ε).
(4) As the final stage, the re-scaling and offsetting of the input is performed. Here two components of the BN algorithm is used, γ(gamma) and β (beta). These parameters are used for re-scaling (γ) and shifting(β) the vector contains values from the previous operations.

These two parameters are learnable parameters, Hence during the training of neural network, the optimal values of γ and β are obtained and used. Hence we get the accurate normalization of each batch.