ELEC60015 Deep Learning Revision Questions

Timothy Chung

Spring 2024

Contents

Chapter 1

Questions from Mentimeter and Slides

2-3 CNNs and Network Training

1. In CNN, a filter is applied to:
    a. Each channel separately
    b. All channels across the layer
    c. All layers across the network
2. Stride is:
    a. Step with which a filter is applied
    b. Slide of the receptive field
    c. Number of channels a filter is applied to
3. Learning Rate is:
    a. Rate of convergence
    b. Step of weight's update
    c. Param learnt by SGD
4. Weights are not updated once per:
    a. Batch
    b. Iteration
    c. Epoch
5. All training data is used to update weights in one:
    a. Epoch
    b. Batch
    c. Iteration
6. Averaging updates over iterations is referred to as:
    a. Decay
    b. Momentum
    c. Dropout
7. First and second order moments of gradients are used in:

a. Adam
b. RMSProp
c. Adagrad
d. Nesterov
e. Adadelta
f. SGD momentum

8. Second order moments of gradients are used in:

a. Adam
b. RMSProp
c. Adagrad
d. Nesterov
e. Adadelta
f. SGD momentum

9. Batch Normalisation is applied to:

a. Weights
b. Channels
c. Input Data

10. Dropout is an effective regularisation of:

a. Layers with small filters
b. Conv Layers
c. Fully Connected Layers

11. (*) L2 regularisation of weights is called:

a. Absolute norm
b. Momentum
c. Decay

12. Finetuning is a process of:

a. Adjusting hyperparameters on validation set
b. Updating parameters pretrained on another dataset
c. Updating parameters near convergence

13. Data Augmentation consists of:

a. Collecting more data samples
b. Generating new samples from existing data
c. Increasing the size of data samples

14. A hard negative is a:

a. Positive example similar to a negative one
b. Negative example dissimilar to a positive one
c. Negative example similar to a positive one

15. A hard positive is a:

a. Positive example similar to a negative one
b. Positive example dissimilar to a positive one
c. Positive example dissimilar to a negative one

16. To debug a model:

a. Reduce batch size and learning rate
b. Add more data, train longer
c. Overfit to a small dataset

17. Bias in a dataset is:

a. Constant offset introduced during normalisation
b. Confusing noise introduced during data collection
c. Constant offset introduced to avoid overfitting

4 CNN Architecture

1. VGG uses:

a. 5x5 filters avg pool
b. 3x3 filters max pool
c. 3x3 filters avg pool

2. VGG is used because:

a. small model size
b. computation efficiency
c. effective feature representation

3. Efficiency of 1x1 conv filters are used in:

a. VGG
b. Resnet
c. Inception

4. Inception Block uses:

a. Parallel filters with concatenated outputs
b. Parallel streams combined with FC layers
c. Parallel filters same size

5. Skip connections are used in:

a. ResNet
b. VGG
c. InceptionNet

6. Skip connections:

a. Apply only to ReLU activation
b. Apply skip operation to data
c. Do not change the data

5 RNNs

1. Best performing word embedding is:
    a. GloVe
    b. Elmo
    c. Bert
2. Which unit is least effective in remembering sequences:
    a. RNN
    b. LSTM
    c. GRU
3. Gating mechanism uses:
    a. Sigmoid
    b. Tanh
    c. ReLU
4. In GRU, hidden state and input are:
    a. Averaged
    b. Multiplied
    c. Concatenated
5. Language modelling uses architecture type:
    a. One to many
    b. Many to Many
    c. Many to One
6. Transformer's self-attention uses:
    a. LSTM units
    b. GRU units
    c. Linear projections

6 Representation Learning

True or False?

1. Dim of representation space Z should be smaller than that of input space X.
2. In an autoencoder, encoder and decoder can be asymmetric.
3. The human brain can be seen as a universal representation learner.
4. Representation can be learned using supervised learning.
5. Ideally, we would like to have the decoder to be the inverse of the encoder.
6. Which of these losses below are not robust to outliers:
    a. MSE
    b. Absolute value loss
    c. Cross Entropy

7 Generative Models

True or False?

1. All generative models, in one way or another, directly learn the data distribution
2. Likelihood-based generative models are not suitable for inpainting.
3. Generative models can be used to classify data.
4. Autoencoders can be used to generate data.

8 Reinforcement Learning

1. How to find an optimal policy using value iteration? (From slides)

2. Let $A(s) = \mathbb{E}_\pi[G_t|S_t = s]$ and $B(s) = \mathbb{E}_\pi[G_{t+1}|S_{t+1} = s]$. What is true? (From slides)
   a. $A > B$
   b. $A = B$
   c. $A < B$
   d. Not enough information to decide.

3. Taken from the 2024 past paper: Consider a Markov Decision Process with states $\{A, B\}$. The transition probabilities and rewards are given as:

   - Transition probabilities: $P(A \to A) = 0.8$, $P(A \to B) = 0.2$, $P(B \to A) = 0.4$, $P(B \to B) = 0.6$
   - Rewards: $R(A) = +5$, $R(B) = +2$.

   Assume a discount factor γ of 0.5. What are the values for states A and B?
   a. $v(A) = 9.25, v(B) = 5.5$
   b. $v(A) = \infty, v(B) = \infty$
   c. $v(A) = 0.6, v(B) = 1.2$
   d. $v(A) = 6.5, v(B) = 3.25$

Chapter 2

Custom Questions

1 Intro To Machine Learning

1. What type of noise is usually caused from measuring data?
    a. Deterministic
    b. Stochastic
2. Fitting to noise instead of the underlying target function is a sign of:
    a. Overfitting
    b. Underfitting
    c. Regularisation
3. What norm does Ridge regression use?
    a. L0
    b. L1
    c. L2

2 CNNs

1. Which of the following is not usually represented as one of the 4 dimensions in a CNN tensor?
    a. samples
    b. depth
    c. height
    d. width
    e. channel
2. Filter depth is:
    a. Step with which a filter is applied
    b. Slide of the receptive field
    c. Number of channels a filter is applied to
3. Filter padding is :
    a. The step of the shift when convolving a filter with the image input

b. Adding space around the borders of an image input to modify output feature map size
c. A layer to perform non-linear downsampling via a sliding window across the feature map

4. The pooling layer is:
   a. The step of the shift when convolving a filter with the image input
   b. Adding space around the borders of an image input to modify output feature map size
   c. A layer to perform non-linear downsampling via a sliding window across the feature map
5. The pooling layer is used for:
   a. Dimensionality increase
   b. Dimensionality reduction
   c. Adding parameters to be learnt
6. Which error has the best performance for multiclass classification with CNNs?
   a. Mean Squared Error
   b. Categorical Cross Entropy Loss
   c. Classification Error
7. Generally, it is better to use larger layers first or smaller filters first?
   a. Larger
   b. Smaller

3 Network Training

1. For a neural network with all sigmoid activation functions, what can result from saturated gradients?
   a. Exploding gradients
   b. Vanishing gradients
2. What is the difference between normal momentum-based SGD and Nesterov momentum?
   a. Nesterov momentum uses decaying moving average of the gradients of projected positions
   b. Nesterov momentum accumulates the contribution of past gradients with an additional momentum term
3. Which optimisers use the hadamard product?
   a. AdaGrad
   b. RMSProp
   c. SGD with Nesterov
   d. AdaDelta
   e. Adam
4. Which optimisers use the second moment of the gradient?
   a. AdaGrad
   b. RMSProp
   c. SGD withNesterov
   d. AdaDelta

e. Adam

5. Which form of regularisation randomly zeroes out nodes during training and scales outputs during testing?

   a. Batch Normalisation
   b. Dropout

6. Dropout: If fraction p of nodes are zeroed out during training, what should their outputs be scaled by during testing?

   a. p/l, where l refers to the count of layers
   b. p

7. Batch Normalisation: The Xavier initialisation and Kaiming He initialisation are used for what kinds of activation functions?

   a. Sigmoid for Xavier, ReLU for Kaiming He
   b. ReLU for Xavier, Sigmoid for Kaiming He
   c. Both initialisations can use either
   d. Sigmoid for both initialisations

8. To prevent the symmetry problem caused by identical gradients during backpropagation, weights should be initialised to:

   a. Zero
   b. A random sample from the uniform distribution
   c. A random sample from the Gaussian distribution
   d. A random sample from both distributions above

9. Usually it is best to place a Batch Normalisation layer:

   a. Between the Conv and Activation Layers
   b. After the FC layer
   c. Before the pooling layer

10. For 4D tensor input $[N, H, W, C]$, Batch Normalisation in convolutional networks (ConvNets) calculate the mean and standard deviations across:

    a. H
    b. W
    c. C

11. When fine-tuning a network, it is normal to use learning rate that is:

    a. larger than the original
    b. smaller than the original

12. Data splitting: What is the best ratio below for partitioning the training, validation, and test sets respectively?

    a. 60:20:20
    b. 20:60:20
    c. 20:20:60

13. Does data augmentation require collecting new data?

    a. Yes
    b. No

14. When training error is high, what should be done?

    a. Train longer, try different model
    b. Regularise, get more data

15. When validation-training error is high, what should be done?

    a. Train longer, try different model
    b. Regularise, get more data

16. When validation-testing error is high, what should be done?

    a. Allocate more validation-training data
    b. Redo hyper-parameter search

17. Which optimiser adapts the learning rate based on the accumulated square of gradients and is particularly useful for sparse data?

    a. Adam
    b. Adagrad
    c. Adadelta
    d. SGD with Nesterov momentum

18. Adadelta is an extension of which of the following optimisers?

    a. Adam
    b. Adagrad
    c. RMSprop
    d. SGD with Nesterov momentum

19. Which optimiser combines the advantages of Adagrad and RMSprop and also utilises momentum?

    a. Adadelta
    b. SGD
    c. Adam
    d. SGD with Nesterov momentum

20. Stochastic Gradient Descent (SGD) with Nesterov momentum differs from classic momentum because:

    a. It computes the gradient at the updated position rather than the current position.
    b. It uses a fixed learning rate throughout the training.
    c. It does not accumulate the gradient.
    d. It computes the gradient at a position ahead in the direction of the momentum.

21. RMSprop is designed to solve the diminishing learning rates problem found in which optimiser?

    a. Adam
    b. Adagrad
    c. Adadelta
    d. SGD with Nesterov momentum

4 CNN Architectures

1. What features did VGG have that were different from AlexNet?
    a. Number of FC layers
    b. Number of filter kernels
    c. Number of channels
    d. Activation function
2. What type of pooling layer does GoogLeNet use more often than ResNet?
    a. Average
    b. Max Pooling
        - GooGLeNet used max pooling followed by 1×1 convolution to reduce depth
3. Which model **improved** on ResNet's skip connections?
    a. ResNeXt
    b. Densenet
4. What features does ResNeXt have?
    a. Google's Inception module
    b. Kaiming He's skip connections
    c. Both Inception modules and skip connections

5 RNNs

1. What activation functions do RNNs generally use?
    a. tanh
    b. ReLU
    c. Softmax
2. What activation functions are generally used with cell states in LSTMs?
    a. tanh
    b. ReLU
    c. Sigmoid
3. What activation functions do LSTM are generally used with update and forget gates?
    a. tanh
    b. ReLU
    c. Sigmoid
4. Which is a use case of the Many-to-Many models?
    a. Image captioning
    b. Sentiment analysis
    c. Video annotation
5. What is a key feature of LSTM networks that distinguishes them from basic RNNs?

a. They have a simpler architecture that is easier to train.
b. They have feedback loops within the network layers.
c. They use gating mechanisms to control the flow of information.
d. They can only process sequence data in one direction.

6. Why are LSTMs particularly suited for processing sequences with long-range dependencies?

a. Because they can process data in parallel.
b. Because their gating mechanisms help mitigate the vanishing gradient problem.
c. Because they have a fixed-size hidden layer.
d. Because they require less training data than other RNNs.

7. Which of the following tasks is LSTM least likely to excel at?

a. Language translation.
b. Sentiment analysis.
c. Short-term time series prediction.
d. Real-time image classification.

8. In an LSTM unit, which gate is responsible for deciding what information to throw away from the cell state?

a. Input gate.
b. Output gate.
c. Forget gate.
d. Update gate.

9. LSTMs can be utilised in which of the following applications?

a. Time series forecasting.
b. Text generation.
c. Speech recognition.
d. All of the above.

6 Representation Learning

1. What kind of data does Supervised learning use?

a. Labelled data
b. Unlabelled data

2. What shows there exists no universally best single feature representation?

a. Hoeffding's inequality
b. No free lunch theorem
c. Chebyshev inequality

3. What is a solution to the linear autoencoder problem?

a. Principal Component Analysis
b. Least Squares Approximation
c. Inverse covariance matrix

4. What is an example of a Many-to-Many model?
    a. Image captioning
    b. Sentiment analysis
    c. Video annotation
5. Which type of autoencoder requires regularisation?
    a. Undercomplete
    b. Overcomplete
6. An overcomplete autoencoder only tuned to specific patterns in the data would be a:
    a. Contractive autoencoder
    b. Sparse autoencoder
7. What activation functions can a contractive autoencoder not use?
    a. Sigmoid
    b. ReLU
8. What is the primary purpose of a linear autoencoder?
    a. To classify input data into predefined categories.
    b. To learn a compressed representation of the input data.
    c. To predict future data points in a time series.
    d. To increase the dimensionality of the input data.
9. In the context of linear autoencoders, what does the term "overcomplete" refer to?
    a. The autoencoder has more layers than necessary.
    b. The autoencoder has more neurons in the hidden layer than inputs.
    c. The autoencoder has a hidden layer larger than the input layer.
    d. The autoencoder is trained on more data than necessary.
10. Which of the following is a key characteristic of a linear autoencoder for dimensionality reduction?
    a. Non-linear activation functions.
    b. Recurrent connections in the network.
    c. Absence of activation functions, resulting in a linear transformation.
    d. Convolutional layers to capture spatial hierarchies.

7 Generative Models

1. Which generative model produces sharper examples?
    a. Generative Adversarial Networks
    b. Variational Autoencoders
2. What prevents mode collapse in GANs?
    a. Wasserstein Distance
    b. Leaky ReLUs
    c. Using CNNs

3. Do GANs have an encoder?
    a. Yes
    b. No
4. Does a GAN's generator sees the data during training?
    a. Yes, it needs comparisons to produce similar data to the training data to fool the discriminator
    b. No, it only relies on the discriminator's feedback in the form of probabilities
5. What is a distinctive feature of Variational Autoencoders (VAEs) compared to traditional autoencoders?
    a. They use backpropagation for training.
    b. They only require labeled data for training.
    c. They learn a probabilistic latent space and can generate new data.
    d. They always produce binary outputs.
6. Generative Adversarial Networks (GANs) are composed of two main components, what are they?
    a. Encoder and Decoder.
    b. Generator and Discriminator.
    c. Predictor and Classifier.
    d. Supervisor and Operator.
7. Generative Adversarial Networks (GANs) are composed of two main components, what are they?
    a. Encoder and Decoder.
    b. Generator and Discriminator.
    c. Predictor and Classifier.
    d. Supervisor and Operator.
8. What role does the discriminator play in a Generative Adversarial Network (GAN)?
    a. It generates new data instances.
    b. It evaluates the authenticity of samples, distinguishing between real and generated data.
    c. It classifies data into predefined categories.
    d. It compresses the input data into a latent space.
9. In the training of a VAE, what is the purpose of the reparameterization trick?
    a. To speed up the training process.
    b. To allow backpropagation through random sampling.
    c. To reduce the number of parameters in the model.
    d. To enhance the resolution of generated images.
10. Which of the following best describes the training process of GANs?
    a. The generator maximizes the probability of the discriminator being correct.
    b. The discriminator maximizes the probability of identifying real and generated data correctly, while the generator maximizes the probability of the discriminator making a mistake.
    c. The generator and discriminator are trained in separate phases, not affecting each other.
    d. The discriminator is trained to distinguish real data from fake, while the generator is trained to fool the discriminator into thinking the fake data is real.

8 Reinforcement Learning

1. What does the Markov Decision Process tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$ contain?
    a. State set, Action set, Probability function, Reward function
    b. State set, Agent set, Probability function, Reward function
    c. State set, Action set, Probability distribution, Reward function
2. Which attributes below describe an MDP?
    a. time-independent
    b. depends on past states
    c. random process
3. Which training method looks one step ahead?
    a. Temporal Difference
    b. Monte Carlo
4. Can TD be applied to non-episodic MDPs?
    a. Yes
    b. No
5. Can MC be applied to non-episodic MDPs?
    a. Yes
    b. No
6. Which of the below works better in complex unpredictable scenarios where the environment is not known?
    a. MC
    b. TD
7. Which of the following statements best describes the difference between Q-learning and SARSA in reinforcement learning?
    a. Q-learning is an on-policy algorithm, while SARSA is an off-policy algorithm.
    b. Q-learning updates its Q-values using the greedy action selection, while SARSA updates its Q-values using the actual next action taken.
    c. Q-learning and SARSA both update their Q-values based on the actual next action taken and are considered on-policy algorithms.
    d. There is no difference; Q-learning and SARSA are simply different names for the same algorithm.
8. What distinguishes Q-learning in terms of policy type?
    a. Q-learning is an on-policy control algorithm.
    b. Q-learning is an off-policy control algorithm.
    c. Q-learning does not utilize any policy for learning.
    d. Q-learning switches between on-policy and off-policy during training.
9. In Q-learning, when is the Q-value updated?
    a. After the completion of an episode.

b. After each step within an episode.

c. Only at the end of a learning session.

d. Before the agent takes an action.

10. In SARSA, how are action values updated?

    a. Using the expected value of the next state's best action.

    b. Using the value of the next action selected by the current policy.

    c. Independently of the actions selected by the current policy.

    d. Using a random action from the next state.

11. SARSA is considered what type of learning algorithm?

    a. Deterministic

    b. Off-policy

    c. On-policy

    d. Model-based

Chapter 3

Questions from Mentimeter and Slides with Answers

2-3 CNNs and Network Training

1. In CNN, a filter is applied to:
    a. Each channel separately
    b. All channels across the layer
    c. **All layers across the network**
2. Stride is:
    a. **Step with which a filter is applied**
    b. Slide of the receptive field
    c. Number of channels a filter is applied to
3. Learning Rate is:
    a. Rate of convergence
    b. **Step of weight's update**
    c. Param learnt by SGD
4. Weights are not updated once per:
    a. Batch
    b. Iteration
    c. **Epoch**
5. All training data is used to update weights in one:
    a. **Epoch**
    b. Batch
    c. Iteration
6. Averaging updates over iterations is referred to as:
    a. Decay
    b. **Momentum**
    c. Dropout

7. First and second order moments of gradients are used in:
    a. **Adam**
    b. RMSProp
    c. Adagrad
    d. Nesterov
    e. Adadelta
    f. SGD momentum
8. Second order moments of gradients are used in:
    a. **Adam**
    b. **RMSProp**
    c. **Adagrad**
    d. Nesterov
    e. **Adadelta**
    f. SGD momentum
9. Batch Normalisation is applied to:
    a. Weights
    b. **Channels**
    c. Input Data
10. Dropout is an effective regularisation of:
    a. Layers with small filters
    b. Conv Layers
    c. **Fully Connected Layers**
11. (*) L2 regularisation of weights is called:
    a. Absolute norm
    b. Momentum
    c. **Decay**
12. Finetuning is a process of:
    a. Adjusting hyperparameters on validation set
    b. **Updating parameters pretrained on another dataset**
    c. Updating parameters near convergence
13. Data Augmentation consists of:
    a. Collecting more data samples
    b. **Generating new samples from existing data**
    c. Increasing the size of data samples
14. A hard negative is a:
    a. Positive example similar to a negative one
    b. Negative example dissimilar to a positive one
    c. **Negative example similar to a positive one**

15. A hard positive is a:
    a. Positive example similar to a negative one
    b. **Positive example dissimilar to a positive one**
    c. Positive example dissimilar to a negative one
16. To debug a model:
    a. Reduce batch size and learning rate
    b. Add more data, train longer
    c. **Overfit to a small dataset**
17. Bias in a dataset is:
    a. Constant offset introduced during normalisation
    b. **Confusing noise introduced during data collection**
    c. Constant offset introduced to avoid overfitting

4 CNN Architecture

1. VGG uses:
    a. 5x5 filters avg pool
    b. **3x3 filters max pool**
    c. 3x3 filters avg pool
2. VGG is used because:
    a. small model size
    b. computation efficiency
    c. **effective feature representation**
        - it is not efficient(note the on the CNN performance graph)
        - but we like to use it for pretrained models
3. Efficiency of 1x1 conv filters are used in:
    a. VGG
    b. Resnet
    c. **Inception**
4. Inception Block uses:
    a. **Parallel filters with concatenated outputs**
    b. Parallel streams combined with FC layers
    c. Parallel filters same size
5. Skip connections are used in:
    a. **ResNet**
    b. VGG
    c. InceptionNet
6. Skip connections:
    a. Apply only to ReLU activation
    b. Apply skip operation to data
    c. **Do not change the data**

5 RNNs

1. Best performing word embedding is:
    a. GloVe
    b. Elmo
    c. **Bert**
2. Which unit is least effective in remembering sequences:
    a. **RNN**
    b. LSTM
    c. GRU
3. Gating mechanism uses:
    a. **Sigmoid**
    b. Tanh
    c. ReLU
4. In GRU, hidden state and input are:
    a. Averaged
    b. Multiplied
    c. **Concatenated**
5. Language modelling uses architecture type:
    a. One to many
    b. **Many to Many**
    c. Many to One
6. Transformer's self-attention uses:
    a. LSTM units
    b. GRU units
    c. **Linear projections**

6 Representation Learning

True or False?

1. Dim of representation space Z should be smaller than that of input space X. **False**
2. In an autoencoder, encoder and decoder can be asymmetric. **True**
3. The human brain can be seen as a universal representation learner. **False**
    - There is no single representation feature, e.g., optical limitations.
    - Mathematically, there is no universal representation learner.
4. Representation can be learned using supervised learning. **True**
5. Ideally, we would like to have the decoder to be the inverse of the encoder. **False**
6. Which of these losses below are not robust to outliers:
    a. **MSE**
    b. Absolute value loss
    c. Cross Entropy

7 Generative Models

True or False?

1. All generative models, in one way or another, directly learn the data distribution **True**
2. Likelihood-based generative models are not suitable for inpainting. **False**
3. Generative models can be used to classify data. **True. Indirectly so, by estimating the likelihood of whether a point should belong to which class.**
4. Autoencoders can be used to generate data. **True, Variational Autoencoders especially**

8 Reinforcement Learning

1. How to find an optimal policy using value iteration? (From slides)
   Value iteration is an algorithm that finds the optimal policy π^* for a Markov Decision Process (MDP). The algorithm iteratively improves the value function $V(s)$ for all states s until it converges to the optimal value function $V^*(s)$. Once convergence is reached, the optimal policy π^* can be derived from $V^*(s)$. The steps to perform value iteration are:

   (a) Initialise $V(s)$ arbitrarily for all states s. For a terminal state, if any, set $V(s)$ to 0.

   (b) Repeat:

      i. For each state s, update the value $V(s)$ based on the expected utility of taking each possible action a, and then transitioning to the next state s', considering the reward received r and the discounted value of the next state $\gamma V(s')$. The value update rule is:

      $$V(s) \leftarrow \max_a \sum_{s',r} p(s', r|s, a)[r + \gamma V(s')]$$

      where $p(s', r|s, a)$ is the state transition probability, and γ is the discount factor.

   (c) Until $V(s)$ converges, i.e., the change in $V(s)$ is smaller than a threshold θ for all states.

   (d) Output a deterministic policy $\pi^*(s)$ that selects the action a that maximizes the expected utility for each state s:

   $$\pi^*(s) = \arg\max_a \sum_{s',r} p(s', r|s, a)[r + \gamma V(s')]$$

   The resulting policy π^* is the optimal policy for the MDP, and the value function $V^*(s)$ gives the expected return when starting in state s and following π^*.

2. Let $A(s) = \mathbb{E}_\pi[G_t|S_t = s]$ and $B(s) = \mathbb{E}_\pi[G_{t+1}|S_{t+1} = s]$. What is true? (From slides)

   a. $\mathbf{A} > \mathbf{B}$
   b. $A = B$
   c. $A < B$
   d. Not enough information to decide.

   Solution: Given that G_t is the return from time t and can be expressed as a sum of rewards as follows:

   $$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$$

   The return at time $t + 1$ is:

   $$G_{t+1} = R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \dots$$

Notice that G_{t+1} is the sum of rewards from time $t+1$ onwards and G_t includes the reward R_{t+1} plus G_{t+1} discounted by γ. Thus, we can write:

$$G_t = R_{t+1} + \gamma G_{t+1}$$

Taking the expectation of both sides given $S_t = s$, we have:

$$A(s) = \mathbb{E}_\pi[R_{t+1}|S_t = s] + \gamma \mathbb{E}_\pi[G_{t+1}|S_t = s]$$

Since $\mathbb{E}_\pi[G_{t+1}|S_t = s]$ is not conditioned on $S_{t+1} = s$, it cannot be simplified directly to $B(s)$. However, because the expectation of G_{t+1} given $S_t = s$ would include transitions from state s at time t to all possible next states at time $t+1$, and because the immediate reward R_{t+1} is not included in $B(s)$, it follows that:

$$A(s) > B(s)$$

Hence, the correct answer is:

$$A > B$$

3. Taken from the 2024 past paper: Consider a Markov Decision Process with states $\{A, B\}$. The transition probabilities and rewards are given as:

   - Transition probabilities: $P(A \to A) = 0.8$, $P(A \to B) = 0.2$, $P(B \to A) = 0.4$, $P(B \to B) = 0.6$
   - Rewards: $R(A) = +5$, $R(B) = +2$.

   Assume a discount factor γ of 0.5. What are the values for states A and B?

   a. $v(A) = 9.25, v(B) = 5.5$✓
   b. $v(A) = \infty, v(B) = \infty$
   c. $v(A) = 0.6, v(B) = 1.2$
   d. $v(A) = 6.5, v(B) = 3.25$

   Solution: Given the Bellman equations for the value of states A and B:

   For state A:

   $$v(A) = R(A) + \gamma \left[P(A \to A) \cdot v(A) + P(A \to B) \cdot v(B)\right]$$

   For state B:

   $$v(B) = R(B) + \gamma \left[P(B \to A) \cdot v(A) + P(B \to B) \cdot v(B)\right]$$

   Substituting the given values:

   For state A:

   $$v(A) = 5 + 0.5 \left[0.8 \cdot v(A) + 0.2 \cdot v(B)\right]$$

   For state B:

   $$v(B) = 2 + 0.5 \left[0.4 \cdot v(A) + 0.6 \cdot v(B)\right]$$

   Solving the system of equations we find:

   $$v(A) = 9.25$$
   $$v(B) = 5.5$$

   This results in the values for states A and B with a discount factor γ of 0.5.

Chapter 4

Custom Questions With Answers

1 Intro To Machine Learning

1. What type of noise is usually caused from measuring data?
    a. Deterministic
    b. **Stochastic**
2. Fitting to noise instead of the underlying target function is a sign of:
    a. **Overfitting**
    b. Underfitting
    c. Regularisation
3. What norm does Ridge regression use?
    a. L0
    b. L1
    c. **L2**

2 CNNs

1. Which of the following is not usually represented as one of the 4 dimensions in a CNN tensor?
    a. samples
    b. **depth**
    c. height
    d. width
    e. channel
2. Filter depth is:
    a. Step with which a filter is applied
    b. Slide of the receptive field
    c. **Number of channels a filter is applied to**
3. Filter padding is :
    a. The step of the shift when convolving a filter with the image input

b. **Adding space around the borders of an image input to modify output feature map size**

c. A layer to perform non-linear downsampling via a sliding window across the feature map

4. The pooling layer is:

a. The step of the shift when convolving a filter with the image input

b. Adding space around the borders of an image input to modify output feature map size

c. **A layer to perform non-linear downsampling via a sliding window across the feature map**

5. The pooling layer is used for:

a. Dimensionality increase

b. **Dimensionality reduction**

c. Adding parameters to be learnt

6. Which error has the best performance for multiclass classification with CNNs?

a. Mean Squared Error

b. **Categorical Cross Entropy Loss**

c. Classification Error

7. Generally, it is better to use larger layers first or smaller filters first?

a. **Larger**

b. Smaller

3 Network Training

1. For a neural network with all sigmoid activation functions, what can result from saturated gradients?

a. Exploding gradients

b. **Vanishing gradients**

2. What is the difference between normal momentum-based SGD and Nesterov momentum?

a. **Nesterov momentum uses decaying moving average of the gradients of projected positions**

b. Nesterov momentum accumulates the contribution of past gradients with an additional momentum term

3. Which optimisers use the hadamard product?

a. **AdaGrad**

b. **RMSProp**

c. SGD with Nesterov

d. **AdaDelta**

e. **Adam**

4. Which optimisers use the second moment of the gradient?

a. AdaGrad

b. **RMSProp**

c. SGD withNesterov

d. **AdaDelta**

e. **Adam**

5. Which form of regularisation randomly zeroes out nodes during training and scales outputs during testing?

   a. Batch Normalisation

   b. **Dropout**

6. Dropout: If fraction p of nodes are zeroed out during training, what should their outputs be scaled by during testing?

   a. p/l, where l refers to the count of layers

   b. **p**

7. Batch Normalisation: The Xavier initialisation and Kaiming He initialisation are used for what kinds of activation functions?

   a. **Sigmoid for Xavier, ReLU for Kaiming He**

   b. ReLU for Xavier, Sigmoid for Kaiming He

   c. Both initialisations can use either

   d. Sigmoid for both initialisations

8. To prevent the symmetry problem caused by identical gradients during backpropagation, weights should be initialised to:

   a. Zero

   b. A random sample from the uniform distribution

   c. A random sample from the Gaussian distribution

   d. A random sample from either distribution listed above

9. Usually it is best to place a Batch Normalisation layer:

   a. **Between the Conv and Activation Layers**

   b. After the FC layer

   c. Before the pooling layer

10. For 4D tensor input $[N, H, W, C]$, Batch Normalisation in convolutional networks (ConvNets) calculate the mean and standard deviations across:

    a. H

    b. W

    c. **C**

11. When fine-tuning a network, it is normal to use learning rate that is:

    a. larger than the original

    b. **smaller than the original**

12. Data splitting: What is the best ratio below for partitioning the training, validation, and test sets respectively?

    a. **60:20:20**

    b. 20:60:20

c. 20:20:60

13. Does data augmentation require collecting new data?

    a. Yes
    b. **No**
        - Data augmentation adds noise, generates variations of test samples (e.g. rotated or transformed images), replacing with synonyms (text data), but does not involve collecting new data.

14. When training error is high, what should be done?

    a. **Train longer, try different model**
    b. Regularise, get more data

15. When validation-training error is high, what should be done?

    a. Train longer, try different model
    b. **Regularise, get more data**

16. When validation-testing error is high, what should be done?

    a. **Allocate more validation-training data**
    b. Redo hyper-parameter search

17. Which optimiser adapts the learning rate based on the accumulated square of gradients and is particularly useful for sparse data?

    a. Adam
    b. **Adagrad**
    c. Adadelta
    d. SGD with Nesterov momentum

18. Adadelta is an extension of which of the following optimisers?

    a. Adam
    b. Adagrad
    c. **RMSprop**
    d. SGD with Nesterov momentum

19. Which optimiser combines the advantages of Adagrad and RMSprop and also utilises momentum?

    a. Adadelta
    b. SGD
    c. **Adam**
    d. SGD with Nesterov momentum

20. Stochastic Gradient Descent (SGD) with Nesterov momentum differs from classic momentum because:

    a. It computes the gradient at the updated position rather than the current position.
    b. It uses a fixed learning rate throughout the training.
    c. It does not accumulate the gradient.
    d. **It computes the gradient at a position ahead in the direction of the momentum.**

21. RMSprop is designed to solve the diminishing learning rates problem found in which optimiser?

    a. Adam
    b. **Adagrad**
    c. Adadelta
    d. SGD with Nesterov momentum

4 CNN Architectures

1. What features did VGG have that were different from AlexNet?
    a. Number of FC layers
    b. **Number of filter kernels**
    c. **Number of channels**
    d. **Activation function**
2. What type of pooling layer does GoogLeNet use more often than ResNet?
    a. Average
    b. **Max Pooling**
        - GooGLeNet used max pooling followed by 1×1 convolution to reduce depth
3. Which model **improved** on ResNet's skip connections?
    a. ResNeXt
    b. **Densenet**
4. What features does ResNeXt have?
    a. Google's Inception module
    b. Kaiming He's skip connections
    c. **Both Inception modules and skip connections**

5 RNNs

1. What activation functions do RNNs generally use?
    a. **tanh**
    b. ReLU
    c. Softmax
2. What activation functions are generally used with cell states in LSTMs?
    a. **tanh**
    b. ReLU
    c. Sigmoid
3. What activation functions do LSTM are generally used with update and forget gates?
    a. tanh
    b. ReLU
    c. **Sigmoid**
4. Which is a use case of the Many-to-Many models?
    a. Image captioning
    b. Sentiment analysis
    c. **Video annotation**
5. What is a key feature of LSTM networks that distinguishes them from basic RNNs?

a. They have a simpler architecture that is easier to train.
b. They have feedback loops within the network layers.
c. **They use gating mechanisms to control the flow of information.**
d. They can only process sequence data in one direction.

6. Why are LSTMs particularly suited for processing sequences with long-range dependencies?

a. Because they can process data in parallel.
b. **Because their gating mechanisms help mitigate the vanishing gradient problem.**
c. Because they have a fixed-size hidden layer.
d. Because they require less training data than other RNNs.

7. Which of the following tasks is LSTM least likely to excel at?

a. Language translation.
b. Sentiment analysis.
c. Short-term time series prediction.
d. **Real-time image classification.**

8. In an LSTM unit, which gate is responsible for deciding what information to throw away from the cell state?

a. Input gate.
b. Output gate.
c. **Forget gate.**
d. Update gate.

9. LSTMs can be utilised in which of the following applications?

a. Time series forecasting.
b. Text generation.
c. Speech recognition.
d. **All of the above.**

6 Representation Learning

1. What kind of data does Supervised learning use?

a. **Labelled data**
b. Unlabelled data

2. What shows there exists no universally best single feature representation?

a. Hoeffding's inequality
b. **No free lunch theorem**
c. Chebyshev inequality

3. What is a solution to the linear autoencoder problem?

a. **Principal Component Analysis**
b. Least Squares Approximation
c. Inverse covariance matrix

4. What is an example of a Many-to-Many model?
    a. Image captioning
    b. Sentiment analysis
    c. **Video annotation**
5. Which type of autoencoder requires regularisation?
    a. Undercomplete
    b. **Overcomplete**
6. An overcomplete autoencoder only tuned to specific patterns in the data would be a:
    a. Contractive autoencoder
    b. **Sparse autoencoder**
7. What activation functions can a contractive autoencoder not use?
    a. Sigmoid
    b. **ReLU**
        - Contractive autoencoders have a penalty that takes the squared Frobenius norm of the Jacobian of ϕ in x. It requires activation functions that are differentiable twice, and ReLU is not.
8. What is the primary purpose of a linear autoencoder?
    a. To classify input data into predefined categories.
    b. **To learn a compressed representation of the input data.**
    c. To predict future data points in a time series.
    d. To increase the dimensionality of the input data.
9. In the context of linear autoencoders, what does the term "overcomplete" refer to?
    a. The autoencoder has more layers than necessary.
    b. The autoencoder has more neurons in the hidden layer than inputs.
    c. **The autoencoder has a hidden layer larger than the input layer.**
    d. The autoencoder is trained on more data than necessary.
10. Which of the following is a key characteristic of a linear autoencoder for dimensionality reduction?
    a. Non-linear activation functions.
    b. Recurrent connections in the network.
    c. **Absence of activation functions, resulting in a linear transformation.**
    d. Convolutional layers to capture spatial hierarchies.

7 Generative Models

1. Which generative model produces sharper examples?
    a. **Generative Adversarial Networks**
    b. Variational Autoencoders
2. What prevents mode collapse in GANs?

a. **Wasserstein Distance**
b. Leaky ReLUs
c. Using CNNs

3. Do GANs have an encoder?

a. Yes
b. **No**

4. Does a GAN's generator sees the data during training?

a. Yes, it needs comparisons to produce similar data to the training data to fool the discriminator
b. **No, it only relies on the discriminator's feedback in the form of probabilities**

5. What is a distinctive feature of Variational Autoencoders (VAEs) compared to traditional autoencoders?

a. They use backpropagation for training.
b. They only require labeled data for training.
c. **They learn a probabilistic latent space and can generate new data.**
d. They always produce binary outputs.

6. Generative Adversarial Networks (GANs) are composed of two main components, what are they?

a. Encoder and Decoder.
b. **Generator and Discriminator.**
c. Predictor and Classifier.
d. Supervisor and Operator.

7. Generative Adversarial Networks (GANs) are composed of two main components, what are they?

a. Encoder and Decoder.
b. **Generator and Discriminator.**
c. Predictor and Classifier.
d. Supervisor and Operator.

8. What role does the discriminator play in a Generative Adversarial Network (GAN)?

a. It generates new data instances.
b. **It evaluates the authenticity of samples, distinguishing between real and generated data.**
c. It classifies data into predefined categories.
d. It compresses the input data into a latent space.

9. In the training of a VAE, what is the purpose of the reparameterization trick?

a. To speed up the training process.
b. **To allow backpropagation through random sampling.**
c. To reduce the number of parameters in the model.
d. To enhance the resolution of generated images.

10. Which of the following best describes the training process of GANs?

a. The generator maximizes the probability of the discriminator being correct.

b. The discriminator maximizes the probability of identifying real and generated data correctly, while the generator maximizes the probability of the discriminator making a mistake.

c. The generator and discriminator are trained in separate phases, not affecting each other.

d. **The discriminator is trained to distinguish real data from fake, while the generator is trained to fool the discriminator into thinking the fake data is real.**

8 Reinforcement Learning

1. What does the Markov Decision Process tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$ contain?

   a. **State set, Action set, Probability function, Reward function**

   b. State set, Agent set, Probability function, Reward function

   c. State set, Action set, Probability distribution, Reward function

2. Which attributes below describe an MDP?

   a. **time-independent**

   b. depends on past states

   c. **random process**

3. Which training method looks one step ahead?

   a. **Temporal Difference**

   b. Monte Carlo

4. Can TD be applied to non-episodic MDPs?

   a. **Yes**

   b. No

5. Can MC be applied to non-episodic MDPs?

   a. Yes

   b. **No**

6. Which of the below works better in complex unpredictable scenarios where the environment is not known?

   a. **MC**

   b. TD

7. Which of the following statements best describes the difference between Q-learning and SARSA in reinforcement learning?

   a. Q-learning is an on-policy algorithm, while SARSA is an off-policy algorithm.

   b. **Q-learning updates its Q-values using the greedy action selection, while SARSA updates its Q-values using the actual next action taken.**

   c. Q-learning and SARSA both update their Q-values based on the actual next action taken and are considered on-policy algorithms.

   d. There is no difference; Q-learning and SARSA are simply different names for the same algorithm.

8. What distinguishes Q-learning in terms of policy type?

   a. Q-learning is an on-policy control algorithm.

b. **Q-learning is an off-policy control algorithm.**

c. Q-learning does not utilize any policy for learning.

d. Q-learning switches between on-policy and off-policy during training.

9. In Q-learning, when is the Q-value updated?

a. After the completion of an episode.

b. **After each step within an episode.**

c. Only at the end of a learning session.

d. Before the agent takes an action.

10. In SARSA, how are action values updated?

a. Using the expected value of the next state's best action.

b. **Using the value of the next action selected by the current policy.**

c. Independently of the actions selected by the current policy.

d. Using a random action from the next state.

11. SARSA is considered what type of learning algorithm?

a. Deterministic

b. Off-policy

c. **On-policy**

d. Model-based