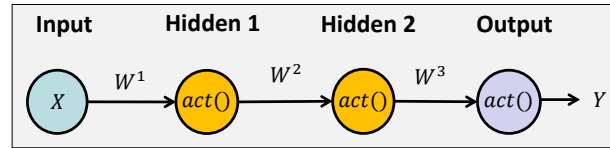**Questions:**

1) Which equation gives the output of the network shown below:

- ☐ $Y = act(W^3 \times act(W^2 \times act(W^1 \times X)))$
- ☐ $Y = act(W^1 \times act(W^2 \times act(W^3 \times X)))$
- ☐ $Y = act(W^3) + act(W^2) + act(W^1) + X$
- ☐ $Y = act(W^3) \times act(W^2) \times act(W^1) \times X$

2) Gradient Descent learning is based on the derivative of the which two factors:

- ☐ Output and Error
- ☐ Weight and Error
- ☐ Input and Error
- ☐ Input and Output

3) How is this derivative used to minimise the error?

- ☐ The weights are adjusted in the opposite direction of the derivative
- ☐ The weights are adjusted in the same direction of the derivative
- ☐ The weights are adjusted according to the inverse of the derivative
- ☐ The weights are adjusted according to the square of the derivative

4) Which of the following equations shows the concept for one step of a gradient-based training algorithm using the learning rate α, the trainable parameters $\boldsymbol{\theta}_t$ of at different time steps t and the gradient $\nabla_{\boldsymbol{\theta}_t}$:

- ☐ $\boldsymbol{\theta}_{t+1} = -\alpha\nabla_{\boldsymbol{\theta}_t}.$
- ☐ $\boldsymbol{\theta}_{t+1} = \alpha\boldsymbol{\theta}_t - .\nabla_{\boldsymbol{\theta}_t}$
- ☐ $\boldsymbol{\theta}_{t+1} = -\alpha.\boldsymbol{\theta}_t\nabla_{\boldsymbol{\theta}_t}$
- ☐ $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha\nabla_{\boldsymbol{\theta}_t}.$

5) In gradient-based training, what is the role of the learning rate α?

☐ Helps minimise convergence effects when the input is small

☐ Helps minimise divergence effects when the input is large

☐ Helps maximise convergence effects when the input is small

☐ Helps maximise divergence effects when the input is large

6) The role of a *cost function* when training a deep learning model is to:

☐ Provide a score reflecting model performance

☐ Decrease the generalisation errors

☐ Determine the optimal complexity of the model

☐ Determine the adequate amount of training data

7) Stacked linear networks represent:

☐ Negates the need to induce correlation between the input and the output

☐ A highly effective regularisation technique

☐ A computationally inexpensive version of a single layer linear network

☐ A computationally expensive version of a single layer linear network

8) Which is the following is **not** a core property of an activation function

☐ Monotonic

☐ Discrete

☐ Nonlinear

☐ Computational Efficient

9) The *Sigmoid* and *Hyperbolic Tangent* activation functions:

☐ Both have a range of 0 to 1

☐ Both have a range of -1 to 1

☐ Have a range of 0 to 1 and -1 to 1 respectively

☐ Have a range of -1 to 1 and 0 to 1 respectively

10) Which of the following mathematical operations is characteristic for backpropagation?

☐ Polynomial interpolation.

☐ Chain rule of derivation.

☐ Integration by substitution.

☐ Higher order derivatives.

11) Which equation gives the gradient with respect to W^1 in the network shown below?
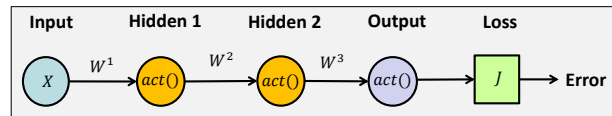
☐ $\frac{\partial Error}{\partial W^1} = \frac{\partial Error}{\partial X} \times \frac{\partial X}{\partial W^1}$

☐ $\frac{\partial Error}{\partial W^1} = \frac{\partial Error}{\partial Output} + \frac{\partial Output}{\partial W^1}$

☐ $\frac{\partial Error}{\partial W^1} = \frac{\partial Error}{\partial Output} + \frac{\partial Output}{\partial hidden2} + \frac{\partial hidden2}{\partial hidden1} + \frac{\partial hidden1}{\partial W^1}$

☐ $\frac{\partial Error}{\partial W^1} = \frac{\partial Error}{\partial Output} \times \frac{\partial Output}{\partial hidden2} \times \frac{\partial hidden2}{\partial hidden1} \times \frac{\partial hidden1}{\partial W^1}$

12) In weighted-based regularisation, what is the main **advantage** in ensure smaller weights?

☐ Reduce bias errors

☐ Reduce the complexity of the model

☐ Improve training speed

☐ Reduce both bias and variance errors

13) What is the main **disadvantage** of *data augmentation* as a regularisation strategy?

☐ Increased training time

☐ Larger mini-batch size

☐ Increased model complexity

☐ Applying cross-class transformation

14) Which of the following properties makes a signal **less** suitable for analysis with a CNN

☐ Similar patterns appear in different regions of the signal

☐ The signal is sequential (time-dependent) in nature

☐ Key patterns can be much smaller than the whole signal

☐ Downsampling the signal does not change the key properties

15) The output of a convolutional kernel is found by doing which operation on the input signal:

☐ Dot product

☐ Conventional Convolution

☐ Flipped Convolution

☐ Fully connected summation

16) How many learnable parameters are associated with a max pooling layer?

☐ Zero

☐ One, the size of the pooling field

☐ Two, the size of the pooling field and the stride

☐ The same as size of the pooling field

17) What is the role of the forget gate in a *Long Short-Term Memory* (LSTM) cell is to:

☐ Reset the previous hidden state value

☐ Determine what old information should be thrown away or kept

☐ Determine what old information to throw away and what new information to add

☐ Learn when to forget all information and reset weights to identity matrix

18) Which of the following statements is **true**:

☐ GRU and LSTM networks never overfit

☐ LSTM cells are less likely to overfit than GRUs

☐ GRU and LSTM networks are equally as likely to overfit

☐ GRUs are less likely to overfit than LSTM cells

19) What is the disadvantage of conventional sequence-to-sequence modelling

☐ They can only be used in combination with LSTM networks

☐ Only the final decoder state is used to initialise the encoder

☐ Only the final encoder state is used to initialise the decoder

☐ The context vector increases the number of learnable parameters

20) Which of the following options uses exploration as a learning strategy in a reward-based environment?

☐ Supervised Learning

☐ Reinforcement Learning

☐ Unsupervised Learning

☐ Semi-Supervised Learning