Deep Learning 2021: PRACTICE EXAM

Name and student ID: _____
DATE: _____

1 Answer sheet

	A	B	C	D
1	x			
2	x			
3	x			
4	x			
5	x			
6	x			
7	x			
8	x			
9	x			
10	x			
11	x			
12	x			
13	x			
14	x			
SUM:				

2 Questions

1. We build a two-layer feed-forward network, and don't include any activation function. Which is false?
   A. This network suffers from vanishing gradients because we didn't include sigmoid activations.
   B. This network can also be represented by a one-layer network.
   C. This network can also be represented by a linear function.
   D. This network can also be represented by a multi-layer network with activation functions.

2. Backpropagation is based on the chain rule of calculus.

   If we apply the chain rule to a simple feedforward neural network, the factors we get require things like the derivative of a function with matrix in put and vector output. Such derivates are most naturally expressed as 3-tensors, which require a lot of memory to store.

   Why is it that we can still apply backpropagation to a feedforward network efficiently?
   A. We only ever focus only on the derivative of the loss, and accumulate the product of the chain rule from the loss to the inputs.
   B. We unroll the network in the time dimension.
   C. We flatten all matrices into vectors, making all modules vector-to-vector functions.
   D. We focus only on the scalar view of backpropgation, looping over the individual elements of the tensors.

3. If we build a multilayer perceptron, but don't add any activation functions, what is the result?
   A. A network that can also be represented by a single-layer network.
   B. A network that suffer strongly from vanishing gradients.
   C. A network that is computationally unstable.
   D. A network to which we cannot apply gradient descent.

4. Neural networks were originally based on the neurons in (human) brains. What part of a modern Neural network closest resembles a brain neuron?
   A. a node in a non-linear layer
   B. a node in a linear layer
   C. a node in a hidden layer
   D. the weight of a node in a hidden layer

5. It is known that GANs tend to generate images of high quality, however, they can suffer from serious problems. Which of the following scenarios is one of such problems:
   A. For handwritten digits, GANs can generate only a subset of images (e.g., only digits 1, 4, and 7).
   B. For colorful images, GANs can skip some channels (e.g., generate only gray images).
   C. For images of faces, GANs can generate cars instead.
   D. GANs do not suffer from any problem.

6. When we derived backpropagation, we introduced the multivariate chain rule. Why did we need this?
   A. To ensure that backprogation is purely symbolic.
   B. For when a node in the computation graph depends on a parameter through multiple paths.
   C. To ensure that backpropagation is purely numeric.
   D. Because backpropagation is a hybrid of numeric and symbolic differentiation.

7. We build a sequence classifier by stacking a number of RNN layers followed by a global pooling operation. This results in a single vector which we then project down to the number of classes.

   Why is global pooling preferable to just picking the last element of the output sequence to represent the whole?

   A. It means every token in the input has an equally short path to the output.
   B. It is cheaper to compute.
   C. It reduces the output resolution so we can add more channels.
   D. It mean we will not need an activation in the final layer.

8. Please indicate components of GANs:

   A. Discriminator, generator, adversarial loss
   B. Painter, expert, Pablo Picasso.
   C. Critic, discriminator, generator, Wasserstein loss.
   D. Discriminator, generator, integrator.

9. In the lecture we have showed that an MLP can be used as a sequence-to-sequence layer.

   Why will it not be very effective if we build a sequence-to-sequence model out of only MLP sequence-to-sequence layers?

   A. Such layers to not propagate information along the time dimension.
   B. We would suffer from vanishing gradients along the time dimension.
   C. They are too expensive to compute compared to RNNs.
   D. They use too much memory, compared to RNNs.

10. What do World Models model that Actor-Critic methods do not model?

    A. Transition dynamics and rewards
    B. Rewards
    C. Transition dynamics
    D. Transition dynamics and value functions

11. Which one is true?

    A. Self attention by itself is permutation equivariant.
    B. Self attention with position embeddings becomes permutation equivariant.
    C. Key, query and value transformations break permutation equivariance.
    D. Multi-head attention breaks permutation equivariance.

12. Which of the following models is an autoregressive model:

    A. WaveNet
    B. RealNVP
    C. Generative Adversarial Networks
    D. Probabilistic PCA

13. What Policy Gradient method best describes this reasoning process:

    "I think I should reinforce this action because I think I'll get 8 more dollars when executing it"

    A. Advantage Actor-Critic
    B. REINFORCE
    C. REINFORCE with baseline

D. Actor-Critic

14. Let us assume that x is a random variable that can take only binary values, i.e., $x \in \{0, 1\}$. In the VAE framework, what distribution we can use for the conditional likelihood $p(x|z)$:

    A. Bernoulli distribution

    B. Gaussian distribution

    C. Laplace distribution

    D. Poisson distribution