Midterm Exam
Deep Learning (Fall 2021)

1:20 pm - 3:10 pm, Dec. 14, 2021

Write your name and ID on both the Problem Sheet and the Answer Sheet. Limit your answer to 3 sentences per question. Closed book. No cell phones, calculators, laptops, iPads, etc. Submit both your Problem Sheet and Answer Sheet.

1) Neural network architecture (20 points in total)
   a) (4 points) Under what conditions sigmoid units would be preferred for output and under what conditions softmax units would be preferred instead?
   b) (4 points) Why are rectified linear units (ReLU) a preferred choice of activation functions for hidden units?
   c) (4 points) Explain the purpose of skip connections.
   d) (4 points) What does the "universal approximation theorem" state for neural networks?
   e) (4 points) What is the similarity and what is the difference in ideas between CNNs and RNNs?

2) Regularization (16 points in total)
   a) (4 points) Explain why regularization is required for machine learning.
   b) (4 points) What is dataset augmentation?
   c) (4 points) How does early stopping work?
   d) (4 points) What is dropout?

3) Convolutional neural network (CNN) (16 points in total)
   a) (4 points) Calculate the output of the CNN given the input and the kernel shown below. Use cross-correlation. Consider a simple case without stride and without zero padding.

Input

0	1	1	1	0	0
0	0	1	1	1	0
0	0	0	1	1	1
0	0	0	1	1	0
0	0	1	1	0	0
0	1	1	0	0	0

Kernel

1	0	1
0	1	0
1	0	1

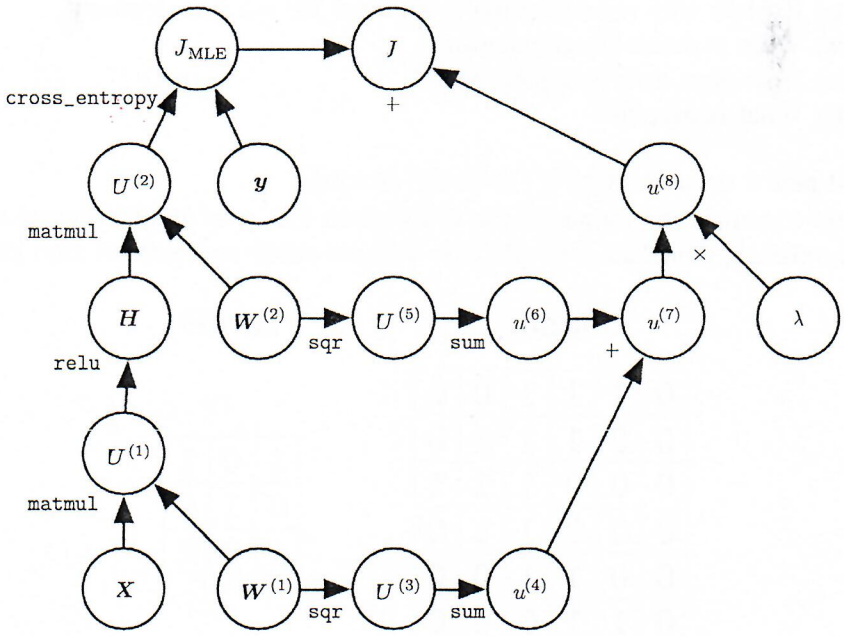
   b) (4 points) Convolution leverages three important ideas that can help improve a machine learning system. What are they? Briefly explain each of them.
   c) (4 points) What is the advantage of using pooling?
   d) (4 points) Explain why zero padding is needed in CNNs. Draw and describe valid convolution, same convolution, and full convolution.

4) Network training (36 points in total)

a) (4 points) What is wrong with the description "Back-propagation refers to the method for exhaustively computing the gradient and performing learning using this gradient."?

b) (8 points) Refer to the computational graph for the gradient of a neural network shown below and fill out the blanks (with circled number) in the following paragraph. Note that *all answers should be a matrix, vector, or scalar.*

To train a single layer MLP with the total cost $J = J_{\text{MLE}} + \lambda \left(\sum_{i,j} \left(W^{(1)}_{i,j} \right)^2 + \sum_{i,j} \left(W^{(2)}_{i,j} \right)^2 \right)$, we wish to compute both $\nabla_{W^{(1)}} J$ and $\nabla_{W^{(2)}} J$. There are two different paths leading backward from J to the weights: one through the cross-entropy cost, and one through the weight decay cost. The weight decay cost is relatively simple; it will always contribute ① to the gradient on $W^{(i)}$. The other path through the cross-entropy cost is slightly more complicated. Let G be the gradient on the unnormalized log probabilities $U^{(2)}$ provided by the cross_entropy operation. The back-propagation algorithm now needs to explore two different branches. On the shorter branch, it adds ② to the gradient on $W^{(2)}$, using the back-propagation rule for the second argument to the matrix multiplication operation. The other branch corresponds to the longer chain descending further along the network. First, the back-propagation algorithm computes $\nabla_H J =$ ③ using the back-propagation rule for the first argument to the matrix multiplication operation. Next, the ReLU operation uses its back-propagation rule to zero out components of the gradient corresponding to entries of $U^{(1)}$ that are less than 0. Let the result be called G'. The last step of the back-propagation algorithm is to use the back-propagation rule for the second argument of the matmul operation to add ④ to the gradient on $W^{(1)}$.



c) (4 points) Explain the gradient descent algorithm and the stochastic gradient descent (SGD) algorithm.

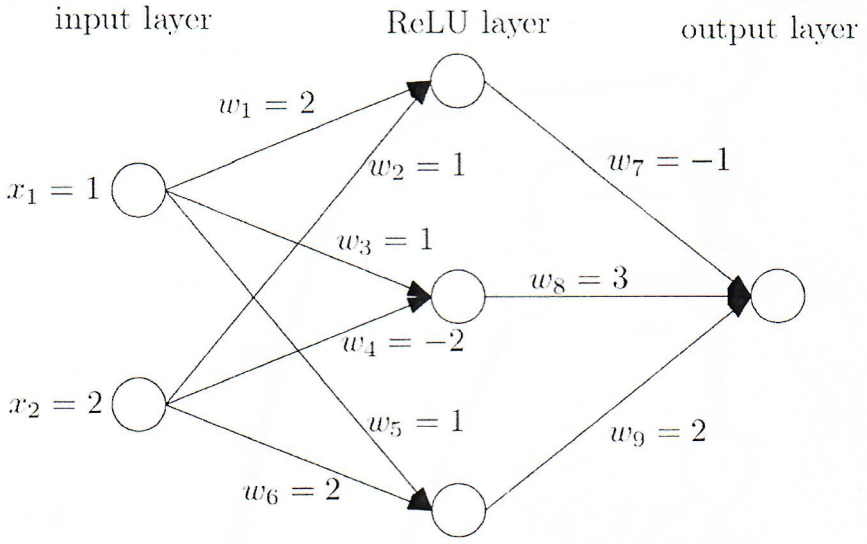
d) (4 points) What are the ideas behind the momentum algorithm and the Nesterov momentum algorithm?

e) (4 points) What are the advantages and disadvantages of first-order and second-order training algorithms?

f) (4 points) Briefly explain the idea of the Adam algorithm.

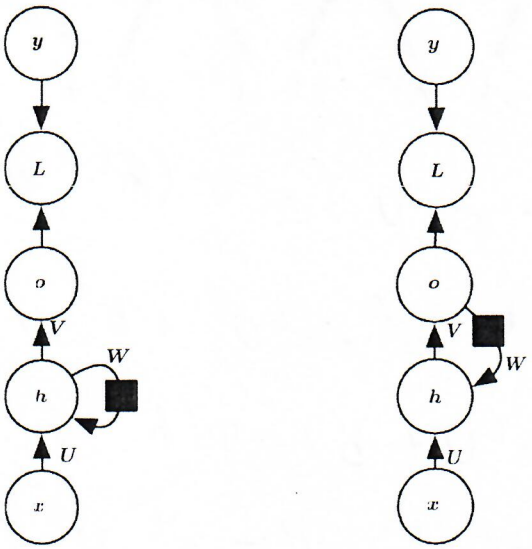
g) (8 points) Consider the neural network shown below. Mean squared error is used for the cost function and the output is 3. Compute the values of all weights ω_i after performing a SGD update with learning rate 0.1. Note that the derivative of ReLU is:

$$f'(x) = \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$



5) Recurrent neural network (RNN) (12 points in total)

a) (4 points) Draw the unfolded graphs of the following two RNNs.



b) (8 points) The figure below shows the LSTM cell. Derive the mathematical expression for both the internal state $s^{(t)}$ and the current hidden layer $h^{(t)}$. Let the current input be $x^{(t)}$. Denote b^f, U^f, and W^f respectively as the bias, input weight, and recurrent weight for the *forget gate*, b^g, U^g, and W^g respectively as the bias, input weight, and recurrent weight for the *input gate*, b^o, U^o, and W^o respectively as the bias, input weight, and recurrent weight for the *output gate*, and b, U, and W respectively as the bias, input weight, and recurrent weight into the LSTM cell. Let the activation function be $\sigma(\cdot)$. To make your life easier, all variables are scalars.