24-788 Introduction to Deep Learning **Name (Print):** _____
Spring 2025
Midterm Exam **Andrew ID:** _____
2/25 6:00 PM - 2/27 11:59 PM (EST)

This exam contains 6 pages (including this cover page) and 3 problems. Check to see if any pages are missing. Submit your work to Gradescope, making sure you assign the correct page that contains the answer to each question

You are required to show your work on each problem on this exam. The following rules apply:

- **Organize your work**, in a reasonably neat and coherent way. Scattered work without a clear order will receive very little credit.
- **Mysterious or unsupported answers will not receive full credit**. A correct answer, unsupported by calculations, explanation, or algebraic work will receive no credit; an incorrect answer supported by substantially correct calculations and explanations might still receive partial credit.

Problem	Points	Score
1	35	
2	35	
3	30	
Total:	100	

1. (35 points) Short Answer Questions

   (a) (5 points) Why do modern deep learning models no longer use threshold function (like shown in Eq. 1) as an activation function?

   $$f(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases} \tag{1}$$

   (b) (5 points) Name 2 reasons why ReLU (Rectified Linear Unit) is better than Sigmoid and Tanh (hyperbolic tangent) activation functions in deep neural networks.

   (c) (5 points) Assume you are training a fully connected neural network for classification and there is no bug in your code. After the loss function converges, although the prediction accuracy is high on the training set, the prediction accuracy on the validation set is low. List 2 things you could try to improve the performance, and explain why they should work.

   (d) (5 points) Assume you are training a fully connected neural network for classification and there is no bug in your code. This time, the loss function converges after only a few epochs but the prediction accuracy on both the training set and validation set is low. List 2 things you could try to improve the performance, and explain why they should work.

   (e) (5 points) What are the two central features of the Adam optimizer that make it a successful optimizer for training neural networks?

   (f) (5 points) Name 2 advantages of CNNs (Convolutional Neural Networks) compared to fully connected neural networks.

   (g) (5 points) Explain the difference between Max pooling and Average pooling. When would you prefer one over the other while working with an image-based dataset?

2. (35 points) Calculation

Please give your calculation steps and necessary descriptions. You will lose points if only results are given.

(a) (5 points) Your friend wants to build a fully-connected one-hidden-layer Neural Network. Her inputs have 99 features. She wants 7 hidden neurons in the first layer and 2 neurons in the output layer. The output is passed through a softmax function. Each layer is a linear layer which is fully-connected to the previous layer and includes bias terms. What is the number of parameters she will need to create this neural network? Each scalar counts as one parameter.

(b) (5 points) A 7×7 feature map is input into a layer of a CNN model which has a 3×3 filter with padding size 2 and stride size 2. What is the size of the new feature map output from this convolutional layer?

If the same 7×7 feature map is passed into a CNN layer which has a 5×5 filter with stride 1, what padding size does the layer need to use to make the size of the outputted feature map the same size as the original input (7×7)?

(c) (5 points) Suppose in one layer of a CNN, you get the feature map given in Fig. 1. and want to conduct convolution using the filter given in Fig. 1., with stride 1 and padding 0. Calculate the new feature map after the convolution operation.

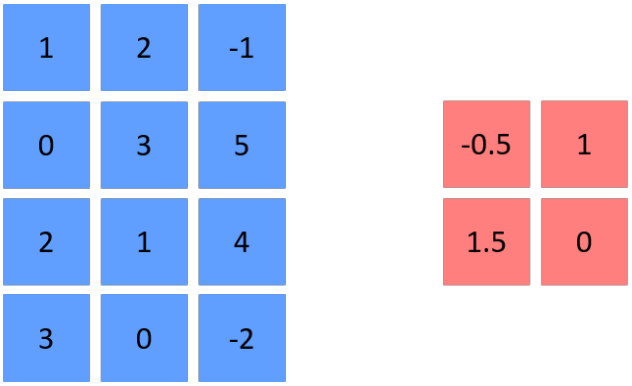

Figure 1: Left: feature map, right: convolutional filter

(d) (10 points) Fig. 2 shows the architecture of LeNet, where in each convolutional layer, $C@W \times W$ represents the size of the output feature maps (C for number of channels and W for height and width). Calculate the number of parameters in this CNN model. (hint: don't forget biases in both convolutional filters and fully connect layers, also you should only consider trainable parameters)

Size of filter and output feature map in each layer is given:

input: 1x32x32

conv1: f(6x5x5)@stride1 $\rightarrow$ 6x28x28

s2: pooling(2x2) $\rightarrow$ 6x14x14

conv3: f(16x5x5)@stride1 $\rightarrow$ 16x10x10

s4: pooling(2x2) $\rightarrow$ 16x5x5

conv5: f(120x5x5)@stride1 $\rightarrow$ 120x1x1

fc6: 120 $\rightarrow$ 84

fc7: 84 $\rightarrow$ 10

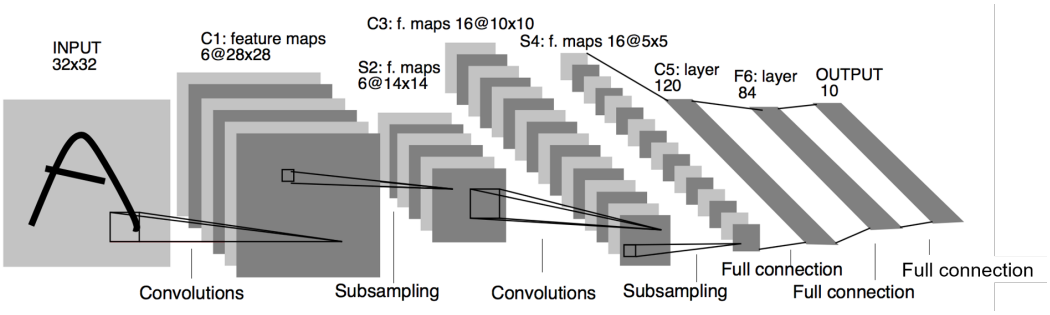

Figure 2: LeNet

(e) (10 points) Let $f(\cdot)$ be the objective function you would like to minimize. At time step t, you get $x_t = 3$, $f(x_t) = 6$ and $f'(x_t) = 1$. After one iteration of gradient descent with learning rate $\eta = 0.1$, what is the value of x_{t+1}.

Now assume you get $f'(x_{t+1}) = 0$, does it guarantee x_{t+1} is the global optimum? If yes, explain why. If not, give a solution that will help find the global optimum.

3. (30 points) Multiple choice questions

Please select ALL that apply! You don't need to give an explanation of your choices. To get full points for each question, you have to select all the correct option(s). You lose 1 point per correct option that is missing and lose all 3 points if you select a wrong option. You get 0 points if you do not select any of the correct answers.

(a) (3 points) In CNN training, when backpropagating through a max pooling layer, the gradient at the input locations of the non-maximum values is:

A) Identical to the derivative at the location of the maximum value

B) Identical to the input value at the location of the maximum value

C) Identical to the input value at the location of the non maximum value

D) 0

(b) (3 points) In CNN training, when backpropagating through a mean pooling layer, the gradient at the output of a mean pool filter is

A) equally distributed over the input pool

B) assigned to the input location of the maximum value

C) 0

D) distributed over the inputs in proportion to their values

(c) (3 points) Consider applying a 5x5x5 convolution filter with a stride of 1 and padding of 2 on an RGB image input of dimension (3,H,W). Which of the following statements are true?

A) The output is of dimension (5,H,W)

B) The output is of dimension (3,H,W)

C) The output is of dimension (5,5,5)

D) The convolution is not valid

(d) (3 points) When using gradient descent to find the minimum of a function, which of the following is correct about gradient descent:

A) Accurately solves for the location that has minimum gradient.

B) Uses the Hessian of the function to solve for the location of the minimum

C) Starts at some initial location and iteratively moves in the opposite direction of the gradient, until it finds the minimum.

D) Starts at some initial location and iteratively moves in the direction of the gradient, until it finds the minimum.

(e) (3 points) Select all the statements that are correct:

A) The output of a neuron has a probabilistic interpretation when using sigmoid activations.

B) A multi-layer fully-connected neural network with linear activation functions is equivalent to a single-layer fully-connected neural network with linear activation.

C) For a multilayer fully connected neural network with Tanh as an activation function at each layer, if you scale up all the weights (including biases) by 2, the output of the model remains exactly the same.

D) Threshold activations cannot be used for backpropagation.

(f) (3 points) What are the benefits of using L2 regularization of weights in a network?

A) It ensures that neuron activations are sparse (i.e. only a few are non-zero), improving the interpretability of their output

B) It prevents activation functions from becoming too steep, which helps keep the network output from changing too steeply with respect to changes in the input.

C) It prevents overfitting to training data

D) It guarantees to restrict the weights to lie within a unit hypersphere, preventing floating point overflow

(g) (3 points) What are the improvements of the momentum learning rule compared to vanilla gradient descent?

A) The magnitude of the changes to the parameters can increase without bound if the gradients don't change very much across iterations.

B) Momentum learning allows us to forget about the learning rate since the learning rule will automatically adjust the step size.

C) Momentum learning guarantees the network will find global minima as opposed to local minima.

D) Momentum learning smooths noisy gradients, reduces oscillations, and encourages updates in directions with smooth convergence behaviors.

(h) (3 points) Which of the following guidelines is applicable to the initialization of the weight vector in a fully connected neural network (assume ReLU activation)?

A) Should not set it to zero since it will cause overfitting.

B) Should not set it to zero since it will cause all neurons to have the same gradient.

C) Should set it to zero to avoid introducing an unlearned bias into the network

D) Should set it to zero in order to preserve symmetry across all neurons.

(i) (3 points) Which of the following statements are **true** regarding Multilayer perceptrons:

A) If there is an upper limit on the width of the layers but no limit on the depth of the network (number of layers), the MLP can model/approximate any function.

B) A one-hidden layer network with infinite neurons in the hidden layer can compose an exact model of any real-valued function.

C) A one-hidden layer network with infinite neurons in the hidden layer can compose an approximate model of any real-valued function.

D) For a deep neural network, if all the weight parameters are initialized with an appropriate initialization method, biases can be initialized with zeros or constants.

(j) (3 points) Select all the true statements about a convolution layer:

A) The number of "channels" in any filter equals the number of input maps.

B) The number of "channels" in any filter equals the number of output maps. (Affine maps output by the layer)

C) The number of filters equals the number of input maps.

D) The number of filters equals the number of output maps.