Seunghyup Alex Oh (916201993)

Problem Set#4

1. Short answer problems [60 points] susceptible

1. In the boosting algorithm AdaBoost, an ensemble of classifiers is selected in sequence, and the weight wi on each labeled training example xi is adjusted from iteration to iteration. What is the purpose of these weights? How do they influence which feature+classifier combination is selected in the next round? [10 points]

An AdaBoost, 'Adaptive Boosting' algorithm, helps to combine multiple weak classifiers and convert into a single strong classifier. The AdaBoost set weights to the training example xi to indicate the amount of difficulty to correctly classify it, namely, weight of each training example is directly proportional to its accuracy – more weight on incorrectly classified and less weight on correctly classified. For every round, the AdaBoost reweights the examples according to errors, updates the weights, and choose the best feature+classifier combination, the classifier with the lowest weighted classification error, for the next round.

2. Invariant interest point detection and geometric verification using a parametric transformation are both commonly used in "instance" recognition (e.g., to recognize a landmark building in a tourist photo), but not in object category recognition (e.g., to recognize any dog). Explain why. [10 points]

Both invariant interest point detection and geometric verification using a parametric transformation aim for matching specific instances, local features. They are based on an assumption that they are able to find reliable features within clutter. The instance based parametric transformation is not suited for generic category recognition. On the other hand, the goal of an object category recognition is given a small number of training images of a category, to recognize a-priori 'unknown instances' of that category and assign the correct label of category to it. The assumption of 'instance' recognition is not applicable to the object category recognition.

3. How does a k-nearest neighbor classifier use the k nearest neighbors to make a label prediction? [10 points]

A k-nearest neighbor, the algorithm which stores all the available cases and classifies the new data based on a measure of the similarity, is used to classify a data point based on how its neighbors are classified. In order to make a label prediction, k-nearest neighbor classifier finds the k nearest neighbors from a new data point and classifies the new data point by forming a majority vote from the (labels of) k nearest neighbors to the new data point.

4. A deep neural network has multiple layers with non-linear activation functions (e.g., ReLU) in between each layer, which allows it to learn a complex non-linear function. Suppose instead we had a deep neural network without any non-linear activation functions. Concisely describe what effect this would have on the network. (Hint: can it still be considered a deep network?) [10 points]

The non-linear activation functions allow the model to create complex mappings between the network's inputs and outputs, which are the essential for learning and modeling complex data. A deep neural network without any non-linear activation functions in the network will just operate like a single-layer perceptron because the adding up all the layers, linear combinations of linear functions, is another linear function, a line again – not the complex relations. Whereas, the real deep neural networks involve complex relations between each layer and the non-linear activation functions enable that complex relations. Namely, without non-linear activation functions, the 'deep neural network' is no longer considered as a 'deep network.'

5. One module of a standard convolutional neural network is the max-pooling operation. Given the 4x4 image below, perform max-pooling with a stride of 2 and a pooling window of 2x2. [10 points]

2	5	3	56
20	5	1	32
3	3	7	46
4	3	12	23

Max pooling is a sample-based discretization process with the purpose of down-sampling an input representation. We take the maximum of each region represented by the filter and create new, output matrix where each element is the maximum of the region in the original input.

20	56
4	46

6. Determine whether each statement is true or false (no need to explain your answer) [10 points]:

a. To detect profile (side) views of faces with the boosting-based Viola-Jones face detector, one should re-train and select a new set of discriminative features using profile face images. - True

b. A key idea of deep learning for visual recognition is to learn a feature hierarchy all the way from pixels to classifier. - True

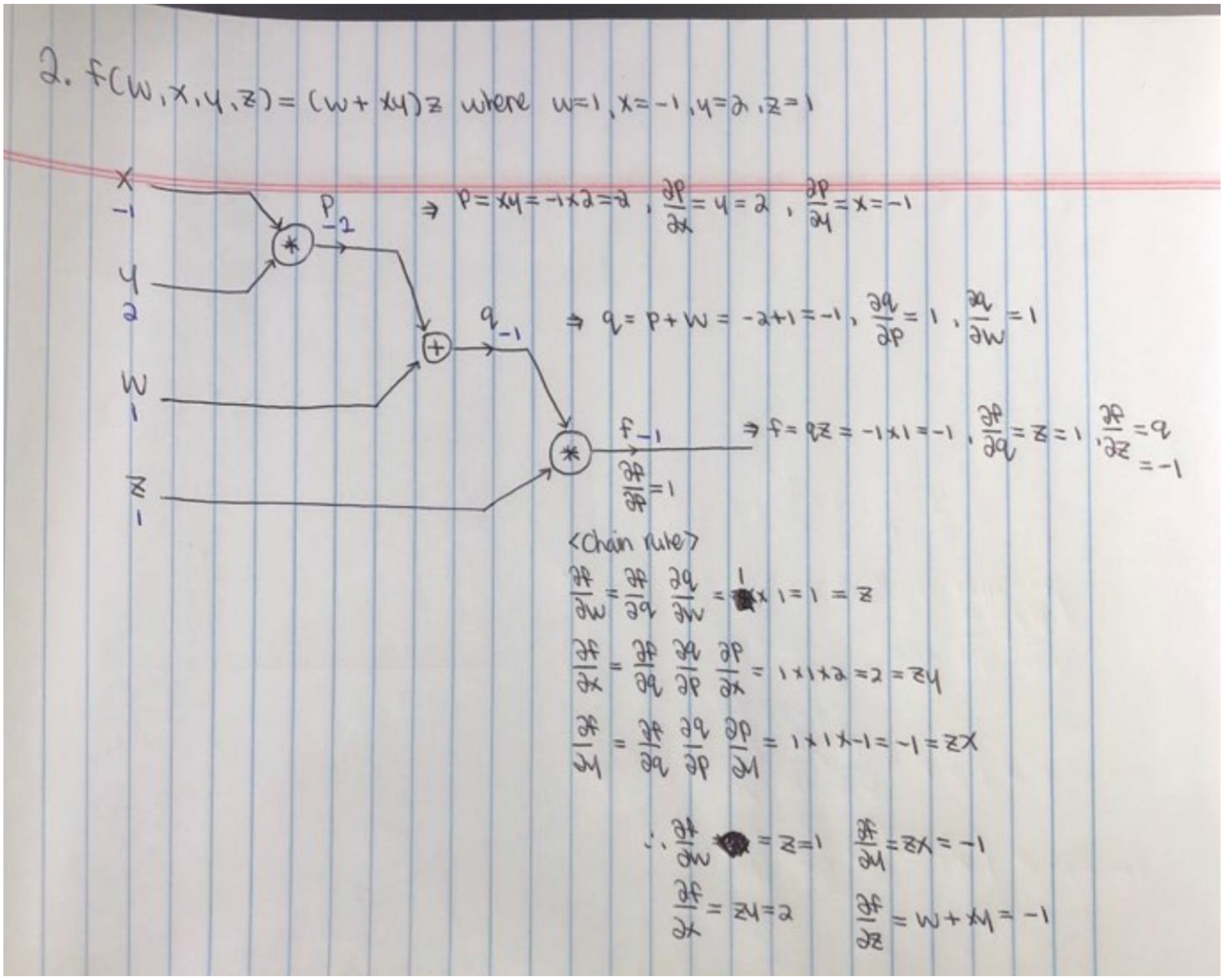
2. Backpropagation [40 points]

Consider the following function:

$$f(w, x, y, z) = (w + xy)z \qquad \text{where} \quad w = 1, x = -1, y = 2, z = 1$$

Draw the computational graph for this function and write all the intermediate variable output values as well as their partial derivatives (i.e.,). Clearly define any intermediate variables that you use (Hint: there should be two). Ultimately, we are interested in computing the following four quantities:

$\partial f/\partial w, \partial f/\partial x, \partial f/\partial y, \partial f/\partial z$


x
-1
y
2
w
1
z
1
*
p
-2
+
q
-1
*
f
-1
$\frac{\partial f}{\partial f}=1$

2. f(w,x,y,z) = (w+xy)z where w=1, x=-1, y=2, z=1

$\Rightarrow p = xy = -1\times2 = 2, \ \frac{\partial p}{\partial x} = y = 2, \ \frac{\partial p}{\partial y} = x = -1$

$\Rightarrow q = p + w = -2+1 = -1, \ \frac{\partial q}{\partial p} = 1, \ \frac{\partial q}{\partial w} = 1$

$\Rightarrow f = qz = -1\times1 = -1, \ \frac{\partial f}{\partial q} = z = 1, \ \frac{\partial f}{\partial z} = q = -1$

<Chain rule>

$\frac{\partial f}{\partial w} = \frac{\partial f}{\partial q}\frac{\partial q}{\partial w} = \overset{1}{}\times 1 = 1 = z$

$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q}\frac{\partial q}{\partial p}\frac{\partial p}{\partial x} = 1\times1\times2 = 2 = zy$

$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial q}\frac{\partial q}{\partial p}\frac{\partial p}{\partial y} = 1\times1\times-1 = -1 = zx$

$\therefore \frac{\partial f}{\partial w} = z = 1 \qquad \frac{\partial f}{\partial y} = zx = -1$

$\frac{\partial f}{\partial x} = zy = 2 \qquad \frac{\partial f}{\partial z} = w + xy = -1$