CS182/282A Designing, Visualizing & Understanding Deep Neural Networks SP2019

Midterm 1

This is a closed-book exam with 7 questions. You can use one two-sided sheet of notes. Please write all your answers in this book. There are a total of 80 points for all questions, and you have 80 minutes to complete the exam. Please budget your time accordingly. Good luck!

NAME____ **SID**____________________

NAME TO YOUR LEFT______________________________________

NAME TO YOUR RIGHT____

1. **(40 points)** Give short answers for each of the following questions – 2 points each

a) Describe one architectural difference (other than difference in number and size of its layers) between the LeNet text recognizer from 1989 and AlexNet from 2012. (2 points)

b) What is multi-task learning, and why does it often improve the performance of deep network models? (2 points)

c) Give one advantage of generative ML models and one advantage of discriminative models. (2 points)

d) Squared error is a loss commonly used for real-valued predictors. What loss is commonly used with predictors of discrete values? (2 points)

e) Like SVMs, logistic regression classifiers typically exhibit a margin around their decision boundary. Explain why with a picture. (2 points)

f) Why is there a tradeoff between the bias and variance of a prediction model? Where do deep networks lie on the spectrum of this tradeoff? (2 points)

g) When would you expect a naïve bayes model to be more accurate than a logistic regression model? When would you expect the logistic regression model to be more accurate? (2 points)

h) For a function $f(X)$, at what kinds of points does the gradient of f vanish? List all that apply. (2 points)

i) If an input data block in a convolutional network has dimension $C \times H \times W = 96 \times 128 \times 128$, (96 channels, spatial dim 128x128) and we apply a convolutional filter to it of dimension $D \times C \times H_F \times W_F = 128 \times 96 \times 7 \times 7$, (i.e. a block of D=128 filters) with stride 2 and pad 3, what is the dimension of the output data block? (2 points)

j) With the aid of a picture, explain how cross-validation is used to estimate model performance. (2 points)

k) Given that matrix multiplication is associative and the derivative of the loss with respect to a weight layer in a deep network is a product of Jacobian matrices, why is the derivative computed backwards from the output (i.e. using backpropagation)? (2 points)

l) Consider the image below. Apply max pooling to it with a 2x2 max pool filter with stride 2 and no padding. Draw the result to the right of the image. (2 points)

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

m) ReLU layers have non-negative outputs. Given one negative consequence of this. Give an example of another layer type that was developed to address this issue? (2 points)

n) What is Xavier initialization and why is it used? (2 points)

o) What other benefits does batch normalization exhibit, besides scaling/normalizing activations? (2 points)

p) What is inverted dropout and what is its advantage? (2 points)

q) What are bagging and boosting? Which method is more commonly used with deep networks and why? (2 points)

r) What is the main difference between Fast R-CNN and Faster R-CNN? What's the reason for this difference? (2 points)?

s) What is a limitation of vanilla recurrent networks (RNNs) and how is this addressed in LSTM networks? (2 points)

t) When using activation maximization to visualize a neuron, why is initialization of the image important? (2 points).

2. **(6 points)** A soft-margin SVM minimizes the following loss function:

$$L = \sum_{i=1}^{n} (0, 1 - y_i f(x_i)) \ + \lambda||w||^2$$

where $f(x) = w^T x + b$, and $y_i \in \{-1, 1\}$ is a label, and n is the number of training samples.

a) Ignoring the regularizing term (with factor λ), how does the loss vary as a function of $||w||$?

b) How does the margin of this classifier vary as a function of λ ?

c) What is the derivative of the loss with respect to w ?

3. **(8 points)** A deep network implements a parametric function $f(X, W)$ where X is the input and W is a vector of weights. Assume that the loss being optimized is symmetric (spherical) about an optimum weight W_0.
   Now consider a function $g(X, V) = f(X, D\ V)$ where D is a diagonal matrix, with $D_{\{ii\}} = i$. Suppose you have successfully trained the function f to some loss L_{final} after N iterations. What if anything can you say about training g assuming both networks were trained with:
   a) Vanilla SGD (no momentum)
   b) SGD with momentum
   c) RMSprop
   d) ADAM

4. **(8 points)** A deep network contains a new layer type called "BiasScale" that implements the following operation:

$$y_i = s_i x_i + b_i$$

where y is an output vector, x is the input vector, and s and b are scale and bias weight vectors respectively. Subscripts denote the i^{th} elements of these vectors, which are of equal dimension.

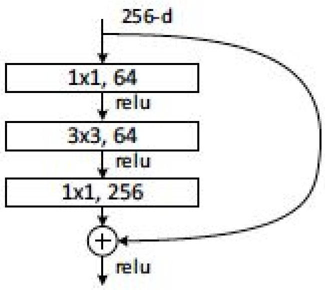
Given the output loss gradient

$$\frac{dL}{dy_i}$$

compute the loss gradient wrt to the parameters s and b.

Would it make sense to place this layer before or after an affine (matrix multiply) layer? Why?

5. **(6 points)** A residual network features bottleneck layers like the one below that use 1x1 convolutions. What is the effect of this design on:
   a) Training and test time (forward and backward passes)
   b) Model complexity (number of parameters)
   c) Overall network accuracy



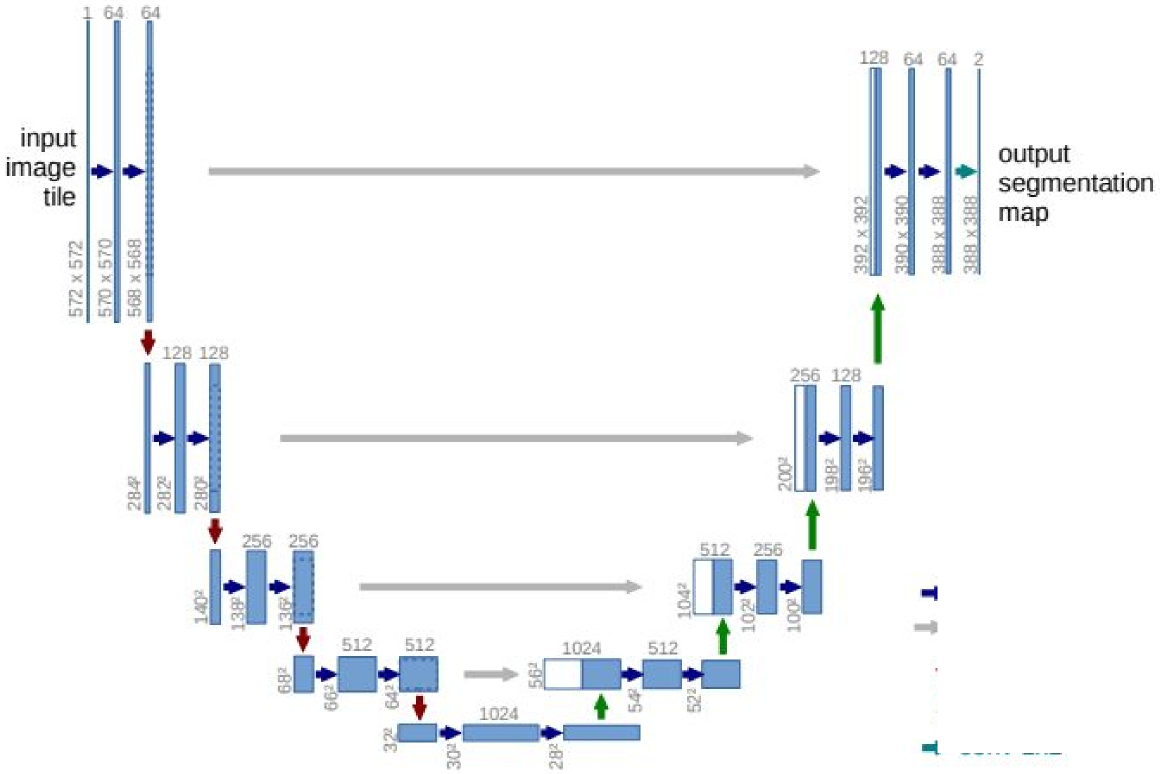
6. **(6 points)** Consider the following collections of models. For each collection, say how would you combine their outputs to improve accuracy: prediction averaging or parameter averaging, and then say what kinds of improvements you would expect to see: reduced bias or reduced variance:

   a) An ensemble of CNN models, all with the same number and dimensions of layers, with different initialization.

   b) An ensemble of CNN models, with different numbers and dimensions of layers.

   c) A series of checkpoints (snapshots) of the same model at different points during training. Assume training has reached the optimum loss, but is continuing.

7. **(6 points)** The figure below shows a U-Net which is used to compute a semantic segmentation of an image.
   a) What operations are represented by the upward arrows in the figure ?
   b) What is the role of the rightward arrows ?
   c) Which of the transformations (arrows in any direction) in the network have learned parameters ? Circle all the arrows that have learned parameters.



1
64
64
input image tile
572 x 572
570 x 570
568 x 568
128
128
284²
282²
280²
256
256
140²
138²
136²
512
512
68²
66²
64²
1024
32²
30²
28²
1024
512
56²
54²
52²
512
256
104²
102²
100²
256
128
200²
198²
196²
128
64
64
2
392 x 392
390 x 390
388 x 388
388 x 388
output segmentation map