CS194-129 Designing, Visualizing & Understanding Deep Neural Networks F2016

Sample Midterm Questions

This will be a closed-book exam of 80 minutes duration, except that you can use one page of notes. These sample questions have not been calibrated to be done in 80 minutes, they are for study purposes only.

Sample Questions: mostly short, some longer

(a) Why do deep neural networks typically outperform shallow networks?

(b) For a k-nearest neighbor (kNN) classifier, increasing the number of neighbors k should have what effect on bias and variance? Explain briefly.

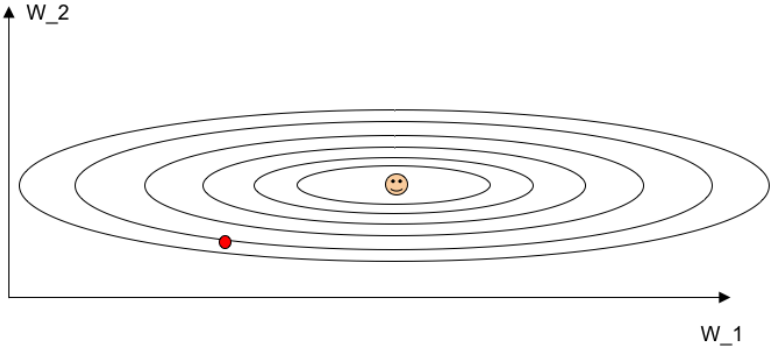
(c) What is a validation dataset for?

(d) Which of the following are true for the SVM loss from class, select all that apply:

   a. Loss is positive when the correct class score is lower than some other class score for the same image.
   b. Loss is positive when the correct class score is the same as the maximum of other class score for the same image.
   c. Loss is positive when the correct class score is slightly larger than the maximum of other scores for that image.

(e) Explain why SGD with O(1/n) convergence is usually preferred over methods with linear O(log n) or quadratic O(log log n) theoretical convergence?

(f) Sketch trajectories on the following contour plot of loss for:

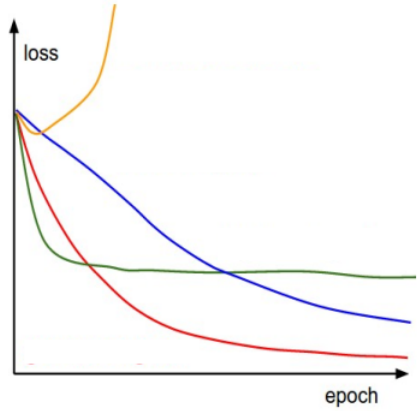


   a. Simple SGD without momentum
   b. SGD with momentum, under-damped
   c. SGD with momentum, critically-damped (the boundary between over-damping and under-damping).

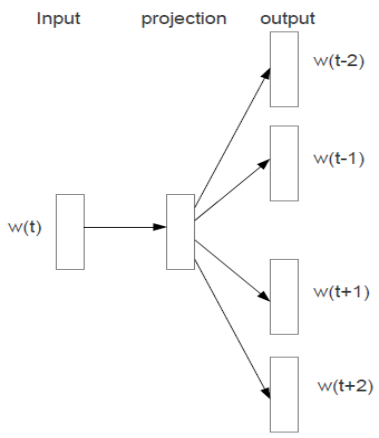
(g) The following learning rate plots show the loss of an algorithm versus time. Label each plot with one of these labels:

   a. Low learning rate
   b. Optimal learning rate

c. High learning rate
d. Very high learning rate



(h) What is leaky-RELU activation, and why is it used? Sketch its activation curve.
(i) Derive a good weight initialization (e.g. Xavier initialization) for an n x m FC layer.
(j) For a minibatch with mean M and variance V, what does batch normalization do?
(k) In one or more sentences, and using sketches as appropriate, contrast: AlexNet, VGG-Net, GoogleNet and ResNet. What was one defining characteristic of each?
(l) In a sentence, explain: gradient clipping, one-bit gradients, gradient noise.
(m) Which of the following systems use only deep network components (and are trained end-to-end)?
    a. R-CNN
    b. Fast R-CNN
    c. Faster R-CNN
(n) Using LSTM blocks as single elements, draw a simple (depth-1) recurrent network for language translation. Show sample input and output words (both can be in English for the purpose of this answer). What loss is used? Explain how to increase the depth of your network.
(o) GANs (Generative Adversarial Networks) include a generator and a discriminator. Sketch a basic GAN using those elements, a source of real images, and a source of randomness.
(p) Contrast t-SNE projection with PCA projection for visualizing a set of high-dimensional points.
(q) For the sentence “morning fog, afternoon light rain,” place the words on the skip-gram Word2Vec model below. Draw a CBOW model using the same words.



(r) Define the following terms:
   a. Hard attention
   b. Soft attention

   Which attention models can be trained with backpropagation only? What other training method is required? Briefly explain why.

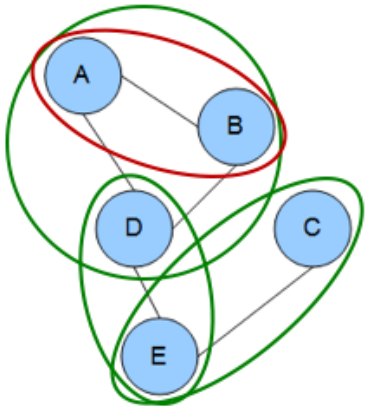
(s) The diagram below shows the steps in DAGGER imitation learning. Modify two steps in the cycle to implement PLATO (Policy Learning with Adaptive Trajectory Optimization). Write the two modified steps under this list with appropriate step numbers.

1. train $\pi_\theta(\mathbf{u}_t|\mathbf{o}_t)$ from human data $\mathcal{D} = \{\mathbf{o}_1, \mathbf{u}_1, \ldots, \mathbf{o}_N, \mathbf{u}_N\}$
2. run $\pi_\theta(\mathbf{u}_t|\mathbf{o}_t)$ to get dataset $\mathcal{D}_\pi = \{\mathbf{o}_1, \ldots, \mathbf{o}_M\}$
3. Ask human to label $\mathcal{D}_\pi$ with actions $\mathbf{u}_t$
4. Aggregate: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$

(t) List some weaknesses of the vanilla policy gradient method? What are some solutions for those weaknesses?

(u) List 4 transformations that can be applied to images in a dataset to augment the dataset for CNN training.

(v) For the Markov random field below, the clique potentials are highlighted. Draw the corresponding factor graph.


A
B
D
C
E