Homework 1 Quiz

- Due Sep 6 at 11:59pm
- Points 10
- Questions 10
- Available Aug 29 at 12am - Sep 6 at 11:59pm
- Time Limit None
- Allowed Attempts Unlimited

Take the Quiz Again

Attempt History

	Attempt	Time	Score
LATEST	Attempt 1	1,496 minutes	10 out of 10

Correct answers are hidden.

Score for this attempt: 10 out of 10
Submitted Sep 6 at 4:51pm
This attempt took 1,496 minutes.

Question 1
1 / 1 pts
You are backpropating through a sigmoid activation layer and you have to compute its local gradient. Say the pre-activation and post-activation values are represented by Z and A respectively, What's the correct formula for computing the sigmoid activation layer gradient?

- $\frac{\partial A}{\partial Z} = 1 - 2(A \odot A)$
- $\frac{\partial A}{\partial Z} = A - A \odot A$
- $\frac{\partial A}{\partial Z} = 1 - A \cdot A$
- $\frac{\partial A}{\partial Z} = A - A^2 \odot A$
- Sigmoid is not differentiable

Question 2
1 / 1 pts
Which of the following is NOT true about SGD with momentum?

- The learning rate is adapted according to the stage the algorithm is in the optimization process.
- The momentum term incorporates the effect of gradients from previous iterations.
- The parameters are updated with the new momentum term and not just the current iteration's gradient.
- It has lesser oscillations than simple SGD

Question 3

1 / 1 pts

Which of the following statements about batch normalization is incorrect?

To compute the derivative of the mini-batch of any normalized data , we will only consider the respective instance Z from the batch.

- The shift and scaling are applied to the individual $\hat{Z}$ to compute the corresponding $\tilde{Z}$
- Batch Normalization is primarily used to improve the stability and convergence of the neural network during training.
- Batch Normalization introduces additional trainable parameters known as scale and shift.

Question 4

1 / 1 pts

Which of the following statements about activation functions is incorrect?

- The jacobian of a scalar activation function is a diagonal matrix

Backpropagating gradients through an output element of a vector activation function requires only the corresponding input element

- Each output element of a vector activation function depends on all of the input elements.
- Scalar activations when applied to a vector, operate elementwise

Question 5

1 / 1 pts

The shape of your activations Z after forwarding the input data A through the network is N X Cout where N is batch size and Cout is the number of output features. What will be the shape of the derivative of loss L with respect to Z, dLdZ?

- The shape of dLdZ is Cout X N
- The shape of dLdZ is 1 X Cout
- The shape of dLdZ is N X 1
- The shape of dLdZ is N X Cout

Question 6

1 / 1 pts

The "dev" set for hw1p2 has 2703 recordings. The most common phoneme in the transcript is "SIL", denoting silence. What is the least common phoneme? Please write your answer in all caps, exactly how it is in the transcript.

ZH

Question 7

1 / 1 pts

X denotes the number of audio clips or recordings. For every clip, you have a (Y, Z) dimensions where y represents time, and z represents frequency (number of frequency bins = 28 in this case). However, each Y is a fraction of a second, so it is better to have more context (looking at the timesteps before and after) for that given timestep. What do you do at y = 1 (the first timestep)?

- (●) You would need to pad prior to the first Y with size context and with zeros but do not need to pad the training labels.
- ( ) You would need to pad prior to the first Y with size context and use the prior audio clip values.
- ( ) You would need to pad prior to the first Y with size context and with zeros and similarly with the training labels.

Question 8

1 / 1 pts

X denotes the number of audio clips or recordings. For every clip, you have a (Y, Z) dimensions where y represents time, and z represents frequency (number of frequency bins = 28 in this case). What is the shape for a single training instance with a context size of 12 before flattening?

- ( ) (27,28)
- ( ) (24, 28)
- (●) (25, 28)
- ( ) (12, 40)

Question 9

1 / 1 pts

The "dev" set for hw1p2 has 2703 recordings. The most common phoneme in the transcript is "SIL", denoting silence. How many instances of SIL are there in the entire dev set?

319,908

Question 10

1 / 1 pts

David started developing his hw1p2 model starting from the simple model provided in the example notebook and 2-layer baseline and passed the very low cutoff required for the early submission.

Looking at the writeup and Piazza, he found that there were a lot of things to try next, from increasing the number of layers, changing the # of neurons in each layer, and different schedulers (in addition to many other things). What is the first thing he should do to go about improving his model?

Looking at the writeup, David notices that one of the things to try is larger model size. To improve his model, David tries adding 2 more layers his model and finds that the accuracy had increased a great deal! Confident that his change had been beneficial, David continues building on the updated model, trying other changes.

David finds that his model is overfitting, with a training accuracy in the 90s but a validation accuracy in the 80s. Learning that dropout is a way to mitigate overfitting, David adds some dropout layers to his model and trains his model. However, looking at the training loss and validation accuracy curves as the training progresses, David is surprised to find out that although his training accuracy decreased, his validation accuracy also decreased. Finally, at the end of training the same 10 epochs, the model's validation accuracy is indeed worse. David concludes that dropout is not effective and tries something else.

David knows from talking with TAs and referencing the writeup that adding context and adding some dropout are good ways that he can improve his model. Reasoning along this line, he tried increasing context and added a little bit of dropout between the layers to reduce overfitting.

David first chooses to experiment with increasing context size, as suggested in the writeup. He tries increasing the context size to 6 and notices a remarkable boost in performance. He then goes on to test increasing model sizes, adding batchnorm, etc.

Quiz Score: 10 out of 10