#### **Shallow Neural Networks**

Logistic regression’s weights w should be initialized randomly rather than to all zeros, because if you initialize to all zeros, then logistic regression will fail to learn a useful decision boundary because it will fail to “break symmetry”, True/False?

Correct Answer – False

10.

Question 10

In the same network as the previous question, what are the dimensions of Z^{[1]}*Z*[1] and A^{[1]}*A*[1]?

1 point



Z^{[1]}*Z*[1] and A^{[1]}*A*[1] are (1,4)



Z^{[1]}*Z*[1] and A^{[1]}*A*[1] are (4,2)



Z^{[1]}*Z*[1] and A^{[1]}*A*[1] are (4,m)



Z^{[1]}*Z*[1] and A^{[1]}*A*[1] are (4,1)

#### Key concepts on Deep Neural Networks

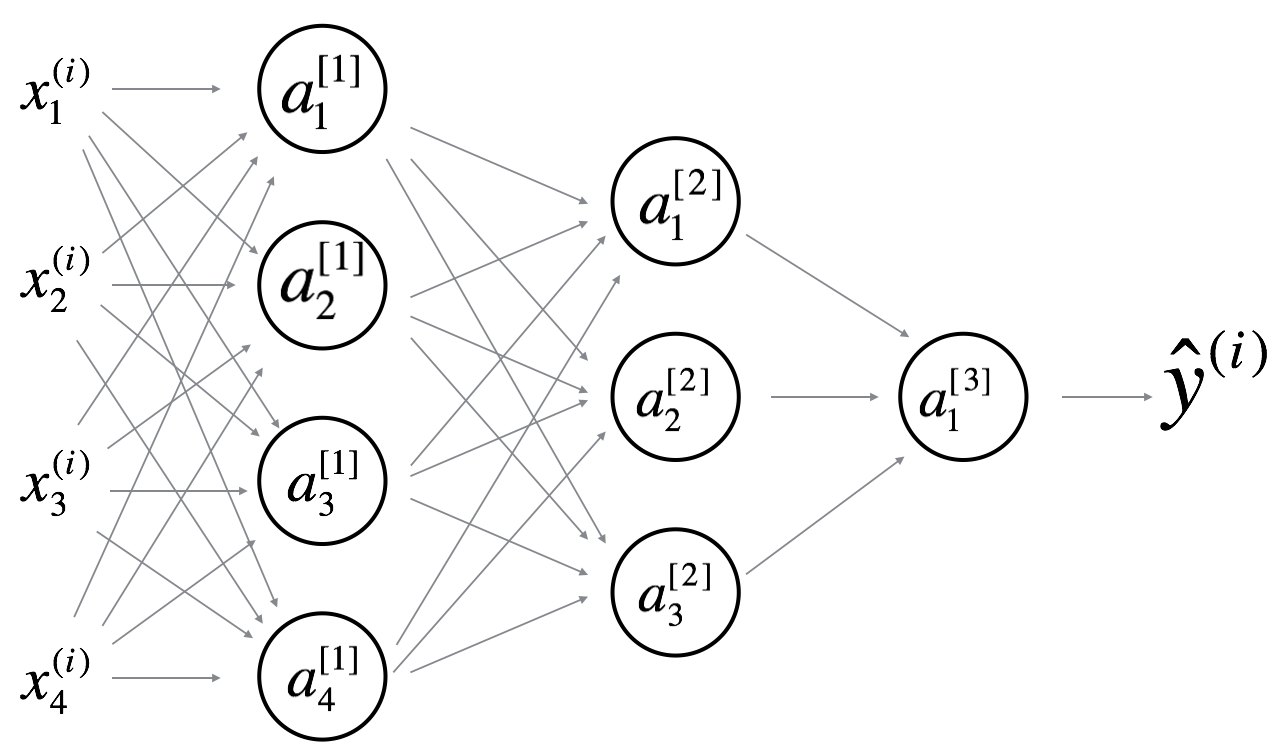
4. Vectorization allows you to compute forward propagation in an L*L*-layer neural network without an explicit for-loop (or any other explicit iterative loop) over the layers l=1, 2, …,L. True/False?

False

9.

Question 9

Consider the following 2 hidden layer neural network:



Which of the following statements are True? (Check all that apply).

1 point



W^{[1]}*W*[1] will have shape (4, 4)



b^{[1]}*b*[1] will have shape (4, 1)



W^{[1]}*W*[1] will have shape (3, 4)



b^{[1]}*b*[1] will have shape (3, 1)



W^{[2]}*W*[2] will have shape (3, 4)



b^{[2]}*b*[2] will have shape (1, 1)



W^{[2]}*W*[2] will have shape (3, 1)



b^{[2]}*b*[2] will have shape (3, 1)



W^{[3]}*W*[3] will have shape (3, 1)



b^{[3]}*b*[3] will have shape (1, 1)



W^{[3]}*W*[3] will have shape (1, 3)



b^{[3]}*b*[3] will have shape (3, 1)

#### Optimization algorithms

Which of these statements about mini-batch gradient descent do you agree with?

1 point



You should implement mini-batch gradient descent without an explicit for-loop over different mini-batches, so that the algorithm processes all mini-batches at the same time (vectorization).



One iteration of mini-batch gradient descent (computing on a single mini-batch) is faster than one iteration of batch gradient descent.



Training one epoch (one pass through the training set) using mini-batch gradient descent is faster than training one epoch using batch gradient descent.

9.

Question 9

Suppose batch gradient descent in a deep network is taking excessively long to find a value of the parameters that achieves a small value for the cost function \mathcal{J}(W^{[1]},b^{[1]},..., W^{[L]},b^{[L]})J(*W*[1],*b*[1],...,*W*[*L*],*b*[*L*]). Which of the following techniques could help find parameter values that attain a small value for\mathcal{J}J? (Check all that apply)

1 point



Try using Adam



Try tuning the learning rate \alpha*α*



Try initializing all the weights to zero



Try mini-batch gradient descent



Try better random initialization for the weights

#### Bird recognition in the city of Peacetopia (case study)

5.

Question 5

After setting up your train/dev/test sets, the City Council comes across another 1,000,000 images, called the “citizens’ data”. Apparently the citizens of Peacetopia are so scared of birds that they volunteered to take pictures of the sky and label them, thus contributing these additional 1,000,000 images. These images are different from the distribution of images the City Council had originally given you, but you think it could help your algorithm.

Notice that adding this additional data to the training set will make the distribution of the training set different from the distributions of the dev and test sets.

Is the following statement true or false?

"You should not add the citizens' data to the training set, because if the training distribution is different from the dev and test sets, then this will not allow the model to perform well on the test set."

1 point



True



False

6.

Question 6

One member of the City Council knows a little about machine learning, and thinks you should add the 1,000,000 citizens’ data images to the test set. You object because:

1 point



A bigger test set will slow down the speed of iterating because of the computational expense of evaluating models on the test set.



The 1,000,000 citizens’ data images do not have a consistent x-->y mapping as the rest of the data (similar to the New York City/Detroit housing prices example from lecture).



This would cause the dev and test set distributions to become different. This is a bad idea because you’re not aiming where you want to hit.



The test set no longer reflects the distribution of data (security cameras) you most care about.

8.

Question 8

You ask a few people to label the dataset so as to find out what is human-level performance. You find the following levels of accuracy:

|  |  |
| --- | --- |
| Bird watching expert #1 | 0.3% error |
| Bird watching expert #2 | 0.5% error |
| Normal person #1 (not a bird watching expert) | 1.0% error |
| Normal person #2 (not a bird watching expert) | 1.2% error |

If your goal is to have “human-level performance” be a proxy (or estimate) for Bayes error, how would you define “human-level performance”?

1 point



0.0% (because it is impossible to do better than this)



0.3% (accuracy of expert #1)



0.4% (average of 0.3 and 0.5)



0.75% (average of all four numbers above)

11.

Question 11

You also evaluate your model on the test set, and find the following:

|  |  |
| --- | --- |
| Human-level performance | 0.1% |
| Training set error | 2.0% |
| Dev set error | 2.1% |
| Test set error | 7.0% |

What does this mean? (Check the two best options.)

1 point



You should get a bigger test set.



You have overfit to the dev set.



You have underfit to the dev set.



You should try to get a bigger dev set.

14.

Question 14

You’ve handily beaten your competitor, and your system is now deployed in Peacetopia and is protecting the citizens from birds! But over the last few months, a new species of bird has been slowly migrating into the area, so the performance of your system slowly degrades because your data is being tested on a new type of data.



You have only 1,000 images of the new species of bird. The city expects a better system from you within the next 3 months. Which of these should you do first?

1 point



Use the data you have to define a new evaluation metric (using a new dev/test set) taking into account the new species, and use that to drive further progress for your team.



Put the 1,000 images into the training set so as to try to do better on these birds.



Try data augmentation/data synthesis to get more images of the new type of bird.



Add the 1,000 images into your dataset and reshuffle into a new train/dev/test split.

15.

Question 15

The City Council thinks that having more Cats in the city would help scare off birds. They are so happy with your work on the Bird detector that they also hire you to build a Cat detector. (Wow Cat detectors are just incredibly useful aren’t they.) Because of years of working on Cat detectors, you have such a huge dataset of 100,000,000 cat images that training on this data takes about two weeks. Which of the statements do you agree with? (Check all that agree.)

1 point



Needing two weeks to train will limit the speed at which you can iterate.



If 100,000,000 examples is enough to build a good enough Cat detector, you might be better of training with just 10,000,000 examples to gain a \approx≈10x improvement in how quickly you can run experiments, even if each model performs a bit worse because it’s trained on less data.



Buying faster computers could speed up your teams’ iteration speed and thus your team’s productivity.



Having built a good Bird detector, you should be able to take the same model and hyperparameters and just apply it to the Cat dataset, so there is no need to iterate.

#### Autonomous driving (case study)

10.

Question 10

You decide to use data augmentation to address foggy images. You find 1,000 pictures of fog off the internet, and “add” them to clean images to synthesize foggy days, like this:



Which of the following statements do you agree with?

1 point



Adding synthesized images that look like real foggy pictures taken from the front-facing camera of your car to training dataset won’t help the model improve because it will introduce avoidable-bias.



There is little risk of overfitting to the 1,000 pictures of fog so long as you are combing it with a much larger (>>1,000) of clean/non-foggy images.



So long as the synthesized fog looks realistic to the human eye, you can be confident that the synthesized data is accurately capturing the distribution of real foggy images (or a subset of it), since human vision is very accurate for the problem you’re solving.

#### The basics of ConvNets

1.

Question 1

What do you think applying this filter to a grayscale image will do?

⎡⎣⎢⎢01101331−1−3−3−10−1−10⎤⎦⎥⎥

⎣⎢⎢⎢⎡​0110​1331​−1−3−3−1​0−1−10​⎦⎥⎥⎥⎤​

1 point



Detect image contrast



Detect horizontal edges



Detect 45 degree edges



Detect vertical edges

2.

Question 2

Suppose your input is a 300 by 300 color (RGB) image, and you are not using a convolutional network. If the first hidden layer has 100 neurons, each one fully connected to the input, how many parameters does this hidden layer have (including the bias parameters)?

1 point



9,000,001



9,000,100



27,000,001



27,000,100

3.

Question 3

Suppose your input is a 300 by 300 color (RGB) image, and you use a convolutional layer with 100 filters that are each 5x5. How many parameters does this hidden layer have (including the bias parameters)?

1 point



2501



2600



7500



7600

8.

Question 8

Because pooling layers do not have parameters, they do not affect the backpropagation (derivatives) calculation.

1 point



True



False

#### Deep convolutional models

**TOTAL POINTS 10**

3.

Question 3

In order to be able to build very deep networks, we usually only use pooling layers to downsize the height/width of the activation volumes while convolutions are used with “valid” padding. Otherwise, we would downsize the input of the model too quickly.

1 point



True



False

7.

Question 7

Suppose you have an input volume of dimension 64x64x16. How many parameters would a single 1x1 convolutional filter have (including the bias)?

1 point



1



17



4097



2

8.

Question 8

Suppose you have an input volume of dimension n\_H*nH*​ x n\_W*nW*​ x n\_C*nC*​. Which of the following statements you agree with? (Assume that “1x1 convolutional layer” below always uses a stride of 1 and no padding.)

1 point



You can use a 1x1 convolutional layer to reduce n\_H*nH*​, n\_W*nW*​, and n\_C*nC*​.



You can use a pooling layer to reduce n\_H*nH*​, n\_W*nW*​, but not n\_C*nC*​.



You can use a pooling layer to reduce n\_H*nH*​, n\_W*nW*​, and n\_C*nC*​.



You can use a 1x1 convolutional layer to reduce n\_C*nC*​ but not n\_H*nH*​, n\_W*nW*​.

9.

Question 9

Which ones of the following statements on Inception Networks are true? (Check all that apply.)

1 point



A single inception block allows the network to use a combination of 1x1, 3x3, 5x5 convolutions and pooling.



Inception networks incorporates a variety of network architectures (similar to dropout, which randomly chooses a network architecture on each step) and thus has a similar regularizing effect as dropout.



Making an inception network deeper (by stacking more inception blocks together) should not hurt training set performance.



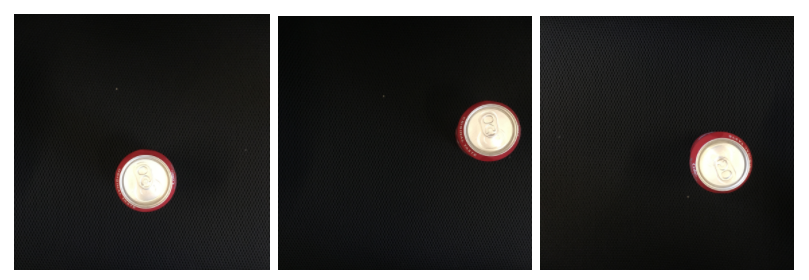
Inception blocks usually use 1x1 convolutions to reduce the input data volume’s size before applying 3x3 and 5x5 convolutions.

#### Detection algorithms

3.

Question 3

You are working on a factory automation task. Your system will see a can of soft-drink coming down a conveyor belt, and you want it to take a picture and decide whether (i) there is a soft-drink can in the image, and if so (ii) its bounding box. Since the soft-drink can is round, the bounding box is always square, and the soft drink can always appears as the same size in the image. There is at most one soft drink can in each image. Here’re some typical images in your training set:



What is the most appropriate set of output units for your neural network?

1 point



Logistic unit (for classifying if there is a soft-drink can in the image)



Logistic unit, b\_x*bx*​ and b\_y*by*​



Logistic unit, b\_x*bx*​, b\_y*by*​, b\_h*bh*​ (since b\_w*bw*​ = b\_h*bh*​)

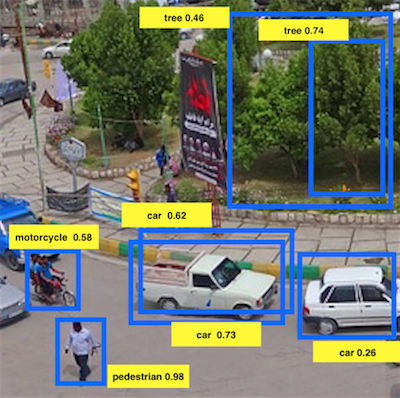


Logistic unit, b\_x*bx*​, b\_y*by*​, b\_h*bh*​, b\_w*bw*​

9.

Question 9

Suppose you run non-max suppression on the predicted boxes above. The parameters you use for non-max suppression are that boxes with probability \leq≤ 0.4 are discarded, and the IoU threshold for deciding if two boxes overlap is 0.5. How many boxes will remain after non-max suppression?



1 point



3



4



5



6



7

#### Special applications: Face recognition & Neural style transfer

3.

Question 3

In order to train the parameters of a face recognition system, it would be reasonable to use a training set comprising 100,000 pictures of 100,000 different persons.

1 point



True



False

7.

Question 7

Neural style transfer is trained as a supervised learning task in which the goal is to input two images (x*x*), and train a network to output a new, synthesized image (y*y*).

1 point



True



False

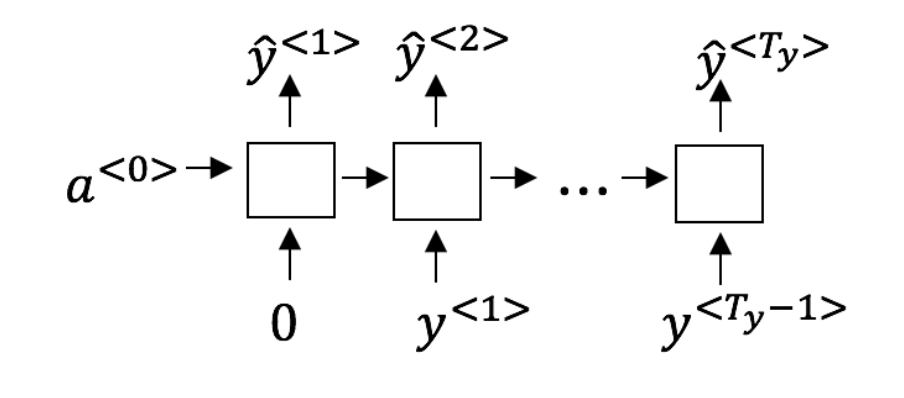
Recurrent Neural Networks

Graded Quiz • 30 min

4.

Question 4

You are training this RNN language model.



At the t^{th}*tth* time step, what is the RNN doing? Choose the best answer.

1 point



Estimating P(y^{<1>}, y^{<2>}, …, y^{<t-1>})*P*(*y*<1>,*y*<2>,…,*y*<*t*−1>)



Estimating P(y^{<t>})*P*(*y*<*t*>)

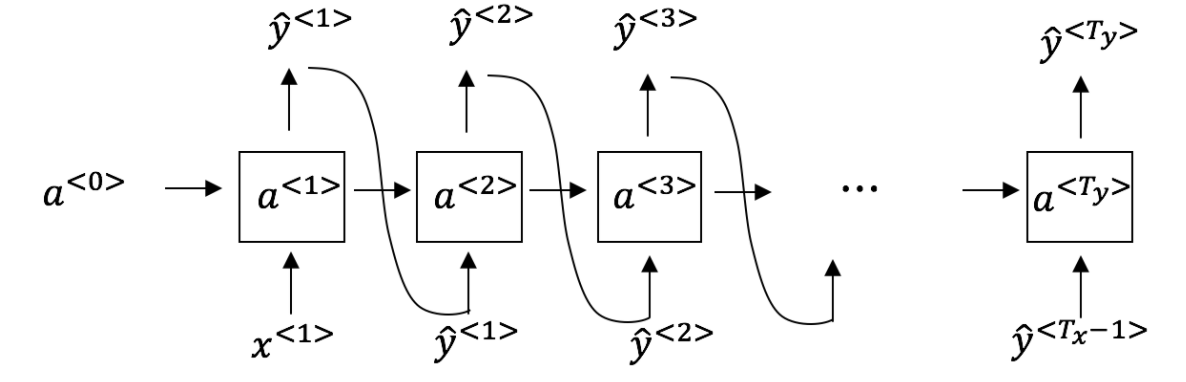


Estimating P(y^{<t>} \mid y^{<1>}, y^{<2>}, …, y^{<t-1>})*P*(*y*<*t*>∣*y*<1>,*y*<2>,…,*y*<*t*−1>)

5.

Question 5

You have finished training a language model RNN and are using it to sample random sentences, as follows:



What are you doing at each time step t*t*?

1 point



(i) Use the probabilities output by the RNN to pick the highest probability word for that time-step as \hat{y}^{<t>}*y*^​<*t*>. (ii) Then pass the ground-truth word from the training set to the next time-step.



(i) Use the probabilities output by the RNN to randomly sample a chosen word for that time-step as \hat{y}^{<t>}*y*^​<*t*>. (ii) Then pass the ground-truth word from the training set to the next time-step.



(i) Use the probabilities output by the RNN to pick the highest probability word for that time-step as \hat{y}^{<t>}*y*^​<*t*>. (ii) Then pass this selected word to the next time-step.



(i) Use the probabilities output by the RNN to randomly sample a chosen word for that time-step as \hat{y}^{<t>}*y*^​<*t*>. (ii) Then pass this selected word to the next time-step.

Natural Language Processing & Word Embeddings

Graded Quiz • 30 min

6.

Question 6

When learning word embeddings, we create an artificial task of estimating P(target \mid context)*P*(*target*∣*context*). It is okay if we do poorly on this artificial prediction task; the more important by-product of this task is that we learn a useful set of word embeddings.

1 point



True



False

8.

Question 8

Suppose you have a 10000 word vocabulary, and are learning 500-dimensional word embeddings. The word2vec model uses the following softmax function:

P(t \mid c) = \frac{e^{\theta\_t^T e\_c}}{\sum\_{t’=1}^{10000} e^{\theta\_{t’}^Te\_c}}*P*(*t*∣*c*)=∑*t*’=110000​*eθt*’*T*​*ec*​*eθtT*​*ec*​​

Which of these statements are correct? Check all that apply.

1 point



\theta\_t*θt*​ and e\_c*ec*​ are both 500 dimensional vectors.



\theta\_t*θt*​ and e\_c*ec*​ are both 10000 dimensional vectors.



\theta\_t*θt*​ and e\_c*ec*​ are both trained with an optimization algorithm such as Adam or gradient descent.



After training, we should expect \theta\_t*θt*​ to be very close to e\_c*ec*​ when t*t* and c*c* are the same word.

9.

Question 9

Suppose you have a 10000 word vocabulary, and are learning 500-dimensional word embeddings.The GloVe model minimizes this objective:

\min \sum\_{i=1}^{10,000} \sum\_{j=1}^{10,000} f(X\_{ij}) (\theta\_i^T e\_j + b\_i + b\_j’ - log X\_{ij})^2min∑*i*=110,000​∑*j*=110,000​*f*(*Xij*​)(*θiT*​*ej*​+*bi*​+*bj*​’−*logXij*​)2

Which of these statements are correct? Check all that apply.

1 point



\theta\_i*θi*​ and e\_j*ej*​ should be initialized to 0 at the beginning of training.



\theta\_i*θi*​ and e\_j*ej*​ should be initialized randomly at the beginning of training.



X\_{ij}*Xij*​ is the number of times word j appears in the context of word i.



The weighting function f(.)*f*(.) must satisfy f(0) = 0*f*(0)=0.

10.

Question 10

You have trained word embeddings using a text dataset of m\_1*m*1​ words. You are considering using these word embeddings for a language task, for which you have a separate labeled dataset of m\_2*m*2​ words. Keeping in mind that using word embeddings is a form of transfer learning, under which of these circumstance would you expect the word embeddings to be helpful?

1 point



m\_1*m*1​ >> m\_2*m*2​



m\_1*m*1​ << m\_2*m*2​

Sequence models & Attention mechanism

Graded Quiz • 30 min

3.

Question 3

In machine translation, if we carry out beam search without using sentence normalization, the algorithm will tend to output overly short translations.

1 point



True



False

9.

Question 9

Under the CTC model, identical repeated characters not separated by the “blank” character (\_) are collapsed. Under the CTC model, what does the following string collapse to?

\_\_c\_oo\_o\_kk\_\_\_b\_ooooo\_\_oo\_\_kkk

1 point



cokbok



cookbook



cook book



coookkboooooookkk