DPL-Course1->Course4

HÍc trñc tuy¿n t¡i https://quizlet.com/_djkr74

1. **Which of the following are some aspects in which AI has transformed business?**

   A-Web searching and advertisement.
   B-Creating an AI-powered society.
   C-Eliminating the need for health care services.
   D-AI has not been able to transform businesses.

   A-Web searching and advertisement.

   AI has helped to make a fit between services or results and consumers or queries.

2. **Which of the following play a major role to achieve a very high level of performance with Deep Learning algorithms?**
   A-Large amounts of data.
   B-Better designed features to use.
   C-Large models.
   D-Deep learning has resulted in significant improvements in important applications such as online advertising, speech recognition, and image recognition.
   E-Smaller models.

   A-Large amounts of data.

   C-Large models
   D-Deep learning has resulted in significant improvements in important applications such as online advertising, speech recognition, and image recognition.

3. **Recall the diagram of iterating over different ML ideas. Which of the stages shown in the diagram was improved with the use of a better GPU/CPU?**

   A-Experiments finish faster, producing better ideas through increased iteration tempo.
   B-Without better hardware, there is no way to train models faster.
   C-Some algorithms are specifically designed to run experiments faster.
   D-With larger datasets, the iteration process is faster.

   A-Experiments finish faster, producing better ideas through increased iteration tempo.

   C-Some algorithms are specifically designed to run experiments faster.

4. **When experienced deep learning engineers work on a new problem, they can usually use insight from previous problems to train a good model on the first try, without needing to iterate**

   False

   Finding the characteristics of a model is key

	multiple times through different models. **True/False?**	to having good performance. Although experience can help, it requires multiple iterations to build a good model.
5.	**Images for cat recognition is an example of "structured" data, because it is represented as a structured array in a computer.** **True/False?**	False Images for cat recognition are examples of "unstructured" data.
6.	**A demographic dataset with statistics on different cities' population, GDP per capita, and economic growth is an example of "unstructured" data because it contains data coming from different sources.** **True/False?**	False A demographic dataset with statistics on different cities' population, GDP per capita, and economic growth is an example of "structured" data in contrast to image, audio or text datasets.
7.	**Why is an RNN (Recurrent Neural Network) used for machine translation, say translating English to French? (Check all that apply.)** **A-It can be trained as a supervised learning problem.** **B-RNNs represent the recurrent process of Idea->Code->Experiment->Idea->....** **C-It is strictly more powerful than a Convolutional Neural Network (CNN).** **D-It is applicable when the input/output is a sequence (e.g., a sequence of words).**	A-It can be trained as a supervised learning problem. D-It is applicable when the input/output is a sequence (e.g., a sequence of words).
8.	**diagram: Performanc <- Amount of data** **Suppose the information given in the diagram is accurate. We can deduce that when using large**	True the graph shows that after a certain amount of

	training sets, for a model to keep improving as the amount of data for training grows, the size of the neural network must grow. True/False?	data is fed to a NN it stops increasing its performance. To increase the performance it is necessary to use a larger model.
9.	**diagram: Performanc <- Amount of data Assuming the trends described in the figure are accurate. Which of the following statements are true? Choose all that apply. A-Increasing the training set size of a traditional learning algorithm stops helping to improve the performance after a certain size. B-Decreasing the training set size generally does not hurt an algorithm's performance, and it may help significantly. C-Increasing the training set size of a traditional learning algorithm always improves its performance. D-Increasing the size of a neural network generally does not hurt an algorithm's performance, and it may help significantly.**	A-Increasing the training set size of a traditional learning algorithm stops helping to improve the performance after a certain size. D-Increasing the size of a neural network generally does not hurt an algorithm's performance, and it may help significantly.
10.	****Which of these are reasons for Deep Learning recently taking off? (Check the three options that apply.) A-Neural Networks are a brand new field. B-Deep learning has resulted in significant improvements in important applications such as online advertising, speech recognition, and image recognition. C-We have access to a lot more data. The digitalization of our society has played a huge role in this.**	B-Deep learning has resulted in significant improvements in important applications such as online advertising, speech recognition, and image recognition. C-We have access to a lot more data. The digitalization of our society has played a huge role in this. D-We have access to a

	D-We have access to a lot more computational power.	lot more computational power.
11.	****Recall this diagram of iterating over different ML ideas. Which of the statements below are true? (Check all that apply.)** **Idea->Code->Experiment->Idea** **A-Being able to try out ideas quickly allows deep learning engineers to iterate more quickly** **B-Recent progress in deep learning algorithms has allowed us to train good models faster (even without changing the CPU/GPU hardware). For example, we discussed how switching from sigmoid to ReLU activation functions allows faster training.** **C-It is faster to train on a big dataset than a small dataset.** **D-Faster computation can help speed up how long a team takes to iterate to a good idea.**	A-Being able to try out ideas quickly allows deep learning engineers to iterate more quickly B-Recent progress in deep learning algorithms has allowed us to train good models faster (even without changing the CPU/GPU hardware). For example, we discussed how switching from sigmoid to ReLU activation functions allows faster training. D-Faster computation can help speed up how long a team takes to iterate to a good idea.
12.	**Neural networks are good at figuring out functions relating an input x to an output y given enough examples.** **True/False?**	True. with neural networks, we don't need to "design" features by ourselves. The neural network figures out the necessary relations given enough data.
13.	****Which of the following are examples of unstructured data? Choose all that apply.** **A-Information about elephants' weight, height, age, and the number of offspring.** **B-Sound files for speech recognition.**	B-Sound files for speech recognition. C-Images for bird recognition. D-Text describing size

	C-Images for bird recognition. **D-Text describing size and number of pages of books.**	and number of pages of books.
14.	**Which of the following are examples of structured data? Choose all that apply.** **A-A set of audio recordings of a person saying a single word.** **B-A dataset with zip code, income, and name of a person.** **C-A dataset of weight, height, age, the sugar level in the blood, and arterial pressure.** **D-A dataset with short poems.**	B-A dataset with zip code, income, and name of a person. C-A dataset of weight, height, age, the sugar level in the blood, and arterial pressure.
15.	**Assuming the trends described in the figure are accurate.The performance of a NN depends only on the size of the NN.** **True/False?**	False. According to the trends in the figure above, It also depends on the amount of data.
16.	**What does the analogy "AI is the new electricity" refer to?** **A-AI runs on computers and is thus powered by electricity, but it is letting computers do things not possible before.** **B-AI is powering personal devices in our homes and offices, similar to electricity.** **C-Similar to electricity starting about 100 years ago, AI is transforming multiple industries.** **D-Through the "smart grid", AI is delivering a new wave of electricity.**	C-Similar to electricity starting about 100 years ago, AI is transforming multiple industries. AI is transforming many fields from the car industry to agriculture to supply-chain...
17.	**When building a neural network to predict housing price from features like size, the number of bedrooms, zip code, and wealth, it is necessary to come up with other features in between input and output like family size and school quality.** **True/False?**	False. A neural network figures out by itself the "features" in between using the samples used to train it.

18. **Why can an RNN (Recurrent Neural Network) be used to create English captions to French movies? Choose all that apply.**

A-The RNN is applicable since the input and output of the problem are sequences.
B-RNNs are much more powerful than a Convolutional neural Network (CNN).
C-The RNN requires a small number of examples.
D-It can be trained as a supervised learning problem.

A-The RNN is applicable since the input and output of the problem are sequences.

D-It can be trained as a supervised learning problem.

19. **Which of the following are reasons that didn't allow Deep Learning to be developed during the '80s?**
A-People were afraid of a machine rebellion.
B-Limited computational power.
C-The theoretical tools didn't exist during the 80's.
D-Interesting applications such as image recognition require large amounts of data that were not available.

B-Limited computational power.

D-Interesting applications such as image recognition require large amounts of data that were not available.

20. **ReLU stands for which of the following?**
A-Rectified Last Unit
B-Rectified Linear Unit
C-Recognition Linear Unit
D-Representation Linear Unit

B-Rectified Linear Unit

21. **diagram: Performanc <- Amount of data**

Assuming the trends described in the previous question's figure are accurate (and hoping you got the axis labels right), which of the following are true? (Check all that apply.)

A-Increasing the training set size generally does not hurt an algorithm's performance, and

A-Increasing the training set size generally does not hurt an algorithm's performance, and it may help significantly.

C-Increasing the size of a neural network generally does not hurt an al-

	it may help significantly. **B-Decreasing the training set size generally does not hurt an algorithm's performance, and it may help significantly.** **C-Increasing the size of a neural network generally does not hurt an algorithm's performance, and it may help significantly.** **D-Decreasing the size of a neural network generally does not hurt an algorithm's performance, and it may help significantly.**	gorithm's performance, and it may help significantly.
22.	**RNNs (Recurrent Neural Networks) are good for data with a temporal component.** **True/False?**	True. RNN are good to work with sequences, and the elements of the sequence can be sorted by a temporal component.
23.	**diagram: Performanc <- Amount of data** **From the given diagram, we can deduce that Large NN models are always better than traditional learning algorithms.** **True/False?**	False. when the amount of data is not large the performance of traditional learning algorithms is shown to be the same as NN.
24.	**Which of the following best describes the role of AI in the expression "an AI-powered society"?** **A-AI controls the power grids for energy distribution, so all the power needed for industry and in daily life comes from AI.** **B-AI is an essential ingredient in realizing tasks, in industry and in personal life.** **C-AI helps to create a more efficient way of pro-**	B-AI is an essential ingredient in realizing tasks, in industry and in personal life. In an AI-powered society AI plays a fundamental role to complete most tasks, in industry and personal life.

ducing energy to power industries and personal devices.

25. **Features of animals, such as weight, height, and color, are used for classification between cats, dogs, or others. This is an example of "structured" data, because they are represented as arrays in a computer. True/False?**

FalseNo. The data can be represented by columns of data. This is an example of structured data, unlike images of the animal.

TrueYes. The data can be represented by columns of data. This is an example of structured data, unlike images of the animal.

TrueYes. The data can be represented by columns of data. This is an example of structured data, unlike images of the animal.

26. **A dataset is composed of age and weight data for several people. This dataset is an example of "structured" data because it is represented as an array in a computer.**

True/False?

True

Yes, the sequences can be represented as arrays in a computer. This is an example of structured data.

27. **Recall this diagram of iterating over different ML ideas. Which of the statements below are true? (Check all that apply.)**

Graph:
Idea->Code->Experiment->Idea

A-Better algorithms allow engineers to get more data and then produce better Deep Learning models.
B-Improvements in the GPU/CPU hardware enable the discovery of better Deep Learning algorithms.
C-Larger amounts of data allow researchers to

B-Improvements in the GPU/CPU hardware enable the discovery of better Deep Learning algorithms.

D-Better algorithms can speed up the iterative process by reducing the necessary computation time.

	try more ideas and then produce better algorithms in less time. **D-Better algorithms can speed up the iterative process by reducing the necessary computation time.**	
28.	**What does a neuron compute?** **A-A neuron computes a linear function z=Wx+b followed by an activation function** **B-A neuron computes the mean of all features before applying the output to an activation function** **C-A neuron computes an activation function followed by a linear function z=Wx+b** **D-A neuron computes a function g that scales the input x linearly (Wx + b)**	A-A neuron computes a linear function z=Wx+b followed by an activation function we generally say that the output of a neuron is a = g(Wx + b) where g is the activation function (sigmoid, tanh, ReLU, ...).
29.	**Which of these is the "Logistic Loss"?** **A-L(i)(^y(i),y(i)) = max(0,y(i) ^y(i))** **B-L(i)(^y(i),y(i)) =** **(y(i)log(^y(i))+(1 y(i))log(1 ^y(i))** **C-L(i)(^y(i),y(i)) = #y(i) ^y(i)#2** **D-L(i)(^y(i),y(i)) = #y(i) ^y(i)#**	B-L(i)(^y(i),y(i)) = (y(i)log(^y(i))+(1 y(i))log(1 ^y
30.	**Suppose that y^=0.5 and y=0. What is the value of the "Logistic Loss"? Choose the best option.** **A-0.693** **B-+** **C-L(^y,y)= (^ylogy+(1 ^y) log(1 y))** **D-0.5**	A-0.693 Yes. Given the values of y^ and y we get L(0.5,0)= (0log0.5+1log(0.5)
31.	**Suppose that y^=0.9 and y=1. What is the value of the "Logistic Loss"? Choose the best option.** **A-0.005** **B-L(^y,y)= (^ylogy+(1 ^y) log(1 y))** **C-0.105** **D-+**	C-0.105 Yes. Given the values of y^ and y we get L(0.9,1)= (1.log0.9+0.log(0.1

32.	**Suppose img is a (32,32,3) array, representing a 32x32 image with 3 color channels red, green and blue. How do you reshape this into a column vector x?
	A-x = img.reshape((1,32*32,3))
	B-x = img.reshape((32*32*3,1))
	C-x = img.reshape((3,32*32))
	D-x = img.reshape((32*32,3))**
		B-x = img.reshape((32*32*3,1))

33.	**Suppose x is a (8, 1) array. Which of the following is a valid reshape?
	A-x.reshape(-1, 3)
	B-x.reshape(2, 4, 4)
	C-x.reshape(2, 2, 2)
	D-x.reshape(1, 4, 3)**
		C-x.reshape(2, 2, 2)

34.	**Consider the two following random arrays a and b:
	a=np.random.randn(1,3) #a.shape=(1,3)
	b=np.random.randn(3,3) #b.shape=(3,3)
	c=a b
	What will be the shape of c?
	
	A-c.shape = (1, 3)
	B-c.shape = (3, 3)
	C-The computation cannot happen because the sizes don't match.
	D-The computation cannot happen because it is not possible to broadcast more than one dimension.**
		B-c.shape = (3, 3)

35.	**Consider the two following random arrays a and b:
	a=np.random.randn(1,3) #a.shape=(3,3)
	b=np.random.randn(3,3) #b.shape=(2,1)
	c=a b
	What will be the shape of c?
	
	A-c.shape = (1, 3)**
		D-The computation cannot happen because it is not possible to broadcast more than one dimension.

B-c.shape = (3, 3)
C-The computation cannot happen because the sizes don't match.
D-The computation cannot happen because it is not possible to broadcast more than one dimension.

36. **Suppose you have n_x input features per example. If we decide to use row vectors x_j for the features and X =**
[x_1
x_2
...
x_m]
What is the dimension of X?
A-(n_x, m)
B-(n_x, n_x)
C-(m, n_x)
D-(1, n_x)

C-(m, n_x)

Each x_j has dimension 1 x n_x, X is built stacking all rows together into a m x n_x array.

37. **Suppose you have n_x input features per example. Recall that X = [x_1 x_2 ... x_m]**
What is the dimension of X?
A-(n_x, m)
B-(n_x, n_x)
C-(m, n_x)
D-(1, n_x)

A-(n_x, m)

38. **Considering the following array:**
a = np.array([[2,1], [1,3]])
What is the result of a*a?

A-The computation cannot happen because the size don't match. It's going to be "Error"!
B-[[5,5], [5,10]]
C-[[4,2], [2,6]]
D-[[4,1], [1,9]]

D-[[4,1], [1,9]]

39. **Considering the following array:**
a = np.array([[2,1], [1,3]])

B-[[5,5], [5,10]]

What is the result of np.dot(a, a)?

A-The computation cannot happen because the size don't match. It's going to be "Error"!
B-[[5,5], [5,10]]
C-[[4,2], [2,6]]
D-[[4,1], [1,9]]

40. **Recall that "np.dot(a,b)" performs a matrix multiplication on a and b, whereas "a*b" performs an element-wise multiplication.**
Consider the two following random arrays "a" and "b":

```
a = np.random.randn(12288, 150) # a.shape = (12288, 150)
b = np.random.randn(150, 45) # b.shape = (150, 45)
c = np.dot(a,b)
```

What is the shape of c?

A. c.shape = (150,150)
B. c.shape = (12288, 150)
C. c.shape = (12288, 45)
D. The computation cannot happen because the sizes don't match. It's going to be "Error"!

C. c.shape = (12288, 45)

remember that a np.dot(a, b) has shape (number of rows of a, number of columns of b). The sizes match because: "number of columns of a = 150 = number of rows of b"

41. **consider the following code snippet:**

```
a.shape = (4,3)
b.shape = (4,1)

for i in range(3):
for j in range(4):
c[i][j] = a[i][j] + b[j]
```

How do you vectorize this?
A-c = a + b
B-c = a.T + b

B-c = a.T+b. False. Notice that b is a column vector; but we are using it to fill the row i of c.

maybe D

C-c = a + b.T
D-c = a.T + b.T

42. **consider the following code snippet:**
a.shape = (4,3)
b.shape = (4,1)

for i in range(3):
for j in range(4):
c[i][j] = a[j][i] + b[j]

How do you vectorize this?
A-c = a + b
B-c = a.T + b
C-c = a + b.T
D-c = a.T + b.T

D-c = a.T + b.T

43. **consider the following code snippet:**
a.shape = (3,4)
b.shape = (4,1)

for i in range(3):
for j in range(4):
c[i][j] = a[i][j] * b[j]

How do you vectorize this?
A-c = a * b
B-c = a.T * b
C-c = a * b.T
D-c = np.dot(a,b)

C-c = a * b.T

44. **consider the following code snippet:**
a.shape = (3,4)
b.shape = (4,1)

for i in range(3):
for j in range(4):
c[i][j] = a[j][i] * b[j]

C-c = a.T * b.T

How do you vectorize this?
A-c = a * b
B-c = a.T * b
C-c = a.T * b.T
D-c = a * b.T

45. **Considering the following array:**
a = np.array([[1,1], [1,-1]])
b = np.array([[2], [3]])
c = a + b
Which of the following arrays is stored in c?
A- [[3,3], [3,1], [4,4], [5,2]]
B- The computation cannot happen because the sizes don't match. It's going to be an "Error"!
C- [[3,4], [3,2]]
D- [[3,3], [4,2]]

D- [[3,3], [4,2]]

46. **Consider the code snippet:**
a.shape=(3,3)
b.shape=(3,3)
c=a 2+b.T 2
Which of the following gives an equivalent output for c?

A-for i in range(3):
for j in range(3):
c[i][j] = a[i][j]2 + b[j][i]**2**
B-for i in range(3):
for j in range(3):
c[i][j] = a[i][j]2 + b[i][j]**2**
C-The computation cannot happen because the sizes don't match. It's going to be an "Error"!
D-for i in range(3):
c[i] = a[i]2 + b[i]**2**

A-for i in range(3):
for j in range(3):
c[i][j] = a[i][j]**2 + b[j][i]**2

This code squares each entry of a and adds it to the transpose of b square.

47. **Consider the following computational graph.**
+u = a*b
+v = a+c
+w= b*c

A-(a+c),(b 1)

J=u v+w=ab (a+c)+bc=ab a-

	**--> J = u-v+w
	What is the output of J?
	A-(a+c),(b 1)
	B-ab+bc+ac
	C-(c 1),(a+c)
	D-(a 1),(b+c)**	

48.	**Consider the following computational graph. +u = a*b +v = a*c +w= b+c --> J = u+v-w What is the output of J? A- a*b + b*c + c*a B- (a-1)*(b+c) C- (c-1)*(b+a) D- (b-1)*(c+a)**	B- (a-1)*(b+c)

49.	**In logistic regression given x and parameters w R^n_x , b R. Which of the following best expresses what we want yhat to tell us?
	A-P(y=yhat|x)
	B-ÃW(x+b)
	C-ÃW(x)
	D-P(y=1 | x)**
		D-P(y=1 | x)
		
		We want the output yhat to tell us the probability that y=1 given x.

50.	**Suppose our input batch consists of 8 grayscale images, each of dimension 8x8. We reshape these images into feature column vectors xj. Remember that X=[x(1)x(2)ïx(8)]. What is the dimension of X?
	
	A-(8, 64)
	B-(8, 8, 8)
	C-(512, 1)
	D-(64, 8)**
		D-(64, 8)
		
		B-No. After converting the 8x8 gray scale images to a column vector we get a vector of size 64, thus X has dimension (64,8).

51.	**Consider the following random arrays a and b, and c:**	B-c.shape = (3, 4) Yes. Broadcasting is

	**a=np.random.randn(3,4) # b=np.random.randn(1,4) #
	c=a+b
	What will be the shape of c?
	
	A-c.shape = (1, 4)
	B-c.shape = (3, 4)
	C-c.shape = (3, 1)
	D-The computation cannot happen because it is not possible to broadcast more than one dimension.**
		used, so row b is copied 3 times so it can be summed to each row of a.

52.	**Consider the two following random arrays a and b:
	a=np.random.randn(1,3) #a.shape=(4,3)
	b=np.random.randn(3,3) #b.shape=(1,3)
	c=a b
	What will be the shape of c?
	
	A-c.shape = (1, 3)
	B-c.shape = (4, 3)
	C-The computation cannot happen because the sizes don't match.
	D-The computation cannot happen because it is not possible to broadcast more than one dimension.**
		B-c.shape = (4, 3)
		
		Yes. Broadcasting is invoked, so row b is multiplied element-wise with each row of a to create c.

53.	**In logistic regression given the input x, and parameters w Rnx , b R, how do we generate the output y^ ?
	
	A-Ã(Wx)
	B-Wx+b
	C-tanh(Wx+b)
	D-Ã(Wx+b)**
		D-Ã(Wx+b)

54.	**Which of the following are true? (Check all that apply.)**	E-w[4]_3 is the column vector of parameters of the fourth layer and third

A-w[4]_3 is the column vector of parameters of the third layer and fourth neuron.
B-W_1 is a matrix with rows equal to the parameter vectors of the first layer.
C-w[4]_3 is the row vector of parameters of the fourth layer and third neuron.
D-W[1] is a matrix with rows equal to the parameter vectors of the first layer.
E-w[4]_3 is the column vector of parameters of the fourth layer and third neuron.
F-W[1] is a matrix with rows equal to the transpose of the parameter vectors of the first layer.

neuron.
F-W[1] is a matrix with rows equal to the transpose of the parameter vectors of the first layer.

55. **The tanh activation is not always better than sigmoid activation function for hidden units because the mean of its output is closer to zero, and so it centers the data, making learning complex for the next layer.**

True/False?

False.

As seen in lecture the output of the tanh is between -1 and 1, it thus centers the data which makes the learning simpler for the next layer.

56. **Which of the following are true about the tanh function?**

A-For large values the slope is larger.
B-The derivative at c=0 is not well defined.
C-For large values the slope is close to zero.
D-The tanh is mathematically a shifted version of the sigmoid function.
E-The slope is zero for negative values.

C-For large values the slope is close to zero.
D-The tanh is mathematically a shifted version of the sigmoid function.

57. **In which of the following cases is the linear (identity) activation function most likely used?**

A-For binary classification problems.
B-The linear activation function is never used.
C-As activation function in the hidden layers.
D-When working with regression problems.

D-When working with regression problems.

In problems such as predicting the price of a house it makes sense to use the linear activation function as output.

58. **The sigmoid function is only mentioned as an activation function for historical reasons. The tanh is always preferred without exceptions in all the layers of a Neural Network.**

True/False?

False.

Although the tanh almost always works better than the sigmoid function when used in hidden layers, thus is always proffered as activation function, the exception is for the output layer in classification problems.

59. **A single output and single layer neural network that uses the sigmoid function as activation is equivalent to the logistic regression.**

True/False

True

60. **You are building a binary classifier for recognizing cucumbers (y=1) vs. watermelons (y=0). Which one of these activation functions would you recommend using for the output layer?**

A-tanh
B-sigmoid
C-Leaky ReLU
D-ReLU

B-sigmoid

Sigmoid outputs a value between 0 and 1 which makes it a very good choice for binary classification. You can classify as 0 if the output is less than 0.5 and classify as 1 if the output is more than 0.5. It can be done with tanh as well but it is less convenient as the output is between -1 and 1.

61. **When building a binary classifier for recognizing cats (y=1) vs raccoons (y=0). Is better to use**

False.

	the sigmoid function as activation function for the hidden layers. **True/False?**	Using tanh almost always works better than the sigmoid function for hidden layers.
62.	**Which of the following represents the activation output of the second neuron of the third layer applied to the fourth example?**	a[3](4)_2 The superscript in brackets indicates the layer number, the superscript in parenthesis represents the number of examples, and the subscript the number of the neuron.
63.	**Suppose you have built a neural network with one hidden layer and tanh as activation function for the hidden layer. You decide to initialize the weights to small random numbers and the biases to zero. The first hidden layer's neurons will perform different computations from each other even in the first iteration. True/False?** **False No. Since the weights are most likely different, each neuron will do a different computation.** **True Yes. Since the weights are most likely different, each neuron will do a different computation.**	True Yes. Since the weights are most likely different, each neuron will do a different computation.
64.	**Using linear activation functions in the hidden layers of a multilayer neural network is equivalent to using a single layer.** **True/False?**	True. When the identity or linear activation function g(c)=c is used the output of composition of layers is equivalent to the com-

		putations made by a single layer.
65.	**The use of the ReLU activation function is becoming more rare because the ReLU function has no derivative for c=0.** **True/False?**	False. Although the ReLU function has no derivative at c=0 this rarely causes any problems in practice. Moreover it has become the default activation function in many cases, as explained in the lectures.

66.	**Consider the following code:**
	A = np.random.randn(4,3)
	B = np.sum(A, axis = 1, keepdims = True)
	What will be B.shape? (If you're not sure, feel free to run this in python to find out).
	A-(4, 1)
	B-(4,)
	C-(1, 3)
	D-(3,)
		A-(4,1)
		we use (keepdims = True) to make sure that A.shape is (4,1) and not (4,). It makes our code more robust.

67.	**Consider the following code:** **#+begin_src python** **x = np.random.rand(3, 2)** **y = np.sum(x, axis=0, keepdims=True)** **#+end_src** **What will be y.shape?** **A-(1, 2)** **B-(3,)** **C-(2,)** **D-(3, 1)**	A-(1, 2)
68.	**Consider the following code:** **#+begin_src python**	D-(4,)

```
x = np.random.rand(4, 5)
y = np.sum(x, axis=1)
#+end_src
```

What will be y.shape?

A-(1, 5)
B-(4, 1)
C-(5,)
D-(4,)

69. **Suppose you have built a neural network with one hidden layer and tanh as activation function for the hidden layers. Which of the following is a best option to initialize the weights?**

A-Initialize all weights to a single number chosen randomly.
B-Initialize the weights to small random numbers.
C-Initialize all weights to 0.
D-Initialize the weights to large random numbers.

B-Initialize the weights to small random numbers.

The use of random numbers helps to "break the symmetry" between all the neurons allowing them to compute different functions. When using small random numbers the values z[k] will be close to zero thus the activation values will have a larger gradient speeding up the training process.

70. **Suppose you have built a neural network. You decide to initialize the weights and biases to be zero. Which of the following statements is true?**

A-Each neuron in the first hidden layer will perform the same computation in the first iteration. But after one iteration of gradient descent they will learn to compute different things because we have "broken symmetry".
B-Each neuron in the first hidden layer will compute the same thing, but neurons in different layers will compute different things, thus

C-Each neuron in the first hidden layer will perform the same computation. So even after multiple iterations of gradient descent, each neuron in the layer will be computing the same thing as other neurons.

we have accomplished "symmetry breaking" as described in the lecture.
C-Each neuron in the first hidden layer will perform the same computation. So even after multiple iterations of gradient descent, each neuron in the layer will be computing the same thing as other neurons.
D-The first hidden layer's neurons will perform different computations from each other even in the first iteration; their parameters will thus keep evolving in their own way.

71. **Logistic regression's weights 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?

False.

Logistic Regression doesn't have a hidden layer. If you initialize the weights to zeros, the first example x fed into the logistic regression will output zero but the derivatives of the Logistic Regression depend on the input x (because there's no hidden layer) which is not zero. So at the second iteration, the weights' values follow x's distribution and are different from each other if x is not a constant vector.

72. **You have built a network using the tanh activation for all the hidden units. You initialize the weights to relatively large values, using np.random.randn(..,..)*1000. What will happen?**

A-This will cause the inputs of the tanh to also

A-This will cause the inputs of the tanh to also be very large, thus causing gradients to be close to zero. The optimization algorithm will thus be-

	be very large, thus causing gradients to be close to zero. The optimization algorithm will thus become slow. **B-This will cause the inputs of the tanh to also be very large, causing the units to be "highly activated" and thus speed up learning compared to if the weights had to start from small values.** **C-So long as you initialize the weights randomly gradient descent is not affected by whether the weights are large or small.** **D-This will cause the inputs of the tanh to also be very large, thus causing gradients to also become large. You therefore have to set ±to a very small value to prevent divergence; this will slow down learning.**	come slow. tanh becomes flat for large values; this leads its gradient to be close to zero. This slows down the optimization algorithm.
73.	**Which of the following are true? (Check all that apply.)** **A-w[4]_3 is the row vector of parameters of the fourth layer and third neuron.** **B-w[4]_3 is the column vector of parameters of the fourth layer and third neuron.** **C-a[3](2) denotes the activation vector of the second layer for the third example.** **D-a[2]_3 denotes the activation vector of the second layer for the third example.** **E-a[2] denotes the activation vector of the second layer.** **F-w[4]_3 is the column vector of parameters of the third layer and fourth neuron.**	B-w[4]_3 is the column vector of parameters of the fourth layer and third neuron. E-a[2] denotes the activation vector of the second layer.

74.	**Which of the following is a correct vectorized implementation of forward propagation for layer 2?**
	
	A-
	Z[2]=W[2]A[1]+b[2]
	A[2]=g[2](Z[2])
		A-
		Z[2]=W[2]A[1]+b[2]
		A[2]=g[2](Z[2])

	**B-
	Z[2]=W[2]A[1]+b[2]
	A[2]=g(Z[2])
	C-
	Z[1]=W[1]X+b[1]
	A[1]=g[1](Z[1])
	D-
	Z[2]=W[2]X+b[2]
	A[2]=g[2](Z[2])**	

75.	**Which of the following is true about the ReLU activation functions? A-They are increasingly being replaced by the tanh in most cases. B-They are only used in the case of regression problems, such as predicting house prices. C-They cause several problems in practice because they have no derivative at 0. That is why Leaky ReLU was invented. D-They are the go to option when you don't know what activation function to choose for hidden layers.**	A,C-False Maybe D?
76.	**What is the "cache" used for in our implementation of forward propagation and backward propagation? A-We use it to pass Z computed during forward propagation to the corresponding backward propagation step. It contains useful values for backward propagation to compute derivatives. B-It is used to keep track of the hyperparameters that we are searching over, to speed up computation. C-It is used to cache the intermediate values of the cost function during training. D-We use it to pass variables computed dur-**	A-We use it to pass Z computed during forward propagation to the corresponding backward propagation step. It contains useful values for backward propagation to compute derivatives. Correct, the "cache" records values from the forward propagation units and are used in backward propagation units because it is need-

ing backward propagation to the corresponding forward propagation step. It contains useful values for forward propagation to compute activations.

ed to compute the chain rule derivatives.

77. **We use the "cache" in our implementation of forward and backward propagation to pass useful values to the next layer in the forward propagation.**

True/False?

False.

Correct. The "cache" is used in our implementation to store values computed during forward propagation to be used in backward propagation.

78. **What is stored in the 'cache' during forward propagation for latter use in backward propagation?**

A-Z[l]
B-b[l]
C-W[l]
D-A[l]

A-Z[l]

This value is useful in the calculation of dW[l] in the backward propagation.

79. **Which of the following statements is true?**

A-The earlier layers of a neural network are typically computing more complex features of the input than the deeper layers.
B-The deeper layers of a neural network are typically computing more complex features of the input than the earlier layers.

B-The deeper layers of a neural network are typically computing more complex features of the input than the earlier layers.

80. **Which of the following are "parameters" of a neural network? (Check all that apply.)**

A-L the number of layers of the neural network.
B-g[l] the activation functions.
C-b[l] the bias vector.
D-W[l] the weight matrices.

C-b[l] the bias vector.
D-W[l] the weight matrices.

81. **Among the following, which ones are "hyperparameters"? (Check all that apply.)**

A-size of the hidden layers n[l]
B-number of layers L in the neural network
C-number of iterations
D-learning rate ±
E-weight matrices W[l]
F-activation values a[l]
B-bias vectors b[l]

A-size of the hidden layers n[l]
B-number of layers L in the neural network
C-number of iterations
D-learning rate ±

82. **During the backpropagation process, we use gradient descent to change the hyperparameters.**

True/False?

False.

During backpropagation, we use gradient descent to compute new values of W[l] and b[l]. These are the parameters of the network.

83. **During forward propagation, in the forward function for a layer l you need to know what is the activation function in a layer (sigmoid, tanh, ReLU, etc.). During backpropagation, the corresponding backward function also needs to know what is the activation function for layer l, since the gradient depends on it.**

True/False?

True.

each activation has a different derivative. Thus, during backpropagation you need to know which activation was used in the forward propagation to be able to compute the correct derivative.

84. **We can not use vectorization to calculate da[l] in backpropagation, we must use a for loop over all the examples.**

True/False?

False.

We can use vectorization in backpropagation to calculate dA[l] for each layer. This computation is done over all the training examples.

85. **Vectorization allows you to compute forward propagation in an 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.

Forward propagation propagates the input through the layers, although for shallow networks we may just write all the lines a[2]=g[2](z[2]), z[2]=W[2]a[1]+b[2], ...) in a deeper network, we cannot avoid a for loop iterating over the layers: (a[l]=g[l](z[l]), z[l]=W[l]a[l 1]+b[l], ...).

86. **Vectorization allows us to compute a[l] for all the examples on a batch at the same time without using a for loop.**

True/False?

True.

Vectorization allows us to compute the activation for all the training examples at the same time, avoiding the use of a for loop.

87. **Suppose W[i] is the array with the weights of the i-th layer, b[i] is the vector of biases of the i-th layer, and g is the activation function used in all layers. Which of the following calculates the forward propagation for the neural network with L layers.**

```
A-for i in range(L):
Z[i] = W[i]*X + b[i]
A[i] = g(Z[i])
B-for i in range(1, L+1):
Z[i] = W[i]*A[i-1] + b[i]
A[i] = g(Z[i])
```

```
B-for i in range(1, L+1):
Z[i] = W[i]*A[i-1] + b[i]
A[i] = g(Z[i])
```

Remember that the range omits the last number thus the range from 1 to L calculates only the A up to the L-1 layer.

C-for i in range(1, L):
Z[i] = W[i]*A[i-1] + b[i]
A[i] = g(Z[i])
D-for i in range(L):
Z[i+1] = W[i+1]*A[i+1] + b[i+1]
A[i+1] = g(Z[i+1])

88. **Assume we store the values for n[l] in an array called layer_dims, as follows: layer_dims = [n_x , 4, 3, 2, 1]. So layer 1 has four hidden units, layer 2 has 3 hidden units and so on. Which of the following for-loops will allow you to initialize the parameters for the model?**

A-for i in range(1, len(layer_dims)):
parameter['W' + str(i)] = np.random.randn(layer_dims[i-1], layer_dims[i]) * 0.01
parameter['b' + str(i)] = np.random.randn(layer_dims[i], 1) * 0.01
B-for i in range(1, len(layer_dims)/2):
parameter['W' + str(i)] = np.random.randn(layer_dims[i], layer_dims[i-1]) * 0.01
parameter['b' + str(i)] = np.random.randn(layer_dims[i], 1) * 0.01
C-for i in range(1, len(layer_dims)):
parameter['W' + str(i)] = np.random.randn(layer_dims[i], layer_dims[i-1]) * 0.01
parameter['b' + str(i)] = np.random.randn(layer_dims[i], 1) * 0.01
D-for i in range(1, len(layer_dims)/2):
parameter['W' + str(i)] = np.random.randn(layer_dims[i], layer_dims[i-1]) * 0.01
parameter['b' + str(i)] = np.random.randn(layer_dims[i-1], 1) * 0.01

C-for i in range(1, len(layer_dims)):
parameter['W' + str(i)] = np.random.randn(layer_dims[i], layer_dims[i-1]) * 0.01
parameter['b' + str(i)] = np.random.randn(layer_dims[i], 1) * 0.01

không có /2
dims[i] ... dim[i-1]

89. **During forward propagation, for the value of A[l] the value is used of Z[l] with the activation function g[l]. During backward propagation we calculate dA[l] from Z[l].**

False.

During backward propagation we are interested

	True/False?	in computing dW[l] and db[l]. For that we use g2L, dZ[l], Z[l], and W[l].
90.	**If L is the number of layers of a neural network then dZ[L]=A[L] Y.** **True/False?**	True Yes. The gradient of the output layer depends on the difference between the value computed during the forward propagation process and the target values. False No. The gradient of the output layer depends on the difference between the value computed during the forward propagation process and the target values.
91.	**A shallow neural network with a single hidden layer and 6 hidden units can compute any function that a neural network with 2 hidden layers and 6 hidden units can compute.** **True/False?**	False. As seen during the lectures there are functions you can compute with a "small" L-layer deep neural network that shallower networks require exponentially more hidden units to compute.

92.	**In the general case if we are training with m examples what is the shape of A[l]?**
	A-(m, n[l+1])
	B-(m, n[l])
	C-(n[l+1], m)
	D-(n[l], m)
		D-(n[l], m)
		The number of rows in A[1] corresponds to the number of units in the l-th layer.

93. **Whereas the previous question used a specific network, in the general case what is the dimension of W^{[l]}, the weight matrix associated with layer l?**

A-W[l] has shape (n[l],n[l+1])
B-W[l] has shape (n[l 1],n[l])
C-W[l] has shape (n[l+1],n[l])
D-W[l] has shape (n[l],n[l 1])

D-W[l] has shape (n[l],n[l 1])

94. **Whereas the previous question used a specific network, in the general case what is the dimension of b[l], the bias vector associated with layer l?**

A-b[l] has shape (n[l+1],1)
B-b[l] has shape (1,n[l])
C-b[l] has shape (1,n[l 1])
D-b[l] has shape (n[l],1)

D-b[l] has shape (n[l],1)

95. **For any mathematical function you can compute with an L-layered deep neural network with N hidden units there is a shallow neural network that requires only logN units, but it is very difficult to train.**

True
False

False

some mathematical functions can be computed using an L-layered neural network and a given number of hidden units; but using a shallow neural network the number of necessary hidden units grows exponentially.

96. **If you have 10,000 examples, how would you split the train/dev/test set? Choose the best option.**

A-33% train. 33% dev. 33% test.

B-60% train. 20% dev. 20% test.
This might be considered a small data set, not in the range of big data. Thus a more clas-

	B-60% train. 20% dev. 20% test. **C-98% train. 1% dev. 1% test.**	sical (old) best practice should be used.
97.	**If you have 20,000,000 examples, how would you split the train/dev/test set? Choose the best option.** **A-90% train. 5% dev. 5% test.** **B-99% train. 0.5% dev. 0.5% test.** **C-60% train. 20% dev. 20% test.**	B-99% train. 0.5% dev. 0.5% test. Given the size of the dataset, 0.5% of the samples are enough to get a good estimate of how well the model is doing.
98.	**If you have 10,000,000 examples, how would you split the train/dev/test set?** **A-33% train. 33% dev. 33% test** **B-60% train. 20% dev. 20% test** **C-98% train. 1% dev. 1% test**	C-98% train. 1% dev. 1% test
99.	**When designing a neural network to detect if a house cat is present in the picture, 500,000 pictures of cats were taken by their owners. These are used to make the training, dev and test sets. It is decided that to increase the size of the test set, 10,000 new images of cats taken from security cameras are going to be used in the test set. Which of the following is true?** **A-This will increase the bias of the model so the new images shouldn't be used.** **B-This will reduce the bias of the model and help improve it.** **C-This will be harmful to the project since now dev and test sets have different distributions.**	C-This will be harmful to the project since now dev and test sets have different distributions.
100.	**In a personal experiment, an M.L. student decides to not use a test set, only train-dev sets. In this case which of the following is true?**	C-He might be overfitting to the dev set. Although not recom-

	A-He won't be able to measure the bias of the model. B-Not having a test set is unacceptable under any circumstance. C-He might be overfitting to the dev set. D-He won't be able to measure the variance of the model.	mended, if a more accurate measure of the performance is not necessary it is ok to not use a test set. However, this might cause an overfit to the dev set.
101.	**The dev and test set should: A. Come from the same distribution B. Come from different distributions C. Be identical to each other (same (x,y) pairs) D. Have the same number of examples**	A. Come from same distributions
102.	**If your Neural Network model seems to have high bias, what of the following would be promising things to try? (Check all that apply.) A-Get more training data B-Make the Neural Network deeper C-Increase the number of units in each hidden layer D-Add regularization**	B-Make the Neural Network deeper C-Increase the number of units in each hidden layer
103.	**If your Neural Network model seems to have high variance, what of the following would be promising things to try? A-Increase the number of units in each hidden layer B-Get more training data C-Make the Neural Network deeper D-Get more test data E-Add regularization**	B-Get more training data E-Add regularization
104.	**Working on a model to classify bananas and oranges your classifier gets a training set error of 0.1% and a dev set error of 11%. Which of the following two are true?**	C-The model has a high variance. D-The model is overfitting the train set.

A-The model is overfitting the dev set.
B-The model has a very high bias.
C-The model has a high variance.
D-The model is overfitting the train set.

This model has a low bias and high variance.

105. **You are working on an automated check-out kiosk for a supermarket, and are building a classifier for apples, bananas and oranges. Suppose your classifier obtains a training set error of 0.5%, and a dev set error of 7%. Which of the following are promising things to try to improve your classifier? (Check all that apply.)**

A-Increase the regularization parameter lambda
B-Decrease the regularization parameter lambda
C-Get more training data
D-Use a bigger neural network

A-Increase the regularization parameter lambda
C-Get more training data

106. **You are working on an automated check-out kiosk for a supermarket and are building a classifier for apples, bananas, and oranges. Suppose your classifier obtains a training set error of 19% and a dev set error of 21%. Which of the following are promising things to try to improve your classifier? (Check all that apply, suppose the human error is approximately 0%)**

A-Get more training data.
B-Use a bigger network.
C-Increase the regularization parameter lambda.

B-Use a bigger network.

This can be helpful to reduce the bias of the model, and then we can start trying to reduce the high variance if this happens.

107. **What is weight decay?**

A-Gradual corruption of the weights in the neural network if it is trained on noisy data.

C-A regularization technique (such as L2 regularization) that results in gradient descent shrink-

	B-A technique to avoid vanishing gradient by imposing a ceiling on the values of the weights. C-A regularization technique (such as L2 regularization) that results in gradient descent shrinking the weights on every iteration. D-The process of gradually decreasing the learning rate during training.	ing the weights on every iteration.
108.	**Which of the following are regularization techniques? A-Weight decay. B-Increase the number of layers of the network. C-Dropout. D-Gradient Checking.**	A-Weight decay. C-Dropout.
109.	**To reduce high variance, the regularization hyperparameter lambda must be increased. True/False?**	True. By increasing the regularization parameter the magnitude of the weight parameters is reduced. This helps avoid overfitting and reduces the variance.
110.	**The regularization hyperparameter must be set to zero during testing to avoid getting random results. True/False?**	False The regularization parameter affects how the weights change during training, this means during backpropagation. It has no effect during the forward propagation that is when predictions for the test are made.
111.	**With the inverted dropout technique, at test time: A. You do not apply dropout (do not ran-**	A. You do not apply dropout (do not randomly eliminate units) and do not keep the

	domly eliminate units) and do not keep the 1/keep_prob factor in the calculations used in training **B. You do not apply dropout (do not randomly eliminate units), but keep the 1/keep_prob factor in the calculations used in training.** **C. You apply dropout (randomly eliminating units) and do not keep the 1/keep_prob factor in the calculations used in training** **D. You apply dropout (randomly eliminating units) but keep the 1/keep_prob factor in the calculations used in training.**	1/keep_prob factor in the calculations used in training
112.	**Which of the following are true about dropout?** **A-In practice, it eliminates units of each layer with a probability of keep_prob.** **B-It helps to reduce the bias of a model.** **C-In practice, it eliminates units of each layer with a probability of 1- keep_prob.** **D-It helps to reduce the variance of a model.**	C-In practice, it eliminates units of each layer with a probability of 1-keep_prob. D-It helps to reduce the variance of a model.
113.	**Increasing the parameter keep_prob from (say) 0.5 to 0.6 will likely cause the following: (Check the two that apply)** **A-Increasing the regularization effect** **B-Reducing the regularization effect** **C-Causing the neural network to end up with a higher training set error** **D-Causing the neural network to end up with a lower training set error**	B-Reducing the regularization effect D-Causing the neural network to end up with a lower training set error
114.	**Decreasing the parameter keep_prob from (say) 0.6 to 0.4 will likely cause the following:** **A-Causing the neural network to have a higher variance.**	C-Increasing the regularization effect. This will make the dropout have a higher probability of eliminating

B-Reducing the regularization effect.
C-Increasing the regularization effect.

a node in the neural network, increasing the regularization effect.

115. **Which of the following actions increase the regularization of a model? (Check all that apply)**

A-Decrease the value of the hyperparameter lambda.
B-Increase the value of keep_prob in dropout.
C-Decrease the value of keep_prob in dropout.
D-Increase the value of the hyperparameter lambda.
E-Use Xavier initialization.

C-Decrease the value of keep_prob in dropout.
D-Increase the value of the hyperparameter lambda.

116. **Which of the following actions increase the regularization of a model? (Check all that apply)**

A'-Make use of data augmentation.
B'-Normalizing the data.
C'-Increase the value of the hyperparameter lambda.
D'-Increase the value of keep_prob in dropout.
E'-Decrease the value of the hyperparameter lambda.

A'-Make use of data augmentation.

C'-Increase the value of the hyperparameter lambda.

Data augmentation has a way to generate "new" data at a relatively low cost. Thus making use of data augmentation can reduce the variance. When increasing the hyperparameter lambda we increase the effect of the L_2 penalization.

117. **Why do we normalize the inputs x?**

A-It makes the parameter initialization faster
B-It makes it easier to visualize the data
C-Normalization is another word for regulariza-

D-It makes the cost function faster to optimize

tion--It helps to reduce variance
D-It makes the cost function faster to optimize

---

118. **Suppose that a model uses, as one feature, the total number of kilometers walked by a person during a year, and another feature is the height of the person in meters. What is the most likely effect of normalization of the input data?**

A-It will make the data easier to visualize.
B-It will increase the variance of the model.
C-It won't have any positive or negative effects.
D-It will make the training faster.

D-It will make the training faster.

Since the difference between the ranges of the features is very different, this will likely cause the process of gradient descent to oscillate, making the optimization process longer.

---

119. **In every case it is a good practice to use dropout when training a deep neural network because it can help to prevent overfitting.**

True/False?

False

In most cases, it is recommended to not use dropout if there is no overfit. Although in computer vision, due to the nature of the data, it is the default practice.

---

120. **Which of these techniques are useful for reducing variance (reducing overfitting)? (Check all that apply.)**

A-Exploding gradient
B-Gradient Checking
C-Xavier initialization
D-Data augmentation
E-Dropout
F-Vanishing gradient
G-L2 regularization

D-Data augmentation
E-Dropout

G-L2 regularization

---

121. **During training a deep neural network that uses the tanh activation function, the value of the gradients is practically zero. Which of the fol-**

C-false
Maybe B?

lowing is most likely to help the vanishing gradient problem?

A-Increase the number of layers of the network.
B-Use Xavier initialization.
C-Use a larger regularization parameter.
D-Increase the number of cycles during the training.

122. **A model developed for a project is presenting high bias. One of the sponsors of the project offers some resources that might help reduce the bias. Which of the following additional resources has a better chance to help reduce the bias?**

A-Use different sources to gather data and better test the model.
B-Gather more data for the project.
C-Give access to more computational resources like GPUs.

C-Give access to more computational resources like GPUs.

Yes. This can allow the developers to try bigger networks, train for more cycles, and test different architectures.

123. **What happens when you increase the regularization hyperparameter lambda?**

A-Doubling lambda should roughly result in doubling the weights
B-Weights are pushed toward becoming bigger (further from 0)
C-Weights are pushed toward becoming smaller (closer to 0)
D-Gradient descent taking bigger steps with each iteration (proportional to lambda)

C-Weights are pushed toward becoming smaller (closer to 0)

124. **Which notation would you use to denote the 3rd layer's activations when the input is the 7th example from the 8th minibatch?**

A-a[3]{7}(8)

C-a[3]{8}(7)

B-a[8]{7}(3)
C-a[3]{8}(7)
D-a[8]{3}(7)

125. **Suppose you don't face any memory-related problems. Which of the following make more use of vectorization.**

A-Stochastic Gradient Descent
B-Batch Gradient Descent
C-Stochastic Gradient Descent, Batch Gradient Descent, and Mini-Batch Gradient Descent all make equal use of vectorization.
D-Mini-Batch Gradient Descent with mini-batch size m/2.

B-Batch Gradient Descent

If no memory problem is faced, batch gradient descent processes all of the training set in one pass, maximizing the use of vectorization.

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

A-Training one epoch (one pass through the training set) using mini-batch gradient descent is faster than training one epoch using batch gradient descent.
B-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).
C-When the mini-batch size is the same as the training size, mini-batch gradient descent is equivalent to batch gradient descent.

C-When the mini-batch size is the same as the training size, mini-batch gradient descent is equivalent to batch gradient descent.

127. **Which of the following is true about batch gradient descent?**

A-It is the same as the mini-batch gradient descent when the mini-batch size is the same as the size of the training set.
B-It has as many mini-batches as examples in

A-It is the same as the mini-batch gradient descent when the mini-batch size is the same as the size of the training set.

the training set.
C-It is the same as stochastic gradient descent, but we don't use random elements.

128. **We usually choose a mini-batch size greater than 1 and less than m, because that way we make use of vectorization but not fall into the slower case of batch gradient descent.**

True/False?

True.
Precisely by choosing a batch size greater than one we can use vectorization; but we choose a value less than m so we won't end up using batch gradient descent.

129. **Why is the best mini-batch size usually not 1 and not m, but instead something in-between? Check all that are true.**

A-If the mini-batch size is m, you end up with stochastic gradient descent, which is usually slower than mini-batch gradient descent.
B-If the mini-batch size is 1, you end up having to process the entire training set before making any progress.
C-If the mini-batch size is m, you end up with batch gradient descent, which has to process the whole training set before making progress.
D-If the mini-batch size is 1, you lose the benefits of vectorization across examples in the mini-batch.

C-If the mini-batch size is m, you end up with batch gradient descent, which has to process the whole training set before making progress.
D-If the mini-batch size is 1, you lose the benefits of vectorization across examples in the mini-batch.

130. **Suppose the temperature in Casablanca over the first two days of March are the following:**

March 1st: ¸1 =10C
March 2nd: ¸2 =25C

Say you use an exponentially weighted average with ²=0.5to track the temperature: v0 =0, vt =²vt 1 +(1²)¸t.

A-v2=15, v2^corrected=20

v2 =²vt 1 +(1²)¸ thus v1 =5, v2 =15. Using the bias correction vt/(1-²^t)   we get 15/(1 (0.5)^2) =20.

If v2 is the value computed after day 2 without bias correction, and v2^corrected is the value you compute with bias correction. What are these values?

A-v2=15, v2^corrected=20
B-v2=20, v2^corrected=20
C-v2=20, v2^corrected=15
D-v2=15, v2^corrected=15

131. **Suppose the temperature in Casablanca over the first two days of March are the following:**

March 1st: ¸1 =30C
March 2nd: ¸2 =15C

Say you use an exponentially weighted average with ²=0.5to track the temperature: v0 =0, vt ²=-vt 1 +(1²)¸t.
If v2 is the value computed after day 2 without bias correction, and v2^corrected is the value you compute with bias correction. What are these values?

A-v2=20, v2^corrected=20
B-v2=20, v2^corrected=15
C-v2=15, v2^corrected=15
D-v2=15, v2^corrected=20

D-v2=15, v2^corrected=20

v2 =²vt 1 +(1²)¸t thus v1 =15, v2 =15. Using the bias correction vt/(1-²^t) we get 15/(1 (0.5)^2) =20.

132. **Suppose the temperature in Casablanca over the first two days of March are the following:**

March 1st: ¸1 =10C
March 2nd: ¸2 =10C

Say you use an exponentially weighted average with ²=0.5to track the temperature: v0 =0, vt ²=-vt 1 +(1²)¸t.
If v2 is the value computed after day 2 without

D-v2=7.5, v2^corrected=10

v2 =²vt 1 +(1²)¸t thus v1 =5, v2 =7.5. Using the bias correction vt/(1-²^t) we get 7.5/(1 (0.5)^2) =10.

bias correction, and v2^corrected is the value you compute with bias correction. What are these values?

A-v2=10, v2^corrected=10
B-v2=10, v2^corrected=7.5
C-v2=7.5, v2^corrected=7.5
D-v2=7.5, v2^corrected=10

133. **You use an exponentially weighted average on the London temperature dataset. You use the following to track the temperature: vt =²vt 1 +(1²)¸t. The red line below was computed using ²=0.9. What would happen to your red curve as you vary ²? (Check the two that apply)**

A-Decreasing ²will shift the red line slightly to the right.
B-Increasing ²will shift the red line slightly to the right.
C-Decreasing ²will create more oscillation within the red line.
D-Increasing ²will create more oscillations within the red line.

B-Increasing ²will shift the red line slightly to the right.
C-Decreasing ²will create more oscillation within the red line.

134. **Which of the following is true about learning rate decay?**

A-The intuition behind it is that for later epochs our parameters are closer to a minimum thus it is more convenient to take larger steps to accelerate the convergence.
B-We use it to increase the size of the steps taken in each mini-batch iteration.
C-It helps to reduce the variance of a model.
D-The intuition behind it is that for later epochs our parameters are closer to a minimum thus it is more convenient to take smaller steps to prevent large oscillations.

D-The intuition behind it is that for later epochs our parameters are closer to a minimum thus it is more convenient to take smaller steps to prevent large oscillations.

135.	**Which of the following are true about gradient descent with momentum?** **A-Increasing the hyperparameter ²smooths out the process of gradient descent.** **B-It decreases the learning rate as the number of epochs increases.** **C-It generates faster learning by reducing the oscillation of the gradient descent process.** **D-Gradient descent with momentum makes use of moving averages.**	A-Increasing the hyperparameter ²smooths out the process of gradient descent. C-It generates faster learning by reducing the oscillation of the gradient descent process. D-Gradient descent with momentum makes use of moving averages.
136.	**** 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 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 J? (Check all that apply)** **A-Try mini-batch gradient descent** **B-Try better random initialization for the weights** **C-Try initializing all the weights to zero** **D-Try tuning the learning rate ±** **E-Try using Adam**	A-Try mini-batch gradient descent B-Try better random initialization for the weights D-Try tuning the learning rate ± E-Try using Adam
137.	**** 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 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 J? (Check all that apply)** **A-Try using gradient descent with momentum.** **B-Normalize the input data.** **C-Add more data to the training set.**	A-Try using gradient descent with momentum. B-Normalize the input data. D-Try better random initialization for the weights

	D-Try better random initialization for the weights	
138.	**** 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 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 J? (Check all that apply)** **A-Normalize the input data.** **B-Try initializing the weight at zero.** **C-Try using Adam.** **D-Try mini-batch gradient descent.**	A-Normalize the input data. C-Try using Adam. D-Try mini-batch gradient descent.
139.	**Which of the following are true about Adam?** **A-Adam combines the advantages of RMSProp and momentum.** **B-Adam automatically tunes the hyperparameter ±.** **C-The most important hyperparameter on Adam is õand should be carefully tuned.** **D-Adam can only be used with batch gradient descent and not with mini-batch gradient descent.**	A-Adam combines the advantages of RMSProp and momentum. Precisely Adam combines the features of RMSProp and momentum that is why we use two-parameter ²1and ²2 , besides õ.
140.	**Which of the following statements about Adam is False?** **A-We usually use "default" values for the hyperparameters ²1,²2and µin Adam (²1=0.9, ²2=0.999, µ=10 8)** **B-Adam combines the advantages of RMSProp and momentum** **C-Adam should be used with batch gradient computations, not with mini-batches.**	C-Adam should be used with batch gradient computations, not with mini-batches.

D-The learning rate hyperparameter ±in Adam usually needs to be tuned.

141. **In very high dimensional spaces it is most likely that the gradient descent process gives us a local minimum than a saddle point of the cost function.**

True/False?

False.
Due to the high number of dimensions it is much more likely to reach a saddle point, than a local minimum.

142. **If searching among a large number of hyperparameters, you should try values in a grid rather than random values, so that you can carry out the search more systematically and not rely on chance.**
True or False?

False

143. **Which of the following are true about hyperparameter search?**

A-Choosing values in a grid for the hyperparameters is better when the number of hyperparameters to tune is high since it provides a more ordered way to search.
B-Choosing random values for the hyperparameters is convenient since we might not know in advance which hyperparameters are more important for the problem at hand.
C-When sampling from a grid, the number of values for each hyperparameter is larger than when using random values.
D-When using random values for the hyperparameters they must be always uniformly distributed.

B-Choosing random values for the hyperparameters is convenient since we might not know in advance which hyperparameters are more important for the problem at hand.

Different problems might be more sensitive to different hyperparameters.

144. **With a relatively small set of hyperparameters, it is OK to use a grid search.**

True/False?

True
When the set of hyperparameters is small like a range for n_l =1,2,3 grid search works fine.

145.	**Every hyperparameter, if set poorly, can have a huge negative impact on training, and so all hyperparameters are about equally important to tune well.** **True or False?**	False. We've seen in the lecture that some hyperparameters, such as the learning rate, are more critical than others.
146.	**Once good values of hyperparameters have been found, those values should be changed if new data is added or a change in computational power occurs.** **True/False?**	True The choice of some hyperparameters such as the batch size depends on conditions such as hardware and quantity of data.
147.	**If it is only possible to tune two parameters from the following due to limited computational resources. Which two would you choose?** **A-±** **B-²1, ²2in Adam.** **C-The ²parameter of the momentum in gradient descent.** **D-õin Adam.**	A-± C-The ²parameter of the momentum in gradient descent.
148.	**In a project with limited computational resources, which three of the following hyperparameters would you choose to tune? Check all that apply.** **A-õin Adam.** **B-±** **C-mini-batch size** **D-The ²parameter of the momentum in gradient descent.** **E-²1, ²2in Adam.**	B-± C-mini-batch size D-The ²parameter of the momentum in gradient descent.
149.		

**Even if enough computational power is available for hyperparameter tuning, it is always better to babysit one model ("Panda" strategy), since this will result in a more custom model.
True/False?**

False.
Although it is possible to create good models using the "Panda" strategy, obtaining better results is more likely using a "caviar" strategy due to the number of tests and the nature of the deep learning process of ideas, code, and experiment.

150. **During hyperparameter search, whether you try to babysit one model ("Panda" strategy) or train a lot of models in parallel ("Caviar") is largely determined by:**

**A-The number of hyperparameters you have to tune
B-The presence of local minima (and saddle points) in your neural network
C-The amount of computational power you can access
D-Whether you use batch or mini-batch optimization**

C-The amount of computational power you can access

151. **Using the "Panda" strategy, it is possible to create several models.
True/False?**

True.
Following the "Panda" analogy, it is possible to babysit a model until a certain point and then start again to produce a different one.

152. **Knowing that the hyperparameter ± should be in the range of 0.001 and 1.0. Which of the following is the recommended way to sample a value for ±?**

D-r = -3*np.random.rand()alpha = 10**r

This gives a ran-

A-r = -5*np.random.rand()alpha = 10r**
B-r = 4*np.random.rand()alpha = 10r**
C-r = np.random.rand()alpha = 0.001 + r*0.999
D-r = -3*np.random.rand()alpha = 10r**

dom number between 0.001=10^ 3 and 10^0.

153. **If you think ² (hyperparameter for momentum) is between 0.9 and 0.99, which of the following is the recommended way to sample a value for beta?**

A-r = np.random.rand()
beta = r*0.9 + 0.09
B-r = np.random.rand()
beta = 1-10(- r + 1)**
C-r = np.random.rand()
beta = r*0.09 + 0.9
D-r = np.random.rand()
beta = 1-10(- r - 1)**

D-r = np.random.rand()
beta = 1-10**(- r - 1)

154. **Finding good hyperparameter values is very time-consuming. So typically you should do it once at the start of the project, and try to find very good hyperparameters so that you don't ever have to tune them again.**

True or false?

False

155. **Finding new values for the hyperparameters, once we have found good ones for a model, should only be done if new hardware or computational power is acquired.**

True/False?

False

Correct. As the data changes for the model, it might be beneficial to tune some of the hyperparameters again.

156. **In batch normalization as presented in the videos, if you apply it on the llth layer of your neural network, what are you normalizing?**

maybe C

A. b[l]
B. a[l]
C. z[l]
D. W[l]

157. **When using batch normalization it is OK to drop the parameter b[l] from the forward propagation since it will be subtracted out when we compute z~[l]=³z_normalize[l] +²[].**
True/False?

True.
Since in the normalization process the values of z[l] are re-centered at the origin, it is irrelevant to add the b[l] parameter.

158. **When using batch normalization it is OK to drop the parameter W[l] from the forward propagation since it will be subtracted out when we compute z~[l]=³z_normalize[l] +²[].**
True/False?

False.
The parameter W[l] doesn't get subtracted during the batch normalization process, although it gets re-scaled.

159. **Which of the following are true about batch normalization?**

A-The parameters ²and ³of batch normalization can't be trained using Adam or RMS prop.
B-The parameter õn the batch normalization formula is used to accelerate the convergence of the model.
C-There is a global value of ³and ²that is used for all the hidden layers where batch normalization is used.
D-One intuition behind why batch normalization works is that it helps reduce the internal covariance.

D-One intuition behind why batch normalization works is that it helps reduce the internal covariance.

160. **Which of the following is true about batch normalization?**

A-The parameters ³[] and ²[] set the variance and mean of ˜z[l].
B-z(i)norm=[z(i)¼]/ (Ã^2).

A-The parameters ³[] and ²[] set the variance and mean of ˜z[l].

Correct. When applying the linear transforma-

C-The parameters ³[] and ²[] can be learned only using plain gradient descent.
D-The optimal values to use for ³and ²are ³= (Ã^2+õ) and ²=¼.

tion ~z(l)=²[].znorm(l) +³[] we set the variance and mean of ~z[l].

161. **When using normalization: z_norm(i) =(z(i) -¼)/(Ã^2+õ)**
In case Ã is too small, the normalization of z(i) may fail since division by 0 may be produced due to rounding errors.

True/False?

False.
The normalization formula uses a smoothing parameter õ so in z_norm(i) =(z(i)-¼)/(Ã^2+õ use of the õ parameter prevents that the denominator be 0.

162. **In the normalization formula z_norm(i) =(z(i) -¼)/(Ã^2+õ) , why do we use epsilon?**

A-To speed up convergence
B-In case ¼ is too small
C-To have a more accurate normalization
D-To avoid division by zero

D-To avoid division by zero

163. **A neural network is trained with Batch Norm. At test time, to evaluate the neural network on a new example you should perform the normalization using ¼ and Ã^2 estimated using an exponentially weighted average across mini-batches seen during training.**

True/false?

True

This is a good practice to estimate the ¼ and Ã^2 to use since at test time we might not be predicting over a batch of the same size, or it might even be a single example, thus using the ¼ and Ã^2 of a single sample doesn't make sense.

164. **Which of the following are true about batch normalization?**

A-²[] and ³[] are hyperparameters that must be tuned by random sampling in a logarithmic

C-When using batch normalization we introduce two new parameters ³[], ²[] that must be "learned" or trained.

	scale. **B-z(i)_norm=z(i)¼ Ã2.** **C-When using batch normalization we introduce two new parameters ³[], ²[] that must be "learned" or trained.** **D-The parameters ³[] and ²[] set the variance and mean of ˜z[l].**	D-The parameters ³[] and ²[] set the variance and mean of ˜z[l]. Batch normalization uses two parameters ² and ³to compute ˜z(i)=²-z(i)norm+³. When applying the linear transformation ˜z(l)=²[-l]z(l)norm+³[] we set the variance and mean of ˜z[l].
165.	**Which of the following statements about ³and ² in Batch Norm are true?** **A-²and ³are hyperparameters of the algorithm, which we tune via random sampling.** **B-They can be learned using Adam, Gradient descent with momentum, or RMSprop, not just with gradient descent.** **C-The optimal values are ³= Ã2and µ, ²=¼.** **D-There is one global value of ³R and one global value of ²R for each layer, and these apply to all the hidden units in that layer.** **E-They set the variance and mean of the linear variable ˜z[l] of a given layer.**	B-They can be learned using Adam, Gradient descent with momentum, or RMSprop, not just with gradient descent. E-They set the variance and mean of the linear variable ˜z[l] of a given layer.
166.	**After training a neural network with Batch Norm, at test time, to evaluate the neural network on a new example you should:** **A-Skip the step where you normalize using ¼ and Ã2since a single test example cannot be normalized.** **B-Use the most recent mini-batch's value of ¼ and Ã2to perform the needed normalizations.** **C-If you implemented Batch Norm on**	D-Perform the needed normalizations, use ¼and Ã2estimated using an exponentially weighted average across mini-batches seen during training.

mini-batches of (say) 256 examples, then to evaluate on one test example, duplicate that example 256 times so that you're working with a mini-batch the same size as during training.
D-Perform the needed normalizations, use ¼and Ã2estimated using an exponentially weighted average across mini-batches seen during training.

167. **A neural network is trained with Batch Norm. At test time, to evaluate the neural network we turn off the Batch Norm to avoid random predictions from the network.**

True/False?

False

During the test, the parameters ¼and Ã^2are estimated using an exponentially weighted average across mini-batches used during training.

168. **Which of the following are some recommended criteria to choose a deep learning framework?**

A-It must run exclusively on cloud services, to ensure its robustness.
B-It must be implemented in C to be faster.
C-Running speed.
D-It must use Python as the primary language.

C-Running speed.

The running speed is a major factor, especially when working with large datasets.

169. **Which of these statements about deep learning programming frameworks are true? (Check all that apply)**

A-Deep learning programming frameworks require cloud-based machines to run.
B-A programming framework allows you to code up deep learning algorithms with typically fewer lines of code than a lower-level language such as Python.
C-Even if a project is currently open source, good governance of the project helps ensure

B-A programming framework allows you to code up deep learning algorithms with typically fewer lines of code than a lower-level language such as Python.
C-Even if a project is currently open source, good governance of the project helps ensure that it remains open even

that it remains open even in the long term, rather than become closed or modified to benefit only one company.

in the long term, rather than become closed or modified to benefit only one company.

170. **If a project is open-source, it is a guarantee that it will remain open source in the long run and will never be modified to benefit only one company.**

True/False?

False

To ensure that a project will remain open source in the long run it must have a good governance body too.

171. **Having three evaluation metrics makes it harder for you to quickly choose between two different algorithms, and will slow down the speed with which your team can iterate.**

True/False?

True

172. **You meet with them and ask for just one evaluation metric.**

True/False?

True
More than one metric expands the choices and tradeoffs you have to decide for each with unknown effects on the other two.

173. **You are delighted because this list of criteria will speed development and provide guidance on how to evaluate two different algorithms.**

True/False?

False
The goal is to have one metric that focuses the development effort and increases iteration velocity.

More than one metric expands the choices and tradeoffs you have to decide for each with unknown effects on the other two.

174. **The city asks for your help in further defining the criteria for accuracy, runtime, and memory. How would you suggest they identify the criteria?**

A-Suggest to them that they focus on whichever criterion is important and then eliminate the other two.
B-Suggest to them that they define which criterion is most important. Then, set thresholds for the other two.
C-Suggest that they purchase more infrastructure to ensure the model runs quickly and accurately.

B-Suggest to them that they define which criterion is most important. Then, set thresholds for the other two.

The thresholds provide a way to evaluate models head to head.

175. **The city revises its criteria to:**
"We need an algorithm that can let us know a bird is flying over Peacetopia as accurately as possible."
"We want the trained model to take no more than 10 sec to classify a new image."
"We want the model to fit in 10MB of memory."
Given models with different accuracies, runtimes, and memory sizes, how would you choose one?

A-Find the subset of models that meet the runtime and memory criteria. Then, choose the highest accuracy.
B-Take the model with the smallest runtime because that will provide the most overhead to increase accuracy.
C-Accuracy is an optimizing metric, therefore the most accurate model is the best choice.
D-Create one metric by combining the three metrics and choose the best performing model.

A-Find the subset of models that meet the runtime and memory criteria. Then, choose the highest accuracy.

Once you meet the runtime and memory thresholds, accuracy should be maximized.

176. **The essential difference between an optimizing metric and satisficing metrics is the priority assigned by the stakeholders. True/False?**

False.
Stakeholders must define thresholds for satisficing metrics, leaving the optimizing metric unbounded.

177. **Based on the city's requests, which of the following would you say is true?**

A-Accuracy is a satisfying metric; running time and memory size are an optimizing metric.
B-Accuracy, running time and memory size are all optimizing metrics because you want to do well on all three.
C-Accuracy is an optimizing metric; running time and memory size are satisfying metrics.
D-Accuracy, running time and memory size are all satisfying metrics because you have to do sufficiently well on all three for your system to be acceptable.

C-Accuracy is an optimizing metric; running time and memory size are satisfying metrics.

178. **Which of the following best answers why it is important to identify optimizing and satisficing metrics?**

A-It isn't. All metrics must be met for the model to be acceptable.
B-Knowing the metrics provides input for efficient project planning.
C-Identifying the metric types sets thresholds for satisficing metrics. This provides explicit evaluation criteria.
D-Identifying the optimizing metric informs the team which models they should try first.

C-Identifying the metric types sets thresholds for satisficing metrics. This provides explicit evaluation criteria.

Thresholds are essential for evaluation of key use case constraints.

179. **You propose a 95/2.5%/2.5% for train/dev/test splits to the City Council. They ask for your reasoning. Which of the following best justifies**

B-With a dataset comprising 10M individual samples, 2.5% repre-

your proposal?

A-The emphasis on the training set will allow us to iterate faster.
B-With a dataset comprising 10M individual samples, 2.5% represents 250k samples, which should be more than enough for dev and testing to evaluate bias and variance.
C-The emphasis on the training set provides the most accurate model, supporting the memory and processing satisficing metrics.
D-The most important goal is achieving the highest accuracy, and that can be done by allocating the maximum amount of data to the training set.

sents 250k samples, which should be more than enough for dev and testing to evaluate bias and variance.

The purpose of dev and test sets is fulfilled even with smaller percentages of the data.

180. **Now that you've set up your train/dev/test sets, the City Council comes across another 1,000,000 images from social media and offers them to you. These images are different from the distribution of images the City Council had originally given you, but you think it could help your algorithm. You should add the citizens' data to the training set.**

True/False?

True

Adding this data to the training set will change the training set distribution. However, it is not a problem to have different training and dev distributions. In contrast, it would be very problematic to have different dev and test set distributions.

181. **Now that you've set up your train/dev/test sets, the City Council comes across another 1,000,000 images from social media and offers them to you. These images are different from the distribution of images the City Council had originally given you, but you think it could help your algorithm. Which of the following is the best use of that additional data?**
A-Do not use the data. It will change the distribution of any set it is added to.

D-Add it to the training set.

B-Split it among train/dev/test equally.
C-Add it to the dev set to evaluate how well the model generalizes across a broader set.
D-Add it to the training set.

182. **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 dev set. You object because: (Choose all that apply)**

A-A bigger test set will slow down the speed of iterating because of the computational expense of evaluating models on the test set.
B-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.
C-The dev set no longer reflects the distribution of data (security cameras) you most care about.
D-The 1,000,000 citizens' data images do not have a consistent x-->y mapping as the rest of the data.

B-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.
C-The dev set no longer reflects the distribution of data (security cameras) you most care about.

183. **One member of the City Council knows a little about machine learning and thinks you should add the 1,000,000 citizens' data images proportionately to the train/dev/test sets. You object because:**

A-If we add the images to the test set then it won't reflect the distribution of data expected in production.
B-The training set will not be as accurate because of the different distributions.
C-The additional data would significantly slow down training time.
D-The 1,000,000 citizens' data images do not

A-If we add the images to the test set then it won't reflect the distribution of data expected in production.

Yes. Using the data in the training set could be beneficial, but you wouldn't want to include such images in your test set as they are not from the expected distribution of data you'll see in production.

have a consistent x-->y mapping as the rest of the data.

184. **You train a system, and the train/dev set errors are 3.5% and 4.0% respectively. You decide to try regularization to close the train/dev accuracy gap. Do you agree?**

A-No, because you do not know what the human performance level is.
B-Yes, because having a 4.0% training error shows you have a high bias.
C-Yes, because this shows your bias is higher than your variance.
D-No, because this shows your variance is higher than your bias.

A-No, because you do not know what the human performance level is.

Yes. You need to know what the human performance level is to estimate avoidable bias.

185. **You train a system, and its errors are as follows (error = 100%-Accuracy):**
Training set error 4.0%
Dev set error 4.5%
This suggests that one good avenue for improving performance is to train a bigger network so as to drive down the 4.0% training error. Do you agree?

A-No, because this shows your variance is higher than your bias.
B-No, because there is insufficient information to tell.
C-Yes, because having a 4.0% training error shows you have a high bias.
D-Yes, because this shows your bias is higher than your variance.

B-No, because there is insufficient information to tell.

no information about human performance

186. **Human performance for identifying birds is < 1%, training set error is 5.2% and dev set error is 7.3%. Which of the options below is the best next step?**

A-Train a bigger network to drive down the >4.0% training error.

A-Train a bigger network to drive down the >4.0% training error.
B-Try an ensemble model to reduce bias and variance.
C-Get more data or apply regularization to reduce variance.
D-Validate the human data set with a sample of your data to ensure the images are of sufficient quality.

Avoidable bias is >4.2% which is larger than the 2.1% variance.

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

A-The performance of the head of the City Council.
B-The best performance of a specialist (ornithologist) or possibly a group of specialists.
C-The performance of their volunteer amateur ornithologists.
D-The performance of the average citizen of Peacetopia.

B-The best performance of a specialist (ornithologist) or possibly a group of specialists.

This is the peak of human performance in this task.

188. **You want to define what human-level performance is to the city council. Which of the following is the best answer?**

A-The average performance of all their ornithologists (0.5%).
B-The average of regular citizens of Peacetopia (1.2%).
C-The performance of their best ornithologist (0.3%).
D-The average of all the numbers above (0.66%).

C-The performance of their best ornithologist (0.3%).

The best human performance is closest to Bayes' error.

189.

Which of the below shows the optimal order of accuracy from worst to best?

A-Human-level performance -> Bayes error -> the learning algorithm's performance.
B-The learning algorithm's performance -> human-level performance -> Bayes error.
C-Human-level performance -> the learning algorithm's performance -> Bayes error.
D-The learning algorithm's performance -> Bayes error -> human-level performance.

C-Human-level performance -> the learning algorithm's performance -> Bayes error.

A learning algorithm's performance can be better than human-level performance but it can never be better than Bayes error.

190. **A learning algorithm's performance can be better than human-level performance but it can never be better than Bayes error.**
True/False?

True.
By definition, human level error is worse than Bayes error.

191. **Which of the following statements do you agree with?**

A-A learning algorithm's performance can be better than human-level performance and better than Bayes error.
B-A learning algorithm's performance can be better than human-level performance but it can never be better than Bayes error.
C-A learning algorithm's performance can never be better than human-level performance but it can be better than Bayes error.
D-A learning algorithm's performance can never be better than human-level performance nor better than Bayes error.

B-A learning algorithm's performance can be better than human-level performance but it can never be better than Bayes error.

192. **After working on your algorithm you have to decide the next steps. Currently, human-level performance is 0.1%, training is at 2.0% and the dev set is at 2.1%. Which statement below best describes your thought process?**

A-Address bias first through a larger model to get closest to human level error.

D-Decrease regulariza-

A-Address bias first through a larger model to get closest to human level error.
B-Decrease variance via regularization so training and dev sets have similar performance.
C-Get a bigger training set to reduce variance.
D-Decrease regularization to boost smaller signals.

tion to boost smaller signals.

193. **You've now also run your model on the test set and find that it is a 7.0% error compared to a 2.1% error for the dev set. What should you do? (Choose all that apply)**

A-Try decreasing regularization for better generalization with the dev set.
B-Try increasing regularization to reduce overfitting to the dev set.
C-Increase the size of the dev set.
D-Get a bigger test set to increase its accuracy.

B-Try increasing regularization to reduce overfitting to the dev set.
C-Increase the size of the dev set.

194. **Which of the following best expresses how to evaluate the next steps in your project when your results for human-level performance, train, and dev set error are 0.1%, 2.0%, and 2.1% respectively?**

A-Keep tuning until the train set accuracy is equal to human-level performance because it is the optimizing metric.
B-Port the code to the target devices to evaluate if your model meets or exceeds the satisficing metrics.
C-Evaluate the test set to determine the magnitude of the variance.
D-Based on differences between the three levels of performance, prioritize actions to decrease bias and iterate.

D-Based on differences between the three levels of performance, prioritize actions to decrease bias and iterate.

195.

After working on this project for a year, you finally achieve:
Human-level performance 0.10%
Training set error 0.05%
Dev set error 0.05%
What can you conclude? (Check all that apply.)

A'-You are close to Bayes error and possible overfitting.
B'-With only 0.05% further progress to make, you should quickly be able to close the remaining gap to 0%
C'-This is a statistical anomaly (or must be the result of statistical noise) since it should not be possible to surpass human-level performance.
D'-All or almost all of the avoidable bias has been accounted for.

A'-Yes. By definition, Bayes error cannot be exceeded except for overfitting.
D'-Yes. Exceeding human performance makes the identification of avoidable bias very challenging.

196. **After working on this project for a year, you finally achieve:**
Human-level performance 0.10%
Training set error 0.05%
Dev set error 0.05%
What can you conclude? (Check all that apply.)

A-It is now harder to measure avoidable bias, thus progress will be slower going forward.
B-With only 0.05% further progress to make, you should quickly be able to close the remaining gap to 0%
C-This is a statistical anomaly (or must be the result of statistical noise) since it should not be possible to surpass human-level performance.
D-If the test set is big enough for the 0.05% error estimate to be accurate, this implies Bayes error is d0.05d0.05

A-It is now harder to measure avoidable bias, thus progress will be slower going forward.
D-If the test set is big enough for the 0.05% error estimate to be accurate, this implies Bayes error is d0.05d0.05

197. **After running your model with the test set you find it is a 7.0% error compared to a 2.1% error**

B-You have overfitted to the dev set.

	for the dev set and 2.0% for the training set. What can you conclude? (Choose all that apply) **A-You have underfitted to the dev set.** **B-You have overfitted to the dev set.** **C-You should try to get a bigger dev set.** **D-Try decreasing regularization for better generalization with the dev set.**	C-You should try to get a bigger dev set.
198.	**Your system is now very accurate but has a higher false negative rate than the City Council of Peacetopia would like. What is your best next step?** **A-Expand your model size to account for more corner cases.** **B-Reset your "target" (metric) for the team and tune to it.** **C-Pick false negative rate as the new metric, and use this new metric to drive all further development.** **D-Look at all the models you've developed during the development process and find the one with the lowest false negative error rate.**	B-Reset your "target" (metric) for the team and tune to it. The target has shifted so an updated metric is required.
199.	**It turns out Peacetopia has hired one of your competitors to build a system as well. Your system and your competitor both deliver systems with about the same running time and memory size. However, your system has higher accuracy! However, when Peacetopia tries out your and your competitor's systems, they conclude they actually like your competitor's system better, because even though you have higher overall accuracy, you have more false negatives (failing to raise an alarm when a bird is in the air). What should you do?** **A-Rethink the appropriate metric for this task,**	A-Rethink the appropriate metric for this task, and ask your team to tune to the new metric.

**and ask your team to tune to the new metric.
B-Pick false negative rate as the new metric, and use this new metric to drive all further development.
C-Ask your team to take into account both accuracy and false negative rate during development.
D-Look at all the models you've developed during the development process and find the one with the lowest false negative error rate.**

---

200. **It turns out Peacetopia has hired one of your competitors to build a system as well. Your system and your competitor both deliver systems with about the same running time and memory size. However, your system has higher accuracy! However, when Peacetopia tries out your and your competitor's systems, they conclude they actually like your competitor's system better, because even though you have higher overall accuracy, you have more false negatives (failing to raise an alarm when a bird is in the air). What should you do?**

**A'-Apply regularization to minimize the false negative rate.
B'-Pick false negative rate as the new metric, and use this new metric to drive all further development.
C'-Brainstorm with your team to refine the optimizing metric to include false negatives as they further develop the model.
D'-Ask your team to take into account both accuracy and false negative rate during development.**

C'-Brainstorm with your team to refine the optimizing metric to include false negatives as they further develop the model.

---

201. **You've handily beaten your competitor, and your system is now deployed in Peacetopia and is protecting the citizens from birds! But**

A-Augment your data to increase the images of the new bird.

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 model is being tested on a new type of data.

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

A-Augment your data to increase the images of the new bird.
B-Add the new images and split them among train/dev/test.
C-Add hidden layers to further refine feature development.
D-Put them into the dev set to evaluate the bias and re-tune.

202. **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 model is being tested on a new type of data.**

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

A'-Try data augmentation/data synthesis to get more images of the new type of bird.
B'-Add the 1,000 images into your dataset and reshuffle into a new train/dev/test split.
C'-Put the 1,000 images into the training set so

D'-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.

as to try to do better on these birds.
D'-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.

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

A-Add pooling layers to downsample features to accommodate the new species.
B-Split them between dev and test and re-tune.
C-Augment your data to increase the images of the new bird.
D-Put the new species' images in training data to learn their features.

C-true-Augment your data to increase the images of the new bird.

A sufficient number of images is necessary to account for the new species.

204. **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. You have a huge dataset of 100,000,000 cat images. Training on this data takes about two weeks. Which of the statements do you agree with? (Check all that agree.)**

A-This significantly impacts iteration speed.
B-Reducing the model complexity will allow the use of the larger data set but preserve accuracy.
C-Lowering the number of images will reduce training time and likely allow for an acceptable tradeoff between iteration speed and accuracy.

A-This significantly impacts iteration speed.

C-Lowering the number of images will reduce training time and likely allow for an acceptable tradeoff between iteration speed and accuracy.

205. **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. You have a huge dataset of 100,000,000 cat images. Training on this data takes about two weeks. Which of the statements do you agree with? (Check all that agree.)**

A-Given a significant budget for cloud GPUs, you could mitigate the training time.
B-With the experience gained from the Bird detector you are confident to build a good Cat detector on the first try.
C-You could consider a tradeoff where you use a subset of the cat data to find reasonable performance with reasonable iteration pacing.
D-Accuracy should exceed the City Council's requirements but the project may take as long as the bird detector because of the two week training/iteration time.

A-Given a significant budget for cloud GPUs, you could mitigate the training time.

C-You could consider a tradeoff where you use a subset of the cat data to find reasonable performance with reasonable iteration pacing.
D-Accuracy should exceed the City Council's requirements but the project may take as long as the bird detector because of the two week training/iteration time.

206. **You are getting started with this project. What is the first thing you do? Assume each of the steps below would take about an equal amount of time (a few days).**

A-Spend a few days collecting more data using the front-facing camera of your car, to better understand how much data per unit time you can collect.
B-Spend some time searching the internet for the data most similar to the conditions you expect on production.
C-Train a basic model and do error analysis.
D-Invest a few days in thinking on potential difficulties, and then some more days brainstorming about possible solutions, before training any model.

C-Train a basic model and do error analysis.

As discussed in lecture, applied ML is a highly iterative process. If you train a basic model and carry out error analysis (see what mistakes it makes) it will help point you in more promising directions.

207. **Your goal is to detect road signs (stop sign, pedestrian crossing sign, construction ahead sign) and traffic signals (red and green lights) in images. The goal is to recognize which of these objects appear in each image. You plan to use a deep neural network with ReLU units in the hidden layers. For the output layer, a softmax activation would be a good choice for the output layer because this is a multi-task learning problem.**

True/False?

False

a softmax activation is not suitable for multi-task learning problem.

Softmax would be a good choice if one and only one of the possibilities (stop sign, speed bump, pedestrian crossing, green light and red light) was present in each image.

208. **Your goal is to detect road signs (stop sign, pedestrian crossing sign, construction ahead sign) and traffic signals (red and green lights) in images. The goal is to recognize which of these objects appear in each image. You plan to use a deep neural network with ReLU units in the hidden layers. For the output layer, which of the following gives you the most appropriate activation function?**

A-Linear
B-ReLU
C-Softmax
D-Sigmoid

D-Sigmoid

Correct. This works well since the output would be valued between 0 and 1 which represents the probability that one of the possibilities is present in an image.

209. **You are carrying out error analysis and counting up what errors the algorithm makes. Which of these datasets do you think you should manually go through and carefully examine, one image at a time?**

A-10,000 images on which the algorithm made a mistake
B-10,000 randomly chosen images

D-500 images on which the algorithm made a mistake

Focus on images that the algorithm got wrong. Also, 500 is enough to give you a good initial sense of the error statis-

	C-500 randomly chosen images **D-500 images on which the algorithm made a mistake**	tics. There's probably no need to look at 10,000, which will take a long time.
210.	**You are working out error analysis and counting up what errors the algorithm makes. Which of the following do you think you should manually go through and carefully examine, one image at a time?** **A-500 images of the training-dev set, on which the algorithm made a mistake.** **B-500 images of the train set, on which the algorithm made a mistake.** **C-500 images of the dev set, on which the algorithm made a mistake.** **D-500 images of the test set, on which the algorithm made a mistake.**	C-500 images of the dev set, on which the algorithm made a mistake.
211.	**When trying to determine what strategy to implement to improve the performance of a model, we manually check all images of the training set where the algorithm was successful.** **True/False?**	False This set should be too large to manually check all the images. It is better to focus on the images that the algorithm got wrong from the dev set. Also, choose a large enough subset that we can manually check.
212.	**After working on the data for several weeks, your team ends up with the following data:** **100,000 labeled images taken using the front-facing camera of your car.** **900,000 labeled images of roads downloaded from the internet.** **Each image's labels precisely indicate the presence of any specific road signs and traffic sig-**	1) False Multi-task learning can still be effective even if some images are labeled only for a subset of the tasks. The loss function is adjusted to exclude the unla-

	nals or combinations of them. For example, y(i) = [10010]^T means the image contains a stop sign and a red traffic light. **1) Because this is a multi-task learning problem, you need to have all your y(i) vectors fully labeled. If one example is equal to [0?11?]^T then the learning algorithm will not be able to use that example. True/False?**	beled tasks when calculating the loss
213.	**After working on the data for several weeks, your team ends up with the following data:** **100,000 labeled images taken using the front-facing camera of your car.** **900,000 labeled images of roads downloaded from the internet.** **Each image's labels precisely indicate the presence of any specific road signs and traffic signals or combinations of them. For example, y(i) = [10010]^T means the image contains a stop sign and a red traffic light.** **2) we can use it if we ignore those entries when calculating the loss function. True/False?**	2) True We can't use the components of the labels that are missing but we can use the ones we have to train the model.
214.	**The distribution of data you care about contains images from your car's front-facing camera, which comes from a different distribution than the images you were able to find and download off the internet. The best way to split the data is using the 900,000 internet images to train, and divide the 100,000 images from your car's front-facing camera between dev and test sets.** **True/False?**	False. 100,000 images are too many to use in dev and test. A better distribution would be to use 80,000 of those images to train, and split the rest between dev and test. As seen in the lecture, it is important that your dev and test set have the closest possible distribution to "real" data.

		It is also important for the training set to contain enough "real" data to avoid having a data-mismatch problem.
215.	**The distribution of data you care about contains images from your car's front-facing camera, which comes from a different distribution than the images you were able to find and download off the internet. Which of the following are true about the train/dev/test split?** **A-The dev and test sets must come from the same distribution.** **B-The train, dev, and test must come from the same distribution.** **C-The dev and test sets must contain some images from the internet.** **D-The dev and test set must come from the front-facing camera.**	A-The dev and test sets must come from the same distribution. D-The dev and test set must come from the front-facing camera. D-Correct Correct. This is the distribution we care about most, thus we should use this as a target.
216.	**The distribution of data you care about contains images from your car's front-facing camera; which comes from a different distribution than the images you were able to find and download off the internet. How should you split the dataset into train/dev/test sets?** **A-Mix all the 100,000 images with the 900,000 images you found online. Shuffle everything. Split the 1,000,000 images dataset into 600,000 for the training set, 200,000 for the dev set and 200,000 for the test set.** **B-Mix all the 100,000 images with the 900,000 images you found online. Shuffle everything. Split the 1,000,000 images dataset into 980,000 for the training set, 10,000 for the dev set and 10,000 for the test set.**	D-Choose the training set to be the 900,000 images from the internet along with 80,000 images from your car's front-facing camera. The 20,000 remaining images will be split equally in dev and test sets. Yes. As seen in the lecture, it is important that your dev and test set have the closest possible distribution to "real" data. It is also important for the training set

	C-Choose the training set to be the 900,000 images from the internet along with 20,000 images from your car's front-facing camera. The 80,000 remaining images will be split equally in dev and test sets. **D-Choose the training set to be the 900,000 images from the internet along with 80,000 images from your car's front-facing camera. The 20,000 remaining images will be split equally in dev and test sets.**	to contain enough "real" data to avoid having a data-mismatch problem.
217.	**Assume you've finally chosen the following split between the data:** **Error of the algorithm:** **Training: 1%** **Training-Dev: 5.1%** **Dev: 5.6%** **Test: 6.8%** **You also know that human-level error on the road sign and traffic signals classification task is around 0.5%. Which of the following is true?** **A-You have a high variance problem.** **B-You have a high bias.** **C-You have a large data-mismatch problem.** **D-The size of the train-dev set is too high.**	A-You have a high variance problem.
218.	**Assume you've finally chosen the following split between the data:** **Error of the algorithm:** **Training: 12%** **Training-Dev: 15.1%** **Dev: 12.6%** **Test: 15.8%** **You also know that human-level error on the road sign and traffic signals classification task is around 0.5%. Which of the following is true?** **A-You have a high variance problem.** **B-You have a large data-mismatch problem.**	C-You have a high bias. The avoidable bias is significantly high since the training error is a lot higher than the human-level error.

C-You have a high bias.
D-You have a too low avoidable bias.

219. **Assume you've finally chosen the following split between the data:**
Error of the algorithm:
Training: 2%
Training-Dev: 2.3%
Dev: 1.3%
Test: 1.1%

You also know that human-level error on the road sign and traffic signals classification task is around 0.5%. Based on the information given, a friend thinks that the training data distribution is much harder than the dev/test distribution. What do you think?

A-Your friend is wrong. (i.e., Bayes error for the dev/test distribution is probably higher than for the train distribution.)
B-Your friend is probably right. (i.e., Bayes error for the dev/test distribution is probably lower than for the train distribution.)
C-There's insufficient information to tell if your friend is right or wrong.

B-Your friend is probably right. (i.e., Bayes error for the dev/test distribution is probably lower than for the train distribution.)

Since the training-dev error is higher than the dev and test errors, the dev/test distribution is probably "easier" than the training distribution.

220. **You decide to focus on the dev set and check by hand what are the errors due to. Here is a table summarizing your discoveries:**
Overall dev set error: 15.3%
Errors due to incorrectly labeled data: 4.1%
Errors due to foggy pictures: 8.0%
Errors due to rain drops stuck on your car's front-facing camera: 2.2%
Errors due to other causes: 1.0%
In this table, 4.1%, 8.0%, etc. are a fraction of the total dev set (not just examples of your algorithm mislabeled). For example, about 8.0/15.3

C-False because it depends on how easy it is to add foggy data. If foggy data is very hard and costly to collect, it might not be worth the team's effort.

= 52% of your errors are due to foggy pictures. The results from this analysis implies that the team's highest priority should be to bring more foggy pictures into the training set so as to address the 8.0% of errors in that category. True/False?

A-True because it is greater than the other error categories added together 8.0>4.1+2.2+1.08.0>4.1+2.2+1.0.
B-First start with the sources of error that are least costly to fix.
C-False because it depends on how easy it is to add foggy data. If foggy data is very hard and costly to collect, it might not be worth the team's effort.
D-True because it is the largest category of errors. We should always prioritize the largest category of errors as this will make the best use of the team's time.

221. **You decide to focus on the dev set and check by hand what the errors are due to. Here is a table summarizing your discoveries:**
Overall dev set error: 15.3%
Errors due to incorrectly labeled data: 4.1%
Errors due to foggy pictures: 3.0%
Errors due to partially occluded elements: 7.2%
Errors due to other causes: 1.0%
In this table, 4.1%, 7.2%, etc. are a fraction of the total dev set (not just examples of your algorithm mislabeled). For example, about 7.2/15.3 = 47% of your errors are due to partially occluded elements.

2) You shouldn't invest all your efforts to get more images with partially occluded elements since 4.1 + 3.0 + 1.0 = 8.1 > 7.2.

2) False
These kinds of arguments don't help us to decide on the strategy to follow. Other factors should be used, such as the tradeoff between the cost of getting new images and the improvement of the system performance.

True/False?

222. **You decide to focus on the dev set and check by hand what the errors are due to. Here is a table summarizing your discoveries:**
Overall dev set error: 15.3%
Errors due to incorrectly labeled data: 4.1%
Errors due to foggy pictures: 3.0%
Errors due to partially occluded elements: 7.2%
Errors due to other causes: 1.0%
In this table, 4.1%, 7.2%, etc. are a fraction of the total dev set (not just examples of your algorithm mislabeled). For example, about 7.2/15.3 = 47% of your errors are due to partially occluded elements.
1) You find out that there is an anti-reflective film guarantee to eliminate the sun reflection, but it is quite costly. Which of the following gives the best description of what the investment in the film can do to the model?

A-The film will reduce the dev set error with 7.2% at the most.
B-The overall test set error will be reduced by at most 7.2%.
C-The film will reduce at least 7.2% of the dev set error.

1) A-The film will reduce the dev set error with 7.2% at the most.

Remember that this 7.2% gives us an estimate for the ceiling of how much the error can be reduced when the cause is fixed.

223. **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?**

A-Adding synthesized images that look like real foggy pictures taken from the front-facing camera of your car to the training dataset won't help

C-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

the model improve because it will introduce avoidable bias.
B-There is little risk of overfitting to the 1,000 pictures of fog so long as you are combining it with a much larger (>>1,000) set of clean/non-foggy images.
C-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.

problem you're solving.

If the synthesized images look realistic, then the model will just see them as if you had added useful data to identify road signs and traffic signals in foggy weather. I will very likely help.

---

224. **After working further on the problem, you've decided to correct the incorrectly labeled data. Your team corrects the labels of the wrongly predicted images on the dev set. Which of the following is a necessary step to take?**

A-Create a train-dev set to estimate how many incorrectly labeled examples are in the train set.
B-Use a correctly labeled version and an incorrectly labeled version to make the model more robust.
C-Correct the labels of the test set.
D-Correct the labels of the train set.

C-Correct the labels of the test set. - True

Recall that the dev set and the test set must come from the same distribution.

---

225. **After working further on the problem, you've decided to correct the incorrectly labeled data. Your team corrects the labels of the wrongly predicted images on the dev set.**
You have to correct the labels of the test so test and dev sets have the same distribution, but you won't change the labels on the train set because most models are robust enough they don't get severely affected by the difference in distributions. True/False?

A-True, as pointed out, we must keep dev and

A-True, as pointed out, we must keep dev and test with the same distribution. And the labels at training should be fixed only in case of a systematic error.

To successfully train a model, the dev set and test set should come from the same distri-

	test with the same distribution. And the labels at training should be fixed only in case of a systematic error. **B-False, the test set shouldn't be changed since we want to know how the model performs in real data.** **C-False, the test set should be changed, but also the train set to keep the same distribution between the train, dev, and test sets.**	bution. Also, the deep learning models are robust enough to handle a small change in distributions, but if the errors are systematic they can significantly affect the training of the model.
226.	**After working further on the problem, you've decided to correct the incorrectly labeled data on the dev set. Which of these statements do you agree with? (Check all that apply).** **A-You do not necessarily need to fix the incorrectly labeled data in the training set, because it's okay for the training set distribution to differ from the dev and test sets. Note that it is important that the dev set and test set have the same distribution.** **B-You should correct incorrectly labeled data in the training set as well so as to avoid your training set now being even more different from your dev set.** **C-You should not correct the incorrectly labeled data in the test set, so that the dev and test sets continue to come from the same distribution.** **D-You should also correct the incorrectly labeled data in the test set, so that the dev and test sets continue to come from the same distribution.**	A-You do not necessarily need to fix the incorrectly labeled data in the training set, because it's okay for the training set distribution to differ from the dev and test sets. Note that it is important that the dev set and test set have the same distribution. D-You should also correct the incorrectly labeled data in the test set, so that the dev and test sets continue to come from the same distribution.
227.	**Your client asks you to add the capability to detect dogs that may be crossing the road to the system. He can provide a relatively small set containing dogs. Which of the following do you agree most with?**	C-You can use weights pre-trained on the original data, and fine-tune with the data now including the dogs.

	A-You should train a single new model for the dogs' task, and leave the previous model as it is. B-You will have to re-train the whole model now including the dogs' data. C-You can use weights pre-trained on the original data, and fine-tune with the data now including the dogs. D-Using pre-trained weights can severely hinder the ability of the model to detect dogs since they have too many learned features.	Since your model has learned useful low-level features to tackle the new task we can conserve those by using the pre-trained weights.
228.	**One of your colleagues at the startup is starting a project to classify road signs as stop, dangerous curve, construction ahead, dead-end, and speed limit signs. Given how specific the signs are, he has only a small dataset and hasn't been able to create a good model. You offer your help providing the trained weights (parameters) of your model to transfer knowledge. But your colleague points out that his problem has more specific items than the ones you used to train your model. This makes the transfer of knowledge impossible. True/False?**	False. The model can benefit from the pre-trained model since there are many features learned by your model that can be used in the new problem.
229.	**So far your algorithm only recognizes red and green traffic lights. One of your colleagues in the startup is starting to work on recognizing a yellow traffic light. (Some countries call it an orange light rather than a yellow light; we'll use the US convention of calling it yellow.) Images containing yellow lights are quite rare, and she doesn't have enough data to build a good model. She hopes you can help her out using transfer learning. What do you tell your colleague? A-You cannot help her because the distribution**	B-She should try using weights pre-trained on your dataset, and fine-tuning further with the yellow-light dataset. Yes. You have trained your model on a huge dataset, and she has a small dataset. Although your labels are different, the parameters of your model have been trained

of data you have is different from hers, and is also lacking the yellow label.
B-She should try using weights pre-trained on your dataset, and fine-tuning further with the yellow-light dataset.
C-Recommend that she try multi-task learning instead of transfer learning using all the data.
D-If she has (say) 10,000 images of yellow lights, randomly sample 10,000 images from your dataset and put your and her data together. This prevents your dataset from "swamping" the yellow lights dataset.

to recognize many characteristics of road and traffic images which will be useful for her problem. This is a perfect case for transfer learning, she can start with a model with the same architecture as yours, change what is after the last hidden layer and initialize it with your trained parameters.

230. **One of your colleagues at the startup is starting a project to classify stop signs in the road as speed limit signs or not. He has approximately 30,000 examples of each image and 30,000 images without a sign. He thought of using your model and applying transfer learning but then he noticed that you use multi-task learning, hence he can't use your model.**

True/False?

False

Correct. When using transfer learning we can remove the last layer. That is one of the aspects that is different from a binary classification problem.

231. **When building a system to detect cattle crossing a road from images taken with the front-facing camera of a truck, the designers had a large dataset of images. Which of the following might be a reason to use an end-to-end approach?**

A-That is the default approach on computer vision tasks.
B-This approach will make use of useful hand-designed components.
C-It requires less computational resources.
D-There is a large dataset available.

D-There is a large dataset available.

Correct. To get good results when using an end-to-end approach, it is necessary to have a big dataset.

232.

Another colleague wants to use microphones placed outside the car to better hear if there are other vehicles around you. For example, if there is a police vehicle behind you, you would be able to hear their siren. However, they don't have much to train this audio system. How can you help?

A-Neither transfer learning nor multi-task learning seems promising.
B-Multi-task learning from your vision dataset could help your colleague get going faster. Transfer learning seems significantly less promising.
C-Transfer learning from your vision dataset could help your colleague get going faster. Multi-task learning seems significantly less promising.
D-Either transfer learning or multi-task learning could help our colleague get going faster.

A-Neither transfer learning nor multi-task learning seems promising.

233. **To recognize a stop sign you use the following approach: First, we localize any traffic sign in an image. After that, we determine if the sign is a stop sign or not. We are using multi-task learning.**
True/False?

False.
Multi-task learning is about joining several tasks that can benefit from each other.

234. **To recognize a stop sign you use the following approach:**
First, we localize any traffic sign in an image.
After that, we determine if the sign is a stop sign or not.
This is a better approach than an end-to-end model for which of the following cases? Choose the best answer.

A-The problem has a high Bayes error.
B-There is a large amount of data.

C-There is not enough data to train a big neural network.

Correct. This might be the most important factor when deciding whether to use an end-to-end approach.

	C-There is not enough data to train a big neural network. **D-There are available models which we can use to transfer knowledge.**	
235.	**Consider the following two approaches, A and B:** **· (A) Input an image (x) to a neural network and have it directly learn a mapping to make a prediction as to whether there's a red light and/or green light (y).** **· (B) In this two-step approach, you would first (i) detect the traffic light in the image (if any), then (ii) determine the color of the illuminated lamp in the traffic light.** **Approach A tends to be more promising than approach B if you have a ________ (fill in the blank).** **A-Large training set** **B-Problem with a high Bayes error.** **C-Multi-task learning problem.** **D-Large bias problem.**	A-Large training set In many fields, it has been observed that end-to-end learning works better in practice, but requires a large amount of data.
236.	**To recognize red and green lights, you have been using this approach:** **(A) Input an image (x) to a neural network and have it directly learn a mapping to make a prediction as to whether there's a red light and/or green light (y).** **A teammate proposes a different, two-step approach:** **(B) In this two-step approach, you would first (i) detect the traffic light in the image (if any), then (ii) determine the color of the illuminated lamp in the traffic light.** **Between these two, Approach B is more of an end-to-end approach because it has distinct steps for the input end and the output end.**	False (A) is an end-to-end approach as it maps directly the input (x) to the output (y).

True/False?

237. **An end-to-end approach doesn't require that we hand-design useful features, it only requires a large enough model.**

True/False?

True
This is one of the major characteristics of deep learning models, that we don't need to hand-design the features.

238. **What do you think applying this filter to a grayscale image will do?**
[[-1,-1,2],
[-1,2,1],
[2,1,1]]

A-Detect horizontal edges.
B-Detecting image contrast.
C-Detect 45-degree edges.
D-Detect vertical edges.

C-Detect 45-degree edges.

Notice that there is a high delta between the values in the top left part and the ones in the bottom right part. When convolving this filter on a grayscale image, the edges forming a 45-degree angle with the horizontal will be detected.

239. **What do you think applying this filter to a grayscale image will do?**
[[0,1,-1,0],
[1,3,-3,-1],
[1,3,-3,-1],
[0,1,-1,0]]

A-Detect vertical edges
B-Detect 45 degree edges
C-Detect horizontal edges
D-Detect image contrast

A-Detect vertical edges

As you can see the difference between values from the left part and values from the right of this filter is high. When convolving this filter on a grayscale image, the vertical edges will be detected.

240. **Suppose your input is a 128 by 128 color (RGB) image, and you are not using a convolutional network. If the first hidden layer has 64 neurons, each one fully connected to the input, how**

D-3145792

Correct, the number of inputs for each unit is

many parameters does this hidden layer have (including the bias parameters)?

A-3145728
B-1048640
C-1048576
D-3145792

128×128×3 since the input image is RGB, so we need 128×128×3×64 parameters for the weights and 64 parameters for the bias parameters, thus 128×128×3×64+64=314579

241. **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)?**

A-27,000,100
B-9,000,001
C-9,000,100
D-27,000,001

A-27,000,100

the number of weights is 300×300×3×100=27,000,00 when you add the bias terms (one per neuron) you get 27,000,10027,000,100.

242. **Suppose your input is a 128 by 128 grayscale image, and you are not using a convolutional network. If the first hidden layer has 256 neurons, each one fully connected to the input, how many parameters does this hidden layer have (including the bias parameters)?**

A-4194560
B-4194304
C-12582912
D-12583168

A-4194560

the number of inputs for each unit is 128×128 since the input image is grayscale, so we need 128×128×256 parameters for the weights and 256 parameters for the bias thus 128×128×256+256=419456

243. **Suppose your input is a 256 by 256 grayscale image, and you use a convolutional layer with 128 filters that are each 3×33×3. How many parameters does this hidden layer have (including**

C-1280

Yes, since the input volume has only one

the bias parameters)?

A-75497600
B-1152
C-1280
D-3584

channel each filter has 3×3+1weights including the bias, thus the total is (3×3+1)×128.

244. **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)?**

A-2501
B-7500
C-7600
D-2600

C-7600

Correct, you have 25×3=75 weights and 1 bias per filter. Given that you have 100 filters, you get 7,600 parameters for this layer.

245. **Suppose your input is a 256 by 256 color (RGB) image, and you use a convolutional layer with 128 filters that are each 7×7. How many parameters does this hidden layer have (including the bias parameters)?**

A-18944
B-1233125504
C-18816
D-6400

D-6400-false
recall that the filter must have matching channels with the input volume.

maybe A : 128x(49x3+1)

246. **You have an input volume that is 121×121×16, and convolve it with 32 filters of 4×4, using a stride of 3 and no padding. What is the output volume?**

A-40×40×16
B-118×118×16
C-118×118×32
D-40×40×32

D-40×40×32

n_out = (n_in+2xp-f)/s + 1
with p=0, f=4, s=3, n_in=121

247.

You have an input volume that is 31x31x32, and pad it using "pad=1". What is the dimension of the resulting volume (after padding)?

A-33x33x33
B-32x32x32
C-33x33x32
D-31x31x34

C-33x33x32

Yes, if the padding is 1 you add 2 to the height dimension and 2 to the width dimension.

248. **You have an input volume that is 63x63x16, and convolve it with 32 filters that are each 7x7, using a stride of 2 and no padding. What is the output volume?**

A-29x29x16
B-29x29x32
C-16x16x16
D-16x16x32

B-29x29x32

Yes, (63 7+0×2)/2+1=29 and the number of channels should match the number of filters.

249. **You have an input volume that is 61x61x32, and pad it using "pad=3". What is the dimension of the resulting volume (after padding)?**

A-64x64x35
B-67x67x32
C-64x64x32
D-61x61x35

B-67x67x32

if the padding is 3 you add 6 to the height dimension and 6 to the width dimension.

250. **You have a volume that is 64×64×32, and convolve it with 40 filters of 9×9, and stride 1. You want to use a "same" convolution. What is the padding?**

4
6
8
0

4

Yes, when using a padding of 4 the output volume has n_H =(64 9+2×4)/1 +1.

251. **You have a volume that is 121×121×32, and convolve it with 32 filters of 5×5, and a stride of 1.**

B-2

You want to use a "same" convolution. What is the padding?

A-3
B-2
C-5
D-0

Yes, when using a padding of 2 the output volume has n_H =(121 5+4)/1+1.

252. **You have an input volume that is 66x66x21, and apply max pooling with a stride of 3 and a filter size of 3. What is the output volume?**

A-22×22×7
B-66×66×7
C-21×21×21
D-22×22×21

D-22×22×21

n_out = (n_in+2xp-f)/s + 1
with p=0, f=3, s=3, n_in=66

253. **You have an input volume that is 128x128x12, and apply max pooling with a stride of 4 and a filter size of 4. What is the output volume?**

A-64×64×12
B-32×32×12
C-128×128×3
D-32×32×3

B-32×32×12

Yes, using the formula n_H[l] = (n_H[l 1] +2×p f)/s+1 with p=0, f=4, s=4 and n_H[l 1] =32.

254. **You have an input volume that is 15x15x8, and pad it using "pad=2". What is the dimension of the resulting volume (after padding)?**

A-17x17x8
B-17x17x10
C-19x19x8
D-19x19x12

C-19x19x8

Correct, padding is applied over the height and the width of the input image. If the padding is two, you add 4 to the height dimension and 4 to the width dimension.

255. **You have an input volume that is 32x32x16, and apply max pooling with a stride of 2 and a filter size of 2. What is the output volume?**

D-16x16x16

	A-16x16x8 B-15x15x16 C-32x32x8 D-16x16x16	
256.	**Which of the following are hyperparameters of the pooling layers? (Choose all that apply) A-Number of filters. B-Whether it is max or average. C-Filter size. D-Average weights.**	B-Whether it is max or average. C-Filter size.
257.	**Which of the following are hyperparameters of the pooling layers? (Choose all that apply) A'-Whether it is max or average. B'-b[l] bias. C'-W[l] weights. D'-Stride**	A'-Whether it is max or average. D'-Stride
258.	**Which of the following are true about convolutional layers? (Check all that apply) A-It allows a feature detector to be used in multiple locations throughout the whole input volume. B-Convolutional layers provide sparsity of connections. C-It speeds up the training since we don't need to compute the gradient for convolutional layers.**	A-It allows a feature detector to be used in multiple locations throughout the whole input volume. B-Convolutional layers provide sparsity of connections.
259.	**Because pooling layers do not have parameters, they do not affect the backpropagation (derivatives) calculation. True False**	False Everything that influences the loss should appear in the backpropagation because we are computing derivatives. In

fact, pooling layers modify the input by choosing one value out of several values in their input volume. Also, to compute derivatives for the layers that have parameters (Convolutions, Fully-Connected), we still need to backpropagate the gradient through the Pooling layers.

260. **Which of the following are the benefits of using convolutional layers? (Check all that apply)**

A-Convolutional layers are good at capturing translation invariance.
B-It reduces the computations in backpropagation since we omit the convolutional layers in the process.
C-It reduces the total number of parameters, thus reducing overfitting through parameter sharing.

A-Convolutional layers are good at capturing translation invariance.

C-It reduces the total number of parameters, thus reducing overfitting through parameter sharing.

261. **The following image depicts the result of a convolution at the right when using a stride of 1 and the filter is shown right next.**

On which pixels does the circled pixel of the activation at the right depend?

A-It depends on the pixels enclosed by the blue square.
B-It depends on the pixels enclosed by the red square.
C-It depends on all the pixels of the image on the left.

D-It depends on the pixels enclosed by the green square.

this is the position of the filter when we move it two pixels down and one to the right.

D-It depends on the pixels enclosed by the green square.

262. **The sparsity of connections and weight sharing are mechanisms that allow us to use fewer parameters in a convolutional layer making it possible to train a network with smaller training sets.**

True/False?

True
weight sharing reduces significantly the number of parameters in a neural network, and sparsity of connections allows us to use a smaller number of inputs thus reducing even further the number of parameters.

263. **In lecture we talked about "parameter sharing" as a benefit of using convolutional networks. Which of the following statements about parameter sharing in ConvNets are true? (Check all that apply)**

A-It reduces the total number of parameters, thus reducing overfitting.
B-It allows gradient descent to set many of the parameters to zero, thus making the connections sparse.
C-It allows parameters learned for one task to be shared even for a different task (transfer learning).
D-It allows a feature detector to be used in multiple locations throughout the whole input image/input volume.

A-It reduces the total number of parameters, thus reducing overfitting.

D-It allows a feature detector to be used in multiple locations throughout the whole input image/input volume.

264. **Which of the following are hyperparameters of the pooling layers? (Choose all that apply)**

A-Whether it is max or average.
B-Average weights.
C-Filter size.
D-Number of filters.

A-Whether it is max or average.

C-Filter size.

265.	**Which of the following are the benefits of using convolutional layers? (Check all that apply)** **A-It reduces the total number of parameters, thus reducing overfitting through parameter sharing.** **B-Convolutional layers are good at capturing translation invariance.** **C-It reduces the computations in backpropagation since we omit the convolutional layers in the process.**	A-It reduces the total number of parameters, thus reducing overfitting through parameter sharing. B-Convolutional layers are good at capturing translation invariance.
266.	**In lecture we talked about "sparsity of connections" as a benefit of using convolutional layers. What does this mean?** **A-Each filter is connected to every channel in the previous layer.** **B-Each activation in the next layer depends on only a small number of activations from the previous layer.** **C-Each layer in a convolutional network is connected only to two other layers** **D-Regularization causes gradient descent to set many of the parameters to zero.**	B-Each activation in the next layer depends on only a small number of activations from the previous layer. Yes, each activation of the output volume is computed by multiplying the parameters from with a volumic slice of the input volume and then summing all these together.
267.	**When building a ConvNet, typically you start with some POOL layers followed by some CONV layers.** **True/False?**	False Correct. It is typical for ConvNets to use a POOL layer after some Conv layers; sometimes even one POOL layer after each CONV layer; but is not common to start with POOL layers.
268.		

Which of the following do you typically see in a ConvNet? (Check all that apply.)

A-Multiple CONV layers followed by a POOL layer
B-FC layers in the first few layers
C-Multiple POOL layers followed by a CONV layer
D-FC layers in the last few layers

A-Multiple CONV layers followed by a POOL layer

D-FC layers in the last few layers

269. **Which of the following do you typically see in ConvNet? (Check all that apply.)**

A-Multiple FC layers followed by a CONV layer.
B-Use of FC layers after flattening the volume to generate output classes.
C-ConvNet makes exclusive use of CONV layers.
D-Use of multiple POOL layers followed by a CONV layer.

B-Use of FC layers after flattening the volume to generate output classes.

FC layers are typically used in the last few layers after flattening the volume to generate the output in classification.

270. **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.**

True/False?

False

271. **In LeNet - 5 we can see that as we get into deeper networks the number of channels increases while the height and width of the volume decreases.**

True/False?

True

since in its implementation only valid convolutions were used, without padding, the height and width of the volume were reduced at each convolution. These were also reduced by the POOL layers, whereas the number of channels was increased from 6 to 16.

272. **LeNet - 5 made extensive use of padding to create valid convolutions, to avoid increasing the number of channels after every convolutional layer.**

True/False?

False

back in 1998 when the corresponding paper of LeNet - 5 was written padding wasn't used.

273. **Training a deeper network (for example, adding additional layers to the network) allows the network to fit more complex functions and thus almost always results in lower training error. For this question, assume we're referring to "plain" networks.**

True/False?

False

Correct, Resnets are here to help us train very deep neural networks.

274. **The motivation of Residual Networks is that very deep networks are so good at fitting complex functions that when training them we almost always overfit the training data.**

True/False?

False

very deep neural networks are hard to train and a deeper network does not always imply lower training error. Residual Networks allow us to train very deep neural networks.

275. **The computation of a ResNet block is expressed in the equation:**
a[l+2] = g(W[l+2].g(W[l+1]a[l] + b[l+1]) + b[l+2] + a[l])

Box C: W[l+2] - Box A: b[l+1] - Box B: a[l]

B-The term in the orange box, marked as B.

Which part corresponds to the skip connection?

A-The term in the red box, marked as C.
B-The term in the orange box, marked as B.
C-The term in the blue box, marked as A.
D-The equation of ResNet.

276. **When having a small training set to construct a classification model, which of the following is a strategy of transfer learning that you would use to build the model?**

A-Use an open-source network trained in a larger dataset. Use these weights as an initial point for the training of the whole network.
B-Use an open-source network trained in a larger dataset freezing the layers and re-train the softmax layer.
C-It is always better to train a network from a random initialization to prevent bias in our model.
D-Use an open-source network trained in a larger dataset, freeze the softmax layer, and re-train the rest of the layers.

B-Use an open-source network trained in a larger dataset freezing the layers and re-train the softmax layer.

277. **In the best scenario when adding a ResNet block it will learn to approximate the identity function after a lot of training, helping improve the overall performance of the network. True/False?**

False

When adding a ResNet block it can easily learn to approximate the identity function, thus in a worst-case scenario, it will not affect the performance of the network at all.

278. **Adding a ResNet block to the end of a network makes it deeper. Which of the following is true?**

C-The performance of the networks doesn't get

	A-The performance of the networks is hurt since we make the network harder to train. B-The number of parameters will decrease due to the shortcut connections. C-The performance of the networks doesn't get hurt since the ResNet block can easily approximate the identity function. D-It shifts the behavior of the network to be more like the identity function.	hurt since the ResNet block can easily approximate the identity function.
279.	**Suppose you have an input volume of dimension nH x nW x nC . Which of the following statements do you agree with? (Assume that the "1x1 convolutional layer" below always uses a stride of 1 and no padding.) A-You can use a 2D pooling layer to reduce nH, nW, and nC. B-You can use a 1x1 convolutional layer to reduce nH, nW, and nC. C-You can use a 2D pooling layer to reduce nH, nW, but not nC. D-You can use a 1x1 convolutional layer to reduce nC but not nH and nW.**	C-You can use a 2D pooling layer to reduce nH, nW, but not nC. D-You can use a 1x1 convolutional layer to reduce nC but not nH and nW.
280.	**1×1 convolutions are the same as multiplying by a single number. True/False?**	False a 1×1 layer doesn't act as a single number because it makes a sum over the depth of the volume.
281.	**For a volume of 125×125×64 which of the following can be used to reduce this to a 125×125×32 volume? A-Use a 1×1 convolutional layer with a stride of**	A-Use a 1×1 convolutional layer with a stride of 1, and 32 filters. since using 1×1 convo-

1, and 32 filters.
B-Use a POOL layer of size 2×2 but with a stride of 1.
C-Use a 1×1 convolutional layer with a stride of 2, and 32 filters.
D-Use a POOL layer of size 2×2 with a stride of 2.

lutions is a great way to reduce the depth dimension without affecting the other dimensions.

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

A-A ResNet with L layers would have on the order of L2 skip connections in total.
B-The skip-connection makes it easy for the network to learn an identity mapping between the input and the output within the ResNet block.
C-Using a skip-connection helps the gradient to backpropagate and thus helps you to train deeper networks
D-The skip-connections compute a complex non-linear function of the input to pass to a deeper layer in the network.

B-The skip-connection makes it easy for the network to learn an identity mapping between the input and the output within the ResNet block.
C-Using a skip-connection helps the gradient to backpropagate and thus helps you to train deeper networks

283. **Which ones of the following statements on Inception Networks are true?**

A-A single inception block allows the network to use a combination of 1x1, 3x3, 5x5 convolutions and pooling.
B-Inception networks incorporate 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.
C-Making an inception network deeper (by stacking more inception blocks together) can improve performance, but can also lead to over-

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

C-Making an inception network deeper (by stacking more inception blocks together) can improve performance, but can also lead to overfitting and increase in computational cost.

	fitting and increase in computational cost. **D-Inception blocks usually use 1x1 convolutions to reduce the input data volume's size before applying 3x3 and 5x5 convolutions.**	D-Inception blocks usually use 1x1 convolutions to reduce the input data volume's size before applying 3x3 and 5x5 convolutions.
284.	**Which of the following are true about the inception Network? (Check all that apply)** **A-Inception blocks allow the use of a combination of 1x1, 3x3, 5x5 convolutions and pooling by stacking up all the activations resulting from each type of layer.** **B-Inception blocks allow the use of a combination of 1x1, 3x3, 5x5 convolutions, and pooling by applying one layer after the other.** **C-One problem with simply stacking up several layers is the computational cost of it.** **D-Making an inception network deeper won't hurt the training set performance.**	A-Inception blocks allow the use of a combination of 1x1, 3x3, 5x5 convolutions and pooling by stacking up all the activations resulting from each type of layer. C-One problem with simply stacking up several layers is the computational cost of it.
285.	**Which of the following are true about bottleneck layers? (Check all that apply)** **A-By adding these layers we can reduce the computational cost in the inception modules.** **B-The use of bottlenecks doesn't seem to hurt the performance of the network.** **C-Bottleneck layers help to compress the 1x1, 3x3, 5x5 convolutional layers in the inception network.** **D-The bottleneck layer has a more powerful regularization effect than Dropout layers.**	A-By adding these layers we can reduce the computational cost in the inception modules. B-The use of bottlenecks doesn't seem to hurt the performance of the network.
286.	**Which of the following are common reasons for using open-source implementations of ConvNets (both the model and/or weights)? Check all that apply.**	B-It is a convenient way to get working with an implementation of a complex ConvNet archi-

A-A model trained for one computer vision task can usually be used to perform data augmentation for a different computer vision task.
B-It is a convenient way to get working with an implementation of a complex ConvNet architecture.
C-The same techniques for winning computer vision competitions, such as using multiple crops at test time, are widely used in practical deployments (or production system deployments) of ConvNets.
D-Parameters trained for one computer vision task are often useful as pre-training for other computer vision tasks.

tecture.

D-Parameters trained for one computer vision task are often useful as pre-training for other computer vision tasks.

287. **Models trained for one computer vision task can't be used directly in another task. In most cases, we must change the softmax layer, or the last layers of the model and re-train for the new task.**

True/False?

True

Yes, this is a good way to take advantage of open-source models trained more or less for the task you want to do. This may also help you save a great number of computational resources and data.

288. **Which of the following are true about Depthwise-separable convolutions? (Choose all that apply)**

A-The depthwise convolution convolves the input volume with 1×1 filters over the depth dimension.
B-The pointwise convolution convolves the output volume with 1×1 filters.
C-The depthwise convolution convolves each channel in the input volume with a separate

B-The pointwise convolution convolves the output volume with 1×1 filters.
C-The depthwise convolution convolves each channel in the input volume with a separate filter.
D-Depthwise-separable convolutions are

	filter. **D-Depthwise-separable convolutions are composed of two different types of convolutions.**	composed of two different types of convolutions.
289.	**Which of the following are true about Depth wise-separable convolutions? (Choose all that apply)** **A-The result has always the same number of channels nc as the input.** **B-They are just a combination of a normal convolution and a bottleneck layer.** **C-They combine depthwise convolutions with pointwise convolutions.** **D-They have a lower computational cost than normal convolutions.**	C-They combine depthwise convolutions with pointwise convolutions. D-They have a lower computational cost than normal convolutions.
290.	**In Depthwise Separable Convolution you:** **A-Perform one step of convolution.** **B-For the "Depthwise" computations each filter convolves with all of the color channels of the input image.** **C-For the "Depthwise" computations each filter convolves with only one corresponding color channel of the input image.** **D-You convolve the input image with a filter of nf x nf x nc where nc acts as the depth of the filter (nc is the number of color channels of the input image).** **E-The final output is of the dimension nout x nout x n2c (where n2c is the number of filters used in the pointwise convolution step).** **F-Perform two steps of convolution.** **G-The final output is of the dimension nout x nout x nc (where nc is the number of color channels of the input image).** **H-You convolve the input image with nc number**	C-For the "Depthwise" computations each filter convolves with only ONE corresponding color channel of the input image. E-The final output is of the dimension nout x nout x n2c (where n2c is the number of filters used in the pointwise convolution step). F-Perform two steps of convolution. H-You convolve the input image with nc number of nf x nf filters (nc is the number of color channels of the input image).

of nf x nf filters (nc is the number of color channels of the input image).

291. **Suppose that in a MobileNet v2 Bottleneck block the input volume has shape 64×64×16. If we use 32 filters for the expansion and 1616 filters for the projection. What is the size of the input and output volume of the depthwise convolution, assuming a pad='same'?**

A-64×64×32 64×64×32
B-64×64×32 64×64×16
C-64×64×16 64×64×32
D-32×32×32 32×32×32

A-64×64×32 64×64×32
Correct, the size of the input and output volume of the depthwise convolution is determined by the number of filters in the expansion.

C-Incorrect, the input and output volume of the depthwise convolution are the same.

292. **Suppose that in a MobileNet v2 Bottleneck block we have an n×n×5 input volume, we use 30 filters for the expansion, in the depthwise convolutions we use 3×3 filters, and 20 filters for the projection. How many parameters are used in the complete block, suppose we don't use bias?**

A-1020
B-80
C-1101
D-8250

B false
A-1020

Yes, the expansion filters use 5 × 30 = 150 parameters, the depthwise convolutions need 3 × 3 × 30 = 270 parameters, and the projection part 30 × 20 = 600 parameters.

293. **You are building a 3-class object classification and localization algorithm. The classes are: pedestrian (c=1), car (c=2), motorcycle (c=3). What should y be for the image below? Remember that "?" means "don't care", which means that the neural network loss function won't care what the neural network gives for that component of the output. Recall y=[pc ,bx ,by ,bh ,bw ,c1 ,c2 ,c3].**

A-y=[1,0.66,0.5,0.75,0.16,1,

p_c =1 since there is a pedestrian in the picture. We can see that bx ,by as percentages of the image are approximately correct as well bh ,bw , and the value of c1 =1 for a pedestrian.

A-y=[1,0.66,0.5,0.75,0.16,1,0,0]
B-y=[1,?,?,?,?,1,?,?]
C-y=[1,0.66,0.5,0.75,0.16,0,0,0]
D-y=[1,0.66,0.5,0.16,0.75,1,0,0]

https://www.pexels.com/es-es/foto/mujer-vestida-con-falda-azul-y-blanca-caminando-cerca-de-la-hierba-verde-durante-el-dia-144474/

294. **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 appear the same size in the image. There is at most one soft drink can in each image. Here are some typical images in your training set:**
What are the most appropriate (lowest number of) output units for your neural network?

A-Logistic unit, bx and by
B-Logistic unit, bx, by, bh, bw
C-Logistic unit (for classifying if there is a soft-drink can in the image)
D-Logistic unit, bx, by, bh (since bw = bh)

A-Logistic unit, bx and by

295. **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 appear the same size in the image. There is at most one soft drink can in each image. Here're some typical images in**

D-We can approach the task as an image classification with a localization problem.

We can use a network to combine the two tasks similar to that described in the lectures.

**your training set:
To solve this task it is necessary to divide the task into two: 1. Construct a system to detect if a can is present or not. 2. Construct a system that calculates the bounding box of the can when present. Which one of the following do you agree with the most?**

**A-We can't solve the task as an image classification with a localization problem since all the bounding boxes have the same dimensions.
B-An end-to-end solution is always superior to a two-step system.
C-The two-step system is always a better option compared to an end-to-end solution.
D-We can approach the task as an image classification with a localization problem.**

296. **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 appear the same size in the image. There is at most one soft-drink can in each image. Here are some typical images in your training set:**

The most adequate output for a network to do the required task is y=[pc ,bx ,by ,bh ,bw ,c1]. (Which of the following do you agree with the most?)

**A-False, since we only need two values c1 for no soft-drink can and c2 for soft-drink can.
B-False, we don't need bh, bw since the cans are all the same size.**

B-False, we don't need bh, bw since the cans are all the same size.

Correct. With the position bx ,by we can completely characterize the position of the object if it is present. We should use only one additional logistic unit to indicate if the object is present or not.

C-True, pc indicates the presence of an object of interest, bx,by,bh,bw indicate the position of the object and its bounding box, and c1 indicates the probability of there being a can of soft-drink.
D-True, since this is a localization problem.

297. **When building a neural network that inputs a picture of a person's face and outputs N landmarks on the face (assume that the input image contains exactly one face), we need two coordinates for each landmark, thus we need 2N output units.**

True/False?

True

Recall that each landmark is a specific position in the face's image, thus we need to specify two coordinates for each landmark.

298. **When building a neural network that inputs a picture of a person's face and outputs N landmarks on the face (assume that the input image contains exactly one face), which is true about y^ (i)?**

A-^y(i) has shape (N, 1)
B-^y(i) stores the probability that a landmark is in a given position over the face.
C-^y(i) has shape (1, 2N)
D-^y(i) has shape (2N, 1)

D-^y(i) has shape (2N, 1)

299. **You are working to create an object detection system, like the ones described in the lectures, to locate cats in a room. To have more data with which to train, you search on the internet and find a large number of cat photos.**
Which of the following is true about the system?

A-We should add the internet images (without the presence of bounding boxes in them) to the train set.

D-We can't add the internet images unless they have bounding boxes.

As this is a localization model, we also need the coordinates of the bounding boxes, not just the images.

B-We should use the internet images in the dev and test set since we don't have bounding boxes.
C-We can't use internet images because it changes the distribution of the dataset.
D-We can't add the internet images unless they have bounding boxes.

300. **When training one of the object detection systems described in the lectures, you need a training set that contains many pictures of the object(s) you wish to detect. However, bounding boxes do not need to be provided in the training set, since the algorithm can learn to detect the objects by itself.**

True/False

False

you need bounding boxes in the training set. Your loss function should try to match the predictions for the bounding boxes to the true bounding boxes from the training set.

301. **When training one of the object detection systems described in the lectures, each image must have zero or exactly one bounding box.**

True/False?

False.

In a single image, there might be more than only one instance of the object we are trying to localize, so it must have several bounding boxes.

302. **Suppose you are using YOLO on a 19x19 grid, on a detection problem with 20 classes, and with 5 anchor boxes. During training, for each image you will need to construct an output volume y as the target value for the neural network; this corresponds to the last layer of the neural network. (y may include some "?", or "don't cares"). What is the dimension of this output volume?**

A-19x19x(25x20)

D-19x19x(5x25)

Correct, you get a 19x19 grid where each cell encodes information about 5 boxes and each box is defined by a confidence probability (p_c), 4 coordinates (bx ,by ,bh ,bw) and classes (c1 ,...,c20).

B-19x19x(5x20)
C-19x19x(20x25)
D-19x19x(5x25)

303. **If we use anchor boxes in YOLO we no longer need the coordinates of the bounding box bx ,by ,bh ,bw since they are given by the cell position of the grid and the anchor box selection.**

True/False?

False

We use the grid and anchor boxes to improve the capabilities of the algorithm to localize and detect objects, for example, two different objects that intersect, but we still use the bounding box coordinates.

304. **Which of the following do you agree with about the use of anchor boxes in YOLO? Check all that apply.**

A-Each object is assigned to any anchor box that contains that object's midpoint.
B-Each object is assigned to an anchor box with the highest IoU inside the assigned cell.
C-They prevent the bounding box from suffering from drifting.
D-Each object is assigned to the grid cell that contains that object's midpoint.

B-Each object is assigned to an anchor box with the highest IoU inside the assigned cell.

D-Each object is assigned to the grid cell that contains that object's midpoint.

305. **What is Semantic Segmentation?**

A-Locating an object in an image belonging to a certain class by drawing a bounding box around it.
B-Locating objects in an image belonging to different classes by drawing bounding boxes around them.
C-Locating objects in an image by predicting each pixel as to which class it belongs to.

C-Locating objects in an image by predicting each pixel as to which class it belongs to.

306. **Semantic segmentation can only be applied to classify pixels of images in a binary way as 1 or 0, according to whether they belong to a certain class or not.**

True/False?

False

The same ideas used for multi-class classification can be applied to semantic segmentation.

307. **Using the concept of Transpose Convolution, fill in the values of X, Y and Z below.**
(padding = 1, stride = 2)

Input: 2x2 : [[1, 2], [3, 4]]
Filter: 3x3 : [[1, 0, -1], [1, 0, -1], [1, 0, -1]]
Result: 6x6: [[0, 1, 0, -2], [0, X, 0, Y], [0, 1, 0, Z], [0, 1, 0, -4]]

A-X = -2, Y = -6, Z = -4
B-X = 2, Y = -6, Z = 4
C-X = 2, Y = -6, Z = -4
D-X = 2, Y = 6, Z = 4

C-X = 2, Y = -6, Z = -4

308. **Suppose your input to a U-Net architecture is x w x 3, where 3 denotes your number of channels (RGB). What will be the dimension of your output ?**

A- x w x n, where n = number of input channels
B- x w x n, where n = number of filters used in the algorithm
C- x w x n, where n = number of output classes
D- x w x n, where n = number of of output channels

C- x w x n, where n = number of output classes

309. **When using the U-Net architecture with an input h×w×c, where c denotes the number of channels, the output will always have the shape h×w×c.**

True/False?

False

The output of the U-Net architecture can be h×w×k where k is the number of classes.

		The number of channels doesn't have to match between input and output.
310.	**When using the U-Net architecture with an input h×w×c, where c denotes the number of channels, the output will always have the shape h×w.** **True/False?**	False Correct. The output of the U-Net architecture can be h×w×k where k is the number of classes.
311.	**We are trying to build a system that assigns a value of 1 to each pixel that is part of a tumor from a medical image taken from a patient. This is a problem of localization?** **True/False?**	False Correct. This is a problem of semantic segmentation since we need to classify each pixel from the image.
312.	**Face verification and face recognition are the two most common names given to the task of comparing a new picture against one person's face.** **True/False?**	False This is the description of face verification, but not of face recognition.
313.	**Which of the following do you agree with?** **A-Face verification requires K comparisons of a person's face.** **B-Face recognition requires K comparisons of a person's face.** **C-Face recognition requires comparing pictures against one person's face.**	B-Face recognition requires K comparisons of a person's face. in face recognition we compare the face of one person to K to classify the face as one of those K or not.
314.	**Face verification requires comparing a new picture against one person's face, whereas face recognition requires comparing a new picture**	True

	against K persons' faces. **True/False?**	
315.	**Why do we learn a function d(img1,img2) for face verification? (Select all that apply.)** **A-Given how few images we have per person, we need to apply transfer learning.** **B-We need to solve a one-shot learning problem.** **C-This allows us to learn to predict a person's identity using a softmax output unit, where the number of classes equals the number of persons in the database plus 1 (for the final "not in database" class).** **D-This allows us to learn to recognize a new person given just a single image of that person.**	B-We need to solve a one-shot learning problem. D-This allows us to learn to recognize a new person given just a single image of that person.
316.	**Why is the face verification problem considered a one-shot learning problem? Choose the best answer.** **A-Because we have only have to forward pass the image one time through our neural network for verification.** **B-Because we might have only one example of the person we want to verify.** **C-Because we are trying to compare to one specific person only.** **D-Because of the sensitive nature of the problem, we won't have a chance to correct it if the network makes a mistake.**	B-Because we might have only one example of the person we want to verify. Correct. One-shot learning refers to the amount of data we have to solve a task.
317.	**You want to build a system that receives a person's face picture and determines if the person is inside a workgroup. You have pictures of all the faces of the people currently in the workgroup, but some members might leave, and**	A-It will be more efficient to learn a function d(img1,img2) for this task. B-This can be consid-

some new members might be added. Which of the following do you agree with?

A-It will be more efficient to learn a function d(img1,img2) for this task.
B-This can be considered a one-shot learning task.
C-It is best to build a convolutional neural network with a softmax output with as many outputs as members of the group.
D-This can't be considered a one-shot learning task since there might be many members in the workgroup.

ered a one-shot learning task.

318. **You want to build a system that receives a person's face picture and determines if the person is inside a workgroup. You have pictures of all the faces of the people currently in the workgroup, but some members might leave, and some new members might be added. To train a system to solve this problem using the triplet loss you must collect pictures of different faces from only the current members of the team.**

True/False?

False

Although it is necessary to have several pictures of the same person, it is not absolutely necessary that all the pictures only come from current members of the team.

319. **You want to build a system that receives a person's face picture and determines if the person is inside a workgroup. You have pictures of all the faces of the people currently in the workgroup, but some members might leave, and some new members might be added. To train a system to solve this problem using the triplet loss you get many persons and take several pictures of each one. Which of the following do you agree with? (Select the best answer.)**

A-You shouldn't use persons outside the workgroup you are interested in because that might

C-You take several pictures of the same person to train d(img1,img2) using the triplet loss.

create a high variance in your model.
B-You take several pictures of the same person because this way you can get more pictures to train the network efficiently since you already have the person in place.
C-You take several pictures of the same person to train d(img1,img2) using the triplet loss.
D-It would be best to increase the number of persons in the dataset by taking only one picture of each person to have a more representative set of the population.

320. **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.**

True/ False?

False

to train a network using the triplet loss you need several pictures of the same person.

321. **In the triplet loss:**
max(%f(A) f(P)%2 %f(A) f(N)%2+±,0)
Which of the following are true about the triplet loss? Choose all that apply.

A-f(A) represents the encoding of the Anchor.
B-±is a trainable parameter of the Siamese network.
C-We want that %f(A) f(P)%^2<%f(A) f(N)%^2 so the negative images are further away from the anchor than the positive images.
D-A the anchor image is a hyperparameter of the Siamese network.

A-f(A) represents the encoding of the Anchor.

C-We want that
%f(A) f(P)%^2<%f(A) f(N)%
the negative images are further away from the anchor than the positive images.

322. **Triplet loss:**
max(%f(A) f(P)%2 %f(A) f(N)%2+±,0)
is larger in which of the following cases?

A-When the encoding of A is closer to the encoding of P than to the encoding of N.

C-When the encoding of A is closer to the encoding of N than to the encoding of P.

B-When A=P and A=N.
C-When the encoding of A is closer to the encoding of N than to the encoding of P.

323. **Consider the following Siamese network architecture:**
Which of the following do you agree with the most?

A-This depicts two *different* neural networks with different architectures, although we use the same drawing.
B-Although we depict two neural networks and two images, the two images are combined in a single volume and pass through a single neural network.
C-The upper and lower neural networks depicted have exactly the same parameters, but the outputs are computed independently for each image.
D-The two neural networks depicted in the image have the same architecture, but they might have different parameters.

C-The upper and lower neural networks depicted have exactly the same parameters, but the outputs are computed independently for each image.

Both neural networks share the same weights, and each image passes through the neural network in an independent manner.

324. **You train a ConvNet on a dataset with cats, dogs, birds, and other types of animals. You try to find a filter that strongly responds to horizontal edges. You are more likely to find this filter in layer 6 of the network than in layer 1.**

True/False?

False

Edges are a very low-level feature, thus it is more likely to find such a feature detector in the first layers of the network.

325. **You train a ConvNet on a dataset with 100 different classes. You wonder if you can find a hidden unit which responds strongly to pictures of cats. (I.e., a neuron so that, of all the input/training images that strongly activate that neuron, the majority are cat pictures.) You are**

True

Yes, this neuron understands complex shapes (cat pictures) so it is more likely to be in a

	more likely to find this unit in layer 4 of the network than in layer 1 **True/False?**	deeper layer than in the first layer.
326.	**Our intuition about the layers of a neural network tells us that units that respond more to complex features are more likely to be in deeper layers.** **True/False?**	True Neurons that understand more complex shapes are more likely to be in deeper layers of a neural network.
327.	**In neural style transfer, we define style as:** **A-The correlation between activations across channels of an image.** **B-%a[l](S) a[l](G)%^2 the distance between the activation of the style image and the content image.** **C-The correlation between the activation of the content image C and the style image S.** **D-The correlation between the generated image G and the style image S.**	A-The correlation between activations across channels of an image. this correlation is represented by Gkk2[l](I) for the image I.
328.	**Neural style transfer uses images Content C, Style S. The loss function used to generate image G is composed of which of the following: (Choose all that apply.)** **A-Jstyle that compares S and G.** **B-Jcorr that compares C and S.** **C-T that calculates the triplet loss between S, G, and C.** **D-Jcontent that compares C and G.**	A-Jstyle that compares S and G. D-Jcontent that compares C and G.
329.	**Neural style transfer is trained as a supervised learning task in which the goal is to input two images (x), and train a network to output a new, synthesized image (y).**	False Yes, Neural style transfer is about training the pixels of an image to

True/False?

make it look artistic, it is not learning any parameters.

330. **In neural style transfer, we train the pixels of an image, and not the parameters of a network.**

True/ False?

True

Neural style transfer compares the high-level features of two images and modifies the pixels of one of them in order to look artistic.

331. **In neural style transfer the content loss Jcont is computed as:**
%2J_cont (G,C)=% a[l](C) a[l](G)%^ 2
Where a[l](k) is the activation of the l-th layer of a ConvNet trained for classification. We choose l to be a very high value to use compared to the more abstract activation of each image.

True/False?

False

We don't use a very deep layer since this will only compare if the two images belong to the same category.

332. **In neural style transfer, what is updated in each iteration of the optimization algorithm?**

A-The neural network parameters
B-The pixel values of the generated image G
C-The pixel values of the content image C
D-The regularization parameters

B-The pixel values of the generated image G

neural style transfer is different from many of the algorithms you've seen up to now, because it doesn't learn any parameters; instead it learns directly the pixels of an image.

333. **In the deeper layers of a ConvNet, each channel corresponds to a different feature detector. The style matrix G[l] measures the degree to which the activations of different feature detectors in layer l vary (or correlate) together with each**

True

Yes, the style matrix G[l] can be seen as a matrix of cross-correlations be-

	other. **True/False?**	tween the different feature detectors.
334.	**In neural style transfer, we can't use gradient descent since there are no trainable parameters.** **True/False?**	False We use gradient descent on the cost function J(G) and we update the pixel values of the generated image G.
335.	**In neural style transfer, which of the following better express the gradients used?** **A-Neural style transfer doesn't use gradient descent since there are no trainable parameters.** **B- J/ G** **C- J/ W[l]** **D- J/ S**	B- J/ G we use the gradient of the cost function over the value of the pixels of the generated image.
336.	**You are working with 3D data. The input "image" has size 32×32×32×3, if you apply a convolutional layer with 16 filters of size 4×4×4, zero padding and stride 1. What is the size of the output volume?** **A-31×31×31×16** **B-29×29×29×16** **C-29×29×29×13** **D-29×29×29×3**	B-29×29×29×16
337.	**You are working with 3D data. The input "image" has size 64×64×64×3, if you apply a convolutional layer with 16 filters of size 4×4×4, zero padding and stride 2. What is the size of the output volume?** **A-64×64×64×3.** **B-31×31×31×3.**	C-31×31×31×16.

	C-31×31×31×16. **D-61×61×61×14.**	
338.	**You are working with 3D data. You are building a network layer whose input volume has size 32x32x32x16 (this volume has 16 channels), and applies convolutions with 32 filters of dimension 3x3x3x16 (no padding, stride 1). What is the resulting output volume?** **A-30x30x30x32** **B-30x30x30x16** **C-Undefined: This convolution step is impossible and cannot be performed because the dimensions specified don't match up.**	A-30x30x30x32