# 深度学习测试题

机器学习初学者

算法讲解　　论文解读

学习路线　　学术技巧

长按二维码
关注公众号

# 目录

# Lesson1 Neural Networks and Deep Learning (第一门课 神经网络和深度学习)

## Week 1 Quiz - Introduction to deep learning（第一周测验 - 深度学习简介）

1. What does the analogy "AI is the new electricity" refer to?(和"AI 是新电力"相类似的说法是什么？)

【 】AI is powering personal devices in our homes and offices, similar to electricity.(AI 为我们的家庭和办公室的个人设备供电，类似于电力。)

【 】Through the "smart grid", AI is delivering a new wave of electricity.(通过"智能电网"，AI 提供新的电能。)

【 】AI runs on computers and is thus powered by electricity, but it is letting computers do things not possible before.(AI 在计算机上运行，并由电力驱动，但是它正在让以前的计算机不能做的事情变为可能。)

【★】Similar to electricity starting about 100 years ago, AI is transforming multiple industries.(就像 100 年前产生电能一样，AI 正在改变很多的行业。)

Note: Andrew illustrated the same idea in the lecture.(注: 吴恩达在视频中表达了同样的观点。)

2. Which of these are reasons for Deep Learning recently taking off? (Check the two options that apply.)(哪些是深度学习快速发展的原因？ (两个选项))

【★】We have access to a lot more computational power.(现在我们有了更好更快的计算能力。)

【 】Neural Networks are a brand new field.(神经网络是一个全新的领域。)

【★】We have access to a lot more data.(我们现在可以获得更多的数据。)

【 】Deep learning has resulted in significant improvements in important applications such as online advertising, speech recognition, and image recognition.(深度学习已经取得了重大的进展，比如在在线广告、语音识别和图像识别方面有了很多的应用。)

3. Recall this diagram of iterating over different ML ideas. Which of the statements below are true? (Check all that apply.)(回想一下关于不同的机器学习思想的迭代图。下面哪(个/些)陈述是正确的？)
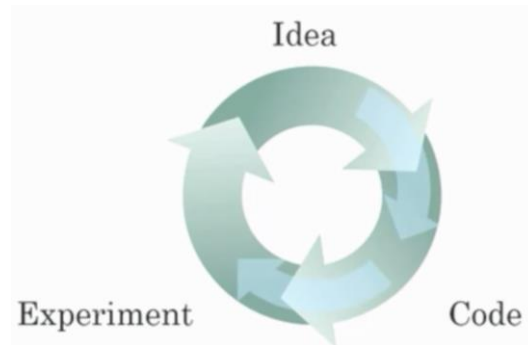


【★】 Being able to try out ideas quickly allows deep learning engineers to iterate more quickly.(能够让深度学习工程师快速地实现自己的想法。)

【★】 Faster computation can help speed up how long a team takes to iterate to a good idea.(在更好更快的计算机上能够帮助一个团队减少迭代(训练)的时间。)

【 】 It is faster to train on a big dataset than a small dataset.(在数据量很多的数据集上训练上的时间要快于小数据集。)

【★】 Recent progress in deep learning algorithms has allowed us to train good models faster (even without changing the CPU/GPU hardware).(使用更新的深度学习算法可以使我们能够更快地训练好模型（即使更换 CPU / GPU 硬件）。)

Note: A bigger dataset generally requires more time to train on a same model.( 请注意: 同一模型在较大的数据集上通常需要花费更多时间。)

4. When an experienced deep learning engineer works 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 multiple times through different models. True/False?( 当一个经验丰富的深度学习工程师在处理一个新的问题的时候，他们通常可以利用先前的经验来在第一次尝试中训练一个表现很好的模型，而不需要通过不同的模型迭代多次从而选择一个较好的模型，这个说法是正确的吗？)
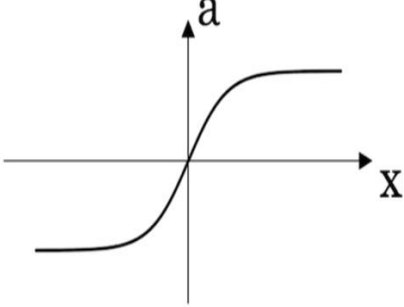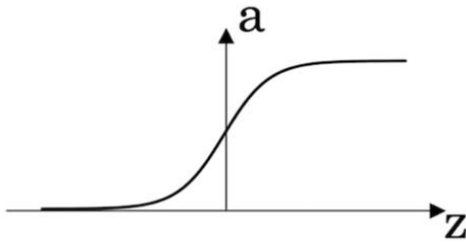
【 】 True(正确)
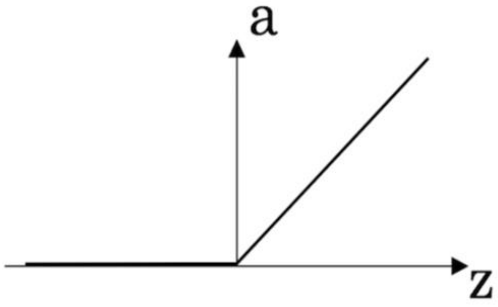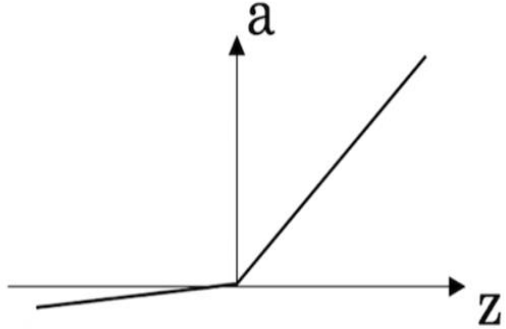
【★】 False(错误)

Note: Maybe some experience may help, but nobody can always find the best model or hyperparameters without iterations.(注：也许之前的一些经验可能会有所帮助，但没有人总是可以找到最佳模型或超参数而无需迭代多次。)

5. Which one of these plots represents a ReLU activation function? (这些图中的哪一个表示 ReLU 激活功能？)

**Answer(回答)：**

| | |
|---|---|
|  【 】Figure1(图一) |  【 】Figure2(图二) |

| | |
|---|---|
|  【★】Figure3(图三) |  【 】Figure4(图四) |

6. Images for cat recognition is an example of "structured" data, because it is represented as a structured array in a computer. True/False?(用于识别猫的图像是"结构化"数据的一个例子，因为它在计算机中被表示为结构化矩阵，是真的吗？)

【 】True(正确)

【★】 False(错误)

7. A demographic dataset with statistics on different cities' population, GDP per capita, economic growth is an example of "unstructured" data because it contains data coming from different sources. True/False? (统计不同城市人口、人均 GDP、经济增长的人口统计数据集是"非结构化"数据的一个例子，因为它包含来自不同来源的数据，是真的吗？)

【 】True(正确)

【★】 False(错误)

8. Why is an RNN (Recurrent Neural Network) used for machine translation, say translating English to French? (Check all that apply.)(为什么在上 RNN（循环神经网络）可以应用机器翻译将英语翻译成法语？)

【★】It can be trained as a supervised learning problem. (因为它可以被用做监督学习。)

【 】It is strictly more powerful than a Convolutional Neural Network (CNN).(严格意义上它比卷积神经网络（CNN）效果更好。)

【★】It is applicable when the input/output is a sequence (e.g., a sequence of words). (它比较适合用于当输入/输出是一个序列的时候（例如：一个单词序列）)

【 】RNNs represent the recurrent process of Idea->Code->Experiment->Idea->….(RNNs 代表递归过程：想法->编码->实验->想法->...)

9. In this diagram which we hand-drew in lecture, what do the horizontal axis (x-axis) and vertical axis (y-axis) represent?(在我们手绘的这张图中，横轴（x 轴）和纵轴（y 轴）代表什么?)
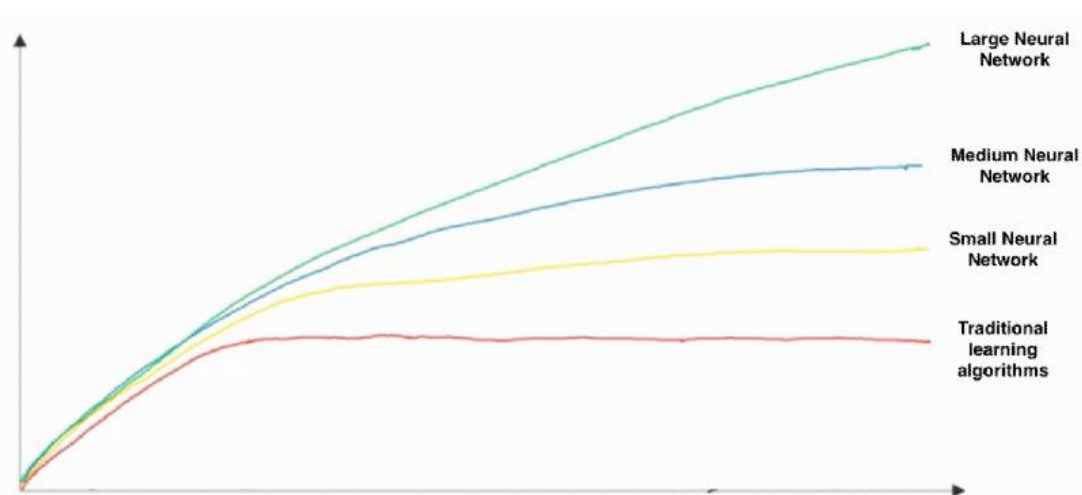


**Answer(回答):**

x-axis is the amount of data(x 轴是数据量)

y-axis (vertical axis) is the performance of the algorithm.(y 轴（垂直轴）是算法的性能)

10. 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.) (假设上一个问题图中描述的是准确的（并且希望您的轴标签正确），以下哪一项是正确的?

【★】Increasing the training set size generally does not hurt an algorithm performance, and it may help significantly. (增加训练集的大小通常不会影响算法的性能，这可能会有很大的帮助。)

【★】 Increasing the size of a neural network generally does not hurt an algorithm performance, and it may help significantly.(增加神经网络的大小通常不会影响算法的性能，这可能会有很大的帮助。)

【 】Decreasing the training set size generally does not hurt an algorithm performance, and it may help significantly.(减小训练集的大小通常不会影响算法的性能，这可能会有很大的帮助。)

【 】Decreasing the size of a neural network generally does not hurt an algorithm performance, and it may help significantly.(减小神经网络的大小通常不会影响算法的性能，这可能会有很大的帮助。)

# Week 2 Quiz - Neural Network Basics（第二周测验 - 神经网络基础）

1. What does a neuron compute?(神经元节点计算什么？)

【 】A neuron computes an activation function followed by a linear function (z = Wx + b)(神经元节点先计算激活函数，再计算线性函数(z = Wx + b))

【★】A neuron computes a linear function (z = Wx + b) followed by an activation function(神经元节点先计算线性函数（z = Wx + b），再计算激活。)

【 】A neuron computes a function g that scales the input x linearly (Wx + b)(神经元节点计算函数 g，函数 g 计算(Wx + b))

【 】A neuron computes the mean of all features before applying the output to an activation function(在 将输出应用于激活函数之前，神经元节点计算所有特征的平均值)

Note: The output of a neuron is a = g(Wx + b) where g is the activation function (sigmoid, tanh, ReLU, ...).(注：神经元的输出是 a = g（Wx + b），其中 g 是激活函数（sigmoid，tanh，ReLU，...）)

2. Which of these is the "Logistic Loss"?(下面哪一个是 Logistic 损失？)

【★】损失函数：$L(\hat{y}^{(i)}, y^{(i)}) = -y^{(i)}\log\hat{y}^{(i)} - (1 - y^{(i)})\log(1 - \hat{y}^{(i)})$

Note: We are using a cross-entropy loss function.(注：我们使用交叉熵损失函数。)

3. 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?(假设 img 是一个（32,32,3）数组，具有 3 个颜色通道：红色、绿色和蓝色的 32x32 像素的图像。 如何将其重新转换为列向量？)

Answer（答）：

```
x = img.reshape((32 * 32 * 3, 1))
```

4. Consider the two following random arrays "a" and "b":(看一下下面的这两个随机数组"a"和"b"：)

```
a = np.random.randn(2, 3) # a.shape = (2, 3)
b = np.random.randn(2, 1) # b.shape = (2, 1)
c = a + b
```

What will be the shape of "c"?(请问数组 c 的维度是多少？)

Answer（答）：

c.shape = (2, 3)

b (column vector) is copied 3 times so that it can be summed to each column of a. Therefore, c.shape = (2, 3).( B（列向量）复制 3 次，以便它可以和 A 的每一列相加，所以：c.shape = (2, 3))

5. Consider the two following random arrays "a" and "b":(看一下下面的这两个随机数组"a"和"b")

```
a = np.random.randn(4, 3) # a.shape = (4, 3)
b = np.random.randn(3, 2) # b.shape = (3, 2)
c = a * b
```

What will be the shape of "c"?(请问数组"c"的维度是多少？)

Answer（答）：

The computation cannot happen because the sizes don't match. It's going to be "error"!

Note:"*" operator indicates element-wise multiplication. Element-wise multiplication requires same dimension between two matrices. It's going to be an error.(注：运算符 "*" 说明了按元素乘法来相乘，但是元素乘法需要两个矩阵之间的维数相同，所以这将报错，无法计算。)

6. Suppose you have $n_x$ input features per example. Recall that $X = [x^{(1)}, x^{(2)} \ldots x^{(m)}]$. What is the dimension of X?(假设你的每一个样本有$n_x$个输入特征，想一下在 $X = [x^{(1)}, x^{(2)} \ldots x^{(m)}]$ 中，X 的维度是多少？)

Answer（答）：

$(n_x, m)$

Note: A stupid way to validate this is use the formula $Z^{(l)} = W^{(l)}A^{(l)}$ when $l = 1$, then we have(请注意：一个比较笨的方法是当$l = 1$的时候，那么计算一下$Z^{(l)} = W^{(l)}A^{(l)}$，所以我们就有：)

$$A^{(1)} = X$$

$$X.shape = (n_x, m)$$

$$Z^{(1)}.shape = (n^{(1)}, m)$$

$$W^{(1)}.shape = (n^{(1)}, n_x)$$

7. Recall that np.dot(a,b) performs a matrix multiplication on a and b, whereas a*b performs an element-wise multiplication.(回想一下，np.dot（a，b）在 a 和 b 上执行矩阵乘法，而"a*b"执行元素方式的乘法。)Consider the two following random arrays "a" and "b":(看一下下面的这两个随机数组"a"和"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?(请问 c 的维度是多少？)

Answer（答）：

c.shape = (12288, 45), this is a simple matrix multiplication example.( c.shape = (12288, 45), 这是一个简单的矩阵乘法例子。)

8. Consider the following code snippet:(看一下下面的这个代码片段：)

```
# 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?(请问要怎么把它们向量化？)

Answer（答）:

```
c = a + b.T
```

9. Consider the following code:(看一下下面的代码：)

```
a = np.random.randn(3, 3)
b = np.random.randn(3, 1)
c = a * b
```

What will be c?(请问 c 的维度会是多少？)

Answer（答）:

```
c.shape = (3, 3)
```

This will invoke broadcasting, so b is copied three times to become (3,3), and * is an element-wise product so c.shape = (3, 3).(这将会使用广播机制，b 会被复制三次，就会变成 (3,3)，再使用元素乘法。所以：c.shape = (3, 3).)

10. Consider the following computation graph,What is the output J.(看一下下面的计算图，J 输出是什么：)



```
J = u + v - w
  = a * b + a * c - (b + c)
```

```
= a * (b + c) - (b + c)
= (a - 1) * (b + c)
```
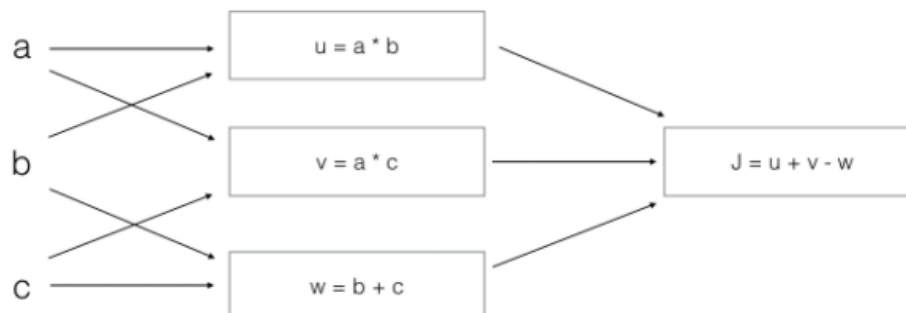
**Answer**（答）:J=(a - 1) * (b + c)

# Week 3 Quiz - Shallow Neural Networks（第三周测验 - 浅层神经网络）

1. Which of the following are true? (Check all that apply.) Notice that I only list correct options(以下哪一项是正确的？只列出了正确的答案)

【★】$X$ is a matrix in which each column is one training example.($X$是一个矩阵，其中每个列都是一个训练样本。)

【★】$a_4^{[2]}$ is the activation output by the 4th neuron of the 2nd layer($a_4^{[2]}$是第二层第四层神经元的激活的输出。)

【★】$a^{[2](12)}$ denotes the activation vector of the 2nd layer for the 12th training example.($a^{[2](12)}$表示第二层和第十二层的激活向量。)

【★】$a^{[2]}$ denotes the activation vector of the 2nd layer.($a^{[2]}$ 表示第二层的激活向量。)

2. The tanh activation usually works better than sigmoid activation function for hidden units because the mean of its output is closer to zero, and so it centers the data better for the next layer. True/False?(tanh 激活函数通常比隐藏层单元的 sigmoid 激活函数效果更好，因为其输出的平均值更接近于零，因此它将数据集中在下一层是更好的选择，请问正确吗？)

【★】True(正确)

【 】 False(错误)

 Note: You can check this post and (this paper)[http://yann.lecun.com/exdb/publis/pdf/lecun-98b.pdf].(请注意，你可以看一下这篇文章 和这篇文档.)

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.(tanh 的输出在-1 和 1 之间，因此它将数据集中在一起，使得下一层的学习变得更加简单。)

3. Which of these is a correct vectorized implementation of forward propagation for layer $l$, where $1 \le l \le L$? Notice that I only list correct options(其中哪一个是第 l 层向前传播的正确向量化实现，其中$1 \le l \le L$)(以下哪一项是正确的？只列出了正确的答案)

【★】$Z^{[l]} = W^{[l]}A^{[l-1]} + b^{[l]}$

【★】$A^{[l]} = g^{[l]}(Z^{[l]})$

4. 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?(您正在构建一个识别黄瓜（y = 1）与西瓜（y = 0）的二元分类器。你会推荐哪一种激活函数用于输出层？)

【 】 ReLU

【 】 Leaky ReLU

【★】 sigmoid

【 】 tanh

Note: The output value from a sigmoid function can be easily understood as a probability.(注意：来自 sigmoid 函数的输出值可以很容易地理解为概率。)

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.(Sigmoid 输出的值介于 0 和 1 之间，这使其成为二元分类的一个非常好的选择。 如果输出小于 0.5，则可以将其归类为 0，如果输出大于 0.5，则归类为 1。 它也可以用 tanh 来完成，但是它不太方便，因为为输出在 -1 和 1 之间。)

5. Consider the following code:(看一下下面的代码：)

```
A = np.random.randn(4,3)
B = np.sum(A, axis = 1, keepdims = True)
```

What will be B.shape?(请问 B.shape 的值是多少?)

```
B.shape = (4, 1)
```

we use (keepdims = True) to make sure that A.shape is (4,1) and not (4, ). It makes our code more rigorous.(我们使用（keepdims = True）来确保 A.shape 是（4,1）而不是（4，），它使我们的代码更加严格。)

6. Suppose you have built a neural network. You decide to initialize the weights and biases to be zero. Which of the following statements are True? (Check all that apply)(假设你已经建立了一个神经网络。 您决定将权重和偏差初始化为零。 以下哪项陈述是正确的？)

【★】 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.(第一个隐藏层中的每个神经元节点将执行相同的计算。 所以即使经过多次梯度下降迭代后，层中的每个神经元节点都会计算出与其他神经元节点相同的东西。)

【 】 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".( 第一个隐藏层中的每个神经元将在第一次迭代中执行相同的计算。 但经过一次梯度下降迭代后，他们将学会计算不同的东西，因为我们已经"破坏了对称性"。)

【 】 Each neuron in the first hidden layer will compute the same thing, but neurons in different layers will compute different things, thus we have accomplished "symmetry breaking" as described in lecture.(第一个隐藏层中的每一个神经元都会计算出相同的东西，但是不同层的神经元会计算不同的东西，因此我们已经完成了"对称破坏"。)

【 】 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.(即使在第一次迭代中，第一个隐藏层的神经元也会执行不同的计算， 他们的参数将以自己的方式不断发展。)

7. Logistic regression's weights w should be initialized randomly rather than to all zeros, because if you initialize to all zeros, then logistic regression will fail to learn a useful decision boundary because it will fail to "break symmetry", True/False?(Logistic 回归的权重 w 应该随机初始化，而不是全零，因为如果初始化为全零，那么逻辑回归将无法学习到有用的决策边界，因为它将无法"破坏对称性"，是正确的吗？)

【 】True(正确)

<span style="color:red">神经网络（深度>=2）：需要随机初始化来打破同一层内神经元之间的对称性。
逻辑回归（单层）：可以将权重初始化为零。虽然随机初始化在实践中可能对收敛速度有微小影响，但全零初始化并不会阻止它学习一个有用的决策边界。对称性问题在这里不适用。</span>

【★】False(错误)

Note: Logistic Regression doesn't have a hidden layer. If you initialize the weights to zeros, the first example x fed in 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.(Logistic 回归没有隐藏层。 如果将权重初始化为零，则 Logistic 回归中的第一个样本 x 将输出零，但 Logistic 回归的导数取决于不是零的输入 x（因为没有隐藏层）。 因此，在第二次迭代中，如果 x 不是常量向量，则权值遵循 x 的分布并且彼此不同。)

8. You have built a network using the tanh activation for all the hidden units. You initialize the weights to relative large values, using np.random.randn(..,..)*1000. What will happen?(您已经为所有隐藏单元使用 tanh 激活建立了一个网络。 使用 np.random.randn（..，..）* 1000 将权重初始化为相对较大的值。 会发生什么？)

【 】It doesn't matter. So long as you initialize the weights randomly gradient descent is not affected by whether the weights are large or small.(这没关系。只要随机初始化权重，梯度下降不受权重大小的影响。)

【 】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 伪 to be very small to prevent divergence; this will slow down learning.(这将导致 tanh 的输入也非常大，因此导致梯度也变大。因此，您必须将 α 设置得非常小以防止发散; 这会减慢学习速度。)

【 】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.(这会导致 tanh 的输入也非常大，导致单位被"高度激活"，从而加快了学习速度，而权重必须从小数值开始。)

【★】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 become slow.(这将导致 tanh 的输入也很大，因此导致梯度接近于零， 优化算法将因此变得缓慢。)

Note:tanh becomes flat for large values, this leads its gradient to be close to zero. This slows down the optimization algorithm.(注：tanh 对于较大的值变得平坦，这导致其梯度接近于零。 这减慢了优化算法。)

9. Consider the following 1 hidden layer neural network:Notice that I only list correct options(看一下下面的单隐层神经网络：只列出了正确的答案)

【★】$b^{[1]}$ will have shape (4, 1)($b^{[1]}$的维度是(4, 1))

【★】$W^{[1]}$ will have shape (4, 2)($W^{[1]}$的维度是 (4, 2))

【★】$W^{[2]}$ will have shape (1, 4)($W^{[2]}$的维度是 (1, 4))

【★】$b^{[2]}$ will have shape (1, 1)($b^{[2]}$的维度是 (1, 1))
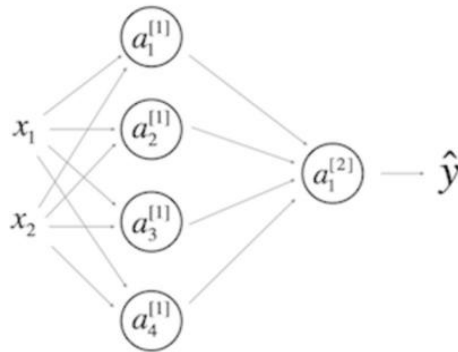


Note: Check here for general formulas to do this.(注:来看一下公式)



10. In the same network as the previous question, what are the dimensions of $z^{[1]}$ and $A^{[1]}$?(I在 和上一个相同的网络中，$z^{[1]}$ 和 $A^{[1]}$的维度是多少？只列出了正确的答案)

【★】$z^{[1]}$ and $A^{[1]}$ are (4,m)($z^{[1]}$和 $A^{[1]}$ 的维度都是 (4,m))

Note: For general formulas to do this.(请注意: 来看一下公式)

# Parameters $W^{[l]}$ and $b^{[l]}$



$L = 5$

$z^{[l]} = g^{[l]}(a^{[l]})$

$a^{[l]}$

$n^{[0]} = n_x = 2$

$n^{[1]} = 3$

$n^{[2]} = 5$

$n^{[3]} = 4$

$n^{[4]} = 2$

$n^{[5]} = 1$

$z^{[1]} = W^{[1]} \cdot x + b^{[1]}$

$(3,1) \leftarrow (3,2) \quad (2,1) \qquad (3,1)$

$(n^{[1]},1) \quad (n^{[1]},n^{[0]}) \quad (n^{[0]},1) \qquad (n^{[1]},1)$

$$\begin{bmatrix} \cdot \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot & \cdot \\ \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ \cdot \end{bmatrix}$$

$W^{[1]} : (n^{[1]}, n^{[0]})$

$W^{[2]} : (5, 3) \quad (n^{[2]}, n^{[1]})$

$z^{[2]} = W^{[2]} \cdot a^{[1]} + b^{[2]}$

$(5,1) \quad (5,3) \quad (3,1) \quad (5,1)$

$(n^{[2]},1)$

$W^{[3]} : (4, 5)$

$W^{[4]} : (2,4) \quad , \quad W^{[5]} : (1,2)$

$W^{[l]} : (n^{[l]}, n^{[l-1]})$

$b^{[l]} : (n^{[l]}, 1)$

$dW^{[l]} : (n^{[l]}, n^{[l-1]})$

$db^{[l]} : (n^{[l]}, 1)$

Andrew Ng

**14**

# Week 4 Quiz - Key concepts on Deep Neural Networks（第四周测验 – 深层神经网络）

1. What is the "cache" used for in our implementation of forward propagation and backward propagation?(在实现前向传播和反向传播中使用的"cache"是什么？)

【 】It is used to cache the intermediate values of the cost function during training.(用于在训练期间缓存成本函数的中间值。)

【★】We use it to pass variables computed during forward propagation to the corresponding backward propagation step. It contains useful values for backward propagation to compute derivatives.(我们用它传递前向传播中计算的变量到相应的反向传播步骤，它包含用于计算导数的反向传播的有用值。)

【 】It is used to keep track of the hyperparameters that we are searching over, to speed up computation.(它用于跟踪我们正在搜索的超参数，以加速计算。)

【 】We use it to pass variables computed during backward propagation to the corresponding forward propagation step. It contains useful values for forward propagation to compute activations.(我们使用它将向后传播计算的变量传递给相应的正向传播步骤，它包含用于计算计算激活的正向传播的有用值。)

Note: the "cache" records values from the forward propagation units and sends it to the backward propagation units because it is needed to compute the chain rule derivatives.(请注意："cache"记录来自正向传播单元的值并将其发送到反向传播单元，因为需要链式计算导数。)

2. Among the following, which ones are "hyperparameters"? (Check all that apply.) I only list correct options.(以下哪些是"超参数"？只列出了正确选项)

【★】size of the hidden layers $n^{[l]}$(隐藏层的大小$n^{[l]}$)

【★】learning rate α(学习率 α)

【★】number of iterations(迭代次数)

【★】number of layers $L$ in the neural network(神经网络中的层数$L$)

Note: You can check this Quora post or this blog post.(请注意：你可以查看 Quora 的这篇文章 或者这篇博客.)

3. Which of the following statements is true?(下列哪个说法是正确的？)

【★】The deeper layers of a neural network are typically computing more complex features of the input than the earlier layers. (神经网络的更深层通常比前面的层计算更复杂的输入特征。)

【 】The earlier layers of a neural network are typically computing more complex features of the input than the deeper layers.(神经网络的前面的层通常比更深层计算更复杂的输入特征。)

Note: You can check the lecture videos. I think Andrew used a CNN example to explain this.(注意：您可以查看视频，我想用吴恩达的用美国有线电视新闻网的例子来解释这个。)

4. 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?(向量化允许您在$L$层神经网络中计算前向传播，而不需要在层(l = 1,2，…，L)上显式的使用 for-loop（或任何其他显式迭代循环），正确吗？)

【　】 True(正确)

【★】 False(错误)

Note: We cannot avoid the for-loop iteration over the computations among layers.(请注意：在层间计算中，我们不能避免 for 循环迭代。)

5. Assume we store the values for $n^{[l]}$ in an array called layers, 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?(假设我们将$n^{[l]}$的值存储在名为 layers 的数组中，如下所示：layer_dims = [$n_x$, 4,3,2,1]。 因此，第 1 层有四个隐藏单元，第 2 层有三个隐藏单元，依此类推。 您可以使用哪个 for 循环初始化模型参数？)

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

6. Consider the following neural network.(下面关于神经网络的说法正确的是：只列出了正确选项)
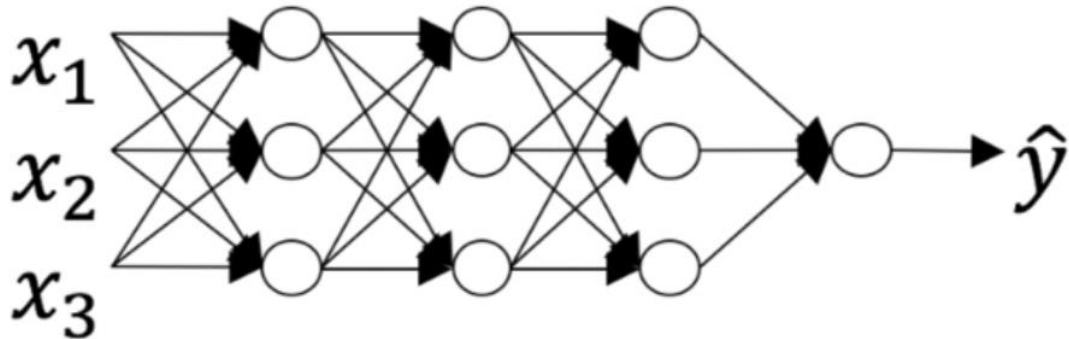


【★】 The number of layers$L$ is 4. The number of hidden layers is 3.(层数$L$为 4，隐藏层数为 3)

Note: The input layer ($L^{[0]}$) does not count.(注意：输入层（$L^{[0]}$）不计数。)

As seen in lecture, the number of layers is counted as the number of hidden layers + 1. The input and output layers are not counted as hidden layers.(正如视频中所看到的那样，层数被计为隐藏层数 +1。输入层和输出层不计为隐藏层。)

7. 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?(在前向传播期间，在层$l$的前向传播函数中，您需要知道层$l$中的激活函数（Sigmoid，tanh，ReLU 等）是什么， 在反向传播期间，相应的反向传播函数也需要知道第$l$层的激活函数是什么，因为梯度是根据它来计算的，正确吗？)

【★】True(正确)

【 】False(错误)

Note: During backpropagation you need to know which activation was used in the forward propagation to be able to compute the correct derivative.(注：在反向传播期间，您需要知道正向传播中使用哪种激活函数才能计算正确的导数。)

8.There are certain functions with the following properties:(有一些函数具有以下属性：)

(i) To compute the function using a shallow network circuit, you will need a large network (where we measure size by the number of logic gates in the network), but (ii) To compute it using a deep network circuit, you need only an exponentially smaller network. True/False?((i) 使用浅网络电路计算函数时，需要一个大网络（我们通过网络中的逻辑门数量来度量大小），但是（ii）使用深网络电路来计算它，只需要一个指数较小的网络。真/假？)

【★】True(正确)

【 】False(错误)

Note: See lectures, exactly same idea was explained.(参见视频，完全相同的题。)

9. Consider the following 2 hidden layer neural network: Which of the following statements are True? (Check all that apply).((在 2 层隐层神经网络中，下列哪个说法是正确的？ 只列出了正确选项))

【★】$W^{[1]}$ will have shape (4, 4)($W^{[1]}$的维度为 (4, 4))

【★】$b^{[1]}$ will have shape (4, 1)($b^{[1]}$的维度为 (4, 1))

【★】$W^{[2]}$ will have shape (3, 4)($W^{[2]}$的维度为 (3, 4))

【★】$b^{[2]}$ will have shape (3, 1)($b^{[2]}$ 的维度为 (3, 1))

【★】$b^{[3]}$ will have shape (1, 1)($b^{[3]}$ 的维度为 (1, 1))

【★】$W^{[3]}$ will have shape (1, 3)($W^{[3]}$ 的维度为 (1, 3))

Note: See [this image] for general formulas.(注：请参阅图片。)
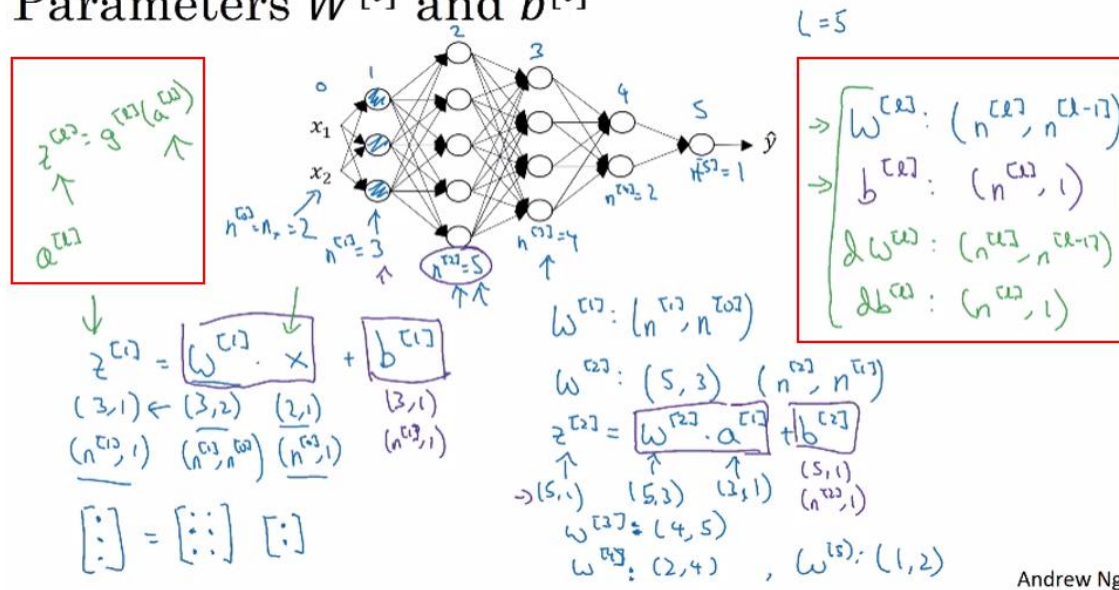


10. Whereas the previous question used a specific network, in the general case what is the dimension of $W^{[1]}$, the weight matrix associated with layer $l$?(前面的问题使用了一个特定的网络，与层 ll 有关的权重矩阵在一般情况下，$W^{[1]}$ 的维数是多少，只列出了正确选项)

【★】$W^{[1]}$ has shape $(n^{[l]}, n^{[l-1]})$($W^{[1]}$的维度是 $(n^{[l-1]}, n^{[l-1]})$)

Note: See this image for general formulas.(注：请参阅图片)
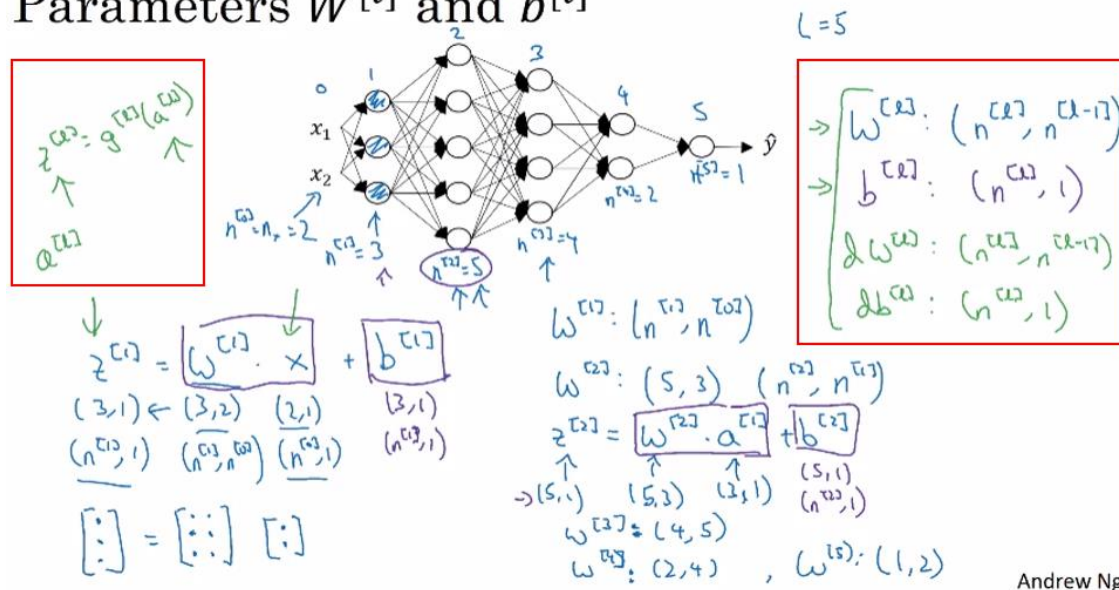
# Lesson2 Improving Deep Neural Networks:Hyperparameter tuning, Regularization and Optimization(第二门课 改善深层神经网络：超参数调试、正则化以及优化)

## Week 1 Quiz - Practical aspects of deep learning（第一周测验 - 深度学习的实践）

1. If you have 10,000,000 examples, how would you split the train/dev/test set? (如果你有 10,000,000 个样本，你会如何划分训练/开发/测试集？)

【★】98% train . 1% dev . 1% test(训练集占 98%， 开发集占 1%， 测试集占 1%)

2. The dev and test set should: (开发和测试集应该)

【★】Come from the same distribution (来自同一分布)

3.If your Neural Network model seems to have high variance, what of the following would be promising things to try? (如果你的神经网络模型似乎有很高的方差，下列哪个尝试是可能解决问题的？)

【★】Add regularization(添加正则化)

【★】Get more training data (获取更多的训练数据)

4. 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.) (你在一家超市的自动结帐亭工作，正在为苹果，香蕉和橘子制作分类器。 假设您的分类器在训练集上有 0.5％的错误，以及开发集上有 7％的错误。 以下哪项尝试是有希望改善你的分类器的分类效果的？)

【★】Increase the regularization parameter lambda (增加正则化参数 lambda)

【★】Get more training data (获取更多的训练数据)

5. What is weight decay? (什么是权重衰减？)

【★】A regularization technique (such as L2 regularization) that results in gradient descent shrinking the weights on every iteration. (正则化技术（例如 L2 正则化）导致梯度下降在每次迭代时权重收缩。)

6. What happens when you increase the regularization hyperparameter lambda? (当你增加正则化超参数 lambda 时会发生什么？)

【★】Weights are pushed toward becoming smaller (closer to 0) (权重会变得更小（接近 0）)

7. With the inverted dropout technique, at test time: (在测试时候使用 dropout)

【★】You do not apply dropout (do not randomly eliminate units) and do not keep the 1/keep_prob factor in the calculations used in training(不要随机消除节点，也不要在训练中使用的计算中保留 1 / keep_prob 因子)

8. Increasing the parameter keep_prob from (say) 0.5 to 0.6 will likely cause the following: (Check the two that apply) (将参数 keep_prob 从（比如说）0.5 增加到 0.6 可能会导致以下情况)

【★】Reducing the regularization effect (正则化作用减弱)

【★】Causing the neural network to end up with a lower training set error(使神经网络在结束时会在训练集上表现好一些。)

9. Which of these techniques are useful for reducing variance (reducing overfitting)? (Check all that apply.) (以下哪些技术可用于减少方差（减少过拟合）)

【★】Dropout

【★】L2 regularization (L2 正则化)

【★】Data augmentation(数据增强)

10. Why do we normalize the inputs x? (为什么我们要归一化输入 x？)

【★】It makes the cost function faster to optimize(它使成本函数更快地进行优化)

# Week 2 Quiz - Optimization algorithms(第二周测验-优化算法)

1. Which notation would you use to denote the 3rd layer's activations when the input is the 7th example from the 8th minibatch?(当输入从第八个 mini-batch 的第七个的样本的时候，你会用哪种符号表示第三层的激活？)

【★】$a^{[3]\{8\}(7)}$

Note: $[i]\{j\}(k)$superscript means $i$-th layer, $j$-th minibatch, $k$-th example(注意：$[i]\{j\}(k)$上标表示 第 $i$层，第$j$小块，第$k$个样本)

2. Which of these statements about mini-batch gradient descent do you agree with?(关于 mini-batch 的说法哪个是正确的？)

【 】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).(在不同的 mini-batch 下，不需要显式地进行循环，就可以实现 mini-batch 梯度下降，从而使算法同时处理所有的数据（向量化）)

【 】 Training one epoch (one pass through the training set) using mini-batch gradient descent is faster than training one epoch using batch gradient descent.(使用 mini-batch 梯度下降训练的时间（一次训练完整个训练集）比使用梯度下降训练的时间要快。)

【★】One iteration of mini-batch gradient descent (computing on a single mini-batch) is faster than one iteration of batch gradient descent.(mini-batch 梯度下降（在单个 mini-batch 上计算）的一次迭代快于梯度下降的迭代。)

Note: Vectorization is not for computing several mini-batches in the same time.(注意：向量化不适用于同时计算多个 mini-batch。)

3. Why is the best mini-batch size usually not 1 and not m, but instead something in-between?(为什么最好的 mini-batch 的大小通常不是 1 也不是 m，而是介于两者之间？)

【★】If the mini-batch size is 1, you lose the benefits of vectorization across examples in the mini-batch.(如果 mini-batch 大小为 1，则会失去 mini-batch 示例中向量化带来的的好处。)

【★】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.(如果 mini-batch 的大小是 m，那么你会得到批量梯度下降，这需要在进行训练之前对整个训练集进行处理)

4. Suppose your learning algorithm's cost J, plotted as a function of the number of iterations, looks like this:(如果你的模型的成本$J$随着迭代次数的增加，绘制出来的图如下，那么： )
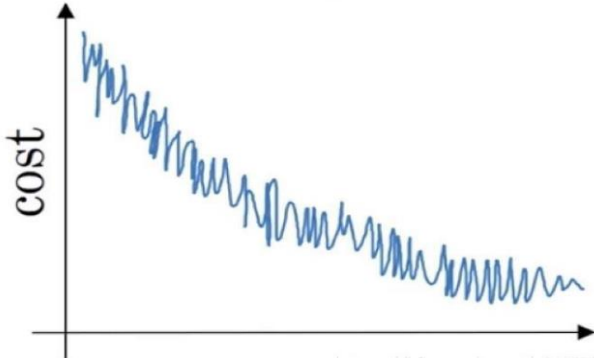


【★】If you're using mini-batch gradient descent, this looks acceptable. But if you're using batch gradient descent, something is wrong.(如果你使用的是 mini-batch 梯度下降，这看起来是可以接受的。但是如果你使用的是批量梯度下降，那么你的模型就有问题。)

Note: There will be some oscillations when you're using mini-batch gradient descent since there could be some noisy data example in batches. However batch gradient descent always guarantees a lower J before reaching the optimal.(注：使用 mini-batch 梯度下降会有一些振荡，因为 mini-batch 中可能会有一些噪音数据。 然而，批量梯度下降总是保证在到达最优值之前达到较低的$J$。)

5. Suppose the temperature in Casablanca over the first three days of January are the same:(假设一月的前三天卡萨布兰卡的气温是一样的： )

Jan 1st: $\theta_1 = 10$(一月第一天: $\theta_1 = 10$)

Jan 2nd: $\theta_2 * 10$(一月第二天: $\theta_2 * 10$)

Say you use an exponentially weighted average with $\beta = 0.5$ to track the temperature: $v_0 = 0$, $v_t = \beta v_t - 1 + (1 - \beta)\theta_t$. If $v_2$ is the value computed after day 2 without bias correction, and $v_2^{corrected}$ is the value you compute with bias correction. What are these values?(假设您使用$\beta = 0.5$的指数加权平均来跟踪温度：$v_0 = 0, v_t = \beta v_t - 1 + (1 - \beta)\theta_t$。 如果$v_2$是在没有偏差修正的情况下计算第 2 天后的值，并且$v_2^{corrected}$是您使用偏差修正计算的值。 这些下面的值是正确的是？)

【★】$v_2 = 7.5, v_2^{corrected} = 10$

6. Which of these is NOT a good learning rate decay scheme? Here, t is the epoch number.(下面哪一个不是比较好的学习率衰减方法？)

【★】$\alpha = e^t * \alpha_0$

Note: This will explode the learning rate rather than decay it.(注：这会使得学习率出现爆炸，而没有衰减。)

7. You use an exponentially weighted average on the London temperature dataset. You use the following to track the temperature: $v_t = \beta v_t - 1 + (1 - \beta)\theta_t$. The red line below was computed using $\beta = 0.9$. What would happen to your red curve as you vary $\beta$? (Check the two that apply)(您在伦敦温度数据集上使用指数加权平均值，您可以使用以下公式来追踪温度：$v_t = \beta v_t - 1 + (1 - \beta)\theta_t$。下面的红线使用的是$\beta = 0.9$来计算的。当你改变$\beta$时，你的红色曲线会怎样变化？)
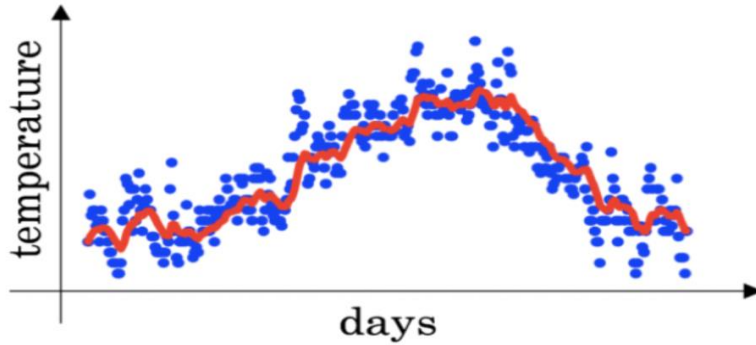


【★】Increasing $\beta$ will shift the red line slightly to the right.(增加$\beta$会使红线稍微向右移动。)

【★】Decreasing $\beta$ will create more oscillation within the red line.(减少$\beta$会在红线内产生更多的振荡。)

8. Consider this figure:(看一下这个图: )These plots were generated with gradient descent; with gradient descent with momentum ($\beta = 0.5$) and gradient descent with momentum ($\beta = 0.9$). Which curve corresponds to which algorithm?(这些图是由梯度下降产生的; 具有动量梯度下降（$\beta = 0.5$）和动量梯度下降（$\beta = 0.9$）。哪条曲线对应哪种算法？)
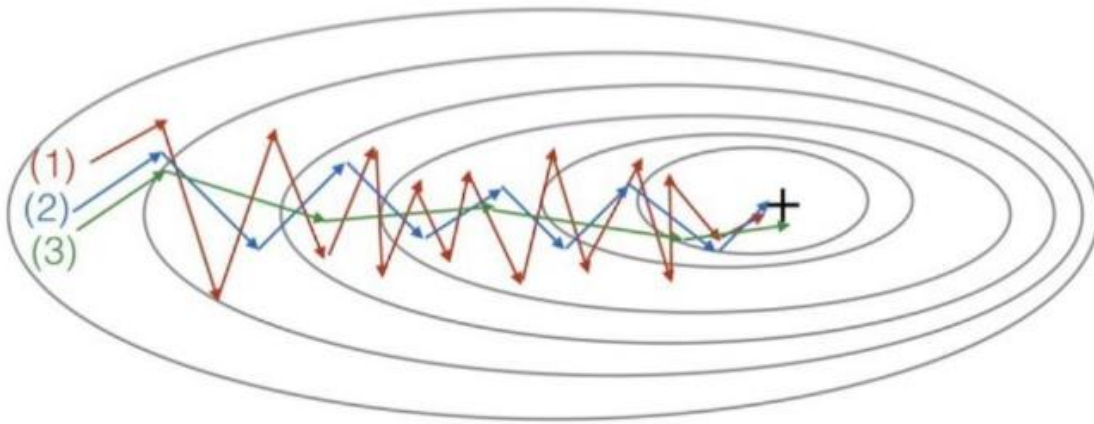


【★】 (1) is gradient descent. (2) is gradient descent with momentum (small $\beta$). (3) is gradient descent with momentum (large $\beta$)((1）是梯度下降。 （2）是动量梯度下降（$\beta$值比较小）。（3）是动量梯度下降（$\beta$比较大）)

9. Suppose batch gradient descent in a deep network is taking excessively long to find a value of the parameters that achieves a small value for the cost function $J(W^{[1]}, b^{[1]}, \ldots, W^{[L]}, b^{[L]})$. Which of the following techniques could help find parameter values that attain a small value for $J$? (Check all that apply)(假设在一个深度学习网络中批处理梯度下降花费了太多的时间来找到一个值的参数值，该值对于成本函数$J(W^{[1]}, b^{[1]}, \ldots, W^{[L]}, b^{[L]})$来说是很小的值。 以下哪些方法可以帮助找到$J$值较小的参数值？)

【★】Try using Adam(尝试使用 Adam 算法)

【★】Try better random initialization for the weights(尝试对权重进行更好的随机初始化)

【★】Try tuning the learning rate $\alpha$(尝试调整学习率$\alpha$)

【★】Try mini-batch gradient descent(尝试 mini-batch 梯度下降)

【 】Try initializing all the weights to zero(尝试把权值初始化为 0)

10. Which of the following statements about Adam is False?(关于 Adam 算法，下列哪一个陈述是错误的？)

【★】Adam should be used with batch gradient computations, not with mini-batches.(Adam 应该用于批梯度计算，而不是用于 mini-batch。)

Note: Adam could be used with both.(注: Adam 可以同时使用。)

# Week 3 Quiz - Hyperparameter tuning, Batch Normalization, Programming Frameworks(第三周测验 - 超参数调整，批量标准化，编程框架)

1. 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? (如果在大量的超参数中搜索最佳的参数值，那么应该尝试在网格中搜索而不是使用随机值，以便更系统的搜索，而不是依靠运气，请问这句话是正确的吗？)

【★】False (错误)

【 】True（正确）

Note: Try random values, don't do grid search. Because you don't know which hyperparamerters are more important than others.And to take an extreme example, let's say that hyperparameter two was that value epsilon that you have in the denominator of the Adam algorithm. So your choice of alpha matters a lot and your choice of epsilon hardly matters. (请注意：应当尝试随机值，不要使用网格搜索，因为你不知道哪些超参数比其他的更重要。举一个很极端的例子，就比如在 Adam 算法中防止除零操作的 ε 的值，一般为 1 的负 8 次方，但是和学习率 α 相比，ε 就显得不那么重要了。)

2. 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? (每个超参数如果设置得不好，都会对训练产生巨大的负面影响，因此所有的超参数都要调整好，请问这是正确的吗？)

【★】False（错误）

【 】True (正确)

Note:We've seen in lecture that some hyperparameters, such as the learning rate, are more critical than others. (注意：我们在视频中讲到的比如学习率这个超参数比其他的超参数更加重要。)

3. 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: (在超参数搜索过程中，你尝试只照顾一个模型（使用熊猫策略）还是一起训练大量的模型（鱼子酱策略）在很大程度上取决于：)

【 】Whether you use batch or mini-batch optimization (是否使用批量（batch）或小批量优化（mini-batch optimization）)

【 】The presence of local minima (and saddle points) in your neural network (神经网络中局部最小值（鞍点）的存在性)

【★】The amount of computational power you can access (在你能力范围内，你能够拥有多大的计算能力)

【 】The number of hyperparameters you have to tune(需要调整的超参数的数量)

4. If you think $\boldsymbol{\beta}$ (hyperparameter for momentum) is between on 0.9 and 0.99, which of the following is the recommended way to sample a value for beta? (如果您认为$\boldsymbol{\beta}$（动量超参数）介于 0.9 和 0.99 之间，那么推荐采用以下哪一种方法来对$\boldsymbol{\beta}$值进行取样？)

【★】

```
r = np.random.rand()

beta = 1 - 10 ** (-r - 1)
```

5. 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 revisit tuning them again. True or false? (找到好的超参数的值是非常耗时的，所以通常情况下你应该在项目开始时做一次，并尝试找到非常好的超参数，这样你就不必再次重新调整它们。请问这正确吗？)

【★】False (错误)

【 】True（正确)

Note: Minor changes in your model could potentially need you to find good hyperparameters again from scratch. (请注意：模型中的细微变化可能导致您需要从头开始重新找到好的超参数。)

6. In batch normalization as presented in the videos, if you apply it on the $l$ th layer of your neural network, what are you normalizing? (在视频中介绍的批量标准化中，如果将其应用于神经网络的第$l$层，那么您怎样进行标准化？)

【★】$z^{[l]}$

7. In the normalization formula $z_{norm}^{(i)} = \frac{z^{(i)}-\mu}{\sqrt{\sigma^2+\varepsilon}}$, why do we use epsilon? (在标准化公式$z_{norm}^{(i)} = \frac{z^{(i)}-\mu}{\sqrt{\sigma^2+\varepsilon}}$中，为什么要使用 epsilon)

【★】To avoid division by zero(为了避免除零操作)

8.Which of the following statements about γ and β in Batch Norm are true? Only correct options listed(Batch Norm 中关于 γ 和 β 的以下哪些陈述是正确的？)

【★】They can be learned using Adam, Gradient descent with momentum, or RMSprop, not just with gradient descent. (它们可以在 Adam、具有动量的梯度下降或 RMSprop 使中用，而不仅仅是用梯度下降来学习。)

【★】They set the mean and variance of the linear variable $z^{[l]}$ of a given layer.( 它们设定给定层的线性变量$z^{[l]}$ 的均值和方差)

9. After training a neural network with Batch Norm, at test time, to evaluate the neural network on a new example you should: (在训练具有 Batch Norm 的神经网络之后，在测试时间，在新样本上评估神经网络，您应该)

【★】Perform the needed normalizations, use $\mu$ and $\sigma^2$ estimated using an exponentially weighted average across mini-batches seen during training. (执行所需的标准化，在训练期间使用了 $\mu$ 和 $\sigma^2$ 的指数加权平均值来估计 mini-batches 的情况。)

10. Which of these statements about deep learning programming frameworks are true? (Check all that apply) (关于深度学习编程框架的这些陈述中，哪一个是正确的？)

【★】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. (通过编程框架，您可以使用比低级语言（如 Python）更少的代码来编写深度学习算法。)

【★】Even if a project is currently open source, good governance of the project helps ensure that the it remains open even in the long term, rather than become closed or modified to benefit only one company. (即使一个项目目前是开源的，项目的良好管理有助于确保它即使在长期内仍然保持开放，而不是仅仅为了一个公司而关闭或修改)

【 】Deep learning programming frameworks require cloud-based machines to run.( 深度学习编程框架的运行需要基于云的机器。 )

# Lesson3 Structuring Machine Learning Projects (第三门课 结构化机器学习项目)

## Week1 Bird recognition in the city of Peacetopia (case study)( 和平之城中的鸟类识别(案例研究))

1.Problem Statement

This example is adapted from a real production application, but with details disguised to protect confidentiality. (问题陈述：这个例子来源于实际项目，但是为了保护机密性，我们会对细节进行保护。)

You are a famous researcher in the City of Peacetopia. The people of Peacetopia have a common characteristic: they are afraid of birds. To save them, you have to build an algorithm that will detect any bird flying over Peacetopia and alert the population.

The City Council gives you a dataset of 10,000,000 images of the sky above Peacetopia, taken from the city's security cameras. They are labelled:

(现在你是和平之城的著名研究员，和平之城的人有一个共同的特点：他们害怕鸟类。为了保护他们，你必须设计一个算法，以检测飞越和平之城的任何鸟类，同时警告人们有鸟类飞过。市议会为你提供了 10,000,000 张图片的数据集，这些都是从城市的安全摄像头拍摄到的。它们被命名为:)

y = 0: There is no bird on the image (图片中没有鸟类)

y = 1: There is a bird on the image (图片中有鸟类)

Your goal is to build an algorithm able to classify new images taken by security cameras from Peacetopia.(你的目标是设计一个算法，能够对和平之城安全摄像头拍摄的新图像进行分类。)

There are a lot of decisions to make:

What is the evaluation metric? How do you structure your data into train/dev/test sets? Metric of success The City Council tells you the following that they want an algorithm that

(有很多决定要做：评估指标是什么？ 你如何将你的数据分割为训练/开发/测试集? 成功的指标 市议会告诉你，他们想要一个算法：)

Has high accuracy Runs quickly and takes only a short time to classify a new image. Can fit in a small amount of memory, so that it can run in a small processor that the city will attach to many different security cameras. Note: 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?( 拥有较高的准确度 快速运行，只需要很短的时间来分类一个新的图像。 可以适应小内存的设备，这样它就可以运行在一个小的处理器上，它将用于城市的安全摄像头上。 请注意: 有三个评估指标使您很难在两种不同的算法之间进行快速选择，并且会降低您的团队迭代的速度，是真的吗？)

【★】True(正确 )

【 】False(错误)

2.After further discussions, the city narrows down 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 10sec to classify a new image." "We want the model to fit in 10MB of memory." (经过进一步讨论，市议会缩小了它的标准："我们需要一种算法，可以让我们尽可能精确的知道一只鸟正飞过和平之城。""我们希望经过训练的模型对新图像进行分类不会超过 10 秒。""我们的模型要适应 10MB 的内存的设备.")

If you had the three following models, which one would you choose?*( 如果你有以下三个模型，你会选择哪一个？)

| 【 】 | Test Accuracy(测试准确度) | Runtime(运行时间) | Size(内存大小) |
|---|---|---|---|
| | 97% | 1 sec(秒) | 3MB |

| 【 】 | Test Accuracy(测试准确度) | Runtime(运行时间) | Size(内存大小) |
|---|---|---|---|
| | 99% | 1 3sec(秒) | 9MB |

| 【 】 | Test Accuracy(测试准确度) | Runtime(运行时间) | Size(内存大小) |
|---|---|---|---|
| | 97% | 3 sec(秒) | 2MB |

| 【★】 | Test Accuracy(测试准确度) | Runtime(运行时间) | Size(内存大小) |
|---|---|---|---|
| | 98% | 9 sec(秒) | 9MB |

3.Based on the city's requests, which of the following would you say is true?( 问题 3 根据城市的要求，您认为以下哪一项是正确的？)

【★】Accuracy is an optimizing metric; running time and memory size are a satisficing metrics.( 准确度是一个优化指标; 运行时间和内存大小是令人满意的指标。)

【 】Accuracy is a satisficing metric; running time and memory size are an optimizing metric. (准确度是一个令人满意的指标; 运行时间和内存大小是一个优化指标。)

【 】Accuracy, running time and memory size are all optimizing metrics because you want to do well on all three. (准确性、运行时间和内存大小都是优化指标，因为您希望在所有这三方面都做得很好。)

【 】 Accuracy, running time and memory size are all satisficing metrics because you have to do sufficiently well on all three for your system to be acceptable. (准确性、运行时间和内存大小都是令人满意的指标，因为您必须在三项方面做得足够好才能使系统可以被接受。)

4.Structuring your data Before implementing your algorithm, you need to split your data into train/dev/test sets. Which of these do you think is the best choice? (结构化你的数据 在实现你的算法之前，你需要将你的数据分割成训练/开发/测试集，你认为哪一个是最好的选择？)

| 【 】 | Train(训练集) | Dev(开发集) | Test(测试集) |
|------|--------------|-------------|--------------|
|      | 6,000,000    | 1,000,000   | 3,000,000    |

| 【 】 | Train(训练集) | Dev(开发集) | Test(测试集) |
|------|--------------|-------------|--------------|
|      | 6,000,000    | 3,000,000   | 6,000,000    |

| 【★】 | Train(训练集) | Dev(开发集) | Test(测试集) |
|------|--------------|-------------|--------------|
|      | 9,500,000    | 250,000     | 250,000      |

| 【 】 | Train(训练集) | Dev(开发集) | Test(测试集) |
|------|--------------|-------------|--------------|
|      | 3,333,334    | 3,333,333   | 3,333,333    |

5.After setting up your train/dev/test sets, the City Council comes across another 1,000,000 images, called the "citizens' data". Apparently the citizens of Peacetopia are so scared of birds that they volunteered to take pictures of the sky and label them, thus contributing these additional 1,000,000 images. These images are different from the distribution of images the City Council had originally given you, but you think it could help your algorithm. (在设置了训练/开发/测试集之后，市议会再次给你了 1,000,000 张图片，称为"公民数据"。 显然，和平之城的公民非常害怕鸟类，他们自愿为天空拍照并贴上标签，从而为这些额外的 1,000,000 张图像贡献力量。 这些图像与市议会最初给您的图像分布不同，但您认为它可以帮助您的算法。)

You should not add the citizens' data to the training set, because this will cause the training and dev/test set distributions to become different, thus hurting dev and test set performance. True/False? (你不应该将公民数据添加到训练集中，因为这会导致训练/开发/测试集分布变得不同，从而损害开发集和测试集性能，是真的吗？)

【 】 True(正确)

【★】 False(错误)

6. One member of the City Council knows a little about machine learning, and thinks you should add the 1,000,000 citizens' data images to the test set. You object because: (市议会的一名成员对机器学习知之甚少，他认为应该将 1,000,000 个公民的数据图像添加到测试集中，你反对的原因是：)

【★】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. (这会导致开发集和测试集分布变得不同。这是一个很糟糕的主意，因为这会达不到你想要的效果。)

【 】The 1,000,000 citizens' data images do not have a consistent x–>y mapping as the rest of the data (similar to the New York City/Detroit housing prices example from lecture). (公民的数据图像与其他数据没有一致的 x- >y 映射(类似于纽约/底特律的住房价格例子)。)

【 】A bigger test set will slow down the speed of iterating because of the computational expense of evaluating models on the test set. (一个更大的测试集将减慢迭代速度，因为测试集上评估模型会有计算开销。)

【★】The test set no longer reflects the distribution of data (security cameras) you most care about. (测试集不再反映您最关心的数据(安全摄像头)的分布。（博主注：训练集是摄像头拍的，用他人拍的数据去测试摄像头拍的，势必会导致准确度下降，要添加也应该添加到整个数据集中，保证同一分布。）)

7.You train a system, and its errors are as follows (error = 100%-Accuracy):

(你训练了一个系统，其误差度如下（误差度 = 100％ - 准确度）)

| Training set error(训练集误差) | 4.0% |
|---|---|
| Dev set error(测试集误差) | 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? (这表明，提高性能的一个很好的途径是训练一个更大的网络，以降低 4%的训练误差。你同意吗？)

【 】Yes, because having 4.0% training error shows you have high bias. (是的，因为有 4%的训练误差表明你有很高的偏差。)

【 】Yes, because this shows your bias is higher than your variance. (是的，因为这表明你的模型的偏差高于方差。)

【 】No, because this shows your variance is higher than your bias. (不同意，因为方差高于偏差。)

【★】No, because there is insufficient information to tell. (不同意，因为没有足够的信息，这什么也说明不了。)

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

（你让一些人对数据集进行标记，以便找出人们对它的识别度。你发现了准确度如下：）

| bird watching expert #1 (鸟类专家 1) | 0.3% Error(误差) |
|---|---|
| bird watching expert #2 (鸟类专家 2) | 0.5% Error(误差) |
| Normal person #1 (not a bird watching expert) (普通人 1) | 1.0% Error(误差) |
| Normal person #2 (not a bird watching expert)(普通人 2) | 1.2% Error(误差) |

If your goal is to have "human-level performance" be a proxy (or estimate) for Bayes error, how would you define "human-level performance"? (如果您的目标是将"人类表现"作为贝叶斯错误的基准线（或估计），那么您如何定义"人类表现"？)

【 】0.0% (because it is impossible to do better than this) (0.0% (因为不可能做得比这更好))

【★】 0.3% (accuracy of expert #1) (0.3% (专家 1 的错误率))

【 】0.4% (average of 0.3 and 0.5) (0.4% (0.3 到 0.5 之间))

【 】0.75% (average of all four numbers above) (0.75% (以上所有四个数字的平均值)

9.Which of the following statements do you agree with? (您同意以下哪个观点？)

【★】 A learning algorithm's performance can be better than human-level performance but it can never be better than Bayes error. (学习算法的性能可以优于人类表现，但它永远不会优于贝叶斯错误的基准线。)

【 】A learning algorithm's performance can never be better than human-level performance but it can be better than Bayes error. (学习算法的性能不可能优于人类表现，但它可以优于贝叶斯错误的基准线。)

【 】A learning algorithm's performance can never be better than human-level performance nor better than Bayes error. (学习算法的性能不可能优于人类表现，也不可能优于贝叶斯错误的基准线。)

【 】A learning algorithm's performance can be better than human-level performance and better than Bayes error. (学习算法的性能可以优于人类表现，也可以优于贝叶斯错误的基准线。)

10.You find that a team of ornithologists debating and discussing an image gets an even better 0.1% performance, so you define that as "human-level performance." After working further on your algorithm, you end up with the following: (你发现一组鸟类学家辩论和讨论图像得到一个更好的 0.1%的性能，所以你将其定义为"人类表现"。在对算法进行深入研究之后，最终得出以下结论：)

| Human-level performance(人类表现) | 0.1% |
|---|---|
| Training set error(训练集误差) | 2.0% |
| Dev set error(开发集误差) | 2.1% |

Based on the evidence you have, which two of the following four options seem the most promising to try? (Check two options.) (根据你的资料，以下四个选项中哪两个尝试起来是最有希望的？（两个选项。）)

【 】 Try increasing regularization. (尝试增加正则化。)

【 】 Get a bigger training set to reduce variance. (获得更大的训练集以减少差异。)

【★】 Try decreasing regularization. (尝试减少正则化。)

【★】 Train a bigger model to try to do better on the training set. (训练一个更大的模型，试图在训练集上做得更好。)

11.You also evaluate your model on the test set, and find the following:( 你在测试集上评估你的模型，并得到以下内容)

| Human-level performance(人类表现) | 0.1% |
|---|---|
| Training set error(训练集误差) | 2.0% |
| Dev set error(开发集误差) | 2.1% |
| Test set error(开发集误差) | 7.0% |

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

【 】 You have underfit to the dev set. (你的开发集欠拟合了。)

【★】 You should try to get a bigger dev set. (你应该尝试获得更大的开发集。)

【 】 You should get a bigger test set. (你应该得到一个更大的测试集。)

【★】 You have overfit to the dev set. (你的开发集过拟合了。)

12.After working on this project for a year, you finally achieve:( 在一年后，你完成了这个项目，你终于实现了：)

| Human-level performance(人类表现) | 0.10% |
|---|---|
| Training set error(训练集误差) | 0.05% |
| Dev set error(开发集误差) | 0.05% |

What can you conclude? (Check all that apply.) (你能得出什么结论？ （检查所有选项。）)

【★】 It is now harder to measure avoidable bias, thus progress will be slower going forward. (现在很难衡量可避免偏差，因此今后的进展将会放缓。)

【 】 This is a statistical anomaly (or must be the result of statistical noise) since it should not be possible to surpass human-level performance. (统计异常(统计噪声的结果)，因为它不可能超过人类表现。)

【　】With only 0.09% further progress to make, you should quickly be able to close the remaining gap to 0%(只有 0.09％的进步空间，你应该很快就能够将剩余的差距缩小到 0％)

【★】If the test set is big enough for the 0.05% error estimate to be accurate, this implies Bayes error is ≤0.05(如果测试集足够大，使得这 0.05%的误差估计是准确的，这意味着贝叶斯误差是小于等于 0.05 的。)

13.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? (事实证明，和平之城也雇佣了你的竞争对手来设计一个系统。您的系统和竞争对手都被提供了相同的运行时间和内存大小的系统，您的系统有更高的准确性。然而，当你和你的竞争对手的系统进行测试时，和平之城实际上更喜欢竞争对手的系统，因为即使你的整体准确率更高，你也会有更多的假阴性结果(当鸟在空中时没有发出警报)。你该怎么办？)

【　】Look at all the models you've developed during the development process and find the one with the lowest false negative error rate. (查看开发过程中开发的所有模型，找出错误率最低的模型。)

【　】Ask your team to take into account both accuracy and false negative rate during development. (要求你的团队在开发过程中同时考虑准确性和假阴性率。)

【★】Rethink the appropriate metric for this task, and ask your team to tune to the new metric. (重新思考此任务的指标，并要求您的团队调整到新指标。)

【　】Pick false negative rate as the new metric, and use this new metric to drive all further development. (选择假阴性率作为新指标，并使用这个新指标来进一步发展。)

14.You've handily beaten your competitor, and your system is now deployed in Peacetopia and is protecting the citizens from birds! But over the last few months, a new species of bird has been slowly migrating into the area, so the performance of your system slowly degrades because your data is being tested on a new type of data. (你轻易击败了你的竞争对手，你的系统现在被部署在和平之城中，并且保护公民免受鸟类攻击！但在过去几个月中，一种新的鸟类已经慢慢迁移到该地区，因此你的系统的性能会逐渐下降，因为您的系统正在测试一种新类型的数据。（博主注：以系统未训练过的鸟类图片来测试系统的性能）)

You have only 1,000 images of the new species of bird. The city expects a better system from you within the next 3 months. Which of these should you do first? (你只有 1000 张新鸟类的图像，在未来的 3 个月里，城市希望你能更新为更好的系统。你应该先做哪一个？)

【★】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. (使用所拥有的数据来定义新的评估指标（使用新的开发/测试集），同时考虑到新物种，并以此来推动团队的进一步发展。)

【 】Put the 1,000 images into the training set so as to try to do better on these birds. (把 1000 张图片放进训练集，以便让系统更好地对这些鸟类进行训练。)

【 】Try data augmentation/data synthesis to get more images of the new type of bird. (尝试数据增强/数据合成，以获得更多的新鸟的图像。)

【 】Add the 1,000 images into your dataset and reshuffle into a new train/dev/test split. (将 1,000 幅图像添加到您的数据集中，并重新组合成一个新的训练/开发/测试集)

15.The City Council thinks that having more Cats in the city would help scare off birds. They are so happy with your work on the Bird detector that they also hire you to build a Cat detector. (Wow Cat detectors are just incredibly useful aren't they.) Because of years of working on Cat detectors, you have such a huge dataset of 100,000,000 cat images that training on this data takes about two weeks. Which of the statements do you agree with? (Check all that agree.)( 问题 15 市议会认为在城市里养更多的猫会有助于吓跑鸟类，他们对你在鸟类探测器上的工作感到非常满意，他们也雇佣你来设计一个猫探测器。（哇~猫探测器是非常有用的，不是吗？）由于有多年的猫探测器的工作经验，你有一个巨大的数据集，你有 100,000,000 猫的图像，训练这个数据需要大约两个星期。你同意哪些说法？（检查所有选项。）)

【★】 Needing two weeks to train will limit the speed at which you can iterate.( 需要两周的时间来训练将会限制你迭代的速度。)

【★】 Buying faster computers could speed up your teams' iteration speed and thus your team's productivity.( 购买速度更快的计算机可以加速团队的迭代速度，从而提高团队的生产力。)

【★】 If 100,000,000 examples is enough to build a good enough Cat detector, you might be better of training with just 10,000,000 examples to gain a ≈10x improvement in how quickly you can run experiments, even if each model performs a bit worse because it's trained on less data.( 如果 100,000,000 个样本就足以建立一个足够好的猫探测器，你最好用 100,000,00 个样本训练，从而使您可以快速运行实验的速度提高约 10 倍，即使每个模型表现差一点因为它的训练数据较少。)

【 】Having built a good Bird detector, you should be able to take the same model and hyperparameters and just apply it to the Cat dataset, so there is no need to iterate.( 建立了一个效果比较好的鸟类检测器后，您应该能够采用相同的模型和超参数，并将其应用于猫数据集，因此无需迭代。)

# Week2 Autonomous driving (case study) (case study)( 自动驾驶（案例研究）)

1. To help you practice strategies for machine learning, in this week we'll present another scenario and ask how you would act. We think this "simulator" of working in a machine learning project will give a task of what leading a machine learning project could be like! (为了帮助你练习机器学习策略，本周我们将介绍另一种场景并询问你将如何做。我们认为这个在机器学习项目中工作的"模拟器"将给出一个引导机器学习项目的任务。)

You are employed by a startup building self-driving cars. You are in charge of detecting 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. As an example, the above image contains a pedestrian crossing sign and red traffic lights.( 你受雇于一家创业的自动驾驶的创业公司。您负责检测图片中的路标（停车标志，行人过路标志，前方施工标志）和交通信号标志（红灯和绿灯），目标是识别哪些对象出现在每个图片中。例如，上面的图片包含一个行人过路标志和红色交通信号灯标志。)
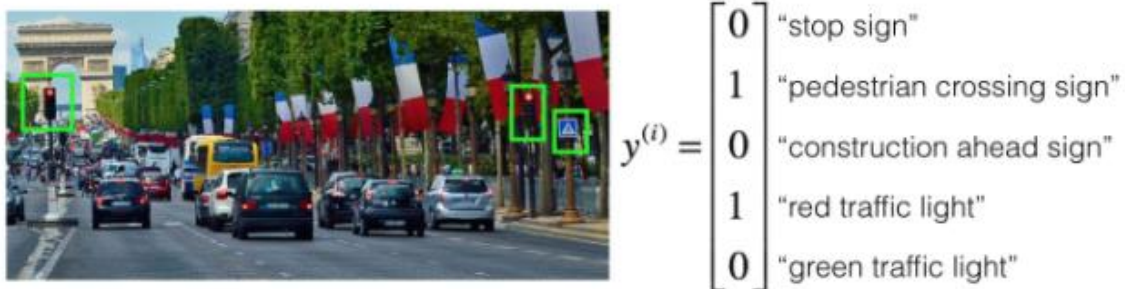


Your 100,000 labeled images are taken using the front-facing camera of your car. This is also the distribution of data you care most about doing well on. You think you might be able to get a much larger dataset off the internet, that could be helpful for training even if the distribution of internet data is not the same.

You are just getting started on 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).

(您的 100,000 张带标签的图片是使用您汽车的前置摄像头拍摄的，这也是你最关心的数据分布，您认为您可以从互联网上获得更大的数据集，即使互联网数据的分布不相同，这也可能对训练有所帮助。你刚刚开始着手这个项目，你做的第一件事是什么？假设下面的每个步骤将花费大约相等的时间（大约几天）。)

【★】Spend a few days training a basic model and see what mistakes it makes. (花几天时间训练一个基本模型，看看它会犯什么错误。)

【  】 Spend a few days checking what is human-level performance for these tasks so that you can get an accurate estimate of Bayes error. (花几天的时间检查这些任务的人类表现，以便能够得到贝叶斯误差的准确估计。)

【 　】 Spend a few days getting the internet data, so that you understand better what data is available. (花几天时间去获取互联网的数据，这样你就能更好地了解哪些数据是可用的。)

【 　】 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. (花几天的时间使用汽车前置摄像头采集更多数据，以更好地了解每单位时间可收集多少数据。)

Note:As seen in the lecture multiple times , Machine Learning is a highly iterative process. We need to create, code, and experiment on a basic model, and then iterate in order to find out the model that works best for the given problem. (注意：正如在视频中多次看到的，机器学习是一个高度迭代的过程。我们需要在基本模型上创建、编码和实验，然后迭代以找出对给定问题最有效的模型。)

2. 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?( 您的目标是检测道路标志（停车标志、行人过路标志、前方施工标志）和交通信号（红灯和绿灯）的图片，目标是识别这些图片中的哪一个标志出现在每个图片中。 您计划在隐藏层中使用带有 ReLU 单位的深层神经网络。 对于输出层，使用 Softmax 激活将是输出层的一个比较好的选择，因为这是一个多任务学习问题，对吗？)

【 　】 True (正确)

【★】 False(错误)

Note: Softmax would have been 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. Since it is not the case , softmax activation cannot be used. (注意：如果每个图片中只有一个可能性：停止标志、减速带、人行横道、红绿灯， 那么 SoftMax 将是一个很好的选择。由于不是这种情况，所以不能使用 Softmax 激活函数。)

3. 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?( 你正在做误差分析并计算错误率，在这些数据集中，你认为你应该手动仔细地检查哪些图片（每张图片都做检查）？)

【 　】 10,000 randomly chosen images (随机选择 10,000 图片)

【★】 500 images on which the algorithm made a mistake (500 张算法分类错误的图片。)

【 　】 10,000 images on which the algorithm made a mistake (10,000 张算法分类错误的图片。)

【 　】 500 randomly chosen images(随机选择 500 图片)

Note:It is of prime importance to look at those images on which the algorithm has made a mistake. Since it is not practical to look at every image the algorithm has made a mistake on, we need to randomly choose 500 such images and analyse the reason for such errors. (注意：查看

算法分类出错的那些图片是非常重要的，由于查看算法分类错误造成的每个图片都不太实际，所以我们需要随机选择 500 个这样的图片并分析出现这种错误的原因。)

4. 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, $y^{(i)} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$ means the image contains a stop sign and a red traffic light. 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 $\begin{bmatrix} 0 \\ ? \\ 1 \\ 1 \\ ? \end{bmatrix}$ then the learning algorithm will not be able to use that example.

True/False?( 每张图片的标签都精确地表示任何的特定路标和交通信号的组合。 例如，$y^{(i)} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$表示图片包含了停车标志和红色交通信号灯。 因为这是一个多任务学习问题，你需要让所有$y^{(i)}$向量被完全标记。 如果一个样本等于$\begin{bmatrix} 0 \\ ? \\ 1 \\ 1 \\ ? \end{bmatrix}$，那么学习算法将无法使用该样本，是正确的吗？)

【  】 True 正确

【★】 False 错误

Note: In the lecture on multi-task learning, you have seen that you can compute the cost even if some entries haven't been labeled. The algorithm won't be influenced by the fact that some entries in the data weren't labeled. (注意：在多任务学习的视频中，您已经看到，即使某些条目没有被标记，您也可以计算成本。该算法不会受到数据中某些条目未标记的样本的影响。)

5.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? (你所关心的数据的分布包含了你汽车的前置摄像头的图片，这与你在网上找到并下载的图片不同。如何将数据集分割为训练/开发/测试集?)

【★】 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. (选择从互联网上的 90 万张图片和汽车前置摄像头的 8 万张图片作为训练集，剩余的 2 万张图片在开发集和测试集中平均分配。)

【   】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. (将 10 万张前摄像头的图片与在网上找到的 90 万张图片随机混合，使得所有数据都随机分布。 将有 100 万张图片的数据集分割为：有 60 万张图片的训练集、有 20 万张图片的开发集和有 20 万张图片的测试集。)

【   】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. (选择从互联网上的 90 万张图片和汽车前置摄像头的 2 万张图片作为训练集，剩余的 8 万张图片在开发集和测试集中平均分配。)

【   】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. (将 10 万张前摄像头的图片与在网上找到的 90 万张图片随机混合，使得所有数据都随机分布。将有 100 万张图片的数据集分割为：有 98 万张图片的训练集、有 1 万张图片的开发集和有 1 万张图片的测试集。)

Note: As seen in lecture, it is important to distribute your data in such a manner that your training and dev set have a distribution that resembles the "real life" data. Also , the test set should contain adeqate amount of "real-life" data you actually care about.( 正如在课堂上看到的那样，分配数据的方式非常重要，您的训练和开发集的分布类似于"现实生活"数据。此外，测试集应包含您实际关心的足够数量的"现实生活"数据。)

6. Assume you've finally chosen the following split between of the data:( 假设您最终选择了以下拆分数据集的方式: )

| Dataset:(数据集) | Contains:(图片数量) | Error of the algorithm:(算法产生的错误) |
| --- | --- | --- |
| 训练集 | 940,000 images randomly picked from (900,000 internet images + 60,000 car's front-facing camera images)(随机抽取 94 万张图片 （从 90 万张互联网图片 + 6 万张汽车前摄像头拍摄的图片中抽取）) | 8.8% |
| 训练-开发集 | 20,000 images randomly picked from (900,000 internet images + 60,000 car's front-facing camera images)(随机抽取 2 万张图片 （从 90 万张互联网图片 + 6 万张汽车前摄像头拍摄的图片中抽取）) | 9.1% |
| 开发集 | 20,000 images from your car's front-facing camera(2 万张汽车前摄像头拍摄的图片) | 14.3% |
| 测试集 | 20,000 images from the car's front-facing camera(2 万张汽车前摄像头拍摄的图片) | 14.8% |

You also know that human-level error on the road sign and traffic signals classification task is around 0.5%. Which of the following are True? (Check all that apply).(您还知道道路标志和交通信号分类的人为错误率大约为 0.5％。以下哪项是真的（检查所有选项）？)

【　】 Your algorithm overfits the dev set because the error of the dev and test sets are very close.(由于开发集和测试集的错误率非常接近，所以你的算法在开发集上过拟合了。)

【★】 You have a large data-mismatch problem because your model does a lot better on the training-dev set than on the dev set(你有一个很大的数据不匹配问题，因为你的模型在训练-开发集上比在开发集上做得好得多。)

【★】 You have a large avoidable-bias problem because your training error is quite a bit higher than the human-level error.(你有一个很大的可避免偏差问题，因为你的训练集上的错误率比人为错误率高很多。)

【　】 You have a large variance problem because your training error is quite higher than the human-level error.(你有很大的方差的问题，因为你的训练集上的错误率比人为错误率要高得多。)

【　】 You have a large variance problem because your model is not generalizing well to data from the same training distribution but that it has never seen before.(你有很大的方差的问题，因为你的模型不能很好地适应来自同一训练集上的分布的数据，即使是它从来没有见过的数据。)

7. Based on table from the previous question, a friend thinks that the training data distribution is much easier than the dev/test distribution. What do you think? (根据上一个问题的表格，一位朋友认为训练数据分布比开发/测试分布要容易得多。你怎么看？)

【　】 Your friend is right. (I.e., Bayes error for the training data distribution is probably lower than for the dev/test distribution.) (你的朋友是对的。（即训练数据分布的贝叶斯误差可能低于开发/测试分布）。)

【　】 Your friend is wrong. (I.e., Bayes error for the training data distribution is probably higher than for the dev/test distribution.) (你的朋友错了。（即训练数据分布的贝叶斯误差可能比开发/测试分布更高）)

【★】 There's insufficient information to tell if your friend is right or wrong. (没有足够的信息来判断你的朋友是对还是错。)

Note: To get an idea of this, we will have to measure human-level error separately on both distributions.The algorithm does better on the distribution data it is trained on. But we do not know for certain that it was because it was trained on that data or if it was really easier than the dev/test distribution. (注：为了了解这一点，我们必须在两个分布上分别测量人类水平误差，该算法对训练过的分布数据有更好的效果。但我们不确定这是因为训练数据分布比开发/测试分布要容易得多。)

8. 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(开发集总误差) | 14.3% |
|---|---|

| Errors due to incorrectly labeled data(由于数据标记不正确而导致的错误) | 4.1% |
|---|---|
| Errors due to foggy pictures(由于雾天的图片引起的错误) | 8.0% |
| Errors due to rain drops stuck on your car's front-facing camera(由于雨滴落在汽车前摄像头上造成的错误) | 2.2% |
| Errors due to other causes(其他原因引起的错误) | 1.0% |

In this table, 4.1%, 8.0%, etc.are a fraction of the total dev set (not just examples your algorithm mislabeled). I.e. about 8.0/14.3 = 56% of your errors are due to foggy pictures. (在这个表格中，4.1％、8.0％这些比例是总开发集的一小部分（不仅仅是您的算法错误标记的样本），即大约8.0 / 14.3 = 56％的错误是由于雾天的图片造成的。)

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? (从这个分析的结果意味着团队最先做的应该是把更多雾天的图片纳入训练集，以便解决该类别中的 8%的错误，对吗？)

【 　 】 True because it is the largest category of errors. As discussed in lecture, we should prioritize the largest category of error to avoid wasting the team's time. (是的，因为它是错误率最大的类别。正如视频中所讨论的，我们应该对错误率进行按大小排序，以避免浪费团队的时间。)

【 　 】 True because it is greater than the other error categories added together (8.0 > 4.1+2.2+1.0). (是的，因为它比其他的错误类别错误率加在一起都大(8.0 > 4.1+2.2+1.0)。)

【★】 False because this would depend on how easy it is to add this data and how much you think your team thinks it'll help. (错误，因为这取决于添加这些数据的容易程度以及您要考虑团队认为它会有多大帮助。)

【 　 】 False because data augmentation (synthesizing foggy images by clean/non-foggy images) is more efficient. (错误，因为数据增强(通过清晰的图像+雾的效果合成雾天的图像)更有效。)

9. You can buy a specially designed windshield wiper that help wipe off some of the raindrops on the front-facing camera. Based on the table from the previous question, which of the following statements do you agree with? (如果合成的图像看起来逼真, 就好像您在有雾的天气中添加了有用的数据来识别道路标志和交通信号一样。)

【★】 2.2% would be a reasonable estimate of the maximum amount this windshield wiper could improve performance. (对于挡风玻璃雨刷可以改善模型的性能而言，2.2％是改善的最大值。)

【 　 】 2.2% would be a reasonable estimate of the minimum amount this windshield wiper could improve performance. (对于挡风玻璃雨刷可以改善模型的性能而言，2.2％是改善最小值。)

【　】2.2% would be a reasonable estimate of how much this windshield wiper will improve performance. (对于挡风玻璃雨刷可以改善模型的性能而言，改善的性能就是 2.2％。)

【　】2.2% would be a reasonable estimate of how much this windshield wiper could worsen performance in the worst case.( 在最坏的情况下，2.2%将是一个合理的估计，因为挡风玻璃刮水器会损坏模型的性能。)

Note:You will probably not improve performance by more than 2.2% by solving the raindrops problem. If your dataset was infinitely big, 2.2% would be a perfect estimate of the improvement you can achieve by purchasing a specially designed windshield wiper that removes the raindrops. (注意：一般而言，解决了雨滴的问题你的错误率可能不会完全降低 2.2%，如果你的数据集是无限大的, 改善 2.2% 将是一个理想的估计, 买一个雨刮是应该可以改善性能的。)

10.You decide to use data augmentation to address foggy images. You find 1,000 pictures of fog off the internet, and "add" them to clean images to synthesize foggy days, like this: (您决定使用数据增强来解决雾天的图像，您可以在互联网上找到 1,000 张雾的照片，然后拿清晰的图片和雾来合成雾天图片，如下所示：)

Which of the following statements do you agree with? (你同意下列哪种说法？ （检查所有选项）)

【★】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. (只要合成的雾对人眼来说看起来逼真，你就可以确信合成的数据和真实的雾天图像差不多，因为人类的视觉对于你正在解决的问题是非常准确的。)

【　】Adding synthesized images that look like real foggy pictures taken from the front-facing camera of your car to training dataset won't help the model improve because it will introduce avoidable-bias. (将合成的看起来像真正的雾天图片添加到从你的汽车前摄像头拍摄的图片的数据集对与改进模型不会有任何帮助，因为它会引入可避免的偏差。)

【　】There is little risk of overfitting to the 1,000 pictures of fog so long as you are combing it with a much larger (>>1,000) of clean/non-foggy images. (只要你把它与一个更大（远大于 1000）的清晰/不模糊的图像结合在一起，那么对雾的 1000 幅图像就没有太大的过拟合的风险。)

Note:Yes. 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 a foggy weather. I will very likely help. (注意：如果合成的图像看起来逼真, 就好像您在有雾的天气中添加了有用的数据来识别道路标志和交通信号一样。)

11.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). (在进一步处理问题之后，您已决定更正开发集上错误标记的数据。 您同意以下哪些陈述？ （检查所有选项）)

【★】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 (您还应该更正测试集中错误标记的数据，以便开发和测试集来自同一分布。)

【　】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. (您应该更正训练集中的错误标记数据, 以免您现在的训练集与开发集更不同。)

【　】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 (您不应该更正测试集中错误标记的数据，以便开发和测试集来自同一分布。)

【★】You should not 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. (您不应更正训练集中的错误标记的数据, 以免现在的训练集与开发集更不同。)

Note: Because you want to make sure that your dev and test data come from the same distribution for your algorithm to make your team's iterative development process is efficient. (注意：因为你想确保你的开发和测试数据来自相同的分布，以使你的团队的迭代开发过程高效。)

12.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. (到目前为止，您的算法仅能识别红色和绿色交通灯，该公司的一位同事开始着手识别黄色交通灯（一些国家称之为橙色光而不是黄色光，我们将使用美国的黄色标准），含有黄色灯的图像非常罕见，而且她没有足够的数据来建立一个好的模型，她希望你能用迁移学习帮助她。)

What do you tell your colleague? (你告诉你的同事怎么做？)

【★】She should try using weights pre-trained on your dataset, and fine-tuning further with the yellow-light dataset. (她应该尝试使用在你的数据集上预先训练过的权重，并用黄灯数据集进行进一步的微调。)

【　】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. (如果她有 10,000 个黄灯图像，从您的数据集中随机抽取 10,000 张图像，并将您和她的数据放在一起，这可以防止您的数据集"淹没"她的黄灯数据集。)

【　】You cannot help her because the distribution of data you have is different from hers, and is also lacking the yellow label. (你没办法帮助她，因为你的数据分布与她的不同，而且缺乏黄灯标签的数据。)

【　】Recommend that she try multi-task learning instead of transfer learning using all the data. (建议她尝试多任务学习，而不是使用所有数据进行迁移学习。)

Note: 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 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. (注：你已经在一个庞大的数据集上训练了你的模型，并且她有一个小数据集。尽管您的标签不同，但您的模型参数已经过

训练，可以识别道路和交通图像的许多特征，这些特征对于她的问题很有用。 这对于转移学习来说是一个完美的例子，她可以从一个与您的架构相同的模型开始，改变最后一个隐藏层之后的内容，并使用您的训练参数对其进行初始化。)

13.Another colleague wants to use microphones placed outside the car to better hear if there's 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? (你已经在一个庞大的数据集上训练了你的模型，并且她有一个小数据集。 尽管您的标签不同，但您的模型参数已经过训练，可以识别道路和交通图像的许多特征，这些特征对于她的问题很有用。 这对于转移学习来说是一个完美的例子，她可以从一个与您的架构相同的模型开始，改变最后一个隐藏层之后的内容，并使用您的训练参数对其进行初始化。 另一位同事想要使用放置在车外的麦克风来更好地听清你周围是否有其他车辆。 例如，如果你身后有警车，你就可以听到警笛声。 但是，他们没有太多的训练这个音频系统，你能帮忙吗？)

【　】Transfer learning from your vision dataset could help your colleague get going faster. Multi-task learning seems significantly less promising. (从视觉数据集迁移学习可以帮助您的同事加快步伐，多任务学习似乎不太有希望。)

【　】Multi-task learning from your vision dataset could help your colleague get going faster. Transfer learning seems significantly less promising. (从您的视觉数据集中进行多任务学习可以帮助您的同事加快步伐，迁移学习似乎不太有希望。)

【　】Either transfer learning or multi-task learning could help our colleague get going faster. (迁移学习或多任务学习可以帮助我们的同事加快步伐。)

【★】Neither transfer learning nor multi-task learning seems promising. (迁移学习和多任务学习都不是很有希望。)

Note: The problem he is trying to solve is quite different from yours. The different dataset structures make it probably impossible to use transfer learning or multi-task learning. (他试图解决的问题与你的问题完全不同，不同的数据集结构可能无法使用迁移学习或多任务学习。)

14.To recognize red and green lights, you have been using this approach: (要识别红灯和绿灯，你使用这种方法：)

(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). (将图像 xx 输入到神经网络，并直接学习映射以预测是红灯(和/或)绿灯(y)。)

A teammate proposes a different, two-step approach: (一个队友提出了另一种两步走的方法：)

(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. (在这个两步法中，您首先要检测图像中的交通灯（如果有），然后确定交通信号灯中照明灯的颜色。)

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. True/False? (在这两者之间，方法 B 更多的是端到端的方法，因为它在输入端和输出端有不同的步骤，这种说法正确吗？)

【　】True (正确)

【★】错误(False)

Note: (A) is an end-to-end approach as it maps directly the input (x) to the output (y). (A 是一种端到端的方法，因为它直接将输入（x）映射到输出（y）。)

15.Approach A (in the question above) tends to be more promising than approach B if you have a __ (fill in the blank). (如果你有一个__,在上面的问题中方法 A 往往比 B 方法更有效。)

【★】Large training set  (大训练集)

【　】Multi-task learning problem. (多任务学习的问题。)

【　】Large bias problem. (偏差比较大的问题)

【　】Problem with a high Bayes error. (高贝叶斯误差的问题。)

Note:In many fields, it has been observed that end-to-end learning works better in practice, but requires a large amount of data. Without a larger amout of data , the application of End-To-End Deep Learning is futile.( 注意：在许多领域，据观察，端到端学习在实践中效果更好，但需要大量数据。如果没有大量的数据，端到端深度学习的应用是效果比较差的。)

# Lesson4 Convolutional Neural Networks（第四门课 卷积神经网络）

## Week 1 quiz - The basics of ConvNets(第一周测验 - 卷积神经网络的基本知识)

1. What do you think applying this filter to a grayscale image will do?( 你认为把下面这个过滤器应用到灰度图像会怎么样？)

$$\begin{bmatrix} 0 & 1 & -1 & 0 \\ 1 & 3 & -3 & -1 \\ 1 & 3 & -3 & -1 \\ 0 & 1 & -1 & 0 \end{bmatrix}$$

【 】 Detect 45 degree edges(会检测 45 度边缘)

【★】Detect vertical edges(会检测垂直边缘)

【 】 Detect horizontal edges(会检测水平边缘)

【 】 Detect image contrast(会检测图像对比度)

(注:因为左边的部分是正的，右边的部分是负的。)

2. Suppose your input is a 300 by 300 color (RGB) image, and you are not using a convolutional network. If the first hidden layer has 100 neurons, each one fully connected to the input, how many parameters does this hidden layer have (including the bias parameters)? (假设你的输入是一个 300×300 的彩色（RGB）图像，而你没有使用卷积神经网络。 如果第一个隐藏层有 100 个神经元，每个神经元与输入层进行全连接，那么这个隐藏层有多少个参数（包括偏置参数）？ )

【 】 9,000,001

【 】 9,000,100

【 】 27,000,001

【★】27,000,100

(注：先计算$W^{[1]} = [l^{[1]}, X] = [100,300 * 300 * 3] = 100 * 300 * 300 * 3 = 27,000,000$，然后计算偏置$b$，因为第一隐藏层有 100 个节点，每个节点有 1 个偏置参数，所以$b = 100$，加起来就是 $27,000,000 + 100 = 27,000,100$。)

3. Suppose your input is a 300 by 300 color (RGB) image, and you use a convolutional layer with 100 filters that are each 5x5. How many parameters does this hidden layer have (including the bias parameters)? (假设你的输入是 300×300 彩色（RGB）图像，并且你使用卷积层和 100 个过滤器，每个过滤器都是 5×5 的大小，请问这个隐藏层有多少个参数（包括偏置参数）？)

【 】2501

【 】2600

【 】7500

【★】7600

(注：首先，参数和输入的图片大小是没有关系的，无论你给的图像像素有多大，参数值都是不变的，在这个题中，参数值只与过滤器有关。我们来看一下怎么算：单个过滤器的大小是 $5*5$，由于输入的是 RGB 图像，所以通道$n_c = 3$，由此可见，一个完整的过滤器的组成是：$5*5*n_c = 5*5*3$，每一个完整的过滤器只有一个偏置参数$b$，所以，每一个完整的过滤器拥有$5*5*3+1 = 76$个参数，而此题中使用了 100 个过滤器，所以这个隐藏层包含了$76*100 = 7600$个参数。)

4. 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? (你有一个 63x63x16 的输入，并使用大小为 7x7 的 32 个过滤器进行卷积，使用步幅为 2 和无填充，请问输出是多少？)

【★】29x29x32

【 】16x16x32

【 】29x29x16

【 】16x16x16

(注：n = 63, f = 7, s = 2, p = 0, filters(过滤层)=32.我们先来看一下这个输出尺寸的公式：$\left\lfloor \frac{n_h+2p-f}{s}+1 \right\rfloor \times \left\lfloor \frac{n_w+2p-f}{s}+1 \right\rfloor$，我们就直接代入公式：$\left\lfloor \frac{63+2\times0-7}{2}+1 \right\rfloor \times \left\lfloor \frac{63+2\times0-7}{2}+1 \right\rfloor = \left\lfloor \frac{56}{2}+1 \right\rfloor \times \left\lfloor \frac{56}{2}+1 \right\rfloor = 29 \times 29$，由于有 32 个过滤器，所以输出为$29 \times 29 \times 32$。)

5. You have an input volume that is 15x15x8, and pad it using "pad=2." What is the dimension of the resulting volume (after padding)? (你有一个 15x15x8 的输入，并使用"pad = 2"进行填充，填充后的尺寸是多少？)

【 】17x17x10

【★】19x19x8

【 】19x19x12

【 】17x17x8

(长和宽左右各填充 2，即，变为(2+15+2)* (2+15+2))

6. You have an input volume that is 63x63x16, and convolve it with 32 filters that are each 7x7, and stride of 1. You want to use a "same" convolution. What is the padding? (你有一个 63x63x16 的输入，有 32 个过滤器进行卷积，每个过滤器的大小为 7x7，步幅为 1，你想要使用"same"的卷积方式，请问 pad 的值是多少？)

【 】 1

【 】 2

【★】 3

【 】 7

(注："same"的卷积方式就是卷积前后的大小不变，也就是 63x63x16 的输入进行卷积后的大小依旧为 63x63x16，这需要我们对输入过来的数据进行填充处理。我们来看一下这个输出尺寸的公式 (假设输入图像的宽、高相同)：：$\left\lfloor \frac{n+2p-f}{s} + 1 \right\rfloor$，由此我们可以推出来$p$的值：$p = \frac{s \times n - n - s + f}{2} = \frac{1 \times 63 - 63 - 1 + 7}{2} = \frac{6}{2} = 3$)

7. 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?( 你有一个 32x32x16 的输入，并使用步幅为 2、过滤器大小为 2 的最大池化，请问输出是多少？)

【 】 15x15x16

【 】 16x16x8

【★】 16x16x16

【 】 32x32x8

8. Because pooling layers do not have parameters, they do not affect the backpropagation (derivatives) calculation. (因为池化层不具有参数，所以它们不影响反向传播的计算。)

【 】 True(正确)

【★】 False(错误)

(注：由卷积层->池化层作为一个 layer，在前向传播过程中，池化层里保存着卷积层的各个部分的最大值/平均值，然后由池化层传递给下一层，在反向传播过程中，由下一层传递梯度过来，"不影响反向传播的计算"这意味着池化层到卷积层（反向）没有梯度变化，梯度值就为 0，既然梯度值为 0，那么例如在$W^{[l]} = W^{[l]} - \alpha \times dW^{[l]}$的过程中，参数$W^{[l]} = W^{[l]} - \alpha \times 0$，也就是说它不再更新，那么反向传播到此中断。所以池化层会影响反向传播的计算。)

9. 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.) (在课程中，我们谈到了"参数共享"是使用卷积网络的好处。关于参数共享的下列哪个陈述是正确的？（检查所有选项。）)

【 】 It reduces the total number of parameters, thus reducing overfitting. (减少了参数的总数，从而减少过拟合。)

【★】It allows a feature detector to be used in multiple locations throughout the whole input image/input volume. (它允许在整个输入值的多个位置使用特征检测器。)

【 】It allows parameters learned for one task to be shared even for a different task (transfer learning). (它允许为一项任务学习的参数即使对于不同的任务也可以共享（迁移学习）。)

【★】It allows gradient descent to set many of the parameters to zero, thus making the connections sparse. (它允许梯度下降将许多参数设置为零，从而使得连接稀疏。)

10. In lecture we talked about "sparsity of connections" as a benefit of using convolutional layers. What does this mean? (在课堂上，我们讨论了"稀疏连接"是使用卷积层的好处。这是什么意思?)

【 】Regularization causes gradient descent to set many of the parameters to zero. (正则化导致梯度下降将许多参数设置为零。)

【 】Each filter is connected to every channel in the previous layer. (每个过滤器都连接到上一层的每个通道。)

【★】Each activation in the next layer depends on only a small number of activations from the previous layer. (下一层中的每个激活只依赖于前一层的少量激活。)

【 】Each layer in a convolutional network is connected only to two other layers. (卷积网络中的每一层只连接到另外两层。)

# Week 2 quiz-Deep convolutional models: case studies) (第二周测验-深度卷积模型：实例探究)

1. Which of the following do you typically see as you move to deeper layers in a ConvNet? (在典型的卷积神经网络中，随着网络的深度增加，你能看到的现象是？)

【 】 $n_H$ and $n_W$ increases, while $n_C$ decreases($n_H$和$n_W$增加，同时$n_C$ 减少。)

【 】 $n_H$ and $n_W$ decreases, while $n_C$ also decreases($n_H$和$n_W$减少，同时$n_C$也减少。)

【 】 $n_H$ and $n_W$ increases, while $n_C$ also increases($n_H$和$n_W$ 增加，同时$n_C$也增加。)

【★】 $n_H$ and $n_W$ decrease, while $n_C$ increases($n_H$和$n_W$减少，同时$n_C$ 增加。)

2. Which of the following do you typically see in a ConvNet? (Check all that apply.) (在典型的卷积神经网络中，你能看到的是？)

【★】 Multiple CONV layers followed by a POOL layer(多个卷积层后面跟着的是一个池化层。)

【 】 Multiple POOL layers followed by a CONV layer(多个池化层后面跟着的是一个卷积层。)

【★】 FC layers in the last few layers(全连接层（FC）位于最后的几层。)

【 】 FC layers in the first few layers(全连接层（FC）位于开始的几层。)

3. In order to be able to build very deep networks, we usually only use pooling layers to downsize the height/width of the activation volumes while convolutions are used with "valid" padding. Otherwise, we would downsize the input of the model too quickly. (为了构建一个非常深的网络，我们经常在卷积层使用"valid"的填充，只使用池化层来缩小激活值的宽/高度，否则的话就会使得输入迅速的变小。)

【 】 True(正确)

【★】 False(错误)

(注：我们经常使用"SAME"的 padding 方式。)

4. 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. (我们使用普通的网络结构来训练一个很深的网络，要使得网络适应一个很复杂的功能（比如增加层数)，总会有更低的训练误差。)

【 】 True(正确)

【★】 False(错误)

(注：在没有残差的普通神经网络中，理论上是误差越来越低的，但是实际上是随着网络层数的加深，先减小再增加；在有残差的 ResNet 中，即使网络再深，训练误差都会随着网络层数的加深逐渐减小。)

5. The following equation captures the computation in a ResNet block. What goes into the two blanks(?) above? (下面计算残差(ResNet)块的公式中，横线上应该分别填什么？)

$$a^{[l+2]} = g(W^{[l+2]}g(W^{[l+1]}a^l + b^{[l+1]}) + b^{[l+2]} + \underline{\quad?\quad}) + \underline{\quad?\quad}$$

【 】 0 and $z^{[l+1]}$, respectively (分别是 0 和 $z^{[l+1]}$)

【★】 $a^{[l]}$ and 0, respectively (分别是 $a^{[l]}$ 和 0)

【 】 $z^{[l]}$ and $a^{[l]}$, respectively (分别是 $z^{[l]}$ 和 $a^{[l]}$)

【 】 0 and $a^{[l]}$, respectively(分别是 0 和 $a^{[l]}$ )

(注：推导：

$$a^{[l+2]} = g(z^{[l+2]} + a^{[l]})$$

$$= g(W^{[l+2]} \times a^{[l+1]} + b^{[l+2]} + a^{[l]})$$

$$= g(W^{[l+2]} \times g(z^{[l+1]}) + b^{[l+2]} + a^{[l]})$$

$$= g(W^{[l+2]} \times g(W^{[l+1]} \times a^{[l]} + b^{[l+1]}) + b^{[l+2]} + \underline{\quad a^{[l]} \quad}) + \underline{\quad 0 \quad}$$

)

6. Which ones of the following statements on Residual Networks are true? (Check all that apply.) (关于残差网络下面哪个（些）说法是正确的？)

【 】 Using a skip-connection helps the gradient to backpropagate and thus helps you to train deeper networks(使用跳越连接能够对反向传播的梯度下降有益且能够帮你对更深的网络进行训练。)

【★】 The skip-connections compute a complex non-linear function of the input to pass to a deeper layer in the network. (跳跃连接计算输入的复杂的非线性函数以传递到网络中的更深层。)

【 】 A ResNet with L layers would have on the order of $L^2$ skip connections in total. (有 L 层的残差网络一共有 $L^2$ 种跳跃连接的顺序。)

【★】 The skip-connection makes it easy for the network to learn an identity mapping between the input and the output within the ResNet block. (跳跃连接能够使得网络轻松地学习残差块类的输入输出间的身份映射。)

7. Suppose you have an input volume of dimension 64x64x16. How many parameters would a single 1x1 convolutional filter have (including the bias)? (假设你的输入的维度为 64x64x16，单个 1x1 的卷积过滤器含有多少个参数（包括偏差）)

【 】2

【 】17

【★】4097

【 】1

(注：64x64x1 + 1 = 4097)

8. Suppose you have an input volume of dimension $n_H \times n_W \times n_C$. Which of the following statements you agree with? (Assume that "1x1 convolutional layer" below always uses a stride of 1 and no padding.) (假设你有一个维度为$n_H \times n_W \times n_C$的卷积输入，下面哪个说法是正确的（假设卷积层为 1x1，步伐为 1，padding 为 0）？)

【 】 You can use a 1x1 convolutional layer to reduce $n_C$ but not $n_H$, $n_W$. (你能够使用 1x1 的卷积层来减少$n_C$，但是不能减少 $n_H$、$n_W$)

【★】 You can use a pooling layer to reduce $n_H$, $n_W$, but not $n_C$. (你可以使用池化层减少 $n_H$、$n_W$，但是不能减少$n_C$)

【★】 You can use a 1x1 convolutional layer to reduce $n_H$, $n_W$, and $n_C$. (可以使用一个 1x1 的卷积层来减少$n_H$、$n_W$和$n_C$.)

【 】 You can use a pooling layer to reduce $n_H$, $n_W$, and $n_C$. (你可以使用池化层减少 $n_H$、$n_W$和$n_C$.)

9. Which ones of the following statements on Inception Networks are true? (Check all that apply.) (关于 Inception 网络下面哪些说法是正确的？)

【 】 Inception networks incorporates a variety of network architectures (similar to dropout, which randomly chooses a network architecture on each step) and thus has a similar regularizing effect as dropout. (Inception 网络包含了各种网络的体系结构（类似于随机删除节点模式，它会在每一步中随机选择网络的结构），因此它具有随机删除节点的正则化效应。)

【★】 Inception blocks usually use 1x1 convolutions to reduce the input data volume's size before applying 3x3 and 5x5 convolutions. (Inception 块通常使用 1x1 的卷积来减少输入卷积的大小，然后再使用 3x3 和 5x5 的卷积。)

【★】 A single inception block allows the network to use a combination of 1x1, 3x3, 5x5 convolutions and pooling. (一个 inception 块允许网络使用 1x1, 3x3, 5x5 的和卷积个池化层的组合。)

【★】 Making an inception network deeper (by stacking more inception blocks together) should not hurt training set performance. (通过叠加 inception 块的方式让 inception 网络更深不会损害训练集的表现。)

10. Which of the following are common reasons for using open-source implementations of ConvNets (both the model and/or weights)? Check all that apply.( 下面哪些是使用卷积网络的开源实现（包含模型/权值）的常见原因？)

【 】 A model trained for one computer vision task can usually be used to perform data augmentation even for a different computer vision task. (为一个计算机视觉任务训练的模型通常可以用来数据扩充，即使对于不同的计算机视觉任务也是如此。)

【★】 Parameters trained for one computer vision task are often useful as pretraining for other computer vision tasks. (为一个计算机视觉任务训练的参数通常对其他计算机视觉任务的预训练是有用的。)

【★】 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. (使用获得计算机视觉竞赛奖项的相同的技术，广泛应用于实际部署。)

【★】 It is a convenient way to get working an implementation of a complex ConvNet architecture. (使用开源实现可以很简单的来实现复杂的卷积结构。)

# Week3 Quiz: Detection algorithms (第三周测验：检测算法)

1. You are building a 3-class object classification and localization algorithm. The classes are: pedestrian (c=1), car (c=2), motorcycle (c=3). What would be the label for the following image? Recall $y = [p_c, b_x, b_y, b_h, b_w, c_1, c_2, c_3]$ (检测算法 现在你要构建一个能够识别三个对象并定位位置的算法，这些对象分别是：行人（c=1），汽车（c=2），摩托车（c=3）。下图中的标签哪个是正确的？注：$y = [p_c, b_x, b_y, b_h, b_w, c_1, c_2, c_3]$)



【★】 y=[1, 0.3, 0.7, 0.3, 0.3, 0, 1, 0]

【 】 y=[1, 0.7, 0.5, 0.3, 0.3, 0, 1, 0]

【 】 y=[1, 0.3, 0.7, 0.5, 0.5, 0, 1, 0]

【 】 y=[1, 0.3, 0.7, 0.5, 0.5, 1, 0, 0]

【 】 y=[0, 0.2, 0.4, 0.5, 0.5, 0, 1, 0]

2. Continuing from the previous problem, 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. As before, $y = [p_c, b_x, b_y, b_h, b_w, c_1, c_2, c_3]$. (继续上一个问题，下图中 y 的值是多少？注："？"是指"不关心这个值"，这意味着神经网络的损失函数不会关心神经网络对输出的结果，和上面一样，$y = [p_c, b_x, b_y, b_h, b_w, c_1, c_2, c_3]$)



【 】 y=[1, ?, ?, ?, ?, 0, 0, 0]

【★】y=[0, ?, ?, ?, ?, ?, ?, ?]

【 】 y=[?, ?, ?, ?, ?, ?, ?, ?]

【 】 y=[0, ?, ?, ?, ?, 0, 0, 0]

【 】 y=[1, ?, ?, ?, ?, ?, ?, ?]

3. You are working on a factory automation task. Your system will see a can of soft-drink coming down a conveyor belt, and you want it to take a picture and decide whether (i) there is a soft-drink can in the image, and if so (ii) its bounding box. Since the soft-drink can is round, the bounding box is always square, and the soft drink can always appears as the same size in the image. There is at most one soft drink can in each image. Here're some typical images in your training set: (你现在任职于自动化工厂中，你的系统会看到一罐饮料从传送带上下来，你想要对其进行拍照，然后确定照片中是否有饮料罐，如果有的话就对其进行包装。饮料罐头是圆的，而包装盒是方的，每一罐饮料的大小是一样的，每个图像中最多只有一罐饮料，现在你有下面的方案可供选择，这里有一些训练集图像：)
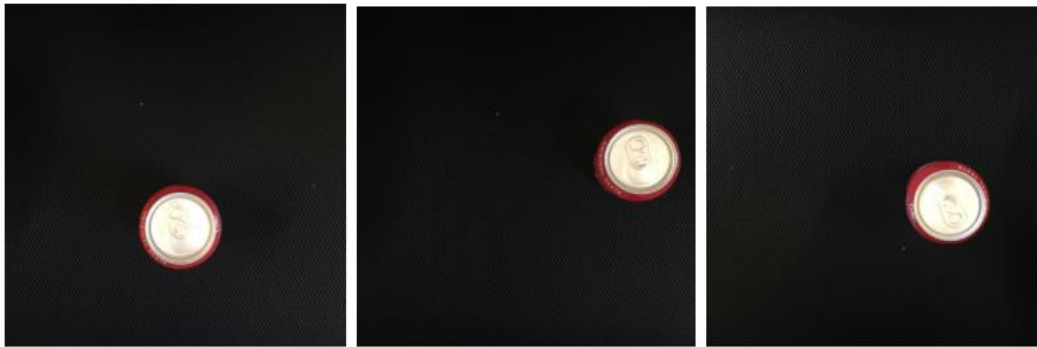


【 】 Logistic unit (for classifying if there is a soft-drink can in the image)( 用于分类图像中是否有罐装饮料)

【★】 Logistic unit,$b_x$ and $b_y$(Logistic 单元,$b_x$ 和 $b_y$)

【 】 Logistic unit,$b_x$,$b_y$,$b_h$ (since $b_w = b_h$)(因为$b_w = b_h$，Logistic 单元,$b_x$,$b_y$,$b_h$)

【 】 Logistic unit,$b_x$,$b_y$,$b_h$,$b_w$(,$b_x$,$b_y$,$b_h$)

(注：因为每个罐装饮料大小是一定的，所以我们只需要知道它的中心位置就好了。)

4.If you build a neural network that inputs a picture of a person's face and outputs $N$ landmarks on the face (assume the input image always contains exactly one face), how many output units will the network have? (如果你想要构建一个能够输入人脸图片输出为$N$个标记的神经网络（假设图像只包含一张脸），那么你的神经网络有多少个输出节点？)

【 】 $N$

【★】$2N$

【 】 $3N$

【 】$N^2$

(注：图像是二维的，指定一个位置应该是(x,y)，那么，一个标记就需要两个节点。)

5.When training one of the object detection systems described in lecture, 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. (当你训练一个视频中描述的对象检测系统时，里面需要一个包含了检测对象的许多图片的训练集，然而边界框不需要在训练集中提供，因为算法可以自己学习检测对象，这个说法对吗？)

【 】 True(正确)

【★】False(错误)

6. Suppose you are applying a sliding windows classifier (non-convolutional implementation). Increasing the stride would tend to increase accuracy, but decrease computational cost. (假如你正在应用一个滑动窗口分类器（非卷积实现），增加步长不仅会提高准确性，也会降低成本。)

【 】 True(正确)

【★】 False(错误)

7. In the YOLO algorithm, at training time, only one cell —the one containing the center/midpoint of an object— is responsible for detecting this object. (在 YOLO 算法训练时候，只有一个包含对象的中心/中点的一个单元负责检测这个对象。)

【★】 True(正确)

【 】 False(错误)

8. What is the IoU between these two boxes? The upper-left box is 2x2, and the lower-right box is 2x3. The overlapping region is 1x1. (这两个框中 IoU 大小是多少？左上角的框是 2x2 大小，右下角的框是 2x3 大小，重叠部分是 1x1。)
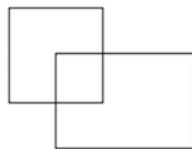


【 】1/6

【★】 1/9

【 】1/10

【 】 None of the above (以上都不是)

(注：(1×1)/(2×2+2×3−1×1)=1/9)

9. Suppose you run non-max suppression on the predicted boxes above. The parameters you use for non-max suppression are that boxes with probability ≤ 0.4 are discarded, and the IoU threshold for deciding if two boxes overlap is 0.5. How many boxes will remain after non-max suppression? (假如你在下图中的预测框中使用非最大值抑制，其参数是放弃概率≤ 0.4 的框，并决定两个框 IoU 的阈值为 0.5，使用非最大值抑制后会保留多少个预测框？)
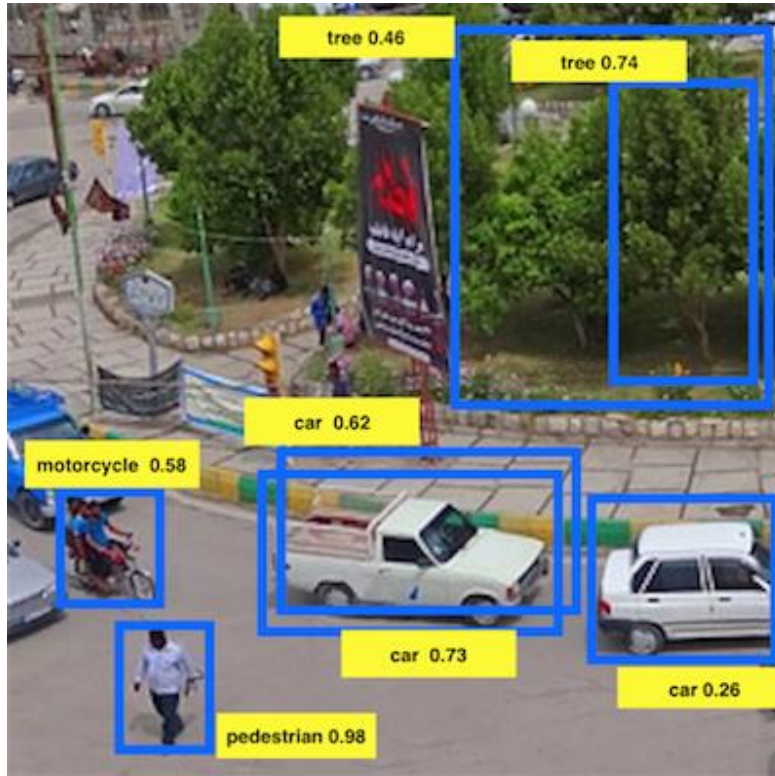


【 】3

【 】4

【★】5

【 】6

【 】7

10.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? (假如你使用 YOLO 算法，使用 19x19 格子来检测 20 个分类，使用 5 个锚框（anchor box）。在训练的过程中，对于每个图像你需要输出卷积后的结果 y 作为神经网络目标值（这是最后一层），y 可能包括一些"？"或者"不关心的值"。请问最后的输出维度是多少？)

【 】 19x19x(25x20)

【 】 19x19x(20x25)

【★】 19x19x(5x25)

【 】 19x19x(5x20)

# Week 4 Quiz: Face recognition & Neural style transfer(第四周测验：面部识别和神经风格转移)

1. Face verification requires comparing a new picture against one person's face, whereas face recognition requires comparing a new picture against K person's faces.( 面部验证只需要将新图片与 1 个人的面部进行比较，而面部识别则需要将新图片与 K 个人的面部进行比较)

【★】True(正确)

【 】False(错误)

2. Why do we learn a function $d(img1, img2)$ for face verification? (Select all that apply.) (在人脸验证中函数 $d(img1, img2)$ 起什么作用？)

【★】We need to solve a one-shot learning problem. (为了解决一次学习的问题。)

【 】Given how few images we have per person, we need to apply transfer learning. (鉴于我们拥有的照片很少，我们需要将它运用到迁移学习中。)

【 】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). (这可以让我们使用 softmax 输出单元来学习预测一个人的身份，在这个单元中分类的数量等于数据库中的人的数量加 1。)

【★】This allows us to learn to recognize a new person given just a single image of that person.( 只需要给出一个人的图片就可以让网络认识这个人。)

(注：第三个选项中，Softmax 输出单元在这里已经被去掉了。第四个选项中，我们不需要使用迁移学习。)

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

【 】True(正确)

【★】False(错误)

(注：每个人需要多张照片的。)

4. Which of the following is a correct definition of the triplet loss? Consider that $\alpha > 0$. (We encourage you to figure out the answer from first principles, rather than just refer to the lecture.) (下面哪个是三元组损失的正确定义（请把 $\alpha > 0$.也考虑进去）)
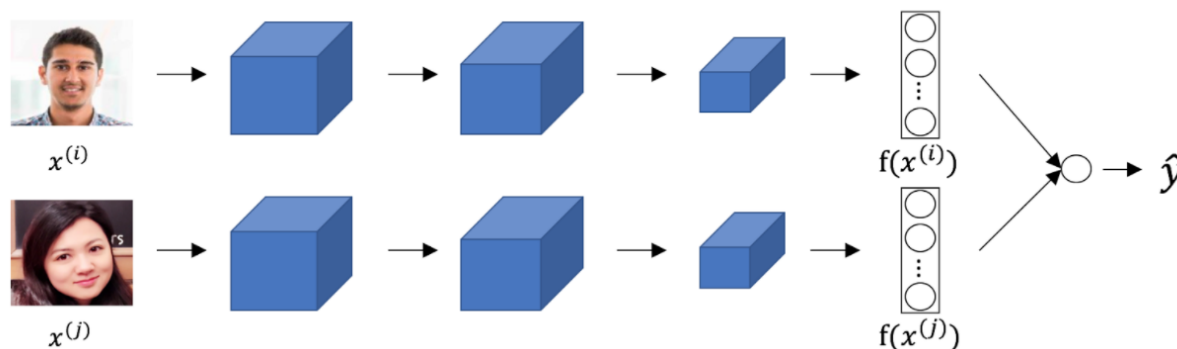
【★】 $max(||f(A)-f(P)||^2 - ||f(A)-f(N)||^2 + \alpha, 0)$

【 】 $max(||f(A)-f(N)||^2 - ||f(A)-f(P)||^2 - \alpha, 0)$

【 】 $max(||f(A)-f(N)||^2 - ||f(A)-f(P)||^2 + \alpha, 0)$

【 】 $max(||f(A)-f(P)||2-||f(A)-f(N)||^2-\alpha,0)$

5. Consider the following Siamese network architecture:



The upper and lower neural networks have different input images, but have exactly the same parameters. ((在下图中的 (Siamese network)结构图中，上下两个神经网络拥有不同的输入图像，但是其中的网络参数是完全相同的。)

【★】 True(正确)

【 】 False(错误)

(注：我们需要相同的参数来获得$f(x^{[i]})$。)

6.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 more likely to find this unit in layer 4 of the network than in layer 1. (你在一个拥有 100 种不同的分类的数据集上训练一个卷积神经网络，你想要知道是否能够找到一个对猫的图片很敏感的隐藏节点（即在能够强烈激活该节点的图像大多数都是猫的图片的节点），你更有可能在第 4 层找到该节点而不是在第 1 层更有可能找到。)

【 】 True(正确)

【★】 False(错误)

7. 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). (神经风格转换被训练为有监督的学习任务，其中的目标是输入两个图像 (x)，并训练一个能够输出一个新的合成图像(y)的网络。)

【 】 True(正确)

【★】 False(错误)

(注：监督学习需要标签，但是这里的图像没有标签。)

8. 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 other. (在一个卷积网络的深层，每个通道对应一个不同的特征检测器，风格矩阵$G^{[l]}$度量了 l 层中不同的特征探测器的激活（或相关）程度。)

【 】 True(正确)

【★】 False(错误)

9. In neural style transfer, what is updated in each iteration of the optimization algorithm? (在神经风格转换中，在优化算法的每次迭代中更新的是什么？)

【★】 The pixel values of the generated image $G$(生成图像 $G$ 的像素值)

【 】 The neural network parameters(神经网络的参数)

【 】 The pixel values of the content image $C$(内容图像 $C$ 的像素值)

【 】 The regularization parameters(正则化参数)

10. 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 3x3x3 (no padding, stride 1). What is the resulting output volume? (你现在用拥有的是 3D 的数据，现在构建一个网络层，其输入的卷积是 32×32×32×16（此卷积有 16 个通道），对其使用 32 个 3×3×3 的过滤器（无填充，步长为 1）进行卷积操作，请问输出的卷积是多少？)

【★】 30x30x30x32

【 】 30x30x30x16

【 】 Undefined: This convolution step is impossible and cannot be performed because the dimensions specified don't match up.( 不能操作，因为指定的维度不匹配，所以这个卷积步骤是不可能执行的。)

# Lesson5 Sequence Models (第五课：序列模型)

## Week 1 Quiz: Recurrent Neural Networks(第一周测验：循环神经网络)

1. Suppose your training examples are sentences (sequences of words). Which of the following refers to the jth word in the ith training example?( 假设你的训练样本是句子(单词序列)，下面哪个选项指的是第$i$个训练样本中的第$j$个词?)
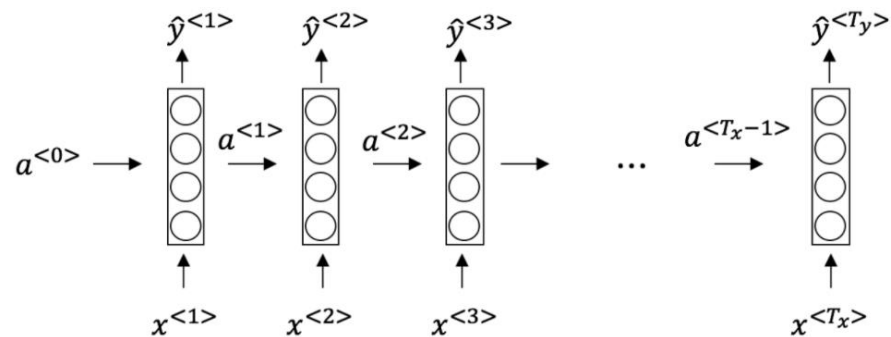
【★】 $x^{(i)<j>}$

【 】 $x^{<i>(j)}$

【 】 $x^{(j)<i>}$

【 】 $x^{<j>(i)}$

(注：首先获取第$i$个训练样本(用括号表示)，然后到$j$列获取单词(用括尖括号表示)。)

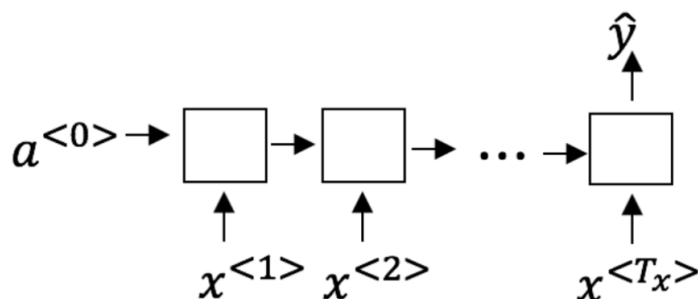2. Consider this RNN: This specific type of architecture is appropriate when: (看一下下面的这个循环神经网络：在下面的条件中，满足上图中的网络结构的参数是)



【★】 $T_x = T_y$

【 】 $T_x < T_y$

【 】 $T_x > T_y$

【 】 $T_x = 1$

3. To which of these tasks would you apply a many-to-one RNN architecture? (Check all that apply). (上图中每一个输入都与输出相匹配。 这些任务中的哪一个会使用多对一的 RNN 体系结构？)



【　】 Speech recognition (input an audio clip and output a transcript) (语音识别（输入语音，输出文本）。)

【★】 Sentiment classification (input a piece of text and output a 0/1 to denote positive or negative sentiment) (情感分类（输入一段文字，输出 0 或 1 表示正面或者负面的情绪）。)

【　】 Image classification (input an image and output a label) (图像分类（输入一张图片，输出对应的标签）。)

【★】 Gender recognition from speech (input an audio clip and output a label indicating the speaker's gender) (人声性别识别（输入语音，输出说话人的性别）。)

4. You are training this RNN language model.
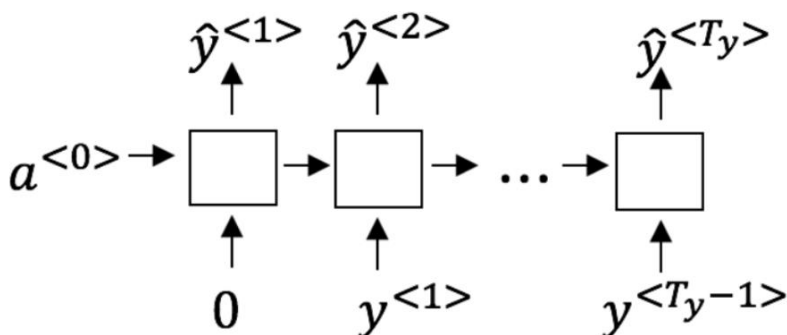


At the $t$th time step, what is the RNN doing? Choose the best answer.( 假设你现在正在训练下面这个 RNN 的语言模型： 在$t$时，这个 RNN 在做什么？)

【　】 Estimating $P(y^{<1>}, y^{<2>}, \ldots, y^{<t-1>})$( 计算 $P(y^{<1>}, y^{<2>}, \ldots, y^{<t-1>})$ )

【　】 Estimating $P(y^{<t>})$ (计算 $P(y^{<t>})$ )

【★】 Estimating $P(y \mid y^{<1>}, y^{<2>}, \ldots, y^{<t>})$ ( 计算 $P(y \mid y^{<1>}, y^{<2>}, \ldots, y^{<t>})$ )

【　】 Estimating $P(y \mid y^{<1>}, y^{<2>}, \ldots, y^{<t-1>})$( 计算 $P(y \mid y^{<1>}, y^{<2>}, \ldots, y^{<t-1>})$)

Yes, in a language model we try to predict the next step based on the knowledge of all prior steps.

5. You have finished training a language model RNN and are using it to sample random sentences, as follows: What are you doing at each time step $t$?( 你已经完成了一个语言模型 RNN 的训练，并用它来对句子进行随机取样，如下图： 在每个时间步$t$都在做什么？)
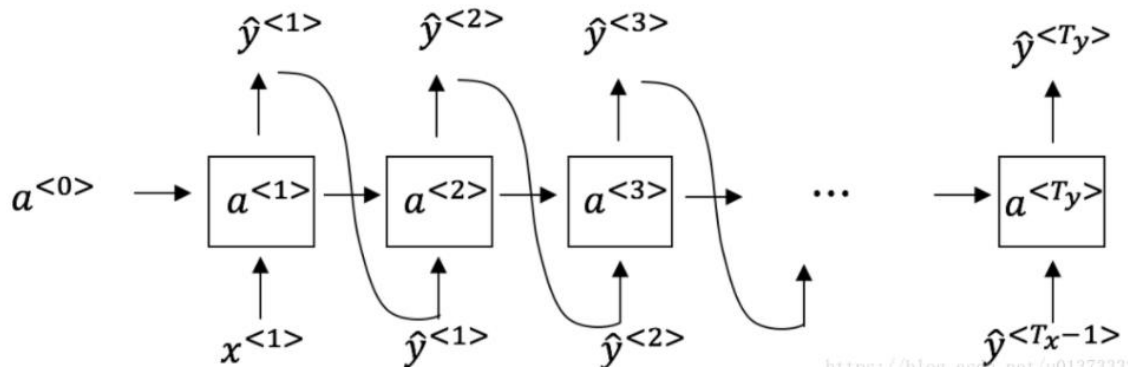


【 】(i) Use the probabilities output by the RNN to pick the highest probability word for that time-step as $y^{<t>}$. (ii) Then pass the ground-truth word from the training set to the next time-step. ((1)使用 RNN 输出的概率，选择该时间步的最高概率单词作为$y^{<t>}$，(2)然后将训练集中的正确的单词传递到下一个时间步。)

【 】(i) Use the probabilities output by the RNN to randomly sample a chosen word for that time-step as $y^{<t>}$. (ii) Then pass the ground-truth word from the training set to the next time-step. ((1)使用由 RNN 输出的概率将该时间步的所选单词进行随机采样作为 $y^{<t>}$，(2)然后将训练集中的实际单词传递到下一个时间步。)

【 】(i) Use the probabilities output by the RNN to pick the highest probability word for that time-step as $y^{<t>}$. (ii) Then pass this selected word to the next time-step. ((1)使用由 RNN 输出的概率来选择该时间步的最高概率词作为 $y^{<t>}$，(2)然后将该选择的词传递给下一个时间步。)

【★】(i) Use the probabilities output by the RNN to randomly sample a chosen word for that time-step as $y^{<t>}$. (ii) Then pass this selected word to the next time-step. ((1)使用 RNN 该时间步输出的概率对单词随机抽样的结果作为$y^{<t>}$，(2)然后将此选定单词传递给下一个时间步。)

6. You are training an RNN, and find that your weights and activations are all taking on the value of NaN ("Not a Number"). Which of these is the most likely cause of this problem?

【 】Vanishing gradient problem. (梯度消失。)

【★】Exploding gradient problem. (梯度爆炸。)

【 】ReLU activation function g(.) used to compute g(z), where z is too large. (**ReLU** 函数作为激活函数$g(.)$，在计算$g(z)$时，$z$的数值过大了。)

【 】Sigmoid activation function g(.) used to compute g(z), where z is too large. (**Sigmoid** 函数作为激活函数$g(.)$，在计算$g(z)$时，$z$的数值过大了。)

7.Suppose you are training a LSTM. You have a 10000 word vocabulary, and are using an LSTM with 100-dimensional activations a. What is the dimension of Γu at each time step?( 假设你正在训练一个 **LSTM** 网络，你有一个 10,000 词的词汇表，并且使用一个激活值维度为 100 的 **LSTM** 块，在每一个时间步中，$\Gamma_u$的维度是多少？)

【 】1

【★】100

【 】300

【 】10000

(注:$\Gamma_u$的向量维度等于 **LSTM** 中隐藏单元的数量。)

8. Here're the update equations for the GRU.

Alice proposes to simplify the GRU by always removing the $\Gamma_u$. I.e., setting $\Gamma_u = 1$. Betty proposes to simplify the GRU by removing the $\Gamma_r$. I. e., setting $\Gamma_r = 1$ always. Which of these models is more likely to work without vanishing gradient problems even when trained on very long input sequences?( 这里有一些 **GRU** 的更新方程： 爱丽丝建议通过移除$\Gamma_u$来简化 **GRU**，即设置$\Gamma_u = 1$。贝蒂提出通过移除$\Gamma_r$来简化 **GRU**，即设置$\Gamma_r = 1$。哪种模型更容易在梯度不消失问题的情况下训练，即使在很长的输入序列上也可以进行训练？)

## GRU

$$\tilde{c}^{<t>} = \tanh(W_c[\Gamma_r * c^{<t-1>}, x^{<t>}] + b_c)$$

$$\Gamma_u = \sigma(W_u[c^{<t-1>}, x^{<t>}] + b_u)$$

$$\Gamma_r = \sigma(W_r[c^{<t-1>}, x^{<t>}] + b_r)$$

$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + (1 - \Gamma_u) * c^{<t-1>}$$

$$a^{<t>} = c^{<t>}$$

【 】 Alice's model (removing $\Gamma_u$), because if $\Gamma_r \approx 0$ for a timestep, the gradient can propagate back through that timestep without much decay. (爱丽丝的模型（即移除 $\Gamma_u$），因为对于一个时间步而言，如果$\Gamma_r \approx 0$，梯度可以通过时间步反向传播而不会衰减。)

【 】 Alice's model (removing $\Gamma_u$), because if $\Gamma_r \approx 1$ for a timestep, the gradient can propagate back through that timestep without much decay. (爱丽丝的模型（即移除 $\Gamma_u$），因为对于一个时间步而言，如果$\Gamma_r \approx 1$，梯度可以通过时间步反向传播而不会衰减。)

【★ 】 Betty's model (removing $\Gamma_r$), because if $\Gamma_r \approx 0$ for a timestep, the gradient can propagate back through that timestep without much decay. (贝蒂的模型（即移除$\Gamma_r$），因为对于一个时间步而言，如果$\Gamma_r \approx 0$，梯度可以通过时间步反向传播而不会衰减。)

【 】 Betty's model (removing $\Gamma_r$), because if $\Gamma_r \approx 1$ for a timestep, the gradient can propagate back through that timestep without much decay. (贝蒂的模型（即移除$\Gamma_r$），因为对于一个时间步而言，如果$\Gamma_r \approx 1$，梯度可以通过时间步反向传播而不会衰减。)

(注：要使信号反向传播而不消失，我们需要$c^{<t>}$高度依赖于$c^{<t-1>}$。)

9. Here are the equations for the GRU and the LSTM:

From these, we can see that the Update Gate and Forget Gate in the LSTM play a role similar to _ and _ in the GRU. What should go in the the blanks?( 这里有一些 **GRU** 和 **LSTM** 的方程: 从这些我们可以看到，在 **LSTM** 中的更新门和遗忘门在 **GRU** 中扮演类似__与_的角色，空白处应该填什么？ )

## GRU

$$\tilde{c}^{<t>} = \tanh(W_c[\Gamma_r * c^{<t-1>}, x^{<t>}] + b_c)$$

$$\Gamma_u = \sigma(W_u[c^{<t-1>}, x^{<t>}] + b_u)$$

$$\Gamma_r = \sigma(W_r[c^{<t-1>}, x^{<t>}] + b_r)$$

$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + (1 - \Gamma_u) * c^{<t-1>}$$

$$a^{<t>} = c^{<t>}$$

## LSTM

$$\tilde{c}^{<t>} = \tanh(W_c[a^{<t-1>}, x^{<t>}] + b_c)$$

$$\Gamma_u = \sigma(W_u[a^{<t-1>}, x^{<t>}] + b_u)$$

$$\Gamma_f = \sigma(W_f[a^{<t-1>}, x^{<t>}] + b_f)$$

$$\Gamma_o = \sigma(W_o[a^{<t-1>}, x^{<t>}] + b_o)$$

$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>}$$

$$a^{<t>} = \Gamma_o * c^{<t>}$$

【★】 $\Gamma_u$ 与 $1 - \Gamma_u$ ($\Gamma_u$ 与 $1 - \Gamma_u$ )

【 】 $\Gamma_u$ 与 $\Gamma_r$ ($\Gamma_u$ 与 $\Gamma_r$ )

【 】 $1 - \Gamma_u$ 与 $\Gamma_u$ ($1 - \Gamma_u$ 与 $\Gamma_u$ )

【 】 $\Gamma_r$ 与 $\Gamma_u$ ($\Gamma_r$ 与 $\Gamma_u$ )

10. You have a pet dog whose mood is heavily dependent on the current and past few days' weather. You've collected data for the past 365 days on the weather, which you represent as a sequence as $x^{<1>}, \ldots, x^{<365>}$. You've also collected data on your dog's mood, which you represent as $y^{<1>}, \ldots, y^{<365>}$. You'd like to build a model to map from x→y. Should you use a Unidirectional RNN or Bidirectional RNN for this problem?( 你有一只宠物狗，它的心情很大程度上取决于当前和过去几天的天气。你已经收集了过去 365 天的天气数据 $x^{<1>}, \ldots, x^{<365>}$，这些数据是一个序列，你还收集了你的狗心情的数据 $y^{<1>}, \ldots, y^{<365>}$，你想建立一个模型来从 $x$ 到 $y$ 进行映射，你应该使用单向 RNN 还是双向 RNN 来解决这个问题？)

【 】Bidirectional RNN, because this allows the prediction of mood on day $t$ to take into account more information. (双向 RNN，因为在 $t$ 日的情绪预测中可以考虑到更多的信息。)

【 】 Bidirectional RNN, because this allows backpropagation to compute more accurate gradients.( 双向 RNN，因为这允许反向传播计算中有更精确的梯度。)

【★】 Unidirectional RNN, because the value of $y^{<t>}$ depends only on $x^{<1>}, \ldots, x^{<t>}$ ,but not on $x^{<t+1>}, \ldots, x^{<365>}$ (单向 RNN，因为 $y^{<t>}$ 的值仅依赖于 $x^{<1>}, \ldots, x^{<t>}$，而不依赖于 $x^{<t+1>}, \ldots, x^{<365>}$。)

【 】Unidirectional RNN, because the value of $y^{<t>}$ depends only on $x^{<t>}$, and not other days' weather.( 单向 RNN，因为 $y^{<t>}$ 的值只取决于 $x^{<t>}$，而不是其他天的天气。)

# Week 2 Quiz: Natural Language Processing and Word Embeddings (第二周测验：自然语言处理与词嵌入)

1.Suppose you learn a word embedding for a vocabulary of 10000 words. Then the embedding vectors should be 10000 dimensional, so as to capture the full range of variation and meaning in those words. (假设你学习了一个为 10000 个单词的词汇表嵌入的单词，那么嵌入向量应该为 10000 维，这样就可以捕捉到所有的变化和意义)

【 】 True(正确)

【★】 False(错误)

Note: The dimension of word vectors is usually smaller than the size of the vocabulary. Most common sizes for word vectors ranges between 50 and 400. (注：词向量的维数通常小于词汇表的维数，词向量最常见的大小在 50 到 400 之间。)

2.What is t-SNE?( t-SNE 是什么？)

【 】 A linear transformation that allows us to solve analogies on word vectors(一种可以让我们解决词向量的相似性的线性变换)

【★】 A non-linear dimensionality reduction technique(一种非线性降维技术)

【 】 A supervised learning algorithm for learning word embeddings(一种学习词嵌入的监督学习算法)

【 】 An open-source sequence modeling library(一个开源序列建模库)

3.Suppose you download a pre-trained word embedding which has been trained on a huge corpus of text. You then use this word embedding to train an RNN for a language task of recognizing if someone is happy from a short snippet of text, using a small training set. (假设你下载了一个经过预先训练的词嵌入模型，该模型是在一个庞大的语料库上训练的出来的。然后使用这个词嵌入来训练一个 RNN 来完成一项语言任务，即使用一个小的训练集从一小段文字中识别出某人是否快乐。)

| x (input text)(输入文本) | y (happy?) (快乐吗？) |
|---|---|
| I'm feeling wonderful today! (我今天感觉很好) | 1 |
| I'm bummed my cat is ill. (我很不高兴我的猫病了) | 0 |
| Really enjoying this! (很享受这个) | 1 |

【★】 True(正确)

【 】 False(错误)

Note: Then even if the word "ecstatic" does not appear in your small training set, your RNN might reasonably be expected to recognize "I'm ecstatic" as deserving a label $y = 1$. (注：即使"欣喜若

狂"这个词没有出现在你的小训练集中，你的 RNN 也会理所当然的认为"我欣喜若狂"应该被贴上"y=1"的标签。)

4.Which of these equations do you think should hold for a good word embedding? (Check all that apply) (你认为以下哪些公式是合适的词嵌入？)

【★】 $e_{boy} - e_{girl} \approx e_{brother} - e_{sister}$

【 】 $e_{boy} - e_{girl} \approx e_{sister} - e_{brother}$

【★】 $e_{boy} - e_{brother} \approx e_{girl} - e_{sister}$

【 】 $e_{boy} - e_{brother} \approx e_{sister} - e_{girl}$

5. Let $E$ be an embedding matrix, and let $o1234$ be a one-hot vector,corresponding to word 1234. Then to get the embedding of word 1234, Why don't we call $E^T * o1234$ in Python?( 设 $E$为嵌入矩阵，$o1234$为独热向量，对应单词 1234，那么为了得到嵌入单词 1234，为什么不在 Python 中调用$E^T * o1234$)

【★】 It is computationally wasteful(这非常耗费计算资源)

Note:Yes, the element-Wise multiplication will be extremely inefficient(注：是的，基于元素乘法效率非常低)

【 】 The correct formula is $E^T * o1234$()

【 】 This doesnt handle unknown words <Unk>(这不能处理未知单词<Unk>)

【 】 None of the above: calling the python snippet as described above is fine(以上都不是，调用上面描述的 Python 代码片段是可以的)

6. When learning word embeddings, we create an artificial task of estimating $P(target|context)$. It is okay if we do poorly on this artificial prediction task; the more important by-product of this task is that we learn a useful set of word embeddings. (在学习词嵌入时，我们创建了一个估算$P(target|context)$的人工任务，如果我们在这个人工预测上做得不好也没关系：这个任务一个重要的副产品就是我们学习了一组有用的词嵌入)

【★】 True(正确)

【 】 False(错误)

7.In the word2vec algorithm, you estimate $P(t|c)$, where $t$ is the target word and $c$ is a context word. How are $t$ and $c$ chosen from the training set? Pick the best answer. (在 word2vec 算法中，估计$P(t|c)$，其中$t$是目标词，$c$是上下文词，如何从训练集中选择$t$和$c$？选择最好的答案)

【★】 $c$ and $t$ are chosen to be nearby words. ($c$和$t$被选为邻近词)

【 】 $c$ is a sequence of several words immediately before $t$.( $c$是紧接在$t$前面的几个词的序列)

【 】 $c$ is the sequence of all the words in the sentence before $t$.( $c$是在$t$之前所有单词的序列)

【 】 $c$ is the one word that comes immediately before $t$. ($c$是在$t$之前出现的一个词)

8.Suppose you have a 10000 word vocabulary, and are learning 500- dimensional word embeddings. The word2vec model uses the following softmax function: Which of these statements are correct? Check all that apply. (假设你有 10000 个词汇表，并且正在学习 500 维的词嵌入， word2vec 模型使用了以下 softmax 函数：这些语句中哪一个是正确的？多选题)

【★】 $\theta_t$ and $e_c$ are both 500 dimensional vectors.( $\theta_t$ 和$e_c$ 都是 500 维向量)

【 】 $\theta_t$ and $e_c$ are both 10000 dimensional vectors.( $\theta_t$ 和$e_c$ 都是 1000 维向量)

【★】 $e_c\theta_t$ and $e_c$ are both trained with an optimization algorithm such as Adam or gradient descent.( $e_c\theta_t$和$e_c$都使用了优化算法的训练，比如 Adam 或者梯度下降)

【 】 After training, we should expect $\theta_t$ to be very close to $e_c$ when $t$ and $c$ are the same word.( 当$t$和$c$是同一个单词，我们期望经过训练 $\theta_t$应该非常接近$e_c$)

9.Suppose you have a 10000 word vocabulary, and are learning 500- dimensional word embeddings. The GloVe model minimizes this objective: Which of these statements are correct? Check all that apply. (假设你有 10000 个词汇表，并且正在学习 500 维的词嵌入，GloVe 模型最小化了这个目标：这些表述中哪一个是正确的？多选题)

【 】 $\theta_i$ and $e_j$ should be initialized to 0 at the beginning of training. ($\theta_i$ 和 $e_j$在训练开始时应该初始化为 0)

【★】 $\theta_i$ and $e_j$ should be initialized randomly at the beginning of training. ($\theta_i$ 和 $e_j$在训练开始时应该随机初始化)

【★】 $X_{ij}$ is the number of times word i appears in the context of word j. ($X_{ij}$是单词 i 在单词 j 上下文中出现的次数)

【★】 The weighting function $f(.)$ must satisfy $f(0) = 0$. Note:The weighting function helps prevent learning only from extremely common word pairs. It is not necessary that it satisfies this function. (权重函数$f(.)$必须满足$f(0) = 0$，注：权重函数有助于防止学习到极端常见的单词对，它不需要满足这个函数)

10.You have trained word embeddings using a text dataset of $m_1$words. You are considering using these word embeddings for a language task, for which you have a separate labeled dataset of $m_2$ words. Keeping in mind that using word embeddings is a form of transfer learning, under which of these circumstance would you expect the word embeddings to be helpful? (你已经使用$m_1$的文本数据集训练了词嵌入。当你有一个单独标记的 $m_2$ 数据集你正在考虑使用这些词嵌入到语言任务中。记住，使用使用词嵌入是一种迁移学习的形式，你希望哪种情况对词嵌入有帮助？)

【★】 $m_1 \gg m_2$

【 】 $m_1 \ll m_2$

# Week 3 Quiz: Sequence models & Attention mechanism (第三周测验：序列模型和注意力机制)

1.Consider using this encoder-decoder model for machine translation. (看下使用这种编码器-解码器模型进行机器翻译)
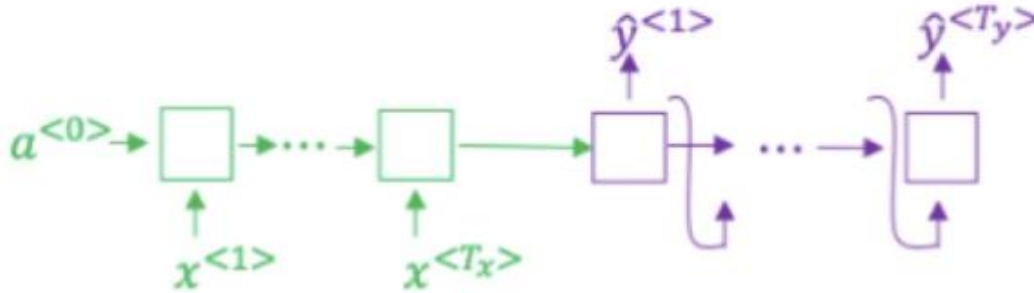


This model is a "conditional language model" in the sense that the encoder portion (shown in green) is modeling the probability of the input sentence $x$. (这个模型是一个条件语言模型，因为编码器部分（用绿色表示）是对输入句子$x$的概率进行建模)

【 】 True(正确)

【★】 False(错误)

2.In beam search, if you increase the beam width $B$, which of the following would you expect to be true? Check all that apply. (在集束搜索中，如果你增加集束宽 $B$，下面哪个是正确的？多选题)

【★】 Beam search will run more slowly. (集束将运行的更慢)

【★】 Beam search will use up more memory. (集束搜索会消耗掉更多内存)

【★】 Beam search will generally find better solutions (i.e. do a better job maximizing $P(y|x)$). (集束搜索会找到更好的解决方案（即更好的最大化 $P(y|x)$）)

【 】 Beam search will converge after fewer steps. (集束搜索经过较少的步数后会收敛)

3.In machine translation, if we carry out beam search without using sentence normalization, the algorithm will tend to output overly short translations. (在机器翻译，如果我们在不使用句子归一化的情况下进行集束搜索，算法会输出过短的翻译)

【★】 True(正确)

【 】 False(错误)

4.Suppose you are building a speech recognition system, which uses an RNN model to map from audio clip $x$ to a text transcript $y$. Your algorithm uses beam search to try to find the value of $y$ that maximizes $P(y|x)$. (假设你正在构建一个语音识别系统，该系统适用 RNN 模型将音频片段$x$映射到文本$y$。你的算法使用集束搜索找到最大化$P(y|x)$的$y$值)

On a dev set example, given an input audio clip, your algorithm outputs the transcript $\hat{y}$ = "I'm building an A Eye system in Silly con Valley", whereas a human gives a much superior transcript $y^*$ = "I'm building an Al system in Silicon Valley." (在一个开发集中，给定一个输入的音频剪辑，$\hat{y}$ = "I'm building an A Eye system in Silly con Valley"，而人类给出的更好的答案：$y^*$ = "I'm building an Al system in Silicon Valley.")

According to your model, $P(\hat{y}|x) = 1.09 * 10^{-7}$     $P(y^*|x) = 7.21 * 10^{-8}$

Would you expect increasing the beam width B to help correct this example? (根据你的模型，$P(\hat{y}|x) = 1.09 * 10^{-7}$     $P(y^*|x) = 7.21 * 10^{-8}$，你希望通过增加集束宽 B 来帮助纠正这个示例吗？)

【★】No, because $P(y^*|x) \leq P(\hat{y}|x)$ indicates the error should be attributed to the RNN rather than to the search algorithm. (没有，因为$P(y^*|x) \leq P(\hat{y}|x)$表示错误应该归咎于 RNN 而不是搜索算法)

【 】No, because $P(y^*|x) \leq P(\hat{y}|x)$ indicates the error should be attributed to the search algorithm rather than to the RNN. (没有，因为$P(y^*|x) \leq P(\hat{y}|x)$表示错误应该归咎于搜索算法而不是 RNN)

【 】Yes, because $P(y^*|x) \leq P(\hat{y}|x)$ indicates the error should be attributed to the RNN rather than to the search algorithm. (是的，因为$P(y^*|x) \leq P(\hat{y}|x)$表示错误应该归咎于 RNN 而不是搜索算法)

【 】Yes, because $P(y^*|x) \leq P(\hat{y}|x)$ indicates the error should be attributed to the search algorithm rather than to the RNN. (是的，因为$P(y^*|x) \leq P(\hat{y}|x)$表示错误应该归咎于搜索算法而不是 RNN)

5. Continuing the example from Q4, suppose you work on your algorithm for a few more weeks, and now find that for the vast majority of examples on which your algorithm makes a mistake,$P(y^*|x) > P(\hat{y}|x)$. This suggest you should focus your attention on improving the search algorithm. (继续 Q4 中的示例。假设你再花几周时间研究你的算法。现在你会发现，在大多数例子中你的算法都会出错，$P(y^*|x) > P(\hat{y}|x)$表示你应该吧注意力集中在改进搜索算法上)

【★】True(正确)

【 】False(错误)

6.Consider the attention model for machine translation. (看下机器翻译的注意力模型。)
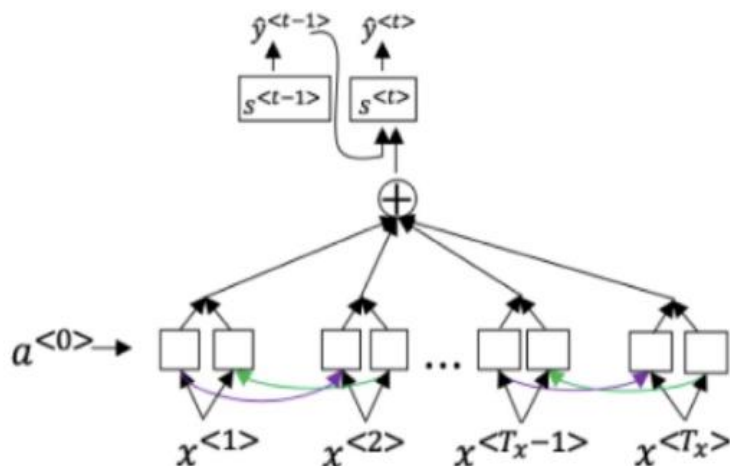


Further, here is the formula for $a^{<t,t'>}$. (接着来看，这是 $a^{<t,t'>}$ 的公式)

$$a^{<t,t'>} = \frac{\exp(e^{<t,t'>})}{\sum_{t'=1}^{Tx} \exp(e^{<t,t'>})}$$

Which of the following statements about $a^{<t,t'>}$ are true? Check all that apply. (以下哪个关于公式 $a^{<t,t'>}$ 的说法是正确的？多选题)

【★】We expect $a^{<t,t'>}$ to be generally larger for values of $a^{<t'>}$ that are highly relevant to the value the network should output for $y^{<t'>}$. (Note the indices in the superscripts.) (我们期望 $a^{<t,t'>}$ 的值一般会更大，因为 $a^{<t'>}$ 的值与 $y^{<t'>}$ 的网络输出值高度相关）（注意上面的索引)

【 】We expect $a^{<t,t'>}$ to be generally larger for values of $a^{<t'>}$ that are highly, relevant to the value the network should output for $y^{<t'>}$. (Note the indices in the superscripts.) (我们期望 $a^{<t,t'>}$ 的值一般更大，与 $y^{<t'>}$ 的网络输出值相关）（注意上面的索引）)

【 】$\sum_t a^{<t,t'>} = 1$ (Note the summation is over $t$.) (注意求和到 $t$)

【★】$\sum_{t'} a^{<t,t'>} = 1$ (Note the summation is over $t'$.）(注意求和到 $t'$)

7.The petwork learns where to "pay attention" by learning the values $e^{<t,t'>}$, which are computed using a small neural network: (petwork 通过学习 $e^{<t,t'>}$ 的值来学习"注意"的位置，该值是通过一个小型神经网络计算出来的)

We can't replace $s^{<t-1>}$ with $s^{<t>}$ as an input to this neural network. This is because $s^{<t>}$ depends on $a^{<t,t'>}$ which in turn depends on $e^{<t,t'>}$; so at the time we need to evalute this network, we haven't computed $s^{<t>}$ yet. (我们不能用 $s^{<t>}$ 替代 $s^{<t-1>}$ 作为神经网络的输入，这是因为 $s^{<t>}$ 依赖于 $a^{<t,t'>}$，所以在我们需要评估这个网络时，我们还没有计算 $s^{<t>}$)

【★】True(正确)

【 】False(错误)

8.Compared to the encoder-decoder model shown in Question 1 of this quiz (which does not use an attention mechanism), we expect the attention model to have the greatest advantage when: (与本周测验问题 1 所示的编译器—译码器模型（不使用注意机制）相比，我们期望注意力模型在以下情况下具有最大的优势)

【★】The input sequence length $T_x$ is large. (输入序列$T_x$长度较大)

【 】The input sequence length $T_x$ is small. (输入序列$T_x$长度较小)

9. Under the CTC model, identical repeated characters not separated by the "blank" character (_) are collapsed. Under the CTC model, what does the following string collapse to? (在 CTC 模型下，未被空白字符 (_) 分隔的相同重复字符被折叠，在 CTC 模型下，下面的字符串折叠到什么位置)

__*c*oo_*o*_kk__**b_ooooo_**oo_kkk

【 】cokbok

【★】cookbook

【 】cook book

【 】coookkboooooookkk

10. In trigger word detection, $x^{<t>}$ is: (在触发字检测中，, $x^{<t>}$为：)

【★】Features of the audio (such as spectrogram features) at time $t$. ($t$时刻的音频特征（如光谱特征）)

【 】The $t$-th input word, represented as either a one-hot vector or a word embedding. (第$t$个输入词，表示为一个独热向量或一个词嵌入)

【 】Whether the trigger word is being said at time $t$. (是否在$t$时刻触发这个词)

【 】Whether someone has just finished saying the trigger word at time $t$. (是否有人在$t$时刻刚刚讲完触发词)