

- 广告位
 - 一起刷题群：
 - 添加微信heywhaleshequ，备注“学习”，带你认识更多一起刷题的小伙伴
- ~

大家好，吴恩达《深度学习》多有名多经典我就不赘述啦，此次完成的是配套测验（quiz）的翻译。

但是因为本身是付费教程，github上翻了一下，完整的题目资源都很少，更别提翻译啦~所以就快马加鞭地给大家安排上了。感谢神仙博主@何宽 部分翻译都是从他的文章中直接照搬的，以及慷慨解囊的@黄海广 在我全部写完之后，从兜里抠出一个早就翻好的文件

参考资料：

- [何宽 - 【deplearning.ai】【吴恩达课后作业】 - CSDN](#)
- [ilarum19 - coursera-deeplearning.ai-NNDeepLearning - github](#)
- [黄海广 - 知乎](#)

其他x题系列：

- [50道练习带你玩转Pandas](#)
- [这100道练习，带你玩转Numpy](#)
- [35题初探scikit-learn库，get机器学习好帮手\](#)
- [50题matplotlib从入门到精通](#)
- [40题刷爆Keras，人生苦短我选Keras](#)
- [60题PyTorch简易入门指南，做技术的弄潮儿](#)
- [90题细品吴恩达《机器学习》，感受被刷题支配的恐惧](#)
- [170题吴恩达《深度学习》面面观，一套更比三套强](#)

课程一 - 神经网络和深度学习

第一周 - 深度学习简介

第 1 题

“人工智能是新电力”这个比喻指的是什么？

- A.人工智能为我们的家庭和办公室的个人设备供电，类似于电力。
- B.通过“智能电网”，人工智能正在传递新一波的电力。
- C.人工智能在计算机上运行，因此由电力驱动，但它让计算机做以前不可能做的事情。

D.与100年前开始的电力类似，人工智能正在改变多个行业。

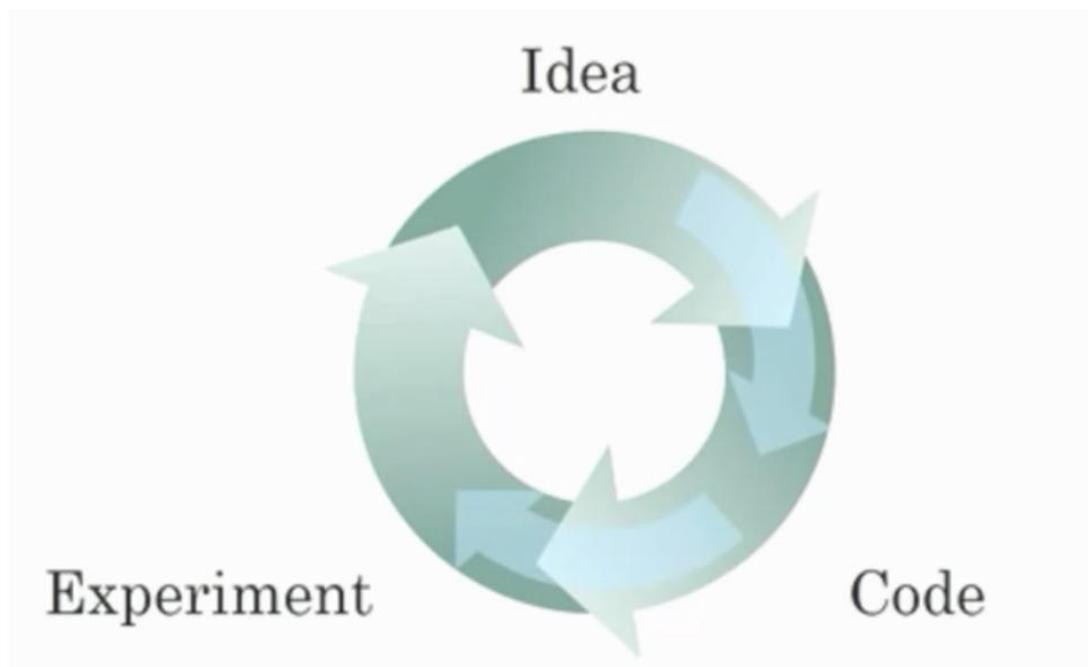
第 2 题

以下哪些是最近深度学习开始崛起的原因？（选2个答案）

- A.我们拥有了更多的计算能力
- B.神经网络是一个崭新的领域。
- C.我们有了更多的数据。
- D.深度学习在诸如在线广告、语音识别和图像识别等重要应用方面取得了显著的改进。

第 3 题

回想一下这个机器学习迭代的图。以下哪项陈述是正确的？（选出所有正确项）



- A.能够快速地尝试各种想法可以让深入学习的工程师更快地迭代。
- B.更快的计算有助于加快团队迭代一个好主意所需的时间。
- C.在大数据集上训练比在小数据集上训练更快。
- D.深度学习算法的最新进展使我们能够更快地训练好的模型（即使不改变CPU/GPU硬件）。

第 4 题

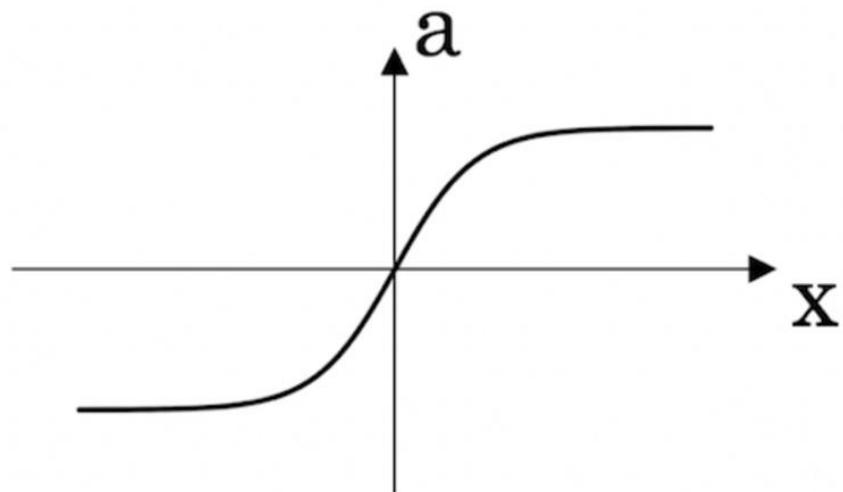
当一个有经验的深度学习工程师处理一个新问题时，他们通常可以在第一次尝试时利用以前问题的洞察力来训练一个好的模型，而不需要在不同的模型中重复多次。

- A. 对
B. 不对

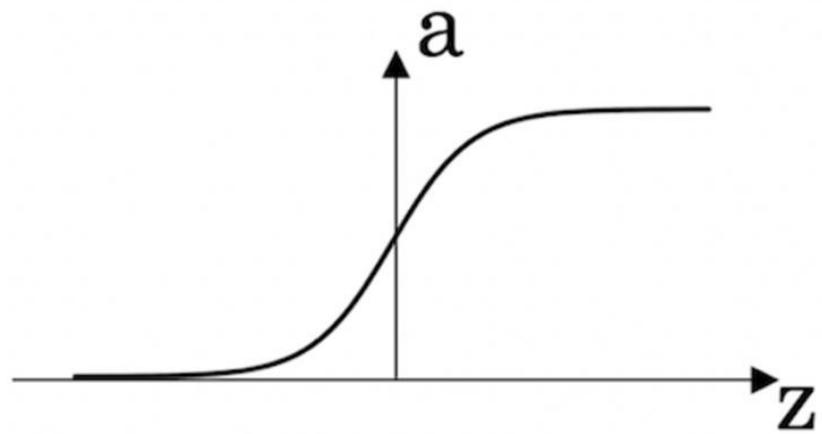
第 5 题

这些图中的哪一个表示ReLU激活函数？

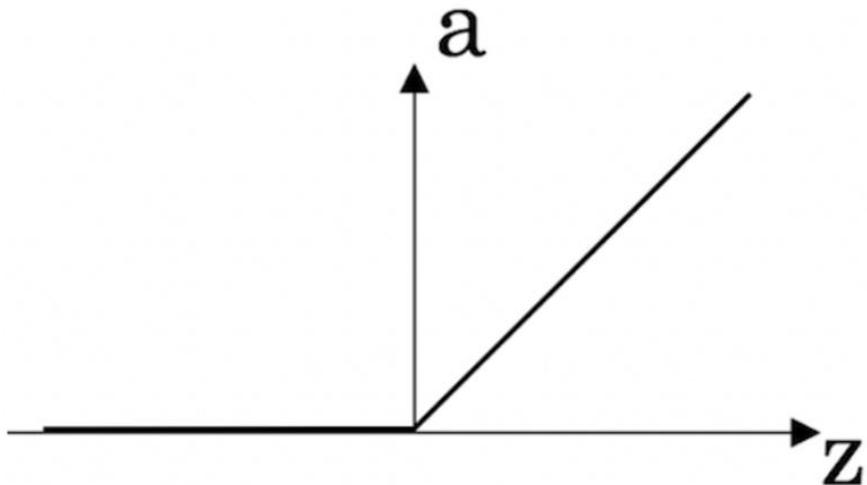
A.



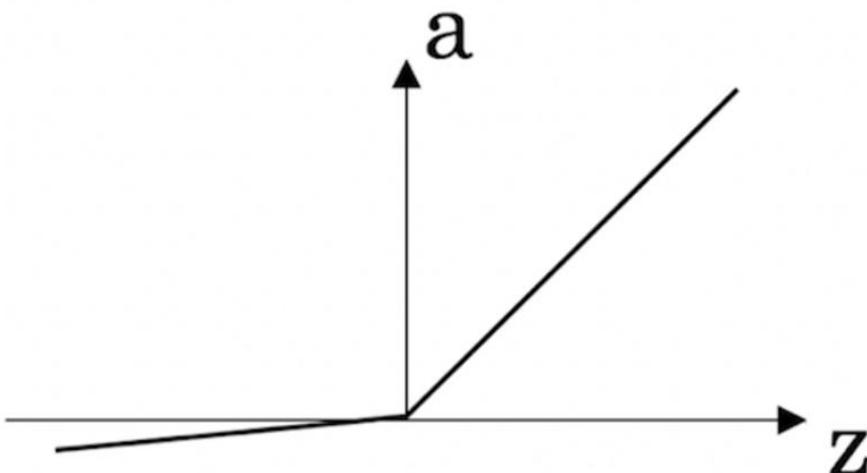
B.



C.



D.



第 6 题

用于猫识别的图像是“结构化”数据的一个例子，因为它在计算机中表示为结构化的数组。

- A.对
- B.不对

第 7 题

人口数据集包含不同城市人口、人均GDP、经济增长的统计数据，这是“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。

- A.对
- B.不对

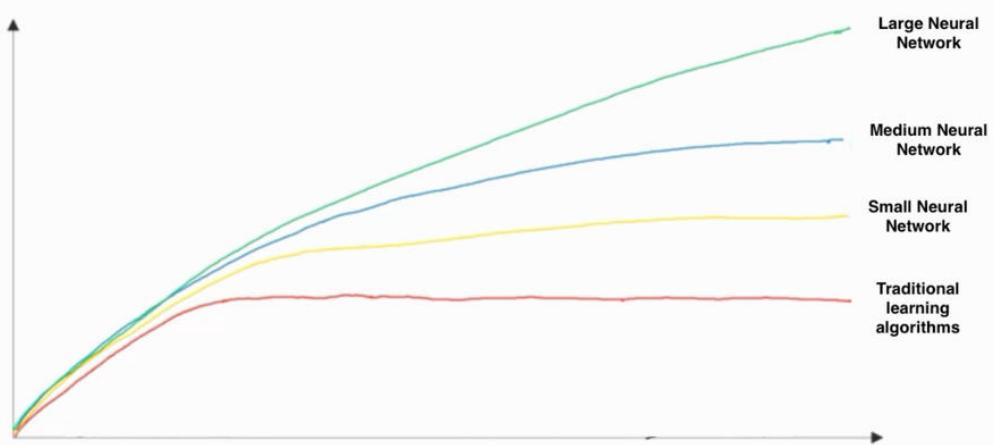
第 8 题

为什么RNN（递归神经网络）被用于机器翻译，比如说将英语翻译成法语？（选出所有正确项）

- A.它可以训练成一个有监督的学习问题
- B.它比卷积神经网络（CNN）更强大
- C.当输入/输出是一个序列（例如，一个单词序列）时适用
- D.RNN表示 想法->代码->实验->想法->... 的循环过程

第9题

在这PPT截图中，水平轴（X轴）和垂直轴（Y轴）代表什么？



- A.X轴代表数据量，Y轴代表模型规模
- B.X轴代表数据量，Y轴代表模型表现
- C.X轴代表模型表现，Y轴代表数据量
- D.X轴代表模型的输入，Y轴代表输出

第10题

假设前一个问题中所描述的趋势是准确的（并且希望你的坐标轴标签正确），下列哪一个是正确的？（选出所有正确项）

- A.增加训练集的大小通常不会影响算法的性能，而且可能会有很大帮助。
- B.增加神经网络的规模通常不会影响算法的性能，而且可能会有很大帮助。
- C.减小训练集的大小通常不会影响算法的性能，而且可能会有很大帮助。
- D.减小神经网络的规模通常不会影响算法的性能，而且可能会有很大帮助。

1-10题 答案

1.D 2.ACD 3.ABD 4.B 5.C 6.B 7.B 8.AC 9.B 10.AB

第二周 - 神经网络基础

第 11 题

神经元计算什么？

- A. 神经元计算激活函数后，再计算线性函数 ($z=Wx+b$)
- B. 神经元计算一个线性函数 ($z=Wx+b$)，然后接一个激活函数
- C. 神经元计算一个函数 g ，它线性地缩放输入 x ($Wx+b$)
- D. 神经元先计算所有特征的平均值，然后将激活函数应用于输出

第 12 题

以下哪一个是逻辑回归的损失函数？

- A. $L^{(i)}(\hat{y}^{(i)}, y^{(i)}) = |y^{(i)} - \hat{y}^{(i)}|$
- B. $L^{(i)}(\hat{y}^{(i)}, y^{(i)}) = \max(0, y^{(i)} - \hat{y}^{(i)})$
- C. $L^{(i)}(\hat{y}^{(i)}, y^{(i)}) = |y^{(i)} - \hat{y}^{(i)}|^2$
- D. $L^{(i)}(\hat{y}^{(i)}, y^{(i)}) = -(y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}))$

第 13 题

假设 img 是一个 $(32, 32, 3)$ 数组，表示一个 32×32 图像，它有三个颜色通道：红色、绿色和蓝色。如何将其重塑为列向量？

- A. $x = \text{img.reshape}((1, 32*32, 3))$
- B. $x = \text{img.reshape}((32*32*3, 1))$
- C. $x = \text{img.reshape}((3, 32*32))$
- D. $x = \text{img.reshape}((32*32, 3))$

第 14 题

考虑以下两个随机数组 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
```

c 的维度是什么？

- A. $c.\text{shape} = (3, 2)$
- B. $c.\text{shape} = (2, 1)$
- C. $c.\text{shape} = (2, 3)$

D.计算不成立因为这两个矩阵维度不匹配

第 15 题

考虑以下两个随机数组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
```

c的维度是什么？

A.c.shape = (4, 3)

B.c.shape = (3, 3)

C.c.shape = (4, 2)

D.计算不成立因为这两个矩阵维度不匹配

第 16 题

假设每个示例有 n_x 个输入特性， $X = [X^{(1)}, X^{(2)} \dots, X^{(m)}]$ 。 X 的维数是多少？

A.(m, 1)

B.(1, m)

C.(n_x , m)

D.(m, n_x)

第 17 题

np.dot(a, b)对a和b的进行矩阵乘法，而a*b执行元素的乘法，考虑以下两个随机数组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)
```

c的维度是什么？

A.c.shape = (12288, 150)

B.c.shape = (150, 150)

C.c.shape = (12288, 45)

D.计算不成立因为这两个矩阵维度不匹配

第 18 题

请考虑以下代码段：

```
# 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]
```

如何将之矢量化？

- A. $c = a + d$
- B. $c = a + b.T$
- C. $c = a.T + b.T$
- D. $c = a.T + b$

第 19 题

请考虑以下代码段：

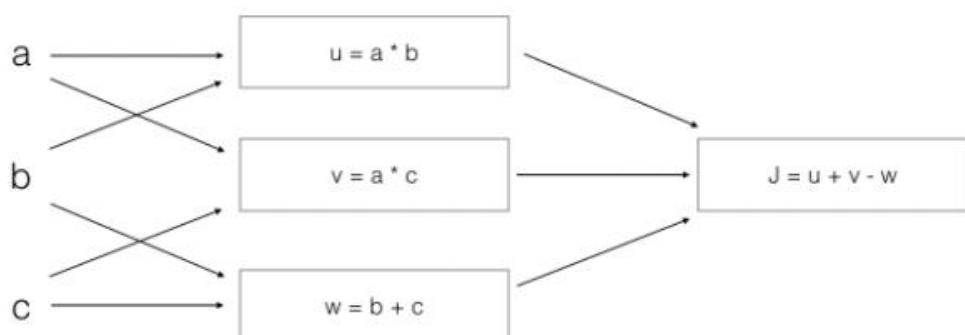
```
a = np.random.randn(3, 3)
b = np.random.randn(3, 1)
c = a * b
```

c的维度是什么？

- A. 这会触发广播机制，b会被复制3次变成 (3×3) ，而*操作是元素乘法，所以 $c. shape = (3, 3)$
- B. 这会触发广播机制，b会被复制3次变成 (3×3) ，而*操作是矩阵乘法，所以 $c. shape = (3, 3)$
- C. 这个操作将一个 3×3 矩阵乘以一个 3×1 的向量，所以 $c. shape = (3, 1)$
- D. 这个操作会报错，因为你不能用*对这两个矩阵进行操作，你应该用 $\text{np. dot}(a, b)$

第 20 题

请考虑以下计算图：



输出J是?

- A.J = (c - 1) * (b + a)
- B.J = (a - 1) * (b + c)
- C.J = a*b + b*c + a*c
- D.J = (b - 1) * (c + a)

11-20题 答案

11.B 12.D 13.B 14.C 15.D 16.C 17.C 18.B 19.A 20.B

第三周 - 浅层神经网络

第 21 题

以下哪项是正确的? (选出所有正确项)

- A. $a^{[2](12)}$ 是第12层, 第2个训练数据的激活向量
- B. X 是一个矩阵, 其中每个列是一个训练数据
- C. $a_4^{[2]}$ 是第2层, 第4个训练数据的激活输出
- D. $a_4^{[2]}$ 是第2层, 第4个神经元的激活输出
- E. $a^{[2]}$ 表示第2层的激活向量
- F. $a^{[2](12)}$ 是第2层, 第12个数据的激活向量
- G. X 是一个矩阵, 其中每个行是一个训练数据

第 22 题

对于隐藏单元, tanh激活通常比sigmoid激活函数更有效, 因为其输出的平均值接近于零, 因此它可以更好地将数据集中到下一层。

- A.对
- B.不对

第 23 题

以下哪一个 l 层的正向传播的正确矢量化实现, 其中 $1 \leq l \leq L$

- A.
 $Z^{[l]} = W^{[l]} A^{[l]} + b^{[l]}$
 $A^{[l+1]} = g^{[l]}(Z^{[l]})$
- B.
 $Z^{[l]} = W^{[l]} A^{[l]} + b^{[l]}$
 $A^{[l+1]} = g^{[l+1]}(Z^{[l]})$

C.

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

$$A^{[l]} = g^{[l]}(Z^{[l]})$$

D.

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

$$A^{[l]} = g^{[l]}(Z^{[l]})$$

第 24 题

您正在构建一个用于识别黄瓜 ($y=1$) 与西瓜 ($y=0$) 的二进制分类器。对于输出层，您建议使用哪一个激活函数？

A.ReLU

B.Leaky ReLU

C.sigmoid

D.tanh

第 25 题

考虑以下代码：

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

B. shape是多少？

A.(4,)

B.(1, 3)

C.(, 3)

D.(4, 1)

第 26 题

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？（选出所有正确项）

A.第一隐藏层中的每个神经元将执行相同的计算。因此，即使在梯度下降的多次迭代之后，层中的每个神经元将执行与其他神经元相同的计算。

B.第一隐层中的每个神经元在第一次迭代中执行相同的计算。但是在梯度下降的一次迭代之后，他们将学会计算不同的东西，因为我们已经“破坏了对称性”。

C.第一个隐藏层中的每个神经元将执行相同的计算，但不同层中的神经元执行不同的计算，因此我们完成了课堂上所描述的“对称性破坏”。

D.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，因此，它们的参数会以自己的方式不断演化。

第 27 题

逻辑回归的权重w应该随机初始化，而不是全部初始化为全部零，否则，逻辑回归将无法学习有用的决策边界，因为它将无法“打破对称”

- A.对
- B.不对

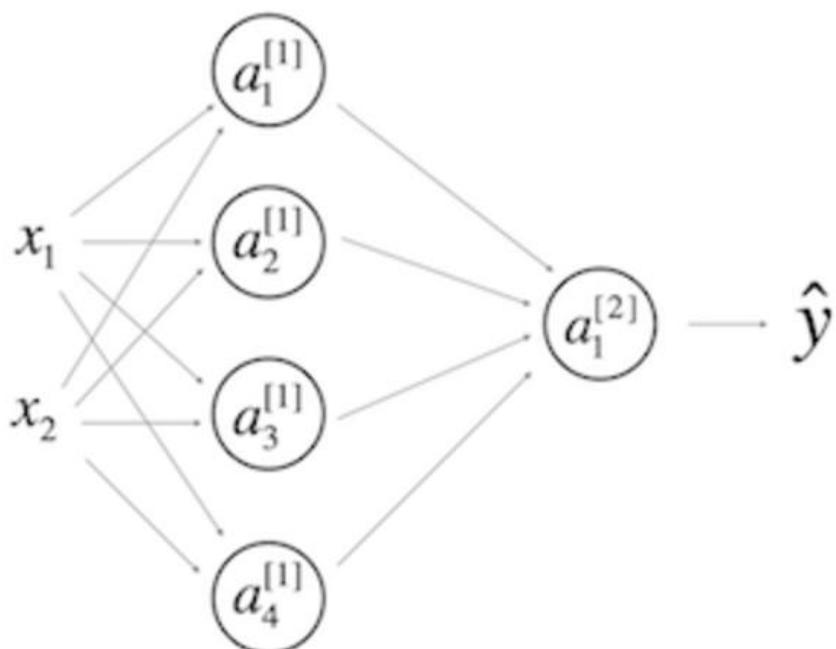
第 28 题

你已经为所有隐藏的单位建立了一个使用tanh激活的网络。使用`np.random.randn(..., ...)*1000`将权重初始化为相对较大的值。会发生什么？

- A.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。
- B.这将导致tanh的输入也非常大，从而导致梯度也变大。因此，你必须将 α 设置得非常小，以防止发散；这将减慢学习速度。
- C.这将导致tanh的输入也非常大，导致单元被“高度激活”。与权重从小值开始相比，加快了学习速度。
- D.这将导致tanh的输入也非常大，从而导致梯度接近于零。因此，优化算法将变得缓慢。

第 29 题

考虑以下1个隐层的神经网络：



- A. $W^{[1]}$ 的形状是(2, 4)
- B. $b^{[1]}$ 的形状是(4, 1)
- C. $W^{[1]}$ 的形状是(4, 2)
- D. $b^{[1]}$ 的形状是(2, 1)
- E. $W^{[2]}$ 的形状是(1, 4)
- F. $b^{[2]}$ 的形状是(4, 1)
- G. $W^{[2]}$ 的形状是(4, 1)
- H. $b^{[2]}$ 的形状是(1, 1)

第 30 题

在和上一问相同的网络中， $Z^{[1]}$ 和 $A^{[1]}$ 的维度是多少？

- A. $Z^{[1]}$ 和 $A^{[1]}$ 是(4, 1)
- B. $Z^{[1]}$ 和 $A^{[1]}$ 是(1, 4)
- C. $Z^{[1]}$ 和 $A^{[1]}$ 是(4, m)
- D. $Z^{[1]}$ 和 $A^{[1]}$ 是(4, 2)

21-30题 答案

21.BDEF 22.A 23.D 24.C 25.D 26.A 127.B 28.D 29.BC EH 30.C

第四周 - 深度神经网络的核心概念

第 31 题

在我们的前向传播和后向传播实现中使用的“缓存”是什么？

- A. 它用于在训练期间缓存成本函数的中间值。
- B. 我们用它将在正向传播过程中计算的变量传递到相应的反向传播步骤。它包含了反向传播计算导数的有用值。
- C. 它用于跟踪我们正在搜索的超参数，以加快计算速度。
- D. 我们用它将反向传播过程中计算的变量传递到相应的正向传播步骤。它包含用于计算正向传播的激活的有用值。

第 32 题

以下哪些是“超参数”？(选出所有正确项)

- A. 隐藏层规模 $n^{[l]}$
- B. 神经网络的层数 L

C. 激活向量 $a^{[l]}$

D. 权重矩阵 $W^{[l]}$

E. 学习率 α

F. 迭代次数

G. 偏置向量 $b^{[l]}$

第 33 题

下列哪个说法是正确的?

A. 神经网络的更深层通常比前面的层计算更复杂的特征

B. 神经网络的前面的层通常比更深层计算更复杂的特性

第 34 题

向量化允许您在 L 层神经网络中计算前向传播时，不需要在层 $l = 1, 2, \dots, L$ 间显式的使用 for 循环（或任何其他显式迭代循环）

A. 对

B. 不对

第 35 题

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 4 个隐藏单元，第 2 层有 3 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？

A.

```
for(i in range(1, len(layer_dims/2))):  
    parameter['W' + str(i)] = np.random.randn(layers[i], layers[  
    parameter['b' + str(i)] = np.random.randn(layers[i], 1) * 0.
```

B.

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

C.

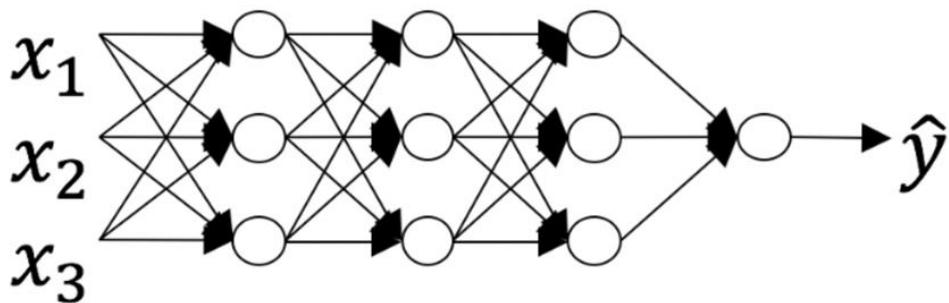
```
for(i in range(1, len(layer_dims))):  
    parameter['W' + str(i)] = np.random.randn(layers[i-1], layer  
    parameter['b' + str(i)] = np.random.randn(layers[i], 1) * 0.
```

D.

```
for(i in range(1, len(layer_dims))):  
    parameter['w' + str(i)] = np.random.randn(layers[i], layers[  
    parameter['b' + str(i)] = np.random.randn(layers[i], 1) * 0.
```

第 36 题

考虑以下神经网络：



该神经网络有几层？

- A. 层数 L 是 4，隐藏层数是 3
- B. 层数 L 是 3，隐藏层数是 3
- C. 层数 L 是 4，隐藏层数是 4
- D. 层数 L 是 5，隐藏层数是 4

第 37 题

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

- A. 对
- B. 不对

第 38 题

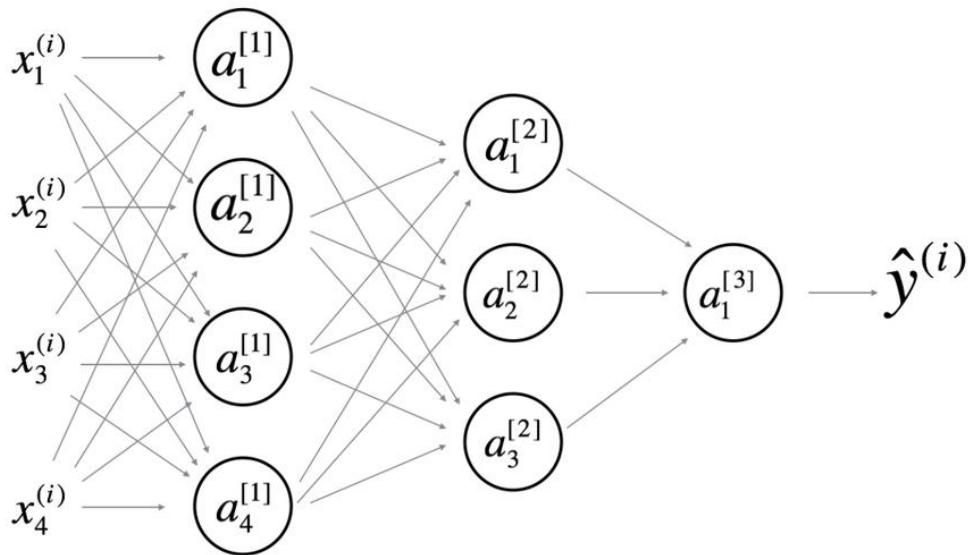
有一些函数具有以下特性：

- (i) 当使用浅网络计算时，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) 当使用深网络来计算时，我们只需要一个指数级小的网络

- A. 对
- B. 不对

第 39 题

在以下2层隐藏层的神经网络中，以下哪句话是正确的？



- A. $W^{[1]}$ 的形状是 $(4, 4)$
- B. $b^{[1]}$ 的形状是 $(4, 1)$
- C. $W^{[2]}$ 的形状是 $(3, 4)$
- D. $b^{[2]}$ 的形状是 $(3, 1)$
- E. $b^{[3]}$ 的形状是 $(1, 1)$
- F. $W^{[3]}$ 的形状是 $(1, 3)$

第 40 题

前面的问题使用了一个特定的网络，一般情况下，层 l 的权重矩阵 $W^{[l]}$ 的维数是多少？

- A. $W^{[l]}$ 的形状是 $(n^{[l]}, n^{[l-1]})$
- B. $W^{[l]}$ 的形状是 $(n^{[l-1]}, n^{[l]})$
- C. $W^{[l]}$ 的形状是 $(n^{[l+1]}, n^{[l]})$
- D. $W^{[l]}$ 的形状是 $(n^{[l]}, n^{[l+1]})$

31-40题 答案

31.B 32.ABEF 33.A 34.B 35.D 36.A 37.A 38.A 39.ABCDEF 40.A

课程二 - 改善深层神经网络

第一周 - 深度学习的实践

第 41 题

如果你有10,000,000个例子，你会如何划分训练/开发/测试集？

- A.33%训练, 33%开发, 33%测试
- B.60%训练, 20%开发, 20%测试
- C.98%训练, 1%开发, 20%测试

第 42 题

开发和测试集应该：

- A.来自同一分布
- B.来自不同分布
- C.完全相同（一样的(x, y)对）
- D.数据数量应该相同

第 43 题

如果你的神经网络方差很高，下列哪个尝试是可能解决问题的？

- A.添加正则项
- B.获取更多测试数据
- C.增加每个隐藏层的神经元数量
- D.用更深的神经网络
- E.用更多的训练数据

第 44 题

你正在为苹果，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？

- A.增大正则化参数 λ
- B.减小正则化参数 λ
- C.获取更多训练数据
- D.用更大的神经网络

第 45 题

什么是权重衰减？

- A.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
- B.在训练过程中逐渐降低学习率的过程
- C.如果神经网络是在噪声数据下训练的，那么神经网络的权值会逐渐损坏
- D.通过对权重值设置上限来避免梯度消失的技术

第 46 题

当你增大正则化的超参数 λ 时会发生什么？

- A.权重变小（接近0）
- B.重量变大（远离0）
- C.2倍的 λ 导致2倍的权重
- D.每次迭代，梯度下降采取更大的步距（与 λ 成正比）

第 47 题

在测试时候使用dropout：

- A.不随机关闭神经元，但保留 $1/keep_prob$ 因子
- B.随机关闭神经元，保留 $1/keep_prob$ 因子
- C.随机关闭神经元，但不保留 $1/keep_prob$ 因子
- D.不随机关闭神经元，也不保留 $1/keep_prob$ 因子

第 48 题

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（选出所有正确项）：

- A.正则化效应被增强
- B.正则化效应被减弱
- C.训练集的误差会增加
- D.训练集的误差会减小

第 49 题

以下哪些技术可用于减少方差（减少过拟合）？（选出所有正确项）

- A.梯度消失
- B.数据扩充
- C.Dropout
- D.梯度检查
- E.Xavier初始化

F.L2正则化

G.梯度爆炸

第 50 题

为什么要对输入 x 进行归一化?

- A.让参数初始化更快
- B.让代价函数更快地优化
- C.更容易做数据可视化
- D.是另一种正则化——有助减少方差

41-50题 答案

41.C 42.A 43.AE 44.AC 45.A 46.A 47.D 48.BD 49.BCF 50.B

第二周 - 优化算法

第 51 题

当输入从第8个mini-batch的第7个的例子的时候，你会用哪种符号表示第3层的激活?

- A. $a^{[3]\{8\}(7)}$
- B. $a^{[8]\{7\}(3)}$
- C. $a^{[8]\{3\}(7)}$
- D. $a^{[3]\{7\}(8)}$

第 52 题

关于mini-batch的说法哪个是正确的?

- A.mini-batch迭代一次（计算1个mini-batch），要比批量梯度下降迭代一次快
- B.用mini-batch训练完整个数据集一次，要比批量梯度下降训练完整个数据集一次快
- C.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（矢量化）

第 53 题

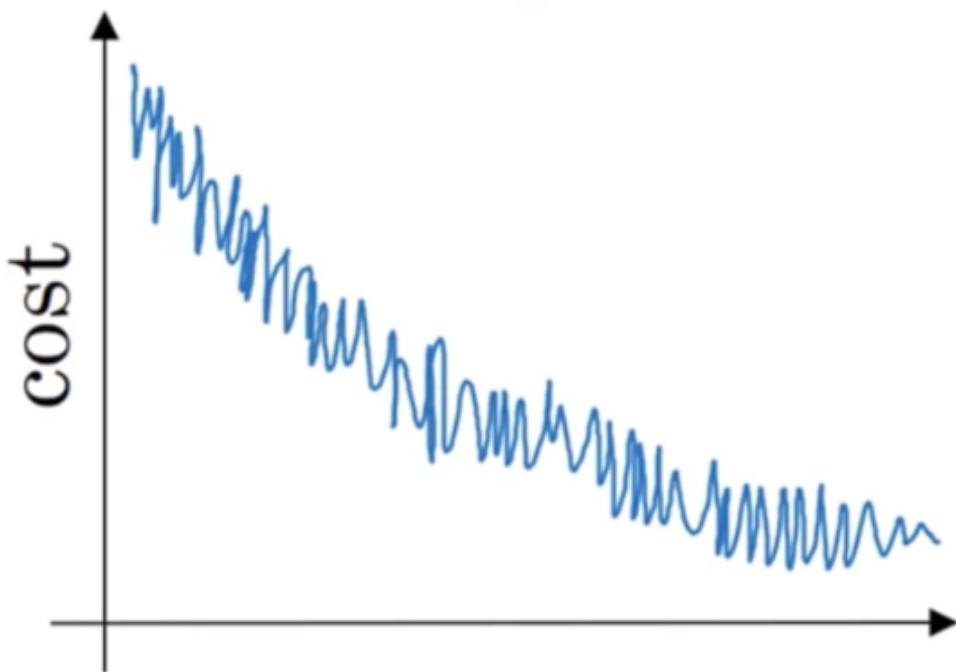
为什么最好的mini-batch的大小通常不是1也不是m，而是介于两者之间?

- A.如果mini-batch的大小是1，那么在你取得进展前，你需要遍历整个训练集

- B.如果mini-batch的大小是m，就会变成批量梯度下降。在你取得进展前，你需要遍历整个训练集
- C.如果mini-batch的大小是1，那么你将失去mini-batch将数据矢量化带来的的好处
- D.如果mini-batch的大小是m，就会变成随机梯度下降，而这样做经常会比mini-batch慢

第 54 题

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：



- A.如果你正在使用mini-batch梯度下降，那可能有问题；而如果你在使用批量梯度下降，那是合理的
- B.如果你正在使用mini-batch梯度下降，那看上去是合理的；而如果你在使用批量梯度下降，那可能有问题
- C.无论你在使用mini-batch还是批量梯度下降，看上去都是合理的
- D.无论你在使用mini-batch还是批量梯度下降，都可能有问题

第 55 题

假设一月的前三天卡萨布兰卡的气温是一样的：

一月第一天: $\theta_1 = 10$

一月第二天: $\theta_2 = 10$

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？

- A. $v_2 = 10, v_2^{corrected} = 10$

- B. $v_2 = 10, v_2^{corrected} = 7.5$
- C. $v_2 = 7.5, v_2^{corrected} = 7.5$
- D. $v_2 = 7.5, v_2^{corrected} = 10$

第 56 题

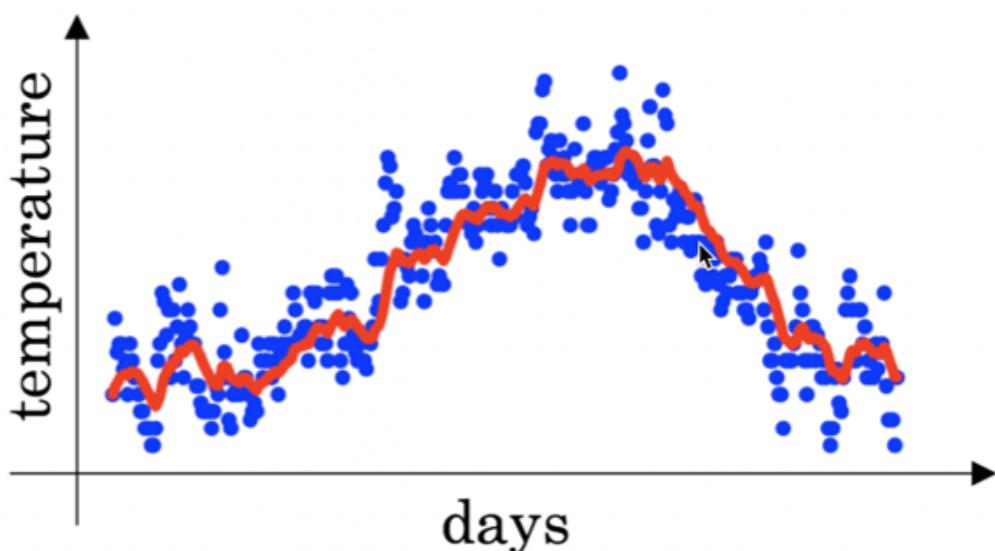
下面哪一个不是比较好的学习率衰减方法?

- A. $\alpha = \frac{1}{1+2*t} \alpha_0$
- B. $\alpha = \frac{1}{\sqrt{t}} \alpha_0$
- C. $\alpha = 0.95^t \alpha_0$
- D. $\alpha = e^t \alpha_0$

第 57 题

您在伦敦温度数据集上使用指数加权平均， 使用以下公式来追踪温度：

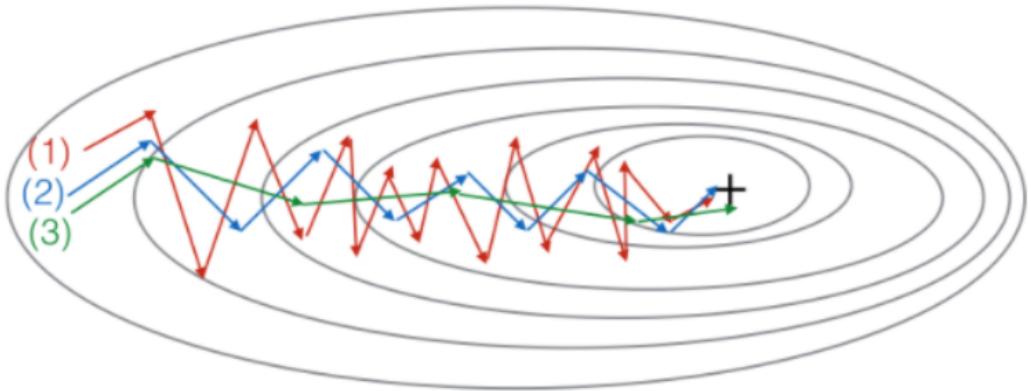
$v_t = \beta v_{t-1} + (1 - \beta) \theta_t$ 。下图中红线使用的是 $\beta = 0.9$ 来计算的。当你改变 β 时，你的红色曲线会怎样变化？（选出所有正确项）



- A. 减小 β , 红色线会略微右移
- B. 增加 β , 红色线会略微右移
- C. 减小 β , 红色线会更加震荡
- D. 增加 β , 红色线会更加震荡

第 58 题

下图中的曲线是由：梯度下降，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？



- A.(1)是梯度下降；(2)是动量梯度下降 ($\beta = 0.9$) ；(3)是动量梯度下降 ($\beta = 0.5$)
 B.(1)是梯度下降；(2)是动量梯度下降 ($\beta = 0.5$) ；(3)是动量梯度下降 ($\beta = 0.9$)
 C.(1)是动量梯度下降 ($\beta = 0.5$) ；(2)是动量梯度下降 ($\beta = 0.9$) ；(3)是梯度下降
 D.(1)是动量梯度下降 ($\beta = 0.5$) ；(2)是梯度下降；(3)是动量梯度下降 ($\beta = 0.9$)

第 59 题

假设在一个深度学习网络中，批量梯度下降花费了大量时间时来找到一组参数值，使成本函数 $J(W^{[1]}, b^{[1]}, \dots, W^{[L]}, b^{[L]})$ 小。以下哪些方法可以帮助找到 J 值较小的参数值？

- A.令所有权重值初始化为0
- B.尝试调整学习率
- C.尝试mini-batch梯度下降
- D.尝试对权重进行更好的随机初始化
- E.尝试使用 Adam 算法

第 60 题

关于Adam算法，下列哪一个陈述是错误的？

- A.Adam结合了Rmsprop和动量的优点
- B.Adam中的学习率超参数 α 通常需要调整
- C.我们经常使用超参数的“默认”值 $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$
- D.Adam应该用于批梯度计算，而不是用于mini-batch

51-60题 答案

51.A 52.C 53.BC 54.B 55.D 56.D 57.BC 58.B 59.BCDE 60.D

第三周 - 超参数调整，批量标准化，编程框架

第 61 题

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？

- A. 对
- B. 不对

第 62 题

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？

- A. 对
- B. 不对

第 63 题

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：

- A. 是否使用批量（batch）或小批量优化（mini-batch optimization）
- B. 神经网络中局部最小值（鞍点）的存在性
- C. 在你能力范围内，你能够拥有多大的计算能力（博主注：就是高性能电脑和低性能电脑的区别）
- D. 需要调整的超参数的数量

第 64 题

如果您认为 β （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 β 值进行取样？

A.

```
r = np.random.rand()  
beta = r * 0.09 + 0.9
```

B.

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

C.

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

D.

```
r = np.random.rand()  
beta = r * 0.9 + 0.09
```

第 65 题

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？

- A.对
- B.不对

第 66 题

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，您应该对谁进行标准化？

- A. $z^{[l]}$
- B. $W^{[l]}$
- C. $a^{[l]}$
- D. $b^{[l]}$

第 67 题

在标准化公式 $z_{norm}^{(i)} = \frac{z^{(i)} - \mu}{\sqrt{\sigma^2 + \epsilon}}$ ，为什么要使用epsilon (ϵ) ？

- A.为了更准确地标准化
- B.为了避免除零操作
- C.为了加速收敛
- D.防止 μ 太小

第 68 题

批标准化中关于 γ 和 β 的以下哪些陈述是正确的？

- A.对于每个层，有一个全局值 $\gamma \in \mathbb{R}$ 和一个全局值 $\beta \in \mathbb{R}$ ，适用于该层中的所有隐藏单元。
- B. γ 和 β 是算法的超参数，我们通过随机采样进行调整
- C.它们确定了给定层的线性变量 $z^{[l]}$ 的均值和方差

D. 最佳值是 $\gamma = \sqrt{\sigma^2 + \epsilon}$, $\beta = \mu$

E. 它们可以用Adam、动量的梯度下降或RMSprop，而不仅仅是用梯度下降来学习

第 69 题

在训练了具有批标准化的神经网络之后，在用新样本评估神经网络的时候，您应该：

A. 如果你在256个例子的mini-batch上实现了批标准化，那么如果你要在一个测试例子上进行评估，你应该将这个例子重复256次，这样你就可以使用和训练时大小相同的mini-batch进行预测。

B. 使用最新的mini-batch的 μ 和 σ^2 值来执行所需的标准化

C. 跳过用 μ 和 σ^2 值标准化的步骤，因为一个例子不需要标准化

D. 执行所需的标准化，使用在训练期间，通过指数加权平均值得出的 μ 和 σ^2

第 70 题

关于深度学习编程框架的这些陈述中，哪一个是正确的？（选出所有正确项）

A. 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。

B. 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。

C. 深度学习编程框架的运行需要基于云的机器。

61-70题 答案

61.B 62.B 63.C 64.B 65.B 66.A 67.B 68.CE 69.D 70.AB

课程三 - 结构化机器学习项目

第一周 - 和平之城中的鸟类识别（案例研究）

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。

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：

- $y = 0$: 图片中没有鸟类
- $y = 1$: 图片中有鸟类

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。

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：

1. 拥有较高的准确度
2. 快速运行，只需要很短的时间来分类一个新的图像。
3. 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。

第 71 题

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

- A.对
B.不对

第 72 题

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。”

如果你有以下三个模型，你会选择哪一个？

- A. 测试准确度：97%；运行时间：1 sec；内存大小：3MB
B. 测试准确度：99%；运行时间：13 sec；内存大小：9MB
C. 测试准确度：97%；运行时间：3 sec；内存大小：2MB
D. 测试准确度：98%；运行时间：9 sec；内存大小：9MB

第 73 题

根据城市的要求，您认为以下哪一项是正确的？

- A. 准确度是一个优化指标；运行时间和内存大小是满意指标。
B. 准确度是一个满意指标；运行时间和内存大小是一个优化指标。
C. 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。

D.准确性、运行时间和内存大小都是满意指标，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。

第 74 题

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？

- A.训练集:3,333,334；开发集：3,333,333；测试集：3,333,333
- B.训练集:9,500,000；开发集：250,000；测试集：250,000
- C.训练集:6,000,000；开发集：3,000,000；测试集：1,000,000
- D.训练集:6,000,000；开发集：1,000,000；测试集：3,000,000

第 75 题

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。

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

- A.对
- B.不对

第 76 题

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：

- A.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
- B.公民的数据图像与其他数据没有一致的 $x \rightarrow y$ 映射(类似于纽约/底特律的住房价格例子)
- C.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
- D.测试集不再反映您最关心的数据(安全摄像头)的分布

第 77 题

你训练了一个系统，其误差度如下（误差度 = 100% - 准确度）：

训练集误差：4.0%

开发集误差：4.5%

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？

- A.是的，因为有4%的训练误差表明你有很高的偏差。
- B.是的，因为这表明你的模型的偏差高于方差。
- C.不同意，因为方差高于偏差。
- D.不同意，因为没有足够的信息，这什么也说明不了。

第 78 题

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：

鸟类专家1 错误率：0.3%

鸟类专家2 错误率：0.5%

普通人1 (不是专家) 错误率：1.0%

普通人2 (不是专家) 错误率：1.2%

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”？

- A.0.0% (因为不可能做得比这更好)
- B.0.3% (专家1的错误率)
- C.0.4% (0.3 到 0.5 之间)
- D.0.75% (以上所有四个数字的平均值)

第 79 题

您同意以下哪项陈述？

- A.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。
- B.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。
- C.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。
- D.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。

第 80 题

你发现一组鸟类学家辩论和讨论图像，可以得到一个更好的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：

人类表现 0.1%

训练集误差 2.0%

开发集误差 2.1%

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）

- A.尝试增加正则化。

- B. 获得更大的训练集以减少差异。
- C. 尝试减少正则化。
- D. 训练一个更大的模型，试图在训练集上做得更好。

第 81 题

你在测试集上评估你的模型，并找到以下内容：

人类表现 0.1%

训练集误差 2.0%

开发集误差 2.1%

测试集误差 7.0%

这意味着什么？（两个最佳选项）

- A. 你没有拟合开发集
- B. 你应该尝试获得更大的开发集
- C. 你应该得到一个更大的测试集
- D. 你对开发集过拟合了

第 82 题

在一年后，你完成了这个项目，你终于实现了：

人类表现 0.10%

训练集误差 0.05%

开发集误差 0.05%

你能得出什么结论？（检查所有选项。）

- A. 现在很难衡量可避免偏差，因此今后的进展将会放缓。
- B. 统计异常（统计噪声的结果），因为它不可能超过人类表现。
- C. 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%。
- D. 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。

第 83 题

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？

- A. 查看开发过程中开发的所有模型，找出错误率最低的模型。

- B.要求你的团队在开发过程中同时考虑准确性和假阴性率。
- C.重新思考此任务的指标，并要求您的团队调整到新指标。
- D.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。

第 84 题

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。

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？

- A.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。
- B.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。
- C.尝试数据增强/数据合成，以获得更多的新鸟的图像。
- D.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

第 85 题

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。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？（选出所有正确项）

- A.需要两周的时间来训练将会限制你迭代的速度。
- B.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。
- C.如果10,000,000个样本就足以建立一个足够好的猫探测器，你最好用10,000,00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。
- D.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。

71-85题答案

71.A 72.D 73.A 74.B 75.B 76.AD 77.D 78.B 79.A 80.CD 81.D 82.AD 83.C 84.A 85.ABC

第二周 - 自动驾驶（案例研究）

为了帮助你练习机器学习的策略，本周我们将介绍另一个场景，并询问你将如何行动。我们认为这个在机器学习项目中工作的“模拟器”将给出一个任务，即领导一个机器学习项

目可能是什么样的！

您受雇于一家初创公司，制造自动驾驶汽车。您负责检测图像中的道路标志（停车标志、人行横道标志、施工先行标志）和交通信号（红绿灯）。目标是识别每张图像中的这些对象。例如，上面的图像包含人行横道标志和红色交通灯



$$y^{(i)} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \begin{array}{l} \text{"stop sign"} \\ \text{"pedestrian crossing sign"} \\ \text{"construction ahead sign"} \\ \text{"red traffic light"} \\ \text{"green traffic light"} \end{array}$$

第 86 题

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）。

- A.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。
- B.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。
- C.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。
- D.花几天时间训练一个基本模型，看看它会犯什么错误。

第 87 题

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？

- A.对
- B.不对

第 88 题

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）？

- A.随机选择10,000图片
- B.随机选择500图片

C.500张算法分类错误的图片

D.10,000张算法分类错误的图片

第 89 题

在处理了数据几周后，你的团队得到以下数据：

100,000 张使用汽车前摄像头拍摄的标记了的图片。

900,000 张从互联网下载的标记了道路的图片。

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

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， $y^{(i)} =$

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 $\begin{bmatrix} 1 \\ ? \\ 0 \\ 1 \\ ? \end{bmatrix}$ 那么学习算法将无法使用该样本，是正确的吗？

A.对

B.不对

第 90 题

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？

A.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。

B.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。

C.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。

D.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。

第 91 题

假设您最终选择了以下拆分数据集的方式：

数据集	图片数量	算法产生的错误
训练集	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	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	2万张汽车前摄像头拍摄的图片	14.3%
测试集	2万张汽车前摄像头拍摄的图片	14.8%

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%。以下哪项是对的？(选出所有正确项)

- A.由于开发集和测试集的错误率非常接近，所以你过拟合了开发集。
- B.你有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。
- C.你有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。
- D.你有很大的方差问题，因为你的训练集上的错误率比人为错误率要高得多。
- E.你有很大的方差问题，因为你的模型不能很好地适应它从来没有见过，但是来自训练集同一分布的数据

第 92 题

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？

- A.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）。
- B.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）。
- C.没有足够的信息来判断你的朋友是对还是错。
- D.无论你的朋友是对还是错，这些信息都对你没有用。

第 93 题

您决定将重点放在开发集上，并手动检查是什么原因导致的错误。下面是一个表，总结了您的发现：

开发集总误差 14.3%

由于数据标记不正确而导致的错误 4.1%

由于雾天的图片引起的错误 8.0%

由于雨滴落在汽车前摄像头造成错误 2.2%

其他原因引起的错误 1.0%

在这个表格中，4.1%、8.0%这些比例是占总开发集的比例（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。

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？

- A. 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。
- B. 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。
- C. 是的，因为它比其他的错误类别错误率加在一起都大($8.0 > 4.1+2.2+1.0$)。
- D. 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。

第 94 题

你可以买一个专门设计的雨刮，帮助擦掉正面相机上的一些雨滴。根据上一个问题的表格，您同意以下哪些陈述？

- A. 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。
- B. 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。
- C. 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%。
- D. 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。

第 95 题

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：



你同意下列哪种说法？（选出所有正确项）

- A. 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幅图片就没有太大的过拟合的风险。
- B. 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到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。
- C. 只要合成的雾对人眼来说是真实的，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。

第 96 题

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？（选出所有正确项）

- A.您不应更正训练集中的错误标记的数据，因为这不值得
- B.您应该更正训练集中的错误标记数据，以免您训练集与开发集差距更大
- C.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
- D.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

第 97 题

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。

你告诉你的同事怎么做？

- A.她应该尝试使用在你的数据集上预先训练过的权重，并用黄光数据集进行进一步的微调。
- B.如果她有10,000个黄光图像，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。
- C.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。
- D.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。

第 98 题

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？

- A.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。
- B.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。
- C.迁移学习或多任务学习可以帮助我们的同事加快步伐。
- D.迁移学习和多任务学习都不是很有希望。

第 99 题

要识别红色和绿色的灯光，你一直在使用这种方法：

(A)将图像 x 输入到神经网络，并直接学习映射以预测是否存在红光(和/或)绿光 y 。

一个队友提出了另一种两步的方法：

(B)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。

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？

- A.对
- B.不对

第 100 题

上一题中，A方法似乎比B方法更有效，如果你有一个__

- A.大训练集
- B.多任务学习的问题
- C.偏差比较大的问题
- D.高贝叶斯误差的问题

86-100题 答案

86.D 87.B 88.C 89.B 90.C 91.BC 92.C 93.A 94.A 95.C 96.AD 98.A 99.D 99.B 100.A

课程四 - 卷积神经网络

第一周 - 卷积神经网络的基本知识

第 101 题

你认为把下面这个过滤器应用到灰度图像会怎么样？

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

- A.会检测45度边缘
- B.会检测垂直边缘
- C.会检测水平边缘
- D.会检测图像对比度

第 102 题

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）？

- A.9,000,001

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

第 103 题

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）？

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

第 104 题

你有一个 $63 \times 63 \times 16$ 的输入，并使用大小为 7×7 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？

- A. $29 \times 29 \times 32$
- B. $16 \times 16 \times 32$
- C. $29 \times 29 \times 16$
- D. $16 \times 16 \times 16$

第 105 题

你有一个 $15 \times 15 \times 8$ 的输入，并使用“pad = 2”进行填充，填充后的尺寸是多少？

- A. $17 \times 17 \times 10$
- B. $19 \times 19 \times 8$
- C. $19 \times 19 \times 12$
- D. $17 \times 17 \times 8$

第 106 题

你有一个 $63 \times 63 \times 16$ 的输入，有32个过滤器进行卷积，每个过滤器的大小为 7×7 ，步幅为1，你想要使用“same”的卷积方式，请问pad的值是多少？

- A.1
- B.2
- C.3

第 107 题

你有一个 $32 \times 32 \times 16$ 的输入，并使用步幅为2、过滤器大小为2的最大化池，请问输出是多少？

- A. $15 \times 15 \times 16$
- B. $16 \times 16 \times 8$
- C. $16 \times 16 \times 16$
- D. $32 \times 32 \times 8$

第 108 题

因为池化层不具有参数，所以它们不影响反向传播的计算。

- A. 对
- B. 不对

第 109 题

在视频中，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？（选出所有正确项）

- A. 它减少了参数的总数，从而减少过拟合。
- B. 它允许在整个输入值的多个位置使用特征检测器。
- C. 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）。
- D. 它允许梯度下降将许多参数设置为零，从而使得连接稀疏。

第 110 题

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？

- A. 正则化导致梯度下降将许多参数设置为零。
- B. 每个过滤器都连接到上一层的每个通道。
- C. 下一层中的每个激活只依赖于前一层的少量激活。
- D. 卷积网络中的每一层只连接到另外两层。

101-110题 答案

101.B 102.D 103.B 104.A 105.B 106.C 107.C 108.B 109.BD 110.C

第二周 - 深度卷积模型

第 111 题

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？

- A. n_H 和 n_W 增加，同时 n_C 减少
- B. n_H 和 n_W 减少，同时 n_C 也减少
- C. n_H 和 n_W 增加，同时 n_C 也增加
- D. n_H 和 n_W 减少，同时 n_C 增加

第 112 题

在典型的卷积神经网络中，你能看到的是？

- A. 多个卷积层后面跟着的是一个池化层
- B. 多个池化层后面跟着的是一个卷积层
- C. 全连接层 (FC) 位于最后的几层
- D. 全连接层 (FC) 位于开始的几层

第 113 题

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。

- A. 对
- B. 不对

第 114 题

训练更深层的网络（例如，在网络中添加额外的层）可以使网络适应更复杂的功能，因此几乎总是会导致更低的训练错误。对于这个问题，假设是指“普通”网络

- A. 对
- B. 不对

第 115 题

下面计算残差(ResNet)块的公式中，横线上应该分别填什么？

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

- A. 分别是 0 与 $z^{[l+1]}$
- B. 分别是 $a^{[l]}$ 与 0

C. 分别是 $z^{[l]}$ 与 $a^{[l]}$

D. 分别是 0 与 $a^{[l]}$

第 116 题

关于残差网络下面哪个（些）说法是正确的？

A. 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

B. 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

C. 有 L 层的残差网络一共有 L^2 种跳跃连接的顺序

D. 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

第 117 题

假设你的输入的维度为 $64 \times 64 \times 16$ ，单个 1×1 的卷积过滤器含有多少个参数（包括偏差）？

A. 2

B. 17

C. 4097

D. 1

第 118 题

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 1×1 ，步长为 1，padding 为 0）？

A. 你能够使用 1×1 的卷积层来减少 n_C ，但是不能减少 n_H, n_W

B. 你可以使用池化层减少 n_H, n_W ，但是不能减少 n_C

C. 你可以使用一个 1×1 的卷积层来减少 n_H, n_W 和 n_C

D. 你可以使用池化层减少 n_H, n_W 和 n_C

第 119 题

关于 Inception 网络下面哪些说法是正确的

A. 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。

B. Inception 块通常使用 1×1 的卷积来减少输入卷积的大小，然后再使用 3×3 和 5×5 的卷积。

C. 一个 Inception 块允许网络使用 $1 \times 1, 3 \times 3, 5 \times 5$ 的和卷积个池化层的组合。

D.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。

第 120 题

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.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。
- B.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。
- C.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。
- D.使用开源实现可以很简单的来实现复杂的卷积结构。

111-120题 答案

111.D 112.AC 113.B 114.B 115.B 116.BD 117.B 118.AB 119.BC 120.BCD

第三周 - 检测算法

第 121 题

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：

$$y = [p_c, b_x, b_y, b_h, b_w, c_1, c_2, c_3]$$



A. $y=[1, 0.3, 0.7, 0.3, 0.3, 0, 1, 0]$

B. $y=[1, 0.7, 0.5, 0.3, 0.3, 0, 1, 0]$

C. $y=[1, 0.3, 0.7, 0.5, 0.5, 0, 1, 0]$

D.y=[1, 0.3, 0.7, 0.5, 0.5, 1, 0, 0]

E.y=[0, 0.2, 0.4, 0.5, 0.5, 0, 1, 0]

第 122 题

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，

$$y = [p_c, b_x, b_y, b_h, b_w, c_1, c_2, c_3]$$



A.y=[1, ?, ?, ?, ?, 0, 0, 0]

B.y=[0, ?, ?, ?, ?, ?, ?, ?]

C.y=[?, ?, ?, ?, ?, ?, ?, ?]

D.y=[0, ?, ?, ?, ?, 0, 0, 0]

E.y=[1, ?, ?, ?, ?, ?, ?, ?]

第 123 题

你现在任职于自动化工厂中，您的系统将看到一罐饮料沿着传送带向下移动，你要对其进行拍照，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：



你的神经网络最合适的输出单位是什么？

- A.逻辑单元(用于分类图像中是否有罐头)
- B.逻辑单元, b_x 和 b_y
- C.逻辑单元, b_x, b_y, b_h (因为 b_w, b_h , 所以只需要一个就行了)
- D.逻辑单元, b_x, b_y, b_h, b_w

第 124 题

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？

- A.N
- B.2N
- C.3N
- D. N^2

第 125 题

在训练课程中描述的一个对象检测系统中，您需要一个训练集，其中包含许多要检测的对象的图片。但是，由于该算法可以学习自检测对象，因此不需要在训练集中提供边界框。

- A.正确
- B.错误

第 126 题

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。

- A.正确
- B.错误

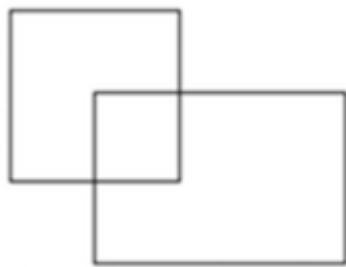
第 127 题

在YOLO算法中，在训练时，只有一个单元（该单元包含对象的中心/中点）负责检测这个对象

- A.正确
- B.错误

第 128 题

这两个框中IoU大小是多少？左上角的框是 2×2 大小，右下角的框是 2×3 大小，重叠部分是 1×1



- A.1/6
- B.1/9
- C.1/10
- D.以上都不是

第 129 题

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？



- A.3
- B.4
- C.5
- D.6
- E.7

第 130 题

假如你使用YOLO算法，使用 19×19 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？

- A. $19 \times 19 \times (25 \times 20)$
- B. $19 \times 19 \times (20 \times 25)$
- C. $19 \times 19 \times (5 \times 25)$
- D. $19 \times 19 \times (5 \times 20)$

121-130题 答案

121.A 122.B 123.B 124.B 125.B 126.B 127.A 128.B 129.C 130.C

第四周 - 特殊应用：人脸识别和神经风格转换

第 131 题

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。

- A.正确
- B.错误

第 132 题

在人脸验证中函数 $d(\text{img1}, \text{img2})$ 起什么作用？

- A.只需要给出一个人的图片就可以让网络认识这个人
- B.为了解决一次学习的问题
- C.这可以让我们使用softmax函数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
- D.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

第 133 题

为了训练人脸识别系统的参数，使用包含了10万个不同的人的10万张图片的数据集进行训练是合理的。

- A.正确
- B.错误

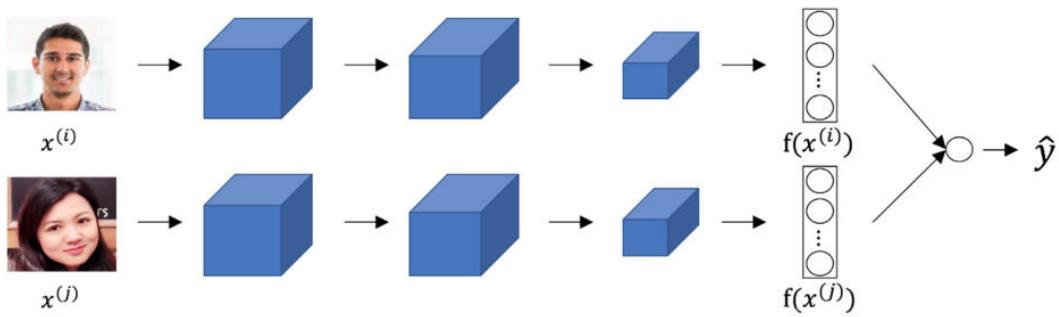
第 134 题

下面哪个是三元组损失的正确定义（请把 α 也考虑进去）？

- A. $\max(\|f(A) - f(P)\|^2 - \|f(A) - f(N)\|^2 + \alpha, 0)$
- B. $\max(\|f(A) - f(N)\|^2 - \|f(A) - f(P)\|^2 + \alpha, 0)$
- C. $\max(\|f(A) - f(N)\|^2 - \|f(A) - f(P)\|^2 - \alpha, 0)$
- D. $\max(\|f(A) - f(P)\|^2 - \|f(A) - f(N)\|^2 - \alpha, 0)$

第 135 题

在下图中的孪生卷积网络(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

- A.正确
- B.错误

第 136 题

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。

- A.正确
- B.错误

第 137 题

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

- A.正确
- B.错误

第 138 题

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

- A.正确
- B.错误

第 139 题

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？

- A.神经网络的参数
- B.生成图像G的像素值

- C.正则化参数
- D.内容图像C的像素值

第 140 题

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
 $32 \times 32 \times 32 \times 16$ ($32 \times 32 \times 16$ 有16个通道)， 对其使用 $32 \times 32 \times 3 \times 3 \times 3 \times 3$ 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？

- A. $30 \times 30 \times 30 \times 32$
- B.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
- C. $30 \times 30 \times 30 \times 16$

131-140题 答案

131.A 132.AB 133.B 134.A 135.A 136.A 137.B 138.A 139.B 140.A

课程五 - 序列模型

第一周 - 循环神经网络

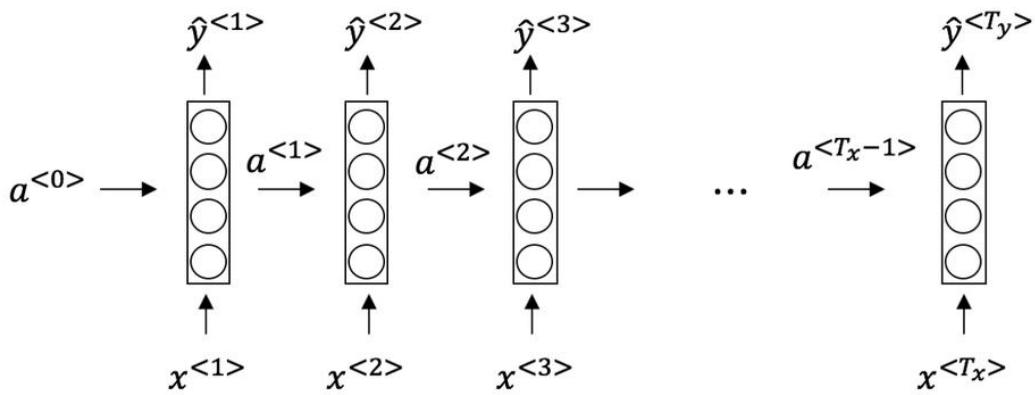
第 141 题

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？

- A. $\$x^{(i)}\$$
- B. $\$x^{(j)}\$$
- C. $\$x^{(j)}\$$
- D. $\$x^{(i)}\$$

第 142 题

看一下下面的这个循环神经网络：

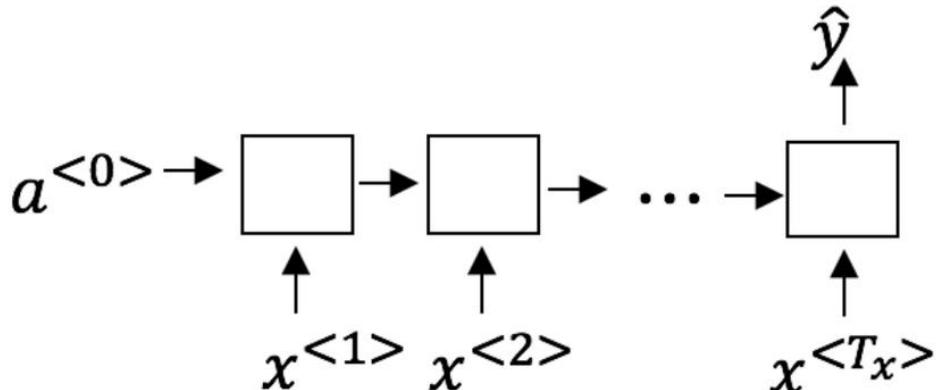


在下面的条件中，满足上图中的网络结构的参数是：

- A. $T_x = T_y$
- B. $T_x > T_y$
- C. $T_x < T_y$
- D. $T_x = 1$

第 143 题

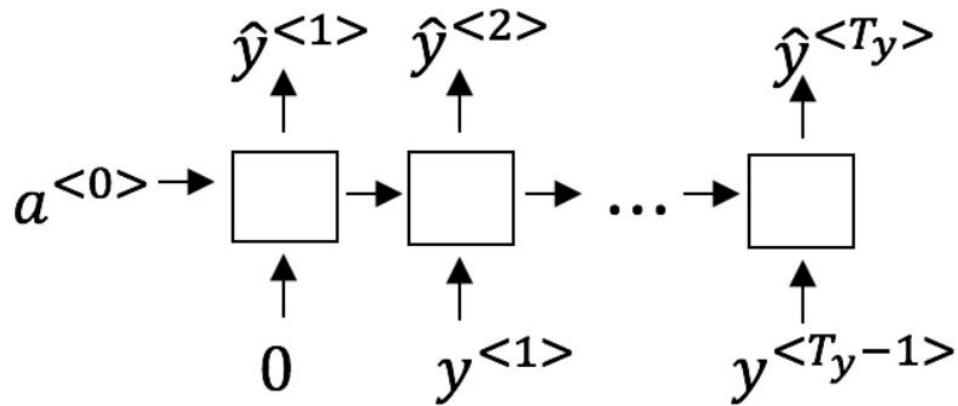
这些任务中的哪一个会使用多对一的RNN体系结构？



- A. 语音识别（输入语音，输出文本）
- B. 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）
- C. 图像分类（输入一张图片，输出对应的标签）
- D. 人声性别识别（输入语音，输出说话人的性别）

第 144 题

假设你现在正在训练下面这个RNN的语言模型：

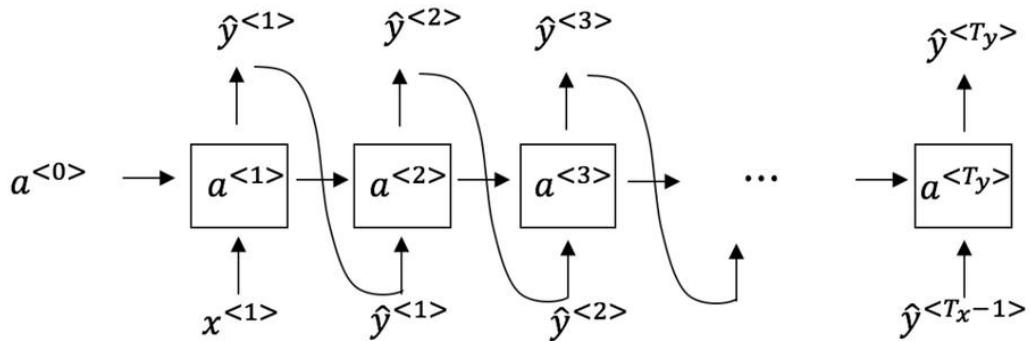


在 t 时，这个RNN在做什么？

- A. 计算 $P(y^{<1>}, y^{<2>}, \dots, y^{\{t\}})$
- B. 计算 $P(y^{\{t\}})$
- C. 计算 $P(y^{\{t\}} | y^{<1>}, y^{<2>}, \dots, y^{\{t-1\}})$
- D. 计算 $P(y^{\{t\}} | y^{<1>}, y^{<2>}, \dots, y^{<t-1>})$

第 145 题

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？

- A. (1) 使用RNN输出的概率，选择该时间步的最高概率单词作为 $\hat{y}^{\{t\}}$ ，(2)然后将训练集中的正确的单词传递到下一个时间步
- B. (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 $\hat{y}^{\{t\}}$ ，(2)然后将训练集中的实际单词传递到下一个时间步
- C. (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 $\hat{y}^{\{t\}}$ ，(2)然后将该选择的词传递给下一个时间步
- D. (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 $\hat{y}^{\{t\}}$ ，(2)然后将此选定单词传递给下一个时间步

第 146 题

你正在训练一个RNN网络，你发现你的权重与激活值都是“NaN”，下列选项中，哪一个是导致这个问题的最有可能的原因？

- A. 梯度消失
- B. 梯度爆炸
- C. ReLU函数作为激活函数 $g(\cdot)$ ，在计算 $g(z)$ 时， z 的数值过大了
- D. Sigmoid函数作为激活函数 $g(\cdot)$ ，在计算 $g(z)$ 时， z 的数值过大了

第 147 题

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， Γ_u 的维度是多少？

- A. 1
- B. 100
- C. 300
- D. 10000

第 148 题

这里有一些GRU的更新方程：

GRU

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

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

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

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

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

爱丽丝建议通过移除 Γ_u 来简化GRU，即设置 $\Gamma_u = 1$ 。贝蒂提出通过移除 Γ_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？

- A. 爱丽丝的模型（即移除 Γ_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。
- B. 爱丽丝的模型（即移除 Γ_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。
- C. 贝蒂的模型（即移除 Γ_r ），因为对于一个时间步而言，如果 $\Gamma_u \approx 0$ ，梯度可以通过时间步反向传播而不会衰减。

D. 贝蒂的模型（即移除 Γ_r ），因为对于一个时间步而言，如果 $\Gamma_u \approx 1$ ，梯度可以通过时间步反向传播而不会衰减。

第 149 题

这里有一些GRU和LSTM的方程：

GRU	LSTM
$\tilde{c}^{<t>} = \tanh(W_c[\Gamma_r * c^{<t-1>}, x^{<t>}] + b_c)$	$\tilde{c}^{<t>} = \tanh(W_c[a^{<t-1>}, x^{<t>}] + b_c)$
$\Gamma_u = \sigma(W_u[c^{<t-1>}, x^{<t>}] + b_u)$	$\Gamma_u = \sigma(W_u[a^{<t-1>}, x^{<t>}] + b_u)$
$\Gamma_r = \sigma(W_r[c^{<t-1>}, x^{<t>}] + b_r)$	$\Gamma_f = \sigma(W_f[a^{<t-1>}, x^{<t>}] + b_f)$
$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + (1 - \Gamma_u) * c^{<t-1>}$	$\Gamma_o = \sigma(W_o[a^{<t-1>}, x^{<t>}] + b_o)$
$a^{<t>} = c^{<t>}$	$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>}$
	$a^{<t>} = \Gamma_o * c^{<t>}$

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？

- A. Γ_u 与 $1 - \Gamma_u$
- B. Γ_u 与 Γ_r
- C. $1 - \Gamma_u$ 与 Γ_u
- D. Γ_r 与 Γ_u

第 150 题

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 $x^{<1>} , \dots, x^{<365>}$ ，这些数据是一个序列，你还收集了你的狗心情的数据 $y^{<1>} , \dots, y^{<365>}$ ，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？

- A. 双向RNN，因为在 t 日的情绪预测中可以考虑到更多的信息。
- B. 双向RNN，因为这允许反向传播计算中有更精确的梯度。
- C. 单向RNN，因为 $y^{<1>}$ 的值仅依赖于 $x^{<1>} , \dots, x^{<1>}$ ，而不依赖于 $x^{<1>} , \dots, x^{<365>}$ 。
- D. 单向RNN，因为 $y^{<1>}$ 的值只取决于 $x^{<1>}$ ，而不是其他天的天气。

141-150题 答案

141.A 142.A 143.BD 144.C 145.D 146.B 147.B 148.C 149.A 150.C

第二周 - 自然语言处理与词嵌入

第 151 题

假设你为10000个单词学习词嵌入，为了捕获全部范围的单词的变化以及意义，那么词嵌入向量应该是10000维的。

- A. 正确
- B. 错误

第 152 题

什么是t-SNE？

- A. 一种非线性降维算法
- B. 一种能够解决词向量上的类比的线性变换
- C. 一种用于学习词嵌入的监督学习算法
- D. 一个开源序列模型库

第 153 题

假设你下载了一个已经在很大的文本语料库上训练过的词嵌入的数据，然后你要用这个词嵌入来训练RNN并用于识别一段文字中的情感，判断这段文字的内容是否表达了“快乐”。

x(输入文本)	y (是否快乐)
我今天感觉很好！	1
我觉得很沮丧，因为我的猫生病了。	0
真的很享受这个！	1

那么即使“欣喜若狂”这个词没有出现在你的小训练集中，你的RNN也会认为“我欣喜若狂”应该被贴上 $y = 1$ 的标签。

- A. 正确
- B. 错误

第 154 题

对于词嵌入而言，下面哪一个（些）方程是成立的？

- A. $e_{boy} - e_{girl} \approx e_{brother} - e_{sister}$
- B. $e_{boy} - e_{girl} \approx e_{sister} - e_{brother}$

C. $e_{boy} - e_{brother} \approx e_{girl} - e_{sister}$

D. $e_{boy} - e_{brother} \approx e_{sister} - e_{girl}$

第 155 题

设 E 为嵌入矩阵， e_{1234} 对应的是词“1234”的独热向量，为了获得1234的词嵌入，为什么
不直接在Python中使用代码 $E * e_{1234}$ 呢？

- A. 因为这个操作是在浪费计算资源
- B. 因为正确的计算方式是 $E^T * e_{1234}$
- C. 因为它没有办法处理未知的单词 ($<UNK>$)
- D. 以上全都不对，因为直接调用 $E * e_{1234}$ 是最好的方案

第 156 题

在学习词嵌入时，我们创建了一个预测 $P(\text{target} | \text{context})$ 的任务，如果这个预测做的
不是很好那也是没有关系的，因为这个任务更重要的是学习了一组有用的嵌入词。

- A. 正确
- B. 错误

第 157 题

在 word2vec 算法中，你要预测 $P(t | c)$ ，其中 t 是目标词 (target word)， c 是语境词
(context word)。你应当在训练集中怎样选择 t 与 c 呢？

- A. c 与 t 应当在附近词中
- B. c 是在 t 前面的一个词
- C. c 是 t 之前句子中所有单词的序列
- D. c 是 t 之前句子中几个单词的序列

第 158 题

假设你有 1000 个单词词汇，并且正在学习 500 维的词嵌入，word2vec 模型使用下面的
softmax 函数：

$$P(t | c) = \frac{e^{\theta_t^T e_c}}{\sum_{t'=1}^{10000} e^{\theta_{t'}^T e_c}}$$

以下说法中哪一个 (些) 是正确的？

- A. θ_t 与 e_c 都是 500 维的向量
- B. θ_t 与 e_c 都是 10000 维的向量
- C. θ_t 与 e_c 都是通过 Adam 或 梯度下降等优化算法进行训练的

D. 训练之后, θ_t 应该非常接近 e_c , 因为 ttt 和 ccc 是一个词

第 159 题

假设你有 10000 个单词词汇, 并且正在学习 500 维的词嵌入, GloVe 模型最小化了这个目标:

$$\min \sum_{i=1}^{10,000} \sum_{j=1}^{10,000} f(X_{ij})(\theta_i^T e_j + b_i + b'_j - \log X_{ij})^2$$

以下说法中哪一个 (些) 是正确的?

- A. θ_i 与 e_j 应当初始化为 0
- B. θ_i 与 e_j 应当使用随机数进行初始化
- C. X_{ij} 是单词 i 在 j 中出现的次数
- D. 加权函数 $f(\cdot)$ 必须满足 $f(0) = 0$

第 160 题

你已经在文本数据集 m_1 上训练了词嵌入, 现在准备将它用于一个语言任务中, 对于这个任务, 你有一个单独标记的数据集 m_2 , 请记住, 使用词嵌入是一种迁移学习的形式。在以下那种情况中, 词嵌入会有帮助?

- A. $m_1 \gg m_2$
- B. $m_1 \ll m_2$

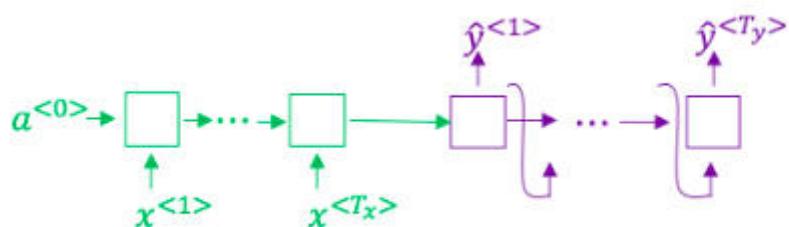
151-160 题 答案

151.B 152.A 153.A 154.AC 155.A 156.B 157.A 158.AC 159.BCD 160.A

第三周 - 序列模型与注意力机制

第 161 题

想一想使用如下的编码-解码模型来进行机器翻译:



这个模型是“条件语言模型”，编码器部分(绿色显示)的意义是建模中输入句子 x 的概率

- A. 正确
- B. 错误

第 162 题

在集束搜索中，如果增加集束宽度 b ，以下哪一项是正确的？

- A. 集束搜索将运行的更慢
- B. 集束搜索将使用更多的内存
- C. 集束搜索通常将找到更好地解决方案（比如：在最大化概率 $P(y|x)$ 上做的更好）
- D. 集束搜索将在更少的步骤后收敛

第 163 题

在机器翻译中，如果我们在不使用句子归一化的情况下使用集束搜索，那么算法会输出过短的译文。

- A. 正确
- B. 错误

第 164 题

假设你正在构建一个能够让语音片段 x 转为译文 y 的基于RNN模型的语音识别系统，你的程序使用了集束搜索来试着找寻最大的 $P(y|x)$ 的值 yyy 。在开发集样本中，给定一个输入音频，你的程序会输出译文 \hat{y} =“I'm building an A Eye system in Silly con Valley.”，人工翻译为 y^* =“I'm building an AI system in Silicon Valley.”

在你的模型中，

$$P(\hat{y} | x) = 1.09 * 10^{-7}$$

$$P(y^* | x) = 7.21 * 10^{-8}$$

那么，你会增加集束宽度 B 来帮助修正这个样本吗？

- A. 不会，因为 $P(y^* | x) \leq P(\hat{y} | x)$ 说明了问题在RNN，而不是搜索算法
- B. 不会，因为 $P(y^* | x) \leq P(\hat{y} | x)$ 说明了问题在搜索算法，而不是RNN
- C. 会的，因为 $P(y^* | x) \leq P(\hat{y} | x)$ 说明了问题在RNN，而不是搜索算法
- D. 会的，因为 $P(y^* | x) \leq P(\hat{y} | x)$ 说明了问题在搜索算法，而不是RNN

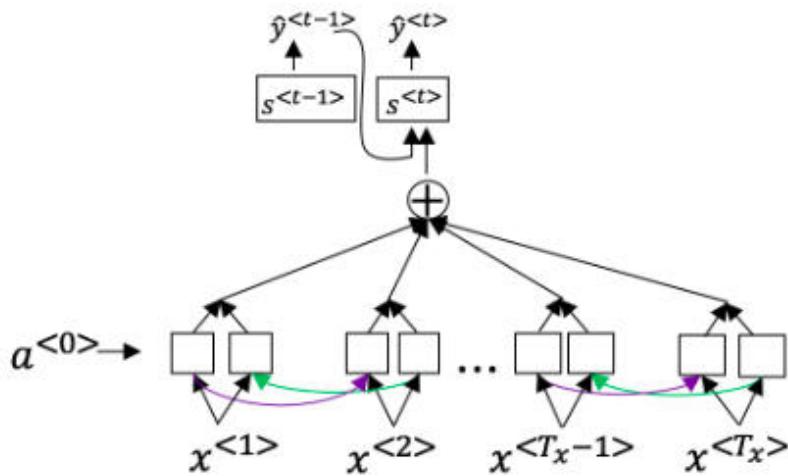
第 165 题

接着使用第4题的样本，假设你花了几周的时间来研究你的算法，现在你发现，对于绝大多数让算法出错的例子而言， $P(y^* | x) \leq P(\hat{y} | x)$ ，这表明你应该将注意力集中在改进搜索算法上，对吗？

- A. 正确
- B. 错误

第 166 题

回想一下机器翻译的模型：



除此之外，还有个公式 $\alpha^{<t,t'>} = \frac{\exp(e^{<t,t'>})}{\sum_{t'=1}^{T_x} \exp(e^{<t,t'>})}$

下面关于 $\alpha^{<t,t'>}$ 的选项那个（些）是正确的？

- A. 对于网络中与输出 $y^{<t>}$ 高度相关的 $\alpha^{<t,t'>}$ 而言，我们通常希望 $\alpha^{<t,t'>}$ 的值更大（请注意上标）
- B. 对于网络中与输出 $y^{<t>}$ 高度相关的 $\alpha^{<t,t'>}$ 而言，我们通常希望 $\alpha^{<t,t'>}$ 的值更大（请注意上标）
- C. $\sum_t \alpha^{<t,t'>} = 1$ （注意是和除以 t ）
- D. $\sum_{t'} \alpha^{<t,t'>} = 1$ （注意是和除以 t' ）

第 167 题

网络通过学习的值 $e^{<t,t'>}$ 来学习在哪里关注“关注点”，这个值是用一个小的神经网络的计算出来的：

这个神经网络的输入中，我们不能将 $s^{<t>}$ 替换为 $s^{<t-1>}$ 这是因为 $s^{<t>}$ 依赖于 $\alpha^{<t,t'>}$ ，而 $\alpha^{<t,t'>}$ 又依赖于 $e^{<t,t'>}$ ；所以在我们需要评估这个网络时，我们还没有计算出 s^t

- A. 正确
- B. 错误

第 168 题

与题1中的编码-解码模型（没有使用注意力机制）相比，我们希望有注意力机制的模型在下面的情况下有着最大的优势：

- A. 输入序列的长度 T_x 比较大
- B. 输入序列的长度 T_x 比较小

第 169 题

在CTC模型下，不使用“空白”字符（_）分割的相同字符串将会被折叠。那么在CTC模型下，以下字符串将会被折叠成什么样子？_c_oo_o_kk__booooo_oo_kkk

- A. cokbook
- B. cookbook
- C. cook book
- D. cooookboooooooookkk

第 170 题

在触发词检测中， $x^{<t>}$ 是：

- A. 时间 t 时的音频特征（就像是频谱特征一样）
- B. 第 t 个输入字，其被表示为一个独热向量或者一个字嵌入
- C. 是否在第 t 时刻说出了触发词
- D. 是否有人在第 t 时刻说完了触发词

161-170题 答案

161.B 162.ABC 163.A 164.A 165.A 166.AD 167.A 168.A 169.B 170.A