深度学习—大厂笔试题

一、判断

1、训练 CNN 时，可以对输入进行旋转、平移、缩放等预处理提高模型泛化能力。这么说是对，还是不对？（对）

> 输入进行旋转、平移、缩放等预处理相当于做了数据增强，提升泛化能力

2、深度学习与机器学习算法之间的区别在于，后者过程中无需进行特征提取工作，也就是说，我们建议在进行深度学习过程之前要首先完成特征提取的工作。这种说法是：（错）

> 机器学习和深度学习都需要特征提取工作
>
> 深度学习是直接从数据中自动学习到特征，无需提前人工进行特征提取

二、单选

1、下列哪个函数**不可以**做激活函数 （ D ）

A、y = tanh(x)
B、y = sin(x)
C、y = max(x,0)
D、y = 2x

> 神经网络的激活函数需要代入非线性性，这样能使神经网络能拟合任何线性和非线性函数的能力。

2、列哪些项所描述的相关技术是**错误**的 （ C ）

A、AdaGrad 使用的是一阶差分(first order differentiation)
B、L-BFGS 使用的是二阶差分(second order differentiation)
C、AdaGrad 使用的是二阶差分

> AdaGrad 是梯度下降法，用的是一阶导数信息，
>
> L-BFGS 是拟牛顿法，在牛顿法的基础上发展而来的，用到了二阶导数信息。

3、假设我们有一个使用 ReLU 激活函数(ReLU activation function)的神经网络，假如我们把 ReLU 激活替换为线性激活，那么这个神经网络能够模拟出同或函数(XNOR function)吗？（ D ）

A、可以
B、不好说
C、不一定
D、不能

> 加入激活函数，模型具有了模拟非线性函数的作用，如果被替换成了线性，那么模型就不能进行其他非线性函数的模拟。

4、深度学习是当前很热门的机器学习算法，在深度学习中，涉及到大量的矩阵相乘，现在需要计算三个稠密矩阵 A，B，C 的乘积 ABC,假设三个矩阵的尺寸分别为 m* n，n* p，p* q，且 m <n < p < q，以下计算顺序效率最高的是 （ A ）

A、(AB)C
B、AC(B)
C、A(BC)
D、所以效率都相同

> 注：不论矩阵大小如何，A 选项划分是效率最高的
> 考查矩阵相乘的效率问题，即需要计算的乘法和加法的次数之和。当 m,n,p,q 较大时，加法忽略不计。任意两个大小分别为 a*b*,*b*c 的矩阵相乘，需要乘法的次数为 a * b * c
> A 选项的乘法计算次数，m * n * p+m * p * q；
> B 选项不满足矩阵相乘规则，舍弃；
> C 选项的乘法计算次数，n * p * q + m * n * q

5、梯度下降算法的正确步骤是什么 （ D ）

a.计算预测值和真实值之间的误差

b.重复迭代，直至得到网络权重的最佳值

c.把输入传入网络，得到输出值

d.用随机值初始化权重和偏差

e.对每一个产生误差的神经元，调整相应的（权重）值以减小误差

A、abcde
B、edcba
C、cbaed
D、dcaeb

6、下列哪一项属于**特征学习算法**（representation learning algorithm）（ C ）

A、K 近邻算法
B、随机森林
C、神经网络
D、都不属于

> 特征学习能够替代手动特征工程，而只有神经网络是自动学习特征

7、下面哪项操作能实现跟神经网络中 Dropout 的类似效果 （ B ）

A、Boosting
B、Bagging
C、Stacking
D、Mapping

- Bagging：独立的集成多个模型，每个模型有一定的差异，最终综合有差异的模型的结果，获得学习最终的结果；
- Boosting（增强集成学习）：集成多个模型，每个模型都在尝试增强（Boosting）整体的效果；
- Stacking（堆叠）：集成 k 个模型，得到 k 个预测结果，将 k 个预测结果再传给一个新的算法，得到的结果为集成系统最终的预测结果；

8、caffe 中基本的计算单元为 （ B ）

A、blob
B、layer
C、net
D、solver

Blob 是 Caffe 的**基本存储单元**
Layer 是 Caffe 的**基本计算单元**

9、阅读以下文字：假设我们拥有一个已完成训练的、用来解决车辆检测问题的深度神经网络模型，训练所用的数据集由汽车和卡车的照片构成，而训练目标是检测出每种车辆的名称（车辆共有 10 种类型）。现在想要使用这个模型来解决另外一个问题，问题数据集中仅包含一种车（福特野马）而目标变为定位车辆在照片中的位置 （ B ）

A、除去神经网络中的最后一层，冻结所有层然后重新训练
B、对神经网络中的最后几层进行微调，同时将最后一层（分类层）更改为回归层
C、使用新的数据集重新训练模型
D、所有答案均不对

一个是分类任务，一个是检测任务

10、有关深度学习加速芯片，以下的说法中不正确的是：（ C ）

A、GPU 既可以做游戏图形加速，也可以做深度学习加速
B、用于玩游戏的高配置显卡，也可以用于深度学习计算。
C、Google TPU 已经发展了三代，它们只能用于推断（Inference）计算，不能用于训练（Training）计算
D、FPGA 最早是作为 CPLD 的竞争技术而出现的

11、考虑以下问题：假设我们有一个 5 层的神经网络，这个神经网络在使用一个 4GB 显存显卡时需要花费 3 个小时来完成训练。而在测试过程中，单个数据需要花费 2 秒的时间。 如果我们现在把架构变换一下，当评分是 0.2 和 0.3 时，分别在第 2 层和第 4 层添加 Dropout，那么新架构的**测试所用时间**会变为多少？（ C ）

A、少于 2s
B、大于 2s
C、仍是 2s
D、说不准

> 在架构中添加 Dropout 这一改动仅会影响训练过程，而并不影响测试过程。
> Dropout 是在训练过程中以一定的概率使神经元失活，即输出为 0，以提高模型的泛化能力，减少过拟合。Dropout 在训练时采用，是为了减少神经元对部分上层神经元的依赖，类似将多个不同网络结构的模型集成起来，减少过拟合的风险。而在测试时，应该用整个训练好的模型，因此不需要 dropout。
> Batch Normalization（BN），就是在深度神经网络训练过程中使得每一层神经网络的输入保持相近的分布。
> 对于 BN，在训练时，是对每一批的训练数据进行归一化，也即用每一批数据的均值和方差。而在测试时，比如进行一个样本的预测，就并没有 batch 的概念，因此，这个时候用的均值和方差是全量训练数据的均值和方差，这个可以通过移动平均法求得。

12、关于 Attention-based Model，下列说法正确的是 （ A ）

A、相似度度量模型
B、是一种新的深度学习网络
C、是一种输入对输出的比例模型
D、都不对

> Attention-based Model 其实就是一个相似性的度量，当前的输入与目标状态越相似，那么在当前的输入的权重就会越大，说明当前的输出越依赖于当前的输入。严格来说，Attention 并算不上是一种新的 model，而仅仅是在以往的模型中加入 attention 的思想，所以 Attention-based Model 或者 Attention Mechanism 是比较合理的叫法，而非 Attention Model。

13、下列的哪种方法可以用来降低深度学习模型的过拟合问题？（ D ）

①增加更多的数据　　②使用数据扩增技术(data augmentation)

③使用归纳性更好的架构　④ 正规化数据　　⑤ 降低架构的复杂度

A、1 4 5
B、1 2 3
C、1 3 4 5
D、所有项目都有用

> 防止过拟合的几种方法：
> 引入正则化
> Dropout
> 提前终止训练
> 增加样本量

14、假设我们有一个如下图所示的隐藏层。隐藏层在这个网络中起到了一定的降纬作用。假如现在我们用另一种维度下降的方法，比如说主成分分析法（PCA）来代替这个隐藏层，那么，这两者的输出效果是一样的吗 （ B ）

A、是
B、否

> PCA 降维的特点在于使用矩阵分解求特征值的方式，提取的是数据分布方差比较大的方向，提取的是主要成分；hidden layer 主要是点乘+非线性变换，目的是特征的提取，转换

15、假设你需要调整超参数来最小化代价函数（cost function），会使用下列哪项技术？（ D ）

A、穷举搜索
B、随机搜索
C、Bayesian 优化
D、都可以

> 穷举搜索法，随机搜索法，贝叶斯优化都可以优化超参数，各有优劣。
> 所以 ABC 三种都可实现调整优化超参数。

16、现有一 1920 * 1080 的单通道图像，每个像素用 float32 存储，对其进行 4 个 3 * 3 核的卷积(无 padding)，卷积核如下：

```
1 1 1 1 1 1 0 0 1 1 0 1 0
2 1 0 1 1 1 1 1 1 1 1 1 1
3 1 1 1 0 1 1 1 1 0 0 1 0
```

若原图像由于量化问题出现了 100 个 INFINITY（无穷），而其他的值都在(-1,1)区间内，则卷积的结果至少有多少个 NaN? （ B ）

A、256
B、284
C、296
D、324

> 因为 INFINITY 与 0 相乘为 NaN，与除了乘以 0 以外的任何四则运算，得到的结果仍然是 INFINITY。
> 题目问的是至少有多少个，那么我们就考虑 nan 最少的情况下的 INFINITY 分布位置，

17、提升卷积核(convolutional kernel)的大小会显著提升卷积神经网络的性能，这种说法是 （ B ）

A、正确的
B、错误的

> 卷积核的大小是一个超参数，也就意味着改变它有可能增强也可能降低模型的性能。
> 卷积核越大，特征越多，参数也越多，性能是下降的，因此才有了以多个小卷积核代替大卷积核的做法

18、如果我们用了一个过大的学习速率会发生什么？（ D ）

A、神经网络会收敛
B、不好说
C、都不对
D、神经网络不会收敛

> 学习率过小，收敛太慢，学习率过大，震荡不收敛
> 如果使用自适应优化器，训练到后期学习率是会变小的

19、神经网络模型（Neural Network）因受人类大脑的启发而得名，神经网络由许多神经元（Neuron）组成，每个神经元接受一个输入，对输入进行处理后给出一个输出，如下图所示。请问下列关于神经元的描述中，哪一项是正确的？（ E ）

A、每个神经元可以有一个输入和一个输出
B、每个神经元可以有多个输入和一个输出
C、每个神经元可以有一个输入和多个输出
D、每个神经元可以有多个输入和多个输出
E、上述都正确

20、BatchNorm 层对于 input batch 会统计出 mean 和 variance 用于计算 EMA。如果 input batch 的 shape 为(B, C, H, W)，统计出的 mean 和 variance 的 shape 为:（ B ）

A、B * 1 * 1 * 1
B、1 * C * 1 * 1
C、B * C * 1 * 1
D、1 * 1 * 1 * 1

> B 代表图像的 batch，即多少张图像一个 batch。C 代表图像的通道数。
> BN 是对多张图像的同一通道做 Normalization
> 所以有多少通道就有多少个 mean 和 variance

21、ResNet-50 有多少个卷积层?（ B ）

A、48
B、49
C、50
D、51

> ResNet-50（50-layer 指的是 50 层网络）首先有个输入 7x7x64 的卷积，1 层
> 然后经过 3 + 4 + 6 + 3 = 16 个 building block，每个 block 为 3 层，所以有 16 x 3 = 48 层，
> 最后有个 fc 层(用于分类) 1 层
> 这 50 层里只有一个全连接层，剩下的都是卷积层，所以是 50-1=49

22、下列哪一项在神经网络中引入了非线性 （ B ）

A、随机梯度下降
B、修正线性单元（ReLU）
C、卷积函数
D、以上都不正确

> 线性修正单元其实是非线性激活函数
> 卷积运算为线性，单层的神经网络都是线性运算，只有加了激活函数之后，网络才是非线性的运算。
> 多层神经网络不加激活函数也是线性变换。

23、如果增加多层感知机（Multilayer Perceptron）的隐藏层 层数，分类误差便会减小。这种陈述正确还是错误？（ B ）

A、正确
B、错误

> 过拟合可能会导致错误增加

24、考虑某个具体问题时，你可能只有少量数据来解决这个问题。不过幸运的是你有一个类似问题已经预先训练好的神经网络。可以用下面哪种方法来利用这个预先训练好的网络？ （ C ）

A、把除了最后一层外所有的层都冻结，重新训练最后一层
B、对新数据重新训练整个模型
C、只对最后几层进行调参(fine tune)
D、对每一层模型进行评估，选择其中的少数来用

> 不同数据集下使用微调：
> 数据集 1-数据量少，但数据相似度非常高：在这种情况下，我们所做的只是修改最后几层或最终的 softmax 图层的输出类别。
> 数据集 2-数据量少，数据相似度低：在这种情况下，我们可以冻结**预训练模型**的初始层（比如 k 层），并再次训练剩余的（n-k）层。由于新数据集的相似度较低，因此根据新数据集对较高层进行重新训练具有重要意义。
> 数据集 3-数据量大，数据相似度低：在这种情况下，由于我们有一个大的数据集，我们的神经网络训练将会很有效。但是，由于我们的数据与用于训练我们的预训练模型的数据相比有很大不同，使用预训练模型进行的预测不会有效。因此，最好根据你的数据**从头开始训练神经网络**（Training from scatch）。
> 数据集 4-数据量大，数据相似度高：这是理想情况。在这种情况下，预训练模型应该是最有效的。使用模型的最好方法是保留模型的体系结构和模型的初始权重。然后，我们可以使用在预先训练的模型中的权重来重新训练该模型。

25、下列哪个神经网络结构会发生权重共享？（ D ）

A、卷积神经网络
B、循环神经网络
C、全连接神经网络
D、选项 A 和 B

> 权值共享就是说，给一张输入图片，用一个卷积核去扫这张图，卷积核里面的数就叫权重，这张图每个位置是被同样的卷积核扫的，所以权重是一样的，也就是共享。

26、输入图片大小为 200×200，依次经过一层卷积（kernel size 5×5，padding 1，stride 2），pooling（kernel size 3×3，padding 0，stride 1），又一层卷积（kernel size 3×3，padding 1，stride 1）之后，输出特征图大小为 （ C ）

A、95
B、96
C、97
D、98

> kernel_size 就是卷积核的长度
> padding：认为的扩充图片， 在图片外围补充一些像素点，把这些像素点初始化为 0
> stride 是卷积步长
> 卷积向下取整，池化向上取整，
> padding = “value”，N = [(W-K+2P)/S]+1，这里表示的是向下取整再加 1
> 输出高度 = （输入高度 - Kernel 高度 + 2 * padding）/ 步长 stride + 1
> 输出宽度 = （输入宽度 - Kernel 宽度 + 2 * padding）/ 步长 stride + 1
> 除法都为向下取整
> （200 - 5 + 2 * 1）/2+1 为 99.5，取 99
> （99 - 3 +2 * 0）/1+1 为 97
> （97 - 3 + 2 * 1）/1+1 为 97

27、已知：（1）大脑是有很多个叫做神经元的东西构成，神经网络是对大脑的简单的数学表达。（2）每一个神经元都有输入、处理函数和输出。（3）神经元组合起来形成了网络，可以拟合任何函数。（4）为了得到最佳的神经网络，我们用梯度下降方法不断更新模型。给定上述关于神经网络的描述，什么情况下神经网络模型被称为深度学习模型？（ A ）

A、加入更多层，使神经网络的深度增加
B、有维度更高的数据
C、当这是一个图形识别的问题时
D、以上都不正确

28、下图显示了训练过的 3 层卷积神经网络准确度，与参数数量(特征核的数量)的关系。从图中趋势可见（先上升，后下降），如果增加神经网络的宽度，精确度会增加到一个特定阈值后，便开始降低。造成这一现象的可能原因是什么？（ C ）

A、即使增加卷积核的数量，只有少部分的核会被用作预测
B、当卷积核数量增加时，神经网络的预测能力（Power）会降低
C、当卷积核数量增加时，导致过拟合
D、以上都不正确

> 过拟合在训练过程中不会导致精度下降，只有在验证阶段，用验证集和训练集做对比的时候，会导致训练集的精度高于验证集的精度，导致模型的泛化能力下降。
> 卷积核的作用为提取图像特征，当卷积核增加时，过多的学习了图象中的特征，会导致过拟合

29、假设你有 5 个大小为 7x7、边界值为 0 的卷积核，同时卷积神经网络第一层的深度为 1。此时如果你向这一层传入一个维度为 224x224x3 的数据，那么神经网络下一层所接收到的数据维度是多少？（ A ）

A、218x218x5
B、217x217x8
C、217x217x3
D、220x220x5

（W−F+2P）/S + 1
其中 W 为输入，F 为卷积核，P 为 pading 值，S 为步长
（224 - 7 + 2 * 0）/ 1 + 1 为 218，取 218

30、混沌度(Perplexity)是一种常见的应用在使用深度学习处理 NLP 问题过程中的评估技术，关于混沌度，哪种说法是正确的？（ B）

A、混沌度没什么影响
B、混沌度越低越好
C、混沌度越高越好
D、混沌度对于结果的影响不一定

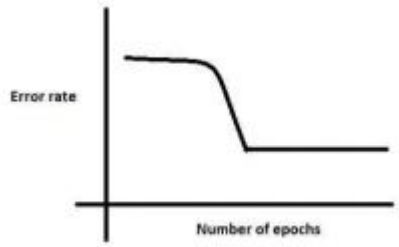
混沌度（不确定性程度），越低越好。

31、在 CNN 网络中，图 A 经过核为 3x3，步长为 2 的卷积层，ReLU 激活函数层，BN 层，以及一个步长为 2，核为 2 * 2 的池化层后，再经过一个 3 * 3 的的卷积层，步长为 1，此时的感受野是（ D ）

A、10
B、11
C、12
D、13

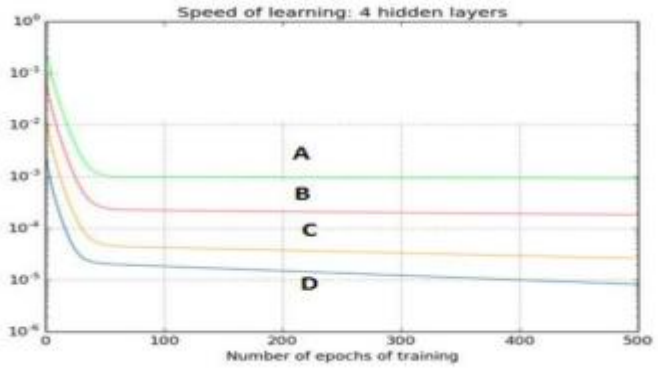
感受野：现在的一个像素对应原来的多少个像素
倒推上一层感受野 L_{n-1} =（ L_n -1）* S_{n-1}+ K_{n-1} ，S 和 K 分别是 stride（步长）和 kernel size（卷积核大小）
卷积层 3x3，步长 1： 1（1-1）+3=3*3；
池化层 2x2，步长 2：2（3-1）+2=6*6
卷积层 3x3，步长 2：2（6-1）+3=13*13

32、在训练神经网络时，损失函数(loss)在最初的几个 epochs 时没有下降，可能的原因是？（ D）



A、学习率(learning rate)太低
B、正则参数太高
C、陷入局部最小值
D、以上都有可能

33、下图是一个利用 sigmoid 函数作为激活函数的含四个隐藏层的神经网络训练的梯度下降图。这个神经网络遇到了梯度消失的问题。下面哪个叙述是正确的？（ A ）



A、第一隐藏层对应 D，第二隐藏层对应 C，第三隐藏层对应 B，第四隐藏层对应 A
B、第一隐藏层对应 A，第二隐藏层对应 C，第三隐藏层对应 B，第四隐藏层对应 D
C、第一隐藏层对应 A，第二隐藏层对应 B，第三隐藏层对应 C，第四隐藏层对应 D
D、第一隐藏层对应 B，第二隐藏层对应 D，第三隐藏层对应 C，第四隐藏层对应 A

由于梯度反向传播，在梯度消失情况下越接近输入层，其梯度越小；在梯度爆炸的情况下越接近输入层，其梯度越大。
损失函数曲线越平，就说明梯度消失越厉害，在前向传播中，越接近输出层，梯度越接近 0，所以最先消失的一定是离输出层最近的。
由于反向传播算法进入起始层，学习能力降低，这就是梯度消失。换言之，梯度消失是梯度在前向传播中逐渐减为 0，按照图标题所说，四条曲线是 4 个隐藏层的学习曲线，那么第一层梯度最高(损失函数曲线下降明显)，最后一层梯度几乎为零(损失函数曲线变成平直线). 所以 D 是第一层, A 是最后一层。

34、基于二次准则函数的 H-K 算法较之于感知器算法的优点是 (B)

A、计算量小
B、可以判别问题是否线性可分
C、其解完全适用于非线性可分的情况

> HK 算法的思想是在最小均方误差准则下求得权矢量。相对于感知器算法的优点在于，它适用于线性可分和非线性可分的情况。
>
> 对于线性可分的情况，给出最优权矢量；对于非线性可分的情况，能够判别出来，以退出迭代过程。

35、有关深度神经网络的训练（Training）和推断（Inference），以下说法中不正确的是：（ B ）

A、将数据分组部署在不同 GPU 上进行训练能提高深度神经网络的训练速度。
B、TensorFlow 使用 GPU 训练好的模型，在执行推断任务时，也必须在 GPU 上运行。
C、将模型中的浮点数精度降低，例如使用 float16 代替 float32，可以压缩训练好的模型的大小。
D、GPU 所配置的显存的大小，对于在该 GPU 上训练的深度神经网络的复杂度、训练数据的批次规模等，都是一个无法忽视的影响因素。

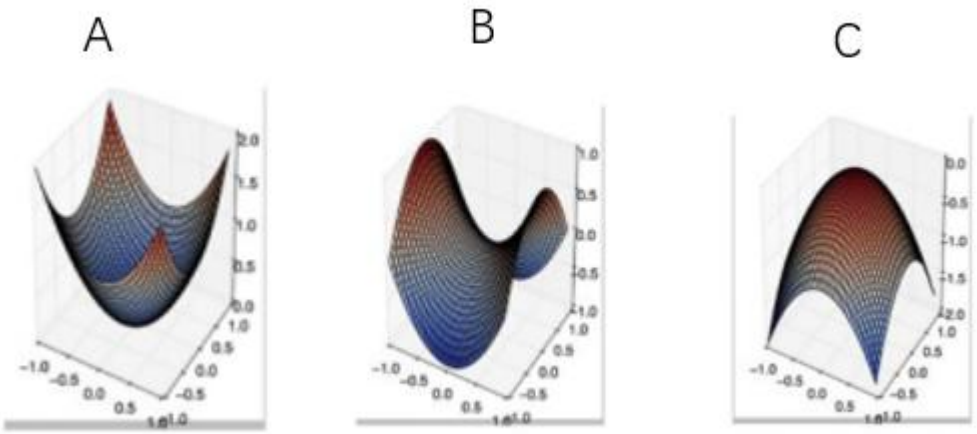
> GPU 是为了大量并行计算，减少训练花费时间。并不存在 GPU 训练的模型和 CPU 训练的模型不同的说法，只是计算方式的不同。

36、当在卷积神经网络中加入池化层(pooling layer)时，变换的不变性会被保留，是吗？（ C ）

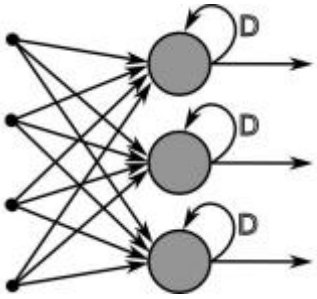
A、不知道
B、看情况
C、是
D、否

> 池化算法比如取最大值/取平均值等，都是输入数据旋转后结果不变，所以多层叠加后也有这种不变性。

37、在下面哪种情况下，一阶梯度下降不一定正确工作（可能会卡住）？（ B ）



38、构建一个神经网络，将前一层的输出和它自身作为输入。（ A ）



下列哪一种架构有反馈连接？

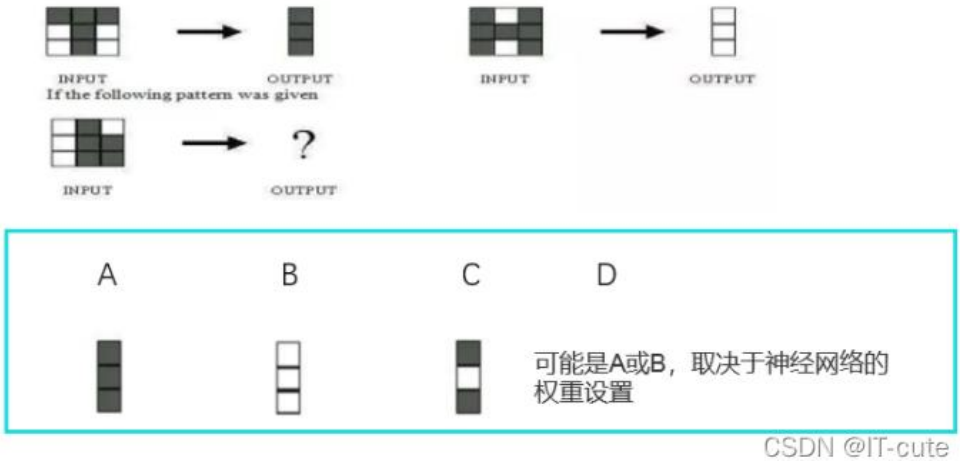
A、循环神经网络
B、卷积神经网络
C、限制玻尔兹曼机
D、都不是

反馈即是把当前输出回传到输入项，和下一时刻的输入共同决定下一时刻的输出。

39、sigmoid 导数为 （ D ）

A、f（z）
B、f(1-z)
C、f(1+z)f(1-z)
D、f(z)(1-f(z))

40、下图所示的网络用于训练识别字符 H 和 T，如下所示 （ D ）



训练出来的就是个二分类器，简单举例，如果分类器认为中间一列全为黑就是 T，否则就是 H，那么输出全黑；如果分类器认为右下角为黑就是 H，否则就是 T，那么输出全白，具体还得看分类器权重是怎么学的

41、CNN 常见的 Loss 函数不包括以下哪个 （ D ）

A、softmax_loss
B、sigmoid_loss
C、Contrastive_Loss（对比损失）
D、siamese_loss

> 在传统的 siamese network 中一般使用 Contrastive Loss 作为损失函数，这种损失函数可以有效的处理孪生神经网络中的 paired data 的关系。

42、在感知机中（Perceptron）的任务顺序是什么？（ D ）

1、随机初始化感知机的权重 2、去到数据集的下一批（batch）
3、如果预测值和输出不一致，则调整权重 4、对一个输入样本，计算输出值

A、1, 2, 3, 4
B、4, 3, 2, 1
C、3, 1, 2, 4
D、1, 4, 3, 2

43、在一个神经网络中，下面哪种方法可以用来处理过拟合？（ D ）

A、Dropout
B、分批归一化(Batch Normalization)
C、正则化(regularization)
D、都可以

44、在选择神经网络的深度时，下面哪些参数需要考虑？（ C ）

1 神经网络的类型(如 MLP,CNN) 2 输入数据 3 计算能力(硬件和软件能力决定)
4 学习速率 5 映射的输出函数

A、1,2,4,5
B、2,3,4,5
C、都需要考虑
D、1,3,4,5

45、当数据过大以至于无法在 RAM 中同时处理时，哪种梯度下降方法更加有效？（ A ）

A、随机梯度下降法(Stochastic Gradient Descent)
B、不知道
C、整批梯度下降法(Full Batch Gradient Descent)
D、都不是

> 梯度下降法：随机梯度下降(每次用一个样本)、小批量梯度下降法(每次用一小批样本算出总损失，因而反向传播的梯度折中)、全批量梯度下降法则一次性使用全部样本。
> 这三个方法，对于全体样本的损失函数曲面来说，梯度指向一个比一个准确．但是在工程应用中,受到内存/磁盘 IO 的吞吐性能制约，若要最小化梯度下降的实际运算时间，需要在梯度方向准确性和数据传输性能之间取得最好的平衡．所以，对于数据过大以至于无法在RAM中同时处理时，RAM每次只能装一个样本，那么只能选随机梯度下降法。

46、批规范化(Batch Normalization)的好处都有啥 （ A ）

A、让每一层的输入的范围都大致固定
B、它将权重的归一化平均值和标准差
C、它是一种非常有效的反向传播(BP)方法
D、这些均不是

> BN 是对数据进行归一化，而不是权重
>
> batch normalization 的作用是将经过 activation 的输出特征图归一化接近均值为 0，方差为 1 的正太分布。

47、在一个神经网络中，知道每一个神经元的权重和偏差是最重要的一步。如果知道了神经元准确的权重和偏差，便可以近似任何函数，但怎么获知每个神经的权重和偏移呢？（ B ）

A、搜索每个可能的权重和偏差组合，直到得到最佳值
B、赋予一个初始值，然后检查跟最佳值的差值，不断迭代调整权重
C、随机赋值，听天由命
D、以上都不正确的

> 深度学习是根据梯度下降优化参数的

三、多选

1、深度学习中的激活函数需要具有哪些属性 （ A B D ）

A、计算简单
B、非线性
C、具有饱和区
D、几乎处处可微

1. 非线性：即导数不是常数
2. 几乎处处可微（即仅在有限个点处不可微）：保证了在优化中梯度的可计算性
3. 计算简单：激活函数在神经网络前向的计算次数与神经元的个数成正比，因此简单的非线性函数自然更适合用作激活函数。
4. 非饱和性（saturation）：饱和指的是在某些区间梯度接近于零（即梯度消失）,使得参数无法继续更新的问题。
5. 单调性（monotonic）：即导数符号不变。个人理解，单调性使得在激活函数处的梯度方向不会经常改变，从而让训练更容易收敛。
6. 输出范围有限：有限的输出范围使得网络对于一些比较大的输入也会比较稳定，但这导致了前面提到的梯度消失问题，而且强行让每一层的输出限制到固定范围会限制其表达能力。
7. 接近恒等变换（identity）：即约等于 x。这样的好处是使得输出的幅值不会随着深度的增加而发生显著的增加，从而使网络更为稳定，同时梯度也能够更容易地回传。
8. 参数少：大部分激活函数都是没有参数的。

2、googlenet 提出的 Inception 结构优势有（ A D ）

A、保证每一层的感受野不变，网络深度加深，使得网络的精度更高
B、使得每一层的感受野增大，学习小特征的能力变大
C、有效提取高层语义信息，且对高层语义进行加工，有效提高网络准确度
D、利用该结构有效减轻网络的权重

3、下列是 caffe 支持的 loss 优化的方法的是 （ A B C D ）

A、Adam
B、SGD
C、AdaDelta
D、Nesterov

> caffe 六种优化方法：
> Stochastic Gradient Descent (type: “SGD”), 随机梯度下降
> AdaDelta (type: “AdaDelta”) 自适应学习率
> Adaptive Gradient (type: “AdaGrad”) 自适应梯度
> Adam (type: “Adam”) 自适应学习，推荐使用
> Nesterov’s Accelerated Gradient (type: “Nesterov”) 加速梯度法
> RMSprop (type: “RMSProp”)

4、深度学习中，以下哪些方法可以降低模型过拟合？（ A B D ）

A、增加更多的样本
B、Dropout
C、增大模型复杂度，提高在训练集上的效果
D、增加参数惩罚

> 放置模型过拟合：
> 1.引入正则化（参数范数惩罚）
> 2.Dropout
> 3.提前终止训练
> 4.增加样本量
> 5.参数绑定与参数共享
> 6.辅助分类节点(auxiliary classifiers)
> 7.Batch Normalization