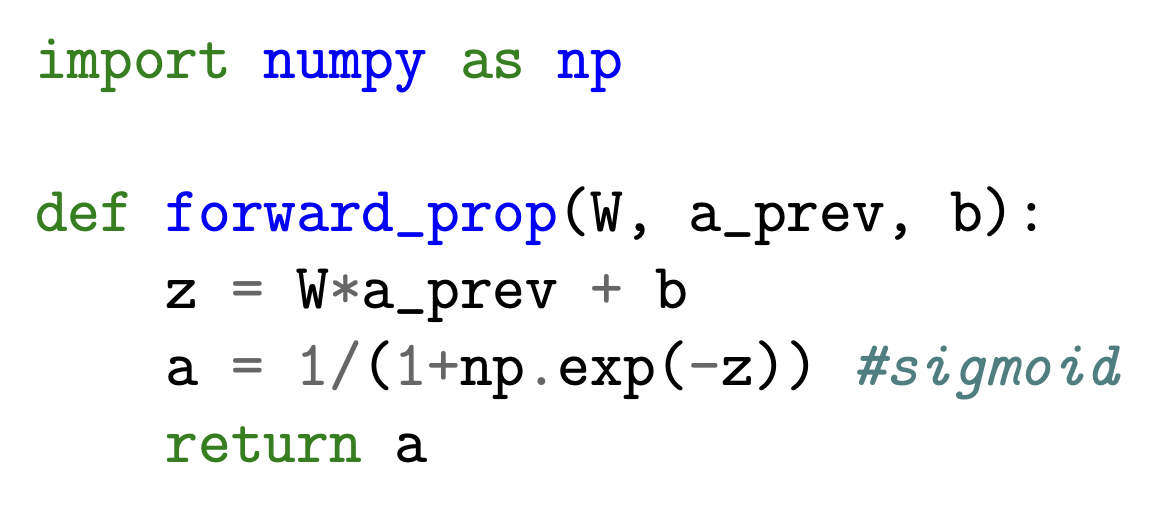
作业3

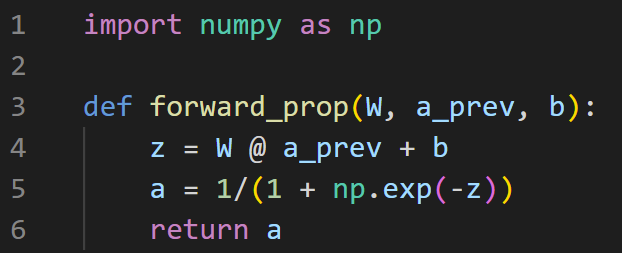
211300024 石睿

1，下面是计算一个layer的“forward propagation”操作的函数，这一层使用了sigmoid activation函数，请找出代码的错误并写出修正的结果



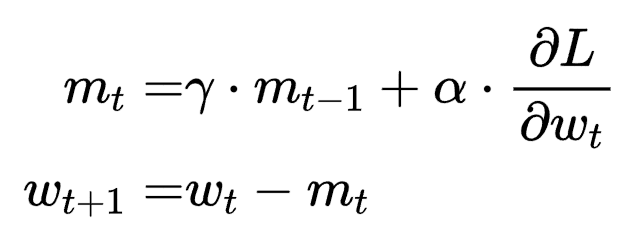
**答案：**

由神经网络forward propagation操作可知，如若抽象成一个函数，则其中的W和a\_prev应分别是一个向量，b是一个标量。而向量间点积应使用矩阵乘法操作符“@”。综上，代码应修改成



2，请解释Momentum相比于标准的Gradient Descent, 它是如何加速学习过程的，并解释它在优化求解中的优势/作用？

**答案：**



本式即引入Momentum后的参数更新情况。其中m是动量，同时受到历史动量以及当前梯度的影响。下标t表示是t时刻的动量，γ代表对历史动量信息的参考程度，一般取0.9。

1. **加速**

1.1：t时刻动量可以不断累积t时刻之前的动量信息，期望情况下动量会逐步增大（同时取决于当前梯度）

1.2：相比于标准的梯度下降算法，如若遇到函数较平坦的位置（梯度很小），则更新较慢。但在动量法中，尽管遇到函数较为平坦处，但仍有相对可观的动量可以更新参数

1. **优势**

2.1 稳定性高：其他批次的梯度信息或本次更新中的历史信息，使参数更新更接近真实的梯度，可以增加稳定性。同时如果梯度方向发生改变（和动量方向相反），参数更新幅度下降，可以有效避免参数更新时的震荡而不收敛的困境。

2.2 跳出局部最优解：相比于标准的梯度下降算法，如若遇到局部最小值以及函数波谷则很难跳出。但动量法中在梯度方向不正确（朝局部最优方向），依靠动量的正确方向有可能跳脱出局部最优解。

2.3 加速学习过程：由（1）中分析

3，有人认为“single hidden layer”的神经网络已经足够逼近任何连续的函数，那我们为什么很多时候却会使用“multiple layers”的神经网络？请分析两种设计的优缺点。

**答案：**

这是神经网络“万有逼近性”的一个错误理解。【看西瓜书】

4，请简要描述batch normalization，以及在standard normalization后为什么要做“scaling” 和“shifting” ？

**答案：**

5，假设你决定在你的神经网络中使用ReLU 作为hidden layer的activation函数，你在sigmoid activation前面也插入了一个ReLU函数。请分析这样做会有什么问题？

**答案：**

6，为什么sigmoid activation function可能会导致gradient vanishing的问题？请列举至少2个可以克服这个问题的activation function。

**答案：**

7，假设你为一位客户开发一个基于指纹的验证系统使客户能够用自己的指纹登录各种设备。你收集了一个开发用的数据集，训练出一个在测试集上能达到>99%准确率的模型。但是为客户部署模型后，你却遭到了客户的投诉说模型的准确率很糟。请列出一些可能导致以上情况发生的因素，并提出可能的解决办法。

**答案：**

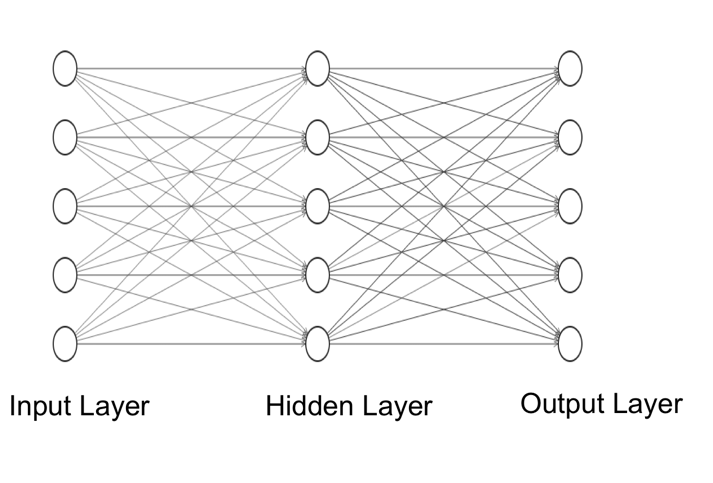
8，训练完一个神经网络模型后，你发现training accuracy (100%) 和 test accuracy (42%)之间有一个巨大的差异。请列举可能的原因和可能的解决办法？

**答案：**

9，什么时候可以使用multi-task learning技术？

**答案：**

10，假设你构建了如下结构的2层神经网络，请问它一共有多少参数（包括bias）？



**答案：**

11，请列举3个能够克服模型overfitting的方法？

答案：

12，请列举2个能够解决或减轻gradient exploding问题的方法？

答案：

13，尽管Pooling操作会使模型丢失掉一些feature map信息，那我们为什么常在CNN里面使用Pooling操作？

答案：

14，假设你在开发一个分类模型，你首先在一个有20个样本的数据上训练模型，模型训练收敛后，你发现training loss仍然挺高；所以你打算到一个有10000个样本的数据上重新训练模型。请问这样你能够获得满意的训练模型么？如果可以请解释原因和可能会得到的结果；如果不能，请解释原因并提供解决方法。

答案：

15，请介绍Self-Attention，并比较其和LSTM的相同点和不同点。

答案：

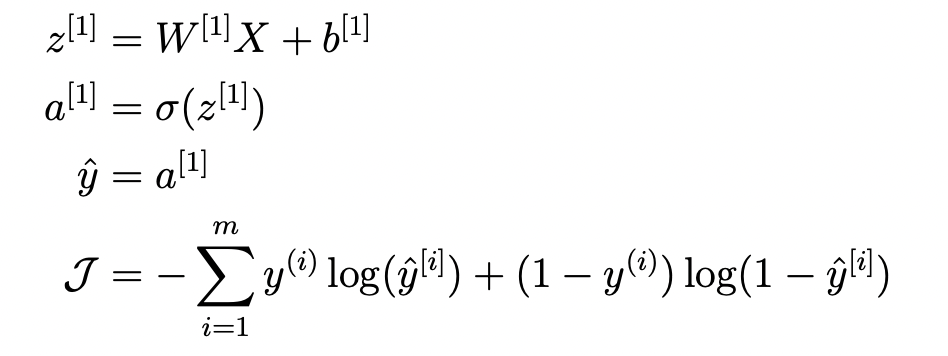
16，为什么deep network需要non-linear 的activation function?

答案：

17，请列举mini-batch的优点和缺点。

答案：

18，我们有一个做binary classification的单层神经网络。Input , Output  , true lable 以及具体的计算如下：



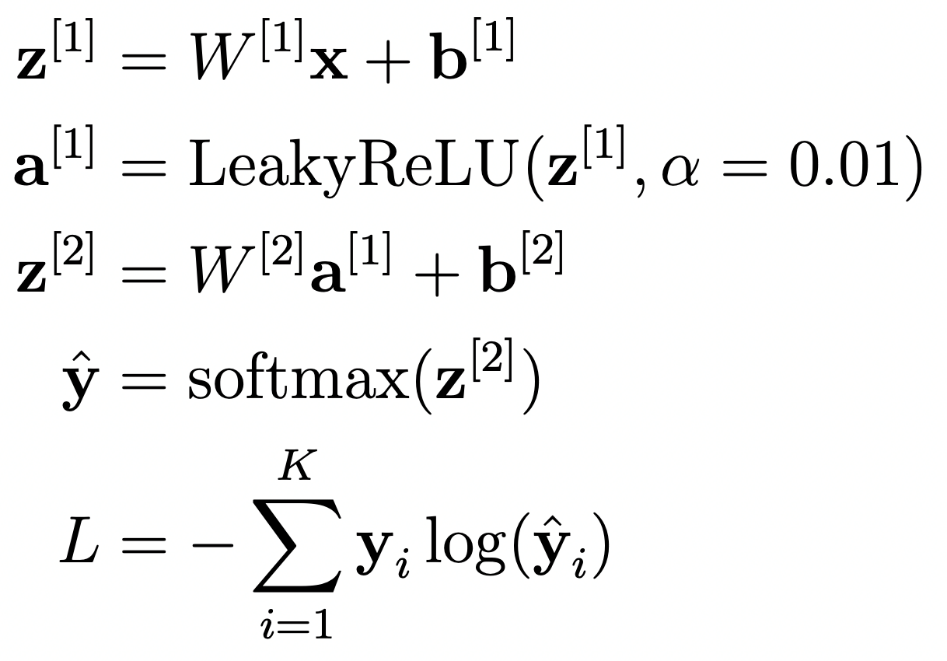
请以矩阵向量计算的形式写出的计算结果。

答案：

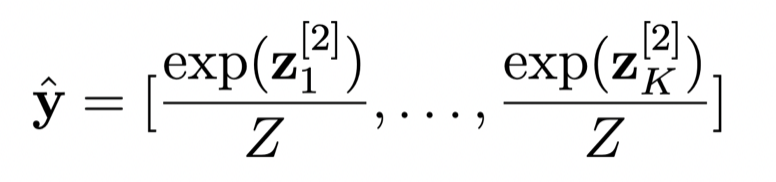
19，假设你训练一个GAN模型，在每一个epoch的结尾你都记录了Generator和Discriminator的loss值。 你发现第1个和第100个epoch结尾的losses非常接近，没啥变化。请问在这种情况下生成的图片的质量是否是一样的？为什么？

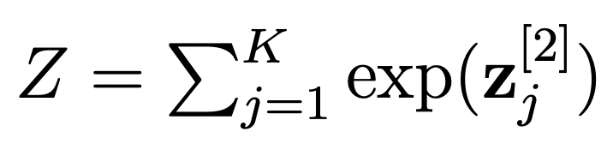
答案：

20，假设你用一个2层hidden layer的神经网络来解决K-class的分类问题。具体定义如下：



其中输入 , one-hot encoded标签，。softmax定义如下：





1. 请问和的维度是多少（dimensionality）？如果我们一次给一个batch的输入数据 （m个样本）,请问最后一层输出的结果的维度是多少？

答案：

1. 请计算？请尽量用里的元素表示最终结果。

答案：

1. 如果, 请给出的计算结果，请尽量用里的元素表示最终结果。

答案：

1. 假设y向量里的第k个元素值是1，其它都是0。请分情况计算? 请尽量用里的元素表示最终结果。

答案：