1. dropout作用是什么

dropout会随机封锁一些神经元权重的训练，让这些神经元“失效”，在每一个batch中，都会以参数p的比例使神经元失活，进而使得每一次训练的神经网络架构都不尽相同，在梯度下降的方向不一致基础上进一步加大每一次训练模型的差异性，虽然可能会使收敛变慢，但是会更有可能跳出局部最小值，防止模型过拟合。

1. 谈谈dropout为什么有效？（提示：bagging，贝叶斯概率）

**Bagging**：Dropout相当于通过随机失活神经元，每一次训练的模型都不一样，相当于bagging过程，模型集成可以有效降低过拟合。

**生物学解释**：dropout相当于有性繁殖。使用dropout可以减小和固有神经元之间的依赖，减少co-adaptation，使得它们可以随机地和其它神经元来共同学习，这可以使得神经元更加鲁棒，能够学到更多有用的特征。

**贝叶斯概率**：没找到（回想一下使用Bagging学习，我们定义 k 个不同的模型，从训练集有替换采样构造 k 个不同的数据集，然后在训练集上训练模型 i。Dropout的目标是在指数级数量的神经网络上近似这个过程。Dropout训练与Bagging训练不太一样。在Bagging的情况下，所有模型是独立的。在Dropout的情况下,模型是共享参数的，其中每个模型继承的父神经网络参数的不同子集。参数共享使得在有限可用的内存下代表指数数量的模型变得可能。在Bagging的情况下，每一个模型在其相应训练集上训练到收敛。在Dropout的情况下,通常大部分模型都没有显式地被训练,通常该模型很大，以致到宇宙毁灭都不能采样所有可能的子网络。取而代之的是，可能的子网络的一小部分训练单个步骤,参数共享导致剩余的子网络能有好的参数设定。）网上都是这么写的，我感觉没什么关系。

1. 谈一下LSTM与GRU的异同

LSTM与GRU都是在RNN基础之上的改进，可以避免rnn在对更远神经元状态的保留时产生的梯度消失和梯度爆炸的问题。LSTM通过引入一个单元状态C保留更远的神经元状态，通过遗忘门控制上一个Ct-1对Ct的影响，通过输入门控制当前xt对Ct的输入，通过输出门控制在遗忘门和输入门调控后的Ct有多少可以输出到ht中；GRU则简化成两个门，分别是重置上一个单元记忆状态ht-1的重置门（我认为类似于LSTM的遗忘门，决定了上一个时刻多少状态可以保留），和决定当前记忆状态xt有多少可以保留并送入下一个ht的更新门（类似于LSTM的输入门与输出门）。也因此GRU的参数比LSTM更少，更容易训练。但二者达到的训练水平基本一致

1. CNN具备怎样的特性，为什么对于视觉信号处理这么有效？

CNN具有稀疏交互和参数共享的特性，稀疏交互要求下一层卷积网络仅包含上一层网络局部的信息，这种方式可以提取图像中的局部特征，而非整体图像按照固定顺序排列的特征，可以有助于提取更深层次的特征（例如更远距离特征之间的组合），从而对视觉信号处理十分有利；参数共享的特性可以有效降低网络中神经元数量，对于含有大量像素点多通道的图片，可以减少网络中的参数，从而提升网络的学习能力，并且降低运算难度，提升运算时间。

1. dropout层和激活层的顺序应该怎么设计，为什么？

Dropout放在激活层之后，激活层如果是relu本身就会沉默一些神经元，需要在这基础之上再次沉默，另外，对于已经沉默的神经元再次进行激活函数的处理没有意义。

1. 阅读代码401 CNN MNIST，最后一层输出可视化给你什么启示？

最后一层使用TSNE作为可视化工具，提取最后一层卷积层输出数据，（已拉平至（2000，,32\*7\*7）（2000张图片,通道32，特征图大小7\*7）），取前500张，降维图片数据（32\*7\*7维）至二维，根据降维的意义，当相同的数字更多的聚集在一起即为误差更小。

在数据降维之前一般需要进行标准化，可以加速收敛速度，降低偏差过大的数据点对降维过程的影响，我将图片数据按行标准化后发现没有什么变化，然后才发现原因是minst手写数字数据集（训练集）已经标准化了。。。。