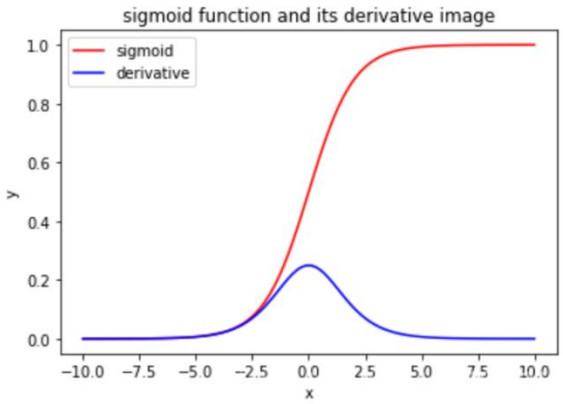
神经网络与深度学习习题解答

1，计算 Sigmoid 函数、双曲正切函数和 ReLU 函数的导数函数，分析这三个函数作为激活函数时的优缺点。

解：（1）对于 Sigmoid 函数，$\theta(x)=\frac{1}{1+e^{-x}}$

$$\frac{\partial\theta(x)}{\partial x}=\frac{e^{-x}}{(1+e^{-x})^2}=\frac{1+e^{-x}-1}{1+e^{-x}}\frac{1}{1+e^{-x}}=(1-\theta(x))\theta(x)$$

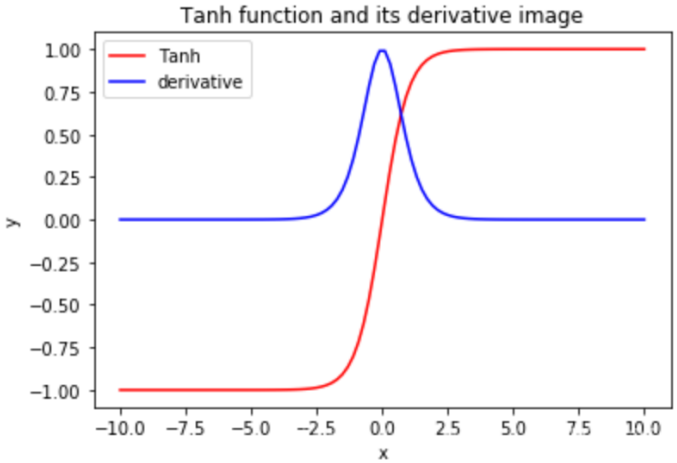
Sigmoid 函数作为激活函数的优点是连续、单调、可导且具有非线性特点。但其缺点是非中心对称，输出不是 0 均值的，同时它的导数函数曲线见右图，x 变化很小的范围内，导数才有值且不大，在神经网络应用中，训练过程使用的是反向传播算法，通过链式法则回传的梯度不断相乘，因此，Sigmoid 函数作为激活函数时会导致梯度消失问题，尤其在网络比较深时会达不到训练效果。



（2）对于双曲正切函数，$tanh(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$

$$\frac{\partial\tanh(x)}{\partial x}=\frac{(e^{x}+e^{-x})(e^{x}+e^{-x})-(e^{x}-e^{-x})(e^{x}-e^{-x})}{(e^{x}+e^{-x})^2}=1-tanh^2(x)$$

双曲正切函数作为激活函数的优点是中心对称、单调、连续、可导且具有非线性特点。但其缺点是它的导数函

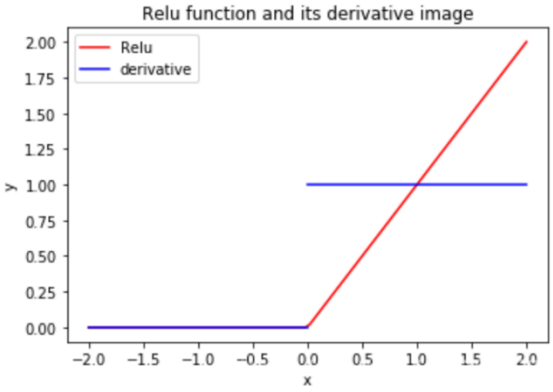


数曲线见右图，x 变化更小的范围内，导数才有值且不大于 1，在神经网络应用中，训练过程使用的是反向传播算法，通过链式法则回传的梯度不断相乘，因此，双曲正切函数作为激活函数时同样会导致梯度消失问题，尤其在网络比较深时会达不到训练效果。

（3）对于 ReLU 函数，$ReLU(x) = \max(0, x)$

$$\frac{\partial \mathrm{ReLU}(x)}{\partial x} = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$$

ReLU 函数作为激活函数的优点是梯度不会饱和，解决了梯度消失问题，没有指数计算，计算复杂度低。它的缺点是非中心对称，负数部分的梯度为 0，导致相应参数不会更新。



2，假设用训练样本集为 $\mathrm{D} = \{(\mathbf{x}_1, y_1) = ((0,1)^T, 1), (\mathbf{x}_2, y_2) = ((1,0)^T, 1), (\mathbf{x}_3, y_3) = ((-1,0)^T, -1)\}$设计单层感知机（只有一层、一个神经元），当无激活函数时，采用感知器算法、对权重的初始值进行随机初始化，得到的分类面是：$\mathrm{g}_{PLA}(\mathbf{x}) = 1 + 10x_1 + 10x_2$，若采用 Sigmoid 激活函数，用 BP 算法（反向传播算法），经过多次迭代后得到的分类面是：$\mathrm{g}_{BP}(\mathbf{x}) = x_1 + x_2$。现增加一个训练样本$(\mathbf{x}_4, y_4) = ((0, -t)^T, 1)$，且$t \to \infty$，再进行训练得到新的分类面，请问（1）当无激活函数时，采用感知器算法能否得到能使训练样本都能正确分类的分类面$\mathrm{g}_{PLA}(\mathbf{x})$？（2）若采用 Sigmoid 激活函数，用 BP 算法得到的

分类面会和$\mathrm{g}_{BP}(\mathbf{x})=x_1+x_2$基本一致吗？

解：

（1）增加了样本$\mathbf{x}_4$后，当无激活函数时，这四个训练样本仍然是线性可分的，因此，用感知器算法仍然能够得到最优分类面，使得四个样本都能正确分类，但此时的分类面与$\mathrm{g}_{PLA}(\mathbf{x})=1+10x_1+10x_2$不同，经过训练后，它可能得到的最优解为：$\mathrm{g}_{PLA}(\mathbf{x})=1.6+2x_1-x_2$

（2）增加了样本$\mathbf{x}_4$后，当采用 Sigmoid 激活函数时，用 BP 算法得到的分类面会和$\mathrm{g}_{BP}(\mathbf{x})=x_1+x_2$基本一致。原因是：

对于 Sigmoid 激活函数，其损失函数为$\ell_{in}(\boldsymbol{w})=(\theta(y_n\boldsymbol{w}^T\mathbf{x}_n)-1)^2$，其中，$\theta(y_n\boldsymbol{w}^T\mathbf{x}_n)=\frac{1}{1+exp(-y_n\boldsymbol{w}^T\mathbf{x}_n)}$，计算其梯度为：

$$\frac{\partial\ell_{in}(\boldsymbol{w})}{\partial\boldsymbol{w}}=2(\theta(y_n\boldsymbol{w}^T\mathbf{x}_n)-1)\theta(y_n\boldsymbol{w}^T\mathbf{x}_n)(1\text{-}\theta(y_n\boldsymbol{w}^T\mathbf{x}_n))y_n\mathbf{x}_n^T$$

由于$(\mathbf{x}_4,y_4)=((0,-t)^T,1)$，且$t\to\infty$，即$y_4\boldsymbol{w}^T\mathbf{x}_4$在迭代过程中会很大或者很小，此时$\frac{\partial\ell_{in}(\boldsymbol{w})}{\partial\boldsymbol{w}}$都近似为 0，也就是说样本$\mathbf{x}_4$对于权重更新几乎没有贡献，因此，在原有三个样本点获得的最有边界附近梯度下降法无法对参数进行大幅修正，因此，得到的最优解和$\mathrm{g}_{BP}(\mathbf{x})=x_1+x_2$基本一致。

3，给定输入$\mathbf{x}$，假设一个包含两个隐含层（参数分别为$w_{ij}^{(1)}$和$w_{ij}^{(2)}$）和一个输出层$\hat{y}$（参数为$w_{ij}^{(3)}$）的神经网络，所有神经元的激活函数均为 Sigmoid 函数，给定可导的损失函数ℓ_n，请推导损失函数对各参数的梯度，并分析梯度消失现象。

解：

根据定义，网络结构为：

网络第1层第j个神经元的输入为：

$$S_j^{(1)} = \sum_{i=0}^{d^{(0)}} w_{ij}^{(1)} x_i^{(0)}$$

网络第1层第j个神经元的输出为：

$$x_j^{(1)} = \theta(S_j^{(1)})$$

网络第2层第j个神经元的输入为：

$$S_j^{(2)} = \sum_{i=0}^{d^{(1)}} w_{ij}^{(2)} x_i^{(1)}$$

网络第2层第j个神经元的输出为：

$$x_j^{(2)} = \theta(S_j^{(2)})$$

网络输出层第j个神经元的输入为：

$$S_j^{(3)} = \sum_{i=0}^{d^{(2)}} w_{ij}^{(3)} x_i^{(2)}$$

网络输出层第j个神经元的输出为：

$$\hat{y} = \theta(S_j^{(3)})$$

其中：$\theta(u) = \frac{1}{1+e^{-u}}$为 Sigmoid 函数

根据链式法则：

$$\frac{\partial \ell_n}{\partial w_{ij}^{(3)}} = \frac{\partial \ell_n}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial S_j^{(3)}} \cdot \frac{\partial S_j^{(3)}}{\partial w_{ij}^{(3)}} = \frac{\partial \ell_n}{\partial \hat{y}} \cdot \theta'(S_j^{(3)}) \cdot \frac{\partial S_j^{(3)}}{\partial w_{ij}^{(3)}}$$

$$\frac{\partial \ell_n}{\partial w_{ij}^{(2)}} = \frac{\partial \ell_n}{\partial S_j^{(2)}} \cdot \frac{\partial S_j^{(2)}}{\partial w_{ij}^{(2)}} = \sum_{k=1}^{d^{(3)}} \frac{\partial \ell_n}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial S_k^{(3)}} \cdot \frac{\partial S_k^{(3)}}{\partial x_j^{(2)}} \cdot \frac{\partial x_j^{(2)}}{\partial S_j^{(2)}} \cdot \frac{\partial S_j^{(2)}}{\partial w_{ij}^{(2)}}$$

$$= \sum_{k=1}^{d^{(3)}} \frac{\partial \ell_n}{\partial \hat{y}} \cdot \theta'(S_k^{(3)}) \cdot \frac{\partial S_k^{(3)}}{\partial x_j^{(2)}} \cdot \theta'(S_j^{(2)}) \cdot \frac{\partial S_j^{(2)}}{\partial w_{ij}^{(2)}}$$

$$\frac{\partial \ell_n}{\partial w_{ij}^{(1)}} = \frac{\partial \ell_n}{\partial S_j^{(1)}} \cdot \frac{\partial S_j^{(1)}}{\partial w_{ij}^{(1)}} = \sum_{m=1}^{d^{(2)}} \frac{\partial \ell_n}{\partial S_m^{(2)}} \cdot \frac{\partial S_m^{(2)}}{\partial x_j^{(1)}} \cdot \frac{\partial x_j^{(1)}}{\partial S_j^{(1)}} \cdot \frac{\partial S_j^{(1)}}{\partial w_{ij}^{(1)}}$$

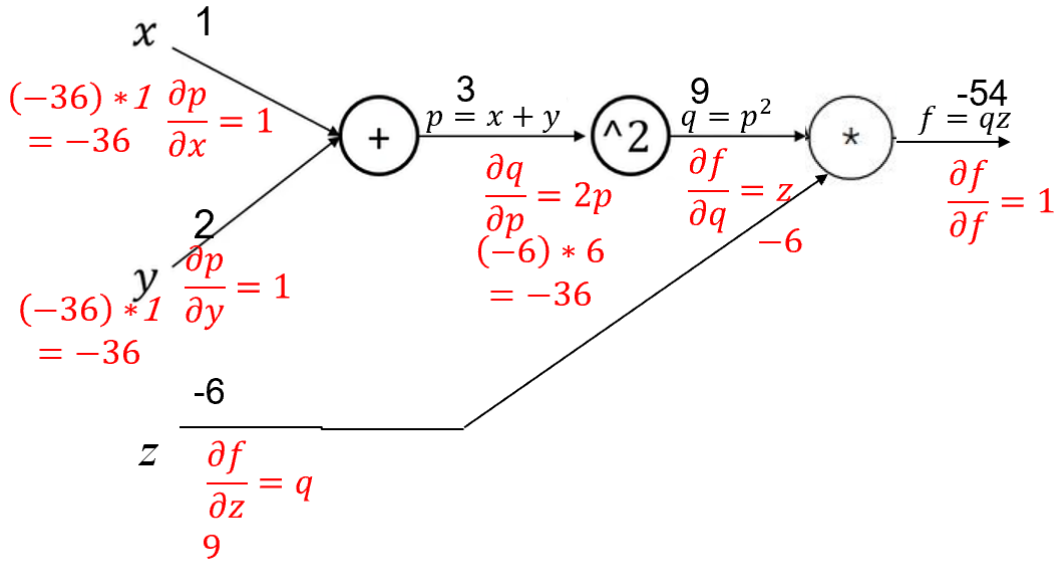
$$= \sum_{m=1}^{d^{(2)}} \sum_{k=1}^{d^{(3)}} \frac{\partial \ell_n}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial S_k^{(3)}} \cdot \frac{\partial S_k^{(3)}}{\partial x_m^{(2)}} \cdot \frac{\partial x_m^{(2)}}{\partial S_m^{(2)}} \cdot \frac{\partial S_m^{(2)}}{\partial x_j^{(1)}} \cdot \frac{\partial x_j^{(1)}}{\partial S_j^{(1)}} \cdot \frac{\partial S_j^{(1)}}{\partial w_{ij}^{(1)}}$$

$$= \sum_{m=1}^{d^{(2)}} \sum_{k=1}^{d^{(3)}} \frac{\partial \ell_n}{\partial \hat{y}} \cdot \theta'(S_k^{(3)}) \cdot \frac{\partial S_k^{(3)}}{\partial x_m^{(2)}} \cdot \theta'(S_m^{(2)}) \cdot \frac{\partial S_m^{(2)}}{\partial x_j^{(1)}} \cdot \theta'(S_j^{(1)}) \cdot \frac{\partial S_j^{(1)}}{\partial w_{ij}^{(1)}}$$

可以观察到损失函数ℓ_n 关于各层参数的梯度包含该层之后每一个Sigmoid层的梯度，如果某一层Sigmoid函数的梯度θ'接近0，那么其之前各层都会因为这一乘数因子的存在导致梯度接近于0，出现梯度消失。同时可以看到，随着网络层次的增加，梯度在反传过程中会逐渐减小，也会导致梯度消失。

4，画出（$x+y$）$^2 z$的计算图，当 x=1,y=2,z=-6 时，写出前向传播的数值和反向传播的梯度值。

解：



5，对于一幅 300*300 大小的彩色（RGB）图像，（1）如果输入端与有 100 个神经元的第一层隐含层用全链接方式（Fully Connected neural Network）连接时，请问这一层会包含多少参数？（2）如果用 100 个 5*5*3 大小的滤波器作卷积操作，那么这一层的参数为多少？如果滤波器移动步长（stride=1）为 1，经过卷积计算后的输出端神经元个数有多少？

解：（1）因为是 RGB 图像，所以共有 3 个通道，全连接情况下包含的参数是：300*300（图像大小）*3（颜色通道数）*100（第一层神经元个数）+1*100（每个神经元都要与输入层的常数项连接）=27000100；

（2）卷积操作时，第一层的参数是由卷积滤波器大小和常数项共同确定的，因此其包含的参数为：（5*5（滤波器大小）*3（颜色通道数）+1（常数项））*100（滤波器个数）=7600；（即：F=5，K=100，

D1=3，参数量=（F*F*D1）*K+K=5*5*3*100+100=7600）

因为卷积核为 5*5*3，且填充值为 0，卷积后第一层神经元的个数为（（300（图像长或宽）-5（滤波器大小）/1（移动步长））+1）^2*100（滤波器个数）=8761600（相当于 W1=300，H1=300，D1=3，F=5，K=100，S=1，P=0，W2=（W1-F+2P）/S+1=（300-5+0）/1+1=296；H2=（H1-F+2P）/S+1=（300-5+0）/1+1=296；D2=K=100，神经元个数为：W2*H2*D2=296*296*100=8761600）

6，某一个卷积神经网络结构如下：

（i）输入层 Input 的 RGB 图像大小是 227*227*3。

（ii）第 1 层卷积层 Conv-1 是通过对输入图像用 96 个 11*11*3 大小的滤波器通过步长(stride)为 4，不做边缘填充（padding）得到的。

（iii）接下来是池化层 MaxPool-1，它用 3*3 尺寸、步长为 2 对 Conv-1 做 Max Pooling 操作。

（iv）然后我们对图像进行边缘填充，填充值为 2（如原来图像大小为 7*7 时，做填充值为 2 的填充后，图像大小变为 11*11），用 256 个 5*5 大小的滤波器按步长为 1，做第二次卷积操作，得到 Conv-2 层。

（v）再接一个池化层 MaxPool-2，它用 3*3 尺寸、步长为 2 做一次 Max Pooling 操作。

（vi）MaxPool-2 层输出去接一个有 4096 个神经元的全连接层 FC-1。

（vii）再接一个全连接层 FC-2 实现对 1000 个类别的分类。

请计算：（1）输入层到 Conv-1 层的参数量有多少？（2）经过池化层 MaxPool-1 后的神经元是多少？（3）经过第二次卷积操作后的图像大小为多少？（4）MaxPool-2 层到 FC-1 层的参数量是多少？（5）FC-1 层到 FC-2 层的参数量是多少？

解：输入图像大小为 227*227*3（即：W1=227，H1=227，D1=3），第一层卷积核为 11*11（即：F=11），共 96 个滤波器（即：K=96），步长为 4（即：S=4），边缘填充为 0（即：P=0），则卷积以后的图像边长为：（（227-11+2*0）/4）+1=55，大小为 55*55（即 W2=H2=55），与 96 个滤波器构成特征图，所以卷积层 Conv-1 的神经元个数为 55*55*96=290400，（1）输入层到 Conv-1 层的参数量为 F*F*D1*K+1*K，即：11*11*3*96+1*96=34944；对 55*55 大小的图像做第一次 Maxpooling，这时候通道数 96 保持不变，因为它用 3*3 大小的尺寸以步长为 2 做 Maxpooling，则得到的图像边长为（（55-3）/2）+1=27，图像大小为 27*27，（2）经过池化层 MaxPool-1 后的神经元为 27*27*96；再做第二次卷积，此时是对 27*27 大小的图像，用 5*5 大小的滤波器按步长为 1，边缘填充值为 2 做卷积，滤波器个数为 256 个，所以，卷积后图像的边长为（（27-5+2*2）/1）+1=27，大小为 27*27，与 256 个滤波器构成特征图，所以卷积层 Conv-2 的神经元个数为 27*27*256=186624，（3）经过第二次卷积操作后的图像大小为 27*27，MaxPool-1 层到 Conv-2 层的参数量为：5*5*96（池化层的通道数）*256（滤波器个数）+256（常数项）=614656；再经过池化层 MaxPool-2，3*3 尺寸、步长为 2，则得到的图像边长为（（27-3）/2）+1=13，图

像大小为 13*13，上一层的通道数是 256，所以，MaxPool-2 层的神经元个数为 13*13*256=43264；MaxPool-2 层输出去接一个有 4096 个神经元的全连接层 FC-1，所以，（4）MaxPool-2 层到 FC-1 层的参数量是 13*13*256*4096+4096（常数项）=177213440；最后的输出层要对 1000 个类别进行分类，即 FC-2 层的神经元个数是 1000 个，而输入是 4096 个神经元，所以，（5）FC-1 层到 FC-2 层的参数量是 4096*1000+1000=4097000