问题

在将图像输入到深度学习网络之前，一般先对图像进行预处理，即图像归一化，为什么需要这么做呢？

问题背景

在面试的时候，面试官先问的问题是“机器学习中为什么要做特征归一化”，我的回答是“特征归一化可以消除特征之间量纲不同的影响，不然分析出来的结果显然会倾向于数值差别比较大的特征，另外从梯度下降的角度理解，数据归一化后，最优解的寻优过程明显会变得平缓，更容易正确的收敛到最优解”。接着面试官又问“图像的像素值都是在0到255之间，并不存在量纲的差别，那为什么还需要做归一化呢？”是啊，为什么还要呢，被问住了……

拓展

既然是从机器学习特征归一化引出的图像的归一化问题，那么我们先讨论下“为什么要对数值型特征做归一化？”吧。

很多资料例如《百面机器学习》都是从梯度下降的角度来分析这个问题的，讨论地还不错的一篇是这个知乎的回答，已经写得比较清晰了，所以这里就不再整理了，直接点开链接看。

不过这个回答里面未归一化时的损失函数等高线图中椭圆的方向应该是横向的而不是纵向的，因为 θ_2 前的系数比 θ_1 的大，所以在损失函数等高图上 θ_2 的变化范围比 θ_1 小才对，另外对于为什么圆形的等高线相对于椭圆形的等高线，更新方向更加平滑，所以更快也更容易收敛到最优解呢？一句话简单解释就是因为归一化后，等高图大致为圆形，更新方向与等高线垂直，所以理想的更新方向是直指圆心的一条直线。

所以对于“为什么机器学习中要进行特征归一化”这个问题，总结起来可以从三个点去回答：

1. **消除特征之间量纲的影响**，使得不同特征之间具有可比性
2. 在使用随机梯度下降求解的模型中，**能加快模型收敛速度**
3. **归一化还有可能提高精度**：一些分类器需要计算样本之间的距离（如欧氏距离），例如KNN。如果一个特征值域范围非常大，那么距离计算就主要取决于这个特征，从而与实际情况相悖（比如这时实际情况是值域范围小的特征更重要）。

问题解答

对于这个问题，网上看了很多博客，没有找到一个很全面权威的解释，所以这里把几个自认为讲得比较合理的解释列出来：

1. 灰度数据表示有两种方法：一种是uint8类型、另一种是double类型。其中uint8类型数据的取值范围为 [0,255]，而double类型数据的取值范围为[0,1]，两者正好相差255倍。对于double类型数据，其取值大于1时，就会表示为白色，不能显示图像的信息，故当运算数据类型为double时，为了显示图像要除255。
2. 图像深度学习网络也是使用gradient descent来训练模型的，使用gradient descent的都要在数据预处理步骤进行数据归一化，主要原因是，根据反向传播公式：

$$\frac{\partial J}{\omega_{11}} = x_1 * \text{后面层梯度的乘积} \tag{1}$$

   如果输入层 x 很大，在反向传播时候传递到输入层的梯度就会变得很大。梯度大，学习率就得非常小，否则会越过最优。在这种情况下，学习率的选择需要参考输入层数值大小，而直接将数据归一化操作，能很方便的选择学习率。在未归一化时，输入的分布差异大，所以各个参数的梯度数量级不相同，因此，它们需要的学习率数量级也就不相同。对 w1 适合的学习率，可能相对于 w2 来

说会太小，如果仍使用适合 w1 的学习率，会导致在 w2 方向上走的非常慢，会消耗非常多的时间，而使用适合 w2 的学习率，对 w1 来说又太大，搜索不到适合 w1 的解。

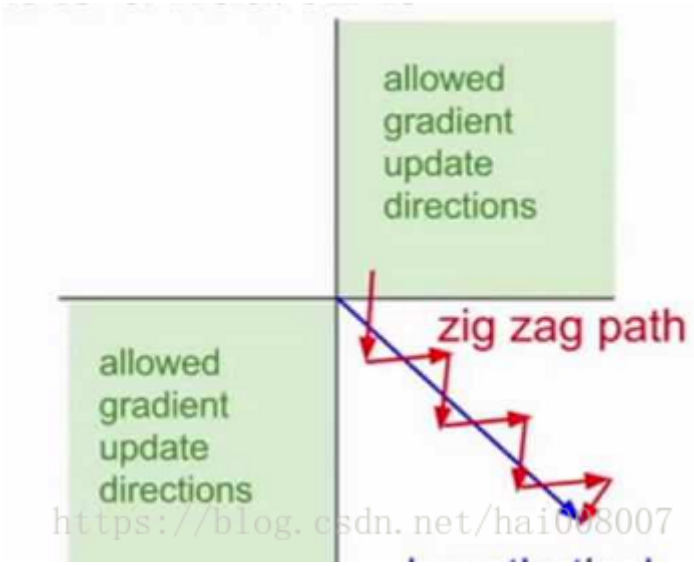
3. 通过标准化后，实现了数据中心化，数据中心化符合数据分布规律，能增加模型的泛化能力

问题深入

那么深度学习中在训练网络之前应该怎么做图像归一化呢？有两种方法：

1. **归一化到 0 - 1**：因为图片像素值的范围都在0~255，图片数据的归一化可以简单地除以255. 。 (注意255要加 . ，因为是要归一化到double型的 0-1)
2. **归一化到 [-1, 1]**：在深度学习网络的代码中，将图像喂给网络前，会先统计训练集中图像RGB这3个通道的均值和方差，如：`mean=[123.675, 116.28, 103.53], std=[58.395, 57.12, 57.375]`，接着对各通道的像素做减去均值并除以标准差的操作。不仅在训练的时候要做这个预处理，在测试的时候，同样是使用在训练集中算出来的均值与标准差进行的归一化。

注意两者的区别：归一化到 [-1, 1] 就不会出现输入都为正数的情况，如果输入都为正数，会出现什么情况呢？：根据求导的链式法则，w的局部梯度是X，当X全为正时，由反向传播传下来的梯度乘以X后不会改变方向，要么为正数要么为负数，也就是说w权重的更新在一次更新迭代计算中要么同时减小，要么同时增大。



其中，w的更新方向向量只能在第一和第三象限。假设最佳的w向量如蓝色线所示，由于输入全为正，现在迭代更新只能沿着红色路径做zig-zag运动，更新的效率很慢。

基于此，当输入数据减去均值后，就会有负有正，会消除这种影响。

参考资料：

1.为什么要对数据进行归一化处理？
2. 深度学习中图像为什么要归一化？
3. 深度学习中，为什么需要对数据进行归一化
4. 深度学习的输入数据集为什么要做均值化处理

By Yee

2020.05.10