Deep Learning Technology and Application

Ge Li

Peking University

权重初始化方法

权重初始化方法

初始化参数的目的：

- 初始化是消除神经网络隐藏层同层神经元之间存在的对称性;
- 不合理的初始化将导致怪异的输出结果：
  - 若采用统一的初始化值，梯度下降又促使所有参数按照相同的方式进行调整，将导致训练实效；
  - 不合理的初始化，可能导致训练结果持续变大，或持续变小，从而出现怪异的输出结果；
  - 不合理的分布，可能导致权重参数出现方差越来越大的现象；
- 应该尽量控制参数的变化范围：使其方差变化不至于太大。

权重初始化方法

常见的初始化方法有如下几种：

1. **论文 [1] 按照标准差为** 0.01，**均值为零的高斯分布随机生成数值，初始化权重** W，**并将第** 2，**第** 4 **和第** 5 **个卷积层，及所有全连接层的偏置项设置为常数**；
2. **论文 [2] 提出一种** Xavier **方法**；**论文 [3] 又给出了一种考虑** ReLU **函数的改进方法**；

[1] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: NIPS, 2012.

[2] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: AISTATS, 2010.

[3] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: ICCV, 2015.

权重初始化方法

根据 Xavier 方法：当激活函数在 0 值附近，导数接近 1 的条件下，其初始化参数可以按照如下范围内的均匀分布提取：

$$\left[-\sqrt{\frac{6}{n^{k+1}+n^k}}, \sqrt{\frac{6}{n^{k+1}+n^k}}\right]$$

接下来推证其合理性。先来回顾预备知识：
方差：如果随机变量 X 的数学期望 $\mu = EX$ 有限，则称：

$$E(X-\mu)^2$$

为 X 的方差，记作 $var(X)$.
若随机变量 X 和 Y 均服从均值为 0，方差为 σ 的分布，则：

- $X * Y$ 服从均值为 0，方差为 σ^2 的分布；
- $X * Y + X * Y$ 服从均值为 0，方差为 $2\sigma^2$ 的分布；

权重初始化方法

设输入数据 $x \in R^n$ 服从均值为 0，方差为 σ_x 的分布；
设 $w \in R^n$ 为输入层 n 个神经元与输出层第 j 个神经元之间的连接权重；
设 w 服从均值为 0，方差为 σ_w 的分布，则：

$$z_j = \sum_{i}^{n} w_i * x_i$$

根据方差的性质，z_j 满足均值为 0，方差为 $n\sigma_x\sigma_w$ 的分布；

注意到，在 0 值附近，神经网络激活函数近似服从 $a = x$，因此，在不考虑 w 的情况下有：$a_j = x$. 那么，若要使神经网络输入层与其下一层的输出值保持方差不变，则应令：

$$\sigma_w = \frac{1}{n}$$

权重初始化方法

若神经网络有 k 层，则根据前向传播公式，第 k 层神经元的 z 值方差为：

$$\sigma_x^k = \sigma_x^1 * \prod_{i=2}^{k} n^i * \sigma_w^i$$

可见，若要使神经网络输入层与其下一层的输出值保持方差不变，需令：

$$\sigma_w^k = \frac{1}{n^{k-1}}$$

即：第 k 层的权值的方差 σ_w^k 应为第 $k-1$ 层神经数目的倒数；
接下来，看反向传播过程中的约束关系。

权重初始化方法

因为：

$$x_i^k = f(\sum_{j=1}^{n^{k-1}} w_j^k * x_j^{k-1} + b)$$

则：

$$\frac{\partial J}{\partial x_j^{k-1}} = \sum_{i=1}^{n^k} \frac{\partial J}{\partial x_i^k} * w_j^k$$

则，根据方法性质公式，得：

$$\text{var}\left(\frac{\partial J}{\partial x_j^{k-1}}\right) = n^k * \text{var}\left(\frac{\partial J}{\partial x_i^k}\right) * \sigma_w^k$$

权重初始化方法

若神经网络有 k 层，则在反向传播中，有：

$$\text{var}\left(\frac{\partial J}{\partial x_j^1}\right) = \text{var}\left(\frac{\partial J}{\partial x_i^k}\right) * \prod_{i=1}^{k-1} n^i * \sigma_w^i$$

可见，若要使神经网络输入层与其下一层的输出值保持方差不变，需令：

$$\sigma_w^k = \frac{1}{n^k}$$

即：第 k 层的权值的方差 σ_w^k 应为第 k 层神经数目的倒数；

权重初始化方法

综上所述，我们得到如下两个约束关系：

$$\sigma_w^k = \frac{1}{n^k} \qquad \sigma_w^k = \frac{1}{n^{k-1}}$$

对上述两个约束条件进行融合，得第 k 层 w 的约束条件为：

$$\sigma_w^k = \frac{2}{n^{k-1} + n^k}$$

这是初始化应满足的第一个条件。

设，我们对权重进行初始化的取值条件是：$[-u, u]$ 之间的均匀分布，则依据均匀分布的方差公式，又有初始化满足的第二个条件：

$$\text{var(uniform)} = \frac{(u-(-u))^2}{12} = \frac{u^2}{3}$$

权重初始化方法

联合上述两个条件，于是有：

$$\sigma_w^k = \frac{2}{n^{k-1} + n^k} = \frac{u^2}{3}$$

得：

$$u = \sqrt{\frac{6}{n^{k-1} + n^k}}$$

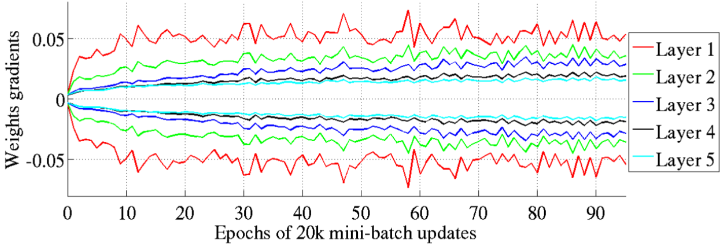
即得到 Xavier 方法：

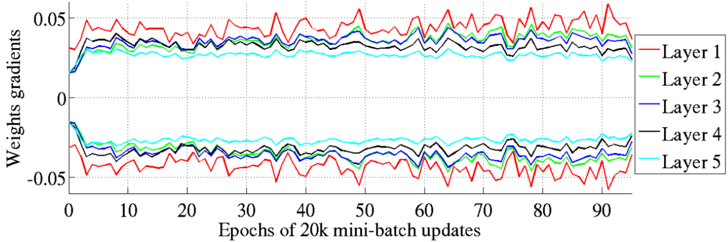
当激活函数在 0 值附近，导数接近 1 的条件下，其初始化参数可以按照如下范围内的均匀分布提取：

$$\left[-\sqrt{\frac{6}{n^{k+1} + n^k}}, \sqrt{\frac{6}{n^{k+1} + n^k}}\right]$$

权重初始化方法

Xavier 方法的效果：





Thanks.