Batch Normalization的作用和原理

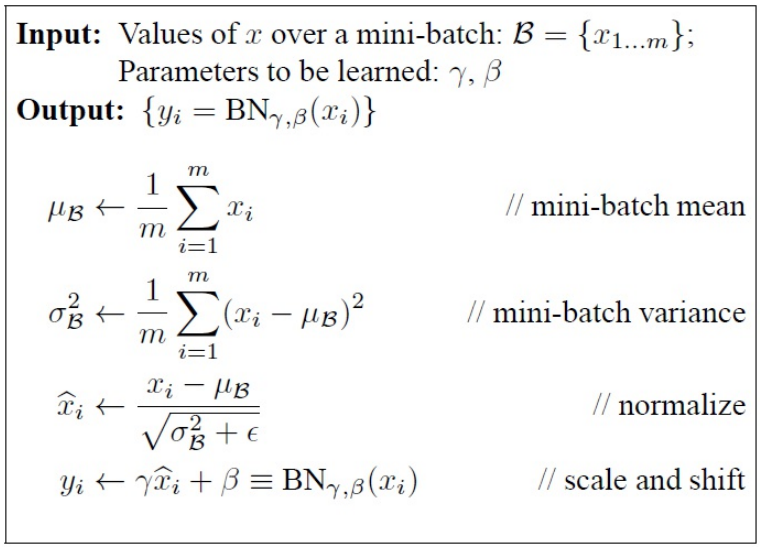
从实用性出发

使用batch normalization的目的是加速训练。keras中对Batchnormalization的解释为：

> Normalize the activations of the previous layer at each batch, i.e. applies a transformation that maintains the mean activation close to 0 and the activation standard deviation close to 1.

标准化前一层的激活输出，将其变换到均值接近0，标准差接近1的分布范围上。

Batch normalization的论文对该算法的描述如下：Batch Normalization



从算法中我们可以看出，上一层的激活值首先会被统计求出均值 μ 和方差 $\sigma^{2}}$, 然后进行标准化，注意这里标准化时方差需要加上极小值 ϵ，目的是防止除数为0.

注意最后输出的:
$$ y = \gamma \cdot x + \beta $$
这里 γ和 β 是两个可以更新的参数，这样做的原因原论文的结束如下：

> Note that simply normalizing each input of a layer may
> change w hat the layer can represent. For instance, normalizing
> the inputs of a sigmoid w ould constrain them to
> the linear regime of the nonlinearity. To address this, w e
> make sure that the transformation inserted in the netw ork
> can represent the identity transform.

也就是说，如果只是简单的将输入标准话，有可能会改变它原本想要表达的内容，例如标准话sigmoid函数的输入会使他们强行落在线性范围内，为了解决这一问题而引入的这两个可以被训练的参数。

在训练过程中Batch normalization层发生了什么？

通过一个实验对比