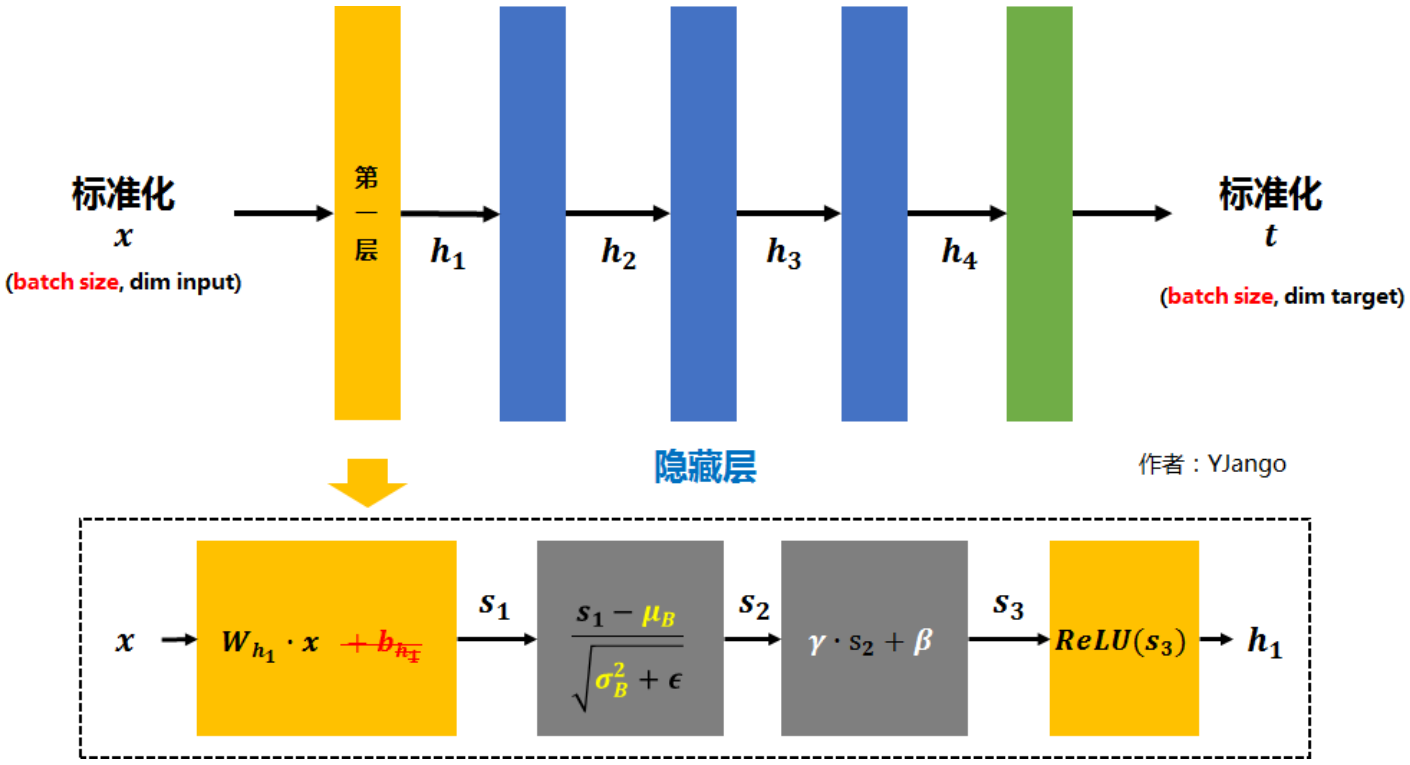


YJango的Batch Normalization--介绍

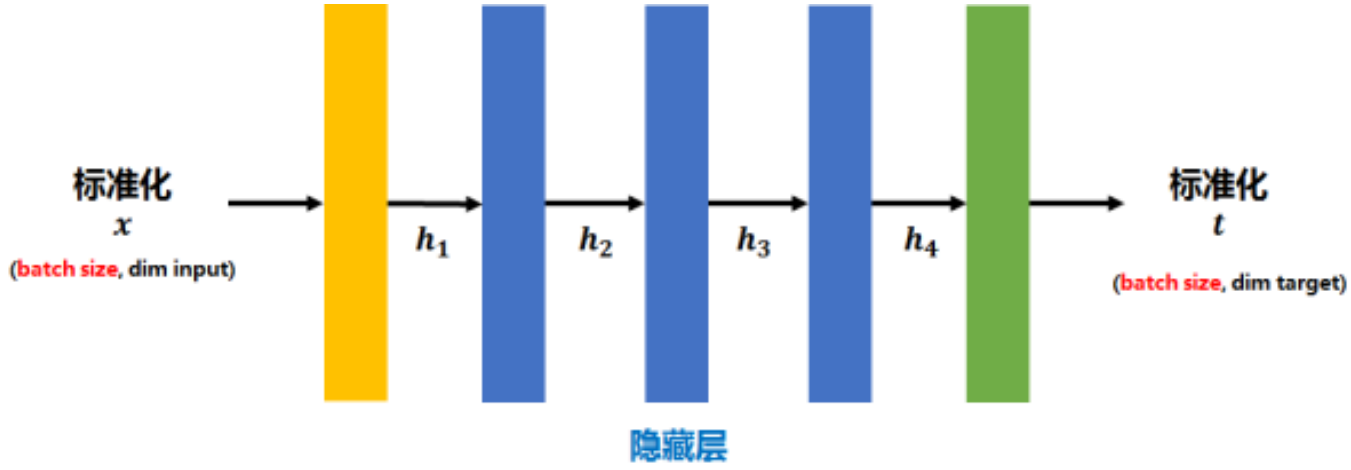
YJango · 2 个月前

思考

YJango的前馈神经网络--代码LV3的数据预处理中提到过：在数据预处理阶段，数据会被标准化（减掉平均值、除以标准差），以降低不同样本间的差异性，使建模变得相对简单。

我们又知道神经网络中的每一层都是一次变换，而上一层的输出又会作为下一层的输入继续变换。如下图中，x 经过第一层 $\phi(W_{h_1} \cdot x + b_{h_1})$ 的变换后，所得到的 h_1 ；而 h_1 经过第二层 $\phi(W_{h_1} \cdot h_1 + b_{h_1})$ 的变换后，得到 h_2 。

h_1 在第二层所扮演的角色就是 x 在第一层所扮演的角色。 我们将 x 进行了标准化，**那么，为什么不对 h_1 也进行标准化呢？**



Batch Normalization论文便首次提出了这样的做法。

Batch Normalization（BN）就是将每个**隐藏层的输出结果**（如 h_1, h_2, h_3 ）在batch上也进行**标准化后再送入下一层**（就像我们在数据预处理中将 x 进行标准化后送入神经网络的第一层一样）。

优点

那么Batch Normalization（BN）有什么优点？BN的优点是多个并存，但这里只提一个最容易理解的优点。

训练时的问题

尽管在讲解神经网络概念的时候，神经网络的输入指的是一个向量 x_i 。

但在实际训练中有：

- 随机梯度下降法（Stochastic Gradient Descent）：用一个样本的梯度来更新权重。
- 批量梯度下降法（Batch Gradient Descent）：用多个样本梯度的平均值来更新权重。

如下图所示，绿、蓝、黑的箭头表示三个样本的梯度更新网络权重后loss的下降方向。

若用**多个梯度的均值**来更新权重的批量梯度下降法可以用相对少的训练次数遍历完整个训练集，其次可以使更新的方向更加贴合整个训练集，避免单个噪音样本使网络更新到错误方向。

然而**也正是因为平均了多个样本的梯度**，许多样本对神经网络的贡献就被其他样本平均掉了，相当于在每个epoch中，训练集的样本数被缩小了。batch中每个样本的差异性越大，这种弊端就越严重。

一般的解决方法就是在每次训练完一个epoch后，将训练集中样本的顺序打乱再训练另一个epoch，不断反复。这样重新组成的batch中的样本梯度的平均值就会与上一个epoch的不同。而这显然增加了训练的时间。

同时因为没办法保证每次更新的方向都贴合整个训练集的大方向，只能使用较小的学习速率。这意味着训练过程中，一部分steps对网络最终的更新起到了促进，一部分steps对网络最终的更新造成了干扰，这样“磕磕碰碰”无数个epoch后才能达到较为满意的结果。

注：一个epoch是指训练集中的所有样本都被训练完。一个step或iteration是指神经网络的权重更新一次。

为了解决这种“不效率”的训练，BN首先是把所有的samples的统计分布标准化，降低了batch内不同样本的差异性，然后又允许batch内的各个samples有各自的统计分布。所以，

> BN的优点自然也就是允许网络使用较大的学习速率进行训练，加快网络的训练速度（减少epoch次数），提升效果。

做法

设，每个batch输入是 $\boldsymbol{x} = [x_0, x_1, x_2, \ldots, x_n]$ （其中每个 x_i 都是一个样本， n 是batch size）假如在第一层后加入Batch normalization layer后， h_1 的计算就倍替换为下图所示的那样。

- 矩阵 x 先经过 W_{h_1} 的线性变换后得到 s_1
  - **注**：因为减去batch的平均值 μ_B 后， b 的作用会被抵消掉，所以没必要加入 b （红色删除线）。
- 将 s_1 再减去batch的平均值 μ_B ，并除以batch的标准差 $\sqrt{\sigma_B + \epsilon}$ 得到 s_2 。 ϵ 是为了避免除数为0的情况所使用的微小正数。
  - $\mu_B = \frac{1}{m}\sum_{i=0}^{m} W_{h_1} x_{i,:}$
  - $\sigma_B^2 = \frac{1}{m}\sum_{i=0}^{m}(W_{h_1} x_{i,:} - \mu_B)^2$
  - **注**：但 s_2 基本会被限制在正态分布下，使得网络的表达能力下降。为解决该问题，引入两个新的parameters： γ 和 β 。 γ 和 β 是在训练时网络自己学习得到的。
- 将 s_1 乘以 γ 调整数值大小，再加上 β 增加偏移后得到 s_3 。
- 为加入非线性能力， s_3 也会跟随着ReLU等激活函数。
- 最终得到的 h_1 会被送到下一层作为输入。

需要注意的是，上述的计算方法用于在**训练**。因为测试时常会只预测一个新样本，也就是说batch size为1。若还用相同的方法计算 μ_B ， μ_B 就会是这个新样本自身， $s_1 - \mu_B$ 就会成为0。

所以在测试时，所使用的 μ 和 σ^2 是整个训练集的均值 μ_P 和方差 σ_P^2 。

而整个训练集的均值 μ_P 和方差 σ_P^2 的值通常也是在训练的同时用移动平均法来计算，会在下一篇代码演示中介绍。

「真诚赞赏，手留余香」

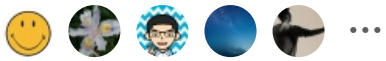
赞赏

1 人赞赏

机器学习 神经网络 深度学习（Deep Learning）

☆ 收藏 分享 举报

79



13 条评论

写下你的评论...

Kevin 回复 **郭家琪** 查看对话

参考中心极限定理

2 个月前 1 赞

张嘉伟

一个小建议：应该详细解释一下为什么需要作用于s1的alpha和beta，与全面被省略的bias又是什么关系

2 个月前 1 赞

以上为精选评论

恐成最大赢家

讲得很好

2 个月前 1 赞

猫狗大战

报告，发现一个错别字。normalization layer后， 的计算就倍替换为下图所示的那样。

2 个月前 1 赞

郭家琪

求问，为什么 s_2 会被限制在正态分布下？谢谢！

2 个月前

王赟 Maigo

为什么BN「把所有的samples的统计分布标准化」，就能够「降低batch内不同样本的差异性」呢？施加在同一个batch内的样本上的是相同的变换呀！

2 个月前

YJango（作者） 回复 **王赟 Maigo** 查看对话

值域的差异性

2 个月前

YJango（作者） 回复 **王赟 Maigo** 查看对话

其实我也不知道该怎么描述这个意思，就是像GMM去建模的话，如果不标准化，就需要更多的components of gaussian去mix的那种意思。

2 个月前

岁月流觞

请问下，公式里的beta和gamma是怎么求出来的啊？

2 个月前

YJango（作者） 回复 **岁月流觞** 查看对话

是学习出来的。反向传播算法，将输入和输出都送入网络后计算出loss，分别把beta和gamma看成是输入来求梯度（各个维度偏导数组成的向量），来不断的用beta = beta-learning*gradient更新beta到一个尽量让loss最小的数值（gamma同理）。

2 个月前

下一页

文章被以下专栏收录

超智能体

分享最通俗易懂的深度学习教程

进入专栏

推荐阅读

律师实习的那些事儿-4

上周跟助理进行了沟通，我问她在实习过程中有什么困惑没有？今天她告诉我，她把自己的困惑和收获写出来了，我前段时间太忙，今天下午抽空进行了交流，经助理同意，把她的困惑和收获的电子版… 查看全文 >

袁海华律师 · 6 天前 · 编辑精选

担心败诉？那就选择仲裁吧！

*本文经授权发布，谢绝无授权转载*没有任何代理律师敢承诺百分之百的胜诉，那么提高胜诉率的… 查看全文 >

建纬（北京）律师事务所 · 7 天前 · 编辑精选

第19篇 肠道里的那些事，就交给益生菌来打理吧

引子事情要从三个月前说起。当时，我带着此前从潭柘寺捡来的猫主子，奔袭几千里，从北京搭着… 查看全文 >

孙亚飞 · 20 天前 · 编辑精选 · 发表于 秀色不可餐，智慧尚能饭

律师江湖师生情

每年教师节,想必是老师们最欣慰的日子:学生的感恩与祝福纷至沓来,教师群体成为大小媒体最为… 查看全文 >

易胜华 · 10 天前 · 编辑精选 · 发表于 人云易云