BAT面试题15：梯度消失与梯度膨胀，以及6种解决措施

编辑: zhenguo

1.梯度消失

根据链式法则，如果每一层神经元对上一层的输出的偏导乘上权重结果都小于1的话，那么即使这个结果是0.99，在经过足够多层传播之后，误差对输入层的偏导会趋于0。

这种情况会导致靠近输入层的隐含层神经元调整极小。

2.梯度膨胀

根据链式法则，如果每一层神经元对上一层的输出的偏导乘上权重结果都大于1的话，在经过足够多层传播之后，误差对输入层的偏导会趋于无穷大。

这种情况又会导致靠近输入层的隐含层神经元调整变动极大。

3. 梯度消失和梯度膨胀的解决方案

本文提供6种常见的解决梯度消失和膨胀的方法，欢迎阅读学习。

3.1 预训练加微调

此方法来自Hinton在2006年发表的一篇论文，Hinton为了解决梯度的问题，提出采取无监督逐层训练方法，其基本思想是每次训练一层隐节点，训练时将上一层隐节点的输出作为输入，而本层隐节点的输出作为下一层隐节点的输入，此过程就是逐层“预训练”（pre-training）；在预训练完成后，再对整个网络进行“微调”（fine-tunning）。

Hinton在训练深度信念网络（Deep Belief Networks）中，使用了这个方法，在各层预训练完成后，再利用BP算法对整个网络进行训练。此思想相当于是先寻找局部最优，然后整合起来寻找全局最优，此方法有一定的好处，但是目前应用的不是很多了。

3.2 梯度剪切、正则

梯度剪切这个方案主要是针对梯度爆炸提出的，其思想是设置一个梯度剪切阈值，然后更新梯度的时候，如果梯度超过这个阈值，那么就将其强制限制在这个范围之内，通过这种直接的方法就可以防止梯度爆炸。

注：在WGAN中也有梯度剪切限制操作，但是和这个是不一样的，WGAN限制梯度更新信息是为了保证lipchitz条件。

关于WGAN(Wasserstein GAN) 的介绍

We introduce a new algorithm named WGAN, an alternative to traditional GAN training. In this new model, we show that we can improve the stability of learning, get rid of problems like mode collapse, and provide meaningful learning curves useful for debugging and hyperparameter searches. Furthermore, we show that the corresponding optimization problem is sound, and provide extensive theoretical work highlighting the deep connections to other distances between distributions.

另外一种解决梯度爆炸的手段是采用权重正则化（weithts regularization）比较常见的是l1正则，和l2正则，在各个深度框架中都有相应的API可以使用正则化，比如在tensorflow中，搭建网络的时候已经设置了正则化参数，则调用以下代码可以直接计算出正则损失：

```
regularization_loss = tf.add_n(tf.losses.get_regularization_losses(scope='my_resnet_50'))
```

如果没有设置初始化参数，也可以使用以下代码计算l2 正则损失：

```
l2_loss = tf.add_n([tf.nn.l2_loss(var) for var in tf.trainable_variables() if 'weights' in var.name])
```

正则化是通过对网络权重做正则限制过拟合，仔细看正则项在损失函数的形式：

$$Loss = (y - W^T x)^2 + \alpha ||W||^2$$

其中，α 是指正则项系数，因此，如果发生梯度爆炸，权值的范数就会变的非常大，通过正则化项，可以部分限制梯度爆炸的发生。

注：事实上，在深度神经网络中，往往是梯度消失出现的更多一些。

3.3 relu、leakrelu、elu等激活函数

Relu:思想也很简单，如果激活函数的导数为1，那么就不存在梯度消失爆炸的问题了，每层的网络都可以得到相同的更新速度，relu就这样应运而生。

Relu的主要贡献在于：

1. 解决了梯度消失、爆炸的问题
2. 计算方便，计算速度快
3. 加速了网络的训练

同时也存在一些缺点：

1. 由于负数部分恒为0，会导致一些神经元无法激活（可通过设置小学习率部分解决）
2. 输出不是以0为中心的

leakrelu就是为了解决relu的0区间带来的影响，其数学表达为：leakrelu=max(k*x,x)

其中k是leak系数，一般选择0.01或者0.02，或者通过学习而来。leakrelu解决了0区间带来的影响，而且包含了relu的所有优点

3.4 batchnorm

Batchnorm是深度学习发展以来提出的最重要的成果之一了，目前已经被广泛的应用到了各大网络中，具有加速网络收敛速度，提升训练稳定性的效果，Batchnorm本质上是解决反向传播过程中的梯度问题。

batchnorm全名是batch normalization，简称BN，

通过规范化操作将输出x规范化以此来保证网络的稳定性。

batchnorm就是通过对每一层的输出规范为均值和方差一致的方法，消除了w带来的放大缩小的影响，进而解决梯度消失和爆炸的问题。

详情可参考文章：

http://blog.csdn.net/qq_25737169/article/details/79048516

3.5 残差结构

事实上，就是残差网络的出现导致了image net比赛的终结，自从残差提出后，几乎所有的深度网络都离不开残差的身影，相比较之前的几层，几十层的深度网络，在残差网络面前都不值一提，残差可以很轻松的构建几百层，一千多层的网络而不用担心梯度消失过快的问题，原因就在于残差的捷径（shortcut）部分。

残差结构说起残差的话，不得不提这篇论文了：

> Deep Residual Learning for Image Recognition

3.6 LSTM

LSTM全称是长短期记忆网络（long-short term memory networks），是不那么容易发生梯度消失的，主要原因在于LSTM内部复杂的“门”(gates)，LSTM通过它内部的“门”可以接下来更新的时候“记住”前几次训练的”残留记忆“，因此，经常用于生成文本中。

4. 总结

文章总结了什么是梯度消失和梯度膨胀；文章大部分篇幅总结了解决这些问题的常用方法，提到了一些经典的论文，有兴趣的可以学习。

本文主要参考博客如下：

https://blog.csdn.net/qq_25737169/article/details/78847691