

工作汇报

Shilong Zhang

9.24至 10.7

1.深度学习理论相关

1.将基本所有GAN的基础知识补上了。

2.读了几个经典GAN网络的代码。

3.CS231。

4.一些机器学习理论相关

1 deep learning的一些相关理论

并非浅层网络表达能力差 并非深层网络相比浅层网络有更好的表示能力，能拟合更复杂的函数，实际上，浅层网络只要有足够的参数，亦可以拟合任意复杂的函数(前提是具有一定的平滑性，是L-Lipschitz的,给定任意小的误差 ϵ ,利用的足够多的Relu激活函数就可以组合出来)，其实，深层网络真正的作用的在拟合相同复杂的函数时候，需要的参数更少。

深层网络的好处 之前证明利用浅层网络可以用Relu来拟合任意L-Lipschitz时候，有一个重要的指标就是网络可以组合出来的分段数目，我们可以证明相同神经元数目下，两层之间分段可以按指数级别增加，而每个层仅仅可以让分段以线性增加。即有关理论研究给出了网络可以产生的Linear Pieces给出的lower Bound，K是每层的宽度，H是深度的话，可以至少产生 K^H 个piece,有理论表明，如果想用一个浅层网络逼近深层网络，往往浅层网络需要非常宽(即参数需要非常多)。

lowlayer 的参数是最重要的 后面网络仿佛是在对折lowlayer产生的pieces，nips上有验证实验，1 不同层的参数加噪声，越往前的层数对噪声更敏感。

2.单层训练，即除去某一层外都随机产生，那么仅仅训练第一层比仅仅训练其他层有高的多的正确率。

通过NN的到底是如何拟合function的？ 可以找到 x^2 的一种通过ReLU进行拟合的方法，通过 x^2 组合出multiply功能，从而可以拟合任意多项式，那么就可以拟合任意的有多项式展开的函数。

optimization deep learning 即如何在网络能表达的函数集中选出最接近target function的,nn的loss function 并不是convex的，即寻求最有解是困难的，但是很幸运，虽然不是convex的，但是近期研究表明，loss function似乎具有很好的性质。有实验表明，deep learning 的local minima 与 global minima几乎是差不多的，loss function 的梯度为零的点在 training error 较大的点有较大的可能是saddle point，在training error较小的点有较大的可能是local minima,但是在large network 里面local minima是很少见的。

影响最后收敛的结果的因素 发现采用不同的initialization与不同的优化算法，经过训练之后，虽然最后的loss function的值差不多，但是最后的实际上网络参数差别非常多，并且哪怕是在训练的最后阶段更换优化算法，最后的收敛的loss都会不同，也可以说，似乎每种优化算法都会有他自己的特性，在不知道哪种优化算法更是和问题时候，可以采用不同的算法进行优化。

Generation Gap 原本机器学习理论在训练集上loss差不多的时候，我们应该选择capacity更小一些的model，这样trainin loss 与eneralization loss在概率上差别更小。大门时deep learning 中有神奇的现象，在training loss已经收敛时候(比如降到0)，这时候继续增大paramter的数目，在 test set上的误差还会下降。有人认为deep network是自带regularization的，deep network 会倾向于用一种比较简单的方式去拟合数据，也就是说虽然deep network的capacity非常强，代表的function space非常大，但是它并没有我们想的那么容易 over fitting。但是，当数据比较差的时候，还是会出现deep network 强记数据的情况，即over fitting，发现L2 norm 在network 过拟合的时候会比较大，可以用来防止过拟合。

泛化能力与sensitivity 可以用jacobian矩阵来表示出网络对数据的sensitivity, 实验表明sensitivity与网络对数据的泛化能力是正相关的, network对分布在training data附近的数据sensitvoity比较低, 也就是对training 附近的数据预测比较稳定, 因此做数据扩充是有必要的。那么对一批test data,就可以先计算数据的sensitivity,让网络仅仅预测sensitivity比较低的数据(答对的概率较大), 而sensertivity比较大的数据就单独交给人来标记预测, 还是有很大实用性的。

泛化能力与sharpness 现在认为loss函数的local minima周围比较sharp的泛化能力会比较差, 而比较flat的泛化能力会比较好, 这貌似可以解释deep learning的泛化能力会比较好, 因为initialize时候很容易初始化在把比较平坦的地方, 因为flat的相比sharp的还是在高维空间中占据更大的空间。而sharp还是flat一个重要影响因素是训练时候的batch size, 小batch size的loss local minima 一般比较flat, 泛化能力就比较好(是有实验做这样的事情), 但是最近有文章论证shape的local minima 也有很好的泛化能力。