MLP

林子清
2212279
智能科学与技术
nkuailzq@gmail.com

1 Training

我们首先使用最基础的配置进行训练:
`Linear(768*256)+ReLU+Linear(256*10)+CrossEntropyLoss+SGD(lr=1e-1)`

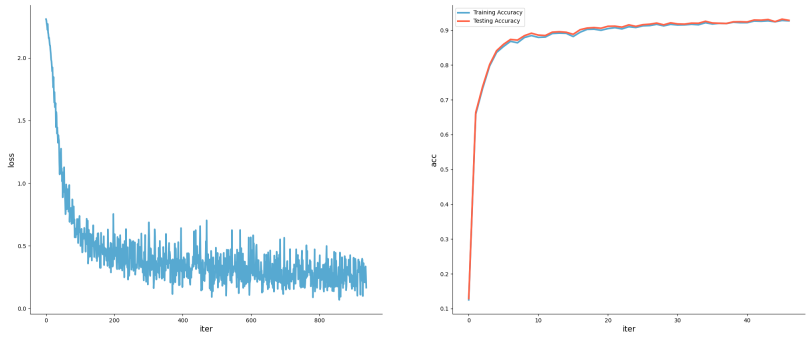

图 1: Base Model

1.1 Activation Function

这里我们尝试改用`GELU` 这种常见的激活函数和较为罕见的 `Maxout` 激活函数并给出了一种基于`pytorch` 的`Maxout` 实现。

`GELU` 激活函数解决了`ReLU` 激活函数再零点处导数不连续的问题，其往往在`transformer` 架构中被使用，但在本例中模型为较小的全连接网络，看不出明显优势。

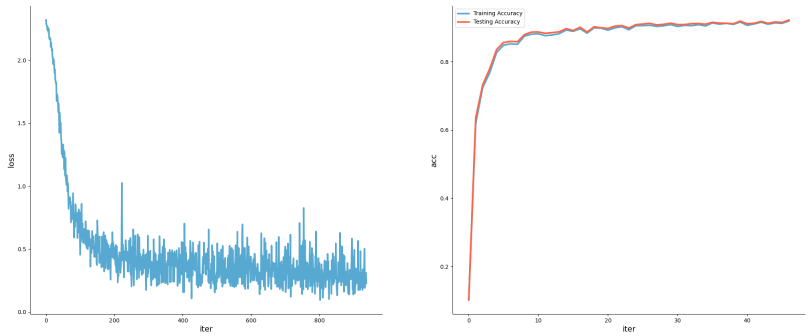

图 2: GELU

使用Maxout 激活函数在本问题中得到了 2% 的提升, 这也证明了其分段拟合能力相对ReLU 的优越性。但这里采用k=2 的maxout 使模型增加了 $(256 \times 256 + 256) \times 2 = 131584$ 个参数，训练达到稳定的时间变长，同时其对优化器要求也更高，很容易出现梯度爆炸的情况

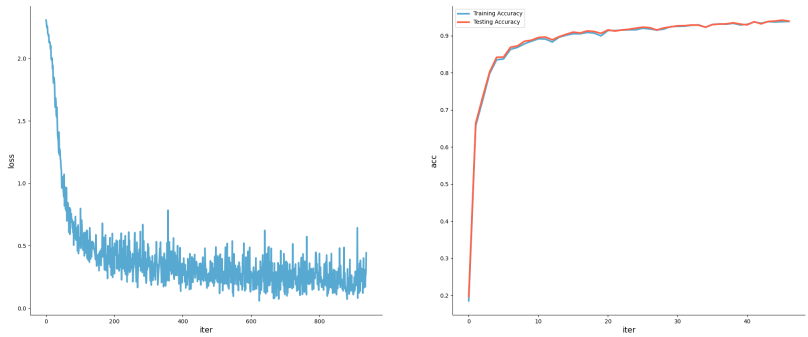

图 3: maxout

1.2 Loss Function

这里我们尝试改用KLDivLoss 损失函数。

KLDivLoss 可以比对两个概率分布，因此我们先使用log_softmax 将输出转化为概率分布，但目标概率分布只能使用独热编码生成概率分布，同时将其松弛化。但这种方法终究不如CrossEntropyLoss，训练速度和稳定精度都不如。

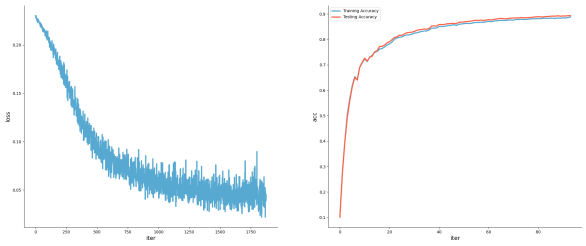

图 4: KLDivLoss

2 Parameters

基础网络有 $768 \times 256 + 256 + 256 \times 10 + 10 = 199434$ 个参数。

2.1 Size of Hidden Layers

一般情况下隐藏层神经元个数越多，模型总参数越多，模型拟合能力越强，从 24 个神经元提升至 2048 个神经元可以提升 3.5%。这一般是因为在中间层神经元过少时，信息会被过度压缩。

$784 \times 12 + 12 + 12 \times 10 + 10 = 9550$

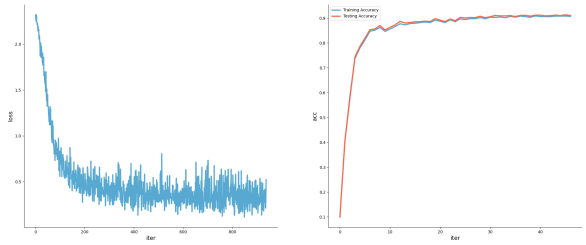

图 5: 784

$784 \times 784 + 784 + 784 \times 10 + 10 = 623290$

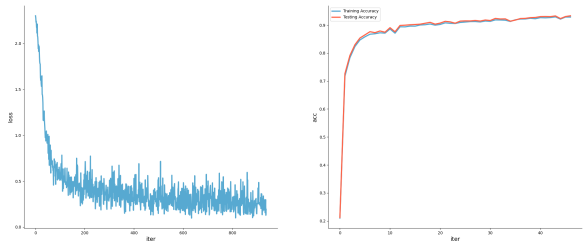

图 6: 784

$784 \times 2048 + 2048 + 2048 \times 10 + 10 = 1628170$

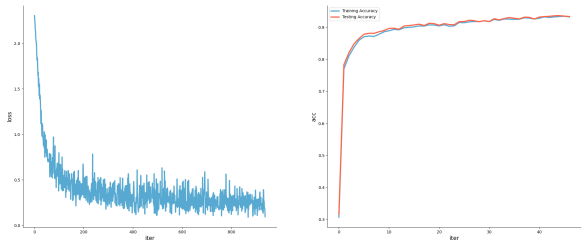

图 7: 2048

2.2 Number of Hidden Layers

增加隐藏层数和增加隐藏层大小都可以增大模型总参数，但是往往增加模型深度更有效，例如本例中与 784 隐藏层相比增加隐藏层数减小了模型参数量但使得效果提升了 1%. $784 \times 512 + 512 + 512 \times 256 + 256 + 256 \times 10 + 10 = 535818$

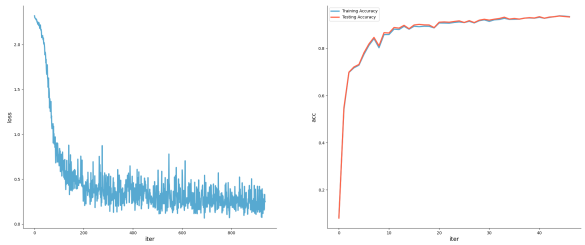

图 8: 3

3 Generalization

泛化性能是一个模型好坏的重要标准，在训练集上训练的模型若出现过拟合现象，会导致训练精度高于测试精度的问题，但在本例中，这种现象并不存在。虽然如此，但我们依然可以尝试一些正则化技术，其主要思想便是将信息均匀的训练到网络中。

3.1 Regularization

在模型的损失函数中添加模型参数的正则化项是最简单的方法，其阻值某些参数变得过大从而使其在网络中占据主导地位，这里使用L2 范数。同时可以证明的是L2 正则化与权重衰减是等价的。

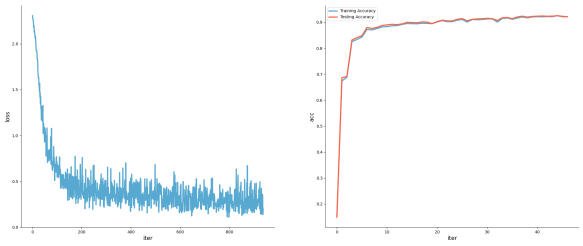

图 9: L2

3.2 Dropout

添加`Dropout` 层则从另一种方式减小过拟合，其每次只使用部分网络进行训练，减小了网络对于某些参数的过度依赖。

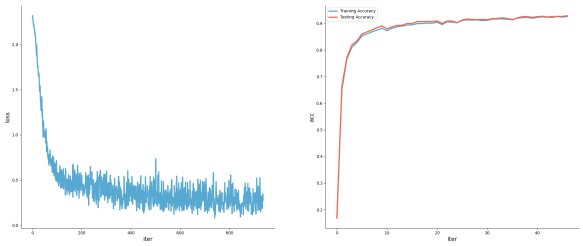

图 10: dropout

3.3 Batch Normalization

考虑到中间层的数据分布在训练过程中很可能产生漂移，我们可以添加`BN` 层来缓解这一现象。但一些研究表明其减小过拟合的有效性并非是这个原因，而是其让训练中每个目标的结果与其所在 batch 相关联从而不会过分关注某个特定样本点。

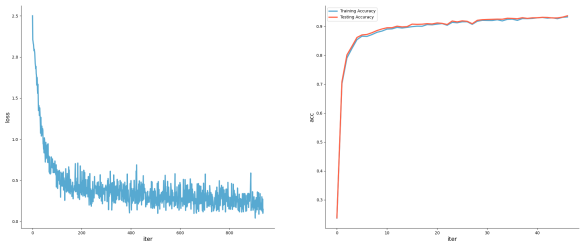

图 11: BN

4 XOR

我们用同样的方法训练了一个 XOR 模型。

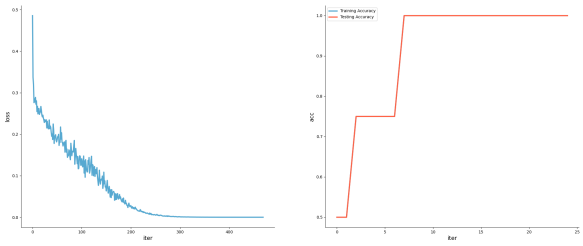

图 12: XOR