**Bag of Tricks for Image Classification with Convolutional Neural Networks**

**Bag of Tricks for Image Classification**

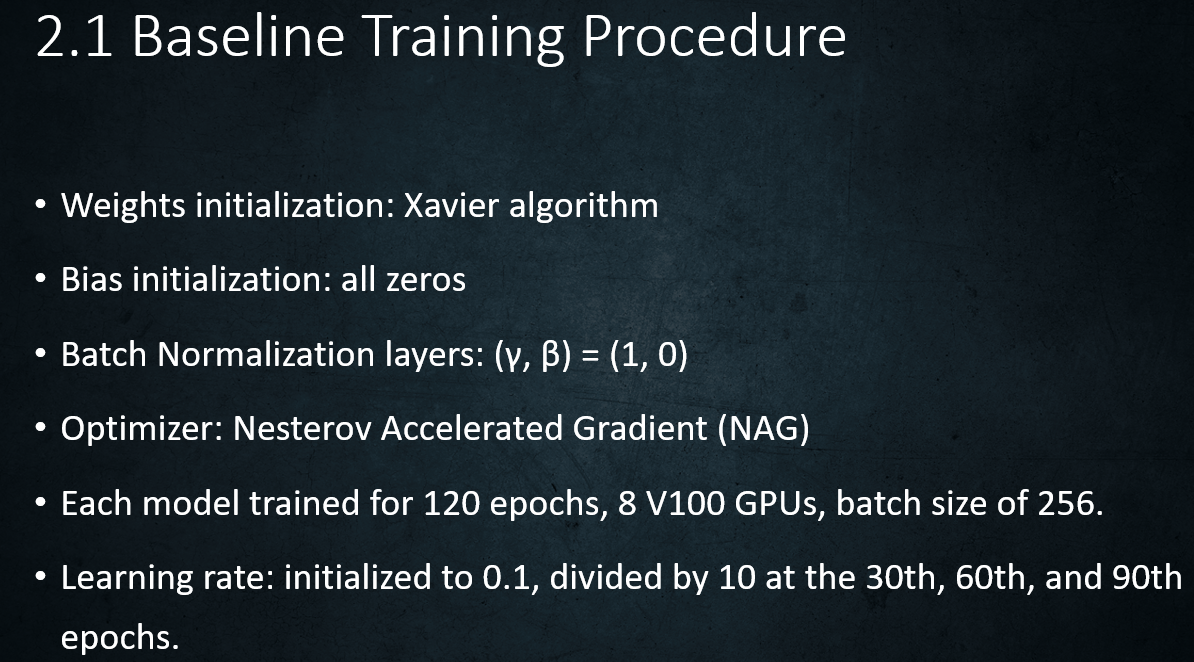
一些图像分类训练的技巧

baseline benchmark是什么？

baseline是一个基准，比如论文中常常会有一个基础架构的网络，看它的精度是多少，这个基本网络称之为baseline，而后加了一些技巧（trick）算法优化和调参观察它与baseline相比提升了多少的精度，可以说是性能的基准点。

benchmark是一种目前领域中的标准，人们对benchmark研究的很深，用能于跟其它方法作为基准，如果benchmark是SOTA算法是最好的。所以总而言之，benchmark衡量就是要用于说明方法的突破性、好效果。

224像素图片: imageNet是224\*224 ，最后的特征图 7\*7大小比较好，那么原图大小就要7\*2的k次方。常常输入的图片300的比较多，而224是7\*2的k次方中最接近300的了



说明了训练需要考虑的过程

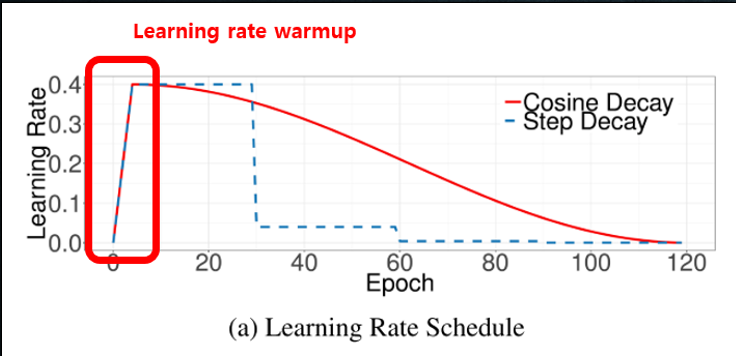
**Linear Scaling Rule，**当 Batch size 变为 K 倍时，Learning rate 需要乘以 K 就能够达到同样的训练结果。

**Learning Rate Warmup:**

[**什么是warm-up**](https://stackoverflow.com/questions/55933867/what-does-learning-rate-warm-up-mean)

为了防止一开始的时候网络就像某个方向拟合，将初始的学习率调低。

而warm-up便用来在最初的训练阶段设置较小的学习率，而后逐渐增加至需要的学习率。线性: 公式： 最终需要的学习率p，一共有n个阶段进行warm-up，那么第i个epoch的学习率设置为 p\*i/n。 其中n是一个超参数



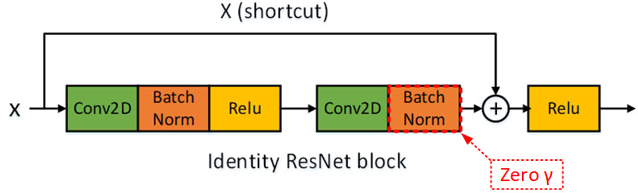
**zero γ**

ResNet由多个残差块组成，可以简略表示为x + block(x)

在它们之后是一个批处理的归一化层，这些曾进一步按比例转化为γ \* (x) + bias

训练技巧，将γ初始化为0，bias也初始化为0.（默认情况下γ为1，bias为0）

这样残差块仅返回其输入，模拟层数较少且在初始阶段更容易训练。



**No Bias decay**

decay衰减通常会被应用到所有可训练参数中，但是这会导致过拟合。所以一个策略是将正则化仅仅应用于卷积层和全连接层。对于bias和批处理参数则不进行decay

[**正则化的理解**](https://morvanzhou.github.io/tutorials/machine-learning/ML-intro/3-09-l1l2regularization/)

**low precision training**

**mixed precision training 混合精度训练**

在某些地方用较低精度训练可以加快速度，而又不会损失准确率

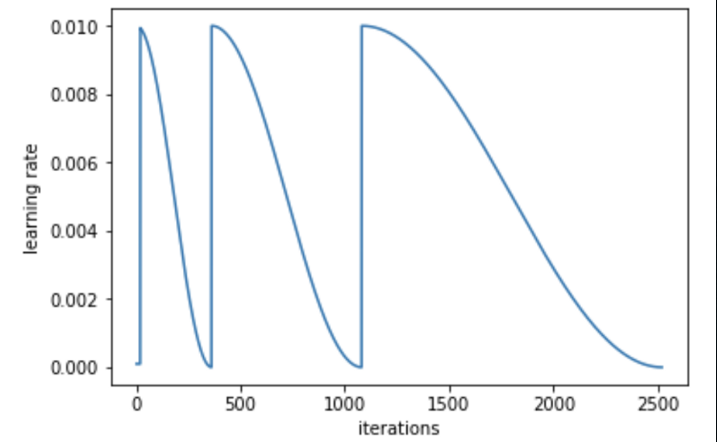
**training refinement**

**Cosine Learning Rate Decay**

learning rate decay 学习率衰减也是需要考虑的一种调参技巧，学习率在开始时较大，从而能够快速接近一个局部最优的位置，接着学习率慢慢江地，从而愈发接近这个最优值。

相较于step decay而言，cosine learning rate decay策略能够让这个寻找过程比较平滑

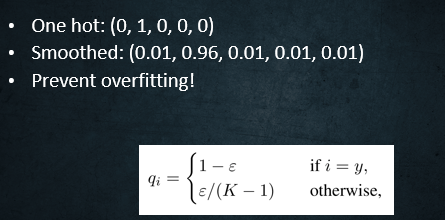
**SGDR: Stochastic Gradient Descent with Warm Restarts**



为了找到更稳定的局部最优解，可以时不时提高学习率，从而有从一个局部到另一个局部的能力，这个过程称为restart

**Label Smoothing 标签平滑**

一种正则化表达方式。



如果损失函数是交叉熵的话，网络最终会取向于让某一类的概率尽可能大，而其它类的概率尽可能小，从而导致了过拟合。

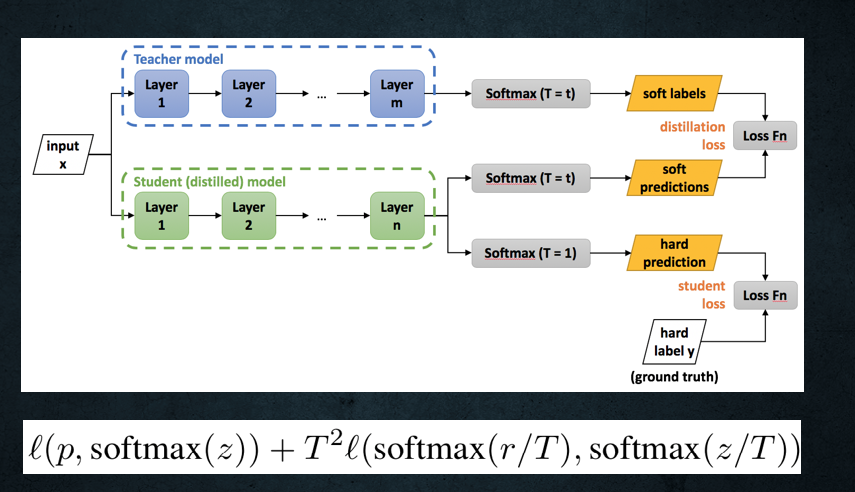
因此采用label smoothing使得最终网络有更好的泛化能力。

**knowledge distillitation**

知识蒸馏是一种模型压缩方法，有些像迁移学习，可以将复杂、学习能力强的网络作为teacher model，而后用这个model学习到的知识传递给参数少、学习能力较弱的student model的网络。

有很多已经训练的很好的网络，就不用重复训练了。从预训练(pre-trained)的teacher model转移到当前的student network

加入一个蒸馏损失函数用于衡量学生网络的预测结果与老师网络结果的偏离。

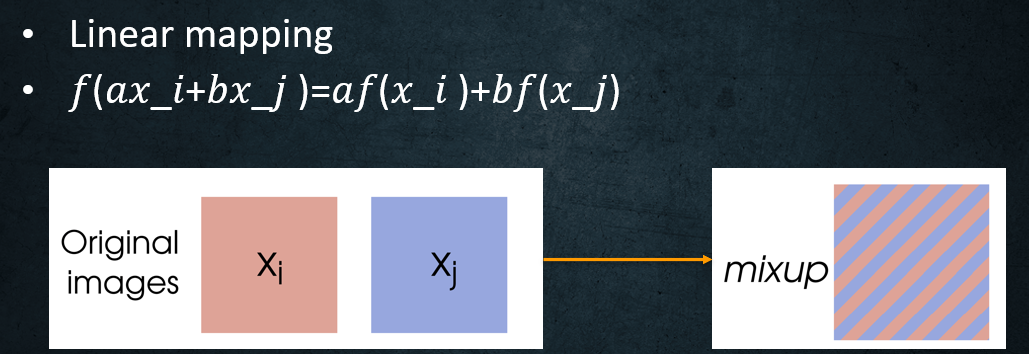


**mix-up**

**图像混合技术。**

常用的数据增强有翻转、小角度旋转等

该方法就是两图混合





防止过拟合，增加一些扰动，提升模型的泛化能力。

**其它讲这篇博客的博文**

[**文章一**](https://zhuanlan.zhihu.com/p/66393448)

[**https://zhuanlan.zhihu.com/p/66612268**](https://zhuanlan.zhihu.com/p/66612268)