模型剪枝

- 模型剪枝

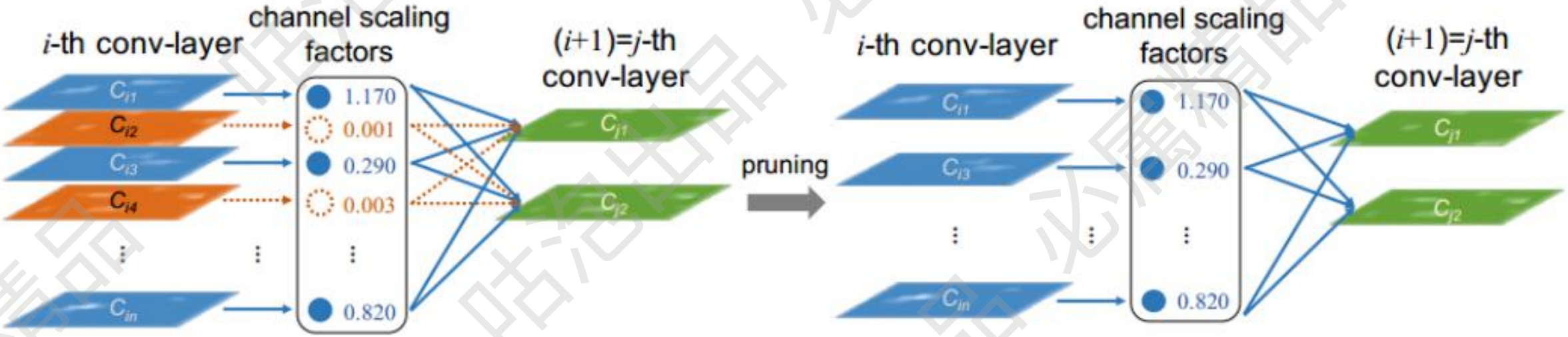
  - ICCV经典论文，通俗易懂！
  - 卷积后能得到多个特征图，这些图一定都重要吗？
  - 训练模型的时候能否加入一些策略，让权重参数提现出主次之分？
  - 这两点就是论文的核心，先看论文再看源码其实并不难！

参考论文：Learning Efficient Convolutional Networks through Network Slimming

模型剪枝

- 模型剪枝

  - 基本思想，这一张图就足够解释了



参考论文：Learning Efficient Convolutional Networks through Network Slimming

模型剪枝

- 如何得到每个特征图的重要性呢?

  - Network slimming，就是利用BN层中的缩放因子γ

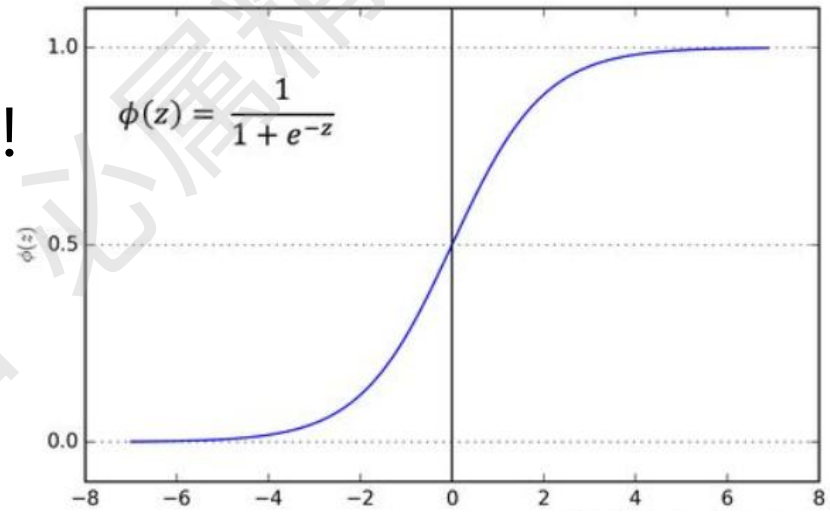
  - 先来回顾下BN是做什么的：$$\hat{x}^{(k)} = \frac{x^{(k)} - E[x^{(k)}]}{\sqrt{Var[x^{(k)}]}}$$

  - 整体感觉就是一个归一化操作，但是BN中还额外引入了两个可训练的参数：γ和β。

参考论文：Learning Efficient Convolutional Networks through Network Slimming

模型剪枝

- BatchNorm
  - 如果训练时候输入数据的分布总是改变，网络模型还能学的好吗？
  - 对于卷积层来说，它的输入可不是只有原始输入数据
  - 如果对每层的学习结果不加以限制可能会出问题！
  - 以sigmoid为例，很多输出值越来也偏离，导致模型收敛越来越难！



参考论文：Learning Efficient Convolutional Networks through Network Slimming

模型剪枝

- BatchNorm的本质
  - BN要做的就是把越来越偏离的分布给他拉回来！
  - 再重新规范化到均值为0方差为1的标准正态分布
  - 这样能够使得激活函数在数值层面更敏感，训练更快
  - 有一种感觉：经过BN后，把数值分布强制在了非线性函数的线性区域中

参考论文：Learning Efficient Convolutional Networks through Network Slimming

模型剪枝

- BatchNorm参数
  - 如果都是线性的了，神经网络还有意义吗？
  - BN另一方面还需要保证一些非线性，对规范化后的结果再进行变换
  - 这两个参数是训练得到的：$y^{(k)} = \gamma^{(k)} \hat{x}^{(k)} + \beta^{(k)}$
  - 感觉就是从正太分布进行一些改变，拉动一下，变一下形状！

参考论文：Learning Efficient Convolutional Networks through Network Slimming

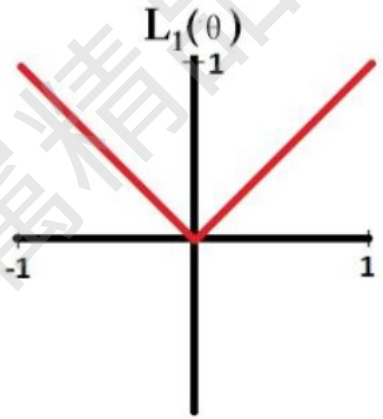
模型剪枝

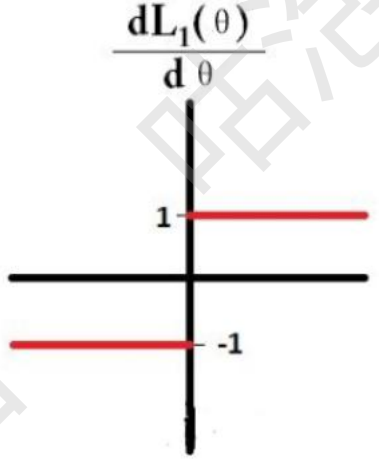
- L1与L2正则化
  - 论文中提出：训练时使用L1正则化能对参数进行稀疏作用
  - L1：稀疏与特征选择；L2：平滑特征
  - L1正则化：$J(\vec{\theta}) = \frac{1}{2}\sum_{i=1}^{m}\left(h_{\vec{\theta}}\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}|\theta_j|$
  - L2正则化：$J(\vec{\theta}) = \frac{1}{2}\sum_{i=1}^{m}\left(h_{\vec{\theta}}\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\theta_j^2$

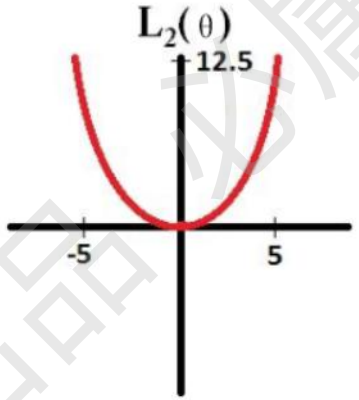
参考论文：Learning Efficient Convolutional Networks through Network Slimming

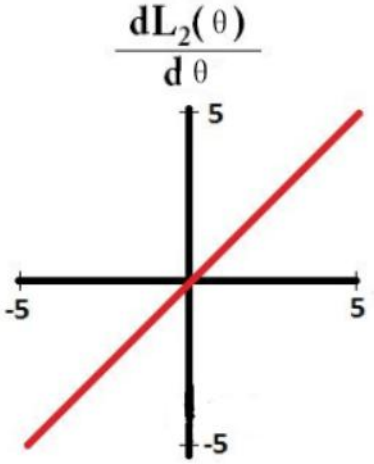
模型剪枝

- L1与L2正则化
  - L1求导后为：sign(θ)
  - 相当于稳定前进，都为1，最后学成0了
  - L2求导为：θ
  - 相当于越来越慢，很多参数都接近0，平滑









参考论文：Learning Efficient Convolutional Networks through Network Slimming

模型剪枝

- 论文核心点

  - 以BN中的γ 为切入点，即γ越小，其对应的特征图越不重要

  - 为了使得γ 能有特征选择的作用，引入L1正则来控制γ

$$L = \sum_{(x,y)} l(f(x, W), y) + \lambda \sum_{\gamma \in \Gamma} g(\gamma)$$

参考论文：Learning Efficient Convolutional Networks through Network Slimming

模型剪枝

论文核心点

训练-剪枝-再训练，整体流程如下图所示：

参考论文：Learning Efficient Convolutional Networks through Network Slimming

模型剪枝

论文核心点

部分实验结果：

(a) Test Errors on CIFAR-10

Model	Test error (%)	Parameters	Pruned	FLOPs	Pruned
VGGNet (Baseline)	6.34	20.04M	-	7.97×10^8	-
VGGNet (70% Pruned)	**6.20**	2.30M	88.5%	3.91×10^8	51.0%
DenseNet-40 (Baseline)	6.11	1.02M	-	5.33×10^8	-
DenseNet-40 (40% Pruned)	**5.19**	0.66M	35.7%	3.81×10^8	28.4%
DenseNet-40 (70% Pruned)	5.65	0.35M	65.2%	2.40×10^8	55.0%
ResNet-164 (Baseline)	5.42	1.70M	-	4.99×10^8	-
ResNet-164 (40% Pruned)	**5.08**	1.44M	14.9%	3.81×10^8	23.7%
ResNet-164 (60% Pruned)	5.27	1.10M	35.2%	2.75×10^8	44.9%

参考论文：Learning Efficient Convolutional Networks through Network Slimming