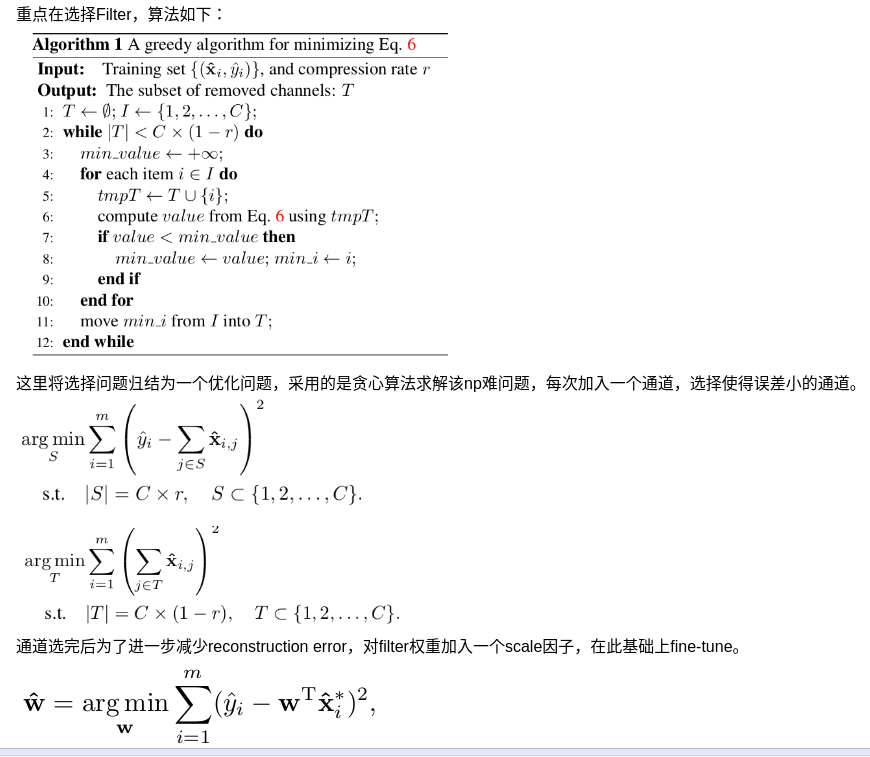
通道剪枝（channel Pruning）

1.《ThiNet: A Filter Level Pruning Method for Deep Neural Network Compression》

南大：罗建豪 吴建新

没有开源code

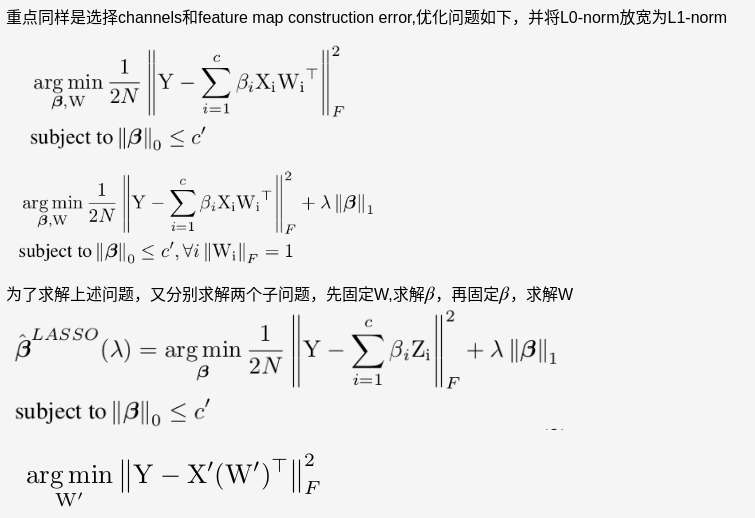
大致思想：贪心算法求解



2.《Channel Pruning for Accelerating Very Deep Neural Network》

Face++出品

作者开源code: <https://github.com/yihui-he/channel-pruning>(基于caffe)



3.《Pruning Convolutional Neural Networks For Resource Efficient Inference》

Nvidia出品

复现code：<https://github.com/jacobgil/pytorch-pruning>（Pytorch）

利用泰勒公式将feature map和loss联系起来，得出pruning标准

4. 《Pruning Filters For Efficient ConvNets》

无开源code

标准简单粗暴：L1-norm

5.《Auto-balanced Filter Pruning for Efficient Convolutional Neural Networks》

无开源实现

基本思想：通过正则化方法，将Filter表达能力转移到一小部分，剪掉其余部分。

参数剪枝（parameter pruning）

1.《Dynamic Network Surgery for Efficient DNNs》

作者开源：https://github.com/yiwenguo/Dynamic-Network-Surgery（caffe）

基本思想：本文解决方法为剪枝和嫁接(pruning and splicing)，通过更新参数的重要程度来实现的。

### 量化

1.《Incremental Network Quantization: Towards Lossless CNNs with Low-Precision Weights》（ICLR 2017）

作者开源：<https://github.com/Zhouaojun/Incremental-Network-Quantization（caffe>）

基本思想：现有工作都是采用全局量化即同时对所有权重都进行量化，而本文提出对预训练好的模型权重采取逐渐量化方式。分为3步：1.将每层权重分成两组(weight partition)，2.第一组被量化形成低精度weights(group-wise quantization)，3.另一组再训练用来补偿量化后的精度损失(re-training)。第一组量化完成，接下来对第二组重复以上操作直到所有的权重都量化成低精度。（量化成2的幂次方和0）

2.《XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks》

作者开源：<https://github.com/allenai/XNOR-Net> （torch 7.0）

复现code: <https://github.com/jiecaoyu/XNOR-Net-PyTorch> (Pytorch)

*This is a deep learning framework for cnn which using xnor method to accelerate calculating convolution layer on cpu*

<https://github.com/luhaofang/CACU>

3. 《Trained Ternary Quantization (TTQ)》

作者开源：<https://github.com/czhu95/ternarynet>（Tensorflow, TensorPack）

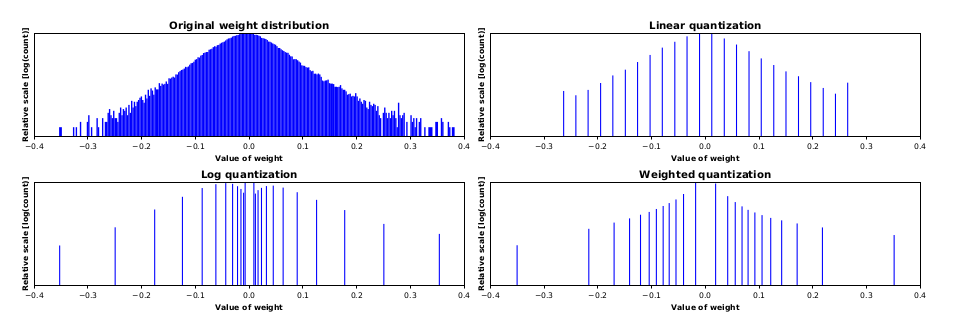
复现code: <https://github.com/TropComplique/trained-ternary-quantization> (Pytorch)

基本思想：用可训练三元组量化整个网络

4. 《Weighted-Entropy-based Quantization for Deep Neural Networks》

无开源实现

基本思想：提出weighted-entropy量化



5. 《Deep Learning with Low Precision by Half-wave Gaussian Quantization》

作者开源：<https://github.com/zhaoweicai/hwgq（caffe>）

基本思想：提出用ReLU-variant作为激活函数，使用k-means离线量化

6. 《Network Sketching: Exploiting Binary Structure in Deep CNNs》

没有开源

基本思想：对weights进行高阶残差量化，并进一步提出加速收敛方法

7.《PACT: Parameterized Clipping Activation for Quantized Neural Networks》  
没有开源

基本思想：提出类ReLU激活函数，用可训练参数来调节阈值减少量化误差。

8. 《DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients》

作者开源code: <https://github.com/ppwwyyxx/tensorpack/tree/master/examples/DoReFa-Net> (tensorpack, tensorflow) 还包含了BWN, TTQ，BNN实现。

基本思想：首次量化了梯度

9. 《Deep Neural Network Compression with Single and Multiple Level Quantization》

作者没开源（上交的，可以拿到caffe实现）

基本思想：和《Incremental Network Quantization: Towards Lossless CNNs with Low-Precision Weights》一样，提出一部分量化，一部分训练，逐渐量化所有参数。量化采用kmeans先分组，量化误差大的先量化，等等。

