第0周工作汇报

核心思想

减少网络模型中参数量和计算量，同时尽量保证模型的性能不受影响。

太大的模型产生的问题

- 神经元数量太多，计算延迟难以达到应用的标准
- 模型部署的时候，部分设备对计算能力和内存空间有严格限制

提出的原因

- 实际的思考：类似于神经网络提出的原因，人脑中的神经元突触会随着年龄的增长，先增加后减少一些无用突触，根据这个现象，引出了模型剪枝的思考
- 较为科学的解释：在训练阶段，在数学上需要进行大量的微分求解，去捕抓数据中的微小变化信息，一旦完成迭代式的训练之后，网络模型推理的时候就不需要这么多参数，那么可不可以直接不管那些无用的参数，用可靠的参数来训练

剪枝算法的可行性

彩票假说

论文《The Lottery Ticket Hypothesis: Training Pruned Neural Networks》有证明

训练成功的大网络包含子网络，这个子网络从头开始训练可以达到大网络的精度，这个网络也可以叫做中奖彩票。至于，为什么叫彩票假说，因为一群人如果有一人中奖，那么把这个人单独拿出来，他得的奖和这群人的总奖金相同。两件事原理相通。

但是《Rethinking the value of Network Pruning》提出了相反的意见，到时候细读看看

剪枝算法分类

实际上大部分刚接触剪枝算法的时候，都会从从宏观层面去划分剪枝技术，主要是分为Drop Out和Drop Connect两种经典的剪枝算法。

1）Drop Out：随机的将一些神经元的输出置零，称之为神经元剪枝。

2）Drop Connect：随机将部分神经元间的连接Connect置零，使得权重连接矩阵变得稀疏。

如何去剪枝

细粒度划分： 根据修剪的最小单位区分方法

- 非结构性剪枝：权重剪枝、向量剪枝、kernel剪枝，这些剪枝算法可以造成模型结构的不规则化，所以这些方法需要特殊的硬件设计来支持稀疏操作，但是这些模型剪枝较为精细，所以剪枝后精度较高。
- 结构性剪枝：卷积核剪枝、通道剪枝和层级剪枝，只需改变网络中卷积核和特征通道的数目，所得到的模型就可以运行，无需特殊的算法设计（一般用这种）

解决方式划分：启发式、优化式

- 启发式：手动定义每个修剪单元的重要性，以删除神经网络中不重要的单元

  基于幅度的剪枝方法：通过大小直接衡量每个单独权重的重要性

  激活值，APoZ:零元素占比，BN层的规模因子，引入浮点

  剪枝效率很低

  sol:文章将每层的浮点数计算量引入进来，来引导剪枝算法对计算量较多的层的剪枝：

- 优化式剪枝：将模型剪枝看作一个优化问题来自动寻找剪枝位置

  L1，L2正则化，BN层的规模化因子

  会有模型残留问题

  Zeroing-out:提出截断冗余参数的目标函数反传的梯度，只允许参数的衰减，最终直至为零。

  Centripetal constraint:将多个相近的卷积核推到一点，导致一个冗余的模式，当多个卷积核在参数超空间中被约束得越来越近时，我们称之为向心约束，尽管它们开始产生越来越相似的信息，即下一层相应输入通道传递的信息仍在充分利用，因此模型的表示能力比滤波器为零的对应模型更强输出（没看懂）

从哪些方面剪枝

宽度决定了输出维度（特征的丰富程度），网络的深度代表了模型的非线性转换能力。大多数修剪工作都选择一个维度进行修剪。然而，很容易达到极限，这限制了模型的压缩率和精度。论文《Accelerate cnns from three dimensions: A comprehensive pruning framework》从三个维度depth、width、resolution对模型进行剪枝，上图展示了不同维度对于剪枝的敏感度不同，所以最大的问题是如何分配三个维度的剪枝率，然后利用了拉格朗日乘子法获得最优解

剪枝的流程

- 标准剪枝：主要包含三个部分：训练、剪枝、微调、再剪枝。
- 基于子模型采样的剪枝：对训练好的原模型中可修剪的网络结构，按照剪枝目标进行采样，对采样后的网络结构进行剪枝，得到采样子模型。
- 基于搜索的剪枝：依靠强化学习等一系列无监督学习或者半监督学习算法

一些问题

- 为什么剪枝之后模型压缩率变低了，但是最终精确度却变高了
- 两篇论文种提到的不同的观点（下周看完论文再仔细说明）
- 老师有没有推荐的模型
- 模型残留问题的第二个
- 我应该具体做什么，是提出新算法压缩老模型，还是用老算法压缩新模型
- 子模型是什么意思

总结

了解了剪枝算法在模型压缩的中的意义，为什么要进行模型压缩，剪枝算法对模型压缩中的核心思想是什么。

剪枝算法的合理性解释，分生物解释和数学解释。目前粗略的对剪枝算法有个大概的了解，包括分类，流程，从哪些方面进行剪枝等等

完成了github仓库的创建，后续相关内容和进展会更新到github上。Alizen-1009/HHU_Graduation-Project_mode-pruning: 河海大学智能科学与技术专业，本科毕业设计 (github.com)

下周安排

1.要阅读的论文

- 《Accelerate cnns from three dimensions: A comprehensive pruning framework》这篇论文探讨了三个维度的对模型精度的影响
- 《The Lottery Ticket Hypothesis: Training Pruned Neural Networks》和《Rethinking the value of Network Pruning》中相反的意见

2.在网上搜罗相关模型，找到一个合适的，了解掌握这款模型的结构，为后续剪枝做准备

3.阅读老师发的github代码和相关论文（MIM）