# 目标

现状fp16 30ms 目标 5ms

# 模型压缩技术实现

当我们利用深度学习训练出我们的模型之后，下一步会做什么呢？无疑是去测试我们的模型，精度达到标准既可发布上线，那发布之后我们会发现的问题是->推理时间慢，那我们该怎么做呢？以下几个思路可供借鉴，这里以yolov5为例：

更改网络为轻量级，如YOLOv5s、YOLOv5n

通过技术手段进行模型加速，如tensorrt，openvino等

资金雄厚的情况下，可以考虑使用更高算力的卡

模型压缩

接下来，就给大家介绍下发现的一个用于模型压缩的工具–>PaddleSlim

剪枝，量化，蒸馏

<https://blog.csdn.net/qq_55068938/article/details/128132402>

这里我们需要提前准备好要转换的pt模型，并提前转换为onnx或paddle的格式。

学术界的 SOTA 模型在落地部署到工业界应用到过程中，通常是要面临着低延迟（Latency）、高吞吐（Throughpout）、高效率（Efficiency）挑战的。而模型压缩算法可以将一个庞大而复杂的预训练模型转化为一个精简的小模型，从而减少对硬件的存储、带宽和计算需求，以达到加速模型推理和落地的目的。

近年来主流的模型压缩方法包括：数值量化（Data Quantization，也叫模型量化），模型稀疏化（Model sparsification，也叫模型剪枝 Model Pruning），知识蒸馏（Knowledge Distillation）， 轻量化网络设计（Lightweight Network Design）和 张量分解（Tensor Decomposition）。

其中模型剪枝是一种应用非常广的模型压缩方法，其可以直接减少模型中的参数量。本文会对模型剪枝的定义、发展历程、分类以及算法原理进行详细的介绍。

# 剪枝方法分类

剪枝

一般先训好，然后剪枝，在微调

如果调用工具直接压缩，可以直接测试推理时间和模型大小

技术实现：调研方法，实验看效果

（pruning）是一种模型压缩的方法，它可以通过去除神经网络中冗余的参数或结构，来减少模型的存储和计算开销，同时保持模型的性能1。剪枝的方法可以分为三类：

非结构化剪枝（unstructured pruning）：按照参数的大小或重要性，去除单个的权重或神经元。

结构化剪枝（structured pruning）：按照参数的分布或敏感度，去除整个的通道或过滤器23。

自动化剪枝（automated pruning）：利用强化学习或贝叶斯优化等方法，自动确定每层的剪枝比例。

下表列出了一些常见的剪枝方法和它们的分类：

剪枝方法 分类 参考

Optimal Brain Damage 非结构化 LeCun et al., 1990

Network Slimming 结构化 Liu et al., 2017

AMC 自动化 He et al., 2018

非结构化剪枝：最简单的，通过权重排序，删除一些大于threshold的权重

## 结构化剪枝

（structured pruning）是一种模型压缩的方法，它可以通过移除神经网络中冗余的结构，如通道或过滤器，来减少模型的参数量和计算量1。结构化剪枝的优点是可以在现有的框架上实现模型的加速，而不需要额外的稀疏矩阵运算2。

结构化剪枝的方法可以根据剪枝的粒度分为三类：

滤波器剪枝（filter pruning）：滤波器剪枝是移除卷积层中整个的过滤器，从而减少输出通道的数量1。这样可以减少模型的参数量和计算量，但也会降低模型的表达能力2。。

通道剪枝（channel pruning）：通道剪枝是移除卷积层中整个的输入通道，从而减少输入通道的数量1。这样可以减少模型的内存占用和数据传输量，但也会改变模型的输入输出结构3。

层剪枝（layer pruning）：移除神经网络中整个的层，从而减少网络的深度2。

下表列出了一些常见的结构化剪枝方法和它们的分类：

剪枝方法 分类 参考

Network Slimming 滤波器剪枝 Liu et al., 2017

CP-Decomposition 通道剪枝 Lebedev et al., 2014

SkipNet 层剪枝 Wang et al., 2018

### 通道剪枝

是针对卷积层的通道，而不是输入图像的通道。卷积层的通道是指卷积核的个数，也就是输出特征图的深度。通道剪枝是通过删除一些不重要的卷积核，来减少输出特征图的深度，从而降低模型的复杂度。

Channel/Filter 剪枝

channel 剪枝的工作是最多的，channel 剪枝和 filter 剪枝其实意义是一样的，一个过滤器移除了，对应输出 feature map 的一个通道自然也被移除，反之一样。

filter (channel) pruning (FP) 属于粗粒度剪枝（或者叫结构化剪枝 Structured Pruning），基于 FP 的方法修剪的是过滤器或者卷积层中的通道，而不是对个别权重，其原始的卷积结构不改变，所以剪枝后的模型不需要专门的算法和硬件就能够加速运行。

CNN 模型中通道剪枝的核心在于如何减少中间特征的数量，其中一个经典思路是基于重要性因子，即评估一个通道的有效性，再配合约束一些通道使得模型结构本身具有稀疏性，从而基于此进行剪枝。

基于重要性因子的方法进行通道剪枝，和前面非结构化剪枝中的基于权重幅度的方法来进行连接剪枝类似，都有点主观性太强。

一个filter等同于一个卷积核：只是指定了卷积核的长宽深；

而有的情况（例如tensorflow等框架中，filter参数（一个卷积层）通常指定了卷积核的长、宽、深、个数四个参数），filter包含了卷积核形状和卷积核数量的概念：即filter既指定了卷积核的长宽深，也指定了卷积核的数量。

通道剪枝的主要步骤有：

通道选择（channel selection）：根据一定的标准或指标，选择要保留或删除的通道。常用的标准或指标有：

L1范数：计算每个通道的权重矩阵的L1范数，越小的通道越容易被剪掉3。

BN层系数：利用BN层中的缩放系数（scale factor）或者标准差（standard deviation）来衡量每个通道的重要性，越小的通道越容易被剪掉4。

梯度信息：利用反向传播过程中每个通道对损失函数的梯度来衡量每个通道的重要性，越小的通道越容易被剪掉5。

重构（reconstruction）：在保持输出特征图不变的约束下，重新训练或优化剪枝后的模型，以恢复模型性能。常用的方法有：

最小二乘法（least squares）：最小化剪枝后输出特征图和原始输出特征图之间的均方误差3。

知识蒸馏（knowledge distillation）：利用原始模型作为教师模型，指导剪枝后的模型作为学生模型进行学习。

微调（fine-tuning）：在原始数据集上对剪枝后的模型进行再训练，以调整参数。