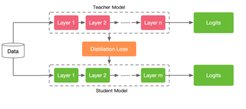
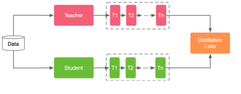
推理加速与部署

- 推理系统
- 模型小型化
  - CNN小型化
    - 论文
      - SqueezeNet
      - ShuffleNet
      - MobileNet
      - ESPnet
      - FBNet
      - EfficientNet
      - GhostNet
    - 小型化维度
      - 卷积核
        - 大卷积核用小卷积核代替
        - 单一尺寸卷积核使用多尺度卷积核代替
        - 固定形状卷积核趋于使用可变性卷积核
        - 使用1x1卷积，构造bottlenet结构
      - 卷积层通道
        - 标准卷积用depthwise卷积代替
        - 使用分组卷积
        - 分组卷积前使用channel shuffle
        - 通道加权计算
      - 卷积层连接
        - 使用skip connection，让模型更深
        - densely connection，融合其他层特征输出
  - transformer小型化
    - 论文
      - Q8BERT
      - DistilBERT
      - TinyBERT
      - Training data-efficient image transformers & distillation through attention
      - MiniViT
      - TinyViT
      - DynamicViT
      - Compressing Visual-linguistic Model via Knowledge Distillation
      - MiniVLM
- 离线优化压缩
- 模型转换与优化
  - 低比特量化
    - 目标
      - 寻找数据向低比特映射的scale和zero
    - 分类
      - 感知量化Quantization Aware Traning(QAT)
        - QAT在训练过程中将量化引入的误差以噪声的形式加入训练中，然后进行训练。可以得到量化后较高的准确率
      - 训练后量化Post-Training Quantization(PTQ)
  - 二值化网络
  - 模型剪枝
    - 剪枝分类
      - 非结构化剪枝
        - 随机对独立的权重或者神经元连接进行剪枝
        - 优劣势
          - 剪枝算法简单，模型压缩比高
          - 精度不可控，剪枝后权重矩阵稀疏，需要专用硬件来实现压缩和加速
      - 结构化剪枝
        - 对filter/channel/layer进行剪枝
        - 优劣势
          - 大部分算法在channel或者layer上进行剪枝，保留原始卷积结构，不需要专用硬件
          - 剪枝算法相对复杂
    - 剪枝子模块
      - 训练：训练参数模型，得到最佳网络性能，得到基础模型
      - 剪枝：根据算法对模型剪枝，调整网络结构中通道或层数，得到剪枝后的网络
      - 微调：在原始数据集上进行微调，用于重新弥补因剪枝后的稀疏模型丢失的精度性能
    - 举例
      - L1 norm剪枝
        - 对每个卷积核，计算权重绝对值
        - 根据卷积核的权重绝对值进行排序
        - 对m个权重绝对值之和最小的卷积核及对应的feature map进行剪枝
        - 下一个卷积层中与剪掉的feature map相关的卷积核也要剪枝
  - 模型蒸馏
    - 蒸馏知识分类
      - Response-Based Knowledge
        - 将teacher model的输出层特征作为student model的学习对象
        - 

      - Feature-Based Knowledge
        - 将teacher model的各层特征作为student model的学习对象
        - 

      - Relation-Based Knowledge
        - 将样本之间或特征层之间的关系蒸馏到student model中
        - 

    - 蒸馏方法分类
      - offline distillation
      - online distillation
      - self-distillation
- kernel优化
- runtime优化