- NLP方向总结-模型蒸馏
  - 知识蒸馏
    - 多层对多层映射的BERT蒸馏
      - BERT-EMD
        - https://zhuanlan.zhihu.com/p/266602585
        - 特点
          - 多层对多层映射策略
          - Earth Mover's Distance（EMD）衡量teacher与student网络之间的距离
          - 几部分的蒸馏的损失之和
        - 假设teacher网络具有M个Transformer层，而student网络具有N个Transformer层。每个Transformer层都包含一个注意力层和一个隐藏层
          - 嵌入层蒸馏
            - 均方误差（MSE）最小化
            - 一对一蒸馏
          - transformer蒸馏
            - 将网络层视为分布，并且所需的转换应使两个分布（teacher层和student层）接近
              - 基于注意力的蒸馏，是学生网络的注意力和教师网络注意力之间的EMD距离
                - $\mathcal{L}_{\mathrm{attn}} =\operatorname{EMD} \left(\mathbf{A}^{S}, \mathbf{A}^{T}\right)$
              - 基于隐藏的蒸馏，是学生网络的隐藏层和教师网络隐藏层之间的EMD距离
                - $\mathcal{L}_{\text {hidden }} =\operatorname{EMD} \left(\mathbf{H}^{S}, \mathbf{H}^{T}\right)$
          - 预测层蒸馏
            - 一对一蒸馏
              - $\mathcal{L}_{\text {pred }}=-\operatorname{softmax} \left(\mathbf{z}^{\mathrm{T}}\right) \cdot \log _{-\operatorname{softmax}} \left(\mathbf{z}^{\mathrm{S}} / \mathrm{t}\right)$
            - 学生网络从teacher网络提供的logits概率中学习
    - Bert 蒸馏到简单的BiLSTM
      - DistilBert
        - https://zhuanlan.zhihu.com/p/273543240
    - student模型耐心地从teacher模型的多个中间层中学习以进行增量知识提取
      - BERT-PKD
        - https://zhuanlan.zhihu.com/p/274329168
    - 一般蒸馏+特定知识蒸馏的
      - TinyBERT
        - 预训练和fine-tuning都进行了transformer蒸馏
        - https://zhuanlan.zhihu.com/p/273467698
    - 通过逐步更换模块压缩BERT模型
      - BERT-of-Theseus
        - https://zhuanlan.zhihu.com/p/283118184
    - 用于资源限制设备的紧凑型任务型BERT
      - MobileBERT
        - https://zhuanlan.zhihu.com/p/365329984
  - 模型压缩技术
    - 低秩矩阵分解
    - 剪枝：权重剪枝
    - 知识蒸馏
    - 量化
  - 知识蒸馏工具包
    - TextBrewer
      - https://zhuanlan.zhihu.com/p/275722016
    - mmrazor
      - 支持剪枝，知识蒸馏，量化，神经架构搜索
    - mmdeploy
      - 模型中间格式转换
    - repDistiller
      - 12个知识蒸馏算法的模型
  - 论文
    - AdaShare: Learning What To Share For Efficient Deep Multi-Task Learning
      - AdaShare
        - 本文通过自动寻找适合的多任务得网络配置
          - 为每个层l和任务Tk寻求一个二元随机变量ul,k（又称策略），它决定了在解决Tk时是选择执行还是跳过深度神经网络中的第l层，以获得最佳共享模式，在任务集T上产生最佳的整体性能。
        - 结构学习策略
          - 这个层对于一个任务是跳过还是执行
            - 鼓励任务之间通过共享块进行积极的共享，必要时通过使用特定任务的块来尽量减少消极的干扰
          - 引入了两个策略正则化，以实现紧凑的多任务网络中的有效知识共享
            - 稀疏性正则化Lsparsity
              - 提高效率，提高模型的紧凑性。
            - Lsharing损失
              - 鼓励跨任务的残差块共享
          - 一个课程学习策略，以稳定早期阶段的优化
            - 随着epoch的增加，逐渐学习每个块的决策策略
          - 通过反向传播优化选择或跳过策略U和网络权重W
            - Gumbel-Softmax Sampling 解决不可微问题
              - 用相应的Gumbel-Softmax分布中的可微分样本替代离散分布中的原始非可微分样本
        - 实验
          - 数据集
            - NYU，CityScapes，TinyTaskonomy，DomainNet，sogou_news
          - 任务
            - 语义分割、表面正常预测、深度预测、关键点检测和边检测， 图像分类，文本分类
          - 结果对比：表1
            - Cross-Stitch网络，Sluice网络，NDDR-CNN，MTAN，DEN
          - 参数量下降，计算开销（FLOPs）下降，性能提升