- NLP方向总结-对抗学习
  - 论文
    - Adversarial Training for Large Neural Language Models
      - ALUM
        - 定义
          - 模型结构名字ALUM：Adversarial training for large neural LangUage Models
          - 鲁棒性：模型具有较高的精度或有效性，对于噪声的容忍力高
          - 泛化性：指学习到的模型对未知数据的预测能力
          - 对抗性训练：对抗训练就是通过添加扰动构造一些对抗样本，放给模型去训练，提高模型的鲁棒性和抵御对抗性攻击，同时一定程度也能提高模型的表现和泛化能力。
        - 亮点
          - 过去的工作都是对抗性训练提高鲁棒性，降低泛化性
          - 本方法对抗性训练不仅提高了鲁棒性还提高了泛化性
        - 标准训练和对抗性训练结构
          - 标准训练目标（公式1）
            - 对训练数据的标准误差最小化：训练目标分别来自自监督（预训练中的MLM和NSP）或直接监督（特定任务微调中的标注实例）
          - 对抗性训练目标
            - CV领域是通过对输入图像施加小的扰动来修改训练目标，使对抗性损失最大化
            - NLP领域是离散的，不太好直接添加扰动
          - ALUM的对抗结构
            - 原理：不是直接对输入文本进行扰动，而是对嵌入空间进行扰动
            - 目标：使用虚拟对抗性训练来正则化标准目标（公式3）
              - 公式3是标准误差和稳健误差的和
            - 训练方式：采用课程学习方法
              - 首先使用标准目标训练模型
              - 然后用虚拟对抗训练继续训练。
        - 实验
          - 测试BERT模型和RoBERTa模型
          - ALUM使用方式
            - 从头开始的预训练模型时使用ALUM对抗性训练
              - BERT：一台有16个V100 GPU训练10天，训练了100万步
              - RoBERTa：一台有16个V100 GPU训练7天，训练了10万步
              - 部分任务性能提升1.2%到2.3%
            - 继续预训练模型时使用：部分任务性能提升+0.5%到0.7%
            - 特定任务的微调时使用：部分任务性能提升2.1%到7.3%
            - 组合方式：训练预训练模型+微调时使用：部分任务性能提升+1.1%到5.5%