NLP方向总结-对比学习

- 定义
  - 在嵌入空间中把一个锚和一个“正面”样本拉在一起，并把锚与许多“负面”样本推开
  - 给定一个锚点，使其与正样本的相似度最大化，并使其与小批次中许多负样本的预期相似度最小化。
  - 指标学习
  - Contrastive Learning
  - https://github.com/KevinMusgrave/pytorch-metric-learning
- 优化
  - N-pairs损失
    - https://github.com/ChaofWang/Npair_loss_pytorch
    - In Proceedings of the 30th International Conference on Neural Information Processing Systems
    - 每次采样一个正样本以及多个负样本，使用对比学习loss计算损失。
    - 采样的负样本数量受到GPU内存限制，因此这种方法的每次更新能够使用到的负样本数量非常有限，影响了对比学习效果。
    - 使用一个正例和许多负例
  - Memory Bank
    - 解决End-to-End负样本采样数量受GPU内存限制，某一轮模型对数据集中所有样本的表示存储起来，这些样本在作为负样本时，可以不进行梯度更新，极大提升了每个batch负样本数量
  - Momentum Encoder
    - 主要为了解决Model Bank中每个样本缓存的表示和Encoder更新不一致的问题
    - 论文：MoCo
    - 跨批次的负样本对，在队列中存储以前批次的表示，并在损失中使用它们作为负样本,允许较小的批次规模，但稍微改变了损失。
  - In-Batch Negtive
    - 假设每个batch内有N个正样本对，那么让这N个正样本之间互为负样本，这样每个样本就自动生成了2*(N-1)个负样本。
    - 缺点是负样本只能使用每个batch内的数据，是随机采样的，无法针对性的构造负样本。而且需要极大的批次规模才能很好地工作。
    - eg: CLIP模型
  - 三元组hinge损失
    - 缺点：负样本对是从其他类别中选择的，几乎总是需要困难负样本挖掘以获得良好的性能
    - 每个锚只使用一个正例和一个负例样本,即使用一个正负样本对
    - 三元祖损失是对比性损失的一个特例
  - 最大边际损失
  - SupCon损失
    - 论文：Supervised Contrastive Learning
    - 当使用一个以上的负样本时，我们表明SupCon损失变得等同于N-pairs损失
  - lifted structures
    - Deep metric learning via lifted structured feature embedding.
  - quadruplets
    - a deep quadruplet network for person re-identification
  - angular loss
    - Deep metric learning with angular loss. In Proceedings of the IEEE International Conference on Computer Vision
  - adapted triplet loss
    - Taskonomy: Disentangling task transfer learning.
  - 多相似度损失
    - Multi-similarity loss with general pair weighting for deep metric learning.
- 论文：
  - Supervised Contrastive Learning
    - SupContrast：有监督的对比学习
      - 定义
        - SupCon损失
          - 新的对比损失函数，是对三元祖损失和N-pairs损失的融合
          - 对每个锚使用许多正例和许多负例样本使我们能够达到最先进的性能
          - 超参数稳定性，不敏感
          - 公式2或公式3
            - 推广到任意数量的正样本
            - 对比力随着负样本的增多而增加
            - 执行困难正/负样本挖掘的内在能力
          - 代码：https://github.com/HobbitLong/SupContrast
      - 测试模型
        - 数据增强模块
          - 对于每个输入样本x，产生两个随机的增强
        - 编码器网络
          - 两个增强的样本分别输入到同一个编码器，产生一对表示向量
        - 投影网络
          - 2个表示向量输入到多层感知器，然后归一化，再家计算它们的距离（内积）
  - A Simple Framework for Contrastive Learning of Visual Representations
    - SimCLR:视觉表示对比学习的简单框架
      - 定义
        - SimCLR: Simple Framework for Contrastive Learning of Visual Representations
        - https://github.com/google-research/simclr
        - 对比学习
          - 数据增强很重要，在对比学习中
          - 在表示之后加入可学习的非线性转换，例如MLP和线性层，然后进行使用对比损失，大大改善性能
          - 正则化和适当温度调整更好
          - 对比性学习受益于更大的批次规模和更多的训练
      - 框架（图2）
        - 随机的数据增强模块
          - 三种简单的增强：随机裁剪，然后调整到原始尺寸，随机颜色扭曲，以及随机高斯模糊
          - 随机抽取N个例子的mini-batch，并将对比性预测任务定义在来自mini-batch的一对增强的例子上，从而得到2N个数据点，增强后的同一个样本是一个正样本对，将mini-batch中的其他2（N-1）个增强的例子视为负样本
        - 神经网络基础编码器
          - 例如ResNet
        - 一个小的神经网络投影头$g(\cdot)$，将表示映射到应用对比性损失的空间
      - 其它
        - LARS优化器
  - Sentence-BERT：Sentence Embeddings using Siamese BERT-Networks
    - Sentence-BERT：使用Siamese BERT 网络的句子嵌入
      - 定义
        - 代码
          - https://github.com/UKPLab/sentence-transformers
        - SBERT：Sentence-BERT
        - 使用孪生网络和三元组网络结构生成句子嵌入，可以使用余弦相似度进行比较
      - 模型（图1）
        - 句子嵌入的3种策略
          - 取CLS的token的向量
          - 默认策略：计算所有输出向量的平均值（MEAN 策略）
          - 计算输出向量的最大随时间推移（MAX 策略）
        - 损失
          - 分类目标函数
            - 句子嵌入 u 和 v 与元素差异 |u - v| 连接起来并将其乘以可训练权重
            - 例如句子对之间的关系：句子对带有 contradiction、eintailment 和 neutral 标签标注
          - 回归目标函数
            - 计算两个句子嵌入 u 和 v 之间的余弦相似度，使用均方误差损失作为目标函数
            - 也可以使用负曼哈顿距离和负欧几里得距离作为相似性度量，结果与余弦相似度类似
          - 三元组目标函数
            - 给定一个锚句 a、一个正样本 p 和一个负样本 n，triplet loss 调整网络，使 a 和 p 之间的距离小于 a 和 n 之间的距离
  - SimCSE: Simple Contrastive Learning of Sentence Embeddings
    - SimCSE：句子嵌入的简单对比学习
      - 定义
        - Contrastive Learning，目的是通过将语义相近的实例拉到一起，将非邻居推开来学习有效的表示
        - 对比学习
          - 一个样本对（xi, x+i），其中xi和x+i在语义上相关
        - 生成样本对
          - CV中
            - 将同一图像的两个随机变换（例如裁剪、翻转、变形和旋转）作为xi和x+i
          - NLP中
            - 应用了删词、重排和替换等增强技术
            - 简单地在中间的表达上使用标准的dropout
              - 本文证明此方法更好
        - 衡量对比学习的表示的好坏
          - 对齐（公式2）
            - 计算配对实例的嵌入之间的预期距离
          - 均匀性（公式3）
            - 均匀性衡量嵌入的均匀分布程度
        - anisotropic
          - 各向异性：在不同的方向不同的特性
          - 例如光照射在紫砂壶上，光的反射就是各项异性的，因为紫砂壶的表面是粗糙的，光反射是各个方向的
          - 优化对比学习目标可以提高均匀性（或缓解各向异性问题）
        - 其它基线模型
          - 平均 GloVe 嵌入，SBERT，SRoBERTa，DeCLUTR，BERT-flow, BERT-whitening,BERT-flow
      - 模型
        - 无监督的方法
          - 只使用标准的 dropout 作为噪声
          - 全连接层上有dropout mask，只需将相同的输入送入编码器两次，得到一对嵌入，然后利用公式4优化
        - 有监督的方法
          - 例如STS数据集，使用“蕴含”对作为正例，将“矛盾”对作为困难负例
          - 如果只有正样本对，使用公式1优化
          - 如果有正样本和负样本，使用公式5优化，其中xi是前提，x+i和x-i是包含性和矛盾性假设