Momentum Contrast for Unsupervised Visual Representation Learning

---

MoCo，其实就是Momentum和Contrastive。前者是动量，涉及到的是这个公式：$y_t = my_{t-1} + (1-m)x_t$，后者指的就是对比学习。

本文为CLIP（多模态、对比学习）的工作打下了基石。

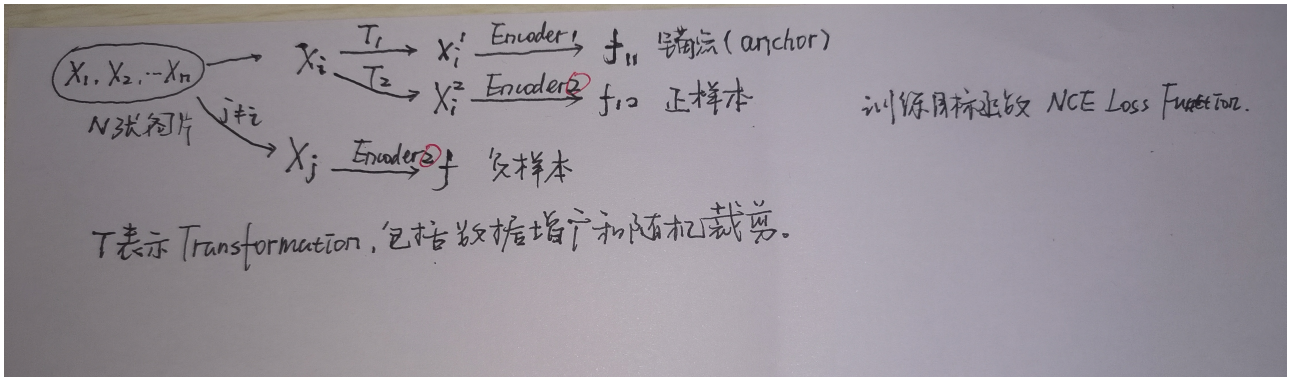
预备知识：

1. 无监督学习（unsupervised learning）

自监督学习（self-supervised learning）（self-supervised learning）是无监督学习的一种。

有两个重点，一是Loss Function（分类，① 生成式Loss、 ② 判别式Loss、③ Contrastive Loss和④ Adversarial Loss），本文的无监督学习要是Contrastive Loss；二是Pretext Task，目的是得到一个“通用”的presentation，之后再transfer到下游任务中。

在计算机领域中，自监督学习是通过设计Pretext Tasks来生成监督信号，从而不需要人为标注的数据集来作为监督信号。Pretext Task例子，个体判别（Instance Discrimination）



2. 对比学习（Contrastive Learning）

动机：

无监督学习在NLP领域中取得了显著的成功，但在CV领域中效果不佳。

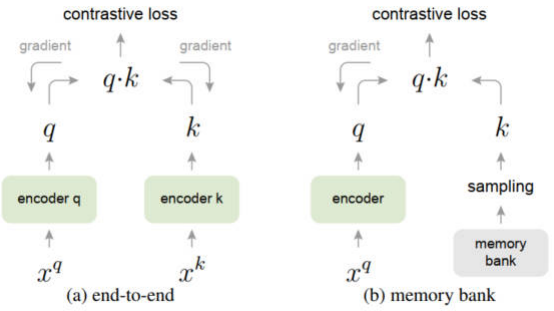
方法：

以对比学习为切入点，将其转换为动态字典查找问题。

前人工作不足之处：

动态字典，有如下两个重点，1) 大小，越大越好；2) 一致性，越能保持一致越好。

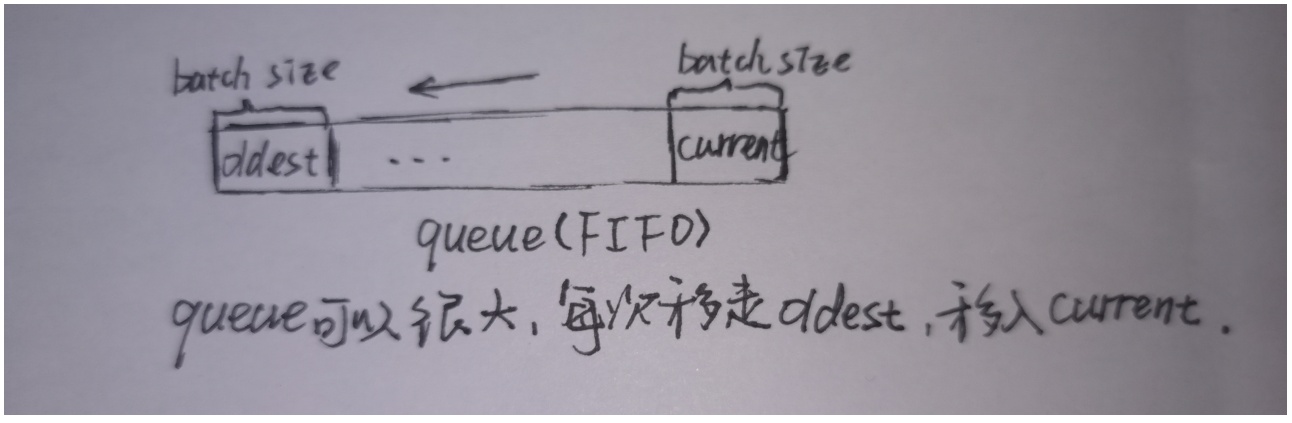
前面工作的研究，只兼顾了其中之一（如，未兼顾到大小——end-to-end；未能兼顾到一致性——memory bank），未能两者都兼顾上。



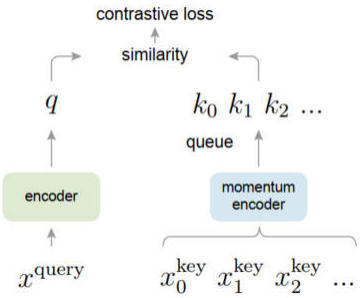
创新点：

1. 兼顾了动态字典的大小和一致性。

大小：通过队列（queue）这种数据结构，将字典大小和batch size大小解耦，从而解除GPU内存大小的限制。



一致性：记encoder的参数是θ_q，momentum encoder的参数是θ_k。θ_q的通过back propagation更新，通过这个式子$\theta_k = m\theta_{k-1} + (1-m)\theta_q$来更新$\theta_k$，为了保持一致性，则要求$\theta_k$和$\theta_{k-1}$尽可能保持一致，所以m的值取得很近似于1。



2. 可以transfer到不同的downstream tasks上。如classification、detection、segmentation等。