MOCO Reading note

1. Method

```
# f_q, f_k : encoder networks for query and key
# queue : dictionary as a queue of K keys    (C×K)
# m: momentum
# t: temperature
f_k.params = f_q.params  # initialize
for x in loader :  # load a minibatch x with N samples      (N =256)
    x_q = aug(x)    # a randomly augmented version
    x_k = aug(x)    # another randomly augmented version

    q = f_q.forward(x_q)  # queries: N×C   256×128
    k = f_k.forward(x_k)  # keys: N×C   256×128
    k = k.detach()  # no gradient to keys   队列里的样本无需梯度回传

    # positive logits: N×1     相当于reshape     256×1
    l_pos = bmm(q.view(N,1,C), k.view(N,C,1))   q·k+
            batch matrix multiplication

    # negative logits: N×K   256 × 65536 默认字典大小
    l_neg = mm(q.view(N,C), queue.view(C,K))   Σ_{i=1}^{K} q k_i

    # logits: N×(1+K)   256×65537
    logits = cat([l_pos, l_neg], dim=1)

    # contrastive loss
    labels = zeros(N)
    loss = CrossEntropyLoss(logits/t, labels)

    # SGD update: query network
    loss.backward()
    update(f_q.params)

    # momentum update: key network
    f_k.params = m·f_k.params + (1-m) * f_q.params
```

# update dictionary
enqueue(queue, k)
enqueue(queue)

$x_1 \longrightarrow \xrightarrow{\text{Transform1}} \underline{x_1^1} \longrightarrow E_{11} \longrightarrow f_{11}$
anchor

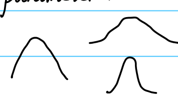
$x_1 \longrightarrow \xrightarrow{\text{Transform2}} \underline{x_1^2} \longrightarrow E_{12} \longrightarrow \underline{f_{12}, f_2, f_3, \cdots, f_N}$
positive　　　　queue

$\underline{x_2, x_3, \cdots, x_N}$ negative

2. contrastive loss function Info$\underline{\text{NCE}}$
noise contrastive estimation

$$L_q = -\log \frac{\exp(q \cdot k_+ / \tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i / \tau)}$$ ⟶ 1个正样本 + K个负样本

① 多分类 ⟶ 二分类：data sample, noise sample
② 每次选K个负样本参与计算，而不是数据中的所有负样本 ⟶ 取近似，K足够大
③ τ是 temperature hyper-parameter：
τ变大 ⟶ 分布平滑 $\xrightarrow{+\infty}$ 对所有样本一视同仁
τ变小 ⟶ 分布集中 $\xrightarrow{0}$ 只关注<u>特别困难的负样本</u> 可能是潜在的正样本



④ 在 cross entropy loss 中，K指数据集里类别的多少
在 infoNCE loss 中，K指负样本数量

3. Views
pretext task
instance discrimination task

contrastive learning ⟶ $\underline{\text{dynamic dictionary}}$ look-up
① large　② consistent

$$\theta_k \longleftarrow m\theta_k + (1-m)\underline{\theta_q}$$
updated by back propogation

decouple 字典大小和 mini-batch 大小