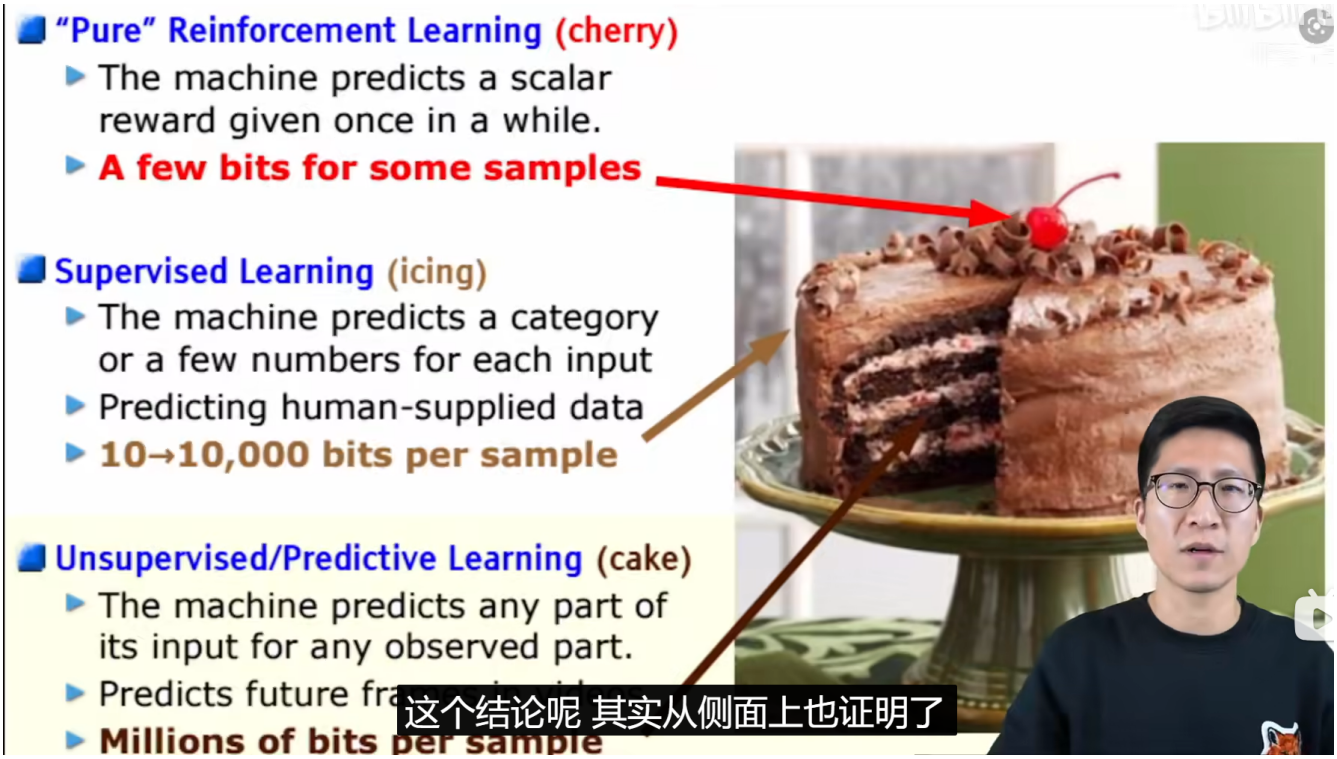
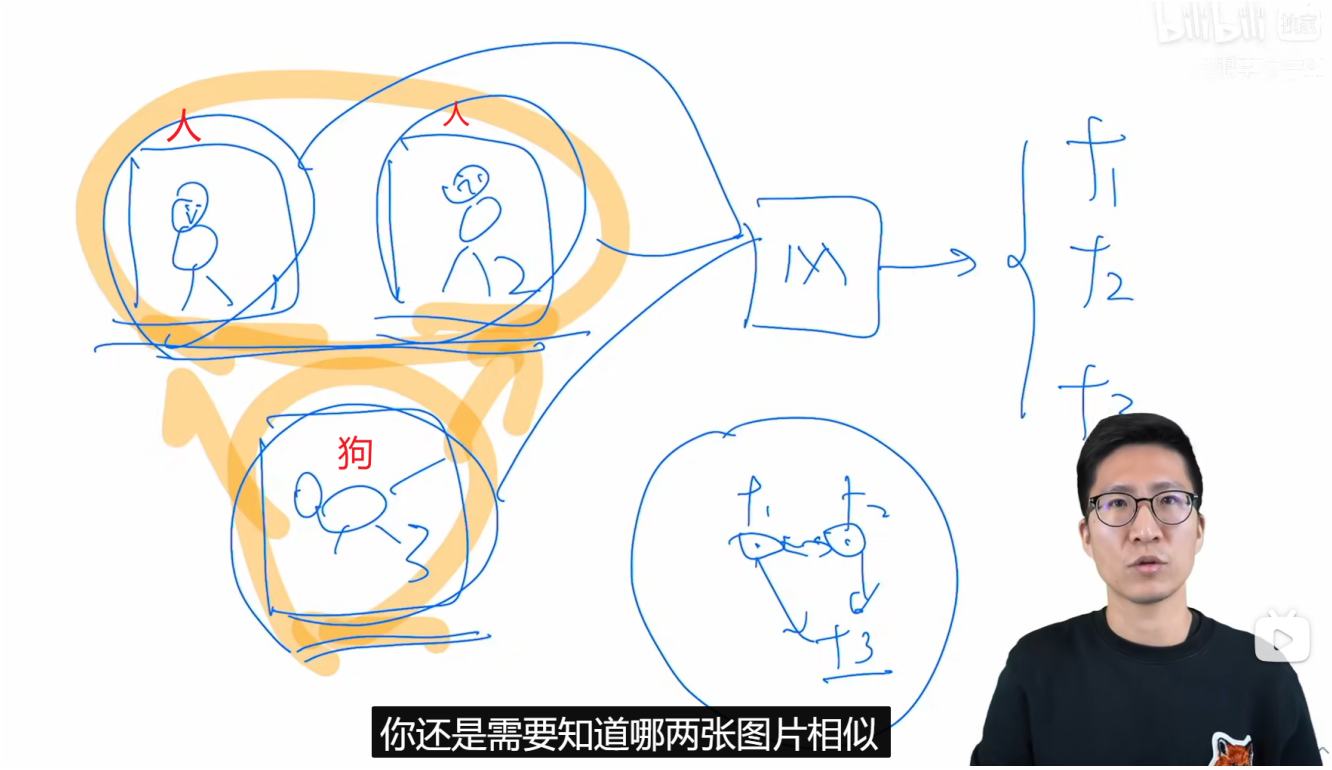
# 重读经典：MOCO

CVPR里2020最佳论文提名，是视觉领域里使用对比学习的里程碑。

Yann LeCun在NeurIPS 2016演讲时说到：




# 对比学习科普：

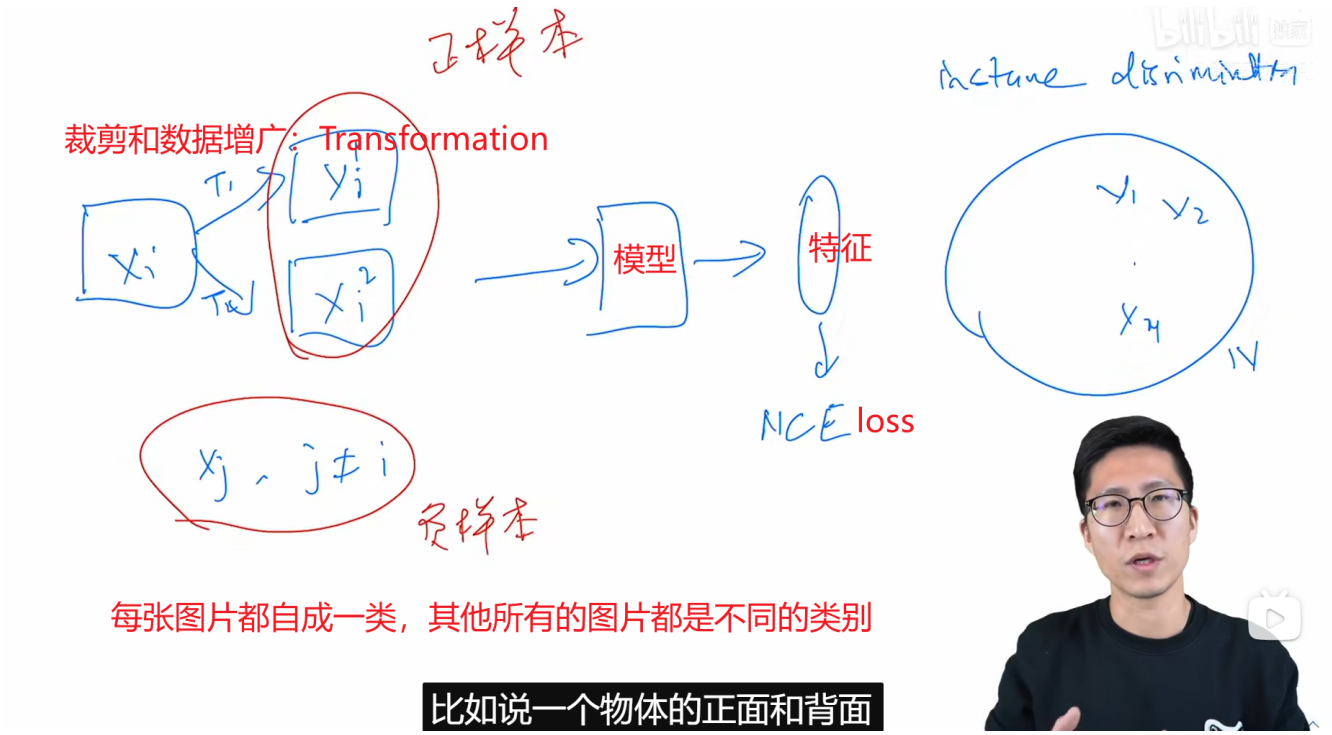



让f1和f2尽量靠近，让f3和f1、f2尽量远离

为什么这会被认为是无监督学习呢？不是还需要标签信息告知你这两张图片是相似的，那两张图片是不相似的？

代理任务（pretext task）登场：

举例：instance discrimination




对比学习的灵活性就在于，你只要制定一个正样本、负样本的规则，后面都是标准操作。

举例：

视频：正样本：同一个视频里的任意两帧，负样本：其他视频里的所有帧

SimCSE：正样本：同样的句子给模型做两次forward，每次forward使用不同的dropout， 负样本：其他所有句子的特征

CMC：正样本：一个物体的不同视角（正面和侧面，rgb图像和深度图像）

# Momentum

加权移动平均

$$y_t = m * y_{t-1} + (1-m) * x_t$$

m是动量的超参数

$y_{t-1}$是上一时刻的输出

$y_t$是这一时刻想要改变的输出

$x_t$是当前时刻的输入

目的就是不让当前时刻的输出完全依赖于当前时刻的输入

因此当m很大时，$y_t$就会改变得很缓慢；当m很小时，$y_t$就比较依赖于$x_t$

MOCO使用这种特性就可以缓慢地更新编码器，从而让中间字典中学习的特征尽可能地保持一致

# 摘要

把对比学习看成是一个字典查询的任务，动态字典，由一个队列和一个移动平均的编码器组成。

队列：因为队列里的样本不需要做梯度回传，所以就可以往队列里放很多负样本，让字典变得很大

移动平均的编码器：让字典里的特征尽量保持一致

结果：

分类任务：好

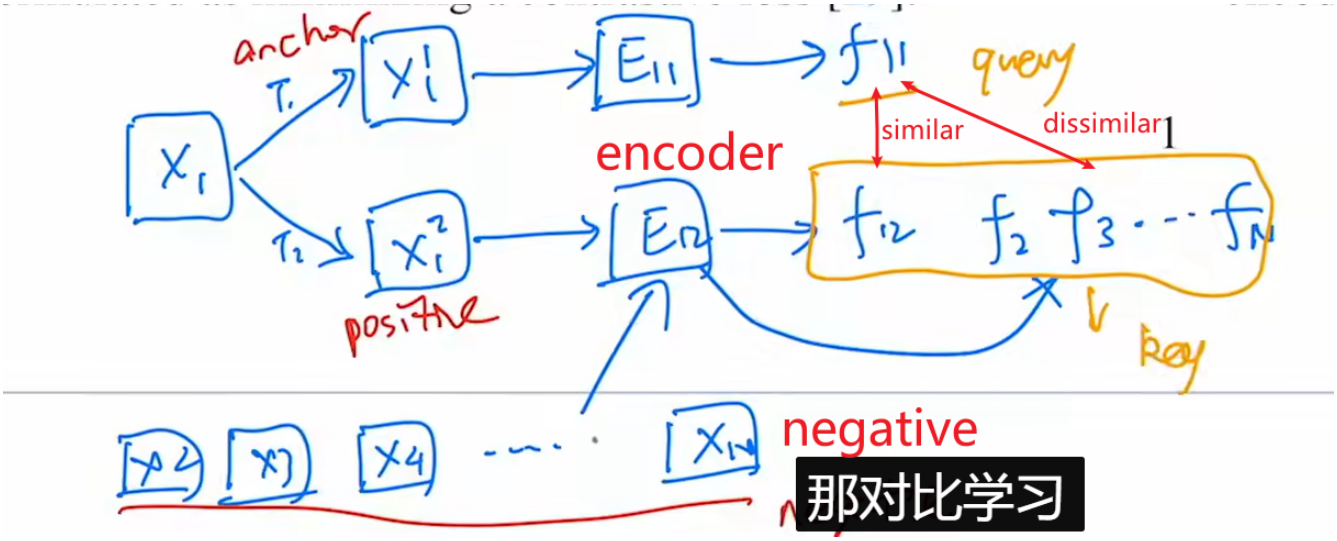
> **linear protocol**
>
> 先预训练一个骨干网络，把它用到不同的数据集上时，backbone freeze，只学习最后的那个全连接层（分类头），这样就相当于把那个提前学习好的预训练模型当做特征提取器，这样就可以证明之前预训练的模型学得好不好

MOCO学到的特征能够很好地迁移到下游任务上：

7个检测/分割任务上进行同样的无监督学习

# 引言

把对比学习当成动态字典，$E_{11}$和$E_{12}$是不同的encoder

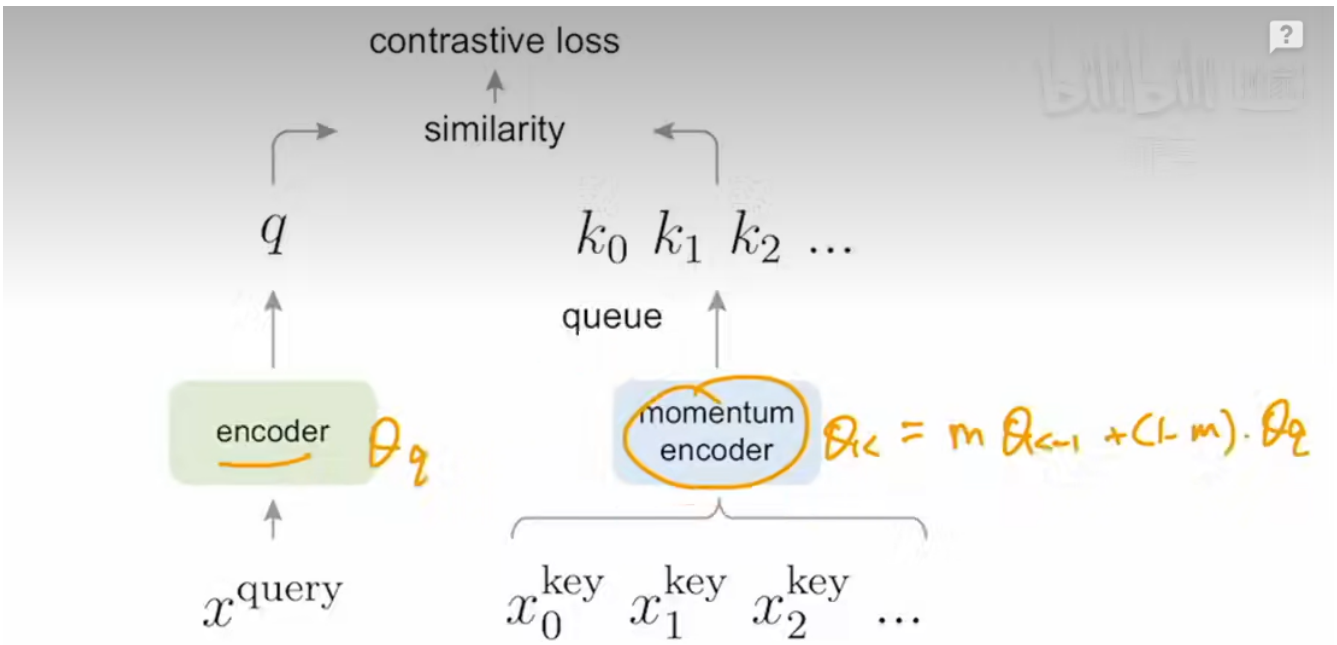



这个动态字典要具备以下两个特点：

1）大

2）在训练时要保持尽可能的一致性

因此MOCO给无监督的对比学习构造了一个又大又一致的字典：




1. **Query encoder**：用于编码当前mini-batch的图像，生成query向量。
2. **Key encoder**：用于编码队列中的样本，生成key向量。这个编码器的参数是通过动量更新机制逐渐变化的，而不是直接复制query encoder的参数，刚开始是由query encoder初始化而来。

用一个队列去表示字典的好处：如果这个字典很大，那么意味着输入的图片也很多，显卡的内存会吃不消。队列可以将字典的大小和模型每次forward时batch size的大小剥离开。

只有当前的mini-batch是由当前的encoder得到的，而之前的key都是不同时刻的encoder抽取的特征，那么这些key不就不一致了吗？ **（因为编码器的参数在训练过程中会不断更新）**

为了解决这个问题，作者使用了momentum encoder。

**代理任务**：本文就选择了一个比较简单的instance discrimination任务（个体判别任务）

正样本：同一张图片的不同视角

**结果**：预训练的特征在下游任务上有很好的迁移性

大家对无监督学习的另一个期待就是，希望它不要有性能保护，

因此，除了在ImageNet数据集上做预训练之外，还在1billion的Instagram数据集上做了预训练。

# 相关工作

> 自监督学习是无监督学习的一种，一般不做区分

**loss function**：

生成式网络：衡量prediction和fixed target直接的区别（如auto-encoder），使用L1 or L2 losses

判别式网络：例如eight positions，使用cross-entropy、margin-based losses

> 九宫格
>
> 能不能预测出随机挑选的格子位于5这个格子的哪个方位（8分类任务）

对比学习的目标函数：没有fixed target，目标在训练的过程中不断改变，是由一个encoder抽取出的特征（即动态字典）决定的。

对抗性的目标函数：衡量两个概率分布之间的差异（GAN）

**pretext tasks**:

脑洞大开的领域

**Contrastive learning *vs*. pretext tasks.**

配对

# 方法

## Contrastive Learning as Dictionary Look-up

InfoNCE loss

这个loss应该满足以下条件：

当q和唯一的正样本k+相似时，loss应该小

当q和其他负样本k都不相似时，loss应该小

反之，

当q和唯一的正样本k+不相似时，loss应该大

当q和其他负样本k都相似时，loss应该大

因为类别太多，没法算softmax，所以发明了NCE loss（noise constrastive estimation）

现在只看做两个类别：data sample和noisy sample

不在整个数据集上计算loss，只是在数据集里选一些负样本

infoNCE：NCE的变体，二分类可能不太合理

温度超参数$\tau$，控制分布形状

$$\mathcal{L}_q = -\log \frac{\exp(q \cdot k_+/\tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i/\tau)}$$

K：负样本的个数 本质上就是k+1分类任务

# Momentum Contrast

这个字典是动态的，key、key encoder在训练过程中都是变化的

**Dictionary as queue**：使用了queue之后，就把字典的大小和mini-batch的大小剥离开了

**Momentum update**：因为字典太大了，所以无法对队列里所有的参数做梯度回传了，也就是key encoder无法通过反向传播更新参数

一个简单的想法是，每轮训练结束后，直接把fq拿过来作为fk

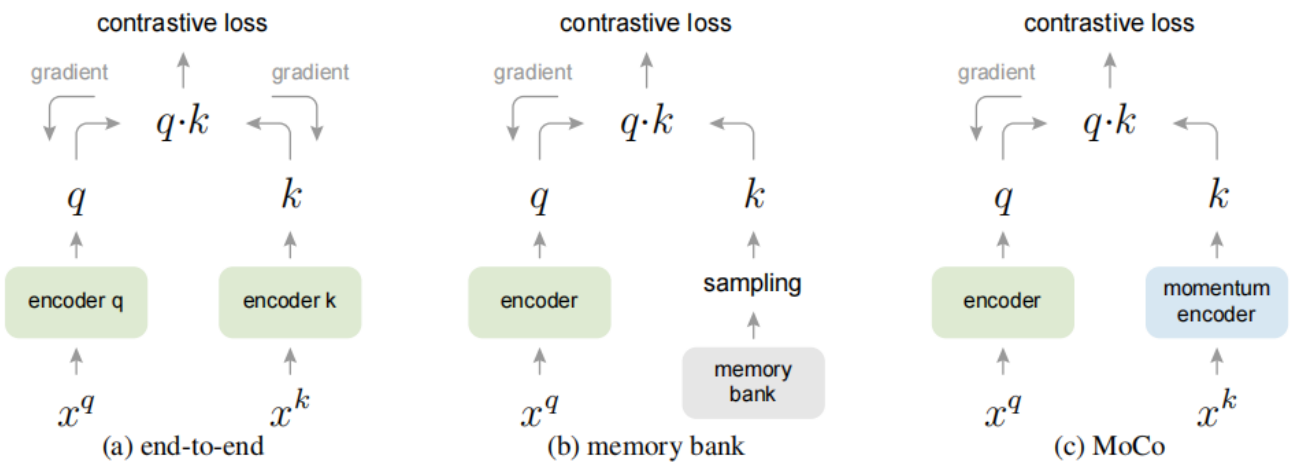
但是一个快速改变的key encoder降低了队列里所有key的一致性

因此提出了动量更新：

$$\theta_{\mathrm{k}} \leftarrow m\theta_{\mathrm{k}} + (1-m)\theta_{\mathrm{q}}.$$

本文中使用的m=0.999

**Relations to previous mechanisms.**:



a)正负样本都是从同一个mini-batch中采样出来的

字典的大小和mini-batch的大小是一样的

encoder q和encoder k的参数都可以更新

具备一致性，但不具备大

比如SimCLR就是用的这种端到端

b）只有encoder q是可以更新的，k这边没有编码器，用memory bank把整个数据集的特征都存起来，ImgeNet128w个key，只要600M

具备大，但是不具备一致性

# 结论

我们的方法在计算机视觉无监督学习任务上效果都很好。

但是，从ImageNet1M的数据到Ins1B数据上的改进比较小，可能是因为这个很大的数据集没有被完全应用。

展望未来：MAE作为对比学习中的代理任务