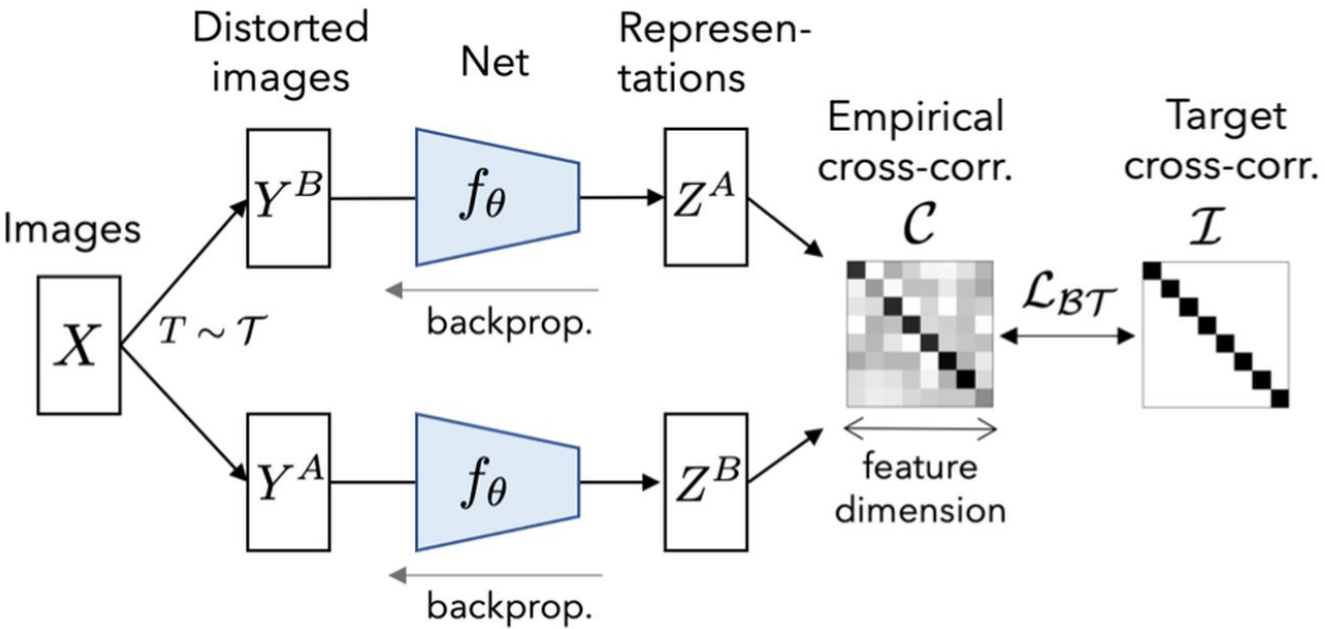


我们检测到你可能使用了 Adblock 或 Adblock Plus，它的部分策略可能会影响到正常功能的使用（如关注）。

×

你可以设定特殊规则或将知乎加入白名单，以便我们更好地提供服务。（为什么？）



最简单的self-supervised方法



王珣


路漫漫其修远兮 吾将上下而求索

关注他

郑华滨等 407 人赞同了该文章

从Kaiming的MoCo和Hinton组Chen Ting的SimCLR开始，自监督学习（SSL）成了计算机视觉的热潮显学。凡是大佬大组（Kaiming, VGG, MMLAB等），近两年都是搞了几个自监督方法的。从一开始的新奇兴奋地看着Arxiv上新发布的SSL方法（像MoCo, SwAV, BYOL, SimSiam等这些方法着实有趣），但是有一些相关的文章多少有些泛滥了，让人有些眼花缭乱。最近FAIR的一个工作，着实让我眼前一亮，觉得好有意思，颇为叹服。关键的是这个方法特别简单，应当可以称之为最简单的SSL。文章名字是：《Barlow Twins: Self-Supervised Learning via Redundancy Reduction》

Barlow Twins: Self-Supervised Learning via Redundancy Reduction

 arxiv.org

藉此机会，我也自己梳理一下SSL在这不到两年的时间里的个人认为比较重要的认知变化的节点：从SimCLR, MoCo为起点，以这篇BarLow Twins为暂时的终点。从这个历史线上去看SSL的发展非常有趣：计算机视觉圈子对于SSL的认知在不断打脸的过程中不断深入。

1. 首先是2020年初的**SimCLR**, 这个文章的核心贡献有二: 一是提供了使用google的丰富的计算资源和强大的工程能力, 使用高达4096的mini-batch size, 把SSL的效果推到了supervised方法差不多的效果(预训练模型做下游任务); 二是细致地整理了一些对SSL效果提升很有用的**tricks**: 如更长的训练, 多层MLP的projector以及更强的data augmentations。这些有用的trick在后来的SSL的论文中一直被沿用, 是SSL发展的基石, 而第一个点, 则是指出了大batch-size出奇迹, 为未来的论文指出了改进的路, 或者树立了一个进击的靶子。

2. **MoCo** 共有两版本, 原始版本是2019年末放出来的。在SimCLR出现后之后, 又吸收SimCLR的几个SSL小技巧, 改进出了V2版, 但是整体方法的核心是没有变化的, V2仅仅是一个2页试验报告。相比于SimCLR大力出奇迹, 恺明设计了一个巧妙的momentum encoder 和 dynamic queue 去获得大量的负样本。这里的momentum encoder 采用了动量更新机制, 除了文章本身的分析, 另一层的理解是: 其实momentum encoder相当于是teacher, 而dynamic里是来自不同mini-batch的样本, 所以teacher需要在时间维度上对于同一个样本的输出具有一致性, 否则, 要学习的encoder也就是student, 会没有一个稳定的学习目标, 难以收敛; 当然另一方面, teacher也不能一直不变, 如果teacher一直不变, student就是在向一个随机的teacher学习。综上, 动量更新机制是一个相当好理解的选择。

**阶段小结:** 抛开细节, SimCLR和MoCo的核心点, 都是认为**negatives (负样本)**非常重要, 一定要有足够多的负样本, 只不过实现方式略有不同。SimCLR 拿着TPU, 直接把batch size搞到4096, 一力降十会; 恺明则是巧妙设计Momentum机制, 避开了硬件工程的限制, 做出了可以飞入寻常百姓家的MoCo。再次重申, 这时候的认识, 还是停留在需要大量的负样本, 来提升SSL model的效果这个历史局限里。

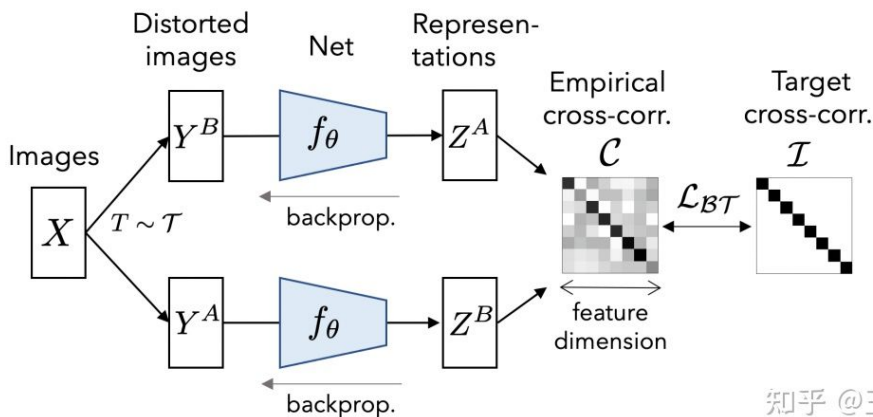
3. **BYOL** 是Deep Mind 在2020年发布的工作, 文章的核心点就是要破除“**负样本迷信**”, BYOL认为不使用负样本, 照样可以训练出效果拔群的SSL model。但是如果直接抛弃负样本, 只拉近正样本对的话, model 会容易陷入**平凡解**: 对于任意样本, 输出同样的embedding。为了在没有负样本的帮助下, 解决这个问题。BYOL 在Projector之上, 增加了一个新的模块, 取名Predictor。整体可以理解为在MoCo的基础上, 但是不再直接拉近正样本对(即同一个样本, 不同增强后的输出)的距离, 而是通过Predictor去学习online encoder 到 target encoder (即moco里的momentum encoder)的映射。另外, 对target network梯度不会传递, 即**Stop-Gradient**。(注: 在MoCo中, momentum encoder也是没有梯度回传的, 不过MoCo这么没有给momentum encoder回传梯度是因为queue里面的负样本来自过去的mini-batch, 其计算图已经丢失, 没有办法回传梯度, 而如果只回传正样本对的梯度, 会很不合理。而BYOL是只考虑正样本对, 如果梯度对于online encoder 和 target encoder都回传, 不存在这个不合理的点, 因此, Stop-Gradient是BYOL的一个特别的设计。)

4. **SimSiam** 是在BYOL的再次做减法, 这里在BYOL的基础上去除了momentum更新的target encoder, 直接让target encoder = online encoder。指出了predictor+stop-gradient 是训练出强大SSL encoder的一个充分条件。

再次的阶段小结: 在这个阶段, 训练模型到了可以没有负样本的阶段, 但是不使用负样本, 模型就gradient技巧;

模型陷入平凡解。BYOL 和 SimSiam 在方法上都是很不错的，试验也做得很可信充分，可是对于方法的解释并没有那么深刻置信，可能要寻求一个扎实的解释也确实很难。可以参见[从动力学角度看优化算法（六）：为什么SimSiam不退化？ - 科学空间|Scientific Spaces](#)，也是另一个角度的解释，颇为有趣合理。此时已经进入到了摆脱了负样本了，但是在不使用负样本的情况，要想成功训练好一个SSL model，需要引入新的trick: 即predictor+stop-gradient。这样子来看，难免有点像左手换右手的无用功，但是整体的技术认识是进步了很多的。

5. 最后，终于到了这次的主角：**Barlow Twins**。在不考虑数据增强这种大家都有的trick的基础上，Barlow Twins 既没有使用负样本，没有动量更新，也没有predictor和stop gradient的奇妙操作。Twins 所做的是换了一种视角去学习表示，从embedding本身出发，而不是从样本出发。优化目标是使得不同视角下的特征的相关矩阵接近恒等矩阵，即让不同的维度的特征尽量表示不同的信息，从而提升特征的代表能力。这种做法，和以前传统降维（如PCA）的方法是有共通之处的，甚至优化的目标可以说非常一致。



Barlow Twins 模型整体图

设 *Embedding* 模型为  $f$ , 其模型参数记为  $\theta$ .

对于  $X$  不同的视角  $a, b$  下的输入  $Y_a, Y_b$ , 分别输出的特征  $Z_a = f_\theta(Y_a), Z_b = f_\theta(Y_b)$ .

其中  $Z_a, Z_b \in \mathcal{R}^{N \times D}$ .

那么Twins 方法和以上的基于正负样本对的所有方法的区别，不严格（抛去特征normalize，BN等操作来说）的来说，可以用一句话，或者说两个式子来概括。

过去的方法大多基于InfoNCE loss 或者类似的对比损失函数，其目的是为了是样本相关阵接近恒等矩阵，即

$$Z_a * Z_b^T \rightarrow \mathcal{I}_N$$

而Twins的目的是为了让特征相关阵接近恒等，即：

$$Z_a^T * Z_b \rightarrow \mathcal{I}_D$$

对于对比损失类方法，比如SimCLR或MoCo需要很大的Batchsize或者用queue的方式去模拟很大的batchsize，而Twins需要极大的特征维度（8192）。这种特性和以上两个公式是完全对应且对称的。一个需要大  $N$ ，一个需要大  $D$ 。

知乎

首发于  
计算机视觉论文速递

$$\mathcal{L}_{BT} \triangleq \underbrace{\sum_i (1 - C_{ii})^2}_{\text{invariance term}} + \lambda \underbrace{\sum_i \sum_{j \neq i} C_{ij}^2}_{\text{redundancy reduction term}}$$

Barlow Twins 的核心在于提出了图中新的损失函数

另外，另一个有意思的点是Loss里面的超参数，在论文里，超参数  $\lambda$  是通过搜索得到的，然后发现在等于0.0002=1/5000是效果不错。其实，这里的loss略微改写，是可以省却这个不必要的超参数的。损失函数的第二项里的求和换成平均即可。首先，里面的求和换成平均等价于原来公式中  $\lambda = 1/8191$ ，虽然数字和搜出来并非完全相等，但是，这种超参数，从经验来说，在数量级上可以是完全一致了。可以合理的想像猜测在搜索这个超参数时，作者本人也是从数量级跨度去搜的。效果上， $\lambda = 1/8191$ 和 $\lambda = 1/5000$  应当不会有差。那么，一个有意思的问题，为什么是平均呢？我认为是平衡“正负样本”（对于Twins其实没有这个概念了，为了方便，类别来说，指的其实是对角线和非对角线）的梯度，*InfoNCE* 其实是通过softmax形式来隐式的获得了梯度之间的平衡，而这里是直接累加，对应的梯度回传也是直接累加，如果不用平均，或者说没有极小且合适的  $\lambda$ 。“负样本对”梯度将会占据主导，结果就是，我们的相关矩阵的非对角线大多已经接近0，loss第二项确实优化得很好，但是第一项没有长进。也就是说对角线元素距离“梦想中的1”会比较远。如果我以上的臆测分析是对的，那么就可以用平均去换掉loss内部求和，为保证公式的对称性，左一项也可以稍作等价改写，具体的Loss形式可以如下：

$$\mathcal{L}_{BT} = \sum_i \frac{1}{|\{j|j=i\}|} \sum_{j=i} (1 - C_{ij})^2 + \sum_i \frac{1}{|\{j|j \neq i\}|} \sum_{j \neq i} C_{ij}^2$$


这样子，省掉一个较为难调的超参数，公式上更加对称，会让Twins显得更简洁合理。

总结：从历史线上来看，从SimCLR和MoCo说一定要有大量的负样本，到BYOL和SimSiam通过神奇操作（stop-grad+predictor）验证了负样本并非不可或缺，最终到了Twins切换了一直以来从对比学习去训练SSL的视角，转向从特征本身出发，推开了另一扇大门。对比而言，相比于最简单的裸InfoNCE，Twins仅仅是换了一个loss function (+大维度的特征)。不过，大的维度相比于增加batchsize的代价要小得多，就是多占一点的显存。


编辑于 03-11

无监督学习

文章被以下专栏收录



计算机视觉论文速递  
欢迎关注微信公众号：CVer



CVer计算机视觉  
CVer：一个专注于分享计算机视觉的平台

关注专栏

关注专栏

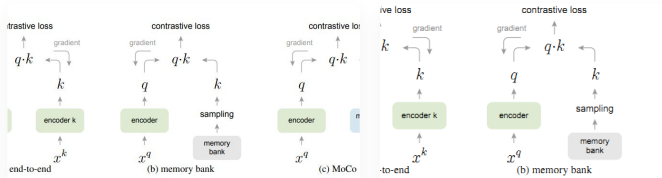


### BYOL与SimSiam

近年self-supervised learning最知名的两个方法是Google Brain的SimCLR与Facebook AI Research的MoCo，不过这两个方法都有各自缺陷：SimCLR需要够大的batch提供互斥的样本，而MoCo在内存中...

yanwa...

发表于AI约读社



### CV中的无监督学习方法：MoCo

小小将

发表于机器学习算...

### 学习笔记（三）对于moco的一些自我理解

小白

31 条评论

切换为时间排序

写下你的评论...

😊

#### 精选评论（3）

Shaohua Yang

03-09

很有意思的论文。  
让embedding的不同特征维度之间相关性为0，确实可以避免产生比较trivial的解(假定trivial解是让Za和Zb在normalization后近似相同的话)，但是反过来，假定不同特征相关度为0，相同特征相关度为1，会不会存在embedding其实很不一样(不知道文章中有没有相关证明)？假如lambda可以理解为平均值的话，第一项为什么不需要平均值？

👍 1

查看回复

疯狂的大液泡

03-09

还有一个点，Twins里边用了类似于BYOL或simsiam的stop-grad + predictor来打破对称结构，不仅没有效果提升，反而会对效果有比较大的损失，作者在论文中也没有就这个现象作进一步的解释，只解释了文中的方法能起到stop-grad+predictor一样的效果，来避免trivial solution

👍 1

查看回复

齐国君

03-09

看了下原文，感觉和InfoNCE基本是类似对偶（duality）的关系，相当于把InfoNCE里面batch样本index（b）换成这个方法里的dimension index（i）。换言之，这个方法里面embedding的不同dimension起到的作用和InfoNCE里面的负样本是类似的。从这个角度，也不难理解为什么用BYOL里面的非对称结构训练效果就不好了，因为非对称结构算出来的correlation C是没有什么物理意义的：基于不同embedding网络提取出来的特征，不同dimension之前自然无法去计算有意义的cross correlation。

不过更意思的一个地方是，他们最后用了8192维特征做embedding，按上面的duality关系，相当于InfoNCE里面用8192个负样本。我们在Adversarial Contrast ([adco: Adversarial Contrast for Efficient Learning of Unsupervised Representations from Self-Trained Negative Adversaries](#)) 里面也发现用8196个负样本就足够来训练InfoNCE loss了,看来真是条条大路通罗马。

👍 14

查看回复

评论（31）



👍 赞

王珣 (作者) 回复 郑华滨

03-08

哈哈，我就是慢慢写，这个论文确实很有意思，感觉是个重要的里程碑。

👍 赞

郑华滨

03-08

我在MoCo V2之后试过和Barlow Twin类似的思路，但是训崩了，看来不是这思路不对，是有什么魔鬼细节没实现好

👍 2

王珣 (作者) 回复 郑华滨

03-08

从文章来看，有两个非常重要的细节，一个是1/5000的那个权重系数，二是要是用极大的feature dimension（8K）。这两个对于效果都是影响巨大的，也是不太容易自己想到的。

👍 3

elvis 回复 郑华滨

03-09

自训练只要有一个小地方有diff 都很容易训不好

👍 2

展开其他 1 条回复

疯狂的大液泡

03-08

用同样的数据训过simclr、moco V1 V2、BYOL和simsiam，用backbone来做pretrained model，目前感觉moco的效果最稳定

👍 7

elvis 回复 疯狂的大液泡

03-09

byol最精巧 moco最皮实 改进空间也最多

👍 3

猪猪侠和狗子

03-08

Correction： MoCo应该是比SimCLR早的。MoCo是19年11月，SimCLR是20年2月。而且SimCLR也大方承认了他们对于MoCo的借鉴。

👍 4

王珣 (作者) 回复 猪猪侠和狗子

03-09

已更正内容，感谢！



👍 1

Shaohua Yang

03-09

很有意思的论文

知乎

首发于  
计算机视觉论文速递

同特征相关度为1，会不会存在embedding其实很不一样(不知道文章中有没有相关证明)? 假如lambda可以理解为平均值的话，第一项为什么不需要平均值?

1

 王珣 (作者) 回复 Shaohua Yang

03-09

第一项也可以是平均，只不过正好每一行就一个元素在对角线上，那么一个元素的平均和一个元素的求和是一回事[酷]

赞

 WonderSeven 回复 Shaohua Yang

03-09

意思是两个view学到的表征没必要类似吗?

赞

展开其他 3 条回复

 疯狂的大液泡

03-09

还有一个点，Twins里边用了类似于BYOL或simsiam的stop-grad + predictor来打破对称结构，不仅没有效果提升，反而会对效果有比较大的损失，作者在论文中也没有就这个现象作进一步的解释，只解释了文中的方法能起到stop-grad+predictor一样的效果，来避免trivial solution

1

 elvis 回复 疯狂的大液泡

03-09

这个不见得没有提升 而是很难调出来了 我实验过类似的，调了好几版才能出来

赞

 疯狂的大液泡 回复 elvis

03-10

调这个可能确实比较玄学[捂脸]

赞

 齐国君

03-09

看了下原文，感觉和InfoNCE基本是类似对偶（duality）的关系，相当于把InfoNCE里面batch样本index（b）换成这个方法里的dimension index（i）。换言之，这个方法里面embedding的不同dimension起到的作用和InfoNCE里面的负样本是类似的。从这个角度，也不难理解为什么用BYOL里面的非对称结构训练效果就不好了，因为非对称结构算出来的correlation C是没有物理意义的：基于不同 embedding网络提取出来的特征，不同dimension之前自然无法去计算有意义的cross correlation。

不过更意思的一个地方是，他们最后用了8192维特征做embedding，按上面的duality关系，相当于InfoNCE里面用8192个负样本。我们在Adversarial Contrast ([adco: Adversarial Contrast for Efficient Learning of Unsupervised Representations from Self-Trained Negative Adversaries](#)) 里面也发现用8196个负样本就足够来训练InfoNCE loss了,看来真是条条大路通罗马。

14

 当当当 回复 齐国君

03-10

样本和维度还是有不同的地方，比如会把同类的样本当做负样本处理了，维度相比来说更合理一些

赞



👍 1



礼拜天

03-10

localfeature 深度学习如l2-net 有类似的loss 去相关loss 目的是避免过拟合

👍 赞



王珣 (作者) 回复 礼拜天

03-10

单看损失其实并没有很新，但是很难想到这种会直接在 SSL里这么简单的work。其实最经典的PCA的目标函数和这个也挺像了。

👍 赞



心知

03-10

根据文章里的算法流程总结起来，通过不同data augmentation后获得初步representation  $z, z'$ ，然后将 $z, z'$ 映射到单位超球面( $l2\ norm = 1$ )，接着获得feature的similarity matrix (基于 $z\_norm\ T, z'\_norm$ 的内积)，最后减去eye matrix，求MSE loss。这个MSE大体上对应于InfoNCE loss 里的那个局部cross entropy。区别应该是在于此处是feature-level similarity, 而infoNCE中是instance-level similarity。

这让我联想到了clustering里，也会对feature-level similarity进行操作。不过，那里一般叫做clustering-level。此时，features的个数是类的个数，分别对应为类的logits。

👍 1



这大概就是人生吧

03-11

这个和Contrastive Clustering中的Cluster-level Contrastive Head不是一样的吗？

👍 赞



王珣 (作者) 回复 这大概就是人生吧

03-11

[pengxi.me/wp-content/up...](https://pengxi.me/wp-content/up...)

这也是一个很好的工作，在论文放出来的时候，就读了，挺不错的文章，但是说实话，并没有很惊叹。在twins中，它是单独存在，是不用contrastive learning 去做SSL，而且做得任务也是更加难调的，比如 $\lambda$ 超参数，以及超高的维度。

在Contrastive Clustering中是做一个辅助的分支，想法来说，是有点局限看了。要是能再大胆一点，努力调调，确实就是Twins。但是正是这大胆的一步，决定了文章的立意高低。

另外，这种特征间去冗余的loss真的很常见，尤其是在深度学习前的时代，这个公式基本上是教材前几章都会出现很多次的，公式一样，或者像真的没啥，能把方法做work，理清楚透彻才是真的厉害。

👍 1



这大概就是人生吧

03-11

非常感谢您的回答~我知道工程实践也是超级重要的。但是即使使用InfoNCE或者Donsker-Varadhan, NWJ, MINE的方法去优化互信息结果应该也不会很差。由于自监督学习现在都只能作为预训练，几个点的差别真的很有意义吗~

因为我一直致力于将其放在变化检测异常检测上，直接比较特征来检验效果，发现差别好小，



该评论已删除



这大概就是人生吧 回复 王珣 (作者)

03-11

谢谢

👍 赞

展开其他 1 条回复

