论文及工具理解

Contrastive Learning with Adversarial Example


```
**Chih-Hui Ho, Nuno Vasconcelos**, Department of Electrical and Computer
Engineering, University of California, San Diego, {chh279,
nvasconcelos}@ucsd.edu
```


1. 研究背景

1.1 自监督学习和对比学习

在日常生活中，很多物体的样貌难以被我们短时间记忆；即便如此，当再次遇见时，多数物体依然很容易被识别出来；在学习过某类事物的某些个体的特征后，就很容易对同类事物的其他个体加以辨识。自监督学习即出于这样的思想，不依赖于特定的标签值，而是自行挖掘并学习数据所蕴含的内在组织结构与特征，寻找样本之间不同层次的联系。

目前机器学习技术采用的主流学习范式大多是依然于人工标签的监督学习。然而数据本身蕴含的信息往往远比标签更丰富，有时即便大量的标签也难以训练出坚固的模型；此外，监督学习往往将模型拘泥于对特定事物的处理和分析，使其难以学到通用的知识，因此不具有较好的迁移性。自监督学习利用数据本身来指引学习过程，因此较好地解决了这些问题。

作为一种主流的自监督学习策略，对比学习通过使模型学习不同数据的相似和不同之处，从而学得数据的较为通用的高阶特征。对比学习亦出于这样的思想，即不追求图像的像素细节，而使更关注关键、抽象的语义信息，在特征空间中学得不同数据的区分性。

1.2 对比学习的问题

对比学习的一般目标就是学习生成一个编码器，使相似正样本通过编码器的映射值之间的相似度远远高于不相似负样本通过编码器的映射值之间的相似度。为了学习得到一个较好的深度表示，对比学习使用不带标签的训练数据来生成彼此成对的增广数据。如何设计正负样本对，是对比学习的核心问题之一。

正样本对的设计方法可包括单模态方法、多模态方法等。常用的数据增强技术可以用于单模态方法，例如常用的原样本扩增、单样本扩增、多样本扩增等图像扩增技术。针对不同类型的数据，采取恰当的扩增策略对于对比学习至关重要。例如，对于视频数据，正样本对往往来自于时间的一致性约束；对于多视图数据，扩增可以更加精细。

而与正样本对的设计不同，负样本对并未在已有的研究中受到较多的重视。对自监督学习的增强策略已有过很多研究，但多数对比学习算法难以对负样本对加以挖掘，也无法对图像实例加以关联。目前依然没有一个较为系统、全面的算法来做到这一点。

本文将对抗样本融入进了对比学习的框架中，提出了一种基于自监督学习的对抗样本训练算法`CLAE`。`CLAE`的思想来源于对抗样本的特点与在对比学习中的应用。相比于传统的对比学习方法，采用对抗样本可以产生更加困难复杂的正样本对；此外，在对抗样本训练的优化过程中，可以将整个批次的所有图像纳入考虑，从而产生更加困难复杂的负样本对。

2. 主要思想算法理解

对比学习通过使用锚点与正负样本表示等，实现相似数据与不同数据的区分。但由于在自监督学习中的数据不具备标签，因此需要采用数据增强的方法。与其他相似工具相比的创新之处：在一般的自监督学习中，基于对比学习的方法不一定会在硬负对上加以优化；此项研究则利用对抗性的例子，动态地生成更具有挑战性的正负对。

从总体上说，更困难的样本对产生的优化损失更大，因此可以使对比学习更有效。在标准的对比学习的流程框架的基础之上，此项研究将对抗样本作为一种数据增强和扩增的手段引入进来。在标准的对比损失函数的基础上，额外添加了一个新的正则项，即对抗对比损失；这种方法提高了对比学习流程的总体性能。给定经过数据增强后的样本，根据对抗对比可以计算出样本的梯度；利用`FGSM`的方法可以生成相应的对抗样本。最终，对比损失可以包含两项，即标准的对比损失以及对抗对比损失；可以制定参数来调节两者之间的权重。

为了寻找图像变换的更好方法，可以尝试将对抗损失最大化，即求出参数满足某一图像变换规律的前提下的对比学习损失函数的最大值；为了消除损失函数定义的模糊性，可以固定一个参数的变换。在此项研究中，再将该变换替换成了寻找给定变换的样本的对抗扰动。研究中的对抗增强方法的主要流程，是给定一种数据变换的规则，生成变换后的数据样本，通过计算和比较变换后的样本的对抗扰动，可以获得更加多样化的正样本对。通过这种方法产生的正样本可以增加学习的难度，从而在一定意义上提高学习算法所产生的特征不变性。同时，这种对抗增强的方法也进行了困难负样本的挖掘；对比损失中的对抗扰动，实际上是分类器的权重。通过对对比学习损失产生的对抗样本加以优化，产生了具备挑战性的正负样本对。

研究者分析说明，对抗性扰动的最佳扰动通常通过最大化交叉熵损失函数来获得。通过试图最小化由损失定义的风险，可能可以学习得到图像的不变表示。先前关于对比学习的研究，已经考虑了转换方式的很多种可能性。转换方式在自适应学习的表现中起关键作用，但先前工作只是在转换集上简单随机地采样。在此项研究中，研究者则试图为每个图像寻找一个最佳变换，来最大化由损失定义的风险。总的来说，此问题定义模糊，但可以修复其模糊性。在研究中，用对抗扰动代替了其中的一个参数，使得在自监督学习的情况下增加了挑战学习，从而使得算法更容易产生更不变的表示。

首先，对比损失可以写成交叉熵损失。通过此方法，易于获得扰动参数的最优集合。这需要确定增强的最佳对抗性扰动，最好将类似的正对嵌入在一起，分离所有不同的样本；即优化寻找具有挑战性的正负对，同时进行硬负对的挖掘。为了进行对抗性训练，研究者采用了`AdvProp`，它使用两个单独的批处理规范化层进行了洁净的和敌对的示例。与洁净示例相关联的动量被固定为原始对比学习算法所采用的值，而与敌对示例相关联的动量则根据经验而被增大。通过将对比损失的梯度反向传播到网络输入，可以创建对抗示例。计算标准扩增对和由扩增和对抗实例组成的对的对比损失项，可以训练网络。

研究者对对抗性对比学习进行了一定的实验评估。通过研究对抗性例子对自监督学习和下游任务的影响，可以发现，对于给定的扰动强度，对抗性扰动可以比随机扰动引起更大的损失。因此，对抗性增强可以比随机扰动产生更具有挑战性的对。通过研究将`CLAE`框架应用于自监督学习的好处，以及这些好处在对比学习方法中的一致性，可以发现，对抗性训练提高了对比学习算法在所有数据集上的性能。

通过实验评估表明，几个对比学习基线的性能可以得到提高。此项研究推进了深度学习技术的广泛使用，特别是在数据集标签难以获得的情况下的模型训练。尽管在先前的自监督学习技术中，更大的神经网络、更大的批量和更长的训练时期是一向被推崇的；但本研究的实验表明，如能通过在有效的训练对上进行优化，那诸如此类的因素就会变得不再重要。