Representation Learning with Contrastive Predictive Coding



1. 摘要

虽然监督学习在许多应用中取得了巨大的进展，但无监督学习却没有得到如此广泛的采用，它仍然是人工智能的一项重要而具有挑战性的工作。在这项工作中，我们提出了一种通用的无监督学习方法，以从高维数据中提取有用的表征，我们称之为对比预测编码（Contrastive Predictive Coding，CPC）。我们模型的关键见解是通过使用强大的自回归模型预测潜在空间中的未来来学习这种表征。我们使用概率对比损失，诱导潜伏空间捕捉对预测未来样本有最大作用的信息。它还通过使用负抽样使模型具有可操作性。虽然大多数先前的工作都集中在评估某一特定模式的表征上，但我们证明了我们的方法能够学习有用的表征，在四个不同的领域取得强大的性能：语音、图像、文本和三维环境中的强化学习。

1. 问题重述

以端到端的方式从标记的数据中学习高级表征，是迄今为止人工智能领域最大的成功之一。这些技术使人工指定的特征在很大程度上变得多余，并在一些现实世界的应用中大大提高了最先进的水平[1, 2, 3]。然而，许多挑战仍然存在，如数据效率、鲁棒性或概括性。

改善表征学习需要那些对解决单一监督任务不太专业的特征。例如，当预先训练一个模型进行图像分类时，诱导的特征可以很好地转移到其他图像分类领域，但也缺乏某些信息，如颜色或计数能力，这些信息与分类无关，但与图像说明等有关[4]。同样地，对转录人类语音有用的特征可能不太适合用于说话人的识别或音乐流派的预测。因此，无监督学习是实现稳健和通用表征学习的重要垫脚石。

尽管它很重要，但无监督学习还没有看到类似于有监督学习的突破：从原始观察中建立高层次的表征模型仍然难以实现。此外，理想的表征是什么并不总是很清楚，以及是否有可能在没有额外监督或专门针对特定数据模式的情况下学习这样的表征。

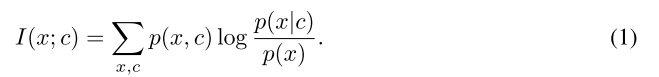
无监督学习最常见的策略之一是预测未来的、缺失的或有背景的信息。这种预测性编码的想法[5, 6]是信号处理中最古老的数据压缩技术之一。在神经科学中，预测性编码理论表明，大脑在不同的抽象层次上预测观察结果[7, 8]。最近在无监督学习方面的工作已经成功地利用这些想法，通过预测相邻的单词来学习单词表征[9]。对于图像来说，从灰度或图像斑块的相对位置来预测颜色也已经被证实。我们假设这些方法是富有成效的，部分原因是我们预测相关价值的背景往往有条件地依赖于相同的共享的高层次潜在信息。而通过将其作为一个预测问题，我们可以自动推断出这些对表征学习感兴趣的特征。

本文提出以下建议：首先，我们将高维数据压缩到一个更紧凑的潜伏嵌入空间，在这个空间中，条件预测更容易建模。其次，我们在这个潜伏空间中使用强大的自回归模型来进行未来许多步骤的预测。最后，我们依靠Noise-Contrastive Estimation[12]的损失函数，其方式类似于自然语言模型中学习词嵌入的方式，允许整个模型进行端到端的训练。我们将由此产生的模型--对比预测编码（CPC）应用于广泛不同的数据模式，包括图像、语音、自然语言和强化学习，并表明同一机制在这些领域中的每一个领域都能学习到有趣的高层次信息，其表现优于其他方法。

1. 详细阐述
   1. Motivation and Intuitions

我们的模型背后的主要直觉是学习编码（高维）信号的不同部分之间的基本共享信息的表征。同时，它摒弃了低层次的信息和更局部的噪音。在时间序列和高维建模中，使用下一步预测的方法利用了信号的局部平稳性。当进一步预测未来时，共享信息量变得更低，模型需要推断出更多的全局结构。这些跨越许多时间步骤的'慢速特征'[13]往往更有趣（例如，语音中的音素和语调，图像中的物体，或书籍中的故事线。）

预测高维数据的挑战之一是，诸如均方误差和交叉熵的单模损失不是很有用，通常需要强大的条件生成模型，需要重构数据中的每一个细节。但是这些模型的计算量很大，而且浪费了对数据x中复杂关系建模的能力，往往忽略了上下文c。例如，图像可能包含数千比特的信息，而高层次的潜在变量如类别标签包含的信息要少得多（10比特用于1024个类别）。这表明，直接对p(x|c)进行建模对于提取x和c之间的共享信息来说可能不是最佳选择。当预测未来信息时，我们将目标x（未来）和背景c（现在）编码为一个紧凑的分布式向量表征（通过非线性学习映射），其方式是最大限度地保留原始信号x和c的相互信息，定义如下：



通过最大化编码表征之间的相互信息（其边界是输入信号之间的MI），我们提取了输入的潜在变量的共同点。

* 1. Contrastive Predictive Coding

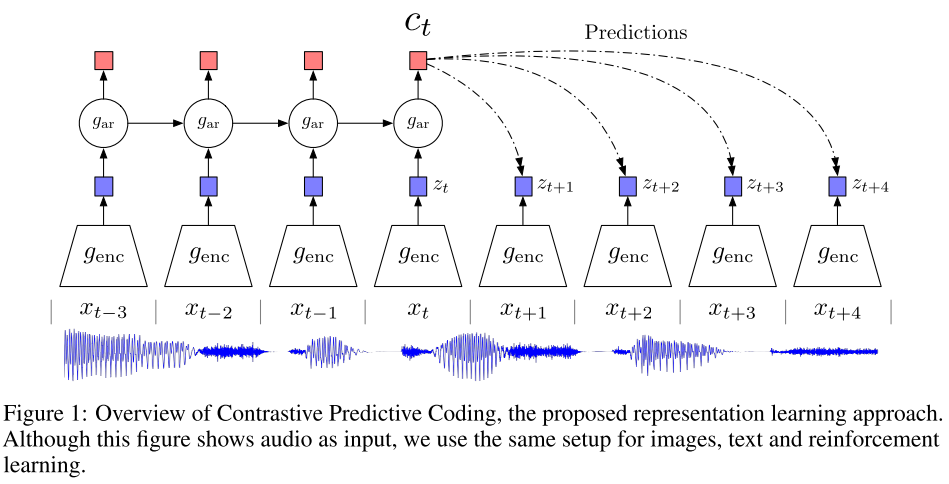


图1显示了对比性预测编码模型的结构。首先，一个非线性编码器genc将输入的观察值xt序列映射到潜像空间的潜像序列zt = genc(xt)，并产生一个上下文潜像ct = gar(z≤t)，可能具有较低的时间分辨率。接下来，一个自回归模型gar总结了所有的z≤t，正如上一节所论述的，我们不直接用生成模型pk(xt+k|ct)预测未来的观测值xt+k。相反，我们建立了一个密度比的模型，它保留了xt+k和ct之间的相互信息（公式1），具体如下（进一步的细节见下一小节）。



请注意，密度比f可以是非正态化的（不一定要积分到1）。尽管这里可以使用任何正的实数分数，但我们使用一个简单的对数双线性模型：



在我们的实验中，线性变换WT k ct被用于预测，每一步k都有不同的Wk。

通过使用密度比f(xt+k, ct)和用编码器推断zt+k，我们将模型从高维分布xtk的建模中解脱出来。

尽管我们不能直接评估p(x)或p(x|c)，我们可以使用这些分布的样本，允许我们使用诸如噪声对比估计[12, 14, 15]和重要性抽样[16]等技术，这些技术是基于将目标值与随机抽样的负值进行比较。

在提议的模型中，zt和ct中的任何一个都可以作为下游任务的代表。如果来自过去的额外背景是有用的，自回归模型的输出ct可以被使用。其中一个例子是语音识别，zt的接受域可能不包含足够的信息来捕捉语音内容。在其他情况下，如果不需要额外的背景，zt可能反而更好。如果下游任务需要整个序列的一个表征，例如在图像分类中，我们可以把来自zt或ct的表征集中到所有的位置上。

最后，请注意，任何类型的编码器和自回归模型都可以用在拟议的框架中。为了简单起见，我们选择了标准的架构，如带有resnet块的串联卷积层作为编码器，GRU[17]作为自回归模型。最近在自回归模型方面取得的更多进展，如屏蔽卷积架构[18, 19]或自我注意网络[20]，可以帮助进一步改善结果。

1. 实验
2. 评析