# 对比预测编码的表示学习

摘要：在监督学习使很大的进步在许多应用程序中,无监督学习没有见过这样的广泛采用,并仍然是人工智能的一个重要和富有挑战性的努力。在这项工作中，我们提出了一种通用的无监督学习方法来从高维数据中提取有用的表示，我们称之为对比预测编码。看见我们的模型的关键是学习这种表征预测未来的潜在空间利用强大的自回归模型。我们使用一种概率对比损失，它诱导潜在空间来捕获对预测未来样本最有用的信息。采用负抽样的方法，使模型易于处理。虽然大多数之前的工作都集中在评估一个特定模态的表征，我们证明我们的方法能够学习有用的表征，在四个不同的领域实现强大的性能:语音，图像，文本和增强学习在3D环境。

1. 介绍

学习高层表示标记与分层可微的模型数据的端到端潮流是一个迄今为止最大的成功的人工智能技术。这些技术使得手工指定的特性在很大程度上是多余的，并且极大地改善了一些现实世界的应用程序。然而，仍然存在许多挑战，如数据效率、健壮性或泛化。

改进表示学习需要的特征不是专门用于解决单一的监督任务。例如，在对一个模型进行图像分类的预训练时，诱导的特征可以很好地转移到图像的其他分类域，但也缺少一些与分类无关但与图像字幕相关的信息，如颜色或计算能力。类似地，对人类语言转录有用的特征可能不太适合于扬声器识别或音乐类型预测。因此，无监督学习是实现鲁棒和泛型表示学习的重要一步。

尽管无监督学习很重要，但与有监督学习类似，无监督学习尚未取得突破:从原始观察中建模高级表示仍然是难以实现的。此外，我们并不总是清楚理想的表示是什么，也不知道是否可以在没有额外监督或对特定数据模态专门化的情况下学习这样的表示。

无监督学习最常见的策略之一是预测未来信息、缺失信息或上下文信息。这种预测编码的想法是数据压缩信号处理中最古老的技术之一。在神经科学中，预测编码理论认为大脑在不同的抽象层次上预测观察结果。最近的无监督学习的工作已经通过预测邻近的单词，成功地使用这些思想来学习单词表征。对于图像，从灰度或图像斑块的相对位置预测颜色也被证明是有用的。我们假设这些方法是卓有成效的，部分原因是我们预测相关价值的环境通常有条件地依赖于相同的共享的高级潜在信息。通过把这个问题作为一个预测问题，我们自动推断出这些特征，这些特征对表征学习来说是有意义的。

在本文中我们提出以下建议:首先，我们将高维数据压缩到一个更紧凑的潜在嵌入空间中，在这个空间中条件预测更容易建模。其次，我们在这个潜在空间中使用强大的自回归模型对未来的许多步骤进行预测。最后，我们使用与自然语言模型中学习单词嵌入类似的方法来对损失函数进行噪声对比估计，从而使整个模型能够端到端进行训练。我们将所得到的模型，对比预测编码(CPC)应用于不同的数据模式、图像、语音、自然语言和强化学习，并表明相同的机制在这些领域学习有趣的高级信息，优于其他方法。

1. 对比预测编码

我们从激励和给予我们方法背后的直觉开始这个部分。接下来，我们介绍了对比预测编码(CPC)的结构。然后解释了基于噪声对比估计的损耗函数。最后，讨论了对比预测编码的相关工作。

2.1动机和思路

我们的模型背后的主要思路是学习编码(高维)信号不同部分之间的基础共享信息的表示。同时，它抛弃了低层次的信息和局部的噪声。在时间序列和高维建模中，采用下一步预测的方法，利用信号的局部平滑性。在进一步预测未来时，共享信息的数量会大大减少，模型需要推断出更多的全局结构。预测未来输出是不可能的，你怎么知道下一步的输出是什么。那么，要做的是就是预测的输出与实际的下一步输出尽可能相似，要做到这一点，就需要Prediction中包含尽可能invariance和global的structure。作者将这些部分定义为“slow features”，并用语音任务中的音素和语调，图像任务中的物体或书籍中的故事情节来举例。

高维数据预测的挑战之一是单峰的损失意味着平方误差和交叉熵等不是很有用,而强大的条件生成模型需要重建每个细节的数据通常是必需的。但这些模型计算激烈,浪费能力在复杂关系数据x（data x）的建模上,往往忽略了context c。例如,图像可能包含成千上万比特的信息和高层潜变量,如类标签包含更少的信息(10比特为1024类别)。这表明建模p (x | c)直接为目的的可能不是最优的提取x和c之间的共享信息。在预测未来信息时，我们将目标x(未来)和上下文c(现在)编码为一个紧凑的分布式向量表示(通过非线性学习映射)，以最大限度地保留原始信号x和c的相互信息，定义为： *（1）*

通过最大化编码表示之间的相互信息(输入信号之间的MI限制)，我们提取了潜在变量的共同输入。

2.2对比预测编码

图1显示了对比预测编码模型的架构。首先，一个非线性编码器映射输入序列的观察到序列的潜在表示，潜在地具有较低的时间分辨率。然后，通过自回归模型汇总潜在空间中所有z≤t的信息，得到上下文潜在表示≤。

正如在前一节中争论的那样，我们不能直接用生成模型来预测未来的观察结果。相反，我们建立了一个密度比模型，该密度比保持了与之间的相互信息(公式1)，如下所示(详见下一小节):

屏幕快照 2020-09-22 下午8.25.31（2）

其中，∝代表“与成比例”(即，直到一个乘法常数)。注意，密度比f可以不归一化(不必积分到1)。虽然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:

屏幕快照 2020-09-22 下午7.16.44（3）

在我们的实验中，我们使用线性变换对每一步的值进行预测，也可以使用非线性网络或递归神经网络。

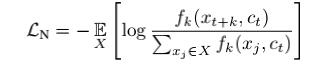
通过使用密度比f(, )和用编码器推断，我们减轻了模型建模高维分布。虽然我们不能直接评估p(x)或p(x|c)，但我们可以使用来自这些分布的样本，这使我们能够使用诸如噪声对比估计和重要性抽样等技术，这些技术是基于比较目标值和随机抽样的负值。

在该模型中，和都可以作为下游任务的表示。如果来自过去的额外背景是有用的，自回归模型输出可以被使用。其中一个例子是语音识别，的接受域可能没有包含足够的信息来捕获语音内容。在不需要附加上下文的其他情况下，可能更好。如果下游任务需要对整个序列使用一种表示，如图像分类，则可以将所有位置的或表示集合起来。

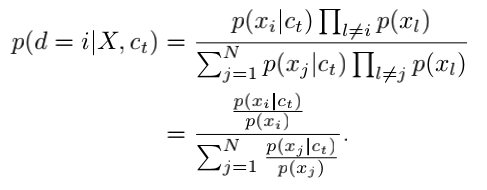
最后，请注意，任何类型的编码器和自回归模型都可以在建议的框架中使用。为了简单起见，我们选择了标准架构，比如带有清零模块的条纹卷积层用于编码器，GRUs用于自回归模型。自回归建模的最新进展，如掩蔽卷积架构或自注意网络，可以帮助进一步改善结果。

2.3InfoNCE损失和互信息估计

编码器和自回归模型都经过训练来共同优化基于NCE的损失，我们称之为InfoNCE。给定一个包含N个随机样本的集合X = {x1，…xN}，其中通过挑选从p(|)中得到的正样本，以及（N-1）个来自分布p()的负样本，我们优化:

（4）

可以让p(|)模拟Eq.2中的密度比。同时上式是一个分类交叉熵，来正确分类正样本。当d=i，即是正样本的时候，最优化概率是。而该样本来自于条件分布p(|)而不是p()的概率是：

（5）

可以看出，公式4中f(|)的最优值与p(|)/p()成正比，且与负样本数N-1的选择无关。

虽然不需要训练，但我们可以对变量和之间的互信息进行如下评估:

屏幕快照 2020-09-22 下午8.00.17

即得到交互信息的下界.

2.4相关工作

CPC是一种将预测未来观测(预测编码)与概率对比损失(公式4)相结合的新方法。这使我们能够提取慢速特征，从而最大化长期观测的相互信息。对比损失和预测编码以前分别以不同的方式被使用，我们现在将讨论。

对比损失函数在过去曾被许多学者所使用。例如，所提出的技术是基于使用最大边际方法来分离正例子和负例子的三重损失。更近期的工作包括时间对比网络，该网络建议最小化来自同一场景的多个视点的嵌入之间的距离，同时最大化来自不同时间步长的嵌入之间的距离。在时间对比学习中，利用对比损失预测多变量时间序列的分段ID，作为提取特征和进行非线性独立分析的一种方法。

在从相关的观察中定义预测任务以提取有用的表示法方面也有一些工作和进展，其中许多已经应用到语言中。在Word2Vec中，相邻单词的预测使用对比损失。跳跃式思维向量和字节mLSTM是替代方案，它们超越了使用递归神经网络进行单词预测，并在观察序列上使用最大似然度。在计算机视觉中，对被跟踪的视频补丁使用三联丢失，使来自同一目标的不同时间步长的补丁比随机的补丁更相似。提出预测图像中斑块的相对位置，并从灰度图像中预测颜色值。

1. 实验

我们提供了四个不同应用领域的基准:语音、图像、自然语言和强化学习。对于每个领域，我们训练CPC模型并通过线性分类任务或定性评估来探测表示包含什么，在强化学习中，我们测量辅助CPC损失如何加速agent的学习。

3.1语音

对于音频，我们使用公共可用的LibriSpeech数据集的100小时子集。尽管数据集不提供原始文本以外的标签，但我们使用Kaldi工具包和Librispeech上的预训练模型获得了强制对齐的电话序列。我们已经使对齐电话标签和我们的火车/测试分裂可下载在谷歌驱动器。数据集包含251个不同的发言者的语音。

在我们的实验中使用的编码器结构包括一个条纹卷积神经网络，它直接运行在16KHz PCM音频波形上。我们使用5个卷积层，带有步长[5,4,2,2,2]，过滤器大小[10,8,4,4,4,4]，以及512个带有ReLU激活的隐藏单元。该网络的总向下采样因子为160，因此每10ms的语音就有一个特征向量，这也是Kaldi得到的音素序列标签的比率。然后对模型的自回归部分使用GRU RNN，即256维隐藏状态的。GRU在每个时间步长上的输出被用作上下文c，从中我们利用对比损失预测未来12个时间步长。我们对长度为20480的采样音频窗口进行训练。我们使用学习率为2e-4的Adam优化器，并使用8个gpu，每个gpu有8个小批量示例，从中提取对比损失中的负样本。这个模型被训练到收敛，大约需要30万次更新。

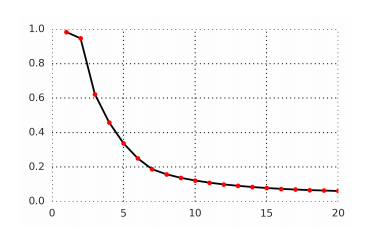
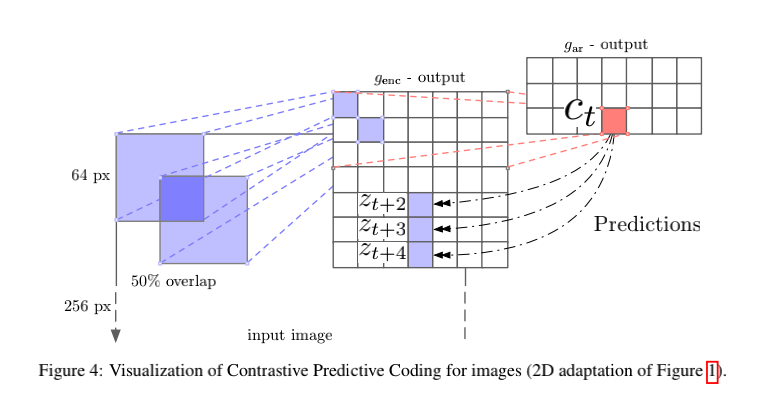
(图3)

图3:在语音波形未来1到20个潜在步长的对比损失中预测正样本的平均准确率。该模型预测，未来将达到200ms，因为每一步都包含10ms的音频。

图3显示了模型预测未来延迟的准确性，从1到20个时间步长。我们报告了在概率对比损失中，正样本的对数的平均次数比负样本的对数要高。这个图还表明，目标不是微不足道的，也不是不可能实现的，正如预期的那样，随着目标越来越远，预测任务会变得越来越困难。



（图4:图像对比预测编码可视化(图1的二维自适应)）

为了理解CPC提取的表征，我们使用基于这些特征训练的线性分类器来测量手机预测性能，从而显示出在这些特征下相关类是如何线性可分的。在模型收敛后，提取整个数据集的256维GRU(即ct)输出，并训练一个多类线性logistic回归分类器。结果如表1 (top)所示。我们用三个基线来比较准确性:来自一个随机初始化模型的表示(例如，和未经过训练)，MFCC特征，以及一个使用标记数据进行端到端监督训练的模型。这两个模型与用于提取CPC表示的模型具有相同的架构。完全监督的模型可以作为这个体系结构可实现的目标的指示。我们还发现并非所有编码的信息都是线性可访问的。当我们使用一个隐藏层代替时，准确率从64.6提高到72.5，这更接近于完全监督模型的准确率。

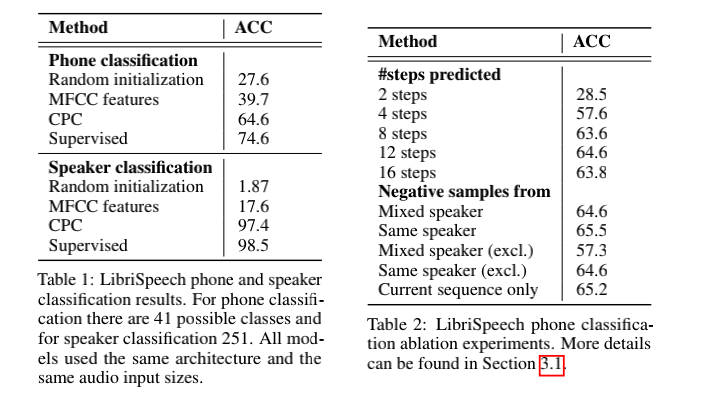
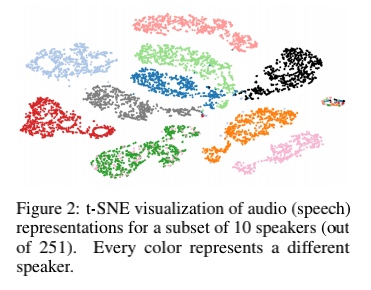


表2给出了两项用于电话分类的CPC消融研究的概述。在第一组中，我们改变模型预测的步骤数，表明预测多个步骤对于学习有用的特征是重要的。在第二组中，我们比较绘制阴性样本的不同策略，所有预测12个步骤(这在第一次消融中给出了最好的结果)。在混合说话者实验中，阴性样本包含了不同说话者的例子(第一行)，与之形成对比的是同一个说话者实验(第二行)。在第三和第四次实验中，我们排除了当前序列来提取负样本(因此X中只存在其他的小批量样本)，在最后一次实验中，我们只在序列中提取负样本(因此所有的样本都来自同一个speaker)。

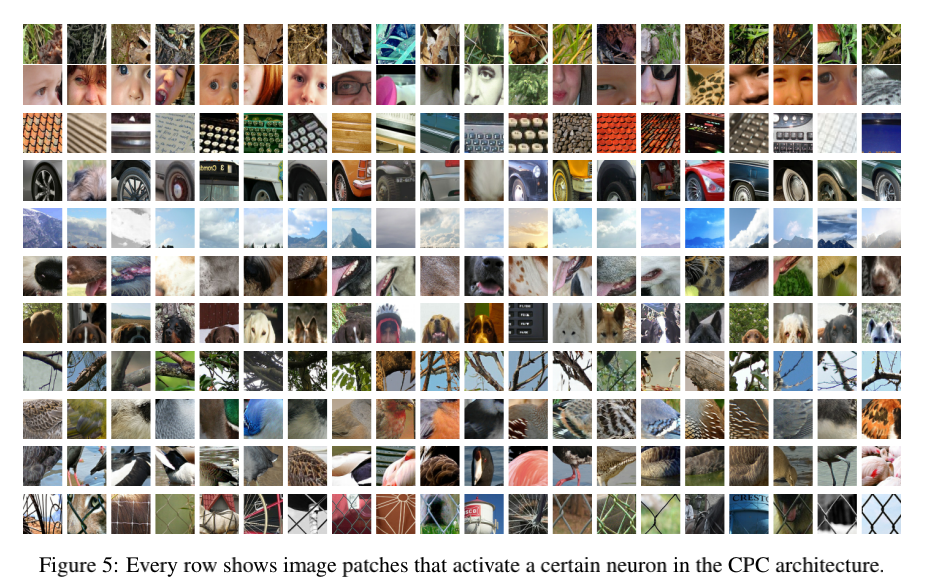
除了电话分类之外，表1(下)显示了使用来自同一表示的线性分类器(我们不按时间平均话语)进行说话人识别的准确性(251个中的251个)。有趣的是，CPCs既能捕获说话人身份，也能捕获讲话内容，这一点可以通过简单线性分类器获得的良好精度得到证明，它也接近oracle，完全监督的网络。

另外，图2显示了t-SNE可视化，显示了嵌入对于扬声器语音特性的区分性。需要注意的是，窗口大小(GRU的最大上下文大小)对性能有很大的影响，并且较长的段将获得更好的结果。我们的模型最多有20480个时间步要处理，略长于一秒。



3.2图像

在我们的视觉表示实验中，我们使用ILSVRC ImageNet竞争数据集。ImageNet数据集已经被许多作者用来评估无监督视觉模型。我们采用与[36]相同的设置，并使用ResNet v2 101架构作为图像编码器来提取CPC表示(注意，该编码器没有经过预先训练)。我们没有使用批量标准。经过无监督训练后，训练一个线性层来测量ImageNet标签的分类精度。



训练过程如下:从一幅256x256的图像中提取一个7x7的网格，包含64x64个作物，重叠点为32个像素。简单的数据增强对256x256的图像和64x64的作物都很有帮助。从300x300的图像中随机裁剪出256x256的图像，以50%的概率水平翻转并转换为灰度。对于每一个64x64的作物，我们随机取一个60x60的子作物，并将它们填充回64x64的图像。

然后由ResNet-v2-101编码器对每个作物进行编码。我们使用第三个残差块和空间均值池的输出来获得每个64x64 patch的单个1024-d向量。这样就得到了一个7x7x1024张量。接下来,我们使用一个PixelCNN-style自回归模型[19](卷积行格勒乌PixelRNN[39]给出了类似的结果)进行预测的潜在激活行全面后,显示在图4。我们从7x7的网格中预测最多5行，并对行中的每个补丁应用对比损失。我们使用学习率为2e-4的Adam optimizer，并在32个gpu上进行训练，每个gpu的批大小为16。

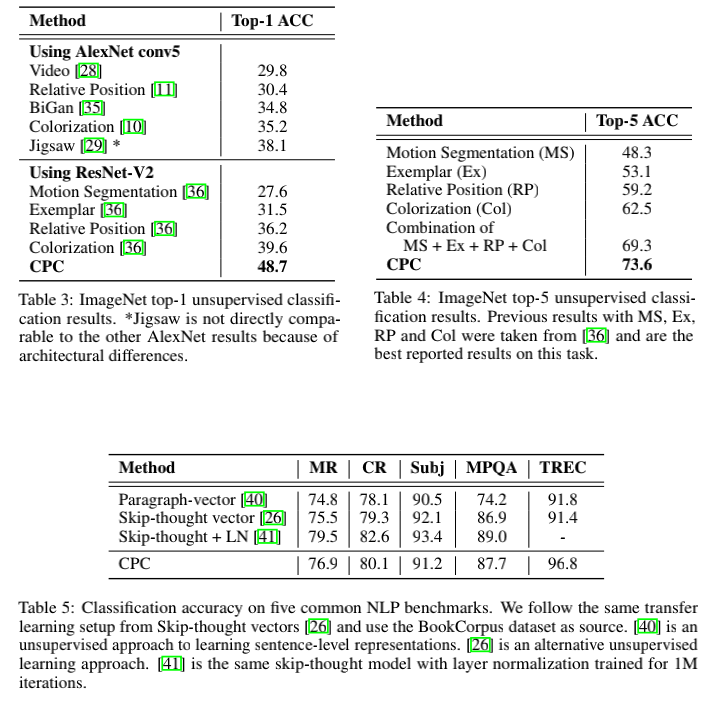
对于在CPC特征上训练的线性分类器，我们使用了动量为0.9的SGD，对于50k、25k和10k更新的学习率调度为0.1、0.01和0.001，单个GPU上的批处理规模为2048。注意，在训练线性分类器时，我们首先在空间上将7x7x1024表示均值池为单个1024维向量。这与[36]略有不同，后者使用没有池的3x3x1024表示，因此在受监督的线性映射中有更多的参数(这可能是有利的)。

表3和表4显示了top-1和top-5的分类准确率与最新的。尽管相对而言领域不确定，CPCs在最高1的精度上提高了9%，在最高5的精度上提高了4%。

3.3自然语言

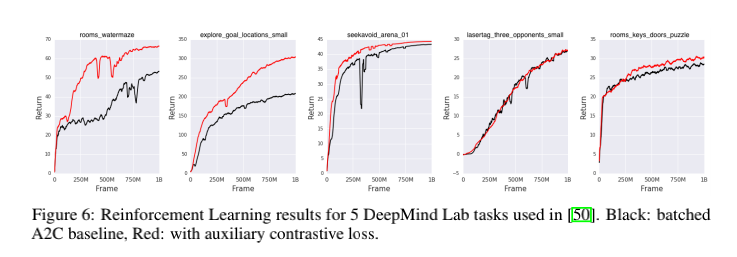
我们的自然语言实验遵循了用于跳跃思维向量模型的[26]的过程。我们首先在图书库数据集[42]上学习我们的无监督模型，并通过对一组分类任务使用CPC表示来评估我们的模型作为通用特征提取器的能力。为了处理训练中没有看到的单词，我们使用了与[26]相同的词汇扩展方法，其中word2vec和通过模型学习到的嵌入单词之间构建了一个线性映射。

对于分类任务，我们使用了以下数据集:电影评论情绪(MR)[43]、客户产品评论(CR)[44]、主观性/客观性[45]、意见极性(MPQA)[46]和问题型分类(TREC)[47]。与[26]一样，我们对MR、CR、Subj、MPQA使用10倍交叉验证的logistic回归分类器进行训练，并对TREC使用训练/测试分割。L2正则化权重通过交叉验证选择(因此前4个数据集的嵌套交叉验证)。



我们的模型由一个简单的句子编码器genc (1 d-convolution + ReLU + mean-pooling)嵌入整个句子到2400维向量z,后跟一个格勒乌(2400隐藏单位),预计未来3句子与对比损失嵌入形成c。我们用亚当优化器的学习速率2 8日训练gpu的军医,每个国家都有一个批处理大小为64。我们发现，更高级的句子编码器并没有显著提高结果，这可能是由于迁移任务的简单性(例如，在MPQA中大多数数据点只包含一个或几个单词)，而且单词包模型通常在许多NLP任务[48]上执行得很好。

评估任务的结果如表5所示，我们将我们的模型与使用相同数据集的其他模型进行了比较。我们的方法的性能与跃思向量模型非常相似，其优点是它不需要一个强大的LSTM作为字级解码器，因此训练起来要快得多。尽管这是一个标准的迁移学习基准，我们发现在儿童书籍中学习到更好关系的模型不一定在目标任务中表现得更好(这是非常不同的:电影评论等)。我们注意到更好的[49,27]结果已经在这些目标数据集上发布，通过转移学习从不同的源任务。



3.4强化学习

最后,我们评估拟议的无监督学习方法在五个强化学习 DeepMind实验室的荷兰国际集团(ing)3D环境中:rooms\_watermaze,explore\_goal\_locations\_small, seekavoid\_arena\_01,lasertag\_three\_opponents\_small rooms\_keys\_doors\_puzzle。

此设置与前三种不同。本文以标准成批A2C[52]代理为基础机型，增加CPC作为辅助损耗。我们不使用重播缓冲区，因此预测必须适应策略不断变化的行为。习得的表征对其未来观察的分布进行编码。

采用与[50]相同的方法，我们对RMSProp[53]的熵正则化权值、学习率和epsilon超参数进行随机搜索。A2C的展开长度为100步，我们预测未来将达到30步，从而得出对比损失。基线代理由一个卷积编码器组成，它将每个输入帧映射成一个单独的矢量，后面跟着一个暂时的LSTM。我们使用与基线代理中相同的编码器，仅为对比损失添加线性预测映射，从而使开销最小化，这也显示了在针对特定任务设计和调优的现有架构之上实现我们的方法的简单性。我们参考[50]了解所有其他超参数和实现细节。

图6显示，在经过10亿帧的训练后，5个游戏中有4个游戏的代理性能随着对比损失的增加而显著提高。对lasertag\_three\_forenents\_small来说，对比损失既没有帮助也没有伤害。我们怀疑这是由于任务设计造成的，它不需要内存，因此产生了一个纯粹的反应策略。

1. 总结

在本文中，我们提出了对比预测编码(CPC)，一个框架，以提取紧凑的潜在表示法编码预测的未来观察。CPC将自回归建模和噪声对比估计与预测编码的直觉相结合，以无监督的方式学习抽象表示法。我们测试了这些表示在各种各样的领域:音频、图像、自然语言和强化学习和实现强或状态时的艺术性能作为独立的特性。训练模型的简单性和低计算要求，以及在具有挑战性的强化学习领域中令人鼓舞的结果，当与主要损失一起使用时，是朝着普遍适用于更多数据模式的有用的无监督学习的令人兴奋的方向发展。