Abstract-摘要，

在GPU上加速卷积神经网络（CNN）通常包括两个阶段：训练和推理。传统上，这个两阶段过程部署在高端 GPU配备的服务器。在终端计算能力的提高和移动GPU的驱使下，人们对在各种各样的平台上performing inference的兴趣越来越大。在训练阶段要满足高吞吐量和精度的要求，与此相反的是，终端用户会面临与inference tasks相关的不同的要求。

为了解决这个兴起的趋势和新的需求，我们提出Pervasive CNN，一种用户满意的CNN inference 框架。P-CNN是由两个阶段组成的：跨平台离线编译和运行时管理**。**基于用户的需求，离线编译使用架构生成最佳内核独立技术，如自适应批量大小选择和协调精调。 运行时管理阶段 包括精度调节，执行和校准。 第一， 精度调节动态识别最快的内核同时保证可接受的精度。 接下来，运行时内核调度器给每个层划分最优计算资源，并且负责调度GPU线程块。 如果它的精度不是最终用户可以接受的，校准阶段选择较慢的但更精确的内核以提高精度。 最后，我们为CNN设计了一个用户满意度度量来评估我们的普适设计。我们的评估结果表明P-CNN可以为不同的推理任务提供最佳的用户满意度**。**

在GPU 上加速卷积神经网络通常包含两个步骤：训练和推理（Inference）。传统上，这两步处理都在含高端GPU配置的服务器中进行。台式机和手机GPU计算力的不断增长，使得在各种各样的平台上进行推理的兴趣不断地增长。与训练阶段所需要的高吞吐量和高精度不同，在与推理相关的任务中，终端用户面临着更多样化的要求。

为了跟进这一新兴趋势，满足新的要求，我们提出了 Pervasive CNN （P-CNN），这是一个注重用户满意度的CNN 推理框架。P-CNN 有两个组成部分：跨平台离线编译（compilation）和运行时间管理。根据用户的需求，离线编译使用架构独立技术，比如自适应批处理量大小选择（adaptive batch size selection ）和协同微调（coordinated fine-tuning）来生成最佳内核（optimal kernel）。运行时间管理阶段由准确率修正、执行和校正组成。首先，准确率修正会在可接受的准确率下动态地确定最快的内核。然后，运行时间内核调度器会为每一层神经网络分配出最优的计算资源，调度GPU 线程。如果获得的准确率对于终端用户来说是不能接受的，校正阶段会选择一个更慢但是更加精确的内核来提升准确率。最后，我们为 CNN 设计了一个用户满意度衡量指标，用于评估我们的 Pervasive 设计。我们的测试结果显示，P-CNN 在不同的推理任务中提供最佳的用户满意度。

**1.**引子

近年来，深度学习based方法，特别是深度卷积神经网络，已经崛起，在很多领域都成为不可或缺的工具，从图像识别到自然语言处理。这主要是由于他们的增长的能力达到高准确率在许多挑战性的机器学习问题。例如，微软最近宣布，它的最新的深度学习网络（PReLU-nets）已经超过了人类水平的准确率在一个分类任务中，使用1000中ImageNet数据集。提高的准确率伴随着显著的计算量的增长。例如，VGGNet（在深度学习社区最受欢迎的CNN之一）要求1.5x（10\*\*10）浮点数乘法每个图片去执行物体识别。为了解决这个挑战，研究人员已经开始寻求CNN加速器的帮助。The GPGPU，拥有大规模并行处理能力，已经成为最受欢迎的深度学习加速器。例如，自从2013年，超过80%参加ImageNet图像识别竞赛的团队使用了GPU。

在GPU上加速CNN通常包括两个阶段。首先，一个CNN模型要在一个大规模的已经打好标签的训练数据集上进行训练。训练阶段，是最耗时的，（例如，在four high-end NVIDIA Titan Black GPUs上训练VGGNet花了3个星期），会交给高端的GPU设备去执行。一旦完成，网络就会被部署去运行inference，用它的先前已经被训练好的参数去分类、识别和处理未知输入。CNN的高准确率催生了一大批人工智能应用，比如DeepFace和Prisma。DeepFace，一款Facebook开发的深度学习面部识别系统，在识别人脸方面接近人工的准确率。Prisma能够把普通的照片转换成著名的艺术作品的风格。由Prisma生成的照片正在占领Instagram。通常来说，这些应用把从移动平台上获得的数据发送到数据中心，在那儿用高端GPU去perform the inference。最近，随着移动GPU的计算能力不断增长，人们对在移动平台上perform inference的兴趣越来越大（比如，loT sensors，移动智能手机）。这使得一些应用，像Facebook Moments（识别一张照片里有哪些朋友）和Google翻译app（通过手机摄像头的实时取景翻译菜单和标识）。结果是，以CNN为基础的应用正在所有的GPU平台上普及。但是，由于GPU硬件资源的差异，在一个GPU上的CNN的最佳结构用在另一个GPU上可能并不适用。这激发了我们研究CNN在GPU上的具体实现，以及制作一个分析的模型去预测在不同GPU架构上的CNN的性能。通过使用一个平台无关的分析模型，在高端GPU上训练得到的CNN模型可以被有效地部署在不同的平台上，不需要经过耗时的再训练。

需要指出很重要的一点，这两个阶段（训练和inference）的最优目标是不同的。训练阶段的目标是尽快获得高准确率。然而在The inference阶段，因为更接近终端用户，它的目标多种多样，要基于具体的任务。实时任务执行inference时延时要小，比如实时监控和自动驾驶。后台和批处理任务，比如部署在数据中心的DeepFace，对实时性不敏感，但是更关心能耗。一些互动任务，比如Google翻译app和Prisma，对实时性敏感，但是可以忍受一些延时。除了运行时间和能量，基于CNN的应用都有一个鲜明特征：准确率。有趣的是，在inference阶段并不总是追求高准确率，特别是当一个CNN的准确率可以超过人类水平的准确率的时候。举个例子，一些娱乐应用不需要高精度，比如Moments。降低准确度甚至可以提高用户体验，因为不显著的精度降低可以使得响应更快。对于一些在安全和科学研究方面的应用，高精度是必要的。因此，准确度对于用户体验的影响同样是由具体任务决定的。为了评估inference阶段的用户体验，我们首先研究了，，，然后提供了

为了保证在各种各样的平台上的理想的用户满意度，我们提出Pervasive CNN，a user satisfaction-aware CNN inference框架。P-CNN有两个阶段：跨平台线下编译和运行管理。基于使用了CNN的应用的规格，P-CNN可以推断出终端用户的需求。然后，根据这些需求，线下编译会针对要部署的平台生成最佳内核，还会为运行管理阶段提供调度信息。运行管理阶段包括精调，执行和和校准。首先，精调动态的标识出在可接受的精度范围的最快的内核。在此阶段，会为接下来的校准产生一系列调谐表。在运行态，当输入数据改变时，P-CNN监控的输出精度。如果终端用户不能接受这个精度，那么校准会选择一个l不太激进的调谐表（对应一个较慢但更精确的内核）来提高输出精度。

为了保证基于CNN的应用程序的实际响应时间在终端用户的忍受范围之内，我们首先开发了一个平台独立的时间模型来估计CNN的执行时间。然后用这个时间模型来指导跨平台的离线编译。对于特定的GPU架构生成最优内核，我们首先介绍深度学习库两个主要调优参数：子矩阵的大小和每个线程的寄存器消耗，这也决定了卷积核的性能。然后我们设计了一个分析指标并通过协调地微调这两个参数来选择最优内核。

通过与多个运行时调度器方案比较， P-CNN在各种GPU平台上执行不同的推理任务时获得了最佳用户满意度。 评价结果还表明我们的基于熵的近似方法达到相同的效果（即在10％精度损失内可以获得1.8倍加速），作为基于精度的方法。

总之，这项工作的主要贡献是：

我们提供普及CNN，一种提高用户满意度的 CNN推理框架，能够有效地在各种平台上部署经过CNN训练的模型而不需要耗时的再训练。

我们提出了平台独立的分析模型（即时间和资源模型），以指导跨平台的离线编译。

我们的第一步是帮助计算机架构团队去理解现有的GPU架构在支持基于CNN的应用程序的不足。

我们开发了一种运行时精度调节技术，可以动态的根据部署环境改变CNN的精度。

本文的其余部分安排如下。 第二节介绍背景。 第三部分深刻 基于学习的应用程序。 第四节描述了我们的普遍CNN，包括跨平台平台离线编译和runYtime管理。 第五节评估我们的普适设计。 相关工作和结论分别在第六节和第七节中讨论。

2.背景

A．CNN

我们以AlexNet为例介绍CNN。如图1所示。它由五个卷积层，三个max池化层，和两个分类器层组成。 卷积层执行卷积滤波器和输入特征图的局部区域的点积。 这些操作决定了CNN计算的执行时间。 卷积层中的卷积操作受益于优化的矩阵乘法库。 cuBLAS [9]是一个这样的库，实现基本线性代数子程序（BLAS），在Nvidia CUDA运行时库的顶部。

我们在图2中展示了一种典型的基于矩阵乘法的卷积运算。

在步骤1中，一个被称为 im2col的操作伸出本地

区域在输入图像中成columnYmajor矩阵（Dm），类似地，在步骤2中，CONV层的权重为 伸出到过滤器矩阵 （Fm ）。 然后原来的卷积运算可以转换成步骤3中的矩阵乘法（Fm\*Dm）。在图2中，过滤器矩阵Fm的维度是——，而数据矩阵Dm的维度是。输出矩阵Om的维度是——。所以我们可以通过Fm\*Dm的乘累加操作计算出一个卷积层的浮点运算量（即FLOPs）

Conv = 2Nf，，，，

一个乘累加操作计为2个浮点运算量。该转化率通常用来衡量卷积层的计算强度。

B.用户对CNN的满意度

虽然我们知道用户对CNN的满意度是和运行时、能耗和精度相关的，但是这三者对用户满意度的影响在不同的任务中是不同的。为了准确的反映出这三个因素如何决定用户满意度，我们将CNN应用程序分为三类——交互式任务，实时任务，和后台任务。首先，我们分别用这三类任务来分析运行时和能耗对用户满意度的影响。然后我们将深入讨论CNN应用的独特因素（即精度）。

1,。交互式任务

由于大多是交互式任务是面向用户的，，比如人工智能的手机应用（例如，Prisma和Google翻译），他们对运行时敏感。如果应用程序的响应时间超过人类可接受的限度，用户会放弃使用该应用程序。然而，用户也可以容忍一些延迟，因为他们

理解人工智能应用程序涉及复杂计算。 我们在图3中阐释了用户 满意度，运行时间和能耗之间的关系。运行时间在x坐标轴上从左到右增加。之前的工作[10]表明，交互式的应用程序的用户满意度可以分为三个不同的状态：随着应用程序运行时增加：感觉不到的、可容忍的和不可用的。

在感觉不到的区域，运行时增加不会危害用户满意度，但可以节约更多的能量。因此，在这一阶段没有必要去追求最快的响应速度。相反，我们应该努力

通过降低性能来最小化能耗，使运行时接近Ti。超过Ti之后，用户满意度进入可容忍区域，用户满意度会不断降低因为用户开始注意到缓慢的响应时间。然而，运行时在此区域仍然可以接受。当运行时进一步增加，用户满意度最终在Tt时被摧毁。在Tt和之后，用户会放弃这个任务。结果是，用户满意度在不可用区域降到0。我们应该努力避免在这个区域运行任务。

2.实时任务

实时任务有严格的运行时要求。例如，自动驾驶车辆的行人检测有严格的执行时间限制，如果不满足这一限制，可能会导致严重的交通事故。因此，实时任务的曲线没有图3中那样的可容忍区域。它的感觉不到的和不可用的区域几乎和交互式任务相同。但是，在感觉不到区域的结束（即Ti）应小于最后期限。

3.后台任务

后台任务通常对运行时不敏感，因为用户不期望结果具有更长的持续时间。相反，这些任务更关心能耗。例如，在获取许多照片之后，用户会切换到其他应用程序，如网页浏览等，而Moments正在识别在这些照片里有哪些朋友。虽然用户不需要立刻发送照片给他们的朋友，他们更加在意Moments的耗电量因为他们想要更长的电池寿命。因此，后台任务的运行时间不会影响用户满意度。在图3中，后台任务的整个区域都是感觉不到的区域。注意，随着运行时间的增加，能耗首先降低，然后平稳。这是因为能耗是product of power and runtime。在Tc和之后的时间里，power的decrease被runtime的increase抵消了。

4.基于CNN的应用的精度

除了运行时和能耗，CNN应用程序有一个独特的标识：精度。有趣的是，高精度在推理中不总是受青睐的，特别是当最先进的CNN的精度可以超越人类水平的精度。 例如，一些娱乐应用不需要高精度，如Moments。 降低准确度甚至可以提高用户体验。不明显的精度下降可以带来更快的反应时间。在这种情况下，找到最佳精度很重要。低精度会影响用户满意度。高精度会导致更重的计算负担。

传统上，研究人员使用标记的测试数据集来评价CNNY应用的准确性。 然而，不可能创建一个标记数据集包含所有的应用程序。 一个可行的选择是在CNN应用中实现自我诊断。在运行时，他们可以利用其输出来诊断自己。由于缺少标记的测试数据集，我们使用网络不确定性[12]而不是精度作为度量。我们以一个分类任务为例。每个输入数据应该属于一个由Y {1，...，N}表示的特定类别。但是，在运行时我们不知道正确的类标签。CNN分类器层（Y）的输出是概率分布P。例如，P（0.6,0.3,0.1）表示输出属于class1的可能性是60％，属于class2的可能性是30%。我们用输出的熵去衡量输出的不准确性。

熵值表达式是

H（Y）

更高的熵值意味着分布P具有更多的不确定性。例如，P1（0.4,0.4,0.2）的熵高于P2（0.7,0.2,0.1），因为P1在class1和class2两者中高度混淆，而P2更加确定输出属于class1。高不确定性（即高熵值）意味着低精度。如表1所示，网络的精度随着熵的减小而提高。因此，我们利用网路熵作为精度的度量来判断一个CNN的输出是否可靠。

3.characterization

A．实验设置和方法

为了表征CNN的实际部署和覆盖所有可能你的应用场景，我们部署了三个ImageNet获胜者模型（AlexNet [1]，GoogLeNet [13]和VGGNet [4]）到四个不同的平台（服务器，台式机 ，笔记本，和移动端）。详细参数列于表二。

我们使用Nvidia Visual Profiler收集GPU运行时信息。CNN模型由Caffe [15]训练，一个学术界和工业界都广泛使用的开源深度学习框架。我们characterize三个最新的深度学习库：cuBLAS（Caffe采用），cuDNN [16]（由NVIDIA开发的主要工业库）和Nervana [17]（最快的实现），在CNN应用的加速上。测试集是从ImageNet [3]上选取的。

B．inference prefers a small batch size

对于训练阶段，通常是批量输入几百个数据并同时处理。这种批处理方法不仅防止过拟合，而且通过摊销从GPU内存加载权重的开销，能显著提高GPU效率。例如，Karen et al. [4]使用随机梯度下降法（SGD）和一批256的样本量来训练VGGNet。批处理方法也有利于推理时的处理吞吐量，它显著增加网络延迟（即网络端到端响应时间），特别是在移动设备平台上。 表III比较了不同的网络延迟.

考虑到移动GPU的计算能力有限，我们在训练时选择比256小的样本量。AlexNet用128，GoogLeNet用64，VGGNet用32。如表三所示，尽管Nervana在这三种最先进的深度学习库中获得了最快的性能，它也需要大约400ms在TX1上使用小批量训练幼稚的CNN（即AlexNet）。

这远远无法满足最终用户的实际需求。更糟的是，cuDNN和Nervana不能够在移动GPU上使用批处理方法运行GoogLeNet和VGGNet，因为out-of-memory问题（表三中列为x）。

更重要的是，没有足够的数据。。。例如，大多数互动式任务一次只提交一个请求，然后等待响应。因此，推理倾向于较小的批处理量，因为要求响应时间快而且只有有限可用的输入数据，特别是对实时任务和交互式任务。

C．当前GPU实现的低效率

尽管推理优选小的批处理量来获得较好的用户满意度，高度优化的深度学习库，比如Nervana和cuDNN，对较小的批量表现不佳。例如，Nervana的批处理大小必须是32的倍数。因此，批处理（即，表III中的fontYbold）对应到支持的最小批量大小：32。支持最小批量大小，GoogLeNet的最小延迟TX1仍然大于500ms。 虽然nonY的延迟在笔记本电脑和desNtop平台上的批处理方法可以满足最终用户的要求（例如，<30ms），它们的吞吐量（即每秒处理的图像的数量）最佳。图4显示了没有配料的生产率比到使用分批方法的。