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inference。传统上，这两个步骤的过程是在装配了GPU的高性能服务器上进行的。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。在训练阶段要求高吞吐量和准确率，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：跨平台的线下编译和运行管理。基于用户的需求，通过architecture-independent，线下编译会产生the optimal kernel，比如说adaptive batch size和coordinated fine-tuning。运行管理阶段由精调，执行和calibration组成。第一步，精调动态的定义最快的kernels。然后，运行kernel。。。

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。降低精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，精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