近年来，卷积神经网络（CNN）因其具有高推断精度和强自适应性而被广

泛应用于各种领域，例如：计算机视觉、语音识别等。手机当前已经成为了人们

日常生活中的随身携带之物，它每天都产生着大量与人们相关的传感数据，因此

许多工程项目也尝试着将卷积神经网络应用于移动平台来处理这些数据为人类

提供更加智能的服务。然而，由于受到当前移动平台的资源限制（内存、计算能

力、电池容量等），基于 CNN 模型的应用并没有在嵌入式平台中成为主流。

目前，手机上基于 CNN 模型的应用绝大部分都是采用“客户端-服务器”模

式。然而，这种模式不仅强依赖于网络性能而且会导致用户隐私泄露，所以于手

机移动端直接进行卷积神经网络的高能效前向推断已然成为学术界和工业界所

需共同面对的严峻挑战。本文直面这些挑战，采用一系列优化策略开发出一套可

以高能效运行在 Android 平台的卷积神经网络推断时库，并将其运用在一个生活

日志型应用中，以探索从系统层进一步降低该类场景应用的运行时功耗。论文的

主要工作包括：

1. 基于 OpenCL 异构编程框架使用手机移动端 GPU 进行卷积神经网络的前

向推断过程。通过解析 Caffe、Tensorflow 等深度学习框架训练出来的卷积神经

网络模型权重，在手机端重构网络，并将卷积、内积等计算密集型的操作转移到

手机 GPU 上进行。

2. 通过对卷积神经网络全连接层权重的剪枝、重训操作，降低网络模型的

内存占用。使用稀疏矩阵向量乘代替密集矩阵的内积操作，进一步降低卷积神经

网络的前向推断时间。

3. 为了充分利用当前以及未来移动设备 SoC 所提供的异构计算特征，本文

提出了一种简单有效的方法，使得基于 CNN 模型的应用在运行中可以根据其所

处运行环境中 CPU、GPU 等异构处理器的性能差异，自适应地调整 CPU/GPU 的

计算任务分配量。

4. 针对基于 CNN 的生活日志型应用，本文从系统层分析并建模预测这类应

用的运行时负载，探索使用动态电压频率调节技术（DVFS）调整 CPU/GPU 的

电压、频率来进一步降低基于 CNN 模型应用的运行时功耗之可能性。