广泛部署DL服务，特别是移动服务DL，需要边缘计算的支持。这种支持不仅仅是网络架构层面的设计，边缘软硬件的适配与优化同样重要。具体来说，

1）定制边缘硬件以及相应的优化软件框架和库可以帮助更有效地执行DL；

2）边缘计算体系结构可以减轻DL计算的负担；

3） 设计良好的边缘计算框架可以更好地维护边缘运行的DL服务；

4）公平用于评估Edge DL性能的平台有助于进一步改进上述实现。

A用于DL的边缘硬件

1）移动cpu和gpu:DL应用程序更多如果直接在轻量级边缘设备（如作为移动电话、可穿戴设备和监控摄像头，靠近事件发生的地点。低功耗物联网边缘设备可用于进行轻量级DL计算，以及因此避免了与云的通信，但它仍然需要面对有限的计算资源、内存占用和能源消耗。为了突破这些瓶颈[143]，作者关注ARM Cortex-M微控制器并开发CMSIS-NN，一个有效的NN核集合。用CMSIS-NN研究NNs在臂皮层的记忆足迹-M个处理器核可以最小化，然后DL模型可以安装到物联网设备中，同时实现正常性能和能效。

关于运行CNN层时的瓶颈移动gpu，DeepMon[164]分解使用的矩阵在CNN层中加速高维矩阵。通过这种方式，高维CNN层中的矩阵运算（特别是乘法运算）在移动GPU中可用，并且可以加速。在视图中在这项工作中，各种移动gpu已经部署在edge中设备，可以使用特定的DL模型进行潜在的探索并在实现edge-DL方面发挥更重要的作用。

除了DL推断[143]，[164]，重要因素影响移动cpu上DL训练的性能在[168]中讨论了gpu。自常用DLVGG[169]等型号的内存太大主流边缘设备的大小，相对较小的学员采用网络[170]对DL训练进行评价。评价结果表明，DL模型的大小对于训练性能与移动cpu的有效融合gpu对加快训练进程具有重要意义。

2）基于FPGA的解决方案：尽管GPU解决方案是广泛应用于云计算中的DL训练和推理，然而，由于中国在能源和成本预算方面的困难边缘，这些解决方案可能不可用。另外，边缘节点应该能够为多个DL计算服务一次请求一次，它使得只需使用轻量级CPUGPU不切实际。因此，边缘硬件基于现场可编程门阵列（FPGA）的研究他们对edge-DL的可行性。

基于FPGA的边缘设备可以实现CNN加速任意卷积和可重构池[143]，而且它们的性能比最先进的CPU还要快和基于RNN的GPU实现[145]实现更高能量的语音识别应用效率。在[52]中，一种基于FPGA的开发边缘平台以允许DL计算卸载从移动设备。论fpgabase的实现edge平台、无线路由器和FPGA板是结合在一起的。使用典型的视觉应用，基于FPGA的边缘平台显示了它的优势，在能源消耗方面硬件成本，超过基于GPU（或CPU）的。

尽管如此，仍有待于确定FPGAs是否或者GPU/CPU更适合于边缘计算，如如表四所示。详细的实验在[171]研究FPGAs相对于gpu的优势：1）能够提供不受工作负载影响的吞吐量；2）保证高并发性的高性能DL计算；3）更好的能源效率。然而FPGAs的缺点在于开发高效的DL大多数程序员对FPGA上的算法并不熟悉。尽管像Xilinx SDSoC这样的工具可以大大减少难度[52]，至少现在，额外的工作是仍然需要移植最先进的DL模型，为gpu编程，进入FPGA平台。

B. Edge-DL的通信和计算模式

尽管在设备上进行了DL计算，如第。五，可以满足轻量级的DL服务。然而，一个独立的终端设备仍然无法承受密集的DL计算任务。边缘计算的概念可能通过将DL计算从终端设备到边缘或（和）云。伴随着边缘架构，以DL为中心的边缘节点可以成为云计算基础设施的重大扩展有大量的DL任务。在本节中，我们将对四种模式进行分类对于边缘DL计算，如图16所示。

1. 整体卸载：DL最自然的模式计算卸载类似于现有的“端云”计算，即终端设备发送其计算请求到云端进行DL推断结果（如图。16（a））。这种卸载通过释放从DL任务分解到组合问题资源优化，可能会带来额外的计算成本和调度延迟，因此实施。在[172]中，提议的分布式基础设施DeepDecision用较少的资源将强大的边缘节点连接在一起强大的终端设备。在DeepDecision中，DL推理可以在端点或边缘执行，这取决于权衡在推理精度、推理延迟和型号大小、电池电量和网络状况。与对于每个DL任务，终端设备决定是否在本地处理或卸载到边缘节点。

1 考虑到型号精度、视频质量、电池约束、网络数据使用和网络之间的复杂的相互作用，确定最优值的条件卸载策略

实现约15 FPS的视频分析同时具有较高的精度

2. 优化工作负载分配权重和计算容量

3. 将实时视频分析任务卸载到使用EdgeEye API而不是使用特定于DL框架的api提供更高的推理性能

4. 提供上下文感知卸载，策略性模型隔板和管道支架有效利用加工边缘容量

执行5.08和23.0加速，以及53.5%节能85.5%仅可穿戴和手持

5. 对DNN层进行分区并递增上载分区，允许边缘计算和终端协作，改进查询性能和能耗

保持几乎相同的上传整，大大改进查询执行时间

6. 执行数据预处理和初步在边缘学习以减少网络流量，以加快云计算

达到90%的精度，同时减少执行时间和数据传输

7. 适应各种DNN架构、硬件平台、无线连接和服务器负载级别，并选择分区最佳延迟和最佳移动点能源消耗

通过以下方式改善端到端延迟平均3.1 ，最高40.7 ，降低移动能耗平均提高59.5%达到94.7%，并改进了数据中心平均吞吐量为1.5 高达6.7

8. 最小化通信和资源允许低级别使用设备通过EEoI分类

降低通讯成本达到20以上总准确率95%

9. 通过高效的识别，通过交互实现无线通信，在末端、边缘和云层之间

仅损失2.5%的检测精度在图像压缩比下60%，同时显著改善图像传输效率

10. 引入主导需求以确定多种资源之间的最佳权衡和准确性

相比之下，精确度提高了5.4 只差最佳值6%

11. 分区已训练的DNN模型在几个移动设备上加速通过减少偏差来计算DNN计算成本和内存使用

当工作节点的数量从2增加到4，MoDNN可以通过以下方法加速DNN计算2.17-4.28

12. 融合多个CNN层的处理并启用中间层的缓存保存数据传输的数据（带宽）

将总数据传输减少95%，从77MB降至36 MB每张图片

13. 采用可伸缩的熔合瓷砖分区最小化内存的CNN层暴露并行性时的占用空间和一种新的工作调度过程

总体执行延迟将内存占用减少更多比68%还要精确，吞吐量提高1.7 -2.2 加速CNN推断比1.7 -3.5

14. 协调参与的移动用户协同训练域感知改进目标识别的自适应模型精确提高目标识别精度

相比之下是150%只需使用通用DL模型

15. 设计各种任务分配方案为边缘间协作而跟踪的

最小化服务响应时间加速范围从1.3 至4 （1.2 至1.7 ）反对在本地运行（clientcloud

配置）