一些观点：

观点一：

现阶段 app 上使用的深度学习主要有两种模式：

一种是 online 方式：移动端做初步预处理，把数据传到服务器执行深度学习模型，优点是这个方式部署相对简单，现成的框架（caffe，theano，mxnet，Torch）做下封装就可以直接拿来用，服务器性能大，能够处理比较大的模型，缺点是必须联网。

另外一种是 offline 方式：在服务器进行训练的过程，在手机上进行预测的过程。

观点二：

做模型压缩。

观点三：

在移动端，直接拿模型来跑肯定不行，除了软件环境，最重要的其实有两个方面：

第一是存储。为了解决这个问题，在DL领域有相当一部分人在做相关的问题，比如二值网络、网络压缩。其目的就是减少模型的参数规模利于存储。将权重怎么存可能也需要各种技巧吧。

第二是计算。这个跟存储和平台也有一定的关系。以上我提到的这两个工作是object detection领域比较新的且能实时处理的工作。在mobile设备上不可能拿模型直接用，要做一下简化和妥协。另外可行的方案是在服务器端跑模型。

另外还有一个比较有意思的工作是最近有一小部分人在做针对神经网络输入的图像采集设备，直接生成神经网络图片。这样就可以直接在输入端上解决存储的问题。

观点四：

注意三点：1、权重存储；2、计算速度；3、功耗。
三点都很重要，可以看看斯坦福William dally组的工作（删减连接、编码存储、混合精度）

观点五：

可是真正的工业界哪里有这么简单？哪里有现成的数据集？数据集怎么构造？如何清洗？数据体量都是上亿 代码写不好 跑不死你？

部门大量工具都是C写的不会C行只会python行吗？这么大数据 不会Hadoop spark搞得定？这么大数据训练 不会参数服务器 不会分布式搞得定？

模型调参就简单？深度学习是黑箱，就更要有比较高的理论水平，否则你连怎么调参的门道都找不到，为什么不收敛，你都想不出来。

模型训练出来就万事大吉了？模型太重速度太慢 不符合业务需求 怎么办？怎么模型压缩？模型怎么上移动端？移动端没有合适的机器学习架构，你能不能写一个？

一个模型上线就完事了？想进一步提高性能，怎么办？是不是要紧跟学术前沿，能读懂paper，快速实验？这对英语和数学编程都有很高要求。

这说的都是深度学习的，然而很多场景是不适合用深度学习的。那决策树模型，进一步的集成学习，随机森林和GBDT你得懂。统计学习的你得懂，贝叶斯SVM和LR你得懂。有些业务还得用HMM，CRF或者NLP的东西。所以还是要先虚心学习一个。

观点六：

1、移动端的训练问题，一般都是用现成的训练好的模型。
2、模型的参数体积问题
3、模型的计算时间问题

当下还是流行使用服务器训练模型，移动端部署模型前向计算。用移动端进行训练现在的移动设备的内存及计算能力都有点吃力。自己大胆做个猜想：将来可以做成通用模型（服务器训练）加个性化模型（手机上训练）。

至于部署到移动端当然还是会涉及到模型压缩及框架选择。

观点七：

个人总结：

1、服务器端跑模型，移动端给结果，这样做必须联网；

2、压缩模型，将参数权重存储在本地。存储问题需要考虑；

3、框架轻量化；

4、卷积硬件计算加速；

5、优化移动端深度学习引擎；

6、一些优化：使用低层语言；缓存友好：少用内存、连续访问、对齐访问、合并访问；多线程：线程开销、动态调度；稀疏化；定点化；包大小；编译优化；代码精简；模型压缩；内存精简；兼容性与可靠性；