蒋豆芽

深度学习——模型优化 已购

来自 【机器学习面试题汇总与解析（蒋豆芽面试题总结）】 | 63 浏览 | 0 回复 | 2021-06-05

蒋豆芽

+关注

机器学习面试题汇总与解析——模型优化

1. **若CNN网络很庞大，在手机上运行效率不高，对应模型压缩方法有了解吗** ☆☆☆☆☆
2. **介绍一下模型压缩常用的方法？为什么用知识蒸馏？** ☆☆☆☆☆
3. **知道模型蒸馏吗？谈下原理** ☆☆☆☆☆
4. **做过模型优化吗？模型蒸馏和模型裁剪？** ☆☆☆☆☆
5. **squeezeNet的Fire Module有什么特点？** ☆☆☆☆
6. **降低网络复杂度但不影响精度的方法** ☆☆☆☆☆
7. **如果让模型速度提高一倍，有什么解决方案？** ☆☆☆☆☆

==========

- 本专栏适合于**Python已经入门**的学生或人士，有一定的编程基础。
- 本专栏适合于**算法工程师、机器学习、图像处理求职**的学生或人士。
- 本专栏针对面试题答案进行了**优化，尽量做到好记、言简意赅。这才是一份面试题总结的正确打开方式。这样才方便背诵**
- 如专栏内容有错漏，欢迎在评论区指出或私聊我更改，一起学习，共同进步。
- **相信大家都有着高尚的灵魂，请尊重我的知识产权，未经允许严禁各类机构和个人转载、传阅本专栏的内容。**

==========

关于**机器学习算法**书籍，我强烈推荐一本**《百面机器学习算法工程师带你面试》**，这个就很类似面经，还有讲解，写得比较好。私聊我进群。

关于**深度学习算法**书籍，我强烈推荐一本**《解析神经网络——深度学习实践手册》**，简称CNN book，通俗易懂。私聊我进群。

关于**模型优化**，先学习一下理论知识：推荐看看文章，**《解析神经网络——深度学习实践手册》第四章——卷积神经网络的压缩**

常见面试题汇总

蒋豆芽

参考回答

1. 低秩近似
2. 剪枝与稀疏约束
3. 参数量化
4. 二值网络
5. 知识蒸馏
6. 紧凑的网络结构

答案解析

1. **低秩近似**

   神经网络的基本运算卷积，实则就是**矩阵运算**，低秩近似的技术是通过一系列小规模矩阵将**权重矩阵**重构出来，以此降低运算量和存储开销。目前有两种常用的方法：一是$Toeplitz$矩阵直接重构**权重矩阵**，二是**奇异值分解（SVD）**，将权重矩阵分解为若干个小矩阵。

2. **剪枝与稀疏约束**

   剪枝是模型压缩领域中一种经典的后处理技术，典型应用如**决策树**的前剪枝和后剪枝。剪枝技术可以减少模型参数量，防止过拟合，提升模型泛化能力。

   剪枝被应用于神经网络中，遵循四个步骤：

   - 衡量神经元的重要程度
   - 移除掉一部分不重要的神经元
   - 对网络进行微调
   - 返回第一步，进行下一轮剪枝

   稀疏约束与直接剪枝在效果上有着异曲同工之妙，其思路是在网络的优化目标中加入权重的稀疏正则项（如l_1和l_2正则化），使得训练网络的部分权重趋向于0，而这些0值元素正是剪枝的对象。因此，稀疏约束可以被视为动态的剪枝。

   相对剪枝，稀疏约束的优点很明显：只需进行一遍训练，便能达到网络剪枝的目的。

3. **参数量化**

   相比于剪枝操作，**参数量化**则是一种常用的后端压缩技术。量化就是从权重中归纳出若干个有代表性的权重，由这些代表来表示某一类权重的具体数值。这些“代表”被存储在码本（codebook）中，而原权重矩阵只需记录各自“代表”的索引即可，从而极大地降低了存储开销。所以可以看出来，**参数量化就是换一种存储方式从而降低模型的存储。**

4. **二值网络**

蒋豆芽

现有神经网络大多基于梯度下降来训练，但二值网络的权重只有+1或-1，无法直接计算梯度信息，也无法更新权重。**一个折中的方法是，网络的前向与反向回传是二值的，而权重的更新则是对单精度权重进行。**

5. **知识蒸馏**

   知识蒸馏指的是模型压缩的思想，通过一步一步地使用一个较大的已经训练好的网络去教导一个较小的网络确切地去做什么。将从大而笨重中需要的知识转换到一个小但是更合适部署的模型。

   知识蒸馏的目的是将一个高精度且笨重的teacher转换为一个更加紧凑的student。具体思路是：训练teacher模型softmax层的超参数获得一个合适的soft target集合（“软标签”指的是大网络在每一层卷积后输出的feature map。），然后对要训练的student模型，使用同样的超参数值尽可能地接近teacher模型的soft target集合，作为student模型总目标函数的一部分，以诱导student模型的训练，实现知识的迁移。

6. **紧凑的网络结构**

   以上的方法都可以理解为后处理。而**紧凑的网络结构方法**则是另起炉灶，直接设计短小精悍的结构保证网络的速度与精度。如squeezeNet的Fire Module模块和GoogLeNet的Conv-M模块。

类似的问题还有：

2. **介绍一下模型压缩常用的方法？为什么用知识蒸馏？** ☆☆☆☆☆

   参考回答

   1. 低秩近似
   2. 剪枝与稀疏约束
   3. 参数量化
   4. 二值网络
   5. 知识蒸馏
   6. 紧凑的网络结构

   深度学习在计算机视觉、语音识别、自然语言处理等众多领域取得了令人难以置信的成绩。然而，这些模型中的大多数在移动电话或嵌入式设备上运行的计算成本太过昂贵。所以需要采用模型压缩的方法，如知识蒸馏。

   知识蒸馏指的是模型压缩的思想，通过一步一步地使用一个较大的已经训练好的网络去教导一个较小的网络确切地去做什么。**将从大而笨重中需要的知识转换到一个小但是更合适部署的模型。**

   答案解析

蒋豆芽

3. **知道模型蒸馏吗？谈下原理** ☆ ☆ ☆ ☆

   参考回答

   知识蒸馏指的是模型压缩的思想，通过一步一步地使用一个较大的已经训练好的网络去教导一个较小的网络确切地去做什么。**将从大而笨重中需要的知识转换到一个小但是更合适部署的模型。**

   知识蒸馏的目的是将一个高精度且笨重的teacher转换为一个更加紧凑的student。具体思路是：训练teacher模型softmax层的超参数获得一个合适的soft target集合（“软标签”指的是大网络在每一层卷积后输出的feature map。），然后对要训练的student模型，使用同样的超参数值尽可能地接近teacher模型的soft target集合，作为student模型总目标函数的一部分，以诱导student模型的训练，实现知识的迁移。

   答案解析

   无。

4. **做过模型优化吗？模型蒸馏和模型裁剪？** ☆ ☆ ☆ ☆ ☆

   参考回答

   参考上面回答。

   答案解析

   无。

5. **squeezeNet的Fire Module有什么特点？** ☆ ☆ ☆ ☆

   参考回答

   1. **挤压**：高维特征有着更好的表示能力，但使得模型参数急剧膨胀。为追求模型容量与参数的平衡，可使用1×1的卷积来对输入特征进行降维。同时，1×1的卷积可以综合多个通道的信息，得到更加紧凑的输入特征，从而保证了模型的泛化性。
   2. **扩张**：通常3×3的卷积占用了大量的计算资源。可以使用1×1的卷积来拼凑出3×3的卷积。

   答案解析

   无。

6. **降低网络复杂度但不影响精度的方法** ☆ ☆ ☆ ☆ ☆

蒋豆芽

答案解析

无。

7. **如果让模型速度提高一倍，有什么解决方案？** ☆☆☆☆☆

参考回答

1. 低秩近似
2. 剪枝与稀疏约束
3. 参数量化
4. 二值网络
5. 知识蒸馏
6. 紧凑的网络结构

答案解析

无。

资源分享 python 机器学习 算法工程师 春秋招 面试题 软件开发 面经 举报

收藏 赞

相关专栏

机器学习面试题汇总与解析（蒋豆芽面试题总结）

27篇文章 | 90订阅 已订阅

0条评论 默认排序

没有回复

蒋豆芽

发布

牛客博客，记录你的成长

关于博客 | 意见反馈 | 免责声明 | 牛客网首页