人工智能

算法岗

江湖武林秘籍（上）

武林盟

江湖之中风云莫测

现有一本

算法岗位面试必备

献给各路豪杰

自序

历时 1 年零 3 个月，修整 7 个版次的大厂算法岗江湖秘籍，终于成型。最近几年，江湖中一直流传，算法岗内卷严重的传闻。让无数想转行、或者毕业的侠士，望而却步。但机会是留给有准备的人的，但如何有计划的准备，却成为无数侠士头痛的难题。

因此，大白才产生构建一套算法岗武林秘籍的想法，将江湖中各个大厂可以搜集到的面试资料，按照逻辑框架，每个大厂都整理成一篇面经。

让大家有计划，有目标的准备，希望为大家行走江湖，提供一些便利。

天涯未远，江湖再见！

江大白

公众号：江大白

微信：wxqzy68

www.jiangdabai.com

人工智能算法岗江湖目录

武林秘籍（上册）

武林秘籍（中册）

武林秘籍（下册）

练功必读

在写这个版块的内容时，大白内心其实五谷杂陈。因此，特想聊一下，为什么要整理这一套《人工智能算法岗江湖武林秘籍》？

《缘由》

当从学生时代毕业求职，或者说，在工作岗位中，想跳槽到心仪的公司，其实大多数人，内心都是惶恐的。

因为对未知的不确定，不知是否可以求职成功，或者跳槽成功。

所以我们常常会查看网上的各种面试经验，但有两个体验不太好的问题：

① **每个人分享的面试经验都是分散的，不成体系的。**因此有的时候，看了很多的内容，但回过头来想一想，却又无法记住，哪些岗位需要哪些经验？会问哪些问题？

② **网上的人工智能面试资料很多，比如列出几百个常问的知识点。**

但不同的公司，需要了解的知识点？甚至不同的岗位，会问的知识点，都不相同。这两个问题，对我们在人工智能算法岗的求职中，造成了很大的困扰。

而成年人的时间，本身就是稀缺的。如何花最少的时间，学习最重要的内容，明白心仪的公司，具体的岗位需要哪些准备？非常关键。

这也是大白想做这套算法岗江湖秘籍的缘由，**有人的地方，就是江湖。**就像阿里巴巴的江湖文化，而算法的领域，也有它自身的江湖。

《剑起》

那么，如何消除求职和跳槽的未知惶恐感呢？

每个人都有自己的方式，比如最直接的，就是找到内部相关岗位的员工，进行咨询。

但比较尴尬，并不是每个人身边正好都有这样的朋友，因此这也是大白写秘籍的第二个缘由，让秘籍成为大家身边的朋友，一起仗剑行走江湖。

① 算法岗面经秘籍

当然算法岗面经秘籍，并不是小抄。大白更想做的是，让大家可以更好的梳理自己的知识框架，明白不同的公司，更在意面试者的哪些经验积累？

更会在学习中督促反思自己，如果面试心仪的公司，当问到自己相关问题的时候，该如何作答？是否掌握相关知识，这样才能不断提升自己的能力？

② 专利核心秘籍

当然，了解公司的实际需求。除了面经，可能很多人，都不会关注一个点：专利。

科技型公司，在做不同的项目，或者技术上有突破时，通常会将核心技术，申请为专利。而了解科技型公司的实际需求，调研竞争力，都可以从专利上直接反馈出来。

当然核心的算法岗，更是关联公司的很多专利技术。那么心仪的公司，做过哪些项目？掌握哪些知识？面试中会常问哪些技术点？都可以从专利中获取。

比如面试某 CV 图像类的公司，通过专利检索了解到，做了很多目标检测、人脸识别相关的项目，那么这些项目涉及哪些知识点？通常也是公司技术面试官常问的问题？

在这里，大白也搜集整理了 8 家公司，每一家 200 篇最新的专利资料，相信大家从专利名称上，就可以了解公司的项目倾向，微信扫描下方二维码即可获取。

《磨剑》

为了整理这一套秘籍，大白放弃了很多陪伴家人的时间。从刚开始的准备，到目前的资料，迭代了 7 次。整理资料->梳理面经框架->再整理资料，一遍又一遍，每一次迭代都需要一个多月的时间。

比如字节跳动有三百多篇的面经，知识量很大。按照知识框架，最终整理成一篇将近四万字的面经。

当然，大家在查看每个公司面试题时，并不需要全都看完。

比如你的工作经验主要以 CV 图像为主，那么该公司面经中，第二节基础知识中的图像处理、CNN 卷积神经网络、以及常问的基础知识点需要重点关注。

此外，第三节到第七节中的项目经验、数据结构、编程、操作系统、开放性问题也都需要对应查看。

大家将自己的项目经验，代入到心仪公司的面经中。根据会问的问题，反思解答的思路，便于查找自己的答案。

① 思维导图：每个公司都是按照知识框架进行整理的，知识框架的思维导图，大家可以看后面一页，有详细的图示说明。

② 抱团取暖：此外，算法江湖，一个的力量是渺小的，群体的力量是巨大的，大家也可以扫描前面的二维码，加入相应公司的面试群，和大家一起准备，驰骋江湖。

③ 其他秘籍：当然除了面经外，大白在 www.jiangdabai.com 中还制作了很多资料。

视频课程：《30 天入门深度学习》课程、《深入浅出人脸识别基础及项目应用》、《深入浅出人脸特效之 Mask 实战应用》、《深入浅出人工智能平台 Api 项目应用》等。

项目实战：N 个酷炫项目的实战，比如人脸墨镜特效、人像分割等项目。每个项目都包括应用场景、项目实践、详细安装说明等。

资源下载：数百个不同类别的数据集，按照不同的场景，不同的任务，非常详细的划分，项目中需要什么数据集，直接查看即可。

当然除了资料，大白还组建了 AI 行业陪伴成长的社群《AI 未来星球》，点击查看。

可以添加大白的微信：wxqzy68，回复关键词：星球，即可加入。

此外，加入星球还可以领取：

① 腾讯课堂上 198 元的《30 天入门人工智能》视频课程，点击查看。

② 大白自费一万多购买的各类数据集，以及公开数据集，数十个，点击查看。

③ 365 元的《AI 未来星球》知识星球，目前 198 元/年限量活动，各类直播分享，一对一提问。

④ 智慧城市、工业视觉、AI 芯片、自动驾驶，行业经验宝藏文档。

⑤ 邀请加入《AI 未来星球》内部群，以及参与星球内每月的各种活动（例如每天自习打卡群、星球会员日等）。

大白希望用自己的力量，为大家在浪迹算法江湖中，提供一些便捷，足已。

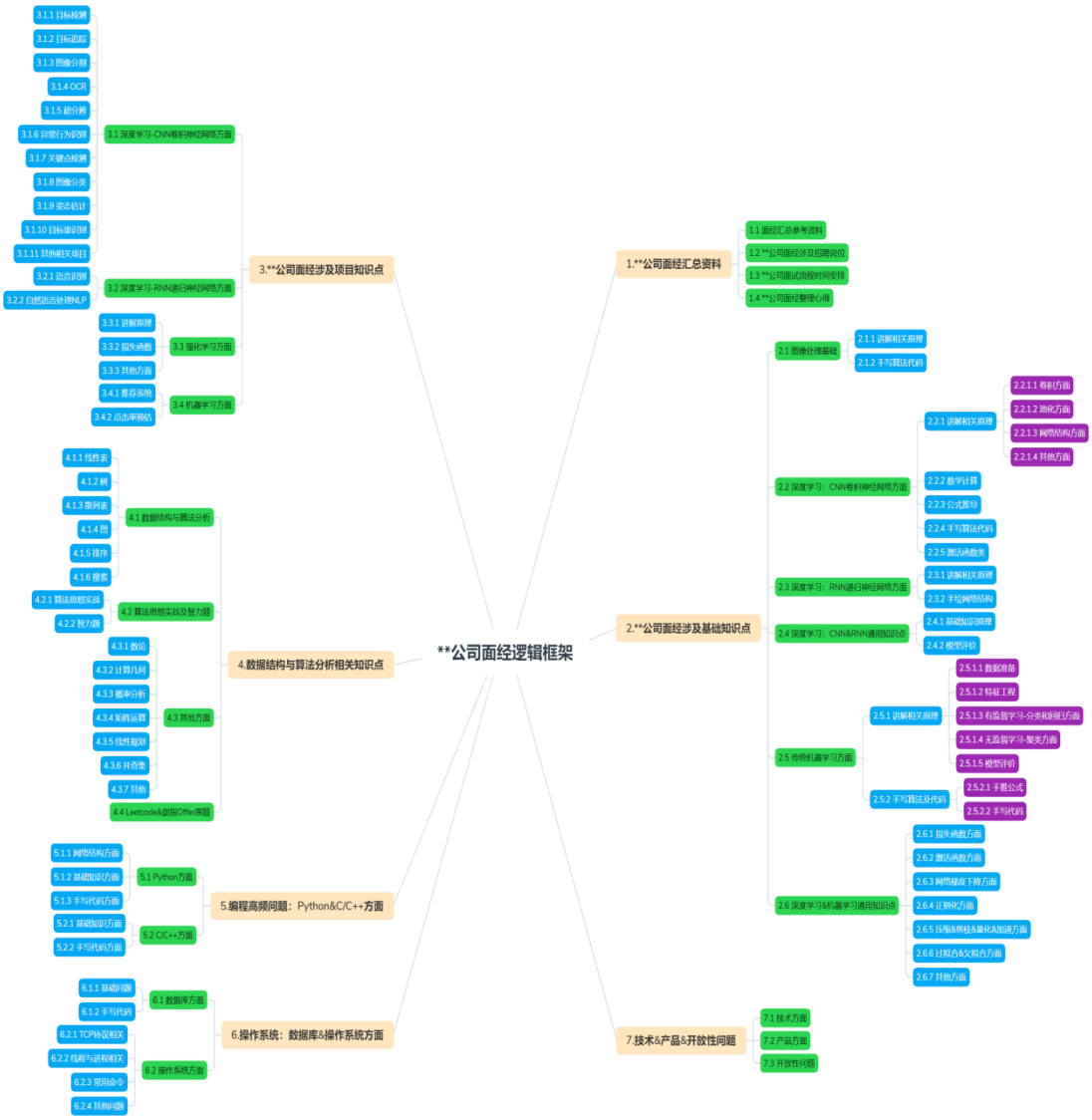
江大白

公众号：江大白

微信：wxqzy68

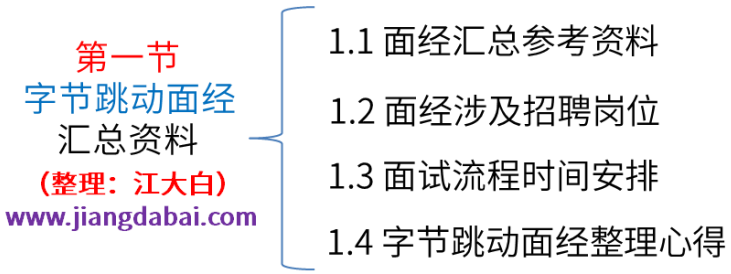
www.jiangdabai.com

知识框架


**公司面经逻辑框架
1.**公司面经汇总资料
1.1 面经汇总参考资料
1.2 **公司面经涉及招聘岗位
1.3 **公司面试流程时间安排
1.4 **公司面经整理心得
2.**公司面经涉及基础知识点
2.1 图像处理基础
2.1.1 讲解相关原理
2.1.2 手写算法代码
2.2 深度学习：CNN卷积神经网络方面
2.2.1 讲解相关原理
2.2.1.1 卷积方面
2.2.1.2 池化方面
2.2.1.3 网络结构方面
2.2.1.4 其他方面
2.2.2 数学计算
2.2.3 公式推导
2.2.4 手写算法代码
2.2.5 激活函数类
2.3 深度学习：RNN递归神经网络方面
2.3.1 讲解相关原理
2.3.2 手绘网络结构
2.4 深度学习：CNN&RNN通用知识点
2.4.1 基础知识原理
2.4.2 模型评价
2.5 传统机器学习方面
2.5.1 讲解相关原理
2.5.1.1 数据准备
2.5.1.2 特征工程
2.5.1.3 有监督学习-分类回归方面
2.5.1.4 无监督学习-聚类方面
2.5.1.5 模型评价
2.5.2 手写算法及代码
2.5.2.1 手推公式
2.5.2.2 手写代码
2.6 深度学习&机器学习通用知识点
2.6.1 损失函数方面
2.6.2 激活函数方面
2.6.3 网络梯度下降方面
2.6.4 正则化方面
2.6.5 压缩&剪枝&量化&加速方面
2.6.6 过拟合&欠拟合方面
2.6.7 其他方面
3.**公司面经涉及项目知识点
3.1 深度学习-CNN卷积神经网络方面
3.1.1 目标检测
3.1.2 目标追踪
3.1.3 图像分割
3.1.4 OCR
3.1.5 超分辨
3.1.6 异常行为识别
3.1.7 关键点检测
3.1.8 图像分类
3.1.9 姿态估计
3.1.10 目标重识别
3.1.11 其他相关项目
3.2 深度学习-RNN递归神经网络方面
3.2.1 语音识别
3.2.2 自然语言处理NLP
3.3 强化学习方面
3.3.1 讲解原理
3.3.2 损失函数
3.3.3 其他方面
3.4 机器学习方面
3.4.1 推荐系统
3.4.2 点击率预估
4.数据结构与算法分析相关知识点
4.1 数据结构与算法分析
4.1.1 线性表
4.1.2 树
4.1.3 散列表
4.1.4 图
4.1.5 排序
4.1.6 搜索
4.2 算法思想实战及智力题
4.2.1 算法思想实战
4.2.2 智力题
4.3 其他方面
4.3.1 数论
4.3.2 计算几何
4.3.3 概率分析
4.3.4 矩阵运算
4.3.5 线性规划
4.3.6 并查集
4.3.7 其他
4.4 Leetcode&剑指Offer原题
5.编程高频问题：Python&C/C++方面
5.1 Python方面
5.1.1 网络结构方面
5.1.2 基础知识方面
5.1.3 手写代码方面
5.2 C/C++方面
5.2.1 基础知识方面
5.2.2 手写代码方面
6.操作系统：数据库&操作系统方面
6.1 数据库方面
6.1.1 基础问题
6.1.2 手写代码
6.2 操作系统方面
6.2.1 TCP协议相关
6.2.2 线程与进程相关
6.2.3 常用命令
6.2.4 其他问题
7.技术&产品&开放性问题
7.1 技术方面
7.2 产品方面
7.3 开放性问题

1|字节跳动算法岗武功秘籍

1 字节跳动面经汇总资料



1.1 面经汇总参考资料

① 参考资料：

（1）牛客网：字节跳动面经-340 篇，网页链接

（2）知乎面经：点击进入查看

（3）面试圈：点击进入查看

② 面经框架&答案&目录&心得：

（1）面经框架及参考答案：点击进入查看

（2）大厂目录及整理心得：点击进入查看

1.2 面经涉及招聘岗位

（1）实习岗位类

【图像与多媒体算法实习】、【Data 搜索部（数据挖掘）实习】、【三维视觉实习】、【自然语言处理实习】、【数据挖掘/搜索/推荐实习】、【效率工程算法实习】、【广告算法实习】、【AI Lab 机器学习实习生】、【商业变现部门推荐算法】、【编解码算法工程师实习】

（2）全职岗位类

【AI Lab 计算机视觉与深度学习岗】、【抖音互娱图形图像算法工程师】、【搜索团队算法工程师】、【研发算法工程师】、【视频基础架构组】，【搜索部门算法工程师】、【广告算法工程师】、【企业应用算法工程师】、【抖音算法工程师图像增强方向】、【飞书算法工程师】、【Data 部门算法工程师】、【推荐算法工程师】、【抖音算法工程师】、【NLP 算法工程师】、【自然语言处理算法工程师】、【机器学习中台算法工程师】、【多媒体视频算法工程师】、【Data 推荐算法工程师】、【字节教育工程师】、【电商 NLP 算法工程师】、【大数据开发工程师】、【教育部门算法工程师】

1.3 面试流程时间安排

字节跳动面试流程-整理：江大白			
	面试类型	面试流程	备注（侧重点）
第一面	技术面	自我介绍+项目/实习经验 +技术问答+算法编程	主要是项目基础知识点， 算法能力也很看重
第二面	技术面	自我介绍+项目/实习经验 +技术问答+算法编程	深挖项目知识细节
第三面	技术Leader面	自我介绍+项目经验+公司发展	/
第四面	HR面	基础人力问题	/

PS：以上流程为大白总结归纳所得，以供参考。

其他注意点：

● 在面试的流程中，需要注意的是，有的人是三轮技术面试，有的人是两轮技术面试。

1.4 字节跳动面试心得汇总

字节跳动的面经超级多，说明机会和发展都是很不错的。以下是大白整理的，几百篇面经中的面试者的心得感悟，将此提炼出来，便于大家体会：

★ 一面二面的面试官比较偏理论基础，三面的大佬比较偏业务，不过总体来说，头条的面试官总的来说都挺好的，编程题检查时也会指错并引导。

★ 阿里更注重底层基础和深度，源码级别，头条更注重算法，手撕代码。常常自我介绍完啥也不说基本都先来一道编程题，然后发问，问到最后，再以一个编程题收尾。

★ 面试时对简历上自己实习或者项目的细节要很清楚，会问的很深。（比如为什么用欧式距离算样本之间的距离）而且涉及的面也会比较广，比如做图像/视频增强的，应该对于超分，去模糊，去噪，去雾，HDR 甚至图像 translation 等问题都应该有一个比较深的了解，希望各位还不急着找工作的同学们能坚持努力。

★ 注重原理的理解而不是方法看起来有多 fancy，原理至上，所以一定要理解透彻。

★ 一定要回忆所有细节，并站着面试官的角度思考他会问什么问题。项目中你做工作时的流程一定要清楚，比如业界一般是怎样解决该问题的，你是怎样做的，你遇到了什么困难，你如何解决这些困难的。如果有的问题不会，或是只有个模糊的答案，直接说出来或这说不会就行了，没有事的。

★ 划重点!!! 项目一定要挑自己熟悉的说，简历上放一些和岗位相关的项目。

★ 心态部分，战略上藐视，战术上重视。

★ 面试官似乎很看重工程能力 而我的经历都很 Research 所以除了表达自己算法能力的同时，把工程上的东西说一说也是很加分的，哪怕是脏活累活。

★ 大家面试的时候放松心态，做足准备就好，谋事在人，成事在天，不必太过紧张。如果面试中遇到思路卡壳可以一点一点解释，不用着急。

★ 算法工程师的自我修养：数据建模，C/C++，思考、解决问题的方向和逻辑（建立在足够的理论基础和实践基础上）

★ 多运用到工程，想一想工程方向的优化。在有算法的基础做一做开发对于自己的成长有帮助。

★ 项目中每一个创新点一定要清楚：为什么用、怎么用的、好处在哪里。

★ 一些基本的概念一定要熟悉，不能只是知道。比如 ROC 曲线和 PR 曲线，面试官的要求不仅仅是横轴纵轴是什么，往往会有进一步的 follow up：比如说样本分布不平衡的话 PR 和 ROC 会有差别类似的；不能只关注于某一个概念是什么，往往最基础的 follow up 是考察的重点。

★ 各种基本算法除了原理要了解他们的优缺点，这里应该是面试官考察比较核心的地方。平

时不能只顾着刷题，而没有去练习怎么在人前把一个算法表述清楚。

★ 感觉算法岗真的没有那么难，没必要散播焦虑，感觉年轻人有无限可能，大家多花点时间真正去做事情了，基本结果都不会差。

★ 可以感受一下面试官问问题的思路和感兴趣的角度。前两轮面试更加注重于专业知识，技术细节，了解你做的多深。后两面更注重了解你思考问题的方式以及想做的内容和团队是否match。

★ 字节跳动是一个很注重基础的公司，他不会要求你有竞赛有 paper 有多么强的工业界能力，但是基础一定要好，至少对于算法工程师来说现场手撕 code 的时候 bug free 是必须的，因为面试官不会留太长的时间给你debug，除非能够精准定位bug并快速解决，不然凉的几率很大。

★ 字节的面试是我面的最硬核的，就是会一直问到底，看你到底掌握到什么程度，如果掌握的不深刻很容易就被问出来了 。

★ 在面试时，针对项目，面试官会假设他不懂这个项目，将项目从头讲到尾

★ 春招的笔试也很难（不要以为实习的笔试就简单了）

★ 早点准备什么时候都不会错的（毕竟机会是留给有准备的人）

★ 发现形式不对,要快刀斩乱麻；今年算法岗从春招的形式就可以看到秋招的严峻了

★ 认清自己（这一点确实有点难），但这确实是重中之重

★ 好好刷题，不要抱着内推免笔试提起批免笔试的心态，即使免笔试，coding 也是你没办法跳过的。

★ 投一个公司之前多问问自己为什么要投，目标一定要明确，不要学我做个无头苍蝇到处乱撞，然后撞的头破血流。

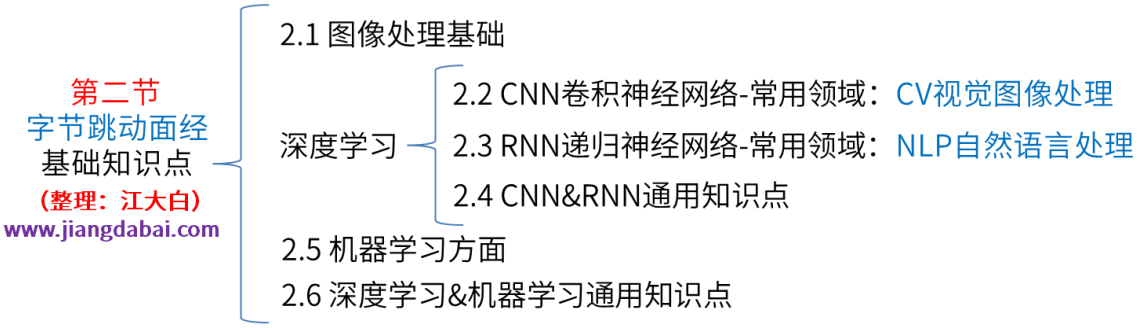
★ 虽然字节疯狂的招人，但是也请看好招人的部门，并不是每个部分都疯狂招人的，所以还得好好甄别，多看看，别头铁第一天出来就投，浪费机会。

★ 一面考核的是思想层面，比如考核面对大数据如何增量训练，面对多指标如何多任务学习，从离线实验到部署上线的流程，线上测试需要关注的信息，模型选择的依据等。

二面考核的主要是宽度，几乎都是横向的问，很少纵向的深挖，而我恰恰缺乏宽度的积累，所以面试过程还是有点难受。

三面面试官的考核主要是深度，一直深挖每一个细节，基本上问到口述代码的程度.不过所有代码都是自己写的就问题不大。

2 字节跳动面经涉及基础知识点



2.1 图像处理基础

2.1.1 讲解相关原理

- 是否了解图像降噪的一些方法?
- 了解常用图像增强的一些方法吗?
- 是否了解各种边缘检测算子?

介绍下 Sobel 算子，sobel 核的参数由-1->2,改变后会发生什么?

面试官的意思：比如实现不同的功能效果，高斯模糊、腐蚀、膨胀、锐化等

- 了解 Hog 吗？讲解下 Hog 特征的原理，步骤流程是什么?
- 知道图像里面的插值算法有哪些？（三次样条和线性插值），用过什么图像的库函数?
- 解释下 Raw 图像和 rgb 图像的区别？了解其他色彩空间格式吗？或者饱和度、亮度这些吗?

2.1.2 手写算法代码

- 手写马赛克算法
- 手写高斯滤波算法
- 手写均值滤波及优化
- 手写下中值滤波器

2.2 深度学习：CNN 卷积神经网络方面

2.2.1 讲解相关原理

2.2.1.1 卷积方面

- depthwise 卷积
- 1*1 的卷积核有什么用?
- 反卷积相比其他上采样层（pixelshuffle）的缺点，棋盘格现象怎么产生的?
- 3D 卷积和 2D 卷积的区别，主要存在问题，如何加速运算，视频理解的 sota 方法，还有什么方向可以改进?
- 卷积核大小如何选取
- 卷积层减少参数的方法？使用 1*3,3*1 代替 3*3 的原理是什么?
- 设计一个在 CNN 卷积核上做 dropout 的方式
- 反卷积/转置卷积的实现原理?
- Dropout 的原理?
- 直接转置卷积和先上采样再卷积的区别?

2.2.1.2 池化方面

- maxPooling 怎么传递导数?
- CNN 里面池化的作用
- 反向传播的时候怎么传递 pooling 的导数

- 卷积神经网络在 maxpooling 处怎么反向传播误差

2.2.1.3 网络结构方面

- shufflenet 的结构
- 深度网络 Attention 是怎么加?
- ResNet 的结构特点以及解决的问题是什么?
- 图神经网络的理解，讲了发展史，应该从基于图谱和基于空间来讲
- unet 结构，为什么要下采样，上采样?
- ResNet V1 到 V2 的改进有了解吗?

  那 ResNet 的下采样过程是怎么样的?

  讲了 res-block 的跳跃连接，以及连接前后的 shape 保持（通过 padding 保持 shape 不变）
- fpn 的结构
- roi pooling 和 roi align 的区别
- Resnet 的理解、和全连接相比有什么区别?
- 简单说一下 Alexnet、Vgg、Resnet、Densenet、和 GoogleNet，它们的特色是什么?
- 问了很多轻量级网络，mobileNet v1 v2，shuffleNet v2，Xception，denseNet 等等

2.2.1.4 其他方面

- 有上过神经网络的课程吗，是自学的吗？了解感受野吗？怎么计算感受野？怎么增加感受野？（增加感受野和网络深度，压缩图像尺寸）
- 为什么卷积神经网络适用于图像和视频，还能用于其他领域吗?
- CNN 反向传播细节，怎么过全连接层、池化层、卷积层?
- CNN 里面能自然起到防止过拟合的办法
- CNN 中感受野/权值共享是什么意思?
- BN 层的作用，为什么有这个作用？测试和训练时有什么不同，在测试时怎么使用?
- BN 层做预测的时候，方差均值怎么算，online learning 的时候怎么算?

- BN 机制，BN 怎么训练；
- 发生梯度消失，梯度爆炸问题的原因？如果发生梯度爆炸、梯度消失，怎么解决？
- 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？

2.2.2 数学计算

- 如何计算卷积的复杂度、卷积层的参数量
- 计算 Feature Map 的 size
- 输入为 L*L，卷积核为 k*k，还有步长 s 和 padding p，求输出尺寸？（L1 = (L-k+2*p)/s + 1）

接上题，求操作的 FLOPs？（FLOPS = k*k*c1*c2*L1*L1）

- 在同时考虑 pooling，stride，padding 的情况下，计算 depthwise conv 和 pointwise conv 过程中每一步的计算量和 feature map 的尺寸
- CNN 中给定输入数据维度[c,w,h]，卷积核[k,k]，则输出维度，如何 padding=p，输出维度是什么？

2.2.3 公式推导

- BP 神经网络反向传播推导
- max pooling 梯度求导？

2.2.4 手写算法代码

- 说一些卷积、用代码实现卷积，并再改成有通道的三维卷积
- 写一个单通道的图像卷积（带 padding）
- 手写前向传播、反向传播代码
- 面试官轻描淡写地说，BP 你会吧，写一下吧，正向传播、反向传播都推了一个遍，交给面试官看了一眼，说用代码实现一下吧，用 numpy 写了一个单层神经元的反向传播，给面试官看了，问他还用不用写完整的传播过程，他说不用了。
- 让写代码或者数学公式展示 BN 的内部实现，为什么要用 GN，你知道 GN，BN，LN 和 IN 的区别吗？（这里 BN 内部实现回答错了，还好面试官非常 nice，一直知指导，最后给我讲了

BN 的内部实现，豁然开朗，回来看了一下代码，有了进一步的认识)。

- 用代码展示 shuffleNet v2 的结构
- 实现一维数组的 maxpool

2.2.5 激活函数类

- 说一下 Softmax 多分类器的作用？和二分类相比有什么特点?
- Softmax 的计算公式写一下,并进行解释
- Softmax 的 Loss function、写一下损失函数
- 写一个 Softmax 实现，注意上下溢出问题
- Softmax 在数值计算上可能会出现的上溢和下溢的问题

2.3 深度学习：RNN 递归神经网络方面

2.3.1 讲解相关原理

- 讲一下隐马、CRF、RNN、LSTM 的区别?
- RNN 为什么会出现梯度消失?
- BPTT 的推导?
- LSTM 和 GRU 和传统 RNN 的对比?
- LSTM 减弱梯度消失的原理，项目里用了 LSTM，问了 LSTM 的结构，三个门的作用，每个门用什么激活函数?
- LSTM 的输入，输出，遗忘门分别是做什么的，整个计算流程怎么样
- RNN 梯度弥散和爆炸的原因，lstm 为什么不会这样
- RNN/LSTM 解释，你知道哪些时间序列预测，举一个例子，写出伪代码(写了 HMM)。
- RNN 如何防止梯度爆炸（LSTM 原理)。
- LSTM 和 RNN 的区别， 遗忘门的具体实现?
- BN 和 LN 的区别，以及 BN 一般怎么用，LSTM 中有没有用 BN?

2.3.2 手绘网络原理

- 手画 gru，并解释门的原理
- 写一下 LSTM 的公式?

2.4 深度学习：CNN&RNN 通用知识点

2.4.1 基础知识点

- 详解梯度消失、爆炸原因及其解决方法
- 你用过 dropout 么？介绍一下 ,Dropout 的作用 ？
- 梯度消失的表现是什么，该怎么处理
- 神经网络权重怎么初始化，说一下自己知道的方法
- dropout 机制，为什么 dropout 能够抑制过拟合?
- 神经网络中网络权重 W 初始化为 0 有什么问题？为什么不能初始权重为 0?
- 如何解决模型不收敛问题 以及如何加快模型的训练速度
- 你知道哪几种 normlize 的方法？请着重介绍一种（BatchNormalization）.这个方法在深度学习网络中有什么用？为什么可以加速模型收敛?
- Attention 怎么做，self-attention 怎么做？self-attention 原理公式，为什么有效?
- Encoder-Decoder 模型里，如果 Decoder 是基于 Attention 做的，该怎么做，是一个什么结构?
- attention 机制是什么解释一下，啥是 soft attention 和 hard attention?

2.4.2 模型评价

- 有哪些评价指标？-比如 ROC、AUC、F1-Score
- 解释下深度学习中的评价指标：Map、PR 曲线、AUC、Recall?
- AUC 怎么计算？它刻画的是什么?实现求 AUC 的过程？（输入就是 instance 的 score 和对应 label）
- 给你 M 个正样本，N 个负样本，以及他们的预测值 P，求 AUC。（写完之后接问：AUC 究

竟在衡量模型什么能力？如果现在所有预测值都*1.2，AUC 是否会变化？）这一题印象深刻是因为平时在计算 auc 的时候，很多同学都知道是 roc 曲线的面积，但是对 auc 具体的含义了解不多。

- ROC 曲线的含义和其他评价指标的区别？
- 分类问题的指标是什么？准确度、召回率、PR 曲线
- 相关系数是怎么计算的？讲一下协方差和它的意义？
- 做视频用的是何种评价指标？
- 计算广告中 CPM、CPC、ROI 的含义，计算方式

2.5 传统机器学习方面

2.5.1 讲解相关原理

2.5.1.1 数据准备

- 采样一般有哪些方法？讲一下
- 解释下 MCMC 采样？

2.5.1.2 特征工程

① 特征降维

- 看项目中有数据降维的项目，讲一下 PCA 原理？PCA 与 SVD 的联系与区别？SVD 分解是怎么回事？
- PCA 了解吗？怎么推导？SVD 怎么求？
- 简单说一下 LDA 的思想？并说一下公式
- T-SNE 算法了解吗？

② 特征选择

- 特征选择有哪些方法？什么是特征向量与特征值？怎么理解它们代表的意义。（介绍项目时涉及到特征相关性分析，所以问了这个）
- 介绍下特征选择的 Lasso 回归？

- 特征选择里提到的互信息选择，互信息的计算公式是什么？
- 树模型中分叉的判断有哪些：信息增益，信息增益比，Gini 系数；他们有什么区别？
- 写出信息增益的表达式
- 在做特征工程时采用了哪些方法呢？ 常见的筛选特征的方法有哪些？

2.5.1.3 有监督学习-分类和回归方面

① 分类回归树（集成学习）

- 机器学习的集成方法有哪些？
- Boosting 与 Bagging 的原理以及异同点?为什么说 bagging 降低方差 boosting 降低偏差？谁是更关注方差 ，谁是更关注偏差？
- bagging 中随机有放回采样，假如一共有 N 个样本 采样了 N 次，得到 N 个采样数据，去重后有 X 个数据 求 E（X），我只列出了暴力计算的方法。
- 决策树，熵的公式、如何分裂，如何剪枝，回归树、分类树的做法

A.基于 bagging：随机森林

- 随机森林的随机性怎么体现？
- 为什么 bagging 能降低方差？

B.基于 boosting：Adaboost、GDBT、XGBoost

- xgboost 与 lgbm 的原理讲一下？是如何进行调参的？
- xgboost和gbdt的优势？两者的区别？并行怎么做？他们的应用场景有哪些呢？其他模型的能说一下吗？
- Xgboost 和 GBDT 的区别 以及如何改进和提升 Xgboost 模型
- Random Forest 和 GDBT、 XGBoost 、LR 有什么区别？
- GBDT 的原理，怎么做多分类问题？
- gbdt 的 gb 是什么意思，如何体现。gbdt 里如何知道每个特征的重要性。
- 因为实习用到了 xgb，让写一下 xgb 的 loss func？问 xgb 到底是怎么预测的？

- 问 GBDT 原理，然后具体问了下每个叶子节点是怎么分裂的，用什么标准决定最优特征，答曰和 CART 一样，用 Gini 指数，然后写了下 Gini 指数的公式。老板看起来不甚满意，估计他本来想让我写的是 xgboost 那种带正则项的节点分裂方式吧。
- CART 了解吗？怎么做回归和分类的？ 哈希表了解吗？有哪些解决冲突方法？ 堆空间栈空间了解吗？
- CART 树的原理，和 ID3 以及 C4.5 有什么区别，回归树与分类树有什么区别。
- GBDT 中 G 是什么？怎么拟合树的？梯度拟合了怎么和原来的树合并的
- 为什么 XGBOOST 在大赛上表现很好/与 GBDT 相比优势
- lightgbm GBDT xgb，问的超级细，可能持续了 7 8 分钟，XGB 残差怎么用一次和二次梯度求，分裂点怎么求，思想原理是什么。XGB 实际使用中重要的超参数，你们比赛中用的目标函数是什么，为什么 lightgbm 速度更快，其并行计算如何实现？
- xgboost 的特征重要性怎么计算的？设计能适应测试集里有缺失值的训练集没有的 GBDT， 要求不能从填充数据的角度来做？
- LightGBM 和 xgboost 的区别 ，LightGBM 的直方图排序后会比 xgboost 的效果差吗，为什么？
- xgb 怎么并行运算（除了自带的并行找特征分裂点，还说了一般模型的按数据和按特征并行） ，但是面试官一直追问详细的并行方法
- xgb 与 LR 各自的优缺点，LR 为什么更容易并行？

② 线性回归

- 能否详细的讲解一下，线性回归的原理？具体讲解一下线性回归的底层原理，比如说如何训练，如何得到参数，如何调整参数等？
- 线性回归 R^2 公式及意义

③ K 近邻（KNN）

- knn 算法了解吗，和传统的 LR 和 SVM 有什么区别？
- 怎么优化 knn 呢？

④ 逻辑回归 LR

- LR 为什么要用 sigmoid？（经过面试官提示，是来自于最大熵模型，建议不明白的同学去查一下，下次面试给面试官露一手）
- 逻辑回归特征之间关联程度大会有什么问题？
- 讲解一下逻辑回归的原理？再详细的讲解一下朴素贝叶斯的底层原理，比如说，如何选参数，如何训练模型，如何做分类？
- 对于 LR 来说，LR 如果多了一维冗余特征，其权重和 AUC 会怎样变化（权重变为 1/2, AUC 不会变化）
- 逻辑斯蒂回归里面，输出的那个 0-1 之间的值，是概率值吗？你看它又叫对数几率回归，怎么理解几率这个概念？
- 什么是线性模型？项目为什么使用 LR，介绍 LR？LR 为什么是线性模型？ 如何提升 LR 的模型性能？
- FM 与 LR 对比一下，FM 是否也能起到自动特征选择的作用，为什么？
- LR 和 FM 的区别？
- LR 的 w 可不可能是负的，正负样本 10：1 的情况下？
- LR 一个特征重复会怎么样？

⑤ SVM（支持向量机）

- 为什么 svm 的 loss 不能直接用梯度下降要用对偶？说说你知道的优化算法。
- SVM 原理，与感知机的区别？还问了 SVM 如果不用对偶怎么做？
- SVM 对于异常值的处理，敏感程度？
- Svm 和 LR 的区别和各自优缺点？
- SVM 最后的形表达形式是什么？
- KKT 条件是什么？在 SVM 中起到什么样的作用；
- SVM 中 SMO 具体的操作以及原理。
- 熟悉什么机器学习算法（SVM），写损失函数（hinge+正则）

- SVM 原理？为什么能转化为对偶问题？能不能推导
- SVM 怎么解决不容易找到超平面的问题？
- SVM 有哪些核函数？

⑥ 朴素贝叶斯（Naive Bayes）

- 贝叶斯模型知道吗？问贝叶斯网络的原理，贝叶斯估计和极大似然估计原理？朴素贝叶斯公式？
- 讲一下最大似然的原理？
- 朴素贝叶斯的算法实现？

⑦ 决策树（DT）

- 了解其他机器学习模型吗，说了决策树，为什么用信息熵？
- 决策树的 ID3 和 C4.5 介绍一下？决策树模型的类别？
- 决策树分裂节点的标准与对应的算法
- 代码写一个决策树，给定数据，启发函数是信息增益，假设所有特征的值都是数值类型的：定义节点类、构建节点、选取当前节点的最优划分特征（计算所有特征的信息增益）、数据划分、构建子节点、考虑停止划分的条件。花了好长时间写了个代码框架，然后和面试官讲了思路。
- 写的决策树是几叉树？暂时考虑的是有多少种不同的取值就有多少个分支，意识到这肯定是不对的，优化的话可以将所有取值进行划分，比如二划分就可以改成二叉树。

2.5.1.4 无监督学习-聚类方面

① Kmeans 均值聚类

- KMeans 和 GMM 联系与区别，kmeans 原理，怎么做的，你是怎么并行的？
- kmeans 原理，怎么做的，你是怎么并行的？k-means 是否一定收敛？

② 高斯混合模型（GMM）

- 高斯混合模型和 K-means 的区别和联系

2.5.1.5 模型评价

- 信息检索中为什么使用 Recall 和 Precision？

- 机器学习中一般怎么衡量模型效果？AUC 值怎么理解？
- 怎么衡量两个分布的差异？KL 散度和交叉熵损失有什么不同？关系是啥？
- 一些统计学的原理比如 t-test, AUC curve 的意义是啥，为什么要用 AUC 去衡量机器学习模型的好坏。
- 为什么召回的数量级小，排序模型的效果就好
- 交叉熵 和 相对熵(kl 散度)的关系

2.5.2 手推算法及代码

2.5.2.1 手推公式

- LR 推导，手写 LR 前向传播和反向传播
- 写出 LR 的损失函数（交叉熵损失函数）
- FM 的推导?
- 上来就让手写个 LogisticRegression 你了解元学习吗？说一下你的理解。 权值初始化方式对 LR 的收敛有影响吗？ 你对权值初始化有什么了解？怎样才算是好的初始化?
- 问了 xgboost 并且手推
- 问能推哪些算法的公式，只敢说 LR，因为没准备 SVM。然后就是推 LR 的梯度下降，接着让我写 sigmoid 函数，最后就是 sigmoid 求导。
- LR、SVM 的公式推导
- 朴素贝叶斯写公式。
- 介绍一个熟悉的算法（LR)，推导 sigmoid 求导过程
- 手写 LR 的实现过程，然后聊了聊 L1 以及 L1 的扩展
- 手推 LR 的损失函数、损失函数怎么来的、梯度如何计算，写成一个完整的类

2.5.2.2 手写代码

- 问了 adaboost 的原理，模型的权重以及数据的权重各自有什么意义，写出 adaboost 的伪代码。

- 手写 kmeans 聚类算法（代码）

2.6 深度学习&机器学习面经通用知识点

2.6.1 损失函数方面

- 交叉熵，相对熵
- 回归和分类的常用损失函数
- Logistic Regression 损失函数，怎么来的?
- 常见损失函数有哪些
- 逻辑回归中损失函数的实际意义?
- Smooth l1 loss 公式以及为什么是这样的?
- 为什么分类问题用交叉熵，怎么来的?

2.6.2 激活函数方面

- 各种常用激活函数对比下？（sigmoid, tanh, relu, lrelu 等）
- sigmoid 的优缺点?
- sigmoid 和 relu 的区别?
- 平时用什么用的多？为什么用 relu 多呢?
- 写逻辑回归的 logloss 损失函数

2.6.3 网络优化梯度下降方面

- SGD 每步做什么，为什么能 online learning
- l1 是损失函数，有哪些优化方法，能用 sgd 么？为什么?
- Adam 优化器的迭代公式
- 4adam 用到二阶矩的原理是什么
- 几种梯度下降的方法和优缺点?
- 梯度下降系列算法有哪些，有点蒙住了，后来才想起来应该问问 momentum adam 之类的算

不算?

- 讲一下你熟悉的优化器，说一下区别或发展史
- 有哪些优化算法，Adam 的默认参数有哪些?
- 介绍方向导数和梯度，方向导数和梯度的关系？为什么梯度在机器学习中的优化方法中有效?
- 神经网络权重初始化方法和优化方法
- 介绍一下你了解的优化器和各自的优缺点?
- Adam 和 Adagrad 的区别?

2.6.4 正则化方面

- 正则化的本质?
- L1 正则化和 L2 正则化的区别，从数学角度说
- L1 有什么缺点？L2 呢？平时用 L1 多还是用 L2 多？为什么正则化选 L2 呢？为什么不选 L1？

L1 为什么产生稀疏解?

- L1、L2 的区别， L1 为什么图像是菱形
- L1 范数和 L2 范数的区别，作用。为什么 bias 不正则
- 为什么要用正则化？解释了奥卡姆剃刀
- L1 正则化与 L2 正则化的区别？解释了参数先验和拉格朗日乘子法

2.6.5 压缩&剪枝&量化&加速

- 了解模型蒸馏吗?
- 怎么做模型压缩？-使用知识蒸馏、设计小的网络，得到 End-to-End 模型。
- 介绍了量化（8bit,4bit,二值化）的项目?训练后量化和量化感知训练分别是怎么实现的?

2.6.6 过拟合&欠拟合方面

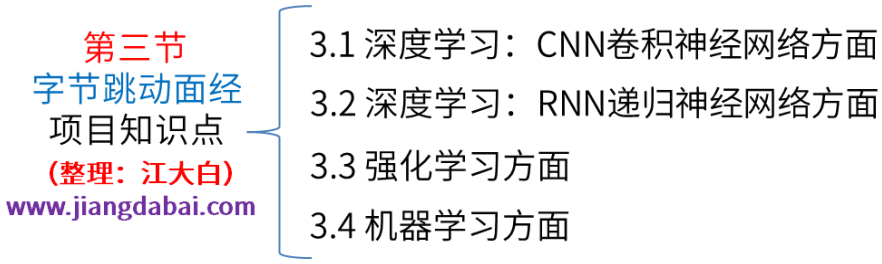
- 什么是过拟合，过拟合怎么解决?
- 深层网络容易过拟合还是浅层网络容易过拟合 ？

- 防止欠拟合的方法?
- 过拟合要怎么解决? （减少模型参数、早停、正则化、数据增强、GAN 合成数据、dropout、few shot learning，等等等等）
- BN 为什么防止过拟合呢?

2.6.7 其他方面

- 深度学习与机器学习的异同及联系?
- 数据不均衡的处理方法-过多数据欠采样，过少数据过采样，另外还有一些基于模型的方法比如 SMOTE 方法等。
- 样本不均衡怎么解决，我说人为对采样少的样本重复几次，然后他问这样 auc 会不会变并解释，我说不会变，解释的不太清楚但他好像听懂了
- 当模型的性能不好时，如何分析模型的瓶颈?
- 数据不均衡有什么解决方式，从数据，模型选择，以及损失函数选择角度?

3 字节跳动面经涉及项目知识点



3.1 深度学习-CNN 卷积神经网络方面

3.1.1 目标检测方面

3.1.1.1 讲解原理

- RoI Pooling 和 RoI Align， 怎么做插值，线性插值，spline 插值，写插值公式,这个问题二面和三面都被问到了
- 比如 ROI Align 相较于 ROI Pooling 有什么改进之类的

- detection 的发展，从 RCNN 到 CenterNet
- 着重讲 Faster RCNN，问的非常细, RPN 原理,。9 种 Anchor 怎么来的，为什么这样设计 Anchor。哪些为正类，哪些为负类。Loss 怎么设计的，tx，ty，tw，th。
- 说下 R-FCN 与 Faster RCNN 的区别
- 怎样理解 R-FCN 的 PS RoiPooling?
- RPN 中正负样本的阈值为（0.7、0.3），中间（0.3-0.7）的不选用会有什么后果？为什么需要把正负样本的阈值设定在这两个相隔较远的值?
- 一直在问 faster rcnn，faster rcnn RPN 流程，anchor 选的太大或太小有什么影响，正负样本选取的时候为什么用 0.7 和 0.3 的超参，与 nms 使用 0.3 的超参是否有关，smooth l1 损失为什么不用 l1，为什么不用 l2，w，h 回归的时候为什么要用 log，x，y 回归的时候为什么用除法等等。
- 讲一下 SSD 流程
- faster rcnn、yolo、ssd 的区别?
- ssd 各种变体
- Yolov2 v3 的提升：小物体检测上的提升
- Anchor free 检测算法了解吗，怎么回事
- 发展过来的前世今生，yolo 全套,ssd,faster rcnn 具体细节，代码实现，工程中需要考虑的实际问题
- anchor free 框架，基本思想，不同网络的具体对比，hourglass 结构的好处，损失函数，我自己的框架具体结构，和 sota 比性能如何（map 更高速度更快），新的损失函数为什么这么设计?
- 介绍下 cascade rcnn
- 目标检测框架，two-stage， one-stage；
- 对目标检测问了很大细节，包括 BN 的训练，ROI Pooling 和 Allign 的细节，FPN 网络的细节
- YOLOv3 和其他的目标检测有研究吗？为什么选 YOLOv3？这个和其他 Fast-RCNN 类型的有什么区别？SSD 和 YOLOv3 的区别?

- 问检测中能提升速度但不损失性能的操作有哪些？用过的没用过都行。
- 介绍一下 retinaface?
- 介绍了一下 Centernet?
- 单阶段的检测方法如 YOLO 为什么对负样本需求更大?
- Yolov3 中 anchor 尺寸的设置?
- 数据增强的常用方法，以及项目里用的数据增强，目标检测中的数据增强?
- 两阶段方法与一阶段方法的对比及其优缺点，focal loss 的表达式，anchor free 方法的优缺点?

3.1.1.2 损失函数

- 讲一下 Focal loss，公式是什么?它解决了一个什么东西？答：难易样本不平衡

再问:如何解决的,和难例挖掘 OHEM 有什么区别?

- Focal Loss 和 OHEM 的区别，Focal Loss 解决了正负样本不均衡吗?
- faster rcnn 损失函数的构成?

3.1.1.3 手写代码

- 手写 NMS（真的是非常简单的一个题了），我是把官方的 NMS 背下来然后写上去了
- 画出 faster rcnn 流程以及 RPN 的具体过程
- 实现 NMS 的全过程：包括按 score 排序（我当时选的归并），IOU 的计算，NMS 的操作（选出最后的框）
- 实现计算 IOU 的函数，扩展：当 bounding box 与坐标轴不平行时怎么计算，说出思路?

3.1.2 目标追踪

- 目标跟踪里匈牙利算法的流程?

3.1.3 图像分割

- 讲一下 unet 和 deeplabv2 的流程，顺便问了下 deeplabv3 是否了解?
- 介绍一下 Unet？为啥要这么设计，好处是什么?

- FPN 和 Unet 的上采样用了直接想加和 concat，有什么区别，从反向传播的角度来说说
- 基础知识，比如分割的网络有哪些，网络是如何优化的等等？
- 语义分割中 miou 计算公式？
- 了解哪些语义分割算法？

3.1.4 OCR

- 我有一个中文的 OCR 识别模型，现在如何得到一个日文的识别模型？

（这里我想说翻译，但是感觉 OCR 本身就有自己的误差，加上翻译的误差可能会导致准确率很低， 所以我说的是中文训练之后再做迁移。但是感觉他想听的是翻译模型。）

- OCR 可以在不同的场景下识别出文字，但是在我们做 AR 实景翻译的时候，由于场景很复杂，所以识别出来的文字送到翻译模型中的时候就无法确定位置关系

例如

离野

离火

原烧

上不

草尽

，，

一春

岁风

一吹

枯又

荣生

翻译的时候可能会翻译 离野 这种情况该怎么设计一个解决方案？

3.1.5 超分辨

- 自己做的超分辨项目有没有什么创新点
- 超分辨今年有什么改进，有没看过今年的超分辨 paper
- 超分辨率用的什么损失函数？（MSE, RMSE，感知损失等）

3.1.6 关键点检测

- 检测到的人脸如何对齐，warpaffine 参数，转换矩阵 M 有几维？

3.1.7 图像分类

- 分类网络训练样本有噪声(错误标注)怎么办
- 分类网络样本不均衡怎么办
- 分类网络想要分很细的类(比如阿拉斯加和哈士奇)怎么办
- 图像分类一般用什么损失函数？（回答交叉熵） 那说一下交叉熵的形式吧？可以写下来，讲一讲怎么来的？ 举了逻辑回归中的交叉熵损失，然后讲了公式变换以及对数似然等
- 如果数据集有 20%的噪声数据，会有什么影响？可以按照上面写的损失函数来想？
- 对图像分类网络的发展历程和进展有了解过吗？比如 resnet, inception 这些

3.1.8 姿态估计

- 姿态估计方向的一些知识，比如 openpose 的实现过程，PAF 的原理？
- 姿态估计除了 MSE 还能用什么 loss，还讲了些其他相关论文里用到的一些方法，面试官对相关方向很懂，问了很多问题？

3.1.9 目标重识别

- Reid 里 MGN 网络的设计，为什么有三条支路？将局部特征加入的会有特征冗余，为什么还要加入局部特征？

3.1.10 人脸识别

- FaceNet 里的 triplet loss 的公式，反向传播如何更新？

3.1.11 图形图像方面

- Phong 模型，如果能量不守恒了怎么办
- 渲染管线
- 迟渲染和前向渲染的区别，Tile-Based Rendering
- MRT
- FrameBuffer
- Z fighting
- 光线与三角形求交
- 你所理解的 PBR
- 抗锯齿技术(MSAA TAA 等等)
- OpenGL 中 Blend 方式
- 渲染半透明物体，次序无关的半透明(Depth Peeling 和 Per-Pixel Linked Lists

3.1.12 视频编解码

- VVC 都了解哪些部分?
- VVC affine AMVP 的参数获取流程（讲了构建候选列表，利用梯度计算偏移值做运动搜索，cost 比常规 amvp 小再做八个点的小范围搜索）
- 前面说的梯度计算偏移值，公式推导背景原理知道吗?
- 运动补偿的差值怎么做的知道吗?
- merge 列表构建过程（空域相邻、时域同位块、基于历史、成对平均、零 mv）
- DMVR、PROF 和 BDOF 这些了解多少?
- 编写代码：PSNR 计算函数
- 编写代码：十进制转十六进制函数

3.2 深度学习-RNN 递归神经网络方面

3.2.1 自然语言处理 NLP

3.2.1.1 讲解原理

① Bert

- Bert 的输入是什么？讲一些细节，你对 BERT 有什么可以改进的地方？
- 说一下 Bert 的嵌入层，然后就是各种关于 Bert 的细节问题？
- Elmo 和 Bert 的区别？ Bert 细节（多头和缩放）
- Bert 的原理和 Word2vec 的区别？
- 解释 Bert， Bert 好在哪里？
- ELMO 模型和 Bert 模型的比较，position embedding|sentence embedding|mask embedding 如何拼接的，position embedding 细节？

② Transformer

- 解释下 Transformer 结构？
- Transformer 的结构。bert 里的编码方式。
- Transformer 中为什么要除以根号 dk，为什么能加快收敛速度，为什么加了根号？
- 说一下 Transformer 中为什么 decoder 比 LSTM 慢？

③ CRF

- 阐述 CRF 的原理？
- CRF 与 HMM 关系，区别？
- 阐述 BiLSTM 的 BP 过程，为何 BiLSTM 后接一层 CRF 会有提升？CRF 层自己是怎么实现的

④ Word2vec

- 神经网络中的 word2vec 了解么？详细讲解一下它们的原理？包括两种训练方式及效率等
- W2V 的原理 ，两种生成方式，W2V 的思想到底是什么，为什么要这样做，W2V 的缺点，W2V 中所用的 softmax 比起普通 softmax 有何区别，为什么能减少计算量（我并不是搞自然

语言的，这一波问的我有点捉襟见肘，只是勉强回答了，面试官很好，我没回答清楚地就给我讲，引导我）

- word2vec 如何训练，hierarhical softmax 和 negative sampling（后面的没印象了）
- Word2vec 两种方式，怎么优化，负采样
- 除了 word2vec 的其他 embedding 方法
- word2vec 如何训练的，细节，权值矩阵如何训练
- word2vec 训练时如何加速
- word2vec 原理，如何得到词向量？如何分词，分词原理？
- 说一下 word2vec，为什么通过对单词预测可以学习到单词的 Embedding?

⑤ CNN 的应用

- CNN 在文本中的用法，pooling 的作用，有哪些 pooling?
- CNN，RNN，Tansformer 分别如何编码文本

⑥ 其他

- 词向量 onehot 的缺点 word2vec，glove，elmo，bert 区别
- 你能详细的说一下 CBOW 和 skip-garm 它们的区别么？分别适用于什么场景？
- 千万向量中如何找到和单个向量相似的那个
- 给你一段文字，如何提取文本特征

答：TF-IDF（解释了一下原理）；word2vec；one-hot；（其实 GLOVE ELMO 这些也可以讲一讲，不过一时间没想起来）

- 现在最火的 NLP 模型是啥（项目中用了 BERT， 然后上个月的 XLNet 目测效果更好）
- 聊了 textcnn，lstm attention
- TF-IDF 分别表示什么及公式?
- Fasttext 模型为什么比 word2vec 快，隐层怎么处理的?
- 知识图谱学习得到的 Graph Embedding 是用于召回还是排序（召回）（1.有噪声；2.因为对于传统观点的召回来说，精准并不是最重要的目标，找出和用户兴趣有一定程度相关性但是

又具备泛化性能的物品是召回侧的重点，所以可能知识图谱的模式更适合将知识图谱放在召回侧。）

3.2.1.2 损失函数类

- 讲一下 NLP 中常用的损失函数?
- 项目里有 CTCLoss，问了一下 CTC loss 有什么用，不用 CTC 的话怎么办?

3.3 强化学习

3.3.1 讲解原理

- 强化学习 DQN，DDQN，AC，DDPG 的区别
- 介绍一下什么是 GAN? GAN 用来干啥的,是怎么训练的?
- GAN 是怎么训练生成器、判决器的？（楼主就说了最原始的 GAN 的交替训练的方法）
- 因为项目中用到了强化学习 DDPG，介绍了 ddpg 的原理，训练细节等
- 介绍 GAN 中的生成器和判别器
- 通常 GAN 有不收敛、模式崩溃的问题。怎么让 GAN 更稳定更好。不收敛的原因分析：在最优解附近震荡，需要约束梯度。
- 让 GAN 稳定的 trick:

WGAN 的地球移动距离衡量数据分布差异

零中心梯度惩罚。权重的 L2 正则化

权重平滑移动（EMA)。

均衡学习率。权重归一化。

- 主要介绍了实习项目基于 GAN 做图像补全的项目：包括网络结构、损失函数、实际效果、指标
- 介绍强化学习的项目（背景、动机、如何建模、输入输出和训练算法说了一遍，说完后面试官问了一些细节）

3.3.2 损失函数

- 我实习中有用过 GAN 生成人脸，所以问我：GAN 的损失函数形式是什么样的？（楼主写出来了，不过忘了写前面的 min、max 目标，面试官表示没关系，知道我是懂的就行了）
- GAN 的损失函数形式是什么样的?

3.3.3 其他方面

- 强化学习 PG 的推导
- 新闻推荐如果用强化学习，怎么设计。
- 有木有试过 StarGAN 之外的方法?
- 文字生成可以用除了 GAN 其他的吗（说了说 RNN，但是感觉 NLP 这边很多自然语言模型都可以做文本生成啊，根据模型调整输出层？）
- 问在用 StarGAN 合成人脸表情的时候训练有没有遇到什么困难(额感觉面试官都好喜欢问 GAN 方向的问题)

3.4 机器学习方面

3.4.1 推荐系统

3.4.1.1 讲解原理

- xdeepfm 模型结构公式等
- user-cf、item-cf 公式，原理以及区别?
- embeding 的方法？FM FFM deep FM
- DeepFM 与 FM 的关联，并描述 DeepFM 的结构
- 问了下 fm，ffm，deepfm 的区别
- 讲述一下 FFM 和 FM，之后问了如何处理特征?
- FM 了解么，具体怎么做的，怎么解决权重系数难训练的问题，梯度怎么更新的 ？
- DeepFM 了解么，embedding 层是怎么训练的，结构是什么样的，比赛的 DeepFM 是自

己写的么（用的 DeepCTR） ？

● 做用户商品交互特征的时候，你知道业界是怎么做的？扯了一下 DIN 模型的和目标商品的 attention 做法?

● 排序阶段你知道业界是怎么做的？说了一下点击率模型：deepfm，nfm，wide & deep，dcn，deepcrossing?

3.4.1.2 手写代码

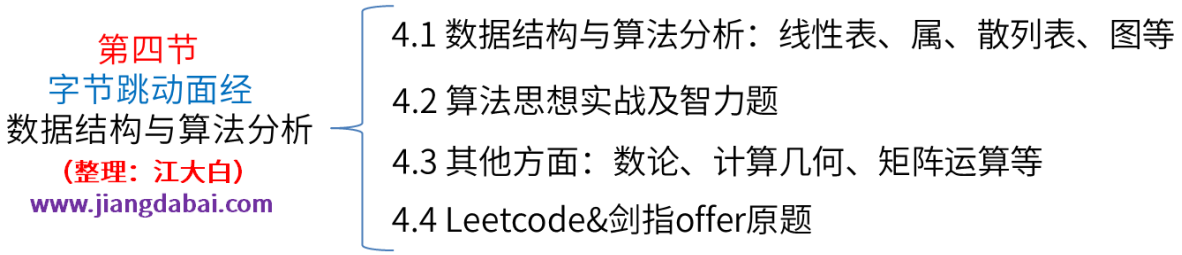
● 了解 Deep 模型吗？说了 deepFM，手写 deepFM 的结构（代码）

3.4.1.3 其他方面

● 推荐系统模型收敛的很好，但是多样性可能不好的情况下如何解决

● 搜索引擎算法：搜索引擎的流程是什么样的（不太会，只说了 query 分析，然后匹配 doc）

4 数据结构与算法分析相关知识点



4.1 数据结构与算法分析

4.1.1 线性表

4.1.1.1 数组

● 排序数组中绝对值不同的个数

● 手撕代码，1.给定一个数组，前面一部分已经排好序，后面一部分也排好序，将整个数组排序

● 数组前半部分和后半部分有序，然后排序(用的归并)

● S1, S2 两个整数数组，已经从大到小排序。 输出最大的 K 个数，时间复杂度：通过 k/2 的

思想，直到把 k 为到 1 为止。

● 给定一个数组，找出数组的最长连续子序列。例：3,3,4,7,5,6,8，最长的连续子序列（这里的连续是说连续整数，整个子序列是连续整数，我一开始题都没看明白）应该是(3,4,5,6)，需要返回它们的下标(1,2,4,5)。如果存在多种答案，只需给出任意一组下标。面试官看我不会，让我先写一个暴力的方法，我还是不会啊，然后一个小时过完了，凉凉。

● n*n 的二维数组，求最长上升序列，每个位置都可以上下左右走。

例如：

6，9，9

4，6，8

3，1，3

最长就是 1-3-4-6-8-9？解法：DFS

问复杂度是多少，我说 n 四次方

问怎么优化？用缓存保存 dfs 过的值，减少重复递归。

● 求无序数组的不连续递增子序列？

● 给定 2 个数组，一个数组是 PID ，里面存的进程号。一个数组是 PPID 代表着 PID 对应位置进程的父进程号。然后删除一个进程，问因为这个进程被删除，而牵连都被删除的进程号打印出来。

其实就是个简单的 DFS，建立一个 graph，然后 dfs 直接输出即可。

● 有一个长度为 n 的数组，求一个数 k，k 的取值区间为[1, n-1]，使得数组的前 k 个数和后 n-k 个数的方差和最小。

要求化简方差公式，达到计算子序列方差的时间复杂度为 O(n)。化简后要求空间复杂度为常数级别。

● 将一个数分成给定的一些数的组合，给出所有这样的组合。比如将 10 分成[1,2,3],其中一种为[1,1,1,1,1,1,1,1,1,1]

● 给你一个数组,[1 2 3 5 5 3]，每次可以溢出连续的重复数字，移除后得到的分数为连续数字的长度的平方，求让数组全为空时获得的最大分数

1 2 3 5 5 3 第一次移除 5 5 得到 1 2 3 3 ，得到分数是 2*2

第二次移除 3 3 ，得到 1 2 ，得到分数是 2*2

第三次第四次分别移除 1 和 2，最后得到的总分数为 2*2+2*2+1+1

第一反应是区间 dp,每次遇到连续的重复数字就 dp 求一次,取最大,但是没考虑 2 1 1 1 3 1 1 ，去除 3 后又变连续的情况，到最后也没调出来

- 给一个无序数组，找到其中位数（快排，O(n)），问时间复杂度?
- 给定无序的数组，求出连续相邻的子数组中最小值乘以长度，使得值最大的连续数组?
- 给一个无序数组，输出最小的不在数组中的正数
- 非递减数组中查询某个目标值出现个数。解法：二分查找左右边界。
- 扭转有序数组中查找，比如说[5,6,7,8,1,2,3,4]{ [5,6,7,8,1,2,3,4] }[5,6,7,8,1,2,3,4]这种，要求时间复杂度 O(lgN)O(lgN)O(lgN)

先看是否扭转，没有扭转用普通的二分做。否则用二分查找查找出扭转点， 然后看要查询的 keykeykey 在哪一边， 再二分查找就行。

- 数组中元素两两成对，除了一个元素只出现一次，找出该元素。异或有交换律吗?
- 数组中元素两两成对，除了两个个元素只出现一次，找出这两个元素。
- 两个有序数组的中位数?
- [2,3,4,6,2,3],找出数组中每个数字的后边比它大的第一个数?
- 输出数组中位数，第 k 个元素为前 k 个数的中位数。
- 连续数组，给定 k，求连续数组最小区间。
- 空间上如何优化：滚动数组
- 一维 01 数组中，求最长的区间，其中 0 和 1 数量相等。
- 一个非负数组，求出和为 m 的最长连续子序列的长度
- vector<vector<int> > x 里面，求 min(Σi=0n－1丨xi+1[k]－xi[m]丨)min(\sum _{i=0}^{n-1}{|x_{i+1}[k]-x_i[m]|})min(Σi=0n－1丨xi+1[k]－xi[m]丨),就是每个数组任选一个数字，相邻求差的绝对值，然后再求和求最小。

● 编程题：给定一个数组，数组长度为 n,数组所有元素 x 满足：0<x<=n<=1000。求数组中出现次数最多的元素，若多个元素出现次数相同，输出元素值较大的，要求时间复杂度 O(n),空间 O(1)。

● 给一个数组，求其所有数都平方后，共有多少个唯一的值。

● 数组的全排列（空间复杂度 O（1））

● 判断是否存在个数超过数组长度一半的数

● 给一个数组，长度是 N，里面的大小也是 0～N，用 O(n)的时间，O(1)的空间复杂度统计里面数字的个数？

● 旋转数组查找，一个二维 DP

● 旋转数组中搜索某个目标值

● 查找一个有序数组旋转后中有无 key 值。

● 单调不减数组找出一个数最后出现的位置（二分变形）

● 两个排好序的数组，找中位数。这个题如果复杂度 O(n)就没意义了，显然是要求写 O(log n)的二分查找

● 整数数组找两个相加为 k 的两个数，先写了扫一遍数组用 map 存扫描过的元素，她说你这个空间复杂度高，能不能换一个，后来又写个排序后用双指针的版本。

● 给定数组返回任意满足（当前数大于左右两个数），要求时间小于 O(n) ，开始想返回随机数，面试官提示小于 O(n)就是 logn 就是二分

● K 个有序数组，找一个长度最小的区间，在这个区间里至少包含每个数组各一个数。

分析：初始化带下为 K 的最小堆，K 个数字是每个数组中的最小值，设置变量 max 记录 k 个数字中的最大值，删除堆顶元素，将原堆顶元素对应的数组中下一个元素加入到堆中，调整堆，并且记录当前区间范围为（max-min），重复执行直到某个数组所有值都被删除。

● 问怎么实现一个字符串中找最小的包含所有不同字符的子串，回答用双指针，让证明双指针的正确性。

● 长度为 n 的数组中有一个数字出现了 n/2 次，快速找到这个数

- 输入一个二维数组和一个一维数组，例如

[1,2,3,4,5]

[1,2,3,4,6]

[6,1,2,3,7]

[0,0,5,3,2]

和 [1,2,3]

输出[1,2,3]在二维数组中的位置

(0,0)

(0,1)

(1,2)

直接遍历，然后代码行云流水的写出，然后问我时间复杂度。

- 给定两个 unordered 数组，数组中每个元素都包含一个 int 和一个 bool,bool 表示这个 int 数值是否应该被 delete，每个元素的值可能会出现多次，和不同的状态，将 2 个数组合并成一个

要求：merged 后返回一个数组，每个元素智能出现一次

不能包含曾经被 delete 过的元素，这道题我用 hashmap 做的，做出来分析时间复杂度

- 从 n 个数字的数组中任取 m 个为一个组合，返回所有组合，顺序不一样的算一个组合（递归遍历+回溯）

- 给出一个数组，数组中有正数和负数，要求重新排列这个数组，使得原始数组中的正负数交替排列，且保证各自的相对顺序不变。

- 合并 n 个有序数组

- 二维数组逆时针螺旋打印

比如输入

1234

5678

9abc

输出

159abc843267

- 最大数组连乘值
- 给定一个数组，找到这个数组中，和等于 0 的所有三元组。 我说我只能想到 n^2logn 复杂度的，在面试官提醒下写了，依然没提交，大概思路对就过了
- 一个正整数数组，寻找连续区间使得和等于 target，简单的用两个指针做出来了，不过让我证明一下解法的正确性，纠结了一会儿也算是证明出来了。然后如果里面有负数怎么做?
- 数组连续元素最大值的和，动态规划解决
- 求和为 k 的子数组个数
- 两个有序数组求交集
- 两个数组自身元素一样，各自内部的元素不能比较，实现两个数组的排序
- 给了一个字符数组，求这些数组的组合，例如{a,b,c}的答案是{a,b,c,ab,ac,bc,abc}
- 一个数组，一个数出现一次，其他数出现两次，求出现一次的那个数。
- 一个数组，两个数出现一次，其他数出现两次，求那两个数。
- O(n)时间复杂度和 O(1)空间复杂度删除重复元素，数组有序，输入 a=[1,1,1,2,5,6,6]，输出 a=[1,2,5,6]
- 双指针：比较 i，j，a[i]==a[j] ，j++；不等就交换 i+1 和 j，然后 j++，最后返回有效长度。
- 长度为 n 的字符串中包含 m 个不同的字符，找出包含这 m 个不同字符的最小子串。
- 如果用数组实现，数组初始容量为 n，每次 push 到容量上限之后都扩容到原来的两倍，现在 push 进去 m 个数，m 远大于 n，求相比于 m 的时间复杂度
- 无序数组 Top K，时间复杂度，给出一个最坏复杂度的样例
- 给定一个数组，返回每个对应位置右边第一个比他大的数，没有就是-1，如【4，1，2，5，8】返回【5，2，5，8，-1】

4.1.1.2 链表

- 递归和迭代的反转链表
- 链表反转，二叉树中序遍历递归+非递归
- 二叉树转双向链表（中序非递归遍历修改指针）
- 反转链表中偶数位置的值，例如 1-2-3-4-5-6-7 变为 1-6-3-4-5-2-7
- 链表判断是否有环
- 找一个链表中的环
- 链表找环并证明?
- 删除链表 A 中出现在链表 B 的元素
- 删除链表重复节点。如 1-2-3-3-5-5-6 变为 1-2-6
- 删除倒数第 k 个链表节点?
- 如何用链表实现一个栈，O(1)获取最小值，get_min、如何节省空间，存放最小值，如果有多个，不想多次存放
- 有两个数字，用链表表示，如第一个数 123 的链表是 1-2-3，第二个数 4123 的链表表示是 4-1-2-3，输出两个数求和后的链表表示。
- 求两个链表的第一个公共结点。
- 如何判断两个单链表是否相交并找到交点，这个题没反应过来，所以答得并不好，只大概给了个思路。用两个队列实现栈的入栈和出栈操作。
- 链表的冒泡排序
- k 个有序链表合并，问时间复杂度
- 一个链表，奇数下标递增，偶数下标递减，排序使其总体递增。
- 已知单链表，要求奇数位置降序，偶数位置升序，说下思路，再写代码?

4.1.1.3 栈

- 栈和堆的原理和区别?

4.1.1.4 队列

- 两个栈模拟一个队列
- 给定栈，保证栈的效率的同时能够在 O(1)返回栈的最大值

4.1.1.5 字符串

- 将一个字符串通过 增、删、改 三种操作得到另一个字符串的最少操作，我当时选的动态规划。
- 有两个字符串，你只可以进行删除操作，问最少进行多少次操作可以使两个字符串相等？

例：sea，eat 需要两次删除操作

A：这个简单，思路就是用动态规划求两个字符串的最大公共字串的长度。然后使用每一个字符串的长度减去公共子字符串的长度。

Q：那咱们再加一点，如果我想要知道每个字符串需要删除的字符是那些呢？

A：那我们就需要求出最大公共字串具体是由什么字符构成的，思路也是动态规划。(很快就写完了)

Q：好的，那你有什么想要问我的么？

- 字符串最小编辑路径
- 字符串转数字，及边界条件
- 删除字符串中连续的重复字符
- isUniqueChar 一个字符串内是否有重复字符
- 1，2，...，N 中，字符 1 出现的次数？
- 给定字符串，求最大不重复子串长度
- 最长公共子序列
- 两个字符串，求最短编辑次数使相等
- 给一个字符串，列出所有可能的 ip 组合
- 给出字符串 x，和字符集合 y，求 x 中包含所有 y 中元素的最短字串？
- 我们输入两个值 n 和 k，n 表示我们有从 1 到 n 个整数，然后将这些整数都字符串化之后

按字典排序，找出其中第 K 大的。例如:n=15,k=5.那么 1-15 字符串化之后排序如下:1,10,11,12,13,14,15,2,3,4,5,6,7,8,9。其中第 5 大的就为 13。

- 给定一个字符串 S[0...N-1],要求把 S 的前 K 个字符移动到 S 的尾部，比如字符串"abcdef"，前面两个字符 'a' 'b'移动到字符串的尾部，得到新字符串"cdefab"，即字符串循环左移 K。要求：时间复杂度 O(n)，空间复杂度。
- 给定 random7 返回 random10
- 用 [1,5][1,5][1,5]的随机数生成器生成 [1,7][1, 7][1,7].

我给出了一种非常 naiive 的做法。就是调用两次随机数生成器，变成 1 - 25 的数。1-21 每 3 个对应 1-7 里面的一个数， 剩下 4 个重来。

- [9,91,87]，变成字符串组合成最大数?
- 判断输入字符串开始或者结尾是否包含非法字符串(前缀树)
- 给定一个 n*n 的字符盘，和一个字符串，看改字符串是否出现在字符盘中?
- 有一个字符串，判断 QWER 出现次数是否相同？如果次数不相同，如何修改可以让他们相同。
- 设计一个可动态扩展的字符串类，尽可能降低占用空间和时间复杂度。
- 字符串的全排列，问了时间复杂度（O（n*n!)），以及详细的时间复杂度推导（n 是怎么来的，n！是怎么来的)，怎么优化（DFS 剪枝)。

4.1.1.6 列表

- 列表数字排列可组成的最大数字?

4.1.2 树

4.1.2.1 二叉树

- 什么是二叉树?
- 构建哈夫曼树：本身不了解哈夫曼树，但是了解哈夫曼编码的一些思想，讲出来后，小哥哥引导着思路，然后我写出了代码。代码本身还有优化空间，但是小哥哥也说通过了。
- 二叉树子路径和为 k 的路径个数

- 给一个类似树的结构，每个节点都可以有多个节点（不止两个树）然后每个根节点和字节点间的路径不一样，求叶子结点到叶子结点的最大路径？
- 两个树节点的最近公共祖先节点？时间复杂度？
- 给定一颗二叉数，每两个结点路径为 1，求相隔最远的两个结点的距离
- 蛇形打印二叉树？
- 给出二叉树的层次遍历和先序遍历，求二叉树的后序遍历
- 求两棵二叉树最大公共子结构的节点数目
- 红黑树描述？

答：节点是红色或者黑色的，红色节点的子节点必须是黑色的，根节点为黑色，叶子节点（Nil 节点）为黑色的，根节点到叶子节点的路径上的黑色节点一样多。

- 加问为什么要使用不同颜色的点？

答：红黑树是平衡二叉树的变形，由于平衡二叉树插入删除操作复杂特别是如果插入有序的数字时，二红黑树只要满足节点的颜色要求，在插入删除过程中满足红黑树定义要求，就能满足二叉树的相对平衡。

- 红黑树和 AVL 的区别？
- 二叉树最大深度，地图中找出大陆的个数（一道 BFS 题）
- 二叉树最远节点距离
- 给定一个二叉树，求出这个二叉树的宽度和高度
- 什么是平衡二叉树？平衡二叉树的应用都有哪些？
- 判断是否是二叉平衡树？
- 二叉树前序中序遍历，重建二叉树
- 非递归中序遍历二叉树
- 二叉搜索树已知先序求后序(代码实现)
- 二叉树层次遍历
- 二叉树之字型遍历，每行打印

- 打印二叉树中最左边节点
- 一个二叉树的所有右叶子结点之和
- 判断给定序列是否为二叉搜索树的前序遍历
- 给 N 个数字，返回这 N 个数字能组成的所有二叉搜索树
- 二叉树输出给定节点到目标节点的路径，寻找两个字符串中只有首尾字符相同的所有子串？例如 ABCDE 和 ADCAE 中包含（ABC--ADC）以及（CDE--CAE）
- 给一个二叉树，输出所有完全一样的（重复的）子树。 要求 O（N）（code）
- 设计一个数据库的表来存储树形结构（不一定是二叉树），要求（1）输入父节点返回所有子节点（2）输入某节点返回所有兄弟节点
- 判断二叉树上是否存在一条从根结点到叶结点的路径，满足其上的元素之和等于 target。
- 给定 m 个不重复的字符 [a, b, c, d ...]，以及一个长度为 n 的字符串 tbcacbdata，

问能否在这个字符串中找到最长的连续子串，使得这个子串由上面 m 个字符组成。

return: 子串和起始位置

4.1.2.2 多路查找树

- 给你一个二叉查找树，还有一个数 K。如果能找到，就返回节点，如果找不到，就返回空。

4.1.2.3 堆

- 最小 K 堆，只写伪代码就行
- 找出一颗完全二叉树最后一个节点，时间复杂度要求 logN 的平方
- 有 M 个有序链表（从大到小）。现在我们要取出前 K 大的元素。

A:（这个我见过，内心美滋滋）我们应该把 M 个链表的头节点做成一个大小为 M 的最大堆，每次取出堆中最大的节点，然后将这个节点的后序节点放进来，重新对堆进行排序。

Q：好的，那这个算法的时间复杂度和空间复杂度是多少呢

A：时间复杂度，每次需要 O(logm)O(log^{m})O(logm) ，需要 k 次，那么总的时间复杂度为 O(klogm)O(klog^{m})O(klogm) 。空间复杂度为 O(m)O(m)O(m)

Q：那建立这个堆的时候时间复杂度是多少？

A： O(mlogm)O(mlog^{m})O(mlogm) ，那总的时间复杂度应该为 O((k+m)logm) O((k+m)log^{m})O((k+m)logm) 。

4.1.2.4 其他

- 线段树和树状数组的异同?
- 树的路径和

4.1.3 图

4.1.3.1 拓扑排序

- 写一下拓扑排序

4.1.3.2 最短路径

- 无向无环图中，最短路径的最大值（O(n^3)的解法)，这里考察的其实就是 Floyd 算法。

4.1.4 排序

- 说几个常用排序算法的时间复杂度、空间复杂度、稳定性?
- 说一下所有的排序方法，并给出他们的时间复杂度?
- 为什么归并排序、快速排序和堆排序都是 o（nlogn）的时间复杂度，大家都习惯用快速排序，归并排序和堆排序差在哪?
- 希尔排序知道吗？为什么这么操作?

答：改良的插入排序。插入排序在数组基本有序的时候可大大降低时间复杂度，希尔排序通过将数组分块后对每块数组进行插入排序，每次排序完成，块数减少一倍，数组也相对变得有序，知道最后对整个数组进行插入排序，则排序完成。

- 写一个归并排序
- 快速排序和归并排序描述一下，优缺点（描述）

答：快速排序：先选定一个基准元素，按照这个基准元素将数组划分，再在被划分的数组上重复上过程，最后可以得到排序结果。

归并排序：将数组不断细分成最小的单位，然后每个单位分别排序，排序完以后合并，重复这

个过程就得到了排序结果 。

优缺点：归并排序稳定且最高最低时间复杂度都是 nlogn，但是占用额外空间；不稳定，最高时间复杂度 n2，最低时间复杂度 nlgn，不占用额外空间。

- 快速排序很多重复数字如何优化

答：返回基准元素位置时返回基准元素的最左和最有索引，减少排序次数。

- input：无序的实数数组

output：求大小相邻两个数之间的最大差?

排序可以 o(nlogn)解决，然后问我怎么优化?

- 手写快排,求 TOPK
- 找第 K 大的数（快排）
- 一亿个浮点数，大小不超过 2^32，均匀分布在值域内，求最快的排序方法；分析排序方法的复杂度。
- 有两个数列，将两个数列排序，但是自己数列里面的数字不能和自己数列里面的相比较（快速排序变种）
- 前序中序，求后序
- 手写堆排序
- 海量数据 TopK 问题。一般这种问题都是用哈希表分治+堆排序，但是当时不会，所以挂了。
- 写一下桶排序

4.1.5 搜索

4.1.5.1 深度优先搜索（DFS）

- 求矩阵中连通域的个数，用 bfs 或 dfs 做，很简单。先说思路，然后用自己喜欢的语言做
- 二维数组找最长递增路径。很简单的题，DFS 即可

4.1.5.2 广度优先搜索（BFS）

- 寻找迷宫中的最短路径，迷宫中 1 表示有墙，路不通，0 表示可以走。我脑子不知道怎么

抽了，直接想用 DFS 来解，给面试官讲了一下思路。面试官提醒我，DFS 和 BFS 你是怎么考虑用哪个的。然后我就明白了，应该用 BFS，讲了一下 BFS 和 DFS 适用的场景。然后用 BFS 比较顺利的写出了程序。

4.2 算法思想实战及智力题

4.2.1 算法思想实战

- 岛计数问题 dfs
- 两堆钞票，尽可能均分（利用背包问题的思想）
- 著名的 小兔的棋盘，我后来查了一下，是什么卡特兰数。然而面试的时候我没听说过这一道题，不过还是磕磕绊绊地用 DFS 解出来了，面试官说可以了，也没让我继续用 DP 来解。哎，算法还是有点菜的。
- 给一个表只有 id 和时间，如何估算平均访问时长 【撕代码】
- 从用户的访问日志中，选出访问次数最多的 topK 的用户

我的思路：

因为是实际应用，所以我会想到对数据进行预处理，比如对同一用户在同一分钟内（时间阈值可以自己设定）的多条相同访问日志就保留一条的情况，因为会有因为网络异常或者服务器异常的情况，所以对于用户来说真实的访问记录应该是一条记录。 第二步在用快速排序的思想实现

- 过河问题

小明需要踩着石头过河，下一步只能到达距离为 3、4、5 石头。

给一个数组，里面是 n 个石头，以及石头到岸边的距离。

假设从小到大排序，问能否到达对岸（到达第 n 块石头）？

我的方法：用深搜+剪枝的递归实现。复杂度 O(n)

- 给定一个词典，词典里是分好的词，再给定一串文本，找到所有可能的分词序列。我用的 dfs，优化问题整了我好久。

- 棋盘上的连通棋子团数（最基本的 dfs）
- 经典的接雨水问题，然后问能不能空间优化？进一步提问，再给你 k 个格子，你可以随意安防他们，问安放后最多能接多少雨水？
- 不用 api 实现 sqrt(x)，要求时间复杂度小于 O(n)
- 不用 api 实现 pow(x, n) 时间复杂度小于 O(n)
- 假设现在你全中国的一亿个商家的位置信息，现在输入一个坐标 x，y 找出距离你最近的 k 个商家，收一下思路，写一下代码？
- 小机器人从左下角走到右上角，只能向上或者向右，有多少种走法？
- 如果把一次拐弯当做 k，现在传入一个参数 k，问在 k 次转弯下，有多少种走法？
- 给一个数组，进行两次股票交易，最大收益是多少？
- 给定 k 和 n，A 和 B 先后从 1-k 之间挑出一个数，不可以重复挑，然后每次挑出来都加在一起，当当前的和大于等于 n 时，当前选手获胜，求给定 k 和 n 时，A 先手是否能赢（假设两个人每次都是最优策略）
- 一个无序数组，数字代表挡板的高度，挡板没有厚度最多可以盛多少水，如[3, 1, 2] 输出:4

4.2.2 智力题

- 在一片草原上有 1 只羊和若干只狼，狼可以吃羊或不吃羊，但狼吃羊后会变成羊，从而被其它狼吃掉，已知羊不能被两只或以上的狼分着吃掉，并且每一只狼都会先保证自己不被吃掉，而在此前提下每一只狼又都想吃到羊，那么羊是否会被吃掉？
- 43 个石头，A,B 轮流拿，每次可以拿 1~3 个，A 先拿能否保证自己获胜？
- 1000 盏灯开着，1000 个人标号 1~1000 依次进入，每个人进去按一下自己标号倍数的开关，问最后哪些灯亮着？
- 一个人初始体力为 m，起始点为 0，终点为 n。一路上有毒蘑菇和体力蘑菇，用 v_i 表示，吃了会有体力的增加或减少，走多少步就消耗多少体力。问这个人能否到达终点，如果能，最小体力是多少。
- 我手中有一堆扑克牌， 但是观众不知道它的顺序。

第一步， 我从牌顶拿出一张牌， 放到桌子上。

第二步， 我从牌顶再拿一张牌， 放在手上牌的底部。

第三步， 重复第一/二步的操作， 直到我手中所有的牌都放到了桌子上。

最后， 观众可以看到桌子上牌的顺序是：13\12\11\10\9\8\7\6\5\4\3\2\1 请问， 我刚开始拿在手里的牌的顺序是什么？

- 求股票的最大利润，例如[1, 3, 1, 8, 10, 3]，只能买卖一次，计算最大收益

能买卖无数次，计算最大收益

只能买卖两次，计算最大收益

- 十个红球十个白球，无放回抽出 10 个然后红球互不相邻的可能性。没想好，不过具体思想就是一红一白相间地摆好先，然后再在白球红球之间插入白球，面试官说时间关系就先这样了，但是很接近了。

- A 和 B 比赛，A、B 获胜的概率分别是 0.6、0.4，如果你是 A，3 局 2 胜和 5 局 3 胜你会选择哪个。

- n 个人之间存在 m 个关系对，关系具有传递性，假如 A 关注 B，B 关注 C，那么 A 就间接关注了 C。如果一个人被除他之外的所有人都直接或间接关注，那么这个人就是抖音红人，求抖音红人的总数。

- 25 匹马赛跑，5 个跑道，怎么以最少的比赛次数来决出最快的 3 匹(思路分析)

- 三扇门问题。三个门里其中一个有宝石，刚开始你选了一个门，然后主持人开了一个没有宝石的门，问你要不要换？给出具体求解方法。

- 疯狗问题

- 找山顶元素

- 一个岛上有若干人，每个人都戴一顶帽子，不是绿帽子就是白帽子，每个人看不见自己的帽子颜色，可以看见别人的帽子颜色，不能交流。现在知道至少有一顶绿帽子。

- 一个人确定知道自己帽子颜色的时候就会离开，请问岛上会发生什么？

4.3 其他方面

4.3.1 数论

- 求几何分布的期望?
- 泰勒公式，用泰勒公式实现 e 的计算求值?
- x,y 属于[0,1]的均匀分布，求 max（x,y）的期望?
- 丢硬币，连续丢出 2 次正面才停止，求丢硬币次数的期望?
- 一辆巴士载了 25 人，路经 10 个车站。每个乘客以相同的概率在各个车站下车。如果某个车站有乘客要下车，则大巴在该站停车。每个乘客下车的行为是独立的。记大巴停车次数为 X，求 X 的数学期望（要求通过编程求数学期望）。
- 求 max(x1, x2)期望?
- 已知 var(x),var(y),E(x),E(y)。求 Var(x*y)
- a , b ~ U(0,1), a 和 b 独立。求 E(max(a,b))
- 假设有一组基向量 b1,b2,...,bn,现在有一个向量 x，希望能用这组基向量中的三个表示，也即 x=w1bi+w2bj+w3bkx = w_1b_i + w_2b_j + w_3b_kx=w1bi+w2bj+w3bk,问如何求解这个问题?
- 定义域值域都是正整数的单调递增函数 f，给一个值 y，找到使|f(x)-y|最小的 x。

（肯定是二分，但其实有很多细节值得注意。如定义域值域都是正整数，所以可以推出 f（x）是不可能小于 x 的，应该是 x^n 的形式，所以开始搜索的范围就是 ed=y）。

- 给一个中文数字，比如一百二十，如何转换为整形数字
- 由长度为 length 的 array 表示的整数，允许相邻位数交换，求 n 步交换内能得到的最小整数
- 编程题：3*x+7*y=125,x,y 为质数。如何快速求解?
- 把只包含质因子 2、3 和 5 的数称作丑数。例如 6、8 都是丑数，但 14 不是，因为它包含质因子 7。 习惯上我们把 1 当做是第一个丑数。求按从小到大的顺序的第 N 个丑数。
- 从 K 个整数中，组合出能被 3 整除的最大数，例如: [1, 2, 3]，组合出能最大能被 3 整除的

数是 321

4.3.2 计算几何

- 一个图片中心逆时针旋转 30 度后，求最小外接矩形长和宽，说一下有哪些解决方法？

第一种初中数学，几何知识；第二种，求解仿射变换矩阵（2x3），然后和原图相乘，就得到变换后的图片，也就知道了最小外接矩形的长和宽。

- 现在有一堆点，求一个点到每个点的距离之和最小，证明这个点是质心。

4.3.3 概率分析

- 给一个 01 二项分布的随机器，参数为 p，用它设计一个 0-1 的均匀分布的随机器（连续的）
- 已知 x.y 的概率分布，求 max x,y 的分布？
- 一个圆上随机三点，求形成锐角三角形概率，要求数学推导
- 现在我有抛一枚硬币，正面朝上的概率是 p,反面是 1-p。那么第 k 次抛的时候出现第一次正面的概率是多少？

A: $P(1-p)^{k-1}$

Q：好的，那么我们设 $f(z=k)=p(1-p)^{k-1}$ ，那你计算一下 $E(z)$（求个均值）

A：(想了一会) $E(z)=p+2p(1-p)+3p(1-p)^2+...+mp(1-p)^{m-1}$

Q：能不能计算一下 $E(z)$ 的数学表达式

A：好的，思考了一会，可以使用 $E(z)-(1-p)E(z)=A$ 。其中A是一个等比数列。然后就可以求出E(z)。

- 抛 2k+1 次硬币，问正面次数比背面多的概率是多大，并讲出数学证明思路。
- 3 种颜色砖块，单位长宽，铺满单位宽，长 m 的地板有多少种铺法？
- A、B 交替抛硬币，正面概率为 1/2，谁先抛到正面谁胜。问 A 先抛并获胜的概率
- 13 个人生日都不是同一天的概率，要求给出表达式和最终结果（不用计算器估算）
- 给你一个圆，让你在上面画三角形，要求三个顶点在圆周上，问画的三角形是锐角三角形的概率是多少。怎么求解？
- 2 个人玩游戏，每局获胜的概率都是 50%，A 赢 3 次胜利，B 赢 2 次胜利，求 A B 的

获胜概率（就是一个状态转移问题，画了图，秒掉）

● 斗地主农民拿到炸弹的概率是多少，听到这个题都蒙了，完全不知道怎么解，面试官也说他自己都不知道答案，然后就在面试官的一次次提问下写了个大概思路

● 一条线段分成三段能够组成三角形的概率，这个题目碰到过，不过当时没想起来，面试官提示了下解答出来了

● 抛一个不均匀硬币五次，两次正三次反，下一次正的概率 p1 是多少？

● 抛一个不均匀硬币五十次，二十次正三十次反，下一次正的概率 p2 是多少？

● 概率题：x, y 服从 0-1 均匀分布，求 x+y<1 的概率？x, y, z 服从 0-1 均匀分布，求 x+y+z<1 的概率？

● 一道数学题，A、B 两人投硬币，谁先投到正面谁就赢，求先投的人赢的概率

● 问：我们来讨论一个下雨的问题

今天下雨的概率是 0.2，天气预报的准确率是 0.8. 问已知今天下雨，天气预报预测下雨的概率？

● 在[a,b]之间，a，b 为正整数，问其中不包含数字 3，5，7 数字的个数；

● 一个硬币，正面向上是 p，投 2k+1 次，正面比反面多的概率，写出表达式

● 真硬币 m 个，假币 n 个。假币只有正面。真币投掷正面概率为 p。其中某硬币投掷 k 次都是正面，求它为真币概率。

● 斗地主中一个农民抓到王炸的概率

● 一分钟看到红车的概率是 0.2，一个小时看到 1 辆红车的概率是多少

● N 个相同的球，取其中 M 个(M<N)，如何保证每个球取的概率一致？

我答了有放回取样，已取样的做标记，若再次取到有标记的则放回重新取，直到取得 M 个。

考官让算一下，这么做的期望复杂度是多大。没算出来。

● 如果是头条的用户场景，每天用户总数量是不确定的，但是要抽 M 人，如何保证概率一致。

考官提示：如果已经有一个函数，使 N-1 个人中等概率抽取了 M-1 个人，那么下一个人加入的时候如何保证等概率。

在这个提示下我想到了需要列式使新加入进来的人概率和前一个人上一次中的概率和这一次的概率之和是一致的，以此类推。其实一下子还是没有写出完整表达式，因为时间比较捉急了，我直接用 N 个人抽 N-1 个的特殊情况写了递推式，表示取 M 个的话需要进行变形。考官应该是认为我已经理解了思路，所以结束了这个问题。

总结：保持和考官的交流，有思路及时沟通，一下写不出答案可以先考虑特化情况。

- x, y 服从 0-1 均匀分布，求 x+y<1 的概率？x, y, z 服从 0-1 均匀分布，求 x+y+z<1 的概率？
- 一个概率问题，每个人投票给 a 概率 51%，每个人投票给 b 概率 49%，问投票人数和 a 最终获得更多票数之间的关系
- 概率题：飞机上有 100 个座位，有 100 个乘客准备登机，每个乘客按顺序上飞机，但是第一个乘客喝醉了，随机挑了一个座位来坐。每个乘客的选座位规则：1）如果自己的座位没被坐，则坐自己的位置；2）如果自己的座位被坐了，则从剩下的座位中随机选一个来坐。则第 100 个人能做到自己座位的概率是？
- 甲乙射击比赛，单局甲胜率 0.6，3 局 2 胜和 5 局 3 胜两种赛制甲如何选择？
- 网游中杀死小怪时候，有 P=0.2 的概率掉落一把宝剑，野猪的死亡是独立事件，某玩家杀了 10 个小怪，求掉落 4 把宝剑的概率？
- 你有 1000 瓶饮料，其中有 1 瓶有毒，你有许多老鼠，老鼠喝完饮料之后 24 小时会死，请问你平均需要多少天找出这个有毒的饮料？需要多少老鼠？
- 你有 1000 法力值，有 4 个技能，技能伤害值与消耗魔法值成正比，请问你怎样用技能，才能做到伤害输出最大？
- 打怪有 80%概率掉落 a 装备，20%概率掉落 B 装备，请问一个人平均要打几次怪，才可以凑齐 ab 装备？
- 人群中男人色盲的概率为 5%，女人为 0.25%。从男女人数相等的人群中随机选一人，恰好是色盲。求此人是男人的概率。
- 甲扔 n 次骰子，取其中最大的点数作为它的最终点数，乙扔一次骰子得到点数，求乙的点数大于甲的概率。
- 某种病的发病率为 1/100，某种检测该病的技术检测正确率为 99/100，现有一人被检测到

生病的概率为 p，求他真实生病的概率是多少？

在上一问的基础上，现在连续两次检测为有病才会停止检测，求检测次数的期望值。

● 10 个人里每个人在 10 分钟内的任何一个分钟到达的概率是均匀分布的，问所有人都到达的时刻在几分钟时概率最大。

● 问比赛甲获胜概率 0.6，乙获胜概率 0.4，该选三局两胜还是五局三胜。再问不通过计算怎么判断？当 n 为一个趋近于无穷大的奇数时，甲乙获胜概率如何？

● 斗地主有人拿到 2 张王的概率

● 已知 1-5 的随机数发生器，怎么生成 1-7 的随机数发生器

● 假定一个人胜的概率是 p，判断五局三胜有利还是七局四胜有利

● 有 100 个乘客，每个乘客手里有一张座位票，座位是 1 到 100 标号。正常情况下应该对号入座。但是第一个乘客喝醉了，他没坐在 1 号位置，而是从其他 99 个座位随机选了一个座位。从第二个乘客开始，如果他的座位没有被前面的人占领，他就对号入座，如果被前面的人占了，他就在剩下的座位里随机选一个，问第 100 个人正确坐到自己座位上的概率是多少？

● 一个圆，问走 n 步回到原点多少种方法？

● 10 个球放到 12 个盒子里 空盒子=5 的概率 P=？用代码模拟 10000 次的概率是多少？

● 10 个小球，随机分到 12 个盒子里，求恰好 10 个盒子都为空的概率。要求用 Python 程序模拟十万次，暴力求出该概率。

● 假设现在有一个函数 random(), n 为未知数，1/n 的概率返回 0，2/n 的概率返回 1，写一个 newRandom()，让返回 0,1 的概率各为 1/2。medium。

● 给定 N 种不同颜色的球以及每种颜色的球的数量，把它们放进一个容器里面，随机抓取。要求写程序实现该功能，并且要按照每种颜色球的概率返回对应的球的编号。

● 比如，有 A, B, C 3 种颜色的球，数量分别是 1，2，3。然后把它们统一放入盒子里，随机抓取（使用 random 随机生成(0, 1)之间的小数），要求按照它们各自的频数返回对应的颜色的球。

● 有两张表，第一张表有 n 个专有名词，比如今日头条、抖音等，第二张表有 m 条 query，比如今日头条是怎样的应用、有多少人喜欢刷抖音等，如何统计表 1 中所有名词在表 2 中出现

的频次。

● 有一个生成 0-4 的均匀分布的整数随机数生成函数，利用这个函数生成 0-9 均匀分布的整数随机数生成函数。

● n 个[0,n)的数，求每个数的出现次数（不能开辟额外空间）

● 有一个 0-1 的均匀分布随机器，用它实现一个 N(0, 1)的正太分布随机器?

● 一根木棍, 随机切成三段, 求能围成三角形的概率?

● 三个盒子分别放的球为：“红 红”，“红 蓝”， “蓝 蓝”，第一次取出一个红球后，取出两个红球且为第一个盒子的概率?

● 给一个现成的生成器，可以以概率 p 生成 1，概率 1-p 生成 0，让我用这个生成器构造一个新的生成器，满足每次均匀返回 0-1 之间的一个浮点数?

4.3.4 矩阵运算

● 二维矩阵，求连通区域数量（连通的定义： 两个像素是四邻接的邻居，并且像素值的差的绝对值小于等于 16，那么这两个像素是连通的）。

● 蛇形打印 n*n 的矩阵

● 给出一个数字矩阵，寻找一条最长上升路径，每个位置只能向上下左右四个位置移动。

● m*n 矩阵从左上角走到右下角一共有多少种走法？如果有障碍物的话怎么求？求最大的路径和？先从左上到右下在从右下返回到左上，重复走的节点值为 0，求两条路径加和最大值？（都是 DP）-问时间复杂度

● 矩阵的转置和回旋输出

● 写个矩阵乘法，不让用 numpy，再优化下

● 由 0 和 1 组成的二维矩阵，找出 1 的最大连通域，计算其面积。

● 矩阵 TopK 问题？先说思路，思路不对、复杂度太高的话就不用写了。

● [(“A”， “B"), ("C", "D"), ("B", "A")]找出该列表中所有的环路?

● 一个二维矩阵由小到大排列，找 target 数字?

● 给定一个矩阵里面只包含 0 和 1 两种数字，给出每个单元距离最近的 0 的距离？上下左右

都算作相邻，相邻的距离为 1。

- 有序矩阵中第 k 小个元素
- 三数之和 = target 找出所有可能三元组

4.3.6 其他

- 一个无序数字序列，每次只能左旋操作 3 个数，求要求有序下，证明能否通过有限次数能否有序。
- 如何判断一个算法是线性的还是非线性的?
- 中位数与平均数什么时候会相等？貌似是数据分布对称就行?
- 手撕代码，开根号，以前没遇到过这个问题，于是写了二分查找。面试官问会不会牛顿法，现场推了下公式，结果牛顿法太久不用公式都忘了，用泰勒展开推了一下写成了拟牛顿法
- 求 2 的平方根，精度 0.00001
- 分解质因数【撕代码】
- 输出幂集（比较简单，给了两种方法，迅速的过了）
- 给出 6 * n 的方块，用 1 * 2 或者 2 * 1 的方块覆盖它。不要求求出具体的个数，证明该方法时多项式级数还是指数级数?

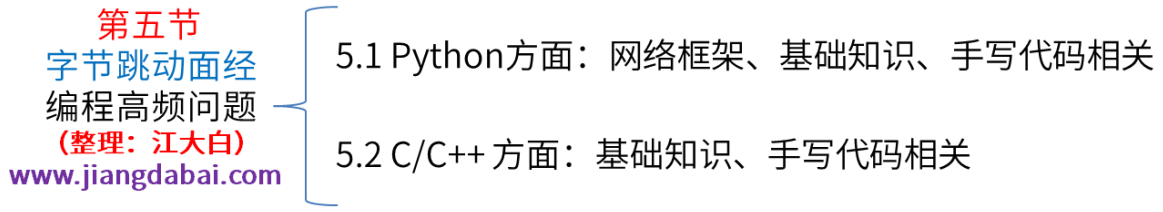
4.4 Leetcode&剑指 offer 原题

- Leetcode 3 最长不重复子串
- Leetcode 11
- Leetcode 32
- Leetcode 34（简单变形）
- Leetcode 42
- Leetcode 72：字符串的编辑距离
- LeetCode 76：Minimum Window Substring
- Leetcode 123

- Leetcode 124
- Leetcode 148
- Leetcode 206
- Leetcode 224：hard，实现一个基本的计算器来计算一个简单的表达式字符串。表达式字符串只包含非负整数、+, -, *, /操作符。可以假设给定的表达式总是有效的。
- Leetcode 284：一个类 A 有 next，has_next 两个方法，其中 next 调用会返回值，但索引会自增。实现一个 peek 访问只返回值，索引不自增。
- Leetcode 306
- Leetcode 315
- Leetcode 378
- Leetcode 448：要求时间 o(n)，空间 o(1)。
- Leetcode 560
- Leetcode 786
- Leetcode958：判断一棵树是不是完全二叉树
- Leetcode1047
- Leetcode 原题：链表采样，resevior sampling。要求在一个无限长的单向链表中采样，当遍历的节点数量充分多时，每个节点被采样到的概率应相等。
- Leetcode 原题：正则表达式匹配
- Leetcode 原题：求数 x 的开方，精确到小数点后一位
- Leetcode 原题：旋转数组查找 target
- Leetcode 原题：字符串左移 K 位
- Leetcode 原题：股票价格
- Leetcode 原题：矩阵右旋
- Leetcode 原题：01 矩阵找最大子矩阵大小
- Leetcode 原题：n 的平方根，精度十位小数

- Leetcode 原题：计数质数
- Leetcode 原题：蚂蚁爬杆
- Leetcode 原题：屋舍打劫
- Leetcode 原题：数据流中，按照某个窗口大小，找窗口中的最大值
- Leetcode 原题：最长回文子串
- Leetcode 原题：medium，假设现在有一个函数 random(), n 为未知数，1/n 的概率返回 0，2/n 的概率返回 1，写一个 newRandom()，让返回 0,1 的概率各为 1/2。
- Leetcode 原题：一个字符串，一个单词字典，把字符串分成若干个子串，每个子串都包含在字典中，返回多少种分割法？
- Leetcode 原题：hard，现在有一个每行每列递增的 2D 数列，比如[[1,2,3,4], [2,3,4,5], [4,5,6,7]]，在 O(nm)的时间复杂度返回最小的 k 个数。.
- 剑指 offer 41：数据流中的中位数，设计一个数据结构，有插入和删除操作，并且能随时得到数据中的中位数。

5 编程高频问题：Python&C/C++方面



5.1 python 方面

5.1.1 网络框架方面

5.1.1.1 Pytorch 相关

- torch.Tensor 和 torch.tensor 有什么区别？
- torch.no_grad 和 required_grad=False 的区别？
- pytorch 里 function 和 module 有什么区别？

- pytorch 里 dataset、dataloader、sampler 有什么区别?

5.1.1.2 Tensorflow 相关

- 问 TensorFlow 如何实现并行，我讲了下 PS 架构，又问梯度更新时是同步还是异步，同步异步的优缺点。最后问到优缺点说明面试官经验还是非常丰富的，刨根问底可以看出你是否真的深入思考过这个问题，抑或是从别人博客看来的人云亦云。
- 用 TensorFlow 多吗？它的优化器都是什么？分别介绍一下
- TensorFlow 怎么实现梯度传递?
- Tf、keras、pytorch 区别?

5.1.2 基础知识

5.1.2.1 线程相关

- 你比较熟悉 Python?你了解过他的多线程么？ 为什么多线程比较鸡肋
- python GIL 解释一下
- python 里的多线程，怎么让它占满核呢?
- 问了 python 中的进程和线程？python 中线程有什么缺点吗（全局锁）？
- 既然 python 中的线程有全局锁是不是没有啥用（不是，虽然有全局锁但是对于一些 I/O 操作较大的应用影响不大，因为他们并不需要真正的并行运算）？
- Python 中多进程多线程的应用

5.1.2.2 内存相关

- STL 中 vector 的低层实现，STL 中插入的操作时间复杂度，要考虑内存复制扩充，
- 如何实现一个栈，支持动态扩充
- python 里面的深拷贝和浅拷贝
- python 的回收机制

5.1.2.3 区别比较

- python 是解释语言还是编译语言

- xrange 与 range 的区别，xrange 通过什么关键字实现的？yield 语句底层如何实现？
- python 中 is 与== 的区别
- new 和 malloc 区别
- map 和 unordered_map 区别
- 了解 utf8 吗？utf8 有几位？英文在 utf8 中占几位？utf8 和 utf16 有什么区别？（utf8 四连把我给问蒙了，出师不利）
- python2 和 python3 的区别，python2 为什么要更新成 python3，print 为什么要加括号？

5.1.2.4 讲解原理

- python 装饰器，python 迭代器和生成器介绍一下
- python 全局锁是什么？
- python 的动态数组是如何实现的
- python 中 dict 的底层是啥
- 问 python 中 list 的底层怎么实现
- python 有多线程吗？
- python 里的生成器是什么？

5.1.2.5 讲解应用

- Python 数据结构有哪些？
- python 中的 C 拓展的具体例子

5.1.3 手写代码相关

- 交换 a 和 b 两个数，不借助第三个变量
- 用 lambda 表达式生成奇数的数组
- 实现 sqrt 函数，结果保留 5 位小数

5.2 C/C++方面

5.2.1 基础知识

5.2.1.1 内存相关

- C 语言的动态内存分配器
- 是否了解虚函数，虚继承，多态。 vector 底层如何实现插入的时候不改变内存空间
- C++共享内存实现原理，多线程通信，互斥锁

5.2.1.2 区别比较

- 参数传递时，传值、传引用和传指针的区别
- 函数返回时，返回值、返回引用和返回指针的区别
- dynamic_cast、static_cast 和 reinterpret_cast 区别
- const A&func(const B& b) const 中三个 const 的区别
- new 和 malloc 有什么区别
- C++的虚函数和虚继承的作用
- 数组和链表的优缺点?

5.2.1.3 讲解原理

- 重载和重写
- 什么是多态，如何实现多态
- 多态；虚函数表；虚函数表指针大小；gcc 编译过程；
- c++11 新特性；const；new、malloc 区别; 虚继承；四种类型转换；智能指针
- C++ static 关键词的作用，初始化参数列表有什么用
- 介绍 C++的虚函数，析构函数一定要是虚函数吗?
- C++面向对象介绍下?

- static 修饰符有什么用？如果不加会出现什么后果?
- List 的底层实现?
- Hashmap 特点?HashMap 的实现原理?如何解决 Hash 冲突?
- LRU 算法讲一下?
- FFmpeg 熟悉程度?
- 研究过 h264，h265 和 h266 吗？写过解码器吗?
- 设计一个带 TTL 的容量有限的 hash map，在 TTL=0 后从 map 中移除。在装满后，如果还有元素进入，将 TTL 最小的元素移除。

5.2.1.4 讲解应用

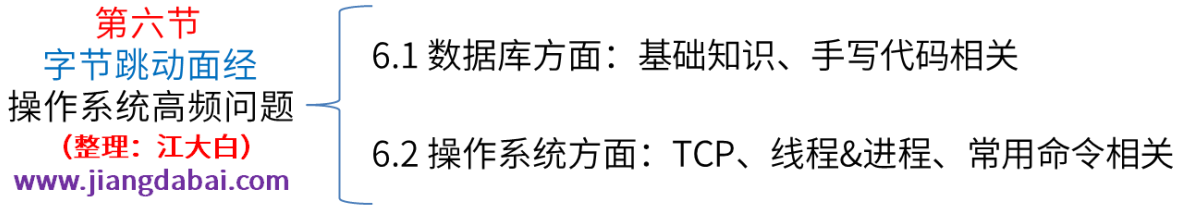
- C 语言中结构体 struct{int i; bool b}一共占几个字节?
- C++的析构函数一定要是虚函数吗?
- 如何从用户态进入内核态?
- C++定义图和节点
- 这样声明变量有没有问题：int a[10000000]?
- C++ map 中自定义的 class 为什么不能作为 key?

5.2.2 手写代码相关

- 实现一个 C++双向链表类
- 手写双向链表插入新元素；
- 稀疏向量的点乘。先要我自定义存储的结构体，然后写函数头，再编程，本来要我用 template 但我不会就算了
- 如果实现 c++中的 vector，只需 push_back 和查找两个功能，底层如何实现。
- 实现 memcpy
- 手写拓扑排序?
- 底层基于 xnor 运算的二值化卷积算子你是怎么实现的？ 可以简单实现一下吗？现场用

CUDA 写了底层基于位运算的 GEMM kernel ，实现 im2col 的矩阵乘法。

6 操作系统高频问题：数据库&线程&常用命令等



6.1 数据库方面

6.1.1 基础问题

6.1.1.1 区别比较

- SQL 中 count(1), count(*), count(列） 区别

6.1.1.2 讲解原理

- Mysql 的数据存储结构
- MySQL 底层是什么、B 树和 B 树的区别（因为我回答 MySQL 用 B 树）

6.1.2 手写代码

- SQL 如何取出成绩表中各科的前三名?

6.2 操作系统方面

6.2.1 TCP 协议相关

- TCP 和 UDP 的区别:一个面向连接，一个面向无连接。
- 为什么 TCP 比较慢，另外怎么保证可靠性
- TCP 如何保证消息安全?
- 拥塞控制，流量控制，然后问流量控制会不会发生死锁?
- 三次握手和四次挥手的原理?

6.2.2 线程和进程相关

6.2.2.1 区别比较

- 线程和进程的区别和联系,进程和线程相比有什么好处?
- 同步 IO 和异步 IO

6.2.2.2 讲解原理

- 进程之间的通信方式 ,进程之间的序列
- 常见的进程调度算法? 公平调度? 时间片调度?
- 32 位操作系统和 64 位操作系统的寻址范围?
- 线程如何保证原子性?

6.2.2.3 讲解应用

- 讲一下进程和线程的应用场景（多线程爬虫）。
- 加问死循环发生原因:
- 多个线程执行 put 操作时同时触发了 rehash 方法，可能会生成环形链
- 两个线程，每个线程有一个数组，交替输出，怎么做?

（问操作系统、数据库、网络哪个比较会。。答差不多都忘了，随便问吧，应该就是跪在这里了吧? 所以说投算法的同学还是要复习计算机基础的，基础不能放下啊）

6.2.3 常用命令

- Linux 系统怎么查看进程 CPU 使用率?

7 技术&产品&开放性问题

7.1 技术方面

- 电梯算法设计，机器学习怎么搞?
- 听音乐，怎么推荐，10 亿首，怎么设计方案（K-D 树以及实现）

- 有一个 1000w 的视频库，每个视频 3-5 分钟。 新来一个视频，我们需要去这 1000w 的视频库里查询，是否存在相同视频。 1.选择使用什么样的特征 。 2.设计一个好的 index，使得查询尽可能快 （这 1000w 视频可以离线处理）-抽帧+Pooling+相似性哈希
- 工程中遇到了哪些优化复杂度的方法。
- 正负样本数据不均衡怎么处理。说了几种方案，欠采样，负采样，生成负样本等。接着问我怎么生成负样本，说了图像处理，GAN 等等。就问我生产样本需要注意什么，我说最重要的是要和原始样本的分布保持一致。接着问我怎么能保证分布一致，到这基本就是我的知识盲区了，靠着经验和理解开始扯，期间他一直盯着屏幕打字，中间还打了几个哈欠，让我一直纠结还要不要接着说下去，最后实在扯不下去了，经历了十几秒尴尬的沉默之后。他说，问你道题吧。
- 如果遇到一个模型的归回效果特别差，怎么考虑和解决？A:首先考虑模型的选取是否符合问题，然后考虑数据样本映射到更易回归的空间中（后来觉得应该是：先看数据是否有很多异常值、离群点影响模型效果，然后再看模型的复杂度够不够，通过增加模型复杂度来提升模型拟合能力，好了，我不瞎说了，说的都是皮毛，纸上谈兵罢了，自行百度，真的可以很复杂）
- 如何判断机器是大端模式还是小端模式
- 优化搜索引擎时，如何从用户的行为上判断用户对我们提供的搜索结果的满意程度？很实际的问题，但对我而言是完全陌生的领域，估计研究推荐系统的人都懂这个，我当时绞尽脑汁：如果用户经常点开的链接不是第一个结果，而是后面的若干结果，那么说明用户对结果不满意。
- 三维视觉,你的研究方向在我们的产品中可以用到哪里？

回答了抖音的动态贴图和头部三维建模

现在大部分手机都是单目摄像头没有深度摄像头，怎么解决，来应用你的算法

回答了模型需要通过三维数据进行训练，但是应用模型时可以直接单张图片重建三维模型

了解动作捕捉吗？现在有一个人在做表情，有一个动画模型，怎么让动画模型做出人脸同样的表情？

回答了 3DMM 的表情系数

- 现代 cpu 算力在什么量级

● 开放题：有一个 1000w 的视频库，新来一个视频，我们需要去库里查询是否存在相同视频。存储这么多视频时应该选择什么样的特征，查询要用什么方法。（这个回答的还行，楼主大致说了一下用 LSTM 提特征再哈希的思路，面试官没说什么就过了）

● 开放题：磁盘上有 M 个数组（M 很大），每个数组长度不确定，数组内数值不会重复。现在要求其中 n 个数组的交集。设计算法使这个求交集的速度最快，另外有内存限制。（这题真的雪崩！楼主在看题的时候把交集当成了并集来做，所以思路完全跑偏了。。。后面是快结束的时候面试官一提醒才发现的。然后赶紧说了一个两个数组求交集的实现就结束了）

● 开放题：你觉得影响模型效果的因素有什么，并排个序。（我回答的是特征>模型>优化器，才疏学浅只答了这些）

● 开放式问题：怎么论证现有的模型需要多少额外的标注数据

● 给你一亿个特征，现在来了一个新的特征如何处理。

● 2.5 亿个整数找不重复的整数，内存无法一下存下这 2.5 亿个数，怎么做。

● 写一个函数，输入是 N 个文件，每个文件中是很多 float 的数值，文件内部无序。输出是一个有序的大文件，内存约束 2 个 G，磁盘可以随便用，具体怎么实现比较高效？

● 设计一个函数，判断传入的 IP 地址在过去一个小时之内的访问次数是否大于 10000 次。这个问题没有做出来。（还是没有 get 到面试官出这个题的点，当时自己想的太过于复杂，就不知如何去实现了）

● 推荐系统方面：具体场景应用是由词向量引出的，首先问了词向量负采样的细节，数学原理，然后问大规模的用户导致维度爆炸怎么处理？如果让你设计一个推荐系统怎么做？

● NLP 方面：设计一个系统来筛选抖音中的低俗视频？（从视频帧，图片，文字三方面来提取特征）

● NLP 方面：现在有一些新闻，包含军事、体育、经济等，想分出它属于哪个类，该怎么做？

● 开放题：在一个只能传输 0，1 的信道里面,如何传输两个数字？（面试 AI Lab 机器学习实习生时问的，非常规）

7.2 产品方面

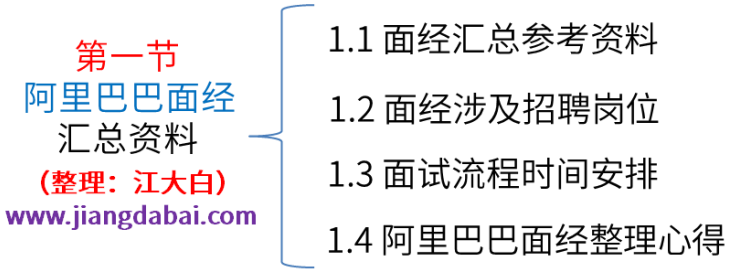
- 模型升级后，放到线上使用，发现线上效果不好，此时你该怎么办？
- 给图片去水印怎么去？
- 面试官解释题目：假如抖音里面有 5 亿用户，那么每个用户打开一次抖音就有 5 亿条记录，如果每个用户打开两次抖音，就有 10 亿条记录。也就是说，用户每打开一次抖音，就记录一下他的 uid。请找出打开抖音次数最频繁的前 10 个用户。
- 聊了一些开放性问题，问抖音 APP 的冷启动怎么做。如何推测用户的性别，怎么手机训练模型用的数据。你觉得你使用抖音有什么问题，有什么解决方案。
- 怎么推测抖音用户是男性还是女性？怎么确认判断结果的靠谱程度？
- 从一段长视频中截取或者拼凑 10s-20s 的短视频用于广告投放，吸引用户点击下载 app，怎么得到目标短视频？
- 任务，找出大量数据中的宠物猫

7.3 开放性问题

- 抖音你觉得如何刻画用户画像？用户在不同时间段刷，怎么理解用户的属性？如果一个时间段用户活跃度不高，怎么从数据挖掘角度看呢？
- 如何给主播打标签？
- 给你一个产品，作为一个算法工程师，你会做些什么，来让你的产品变得更好？

2|阿里巴巴算法岗武功秘籍

1 阿里巴巴面经汇总资料



1.1 面经汇总参考资料

① 参考资料：

（1）牛客网：阿里巴巴面经-201 篇，网页链接

（2）知乎面经：点击进入查看

（3）面试圈：点击进入查看

② 面经框架&答案&目录&心得：

（1）面经框架及参考答案：点击进入查看

（2）大厂目录及整理心得：点击进入查看

1.2 面经涉及招聘岗位

（1） 实习岗位类

【蚂蚁金服机器学习实习】、【达摩院研究型实习生】、【达摩院自然语言实习生】、【蚂蚁金服风控算法实习生】、【阿里云实习】

（2） 全职岗位类

【机器视觉算法工程师】、【智能事业服务部算法工程师】、【NLP 算法工程师】、【达摩院计算机视觉】、

【onsite 算法工程师】、【消费者 bg 的软件算法工程师】、【蚂蚁金服算法工程师】、【机器学习算法工程师】、【大文娱算法工程师】、【阿里 UC 神马搜索】、【阿里口碑机器学习】、【阿里新零售天猫机器学习】、【阿里搜索部算法工程师】、【新零售技术事业群 CCO 技术部自然语言处理】、【阿里大文娱机器学习】、【淘宝算法岗】、【阿里优酷用户增长组】、【阿里飞猪算法工程师】、【企业金融工程师】、【达摩院 NLP 工程师】、【数据分析/机器学习工程师】、【阿里健康算法工程师】、【推荐算法工程师】、【lazada 算法工程师】、【蚂蚁金服人工智能部门算法工程师】、【淘宝技术部算法岗】

1.3 面试流程时间安排

阿里巴巴面试流程-整理：江大白			
	面试类型	面试流程	备注（侧重点）
第一面	技术面	自我介绍+项目/实习经验 +技术问答+算法编程	项目基础&细节问的很细， 项目经常一个个问
第二面	技术面	自我介绍+项目/实习经验 +技术问答+算法编程	主要围绕项目落地，常问 创新点以及为什么这样做？
第三面	技术Leader面	自我介绍+项目经验+公司发展	偏发散性考察，解决问题的思路，对知识的理解以及延伸
第四面	交叉面	自我介绍+项目/实习经验 +技术问答+算法编程	项目方面的应用 以及技术领域的宽度
第五面	HR面	基础人力问题	权利很大，价值观为主 比较重要，有可能会挂人

PS：以上流程为大白总结归纳所得，以供参考。

其他注意点：

- 有时是 2 技术+交叉面+HR 面、有的也是 3 技术+交叉+HR
- 第一面也可能是简历面
- 个别部门参照：

达摩院实习生是 3 面技术+1 交叉面+HR 面

蚂蚁金服是 1 笔试+6 面试

搜索部是 3 面+交叉面+HR 面

1.4 阿里巴巴面试心得汇总

阿里的面试资料很多，下面大白将很多面试者的心得也提取出来，便于大家在准备面试时，有所倾重：

★ 项目经历很重要，但是一定要对自己项目的每个细节了如指掌，挨个聊项目，每一个都问得非常细,要有一些有深度的思考，这是面试官期待看到的

★ 阿里流程会比较长，需要耐心一些，如果可以的话，大家一定要找一个靠谱的师兄师姐内推，通过他们可以提前知道接下来的流程。

★ 自己的论文，项目一定要讲清楚（背的滚瓜烂熟，阿里还是很重视论文项目的，对算法题要求没有那么高）

★ 每一轮面试都要认真对待，即使是交叉面和 hr 面也要好好准备 （我就是到了交叉面有点浮躁，面试完感觉要凉凉了），hr 面我就老实了（毕竟 hr 具有一票否决权），提前准备了好多、常问的问题（你的优点缺点，最近看什么书，对福报的理解，部门的职责，团队的职责，价值观相关，还好问的我问题比较常规）

★ 阿里笔试要认真做（题目灵活性还是很大），然后积极找找师兄内推还是很有机会参加面试的。面试官很看重基本功，和逻辑思考能力。有些项目的问题可能自己也没考虑过，不过也不要担心，重要的是如何去分析和解决那个问题。所以感觉考察的方面还是挺综合的。

★ 对于简历上的知识点，在深挖的同时，注意广度，并且还要注意凸出自己的方向，不能因为广而浅尝辄止，那样容易给面试官留下不好的感受。

★ 阿里的考察非常全面，一二面考察偏重基础，四面考察偏重业务，交叉面比较简短。笔者为通信专业，非科班，这里要提醒非科班的同学一定要重视基础，尤其是数据结构，不能光刷题，要系统的仔细的学习这门课程!!!

★ 面试官的问题更加全局化，站在业务和整体的角度看问题，比起算法细节，更多的问的是一种思维方式，为什么选择这个办法而不选择那个办法？为什么这个方法的结果比那个好？还是要多学习多思考。

★ 阿里的感觉是比较重视工程的能力，你是不是可以应对量级较大的数据，很多模型的存在

很依赖于要达到什么样的目的，数据的关注的重点：数据质量，数据你怎么处理，怎么使用，比模型调参更体现了你的能力。

★ 不会的千万不要往简历上放，会的能说的就多说点，阿里的题比较开放，给你大的发挥空间，也可以很快的测出你的能力，阿里的实习给我实践性的启迪很大，就是让我觉得，我是来解决问题的，那解决问题比较重要的点在哪，不能只关注部分环节。

★ 简历首先最好有点能了解透彻的项目，简历上涉及的基础算法都要搞懂，面试有一定技巧性所以可以先去面几家练练手，算法题肯定要刷的，统计学习方法要看的，深度学习模型调优什么的。

★ 既要深度也要广度，发散性的问题，常常会问，为什么要这样？为什么不要那样？有什么好处？

★ 简洁，不要扯东扯西，不要发散思维，这样会妨碍面试官了解到他想得到的信息。当然，如果你知道他想问什么，往那方面发散是最好不过的了。但是一开始最好还是让面试官引导，自己先摸清面试官的意图，面试官的技术栈。

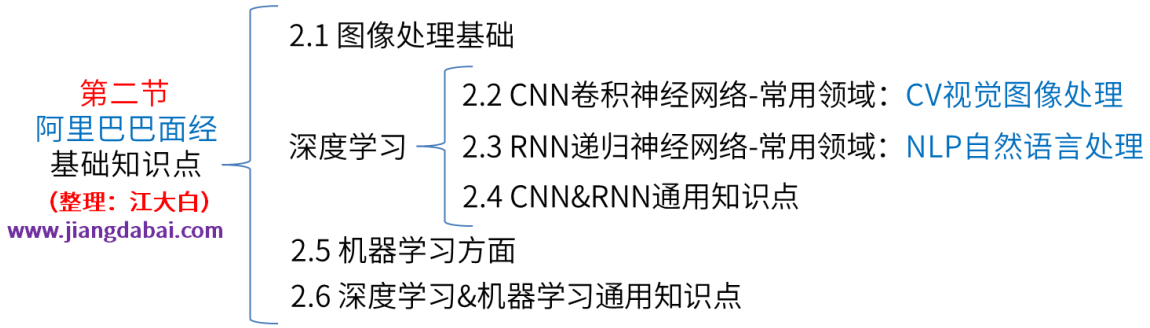
★ 蚂蚁非常看重的是你对于自己研究领域的思考和实践，比较注重你是否具有较高的学习热情和个人潜力

★ 阿里巴巴在编程语言上甚至落地分布式的问题都可能问的很深。。一个学弟搞图像检测的，都聊到分布式负载均衡和反向代理了。

★ 深挖你的项目经历，首先让你概括性描述一下你的项目主要内容，以及你在里边儿承担的具体工作，然后开始一点一点扣项目经历。(真的要好好准备，而且还要思考一下为什么这么做，以及还有其他什么方法，各种不同方法之间的对比，问的特别细致反正，但感觉这也看面试官吧)

★ 阿里的 HR 权利很大，即使面到了最后一面 HR 面还是有很大可能会被挂掉，一句话夸张的概括一下阿里 HR 的存在吧，阿里技术部的面试官是在给 HR 部门招人，而不是 HR 再给技术部招人。

2 阿里巴巴面经涉及基础知识点



2.1 图像处理基础

2.1.1 讲解相关原理

- 计算连通域的个数？算法复杂度？
- 如何求边缘，45°边缘，高斯滤波喝双边滤波？

2.1.2 手写算法代码

- 灰度图求直方图
- 给你个 0 1 二维矩阵，寻找 1 的连通域有几个？算法复杂度？怎么优化加速呢？

2.2 深度学习：CNN 卷积神经网络方面

2.2.1 讲解相关原理

2.2.1.1 卷积方面

- 卷积的作用，以及局限
- 比较 VGG 和 LeNet，VGG 使用 3*3 的卷积核最大的优势是什么？
- CNN，CNN 的核心是什么？卷积操作是怎样进行的？卷积反向传播过程？Tensorflow 中卷积操作是怎样实现的？（感觉这种问题是逃不掉的），池化又是怎样操作的，反向传播过程中池化层怎么接受后面传过来的损失的？
- CNN 参数共享什么意思？

- 反卷积和上采样什么意思?
- Dropout 跟浅层网络区别，dropout 后验证集如何处理?

2.2.1.2 池化方面

- 池化层的作用?
- max pooling 和 ave pooling 有啥区别，说一下?
- pooling 层为什么用 max pool？有哪些 pooling 层技术？各自对比、使用场合等等?
- Max Pooling 是如何反向传递梯度的?

2.2.1.3 网络结构方面

- ResNet，让我介绍了一下 ResNet 主要解决的问题是什么，然后又问我对看完 ResNet 有什么看法
- Alexnet 的问题，大概是什么结构，和 LeNet 比有什么改进的地方，问了 Relu 比 sigmoid 好在哪?
- skip connection 有什么好处?
- mobileNet、shufflenet 的原理？说了下原理
- 为什么 mobileNet 在理论上速度很快，工程上并没有特别大的提升？ 先说了卷积源码上的实现，两个超大矩阵相乘，可能是 group 操作，是一些零散的卷积操作，速度会慢。
- 说下 VGG ResNet 网络结构?
- 介绍一下各种轻量级网络?
- 讲一下 MobileNet 的原理？深度可分离卷积，从参数数量以及计算量角度与传统卷积对比分析。
- MobileNet 与 Xception 以及 ShuffleNet 的对比？是否测试过 MobileNet 在不同计算设备上的运行速度?

2.2.1.4 其他方面

- 简单谈了一下对梯度消失和如何防止过拟合的看法?

- CNN 为什么比 DNN 在图像识别上更好？.
- 神经网络的反向传播机制？pooling 和卷积。
- 解释一下 BN?为什么用 BN?BN 层，归一化后的操作?
- 梯度爆炸和梯度消失的原因?
- 神经网络为什么要加 sigmod 函数?
- 如何防止过拟合?
- 初始学习率怎么设？这个没有总结过，只是说一般使用 0.01～0.1
- 迁移学习是什么？怎么迁移？怎么选择迁移的模型?
- softmax、多个 logistic 的各自的优势?
- 深度学习里面解决梯度消失的办法?
- CV 中做数据增强的方法?
- 工程上如何对卷积操作进行优化?
- 样本不均衡怎么处理？一个 batch 类别均等采样，修改 loss 对不同样本的权重。
- 学习率过大会出现什么问题，怎么解决?

2.2.2 数学计算

- 卷积时间复杂度？CNN 的卷积计算，参数计算?

2.3 深度学习：RNN 递归神经网络方面

2.3.1 讲解相关原理

- LSTM 的原理大概说一下，解决 RNN 的哪些问题。
- 对于深度学习，例如 LSTM 等的理解，还有它的改进 GRU，还有梯度消失、梯度爆炸等等
- BN 细节，有什么好处，能不能在 RNN 中用?
- LSTM 为什么可以缓解梯度消失
- 讲一下 LSTM 各个门?

- LSTM 原理，其中的参数是否相同?
- LSTM 的原理知道吗？LSTM 与传统 RNN 相比有什么优势?
- 为什么 LSTM 可以解决梯度消失的问题？LSTM 可以解决梯度爆炸吗?
- GRU 原理（几乎是要手写公式了）
- GRU 两个 gate 的作用分别是什么?
- GRU 和 transformer 各自的优势?
- LSTM/Bert 的结构，优劣势?
- 序列标注时数据量太少的时候怎么办?
- LSTM 相比 RNN 优点缺点?
- 讲一下 LSTM，LSTM 相对于 RNN 有哪些改进？LSTM 为什么可以解决长期问题，相对与 RNN 改进在哪？梯度消失和梯度爆炸？LSTM 如何解决梯度消失的问题?

2.4 深度学习：CNN&RNN 常问通用知识点

2.4.1 基础知识点

- 解释下 CNN 与 RNN 的区别?
- 对比讲了 cnn,rnn 和 lstm,并讲了 transformer 相对于他们的优点，transformer 有啥缺点，transformer 里面的两种 mask 操作，反正问了很多 transformer 里面具体的实现细节?
- 为什么要用 CNN，Bi-LSTM？如何用 Attention?
- Attention 原理?

主要讲的是 Transformer 中 Multi-Head Scaled Dot-Product Attention。注意，这里有一个 Mask Attention 机制，它对于 Transformer Decoder 和 XLNet 的实现原理非常重要，同学们如果了解相关知识点，一定要对这个 Mask Attention 知识点进行深入的理解。

- Multi-Head Attention 中如何优化 Muti-Head 的计算?

没有相关底层优化经验，所以回答：借助 CNN 底层计算原理，将多头变换展开为二维矩阵（填充大量 0），将多头变换转为矩阵乘法运算。

2.4.2 模型评价

- logloss 和 auc 的区别、为什么业务中喜欢用 auc?
- 二分类的评价指标都有哪些?
- acc 与 auc 的选择?
- 如果线下 auc 很高，线上各项指标都不好，可能是因为什么，怎么解决
- AUC 的计算?
- 准确率和召回率的概念
- 了解统计学的指标吗? AUC,ROC,F1 都是干嘛的
- 混淆矩阵角度解读召回率和准确率

2.5 机器学习方面

2.5.1 讲解相关原理

2.5.1.1 数据准备

- 常用采样的方法?
- MCMC 采样?

2.5.1.2 特征工程

① 特征降维

- PCA 的主成分是怎么得到的?
- 特征变换做什么? 特征处理?
- svd 怎么实现图像降维，怎么确定不会影响训练效果?
- LDA 知道不? 说一下 LDA 的原理?
- PCA 原理及涉及的公式?

② 特征选择

- 特征相关怎么处理? 好几个特征都相关怎么处理?

- 高维数据，其中有一维是时间，有缺失，如何处理？
- 如果选出好特征，去掉不好的特征？
- 特征工程中具体衍生出来的特征进行了详细的询问，为什么要生成这样的特征，依据是什么，为什么要使用 PCA 进行降维，如何存在多个特征高度共线会有什么问题？
- 机器学习中的特征，是选择重要的特征，还是特征的相互关联？
- 对相似度的理解？如何进行特征筛选？如何衡量特征之间的相关性？
- 如果要用树模型的话，可以做哪些特征工程？（n-gram，tf-idf，w2v）
- 假如说句子长度差别很大的话，tf-idf 这个指标会有什么问题？one-hot encoding 这个指标又会有什么问题？
- 树模型怎么分裂？怎么理解信息增益？

2.5.1.3 有监督学习-分类和回归方面

① 分类回归树（集成学习）

- 集成学习分为几大类，rf 和 bagging 的区别？
- 从方差和偏差角度比较 bagging 和 boosting？
- boosting 和 bagging 的差异？两者的思想说一下？bagging 里面树的深度和 boosting 里面的不同，为什么？

A.基于 bagging：随机森林

- bagging 的思想是什么，本质是什么；
- 随机森林里面的两个随机，随机森林为什么是减小方差 ？和 Adaboost 的区别？
- 随机森林与决策树关系，防止过拟合的原理，随机性的体现。
- 随机森林如何选择 feature？
- 从方差偏差的角度解释 bagging？为什么随机森林泛化能力强？
- dropout 是否了解？随机森林是否也可以用 dropout？
- 随机森林的随机体现在哪里？样本/特征随机采样的目的是什么？

- 随机森林和 GBDT 的区别和联系
- 什么样的数据你会选择使用随机森林
- 随机森林相比决策树的优点有什么
- 随机森林怎么提高泛化能力的?
- 随机森林原理，和决策树有什么区别，追问投票是怎么投票的?

B.基于 boosting：Adaboost、GDBT、XGBoost

- 用到了哪些模型（LGB+XGB 的 bagging），为何？有什么好处？LGB 里面你用到了哪些参数？你怎么调的参数?
- GBDT、XGBoost、LightGBM 三个算法的原理和区别？XGBoost 和 GBDT 哪个模型性能好，为什么?
- RF、GBDT、XGBoost、AdaBoost 的区别?
- GBDT+LR 原理
- 讲一下 boosting 算法？说一下 adaboost 是怎么更新的?
- 介绍XGB对GBDT的提升，LGB对XGB的提升，以及既然使用了LGB为什么还要使用XGB?
- XGB 如何处理缺失值，LGB 的差加速和直方图算法的底层代码是否有过了解
- XGBoost 的优化点，与传统 GDBT 的区别?
- GBDT 的特征组合原理?
- 问了 gbdt 在分类时节点输出是什么？怎么拟合残差?
- XGB 为什么要做泰勒展开？正则项的内容？为什么要拟合二阶梯度?
- XGboost 的底层算法是什么（CART 树）
- Xgboost 的应该着重调哪些参数?
- 介绍随机森林和 GBDT 的区别，为什么 Bagging 降方差，Boosting 降偏差
- 讲一下 GBDT，Gradient 代表什么?

② 线性回归

- 线性回归和逻辑回归关系，区别?

③ K 近邻（KNN）

- KNN 复杂度高，怎么解决?

④ 逻辑回归 LR

- LR 的原理，LR 参数的意义？LR 的损失函数，怎么求解.
- 为什么 LR 的目标函数是最大化似然函数?
- LR 加上正则化项后怎么求解?
- LR 如何引入非线性?
- LR 的损失函数是什么？lr 为什么不用 min square loss？它的导数是啥？加了正则化之后它的导数又是啥?
- LR 里面，损失函数能不能把交叉熵换成 MSE?
- LR 和 SVM 的时间代价比较过没有。
- 问我逻辑回归中 sigmoid 函数的好处？以及为什么用极大似然?
- 逻辑回归的思想和过程，损失函数是什么，如何训练得到最优参数
- 说说逻辑回归吧，适用于什么场景呢，和 knn 区别?

⑤ SVM（支持向量机）

- 详细说一下 LR，SVM 和 DT 的原理
- 讲一下 LR、SVM、XGBoost 模型的区别
- 讲一下 SVM 的原理，核函数，以及和 LR 的区别，哪个是参数模型?
- SVM 目标函数，为什么转为对偶，SVM 的核函数的本质是什么?
- SVM 对偶问题介绍一下，从函数间隔，几何间隔开始介绍。
- SVM 优化的目标是啥？问了 SVM 推导以及拉格朗日对偶法，从数学角度来说明
- SVM 当线性不可分的时候怎么办？(楼主答用核函数升维)
- 说一下 SVM 适合什么场景？或者说有什么限制?
- SVM 的 KTT 条件?

- 知道哪几种核函数？介绍一下高斯核函数？
- 核函数的作用，核函数为什么有用？从数学角度说明

⑥ 朴素贝叶斯（Naive Bayes）

- 朴素贝叶斯的原理？
- 最大似然估计和贝叶斯估计的联系和区别
- 贝叶斯估计和似然估计的区别

⑦ 决策树（DT）

- 决策树的原理，前后剪枝，评价指标。
- 决策树分裂节点的选择？
- 各种决策树:id3 c4.5 cart
- 朴素贝叶斯和决策树的差别，各有什么缺点？再加上 SVM 呢？
- rf 和 gbdt 基分类器区别，里面的决策树分别长啥样，怎么剪枝
- 决策树的一下细节，GBDT，各种熵的计算，
- 决策树相比其他算法有什么优势？
- 决策树中有哪些参数，如何避免决策树的过拟合
- 决策树具体讲讲，原理，如何选取特征的，怎么进行分类预测的？

⑧ 其他

- 说一下生成式模型？生成式模型和判别式模型有什么特点？
- 生成式和判别式模型哪个使用极大似然估计？

2.5.1.4 无监督学习-聚类方面

- 非监督学习有哪些？ 介绍了 Kmeans，k-means 的时间复杂度
- 怎么选取 K 值？ 手肘法
- K-means 的缺点？
- 如果没有先验知识，如何确定 K-means 的参数

- 衡量 K-means 效果好坏的方法
- kmeans 原理，优化目标是什么?
- k-means 和 knn 有什么关系，区别?
- k-means 聚类效果会不会因为初始选取点不同，聚类效果完全不一样？k-means 聚类效果和 k 的大小有什么关系?
- 简述 kmeans 算法，有什么缺点，如何改进，kmeans++是如何改进的?

2.5.1.5 模型评价

- 如何选择模型？从数据量，特征量方面分析了一遍

2.5.2 手写算法及代码

2.5.2.1 手写公式

- 贝叶斯公式知道吗，什么含义?
- LR 是否知道，讲一下数学原理（公式层面）
- 推导一下 SVM?

2.5.2.2 手写代码

- 手写一个 kmeans
- 求两个点的欧几里得距离？并实现一个 kmeans?

2.6 深度学习&机器学习面经通用知识点

2.6.1 损失函数方面

- 讲一下交叉熵的公式和意义
- 为什么分类问题的损失函数用交叉熵?

2.6.2 激活函数方面

- 激活函数作用，有哪些，都怎么改进?

- ReLu 函数所有背景、原理、应用?
- RELU 和 Sigmoid 相比，优点有哪些？ReLU 解决了什么问题?

2.6.3 网络优化梯度下降方面

- 梯度下降、牛顿法优缺点?
- 梯度下降和随机梯度下降讲一下?
- SGD,ADAM 区别
- SGD 和 BGD 区别，还知道哪些优化算法？动量的作用是什么?
- 梯度下降过程中如果不按正确的方向进行怎么办?
- 用什么优化方法，梯度下降的种类，各有什么优点
- 梯度下降法和牛顿法是如何实现的。优化问题分哪些种，无约束的优化问题怎么处理，有约束的优化问题怎么处理。
- 凸优化了解吗？ 牛顿法、SGD、最小二乘法，各自的优势，牛顿法与 SGD 的区别？牛顿法能用于非凸函数吗？拟牛顿法能说说吗?
- 说一下牛顿法，为什么深度学习很少用牛顿法？牛顿法一般应用于什么场景，有什么好处?

2.6.4 正则化方面

- L1，L2 正则化的原理讲一下?
- L1 和 L2 的区别，数学上解释（等高线）
- L1 正则不是连续可导的，那么还能用梯度下降么，如果不能的话如何优化求解?
- 正则化有哪几种，分别有什么作用?
- L1 和 L2 的区别？从贝叶斯估计的角度看？它们的先验分布是什么?
- 问了 l1 正则 产生稀疏解的原因（一画图准备扯拉普拉斯分布 0 点值就把我打住了，知道我想说什么），不可导点的处理（说了三种方法，问了不了解针对大规模数据的方法，这些实际都不常用）
- L1、L2 正则化的区别和应用场景？你知道哪些激活函数？简单说说区别？为什么需要激活

函数，它解决了一些什么问题？为什么 ReLU 比 sigmoid 更能解决梯度消失的问题？

2.6.5 压缩&剪枝&量化&加速

- 剪枝与正则化的联系，笔者从结构化剪枝与非结构化剪枝分别对应 Lasso 和 Group Lasso 的角度来回答
- 结构化剪枝和非结构化剪枝
- 三大角度：蒸馏，剪枝，量化。笔者分别介绍了三大角度的基本原理。

2.6.6 过拟合&欠拟合方面

- 如何检验过拟合，数据量很小怎么办？

就项目中数据处理方式做了详细的询问，生成的多张数据集如何使用，缺失值的处理需要考察到哪些问题，均值填充是否科学等

- 如何防止过拟合？为什么会过拟合？（从数据、模型、指标三个角度，提到了 dropout、正则，后面正好顺着问。）
- dropout 是否了解？讲一下 dropout 原理，为什么能防止过拟合？对训练数据和预测数据有什么区别？
- 神经网络怎么避免过拟合？

2.6.7 其他方面

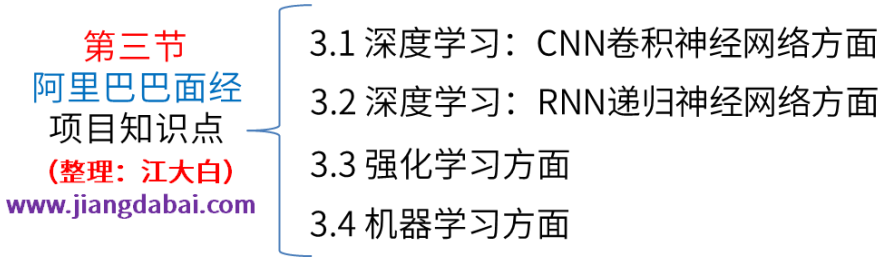
- 监督学习，无监督学习区别，半监督是什么？
- 怎么做数据增强(结构化数据，图像，文本）
- 为什么会产生梯度震荡、学习率是干嘛的
- 怎么处理非平衡问题（除了我说的欠采样过采样，小哥哥说可以对损失函数进行改进，比如令正类的损失函数为 1，负类的损失函数为 9）
- 数据不平衡问题？

从欠采样过采样等经典解决办法的角度回答。另外回答了一些其他方法：GAN（ICCV 2019 best paper：SinGAN），特征空间增广，改进训练方式（源数据训练特征提取 backbone，欠采样或

过采样训练分类器），Loss 加权，使用 AdaGrad 优化器等。

- 如何可视化和理解你的模型？（遮挡实验，attention score）
- 问平常训练模型有没有遇到什么问题，说了显存和 batch-size,然后面试官一路问到底，就各种问题如何解决，有个地方我说到了静态 rnn 和 dynamic rnn,例如如何提高显存利用率，一个模型如何发现性能瓶颈在哪
- 如何降低模型复杂度
- 如果数据不充足，或者说非常不平均，要怎么解决？从数据增强和建模来讲
- 偏差和方差的区别？
- Graph Embedding 和 GNN 的区别？

3 阿里巴巴面经涉及项目知识点



3.1 深度学习：CNN 卷积神经网络方面

3.1.1 目标检测方面

3.1.1.1 讲解原理

- faster rcnn 的 RPN 和 ROI pooling？
- 单阶段和两阶段的优缺点？
- 非极大值抑制？
- 交通标志检测和识别的关键？
- 基于 SSD 的，问训练阶段 SSD 的 gt 怎么和预测框发生联系。
- 在目标检测的基础上讲了讲跟踪方面的项目，深挖了几个点比如 deformable CNN 的具体

实现，Siamese-RPN 的具体实现

- 如何解决小目标检测的问题?
- 为什么要深层、浅层 featureMap concat？提了点细节和我踩的坑，需要数量级上的调整，不然深层的 feature 可能会被压制。
- Cascade 的思想？说了下我的摸索的一个过程。改变样本分布，困难样本挖掘，能达到比较好的效果。
- 介绍网络：Faster-RCNN、YOLO、SSD、YOLOv1、YOLOv2、YOLOv3、Masker-RCNN、GAN
- CenterNet 的实现细节（argmax）
- RCNN、Fast RCNN 和 Faster RCNN 的区别
- 为什么选择 RetinaNet
- 特征金字塔 FPN 的作用?

3.1.1.2 损失函数

- Focal loss 原理?

3.1.2 OCR

- 就实习阶段所做的超分辨率算法工作进行了详细的询问：数据如何生成，从概率的角度解释网络为何能够学到 LR 和 SR 的映射关系，如何搭建和训练网络，如何解决模型落地问题
- 了解到答主在做超分时遇到的问题后，对业界前沿的技术做了相关询问，用了哪些 GAN 模型，GAN 模型的 loss 函数如何设计，为什么这么设计?

3.1.3 图像分类

- 分类器有了解吗？对哪些分类算法有研究?

3.2 深度学习：RNN 递归神经网络方面

3.2.1 自然语言处理 NLP

① Bert

- Bert 详解？损失函数？bert 的 mask 相对于 CBOW 有什么相同与不同
- bert 当时是用预训练模型还是自己重新训练的？用 bert 来干嘛了？
- 了解过 BERT 吗，里面的三种 embedding 分别是什么，为什么要这样做？
- Bert 和 Elmo 在工程中存在的一些 Trick？

② Transformer

- Transformer 模型架构说一下？ 为什么用 transformer，不用 rnn 和 LSTM、transformer 的优势是什么 ？
- Transformer 和 BERT 的位置编码有什么区别？
- Transformer 用的 Layer Normalize 还是 Batch Normalize？Layer，有什么区别？...
- 用 python 写一个 multi-head attention

③ CRF

- CRF 的作用？维特比详细过程

④ HMM 隐马尔科夫模型

- 隐马原理？如何应用在分词当中的？

⑤ Word2vec

- 讲一下 word2vec，word2vec 网络模型？怎么训练？讲一下 word2vec 的霍夫曼树的原理

⑥ CNN 方面

- NLP 哪个模型最熟悉？Text-CNN，讲一下？
- TextCNN 模型背景、原理、应用；当时毕设/美团评论情感分析比赛的数据集、评价体系、任务要求？
- TextCNN 当时用了哪些卷积核？数值、尺寸？为什么用这种？

⑦ 其他

- 介绍预训练语言模型

ELMo，BERT，Transforler-XL，XLNET，ERNIE，RoBERTa，ALBERT，ELECTRA。。。笔者从 BERT 的 mask LM 以及 NSP 任务出发讲解了 BERT 后续各大预训练的改进。

各大预训练语言模型可能不能从头到尾讲起，笔者线是介绍了 BERT，然后从 BERT 的预训练任务出发，比如介绍了 ERNIE 中对 mask LM 的改进，ALBERT 中将 NSP 任务替换为 SOP 任务等。

- CNN，RNN 在处理文本上有什么区别?
- CNN 在文本分类上的应用与什么比较像？（意思就是卷积核的作用与什么的作用很像，答案是 n-gram，没答上来，我说了一下在文本分类中卷积核是怎么运作的）
- Dropout 有什么作用？类似于 Bagging 。在 Transformer 模型中 dropout 主要用在哪里?
- 如何衡量两个句子的相似度，sentence embedding 的方法
- 从长文本到短文本的生成 怎么做
- 生成文本或者说生成模型，最主要的因素是什么
- 对 NLP 的理解，讲了一下文本分类的发展史，主流分类方法的发展
- 说一说 nlp 的基本预训练模型（说了 ELMO、GPT、BERT），然后这些预训练模型有什么特点?
- 说一个熟悉的文字识别模型（CRNN 的结构、CTC-Loss）
- 怎么处理非平衡？欠采样的时候其实可以考虑文本相似度，了解怎么做文本相似度吗?
- 讲一下训练词向量的方法（w2v，skip-gram，CBOW，glove）
- 文本里面的“服务”、“食物好吃”等，如何抽取作为特征？（这问题当时理解有歧义，所以了有点久才知道问的，词性关联模板）
- 如何判断一段文字的时效性?

3.3 强化学习方面

3.3.1 讲解原理

- 你觉得强化学习和推荐有那些关联?
- 强化学习模型 和 CTR 预估模型的区别?
- 强化学习有哪些评价指标?
- gan 中的转置卷积
- 论文相关，gan 的对抗样本对判别器有什么效果， gan 生成的样本只考虑了一部分特征，如何考虑业务数据下游的下游特征，他们如何增强?
- gan 在淘宝场景中的应用（从数据，样本，模型，评估层面分析）
- 面试官在手淘那边是负责做推荐搜索的，而我的方向是做强化学习的，而用强化学习做推荐搜索很可能是近几年的一个趋势，让我用强化学习对推荐搜索进行建模，包括 state、action、agent 的选取，reward 的设计，以及如何训练。以及和目前现有的推荐搜索技术相比，用强化学习做有什么优势呢。
- 介绍强化学习都有哪些方法?
- off policy 和 on policy 都有哪些应用场景? 区别是什么?

3.3.2 损失函数

- 强化学习 loss 函数说一下?

3.4 机器学习方面

3.4.1 推荐系统

- 介绍下基本的推荐算法
- CTR 比赛中如何做的特征?
- 说一下协同过滤公式，两种协同过滤额应用场景有啥不同?
- user 向量和 item 向量，协同过滤，Neural CF

- user 的嵌入向量怎么得到的?
- 冷启动：用户很少交互怎么解决？商品数目很多训练嵌入向量怎么训练?
- CTR 预估中，如果有两个模型 CTR/CVR, 怎么做最后的 item 排序。猜测是问 ensemble 方法，回答了 bagging，然后让我具体描述一下思路。

召回策略的方法： CF -> Nerual CF -> Youtube DNN

讲一下 CF 的思路，UserCF, ItemCF

- 假如说处理一个多级分类的问题，有没有什么办法只用一个模型？-multitask

如果说 multitask 的输出 y 之间互相有制约关系，要怎么处理？（之前还不知道有 CRF 这个东西，就说了如果输出的 y 之间冲突了就引入一个 loss。面试官说看来你还不知道 CRF，下去好好学习一下）

- 问了 netflix 的电影评分预测 讲了怎么基于矩阵分解求出评分 svd 的 k 值代表什么？怎么确定?
- 问了 deepfm，讲了 FM 的原理，怎么缩减计算量
- point-wise，pair-wise, list-wise 的优缺点，对这些 loss 的常用设计形式了解吗?
- 对召回算法有了解吗？常用的召回算法的优缺点?
- CTR 中为什么经常用 LR?
- 介绍 FM FFM 原理
- FM 推导
- 召回和精排的区别以及各自的特点?
- 召回和精排的负样本有何不同?
- 精排特征有哪些，点击序列是怎么作为特征放进去的?

4 数据结构与算法分析相关知识点

第四节
阿里巴巴面经
数据结构与算法分析
（整理：江大白）
www.jiangdabai.com

- 4.1 数据结构与算法分析：线性表、属、散列表、图等
- 4.2 算法思想实战及智力题
- 4.3 其他方面：数论、计算几何、矩阵运算等
- 4.4 Leetcode&剑指offer原题

4.1 数据结构与算法分析

4.1.1 线性表

4.1.1.1 数组

- 2*N 数组，将奇数放到奇数位置，偶数放到偶数位置
- 给定一个长度为 n 的数组，如【12123】，相邻值为 1、-1，现在有一个值 x=100，查找 x 的位置?
- 给一个数组，如何判断这是数组是不是一棵树的后序遍历？后序遍历有什么特点?
- 一个 1000W 的 64bit 整型数组，无需、可重复，找第 100W 的数字。 【快排思想（非快排）O(2n) > 堆排 O(n logk) > 快排 O(n logn)】
- 一个数组按从大到小排列，但是有重复的元素，利用二分查找查找到指定的元素，如果有多个就返回最大的那个索引。我先写了递归算法，然后他让我非递归写一下。
- 如何在 n 个数组中找出它的中位数（n 个数组无法完全放在内存中）
- 求一个数组的连续区间，使得和最大?
- 两个有序数组怎样最快的确定差集，复杂度是多少?
- 两个无序数组去重复
- 给定一个数组，求最大连乘子区间（可以包含小数、负数）
- 找到一个无序数组里面连续的最长整数数组长度。顺带考察了基数排序和快速排序
- 链表和数组的区别?

- 可重复排列，数组第 k 个数求时间复杂度？

4.1.1.2 链表

- 说一下链表反转？
- 把一个链表转化成平衡二叉树
- 说一下链表操作的时间复杂度：查找、查询、删除、增加。和顺序表的区别？
- 链表插入的复杂度什么时候是 O(1)？
- 链表的插入和查找的复杂度
- 2*N 数组，将奇数放到奇数位置，偶数放到偶数位置

4.1.1.3 字符串

- 一个是单词级别的翻转字符串，比如“I love you”翻转成“you love I”？
- 反转字符串 in place
- 把一个字符串的小写字母放到前面，大写放到后面，保持原有的顺序？
- 两个字符串的最小距离（插入，删除，改变一个字符）说一下思路和复杂度？
- 一个求最长不重复字符串长度的代码题？
- 给定一个字符串（例如 abc），和一个文档，给出每个字符在文档中的倒排索引，在文档中找到一个最小窗口（adebdc），使给定字符串是其子串。（解法：使用归并思想，从第一个字符（a）的倒排索引开始，找仅比当前索引大的第二个字符（b）的索引，直到最后一个字符，计算窗口大小，保留最小值，时间复杂度：各字符倒排索引数组大小相加（线性），空间复杂度，O(1)？
- 从字符串中提取所有有效的 ip 地址？
- 给定一个字符串，和字符串列表，判断能否用字符串列表拼接生成字符串？
- 两个字符串的公共子串，动态规划？

4.1.2 树

4.1.2.1 二叉树

- 求解二叉树的最大高度，用一个非递归的做法
- 二叉树最长叶子结点路径
- 如何判断一棵树是另一棵树的子树?
- 如何判断一棵树是不是平衡二叉树?
- 开发一颗字典树，实现建树和搜索功能
- 有一个词库，如何像纸制字典一样建立索引？ (B+树)
- 红黑树了解吗，介绍一下?
- 红黑树的查询复杂度?
- 二叉平衡查找树怎么实现平衡?
- 给出二叉树的前序和中序序列恢复二叉树的结构
- 不用递归遍历一棵树
- 二叉树层次遍历
- 给出二叉树的前序遍历和中序遍历结果，构建这棵树
- 描述一下二叉搜索树?
- 时间复杂度为 O(n)、空间复杂度为 O(k)的树的搜索方法

4.1.2.2 堆

- 什么是堆，构建堆的复杂度，堆找出第 k 大元素的复杂度?
- 描述一下堆排序、什么是大顶堆、什么是小顶堆
- 一个数组找 top-K，堆遍历数组。（k 最大）为啥用小顶堆。大顶堆小顶堆的区别?

4.1.3 排序

- 有哪些排序算法，他们的时间复杂度是多少

- 海量数据前 k 大
- 只给一分钟思考口述如何在时间复杂度最低的情况下找到无序数组中的第 k 个数。（快排+剪枝）
- 一亿个浮点数，大小不超过 2^32，均匀分布在值域内，求最快的排序方法；分析排序方法的复杂度（面试官全程提示）
- 手写快排

4.1.4 搜索

- 题目：最少新建道路条数

  已知有 N 个城市，城市编号 1...N

  已知 M 条路，每条路表示为(x_i, y_i), xy 分别为城市编号

  现在需要新建道路，确保任意两个城市之间，是可以通过一条或者多条道路联通

  求解能够达到此目的的最小道路条数

思路：BFS 或者 DFS 应该都可以找到城市簇即可

4.2 算法思想实战及智力题

4.2.1 算法思想实战

- 一条无限长的 x 轴，假设你站在原点，可以向左走也可以向右走，第 n 次走 n 步或者-n 步，给一个 target，问走到 target 的最小次数
- 一个黑盒里有 n 个球，球分为三种颜色，RGB，乱序，每种颜色的球 n/3 个。这个黑盒有两个接口，一个接口可以获取第 i 个位置的球颜色，一个接口可以交换两个位置的球。通过这两个接口将球排序成 RGBRGB 这种的顺序。

面试官先上让我不考虑时间复杂度说出一个解法，我说了一个 O(n^2)的解法。

然后他让我想想有没有更好的解法，我想了大概 5 分钟，恍然大悟，说了一个 O(n)的解法

- 一个快递员送快递，有 n 个城市，怎么选择路线，使得走的路程最短。
- 给一个数，怎么快速求这个数的二进制中有多少个 1?

- 实现一个数据结构，满足 set(index,value),get(index),setall(value)的操作尽可能高效，使得 get(index)返回正确的结果。

4.2.2 智力题

- 有一座桥，A 通过需要 25 分钟，B 通过需要 20 分钟，C 通过需要 10 分钟，D 通过需要 5 分钟，一个桥同时只能走两人，且快的人需要等慢的人到达才能一起到达。走桥时必须要有手电筒才能经过，且手电筒只有一个，问如何在 60 分钟内使得四人均通过?
- 长 m，宽 n 的长方形，每长度为 1 画一条线，问可以找到多少正方形（m*n+(m-1)(n-1)+(m-2)(n-2)+...+(m-n+1)*1）？多少长方形（C_{m+1}^2 * C_{n+1}^2，即在 m+1 个点中随机选 2 个，在 n+1 个点中随机选 2 个）？
- 两个人 A,B；A 红绿色盲，B 声称自己不是红绿色盲，现在 B 两只手上分别有红绿两颗球。

问 A 怎么分辨出 B 是不是红绿色盲。

- 你有两根均匀的绳子，和一个无限燃料的打火机。每根绳子若点燃一头可燃烧 1 分钟。问如何用这些东西准确测量出 45 秒的时间?
- 在面试官快要不耐烦的时候想出来了正解...绳 A 点燃两头，绳 B 点燃一头。在绳 A 燃尽时点燃绳 B 的另一头，两根绳子全部燃尽总耗时 45 秒。
- 8 个小球称重；如果不知道次品轻重怎么办?

4.3 其他方面

4.3.1 数论

- 在 A 地有两辆公交车，一辆间隔 5 分钟，一辆间隔 7 分钟，问等车时间的期望。
- 掷骰子，问掷到 6 个面全出现的期望?

4.3.2 计算几何

- 拉格朗日乘子法，hessian 矩阵

4.3.3 概率分析

- 长度为 1 的线段，随机地取两点 A 和 B，求 AB 长度的概率密度函数?
- 8 个球有一个重一点，最少称几次能找出来，我说 3 次，后来鼓起勇气问他要几次，他说两次，让我自己想怎么称?
- 给定随机函数 R，以 p 概率产生 1,1-p 产生 0。生成随机函数 R'，以 1/2 概率产生 0,1?
- 54 张扑克牌，分成三等份，大小王在同一组的概率?
- 求一根绳子被切两刀能组成一个三角形的概率。
- 有一苹果，两个人抛硬币来决定谁吃，先抛到正面的先吃，问先抛者吃到苹果的概率 ？
- 0 1 2 3 4 5 6 7 8 9 下面写一个数，使得下面这个数刚好是这个数字在下面一行出现的次数
- 投篮命中率 10%，那么 a.投 10 次中 1 次 b.投 10000 次中<1000 次，哪个概率大?
- 一段绳子分三段，组成三角形概率?
- 棋子在规定走法，规定大小的棋盘上，N 步后还在棋盘上的概率，主要考察动态规划
- 一个 NxN 的棋盘，一个棋子可以等概率地跳八个方向（和象棋中马一样的跳法）。当这个棋子跳出棋盘范围的时候，就停止。问棋子跳了 k 步之后，棋子还留在棋盘的概率。
- 学校男生的概率 2/3，女生的概率 1/3，男生穿牛仔的概率 2/3，女生穿牛仔的概率 1/3，你看到一个穿牛仔的，问他是男生的概率是?
- 六边形，顶点上各有 6 只毛毛虫，可以沿着边走，两个毛毛虫相遇的概率，n 边形呢
- 有 A、B 两枚不同的硬币，它们正面朝上的概率不一定是 0.5，且两枚硬币正面朝上的概率不一定相同。现在做 1000 次这样的实验：从两枚硬币中随机抽一枚抛一下，记录下正面，重复 100 次/10 次。问如何通过这 1000 次实验的结果求出 A、B 两枚硬币正面向上的概率?
- 一个总数很大的数据流，希望以等概率的方式进行样本数为一的抽样，怎么做?
- 将一根木棍分成三段，求这三段构成三角形的概率?

4.3.4 矩阵运算

- 一个 m*n 矩阵，从左往右逐步递增，下一行比上一行数都打，查找某一个值是否出现?

- 说一下矩阵的秩?再说一下矩阵的特征值和特征向量?(数学专业背景)

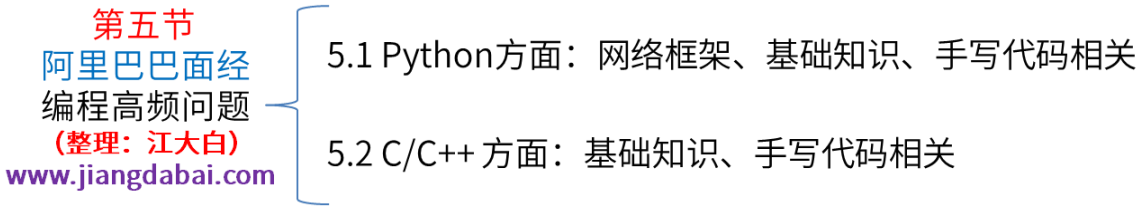
4.3.5 其他

- 贪心和 DP 介绍一下?
- DP 的一般做法流程?
- 布隆过滤器知道吗？用在什么场景下？推导会么（加分项）
- 线性回归多变量求解的过程，为什么这样求解？这样求解为什么是最优解？
- 纳什均衡看过吗?

4.4 Leetcode&剑指 offer 原题

- Leetcode 原题：接雨水
- Leetcode 23：合并 K 个升序链表
- Leetcode 236 ：判断围棋死活
- Leetcode139：单词拆分
- Leetcode 382：链表随机节点，并口述蓄水池采样算法的推导

5 编程高频问题：Python&C/C++方面



5.1 python 方面

5.1.1 网络框架方面

5.1.1.1 Pytorch 相关

- tensorflow 和 torch 的区别?

5.1.1.2 Tensorflow 相关

- Tensorflow 如何读取数据?
- TensorFlow 的参数初始化机制?
- Tensorflow 写一个全连接层的代码?

5.1.1.3 Caffe 相关

- 对 caffe 源码熟悉程度。(我扯了扯源码的底层设计模式，数据流怎么流的，如何添加新层、cuda 代码的细节)

5.1.2 基础知识方面

5.1.2.1 线程相关

- Pyhton 线程和协程的区别?

5.1.2.2 内存相关

- python 语言怎么处理内存溢出的情况，怎么设计内存回收?
- Python 的内存管理
- Python 垃圾回收机制

5.1.2.3 区别比较

- 说说 list 和 tuple 的区别? list 和 set 的区别，装饰器
- 生成器、迭代器的区别?
- copy，deepcopy，赋值的区别?

5.1.2.4 讲解原理

- Python 里面的字典的 key 可以用 list 吗？可以用 tuple 吗？可以用 set 吗？为什么？从底层实现原理说一下?
- Python 里面的循环很慢，为什么?
- Python 怎么生成一个迭代器?
- 讲一下 yield 关键字？它的作用是啥?

- 什么是装饰器?
- 说一说 python 重载(面像对象)

5.1.2.5 讲解应用

- Python、C++、Java 哪个用的多一点？值传递和引用传递区别。

5.1.3 手写代码方面

- Python 字典的删除实现

5.2 C/C++方面

5.2.1 基础知识

5.2.1.1 线程相关

- C++多线程熟不，有什么库？你使用过什么库?

5.2.1.2 内存相关

- C++中的内存泄漏是怎么发生的?
- 如何避免 C++中发生内存泄漏?

5.2.1.3 区别比较

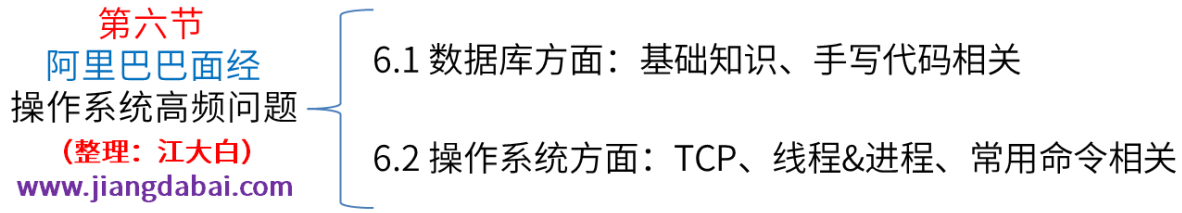
- C 的动态库和静态库?
- C++中引用和指针的区别?
- C++ operator new 和 new operator 的区别
- C++的继承和 Java 继承的区别?
- 接口和类的区别?
- C++、Java、Python 的主要区别 (编译型语言和解释性语言)

5.2.1.4 讲解原理

- C 的相关特性，多态?
- C++中 map、hash_map 底层实现及增删改查的复杂度

- C++智能指针了解吗?

6 操作系统高频问题：数据库&线程&常用命令等



6.1 数据库方面

6.1.1 基础知识

- 数据库：什么是主键？主键的作用?

6.2 操作系统方面

6.2.1 TCP 协议相关

- TCP 与 UDP 的区别
- TCP/UDP 简单介绍一下，如何用 UDP 来实现 TCP

6.2.2 线程和进程相关

6.2.2.1 区别比较

- 线程和进程的区别?
- 进程、线程的区别和联系，线程共享哪些资源

6.2.2.2 讲解原理

- 线程进程是什么、什么关系、什么时候用线程，什么时候用进程?
- Linux 多个进程如何通信的?

6.2.2.3 讲解应用

- 详细讲讲线程进程的区别，还有一些具体场景下的变化

6.2.3 常用命令

- Linux 如何查看进程
- Linux 中查看进程状态和查看开放端口的命令
- 常见的操作指令，还有一些场景题：例如 shell 里面写查询一个服务器日志文件中访问最多的那个 ip

7 技术&产品&开放性问题

7.1 技术方面

- 问一个区域确定人口中心划区域用什么方法，开始我说用 k-means，然后他问我具体怎么实现。我说先设定一个 k 值，他问 k 值怎么定，拍脑门吗？不过在没有先验信息的情况下，确实没法选，就多试几个不行吗。然后我说那用密度聚类。那密度聚类的密度函数用什么？
- 由于拍摄焦距的原因，有些图片的前景很清晰，但是背景很模糊。导致分类的之后分成了不清晰的图片。你觉得有什么解决的办法？

学一个 mask 的网络，把图像中清晰的部分扣出来。再去分类。

我说了用一些传统的手工特征如 SIFT，HOG 之类。去比较每帧之间的差异。

- 大量数据，亿为单位，找出与给定数据最相似的一个？
- 如何检测视频转场的时间，转场就是拍摄的场景变化了。

用光流算法？或者用 LSTM（这里问了很多，比如光流的好处，LSTM 的好处之类的，如何使用 LSTM）

- 给饭店确定菜系，是鲁菜、川菜、西餐、混合菜系。。。你需要收集哪些数据，用什么方法？
- 对一个数据，比如点击率，应该做哪些处理？
- 现在的搜索技术很少上深度学习或者说很深的网络，你觉得是为什么？如果要用深度学习，你觉得应该往哪些方向思考？deep learning 比较吃资源，不太适合业务规模比较大的系统，比如：双十一的压力，如果一定要用，可以考虑深度模型压缩，量化，矮胖网络的并行计算等方

向。

● 在一个坐标系内，用户和商户都有自己的坐标（x,y），那么我想找到距离用户最近的 k 个商户，如何最快的得到？

● 比如淘宝搜索时的自动补全该怎么做，用什么模型或者算法（之后查了一下用模糊匹配）

● 原始数据被污染了怎么办，这时候怎么判断是模型的问题还是数据的问题。

● 通过一个单目固定相机，如何获得室内桌子等物体的 3D 信息（回答的是 3D CNN + 卡尔曼滤波）

● 给了一个情景，如何训练模型、调优。（题目很空，主要考察你对深度学习的理解）

-根据需求（前向传播时间、模型大小），确定模型和基础网络，跑第一版模型。（举了个栗子）

-判断模型是否出现过拟合的情况，来决定下一步的优化方向。

-结果分析(confusionMatrix 等)，分析问题，将论文中的方法套上去，如果没有自己创造。（又举了个栗子）

● 设计一个情景，倾斜字体检测，问我有什么好的想法？（我觉得应该是他现在遇到的问题）

数据增强，加入形变扰动。

非 end-to-end 版本：分别训练检测和分类，举了之前做过的一个文字识别的项目的实现。

end-to-end 版本：加入仿射变换学习因子，学习字体倾斜的角度和形变。

● 如果让你设计一个推荐系统，你会设计一个什么样的架构？你设计的重点是什么？

● 在广告推荐的时候，我们常将展示出来但是用户没有点击的广告作为负样本，然而其实有的时候真实的情形是用户没有注意或者没有看见这个广告，

也就是说这条广告不是真正的负样本，用户对其可能是感兴趣的，这种情况该如何处理？

● 对抗学习有了解吗？你觉得该如何将 NLP 和推荐相互结合？

● 为什么人工智能在图像里应用落地更好，在 nlp 不行。谈谈你的看法。

● 双十一时向用户发放优惠券，希望在成本一定的前提下，使得盈利最大化，该如何建模发放给用户？用户无法做 AB 测试，该怎样划定正负样本？

● 模型上线时应该注意的事，如果请求过高模型服务挂了怎么办？

● 假设我们有很多数据给用户看到，但是我们只知道一部分的数据用户是否点击了，其他的我们都不知道用户是否点击了（相当于没有标签），如果是你的话，你怎么解决这个问题？

● 根据你的经验，你认为深度学习比传统机器学习好在哪里？（1. 特征工程：学习能力更强，同样多的数据，深度学习可以更准确地提取特征，而且深度学习可以自动提取特征，而传统机器学习很多需要手动提取特征 2. 比较灵活，我可以自己决定 CNN 有多少个卷积层，多少个池化层等等）

● 有没有非网络技术上的攻击，就是对模型本身的攻击？如果有，怎么应对？

● 怎么防止新来的错误数据对原有模型进行特别大的改变，怎么防数据污染，蓄意破坏？

● 双目相机识别目标深度的原理，以及 PnP 算法的原理

● 支付宝年末要出一个年终总结，那么我要对所有用户的交易额度进行全量的排序，那么内存肯定是不够用的，这种情况下应该怎么做？

● 输入是一个短文本，判断是否是服装相关内容

● 阿里有很多商品都是由一段话来描述的，现在让我来设计一个模型，输入是商品的描述，输出是描述商品的一句简短的话，要求用户看到这句话后尽可能保证不降低用户的体验（包括购买率、点击率、浏览量不要下降）

● 给定一堆商品 C，每个商品 c 对应有标题、描述，给定标签：时尚风、复古风、百搭风。设计技术方案，把每个商品 c 的主题标签给分类/聚类 等弄出来。没有训练数据，有哪些方案？

● 做分类的时候如何评估好坏呢？追问比如:你分类总数为 2，但是现在出现第 3 种类别，怎么办，例如做的分类是男女分类，现在有个中性的人，那你怎么办呢？如何应对这种情况？

● 场景题:如何从百万数据中找出最大的 k 个数？分治+堆排序

● 一个很大的日志文件里面存了各个 ip 访问的信息，几百 G，如何统计里面某个 ip 地址访问的次数？

答：不一次性读取，分批读，缓存，然后分别统计，最后加起来。

● 开放式探讨：如何使用强化学习去实现个性化商品推荐？

● 一个超级大文件，每一行有一个 ip 地址，内存有限，如何找出其中重复次数最多的 ip 地

址

7.2 产品方面

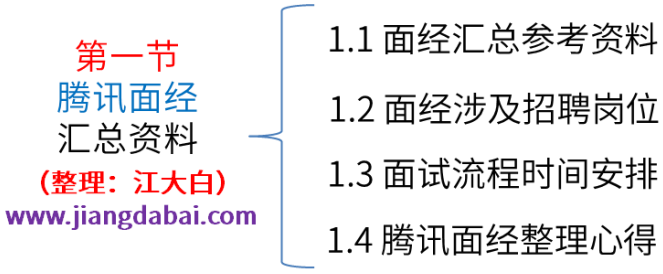
- 怎么给新上架的商品做冷启动?
- 口碑要拉新客，我们的策略是发红包，怎么如何在预算有限的情况下发红包能让最多的用户来安装口碑呢?
- 给你一些用户每天的相对位置信息，怎么区分它们的职业？怎么判断上线的现金贷产品的盈利能力?
- 天猫有 1000 万条数据，如何找到使用人数最多的前 1000 个?
- 给你淘宝的商品总量，怎么预测拼多多的商品总量
- 给出用户的数据和交易记录，如何判断是否给他开通花呗?
- 在有各种用户的数据，上网状态，手机型号，用户照片等数据的情况下，怎么判断支付是否存在欺诈?

7.3 开放性问题

- 当碰到难题时，团队士气低落的时候，作为团队的一员，该怎么去做?
- 如何把 AI 技术用到素材生成和智能分发场景中?
- AI 还有哪些能够应用到视频上的?
- 你觉得在公司或企业里，工程和算法的界限是什么?

3|腾讯算法岗武功秘籍

1 腾讯面经汇总资料



1.1 面经汇总参考资料

① 参考资料：

（1）牛客网：腾讯面经-212 篇，网页链接

（2）知乎面经：点击进入查看

（3）面试圈：点击进入查看

② 面经框架&答案&目录&心得：

（1）面经框架及参考答案：点击进入查看

（2）大厂目录及整理心得：点击进入查看

1.2 面经涉及招聘岗位

（1） 实习岗位类

【微信 nlp 应用研究实习】、【光子技术中心实习生】、【腾讯游戏算法实习】、【PCG 推荐岗实习】、【腾讯 AI 暑假实习岗】、【腾讯新闻 NLP 实习】、【腾讯音乐媒体研发中心算法实习】、【腾讯机器学习暑假实习】、【视频推荐算法实习】、【腾讯应用宝推荐算法实习】

（2） 全职岗位类

【omg 事业群计算机视觉工程师】、【腾讯优图算法工程师】、【SNG 云部门算法工程师】、【腾讯社交广告部算法工程师】、【IEG 安全部门算法工程】、【IEG（数据挖掘部）机器学习】、【腾讯游戏数据挖掘实现】、【微信事业部机器学习】、【PCG 机器学习】、【TEG 事业群机器学习岗】、【腾讯量子研究室】、【IEG 事业群光子工作室】、【腾讯短视频推荐算法工程师】、【wxg 推荐算法工程师】、【腾讯 AI Lab】、【腾讯北京应用研究路径规划和车辆流量监测】、【PCG 看点用户增长算法工程师】、【TEG 应用研究岗】、【PCG 应用研究岗】、【PCG 微视业务组】、【微信搜一搜算法工程师】、【PCG 腾讯视频搜索】、【腾讯云 NLP 算法工程师】、【CSIG 腾讯云算法工程师】、【腾讯动漫算法工程师】、【PCG 创新推荐算法工程师】、【机器学习技术研究算法工程师】、【WXG 开发平台基础部机器学习工程师】、【CSIG 云与智慧产业事业群 CV 算法工程师】、【后台策略安全岗机器学习工程师】、【QQ 音乐算法工程师】

1.3 面试流程时间安排

腾讯面试流程-整理：江大白			
	面试类型	面试流程	备注（侧重点）
第一面	技术面	自我介绍+项目/实习经验 +技术问答+算法编程	围绕简历上的项目问， 再考一些基础知识点
第二面	技术面	自我介绍+项目/实习经验 +技术问答+算法编程	简历上的项目扣得很细， 关注为什么？优点？
第三面	技术面	自我介绍+项目/实习经验 +技术问答+算法编程	围绕简历上的项目及基础提问
第四面	技术Leader面	自我介绍+项目经验+公司发展	考察抗压或者应急反应能力 以及解决问题的思路
第五面	HR面	基础人力问题	/

PS：以上流程为大白总结归纳所得，以供参考。

其他注意点：

- 实习岗是两面+HR，正式岗是四面+HR

● 第一面，也有人说是简历面，主要问基础

1.4 腾讯面经面试心得汇总

★ 关于项目的问题问的都很一针见血，面试官真的会很关注你简历上面写的东西，你写上去的一定要特别熟悉才行。

★ 感觉腾讯面试官更侧重应用场景，然后根据项目中的方法进行拓展问。

★ 面试的问题都是从简历内容出发延伸的，注意细节要完全弄清楚，面试官会问到底的。

★ 简单来说的话，可能对于有论文和比较好比赛成绩的同学，有专精方向的话，基础问的不会那么多，比如什么 lr 推导，svm 推导，bp 推导。

★ 腾讯的笔试题还是相对容易的，好好复习、认真做题应该问题不大。尤其最后的两三道编程题，其实腾讯出的都是常规题，只要数据结构和算法基础扎实，AC 两三道应该没问题。

★ 语言组织能力也很重要，逻辑能力好点，做过的事给面试官讲清楚。就算很水的项目，多介绍下原理，多说说自己的理解，多讲讲自己的改进，还是有很多谈资的。

★ 腾讯机器学习算法岗的面试算是非常正规的了，整套面试流程下来几乎能把你几年所学的东西都问到。所以，不要存在侥幸心理，踏踏实实的刷题，复习好常规机器学习算法，尤其是算法的原理和应用场景。

★ 项目和比赛经历非常的重要，往往面试官都是根据项目里用到的方法拓展提问，对项目的优化和改进也问的比较多。还有就是能内推的一定去找学长学姐或是其它资源去内推。

★ 面试过程中如果实在写不出来代码的话，就给面试官讲思路，尽量把自己的想法和思考过程表达出来。

★ 关于简历的技术经历的考察，面试官会问：你觉得简历里面的哪个项目做的最好？然后追问这个项目当中的技术细节，cnn，loss function，参考指标之类的。所以，建议大家把自己做过的项目好好的总结复盘一下，可以尝试着自己提问自己。

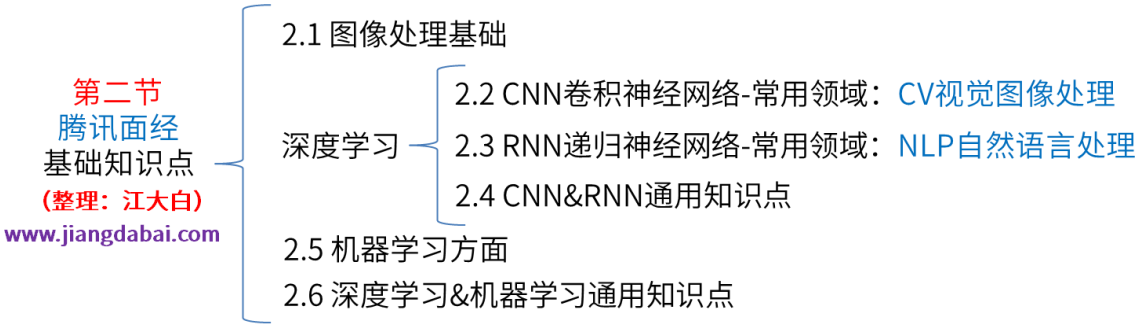
★ 找实习的话，由于校招同学没有工作经验，所以项目和实习是展示自身能力的最好亮点（一堆 CCF-A 的大佬请忽略并接受我的膝盖）。面试官可以从介绍中了解你的工程能力、抗压能力、沟通能力、思维方式等等，所以对于简历上的内容要滚瓜烂熟。

切忌在简历上洋洋洒洒地写"熟悉 Xgboost、SVM、Bayes、HMM、CRF、KNN、LR、CNN、Attention 等算法"。是听过名字就算熟悉？还是啃过源码？

而对于没有项目/实习/比赛经历的同学，这一部分时间只能由考察基础算法来弥补了，这样一来随机性、难度都提升了不少；所以平时尽可能多积累这方面的经验，让简历充实起来。

★ 腾讯一面难度中规中矩，量比较多但基本都在考察基础。

2 腾讯面经涉及基础知识点



2.1 图像处理基础

2.1.1 讲解相关原理

- Canny、Scharr、Sobel 边缘检测
- HOG 特征
- Hough 变换
- 图像匹配算法了解哪些？
- SIFT 特征了解吗？是怎么生成的？SIFT 特征是怎么进行匹配的？（OpenCV 中常用的是暴力匹配函数，即对两组 SIFT 特征向量，将一组中的每个向量依次与另一组的所有向量进行欧式距离度量，选择距离最近的作为匹配点）
- 如何根据局部特征去检索更大的图片？
- Opencv 如何读取数据流中的图片？如何生成图片？
- Opencv 中高斯滤波的函数是什么？

2.2 深度学习：CNN 卷积神经网络方面

2.2.1 讲解相关原理

2.2.1.1 卷积方面

- 谈谈 1*1 卷积？1*1 卷积有什么作用?
- 卷积原理，工程实现
- 为什么深层神经网络里的卷积核都改成了 3x3，而不是 5x5?
- 感受野的原理，CNN 的局限性?
- 为什么 Dropout 可以防止过拟合？随机失效会不会导致单次 Forward 的神经网络欠拟合?
- Dropout 作用与原理 ？dropout 是怎么调整参数?
- 神经网络利用 dropout 和多项式回归利用正则项减轻过拟合的本质是什么？（为什么就能减轻过拟合了）
- 介绍 Dropout 后，问缺点是什么？或者说，低层还是高层使用，为什么?

2.2.1.2 池化方面

- Pooling 种类，区别以及适用场景?
- Maxpool 和 Average Pool 哪个好?
- pooling 层前向计算、反向传播?

2.2.1.3 网络结构方面

- 写了 vgg-16 网络结构，问的还比较详细，又问我还用过其他什么网络没有?
- 图像分类 ResNet 和 DenseNet 原理？DenseNet 相比 ResNet 提升的效果怎么样？DenseNet 为什么比 ResNet 有更强的表达能力?
- 了解胶囊网络么?
- SENet 的 Squeeze-Excitation 结构是怎么实现的?
- mobileNet、shuffleNet 的结构知道吗?

- Resnet，讲一下结构，优点，为啥有效？为什么效果不会随着深度变差？
- Mobilnetv2 和 v1 的区别？
- 介绍一下图表示学习 GCN 以及 GCN 的缺点以及针对缺点前沿的扩展模型？
- VGG，ResNet 这种模型的优势的改进点？

2.2.1.4 其他方面

- 怎么解决图像细节不足问题的？（增强特征提取骨干网络的表达能力）
- BN 层怎么实现？有什么具体的好处？有哪些缺陷，怎么改进，了解其他的归一化算法吗？
- Batch Normalization 的原理及作用 （问的最多，感觉被问了至少 3 次）
- 介绍 BN 后：问 BN 有什么缺陷？在模型中用过吗？回答效果不太好。分析原因？
- 了解 Layer normalization 吗？（这个应该才是想问的问题...感觉上面的 BN 缺点都是铺垫）
- BN 每层的参数一样吗？
- 实际场景下做 softmax 容易出现一些问题，怎么解决（面试的时候没明白什么意思，面试结束后询问，他是说实际场景做 softmax 很容易出现下溢问题，这个可以用每个维度减去一个固定值就可以了）
- 普通的 DNN 网络如何设计的，隐层规模、隐藏单元，我说根据样本大小和输入的特征来设计。
- Batch size 如何选择？
- 梯度消失和梯度爆炸的原因和解决方法？

2.2.2 数学计算

- 输入是三通道的 WH 的图像，3×3 的卷积核，步长是 1，有 padding，输出是 512 通道的 feature maps，问输出的 feature map 大小和总共的卷积参数？
- 卷积输出大小的公式写一下？
- 输入 25*25*10 的 feature map， 用 3*3 卷积， 输出成 25*25*30， 问参数量？

2.2.3 公式推导

- Res 和 Densenet 中的融合有什么区别？用公式解释。
- DNN 反向传播公式推导?
- CNN 反向传播公式推导?

2.2.4 手写算法代码

- 定义函数实现一下 softmax（不要用 numpy） ，数值太大溢出怎么办?

2.2.5 激活函数类

- DNN 中如果把中间层的激活函数去掉会怎样？去掉激活函数的 DNN 与逻辑回归有什么区别?

2.3 深度学习：RNN 递归神经网络方面

2.3.1 讲解相关原理

- RNN 的梯度爆炸怎么解决?
- LSTM 对于 RNN 的改进地方在哪?解决了 RNN 的什么问题?(梯度弥散)
- LSTM 的结构，里面的遗忘门是一个数值还是向量?维度是多少?
- LSTM 为什么可以解决梯度弥散的问题?
- 为什么用 Relational-RNN 代替 LSTM?
- LSTM 对于 one-to-one，many-to-many 等这些场景如何选择模型?
- LSTM 里面有哪些门，为什么用这些门?
- LSTM 和 GRU 的区别，GRU 具体简化了哪个门?
- LSTM 与 RNN 的区别?
- LSTM 你用的优化方法是哪个，我说是 Adam，他问还有哪些，我说 SGD 等等，他问 GD 和 SGD 的区别?
- LSTM 里面为什么有些激活函数用 sigmoid，有些用 tanh?

- LSTM 门控机制是怎样的?LTSM 主要解决了 RNN 什么问题，为什么能解决?

2.3.2 手绘网络原理

- LSTM 能解决 rnn 什么问题，并通过数学公式推导?
- 手写 BiLSTM 的公式?

2.4 深度学习 CNN&RNN 通用的问题

2.4.1 基础知识点

- 了解 Attention 机制吗？Attention 的几种实现方式?
- CNN、LSTM 区别、文本里怎么应用?
- 讲一下对 CNN 和 RNN 的认识?
- 解释一下模型训练里面的偏差和方差对于训练的意义?
- CNN，RNN，Attention 在文本上怎么用?
- 实现一下 Attention，不准使用任何库函数?
- 一个二分类任务，假设只有一个维度的特征，取值范围是 0～正无穷，如何实现二分类?

2.4.2 模型评价

- 多分类有哪些评判指标，micro-f1 和 macro-f1 有哪些不同，做多分类的时候 precision、recall 和 f1 往往是相等的，为什么?
- 分类模型评价标准有哪些？追问，AUC 可以用于多分类模型么?
- ROC 怎么画 ？
- 解释一下 AUC(ROC 的面积)，从概率上面的解释一下?
- AUC 为什么适合在二分类任务下，为什么能解决正负样本不均衡?

2.5 传统机器学习方面

2.5.1 讲解相关原理

2.5.1.1 数据准备

- 常用采样的方法?
- MCMC 采样?
- 样本不平衡时怎么处理？过采样和欠采样等。

2.5.1.2 特征工程

① 特征降维

- PCA 的原理说一下?
- 自己对 PCA 的理解（最小方差、最大误差、信噪比最大、基于特征值分解、PCA 和 SVD 都是基于特征值分解，然后继续说 SVD 可以用于做推荐）。
- SVD 怎么实现降维?
- KSVD 和 SVD 的区别；因为楼主的项目里有用到 KSVD，稀疏编码的一种字典学习算法；
- 特征维度很高时你是怎样做的操作?
- 特征值的大小有什么含义，为什么 PCA 中分解协方差矩阵要按特征值排序?

② 特征选择

- 特征工程的作用?
- 特征选择方法都有哪些?
- lasso 和 ridge 的区别以及如何做特征选择?

2.5.1.3 有监督学习-分类和回归方面

① 分类回归树（集成学习）

- 知道哪些分类算法？（回答了决策树、RF、SVM、贝叶斯、神经网络），讲讲 RF 和 XGboost 区别

A.基于 bagging：随机森林

- 为什么使用随机森林?
- 随机森林的基本原理?
- 为什么随机森林效果比较好?
- 随机森林怎么进行特征选择?
- 随机森林和 XGboost 有什么区别?
- 为什么随机森林和 SVM 的效果比其他算法好?
- 随机森林怎么得到特征复杂度（简历中项目）

B.基于 boosting：Adaboost、GDBT、XGBoost

- XGB，RF，LR 优缺点场景？逢面必问。
- 看到你用过 LightGBM，你可以说一下 GBDT、XGBoost、LightGBM 的特点?
- XGBoost、LightGBM 哪种快一点，或者结果好一点？lightGBM 说一下改进点?
- GBDT 和 XGBOOST 的区别，gain 函数和 gbdt 的不同?
- XGBoost 如何给出特征重要性（我说根据推导的目标函数，分裂前后相减得到增益值，对这个增益值进行排序，从而给出特征重要性，面试官说可能不是这样子，让自己查一下）
- GBDT 怎么分裂的？XGB 原理？说说 XGB 相比 GDBT 的优势?
- RBF 和 XGB 同等效果下，哪个更深（RBF 更深）
- 说一下 XGBoost 的列采样?
- GBDT 的梯度是什么，为什么要算这个，怎么算?
- 用 GBDT 的时候，主要调节的参数？如何判断判断过拟合?
- RF 和 GBDT 的不同，GBDT 和 Xgboost 的不同，Xgboost 为什么更快，做了哪些优化?

② 逻辑回归 LR

- 逻辑回归和线性回归的区别?
- LR 和 XGB 关于特征处理有什么区别吗?

- LR 和树模型区别?LR 模型离散化的好处?
- LR 模型能够用来回归么，不设阈值可以吗?
- LR 模型为什么要用交叉熵而不是 MSE？刚刚说了 MSE 那么多缺点，那为什么回归还要用 MSE。
- 以 LR 为例，解释一下为什么权重接近 0 的时候能防止过拟合?
- LR 和 SVM 的区别，应用起来有什么不同?当聊到损失函数，问分别是什么?hinge 损失函数里面的 z 是什么?
- LR 和 SVM 对于离群点的敏感性？LR 有什么特点，适合做哪些?
- LR 强行引入非线性和 FM 的差别?
- LR 可以处理非线性问题吗?
- LR 怎么缓解过拟合？（L1，L2）L1 和 L2 之后的解有什么不同?

③ SVM（支持向量机）

- 问 SVM 的原理和优点?
- SVM 的物理意义是什么?
- SVM 最后训练好的网络是什么结构?
- SVM 的硬间隔，软间隔表达式?
- SVM 使用对偶计算的目的是什么，如何推出来的，手写推导?
- SVM 的核函数使用过哪些？SVM 为什么要加核函数?
- SVM 和 GBDT 的比较?
- SVM 应用于哪个领域?
- SVM 可以处理非线性问题吗?
- SVM 的基本假设，（存在一个超平面可以分开正负样本）如果分不开怎么办？（软间隔，核函数）
- 有哪些常用的核函数?

④ 朴素贝叶斯（Naive Bayes）

- 朴素贝叶斯与贝叶斯有什么区别?
- 朴素贝叶斯适用哪些场景?
- 极大似然估计是什么意思?
- 最大似然估计和贝叶斯估计的区别?

⑤ 决策树（DT）

- 决策树都讲讲，ID3，C4.5，CART 树是什么？各有什么优势?
- 决策树有哪些选择最优划分节点的方法？决策树节点确定?有了信息增益为什么还提出了增益比?
- 决策树里面的分类树怎么选择划分属性，给了两个属性分布情况，问选择哪个?
- 决策树做回归时候划分点怎么选择？决策树启发函数?
- 讲一下信息增益,信息增益比,Gini 系数的关系?
- 有真实的场景吗？为什么树的深度会影响过拟合?
- 决策树的分裂方式是什么，根据什么变量来决定分裂变量?
- 决策树怎么选择分裂点?

⑥ 其他

- 分类与回归的区别（自己说区别是目标变量是连续还是回归，面试官提出质疑，后面和面试官讨论了一下，今天搜答案，有个答案说最本质的区别是度量空间的不同）

浅层： 两者的的预测目标变量类型不同，回归问题是连续变量，分类问题连续变量。

中层：回归问题是定量问题，分类问题是定性问题。

高层： 回归与分类的根本区别在于输出空间是否为一个度量空间。

2.5.1.4 无监督学习-聚类方面

- 聚类的算法有哪些？评价方法？优化算法?
- 聚类了解那些算法？只说了 K-Means，介绍了 K-means 的步骤?
- 聚类 kmeans 的原理、缺陷以及改进方式，初始点的选择?

- Kmeans 与 kNN 什么区别?

2.5.1.5 模型评价

- 模型评价标准，RMSE\ROC\AUC 等等
- 二分类混淆矩阵及相关指标你能想起来哪些，都介绍下?
- 怎么衡量两个分类的相似度，说下混淆矩阵和 softmax 输出的两个概率之差这两种方法确定分类相似度的异同?
- 机器学习分类指标有哪些? AUC 如何计算?

2.5.2 手推算法及代码

2.5.2.1 手推公式

- 推一下 FM 的公式，然后推一下 FM 的梯度更新方式
- 说下 LR 模型的推导，指定 LR 模型之后你怎么判断这个问题是回归还是分类?
- LR 写一下损失函数
- 介绍核岭回归算法，并推导?

2.5.2.2 手写代码

- 一个 M 的空间，对其采样（采样有误差），在 2 范数的情况下，映射到 N 维空间，输出 N 维正交基?

2.6 深度学习&机器学习面经通用知识点

2.6.1 损失函数方面

- 常用的损失函数和适用场景? 物理意义?
- 从理论角度讲解一下，基于不确定性平衡多任务各个 loss 的原理以及合理性
- Logistic Regression 损失函数
- 逻辑回归中损失函数的意义?
- 为何分类问题用交叉熵而不用平方损失?

- 二分类交叉熵和多分类交叉熵的公式?
- 手写交叉熵公式、核函数解决什么问题?
- 交叉熵和最大似然损失函数的区别?

2.6.2 激活函数方面

- 激活函数的作用是什么？我说最开始的感知机那儿是可以让神经元的输出是一个 0~1 之间的数，这样可以判别属于哪一类的概率大，他问我还有没有，我想了一会儿说不知道了（面试的最后我问了他，他说是让神经元非线性化，这个非常重要）
- 常用什么激活函数，有什么作用?
- Leaky-ReLU 和 ReLU 的区别?
- 激活函数为什么会梯度消失、Relu 有什么改进?

2.6.3 网络优化梯度下降方面

- 说一下常见的优化器？优化方法？Adam 和 sgd?
- SGD 优化方法，批数量的影响?
- minibach SGD 的 minibatch 怎么选择，如果给 1000 万的数据，minibach 应该选多少?

2.6.4 正则化方面

- 正则化有哪些方法？作用？原理?
- 从原理上解释 L1, L2 正则（如 L1 正则为什么能够起到特征选择的作用）
- L1 和 L2 正则化区别，为什么防止过拟合?
- L1 的形状，无穷范数的形状?

2.6.5 压缩&剪枝&量化&加速方面

- 深度学习模型压缩有哪些方法介绍一下?
- 量化和剪枝讲一下？不了解量化，主要说了剪枝，主要是随机取消某些权重，如果模型跑的结果和原完整模型跑的结果在我们容忍（阈值）范围内，剪掉这些参数，模型压缩会降低原

模型精度。

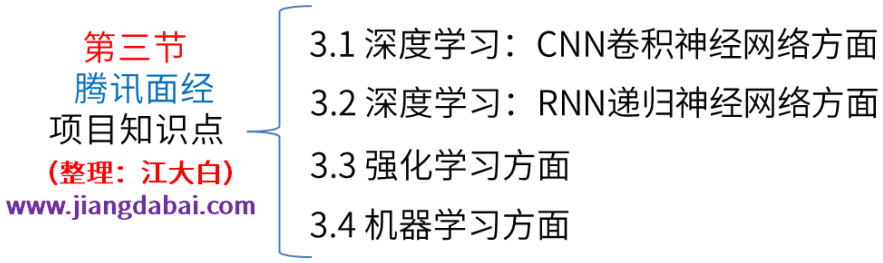
2.6.6 过拟合&欠拟合方面

- 过拟合的本质?
- 防止过拟合的方法，具体怎么实现的？正则项为什么能减缓过拟合?
- Loss 不降怎么办？vali loss 不升（过拟合）怎么办?
- 少样本问题怎么办？从模型的角度：添加正则项防止过拟合、提前停止训练等。
- 如何解决过拟合，L1 和 L2 正则原理，区别?

2.6.7 其他方面

- 数据类别不平衡问题如何解决?
- 模型训练的停止标准是什么？如何确定模型的状态（指标不再提升）

3 腾讯面经涉及项目知识点



3.1 深度学习-CNN 卷积神经网络方面

3.1.1 目标检测方面

3.1.1.1 讲解原理

- 说一下目标检测有哪些网络？挑一个说一下
- SSD 原理讲一下
- 讲一下 faster rcnn、yolo、SSD?
- Faster-RCNN 结构，与其他 RCNN 对比?

- FasterRCNN 细节，怎么筛选正负 anchor?
- softnms 原理，比 nms 好吗?
- 简单介绍 Fast RCNN -> Faster RCNN -> mask RCNN (这个真的好高频)、有没有自己搭一个 faster rcnn 的网络?
- YOLO 为什么比其他检测算法快?
- YOLO 每一代的不同（问的特别细）

3.1.1.2 手写代码

- 实现简单的 NMS（已提供计算 IOU 的函数）
- IOU 的实现写一下?

3.1.2 目标追踪

- 追踪结果与检测结果怎么融合?
- EKF 多目标跟踪的原理?

3.1.3 图像分割

- 问了解其他图像领域的任务吗？我说了解图像分割，然后叫介绍一下图像分割领域的常用结构，我说了 FCN, U-net, deep lab 系列之类的，然后稍微介绍了这些网络的结构，优点之类的
- Mask R-CNN 相比于 Faster R-CNN 有哪些改进?

3.1.4 超分辨

- 视频超分如何应对异常可能，即运动误差大的如何处理?

3.1.5 图像分类

- 介绍分类经典网络;当时我就只想到 AlexNet,VggNet,ResNet,着重讲了下 Alex 和 ResNet。
- 分类问题：几万个类怎么做，如果用 softmax 的话有什么问题，类之间如果很相关又如何做?

3.2 深度学习-RNN 递归神经网络方面

3.2.1 自然语言处理 NLP

① Bert

- BERT 的细节、优势？与 GPT、ElMo 比较
- BERT 模型的个人理解，有那些 tricks，为什么会好用?
- 讲下 bert，讲着讲着面试官打断了我，说你帮我估算一下一层 bert 大概有多少参数量?
- Bert 里 add&norm 是什么以及作用?
- Bert 里面位置向量的作用是什么？有哪些生成方式?
- 了解 bert 的扩展模型，roberta，albert，XLnet 吗？为什么 XLnet 效果比 bert 效果好?
- ELMO、BERT、GPT 模型彼此之间有什么区别?
- Bert、roberta、xlnet 异同点?
- Bert 参数如何计算?
- Bert mask 策略，作用分别是什么?

② Transformer

- Transform 和 RNN 的优缺点?
- Transformer 结构，input_mask 如何作用到后面 self-attention 计算过程。
- 讲下 transformer 相对于其他 RNN 的优点，讲下 self-attention 和 attention。聊了下 bert 的文本分类，然后扩展到场景题，问一个新的短文本，如何正确分到正确的类别里取，不可能去重新训练模型（包括增量训练），也不是用规则这种。

③ HMM 隐马尔科夫模型

- 隐马尔科夫了解吗?
- CRF 和 HMM 的区别?
- CRF 与 HMM，特征函数，有向图无向图，因子分解。
- 命名实体识别模型介绍？评价标准 ROC?

- 命名实体识别模型的参数量?
- 命名实体识别过程做了哪些参数的调整?

④ Word2vec

- word2vec 原理讲一下?word2vec 的训练过程?
- 从 word2vec 讲到分层 softmax，再到负采样。
- 简单介绍一下 word2vec 和 fasttext?
- word2vec 两个模型的损失函数是什么?
- 关于 NLP 的 word2vec，怎么实现对单词的编码?
- w2v 怎么用的，为什么不和 NN 一起训练而是固定住，训练开销很大么，embedding 矩阵很大么，会占用很多内存么?
- NN 用的框架是什么，画出框架图，为什么选择双向 LSTM，序列与时间有关么？如果和序列没有关系，那双向 LSTM 会不会引入噪声?
- word2vec 和 ELMO 主要有什么不同，为什么 elmo 效果更好?

⑤ 其他

- 情感分析用什么数据集?
- 你了解 network embedding 方法吗？deepwalk 和 node2vec
- 问了 gcn、deepwalk 之类的 graph embedding?
- one-hot 编码的原理及意义?
- cbow 和 skip gram 的区别？你觉得用哪个训练的词向量结果好？为什么?
- 序列标注常见的算法有什么?
- 问了一下简历里面关键词提取是怎么实现的?觉得我只是用 TFIDF 和 TextRank 没有什么亮点，他们会根据业务需求去设计 learning 模型。
- tf-idf 原理，还知道其他关键词提取技术吗?
- 如何理解双塔模型中 cosine similarity 的计算？如何理解粗排和精排的不同需求?

- 介绍 GPT2 如何写诗、写对联?
- GPT2 如何围绕一个主题/关键词写诗?

3.3 强化学习

3.3.1 讲解原理

- 介绍一下强化学习的策略梯度?
- 生成对抗网络用在文本中如何梯度估计?
- 判别学习和生成学习分别有哪些?
- Policy Gradient 和 Q Learning 的区别?
- GAN 训练过程一定能收敛吗?
- GAN 相关。描述、介绍优缺点、知道哪些 variant、WGAN 的优点和不足、conditionalGAN 的应用场景
- 你用 GAN 做数据增强?你有对比过 GAN 和 VAE 的生成效果吗?

3.4 机器学习方面

3.4.1 推荐系统

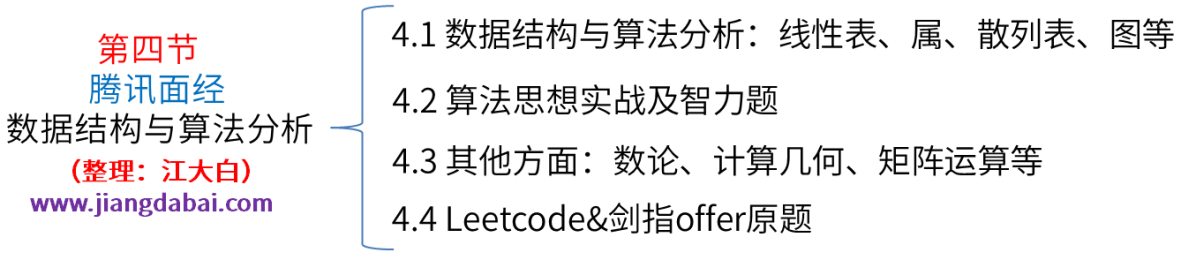
- CTR 预估的深度模型以及一些扩展从 wide&deep 到 deepfm 到 dcn 到 xdeepfm 都解决什么问题?
- 问推荐算法，fm，lr，embedding?
- 谈谈 FM 与 DeepFM?FM 和 FFM 的区别在哪里?
- DIN 结构，DIN 提出动机以及与之前模型的区别?
- DIN 中 Attention 机制实现
- CTR 预估模型的演化过程中的着手点?
- 协同过滤的 itemCF，userCF 区别适用场景?
- 推荐系统的大概步骤，如何解决冷启动?

- 推荐系统有几个步骤，为什么要召回?
- 深度学习用在推荐里的例子有哪些？能描述一下 ncf 的基本框架吗?
- 推荐系统里的排序算法有哪些?
- 如果让你给用户推荐页面的排序评分，你会怎么设计这种评分机制?
- 介绍 din、dien?
- 两种协调过滤，区别比较，如何解决冷启动?
- 推荐系统如何离开局部最优?
- 基于 w2v 的 ANN 怎么做的?
- 推荐系统里面是如何考虑冷门商品?
- 新增一路召回，在排序阶段需要做什么改进?

3.4.2 点击率预估

- 广告点击率预测怎么做？怎么特征选择?

4 数据结构与算法分析相关知识点



4.1 数据结构与算法分析

4.1.1 线性表

4.1.1.1 数组

- 数组有序，但是循环右移了几位，问新数组中原数组起始位子的下标是多少?
- 有序含重复值数组找某个值第一次出现的位置

- 无序数组中找第 k 大的数
- 二维矩阵的查找（每次去掉一行一列，以及分治法）
- 有序数组查找给定值出现的左右位置?
- 找出数组唯一的重复数
- 一个长度为 n 的数组里面可能有 1...n-1 出现，一个数字出现多次，其余数字出现 0 次或 1 次，怎么判断出现多次的数字是哪个?
- 判断两个数组是否有相同数字?
- 旋转数组的二分查找
- 旋转数组的最小数字
- 一个有序数组[0,0,0,1,1,1,1,1,1,3,3,3,3,,3]，给一个 k，找出 k 的左边界，不存在就返回-1（二分）
- 一个整数数组，找最右边数与最左边数的最大差值
- 无序数组找前 k 大的数，描述思路+复杂度分析
- 给一个长度为 n 的数组，输出其中任意 n-1 个数的乘积，可不可以不用除法?
- 给定一个整数数组[a1,a2,.....aN] ，N 个数, 现在从里面选择若干数使得他们的和最大，同时满足相邻两数不能同时被选， a1 和 aN 首尾两个也认为是相邻的?
- 子数组最大和
- 给一个数 N，k，每一轮可以进行两种操作的其中一种

a. 所有的数拆分成两个更小的数

b. 所有的数-1

已知拆分操作只能进行 k 次，最少需要多少次把所有数都消去?

- 给一个数组，求连续子数组和为 k 的倍数的所有子数组
- 给定一个整数数组 A，找到 min(B) 的总和? 其中 B 为 A 的每个连续子数组。(连续子数组的最小值之和)
- 一个二维数组，表示地图上该处的高度，一个人滑雪只能从高处滑向低处，方向为相邻四

个方向，求最长滑行距离?

4.1.1.2 链表

- 单向链表的反转
- 逆序双向链表
- 单链表判断是否有环（leetcode easy)，以及判断环入口
- 链表判断有环&不在环里的链的长度
- 链表排序（用 O（nlogn）的归并写的）
- 合并两个有序链表
- 单向链表输出倒数第 k 个数
- 写一个函数把两个单调递增的链表连接起来，要求连接之后还是单调递增?

4.1.1.3 栈

- 堆和栈的区别?

4.1.1.4 队列

- 队列和栈的区别？再写代码：用栈实现队列

4.1.1.5 字符串

- 反转字符串
- 去掉字符串中违法的单括号
- 字符串编辑距离
- 求 A 的长度为 L 的各个连续子串在 B 中出现的次数总和?
- 找到字符串中第一个只出现一次的字符？（说明时间复杂度）
- 找数字中只出现一次的字符
- 数字的二分查找，找到始末位置
- 寻找回文串

4.1.2 树

- 二叉树搜索的复杂度？与红黑树的区别？
- 按照左根右根的方式序列化二叉树？再根据的序列，反序列化一颗二叉搜索树？
- 红黑树怎么调整？红黑树的左旋，右旋？
- 二叉树最长路径长度
- 二叉树最低公共祖先，非递归方法
- 判断平衡二叉树？
- 根据前序，中序创建二叉树
- 二叉树左右子树的翻转
- 二叉搜索树的性质？
- 二叉搜索树的第 K 大节点？
- 对堆排序的介绍

4.1.3 排序

- 排序算法哪些时间复杂度比较低？
- 最长上升子序列并分析复杂度，O(N^2)和 O(NlogN)的方法都说一下？
- 最大递增子序列如果是个环，怎么计算？
- 快排的思想是什么？
- 各种排序介绍一下，手写快排，快排复杂度是多少，最差的情况下怎么优化？
- 如何在海量数据中快速查找最相似的文本？
- TopK 给出 3 种解法
- n 个数求第 k 大数（我说用最大堆或者锦标赛排序，但面试官说不是最优解，说用快排，复杂度是 n）
- TOPk 问题，如果可以并行，如何优化？
- 堆排序的细节

4.1.4 搜索

- 0-1 矩阵，DFS 判断两点是否可达?

4.2 算法思想实战及智力题

4.2.1 算法思想实战

- 疯子坐飞机的算法题
- 高楼扔鸡蛋问题，不需要数学最优解，能把 DP 解法讲出来就够了
- 圆桌问题（约瑟夫环）
- 数岛屿数目的变体
- 名人问题，给出最优解法
- 上 N 阶台阶的种类（答完 O(n)解后，要求 O(logn)复杂度。受到启发，用类似马氏链中状态转移矩阵的方法去做）
- List 和 set 的区别查找插入的时间复杂度，set 如何实现 O（1）的查找复杂度?
- 三角形数的最大路径和? （dp 思想)

4.2.2 智力题

- 有一栋 100 层高的大楼，给你两个完全相同的玻璃球。假设从某一层开始，丢下玻璃球会摔碎。那么怎么利用手中的两个球，用什么最优策略知道这个临界的层是第几层?
- 有 12 张生肖卡，每个人吃一顿饭集齐一张，平均吃多少顿能全部集齐?
- 25 匹马，5 条赛道，无计时工具，比出前三名最少多少场比赛?
- 64 匹马，每次最多可以赛 8 匹，可以知道结果的相对顺序，测多少次可以选出前 4 名?
- 给一串数列，这串数列有正有负，但是总和为 0。每个数 xi 代表一个村庄，正的表示村庄想卖出 xi 份水果，负的表示想买入 xi 份水果。两相邻村庄间的距离是相同的，单位距离运送一份水果的运费均相同，每份都是 k。问，把每个村庄的需求和供给都解决掉需要的最少运送费是多少?

- 一款修仙游戏，设定从凡人到飞升一共 100 个等级，凡人为 0 级，飞升对应 99 级。 从凡人升级到 1 级需要消耗资源 1 个单位，成功概率 99%，失败还是凡人。 从 1 级升级到 2 级，需要消耗资源 2 个单位，成功概率 98%， 失败则降 1 级为凡人。 类似地，从 k（k 大于 0，小于 99)升级到 k+1 级需要消耗资源 k+1 个单位，成功率（1-(k+1)/100), 失败则降一级。问从凡人到飞升平均需要消耗多少单位资源 ？
- 已知 n 个人（以编号 1，2，3...n 分别表示）围坐在一张圆桌周围。从编号为 k 的人开始报数，数到 m 的那个人出列；他的下一个人又从 1 开始报数，数到 m 的那个人又出列；依此规律重复下去，直到圆桌周围的人都出列?
- 有 N 个人轮流进去一个小黑屋，可开灯、关灯或不操作，每此随机地有一个人进去，互相之间无任何交流，问某人如何知道自己是最后一个进去的人?
- 三个瓶子倒水问题:

11 升，5 升，6 升的瓶子，其中 11 升的瓶子里装满了水，请倒出 8 升的水。

4.3 其他方面

4.3.1 数论

- 100 块钱随机分给 10 个人，要求每个人分到的数额在期望上相等（即都是 10）
- 求数字的平方根?

4.3.2 计算几何

- KKT 条件
- 拉格朗日乘子法对偶
- 说一下凸优化方法?
- 给出一些点坐标，判断这些点是否关于与 y 轴平行的轴对称

4.3.3 概率分析

- 给一个实时更新的数据流，最开始大小为 N，从里面需要抽取 m 个数据，问新加入的数据怎么能够保证仍然是等概率抽样？即新加入的数据以多大的概率保留才能保证你所选的样本都

是等概率的？（涉及到的蓄水池抽样算法）

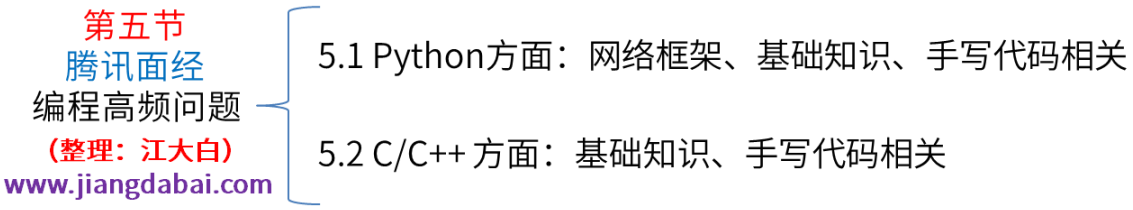
- 给 100 个红球 100 个篮球 两个篮子，怎么分配这些球让你在任意一个篮子中摸到红球的概率最大？
- 甲乙扔骰子，获胜概率相同，投 10 次，已经 5 次了，甲已经赢了 3 次，问甲获胜概率？
- 线段分三段，成三角形的概率
- 给一个均匀生成 1-7 随机数的生成器，怎样均匀生成一个 1-10 随机数的生成器？
- 一个函数 f()返回 0 的概率是 0.6，返回 1 的概率是 0.4，写一个函数，利用 f() 使返回 0 和 1 的概率均等？

4.4 Leetcode&剑指 offer 原题

- Leetcode 3，medium：求最长的不含重复字符的子串
- Leetcode 53：最大子序和
- Leetcode 72
- Leetcode 75
- Leetcode 215：在 N 个无序无重复的整数中，找到第 K 大的数字
- Leetcode 422
- Leetcode 458：小白鼠试有毒药水
- Leetcode 887
- LeetCode 中等题：开关灯泡：初始时有 n 个灯泡关闭。 第 1 轮，你打开所有的灯泡。第 2 轮，每两个灯泡你关闭一次。 第 3 轮，每三个灯泡切换一次开关（如果关闭则开启，如果开启则关闭）。第 i 轮，每 i 个灯泡切换一次开关。 对于第 n 轮，你只切换最后一个灯泡的开关，找出 n 轮后有多少个亮着的灯泡？
- LeetCode 中等题：朋友圈：班上有 N 名学生。其中有些人是朋友，有些则不是。他们的友谊具有是传递性。如果已知 A 是 B 的朋友，B 是 C 的朋友，那么我们可以认为 A 也是 C 的朋友。所谓的朋友圈，是指所有朋友的集合。输出所有学生中的已知的朋友圈总数。裸题， 需要介绍一下并查集，并且问我并查集是图论的什么问题？

- Leetcode 原题：找出一个字符串中所有的回文串
- Leetcode 原题：打家劫舍 II
- Leetcode 原题：丢骰子 https://leetcode-cn.com/problems/nge-tou-zi-de-dian-shu-lcof/
- leetcode 原题： https://leetcode-cn.com/problems/container-with-most-water/
- 剑指 Offer 60：n 个骰子的点数
- 剑指 offer 原题：数组的全排列
- 剑指 offer 原题：给定一个未知长度的单链表，找到倒数第 K 个节点
- 剑指 offer 原题：输入一个正整数数组，把数组里所有数字拼接起来排成一个数，打印能拼接出的所有数字中最小的一个

5 编程高频问题：Python&C/C++方面



5.1 python 方面

5.1.1 网络框架方面

5.1.1.1 Pytorch 相关

- 深度学习框架之间的差别?
- 如何做 pytorch 的部署与热更新?

5.1.1.2 Tensorflow 相关

- Tensorflow 与 Pytorch 的区别？说几个你印象比较深刻的函数 API

5.1.2 基础知识

5.1.2.1 线程相关

- Python 多线程多进程了解吗?

5.1.2.2 区别比较

- Python 的浅拷贝和深拷贝?
- Python 中 lambda 与 map 的作用?
- Python 中，a is b 和 a=b 的区别?
- Python copy() deepcopy() 普通赋值，有啥区别?
- Python2 和 Python3 的区别有哪些?

5.1.2.3 讲解原理

- 可变/不可变类型，函数参数传递是否改变?

可变：int、float、list、dict.values 不可变：str、tuple、dict.keys

- Sort 原理是什么?
- 如果 a=b，则 a is b 是否成立?
- Python 中使用字典的好处？字典是哈希表，那字典查找跟列表查找的时间复杂度分别是多少?
- Python 的生成器和迭代器了解吗？Python 的内存管理了解吗?
- Python：callable，垃圾回收
- Python 列表合并方法有哪些：加法、extend，区别，旧内存如何处理;
- Python：Python 中字典的底层实现（哈希表）
- 了解 Python 的 GIL（全局解释器）吗?
- Python 怎么定义一个类的成员变量?

5.1.3 手写代码相关

- 如何在 Python 中调用 linux 系统程度 os?
- 如何在 Python 中遍历文件夹 os.walk?
- 字典如何按 value 排序？sorted(dict.items(), key=lambda x:x[1]) ？
- Python 写一个函数，实现给定一个列表，把列表所有 0 移到列表最后面，其余相对顺序不变，要求时间 o(n)，空间 o(1)
- Python 写一个函数，实现有 1T 的数据，10 亿个不重复单词，给你一台机器，16G 的内存和 5T 的内存，怎么统计每个单词的个数?
- 我有一个文本，那么我要统计每个词出现的频率，Python 上应该怎么做?
- 不用库函数的情况下实现 split 函数
- Python 里的正则表达式
- Python 的 list 怎么去重?
- range(10)[::-1]是什么？xrange(10)呢?
- [i for i in range(10)]和(i for i in range(10))区别?
- Python 对一个列表删除所有为 0 的数字?

5.2 C/C++方面

5.2.1 基础知识

5.2.1.1 内存相关

- 清除 size 可以使用 clear()，那释放内存用什么（后来查了下，应该是 swap()）
- 说一下智能指针，怎么实现的自动释放内存?

5.2.1.2 区别比较

- vector 底层实现，size 和 capacity 区别?

5.2.1.3 讲解原理

- 说说 C 实现多态的好处，有哪些实现多态的方法？
- 纯虚函数的用处？虚函数的好处？
- C 类里面有一个静态成员，那么有什么特性？
- C 的虚基类，stl 中的 map 怎么实现的，复杂度是多少，智能指针讲一下
- C++：多态、虚函数、构造函数和析构函数、继承等等
- 智能指针，static 关键字的作用。
- 为什么析构函数要写成虚函数的形式吗？
- 操作系统是如何执行 C++代码的？C++智能指针介绍一下？

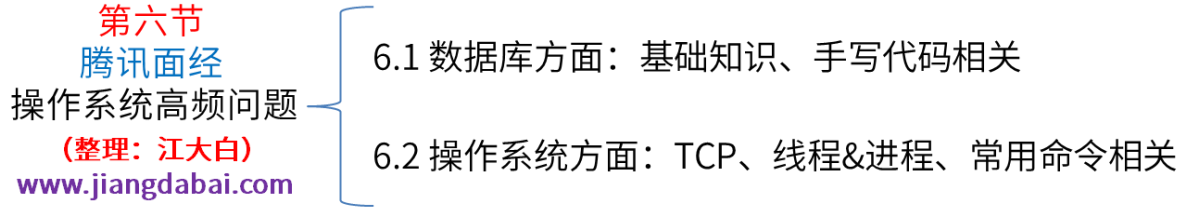
5.2.1.4 讲解应用

- C/C++为什么比 python 语言执行速度快？讲讲 CMAKE？在 CMAKE 中如何交叉编译？
- 我说是 linux 底层时 c 语言，c/c++调用接口可以直接调用系统的，而 python 语言调用 python，然后 python 调用 c 接口
- 面试官说 python 也可以直接调用 c 接口，主要是编译的时候，编译 python 为机器语言会产生一些冗余的机器代码，而 c/c++不会或产生很少冗余的机器代码，所以执行效率高。

5.2.2 手写代码相关

- 创建一个 N*M 的数组，再删除它，释放内存

6 操作系统高频问题：数据库&线程&常用命令等



6.1 数据库方面

无

6.2 操作系统方面

6.2.1 TCP 协议相关

- TCP 和 UDP 的区别?

6.2.2 线程和进程相关

- 进程和线程的区别?

6.2.3 常用命令

- linux 怎么删除一个进程?那么进程号怎么知道呢?
- linux 命令怎么查看硬盘太小?
- linux 中查找符合一定规则的文件名怎么查找，或者用脚本也行?
- 对操作系统的了解，Linux 命令查看端口 netstat -ntlp|grep 80

7 技术&产品&开放性问题

7.1 技术方面

- 10 亿个 32 位正整数，求不同值，只给 1GB 内存？（我只答出来 4GB 的情况，时间负责度还不是最优的）
- 文件 40 亿+无重复数字，排序到新文件。（好像要用 bitmap 结构）
- 海量数据找其中唯一一个不重复的字符（bitmap）
- 从大数据中抽取 m 个样本，怎么保证可以代表原数据集?
- 如果给你一些数据集，你会如何分类（我是分情况答的，从数据的大小，特征，是否有缺

失，分情况分别答的）；

- 如果数据有问题，怎么处理；
- 如果出现重合的人像怎么区分？
- 思维题：A，B 为两个文本文件，大小为 1G，内容均为数字，给一个机器，硬盘为 1T，内存为 256M，如何设计来找出 A，B 中重复的数字（A，B 自身也需去重），输出到 C 文件。
- 如何使用机器学习对代码的相似性进行分析？
- 就一个简单的短文本，怎么挖掘对应的相关的文本，或者挖掘相关的信息？
- 一堆恶意文本 case，怎么检测去除（一些网页上的广告评论），传统方法、AI 方法？

7.2 产品方面

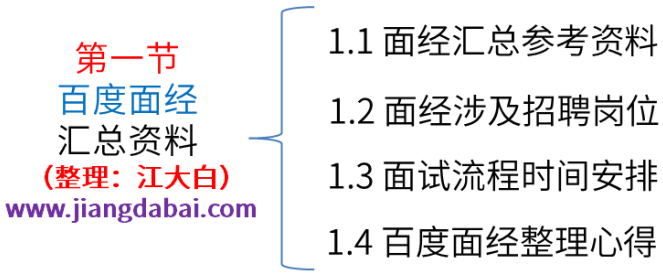
- 一排货架上全是饮料，要专门把可乐找出来，而不要找其它的饮料，神经网络需要怎么玩，需要注意一些什么；以及这个模型如果放到一般室内环境中，预计的工作效果如何。
- 场景分析：如何给玩家的商城界面的皮肤排序（答了一些 RNN 方法）？如果是新玩家呢？（答了用 pre-trained model 直接给排序）
- 场景分析：如何给游戏商品定价？结合道具图片、描述。（答了 CNN、NLP 相关）如何提取道具图片的特征？（答了用 CNN）
- 情景题：有上亿的邮件，如何聚类？应该从哪些方面考虑？
- 思维题：如何利用人工智能技术检测两篇文章的相似度？请设计具体的方案，详细介绍用到的技术，及相似度衡量标准选择等。
- 思维题：如何测试两步电梯的性能？请写出详细的测试用例。
- 在应用宝里面用户搜索 APP，你如何利用每个 APP 的标题、描述、历史点击情况等属性为结果排序呢？我就简单想了一下利用历史点击率去设计 learning 模型，但是他说用户可能会有错别字呀等情况，你应该考虑这些该怎么处理。
- 场景设计：现在你是 iphone 的设计师，我们需要在锁屏界面左下角的按键处把手电筒替换成一个用户下一个可能用的功能，如何设计一个系统来收集信息并预测这个按键应该放哪个功能。

7.3 开放性问题

- 开放问题：时针分针一天重合多少次
- 讲一下哲学家就餐问题？
- 开放题：如果让你去解决一个陌生领域的问题，从分析问题到设计模型以及评价指标，你会怎么做？
- 开放题：给用户的搜索日志、记录、点击曝光记录等可以解决什么问题，我答了构建用户画像，进一步针对用户画像的问题，利用哪些特征，为什么，特征都如何 embedding？

4 | 百度算法岗武功秘籍

1 百度面经汇总资料



1.1 面经汇总参考资料

① 参考资料：

（1）牛客网：百度面经-218 篇，网页链接

（2）知乎面经：点击进入查看

（3）面试圈：点击进入查看

② 面经框架&答案&目录&心得：

（1）面经框架及参考答案：点击进入查看

（2）大厂目录及整理心得：点击进入查看

1.2 面经涉及招聘岗位

（1） 实习岗位类

【NLP 实习工程师】、【视觉技术部算法实习岗】、【百度健康业务部算法实习岗】

（2） 全职岗位类

【计算机视觉算法工程师】、【无人驾驶算法工程师】、【凤巢算法工程师】、【展示广告部机器学习/数据挖掘/自然语言处理工程师】、【自动驾驶车联网部门工程师】、【百度推荐平台算法工程

师】、【百度搜索岗视频内容搜索算法】、【百度原声商业推广部算法】、【百度 SRE 工程师】、【百度度秘推荐算法】、【百度地图算法岗位】、【nlp 算法工程师】、【百度大搜实习算法工程师】、【推荐策略部算法工程师】、【feed 推荐算法工程师】、【搜索团队机器学习工程师】、【度秘算法工程师】、【语音部门算法工程师】、【百度大搜算法工程师】】、【智能交通高级算法工程师】、【商业场景研发部算法工程师】、【百度增强现实技术部_高级动作捕捉、机器人运动学算法工程师】、【百度 AGG 智慧城市事业部的计算机视觉算法工程师】

1.3 面试流程时间安排

百度面试流程-整理：江大白			
	面试类型	面试流程	备注（侧重点）
第一面	技术面	自我介绍+项目/实习经验 +技术问答	项目细节问的很细， 对基础知识发散分析
第二面	技术面	自我介绍+项目/实习经验 +技术问答+算法编程	偏压力面，也会问 算法的底层实现和项目
第三面	技术Leader面	自我介绍+项目经验+公司发展	偏实际和业务场景的问题 以及合作能力
第四面	HR面	基础人力问题	/

PS：以上流程为大白总结归纳所得，以供参考。

其他注意点：

- 二面面试官经常是以后的直系领导，主要是问项目还有一些临场发挥的其它问题
- 有的第三面后，还会增加性格面试，不过算是加面，有可能给 SP 级别。

1.4 百度面经面试心得汇总

★ 百度的面试过程还是挺有水平的。基础知识+项目+发散性的问题都有涉及，还是挺考验基本功与反应能力的。

★ 百度的笔试就令人印象深刻：选择题啥都考，很杂，操作系统，数据库，c++，python，机器学习，深度学习啥都考

★ 百度面试很注重计算机基础和数学原理，而且比较注重论文发表情况或者项目经历，虽然问的比较难，但是面试官会给提示，逐步引导你走向正确答案，如果回答不上了面试官会做出讲解。

★ 偏好 C++，直接让你敲 c++也反映出他们的工程需求比较大，内部的需求也是项目落地，而不是模型的建模。

★ 面试时，会循序渐进，不断挖掘自己知识库中知识储备以及灵活应用能力，引导你发散思维，大胆进行业务处理，总结来说，技术方向对求职者的考量确实很到位

★ 面试官建议：夯实基本功，算法原理、数据结构、代码功底线上 C++线下 python，hadoop 等工具，tf 等框架的使用等，第二提高系统思维，从解决问题角度从头至尾分析，第三，了解业务方面。

★ 无论是找算法还是开发，代码能力都很重要，刷题必备。

★ 目标公司最好能加到 HR 或者面试官的联系方式，不然面完就失联很被动。

★ 百度的面试非常重视基础，问得很细。

★ 基础知识复习复习好（包括机器学习、深度学习还有最重要的刷题），尽量提早做，说实话这些事我也是到春招末和秋招中才做得比较好，因此错失了很多机会。

★ 简历上的内容一定要很熟悉，这个不解释；

★ 准备一个漂亮的简历，所以春招找实习很重要！拿到一个大中厂的实习，你会发现事事都会顺利很多；

★ 平时学习模型的时候还是应该再深入一些，了解模型的细节，以及做法的原因。还有一些主流推荐模型的优缺点和适用场景、改进方法等等。

★ 因为机器学习和深度学习项目，大家很容易盲目的去做，经常是随便试好几个模型，哪个好就用哪个。 但是面试官更希望你能了解模型的细节和原理，有针对性地有充分理由地使用模型、调整参数。我觉得大家在面试之前可以多准备这方面，就算面试官没有问也可以主动地说出来项目中一些做法的原因。这样可以让面试官觉得你是有独立思考的，而不是随便调包。不过最好是有深度一点，像 “我发现了过拟合，用了 dropout，最后缓解了过拟合“ 这种比较普通的说出来可能也没什么帮助。

★ 面试官会对项目问的超级细，甚至问了当时结果具体数字是多少？从侧面反映他其实是想看你是不是真的做了这个项目，是不是熟悉！

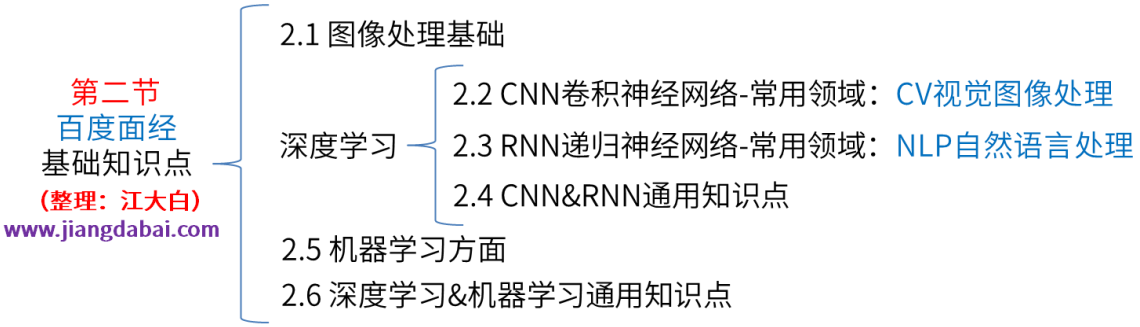
★ 很多问题是发散思维的，没有标准的答案，网上都难搜到，就是看平时有木有思考的习惯。

★ 百度是纯粹的 C++厂，code 能用 C++别用 python，当然事先最好准备一些 C++的常见问题。

★ 百度的面试除了考察基础知识之外，更重要的是知识的运用。考察题目灵活且仔细，所以没有一些运用经验或者较为细致全面的学习很难答得理想。算法题以及代码能力较为被看中，有想法的同学可以多多注意一下。

★ 有一个面试官说，主要从四个方面来考量候选人：是否有 ACM 奖项，是否有 CCF-B 级别的论文，是否参与过实验室的省级、国家级项目，是否参与过 Kaggle 竞赛。

2 百度面经涉及基础知识点



2.1 图像处理基础

2.1.1 讲解相关原理

- 边缘检测算法用过哪些?
- SIFT 算子的原理讲一下?
- 介绍 canny 边缘检测?
- 描述一下 ORB 特征?

2.1.2 手写算法代码

- OpenCV，截取 Mat 矩阵的一部分区域的数据的具体实现，以及 Mat 内存管理的机制?
- 对 M*N 图像实现均值滤波（k*k）？

2.2 深度学习：CNN 卷积神经网络方面

2.2.1 讲解相关原理

2.2.1.1 卷积方面

- CNN 的原理，3DCNN 的原理?
- CNN padding 的原理?
- CNN 中的平移不变性是什么意思?
- CNN 权重的初始化方法？CNN 权重如果全部初始化为 0，1 等任意常数可以么，为什么?
- Dropout 原理和实现方式？有什么优点?
- Dropout 的工作原理，以及在预测的时候他的作用是什么？怎么操作的?
- Dropout 是训练中的，那在预测的时候，是使用 dropout 训练出的权重还是要乘以 keep-prib 呢，为什么?
- CNN 为什么能提特征?
- 卷积神经网络是怎样工作的?
- CNN 了解么？CNN 在模型并行和数据并行上有什么差别?
- 空洞卷积优缺点

2.2.1.2 池化方面

- Pooling 原理和实现方式?

2.2.1.3 网络结构方面

- 讲一下 Resnet 的 Bottelneck 结构?
- Incepetion 的结构?

- 画了下 googlenet 的结构？讲了下 googlenet 跟之前的网络的不同。
- googlenet 中为什么采用小的卷积核?
- Xception 网络参数减少量?
- Vgg，Resnet，Densenet 了解吗?

2.2.1.4 其他方面

- 什么是一个端到端的学习?
- BN 的全称，BN 的作用，BN 一般放在哪里？为什么能解决梯度爆炸?
- Batch_Normal 为什么需要还原?
- 为什么 DNN 跟传统的机器学习方法有什么不同，为什么?
- 如何解决梯度消失和梯度爆炸？怎么解决过拟合（首先可以从数据角度下手，比如数据增强。其次有 dropout l1 l2 正则，另外还可以根据实际场景考虑减小模型大小，更换损失函数，例如 triplelet loss focal loss 等，面试官表示很满意）。
- 了解图模型吗？了解 Graph Embedding 吗?
- autoML 中知道什么算法?

2.2.2 公式推导

- 手推 BP 前向传播
- 手写 BN 公式
- 卷积维度变换的公式推导?

2.2.3 手写算法代码

- Softmax 的梯度公式推导和代码实现?
- 卷积是如何编程实现的?

2.2.4 激活函数类

- Softmax 层的 label 是什么？我回答 one-hot 向量。

2.3 深度学习：RNN 递归神经网络方面

2.3.1 讲解相关原理

- RNN，LSTM，GRU 各种的都介绍了一遍?
- RNN 求导，与线性层有什么区别?
- Transformer 跟 RNN 相比有哪些优势?
- LSTM 和 RNN 相比的区别有哪些？RNN 有哪些缺点?
- 知道哪些 RNN 模型?了解 GRU 不，了解 LSTM 不？LSTM 参数有几个矩阵，它们的维度分别是多少?
- LSTM 网络参数数量，计算公式的推导?
- 项目中用到了 LSTM，为什么可以用 LSTM，它的主要用处是什么，以及 LSTM 的梯度消失问题?
- LSTM 为什么要用 Tanh?
- LSTM 的优点，记忆单元是怎么工作的，他为什么可以克服梯度消失?
- 为什么把 CNN 结构放到 LSTM 之前，效果为什么比单独使用 LSTM 差，为什么不考虑 CNN+LSTM+CNN，论文里提到 CNN 对单字特征提取效果较好?
- CNN 与 RNN 有什么区别？RNN 为什么难以训练，LSTM 又做了什么改进?
- CNN，RNN，LSTM，Transformer 之间的优缺点?
- LSTM 为什么能捕获长期依赖关系?
- LSTM 中的 Attention 是怎么操作的?
- LSTM 结构，输入门，输出门，遗忘门怎么计算的，他们的作用分别是什么?
- CNN/RNN 是如何提取特征的，LSTM 和 RNN 之间的区别?
- LSTM 有几个门，讲一下？GRU 呢?
- 介绍 LSTM 结构，为什么这么设计？为什么三个门的激活函数是 sigmod？生成候选记忆的时候为什么用 Tanh?
- 度秘部门：LR，GBDT 的使用率要远远高于 NN，需有所准备。NN 的重点也应该放在 LSTM，

transformer，BERT 上。度秘的核心是对话系统，建议了解意图识别及常见做法，比如基本的 LSTM+CRF 模型。

2.3.2 手绘网络原理

- 画 LSTM 原理图?
- 画 GRU 的图，公式，与 LSTM 的区别。
- 手推 GRU

2.4 深度学习 CNN&RNN 通用的问题

2.4.1 基础知识点

- 数据增强有哪些方法?
- self-attention 机制原理?
- multi-attention 多头注意力机制的原理?
- attention 的概念，attention 的本质是什么?
- attention 里面的 QKV 都是什么，怎么计算的
- 讲一下对深度学习的理解，从 CNN、RNN 等多个方面介绍自己掌握的?
- 网络初始化有哪些方式，他们的公式，初始化过程?
- 神经网络初始化方法，我回答了随机初始化和 He 初始化，面试官问我随机初始化有什么问题，He 初始化解决了什么问题?
- 训练时样本不平衡问题如何解决，小样本问题如何解决?

2.4.2 模型评价

- 评估指标，P、R、F1，还有哪些（ROC 曲线，AUC 值），为什么不平衡时用 AUC，怎么算，代表含义，F 值的公式，还有没有其他的 F 值?
- 分类问题的评价方法，F1score，ROC，AUC，准确率和召回率?
- AUC 相关概念，是怎么做指标评价的?

- 常用的模型评估指标有哪些，也问到了对 AUC 的理解?
- 分类的评价指标，准确率，精准率、召回率是什么?
- AUC 原理介绍一下?
- AUC 评价指标具体意义，比如 auc=0.85 一个正例一个负例那这个 0.85 表示什么?

2.5 传统机器学习方面

2.5.1 讲解相关原理

2.5.1.1 数据准备

- N 个样本，N 很大，怎么做抽样?

2.5.1.2 特征工程

① 特征降维

- 知道哪些降维的方法，具体讲讲?
- 是否了解降维方法，我说了 PCA，他接着问我 SVD?
- PCA 推导?
- PLSA、LDA 有啥差别?
- LDA 和 PCA 的区别?
- 聚类 pca 懂么，讲一下，怎么操作的，为什么要算特征值? SVD 懂么（不懂），LDA 讲一下，跟 PCA 有什么异同。

② 特征选择

无

2.5.1.3 有监督学习-分类和回归方面

① 分类回归树（集成学习）

- Boosting 和 Bagging 的概念和区别?
- Bagging 和 Boosting 分别改善了偏差和方差的哪一个?偏差和过拟合的关系?

- XGboost、lightgbm、Catboost 三者介绍?
- RF 和 XGBoost 的区别?
- 问 GBDT 和 RF 的区别，我讲了以下 bagging 和 boosting，然后详细的说了以下个自的特点?
- 接着问了一下这两个个自的应用场景，我一下子答不上来，我感觉 GBDT 好像哪里都能用。RF 的话应该是用在特征不多的地方，然后我嘴欠说了特征不多，样本少，回头才发现 RF 只能用在大样本，小样本不适用

A.基于 bagging：随机森林

- 随机森林讲一下，随机森林的优点?
- 随机森林的随机性怎么体现?

B.基于 boosting：Adaboost、GDBT、XGBoost

- GBDT 不擅长处理离散特征，你在应用的时候是怎么处理的?
- 讲一下 GBDT 的原理？GBDT 在回归和多分类当中有什么不同，在预测的时候的流程是怎样的?
- GBDT 是否只能用 CART 树，GBDT 中残差计算公式?
- GBDT 跟 stacking 的区别?stacking 的流程
- GBDT 的实现、前向后向怎么做的，loss 怎么算的，梯度下降怎么求得。
- XGB 和 GBDT 的区别?
- GBDT 原理，树的权重，损失函数（分类树和回归树）
- GBDT 为什么用负梯度去拟合新的树？GBDT 的基分类器是什么?
- 结合项目，讲下怎么进行特征选择，特征工程方面，比如说非数值型怎么处理，one-hot 后维度高怎么。问你用的 xgboost 也许要对特征进行标准化吗？这里也考得 xgboost 吧，我感觉 xgboost 其实对特征的的预处理要求不是那么高，讲了下原理。
- 根据项目问了 XGBoost，LightGBM 原理上的区别，应用中的区别，遇到的难点，为什么会出现这样的情况?

- 传统的机器学习算法了解吗，xgboost 原理，为什么训练快?
- xgboost 和 gbdt 的区别？这方面问的很细，比如说 xgboost 可以并行加速是怎么进行的，每次分裂叶子节点是怎么决定特征和分裂点的。
- 为什么 xgboost 效果不如随机森林?
- xgboost 和 lightgbm 的区别?

xgboost 如何处理缺失数据?

- xgboost 如何防止过拟合，预剪枝和后剪枝?
- 介绍下 xgb 是如何调参的，哪一个先调，哪一个后调，为什么？哪几个单独调，哪几个放在一组调，为什么？哪些是处理过拟合的，哪些是增加模型复杂程度的，为什么?
- xgb 分裂节点的依据？xgb 如何处理离散值、连续值?
- 如果相邻的连续值比较接近，比如只有小数点后三位的差距，xgb 会遇到什么问题吗?
- 问了 lightgbm 有什么优势。我就把 gdbt，xgboost，lightgbm 从头到尾讲了一遍
- 简历写了集成学习，是项目中的 xgboost 吗？gbdt 讲一下，xgboost、gbdt 比较，为什么 gbdt 用负梯度不用残差。

② 线性回归

- 线性回归的共线性，如何解决，为什么深度学习不强调?

③ K 近邻（KNN）

- KNN 原理，kd 树的构建与搜索，讲原理，还有没有其他的树结构能实现 kd 树的效果

④ 逻辑回归 LR

- LR 和 SVM 的区别，从 Loss func 来说，LR 和数据分布有没有关系?
- LR 和 SVM，两个算法各自的使用场景以及它们之间的区别。
- GBDT+LR 的原理，如果 GBDT 有有 1 万颗树，每个树有 100 个叶子节点，那么输入到 LR 的特征会是一个高维稀疏的向量，那么应该如何处理，使用 PCA 降维的话会造成损失，如果不想有损失的话应该怎么办?
- GBDT+LR 中 LR 输入的特征都有哪些，除了 GBDT 输出的特征 有没有加入原始特征

- 项目中 LR 用的优化方法是什么，有没有用正则化，有没有调整 sgd 的步长
- 逻辑回归的特征处理，连续值、离散值，离散化连续特征的好处？
- LR 的损失函数，问了先验概率和后验概率的区别，让求了一下交叉熵损失函数的导数？
- 逻辑回归的原理
- 最大似然估计的作用
- 逻辑回归介绍一下？怎么解决过拟合？
- LR 最初是怎么提出来的，或者说为什么会出现 LR？
- LR 的目标函数是怎么得来的？
- 大规模 LR 参数稀疏解怎么求？
- 实际情况考察：如果 LR 训练中有 100 个变量，但是其中有 90 个变量是高度相似的，会对最终结果有什么影响呢？

⑤ SVM（支持向量机）

- SVM 的原理?如何用通俗的方式给父母讲 SVM？
- SVM 多分类怎么做到的（OVR、OVO、层次 SVM），分析各自的特点？
- SVM 损失函数、原始问题形式、对偶问题形式、引入核函数
- SVM 为什么可以处理非线性问题，怎么解决线性不可分的问题？
- SVM 为什么二分类效果最好？
- 知道 SVM 吗，怎么推导的? SVM 的目标函数是凸函数吗？有唯一解吗？
- SVM 能不能看成一种特殊的神经网络，或者说 SVM 和神经网络有木有什么联系？
- 介绍一下 SVM，遇到线性不可分怎么办，核函数有什么特点？

⑥ 朴素贝叶斯（Naive Bayes）

- 能不能用朴素贝叶斯某个特征在某一类的概率来选特征？（只衡量一个特征在一个类里出现的概率大小并不能用来筛选特征，个人感觉可以参考互信息法来回答）
- BPR（贝叶斯个性化排序）的理解？

- 贝叶斯公式，贝叶斯估计和极大似然估计的区别?
- 朴素贝叶斯思想

⑦ 决策树（DT）

- 说一下决策树吧，什么是决策树？节点划分有哪些方法，如何剪枝?
- 决策树深度很深会怎么样，叶子节点很多会怎么样?
- 回归决策树和分类决策树分裂节点的时候怎么处理?
- ID3,C4.5.CART 树的区别(一个个讲的，谁解决了谁不能解决的问题,如何解决的,缺失值如何处理,说了 scikit-learn 里面的 GBDT，原理上可以处理 missing value，但其实里面并未实现)
- ID3、C4.5、CART（比较具体）
- 决策树有哪些分裂方式，怎么计算的?
- 决策树的叶节点的怎么得到分数的?
- 树模型，ID3,C4.5,CART 怎么计算分割点的，信息增益和信息增益率的区别?
- 说一说决策树以及相关算法（gbdt、xgboost），区别以及各自的优势。
- CART 树怎么分裂节点？（分类和回归都要说）

⑧ 其他

- 说一下机器学习训练的过程，其中要考虑哪些因素（模型输入、损失函数的选择、优化函数）?
- 最小二乘法在什么条件下与极大似然估计等价?

2.5.1.4 无监督学习-聚类方面

- 简历上写了 k-means，面试官问我聚类有哪些具体的应用，聚类还有哪些高级的方法，各方法分别在什么情况下使用
- 改进现有的分类算法、聚类算法等，提出一种新算法，从哪个角度切入？（讲了感知机到 SVM 的改进，GBDT 到 xgb 的改进，kmeans 到结合语义的改进）
- Kmean 和 GMM 原理、区别、应用场景?
- 聚类算法了解程度、kmeans 介绍、K 值选择、kmeans++算法？kmean 如何选择 k?

- Kmeans 和其他聚类算法有啥优缺点?
- kmeans 是有监督还是无监督?
- 你说用聚类，你会怎么选择 K-means 的 K 值呢？K-means 会分类错吗？错了怎么处理?
- 为什么选择 DBSCAN，它的优缺点是什么。还知道那些聚类算法，各有什么优缺点?

2.5.2 手推算法及代码

2.5.2.1 手推公式

- 详细说明 xgboost，部分推导?
- 有什么常用机器学习算法，回答了几种，面试官问哪种能推导，回答 LR 和 SVM 都行，让推 SVM
- 讲解并推导一下 SVM?
- LR 详细的推导
- 手推 SVM, GBDT, Xgboost

2.5.2.2 手写代码

- 写一下 kmeans 的算法实现

2.6 深度学习&机器学习面经通用知识点

2.6.1 损失函数方面

- 损失函数都有哪些？（指数损失、平方损失、绝对值损失、对数损失等）
- Logistic 回归损失函数的公式和含义?
- 知道哪些损失函数？为什么分类问题不能用均方差?
- 谈一下交叉熵吧，公式怎么写的?交叉熵有什么问题?
- 二分类问题为什么不直接用 0-1 损失函数?
- 常见激活函数有哪些？为什么 sigmoid 要以这种形式呈现?
- 交叉熵损失为什么有 log 项?

2.6.2 激活函数方面

- 知道哪些激活函数，它们的优缺点分别是什么?
- 写出常用的激活函数及其导数?
- Relu 函数是做什么的，作用是什么?
- 激活函数讲一下，tanh 和 sigmoid 的关系?
- 介绍一下 sigmoid 和 relu，relu 有什么缺点?

2.6.3 网络优化梯度下降方面

- 优化方法的比较和选择，为什么 Adam 比 SGD 好?
- 不同梯度下降的方法，还有哪些降低损失函数值的方法，有哪些模型不是利用梯度下降迭代的?
- 讲一下梯度和导数的区别?
- 随机梯度下降和批量梯度下降的区别?
- 为什么神经网络要用梯度下降法优化，而不用乘子法，牛顿法等优化?

2.6.4 正则化方面

- L1、L2 正则化的区别是什么？应用场景?
- L1、L2 正则为什么可以防止过拟合?
- L1 正则相当于拉普拉斯先验，那么在损失函数为最小二乘法的时候，如何通过拉普拉斯先验推导出 L1 正则?
- L1 正则是不可导的，那么在这种情况下如何优化求解损失函数?
- L1 为什么能使得特征变得稀疏？L1 解空间为什么长成菱形?
- L1/L2 正则化及对损失函数造成影响的区别?

2.6.5 压缩&剪枝&量化&加速

- gru 与网络轻量化？叙述了 gru 结构，网络轻量化提到了 sgru 中的权值共享，提到了两个 3*3 的卷积核可以带替一个 5*5 的卷积核。

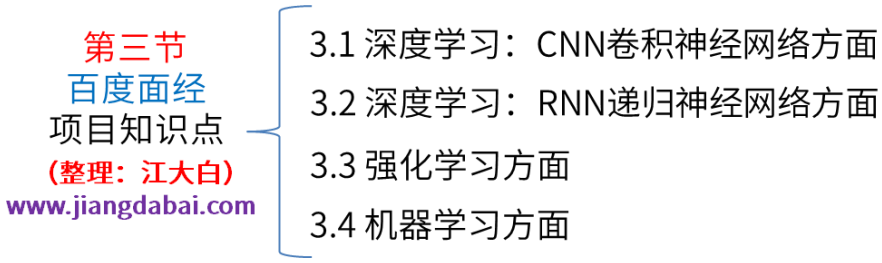
2.6.6 过拟合&欠拟合方面

- 过拟合怎么判断，如何解决过拟合?
- 过拟合、欠拟合原因及解决办法?

2.6.7 其他方面

- 代价函数、目标函数、损失函数的区别?
- 怎么计算相似度 ？
- 样本不平衡怎么处理?
- 算距离时余弦相似度和欧式距离，什么情况下两者可以等同?
- 对于不同场景机器学习和深度学习你怎么选择，你更习惯机器学习还是深度学习?
- 在模型训练的时候有什么技巧和经验，比如 loss 不收敛等情况?

3 百度面经涉及项目知识点



3.1 深度学习-CNN 卷积神经网络方面

3.1.1 目标检测方面

3.1.1.1 讲解原理

- 说一下 Faster R-CNN，要详细画出图，说一下 ROI polling 与 RPN?
- Rcnn，Fast-Rcnn，Faster-Rcnn，SSD，YOLO，FPN，MASK RCNN，Cascade RCNN，都简单的介绍了一下?

- 讲一下 Two-stage 和 One-stage 的异同?
- 目标检测的时候出现一半的物体的处理方式?
- 讲一下近两年比较新的结构上的改进 ？
- 小目标检测怎么解决?
- 从 yolo v1 开始讲一下 yolo 的历程？因为 yolo v3 比较熟，能不能从 yolo v3 开始，被允许。介绍了 yolo v3 的网络结构，draknet53，多尺度特征图上的预测，损失函数。
- Yolo 怎么从 Anchor 变成具体坐标的？w，h 的预测为什么要乘缩放系数?
- 介绍下 yolov4 相对于 yolov3 的改进？mosaic 有效的原因?
- Anchor free 的方法了解多少？讲一篇比较熟悉的
- YOLOV4 的改进在哪里?
- 解决难样本问题的方法，ohem 与 focal loss 的相同点和不同点
- ssd 对小目标不好的原因
- DCN 的原理与实现，感受野的含义
- fpn 的实现细节，anhcor 在不同层分配方式
- anchor 聚类的具体实现与原理
- 介绍自己对整个目标检测领域的看法(不是宏观) 基本上从两阶段讲到一阶段再到 anchor-free 再到 trasformer。
- yolov5 的改进工作
- Giou Diou Ciou 改进方法

3.1.1.2 损失函数

- Focal loss 的共识写一下?

3.1.1.3 手写代码

- 写一下非极大值抑制（NMS）
- 手撕 SoftNMS

3.1.2 图像分割

- Bisenet 网络的改进目的?
- 分类的预训练模型如何应用到语义分割上?
- 语义分割上采样的方法?
- 问了反卷积是怎么做的， unpooling 中 maxPooling 怎么实现?

3.1.3 OCR

- OCR 识别有哪些算法模型?

3.1.4 图像分类

- 如果图像分类有百万个 class，你会怎么设计模型?

3.2 深度学习-RNN 递归神经网络方面

3.2.1 自然语言处理 NLP

① Bert

- 文本分类，bert 了解吗? 输入有什么改进?
- 讲一下 Bert 原理?
- Bert 模型结构，分类和句子翻译如何微调?
- Bert 的输入，token embedding 是怎么得到的? 输出是什么，维度是什么?
- Bert 里面的 mask 为什么有效?
- 做两个句子的语义相似度，Bert 结构怎么 fine-tuning?
- Bert 的 embedding 向量怎么来的?
- Bert 细节（mask 原理，训练过程的两个任务，输入输出，为什么要遮蔽部分单词，80%和 20%等具体问题）
- 怎么用 Bert 来做搜索?
- 介绍 Bert，写 attention 公式

- Bert 怎么做专有名词识别和词重要性识别?
- 问 Bert 和 Xlnet 的互相弥补与改进?
- Bert 的 word piece 对于中文词表的一个好处，词表变小也可防止 OOV?
- Bert 中用的 LN，LN 和 BN 有什么区别，为什么 Bert 用 LN?

② Transformer

- Transformer/GPT/BERT 的原理简单讲解?
- GPT/BERT 中分别是怎么用 Transformer 的?
- Transformer? 比 RNN 优越在哪?

③ Attention

- 简单的介绍下注意力机制的原理?

④ CRF

- CRF 介绍，CRF 是怎么优化的（L-BFGS)，L-BFGS 是什么，为什么用这个?
- CRF 特征函数是什么？输入是什么？和 HMM 有什么不同？什么原理？怎么训练?

⑤ HMM 隐马尔科夫模型

- HMM 和 CRF 的区别以及原因，HMM 参数有哪些?
- HMM 的原理以及公式?
- HMM 了解吗？什么原理？怎么训练?

⑥ Word2vec

- 解释 Word2vec 原理，两种模型结构，两种改进方案?
- Word2Vec 怎么对词向量进行特征的抽取?
- Word2vec 怎么训练的，有没有没有得到的词向量，比例多少?
- Word2vec 训练方式哪种更好?
- Word2vec:训练目标是什么，skip-gram ,cbow, 层次 softmax, 高频词和低频词训练出的表示有什么特点， 怎么解决 ？

- Word2vec 里面为啥要负采样?
- word2vector 如何做负采样？是在全局采样？还是在 batch 采样？如何实现多 batch 采样？怎么确保采样不会采到正样本？word2vector 负采样时为什么要对频率做 3/4 次方?
- Word2Vec 的输出是什么？损失函数是什么？怎么训练？怎么优化?

⑦ 其他

- 关键词提取的方法，TF-IDF 会不会？TF-IDF 公式?
- TF-在提取关键字的时候有没有遇到问题?
- tf-idf 公式是什么,对于低频词和高频词有处理么,高频词算出来的 tf-idf 的值会更大还是更小?
- fasttext 原理，有什么好处?为什么可以解决未登录词?
- TextCNN 原理以及和 CNN 的区别?
- 问了不了解 graph embedding？讲了 deepwalk，node2vec 区别和具体细节。
- Node2vec 中分别以 BFS 和 DFS 的方式游走会对最终的推荐结果产生什么影响?
- 稀疏词向量，用 skip-gram 还是 cbow 训练好?
- NLP 怎么处理文本特征?
- n-gram 模型原理，有什么作用？使用中有哪些缺点?
- 实体抽取的相关算法（bilstm+crf），（面试官补充现在最新的是 bert+crf）
- fasttext 和 word2vec 的区别（项目中涉及 fasttext）
- 如果关键词和所有商品全部用 Dense Vector 来表征，如何快速匹配最相似的?
- 如果商品可以标注，你会标注什么信息?
- 商品的什么特征是最重要的，你如何提取这些特征来做关键词匹配?
- 对话系统中，DST 的作用，DST 和 NLU 的区别是什么?

3.3 强化学习

3.3.1 讲解原理

- 结合项目，用 GAN 网络去噪，解释了下 conditional gan 的原理。损失函数的创新点?

3.3.2 损失函数

- GAN 网络的 Loss 讲一下?

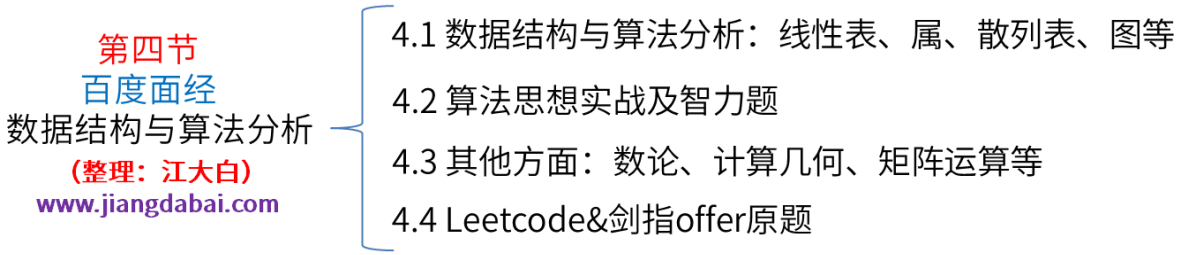
3.4 机器学习方面

3.4.1 推荐系统

- CTR 和 CVR 怎么实现?
- 不定长文本数据如何输入 deepFM，如果不截断补齐又该如何输入?
- FM、ffm、DeepFM、wide&deep 的区别和联系，然后让我用 TensorFlow 实现 DeepFM?
- UserCF 在现实场景中实现遇到的问题，如何解决?
- 介绍了一些经典的推荐系统算法，再介绍了一些基于深度学习的推荐算法
- 协同过滤（基于用户，基于内容），矩阵分解及其后续改进
- 异质信息网络中的异质信息是什么，如何构建异质信息网络
- 如何从异质信息网络中提取 user，item 的 Embedding
- BPR（贝叶斯个性化排序）系列，CDL（基于 MF 架构引入自编码器提取 item 特征），CML（度量学习范畴），NCF，RRN（基于 RNN 建模用户历史偏好），基于强化学习的推荐算法等算法的了解?
- 协同过滤了解吗?
- 基于用户和基于 item 的协同过滤讲一下
- 用户冷启动和 item 冷启动应该用什么策略?
- widedeep 与 deepfm 区别?
- FM 相比于 LR 的优势（自交叉、稀疏特征不太影响训练、可以得到 embedding，进行高维交叉，推理未出现过的特征组合）
- 介绍一下 DIN 和 DIEN?
- 知道哪些推荐算法? （提了很多，后来点了一下百度比较出名的双塔模型）

- 双塔模型有什么问题？（我说不能实时反应客户的行为？）
- 如何改进双塔模型？
- 10 个字以内关键词搜索，如何从 一千万个商品（只有标题和长文叙述）中快速检索 Top10？
- 推荐系统如何解决马太效应？除了挖掘长尾
- 离散、连续特征如何拼接？多模态特征怎么融合？多路召回怎么融合？
- 如何解决广告位置 bias？单点预估(无位置信息)怎么预测 ctr？
- 如何在不降低总体指标的情况下增强 ctr 模型实时性？除了增量学习
- 如何填充曝光未点击样本的点击率？
- 如何 evaluate 新 feature 是否 work 带来提升？除了 abtest
- 场景题：搜索场景下有监督无监督时候 query 匹配如何融入 ctr 到词重要性任务？

4 数据结构与算法分析相关知识点



4.1 数据结构与算法分析

4.1.1 线性表

4.1.1.1 数组

- 给定一个数组，写一个函数来随机打乱这个数组？
- 给定两个整数数组，对第一个数组进行排序，整数顺序由其在第二个数组中的位置决定，对于没有出现在第二个数组中的整数，应排在末尾，其之间顺序无限制。这里整数的取值范围是[0, 2^32-1]，例如第一个数组为 5 1 6 2 1 2 3，第二个整数数组为 2 1 3，则排序结果为 2 2 1 1 3 6 5 或 2 2 1 1 3 5 6

- 两个序列的最长公共子序列
- 二维有序数组的二分查找某个数?
- 给一个数组，找出最小缺失正整数（不能排序，O(n)时间复杂度）
- 从数组中找到两个和为给定值的数字?
- 搜索旋转排序数组
- 二维数组路径和最小
- 连续子数组的最大和
- 统计一个文本中的出现次数最多的 k 个单词?
- 旋转数组找 K 值?
- 左右有序上下有序的数组中查找指定数?
- 无序数组找到第 k 大的数，写出两种做法
- 找出一个数组中的第 k 小元素，我是用快排的变种做的。平均时间复杂度是 O(n).但是面试官的理想答案是用堆。
- 统计一个文本中的出现次数最多的 k 个单词?
- 返回数组中和为指定值的两个元素，如果有多个，返回成绩最小的那个?
- 有一个特别长的数组，放不进内存的情况下，找出最小 k 个数?
- 有序等长数组找中位数
- 给一个数组，一个目标数，找出数组的两个数相加等于目标数
- python：两个无序数组去重合成一个升序数组
- 两个数组找乘积最大的
- 最大连续子数组的和
- 两个有序数组，随意挑选两个值，求其和，求第 k 大的组合
- 动态求区间和(树状数组)
- 两个排好序的数组求交?

- 排好序的数组，里面有重复的元素删除重复元素，只保留其中的一部分?
- 一个有序数组，找到两个数相加为 x？问空间时间复杂度。
- 1000 个数的数组，每个数在 1-999 之间，有两个数是一样的，找出。要求时空复杂度最优。时间 O(n)，空间 O(1)。
- 1，0，1，0，0，1.....】1 可以变成 0，0 可以变成 1 求最少多少次操作可以让 0 后没有 1?
- 数组，可以分别从最左边最右边取个数字，求取得 k 个数的最大值，O（1）空间呢，k 的取值范围的条件?
- 给你一个数组 A，数组里按顺序存的一组点，表示一个多边形，再给你一个点 B，问如何判断点在多边形内部？（面试官说把多边形分解成有限个三角形，去判断点是不是在三角形内）
- 找数组的众数
- 二维数组路径和最小
- 数组往右移动 k 位
- 两个有序数组的最小的共同数

4.1.1.2 链表

- 翻转单链表
- 判断链表是否有环以及环的位置?
- 找到环形链表的入口
- 链表公共节点
- 删除链表节点
- 两个单向链表，怎么判断他们是否相交，交点在哪里?
- 两个链表是否相交，自建链表测试
- 手写链表排序
- 链表的排序：merge sort
- 两个有序链表的归并

- 合并两个无序链表成为有序链表
- 判断相交链表
- 在一个排序的链表中，存在重复的结点，请删除该链表中重复的结点，重复的结点不保留，返回链表头指针？
- 两个链表分别表示两个数，头指针为低位尾指针为高位，求和返回新链表？
- 数组和链表的区别
- 链表是否回文
- 两个链表相加
- 对倒排索引的认识，在此基础上，两个倒排索引得到的链表，得到其公共部分转换成两个有序链表合并的问题

4.1.1.3 栈

- 两个栈实现一个队列（优化方法）

4.1.1.4 队列

4.1.1.5 字符串

- 字符串切割（raw:adbacaf　sep:bac　　result 就是 ad 和 af）
- 字符串反转
- 如何将（a,(b,c,null),(d,(e,f,g),(h,null,(i,j,k)))）这样的字符串转为一颗二叉树？
- 字符串转 float
- 有 10G 的数据，每行数据是一个字符串，得到前 K 多的字符串，最后去重？
- 给定一个字符串，一个子串集合，要求不断删除字符串中的子串，直到没有可以删的为止？
- 给一个字符串 s,一个目标子串 t,写一个函数判断 s 中是否包含 t 的旋转串？ 旋转串的定义是，比如 t = "abcd",则"abcd"、"bcda"、"cdab"、"dabc"都是 t 的旋转串。但是因为时间仓促，用暴力判别做的。 写完之后被问优化，答的是遍历 s 中的每一个字符，只要首字母确定，旋转

串就是固定的，只需要直接判定就好。

- 最长公共字符串，动态规划思路
- 求和大于 k 的子串的最小长度?
- 一个字符串的最长回文串，有没有效率优于平方的算法?
- 给一个字符串和一个字符串集合，问给定字符串是否能被集合中的字符串组成
- 在两个大文件中，找单词的交集?
- 最长公共子串
- 给一个字符串 s，再给一个目标字符串 t，你可以对字符串中的每一个字母进行以下三种操作之一，问从 s 变成 t 最少需要多少步。

操作 1：删除改字符

操作 2：增添一个任意字符

操作 3：将该字符换成别的任意字符

- 给定 txt 文件，里面每个都是以逗号分隔的字符，要求按照某列来对另外两列求和，并按照其中一列逆序，如下：

A B C

1 2 3

1 4 5

1 6 7

2 3 8

2 7 9

按 A 列对 B 和 C 进行求和，并将结果按照 B 逆序排列，不能用 pandas 中的 groupby 功能，只能用 list，dict 和 tuple

- 求字符串的子集
- 判断字符串中左右括号是否合法
- 给定一个字符串和数组，判定字符串是否能由数组中的字符串组成

- "today is good"转换为"good is today",要求 O(1)空间（只说思路）
- 最长回文子串
- 口述给定一个字符串，找出第一个重复字符，时间/空间复杂度，如何优化？
- 字符串中两两成对连续出现，只有一个单独出现的字符，找出这个字符？

4.1.2 树

4.1.2.1 二叉树

- 二叉排序树特点，平衡树特点，最小生成树算法有哪些？
- 实现二叉树查找，删除节点和反转
- 判断两个二叉树是否一样？
- 逆序对和判断一个树是 bst？

Z 字型打印二叉树

- 求二叉树的所有左叶子节点和？
- 二叉树的右视图
- 层序遍历二叉树
- 判断镜像(对称)二叉树
- 手动构建二叉树，dfs 遍历
- 旋转二叉树+二分查找（递归非递归）
- 二叉树层次打印
- 二叉树最近公共父节点、并查集
- 二叉树判断有没有和等于 sum 的路径
- 求解完全二叉树节点个数，要求是必须要用到完全二叉树的性质
- 实现前缀树的插入，查找以及某一个前缀的所有词查找？
- 字典树如何构建以及使用场景？
- 简单的二叉树最大最小值

- 红黑树的几个特点?
- 二叉树两节点最长距离?
- 二叉树中的最长路径
- 求二叉树最大深度
- 平衡二叉树是什么?
- 二叉树的先序遍历，中序遍历，后序遍历（要非递归），时间复杂度
- 二叉树的 Z 字形遍历
- 判断是不是后续遍历中序二叉树?
- 二叉搜索树给定一个节点查找下一个节点?
- 两个节点的最近公共祖先
- 找一棵二叉搜索树里面给定两个节点的公共祖先?
- 排序数组能够构成多少个二叉搜索树?
- 判断一个树是否为二叉搜索树？递归或中序遍历后看是否为递增序列。
- 二叉树是否存在一个子树，子树节点和为整个树节点和的一半?

4.1.2.2 堆

- 介绍堆排序，以及其它排序?
- 最大堆的插入

4.1.3 排序

- 排序算法，从稳定性分析?
- 排序算法有哪些?各种排序算法时间复杂度?空间复杂度?稳定性?什么实际情况下考虑稳定性 ?
- 快排要稳定的话怎么写
- 写个归并排序
- 归并排序（使用 C++完成）

- 文件太大，内存太小时候，最大的 K 个数怎么样的求解过程?
- 海量数据中选取 topK 大的数（堆排序和快排实现），给出时间复杂度分析
- 百度有海量的搜索词记录，返回 TopK 个高频词?
- 从 1-1000 中找到缺失的值（用字典），一堆乱序数中找到第 k 大的数（快排或堆排序，如何实现，复杂度为多少 klogn），传统快排复杂度?
- 手写快速排序
- 给定一个数组[-1,2,3,4,7,-4,-5,8,0],使用时间复杂度为 O（N）的算法求解?
- 排序，奇数在前，偶数在后
- 为什么快排的时间复杂度是 nlog(n),快排和归并排序有什么区别?
- 手写堆排序
- 堆和栈的区别?
- Top-K 问题，手写堆排

4.2 算法思想实战及智力题

4.2.1 算法思想实战

- 青蛙跳台阶，变态青蛙跳台阶
- 爬楼梯问题（动态规划）
- 蓄水池抽样
- 100 层楼 2 个小球问题
- 硬币兑换最小数
- 零钱兑换
- 类似登山问题
- 给定高考全部人分数，查找一个分数的排名多少?
- 写一个 string to float 函数，进行浮点数转化，但是要考虑浮点数各种表示方法，比如科学计数法等。

- 开根号（经典题目，牛顿法和二分法）
- 返回数字 1-n 的全排列
- 找出岛屿数量
- 根据词频随机出 K 个词
- 给定 n 个数，找出 pair 的个数，满足 x&y >= x^y
- 翻译数字为英文
- 例如，1001，翻译成 one thousand one

1101，翻译成 one thousand one hunderd one

0~4 随机函数，怎么生成 0~6 的随机函数?

- 有效括号用了什么数据结构（栈）
- 编辑距离，并且回溯出路径。进一步扩展问题：如果有很长的两个串要计算编辑距离，但是如果发现他们的编辑距离>5 就不需要继续计算了，直接返回过大。可以优化编辑距离算法吗?
- 图-路径问题：假设现在有很多海岛，有些海岛之间有桥连接，你已知海岛连接情况。

① 我现在想从到 A 岛 去 B 岛，问是否能通过陆路到达（过桥）

②如果可以，最少需要过几次桥

③输出一条最短的路径（P.S. 最好可以写一个非递归的形式）

- 输入是一个字符串流每一行长度在 2～400 之间，要求按照 batch size 中所有字符串的长度差异<=2，每个 batch 之间输出一个空行，输入的每行在输出中不变，顺序无要求，请给出尽量最小的缓存需求下的解决算法，并给出保障所有情况可运行的所需最小缓存大小的计算方法。
- 给一组宽度相同、高度不同的图片，将这组图片首尾拼接为两部分，使这两部分的高度尽可能接近?（其实是套着业务外衣的一道 leetcode 算法题，就是给一组数（高度），把这组数分成两部分，让这两部分的和尽可能相近）
- 外存上有大量文件，文件里有大量已排序的数字，而内存又容不下一个文件，问如何全局排序（给思路就行）。这属于很经典的外部排序问题。

- 在内部排序里有什么方法很快能找到当前最小的数?
- 数组 b, c 含有相同的元素集合(顺序不一定相同), 各自内部不可比较,求 b、c 元素的对应,要求时间复杂度<=O(NLog(N))
- N*M 的矩阵中，有两种齿轮，一种是初始状态朝上，每次触发会顺时针旋转 90 度；另一种是初始状态朝右，每次触发会逆时针旋转 90 度，找到矩阵第一列的元素，使得触发它后可以走到矩阵的右下角并且以朝右的方向出去。
- 一个人从家往公司走，只有一条路径，有东西南北四个方向，现在求这个人往东南走的最远的距离?
- a~z 和 A~Z 代表 0 到 51，如何将 52 进制转化为 10 进制?

4.2.2 智力题

- 64 匹马,8 个赛道,找出前 4 名最少比赛多少场?
- 有 N 个甘蔗，甘蔗有多个节，每个节的长度不一样，所有甘蔗头对齐，问随便切一刀，做多能同时砍到几个节点?
- 10 个人打牌，怎么样迅速分配输赢的钱?
- 一千瓶水，一瓶有毒，十只老鼠做实验，老鼠服用有毒的水之后一周后死亡，一周之内测出哪一瓶水有毒?
- 给两个砝码 7g 和 2g，有 140g 盐，分成 50g 和 90g 两堆盐，天平用三次分出这两堆来?
- 100 块钱，每次可以花 1、2 或者 3 块，有多少种花法?
- 10 只老鼠 1024 瓶药水，一次找到其中的唯一一瓶毒药 ？（利用编码的思想，将药水编码成 01 的格式，某一位置 1 代表对应老鼠喝该瓶药水）
- 有 N 个硬币排成一排,每次要你从最左边或者最右边拿出一个硬币。总共拿 k 次,写一个算法,使能拿到的硬币和最大?
- 区域中有 m 个人和 n 个怪，当人走入怪的警戒范围内的时候就会报警，描述一个算法，判断是否会报警?
- 想象桌面上有一堆点，把手机放上去，如何快速找到被手机压到的点?

- 智力题：给一个很大的数字，以 5 为结尾，如何快速知道这个数是哪个数字的六次方
- 智力题：有容量为 3、5、8 升的被子，怎么量出 4 升，代码怎么实现(不会，面试官提示 bfs)。

4.3 其他方面

4.3.1 数论

- 求解优化问题的方法，讲出拉格朗日函数，对偶问题?
- 判断质数
- 1-n 中数字 1 的个数?
- 求 1 到 n 之间的素数个数
- 10 进制转 8 进制
- 求一个整数的平方根

4.3.2 计算几何

- 给一个点和一个矩形（由长、宽和旋转角度表示），判断点是否在矩形中?
- x^2+y^2+z^2=1 的球面随机取点，怎么取，应该是想让我说极坐标的方法?

4.3.3 概率分析

- A,B 约定在 12:00 到 13:00 见面，先到者等待 15 分钟，求见面概率?
- 一个期望的题 每次抽球概率是 1/4，前三次抽中则第四次必中，求抽一百次的期望?
- 一个箱子，里面不知道有多少个球（球按照从大到小编号了 1-n，但是 n 未知），现在取 k 次，每次取 m 个球（有放回），问你 估计下里面一共多少个球?
- 进阶问题：如果不知道每次取 m 个球是哪些球，只知道他们的方差，怎么估计箱子里面的球?
- 圆上三个点组成锐角三角形的概率?
- 一副扑克 52 张，没有大小王，随机取两张，都是红桃的概率（1/17）
- 2 个盒子，50 个红球和 50 个白球，怎么放使得摸到红球概率最大（计算步骤）

- 给 N 个数，每个数有一个概率值，要怎么依据概率得到每个数?
- n 个数中等概率抽取 m 个数?
- N 个球取 M 个，每个球被取走的概率
- 抽小球这样的概率题，用递归做的?
- 男性得色盲的概率是 5%，女性得色盲的概率是 0.25%，求一个人是色盲的前提下是男性的概率。(贝叶斯公式推导一下就可以,当时不知道 P(男)怎么算,后来面试官说可以认为是 1/2)
- 一个长度为 n 的 list，元素不重复,从中取出 m 个数，问这 m 个数中某一个元素被取出的概率是多少?

4.3.4 矩阵运算

- 从矩阵左上走到右下的最长路径，只能向右或者向下？用动态规划 code 的?
- 矩阵中给定两个点作为对角线组成的矩阵里的数的和（复杂度 O（1），可简单预处理）
- 最低公共祖先，01 矩阵中最大全 1 矩形的面积?
- 矩阵逆时针旋转 90 度
- 有一个 n*n 的数字矩阵，我要将其向右旋转 90 度，应该如何实现，时间复杂度是多少?
- 将 mxn 的矩阵 0 元素对应的行和列的其他元素变为 0

4.3.5 其他

- 给定一个带有权重的无向图（网）,求初始点 A 到其余各个顶点的最短距离?
- 双数组树的原理?
- 求一个数的平方根（牛顿迭代或者二分）
- 证明一个式子是否收敛，给定一个长串式子，先求出通项，然后判断是否收敛 sqrt(asqrt(asqrt(a*sqrt(a)...)))
- 求 1+2+3+...+n。不能使用乘除，for，while，if，else，switch。case 等关键字以及条件判断语句(A?B:C)
- k 个的列表反转?

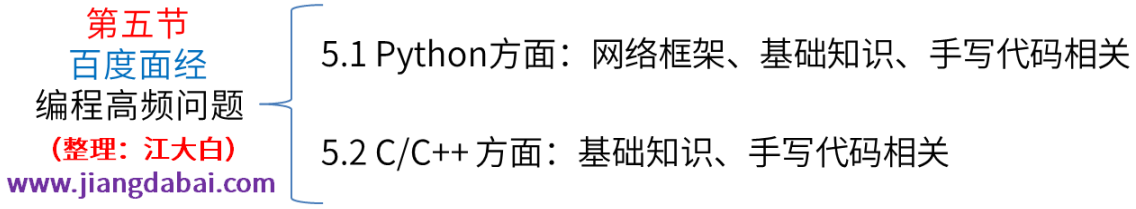
- 给你 2 个分好词的文件 A 和 B，A 是亿级，B 是万级，内存 1G，判断 B 是不是 A 的真子集?
- 给定数据找两个数之和等于 target，分析复杂度?
- 将 IP 字符转化为整数(127.0.0.1 转化为 16777343)?
- 集合的子集，无重复和重复。
- 给定一串数字，找到其中连续最长的序列，使其求和为 0。解题思路用积分图的思想解决?
- 读取文件（一行一个，有顺序 a-z），统计元素词频
- 长度为 n 的数字，找出最大值最小值，直观为 O(2n)，另一种解法：两个元素比较，大的最大值比，小的和最小值比，共 O(1.5n)
- 贝叶斯公式你给我写一下、
- 卡方检验知道吗，你知道里面的 P 值是什么吗?

4.4 Leetcode&剑指 offer 原题

- Leetcode120
- Leetcode 279：完全平方数
- Leetcode 2：链表相加，需要注意的是最好提前练习如何写链表的示例，因为我们刷题只用写个函数
- Leetcode 原题：二叉树的最短路径
- Leetcode 原题：岛屿问题
- Leetcode 原题：一个有序数组，旋转后，判断是否存在某元素
- Leetcode 原题：找到链表交叉点、
- Leetcode 原题：数组全排列
- Leetocde 原题：threeSum
- Leetocde 原题：爬梯子问题，变态的爬梯子问题
- Leetocde 原题：股票买卖问题（三种情况）
- Leetcode 原题：数字组成的字符串，折算成小写字母（1-26）的所有可能性个数

- Leetcode 原题：数字组成的字符串，折算成小写字母（1-26）的所有可能性个数
- Leetcode 原题：给一个 m * n 的矩阵，求左上角到右下角共有几条路径？只可以往右和往下走。
- Leetcode 原题：给定一个数组，其中有多个单词，判断一个给定字符串中是否包含所有的单词。
- 剑指剑指 offer 4：二维数组的查找，有无别的思路(提示递归)。
- 剑指 offer 原题：字符串旋转
- 剑指 offer 原题：数组中出现次数超过一半的数字，给了两种解法
- 剑指 offer 原题：判断五张扑克牌是不是顺子
- 剑指 offer 原题：在一个数组中找两个不同的数字
- 剑指 offer 原题：判断单链表是否有环？找到入口节点

5 编程高频问题：Python&C/C++方面



5.1 python 方面

5.1.1 网络框架方面

5.1.1.1 Pytorch 相关

- 主要用的是 pytorch 还是 tf，介绍一下 torch 构建一个简单的 dnn 的流程

5.1.1.2 Tensorflow 相关

- tensorflow 中动态图静态图
- tensorflow 模型剪枝

- tensorflow 中有哪些关键技术，tensorflow 如何的动态图计算框架

5.1.1.3 其他

- Paddle 用过吗？静态图与动态图的优缺点？

5.1.2 基础知识

5.1.2.1 线程相关

- Python 进程和线程的知识点？
- Python 内部实现的多线程有什么问题？

5.1.2.2 内存相关

- Python 的内存管理机制，优缺点
- Python 内存管理，内存池最大？

5.1.2.3 区别比较

- Python 深、浅拷贝的区别，如何使用？
- Python2、3 的区别？
- Python 当中迭代器和生成器的区别？
- Python yeild 和 return 的区别？

5.1.2.4 讲解原理

- Python 的装饰器都有什么作用，如何计算多个函数花费的时间？
- Python 的类中加下划线的含义？
- Python 闭包、反射、装饰器、GIL
- Python 中的闭包，python 如何在类中修改全局变量
- yeild 是什么？
- python 深浅拷贝

5.1.3 手写代码相关

- 读取文件（一行一个，有顺序 a-z），统计元素词频
- Readline 和 Readlines 的区别
- Python lamba 与 def 定义函数的区别
- 写一个函数，实现 python 的继承，数据的交换，类中的全局变量等等
- Python 中 +和 join 的区别。
- Python 写一个生成器:range(start, end, step),应该就是考知不知道 yield;然后问相比于 list，生成器优点

5.2 C/C++方面

5.2.1 基础知识

5.2.1.1 内存相关

- 线程通信的方式有哪些（共享内存）
- C++内存泄漏的原理，C++如何解决 C++内存泄漏的

5.2.1.2 区别比较

- C++的 vector 与 arrylist?
- new、delete、malloc、free 的区别
- 虚函数与纯虚函数的区别引用和指针的区别?
- 指针和引用的区别?
- 问了 C++，指针，函数指针，变量存储位置，垃圾回收机制，虚函数等等
- 死锁如何产生，如何解决?
- 深拷贝浅拷贝，is 和==之间的区别?
- Structure 和 Class 的区别?
- C++ vector reserve&resize?

- C++ class && struct?

5.2.1.3 讲解原理

- C++是否支持多继承?
- C++ 纯虚函数，抽象类能不能被实例化?
- Vector 分配的内存不足时的底层操作
- C++智能指针
- C++ 变量定义在函数内，存储在什么区?
- Static 讲一下?
- const 讲一下，const char *&a
- 虚函数在什么阶段创建?
- C++结构体初始化时，什么时候不写构造函数会报错
- C++的 vector 底层实现
- C++的 sort 底层实现

5.2.1.4 讲解应用

- LRU 算法实现?
- Set 的查找时间复杂度

6 操作系统高频问题：数据库&线程&常用命令等

第六节
百度面经
操作系统高频问题
（整理：江大白）
www.jiangdabai.com

6.1 数据库方面：基础知识、手写代码相关

6.2 操作系统方面：TCP、线程&进程、常用命令相关

6.1 数据库方面

无

6.2 操作系统方面

6.2.1 TCP 协议相关

- Tcp 的三次握手?
- TCP 三次握手，为什么要三次，和四次有什区别?

6.2.2 线程和进程相关

6.2.2.1 区别比较

- Windows 和 Linux 内存管理对比?
- 深拷贝和浅拷贝的区别?
- 进程和线程有什么区别?
- 进程间通信、线程间通信有哪些?

6.2.2.2 讲解原理

- 进程之间如何通信?
- 操作系统的锁了解吗？有哪些种类，自旋锁了解吗？和其他锁的区别
- 死锁产生的原因，死锁避免的方法有哪些?

6.2.3 常用命令

- 对 Linux 的系统的了解，如何统计磁盘使用情况的命令?
- 用过最复杂的 linux 命令是什么?

7 技术&产品&开放性问题

7.1 技术方面

- 向量相似度计算方法？（欧式距离、余弦距离）
- 关于神经网络的一些问题

①陷入局部最优解如何进行解决，这个问题应该答加入带有动量的优化器，或者求二阶导数

②求解二阶导数的时候的时间复杂度变成多少

③计算机矩阵求逆的计算复杂度

● 12G float 数据，256M 内存，怎么把这些数据排序?

● 模型在线下可以得到很好的效果，但是上线后效果不好，有哪些原因？项目中如何判断是否拟合、如果离线数据不能很好反映全集的情况如何处理

● 调参一般调试哪些参数 ？

● 从度秘的一个场景题引入，问了我如何设计度秘的音乐推荐的 embedding。如果用简单的 word2vec 怎么做， 存在什么缺点，面试官说音乐可能不像 nlp 任务有那么强的顺序性，怎么改进？【后来我查了以下就是微软 item2vec 的思想】

● 现在有一亿个样本，你如何找到单词最相似的?

● 一亿数据量的文本怎么做到随机抽取和重复统计?

● 假设现在有一万个数据，每个数据被取到的概率是不同的，1/2,1/3,1/100,现在如果是你，你会怎么取这一批数据？，数据有好有坏

● 给我看一个数据集（200*2），然后现场搜 api，来做一个分类。我快速的口述了一下 ml 的流程，数据分析，数据清洗，选用模型，评价标准。

然后选用了 LR，搜 api 一半，面试官说 ok，看来流程是 ok，不用写了。然后面试官引导我，这个数据集比较简单是不是可以用__比较简单，我扯了半天愣是没说到他想听的 knn。

● 情景题：预判某所大学的某个专业，在某个省份下一年的分数线（从特征选择到模型的选择，以及为什么用这个模型，为什么不用其他某个模型）

● 天池大数据骗保现象的原因，数据比例分布，数据不平衡怎么做，SMOTE 采样怎么做的，SMOTE 原理，采样完的比例分布，用的算法（RF、xgboost）

● 从头至尾介绍垃圾邮件识别的过程，从数据获取至生成报告，重点考察思考问题的广度，数据不平衡时怎么办（如数据量很大时怎么处理，数据量小时怎么处理，如何将文本数据转化为特征向量，用什么方法，对于邮件的标题及正文是否进行相同的处理，权重是否相同，如何

找到关键特征等），数据特征维度远高于数据量时，从样本采样、向量维度降维、分类算法的重新选择三方面改进，尽量多考虑系统性设计的思想

● 设计一个系统：可以写出春联，必须满足他的要求，平仄音节都要对上，我直接 BERT+CRF+GPT 一顿乱写。

● 有一系列人的属性相关数据，要探索他们的健康状况（0，1 问题），可以用 DNN 做吗，和普通的学习方法如 GBDT 比有什么区别么（我从特征相关，特征组合的角度回答的）

● 100 亿数据，100 个节点，如何实现并行聚类？

● 4g 内存，10g 的文件数据进行排序输出到新的文件中。

● 给定 10G 的文件，只有 2G 的内存，如何将文件放到内存中

● 给很多文件排序，内存不够用，但磁盘容量充足怎么办？

● 一个文件有 9998 个数，对应[1-10000]的范围，少了两个数。内存 1k，怎么查找？

● 开放性问题：

● 没有任何用户行为的用户如何做推荐？

● 大量用户行为的用户如何做推荐？

● 思考题：排序曝光的数据直接拿回来做召回训练可以吗？

● 抖音这类 app 怎么做推荐系统？分类问题还是回归问题？当作回归问题的话会不会有什么问题？（提了一下 youtube 模型，它是对观看时长进行建模）

● 给你一个海量图片数据集和同一个人的多张照片，如何快速从中找到这个人的其他照片？

● 海量数据中给定 query 搜索出 N 个答案

● 业务题：给一个 txt 文件，里边每行是一个 img_id 和 img_type，处理该 txt 文件得到另外一个 txt 文件，每行为一个 img_id 和另一个 img_id 和他们对应的 img_type

● 两个很大的表，一个是[用户,搜索词]，一个是[搜索词,词分类]，如何知道每个用户的词分类

没答出来，在面试官的提示下想到了思路，但是没来得及说，把第一个表转置，变成[搜索词,用户]，然后按照搜索词将两个大文件分成许多小文件，map reduce，分布式计算

● 凤巢算法开放题: 广告系统设计，图对广告的点击率是印象很大的，因此如果从知识图谱的角度考虑这个问题

● 图像中水印如何识别?

● 场景题：一个系统，当一个 id 登陆和退出的时候的时间都是记录的，如何统计当天每个时段的人数

● 开放问题：如何设计电影垂直搜索系统（数据与搜索两个角度答）

● 开放问题：如何设计一套搜索引擎(我主要回答了查询理解部分)

● 场景题：给一个文件，里面存储着一个类别 ID 以及该类别的父类别 ID，要求写函数处理文件，并能够根据查询的类别 ID 输出其所有子类别的 ID，个人理解是把文件构造成多叉树，根据输入的节点输出其所有子节点

7.2 产品方面

● 百度账号有男女之分，设有唯一标识 ID（数字形式，0-2 亿之间），数据量有 1 亿条。要求输入数字，输出是男？还是女?

● 场景：如何判断一个短视频中的内容有没有出现在另一个较长的视频中。

● 场景题: 一些人可能在视频网站的弹幕上发一些不好的话，如何找出这些话，如何设计评价标准。

● 场景题：更换路灯灯泡，假设以前的路灯更换是人，如何获得任何数据

● 怎么给用户推荐内容（可以利用浏览历史等等）

● 如何从百度页面抽取出娱乐明星的代表作，讲下你的思路、方案、实现流程?

● 最后是一个场景设计题，如何开发一个时间管理的程序？分析主要的功能，需要解决用户的哪些痛点?

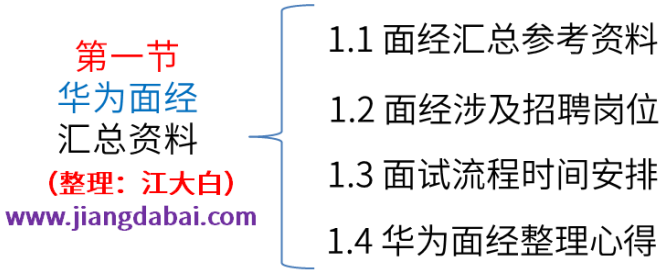
答：我主要从任务管理的角度回答，包括按时间段设置任务，每日完成情况分析，好友监督等。然后强扯了一下推荐，比如把其他人的一些好的时间管理经验推送给用户（针对设定的任务以及用户的匹配程度）

7.3 开放性问题

- 一个东西上线了跟线下实验不一样咋整。
- 高考满分 750，有 100w 个考生的成绩，求第一百名的成绩（要求最优）。
- 如何检测藏头诗？机器学习可以吗？神经网络呢？为什么可以？如果是不是诗句，普通的文本如何检测是否有人为的编码信息？有什么思路？（情景题是考验思路，一定要灵活，不对也没关系，找到问题的几个关键点，先弱化其中几个关键点，再提出思路，一步一步解决）
- 搜苹果的时候要投放苹果手机的广告，怎么弄？
- 苹果可以是苹果手机，也可以是水果，怎么区别？
- 给一个广告素材（里面可能有图片视频文本），如何生成一个标题？
- 给你三个场景，你分别会怎么做，最讨厌哪个，最喜欢哪个：1、这周五布置的任务，下周五交，但是任务量巨大，时间很紧；2、同样这周五布置，下周五交，但是任务方向流程 leader 都没有跟你交代清楚；3、周围的同事比你优秀。
- 开放题：如何推广飞桨

5|华为算法岗武功秘籍

1 华为面经汇总资料



1.1 面经汇总参考资料

① 参考资料：

（1）牛客网：华为面经-172 篇，网页链接

（2）知乎面经：点击进入查看

（3）面试圈：点击进入查看

② 面经框架&答案&目录&心得：

（1）面经框架及参考答案：点击进入查看

（2）大厂目录及整理心得：点击进入查看

1.2 面经涉及招聘岗位

（1） 实习岗位类

【华为云 EI 实习岗】、【计算机视觉实习生】、【华为杭研院 Cloud&AI 昇腾计算产品部算法实习】

（2） 全职岗位类

【机器学习算法工程师】、【终端部门算法工程师】、【开发硬件算法工程师】、【华为上研算法工

程师】、【AI 应用研究中心工程师】、【华为云视频内容分析】、【华为消费者 bg 算法工程师】、【Cloud Bu 人工智能工程师】、【华为南京研究院算法工程师】、【华为成都研究院算法工程师】、【华为 AI 算法工程师】、【华为西安研究院算法工程师】、【华为南京 NLP 算法工程师】、【华为自动驾驶算法工程师】、【华为射频算法工程师】、【华为消费云服务部 AI 工程师】、【华为数据存储与机器视觉产品线智能协作产品部 AI 工程师】、【华为智能车 BU AI 算法工程师】、【图像算法工程师】、【音频算法工程师】、【搜索推荐算法工程师】、【圣无线的通用软件开发工程师】、【成都传送/无线部门通信算法工程师】、【华为数据存储 AI 工程师】、【昇腾计算产品部 AI 工程师】（机器学习方向）】

1.3 面试流程时间安排

华为面试流程-整理：江大白			
	面试类型	面试流程	备注（侧重点）
第一面	技术面	自我介绍+项目/实习经验 +技术问答	主要问项目+基础知识
第二面	技术面	自我介绍+项目/实习经验 +技术问答+算法编程	关注项目广度和理解
第三面	HR面	基础人力问题	/
第四面	综合面	自我介绍+项目经验+公司发展	相当于boss面，问的更全面，从宏观到细节，以及项目落地

PS：以上流程为大白总结归纳所得，以供参考。

其他注意点：

- 有些人在第一面之前，还会有机试和性格测试
- 有些区域是技术面+机试性格测试+三面 BOSS 面
- 有些人是先综合面，再 HR 面

1.4 华为面经面试心得汇总

★ 华为特别重视底层原理，和其他互联网公司不一样。

★ 华为的面试看面试官吧！有的人会被很多技术的，有的只是聊聊人生和项目。

★ 总结一下三场面试，需要准备好编程相关的问题，机器学习相关的问题，自己方向最新的技术。另外，三场面试都着重问了项目，可能我比较菜，没有发过论文。自己对项目的细节一定要十分了解，这样就不用慌了，随便问都能答上来。

★ 每个区域的招聘流程稍微有点差异，不过一般分为基础面试、综合面试：

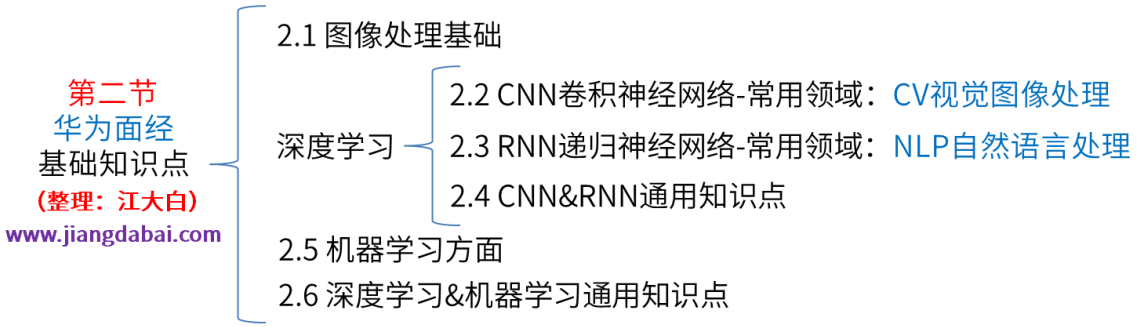
基础面试基本就是聊项目经历或者实习经历，另外有些会从产品的角度出发，出一些发散性思维的题目，不怎么为难你 ，主要问项目经历

综合面试主要谈性格、对华为的认识、为什么想加入华为；主要看重承担压力的能力，表现的性格开朗就 Okay 了。

★ 有的时候，面试很难，有的时候很简单，所以还是看人，但是最好认真准备，以不变应万变。

★ 聊简历上的项目，每次说到某个点会继续深入问一下，但挖的不深。我面的那个面试官是做人脸识别和指纹识别的，最后问我怎么识别是照片还是真人，我以为都只能拍一张照片，就说了一些用深度，或者阴影和光照等解决之类的，但其实是可以拍很多张的，可以根据运动判断，所以跟面试官好好沟通真的很重要！

2 华为面经涉及基础知识点



2.1 图像处理基础

2.1.1 讲解相关原理

- 传统图像处理的 canny 算子

2.2 深度学习：CNN 卷积神经网络方面

2.2.1 讲解相关原理

2.2.1.1 卷积方面

- CNN 中 1*1 卷积核的作用?
- 介绍一个熟悉的 CNN 模型，卷积怎么反向传播?
- CNN 基本组成，什么是感受野， 反向传播原理?
- 膨胀卷积原理
- 空洞卷积相比普通卷积的不同之处，如果特征图很小，这时要用空洞卷积就会加很多 padding，增加很多无用信息，怎么处理这种情况?

2.2.1.2 池化方面

- 池化层的作用? （拓展讲了种类、反向传播，以及 pytorch 特有的自适应池化）

2.2.1.3 网络结构方面

- 简述 MobileNet 的 V1,V2,V3 的区别?
- vgg、resnet、densenet 之间的比较?
- 画一下 MobileNet 网络结构
- resnet 和 denseNet 的网络结构，以及为什么这样设计?
- ResNet 的作用?
- 认识哪些常用网络，是为了解决什么问题所提出的?
- 为什么要用轻量级的网络? shufflenetv2 相比 v1 有什么改进?

2.2.1.4 其他方面

- 简单的介绍一下 CNN，及它的发展和应用?
- 自己写网络模型时，是手动搭，还是复现或调库？自己有没有优化或者自己搭建新模型，描述一下?
- 梯度消失/爆炸产生原因，及解决方法?

2.2.2 公式推导

- 写一下了 batch norm 的公式?
- Softmax 等层的原理（公式）写一下?
- 推导神经网络链式法则

2.2.3 手写算法代码

- 手推卷积过程

2.3 深度学习：RNN 递归神经网络方面

2.3.1 讲解相关原理

- 简单的介绍一下 RNN，及它的发展和应用?
- RNN，LSTM，GRU 的异同?
- 介绍 LSTM 及其变种?
- 解释 LSTM 原理，LSTM 的结构描述一下，超参数说一下?
- LSTM 为了解决长依赖问题，引入了三个门，分别啥意思?
- 能否详细的介绍 LSTM 模型的结构和内部的运行过程?
- 双向 LSTM 比 LSTM 到底好在哪?
- LSTM 为什么可以避免过拟合?
- LSTM 哪个门用到了上一状态?

2.3.2 手绘网络原理

- 画出 LSTM 的结构图，写公式

2.4 深度学习 CNN&RNN 通用的问题

2.4.1 基础知识点

- 不平衡样本怎么处理?
- Transformer 相比于 RNN 你认为有哪些改进?
- 怎么做的数据增广?
- attention 怎么做?

2.5 传统机器学习方面

2.5.1 讲解相关原理

2.5.1.1 数据准备

无

2.5.1.2 特征工程

① 特征降维

- SVD 与 PCA 的关系?

② 特征选择

- 特征选择的方法？(这里建议分 filter, wrapper, embedded 来讲，我只是说了 PCA，LDA，L1)

2.5.1.3 有监督学习-分类和回归方面

① 分类回归树（集成学习）

- 集成学习了解吗？XGBOOST/GBDT 简单介绍，区别?
- bagging 和 boosting，stacking 区别，分别的原理?

A.基于 bagging：随机森林

- 为什么随机森林能降低方差?

B.基于 boosting：Adaboost、GDBT、XGBoost

- 树模型和熵介绍，为什么 xgboost 效果好?
- xgb 和 gbdt 的区别?
- GBDT、RF 有什么异同？各适用于什么样的情况?
- 介绍 xgb,lgb?

② 逻辑回归 LR

- 线性回归解析解的推导（三种方法）

③ SVM（支持向量机）

- 介绍一下 SVM，介绍了核函数的种类、支持向量、超平面、软间隔、Hinge Loss?
- svm 优缺点

④ 朴素贝叶斯（Naive Bayes）

- 贝叶斯模型？（这里我顺着讲了朴素贝叶斯、逻辑回归最大似然推损失函数的过程）
- 解释极大似然估计，最大后验概率估计，解释核函数及其应用?

⑤ 决策树（DT）

- 决策树划分选择、树的复杂度、剪枝?
- 决策树，随机森林原理?

2.5.1.4 无监督学习-聚类方面

- knn 与 k-means 的区别?
- k-means 和 DBSCAN 的对比，k 的选取，提速，聚类方法的评估?
- 聚类算法如何提升性能?
- K-mean 算法的优缺点（简历中有提到 K-means++）
- DBSCAN 的原理?

2.5.2 手推算法及代码

- 手推 SVM
- 写一下贝叶斯公式
- 写一下 KL 散度公式

2.6 深度学习&机器学习面经通用知识点

2.6.1 损失函数方面

- 深度学习有哪些激活函数？为啥会有激活函数？
- 为什么 Relu 的结构小于 0 的输出为 0？这样有什么优点（防止梯度消失，稀疏性以及加快计算，当时没想到）什么缺点，如何改进，改进版 relu 的名字是什么（忘记了叫 LRelu）？
- MSE 和交叉熵的区别 ，写交叉熵？
- 手推交叉熵的求导

2.6.2 激活函数方面

- Sigmoid 与 Softmax 的区别与联系？
- 说一下激活函数，relu 和 sigmoid 区别？

2.6.3 网络优化梯度下降方面

- sgd 和 adam 的优缺点？
- 什么是 ADMM，为什么用 ADMM，子问题为什么不用梯度下降求解？
- 梯度下降为什么可以成功？（我回答的是损失函数是凸函数）

2.6.4 正则化方面

- L1，L2 符合哪种分布？

2.6.5 压缩&剪枝&量化&加速

- 模型压缩的几种方法？（量化、剪枝、低秩分解等 ）实际用过吗？

- 量化的理解，有什么好处 ？
- 加速优化的方法有哪些，剪枝如何操作，最近看过的论文，跟进的方法?

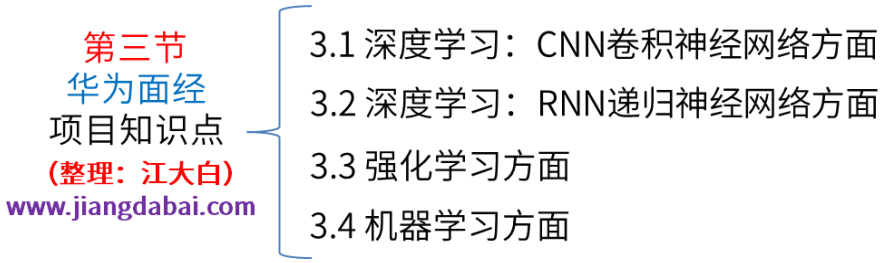
2.6.6 过拟合&欠拟合方面

- 怎么判断过拟合与欠拟合?
- 解决过拟合和欠拟合的办法?
- 机器学习当中可能会有欠拟合过拟合的问题，怎么解决过拟合问题?
- 对于传统的机器学习(rf,lr,svm)来说，一般靠引入正则化项来避免正则化问题，那么应用到深度学习里面，过拟合的解决方式主要有 dropout、early-stopping、数据增强等

2.6.7 其他方面

- 深度学习与传统方法的区别，深度学习为什么效果这么好?
- 根据项目经历解释偏差-方差的权衡?
- 数据不平衡怎么解决?
- 说一下训练模型过程中可能遇到的问题以及解决方法？(这里我详细讲了梯度消失、爆炸，训练曲线不下降，过拟合，欠拟合的产生情况和解决方案)

3 华为面经涉及项目知识点



3.1 深度学习-CNN 卷积神经网络方面

3.1.1 目标检测方面

3.1.1.1 讲解原理

- Rcnn、Fast Rcnn、Faster Rcnn 直接的对比与联系?

3.1.1.2 损失函数

- faster 和 ssd 的损失函数表达式?

3.1.1.3 手写代码

- 写一下 IOU 计算
- 写一下 NMS

3.1.2 目标追踪

- 目标追踪和目标检测区别? kcf?
- 传统目标跟踪方法有什么改进操作?

3.1.3 图像分割

- 对于项目里的语义分割，你还知道哪些语义分割的框架?

3.1.4 图像分类

- 分类器了解哪些，自己写过哪些；最熟悉哪个分类模型?

3.1.5 自动驾驶

- 汽车运动学，动力学
- 传感器硬件(雷达，摄像头等)和相关的算法（欧式聚类）
- fernet 坐标系 S-T 图 lattice planner EM planner
- 自动驾驶分级和一些相关的概念
- 自动驾驶决策规划的相关模块都有问到，主要是从轨迹规划等问题切入问的。

- 问了埃尔米特插值法，实际使用上会出现的什么问题，如何解决?
- 从 autoware 到 Apollo 上的规划模块都有问到，主要是区别还有实际使用上的情况。
- 围绕简历问的 autoware 的优劣势，还有现在主流的决策方法的优劣势（状态机，概率图，强化学习）

3.1.6 音频算法

- 噪声与语音的区分方法，与项目有关?
- 白噪声的特征，如何识别盲源白噪声，如何降噪?
- 时间序列分类算法，用到过那些?
- 时间序列特征
- 时域离散周期的频谱?

3.1.7 通信算法

- 描述奈奎斯特采样定理
- 写出香农公式，说明每一项的含义
- 画出一个你最熟悉的通信系统框图，并简要描述每个部分功能
- 画出 16QAM 调制的星座图，IQ 不平衡时的 16QAM 星座图，带有频偏的 16QAM 星座图
- 说明 FIR 滤波器和 IIR 滤波器的区别
- 列举数字滤波器设计中常用的窗函数
- 说明卷积和相关的区别
- 给出序列[1,0,2 1 3]，计算该序列与[1,0,2,1,3]的线性卷积和循环卷积结果
- 说明什么是“各态历经性”
- 64 阶 FFT 中使用了多少个乘法器?
- 给出有符号定点数 11011011,其中有 5 位为小数，将其转换为十进制数
- 给出信号非整数倍变换采样率的方法，包括频域和时域方法
- 画出 a.格雷映射的 16QAM 星座图，并且写出映射关系 b.画出 IQ 不平衡的 16QAM 星座图 c.

画出加入加性白高斯噪声的 16QAM 星座图 d.画出带有频偏和相偏的 16QAM 星座图

3.2 深度学习-RNN 递归神经网络方面

3.2.1 自然语言处理 NLP

① Transformer

- seq2seq 除了 LSTM 编码，还有 Transformer 的编码结构，了解吗?

② Word2vec

- Word2vec 和 fasttext 区别?
- Word2vec 方法有哪些/区别?
- 介绍一下 Word2vec?

③ 其他

- 根据简历上的 CTR 比赛，问了 fm，ffm，deepfm，dcn，xdewpfm

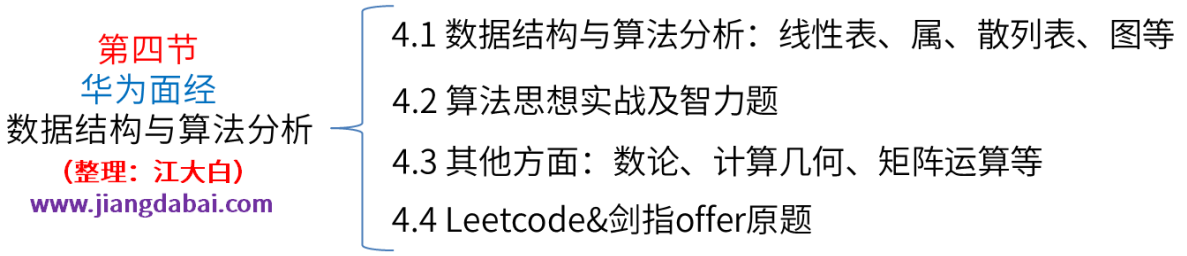
3.3 强化学习

- 孪生为啥起作用?
- 强化学习 Q-learning 和 DQN 写了一下更新公式，然后公式里各个变量的含义啥的，DQN 的伪代码和流程图。
- GAN 怎么训练?
- WCGAN 为什么比 WGAN 好?
- 在普通 WGAN 上做了哪些优化，为什么可以这样优化?
- 问了 WGAN 的优化以及 G 和 D 训练中的平衡?
- 差分隐私怎样引入，证明正确性?

3.4 机器学习方面

无

4 数据结构与算法分析相关知识点



4.1 数据结构与算法分析

4.1.1 线性表

4.1.1.1 数组

- 从未排序的数组里找到第 k 个大的数
- 一个排序数组除了一个元素，其他的元素都是相同的两个，找到这个元素，复杂度 O(logn)（下标奇偶二分）
- 求数组第 K 大的元素，要求 O(n)时间复杂度。
- 找出数组里中每个元素比它小的个数，直接排序+遍历
- 求一个数组中和为 k 的最长连续数组?
- 给你一个数组，给一个 target,从数组中选取任意数量的数字，保证数字之和等于 target，每个数字可以重复取，给出所有的取法?
- 给一个整数组成的 digits 数组，从里面选出一部分数字组合成一个新的数字，要求出能被 3 整除的最大数字?
- 有一个包含正负整数的无需数组，如果数组中存在连续子序列之和为 0，则把这个序列剔除，输出剔除所有符合要求子序列之后的结果?
- 一个长度不超过 10000000 的不重复整数数组，输出其中所有和为 0 的三元组，三元组中元素

可以有单个重复计算？注意尽可能减小时间复杂度。

- 数组访问要注意什么？越界问题
- 给一个数组，求每个元素与后面第一个比他大元素的距离

input :[30,31,25,24,30]

output: [1,-1,2,1,-1]

- 给定一个数组，不改变数组顺序，从前往后依次把所有数取出来，每次取数之和有最大值限制，给定取数的次数，问最大值限制最小是多少？
- 给一个数组，让求和为给定值的最长子数组的长度
- 求数组的最大子序列之和？

4.1.1.2 链表

- 链表反转
- 合并 k 个链表
- Linkedlist 和 Arraylist 的区别？

4.1.1.3 字符串

- 字符串反转
- 实现浮点数转字符串，要注意的点：（1）0.XXX（2）负数
- 进制转换，将输入的数转换成十进制。
- 输入字符串格式有两种：

第一种：base#n，base 表示数字基数(进制)，范围 2-64，超过 10 的数字用 a-z, A-Z, @, _ ，总共 54 个字符表示

第二种： n，没有 base#，0x 开头是十六进制，0 开头是八进制

非法输入，输出 ERROR

- 判断是否为交叉字符串，如：str1 = "abcd", str2 = "1234", str3 = "ab12c3d4", 判断 str3 中是否包含 str1 与 str2 交叉后的字符串？

- 给定字符串（全部是大写字母），给出字符串所有不重复排列数 ？
- 组合无重复最长字符串?
- 字符串的最长公共子串
- 最长公共子序列
- 给定字符串，找出最长的回文子串?
- 求字符串是否是另外字符串的子集?
- 写了个字符串 A=“abcebdfa”，B=“abd”），怎么样剔除 A 中含有的 B 中的字符?
- 给定一行字符串，求出这行字符串中出现频率最高的字符，字符串中含有标点符号，字符不区分大小写。如果出现频率相同时，输出先出现在字符串中的字符?
- 给一个字符串和一个字符，让你找出该字符在字符串中出现的个数，字母的话不区分大小写

4.1.2 树

4.1.2.1 二叉树

- 问斐波那契数列计算的复杂度，分了递归和非递归来讲，但面试官问能不能更快?
- 二叉树，二叉搜索树，二叉平衡树，红黑树
- 二分查找的时间复杂度是多少?
- 二叉排序树的时间复杂度是多少?
- 三叉排序树、四叉排序树的时间复杂度呢?
- 二叉树每个节点的值为 0 或者 1，每个叶子节点所在路径都对应一个二进制数，将其转换为十进制，然后求所有十进制的和?
- 树有几种？分别是什么内涵?
- 寻找二叉树中是否存在值为 k 的路径
- 二叉树最大宽度
- 二叉树删除的时间复杂度，删除后怎么变化，为什么是 logn?

- 完全二叉树的第 7 有 10 个叶子结点，则整个二叉树的结点数可能是多少？
- 给一个二叉树的前序和中序遍历的数组，让你算出它的后续遍历？
- 平衡二叉树的失衡调整
- 树的遍历算法
- 树形运算节点，找出同时最大内存分配？
- 如何把一个搜索二叉树变成排序数组？
- 二分查找

4.1.2.2 堆

- 堆排序算法、冒泡排序的时间复杂度：nlogn、n^2，追问堆排序算法的空间复杂度？
- 大小顶堆如何用数组表示？

4.1.3 排序

- 分别说一下在数据量比较大的情况下最快的查找算法，和数据量比较小的情况下最快的查找算法？
- 数据结构中查找最快的算法是哪个？
- 用 Python 写个归并排序
- 归并排序，但不准调用库函数
- 归并排序的原理
- 如果有非常多数据怎么找出最大的 k 个？
- 找 n 个数里最大的 m 个数
- 快速排序，归并排序，堆排序的思想，复杂度分析？
- 快排，并说下复杂度？
- 在 24h 制下，给定时间字符串数组，求间隔最短时间。

要求不能用自带的排序等功能〔明显想让自己写排序〕

比如〔“12：00”，“12：03”，“15：03”〕

输出就是 3

- 拓扑排序
- 说一下排序算法的稳定性?

4.1.6 搜索

- 图遍历深度优先，广度优先有没有了解，说说这两个方法可以解决什么问题，具体怎么用?

4.2 算法思想实战及智力题

4.2.1 算法思想实战

- 海洋陆地
- 俄罗斯套娃
- 跳台阶
- 找零钱
- 求水仙花数
- 股票收益最大化问题
- 讲一了 BFS 和 DFS
- 有 n 长的钢条，可以任意切割，给定各个长度的价值，求解如何切分可以获得最大价值（动态规划求解）
- 二值矩阵求最大 1 的矩形面积，和面试官说用动态规划做，讲了思路?
- n 级台阶，从某高度往下砸小球，问怎么判断在哪个台阶就会碎，我第一反应动态规划，卡了半天发现并不是，有两个小问题：

a.只有一个小球怎么办，遍历；

b.有两个小球怎么办，一开始说分治，面试官说不是最优解，在提示下答出一个小球用来确定区间，另一个小球用来在区间遍历。问区间取多少，随便答了个 logn，面试官说也可以，最优是根号 n；

● 给定一组温度值序列，返回一个数组，该数组每个点代表当前温度经过多少天以后能够升温，要求用 O(n)时间复杂度?

● 找中间索引，一个数组，[1,3,4,6,5,2]，规定这个中间索引左边的和等于右边的和，如果有多个中间索引，取最左的那个。

先算总和，从左开始遍历数组，每次算左和和右和判断是否相等就可以了。

4.2.2 智力题

● 拆礼物盒， []表示一个盒子，盒子里可以放多个礼物或礼物盒，礼物盒都不为空。

要求拆开所有礼盒，取出小礼盒，仅保留里面的礼物，并摆好礼盒。

礼盒摆放要求:

a.大礼盒在底层，小礼盒在顶层

b.同一级别的礼盒，按照原来从左到右的顺序摆放

c.拆开后，如果大礼盒剩余为空，输出[]

例如:

输入：[[a, b], [c, d], e, f]

输出：[a, b], [c, d][e, f]

输入：[[a, b], [c, d]]

输出：[a, b], [c, d]

● 类似中小学奥林匹克的题，十二个球，其中一个重量与其他不同，用一个天平几次可以找到那个球?

4.3 其他方面

4.3.1 数论

● 给定整数 n，写出其因式分解，因式分解数字从小到大排列?

● 什么是凸函数?

4.3.2 计算几何

- 给定周长，求直角三角形个数
- 单调栈求最大矩形框面积?

4.3.3 概率分析

- 一个无限长的格子，从第一格出发，不停的扔骰子（点数随机 1 到 6），按照骰子数前进几格，问刚好停在第 50 格的概率是多大?

4.3.4 矩阵运算

- 手斯矩阵转秩
- M×N 的矩阵，从左上角走，只能向右或者向下走，要求走过的每个元素的值加起来的和最大，步数不限?
- 4*4 的矩阵，每个位置都有一个 value，求从左上角到右下角的最大累计 value 路径？每次移动只能向右或者向下。
- 给一个二维矩阵 有正有负，求从左下到右上的最大乘积路径，DP BFS
- 给一个二维矩阵 有正有负，求从左下到右上的最大和路径，DP
- 矩阵中有一些数，从左上走到右下，只能往右和往下，最大权值的路径权值是多少?
- 给一个 N,M 的矩阵，由 0，1 组成。其中 1 代表能走，0 不能走，当前小明从左上角(0,0)出发，且初始点必定为 1，他必须用固定的步长 S 走，如果他能走到右下角则输出 1，不能输出 0。 直接用 BFS 很快就可以写出来。

4.3.5 其他

- 开平方，不能用乘，只能用移位做
- 两个列表合并成有序列表
- n 个数的二进制数中 1 的个数？要求一次遍历即可得。
- 打印一个集合的所有子集?

- hash 冲突有哪几种，怎么解决？
- 队列和数组的区别，匹配问题？

4.4 Leetcode&剑指 offer 原题

- Leetcode 3：
- Leetcode 16：
- Leetcode 18：主要是考察双指针
- Leetcode 30：
- Leetcode 72：
- Leetcode 76 题：Minimum Window Substring
- Leetcode 179：
- Leetcode 1144 题
- Leetcode 1162 题：主要是要使用 BFS 算法
- Leetcode 1363 题
- Leetcode 1386 题
- Leetcode 原题：最长乘积子数组
- Leetcode 原题：求一个集合的所有子集，一共有多少个子集？
- Leetcode 原题：O(logN）复杂度找到单次部分旋转后的非减数组最小值？例如：

[1,2,3,4,5]->[4,5,1,2,3] 从后面这个数组中找到最小值 1。

- 剑指 offer 原题：如何判断是否是正确的出栈顺序

5 编程高频问题：Python&C/C++方面

第五节
华为面经
编程高频问题
（整理：江大白）
www.jiangdabai.com

5.1 Python方面：网络框架、基础知识、手写代码相关

5.2 C/C++ 方面：基础知识、手写代码相关

5.1 python 方面

5.1.1 网络框架方面

5.1.1.1 Pytorch 相关

- 问了 Tensorflow&Pytorch 的不同点?

5.1.1.2 Tensorflow 相关

- Tensorflow 如果加载模型?如何加载模型的一部分？具体调用的是哪个接口函数?

5.1.1.3 其他

- 动态图和静态图的区别?
- 把 mxnet，pytorch，tensorflow，caffe 优缺点都讲一下
- caffe 和 tensorflow 的区别?
- Numpy 的数组，pytorch tensor 有什么区别?

5.1.2 基础知识

5.1.2.1 线程相关

- 介绍一下 python 的多线程
- Python 的多线性和进程?
- Python 里面多进程和多线程怎么用，进程之间怎么通信？知道协程吗，你为什么说协程比线程更轻量?

5.1.2.2 内存相关

- python 垃圾回收

5.1.2.3 区别比较

- Python3.5 和 Python2.7 的 map 有何区别?
- 内置数据结构有哪些(tuple, list, dict, set), tuple 与 list 有什么区别?
- 列表和元组的区别?

- is 和 ==和 = 的区别?
- python3 和 python2 的区别?
- for while 循环区别
- 迭代器和集合区别
- Python 静态方法和类方法的区别?
- copy 和 deepcoy

5.1.2.4 讲解原理

- Python 的 map 函数是啥？（list 映射）
- Python 中基本类型有哪些?
- Python 实现单例模式
- Python 设计模式
- Python 面向对象有什么特性?
- 多线程，multiprogress 是真的多线程吗?

5.1.2.5 讲解应用

- Python 动态加载模块怎么做（没用过，后来查了是 importlib）
- 元组的特点，使用场景
- Python 深浅拷贝，一个字典 a，b=a，b=copy.copy(a)，和 b=copy.deepcopy(a)，这时改变 b 的值，a 有什么区别?
- Python 生成器了解吗，在训练数据时，直接加载到一个 list 和用生成器有什么区别?
- Pytorch 里面如果一部分不想参与训练要怎么设置?

5.1.3 手写代码相关

- Python 如何判断将一个句子切分成单词，单词如何判断是否回文?
- Python 正则表达式模块知道吗？写一个
- Python 矩阵乘法怎么写?

5.2 C/C++方面

5.2.1 基础知识

5.2.1.1 区别比较

- 面对对象和面对接口的理解?

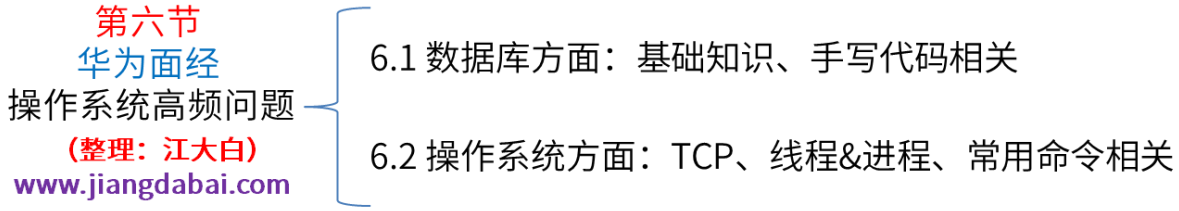
5.2.1.2 讲解原理

- 问 C++的指针知道在操作系统咋实现吗?
- 虚函数知道吗?三种继承方式说一下?
- 聊 c++多态 虚函数 纯虚析构函数 栈解旋
- 多线程，讲了下 CUDA 编程
- 问 redis 接口实现，共享内存接口实现?

5.2.1.3 讲解应用

- C static 有哪些应用场景?

6 操作系统高频问题：数据库&线程&常用命令等



6.1 数据库方面

- Sql: 左连接是什么?
- 数据库会哪些，A 表整体插入到 B 表怎么操作

6.2 操作系统方面

6.2.1 TCP 协议相关

- 知道 tcp/ip 的算法，ip 寻址吗?
- TCP 和 UDP 的区别，设计模式会哪些?

6.2.2 线程和进程相关

6.2.2.1 区别比较

- 进程和线程的区别?

6.2.2.2 讲解原理

- 说一下线程有几种实现方式?
- 进程和线程哪个可以资源共享，另一个为什么不可以?
- 进程和线程，使用线程带来的好处和存在的问题
- 进程通信方式，为什么要通信，线程通信的是什么?

6.2.3 常用命令

- linux 常用命令，查看端口是否被占用
- top 命令的 si 代表什么?

6.2.4 其他问题

- 知道 linux 内核原理、调度吗?
- 是否熟悉封装、继承、多态? STL 数据结构用过那些?
- git 常用命令，如何创建新分支
- ssh 的私钥和公钥的区别、作用

7 技术&产品&开放性问题

7.1 技术方面

- 深度学习网络模型往移动端移植会遇到的挑战以及应用?
- 项目中有遇到数据量大，计算慢的问题吗，有什么瓶颈问题，性能问题吗?
- 如果硬盘容量只有 2G，但是数据量有 10G，应该怎么加载数据?
- deepstream 框架用过吗?
- 现在有 N*n 张照片，请以一个标准对这些图片进行评价。其中，N 表示 N 个拍摄场景，n 表示 n 台终端设备（平板、笔记本、手机等)。(限时 20min 左右)

a.将照片分为 N 组，并对每组照片进行处理，评价，排等级。

b.如何对每组内的照片进行评分?

c.n 设备拍摄照片的 Top 问题（最好最差的情况)，应该如何解决问题，优化照片指标?

（做题的前提是充分理解题目的用意：做题期间，关于图片性能标准、分组评级的内容，我不是很理解，跟面试官沟通了两次，不懂得多问，同时也可以让面试官了解你的沟通能力)

7.2 产品方面

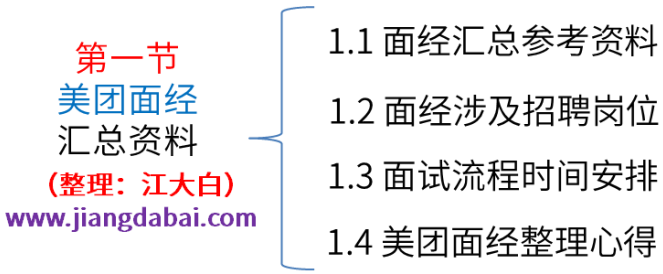
- 问如果分给你的一部分频谱资源被其他的设备所污染，你会采取什么措施?

7.3 开放性问题

- 问以往的经历中有没有遇到什么困难，是怎么解决的?

6|美团算法岗武功秘籍

1 美团面经汇总资料



1.1 面经汇总参考资料

① 参考资料：

（1）牛客网：美团面经-122 篇，网页链接

（2）知乎面经：点击进入查看

（3）面试圈：点击进入查看

② 面经框架&答案&目录&心得：

（1）面经框架及参考答案：点击进入查看

（2）大厂目录及整理心得：点击进入查看

1.2 面经涉及招聘岗位

（1）全职岗位类

【机器学习工程师】、【数据挖掘工程师】、【广告平台的 NLP 工程师】、【AI 算法工程师】、【美团点评到店事业群算法挖掘岗】、【无人驾驶算法工程师】、【美团地图部工程师】、【nlp 和搜索方向工程师】、【智能推荐平台工程师】、【机器学习引擎框架开发】、【机器学习/数据挖掘算法工程师】、【推荐算法工程师】、【美团到店数据挖掘算法工程师】、【美团点评机器学习工程师】、【美团优选 NLP 算法】

1.3 面试流程时间安排

美团面试流程-整理：江大白			
	面试类型	面试流程	备注（侧重点）
第一面	技术面	自我介绍+项目/实习经验 +技术问答	项目细节问的很细， 对应用场景也会发散提问
第二面	技术面	自我介绍+项目/实习经验 +技术问答+算法编程	在一面的基础上， 更多维度提问
第三面	技术Leader面	自我介绍+项目经验+公司发展	和项目相关， 主要关注解决问题的思路
第四面	HR面	基础人力问题	/

PS：以上流程为大白总结归纳所得，以供参考。

1.4 美团面经面试心得汇总

★ 写题比较简单，模型扣的细，细的程度不是理解网上博客写的就可以，而且要看原论文。一面的时候被憋的几乎想拿书包走人了，说的最多的话就是，这个记不清，这个不了解，建议大家深扣细节！

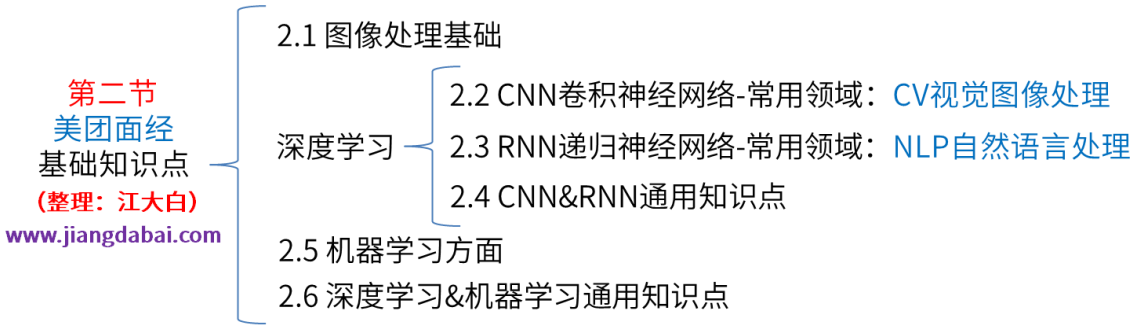
★ 美团的面试官水平真的高，技术面问的我都淌汗了，总监面也很专业。

★ 会经常问算法题目+机器学习相关的，根据项目问

★ 项目问的比较清楚，所有的东西都是由简历的内容进行拓展。一般都会有手写代码题，所以常见的一定得刷一刷。

★ 机器学习的偏多，底层的算法都会涉及到，但是简历上的项目建议要准备好，而且虽然是算法岗，但是也会问很多计算机基础相关的问题。

2 美团面经涉及基础知识点



2.1 图像处理基础

2.1.1 讲解相关原理

- 透视变换和仿射变换讲一下？
- 写出旋转变换矩阵（三维的那种）
- 写出带平移的旋转变换矩阵，写出带缩放的旋转变换矩阵，写出射影变换矩阵

2.2 深度学习：CNN 卷积神经网络方面

2.2.1 讲解相关原理

2.2.1.1 卷积方面

- 是否了解图卷积？
- CNN 1*1 卷积核的作用？
- 卷积神经网络的权重是怎么更新的？
- dropout 的随机因子会对结果的损失有影响吗
- dropout 怎么反向传播？
- 卷积是空间不变性还是时间不变性？
- CNN 网络有哪些层？

- 感受野在 cv 中的作用，大小分别有什么影响 全局感受野和局部感受野的优缺点，哪些论文的方法是从这方面考虑并进行改进的，介绍一下他们的方法？

2.2.1.2 池化方面

- 平均池化和最大持化的反向传播是怎么运作的？
- 池化层如何反向传播？

2.2.1.3 网络结构方面

- Inception V3 和 ResNet50 网络模型用过吗？
- 简单讲一下 Inception 家族(各种优缺点以及改进)
- 讲一下 ResNet 的原理以及它解决了什么问题？怎么解决的？
- Resnet 为什么有效果？为什么能够保证很深的网络具备不错的效果？
- 对感受野的理解？例如 VGG 网络，最后一层卷积网络输出图片对于输入图片的感受野的大小？
- SEnet 的结构？SEnet 如何放到 Resnet 的 backbone 里？
- RoI Pooling 和 ROI Align 区别？

2.2.1.4 其他方面

- 神经网络中的偏置项（b）尺寸应该是什么样的？
- BN 知道嘛？讲一下 BN 的原理，作用？它有四个公式，每一个公式分别是什么，有什么各自的作用？
- BN 为啥可以缓解过拟合，详细讲一下？BN 有哪些需要学习的参数啊，BN 训练和测试是怎么做的？
- 除了 BN，你还知道那些其他的加速收敛的方法(楼主说到了归一化)，面试官说，和 BN 差不多的那些你了解吗？(GN, IN, FN)
- BN 一般用在网络的哪个部分啊？
- BN 底层如何计算，手撕 BN，BN 在训练、测试阶段的计算有什么区别

- 用公式推导小的 batchsize 会对模型训练有什么影响，我回答了 BN 方面的一些影响，面试官说不行，从 BP 角度考虑。
- 如何解决梯度消失问题?
- 梯度消失，梯度爆炸讲一下？怎么解决?
- 分类问题的 loss 为什么选交叉熵，MSE 可以吗？基尼系数的公式为什么这么写?

2.2.2 公式推导

- 手推 BP 算法公式（就只有一层隐含层的那种）
- Softmax 的计算公式是什么？为什么使用指数函数?
- 用公式详解 BP 原理
- 通过公式解释链式法则以及 resnet?

2.3 深度学习：RNN 递归神经网络方面

- RNN 为什么梯度消失？ 答：tanh 激活函数 以及序列过长会导致梯度消失。还有个原因是因为 RNN 是每一步都共享权重的。
- LSTM 跟 RNN 的区别，他和 RNN 相比有什么优势。
- LSTM 的信息传递机制是什么?
- LSTM 原理，与 GRU 区别？使用场景的不同点?
- BiLSTM 相比 LSTM 有哪些 case 上的提升。Attention 是如何加的取得了哪些效果的提升?
- LSTM 结构以及从数学层面谈为啥优于 RNN?

2.4 深度学习：CNN&RNN 通用的问题

2.4.1 基础知识点

- 注意力机制的运行过程是什么样的?
- Local attention 和 global attention 的区别?
- attention 机制的作用以及选用的原因?

- 为什么设计神经网络解决问题，目前网络存在的问题是什么，后续可以怎么优化?
- 介绍 self-attention 计算，为什么用多头注意力?
- Dropout 的解释
- 介绍一下 transformer。有什么可以调整的参数
- 具体讲一下 self attention。
- self attention， attention， 双向 lstm 的区别。
- CNN 和 RNN 的优缺点

2.4.2 模型评价

- 问了 Precision Recall?
- AUC 的作用，AUC 计算方法，AUC 的时间复杂度?

2.5 传统机器学习方面

2.5.1 讲解相关原理

2.5.1.1 数据准备

- 数据探索（拿到数据之后怎么分析的？）

2.5.1.2 特征工程

① 特征降维

- 降维了解吗? PCA 是什么，矩阵的特征值和特征向量的物理意义是什么?
- 为什么用 SVD?
- 说说矩阵分解?

② 特征选择

- 特征选择有哪些方法？（这个也是有很多，从过滤法、包装法、嵌入法来论述，条理依旧很重要）
- 特征筛选都有哪种方式（卡方检验的公式是啥）

- 了解 Embedding 吗？图嵌入的训练集是什么？

2.5.1.3 有监督学习-分类和回归方面

① 分类回归树（集成学习）

- boosting 和 bagging 在不同情况下的选用？
- Xgboost 和随机森林的区别？

A.基于 bagging：随机森林

- 随机森林怎么计算特征重要性？
- 讲下随机森林，随机森林的随机体现在哪里？

B.基于 boosting：Adaboost、GDBT、XGBoost

- GBDT 的原理，以及常用的调参的参数
- GBDT 怎么处理类别特征，例如 ID 特征
- XGBoost 中对 GBDT 有哪些优化？Xgboost 中的行抽样，可以起到哪些作用？
- XGBoost 原理，怎么防过拟合，gbdt 推导？
- XGBoost 和 lightgbm 的区别和适用场景？
- gbdt 推导和适用场景？
- LightGBM 和 XGBoost 是怎么处理缺失值的？
- RF 和 GBDT 的区别？
- Xgboost 跟 RF 的区别?你觉得在什么场景下，哪个性能会更出色？
- XGboost 里面预排序是怎么做的？
- 比较 GBDT、XGBOOST 区别
- xgboost 怎么处理 null 值？null 值分裂的时候在正常值的哪一边？
- 介绍 xgb 和 lgb，改进了什么？
- XGBoost 如果损失函数没有二阶导，该怎么办？

② 线性回归

- 线性回归跟逻辑回归的区别?

③ 逻辑回归 LR

- LR 的推导，损失函数?
- 讲一下 LR，LR 怎么优化的(说到了线性回归，然后面试官又追问了线性回归怎么优化的?说了梯度下降，但面试官好像不满意，他是想让我说极大似然估计?
- 为什么说 LR 是广义线性模型?
- LR 的损失函数写一下，极大似然和最大后验的区别?
- 问了 LR 和 DecisionTree 的区别?
- GBDT+LR 的设计理念是啥，为啥要这样设计，为什么不用 RL，而是 GBDT?

④ SVM（支持向量机）

- 解释 SVM，问到了 SVM 原理，对偶过程等。
- SVM 介绍一下？为什么可以使用对偶来求解原始问题?
- 核函数了解吗？核函数解决什么问题？我说核函数解决了当前特征空间中样本不可分的问题，他说只要是样本不可分就用核函数吗?
- SVM 的松弛因子作用?
- SVM 的理论依据，如何推导?
- One-class svm 和传统 svm 的区别，你看过 one-class svm 的底层代码吗?
- 介绍一下 SVM，SVM 如何扩展维度?

⑤ 朴素贝叶斯（Naive Bayes）

- 解释贝叶斯深度网络，并说明其优缺点？
- 贝叶斯分类的前提假设?

⑥ 决策树（DT）

- 具体说明一下决策树如何划分，写出相应的公式?
- 决策树将一个特征全部乘以 2 会有什么影响?

- 决策树的分裂方式？（id3,gini,gdbt,xgboost）
- 决策树分类划分的选择有哪些?
- 如果你是算法的设计者，你会怎么设计决策树回归?
- 关于决策树的了解?回答了从 ID3，C4.5 到 CART 各自的优缺点
- C4.5 和 CART 如何处理连续变量以及样本量问题?

2.5.1.4 无监督学习-聚类方面

- 对机器学习的了解，项目中视觉词袋有用到 K-means，问 K-means 初始类中心的选择?
- GMM 介绍一下?

2.5.2 手推算法及代码

2.5.2.1 手推公式

- SVM 的推导，挨个步骤说清楚，对核函数的理解
- 推导 LR
- 手推了 GBDT 公式，问的很细，得对公式非常熟悉

2.5.2.2 手写代码

- 手写实现 kmeans 的伪代码?

2.6 深度学习&机器学习面经通用知识点

2.6.1 损失函数方面

- 写一下交叉熵损失函数?

2.6.2 激活函数方面

- 常见激活函数的异同？说一下，楼主说了 sigmoid，tanh，ReLU, leaky ReLU, PReLU, ELU，random ReLU 等。
- Softmax 和 Sigmoid 区别?
- Sigmoid 函数讲一下，ReLU 讲一下？它们俩的区别以及 ReLU 的优点

- ReLU 有哪些缺点啊，怎么解决的？其他的解决方法你知道吗？
- 常见的激活函数有哪些？各自有什么特点？分布应用于场景？leaky relu 公式？

2.6.3 网络优化梯度下降方面

- 梯度下降，何时收敛何时震荡？

2.6.4 正则化方面

- 正则化有什么用，为什么有用，L1 正则为什么能使参数稀疏，为什么能防止过拟合？
- L1、L2 如何选择？区别？
- 为什么 L1 更容易产生稀疏解，从数学公式回答？
- 分别介绍 L1 和 L2 正则化的方式和优缺点

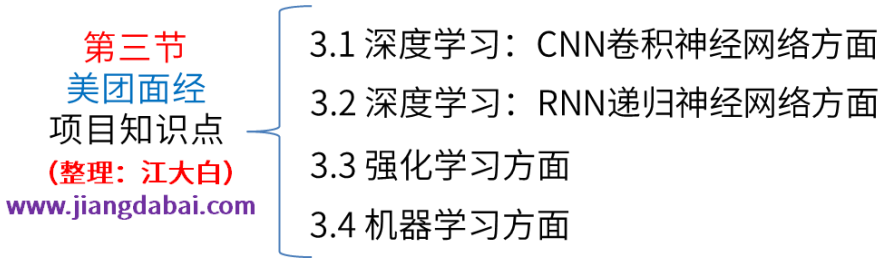
2.6.5 过拟合&欠拟合方面

- 过拟合、欠拟合讲一下，怎么解决？
- 如何防止过拟合，你都采用了哪些方法，还有哪些你没有用到的方法？
- 过拟合和欠拟合哪种比较好处理？答：过拟合，因为欠拟合要么增加数据非线性来提高模型性能，要么增加数据集，做增广（图像方面）。过拟合可以加 dropout，BN，L1，L2 正则。如果实在不行，可能要重新查看数据集，看数据集是否有问题
- 神经网络和树模型的过拟合的一些参数以及解决方法？

2.6.6 其他方面

- 如果样本不平衡，怎么处理？

3 美团面经涉及项目知识点



3.1 深度学习：CNN 卷积神经网络方面

3.1.1 目标检测方面

3.1.1.1 讲解原理

● 讲述一下 SSD 和 YOLO（我把 DSSD、DSOD、FSSD、RFBNet 等全讲了一遍，外带了一些语义分割的网络，简历上写了，全讲了）

● 简述 Cascade R-CNN 的提出为了解决什么问题?

3.1.2 目标追踪

● Deep sort 跟踪算法使用过吗?

3.1.3 图像分割

● 说一下你知道的经典图像分割网络(不限于医学图像)，我说了 FCN，SegNet，Deeplab 系列，顺便也说到了 U-Net。

3.2 深度学习：RNN 递归神经网络方面

3.2.1 自然语言处理 NLP

3.2.1.1 讲解原理

① Bert

● 简单问了 Bert 模型和 Xlnet 模型?

- Bert 里 Transformer 的多头 attention 是怎么做的?
- Bert 和普通的 Word2Vec 模型相比优势在哪里？他为什么会有这样的优势?
- Bert 为什么只用 Transformer 的 Encoder 而不用 Decoder ?
- Bertde 的原理讲一下?

② Transformer

- 问 transformer 的 encoder 输出的 k，v 是不是相同的，如果相同，为什么需要两个?

③ Word2vec

- Word2Vec 讲一下，想用 Word2Vec 构造的特征表达什么信息（项目）
- Word2vec 训练过程的最后一步有什么办法可以优化 softmax 的计算？指数函数的计算会用查表来近似代替。
- Word2vect 与 glove 区别?

④ 其他

- Deepwalk 是手写还是工具包，有没有用 numpy?
- Seq2seq 中 scheduled sampling 如何做的，RL 部分训练过程中数据集如何构造?
- 问到了 RL+Seq2seq 的一些技术，比如 Seq2seq 怎样和 RL 结合，这里的 action 与 state 都是什么，如何设计 reward 等，为什么选取这样的 reward,具体训练流程是怎样的;
- 怎么样做实体抽取，怎样进行 aspect-level 情感分析，你们模型中增强学习的 reward 如何设计的？为什么这样设计?
- One-Hot 编码的优势?

3.2.1.2 损失函数类

- CTC 介绍一下?

3.3 强化学习

- GAN 里面经常会用到 KL 散度，来写一下 KL 散度的公式?

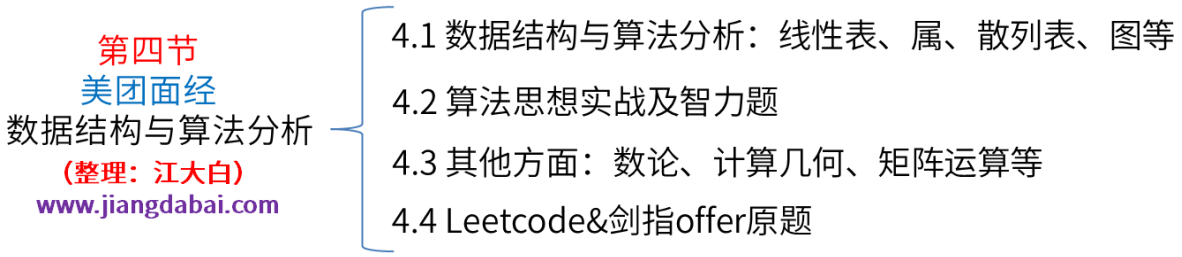
- 为什么要用强化学习，这个问题还能怎么解决，强化学习好处是什么？

3.4 机器学习方面

3.4.1 推荐系统

- 使用深度模型的话为啥使用 deepfm 而不使用 wide&deep 之类的？
- DNN 与 Wide&Deep 与 DeepFM 之间的区别？
- 传统的机器学习和深度学习在推荐上的异同点，深度学习的优势在哪里?。
- 说说推荐系统算法大概可以分为哪些种类：（1）基于内容；（2）基于协同过滤：基于内存（UBIB）；基于模型（MF）
- 推荐系统了解多少，冷启动问题解决什么方案？
- 协同过滤 itemcf 怎么改进？
- 可不可以将深度学习关于图像方面的东西加入到推荐系统中去？
- 文本内容推荐中有哪些内容可以应用到商品团购推荐当中去？
- 因为项目用了 deepfm，所以还发散问了下 fm 的原理，以及 fm 怎么降低复杂度的？
- 对 fm，deepfm，wide&deep，DIN 的看法和了解？
- EE 有哪些别的方法，汤普森采样和 UCB 分别适用什么场景？
- 你还知道哪些 CTR 预估模型，你都用过哪些？
- FM，LR 的区别？FM 的计算复杂度？
- 介绍 Wide&Deep 和 deepFM，有啥区别，有啥优势？

4 数据结构与算法分析相关知识点



4.1 数据结构与算法分析

4.1.1 线性表

4.1.1.1 数组

- 一个数组，找出第 k 大的数、这些方法的时间复杂度是多少？
- 给一个数组，用最小空间复杂度算出数目大于 1 的值
- 二维有序数组 找 target
- 先递增后递减的数组查找最大值（二分法）
- 从给定数组中找到三个和为定值 k 的数？
- 一个升序数组{2,6,7,8,8,9，100}找出其中重复的数字，返回重复数字的起始位置和重复次数？
- 一维数组[1,n],n 可以认为正无穷，然后输入无数个区间[a,b],保证 b>a，求所有区间长度？
- 最长的可整合子数组的长度
- 旋转数组找目标值？
- 有序旋转数组查找？
- 求一个类似旋转数组的拐点位置？由于开始对题目描述理解有点问题，写了一个 o(n)的过程，被问到是否有更快的解决方法，在启发下写了类似下面的二分查找过程。
- 翻转数组，找一个值
- 两个有序数组，求中位数
- 给一个数组，里面只有一个数只出现一次，其余均出现两次，找出那个数（会挖坑，就是给你错误的输入，比如输入的数组大小是偶数）
- 如果只有两个数只出现一次，其余均出现两次，找出这两个数？
- 两个有序数组，找到它们第 K 大的数？
- 两个有序数组合并
- 打印 N 个数组整体最大的 Top K：有 N 个长度不一的数组，所有的数组都是有序的，请

从大到小打印这 N 个数组整体最大的前 K 个数?

- 顺时针打印二维数组：关键考点是边界条件，奇数偶数两种情况如何简化代码，极限情况(例如 1*1 的矩阵)要确保能打印
- 数组的最大连续子数组和
- 一个数组，分为两个子数组，使得它们的和相等
- 对一个数组随机排序，每个位置都随机，o(n)?
- 给定一个数组[3,2,5,0,2,0,0,0,0,7]，实现一个算法，使 0 都放置在数组末尾，其他元素保持顺序不变？要求时间复杂度为 O(n),空间复杂度为常数。
- 给一个数组，将其中的非 0 元素移动到做边，0 移动到右边，不改变非 0 元素的相对顺序?
- 删除一个排序数组的重复项
- 处理 IP 地址，例如 192.18.23.57，将四个数存入数组中（数组范围在 0 到 255 之间，为何是这个范围，因为只有 8 位，可能有字母或无效的情况出现），用的队列，面试官说想到这个很不错，不过可以调函数实现（可能在考察 C++的一些字符串处理库）
- 给定一个目标值 M 的数组，返回数组是否存在和为 M 子集

4.1.1.2 链表

- 判断一个链表是否有环路，并找到环路入口?
- 判断一个链表是否有环？快慢指针可以解决
- 找相交链表的交点
- 合并两个排好序的链表
- 链表问了很多：找中点，是否有环，环的入口，是否有交点，交点在哪里，N 个链表是否有交点，复杂度分析?
- 链表倒数第 k 个值

4.1.1.3 字符串

- 翻转中间由各种符号隔开的字符串?
- 字符串转换为数字，比如'123'变成 123.【要考虑特殊情况 '0123 ‘>123，'12 3'>123,

‘123’>123】

- 输入一个字符串，判断其是否是“（”和“）”的一 一配对
- 两个字符串拼接起来，判断有没有回文?
- 输出最大无重复子串
- 给定一个字符串列表（长度为 n），给定一个滑动窗长度 L，求滑动窗里边最多有多少个不同字符？（要求时间复杂度 O(n)，空间复杂度 O(1)
- 求两个字符串的最长公共子串
- 括号字符串是否能完全匹配
- 求一个字符串的顺序子串的个数
- 一个字符串，找到第一个只出现一次的字符，n 空间 n 时间，只能扫一次
- 字符串把多个连续空格合并成一个，输入是 char*，要求原地空间

4.1.2 树

- 问了满二叉树和完全二叉树，大概画了一下
- 从右边看被遮挡的二叉树，求露出的 node?
- 判断两个二叉树是否相同，递归可用
- 树的层序遍历，每一层单独输出，且从底向上输出
- 翻转一棵二叉树
- 输出二叉树最长的路径?
- 层序遍历二叉树
- 中序遍历二叉树
- 手撕反转二叉排序树递归和非递归 问了时间复杂度?
- 二叉搜索树如何序列化和反序列化，如果是普通的二叉树、多叉树呢?
- 给二叉搜索树插入节点，然后逐个返回父节点的值
- 文件查找树（递归、DFS）

4.1.3 排序

- 冒泡排序的复杂度说一下?
- 写一下归并排序
- 100 万数据中找前 10 大的数，随时找到 100 万数组中的某个数，O（1）复杂度
- topK，用了快排和堆两种思路，面试官让手推快排时间复杂度
- 海量数据，内存不够的情况下如何以最快速度进行排序?
- 排序算法，快排复杂度，最坏情况，怎么优化?
- 手写最小堆代码
- 给一个数组，找出最小的 K 个数，用了堆排序 O(nlogk)，有没有更快的方法?
- 一个很大的数组找出第 k 个数等（基于桶排序回答的）

4.1.4 搜索

- 平面一些点，距离近的算一类，输出可以有几类（我用的 dfs，复杂度高了点）希望复杂度降下来。
- 二叉树 DFS，BFS，如果层次遍历要从右到左

4.2 算法思想实战及智力题

4.2.1 算法思想实战

- 四边形四个顶点 S、A、B、C，每个顶点都与其它三个点相连，所以把它们平铺成平面，就是一个 2*2 的网格，两点之间都有连线（画出来就是一个框框里一个叉×）。把 S 坐标定为（0，0），就是求：走 k 步回到 S 点的路线有多少个?
- 用加减乘除算根号 2
- 写了个 DFS：一个 nxn 的矩阵，一个阈值，找到大于这个阈值的点，若它的上下左右都满足条件，它们则聚为一类，输出每个类别所含点的坐标
- 8 升、5 升、3 升水桶各一个,如何分成两个 4 升,写出允许状态集合和允许决策集

4.2.2 智力题

- 抽蓝球红球，蓝结束红放回继续，平均结束游戏抽取次数?
- 五个外卖，先取后送，一共十个点，有多少种排列组合情况?
- 如何来判断一个骰子是否是均匀的?
- 找出 9*9 的数独的一个解?
- 两瓶墨水，一红一黑，用小勺从红墨水瓶里舀一勺放入黑瓶，搅拌均匀，然后从黑瓶里舀一勺放入红瓶，这时红瓶里的红墨水多还是黑瓶里的黑墨水多？如果不搅匀呢?
- 25 匹马，5 个赛道，最多几次可以知道前三名

4.3 其他方面

4.3.1 数论

- 假设 A 和 B 服从(0,1)的均匀分布，并且 A 和 B 相互独立，求 max(A, B)的数学期望?
- 给定一个数轴，在数轴上放一些点，给定一个长度的标尺，在标尺内最多能有多少个点(当时题意理解不对，说用滑窗，考官说思路对)
- 10 万个数字找最大的 k 个
- 输入两个整数 n 和 m，从数列 1，2，3.....n 中随意取几个数,使其和等于 m ？要求将其中所有的可能组合列出来。
- 给定整数 n，求离根号 n 最近的整数？用了二分查找实现。
- 给定一段升序整数序列，求出积最小且和为定值的两个数，设置头尾两个指针即可
- 一个整数数组，找最长的先增后降的序列

4.3.2 概率分析

- 比如 10 只（5 双）鞋子抽 4 只，不能配对的概率?
- 支付宝集福活动，假设五张福的概率一样，那么平均多少张福可以集齐五福?
- 给定 n 个小球，有放回地采样。当 n 趋向于无穷的时候，某小球不被取到的概率是多少?

- 问个数学题，一副牌 54 张，去掉大小王，还有 52 张，从中抽取 2 张，问是一红一黑的概率多大?
- 一个商店，1 个小时卖出去 5 个包子，问下一个小时卖出 6 个的概率?
- 一个家庭有两个孩子，已知有一个是女孩子，问全是女孩子的概率是多少?
- 一根木棍随机砍两道，构成三角形的概率?
- 某村庄的的习惯是一直生到男生为止，求村庄的男女比例?
- 一个小时平均闯红灯 5 次，问一个小时闯红灯 6 次的概率？（泊松分布）
- 一家人两个孩子，已知一个是女儿，问两个都是女儿的概率？（条件概率）
- 一个人打靶十次命中 7 次，命中率是 70%，这个概率是怎么估算出来的?
- N 枚真硬币是一面图案一面字，M 枚假硬币是两面图案，选了一枚抛 K 次都是图案，问是真硬币的概率——贝叶斯

4.3.3 并查集

- 两个集合求交集

4.3.4 其他

- 实现 Math.sqrt（）
- 用最快的方法计算 2 的次幂?2 的次幂溢出怎么办?
- 查找两个数的中位数（一步步深入，先从空间复杂度 O（n+m）到时间负责度 O（n+m）到 O（log(n+m)）？
- 通过概率未知的非均匀硬币生成 1 到 N 随机数?
- 求因子个数为 n 的最小数字(写了暴力的方案)
- 给一个数列，找到最长上升子序列并输出？（dp+回溯）
- [4,5,6， 1,2,3] 找到两段有序数列的分割点？先说了个顺序查找 O(N) 然后说了个二分查找 O（logN）并实现?
- LRU cache 设计

- DAG 遍历
- 尼系数是什么？为什么用基尼系数不用熵？
- 判断点是否在矩形框内

4.4 Leetcode&剑指 offer 原题

- Leetcode 172
- Leetcode 300
- Leetcode 543
- Leetcode 695：岛屿的最大面积
- Leetcode 1482
- Leetcode 简单题：求 n！后面有几个 0，要求不把 n！求出来？
- Leetcode 简单题：给一个数组[1,2,3,0,0,1,0,2]，把 0 移到数组的后面，非 0 数字保持相对顺序。要求用双指针。
- 剑指 offer 原题：不用加号实现加法

5 编程高频问题：Python&C/C++方面

第五节
美团面经
编程高频问题
（整理：江大白）
www.jiangdabai.com

5.1 Python方面：网络框架、基础知识、手写代码相关

5.2 C/C++ 方面：基础知识、手写代码相关

5.1 python 方面

5.1.1 网络框架方面

5.1.1.1 Pytorch 相关

- PyTorch 里增加张量维度和减少张量维度的函数？

5.1.1.2 Tensorflow 相关

- TensorFlow 参数服务器实现、为什么选 MXNet、参数服务器分布式训练过程、MXNet 中 KVStore 的实现细节？

5.1.1.3 其他

- 平时主要用什么深度学习框架？
- Mxnet 跟 TF 介绍一下

5.1.2 基础知识

5.1.2.1 线程相关

- Python 里面的多线程，多进程说一下？
- 线程之间怎么通信，进程之间怎么通信，python 的多线程有用吗？

5.1.2.2 内存相关

- 知道 python 的垃圾回收机制吗？

5.1.2.3 区别比较

- 数组和链表的区别？
- Python 里面的深拷贝和浅拷贝说一下？
- 创建对象时 new 和 init 有什么区别？
- python 的单下划线和双下划线有什么区别？

5.1.2.4 讲解原理

- Python 的 try except finally？

5.1.3 手写代码相关

- 有一个 n 个数字的序列，现在想把这个序列分成 k 段连续段，想知道分出来的 k 个连续段的段内数字和的最小值最大可以是多少？
- Python 里面的 lambda 表达式写一下，随便写一个？

5.2 C/C++方面

5.2.1 基础知识

5.2.1.1 区别比较

- C++ STL 容器以及命令有哪些，区别是什么?

5.2.1.2 讲解原理

- C++中的 export 讲一下；虚函数讲一下；构造函数能是私有函数吗？C++是面向对象的吗?
- C++学过吧，map 的底层实现是啥?
- 了解 Redis 吗?
- shared ptr 的特点是什么，可以引用传参吗?

5.2.2 手写代码相关

- 编写 C 语言 atoi 函数

6 操作系统高频问题：数据库&线程&常用命令等

第六节
美团面经
操作系统高频问题
（整理：江大白）
www.jiangdabai.com

6.1 数据库方面：基础知识、手写代码相关

6.2 操作系统方面：TCP、线程&进程、常用命令相关

6.1 数据库方面

6.1.1 基础问题

6.1.1.1 区别比较

- count(1),count(*),count(列名) 这三个有什么区别?

6.1.1.2 讲解原理

- 数据库事务，隔离级别，引擎，为什么用 B+树

- 数据库隔离级别、数据库最左匹配
- SQL 熟悉吗？讲一下 SQL 引擎

6.1.2 手写代码

- SQL 了解吗？写个 SQL，一个表字段有学生 id,学科、成绩，选出所有学科成绩都大于 60 分的学生学号。

6.2 操作系统方面

6.2.1 线程和进程相关

- 进程的状态和转换，进程调度算法，进程间通信，消息队列有什么好处？
- 讲一下多线程和多进程的原理和区别？

6.2.2 常用命令

- Linux 查看进程命令

7 技术&产品&开放性问题

7.1 技术方面

- 树模型跟深度学习的区别？垃圾邮件分类哪个更适合？
- 如何判断一个字符串是手机号？
- 场景题，有很多正例，没有负例，然后有很多未标注数据，尽可能的从这些数据中寻找负例（我是从聚类、最近邻以及异常点检测的思路去想。one-class-svm 应该也是可以的）。场景题回答的好是很关键的，其余的都可以背下来，但是只有这部分才会考你对算法的理解和对问题的处理方式。
- 场景题：如何给商家选择头图，以及 topk 的实现？
- 场景题：怎么做美团 app 的猜你喜欢，只能用 LR 模型，（特征，那些特征，怎么获取，怎么处理）特征怎么离散化、怎么设计整个逻辑，在线怎么获取用户的特征？

- 有 100 万条诈骗电话黑名单，现在有个电话来了，快速判断这个电话是否在黑名单里，要求查询 1000 条和 100 万条所消耗的时间一样?
- 场景题：一个 query，一些结果商品，怎么做点击率模型，怎么处理商家恶意点击?
- 场景题，问电商推荐可以用的主要特征有哪些?
- 如何根据美团的商品评论，生成商品的描述。传统抽取方法，语料大后上深度模型。采用类似于 TF-IDF 的思想避免抽取的描述太大众化没有特点?
- 数据样本不平衡问题有哪些解决办法?
- 假设两个分布 A 和 B，我们一般怎么衡量两个分布之间的距离，一般用什么距离?
- 场景题：找 top10 个常出现的 ip?
- 场景题，如果输入某关键词进行搜索 对于返回的结果可以从哪些方面进行设计?
- 开放问题：如何去做门店的推荐匹配

7.2 产品方面

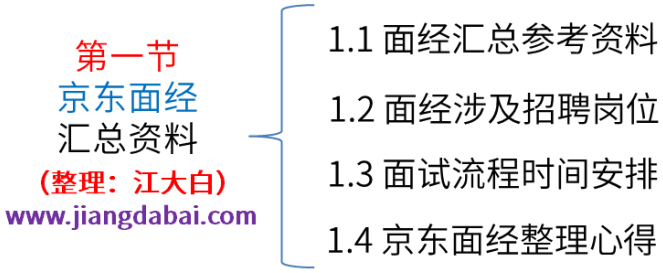
- 场景题：如何判断异常账号（从注册的时候看）
- 如何找到从交易记录中找到最近 10 天最大一笔交易的时间?
- 场景题：如何判断刷单?
- 场景题：判断是不是垃圾信息
- 如何对图片拍摄角度进行纠正?

7.3 开放性问题

- 你会如何去统计一下北京所有大学生在食堂吃饭的比例以及消费情况?

7|京东算法岗武功秘籍

1 京东面经汇总资料



1.1 面经汇总参考资料

① 参考资料：

（1）牛客网：京东面经-105 篇，网页链接

（2）知乎面经：点击进入查看

（3）面试圈：点击进入查看

② 面经框架&答案&目录&心得：

（1）面经框架及参考答案：点击进入查看

（2）大厂目录及整理心得：点击进入查看

1.2 面经涉及招聘岗位

（1） 实习岗位类

【算法工程师实习】、【京东搜索部门算法实习】

（2） 全职岗位类

【京东保险算法岗】、【数据分析工程师】、【搜索与推荐平台算法工程师】、【京东数科算法工程师】、【京东云下的应用研发部算法推荐工程师】、【语音识别算法工程师】、【机器学习算法工程师】、【推荐系统算法岗】、【图像算法工程师】、【京东零售部算法工程师】、【达达京东到家算法工程师】、【京东广告算法工程师】、【京东 NLP 算法工程师】、【广告质量部算法工程师】、【京东

物流算法工程师】

1.3 面试流程时间安排

京东面试流程-整理：江大白			
	面试类型	面试流程	备注（侧重点）
第一面	技术面	自我介绍+项目/实习经验 +技术问答	主要问项目和比赛经历
第二面	技术Leader面	自我介绍+项目经验+公司发展	偏实际和业务场景的问题 以及合作能力
第三面	HR面	基础人力问题	/

PS：以上流程为大白总结归纳所得，以供参考。

其他注意点：

● 有的 HR 问的问题常规，项目中最有成就感的一次 最有挫折感的一次?为什么想要加入 jd?

● 有的 HR 问的很独特，比如：

（1）你认为成为好朋友是契机重要还是相处过程重要

（2）和人相处过程中是否有遇到突破你底线的事情

（3）关注时事吗？说一个时事，以及你的感受

（4）你觉得自己是哪种动物?

1.4 京东面经面试心得汇总

★ 问得比较杂，比较综合，比如操作系统、计算机网络、linux 系统、数据库、机器学习、大叔局，编程。

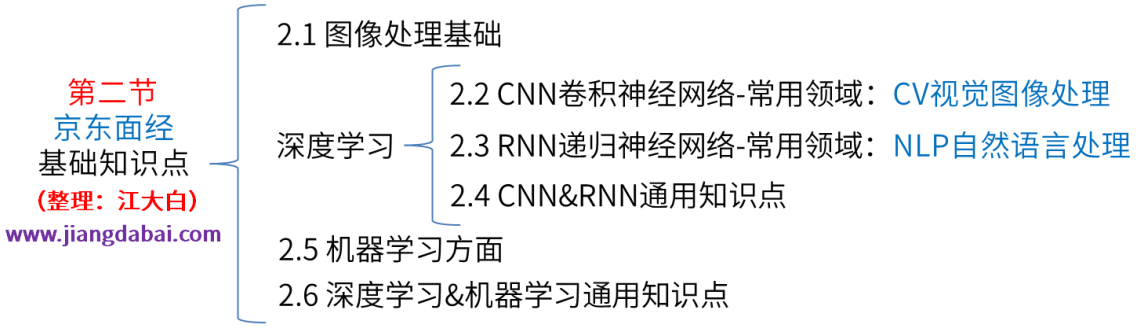
★ 京东的面试专注于考基础知识，基本不涉及特别深入的理解，我的面试时间相对来说比较长，一般都是 20 分钟左右的面试流程，面试的是最后是否会录取你的部门，京东面试的流程比

较快，但是发 offer 可能会比较慢，楼主在二面当天晚上就加到了 leader 的微信，期间一直询问我有没有收到 offer，但是最终时隔 20 多天才收到 offer，可能是校招组和内部沟通还是有时间的延迟吧，不过总体来说京东的面试体验挺好的，守时而且面试难度一般，面试官态度很好。

★ 各个方面都会问一些，针对会的问题会延伸问

★ 对于整个行业的动态，了解的比较多

2 京东面经涉及基础知识点



2.1 图像处理基础

2.1.1 讲解相关原理

- 膨胀腐蚀的原理讲一下？
- 传统去噪算法了解哪些，BM3D，NLM，介绍下？
- 聊了傅立叶变换，小波变换，离散余弦变换？
- 图像的有椒盐噪声用什么滤波器？
- 中值滤波与高斯滤波的原理与运用场合？
- canny 算子和 sobel 算子的原理与运用场合？
- 霍夫变换检测圆的原理？

2.2 深度学习：CNN 卷积神经网络方面

2.2.1 讲解相关原理

2.2.1.1 卷积方面

- 感受野受哪几个参数的影响，给了一个距离例子，计算感受野的大小 ？
- 上采样方式 subpixel，反卷积，resize
- 卷积核大小为什么是奇数?
- CNN 为什么参数共享?
- 简要介绍一下 dropout（训练测试时的步骤，为什么可以防止过拟合）
- Dropout 前向和反向的处理?
- dropout 原理，在测试时需要怎么补偿?
- Dropout 什么原理?
- CNN 为什么比 DNN 好呢?
- CNN 的权重共享平移不变的作用和意义怎么体现的?

2.2.1.2 网络结构方面

- 把 CNN 的发展历史从 2010 年开始按时间轴顺序说一下，说 Dense Net, Xception，胶囊网络这些，然后问为什么 pooling 层不好，哪里不好，要用胶囊网络?
- 画一画 ResNet 的一个 BottleNeck?
- Resnet 说一下 shortcut,两个 mapping 、为啥可以无损传播梯度，为啥可以缓解网络退化
- Inception 网络多层卷积之后是 concat 还是逐像素相加?
- Xception 网络含义?
- ResNet、DenseNet 含义，处理方式，有什么好处，具体 concat 还是逐像素相加?
- 了解哪些模型，讲下它们的原理（VGG，Inception V1-V4，Resnet）
- Vgg 网络名字的由来?

2.2.1.3 其他方面

- 神经网络如何加速?
- CNN 和传统的全连接神经网络有什么区别?
- BN 怎么实现的? inference 时候具体怎么做的?
- BN 的参数，原理说一下?怎么做的标准化，作用是什么？为什么减少过拟合?
- Batch normalization 原理，先归一化然后恢复有何意义?
- BN 和 LN，问的很细，包括二者区别，为什么 BN 不在 RNN 中使用?
- 描述下前向传播、后向传播?
- 样本分布不平衡时，模型效果为什么不好？说明理由?
- DNN 和 CNN 区别?
- 梯度消失和梯度爆炸的原因，怎么解决?
- 解决梯度爆炸的方式（算法层面）?
- 梯度爆炸梯度消失（要求举具体的例子做为说明）
- 你的参数是怎么初始化的。全部为 0？随机初始化？高斯分布中随机取点?

2.2.2 数学计算

- 卷积参数量计算，尺寸计算?

2.2.3 公式推导

- 神经网络分类的 softmax 数学公式，如何计算

2.3 深度学习：RNN 递归神经网络方面

2.3.1 讲解相关原理

- RNN 为什么会出现梯度消失的问题?
- 介绍一下 LSTM? GRU 跟 LSTM 有什么区别?
- LSTM 怎么解决 RNN 的问题?

- GRU 改进了，为什么不用？RNN 上面最近两年有什么新网络改进？
- LSTM 里面怎么处理输入变长的序列？追问那一个 batch 里面长度不一样怎么办，训练会出现什么问题？继续追问选择一批训练的数据满足哪些约束，随机选么?说一下对 LSTM 里面的 Attention 机制的了解？
- RNN，LSTM 原理？区别？为什么 lstm 门用 tanh？
- Transformer 和 lstm 的优缺点？
- LSTM 有几个门？各个门的作用是什么？公式是什么？LSTM 解决了什么问题？

2.3.2 手绘网络原理

- 写一下 LSTM 的结构和前向的传播公式

2.4 深度学习：CNN&RNN 通用的问题

2.4.1 基础知识点

- 如果数据量很大，内存不够怎么办 ？
- transformer：位置编码，为什么用位置编码，self-attention
- self-attention 的作用和功能？
- 加不加 self-attention 在计算效率上有什么不同？

2.4.2 模型评价

- 知道哪些评价指标？
- 介绍下 AUC 和 F1-score?F1 值的计算公式说一下？怎么理解 AUC？
- 手写 AUC 的计算（小矩形积分得到总面积即可）
- 样本不均衡对 roc 曲线有影响吗？对 pr 曲线呢？为什么？
- 画一下混淆矩阵，写一下精确率和召回率的公式？
- 验证集是做什么的，测试集效果怎么评估？
- 介绍回归、分类用到的评价函数？

- 分类器评估标准（准确率，召回率，F1 值，ROC，AUC）
- 手写 recall，precision，f1score 公式以及公式中各个指标代表的含义？

2.5 传统机器学习方面

2.5.1 讲解相关原理

2.5.1.1 数据准备

无

2.5.1.2 特征工程

① 特征降维

- 有没有用过机器学习的降维方法？
- 讲一讲 LDA 算法（线性判别分析）？
- 介绍 PCA
- 特征工程预处理的流程？

② 特征选择

- 特征工程对于连续特征，我们通常有两种处理方式：1. 连续特征离散化；2. 特征缩放，这两种分别在什么情况下做？
- 讲了一些特征工程的技巧？
- 如何对连续特征进行离散化处理，为什么要这样做？
- 做数据分析选特征的时候有哪些评判指标？
- 所有模型都要求对数据进行标准化么？

2.5.1.3 有监督学习-分类和回归方面

① 分类回归树（集成学习）

- 讲一讲树模型（RF, GBDT, XGBOOST）
- GBDT 和随机森林的区别来说一下？

- GBDT 和随机森林的树的深度哪一个比较深？为什么？
- 知道哪些集成方法？
- 如何用回归树实现分类算法？
- 常用的线性分类算法有哪些？
- 常用的非线性分类算法有哪些？
- xgb 和 gbdt 的区别?gbdt+lr 实现细节？
- 讲一下 XGB 的原理，优缺点，推一下公式？
- LGB、XGB 的区别和联系，并行是如何并行的？
- 除了树模型，Bagging 能不能接其他的基模型？
- lightgbm 讲一下，具体是怎么做的？和 xgboost 的区别讲一下？
- lightgbm 的直方图加速讲一下？具体是怎么来做的？叶子节点是怎么分裂的？说一下？
- Xgboost 和 LGB 原理？
- 说一下 GBDT 的原理，boosting 和 bagging 是怎么减少偏差的？
- gbdt 各基学习器之间是如何产生联系的？
- 手写 xgboost 的目标函数，xgboost 构建树时候节点分裂的公式是什么？
- xgboost 如何调参,xgboost 可以自定义损失函数吗？
- 给定一个场景如何自定义损失函数？
- 如果样本的权重不一样如何自定义损失函数？
- sklearn 的 xgboost 支持哪些损失函数？
- 分类和回归算法都有哪些损失函数？
- 模型融合如何做的？
- bagging，boosting 和 stacking 的原理以及他们的区别是是什么？
- XGB+LR，XGB 充当什么角色？
- 为什么 XGB+LR 可以提高模型效果？

- 如何在 XGB 模型选择树的棵树时早停?
- XGB 的损失函数进行了二阶泰勒展开，为什么可以用泰勒展开？为什么用二阶而不是三阶四阶?
- 说一下自己理解的 XGBoost?
- XGBoost 和 GBDT 的区别是什么?

② 逻辑回归 LR

- 说了逻辑回归，在什么情况下你会选择用逻辑回归?
- 用 L-BFGS 来推导一下 logistic regression 的迭代公式?
- LR 损失函数介绍一下，如何优化?
- 讲一下逻辑回归？ 当数据量特别大的时候，逻辑回归(LR)怎么做并行化处理?
- LR 为什么不用 mse，svm 为什么用 hinge 不用 logloss，我不会，面试官耐心画图给我讲原理。问 svm 为什么要用核函数。
- 如果自己写一个 LR 的话，要包含哪些模块（我只说到了训练部分，面试官补充了还有分类预测的模块）
- LR 不做标准化有影响吗，神经网络呢?
- 简单介绍下 LR，写一下极大似然的函数?
- 送入 LR 前，如何处理数据（特征工程）
- 知道最大似然估计和最大后验概率估计么?
- 讲一下最大释然估计的原理？然后给出一个二项分布，让用最大释然估计手推出该分布的参数?
- 逻辑回归背后的数学原理是什么，如何推导的?

③ SVM（支持向量机）

- 讲一下 SVM (建模思想，误差函数推导，核，优化) SVM 的核函数有哪些？你都用过哪些?
- SVM 的推导对偶除了方便计算以外还有什么好处?
- 解释 SVM 的核函数，核函数的含义以及为什么能起作用?

- 为什么不用 SVM 做分类？从原理上讲一下 SVM，SVM 怎么解决多分类问题？
- LR，SVM 的原理，LR 和 SVM 区别，SVM 损失函数
- SVM 中有哪些调参经验？
- SVM 和 LR 的区别？
- SVM 数学上的实现？
- SVM： 拉格朗日乘子，KKT 条件，对偶问题，核方法是什么，用过哪些核函数？

④ 朴素贝叶斯（Naive Bayes）

- 贝叶斯思想了解吗？写一下公式并解释一下？
- 说一下朴素贝叶斯，为什么叫朴素贝叶斯？
- 朴素贝叶斯的好处？为什么那么多人用？

⑤ 决策树（DT）

- 决策树原理，CART 树？
- 决策树 ID3 算法的特征选择指标，口述一下数学公式(信息增益)
- 决策树的启发式算法有哪些，不同算法分别用了什么准则来选择特征？
- 介绍一下 ID3、CART，逻辑树
- 说一下 C4.5 的过程，围绕决策树

2.5.1.4 无监督学习-聚类方面

- 聚类了解吗，说一下 K-means 聚类的原理以及过程？K-means 聚类怎么衡量相似度的？(我说欧式距离)K 的优化方法？
- 说一下 kmeans 聚类算法的原理，对于 k 和中心点怎么确定
- 如何评估我们的聚类结果，以及如何提升？
- 介绍 K-means 聚类，以及每次聚类结果是否一致，为什么？
- 介绍下 kmeans？有什么改进方法么？遇到很多维数据时会发生什么？
- 讲一下混合高斯模型，EM 的核心思想是什么？

2.5.2 手推算法及代码

2.5.2.1 手推公式

- 写一下 Gini 系数、信息增益、信息增益率的公式?

2.5.2.2 手写代码

- 手推 SVM：空间上一点到超平面距离，SVM 整体代价函数，如果进行对偶形。
- 详解 GBDT，用伪代码实现树的生成和 boosting 迭代过程?

2.6 深度学习&机器学习面经通用知识点

2.6.1 损失函数方面

- 用过哪些损失函数，有什么区别?
- 分类和回归都用什么损失函数，分类为什么不用平方损失?
- 描述一下交叉熵 ，求导?
- 交叉熵损失函数定义，举例计算过程?
- 写一下 logstic regression 的损失函数
- 通过交叉熵做 loss，怎样体现，输入 x1、x2，输出 y1，y2 交叉熵如何计算，比如 y1 和 y1’越接近 loss 为何越小?

2.6.2 激活函数方面

- 了解哪些激活函数，为什么要激活函数?
- sigmoid 和 relu 对比?
- 激活函数为什么要零均值输出?
- 为什么要用非线性激活函数，relu 右侧导数是 1，为什么能作为激活函数?

2.6.3 网络优化梯度下降方面

- 参数优化方法说一下(梯度下降的三种方式的优缺点)

- 什么是梯度下降，有哪些优化算法，区别是什么，它们（SGD,BGD,mini-BGD）的区别?
- SGD 和 ADAM 的区别和联系？ADAM 算法比 SGD 优化好在哪儿?
- 深度学习里面的优化方法 momentum 和 Adam 来分别讲一下原理和公式?
- SGD 和 Batch 梯度下降区别?
- 怎么用的动态学习率，人工干预还是自动的?
- 手写 adam 更新公式?
- 平时怎么选择优化器？讲一下 adam 的优点?
- 深度学习常见优化方法有哪些?
- Momentumt 的公式，RMSPorp,adam 的公式以及公式中参数代表的含义，以及他们分别解决了什么问题?

2.6.4 正则化方面

- 正则化有哪些方法?
- 过拟合问题。我答了几个方法。然后着重问了一下正则化的内容，l1、l2 正则化
- 正则化是怎么防止过拟合的?
- 介绍一下正则，L1 L2 的比较，为啥 L1 更稀疏 ？
- L1 和 L2 的数学解释，L1、L2 有什么区别，适用于什么场景?L1 为什么不用于卷积神经网络?
- 从多个角度分析 l1 和 L2 正则化为什么能防止过拟?

2.6.5 过拟合&欠拟合方面

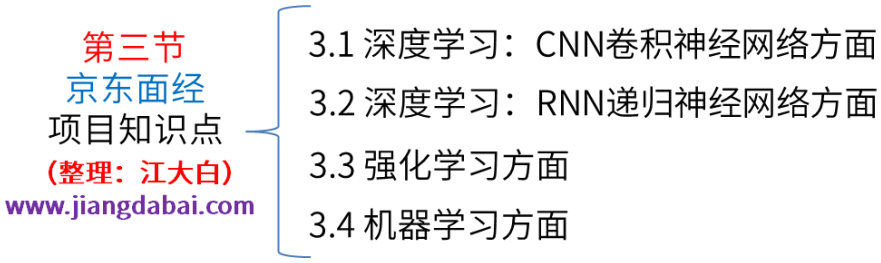
- 过拟合是什么，如何解决，应对措施?
- 讲了讲深度学习训练中过拟合/Loss 不降等常见问题的处理方法?
- 模型效果不好的前提下，如何区分是过拟合还是模型复杂度不够?
- 从模型结构上如何解决过拟合?
- 讲一下偏差和方差(楼主从欠拟合和过拟合来讲的)

- 如何判断一个模型是处于高方差还是高偏差?
- 高方差如何调节模型，高偏差如何调节模型?

2.6.6 其他方面

- 写一下欧式距离的公式?
- 传统机器学习都是一次把全部的数据送进模型，现在深度学习为什么一次就一批?
- 数据类别不平衡怎么处理?

3 京东面经涉及项目知识点



3.1 深度学习：CNN 卷积神经网络方面

3.1.1 目标检测方面

3.1.1.1 讲解原理

- faster rcnn 流程以及 RPN 的具体过程?
- 问项目相关，提升小目标检测效果的方法，kcf 的原理等等
- NMS 和 IOU 的计算
- Two-Stage 和 One-stage 结构的不同?
- 介绍了下 Faster RCNN，问了 ROI Pooling。和卷积中普通的 Pooling 有什么区别？你们做目标检测一般用什么损失函数？写一写 KL 散度和交叉熵函数?

3.1.1.2 损失函数

- 用过 Focal loss 吗?

- Yolo 的损失函数， v1 和 v3 损失函数的区别?

3.1.1.3 手写代码

- 手写 nms

3.1.2 图像分割

- 讲一下 unet 和 deeplabv2 的流程，顺便问了下 deeplabv3?

3.1.3 OCR

- 讲一下文本分类模型
- 给一个新的文本分类任务，会怎么选模型?

3.1.4 图像分类

- 常见的分类算法以及评估指标?

3.2 深度学习：RNN 递归神经网络方面

3.2.1 自然语言处理 NLP

① Bert

- Bert 原理?
- Bert 内部结构(矩阵参数信息等)
- 问 Bert 的 attention 和普通的 attention 的区别，具体怎么做的，多头为什么要多头?
- bert 怎么分词? bert 的输入具体是啥?

② Transformer

- 画一下 transformer，介绍一下结构，说一下维度
- transformer 介绍

③ CRF

- 说一下 CRF 模型?

④ HMM 隐马尔科夫模型

- 讲一下 HMM 模型?

⑤ Word2vec

- 简单介绍下 word2vec 原理，对比下 CBOW 和 skip-gram 的区别 ？
- Word2vec，看过源码吗？源码里面是如何负采样的，为什么要层次化 softmax，sigmod 在源码里面的计算方法是什么?
- Word2vec 三层结构很简单为什么效果这么好，word2vec 激活函数?
- Word2vec 和 bert 区别?
- 为什么 w2v 向量在语义空间内有很好的数学性质，比如相加减?

⑥ Deepwalk&Node2vec

- deepwalk 介绍一下，优缺点？ （本来问的 word2vec，我说没有做过 NLP 的东西，介绍 deepwalk 可以吗，面试官说可以） ，deepwalk 的损失函数?

⑦ 其他

- 如果 onehot 等操作之后维度过高你会怎么做?
- 分词与实体识别的区别关系?
- fasttext 原理，同样要求画框架?
- 在无上下文的情况下如何看两个词是否是同义词?

3.3 强化学习

3.3.1 讲解原理

- 生成式模型和判别式模型的区别，都有哪些?
- 生成式模型和判别式模型具体讲下?
- G 和 D 具体结构?
- GAN 算法的二进制交叉熵函数怎么实现的（极大似然估计）？
- 介绍一下 GAN 算法?

3.3.2 损失函数

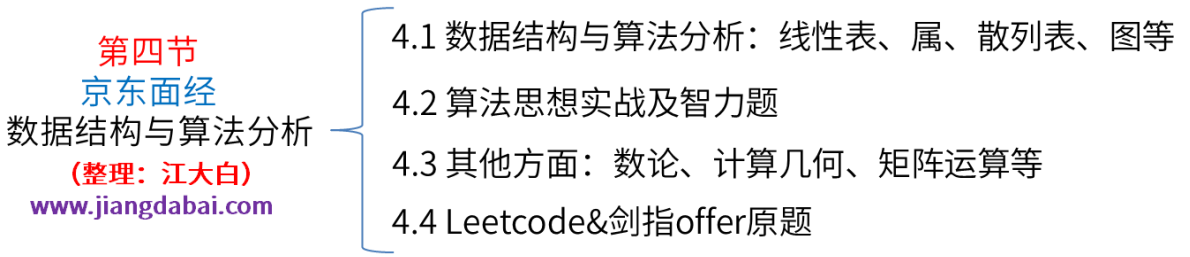
- G 网络的三种 loss 是怎么计算的，即 L1 loss L2 loss gan loss?
- 判别器 loss 如何度量?

3.4 机器学习方面

3.4.1 推荐系统

- 介绍一下 wide&deep 算法的原理?
- FM 模型与 LR 区别？怎么训练？FM 模型的具体公式，FFM 在此基础上有什么改进，如何确定每一个特征所属的 field?
- 搜索引擎的拼写纠正怎么做的？(楼主说了朴素贝叶斯和词袋模型)那如果第一个字母就输错了怎么办？词向量这一块有了解过吗?
- 讲一下 deepfm 的原理?
- 推荐里面的低秩矩阵分解具体是怎么做的?
- 个性化推荐是怎么样过程，组员如何分配任务？如何进行优化改进？如果要考虑输入集的权重，在 fp-growth 中如何实现？如何评判他的一个推荐标准？（用户采纳度、收藏或者点进去看）fp-growth 在这个项目中的优缺点?

4 数据结构与算法分析相关知识点



4.1 数据结构与算法分析

4.1.1 线性表

4.1.1.1 数组

- 求数组里面连续子段和最大值?
- 旋转数组查找 k
- 旋转数组中查找某给定数（二分查找）
- 一个数组中出现次数最多的 K 个数?
- 2^n 个数组，每个数组长度都是 m，每个都是有序，合并成一个有序的数组?
- 数组中最大子数组的和，矩阵中最大子矩阵块的和?
- 一个数组里面，每 K 个数是一个递增的有序数组，将整个数组排序?
- 给定一个有序数组，统计目标值的个数。（二分查找，找到目标值的下界和上界。）

4.1.1.2 链表

- 反转链表
- K 个一组，反转链表
- 单链表的分组翻转（即 k 个一组翻转链表）
- 如何判断一个链表上是否有环?
- 快慢指针如果快指针走 3 步的话呢 还能奏效吗? 如果快慢指针的起点不一样呢，还能奏效吗?
- 链表转化，1->2->3->4->5->6->7 转化为 1->7->2->6->3->5->4?
- 合并两个有序链表，合并 k 个有序链表（不让用递归），最小的 k 个数（指定用 partition，为什么时间复杂度是 O(n)）

4.1.1.3 字符串

- 字符串转 int?

- 字符串拼接成最大的数字的排序？
- 括号匹配：给定 n 对括号，求问总共有多少种排列方式？要求必须满足左右括号的顺序？
- 实现字符串的 lr_trim 算法，去掉字符串首尾的多个空格，函数名是 void lr_trim(char * s)

求字符串的最大回文子串(动态规划)

- 给一个小字符串数组，一个目标字符串，目的是找出数组中是目标字符串的子串的最大长度，我的想法就是维护字典树，加 KMP 模式匹配优化。
- 多个字符串，给定前缀和长度比例阈值，返回符合条件的字符串个数？
- 给出一个字符串，写出该子符串的全部排列组合？
- 反转字符串
- 给定 2 个字符串求最长公共子串的长度

4.1.2 树

4.1.2.1 二叉树

- 什么是二叉树、用伪代码说一下求二叉树的深度
- 二叉树镜像
- 给一个二叉树，和一个节点，找出该节点二叉树中序遍历下的下一个节点，如果树有父节点则个属性，在空间复杂度 O(1)的情况下找出来？
- 判断两棵树是否相同？
- 求二叉树每一层的最大值？
- 用伪代码说一下求二叉树的深度，如果用递归，具体实现，代码？
- 二叉树前序遍历？
- 二叉树的层次遍历

4.1.2.2 堆

- 寻找无序数组中的第 K 大的数，这个只需要说思路和复杂度？用最小堆 O(NlogK)
- 无序数组中找第 K 大的数，时间复杂度是多少？为什么是 O(n)，而不是 O(nlogk)，来推导

并且证明一下你的解法的时间复杂度(级数求和)、O(nlogk)的解法是怎么做的，说一下(堆排序)?

4.1.3 图

- 问最短路径算法有哪些？介绍一下 a*算法。回答：有起点到中间点的距离加上中间点到终点的一个估计距离。面试官又问，如果去掉第一项这个问题会变成什么？

4.1.4 排序

- 各种排序算法说下，写个插入排序
- 冒泡和快排是否稳定？还有哪些稳定的算法？
- 无序数组找第 k 大？快排&堆。写了快排。分别讲了一下两种思路，分析两个的时间复杂度。
- 如果几亿的数分别在不同机器上，怎么找第 k 大？
- 海量数据如何找到中位数和第 200W 个数
- 有 4 亿个数据，内存只能存 1 亿个数，找出第 8000 万大的数？你用堆排序做是吧？堆排序这种方法有什么缺点？如果我要找第 1.3 亿大的数呢？
- 手写快排，推复杂度
- 堆排序（C++）
- 1000w 个数，数范围[-1000, 1000]，写个排序？

4.2 算法思想实战及智力题

4.2.1 算法思想实战

- 上台阶，一次能上一个或者两个，问上 n 个台阶有多少种方法？（思路+撕代码）
- 纸牌博弈
- 最短路径问题
- 优先队列，列表中出现次数最多的数字，复杂度小于 nlog2n。
- 有一个 1G 大小的一个文件，里面每一行是一个词，词的大小不超过 16 字节，内存限制大

小是 1M，返回频数最高的 100 个词。

4.2.2 智力题

- 一维消消乐，红红黄黄绿绿绿绿黄蓝 ->消一次变成　红红蓝　（三个以上的干掉）？

4.3 其他方面

4.3.1 数论

- 两个独立变量满足 0 到 1 均匀分布，求两个变量最大值的期望。

$$\$max\{X_1, X_2\} = \frac{X_1+X_2+|X_1-X_2|}{2} E(max\{X_1, X_2\}) = \frac{1}{2} + \frac{1}{2} \times E|X_1 - X_2|\$$$

4.3.2 计算几何

- 凸函数有什么优点？如何证明一个 n 元函数是凸函数？
- 拉格朗日乘子法能否求解非凸的目标函数问题？
- 凸优化的相关知识。
- 马尔科夫链的相关知识
- 线性代数里面的矩阵分解你知道吗？具体是怎么做的？

4.3.3 概率分析

- 最大似然估计解释下？
- 极大似然估计和最大后验估计的区别是什么？
- 计算抛骰子，抛 1 或者 6 庄家赢，2，3，4，5 你赢，

庄家连续赢了三次，这个概率是多大，这样能说明骰子有问题吗？

如果抛了 100 次，庄家赢了 40 次，能说明有问题吗？

那怎样才能证明这个骰子有没有问题？

- 54 张扑克牌，大小王在同一堆的概率？
- a,b 丢硬币吃苹果，问吃到的概率？

- M 个样本有放回采样 N 次，问某条样本一次没被采中的概率？
- 某人有两个孩子，其中一个是女孩，两个孩子都是女孩的概率是多少？

4.3.4 矩阵运算

- 二维矩阵，从左到右从上到下递增，找 target？
- 给出一个二维矩阵，顺时针由外层到内层打印该二维矩阵？

4.3.5 其他

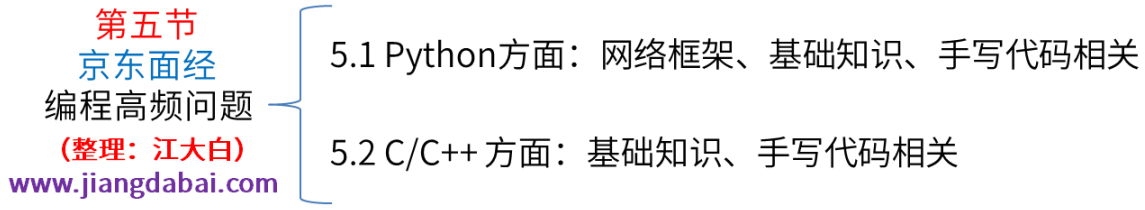
- 求 sinx？
- 实现 int sqrt(int x)函数
- 正则表达式匹配
- 已知有个 rand7() 的函数，返回 1 到 7 随机自然数，怎样利用这个 rand7() 构造 rand10()，随机 1 ~ 10？
- rand1()生成 rand5()？
- 给你均值方差，让你利用正态分布随机生成 1000 个点？
- 乱序数据找第 K 大的数
- 字符流采 10 个字符，保证每个字符的采样概率一样？
- 动态规划，左右，求最大的路径和？

4.4 Leetcode&剑指 offer 原题

- Leetcode 4：两个排序数组的中位数
- Leet code 47：全排列 II
- Leetcode 143：重排链表
- Leetcode 152：求数组的最大连续子序和。时间空间复杂度？空间复杂度优化？空间复杂度优化？如果是连续乘积呢？
- Leetcode 215：求数组中第 K 大的数

- Leetcode 206：反转链表
- Leetcode 279：完全平方数
- 剑指 offer 11：旋转数组的最小数字
- 剑指 offer 51：数组中逆序对

5 编程高频问题：Python&C/C++方面



5.1 python 方面

5.1.1 网络框架方面

5.1.1.1 Pytorch 相关

- 项目用的什么平台，keras、tensorflow、pytorch 都是哪家公司的，为什么喜欢用 pytorch?
- tensorflew 与 pytorch 区别?

5.1.1.2 Tensorflow 相关

- tensorflow 或者 caffe 的底层代码看过吗？卷积是怎么实现的？GPU 进行并行计算时如何计算矩阵卷积的?
- tensorflow 中两个矩阵乘法的区别?

5.1.1.3 其他

- 常用的深度学习框架都有哪些?
- keras sequential 与自定义模型构建区别?

5.1.2 基础知识

5.1.2.1 线程相关

- 讲一下 Python 的多线程

5.1.2.2 内存相关

- Python 需要和 C++一样释放内存吗?
- Python 垃圾回收
- 讲一讲 python 内存

5.1.2.3 区别比较

- python 参数 * 和 ** 区别?
- list 和 tuple 区别?
- xrang 和 rang 的区别?

5.1.2.4 讲解原理

- python 的 with 什么意思?
- python 的 dict 实现，哈希表查找的时间复杂度一定是 O(1)么？为什么？怎么解决?
- 如果哈希表发生大量冲突，怎么解决（想到了二叉搜索树，面试官问了解红黑树么）

5.2 C/C++方面

5.2.1 基础知识

5.2.1.1 内存相关

- C 的内存对齐，给了几个 struct 计算占用内存?

5.2.1.2 区别比较

- C 和 C++中的 struct 和 class 的区别?
- 介绍面向对象和面向过程的区别?

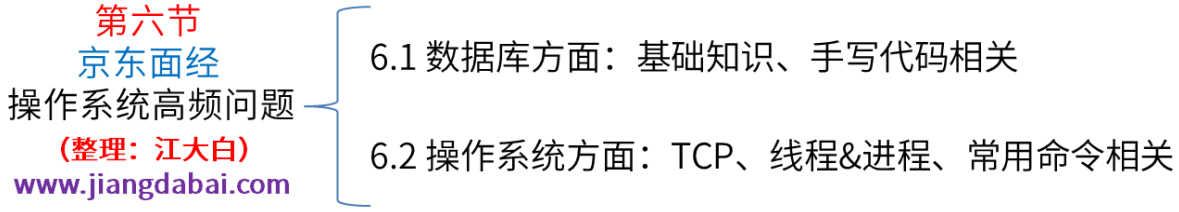
5.2.1.3 讲解原理

- C++继承、重载、虚函数的相关知识
- C　纯虚函数、虚函数表说一下
- 全局变量，静态全局变量存储位置

5.2.2 手写代码相关

- 给定一个数组，相邻元素之差的绝对值<=1，如[1,2,3,2,2,1,2]，如何快速查找某个数?
- a=1，2，b=（1，2），问 ab 输出?不用中间变量交换 ab，用异或?

6　操作系统高频问题：数据库&线程&常用命令等



6.1 数据库方面

- 口述一道 SQL 题，差集?
- 数据库相关，主键和外键的区别?
- 数据库的索引介绍一下（B+树）
- sql 常用关键字的执行顺序

6.2 操作系统方面

6.2.1 TCP 协议相关

- TCP/IP 四层模型
- 计算机网络：TCP 和 UDP 的区别，OSI 七层模型、全双工和单工?

6.2.2 线程和进程相关

- 多线程多进程问题，cpu，磁盘 io 哪种多线程带来效果好?
- 多进程适合处理什么类型的问题？（应该是想让回答 I/O 型？？）适合解决计算复杂的问题么?

6.2.3 常用命令

- Linux 基本命令，说下对管道的理解?
- 强制杀死某个进程的命令、其他的命令?
- Linux 更改文件权限的，那改可执行的是多少？777？详细解释下?
- 查看文件前十行?

7 技术&产品&开放性问题

7.1 技术方面

- 京东有 20 万人，要做通讯录，包括名字和电话，名字可以重复，电话不可以，要实现快速的增删改查的话，用什么数据结构比较好?
- 假设京东有 1000 台服务器，每台上面有 100g 的日志文件，然后现在要在自己的服务器上进行统计出现次数最多的 ip 地址（服务器上不能进行统计）？
- 场景题：有用户 feed 流和点击信息，如何做推荐 ？
- 场景题：每天的用户，商品，销量订单记录，求销量前 10 的商品，用 SQL 或者什么编程语言写一下?
- 场景题：商品销量的时间序列数据如何分解?
- 场景题：解一个在北京地区的不同库房分配某一个商品的混合整数规划问题（要求说出目标函数，限制条件和求解法，也算磕磕碰碰答上来了）
- 场景题：打开京东 APP，点击一个商品，详情页会显示 XXX 也买过，这个具体是怎么做的?(扯到了协同过滤和冷启动)

- 场景题：现在我们有两个排序模型，分别是 A 和 B，他们分别预测出了对应的排序的结果在我们的 APP 上，我们有真实地用户数据，那么怎么来评估这两个模型的好坏呢？用说出数学公式。
- 场景题： 在京东有许多的不同的消费群体，我们如何找到学生群体并把他介绍给我们的用户。（大概就是聚类的详细建模的过程）
- 业务代码：

If (a < 3):

If (b……):

If(c…....):

If(d…….):

If(e……)

If (a >=3):

If (b……):

If(c…....):

If(d…….):

If(e……)

这种 if 语句嵌套太多，条件也太多了，怎么来优化它？（面试官一步一步提示：决策树相关）

- 场景题：京东搜索里，输入一个关键词搜索某件商品，但是现有数据库里没有该关键词，该商品对应的是另一个关键词，这种情况下怎么解决？

7.2 产品方面

- 基于京东的数据（地理位置、活跃度等）来判断黄牛卖家？

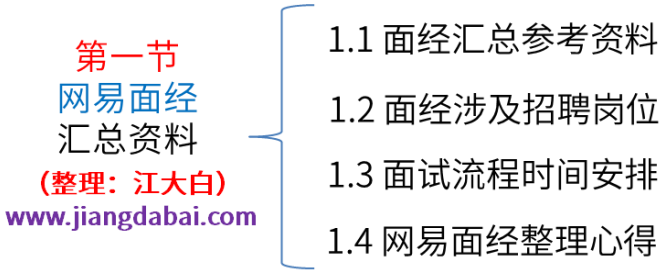
7.3 开放性问题

- 如何给一个完全没有接触过机器学习的人介绍机器学习，机器学习是做什么的？

- 推荐岗位相关：你想找推荐，我们是机器学习+组合优化，偏向运筹学，考虑么？

8|网易算法岗武功秘籍

1 网易面经汇总资料



1.1 面经汇总参考资料

① 参考资料：

（1）牛客网：网易面经-85 篇，网页链接

（2）知乎面经：点击进入查看

（3）面试圈：点击进入查看

② 面经框架&答案&目录&心得：

（1）面经框架及参考答案：点击进入查看

（2）大厂目录及整理心得：点击进入查看

1.2 面经涉及招聘岗位

（1） 实习岗位类

【网易机器学习实习岗】、【杭州研究院图像算法实习岗】

（2） 全职岗位类

【网易游戏研发工程师】、【网易考拉机器学习工程师】、【网易考拉 NLP 工程师】、【数据挖掘工

程师】、【网易有道云算法工程师】、【网易雷火机器学习】、【网易互娱 NLP】、【网易云音乐机器学习】、【网易严选算法工程师】、【网易云音乐机器学习】、【网易云音乐音频算法工程师】

1.3 面试流程时间安排

网易面试流程-整理：江大白			
	面试类型	面试流程	备注（侧重点）
第一面	技术面	自我介绍+项目/实习经验 +技术问答	项目为主
第二面	技术面	自我介绍+项目/实习经验 +技术问答+算法编程	项目为主
第三面	技术Leader面	自我介绍+项目经验+公司发展	/
第四面	HR面	基础人力问题	/

PS：以上流程为大白总结归纳所得，以供参考。

其他注意点：

● 第三面，技术 Leader 总监面，有的人有，有的人没有

1.4 网易面经面试心得汇总

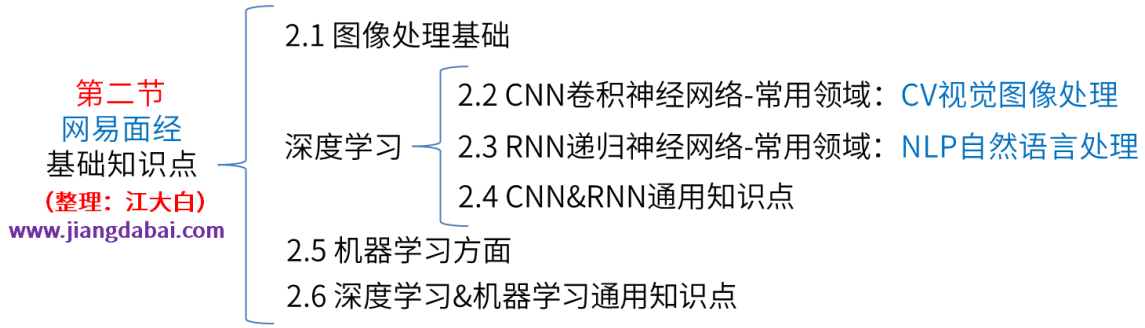
★ 先准备简历，简历干净凝练，最好所有内容都直指一个方向，最重要的是，简历上的任何东西一定要搞的明明白白，并且可以侃侃而谈，这很关键。

★ 相当重视基础，问的比较偏特征工程。先聊项目论文和实习，但是没细问。

★ 主要是问项目，然后根据项目里问一些细的技术点，每个项目都仔细问，确认是不是本人做的。

★ 面试官的风格是连环追问，你说一个答案他会在你基础上再问，为什么用这些？

2 网易面经涉及基础知识点



2.1 图像处理基础

- 传统图像处理：边缘检测、均值滤波、霍夫变换

2.2 深度学习：CNN 卷积神经网络方面

2.2.1 讲解相关原理

2.2.1.1 卷积方面

- 1*1 卷积的作用?
- CNN 卷积核的计算，参数量怎么算?
- 传统卷积核 depthwise conv 的区别?
- 是否了解空洞卷积?
- 卷积核参数怎么算，DW 卷积参数怎么算?
- 简历中提到了 dropout，那具体说下 dropout 操作细节、怎么解决 dropout 之后输出值减小的问题
- Dropout、BN 的原理、训练过程以及测试过程的具体做法?
- CNN 可视化特征?
- CNN 原理，池化的目的?

2.2.1.2 池化方面

- 介绍 CNN 为什么要用池化层?

2.2.1.3 网络结构方面

- shuffleNet v2，mobileNet v2 的结构?
- senet 结构？senet 为什么是这样的？为什么非要加池化呢?
- ResNet50 的 bottleneck 结构（conv1+bn+ReLU+conv2....)，与 ResNet18 差别?
- 了解哪些 backbone，MobileNet v1 v2 v3 的差别?
- resnet 比 alex vgg 一类有什么好处?
- 为什么 resnet 可以深一些?
- VGG 优点，resnet 优点特点?

2.2.1.4 其他方面

- 介绍下 CNN 原理？CNN 之所以成功的原因是什么？CNN 的参数共享是什么?
- BN 的原理，BN 的训练和测试的区别，讲一下过程?
- BN 层介绍，作用，为什么?
- BN 层参数有几个，forward 怎么计算?
- 说一下 BN 的归一化操作放在激活前还是激活后，两种有什么区别?
- 梯度爆炸和梯度消失原因，解决策略?
- 深度学习权重初始化方法有哪些?
- relu 为什么可以解决梯度消除，除了这个好处还有什么好处?

2.2.2 数学计算

- 输入图像 N*N，卷积核 k*k，问计算方式，时间复杂度？补充：单通道情况 O(N*N*k*k)，如何优化？提示了重复计算?
- 输入 n*n*1，卷积和 k*k*6， 问输出尺寸和参数量?

2.2.3 公式推导

- Softmax 公式?

2.3 深度学习：RNN 递归神经网络方面

2.3.1 讲解相关原理

- 介绍下 RNN?
- RNN 的改进有哪些？讲 lstm 和 gru，对于更长的序列怎么处理
- LSTM 和 RNN 的区别？lstm 为什么好?
- LSTM 的思想，原理，结构和公式?

2.3.2 手绘网络原理

- 推一遍 LSTM?
- RNN 原理，画一下?
- 写 RNN 公式

2.4 深度学习：CNN&RNN 通用的问题

2.4.1 基础知识点

- Attention 机制及它是如何聚焦的?
- 交叉验证的原理，作用，与直接划分训练集和测试集相比的优点?
- 数据增强方法?

2.4.2 模型评价

- 如何衡量模型性能，指标如何计算？常用的评价指标及含义?
- AUC 的意义，ROC 的绘制方式，AUC 的优势（不平衡数据集的情况）
- 如何评价一个分类器，AUC 的工程计算方式，ROC 曲线?

2.5 传统机器学习方面

2.5.1 讲解相关原理

2.5.1.1 数据准备

- 数据清洗如何做?
- MCMC 如何实现抽样，MH 和 GIbbs 抽样的区别，详细讲一下他们都是怎么实现的?

2.5.1.2 特征工程

① 特征降维

- 降维方法有哪些? PCA 的原理，作用?
- SVD 的原理?

② 特征选择

- 常用特征选择的方法? 如何把类别型数据转为数值型? 数据缺失值处理?

2.5.1.3 有监督学习-分类和回归方面

① 分类回归树（集成学习）

- 为什么两个重分类分支融合的不做 bagging 和 boosting，两个概念的本质上是什么意思，为什么效果会变好呢? ，boosting 为什么多个分类器合在一起效果就变好了?
- xgb、gbdt、RF 区别与联系、xgb 的优势? RF/AdaBoost/GBDT 怎么做推断?

A.基于 bagging：随机森林

- 随机森林了解吗? 随机森林是什么，每棵树有啥区别? 对比了一下 RF 和 GBDT?
- 随机森林/adaboost/gbdt 的原理，相同点不同点。在训练和测试阶段各是如何进行的?
- 随机森林的树会不会限制它的生长（不会），gbdt 的树呢（会），为什么?
- 随机森林和 GBDT 的差异描述一下?

B.基于 boosting：Adaboost、GDBT、XGBoost

- 介绍一下 GBDT? GBDT 有哪些可调的参数呢? 一般怎么选取参数呢?

- GBDT 如何进行多分类?
- 讲一下 xgboost 的原理 ?xgboost 的过程、损失函数?为啥泰勒展开成二阶，作用?
- 为什么 XGBoost 效果好于随机森林?RF 怎么解决的过拟合问题?
- 如何做特征选择，xgboost 做特征选择的时候，重要性是如何确定的，信息增益做特征选择如何做的，特征选择方法？xgboost 如何一步步构建分类树的?

② 线性回归

- 线性回归和逻辑回归区别?

③ 逻辑回归 LR

- 介绍下逻辑回归原理？逻辑回归推导?
- LR，SVM 怎么算的，损失函数是什么?
- LR 处理的特征是离散的还是连续的？离散化，会有什么影响吗，比如一个特征取值 0-1，需要离散化吗?
- 写 LR 的公式和 loss function
- LR 的 loss 是什么?
- 把 LR 的 loss 改成平方损失可以么？为什么?
- LR 的梯度下降有几种优化器?
- LR 中存在相同特征的话，对模型预估有影响吗?

④ SVM（支持向量机）

- SVM 了解吗？（SVM 是通过最小间隔最大化寻找超平面）；为什么么要最大化最小间隔呢?
- SVM 需不需要做 normalization?
- SVM 和 LR 的区别?
- SVM 有哪些核函数，对应有哪些使用场景和特点?
- 对于 SVM，假如先把数据映射到高维，然后不使用核函数，如何？为什么大多数人选择使用核函数?

⑤ 决策树（DT）

- 说一下决策树的原理?
- 决策树的分裂策略：ID3，C4.5，Gini 指数，选一个讲一下?
- ID3、C4.5、CART 的区别
- 写信息增益、信息增益率、基尼系数的公式，讲解原理
- 树有几种剪枝的方式，各有什么优缺点?

2.5.1.4 无监督学习-聚类方面

- 问了 kmeans 的计算过程?
- 哪些条件对 kmeans 的影响最大？k 值选取?
- kmeans 时间复杂度和空间复杂度?
- kmeans 和 EM 算法的关系，哪一步是 E 步，哪一步是 M 步?

2.5.2 手推算法及代码

2.5.2.1 手推公式

- 完整推导 SVM?
- LR 公式写一下?
- GBDT 推导一下?

2.5.2.2 手写代码

- 手写 Kmeans

2.6 深度学习&机器学习面经通用知识点

2.6.1 损失函数方面

- 常用 loss 函数?
- 写一下多分类交叉熵的公式?
- 解释 triplet loss，详解 triplet loss，triplet loss 反向传播?
- 分类为什么用交叉熵?

2.6.2 激活函数方面

- 各种激活函数介绍一下，优缺点及适用场景？
- 激活函数的作用？
- sigmoid 和 softmax 的区别？
- 为什么使用 sigmoid 激活函数会导致梯度消失？

2.6.3 网络优化梯度下降方面

- 用过那些优化器，SGD 和 Adam 分别在什么情况下使用？
- 牛顿法和拟牛顿法讲一下过程？
- 有哪些二阶优化方法，牛顿法存在什么问题？
- sgd 和 adam 的区别，有啥优点？
- bgd，sgd，mini-bgd 的区别？
- 手推：adam, sgd, momentum,rmsprop，通过公式解释一下为什么他们能优化？
- 最优化方法，ADAM？
- FTRL、牛顿法懂吗？
- SGD 与牛顿法的区别在哪？

2.6.4 正则化方面

- 各种正则化原理、方式及各自优缺点？L1 和 L2 正则的区别？
- 从数学角度讲一下正则项为什么能防止过拟合？
- 写一下 LR 的损失函数，加上 L1 /L2 正则化；然后解释原理，分析不同点，怎么用？

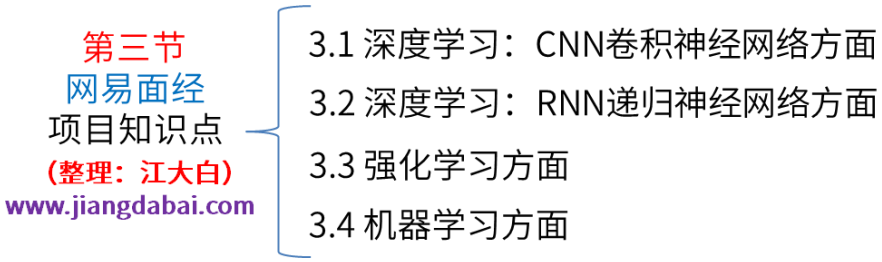
2.6.5 压缩&剪枝&量化&加速

- 剪枝用了哪几种方法，怎么训练的？
- 用过量化吗？，模型训练完成后 32bit 怎么量化为 8bit？

2.6.6 过拟合&欠拟合方面

- 描述下过拟合和欠拟合?
- 如何判断过拟合，如何解决过拟合?
- 为什么会减小过拟合的风险？（惩罚非线性参数的力度较大，减小模型非线性的程度）

3 网易面经涉及项目知识点



3.1 深度学习：CNN 卷积神经网络方面

3.1.1 目标检测方面

3.1.1.1 讲解原理

- 目标检测的整个流程，包括数据处理、模型训练、模型选择?
- 说说 Faster-RCNN，YOLO，SSD，FPN?
- Faster-rcnn 的训练过程是怎样的?
- 目标检测中的 mAP?
- 介绍你用过的一个目标检测算法?
- Faster R-CNN 的具体流程（简历里有提到）?
- Faster R-CNN 训练和测试的流程有什么不一样?
- 如何从 rpn 网络生成的多个候选框中确定出目标候选框?
- ROI Pooling 怎么实现的?
- ROI Align 原理?

- YOLOv3 和 Faster R-CNN 的差异?
- YOLO 系列有几个版本?
- YOLOv4 用到了哪优化方法? transformer 了解过吗?
- 解释 FCOS 正负样本回归的几种方式和特点?
- 多标签分类准确率，数据类别不平衡?

3.1.1.2 损失函数

- focal loss 介绍?

3.1.2 OCR

- CTPN 原理，文字识别的实现，遇到的问题，数据集，做的改进?
- 车牌识别实现，遇到的问题，数据集，数据增强?
- 文字检测的一些 trick 说一下?
- 最新的文字检测方法了解吗?

3.1.3 超分辨

- 超分算法的缺陷及改进方向?
- 超分落地要考虑哪些问题?
- 如何将超分算法与视频压缩相结合?
- 超分算法中不同采样方式的对比?

3.1.4 目标重识别

- ReID 常用的方法? 三元组损失怎么训练?

3.1.5 音频算法

- 用到的模型，以及对 Kaldi 和端到端的了解?
- 在语音降噪过程中，怎样实现噪声估计?
- 使用模型降噪的原理?

- 语音识别用到的模型?
- OMLSA? MCRA 用过吗?
- FIR 和 IIR?
- 16K 重采样到 8K，怎么做?
- MFCC 的原理?
- 频率分辨率
- 如何做卷积?

3.2 深度学习：RNN 递归神经网络方面

3.2.1 自然语言处理 NLP

① Bert

- BERT、GPT、ELMO 之间的区别?（模型结构、训练方式）

② CRF

- CRF 的原理讲一下?

③ Word2vec

- 讲一下 word2vec 怎么实现?
- Word2vec 和 bert 区别?
- word2vec 的缺点，word2vec 的输出是什么?
- Word2vec，负采样，层次归一化?

④ 其他

- 画出 fasttext 的网络结构，描述其在分类和 embedding 时的区别。详述训练过程
- 介绍方面级的情感分析模型? 情感分析任务用哪个数据集?
- tfidf 的计算公式?
- 如何进行句子编码，提取句子的特征向量，有哪几种方式（CNN，LSTM，Attention），各

种方式的优缺点?

3.3 强化学习

3.3.1 讲解原理

- 用过的 CycleGAN 介绍一下
- 了解 GAN 吗，简单介绍一下?
- GAN 在文本生成中如何应用?
- 如何解决 GAN 中文本离散的问题?
- 如何解决 GAN 中生成器与判别器训练不平衡的问题?
- 为什么用 PPO 算法?
- 讲一下 A3C 异步效果为什么可能不收敛?
- 提升样本使用效率还有什么方案?

3.3.2 损失函数

- 介绍 GAN，实际实现中的 loss 是什么?

3.4 机器学习方面

3.4.1 推荐系统

- 讲一下协同过滤的原理?
- FM 公式?
- deepFM 的原理?

4 数据结构与算法分析相关知识点

第四节
网易面经
数据结构与算法分析
（整理：江大白）
www.jiangdabai.com

- 4.1 数据结构与算法分析：线性表、属、散列表、图等
- 4.2 算法思想实战及智力题
- 4.3 其他方面：数论、计算几何、矩阵运算等
- 4.4 Leetcode&剑指offer原题

4.1 数据结构与算法分析

4.1.1 线性表

4.1.1.1 数组

- 在乱序数组当中找出中位数?
- 求数组的最长子集，整个子集满足所有的元素两两模除为 0?
- 两个一样的数组，其中一个数组插入了一个数，如何找到它的位置?
- 一个二维数组从左到右，从上到下都是增大的，找一个数，又问了时间复杂度?
- 数组有 n 个整数，每次对 n-1 个数全部加 1,求最少几次让所有的数相等?
- 16 进制转 10 进制，最好用 C 语言来写，实在不行也可以用 python。

4.1.1.2 链表

- 环形链表 判断有无环
- 找环的入口
- 两条链表求第一个公共节点?
- 两个不等长链表的公共节点?
- 合并两个有序链表?

4.1.1.3 字符串

- 翻转字符串
- 给个字符串，返回最长无重复的子串?

- 找出字符串的所有全排列?
- 求最长匹配括号的长度?
- 给定一个字符串，只保留 k 位，不改变字符间的顺序，使得字符串字典序最小?
- 给定字符串，求长度为 k 的字典序最小的子序列?

4.1.2 树

- 哈夫曼树相关

4.1.3 排序

- 排序算法了解吗？快速排序的时间复杂度怎么样呢？快速排序是稳定的吗？有哪些稳定排序算法呢?
- 解释一下排序的稳定性，冒泡排序是否稳定，复杂度多少?
- Topk 问题
- 有足够多的数据（内存无法一次性装下），如何获得最大的 k 个数?
- 写一下快排，讲一下最好和最坏的情况?
- 写快排里的 partition 函数?
- 介绍堆排序?
- K 个最大值：堆排序

4.2 算法思想实战及智力题

4.2.1 算法思想实战

- 洗牌算法，有多少种可能性以及数学证明（很简单，Knuth 洗牌算法）
- 图找环方法，有向图和无向图找环区别?

4.2.2 智力题

- 问三个囚犯一碗汤,每次都分的不均不开心，设计分法让大家都满意?

4.3 其他方面

4.3.1 数论

- 两枚硬币,依次掷，两枚掷完算一次，到第五次两枚第一次同时出现相同的面的期望值?
- 如何在一堆数里确定是否存在某几个数?

4.3.2 计算几何

- 数轴上某些位置有点，每个点都有一个速度和方向（左或右），在零时刻他们开始运动，求第一次有两点相碰的时间？如果只有相反方向的相碰才算，如何求解？

4.3.3 概率分析

- 10000 个黑球、10000 个白球，混合在一个桶里，无放回的取两个球，异色放白球，同色放黑球，求最后一次是黑球的概率?
- 54 张扑克牌，分三堆，其中 4 张 A 在同一堆的概率?
- 流数据 n 个中随机取 k 个数，每次只能取一个，怎么使取到每个的概率相等?
- 27 个球，有一个轻的，找出来最少需要几次?
- 一条绳子切两刀得到的三段线组成三角形的概率?

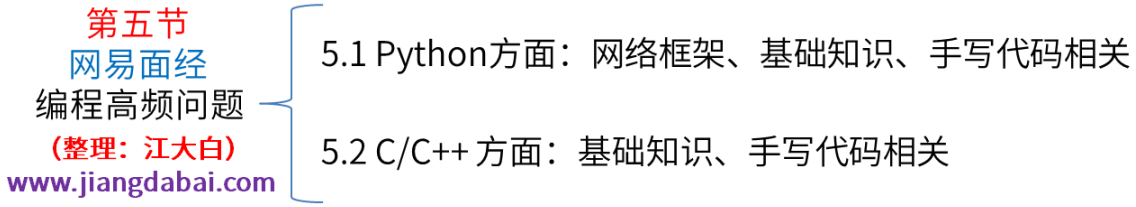
4.3.4 其他

- 给出 n 个点的坐标（x1,y1）...(xn,yn)，找出其中离(a,b)最近的点，要求 xi,yi,a,b 的数字动态变化，实现高频查找?
- 给定一个长度为 n 的序列，将其分割成若干连续子序列，若这些序列构成的数能整除 m，输出相应结果?
- 有序列表合并?
- N 的阶乘后面有几个零?
- 求 0 - N-1 的全排列输出?

4.4 Leetcode&剑指 offer 原题

- Leetcode 33：搜索旋转排序数组
- Leetcode 518
- Leetcode 原题：射气球
- 剑指 offer 41：数据流中的中位数
- 剑指 offer 原题：二维向下向右递增的矩阵查找

5 编程高频问题：Python&C/C++方面



5.1 python 方面

5.1.1 网络框架方面

5.1.1.1 Pytorch 相关

- Pytorch 和 TensorFlow 有什么区别?

5.1.1.2 Tensorflow 相关

- 手写了 tensorflow 的图像分类代码?
- tensorflow 有哪些缺点?

5.1.2 基础知识

5.1.2.1 线程相关

- python 多线程有多少了解？和 c 的多线程最大区别是什么?
- python 多线程缺点?

5.1.2.2 内存相关

- Python 怎么做内存回收?

5.1.2.3 区别比较

- Python xrange 和 range 差别?
- 装饰器及多进程和多线程区别?
- Python 浅拷贝和深拷贝有什么区别?

5.1.2.4 讲解原理

- Python yield 关键字是什么用的?
- 问了 Python 怎么加速?(np 矩阵乘法替代循环)
- Python 的命名规则、self、lambda、with、循环引用

5.2 C/C++方面

5.2.1 基础知识

5.2.1.1 内存相关

- C++的栈区和堆区知道吗？分别是干什么用的？（栈区是存储函数内部变量的内存区，堆区是存动态申请的内存）
- 什么时候要进行动态内存申请？（以前没思考过，没答上来，后来查了一下，当无法事先确定对象需要使用多少内存（这些对象所需的内存大小只有在程序运行的时候才能确定）时就要申请动态内存，比如维护一个动态增长的链表或树）
- 栈和函数调用的关系?

5.2.1.2 区别比较

- C++和 Python 的区别?
- 栈和堆的区别?

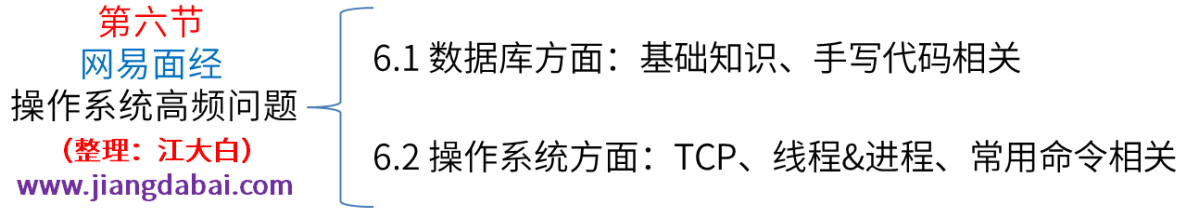
5.2.1.3 讲解原理

- 多态，虚函数
- 为什么 C++比 Python 快?
- Python 比 C++好在哪里（自动能实现内存回收机制）
- C++虚函数（如何实现，有什么功能）、虚函数表
- 类实例化的机制，具体实现是什么?
- stl 容器分类?
- 析构函数的基类为什么是虚函数?

5.2.2 手写代码相关

- 16 进制转 10 进制，最好用 C 语言来写?
- vector 排序

6 操作系统高频问题：数据库&线程&常用命令等



6.1 数据库方面

无

6.2 操作系统方面

6.2.1 线程和进程相关

- 进程和线程区别?

6.2.2 常用命令

- 如何查看某进程关联的相关文件有哪些?

7 技术&产品&开放性问题

7.1 技术方面

- 不平衡数据的解决方式，数据分布改变了怎么办?
- 直播中如何判断人眼关注点区域?
- 视频通话场景如何估计噪声?
- h.264 压缩优化?
- 图像复原与图像增强的区别与联系?
- 如果有一百万个游戏片段，仅有少数有标记，如何利用这些数据?
- 如果数据的维度很高（3 万），如何完成聚类?
- 100W 个起始结束 IP 段以及对应中文名，建一个系统，让它可以很快查找出某个 IP 对应的中文名?
- 开放题：如何设计一个人脸检索系统（从数据、模型、loss 等考虑）
- 开放题：对 CV 类前景的看法

7.2 产品方面

- 场景题：根据游戏用户反馈的问题，进行信息分类与关键词提取，给出一个比较详细的综合方案（从技术到实际运行）

7.3 开放性问题

- 假如面试官是一个零基础的深度学习学习者，那么有什么建议，如何入门?
- 如果某个事业部想要销量提升 20%，作为数据分析人员，给什么方案?