

张胜东

- 性别：男
- 出生年月：1991.11
- 学历：硕士研究生
- QQ：435878393
- 手机：13912998528
- Email：435878393@qq.com
- 个人主页：<https://www.zhangshengdong.com>
- Github：<https://github.com/zhangsheng377>

教育经历

江苏省南京市金陵中学 2007.9-2010.7

长春理工大学光电信息学院本科 2010.9-2014.7

- 物理系,光学专业
- 获奖情况：

2010-2011 学年 一等奖学金
2011-2012 学年 一等奖学金
2012-2013 学年 一等奖学金
2013-2014 学年 国家奖学金
校级 和 院级 优秀毕业论文

- 获证书情况：

全国蓝桥杯软件大赛全国总决赛 二等奖
全国大学生数学建模竞赛吉林赛区 二等奖
吉林省程序设计大赛(acm) 一等奖
吉林省电子设计大赛 二等奖
全国信息技术考试数据库工程师认证
全国计算机四级数据库工程师

南京邮电大学研究生 2015.9-2018.7

- 计算机技术专业

- 获奖情况：

凯易讯软件大赛全国总决赛 第 25 名
中兴软件编程大赛 决赛
黑马大赛全国总决赛 第 2 名
全国物联网设计竞赛 二等奖

- 主要项目：

- i. 卫星鉴权高并发服务器和客户端：

使用了 **epoll, fork, socket, pipe, 命名管道 FIFO** 等的服务器和客户端程序，功能是串口连接北邮的网关设备进行控制，同时接收用户的信息，若是登录信息就进行鉴权，鉴权通过就将此端口转发规则写入网关设备，同时还涉及到信道分配，流量管理，加解密通讯等。

- ii. 大气质量监测及预测系统：

树莓派开 **gpio** 读取传感器并上传云端，同时有触屏显示界面；跨平台客户端从云端读取数据，并作 **3d** 显示；用 **php** 写的微信公众号服务器程序，支持查询、订阅和报警；**用 Python 写了机器学习的线性回归和在线学习，来预测第二天的空气质量。**

- iii. 水产品监控及直播系统：

与中科院南京软件研究所合作，使用单片机连接摄像头采集视频信号，使用 **EasyDarwin** 推送视频流，使用拉流技术建立起流媒体管理云平台，从而实现一对多的直播效果，类似于现在的直播软件。

- iv. 黑马大赛：商品类目预测：

题目：只给商品标题和已有分类，测试集中可能需要分到新的类。

方案：使用 **多线程jieba分词** 得到商品标题的分词向量，然后手撸的朴素贝叶斯算法，同时做了一些修改：**当一个词在某个类中出现的频率低于所设阈值时，则将该词在这类的权重置为 0，以此来避免大众词汇对于分类的干扰，提升小样本类别的识别率；并且当分类得分低于阈值时，则新建分类，并更新矩阵。**

- v. 利用视频关键帧预测中间帧：

教研室项目：视频传输时只传关键帧，中间的补帧利用机器学习预测出。具体是将画面分割成许多 **8*8** 的小块，将每一块丢进神经网络计算运动向量，再将结果平滑，得到运动轨迹，从而生成中间帧。最后，提供训练好的模型和供 **c++** 调用的 **python** 的接口给下游。

- vi. 盲人导盲项目：

单板机放置于盲人帽子上，实时语音识别出口令后，调用摄像头拍照，开启控制信令 **socket** 与服务器建立会话，然后建立数据 **socket**，将压缩后的图片发送给服务器(**图片压缩后可以降低图片中杂项的干扰，提高对于主体的识别率**)。服务器使用 **yolo** 进行物体识别，对于主要物体的方位和距离进行估算，将结果以 **json** 格式返回给单板机。单板机接收到识别结果后，语音合成内容，播报出来。

工作经历

华为 数据通信网络协议开发部 2018.7-2020.5

- 主要项目：

- i. 独立设计 Trunk 软件选口算法，使用分层的 avl 管理端口资源，使板级异常的主备切换耗时从原来的几百毫秒，降低至几毫秒。
- ii. 使用 bitmap 设计并完成网络协议的动态唯一标签申请及分配模块。
- iii. 对开源代码 sprintf 进行整改，成功对 va_list 底层异构所导致的异常进行定位，并做出应对方案。
- iv. 成为代码 committer，负责部门代码微重构，及代码review，并指导新员工编程；
- v. 被选拔进入软件学院进修，并成为部门第一个通过软件认证考试的人；
- vi. 日常分享技术总结，并在内网发博客 10 余篇，累计 UV 阅读量 3000 多人。

- 个人项目：

- i. 参加 科赛Kesci 的「二分类算法」提供银行精准营销解决方案 比赛，**取得全球100+名次**，获得一张100美元AWS代金券。

华为 消费者云服务HiCloud开发部 2020.5-2021.9

在浏览器算法团队，负责NLP模型和搜索直达的排序模型。

- 主要项目：

- i. 优质文章模型（浏览器每日精选栏目）：经bert得文章embedding(即文章的语义信息)，再拼接上文章的结构信息（段落、字数、图片数、来源、作者等，经过embedding和标准化），接上双塔网络，判断文章优质与否，以及属于哪个优质类别。
- ii. 层次文章分类模型：根据各层次分类之间天然的关联性，同时学习多个label，设计多任务的 Bert模型（将一级分类网络的最后一级输出，拼上之前bert输出的embedding，再进二级分类网络，以此类推），在学习阶段就可自动进行层次分类校准。并且，在推断时，可以获取各层分类类别的概率，采用beam search，进行有限度的扩展搜索。
- iii. 搜索直达功能的排序模型：打通FTRL模型上线，在产品诞生之初快速赋能；后切换到DCN模型，自动进行特征交叉；现转向ESMM模型，多任务，同时训练 CTR 和 CVR 指标，以期真正提升 CTCVR 业务指标。
- iv. 同时做过 文章地域模型、时效模型、负面文章模型、友商吹捧文章模型等，主管NLP领域的分类模型。

- 个人项目：

- i. 编写股票监控平台。采用docker部署爬虫模块、量化指标算法模块、订阅分发模块等，各模块之间使用rabbitmq和redis，以及mongodb连接。爬虫模块爬取股票数据存入通用数据库接口（mongodb），并缓存至redis；算法模块监控rabbitmq，从而处理数据并将报警信息存入redis；订阅分发模块监控到有报警信息后主动向微信订阅用户推送；同时部署有微信服务器，完成与微信用户的交互。https://www.zhangshengdong.com/post/monitor_stock_system/

华为 NAIE AI模型与训练服务部 2021.9-至今

在用户体验团队，承担算法SE（架构师）角色，负责设计与实现电信领域用户体验相关模型。

- **主要项目：**

- i. 基站信号仿真大模型：输入地形各图层数据及基站位置等参数，输出地图各处基站信号强度。模型结构借鉴cv领域双流网络思想，采用白盒（可微电磁传播公式）和黑盒（大规模卷积网络）双流的思路。
- ii. 大规模表格类数据无监督训练框架：针对电信领域数据基本为表格类无标签数据的特点，结合巧妙地特征工程，设计了以锚点和先验概率分布为约束的无监督训练框架。可直接利用无标签数据进行训练，给出用户的体验打分，同时支持小批量有标签样本（数据不出局点）的在线调优。
- iii. 神经网络关于特征的模型自解释方法：利用模型对于特征的梯度，结合巧妙地特征工程构造出的背景样本和MSE Loss，即可计算出各用户特征输入关于模型打分结果的影响度。

晟腾特战队：

- **主要项目：**

- i. 负责开源项目大模型加速库 <https://gitee.com/ascend/AscendSpeed> 的调优部分，并成为该项目的committer。主要是带领团队进行大模型训练和调优方面的特性开发，并支撑客户训练大模型。多次向Megatron-LM、Megatron-DeepSpeed和PEFT库贡献代码。
- ii. 向PEFT开源社区贡献了一个关键独立特性：分布式LoRA

<https://github.com/huggingface/peft/pull/1092> 已经在peft0.7.0版本发布：

The screenshot shows the GitHub release page for the Hugging Face PEFT library, version 0.7.0. The page title is "v0.7.0: Orthogonal Fine-Tuning, Megatron support, better initialization, safetensors, and more". It indicates that Benjamin Bossan released this version last week, with 17 commits to main since the previous release. The highlights section lists three key features: 1. Orthogonal Fine-Tuning (OFT), a new adapter similar to LoRA, promising better controllability and compositionality for Stable Diffusion. 2. Support for parallel linear LoRA layers using Megatron, which should speed up training with Megatron. 3. LoftQ, a new method to initialize LoRA layers of quantized models to minimize quantization error.

Significant community contributions

The following contributors have made significant changes to the library over the last release:

@zhangsheng377

- Parallel linear Lora by **@zhangsheng377** in [#1092](#)
- Fix parallel linear lora by **@zhangsheng377** in [#1202](#)

<https://github.com/huggingface/peft/releases/tag/v0.7.0>

该分布式LoRA方案已达理论最优，领先于半年之后发表的S-LoRA论文。

- iii. 向Accelerate开源社区贡献了一个关键独立特性：支持nvidia官方megatron的MegatronLMPlugin <https://github.com/huggingface/accelerate/pull/2501> 已经在accelerate0.31.0版本发布：



<https://github.com/huggingface/accelerate/releases/tag/v0.31.0>

主导负责OpenMind生态社区分布式与训练特性的设计与开发工作，成为OpenMind套件的核心差异化竞争力，支持晟腾的社区生态发展。

Accelerate对接megatron特性同步反哺官方开源社区。

- iv. 负责我司行业大模型的微调部分。作为开源蓝军路线的领导人，提出并尝试了模拟退火、dpo、细化二三级子领域能力分类、数据平权采样、基于裁判模型的预料质量评估、重复训练波动分析、数据精细化管理方案等有效实践，并积极分享，主力红军路线成功。
- v. 负责我司自动化评测系统的设计与开发。提出了从nginx的工程设计，到规则与大模型共存的评测方案，以及将rag引入评测系统，以此增强评测系统对于错误多样性的适应性。使用裁判模型作为rm，打通dpo训练。

[pdf版简历](#)

本科时期的旧博客: <http://zhangshengdong29.lofter.com/view>