**陶启萌**

qt2139@columbia.edu **•** [www.linkedin.com/in/qimeng-tao](https://www.linkedin.com/in/qimeng-tao/) **•** 13095523353

**教育**

**哥伦比亚大学 (Columbia University)** New York City

电气工程理学硕士 (M.S. in Electrical Engineering) 2021. 09 - 2023.02

GPA: 3.774/4.0

**华北电力大学科技学院** 保定, 河北省

电气工程及其自动化 2016. 09 - 2020. 06

GPA: 87.18/100; 排名: 5/377

**技能**

**Languages & Frameworks** Python, C, C++, Pytorch, MindSpore, TensorRt, DeepSpeed, Hugging Face, TensorFlow

**Techniques** Multimodal, Large Language Model, Video Understanding, Contrastive Learning, etc.

**Tools** Git, Docker, Linux, AWS (EC2), Google Cloud Platform (GCP)

**工作经历**

**软件开发工程师, 华为** 苏州, 2023. 02 – 至今

* 利用DeepSpeed框架, 采用分布式训练策略, 使用零冗余优化器Zero微调Llama, 训练了ICT第一个大语言模型
* 使用llama.cpp完成模型部署, 推理速度平均提升130%, 显存降低了3倍, 从13GB降至3.9GB
* 通过MindSpore训练电信领域文本二分类模型, 并用TensorRT && MindSpore Lite C++ 部署, 推理速度提升3倍
* 构建文本向量数据库, 利用faiss达到亿级数据毫秒级别的检索时间, 各个用户的向量数据库支持在线增删改查, 结合检索答案优化prompt后使大模型性能提升158%

**计算机视觉工程师, Visla** Palo Alto, USA, 2022. 05 - 2022. 09

* 部署OFA / SwinBert 模型以实现Image Captioning / Video Captioning, 在公司数据集中, BLEU4超过44.2, Cross-Encoder超过0.75, 与之前的模型相比, 模型性能提高了 10.6%
* 搭建RetinaFace和FaceNet实现视频人脸识别, 并检测视频中关键人物出现的时间区间, 在压力测试中, 时间区间的准确率为80.3%, 模型通过TensorRT加速推理, 速度从30 FPS / s提升至90 FPS / s
* 使用YOLOv7和GLIP 检测视频中经常出现的物体, 并检测物体在视频中出现的时间区间, 在公司数据集中, mAP 为 51.4%, 时间区间的准确率为 97.7%, 并用TensorRT部署, 速度从90 FPS / s提升至180 FPS / s

**项目经历**

**基于SMOKE 3D目标检测** 2022. 04 - 2022. 05

* 模拟生成多种车牌，使用DeepStream链接车牌检测和识别模型
* 通过TensorRT部署, 使用C++重写, INT8量化, 速度提升至20 FPS / s
* 使用CUDA重写预处理部分并改为多线程处理, 速度提升至30 FPS / s

**基于EfficientNet 和EfficientNetV2在花朵数据集上的图像分类** 2020.12 - 2021. 01

* 使用Confident Learning (Cleanlab)寻找错误数据, 因为数据符合长尾分布, 所以损失函数使用EQLv2Loss, 改变后性能提升5.7% [Github](https://github.com/qt2139/Rethinking-EfficientNet-and-EfficientNetV2)
* 利用progressive learning, 根据图像大小动态调整正则化方法
* 将浅层网络(Stage1-Stage3)的MBConv 修改为Fused - MBConv, 并使用更小的expansion ratio使浅层网络更强大, 与EfficientNet 相比, 训练时间减少了11 倍