

推理系统系列

推理系统内容简介



ZOMI



BUILDING A BETTER CONNECTED WORLD

Ascend & MindSpore

www.hiascend.com
www.mindspore.cn

Talk Overview

1. 推理系统介绍

- 推理系统与推理引擎区别
- 推理工作流程
- 推理系统介绍
- 推理引擎介绍

2. 模型小型化

- NAS神经网络搜索
- CNN小型化结构
- Transform小型化结构

3. 离线优化压缩

- 低比特量化
- 二值化网络
- 模型剪枝
- 模型蒸馏

4. 部署和运行优化

- 图转换优化（算子融合/重排/替换）
- 并发执行与内存分配
- 动态batch与bin Packing

Talk Overview

I. 推理系统与推理引擎

- Training and Inference – 训练和推理服务的区别
- What is inference system - 什么是推理系统
- Optimization objectives and constraints - 推理系统优化目标与约束
- Difference bet inference system and engine - 推理系统与推理引擎

训练和推理 服务



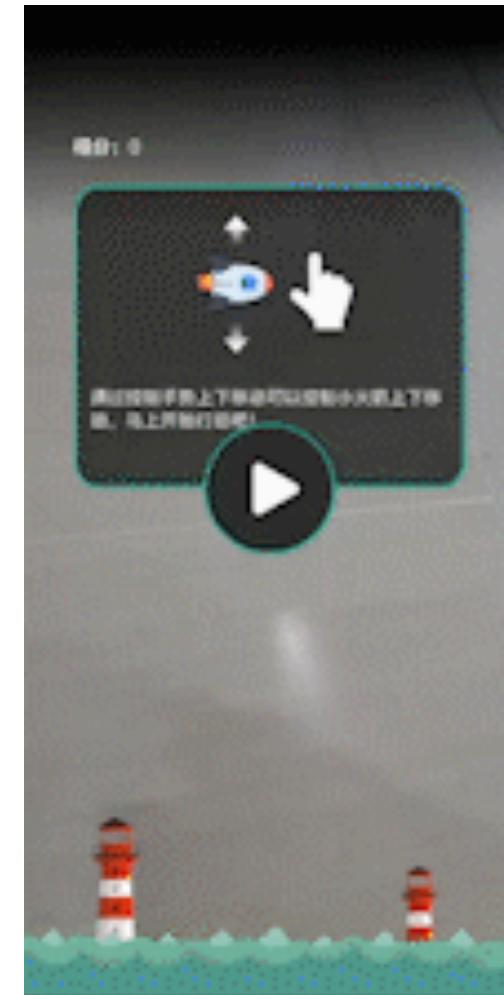
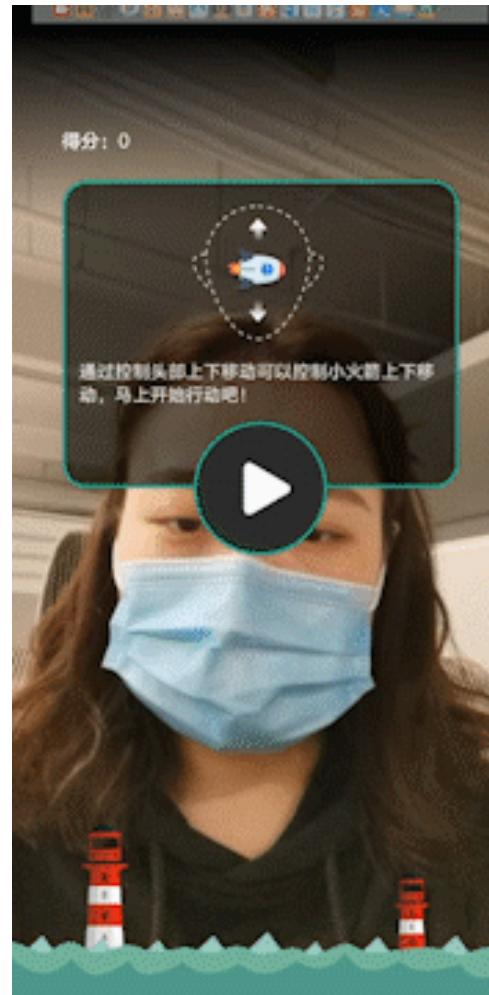
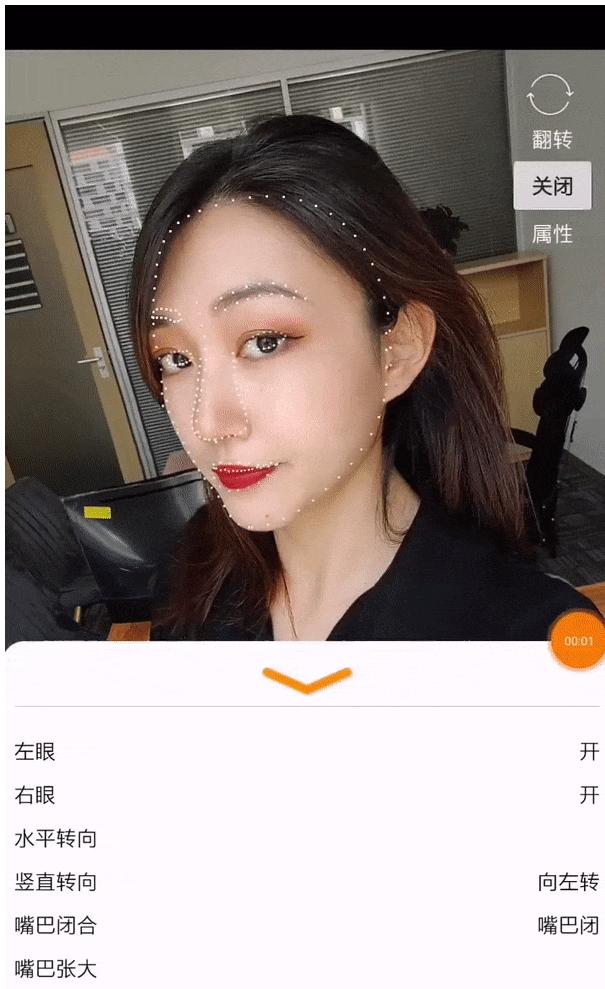
BUILDING A BETTER CONNECTED WORLD

Ascend & MindSpore

4

www.hiascend.com
www.mindspore.cn

典型深度学习推理应用



典型深度学习推理应用

对话机器人

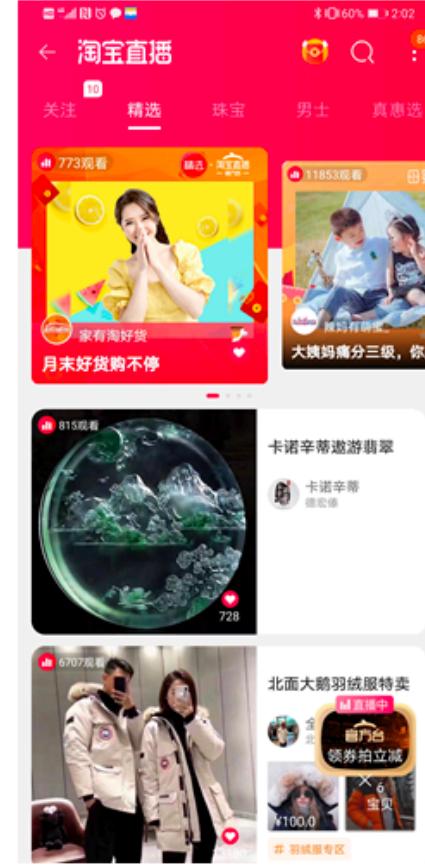
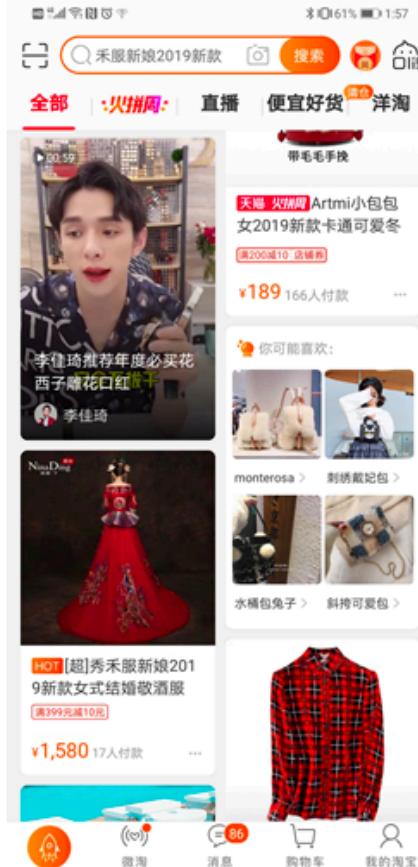
(e.g., Microsoft Xiao Ice, etc.)



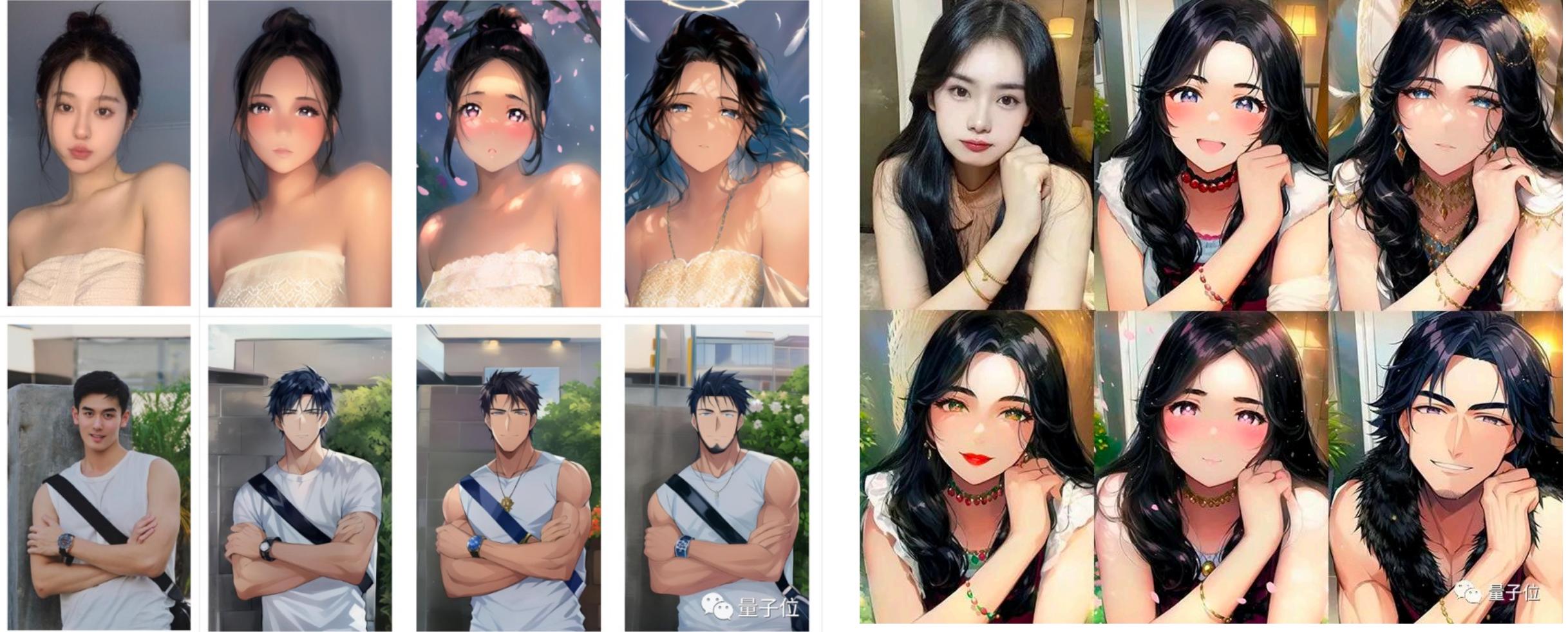
典型深度学习推理应用

推荐系统

(e.g., Bing News, etc.)



典型深度学习推理应用



典型深度学习推理应用

在这个过程中，推理系统需要考虑和提供以下的功能：

- 提供可以被用户调用的接口
- 能够完成一定的数据处理将输入数据转为向量
- 能够在指定低延迟要求下返回用户响应
- 能够利用多样的加速器进行一定的加速
- 能够随着用户的增长保持高吞吐的服务响应和动态进行扩容
- 能够可靠的提供服务，应对软硬件的失效
- 能够支持算法工程师不断更新迭代模型，应对不断变化的新框架



典型深度学习推理应用

- 单纯复用原有的 Web 服务器或者移动端应用软件，只能解决其中一部分问题，在深度学习模型推理的场景下，产生了新的系统设计需求与挑战。
- 从深度学习训练过程和推理过程对比两者的相同点和不同点，以及在生命周期所处的环节，进而便于理解深度学习推理系统所侧重的目标。

参考文献

1. [Deep Learning Inference in Facebook Data Centers: Characterization, Performance Optimizations and Hardware Implications](#)
2. [Clipper: A Low-Latency Online Prediction Serving System](#)
3. [TFX: A TensorFlow-Based Production-Scale Machine Learning Platform](#)
4. [TensorFlow-Serving: Flexible, High-Performance ML Serving](#)
5. [Optimal Aggregation Policy for Reducing Tail Latency of Web Search](#)
6. [A Survey of Model Compression and Acceleration for Deep Neural Networks](#)
7. [CSE 599W: System for ML - Model Serving](#)
8. <https://developer.nvidia.com/deep-learning-performance-training-inference>
9. [DEEP COMPRESSION: COMPRESSING DEEP NEURAL NETWORKS WITH PRUNING, TRAINED QUANTIZATION AND HUFFMAN CODING](#)
10. [Learning both Weights and Connections for Efficient Neural Networks](#)
11. [DEEP LEARNING DEPLOYMENT WITH NVIDIA TENSORRT](#)
12. [Halide: A Language and Compiler for Optimizing Parallelism,Locality, and Recomputation in Image Processing Pipelines](#)
13. [TVM: An Automated End-to-End Optimizing Compiler for Deep Learning](#)
14. [8-bit Inference with TensorRT](#)
15. <https://github.com/microsoft/AI-System>



BUILDING A BETTER CONNECTED WORLD

THANK YOU

Copyright©2014 Huawei Technologies Co., Ltd. All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.