

推理系统-模型小型化

Transformer 小型化



ZOMI



BUILDING A BETTER CONNECTED WORLD

Ascend

www.hiascend.com

Talk Overview

1. 推理系统介绍

- 推理系统与推理引擎
- 推理系统的工作流程
- 推理系统生命周期管理

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

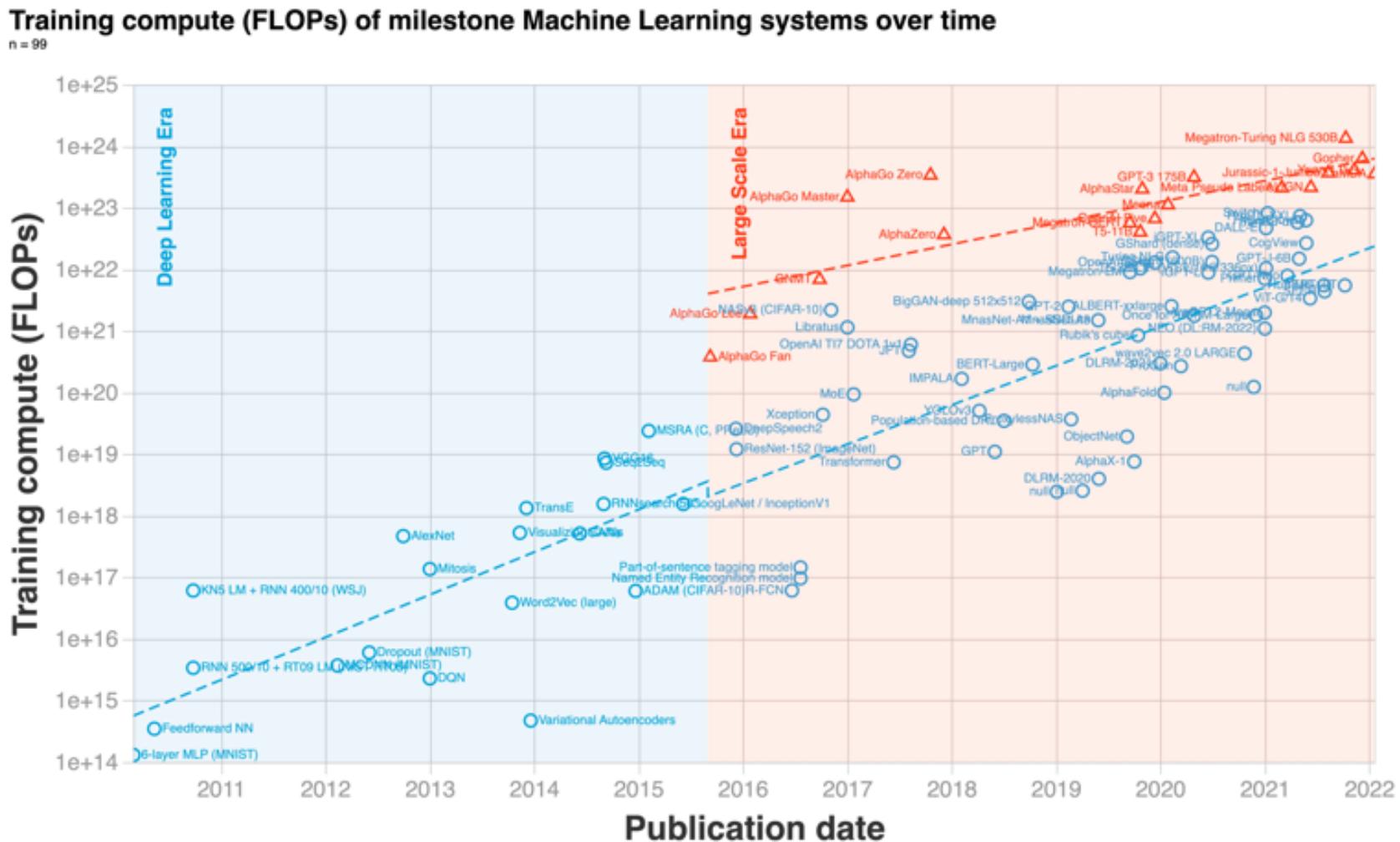
3. 离线优化压缩

- 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

深度学习模型发展



轻量级模型

1. MobileViT (2021)
2. Mobile-Former (2021)
3. EfficientFormer (2022)

轻量化网络总结

如何选择轻量化网络：

1. 不同网络架构，即使 FLOPs 相同，但其 MAC 也可能差异巨大
2. FLOPs 低不等于 latency 低，结合具硬件架构具体分析
3. 多数时候加速芯片算力的瓶颈在于访存带宽
4. 不同硬件平台部署轻量级模型需要根据具体业务选择对应指标



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.