

# 推理系统-模型小型化

# CNN 小型化



# ZOMI



BUILDING A BETTER CONNECTED WORLD

Ascend

[www.hiascend.com](http://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

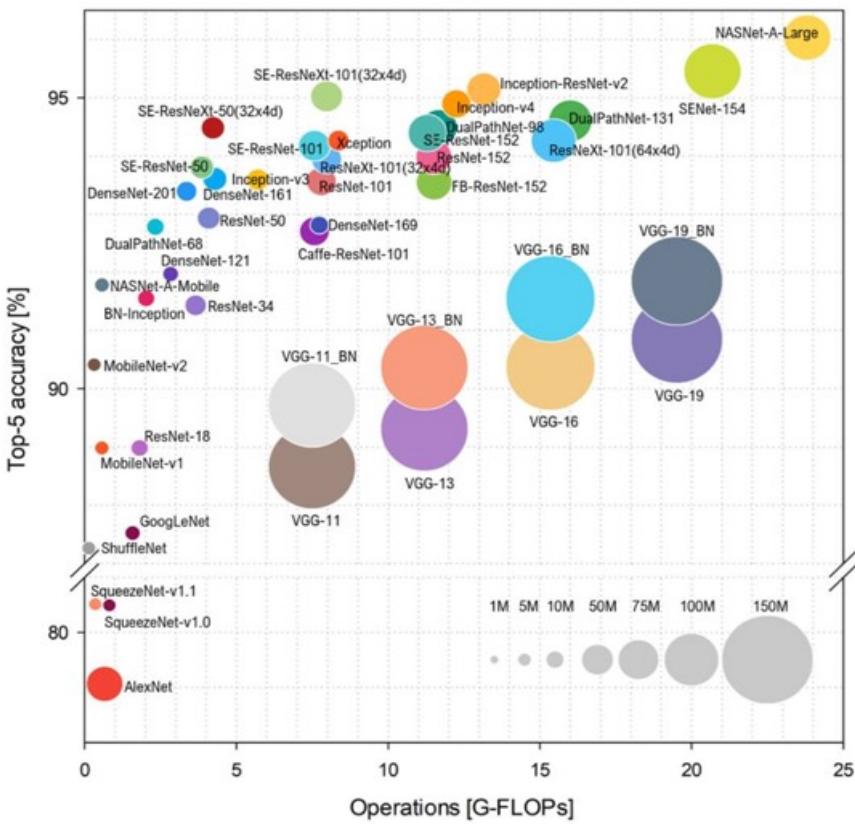
## 3. 离线优化压缩

- 低比特量化
- 二值化网络
- 模型剪枝
- 模型蒸馏

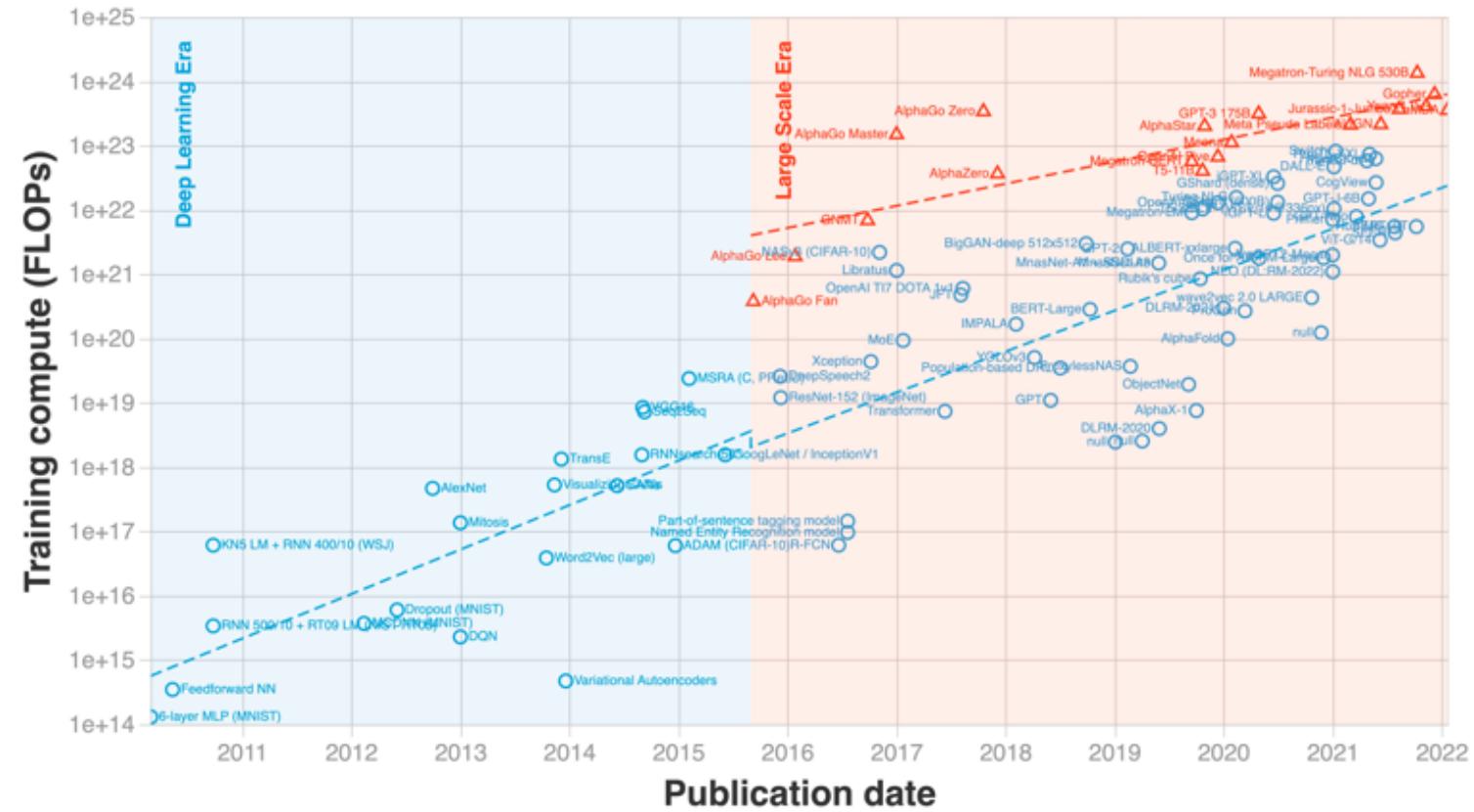
## 4. 部署和运行优化

- 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

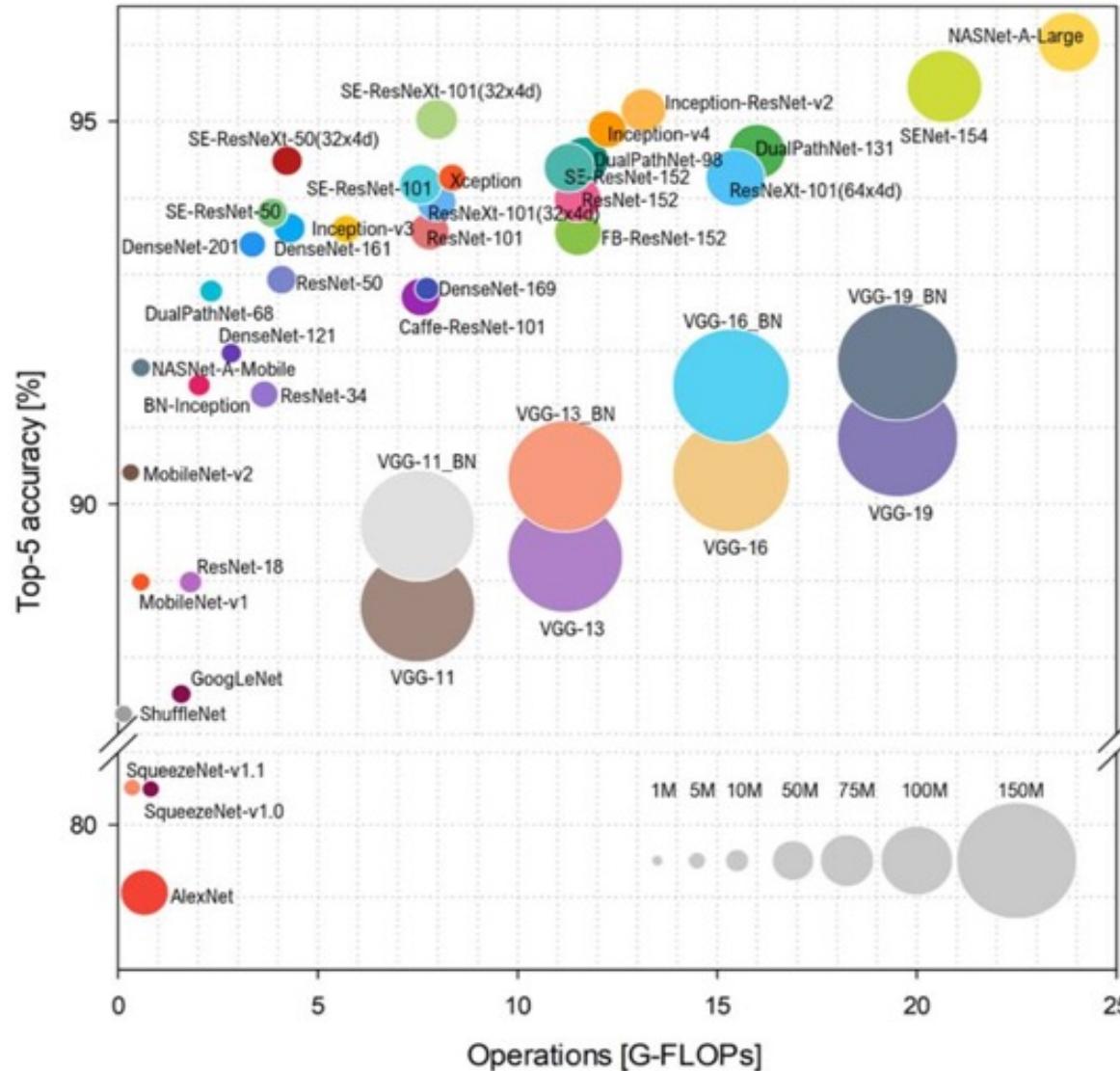
# 深度学习模型发展



Training compute (FLOPs) of milestone Machine Learning systems over time  
n = 99



# 深度学习模型发展



# 轻量级模型

1. SqueezeNet 系列 ( 2016 )
2. ShuffleNet 系列 ( 2017 )
3. MobileNet 系列 ( 2017 )
4. ESPnet 系列 ( 2018 )
5. FBNet 系列 ( 2018 )
6. EfficientNet 系列 ( 2019 )
7. GhostNet 系列 ( 2019 )

# CNN轻量化网络总结

## 卷积核方面：

1. 大卷积核用多个小卷积核代替
2. 单一尺寸卷积核用多尺寸卷积核代替
3. 固定形状卷积核趋于使用可变形卷积核
4. 使用 $1 \times 1$ 卷积核 - bottleneck结构

## 卷积层通道方面：

1. 标准卷积用depthwise卷积代替
2. 使用分组卷积
3. 分组卷积前使用 channel shuffle
4. 通道加权计算

# CNN轻量化网络总结

## 卷积层连接方面：

1. 使用skip connection，让模型更深
2. densely connection，融合其它层特征输出



# 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.