推理系统-模型小型化

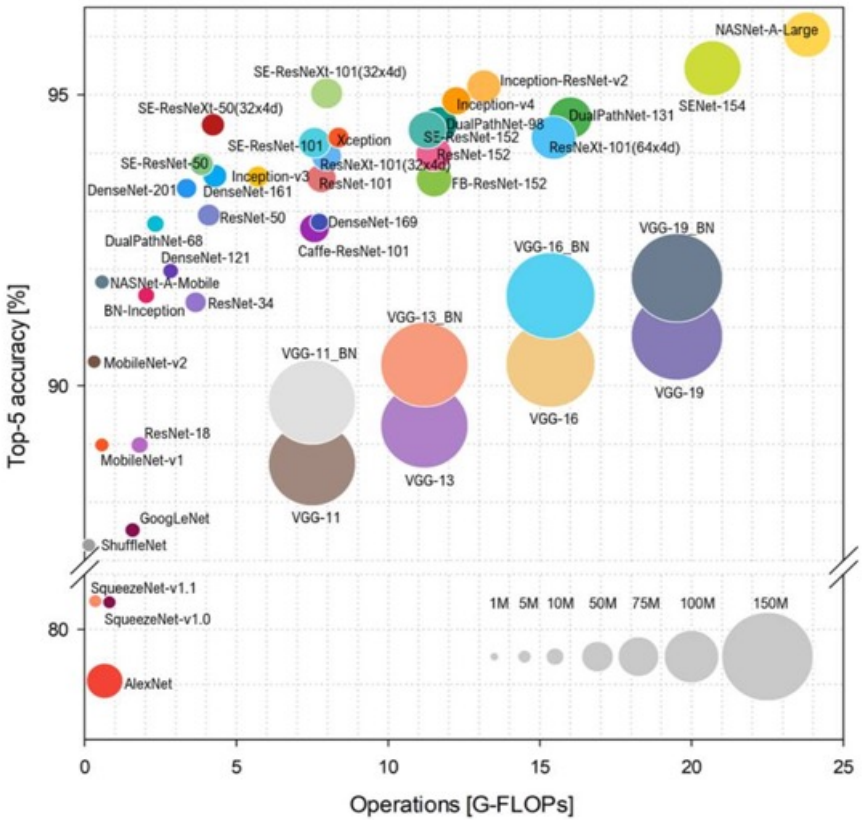
CNN 小型化

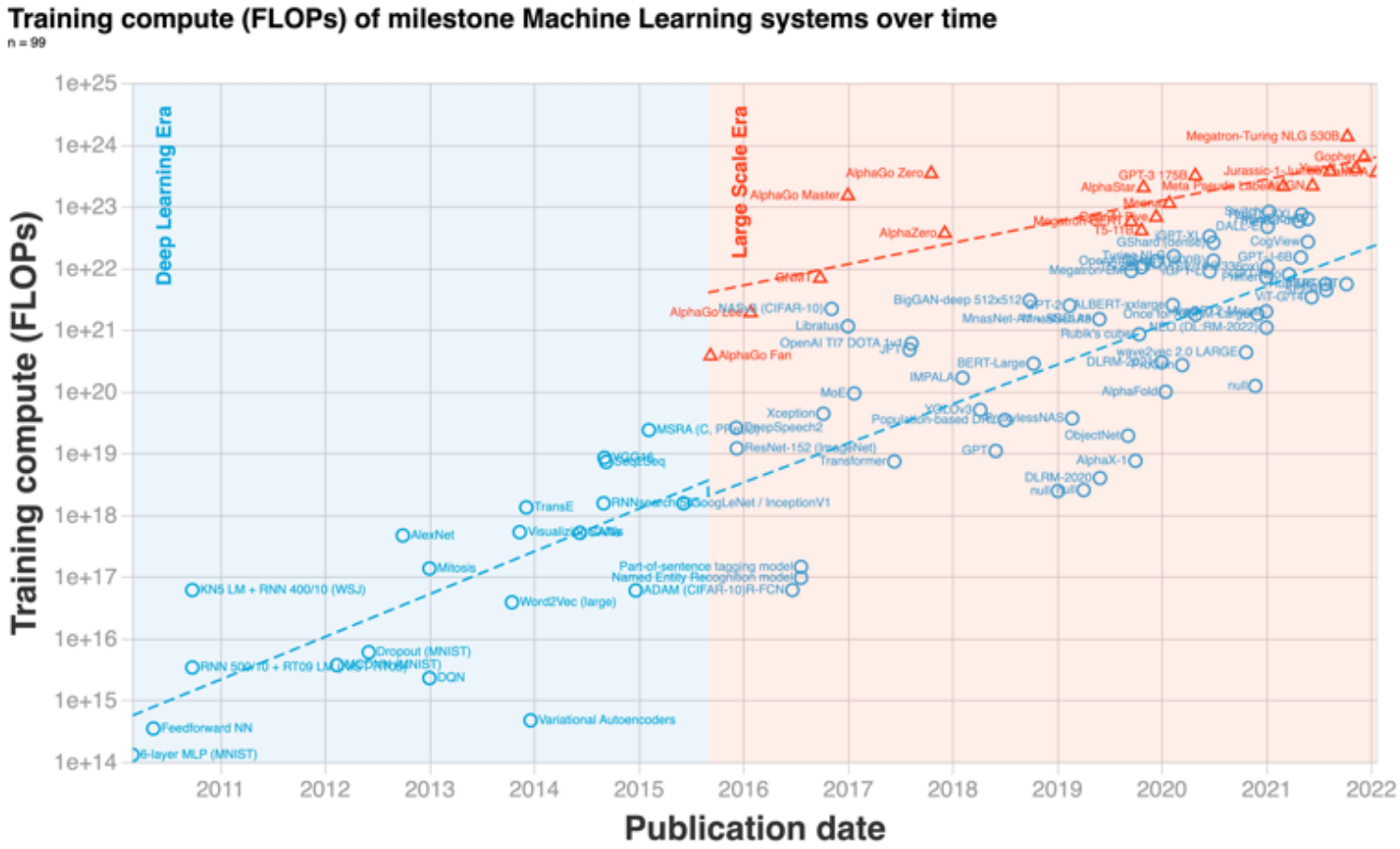
ZOMI

Talk Overview

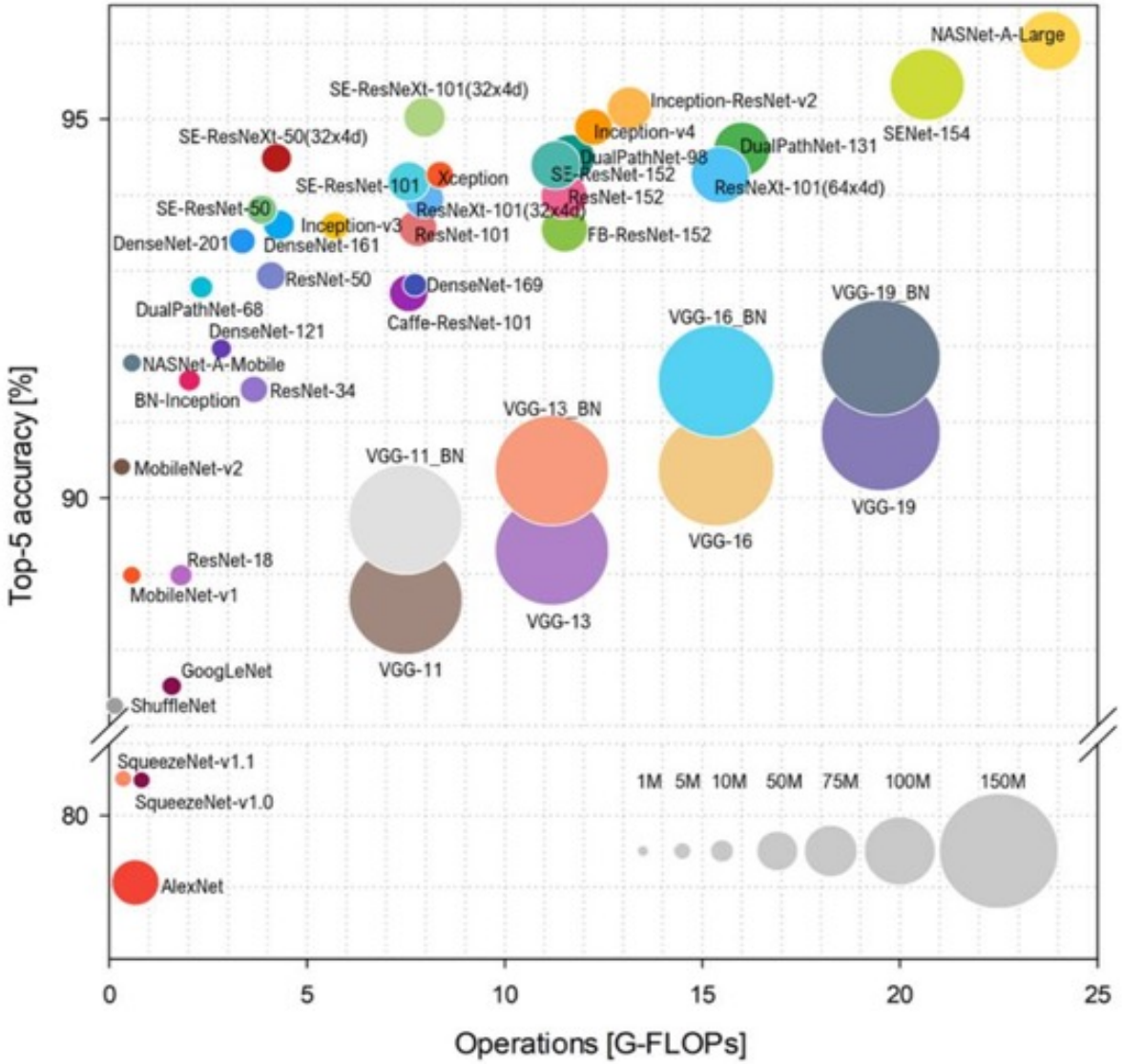
1. **推理系统介绍**
   - 推理系统与推理引擎区别
   - 推理工作流程
   - 推理系统介绍
   - 推理引擎介绍
2. **模型小型化**
   - 基础参数概念
   - CNN小型化结构
   - Transform小型化结构
3. **离线优化压缩**
   - 低比特量化
   - 二值化网络
   - 模型模型剪枝
   - 模型模型蒸馏
4. **部署和运行优化**
   - 图转换优化（算子融合/重排/替换）
   - 并发执行与内存分配
   - 动态batch与bin Packing

深度学习模型发展


NASNet-A-Large
SE-ResNeXt-101(32x4d)
Inception-ResNet-v2
SE-ResNeXt-50(32x4d)
Inception-v4
SENet-154
DualPathNet-131
DualPathNet-98
SE-ResNet-101
Xception
SE-ResNet-152
ResNeXt-101(64x4d)
SE-ResNet-50
ResNeXt-101(32x4d)
Inception-v3
ResNet-101
FB-ResNet-152
DenseNet-201
DenseNet-161
ResNet-50
DenseNet-169
VGG-19_BN
DualPathNet-68
Caffe-ResNet-101
VGG-16_BN
DenseNet-121
NASNet-A-Mobile
BN-Inception
ResNet-34
VGG-13_BN
VGG-11_BN
MobileNet-v2
VGG-19
VGG-16
ResNet-18
MobileNet-v1
VGG-13
VGG-11
GoogLeNet
ShuffleNet
SqueezeNet-v1.1
SqueezeNet-v1.0
AlexNet
1M 5M 10M 50M 75M 100M 150M
Top-5 accuracy [%]
Operations [G-FLOPs]


Training compute (FLOPs) of milestone Machine Learning systems over time
n = 99
Deep Learning Era
Large Scale Era
Training compute (FLOPs)
Publication date
AlexNet
Dropout (MNIST)
DQN
Feedforward NN
Variational Autoencoders
Word2Vec (large)
ADAM (CIFAR-10)
Xception
MoE
BERT-Large
GPT
AlphaGo Zero
AlphaGo Master
AlphaZero
AlphaGo Fan
Megatron-Turing NLG 530B
GPT-3 175B
AlphaFold

深度学习模型发展


NASNet-A-Large
SE-ResNeXt-101(32x4d)
Inception-ResNet-v2
SENet-154
SE-ResNeXt-50(32x4d)
Inception-v4
DualPathNet-131
DualPathNet-98
SE-ResNet-101
Xception
SE-ResNet-152
ResNeXt-101(64x4d)
ResNet-152
SE-ResNet-50
ResNeXt-101(32x4d)
Inception-v3
ResNet-101
FB-ResNet-152
DenseNet-201
DenseNet-161
ResNet-50
DenseNet-169
VGG-19_BN
DualPathNet-68
Caffe-ResNet-101
VGG-16_BN
DenseNet-121
NASNet-A-Mobile
BN-Inception
ResNet-34
VGG-13_BN
VGG-11_BN
MobileNet-v2
VGG-19
VGG-16
ResNet-18
MobileNet-v1
VGG-13
VGG-11
GoogLeNet
ShuffleNet
SqueezeNet-v1.1
SqueezeNet-v1.0
1M 5M 10M 50M 75M 100M 150M
AlexNet
Top-5 accuracy [%]
95
90
80
0
5
10
15
20
25
Operations [G-FLOPs]

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1×1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. 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