Compressing Deep Networks

James Kwok

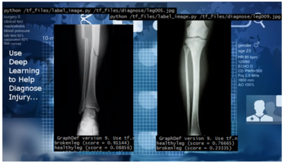
DeepLearn 2019

香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

Machine Learning is Everywhere

Self-Driving

Machine Translation



```
python /tf_files/label_image.py /tf_files/diagnose/leg005.jpg
 GraphDef version 9. Use tf.n
brokenleg (score = 0.91144)
healthyleg (score = 0.08856)
```

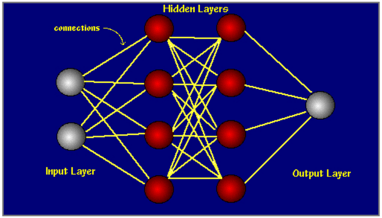
```
python /tf_files/label_image.py /tf_files/diagnose/leg009.jpg
 GraphDef version 9. Use tf.n
healthyleg (score = 0.76665)
brokenleg (score = 0.23335)
```

Healthcare

Game Playing

deep learning: excellent performance in a variety of domains

Deep Learning (Neural Networks)

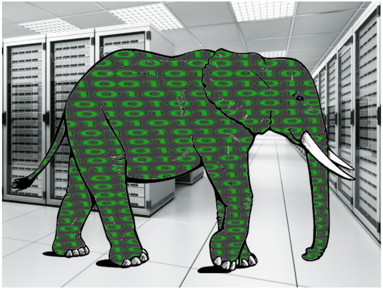


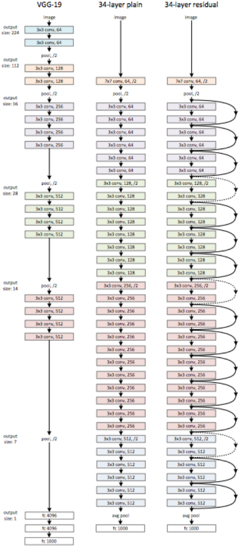
Deeper and Deeper Networks

ImageNet classification

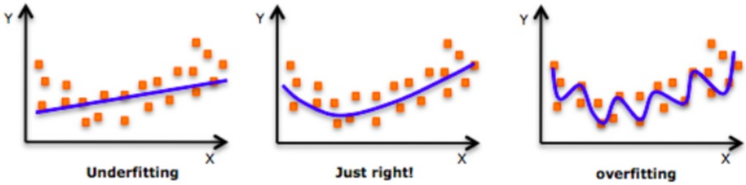
	number of layers	top-5 error (%)
ILSVRC'12 (AlexNet)	8	16.4
ILSVRC'13	8	11.7
ILSVRC'14 (VGG)	19	7.3
ILSVRC'14 (GoogleNet)	22	6.7
ILSVRC'15 (ResNet)	152	3.57

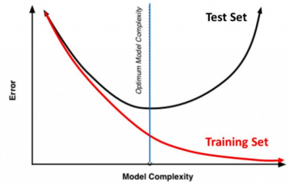
Deep Learning + Big Data + Big Compute



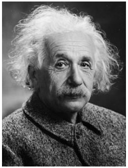


Overfitting





Quest for a Small Model



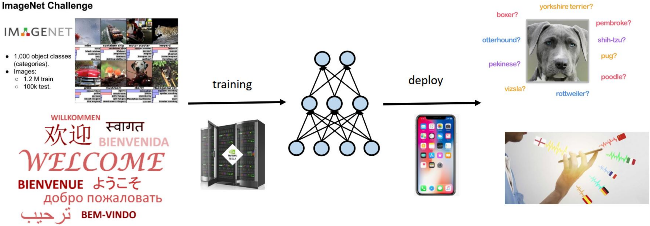
Everything should be made as simple as possible, but no simpler

Occam's razor:
The simplest solution tends to be the right one

Advantages of a small machine learning model

- better generalization
- smaller memory footprint
- faster prediction
- less expensive to collect features

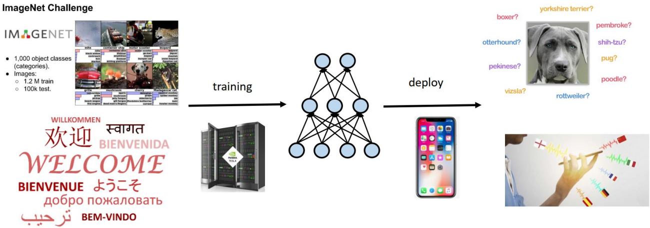
Deep Learning: From Development to Deployment



Example (AlexNet, VGG-16, Resnet)

- hundred of megabytes to store
- billions of high-precision operations on classification
- more operations → more energy

From Development to Deployment

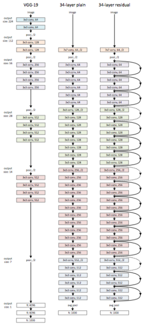


Problem

computation and memory intensive on small computing devices

- cell phones, self-driving cars, internet of things (IoT) devices

Good News!

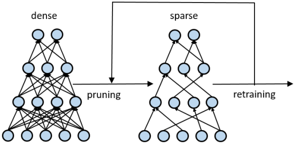


capacity of deep network is usually
larger than needed

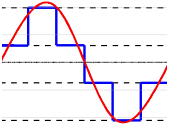
can be compressed without accuracy degradation

Compressing Deep Networks

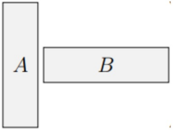
- network sparsification



- quantization

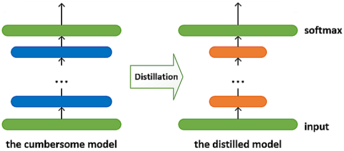


- low-rank approximation

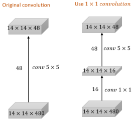


Compressing Deep Networks...

- distillation



- more compact model



- neural architecture search