MOBILE

DEEP BUT SMALL

martin @ reddragon.ai
sam @ reddragon.ai

15 November 2017

`WiFi : SG-Guest`

Problems with Installation? **ASK!**

RED DRAGON AI

PLAN OF ACTION

TODAY

- Mobile : General & Specific
- Personal Project work

PLAN OF ACTION

27-NOV

- Reinforcement Learning(?)
- Finalize Projects

PERSONAL PROJECTS

- Form to fill in
- Plan to finish : **27**-Nov (last session)
- WSG DEADLINE : 30-Nov, including write-ups :
  - "Punchy Headline"
  - Minimum : `README.md` on GitHub page
  - Hosted slides / demo
  - Lightning Talk

SGINNOVATE

SGINNOVATE

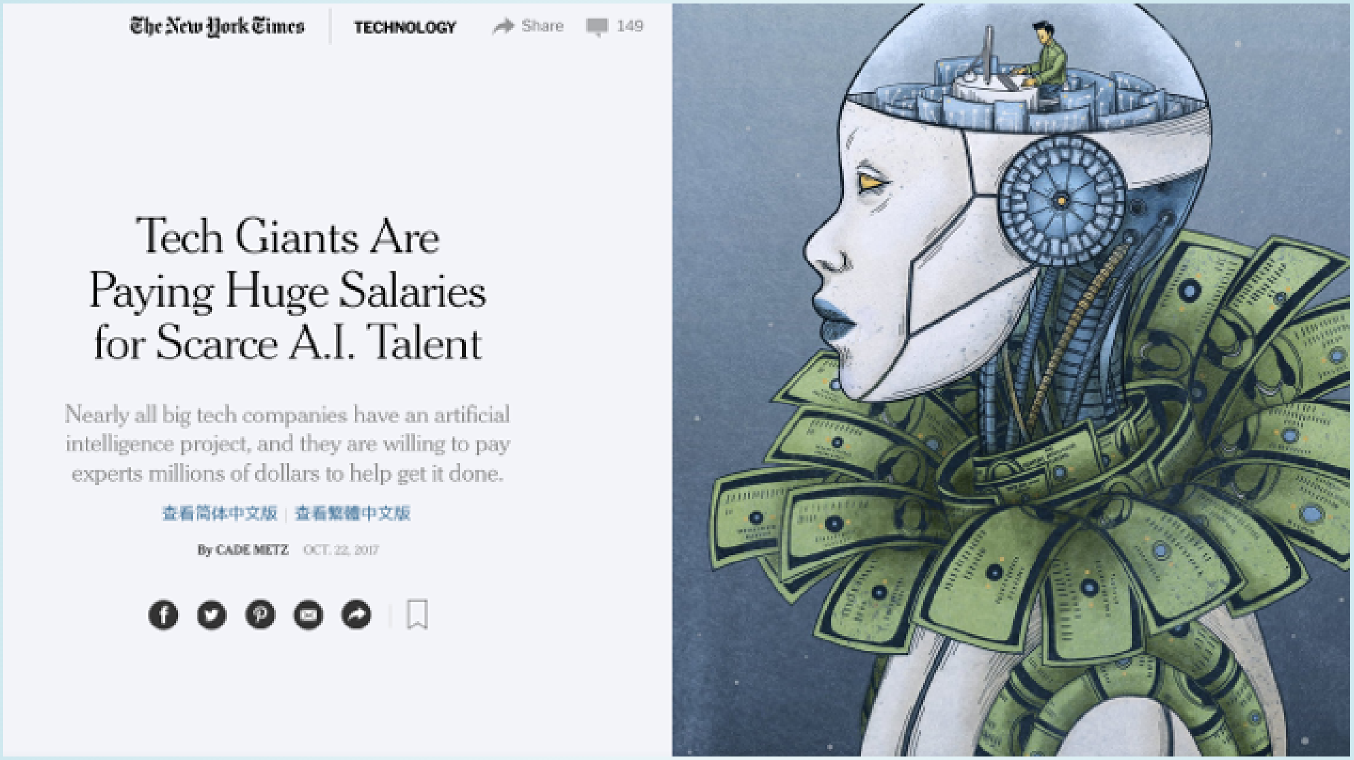
Looking for opportunities?
We find and match you to jobs based on your skills and interests.

+65 | PHONE NUMBER | APPLY

http://bit.ly/2ialg60

nicolette @ sginnovate.com

NEW YORK TIMES


The New York Times | TECHNOLOGY Share 149

Tech Giants Are Paying Huge Salaries for Scarce A.I. Talent

Nearly all big tech companies have an artificial intelligence project, and they are willing to pay experts millions of dollars to help get it done.

查看简体中文版 | 查看繁體中文版

By CADE METZ OCT. 22, 2017

Article Link

RED DRAGON AI

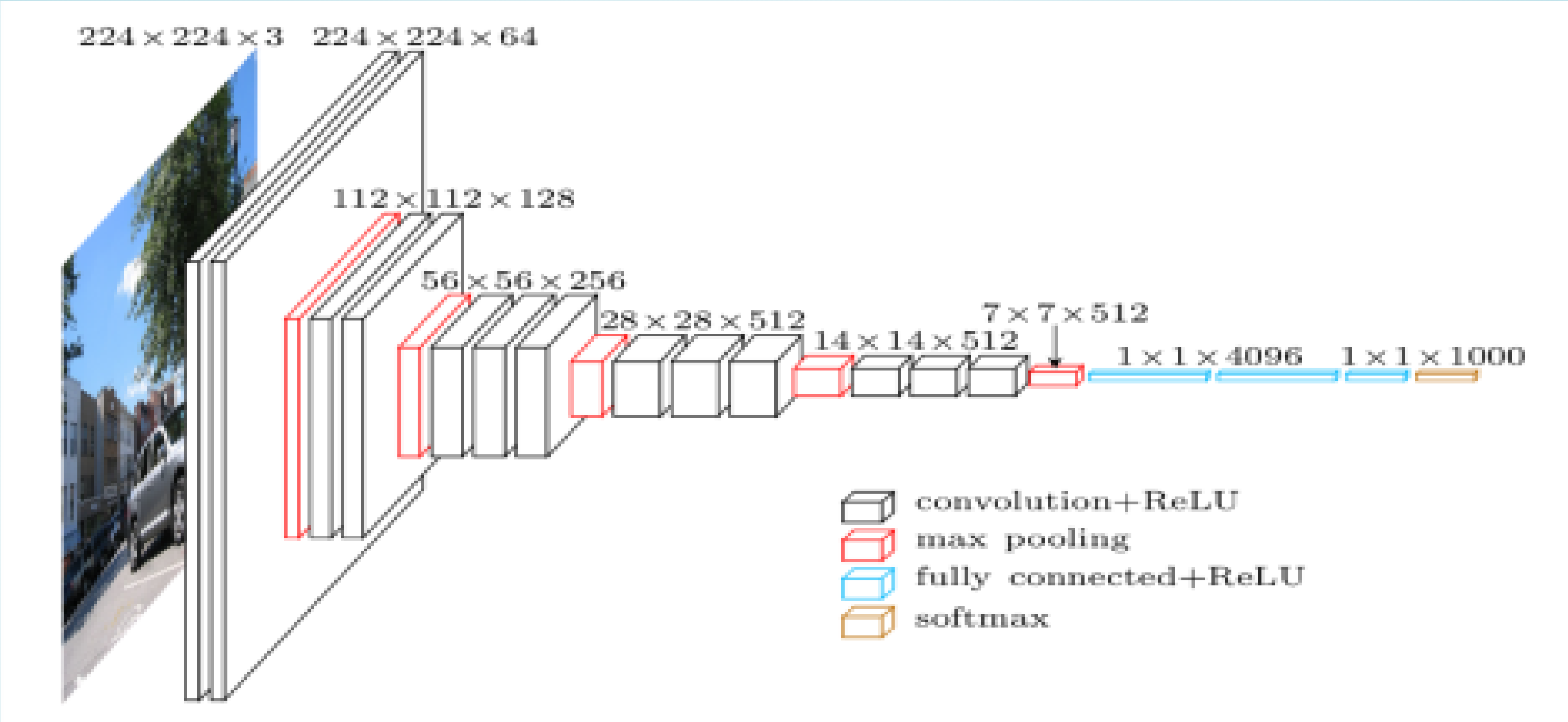
IMAGE COMPETITION

- ImageNet aka ILSVRC
- over 15 million labeled high-resolution images...

  ... in over 22,000 categories

RED DRAGON AI

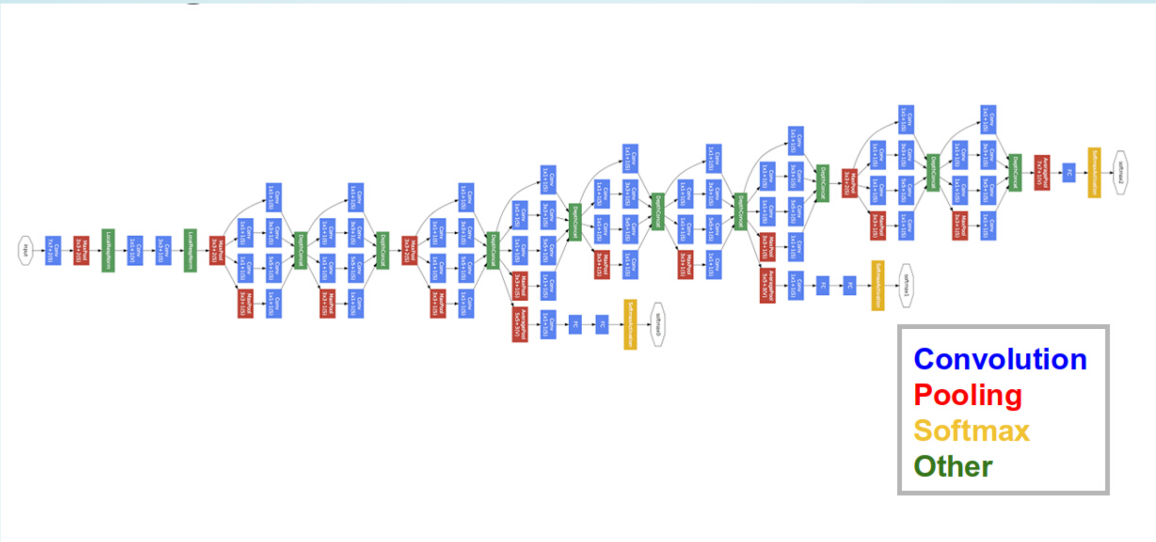
NETWORKS HAVE EVOLVED



VGG 16 (2014)

RED DRAGON AI

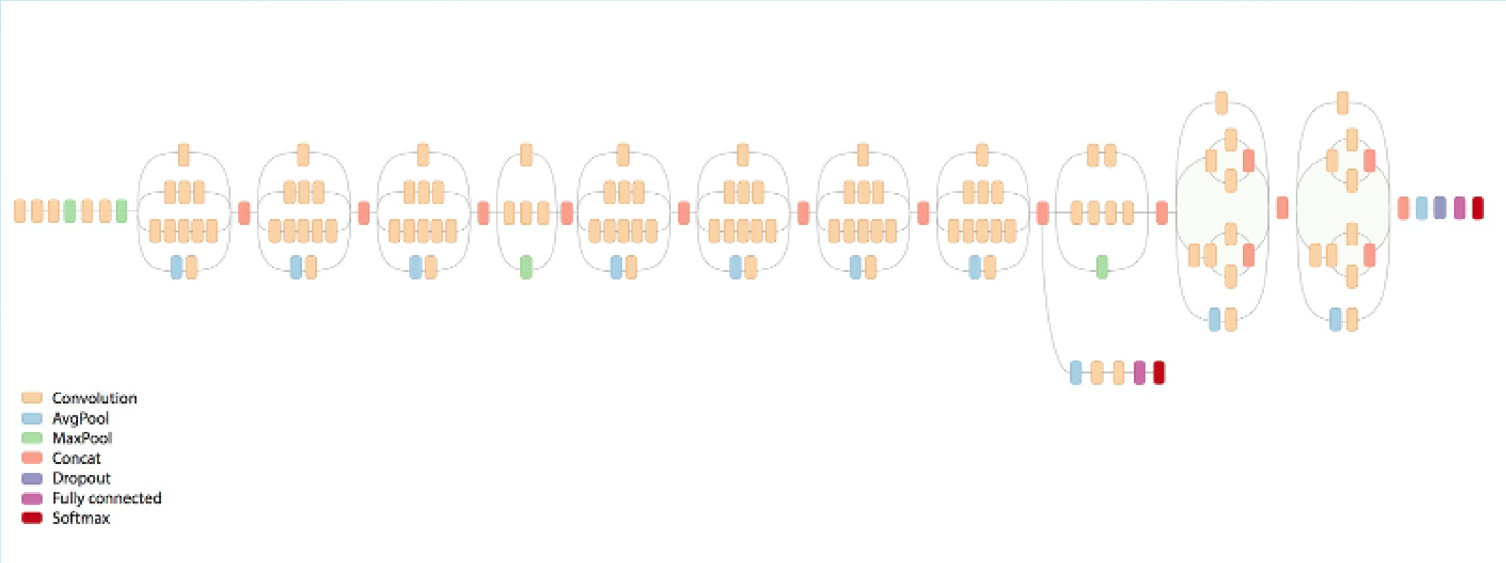
GETTING DEEPER ...



GoogLeNet (2014)

RED DRAGON AI

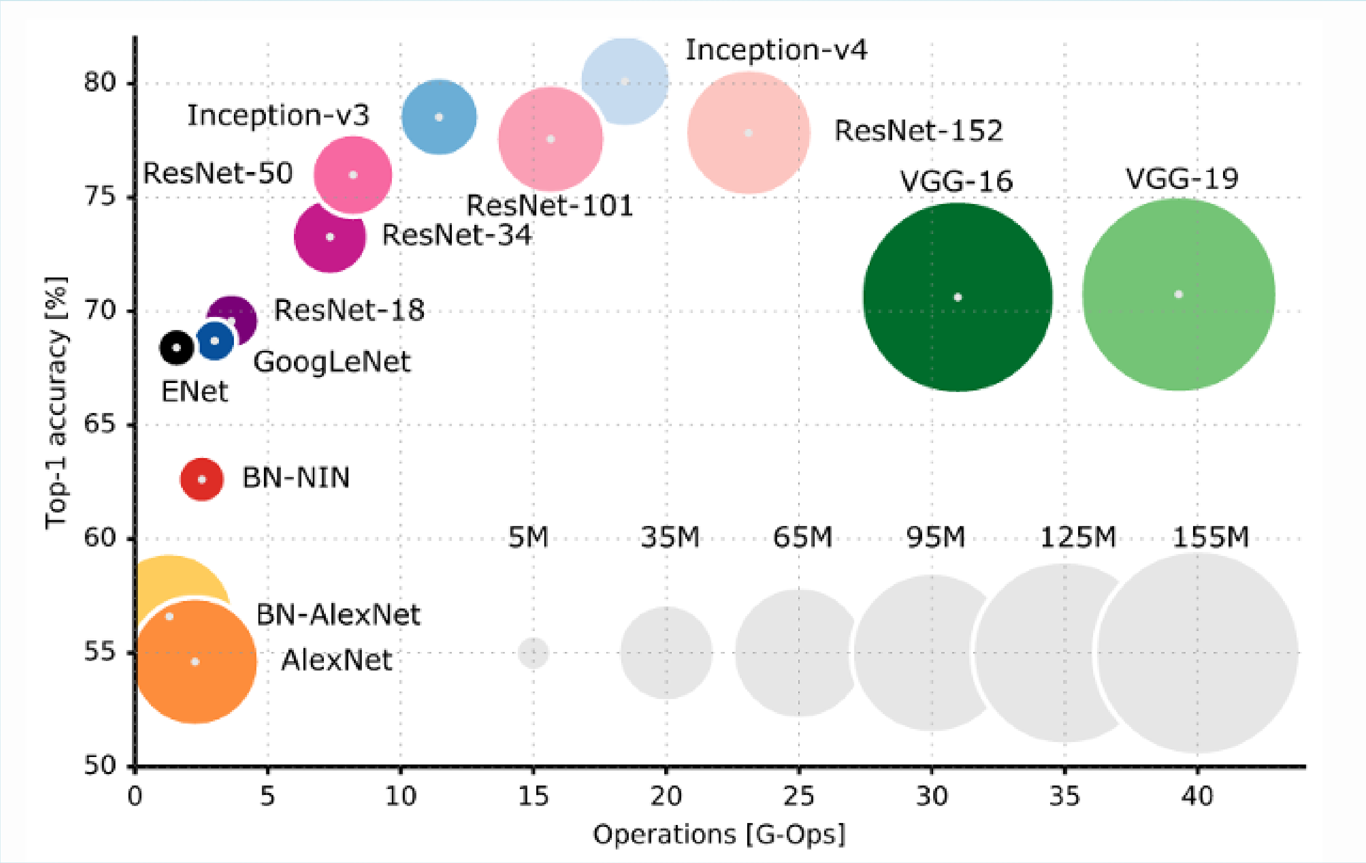
... EVEN MORE COMPLEX



Google Inception-v3 (2015)

RED DRAGON AI

MODEL SIZE / PERFORMANCE



github.com/mdda/deep-learning-workshop/notebooks/2-CNN/4-ImageNet/0-modelzoo-tf-keras.ipynb

RED DRAGON AI

BUT WHAT ABOUT MOBILE?

- Better performance ⇒ larger network
- But mobile needs us to :
  - Compress
  - Downgrade
  - Restructure

ENERGY USAGE

Table 1: Energy table for 45nm CMOS process [20]. DRAM access uses three orders of magnitude more energy than simple arithmetic and 128x more than SRAM.

Operation	Energy [pJ]	Relative Cost
32 bit int ADD	0.1	1
32 bit float ADD	0.9	9
32 bit Register File	1	10
32 bit int MULT	3.1	31
32 bit float MULT	3.7	37
32 bit 32KB SRAM	5	50
32 bit DRAM	**640**	**6400**

EIE: Efficient Inference Engine on Compressed Deep Neural Network (ISCA'16)

COMPRESS / DOWNGRADE

- Sparsity
- High precisions are not required
- Quantisation
- Compressibility

SPARSITY

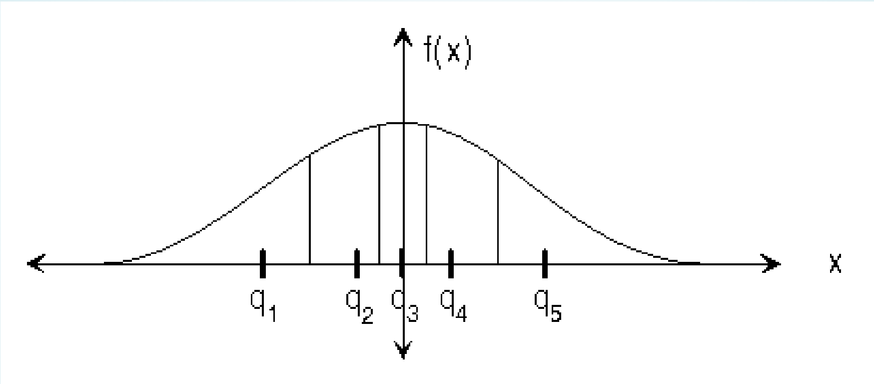
- Weights near zero → Zero
- Clamp these weights during training

LOW-PRECISION TRAINING

- Quantise weights in forward pass
- Use 'full resolution' derivative to do backprop
- 6-bits per parameter seems to work

QUANTISATION

- Bucket weights into a few levels
- Store bucket positions and bucket indexes
- eg: 4 buckets (2 bits per weight index)

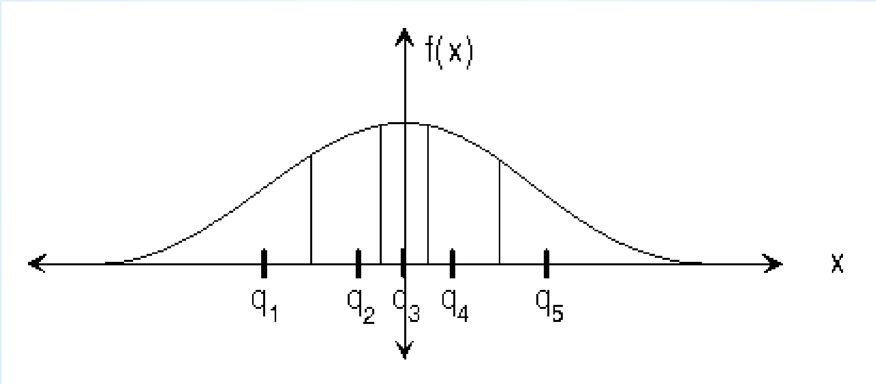


Quantisation in TensorFlow

RED DRAGON AI

COMPRESSIBILITY

- Bucket weights into a few levels
- Store bucket positions and bucket indexes
- But make bucket positions ZIPable



RED DRAGON AI

RESTRUCTURE 5X5

- Normal CNN '5x5' layer operation :
  - Has `(5x5+1)` parameters (per input channel)
- Convert to 2 stacked 3x3 layers :
  - Has 2x`(3x3+1)` parameters (per input channel)
- Parameter count : `26x50=1300` → `20x50=1000`

RED DRAGON AI

RESTRUCTURE 3X3

- Normal '3x3' CNN layer operation :
  - For each output channel:
    - Run separate '3x3' kernels over all input channels
    - Allows anywhere-to-anywhere interactions
- Parameter count : `(3x3+1)x50 = 500`

RED DRAGON AI

SEPARABLE CONVOLUTIONS

- Separable CNN layer operation :
  - For each output channel:
    - Run one '3x3' kernel over all input channels
    - Do a weight sum (a '1x1' convolution) over results
    - Separate *texture* vs *layer* operations
- Parameter count : `(3x3+1)x1 + 1x1x(50+1) = 61`

+VARIATIONS

- Need to be careful that factorisation doesn't destroy performance
- Lots of scope for experimentation :
  - Xception
  - Depthwise Separable Convolutions for Neural Machine Translation

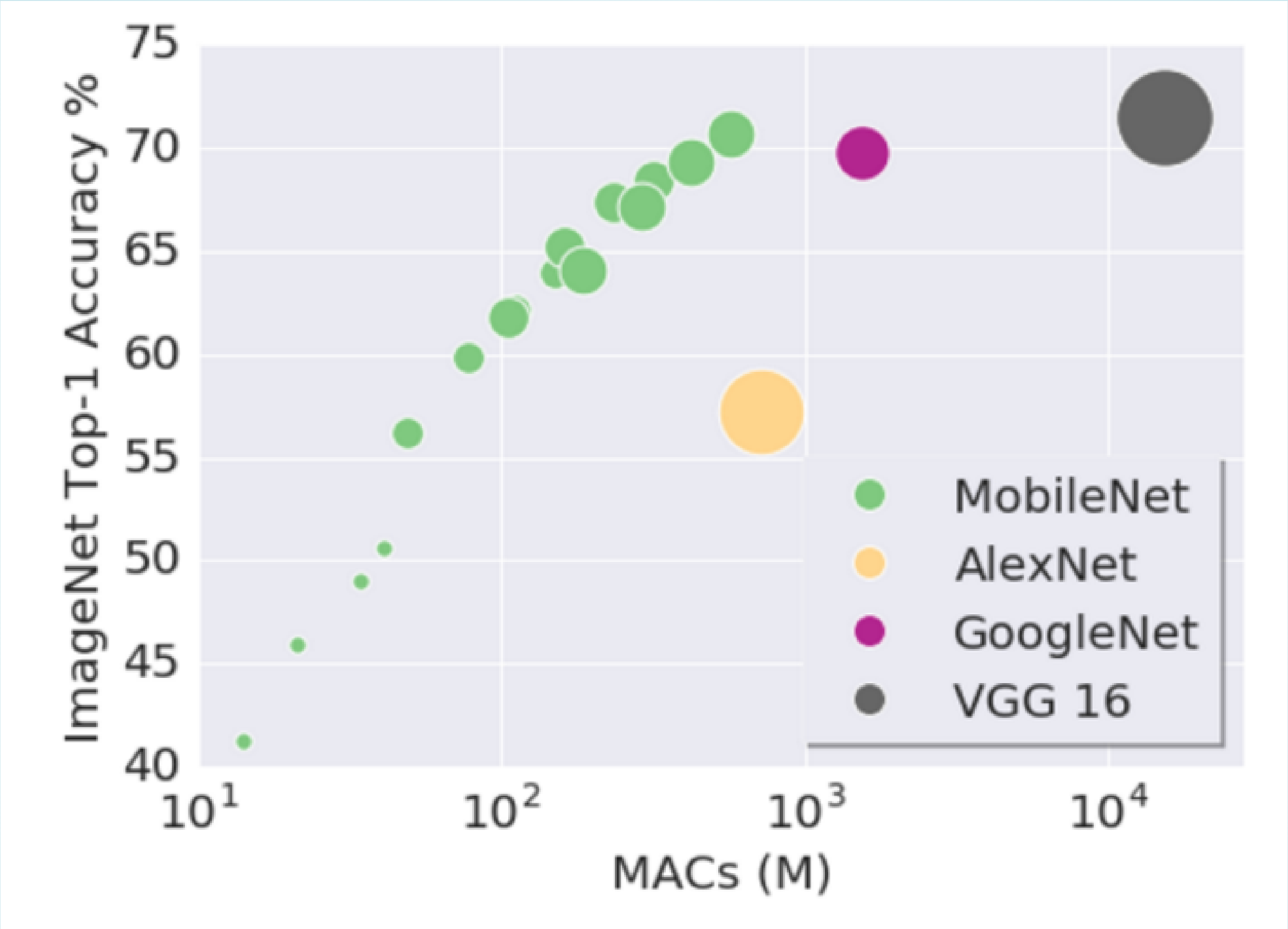
PRACTICALITIES

- Understand tradeoffs
- Use pre-defined models
- Hardware should start to arrive soon

SQUEEZENET

- SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size
- 1x1 and 3x3 layers
- No fully-connected layer

MOBILENETS ~ TF



github.com/tensorflow/research/slim/nets/mobilenet_v1.md

RED DRAGON AI

MOBILENETS ~ KERAS

github.com/keras/applications/mobilenet.py

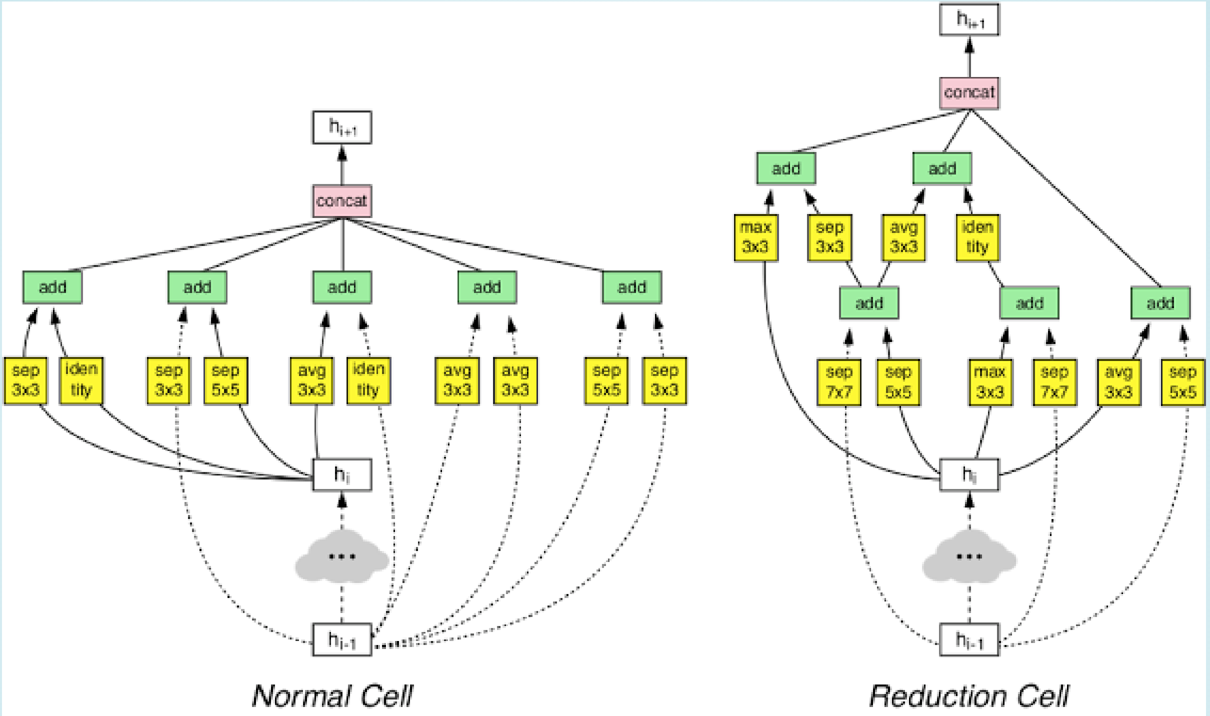
```
from keras.applications.mobilenet import MobileNet
from keras.applications.mobilenet preprocess_input, decode_predictions

img = keras_preprocessing_image.load_img(img_path)
#...
x = preprocess_input(img)

model = MobileNet(weights='imagenet')
preds = model.predict(x)

predictions = decode_predictions(preds, top=1)
```

RED DRAGON AI

AUTOML

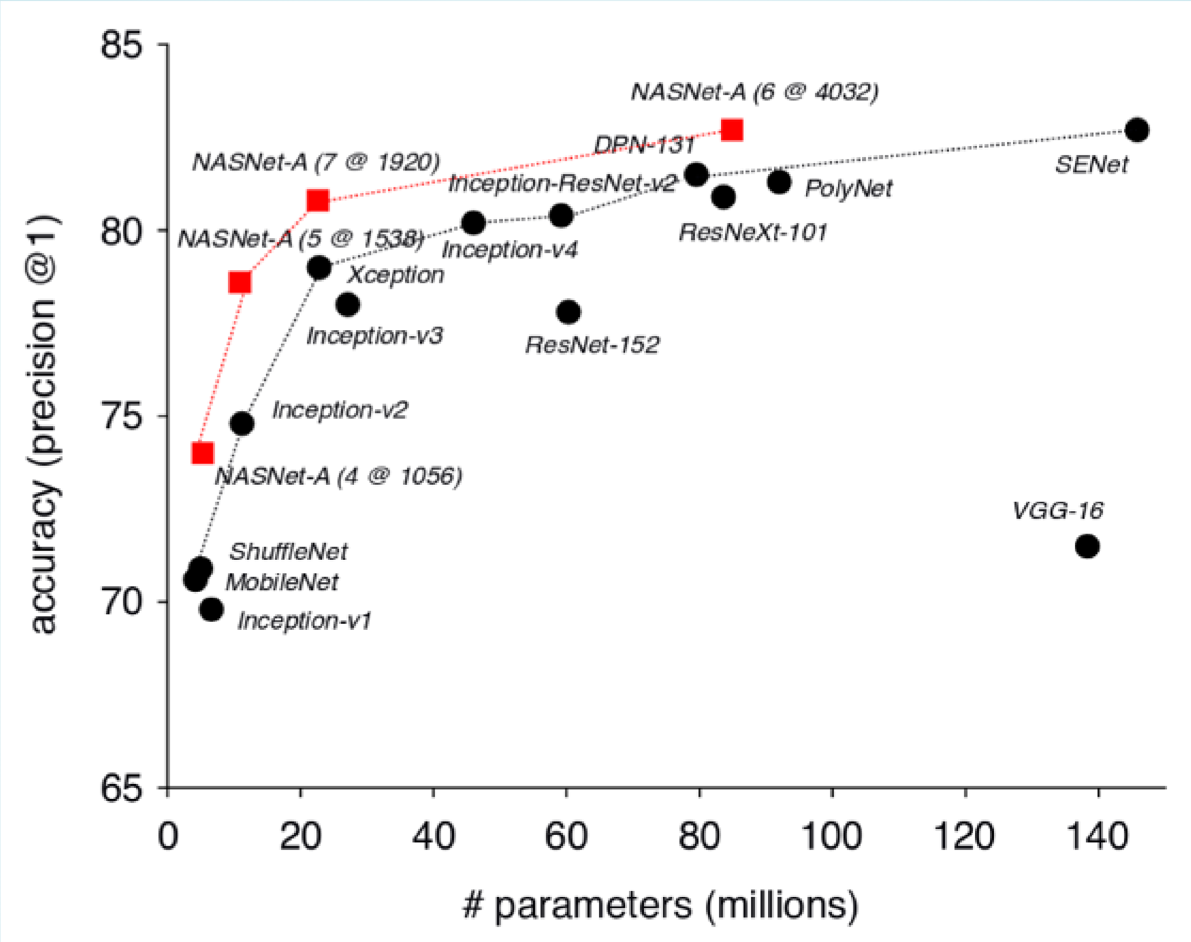
- Recent idea : Search for architecture
  - Optimise for desired performance trade-off
  - Makes use of many GPUs/TPUs
  - Google Blog post
- NB: Not so flexible for 'mash-ups'

NASNET DESIGN



Learning Transferable Architectures for Scalable Image Recognition (Nov-2017)

NASNET ~ TF

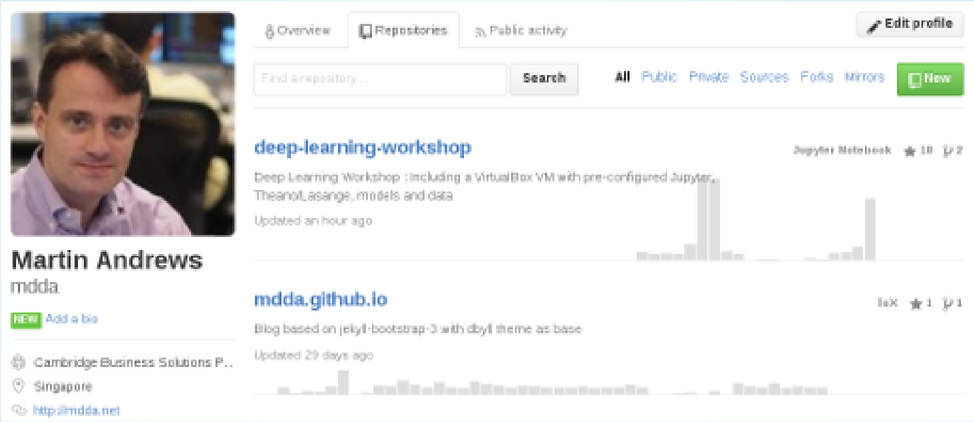


github.com/tensorflow/research/slim/nets/nasnet

RED DRAGON AI

WRAP-UP

- Explore structure vs accuracy tradeoffs
- Even tiny models work 'well enough'
- Lots more behind all this



* Please add a star... *

RED DRAGON AI

- QUESTIONS -

MARTIN.ANDREWS @ REDDRAGON.AI

My blog : http://mdda.net/

GitHub : mdda

RED DRAGON AI