CNN TRAINING & FAMOUS ARCHITECTURES

Dr. Brian Mc Ginley

PREVENTING OVERFITTING

- The usual things should be tried
  - Regularisation
  - Early Stopping
- But sometimes that is enough. Two methods I'll discuss briefly are
  - Data Augmentation
  - Dropout

DATA AUGMENTATION

- If you only have a limited number of images, one way to get "more" is to do data augmentation. The more data we have, the less likely the model is going to be overfit. https://www.tensorflow.org/tutorials/images/data_augmentation
- Where you make multiple copies of the labelled training data and make modifications to it. Then you have many "new" labelled images.

DATA AUGMENTATION

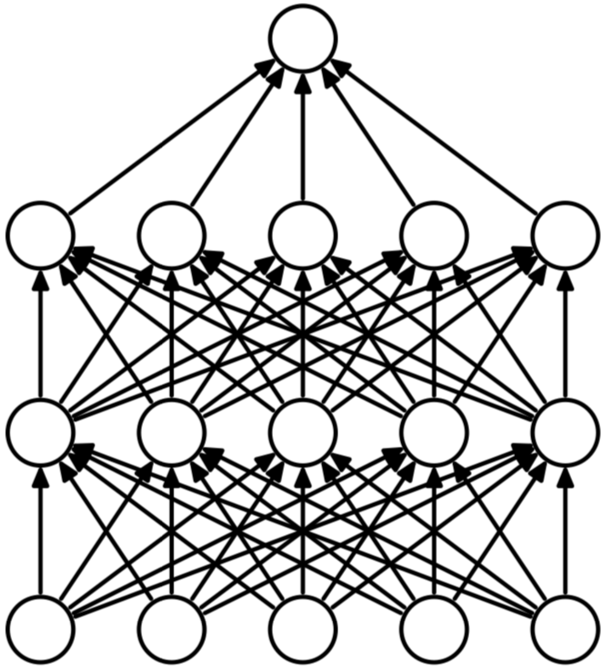
- If, after the modification, it is still obvious to a human what is in the image then a well-trained algorithm should cope too.
- Examples:
  - Random Crops
  - Scaling
  - Rotations
  - Translations
  - Noise
  - Mirror - be careful.
  - Lens and perspective distortions

DROPOUT

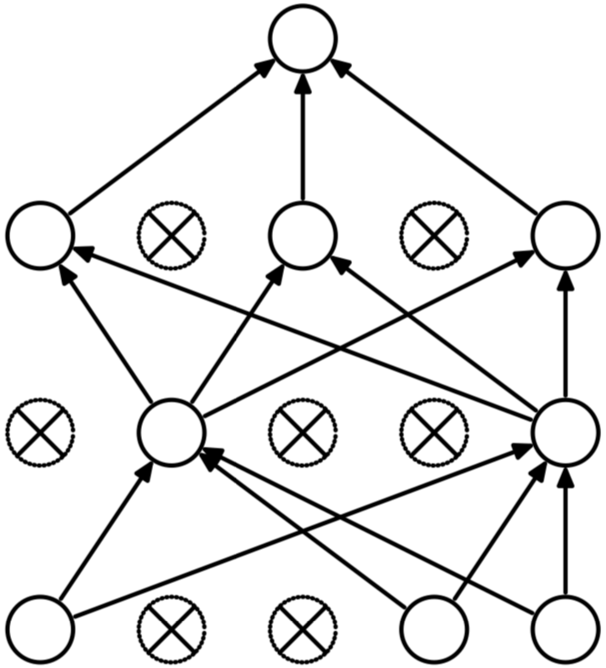
- https://machinelearningmastery.com/dropout-for-regularizing-deep-neural-networks/
- Another method of regularising the parameters, (using a patent by Google) is called Dropout.
- Dropping out 10%, 20% or 40% of the output units randomly from the applied layer during the Training process.
- This ensures the prediction does not become too reliant on some units/filters.
- Important to note: this (and data augmentation) only applies during the training phase. When using a network for inference/prediction, it will always use all the units in every layer.
- Note:
  - Dropout is patented by Google, but it is unclear whether this is to protect it from others doing so or to enforce the patent.

DROPOUT

- Figure: Network with and without dropout
- http://jmlr.org/papers/volume15/srivastava14a.old/srivastava14a.pdf



(a) Standard Neural Net



(b) After applying dropout.

FAMOUS ARCHITECTURES

- To understand how the Architectures have progressed we will take a quick tour of the significant architectures for feed-forward classification and localisation, including:
  - LeNet – 1989 (published in 1998)
  - AlexNet - 2012
  - VGGNet - 2014
  - GoogLeNet - 2014
  - ResNet - 2015
- Don't assume these are the only architectures. There are too many to go through but those mentioned above would be unlikely to be left out of any list.

LENET

- The LeNet (1998) by Yann leCun was the first successful use of a convolutional neural network. It was successfully deployed for use in postal services for reading hand written postal codes. It would be a while before they could be used at scale.

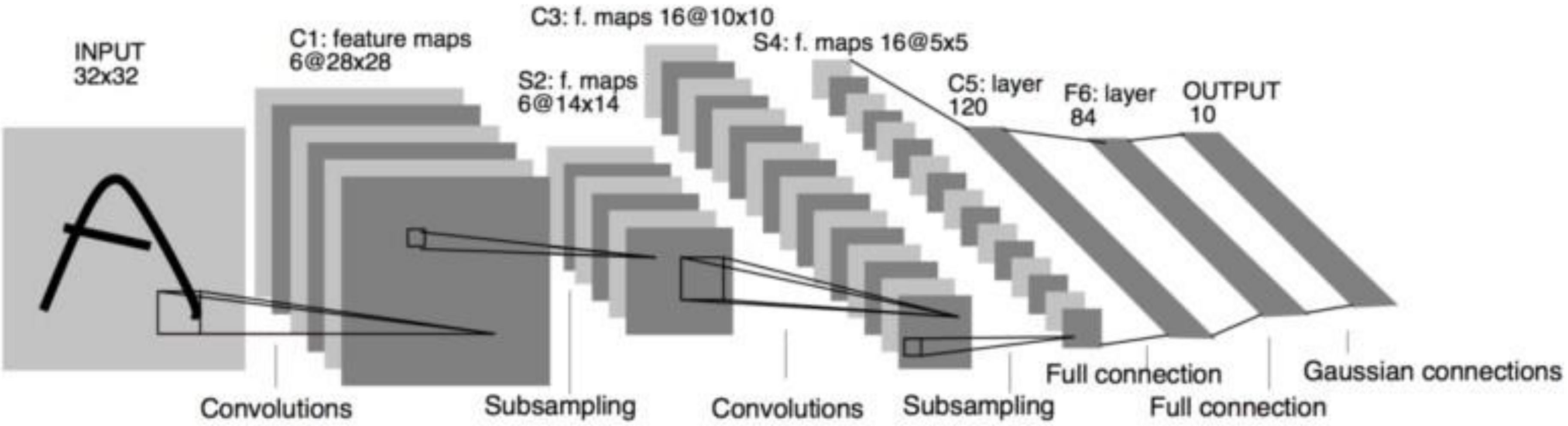


Figure: LeNet Architecture

ALEXNET

- AlexNet (2012) was the break through for CNNs. Alex Krishevsky et al. created an network that won the ImageNet challenge by a huge margin. It has not been won since by anything other than a CNN.

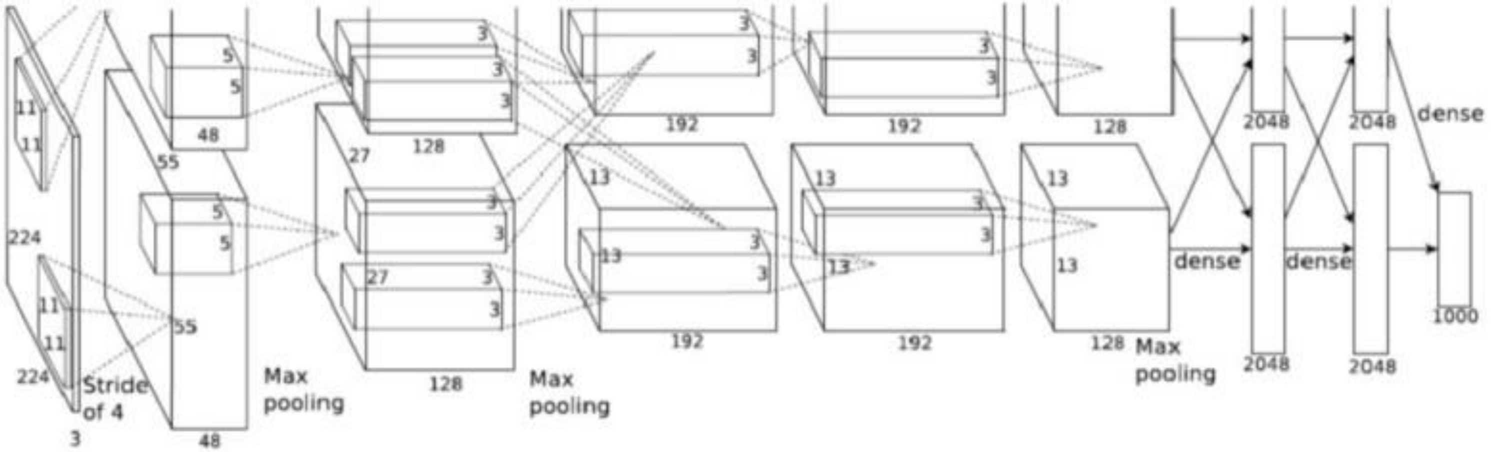


Figure: The AlexNet Architecture

ALEXNET – 8 LAYERS

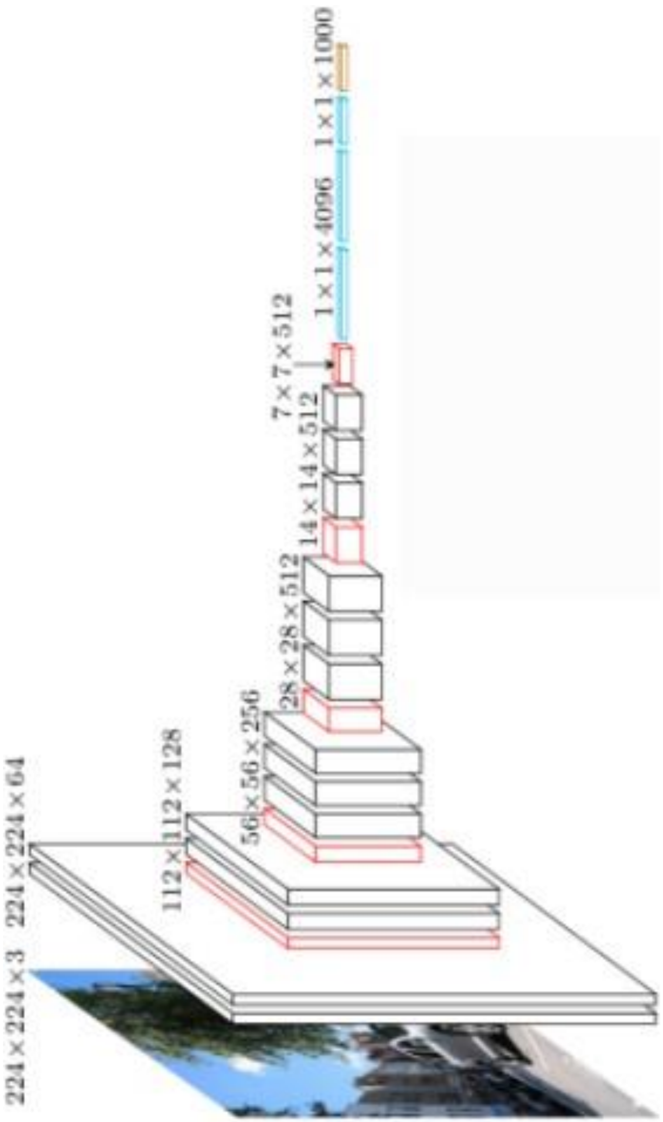
- $227 \times 227 \times 3$ Input
- $55 \times 55 \times 96$ Conv1: 96- 11×11 filters at stride 4, pad 0
- $27 \times 27 \times 96$ MaxPool1: 3×3 filters at stride 2
- $27 \times 27 \times 256$ Norm1: 3×3 Normalisation Layer
- $27 \times 27 \times 256$ Conv2: 256 5×5 filters at stride 1, pad 2
- $13 \times 13 \times 256$ MaxPool2: 3×3 filters at stride 2
- $13 \times 13 \times 256$ Norm2: 3×3 Normalisation Layer
- $13 \times 13 \times 384$ Conv3: 384 3×3 filters at stride 1, pad 1
- $13 \times 13 \times 384$ Conv4: 384 3×3 filters at stride 1, pad 1
- $13 \times 13 \times 256$ Conv5: 256 3×3 filters at stride 1, pad 1
- $6 \times 6 \times 256$ MaxPool3: 3×3 filters at stride 2
- 4096 FC6: 4096 neurons
- 4096 FC6: 4096 neurons
- 4096 FC6: 4096 neurons

ALEXNET

- ReLU
- Norm layers which aren't really used any more
- Seven CNN ensembles
- Training details:
  - Dropout of 0.5
  - Batchsize 128
  - A lot of data augmentation
  - SGD with momentum 0.9
- Learning rate 1e−2, which was reduced by a factor of 10 each time.
- This was carried out manually.
- The architecture looks a little more complicated than it actually was.
  - The problem was that there was not enough memory on GPUs at the time to fit the whole model on a single GPU, so the diagram shows it split across two GPUs.

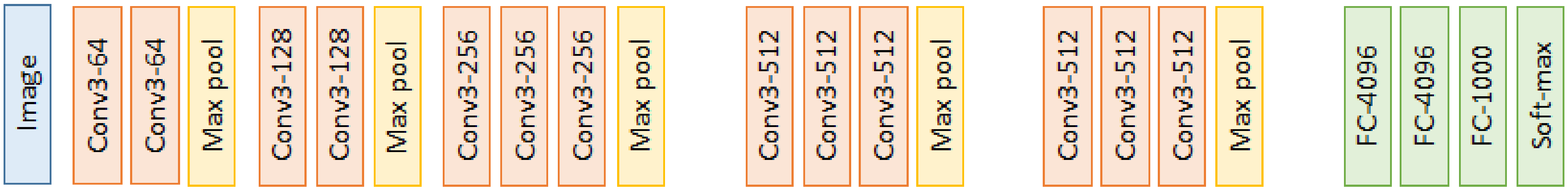
VGG-NET

- VGGNet - 2014 (Visual Geometry Group – Oxford - Davi Frossard K. Simonyan and A. Zisserman).
- Came second in the classification category on ImageNet but first in localisation.
- Main concepts: Go deeper (11 to 19 layers) with more uniform conv-layers - i.e. all 3 × 3 stride 1 and pad 1. 2 × 2 max-pool stride 2.
- With depth, the 3 ×3 filters further into the network have an increased receptive field.



VGG-NET TRAINING

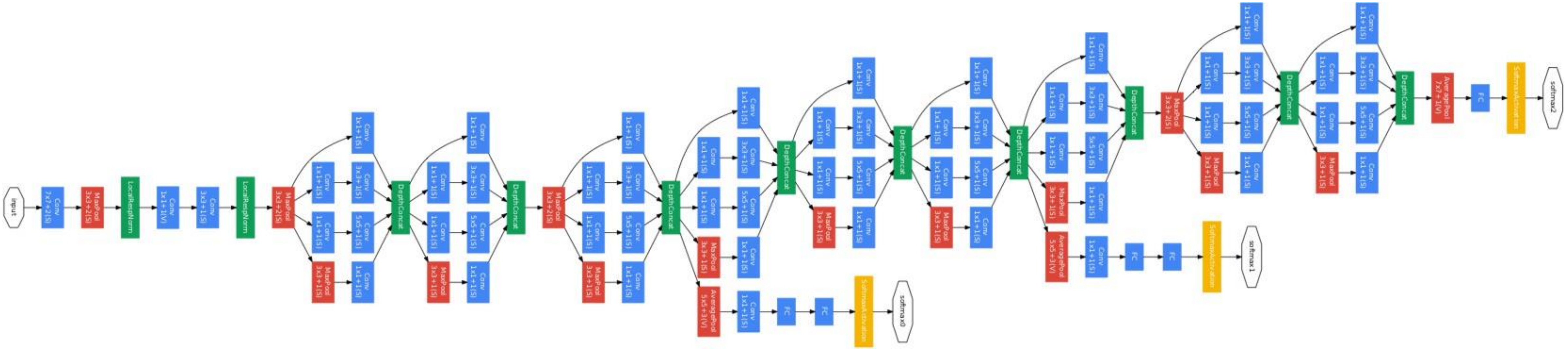
- Similar training to AlexNet.
- There was none of the Normalisation layers. They also use an Ensemble of 7 networks.
- One other thing to note is that the features of the last FC-4096 layer were found to generalise well to other tasks.
- This is useful to note for a technique called Transfer Learning.



VGG-16

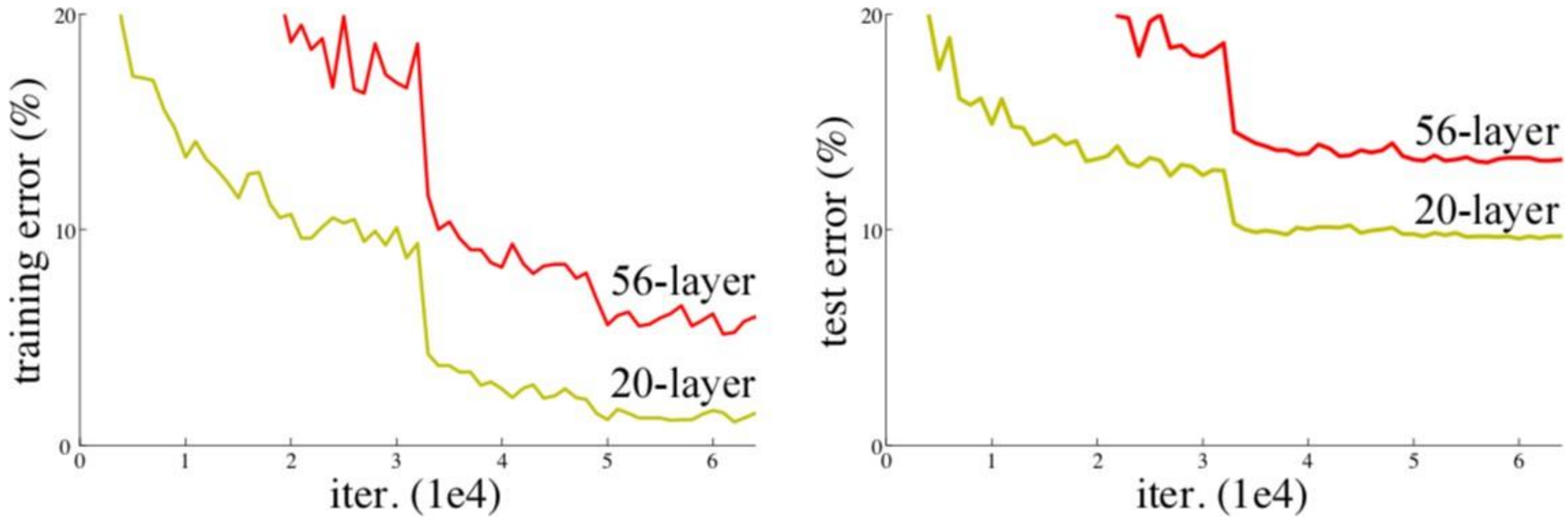
GOOGLENET

- GoogLeNet – later renamed Inception



MORE DEPTH BETTER?

- Evidence that more layers does not necessarily mean better accuracy.
- Deep Residual Learning for Image Recognition paper



RESNET

- Published (2015) by Microsoft - introduced a new architecture called Residual Network.
  - 34 layer ResNet - Deep Residual Learning for Image Recognition - Kaiming He et al.
- ResNets have become associated with solving the *vanishing gradient problem* as they will certainly do this.
- But they created examples at 50 layers, 101 layers and 152 layers.