Chapter 1

Famous Networks

LeNet-5

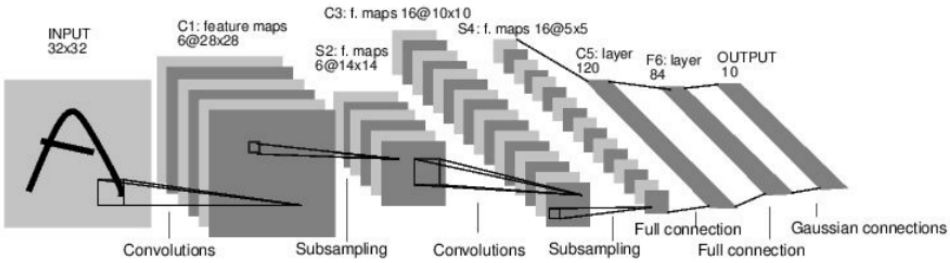

FIGURE 1.1: LeNet

The first successful applications of Convolutional Networks were developed by Yann LeCun in 1990s. Of these, the best known is the LeNet architecture that was used to read zip codes, digits, etc.

AlexNet

The first work that popularized Convolutional Networks in Computer Vision was the AlexNet, developed by Alex Krizhevsky, Ilya Sutskever and Geoff Hinton. The AlexNet was submitted to the ImageNet ILSVRC challenge in 2012 and significantly outperformed the second runner-up (top 5 error of 16% compared to runner-up with 26% error). The Network had a very similar architecture to LeNet, but was deeper, bigger, and featured Convolutional Layers stacked on top of each other (previously it was common to only have a single CONV layer always immediately followed by a POOL layer). 60M parameters.

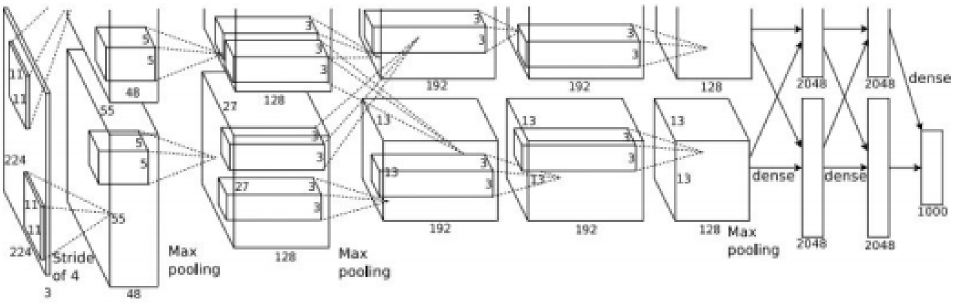

FIGURE 1.2: AlexNet. There is an error in the paper figure, the input image must be of 227 so that the rest of volumes are coherent

VGGNet

The runner-up in ILSVRC 2014 was the network from Karen Simonyan and Andrew Zisserman that became known as the VGGNet. Its main contribution was in showing that the depth of the network is a critical component for good performance. Their final best network contains 16 CONV/FC layers and, appealingly, features an extremely homogeneous architecture that only performs 3×3 convolutions and 2×2 pooling from the beginning to the end. Their pretrained model is available for plug and play use in Caffe. A downside of the VGGNet is that it is more expensive to evaluate and uses a lot more memory and parameters (140M). Most of these parameters are in the first fully connected layer, and it was since found that these FC layers can be removed with no performance downgrade, significantly reducing the number of necessary parameters.

The most of the network parameters are in the FC layer and most of the memory required by the network is used in the first 2 ConvLayers.

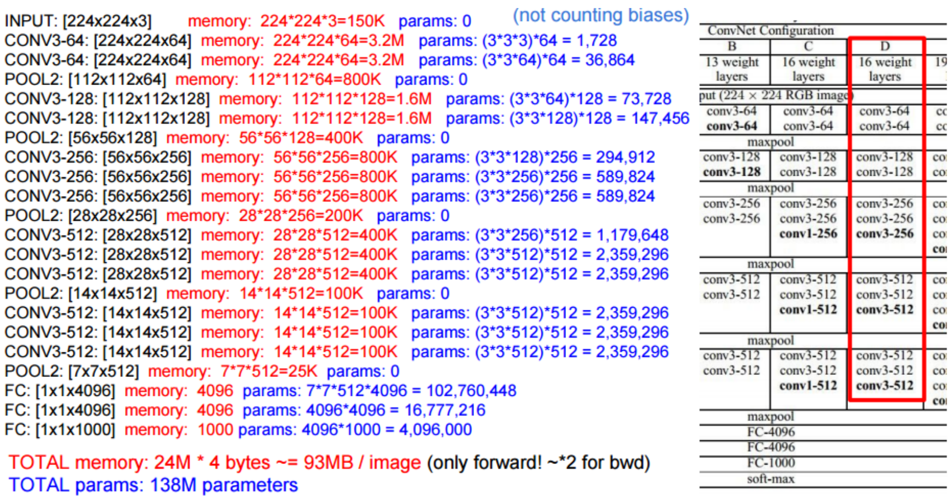

FIGURE 1.3: VGGNet - 7.3% top 5 error in ImageNet

GoogLeNet

The ILSVRC 2014 winner was a Convolutional Network from Szegedy et al. from Google. Its main contribution was the development of an Inception Module that dramatically reduced the number of parameters in the network (4M, compared to AlexNet with 60M). Additionally, this paper uses Average Pooling instead of Fully Connected layers at the top of the ConvNet, eliminating a large amount of parameters that do not seem to matter much. There are also several followup versions to the GoogLeNet, most recently Inception-v4.

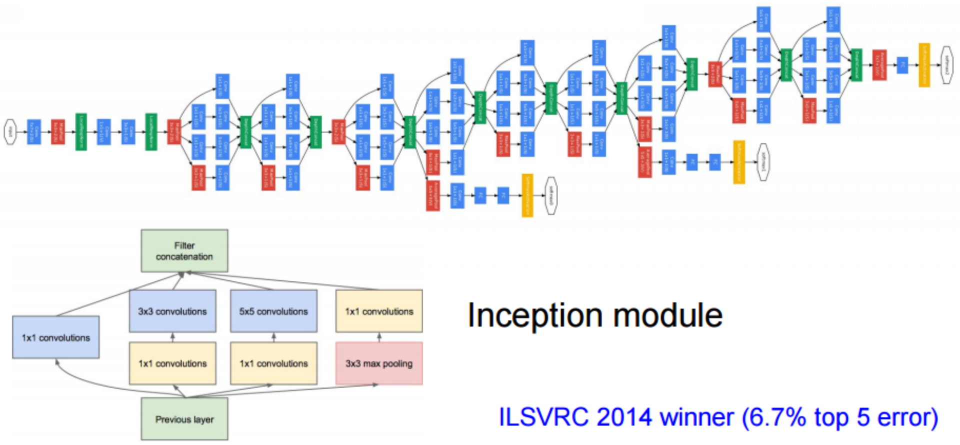

FIGURE 1.4: VGGNet - 6.7% top 5 error in ImageNet

type	patch size/ stride	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj	params	ops
convolution	7×7/2	112×112×64	1							2.7K	34M
max pool	3×3/2	56×56×64	0								
convolution	3×3/1	56×56×192	2		64	192				112K	360M
max pool	3×3/2	28×28×192	0								
inception (3a)		28×28×256	2	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	2	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480	0								
inception (4a)		14×14×512	2	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	2	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	2	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	2	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	2	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832	0								
inception (5a)		7×7×832	2	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	2	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024	0								
dropout (40%)		1×1×1024	0								
linear		1×1×1000	1							1000K	1M
softmax		1×1×1000	0								

FIGURE 1.5: VGGNet structure

ResNet

Let $H(x)$ be a function that you desire to obtian. In a typical net you would compute a squence of steps `ReLu(ReLu(xw1+b1)*w2+b2)` to transform x to $H(x)$. Instead, in a ResNet you compute a delta to be added to the original input to obtain $H(x)$.

What is nice about it is that in *plain nets*, gradients must flow through all the transformations. Instead, in *residual nets* because it is addition (distributes the gradient equally to all its children), the gradient with flow through the (weights, ReLU) but will also skip this transformations and will go directly to the previous part and flow directly to the previous block. So the gradients can skip all the transformations and go directly to the first layer. In this way, you can train very fast the first layer which is doing simple statistics, and the rest of layers will learn to add to the single in between to make it work at the end.

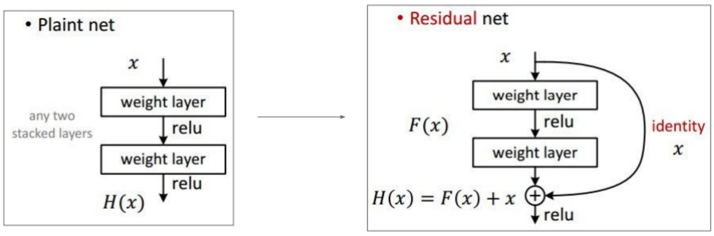

FIGURE 1.6: Plain vs Residual Net

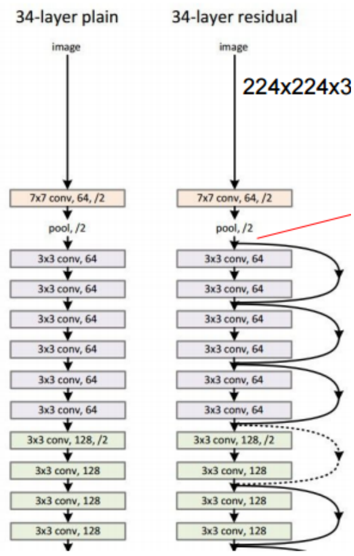

FIGURE 1.7: ResNet (much more layers than the ones on the diagram)

Another way of seeing it, it that ResNets are only computing a delta on top of the identity. So it makes it nice to optimize.

Residual Network developed by Kaiming He et al. was the winner of ILSVRC 2015. It features specialskip connections and a heavy use of batch normalization. The architecture is also missing fully connected layers at the end of the network. ResNets are currently by far state of the art Convolutional Neural Network models and are the default choice for using ConvNets in practice (as of May 10, 2016). In particular, also see more recent developments that tweak the original architecture from Kaiming He et al. Identity Mappings in Deep Residual Networks (published March 2016). It is interesting that after the first layer they do polling (the only polling in all the net) and they scale the input image of 244×244 to 56×56, and the net works that well. Its crazy that all the layer (except the first one) work wit $56 \times 56 \times ?$ and even compressing the data this much it has a high accuracy.

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

FIGURE 1.8: ResNet structure. 3.6% top 5 error in ImageNet, 152 layers, 2-3 weeks training on 8 GPU machine, faster at test time that VGGNet

Should we add infinite layers?

In plot **??** it is clear that networks are getting deeper and deeper But we have to be careful.

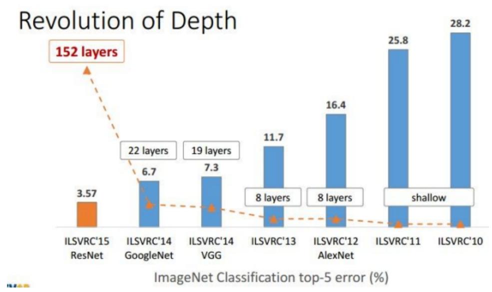

FIGURE 1.9: Depth revolution

Plot **??** shows CIFAR-10 training error. In the left plain nets (weighted layer + ReLU) in the right ResNet. How it is possible to get a higher training error (dashed lines) with higher number of layers? It should not happen, the model is more complex. The explanation is that we are still not capable of optimizing them good enough.

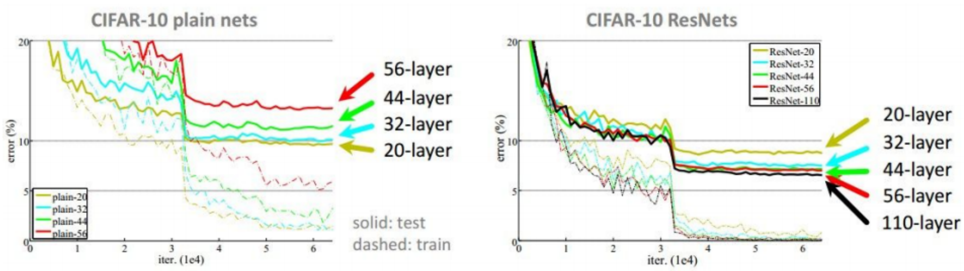

FIGURE 1.10: CIFAR-10 training error

However, ResNets always improve the test and training error.

So the answer to the question is that we should keep adding more layers but not in a naive way, we should do it in a ResNet way