**Week 2 – Notes**

**Case Studies**

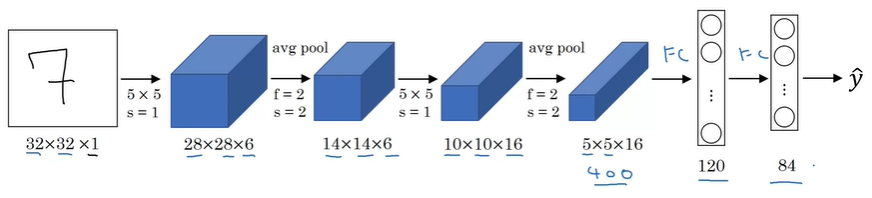
**Why look at case studies?**

There are classic networks, such as: LeNet-5, AlexNet and VGG

More modern networks are the ResNet which has up to 152 layers and Inception

**Classic Networks**

LeNet-5 was published in Yann LeCun in 1998 and it has ~60k parameters



Differences compared to nowadays:

At that time, avg pooling was more popular than the max pooling

Instead of softmax as an activation function of the output layer, another function was used

Sigmoid and tanh were used instead of ReLU

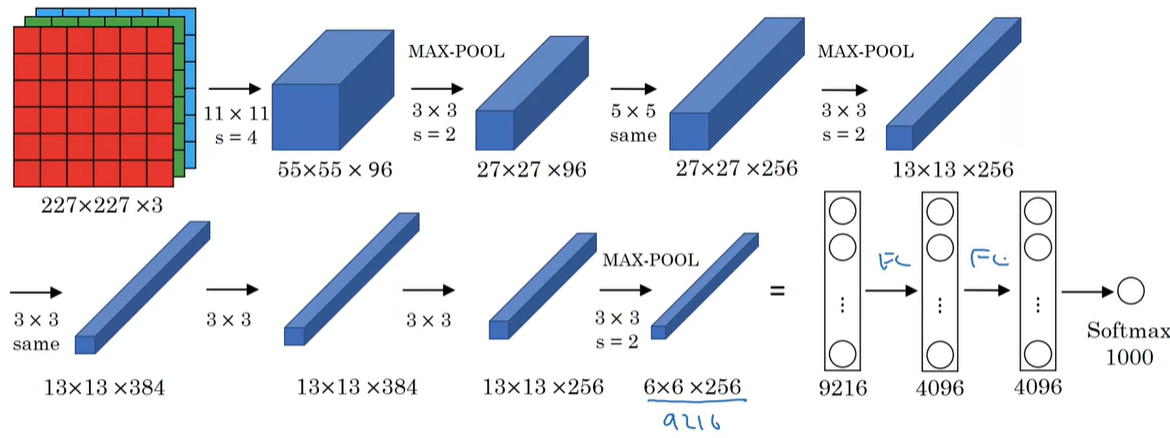
The non-linearity was applied after the pooling layer

Things that are still valid today:

As the data progress through the network, nH and nW decrease and nC increases

The architecture is relevant: conv -> pool -> conv -> pool -> fc -> fc -> output

AlexNet was published in 2012 and it has ~60 million parameters



It uses convolutions with large kernels and sequential convolutions with the same kernel (even with the same padding and stride)

Things that are still valid today:

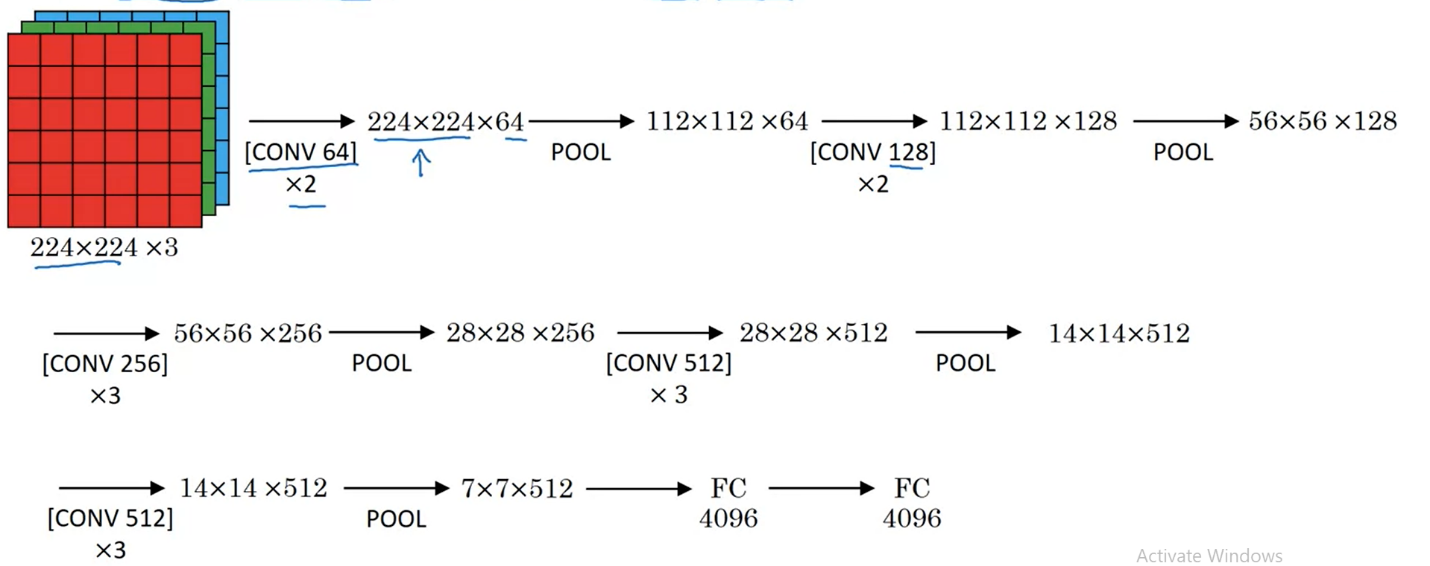
The activation function is ReLU

Differences compared to nowadays:

Used a Local Response Normalization (for each pixel across all channels) – turned out to not be that effective

VGG-16 was published in 2015 and it has ~138 million parameters

There also is a version called VGG-19



This network simplified very much the architecture of CNNs, which became more and more complicated

The authors used only 2 types of layers: Conv2D with 3x3 filters, a stride of 1 and the same padding and max pooling layer with a kernel of 2x2 and a stride of 2

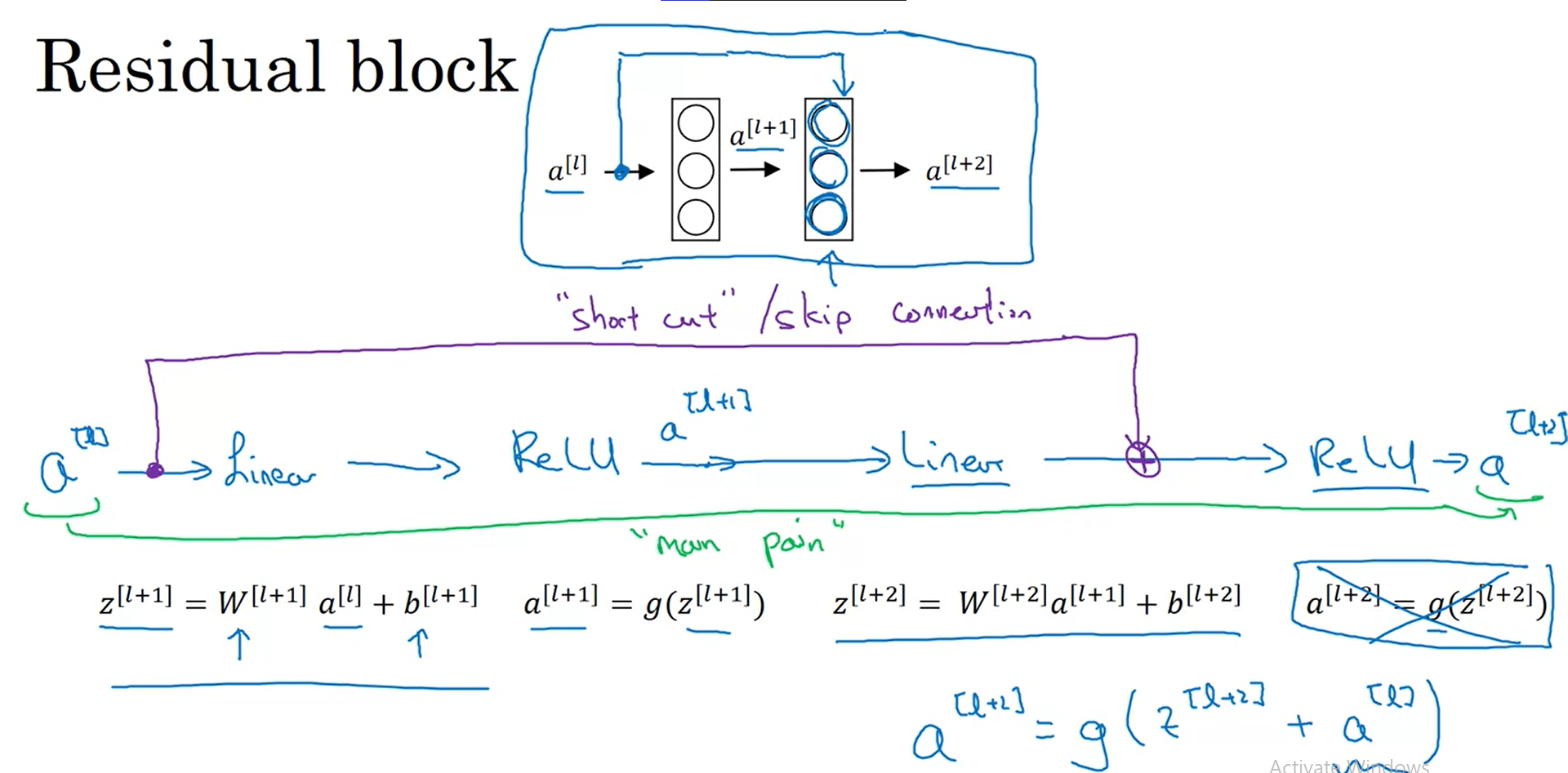
Interesting idea: you can use consecutive identical conv layers

**ResNets**

Deeper networks have problems such as vanishing (tanh or sigmoid activation functions) or exploding gradients (relu activation function), so to solve that we can use short cuts / skip connections

This means we take the outputs of a layer and we add to the output of the layer, but before we apply the activation function (in our case ReLU)

We don’t add after we apply the activation function because we want to avoid the exploding gradients; additionally, we don’t add before the we compute the current Z to not damage the computation of Z of the current layer



A network w/o skip connections is called a plain network

The skip connection is added so that each time we jump over one layer

Plain network: in reality, as we add more layers, after one point the training error increases

ResNets: as we add more layers, the training error decreases (at most there’s a plateau)

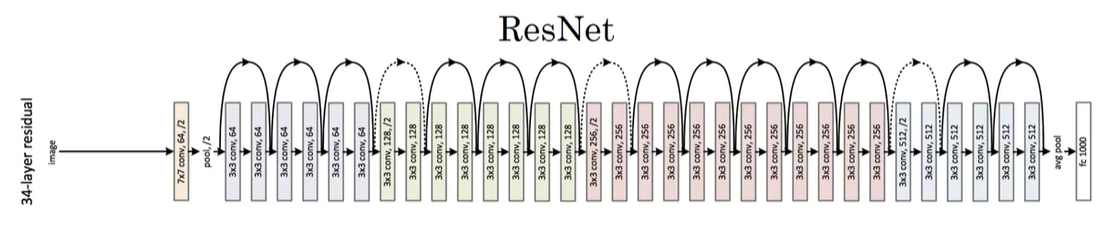
**Why ResNets Work?**

Let’s say you have a big nn that has the output a[l]

If you add 2 new layers and use a skip connection from a[l] to a[l+2] (layer l and l + 2 have outputs with the same size), then in the worst case because you use weight decay (L2 regularization) your network has w[l+2] and b[l+2] as zero values and then you will compute in the l+2 layer g(a[l]) which is equal to a[l] in the case of g = ReLU; thus, there’s an identity function because a[l+2] == a[l]

On the contrary, if you network can learn anything, then the skip connection is useful, but in any case, it doesn’t do any harm to you network

If the size of a[l] and a[l+2] is different, then when adding the skip connection, we multiply a[l] with Ws, so that the Ws \* a[l] has the same shape as a[l+2]; however, mostly same convolutions are used, so that we have outputs with identical shapes



We can see that after applying a pooling layer, the dimensions of the outputs don’t match, so we have to use a skip connection + adjust the output of the forwarded layer