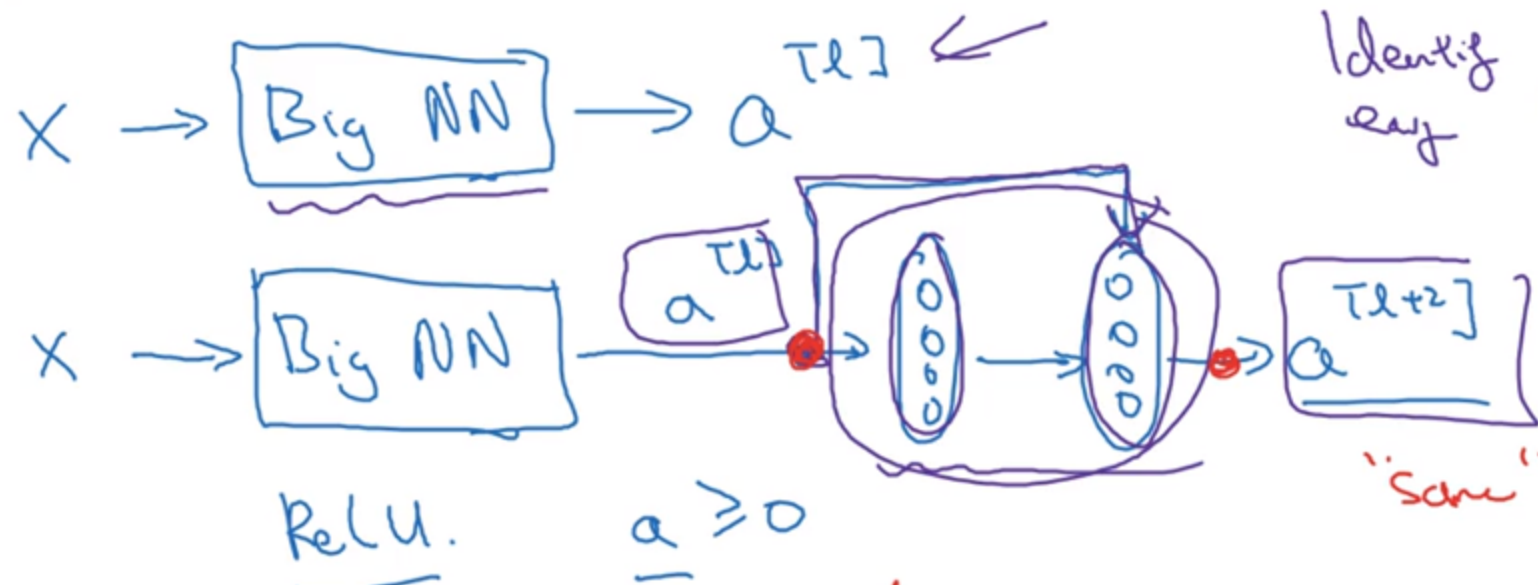


Why do residual networks work?



Identifying for Residual blocks to learn!

a lot of use of same convolutions, so that the d of this equal to input l.

$$\begin{aligned}
 a^{[l+2]} &= g(z^{[l+2]} + a^{[l]}) \\
 &= g(\underbrace{W^{[l+2]} a^{[l+1]} + b^{[l+2]}}_{\text{If } W^{[l+2]}=0, b^{[l+2]}=0} + \underbrace{W_s a^{[l]}}_{\text{256} \times \text{128}}) = g(a^{[l]}) \\
 &= \underline{a^{[l]}}
 \end{aligned}$$

256

256 x 128

128

it could be a matrix of parameters we learned, it could be matrix that just implement zero padding, so that it takes $a[l]$ and then zero pads it to be 256 d.

ResNet

Plain



ResNet

