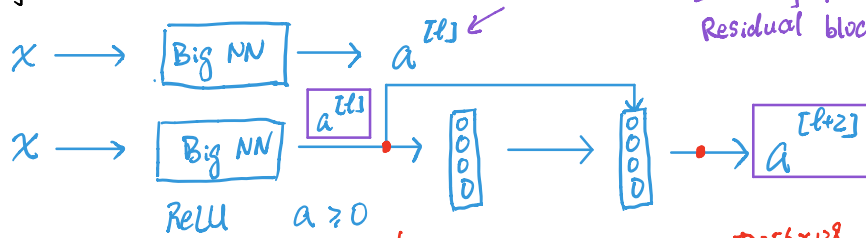


Why do residual networks work?

Identity function is easy for Residual block to learn!



$$\begin{aligned}
 a^{[l+2]} &= g(\underbrace{z^{[l+2]}}_{\text{ReLU } a \geq 0} + a^{[l]}) \\
 &= g(W^{[l+2]} a^{[l+1]} + b^{[l+2]} + W_s a^{[l]}) = g(a^{[l]}) \\
 &\quad \uparrow \quad \quad \quad \uparrow \quad \quad \quad \uparrow \\
 &\quad \text{if } W^{[l+2]} = 0, b^{[l+2]} = 0 \quad \quad \quad \text{if } W_s = 0
 \end{aligned}$$

Annotations:  $256$  (under  $a^{[l+2]}$ ),  $128$  (under  $a^{[l]}$ ),  $\mathbb{R}^{256 \times 128}$  (above  $W_s$ ).