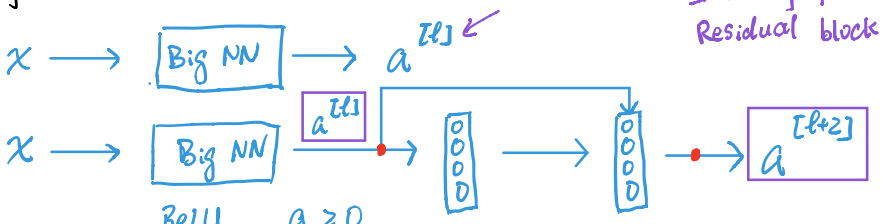
Why do residual networks work?

Indentity function is easy for Residual block to learn!



ReLU $a \geq 0$

$$a^{[l+2]} = g(z^{[l+2]} + a^{[l]})$$

$$= g(w^{[l+2]} a^{[l+1]} + b^{[l+2]} + w_s a^{[l]}) = g(a^{[l]})$$

$$= a^{[l]}$$

$a^{[l+2]}$: 256

$w_s \in \mathbb{R}^{256 \times 128}$

$a^{[l]}$: 128

If $w^{[l+2]} = 0$, $b^{[l+2]} = 0$