

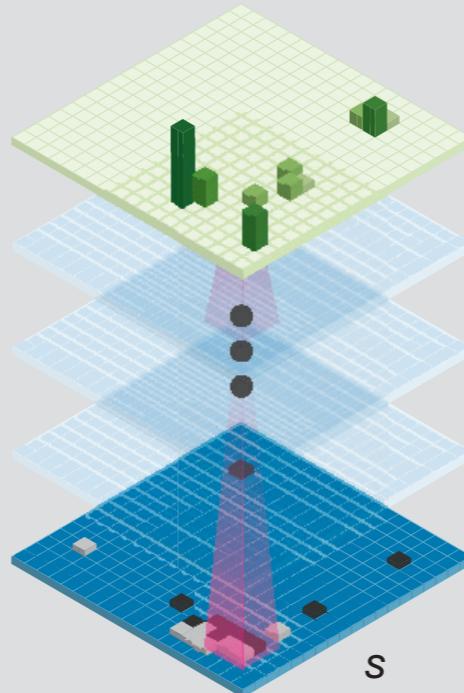
$$a_t = \underset{a}{\operatorname{argmax}}(Q(s_t, a) + u(s_t, a))$$

$$u(s, a) \propto \frac{P(s, a)}{1 + N(s, a)}$$

$$V(s_L) = (1 - \lambda)v_\theta(s_L) + \lambda z_L$$

Policy network

$$p_{\sigma/\rho}(a|s)$$



Value network

$$v_\theta(s')$$

