Algorithm 1 A2C algorithm

Require: Initialize Actor-Critic network G with parameter θ.

for each episode **do**
  Get initial state s
  Initialize a storage buffer S, A, R, V, S'
  for $i = 1, 2, 3.., N$ **do**
    Sample an action $a \sim G(s)_\pi$ and get the associeted value $v \leftarrow G(s)_v$
    Run the action a through the environment, obtain the reward and next state $r, s' \leftarrow ENV(s, a)$
    Collect and store $S, A, R, V, S' \leftarrow s, a, r, v, s'$
    $s \leftarrow s'$
  end for
  Compute the discounted returns $\hat{V} = \sum_{l=0}^{N-1} \gamma^l r_{t+l}$
  Compute an advantage function $\psi(V, R, S')$
  Optimize θ to minimize $-\log(G(A \mid S)_\pi)\psi(V, R, S') + \lambda \|V - \hat{V}\|$
  Empty S, A, R, V, S'