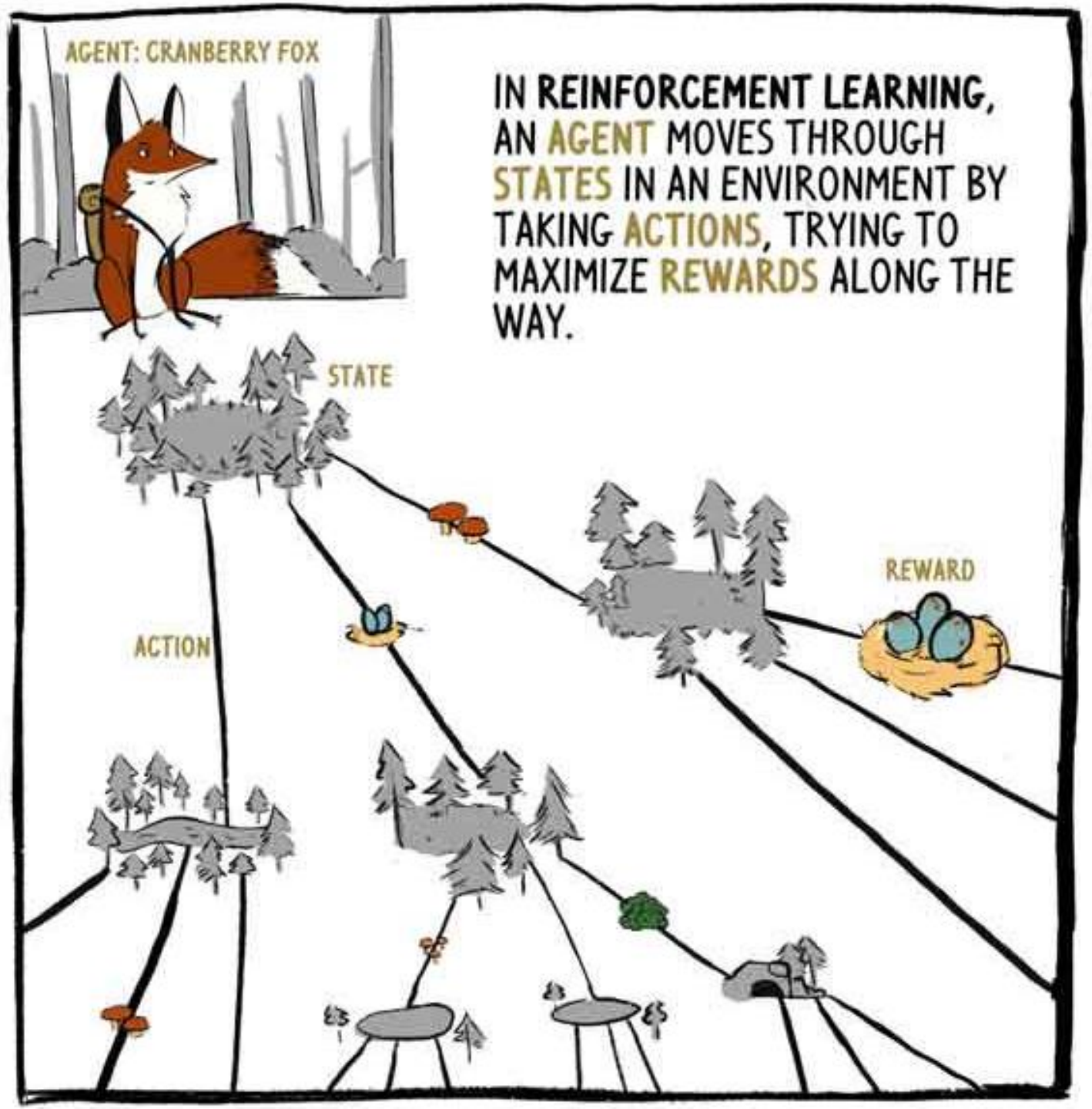
Actor-critic

Сапожников Денис, БПМИ-192

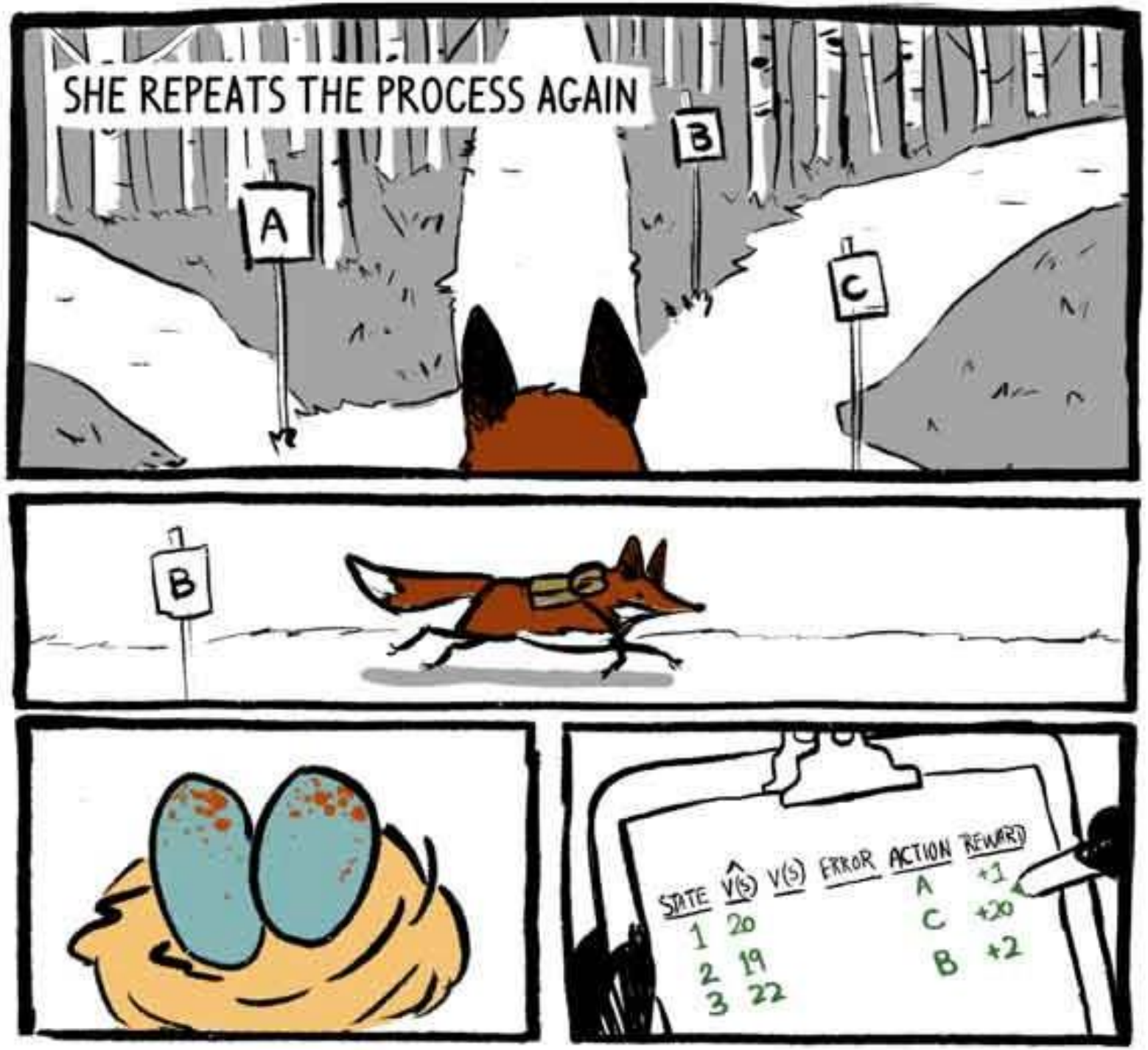
Преимущества и недостатки policy gradient по сравнению с Q-learning

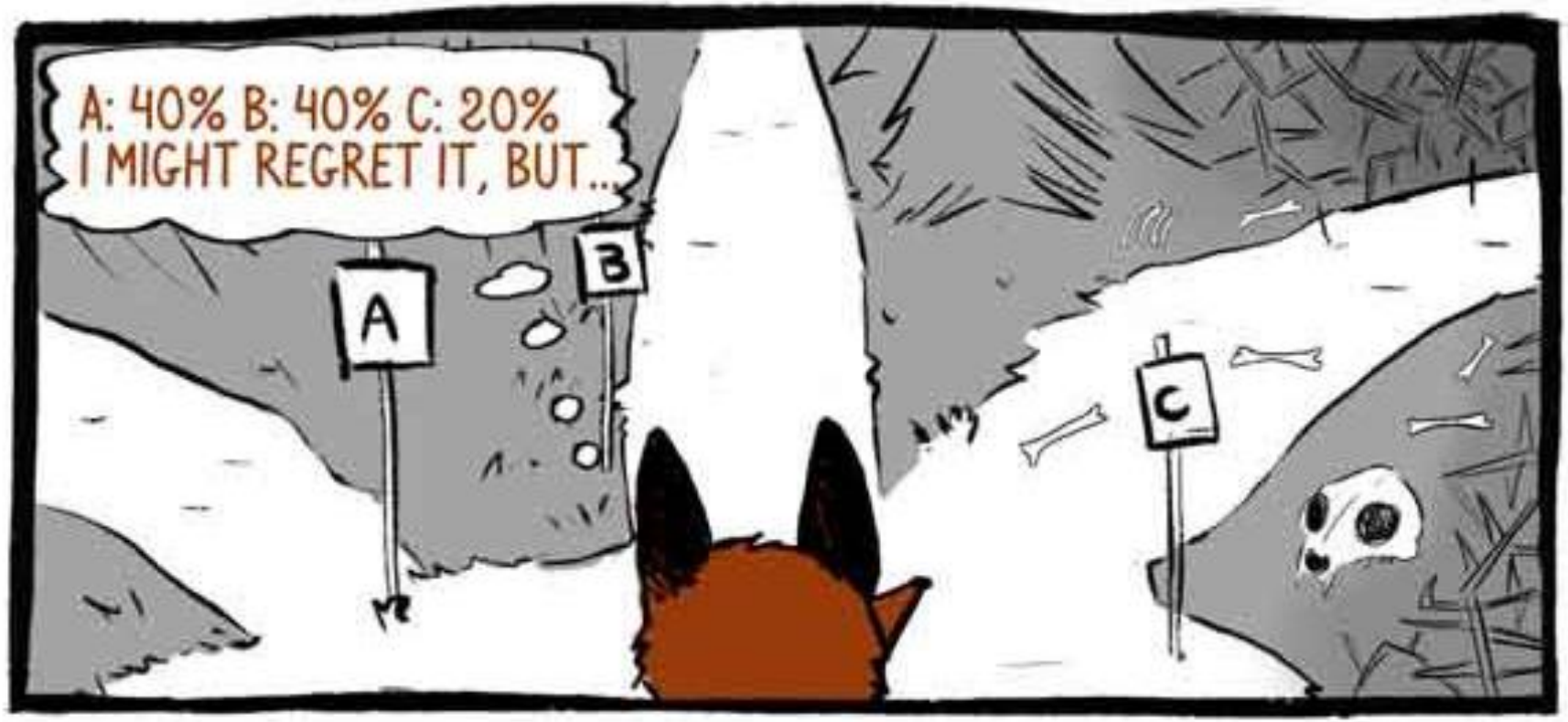
- Преимущества:
  - Легко обобщается на задачи с большим множеством действий, в том числе на задачи с непрерывным множеством действий;
  - По большей части избегает конфликта между эксплуатацией (exploitation) и исследованием (exploration), так как оптимизирует напрямую стохастическую стратегию $\pi_\theta(a|s)$
  - Имеет более сильные гарантии сходимости
- Недостатки:
  - Очень-очень долгий
  - В случае конечных МППР Q-learning сходится к global minimum


AGENT: CRANBERRY FOX
IN REINFORCEMENT LEARNING, AN AGENT MOVES THROUGH STATES IN AN ENVIRONMENT BY TAKING ACTIONS, TRYING TO MAXIMIZE REWARDS ALONG THE WAY.
STATE
ACTION
REWARD

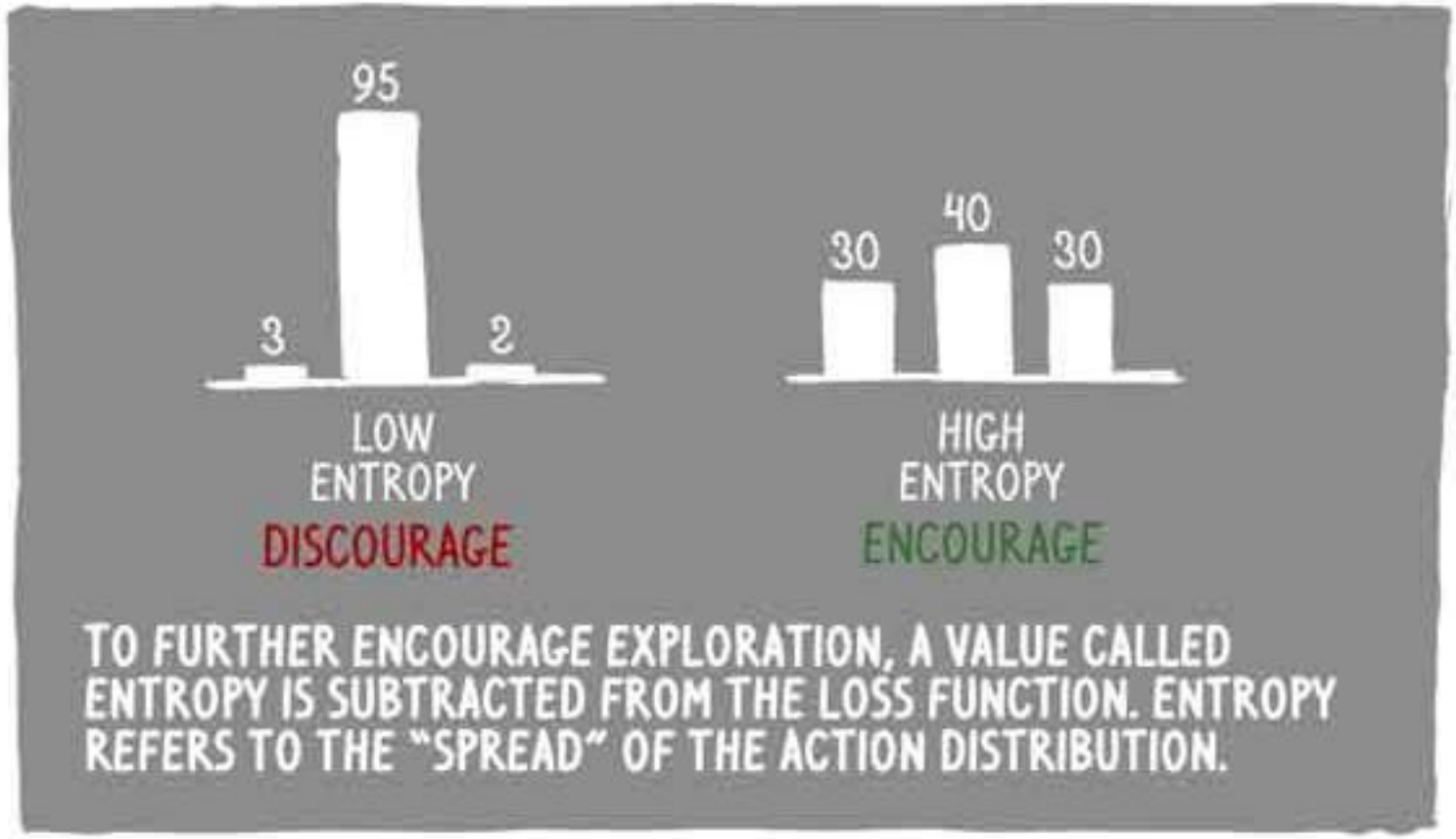

A2CS TAKE IN A STATE—SENSORY INPUTS IN CRANBERRY'S CASE—AND GENERATE TWO OUTPUTS:
1) AN ESTIMATE OF HOW MANY REWARDS THEY EXPECT TO GET FROM THAT POINT ONWARDS, THE STATE VALUE.
2) A RECOMMENDATION OF WHAT ACTION TO TAKE, THE POLICY
THE "CRITIC"
WOW, WHAT A WONDERFUL GLEN! THIS WILL BE A FRUITFUL DAY OF FORAGING. I BET I'LL GATHER 20 REWARD POINTS BEFORE SUNSET TODAY.
$\hat{V}(S) = 20$
THE "ACTOR"
THOSE FLOWERS LOOK PRETTY, I'M FEELING DRAWN TOWARDS "A"...
A: 80%
B: 10%
C: 10%
A
B
C

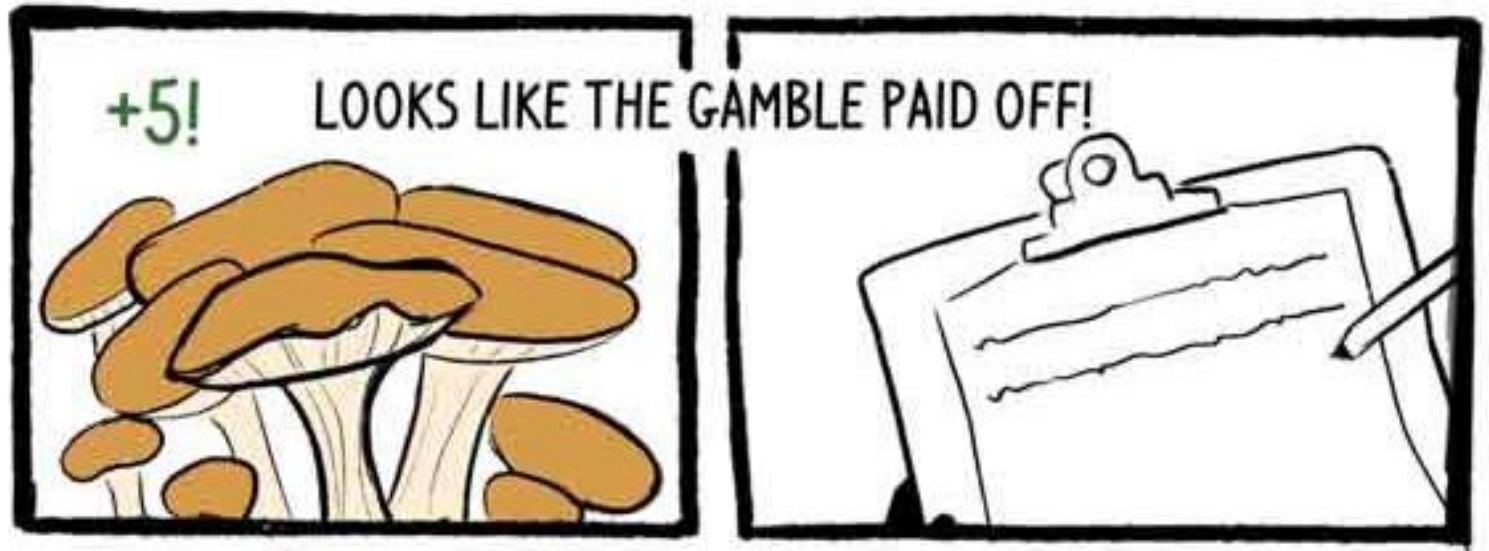

1
2
THIS SET OF STATE-ACTION-REWARD MAKES UP A SINGLE OBSERVATION. SHE'LL RECORD THIS ROW OF DATA IN HER JOURNAL BUT SHE'S NOT GOING TO REFLECT ON IT JUST YET.
STATE
V̂(S)
V(S)
ERROR
ACTION
REWARD
1
20
A
+1
SHE'LL FILL THESE OUT WHEN SHE STOPS TO REFLECT.
SOME AUTHORS ASSOCIATE REWARD 1 WITH TIMESTEP 1, OTHERS ASSOCIATE IT WITH TIMESTEP 2. ALL ARE REFERRING TO THE SAME CONCEPT: A REWARD IS ASSOCIATED WITH THE STATE AND ACTION DIRECTLY PRECEDING IT.


SHE REPEATS THE PROCESS AGAIN
A
B
C
B
STATE
V(s)
V(s)
ERROR
ACTION
REWARD
1
20
A
+1
2
19
C
+20
3
22
B
+2


A: 40% B: 40% C: 20%
I MIGHT REGRET IT, BUT...
A
B
C

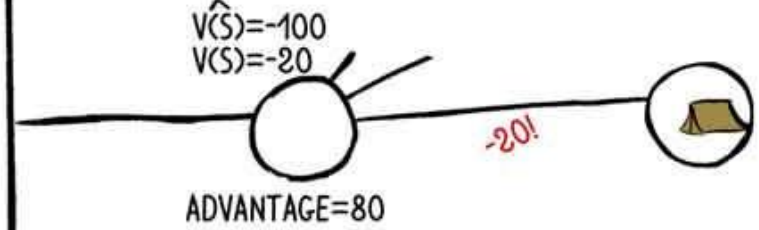

SOMETIMES EXPLORING UNKNOWN THINGS
CAN LEAD TO EXCITING NEW DISCOVERIES...
C


95
3
2
LOW
ENTROPY
DISCOURAGE
40
30
30
HIGH
ENTROPY
ENCOURAGE
TO FURTHER ENCOURAGE EXPLORATION, A VALUE CALLED ENTROPY IS SUBTRACTED FROM THE LOSS FUNCTION. ENTROPY REFERS TO THE "SPREAD" OF THE ACTION DISTRIBUTION.


+5!
LOOKS LIKE THE GAMBLE PAID OFF!

A SIMPLE POLICY GRADIENT FOX WOULD LOOK AT THE ACTUAL RETURNS FOLLOWING AN ACTION AND TUNE HER POLICY TO MAKE GOOD RETURNS MORE LIKELY.

LOOKS LIKE MY POLICY FROM THAT STATE LED TO LOSING 20 POINTS, I GUESS I BETTER MAKE "C" LESS LIKELY IN THE FUTURE...



BUT WAIT!
IT'S NOT FAIR TO PLACE THE BLAME ON ACTION C. THAT STATE HAD AN ESTIMATED VALUE OF -100, SO TAKING "C" AND ENDING UP WITH -20 WAS ACTUALLY A **RELATIVE IMPROVEMENT** OF 80! I SHOULD MAKE "C" **MORE** LIKELY IN THE FUTURE.

INSTEAD OF TUNING HER POLICY IN RESPONSE TO THE TOTAL RETURNS SHE GOT BY TAKING ACTION C, SHE TUNES HER ACTIONS TO THE RELATIVE RETURNS OF TAKING ACTION C. THIS IS CALLED THE "ADVANTAGE".

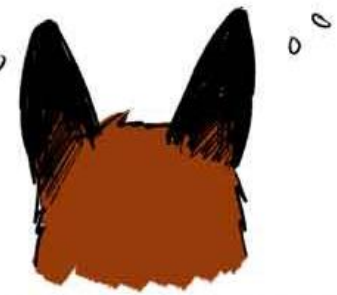
WHAT WE CALLED THE ADVANTAGE IS JUST THE ERROR. AS THE ADVANTAGE, CRANBERRY USES IT TO MAKE ACTIONS THAT WERE SURPRISINGLY GOOD MORE LIKELY. AS THE ERROR, SHE USES THE SAME QUANTITY TO NUDGE HER INNER CRITIC TO MAKE BETTER ESTIMATIONS OF STATE VALUES.

ACTOR USES ADVANTAGE

WOW, THAT WORKED OUT BETTER THAN I THOUGHT, ACTION C MUST HAVE BEEN A GOOD IDEA.

CRITIC USES ERROR

BUT WHY WAS I SURPRISED IN THE FIRST PLACE? I PROBABLY SHOULDN'T HAVE ESTIMATED THAT STATE AS NEGATIVELY AS I DID.



NOW WE CAN SHOW HOW TOTAL LOSS IS COMPUTED—THIS IS THE FUNCTION WE MINIMIZE TO IMPROVE OUR MODEL.

TOTAL LOSS = ACTION LOSS + VALUE LOSS - ENTROPY.

NOTICE WE'RE SHOVING GRADIENTS OF THREE QUALITATIVELY DIFFERENT TYPES BACK THROUGH A SINGLE NN. THIS IS EFFICIENT BUT IT CAN MAKE CONVERGENCE MORE DIFFICULT.

Formal problem

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t^i|s_t^i) Q_{\tau_i,t}$$

то что было в policy gradient

$$Q^\pi(s_t, a_t) = \sum_{t'=t}^{T} E_{\pi_\theta}[r(s_{t'}, a_{t'})|s_t, a_t],$$
$$V^\pi(s_t) = E_{a_t \sim \pi_\theta(a_t|s_t)}[Q^\pi(s_t, a_t)] = \sum_{t'=t}^{T} E_{\pi_\theta}[r(s_{t'}, a_{t'})|s_t]$$

чуть-чуть улучшим,
заменив семплы на матожидание

$$A^\pi(s_t, a_t) = Q^\pi(s_t, a_t) - V^\pi(s_t)$$
$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t^i|s_t^i) A^\pi(s_t^i, a_t^i)$$

заменим Q на A
потому что почему бы и нет?

$$Q^\pi(s_t, a_t) = r(s_t, a_t) + E_{s_{t+1} \sim p(s_{t+1}|s_t, a_t)}[V^\pi(s_{t+1})] \approx r(s_t, a_t) + V^\pi(s_{t+1}),$$
$$A^\pi(s_t^i, a_t^i) = Q^\pi(s_t, a_t) - V^\pi(s_t) \approx r(s_t, a_t) + V^\pi(s_{t+1}) - V^\pi(s_t),$$

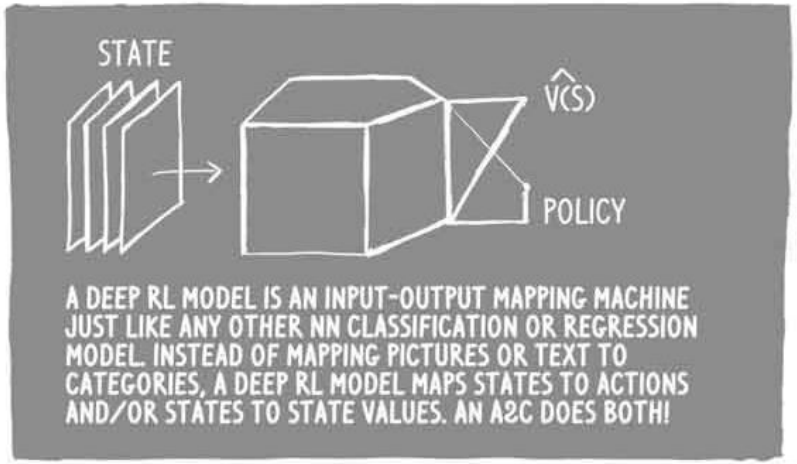
сведём вычисления A к V

$$V^\pi(s_t) = r(s_t, a_t) + V^\pi(s_{t+1})$$
$$V^\pi(s_t) \leftarrow (1 - \beta) V^\pi(s_t) + \beta(r(s_t, a_t) + V^\pi(s_{t+1}))$$

трюк из SARSA

Advantage Actor-Critic (A2C)

1. производим действие $a \sim \pi_\theta(a|s)$, переходим в состояние s' и получаем вознаграждение r;
2. $V^\pi(s) \leftarrow (1-\beta)V^\pi(s) + \beta(r + V^\pi(s'))$;
3. $A_\pi(s,a) \leftarrow r + V^\pi(s') - V^\pi(s)$;
4. $\nabla_\theta J(\theta) \leftarrow \nabla_\theta \log \pi_\theta(a|s) A^\pi(s,a)$;
5. $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$;
6. Если не сошлись к экстремуму, повторить с пункта 1.



Пруфы будут?

Заметим, что если b - константа относительно τ, то

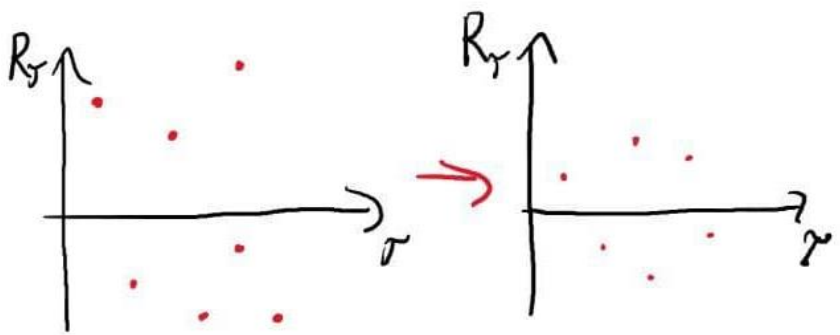
$$E_{\tau\sim p_\theta(\tau)}\left[\nabla_\theta \log p_\theta(\tau)(R_\tau - b)\right] = E_{\tau\sim p_\theta(\tau)}\left[\nabla_\theta \log p_\theta(\tau) R_\tau\right],$$

так как

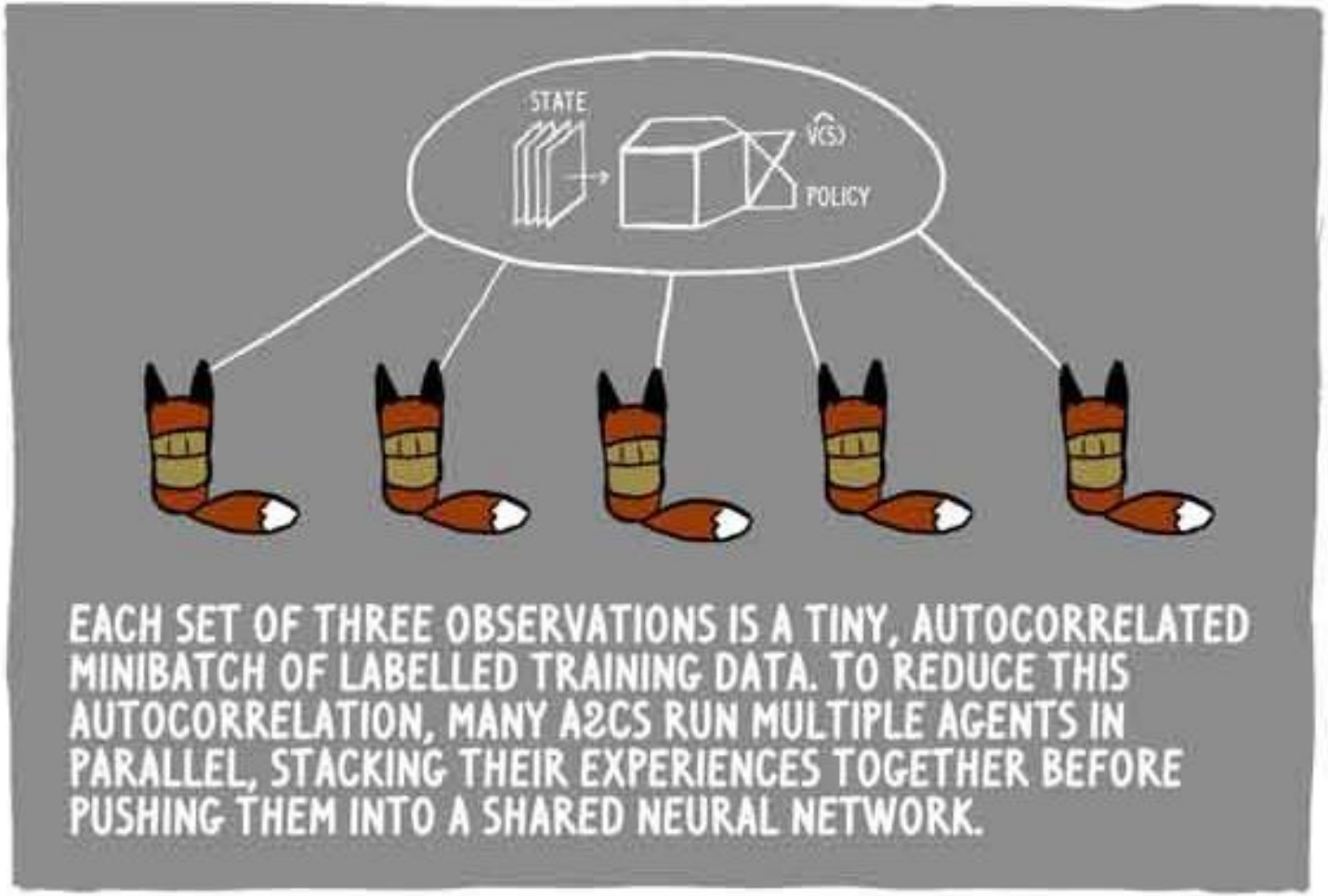
$$E_{\tau\sim p_\theta(\tau)}\left[\nabla_\theta \log p_\theta(\tau) b\right] = \int p_\theta(\tau)\nabla_\theta \log p_\theta(\tau) b d\tau = \int \nabla_\theta p_\theta(\tau) b d\tau = b\nabla_\theta \int p_\theta(\tau) d\tau = b\nabla_\theta 1 = 0,$$

Таким образом, изменение R_τ на константу не меняет оценку $\nabla_\theta J(\theta)$. Однако *дисперсия* $Var_{\tau\sim p_\theta(\tau)}\left[\nabla_\theta \log p_\theta(\tau)(R_\tau - b)\right]$ зависит от b:

$$Var_{\tau\sim p_\theta(\tau)}\left[\nabla_\theta \log p_\theta(\tau)(R_\tau - b)\right] = \underbrace{E_{\tau\sim p_\theta(\tau)}\left[\left(\nabla_\theta \log p_\theta(\tau)(R_\tau - b)\right)^2\right]}_{\text{depends on } b} - \underbrace{E_{\tau\sim p_\theta(\tau)}\left[\nabla_\theta \log p_\theta(\tau)(R_\tau - b)\right]^2}_{=E[\nabla_\theta \log p_\theta(\tau) R_\tau]^2},$$



Asynchronous Advantage Actor-Critic (A3C)



Список источников

- https://habr.com/ru/post/442522/
- https://hackernoon.com/intuitive-rl-intro-to-advantage-actor-critic-a2c-4ff545978752
- https://neerc.ifmo.ru/wiki/index.php?title=Методы_policy_gradient_и_алгоритм_асинхронного_актора-критика
- https://www.machinelearningmastery.ru/the-idea-behind-actor-critics-and-how-a2c-and-a3c-improve-them-6dd7dfd0acb8/
- https://github.com/yandexdataschool/Practical_RL