МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МОЭВМ

ОТЧЕТ
по лабораторной работе №1
по дисциплине «Обучение с подкреплением»
Тема: Реализация DQN для среды CartPole-v1

Студент гр. 0306 _____ Голубев А.Н.

Преподаватель _____ Глазунов С.А.

Санкт-Петербург
2025 г.

Цель работы.

Реализация DQN для среды CartPole-v1. Исследование влияния различных параметров: архитектура сети, значения gamma и epsilon_decay, влияние epsilon на скорость обучения

Задание.

1. Реализация DQN
2. Измените архитектуру нейросети (например, добавьте слои).
3. Попробуйте разные значения gamma и epsilon_decay.
4. Проведите исследование как изначальное значение epsilon влияет на скорость обучения

Выполнение работы.

1. Реализация DQN

Основным классом программы является Agent, который, в свою очередь, использует класс DQN, представляющий собой реализацию нейронной сети. Входными параметрами сети является пространство наблюдений среды CartPole-1 (см. рис. 1). На выходе сеть выдает численный показатель эффективности двух вариантов действий:

- движение вправо (индекс 1)
- движение влево (индекс 0)

Observation Space

The observation is a `ndarray` with shape `(4,)` with the values corresponding to the following positions and velocities:

Num	Observation	Min	Max
0	Cart Position	-4.8	4.8
1	Cart Velocity	-Inf	Inf
2	Pole Angle	~ -0.418 rad (-24°)	~ 0.418 rad (24°)
3	Pole Angular Velocity	-Inf	Inf

Рис. 1 – Описание состояния среды CartPole-v1

Гиперпараметры обучаемой сети считываются из файла «hyperparameters.yml». Пример данного файла представлен на рисунке 2.

```yaml
env_id: CartPole-v1
replay_memory_size: 100000
mini_batch_size: 64
epsilon_init: 1
epsilon_decay: 0.9
epsilon_min: 0.01
network_sync_rate: 100
alpha: 0.001
gamma: 0.3
stop_on_reward: 10000
fc1_nodes: 128
train_episodes: 600
test_episodes: 20
```

Рис. 2 — содержание файла hyperparameters.yml

В файле представлены следующие параметры:

- replay_memory_size – размер буфера ReplayBuffer
- mini_batch_size – размер батча
- epsilon_init – начальное значение epsilon
- epsilon_decay – скорость уменьшения epsilon
- epsilon_min – минимальное значение epsilon
- network_sync_rate – порог количества шагов, при преодолении которого происходит синхронизация нынешней политики (policy_net) и целевой сети (target_net)
- alpha – скорость обучения
- gamma – скорость дисконтирования
- stop_on_reward – порог значения награды, при преодолении которого происходит прекращение обучения
- fc1_nodes – количество узлов в скрытых слоях
- train_episodes – количество эпизодов при тренировке
- test_episodes – количество эпизодов при тестировании модели

Обучение модели происходит с помощью ReplayMemory – буфер, который хранит информацию о изменениях в среде, которые наблюдает агент. Обучение происходит на протяжении нескольких эпизодов.

На каждом эпизоде происходит следующее:

1. В буфер ReplayMemory заносятся данные формата <состояние, действие, новое состояние, награда, симуляция завершена>
2. Если суммарная награда за эпизод больше предыдущего зафиксированного значения, то модель обновляется
3. Если объем данных в ReplayBuffer превосходит размер батча, то из буфера случайным образом отбирается набор данных размером с один батч. Происходит оптимизация сети:

1. С помощью уравнения Беллмана расчитываем целевое Q-значение (награду)
2. Вычисляем нынешнее Q-значение, соответствующее нынешней политике (policy net)
3. Вычисляем значение потери (в данной работе для вычисления потери используется средняя квадратическая ошибка)
4. Оптимизируем модель с помощью обратного распространения ошибки (backpropagation)

4. Если количество совершенных шагов превышает значение network_sync_rate, осуществляется синхронизация целевой сети с нынешней политикой, и счетчик шагов обнуляется

2. Влияние изменения архитектуры сети на результаты

Для выполнения данного пункта было создано две разные архитектуры Deep Q сети (см. рис. 3).

```
class DQN(nn.Module):
    def __init__(self, state_dim, action_dim, hidden_dim=256, size='normal'):
        super(DQN, self).__init__()
        if size == 'normal':
            self.net = nn.Sequential(
                nn.Linear(state_dim, hidden_dim),
                nn.ReLU(),
                nn.Linear(hidden_dim, hidden_dim),
                nn.ReLU(),
                nn.Linear(hidden_dim, action_dim)
            )
        else:
            self.net = nn.Sequential(
                nn.Linear(state_dim, hidden_dim // 2),
                nn.ReLU(),
                nn.Linear(hidden_dim // 2, hidden_dim),
                nn.ReLU(),
                nn.Linear(hidden_dim, hidden_dim),
                nn.ReLU(),
                nn.Linear(hidden_dim, hidden_dim // 2),
                nn.ReLU(),
                nn.Linear(hidden_dim // 2, action_dim)
            )

    def forward(self, x):
        return self.net(x)
```

Рис. 3 — архитектуры сети

Был осуществлен запуск обучения при разных конфигурациях

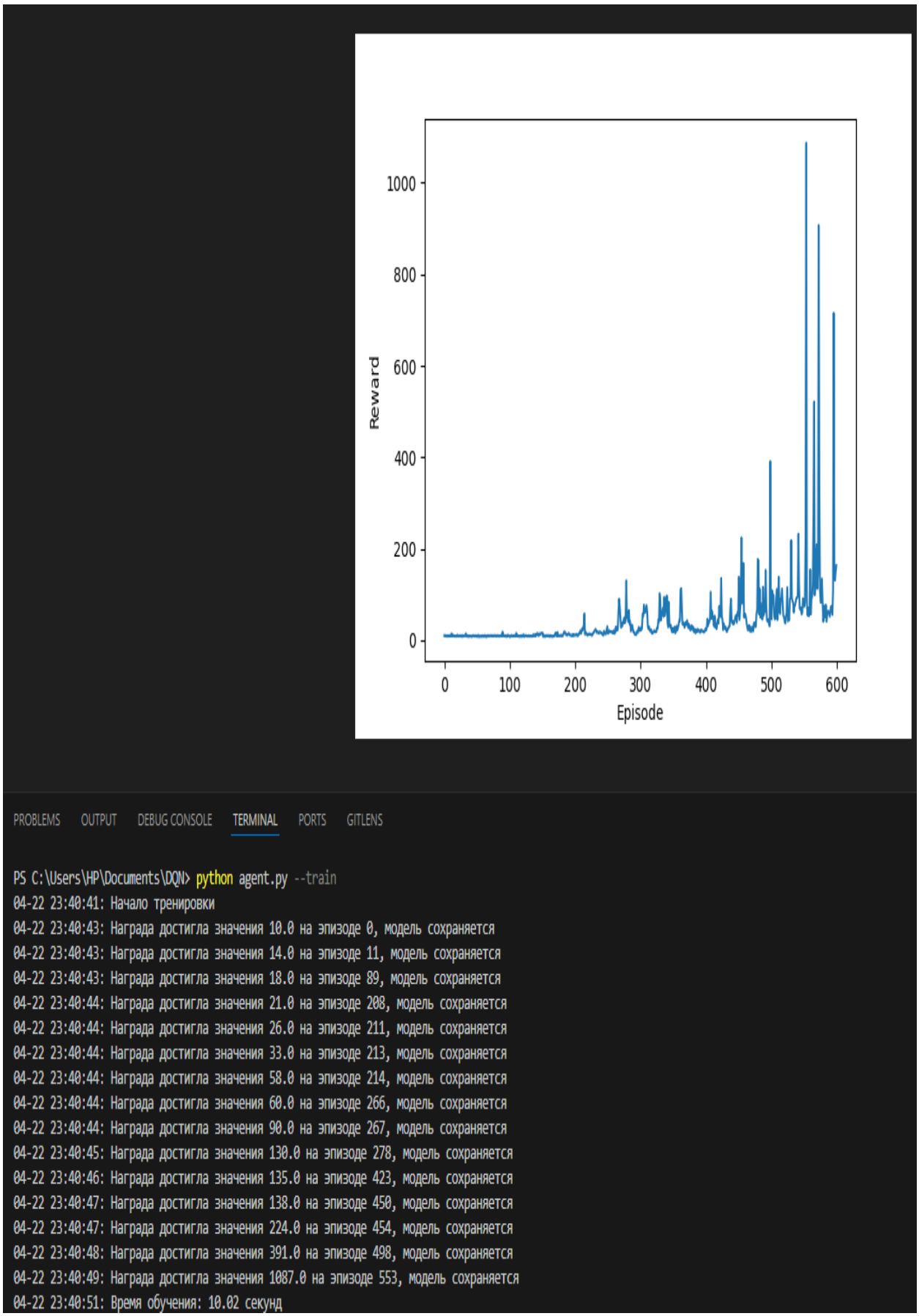
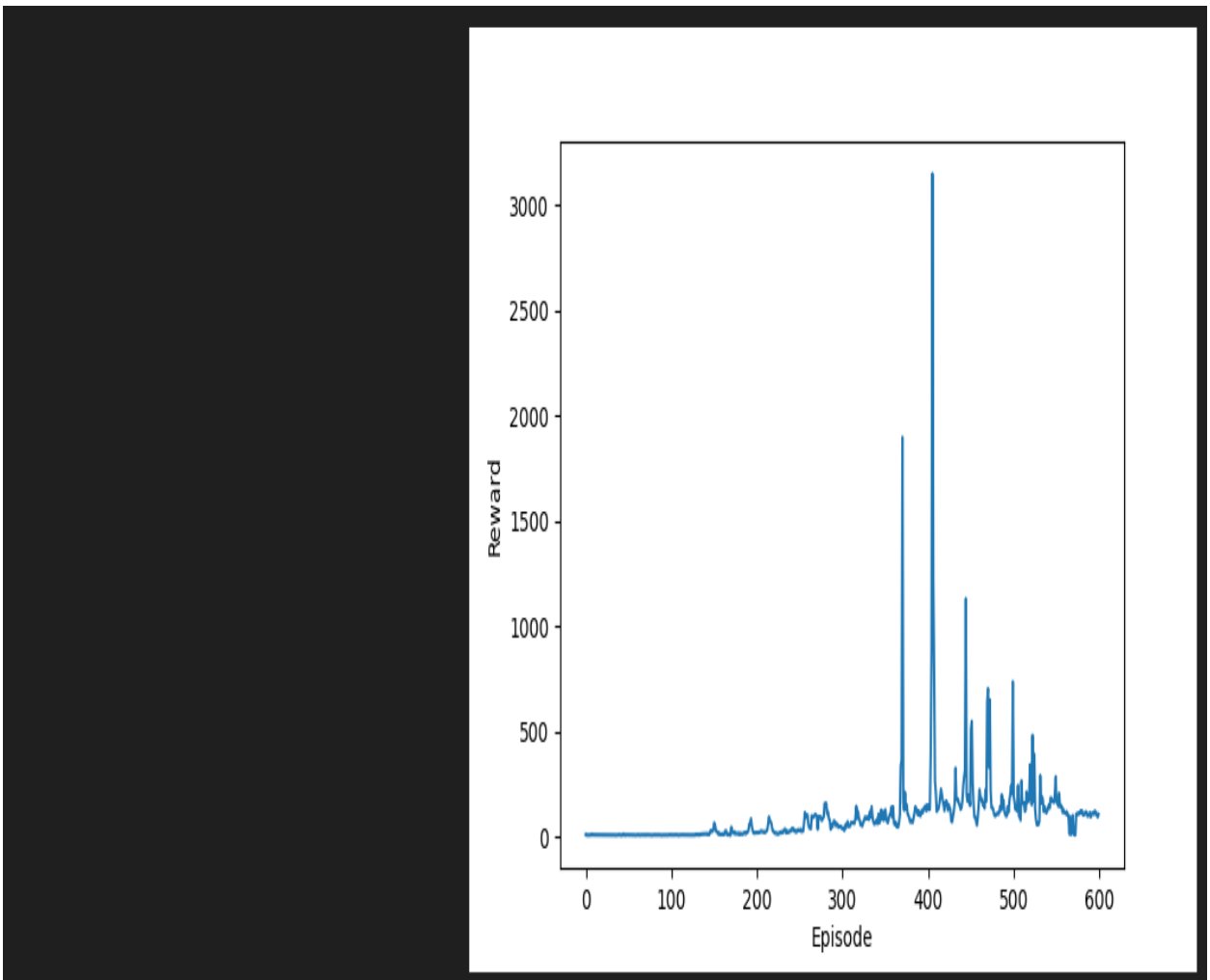


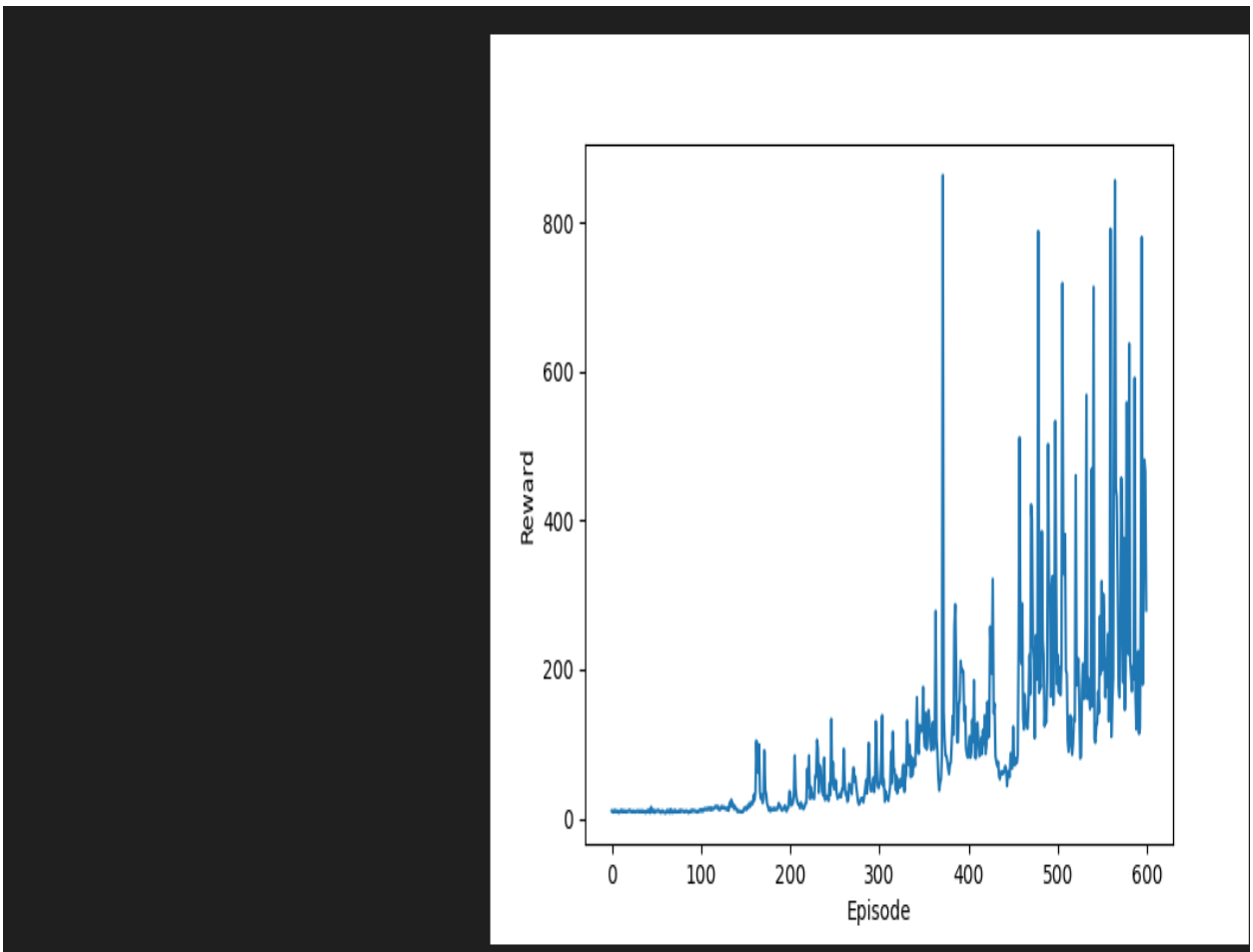
Рис. 4 — результат обучения (size='normal', hidden_dim=128)



PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS GITLENS

```
04-22 23:43:52: Награда достигла значения 11.0 на эпизоде 0, модель сохраняется
04-22 23:43:52: Награда достигла значения 13.0 на эпизоде 7, модель сохраняется
04-22 23:43:53: Награда достигла значения 14.0 на эпизоде 139, модель сохраняется
04-22 23:43:53: Награда достигла значения 15.0 на эпизоде 140, модель сохраняется
04-22 23:43:53: Награда достигла значения 20.0 на эпизоде 145, модель сохраняется
04-22 23:43:53: Награда достигла значения 30.0 на эпизоде 146, модель сохраняется
04-22 23:43:53: Награда достигла значения 34.0 на эпизоде 149, модель сохраняется
04-22 23:43:53: Награда достигла значения 66.0 на эпизоде 150, модель сохраняется
04-22 23:43:53: Награда достигла значения 86.0 на эпизоде 193, модель сохраняется
04-22 23:43:54: Награда достигла значения 95.0 на эпизоде 214, модель сохраняется
04-22 23:43:54: Награда достигла значения 116.0 на эпизоде 256, модель сохраняется
04-22 23:43:55: Награда достигла значения 158.0 на эпизоде 279, модель сохраняется
04-22 23:43:55: Награда достигла значения 159.0 на эпизоде 281, модель сохраняется
04-22 23:43:58: Награда достигла значения 338.0 на эпизоде 368, модель сохраняется
04-22 23:43:58: Награда достигла значения 363.0 на эпизоде 369, модель сохраняется
04-22 23:43:58: Награда достигла значения 1896.0 на эпизоде 370, модель сохраняется
04-22 23:44:01: Награда достигла значения 3147.0 на эпизоде 405, модель сохраняется
04-22 23:44:13: Время обучения: 23.28 секунд
```

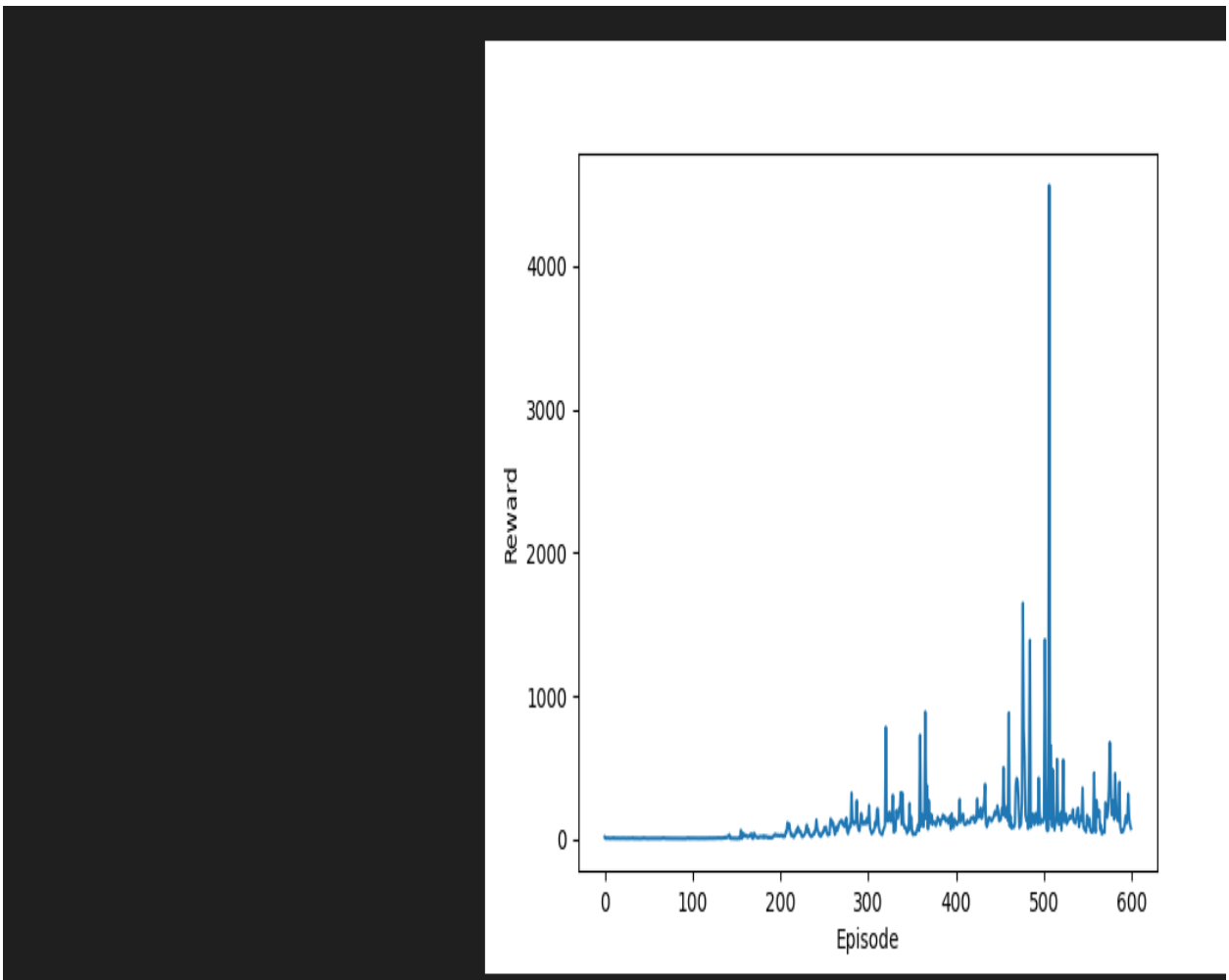
Рис. 5 — результат обучения (size=’big’, hidden_dim=128)



PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS GITLENS

```
04-22 23:47:02: Начало тренировки
04-22 23:47:03: Награда достигла значения 10.0 на эпизоде 0, модель сохраняется
04-22 23:47:03: Награда достигла значения 11.0 на эпизоде 3, модель сохраняется
04-22 23:47:04: Награда достигла значения 15.0 на эпизоде 44, модель сохраняется
04-22 23:47:04: Награда достигла значения 16.0 на эпизоде 116, модель сохраняется
04-22 23:47:04: Награда достигла значения 21.0 на эпизоде 132, модель сохраняется
04-22 23:47:04: Награда достигла значения 25.0 на эпизоде 134, модель сохраняется
04-22 23:47:04: Награда достигла значения 31.0 на эпизоде 159, модель сохраняется
04-22 23:47:04: Награда достигла значения 36.0 на эпизоде 161, модель сохраняется
04-22 23:47:05: Награда достигла значения 104.0 на эпизоде 162, модель сохраняется
04-22 23:47:05: Награда достигла значения 105.0 на эпизоде 230, модель сохраняется
04-22 23:47:06: Награда достигла значения 133.0 на эпизоде 246, модель сохраняется
04-22 23:47:06: Награда достигла значения 138.0 на эпизоде 303, модель сохраняется
04-22 23:47:07: Награда достигла значения 162.0 на эпизоде 342, модель сохраняется
04-22 23:47:07: Награда достигла значения 176.0 на эпизоде 349, модель сохраняется
04-22 23:47:08: Награда достигла значения 278.0 на эпизоде 363, модель сохраняется
04-22 23:47:08: Награда достигла значения 863.0 на эпизоде 371, модель сохраняется
04-22 23:47:21: Время обучения: 19.02 секунд
```

Рис. 6 — результат обучения (size = ‘normal’, hidden_dim=256)



PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS GITLENS

```
04-22 23:47:08: Награда достигла значения 863.0 на эпизоде 371, модель сохраняется
04-22 23:47:21: Время обучения: 19.02 секунд
PS C:\Users\HP\Documents\DQN> python agent.py --train
04-22 23:48:33: Начало тренировки
04-22 23:48:35: Награда достигла значения 17.0 на эпизоде 0, модель сохраняется
04-22 23:48:36: Награда достигла значения 21.0 на эпизоде 140, модель сохраняется
04-22 23:48:36: Награда достигла значения 33.0 на эпизоде 142, модель сохраняется
04-22 23:48:36: Награда достигла значения 64.0 на эпизоде 155, модель сохраняется
04-22 23:48:37: Награда достигла значения 117.0 на эпизоде 208, модель сохраняется
04-22 23:48:38: Награда достигла значения 136.0 на эпизоде 241, модель сохраняется
04-22 23:48:38: Награда достигла значения 142.0 на эпизоде 257, модель сохраняется
04-22 23:48:39: Награда достигла значения 150.0 на эпизоде 275, модель сохраняется
04-22 23:48:39: Награда достигла значения 324.0 на эпизоде 281, модель сохраняется
04-22 23:48:41: Награда достигла значения 783.0 на эпизоде 320, модель сохраняется
04-22 23:48:44: Награда достигла значения 889.0 на эпизоде 365, модель сохраняется
04-22 23:48:51: Награда достигла значения 1647.0 на эпизоде 476, модель сохраняется
04-22 23:48:55: Награда достигла значения 4560.0 на эпизоде 506, модель сохраняется
04-22 23:49:01: Время обучения: 28.25 секунд
```

Рис. 7 — результат обучения (size=’big’, hidden_dim=256)

По полученным результатам можно заметить, что при усложнении архитектуры сети происходит увеличение максимального значения награды (исключением оказался случай на рисунке 6 в сравнении с рисунком 4). Однако, вместе с этим, растет время обучения. Также можно заметить, что при использовании «большой» архитектуры сети на графике значений наград наблюдается меньше крупных скачков значений.

2. Влияния gamma и epsilon_decay

2.1 Изменение gamma

Параметр gamma (или discount rate) отвечает за баланс между важностью сиюминутной и будущих наград. Меньшее значение данного параметра делает стратегию более жадной. На рисунках показаны результаты обучения при разных значениях gamma.

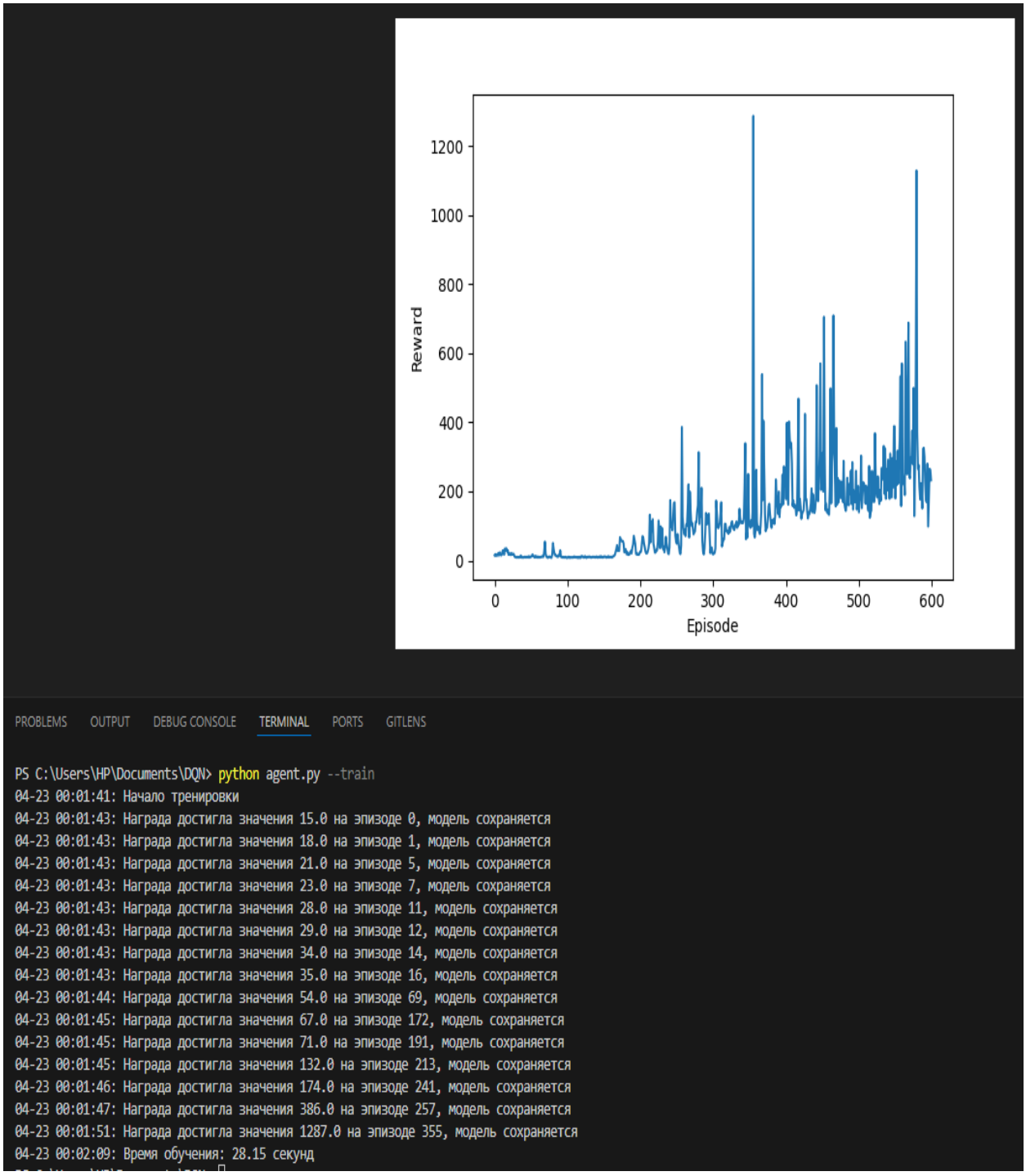
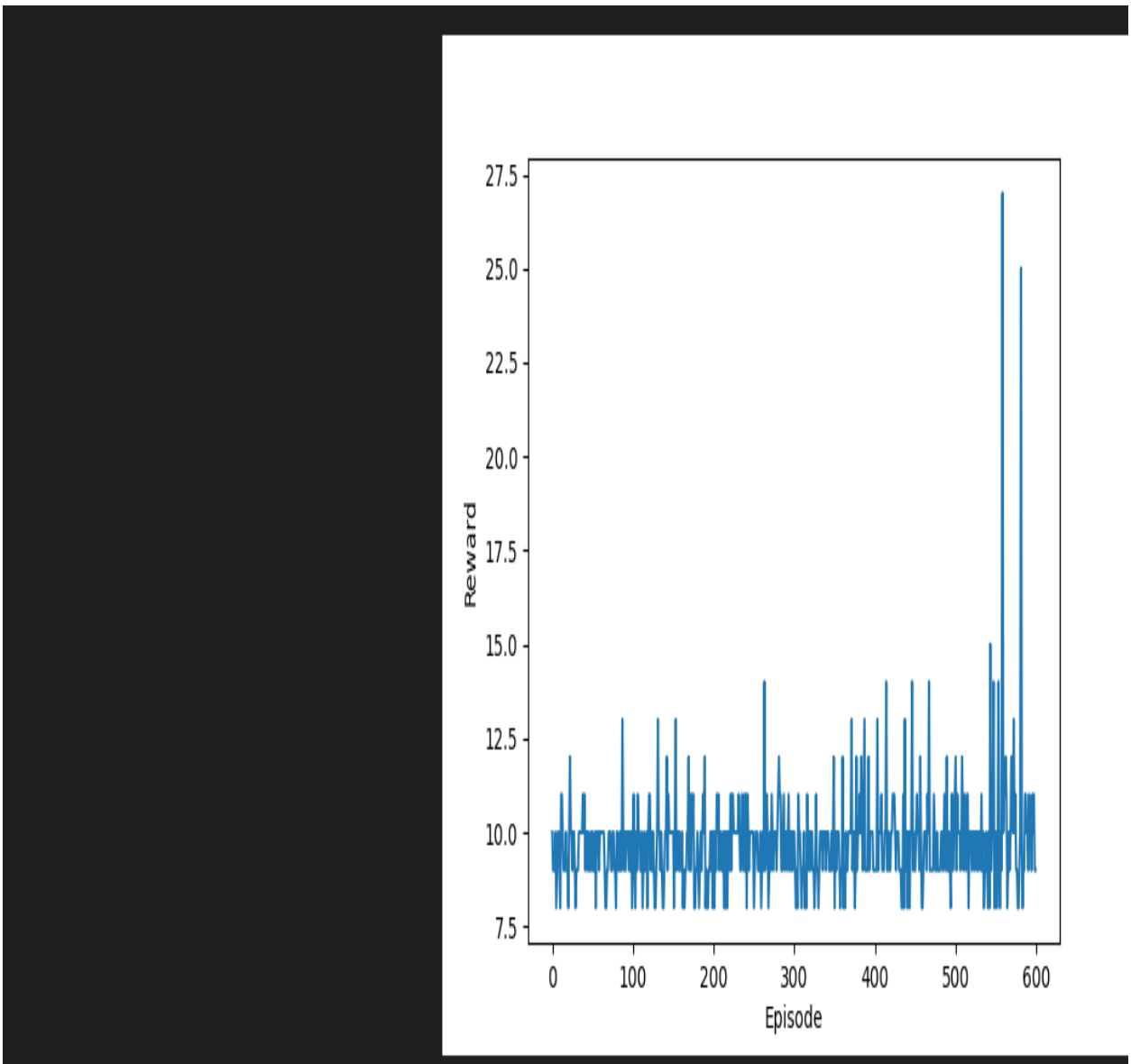


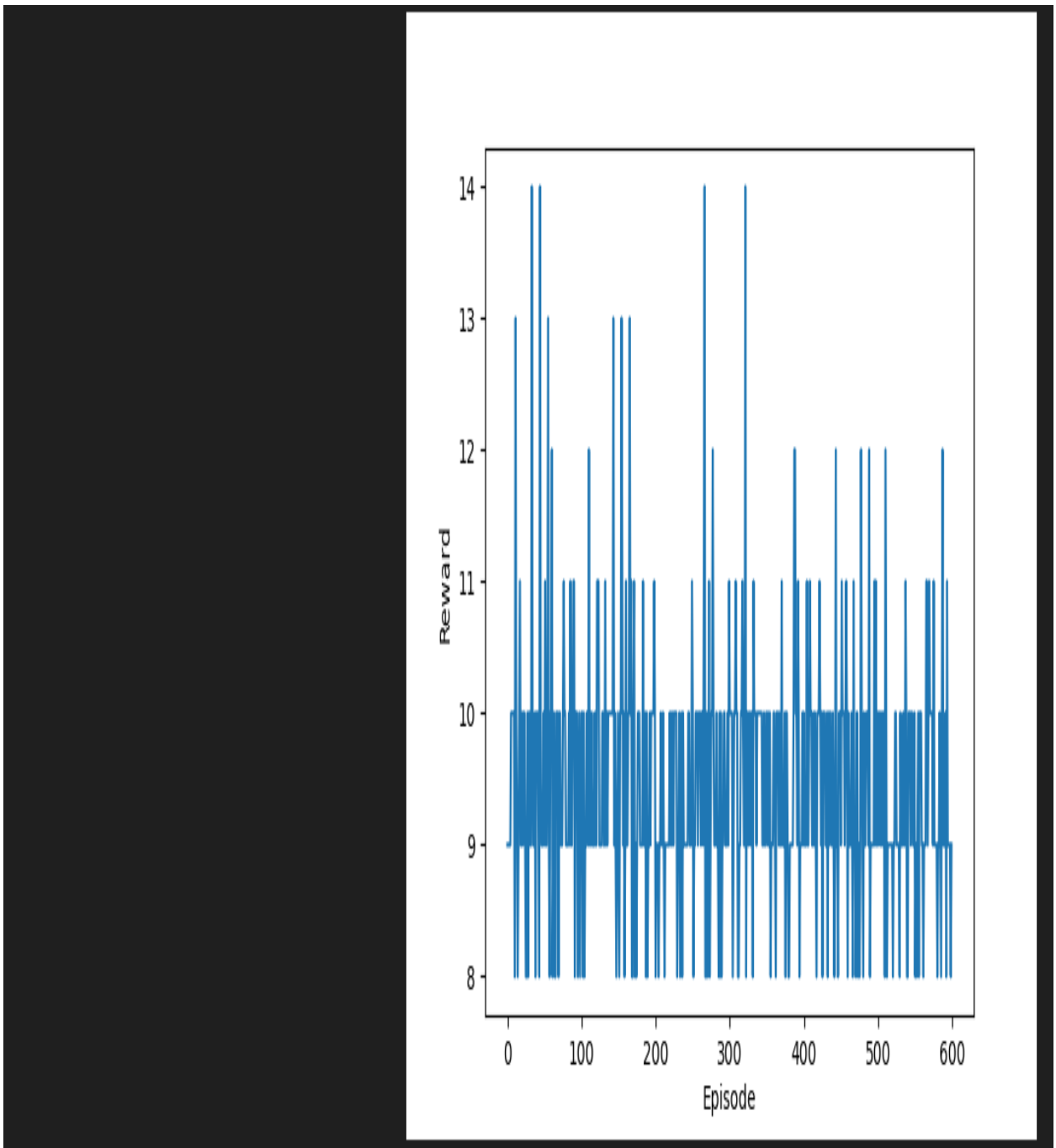
Рис. 8 — результат обучения (gamma=0.99)



PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS GITLENS

```
PS C:\Users\HP\Documents\DQN> python agent.py --train
04-23 00:06:14: Начало тренировки
04-23 00:06:15: Награда достигла значения 10.0 на эпизоде 0, модель сохраняется
04-23 00:06:16: Награда достигла значения 11.0 на эпизоде 11, модель сохраняется
04-23 00:06:16: Награда достигла значения 12.0 на эпизоде 22, модель сохраняется
04-23 00:06:16: Награда достигла значения 13.0 на эпизоде 87, модель сохраняется
04-23 00:06:17: Награда достигла значения 14.0 на эпизоде 263, модель сохраняется
04-23 00:06:19: Награда достигла значения 15.0 на эпизоде 543, модель сохраняется
04-23 00:06:19: Награда достигла значения 27.0 на эпизоде 558, модель сохраняется
04-23 00:06:20: Время обучения: 6.01 секунд
```

Рис. 9 — результат обучения (gamma=0.5)



PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS GITLENS

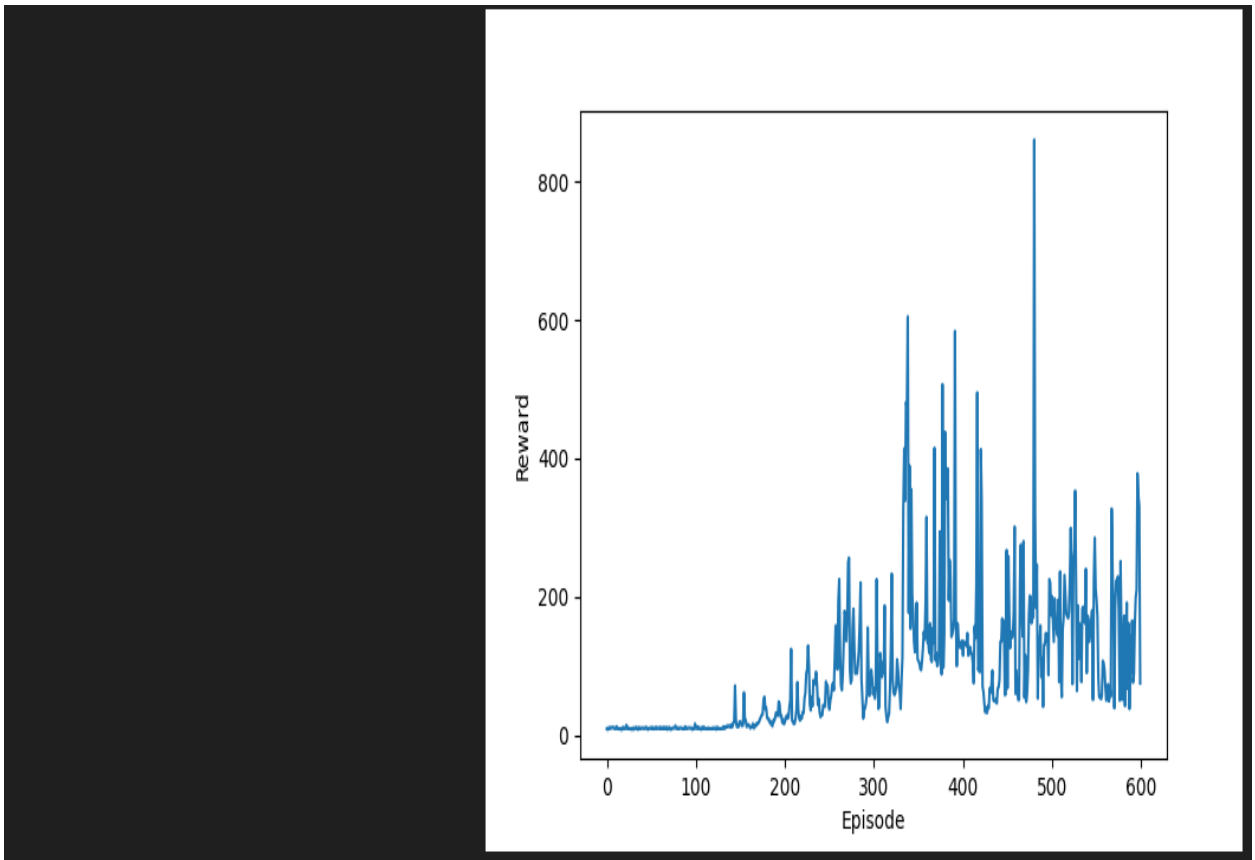
```
PS C:\Users\HP\Documents\DQN> python agent.py --train
04-23 00:08:14: Начало тренировки
04-23 00:08:15: Награда достигла значения 9.0 на эпизоде 0, модель сохраняется
04-23 00:08:16: Награда достигла значения 10.0 на эпизоде 5, модель сохраняется
04-23 00:08:16: Награда достигла значения 13.0 на эпизоде 11, модель сохраняется
04-23 00:08:16: Награда достигла значения 14.0 на эпизоде 33, модель сохраняется
04-23 00:08:20: Время обучения: 6.0 секунд
```

Рис. 10 — результат обучения (gamma=0.1)

По результатам обучения можно заметить, что наилучшие значения награды были достигнуты при gamma=0.99. При снижении значения gamma происходит снижение максимального значения награды, а время обучения снижается.

2.2. Изменение epsilon_decay

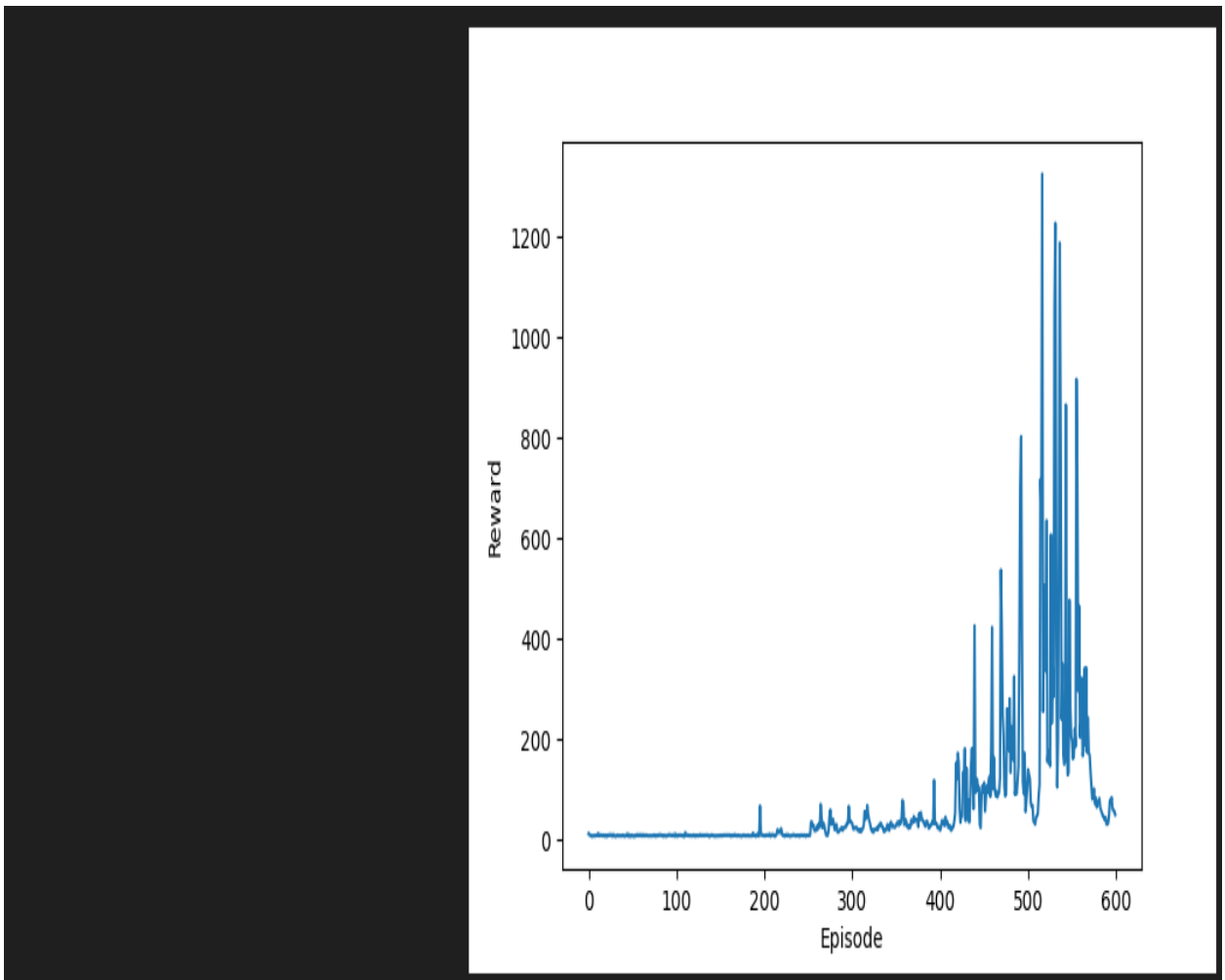
Параметр epsilon_decay отвечает за скорость снижения значения epsilon в процессе обучения. На следующих рисунках представлены результаты обучения при разных значениях epsilon_decay.



PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS GITLENS

```
04-23 00:18:54: Награда достигла значения 9.0 на эпизоде 0, модель сохраняется
04-23 00:18:55: Награда достигла значения 11.0 на эпизоде 3, модель сохраняется
04-23 00:18:55: Награда достигла значения 12.0 на эпизоде 11, модель сохраняется
04-23 00:18:55: Награда достигла значения 14.0 на эпизоде 22, модель сохраняется
04-23 00:18:55: Награда достигла значения 16.0 на эпизоде 99, модель сохраняется
04-23 00:18:56: Награда достигла значения 17.0 на эпизоде 142, модель сохраняется
04-23 00:18:56: Награда достигла значения 19.0 на эпизоде 143, модель сохраняется
04-23 00:18:56: Награда достигла значения 71.0 на эпизоде 144, модель сохраняется
04-23 00:18:56: Награда достигла значения 124.0 на эпизоде 207, модель сохраняется
04-23 00:18:57: Награда достигла значения 129.0 на эпизоде 226, модель сохраняется
04-23 00:18:57: Награда достигла значения 158.0 на эпизоде 257, модель сохраняется
04-23 00:18:58: Награда достигла значения 193.0 на эпизоде 260, модель сохраняется
04-23 00:18:58: Награда достигла значения 225.0 на эпизоде 261, модель сохраняется
04-23 00:18:58: Награда достигла значения 250.0 на эпизоде 271, модель сохраняется
04-23 00:18:58: Награда достигла значения 256.0 на эпизоде 272, модель сохраняется
04-23 00:19:00: Награда достигла значения 328.0 на эпизоде 333, модель сохраняется
04-23 00:19:00: Награда достигла значения 414.0 на эпизоде 334, модель сохраняется
04-23 00:19:01: Награда достигла значения 481.0 на эпизоде 336, модель сохраняется
04-23 00:19:01: Награда достигла значения 605.0 на эпизоде 338, модель сохраняется
04-23 00:19:09: Награда достигла значения 861.0 на эпизоде 480, модель сохраняется
04-23 00:19:15: Время обучения: 22.83 секунд
```

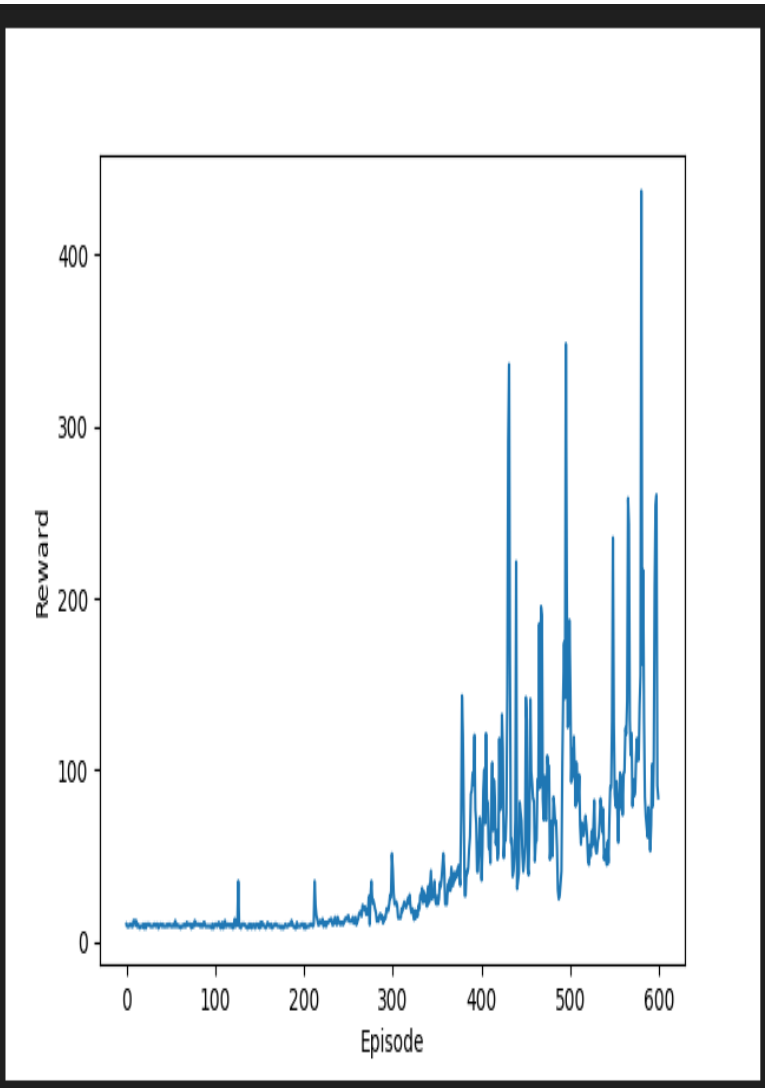
Рис. 11 — результат обучения (epsilon_decay=0.9995)



PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS GITLENS

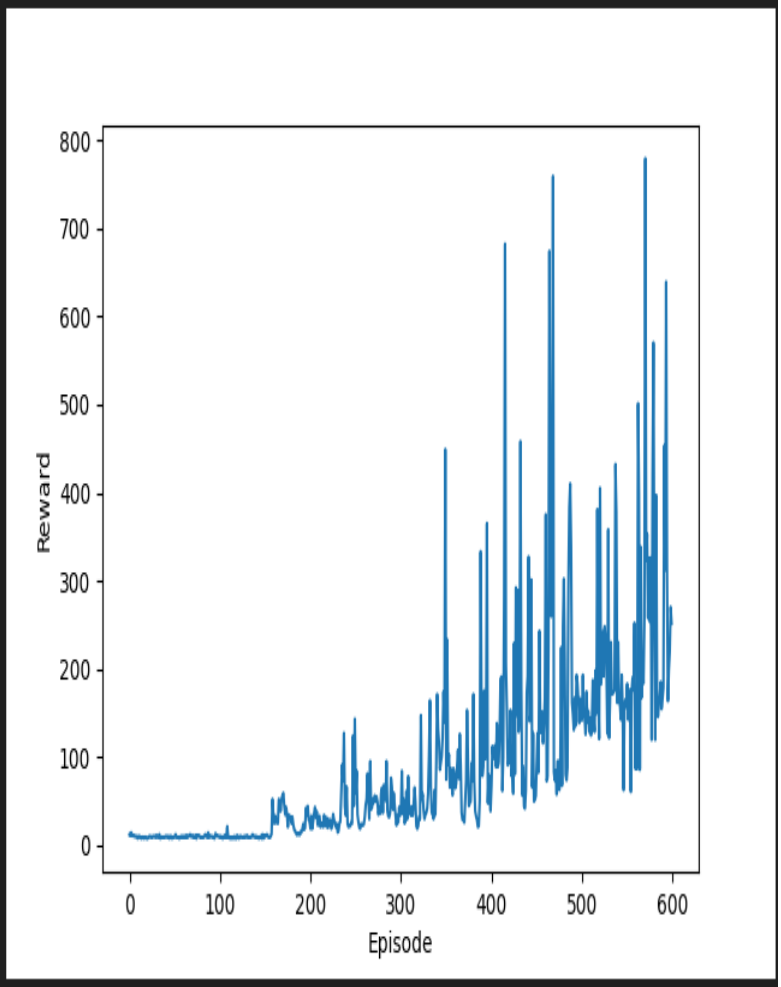
```
04-23 00:19:01: Награда достигла значения 481.0 на эпизоде 336, модель сохраняется
04-23 00:19:01: Награда достигла значения 605.0 на эпизоде 338, модель сохраняется
04-23 00:19:09: Награда достигла значения 861.0 на эпизоде 480, модель сохраняется
04-23 00:19:15: Время обучения: 22.83 секунд
PS C:\Users\HP\Documents\DQN> python agent.py --train
04-23 00:21:02: Начало тренировки
04-23 00:21:04: Награда достигла значения 13.0 на эпизоде 0, модель сохраняется
04-23 00:21:05: Награда достигла значения 15.0 на эпизоде 110, модель сохраняется
04-23 00:21:05: Награда достигла значения 68.0 на эпизоде 195, модель сохраняется
04-23 00:21:06: Награда достигла значения 71.0 на эпизоде 264, модель сохраняется
04-23 00:21:07: Награда достигла значения 79.0 на эпизоде 357, модель сохраняется
04-23 00:21:08: Награда достигла значения 119.0 на эпизоде 393, модель сохраняется
04-23 00:21:08: Награда достигла значения 153.0 на эпизоде 418, модель сохраняется
04-23 00:21:08: Награда достигла значения 173.0 на эпизоде 420, модель сохраняется
04-23 00:21:08: Награда достигла значения 182.0 на эпизоде 428, модель сохраняется
04-23 00:21:09: Награда достигла значения 426.0 на эпизоде 439, модель сохраняется
04-23 00:21:10: Награда достигла значения 537.0 на эпизоде 469, модель сохраняется
04-23 00:21:12: Награда достигла значения 700.0 на эпизоде 491, модель сохраняется
04-23 00:21:12: Награда достигла значения 802.0 на эпизоде 492, модель сохраняется
04-23 00:21:14: Награда достигла значения 1324.0 на эпизоде 516, модель сохраняется
04-23 00:21:21: Время обучения: 19.58 секунд
```

Рис. 12 — результат обучения (epsilon_decay=0.5)




```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS  GITLENS

PS C:\Users\HP\Documents\DQN> python agent.py --train
04-23 00:24:46: Начало тренировки
04-23 00:24:49: Награда достигла значения 10.0 на эпизоде 0, модель сохраняется
04-23 00:24:49: Награда достигла значения 12.0 на эпизоде 8, модель сохраняется
04-23 00:24:49: Награда достигла значения 13.0 на эпизоде 122, модель сохраняется
04-23 00:24:49: Награда достигла значения 35.0 на эпизоде 126, модель сохраняется
04-23 00:24:51: Награда достигла значения 51.0 на эпизоде 299, модель сохраняется
04-23 00:24:52: Награда достигла значения 74.0 на эпизоде 377, модель сохраняется
04-23 00:24:52: Награда достигла значения 143.0 на эпизоде 378, модель сохраняется
04-23 00:24:53: Награда достигла значения 145.0 на эпизоде 429, модель сохраняется
04-23 00:24:53: Награда достигла значения 291.0 на эпизоде 430, модель сохраняется
04-23 00:24:54: Награда достигла значения 336.0 на эпизоде 431, модель сохраняется
04-23 00:24:56: Награда достигла значения 348.0 на эпизоде 495, модель сохраняется
04-23 00:24:58: Награда достигла значения 437.0 на эпизоде 580, модель сохраняется
04-23 00:24:59: Время обучения: 12.84 секунд
```

Рис. 13 — результат обучения (epsilon_decay=0.3)



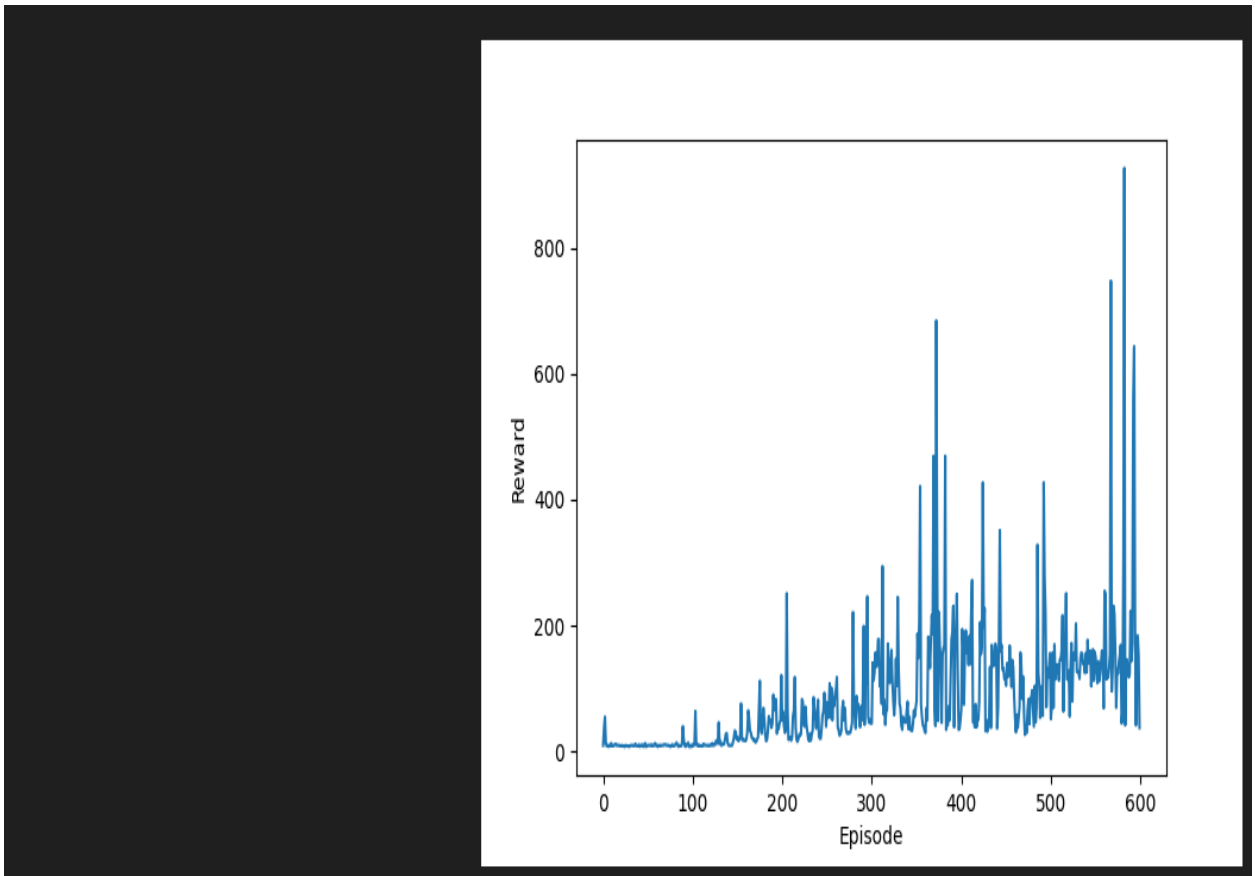

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS  GITLENS

PS C:\Users\HP\Documents\DQN> python agent.py --train
04-23 00:29:22: Начало тренировки
04-23 00:29:24: Награда достигла значения 12.0 на эпизоде 0, модель сохраняется
04-23 00:29:24: Награда достигла значения 14.0 на эпизоде 2, модель сохраняется
04-23 00:29:25: Награда достигла значения 21.0 на эпизоде 108, модель сохраняется
04-23 00:29:25: Награда достигла значения 52.0 на эпизоде 158, модель сохраняется
04-23 00:29:25: Награда достигла значения 56.0 на эпизоде 169, модель сохраняется
04-23 00:29:25: Награда достигла значения 59.0 на эпизоде 170, модель сохраняется
04-23 00:29:26: Награда достигла значения 91.0 на эпизоде 235, модель сохраняется
04-23 00:29:26: Награда достигла значения 127.0 на эпизоде 237, модель сохраняется
04-23 00:29:26: Награда достигла значения 143.0 на эпизоде 249, модель сохраняется
04-23 00:29:28: Награда достигла значения 147.0 на эпизоде 322, модель сохраняется
04-23 00:29:28: Награда достигла значения 164.0 на эпизоде 332, модель сохраняется
04-23 00:29:28: Награда достигла значения 171.0 на эпизоде 340, модель сохраняется
04-23 00:29:29: Награда достигла значения 174.0 на эпизоде 347, модель сохраняется
04-23 00:29:29: Награда достигла значения 449.0 на эпизоде 349, модель сохраняется
04-23 00:29:31: Награда достигла значения 682.0 на эпизоде 415, модель сохраняется
04-23 00:29:35: Награда достигла значения 759.0 на эпизоде 468, модель сохраняется
04-23 00:29:41: Награда достигла значения 779.0 на эпизоде 570, модель сохраняется
04-23 00:29:44: Время обучения: 21.68 секунд
```

Рис. 14 — результаты обучения (epsilon_decay=0.4)

По результатам обучения в плане максимального значения награды лучше всего себя проявил epsilon_decay=0.5; в плане времени обучения — epsilon_decay=0.3.

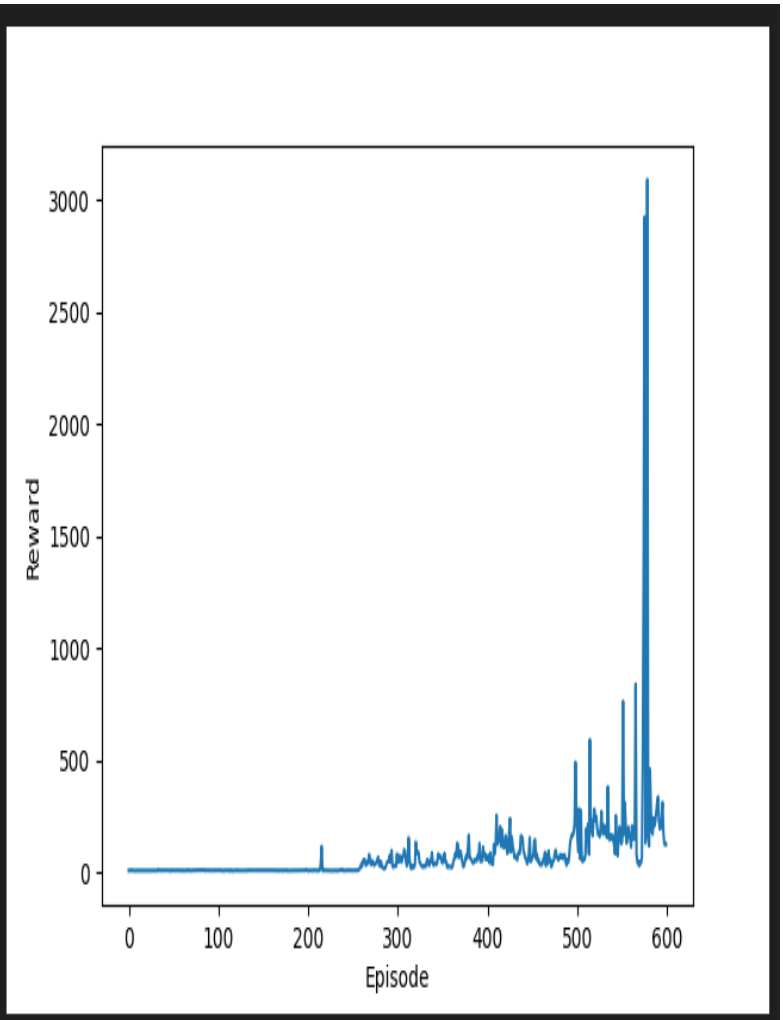
3. Влияние начального значения epsilon на обучение

Параметр epsilon влияет на изначальную частоту выбора случайных действий вместо действий модели. На следующих рисунках представлены результаты обучения при разных значениях epsilon.




```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS  GITLENS

    self.state = np.array((x, x_dot, theta, theta_dot), dtype=np.float64)
                 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
KeyboardInterrupt
PS C:\Users\HP\Documents\DQN> python agent.py --train
04-23 00:56:24: Начало тренировки
04-23 00:56:26: Награда достигла значения 10.0 на эпизоде 0, модель сохраняется
04-23 00:56:26: Награда достигла значения 37.0 на эпизоде 1, модель сохраняется
04-23 00:56:26: Награда достигла значения 55.0 на эпизоде 2, модель сохраняется
04-23 00:56:27: Награда достигла значения 64.0 на эпизоде 103, модель сохраняется
04-23 00:56:27: Награда достигла значения 76.0 на эпизоде 154, модель сохраняется
04-23 00:56:27: Награда достигла значения 112.0 на эпизоде 175, модель сохраняется
04-23 00:56:28: Награда достигла значения 121.0 на эпизоде 199, модель сохраняется
04-23 00:56:28: Награда достигла значения 251.0 на эпизоде 205, модель сохраняется
04-23 00:56:31: Награда достигла значения 294.0 на эпизоде 312, модель сохраняется
04-23 00:56:32: Награда достигла значения 421.0 на эпизоде 354, модель сохраняется
04-23 00:56:33: Награда достигла значения 469.0 на эпизоде 369, модель сохраняется
04-23 00:56:33: Награда достигла значения 684.0 на эпизоде 372, модель сохраняется
04-23 00:56:42: Награда достигла значения 747.0 на эпизоде 567, модель сохраняется
04-23 00:56:43: Награда достигла значения 926.0 на эпизоде 582, модель сохраняется
04-23 00:56:44: Время обучения: 20.75 секунд
```

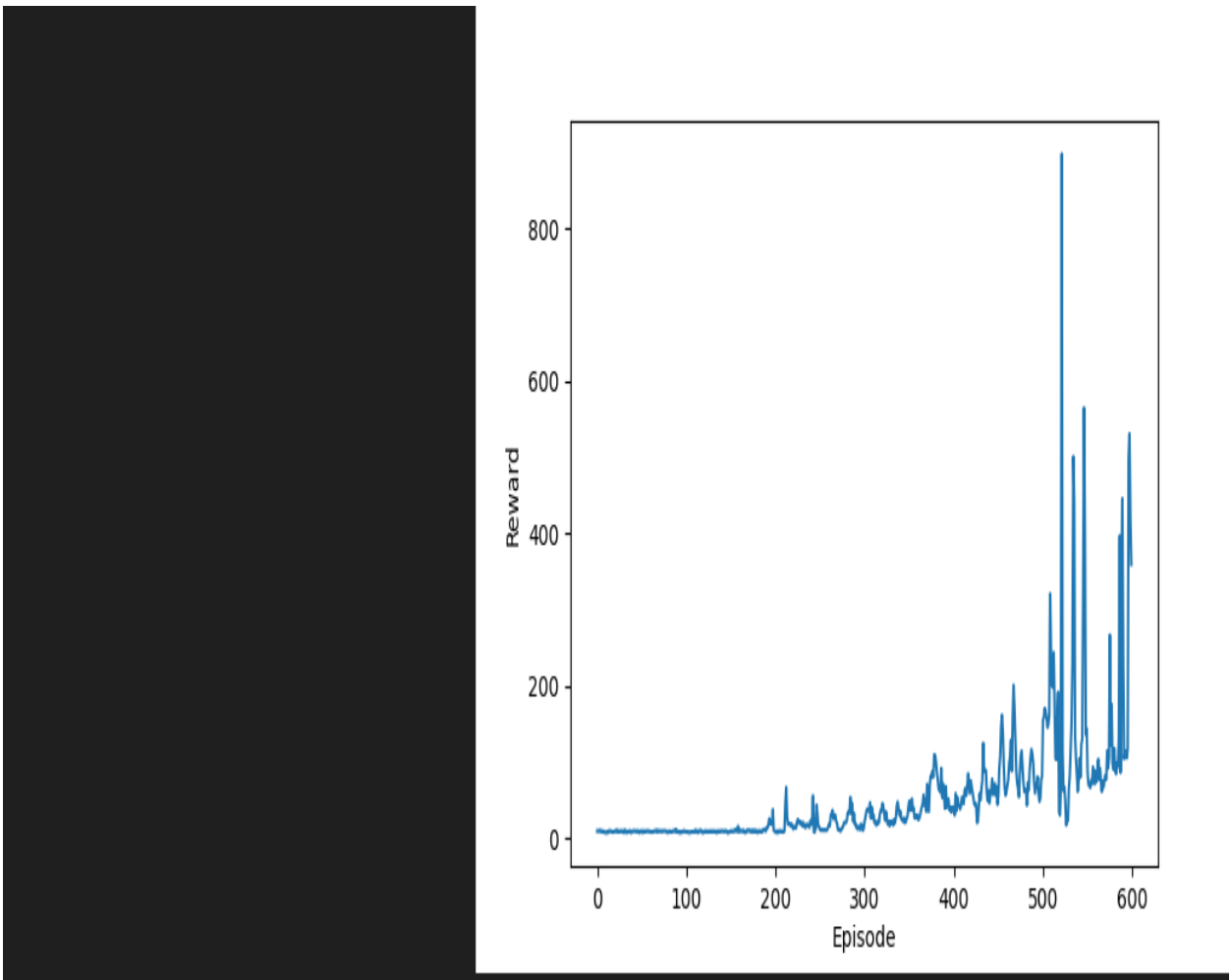
Рис. 15 — результаты обучения (epsilon_init=0.9)



PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS GITLENS

```
04-23 00:56:43: Награда достигла значения 926.0 на эпизоде 582, модель сохраняется
04-23 00:56:44: Время обучения: 20.75 секунд
PS C:\Users\HP\Documents\DQN> python agent.py --train
04-23 01:00:20: Начало тренировки
04-23 01:00:22: Награда достигла значения 9.0 на эпизоде 0, модель сохраняется
04-23 01:00:22: Награда достигла значения 11.0 на эпизоде 3, модель сохраняется
04-23 01:00:22: Награда достигла значения 13.0 на эпизоде 33, модель сохраняется
04-23 01:00:23: Награда достигла значения 38.0 на эпизоде 214, модель сохраняется
04-23 01:00:23: Награда достигла значения 116.0 на эпизоде 215, модель сохраняется
04-23 01:00:25: Награда достигла значения 154.0 на эпизоде 312, модель сохраняется
04-23 01:00:26: Награда достигла значения 166.0 на эпизоде 379, модель сохраняется
04-23 01:00:27: Награда достигла значения 255.0 на эпизоде 410, модель сохраняется
04-23 01:00:30: Награда достигла значения 491.0 на эпизоде 498, модель сохраняется
04-23 01:00:31: Награда достигла значения 592.0 на эпизоде 514, модель сохраняется
04-23 01:00:34: Награда достигла значения 763.0 на эпизоде 551, модель сохраняется
04-23 01:00:35: Награда достигла значения 839.0 на эпизоде 565, модель сохраняется
04-23 01:00:36: Награда достигла значения 1440.0 на эпизоде 574, модель сохраняется
04-23 01:00:37: Награда достигла значения 2922.0 на эпизоде 575, модель сохраняется
04-23 01:00:38: Награда достигла значения 3089.0 на эпизоде 578, модель сохраняется
04-23 01:00:40: Время обучения: 19.75 секунд
```

Рис. 16 — результаты обучения (epsilon_init=0.5)



PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS GITLENS

```
04-23 01:01:45: Награда достигла значения 10.0 на эпизоде 0, модель сохраняется
04-23 01:01:45: Награда достигла значения 11.0 на эпизоде 3, модель сохраняется
04-23 01:01:45: Награда достигла значения 12.0 на эпизоде 88, модель сохраняется
04-23 01:01:46: Награда достигла значения 16.0 на эпизоде 158, модель сохраняется
04-23 01:01:46: Награда достигла значения 18.0 на эпизоде 192, модель сохраняется
04-23 01:01:46: Награда достигла значения 26.0 на эпизоде 193, модель сохраняется
04-23 01:01:46: Награда достигла значения 38.0 на эпизоде 197, модель сохраняется
04-23 01:01:46: Награда достигла значения 45.0 на эпизоде 211, модель сохраняется
04-23 01:01:46: Награда достигла значения 67.0 на эпизоде 212, модель сохраняется
04-23 01:01:48: Награда достигла значения 71.0 на эпизоде 370, модель сохраняется
04-23 01:01:48: Награда достигла значения 82.0 на эпизоде 374, модель сохраняется
04-23 01:01:48: Награда достигла значения 88.0 на эпизоде 376, модель сохраняется
04-23 01:01:48: Награда достигла значения 110.0 на эпизоде 378, модель сохраняется
04-23 01:01:50: Награда достигла значения 125.0 на эпизоде 433, модель сохраняется
04-23 01:01:50: Награда достигла значения 147.0 на эпизоде 453, модель сохраняется
04-23 01:01:50: Награда достигла значения 162.0 на эпизоде 454, модель сохраняется
04-23 01:01:51: Награда достигла значения 201.0 на эпизоде 467, модель сохраняется
04-23 01:01:52: Награда достигла значения 321.0 на эпизоде 508, модель сохраняется
04-23 01:01:53: Награда достигла значения 898.0 на эпизоде 521, модель сохраняется
04-23 01:01:57: Время обучения: 14.59 секунд
```

Рис. 17 — результаты обучения (epsilon_init=0.3)

Согласно результатам наибольшее значение награды было зафиксировано при epsilon_init=0.5. По мере понижения начального значения epsilon происходит понижение времени обучения. При понижении значения epsilon_init произошло уменьшение количества резких скачков в значениях награды.

Выводы.

Была выполнена реализация DQN для среды CartPole-v1. Было проведено исследование влияния изменения некоторых параметров сети на её результат.

Усложненная архитектура сети показала более высокие значения награды, но более длительное время обучения.

Понижение значения gamma показало резкий спад значений награды и времени обучения.

Понижение параметра epsilon_decay дало понижение времени обучения (не такое резкое, как в случае с gamma). Наилучшее значение награды было получено при среднем значении данного параметра.

При понижении начального значения epsilon произошли изменения, схожие с изменением epsilon_decay.