

CLAIMS

1. An intelligent agent reinforcement learning method, comprising:
 - acquiring key visual information on which an intelligent agent makes a policy for a current environment image;
 - 5 acquiring actual key visual information of the current environment image;
 - determining attention variation reward information based on the key visual information and the actual key visual information; and
 - adjusting reward feedback of reinforcement learning of the intelligent agent based on the attention variation reward information.
- 10 2. The method according to claim 1, wherein the key visual information comprises: an attention area of the intelligent agent for the current environment image in making the policy.
- 15 3. The method according to claim 2, wherein acquiring key visual information on which the intelligent agent makes the policy for the current environment image comprises:
 - acquiring a value attention map of the intelligent agent for the current environment image;
 - merging the value attention map and the current environment image to obtain a
 - 20 heat map; and
 - determining the attention area of the intelligent agent for the current environment image based on the heat map.
4. The method according to claim 3, wherein acquiring the value attention map of the

intelligent agent for the current environment image comprises:

acquiring a feature map for the current environment image;

acquiring, based on the feature map, changed feature maps formed by sequentially shielding channels of the feature map;

5 acquiring state value variation amounts of the changed feature maps relative to the feature map respectively; and

generating the value attention map based on the state value variation amounts and the changed feature maps.

10 5. The method according to claim 4, wherein acquiring the feature map for the current environment image comprises:

inputting the current environment image into a convolutional neural network; and

acquiring the feature map output by a last convolutional layer of the convolutional neural network.

15

6. The method according to any one of claims 4 to 5, wherein acquiring state value variation amounts of the changed feature maps relative to the feature map respectively comprises:

for each of the changed feature maps,

20 inputting the changed feature map to a value network of the intelligent agent to acquire a state value of the changed feature map; and

calculating a difference between a state value output by the value network for the feature map and the state value of the changed feature map, to acquire a state value variation amount of the changed feature map relative to the feature map.

7. The method according to any one of claims 1 to 6, wherein the actual key visual information of the current environment image comprises: an area of a target object in the current environment image.

5

8. The method according to claim 7, wherein determining attention variation reward information based on the key visual information and the actual key visual information comprises:

10 determining the attention variation reward information based on a ratio of an intersection between the attention area of the intelligent agent for the current environment image in making the policy and the area of the target object to the area of the target object.

15 9. The method according to any one of claims 1 to 8, wherein the reward feedback of reinforcement learning of the intelligent agent comprises: the attention variation reward information and reward information generated when the intelligent agent makes the policy for the current environment image.

20 10. The method according to any one of claims 1 to 9, further comprising:
determining an exploration degree in an environment exploration cycle based on the key visual information; and

when determining that the exploration degree does not conform to a predetermined exploration degree, performing experience replay with historical training data stored;

wherein the historical training data comprises: training data obtained by filtering

sampled reinforcement learning training data based on a preset requirement.

11. The method according to claim 10, wherein determining the exploration degree in the environment exploration cycle based on the key visual information comprises:

5 determining an attention variation amount in the environment exploration cycle based on variation information between value attention maps of the intelligent agent for the current environment image at multiple adjacent moments in the environment exploration cycle,

 wherein the attention variation amount is used to measure the exploration degree
10 in the environment exploration cycle.

12. The method according to claim 11, further comprising:

 determining an importance level of reinforcement learning training data sampled in the environment exploration cycle based on positive reward and the exploration
15 degree in the environment exploration cycle; and

 storing reinforcement learning training data sampled in the environment exploration cycle which has an importance level conforming to the predetermined requirement, as the historical training data.

20 13. An intelligent agent reinforcement learning apparatus, comprising:

 a key vision acquisition module configured to acquire key visual information on which an intelligent agent makes a policy for a current environment image;

 an actual vision acquisition module configured to acquire actual key visual information of the current environment image;

a variation reward determining module configured to determine attention variation reward information based on the key visual information and the actual key visual information; and

5 a reward feedback adjusting module configured to adjust reward feedback of reinforcement learning of the intelligent agent based on the attention variation reward information.

14. The apparatus according to claim 13, wherein the key visual information comprises: an attention area of the intelligent agent for the current environment image in making
10 the policy;

wherein the key vision acquisition module is configured to:

acquire a value attention map of the intelligent agent for the current environment image;

15 merge the value attention map and the current environment image to obtain a heat map; and

determine the attention area of the intelligent agent for the current environment image based on the heat map;

wherein when acquiring a value attention map of the intelligent agent for the current environment image, the key vision acquisition module is configured to:

20 acquire a feature map for the current environment image;

acquire, based on the feature map, changed feature maps formed by sequentially shielding channels of the feature map;

acquire state value variation amounts of the changed feature maps relative to the feature map respectively; and

generate the value attention map based on the state value variation amounts and the changed feature maps;

wherein when acquiring a feature map for the current environment image, the key vision acquisition module is configured to:

- 5 input the current environment image into a convolutional neural network; and
acquire a feature map output by a last convolutional layer of the convolutional neural network;

wherein the feature map output by the last convolutional layer is the feature map for the current environment image acquired by the key vision acquisition module.

10

15. The apparatus according to claim 13 or 14, wherein the actual key visual information of the current environment image comprises: an area of a target object in the current environment image;

wherein the variation reward determining module is configured to:

- 15 determine the attention variation reward information based on a ratio of an intersection between the attention area of the intelligent agent for the current environment image in making the policy and the area of the target object to the area of the target object.

- 20 16. The apparatus according to any one of claims 13 to 15, further comprising: an experience replay module configured to:

determine an attention variation amount in an environment exploration cycle based on variation information between value attention maps of the intelligent agent for the current environment image at multiple adjacent moments in the environment

exploration cycle, wherein the attention variation amount is used to measure an exploration degree in the environment exploration cycle; and

when determining that the exploration degree does not conform to a predetermined exploration degree, perform experience replay with historical training data stored,

5 wherein the historical training data comprises: training data obtained by filtering sampled reinforcement learning training data based on a preset requirement.

17. The apparatus according to claim 16, further comprising a training data acquisition module configured to:

10 determine an importance level of reinforcement learning training data sampled in the environment exploration cycle based on positive reward and the exploration degree in the environment exploration cycle; and

store reinforcement learning training data sampled in the environment exploration cycle which has an importance level conforming to the predetermined requirement, as

15 the historical training data.

18. An electronic device, comprising:

a memory configured to store a computer program;

a processor configured to execute the computer program stored in the memory,

20 wherein when the computer program is executed, the method of any one of claims 1-12 is implemented.

19. A computer-readable storage medium storing a computer program, wherein when the computer program is executed by a processor, the method of any one of claims 1-

12 is implemented.

20. A computer program comprising computer instructions, wherein when the computer instructions are executed by a processor of a device, the method of any one of claims

5 1-12 is implemented.