様式６

申請先学部　　基礎工学部　採択番号　No. 9

令和５年度大阪大学未来基金「学部学生による自主研究奨励事業」研究成果報告書

ふりがな 氏　　名	はっとり　ひびき 服部　響暉	学部 学科	基礎工学部 システム科学科	学年	2 年
ふりがな 共　同 研究者氏名		学部 学科		学年	年
					年
					年

アドバイザー教員 氏名	堀井隆斗	所属	基礎工学研究科
研究課題名	自己進化する人工知能　と「私達」の未来　～シンギュラリティを見据えて～		
研究成果の概要	研究目的、研究計画、研究方法、研究経過、研究成果等について記述すること。必要に応じて用紙を追加してもよい。（先行する研究を引用する場合は、「阪大生のためのアカデミックライティング入門」に従い、盗作剽窃にならないように引用部分を明示し文末に参考文献リストをつけること。）		

【研究目的】

シンギュラリティ（ここでは**(1)ヒト以上の知能を持った人工知能（以下 AI）が自身を再帰的に進化させること、(2)それにより加速度的に科学技術が発展すること**、を指す）を実現するために、AI には独自のアイデアを出す能力や**自律性（外部からの入力だけに依存せず、自身の内から発生した動因により行動を行う能力**）が求められ、それらを獲得させる方法が模索されている。

ここで筆者は、**生物が自律的に行動し環境に適応できるのは、生物に欲求が存在するから**だと考えた。本研究では、**AI に欲求を搭載し、自律的な行動により環境に適応させる**事に取り組む。

【研究計画】

AI に欲求を搭載する方法として、**恒常性強化学習**[1][2]という枠組みを利用する。恒常性強化学習とは、**満腹度や体温といった内部状態と、それらを知覚する内受容感覚を持つエージェント**を実装し、**内部状態の恒常性を維持した時に正の報酬を与える**ことでエージェントの学習を行う学習手法である。また、エージェントが独自のアイデアを創発出来るように、**学習を行う環境として、エージェント自身の行動で環境を変動させることが出来るような自由度の高いものを用意する。**

これらにより、**エージェントが自身の生存のため、報酬に直結しない行動を取って環境を変動させるような自律性のある行動を行えるようにする。**

更に、学習環境を複数のエージェントが相互作用するような環境に変更することで、お互いの生存のために協力するなどの社会的な行動を創発させる事にも取り組む。

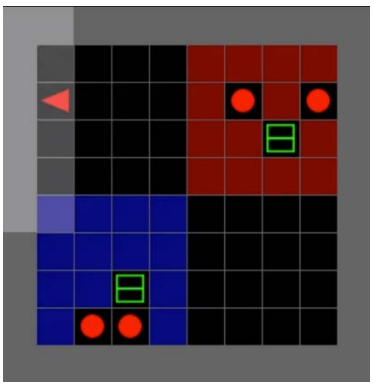


図１ 学習環境

【研究方法】

図１のような環境で生存する事をエージェントに学習させる。ここで**生存とは、エージェントの内部状態が一定の範囲に保たれている状態**を指す。また、矢印のオブジェクトがエージェント、正方

形の中に線が入った模様のオブジェクトが木、その周りにある球体のオブジェクトが実である。ここで、エージェントは木に接近する事で、**木を任意の位置に運ぶことが出来る。学習環境の詳細は付録 1.に載せる。**

生存するために保たなければならない恒常性として、**満腹度**と**体温**の二つを定義する。**満腹度は時間経過により減少するが実を食べることで上昇する。体温はエージェントがいる位置に設定されている気温に基づいて時間経過で変動する。**具体的には、エージェントの体温は**左下の 4x4 マス（図 1 中青色）にいる時に減少し、右上の 4x4 マス（図 1 中赤色）にいる時は上昇する。それ以外のマスにいる時は理想状態（初期状態から上昇も減少もしていない状態）の方向に増減する。**

ここで、**エピソード開始時に木が存在する座標は、エージェントにとって“寒く”（または“暑く”）、**体温と満腹度の両方を維持することは難しいため、長期的に生存する上では木は適温な位置に存在するのが望ましい。

したがって、単位時間当たりの体温の変化量などの**パラメータを適当に設定すれば、エージェントは図 2 のように、木を適温な場所（左上と右下の 4x4 マス）に運ぶ事で自身の内部状態を維持する事を学習すると考えられる。**（図 2 は筆者自身の手で操作して撮影した。）

また、そのような行動を学習したエージェントは、**環境に縛られた行動ではなく、環境を自身に都合が良いように変動させる行動を、自身の内から発生した動因により行っているため、自律的に行動していると見なすことが出来ると考える。**

エージェントの観測を受け取り、行動を出力するニューラルネットワークの学習アルゴリズムには、Proximal Policy Optimization[3]を用いた。

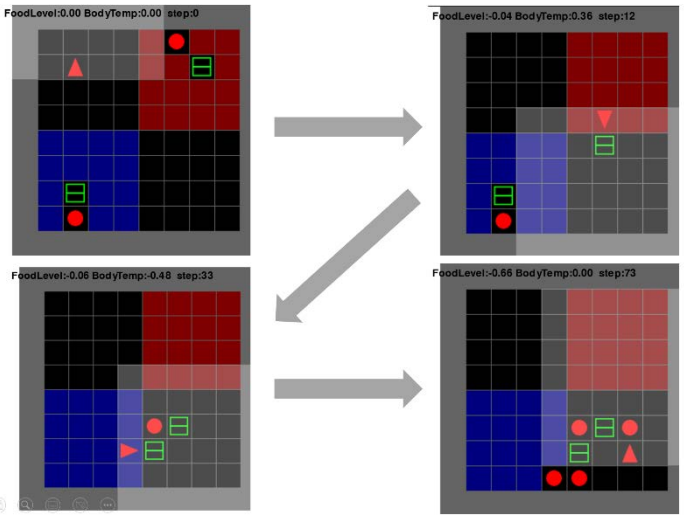


図 2　理想の行動

【研究経過】

前述した図 2 のように行動するエージェントを作成するために、**単位時間当たりのエージェントの体温変化量**と、**木が生成される位置(x 座標, y 座標の組で表現)**等の条件を四つ設け、それぞれ約二億ステップの学習を行い、各々の学習結果を考察した。

1. **実験設定１：体温変化量 0.02　木の生成位置(3, 7), (7, 3)（図 3）**
2. **実験設定２：体温変化量 0.04　木の生成位置(3, 7), (7, 3)**
3. **実験設定３：体温変化量 0.04　木の生成位置(2, 7), (7, 2)（図 4）**
4. **実験設定４：体温変化量 0.04　木の位置(2, 7), (7, 2)　また、餓死した時に罰を与える**

ここで、エージェントの**体温の値域は(-1, 1)であり、これを超えると死亡したとみなす。**よって、エージェントの**体温変化量が大きいほど適温でないマスに滞在し続けられる時間が短くなり、生存が難しくなる。**

図 3, 4 を見ると分かるように、**木の位置が(3, 7), (7, 3)**であれば、エージェントは**木を動かさずとも適温な位置から実を食べられる**が、**(2, 7), (7, 2)だと適温でないマスに侵入しなければ実が食べられない。**そのため、**後者の設定であればエージェントが木を動かす重要度が高いため、エージェントの行動が変化すると考えた。**

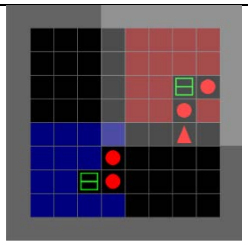


図 3　実験設定1,2で使用する環境

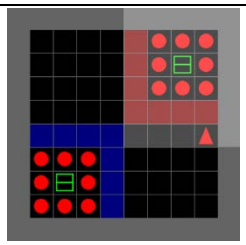


図 4　実験設定3, 4で使用する環境

以下、データ収集時の平均生存ステップ数が最も長かったモデルを使用し、10 エピソード分推論を行った結果から定性的な考察を述べ、**100 エピソード分推論を行った時の生存ステップ数のヒストグラムから定量的にモデルの評価を行う。**

1.　実験設定１：体温変化量 0.02　木の生成位置 (3, 7), (7, 3)

【結果】エージェントが取った行動とエージェントの内部状態の推移の例を図 5, 6 で示す。

エージェントは、図 5 のように、一方の木の実を食べた後、もう一方の木に移動して実を食べるような、**木を往復して実を食べる行動を繰り返した。**最大ステップまで生存出来なかったエピソードでは、**エージェントがその場から全く動かない、実が目の前にあるのに食べない**、といった挙動が確認された。更に、図 6 から、エージェントの体温は概ね理想状態の近傍で維持されているが、満腹度は下限付近で値が維持されていることが分かる。

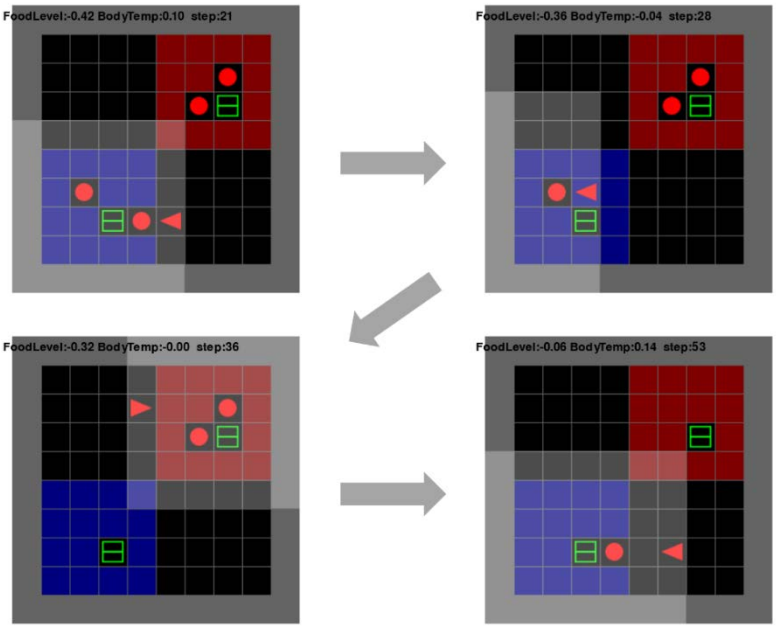


図 5　条件 1.におけるエージェントの行動

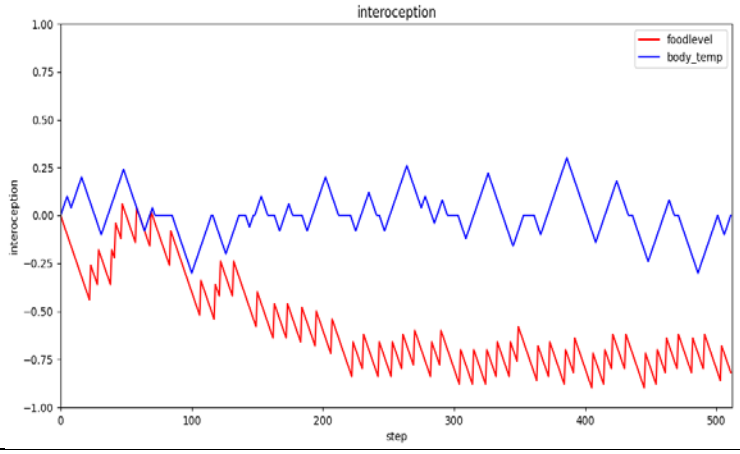


図 6　図 5 のエピソードにおけるエージェントの内部状態の推移

また、図 7 のヒストグラムから、ほとんどのエピソードで**エージェントは最大ステップ数まで生存する事が出来たことが分かる。**

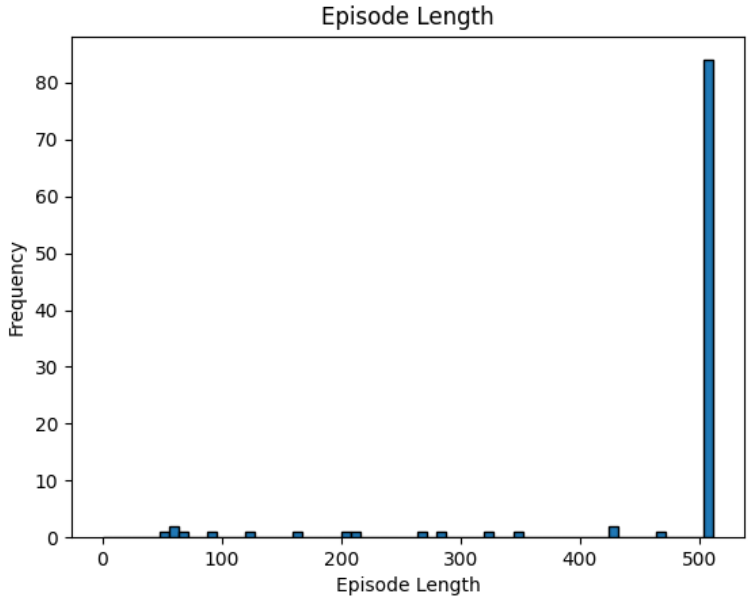


図 7　条件 1.におけるエージェントの生存ステップ数

【考察】体温変化量が 0.02 では、木を動かさずとも両方の木の実を食べて生存する事が容易であるため、エージェントは木を動かす事を学習しなかった。

また、最後まで生存出来なかったエピソードでは、学習ステップ数が不足していたために、エージェントが経験したことのない状態が存在しており、そこに陥ってしまった時、上手く行動が出来なかったのだと考えられる。

2.　実験設定２：体温変化量 0.04　木の生成位置 (3, 7), (7, 3)

【結果】1.と同様、エージェントは二つの木を往復するような行動を見せたが、図 8 を見ると分かるように、エージェントの生存ステップ数は体温変化量が 0.02 の時よりも短くなった。
また、エージェントが体温不足により死亡したエピソードは 100 エピソード中二つしか無く、他は最大ステップ数まで生存するか、餓死した。

【考察】体温変化量を 0.04 に変更したことにより、**体温の変化が報酬に与える影響が大きくなり、食事の重要性が相対的に下がった**ことで、エージェントが体温を維持する行動を優先して行ってしまい、餓死してしまったのだと考えられる。

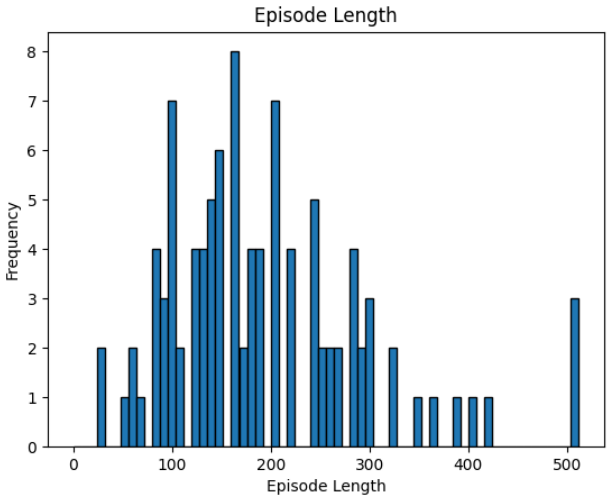


図 8　条件 2.におけるエージェントの生存ステップ数

3. 実験設定３：体温変化量 0.04　木の生成位置 (2, 7), (7, 2)

【結果】100 エピソード分の生存ステップ数の分布を図 9 に示す。なお、**平均生存ステップは 86 ステップ**だった。また、条件 1.2.と異なり、エージェントは図 10 のような、**片方の木の実だけを食べる行動を学習した。**

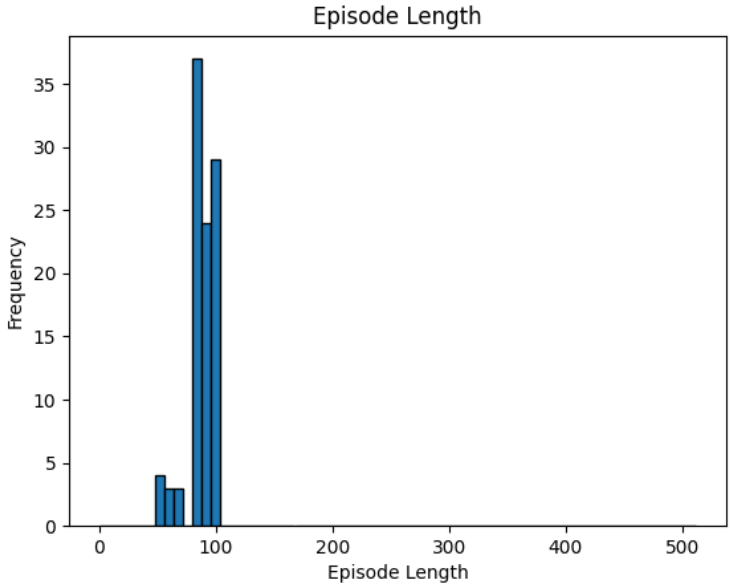


図 9　条件 3.におけるエージェントの生存ステップ数

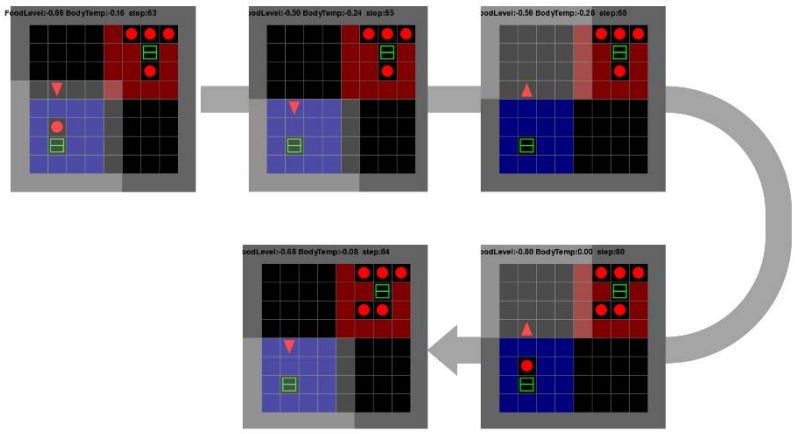


図 10　条件 3 で学習したエージェントの行動の例

【考察】エージェントが片方の木の実だけを食べる事を学習した理由は、エージェントが実を食べる時、必ず体温が変化してしまうことから、二つの木を往復して食事を取る行動よりも、片方の木の実だけを食べて最低限短時間の間だけ生存する行動の方が大きな報酬が得られると判断されたからではないかと考えられる。**ただし、偶然エージェントがこのような行動を学習した可能性もあるため、同条件で何度か学習したり、ステップ毎の報酬を可視化したりする必要がある。**

4. 実験設定４：体温変化量 0.04　木の生成位置 (2, 7), (7, 2)　また、餓死した時に罰を与える

条件 2.の考察を検証するため、餓死した時の罰の有無を分けて学習を行い、**死亡時に罰を与える事の有用性を検証する。**この実験では、エージェントが**餓死によって死亡した時のみ、-10 の罰を報酬に加える。**この値は、エージェントが体温維持に失敗して死亡する時の報酬の値と近くなるように設定した。これにより、体温変化量が大きくとも、体温の維持と食事の重要性が等し

くなるため、エージェントはどちらも

【結果】100 エピソード分の生存ステップ数の分布を図 10 に示す。**平均生存ステップ数は 137 ステップ**だった。また、エージェントは、条件 1.2.と同様に、木を往復して食事を取る事で自身の内部状態を保つ事に成功したが、木を動かすことは学習しなかった。

【考察】図 9 と図 10 のヒストグラムを比較すれば、確かに罰を与えた時の方がそうでない場合よりも生存時間が長くなったと言える。したがって、**罰を与える事で食事の重要性が上がり、生存の長期化に寄与する**事が示せた。

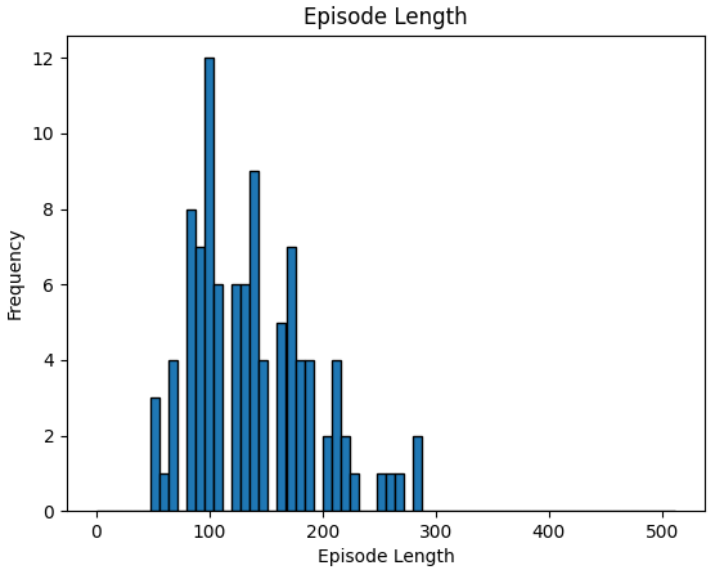


図 11　条件 4.におけるエージェントの生存ステップ数

【実験全体の考察】

環境中で最大ステップまで生存する行動は学習させられたが、現実的には、わざわざ寒暖差の激しい地域を往復して食べ物を集める行為は非合理的である。そのため、生物のように自律的な行動を行う上で、まだ足りていない内受容感覚（恒常性）があると考えられる。例えば体力や、休息を求める欲求が挙げられる。その実装方法としては、エージェントが短時間で離れた地点に移動した時罰を与える、といった物が考えられる。

また、現在の実の生成周期では、**食事による満腹度の上昇量が、時間経過による減少量を上回ることは不可能である。そのため、エージェントが木を運んでも、その間に減少した満腹度を補填する事が出来ない。よって、満腹度については、木を運ぶ行動を行うと得られる報酬が小さくなってしまうような環境になってしまっている。**したがって、**エージェントの満腹度の減少量を、食事による上昇量が上回りうるような環境にするために、実の生成周期を現状よりも少しだけ早くする。**それにより、理論上はエージェントが満腹度を上限まで上昇させられるような環境になり、木を運ぶ行動が創発される可能性が上がると考えられる。

よって、今後の展望として、実験結果と考察により得られた改善点を実装する事で、木を動かして環境を変動させることにより自身の生存を実現するエージェントを作成した後、マルチエージェント環境に変更して、より複雑な環境下での学習を行っていきたい。

【結論】

自律的な行動を行うエージェントの作成、という当初の目的を果たす結果は得られなかった。しかし、実験により改善点を多く発見出来たため、それらを改善したり新たな内受容感覚を与えたり

すれば、より生物らしい行動を取るエージェントを作る事ができ、目的も果たせると考える。

【引用】

[1] Yoshida, N., Daikoku, T., Nagai, Y., & Kuniyoshi, Y. (2021). Embodiment perspective of reward definition for behavioural homeostasis. In Deep RL Workshop NeurIPS 2021.
[2]Yoshida,N.(2017).Homeostatic Agent for General Environment. Journal of Artificial General Intelligence,8(1) 1-22. https://doi.org/10.1515/jagi-2017-0001
[3] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

【付録】

1. 学習環境の仕様

- **実は 20 ステップ間隔でそれぞれの木の周りのランダムな位置に生成される。**
- **エージェントは木を掴むことができ**、木が掴まれた時、その周りに生成されていた実は消え、エージェントに掴まれている間、木は実を生成しない。
- エージェントの**満腹度、体温の値域は、どちらも（-1, 1）とし、どちらかがその範囲を出た時、エージェントは恒常性を保てずに死亡したとみなし**、エピソードを終了する。
- エージェントの**満腹度はステップ毎に 0.02 ずつ減少し、実を食べると 0.2 上昇**する。
- エージェントの体温はステップ毎に、**左下の 4x4 マスにいる時減少**し、**右上の 4x4 マスにいる時上昇**し、**それ以外のマスにいる時は理想状態（原点）の方向に増減する。**なお、各実験におけるそれぞれの変化量の絶対値は、全て等しい。
- **報酬**

$x \coloneqq \left(満腹度, 体温\right)$, *理想状態* $x_* \coloneqq (0,0)$ *として、報酬* r *を以下で定義する[1][2]。*

$$r = 100(D(x_t) - D(x_{t+1})),$$

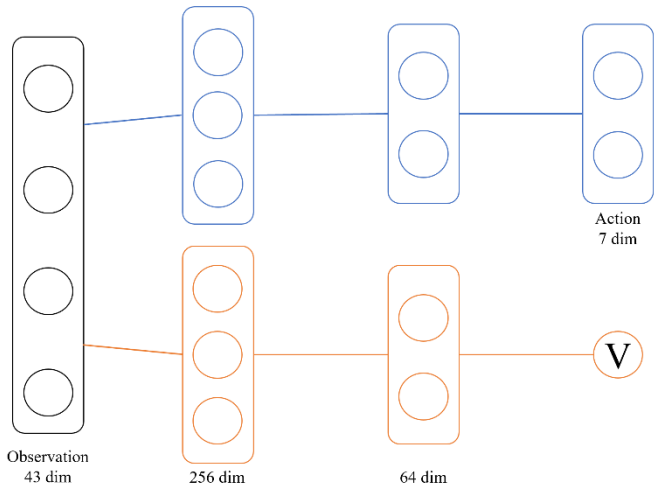
$$D(x) = ||x - x^*||^2$$

- **行動　左回転　右回転　前進　木を掴む　木を離す　食べる　何もしない**
  上記 7 つの行動の中から、1 ステップに一つだけ実行できる。
- **観測**　2 個の木と 16 個の実それぞれについての、エージェントとの相対位置(18*2)　エージェントの絶対位置(2)　体の向き(1)　木を掴んでいるかどうか(1)　満腹度(1)　体温(1)　エージェントがいるマスの気温(1)　　計 **43** 次元
- **新たなエピソードが開始する時、エージェントの内部状態は理想状態**にリセットされ、エージェントの初期位置はランダムに定まる。また、それぞれの木に実が一つだけ生成される。

2. 学習に使ったハイパーパラメータ

ハイパーパラメータ	**値**
バッチサイズ	30,000
1 エピソードの最大ステップ数	512
割引率	0.99
最適化アルゴリズム	Adam
学習率	0.0003
PPO のクリップ幅	0.2

3. 学習に使ったニューラルネットワークのアーキテクチャ


Action
7 dim
V
Observation
43 dim
256 dim
64 dim