The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

内的報酬と敵対的学習によるタスク非依存な注意機構の学習

Task-free Attention Learning with Intrinsic Reward and Adversarial Learning

松嶋 達也 Tatsuya Matsushima　大澤 昇平 Shohei Osawa　松尾 豊 Yutaka Matsuo

東京大学
The University of Tokyo

Recent advances in artificial intelligence, especially deep learning, have enabled us to handle wider range of problems with computers. As for the real-world problem settings, however, there remain some difficulties, for example, inputs for embodied agents are partially observed representation of their states, and building models of their environments is needed for more sample efficient systems. One possible solution for coping with these difficulties is to use attention mechanism, which models the visual system of human and regards its inputs as a sequences, learning to where to attend.

In this paper, we propose a method to train attention mechanism of neural network without external rewards. The proposed method consists of two ideas, one is to use intrinsic reward for attention mechanism and the other is to adopt adversarial learning in the model.

1. 研究の背景と目的

近年，人工知能の活用による社会の生産性向上への期待が高まっている．この期待の背景として，深層学習が人工知能技術のブレイクスルーになったことが挙げられ [Goodfellow 16]，強化学習と組み合わせた深層強化学習を利用することにより，ビデオゲームや囲碁などのタスクにおいて従来の手法を上回る結果を記録している [Mnih 15, Silver 17]

しかし，実世界環境で活用可能な人工知能の構築には，課題が残されている．特に，観測が部分的であること，タスクに対する報酬の設計が難しいことの 2 点が挙げられる [Thrun 05]．

部分的な観測を扱うニューラルネットワークのモデルとして，人間の視覚的注意を模倣した注意機構 (attention mechanism) を持つモデルが提案されている．しかし，これらのモデルでは，注意機構の学習がタスクから定義される外的な報酬信号を用いた強化学習によって行われており，外部からの報酬信号が得られない問題設定下では注意機構の学習を行うことができない．

本研究の目的は，特定のタスクに依存しない方法で注意機構を学習させ，状態の予測を行う手法を構築することである．

提案手法のアイデアは，(i) 内的な報酬の利用と (ii) 敵対的学習の二点である．まず，注意機構を用いたニューラルネットワークのモデルである Recurrent Attention Model (RAM) [Mnih 14] において，タスクから定義され注意機構に与えてられていた外的な報酬の代わりに，1 時間ステップ後における観測の予測誤差を注意機構の内的報酬として与える．次に，観測の予測モデルと注意機構を敵対的な学習により訓練する．

2. 関連研究

2.1 注意機構

RAM は，各時間ステップにおいて画像として得られる観測データをリカレントニューラルネットワークにより時系列方向に統合して，画像のラベルの予測と，次の時間ステップにおいて観測する場所を決定するモデルである．

連絡先: 松嶋達也，東京大学，〒 113-8656 東京都文京区本郷 7-3-1，matsushima@weblab.t.u-tokyo.ac.jp

RAM では画像の識別タスクにおける正誤の情報を報酬として，強化学習を用いて注意機構の学習を行っている． そのため，RAM は状態 (画像全体) のモデルの学習をしているのではない．

2.2 内的な報酬の利用

エージェントにとって環境が未知であり，環境に関する知識を学習する必要のある場合には，エージェントの探索を促進する必要がある．エージェントに環境の探索を促す方法として，内的な動機づけ (intrinsic motivation) や好奇心 (curiosity) を導入する手法が考案されている．

特に，Pathak らの研究 [Pathak 17] では，報酬がスパースな設定の強化学習の問題に対して，好奇心を内的な報酬として導入し探索行動を促している．この研究では，エージェントは入力として受け取る次の観測を予測し，その予測誤差を内的な報酬 (好奇心) として定義している．エージェントの行動は，タスクにおける報酬と好奇心による報酬の重み付き和を報酬とする強化学習により学習されている．実験では，報酬が得られるまでに長い行動の系列が必要なビデオゲーム (Super Mario Bros.) に対し，好奇心を用いることでエージェントが以前に経験していない状態を探索する行動を促し，タスクの達成にかかる時間を短縮できることが示された．

RAM では，画像の分類タスクという特定のタスクに対して，そのタスクから定義される報酬を用いて注意機構の学習を行っている．一方，本研究では，特定のタスクに依らず状態 (RAM でいう画像全体に対応する) のモデルを学習することを目的とするため，注意機構の学習に内的な報酬を利用する．

3. 提案手法

状態のモデルを学習することによってタスク間で共有可能な状態に関する表現を学習できる．

これは，RAM のような単一のタスクを行う問題設定ではなく，訓練時に画像のほかにラベルのような正解データや存在しない場合や，途中でタスクが変化する場合に，一度学習した内容を他のタスクに転移させるときに有効である．

本研究では，RAM を拡張し，注意機構の学習に外的な報酬を用いる代わりに観測に対する予測の誤差を注意機構の内的な

報酬として与え，観測の予測モデルと注意機構を敵対的な学習により訓練する手法を提案する

3.1 問題設定

モデル全体として，状態 (本研究の場合は画像全体)$\boldsymbol{x}$ に関する予測 $\hat{\boldsymbol{x}}$ を行うことを目標とする．つまり，真に最小化を行いたい予測誤差 L は以下のように与えられる．

$$L = \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|^2 . \tag{1}$$

しかし，エージェントは画像全体の情報 $\boldsymbol{x}$ を受け取ることができないため，提案手法では，エージェントが利用可能な情報である各時間ステップ t における観測 $\boldsymbol{o}_t$ を用いて間接的に状態に関する予測を行う．

一様乱数から観測する位置をサンプルし，観測に対する予測を行なった場合，観測に対する予測誤差の期待値は，状態 (画像全体) の予測誤差に比例する．つまり，この最小化問題は観測の予測誤差 $J_p(\theta_p)$ を最小化する問題として考えることができる．

$$J_p(\theta_p) = \frac{1}{T}\sum_{t=1}^{T} \|\hat{\boldsymbol{o}}_t(l_{t-1};\theta_p) - \boldsymbol{o}_t\|^2 . \tag{2}$$

ただし，l_{t-1} は観測 $\boldsymbol{o}_t$ の位置，$\hat{\boldsymbol{o}}_t$ をエージェントによる観測 $\boldsymbol{o}_t$ に対する予測，θ_p は観測を予測するエージェントのモデルパラメータ，T を観測の時間ステップの数とする．

3.2 内的な報酬

RAM における注意機構の学習では，ラベルの識別タスクにおける正誤を報酬として利用する強化学習を行っていた．しかし，ラベルなどの正解データが存在しない場合や，状態に関するモデル自体を学習する場合には，モデルの外部から報酬が与えられず RAM を直接利用することはできない．

そのため，以上に挙げたような問題設定下でも，注意機構の学習に利用することのできる指標として観測の予測に関する誤差 $\|\hat{\boldsymbol{o}}_t - \boldsymbol{o}_t\|^2$ を用い，注意機構を訓練するための報酬として定義することを考える．

しかし，RAM を観測の予測誤差の最小化するというアイデアのみで拡張し，観測の予測誤差が小さくなる位置の決定に対して報酬を多く与えるように設定すると，注意機構によって決定される位置が固定してしまうことが容易に考えられる．

3.3 敵対的な学習

そのため，本論文の提案手法では，注意機構に対して観測の予測誤差が大きくなる行動に対して大きな報酬を与えることで，エージェントの探索を促し，式 1 で表現された真の目的関数を間接的に最適化することを考える．

注意機構に対し，観測の予測誤差が大きくなる行動に対して大きな報酬を与えるためには，報酬として観測の予測誤差の値をそのまま利用すれば良い．時間ステップ t における注意機構に対する報酬を R_t とすると，

$$R_t = \|\hat{\boldsymbol{o}}_t - \boldsymbol{o}_t\|^2 . \tag{3}$$

とできる．

注意機構の入力は，観測を予測するエージェントが持つ状態 (画像全体) に関する予測であり，更新するパラメータは，注意機構が持つパラメータ θ_l のみである．つまり，状態の予測を通じて観測を予測するエージェントには強化学習によるパラメータ更新の勾配を伝播させない．

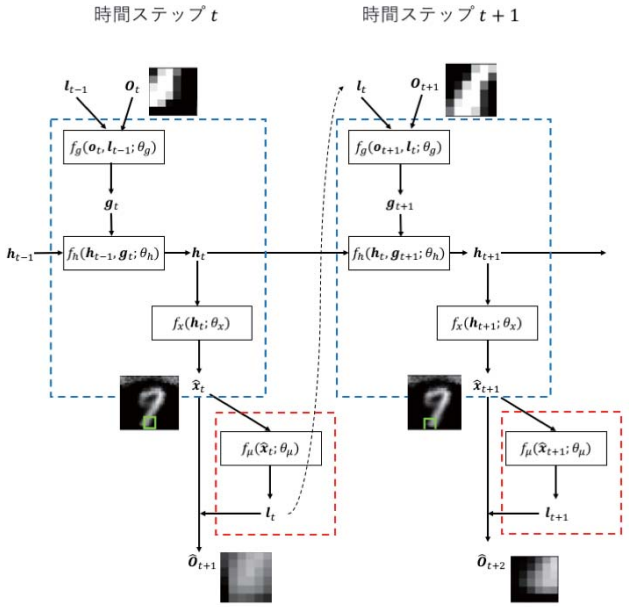

図 1: 提案モデルの概略図

すなわち，注意機構の最適化問題は，式 4 で表される目的関数 $J_l(\theta_l)$ の最小化問題となる．

$$J_l(\theta_l) = -\mathbb{E}_{\pi(\theta_l)}\left[\sum_{t=1}^{T} R_t\right] . \tag{4}$$

ただし，π を注意機構の方策とする．

以上より，提案手法は観測する位置が与えられた上でその観測の予測誤差を最小化するように教師あり学習を行うエージェントと，行動を観測の位置とし，その位置における観測の予測誤差が大きければ大きいほど大きい報酬が与えられる強化学習を行うエージェント (注意機構) の 2 つのエージェントによる学習として考えることができる．

つまり，これらのエージェントは，

$$\min_{\theta_p}\max_{\theta_l} \mathbb{E}_{\pi(\theta_l)}\left[\sum_{t=1}^{T}\left\|\hat{O}_t(\theta_p) - O_t\right\|^2\right] . \tag{5}$$

というミニマックスゲームに基づく，敵対的な学習を行っている．

図 1 は提案モデルの概略図であり，時間ステップ t と $t+1$ における各変数の関係を示している．左上の枠線で囲まれた部分は観測を予測するエージェント，右下の枠線で囲まれた部分は注意機構に相当する．

Algorithm1 は，提案手法におけるモデルの学習の流れを擬似コードを用いて表したものである．1 つの画像に対して，注意機構による観測の位置の決定と，その位置における観測の予測の組み合わせを T 回行ったあと，観測の予測エージェントと注意機構それぞれに対して損失 (loss) を計算し，勾配効果法を用いてそれぞれのモデルのパラメータを更新する．以降，この一連の流れを 1 エポック (epoch) と呼ぶことにする．

Algorithm 1 提案手法における学習

Input: 画像:$\boldsymbol{x}$

$\boldsymbol{l}_0$ を一様乱数を用いて初期化する
for $t = 1, 2, \ldots, T$ **do**
  $\boldsymbol{o}_t \leftarrow Crop(\boldsymbol{x}, \boldsymbol{l}_{t-1})$
  $\boldsymbol{g}_t \leftarrow f_g(\boldsymbol{o}_t, \boldsymbol{l}_{t-1}; \theta_g)$
  $\boldsymbol{h}_t \leftarrow f_h(\boldsymbol{h}_{t-1}, \boldsymbol{g}_t; \theta_h)$
  $\hat{\boldsymbol{x}}_t \leftarrow f_x(\boldsymbol{h}_t; \theta_x)$ ▷ 画像全体に関する予測を行う
  $\boldsymbol{\mu}_t \leftarrow f_\mu(\hat{\boldsymbol{x}}_t; \theta_\mu)$
  $\boldsymbol{l}_t \sim N(\cdot; \boldsymbol{\mu}_t, diag(\sigma^2))$ ▷ REINFORCE
  $\hat{\boldsymbol{o}}_{t+1} \leftarrow Crop(\hat{\boldsymbol{x}}_t, \boldsymbol{l}_t)$
$J_p(\theta_p) \leftarrow \frac{1}{T}\sum_{t=1}^{T} \|\hat{\boldsymbol{o}}_t(\theta_p) - \boldsymbol{o}_t\|^2$ ▷ 損失を計算
$J_l(\theta_l) \leftarrow -\mathbb{E}_{\pi(\theta_l)}\left[\sum_{t=1}^{T} \|\hat{\boldsymbol{o}}_t - \boldsymbol{o}_t\|^2\right]$
$\theta_p \leftarrow \theta_p - \eta_p \nabla_{\theta_p} J_p(\theta_p)$ ▷ パラメータの更新
$\theta_l \leftarrow \theta_l - \eta_l \nabla_{\theta_l} J_l(\theta_l)$

ただし，パラメータ θ_p を観測を予測するエージェントのパラメータ全ての集合 $\theta_p := [\theta_g, \theta_h, \theta_x]$，パラメータ θ_l を注意機構のパラメータ全ての集合 $\theta_l := [\theta_\mu]$，$N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ を平均が $\boldsymbol{\mu}$ で共分散が $\boldsymbol{\Sigma}$ の多変量正規分布，$Crop(\boldsymbol{x}, \boldsymbol{l})$ を画像 $\boldsymbol{x}$ から位置 $\boldsymbol{l}$ を中心とした部分を切り取る操作，η_p を観測を予測するエージェントの学習率，η_l を注意機構の学習率，b を適当なベースライン (定数) とする．

4. 評価実験

提案手法の有効性を検証するために，画像に関するタスクのベンチマークとして利用される MNIST データセット [LeCun 98] を用いて評価実験を行った．また，注意機構の強化学習アルゴリズムとして REINFORCE アルゴリズム [Willia 92] を利用した．

4.1 実験の設定

注意機構のモデルに関する比較実験を行った．実験のタスクは，画像の一部を観測しモデルの内部状態として画像全体の予測を行うことである．以下の 3 つの手法を比較する．

提案手法
注意機構は予測モデルの内部状態を入力として，次に観測を行う位置を決定する．注意機構には，観測の予測誤差が大きければ大きいほど，その位置の選択に対して大きな報酬を与える．このとき，注意機構は観測の予測モデルに対して敵対的な設定となる．

ベースライン (ランダム)
次に観測を行う位置は一様分布からサンプリングされた値を用いる．

ベースライン (協力的)
注意機構には観測の予測誤差が小さくなればなるほど大きい報酬を与える．これは RAM を内部報酬というアイデアのみで拡張したものであり，このとき，注意機構は観測の予測モデルに対して協力的な設定となる．

評価時には，以上のように訓練されたネットワークを用いて，訓練時とは異なる画像に対し，モデルの内部に表現された画像全体に対する予測と正しい画像全体との誤差を評価指標として評価を行う．

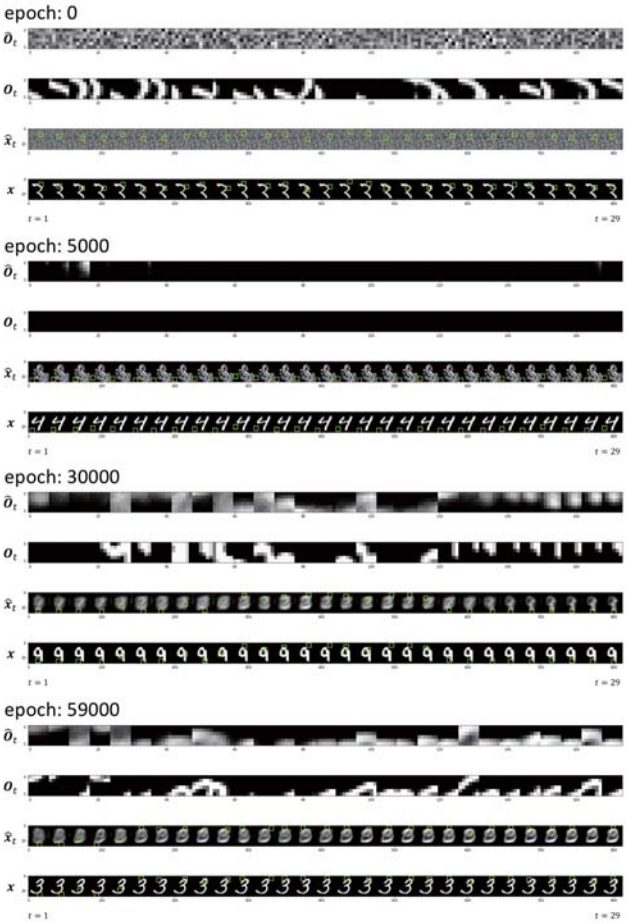

図 2: 提案手法により予測した結果

4.2 結果

図 2 は訓練データに対して，提案手法により予測した結果を表している (図の上からエポックが 0, 5000, 3000, 59000 のときを掲載している)．図の左側から右側に向かって時間ステップが経過している．各エポックに対し，最上段はエージェントによる観測の予測 ($\hat{\boldsymbol{o}}_t$)，2 段目はエージェントが得た実際の観測 ($\boldsymbol{o}_t$)，3 段目はエージェントによる画像全体の予測 ($\hat{\boldsymbol{x}}_t$)，4 段目は実際の画像全体を表している ($\boldsymbol{x}$)．3 段目と 4 段目の緑色の枠は，注意機構により決定された観測の範囲を表している．

学習の初期は，観測の予測モデルが学習されていないため，観測の予測と画像全体に対する予測はランダムなものとなっている (epoch: 0)．ある程度エポックが経過すると，観測の予測をするエージェントの学習が進み，観測の予測は実際の観測と近くなるが，観測の位置は固定してしまっている (epoch: 5000)．その後，注意機構の学習が進み，ステップが進むごとに観測の位置が変化するようになり，画像全体の予測が実際の画像に近づく (epoch: 30000)．学習の終盤では，観測によって画像全体の特徴を反映した予測ができるようになっている (epoch: 59000)．

図 3 は，テストデータに対するステップごとの画像全体に対する予測誤差を示している．提案手法は，ベースライン手法に比べて，少ない観測の回数で高い精度で画像全体を予測することができていることがわかる．

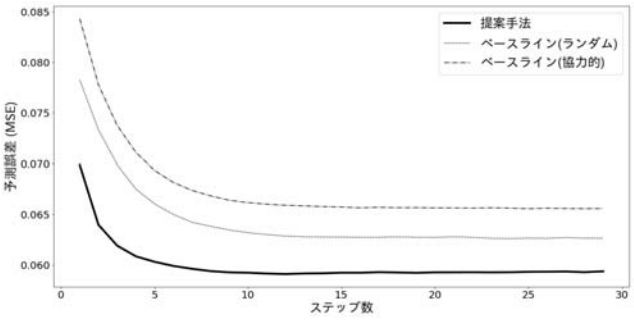

図 3: テストデータ内の画像全体に対する予測誤差

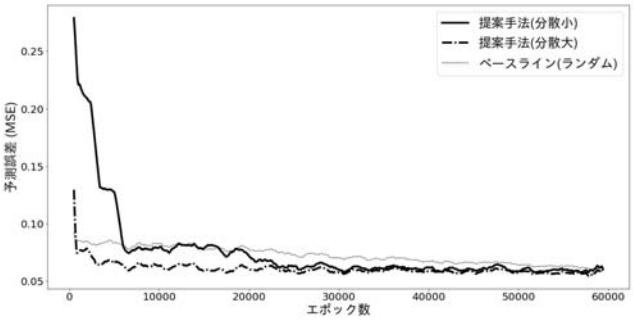

図 4: 提案手法で分散を変更したときの学習曲線

図 4 は，提案手法において，観測の位置の決定に用いる正規分布の分散の値を大きくした場合の学習曲線を表している．分散の値を大きくすると，画像全体に対する予測誤差が学習の早い段階で減少することがわかる

4.3 結果に対する考察

観測の予測誤差を大きくする観測を行うようにする位置の決定に報酬を与えることで，高い効率で状態の予測が可能なことを示した．

提案手法では，学習の初期において観測の予測モデルの学習が進んでおらず，観測の位置が固定してしまう．分散の値を大きくすると，状態の予測誤差が早く減少するようになる．これは，分散の値が大きいほど，実際に観測が行われる位置のランダム性が上がり，もし注意機構の出力する値が固定してしまっても，その値から観測位置が遠くに決定される確率が上昇する．この場合，観測の予測エージェントの予測誤差が大きくなるため，注意機構に対する報酬の値が大きくなり位置の固定から早く抜け出せるようになるためであると考えられる．

以上より，注意機構の出力のランダム性を高め，次第に小さくしてゆくことが有効であると結論づけられる．

5. 考察

今後の研究の方向性として手法のアイデアを動的な状態に適用することが考えられる．これは部分的な観測しか得られない状況下での動画像全体の予測問題となる．

提案手法を実ロボットなどの 3 次元空間内に没入したエージェントの問題へ適用することが考えられる．しかし，実世界の 3 次元空間において，カメラによる画像を入力として用いると，入力の次元が大きく学習が難しくなるため，入力の予測による環境のモデルの構築が難しくなる．そのため，観測やモデル内部での状態表現に対し，何らかの抽象化・離散化が必要になると考えられ，その手法を考案する必要がある．

6. 結論

本研究では，観測に対する予測誤差を注意機構の内的な報酬として与え，観測の予測モデルと注意機構を敵対的な学習により訓練する手法を提案した．

評価実験では，画像のうち一部分が観測として得られる問題設定下で画像全体を予測するタスクにおいて効率的な状態の予測が可能なことを示した．

参考文献

[Goodfellow 16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. MIT Press, 2016.

[Mnih 15] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. Nature, 518(7540):529-533, 2015.

[Silver 17] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George Van Den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go without human knowledge. Nature, 550(7676):354-359, 2017.

[Thrun 05] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. Probabilistic robotics. MIT Press, 2005.

[Mnih 14] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent Models of VisualAttention. In Advances in Neural Information Processing Systems 27, pages 2204-2212, 2014.

[Pathak 17] Pathak, Deepak and Agrawal, Pulkit and Efros, Alexei A. and Darrell, Trevor. Curiosity-Driven Exploration by Self-Supervised Prediction. In Proceedings of the 34th International Conference on Machine Learning, pages 2778-2787, 2017.

[LeCun 98] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner. Gradient-based learning applied to document recognition. In Proceedings of the IEEE, pages 2278-2324, 1998.

[Willia 92] Ronald J. Willia. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. Machine Learning, 8(3):229-256, 1992.