A connection Between GAN(Generative Adversarial Networks) and IRL(Inverse Reinforcement Learning) and Energy-Based Model
GAN と IRL の同義性

Mabonki0725

@AI 論文読会

May 16, 2017

Contents

GAN(Generative Adversial Networks) とは

Generator と Discriminater が相反的に切磋琢磨するモデル

- Generator は Discriminator を騙せる様に訓練
- Disciriminator は Generator の成果を識別できる様に訓練
- 互いが相反的に切磋琢磨しながら双方の精度を上げる仕組み

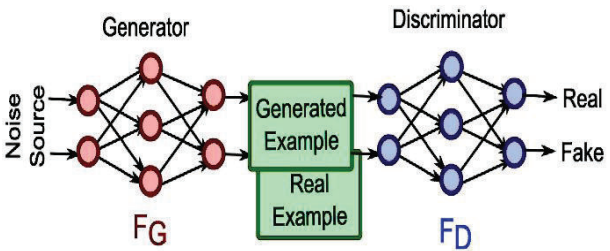
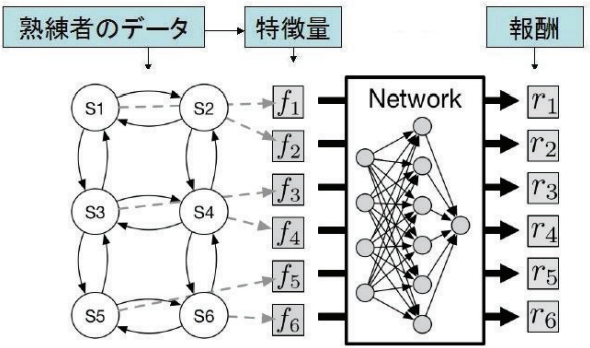


Figure: 1 GAN image

IRL(Inverse Reinforcement Learning) とは

熟練者の選好データより強化学習の報酬を計算するモデル

- 様々な場面の熟練者の選択行動のログを読込む
- 各場面を特徴量に集約する
- 特徴量と選択行動により Network で報酬に変換する



IRL の仮定

仮定 (報酬の仮定)

熟練者の選択過程 τ の確率は報酬 $-c_\theta(\tau)$ の指数に比例する

$$p_\theta(\tau) = \frac{1}{Z(\theta)} \exp\left(-c_\theta(\tau)\right)$$

$$c_\theta(\tau) = \sum_t c_\theta(x_t, u_t)$$

$$\boldsymbol{\tau} = \begin{pmatrix} x_1, x_2, \cdots, x_T \\ u_1, u_2, \cdots, u_T \end{pmatrix}$$

$-c_\theta(\tau)$ は報酬 $c_\theta(\tau)$ は罰則
x_t は時刻 t での局面 x
u_t は時刻 t での行動 u
τ は熟練者の行動過程

IRL の仮定

仮定 (罰則の方が考えやすい)

熟練者の選択過程 τ の確率は報酬 $-c_\theta(\tau)$ の指数に比例する
即ち、確率は罰則 $c_\theta(\tau)$ の指数に反比例する

$$p_\theta(\tau) = \frac{1}{Z(\theta)} \exp\left(-c_\theta(\tau)\right)$$

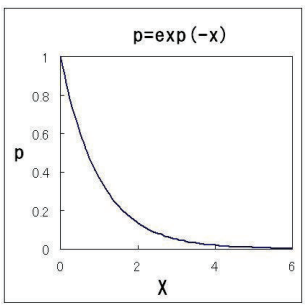


Figure: 罰則 x=0 の時 確率=1 x=∞の時 確率 =0

IRL の解法

熟練者の選択行動があらゆる選択肢の中でどれぐらいの確率かで報酬を決める枠組みは変わらない

IRL の解法の種類

- Max Entropy 法：エントロピー $-\int p_\theta(\log p_\theta)dp_\theta$ の最大化で解く
- Gausian Proces 法：熟練者の選択過程をガウス過程で補間する
- 確率密度比法：熟練者の選択の分布を確率密度比で近似する
- Guide Cost Learning 法

本資料では GAN と構造が似た Guid Cost Learning 法と GAN との同義性を説明する。

Guide cost Learning for IRL

Max Entopy 法で $\mathcal{L}_{cost}(p)$ が最小となる p_θ を Network で訓練して求める

定義 (Cost of IRL)

$$
\begin{aligned}
\mathcal{L}_{cost}(p) &= E_{\tau\sim p}[-\log p_\theta(\tau)] && (1)\\
&= E_{\tau\sim p}[c_\theta(\tau)] + \log Z(\theta) && (2)\\
&= E_{\tau\sim p}[c_\theta(\tau)] + \log \left(E_{\tau\sim q}\left[\frac{\exp(-c_\theta(\tau))}{q(\tau)} \right] \right) && (3)
\end{aligned}
$$

しかし Max Entropy 法では $Z(\theta)$ は不明なので
何らかの確率分布 q を仮定し、q は次の $\mathcal{L}_{sampler}(q)$ で算出する。

Guide cost Learning for IRL

c_θ が最適に近い場合 $Z = \int \exp(c_\theta(\tau))d\theta$ は定数と見做せるので
$q(\tau)$ と $\frac{1}{Z}\exp(-c_\theta(\tau))$ の KL 距離を縮めることを考える
$\mathcal{L}_{sampler}(q)$ を低減させる $q(\tau)$ を Network で訓練して求める

定義 (Sampler of IRL)

$$\begin{aligned}
\mathcal{L}_{sampler}(q) &= KL\left(q(\tau)||\frac{1}{Z}\exp(-c_\theta(\tau))\right) && (4)\\
&= \int q(\tau)\log\frac{\frac{1}{Z}\exp(-c_\theta(\tau))}{q(\tau)}d\tau && (5)\\
&= E_{\tau\sim p}[c_\theta(\tau)] + E_{\tau\sim q}[\log q(\tau)] + \log Z && (6)
\end{aligned}$$

Guide cost Learning では p と q を交互に精緻化する
- $\mathcal{L}_{cost}(p)$ で p_θ
- $\mathcal{L}_{sampler}(q)$ で q

Guide cost Learning for IRL

$q(\tau)$ の擬似的な Importance Sampling は実際の Sample とずれるので
真に近いと考えられる $p(\tau)$ との混合サンプルを使う

$$\mu \sim \frac{1}{2}p(\tau) + \frac{1}{2}q(\tau)$$

$p(\tau)$ は実際にはわからないので、適当な $\tilde{p}(\tau)$ を使う。
例えば、GAN の Generator で生成した $p(\tau)$ とか

定義 (Cost の改善)

$$\mathcal{L}_{cost}(p) = E_{\tau\sim p}[c_\theta(\tau)] + \log\left(E_{\tau\sim\mu}\left[\frac{\exp(-c_\theta(\tau))}{\frac{1}{2}\tilde{p}(\tau) + \frac{1}{2}q(\tau)}\right]\right) \quad (7)$$

GAN の disciminator

本物の $p(\tau)$ と偽者の $q(\tau)$ とを GAN で近似したい場合
Discriminater D^* は以下の定義を使う

定義 (GAN Discriminater)

$$D^*(\tau) = \frac{p(\tau)}{\frac{1}{2}p(\tau) + \frac{1}{2}q(\tau)} \tag{8}$$

ここで $p(\tau)$ は以下の報酬関数とする

$$p(\tau) = \frac{1}{Z}\exp(-c_\theta(\tau))$$

定義 (GAN Discriminater for θ)

$$D_\theta(\tau) = \frac{\frac{1}{Z}\exp(-c_\theta(\tau))}{\frac{1}{2Z}\exp(-c_\theta(\tau)) + \frac{1}{2}q(\tau)} \tag{9}$$

GAN の disciminator Loss

定義 (Loss of Discrimater)

$$\begin{aligned}
\mathcal{L}_{discriminator}(D_\theta) &= E_{\tau\sim p}[\log D_\theta(\tau)] - E_{\tau\sim p}[\log(1 - D_\theta(\tau))] \quad (10)\\
&= E_{\tau\sim p}\left[-\log \frac{\frac{1}{Z}\exp(-c_\theta(\tau))}{\frac{1}{2Z}\exp(-c_\theta(\tau)) + \frac{1}{2}q(\tau)}\right]\\
&- E_{\tau\sim p}\left[-\log \frac{q(\tau)}{\frac{1}{2Z}\exp(-c_\theta(\tau)) + \frac{1}{2}q(\tau)}\right] \quad (11)
\end{aligned}$$

Discriminator の Network はこの Loss が減少する様に訓練する

Estimeate Z

ここで

$$\tilde{\mu} = \frac{1}{2Z}\exp(-c_\theta(\tau)) + \frac{1}{2}q(\tau)$$

とおくと

定義 (Discriminater)

$$\begin{aligned}
\mathcal{L}_{discriminator}(D_\theta) &= E_{\tau\sim p}[\log D_\theta(\tau)] - E_{\tau\sim p}[\log(1 - D_\theta(\tau))] \quad (12)\\
&= E_{\tau\sim\mu}\left[\frac{\frac{1}{Z}\exp(-c_\theta(\tau))}{\tilde{\mu}}\right]\\
&- E_{\tau\sim q}\left[-\log\frac{q(\tau)}{\tilde{\mu}}\right] \quad (13)\\
&= \log Z + E_{\tau\sim p}[c_\theta(\tau)] + E_{\tau\sim p}[\log\tilde{\mu}(\tau)]\\
&- E_{\tau\sim q}[\log q(\tau)] + E_{\tau\sim q}[\log\tilde{\mu}(\tau)] \quad (14)
\end{aligned}$$

Estimeate Z

式 (14) より

$$\begin{aligned}\mathcal{L}_{discriminator}(D_\theta) &= \log Z + E_{\tau\sim p}[c_\theta(\tau)] + E_{\tau\sim p}[\log\tilde{\mu}(\tau)] \\ &- E_{\tau\sim q}[\log q(\tau)] + E_{\tau\sim q}[\log\tilde{\mu}(\tau)]\end{aligned}$$

ここで $\mathcal{L}_{discriminater}(D_\theta)$ を Z で微分して、最適な Z を求める

derivative Discriminater wiht z

$$\partial_z\mathcal{L}_{discriminator}(D_\theta) = \frac{1}{Z} - E_{\tau\sim\mu}\left[\frac{\frac{1}{Z^2}\exp(-c_\theta(\tau))}{\tilde{\mu}}\right] \tag{15}$$

$$\partial_z\mathcal{L}_{discriminator}(D_\theta) = 0 \tag{16}$$

$$Z = E_{\tau\sim\mu}\left[\frac{\exp(-c_\theta(\tau))}{\tilde{\mu}}\right] \tag{17}$$

Derivative Discriminater

式 (14) より

$$\begin{aligned}\mathcal{L}_{discriminator}(D_\theta) &= \log Z + E_{\tau\sim p}[c_\theta(\tau)] + E_{\tau\sim p}[\log\tilde{\mu}(\tau)] \\ &- E_{\tau\sim q}[\log q(\tau)] + E_{\tau\sim q}[\log\tilde{\mu}(\tau)]\end{aligned}$$

こんどは $\mathcal{L}_{discriminater}(D_\theta)$ を θ で微分する。

derivative Discriminater with θ

$$\begin{aligned}\partial_\theta\mathcal{L}_{discriminator}(D_\theta) &= E_{\tau\sim p}[\partial_\theta c_\theta(\tau)] \\ &- E_{\tau\sim\mu}\left[\frac{\frac{1}{Z}\exp(-c_\theta(\tau)\partial_\theta c_\theta(\tau)}{\tilde{\mu}}\right] \qquad (18)\end{aligned}$$

Derivative IRL cost

式 (7) より

$$\mathcal{L}_{cost}(\theta) \quad = \quad E_{\tau\sim p}[c_\theta(\tau)] + \log\left(E_{\tau\sim\mu}\left[\frac{\exp(-c_\theta(\tau))}{\tilde{\mu}(\tau)}\right]\right)$$

IRL のコスト $\mathcal{L}_{cost}(\theta)$ を θ で微分する。ここで式 (17) の Z を使う。

derivative cost with θ

$$\begin{aligned}
\partial_\theta\mathcal{L}_{cost}(\theta) &= E_{\tau\sim p}[\partial_\theta c_\theta(\tau)] + \partial_\theta \log E_{\tau\sim\mu}\left[\frac{\exp(-c_\theta(\tau))}{\tilde{\mu}(\tau)}\right] && (19)\\
&= E_{\tau\sim p}[\partial_\theta c_\theta(\tau)] \\
&- \left(E_{\tau\sim\mu}\left[\frac{\exp(-c_\theta(\tau))\partial_\theta c_\theta(\tau)}{\tilde{\mu}(\tau)}\right] / E_{\tau\sim\mu}\left[\frac{\exp(-c_\theta(\tau))}{\tilde{\mu}(\tau)}\right]\right) \\
&= E_{\tau\sim p}[\partial_\theta c_\theta(\tau)] - \left(E_{\tau\sim\mu}\left[\frac{\exp(-c_\theta(\tau))\partial_\theta c_\theta(\tau)}{\tilde{\mu}(\tau)}\right] / Z\right) && (20)\\
& && (21)
\end{aligned}$$

Conclusion IRS cost and GAN discriminator

Derivative IRL cost = Derivative GAN discriminator

式 (18) より

$$\begin{aligned}\partial_\theta \mathcal{L}_{discriminator}(D_\theta) &= E_{\tau\sim p}[\partial_\theta c_\theta(\tau)] \\ &- E_{\tau\sim\mu}\left[\frac{\frac{1}{Z}\exp(-c_\theta(\tau)\partial_\theta c_\theta(\tau)}{\tilde{\mu}}\right]\end{aligned}$$

$$\begin{aligned}\partial_\theta \mathcal{L}_{cost}(\theta) &= E_{\tau\sim p}[\partial_\theta c_\theta(\tau)] - \left(E_{\tau\sim\mu}\left[\frac{\exp(-c_\theta(\tau))\partial_\theta c_\theta(\tau)}{\tilde{\mu}(\tau)}\right]/Z\right) \\ &= E_{\tau\sim p}[\partial_\theta c_\theta(\tau)] - E_{\tau\sim\mu}\left[\frac{\frac{1}{Z}\exp(-c_\theta(\tau))\partial_\theta c_\theta(\tau)}{\tilde{\mu}(\tau)}\right] \quad (22) \\ &= \partial_\theta \mathcal{L}_{discriminator}(D_\theta) \quad (23)\end{aligned}$$

Conclusion IRS sampler and GAN generator

IRL sampler

式 (6) より

$$\mathcal{L}_{sampler}(q) \quad = \quad E_{\tau\sim p}[c_\theta(\tau)] + E_{\tau\sim q}[\log q(\tau)]$$

GAN generator = IRS sampler + Constant

$$\begin{aligned}
\mathcal{L}_{generater}(q) &= E_{\tau\sim q}[\log(1 - D(\tau)) - \log D((\tau))] && (24)\\
&= E_{\tau\sim q}\left[\log \frac{q(\tau)}{\tilde{\mu}(\tau)} - \log \frac{\frac{1}{Z}\exp(-c_\theta(\tau))}{\tilde{\mu}(\tau)}\right] && (25)\\
&= E_{\tau\sim q}[\log q(\tau) + \log Z + c_\theta(\tau)] && (26)\\
&= \log Z + E_{\tau\sim q}[c_\theta(\tau)] + E_{\tau\sim q}[\log q(\tau)] && (27)\\
&= \log Z + \mathcal{L}_{sampler}(q) && (28)
\end{aligned}$$

Conclusion

network 構造の相似

- IRL は $\mathcal{L}_{cost}$ で q_θ を $\mathcal{L}_{sampler}$ で q を 交互に
  精緻化して報酬関数 $-c_\theta(\tau)$ を求める
- GAN は $\mathcal{L}_{generater}$ と $\mathcal{L}_{discriminator}$ 交互で
  本物 $q(\tau) =$ 偽者 $p(\tau)$ とする

IRL と GAN の同義性

- IRL と GAN の勾配は等しい $\partial_\theta \mathcal{L}_{cost} = \partial_\theta \mathcal{L}_{discriminator}$
- IRL と GAN の差は一定 $\mathcal{L}_{sampler}(q) + \log Z = \mathcal{L}_{generator}(q)$

Proposal

GAN で偽の熟練者行動を多数生成すれば、熟練者の行動の確率の精度が高められる。

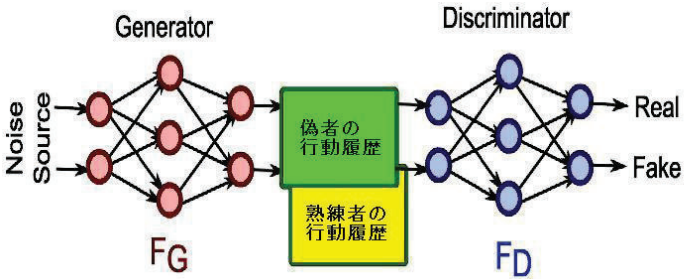


Figure: 熟練者の行動履歴の生成

$$p_\theta(\tau) = \frac{1}{Z(\theta)} \exp\left(-c_\theta(\tau)\right)$$

熟練者の行動パターンの全 $Z(\theta)$ が求まれば c_θ の精度は高められる。