特別賞

人間 GAN：人間の知覚的識別を利用して学習するニューラルネットワーク

徳山工業高等専門学校　情報電子工学専攻 1 年

藤井 一貴

1. 緒　言

1.1　研究背景

生産性や Quality of Life 向上を目指し、人工知能技術の社会実装が進められている。人工知能技術を支える現在の機械学習は、事前にセンシングしたデータ（画像・音声・言語など）に基づいて機械学習モデル（ニューラルネットワークなど）を学習する**データドリブン型**に分類される。データドリブン型では、事前センシングしたデータ（学習データ）に顕在・潜在する情報を機械学習モデルを用いて獲得する。一方で、人工知能の社会実装がさらに進んだ際には、人間と人工知能の間の密な相互通信が行われると予想される。すなわち、現在のチャットボットのような浅いコミュニケーションではなく、人間同士のコミュニケーションがそうであるように、人間と人工知能が互いの性質を理解した上での深いコミュニケーションが行われる。そのような将来を見据え、**新たな機械学習アプローチとして人間ドリブン型を確立**する必要がある。これは、人間が何を感じるか・知っているか（究極的には、人間とはなにか）を用いて機械学習モデルを学習する方法である。本研究では、その第一歩として、**人間の知覚的識別を利用してニューラルネットワークを学習する方法を模索**する。

データドリブン型機械学習における非常に強力な手法が、**敵対的生成ネットワーク**（**generative adversarial network; GAN**）[1]である。GAN は、深層ニューラルネットワーク（deep neural network; DNN）の一種であり、その学習過程において、データの実在する範囲（実在データ分布）を表現する DNN（生成モデル）を学習する。この枠組みでは、生成モデルと別に識別モデルも利用する。生成モデルは識別モデルを詐称するように学習され、識別モデルは実在データと生成データを識別するように学習される。これらを繰り返すことで、最終的に生成モデルは実在データ分布に従うデータをランダムにサンプリングすることができる。この枠組みは、画像[2]、音声[3]、言語[4]などの分野において適用されており、実在人物と見間違えるほど精微な顔画像を生成できることでも知られる。**この枠組みを人間ドリブン型に拡張することで、DNN の強力な表現能力により、人間の知覚を計算機で表現できるようになる**と期待される。

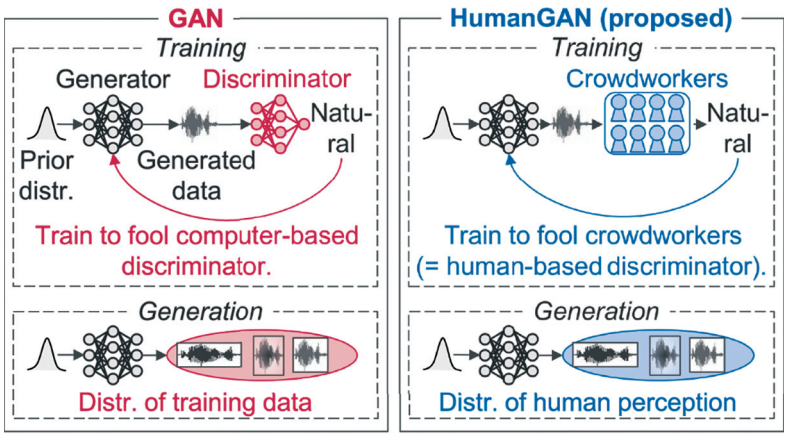

図1　通常の GAN と提案する人間 GAN の比較。通常の GAN では、データから学習された識別モデルを用いて生成モデルを学習する。人間 GAN では、人間を識別モデルとみなして生成モデルを学習する。

1.2 研究目的

本研究では、**人間ドリブン型機械学習における GAN（人間 GAN）**を確立する。すなわち、**通常の GAN のようにデータにアクセスしてその範囲（実在データ分布）を表現する DNN ではなく、人間にアクセスして人間が知覚できる分布（知覚分布）を表現する DNN を学習**する。**Fig. 1**に、通常の GAN と提案する人間 GAN の違いを示す。通常の GAN では、DNN で記述される識別モデルを詐称するのに対して、人間 GAN では、クラウドソーシングサービスの人的リソースから得られる知覚評価を詐称する。人間を「DNN からの生成データに対して、事後確率（すなわち、生成データに対する知覚度合い）の差分を出力する black-box システム」とみなすことで、人間を計算過程に含めた DNN 学習が可能となる。本研究では、この枠組みを理論的に定式化し、その有効性を実験的に確認する。実験的評価の結果、人間に直接的にアクセスして DNN を学習できることを明らかにする。

2. 通常の GAN（従来のデータドリブン型アプローチ）

通常の GAN の目的は、生成モデルの表現する分布を学習に用いられた実在データの分布に一致させることである。そのため、通常の GAN は、データを生成する生成モデル $G(\cdot)$ と、実在データと生成データを識別する識別モデル $D(\cdot)$ を用いる。$G(\cdot)$ と $D(\cdot)$ は DNN で記述される。ここで、N 個の実在データを $x = [x_1, \cdots, x_n, \cdots, x_N]$ とする。生成モデル $G(\cdot)$ は、既知の確率分布（例えば、一様分布）に従う N 個の乱数 $z = [z_1, \cdots, z_n, \cdots, z_N]$ を、生成データ $\hat{x} = [\hat{x}_1, \cdots, \hat{x}_n, \cdots, \hat{x}_N]$ に写像する。識別モデル $D(\cdot)$ は、実在データ x_n もしくは生成データ $\hat{x}_n$ を入力とし、当該入力が実在データである事後確率を出力する。学習時の目的関数 $V(\cdot)$ は次式となる。

$$V(G, D) = \sum_{n=1}^{N} \log D(\boldsymbol{x}_n) + \sum_{n=1}^{N} \log(1 - D(G(\boldsymbol{z}_n))) \tag{1}$$

生成モデル　$G(\cdot)$ はこの関数値を最小化するように、識別モデル $D(\cdot)$ は最大化するように学習される。

3. 人間 GAN（提案する人間ドリブン型アプローチ）

人間 GAN は、**Fig. 1**に示すように、通常の GAN の識別モデルを人間の知覚評価で置換した手法である。生成モデル $G(\cdot)$ が DNN で記述され、既知の確率分布に従う乱数を入力とすることは、通常の GAN と人間 GAN で共通する。一方で、通常の GAN では DNN で記述された識別モデルを詐称して生成モデルを学習するのに対し、**人間 GAN では人間による知覚評価を詐称して生成モデルを学習**する。ここで、$D(\cdot)$ を人間による評価を表す知覚モデルとして再定義する。この $D(\cdot)$ は、$G(\cdot)$ から生成されたデータ $\hat{x}_n$ を入力とし、「当該入力をどの程度知覚できるか」を 0 から 1 の値で事後確率として出力する。学習時の目的関数 $V(\cdot)$ は次式となる。

$$V(G, D) = \sum_{n=1}^{N} D(G(\boldsymbol{z}_n)) \tag{2}$$

$G(\cdot)$ のモデルパラメータ θ_{G} は、Eq.(2) を最大化するように学習される。ここで、勾配法による反復的学習法を考える。すなわち、θ_{G} は次式で反復的に更新される。

$$\theta_{\mathrm{G}}^{(\mathrm{new})} = \theta_{\mathrm{G}} + \alpha \frac{\partial V(G, D)}{\partial \theta_{\mathrm{G}}} \tag{3}$$

ここで、α は学習係数である。第2項の $\partial V(G, D)/\partial \theta_{\mathrm{G}}$ は、$\partial V(G, D)/\partial \hat{x}$ と $\partial \hat{x}/\partial \theta_{\mathrm{G}}$ の積となる。通常の GAN では、$G(\cdot)$ と $D(\cdot)$ の両方の計算過程が微分可能であるため、この式は backpropagation を用いて実行することができる。しかしながら、人間 GAN では、$D(\cdot)$ が微分不可能であり、$\partial V(G, D)/\partial \hat{x}$ を推定できないため、backpropagation による学習は不可能である。そこで、人間を「生成データに対して事後確率差分を出力する black-box システム」とみなした black-box 最適化手法に基づいて $\partial V(G, D)/\partial \hat{x}$ を推定する。本研究では、摂動を用いて勾配を近似する Natural Evolution Strategies（NES）[5] を用いた学習アルゴリズムを提案する。手法の概要図を **Fig. 2** に示す。

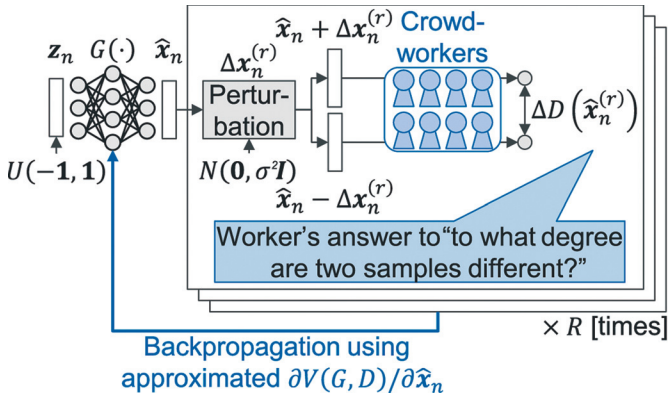

図2　人間 GAN における生成モデルの学習。人間を「生成データに対して事後確率差分を出力する black-box システム」とみなすことで目的関数に対する生成データの勾配を近似できるため、従来のデータドリブン型アプローチと同様に backpropagation を用いて生成モデルを学習できる。

まず、生成データ $\hat{x}_n$ に対し、多変量正規分布 $\mathcal{N}(0,\ \sigma^2 I)$ からランダムに生成した摂動 $\Delta x_n^{(r)}$ を付与する。ここで、σ は定数の標準偏差、r は摂動のインデックス（$1 \leq r \leq R$）、I は単位行列である。次に、摂動後の2つのデータ $\left\{\hat{\boldsymbol{x}}_n + \Delta \boldsymbol{x}_n^{(r)}, \hat{\boldsymbol{x}}_n - \Delta \boldsymbol{x}_n^{(r)}\right\}$ を人間に提示し、それらのデータの事後確率の差分

$$\Delta D(\hat{\boldsymbol{x}}_n^{(r)}) \equiv D\left(\hat{\boldsymbol{x}}_n + \Delta \boldsymbol{x}_n^{(r)}\right) - D\left(\hat{\boldsymbol{x}}_n - \Delta \boldsymbol{x}_n^{(r)}\right) \tag{4}$$

を回答させる。$\Delta D(\hat{x}_n^{(r)})$ は -1 から 1 までの値をとる。例えば、2つのデータのうち $\hat{x}_n + \Delta x_n^{(r)}$ の事後確率が非常に高い場合に人間は $\Delta D(\hat{x}_n^{(r)}) = 1$ を返し、$\hat{x}_n - \Delta x_n^{(r)}$ の事後確率が非常に高い場合に人間は $\Delta D(\hat{x}_n^{(r)}) = -1$ を返すものとする。この摂動・評価を、ある生成データ $\hat{x}_n$ に対して R 回繰り返す。最終的には、N 個の生成データに対して上記の過程を繰り返す。

Backpropagation に用いる勾配 $\partial V(G, D)/\partial \hat{x}$ は、次式で近似される [5]。

$$\frac{\partial V(G,D)}{\partial \hat{\boldsymbol{x}}} = \left[\frac{\partial V(G,D)}{\partial \hat{\boldsymbol{x}}_1}, \cdots, \frac{\partial V(G,D)}{\partial \hat{\boldsymbol{x}}_n}, \cdots, \frac{\partial V(G,D)}{\partial \hat{\boldsymbol{x}}_N}\right] \tag{5}$$

$$\frac{\partial V(G,D)}{\partial \hat{\boldsymbol{x}}_n} = \frac{1}{2\sigma^2 R}\sum_{r=1}^{R} \Delta D(\hat{\boldsymbol{x}}_n^{(r)}) \cdot \Delta \boldsymbol{x}_n^{(r)} \tag{6}$$

ここで、生成モデルの学習に係る人間への問い合わせ回数を整理する。学習の各反復における各生成データに対してR回の摂動を加えるため、最終的な問い合わせ回数は、勾配法の反復回数・生成データ数N・摂動回数Rの積となる。

4. 実験的評価

人間GANが知覚分布を表現できることを検証するために、音声の自然性に関する実験的評価を行う。具体的には、人間が当該音声を人間の音声と知覚できる範囲を知覚分布とし、その知覚分布をDNNで表現できることを確認する。本実験では、身体情報付き男・女・子どもの母音音声データベース(JVPD)[6]に含まれる女性199名の音声特徴量の固有声成分(第1, 2主成分)をデータ空間とした。音声特徴量は、音声分析システムWORLD[7, 8]を用いて抽出した。これ以外の実験条件については、発表文献参照とする。

4.1 実在データ分布と知覚分布の違いの確認

まず、**人間GANの必要性を実験的に確認するために、実在データ分布と知覚分布が異なることを示す。**2次元の空間を1.0毎にグリッド分割し、各グリッドにおける自然性の知覚度合(すなわち、自然性の事後確率)を評価する。人間による評価には、クラウドソーシングサービスLancers[9]を使用する。人間への問い合わせ時には、「人間らしくない音声であれば(1)を、人間らしい音声であれば(5)を選択してください。」というインストラクションを示し、人間に　1つの音声の知覚度合を5段階で回答させる。ここで得られた回答の1から5のそれぞれの値を、事後確率$D(\hat{x}_n)$の0.00, 0.25, 0.50, 0.75, 1.00に対応させる。各グリッドの事後確率の値は、複数の評価値の平均とする。評価者の数は、96人である。

Fig. 3に結果を示す。実在データは白丸で囲んだ範囲にのみ分布する。通常のGANは、この範囲を表現可能である。一方で、知覚分布(**Fig. 3**の明るい箇所)はこの範囲に留まらず、より広い領域に対応することがわかる。この結果は、人間が実在音声より広い範囲を人間の音声と知覚することを表している。**このような知覚分布は、実在データ分布のみを表現する通常のGANでは表現不可能である。すなわち、人間にアクセスして学習する人間GANの枠組みが必要であることが明らかである。**

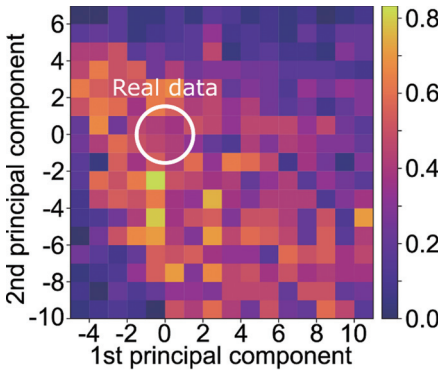

図3　固有声空間の事後確率を表すカラーマップ。グリッドの色が明るいほど、当該音声がより人間らしいと人間に知覚されたことを表す。実在の音声データは白丸で囲われた範囲にのみ分布しているが、明るい色のグリッド（すなわち、人間らしいと評価された音声）は、より広い範囲に分布している。

4.2　生成モデル学習による生成データの変化

次に、**生成モデル学習（DNN 学習）の様子を定性的に確認するために、学習の各反復における生成データをプロット**する。生成モデルに入力する乱数は、2次元の一様分布 $U(-1,\ 1)$ に従うものとする。この乱数は、学習のはじめにランダムに生成され、学習中は固定される。生成モデルは、2ユニットの入力層、2×4ユニットの sigmoid 隠れ層、2ユニットの線形出力層を持つ Feed-Forward neural network である。最適化アルゴリズムとして学習率 $\alpha = 0.0015$ の確率的勾配降下法を用いる。生成する音声特徴量数 N は100、摂動回数 R は5、学習の反復回数は4とする。NES の標準偏差 σ は1.0とする。人間への問い合わせ時には、「2つの音声を聴いて、1個目の方が人間らしいほど(1)を、2個目の方が人間らしいほど(5)を選択してください。両方同程度なら(3)を選択してください。」というインストラクションを示し、人間に2つの音声の知覚度合の差異を5段階で回答させる。ここで得られた回答の1から5の値を、それぞれ事後確率の差分 $\Delta D(\hat{x}_n^{(r)})$ の $1.0,\ 0.5,\ 0.0,\ -0.5,\ -1.0$ に対応させる。

Fig. 4に学習の様子を示す。各点は生成データである。可視化のため、Fig. 3の事後確率を重ねて表示しているが、学習そのものに Fig. 3の結果は使用していないことに注意する。これらの図より、学習の反復により、生成データは事後確率の高い領域に遷移していることがわかる。故に、**人間 GAN は、知覚分布を表現するような生成モデル学習を可能にすることが定性的に確認**できる。

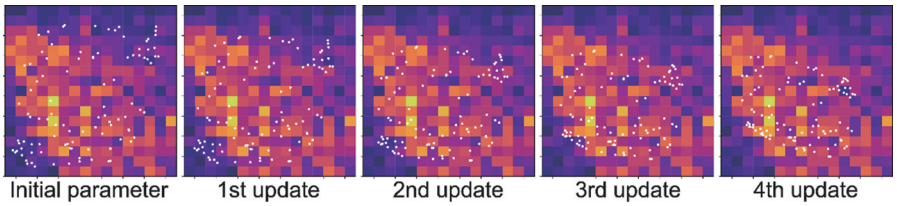

図4　各学習回数の様子。白い点が生成データである。生成データが明るい色のグリッドに移動していることが分かる。ただし、学習時にはグリッドの情報を使用していないことに注意する。

4.3　生成モデル学習による事後確率の増加

最後に、**学習の反復により、目的関数である事後確率が増加することを定量的に確認**する。ここでは、学習に用いた乱数とは別の乱数を発生させ、新たな特徴量を生成する。以降では、学習に使用した乱数で生成した特徴量を closed データ、学習に使用していない乱数で生成した特徴量を open データと呼ぶ。Open データに対する事後確率は、**Section 4.1** と同様に、クラウドソーシングサービス上で人間に評価させる。

Fig. 5 に、各反復における事後確率の箱ひげ図を示す。この図から、学習の反復により、closed データのみならず、open データに対しても事後確率の増加が見られる。故に、**人間 GAN により学習された生成モデルが知覚分布を表現しうることを定量的に確認**できる。

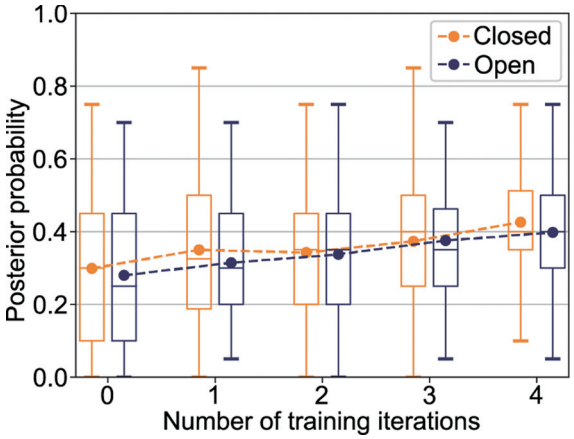

図5　各学習回数における事後確率。値が高いほど、当該音声がより人間らしいと人間に知覚されたことを表す。

5. まとめ

本研究では、人間の知覚に基づいてニューラルネットワークを学習する人間ドリブン型アプローチとして、人間 GAN を提案した。音声を例として実験的評価を行った結果、人間に直接的にアクセスして DNN を学習できることを明らかにした。この研究成果は、人間の知覚を用いて直接的に人工知能を学習するための重要な第一歩である。また、仮想現実感(VR)研究のような、人間の感覚・知覚を計算機で制御するような研究分野に対しても、強く貢献できると期待される。なお、この研究成果と将来性は国際的に評価されており、音声分野の世界最高峰国際会議 ICASSP2020 におけるオーラル発表を予定している。

謝　辞

本研究は、東京大学 大学院情報理工学系研究科 猿渡 洋教授・高道 慎之介助教、筑波大学 情報学群情報科学類 馬場 雪乃准教授の指導のもと実施したものである。

参考文献

［1］ I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Proc. NIPS*, Montreal, Canada, Dec. 2014, pp. 2672-2680.

［2］ M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *arXiv:1411.1784*, 2015.

［3］ Y. Saito, S. Takamichi, and H. Saruwatari, “Statistical parametric speech synthesis incorporating generative adversarial networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84-96, Jan. 2018.

［4］ L. Yu, W. Zhang, J. Wang, and Y. Yu, “SeqGAN: Sequence generative adversarial nets with policy gradient,” in *Proc. AAAI*, San Francisco, U.S.A., Feb. 2017.

［5］ A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, “Black-box adversarial attacks with limited queries and information,” in *Proc. ICML*, vol. 2, Stockholm, Sweden, Jul. 2018, pp. 2137-2146.

［6］ “Vowel database: Five Japanese vowels of males, females, and children along with relevant physical data（JVPD）,” http://research.nii.ac.jp/src/en/JVPD.html.

［7］ M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877-1884, Jul. 2016.

［8］ M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis,” *Speech Communication*, vol. 84, pp. 57-65, Nov. 2016.

［9］ “Lancers,” https://www.lancers.jp/.