2M-03

等価性構造抽出を用いたニューラルネットワークの隠れ層に対する分析

山岸 健太 * 高橋 達二 * 佐藤 聖也 * †

東京電機大学 * 産業技術総合研究所 †

1. はじめに

計算機の発達や多層のニューラルネットワークを用いた機械学習の Deep Learning の影響で, ニューラルネットワークが様々な分野に使われるようになった. しかし, ニューラルネットワークの内部処理はブラックボックスのため, 医療現場などで人命が関わるような重要な場面での活用や判断を人間ができないところに問題がある.

そこで, 人間がニューラルネットワークの隠れ層の内部表現を理解するために二つのニューラルネットワークを用い, 部分的に等価と思われる表現を発見することは有用となりえる.

この発見に等価性構造抽出 [1] が利用できる可能性があるが, ニューラルネットワークの隠れユニットには冗長性等の性質のため直接は利用できない. そのため, 本稿では前処理としてニューラルネットワークの隠れユニットの出力に対して独立成分分析を用いることにより, その後等価性構造抽出を適用できるかを検証する. 等価性構造抽出が適用できれば複数のニューラルネットワーク間の内部表現に等価と思われる構造を発見でき, ニューラルネットワークの内部表現の分析に利用できる可能性がある.

2. 等価性構造抽出

等価性構造抽出技術は, 多次元の系列データの間から, 等価な関係を見出す技術である [1]. この技術はある多次元系列上と, 別の多次元系列上に現れる時系列上のパターンに着目し, 共通のパターンをもっている系列の ID で構成される組を見つけるものである. これをタプルとも呼ぶ. 等価性構造抽出技術の入力は, 系列の集合であり, 出力は等価性構造と呼ばれる集合である. 等価性構造は K 個の系列 ID で構成される, K-タプルを要素とする集合であり, どの K 系列で構成されるタプルを等価と見做すことができるのかを示す. また, どのタプルとどのタプルが等価か決定する方法としては, K 次元系列の部分系列に基づいて決定する.

3. 独立成分分析

3.1 独立成分分析

再構成 ICA は入力から出力に対して線形変換を行う. この変換は出力の特徴量の独立性と再構成能力を合わせるように非線形目的関数を最適化する. 再構成におけるペナルティ項の考慮などをすることは本研究の目的として好ましくないため, 今回は FastICA[2] を用いて実験する.

3.2 使用目的

ニューラルネットワークの隠れユニットは冗長性等の性質により, 各機能（例えば動きや大きさを識別する機能）を一つの隠れユニットにより表現可能であったとしても複数の隠れユニットにより表現するような最適解に多くの場合収束すると思われる. そのため本研究では学習済みのニューラルネットの隠れユニットの出力に対して FastICA を行い, 複数の隠れユニットが表現する複数の機能を分離することを試みる.

4. 実験

4.1 使用したデータセット

本実験では DeepMind が github 上で公開している Disentanglement testing Sprites dataset [3] というデータセットを用いる. このデータセットは 6 つの属性に依存している. この 6 つの属性は色, 形, 大きさ, 角度, x 座標, y 座標である. 色は白の 1 種類, 形はハートと楕円形と四角の三種類, 大きさは 6 種類, 角度は 40 種類用意されている. x 座標と y 座標はどちらも 1 32 まで指定できる.

ここでは上述のデータセットから取り出した 2 つのデータセットを用いる（それぞれデータセット 1 , 2 とする）. データセット 1 の属性の変化をデータセット 2 の属性の変化が内包するように, 大きさ, x 座標, y 座標の 3 つのみが変化するデータセット 1 と形, 大きさ, x 座標, y 座標の 4 つが変化する画像の集合のデータセット 2 を作成する.

データセット 1 , 2 は一つの動画である. だが, ニューラルネットワークの学習時は動画としては扱わず, フレーム番号毎に動画を分割した単に画像の集合として扱い, 学習後に FastICA を適用させる際はこれらのデータセットを時系列データとして用いるため動画として扱う. データセット 1 とデータセット 2 でのフレーム番号毎の属性の大きさの値は同じである.

図 1, 図 2 に実験で用いた画像の上部にフレーム番号、下部と左部に x, y 座標を表示した画像を示す。

4.2 実験内容

学習時, 入力信号としたものは上述の画像群であり, 教師信号にはデータセットに用意されている潜在因子値の整数インデックスである分類ターゲットを使用した. また, FastICA を用いる際に時系列データを用いる必要があるため, これらのデータセットをあたかも時系列データのように扱っている. 本研究での隠れユニット数は, 学習の際に平均二乗誤差が最も小さくなった数の 4 としている.

Equivalence structure extraction on hidden layers of neural networks

*Kenta Yamagishi, Tatsuji Takahashi, Seiya Satoh, Tokyo Denki University

† Seiya Satoh, National Institute of Advanced Industrial Science and Technology

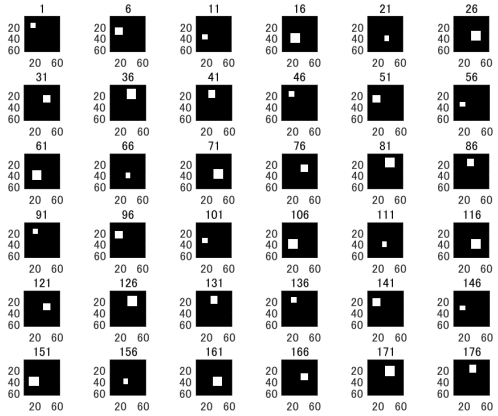

図. 1　データセット 1

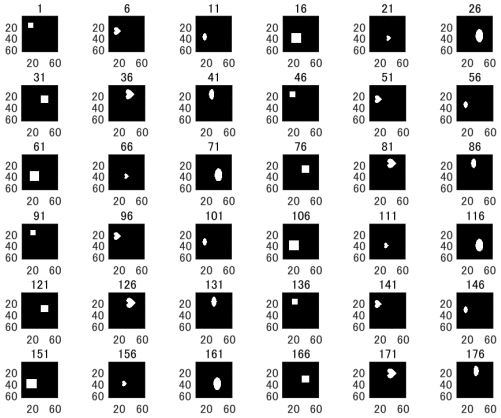

図. 2　データセット 2

5.　結果

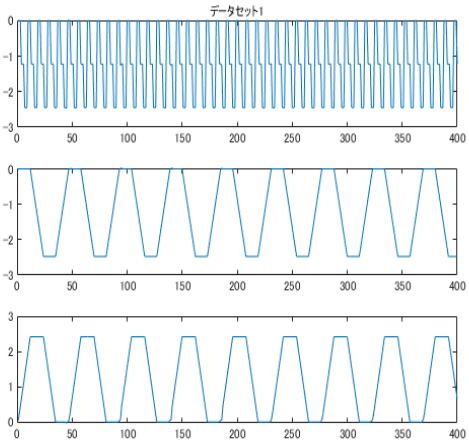

図. 3　隠れユニットの出力に FastICA を適用した結果（データセット 1）

図 1, 図 2 のデータセットを用いて学習した後の隠れ層に対して FastICA を適用させた結果, 図 3, 図 4 のような出力が得られた.

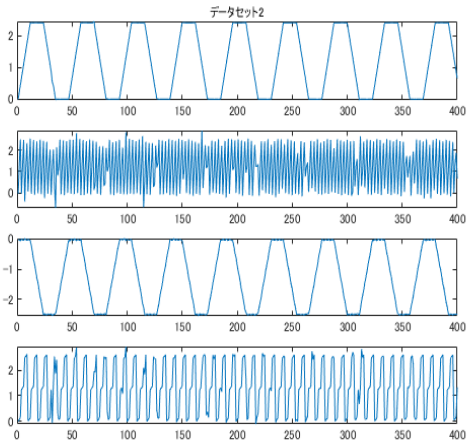

図. 4　隠れユニットの出力に FastICA を適用した結果（データセット 2）

6.　考察

モデル 1 では図形が位置する x 座標, y 座標, 図形の大きさの 3 つの属性を学習させ, モデル 2 ではこの 3 つの属性に加え, 変化する図形の形を学習させた. 図 3, 図 4 の結果を見ると, 図 3 では 3 つ, 図 4 では 4 つの属性に分離されていることがわかる. また, 図 4 では図 3 と似ているグラフが出力されていることから, データセット 2 の隠れ層はデータセット 1 の隠れ層の表現を内包し, 唯一表れていないグラフが図形の形を認識する属性を持っている可能性が高いと考えられる.

7.　おわりに

独立主成分分析を用いた結果, 二つのデータセットから等価性構造抽出を適用できる可能性が高い出力を取得できた.

本研究では二つのデータセットが包含関係になっている場合の実験を行ったが, 今後は二つのデータセットの属性の変化が一部を有している場合について考えることも重要であると考える.

謝辞

本研究の成果は, 国立研究開発法人新エネルギー・産業技術総合開発機構（ＮＥＤＯ）の委託業務の結果得られたものです.

参考文献

[1] Seiya Satoh, Yoshinobu Takahashi, Hiroshi Yamakawa: Accelerated Equivalence Structure Extraction via Pairwise Incremental Search, *KDD'18 Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2160-2169. (2018)

[2] The FastICA package for MATLAB. `https://research.ics.aalto.fi/ica/fastica/`

[3] Deep Mind: dSprites - Disentanglement testing Sprites dataset. `https://github.com/deepmind/dsprites-dataset/`