6C-6

マルチモーダル・データセットを対象とした深層学習の性能評価に関する基礎研究

玉城翔 † 當間愛晃 ‡ 赤嶺有平 ‡ 山田孝治 ‡ 遠藤聡志 ‡
† 琉球大学大学院理工学研究科情報工学専攻 ‡ 琉球大学工学部情報工学科

1 はじめに

1.1 研究背景

2012 年頃から機械学習の分野で、深層学習という言葉が流行りだしてきた。それは、各分野のコンテストにおいて深層学習を使った技術が、これまでの技術を圧倒する精度を誇って優勝をしてきたことにある。現在、音声認識や画像認識といった分野で、高い精度を誇っている深層学習ですが、今後はより複雑なパターン認識の応用が考えられる。複雑なものの例として、マルチモーダルな観測データセットを用いることが挙げられる。ここでのマルチモーダルとは、音声・画像・その他といったようなデータの多様性のことを指す。そこで我々は、マルチモーダル・データセットの深層学習への適用を目的とした研究を行う。本原稿では、代表的な深層学習を使ったニューラルネットワークのマルチモーダル・データセットへの適用を通し、その探索挙動の検証を行った。

1.2 研究内容

今回我々が使用したのは、R. B. Palm[1] が開発した DeepLearnToolbox である。この中に Deep Learning のライブラリが複数用意されており、その中から使用したのが Deep Belief Nets(以下 DBN)[2] である。本研究では、DBN をベースとして、マルチモーダルな情報を扱うニューラルネットワークを構築していくことを考える。

Omid らの研究で、マルチモーダルな情報として Youtube 動画を使った分類問題がある [3]。動画には、画像情報と音声情報が含まれているので、それらをマルチモーダルな情報とする。Omid らは Passive-Aggressive オンラインアルゴリズム (以下 PA) で 1 対他分類器を作り、動画を 31 クラスに分類している。PA は線形分類器であるため、マルチクラスに対しては、クラス数またはクラス数-1(30 クラスなら、30 個または 29 個の線形分類器を用意) の線形分類器を用意する必要がある。また、学習の際には、画像、音声別々で分類器を作り、それらの出力を最後に統合して最終的な分類予測を行う分類器を構築しているため、チューニングの際の手間が複雑になる。このことから、Omid らの分類器構築は、チューニングのためのコストが掛かるため、それだけで手間となる。本実験では、分類器構築への手間を考えて、ニューラルネットワークの使用と、あらかじめデータセットを統合することによって、分類器を 1 つに抑えることとする。ニューラルネットワークでマルチモーダルな情報を扱うことができるようになれば、今後の機械学習領域において、チューニングなどに掛かっていた時間を別の時間に有効活用することもできるようになる。

奈倉ら [4] の研究では、ロボットハンド開発に DBN を使っており、手と把持する物体の画像をマルチモーダルな情報として、ロボットハンドに適切な関節角度を学習させている。奈倉らの研究では、図 1 のようなニューラルネットワークを想定しており、マルチモーダルな入力画像をそれぞれ別々の入力層に配置し、特徴抽出の段階でデータを統合して学習を行う。これに対して今回の我々の実験では、入力層を 1 つにして、マルチモーダル情報をあらかじめ統合し学習する。

深層学習では、多数のパラメータ (層の数、ユニット数、バッチサイズ、学習回数、etc) があり、問題毎に調整する必要がある。今回の実験では、動画分類問題を扱い、Omid らの研究で使われている動画情報を用いる。学習の始めに統合された情報を用いた場合に、パラメータを調整するこ

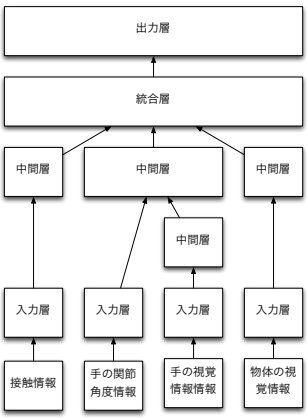

図 1: 奈倉らの提案するニューラルネットワーク

とで学習結果にどのような変化が現れるかを見る。具体的には、中間層の数を 1〜10 までの DBN を構築し、その時のエラー率、適合率、再現率を観察する。

1.3 Deep Belief Nets

DBN とは、多数の層からなるニューラルネットワークであり、その層は Restricted Boltzmann Machine(以下 RBM)[5] による学習を階層的に積み重ねていくことで構築される。図 2 (b) に DBN の例を示す。図 2 では中間層が 3 つだが、深層学習ではこの中間層を多層 (5〜10 層) にしたニューラルネットワークを構成する。RBM とは、ユニット間に対称的な結合荷重を持つ相互結合型のニューラルネットワークである。入力層の各ユニット v_i は、出力層のユニット h_j それぞれに対して対称的な結合荷重 w_{ij} を持つ。

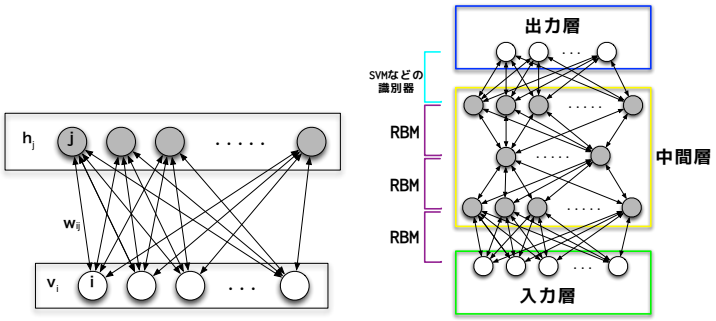

図 2: (a)Restricted Boltzmann Machine (b)Deep Belief Nets

入力層のユニット、出力層のユニットはそれぞれバイアス b_j, c_i を持ち、それぞれ以下の (1),(2) に従って確率的に出力が 1 となる。RBM の学習では、入力層のデータの値と、1 度出力層を通したデータの値が近くなるように結合荷重を調整していく。この RBM によって行われる学習フェーズを、事前学習と呼ぶ。この事前学習は教師なし学習であり、事前学習の後に教師あり学習のバックプロパゲーションを行う。

$$p(h_j = 1) = \frac{1}{1 + \exp(-b_j - \sum_i v_i w_{ij})} \qquad (1)$$

Basic research on performance evaluation of deep learning for multimodal data set

† Kakeru TAMASHIRO・Information Engineering, Graduate School of Science and Engineering, University of the Ryukyus

‡ Naruaki TOMA・Yuhei AKAMINE・Koji YAMADA・Satoshi ENDO・Department of Information Engineering, Faculty of Engineering, University of the Ryukyus

$$p(v_i = 1) = \frac{1}{1 + \exp(-c_i - \sum_j h_j w_{ij})} \quad (2)$$

2 実験

2.1 実験目的

今回の実験では、層を多くすることで学習の精度が上がる深層学習の特徴に着目する。動画というマルチモーダルな情報を扱った場合に、DBN の層の数を増やすことで、学習結果の精度が向上するどうかを見るために、1〜10 の中間層を持つ DBN を用意し、それぞれのエラー率、適合率、再現率を比較する。

2.2 実験設計

今回の実験では、Omid らが提供している YouTube Multiview Video Games Dataset を使用する。

表 1: Youtube Multiview Video Games Dataset

データセット	データ数
学習用	72400 件
テスト用	8800 件
特徴ベクトル	**次元数**
音声	64
画像	64
統合	128
クラスラベル	**ラベル数**
ゲームタイトル	30 件

DBN の入力層のユニット数は、入力画像の次元数が 128 次元なので 128 とし、出力層のユニット数は、ゲームタイトル 30 件を分類するので 30 とする。中間層のユニット数は、すべての層で 100 とする。事前学習、バックプロパゲーションの学習回数は、30、300 とする。活性化関数はシグモイド関数、学習率は 0.2、バッチサイズは 100 とする。

2.3 評価方法

テストデータを使い、1〜10 の中間層を持つ DBN で予測した結果で、間違った件数をテストデータ数で割った値をエラー率とする。それぞれのラベルにおいて正事例、負事例を定義し (ラベル 1 を正事例と置くと、それ以外のラベル 2〜30 を負事例とする)、適合率、再現率を求める。

2.4 実験結果

図 3 は横軸に中間層の数を取り、1〜10 まで増やした時のエラー率の推移を示す。1〜3 層まででエラー率が上昇し、それ以降からはほとんど変化が見られない。表 2,3 は、クラスラベル 1〜10 までの適合率と再現率を示す (ラベルは 1〜30 まであるが、ページ数の関係上 1〜10 までの結果を抜粋)。() 内の数字は中間層の数を示す。表 2,3 においては、エラー率の上昇と共に、適合率と再現率がほとんどのラベルにおいて低下している。中間層 3 の時点では、ラベル 4 のみを出力してしまい、ラベル 4 の適合率が 0.033、再現率が 1 となり、その他のラベルでは適合率、再現率共に 0 となる結果となった。本稿では割愛しているが、中間層 4 以降でも、中間層 3 の場合と同じ現象が起きていた (中間層 4 ではラベル 9、中間層 5 ではラベル 29、中間層 6 ではラベル 8 のみを出力するといった偏ったモデルを構築してしまっているため、再現率が 0 か 1 になっている)。

表 2: クラスラベル 1〜10 の適合率 (中間層の数)

クラスラベル	1	2	3	4	5	6	7	8	9	10
適合率 (1)	0.247	0.365	0.292	0.169	0.333	0.209	0.304	0.298	0.205	0.184
適合率 (2)	0.000	0.128	0.169	0.142	0.113	0.133	0.194	0.242	0.090	0.000
適合率 (3)	0.000	0.000	0.000	0.033	0.000	0.000	0.000	0.000	0.000	0.000
適合率 (4)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.042	0.000

表 3: クラスラベル 1〜10 の再現率 (中間層の数)

クラスラベル	1	2	3	4	5	6	7	8	9	10
再現率 (1)	0.080	0.125	0.287	0.136	0.415	0.525	0.779	0.425	0.212	0.068
再現率 (2)	0.000	0.020	0.362	0.089	0.275	0.179	0.162	0.391	0.273	0.000
再現率 (3)	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
再現率 (4)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000

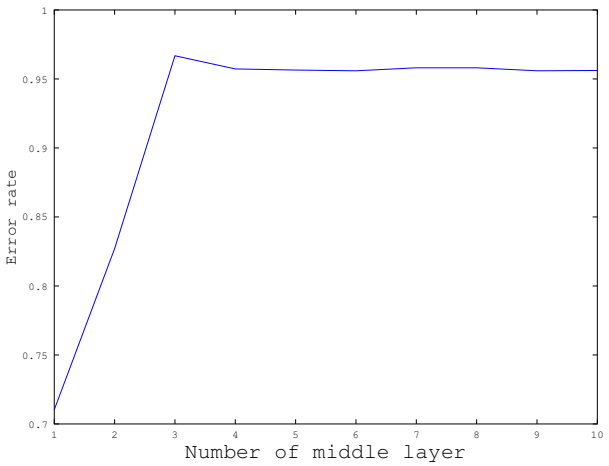

図 3: エラー率

3 現状と今後の課題

本実験では、層を多層にすると学習結果が良くなるという深層学習の特徴は見られず、それとは逆に、精度が悪くなるという結果となった。これは、今回の実験設定において学習した DBN が 1 つのラベルしか選択しないように構築されてしまったことに原因がある。この現象には、学習における内部パラメータに何かしらの因果関係があるのではないかと考える。この実験結果を踏まえて、今後は、今回の実験結果と内部パラメータとの因果関係があるかの実験を行っていきたいと考える。

参考文献

[1] Rasmus Berg Palm, “Prediction as a candidate for learning deep hierarchical models of data”, Technical University of Denmark, DTU Informatics, Master’s thesis, 2012

[2] Geoffrey E. Hinton, Simon Osindero, and Yee W. Teh, “A Fast Learning Algorithm for Deep Belief Nets”, Neural Computation, 18(7):1527—1554 2006.

[3] Omid Madani, Manfred Georg, David A. Ross, “On Using Nearly-Independent Feature Families for High Precision and Confidence”, Machine Learning 92 457-477, 30 May 2013

[4] 奈倉敬典, ジェイコブパイク, 荻野正樹, 浅田稔, “Deep Belief Nets を使った手の姿勢のマルチモーダル表現の獲得”, 日本機械学会ロボティクス・メカトロニクス講演会講演論文集, ROMBUN NO.2 A1-F22, 2009

[5] Geoffrey Hinton, “A Practical Guide to Training Restricted Boltzmann Machine”, UTML TR 2010-003, August 2, 2010