

# 畳み込みニューラルネットワークからのルール抽出 — 教示法によるルール抽出 —

佐藤 優也<sup>†</sup> 月本 洋<sup>‡</sup>

<sup>†</sup> 東京電機大学工学部情報通信工学科 〒120-8551 東京都足立区千住旭町 5 番

E-mail: <sup>†</sup> 19kmc10@ms.dendai.ac.jp, <sup>‡</sup> tsukimoto@mail.dendai.ac.jp

**あらまし** 以前畳み込みニューラルネットワークから論理式を抽出をする手法(近似法:分解法の一つ)を報告したが、そのネットワークの全結合層に中間層が存在する場合に精度が極端に下がることが判明した。本稿では、全結合層を1つのブラックボックスとして論理式を抽出すると良い結果が得られたので、その方法を提示する。また、論理式の入力として畳み込み・プーリング層の出力を最大化する画像を生成した後に、生成した出力最大化画像と全結合層の論理式を統合して、全結合層の論理式が示す特徴を可視化する方法も提示する。本稿では、これらの手法を MNIST データセットを学習させた畳み込みニューラルネットワークに対して行った結果を報告する。

**キーワード** 畳み込みニューラルネットワーク, 特徴抽出, ルール抽出, SmoothGrad, MNIST, 教示法

## Rule Extraction from convolutional neural networks — Rule Extraction with A Pedagogical Method —

Yuya Sato<sup>†</sup> and Hiroshi TSUKIMOTO<sup>‡</sup>

<sup>†</sup> School of Engineering, Tokyo Denki University 5 Senju-Asahi-cho, Adachi-ku, Tokyo, 120-8551 Japan

E-mail: <sup>†</sup> 19kmc10@ms.dendai.ac.jp, <sup>‡</sup> tsukimoto@mail.dendai.ac.jp

**Abstract** We presented a decompositional method of rule extraction from convolutional neural networks in the past. However, we found that the accuracy was extremely low in the case of fully connected layers with hidden layers. This paper presents a pedagogical extracting method. Also, this paper presents a method merging rules and images maximizing the output of convolutional/pooling layers, and visualizing features of the rules extracted from fully connected layers. These methods were applied to convolutional neural networks trained with MNIST dataset.

**Keywords** convolutional neural networks, feature extraction, rule extraction, SmoothGrad, MNIST, pedagogical method

### 1. はじめに

以前の報告[1]では、畳み込みニューラルネットワーク(以下、CNN)の全結合層は筆者の一人が開発した近似法(分解法の一つ)[2][3][4]によって論理式(ルール)を抽出した。その CNN の全結合層の構成は全て中間層なし、つまり出力層1層だけのものであった。その時の精度はおおよそ 0.9 であった。その後 CNN の構成を変えて実験した結果、全結合層に中間層が存在した場合に精度が著しく低下することが判明した。

論理式の抽出方法を CNN の全結合層をブラックボックスとして抽出する教示法に分類される手法に変えた結果、抽出精度が良くなったので本稿ではその手法について説明する。

論理式の入力変数は以前の報告と同じく最後のプーリング層出力なので、それを求める手法も同じく SmoothGrad[5]を用いる。

そして、SmoothGrad の結果(以下、最大化画像)と全結合層で抽出した論理式を統合して出力に対する重要

度を画像化する手法も説明する。

本稿では、上記の手法を MNIST データセットを学習させた CNN に対して行った。全結合層については2層(中間層1つと出力層)と3層(中間層2つと出力層)の場合を報告する。

2 節で近似法による論理式の抽出で中間層が存在するときに精度が低下する原因について述べ、3 節でその解決策である新手法について述べる。4 節では SmoothGrad について説明する。5 節では抽出した論理式と最大化画像の統合手法について述べる。6 節では実験で使用した CNN の構成・学習条件等について述べ、7 節でそのそれぞれについて SmoothGrad, 論理式抽出、及びその2つの統合結果を報告する。

### 2. 近似法による論理式の抽出

中間層が1層存在するネットワークで論理式を抽出すると、中間層の1ユニットあたりの精度は 0.6~0.9 ほどで、出力層の精度は 0.906、テスト用のデータを用

いて求めた論理式の全体としての精度は 0.088 であった。この論理式抽出の中間層と出力層の結果からこの論理式の正答確率を求めると 0.035 となる。このように論理式抽出の精度から求めたルールの正答確率とテスト用データでの論理式の精度に近い値となっている。

このことから、中間層のユニット数が多いほど、そして中間層自体が多いほど指数関数的に精度が低下する、つまり中間層が存在する全結合層から論理式の抽出をして全体としての精度を向上させるには、中間層の論理式の抽出に非常に高い精度が要求されるということを示している。

### 3. 新手法による論理式抽出

近似法を用いた CNN の全結合層に対する論理式の抽出は中間層の精度がボトルネックとなっていた。これにより近似法でのこれ以上の精度向上は期待できないため、近似法のようにユニット 1 つ単位で論理式の抽出をするのではなく、全結合層全体を 1 つのブラックボックスとして、全結合層の入力と出力の関係から論理式を求めることとした。

新手法は近似法と同様に最後のプーリング層出力を入力とし、全結合層の出力層の出力を論理式で表す。

手法の概要としては、学習用データを対象となる CNN に入力し、最後のプーリング層出力(全結合層への入力)と出力層の出力をサンプリングする。そのサンプリングされた入出力データを擬似的な真理値表として解釈することで論理式の抽出を行うというものである。

この新手法の手順は以下の通りである。なお、以下の手順は本稿で用いる CNN の全結合層に対しての手法であるが、入力と出力を対応付けることができるならば他のものにも応用できる。ただし、全結合層への入力は[0,1]に正規化後に閾値 0.5 で離散化されているものとする。

- ① データを CNN に入力し、全結合層への入力と出力を得る。
- ② 出力が”1”である出力ユニットに対して真理値表を用いた論理式の抽出を行う。
- ③ ①と②を繰り返す。

このとき出力ユニットごとに論理式の出現回数を数える。

上記の手順によって出力ユニットごとに出現回数と紐付けされた論理式が得られる。

この論理式を用いて推論する場合は以下の手順で行う。

- ① CNN にデータを入力し、全結合層入力を得る。
- ② 全結合層入力を”0”と”1”に離散化し論理式の出力を計算する。

- ③ 出力が”1”であるユニットが 1 つであったときはそのまま出力する。

2 つ以上あったときは、その論理式に対応する出現回数が最も多いユニットのみ”1”、それ以外を”0”として出力する。

以下、本稿において CNN の全結合層に対して論理式の抽出を行う際に用いるデータは MNIST の学習用画像データ 60000 枚、抽出した論理式の精度の検証に用いるデータは MNIST のテスト用画像データ 10000 枚とする。

### 4. SmoothGrad

CNN の畳み込み・プーリング層については前回の報告と同様に SmoothGrad によって最後のプーリング層出力が最大となる入力画像(最大化画像)を生成する。この最大化画像を求める操作は画像処理での特徴抽出に相当する。

SmoothGrad の手順を以下に示す。

- ① 入力画像に正規分布に則った乱数を加える。
- ② ①で生成した画像から出力各画素の偏微分を算出する。
- ③ ①と②を任意データ数繰り返し、平均を求めてそれを画像化する。

上記のように最大化画像を得る。詳細は参考文献[2]を参照。

本稿では、SmoothGrad で画像を求める際は以下のパラメータで行う。

- ・正規分布：平均 0.0, 標準偏差 0.1
- ・数値偏微分：差分 0.004 で前方差分
- ・使用するデータ：MNIST 学習用画像 10000 枚

### 5. 抽出した論理式と最大化画像の統合

以前報告した実験結果では、SmoothGrad による最大化画像と抽出した論理式から人間の目によって特徴を考えた。しかし、人間が見て人間が考える以上、最後のプーリング層の出力数が多くなると論理式と最大化画像から特徴を判断することが困難になっていく。

そこで、論理演算を以下のように定義して最大化画像と論理式を統合する。

$$A \wedge B = \min(A, B)$$

$$A \vee B = \max(A, B)$$

上記の論理演算に従い、最大化画像の画素ごとに論理演算を行う。

例えば、あるネットワークから「 $x_0 \wedge x_1 \wedge x_2 \wedge x_3 \wedge \bar{x}_4$ 」と「 $x_0 \wedge \bar{x}_1 \wedge x_2 \wedge \bar{x}_3 \wedge \bar{x}_4$ 」( $x_i$ :全結合層の  $i$  番目の入力に対応する最大化画像)という 2 つの論理式が抽出できたとする。このときの論理式における最大化画像は

$$R_1 = \min(x_0, x_1, x_2, x_3, \bar{x}_4)$$

$$R_2 = \min(x_0, \bar{x}_1, x_2, \bar{x}_3, \bar{x}_4)$$

となり，このユニットにおける最大化画像は

$$R = \max(R_1, R_2)$$

と求められる．

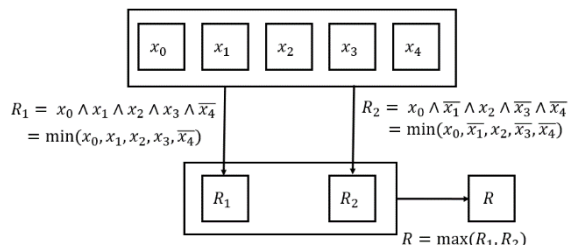


図 1 論理式と最大化画像の統合のイメージ

3 節から 5 節での内容と、前回の手法との比較をまとめると図 2 のようになる．

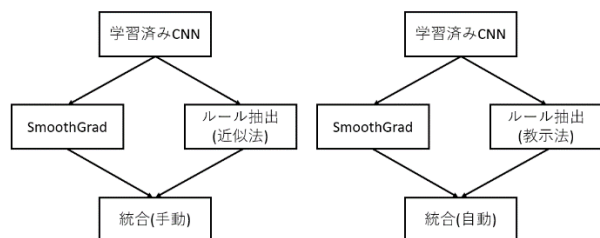


図 2 前回の手法(左)と新手法(右)

## 6. 実験

### 6.1. 全結合層 2 層の場合

まず，全結合層が 2 層で実験を行った．ネットワークの構成と学習条件は以下の通りである．なお，パディングはすべて 0、ストライドは畳み込み層が 1 でプーリング層が 2 である．

[ネットワーク構成]

- ・入力：MNIST(28×28 の 256 階調画像)
- ・畳み込み層 1  
フィルタ数：5  
フィルタサイズ：9×9
- ・プーリング層 1  
フィルタサイズ：2
- ・畳み込み層 2  
フィルタ数：10  
フィルタサイズ：9×9  
活性化関数：シグモイド関数
- ・プーリング層 2  
フィルタサイズ：2×2
- ・全結合層 1(中間層)  
出力数：10  
活性化関数：シグモイド関数
- ・全結合層 2(出力層)  
出力数：10

活性化関数：ソフトマックス関数

[学習条件]

- ・学習データ数：60000
- ・テストデータ数：10000
- ・ミニバッチサイズ：100
- ・エポック数：30
- ・テスト精度：0.972

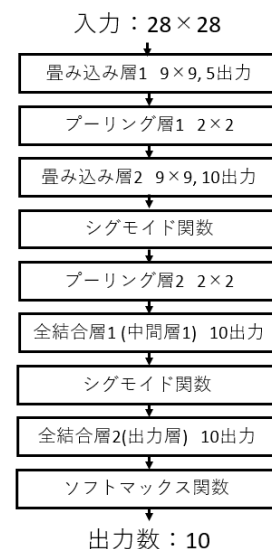


図 3 ネットワークの構成 1

学習した CNN に対して最大化画像と論理式の抽出を行う．全結合層から抽出した論理式については精度も算出する．

### 6.2. 全結合層 3 層の場合

次に，全結合層を 3 層にして同様の実験を行った．ネットワークの構成と学習条件は以下の通りである．ただし，全結合層 1 までのネットワーク構成とテスト精度以外の学習条件は全結合層 2 層の場合と同じであるため省略する．

[ネットワーク構成]

- ・全結合層 1(中間層 1)  
出力数：10  
活性化関数：シグモイド関数
- ・全結合層 2(中間層 2)  
出力数：10  
活性化関数：シグモイド関数
- ・全結合層 3(出力層)  
出力数：10  
活性化関数：ソフトマックス関数

[学習条件]

- ・テスト精度：0.970

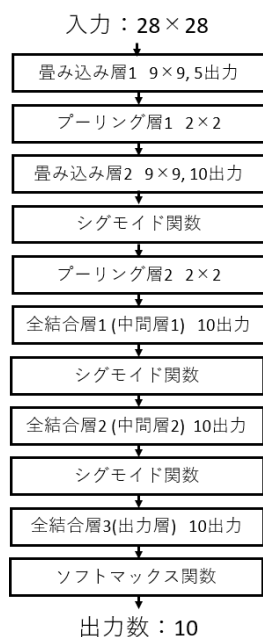


図 4 ネットワークの構成 2

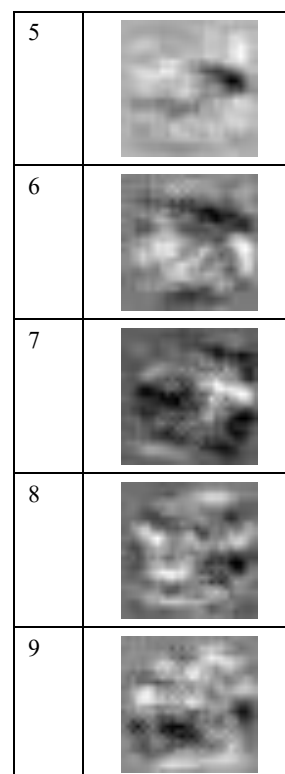


図 5 最大化画像(全結合層 2 層の場合)

## 7. 結果

### 7.1. 全結合層 2 層の場合

まず全結合層 2 層の場合について、実験の結果を図 5(最大化画像)と表 1(論理式抽出結果)、図 6(統合結果)に示す。以下、「No.」はプーリング層の出力番号を表し、論理式の抽出結果は各数字の中で最も出現回数の多かった論理式のみ示す。


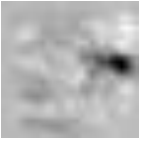





No.	最大化画像
0	
1	
2	
3	
4	

表 1 論理式抽出結果(全結合層 2 層の場合)

数字	最頻出論理式	回数
0	$\bar{x}_0 \wedge x_1 \wedge \bar{x}_2 \wedge x_3 \wedge x_4 \wedge x_5 \wedge \bar{x}_6 \wedge \bar{x}_7 \wedge x_8 \wedge \bar{x}_9$	2728
1	$\bar{x}_0 \wedge x_1 \wedge x_2 \wedge \bar{x}_3 \wedge x_4 \wedge \bar{x}_5 \wedge x_6 \wedge x_7 \wedge \bar{x}_8 \wedge \bar{x}_9$	1807
2	$\bar{x}_0 \wedge x_1 \wedge \bar{x}_2 \wedge x_3 \wedge x_4 \wedge \bar{x}_5 \wedge \bar{x}_6 \wedge \bar{x}_7 \wedge \bar{x}_8 \wedge \bar{x}_9$	3831
3	$\bar{x}_0 \wedge x_1 \wedge \bar{x}_2 \wedge x_3 \wedge \bar{x}_4 \wedge \bar{x}_5 \wedge \bar{x}_6 \wedge x_7 \wedge \bar{x}_8 \wedge x_9$	3954
4	$x_0 \wedge \bar{x}_1 \wedge x_2 \wedge \bar{x}_3 \wedge x_4 \wedge x_5 \wedge \bar{x}_6 \wedge x_7 \wedge \bar{x}_8 \wedge \bar{x}_9$	3990
5	$x_0 \wedge \bar{x}_1 \wedge \bar{x}_2 \wedge x_3 \wedge \bar{x}_4 \wedge x_5 \wedge x_6 \wedge x_7 \wedge x_8 \wedge x_9$	1518
6	$\bar{x}_0 \wedge \bar{x}_1 \wedge \bar{x}_2 \wedge \bar{x}_3 \wedge x_4 \wedge \bar{x}_5 \wedge \bar{x}_6 \wedge x_7 \wedge x_8 \wedge \bar{x}_9$	4244
7	$\bar{x}_0 \wedge x_1 \wedge x_2 \wedge x_3 \wedge \bar{x}_4 \wedge x_5 \wedge \bar{x}_6 \wedge x_7 \wedge \bar{x}_8 \wedge \bar{x}_9$	4179
8	$x_0 \wedge \bar{x}_1 \wedge \bar{x}_2 \wedge x_3 \wedge \bar{x}_4 \wedge \bar{x}_5 \wedge \bar{x}_6 \wedge x_7 \wedge x_8 \wedge \bar{x}_9$	2207
9	$x_0 \wedge \bar{x}_1 \wedge \bar{x}_2 \wedge x_3 \wedge \bar{x}_4 \wedge x_5 \wedge \bar{x}_6 \wedge x_7 \wedge \bar{x}_8 \wedge \bar{x}_9$	2626

数字	統合画像
0	
1	









2	
3	
4	
5	
6	
7	
8	
9	

図 6 統合結果(全結合層 2 層の場合)

論理式の抽出では各出力ユニットで出現回数が多い論理式と少ない論理式に分かれた．表 1 の最頻出論理式は特に多く，抽出に用いたデータのうち約半分が表 1 の論理式のどれかになっていた．

抽出した論理式の精度は 0.9580 であった．

統合結果の画像をそれぞれ見てみると，それぞれの数字でその数字の特徴に類似している．例えば，比較的理解しやすいであろう”7”についての画像を見てみよう．右上から下側中央へ伸びる線分と上側の領域が白く，左側中央と右下，右上の線分の先が黒くなっている．白くなっている領域は”7”を書くために必要な領域，黒くなっている領域は必要のない領域である．その他の画像を見ても同様のことが言える．

## 7.2. 全結合層 3 層の場合

次に全結合層 3 層の場合についても同様に実験結果を図 7(最大化画像)と表 2(論理式抽出結果)，図 8(統合結果)に示す．


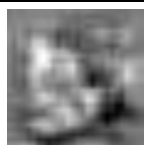
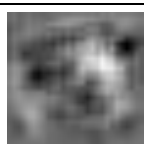















No.	最大化画像
0	
1	
2	
3	
4	
5	
6	
7	
8	
9	

図 7 最大化画像(全結合層 3 層の場合)

表 2 論理式抽出結果(全結合層 3 層の場合)

数字	最頻出論理式	回数
0	$\bar{x}_0 \wedge x_1 \wedge x_2 \wedge \bar{x}_3 \wedge x_4 \wedge x_5 \wedge \bar{x}_6 \wedge x_7 \wedge \bar{x}_8 \wedge \bar{x}_9$	3448
1	$x_0 \wedge x_1 \wedge x_2 \wedge \bar{x}_3 \wedge \bar{x}_4 \wedge \bar{x}_5 \wedge x_6 \wedge x_7 \wedge x_8 \wedge x_9$	2334
2	$\bar{x}_0 \wedge \bar{x}_1 \wedge x_2 \wedge \bar{x}_3 \wedge x_4 \wedge \bar{x}_5 \wedge \bar{x}_6 \wedge x_7 \wedge \bar{x}_8 \wedge x_9$	1863
3	$\bar{x}_0 \wedge \bar{x}_1 \wedge x_2 \wedge x_3 \wedge \bar{x}_4 \wedge \bar{x}_5 \wedge x_6 \wedge x_7 \wedge \bar{x}_8 \wedge \bar{x}_9$	4712
4	$x_0 \wedge \bar{x}_1 \wedge \bar{x}_2 \wedge x_3 \wedge \bar{x}_4 \wedge \bar{x}_5 \wedge \bar{x}_6 \wedge \bar{x}_7 \wedge \bar{x}_8 \wedge x_9$	1776
5	$\bar{x}_0 \wedge \bar{x}_1 \wedge \bar{x}_2 \wedge x_3 \wedge \bar{x}_4 \wedge x_5 \wedge \bar{x}_6 \wedge x_7 \wedge \bar{x}_8 \wedge \bar{x}_9$	2186
6	$x_0 \wedge x_1 \wedge \bar{x}_2 \wedge \bar{x}_3 \wedge x_4 \wedge x_5 \wedge \bar{x}_6 \wedge x_7 \wedge \bar{x}_8 \wedge x_9$	2155
7	$\bar{x}_0 \wedge \bar{x}_1 \wedge x_2 \wedge x_3 \wedge \bar{x}_4 \wedge \bar{x}_5 \wedge \bar{x}_6 \wedge \bar{x}_7 \wedge x_8 \wedge \bar{x}_9$	3329
8	$\bar{x}_0 \wedge x_1 \wedge x_2 \wedge \bar{x}_3 \wedge \bar{x}_4 \wedge \bar{x}_5 \wedge \bar{x}_6 \wedge x_7 \wedge \bar{x}_8 \wedge \bar{x}_9$	2149
9	$x_0 \wedge \bar{x}_1 \wedge x_2 \wedge x_3 \wedge \bar{x}_4 \wedge \bar{x}_5 \wedge \bar{x}_6 \wedge \bar{x}_7 \wedge \bar{x}_8 \wedge x_9$	4020

数字	統合画像
0	
1	
2	
3	
4	
5	
6	
7	

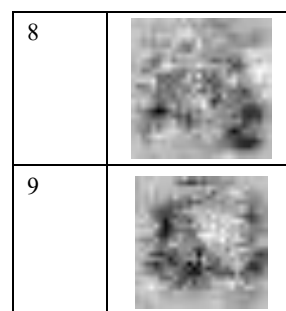


図 8 統合結果(全結合層 3 層の場合)

論理式の抽出において、全結合層 2 層の場合と同様に出力ユニット内で出現回数の偏りがあった。しかし、結果に現れない違いとして、2 層のときより各素子間での論理式の重複が多かった。

抽出した論理式の精度は 0.9588 であった。

SmoothGrad の結果の画像は類似していないにもかかわらず、全結合層 3 層の場合でも統合画像は対応する数字の特徴に類似していた。また、全結合層 2 層の場合での統合画像と比べると、白と黒の濃さやその領域に多少の違いが見られるものの、大まかな分布は一致していた。

## 8. 終わりに

本稿では、近似法を多層化した際の問題点について述べ、その解決策として教示法に分類される論理式の抽出手法を説明した。また、最大化画像と論理式を統合する手法についても提案した。全結合層の論理式抽出手法は層数に依存することなく高い精度で抽出を行うことができた。最大化画像と論理式の統合では、各数字について、特徴らしきものを見出すことができた。本稿では MNIST データセットのみでの検証であったが、今後それ以外のデータセットでも検証をしてゆきたい。

## 文 献

- [1] 月本洋, 佐藤優也, “畳み込みニューラルネットワークからのルール抽出”, 2018 信学技報, PRMU2018-79, pp.23-28, Dec.2018.
- [2] H.Tsukimoto, “Extracting Rules from Trained Neural Networks”, IEEE Transactions on Neural Networks, Vol.11, No.2, pp.377-389, 2000.
- [3] 月本洋, 下郡信宏, 高島文次郎, “多重線形関数を用いたニューラルネットワークの構造分析”, 電子情報通信学会論文誌, Vol.J79-D-II No.7, pp1271-1279, 1996.
- [4] 月本洋, 松本一教, 実戦データマイニング, オーム社, 2018.
- [5] D.Smilkov et al., “SmoothGrad: removing noise by adding noise” arXiv:1706.03825, 2017.