6N-01

深層学習における推論過程の可視化の検討

鈴木 航　孟 林　山崎 勝弘

立命館大学　理工学部

1. はじめに

深層学習は、一般物体認識、連続音声認識、自然言語処理などのさまざまな分野で従来法を大きく上回る性能を発揮しており、近年注目を集めている。しかし、深層学習では何を根拠に推論結果を出力しているのか分からないという問題点がある。本稿ではフィルタの貢献度と重要な認識領域を明らかにし、畳み込みニューラルネットワーク(CNN)の推論根拠を示すことを目指す。

2. MNIST による認識実験

2.1 使用した CNN

本実験で使用する CNN は、2 層の畳み込み、2 層のプーリング、及び 1 層の全結合層から構成される(図 1)。畳み込み層は特徴抽出を行うためのフィルタを 3 枚ずつ持ち、1 層目が 5×5、2 層目が 3×3 である。これはフィルタ枚数が 6, 16 である LeNet を基準として、認識率が 98%以上となる最小のフィルタ枚数を実験により求めた結果である。これにより CNN のパラメータ数は LeNet の 1/5 程度まで削減することができた。

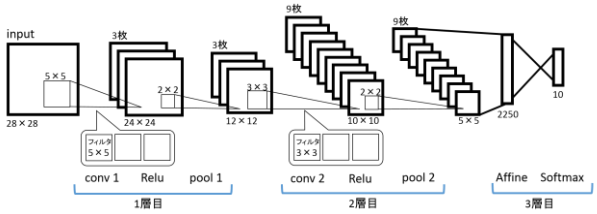

図 1　3 層の CNN

2.2 実験条件

手書き数字文字を対象として、MNIST 画像を用いた。訓練画像を 6 万枚、テスト画像を 1 万枚用意した。訓練はバッチサイズを 100 とし、10 万回繰り返した。初期値はランダム、最適化手法は Adam とした。

3. フィルタの貢献度

3.1 Affine 層の役割

まず本 CNN の結合層に当たる Affine 層の役割について説明する。図 2 に CNN によって画像が処理される過程を示す。入力画像は各層の処理を経て、最終的には(3×3)×10 クラスの計 90 枚のサイズの小さな特徴マップの集合になる。そして特徴マップを構成するデータを数値として取り出し、各クラス毎に全て足し合わせ、クラス毎に持つバイアス b を足す。この処理によって得られた 10 クラスの数値から、最大のものを推論結果として出力する。図 2 の場合、クラス「5」の総和が 10 クラス中最大なので、推論結果は「5」となり正解である。

Toward Visualization of Inference Process in Deep Learning, Kou Suzuki, Lin Meng and Katsuhiro Yamazaki, Graduate Scool of Scince and Engineering, Ritsumeikan University.

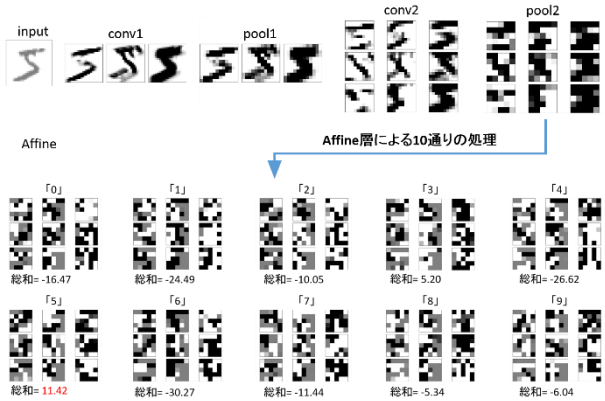

図 2　画像の処理過程(入力が「5」の場合)

3.2 認識貢献フィルタの特定

Affine 層の処理により得られる 9 枚の特徴マップの数値は、図 2 の場合は図 3 のようになる。これら 9 枚の特徴マップの数値により、分類に貢献したフィルタを特定することができる。この場合、9 枚の中央の値が最大であるので、conv1、conv2 共に 2 枚目のフィルタが最も貢献していることとなる。同様にして貢献フィルタを各クラス別に 1000 枚ずつ求めた。それを集計した結果、フィルタの貢献度に大きな偏りが見られた。各クラスの認識貢献フィルタを図 4 に示す。貢献度 1 位のフィルタを赤、2 位のフィルタを青枠で囲み、そのフィルタによって生成される特徴マップも同色の枠で囲っている。「3」「4」については 1 位の数値が突出して大きかったため、2 位の色枠を省略している。

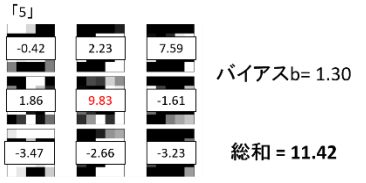

図 3　「5」処理後画像 9 枚の数値

本実験により、クラス毎に貢献度の高いフィルタを明らかにすることができた。貢献度 1 位のフィルタのみを見た場合、使われるフィルタには偏りがあるが、2 位を含めば各クラスによって使われるフィルタの組み合わせは全て違っており、分類クラスによってフィルタを使い分けていることが確認できる。

4. 重要な認識領域の特定

入力画像の一部分を隠した画像を用意し、その画像の出力を観察する。一部を消すことでその画像が誤った認識をした場合、その消した部分が画像の分類にとって重要な特徴であったことになる。逆に一部消しても出力が正しかった場合、その部分は分類にとって重要ではなかったこととなる。本実験では 28×28 の入力画像を 14×14 の 196 領域に分割して、図 5 に示す手順で実験を行った。

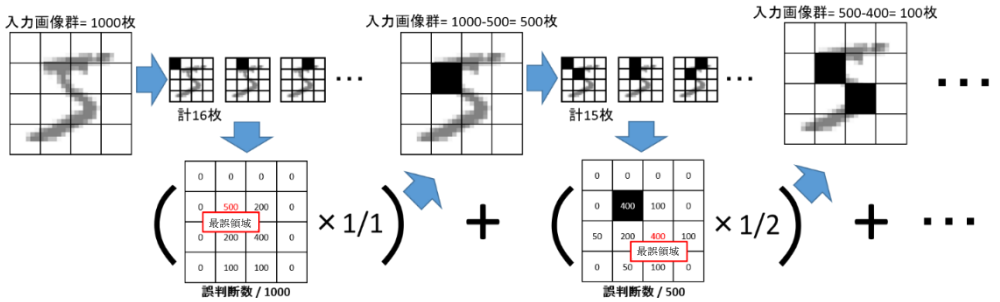

図 5　領域削除実験のイメージ(16 分割の場合)

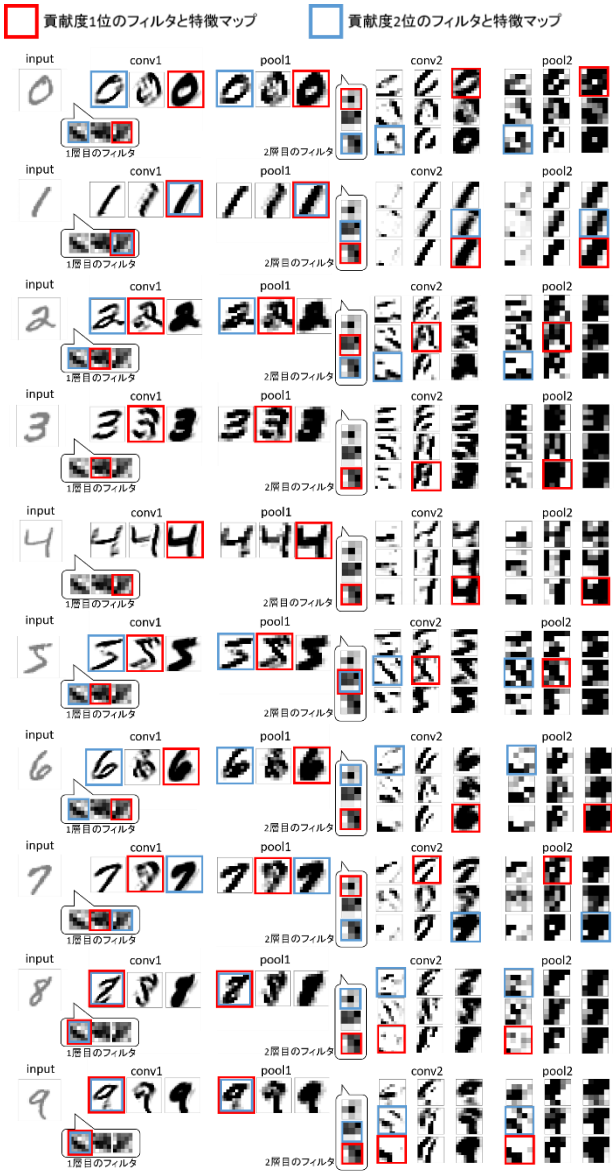

図 4　各クラスの認識貢献フィルタ

① クラス別に 1000 枚の入力画像群を用意する。
② 分割した領域を一つ削除し、CNN を通して出力する作業を全領域分繰り返す。
③ 各領域での誤判断した画像の枚数を保存し、最も多く誤判断した領域を「最誤領域」とする。
④ 最誤領域を削除したもので入力画像群を上書きし、さらに誤判断した画像を入力画像群から取り除く。
⑤ ②～④を全ての画像が誤判断を起こすまで繰り返す。
⑥ ③で記録したデータに重み(1/削除領域数)を掛け、全て足し合わせる。

上記の手順で実験を行った結果が図 6 である。数値が大きい重要な領域を濃く、小さい領域を薄く色付けている。本実験により、各クラス毎に注目している領域が異なることが確認された。

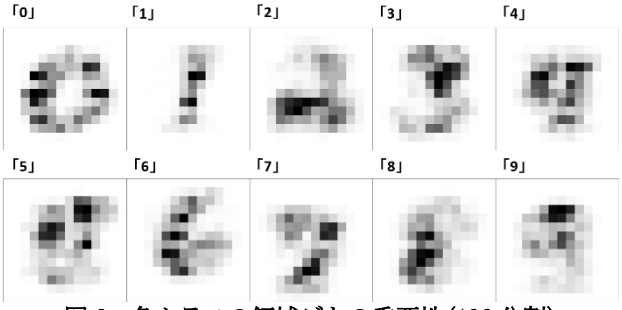

図 6　各クラスの領域ごとの重要性(196 分割)

5.　考察

本実験で CNN の推論において、どのフィルタを使って、どの部分に注目して判断を下しているのかを明らかにした。次の課題として、斜線の方向や線の交差点など、各フィルタが強調する特徴を明らかにすることが挙げられる。これは CNN に対して逆向きの計算を行い、入力画像を生成することによって実現できると考えている。

Affine 層での計算結果が最大となるよう値を設定し、CNN の 2 層の畳み込み層に対して逆畳み込み処理を行う。これによりフィルタごとの特徴が最も表現された入力画像を生成することができる。フィルタの役割が明らかになれば、ネットワークの推論根拠について、画像のどの領域のどの特徴を手がかりとして画像判別を行っているのかを明示できると考えられる。

6.　おわりに

本稿では手書き文字画像である MNIST を対象として、推論の根拠となる要素を、貢献フィルタと重要な認識領域の 2 つの観点から可視化した。さらにフィルタの役割を明らかにする手法を提案した。これらを組み合わせて、推論過程をより直感的に理解できる形へと可視化することが今後の課題である。

参考文献

[1]M. D. Zeiler and R. Fergus : Visualizing and Understanding Convolutional Networks, ECCV 2014, Part I, LNCS 8689, pp. 818-833, 2014.