畳み込みニューラルネットワークの特性

東北大学大学院医学系研究科 博士課程 田中 真一

2024.04.20
2024.04.23 修正

序

前回レポート (2024.04.17) では、多層パーセプトロンモデルによる手書き文字推定において、画像を平行移動させることで適切な推定を行えなくなることを確認した。この平行移動について一部の文献では、畳み込みにより図形的特徴が抽出されたり、プーリングによって平行移動に対するロバスト性が得られたりするかのような記述がみられるが (敢えて引用は示さない) これらは正しくない。

今回は、画像を処理する深層学習モデルで頻用される畳み込みニューラルネットワーク (Convolutional Neural Netowrk; CNN) の代表として 1998 年に提案された LeNet-5[1)] を用い、MNIST[2)] の手書き数字の認識における平行移動に対するロバスト性を検討した。損失関数には交差エントロピー誤差関数を、最適化には確率的勾配降下法を用いた。計算には Python 3.11.2[3)] + PyTorch 1.13[4)] を使用した。

モデルの概略と評価方法

オリジナルの LeNet-5 に加えて、比較のため、以下の 3 個のモデルを用いた計算も行った。Model 1 は多層パーセプトロンモデルであり、畳み込みもプーリングも行わないものである。Model 2 は LeNet-5 と類似の畳み込みニューラルネットワークだが、第 3 層の全ての feature map は 第 2 層で生成された全ての feature map と結びつけられている。すなわち LeNet-5 が採用している非対称性を除去したネットワークである。Model 3 は 4 x 4 のプーリングを行った後に多層パーセプトロンモデルを適用するものである。ここでは畳み込みは行わない。

各モデルについて、まず MNIST の train データセット (60,000 例) を用いて学習を行う。次に MNIST の test データセット (10,000 例) を用いて、数字の認識を適切に行えた例と行えなかった例との数を調べる。さらに、test データセットの各画像について左右方向に -5 ピクセルら +5 ピクセルまで、上下方向に -2 ピクセルから +2 ピクセルまで、計 55 通りに平行移動した画像を用いて計 550,000 通りのパターンについて数字の認識を試みる。

計算時間の都合上、学習率やエポック数の最適化は行わない。

MNIST を用いた評価結果

MNIST test データセットを用いた手書き文字の認識推定結果を表 1 に示す。LeNet-5 では平行移動なしの場合に正答率 97% であったが、平行移動を加えると正答率は 60% に低下した。多層パーセプトロン、対称化 LeNet-5, プーリングあり多層パーセプトロンでは、平行移動なしでは正答率 93%, 96%, 87% であった

が、平行移動を加えることでそれぞれ 41%, 49%, 41% に低下した。

表 1　MNIST を用いた手書き数字画像の認識結果

モデル		平行移動なし			平行移動あり		
		正答	誤答	正答率	正答	誤答	正答率
Model 0	LeNet-5	9696	304	97%	331140	218860	60%
Model 1	多層パーセプトロン	9278	722	93%	224942	325058	41%
Model 2	対称化した LeNet-5	9618	382	96%	268626	281374	49%
Model 3	プーリング多層パーセプトロン	8733	1217	87%	223386	326614	41%

考察

LeNet-5 (Model 0) では多層パーセプトロン (Model 1) に比して、平行移動なしの場合において高い正答率が得られた。これは、LeNet-5 では畳み込み層が含まれているために、局所の相関が重視されているからである。LeNet-5 の第 1 層は 5×5 カーネル 6 チャンネルを用いる畳み込み層であり、$28 \times 28 \times 6 = 4704$ のノードから成っている。原理的には、適切な重み係数の設定が成されれば、同じ 4704 のノードから成る全結合層を用いても、すなわち多層パーセプトロンモデルを用いても、この Lenet-5 第 1 層と等価な層を形成できる。両者の違いは、画像上の「近接したピクセル」と「遠く離れたピクセル」の間に取り扱い上の差異を設けるかどうか、という点のみにある。畳み込み層の場合、カーネルサイズよりも遠方にあるピクセルに相当するノードに対しては重み係数を 0 を設定しているに等しいが、多層パーセプトロンモデルでは互いに離れたノード同士に対しても大きな重み係数が設定される可能性がある。現実の画像、特にアンチエイリアスが加わっている MNIST のような画像では、近接するピクセル同士では色調に強い相関がある一方、遠方のピクセル同士はほぼ独立である。たとえば、あるピクセルが白である場合、その隣のピクセルは白か灰色であって、真っ黒ということはまずない。しかし 10 ピクセル離れた位置については、それが白であっても黒であってもおかしくないから、ほぼ独立といえよう。このような「『近接したピクセル間には強い相関があるが、離れたピクセル間には相関がない』という関係を仮定して重み係数を決定する」というのが畳み込み層の本質である。歴史的経緯から、畳み込み層で得られた値を「特徴量」などと表現することもあるが、これは単に近接するノード同士の重み係数を計算しただけのものであって、何か意味のある特徴を「抽出」しているわけではない。画像に描かれた図形の局所的な幾何学的構造を反映しているわけでもないにもかかわらず、これを意味のある量として強引に解釈しようとする文献が少なくないので、注意を要する。

なお LeNet-5 では、第 1 層で畳み込んだ後に 2×2 の平均プーリング層を加えている。これは畳み込みによって生じた冗長なノードを削減する効果がある。元々の画像は $28 \times 28 = 784$ のピクセルであったが、第 1 層で畳み込んだ後は 4704 にノード数が増加している。しかし畳み込みによって何か新しい情報が付加されたわけではないから、同じ情報が平均 $\frac{4704}{784} = 6$ 回、重複して保持されていることになる。これは無駄であるので、ノード数を 1/4 に減らすことで、重大な情報損失を伴わずに計算効率を高めることができる。理屈としては 1/6 程度にまでノード数を減らしても問題ないはずだが、そのようなプーリングは難しいので LeNet-5 では 1/4 に留めたものと思われる。プーリング層には、このように冗長なノードを削減する効果があるが、それ以外にロバスト性を生むなどの「特殊な」効果は存在しないと考えられる。

LeNet-5 では平行移動ありの場合において、多層パーセプトロンに比して比較的高い正答率が得られた。これは畳み込み層が存在することにより、多少の平行移動に対してはロバスト性が生じるためである。ただし、

これは train データセットの全ての画像データにおいて、数字が画像中央に「正しく」描かれていることが前提である。たとえば数字が左に 3 ピクセル平行移動された場合について考える。教師データにおいては、いずれの数字においても画像の左右両端は空白 (黒) となっているので、左に平行移動された、つまり白い部分が左に偏った test 画像は、いずれの数字にも似ていない、ということになる。しかし畳み込み層の場合、多少の平行移動であれば、畳み込まれた値は平行移動前と比べて全く異なる値になるわけではなく、多少の面影は残した値になる。それゆえに、多層パーセプトロンに比べて、カーネルサイズよりも充分に小さい程度の平行移動に対してはある程度のロバスト性が生じるのである。ただし、これは一部の人が主張するような「特徴量の抽出により、位置が多少変わっても同じような形の部分を選び出して認識できる」というようなものではない。図形の認識や、位置のずれを検出する、というような機構は、畳み込み層にもプーリング層にも含まれていない。

「畳み込み層による平行移動に対するロバスト性は、図形の特徴を認識しているが故ではない」ということは、教師データに多少の平行移動を加える実験により確認できる。具体例として、教師データのうち「3」が描かれているものは右に 3 ピクセル、「2」が描かれているものは左に 3 ピクセルだけ平行移動してから LeNet-5 で学習を行った。このモデルを使い、test データセットからランダムに抽出した画像について上下方向に -2 から +2 ピクセル、左右方向に -5 から +5 ピクセルだけ平行移動してから数字判定を行った結果を示したのが図 1 から図 4 である。図 1 および図 2 は、それぞれ正解が 4, 5 の場合である。平行移動をしない場合 (各図の中央にある画像) はいずれも正しく数字を判定できている。一方、LeNet-5 で採用されている 5×5 のカーネルに比べて小さい量の平行移動であっても、右に平行移動すれば 3 と、左に平行移動すれば 2 と、それぞれ誤判定されている。これは、教師データに平行移動を加えたことにより「右の方に書かれた数字は 3 であり、左の方に書かれた数字は 2 である」という学習が為された結果である。前述のように、畳み込みには図形としての形態を学び取る効果はないので、このような場合には平行移動に対するロバスト性は示されない。また図 3 および図 4 は、それぞれ正解が 2, 3 の場合である。これらの場合も、教師データに加えられた平行移動のため、test データに平行移動を加えなかった場合 (各図の中央にある画像) は正しく判定することができない。

表 1 において、LeNet-5 (Model 0) と対称化した LeNet-5 (Model 2) の間では、平行移動なしの場合には著明な差異はみられなかった。これは、LeNet-5 において第 2 層と第 3 層の間に設けられている非対称性が、学習の効率化にのみ作用しており、最終的に形成されるモデルに本質的な影響を与えないことを示している。平行移動ありの場合には正答率に大きな差異が生じているが、これは対称化した LeNet-5 では学習 (エポック数) が不充分であったことによると考えられる。

プーリングを加えた多層パーセプトロン (Model 3) では、平行移動に対するロバスト性はみられなかった。プーリングは単に冗長なノードを削減するだけの処理であり、ロバスト性には関係しない。

まとめ

畳み込みニューラルネットワークで使われる畳み込み層やプーリング層の働きについて定性的に考察した上で、LeNet-5 による手書き数字の認識を行った。LeNet-5 では、描かれた数字の平行移動に対し軽度のロバスト性が認められたが、これは教師データが「整っている」ことを前提としており、図形の形態的特徴を認識しているわけではないことが示唆された。深層学習を活用する際には、そのモデルの機序や特徴を少なくとも定性的には理解した上で計算することが重要と考えられる。

図 1　平行移動した教師データを用いた手書き数字認識結果の例 (index 4017, 正解 4)

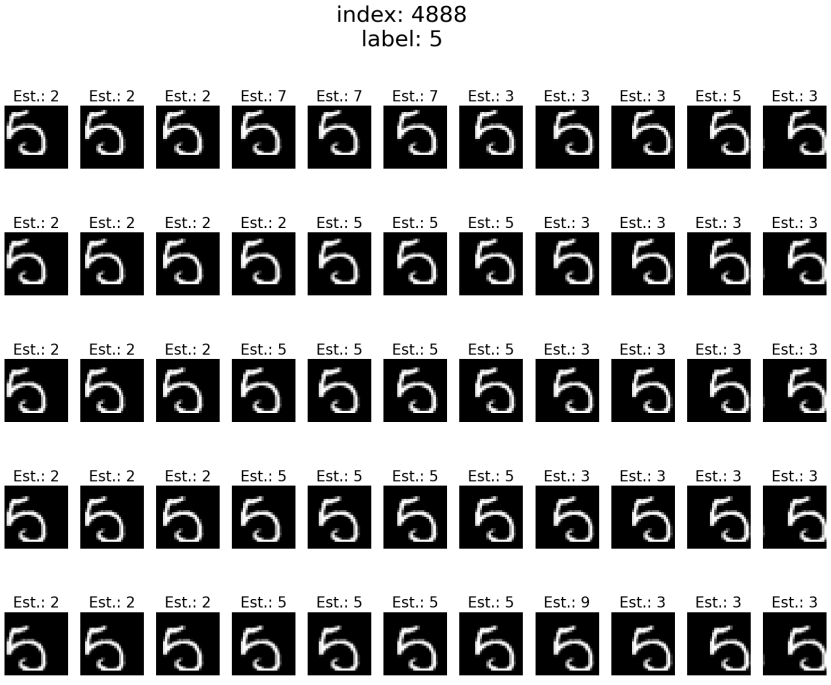

図 2　平行移動した教師データを用いた手書き数字認識結果の例 (index 4888 正解 5)

図 3　平行移動した教師データを用いた手書き数字認識結果の例 (index 5392, 正解 2)

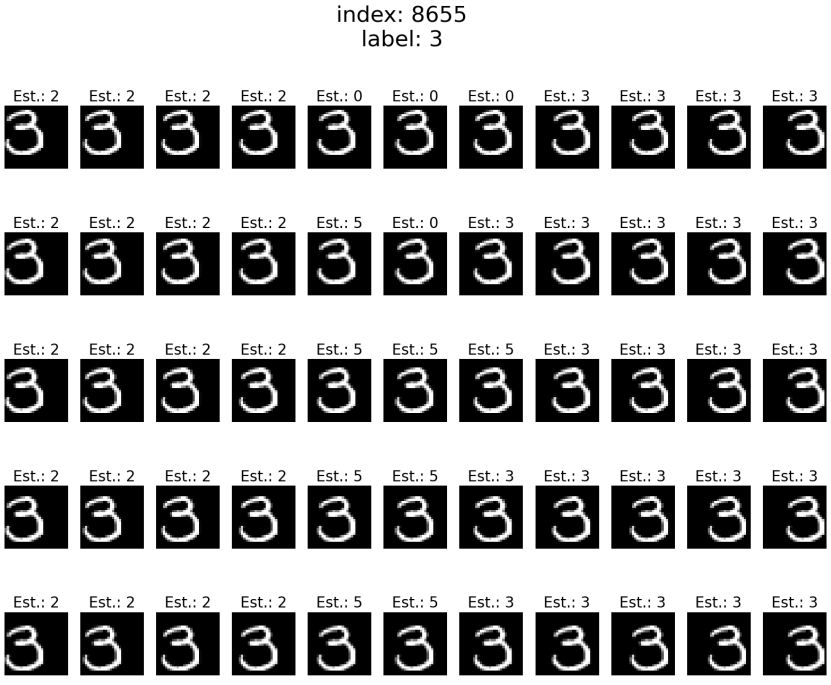

図 4　平行移動した教師データを用いた手書き数字認識結果の例 (index 8655, 正解 3)

参考文献

1) Lecun, Y., *et al.*, 'Gradient-Based Learning Applied to Document Recognition', Proc. IEEE **86**(11), 2278-2324, (1998).
2) The MNIST Database of Handwritten digits, http://yann.lecun.com/exdb/mnist/ (2024.04.17 閲覧)
3) Python.org, https://www.python.org/ (2024.04.17 閲覧)
4) PyTorch, https://pytorch.org/ (2024.04.17 閲覧)