ディープラーニングと画像認識
—基礎と最近の動向—

岡谷　貴之

ディープラーニングは近年，人工知能の色々な分野で大きな成功を収めつつあり，その高い性能は広く知られるようになった．本稿ではディープラーニングの画像への応用，中でも画像認識に焦点を絞り，そこで欠かせない存在となっている畳込みニューラルネットについて，その技術的な基礎から最近の研究の動向までを概観する．

キーワード：画像認識，畳込みニューラルネットワーク，単純型細胞・複雑型細胞，ILSVRC，ネオコグニトロン，LeNet

1. はじめに

現在，ディープラーニングが成功を収めている分野はいくつかあり，画像認識はその 1 つである．ただし画像認識には他と違う点がある．それは，畳込みニューラルネットワーク（convolutional neural network，以下畳込みネット）が，欠かせない存在だということである．畳込みネットは，（最も普通のニューラルネットである）順伝播型ニューラルネットワークの一種であるが，畳込み層とプーリング層と呼ばれる特別な構造を持つ点で独特である．

畳込みネットは，1980 年前後に Fukushima らが発表したネオコグニトロン [1] にルーツを持つ．これは，神経科学の知見 [2] に基づく実験的な画像認識システムであった．80 年代後半，LeCun らは，誤差逆伝播法 (back propagation) に基づく勾配降下法を，ネオコグニトロンの構造を持つネットワークの学習に適用し，これを LeNet と名付けた [3, 4]．彼らはこれが，現実的な文字認識のタスクで高い性能を達成することを示した．LeNet は，今使われている畳込みネットの基本的要素をすべて持っており，このとき，畳込みネットは技術的には完成していたと言える．

2. 畳込みニューラルネットワーク

2.1 単純型細胞と複雑型細胞

先述のとおり畳込みネットは，神経科学の知見に基づく構造を持つ．生物の視覚系では，外界から眼に取り込まれ網膜に結んだ像は，脳の視覚野に電気的な信号として伝達される．そこにある無数の神経細胞の中には，網膜の特定の場所に特定のパタンが入力されると興奮し，それ以外のときは興奮しないという，選択的な振る舞いを示すものがある．それらは，網膜（あるいは視野）の特定の位置に，特定の方向・太さの線分が提示されたときのみ選択的に反応する．

そのような細胞には単純型細胞 (simple cell)，複雑型細胞 (complex cell) と呼ばれる 2 種類があり，それぞれ異なる振る舞いを示す [2, 5, 6]．入力の位置選択性の違いが両者の差であって，前者はそれが厳密だが，後者は一定の寛容性を持つ．

単純型細胞は，図 1 のような構造の単層ネットワークの各ユニットでモデル化できる．左側の層が入力で，右側が出力である．各層のユニットは 2 次元的に並び，同図 (a)，(b) のように右の層のユニットは，左の層の 4×4 のユニット群とのみ結合を持ち，そこに (c) のような特定のパタンが入力されたときのみ，それに反応して活性化するとする．そのパタンは（右の層の）全ユニットで共通である．

複雑型細胞は，図 1 の単層ネットワークの上位に層を追加したとき，そのユニットによってモデル化できる（図 2）．追加した層のユニットは，中間層の 3×3 のユニット群と結合を持ち，これらのユニットのうち 1 つでも活性化すると，自身も活性化するとする．中間層のユニットが活性化するパタンが図 1(c) のとき，全体への入力が図 2(a) から (b) のように変わると，中間層で活性化するユニットは同図のように変化する．一方出力層のユニットは，中間層のユニットがどれか

おかたに　たかゆき
東北大大学院情報科学研究科
〒 980–8579 宮城県仙台市青葉区荒巻字青葉 6–6–01
okatani@vision.is.tohoku.ac.jp

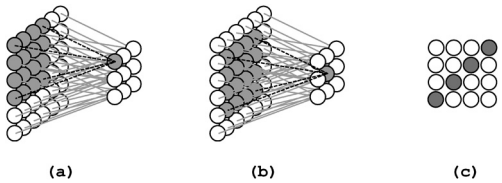

図 1 単純型細胞のモデル．左の層に画像が入力され，右の層から出力される．左の層の各ユニット（単純型細胞）は (a)，(b) のように入力層の限られたユニットとのみ結合を持つ．そして例えば (c) のようなパタンに選択的に反応し，活性化する（図 2 も参照）．

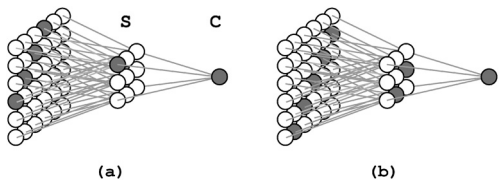

図 2 複雑型細胞のモデル．最も右の層のユニット（複雑型細胞）は，中間層の 3×3 のユニット群のうちどれか 1 つでも活性化していれば，活性化する（最大プーリング）．(a)，(b) のように，入力パタンがわずかにシフトすると中間層のユニットの反応は変わるが，最上位層のユニットは活性化したままである．

1 つでも活性化していれば活性化するため，いずれの入力でも活性化する．このように，中間層のユニット（単純型細胞）は入力パタンの位置変化に敏感だが，出力層のユニット（複雑型細胞）は一定の（この例では 3×3）範囲の位置ずれに鈍感である．

図 2 の中間層と出力層が，畳込みネットを構成する畳込み層およびプーリング層に，それぞれ対応する．以下では，この 2 つを順番に説明する．

2.2 畳込み層

$W \times W$ 画素からなるグレースケールの画像を考える．各画素をインデックス $(i,j)(i = 0,\ldots,W-1, j = 0,\ldots,W-1)$ で表し，画素 (i,j) の画素値を x_{ij} と書く．この画像に適用する $H \times H$ 画素のフィルタ（サイズの小さい画像）を考える．フィルタの画素はインデックス $(p,q)(p = 0,\ldots,H-1,\ q = 0,\ldots,H-1)$ で表し，画素値を h_{pq} と書く．

画像の畳込みとは，画像とフィルタ間で定義される次の積和計算である[1]．

$$a_{ij} = \sum_{p=0}^{H-1} \sum_{q=0}^{H-1} x_{i+p,j+q} h_{pq} \qquad (1)$$

[1] 正確には相関と呼ぶべきだが，フィルタを上下左右を反転すると同じなのでここでは畳込みと呼んでいる．

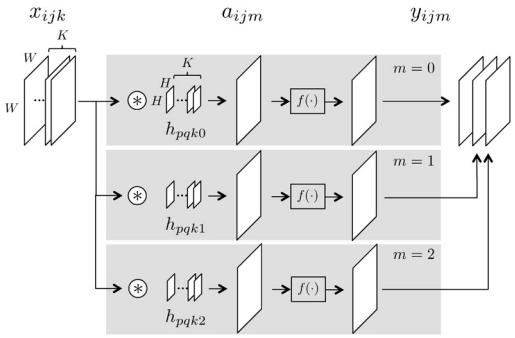

図 3 多チャネルに複数フィルタを適用する畳込み層の概要．K チャネルある入力画像に，3 種類のフィルタ（縦横 $H \times H$ 画素，サイズ $H \times H \times K$）を適用し，3 チャネルの画像（マップ）を出力する場合．

この計算は，フィルタの濃淡パタンと類似した濃淡パタンが入力画像上のどこにあるかを検出する働きがある．つまり，フィルタが表す特徴的な濃淡構造を，画像から抽出する「特徴抽出」の働きがある．

実用的な畳込みネットでは，グレースケールの画像 1 枚に対してではなく，多チャネルの画像に対し，複数個のフィルタを並行して畳込む演算を行う（図 3）．多チャネルの画像とは各画素が複数の値を持つ画像であり，チャネル数が K の画像の各画素は K 個の値を持つ．例えば，グレースケールの画像では $K = 1$，RGB の 3 色からなるカラー画像では $K = 3$ となる．畳込みネットの中間層では，さらにそれ以上のチャネル数（$K = 16$ や $K = 256$ など）の画像を扱う（マップと呼ぶこともある）．以下では，画像の縦横の画素数が $W \times W$ でチャネル数が K のとき，画像のサイズを $W \times W \times K$ と書く．

図 3 を用いて畳込み層での計算を説明する．この畳込み層は直前の層から K チャネルの画像 $x_{ijk}(k = 0,\ldots,K-1)$ を受け取り，これに $M = 3$ 種類のフィルタ $h_{pqkm}(m = 0,\ldots,M-1)$ を適用している．各フィルタ $(m = 0,1,2)$ は通常，入力と同じチャネル数 K を持ち（サイズを $H \times H \times K$ とする），図 3 のようにフィルタごとに計算は並行に実行される．計算の中身は，そのフィルタの各チャネルごとに，これも並行に画像とフィルタの畳込み（(1) 式）を行った後，結果を画素ごとに全チャネルにわたって加算する．

$$a_{ijm} = \sum_{k=0}^{K-1} \sum_{p=0}^{H-1} \sum_{q=0}^{H-1} x_{i+p,j+q,k} h_{pqkm} + b_{ijm} \qquad (2)$$

このように，入力画像のチャネル数によらず，1 つのフィルタからの出力は常に 1 チャネルになる．図では

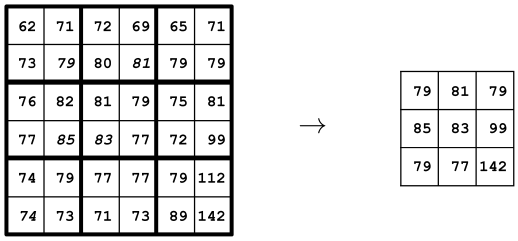

図 4　6×6 の入力画像に 2×2 を 1 つの値にする最大プーリングを，2 画素間隔で適用した例．出力は 3×3 となる．

省略しているが，これにバイアス b_{ijm} が加算される場合もあり，その場合，画像の位置によらず定数（フィルタごとに 1 つ，つまり $b_{ijm} = b_m$）とすることが多い．

こうして得た a_{ijm} に活性化関数を適用する．

$$y_{ijm} = f(a_{ijm}) \tag{3}$$

活性化関数には，正規化線形関数 (rectified linear) すなわち $f(x) = \max(x, 0)$ を使うことが多い．この y_{ijm} が，畳込み層の最終的な出力となりその後の層へと伝わる．これらはフィルタの数 M と同数のチャネル数を持つ多チャネルの画像と見なせる．つまり入力のサイズが $W \times W \times K$ のとき，（畳込み層のフィルタ数を M として）出力のサイズは $W \times W \times M$ になる．

2.3　プーリング層

プーリング層は通常，畳込み層の直後に設置される．プーリング層のユニットは，本章の最初に述べた複雑型細胞のモデルと考えることができ，畳込み層で抽出された特徴の位置感度を（わずかに）低下させる働きがある．

プーリング層での計算は次のとおりである．入力画像上で画素 (i, j) が左上隅に来る $H \times H$ 正方領域をとり（便宜上同じ H を使っているが，畳込み層のフィルタのサイズとは関係ない），この中に含まれる画素の集合を P_{ij} で表す．この P_{ij} 内の画素について，チャネルごとに独立に，H^2 個ある画素値を使って 1 つの画素値を求める．そのやり方にはいくつかあるが，画像認識では，P_{ij} の画素値の最大値を選ぶ最大プーリング (max pooling) が定番である．

P_{ij} は数画素の間隔を空けてとられるのが普通である．したがって，プーリング層では入力よりも出力のサイズが小さくなる（解像度が低下する）．図 4 に，2 × 2 の P_{ij} を縦横方向に 2 画素間隔ずつ選んだ最大プーリングの計算例を示す．なお，プーリングの計算は入力画像の各チャネルで独立に（並行して）行われる．したがって通常，プーリング層の出力のチャネル数は入力画像のチャネル数と一致する．

プーリング層も畳込み層同様，2 層構造のネットワークで表現することができ，畳込み層同様に層間の結合が局所的に限定されたものとなる．ただし結合の重みは畳込み層のフィルタのように調節可能なものではなく，固定されている．故にプーリング層には学習によって変化するパラメータは存在しない．また，プーリング層のユニットには通常，活性化関数を適用しない．

2.4　ネットワークの全体構造

典型的な畳込みネットは，入力側から畳込み層，プーリング層の順で重ね，これを何度か繰り返す構造を持つ（3 節の図 6 も参照）．ただしこの 2 種類の層はいつもペアで使われるわけではなく，畳込み層のみ複数回繰り返した後，プーリング層を 1 層重ねることもある．また，局所コントラスト正規化 (local contrast normalization) と呼ばれる画像濃淡の正規化を行う層が設置される場合もある．ただし最近の研究 [7] にはこれを不要とするものもある．

畳込み層とプーリング層の繰り返しの後には，隣接層間のユニットが全結合した（すべて密に結合した）層が配置される．これは普通の順伝播型ニューラルネットの層間結合であるが，畳込み層などと区別するために，層間が全結合 (fully-connected) であると言う．最後のプーリング層から出力層の間には，通常この全結合層が複数，連続して配置される．最後に位置する出力層は，通常のニューラルネット同様に設計される．例えば目的がクラス分類なら，この層の活性化関数をソフトマックス (softmax) 関数とする．つまり出力層には，分類したいクラス数 K と同数のユニットを並べ，うちユニット $k(= 1, \ldots, K)$ の総入力を a_k と書くとき，このユニットの最終出力を

$$y_k = \frac{\exp(a_k)}{\sum_{j=1}^{K} \exp(a_j)} \tag{4}$$

とする．これがクラス k の尤度を与えると解釈し，入力のクラス分類を行う．

2.5　畳込み層とプーリング層の役割

畳込み層は上述のとおりフィルタが表す特徴を入力から抽出する働きがあり，プーリング層は抽出された特徴の位置感度を低下させる働きがある．これを概観するため，図 5 に手書き数字の認識を目的とする畳込みネットの各層の出力例を示す[2]．この畳込みネットは，入力層から順に畳込み層 (conv1)，プーリング層

[2] データセット MNIST(http://yann.lecun.com/exdb/mnist/) を使用．

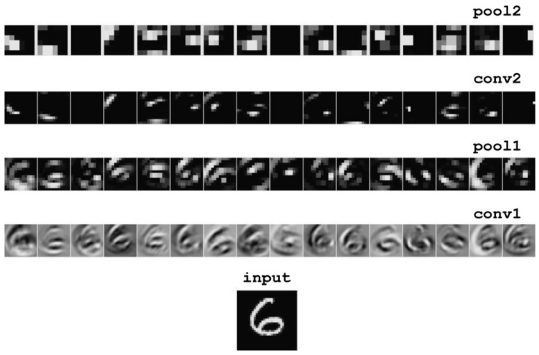

図 5 手書き数字認識のための畳込みネットの振る舞い．学習済みのネットに図一番下の画像を入力したときの各畳込み層・プーリング層の出力．

(pool1)，畳込み層 (conv2)，プーリング層 (pool2)，最後に数字 10 種に対応する 10 個のユニットからなる全結合層を持つ．入力層ではグレースケールの画像 1 枚を受け取り，これに 16 個の 1 チャネルのフィルタを畳込んで 16 チャネルのマップを得，プーリング層を経た後，16 個の 16 チャネルのフィルタを畳込んで 16 チャネルのマップを得，再度プーリング層を経て，最後に全結合層から 10 種のクラス尤度を出力する．

この図より，畳込み層の各マップ (conv1，conv2) では数字の文字形状に対応すると思われる何らかの特徴が抽出されていることが現に見て取れる．その後に続くプーリング層 (pool1，pool2) では，各マップの解像度が一律に低下しており，畳込み層で抽出された特徴の位置感度が低下するだろうことも確かめられる．

2.6 学習最適化

畳込みネットの学習最適化は，一般的な順伝播型ニューラルネットと全く同じように行える．ディープラーニングといえば事前学習 [8] が有名であるが，畳込みネットは通常これを要しない．特殊な層間結合により，多層ニューラルネット最大の問題とも言える勾配消失問題が，回避されているからだと考えられている．

学習データは，入力 $\mathbf{x}$ と望ましい出力 $\mathbf{d}$ のペアの集合 $\{(\mathbf{x}_n, \mathbf{d}_n), n = 1, \ldots, N\}$ として与えられる．$\mathbf{x}_n$ に対する畳込みネットの出力 $\mathbf{y}(\mathbf{x}_n)$ と，その目標値 $\mathbf{d}_n$ のずれの尺度となる交差エントロピーを，畳込みネットのパラメータすなわち，畳込み層の全フィルタおよびバイアスと，全結合層の結合重みおよびバイアスについて最小化する．それには，ミニバッチ単位でパラメータを更新する確率的誤差勾配法を，誤差の勾配を誤差逆伝播法で求めながら実行する．畳込みネットは畳込み層やプーリング層など構造化された層を含むが，誤差逆伝播の考え方は全く同じである．ただし最大プーリングについては，順伝播時に選択された領域内の最大値をとるユニットを記憶しておき，逆伝播時はそのユニットとのみ結合があるとみなすということを行う．

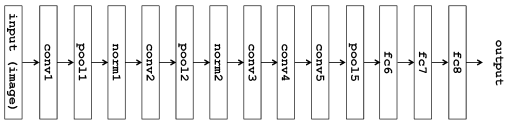

図 6 物体カテゴリ認識に代表される画像認識のための，典型的な畳込みネットの構造．入力から出力へ向けて，畳込み (convolution) 層とプーリング (pooling) 層のペアが何度か繰り返され，その後全結合 (fully-connected) 層を何度か経て，最後にソフトマックス層からカテゴリの尤度が出力される．

3. 事例：1000 カテゴリ物体認識

畳込みネットは様々な画像の問題に応用され，それぞれに成果を挙げている．ここではその中で，従来研究との性能差が著しい代表的な問題である物体カテゴリ認識を選び，実際の使用例を紹介する．物体カテゴリ認識とは，画像 1 枚が与えられ，そこに写る物体が何であるかを認識する問題である（一般物体認識とも呼ばれる）．ここでは，分野内外で高い注目を集めている ILSVRC (ImageNet Large Scale Visual Recognition Challenge) というコンテストでの問題設定を考える．そこでは，1,000 種の物体カテゴリを対象とし，各カテゴリあたり約 1,000 枚の画像が学習データ（つまりサンプル総数は約百万）として与えられている．

図 6 に示す畳込みネットは，2012 年の ILSVRC で優勝し，畳込みネットの高い能力を分野に知らしめたもの（とほぼ同じ）である [9]．以下では，これを用いたときの結果を示す．この畳込みネットは，5 つの畳込み層，3 つのプーリング層，2 つのコントラスト正規化層および 3 つの全結合層から構成される．

この畳込みネットのパラメータ（フィルタと全結合層の重み）をランダムに初期化し，ミニバッチのサイズを 128 として確率的勾配降下法を実行すると，約 200,000 ミニバッチでほぼ収束した（学習サンプルおよび検証用サンプル集合に対する誤差がともにそれ以上減少しなくなった）．全学習サンプルをひと通り処理するのを 1 回と数えると，これは 20 万 × 128/(総学習サンプル=百万)=約 25 回に相当する．学習後の conv1 層と conv2 層のフィルタを図 7 に示す．画像に直接適用される conv1 層のフィルタ（チャネル数 3）には，ガボールウェーブレット状のもの（哺乳類の初期視覚野でも観察される）や，色に反応するものが学習されている．一方，conv2 層のフィルタ（チャネル

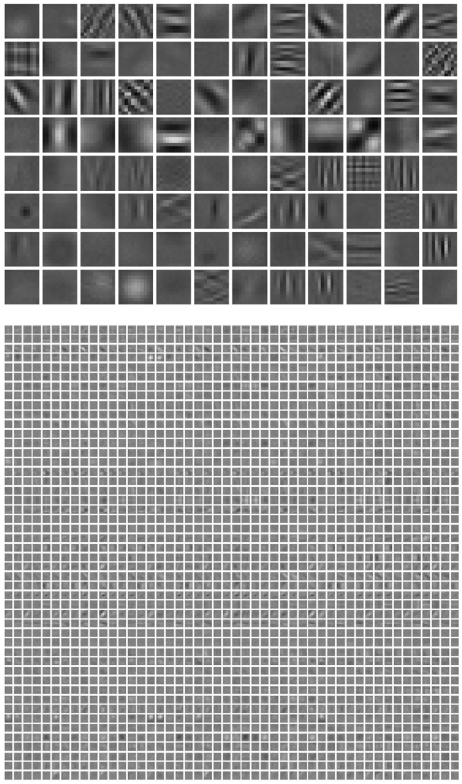

図 7　上：conv1 層の全 96 フィルタ ($11 \times 11 \times 3$)．下：conv2 層の全 256 のうち 48 フィルタ ($5 \times 5 \times 48$)．マトリクスのセルが 5×5 のフィルタの 1 チャネル分を表示し，各列がフィルタに，行がチャネルに対応する．

数 48）はサイズが小さくチャネル数が多いこともあって，直観的な解釈は難しい．

図 8 に画像 1 枚（図 9 の最初の画像）を学習後の畳込みネットに入力したときの各層の出力を示す．(a) から (c) の conv1，conv2，pool5 層ではそれぞれ，各フィルタが入力画像の何らかの特徴を取り出しているらしいことが見て取れるが，詳細な分析は難しい．また，(d) から (f) の fc6，fc7，fc8 の各全結合層の反応は，いっそう解釈が難しい．畳込みネットは，この入力画像の正解クラス（= 'lion'）を正しく答えることができた．同図 (g) のソフトマックス関数の出力に明確なピークが立っている場所（1000 ユニット中の 1 つ）が，クラス 'lion' に対応する．

図 9 に，いくつかの画像に対する畳込みネットの出力を示した．各画像隣の棒グラフは，その画像に対するソフトマックス関数の出力（クラス尤度）の上位 5 クラスを表している．5 番目の入力画像を除き，正解クラスが最上位にきている．誤答となった 5 番目の画像でも，最上位にきたクラス 'papillon' は，正解クラスの 'japanese spaniel' とよく似ており，またその尤度は 2 番目に高く，「惜しい」誤答といえる．このように，自由な背景の前で雑多なオブジェクトが多様な位置・姿勢をとっていても，かなり安定して正しい認識を行うことができる．この畳込みネットの 1,000 カテゴリの認識精度は，正解クラスを最上位に捉える場合で 6〜7 割，5 位までで約 9 割程度に達する．

4. 畳込みネットが起こした革命

4.1 解決へ向かう物体カテゴリ認識

物体カテゴリ認識はかなり以前から研究されていたが，2004 年ころ，テキスト処理における bag-of-words モデルを取り入れた bag-of-features に基づく方法（以下 BoF）が提案されるまで [10]，目立った成果はなかった．物体カテゴリ認識がなぜそんなに難しいかというと，同一カテゴリー内の見えの変動が非常に大きいことが最大の理由である．例えば同じ 'lion' が写った画像であっても，その背景も違えば姿勢も違い，さらには動物の個体差まである．このような大きな見えの変動を乗り越えるには，そんな変動に不変な特徴を画像から取り出す必要がある．その一方で，類似カテゴリーと区別できる必要もあり，それには弁別力（違いに対する敏感さ）も必要である．このような，見えの違いに対する敏感さ（弁別力）と鈍感さ（不変性）という相反する目標を，いかに両立できるかが難しかった [11, 12]．

BoF は，物体の局所的な見えに注目する方法で，逆に言えば大域的な情報，例えば物体のシルエットを扱うようなことは，原理的にできない方法であった．そのことは，上述の見えの変動に対する不変性を向上させるのに貢献し，そのことがそれ以前の従来法と大きな差をつけられた要因であるのだが，弁別力という点で明らかに問題があった．人が物を認識するとき，大域的な形が重要でなかろうはずがない．

これに対し多層の畳込みネットは，局所的な見えはもちろん，大域的な情報も取り出すことができると考えられる．そしてそのことが，BoF との大きな性能差を説明する．ただし，個別の要素に注目すると，畳込みネットと BoF でそれほど大きな違いはない．BoF でも，局所特徴の取り出しにあたって多数のフィルタの畳込みとその直後のプーリングを行っているし，画像の大域表現を求める際，再度プーリングが実行される．畳込みネットとの違いは，まず第一に層の多さであり，第二に各層のフィルタが学習によって決定できることである．

畳込みネットが物体カテゴリ認識でこのような性能

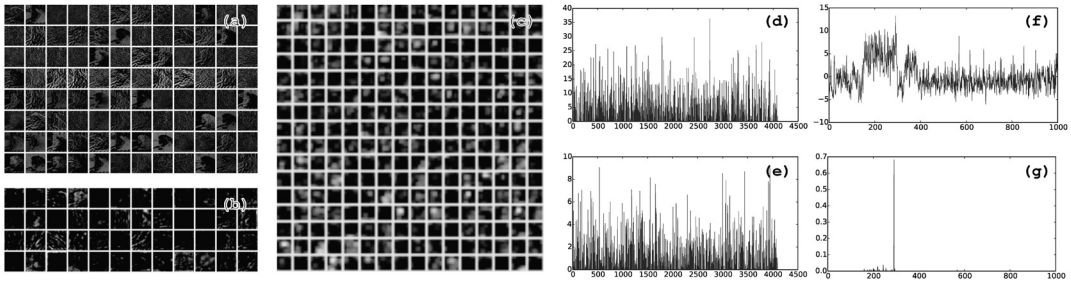

図 8　図 6 のネットワークに図 9 の最初の画像を入力したときの各層の出力．(a) conv1 層の出力である全 96 マップ．図 7 の 96 フィルタに対応する．各マップは 55 × 55．(b) conv2 層の出力である全 256 マップ中の最初の 48 マップ．各マップは 27 × 27．(c) pool5 層の出力である全 256 マップ．各マップは 6 × 6．(d) fc6 層の 4096 ユニットの出力．(e) fc7 層の 4096 ユニットの出力．(f) fc8 層の 1000 ユニットの出力（ソフトマックス適用前）．(g) fc8（適用後）．

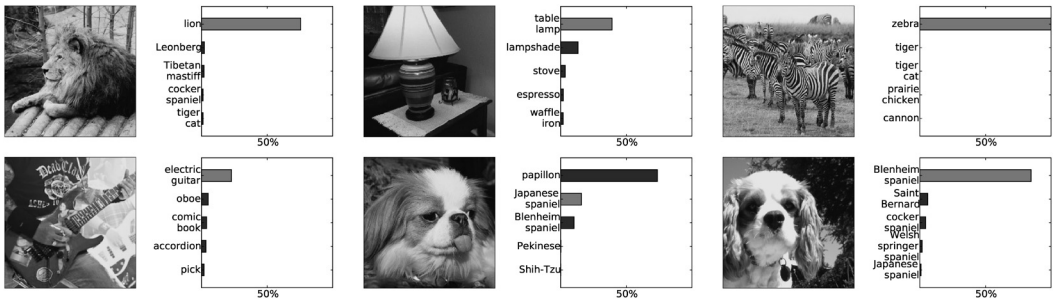

図 9　新しい入力画像に対する畳込みネットの認識結果．棒グラフはソフトマックス関数の出力上位 5 つを示す[3]．

を達成できるとわかったのは，先述のようについ最近，2012 年のことであるが，ILSVRC のコンテストが原動力となる形で，早いペースで性能が向上しつつある．トップ 5 の誤認識率は，2012 年には約 15%だったものが 2013 年には 11%になり，2014 年には 7%を切った．この精度は，人の認識性能にほぼ匹敵するという分析もあり [13]，画像 1 枚からそこに写る物の名前を答える物体カテゴリ認識は，ゴールが見えつつある．なおこれらの性能はすべて，前節で説明した畳込みネットを改良したものによって達成されている．特に最近顕著なトレンドはさらなる多層化で，Google の GoogLeNet[14] や Oxford 大の VGG[7] など，層数が 20 層を超えるものが 2014 年 ILSVRC の上位を占めている．

4.2　神経科学との接点

上述のように畳込みネットは，いくつかの認識タスクに限られるものの，人の視覚に性能で迫ろうとしている．性能だけでなく，畳込みネットはその計算の内容でも，生物の視覚系と類似していることが最近指摘されている [15]．そこでは，霊長類の高次視覚野における神経細胞の活性パタンが，多層畳込みネットの上位層のユニットの活性パタンと，高い相関を持つことが報告されている．人や動物が脳でどのように物体カテゴリ認識を処理しているかは，脳科学でも長年の謎であったが，畳込みネットは少なくともその有力なモデルを与えている．

4.3　残された課題

このように畳込みネットは，工学的な方法として優れているだけでなく，生物の視覚系の有力なモデルとなるなど，多面的に成功しつつある．その一方で，なぜ畳込みネットがそれほど高い性能を示すのか（≒生物はなぜそんな仕組みを採用したのか）は，実はよくわかっていない．なぜ畳込みとプーリングが必要なのか，それを多層化して何度も繰り返すことにどういう意味があるのか，畳込みネットは入力画像の何を「見ている」のか，等疑問はつきない．

これに答えようとする研究はいくつかあり，Poggio らの M-theory や，Mallat らの wavelet scattering network[16]，あるいは Arora らの研究 [17] がある．畳込みネットの可視化の試み [18] や，畳込みネットが思わぬ誤認識を起こす性質 [19] など，興味深い研究は

[3] 画像はすべてクリエイティブ・コモンズ・ライセンス（CC BY あるいは BY-SA 2.0）に従い利用している．上段から下段，左から右へ順にそれぞれのリンク先は次のとおり：
https://www.flickr.com/photos/elpadawan/8238633021,
https://www.flickr.com/photos/38009628@N08/10085782733,
https://www.flickr.com/photos/malczyk/5638610203,
https://www.flickr.com/photos/monavelion/5032771365,
https://www.flickr.com/photos/lostintexas/482312645,
https://www.flickr.com/photos/14511253@N04/4531941062.

いくつかある．しかしながらいずれも，畳込みネットを完全に理解したとは言えない．

中身の理解は置いておいて，畳込みネットの新たな応用先も盛んに探求されている．中でも最も注目を集めているのが，物体検出と動画像認識であろう．物体検出は，与えられた画像の中で，どこにどんな物体が存在するかを言い当てるタスクである．物体のカテゴリを答えるだけでなく，その画像に占める位置を特定する必要がある分，数段難しい．物体検出でもやはり，畳込みネットを使う方法が現状で最も高性能だが，人の視覚には性能面で遠く及ばない．また動画像の認識，つまり動画像（ビデオ画像）が与えられたとき，そこに写る人の行動やシーンの意味を理解することも，まだ難しい問題である．最も高い性能を示しているのはやはり畳込みネットだが，その性能は低い．動画像の認識では時間軸方向の情報が大事なはずだが，これをうまく使える方法が今のところ知られておらず，研究は道半ばである．

参考文献

[1] K. Fukushima and S. Miyake, “Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position,” *Pattern Recognition*, **15**, pp. 455–469, 1982.

[2] D. H. Hubel and T. N. Wiesel, “Receptive fields, binocular interactions, and functional architecture in the cat’s visual cortex,” *Journal of Physiology*, **160**, pp. 106–154, 1962.

[3] Y. Lecun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition, *Neural Computation*, **1**, pp. 541–551, 1989.

[4] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, “Gradient-based learning applied to document recognition,” In *Proceedings of IEEE*, **86**, pp. 2278–2324, 1998.

[5] D. H. Hubel and T. N. Wiesel, “Receptive fields and functional architecture of monkey striate cortex,” *The Journal of Physiology*, **195**, pp. 215–243, 1968.

[6] P. Berkes and L. Wiskott, “Slow feature analysis yields a rich repertoire of compelx cell properties,” *Journal of Vision*, **5**, pp. 579–602, 2005.

[7] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014. *arXiv*.

[8] G. Hinton, S. Osindero and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” *Neural Computation*, **18**, pp. 1527–1544, 2006.

[9] A. Krizhevsky, I. Sutskever and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” In *Proceedings of Neural Information Processing Systems*, 2012.

[10] G. Csurka, C. Dance, L. Fan, J. Willamowski and C. Bray, “Visual categorization with bags of keypoints,” In *Proceedings of European Conference on Conputer Vision*, **1**, 2004.

[11] J. J. DiCarlo, D. Zoccolan and N. C. Rust, “How does the brain solve visual object recognition?” *Neuron*, **73**, pp. 415–434, 2012.

[12] N. C. Rust and J. J. DiCarlo, “Selectivity and tolerance (“invariance”) both increase as visual information propagates from cortical area v4 to it,” *The Journal of Neuroscience*, **30**, pp. 12978–12995, 2010.

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg and L. Fei-Fei, “Imagenet large scale visual recognition challenge,” 2014, *arXiv*.

[14] C. Szegedy, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, “Going deeper with convolutions,” 2014. *arXiv*.

[15] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert and J. J. DiCarlo, “Performance-optimized hierarchical models predict neural responses in higher visual cortex,” In *Proceedings of the National Academy of Sciences of the United States of America*, 2014.

[16] J. Bruna and S. Mallat, “Invariant scattering convolution networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, pp. 1872–1886, 2013.

[17] S. Arora, A. Bhaskara, R. Ge and T. Ma, “Provable bounds for learning some deep representations,” 2013. *arXiv*.

[18] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” In *European Conference on Computer Vision*, 2013.

[19] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow and R. Fergus, “Intriguing properties of neural networks,” 2013. *arXiv*.

