ディープラーニングによる画像認識
2 ―畳込みネットワークの能力と限界―

岡谷貴之（東北大学大学院情報科学研究科 / JST CREST）

画像認識とディープラーニング

ディープラーニングは近年，人工知能の諸問題で軒並み良い成果を挙げており，画像認識はその1つに数えられる．画像認識にも色々あるが，最も大きな成功といえるのは，1枚の画像からそこに写る物体の名前を答える「物体カテゴリ認識」だろう．ディープラーニングの方法は，他の方法を相手にしない高い性能を挙げ，その能力は人に届こうとしている．

物体カテゴリ認識では，畳込みニューラルネットワーク（convolutional neural network，以下CNN）が中心的な役割を果たす．CNNは，フィードフォワードニューラルネットワークの一種だが，畳込み層とプーリング層と呼ばれる特殊な構造の層を内部に持つ．1980年代後半に考案され，当時は文字認識に主に適用されていたが，最近になって物体カテゴリ認識に適用され，きわめて高い能力があることが分かった．

この発見の原動力となったのは，ImageNet Large Scale Visual Recognition Challenge（ILSVRC）というコンテストである．1,000種の物体カテゴリを認識対象とし，各カテゴリあたり約1,000枚の画像，計約百万枚の画像が学習データとして与えられる．図-1に実際の認識の例を示す．CNNは2012年に最初にILSVRCに登場し，その後も順調に性能を向上させてきた．誤認識率（リストアップしたカテゴリ候補5つに，正解が含まれない場合の割合）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．この数字は，人の認識性能に比肩しつつあり，物体カテゴリ認識はゴールが見えつつある．

物体カテゴリ認識での成功に後押しされ，CNNはその他のさまざまな画像認識の問題に適用されてきている．最近の面白い例を1つ挙げると，与えられた画像の情景を自然な文章で記述する方法の研究がある（Vinyalsらの'show and tell'）．そこでは，画像から特徴を取り出すCNNと，文章を生成するRNN（recurrent neural network，再帰的ニューラルネット）を順番に，接続した複合ニューラルネットが用いられており，既存手法を大きく上回る精度の記述が得られるという．

畳込みニューラルネット（CNN）

歴史

CNNのルーツは，1980年前後にFukushimaらが発表したネオコグニトロンにある．これは，1960年ころにネコの脳で発見された単純型細胞・複雑型細胞の働きにヒントを得た実験的なパターン認識システムであった．1980年代後半になってLeCunらは，このネオコグニトロンと同じ構造を持つネットワークの学習に，誤差逆伝播法（back propagation，以下BP）に基づく勾配降下法を適用し，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と名付けられたこのCNNは，現在画像認識で広く使われているCNNとほとんど同じものであった．

CNNの構造

CNNはフィードフォワードニューラルネットワークの一種である．フィードフォワードニューラルネットワークでは，複数のユニットからなる層が何層も重なった構造を持ち，1つの決まった方向に情報が伝播する．ある層がI個のユニットからなり，次の層がJ個のユニットからなるとする．最初の層のi（=1, …, I）番目のユニットがx_iの出力をとるとき，次の層のj（=1, …, J）番目のユニットはこの層から$a_j = \sum w_{ij}x_i + b_j$を受け取り，さらに次の層へ$y_j = f(a_j)$を出力する．$w_{ij}$はユニ

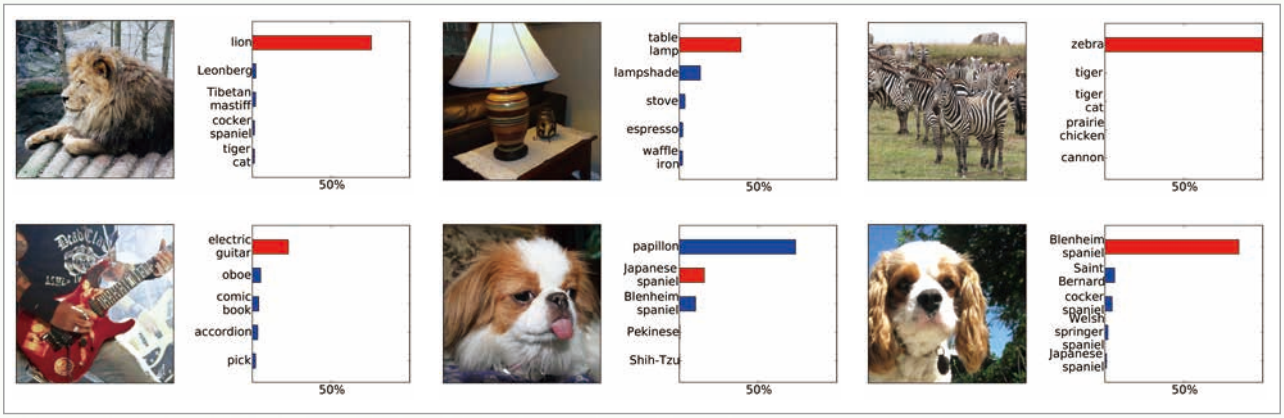

図 -1　CNN による物体認識の例．棒グラフはカテゴリらしさ（尤度）の上位 5 つを示す．赤色のバーが正解カテゴリを示す

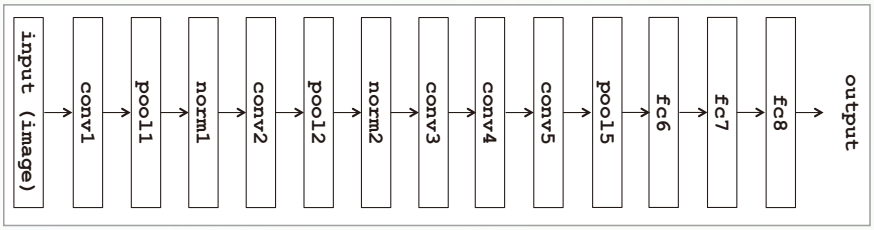

図 -2　典型的な CNN の構造．畳込み層（conv?）とプーリング層（pool?）のペアが何度か繰り返され，その後何層かの全結合（fc?）層を経て，カテゴリ尤度を出力する

ット i と j を結ぶ結合の重み，b_j はユニット j が持つバイアスと呼ぶ．f は非線形関数で活性化関数と呼ばれる．入力層で受け取った入力が以上の計算を繰り返して出力側へと一方向に伝えられ，最後に出力層から出力される．画像認識を行う CNN は画像を入力に受け取るが，画像の各画素が入力層の 1 つのユニットに入力される（したがって，入力層には画素数（カラー画像ならその 3 倍）と同じ数のユニットがある）．

入力層と出力層の間には，畳込み層，プーリング層および全結合層と呼ばれる 3 種類の層が配置される．典型的には，入力側から畳込み層，プーリング層をこの順に重ね，これが何度か繰り返される（図 -2）．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，プーリング層を配置することもある．ほかにも，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使われることもある．

畳込み層とプーリング層の繰り返しの後には，全結合層が（通常，複数連続して）配置される．全結合層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）最も普通の（ニューラルネットの）層である．

最後に位置する出力層は，通常のニューラルネット同様，目的に合わせて設計される．たとえば目的がクラス分類なら，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, …, 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 $y_k = \exp(a_k) / \sum_{j=1}^{K} \exp(a_j)$ 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．回帰が目的であるなら，出力層には，説明したい変数と同数のユニットを配置し，ユニットの活性化関数は変数の値域に合わせ，シグモイド関数や線形関数などを選ぶ．

畳込み層

畳込み層は，画像にフィルタ（＝小さな画像）を畳込む演算を行う層である．畳込みは，画像から特徴を抽出する最も基本的な方法である．入力画像にフィルタと何らかの類似性のある局所パターンがあるとき，その位置と類似度の大きさを出力する．

実際の畳込み層は図 -3 のように，多チャネルの画像を入力に受け取り，また出力する．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．たとえば CNN の入力となる RGB の 3 色からなるカラー画像では K=3 であり，それ以外の層ではそれ以上のチャネル数（K=16 や K=256 など）を扱う．

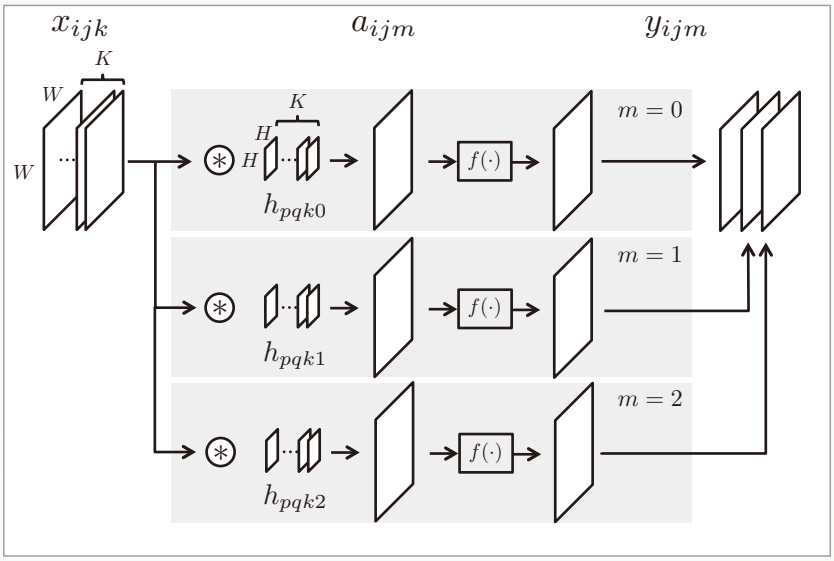

図 -3　畳込み層の概要．K チャネルからなる画像を入力にとり，3 種類のフィルタ（縦横 $H \times H$ 画素，サイズ $H \times H \times K$）を適用し，3 チャネルの画像（マップ）を出力する場合

図のように，これに畳込まれるフィルタは複数あり，それぞれ入力画像と同じ数のチャネルを持ち，フィルタごとに計算は並行に行われる．計算の中身は，そのフィルタのチャネルごと並行に画像とフィルタの畳込みを行った後，結果を画素ごとに全チャネルに渡って加算する．

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

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．

こうして得た a_{ijm} に活性化関数を適用し，出力 $y_{ijm} = f(a_{ijm})$ を得る．活性化関数には，正規化線形関数（rectified linear）すなわち $f(x) = \max(x, 0)$ を使うのが近年のスタンダードである．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．

畳込み層は以上の演算が行われるような単層ネットワークとして構成される．すなわち，x_{ijk} を受け取る入力側の層と，y_{ijm} を出力する層の間で，上の畳込みの計算を実現するようにユニット間の結線がなされる．その結合の重みはフィルタの係数であり，畳込み層ではフィルタの係数が学習の対象となる．

プーリング層

プーリングとは，入力画像の空間解像度を低下させる処理である．通常，畳込み層の直後に設置され，畳込み層で抽出された特徴の位置感度を低下させ，入力パターンの微小並進移動に対する出力の不変性を実現する役割を果たす．具体的には図 -4 のように，入力画像の局所領域（図では 2 × 2 画素）から，それを代表する値を 1 つ選ぶ．代表値の選び方には幾通りかあるが，最大値を選ぶ最大プーリングが最もよく使われる．局所領域は隣と互いに重なるようにとってもよいが，その場合でも出力の空間解像度は入力よりも必ず低下するようにする．また以上の処理は入力画像の各チャネルで並行に行われるのが普通である．

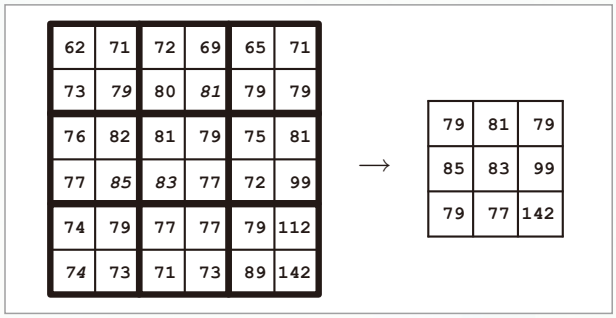

図 -4　プーリング層の概要．6 × 6 の入力画像に 2 × 2 を 1 つの値にする最大プーリングを，2 画素間隔で適用した例．出力は 3 × 3 となる

プーリング層も畳込み層同様，単層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．ゆえ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．

CNN の学習

CNN の学習は，一般的なフィードフォワードニューラルネットとまったく同じように行う．学習データは，入力 $\mathbf{x}$ と CNN 全体の出力の目標値 $\mathbf{d}$ 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 CNN の出力 $\mathbf{y}(\mathbf{x}_n)$ と，その目標値 $\mathbf{d}_n$ のずれ（誤差）が小さくなるように，パラメータ（結合重みとバイアス，畳込み層のフィルタの係数を含む）を決定する．誤差の尺度には，クラス分類では交差エントロピーが，回帰では 2 乗誤差がよく用いられる．

問題の規模の大きさから，誤差の最小化には勾配降下法を使うのが主流である．それも，全サンプルから選んだ 100 個前後のサンプルの集合＝「ミニバッチ」に対する誤差の和を最小化する確率的勾配降下法を使うのが一般的である．ミニバッチ 1 つに対しパラメータ修正を行い，毎回ミニバッチを取り替えながら反復する．必要となる誤差の勾配（パラメータによる微分）は，BP を使って求める．畳込み層やプーリング層など構造化された層を含むが，BP 自体の考え方は通常のネットワークとまったく同じである．なお最大プーリングを行う層では，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見なすということを行う．

脳（視覚皮質）との関係

単純型細胞と複雑型細胞

CNN の畳込み層およびプーリング層は，脳の初期視覚野で発見された単純型細胞および複雑型細胞の振舞いをモデル化したものである．

脳の視覚系では，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ーンが入力されると活性化し，それ以外のときは活性化しないという，選択的な振舞いを示すものが多く見られる．単純型細胞（simple cell）は，そのような細胞の 1 つで，網膜（あるいは視野）の特定の位置に特定の向きの線分が提示されると，選択的に活性化する細胞である．一方，複雑型細胞（complex cell）は，線分の向き選択性はそのままに，位置選択性がわずかに低下したものである．

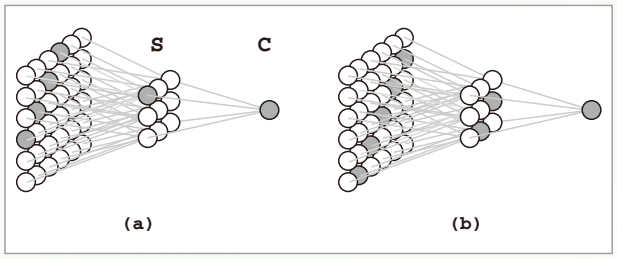

図 -5　単純型細胞と複雑型細胞のモデル．説明は本文を参照

これらの細胞の振舞いは，図 -5 の 2 層ネットワークによってモデル化できる．最も左の層が入力層で，次の中間層の各ユニットは，入力層の 4 × 4 のユニット群のみと結合を持つ．そしてこれらのユニット群に特定のパターンが入ると，それと結合を持つ中間層のユニット 1 つが活性化するようになっている．活性化する入力パターンは中間層のユニットすべてで同じである．この中間層の各ユニットが単純型細胞のモデルであり，その働きは畳込み層のそれである．

このネットワークの最も右に位置するユニットは，中間層の 3 × 3 のユニットすべてと結合を持ち，これらのうち 1 つでも活性化すると，このユニットも活性化するようになっている．ネットワーク全体の入力が図 -5 (a) から (b) のように変わると，それにつれて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，どちらの場合でも活性化したままである．このように，中間層のユニットと異なり，出力層のユニットは一定の（この例では 3 × 3）範囲の位置ずれを許容する（プーリング層の働き）．このユニットが複雑型細胞のモデルである．

高次視覚野

CNN は，図 -5 のような畳込み層とプーリング層のペアが何度か繰り返される多層構造を持つ．このペア 1 つが実現する働きは簡単に理解できる（またそれだけで高度な仕事ができるわけでないことは想像がつく）．しかしながら，これを繰り返す多層構造が行う計算に，どんな意味・働きがあるのかはよく分かっていない．

その一方，物体カテゴリ認識を行う（多層の）CNN

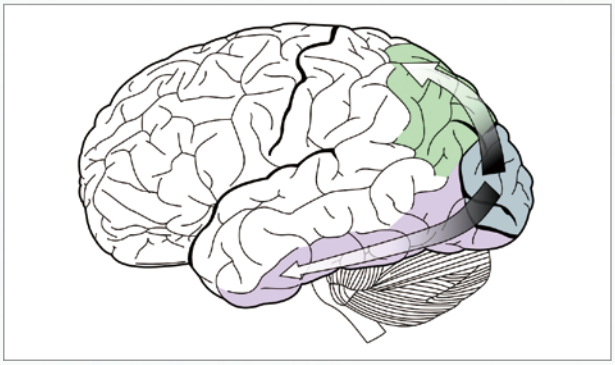

図 -6　腹側視覚皮質路（下側矢印）と背側視覚皮質路（上側矢印）.
(Drawing By Selket, available under CC-BY-SA 3.0.)

が行っている計算は, 同様の認識を行う霊長類の脳（視覚皮質）での計算とよく似ていることが，最近の神経科学の研究で明らかになりつつある．同様の認識とは正確には，ある物体を見たとき（＝物体が中央付近に写る静止画を見たとき），それが何であるかを，まさに上述したような見えの変動，すなわち背景や物の写り方の変化に影響されずに素早く知覚することである．

この機能は，図 -6 に示す脳の下側（下側頭回）にある腹側視覚皮質路と呼ばれる経路で処理されている．この経路は，単純な特徴抽出を行うと言われる低次視覚野を通り，より複雑な計算を行っているとされる高次視覚野に至っている．眼から取り込んだ情報はこの順に経路に沿って流れ，画像提示からこの経路の終端に達するのに 100 ミリ秒のオーダの時間を要する．人やサルは，同程度の時間で物の認識を行うことができる．

DiCarlo らは，この経路にそってサルの皮質に多点電極を複数個埋め込み，特定の画像をこのサルに提示したときの，各電極位置での神経活動を記録できるようにした．そして，同じ画像を CNN に入力したときの各層の活性パターンを，記録された神経活動と比較した．具体的には，同じ画像を見せたときの両者の反応のペアを一定数集め，片方から残りを機械学習の方法で予測し，その精度を比較した．なお記録される神経活動は時系列信号であるが，時間方向の平均活性度を使う．すると上の予測精度は，CNN の上位層とサルの高次視覚野の間で，特に高くなると分かった．たとえば，高次視覚野の記録だけから，同じ画像を入力したときの CNN の上位層の活性パターンを高い精度で予測することができる．一方，CNN の低い層とサルの高次視覚野や，その逆の組合せについての予測精度は低かった．これらの結果は，多層 CNN が行っている計算と，腹側視覚皮質路の特に終端にかけて行われている計算内容が，互いに近いことを意味するものである．

以上のようなことから，CNN は現在，物体認識にかかわる視覚皮質の有力な計算モデルとなっている．CNN が各層で取り出す特徴には層と対応した階層性が見られるが，これが腹側視覚皮質路に関する神経科学の知見とよく合致することも理由である．

まとめと展望

CNN の強みと謎

以上で述べてきたように，CNN は長年の課題であった物体カテゴリ認識を解決しつつある．物体カテゴリ認識を難しくする要因は，同一カテゴリ内の見えの変動がきわめて大きいことにある．まったく同じ物体でさえ，見る方向や照明の違い，背景の違いによって画像は変わる．さらにたとえば椅子にはさまざまな形のものがあるように，同一カテゴリの異なる物体の見えにも大きな変動がある．

したがって物体カテゴリ認識を行うには，そういった変動に不変な特徴を画像から取り出せる必要がある．それと同時に，類似カテゴリとの区別を可能にする弁別力（違いに対する敏感さ）も必要である．難しいのは，この 2 つ（不変性＝鈍感さと弁別力＝敏感さ）が互いに相反することであり，にもかかわらずこれらを両立させないといけないことである．

この両立の難しさは，CNN の登場前に標準的に使われていた Bag of Visual Words（BoVW）の長所と短所を考えると分かりやすい．BoVW は，画像の局所領域の見えを特徴化して符号化し，そして画像内でその頻度をとったものを，画像全体の特徴量とする方法である．見えの大域的な情報を思い切って捨てることで，上述の見えの変動に対する特徴量の不変性を実現し，それが BoVW の成功につながった．しかし同時に，大域情報を捨ててしまったことで BoVW は十分な弁別

力が持てなくなり，そこに限界があった．人が物を認識するとき，大域的な情報―たとえば物の形―を使っていないはずがない．CNN が BoVW より高い性能を示す理由は，この BoVW が捨てていた大域的な情報を捉えつつ，課題であった不変性を実現できていることにあると考えられる（図 -7）．

このような CNN の能力は，CNN が多くの層を持つことで実現されていると考えられる．しかしながらなぜ，畳込み層とプーリング層を何度も繰り返す多層構造が，そんなこと（＝不変性と弁別力の両立）を可能にするのかは，いまだに大きな謎である．たとえば上述の見えの変動に対する不変性が，どういう仕掛けでどのように実現されているかが数学的に記述できれば良いが，上述のような物体カテゴリで問題となる見えの変動そのものを数学的にどう表現できるかが，そもそも分かっていない．

CNN の限界

本稿で主題とした物体カテゴリ認識は，画像認識のあまたあるタスクの 1 つに過ぎず，ほかにも多くのタスクがある．物体カテゴリ認識の成功を受けて，CNN はこれらのタスクにも次々に試されている．物体カテゴリ認識と類似の（と考えてよさそうな）タスク―たとえば顔の認識―では同じように高い（＝人に匹敵する）性能を示している．しかしながら，それ以外のタスクでは，従来手法より良いという程度には一定の成功を収めてはいるものの，人と同等の能力に至るまでには大きなギャップがある．

そんなタスクの例を挙げると，物体検出（どこにその物体があるか），セグメンテーション（物体の背景との境界の特定）や，人の姿勢推定などがある．どれも盛んに CNN が応用され，少なくとも従来手法には勝り，今も性能は上昇過程にある．しかしながら筆者には，これらのタスクについては今の技術の延長線上に人に匹敵する性能があるようには思えない．

そう考える理由の 1 つは，CNN が物体カテゴリ認識で成功した要因，すなわち上述の見えの変動に対する不変性は，上のようなタスクが含む「位置を求める問題」と根本的に相容れないという事実である．もう 1 つの理由は，上で言及した脳の腹側視覚皮質と CNN との高い類似性である．そこでの中心的な働きが，背景や姿勢に影響されない瞬時的な物体認識なのであれば，他のタスクは，脳では別の仕組みで処理されていても不思議はない．実際図 -6 に示すように，脳には背側視覚皮質路と呼ばれる処理経路もあり，これは where 経路と呼ばれ，空間のどこにあるのかを認識する働きを司るとされる．そこで行われている計算はまだはっきりしない．

図 -7　物体カテゴリ認識のための学習済み CNN に対し，その特定カテゴリ出力を最大化する入力画像．左：crane（ツル）．右：starfish（ヒトデ）．CNN が大域的な形を捉える能力があることを示す証左と言える

このほかにも未解決の（＝人ができるようには行えない）画像認識の問題はまだたくさんある．そのうちの 1 つに動画像認識，たとえばビデオから人の行動の内容を認識する問題がある．近年，動画像認識にも CNN が適用され，たとえば畳込み層を時空間方向に拡張する方法などが試されている．しかしながら今のところ，目を見張るような成果は得られていない．

CNN の画像認識への応用は大きなブレークスルーとなった．物体カテゴリ認識を解決に近づけるとともに，多くの画像認識に適用され，その性能向上に貢献した．とは言え，CNN がすべてを解決したわけではない．未解決の画像認識の問題はまだまだたくさんあり，これらに対する「次の一手」が待たれる状況に差し掛かっている．

（2015 年 4 月 8 日受付）

岡谷貴之（正会員）　okatani@vision.is.tohoku.ac.jp

東北大学大学院情報科学研究科教授．1999 年東京大学工学系研究科計数工学専攻博士課程修了．同年東北大学大学院情報科学研究科助手．その後講師，助教授，准教授を経て 2013 年より現職．