The Japanese Journal of Psychonomic Science
2016, Vol. 35, No. 2, 1–14
DOI: http://doi.org/10.14947/psychono.35.●

解　説

深層学習をめぐる最近の熱狂

浅　川　伸　一
東京女子大学

Recent excitement about deep learning

Shin Asakawa
Tokyo Women's Christian University

Brief introduction about current trends in deep learning was intended, including such as convolutional neural networks, and regions with convolutional neural networks. They have features as end-to-end, general purposes, and implementable based on recent advances in computer science. Object recognition of convolutional neural networks overwhelmed human performance. This tidal wave might give deep impact on all the areas in psychology.

Keywords: deep learning, convolutional neural networks, regions with convolutional neural networks, object recognition

1. はじめに

本稿では多層ニューラルネットワーク，深層学習の最近の動向を概観する。従来手法を凌駕する研究成果が発表されたため，多数の人びとの関心を惹きつけている。その結果，環境が整備され，それにより，さらに多くの参入者を呼び込んでいる。環境が整備されたとは，ノートPCでも実行可能なフレームワークが無料で公開され，新しいアルゴリズムを手軽に確認可能である点が大きい。ニューラルネットワークでは学習に合成関数の微分公式であるチェインルールを用いた誤差逆伝播則をもちいるが，公開されているフレームワークでは自動微分機能が実装されている。複雑な関数を用いたとしても，モデルの最適化やパラメータ調整はフレームワークに任せることができる。そのため手軽に試すことができ，参入者には敷居が低い。この状況をマスメディアは第3次ニューロブーム，あるいは第3次人工知能ブームと呼んでいる。加えて，新しい研究成果は査読プロセスを踏襲する科学論文の従来の刊行過程にとらわれない。論文はプレプリントとしてarXiv[1]にアップロードされる。同時に用いたプログラムのソースコードがGitHub[2]で公開されている。このような流れが研究の進展を加速している（4章）。このような状況を背景として，百花繚乱の感のある昨今の趨勢は3つの流れに要約できよう。すなわち，（1）畳込みニューラルネットワーク（CNN），（2）リカレントニューラルネットワーク，（3）強化学習（DQN）である。

知的情報処理を回路に求めるのか，素子に求めるのか，アーキテクチャに求めるのか，アプローチは様々である。CNN（福島，1976; Fukushima, 1987; Krizhevsky, Sutskever, & Hinton, 2012b; LeCun, Bottou, Bengio, & Haffner, 1998）はアーキテクチャに求め（3章），リカレントニューラルネットワークの変種であるLSTM（Long Short-Term Memory）（Gers, Schmidhuber, & Cummins, 2000; Hochreiter, Bengio, Frasconi, & Schmidhuber, 2001; Hochreiter & Schmidhuber, 1997）は素子に求めた。層数，ニューロン数，を固定し，ニューロン間の結合係数を変化させることで学習が進行するとの考え方は，ニューラルネットワークの伝統である。このような方法が花開いた意味は心理学にも影響を与えるもので，看過できないと考える。キーワードとして挙げれば，エンドツーエンド（end-to-end），汎用化，実用化である。CNNでは従来からの画像認識，音声認識，自然言語処理で用いられてき

Corresponding address: Tokyo Women's Christian University, 2–6–1 Zenpukuji, Suginami-ku, Tokyo 167–8585, Japan.
E-mail: asakawa@ieee.org
電子付録（Figure S1–6）はJ-STAGEにて公開しております（論文URL　http://dx.doi.org/10.14947/psychono.35.●）。

[1] http://arxiv.org/
[2] https://github.com/

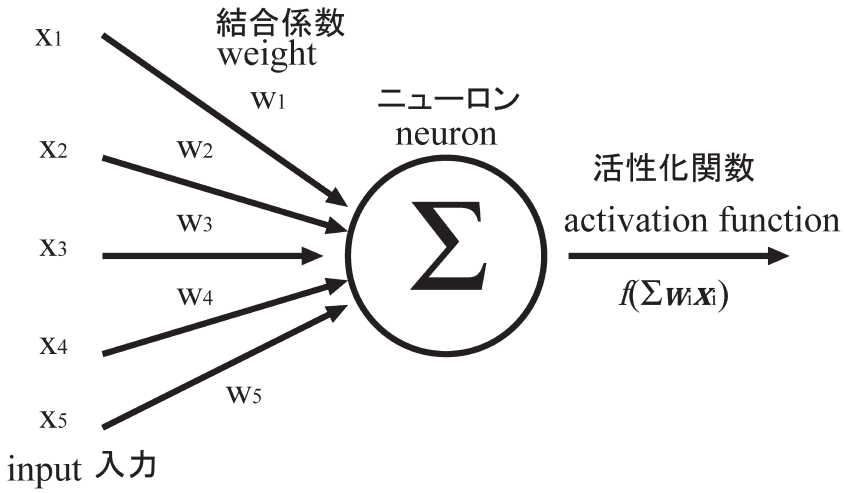

Figure 1. A schematic description of a neuron.

た職人芸的，手工芸的で複雑な前処理を必要としないという特徴が挙げられる。

2. Crickの批判

Watsonと共著でDNAの二重らせん構造を提唱したCrickはおよそ30年前，“ニューラルネットワークをめぐる最近の熱狂”と題した論文をネイチャーに上梓した（Crick, 1989）。Crickは，当時隆盛を極めた誤差逆伝播則（Rumelhart, Hinton, & Williams, 1986）には生物学的妥当性がないことを指摘し，ニューラルネットワークモデルは宇宙人の工学（alian technology）であって，非現実，かつ，日和見主義的であると批判した。ほどなくしてニューラルネットワークは2度目の冬の時代を迎える。

ニューラルネットワークにおける1度目のブームとは1950年代のRosenblattのパーセプトロンによる。1度目のブームはMinskyとPapertがパーセプトロンの限界を指摘したことにより下火となった。Crickの批判したブームが2度目のブームであった。

ところが2012年を境に3度目のニューラルネットワークブームが起こった。誤差逆伝播則は拡張され，微分可能であれば，ほとんどのネットワーク構成を許すようになった。Crickはおそらく不満であろうが，彼が想像しなかった成果を産むに至った。2012年の大規模画像認識チャレンジ[3]コンテストにおいて，深層ニューラルネットワーク（Krizhevsky et al., 2012b）がそれまでのサポートベクターマシン（support vector machines: SVM）（Vapnik, 1995, 1998, 1999）の認識性能を凌駕したことによる。同年Googleの持つ大量の画像からネコと人間を識別する「グーグルのネコ」が話題となった[4]。2015年には人間の認識成績を凌駕した[5]。2016年には思考ゲームの中で探索空間が最も広い囲碁においても人間の世界チャンピオンを破った（Silver et al., 2016）。影響は科学，工学分野に留まらない。人文科学，社会科学への影響も計り知れない。人間の認識を扱ってきた心理学は，自らが依拠する認識機構以外の機構を持つに至った。この知的情報処理機構は内部構造を詳細に同定可能である。

Crickが述べたように第3次ニューロブームも日和見主義的なのであろうか。これに対しては，今度のブームは本物だという見解も存在する。理由は一般物体認識と呼べるまで認識の質と量が向上したこと，汎用人工知能（artificial general intelligence: AGI）と呼べる一般性の高いアルゴリズムを採用していることが理由である。チェスの世界チャンピオンを破った当時のアルゴリズムは，チェスに特化した特化型人工知能であった。これに対してAlphaGoの採用しているアルゴリズムは一般画像認識で用いられてきたCNNを採用し，解の探索に強化学習（Auer, Cesa-Bianchi, & Fischer, 2002; Mnih et al., 2015; Sutton & Barto, 1998）を用いている。

2.1 第3次ニューロブーム

ニューラルネットワークモデル（neural network models）の進展を概説する。ニューラルネットワークモデル

[3] http://image-net.org/challenges/LSVRC/2012/

[4] http://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-machine-learning.html?_r=0

[5] http://image-net.org/challenges/LSVRC/2015/

の歴史は古く，現代的なコンピュータの黎明期1950年代に遡る。現在でも基本的なコンピュータの構造に名を残すvon NeumannがMcCullochとPittsによる形式ニューロン（formal neuron）のモデル（McCulloch & Pitts, 1943）を「雷に打たれたようだ」と形容した（Neumann, 1958）。McCullochとPittsの形式ニューロンモデルは，ニューロンが発火している状態を1,そうでなければ0とする。形式ニューロンは論理演算回路であり，ブール代数（Boolean algebra）（Boole, 1854）の実現と見做しうる。これにより，当時真空管ですらなかったコンピュータとニューロンとが結び付けられた。ニューラルネットワークの第一世代（1956年）Rosenblattが提案したパーセプトロン（perceptron）は形式ニューロンを素子とする（Rosenblatt, 1958, Figure 1）。

同じ時期，コンピュータに知的情報処理を実装する人工知能第一世代も開始された。2016年初頭に亡くなったMinskyやMcCarthy, Shannonらが開催したダートマス会議（1956年）において初めて人工知能という言葉が使われた。1962年MinskyとPapertは線形パーセプトロンには解けない問題があることを指摘した（Minsky & Papert, 1988）。これにより，第一次ニューラルネットワークは終焉を迎え，記号主義的人工知能研究に圧されてニューラルネットワーク研究は下火となった。

再びニューラルネットワーク研究が台頭するのは1986年以降である。この年Rumelhart, Williams, Hinton, McClellandらによりパーセプトロンの限界を打破する誤差逆伝播（back-propagation）則が提案された（Rumelhart et al., 1986）。

第一世代のパーセプトロンは，入力情報を出力情報へと変換する自動学習装置である。パーセプトロンの学習とは，入力層のニューロンと出力層のニューロンとの間のシナプス結合強度をHebbの学習則（Hebb, 1949）によって変化させるモデルである。層内のニューロンに内部結合は存在しない。Hebb則においては，シナプス前ニューロンが活動し，同時にシナプス後ニューロンも活動した場合，そのシナプス結合荷重は強化されると考える。したがって，正解（あるいは報酬）が与えられた場合のみ行動が変容することを仮定する。これに対して誤差逆伝播則は，正解との差異，すなわち誤差の自乗和を，シナプス結合強度で微分し，各結合強度で加重して下位層へ伝播させる。これが誤差逆伝播法と呼ばれる所以である。誤差逆伝播則は入力層と出力層だけでなく，中間層を許す。このため，入出力の表象に拘束されない内部表象を獲得することが期待される。哲学的思弁の対象であって神秘化されたり，データから類推するしかなかった内部表象をモデルとして表現し，シミュレーション可能であることを示した意義は大きい。

1980年代は，この他にも最適化問題や連想記憶モデルであるHopfieldモデル（Hopfield, 1982; Hopfield & Tank, 1985, 1986), Kohonenの自己組織化ネットワーク（self-organizing networks）（Kohonen, 1985; Kohonen, 1996; Oja & Kaski, 1999）などが提案され活況を呈した。一方，古典的人工知能研究ではエキスパートシステムの提案により，チェスなどに特化した人工知能が応用可能性を模索した時代でもあった。

1992年VapnikはSVMを提唱しクラス分類の枠組みを定式化した（Vapnik, 1995, 1998, 1999）。SVMは誤差逆伝播則よりも数学的見通しに優れていた[6]。このため，パターン認識研究の主流はSVMに移行した。

画像認識の分野においてニューラルネットワークモデルがSVMの性能を凌駕したのが2012年である。この年に行われた大規模画像認識コンテストにおいて，前年までのSVMの認識結果を10%以上凌駕して優勝したCNNがAlexNet（Krizhevsky et al., 2012b）であった。以来，畳込み演算を用いた多層ニューラルネットワークが画像認識の主流となった。2015年には残渣ネット（He, Zhang, Ren, & Sun, 2015a）と呼ばれるCNNモデルが152層を数え，人間の成績を凌駕した。残渣ネットは人間の評価者が誤分類するような画像でも正しく分類できる。加えて2016年，Googleが買収したスタートアップDeepMind社のAlphaGoが人間の囲碁世界王者を破った。AlphaGoは碁盤の局面認識に深層ニューラルネットワークを用い，強化学習とモンテカルロ木探索（Monte Carlo tree search）を組み合わせている（Silver et al., 2016）。総説論文としてはLeCun, Bengio, & Hinton（2015), Schmidhuber（2015）がある。

知的情報処理はもはや人間だけが持つものとは言えない。科学，医学，社会，経済，思想，人文，宗教，教育，などに及ぼす影響は計り知れない。

3. C N N

3.1 CNNを用いた画像認識

Figure 2に画像認識技法の変化を模式的に示した。画像認識においては，画像の前処理，特徴抽出，分類，などの技法を割当て処理し分類することが伝統的に行われてきた。ところが近年のCNNに代表される深層学習の発展により，前処理である特徴抽出も最終的な分類判断

[6] 判別境界を与える超局面に対してマージンを適切に設定することを意図していたため理論的性能限界が明確であった。

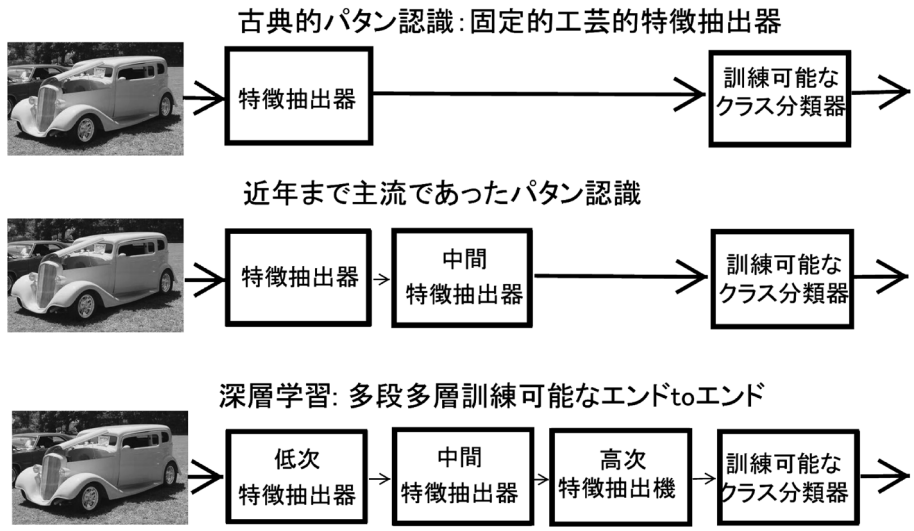

Figure 2.　Change of methods of image recognition.

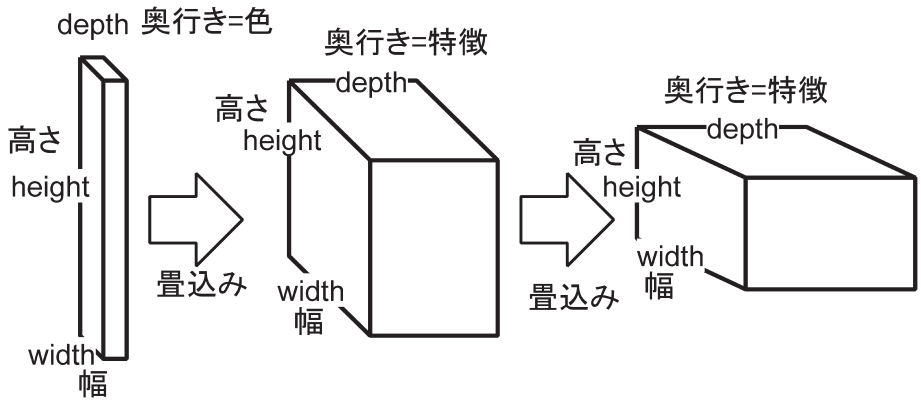

Figure 3.　Data transform of convolutional neural networks.

も職人技のチューニングを必要せず（end-to-end）に，性能向上が示されている。このことは，専門的な知識を必要とせずとも認識性能の向上が期待できることを意味している。これが昨今の流行の一因である。

画像認識においては入力画像を数値化して扱う。画像を構成する最小点は画素（pixel）と呼ばれ，縦横2次元上の一点である。白黒濃淡画像であれば各画素が濃淡値を表す一つの数値で表現される。画像データは幅（width）と高さ（height）に加えてRGB等の奥行き，あるいは深さ（depth）を持つ三次元データである（Figure 3）。

Figure 4に第1畳込み層で抽出されたフィルタの例を示した。Figure 4はDeCAFモデル（Donahue et al., 2013）によって視覚化されたCNN学習済の第1層のフィルタの例である（Caffe（Jia et al., 2014）に付属するサンプル画像[7]）。Figure 4では方位選択性を持つ視覚フィルタが上5行に，色選択性のフィルタが下5行に描かれている。Figure 4は入力画像を用いてCNNを訓練することで，視覚特徴検出器を獲得可能であることを示している。

各層における高さと幅を決めるためには，畳込みの計算に必要な窓（小領域）のサイズと窓のストライド（stride）をあらかじめ定めておく必要がある。窓のサイズとストライドを決めると次の層の高さと幅が定まる。Figure 5に畳込み演算の窓を例示した。窓とは生理学における受容野サイズに対応する。

畳込み演算とは入力情報の窓（小領域）と特徴検出器（核関数またはカーネルkernelと呼ばれる）とを結合係

[7] ファイルとしては00-classification.ipynb

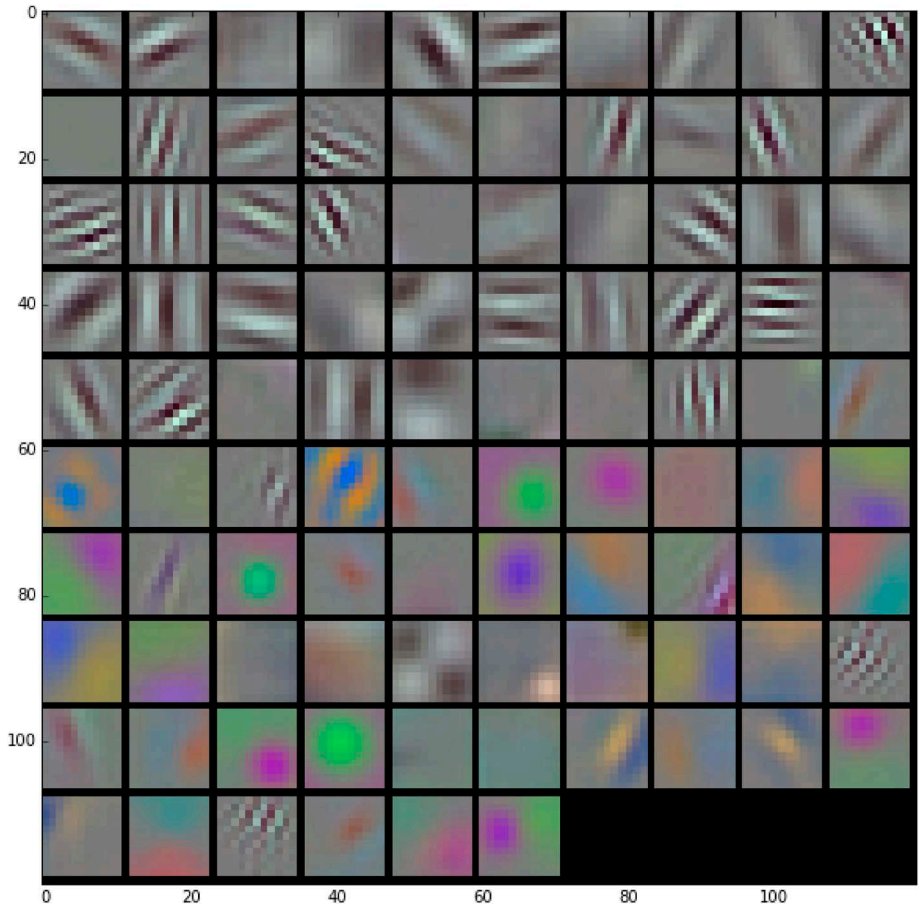

Figure 4. Sample pictures from the first convolutional layer in Caffe.

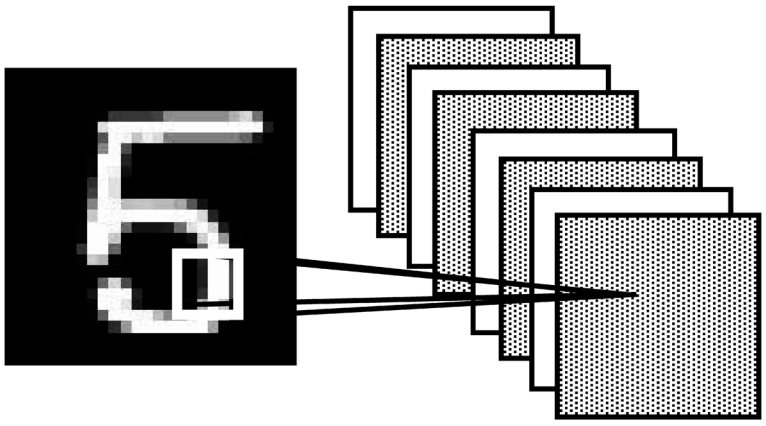

Figure 5. Data transform of convolutional neural networks.

数で重みづけて足し合わせる操作である。入力情報の各窓（小領域）に対して特徴検出器（カーネル）との重み付け和を計算する。

最小の窓サイズは1であり，最小のストライドも1である。簡単のため2次元画像ではなく1次元情報として考える。たとえば［1, 2, 3, 4, 5］と番号付けされた入力領

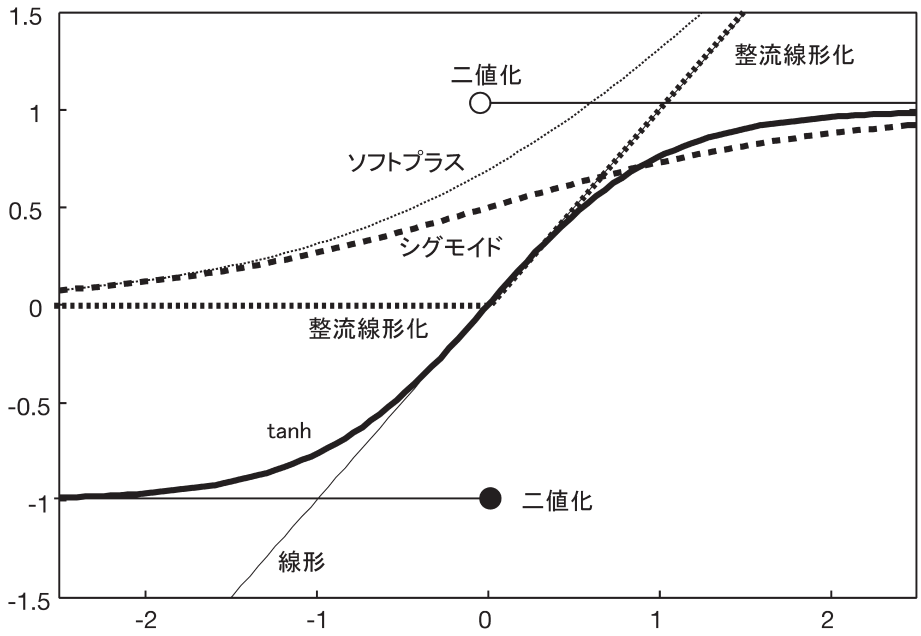

Figure 6.　Various activation functions.

域に対して窓サイズ3，ストライド1であれば，畳込み演算は［1, 2, 3］,［2, 3, 4］,［3, 4, 5］という3つの小領域でそれぞれ同じ計算を行うことになる。したがって入力情報が大きさ5のとき，この畳込み演算によって3つの情報を得ることとなる。

窓幅とストライド値によっては，窓の範囲が入力情報からはみ出る場合が考えられる。この場合，入力が存在しないものとしてゼロパディング（zero padding）することが行われる。上述した例ではストライドを3に設定すると［1, 2, 3］,［4, 5］となり最後の情報が不足する。この不足分をゼロで埋める。窓幅3に対してストライドも3であれば重複無く情報を抽出可能な畳込みとなる。しかし，実際の受容野が重なりを持つのと同じく，窓幅とストライドを同じにせず重複を許すことが行われている。

画像認識に用いられるCNNの特徴を挙げれば次の7点に集約できる。

1. 非線形活性化関数（non-linear activation functions）
2. 畳込み演算（convolutional operation）
3. プーリング処理（pooling）
4. データ拡張（Data augmentation）
5. バッチ正規化（batch normalization）
6. ショートカット（shortcut）
7. GPUの使用

以下ではそれぞれを概説する。

非線形活性化関数：　多層化ニューラルネットワークにおいて非線形活性化関数は決定的な役割を果たす。各層に配置されたニューロンの活動は，結合している下位層ニューロンの活性値を受けとって出力値を計算する多入力一出力の関数である。このときに採用される活性化関数は，Figure 1内に$f(\Sigma w_i x_i)$と表記してあるfに線形関数を用いると層ごとの演算が行列の積で表現されることになる。複数の行列の積はまとめて一つの行列の積として表現可能であるので多層化しても一つの行列の積の演算に等しくなり，多層化する意味がない。出力関数に非線形関数を用いることで多層化ニューラルネットワークの演算に意味を持たせることが可能となる。Figure 6に頻用される活性化関数を示した。線形関数とは入力値をそのまま出力するすなわち何もしない関数である（$y=x$）。シグモイド関数（$y=(1+\exp(-x))^{-1}$）は1980年代から用いられている（Rumelhart et al., 1986）。この関数の微分は$f'(x)=1-f(x)$と簡単なため頻用されてきた。整流線型（Rectified Linear: ReLU）関数（Krizhevsky, Sutskever, & Hinton, 2012a; Nair & Hinton, 2010）とは入力値が負であればゼロ0を出力し，正であれば，その値をそのまま出力する（$y=\max(0, x)$）。したがって正流線型関数と線形関数は，入出力が正の範囲で重なる。ハイパータンジェント関数（tanh）はアルファベットのS字状の曲線であり，最大値と最小値はそれぞれ＋1, －1となる。出力値の範囲はこの間に限定される。LeCun, Bottou, Orr, & Müller（1998）によれば，シグモイド関数に代わり

tanh（x）を使う方がバイアスが発生せず学習時の収束がよい。二値化関数とは出力が −1もしくは +1となる関数である（y=sign（x））。入力が負であれば −1となり，正であれば +1となる。二値化関数はハイパータンジェント関数を簡単にした簡略版とみなしうる。入力xが従来通り浮動小数点であれば今日のコンピュータでは32ビットもしくは64ビットで表現される。二値化により1ビットで表現可能である。すなわち必要なメモリ容量を32倍節約できる（Rastegariy, Ordonezy, Redmon, & Farhadiy, 2016）。性能が落ちなければ二値化してもよいことになる。

畳込み演算：　畳込み演算は，Hubel & Wiesel（1959, 1962, 1968）から得られた事実に基づく脳生理学の知見を数学的に定式化したニューラルネットワークモデルの特徴である。

生理学の知見に基づけば，網膜視細胞に届いた光信号は外側膝状体を経て第一次視覚野から，さらに高次の視覚情報処理過程へと伝達される。視覚情報処理に関与する領野間の結合は複雑ですべてが解明されたとは言いがたいが，階層構造をなしていることが知られている。すなわち，低次の階層で処理された信号は，高次情報処理を司る階層へと伝達される。このとき，高次視覚野から低次視覚野へのフィードバックや，領野内での結合，及び，信号の転送時期や伝達方法などを簡略化のため無視することにすれば，低次層に与えられた信号は，逐次階層を上がるごとに逐次複雑な情報処理が行われると考えられる。

例えば第一次視覚野では線分に応じるニューロンが選好方位ごとに並んでいる（方位選択性）。したがって視覚情報処理の第一段階は線分検出器とみなすことができる。線分検出器や色情報処理に特化した細胞は特徴検出器として振る舞う。これら特徴検出器は，より低次の領野の細胞からの入力範囲が定まっており，受容野と呼ばれる。高次視覚野は低次視覚野の特徴検出器からの情報を受け取り，さらに高次の特徴を表現する複雑な細胞が存在する。

上記のような生理学的事実を，特徴検出器とその受容野と考えると，ある受容野の範囲内に特定の特徴が存在するか否かを検出することが各視覚野の機能であると考えることが可能である。CNNにおける畳込み演算とは受容野内に任意の特徴の存在の有無を検出することに相当する。このとき，受容野の大きさと特徴の種類を決めないと計算できない。

現在のCNNは，受容野サイズと特徴検出機の個数とは，あらかじめ研究者により固定される場合が多い。一方，下位層から上位層のニューロンへの結合係数と特徴検出器がどのような性質を持つかは学習により決定される。

すなわちCNNにおいては，ニューラルネットワークの構造である，層数，各層のニューロン数，受容野のサイズ，個数，受容野の配置間隔，データの種類，学習のための計算手法は固定される。一方，ニューロン間の結合係数と特徴検出器がどのような特徴を表現するようになるかは，データによって定まる。換言すれば，ネットワークが晒される環境によって定まると仮定する。

プーリング：　上述の畳込み演算によってネットワークの結合係数と特徴検出器が学習される。入力層に提示された情報は畳込み演算によって処理され直上層へと伝達される。上位層は下位層の特徴検出器の出力を入力として扱って同様の計算を行う。ここで再び大脳生理学の知見に基づいて特徴検出器の出力を間引く処理を行う。すなわち視覚野においては複雑細胞の受容野幅は広い。受容野内に刺激が存在すればそのニューロンは発火する。

このことと対応してCNNにおいては，あらかじめ定められた範囲の最大値で出力し，その他の情報を捨てる。この処理をマックスプーリング（max pooling）と呼ぶ。場合によっては平均値プーリングなど他の手法も用いられる。ただし，最近の実装（Radford, Metz, & Chintala, 2016）ではプーリング層をストライド付き畳込み演算で代用することが行われている。

データ拡張：　データ拡張とは，データ数を増やすことを言う。例えば，ある物品は画像の中でどの位置に提示されても同じ認識に至るようにするため，画像を上下左右に反転，移動，拡大縮小などを施したデータを入力データに加えることを行う。データ拡張によりデータ数が増加し，安定した認識に至るようになる。データ拡張は限られた訓練データ数から性能のよい認識モデルを作成するために，画像データに変動を加えることでデータ数を増やすことを指す。データ拡張によりモデルの性能向上を意図して行われる。入力画像中に，ある人物の顔が写っているとすれば，この顔が画像中のどこに表示されても，同じ顔であると認識しなければならないので，上述のデータ拡張を行い同一画像から複数の画像を生成する。このデータ拡張によって認識性能，汎化性能が向上することが知られている。いわゆるビッグデータの恩恵による技法である。

バッチ正規化：　ニューラルネットワークモデルが従来モデルと異なる特徴の一つは，Figure 2にも示したとおり，特別な前処理を必要とせず，エンドツーエンドであることである。2015年バッチ正規化（Ioffe & Szegedy, 2015）が提唱される以前では，画像認識における唯一の前処理とは各画像から平均値を引くことだけであった。実際4章で紹介するフレームワークでも画像認識では，

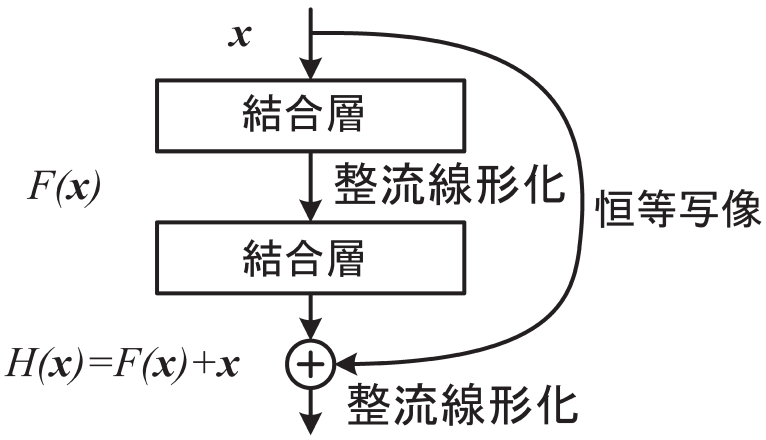

Figure 7. A shotcut of the ResNet.

あらかじめ平均を引く前処理が要求される（例えばCaffeなど）。ところが，バッチ正規化は各層ごとに各訓練ミニバッチごとに，結合係数の重みを正規化する。バッチ正規化は内部共分散シフト（internal covariance shift）を抑制することを意図している。内部共分散シフトとは，訓練中にネットワークのパラメータが変化することで，ネットワークの活性値の分布が変化することと定義されている。各ミニバッチごとに以下の変換を行う。

$$\hat{x}^{(k)} = \frac{x^{(k)} - E[x^{(k)}]}{\sqrt{Var[x^{(k)}]}}, \quad (1)$$

ここで各層ごとに$k \in 1\cdots, K$個の入力画像に対する操作を意味し，ミニバッチごとに平均を引いて標準偏差で除すことを意味している。これにより入力が正規化されるため学習回数の現象と精度の上昇が見られた[8]。

バッチ正規化とは，各層の出力値を分散が一定になるように変換することである（Ioffe & Szegedy, 2015）。この処理は，検査や調査データ処理において，平均を引いて，標準偏差で割ることで平均0，分散1となるようにすることに等しい。これにより外れ値などの影響が小さくなり安定した認識に至る。

ショートカット： ショートカットとは，直下層からの出力に加えて，さらに下位層の出力を加えることを言う（He et al., 2015a; Srivastava, Greff, & Schmidhuber, 2015）。Figure 7にショートカットを図示した。ショートカットでは学習は直下層との結合係数に対してのみ行われ，より下位層からの出力は単に加えられるだけである。したがって学習時の付加的な条件や計算量の増加は発生しない。

より下位層からの出力は固定されているので，学習は固定された出力の残りを学習することとなる。これゆえ残渣ネット（residual net）とも呼ばれる。ショートカットを繰り返した多層ニューラルネットワークの例をFigure S1に示した。残渣ネットは人間の認識性能を凌駕した。Figure S1中の数字は，たとえば7×7は受容野サイズを表している，続いて各層の役割が示され（畳込み，平均プーリング，完全結合），特徴数（奥行き）が記されている。その後/2のような表記は受容野間のストライドを示している。歴史的にはFahlman & Lebiere（1990）が提案したカスケードコリレーション（cascade correlation）モデルが，一般的で適応的なモデルである。現在ではMareschal et al.（2007）; Mareschal, Sirois, Westermann, & Johnson（2007）の形で継承されているものの，CNNへの応用はなされていない。

GPU: GPUとは画面描画を高速化するためにPCに搭載するグラフィック処理ボード（graphical processing units）のことである。元来ゲーム用に開発されたGPU

[8] 本稿ではミニバッチ（mini batch）の説明をしていないが，確率的勾配降下法（SGD: stochastic gradient descent method）が提案している学習回数の改善手法である（Bottou, 2010; Bottou & Bousquet, 2007）。学習則を適用してパラメータの更新を全学習データに対して加算して一度だけパラメータ更新を行うことをバッチ更新とよび，各データごとに逐一パラメータ更新を行うことをオンライン更新とよぶ。一括して学習を行うバッチ処理は局所解に陥る可能性があるが，オンライン更新では収束までに時間を要する。これに対して，バッチ更新とオンライン更新との中間であるSGDでは，データをミニバッチに小分割し，分割されたミニバッチに対してパラメータの更新を行う。SGDは劇的に学習回数を減じることから，近年ではSGDを用いることが前提となっている（浅川，2016）。

はニューラルネットワークの学習に用いられる。ニューラルネットワークの計算は，処理が単純な繰り返しである，単一命令多データ（SIMD: single instruction multi data）処理が特徴である。この特徴から画像描画演算用に開発されたGPUが用いられるようになった。2009年前後からニューラルネットワーク研究に用いられるようになり，活用できるデータサイズが20倍あるいは30倍以上となった。以前ニューラルネットワークモデルはオモチャであり，実問題を解くことができないとの批判を聞いた。しかし，データ規模が今日の規模になると，今度は扱うデータが大きくなっただけだろうと批判を受けるようになった。ではどうすればよいのか途方に暮れる思いだが，どちらの批判もモデルを実際に扱ったことがない人間の批判である。

残渣ネット（He et al., 2015a）とハイウェイネットワーク（Highway network, Srivastava et al., 2015）で関心領域を精度よく解けるのは，CNNの各層で計算される情報の中に，視覚対象の特徴情報だけでなく，対象の位置情報も処理されているのではないかという予想が成りたつ。すなわちCNNにおいては，物体認識に必要な特徴情報と位置情報が抽出可能なことを示していると考えられる。

3.2 CNNの代表的なモデル

ここでは代表的CNNモデルを概説する。

ネオコグニトロン： 現代的な意味でのCNNは福島のネオコグニトロン（NeoCognitron）（福島，1976, 1983; Fukushima, 1980, 1987; Fukushima & Miyake, 1982）が起源である。Figure S2に概略図を示した。ネオコグニトロンはHubelとWieselの生理学実験（Hubel & Wiesel, 1959, 1962, 1968）から得られた事実に基づき，視覚認識を行うモデルである。S層とC層とはそれぞれ単純細胞と複雑細胞から構成される層を示している。生理学的事実に基づき，ネオコグニトロンのニューロンの受容野は層を登るに従って大きくなる。同時に，受容野内に照射された刺激は位置不変性を持っている。すなわち回転，拡大縮小，移動などアフィン変換に対して頑健である。

LeNet: LeCun et al.（1998）は手書き数字認識モデルLeNetを提案した。LeNetは畳込み演算とサブサンプリング（sub sampling）の繰り返しである。サブサンプリングは後にプーリングに置き換えられている。

Figure S3にLeNetを示した。Figure S3中の数字は画像，特徴地図の高さ，幅，奥行き，などを示している。Figure S3最左の入力画像の大きさは高さ，幅すなわち縦横の画素数が32 × 32の濃淡画像（ゆえに奥行きが1）である。続くC1特徴地図層（第1層）は高さと幅が28×28の特徴地図であり，特徴数は6（したがって奥行きは6）である。次のP1特徴地図層は，C1層をプーリングした層であり，高さ14,幅14，奥行き6（したがって特徴数は6）である。同じ処理がもう一度繰り返され，畳込みとプーリングが行われる。さらに10種類の手書き文字を識別するために全結合層が2つ，C5層（ニューロン数120), F6層（ニューロン数84）が用意され最終層は10ニューロンである。これら10個のニューロンはそれぞれ0から9までの10種類の手書き数字に対応している。

AlexNet: 2012年の大規模画像認識コンテストImageNet[9]において当時SOTA（state of the art）であったSVMを10%以上凌駕したモデルがAlexNet（Krizhevsky et al., 2012a）である。第一著者の名前からの呼び名である。AlexNetの特徴としては，畳込み，ドロップアウト，データ拡張，GPUの利用，局所正規化，が挙げられる。ImageNetの分類課題は，画像を1000種のカテゴリに分類することが求められる。このとき上位5候補を出力して，この5カテゴリの中に正解が含まれているか否かで性能を競う。SVMでは上位5候補の誤判別率が約26%であった。一方AlexNetは16%を達成した。ネットワークの構成はLeNetと同様であるが規模が大きい。

ZFNet: ZFNet（Zeiler & Fergus, 2014）はImageNet2013の優勝モデルである。開発者のZeilerとFergus両名の頭文字を取ってZFNetと呼ばれる。ZFNetはAlexNetの畳込み層のニューロン数を調整した進展型とみなしうる。

GoogLeNet: GoogLeNet（Szegedy et al., 2015）はImageNet2014の優勝モデルである。Googleが開発しLeNetに敬意を表してGoogLeNetと表記される。GoogLeNetは複数のカーネルを並列に用いたインセプション（inception）モジュールを基本単位とする。インセプションモジュール内では結合が構造化されているので，全結合を考えるより総結合数が少なくて済む。実際AlexNetでは総結合数（したがって推定すべきパラメータ数）が約600万であったが，GoogLeNetでは40万であった。

Figure S4にインセプションモジュールを連結したGoogLeNetの全体像を示した。Figure S4中の楕円で囲った部分がインセプションモジュールである。楕円の上にモジュールの個数が描かれている。情報は左から右へ向かって流れるフィードフォワードニューラルネットワークである。最左が入力層であり，最右が出力層である。

Figure S5は，Figure S4に9つの楕円で囲まれた領域を

[9] http://image-net.org/challenges/LSVRC/

90度回転させ，拡大した図である。層間畳込みをすべて行うのではなく，インセプションモジュールに示された各カーネルサイズについての畳込みを行うこととなる。これにより総結合数を抑制する効果も期待できる。加えて，Figure S5の最左にカーネルサイズが1 × 1の畳込み演算が含まれている。1 × 1の大きさで畳込むのであるから，結局何もしないで上位層に情報をそのまま伝達することに等しい。CNNに対する批判として，層を飛び越える結合が無いとの批判がある。しかし，インセプションモジュールの最左モジュールは下位層の情報を上位層へ伝達している。これにより擬似的に層を飛び越える意味を含意しているものと考えられる。

VGG（Oxford net）： VGGネットも2014年のモデルである（Simonyan & Zisserman, 2015）。CNNでは多層にすることで成績が向上すると考えてよいのか，絶えず疑問が呈されてきた。VGGネットは均質な構造の繰り返しである点が評価される。VGGネットは，畳込み層，マックスプーリング層を5回繰り返し，その後全結合層を経てソフトマックス層で出力に至る構造であった。畳込み層のカーネルサイズは3 × 3であり，プーリングは2 × 2で一貫していた。GoogLeNetの成績には僅差で及ばなかったが，VGGネットは簡潔な構造のため転移学習に用いられる場合が多い。VGGネットで学習したパラメータを用いて，他の画像認識課題や，応用的課題を行なう研究にVGGネットが用いられる。VGGネットを用いて訓練した，学習済みパラメータファイル（Jia et al., 2014, Caffeモデル）が無料で公開されている。ただしGoogLeNetに比べると推定すべきパラメータ数が多くなる（1400万。先述のGoogLeNetは40万である）。

SPPネット： SPPネットも2014年のモデルである（He, Zhang, Ren, & Sun, 2015b）。GoogLeNet, VGGネットと同等の認識成績を示した。CNNによる画像認識では畳込みとプーリング処理との後に全結合層を連結し最終的な認識へと至る。SPPネットではCNNと全結合層との間に空間ピラミッドプーリング層（spatial pyramidal pooling layer）が挿入された。空間ピラミッドプーリングとは，解像度の異なるプーリングを連結して全結合層へと伝達することである。全領域に渡る1個のプーリング，4分割（2 × 2）のプーリング，16分割（4 × 4）のプーリングを連結する。したがって解像度の異なる情報が全結合層へ転送されることになる。これは入力画像にはさまざまな大きさの物体がさまざまな解像度で存在するため，異なる解像度でプーリングした情報を全て全結合層へ送る方が認識に有利となると考えられるためである。この考え方は後述するRCNN（regions with convolutional neural networks）へ繋がるので，意味がある。

残渣ネット： 残渣ネットについては既述したが，その他にFaster-RCNN（Ren, He, Girshick, & Sun, 2015）の手法を用いて関心領域の切り出しを行っている。残渣ネットはさらにバッチ正則化（Ioffe & Szegedy, 2015）も取り入れている。バッチ正則化を用いたためドロップアウト（Hinton, Srivastava, Krizhevsky, Sutskever, & Salakhutdinov, 2012）は用いられなかった。

3.3 R-CNN

ImageNetコンテストにはクラス分類課題とロケーション課題とが存在する。クラス分類課題とは画像データが与えられたとき，その画像が何であるかを問う課題である。一方，与えられた画像中のどの位置に物体が存在するかを問う課題をロケーション課題という。データ拡張などを用いて訓練してもCNNでは画像の判別は行えても，画像上で物体が占める位置を問うことは難しいと考えられてきた。大規模画像認識コンテストの判別課題である1000種のカテゴリー分類を越えて，CNNにて一般物体認識を行うためにはCNNがロケーション課題を解けなければならない。この関心領域の切り出しには，画像認識の従来手法が用いられてきた（Uijlings, Sande, Gevers, & Smeulders, 2013）。Figure 8はImageNetで用いられた画像である。画像からペリカンとカエルの矩形領域を切りださねばならない。

一般画像認識の難しさは，対象が部分的に他の対象に隠蔽されていたり，画像中に存在する物体の個数についての事前知識を仮定できないことなどが挙げられる。

Girshick, Donahue, Darrell, & Malik（2014）はCNNに領域切り出しを行うことを提案した。具体的には全結合層の直下の畳込み層の出力を領域切り出しに用いることにした。Figure S6にGirshick et al.（2014）の用いた手法を示した。Figure S6では左から右へと処理が流れる。前半部分ではCNNに関心領域の矩形領域を学習させた。矩形領域を切り出すCNNに対して後半は，切りだされた各画像小領域に対して再びCNNを実行して各領域ごとにクラス分類を行わせた。この手法をR-CNNと呼ぶ。

R-CNNはCNNによる関心領域切り出しと，切りだした領域に対してCNNを行う2段階モデルである。各段階はCNNを行うのであるから，CNNを一度で済ませれば速度向上が期待できる。Girshick（2015）は高速R-CNN（Fast-RCNN）を提案した。Figure S 7にFast-RCNNの概要を示した。CNNの畳込み層の最上位層の出力をプーリングし，そのプーリングされた出力に対して矩形領域回帰とクラス分類問題を同時に解くことを行った。

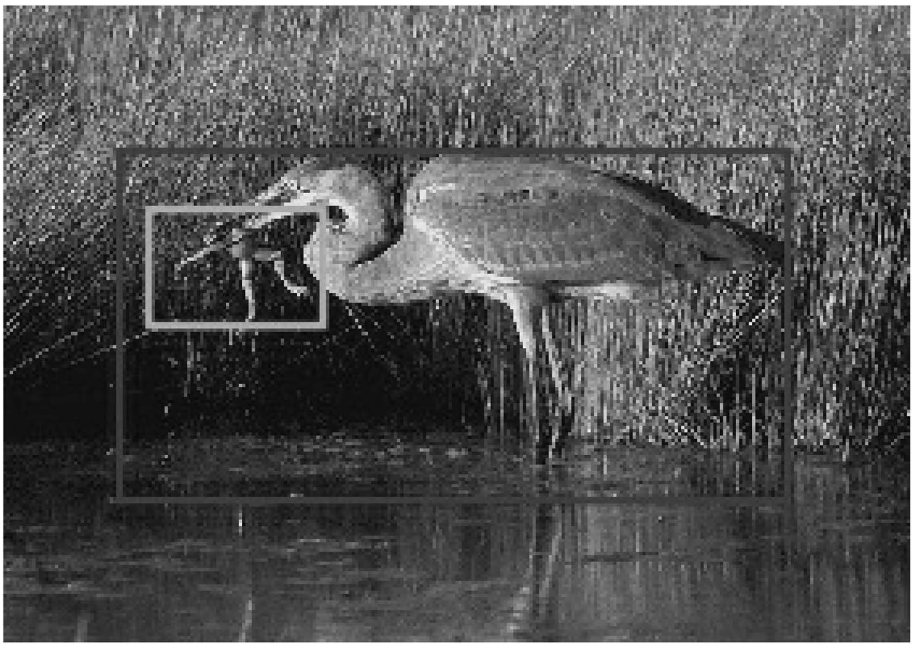

Figure 8. Object recognition and bouding boxes.

最終出力は矩形領域候補とカテゴリ分類候補の2つになる。このため目標関数も両者の出力を勘案せねばならないが，単純に両者の和でよいようである。

Ren et al.（2015）はさらに高速化したR-CNNを提案した。ここではFaster-RCNNと呼ぶ。CNNでは畳込み層が複数回繰り返され，その上位層に2層の全結合層を置くことが多い。Figure S S3, Figure S S4も同じネットワーク構造をしている。畳込み層の中の最上位層と全結合層との間にFigure S S7にも見られる関心領域特徴ベクトル層が加えられた。関心領域特徴ベクトル層の出力を用いて，領域提案ネットワーク（RPN: Region Proposal Network）が構成された。RPNに表現された特徴ベクトルから，物体性得点を出力する層とその時の関心領域のスケールとアスペクト比の領域提案境界位置を出力する層が作られた。R-CNNは，入力画像データ上のどの位置にどの物体が存在するのかを解く。関心領域特徴ベクトル層に現れた表現に対して，矩形領域を設定し，その領域を窓とみなす。関心領域特徴ベクトル層の座標に基づいて逐次窓の位置をスライドさせ，どの窓領域にどの対象が存在するかの回帰問題を解く。これをスライディングウィンドウ（sliding windows）と呼んだ。スライディングウィンドウの様子をFigure 9に示した。候補領域の矩形領域回帰はスライディングウインドウ上で8種類に限定されていた。RPNにより矩形領域回帰問題の高速化が期待できる。RPNはCNNであり，かつ，エンドツーエンドで訓練可能である。Faster-RCNNは実時間で，関心領域の切り出しと切り出した領域の物体認識を行うことが可能である（Ren et al., 2015）。残渣ネット（He et al., 2015a）はFast–RCNNの物体検出ネットワークとRPNを用いて高速化を実現した。

R-CNNの意義として，多層CNNが抽出している情報には，特徴情報だけでなく，物体の位置情報も含まれていると見做せることである。視覚認識に必要な情報は物体の特徴と位置共にCNNが抽出する情報に含まれていたことを意味する。

画像のクラス分類と一般画像認識との間には乖離が存在した。この乖離を埋めることが長らく画像認識研究の目標であったと言える。この目標達成に向けて多くの研究がなされてきた。CNNに基づくR-CNNを用いれば，物体の特徴と位置とを同時に処理する系が可能であることが示された。R-CNNの意義はこの点に認められると考えられる。

4. 実装の公開

近年では，査読論文の投稿，審査，修正，再審査，印刷，というプロセスを経ず誰でもがダウンロードできるarXivなどの論文リポジトリで論文を公開し，同時にプロジェクトページを立ち上げてそのURLをアナウンス

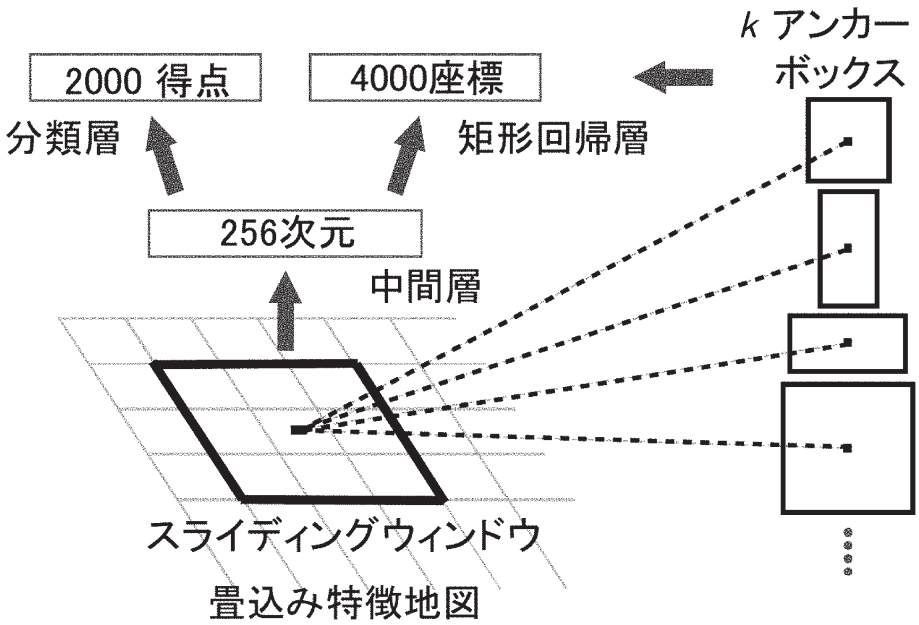

Figure 9.　Faster–RCNN.

する場合が多くなった。このとき，同時に，結果の再現可能性を検証できるようにプログラムのソースコードをGitHub上に公開している。査読プロセスがないために従来の学術権威は存在しないが，ソースコードが動作することで性能が保証されている。進展が早くなるため，このような手法が主流である。

加えて無料で入手可能なフレームワークの存在により，たとえプログラムを書くことができなくても動作は確認できる。このようなフレームワークの中で一般に用いられるものとしては：Caffe（Jia et al., 2014）[10], Chainer（Tokui, Oono, Hido, & Clayton, 2015）[11], TensorFlow[12], Theano（Bastien et al., 2012）[13], Torch[14], CNTK[15] ,などが挙げられる。CaffeはC＋＋で書かれているが，必要な機能を呼び出すためにパラメータをprotobufで記述する。加えてPython APIを持つのでC＋＋を知らなくとも動作確認が可能である。Chainer, TensorFlow, TheanoはPythonベースである。Torchはluaと呼ばれる言語で書かれている。CNTKはマイクロソフトが公開したフレームワークである。これらフレームワークの特徴としては画像データをテンソルに変換し，画像の小領域を切り出して列ベクトルにするなどの（1）行列操作，目標関数の（2）自動微分，（3）GPUへの対応，などの機能が揃っている。浅川（印刷中）はCaffe, Chainer, Theano, TensorFlowを紹介している．

[10] http://caffe.berkeleyvision.org/
[11] http://chainer.org/
[12] https://www.tensorflow.org/
[13] http://deeplearning.net/software/theano/
[14] http://torch.ch/
[15] http://www.cntk.ai/

5. おわりに

最後に日本語で書かれた文献を列挙しておく。人工知能学会による特集は最近の動向を網羅している（麻生ほか，2015）。岡谷（2015）は教科書として有益であろう。浅川（2015a, 2015b, 2016），石橋（2015）はCaffeを詳述している。

冒頭に挙げたCrickの批判は今なお有効であるものの，応用研究がこれだけ花開いた意味は最早無視できない。たとえ宇宙人の工学であっても，知的情報処理を行う機械は心理学研究に携わるものに対して，深い洞察を与えてくれるにちがいない。

引用文献

浅川伸一（2015a）．ディープラーニング，ビッグデータ，機械学習あるいはその心理学　新曜社
浅川伸一（2015b）．ニューラルネットワーク　榊原洋一・米田英嗣（編）　発達科学ハンドブック　Vol. 8（pp. 94–104）　新曜社
浅川伸一（2016）．Pythonで実践する深層学習　コロナ社

麻生秀樹・安田宗樹・前田新一・岡野原大輔・岡谷貴之・久保陽太郎・ボレガラダヌシカ（2015）．深層学習　近代科学社

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, *47*, 235–256.

Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I. J., Bergeron, A., Bouchard, N., & Bengio, Y. (2012). Theano: new features and speed improvements. *Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop*.

Boole, G. L. L. D. (1854). An investigation of the law of thought. London: Walton and Maberly.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Y. Lechevallier & G. Saporta (Eds.), *Proceedings of the 19th International Conference on Computational Statistics* (*COMPSTAT2010*) (pp. 177–187). Paris: Springer.

Bottou, L., & Bousquet, O. (2007). The tradeoffs of large scale learning. In *Advances in neural information processing systems* (Vol. 20). Cambridge: MIT Press.

Crick, F. (1989). The recent excitement about neural network. *Nature*, *337*, 129–132.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2013). DeCAF: A deep convolutional activation feature for generic visual recognition. *arXiv:1310.1531*.

Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture. In D. Touretzky (Ed.), *Advances in neural information processing systems* (Vol. 2, pp. 524–532). San Mateo, CA: Morgan-Kaufman.

福島邦彦（1976）．視覚の生理とバイオニクス　電子通信学会

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*, 193–202.

福島邦彦（1983）．ネオコグニトロンによるパターン認識　トリケップス

Fukushima, K. (1987). A neural network model for selective attention in visual pattern recognition and associative recall. *Applied Optics*, *26*, 4985–4992.

Fukushima, K., & Miyake, S. (1982). Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognition*, *15*, 455–469.

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural Computation*, *12*, 2451–2471.

Girshick, R. (2015). Fast R-CNN. *arXiv:1504.08083*.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of Computer Vision and Pattern Recognition Conference* (pp. 580–587).

He, K., Zhang, X., Ren, S., & Sun, J. (2015a). Deep residual learning for image recognition. *arXiv:1512.033835*.

He, K., Zhang, X., Ren, S., & Sun, J. (2015b). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1–1.

Hebb, D. O. (1949). *Organization of behavior*. New York: Lawrence Erlbaum.
（鹿取廣人・金城辰夫・鈴木光太郎・鳥居修晃・渡邉正孝（訳）(2011). 行動の機構―脳メカニズムから心理学へ―　岩波書店）

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing coadaptation of feature detectors. *The Computing Research Repository (CoRR), abs/1207.0580*.

Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). Gradient flow in recurrent nets the difficulty of learning long-term dependencies. In S. C. Kremer & J. F. Kolen (Eds.), *A field guide to dynamical recurrent neural networks*. Hoboken NJ, IEEE press.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*, 2554–2558.

Hopfield, J. J., & Tank, D. W. (1985). "Neural" computation of decisions in optimization problems. *Biological Cybernetics*, *52*, 141–152.

Hopfield, J. J., & Tank, D. W. (1986). Computing with neural circuits: A model. *Science*, *233*, 625–633.

Hubel, D., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, *148*, 574–591.

Hubel, D., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, *160*, 106–154.

Hubel, D., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, *195*, 215–243.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167*.

石橋崇司（2015）. Caffeをはじめよう　オライリー・ジャパン

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*.

Kohonen, T. (1985). Self-organizing maps. Berlin: Springer-Verlag.

Kohonen, T. (1996). Self-organiziing maps (2nd ed). Berlin: Springer-Verlag.
（コホネンT.（2005）．自己組織化マップ. 徳高平蔵・大藪　又茂・堀尾恵一・藤村喜久郎（監修）丸善出版.）

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012a). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25*. Montréal, Canada.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012b). ImageNet classification with deep convolutional neural networks. In P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems*. Lake Tahoe, Nevada, USA.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*, 2278–2324.

LeCun, Y., Bottou, L., Orr, G. B., & Müller, K. (1998). *Efficient backprops*. Berlin Heidelberg: Springer.

Mareschal, D., Johnson, M. H., Sirois, S., Spratling, M. W., Thomas, M. S. C., & Westermann, G. (2007). *Neuroconstructivism* (Vol. 1). UK: Oxford University Press.

Mareschal, D., Sirois, S., Westermann, G., & Johnson, M. H. (2007). *Neuroconstructivism* (Vol. 2). UK: Oxford University Press.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*, 115–133.

Minsky, M., & Papert, S. (1988). *Perceptrons*（Expanded Edition ed. Cambridge, MA: MIT Press.
（ミンスキー，M., パパート，S. 中野　馨・坂口　豊（訳）（1993）．パーセプトロン　パーソナルメディア）

Mnih, V., Kavukchuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G. … Hassbis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*, 529–533.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In J. Fürnkranz & T. Joachims (Eds.), ~~*In*~~ *Proceedings the 27th International Conference on Machine Learning (ICML)*. Haifa, Israel: Omnipress.

Neumann, J. von. (1958). *The computer and brain*. New Haven/London: Yale Univesity Press.
（ジョン・フォン・ノイマン . 飯島泰蔵・猪俣修二・熊田　衛（訳）電子計算機と頭脳　丸善）

Oja, E., & Kaski, A. (1999). *Kohonen maps*. Amsterdam, Netherland: Elsevier.

岡谷貴之（2015）．深層学習　講談社

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434*.

Rastegariy, M., Ordonezy, V., Redmon, J., & Farhadiy, A. (2016). XNOR-net: Imagenet classification using binary convolutional neural networks. *arXiv:1603.05279*.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *arXiv:1504.01497*.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386–408.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed porcessing: Explorations in the microstructures of cognition* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, **61**, 85–117.

Silver, D., Huang, A., Arthur Guez, C. J. M. an, Sifre, L., Driessche, G. van den, Schrittwieser, J., … Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*, 484–492.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Y. Bengio & Y. LeCun (Eds.), *Proceedings of the International Conference on Learning Representations (ICLR)*. San Diego, CA, USA.

Srivastava, R. K., Greff, K., & Schmidhuber, J. (2015). Training very deep networks. *arXiv:1507.06228*.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning*. Cambridge, MA: MIT Press.

Szegedy, C., Liu, W., Jia, Y., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA.

Tokui, S., Oono, K., Hido, S., & Clayton, J. (2015). Chainer: a next-generation open source framework for deep learning. *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*. Montreal, Canada.

Uijlings, J. R. R., Sande, K. E. A. van de, Gevers, T., & Smeulders, A. W. M. (2013). Selective search for object recognition. *International Journal of Computer Vision*, *104*, 154–171.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

Vapnik, V. N. (1998). *Statistical learning theory*. Hoboken NJ, John Wiley & Sons.

Vapnik, V. N. (1999). An overview of statistical learning theory. IEEE TRANSACTIONS ON NEURAL NETWORKS *10*, 988–999.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *Proceedings of the Computer Vision (ECCV)* (pp. 818–833). Zurich, Switzerland: Springer-Verlag.