解 説

深層学習について

On Deep Learning

岡 谷 貴 之* *東北大学大学院情報科学研究科
Takayuki Okatani* *Graduate School of Information Sciences, Tohoku University

1. はじめに

深層学習（ディープラーニング）は，多層（＝ディープ）のニューラルネット（以降省略して「ディープネット」と記す）を用いた機械学習の方法である．近年，音声認識，画像認識，自然言語処理や製薬（化合物反応予測）など，いくつかの問題で，このディープネットを用いた方法がそれ以前の方法を圧倒する性能を示し，研究開発は活況を呈している．本稿では，深層学習の今に至る研究の経緯と現在の状況について解説する．筆者はコンピュータビジョンを専門とするので，画像認識への応用を中心に述べるが，中身は（画像応用で特異的に高い有用性を持つ畳込みネットを除けば）基本的に音声認識やその他の分野への応用にも当てはまる．

ディープネットがこのように成功している理由は，特徴量を学習する能力にある．音声や画像などを入力とする認識問題（例えば発話内容の認識や画像中の物体認識）は，入力となる音声信号や画像から，まず認識対象を表現する特徴を取り出すステップと，取り出した特徴を分類するステップの二つに分けることができる（**図 1**）．後半の分類のステップは，サポートベクターマシンといった，90 年代に力強く発展した機械学習の方法を使うことで解決できる．問題となるのは，入力からどのような特徴を抽出すればよいかである．

実際，画像認識の分野では，これまでに，手書き数字の認識や顔の検出・認識などの問題を解決してきた一方で，未解決の問題（例えば画像 1 枚からそこに写る物体の名前を答える問題など）もまだ数多く存在している．これらが未解決だったのは，何を特徴として抽出すべきかが不明だったためであると言える．過去，特徴抽出の処理は研究者が手で設計するのが普通であり，それに成功した（性能が出せた）問題は解決され，そうなっていないものが未解決な問題として残されていたと言える．

原稿受付 2014 年 12 月 3 日
キーワード：Deep Learning, Layerwise Pretraining, Convolutional Neural Networks
*〒 980–8579 仙台市青葉区荒巻字青葉 6–6–01
*Aoba–ku, Sendai–shi, Miyagi

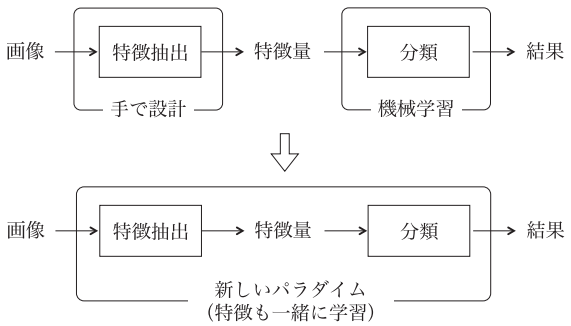

図 1 画像・音声認識研究のパラダイム変化

これに対しディープネットは，特徴そのものをデータから学習する（図 1）．この方法は特徴学習や表現学習とも呼ばれ，手で設計できる範囲を超えた複雑な特徴量を，学習を通じて獲得することを可能にする．この方法は実際，それまで未解決だったいくつかの問題を解決しつつある．この結果，画像認識をはじめとする人工知能の研究は，一つのパラダイムの変化を迎えていると言える．

2. ニューラルネットの基礎

2.1 ネットワークの構造

深層学習といっても，核にある方法は従来からのニューラルネットのそれとまったく同じである．復習を兼ねてここでは，ニューラルネットの基礎を簡単にまとめることにする．画像認識をはじめとする多くの問題で，もっともよく使われているフィードフォワード型のネットワークを対象とする．

最も基本となるのは，**図 2** (a) のように入力 x を受け取り出力 y を返すユニットである．両者の関係は活性化関数と呼ばれる次の関数

$$y = f(x) \tag{1}$$

が与える．活性化関数 f には，昔から使われているシグモイド関数 $f(x) = 1/(1+e^{-x})$ に代わって，$f(x) = \max(x, 0)$ という関数（rectified linear と呼ぶ）が近年，使われるようになった [1]．この関数は計算量低減に効果があるだけでなく，後述の学習最適化がより早く，よりうまく行えるこ

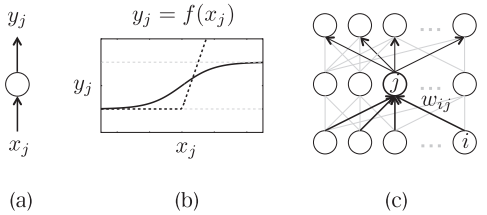

図 2 ユニット，活性化関数，および層間結合

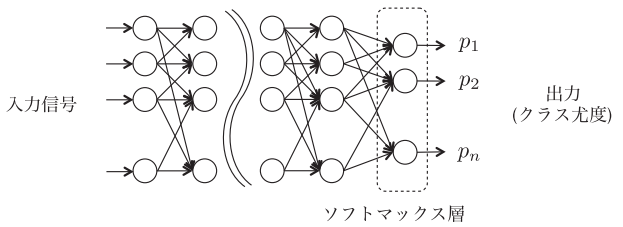

図 3 クラス分類ための多層ニューラルネット．入力信号が n 個のカテゴリのどれらしいかを出力層の n 個のユニットから出力する

とが経験的に分かっており，一般化している．

図 2 (c) のように，このようなユニットを層状に並べて，層間でのみユニット間に結合を与えてネットワークを作る．図中の j 番目のユニットには，直前の層のユニット複数の出力 $y_i (i = 1, \ldots, m)$ にそれぞれ異なる重み w_{ij} をかけて加算した後，バイアスを足した

$$x_j = b_j + \sum_{i=1}^{m} y_i w_{ij} \qquad (2)$$

が入力される．この x_j を活性化関数に入れ，得られる出力 $y_j = f(x_j)$ が，同図 (c) のように上の層の各ユニットに等しく伝えられる．今考えているフィードフォワード型のネットワークでは，**図 3** のようにこのような層を適当な数並べる．最も左の層に入力された信号は，順に層を伝播して行き，最後の層から最終的な結果を出力する．このような多層のニューラルネット一つは，何らかの非線形関数一つを表現することとなる．

2.2 出力層の設計と誤差関数

ニューラルネットの出力層は，問題に応じて決まった形をとる．ここでは音声・画像認識で最も一般的なクラス分類を考える．この場合，出力層には目的のクラス数 n と同数のユニットを配置し，これらのユニットは受け取る入力 $x_j (j = 1, \ldots, n)$ から $p_j = e^{x_j} / \sum_{k=1}^{n} e^{x_k}$ を計算する（ソフトマックス関数と呼ばれる）．$p_1, \ldots, p_n$ を n 個のクラスそれぞれに対する尤度（確からしさ）と見なす．実際に入力を分類するには，p_j が最大値をとるクラス j に分類する．

このように決めたニューラルネットが，入力データに対し望みのクラスを返すように学習を行う．具体的には，多数の学習サンプルを使って，ネットワークの重み w_{ij} およびバイアス b_i を調節する．画像認識などの複雑なタスクでもこれは同じであって，画像と物体のクラスラベルがペアになった学習サンプルを多数準備し，学習サンプルの画像を実際にネットに入力したときの出力 $p_1, \ldots, p_n$ と，望みの出力の「誤差」を小さくする．目標とする出力 $d_1, \ldots, d_n$ は，正解クラス j のみ $d_j = 1$ となり，それ以外の全 $k (\neq j)$ では $d_k = 0$ となるようにするのが普通である．その上で「誤差」は，目標とする出力 $d_1, \ldots, d_n$ と，実際の出力 $p_1, \ldots, p_n$ の乖離を表した交差エントロピー

$$C = -\sum_{j=1}^{n} d_j \log p_j \qquad (3)$$

で測る．各サンプルについて計算されるこの誤差を，以下に述べる方法で小さくする．

2.3 確率的勾配降下法

誤差 C は通常，勾配降下法で最小化する．勾配降下法は，ネットワークのパラメータ（重みとバイアス）を C が最も小さくなる方向，すなわちパラメータによる C の勾配方向に修正することを，反復する方法である．これには微分 $\partial C / \partial w_{ij}$ の計算が必要だが，それには活性化関数が層の数分だけ入れ子になった合成関数の微分を要し，特に出力層から離れた深い層ほど煩雑な計算となる．有名な誤差逆伝播法（back propagation）は，これを系統的かつ効率よく行えるようにした方法である．連続する層の活性化関数の合成関数を微分する際の連鎖規則の適用を，数値計算によって行う．計算は，出力層での誤差を，出力から入力へと逆向きに伝播する形をとる．

なお誤差 C は，学習データの全サンプルに対して評価するのではなく，ミニバッチと呼ばれる数個〜数百程度のサンプルの集合に対して求める．この方法は，サンプルをランダムに選んだものを対象とすることで大域的な収束性能を向上させる（確率的勾配降下法と呼ばれる）ことと，現代の計算機の持つ並列計算資源を有効に活用することを天秤に掛けた結果の産物である．重みとバイアスの修正は，このミニバッチ単位で行う．すなわち

$$\Delta w_{ij}^{(t)} = -\epsilon \frac{\partial C}{\partial w_{ij}^{(t)}} + \alpha \Delta w_{ij}^{(t-1)} - \epsilon \lambda w_{ij}^{(t)} \qquad (4)$$

と更新する（バイアスも同様）．ここで $\Delta w_{ij}^{(t)}$ と $\Delta w_{ij}^{(t-1)}$ はそれぞれ，今回と前回更新時の重みの修正量である．右辺第 1 項は誤差勾配の項で，ϵ は勾配降下のステップ幅を決定する制御パラメータで，学習係数と呼ばれる．第2 項はモメンタムと呼ばれ，前回修正量の何割か $\alpha (\sim 0.9)$ を最新の修正量に加算し，修正量のばらつきを減じる．第3 項は重み減衰項（weight decay）と呼ばれ，重みが過大に大きくならないようにする正則化である．これらパラメータ ϵ, α, λ は適宜うまく選ぶ必要があり，学習の行方を左

右することが少なくない．特に学習係数ϵの決定にはいくつかのノウハウがある [2].

なお一般的な最小化問題では，勾配降下法よりも収束性能に優れる方法は（準ニュートン法など）沢山あるが [3]，大規模なディープネットの学習では，効率面から勾配降下法が今でも最も一般的である [2].

3. 深層学習を支える技術

3.1 勾配消失問題と過学習

現在の深層学習のブームの前，90 年代後半から 00 年代初頭にかけて，ニューラルネットの研究は下火と言える状態であった．その理由の一つに，層を積み重ねて多層化したディープネットの学習がうまく行えなかったことが挙げられる．ネットワークを多層化すると，いわゆる勾配消失問題が生じ，これを解決できなかったのである．勾配消失問題とは，前章で述べた誤差の逆伝播計算を行うとき，層を遡るに従って誤差が急速に小さくなり 0 になる（あるいは急速に大きくなって爆発する）ために，学習が制御不能に陥ることを言う．その一つの現れが過学習，すなわち学習サンプルに対する誤差（訓練誤差）はいくらでも小さくできるのに，汎化誤差を小さくできないことであった.

この勾配消失問題や過学習を克服する手立てが見つかったことが，今のディープネットのブームの根底にある．今ではその方法は何通りかあることが分かっているが，次に述べる Hinton らの研究 [4] が，その一番最初のものである.

3.2 事前学習

勾配降下法は反復法であるので，当たり前だが最初にネットワークの重みを初期化する必要がある．従来，最も一般的な（かつ唯一と考えられていた）方法は，それらをランダムに初期化することであった．これに対しHinton らは，この初期値の決め方が適切でないから上述の諸問題が生じるのであって，事前学習という方法で重みをよりうまく初期化すれば，これらを回避できると示した．具体的には**図 4** のように，目的とするディープネットを単層ネットワークに分離し，それぞれ別々に，ただし入力層側から順番に一つずつ訓練する．各層の訓練は，それぞれをオートエンコーダと見なし，教師なし学習を行う†.

オートエンコーダとは，入力に対する出力がなるべく入力に近づくことを理想とするニューラルネットである．元になるネットワークが，入力 x から出力 y を計算するとき（**図 5** 左），向きが反対の計算，すなわち出力y を入力側 x' に戻す計算を考える．一連の計算$x \to y \to x'$ は，元のネットワークを出力層で折り返した，図5 右のような

†元の Hinton らの論文では，単層オートエンコーダの代わりに制約ボルツマンマシン（restricted Boltzmann machine）が使われている．各層の学習方法は幾分ことなるが，事前学習としての考え方は同じである [5].

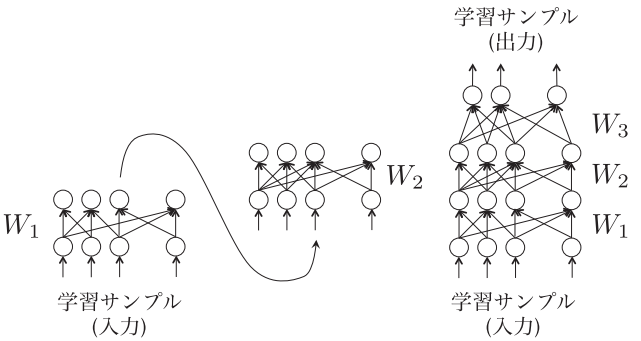

図 4 事前学習の概要（3 層ネットワークの場合）．一番右のネットワークの第 1 層の重み W_1 と第 2 層の重み W_2 について，それらの初期値を各層をオートエンコーダと見て学習することで得る

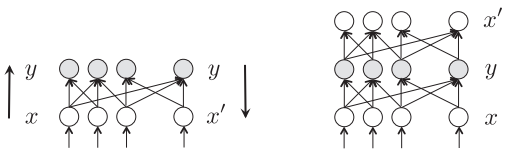

図 5 オートエンコーダの概要．左の単層のネットワークを上の出力層で折り返して 2 層構造とし，入力になるべく近いものを出力するように学習を行う

2 層のネットワークにより実現される．これが，個々の入力になるべく近い出力を返すよう，2 層の重みを学習最適化する．元のネットワークから見れば，学習サンプルはxの集合であり，目標出力は存在せず，したがってその学習は教師なし学習である（折り返して作った 2 層ネットワークは，入力 x がそのまま目標出力となる形である）.

学習には，最終的に目的のディープネットで扱おうとしている学習サンプル（の入力のみ）の集合を使う．自然画像のように，自然界で発生するデータにはそのデータの空間内で強い偏りがあると言える．オートエンコーダの狙いは，そんなデータに内在する偏りをうまく取り出し，表現できるようにしようというものである．なおオートエンコーダは，中間層（元のネットワークの出力層）のユニット数を増やすほどに自由度が高まり，したがって表現能力が増す．ただし自由度があまりに大きいと，恒等写像が学習されてしまい，意味がなくなる．そこでオートエンコーダの学習時には，個々の入力サンプルを表すのに中間層のユニットが少数しか使用されないよう，制約を加える．中間層のユニットがごくまばら（= スパース）にしか活性化しないように制約することから，スパース正則化と呼ばれる.

ディープネットの事前学習とは，以上のオートエンコーダによる各層の学習を，入力層から順に，第2 層，第3 層へという順で繰り返す．最初のオートエンコーダの入力には，上述のようにディープネットが対象とする学習サンプルと同じものをそのまま使う．次の第2 層の入力には，最初の層の学習後決定された重みを使い，得られる学習サン

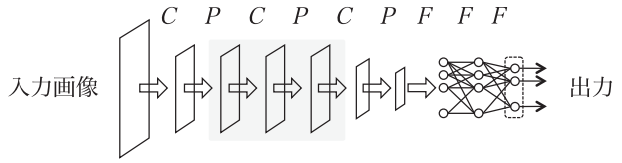

図 6 畳込みネットの全体像．C, P, F はそれぞれ畳込み層，プーリング層，全結合層を表す

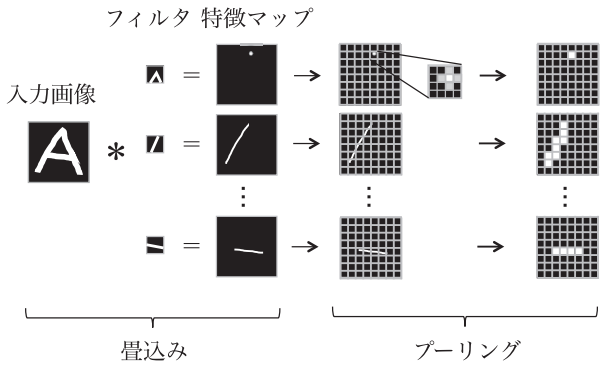

図 7 畳込み層およびプーリング層の概要

プルに対する出力（図 4）を用いる．後はこれを層の数だけ繰り返す．このように定めた各層の重みを，目的とするディープネットの各層の重みに利用する．ただし出力層だけは例外で，重みは従来どおりランダムに初期化する．事前学習を終えた後は，ディープネット全体の学習を教師あり学習によって（誤差逆伝播＋勾配降下法で）で行う．

3.3 畳込みニューラルネット

ここまで考えてきたネットワークは，層間の結合が均一で，隣接層の全ユニット間で結合があるようなものであった．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）とは，畳込み層とプーリング層と呼ぶ 2 種類の層を，交互に積み重ねた構造を持つディープネットである．神経科学の分野での Hubel-Wiesel の発見 [6] と，それに基づいて考案された Fukushima らのネオコグニトロン [7] にルーツがある．

畳込みネットは，**図 6** のように畳込み層とプーリング層のペアを複数回積み重ね，最後に通常の全結合層を数層程度積み重ね，最後に出力を計算する構造である（カテゴリ認識が目的なら出力層は 2 章に述べた構造を与える）．

畳込み層は，多チャネルの画像に同じチャネル数の小サイズの二次元フィルタ（例えば 5×5 画素程度）を畳込む演算を行う（**図 7**）．画像をぼかしたりエッジを強調する目的で行う，一般的な画像処理でのフィルタの畳込みと，基本的に同じである．このフィルタの畳込みによって，局所的な特徴が抽出される．

一方プーリング層は，畳込み層の出力を入力とし，入力された多チャネル画像の小領域（例えば 5×5 画素程度）での値を，一つの値に集約（プーリング）する（図7）．簡単に言えば，入力された多チャネルの画像の解像度を落とす演算である．これによって，入力画像内の特徴の位置が若干変化しても，取り出される特徴はほとんど変化しないという不変性が獲得される．なお，小領域内の複数の値を一つに集約する方法にはいくつかあり，現在最も一般的なのが小領域内の最大値を選んで出力とするマックスプーリングである．

畳込みネットは，以上の畳込み層とプーリング層を結合し，このペアを何度か繰り返す構造を持つ．畳込み層とプーリング層のペア一つで，局所特徴の抽出と位置の微小変位に対する不変性を実現し，これを多層化することで，より複雑な幾何学的な変形に対する不変性を実現している．なお最近では，畳込み層とプーリング層をいつも対にするのではなく，畳込み層を何度か繰り返した後でプーリング層を 1 層入れることも多い．このようにネットワークの設計の考え方は，現状では試行錯誤の段階にある．

畳込み層とプーリング層はいずれも，特殊な配線を持つ単層ネットワークとして表現できる．どちらも，層間の結合は疎らで，層間で全ユニットが結合する層とは異なる．ただし，その構造を除けば畳込みネットも通常のニューラルネット同様，誤差逆伝播による学習が可能である．学習では，畳込み層の重み，すなわちフィルタの係数と上位にある全結合層の重みが調節される．プーリング層の結線は固定で，学習では変化しない．

なお畳込みネットは，多層のネットワークであるにもかかわらず，前節のような事前学習を要しない．つまり，重みをランダムに初期化した状態から直接，誤差逆伝播法で学習を行えてしまう．そしてこのことは，かなり前から知られていた．具体的には，80 年代末にはすでに多層ネットワークの学習に成功し，文字認識のタスクに適用して高い性能が得られていた [8] [9]．なお，畳込みネットが過学習を起こしにくい理由は，結線が局所的に限定されている畳込み層の構造にあると考えられている．層間で全ユニットが結合したネットの場合，先述のように誤差逆伝播時に，層を経るに従って誤差が拡散し小さくなる傾向にあるが，畳込みネットではその局在性ゆえ，同じような誤差の拡散が起こりにくい．

畳込みネットは現在，ディープラーニングの画像認識応用で最も重要な技術である．画像認識の各タスクで良い性能を示しているものは例外なく，畳込みネットである．実はこれら最近のネットは，文字認識に適用された 80 年代の LeCun らのネットとほとんど変わっていない．違いは，適用されるドメインが広がり（文字認識から画像認識全般へ），また学習サンプル数およびネットワークサイズが大きくなったことだけである．畳込みネットに関して言えば，深層学習のブームの中でその価値が再評価されている，というのが正しい．

4. ま と め

以上，ニューラルネットの基礎と，最近の深層学習のブームにつながったディープネットの学習方法について述べた．上ではフィードフォワード型のニューラルネットのみ説明したが，このほかにも重要なニューラルネットがいくつかある．まず出力を入力にフィードバックし，系列データの処理を行うリカレントニューラルネット（RNN）がある．RNN は音声認識や自然言語の諸問題で成果を挙げている．

また，ネットワークの挙動を確率的に記述し，データの生成モデルを与えるボルツマンマシンも重要である．特に上で述べたオートエンコーダと同様に使える制約ボルツマンマシン（RBM）や，それを重ねた構造を持つ多層ボルツマンマシン（DBM）あるいは多層ビリーフネット（DBN）は，深層学習の研究では主要な位置を占める．ただしあくまで応用の中心にあるのは，（ボルツマンマシンではなく）フィードフォワードネット（および勾配降下法によるそれらの学習）である．

最後に，画像認識における最近のディープネットの話題について紹介し，本稿のまとめとしたい．この分野でディープネット（具体的には畳込みネット）が最も威力を発揮しているのが，物体カテゴリ認識である．この問題では，ディープネットとニューラルネット以外の方法との性能差は極めて大きい．

物体カテゴリ認識の性能評価では，ILSVRC（ImageNet Large Scale Visual Recognition Challenge）という毎年開催されているコンテストが最も有名である．主要課題の一つが，1000 クラスの物体カテゴリを認識するタスクである．2012 年のコンテストで初めて，Krizhevsky らのディープネットがこのタスクにエントリし，従来からのニューラルネットを使わない方法を圧倒してみせた．同時にエントリしたニューラルネットを用いない方法は，どれも誤答率が 25%† 前後であったのに対し，彼らの畳込みネットは約 15%の誤答率を示したのである．このディープネットは畳込み層 5 層，全結合層 2 層をもつものであった．

翌年 2013 年には，ほとんどのチームが畳込みネットを使うようになったが，いずれも Krizhevsky らのネットワークを若干改良した程度であった．それでも誤答率は11%程度に上昇した．さらに翌年の2014 年には，参加チームが大幅に増え，それぞれ独自の畳込みネットワークの構造を取り入れるようになってきた．誤答率が最も低かったのはGoogle のチームによる GoogLeNet と呼ぶネットワークで，誤答率は 7%以下であった．2 位はオックスフォード大のチームで，同様に約7%の誤答率を達成した．この2 チームとも，学習の対象となる重みを持つ層（畳込み層と全結合層）の数を，前年度一般的であった 7 層前後から 20 層前後に大幅に増やしている．現在，層数を増やすことで認識精度を向上させようというのがトレンドとなっている．

このように性能向上のペースは上がっているが，深層学習には不明な点も多い．従来方法では望めないほど高い性能を示す反面，なぜそのような高い性能が得られるのか，きちんとした答えは今のところない．また学習係数などチューニングを要するパラメータが多数あり，そのやり方いかんで結果が左右されるなど，使い勝手は以前と同様に悪いままである．また，特徴そのものの学習を可能にする対価として，従来法よりも（その柔軟性・自由度に見合った）大量の学習サンプルを必要とする．これらの問題を解決すべく，全世界で研究が行われているところである．

†候補を五つまで上げたとき，その中に正しいカテゴリが入っていれば正答，そうでなければ誤りという判定の下での誤答率．

参 考 文 献

[1] V. Nair and G.E. Hinton: “Rectified linear units improve restricted boltzmann machines,” Proc. ICML, 2010.

[2] Y. LeCun, L. Bottou, G. Orr and K. Muller: ‘Efficient backprop,’ Neural Networks: Tricks of the trade. Springer, 1998.

[3] Q.V. Le, A. Coates, B. Prochnow and A.Y. Ng: “On Optimization Methods for Deep Learning,” In Lise Getoor and Tobias Scheffer, editors, Proc. ICML, ICML’11, pp.265–272, 2011.

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

[5] Y. Bengio: “Learning deep architectures for ai,” Foundations and Trends in Machine Learning, vol.2, no.1, pp.1–127, 2009.

[6] D.H. Hubel and T.N. Wiesel: “Receptive fields and functional architecture of monkey striate cortex,” The Journal of physiology, vol.195, no.1, pp.215–243, 1968.

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

[8] 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, vol.1, no.4, pp.541–551, 1989.

[9] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner: “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol.86 (issue 11), pp.2278–2324, 1998.

岡谷貴之（Takayuki Okatani）

1999 年東京大学大学院工学系研究科博士課程修了（計数工学）．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（准教授へ名称変更）を経て，2013 年に教授，現在に至る．コンピュータビジョンの研究に従事．数理統計学および数値最適化を理論的基礎に，その多視点幾何から物体認識までの幅広い応用に関心を持つ．電子情報通信学会，情報処理学会，IEEE 等の会員．