2章 機械学習による画像理解

原田達也†

キーワード：物体識別・検出，スパースコーディング，フィッシャーカーネル，オンライン学習，ディープラーニング

1．まえがき

近年の機械学習の発展や，それを支える計算機の進化，画像データセットの整備により，人間のように画像を理解する知能システムの実現が現実味を帯び始めている．「画像を理解する」という意味は幅が広いため，例えば，画像に映る人の心情まで汲み取るようなことは本稿では視野に入れず，画像中の物体識別と物体検出に焦点を当てて述べていく．物体識別とは，一枚の入力画像に対してそこに映る物体のカテゴリーを予測するタスクのことであり，物体検出とは，予め与えられたカテゴリーの物体を画像中から対象物体の領域を含めて発見するタスクのことである．また，何が写っているかわからない画像に対して，物体のカテゴリーとその領域を予測する物体識別と物体検出を同時に行うタスクも存在し，最近ではこの複合課題が画像認識分野のトレンドとなりつつある．

画像認識技術は日進月歩であり，毎年膨大な論文が発表されているため，ここに使われる機械学習に基づく手法をすべて紹介することは現実的ではない．そこで2010年から開催されている大規模な画像を用いた画像認識のコンペティション（ImageNet Large Scale Visual Recognition Challenge: ILSVRC）*1で用いられ，高い性能を示すことが判明している手法を中心に紹介することにする．

2．物体識別・物体検出のパイプライン

物体識別や物体検出の概要を理解するために，物体識別に一般的に用いられるパイプラインを図1に示し，それぞれのモジュールについて順に説明する．

(1) 入力画像に対して局所的な領域の特徴を抽出する．これを局所特徴と呼ぶ．1枚の画像から数百から数万個程度の局所特徴が得られるのが一般的である．従来の画像認識では，局所領域の輝度勾配のヒストグラムを計算したScale-Invariant Feature Transform（SIFT）やテクスチャ情報を表現したLocal Binary Patterns（LBP）などがよく利用される．

(2) 局所特徴を識別に有利な特徴に変換する操作をコーディングと呼ぶ．後段のプーリングを用いて局所特徴群をまとめて画像を代表するベクトルを生成した時に，局所特徴群をモデル化した確率密度分布のパラメータとなるようなコーディング手法がよく用いられる．

(3) 画像空間に配置されたコーディング後の局所特徴群を1本または少数のベクトルにまとめる操作をプーリングと呼ぶ．このプーリングには対象ベクトルの平均値を計算するもの（平均値プーリング）や，ベクトルの各要素の最大値を計算するもの（最大値プーリン

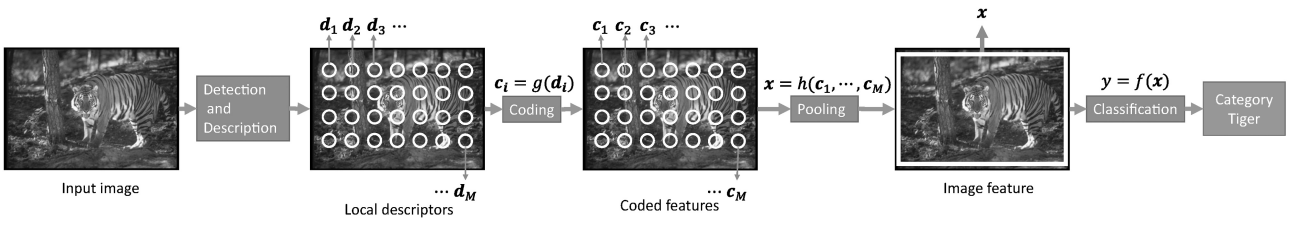


図1　物体識別のパイプライン
（口絵カラー参照）

＊1　http://www.image-net.org/challenges/LSVRC/

† 東京大学 大学院情報理工学系研究科
"Machine Learning for Visual Recognition" by Tatsuya Harada (Graduate School of Information Science and Technology, the University of Tokyo, Tokyo)

グ）などがある．このプーリングの結果，画像一枚を代表するベクトルが得られた場合，これを画像の特徴ベクトルと呼ぶ．

(4) 画像の特徴ベクトルを人，犬，猫などのカテゴリーに分類するモジュールは識別器と呼ばれている．このモジュールを経ることで物体識別が完了する．

従来の画像認識の枠組みでは，それぞれのモジュールを別々の問題として考えて，モジュール毎に機械学習を利用しながら設計するアプローチが取られてきた．一方，1から3のモジュールを多段に重ね，最後に識別器のモジュールを組み入れたパイプラインを考えて，初段から最終段までを一気に学習するのが物体識別におけるディープラーニングの枠組みである．1から3のモジュールを多段に重ねたパイプラインは深いネットワークと呼ばれ，1から3のモジュールが一つのパイプラインは浅いネットワークと呼ばれる．

物体検出は，入力画像の中から物体候補となる画像領域群を抽出し，各画像領域に上記の物体識別のパイプラインを適用することで実現される場合が多い．物体検出のデファクトスタンダードとして利用されてきた手法にDeformable Part Model（DPM）[7]がある．DPMの特徴として，変形可能な物体モデルを扱える点がある．物体の変形具合を隠れ変数と見なし，Latent SVMを適用することで識別器を学習している．物体検出において，負例はほぼ無限に存在するので，困難な負例のみを選別し利用するHard negative mining手法を提案している．画像中から物体候補の領域群を抽出するモジュールは，物体検出の精度と速度を決める重要な部分である．速度は遅いが確実な方法として，ある決まった大きさの小領域を一定のピクセル毎にずらしながら候補領域を抽出する手法が考えられる．これはスライディングウィンドウ法と呼ばれている．少数の物体候補領域を提案する精度の高い手法として最近ではSelective Search[19]がよく利用される．Selective Searchと同等の物体候補領域の提案精度ながら高速に動作するBING[2]も存在する．

以下では上記のモジュールの中でも，コーディング，識別器，また，モジュール群を一気に学習するディープラーニングについて説明する．

3．コーディング

3.1　Bag of Visual Words

局所特徴のコーディング手法としてBag of Features（BoF）[4]が広く利用されている．BoFは訓練集合から代表的ないくつかの局所特徴を取り上げ，画像の中に代表的な局所特徴がいくつ出現するかヒストグラムで表現したものである．BoFはBag of Visual Words（BoVW）とも呼ばれる．BoFの計算プロセスを以下に示す．

(1) 全局所特徴からK個の代表的な局所特徴を選択する．代表的な局所特徴をコードワード，選択されたコードワードの集合をコードブックと呼ぶ．コードワードにそれぞれw_1, ..., w_Kとラベルを付与する．K-meansを利用することが多い．

(2) すべての局所特徴をいずれかのコードワードに対応させる．この操作により，すべての局所特徴にw_1, ..., w_Kのラベルが付与される．

(3) コードワードに関するヒストグラムを計算する．つまり，w_kとラベル付与された局所特徴の数をカウントする．コードワードのヒストグラムをその画像の特徴ベクトルとする．

3.2　スパース符号化

BoFのように一つのコードワードへの割当てでは量子化誤差が大きく，類似した局所特徴であっても量子化後の符号が異なる可能性がある．そこで，局所特徴を少数かつ複数のコードワードで表現するスパース符号化（sparse coding）[22]が提案されている．これにより上記の問題を解決しつつ，量子化誤差を低減可能である．

さらに，局所特徴を近傍に存在するいくつかのコードワードの線形和で局所的に近似する手法が提案されている[23]．この結果，得られた線形和の重みはデータの局所座標符号化（Local Coordinate Coding: LCC）と呼ばれる．文献23）では，ある仮定の下では局所性がスパースネスよりも本質であると述べている．しかしながらスパース符号化と同じように，LCCもL1ノルム最適化問題を解く必要があり，計算コストが高い問題を抱える．そこで，LCCの高速な実装と見なせる局所制約線形符号化（Locality-constrained Linear Coding: LLC）[20]が提案されている．スパース符号化ではコードワードが過剰であるためにスパース性を優先することで類似した局所特徴に対してまったく異なるコードワードを選択する可能性があるが，局所制約線形符号化は類似した局所特徴には類似したコードを出力可能である．図2にBoF，スパース符号化，局所座標符号化の比較を示す．

3.3　BoFの混合ガウス分布による改良

前述の通りにBoFは局所特徴のヒストグラムを計算する手法であるが，局所特徴のヒストグラムを特徴空間におけ

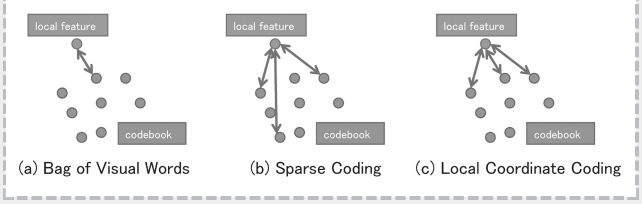


図2　BoF，スパース符号化，局所座標符号化の比較
（口絵カラー参照）

る局所特徴の確率密度分布と考えると，ヒストグラムによる表現は確率密度分布の粗い表現と言える．よって局所特徴の確率密度分布推定をより正確に行えば識別性能の向上につながると考えられる．そこで混合ガウス分布を用いることで，BoF表現を改善する試みが行われている[6]．混合ガウス分布はガウス分布の線形重ね合わせで書ける．

$$p(\boldsymbol{x})=\sum_{k=1}^{K}\pi_k\mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu}_k,\,\Sigma_k)=\sum_{k=1}^{K}\pi_k p_k(\boldsymbol{x}), \tag{1}$$

ここで$\mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu}_k,\,\Sigma_k)$, $\rho_k(\boldsymbol{x})$は混合要素であり，平均$\boldsymbol{\mu}_k$と分散Σ_kを持つ．π_kは混合係数である．混合ガウス分布のパラメータは最尤法によって求めることが可能である．

混合ガウス分布を利用するメリットとして，混合ガウス分布を構成する各ガウス分布がそれぞれ共分散を持つために共分散を考慮した距離計量を利用できることがあげられる．また，BoFは局所特徴が一つのコードワードのみに割当てられるが，混合ガウス分布では局所特徴と多くのコードワードとの関係を表現できるので，特徴空間における局所特徴の位置に関する情報をエンコードできるメリットもある．デメリットして混合ガウス分布表現はBoFと比較してパラメータが多い．そのため，混合ガウス分布は訓練データに対して過剰適合する可能性があり，学習時に正則化や事前知識の導入等を行う必要がある．

3.4　フィッシャーベクトル

確率的生成モデルからカーネル関数を構成する手法がフィッシャーカーネルである．フィッシャーカーネルでは，局所特徴を生成する確率密度分布から導出される勾配ベクトルを計算し，画像を表現する一つの特徴ベクトルとする．

ここで，u_θをあらゆる画像の内容を表現する確率密度関数とし$\boldsymbol{\theta}$を確率密度関数のパラメータとする．局所特徴群を$\mathcal{X}$とすると，このデータを次に示す勾配ベクトルで表現する．

$$G_\theta^{\mathcal{X}}=\frac{1}{N}\nabla_\theta\log u_\theta(\mathcal{X}\,|\,\boldsymbol{\theta}). \tag{2}$$

対数尤度の勾配はデータに最も適合するように確率密度関数のパラメータが修正すべき方向を表現している．また異なるデータサイズの$\mathcal{X}$をパラメータ数に依存した決まった長さの特徴ベクトルに変換する．

この勾配ベクトルはさまざまな識別器に利用できるが，内積を利用する識別器ではベクトルを適切な計量を用いて正規化する必要がある．この正規化にはフィッシャー情報行列が利用できる．

$$F_\theta=E_X[\nabla_\theta\log u_\theta(\mathcal{X}\,|\,\boldsymbol{\theta})\nabla_\theta\log u_\theta(\mathcal{X}\,|\,\boldsymbol{\theta})^\top]. \tag{3}$$

フィッシャー情報行列を用いて正規化された勾配ベクトルは次のように与えられる．

$$\mathcal{G}_\theta^{\mathcal{X}}=F_\theta^{-1/2}\nabla_\theta\log u_\theta(\mathcal{X}\,|\,\boldsymbol{\theta}). \tag{4}$$

このようにしてできた画像の特徴ベクトルを局所特徴群$\mathcal{X}$のフィッシャーベクトル（Fisher vector）と呼ぶ[15]．ILSVRC2011でトップのXerox Europeのチームは，直積量子化[10]を用いてデータを圧縮し，すべての訓練データをメインメモリーに蓄積し，学習時には圧縮されたデータを復号化して重みを更新することで学習の高速化を実現している[17]．

4．識別器

物体認識では識別器としてSupport Vector Machine（SVM）がデファクトスタンダードとして利用されている．大規模データを一括学習するにはメモリーの問題や追加学習の困難さもあり利用しにくいため，オンライン学習が用いられる．物体認識では，確率的勾配降下法（Stochastic Gradient Descent method: SGD method）をSVMに適用したオンライン学習[1]がよく利用される．重み$\boldsymbol{w}$の線形識別器を$y=\boldsymbol{w}^\top\boldsymbol{x}$，ラベル付き訓練画像のペア$\{\boldsymbol{x}_t,\,y_t\}_{t=1}^T$とするとSVMのコスト関数は次式のように表される．

$$L=\lambda\|\boldsymbol{w}\|^2+\max[0,\,1-y\boldsymbol{w}^\top\boldsymbol{x}], \tag{5}$$

ここでλは正則化パラメータである．このコスト関数に確率的勾配降下法を適用した時の重みの更新式は次のようになる．

$$\boldsymbol{w}\leftarrow\boldsymbol{w}-\rho_t\begin{cases}\lambda\boldsymbol{w} & \text{if } y_t\boldsymbol{w}^\top\boldsymbol{x}_t>1\\ \lambda\boldsymbol{w}-y_t\boldsymbol{x}_t & \text{otherwise}\end{cases} \tag{6}$$

標準のSGDでは収束に時間がかかるために重みに対して平均化のスキームを取り入れて収束の高速化が行われる[14]．

オンライン学習により大規模データへの適応が可能になる．しかしながら，データが高次元かつ膨大になると上記の重みの更新式よりも，ハードディスクなどの補助記憶装置から主メモリーへのデータ転送時間が学習速度のボトルネックとなる．ILSVRC2010でトップのNEC lab. Americaのチームは，Hadoopを利用して1-vs-all SVMを並列計算させているが，メインメモリーにロードした学習サンプルをなるべく多くの識別器の学習に同時利用することで，データのロード回数を減らし学習高速化を実現している[14]．

オンライン学習の手法はSGD SVMだけではなく，Passive Aggressive[3]やSoft Confidence Weighted Learning[21]など複数提案されている．Ushikuらは大規模画像データを用いた物体認識におけるこれらのオンライン学習手法を調査し，重みの平均化を用いることでパーセプトロンでも最新のオンライン学習手法に近い識別性能が得られることやマルチクラスの学習手法が1-vs-allよりも学習効率が良いことを示している[18]．

5．ディープラーニング

ディープラーニングは多層に積み重ねた層を重みベクト

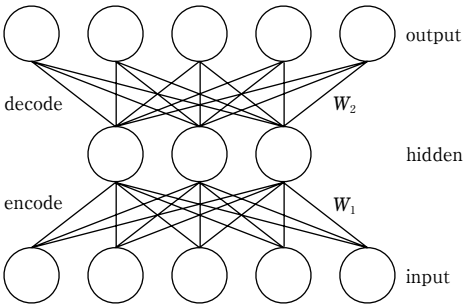


図3　Autoencoder

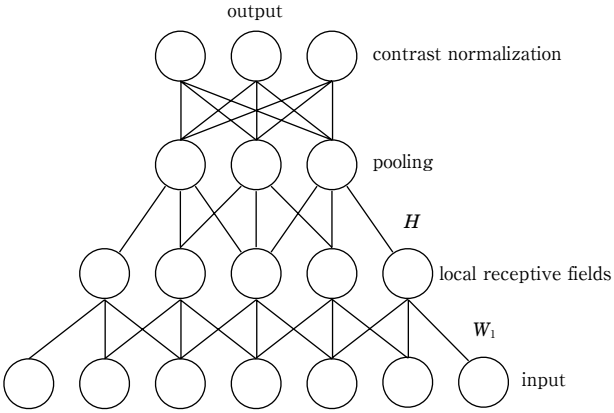


図4　Leらの用いたDeep Autoencoderの1モジュール

ルで接合し，その重みを最下層から最上位層まで一気に学習する手法である．画像認識においてディープラーニングがブレークした年が2012年である．2012年6月にLeらによって，1,000万本のYouTube動画から200×200のサイズの画像を切出し，16,000コアの計算機上に実装したDeep Autoencoderを用いた手法が発表された[12]．この手法は，教師なし学習にもかかわらず，人の顔のみに反応する検出器や猫の顔のみに反応する検出器が構築できたという触れ込みでNew York Timesをはじめメディアに大きく取り上げられた．2012年10月にはILSVRCのコンペティションにおいて，KrizhevskyらのDeep Convolutional Neural Network（Deep CNN）を用いた手法[11]が，図1に示す従来の画像認識のパイプラインの手法を大きく上回り，画像認識分野に大きな影響を与えた．以上より，ディープラーニングにもいくつも手法が存在するが，本稿ではDeep AutoencoderとDeep CNNを紹介する．

5.1　Autoencoder

Autoencoderでは，入力を隠れ層における表現に射影し，さらに隠れ層の表現を出力に射影する．入力と出力がなるべく同じとなるように学習されるのがAutoencoderである（図3）．

入力ベクトルを$\boldsymbol{x}$とすると，隠れ層への射影は以下のようになる．この操作をコーディングと呼ぶ．

$$\boldsymbol{y}=f(W_1\boldsymbol{x}+\boldsymbol{b}_1), \tag{7}$$

ここで，関数fとしてシグモイド関数などが用いられ，$\boldsymbol{y}$が隠れ層での$\boldsymbol{x}$の表現，$\boldsymbol{b}_1$はバイアス，W_1が入力層から隠れ層への重み行列である．隠れ層からの再構築は次にようになる．この操作をデコーディングと呼ぶ．

$$\boldsymbol{z}=f(W_2\boldsymbol{y}+\boldsymbol{b}_2), \tag{8}$$

ここで，$\boldsymbol{z}$が出力ベクトル，W_2は隠れ層から出力層への重み行列，$\boldsymbol{b}_2$はバイアスである．しばしば$W_2=W_1^{\top}$という制約条件を設けることがある．入力$\boldsymbol{x}$と出力$\boldsymbol{z}$との再構築誤差が小さくなるように重みW_1とW_2が学習される．再構築誤差の基準として，クロスエントロピーなどが利用される．また，パラメータの学習にはSGDがよく利用される．

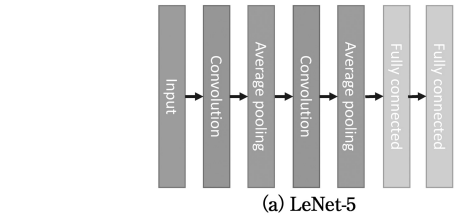
入力層と出力層に同じ画像を設定することで，隠れ層にその画像の本質的な情報が得られると考えられるため，画像認識ではAutoencoderを画像特徴抽出器として用いることが可能である．

次にLeらの用いたSparse Deep Autoencoder[12]を紹介する．この手法では一つの層が図4のような構造となっており，この層を3つ積み重ねたモデルを利用している．この一つの層には局所受容野，局所L2プーリング，局所輝度正規化の3つの重要な要素が備わっている．局所受容野は異なる場所で同じ重みを共有しておらず後述する畳み込みとは異なる．このネットワークは60億もの膨大な数のパラメータを持つ．なるべく重要な情報をエンコードし，かつプーリング特徴として類似した特徴がグループ化されるようにパラメータが学習される．

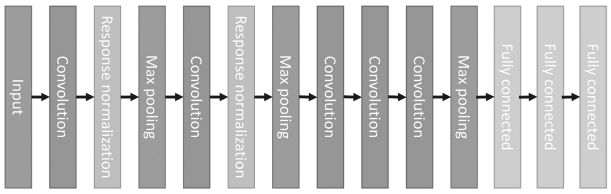
この膨大な数のパラメータの最適化問題を，複数の計算機で効果的に計算させるためにLeらは非同期SGDを提案している[12]．非同期SGDでは，モデルのパラメータを管理するパラメータサーバを設置する．各計算機にモデルを準備し，これをモデルレプリカと呼ぶ．

(1) 各モデルレプリカはパラメータサーバに問合せて，更新されたパラメータのコピーを取得する．
(2) 各モデルレプリカは取得したパラメータを元に，パラメータの勾配をミニバッチにより計算する．
(3) 計算されたパラメータの勾配をパラメータサーバに送信する．

通信のオーバヘッドを減らすために，モデルレプリカのパラメータ取得リクエストはある決まったステップ毎に行い，更新された勾配情報もある決まったステップ毎に送信する．このプロセスを繰り返すことで，計算速度の遅い計算機に足を引っ張られずにパラメータの並列学習が可能となる．この計算はDistBelief[5]と呼ばれる分散計算フレームワークで実行されている．



(a) LeNet-5



(b) AlexNet

図5　LeNet-5とAlexNet
（口絵カラー参照）

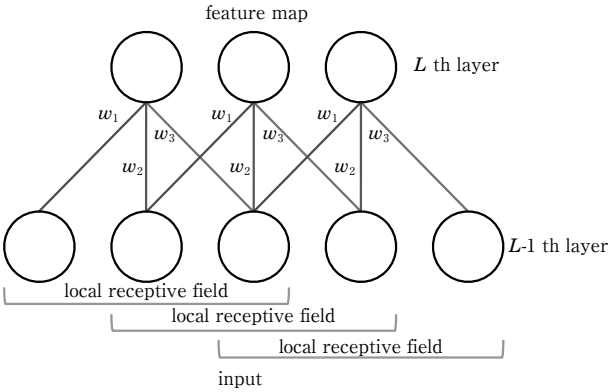
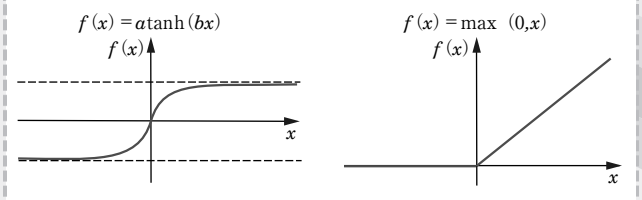


図6　畳み込み層
（口絵カラー参照）



(a) Hyperbolic tangent function　(b) Rectified Linear Units (ReLUs)

図7　tanh関数とReLUs
（口絵カラー参照）

5.2　Convolutional Neural Network

CNNは畳み込み層，プーリング層の組合せにより実現されるネットワークである．現状の高精度な物体識別，物体検出システムの多くはCNNを基盤としている．CNNの例としてLeCunらによって構築された手書き文字認識[13]を説明する（図5（a））．このネットワークはLeNetとも呼ばれる．LeNetは，畳み込み層とプーリング層が2つスタックされ，最後に全結合ネットワークにより出力される．このシステムではプーリングには平均値プーリングを用いて画像が1/4の大きさに縮小される．プーリングの操作により微小な歪み，移動の影響を軽減している．

ここで，畳み込み層の説明をする（図6）．$L-1$番目の層からL番目の層の間で結合を局所に制限する．局所領域を局所受容野と呼ぶ．全結合ネットワークと比較して，パラメータ数を低減させられるので汎化性能の向上が期待できる．また，画像の一部で有効な特徴抽出であれば，画像の他の部分でも有効な特徴抽出と考えて重みの共有を行う．この重みのことをカーネルとも呼ぶ．重みの共有の仮定によってさらにパラメータ数を減らすことができる．このようにして得られた層のことを特徴マップと呼ぶ．もちろん，豊かな特徴抽出をするためにカーネルは複数種類用いることができ，カーネルをK個準備したとすると，同一の入力層から，K枚の異なる特徴マップが得られる．

つまり，ある層におけるk番目の特徴マップ$\boldsymbol{y}^k$は入力画像$\boldsymbol{x}$にk番目のカーネル$\boldsymbol{w}^k$を畳み込むことにより得られる．

$$\boldsymbol{y}^k = f((\boldsymbol{x} * \boldsymbol{w}^k) + \boldsymbol{b}^k), \tag{9}$$

ここで，*は畳み込み演算を表し，$\boldsymbol{b}^k$はバイアス項である．fとしてロジスティックシグモイド関数やハイパボリックタンジェント関数などが利用される．カーネルやバイアス項といったパラメータ群の学習にはSGDによるバックプロパテーションが用いられる．

次にILSVRC2012でトップとなったKrizhevskyらのDeep CNN[11]を紹介する（本稿ではAlexNetと呼ぶことにする）．ネットワーク構造を図5（b）に示す．5つの畳み込み層と3つの全結合層から構成される．第1，2番目の畳み込み層の後に正規化層，各正規化層の後と5番目の畳み込み層の後には最大値プーリング層が用いられている．AlexNetは6,000万のパラメータと650,000個のニューロンを持つ．このネットワークは膨大なパラメータを含むため，局所解になるべく陥らずに高速に学習する工夫が幾つかなされている．

一つ目はReLUs（Rectified Linear Units）である．一般にニューロンの出力関数はロジスティックシグモイド関数やハイパボリックタンジェント関数などが用いられるが，これらの飽和する非線形の関数群（図7（a））を用いた場合，収束が遅いことが知られている．そこで$f(x) = \max(0, x)$というReLUs（図7（b））を用いることで収束を高速化している．AlexNetにおいてReLUsはすべての畳み込み層と全結合層の出力に適用されている．

二つ目はドロップアウト[9]である．多くの異なるネットワークの予測の平均を用いることで，新規データに対する予測誤差を低減できることが知られている．ドロップアウトでは，訓練データが提示される毎に，隠れ層のニューロンの出力を1/2の確率で0とする．つまり，訓練データが

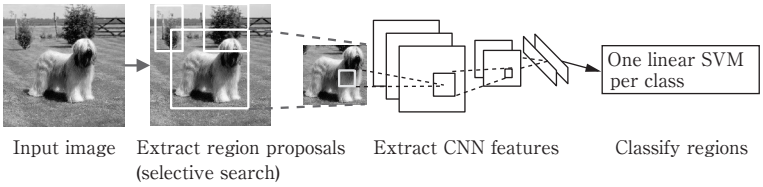


図9　RCNNのパイプライン
（口絵カラー参照）

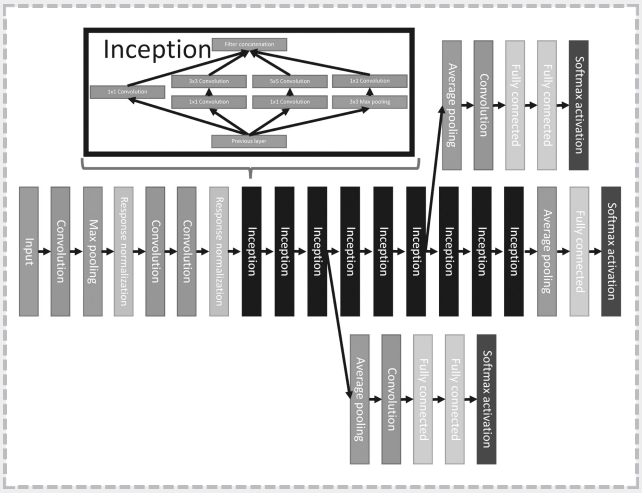


図8　GoogLeNet
（口絵カラー参照）

提示される毎に異なるネットワーク構造が選択される．ただしネットワークの重みは共有している．AlexNetでは第1，2番目の全結合層に適用されている．N個の隠れ層ユニットを持つとすると，2^N個のネットワークによって予測されるカテゴリーの確率の平均を用いて，最終的なカテゴリー予測していることになると言われている．

三つ目はオリジナル画像に左右反転や平行移動等の変換を加えて学習データを拡張していることである．この学習データの拡張によって過学習を防いでいる．また，オリジナル画像の輝度を変化させて学習データを拡張し，画像識別性能の向上を図っている．

四つ目は，上記の手法を効率的に計算可能なように2台のGPGPUを用いて実装した点にある．

ILSVRC2013で物体識別部門でトップになったチームの手法も基本的にAlexNetと同じ構造をしている．ただし，各層の重み（カーネル）を可視化するDeconvolutional Networksを用いることでパラメータ調整のヒントを得て，最適化することでAlexNetよりも性能向上を図っている[24]．

ILSVRC2014で目立った成果を挙げたのはGoogle（GoogLeNet）とOxford大学の2チームである．両チームともAlexNetよりもさらに深い構造にしたDeep CNNを利用している．特にGoogLeNetは，inceptionという層の中にさらにネットワークを持つモジュールを用い，これを9つ接続した深いネットワークを構築している（図8）．inceptionモジュールの中では，1×1，3×3，5×5の3つの畳み込み層と最大値プーリング層が並列となっている．複数の畳み込み層を並列にすることで複数の広がりを持つ局所的な画像の相関を捉えることができる．3×3，5×5の畳み込み層の前段には1×1の畳み込み層が利用されているが，これは次元削減として機能している．GoogLeNetはAlexNetよりも深い構造をしてるが，パラメータ数は1/12となっている．ネットワークのトレーニングにはLeらのSparse Deep AutoencoderとDistBelief[5]を用いてCPUのみで実行されている．ILSVRCのデータにおいて，人の物体識別性能はGoogLeNetの物体識別性能よりも僅かばかり高いが，この性能を人が出すには膨大な学習時間が必要という報告もされている[16]．

CNNを基盤とした物体検出ではRCNN[8]がよく利用される．RCNNでは，はじめにSelective Search[19]を利用して画像内から物体領域の候補群を提案する．各物体領域の候補群をあらかじめ学習しておいたDeep CNNに入力し，中間層の出力をこの領域の画像特徴を見なす．画像特徴は線形SVMに入力され領域のカテゴリーを予測する．線形SVMの学習ではDPM[7]と同様にHardnegative miningを利用する．RCNNのパイプラインを図9に示す．GoogLeNetでもRCNNを用いて物体検出を行っている．

現状では，物体識別，物体検出においてディープラーニングが優位に立っているが，動作認識においてはまだ目立った成果が得られていない．2014年に行われたTHUMOS'14 Challenge＊2ではアムステル大学がトップとなったが，ここで利用されている手法では，ImageNetで事前学習したDeep CNNからの画像特徴と，Fisher Vectorの双方を利用して，SVMで動作識別を実現している．事前学習をしたDeep CNNを用いているものの，動画像からすべてのモジュールを一気に学習しているわけではない．

6．むすび

本稿では，ILSVRCで利用される手法を中心として，物体識別と物体検出に用いられる機械学習を紹介した．従来の画像識別のパイプラインは，局所特徴，コーディング，

＊2　http://crcv.ucf.edu/THUMOS14/home.html

プーリング，識別器という構成であったが，最近ではディープラーニングによりこれらを一気に学習して高い識別性能を示していることを述べた．

ただし，すべての状況で有効なアルゴリズムは存在しないことは心に留めておく必要がある．最適なアルゴリズムは，応用先，利用するデータセット，利用する計算機のアーキテクチャ等に依存する．数年後には，まったく新しい計算機アーキテクチャの出現により，新規のアルゴリズムが生み出されたり，時代遅れと思われた手法が復活する可能性は充分にある．今後のさらなる画像理解手法の発展を期待したい．

（2014年11月5日受付）

〔文　献〕

1) L. Bottou: "Large-scale machine learning with stochastic gradient descent", In COMPSTAT（2010）
2) M.-M. Cheng, Z. Zhang, W.-Y. Lin and P. Torr: "Bing: Binarized normed gradients for objectness estimation at 300fps", In CVPR（2014）
3) K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz and Y. Singer: " Online passive-aggressive algorithms", JMLR, 7, pp.551-585（2006）
4) G. Csurka, C.R. Dance, L. Fan, J. Willamowski and C. Bray: "Visual categorization with bags of keypoints", In ECCV Int. Workshop on SLCV（2004）
5) J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, Q.V. Le and A.Y. Ng: "Large scale distributed deep networks", In NIPS（2012）
6) J. Farquhar, S. Szedmak, H. Meng and J. Shawe-Taylor: "Improving "bag-of-keypoints" image categorisation: Generative models and pdf-kernels", Technical report, University of Southampton（2005）
7) P.F. Felzenszwalb, R.B. Girshick, D. McAllester and D. Ramanan: "Object detection with discriminatively trained part based models", IEEE TPAMI, **32**, 9, pp.1627-1645（2010）
8) R. Girshick, J. Donahue, T. Darrell and J. Malik: "Rich feature hierarchies for accurate object detection and semantic segmentation", In CVPR（2014）
9) G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever and R. Salakhutdinov: "Improving neural networks by preventing coadaptation of feature detectors", arXiv, 1207.0580（2012）
10) H. Jégou, M. Douze and C. Schmid: "Product quantization for nearest neighbor search", IEEE TPAMI, 33, pp.117-128（2011）
11) A. Krizhevsky, I. Sutskever and G. Hinton: "Imagenet classification with deep convolutional neural networks", In NIPS（2012）
12) Q. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean and A. Ng: "Building high-level features using large scale unsupervised learning", In ICML（2012）
13) Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard and L.D. Jackel: "Backpropagation applied to handwrittenzip code recognition", Neural Comput., **1**, 4, pp.541-551（1989）
14) Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao and T. Huang: "Large-scale image classification: Fast feature extraction and svm training", In CVPR（2011）
15) F. Perronnin and C. Dance: "Fisher kernels on visual vocabularies for image categorization", In CVPR（2007）
16) O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M.S. Bernstein, A.C. Berg and L. Fei-Fei: "Imagenet large scale visual recognition challenge", arXiv, 1409.0575（2014）
17) J. Sánchez and F. Perronnin: "High-dimensional signature compression for large-scale image classification", In CVPR（2011）
18) Y. Ushiku, M. Hidaka and T. Harada: "Three guidelines of online learning for large-scale visual recognition", In CVPR（2014）
19) K.E.A. van de Sande, J.R.R. Uijlings, T. Gevers and A.W.M. Smeulders: "Segmentation as selective search for object recognition", In ICCV（2011）
20) J. Wang, J. Yang, K. Yu, F. Lv, T. Huang and Y. Gong: "Localityconstrained linear coding for image classification", In CVPR（2010）
21) J. Wang, P. Zhao and S.C. Hoi: "Exact soft confidence-weighted learning", In ICML（2012）
22) J. Yang, K. Yu, Y. Gong and T. Huang: "Linear spatial pyramid matching using sparse coding for image classification", In CVPR（2009）
23) K. Yu, T. Zhang and Y. Gong: "Nonlinear learning using local coordinate coding", In NIPS（2009）
24) M. Zeiler and R. Fergus: "Visualizing and understanding convolutional networks", In ECCV（2014）

原田（はらだ）　達也（たつや） 2001年，東京大学大学院工学系研究科機械工学博士課程修了．現在，東京大学大学院情報理工学系研究科教授．実世界知能システム，画像認識，コンテンツ自動生成などの研究に従事．