ディープラーニングに用いる畳み込み演算による概念操作の表現
On Represenations for Concept Operations of Convolutional Neural Networks

浅川 伸一
Shin Asakawa
東京女子大学情報処理センター
Center for Information Sciences, Tokyo Woman's Christian University
asakawa@ieee.org

Abstract

This article was aimed to prompt us to understand several key concepts of deep learnings progressing a great deal recently. Those were LeNet, AlexNet, GoogleLeNet, VGG, MSRA(SPP-net), NIN, R-CNN, and Selective search. We also described not only these state-of-the-art algorithms in detail but also the pros and cons of them. Furthermore, these models would show us importand feed backs to understandings of our brains themselves. Such mutual interactions, between progression of neural networks and knowledge about our brains, would give us deeper understanding ourselves.

Keywords — Convolutional Neural Networks, Max-pooling, Higer-order Cognitive Processes, Concept acquisitions

1. はじめに

生体の視覚情報処理 [Hubel 59, Hubel 68] から着想された畳み込みニューラルネットワーク (Convolutional Neural Networks: CNN) [LeCun 98] は性能が保証されたこと [Russakovsky 15] により，逆にこの種の演算が脳内で偏在している可能性が指摘できる．顕著な成功により応用研究も盛んである．例えば画像分類 [Russakovsky 15]，情景認識 [Zhou 14]，画像風情認識 [Karayev 14]，物体検出 [Girshick 14]，系列制御 [Donahue 14]，画像領域分割 [Girshick 14, Donahue 14, Long 15]，視覚運動制御 [Levine 15]，などである．ここでは，本質はネオコグニトロン [Fukushima 82] で実現された位置，回転，拡大，縮小，ノイズといった局所的摂動に対して頑健であるという特徴を外挿して考えることを試みた．高次認知機能のうち抽象化，概念化は CNN+Max-pooling で相同の操作を定式化可能である．これを確認するために，本稿では CNN を高次認知過程における操作と CNN+マックスプーリング (Max-pooling) とを同一視可能か否かを議論した．本稿では [Rogers 04] と [Kemp 09] とに従って [Osherson 90] のデータを説明することを試みた．

以下で歴史的な経緯から CNN と Max-pooling とを概説し，幾つかの提案された手法を述べる．その後，これらの手法と人間の高次認知過程との関連を考える．

2. 画像認識の諸手法

ここでは CNN の起源である LeNet について述べ，正当な更深継承者 (deeper successors, 命名は [Long 15] による) である AlexNet, GoogleLeNet, VGG, MSRA を概説する．

2.1 LeNet

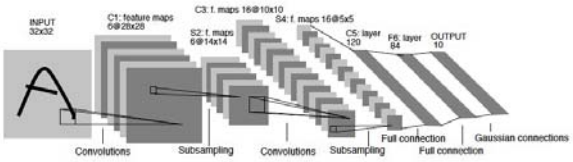

図 1 LeNet の概念図 [LeCun 98] の図 2 を改変．提案時点では Max-pooling ではなくサブサンプリング法であった

CNN[Ranzato 07] は疎性特徴検出器の多層化実装である (図 1)．各層は直下層と結合し受容野が形成される．最下位層においては正弦，余弦ガボール関数で定義される特徴検出器と入力画像との畳込み積分 $\int_d f(x-\chi)\,G(\chi)\,d\chi$ が行われる．ここで $f(x)$ は位置 x における入力信号強度であり，G は $[\sin,\cos](x)\exp(-x)$ である (図 2)．[Chatfield 14] によれば CNN の出力層次元は課題成績の低下を招くことなく大幅に低減することが可能である．画像認識で

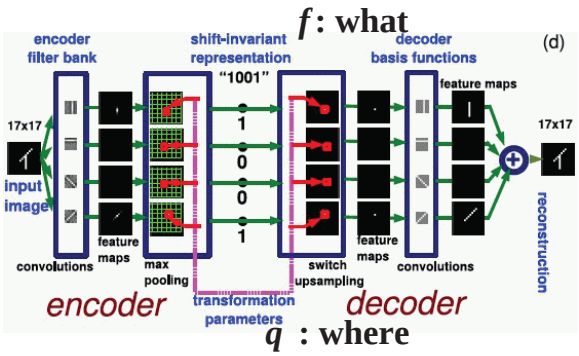

図 2　[Ranzato 07] 図 2 を改変

は入力 2 次元信号であり，CNN に用いる核関数には種々の候補がある．例えば次式のような，

$$G(x,y) = \cos\left(\frac{2\pi}{\lambda}X\right)\exp\left(-\frac{\left(X^2+\gamma^2Y^2\right)}{2\sigma^2}\right), \quad (1)$$

ここで $X = x\cos\theta + y\sin\theta, Y = -x\sin\theta + y\cos\theta$ である．θ は方位，σ は空間解像度，および λ は波長を表すパラメータである．これによって単純細胞を表現できる [Serre 05]．パラメータに依存する核関数集合の種類と量とに関する問題が存在する．如何なる特徴検出器をどれ程用意すれば外界を表現可能で，かつ，最上位層の認識のための十分な情報を提供できるかは事前に知る方法がない．このような議論は残るが，画像認識の分野においては CNN は生理学的事実に裏付けられた演算である．

$\boldsymbol{x}$ を入力画像，$\boldsymbol{w}$ を学習可能なパラメータベクトルとする．$\boldsymbol{f}$ は局所的な位置や回転に不変な特徴ベクトル (“what”)，$\boldsymbol{q}$ を各特徴の位置を指定する変換パラメータベクトル (“where”) とする．CNN+Max-pooling では 2 つの関数 $\boldsymbol{f} = f_f(\boldsymbol{x};\boldsymbol{w}_e)$ 及び $\boldsymbol{q} = f_q(\boldsymbol{x};\boldsymbol{w}_e)$ を考える．f_f は不変特徴ベクトル生成関数であり，f_q は変換パラメータベクトル生成関数である．$\boldsymbol{w}_e$ は学習すべき結合係数行列である．

同様に，復号過程では関数 $f_d(\boldsymbol{f},\boldsymbol{q};\boldsymbol{w}_d)$ を基底関数へ係数パラメータベクトル生成関数とする．入力信号 $\boldsymbol{x}$ と符号，復号過程で再構成された値とのユークリッド距離を復号エネルギー関数として

$$H_d = \|\boldsymbol{x} - f_d(\boldsymbol{f},\boldsymbol{q};\boldsymbol{w}_d)\|^2$$

とする．一方符号過程におけるエネルギー関数を

$$H_e = \|\boldsymbol{f} - f_e(\boldsymbol{x},\boldsymbol{q};\boldsymbol{w}_e)\|^2$$

と定義する．両エネルギー関数を EM アルゴリズムのごとく交互に推定する．$\boldsymbol{f}$ は特徴ベクトルであるが推定すべき未知変数として扱う．各入力毎に $H_d + \alpha H_e, (\alpha > 0)$ を最小にする $\boldsymbol{f}$ の推定値 $\boldsymbol{f}^*$ を求める．一般性を失うこと無く $\alpha = 1$ とおける．すなわち，符号，復号過程の順逆変換によって入力と再構成値との誤差を最小にするように学習させる．以下の手順で $\boldsymbol{w}_e$ と $\boldsymbol{w}_d$ とを推定する．

1. 入力 $\boldsymbol{x}$ を符号器と変換パラメータ $\boldsymbol{q}$ によって $\boldsymbol{f}_0 = f_e(\boldsymbol{x},\boldsymbol{q};\boldsymbol{w}_e)$ へと変換する．$\boldsymbol{q}$ は復号器へコピーする．
2. $\boldsymbol{q}$ を固定し $\boldsymbol{f}_0$ を初期値としてエネルギー関数 $H_d + \alpha H_e$ を $\boldsymbol{f}$ で微分し勾配降下法によって最適値 $\boldsymbol{f}^*$ を得るよう学習する．
3. 復号器のエネルギー:
$$\Delta\boldsymbol{w}_d \propto -\frac{\partial\|\boldsymbol{x} - f_d(\boldsymbol{f}^*,\boldsymbol{q};\boldsymbol{w}_d)\|^2}{\partial\boldsymbol{w}_d}$$
を少させるよう結合係数を 1 ステップ更新する
4. 推定した $\boldsymbol{f}^*$ を目的値として符号器のエネルギー関数を少させるよう結合係数を 1 ステップ更新する
$$\Delta\boldsymbol{w}_e \propto -\frac{\partial\|\boldsymbol{f}^* - f_e(\boldsymbol{x},\boldsymbol{q};\boldsymbol{w}_e)\|^2}{\partial\boldsymbol{w}_e}.$$

復号器は入力画像から計算した推定値 $\boldsymbol{f}^*$ を用いて最適な再構成値を生成するように学習が行われる．同時に，符号器は最適な予測値を出力することを学習する．学習が進行すれば，少ない繰り返しステップで $\boldsymbol{f}^*$ へ到達する．学習終了時には符号器は 1 ステップで最適値 $\boldsymbol{f}^*$ を出力し符号空間の探索を必要としない．ただし図 2 のように CNN の次に Max-pooling が入るので単純な符号復号の関係にはない．しかし初期の [LeCun 98] で採用されていたサブサンプリングより性能が向上した [Ranzato 07, Scherer 10]．

Max-pooling 層は特徴層の最大値を検出し，その値を出力した位置を保持する．値を出力した位置を保持は位置の指標である変換パラメータ (“where”) として作用する．Max-pooling は (1) 最大値のみを取り出し，他の値を捨てることで上位層の計算量を減じることができる．Max-pooling のこの特徴は同一空間には一つの対象物しか存在しないとする外界の拘束条件，視覚環境の制約を表現している．そして他に理由が存在しなければ最大値を出力した検出器を信頼する機構と解釈できる．また (2) 空間的に局在した入力信号の核関数との相関を保持し，(3) 全視野に渡って結合係数を共有するので計算効率が良い，という特徴が挙げられる．

2.2 AlexNet

2010 年に始まった大規模画像認識コンテスト (ILSVRC) で，2012 年に SVM [Haussler 92] に 10% 以上の差をつけて優勝した ディープニューラルネットワークが AlexNet である [Krizhevsky 12]. ILSVRC2012 課題中の上位 1 候補のみを挙げる (すなわち正解を解答する) 課題での誤判断率が 26.2%，上位 5 候補を挙げる課題で誤認識率は 15.3% であった. AlexNet は (1) 各ユニットの出力関数としての線形整

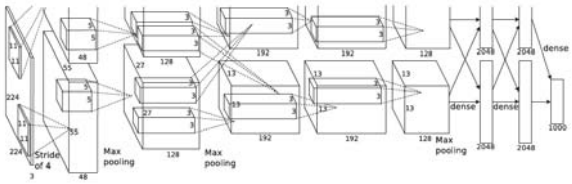

図 3 [Krizhevsky 12] 図 2 を改変

流ユニット (Rectified Linear Unit: ReLU), (2) 局所反応正規化 (Local Responese Normalization: LRN), (3) dropout, (4) オーバーラッププーリング，(5) GPU の複数使用を特徴とする.

出力関数には $f(x) = (1 + \exp(-x))^{-1}$ もしくは $f(x) = \tanh(x)$ が伝統的に用いられてきた．ReLU は $f(x) = \max(0, x)$ である．すなわち ReLU は入力が負であれば他のユニットへ何の貢献もしない．入力が正であればその信号を出力する．従って $x = \infty \Rightarrow f(x) = \infty$ となり飽和しない．そこで ReLU の出力 a を以下の局所的に規格化した b へと変換した.

$$b_i = \frac{a_i}{\left(\kappa + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a_j)^2\right)^{\beta}}, \qquad (2)$$

これを局所反応正規化 (Local Response Normalization: LRN) と呼ぶ．n は隣接核関数地図における隣接する核関数数，N は層内の総核関数数である．ハイパーパラメータは定数であり $\kappa = 2$, $n = 5$, $\alpha = 10^{-4}$, $\beta = 0.75$ であった．すなわち近傍の値を走査して自乗和し，その値を線形変換した値で規格化する．これは応答関数の値 a が過剰に大きくなる場合を抑制する効果と推察される．S 次曲線を仮定した $(1 + \exp(-x))^{-1}$ や $\tanh(x)$ では上下限値に飽和するので LRN を用いる必要がない．ReLU の利点は解釈容易性にある．人間にとっても機械にとっても，ユニット間の通信では，先行するユニットが不活性で，かつ，結合係数が負の時，否定証拠を否定する二重否定となって解釈が困難である．このことは比喩的な意味ばかりではない．二重否定に陥りそうな場合を単純に無視できる方が学習が容易である.

AlexNet では CNN と異なり，畳込み演算の領域を重複させた．s 画素毎にプーリングユニットのグリッドが構成され，各プーリングユニットは $z \times z$ の受容野を持つ．$s = z$ であれば通常のプーリングとなる．$s < z$ であればオーバーラッププーリングとなる．AlexNet では $s = 2$, $z = 3$ であった．オーバーラッププーリングにより過学習が押さえらえるとされる．オーバラッププーリングは冗長性を担保し，認識システムを頑健にするという生物の生存にとっても意味があると考える.

図 3 では畳み込み層までは上下 2 つの流れは 2 枚の GPU に対応している．第 3 層から第 4 層へは相互に情報が流れるが第 4 層から第 5 層からの通信は存在しなかった.

2.3 GoogleLeNet

GoogleLeNet は ILSVRC2014 への登録チーム名であり LeNet–5 [LeCun 98] へのオマージュであると書いてある．ILSVRC2014 のステートオブジアーツであるが，単純な CPU パワーに頼るのではなく，後述の R-CNN を組み込むことでパラメータの数の減少と成績向上とを実現した．GoogleLeNet の概要を図 4 に示した.

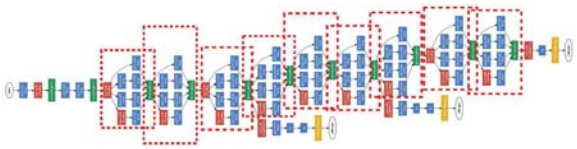

図 4 GoogleLeNet の構成 [Szegedy 15] 図 3 を改変．点線で囲まれた部分がすべて図 5 のインセプションモジュールである

ディープニューラルネットワークでは性能と規模との問題が生じる．ネットワークの深さ（層数）と幅 (層内ユニット数) を増やせば性能は向上する可能性はある．しかし，大規模データ，大規模ネットワークによる解には (1) パラメータ数の増大により過学習の恐れがあり，かつ (2) 計算コストが増大する，という短所がある．CPU や GPU の利用環境が整備されてきたとはいえ，今後，画像認識を越えて，動画，意図，感情，信念，意思，意味，などの高次認知機能の実現を考慮すれば，画像認識ばかりに資源を投入している訳にいかない．可能な限り労力の削減は必要である.

GoogleLeNet ではインセプション構造が採用された (図 5)．図 5 ではインセプションモジュール内で畳み

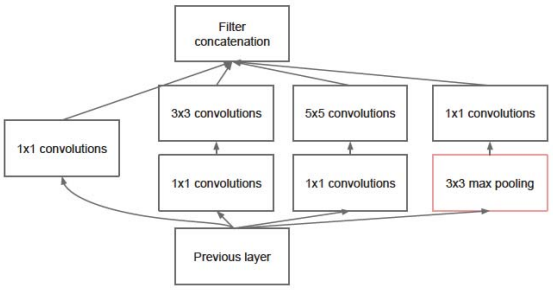

図 5　GoogleLeNet のインセプションモジュール [Szegedy 15] 図 2 を改変

込み演算が連結されている．これにより次元削減と投射の減少による疎性結合とを実現させた．素朴にサイズの異なる核関数と Max-pooling の結果を高次層へと送るよりも構造を作り込んだ方が総結合数は減少する．GoogleLeNet をまとめると以下のようになる．

1. 平均的なプーリング層の核関数の大きさは 5×5 で各核関数の間隔は 3 であった．
2. サイズ 1×1 のフィルターは次元圧縮のために容易された核関数で，ReLU が出力関数として採択された．
3. 完全結合層では 1024 ユニットが使われ，出力関数は ReLU であった．
4. dropout 層では確率 0.7 で出力を脱落させた．
5. softmax よる分類層では 1000 カテゴリーの分類が行われた．
6. 学習は非同期確率勾配降下法 (Stochastic gradient descent) でモーメント係数 0.9, 学習係数は 8 エポック毎に 0.04 減少させた．

2.4　Very Deep (VGG)

VGG は ILSVRC2014 に参加したチーム名 [1] であり，ローカライゼーション課題で 1 位，分類課題では 2 位であった．VGG は 16 層から 19 層の CNN で多層化の可能性を追求した．このため，他のパラメータを全て固定した．畳み込みの核関数としては 3×3 のみを用いた．入力画像は 224×224 ピクセルの RGB 画像に固定し，前処理は各画素の RGB 値から平均値を引いたのみであった．画像の各点は 2 次元格子状に配置されるので畳み込み演算の最小単位は 1×1 である．この場合近傍を考慮しない線形フィルタと同義である．一方画素の 8 近傍を走査する 3×3 は実質的な空間フィルタとしては最小のフィルタである．VGG のフィルタサイズは 3×3 に固定された．畳み込みストライド (フィルタ間隔) は 1 画素であった．畳み込み層への入力における空間充填 (spatial padding) も 1 画素であった．空間充填は 3×3 の空間フィルタに対して後処理において空間解像度を保証する効果がある．Max-pooling は 2×2 画素に対して実行されストライドは 2 であった．VGG はこのように局在，単純な構成で畳み込み層を積み上げ，3 から 5 層おきに Max-pooling 層を挟んで，最上位の 4 層は全結合，最終層は softmax 層で認識に至る．VGG では畳み込み層数 A から E まで 5 種類のアーキテクチャと LRN を介在させた A-LRN の全 6 ネットワークを採用した．

分類課題でステートオブジアーツの結果を示した前述の GoogleLeNet は VGG とは独立に開発されたが，同様の発想で 22 層，畳み込みフィルタサイズは 1×1, 3×3, 5×5 であるので類似したネットワークアーキテクチャと言える．学習時の重み崩壊係数 (L2 ペナルティ) は 5×10^{-4} であり，全結合層の dropout 率は 0.5 であった．

このような単純なネットワークを積み上げて多層化した GoogleLeNet と VGG とは人間の成績に肉薄した．上位 5 カテゴリを挙げる分類課題では GoogleLeNet の誤判別率 6.7%, VGG は 6.7%，人間は 5.1% である [Russakovsky 15]．その後，人間超え 4.94% の論文が arXiv に掲載された [He 15a][2]

2.5　MSRA

MSRA は ILSVRC2014 に参加したマイクロソフトのチームである．彼らのモデル名は SPP-net (空間ピラミッドプーリング Spatial Pyramid Pooling) である [He 15b]．対象検出で 2 位，画像分類で 3 位であった．SPP の概略を図 6 に示した．GoogleLeNet のインセプションモジュール (図 5) との違いは GoogleLeNet が空間的に局在化したモジュールである一方で SPP-net は AlexNet (図 3) の最終畳み込み層 (5 層) と全結合層 (6 層) との間に SPP (図 6) を挿入した構成である．第 5 層の出力を分割したプーリング結果を 6 層への入力信号としている．特徴地図は原理的に位置情報を伝達しないが局在化した情報であるので束ねれば位置情報の概要は伝達可能である．SPP はこの局在化した特徴地図が持つ位置情報が解釈可能となるようなプーリングを行ったと解釈可能である．SPP の介在によっ

[1] 現在は Google DeepMind

[2] 疑う報道もある (http://www.technologyreview.com/view/538111/why-and-how-baidu-cheated-an-artificial-intelligence-test/)

て小さな位置ズレではなく対象が画像上のどの位置にあっても頑健な認識が可能となる．

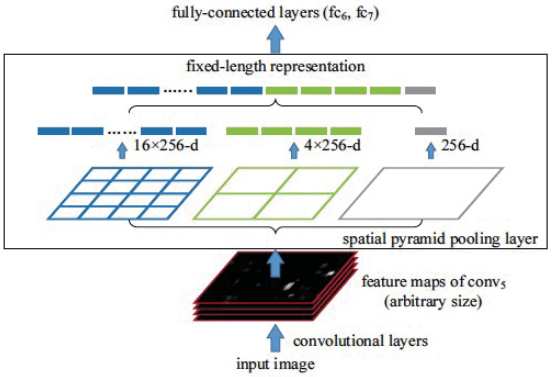

図 6　[He 15b] 図 3 を改変

図 7 左はフォークリストが縦列駐車している画像である．仮に下位層においてフォークリフト認識のための十分な特徴抽出が行われているのであれば，SPP とってフォークリフト複数台縦列駐車は有利な証拠が重なっていると見做しうる．一方，図 7 右は，ペーパーナイフが正解であるが，1. 万年筆，2. ボールペン, 3. 金槌 などと認識し類似した特徴を持つ物体に誤認識した．特徴地図の情報を空間的に束ねる SPP の操作が逆に類似特徴を持つ物体との誤認識率を高めてしまう例となっている．

図 7　[He 15a] 図 5 を改変．グランドトルースと出力とを示した．図左が正分類，図右が誤分類の例

2.6　Network-In-Network (NIN)

Network-in-Network(NIN) は [Lin 14] によって提案された Max-pooling の代替手法である．大域平均プーリング (Global average pooling) を用いる．Max-pooling が最大値を保持し他を捨てることに対して，出力情報に積極的な意味を持たせることを意図した Maxout [Goodfellow 13] がある．Maxout は局所線形近似を行う．しかし出力情報は非線形であっても対応可能なように拡張するとすれば，万能関数近似機である多層パーセプトロンを採用したモデルが NIN である．しかし Vapnik[Vapnik 71] が指摘したように局所的なアルゴリズムは考慮すべきであろう [Vapnik 95].

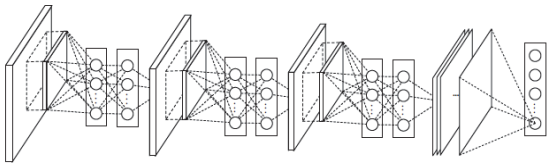

図 8　[Lin 14] 図 2 より

3.　Regional CNN (R-CNN)

SNS に時々刻々投稿される大量の静止画や動画は肖像画とは複雑さが異なる．多尺度かつ多物体の同時識別を行うためには領域切り出しを柔軟に行う必要がある．領域切り出しには，マルコフ確率場 (MRF), 条件確率場 (CRF) を用いたモデルや事前知識を前提とするモデルも提案されてきた．

自然言語処理における BOW (Bag-Of-Words) になぞらえて自然画像は BOVW (Bag-of-Visual-Words) で構成されている．この日常画像からの領域切り出しがの性能が向上してきた．領域の切り出しと物体の認識は相互依存の関係にある．ボトムアップで小領域を結合して矩形領域を切り出すか，トップダウンで画像の概形を捉えることから始めるかで幾種類かの手法が提案されている．心理学的，視覚情報処理と対応のとれるアルゴリズムをあれば，計算効率を追求した提案も存在する．

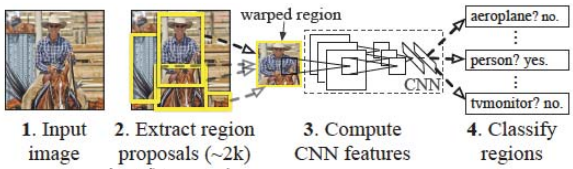

図 9　領域切り出しの R-CNN[Girshick 14] 図 1 より

Girshick ら [Girshick 14] はボトムアップに小領域の特徴を抽出して CNN への入力とし，SVM[Haussler 92] によって領域の分類を行った．

3.1　選択的探索

[Uijlings 13] は領域切り出しに選択的探索と呼ぶ機構を提案した．これはボトムアップ型の階層的領域分割であり，特定の位置，縮尺，物体の構造に依存しな

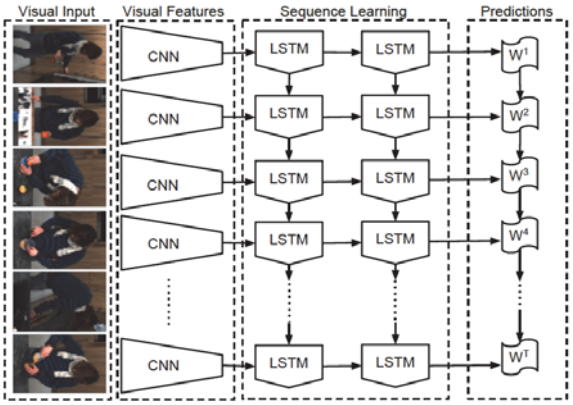

図 10 LRCN (Long Recurrent Convolutional Netowrks) の概念図 [Donahue 14] 図 1 より

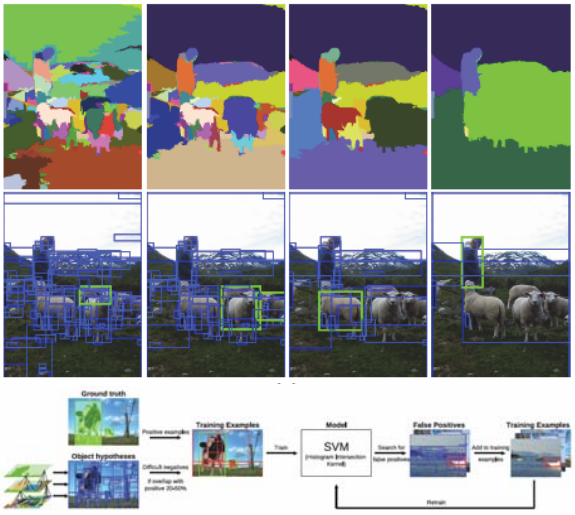

図 11 領域切り出し [Uijlings 13] 図 2, 3 を改変

い．色，テクスチャ，面積，外接矩形の情報を元に，物体が存在する可能性がある領域 “Object Hypothesis” を生成する．

R-CNN の問題点として，領域切り出しを行った後に，認識と領域の補正を行うため計算資源を消費することが挙げられる．また初期の領域切り出しに依存するため CNN だけでは精度が保証されない．これに対して [Long 15] は CNN のみを用いた完全結合の CNN モデル (Fully Connected Networks: FCN) を提案した．

領域切り出し後に CNN による識別過程を仮定するモデルだけでなく，CNN を用いて領域切り出しと対象検出とを同時に推定するモデルも提案されている [Hariharan 14]．R-CNN では原画像の領域の切り出しと畳み込み演算による画像のスキャンを複数回実施する．計算結果を共有すれば速度向上が期待できる [Girshick 15]．

3.2 顔と表情の認識 DeepFace

R-CNN の応用問題として，人物同定，表情認知 [Susskind 08, Taigman 14, Liu 13]，写真から脚注生成 [Vinyals 15, Fang 15] が挙げられる．

DeepFace [Taigman 14] では 4030 人分の顔画像 440 万枚が用いられた．人間に比肩しうる人物識別性能を達成した．DeepFace は検出した顔の矩形領域に対し

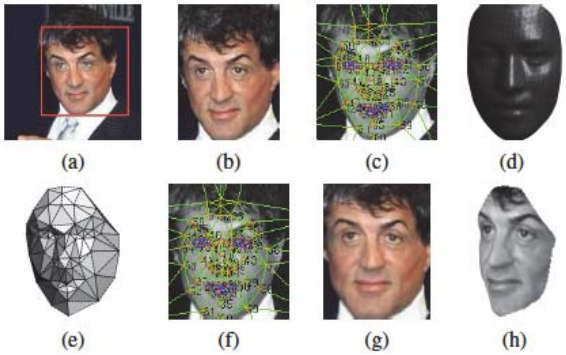

図 12 [Taigman 14] 図 1 より

て，(1) 矩形顔領域の 2 次元配置を確定する．(2) 3 次元モデルを用いて，切り出された顔画像平面を 3 次元画像へと再配置する．(3) 再配置された画像を ディープニューラルネットワークへの入力として顔表現ベクトルを得る．(4) 得られた顔表現ベクトルを比較することで同一人物か否かを判定する．を行った．矩形顔領域の 2 次元配置では，6 箇所の特徴（両目，鼻の先端，両口角，下唇の中心）を計算し，各位置が揃うように画像の位置と縮尺とが調整された．特徴点抽出は Local Binary Pattern と呼ばれる特徴を用いた Support Vector Regression が用いられた．3 次元への再配置は特徴点を 67 点に増やし，用意してあった標準顔の 3D モデル（平均顔）への対応する基準点を用いて特徴点の 3 次元位置を推定した。推定された 3 次元上の特徴点から，カメラ平面上に射影するカメラ行列を推定し，正面正立画像に変換した場合の特徴点の位置を求めた．加えて特徴点を頂点とする多角形（各セルは三角形）を生成し，元画像の画素値を正面正立画像へと三角形上でアファイン変換した。DeepFace では入力画像の顔の位置を独自に計算する．従って高次の位置不変特徴量が必要ではない．そこで DeepFace は 3 層以上の高次層では異なる結合係数行列が用いられた．また 1 層目以降では Max-pooling も行われなかった．結果は LFW (Labeled Faces in the

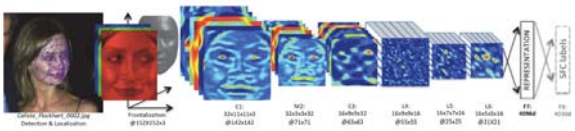

図 13　[Taigman 14] 図 2 より

Wild) データセットで 97.25% の正解率を得た．この正解率は人間の成績 97.53% とほぼ同等であった．誤認識した画像の中には加齢によって大幅に風貌が変化した老人など，人間でも判断が困難な画像が含まれていた．

図 14 には [Fan 14] の用いたピラミッドネットワークの概略が示されている．ネットワークの訓練には

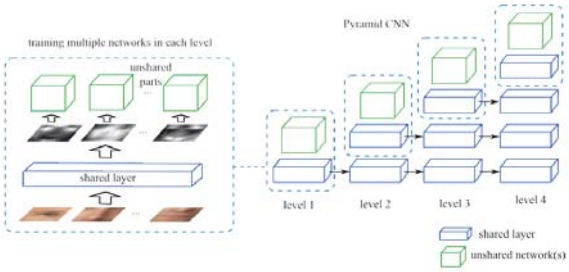

図 14　ピラミッド CNN の概念図 [Fan 14] 図 2 より．左が通常の CNN，右がピラミッド CNN で最上位に描かれている正六面体の結合係数行列は各レベルで共有されない．

シャム猫ネット [Bromley 94] が使用された [3]．2 つの画像が同じ CNN に提示され，出力は 2 つの入力顔が同じ ID を持っているかを予測する出力層が比較して判定した．ピラミッド CNN のネットワークには階層が存在する．ピラミッド CNN には入力サイズが異なる層が存在し，層の一部を共有する。ピラミッド CNN は greedy な訓練がなされた．学習初期ではネットワークは顔の一部を訓練される．その後第 1 層の結合係数は固定される．固定された層は，高次層のフィルタリング，ダウンサンプリング入力に用いらえた．高次ネットワークはあらかじめ低次層で処理された画像上で訓練を受ける．従ってネットワークは層数の増加によって計算コストが発生しない．結果は LFW データセットで 97.3%の認識率（当時ステートオブジアーツ）を示した [4]

以上まとめると自然画像から CNN を経由して認識に至るモデルの精度向上には刮目すべきであるが，依然確立された標準的な手法が存在しない．[Donahue 14] は状況に応じて手法の組み合わせを変更することを考えている (図 15)．どのような状況で

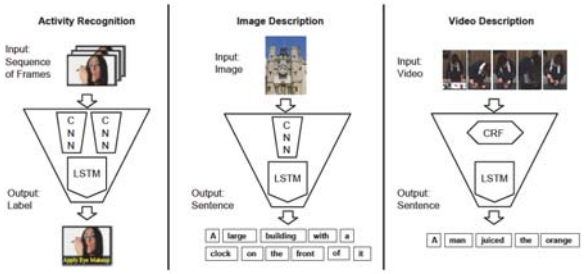

図 15　[Donahue 14] 図 3 を改変

どのようなモジュールを組合わせるのが良いかについては背景となる機構をどのように設定すれば良いのだろうか．

4.　高次認知への示唆

CNN の枠組みで多重解像度，すなわち，階層的な概念構造を表象可能であろう．特徴処理機構 “what” と位置情報 “where” とを分離しうるので，腹側径路と背側径路とによる視覚情報処理に類比しうる．両経路の相互作用と R-CNN による矩形切り出しの修正との相互作用を神経心理学に換言して考える手法が提案できる．すなわち物体失認，無視，相貌失認，などの神経心理学的症状は直接対応がとれるものと思われる．高次認知，概念発達，進行性の病変による意味記憶の崩壊をディープラーニングとの対比から議論可能であろう．基本概念は獲得年齢が低く反応時間も速い (基本概念優位性 [Rosch 76])．一方，意味痴呆患者の動物と人工物の二重乖離，上位概念の頑健性が知られている [Warrington 75]．CNN+Max-pooling を心的演算の相等物と見做せば，概念の抽象化が畳み込み積分で位置不変な特徴抽出に対応し，抽象概念の具体的な例示が下位層への逆照射と見做すことが可能であろう．進行性の意味痴呆患者において，基本概念よりも上位概念が保たれる現象は，基本概念を構成する要素が抜け落ちたとしても Max-pooling において，なお候補となる概念が活性化すると考えれば良い．

人間の情報処理機構に触発されたディープラーニングの諸手法が翻って人間のモデルとして考え得る可能性が指摘できる．図 15 に示されたように概形が類似してはいるが解くべき問題の精度と外界刺激の性質とに依存して生体の採用する機構が変動し，変動は許容範囲内でパラメータの調整を行い適応する．個体が選択した機構によって外部環境が変化するので，直面する課題や障害も多様となる．予め獲得した機構と可能

[3] 構成の同じ双ニューラルネットワークに対して 2 つの入力の異同を学習させるネットワークを双猫ネット，またはシャム猫ネットと呼ぶ．

[4] 開発は FACE++, http://www.faceplusplus.org

な範囲のパラメータ再調整も見極めることは教育，職業訓練，病理診断，リハビリテーション，QOL に有益な示唆を与えるだろう．

参考文献

[Bromley 94] Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R.: Signature Verification using a "Siamese" Time Delay Neural Network, in Cowan, J., Tesauro, G., and Alspector, J. eds., *Advances in Neural Information Processing Systems 6*, pp. 737–744, Morgan-Kaufmann (1994)

[Chatfield 14] Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A.: Return of the Devil in the Details: Delving Deep into Convolutional Nets, in *British Machine Vision Conference* (2014)

[Donahue 14] Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalany, S., Saenko, K., and Darrell, T.: Long-term Recurrent Convolutional Networks for Visual Recognition and Description, . (2014)

[Fan 14] Fan, H., Cao, Z., Jiang, Y., Yin, Q., and Doudou, C.: Learning Deep Face Representation, *CoRR*, Vol. abs/1403.2802, (2014)

[Fang 15] Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Dollar, P., Gao, J., He, X., Mitchell, M., Platt, J., Zitnick, L., and Zweig, G.: From Captions to Visual Concepts and Back, in *The proceedings of CVPR, IEEE Institute of Electrical and Electronics Engineers*, Boston, MA, USA (2015)

[Fukushima 82] Fukushima, K. and Miyake, S.: Neocognitron: A New Algorithm for Pattern Recognition tolerant of Deformations and Shifts in Position, *Pattern Recognition*, Vol. 15, pp. 455–469 (1982)

[Girshick 14] Girshick, R., Donahue, J., Darrell, T., and Malik, J.: Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, in *Proceedings of Computer Vision and Pattern Recognition Conference (CVPR)*, Columbus, Ohio, USA (2014)

[Girshick 15] Girshick, R.: Fast R-CNN (2015)

[Goodfellow 13] Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A., and Bengio, Y.: Maxout Networks, *arXiv:1302.4389v4, stat.ML* (2013)

[Hariharan 14] Hariharan, B., Arbeláaez, P., Girshick, R., and Malik, J.: Simultaneous Detection and Segmentation, in *Proceedings of the European Conference on Computer Vision*, Zurich, Switterland (2014)

[Haussler 92] Haussler, D. ed.: *A Training Algorithm for Optimal Margin Classifiers*ACM press (1992)

[He 15a] He, K., Zhang, X., Ren, S., and Sun, J.: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, Technical report, Microsoft (2015)

[He 15b] He, K., Zhang, X., Ren, S., and Sun, J.: Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2015*, pp. 1–1 (2015)

[Hubel 59] Hubel, D. and Wiesel, T. N.: Receptive Fields of Single Neurones in the Cat's Striate Cortex, *Journal of Physiology*, Vol. 148, pp. 574–591 (1959)

[Hubel 68] Hubel, D. and Wiesel, T. N.: Receptive Fields and Functional Architecture of Monkey Striate Cortex, *Journal of Physiology*, Vol. 195, pp. 215–243 (1968)

[Karayev 14] Karayev, S., Trentacoste, M., Han, H., Agarwala, A., Darrell, T., Hertzmann, A., and Winnemoeller, H.: Recognizing Image Style, in *Proceedings of the British Machine Vision Conference*, BMVA Press (2014)

[Kemp 09] Kemp, C. and Tenenbaum, J. B.: Structured Statistical Models of Inductive Reasoning, *Psychological Review*, Vol. 116, No. 1, pp. 20–58 (2009)

[Krizhevsky 12] Krizhevsky, A., Sutskever, I., and Hinton, G. E.: ImageNet classification with deep convolutional neural networks, in Bartlett, P. L., Pereira, F. C. N., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. eds., *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems*, Lake Tahoe, Nevada, USA (2012)

[LeCun 98] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P.: Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, Vol. 86, pp. 2278–2324 (1998)

[Levine 15] Levine, S., Finn, C., Darrell, T., and Abbeel, P.: End-to-End Training of Deep Visuomotor Policies, Technical report, Berkely Vison and Learning Center BVLC, Report Series 100 (2015)

[Lin 14] Lin, M., Chen, Q., and Yan, S.: Network In Network, *arXiv:1312.4400v3* (2014)

[Liu 13] Liu, M., Li, S., Shan, S., and Chen, X.: AU-aware Deep Networks for facial expression recognition, in *Automatic Face and Gesture Recognition (FG) on the 10th IEEE International Conference and Workshops*, pp. 1–6, Shanhai, China (2013), IEEE

[Long 15] Long, J., Shelhamer, E., and Darrell, T.: Fully Convolutional Networks for Semantic Segmentation, *Computer Vision and Pattern Recognition (CVPR)* (2015)

[Osherson 90] Osherson, D. N., Wilkie, O., Smith, E. E., and López, A.: Category-based induction, *Psychological Review*, Vol. 97, No. 2, pp. 185–200 (1990)

[Ranzato 07] Ranzato, M., Huang, F., Boureau, Y., and LeCun, Y.: Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition, in *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, Minnesota, USA (2007)

[Rogers 04] Rogers, T. T. and McClelland, J. L.: *Semantic Cognition: A Parallel Distributed Processing Approach*, The MIT press, Cambridge, MA (2004)

[Rosch 76] Rosch, E., Mervis, C. B., Gray, W. D., M, D., and Boyes-braem, P.: Basic objects in natural categories, *Cognitive Psychology* (1976)

[Russakovsky 15] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge, *International Journal of Computer Vision* (2015)

[Scherer 10] Scherer, D., Müller, A., and Behnke, S.: Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition, in *20th International Conference on Artificial Neural Networks (ICANN)*, pp.

92–101, Thessaloniki, Greece (2010)

[Serre 05] Serre, T., Wolf, L., and Poggio, T.: Object Recognition with Features Inspired by Visual Cortex, in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 994–1000, San Diego, CA, USA (2005)

[Susskind 08] Susskind, J. M., Anderson, A. K., Hinton, G. E., and Movellan, J. R.: Generating Facial Expressions with Deep Belief Nets, in Or, J. ed., *Affective Computing*, chapter 23, pp. 421–440, INTECH Open Access Publisher (2008)

[Szegedy 15] Szegedy, C., Liu, W., Jia, Y., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A.: Going deeper with convolutions, in *Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA (2015)

[Taigman 14] Taigman, Y., Yang, M., Ranzato, M., and Wolf, L.: DeepFace: Closing the Gap to Human-Level Performance in Face Verification, in *Proceedings of Computer Vision and Pattern Recognition Conference (CVPR)*, Columbus, Ohio, USA (2014)

[Uijlings 13] Uijlings, J. R. R., Sande, van de K. E. A., Gevers, T., and Smeulders, A. W. M.: Selective Search for Object Recognition, *International Journal of Computer Vision*, Vol. 104, No. 2, pp. 154–171 (2013)

[Vapnik 71] Vapnik, V. N. and Chervonenkis, A. Y.: On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities, *Theory of Probability and Its Applications*, Vol. 16, No. 2, pp. 264–280 (1971)

[Vapnik 95] Vapnik, V. N.: *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, NY, USA (1995)

[Vinyals 15] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D.: Show and Tell: A Neural Image Caption Generator, in *Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA (2015)

[Warrington 75] Warrington, E. K.: The Selective impairment of semantic memory, *Quarterly Journal of Experimental Psychology*, Vol. 27, pp. 635–657 (1975)

[Zhou 14] Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A.: Learning Deep Features for Scene Recognition using Places Database, in Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. eds., *Advances in Neural Information Processing Systems 27*, pp. 487–495, Curran Associates, Inc. (2014)