Vol.2014-CVIM-192 No.14
2014/5/15

Convolutional Neural Network を用いた一般物体認識手法の解析

石井 智大[1,a)]　望月 義彦[1,b)]　小山田 雄仁[1,c)]　石川 博[1,d)]

概要：一般物体認識では，近年 Deep Learning を用いた手法が注目されており，その 1 つである Convolutional Neural Network （CNN） は特に優れた結果を示している．しかし，どのような構成の CNN が画像認識に有用であるかは理論的に示されておらず，ノウハウが必要なのが現状である．本研究では CNN を用いた一般物体認識手法において認識精度を変化させる要因の解析を行う．具体的には，Krizhevsky らの手法において，畳込み層のパラメータが認識精度に与える影響を解析するとともに，学習手法の変更が認識精度に与える影響を調べた．

1. はじめに

1.1 研究背景

一般物体認識は，実世界における画像に対して，その画像に含まれる物体を「山」や「スマートフォン」などの一般的な名称で認識することである．一般物体認識の主なタスクとしては，画像全体に 1 つのカテゴリーを付与する画像全体のカテゴリー分類，複数のカテゴリーラベルを画像に付与する画像アノテーション，領域分割された画像の各領域にカテゴリーラベルを付与する画像ラベリング，長方形の矩形で画像中の物体の存在位置を検出するカテゴリー物体検出，物体の領域を正確に切り出すカテゴリー領域抽出が存在する [1].

画像全体のカテゴリー分類は，局所特徴量を用いた Bag-of-features [2] や Fisher ベクトル [3] などの画像表現手法の提案によって研究が進歩している．また，2012 年に大きなインパクトを与えた研究として，Deep Learning という機械学習の方法を用いた研究がある．2012 年開催の一般物体認識のコンテスト ILSVRC（Imagenet Large Scale Visual Recognition Challenge）において，Krizhevsky らが提案した Deep Learning を用いた認識手法 [4] は，従来の Bag-of-features[2] や Fisher ベクトル [3] を用いた認識手法を超えた性能を示している [5].

Deep Learning は基本的に多層のニューラルネットワーク （以下 NN） を用いた方法であり，複数の NN の構成が含まれる [5]．たとえば，音声認識では全結合型 NN が用いられることが多いのに対して，画像認識では Convolutional Neural Network （以下 CNN） が用いられることが多い．

CNN は NN の層の構成やパラメータの設定が学習に大きく影響を及ぼすと言われているが，どのような構成の CNN が画像認識に有用であるかは理論的に示されておらず，ノウハウが必要なのが現状であるとされている [5]．そのため，認識精度の違いも NN の層の構成やパラメータの設定に大きく影響を受けるものであり，NN の層の構成やパラメータの設定を変更していくことで認識精度との関係性を見つけられるのではないかと考えられる．実際に，Krizhevsky らの手法 [4] で示されている CNN も最適なネットワーク構造ではなく，ILSVRC2013 では Zeiler らの手法 [6] などの複数の研究によって Krizhevsky らのネットワーク構造が改良され，認識精度が向上している．

また，CNN による画像認識では学習画像数が少ないと，学習画像に対してはきちんと学習されるが，未知画像（テスト画像）に対しては適合できていない，過学習の状態になることがある．その対策としては，学習画像をランダムに左右反転させて学習させることで，実質的な学習画像の数を増やすことが行われている [4].

1.2 研究目的

本研究では画像全体のカテゴリー分類を対象とし，CNN を用いた一般物体認識手法において認識精度を変化させる要因を解析する．また，一般物体認識に用いられる Deep Learning の手法も複数存在するが，本研究では，画像認識で優れた結果を残している CNN を対象とする．

CNN はネットワーク構造の設計がノウハウの領域とさ

[1] 早稲田大学大学院 基幹理工学研究科 情報理工・情報通信専攻
[a)] tomohiro.ishii@asagi.waseda.jp
[b)] motchy@aoni.waseda.jp
[c)] oyamada@aoni.waseda.jp
[d)] hfs@waseda.jp

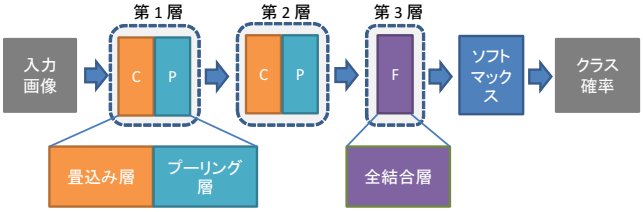

図 1 CNN の全体図．C は畳込み層，P はプーリング層を表し，F は全結合層を表す．

れており，理論的には示されていない．しかし，既存研究には一般物体認識において良い認識率を示す CNN のネットワーク構造が存在する．本研究では，ILSVRC2012 における画像全体のカテゴリー分類で最も良い認識率を示した手法 [4] と ILSVRC2013 で良い認識率を示した手法 [6] を参考にする．

2. CNN の概要

2.1 Neural Network

NN において，第 k 層の i 番目のユニットの出力 h_i^k は以下のように計算される [5].

$$h_i^k = f(b_i^k + \sum_j w_{ij}^k h_j^{k-1}). \quad (1)$$

ここで，f は活性化関数，b_i^k はバイアス，h_j^{k-1} は第 $k-1$ 層の j 番目のユニットの出力，そして，w_{ij}^k は第 $k-1$ 層の j 番目のユニットの出力が第 k 層の i 番目のユニットに入力する際にかかる重みである．

2.2 Convolutional Neural Network

CNN は畳込みとプーリングの 2 つの計算を交互に繰り返す順伝播型のネットワークである．

図 1 に基本的な CNN の構造を示す．始めは入力画像に対するフィルタの畳込み（後述）を行う層と，その出力に対するプーリング（後述）を行う層を交互に何度か繰り返す（図中 1，2 層目）．最終層（図中 3 層目）はタスクに応じた数のユニットを置き，その前の何層かは全結合層とすることが多い．そして，最終層の出力にソフトマックス（後述）を適用し，入力画像のクラス確率を得る．**図 2** は，図 1 のネットワーク構造における NN の結合の概念図である．

CNN は 式 (1) における h_j^{k-1} から h_i^k を計算する点で基本的な NN と同様であるが，その計算方法が畳込みやプーリングなどで行われる点が異なっている．

本研究では，CNN の構成要素のうち畳込み層のフィルタについて，そのサイズと数を変更し解析を行う．

なお，本論文での層の数え方としては，プーリング層と正規化層は層の数に数えず，畳込み層と同一の層として考える．

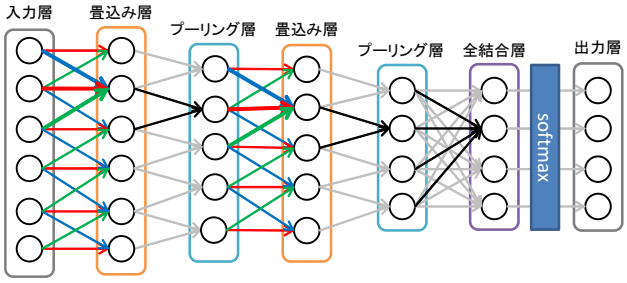

図 2 CNN の概念図．畳込み層は結合重みを共有しており，図中の線の色は同じ結合重みをもっていることを意味している．プーリング層は畳込み層の内の小領域から出力を得てプーリングを行う．全結合層は隣り合う層のすべてのユニットと結合する．

2.2.1 畳込み

畳込み層は特定のフィルタを入力画像に畳込む処理を行う層である．画素数 $n_x \times n_x$ の入力画像 $\boldsymbol{x}$ に対する，画素数 $n_w \times n_w$ のフィルタ $\boldsymbol{w}$ の畳込みを考える．この畳込みの出力を $\boldsymbol{h}$ として，$\boldsymbol{h} = \boldsymbol{x} * \boldsymbol{w}$ と書く．ここで，$*$ は画像の畳込みを表す．このとき，出力 $\boldsymbol{h}$ のサイズは一般的に $n_h \equiv n_x - n_w + 1$ となるが，画像周囲にパディングを行うことで，$n_h = n_x$ とする場合も多い [5]．また，CNN はフィルタを複数持っており，フィルタごとに異なる出力のセット（マップと呼ばれる）を持つ．図 2 では，入力画像の画素が NN の 1 つのユニットに相当し，畳込むフィルタがユニットにかかる重みに相当する．CNN の特徴は，層間の結合がフィルタの範囲の位置ごとに局所的に行われ，その結合重みが上位層のユニット間で共有されていることである．実際の処理においては，フィルタのサイズや数のほかに，フィルタを動かすストライド（間隔）を決める必要がある．

2.2.2 プーリング

プーリング層は，畳込みの出力に対しプーリングを行う層である．プーリングとは，抽出した特徴から認識に余分な情報を捨て，認識に必要な情報を保った新たな表現に変換することと定義されている [7]．本研究で用いるプーリングは，畳込みの出力であるマップの小領域ごとから最大値を取り出す．このプーリングはマックスプーリングと呼ばれ，プーリング層のユニット i の出力 h_i' が小領域 P_i についてプーリングを行ったものとすると，

$$h_i' = \max_{j \in P_i} h_j \quad (2)$$

と表される．本研究では，プーリングを行う小領域のサイズをプーリングのサイズと呼ぶ．また，実際の処理においては，畳込みと同様に，プーリングのサイズのほかに，プーリングのストライドを決定する．

2.2.3 局所コントラスト正規化

CNN で用いられる特徴的な処理として，畳込みとプーリング以外に局所コントラスト正規化 [7] がある．CNN の畳込みでは画像にフィルタを畳込み，出力としてマップを

得るが，そのマップは畳込んだフィルタの数だけ存在する．それらのマップ間で画素ごとに正規化を行う．

2.2.4 全結合

全結合層は，前の層のすべてのユニットと結合する層であり，図 2 に示す様な形となる．

2.2.5 ソフトマックス

CNN を用いてクラス分類を行う場合，最終層に分類すべきクラスと同数のユニットを配置し，あるクラスに対応する画像が入力された場合は，そのクラスに対応するユニットの出力が大きくなるようにすることでクラス分類を行う．実際には最終層のそれぞれのユニットの出力 $\{h_i\}$ に対してソフトマックスを用い，クラス確率 $\{p_i\}$ に変換する．ソフトマックスは以下のように定義される．

$$p_i = \frac{\exp(h_i)}{\sum_j \exp(h_j)}. \tag{3}$$

ソフトマックス関数はマックス関数を滑らかにしたものであり，すべての $i \neq j$ に対して，$h_i \gg h_j$ である場合には，$p_i \simeq 1$ であり，$p_j \simeq 0$ となる．

2.2.6 学習最適化

CNN の学習最適化は，CNN の出力が望んだものとなるように各層の重みを最適化することである．最適化を行う際は，学習画像を CNN に入力した際の出力と望みの出力の差異を以下の様なコスト C として定義し，そのコストを最小化することを考える．

$$C = -\sum_i d_i \log p_i. \tag{4}$$

ここで，入力に対して正しいクラスを k とするとき，$d_k = 1$ でそれ以外は $d_i = 0$ とする．学習画像が複数ある場合は学習画像ごとに C を計算し，それらの総和をコストとする．一般的には学習画像を一定数集めたミニバッチを作り，ミニバッチごとにコストを求めて学習を行う．

学習には確率的勾配降下法を用いる．各層の重み w_{ij} について，以下に示すように，ミニバッチ t に対する重みの修正量 $\Delta w_{ij}(t)$ を，直前のミニバッチ $t-1$ に対する修正量 $\Delta w_{ij}(t-1)$ を用いて滑らかにすることを行う．

$$\Delta w_{ij}(t) = \alpha \Delta w_{ij}(t-1) - \epsilon\lambda w_{ij}(t-1) - \epsilon \frac{\partial C}{\partial w_{ij}(t)}. \tag{5}$$

$$w_{ij}(t) = w_{ij}(t-1) + \Delta w_{ij}(t). \tag{6}$$

ここで，式 5 の第 1 項の α $(0 < \alpha < 1)$ はモメンタムと呼ばれ，平滑化を担う．第 2 項は正則化項であり，λ は任意のパラメータである．また，第 2，3 項の ϵ は更新のステップ幅であり，学習係数と呼ばれる．

バイアスについても同様に学習を行うが，上述の正則化項は含まない．

2.2.7 認識精度の定義

本論文において認識精度を示す指標として認識率と誤認識率を定義する．認識率は，テスト画像を CNN に入力した際に出力されるクラス確率 $\{p_i\}$ のうち最も確率の大きいクラスがどれくらい正解のクラスと一致したかを示すものとする．具体的には，個々のテスト画像について最もクラス確率の大きいクラスが正解クラスであった場合は 1 とし，それ以外の場合は 0 とした上で，全てのテスト画像の場合の総和を取って平均したものとする．ここで，正解であったテスト画像のうち，稀に最もクラス確率の大きいクラスが複数存在する場合がある．その場合は，正解した度合を小さくすることを考え，具体的には，1 を最大のクラス確率のクラス数で平均したものとする．また，誤認識率は，1 から認識率を引いたものとする．

2.3 CNN の一般物体認識への適用

Krizhevsky らの手法 [4] は一般物体認識のコンテスト ILSVRC2012 において，CNN を用いた認識手法で従来の特徴量を用いた手法を大きく上回る性能を示した．これは CNN を最初に一般物体認識へ適用したものではないが，CNN の有用性を示すという点で大きな影響を与えた手法である．以下ではその要素のうち本研究にかかわるものについて述べる．

2.3.1 ReLU Nonlinearity

NN の活性化関数 f には，シグモイド関数を用いるのが一般的である．しかし，勾配降下法による学習時間の観点からすると，シグモイド関数よりも活性化関数 f に以下に定義される修正線形関数を用いることで学習を高速化できる．

$$f(x) = \max(0, x). \tag{7}$$

2.3.2 Local Response Normalization

Local Response Normalization は，修正線形関数の様な制限のない活性化関数のユニットを使用する際に有用であるとされている正規化の手法であり，以下のように定義される．

$$b^i_{x,y} = \frac{a^i_{x,y}}{(k + \alpha \sum_{j=\max(0, i-\frac{n}{2})}^{\min(N-1, i+\frac{n}{2})} (a^j_{x,y})^2)^\beta}. \tag{8}$$

ここで，$a^i_{x,y}$ はマップ i 上の座標 (x, y) のユニットの出力，N は層にあるマップの総数，n は隣接するマップの数である．また，k, α, β は任意のパラメータである．

2.3.3 Overlapping Pooling

プーリングは，プーリングを行う領域が重複しないように行うのが一般的であるとされていた．しかし，領域の重複を許すことで認識率を向上できることが実験的に確認されている．

2.3.4 Data Augmentation

CNN による画像認識では学習画像数が少ないと，学習画像に対してはきちんと学習されるが，未知画像 （テスト

画像）に対して適合できていない，過学習の状態になることがある．その対策として，学習画像の並進移動によるパッチの切り出しと左右反転を行って学習させることで，実質的な学習画像の数を増やしている．

2.4 CNN の可視化による分析

Zeiler らの手法 [6] は，CNN の各層で生成されるマップに現れる入力画像の特徴を明らかにすることで CNN の可視化を行ったものである．また，Krizhevsky らのネットワーク構造 [4] で学習した CNN に対して可視化を行うことで，その問題点を発見・改善し，認識精度の向上を達成している．

Zeiler らの手法で変更された点として，第 1 層のフィルタのサイズとストライドがある．Zeiler らは，Krizhevsky らのネットワーク構造を学習し可視化を行ったところ，いくつかの問題点が存在したことを示している．その問題点の 1 つは第 1 層のフィルタに関するもので，学習されたフィルタの中に特徴抽出の役割を果たさない，単色で変化のないフィルタが存在したことである．また，第 2 層の可視化を行ったところ，エイリアシングを含んだものが存在したことである．エイリアシングに関しては，第 1 層のフィルタのストライドが 4 と大きいことが原因とされている．その対策として，第 1 層のフィルタのサイズを 11×11 から 7×7 にし，また，ストライドを 4 から 2 にすることでこの問題を解決することが出来たとされている．

3. ネットワーク構造の変化とデータセットの画像の変形による認識精度の解析

CNN を用いた一般物体認識の従来手法 [4], [6] から，認識精度に影響を与える複数の要因が考えられる．1 つ目として，畳込み層のフィルタやストライドなどのパラメータが挙げられる．畳込み層のフィルタサイズとストライドを改良することで，画像の特徴をより適切に抽出できるようになったとされている [6]．これより，畳込み層で抽出される入力画像の特徴が重要であることが分かる．2 つ目としてはデータセットの拡張が挙げられる．学習画像のランダムな左右反転と並進移動によるパッチの切り抜きを行うことで認識精度が向上しており [4]，データセットの拡張は有益であると考えられる．CNN を用いた研究においては，画像の回転を加えることで精度を向上させているものもあるため [8]，左右反転と並進移動に回転を加えることで，認識精度の向上につながることが考えられる．

3.1 実験目的

本実験は，CNN のネットワーク構造の変化とデータセットの画像の変形を行い，認識精度への影響を解析することを目的とする．

前述したように，CNN は ネットワーク構造が認識精度に大きく影響を及ぼすとされているが，具体的にどのような構成が有効であるかは理論的に示されておらず，ノウハウが必要とされている．そのため，ネットワーク構造を変化させた際の認識精度への影響を調べることで，ノウハウを蓄積を行う．

また，NN による画像認識では学習画像数が少ないと，学習画像に対してはきちんと学習されるが，未知画像に対しては適合できていない，過学習の状態になることがある．その対策として，学習画像のランダムな左右反転と並進移動にパッチの切り抜きに加えて，画像の回転を行うことで，認識精度への影響を調べる．

表 1 実験環境

CPU	Intel（R） Core（TM） i7-3770K CPU @ 3.50GHz × 8
RAM	16GB
GPU	NVIDIA GeForce GTX680 （2048MB GDDR5）

3.2 実験方法

3.2.1 実験環境

本研究において使用した実験環境を，**表 1** に示す．

3.2.2 データセット

実験対象として，画像認識の分野で広く用いられている CIFAR-10 を使用する．CIFAR-10[9] は，Krizhevsky らによって作成されたデータセットであり，10 種類のカテゴリーについて，画素数 32×32 の 60000 枚の画像から構成されている．それぞれのカテゴリーについては 6000 枚ずつ画像が用意されている．全体の 50000 枚が学習画像として用いられ，残りの 10000 枚がテスト画像として用いられる．

3.2.3 cuda-convnet

今回の実験では，cuda-convnet[*1] を用いて実験を行う．cuda-convnet は Krizhevsky らによって公開されているソフトウェアであり，Krizhevsky らの手法 [4] で用いられたものである．cuda-convnet では，本論文で示している CNN のネットワーク構造のパラメータを任意に設定して実験を行うことが出来る．また，CIFAR-10 用に調整された CNN のネットワーク構造が提供されており，CIFAR-10 については 約 11 % の誤認識率で認識することができる．このネットワーク構造は [4] でも用いられているものであり，本実験では，この提供されているネットワーク構造を基に CNN の解析を行っている．

3.2.4 基とするネットワーク構造

本実験では基本とするネットワーク構造を用意し，その構造に変化を加えることで変化による影響を調査する．基本とするネットワーク構造としては，Krizhevsky らの手法 [4] で CIFAR-10 の実験に用いられ，誤認識率が 11% であったものを用いる．**図 3** はこのネットワーク構造を表し

[*1] https://code.google.com/p/cuda-convnet/

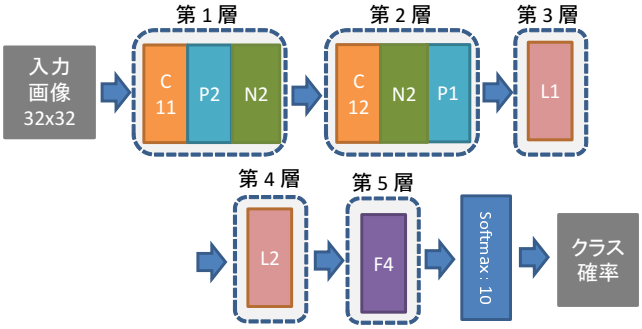

図 3 CIFAR-10 の誤認識率 11% のネットワーク構造．C は畳込み層，P はプーリング層，N は正規化層を表し，F は全結合層を表す．

表 4 全結合層のネットワーク構造のパラメータ．図 3 に示す CNN の全結合層のネットワーク構造を示す．

層番号	ユニットの数	活性化関数
F4	10	-

たものである．また，図 3 のネットワーク構造のパラメータについては，畳込み層，プーリング層，全結合層をそれぞれ，**表 2**，**表 3**，**表 4** に示す．図 3 において，L1，L2 で表記している層は畳込み層の一種であり，通常の畳込み層と違って重みを共有しない層である．

また，図 3 において，N2 と表記している層は，Local Response Normalization を行っているが，[4] において示されている計算式とは定義が異なっている．cuda-convnet で実装されている Local Response Normalization の定義は，

$$b^i_{x,y} = \frac{a^i_{x,y}}{(1 + \frac{\alpha}{n}\sum_{j=\max(0,i-\frac{n}{2})}^{\min(N-1,i+\frac{n}{2})}(a^j_{x,y})^2)^\beta} \qquad (9)$$

である．ここで，今回の構成においては，$n = 9, \alpha = 0.001, \beta = 0.75$ である．

3.2.5 ネットワーク構造の変化

本実験で行うネットワーク構造の変化は以下の通りである．

(1) プーリング （P） と正規化 （N） の順序
(2) 畳込み層のフィルタの数
(3) 畳込み層のフィルタのサイズ
(4) 学習画像の回転
(5) (3) と (4) のそれぞれについてフィルタの数との関係

プーリングと正規化の順序に関して，本実験で用いるネットワーク構造は，第 1 層ではプーリングの後に正規化を行っているのに対して，第 2 層では正規化の後にプーリングを行っている．しかし，Krizhevsky らの手法 [4] では正規化→プーリングの順に統一されている．そのため，本実験では，正規化とプーリングの順序を統一したネットワーク構造の実験を行うことでその順序による認識精度の違いを比較する．

表 7 重みの更新式 (式 (5)) のパラメータ．

層番号	ε	α	λ
C11	0.001	0.9	0.00
C12	0.001	0.9	0.00
L1	0.001	0.9	0.04
L2	0.001	0.9	0.04
F4	0.001	0.9	0.01

畳込み層のフィルタ数に関しては，**表 5** に示すように変更する．また，表 5 に示す変更を行ったネットワーク構造について，それぞれ学習画像にランダムで左右に 10 度回転する操作を加えた実験を行う．

畳込み層のフィルタサイズとフィルタ数に関しては，**表 6** に示すように変更する．

ここで，プーリングと正規化の順序の変更実験について，プーリング （P） と正規化 （N） の順序は，N → P または P → N と表記する．また，フィルタの数の変更と回転操作の実験について，ネットワーク構造は，“(M, N)” という形で記述し，M は第 1 層から第 3 層のフィルタの数，N は第 4 層のフィルタの数を表す．フィルタのサイズと数の変更の実験について，ネットワーク構造は，“$M \times M \times N$” という形で記述し，$M \times M$ はフィルタのサイズ，N はフィルタの数を表す．このとき，図 3 に示す基とするネットワーク構造は，それぞれの表記で，“N → P P → N”，“$(64, 32)$”，“$5 \times 5 \times 64$” となる．

3.2.6 学習方法

3.2.6.1 データセットに対する操作

CNN の学習を行う際に，元々画素数 32×32 である学習画像からランダムに画素数 24×24 のパッチを切り出して学習に用いている．また，50% の確率で切り出したパッチに対して左右反転を行っている．また，学習画像全体の平均画像の画素値を各学習画像の画素値から引いて CNN に入力している．

また，テスト画像を認識する際，テスト画像から画素数 24×24 のパッチを 10 枚切り出して使用し，それらのクラス確率の平均を取っている．ここで，10 枚のパッチとは，元の画像の 4 つの角のパッチと中心のパッチ，そして，それらを左右反転したパッチである．

図 3 のネットワーク構造はこれらの操作を行うことを前提に設計されたものであり，これらの並進移動と左右反転を行ったうえで 誤認識率が 11% となるものである．

3.2.6.2 学習規則

本実験における CNN の学習では確率的勾配降下法を用いており，各層の重みは式 (5)，(6) の更新式で計算される．バイアスについても同様の更新式で計算される．ただし，バイアスの計算では正則化項を含まない．また，式中の各パラメータについては**表 7**，**表 8** に示す．

表 2 畳込み層のネットワーク構造のパラメータ．図 3 に示す CNN の畳込み層のネットワーク構造を示す．

層番号	フィルタ数	フィルタサイズ	チャンネル	ストライド	パディング	活性化関数
C11	64	5×5	3	1	2	ReLU
C12	64	5×5	64	1	2	ReLU
L1	64	3×3	64	1	1	ReLU
L2	32	3×3	64	1	1	ReLU

表 3 プーリング層のネットワーク構造のパラメータ．図 3 に示す CNN のプーリング層のネットワーク構造を示す．

層番号	プーリングのサイズ	ストライド	プーリング方法	活性化関数
P1	3×3	2	Max	ReLU

表 5 フィルタの数の変更と回転操作の実験で使用するネットワーク構造のフィルタ数．この表では表 2 で示したネットワーク構造からの変更点のみを示す．ネットワーク構造は，“(M, N)” という形で記述し，M は第 1 層から第 3 層のフィルタの数，N は第 4 層のフィルタの数を表している．

畳込み層	$(64, 32)$	$(96, 32)$	$(112, 48)$
1,2,3 層目	64	96	112
4 層目	32	32	48

表 8 バイアスの更新式のパラメータ．式 (5) における重みをバイアスに置き換えた場合のバイアスの更新式のパラメータを示す．ただし，バイアスの計算では正則化項を含まない．

層番号	ε	α
C11	0.002	0.9
C12	0.002	0.9
L1	0.002	0.9
L2	0.002	0.9
F4	0.002	0.9

3.2.6.3 学習手順

まず, CIFAR-10 を 6 つのバッチに分割する．CIFAR-10 のデータを D とし，バッチを $D_i (i = 1, ..., 6)$ としたとき，

$$\bigcup_{i=1}^{6} D_i = D, \quad D_i \cap D_j = \varnothing \quad (i \neq j) \tag{10}$$

とする．

次に，これらのバッチを用いて以下のように学習を行う．

(1) CNN の重みの初期値として，平均 0 ，分散 σ^2 の正規分布に従う乱数を与える

$$W_0 = (w_{i,j}^k), \quad w_{i,j}^k \sim \mathcal{N}(0, \sigma^2). \tag{11}$$

(2) バッチ $D_1, ..., D_4$ を使用して 1 回目に 500 epochs の学習を行う．

$$W_1 = \mathrm{train}(W_0, \varepsilon, D_1, ..., D_4). \tag{12}$$

(3) バッチ $D_1, ..., D_5$ を使用して 2 回目に 250 epochs の学習を行う．

$$W_2 = \mathrm{train}(W_1, \varepsilon, D_1, ..., D_5). \tag{13}$$

(4) 学習係数 ε を 0.1 倍して 3 回目に 10 epochs の学習を行う．

$$W_3 = \mathrm{train}(W_2, 0.1\varepsilon, D_1, ..., D_5). \tag{14}$$

(5) 学習係数 ε を更に 0.1 倍して 4 回目に 10 epochs の学習を行う．

$$W_4 = \mathrm{train}(W_3, 0.01\varepsilon, D_1, ..., D_5). \tag{15}$$

(6) W_4 の重みの CNN の認識精度をバッチ D_6 を用いてテストをする．

ここで，$W_0, ..., W_4$ はそれぞれ CNN 全体の重みを表し，関数 train は誤差逆伝播法による学習を表す．また，“epoch” は学習における重みの更新回数のことである．なお，CNN の学習はこれらをすべて行ったうえで 1 回の学習とする．

3.2.6.4 認識精度の評価

CNN の学習は，正規分布に従う乱数で重みを初期化して行うため，各学習ごとに認識精度の差が生じる．本実験では，解析を行う 1 つのネットワーク構造ごとに 5 回の学習を行い，誤認識率の平均と標準偏差を計算する．誤認識率は小数点第 2 位までを考え，評価を行う．

3.3 結果と考察

ネットワーク構造の各要素の変更と学習画像の回転処理を行った場合の実験結果を**表 9**，**表 10**，**表 11** に示す．また，それらの結果をグラフとして**図 4**，**図 5**，**図 6** に示す．ここで，実験結果の値は，誤認識率を百分率で表記したものである．

図 4 において，正規化とプーリングの順序を変更した場合の結果は，第 1 層と第 2 層で順序が異なる基のネット

表 6 フィルタのサイズと数の変更の実験で使用するネットワーク構造のフィルタサイズ (1,2 層目) とフィルタ数 (1,2,3 層目)．この表では表 2 で示したネットワーク構造からの変更点のみを示す．ネットワーク構造は，“$M \times M \times N$” という形で記述し，$M \times M$ はフィルタのサイズ (1,2 層目)，N はフィルタの数 (1,2,3 層目) を表している．

	$5 \times 5 \times 64$	$4 \times 4 \times 64$	$6 \times 6 \times 64$	$5 \times 5 \times 80$	$4 \times 4 \times 80$	$6 \times 6 \times 80$
フィルタサイズ	5	4	6	5	4	6
フィルタ数	64	64	64	80	80	80

表 9 プーリングと正規化の順序の実験結果．プーリング (P) と正規化 (N) の順序について，N → P または P → N と表記している．各値は誤認識率を百分率で表したものであり，5 回の実験の誤認識率とそれらの平均値及び標準偏差である．

ネットワーク構造 1 層目	2 層目	1 回目	2 回目	3 回目	4 回目	5 回目	平均	標準偏差
N → P	P → N	10.90	11.12	11.20	11.09	10.81	11.02	0.14
N → P	N → P	11.25	11.10	10.87	11.44	11.24	11.18	0.18
P → N	P → N	12.32	11.11	11.25	10.80	10.96	11.29	0.53

表 10 フィルタの数の変更と回転操作の実験結果．ネットワーク構造は，“(M, N)” という形で記述し，M は第 1 層から第 3 層のフィルタの数，N は第 4 層のフィルタの数を表している．各値は誤認識率を百分率で表したものであり，5 回の実験の誤認識率とそれらの平均値及び標準偏差である．

ネットワーク構造	1 回目	2 回目	3 回目	4 回目	5 回目	平均	標準偏差
$(64, 32)$	10.90	11.12	11.20	11.09	10.81	11.02	0.14
$(96, 32)$	10.70	11.10	10.70	10.72	10.59	10.76	0.17
$(112, 48)$	10.49	10.72	10.36	10.74	10.62	10.59	0.14
$(64, 32)$ + 回転	11.37	11.42	11.25	11.79	11.12	11.39	0.22
$(96, 32)$ + 回転	10.79	10.96	11.17	10.94	11.03	10.98	0.12
$(112, 48)$ + 回転	10.77	10.82	10.88	11.30	11.14	10.98	0.20

表 11 フィルタのサイズと数の変更の実験結果．ネットワーク構造は，“$M \times M \times N$” という形で記述し，$M \times M$ はフィルタのサイズ，N はフィルタの数を表している．各値は誤認識率を百分率で表したものであり，5 回の実験の誤認識率とそれらの平均値及び標準偏差である．

ネットワーク構造	1 回目	2 回目	3 回目	4 回目	5 回目	平均	標準偏差
$4 \times 4 \times 64$	11.21	11.13	11.31	11.52	10.80	11.19	0.23
$5 \times 5 \times 64$	10.90	11.12	11.20	11.09	10.81	11.02	0.14
$6 \times 6 \times 64$	12.01	11.64	11.52	11.38	11.64	11.64	0.20
$4 \times 4 \times 80$	10.69	10.80	11.18	10.78	11.15	10.92	0.20
$5 \times 5 \times 80$	10.87	10.86	10.85	11.04	10.80	10.88	0.08
$6 \times 6 \times 80$	11.46	11.60	11.45	11.06	11.47	11.41	0.18

ワーク構造が最も精度がよく，標準偏差も小さいことが分かる．また，順序を統一した場合では，正規化を先に行う方の精度が良いことがわかり，逆にプーリングを先に行うと，精度のばらつきが大きくなっていることが分かる．

図 5 において，フィルタの数を変更した場合では，フィルタの数が多い方が認識精度が良い結果となっている．回転処理を加えた場合は，ただ回転処理を加えただけでは認識精度の悪化が見られたが，フィルタ数を増加すると基のネットワーク構造と同程度の認識精度となっていることが分かる．

図 6 において，フィルタのサイズを変更した場合は，基のネットワーク構造の精度が最も良いことが分かり，これより，適切なサイズに調整されていたと考えられる．しかし，フィルタのサイズを小さくした場合でもフィルタの数を増やすと，基のネットワーク構造よりも精度がわずかではあるが向上していることから，複数の要素の関係によっても精度が左右されることが考えられる．

4. おわりに

本研究では，CNN を用いた一般物体認識手法において認識精度を変化させる要因の解析を行った．

その結果，畳込み層のフィルタの数を増加させることで，

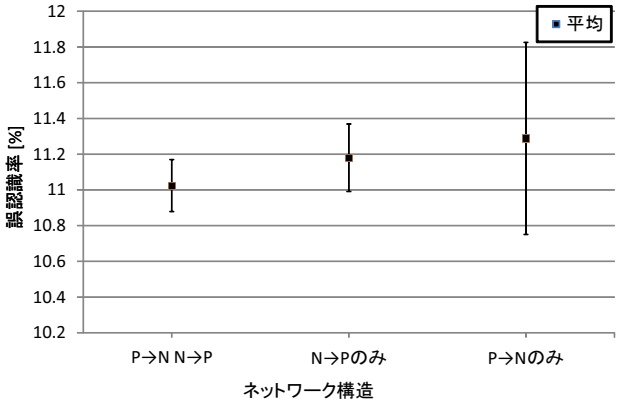

図 4 プーリングと正規化の順序の実験結果．表 9 の結果のうち，平均と標準偏差を表している．プーリング (P) と正規化 (N) の順序について，N → P または P → N と表記している．

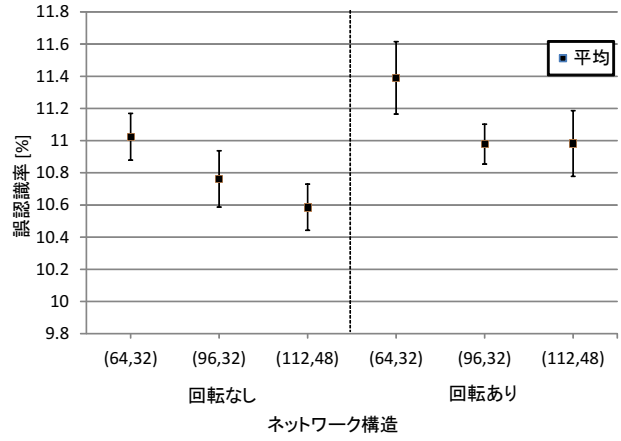

図 5 フィルタの数の変更と回転操作の実験結果．表 10 の結果のうち，平均と標準偏差を表している．ネットワーク構造は，“(M, N)” という形で記述し，M は第 1 層から第 3 層のフィルタの数，N は第 4 層のフィルタの数を表している．

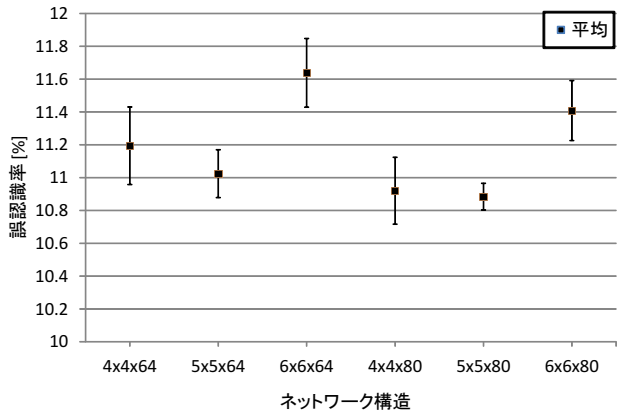

図 6 フィルタのサイズと数の変更の実験結果．表 11 の結果のうち，平均と標準偏差を表している．ネットワーク構造は，“$M \times M \times N$” という形で記述し，$M \times M$ はフィルタのサイズ，N はフィルタの数を表している．

認識精度のある程度の向上が可能であることを確認した．また，学習画像にランダムな回転を加えると認識精度は低下したが，フィルタの数を増やすことで回転を加えていない場合と同程度の認識率に達することを確認した．プーリングと正規化の順序については，正規化を先に行う方が認識精度のばらつきを抑えられることを確認した．

しかし，フィルタの数を増やすと学習に必要な時間も増加するため，今後の研究では，学習時間等も考慮する必要が考えられる．また，本論文では，解析したネットワーク構造の数が少なく，また，学習画像として CIFAR-10 以外のデータセットを用いていないため，CNN 全体に共通した性質を解析するには，より多くの解析を行う必要があると考えられる．

参考文献

[1] 柳井啓司：一般物体認識における機械学習の利用 (特別セッション, 機械学習とその応用)，電子情報通信学会技術研究報告. IBISML, 情報論的学習理論と機械学習， Vol. 110, No. 76, pp. 103–112 (2010).
[2] Csurka, G., Dance, C., Fan, L., Willamowski, J. and Bray, C.: Visual categorization with bags of keypoints, *Workshop on statistical learning in computer vision, ECCV*, pp. 59–74 (2004).
[3] Perronnin, F. and Dance, C.: Fisher kernels on visual vocabularies for image categorization, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007.*, IEEE, pp. 1–8 (2007).
[4] Krizhevsky, A., Sutskever, I. and Hinton, G.: Imagenet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems*, Vol. 25, pp. 1106–1114 (2012).
[5] 岡谷貴之：ディープラーニングと画像認識への応用，第 19 回画像センシングシンポジウム (2013).
[6] Zeiler, M. D. and Fergus, R.: Visualizing and understanding convolutional neural networks, *arXiv preprint arXiv:1311.2901* (2013).
[7] 岡谷貴之，齋藤真樹：コンピュータビジョン最先端ガイド 6，chapter 4, アドコム・メディア株式会社 (2013).
[8] Ciresan, D., Meier, U. and Schmidhuber, J.: Multi-column deep neural networks for image classification, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012*, IEEE, pp. 3642–3649 (2012).
[9] Krizhevsky, A.: Learning multiple layers of features from tiny images, Technical report (2009).