Vol.2016-MPS-111 No.10
2016/12/12

畳み込みニューラルネットワークを用いた画像分類タスクの直感的可視化方法

荒井 敏[†1]　長尾智晴[†1]

概要：深層ニューラルネットワークは画像認識の様々な分野で目覚ましい成果を上げているが，今後の産業応用を考える上では解決すべき課題も残されている．一例として，画像分類のタスクでは分類結果をラベルとして出力するだけではなく，画像中のどの部位に着目して分類がなされたか，分類の根拠を示すよう求められる場合がある．筆者らはこの問題を解決するシンプルな構成のネットワークを提案する．提案手法では分類スコアと直接対応する可視化用のマップが分類タスクの過程で生成され，視覚的に確認可能なマップが分類結果に自然な形で反映される．ベンチマーク用画像を用いて実験を行い，本手法が可視化手法として有効であることを示す．

キーワード：深層学習，畳み込みニューラルネットワーク，画像分類，可視化

Intuitive Visualization Method for Image Classification Using Convolutional Neural Networks

SATOSHI ARAI[†1]　TOMOHARU NAGAO[†1]

Abstract: Deep neural networks show excellent performance in various image recognition field. However, some issues remain for future industrial applications. For example, in image classification tasks, users might request not only to estimate class label but also to answer where the system give attention to classify. We propose novel network architecture to solve this issue. Our method generates 2D maps directly related to classification scores during classification, and generated maps are visually recognizable and reflected to classification result naturally. We empirically indicate effect of our method for existing datasets.

Keywords: deep learning, convolutional neural networks, image classification, visualization

1. はじめに

近年，深層学習[1]の発展にともなって認識処理の性能が急速に向上し，コンシューマ分野から産業分野に至るまでさまざまな活用の機運が高まっている．

画像認識の分野では，長らくヒト視覚系が究極の目標であったが，畳み込みニューラルネットワーク（CNN）と大規模データによる学習を組み合わせることで認識精度が大きく向上し[2]，一般画像の分類においてヒト視覚系の認識精度を超えるような結果も得られるようになった[3].

CNN は畳み込み層を構成要素の一つとして用いるニューラルネットワークであり，特に画像認識や画像生成の分野で広く用いられている．ネットワーク規模に応じて表現力が向上するため，高度なタスクに応用する際はより大規模でより層数の多い（深い）ネットワークが求められる傾向にある．このような大規模なネットワークを学習によって全体最適化できることが深層学習の強みである反面，全体的な動作の把握を難しくしている．

現在，深層学習はその高い性能に牽引される形で普及が進んでいるが，その動作に関しては十分理解が進んでいるとはいえず,未だにブラックボックスであるといってよい．これはコンシューマ分野ではあまり重視されないかもしれないが，産業分野，特に品質検査や医療などの安全性に関わる分野に応用する際は無視できない課題となり得る．

産業分野への応用においては単に正しい認識結果を返すだけでは不十分で，どのような観点で認識処理を行ったのか，何らかの根拠を示すように求められる場合がある．特に画像入力に対してラベルのみを出力するような画像分類のタスクでは，認識処理が想定した対象に対して正しく行われているかという懸念が常にあるため，これを確認する意味でも実際に画像中のどの部位に着目して分類が行われたかという情報は重要である．

このような背景を踏まえ，本稿では CNN を用いて単に画像を分類するだけではなく，認識結果に関する適当な根拠を提示する手法に焦点を当てる．

2. 関連研究

2.1 畳み込みニューラルネットワーク（CNN）

CNN はフィルタの畳み込み演算（convolution）を用いた多層のフィードフォワード型ニューラルネットワークである．畳み込み演算を用いた画像認識系の着想は福島[4]の Neocognitron に端を発する．LeCun ら[5]は手書き数字画像分類用の処理系として，逆伝播法を用いた end-to-end 学習が可能であり現在の CNN の原型となる LeNet を提案した．

†1　横浜国立大学大学院環境情報学府

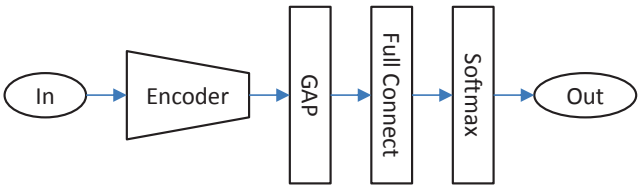

図 1　畳み込みニューラルネットワーク（CNN）の基本的な構成

LeNet 以降様々なバリエーションが提案されており現在も活発な研究が続いているが，画像認識用のネットワークに関しては基本的な骨格は概ね共通している（図 1）．まずネットワークの前半は主に特徴量を算出する役割を担う．ここでは畳み込み演算と活性化関数を用いた非線形変換が多段に適用されるが，その途中，プーリング層によって空間サイズが縮小されるほか，正規化処理などが行われる．特徴量は入力画像に比べて空間サイズが小さく，チャンネル数が大幅に増加するのが普通であり，疎表現の符号化（encode）が行われている．そしてネットワークの後半は特徴量を集約しクラスの判別を行う．前半部分に比べると後半のバリエーションは少なく，GAP 層（Global Average Pooling）や全結合層（Full Connect）などを組み合わせ，最後に Softmax 回帰を行うのが普通である．

2.2 Encoder-Decoder モデル

Encoder-Decoder モデル（図 2）は画像から特徴量を算出（encode）する処理と特徴量から画像を再構成（decode）する処理を組み合わせたネットワークで，画像生成系（generative）の処理などで広く用いられている．一般に CNN は特徴量を算出する過程で空間サイズを縮小するので，再構成処理においては空間サイズの拡大が必要になる．拡大処理には単純なアップサンプリング（upsampling）のほか，逆畳み込み演算（deconvolution）[6]がよく用いられる．

2.3 可視化手法

Zeiler ら[7]は CNN に画像を入力した際に max pooling や ReLU がスイッチのように振る舞うことに着目し，逆畳み込み演算（deconvolution）を繰り返すことでユニットの反応を入力方向に向かって逆伝播させ，入力画像と同じサイズ（解像度）の反応マップを生成している．

Zhou ら[8]は学習済みの全結合層の重みを用いて最終の畳み込み層の出力を重み付き加算し，可視化用のマップを生成する手法（Class Activation Mapping）を提案している．

Selvaraju ら[9]は Zhou らの考え方を発展させた手法（Glad-CAM）を提案している．これはネットワーク出力を最終の畳み込み層の出力で偏微分して各チャンネルの重要度を求め，この重要度を重みとして CAM と同様に可視化用のマップを生成するものである．

Lin ら[10]は画像分類ネットワークの過学習を避ける観点から，全結合層を用いない構成を提案している．これは畳み込み層の最終チャンネル数をクラス数と同一に揃え，

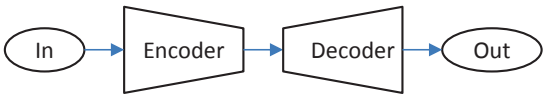

図 2　Encoder-Decoder モデル

その出力を global average pooling を用いて集約してクラス毎のスコアを得るものである．この場合，畳み込み層から出力される特徴マップ（feature maps）が各クラスの信頼度マップ（categories confidence maps）として解釈可能であることが示されている．

これら先行研究はいずれも優れた手法であるが，さらなる研究の余地も残されている．ネットワーク中のユニットの反応を個別に可視化する手法は，可視化対象が単一のユニットであるためネットワーク全体の挙動を総合的に把握することが難しく，最終的な分類結果との関連性を理解するのが困難という問題がある．中間層の出力をそのまま，あるいは重み付き加算して可視化用のマップを生成する手法は，分類処理の中間状態を可視化しているという観点でより直感的であるが，生成されたマップが分類結果とは直接対応しているとは限らず，また位置不変性を高めるために使用されるプーリング層の影響でマップの解像度が低下するという問題がある．

このような問題点を踏まえ，本稿では画像分類の根拠を直感的に可視化できる新たな画像分類手法 Generative Contribution Mappings（GCM）を提案する．

3. Generative Contribution Mappings

本節では提案手法の基本的な考え方をまず説明し，それから実際のネットワーク構成とバリエーションについて記述する．さらに理論的な解釈について解説する．

3.1 基本的な考え方

本手法の目的は，入力画像をクラス分類すると同時にその分類が画像中のどの部位に着目してなされたかという分類の根拠をユーザーに提示することである．これを実現するために以下の方針を採用する．

1. 画像中のどの部位に注目して分類が行われたか，根拠となる情報を二次元のマップとして提示する．マップは入力画像と同じ解像度で生成し，比較を容易にする．
2. 提示する情報は分類結果と直接関連したものとする．情報を見たユーザーがそこから分類結果を直感的に推定できるようなものが好ましい．
3. 方針 1,2 を実現するための構成を初めから画像分類ネットワークに組み込んでおく．これにともなうネットワーク規模の増大は許容する．

3.2 ネットワーク構成

3.1 節の方針を実現するために本手法で用いる画像分類ネットワークの基本的な構成を図 3 に示す．入力画像 I は $R \times C \times D$ の次元数，即ち垂直画素数 R，水平画素数 C，チャンネル数 D を持つとする．特に RGB 画像の場合は D=3

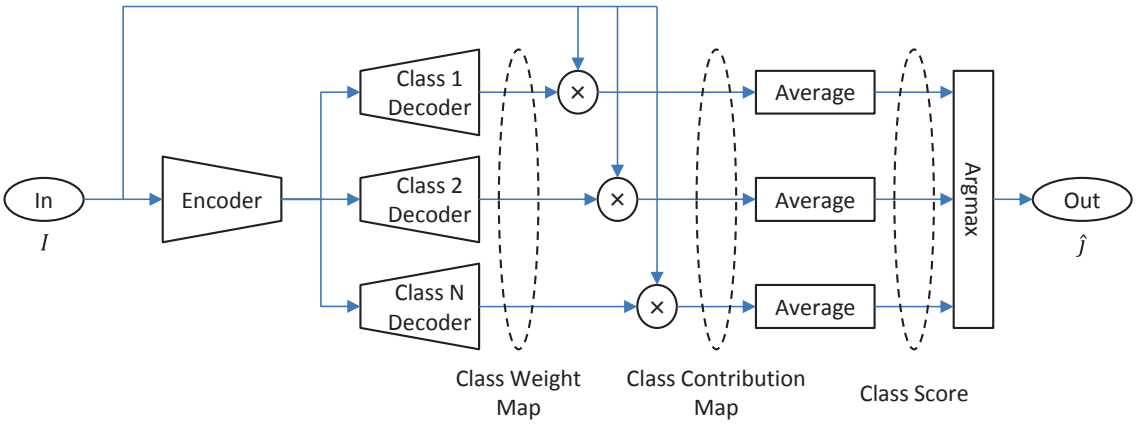

図 3　Generative Contribution Mappings のネットワーク基本構成

である．ネットワークはクラス数 N の画像分類を行い，いずれかのクラスのラベルを出力するものとする．

I はエンコーダ（Encoder）で次元数任意の特徴量に変換された後，デコーダ（Decoder）によって $R\times C\times 1$ の次元数を持つマップに再構成される（式 (1)）．このマップは入力画像の各位置が注目クラスに関してどの程度そのクラスらしいかを表す空間的な重みであり，Class Weight Map（CWM）と呼ぶ．CWM は正負の値をとり，値が負になる場合はそのクラスらしくない程度を表している．デコーダは分類する各クラスに対して 1 つ，合計 N 個が用意され，従って CWM もクラス数 N と同数が生成される．

$$M_{CWM}^{(j)} = F_{Decoder}^{(j)}(F_{Encoder}(I))$$

式（1）

但し，$M_{CWM}^{(j)}$ と $F_{Decoder}^{(j)}$ はクラス j（$j = 1, 2, \ldots, N$）の CWM とデコーダを，$F_{Encoder}$ はエンコーダそれぞれ表す．

つぎに各クラスの CWM をチャンネル方向にコピーし，入力画像 I と同じ次元数 $R\times C\times D$ に拡張する．この演算を Tile と表記する．さらに入力画像 I と要素毎に乗算を行うことで新たなマップを得る（式 (2)）．このマップは入力画像 I からの情報と CWM からのクラスらしさの情報の両方をあわせ持ち，入力画像のどの部位が注目クラスらしいかという情報を一目で把握可能になっている．これを Class Contribution Map（CCM）と呼び，ユーザーに提示するための可視化情報とする．但し，$M_{CCM}^{(j)}$ はクラス j の CCM を，演算子⊗は要素毎の積を表す．

$$M_{CCM}^{(j)} = \mathrm{Tile}\left(M_{CWM}^{(j)}\right) \otimes I$$

式（2）

さらに CCM を空間方向およびチャンネル方向に平均することでクラス j のスコア（Class Score）$V_{SC}^{(j)}$ を得る（式(3)）．

$$V_{SC}^{(j)} = \frac{1}{RCN}\sum^{R}\sum^{C}\sum_{j=1}^{N} M_{CCM}^{(j)}$$

式（3）

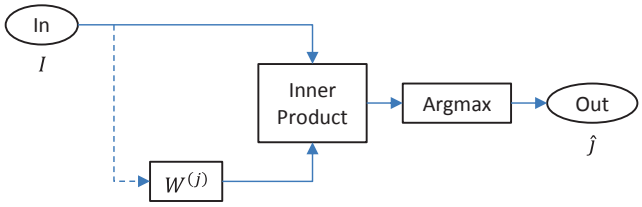

図 4　動的重み生成ネットワーク

最終的にクラススコアの最も高いクラス $\hat{j}$ を分類結果として出力する（式（4））．

$$\hat{j} = \underset{j}{\mathrm{argmax}}\, V_{SC}^{(j)}$$

式（4）

以上が提案手法の基本的な処理の流れであり，Class Contribution Map を生成的（generative）に求めることから Generative Contribution Mappings（GCM）と呼ぶ．

3.3 Shared Decoder

GCM は分類する各クラスに対して 1 つのデコーダを割り当てるため，通常の CNN と比較してネットワーク規模が増大し，これは特にクラス数が多い場合に問題となる．この問題を緩和するため，デコーダの一部をクラス間で共有する構成（Shared Decoder）を用いる．即ち，デコーダを大きく前半と後半に分割し，前半は単一のデコーダを全クラスで共通して使用し，後半はこれまで通り各クラスに 1 つのデコーダを割り当てる構成とする．

3.4 動的な重み生成ネットワークとしての解釈

GCM は画像分類ネットワークとしては複雑な構成に見えるかもしれないが，以下のように変形すると単純な形に整理できる．

まず式（2）右辺第 1 項を $W^{(j)}$ とおき，式（2）を式（3）に代入して整理すると式（5）となる．

$$V_{SC}^{(j)} = \frac{1}{RCN}\sum^{R}\sum^{C}\sum_{j=1}^{N} W^{(j)} \otimes I$$

式（5）

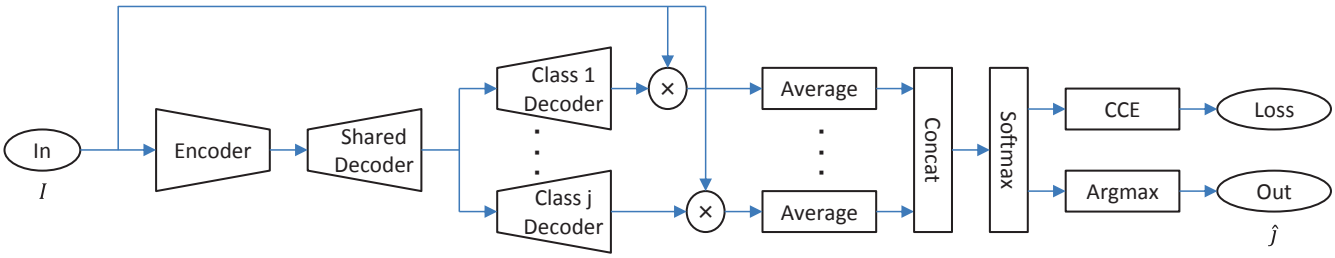

図 5　実験に用いたネットワーク構成

さらに$W^{(j)}$と I の要素毎の積と加算を内積演算としてまとめ，定数除算はクラススコアの大小関係に影響しないことからこれを無視すると，単純な形の式（6）を得る．

$$V_{sc}^{(j)} = \left(W^{(j)}, I\right)$$

式（6）

但し$W^{(j)}$は I と同じ次元数を持ち，入力画像に応じて動的に生成される．従って，GCM とは動的に生成された重み$W^{(j)}$と入力画像 I の内積によってクラススコアを求める動的重み生成型の画像分類ネットワーク（図 4）であると解釈できる．

4. 実験

提案手法の有効性を検証するため, CIFAR-10 画像データセットを用いて実験を行った．

4.1 実験設定

本稿の実験に共通して用いられる設定についてまず説明する．

(1) ネットワーク構成

基本的な構成は図 3 に示す通りであるが，ネットワーク規模抑制のためデコーダを Shared Decoder（3.3 節）に置き換える．また出力の前に Softmax 回帰を加え，損失関数として Categorical Cross-Entropy（CCE）を用いる（図 5）．エンコーダとデコーダの詳細な構成を表 1 に示す．パラメータ欄の 3x3, c16, s1 はフィルタサイズ 3x3，出力チャンネル数 16，ストライド 1 を表す． BN は Batch Normalization の適用を表す．デコーダは 5 層からなるが，層 11 から 14 は Shared Decoder であり全クラス共通で使用される．層 15 はクラス別に用意される．

(2) 学習条件

ネットワークの学習条件を表 2 に示す．これらの学習条件は CIFAR-10 に含まれる 50,000 枚の学習データを 45,000 対 5,000 に分割し，前者を学習データ，後者を検証データとする予備実験を行って事前に決定した．

(3) 実験環境

実験には NVIDIA TITAN X を搭載した Windows 10 Pro マシンを使用した．開発環境は Visual Studio 2013, CUDA 8.0, cuDNN 5.1, Python(Miniconda2) / Theano / Lasagne を用いた．

4.2 CIFAR-10

CIFAR-10 に対して 4.1 節の実験設定を用いて学習とテストを行い，テスト画像に対して 90.43%の分類精度を得た．

表 1　実験に用いたエンコーダ／デコーダの構成

	層番号	種別	パラメータ
Encoder	1	Conv	3x3, c16, s1, BN, ReLU
	2～4	Conv	3x3, c16, s1, BN, ReLU
	5	Conv	3x3, c32, s2, BN, ReLU
	6, 7	Conv	3x3, c32, s1, BN, ReLU
	8	Conv	3x3, c64, s2, BN, ReLU
	9, 10	Conv	3x3, c64, s1, BN, ReLU
Decoder	11	Deconv	3x3, c32, s2, BN, ReLU
	12	Conv	3x3, c32, s1, BN, ReLU
	13	Deconv	3x3, c16, s2, BN, ReLU
	14	Conv	3x3, c16, s1, BN, ReLU
	15	Conv	5x5, c1

表 2　学習条件

エポック数	180
バッチサイズ	128
最適化手法	SGD
学習率	0.1（エポック 100 と 150 でそれぞれ 1/10 に切り下げ）
モーメント	0.9
前処理	なし
データ拡張	画像の上下左右に 4 画素 0 padding. ±4 画素のランダムシフト切り出し＋水平方向ランダムフリップ.

学習パラメータ数は約 162,000 個であった．

図 6 に正しく分類されたテスト画像の例とその CCM を示す．CCM は正負の値をとるため，正の成分と負の成分に分けて表示している．奇数段の一番左が原画像，二番目以降が各クラスに対応する CCM の正の成分である．偶数段は同じく CCM の負の成分を示している．また，正解クラスの CCM には赤枠を付けてある．GCM では CCM の空間平均がそのままクラススコアになるので，どのクラスのスコアが高いかを目視でおおよそ読み取ることができる．

図 8 にさらに幾つかの例について原画像と正解クラスの CCM（但し正の成分のみ）を示す．

4.3 CCM の負の成分をクリップした場合

3.2 節で述べたように CCM は正負の値をとり得るが，あえて負の成分をクリップし，0 または正の成分に限定した場合の挙動を観察した．そのため表 1 に示すネットワーク構成の層 15 に ReLU を追加し，学習をやり直した．このときテスト画像に対する分類精度は 89.91%であった．図 7 に原画像および各クラスの CCM を示す．

4.4 CNN との分類精度の比較

GCM は通常の CNN とは異なりデコーダにもパラメータを割り当てなければならないため，同じパラメータ数の CNN と比較して分類精度が低下する可能性が懸念される．その程度を確認するため，CNN との比較実験を行った．

比較用の CNN は表 1 の Encoder 部に GAP，全結合層，Softmax 回帰を加え，層 2 から 10 のチャンネル数を 1.16 倍した構成とした．パラメータ数と分類精度の比較結果を表 3 に示す．学習条件は表 2 のとおりだが，CNN 側のみ前処理として中心化を行っている．

表 3　CNN との比較結果

手法	パラメータ数	処理時間比	分類精度
CNN	164K	1	90.39%
GCM	162K	1.15	90.43%

5. 考察

(1) GCM を用いた可視化の効果

図 6（A）の画像は複数の物体を含むため，通常の画像分類の場合，どちらの物体を分類対象としたかという疑問が残る．しかし CCM を見れば答えは一目瞭然で，両方の物体を対象としたスコアの合算によって分類していることが容易に読み取れる．

図 6（C）では正解である truck のほか，automobile にも比較的強い反応が見られる．車両前面は両者に共通して反応している一方，荷台部分は truck のみ強く反応し，スコアを増加させている．即ち，処理系が truck へと分類した決め手は荷台部分の有無であったことがわかる．CCM を観察することでこのような分析も可能となる．

(2) CCM における負の成分の効果

図 6 および図 7 を比較すると，正解クラスの CCM にはそれほど大きな差は見られないが，正解以外のクラスでは後者がより強い反応を示し，クラス分類の観点でノイズが多い状態といえる．CCM の負の成分は可視化情報の質を向上させる効果があるといえる．

(3) 分類精度の低下について

表 3 によれば，少なくとも今回の実験設定では GCM には懸念されたような分類精度の低下は見られず，同程度のパラメータ数の CNN と同等の分類精度が得られている．但し，デコーダを持つことで層数が増加するため，処理時間は若干増大している．

(4) 結論

以上の考察を踏まえると，提案手法（GCM）は画像分類の根拠を提示するための可視化手法として有効であり，分類精度の観点でも従来の CNN に劣らず同等であると結論付けられる．

6. まとめ

本稿では画像分類の根拠を提示するための新たな可視化手法を提案した．また，ベンチマークテスト用の画像データを用いてその有効性を確認した．

参考文献

[1] 中山英樹, ディープラーニングの発展と最新動向, 画像電子学会(2016/06/01), 2016.
[2] Krizhevsky, A. et al, ImageNet Classification with Deep Convolutional Neural Networks, Advances in Neural Information Processing Systems 25 (NIPS 2012), 2012.
[3] He, K. et al, Deep Residual Learning for Image Recognition, arXiv:1512.03385v1, 2015.
[4] Fukushima, K., Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position, Biological Cybernetics, Volume 36, Issue 4, pp.193-202, 1980.
[5] LeCun, Y. et al., Backpropagation Applied to Handwritten Zip Code Recognition, Neural Computation, Vol. 1, No. 4, pp.541-551, 1989.
[6] Noh, H. et al, Learning Deconvolution Network for Semantic Segmentation, arXiv:1505.04366v1, 2015.
[7] Zeiler, M. D. and Fergus, R., Visualizing and Understanding Convolutional Networks, arXiv:1311.2901v3, 2013.
[8] Zhou, B. et al., Learning Deep Features for Discriminative Localization, arXiv:1512.04150v1, 2015.
[9] Selvaraju, R. R. et al., Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization, arXiv:1610.02391v1, 2016.
[10] Lin, M. et al., Network In Network, arXiv:1312.4400v3, 2014.

Vol.2016-MPS-111 No.10
2016/12/12

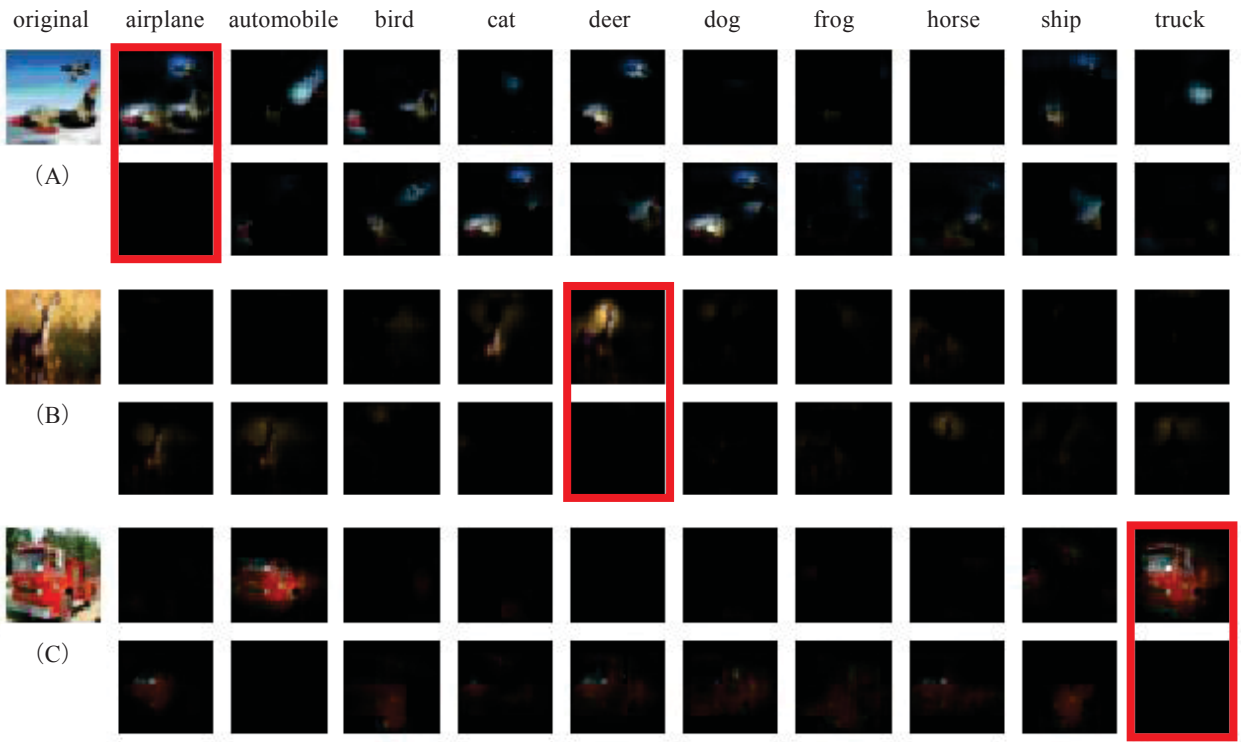

図 6　提案手法を CIFAR-10 画像に適用した場合の Class Contribution Map (CCM)
それぞれ上段が正の成分，下段が負の成分を表す．赤枠は正解クラス．

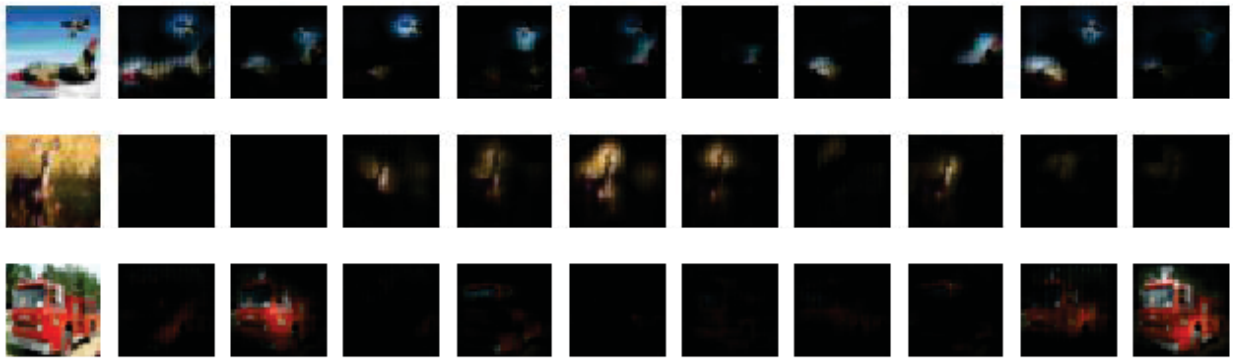

図 7　負の成分をクリップして学習した場合の CCM

図 8　原画像および正解クラスの CCM