Vol.2018-MPS-120 No.7
2018/9/25

Deep Neural Network のモデル逆解析による識別根拠可視化技術

柿下容弓[†1]　服部英春[†1]

概要：近年、Deep Neural Network(DNN)を用いたメディア処理技術が盛んに研究されている。DNN は Convolutional Neural Network(CNN)や、Full Connection 層等、複数種類の層を多層化することで複雑な関数表現を実現している。その反面、識別根拠や識別理由を人間が理解することが難しいという課題があり、誤識別原因の調査や識別精度の向上に多くの労力を要している。この課題を解決するために、DNN の可視化に関する技術が提案されているが、識別器の構成が制限される、識別根拠の解像度が低いといった課題がある。本論文では各層において出力に対する入力の寄与率を算出(モデル逆解析)することで、識別器構成に依存せず、高解像度の識別根拠を可視化する手法を提案する。

キーワード：DNN, Deep Learning, CNN, 可視化, 説明可能性

Classification Reasons Visualization of Deep Neural Network using Model Inverse Analysis

YASUKI KAKISHITA[†1]　HIDEHARU HATTORI[†1]

Abstract: Recently, media processing technologies using deep neural network are actively studied. Deep neural network realizes complicated function by multilayering several kinds of layers such as convolutional neural network, full connection layer, etc. On the other hand, deep neural network has a problem that is difficult to understand classification reasons by human. To solve this problem, visualization methods of deep neural network has been proposed. However, in these methods, the structure of classifier is limited or visualization of classification reasons is shown with low resolution. In this paper, we propose a visualization method of identification reasons with high resolution and without restriction on classifier structure by calculating the contribution ratio of inputs to output (model inverse analysis) in each layer.

Keywords: DNN, Deep Learning, CNN, Visualization, Explainable

1. 緒言

近年、Deep Neural Network(DNN)を用いたメディア処理技術が盛んに研究されている。例えばコンピュータビジョンの分野においては、特に Convolutional Neural Network(CNN)[1]を用いた手法が画像認識や物体検出、セグメンテーション等、様々なタスクで従来手法を大きく上回る成績を収めている。DNN による識別器は、CNN や Full connect 層等の複数種類の層を多層化することで複雑な識別関数を表現し、高い識別性能を実現している。しかしながら、識別器の構造も非常に複雑であり、識別結果に対する根拠を人間が理解することが難しいという課題がある。そのため、誤識別原因の特定や、識別精度向上のための対策検討に多大な労力を要する。そこで我々は、入力画像内のどの領域が、識別結果の根拠となったのかを可視化するために、寄与率解析による識別根拠可視化手法(Contribution Analysis Visualization。以下 CAV と呼ぶ)を提案する。従来も DNN の識別根拠可視化に関する研究はあったが、識別器の構造が制限される、識別根拠の解像度が低いといった課題がある。提案手法は DNN の各層において、入力が出力に対してどの程度寄与したのかを最終層から遡って解析することで、構造上の制限が少なく、高解像度での識別根拠を可視化する。

本論文では 2 章で従来手法について述べた後、3 章で提案手法を説明する。その後、4 章で学習済みの画像識別器を用いた実験結果を報告し、5 章で実験結果について考察する。

2. 従来手法

提案手法に最も関連深い従来手法として、CAM[6]や Grad-CAM[3]がある。CAM は CNN の最終層が出力した各特徴量が、どの程度最終出力に影響するかを可視化するものであり、提案手法の目的に近い。ただし、CAM は CNN の最終層に対して Global Average Pooling[4]を適用しなくてはいけないという、識別器構造上の制約がある。Grad-CAM はこれを一般化した手法であり、Global Average Pooling を適用しないネットワークに対しても、CNN 最終層の特徴量がどの程度最終出力に寄与するかを可視化できる。しかしながら、CAM や Grad-CAM が可視化する識別根拠の解像度は、CNN 最終層の特徴量マップの解像度に依存する。一般に、CNN を用いた識別器は Pooling 等のダウンサンプリ

†1 株式会社 日立製作所 研究開発グループ
Hitachi Ltd. Research & Development Group

ⓒ 2018 Information Processing Society of Japan

ング処理を複数回適用するため、CNN 最終層の特徴量は低解像度である場合が多い。識別器構成によっては CNN 最終層の解像度は 1x1 になる場合もある。このような場合、Grad-CAM では解像度の高い識別根拠を得ることは難しい。

提案手法の目的は、これらの課題を解決し、識別器の構造上の制約無く、高い解像度で特定のクラスに関する識別根拠を可視化することである。

3. 提案手法

3.1 基本的なアイデア

CAV の基本的なアイデアは、Neural Network の各層における出力に対して、各入力値がどの程度の割合で寄与したのかを求めることである。以降、これを寄与率と呼ぶ。説明を簡単にするために、図 1 に示す、入力ユニット数 4、出力ユニット数 1 の Full connection 層を考える。x_i は入力値、w_i は重み、p は識別結果である。ここで入力ユニット x_4 は常に 1.0 が入力されるユニットであり、x_4 に対する重み w_4 はバイアス係数の働きをする。また、活性化関数はこの段階では考慮しない。入力値 x_i と識別スコア p の関係を式 1,2 に示す。

$$\alpha_i = x_i \times w_i \quad \cdots式\ 1$$
$$p = \sum_i \alpha_i \quad \cdots式\ 2$$

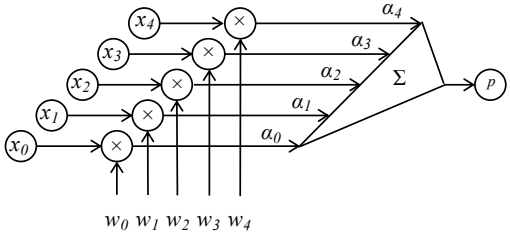

図 1：Full connection 層 1 層の例

ここで、識別スコアpの値を母数として、各入力値がどの程度の割合で寄与しているかを式 3 により算出する。これを識別スコアpに対する入力x_iの寄与率$\beta_{x_i}^{p}$とする。

$$\beta_{x_i}^{p} = \alpha_i / p \quad \cdots式\ 3$$

寄与率$\beta_{x_i}^{p}$が大きい入力値ほど識別スコアpに大きく影響した入力値、すなわち識別根拠となる特徴量を示している。また、寄与率$\beta_{x_i}^{p}$が負の値もとり得、寄与率$\beta_{x_i}^{p}$が負の値の入力値は、識別スコアpを抑制する特徴量であることを示している。識別スコアpの値が低い場合は、識別スコアpを増大する特徴量の不足、または、識別スコアpを抑制する特徴量の存在が原因として考えられるが、提案手法は寄与率の算出により、その両方の原因を確認可能である。

図 1 のように識別器が Full connection 層 1 層で構成されている場合、寄与率$\beta_{x_i}^{p}$は識別結果pに対する各入力値の寄与率であるため、寄与率$\beta_{x_i}^{p}$が識別根拠を表している。しかし多層化した場合、最終層の出力である識別スコアpに対する寄与率を、入力層側に遡って算出する必要がある。次にその方法を説明する。

図 2 に Full connection 層 2 層の例を示す。ここで x_i は Layer 0 への入力値、y_jは Layer 0 の出力値かつ Layer 1 への入力値、pは識別スコア、w_{ji}^{l}は Layer l の出力ユニット j、入力ユニット i に対する重みである。また、図には記載していないが、$\beta_{x_i}^{y_j}$は Layer 0 の出力y_jに対する入力x_iの寄与率、$\beta_{y_j}^{p}$は識別結果pに対する Layer 1 の入力y_jの寄与率を表す。

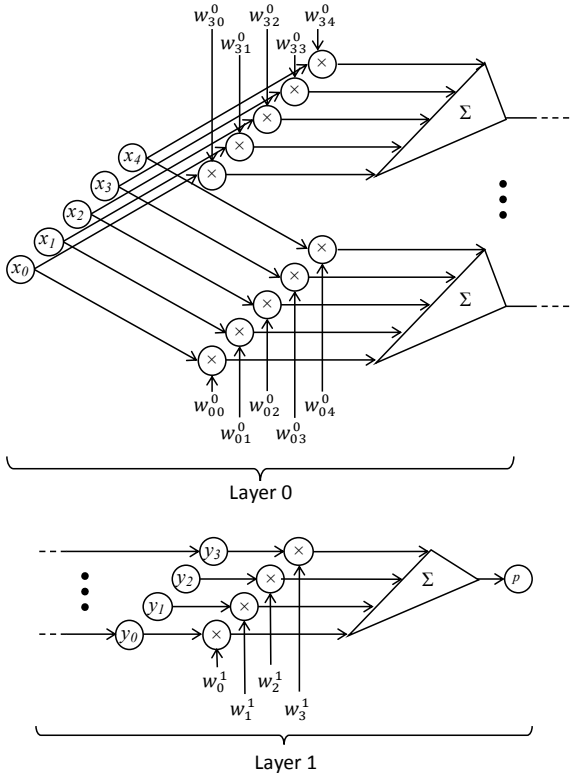

図 2: Full connection 層 2 層の例

図 2 中の Layer 1 は図 1 と同様であるため、先述の方法で寄与率$\beta_{y_j}^{p}$を算出できる。また、図 2 中の Layer 0 は図 1 の出力ユニットを複数設置した状態と考えることができるため、前述の方法で寄与率$\beta_{x_i}^{y_j}$を算出できる。

目的は識別スコアpに対する Layer 0 の入力x_iの寄与率$\beta_{x_i}^{p}$を算出することである。ここで寄与率$\beta_{x_i}^{y_j}$は出力y_jに対する入力x_iの寄与率を表しており、識別スコアpに対する出力y_jの寄与率は$\beta_{y_j}^{p}$であると分かっている。そこで我々は寄与率$\beta_{x_i}^{y_j}$に寄与率$\beta_{y_j}^{p}$を乗算して、出力y_jに関して総和をとることで、寄与率$\beta_{x_i}^{p}$を算出した。寄与率$\beta_{x_i}^{y_j}$の算出式を式 4、寄与率$\beta_{y_j}^{p}$の算出式を式 5、$\beta_{x_i}^{p}$の算出式を式 6 に示す。

$$\beta_{x_i}^{y_j} = (x_i \times w_{ji}^{0}) / y_j \quad \cdots式\ 4$$
$$\beta_{y_j}^{p} = (y_j \times w_j^{1}) / p \quad \cdots式\ 5$$

$$\beta_{x_i}^{p} = \sum_j \beta_{y_j}^{p} \times \beta_{x_i}^{y_j} \quad \cdots式\ 6$$

上記の手法により、最終層の寄与率と最終層の一つ前の層における寄与率を乗算することで、最終層の一つ前の層についても識別スコアに対する寄与率を計算することができる。もし、入力層側に更に層を増やした場合は最終層の一つ前の寄与率を最終層の寄与率と置き換えることで、同様に識別スコアに対する寄与率を計算することが可能である。すなわち、各層毎に寄与率を計算することで、識別スコアに対する任意の入力の寄与率を求めることが可能である。

次章ではコンピュータビジョンの分野で一般的に使われる Activation 層、Max Pooling 層、Convolution 層のそれぞれについて寄与率算出手法を説明する。尚、Full connection 層における寄与率算出手法は上述の通りである。先述の議論から、各層における寄与率算出手法を定義することで、これらの層をどのように組み合わせても識別根拠を可視化することが可能である。本論文では上記 4 種類の層についてのみ説明するが、提案手法はその他の層であっても寄与率算出方法を定義することで適用可能であり、ほぼ識別器構造上の制約無く適用可能な手法である。

3.2 寄与率の算出方法

本章では主に画像処理や画像認識の分野で頻繁に使用する層について、寄与率の算出方法を説明する。全体を通して、x_iを層の入力、y_jを層の出力、$\beta_{x_i}^{y_j}$を出力y_jに対する入力x_iの寄与度とする。また、層によっては入出力が 1 次元配列でなく、複数次元配列（例えばチャンネル、垂直方向位置、水平方向位置の 3 次元配列）の場合もあるが、表記を簡単にするため 1 次元配列として表記する。

3.2.1 Activation 層

Activation 層は入力ユニットに対して ReLU や tanh、sigmoid 等の非線形関数を適用する層である。Activation 層では入力と出力が一対一の関係となるため、出力y_iに対応する入力x_iの寄与率$\beta_{x_i}^{y_j}$のみ 1.0 となる。数式 7 に寄与率$\beta_{x_i}^{y_j}$の算出式を示す。

$$\begin{cases} \beta_{x_i}^{y_j} = 1.0 & (j = i) \\ \beta_{x_i}^{y_j} = 0.0 & (j \neq i) \end{cases} \quad \cdots式\ 7$$

3.2.2 Max Pooling 層

Max Pooling 層は、入力の部分領域内の最大値を出力する層である。一つの出力に寄与する入力は部分領域中の最大値のみである。よって出力y_jに対する入力x_iの寄与率$\beta_{x_i}^{y_j}$は式 7,8 で算出する。ここでR_jは出力y_jの演算対象となる入力x_iの部分領域を表す。

$$max_i = \ argmax(x_i) \ \ (i \in R_j) \quad \cdots式\ 7$$

$$\begin{cases} \beta_{x_i}^{y_j} = 1.0 & (i = max_i) \\ \beta_{x_i}^{y_j} = 0.0 & (i \neq max_i) \end{cases} \quad \cdots式\ 8$$

3.2.3 Convolution 層

Convolution 層は、入力値に対して重みを畳込み、バイアス項を加算した上で出力する層である。図 3 に 1 つの出力値に対する Convolution 処理を示す。図 3 では入力ユニットの部分領域Xを 1 次元に並べ替えてx_0からx_nとして、重みWを 1 次元に並び替えてw_0からw_nとしている。このように並び替えることで Convolution 層の処理は部分的に Full connection 層と同様の処理を行っていると分かる。よって、出力ユニットy_iに対する入力ユニットx_iの寄与率$\beta_{x_i}^{y_j}$は Full connection 層と同様の計算方法で算出できる。寄与率$\beta_{x_i}^{y_j}$の算出方法を式 9 に示す。

$$\beta_{x_i}^{y_j} = (x_i \times w_{ji}\ )/y_j \quad \cdots式\ 9$$

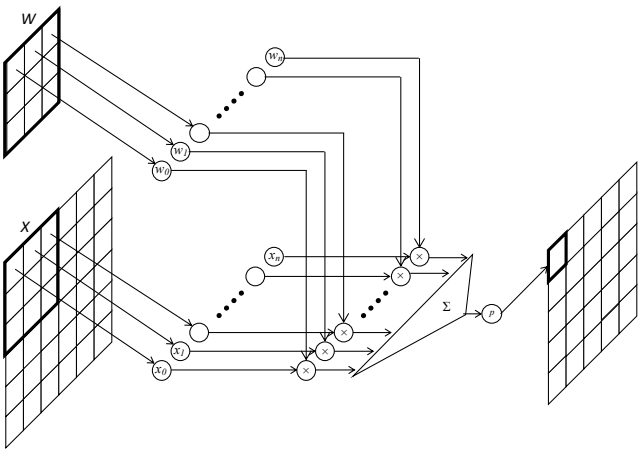

図 3：1 つの出力値に対する Convolution 処理

3.3 可視化方法

上記までで説明した手法により、任意の層における識別スコアpに対する寄与率、すなわち識別根拠を算出できる。ここで、画像識別において、入力画像上の領域毎に識別根拠を可視化する手法を説明する。多くの場合、画像識別器は Convolution 層と Pooling 層を組み合わせて特徴量を抽出した後に、Full Connect 層を 1 層以上通過して出力を得る。Activation 層は全体を通して使用する。

特徴量抽出に属する層の入出力はチャンネル、垂直方向位置、水平方向位置の次元を持つ 3 次元配列である。ただし層によって垂直、水平方向の解像度に違いがあり、Pooling や Convolution 処理の影響で層が深くなるほど解像度が低くなる傾向がある。また、Full connection 層を通過すると水平垂直方向の識別根拠の情報が消失する。よって、本手法では特徴量抽出に属する層において垂直、水平領域毎に上述した寄与率を表示することで、入力画像上に

ⓒ 2018 Information Processing Society of Japan

識別根拠を可視化する。特徴量抽出に属する層の入出力は垂直、水平位置以外にチャンネルの次元を有しているが、本論文では、チャンネル方向の寄与率を総和することで、画像上に識別根拠を可視化している。

4. 実験結果

画像識別器として広く知られているモデルである VGG-16 モデル[5]を使用して、従来手法である Grad-CAM と提案手法である CAV の可視化結果の比較を行った。以降、VGG-16 モデルを単に VGG モデルと呼ぶ。VGG モデルは 2014 年に ImageNet Large-Scale Visual Recognition Challenge にて発表されたモデルであり、入力画像を 1000 種類のオブジェクトカテゴリに分類する識別器である。図 4 に VGG モデルのネットワーク構成図を示す。

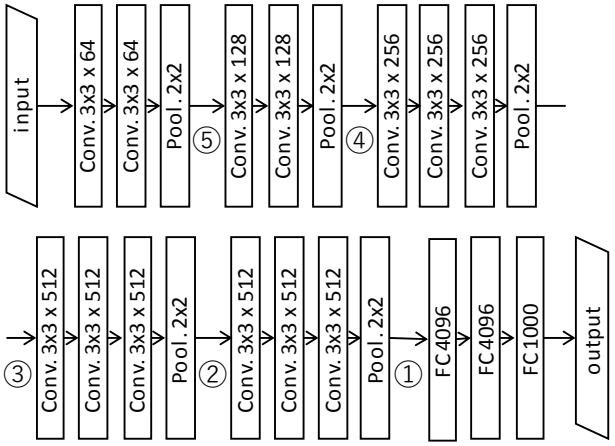

図 4：VGG モデルのネットワーク構成図。①から⑤は提案手法による識別根拠可視化箇所

図 5 に実験に用いた入力画像と、VGG モデルによる識別結果を示す。また、VGG モデルが画像内の犬と猫をどのクラスに識別しているかを確認するため、図 5 とは別に、猫と犬が写っている領域をそれぞれマスクした画像を作成し、VGG モデルに入力した。その結果、画像上側の犬を'boxer'、下側の猫を'tiger cat'と識別した。

図 6 に Grad-CAM による可視化結果[3]と提案手法である CAV による可視化結果を示す。図 6(a)と図 6(c)は Grad-CAM による'tiger cat'クラスと'boxer'クラスに対する可視化結果を示している。図 6(b)と図 6(d)は CAV による'tiger cat'クラスと'boxer'クラスに対する可視化結果であり、ここでは正の寄与率をヒートマップで描画している。図 6 の①列から⑤列は図 5 の①から⑤を表しており、可視化を行う層を示している。

Rank	Class #	Item
1	242	boxer
2	243	bull mastiff
3	246	Great Dane
4	282	tiger cat
5	292	tiger, Panthera tigris

図 5：(左)評価用画像。原画像[6]を 224x224 サイズにリサイズした。(右)識別結果の上位 5 クラス

従来手法である Grad-CAM は、図 4①の特徴量を使用した可視化のみ可能である。一方、提案手法は任意の層において識別根拠を可視化することが可能である。図 4 の①から⑤における可視化結果を図 6(b)および図 6(d)の①列から⑤列に示す。

Grad-CAM による可視化は図 4①の特徴量を使用しているため、図 6(a)および図 6 (c)と、図 6 (b)①および図 6 (d)①の可視化結果は同程度の解像度となる。可視化結果はどちらも猫は下半身、犬は顔の部分に強い反応が出ており、同じ傾向を示している。提案手法では図 6(b)および図 6(d)の①列から⑤列に示すように、更に入力層側に遡って識別根拠を可視化でき、これにより高解像度の識別根拠を可視化可能である。例えば図 6(d)④を見ると、犬の顔の中でも特に口元や額の部分を識別根拠としていることが分かる。また、図 6(b)④を見ると'tiger cat'に対する識別根拠を可視化しているにも関わらず、犬の眉間部分に強い反応が出ていると分かる。この点に関しては次章で考察する。

識別根拠を高解像度で可視化できる一方で、入力層に近すぎると識別根拠が広範囲に分散し、大局的な傾向を判断しにくくなる傾向がある。そのため、識別根拠の大局的な傾向と詳細の両方を把握するためには、複数の層で識別根拠を観察することが有効と考える。

5. 考察

提案手法で可視化した識別根拠について考察する。図 6 では正の寄与率をヒートマップで示したが、提案手法は負の寄与率も算出する。CAV で算出した正の寄与率と負の寄与率を図 7 に示す。図 7(a)は'tiger cat'、図 7(b)は 'boxer'に対する可視化結果であり、図 7 の①列から⑤列は図 4 の①から⑤を表している。図 7 の赤成分は正の寄与率、青成分は負の寄与率を示している。

ⓒ 2018 Information Processing Society of Japan

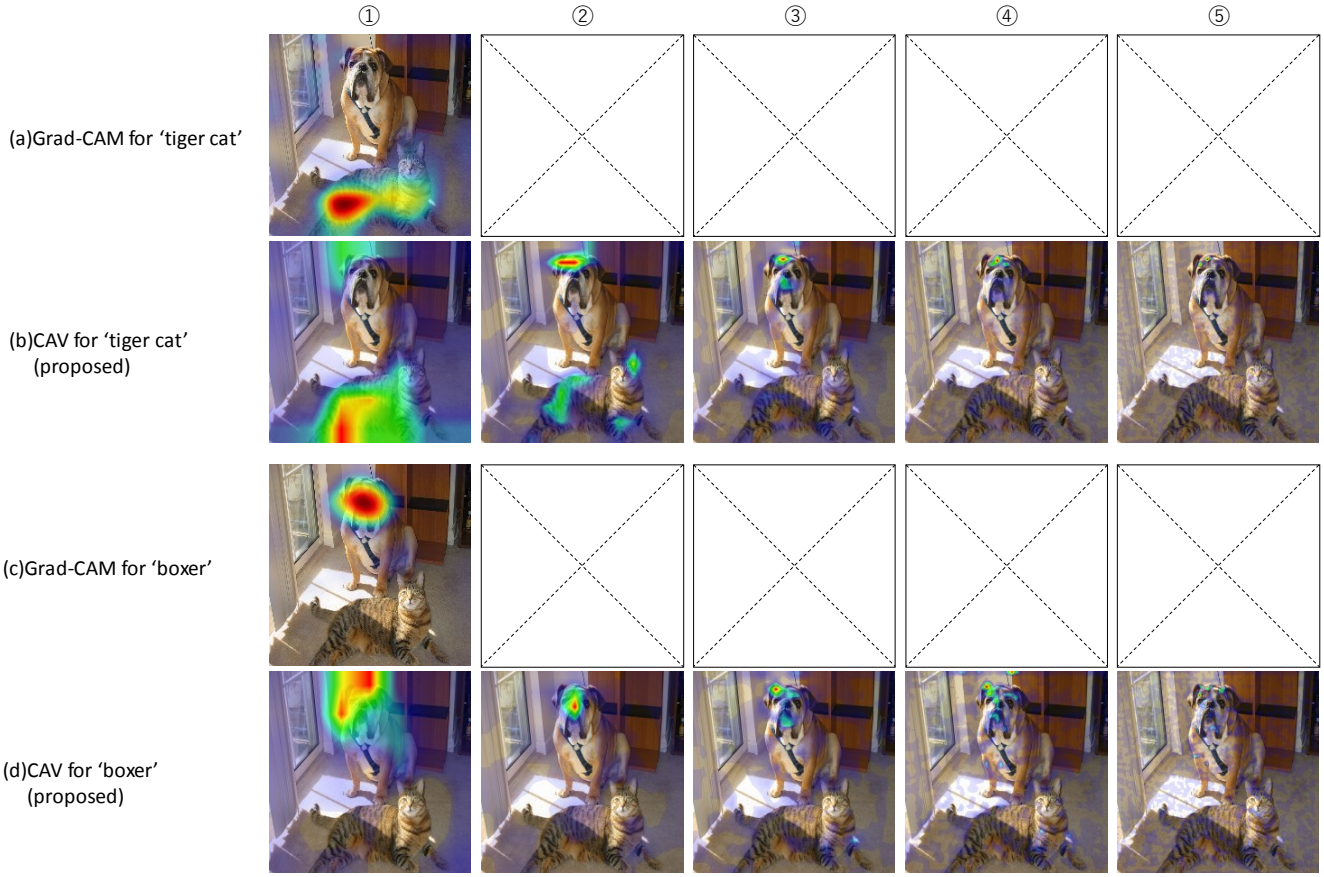

図 6：Grad-CAM と CAV による可視化結果。(a)は Grad-CAM による’tiger cat’に対する可視化結果、(b)は CAV による’tiger cat’に対する可視化結果、(c)は Grad-CAM による’boxer’に対する可視化結果、(d)は CAV による’boxer’に対する可視化結果を示している。①列から⑤列は可視化を行う層を表している

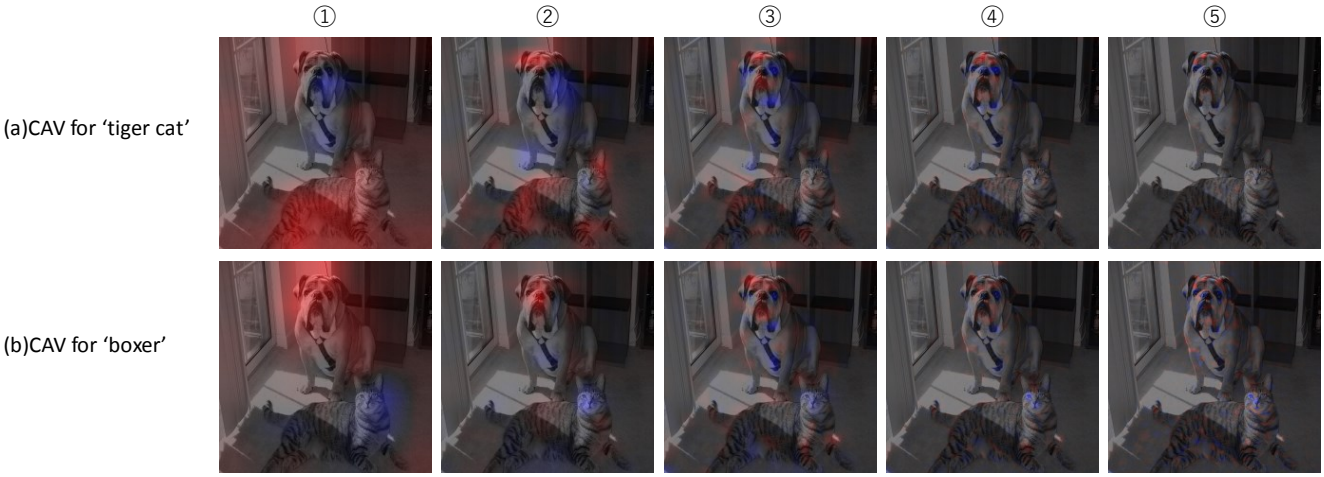

図 7： CAV による正負の寄与率マップ。赤成分は正の寄与率、青成分は負の寄与率を示す。(a)は‘tiger cat’、(b)は‘boxer’に対する可視化結果であり、グレイ画像化した入力画像と重畳している。①列から⑤列は可視化を行う層を表している

図 6(b)④にて、’tiger cat’に対する識別根拠として犬の眉間部分に反応が表れている点を考察する。図 7(a)④が、図 6(b)④の正負の寄与率を示した結果である。確かに犬の眉間部分には正の寄与率(赤)が強く表れているが、目元やあごの部分など、負の寄与率(青)も多く表れていることが分かる。そこで図 8 に示すマスク画像を用いて、識別根拠の妥当性を検証した。図 8(a)は犬が写った領域全体、図 8(b)は図 7(a)④において、犬が写った領域内の正の寄与率が高い箇所を黒で示した図である。図 8(a)(b)の黒塗り部分の特徴量を強制的に無効化し、識別スコアへの影響を観察した。尚、各クラスに対する絶対的なスコアの変化を観察す

ⓒ 2018 Information Processing Society of Japan

るために、VGG モデルの最終層の出力に対して Softmax を適用する前の値を識別スコアとして使用した。

(a) Whole region on dog　(b) High contributed region on dog

図 8：特徴量無効化箇所。(a)は犬全体から正の寄与率が高い箇所を除外した領域、(b)は犬が写った領域内の正の寄与率が高い箇所

無効化前の'tiger cat'クラスの識別スコアは 11.07 であったが、図 8(a)に示す箇所を無効化すると、識別スコアは 11.07 から 13.81 に上昇した。これは、図 8(a)に示す領域には'tiger cat'の識別スコアを抑制する特徴量が多く含まれていることを示している。

一方、図 8(b)に示す箇所を無効化した場合、'tiger cat'に対する識別スコアは 11.07 から 8.34 に低下した。表 1 に図 8(b)の特徴量を無効化した場合の識別結果の変化を示す。表 1 左は特徴量無効化前、表 1 右は図 8(b)の特徴量無効化後の識別結果上位 10 クラスであり、青いセルはイヌ科、緑のセルはネコ科のクラスを示している。特徴量無効化後は'tiger cat'のみならず、ネコ科のクラスが順位を下げ、Top 9 の全てがイヌ科のクラスになっている。このことから VGG モデルは、図 8(b)の黒塗り部分を、猫に無関係の領域にも関わらず、ネコ科の動物の視覚的特徴として使用していると分かる。

表 1：識別結果の比較。左は特徴量無効化前、
右は特徴量無効化後の識別結果の Top10。

before		
Rank	Class #	Item
1	242	boxer
2	243	bull mastiff
3	246	Great Dane
4	282	tiger cat
5	292	tiger, Panthera tigris
6	159	Rhodesian ridgeback
7	172	whippet
8	180	American Staffordshire terrier
9	254	pug, pug-dog
10	247	Saint Bernard, St Bernard

after		
Rank	Class #	Item
1	243	bull mastiff
2	242	boxer
3	246	Great Dane
4	159	Rhodesian ridgeback
5	180	American Staffordshire terrier
6	172	whippet
7	209	Chesapeake Bay retriever
8	254	pug, pug-dog
9	208	Labrador retriever
10	282	tiger cat
⋮		
20	292	tiger, Panthera tigris

以上から、VGG モデルは犬画像の中に、猫の視覚的特徴を発見しているが、その周辺に猫には存在しない視覚的特徴を多く発見しているため、総合的に猫ではないと判定していることが分かる。このように、提案手法を用いることで識別根拠となる領域や特徴を従来よりも詳細に分析することが可能となる。

6. 結言

本論文では、Neural Network を用いた識別器における識別根拠を、モデル逆解析により可視化する手法 CAV を提案し、画像識別に広く知られているモデルである VGG モデルを用いて、その効果を検証した。その結果、提案手法は従来手法よりも詳細に識別根拠となる領域や特徴を可視化できることを示した。提案手法は、どのような層で構成されたネットワークであっても、寄与率の算出方法を定義することで識別根拠の可視化を可能とする。よって、画像識別のみならず、イメージキャプショニングや自然言語処理、音声認識等、様々なタスク向けのネットワークにおいても識別や判断の根拠を提示可能と考える。また、提案手法は層の構成や特別な学習も不要なため、学習済みのネットワークに対しても構成変更や再学習を行うことなく、適用可能な汎用性を持つ。今後、提案技術の展開と応用を検討する。

参考文献

[1]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, 1989.
[2]B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. In CVPR, 2016.
[3]Selvaraju, Ramprasaath R., Das, Abhishek, Vedantam,Ramakrishna,Cogswell, Michael, Parikh, Devi, and Batra,Dhruv. Grad-cam: Why did you say that? Visual explanations from deep networks via gradient-based localization. CoRR, abs/1610.02391, 2016.
[4]M. Lin, Q. Chen, and S. Yan. Network in network. In ICLR, 2014.
[5]K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In ICLR, 2015.
[6]sabianmaggy, flickr, https://www.flickr.com/photos/40765798@N00/202734059, 2005

ⓒ 2018 Information Processing Society of Japan