様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ８ 月 ２６ 日現在

機関番号：５３３０１
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ０３０４６

研究課題名（和文）深層ニューラルネットワーク内部動作可視化システムの開発

研究課題名（英文）A Deep Neural Network Internal Operation Visualization System

研究代表者
金寺　登（Kanedera, Noboru）

石川工業高等専門学校・電子情報工学科・教授

研究者番号：５０１９４９３１

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：　DNNを正しく理解し活用できる人材を育成することを目的に，深層ニューラルネットワーク内部の処理を可視化するシステムを開発し，公開した。本研究では，ネットワーク出力を入力で偏微分した値の変動を指標とすることによって，識別の根拠となる重要な入力を特定する方法を開発した。この指標が大規模なニューラルネットワークにも有効であることを確認した。また，音声認識深層ニューラルネットワークに本研究の方法を適用し，各音韻の識別にとって重要と自動判断された知識と音響学の知識を比較し，本研究の方法の妥当性が明らかになった。

研究成果の学術的意義や社会的意義
　深層ニューラルネットワーク(DNN)を用いることにより，音声認識や画像認識などで高い認識性能を実現できるようになってきた。また災害予測など様々な分野での応用が急速に進んでいる。しかし，ネットワーク内部で何を根拠に判別しているかが不明であるため，判別結果に対する不安が懸念されている。DNN利用例が急増する中で，DNN内部で何が行われているかを正しく理解するためのシステムを開発・公開することは，技術者はもちろん利用する国民にとっても極めて重要である。

研究成果の概要（英文）：With the aim of fostering human resources who can correctly understand and utilize DNNs, we developed and released a system to visualize the internal processing of deep neural networks. In this study, we developed a method to identify important inputs that serve as the basis for discrimination by using the variation of the partial derivative of the network output with the input as an indicator. We confirmed that this index is effective for large-scale neural networks. We also applied our method to a speech recognition deep neural network and compared the knowledge of acoustics with the knowledge automatically determined to be important for the identification of each phoneme, and the validity of our method was demonstrated.

研究分野：音声情報処理

キーワード：深層ニューラルネットワーク　可視化　音声認識　画像認識

１．研究開始当初の背景

深層ニューラルネットワーク(DNN)の高い認識性能や予測性能を用いた各種システムが急速に社会に浸透している。例えば，音声認識や画像認識は人間と同等以上の性能に至っている。海外では心臓移植後の 10 年生存率を予測するシステムにより，心臓移植の順番を決定した例もある。コロンビア大学のアンドリュー･ファーガソン教授は「人工知能がブラックボックスであることが実に危険，このシステムが被害者と加害者を区別せずにピックアップしているため、見方によっては全員容疑者のようなリストを作り出している。」と述べている。

以上のように，「高い性能を示す理由が明らかでない。」「使ってみたらうまくいくがなぜかわからない。」という点が問題になっている。すなわちネットワーク内部で何を根拠に判別しているかが不明であるため，判別結果に対する不安が懸念されている。現在 DNN の内部がわかる人材は不足しており，今後 AI などで活用されるようになったときに，誤った判断が出される危険性がある。これを防ぐための人材育成は将来の AI 社会では不可欠である。

２．研究の目的

本研究の目的は，深層ニューラルネットワーク内部の処理を可視化するシステムを開発し DNN を正しく理解し活用できる人材を育成することである。

畳み込みニューラルネットワーク(CNN)を用いた画像認識等については，ネットワーク出力を feature map 各要素で偏微分したものを平均化することで注目する入力画像部分を可視化することができている(https://arxiv.org/pdf/1610.02391.pdf)。これを利用し，識別対象（犬，猫など）の識別と場所の特定，画像のキャプション付与， Visual Question Answering (VQA，質問に応じた画像位置を特定) が可能である。しかし「猫」がどこにいるかはわかるが，何を根拠として「猫」と判断しているかはわからない。また一般的なシグモイド関数や softmax 関数を用いたニューラルネットワークの場合，ネットワーク内部での処理の解析が困難でありどのような処理を行っているのか不明であることが多い。

そこで，一般的な深層ニューラルネットワーク内部の処理を可視化する方法を検討し，ネットワーク出力を入力で偏微分した値の変動によって，識別の根拠となる重要な入力を特定する方法を提案し有効であることを示した[1]。この方法を音声認識，画像認識，各種予測ニューラルネットワークなど大規模なニューラルネットワークに適用し検証することにより，深層ニューラルネットワークが何を根拠として判断しているか明らかになる点が学術的独自性と創造性を有する点である。

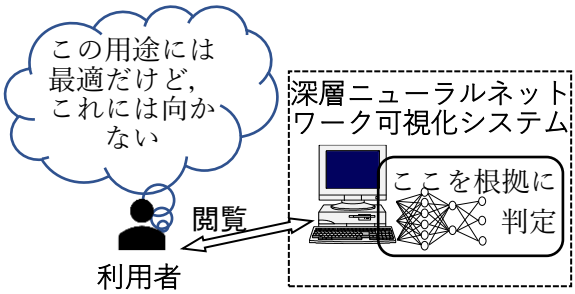

図 1 システム概要

３．研究の方法

ニューラルネットワークの各出力に対して重要な入力特徴量を調査したい場合を考える。重要な入力であれば，入力の変化が出力に影響を及ぼすと想定される。よってニューラルネットワーク出力を入力で偏微分した値が利用できる。この偏微分値は，通常の back propagation アルゴリズムと同様に簡単に導出可能である。

しかし，この偏微分値はクラスが切り替わる境界でしか大きな値を持たず，クラス内やクラス外での偏微分値は小さい。画像認識では，入力画像にノイズを加えて，入力がクラスが切り替わる境界も含まれるようにする方法（SmoothGrad）や，参照値から現在の入力までの偏微分値を積分する方法などが用いられている。画像の識別に関しては，識別対象が移動すると入力画像の位置も移動するため，入力の位置は意味を持ちにくい。

一方，音声認識の場合は，入力特徴量として MFCC（Mel-Frequency Cepstral Coefficient）や MFBANK（Mel-Filter Bank）などのスペクトル情報を数十フレーム使用することが多い。この場合，各入力位置は特定の時刻とスペクトル情報を意味する。よって，ある音素の識別にどの入力ユニットが重要かが分かれば，その音素の識別に重要な時刻とスペクトル情報を知ることができる。

以上のように，ニューラルネットワーク出力を入力で偏微分した偏微分値は，クラス内やクラ

ス外での偏微分値は小さくクラスが切り替わる境界でしか大きな値を持たないため，様々な入力に対して，ニューラルネットワーク出力を入力で偏微分した値の標準偏差が大きい入力が重要な入力であると仮定することとした。なお，標準偏差の代わりに偏微分値の大きさや二乗値も利用可能である。上記の手順で重要と判定された入力の分布を図示することで重要な入力特徴量を可視化できる。

４．研究成果

(1) 音声認識ニューラルネットワークにとって重要なユニットの抽出 [2]

図 2 に示す 9 層の音声認識深層ニューラルネットワークを用いた。入力層(第 0 層)から第 7 層までは日本語話し言葉コーパス（Corpus of Spontaneous Japanese: CSJ)[3]を用いて学習させた。学習には Kaldi[4]における CSJ レシピ[5]を利用した。入力特徴量には 12 次の MFCC と対数パワーもしくは 40CH の MFBANK を用いた。 前後 17 フレーム，計 35 フレームの特徴量をニューラルネットワークへの入力とした。

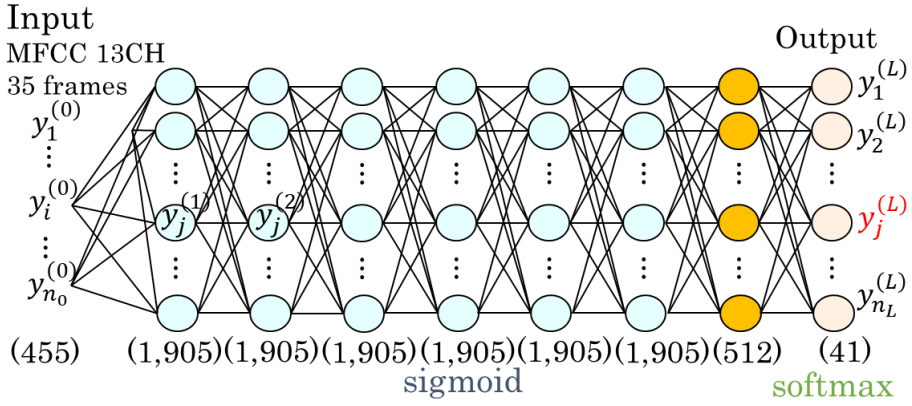

図 2 音声認識ニューラルネットワーク

図 2 の音声認識深層ニューラルネットワークに 3 節の方法を適用し，各音素の識別にとって重要と自動判断された入力を抽出した結果例を図 4 に示す。図 4 は，/s/の状態S_2において，重要な入力を四角形の大きさで図示している。四角形内部の色は，指定された音素環境に対応する入力の平均$\mu(x)$に応じた色相で描かれている。すなわち入力の平均が大きいほど赤色に近づいている。以上のように図示することで，各入力が音素の認識にとってどの程度重要であるかを知ることができる。

各音素にとって重要な入力を比較しにくい場合には，図 5 のように，2 つの音素の変動指標$\sigma\left(\frac{\partial y}{\partial x_i}\right)$の積に応じて四角形の大きさを表示すればよい。四角形の色は，2 つの音素の入力分布の異なり確率に応じた色相に描かれている。図 5 中の各入力x_iに対応する四角形をクリックすると，図 6 のように/s/及び/sh/に対応する入力x_iの確率密度関数$p_s(x_i)$，$p_{sh}(x_i)$が表示される。2 つの確率密度関数が重なる部分εは(2)式で，異なり確率p_eは(2)式で求められる。

$$\varepsilon = \int \min(p_s(x_i), p_{sh}(x_i))\mathrm{d}x_i \qquad (1)$$

$$p_e = \frac{2(1-\varepsilon)}{2-\varepsilon} \qquad (2)$$

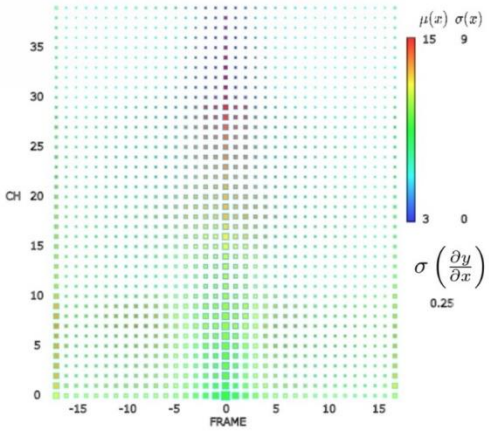

図 4 /s/の識別にとって重要な入力

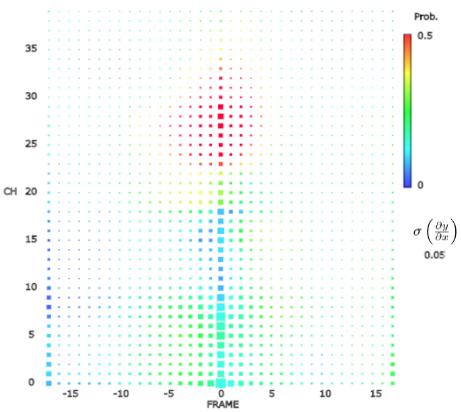

図 5 /s/と/sh/の識別にとって重要な入力

図 6 において，CH26～29 中央フレームの変動指標の積に対応する四角形が大きく，異なり確率が大きいため，CH26～29(3～4kHz)のスペクトルが重要であると推定できる。これは音響学の知見[5]と一致する。よって/s/と/sh/を識別するために深層ニューラルネットワークが注目している所がこれまで人間が注目していた所と一致していることがわかる。また，深層ニューラルネットワークが注目している所を調べることにより，これまで知られていなかった知見を知る切っ掛けになるものと期待される。/ba/と/da/を同様に比較すると，CH13(1kHz),CH17,18(1.5～1.6kHz)のスペクトルが重要であることがわかる。

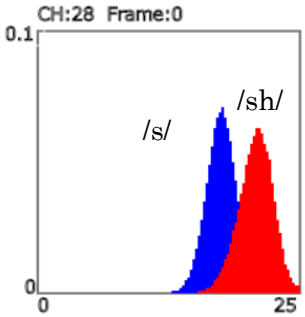

図 6 /s/と/sh/の入力ヒストグラム

(2) 計算量の削減

第 l 層 i 番目のユニットがニューラルネットワークの出力にとって重要かどうかは $Imp(l,i) = \sum_{j=1}^{n_L} \sigma\left(\frac{\partial y_j^{(L)}}{\partial y_i^{(l)}}\right)$ で知ることができる。この値が大きいユニットは残し，逆に小さいユニットを削除しても識別性能が低下しないことが期待される。

図 8 は，$Imp(0,i)$が大きい入力を選択した場合と，$Imp(0,i)$が小さい入力を選択した場合の音素認識率を示している。この結果より，$Imp(0,i)$がどの入力が重要であるかを示唆していることが確認された。

図 9 は，すべての中間層$(0 < l < L)$について$Imp(l,i)$が大きいユニットを選択した場合と，$Imp(l,i)$が小さい大きいユニットを選択した場合の音素認識率を示している。この結果より，$Imp(l,i)$がどの中間ユニットが重要であるかを示唆していることが確認された。

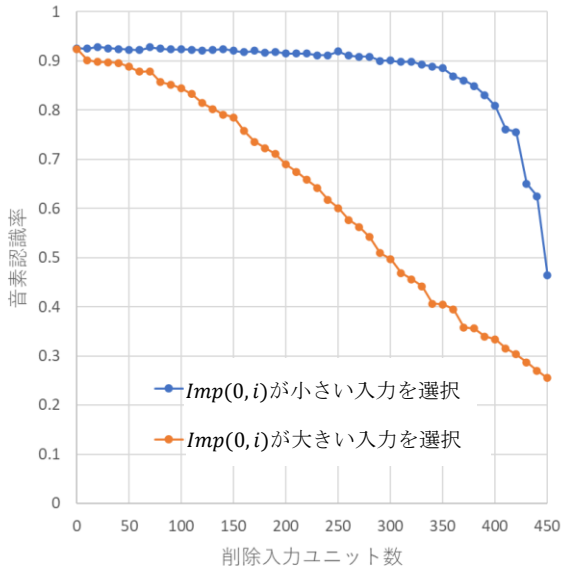

図 8 削除入力ユニット数と音素認識率

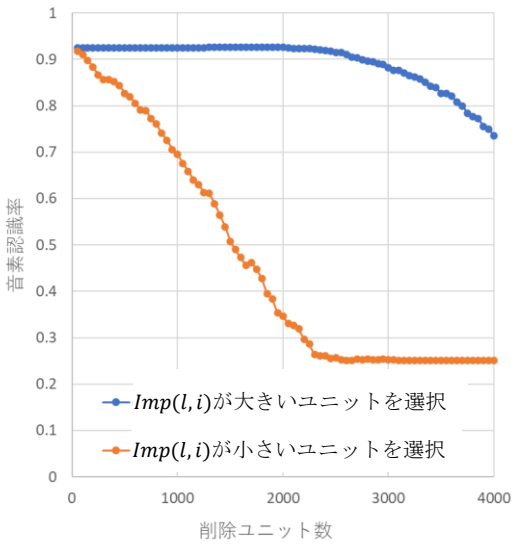

図 9 削除入力ユニット数と音素認識率

(3) 畳み込みニューラルネットワーク

図 10 に示す手書き数字認識ニューラルネットワークを MNIST データベース [7]を用いて学習した。図 13 は，第 2 層のフィルターの重要度が小さいフィルターから順に削除した場合と，重要度が大きいフィルターから順に削除した場合の手書き数字認識率を示している。フィルターを削除後，全結合層の重みのみ再学習した。重要度の小さいフィルターを削除しても識別性能が劣化しないことがわかる。この結果より，重要度によってどのフィルターが重要であるかを示唆していることが確認された。

深層ニューラルネットワーク内部の処理を可視化する方法として，ネットワーク中のユニット出力を各ユニットで偏微分した値の変動によって，識別の根拠となる重要なユニットや畳み込み層のフィルターを特定する重要度を提案し，その有効性を確認した。また，重要度が小さいユニットやフィルターから順に削除することで，認識性能を維持しながらネットワーク規模を縮小できることを示した。

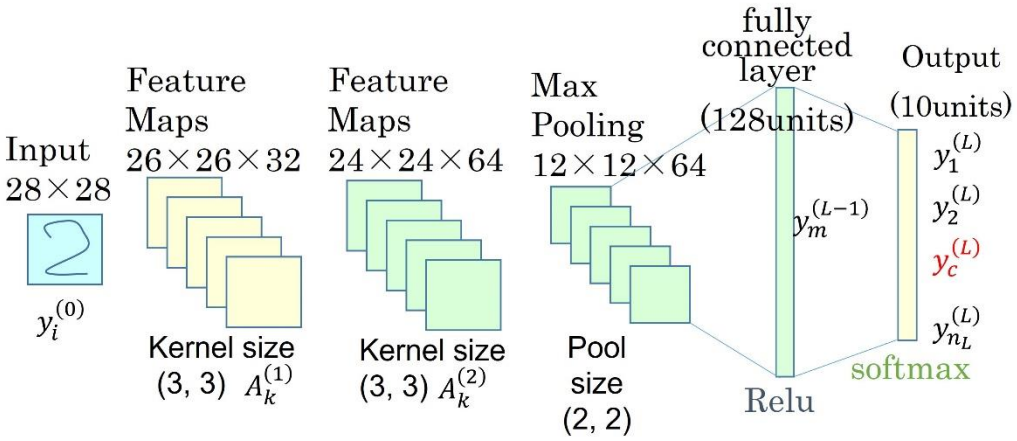

図 10 手書き文字認識ニューラルネットワーク

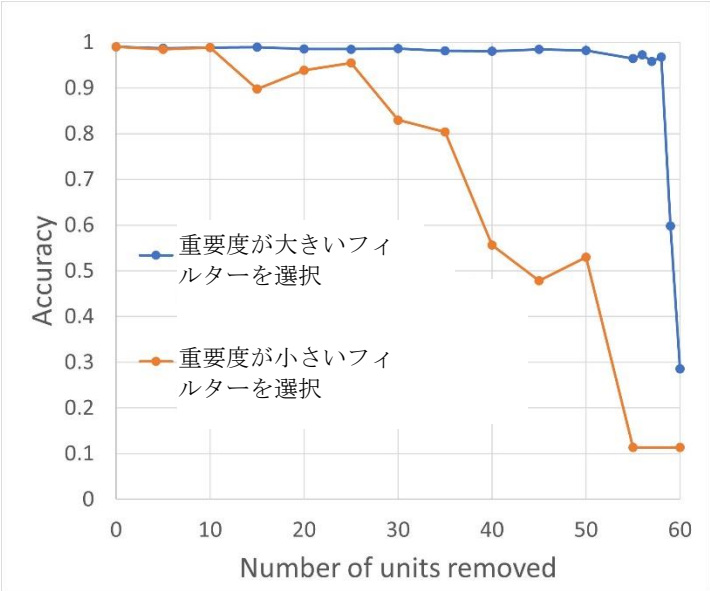

図 13 削除フィルター数と手書き数字認識率

＜参考文献＞

[1] 金寺 登，音声認識音響深層ニューラルネットワークにとって重要な入力の自動抽出，日本音響学会誌，76（6），327-330 (2020).

[2] 音声認識深層ニューラルネットワークにとって重要な入力自動抽出システム，http://sail.i.ishikawa-nct.ac.jp/dnn/ (2019).

[3] S. Furui, K. Maekawa and H. Isahara, “A Japanese National Project on Spontaneous Speech Corpus and Processing Technology,” Proc. ISCA Workshop on Acoustic Speech Recognition, pp.244-248 (2000).

[4] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlcek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, “The Kaldi Speech Recognition Toolkit,” Proc. ASRU (2011).

[5] 篠崎隆宏，森谷崇史，田中智大，渡部晋治, “Kaldi における CSJ レジピの利用法,” 情報処理学会研究報告 2016-SLP-110, 8, pp.1-6 (2016).

[6] 吉田友敬，言語聴覚士の音響学入門（海文堂 ,2005) , p157.

[7] The MNIST Ddatabase of Handwritten Digits, http://yann.lecun.com/exdb/mnist/.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　0件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 金寺 登	４．巻 76
２．論文標題 音声認識音響深層ニューラルネットワークにとって重要な入力の自動抽出	５．発行年 2020年
３．雑誌名 日本音響学会誌	６．最初と最後の頁 327〜330
掲載論文のDOI（デジタルオブジェクト識別子） 10.20697/jasj.76.6_327	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計3件（うち招待講演　1件／うち国際学会　0件）

１．発表者名 金寺 登
２．発表標題 音声認識深層ニューラルネットワークにとって重要なユニットの抽出と計算量削減
３．学会等名 日本音響学会2021年春季研究発表会
４．発表年 2021年

１．発表者名 金寺 登，阿知良 澪，藤井 烈，片桐 寿通，山川 将径，田屋 祐樹，辰橋 浩二，前 正人
２．発表標題 猛禽類の鳴き声自動判別
３．学会等名 日本音響学会2020年春季研究発表会
４．発表年 2020年

１．発表者名 金寺 登
２．発表標題 音声認識深層ニューラルネットワークは何を見ているのか
３．学会等名 2019年度電気・情報関係学会 北陸支部連合大会（招待講演）
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

石川工業高等専門学校電子情報工学科金寺研究室ホームページ http://sail.i.ishikawa-nct.ac.jp/ 音声認識深層ニューラルネットワークにとって重要な入力自動抽出システム http://sail.i.ishikawa-nct.ac.jp/dnn/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関