

# ディープニューラルネットワークにおける 複数特徴量間の関係に着目した解釈手法の提案

京都大学工学部情報学科 広域情報ネットワーク分野

山田 瑛平

指導教員：松原 繁夫 准教授

# 背景

- ▶ 近年, Deep neural networks(DNNs)が様々なタスクにおいてよい性能を示すことが示されている.
- ▶ 画像認識や音声認識, 自然言語処理など, DNNsは様々な分野への適用が進められている. 一方で, DNNsのブラックボックス的側面は実際的な運用にDNNsを適用することの障壁となっている.

# 背景

- ▶ Explainable AIという人工知能に関する研究がある。
- ▶ これは、現在の人工知能の、高い精度で予測や認識ができるが、どのような根拠をもってそれらのタスクを解いているのかが説明できない、という性質に関係する。
- ▶ 人工知能はブラックボックスであり、explainableでない。実際に人工知能を導入する場において、説明が出来ないことはAIへの不信感をもたらす。
- ▶ 例：医療現場において患者の状態から病名を判定するAI。
- ▶ AIは何らかの根拠をもって患者の罹患した病気を判定しているが、具体的にその根拠とは何なのかについて説明を与えられなければ、患者はAIに不信感を募らせるかもしれない。
- ▶ AIに与えられる学習用データに偏りがある場合、AIが誤った判断を下しかねない。このような重大な医療ミスを避けるべく、何を根拠に判定を行うか説明できるようなAIが必要とされている。

# 先行研究

- ▶ XAIに関する研究では、結果が直感的に分かりやすい画像認識やNLPを題材にすることが多い。例えば、ある先行研究では、MNISTの手書き文字分類をするDNNに対して、以下のように各ピクセルに分類への貢献度を割り当てることで、どのピクセルが分類へ貢献したかを一目で理解できるようにしている。

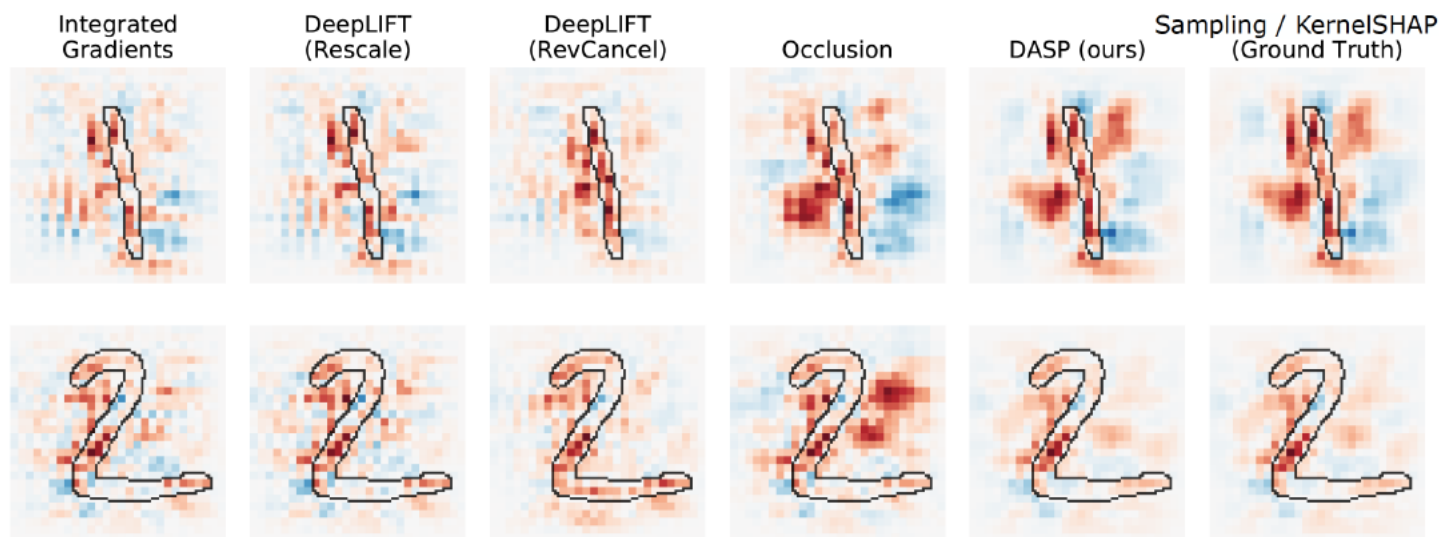


図1：解釈の例

Marco Ancona, Cengiz Oztireli, and Markus Gross: Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Values Approximation.(2019)より引用

# 先行研究

- ▶ Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Values Approximation
- ▶ DNNsを解釈可能にする様々なアプローチと, 研究内のメインテーマである Shapley valueという概念をDNNsに適用した手法との比較を行っている.
- ▶ 逆伝播ベースの手法, 摂動ベースの手法, およびShapley valueを用いた手法を, それぞれMNISTの手書き文字分類に適用し, それらの性能を比較する. 逆伝播ベースの手法はノイズの多い説明をする傾向にあること, 摂動ベースの手法は入力の一部の領域を過大評価または過小評価する傾向にあることを明らかにしている.

# 逆伝播ベースの手法 ;

## LRP(Layer-wise Relevance Propagation)

- ▶ DNNsは, 入力ベクトル  $x$  を最終的なスコア  $f(x)$  に変換する写像  $f$  とみなすことができる.
- ▶ 一般に, ある層  $i$  についての入力ベクトルを  $x = (x_1, x_2, \dots, x_n)$  とおくと, 次の層へは重み  $w$  を用いて  $w^T x = \sum w_t x_t$  が渡される.
- ▶ 出力層のスコア  $R_{i+1}$  を,  $R_{i,t} = \frac{w_t x_t}{\sum w_j x_j} * R_{i+1}$  という式でひとつ前の層に逆伝搬させていく.  $t$  についての総和をとることで, スコア  $R_i$  を計算できる.
- ▶ 最終的には, 入力の各ベクトルが最終的なスコアにどれだけ影響を与えたかが計算できる.
- ▶ LRPは, 最終的なスコアを重みと入力の積の比で分解する手法である.

# Shapley value

- ▶ 協力ゲーム理論における、各エージェントが最終的なスコアにどれだけ影響を与えたのかについての値.
- ▶ エージェントA, B, Cが存在し、あるタスクを協力して実行する状況を考える. ここでは例えば、タスクは物件の家賃の予測とし、A, B, Cは物件に関する特徴量である. 例として、Aは物件の広さ、Bは駅からの距離、Cはペット可かどうかを表しているものとする.
- ▶ 単純に予測をするならば、スコアを各特徴量の線形和で近似する手法が考えられる. しかし線形回帰は、「駅からの距離が近いだけの物件や、広いだけの物件の家賃はそこまで高くないが、駅から近くしかも広い物件の家賃は非常に高くなる」といった、複雑な状況を近似できない.
- ▶ **Shapley valueは、このような状況を表現するための非線形な手法である.**

# Shapley value

- ▶ Shapley valueは,  $n$ 個の各特徴量を用いるか用いないかそれぞれの $2^n$ 通りについてスコア $f$ (特性関数と呼ばれる)を計算し, そのスコアの値を用いて計算される.
- ▶  $n!$ 通りの特徴量の順列について, その順番に特徴量を追加していく状況を考える.
- ▶ 例えばAのShapley valueは, 右の表に従って, 6つの値の平均で計算される.  $f(S)$ は,  $S$ 中の特徴量を用いた場合のスコアである.

$A \rightarrow B \rightarrow C$	$f(A) - 0$
$A \rightarrow C \rightarrow B$	$f(A) - 0$
$B \rightarrow A \rightarrow C$	$f(A, B) - f(B)$
$B \rightarrow C \rightarrow A$	$f(A, B, C) - f(B, C)$
$C \rightarrow A \rightarrow B$	$f(A, C) - f(C)$
$C \rightarrow B \rightarrow A$	$f(A, B, C) - f(B, C)$



# Shapley value

- ▶ 家賃のような例では, ある特徴量のみを用いたときのスコア(家賃)を直接計算することができない.
- ▶ そこで, その特徴量のみが等しい, または近い値であるデータをサンプルとして取り出し, そのサンプル全体でのスコアの平均をとる.
- ▶ インスタンス  $(A, B, C) = (10m^2, 10\text{分}, \text{可})$  において,  $f(A, B)$  を計算したいとする. このとき, スコアは  $(11m^2, 9\text{分}, \text{不可})$   $(10m^2, 11\text{分}, \text{可})$   $(9m^2, 10\text{分}, \text{不可})$ ... といったデータの家賃の平均をとり, その値をスコアとすればよい.
- ▶ DNNsに適用する場合, 各入力をエージェントとみなしてShapley valueを計算すればよい.

# Shapley valueの長所

- ▶ LRPなどのパラメータを用いる手法は, その手法によりある程度にでも正確に貢献度を計算できていることの裏付けはない. 実際に, 恣意的な操作により虚偽の説明を与えることができることが示されている.
- ▶ Shapley valueは5つの好ましい公理に基づいて値を割り振るために, LRPなどよりも優れたアプローチであるとされる.

# 先行研究の問題点

- ▶ 単一特徴量への貢献度の情報だけでは、複数の特徴量が組み合わせることによる、相乗的な貢献度を計算することはできない。
- ▶ 例：分類が**線形分離不可能**であるような場合

A	B
C	D

→ {0, 1}

✓ 条件

A, Bが大きな値→ラベル0  
C, Dが大きな値→ラベル0  
A, Cが大きな値→ラベル1  
B, Dが大きな値→ラベル1

図2：線形分離不可能な条件

- ▶ このような条件のもとでは、各ラベルの貢献度は**すべて0**と計算されてしまう。

# 研究目的

- ▶ 研究目的：複数の特徴量が組み合わさることによる分類や認識への貢献度を計算できるようにすることで、DNNsへのより良い解釈を可能にする。

A	B
C	D

→ {0, 1}

## ✓ 条件

A, Bが大きな値→ラベル0  
C, Dが大きな値→ラベル0  
A, Cが大きな値→ラベル1  
B, Dが大きな値→ラベル1

(再掲)図2：線形分離不可能な条件

- ▶ 複数の特徴量の相乗的な貢献度を計算することにより、AとBが組み合わさることでラベル0への分類が行われている、といった解釈をすることができる。

# 提案手法

- ▶ **提案手法**：2つの特徴量を1つのまとまりとして捉え、 $n-1$ 個の特徴量におけるその特徴量のShapley値から、 $n$ 個の特徴量における2つの特徴量のShapley値2つを引いたもの、を相乗的な貢献度とする。
- ▶ 例：4つのピクセルからなる画像があり、それぞれのピクセルを  $A, B, C, D$  とする。このとき、 $A$  と  $B$  の相乗的な貢献度は、
  - ▶  $A, B$  を1つのピクセル  $E$  とみなした画像  $(C, D, E)$  における、 $E$  の貢献度  $f(E)$
  - ▶  $A, B$  それぞれの貢献度  $f(A), f(B)$の3つの値から、 $f(E) - f(A) - f(B)$  として計算される。
- ▶ 線形分離不可能な場合における例では、ラベル0に対する貢献度は  $f(E) = 1, f(A) = 0, f(B) = 0$  となるために、提案指標の値は1となる。

# Shapley valueを用いる理由

- ▶ 先述の通り, Shapley valueは公理に基づくアプローチであるために, 公理を持たないLRPなどの手法と比較して信頼できる手法であるとされる.
- ▶ LRPを始めとする, DNNの中身に着目した手法は, いずれも線形的なアプローチである. (DNN自体が線形的な手法であるため)
- ▶ 線形的なアプローチで複数の特徴量における貢献度を計算しようとしても, それは2入力の貢献度の和に他ならない.
- ▶ Shapley valueは入力と出力のみに着目した手法であり, 非線形的なアプローチである. そのため, 2入力の貢献度の和として自明でない結果が得られることが予想される.

# 実験

- ▶ 提案指標の有効性の検証のため, MNISTの手書き文字分類タスクにこの指標を適用する.
- ▶ Shapley値の計算には $O(n! + 2^n)$ が必要となる. 計算量を小さくするために,  $7 \times 7$ ピクセルを1ブロックとする16ブロックに分割し, それぞれのブロックでの画素の値を平均したものをデータセットに用いた.
- ▶ 簡単のため, ラベル0および1のものをを用いた2値分類とする.

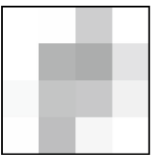
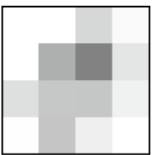
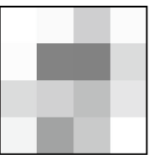

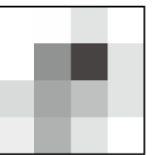
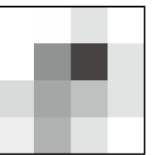
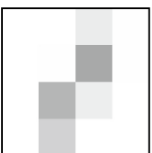
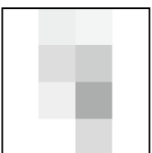
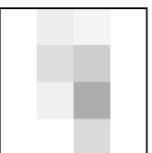
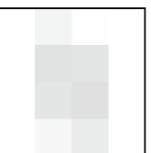
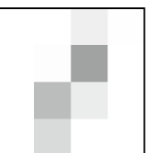
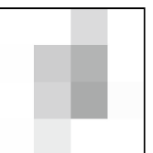
0						
1						

図3：入力データの例

# 実験

- ▶ 以下のような, 4層構造の単純なニューラルネットワークを用いた.
  - ▶ 入力層 : 16変数.
  - ▶ 中間層 1 : 結合層8個. 活性化関数はシグモイド関数.
  - ▶ 中間層 2 : 結合層8個. 活性化関数はシグモイド関数.
  - ▶ 出力層 : ソフトマックス関数により尤度を出力する.
- ▶ 20エポックを学習させたDNNを用いて, 提案指標の有効性を検証する.



# 事前予測

- ▶ 提案指標の有効性の検証のため、事前にどのような結果が出ると指標の有効性を確かめられたことになるかの予測を行う。
- ▶ ラベル0と1とで、入力画像における画素値に差があるような画素(図4における**画素A**)については、別の画素との相乗的な貢献度は低くなるはずである。  
∵DNNは画素Aを見るだけで分類すべきクラスが分かるので、他の画素と組み合わせたところで分類への貢献は変わらない。
- ▶ 一方で、ラベル0と1とで画素値の差が小さいような画素(図4における**画素BやC**など)同士が組み合わせると、相乗的貢献度は、それ以外の組と比較して大きくなるはずである。∵画素Bと画素Cを同時に見ることで、分類すべきクラスへの確信度合いが強まるので。

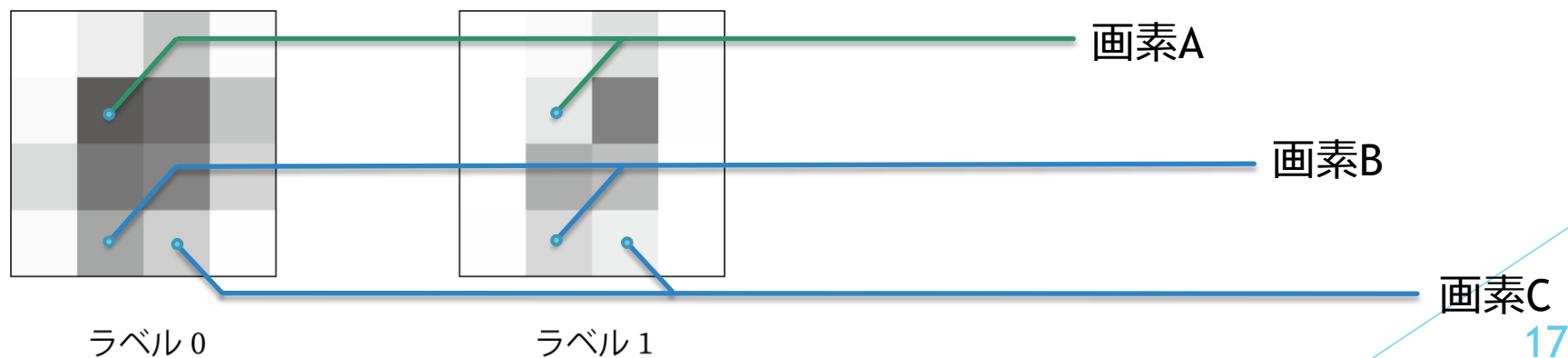


図4：入力画像

# 実験結果

- ▶ 実験結果は、いずれもこれらの予測に反するものとなった。これは、実験に用いたDNNが、事前の予測に沿うものでなかったことが原因だと考えられる。
- ▶ 一方で、下の表1に示す通り、いくつかのピクセルの組について提案指標が特徴的な値を示していることが明らかとなった。

組	提案指標	組	提案指標
X9, X14	<b>0.023985</b>	X6, X9	0.010656
X3, X9	<b>0.019004</b>	X5, X6	0.009897
X6, X10	0.012073	X9, X10	0.009897
X6, X7	0.012027	X5, X8	0.009866
X5, X2	0.01116	X2, X8	0.009075

表1：提案指標の大きさの上位10組。  
上位2組が大きく外れた値をとっている。

X1	X2	<b>X3</b>	X4
X5	X6	X7	X8
<b>X9</b>	X10	X11	X12
X13	<b>X14</b>	X15	X16

図5：実験に用いた画像

# 結論

- ▶ MNIST のデータからラベル0 および1 のものを取り出し, 画像を $4 \times 4$  に圧縮したものを学習させたDNN に対して, 複数ピクセルが組み合わさることによる貢献度を計算する提案指標を適用することで, 指標の有効性を分析した.
- ▶ 提案した指標の有効性を示すことはできなかった一方で, 提案した指標について, いくつかのピクセルの組が特徴的な値を示していた. そのような値をもたらした原因についての分析は, 今後の課題とする.

# 卒論を提出してから考えたこと

- ▶ 提案指標の有効性が確かめられなかったのは、用いたDNNがクラス0と1へ分類するDNNとしては直感に反するものであったから、と結論付けた。
- ▶ こちらが狙ったような挙動をするDNNを選んで用いるのは循環論法になってしまう(もっともらしい値の指標になるようなDNNを恣意的に選んで指標を計算すれば、もっともらしい値が結果として得られるに決まっている)。
- ▶ DNNを選ぶのではなく、エポック数を増やして重みを収束させれば恣意的な選択にはあたらないことに気付いた。
- ▶ 現に、20エポックは分類器の学習として不完全であった。
- ▶ 50エポックまで回すと精度が1%上昇した。

# 参考文献

- ▶ [1] Gregoire Montavon, Wojciech Samek, and Klaus-Robert Muller: Methods for Interpreting and Understanding DeepNeuralNetworks, *Digital Signal Processing*, Vol. 14, pp. 1-15 (2018).
- ▶ [2] Sebastian Bach, Alexander Binder, Gregoire Montavon, Frederick Klauschen, Klaus-Robert Muller, and Wojciech Samek: On pixel-wise explanations for nonlinear classifier decisions by layer-wise relevance propagation, *PLOS ONE* 10 (7) (2015), e0130140.
- ▶ [3] Alexander Binder, Sebastian Bach, Gregoire Montavon, Klaus-Robert Muller, and Wojciech Samek: Layer-wise relevance propagation for deep neural network architectures, *In Information Science and Applications* , vol. 20, pp. 913-922 (2016).
- ▶ [4] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schutt, Sven Dahne, Dumitru Erhan, and Been Kim: The (Un)reliability of saliency methods, *NIPS 2017 - -Workshop on Interpreting, Explaining and Visualizing Deep Learning* (2018).
- ▶ [5] Marco Ancona, Cengiz Oztireli, and Markus Gross: Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Values Approximation, *PMLR* 97, Vol. 10, pp. 272-281 (2019).
- ▶ [6] Shapley L.S: A value for n-person games. *Contributions to the Theory of Games*, Vol. 17, pp. 307 – -317 (1953).