The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

ニューラルネットに基づく時間変化するアトリビューション抽出手法Time-Smoothgradの提案

Time-Smoothgrad: A Method Extracting Time-Varying Attribution for Neural Networks

切通恵介 *1　　泉谷知範 *1
Keisuke Kiritoshi　Tomonori Izumitani

*1NTT コミュニケーションズ株式会社
NTT Communications Corporation

Deep neural networks become applied to industry time-series data for anomaly detection, product quality prediction, and so on. For these applications, temporal interpretations of the neural network models are very important to determine next actions. Attribution map methods, including the saliency map, can be used to interpret relationships between the inputs and outputs of a neural network in terms of partial differential values for image classification tasks. In this paper, to aim application for industry time-series data, we propose Time-Smoothgrad, an extended attribution map method to prevent unstable and noisy influence values caused by complex models and time variation. Experiments on artificial and real data show the efficiency of the method.

1. はじめに

近年，特に，産業の分野においては Internet of Things をキーワードとして各企業で膨大な時系列センサーデータの収集が行われており，これらのデータを利用した異常検知や品質予測へのディープニューラルネット技術の応用が期待されている．

例えば，工場やプラントにおいて回帰モデルの出力値を用いて異常検知を行う場合を考えると，検知された異常の要因分析などを目的として，出力値に影響を与える説明変数の特定が重要となる．各説明変数に対する影響の大きさのことを *Attribution*，その値自体を *Attribution map* と呼ぶことがあり，本研究においてもこの呼称を採用する．

線形モデルを利用した場合は，*Attribution* の抽出に係数の情報を利用することができるが，ニューラルネットを利用する場合は入力と出力の関係が明示的に表現されていない（ブラックボックスになっている）ため，モデルから直接得ることができない．

一般に，産業分野においては，化学プラントなど，分析対象の状態が時間的に大きく変化するなど線形モデルでの表現が困難なものも多く，ニューラルネットを用いた場合の *Attribution* 抽出技術の確立が重要となっている．

ニューラルネットを用いて *Attribution* を抽出する技術としては，画像分類の分野で盛んに行われている．[Ancona 17, Binder 16, Smilkov 17]．一方で，IoT データを用いた異常検知や品質の予測などの応用においては，多変量時系列データを入力して数値を出力する回帰モデルに対する *Attribution* の抽出が重要となるが，研究例は少ない．

本研究では，ニューラルネットのモデルにおいて，出力の各入力値に関する偏微分値を算出し，*Attribution map* として抽出する手法 [Simonyan 13, Smilkov 17] を時系列データに対して拡張した手法 *Time-Smoothgrad* を提案し，人工データと実データによる回帰問題に対して評価を行なった．

本研究の主要な貢献は以下である．

- 線形回帰などの統計的手法と比較して，*Attribution map* を利用した手法の実問題への優位性を指摘し，数値データおよび回帰問題における有用性を示した．
- 状態の時間変化を伴う実問題への応用のため，*Smoothgrad* を時系列に拡張した手法を提案した．人工データ及び実データに対して，勾配をそのまま用いるベースライン手法と比較した有効性を示した．

連絡先: 切通恵介，NTT Communications Corporation. kiritoshi@ntt.com

2. 関連研究

ニューラルネットの入出力の関係性を分析する研究は，可視化や解釈性という文脈で提案されており，その一つとして逆伝播を用いた *Attribution map* 抽出手法が存在する．

Simonyan ら [Simonyan 13] は出力に対する入力の偏微分値をネットワーク構造の逆伝播を用いて計算する手法を提案している．この手法はノイズが多く含まれ，正確性に欠けることが指摘されており [Ancona 17]，このノイズを取り除くために様々な拡張が提案されている．[Binder 16, Shrikumar 17]

Smoothgrad[Smilkov 17] はガウシアンノイズを加えた画像を複製し，それぞれに対して出力に対する入力の偏微分値を計算し，平均することでノイズを抑える手法を提案している．

Ancona ら [Ancona 17] は上記の *Attribution map* 抽出手法が入力の出力に対する勾配を基に計算できることを証明し，各手法を様々なネットワーク構造や画像，文章などのデータセットで比較している．

これらの手法と比較して，本研究は *Smoothgrad* を時系列に拡張した手法を提案している．また，提案手法は画像分類問題に限らない多変量時系列データに対する有用性を実験で評価している．

3. Attribution 分析

3.1 線形回帰

線形回帰は入出力間の影響を論じる際，最も利用される手法である．$x_i(i = 0 \cdots k)$ と y をそれぞれ k 次元の説明変数と一次元の説明変数とすると，線形回帰のモデルは次のように表される．

$$y = \sum_{i=0}^{k} w_i x_i + b_i$$

この学習後のモデルの重み w_i を比較することで出力に対する入力の影響を算出することができる [Cook 77]．しかし，モデルの重みを分析する手法は実問題への応用を行う場合，以下の不足点が存在する．

(1) ニューラルネットのような複雑かつ重みを多層に持つようなモデルに対して重み w_i を解析することは，モデルの複雑性と解釈性はトレードオフであり，一般的に難しい．

(2) モデルの重みを分析することはモデルに対する *Attribution* の分析は可能であるが，状態といったデータ特有の影響を分析できない．例えば異常検知を行うモデルを作成した際に，重みの大きな入力を提示したとしても，正常状態と異常状態のどちらに対して影響をもつかを明らかに出来ない．

以上より，モデルだけではなく，データを考慮した影響度の分析が必要である．

3.2 ニューラルネットの Attribution

Simonyan らは画像分類問題において *Attribution map* としてニューラルネットの出力に対する入力の勾配を取ることを提案している．$S_c(x)$ をモデルの最終層の Softmax 層の前のクラス c の出力，x を説明変数とおくと，*Attribution map*M_c は以下の式で計算される．

$$M_c(x) = \left| \frac{\partial S_c(x)}{\partial x} \right| \tag{1}$$

ここで得られる *Attribution map* は入力の出力に対する影響度を示すが，勾配にノイズが多いことが知られている．

Smoothgrad[Smilkov 17] は，画像にランダムなガウシアンノイズを与えた複製に対してそれぞれ Simonyan らの手法を用いて入力の出力に対する偏微分値を計算し，平均することで最終的に得られる *Attribution map* のノイズを減らす手法である．Smoothgrad $\hat{M}_c$ は以下の式で計算される．

$$\hat{M}_c(x) = \frac{1}{n} \sum_{1}^{n} M_c(x + \mathcal{N}(0, \sigma^2)) \tag{2}$$

これらの手法はモデルの重みを見る手法と比較して，前章で上げた二つの不足点を解消できる可能性がある．

(1) ニューラルネットのような複雑なモデルに対して影響度を算出することができる．特に，*Smoothgrad* はノイズに対する頑健性が高い．

(2) モデルではなく，データに対する動的な Attribution を抽出することができる．式 1 で表されるように，*Attribution map* は各データに対して計算されるため，状態変化が生じてもその状態における *Attribution* を得られる．例えば前章の (2) においても，異常状態に対する *Attribution* を出力することができる．

これらの *Attribution map* 抽出手法は画像分類において有用性が評価されているが，いくつかの不足点が存在する．まず，主に画像データにおける有用性しか示されておらず，産業における実問題で利用される温度や圧力といったセンサーから収集される時系列のセンサーデータに対して有用性を確かめていない．また，分類問題に対してだけでなく，回帰問題に対する有効性を確認できていないことが挙げられる．

4. 提案手法

Smoothgrad のアイデアを用いて，ノイズを減らしつつ *Attribution* を得る手法を時系列データに適用することを考える．また，実問題への対応として，状態変化推定を行うための手法を提案する．

4.1 Time-Smoothgrad

IoT データに代表される実環境で収集されるデータを用いたモデルの構築には，入力データの生成にマルコフ性を有し，入力と出力との関係が（局所的に）線形となる，下式のような仮定を置くのが自然である．

$$x_t = F(x_{t-1}) + v_t \tag{3}$$
$$y_t = H_t(x_t) + w_t \tag{4}$$

ここで，F_γ と H_t は関数，v_t, w_t はノイズである．実環境における時系列データは一般的にノイズを含んでおり，時系列データをモデリングする際，式にノイズを加えることが多い．

我々はニューラルネットワークが時系列データを学習するときに式 3 のような時系列モデルを近似していると仮定し，*Smoothgrad* を説明変数に適用することを考える．

Smoothgrad において，各入力画像はランダムノイズを加算した上で複製される．複製された各画像から入力の出力に対する偏微分値を抽出し，それぞれ入力画像毎に平均されることにより最終的な *Attribution map* を得る．

IoT データにおいては画像に比べ，ノイズを加えた複製の正当性を視覚的に確認することは難しいため，本研究では *Attribution Map* の計算に多数の複製を用いることはせず，時系列データに内在するノイズ成分である v_t の存在を仮定し，長さ k の幅を持つ移動窓内の偏微分値の平均を用いることを提案する．

$$M(x_t) = \frac{1}{k} \sum_{i}^{k} \frac{\partial y_{t-i}}{\partial x_{t-i}} \tag{5}$$

ここで k は影響が生じる期間に依存するハイパーパラメータである．例えば，長い期間における影響を観察する場合は k が大きく，逆に短い期間における影響を抽出する場合は k を小さくすれば良い．また，回帰モデルを利用する場合は，最終層の Softmax 層の一つ前ではなく，最終層の出力そのものを y_t として用いる．式 5 により *Attribution map* を抽出する手法を *Time-Smoothgrad* と名付ける．

5. 実験

我々は回帰問題における提案手法及び既存手法による *Attribution map* 抽出の実験を人工データ，実データについて行い，それぞれ評価を行った．

表 1: 回帰の評価値

	ピアソン相関	平均二乗誤差
状態数 3	0.870	0.679
状態数 10	0.727	1.67

5.1 人工データ

プラントなどで収集される IoT データに特徴的な，入力時系列データの状態および，出力データの生成過程が時間的に変化する環境を表現するため，時系列モデルの一つである *Auto regression*(AR) モデルと式 3 に基づき，時系列データを以下の式で生成する．

$$x_t = \alpha_{1s} x_{t-1} + \alpha_{2s} x_{t-2} + v_t \quad (6)$$

$$y_t = F_s^T x_t + w_t \quad (7)$$

ここで，F は目的変数 y_t に対する説明変数 x_t の係数ベクトル，α は AR モデルの係数を表す．ここで，F は目的変数に対する説明変数の影響を示す重要な係数である．v_t, w_t, x_0 はそれぞれ正規分布から生成する．実験においては，一定期間で状態 s を変動させることで，係数 F，α も変化させる．これにより x_t と y_t は状態が変化する時刻でその関係性が変化する．

実験において，状態数 3，10，説明変数の次元数 10，1000 のデータをそれぞれ用意し，式 6 で目的変数 y_t を生成した．各状態は 5000 データを含み，2500 データずつ学習データとテストデータに分割し，学習とテストデータに対する *Attribution map* の抽出を行なった．

まず，勾配の移動平均を用いない手法をベースラインとして *Time-Smoothgrad* との比較を行った．モデルは多層パーセプトロンを用いて x から y を回帰し，それぞれの手法で抽出される *Attribution map* と，時系列変化する実際の関係を示す F を比較した．回帰の評価を表 1，*Attribution map* の比較をヒートマップとして図 1b に示す．

図 1b を比較すると，*Time-Smoothgrad*，ベースライン手法の共に係数 F_s の変化を表せている．特に，*Time-Smoothgrad* は明らかにノイズが減少しており，F_s とより類似していると考えられる．

係数 F_s と *Attribution map* の値の類似度を測るため，評価尺度として状態毎の *nDCG*[Järvelin 00] を用いる．*nDCG* の値域は 0 から 1 であり，F_s の値による状態毎の要素のランキングと *Time-Smoothgrad* で得られた *Attribution map* の状態毎の平均値のランキングを比較する．2 つのランキングが近いほど，また高い値を持つ要素が上位にいればいるほど高い値を示す．3 状態における結果を表 2 に示す．各状態において 1 に近い値をしてしていることから，係数と提案手法は強く類似していると言える．

また，10 状態における各状態における *nDCG* と予測値と実測値の相関との関係を示すリフトチャートを図 2 に示す．チャートが示すように全ての状態において高い *nDCG* を示す事は出来なかった一方，予測値と実測値の相関が高くなればなるほど係数 F_s とのランキングの相関が高くなる傾向が見られる．

5.2 実データ

実験における実データとして，PAMAP2 Physical Activity Monitoring Data Set[Reiss 12] を使用した．このデータセットは人間が複数の行動をした際の動きをウェアラブルセンサーで取得した 52 次元の時系列データで構成されている．センサーは手，胸，足の三箇所に取り付けられ，それぞれ体温や 3 次元方向の速度，加速度，磁気を取得している．センサーをつけた 8 名の被験者が Walking，Cycling といった 14 個の行動をシナリオに沿って実行している．すなわち，このデータセットは行動の変化という明確な状態変化を持っているといえる．

表 2: 3 状態におけるそれぞれの *nDCG* スコア

状態 1	状態 2	状態 3
0.953	0.950	1.0

実験においては目的変数として 1 つのセンサを選択し，それ以外のセンサーを説明変数としたニューラルネットの回帰モデルを作成する．学習された回帰モデルと入力データを用いて *Attribution* の抽出を行うことで，出力センサ値で表されるような特定の動作に対して大きな影響与えるセンサの特定が期待できる．ただし，目的変数と同じ部位のセンサーは極端に高い相関を示す可能性があるため説明変数から取り除くものとする．今回は，胸の磁気センサーの x 軸方向を目的変数として，手，足の全てのセンサーを説明変数とする回帰問題としてモデルを作成した．八人の被験者のうち一人をテストデータ，七人を学習データとして多層パーセプトロンを用いて学習を行なった．

人工データと同様に．提案手法で抽出した *Attribution map* を図 3 に示す．縦軸はセンサー番号，横軸はユーザの行動を示す．ただし，0 は行動の移行時の *Attribution map* である．

結果として，全ての行動において手の磁気センサーであるセンサー１５，１６が高い値を示した．また，Vacuum Clearning や Lying，Nordic Walking における *Attribution map* は Walking などの他のセンサーと比べてセンサー１５，１６以外の値が高くなるなど，異なる性質を示した．

6. 考察

回帰・数値データへの利用可能性

図 1 で示される人工データの *Attribution map* と係数 F_s の比較や，表 2 に示される *nDCG* の値に基づくと，数値時系列データの回帰問題において，提案手法，既存手法共に重要な特徴を抽出できていると考えられる．また，実データへの実験において時系列センサーデータへの適用可能性を示すことができた．

時間変化する Attribution の抽出

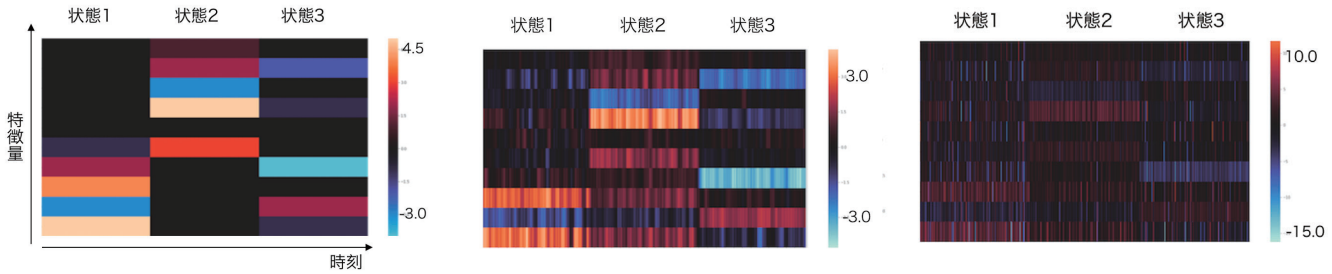
3.2 章で示す通り，実問題においては，時間変化する Attribution を抽出できるという点において，従来手法より *Attribution map* を利用する手法が適していることを述べている．実際に図 1b，1c に示すように，時系列データにおける時間変化する入力の出力に対する Attribution を捉えることができた．すなわち，品質予測や異常検知といった実問題においても *Attribution map* 抽出手法は有効である可能性を示している．

ノイズ除去

図 1c，図 1b にしめす人工データを用いた実験において，Simonyan らの既存手法と提案手法のヒートマップを比較すると，提案手法の方がより鮮明かつノイズを減少させており，重要な入力を選択できていると考えられる．

結果の解釈性

実データにおける実験においては，出力である胸部の磁気センサーに対して，図 3 に示す *Attribution map* は他の部位の



(a) 係数 F_s　(b) Time-Smoothgrad　(c) ベースライン

図 1: 3 状態における *Attribution map* の比較（縦軸: 特徴，横軸: 時刻）

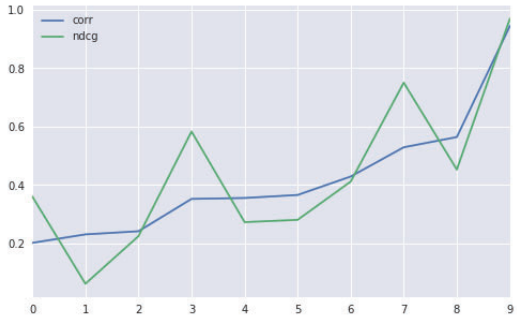


図 2: 相関と *nDCG* のリフトチャート（相関の低い状態の順に並べて nDCG をプロット）

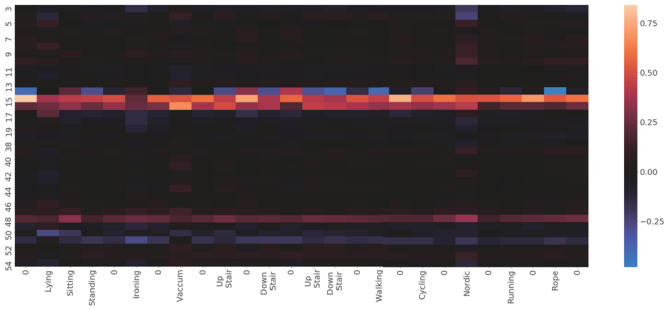


図 3: *Attribution map* のユーザの行動毎の平均

磁気センサーの関連が高いことを示している．磁気センサーは地軸からの向きを示すために，各部位の値は関連を持つ可能性が高く，妥当な結果であると言える．また，Nordic Walking や Vacuum Cleaning は他の行動と比較して手のセンサーにおいて高い値を示している．これらと他の行動との違いは道具を利用していることにある．

これらの解釈はモデルがどのような予測を行なったかを理解する上で非常に有用である．特に産業における異常検知や品質の予測といった実タスクにおいては，*Attribution map* と専門家の知見を照らし合わせることで，モデルの信頼性を得ることができる．さらに，化学プラントをはじめとするセンサー間の関係が不明な複雑な系においても，本手法は *Attribution map* からその分野における新たな発見を得られる可能性がある．

7. 終わりに

本研究で，我々は線形回帰などの手法と比較して *Attribution map* 抽出手法の実問題への有用性について言及し，人工データ，実データにおける実験を行い時系列データ及び回帰問題においての有効性を評価した．さらに，*Smoothgrad* を時系列に拡張した *Time-Smoothgrad* を提案し，実験においては移動平均を用いない素の勾配を用いるベースライン手法への優位性を示した．

今後の課題として，RNN や CNN といった他のネットワークアーキテクチャや，より多様な時系列データにおける有用性の調査が上げられる．また，*Attribution map* を抽出するその他の手法 [Binder 16, Shrikumar 17] との比較を行う必要がある．

参考文献

[Ancona 17] Ancona, M., Ceolini, E., Öztireli, C., and Gross, M.: A unified view of gradient-based attribution methods for Deep Neural Networks, *arXiv preprint arXiv:1711.06104* (2017)

[Binder 16] Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R., and Samek, W.: Layer-wise relevance propagation for neural networks with local renormalization layers, in *International Conference on Artificial Neural Networks*, pp. 63–71Springer (2016)

[Cook 77] Cook, R. D.: Detection of influential observation in linear regression, *Technometrics*, Vol. 19, No. 1, pp. 15–18 (1977)

[Järvelin 00] Järvelin, K. and Kekäläinen, J.: IR evaluation methods for retrieving highly relevant documents, in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 41–48ACM (2000)

[Reiss 12] Reiss, A. and Stricker, D.: Introducing a new benchmarked dataset for activity monitoring, in *Wearable Computers (ISWC), 2012 16th International Symposium on*, pp. 108–109IEEE (2012)

[Shrikumar 17] Shrikumar, A., Greenside, P., and Kundaje, A.: Learning important features through propagating activation differences, *arXiv preprint arXiv:1704.02685* (2017)

[Simonyan 13] Simonyan, K., Vedaldi, A., and Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps, *arXiv preprint arXiv:1312.6034* (2013)

[Smilkov 17] Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M.: Smoothgrad: removing noise by adding noise, *arXiv preprint arXiv:1706.03825* (2017)