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

特徴アテンションを用いた時系列データのアトリビューション抽出

Feature Oriented Attention to Extract Attribution for Multivariate Time-Series Data

浅野秀平 *1　切通恵介 *1　泉谷知範 *1
Shuhei Asano　Keisuke Kiritoshi　Tomonori Izumitani

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

In the application of neural networks to industries such as manufacturing, it is important to extract which features have an important role in model output from the viewpoint of model reliability and understanding the couse of event. Attention mechanisms in neural networks is generally used to handle dynamic dependencies between time series, such as in natural language processing. In this research, we apply attention as a weight of each feature and propose a method to extract attribution of each feature in model output. We also examined its effectiveness using real sensor data.

1. はじめに

近年，ニューラルネットは画像，音声，自然言語といった幅広い分野で研究がなされている．しかし，ニューラルネットの産業への応用には，学習したモデルがブラックボックスであるという大きな課題が存在する．入出力の過程が自明でないニューラルネットでは，線形モデルのように出力に対する入力の要因を特定することが難しく，これはモデルに対する信頼性や，人が取るべき次の行動が分からないなど運用上の問題となる．

ニューラルネットの入出力間の依存関係 (アトリビューション) を抽出する研究は，画像や自然言語処理において広く提案されている [Selvaraju 17, Vaswani 17]．Attention は主に自然言語処理において再帰型ニューラルネットと組み合わせる形で研究がなされており，単語系列間の動的な依存関係を捉えることや，副次的な恩恵として単語系列毎のアトリビューションを可視化することに使われている．

本研究では，センサデータのような多変量時系列データを対象として，attention を特徴量方向に独立性を保ったまま適用する事で，各特徴量のアトリビューションを抽出する手法を提案する．また，実問題での有用性を評価する為に，人の運動を計測したセンサデータセットを使って実験を行なった．

2. 関連研究

ニューラルネットにおいて attention は主に自然言語処理で利用されており [Vaswani 17]，各入力単語系列の影響を時間減衰なく最終的な出力に与える手法として用いられている．また，各時刻における attention を可視化することで，最終的な出力に影響を与えた単語のアトリビューションを抽出することができる．

また，ニューラルネットの出力に対するアトリビューションを分析する研究は，可視化や解釈性という文脈で提案されている．その一つとして，Simonyan ら [Simonyan 13] は画像分類問題において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．また，そのノイズを減らし時系列拡張を行った手法が提案されている [Kiritoshi 18]

連絡先: 浅野秀平，NTT Communications Corporation, shuhei.asano@ntt.com

これらの手法と比較して，本研究は画像分類や自然言語処理ではなく，センサなどの時系列データへの適用を目的としている点が異なる．また一般的なニューラルネットの attention は出力に寄与した時系列方向のアトリビューションを示す一方，本手法では特徴量毎のアトリビューションを抽出している点が異なる．また，Simonyan らの手法は入出力の偏微分値に注目しているが，本研究はネットワークの中間出力として，特徴量毎に計算された feature map と attention の積を与えることで，明示的に出力に寄与した特徴量が得られる点が異なる．

3. 提案手法

提案モデルの概要を図 1 に示す．提案モデルは，特徴量毎に独立した特徴抽出を行うネットワークと，特徴量毎の重み (attention) を変化させるネットワーク (以下 attention network)，およびそれらの出力を元にタスクに応じた最終的な出力を計算するネットワークから構成される．入力データ X に応じて変化するモデル内部の attention を観察することで，どの特徴量がタスクにおいて重要であったか推測する．

3.1 特徴 *attention* によるアトリビューション抽出

本手法はタスクに対する各特徴量の重要度を測る事を目的とする．よって attention はデータの時系列方向ではなく特徴量方向に適用する．入力するデータ X の系列長を T，特徴量の数を C とした時，出力する attention は長さ C のベクトル a となる．本研究ではこれを特徴 attension と呼ぶ．

特徴 attension を生成する attention netowrok は CNN などによって構成し，出力層の活性化関数に softmax 関数を使用する．このネットワークの役割は直接的にタスクの答えを求める事ではなく，入力データの全体の様相を捉え，どの特徴量にタスクを解く上で重要な情報があるかを出力する事にある．

次に attention を掛け合わせる対象である feature map $V \in \mathbb{R}^{F \times C}$ を $V = W_f X^{\mathrm{T}}$ によって得る．ここで $W_f \in \mathbb{R}^{F \times T}$ は学習した重みを表す．この処理は入力 X に対し，特徴量方向にカーネルサイズを 1 とした F 個の畳み込みを行なう操作と同一である．カーネルサイズを敢えて 1 に限定する事で，特徴量間で情報が混ざる事を防ぎ，後のアトリビューションの解釈を容易とする．

次に以下の式によって a と V から長さ F のベクトル m を

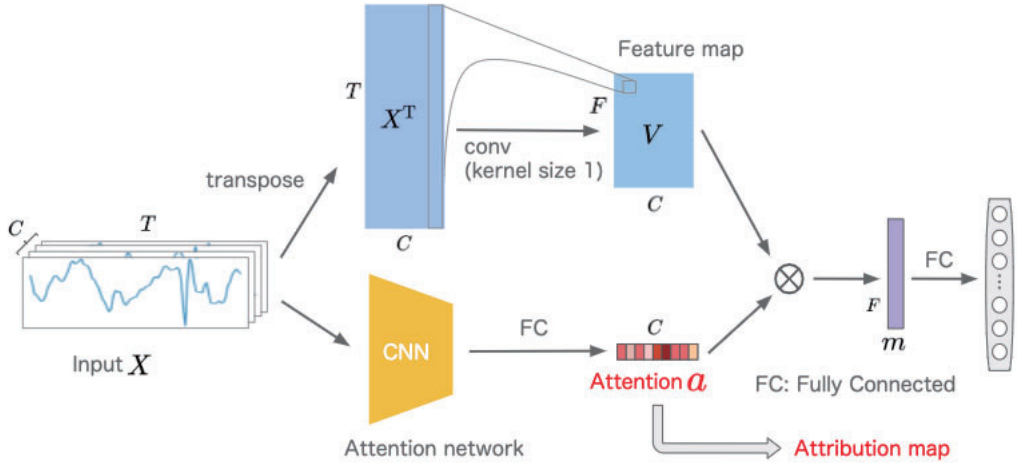

図 1: モデル全体

得る．V を $v = (v_1, v_2, ..., v_C)$ とした時，

$$m = \sum_{i=i}^{C} a_i v_i \tag{1}$$

得られた m は全結合層などを介してタスクに対する最終的な出力に使用される．Attention netwrok や W_f を含むネットワーク全体は，タスクに応じた損失の誤差逆伝播によって学習する．

最後に学習したモデルを使い，入力データ X に対する attention の平均を求める事で，各特徴量のアトリビューションを求める．

4. 実験

時系列センサデータからなる PAMAP2 Physical Activity Monitoring Data Set[Reiss 12] を使い，各センサに対するアトリビューション抽出を行なった．このデータセットは，9 人の被験者に歩行やサイクリングといった 18 種類の異なる行動を行わせ，複数のウェラブルセンサによってそれらの行動の磁気や加速度の時間変化を計測したものである．

今回は，データセットの中から少数の被験者でしか計測されていない watching TV などのデータは取り除き，手足の 34 次元のセンサデータから 12 種類の行動を分類する問題をタスクとして設定した．入力 X は 1 秒間 (100 点) のデータをランダムに切り出すことにより作成した．

上記のタスクを学習後，各行動における各特徴量のアトリビューションを表すために，複数の入力から計算された attention を平均したもの（以下アトリビューションマップと呼ぶ）を用いた．更に学習によるばらつきを抑える為に，異なる初期パラメータを使ってモデルを 30 回学習し，各モデルから得たアトリビューションマップを平均した結果を図 2 に示す．

データセットにおいて 17〜20 番，および 51 番〜54 番のセンサは無効なデータであることが説明されており，これらのセンサの中に高いアトリビューションを示すものは無かった事から，提案法は重要な特徴量の選別に有効である事が示唆される．

更に，より直接的に一部のセンサを被験者の行動と関係のないランダムなノイズに置き換え，同様の実験を行なった結果を図 3 に示す．対象のセンサでは，アトリビューションがゼロに近い値を示した事からも，本手法が入出力間の依存関係の抽出に置いて有効であると考えられる．

5. まとめ

本研究では attention を用いる事で時系列データの特徴量毎のアトリビューションを可視化する手法を提案し，実センサデータによる実験によってその有効性を示した．

次の課題として，回帰問題などの他のタスク条件下での実験や，偏微分値に基づくアトリビューション抽出の手法との比較を行う必要がある．

参考文献

[Kiritoshi 18] Kiritoshi, K., Ito, K., and Izumitani, T.: Capturing Time-Varying Influence Using an Attribution Map Method for Neural Networks, *Workshop on AI for Internet of Things* (2018)

[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)

[Selvaraju 17] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D.: Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization, in *The IEEE International Conference on Computer Vision (ICCV)* (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)

[Vaswani 17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I.: Attention is All you Need, in Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. eds., *Advances in Neural Information Processing Systems 30*, pp. 5998–6008, Curran Associates, Inc. (2017)

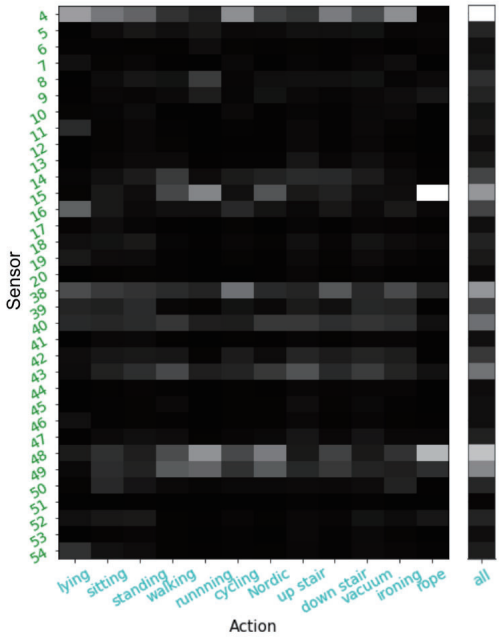

図 2: アトリビューションマップ　(縦軸がセンサの番号を表し，横軸が被験者の行動を表す．明度がアトリビューションの高さを表す．右端のグラフは全ての行動のアトリビューションを足しわせたもの．センサの番号は元のデータセット [Reiss 12] に合わせた)

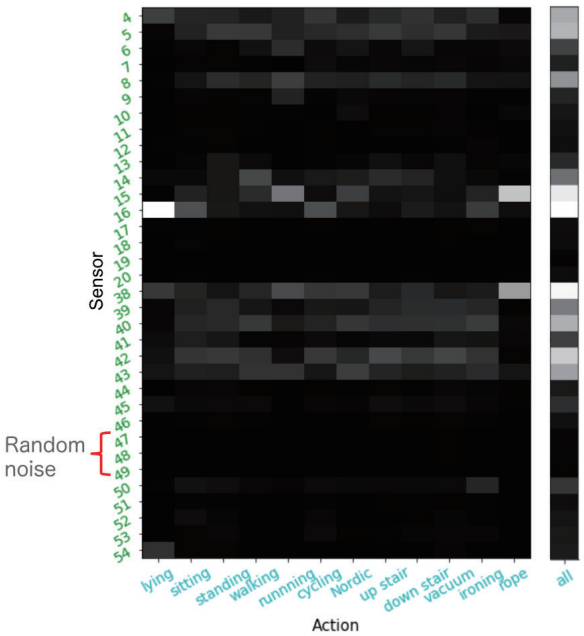

図 3: 一部センサをランダムなノイズに置き換えた条件下でのアトリビューションマップ