Vol.2020-HCI-190 No.13
Vol.2020-UBI-68 No.13
2020/12/9

行動認識ニューラルネットの理解に向けた Activation Maximizaionの活用に関する検討

吉村 直也[1,a)]　前川 卓也[1,b)]　原 隆浩[1,c)]

概要：ウェアラブルセンサを用いた行動認識技術は，高齢者の見守りやヘルスケア，スマートホームなどのアプリケーションを実現するための基盤技術の 1 つであり，近年盛んに研究が行われている．行動認識技術の研究において，行動認識モデルとしてニューラルネットワークの利用が増加しているが，ネットワークの内部動作を解釈するための手法はまだ十分に検討されていない．ネットワークの内部動作が分析できるようになれば，学習したモデルの検証やモデル構成の改良などに活用することができる．画像処理分野では，ニューラルネットワークの内部的な機能を可視化する Activation Maximization (AM) という手法が提案されている．AM では，ネットワーク中の注目したユニットの出力を最大化する入力信号を生成することで，注目したユニットが抽出する特徴を可視化する．しかし，既存の AM は画像信号に対して最適化されており，加速度データを入力とするネットワークに直接適用すると，ノイズが多く解釈が難しい可視化結果となってしまう．本研究では，ノイズが少なく，実在する加速度データに近い信号を生成することを目指し，加速度信号に適した AM の手法を提案する．これを実現するため，本研究ではネットワーク中のユニットの出力値を用いて正則化を行う手法を提案する．提案手法は 2 つのデータセットを用いて，定性的・定量的に評価を行い，有効性を確認した．

キーワード：行動認識, ニューラルネットワーク, 可視化技術

1. はじめに

ウェアラブルデバイスの普及に伴い，ウェアラブルデバイスに搭載された慣性センサを用いた行動認識技術・ジェスチャ認識技術の研究が活発に行われている．これらの技術は高齢者の見守りや，ヘルスケア，スマートホームなどに応用されることが期待されている．近年の行動認識・ジェスチャ認識技術の研究ではニューラルネットワークを用いた手法が盛んに研究されている一方で，行動認識ニューラルネットワークの推論過程はブラックボックスとされており，どのような処理が行われているか明らかでない．

ニューラルネットワークの内部動作を分析するための技術は，画像処理の分野において活発に研究されている [5], [8], [11], [12], [14], [17], [19], [23], [24]．しかし，加速度信号を入力とする行動認識ニューラルネットワークに対する分析技術は十分に研究されていない．行動認識ニューラルネットワークの内部動作に関して理解を深めるための技術は，次のような点から行動認識技術の研究・開発において非常に有益な情報をもたらすと考えられる．

- **学習した行動認識モデルの検証**: 畳み込みニューラルネットワーク中のあるユニットが抽出する特徴を可視化することができれば，ネットワークを適切に学習することができたか検証することができる．例えば，ネットワーク中の多くのユニットで類似した特徴ばかりが抽出されていると分かった場合，学習したネットワークの構成は冗長であり，想定するデータセットに対しては，より規模が小さいネットワークの方が適していると考えられる．このような情報は，各層におけるユニットの数やネットワークの深さなど，行動認識ニューラルネットワークのチューニングに有用であると考えられる．
- **行動認識モデルに対する特徴地図 (Activation Atlas)**: 特徴地図 [3] は，画像認識モデルが抽出した特徴を 2 次元平面上に表現したものである．本研究で注目する行動認識ニューラルネットワークにおいては，本研究で提案する手法を用いて生成した加速度信号を用いることで特徴地図を作ることができる．作成した行動認識版の特徴地図を用いることで，ある行動に特有な波形や，行動とそれを構成する基本動作などの関係

[1] 大阪大学大学院情報科学研究科
[a)] yoshimura.naoya@ist.osaka-u.ac.jp
[b)] maekawa@ist.osaka-u.ac.jp
[c)] hara@ist.osaka-u.ac.jp

© 2020 Information Processing Society of Japan

Vol.2020-HCI-190 No.13
Vol.2020-UBI-68 No.13
2020/12/9

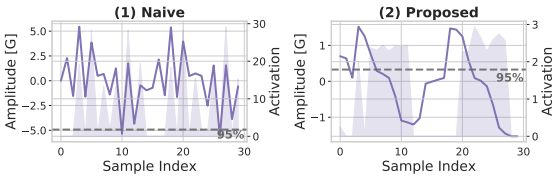


図 1 Activation Maximization を行動認識を行う CNN のあるユニットに対して適用した例．左右の図はそれぞれナイーブな Activation Maximization と提案手法を適用した結果である．横軸はサンプルのインデックス．背景のヒストグラムと右側縦軸は対応するサンプルをネットワークに入力した際の，ユニットの出力値を表す．

を視覚的に把握することができる．また，ジェスチャを用いた認証技術など，推定過程の透明性が求められるアプリケーションにおいて技術の信頼性を分析するためにも活用できると考えられる．

ネットワークが学習したコンセプトを明らかにするため，画像処理分野において Activation Maximization (AM) と呼ばれる注目したユニットの出力値を最大にする入力信号を探索する手法が用いられている [5], [23]．AM を用いることで，非常に大きな出力値を示す入力信号を生成することができるが，ナイーブな AM では人間には理解することが難しい信号が生成される問題が指摘されている [9], [12], [13]．この問題の解決策として，正則化手法 [5], [23] や，事前に学習した情報を活用する手法 [11], [12] などが検討されている．図 1(1) は，ナイーブな AM で生成した加速度信号と対応するユニットの出力値を示したものである．また，図中の点線は全ての学習データをモデルに入力して得られた注目ユニットの出力値の集合における 95%にあたる値を示す．ナイーブな AM ではノイズを多く含む加速度信号が生成されていることがわかる．さらに，生成した信号をネットワークに入力すると，対応する出力値は学習データを入力した場合の 95%点に対して 15 倍と非常に大きな値となっており，対応する加速度信号の振幅も非常に大きな値になっている．画像信号であれば各ピクセルの値は $0-255$ に限定されている．この場合 AM ではクリッピングを用いて限定された値域の中で最適化を行えば，異常値を抑制することができる．一方で，加速度信号には基本的に値域の制限がないため，生成する信号の値域を制限することが難しい．したがって，異常に大きな振幅値などといった，加速度データにおけるノイズを抑制する手法が必要である．

本研究では，加速度データを用いた行動認識に適した AM の最適化手法の検討を行う．異常に大きな振幅値を抑制するため，本研究ではネットワークの出力値を用いた新たな正則化手法を提案する．本研究の技術的貢献は以下の通りである．(i) 行動認識ニューラルネットワークに対する Activation Maximization において，ネットワークの出力値を用いた正則化手法を提案する．提案する正則化手法を用いることで，図 1(2) のように，実在するデータと比較して妥当な振幅値の大きさを持つ信号を生成することができた．(ii) 提案手法に関して，公開データセットを用いて定性的かつ定量的に評価を行い，学習データに存在する加速度信号に似た信号を生成できていることを確認した．

以降の構成は次の通りである．2 節では，行動認識ニューラルネットワークの可視化および，ニューラルネットワークの可視化に関する研究を紹介する．3 節では提案手法を説明し，4 節・5 節において評価を行う．6 節では，提案手法を用いた AM の応用例として特徴地図を紹介する．

2. 関連研究

ニューラルネットワークの推論過程を解明することは，学習されたモデルの検証やパフォーマンスの向上のために重要である．このブラックボックスと考えられているネットワークの推論過程を明らかにするために，画像処理分野においてニューラルネットワークの可視化技術が活発に研究されてきた [8], [18], [19], [24]．行動認識の研究分野においても，同様にニューラルネットワーク可視化技術が研究されている [6], [15], [25]．しかし，これらの手法はネットワーク内で特徴抽出を行う基本単位であるユニットの粒度で，学習されたコンセプトを説明することはできない．Activation Maximization (AM) は，勾配法を用いてネットワーク中のあるユニットの出力値を最大にする入力信号を探索する手法である (詳細は 3.2 節で説明する)．AM は，モデルの出力に対応する分類クラスに関するコンセプト [11], [12], [23] や，中間層のユニットの機能 [5], [23] を可視化するために利用されている．本研究では，加速度の時系列データに注目し，学習したモデルの中間層におけるユニットの機能の解明を AM を用いて試みる．

AM は，高周波のノイズを多く含む信号を生成する傾向がある．吉村ら [22] は，AM の行動認識ネットワークへの適用に関して検討を行っており，実際のデータには存在しない高周波ノイズを除去するため，ローパスフィルタを用いる手法を提案している．しかし，この手法では異常に大きな振幅の生成などを抑制することはできない．本研究では，ネットワークの出力値を利用した正則化手法を提案し，より実在する信号に近い信号の生成を試みる．

先行研究では，敵対的生成モデルを用いて加速度信号の生成を試みたものがあり，生成した信号を追加の学習データとして利用している [20], [21]．しかしこれらの研究では，単に実在するデータに類似したデータを作成しているだけであり，この信号から本研究が目標とするネットワークが学習した特徴を分析することは難しい．

© 2020 Information Processing Society of Japan

Vol.2020-HCI-190 No.13
Vol.2020-UBI-68 No.13
2020/12/9

3. Activation Maximization と正則化手法

3.1 想定環境

本研究では，ユーザの身体に装着された 3 軸加速度センサからのデータを入力として，ユーザの日常行動やエクササイズなどを認識する CNN ベースの行動認識モデルを想定する．ただし，行動認識モデルは事前に行動認識用データセットを用いて学習されたものとする．ニューラルネットワークへの入力 $\boldsymbol{X}$ は，長さ N_T のスライディングウィンドウによって切り出された加速度信号である．また，第 l 層のユニット u の出力を，$\boldsymbol{X}^{(l,u)}$ と表記する．入力 $\boldsymbol{X}$ は $(N_S \times N_T)$ 行列であり，注目ユニット u の出力 $\boldsymbol{X}^{(l,u)}$ は，$(N'_S \times N_T)$ 行列である．N_S は入力となるセンサ (軸) の合計，N'_S は畳み込みフィルタの大きさなどによって決まる定数である．

3.2 Activation Maximization (AM)

前述の通り，AM は勾配法を用いてネットワーク中のあるユニット u の出力を最大にする入力信号 $\boldsymbol{X}^*$ を生成する手法であり，この生成された入力信号からそのユニットが抽出する特徴に関する情報を得ることができる．注目したユニット u の出力を最大にする入力信号 $\boldsymbol{X}^*$ は以下の式を解くことで得ることができる．

$$\boldsymbol{X}^* = \underset{\boldsymbol{X}}{\operatorname{argmax}} f^{(l,u)}(\boldsymbol{X}) \tag{1}$$

$f^{(l,u)}(\boldsymbol{X})$ は，l 層目のユニット u が出力する特徴マップの代表値を計算する関数であり，本研究では平均値を用いる．

AM では，乱数などを用いて適当に初期化した $\boldsymbol{X}$ を繰り返し更新することで目的の $\boldsymbol{X}^*$ を生成するが，次の 2 つの問題のために解釈しやすい信号の生成が難しい．(i) 生成された信号 $\boldsymbol{X}^*$ は，高周波ノイズを多く含む傾向がある [23][*1]．(ii) 注目したユニットの出力値は非常に大きいが，実在するデータとは大きく異なる信号が生成される可能性がある．

この問題に対処するため，Le ら [5] は，式 1 を解く際に，生成している信号 $\boldsymbol{X}$ のノルムを 1 にする制約を導入することでノイズの抑制を試みた．また Yosinski ら [23] は，次の式 2 のように，高周波数成分を抑制するために式 1 に正則化項を導入した．

$$\boldsymbol{X}^* = \underset{\boldsymbol{X}}{\operatorname{argmax}} \Big(f^{(l,u)}(\boldsymbol{X}) - wR(\boldsymbol{X}) \Big) \tag{2}$$

$R(\boldsymbol{X})$ は正則化項を表す．Yosinski らは，Lp ノルムと Total Variation [16] を正則化項として用いた．Lp ノルム ($R_{Lp}(\boldsymbol{X})$) と Total Variation ($R_{TV}(\boldsymbol{X})$) は次の式で表される．

$$R_{Lp}(\boldsymbol{X}) = \left(\sum_{s}^{N_S-1} \sum_{t}^{N_T-1} (x_{s,t})^p \right)^{1/p} \tag{3}$$

$$R_{TV}(\boldsymbol{X}) = \left(\sum_{s}^{N_S-1} \sum_{t}^{N_T-2} |x_{s,t} - x_{s,t+1}| \right) \tag{4}$$

$x_{s,t} \in \boldsymbol{X}$ は，センサ s の時刻 t における加速度データの値を表す．本研究では，加速度データに適した正則化項と式 2 の修正方法を提案する．

式 2 は，ニューラルネットワークの学習のように勾配法によって解くことができる．ニューラルネットワークの学習の場合は，勾配法によって損失関数を最小化するようにネットワークの重みを更新する．一方で AM は，勾配法によって $f^{(l,u)}(\boldsymbol{X}) - wR(\boldsymbol{X})$ を最大化する．まず，信号 $\boldsymbol{X}$ を乱数など適当な値で初期化する．次に，以下の式によって $\boldsymbol{X}$ を更新する．

$$\boldsymbol{X} \leftarrow \boldsymbol{X} + \eta \frac{\partial}{\partial \boldsymbol{X}} \Big(f^{(l,u)}(\boldsymbol{X}) - wR(\boldsymbol{X}) \Big) \tag{5}$$

η は学習係数である．この更新操作を収束するまで繰り返すことで，最終的な結果 $\boldsymbol{X}^*$ が得られる．

3.3 異常に大きな出力値を抑制するための正則化手法

既存の AM では，注目したユニットの出力値が非常に大きいが，理解が困難な信号を生成することがある．例えば，1 層目のある畳み込みフィルタの重みの値が全て正の値かつ，活性化関数として ReLU を用いる場合，注目したユニットの出力値は入力信号 $\boldsymbol{X}$ の振幅に比例する．この結果として，AM は非常に大きな振幅を持つ信号を生成する．生成している信号の振幅値をクリッピングすることで加速度信号の値を制限することはできるが，入力信号の振幅値に対する感度はユニットごとに異なるため，クリッピングを行う範囲を適切に設定することは難しい．本研究ではこの異常な振幅値を抑制するために，注目ユニットの出力値 ($\boldsymbol{X}^{(l,u)}$) を用いる手法を提案する．学習データをネットワークに入力することで，各ユニットの出力値の分布を得ることができる．また，これを用いることで各ユニットに対して出力値が取りうる値の範囲を個別に設定することができる．AM の最適化の過程において，ユニットの出力に設定された値域から外れる異常に大きな値が観測された場合，本研究ではその異常な出力値に対応する入力信号にも学習データには存在しない異常値が発生していると考え，これを抑制することを試みる．

本研究では閾値 $th_{eap}^{(l,u)}$ を設定することで，ユニットの出力値の異常を定義する．例えば，学習データを全て入力して得られた出力値を用いて，2σ 法でこの閾値 $th_{eap}^{(l,u)}$ を設定する [*2]．ユニットの出力値の分布を正規分布と仮定すると，この $th_{eap}^{(l,u)}$ は学習データの出力値におけるおよそ 95%

[*1] 筆者の知る限り，AM で生成された信号に高周波成分が多く含まれる原因は解明されていない．

[*2] 2σ 法では，0 でない出力値の集合に対して平均 μ と標準偏差 σ を計算し，$th_{eap}^{(l,u)} = \mu + 2\sigma$ として定める．

© 2020 Information Processing Society of Japan

Vol.2020-HCI-190 No.13
Vol.2020-UBI-68 No.13
2020/12/9

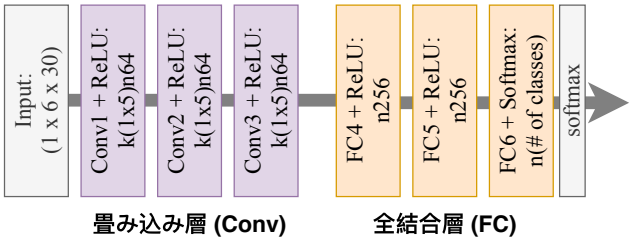


図 2 本研究で用いたニューラルネットワークの構成．“$k(1\times5)$” と “$n64$” は，各層における畳み込みフィルタのサイズとユニット数である．

点に対応する．

本研究では，この閾値を用いた正則化項「Extreme Activation Penalty (**EAP**)」を提案する．EAP は以下の式で定義される．

$$R_{EAP}(\boldsymbol{X}^{(l,u)}) = \mathrm{Mean}\left(\left\{a_{s,t}^{(l,u)} - th_{eap}^{(l,u)} \mid a_{s,t}^{(l,u)} \in \boldsymbol{X}^{(l,u)}; a_{s,t}^{(l,u)} > th_{eap}^{(l,u)}\right\}\right) \quad (6)$$

$a_{s,t}^{(l,u)} \in \boldsymbol{X}^{(l,u)}$ は，センサ s の時刻 t に対応するユニットの出力値を示す．ユニットの出力値が設定した閾値を上回った場合，閾値を超過した分の平均値がペナルティとして与えられる．このように，EAP は出力値を正則化項に直接用いて信号の異常値を抑制する．

また EAP では対応することが難しい異常に大きな出力値を抑制するため，2 つ目の手法である「Activation Clipping」を提案する．通常の AM (式 1) では，注目ユニットの出力の代表値 $f^{(l,u)}(\boldsymbol{X})$ は出力された特徴マップ $\boldsymbol{X}^{(l,u)}$ の値に対する平均値が用いられる．この平均を用いた操作では，$\boldsymbol{X}^{(l,u)}$ に含まれる局所的な異常値を抑制することが難しい．この問題を解決するため，平均を計算する前に，特徴マップ $\boldsymbol{X}^{(l,u)}$ の各値に対して以下の式 7 で表されるクリッピングを適用する．

$$clipping(a_{s,t}^{(l,u)}) = \begin{cases} th_{max}^{(l,u)} & (a_{s,t}^{(l,u)} > th_{eap}^{(l,u)}) \\ a_{s,t}^{(l,u)} & (otherwise) \end{cases} \quad (7)$$

このクリッピングを行うことで閾値を超えた分のユニットの出力値を無視することができ，AM の更新において生成している信号の振幅が大きくなり続けることを抑制することができる．

提案手法では，Lp ノルムと Total Variation に加えて，提案した EAP と Activation Clipping を組み合わせて利用する．

4. 評価方法

4.1 データセットと行動認識モデル

本研究では，Daily and Sports Activities Dataset (**DSADS**) [2] と mHEALTH Dataset (**MHEALTH**) [1] の 2 種類の公開データセットを用いて提案手法の評価を行う．

4.1.1 DSADS

Barshan ら [2] は，日常生活とスポーツに関連する 19 種類の行動に関して 8 人の被験者からデータを収集した．センサは胸・腕・脚に装着されている．本研究では，7・8 番目の被験者のデータをテストに，6 番目の被験者のデータをバリデーションに使用し，残りの被験者のデータを学習に用いた．また，右手と左足に装着された 3 軸加速度センサのデータ，合計 6 軸 ($N_S = 6$) を用いた．

4.1.2 MHEALTH

Banos ら [1] は，日常生活とエクササイズに関連する合計 12 種類の行動に関して 10 人の被験者からデータを収集した．本研究では，9・10 番目の被験者のデータをテストに，8 番目の被験者のデータをバリデーションに使用し，残りの被験者のデータを学習に用いた．また，右下腕と左足首に装着された 3 軸加速度センサのデータ，合計 6 軸 ($N_S = 6$) を用いた．

4.2 行動認識モデル

センサデータの前処理として，まずセンサデータを 30 Hz にダウンサンプリングをしたあと，正規化を行った．その後，長さ 30 pt ($N_T = 30\mathrm{pt} = 1\mathrm{sec}$)，50% オーバーラップのスライディングウィンドウを適用し，これをニューラルネットワークの入力とした．

本研究で用いるネットワークの構成を図 2 に示す．この構成は，Shared-Filter Hybrid Fusion (SF-HF) Model [10] を参考にした．ネットワークの学習には，Adam Optimizer [4] を用い，初期の学習係数 $lr = 10^{-4}$ で 200 epoch 学習した．各設定において，ネットワークは異なる乱数シードを用いて 5 回ずつ学習を行った．学習後のモデルの F 値 (macro average) の 5 回分の平均は，DSADS が 0.737 ($\sigma = 0.005$)，MHEALTH が 0.955 ($\sigma = 0.004$) であった．

4.3 Activation Maximization の実行手順

AM は基本的に，3.2 節で説明した通りに実行する．SF-HF モデルはセンサごと個別に同じ特徴フィルタを適用して特徴抽出を行う [10] ため，本研究では AM を 1 つのセンサ (1 軸) に対象を絞って行った．また，初期値として振幅 0.3 G の正弦波を $\boldsymbol{X}$ の初期値として用いた．式 5 の更新操作は，学習係数 $\eta = 10^{-4}$ で 5000 回行った．学習係数は更新を 1 回行うごとに $\gamma = 0.995$ を乗じて減衰させた．正則化項の強さを制御する重みは，$w = 0.1$ を用いた．これらのパラメータは事前の検証に基づき設定した．

4.4 比較手法

ユニットの出力値を用いて正則化を行う提案手法の有効性を検証するため，生成している信号の振幅の大きさに応じて正則化を行う手法 (**AAP**) を用意し比較する．AAP は，「Abnormal Amplitude Penalty」の頭文字であり，振幅

© 2020 Information Processing Society of Japan

の値に対して閾値 th_{aap} を設定し，生成している信号の振幅がこの閾値を超えた場合，その超過分をペナルティとして与える．閾値は両方のデータセット共通で，$th_{aap} = 1.5$ G と設定した．**DSADS** と **MHEALTH** の 2σ 法で定める閾値がそれぞれ 1.37 G と 1.57 G であり，本研究では両方のデータセット共通で，$th_{aap} = 1.5$ G とした．

評価では，AAP を含めた以下の手法と比較を行う．(1) L2 ノルムと Total Variation のみ用いる手法 (**L2+TV**)．(2) L2 ノルムと Total Variation, AAP を用いる手法 (**L2+TV+AAP**). (3) L2 ノルムと Total Variation, EAP を用いる手法 (**L2+TV+EAP**). (4-5) Activation Clipping を手法 (1) と手法 (3) に追加した手法 (**L2+TV+clip**，**L2+TV+EAP+clip**).

ベースライン手法は **L2+TV**，提案手法は **L2+TV+EAP+clip** である．

4.5 評価方法

提案手法を用いて生成した信号を評価するため，定性的評価に加えて定量的評価を試みる．生成した信号が学習データに含まれるデータと大きく異なる場合，その AM の手法は現実には存在しない不自然な加速度信号を生成していると判断することができ，ニューラルネットワークの分析手法としてのパフォーマンスが低いと考えられる．したがって，本研究では学習データに含まれる信号と生成した信号を比較することで，生成した信号を定量的に評価するための評価指標 (S_{sim}) を作成し，これを用いて評価を行う．評価指標 S_{sim} では，平均 2 乗誤差 (MSE) を用いて各ユニットに対する AM の結果 X^* に類似した学習データを，top-N_{sim} 件抽出し，抽出したサンプルに対する MSE の平均値を評価指標とする．AM で生成した信号が学習データに存在する信号に似ているのであれば，MSE の平均値は小さくなる．

評価指標 S_{sim} の計算においては，次の 2 点に注意する必要がある．(1) AM で生成した信号は，学習データに含まれる特徴的な波形を反映するものであると考えられ，その特徴的な波形の長さは入力信号のウィンドウサイズより短い．したがって，AM で生成した信号の中からユニットの出力値が大きい領域を切り出し，MSE に基づいた類似度の計算を行うべきである．(2) また，ナイーブな AM などでは最適化に失敗し，非常に大きな振幅を持つ信号を生成することがある．そのような信号は平均した MSE の値に大きな影響を与える．したがって，評価指標の計算時にはこのような AM の最適化に失敗した信号を除外して計算すべきである．具体的には，閾値 th_{fail} を設定し，信号に含まれる最大の振幅値がこの閾値 th_{fail} を超えた場合は，最適化に失敗したとみなし計算から除外する．

上記の 2 点に注意し，AM によって生成した信号 $\boldsymbol{X}^*$ と，学習データからスライディングウィンドウによって切り出した信号 $\boldsymbol{X}_i$ の類似度は次のように計算される．

(1) 生成した信号 $\boldsymbol{X}^*$ に対応する特徴マップ $\boldsymbol{X}^{(l,u)}$ に対して，長さ s_w のスライディングウィンドウを適用し，ウィンドウ毎にユニットの出力値の総和を計算する．出力値の総和がもっと大きいウィンドウに対応する信号を $\boldsymbol{X}^*$ から切り出し，$\hat{\boldsymbol{x}}^*$ とする．

(2) 学習データのセグメント $\boldsymbol{X}_i$ に，長さ s_w のスライディングウィンドウを適用し信号を切り出す．$\boldsymbol{X}_i$ において，時刻 j から時刻 $(j + s_w - 1)$ で切り出されたセグメントを $\hat{\boldsymbol{x}}_{i,j}$ とする．

(3) 切り出した $\hat{\boldsymbol{x}}^*$ と，各 $\hat{\boldsymbol{x}}_{i,j}$ に対して MSE を計算する．その中で最も小さい MSE を $\boldsymbol{X}^*$ と $\boldsymbol{X}_i$ の距離 $d(u,i)$ として定義する．

上記の手順をまとめると，$\boldsymbol{X}^*$ と $\boldsymbol{X}_i$ の距離 $d(u,i)$ は次のように書くことができる．

$$d(u,i) = \min_j \{MSE(\hat{\boldsymbol{x}}^*, \hat{\boldsymbol{x}}_{i,j})\} \tag{8}$$

最後に，注目した層の全てのユニットに対して計算した $d(u,i)$ を，距離 $d(u,i)$ の小さい top-N_{sim} 件に対して平均することで評価指標 S_{sim} を計算する．

$$S_{sim} = \frac{1}{U}\sum_{u=0}^{U-1}\left[\frac{1}{N_{sim}}\sum_{i=0}^{N_{sim}-1} d(u,i)\right], \tag{9}$$

U は注目した層に属するユニット数である．ただし最大の振幅が閾値 th_{fail} より大きいものは生成に失敗したものとして除外する．本研究では，$N_{sim} = 50$, $s_w = 10$ pt, $th_{fail} = 10$ G とした．

5. 結果

提案した EAP と Activation Clipping の有効性を示すため，定量的評価と定性的評価を行う．行動認識モデルに対する AM の先行研究 [22] において提案されれたローパスフィルターを用いる手法を本研究でも実験したが，生成した信号の多くが閾値 th_{fail} を超える振幅値を含んでいた．したがって，本節の比較手法からは除外した．

5.1 定量的評価

定量的評価の結果を図 3 に示す．図中の S_{sim} は乱数シードを変えて実行した 5 回の試行の平均値である．EAP と Activation Clipping を導入することで，評価指標 S_{sim} を大きく低減していることがわかる．特に，2 層目・3 層目に対応する「Conv2，Conv3」では，ベースライン手法の L2+TV に比べて非常に大きくスコアが改善しており，提案手法を用いることで学習データに実在する信号により類似した信号を生成できたと考えられる．

EAP と Activation Clipping は，ユニットの出力値を制御するために導入したものである．この効果を検証するために，各ユニットに対して AM で生成した信号をモデル

© 2020 Information Processing Society of Japan

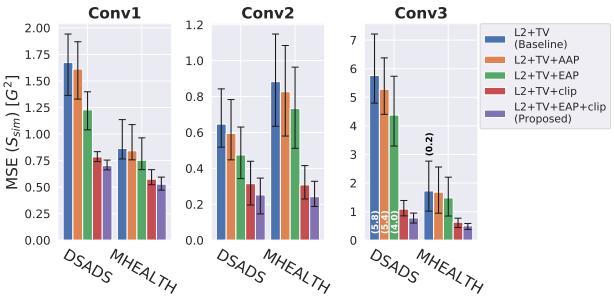


図 3 層毎に計算した評価指標 S_{sim}．「Conv1」は 1 層目の畳み込み層を表す．棒グラフに添えられている数字は，5 回の実行において最適化に失敗したユニット数の平均値である．エラーバーは 5 回の実行における最小値と最大値である．

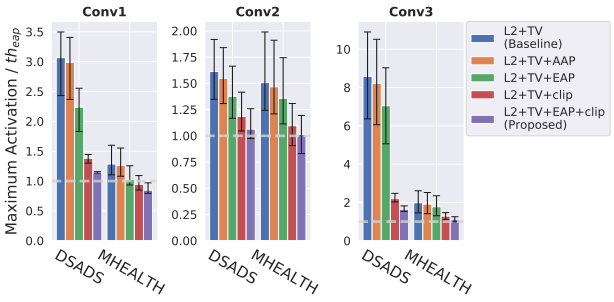


図 4 生成された信号における最大の出力値と閾値 $th_{eap}^{(l,u)}$ の比率．値は層毎に平均したものである．エラーバーは 5 回の実行における最小値と最大値である．

に入力したときに得られる最大の出力値の，閾値 $th_{eap}^{(l,u)}$ に対する比率を図 4 に示す．図中の値は，計算した比率を層ごとに平均した値である．多くの場合，提案手法では 1.0 に近い値となっており，提案手法が生成した信号には閾値 $th_{eap}^{(l,u)}$ を超える出力値を出す結果がほとんどないことを示している．このことから，提案手法がユニットの出力値を制御できていることがわかり，正則化にユニットの出力値を用いる手法の有効性が確認できる．

図 5 に，提案手法で生成した信号と MSE ($d(u,i)$) によって選択された学習データを示す．図の下段は上段の信号をモデルに入力したときに得られる特徴マップ $\boldsymbol{X}^{(l,u)}$ である．紫と青の背景色が入っている部分は MSE を計算したセグメントである．図から，「playing basketball」における右上がり波形や，「running」における山型の波形など，提案手法を用いて生成した信号は実在する信号の特徴を捉えられていることがわかる．

5.2 定性的評価

図 6 と図 7 に，あるユニットに対して生成した信号の例を示す．提案手法の効果が分かりやすいように，右上がりと右下がりの波形を抽出していると思われるユニットを 1 層目から選択した．

図 6 に，L2+TV+AAP と提案手法のそれぞれによって

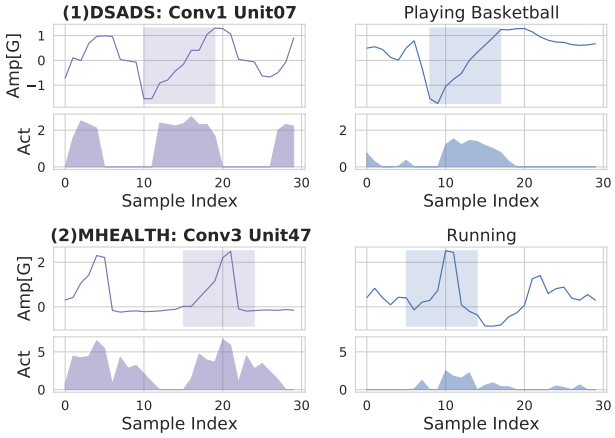


図 5 提案手法 (L2+TV+EAP+Clip) を用いて生成した信号とそれに類似した学習データ (上段)．それぞれの信号をモデルに入力したときに得られる特徴マップ (下段)．「Amp」は振幅 (Amplitude)，「Act」はユニットの出力値 (activation) を表す．これらのペアは，MSE ($d(u,i)$) に基づいて選択されたものである．

生成した信号の例を示す．図中の点線は学習データから計算した注目ユニットの出力値の 95%点を示す．注目したユニットは右下がりの波形を抽出していると考えられる．L2+TV+AAP を用いた手法では，サンプルインデックス (横軸) が $5, 10, 20, 25$ において，非常に大きな出力値を出しており，この値は 95%点のおよそ 2 倍の値である．つまり，L2+TV+AAP を用いた AM では異常に大きなユニットの出力値が発生しており，学習データとは大きく異なる波形が生成されたと言える．一方で，EAP を用いることで異常な出力値の発生が抑えられており，EAP の有効性を確認できる．

Activation Clipping の効果を確認するため，図 7 に Activation Clipping 無しの結果 (左) と，Activation Clipping を用いた結果 (右) を示す．左の図からわかるように，EAP を用いても異常な出力値を抑制することができない場合がある．一方で，Activation Clipping を導入することで，異常に大きな出力値が出ている場所の振幅がさらに大きくなることを防ぐことができ，より滑らかな信号を生成することができた．

6. 特徴地図

本節では応用例を用いて，AM の行動認識モデルの分析・開発における有効性を示す．規模が大きなモデル全体の分析においては，提案手法で生成した大量の信号を 1 つ 1 つ確認することは，効率的ではない．そこで，これらの分析を行うにあたって，提案手法を用いて生成した信号を用いて「特徴地図 (Activation Atlas)」を構築することで抽出された特徴を整理することで，モデル全体の分析を行えるようにする．「特徴地図」は，画像認識モデルに対して，そのモデルが学習した特徴を類似度に基づいてに配置した図で

© 2020 Information Processing Society of Japan

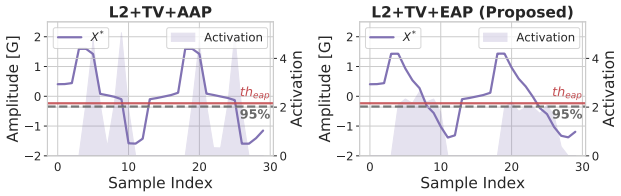


図 6 Examples of generated signals (Conv1 Unit03 of DSADS model).

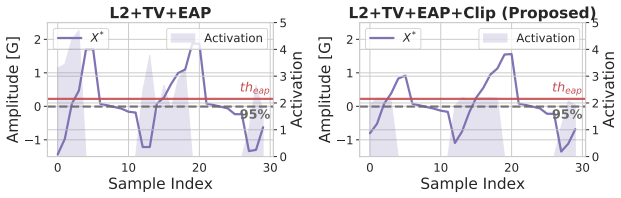


図 7 Examples of signals generated by methods with and without clipping for the same unit (Conv1 Unit29 of DSADS model).

ある．本章では，提案手法で生成した信号を t-SNE [7] を用いて 2 次元に圧縮することで，行動認識モデルにおける特徴地図を作成する．また作成した地図を用いて，ニューラルネットワークによって自動的に抽出された特徴の分析を行う．

6.1 ニューラルネットワークによって抽出された特徴の分析

ニューラルネットワークを用いるメリットの一つは，特徴抽出を自動的に行う点である．特徴地図を使うことで，抽出された特徴の分布を簡単に把握することができる．

図 8 に，MHEALTH データセットを学習したモデルの 1 層目と 3 層目のユニットに対して作成した特徴地図と生成された信号の例を示す．各ポイントは 1 つのユニット (AM で生成した信号) に対応しており，データ点の色は MSE ($d(u,i)$) によって選択された最も類似した学習データの行動クラスに応じて割り当てられている．

左の図では，データ点がクラスタを形成されている．1 層目では，エッジや平坦な波形など比較的単純な信号が生成されているとみられる．一方，右図の特徴地図ではデータ点がばらけて分布しており，生成された波形も様々な種類があることがわかる．行動認識モデルにおいても，深い層では浅い層よりも複雑な特徴が抽出されていると考えられていたが [6]，本研究では AM を用いることでこの事実を確認することができた．

図 8 の左 (3 層目) において，特徴地図の左上・中央・右下に 3 つの大きなクラスタが形成されているのがわかる．また，これらのクラスタは異なる行動クラスに対応するユニットから形成されている．多くのユニットを割り当てて少しずつ異なる特徴を抽出することで，類似した行動クラスの識別を試みているのではないかと考えられる．

このように，AM を用いることで特徴地図が作成でき，行動認識ニューラルネットワークの理解に活用できる．

7. 結論

本研究では，Activation Maximization を用いることで行動認識ニューラルネットワークの特徴抽出機能の理解を試みた．注目したユニットの出力値を最大にする信号を生成するために，学習データから推定したユニットの出力値の分布から得られる情報を用いた新たな正則化手法を提案した．公開データセットを用いて，提案手法の有効性を評価し，提案手法を用いることで実在する信号に近い信号が生成できることを定性的・定量的に確認した．今後は提案手法や特徴地図を用いた，行動認識モデルの設計指針を検討していきたい．

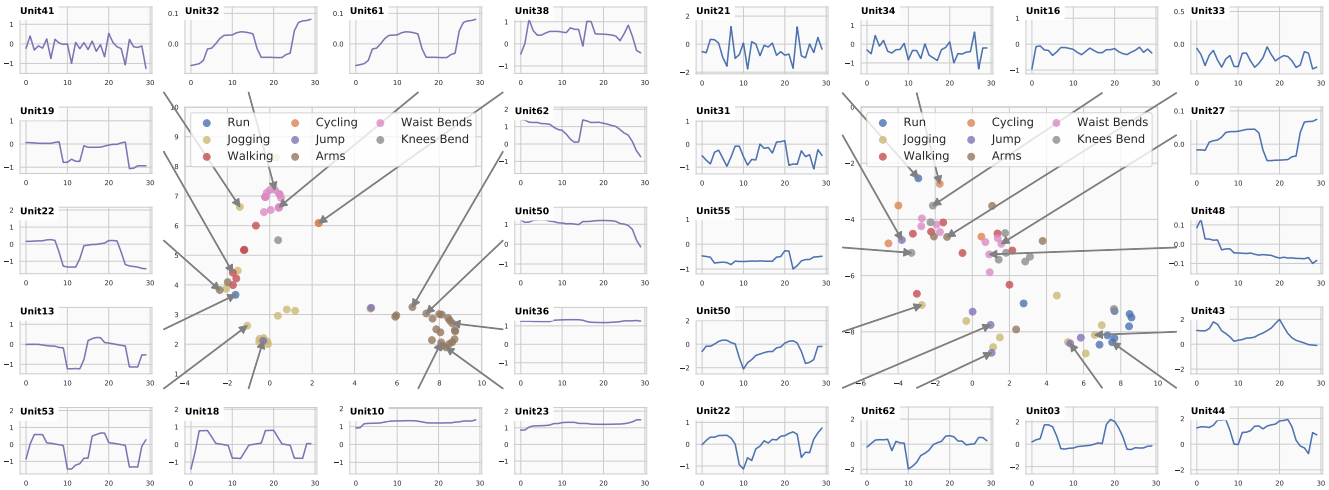
謝辞

本研究の一部は JST CREST JPMJCR15E2，JSPS 科研費 JP16H06539，JP 17H04679，JST ACT-X JPMJAX200T の助成を受けて行われたものである．

参考文献

[1] O. Banos, M. Toth, M. Damas, H. Pomares, and I. Rojas, “Dealing with the effects of sensor displacement in wearable activity recognition,” Sensors, vol.14, no.6, pp.9995–10023, 2014.
[2] B. Barshan and M.C. Yüksek, “Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units,” The Computer Journal, vol.57, no.11, pp.1649–1667, 2014.
[3] S. Carter, Z. Armstrong, L. Schubert, I. Johnson, and C. Olah, “Activation atlas,” Distill, 2019.
[4] D.P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” Proceedings of the 3rd International Conference on Learning Representations, 2015.
[5] Q.V. Le, M. Ranzato, R. Monga, M. Devin, G. Corrado, K. Chen, J. Dean, and A.Y. Ng, “Building high-level features using large scale unsupervised learning,” Proceedings of the 29th International Conference on Machine Learning, 2012.
[6] X. Li, X. Si, L. Nie, J. Li, R. Ding, D. Zhan, and D. Chu, “Understanding and improving deep neural network for activity recognition,” CoRR, vol.abs/1805.07020, 2018.
[7] L.v.d. Maaten and G. Hinton, “Visualizing data using t-SNE,” Journal of Machine Learning Research, vol.9, no.Nov, pp.2579–2605, 2008.
[8] A. Mahendran and A. Vedaldi, “Understanding deep image representations by inverting them,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.5188–5196, 2015.
[9] G. Montavon, W. Samek, and K.R. Müller, “Methods for interpreting and understanding deep neural networks,” Digital Signal Processing, vol.73, pp.1–15, 2018.
[10] S. Münzner, P. Schmidt, A. Reiss, M. Hanselmann, R. Stiefelhagen, and R. Dürichen, “CNN-based sensor fusion techniques for multimodal human activity recognition,” Proceedings of the 2017 ACM International Symposium on Wearable Computers, pp.158–165, 2017.

© 2020 Information Processing Society of Japan



(1) Conv1 (1st layer)　　(2) Conv3 (3rd layer)

図 8 行動認識ニューラルネットワークに対する特徴地図．MHEALTH を学習したモデルの 1 層目 (左) と 3 層目 (右)．中央が特徴地図であり，図中の 1 つ 1 つのデータポイントは各ユニットに対応している．また，各データポイントは MSE に基づく距離指標 ($d(u,i)$) に基づいて選択された最も類似したサンプルの行動クラスによって色が割り当てられている．特徴地図の周辺の波形は，矢印で示されたユニットに対して AM で生成した波形である．横軸はサンプルのインデックス，縦軸は振幅 [G] である．

[11] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski, “Plug & play generative networks: Conditional iterative generation of images in latent space,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.4467–4477, 2017.

[12] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, “Synthesizing the preferred inputs for neurons in neural networks via deep generator networks,” Proceedings of the Advances in Neural Information Processing Systems, pp.3387–3395, 2016.

[13] A. Nguyen, J. Yosinski, and J. Clune, “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.427–436, 2015.

[14] A.M. Nguyen, J. Yosinski, and J. Clune, “Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks,” CoRR, vol.abs/1602.03616, 2016.

[15] A. Saeed, T. Ozcelebi, and J. Lukkien, “Multi-task self-supervised learning for human activity detection,” Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol.3, no.2, p.61, 2019.

[16] S. Saks, Theory of the integral. 2. ed. English translation by *L. C. Young*. With two additional notes by S. Banach., vol.7, 1937.

[17] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.R. Müller, “Evaluating the visualization of what a deep neural network has learned,” IEEE Transactions on Neural Networks and Learning Systems, vol.28, no.11, pp.2660–2673, 2016.

[18] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” Proceedings of the IEEE International Conference on Computer Vision, pp.618–626, 2017.

[19] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” Proceedings of the 2nd International Conference on Learning Representations, 2014.

[20] E. Soleimani and E. Nazerfard, “Cross-subject transfer learning in human activity recognition systems using generative adversarial networks,” CoRR, vol.abs/1903.12489, 2019.

[21] J. Wang, Y. Chen, Y. Gu, Y. Xiao, and H. Pan, “Sensorygans: an effective generative adversarial framework for sensor-based human activity recognition,” Proceedings of the 2018 International Joint Conference on Neural Networks, pp.1–8IEEE, 2018.

[22] N. Yoshimura, T. Maekawa, and T. Hara, “Preliminary investigation of visualizing human activity recognition neural network,” 2019 Twelfth International Conference on Mobile Computing and Ubiquitous Network, pp.1–2IEEE, 2019.

[23] J. Yosinski, J. Clune, A.M. Nguyen, T.J. Fuchs, and H. Lipson, “Understanding neural networks through deep visualization,” CoRR, vol.abs/1506.06579, 2015.

[24] M.D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” Proceedings of the European Conference on Computer Vision, pp.818–833, 2014.

[25] M. Zeng, H. Gao, T. Yu, O.J. Mengshoel, H. Langseth, I. Lane, and X. Liu, “Understanding and improving recurrent networks for human activity recognition by continuous attention,” Proceedings of the 2018 ACM International Symposium on Wearable Computers, pp.56–63, ACM, 2018.

© 2020 Information Processing Society of Japan