ビット重要度の違いに着目した符号アテンションによる説明能力向上

田代 悠馬[1,a)]　粟野 皓光[2,b)]

概要：現代の深層学習アルゴリズムは非常に複雑な人工神経網から成り立っており，人間による推論過程の追跡は困難を極める．深層学習の社会実装が進む一方で，推論誤りがもたらす人的・経済的損失が問題視されており，深層学習アルゴリズムの判断根拠を説明する手法が求められている．例えば，自動運転タスクでは，ステアリング操舵角予測に寄与する領域をアテンション機構によって視覚化する手法が提案されているが，説明能力は依然として低かった．本研究では，アクティベーションを構成するビット毎の重要度の異なり（つまり LSB が重みが小さく，MSB が最も重みが大きい）に着目し，符号ビットに限定してアテンションを付加する手法を提案し，説明能力の更なる向上を図る．提案手法を用いて，ネットワークが出力する着目領域と予測誤差の関係を検証し，提案手法の有効性を示す．

Improving Explanation Ability by Sign Attention Exploiting Difference in Bit Significance.

Tashiro Yuma[1,a)]　Awano Hiromitsu[2,b)]

Abstract: Modern deep learning algorithms consist of extremely complex artificial neural networks, making it extremely difficult for humans to track the inference process. While the social implementation of deep learning is progressing, the human and economic losses caused by inference errors are becoming more and more problematic, and there is a need for methods to explain the basis for the decisions of deep learning algorithms. For example, in the task of self-driving, a method has been proposed to visualize the regions that contribute to steering angle prediction using an attention mechanism, but its explanatory power was still low. In this study, we focus on the difference in the importance of each bit of activation (i.e., LSBs have the lowest weight and MSBs have the highest weight), and propose a method to add attention only to the sign bits to further improve the explanatory power. The effectiveness of the proposed method is demonstrated by examining the relationship between the prediction error and the region of interest output by the network.

1. 序論

ディープニューラルネットワーク (DNN) は，コンピュータの性能やアルゴリズムの向上により，近年大きな進歩を遂げており，音声や自然言語，画像処理などの分野で実用化が進んでいる．2012 年に開催された画像認識コンテストの ILSVRC において，「AlexNet」と呼ばれる 8 層構造のニューラルネットワークが驚異的な精度を示したことで，DNN は注目を集めている．それに伴って DNN は安全性が重視される領域での応用も始まっている．その中で特に注目されているのが自動運転の分野で，正面のカメラの画像からステアリングやスロットルの操作を予測するために DNN が用いられている [1].

深層学習は，ネットワークの複雑さと引き換えに，さまざまな問題に対して非常に高い性能を実現している，最新の DNN は膨大な量のパラメータを機械が計算して最適化することで構成されている．例えば，VGG-19 という有名なネットワークでは，約 1 億 4400 万個のパラメータが組

1　大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University
2　京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University
a)　y-tashir@ist.osaka-u.ac.jp
b)　awano@i.kyoto-u.ac.jp

み込まれており，数百万枚の画像を使って最適化されている [2]．このような複雑な DNN の意思決定する過程は，人間が追跡することは困難である．この問題は「ブラックボックス AI」と呼ばれ，DNN のミッションクリティカルなアプリケーションへの適用を妨げる大きな要因となっている．そこで，なぜその判断に至ったのかを説明できる AI である「説明可能 AI (XAI)」の研究が盛んに行われている [3], [4], [5]．XAI の目的は，推論プロセスを明らかにすることでモデルを解釈可能にし，行動の根拠を人間が理解できるように提供することである．解釈の可能性を高めることで，誤った判断の原因を調べたり，モデルの判断をより確かなものにすることができる．

DNN の解釈性を向上させるためのアプローチの 1 つに，予測を行う上で「重要な」入力画像の領域を視覚化することが挙げられる．例えば，Ribeiro らは，分類器の予測値を解釈可能かつ忠実に説明する，LIME と呼ばれる手法を提案している [6]．LIME では，ブラックボックスで複雑なモデルを，SVM などの線形カーネルを用いた単純で解釈可能な分類器を用いて近似することで，モデルの解釈可能性を高めている．各入力画像に対して，ランダムな摂動を加えてサンプルを生成し，それをもとに解釈可能なローカル分類器を学習する．したがって，LIME は膨大な計算量を必要とし，実時間アプリケーションには適していない．その後，2017 年に Selvaraju らは “Grad-CAM” と名付けられた手法を提案した．CNN の最後の畳み込み層に流れ込む勾配の情報を利用することで，時間のかかるローカル分類器の反復学習を行うことなく，最終的なクラスラベル予測に対する各入力ピクセルの重要性を評価することを実現した [7]． Grad-CAM は計算量が少ないことが特徴であるが，デコンボリューションに基づくアプローチでは，説明された注目度マップの空間的な解像度が低いという問題がある．また，ネットワークが予測だけでなく，予測を行う際にネットワークを行う際にネットワークが焦点を当てた部分も出力するように構造を変更することで，モデルの判断根拠を説明できるようにする方法もある．この仕組みは「アテンション」と呼ばれ，特に自然言語処理のニューラルネットワークで広く利用されている [8]． アテンションに基づいたアプローチでは，ネットワークの出力に大きな影響を与える画像の領域を直接抽出することができる [9], [10]．

アテンション機構は，画像認識や自然言語処理で広く適用されているが，自動運転などのより実用的なアプリケーションに適用されることはほとんどないことが実情である．2018 年，Kim らは，自動運転システムへのアテンション機構の適用に初めて成功したことを報告した [11]．彼らのモデルは，入力された生の画像からステアリングの操作をエンド・ツー・エンドで予測する．その上で，モデルがどこで何を見ているかを視覚化するアテンションヒートマップを生成している．彼らのアプローチの欠点は，注目マップには重要でない領域が含まれている可能性があるため，ネットワーク予測に影響を与えない領域を選別するための後処理が必要になることである．そこで著者らは，アテンションマップの注目度が類似しているピクセルをグループ化してクラスターを作り，視覚的な顕著性をマスキングすることで，各クラスターが予測精度に因果関係を持つかどうかを検証することを提案した．しかし，このような反復的な処理には追加の計算時間が必要であり，リアルタイムのアプリケーションには適していない．

本論文では，エンド・ツー・エンドでステアリング角度を予測し，予測されたステアリング角度の画素ごとの重要度を説明することができるニューラル・ネットワークを提案する．このネットワークの特徴は，特徴量マップの「符号」ビットにのみ注目することである．最上位ビット（MSB）側を表すビットは，他のビットよりも情報量が多いように，浮動小数点値を構成するビットの重要度は各ビットで異なる．このような重要なビットの変化は，予測される操舵角に大きな影響を与える．例えば，符号ビットの変化は，ステアリングの方向を反転させることになる．このような重要なビットに注目することで，説明可能性の向上を図る．従来のネットワーク [11] では，画像の特徴量を構成する全てのビットが同等に扱われていたが，本論文で提案するネットワークでは，まず特徴量の符号を抽出し，その符号にのみアテンションマップを付加する．また，標準的な逆伝播アルゴリズムでモデルを学習できるようにするため，2 値化ニューラルネットワーク（BNN）の学習に用いられる STE（Straight Through Estimator）を導入し，逆伝播の段階で，決定論的な符号関数をハードシグモイド関数に置き換えて，ネットワークに勾配を伝播させるようにした．

評価実験では，Udacity の自動運転データセット [12] を用いて提案手法を検証する．まず，ステアリング角度予測の精度を計測し，符号アテンションの適用による予測精度の低下が無いことを示す．そして，説明されたアテンションマップの注目度が予測精度への貢献度と相関しているかどうかを，削除メトリック [13] を用いて調べた．アテンションの注目度が大きい順にピクセルがマスクされ，予測誤差の増加を測定する．予測誤差の増加が大きいほど，重要な領域に正しく注目度の重みを割り当てることができることが分かる．提案手法では，既存手法と比較して予測誤差の増加が大きいことから，説明されたアテンションマップが予測に影響する部分に正しく注目できていることが分かった．

2. 関連研究

2.1 自動運転車に向けたエンド・ツー・エンドの学習

自動車の自動制御はここ数年で飛躍的な進歩を遂げており，特にニューラルネットワークを用いたエンドツーエン

ドの学習に基づく自動運転車への関心が高まっている．このようなアプローチでは，まず自動車の前面に搭載されたカメラで撮影された映像や，人間が操作したステアリングやスロットルなどの制御情報を収集する．そして，収集した情報を基にモデルを学習させる．最後に，学習したモデルを実際の環境で予測させる．

ALVINN(Autonomous Land Vehicle In a Neural Network) は，世界で初めてニューラルネットワークを用いて入力画像から車両を制御することに成功した [14] これは，前方のカメラ映像と運転手のハンドルやペダル操作との対応関係をニューラルネットワークに学習させることで，人間の運転操作を模倣するものである．ALVINN の成功によって自動運転の研究が盛んにおこなわれるようになり，ディープラーニング技術の発展も相まって，人間の運転に匹敵する正確な自動運転が実現しつつある．

最近では，DNN を用いた手法の 1 つとして，Bojarski らによる畳み込みニューラルネットワークを用いた手法がある [1]．このシステムは，車の正面，左右側面を撮影した 3 つのカメラを入力とし，ステアリング操作を予測することができる．また，このシステムを公道でテストしたところ，人手を介さずに 10 マイルの距離を走ることに成功した．このように高い性能を示した一方で，ニューラルネットワークは複雑化しているため，ステアリング角度を推定する過程を人間が理解できるように説明することは非常に困難である．

2.2 ニューラルネットワークの説明可能性

ニューラルネットワークのブラックボックス問題がこれらを社会実装していく上で大きな障壁となっていることから, ニューラルネットワークを解釈可能なものにするために様々な手法が提案されている．その先駆けとして，ニューラルネットワークに説明可能性を持たせることができる，LIME (Local Interpretable Model-agnostic Explanation) と呼ばれる手法が提案された．この手法は，DNN のモデルは大局的には非常に複雑であっても，特定の入力の近傍では容易に近似できるという考えを基にして予測結果の解釈を可能にしている．f と $\boldsymbol{X}$ を DNN ベースの分類器と入力サンプルとする．LIME は $\boldsymbol{X}$ にランダムなノイズを加えることで $\boldsymbol{X}$ の近傍のサンプルを生成し，それに対応するニューラルネットワークの出力 $f(\boldsymbol{X}+\epsilon)$ を得る．次に，$g(\boldsymbol{X})$ というスパースな線形モデルを用いて，f を $\boldsymbol{X}$ の近傍で局所的に近似する．g は線形関数，つまり $g=\boldsymbol{w}\cdot\boldsymbol{X}$ なので，重みベクトル $\boldsymbol{w}$ を用いることで，識別に大きな影響を与える $\boldsymbol{x}$ の特徴を特定することができる．

LIME は汎用性が高く，様々なモデルに適用できるが，入力の度に分類器を再学習する必要があり，リアルタイム性が求められるアプリケーションには適していない．そこで，より計算効率の良い手法である Grad-CAM が提案されている [7]．Grad-CAM は，勾配の大きいピクセルは予測値に大きな影響を与えるという考えに基づいて，予測値に対して勾配の重み付けを行うことで，重要なピクセルを可視化する．また，Grad-CAM で得られたカラーマップを，Guided Backpropagation や Deconvolution といった既存の説明手法と組み合わせて可視化する Guided Grad-CAM も提案されている．Grad-CAM や Guided Grad-CAM は，ニューラルネットワークの学習に不可欠な勾配を利用しているため，他の XAI 手法と比較して計算量が少なく，リソースに制約のある機器への実装に適している．しかし，これらはデコンボリューション層に依存しているため，説明される注目度マップの空間解像度が低いという問題があった．

アテンションは，入力データ中の着目箇所をモデルが示す仕組みを導入することで，判断の根拠を理解する手法である．当初は，自然言語処理の分野で，再帰的ニューラルネットワークを用いた機械翻訳モデルに適用することで，翻訳文中の各単語が原文中のどこにフォーカスされているかを学習する仕組みとして用いられていた．近年，画像認識の分野では，RNN や LSTM (Long Short Term Memory) を用いた手法が，入力画像中のどの画素に注目したかという視覚的な説明を得るためのアテンションマップを学習する仕組みとして用いられている．アテンションマップは，CNN によって入力画像から得られた特徴量マップから生成され，特徴量マップと組み合わせることで，注目した領域が特徴量マップに大きく寄与する重み付きマップを得ることができる．

こうした XAI 技術の自動運転への応用も広がっている．Kim らは，CNN と LSTM で構成された車両制御モデルにアテンション機構を導入し，ステアリング動作に注目している画像領域を示すことで，モデルの解釈性を向上させた [11]．このモデルでは，畳み込みで抽出した特徴量と LSTM の過去の出力から着目点を算出し，ヒートマップとして得ている．この手法の問題点は，モデルから直接得られたマップが示す注目領域は，出力結果に大きな影響を与えない領域を示すことが多く，予測に寄与する領域を正確に示すことができないことである．そのため，この方法では，モデルから出力された注目マップが示す注目領域のうち，出力結果に影響を与えないと考えられる領域を削除し，実際に予測に大きく寄与する可能性が高い領域を示すように修正するという後処理を行っている．

3. 提案手法

提案するネットワークの構造を図 1 に示す．まず，入力画像から最初の 5 つの畳み込み層によって特徴量マップが抽出される．そして，抽出された特徴量マップは，アテンション機構によって推定されたアテンションの重みによって重みづけされる．従来のアテンション機構では，全ての

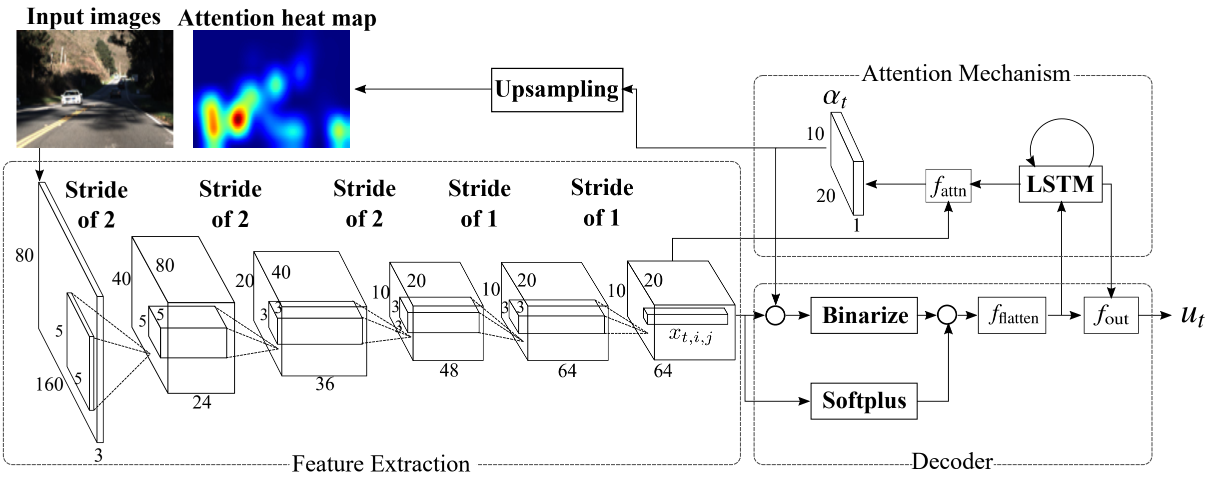

図 1　提案ネットワークの構造

ビットが等しく重みづけされていたが，提案ネットワークでは，符号のみにアテンションが適用される．アテンション機構の出力は平坦になるように整形され，ステアリング角度を予測する LSTM に入力される．以下では、各処理の流れについて説明する。

3.1　前処理

このモデルでは，車両前面のカメラで撮影された画像が与えられると，連続的なステアリング角度値をエンド・ツー・エンドで予測する．まず，計測ノイズを低減して学習の安定性を高めるために，測定されたステアリング角度（と車両速度）に指数平滑化法を適用する．

$$\begin{pmatrix}\hat{\theta}_t\\ \hat{v}_t\end{pmatrix} = \alpha_s \begin{pmatrix}\theta_t\\ v_t\end{pmatrix} + (1-\alpha_s)\begin{pmatrix}\hat{\theta}_{t-1}\\ \hat{v}_{t-1}\end{pmatrix}, \qquad (1)$$

ここで，$\hat{\theta}_t$ と $\hat{u}_t$ は，それぞれ平滑化されたステアリング角度と車両速度の時系列データである．α_s は平滑化の度合いを調整するパラメータであり，0 に近いほど平滑化の効果が大きい．ステアリング角度の大きさはホイールベースなどの車両の構造に依存するため，この手法ではステアリング角度の代わりに逆回転半径 u_t を予測する [1], [11]．ステアリング角度と逆回転半径 u_t の関係は以下の式で近似される．

$$u_t = \frac{\hat{\theta}_t}{d_w K_s \left(1 + K_{\mathrm{slip}} \hat{v}_t{}^2\right)}, \qquad (2)$$

ここで，d_w は前後のタイヤ間の距離，K_s はステアリングの回転と車輪の回転の比，K_{slip} は車輪と路面の間の相対運動を表す．入力データとして使用した画像は，計算コストを削減するために $80 \times 160 \times 3$ のサイズにリサイズする．また，RGB から HSV への色空間の変換を行った．

3.2　特徴抽出エンコーダ

特徴量マップの抽出は，従来手法 [11] に基づいて，5 つの畳み込み層で構成されるネットワークを用いて行う．提案するネットワークの特徴抽出エンコーダの構造を表 1 に示す．1 列目は層の名前，2 列目は出力サイズ，3 列目はフィルタの窓サイズ，チャンネル数，ストライドの大きさを示している．時刻 t の入力画像に一連の畳み込み演算を行うと，高さ H，幅 W，チャンネル C のテンソル X_t が得られる．後の説明のために，X_t の (i,j) 要素を $x_{t,i,j} = (x_{t,i,j,1}, x_{t,i,j,2}, ..., x_{t,i,j,C})$ とする．

表 1　特徴抽出エンコーダの構造

layer name	output size	filter size, # of channel, stride
conv1	40×80	5×5, 24, stride 2
conv2	20×40	3×3, 36, stride 2
conv3	10×20	3×3, 48
conv4	10×20	3×3, 64
conv5	10×20	3×3, 64

前で説明したように，提案手法では説明性を向上させるために，符号にのみアテンションマップを適用する．この処理を行うために，X_t は 2 つの要素に分割され，異なる計算が行われる．1 つはアテンション機構を経て符号を抽出し，もう 1 つは絶対値を抽出する．$\boldsymbol{\alpha}_t = \{\alpha_{t,1,1}, \alpha_{t,1,2}, \cdots, \alpha_{t,W,H}\}$ を時刻 t におけるアテンションの重みとすると，符号の抽出は以下のようにして行われる．

$$x^{\mathrm{sig}}_{t,i,j,c} = \begin{cases} +1 & \text{if } x_{t,i,j,c} \cdot \alpha_{t,i,j} \geq 0, \\ -1 & \text{otherwise.} \end{cases} \qquad (3)$$

ここで，アテンションの重みは $\Sigma_{i,j}\alpha_{t,i,j} = 1$ を満たす．また，X_t の絶対値は “Softplus” 関数を用いて以下のように抽出する．

$$x^{\mathrm{amp}}_{t,i,j,c} = \log(1 + e^{x_{t,i,j,c}}). \qquad (4)$$

最後に，これらの抽出結果を掛け合わせて，符号付き特徴量を再構成する．

$$\hat{\boldsymbol{x}}_{t,i,j} = \left(x^{\mathrm{sig}}_{t,i,j,1} \cdot x^{\mathrm{amp}}_{t,i,j,1}, \cdots, x^{\mathrm{sig}}_{t,i,j,C} \cdot x^{\mathrm{amp}}_{t,i,j,C}\right). \qquad (5)$$

$\hat{x}_{t,i,j}$ を要素とする時刻 t でのテンソル $\hat{X}t$ は，LSTM に入力するために平坦化される（$y_t = f_{\text{flatten}}(\hat{X}_t)$）．ここで，$y_t$ は $W \times H \times C$ の要素を持つ．

3.3 逆回転半径とアテンションの予測

アテンション機構を経た後，次に LSTM に入力する．この LSTM は，1 つ前の隠れ状態である h_{t-1} から逆回転半径 u_t とアテンション α_t を予測する．u_t の予測は以下のように表すことができる．

$$\begin{aligned}\hat{u}_t &= f_{\text{out}}(\boldsymbol{h}_{t-1}, \hat{X}_t)\\ &= \Big(Sigm(W_\alpha \cdot \boldsymbol{h}_{t-1} + \boldsymbol{b}_\alpha) \circ \hat{X}_t\Big) \cdot W_\beta + \boldsymbol{b}_\beta, \qquad (6)\end{aligned}$$

ここで，W_α と W_β は学習可能な重み行列で，b_α と b_β は学習可能なバイアスである．Sigm は Sigmoid 関数，$\circ$ は要素ごとの乗算を表す．

アテンション α を生成するために，まず，追加の隠れ層を計算する．

$$\begin{aligned}\boldsymbol{e}_t &= f_{\text{attn}}(X, \boldsymbol{h}_{t-1})\\ &= \tanh\left(W_\gamma \cdot X_t + W_\delta \cdot \boldsymbol{h}_{t-1}\right) + \boldsymbol{b}_\gamma, \qquad (7)\end{aligned}$$

ここで，W_γ，W_δ は学習可能な重み行列，b_γ は学習可能なバイアスである．そして，$\Sigma_{i,j}\alpha_{t,i,j} = 1$ となるように Softmax 関数を適用する．

$$\begin{aligned}\boldsymbol{e}_t &= f_{\text{attn}}(X, \boldsymbol{h}_{t-1})\\ &= \tanh\left(W_\gamma \cdot X_t + W_\delta \cdot \boldsymbol{h}_{t-1}\right) + \boldsymbol{b}_\gamma, \qquad (8)\end{aligned}$$

最後に，α_t は 2 次元のアテンション行列に再形成される．

時刻 $t = 0$ における LSTM のセル状態と隠れ状態は，隠れ層 $f_{init,c}$ と $f_{init,h}$ を用いて以下の式で初期化される．

$$c_0 = f_{\text{init},c}\left(\frac{1}{L}\sum_{i=1}^{L} x_{0,i}\right), h_0 = f_{\text{init},h}\left(\frac{1}{L}\sum_{i=1}^{L} x_{0,i}\right) \qquad (9)$$

ここで，c_0 と h_0 はそれぞれ，$t = 0$ における LSTM のセル状態と隠れ状態である．

3.4 STE

提案モデルでは，微分不可能な符号関数を利用しており，その微分はほとんどの場所で 0 になるため，モデルを学習する際に一般的な勾配降下法アルゴリズムを直接適用できないという問題が生じる．この問題を解決するために，BNN の学習のために提案された STE（Straight Through Estator）と呼ばれる技術を利用する [15]．具体的には，STE は逆伝播の際に，符号関数をリップした恒等関数 (hard tanh) であるかのように扱う．したがって，勾配は以下のように与えられる．

$$\frac{\partial}{\partial x} Sign(x) = \begin{cases} 1 & \text{if } x \leq 1, \\ 0 & \text{otherwise.} \end{cases} \qquad (10)$$

表 2 予測精度の比較

Methods	MAE [°]
CNN+LSTM w/ Attention [11]	4.94
CNN+LSM w/ Sign only Attention (Proposed)	4.83
CNN+LSTM w/o Attention	4.77

4. 実験

4.1 実験環境

2016 年に開始されたオープンソースの自動運転プロジェクトである Udacity データセット [12] を用いて，提案手法の学習と評価を行った．このデータセットには，フロントガラスの後部に取り付けられたフロントビューカメラで撮影された動画像と，高速道路や市街地を 3.6 時間かけて日中に走行した際に 20 フレーム/秒で記録された，車両速度やステアリング角度などのタイムスタンプ付きのセンサの測定値が含まれている．このデータセットには 263,075 フレームが含まれており，そのうち 257,796 フレームはモデルの学習に，残りの 5,279 フレームはステアリング角度の予測精度の評価に使用する．

モデルの学習には NVIDIA Geforce GTX2080Ti GPU を使用し，コードは PyTorch フレームワークを用いて記述した．ネットワークの重みの初期化には Xavier 初期化を用い，学習には学習率 10^{-4} の Adam オプティマイザを使用した．また，ドロップアウト確率は 0.5 に，LSTM の隠れ状態のサイズは 1024 に設定した．

4.2 性能分析

精度: まず，提案手法のステアリング角度の予測精度を従来手法と比較する．また，アテンション機構が予測性能に与える影響を調べるために，アテンション機構が無い以外は提案手法と同じ構造をした，LSTM と組み合わせた CNN も実装した．各手法で得られた平均絶対誤差を表 2 に示す．表 2 から，提案手法はテストデータセットにおいて予測精度は MAE 換算で 4.83 であり，既存手法と同等の予測精度を達成できることが分かる．

可視化: フロントカメラの画像と，それに対応するアテンションマップの可視化結果を図 2 に示す．ここでは，入力画像とそれに対応するアテンションヒートマップを左から右に並べている．アテンションは畳み込みニューラルネットワークで構成された特徴抽出機構の出力に付加されるため，得られるアテンションは入力画像の 1/8 の解像度となる．入力画像の解像度に合わせるために，アップサンプリングを行った後，ガウシアンフィルタを適用して画像のエッジを柔らかくしている．この図から，提案手法では車を運転する際に重要と考えられる，道路の中央線や対向車に注意が向けられていることがわかる．

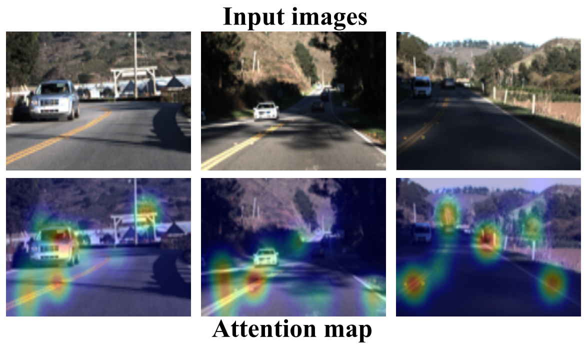

図 2　入力画像と対応するアテンションマップの例

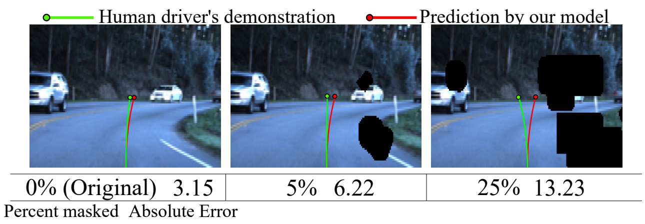

図 3　マスキング度合と予測誤差の増加の関係例

説明能力: 最後に，出力されたアテンションが入力画素のステアリング角度の予測への貢献度をよく反映しているかどうかを削除メトリック [13] によって調べた．この指標は，アテンションの値が高い順に対応する画素をマスクしていき，予測誤差の増加を測定する．予測誤差の増加が大きいほど，説明能力が高いことを示す．ここで，マスキングされた入力画像と予測精度の関係の例を図 3 に示す．この例では，元の入力画像の予測誤差が 3.15° であったのに対して，画像を 25%マスキングした後は，誤差が 13.2° に増加していることが分かる．この手順をテストデータセット 5,279 枚に対して同様に行い，MAE を算出した．マスクされた画素の割合 (パーセント) と MAE の関係を図 4 に示す．黒線と赤線はそれぞれ従来手法 [11] の MAE と提案手法の MAE に対応している．提案手法では誤差が急激に増加していることが分かり，提案手法の説明能力が向上していることを確認できた．さらに，アテンションの信頼性を定量的に比較するために，曲線下面積 (AUC) を計算したところ，提案手法では 8.98，従来手法 [11] では 6.75 となり，提案手法は従来手法と比較して 33%高い AUC を示している．AUC が高いほど，アテンションの重みが入力画像の重要領域とよく相関していることを示すので，提案手法の説明が従来手法 [11] よりも優れていることを確認できた．

5. 結論

本論文では，説明可能性を改善するために，新しいアテンション機構を提案した．MSB は他のビットよりも多くの情報を伝えているというビットの重要度の違いを利用して，「符号」ビットにのみアテンションを付加することを提案した．STE を利用することで，提案も出るは標準的な逆伝播アルゴリズムを用いてエンドツーエンドで学習することができ，符号のみのアテンション機構を導入しても予測精度の低下は無いことを確認した．また，Udacity データセットを用いた評価実験では，入力画像の 1/4 がマスクされた場合，提案モデルの誤差は最大で 13.23° になることがわかった．さらに，アテンションが予測を行う上で入力の重要性をよく反映しているかどうかを調べたところ，AUC が 33 %高かったことから，提案手法は重要な領域に正しく注意を向けることができることが分かった．

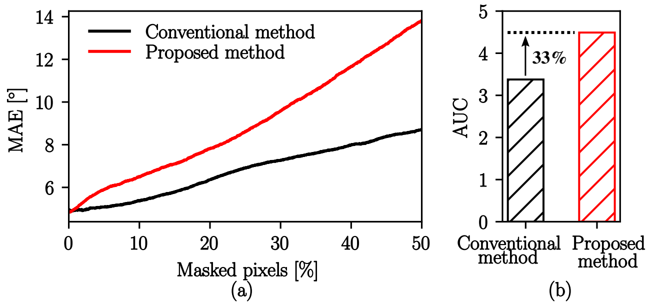

図 4　50%までマスクしたときの予測誤差の増加

謝辞

本研究は，JST，さきがけ，JP-MJPR18M1，JSPS 科研費 21H03409 の支援を受けたものである．

参考文献

[1] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, “End to End Learning for Self-Driving Cars,” *CoRR*, vol. abs/1604.07316, 2016.

[2] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[3] A. Adadi and M. Berrada, “Peeking inside the black-box: a survey on explainable artificial intelligence (XAI),” *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[4] E. Tjoa and C. Guan, “A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI,” *IEEE Trans. Neural Netw. Learn. Syst.*, 2020.

[5] O. Biran and C. Cotton, “Explanation and Justification in Machine Learning: A Survey,” in *Int. Joint Conf. on Artificial Intell. Workshop on Explainable AI (XAI)*, vol. 8, pp. 8–13, 2017.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You?: Explaining the Predictions of Any Classifier,” in *Int. Conf. on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

[7] 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,” in *Int. Conf. on Comput. Vision*, pp. 618–626, 2017.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is All you Need,” in *Neural Information Processing Syst.*, vol. 30, pp. 6000–6010, 2017.

[9] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *International conference on machine learning*, pp. 2048–2057, PMLR, 2015.
[10] S. Sharma, R. Kiros, and R. Salakhutdinov, “Action Recognition using Visual Attention,” *CoRR*, vol. abs/1511.04119, 2015.
[11] J. Kim and J. Canny, “Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention,” in *Int. Conf. on Comput. Vision*, pp. 2942–2950, 2017.
[12] Udacity, “Udacity Self-Driving Car Driving Data.”
[13] V. Petsiuk, A. Das, and K. Saenko, “RISE: Randomized Input Sampling for Explanation of Black-box Models,” in *British Machine Vision Conf.*, 2018.
[14] D. Pomerleau, “ALVINN: An Autonomous Land Vehicle In a Neural Network,” in *Neural Information Processing Syst.*, pp. 305–313, December 1989.
[15] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, “Binarized Neural Networks,” in *Int. Conf. on Neural Information Processing Syst.*, pp. 4114–4122, 2016.