CH-008

Saliency Guided Training を使用した Lambda Attention Branch Networks による視覚的説明生成
Visual Explanation Generation Using Lambda Attention Branch Networks with Saliency Guided Training

小松 拓実[1)] 飯田 紡[1)] 兼田 寛大[1)] 平川 翼[2)]
Takumi Komatsu Tsumugi Iida Kanta Kaneda Tsubasa Hirakawa
山下 隆義[2)] 藤吉 弘亘[2)] 杉浦 孔明[1)]
Takayoshi Yamashita Hironobu Fujiyoshi Komei Sugiura

1 はじめに

Deep Neural Networks (DNNs) は様々なタスクに幅広く用いられている．一方で，その予測に対する判断根拠を解釈することは困難であり，予測結果を十分に信頼できないという問題が指摘されている [19]．特に，医療や自動運転といった分野は，予測結果に対して信頼できる説明が重要である [16]．また，未解明な自然現象を対象とする場合，視覚的説明により重要な領域を可視化することで，理論的な洞察を与えることができる．

CNN に基づくモデルに対する説明生成は盛んに研究されてきた [6,7,18]．一方，視覚的説明の生成タスクは特定のモデルや構造に特化した手法が多く，全てのモデルで汎用的に明瞭な説明を生成できる手法が十分に確立されていない．特に，Lambda Networks [2] に基づく transformer の説明生成手法は十分に確立されておらず，解釈が困難な説明や重要でない領域を注目した説明を生成することがある．

このような背景から，本論文は Lambda Networks に基づく transformer の判断根拠の視覚的説明を生成するタスクを扱う．図 1 に DeFN magnetogram [13] の画像例を示す．本タスクは左図のような入力画像に対して，モデルが分類結果を出力する過程で，右図のような判断根拠の視覚的説明を生成することが望ましい．ここで、右図は出力される視覚的説明を入力画像に重畳した画像である．

Lambda Networks に基づく transformer の説明生成として Lambda Network の注意機構を用いた方法や，LABN [8] が挙げられる．Lambda Network の注意機構を用いた方法は明瞭でない説明が生成されることがある．また，LABN は画像領域全体の注目度が高いという問題点がある．そこで，本研究は LABN を拡張し，解釈性を高めた視覚的説明を生成する LABN-S を提案する．

提案手法における新規性を以下に示す．

- モデルの出力に対する重要度の低い領域の影響を軽減し、視覚的説明の解釈性を高めるために、Saliency Guided Training [1] で提案された損失を導入する．
- マスクによる分布の変化を軽減するために，Insertion-Deletion score および PID score におけるマスク画像としてバイアス画像を導入する．

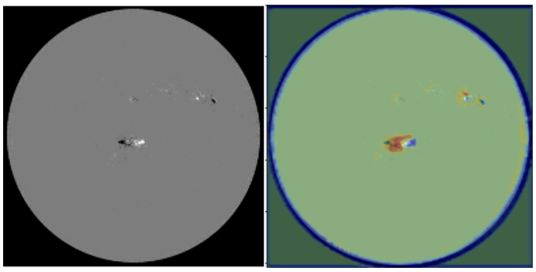

図 1 本タスクにおける入力と視覚的説明の例

2 関連研究

モデルの説明生成を行う手法は多く存在する．[3,4,6,7,15,18] サーベイ論文である [5] は，DNN の説明生成に関して包括的に調査し，説明の生成ごとに各手法の分類・比較を行っている．また，本論文で扱う transformer に基づくモデルにおけるサーベイ論文として [10] があげられる．[10] は，画像認識タスクにおいて transformer に基づく主要なモデルの特徴をまとめている．

モデルの説明生成は、その生成方法によって Back Propagation (BP)， Perturbation (PER) とその他（OT）に分けられる．BP は Back Propagartion 時の勾配に着目して説明を生成する．BP の手法として，CAM [21], Grad-CAM [18]，LRP [4] 等がある．また，PER は入力に摂動を加えて，モデルの出力の変化から説明を生成する．PER の手法として，LIME [19]， RISE[[16], Shapely Sampling [11] 等がある．OT は勾配や摂動以外から説明を生成する手法で，Attention Branch Network [7], IA-RED2 [15] 等がある．

Attention Branch Network [7] は、ブランチ構造を利用した視覚的説明の生成を行ったモデルである。Lambda Attention Branch Network [8] は、ABN を Lambda Networks に導入し、transformer に基づくモデルの視覚的説明の生成に取り組んだモデルである。また、Multimodal Attention Branch Network [12] は，言語と画像のマルチモーダルな情報を用いて生成した指示文の説明に Attention Branch を利用して取り組んだモデルである．ABEN [14] は，Multi-ABN を拡張し，ABN をサブワード単位で注意を向けることによって，指示文に対する詳細な説明を生成したモデルである．このように，ブランチ構造を用いた視覚的説明の生成は，様々なタスクで使用されており，汎用性が高い．

1） 慶應義塾大学 Keio University
2） 中部大学 Chubu University

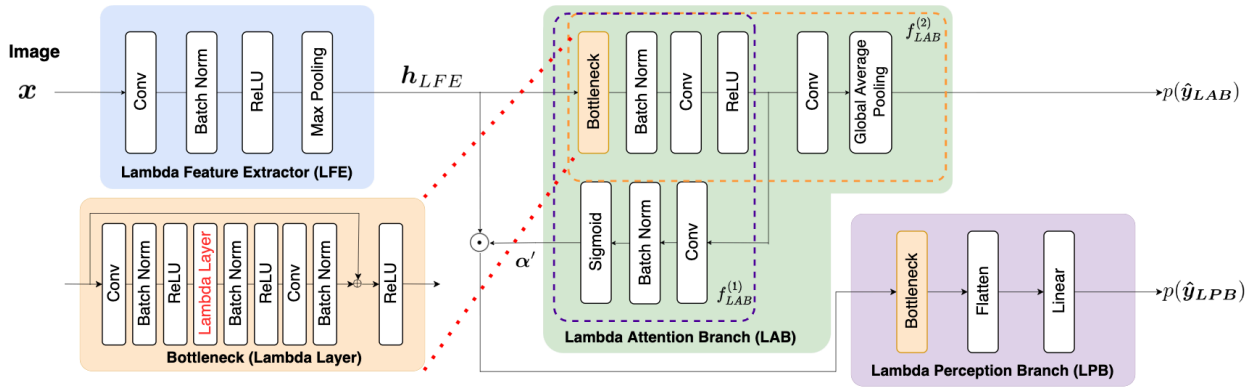

図 2　提案手法のネットワーク構造．“Conv” は畳み込み層を表す

3　問題設定

本論文は，分類問題におけるモデルの判断根拠の視覚的説明の生成として，画像中の重要な領域を可視化するタスクを扱う．特に，Lambda Networks [2] に基づく transformer における説明に焦点をあてる．本タスクにおいては、モデルの予測に貢献した画素に注目した視覚的説明が望ましい．本タスクの入出力を以下のように定義する．

- **入力:** 分類対象の画像 $\boldsymbol{x} \in \mathbb{R}^{c_1 \times w_1 \times h_1}$
- **出力:** 各クラスに属する確率の予測値 $p(\hat{\boldsymbol{y}})$

ここで，c_1, w_1, h_1, $C, \hat{\boldsymbol{y}}$ はそれぞれ入力画像のチャンネル数，横幅，縦幅，クラス数，ラベル（C 次元の 1-of-K 表現）を表す．また，出力に加えて，モデル中の attention map $\boldsymbol{\alpha} \in \mathbb{R}^{w_1 \times h_1}$ として画像の各画素における重要度が得られる．この $\boldsymbol{\alpha}$ を視覚的説明として使用する．

タスクの評価尺度には，Insertion-Deletion score [16] と，スパースな領域を有する画像に特化した Patch Insertion-Deletion (PID) sore [8] を用いる．本論文は Lambda Networks に基づく transformer を前提とし、クラスに依存しない視覚的説明を生成する．さらに，人間による視覚的説明の修正は扱わないものとする．

4　提案手法

4.1　構造

提案手法のネットワーク構造を図 2 に示す．本ネットワークは，LFE，LAB，LPB の 3 モジュールから構成される．はじめに，バックボーンネットワークを内部の Bottleneck 層で分割し，前半を LFE，後半を LPB とする．次に，LFE と LPB の間に並列に LAB を導入する．ここで，バックボーンネットワークは 1 つの bottleneck 層を含む．

LFE の入力は 2 節で定義した画像 $\boldsymbol{x} \in \mathbb{R}^{c_1 \times w_1 \times h_1}$ で，出力は $\boldsymbol{h}_{\mathrm{LFE}} \in \mathbb{R}^{c_2 \times w_2 \times h_2}$ である．ここで，c_2, w_2, h_2 はそれぞれ LFE の出力のチャンネル数，横幅，縦幅を表す．また，図 2 に示すように LFE は 1 つの畳み込み層やバッチ正規化層，Max Pooling 層を含み，画像の特徴量を抽出する．

LAB は，説明生成のための $f^{(1)}_{\mathrm{LAB}}$ と説明と分類を結びつけるための $f^{(2)}_{\mathrm{LAB}}$ の 2 つに分かれる．$f^{(1)}_{\mathrm{LAB}}$ の入力は $\boldsymbol{h}_{\mathrm{LFE}}$ であり，出力は $\tilde{\boldsymbol{\alpha}} \in \mathbb{R}^{w_2 \times h_2}$ である．視覚的説明として使用する $\boldsymbol{\alpha}$ は，$\tilde{\boldsymbol{\alpha}}$ を $w_1 \times h_1$ に拡大して得られる．また，予測に重要でない領域を削除して LPB に入力するために，$\tilde{\boldsymbol{\alpha}}$ のうち $\theta_{\boldsymbol{\alpha}}$ よりも小さな値を 0 として $\boldsymbol{\alpha}' \in \mathbb{R}^{w_2 \times h_2}$ とする．

$$\tilde{\boldsymbol{\alpha}} = f^{(1)}_{\mathrm{LAB}}(\boldsymbol{h}_{\mathrm{LFE}}) \tag{1}$$

$$\alpha'_{ij} = \begin{cases} \tilde{\alpha}_{ij} & (\theta_{\boldsymbol{\alpha}} < \tilde{\alpha}_{ij}) \\ 0 & (\mathrm{otherwise}) \end{cases} \tag{2}$$

$f^{(2)}_{\mathrm{LAB}}$ の入力は $\boldsymbol{h}_{\mathrm{LFE}}$ であり，出力は入力画像がどのクラスに属するかの確率の予測値 $p(\hat{\boldsymbol{y}}_{\mathrm{LAB}})$ である．損失関数に $p(\hat{\boldsymbol{y}}_{\mathrm{LAB}})$ を加えることで，LAB を分類に直接関連付けて学習させることができる．その結果，分類結果と強く関連する視覚的説明を生成できる．

次に，LPB は LFE と LAB の出力を元に分類を行う．LPB の入力は $\boldsymbol{\alpha}' \odot \boldsymbol{h}_{\mathrm{LFE}}$ である．マスク処理をした $\boldsymbol{\alpha}$ と $\boldsymbol{h}_{\mathrm{LFE}}$ を掛け合わせることにより，予測に重要な領域のみを入力とする．また，出力は入力画像がどのクラスに属するかの確率の予測値 $p(\hat{\boldsymbol{y}}'_{\mathrm{LPB}})$ である．ここで，$\odot$ はアダマール積を表す．

4.2　損失関数

損失関数として以下を使用する．

$$\mathcal{L} = \mathcal{L}_{\mathrm{LPB}} + \mathcal{L}_{\mathrm{LAB}} + \lambda \mathcal{L}_{\mathrm{KL}} \tag{3}$$

$$\mathcal{L}_{\mathrm{LAB}} = \mathrm{CE}(\hat{\boldsymbol{y}}_{\mathrm{LAB}}, \boldsymbol{y}) \tag{4}$$

$$\mathcal{L}_{\mathrm{LPB}} = \mathrm{CE}(\hat{\boldsymbol{y}}_{\mathrm{LPB}}, \boldsymbol{y}) \tag{5}$$

$$\mathcal{L}_{\mathrm{KL}} = \mathrm{D}_{\mathrm{KL}}(\hat{\boldsymbol{y}}_{\mathrm{LPB}}(\boldsymbol{x}) || \hat{\boldsymbol{y}}_{\mathrm{LPB}}(\tilde{\boldsymbol{x}})) \tag{6}$$

ここで，$\tilde{\boldsymbol{x}}$，$\boldsymbol{y}_{\mathrm{LPB}}(z)$ はそれぞれ、$\boldsymbol{x}$ をバイアス画像でマスクした画像，z を入力としたときの LPB の出力を表す．ここで，バイアス画像とは，訓練集合に含まれる画像を画素方向に平均化した画像である．また，$\mathrm{CE}(\cdot,\cdot)$，$\mathrm{D}_{\mathrm{KL}}(\cdot||\cdot)$，$\lambda$ はそれぞれ，交差エントロピー誤差，カルバック・ライブラー情報量，損失関数の重みを表す．

4.3　評価指標

本タスクの評価指標として，Insertion-Deletion score [16] およびその拡張である Patch Insertion-Deletion (PID) score [8] を用いる．Insertion-Deletion score は，説明手法の評価指標として標準的な評価指標であるため使用する．また，PID score は，スパースな重要領域を有する画像に特化した評価指標であるため使用する．

表 1　提案手法のパラメータ設定

Optimizer	AdamW	
	$(\beta_1 = 0.9, \beta_2 = 0.999)$	
Learning rate	LAB，Linear	1.0×10^{-3}
	LFE，LPB	1.0×10^{-4}
Weight decay	0.09	
mask ratio	0.8	
θ_α	0.50	
loss weight	0.50	
m_{guided}	1	

標準的な Deletion score は，attention map の値の低いピクセルから順に，各画素値を 0 で置換を行う．一方で，提案手法は，バイアス画像における同じ位置の画素値で置換する．

$$\tilde{x}_{ij} = \begin{cases} x_{ij} & (\alpha_{ij} > \theta) \\ x'_{ij} & (\text{otherwise}) \end{cases} \tag{7}$$

ここで，α_{ij}，θ はそれぞれ，attention map の (i, j) 成分の値，閾値を表す．また，$\tilde{x}_{ij}$，x'_{ij} はそれぞれ，置換後の画像，バイアス画像の $(i,\ j)$ 成分を表す．

本タスクで使用する PID score は次式で表される．

$$\text{PID} = \text{AUC(patch-insertion)} - \text{AUC(patch-deletion)} \tag{8}$$

ここで，AUC は Area Under Curve を表す．

patch-insertion 曲線および patch-deletion 曲線は，以下の手順で得られる．はじめに，入力 $\boldsymbol{x}$ をパッチ（部分行列）$\boldsymbol{p}_{ij} \in \mathbb{R}^{c_1 \times m^2}$ に分割する．ここで，m はパッチの大きさである．次に，$\boldsymbol{\alpha}$ に max pooling を適用し，各パッチごとに $\boldsymbol{\alpha}_p \in \mathbb{R}^{m^2}$ を作成する．$\boldsymbol{\alpha}_p$ の各要素を降順に $\alpha_{i_1 j_1}$，$\alpha_{i_2 j_2}, ..., \alpha_{i_m j_m}$ として，集合 A_n を次のように定義する．

$$A_n = \left\{(i_k,\ j_k) | k \leqq n\right\} \tag{9}$$

ここで，n は挿入・削除したパッチの数を表す．A_n を用いると，patch-insertion，patch-deletion の入力 $\boldsymbol{i}_n$，$\boldsymbol{d}_n$ はそれぞれ次のように表される．

$$(\boldsymbol{i}_n,\ \boldsymbol{d}_n) = \begin{cases} (\boldsymbol{p}_{ij},\ \boldsymbol{p}'_{ij}) & (i,\ j) \in A_n \\ (\boldsymbol{p}'_{ij},\ \boldsymbol{p}_{ij}) & (\text{otherwise}) \end{cases} \tag{10}$$

最後に，$\boldsymbol{i_n}$ と $\boldsymbol{d_n}$ をモデルに入力したときの出力をそれぞれ，$\boldsymbol{y}^{(\text{ins},\ n)}$，$\boldsymbol{y}^{(\text{del},\ n)}$ とする．このとき，n と $\boldsymbol{y}_c^{(\text{ins},\ n)}$，$\boldsymbol{y}_c^{(\text{del},\ n)}$ をプロットして得られる曲線がそれぞれ，patch-insertion 曲線，patch-deletion 曲線である．ここで，c は $\boldsymbol{x}$ が属するクラスを表す．また，$m = 1$ かつ任意の $(i,\ j)$ において，$\boldsymbol{p}'_{ij} = \boldsymbol{0}$ の場合，PID score は標準的な Insertion-Deletion score に一致する．

5　実験

5.1　実験設定

評価実験のデータセットには DeFN magnetograms dataset [13] を使用した．DeFN magnetograms dataset は，Solar Dynamic Observatory [20] のウェブアーカイブより収集した，Helioseismic and Magnetic Imager [17] で撮影された 1 時間間隔の太陽画像および 24 時間以内に発生する最大の太陽フレアクラスをラベルとして含む．また，

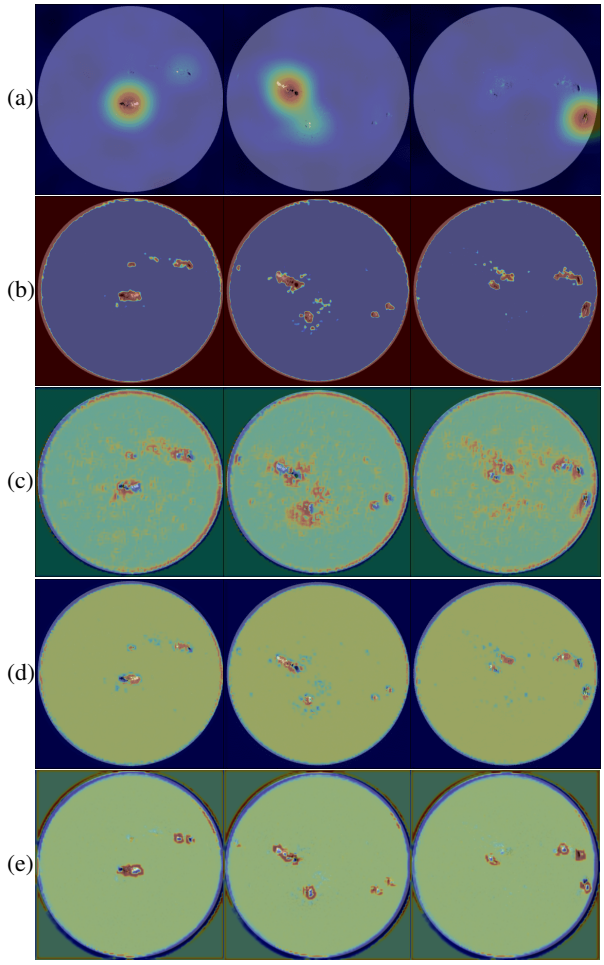

図 3　(a)RISE，(b)Lambda attention，(c)ABN，(d)LABN，(e) 提案手法の定性的結果

DeFN magnetograms dataset は 2010 年 6 月から 2017 年 12 月までの合計 61315 サンプルを含む．

DeFN magnetograms dataset は，最大の太陽フレアクラスのラベルとして，O，C，M，X の 4 段階の太陽フレアクラスを含む．そこで，O・C の 56078 サンプルを “< M”，M・X の 5237 サンプルを “≥ M” とする 2 段階に変換し本タスクで扱うラベル $\boldsymbol{y}$ とし，画像 $\boldsymbol{x}$ は 512×512 にリサイズし標準化を行った．さらに，バイアス画像を “< M” クラスとして訓練集合に追加した．ただし，入力画像とバイアス画像でマスクした画像との比較を行うため，他のデータ拡張は行っていない．また，2010-2015 年の 45530 サンプルを訓練集合に，2016 年の 7795 サンプルを検証集合に，2017 年の 7990 サンプルをテスト集合に割り当てた．

訓練集合，検証集合，テスト集合はそれぞれパラメータの学習，ハイパーパラメータの検証，性能の評価に使用した．検証集合における損失関数の値が 5 回連続で改善しなかった場合に Early Stopping を行った．このとき，検証集合における損失関数の値が最も低いときのテスト集合における精度を最終的な精度とした．

提案手法のパラメータ設定を表 1 に示す．ここで，mask ratio，m_{guided} はそれぞれ，入力画像をバイアス画像でマスクする割合，学習時のマスクのパッチの大きさを表す．また，本実験は，warmup と Cosine-Decay によ

表 2　各手法における定量的結果

Method	TSS	PID				
		$m = 1$	$m = 16$	$m = 32$	$m = 64$	$m = 128$
RISE [16]	0.670 ± 0.044	0.319 ± 0.015	0.179 ± 0.080	0.130 ± 0.045	0.136 ± 0.050	0.101 ± 0.033
Lambda attention	0.670 ± 0.044	−0.101 ± 0.074	−0.105 ± 0.073	−0.116 ± 0.081	−0.123 ± 0.078	−0.093 ± 0.054
LABN	0.653 ± 0.118	0.111 ± 0.273	0.084 ± 0.111	0.150 ± 0.183	0.183 ± 0.253	0.230 ± 0.329
Ours (LABN-S)	**0.795 ± 0.08**	**0.560 ± 0.160**	**0.748 ± 0.102**	**0.755 ± 0.100**	**0.757 ± 0.094**	**0.756 ± 0.096**

表 3　Ablation Studies の定量的結果

Condition	PID				
	$m = 1$	$m = 16$	$m = 32$	$m = 64$	$m = 128$
(i) w/o $\mathcal{L}_{\mathrm{KL}}$	0.124 ± 0.205	0.446 ± 0.175	0.405 ± 0.224	0.388 ± 0.236	0.382 ± 0.248
(ii) w/ LABN scheduler	0.305 ± 0.311	0.559 ± 0.368	0.792 ± 0.086	0.547 ± 0.394	0.538 ± 0.096
(iii) w/o バイアス画像	0.460 ± 0.271	**0.774 ± 0.117**	**0.792 ± 0.086**	**0.808 ± 0.060**	**0.807 ± 0.069**
(iv) Ours (LABN-S)	**0.560 ± 0.160**	0.748 ± 0.102	0.755 ± 0.100	0.757 ± 0.094	0.756 ± 0.096

る学習率のスケジューリングを行った．学習初期の 5 エポックで学習率を線形に増加させる．warmup 終了時の値を基準点として，その後 10 エポックで学習率を減少させ，再度基準点に戻し同様に減少させることを繰り返した．学習にはメモリ 16GB 搭載の GeForce RTX 3080 および Intel Core i9-11980HK を使用した．モデルの訓練時間および 1 サンプルあたりの推論に要した時間はそれぞれ，約 7 時間および約 1 秒であった．

5.2　定量的結果

ベースライン手法として, RISE [16] と Lambda attention, LABN [8] を用いた．transformer の注意機構を説明として利用した [9] を参考に，Lambda attention の視覚的説明を生成した．Lambda attention は transformer に対する説明生成の標準的な手法であり，RISE は汎用的なモデルに適用できる標準的な手法のためベースライン手法として使用した．また，LABN はスパースな重要領域を有する画像に特化した説明生成の標準的な手法であるため使用した．また，DeFN magnetograms dataset は重要領域がスパースな画像を含むため，PID score を主要評価尺度として使用した．$m = 1$ における PID score は，視覚的説明の標準的な評価指標である Insertion-Deletion score と一致する．また，モデルの評価尺度として太陽フレア予測タスクにおいて標準的な評価尺度である TSS を使用した．

表 2 に定量的結果を示す．提案手法においては，実験を 4 回行い，その平均値と標準偏差を示した．また，ベースライン手法の結果は LABN [8] より引用した．ここで，PID score の計算には，“≥ M” のデータのみを使用した．これは，“< M” のデータは，説明として適切な太陽フレアに影響する領域を含まないためである．

表 2 より，$m = 1$ において，提案手法は RISE，Lambda attention の PID をそれぞれ 0.325 ポイント，0.186 ポイント上回っていた．同様に，$m = 16$ において，それぞれ 0.487 ポイント，0.334 ポイント上回っていた．さらに，LABN と提案手法を比較すると，5 つのパッチサイズで PID が向上した．これらの結果は，提案手法が重要領域がスパースな画像に対しても適切な説明を生成したことを示している．

5.3　Ablation Studies

Ablation Studies として，以下の 3 条件を定めた．

(i) $\mathcal{L}_{\mathrm{KL}}$ による性能への寄与を調べるために，$\mathcal{L}_{\mathrm{KL}}$ を損失に含まずに学習を行った．

(ii) scheduler による性能への寄与を調べるために，比較として，LABN と同様の scheduler (3 エポック精度が向上しない場合に学習率の大きさを $\frac{1}{10}$ にする) を用いて学習をした．

(iii) バイアス画像の訓練集合に含めることの性能への寄与を調べるために，バイアス画像を訓練集合から除いて学習を行った．

表 3 に Ablation Studies の定量的結果を示す．(i) より，損失関数から $\mathcal{L}_{\mathrm{KL}}$ の項を削除すると，(iv) と比較して PID score が全体的に約 0.3 ポイント減少した．また，(ii) より scheduler を変更すると，(iv) と比較して全ての PID score が約 0.2 ポイント減少した．また，$m = 128$ を除き，標準偏差は約 0.2 ポイントから 0.3 ポイント増加している．一方で (iii)，(iv) は，大きな差はなかった．これらの結果から，バイアス画像を学習集合に加えたことによる PID score への影響はほとんどない．また，$\mathcal{L}_{\mathrm{KL}}$ を追加することによって，重要な領域を正しく特定することができていると考えられる．

5.4　定性的結果

図 3 に定性的な結果を示す．各行は，上から RISE, Lambda attention，ABN，LABN および提案手法の視覚的説明を表す．また，各列は提案手法において予測に成功した各時刻におけるサンプルを示す．RISE は太陽フレアの原因となる黒点を含む円領域に注目していた．また，Lambda attention は予測に無関係な背景部分に注目し，ABN は注目領域に予測に貢献しない画素を含んでいた．一方，LABN と提案手法は黒点を細かく説明していた．さらに，提案手法は LABN と比較して黒点以外の不適切な領域の注目度を下げることができた．これらの結果から，提案手法はベースライン手法と比較して，重要度の低い領域の影響を軽減し，解釈性の高い視覚的説明を生成することができたと考えられる．

6　おわりに

本論文では，画像分類モデルにおける判断根拠の視覚的説明を生成するタスクに取り組んだ．提案手法の主要な貢献を以下に示す．

- モデルの出力に対する重要度の低い領域の影響を軽減し，視覚的説明の解釈性を高めるために，Saliency Guided Training [1] で提案された損失を導入した．
- マスクによる分布の差を軽減するために，Insertion-Deletion score および PID score におけるマスク画像として，バイアス画像を導入した．
- 提案手法は，Insertion-Deletion score および PID においてベースライン手法を上回った．

謝辞

本研究の一部は，JSPS 科研費 20H04269c と，NEDO の助成を受けて実施されたものである．

参考文献

[1] Soheil Feizi Aya Abdelsalam Ismail , “Improving deep learning interpretability by saliency guided training”, in *NeurIPS* (2021)

[2] Irwan Bello , “LambdaNetworks: Modeling Long-Range Interactions without Attention”, in *ICLR* (2021)

[3] Alexander Binder et al., “Layer-Wise Relevance Propagation for Neural Networks with Local Renormalization Layers”, in *ICANN*, pp. 63–71 (2016)

[4] Hila Chefer , Shir Gur et al., “Transformer Interpretability Beyond Attention Visualization”, in *CVPR*, pp. 782–791 (2021)

[5] Arun Das et al., “Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey”, *arXiv preprint arXiv:2006.11371* (2020)

[6] Wang Haofan , Wang Zifan , Du Mengnan et al., “Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks”, in *CVPR*, pp. 24–25 (2020)

[7] Fukui Hiroshi , Hirakawa Tsubasa et al., “Attention Branch Network: Learning of Attention Mechanism for Visual Explanation”, in *CVPR*, pp. 10705–10714 (2019)

[8] Iida et al., “Lambda Attention Branch Networks による視覚的説明生成”, *JSAI* (2022)

[9] Vig Jesse , “A multiscale visualization of attention in the transformer model”, in *ACL*, pp. 37–42 (2019)

[10] Salman Khan , Muzammal Naseer , Munawar Hayat , Syed Waqas Zamir et al., “Transformers in Vision: A Survey”, *arXiv preprint arXiv:2101.01169* (2021)

[11] Scott Lundberg et al., “A unified approach to interpreting model predictions”, in *NeurIPS*, pp. 4765–4774 (2017)

[12] Aly Magassouba , Komei Sugiura et al., “Multimodal attention branch network for perspective-free sentence generation”, in Leslie Pack Kaelbling , Danica Kragic et al. eds., *CoRL*, Vol. 100, pp. 76–85 (2019)

[13] Sugiura Komei Nishizuka Naoto et al., “Deep Flare Net (DeFN) Model for Solar Flare Prediction”, *The Astrophysical Journal*, Vol. 858, No. 2, p. 113 (8pp) (2018)

[14] Tadashi Ogura , Aly Magassouba , Komei Sugiura et al., “Alleviating the burden of labeling: Sentence generation by attention branch encoder-decoder network”, *RA-L*, Vol. 5, No. 4, pp. 5945–5952 (2020)

[15] Bowen Pan , Rameswar Panda , Yifan Jiang et al., “IA-RED^2: Interpretability-aware redundancy reduction for vision transformers”, in *NeurIPS* (2021)

[16] Vitali Petsiuk , Abir Das et al., “RISE: Randomized input sampling for explanation of black-box models”, in *BMVC*, p. 151(13pp) (2018)

[17] Scherrer Philip , Schou Jesper , Bush Rock et al., “The helioseismic and magnetic imager (HMI) investigation for the solar dynamics observatory (SDO)”, *Solar Physics*, Vol. 275, pp. 207–227 (2012)

[18] Selvaraju Ramprasaath et al., “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”, in *ICCV*, pp. 618–626 (2017)

[19] Marco Ribeiro , Sameer Singh et al., “”Why Should I Trust You?”: Explaining the Predictions of Any Classifier”, in *KDD*, p. 1135–1144 (2016)

[20] Pesnell William , Thompson Barbara et al., “The Solar Dynamics Observatory (SDO)”, *Solar Physics*, Vol. 275, No. 1–2, pp. 3–15 (2012)

[21] Bolei Zhou , Aditya Khosla , Agata Lapedriza , Aude Oliva et al., “Learning deep features for discriminative localization”, in *CVPR*, pp. 2921–2929 (2016)