Layer-Wise Relevance Propagation for ResNet：保全公理を満たす視覚的説明生成

Layer-Wise Relevance Propagation for ResNet:
Visual Explanations Generation with Conservation Property

小槻 誠太郎 *1 Seitaro Otsuki　飯田 紡 *1 Tsumugi Iida　デュブレ フェリックス *1 Félix Doublet　平川 翼 *2 Tsubasa Hirakawa　山下 隆義 *2 Takayoshi Yamashita
藤吉 弘亘 *2 Hironobu Fujiyoshi　杉浦 孔明 *1 Komei Sugiura
*1慶應義塾大学 Keio University　*2中部大学 Chubu University

In the modern era, where deep learning is applied across a wide range of fields, the explainability of deep learning models is crucial, and the generation of explanations with high transparency is desirable. Layer-wise Relevance Propagation (LRP) is mentioned as a method for generating explanations by backpropagating relevance scores, offering high transparency. However, this method has been applied only to models without skip connections, as its application to models with skip connections does not satisfy the conservation property, resulting in poor quality explanations. Therefore, this paper proposes a method for calculating the backpropagation of relevance scores that satisfies the conservation property in models with skip connections. The proposed method outperformed existing methods in terms of Insertion-Deletion Score while satisfying the conservation property.

1. はじめに

ニューラルネットワークの応用が広がる中，これらのモデルの説明可能性の重要性が増している [Shrikumar 17]．欧州議会は 2023 年 12 月に公布された AI 法において，AI システムは安全かつ透明でなければならないと宣言している [Madiaga 23]．ニューラルネットワークの推論における透明性の欠如は，モデルの予測の妥当性を検証する上で重大な課題である．この問題に対処するため，複数の研究がモデルの判断根拠を視覚的説明として可視化する手法を提案している [Iida 22,Ogura 20]．視覚的説明の生成は，生成された視覚的説明を検証するための正解マスクが欠如した状態で重要な領域を過不足なく正確に抽出する必要がある難しいタスクである．Layer-wise Relevance Propagation (LRP) は保全公理を満たす透明性の高い説明手法として一般的なものであるが，ResNet の残差接続は LRP の既存の Relevance 伝播則では扱われておらず，LRP によって生成された ResNet モデルの視覚的説明は，しばしば信頼できる結果を示さない．そこで，本研究では ResNet [He 16] に対して適用可能な LRP の新しい計算方法を提案する．

2. 問題設定

本研究では，モデルの予測の視覚的説明として，画像内の重要な領域を可視化するタスクに焦点を当てる．入力は画像 $\boldsymbol{x} \in \mathbb{R}^{c^{(0)} \times h^{(0)} \times w^{(0)}}$ である．ここに，$c^{(0)}$，$h^{(0)}$，および $w^{(0)}$ はそれぞれ入力画像のチャンネル数，縦幅，および横幅を示す．モデルの出力 $p(\hat{\boldsymbol{y}}) \in [0,1]^C$ は各クラスの予測確率を示し，ここに C はクラス数である．さらに，各ピクセルの重要性は視覚的説明として使用される Attribution $\boldsymbol{\alpha} \in \mathbb{R}^{h^{(0)} \times w^{(0)}}$ として得られる．視覚的説明は，モデルの予測に貢献したピクセルに焦点を当てることが望ましい．図 1 に本タスクの代表例を示す．左図および右図はそれぞれ入力画像および対応する視覚的説明である．本稿では，モデルが ResNet アーキテクチャに基づくことを前提とする．

連絡先: 小槻誠太郎，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，otsu8sei14@keio.jp

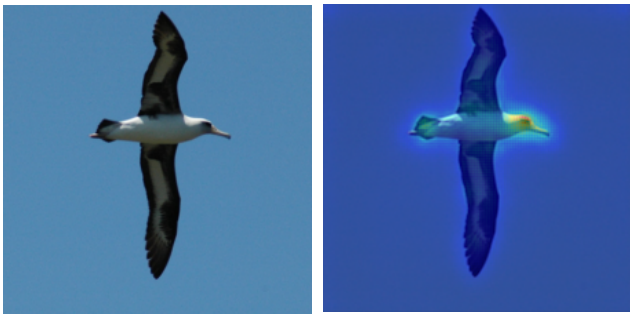


図 1: 入力画像（左）および視覚的説明生成（右）の代表例

3. 提案手法

本研究では，残差接続に対する新しい Relevance 伝播則を導入することで，Layer-wise Relevance Propagation (LRP) [Bach 15] を拡張し，特に ResNet [He 16] モデルに適用可能な LRP を提案する．提案手法による拡張は，残差接続を持つモデルに対する LRP の計算方法を定義するものであり，提案手法は，残差接続を持つモデルに広く適用可能である．図 2 に，ResNet のモデル構造とそれに対する LRP 適用の概要を示す．

3.1 ResNet50 のモデル構造

ResNet50 は，畳み込み層，バッチ正規化 (BN) 層，最大プーリング層，16 個の Bottleneck モジュール，グローバル平均プーリング (GAP) 層，そして全結合層で構成される．各 Bottleneck モジュールは，スキップ接続と残差ブロックで構成される．残差ブロックは，次元削減のための 1×1 畳み込み，3×3 畳み込み，そして次元復元のための別の 1×1 畳み込みの 3 つの畳み込み層で構成される．これらの層には，BN 層および ReLU 活性化関数が順に続く．Bottleneck モジュールには，Simple Bottleneck (S-Bottleneck) モジュールと Downsampling Bottleneck (D-Bottleneck) モジュールの 2 種類がある．S-Bottleneck モジュールは，スキップ接続において恒等写像を使用する．一方，D-Bottleneck モジュールは，次元を一致させるためにスキップ接続において線形射影を使用する．線形射影は 1×1 畳み込みと

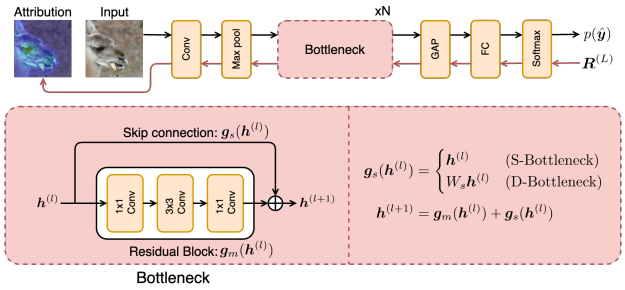


図 2: 提案手法の概略図（上）および Bottleneck モジュールの構造（下）

して実装される．本研究では，これらの Bottleneck モジュールにおいて Relevance がどのように伝播されるべきかについて主に議論し，焦点を当てる．ResNet モデルの学習には，損失関数としてクロスエントロピー損失を使用した．

3.2 連続する層間の LRP

LRP は，モデルの予測をネットワークを通じて Relevance score として逆伝播する説明手法である．その主要な特徴は，ユニットによって受け取られた Relevance score の合計が，同じユニットによって再分配された合計と等しくなるように保証する伝播則である．この特徴は保全特性として知られている．ここに連続する層間の Relevance の伝播則を定式化する．l 番目および $l+1$ 番目の連続する層における中間特徴量をそれぞれ $\boldsymbol{h}^{(l)} \in \mathbb{R}^D$，$\boldsymbol{h}^{(l+1)} \in \mathbb{R}^E$ とする．さらに $\boldsymbol{f_\theta}$ を $\boldsymbol{h}^{(l)}$ から $\boldsymbol{h}^{(l+1)}$ へ写像する関数とし，これは $\boldsymbol{\theta}$ によってパラメータ化されるとする．今，$\boldsymbol{h}^{(l)}$ の Relevance score を $\boldsymbol{R}^{(l)}$，$h_i^{(l)}$ から $h_j^{(l+1)}$ への寄与量を z_{ij} と定義すると，$\boldsymbol{f_\theta}$ を通じて全ての $h_j^{(l+1)}$ から $h_i^{(l)}$ へ伝播する Relevance score は次のように表現できる：

$$R_i^{(l)} = \sum_j^E \frac{z_{ij}}{\sum_k^D z_{kj}} R_j^{(l+1)}. \tag{1}$$

ここに，$R_j^{(l+1)}$ および $R_i^{(l)}$ はそれぞれ $\boldsymbol{h}^{(l+1)}$ の j 番目の要素と $\boldsymbol{h}^{(l)}$ の i 番目の要素の Relevance score を示し，分母の $\sum_k^D z_{kj}$ は 2 つの連続する層間の Relevance score の総量の保全を保証する．z_{ij} の量は様々な方法で定式化できる．

本研究では，$\boldsymbol{f_\theta}$ を線形射影と見なすことができる場合の Relevance 伝播に z^+-Rule [Montavon 17] を採用する．z^+-Rule における線形射影 $\boldsymbol{f}$ を通じた Relevance 伝播は次のように書ける．

$$R_i^{(l)} = \sum_j^E \frac{w_{ji}^+ h_i^{(l)}}{\sum_k^D w_{jk}^+ h_k^{(l)}} R_j^{(l+1)}.$$

ここに，$w_{ji}^+ = \max(0, w_{ji})$，$\boldsymbol{f}(\boldsymbol{h}^{(l)}) = W\boldsymbol{h}^{(l)}$，$w_{ji}$ は $W \in \mathbb{R}^{E\times D}$ の (j,i) 番目の要素を示す．この文脈では，W^+ は W の非負の要素で構成される行列であり，すべての i および j に対して $W_{ij}^+ = \max(0, W_{ij})$ と定義される．上式は次のようにも定式化できる．

$$R_i^{(l)} = \sum_j^E \frac{\frac{\partial f_j^+}{\partial h_i^{(l)}}(\boldsymbol{h}^{(l)}) h_i^{(l)}}{f_j^+(\boldsymbol{h}^{(l)})} R_j^{(l+1)}.$$

ここに，$\boldsymbol{f}^+(\boldsymbol{h}^{(l)}) = W^+\boldsymbol{h}^{(l)}$ である．本研究では畳み込み層，最大プーリング層，GAP 層，全結合層については特定の重み行列を持つ線形射影として定式化できるため，この伝播則を適用する．ただし，ReLU および BN 層については，Relevance score を修正せずに通過させる．

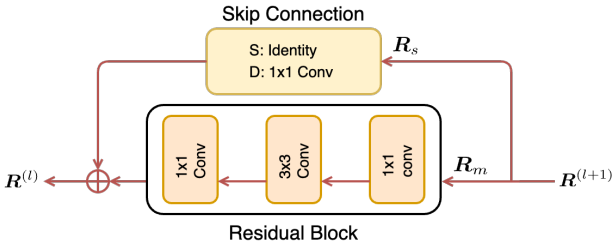


図 3: Bottleneck モジュールにおける Relevance 分割

3.3 Bottleneck モジュールにおける LRP

本節では，Bottleneck モジュールのための複数の潜在的な Relevance 伝播則を議論し定式化する．式 (1) で見たように，LRP は本来 2 つの連続する層間の伝播則を定義しており，非連続な層を橋渡しするスキップ接続の存在を考慮していない．

事前準備として，まず Relevance score $\boldsymbol{R}^{(l+1)}$ をスキップ接続に伝播させるための $\boldsymbol{R}_s$ と残差ブロックに伝播させるための $\boldsymbol{R}_m$ の 2 つに分割する．また，保全特性に従って，以下の制約を課す．

$$\boldsymbol{R}^{(l+1)} = \boldsymbol{R}_s + \boldsymbol{R}_m.$$

ここに，図 3 に示すように $\boldsymbol{R}^{(l+1)}$ を分割する 2 つのアプローチ，即ち Symmetric Splitting アプローチおよび Ratio-Based Splitting アプローチを定式化する．Symmetric Splitting アプローチは $\boldsymbol{R}^{(l+1)}$ を次のように均等に分割する：

$$(R_s)_i = (R_m)_i = \frac{R_i^{(l+1)}}{2}.$$

この方法は直感的だが，中間特徴量 $\boldsymbol{h}^{(l+1)}$ の個々の要素の寄与を考慮していないため，Relevance の伝播が精密でなくなる可能性がある．次に Ratio-Based Splitting アプローチはより精緻なアプローチである．スキップ接続と残差ブロックの出力をそれぞれ $\boldsymbol{h}_s$，$\boldsymbol{h}_m$ とし，$\boldsymbol{R}^{(l+1)}$ を以下の条件を満たすように分割する．

$$(R_s)_i = \frac{R_i^{(l+1)} \cdot |(h_s)_i|}{|(h_m)_i| + |(h_s)_i|}, \quad (R_m)_i = \frac{R_i^{(l+1)} \cdot |(h_m)_i|}{|(h_m)_i| + |(h_s)_i|}.$$

このアプローチは $\boldsymbol{h}_s$ と $\boldsymbol{h}_m$ の各要素の絶対値の比率を考慮している．スキップ接続が恒等写像の場合，$\boldsymbol{h}_s = \boldsymbol{h}^{(l)}$ であり，$\boldsymbol{h}_m$ は $\boldsymbol{h}^{(l+1)} - \boldsymbol{h}^{(l)}$ に等しく，モデルのパラメータによる特徴の変化を表す．この場合，$\boldsymbol{h}_m$ の要素の絶対値が大きいほど，その要素へのモデルの寄与が大きいといえる．このアプローチは以下のように保全特性を満たす．

$$\begin{aligned}(R_s)_i + (R_m)_i &= \frac{R_i^{(l+1)} \cdot |(h_s)_i|}{|(h_m)_i| + |(h_s)_i|} + \frac{R_i^{(l+1)} \cdot |(h_m)_i|}{|(h_m)_i| + |(h_s)_i|} \\ &= \frac{R_i^{(l+1)} \cdot (|(h_m)_i| + |(h_s)_i|)}{|(h_m)_i| + |(h_s)_i|} = R_i^{(l+1)}.\end{aligned}$$

3.4 スキップ接続における LRP

3.1 節で述べたように，Bottleneck モジュールは 2 種類あり，それぞれ異なるスキップ接続を持つ．具体的には，S-Bottleneck モジュールはスキップ接続に恒等写像を持ち，D-Bottleneck モジュールは線形射影を使用する．両者は中間特徴量に対して異なる寄与をするため，2 種類のスキップ接続に異なる Relevance 伝播アプローチを適用する余地がある．線形射影を持つスキップ接続は，スキップ接続の出力と残差ブロックの出力の次元を一致させることを目的としているが，学習可能な射影を通じて重要な入力要素を選択的に強調することができる．一方，恒等写像を持つスキップ接続は変換を行わない．一つのアプローチは，恒等写像を持つスキップ接続の場合に $\boldsymbol{R}_s = 0$ と

表 1: CUB データセットおよび ImageNet データセットにおける定量的比較結果

Method	CUB			ImageNet		
	Insertion ↑	Deletion ↓	ID Score ↑	Insertion ↑	Deletion ↓	ID Score ↑
LRP [Bach 15]	0.058 ± 0.002	0.047 ± 0.001	0.011 ± 0.000	0.095	0.083	0.011
IG [Sundararajan 17]	0.020 ± 0.001	0.015 ± 0.001	0.006 ± 0.000	0.052	0.062	-0.011
Guided BP [Springenberg 15]	0.042 ± 0.002	0.014 ± 0.001	0.028 ± 0.002	0.115	0.057	0.057
Grad-CAM [Selvaraju 17]	0.508 ± 0.015	0.055 ± 0.004	0.453 ± 0.011	0.497	0.126	0.371
Score-CAM [Wang 20]	0.511 ± 0.017	0.054 ± 0.004	0.457 ± 0.014	0.488	0.133	0.355
Ours	$\mathbf{0.595 \pm 0.010}$	$\mathbf{0.014 \pm 0.000}$	$\mathbf{0.582 \pm 0.010}$	**0.563**	**0.018**	**0.545**

設定し，すべての Relevance を残差ブロックを通じて伝播させることである．もう一つのアプローチは，3.3 節で議論された Relevance 分割を恒等写像を持つスキップ接続にも適用することである．予備実験の結果に基づき，提案手法では，Ratio-Based Splitting アプローチを採用し，恒等写像を含むすべてのスキップ接続に Relevance 分割を適用する．さらにその有効性を評価するために，これらの条件に関する Ablation study を実施する．

最終的な Attribution $\boldsymbol{\alpha}$ を生成するため，説明した伝播則を順次適用し，Relevance score を逆伝播させることで，入力 $\boldsymbol{x}$ の Relevance score $\boldsymbol{R}^{(0)}$ を計算する．既存の LRP と同様に，$\boldsymbol{R}^{(0)}$ は $\boldsymbol{x}$ と同じ形状を持ち，そのチャネルごとの合計は $\boldsymbol{\alpha}_R$ として直接 Attribution として使用できる．しかし，$\boldsymbol{\alpha}_R$ は特定のエッジに過度に集中する傾向がある．そのため，$\boldsymbol{\alpha}_R$ の値を量子化することによってこの傾向を軽減した $\boldsymbol{\alpha}$ を得る．本操作を Heat Quantization (HQ) と呼ぶ．

4. 実験

4.1 データセットおよび実験設定

本実験では，データセットとして Caltech-UCSD Birds-200-2011 (CUB) データセット [Wah 11] および ImageNet [Deng 09] (ILSVRC) 2012 の検証集合を用いた．これらは視覚的説明生成タスクの標準的なデータセットである．

ImageNet の検証集合は，1,000 クラスからなる 50,000 枚の画像で構成されており，CUB データセットには，鳥類の種に属する 200 クラスからなる 11,788 枚の画像が含まれる．我々は CUB データセットを，それぞれ 5,394，600，5,794 サンプルからなる学習集合，検証集合，テスト集合に分割した．本研究ではこれらの訓練集合および検証集合をそれぞれパラメータの更新およびハイパーパラメータの選択に使用し，テスト集合を性能の評価に使用した．

また，事前処理として，入力画像を縦幅横幅ともに 224 ピクセルに縮小した．さらに学習時には，画像の反転およびランダム切り抜きによるデータ拡張を行った．本実験では，16GB のメモリを搭載した GeForce RTX 3080 および 64GB のメモリを搭載した Intel Core i9-11980HK を使用した．CUB データセットでの実験では，検証集合における損失関数の値が 6 エポック連続で改善しない場合に早期終了を行った．ResNet50 モデルの学習には 1 時間かかり，1 サンプルあたりの推論時間は約 4.66 ミリ秒であった．また，ImageNet における実験では，ImageNet 上で事前学習済みのモデルを使用した．なお，提案手法による視覚的説明生成における計算コストは誤差逆伝播と同様である．

4.2 定量的結果

表 1 に，ベースライン手法と提案手法の比較に関する定量的結果を示す．CUB データセットにおける実験では，各手法について 5 回の試行を行い，平均値と標準偏差を報告する．ImageNet における実験では，事前学習済みのモデルを使用して 1 回の実験を実施した．ベースラインとして，LRP [Bach 15]，Integrated Gradients (IG) [Sundararajan 17]，Guided BackPropagation (Guided BP) [Springenberg 15]，Grad-CAM [Selvaraju 17]，Score-CAM [Wang 20] を選択した．実験では評価尺度として説明生成タスクのための標準的な Insertion score，Deletion score，および Insertion-Deletion score（ID score）[Petsiuk 18] を使用した．さらに，最も標準的な尺度である ID score を主要な評価尺度とした．

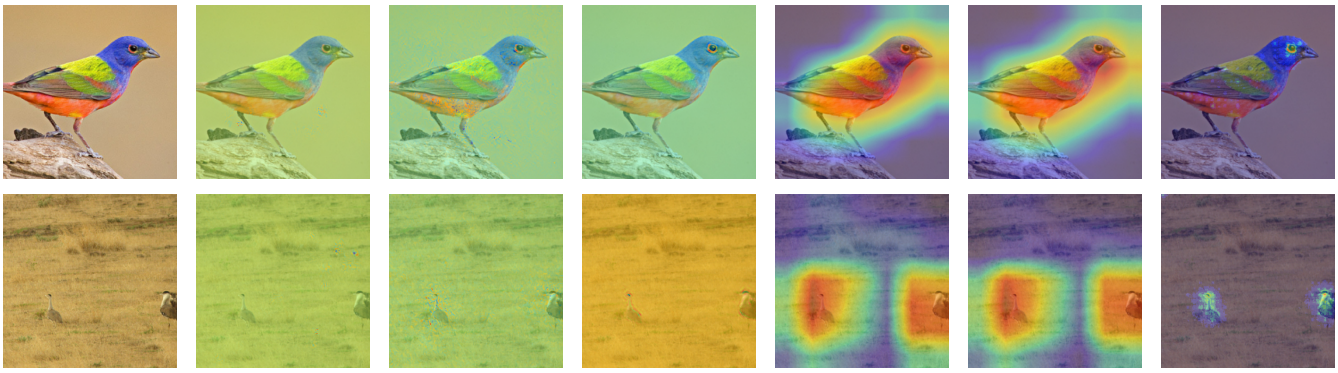
CUB データセットにおける実験では，提案手法の ID score は 0.582 であり，LRP，IG，Guided BP，Grad-CAM，Score-CAM の ID score はそれぞれ 0.011，0.006，0.028，0.453，0.457 であった．提案手法は，ベースラインの中で最も ID score が高い Score-CAM を ID score で 0.125 ポイント上回り，Insertion score と Deletion score の両方で最良であった．さらに，ImageNet における実験においても，提案手法は ID score で全てのベースラインを上回った．具体的には，最も ID score が高いベースラインである Grad-CAM を ID score で 0.174 ポイント上回り，Insertion score と Deletion score の両尺度で最良であった．

4.3 定性的結果

図 4 に定性的結果を示す．列 (a) は入力画像である．列 (b) から (f) にかけては，ベースライン手法によって生成された Attribution を元の画像に重ねて描画したものを示しており，列 (g) は提案手法によって生成された結果を表している．列 (b)，(c)，(d) は，それぞれ LRP，IG，Guided BP によって生成された説明を示す．これらの手法による説明はほとんどの領域に等しい Attribution を割り当ててしまっている．さらに，列 (e) と (f) は，それぞれ Grad-CAM と Score-CAM によって生成された説明を示している．これらの結果は，関連するオブジェクト全体に注目領域が及んでいるが，周囲の背景にも注目している．一方，列 (g) は提案手法によって生成された Attribution を示している．これらの Attribution は，関連するオブジェクトに詳細に集中し，背景領域への注目を最小限に抑え，より適切な説明を提供している．

4.4 Bottleneck モジュールにおける伝播則の Ablation study

4 節で議論した Bottleneck モジュールに対する複数の Relevance 伝播則の重要な設計要素に関して，ImageNet 上で Ablation study を実施した．表 2 にその定量的結果を示す．手法 (iv) の ID score は 0.545 であり，手法 (iii) を 0.035 ポイント上回った．これは，恒等写像を持つスキップ接続に Relevance を割り当てることの有効性を示している．さらに，手法 (iv) の ID score は手法 (ii) を 0.028 ポイント上回った．こ



(a) Original (b) LRP [Bach 15] (c) IG [Sundararajan 17] (d) Guided BP [Springenberg 15] (e) Grad-CAM [Selvaraju 17] (f) Score-CAM [Wang 20] (g) Ours

図 4: CUB データセット (上) および ImageNet データセット (下) における定性的結果

表 2: Bottleneck モジュールにおける伝播則に関する Ablation study の結果

Method	Relevance on Identical Skip Connection	Relevance Splitting approach	Insertion ↑	Deletion ↓	ID Score ↑
(i)		Symmetric Splitting	0.543	0.033	0.510
(ii)	✓	Symmetric Splitting	0.553	0.036	0.517
(iii)		Ratio-Based Splitting	0.543	0.033	0.510
(iv)	✓	Ratio-Based Splitting	**0.563**	**0.018**	**0.545**

れは，スキップ接続と残差接続の出力の要素ごとの比率を考慮する Ratio-Based Splitting アプローチを採用することで，Symmetric Splitting アプローチを使用する場合と比較して，より高品質な Attribution が得られることを示唆している．恒等写像を持つスキップ接続への Relevance の伝播を取り除いた手法 (iii) は，Ratio-Based Splitting を Symmetric Splitting で置換した手法 (ii) よりも ID score が低い．よって，恒等写像を持つスキップ接続に Relevance を割り当てることが，Attribution の品質向上に最も大きく寄与したといえる．

5. おわりに

本研究では，モデルの予測の視覚的説明として，画像内の重要な領域を可視化するタスクに焦点を当て，ResNet [He 16] に対する Layer-wise Relevance Propagation（LRP）の新しい計算方法を提案した．本研究の貢献を以下に示す．

- スキップ接続と残差ブロックからの出力が合流する点での Relevance 分割を導入し，残差接続を持つモデルに対して保全特性を維持しながら LRP を拡張した．
- 特定の画像パターンに対して Attribution が過剰に集中する問題を軽減するため，Heat Quantization を導入した．
- ID score を含む本タスクの標準的な評価指標において，提案手法は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 23H03478，JST CREST，NEDO の助成を受けて実施されたものである．

参考文献

[Bach 15] Bach, S., et al.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation, *PLOS ONE*, Vol. 10, No. 7, pp. 1–46 (2015)

[Deng 09] Deng, J., Dong, W., et al.: ImageNet: A Large-Scale Hierarchical Image Database, in *CVPR*, pp. 248–255 (2009)

[He 16] He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, in *CVPR*, pp. 770–778 (2016)

[Iida 22] Iida, T., Komatsu, T., Kaneda, K., et al.: Visual Explanation Generation Based on Lambda Attention Branch Networks, in *ACCV*, pp. 3536–3551 (2022)

[Madiaga 23] Madiaga, : Artificial intelligence act (2023)

[Montavon 17] Montavon, G., Lapuschkin, S., et al.: Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition, *Pattern Recognition*, Vol. 65, pp. 211–222 (2017)

[Ogura 20] Ogura, T., Magassouba, A., Sugiura, K., 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)

[Petsiuk 18] Petsiuk, V., Das, A., and Saenko, K.: RISE: Randomized Input Sampling for Explanation of Black-box Models, in *BMVC*, pp. 151–164 (2018)

[Selvaraju 17] Selvaraju, R., et al.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, in *ICCV*, pp. 618–626 (2017)

[Shrikumar 17] Shrikumar, A., Greenside, P., and Kundaje, A.: Learning Important Features Through Propagating Activation Differences, in *ICML*, Vol. 70, pp. 3145–3153 (2017)

[Springenberg 15] Springenberg, J., Dosovitskiy, A., Brox, T., and Riedmiller, M.: Striving for Simplicity: The All Convolutional Net, in *ICLR (workshop track)* (2015)

[Sundararajan 17] Sundararajan, M., Taly, A., and Yan, Q.: Axiomatic Attribution for Deep Networks, in *ICML*, Vol. 70, pp. 3319–3328 (2017)

[Wah 11] Wah, C., Branson, S., Welinder, P., et al.: The Caltech-UCSD Birds-200-2011 Dataset, Technical Report CNS-TR-2011-001, California Institute of Technology (2011)

[Wang 20] Wang, H., Wang, Z., Du, M., et al.: Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks, in *CVPR*, pp. 24–25 (2020)