クラック分類タスクにおけるLayer-Wise Relevance Propagation for Branch Networksを用いた視覚的説明生成

Visual Explanation Generation for Road Damage Classification by Using Layer-Wise Relevance Propagation for Branch Networks

飯田 紡 [1*]　小槻 誠太郎 [1]　平川 翼 [2]　山下 隆義 [2]　藤吉 弘亘 [2]　杉浦 孔明 [1]
Tsumugi Iida[1], Seitaro Otsuki[1], Tsubasa Hirakawa[2],
Takayoshi Yamashita[2], Hironobu Fujiyoshi[2], Komei Sugiura[1]

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

Abstract: 深層学習が幅広い分野に応用されている現代において，深層学習モデルの説明性は重要であるが，説明生成のためのモジュールを利用する場合，それ自体が複雑になってしまい透明性が低い．逆伝播により説明を生成する手法は透明性が高いものの，cyclic connection を持たないモデルのみに適用されている．そのため，ブランチ構造を持つモデルにおいては，複数層の寄与度が重複して反映されてしまう．そこで本論文では，ブランチ構造を持つモデルにおける逆伝播の計算方法を新たに提案する．そして，ブランチ構造を持つモデルに，逆伝播による説明生成手法を導入して拡張した，Layer-wise Relevance Propagation for Branch Networks (LRP-BN) を提案する．道路上のクラック有無を分類するモデルに対する視覚的説明を生成するタスクに焦点をあて，LRP-BN により理論的背景が明確で高品質な説明を生成する．実験の結果，提案手法は視覚的説明生成タスクにおける標準的な評価尺度である Insertion-Deletion Score においてベースライン手法を上回り，適切な視覚的説明の生成に成功することが示された．

1　はじめに

深層学習が幅広い分野に応用されている現代において，深層学習モデルの説明性は重要である [Shrikumar 17,Ribeiro 16]．例えば，理論が未解明な自然現象の予測に深層学習を用いた場合，視覚的説明による重要な部分の可視化を通して，理論の洞察を与えることができる．また，複雑な深層学習モデルにおいては，判断根拠を説明することが困難であり，誤った根拠をもとに分類しているかどうかを見分けることが難しい．この場合，クレバーハンス効果 [Pfungst 07] のように，モデルが本質的な特徴ではなく，無関係な特徴に基づいて分類を行い，汎化性能の低下をもたらす可能性がある．そのため，深層学習モデルの説明性を向上させることは有益である．

本論文では，モデルが分類結果を出力する過程に対する判断根拠の視覚的説明生成タスクを扱う．特に，道路上のクラック有無を分類するモデルに対して視覚的説明を生成するタスクに焦点をあてる．この視覚的説明は，道路上のクラックに対するマスクと考えることもできる．この場合，セグメンテーションのマスクが ground truth として与えられず，クラック有無のラベルアノテーションのみを用いてマスクを生成するため，本タスクは image-level weakly supervised semantic segmentation タスクとみなすことができる．

視覚的説明生成タスクは各モデルにおいて，本質的に重要な領域を正確に抽出する必要がある困難なタスクである．実際，人間が作成した道路上のクラックのマスクと，標準的な説明生成手法である GradCAM [Selvaraju 17] が生成した説明との IoU は 0.16 程度しか達成できていない．また，本タスクには明確な正解が存在することが少ないうえ，モデルの特徴や構造によって適切な説明生成手法は異なる．そのため，本タスクは正解マスクを利用せずに過不足なく適切な領域に注目する必要のある，難しいタスクである．図 1 に Road Damage Detection Dataset [Arya 22] の画像例を示す．まず，本

*連絡先：慶應義塾大学理工学部情報工学科
〒 223-8522 神奈川県横浜市港北区日吉 3-14-1
E-mail: tiida@keio.jp

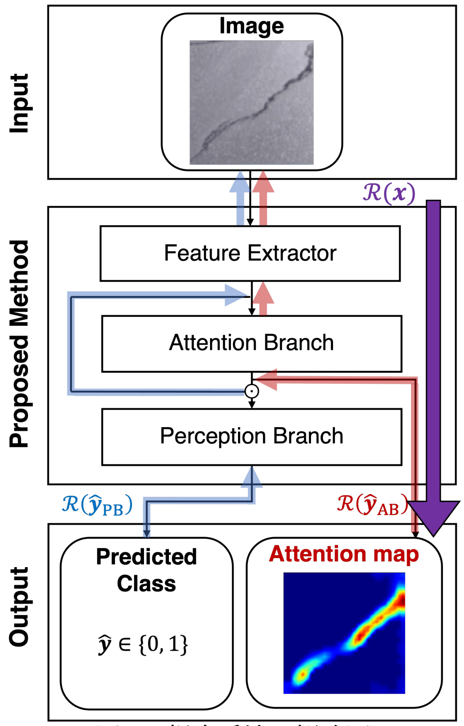


図 1: 提案手法の概略図

タスクにおいては図1中の Input に示すような道路画像をモデルに入力してクラックの有無を分類する．その過程で，図1中の Attention map のような道路上のクラック領域に注目した判断根拠の視覚的説明を生成することが望ましい．

畳み込みニューラルネットワークを基盤とするモデルにおいて，視覚的説明の生成に関する研究は数多く提案されている [Selvaraju 17, Petsiuk 18, Zhang 21a]. これらの手法は，既定の計算方法により説明を生成する．このような手法はモデルの構造に依存しないが，複雑なモデル構造に特化した説明の生成が難しく，不適切な領域に注目する場合がある．また，説明生成のための専用モジュールをブランチとして組み込んだ手法として，Attention Branch Network (ABN) [Fukui 19] や Lambda Attention Branch Networks [Iida 22] などが存在する．しかし，説明生成専用のモジュール自体がブラックボックスになってしまい，透明性が低い．Layer-wise Relevance Propagation (LRP) [Bach 15] は，出力からの逆伝播を利用して説明を生成する手法である．各層での逆伝播の計算方法が定義されているため，拡張性が高い．実際，LSTM や Transformer に対する逆伝播の計算方法も提案されている [Arras 17, Ali 22a]．しかし，これらの方法は cyclic connection を持たないモデルのみに適用されているため，ブランチ構造や skip connection などの cyclic connection を持つモデルへの適用は，新たな計算方法の定義が必要である．

このような背景から，本研究では，ブランチ構造を持つモデルにおいて標準的な説明生成手法である ABN に，説明生成の理論や計算過程が明瞭であり，高い透明性を有する LRP を導入して拡張する．これにより，理論的背景が明確で高品質な説明を生成する．既存研究との違いは，skip connection やブランチ構造を持つモデルにおける LRP の計算方法を新たに提案し，最も注目すべき領域を選択することで説明の品質を向上させる Choice 1 Component (C1C) を導入した点である．ブランチ構造や skip connection に対応した計算方法により，cyclic connection において，複数層の寄与度が重複して反映されてしまうことを防ぎ，適切な説明を生成することができる．また，最も注目度が高い画素を含む領域は，背景などの不適切な領域と連結していないことが多い．そのため，C1C により非連結な領域を除くことで背景を除去することができる．

本研究の独自性は以下の通りである．

- ブランチ構造や skip connection を持つモデルにおける LRP の計算方法を提案する．
- 生成した注目領域を元に，最も注目すべき領域を選択することで説明の品質を向上させる C1C を導入する．

2 関連研究

深層学習モデルの視覚的説明生成に関する研究は広く行われている [Bach 15, Selvaraju 17, Fukui 19, Ali 22a]．先行研究 [Das 20, Zhang 21b, Joshi 21, Ding 22] は，視覚的説明生成を含む深層学習モデルの説明生成に関して，包括的に調査し説明の生成方法ごとに各手法の分類・比較を行っている．視覚的説明生成タスクにおける標準的なデータセットとしては，ImageNet [Deng 09]，CIFAR10，CIFAR100 等の標準的な画像分類データセットが使用されている．

視覚的説明生成の手法は，その生成方法によって Back Propagation (BP)，Perturbation (PER) とその他に分類することができる．BP は逆伝播時の勾配に着目して説明を生成する．BP の手法として，LRP [Bach 15, Binder 16]，Grad-CAM [Selvaraju 17]，Integrated Gradients [Sundararajan 17]，[Chefer 21] 等がある．[Sundararajan 17] は，感度と実装不変の 2 つの公理を満たすように，勾配を積分して説明を生成する手法である．[Ismail 21] は，重要でない領域の勾配をゼロに近づけることでノイズを減らす Saliency Guided Training を考案した．[Bach 15] は，出力からの逆伝播を利用して説明を生成する手法である LRP の基盤となる計算方法を定義した．また，LSTM や Transformer に対する逆伝播の計算方法も提案されている [Arras 17, Ali 22a].

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手法を指す．PER に分類される手法として，LIME [Ribeiro 16]，Shapely Sampling [Lundberg 17]，RISE [Petsiuk 18] 等がある．例えば，[Petsiuk 18]

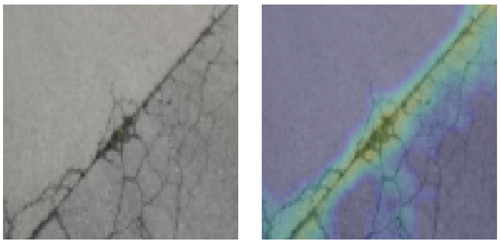


図 2: 視覚的説明生成タスクの例

は，マスクされた画像と出力の関係から説明を生成する手法である．

また，勾配や摂動以外から説明を生成する手法として ABN [Fukui 19]，IA-CNN [Zhang 21c]，IA-RED^2 [Pan 21] 等がある．ABN は，ブランチ構造として説明生成専用のモジュールを導入して説明を生成する拡張性が高い手法であり，Mask A3C [Itaya 21]，PonNet [Magassouba 21]，LABN [Iida 22] 等に応用されている．

サーベイ論文 [Cao 20, Ali 22b] は深層学習を用いた道路上のクラック検出タスクにおける各手法，標準データセット，標準評価尺度を包括的に紹介している．道路上のクラック検出には，Faster-RCNN [Ren 15] や SSD [Liu 16] 等の多くの物体検出モデルが応用されてきた [Yang 20,Yan 21]．[Yang 20] は，SSD に複数のカーネルサイズを持つ畳み込み層を含む Receptive Field を導入し，道路上のクラック検出に応用している．[Yan 21] は Deformable Convolution [Dai 17] を用いてクラックに沿った特徴抽出を行う Deformable SSD を提案している．道路上のクラック検出タスクにおける標準的なデータセットとしては RDD2022 Dataset [Arya 22] や Crack500 dataset [Yang 19] があげられる．

提案手法は説明生成専用のモジュール自体がブラックボックスである ABN とは異なり，ブランチ構造を持つモデルに透明性の高い LRP を導入する．また，cyclic connection に対応していない LRP とは異なり，skip connection やブランチ構造を持つモデルにおける LRP の計算方法を新たに提案する．

3　問題設定

本論文では，道路上のクラック有無分類タスクに対する判断根拠の視覚的説明生成を扱う．図 2 に道路上のクラック有無分類問題の例を示す．左図が入力であり，右図はモデルの注目領域を入力画像に重畳した画像である．本タスクでは，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．

本論文では，画像から道路上のクラックを検出できることを前提とする．標準的な道路上のクラック検出手法は [Arya 22] にあげられている．本論文における用語を以下のように定義する:

- **クラック領域:** 画像における道路上のクラックを示す領域

本タスクの入力と出力はそれぞれ画像 $\boldsymbol{x} \in \mathbb{R}^{c\times h\times w}$ と $\boldsymbol{x}$ がどのクラスに属するかの確率の予測値 $p(\hat{\boldsymbol{y}}) \in \mathbb{R}^C$ である．ここで，C, c, h, w はそれぞれクラス数，入力画像におけるチャンネル数，縦幅，横幅を表す．また，視覚的説明として画像中の各画素に重要度を割り当てた attention map $\boldsymbol{\alpha} \in \mathbb{R}^{h\times w}$ を利用する．

4　提案手法

提案手法はブランチ構造を持つ ABN [Fukui 19] に LRP [Bach 15] を導入して拡張した LRP for Branch Networks (LRP-BN) である．本論文においては，ABN をはじめとするブランチ構造を持つモデルに適用可能な LRP を扱う．本手法で行う拡張は，ブランチ構造を持つモデルにおける LRP の計算方法を定義したものである．そのため，ブランチ構造や cyclic connection をもつ手法一般に適用可能である．提案手法の新規性は以下の通りである．

- ブランチ構造や skip connection を持つモデルにおける LRP の計算方法を提案する．
- 生成した注目領域を元に，最も注目すべき領域を選択することで説明の品質を向上させる C1C を導入する．

4.1　モデル構造

図 3 に提案手法のモデル構造および入力における Relevance $\mathcal{R}$ の計算方法の概略を示す．提案手法は，Feature Extractor (FE)，Attention Branch (AB)，Perception Branch (PB) の 3 モジュールから構成される．

FE は，モデルの注目領域の生成および予測に用いる特徴抽出をするためのモジュールで，畳み込み層，Batch Normalization 層，Max Pooling 層と B 個の Bottleneck 層から構成される．f_{FE} の入力は $\boldsymbol{x}$ で，画像特徴量 $\boldsymbol{h} \in \mathbb{R}^{c_1\times h_1\times w_1}$ を出力する．ここで，c_1, h_1, w_1 はそれぞれ画像特徴量のチャンネル数，縦幅，横幅を表す．

AB は説明生成のための $f^{(1)}_{\mathrm{AB}}$ と，説明と分類を関連づけるための $f^{(2)}_{\mathrm{AB}}$ に分かれる．$f^{(1)}_{\mathrm{AB}}$ は Bottleneck 層，畳み込み層，Batch Normalization 層，Max Pooling 層から構成される．$f^{(1)}_{\mathrm{AB}}$ の入力は $\boldsymbol{h}$ であり，出力は $\tilde{\boldsymbol{\alpha}} \in \mathbb{R}^{w_1\times h_1}$ である．また，予測に重要でない領域を削除して PB に入力するために，$\tilde{\boldsymbol{\alpha}}$ のうち，ハイパーパラメータ $\theta_{\boldsymbol{\alpha}}$ より小さな値を 0 として $\boldsymbol{\alpha}' \in \mathbb{R}^{w_1\times h_1}$ とする．$f^{(2)}_{\mathrm{AB}}$ の入力は $\boldsymbol{h}$ であり，出力は attention loss を計算するための確率の予測値 $p(\hat{\boldsymbol{y}}_{\mathrm{AB}})$ である．$f^{(2)}_{\mathrm{AB}}$ は Bottleneck 層，畳み込み層，Batch Normalization 層，Max Pooling 層，Global Average Pooling 層から構成される．損失関数に $p(\hat{\boldsymbol{y}}_{\mathrm{AB}})$ を加えることで，AB 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，分類結果と関連する attention map を生成できる．

PB は $\boldsymbol{h}$ と $\boldsymbol{\alpha}$ の両方を用いて分類を行うモジュールである．PB は $N_{\mathrm{B}} - B$ 個の Bottleneck 層と全結合層か

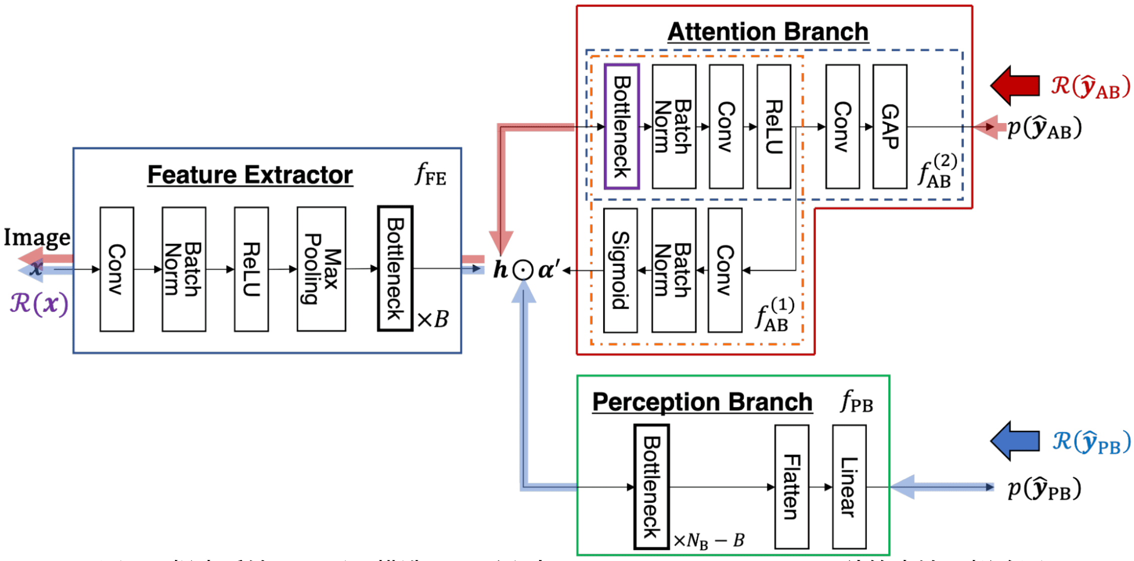


図 3: 提案手法のモデル構造および入力における Relevance $\mathcal{R}$ の計算方法の概略図

ら構成される．ここで，N_{B} はバックボーンネットワークの Bottleneck 層の数を表す．PB の入力は $\boldsymbol{\alpha}'\odot\boldsymbol{h}$ である．マスク処理をした $\boldsymbol{\alpha}'$ と $\boldsymbol{h}$ を掛け合わせることにより，予測に重要な領域を入力することができる．また，PB の出力はどのクラスに属するかの確率の予測値 $p(\hat{\boldsymbol{y}}_{\mathrm{PB}})$ である．

最終的なモデルの予測は以下の式で表される:

$$p(\hat{\boldsymbol{y}}_{\mathrm{PB}}) = f_{\mathrm{PB}}(\boldsymbol{\alpha}'\odot\boldsymbol{h}) \tag{1}$$

$$p(\hat{\boldsymbol{y}}_{\mathrm{AB}}) = f_{\mathrm{AB}}^{(2)}(\boldsymbol{h}) \tag{2}$$

$p(\hat{\boldsymbol{y}}_{\mathrm{PB}})$ はどのクラスに属するかの確率の予測値であり，分類の予測結果を出力するために利用する．また，$p(\hat{\boldsymbol{y}}_{\mathrm{AB}})$ は分類には直接用いないが，損失関数に導入することで説明の品質を向上させることができる．

4.2 LRP の計算方法

通常の構造，ブランチ構造，skip connection の 3 つの構造に分けて，提案手法における LRP の計算方法を説明する．

4.2.1 通常の構造における LRP の計算方法

通常の構造においては，標準的な LRP の z-rule [Bach 15, Binder 16] を適用して計算する．一例として，$\mathcal{R}(z)$ を z の Relevance とすると，Linear 層における LRP の計算は以下で表される:

$$\mathcal{R}\left(z_i^{(\mathrm{I})}\right) = \sum_j \frac{\mathrm{ReLU}\left(w_{ij}z_j^{(\mathrm{I})}\right)}{\sum_k \mathrm{ReLU}\left(w_{kj}z_j^{(\mathrm{I})}\right)}\mathcal{R}\left(z_i^{(\mathrm{O})}\right) \tag{3}$$

ここで，$z_i^{(\mathrm{I})}, z_i^{(\mathrm{O})}$ はそれぞれ Linear 層の入力，出力における i 番目の要素を，w_{ij} は Linear 層の重みにおける (i,j) 要素を表す．しかし，上記の計算方法は cyclic connection を含まないモデルに対して提案されており，ブランチ構造や skip connection などの cyclic connection をもつモデルには対応していない．そのため，本研究ではブランチ構造・skip connection における LRP の計算方法を提案する．

4.2.2 ブランチ構造における LRP の計算方法

本モデルにおけるブランチ構造では，$p(\hat{\boldsymbol{y}}_{\mathrm{AB}})$ と $p(\hat{\boldsymbol{y}}_{\mathrm{PB}})$ それぞれから 2 つの Relevance $\mathcal{R}_{\mathrm{AB}}, \mathcal{R}_{\mathrm{PB}}$ が計算される．そのため，通常の構造とは異なる方法で計算する必要がある．

まず，$\mathcal{R}_{\mathrm{AB}}$ の計算方法を考える．図 3 に示すように，$\mathcal{R}_{\mathrm{PB}}$ の入力に用いる $\boldsymbol{\alpha}'$ は $f_{\mathrm{AB}}^{(1)}$ を通じて計算される．そのため，$\mathcal{R}_{\mathrm{AB}}$ の計算方法として，$\mathcal{R}_{\mathrm{PB}}$ を利用する方法と，$\mathcal{R}_{\mathrm{PB}}$ とは独立に計算する方法が考えられる．ここで，$\boldsymbol{h}$ の Relenvance $\mathcal{R}(\boldsymbol{h})$ を計算する際，$\mathcal{R}_{\mathrm{AB}}$ と $\mathcal{R}_{\mathrm{PB}}$ の両方を利用すると，前者では $\mathcal{R}_{\mathrm{AB}}$ を介して $\mathcal{R}_{\mathrm{PB}}$ の影響が二重に反映される可能性がある．そのため，図 3 に赤・青の矢印で示すように，$\mathcal{R}_{\mathrm{PB}}$ とは独立に計算を行う．この計算方法において，ブランチ構造は $\boldsymbol{\alpha}'$ を用いたゲート構造とみなすことができ，LSTM における LRP [Arras 17] において，ゲート構造を独立に計算する方法と一致する．

次に，$\mathcal{R}(\boldsymbol{h})$ の計算方法に関しては，conservation [Bach 15] を考慮して $\mathcal{R}_{\mathrm{AB}}, \mathcal{R}_{\mathrm{PB}}$ の和を $\mathcal{R}(\boldsymbol{h})$ とする．AB の入力が $\boldsymbol{h}$ で，PB の入力が $\boldsymbol{\alpha}'\odot\boldsymbol{h}$ であるため，$1:\boldsymbol{\alpha}'$ の重みを付けた和も考えられる．しかし，$\boldsymbol{\alpha}'$ の寄与は forward 計算時に既に含まれているため [Arras 17]，重複して寄与を考慮しないために $\mathcal{R}_{\mathrm{AB}}$ と $\mathcal{R}_{\mathrm{PB}}$ の和と定義した．以上より，$\mathcal{R}(\boldsymbol{h})$ は以下の式で表される:

$$\mathcal{R}(\boldsymbol{h}) = \mathcal{R}_{\mathrm{AB}} + \mathcal{R}_{\mathrm{PB}} \tag{4}$$

4.2.3 Skip connection における LRP の計算方法

Residual connection [He 15] はサイクル構造を持つが，z-rule により計算すると skip connection の影響が

考慮されない．また，residual block と skip connection の和を出力とする点で，並列に計算したアダマール積を利用する ABN のブランチ構造とは異なる．そのため，skip connection を考慮した LRP を提案する．

まず，residual connection の入出力をそれぞれ $\boldsymbol{x}_s$, $\boldsymbol{y}_s$ と表し，residual block に z-rule を適用して計算した Relevance，出力の Relevance をそれぞれ $\mathcal{R}(\boldsymbol{x})^-$, $\mathcal{R}(\boldsymbol{y}_s)$ と表す．$\mathcal{R}(\boldsymbol{h})$ の議論と同様に，conservation を考慮すると $\boldsymbol{x}_s$ の Relevance $\mathcal{R}(\boldsymbol{x}_s)$ は $\mathcal{R}(\boldsymbol{x})^-$ と $\mathcal{R}(\boldsymbol{y}_s)$ の加重和で表すことができると考えられる:

$$\mathcal{R}(\boldsymbol{x}_s) = \gamma\mathcal{R}(\boldsymbol{x}_s)^- + (1-\gamma)\mathcal{R}(\boldsymbol{y}_s) \tag{5}$$

ここで，γ は $\mathcal{R}(\boldsymbol{x}_s)^-$ と $\mathcal{R}(\boldsymbol{y}_s)$ の比率である．γ は $\boldsymbol{x}_s$ と $\boldsymbol{y}_s$ を考慮して決定することもできるが，事前実験の結果良好な結果が得られたため，$\gamma = 0.5$ とした．

4.3 Relevance と attention map を用いた視覚的説明の計算方法

提案手法においては，$\mathcal{R}$ と $\boldsymbol{\alpha}'$ を組み合わせ，C1C を導入することで高品質な説明を生成する．本手法で説明として使用する $\boldsymbol{\alpha}$ の計算方法を以下で述べる．

$\mathcal{R}(\boldsymbol{h})$ から FE の入力に対する Relevance を計算することで $\boldsymbol{x}$ に対する Relevance $\mathcal{R}$ が得られる．既存の LRP と同様に，この $\mathcal{R}$ を説明として使用することも可能である．また，既存の ABN と同様に $\boldsymbol{\alpha}'$ も説明として使用できる．しかし，単一の説明生成手法を利用した場合，不適切な領域に注目した説明が生成されることがあり，その際に修正の余地がない．一方，本手法では，LRP と ABN の双方が強く注目した領域をより強調し，高品質な説明を得るために $\mathcal{R}$ と $\boldsymbol{\alpha}'$ のアダマール積を説明に利用する．続いて，背景等の不適切への注目を防ぐため，C1C により最も注目すべき領域を抽出して $\boldsymbol{\alpha}_{\mathrm{C1C}}$ を得る．C1C においては，$\mathcal{R} \odot \boldsymbol{\alpha}'$ を 28×28 に縮小して細かいノイズや不要な情報を削除した上で，注目度が高い画素を含む連結領域を抽出する．$\mathcal{R}$ は多くの場合クラックに最も注目しており，最も注目度が高い画素を含む領域は背景などの不適切な領域と連結していないことが多い．そのため，非連結な領域を除くことで背景を除去することができる．最後に，$\boldsymbol{\alpha}_{\mathrm{C1C}}$ を $w \times h$ に拡大して $\boldsymbol{\alpha}$ を得る．

また，損失関数として，以下を使用する:

$$\mathcal{L} = \mathrm{CE}(\hat{\boldsymbol{y}}_{\mathrm{LPB}}, \boldsymbol{y}) + \lambda\mathrm{CE}(\hat{\boldsymbol{y}}_{\mathrm{LAB}}, \boldsymbol{y}) \tag{6}$$

ここで，$\boldsymbol{y}$, CE,λ はそれぞれ正解ラベル，交差エントロピー誤差関数，損失関数の重みを表す．

5 実験

5.1 データセットと実験設定

本研究で扱う視覚的説明生成タスクのための標準データセットは我々の知る限り存在しない．視覚的説明生成タスクにおいては，教師なしセグメンテーションタスクへの応用が可能で，データ数が十分であることが望ましい．そのため，セグメンテーションのマスクを含まず，人間によるマスク作成が可能でデータ数が十分な Road Damage Detection 2022 Dataset (RDD2022 Dataset) が最も適している．よって，RDD2022 Dataset の訓練集合から，画像選択・画像のクロップ・テスト集合作成の三段階の処理によって RDC Dataset を構築した．

表 1: 実験で用いた設定

エポック		300
バッチサイズ		64
学習率	Feature Extractor Linear	1.0×10^{-4}
	Attention Branch	1.0×10^{-3}
最適化		AdamW

RDC Dataset には，道路画像および，道路上のクラック有無が付与されたラベルが含まれている．RDD2022 Dataset には，日本・インド・チェコ・ノルウェー・アメリカ・中国の合計 6 カ国の道路画像が含まれる．このうち，中国以外の 5 カ国は車内から，中国はドローンやバイクから撮影された画像である．そのため，RDC Dataset においては画角の異なる中国を除いた 5 カ国を採用した．RDD2022 Dataset には，著者らが LabelImg と Computer Vision Annotation Tool によりアノテーションを行ったクラック領域情報が含まれている [Arya 20, Arya 22]．このアノテーションデータに含まれるクラック領域に基づき画像を切り出し，クラック有クラスの画像を作成した．また，クラック無クラスの画像はクラック領域を除いた領域からランダムに切り出して作成した．この際，クラック無クラスの切り出し後の縦・横幅はそれぞれアノテーションデータから計算したクラック領域の縦・横幅分布から選択した．最後に，縮小・標準化・二値化を行った後に画像間の XOR を計算することで類似画像を抽出し，類似画像のない画像群を選択してテスト集合を作成した．これらの処理により RDC Dataset を構築した．

RDC Dataset は日本・インド・チェコ・ノルウェー・アメリカの 5 カ国で撮影された道路画像を含む．クラック有クラスの画像は 47,513 枚，クラック無クラスの画像は 30,430 枚であった．訓練集合，検証集合，テスト集合はそれぞれ 66,641，7,405，3,897 サンプルを含む．本研究では，$\boldsymbol{x}$ を 224×224 にリサイズして，反転・回転・切り抜きによるデータ拡張を行った．訓練集合はモデルの学習に，検証集合はハイパーパラメータを調整するために使用した．また，テスト集合はモデルの性能評価に使用した．

表 1 に提案手法における設定を示す．提案手法のパラメータ数と積和演算数はそれぞれ 3200 万，92.1G で

表 2: 各手法における定量的結果

Method	Acc ↑	Insertion ↑	Deletion ↓	ID Score ↑
RISE	0.958 ± 0.004	0.373 ± 0.042	0.054 ± 0.027	0.319 ± 0.018
GradCAM	0.958 ± 0.004	0.635 ± 0.026	0.052 ± 0.011	0.583 ± 0.020
LRP	0.958 ± 0.004	0.528 ± 0.117	0.301 ± 0.111	0.227 ± 0.010
ABN	0.957 ± 0.004	0.358 ± 0.035	0.090 ± 0.013	0.268 ± 0.039
Ours	0.957 ± 0.004	$\mathbf{0.804 \pm 0.005}$	0.069 ± 0.006	$\mathbf{0.735 \pm 0.007}$

あった．訓練にはメモリ 11GB 搭載 GeForce RTX 2080 Ti， Intel Corei9 9900K および 64GB の RAM を用いて，モデルの訓練時間および 1 サンプルあたりの推論時間は，それぞれ 3 時間および 1.3×10^{-3} 秒であった．検証集合における損失関数の値が 4 回連続改善しなかった場合に早期終了を行った．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．

5.2 実験結果

ベースライン手法として，RISE [Petsiuk 18]，Grad-CAM [Selvaraju 17]，LRP [Bach 15]，ABN [Fukui 19] を使用した．ABN をベースライン手法とした理由は，バックボーンネットワークとして ResNet を用いており，ブランチ構造を有する最も標準的な手法のためである．同様に，RISE・GradCAM・LRP は汎用的なモデルに適用可能な手法の中で標準的であるためベースライン手法とした．

本実験における評価尺度には，Accuracy，Insertion Score，Deletion Score，Insertion-Deletion Score (ID Score) を用いた．また，最も標準的な ID Score を主要評価尺度とした．Accuracy は分類タスクにおけるモデルの標準的な評価尺度であり，Insertion score，Deletion score，ID score は説明生成タスクの標準的な評価尺度であるため使用した．

Insertion Score，Deletion Score は Insertion 曲線，Deletion 曲線の AUC で計算される．また，ID Score は Insertion Score と Deletion Score の差で定義される．ここで，Insertion 曲線，Deletion 曲線はそれぞれ $\boldsymbol{\alpha}$ を基に重要な領域を挿入，削除した際の予測の変化を表す．詳細は以下で定義する．

まず，$\boldsymbol{\alpha}$ の要素を降順に $\boldsymbol{\alpha}_{i_1,j_1}, \boldsymbol{\alpha}_{i_2,j_2}, \cdots, \boldsymbol{\alpha}_{i_w,i_h}$ として，集合 A_n, $\boldsymbol{i}_n$, $\boldsymbol{d}_n$ を次のように定義する．

$$A_n = \{(i_k, j_k) | k \leq n\} \tag{7}$$

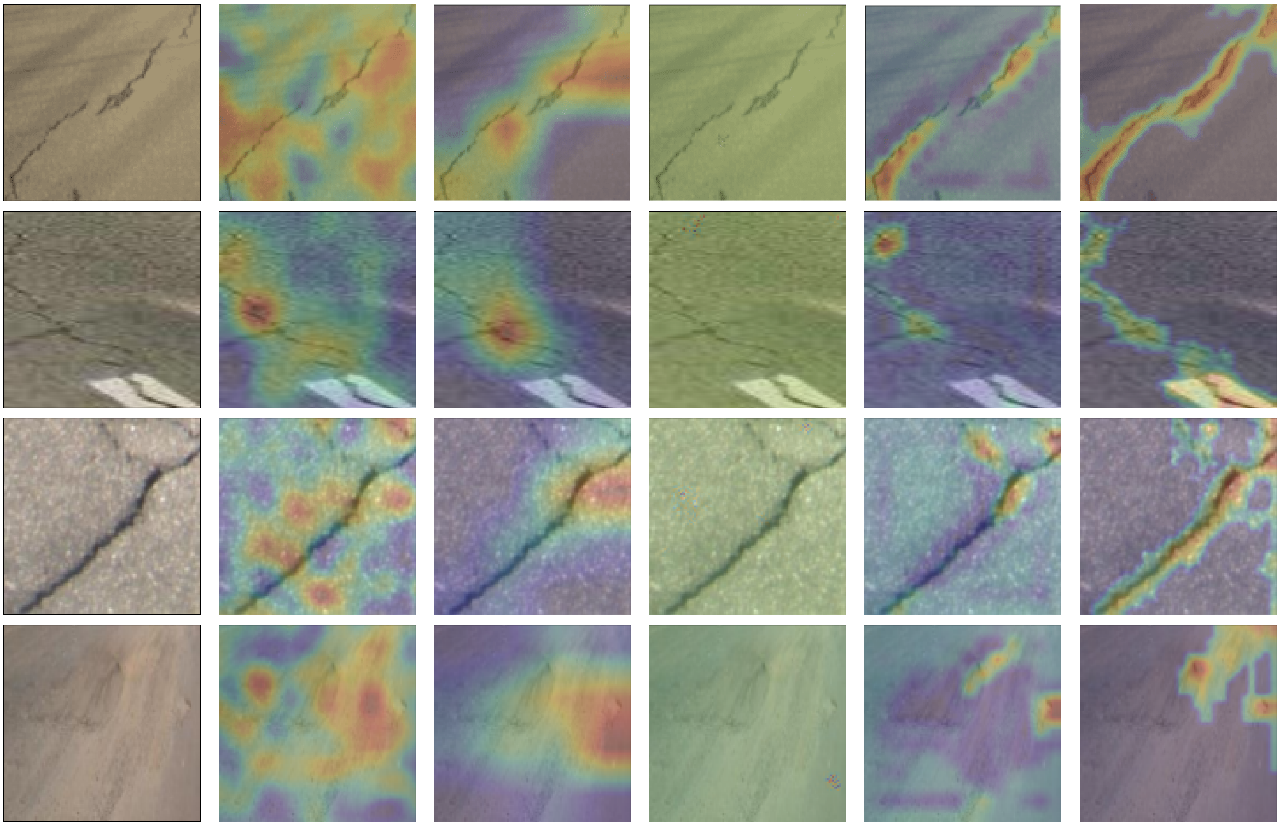
$$(\boldsymbol{i}_n, \boldsymbol{d}_n) = \begin{cases} (x_{ij}, 0), (i,j) \in A_n \\ (0, x_{ij}), (i,j) \notin A_n \end{cases} \tag{8}$$

ここで，n は挿入・削除するピクセル数を表す．$\boldsymbol{i}_n, \boldsymbol{d}_n$ をモデルに入力した際の出力をそれぞれ $\boldsymbol{y}^{(\mathrm{ins},n)}$, $\boldsymbol{y}^{(\mathrm{del},n)}$ とする．このとき，$\left(n, \boldsymbol{y}_{\mathcal{C}}^{(\mathrm{ins},n)}\right)$，$\left(n, \boldsymbol{y}_{\mathcal{C}}^{(\mathrm{del},n)}\right)$ をプロットした曲線が，Insertion 曲線，Deletion 曲線である．ここで，$\mathcal{C}$ は $\boldsymbol{x}$ が属するクラスを表す．

表 2 にベースライン手法と提案手法との比較に関する定量的結果を示す．各手法につき実験を 5 回行い，その平均値および標準偏差を示した．また，表 2 中の太字は，統計的に有意な最良値を表す．表 2 より，主要尺度である ID Score において，RISE，GradCAM，LRP，ABN，および提案手法はそれぞれ 0.319，0.583，0.227，0.268 および 0.735 であり，提案手法はベースラインの中で最も高い GradCAM と比較して 0.152 ポイント上回った．また，Accuracy においては RISE，GradCAM，LRP が 0.958，ABN と提案手法が 0.957 で同程度であった．主要尺度である ID スコアと Insertion スコアにおける性能差は統計有意であった（$p < 0.05$）．

図 4 に定性的結果を示す．(a) 列は元画像を示し，(c)-(e) 列はベースライン手法，(f) 列は提案手法によって生成した説明を元画像に重畳した結果を表す．図 4 の 1-3 行目は説明生成に成功した例で，4 行目は説明生成に失敗した例である．図 4(b) 列より，RISE によって生成された説明は道路上のクラックの周辺に注目領域を有するが，クラック以外の領域にも強く注目していた．また，(b)，(e) 列より GradCAM，ABN によって生成された説明が強く注目していたのはクラックのうちの一部分であった．(d) 列より LRP によって生成された説明は画像中のわずかな画素にのみ強く注目しており，ほとんどの領域の注目度が等しく不適切である．一方で，(f) 列より提案手法は道路上のクラック全体に詳細に注目しており，クラック以外の道路の注目度は低く，適切な説明を生成している．

図 4 の 4 行目に示した失敗例について，(c)，(e)，(f) 列より，GradCAM，ABN，提案手法により生成した説明は全て画像中の右側のクラックに強く注目しており，左側にあるクラックに注目できていない，また，(d) 列より LRP によって生成した説明は画像左下のクラックがない領域にのみ注目している．(b) 列より，RISE によって生成した説明は画像全体を注目している．しかし，中央左の注目度が低い領域にもクラックがあるため，全てのクラックを適切に注目できていない．上記より，全ての手法が道路上のクラックを過不足なく注目できていない．これは，道路が整備されておらず，クラックの無い道路とクラックの境界が曖昧になって



(a) Original　(b) RISE [Petsiuk 18]　(c) GradCAM [Selvaraju 17]　(d) LRP [Bach 15]　(e) ABN [Fukui 19]　(f) Ours

図 4: 各手法における定性的結果

表 3: 被験者が作成した正解マスクと各手法によって生成した説明との IoU 計測実験の結果

	RISE	GradCAM	LRP	ABN	Ours
IoU ↑	0.167 ± 0.004	0.141 ± 0.002	0.111 ± 0.000	0.113 ± 0.107	$\mathbf{0.184 \pm 0.004}$

おり，判別が難しいことが原因だと考えられる．

最後に，被験者実験として，人間が作成したクラックのマスクと，ベースライン手法および提案手法が生成した説明の IoU を計測した．まず，被験者 4 人がそれぞれ異なる 50 サンプルについてクラック領域を示したマスクを作成し，合計 200 サンプルのマスクを得た．これを正解マスクとして，表 3 に，正解マスクと各手法によって生成した説明との IoU を示す．各手法につき実験を 5 回行い，その平均値および標準偏差を示した．また，表 3 の太字は最良値を表す．表 3 より，IoU において，RISE，GradCAM，LRP，ABN，および提案手法はそれぞれ 0.167，0.141，0.111，0.113 および 0.184 であり，提案手法はベースライン手法の中で最も高い RISE と比較して 0.017 ポイント上回った．これらの結果より，提案手法が最も正解マスクと類似した説明を生成できていると示唆される．

6　おわりに

本論文では，道路上のクラック有無分類問題における判断根拠の視覚的説明生成を扱った．提案手法による貢献は以下である．

- Skip connection やブランチ構造を持つモデルにおける LRP の計算方法を提案した．
- 生成した注目領域を元に，最も注目すべき領域を選択することで説明の品質を向上させる C1C を導入した．
- 本タスクの標準的な評価尺度である Insertion Score，ID Score において，提案手法が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 ムーンショット，NEDO の助成を受けて実施されたものである．

参考文献

[Ali 22a] Ali, A., Schnake, T., Eberle, O., et al.: XAI for Transformers: Better Explanations through Conservative Propagation, in *ICML*, Vol. 162, pp. 435–451 (2022)

[Ali 22b] Ali, L., Alnajjar, F., Khan, W., Serhani, M. A., et al.: Bibliometric Analysis and Review of Deep Learning-Based Crack Detection Literature Published between 2010 and 2022, *Buildings*, Vol. 12, No. 4 (2022)

[Arras 17] Arras, L., Montavon, G., Müller, R., et al.: Explaining Recurrent Neural Network Predictions in Sentiment Analysis, in *WASSA*, pp. 159–168 (2017)

[Arya 20] Arya, D., Maeda, H., Kumar, S., Toshniwal, D., et al.: Global Road Damage Detection: State-of-the-art Solutions, in *Big Data*, pp. 5533–5539 (2020)

[Arya 22] Arya, D., Maeda, H., et al.: Crowdsensing-based Road Damage Detection Challenge (CRDDC ' 2022), in *Big Data*, pp. 6378–6386 (2022)

[Bach 15] Bach, S., Binder, A., Montavon, G., 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)

[Binder 16] Binder, A., et al.: Layer-Wise Relevance Propagation for Neural Networks with Local Renormalization Layers, in *ICANN*, pp. 63–71 (2016)

[Cao 20] Cao, M.-T., Tran, Q.-V., Nguyen, N.-M., et al.: Survey on Performance of Deep Learning Models for Detecting Road Damages Using Multiple Dashcam Image Resources, *Adv. Eng. Inform.*, Vol. 46, p. 101182 (2020)

[Chefer 21] Chefer, H., Gur, S., and Wolf, L.: Transformer Interpretability Beyond Attention Visualization, in *CVPR*, pp. 782–791 (2021)

[Dai 17] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., and Wei, Y.: Deformable Convolutional Networks, in *ICCV* (2017)

[Das 20] Das, A. and Rad, P.: Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey, *arXiv preprint arXiv:2006.11371* (2020)

[Deng 09] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database, in *CVPR*, pp. 248–255 (2009)

[Ding 22] Ding, W., Abdel, M., Hawash, H., and Ali, A.: Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey, *Inf. Sci.*, Vol. 615, pp. 238–292 (2022)

[Fukui 19] Fukui, H., Hirakawa, T., et al.: Attention Branch Network: Learning of Attention Mechanism for Visual Explanation, in *CVPR*, pp. 10705–10714 (2019)

[He 15] He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, in *CVPR*, pp. 770–778 (2015)

[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)

[Ismail 21] Ismail, A., Corrada, H., and Feizi, S.: Improving Deep Learning Interpretability by Saliency Guided Training, in *NeurIPS* (2021)

[Itaya 21] Itaya, H., et al.: Visual Explanation using Attention Mechanism in Actor-Critic-based Deep Reinforcement Learning, in *IJCNN*, pp. 1–10 (2021)

[Joshi 21] Joshi, G., Walambe, R., and Kotecha, K.: A Review on Explainability in Multimodal Deep Neural Nets, *IEEE Access*, Vol. 9, pp. 59800–59821 (2021)

[Liu 16] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, Y., and Berg, A.: SSD: Single Shot MultiBox Detector, in *ECCV*, pp. 21–37 (2016)

[Lundberg 17] Lundberg, S. and Lee, I.: A Unified Approach to Interpreting Model Predictions, in *NeurIPS*, pp. 4765–4774 (2017)

[Magassouba 21] Magassouba, A., Sugiura, K., et al.: Predicting and Attending to Damaging Collisions for Placing Everyday Objects in Photo-Realistic Simulations, *Advanced Robotics*, Vol. 35, No. 12, pp. 787–799 (2021)

[Pan 21] Pan, B., Panda, R., Jiang, Y., et al.: IA-RED^2: Interpretability-Aware Redundancy Reduction for Vision Transformers, in *NeurIPS* (2021)

[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)

[Pfungst 07] Pfungst, O.: *Das Pferd des Herrn von Osten: der kluge Hans. Ein Beitrag zur experimentellen Tier- und Menschen-Psychologie*, Barth (1907)

[Ren 15] Ren, S., He, K., Girshick, R., and Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, in *NIPS*, Vol. 28 (2015)

[Ribeiro 16] Ribeiro, M., Singh, S., et al.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in *KDD*, p. 1135–1144 (2016)

[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., et al.: Learning Important Features Through Propagating Activation Differences, in *PMLR*, Vol. 70, pp. 3145–3153 (2017)

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

[Yan 21] Yan, K., et al.: Automated Asphalt Highway Pavement Crack Detection Based on Deformable Single Shot Multi-Box Detector Under a Complex Environment, *IEEE Access*, Vol. 9, pp. 150925–150938 (2021)

[Yang 19] Yang, F., Zhang, L., Yu, S., Prokhorov, D. V., Mei, X., et al.: Feature Pyramid and Hierarchical Boosting Network for Pavement Crack Detection, *IEEE trans Intell Transp Syst.*, Vol. 21, pp. 1525–1535 (2019)

[Yang 20] Yang, J., Fu, Q., et al.: Road Crack Detection Using Deep Neural Network with Receptive Field Block, *Mater. Sci. Eng.*, Vol. 782, No. 4, p. 042033 (2020)

[Zhang 21a] Zhang, Q., et al.: Group-CAM: Group Score-Weighted Visual Explanations for Deep Convolutional Networks, *arXiv preprint arXiv:2103.13859* (2021)

[Zhang 21b] Zhang, Y., Tiňo, P., Leonardis, A., and Tang, K.: A Survey on Neural Network Interpretability, *TETCI*, Vol. 5, No. 5, pp. 726–742 (2021)

[Zhang 21c] Zhang, Z., Chen, Y., et al.: IA-CNN: A generalised interpretable convolutional neural network with attention mechanism, in *IJCNN*, pp. 1–8 (2021)