# 機械学習を用いたWebページスクリーンショットの分類における前処理方法の考察

堀江 傳貴　惣浜 英祐　蜂巣 吉成　吉田 敦　桑原 寛明

Web ページを利用者の目的に合わせて自動で分類することで，Web 検索が使いやすくなる．Web ページのスクリーンショットを機械学習の学習済みモデルを用いて分類したところ，一定の精度は得られた．さらに，適切な前処理を適用することで精度が向上する可能性がある．本研究では，機械学習を用いた Web ページスクリーンショットの分類における前処理方法の検討を行った．学習済みモデルの注目部分の可視化，スクリーンショットを 4 分割したものでの分類精度の確認，過学習の原因の考察から，(1)Canny 法，(2) 平滑化，(3) ヒストグラムの平坦化，(4) 中央部分以外除去，(5) 右側除去，(6) 頻出する Web ページのスクリーンショットをデータセットから外す，の 6 つの前処理方法を検討した．考察した前処理の方法を単体あるいは組み合わせて適用し，分類精度が向上するか確認した．

## 1　はじめに

現在，Web 利用者は検索エンジンを利用することで，膨大な数の Web ページの中から検索キーに関連した Web ページを得ることができる．一方で，一つのキーワードが複数の意味を持つことがあり，検索キーには関連しているが利用者の目的に合わない内容の Web ページが検索結果として表示されることもある．この場合，Web ページのタイトルからの推測や Web ページに一度目を通すなどして検索結果からさらに目的の Web ページを探す必要がある．検索結果で利用者の目的に応じた Web ページのみを表示するには，Web ページを目的に合わせて自動で分類すれば良い．しかし，利用者は自分の目的に合った Web ページかどうかの判断を Web ページを実際に見ることで判断できるが，この判断基準を明示的にルール化することは難しい．我々は深層学習を用いての分類が有効と考え，深層学習を用いた画像分類で，どの程度の精度が得られるかを調査する．具体的には次のように行う．

- 分類の対象を Web ページのスクリーンショットとする．これは，利用者が何らかの特徴を Web ページのファーストビューから捉え，判断しているからである．
- 分類には学習済みのモデルを複数使用し，その精度の違いを確認する．
- 分類の精度を高めるための前処理の方法を検討し，その効果を確認する．

一般に，画像の分類では学習モデルと前処理が重要である．学習モデルについては，様々なタスクに応用できるとされている学習済みモデルを用いることである程度の精度が得られる．前処理は分類対象に応じて適切な前処理を行うことで精度が向上する可能性がある．

本研究では，機械学習を用いた Web ページスクリーンショットの分類における，適切な前処理の方法について検討する．実験にあたっては，歴史上の人物

Data Pre-processing Examinations in Web Page Screenshot Classification Using Machine Learning.
Hiroki Horie,Eisuke Souhama, 南山大学大学院理工学研究科ソフトウェア工学専攻, Software Engineering Major, Graduate School of Science and Technology, Nanzan University.
Yoshinari HachisuAtsushi Yoshida, 南山大学理工学部ソフトウェア工学科, Dept. of Software Engineering, Faculty of Science and Technology, Nanzan University.
Hiroaki Kuwabara, 南山大学理工学部電子情報工学科, Dept. of Electronics and Communication Technology, Faculty of Science and Technology, Nanzan University.

を検索キーとし，利用者の目的を解説ページとした分類において，適切な前処理を加えることで，前処理を行っていないデータセットを用いて分類した結果より精度が向上するか，確認する．前処理の検討方法として,「Web ページの構造や特徴からの検討」,「前処理なしでの分類を行った学習済みモデルの注目部分の可視化からの考察」を行う．これらの方法から Web ページのスクリーンショットの分類に適した前処理方法について検討する．

## 2　関連研究

塩川らの研究 [5] は学術用語解説 Web ページに対して,「文章の分かり易さ」および「Web ページの見易さ」の観点から人手評定を行い，深層学習を用いて，学術用語解説 Web ページの自動評定を行っている.「文章の分かり易さ」では，HTML ファイルから HTML タグを除去し抽出したテキストを深層学習モデルに入力し，文章の分かり易さをふまえて全体評定を行っている.「Web ページの見易さ」では，Web ページを画像化し CNN に入力することにより，ResNet50 モデルを基盤の特徴抽出器として用いて，Web ページの見易さをふまえた全体評定を行っている．検索エンジンを用いて評価用 2 分野の各用語の検索を行った検索結果の上位 10 件以内の Web ページが対象である．本研究では，学術用語解説ページといった限定をせず，Web ページの分類を学習済みモデルを用いて行う．

## 3　学習済みモデルを用いた Web ページスクリーンショットの分類

本節では，3.1 節で本研究で扱う Web ページの分類の概要について，3.2 節では前処理なしの分類実験について記述する．

### 3.1　概要

本研究における Web ページの分類とは，Web 利用者が調べたい事柄に関する検索キーを利用して，検索を行った結果からさらに，利用者の目的に合った Web ページとそうでないものの 2 つに分けることである．例として，歴史上の人物名を検索した利用者の目的が，その歴史上の人物の解説ページ (図 1，図 2) を閲覧することだった場合，次の Web ページは目的に合わない．

- アニメやゲームに関連した Web ページ　(図 3)
- 通信販売サイト (図 4)
- ソーシャルネットワーキングサービス (図 5)
- 人物名が命名由来となったものの Web ページ (図 6)

図 1　解説ページ 1 †1

図 2　解説ページ 2 †2


図 3　アニメ関連サイト †3

図 4　通信販売ページ †4

図 5　ソーシャルネットワーキングサービス †5


図 6　人物名が命名由来となったものの Web ページ †6

本研究における最終的な目標は，図 1 から図 6 で挙げたような Web ページを，解説ページとそうでないページに 2 値分類することである．これは利用者が Web ページを見分けている方法に近い形，つまり，明示的なルール化が難しい特徴を基にした分類である．この分類のために，本研究では，Web ページのスクリーンショットを学習データとして使用する．分類を高精度で行うためには，タスクに合った前処理方法とモデルが必要である．モデルに関しては，様々なタスクに応用できる学習済みモデルを使用することができる．しかし，前処理方法は各タスクに適したものを用意しなければならない．犀川の研究 [4] では，適切な前処理方法を導入することで，平均精度を 12.2%向上させている．

### 3.2 実験

#### 3.2.1 データセットの概要

本研究では，検索キーと分類する目的をそれぞれ

- 検索キー　　：　歴史上の人物名
- 分類する目的　：　歴史上の人物の解説ページとそれ以外のページへの分類

とする．この条件で Google Chrome ブラウザで検索ページの 5 ページ分の Web ページにアクセスし，Web ページの標準表示サイズである $1000 \times 1000$(px) のサイズでスクリーンショットを収集し，手動でラベル付けを行い，データセットを作成する．分類の際には，スクリーンショット総数の 1403 枚のうち，80%を学習&検証データとして，残り 20%をテストデータとする．学習&検証データのうち，80%を学習データ，残り 20%を検証データとする．

†1 https://ja.wikipedia.org/wiki/
†2 https://www.tv-tokyo.co.jp/tohoho/back/040709.html
†3 https://bakumatsu.marv.jp/game/origin/character/toshizo.html
†4 https://www.amazon.co.jp/
†5 https://twitter.com/
†6 https://www.gousa.jp/destination/theodore-roosevelt-national-park

**表 1　前処理なしでの学習済みモデルによる分類結果**

| 使用モデル | 認識率 | 再現率 | 適合率 | F 値 |
|---|---|---|---|---|
| VGG-16 | 0.75 | 0.72 | 0.63 | 0.67 |
| ResNet50 | 0.62 | 0.53 | 0.63 | 0.55 |

#### 3.2.2 使用する学習済みモデルと分類精度の確認方法

本研究で使用する学習済みモデルは，大規模データセット ImageNet によって訓練され，様々なタスクに応用できる，VGG-16 [3] と ResNet50 [1] を用いる．VGG-16，ResNet50 ともに以下の処理を行う．

- 1000 値分類用の全結合層と出力層を切り離し，2 値分類用の出力層を接続
- 5 ブロック目以降の層を学習モードに設定し学習を行う，ファインチューニングを実施．

以上の処理により，接続した出力層と学習モードに設定した畳み込み層は，入力する学習データに対してパラメータの調整を行う．学習における各種パラメータについては，バッチサイズは 16，エポック数は 20，最適化手法は Adam，損失関数は binary cross entropy を採用した．本研究では，Python の深層学習ライブラリである，keras を用いて実装した．分類精度の確認は，分類を学習，検証，テストデータをランダムに組み替えて 10 回行い，認識率，再現率，適合率，F 値の平均を導出し確認する．

#### 3.2.3 分類結果

前処理を行っていないデータセットを用いて，スクリーンショットを用いた Web ページ分類を行い，前処理方法の検討と，検討した前処理方法の効果の確認を行う．スクリーンショットは学習済みモデルへの入力に合わせ，$224 \times 224$(px) にリサイズした．分類の結果を表 1 に示す． 前処理を行ってないデータセットでは高い精度で Web ページの分類を行うことはできないということが判明した．また，損失の上昇が見られたことから過学習が起こっているといえる．これらのことから，Web ページの分類に適した前処理方法を検討する必要がある．

## 4 Web ページスクリーンショットの前処理方法

本節では，4.1 節で検討した前処理を，4.2 節で前

処理方法を適用した実験結果を記述する．

### 4.1 前処理方法の検討

前処理方法の検討のために次のことを行った．

- Grad-CAM [2] による学習モデルの注目部分の可視化：学習時に注目している部分を確認し，その特徴を検討材料とする．
- スクリーンショットを 4 分割してそれぞれを分類：スクリーンショットを上下左右等分に 4 分割し，精度が他と比べて高かった部分に分類に必要な特徴が存在していると推定し検討材料とする．
- 過学習の原因の考察：原因から過学習の抑制方法を検討する．

可視化結果から分類精度の向上を妨げる原因として，次の 2 つを考察した．

- 比較的中央部分に注目しており，中央部分に分類に必要な特徴が集中している．スクリーンショットの端部分は分類にはあまり意味のない特徴が存在する．
- 空白部分には分類に必要な特徴があまり存在しないと考えたが，空白部分に注目することが多くあるので，これが分類に必要な特徴を捉えることの障害になっている．

4 分割しての分類した結果から認識率と F 値を見ると，左上と左下の方が右上と右下と比べて精度が若干ながら高いことから左側に分類に必要な特徴が多く存在するといえる．

以上の検討材料と，Web ページの特徴から次の前処理方法を検討した．

- エッジ検出
- 平滑化
- ヒストグラム平坦化
- 中央部分以外除去
- 右側除去
- 頻出ページをデータセットから外す

エッジ検出は物体の境界を検出する技術である．学習モデルの注目部分の可視化を行った結果，余白のような特徴がないと考えられる部分に注目している可視化結果が多くあった．これにより，Web ページの特徴を捉えきれていないので，分類精度が向上しなかったと考えた．スクリーンショットに対してエッジ検出手法である Canny 法と Laplacian 法を適用し，Web ページの特徴を際立たせる前処理方法を検討する．

平滑化はフィルタ内の画素の平均値で塗りつぶすことで元データをぼかし，ノイズの除去等を目的とする画像処理技術である．解説ページはテキスト，メニュー，ヘッダ，画像の構成で作成されている場合が多く，Web ページの構成要素を抽象化することで，学習モデルが構成を捉えやすくなり，分類精度の向上が見込めると考えた．

ヒストグラム平坦化は画像のヒストグラムを両側に向けて伸ばし，画像のコントラストの改善を行う技術である．Web ページのコントラスト調整を行うことで，学習の補助が見込める．また，ヒストグラムの平坦化は CNN を用いた画像認識タスクによく使用される前処理方法であり，多くの事例で精度の向上が見られることも，検討理由である．

中央部分以外除去はスクリーンショット中央部分のみを残し，それ以外の部分は除去する方法である．Web ページの左右端には，見やすくするための余白や，広告等が配置されていることも多く，Web ページを判断する主要な要素は，中央部分に集約されていると考えた．学習モデルの可視化を行った結果から，どちらの学習済みモデルも比較的中央部分を注視していたことから，Web ページ中央部分に分類に必要な特徴が集中しているといえる．

右側除去はスクリーンショット右側を除去する方法である．Web ページの特徴として右側部分には広告や，SNS などの外部サイトへのリンクといった，Web ページを判断する要素としては関係の薄いものが配置されていることが多いと考えた．スクリーンショットを 4 分割しての分類では，Web ページの左側の方が分類精度が少し高くなっており，Web ページ右側は分類に必要な特徴が少なめであると考えた．

頻出ページをデータセットから外す前処理は，前処理なしでの分類では過学習が起きていると考察したことから検討した．Wikipedia やコトバンクといった，検索結果にほぼ毎回ヒットする Web ページのスクリーンショットに過剰適合していると考えられる．

前処理方法の組み合わせについても検討する．1 つ

目の組み合わせとして，ヒストグラム平坦化とエッジ検出の組み合わせを挙げる．画像のコントラスト調整によりスクリーンショットの明るさを平均化することで，より正確なエッジ検出が見込めると考えた．

2 つ目の組み合わせとして，1 つ目の組み合わせに加え，中央部分以外除去，または右側除去の組み合わせを挙げる．より正確なエッジ検出によって，特徴を際立たせた後，特徴の少ない部分を除去することで，学習モデルがより特徴を捉えやすくなる効果が見込めると考えた．

### 4.2 検討した前処理方法の実験

4 節で述べた前処理方法を実際に適用し，前処理方法適用前の分類結果と前処理方法適用後の結果を比較し精度が向上するか確認する．表記の簡略化のため，本研究では Canny 法を Cn，Laplacian 法を Lap，平滑化を Sm，ヒストグラム平坦化を Equ，中央部分以外除去を Celim，右側除去を Relim，頻出する Web ページのスクリーンショットを外すを Drm と表記する．データセットに検討した各前処理方法を適用し，2 つの学習済みモデル VGG-16，ResNet50 で分類を行い，前処理なしでの分類から精度が向上する前処理方法が存在するか確認する．認識率，F 値が前処理方法適用前と比べて上昇し，損失の上昇が抑制できていれば有効な前処理方法と判断する．

各前処理を単体及び組み合わせて適用し分類を行い，VGG-16 での結果を表 2，ResNet50 での結果を表 3 にそれぞれ示す．

表 2，表 3 から，認識率，F 値が向上している結果は存在しない．また，本稿で挙げていない前処理方法の組み合わせも適用したが認識率，F 値が向上している結果は見られなかった．

## 5 考察

本研究で検討した前処理方法では，認識率，F 値が向上し，損失の上昇を抑制できる有効な前処理とはならなかった．検討した前処理方法が有効とならなかった原因として次のものを挙げる．

- 分類に必要な特徴を失っている
- データ数の減少

**表 2　VGG-16 における各前処理方法の適用結果**

| 前処理 | 認識率 | 再現率 | 適合率 | F 値 |
|---|---|---|---|---|
| 前処理なし | 0.75 | 0.72 | 0.63 | 0.67 |
| Cn | 0.69 | 0.78 | 0.28 | 0.39 |
| Lap | 0.69 | 0.67 | 0.48 | 0.51 |
| Sm | 0.75 | 0.69 | 0.67 | 0.68 |
| Equ | 0.74 | 0.67 | 0.71 | 0.68 |
| Celim | 0.73 | 0.71 | 0.51 | 0.59 |
| Relim | 0.73 | 0.71 | 0.48 | 0.57 |
| Drm | 0.72 | 0.64 | 0.54 | 0.58 |
| Equ Cn | 0.71 | 0.72 | 0.63 | 0.67 |
| Equ Lap | 0.70 | 0.74 | 0.51 | 0.43 |
| Equ Cn Celim | 0.70 | 0.74 | 0.40 | 0.49 |
| Equ Cn Relim | 0.73 | 0.74 | 0.51 | 0.59 |
| Equ Lap Celim | 0.66 | 0.61 | 0.42 | 0.48 |
| Equ Lap Relim | 0.69 | 0.65 | 0.58 | 0.57 |

**表 3　Resnet50 における各前処理方法の適用結果**

| 前処理 | 認識率 | 再現率 | 適合率 | F 値 |
|---|---|---|---|---|
| 前処理なし | 0.61 | 0.53 | 0.63 | 0.55 |
| Cn | 0.63 | 0.59 | 0.58 | 0.50 |
| Lap | 0.64 | 0.62 | 0.35 | 0.36 |
| Sm | 0.63 | 0.64 | 0.17 | 0.24 |
| Celim | 0.61 | 0.52 | 0.58 | 0.51 |
| Relim | 0.60 | 0.52 | 0.51 | 0.44 |
| Drm | 0.62 | 0.64 | 0.57 | 0.51 |
| Equ Cn | 0.62 | 0.57 | 0.58 | 0.55 |
| Equ Lap | 0.66 | 0.67 | 0.38 | 0.43 |
| Equ Cn Celim | 0.51 | 0.45 | 0.87 | 0.58 |
| Equ Cn Relim | 0.63 | 0.56 | 0.63 | 0.55 |
| Equ Lap Celim | 0.66 | 0.61 | 0.42 | 0.48 |
| Equ Lap Relim | 0.60 | 0.54 | 0.70 | 0.56 |

Canny 法と Laplacian 法は空白部分，中央部分以外除去，右側除去は端部分に分類に必要な特徴があまり存在しないと推定し検討した前処理である．しかし，前処理方法を適用した結果，精度が下がっていることから，これらの部分にも分類に必要な特徴が存在していると推測できる．改善点としてソーベル法などの別のエッジ検出手法を用いる，スクリーンショットの一部分除去に関しては，分類に必要のない特徴をより詳細に特定し，除去することが挙げられる．分類に必要のない特徴としては広告が挙げられる．広告を詳細に特定するには，より多くのデータを集め，広告存在がする傾向が強い部分を推定し除去する方法と，広告の位置を自動で特定し除去する方法を挙げる．後者の方法については，深層学習を用いた画像抽出技術を使用し，広告の位置を特定して，そこを黒色などで

**表 4　画像の占める割合の多さを基準とした分類の結果**

| 使用したモデル | 認識率 | 再現率 | 適合率 | F 値 |
|---|---|---|---|---|
| VGG-16 | 0.90 | 0.89 | 0.96 | 0.92 |
| ResNet50 | 0.78 | 0.89 | 0.72 | 0.78 |

塗りつぶすことで除去する方法がある．

平滑化，ヒストグラム平坦化は，認識率，F 値どちらも上昇している結果がないことから，前処理方法の適用によって情報が失われているといえる．改善点としては，本研究では挙げていない前処理方法との組み合わせを試すことが挙げられる．

検索結果に頻出する Web ページのスクリーンショットが過学習の原因と考え，データセットから外したが，結果的に過学習を抑制することはできなかった．全体のデータ数が減少したことで，別のデータに過適合を起こし，過学習が抑えられていないといえる．本研究で用いたデータセットはデータ数 1403 枚であるので，データ数の減少による過学習の発生が起こりやすいといえる．改善点としては，データの追加が挙げられる．総データ数を増加させることで，データの減少に伴う過学習を抑制させることが期待できる．

前処理方法の組み合わせはエッジ検出の効果向上を狙ったものであったが，分類に必要な特徴を無くしているといえる．前処理方法の組み合わせの中でも平滑化，Canny 法中央部分以外除去の組み合わせは適合率が大幅に上昇している．しかし，適合率が上がったことで，再現率が低下している．

Web ページのスクリーンショットを利用した深層学習による分類では，Web ページを分類する目的を，歴史上の人物の解説ページとそうでないページでは，高い精度で分類することはできなかった．Web ページのスクリーンショットを利用した分類自体が可能であるかを確認するために，Web ページを分類する目的を，スクリーンショットにおいて画像が占める割合が 3 割を超えているかいないかに変更して，分類を行った結果を表 4 に示す． この分類目的の場合，VGG-16 による分類では特に高い精度で分類が行えており，スクリーンショットを用いた深層学習による Web ページの分類自体は可能であるといえる．解説ページとそうでないページの分類がうまくいかなかった点から，スクリーンショットを用いた分類には向いている分類目的と，向いていない分類目的があるといえる．

## 6　おわりに

本研究では，機械学習を用いた Web ページスクリーンショットの分類における前処理方法について考察した．前処理なしでの分類を行った学習モデルの注目部分の可視化，スクリーンショットを 4 分割したものでの分類精度の確認，過学習の原因の考察から，分類精度が向上すると見込まれる前処理方法を検討した．検討した前処理方法を単体あるいは組み合わせて適用し，分類精度が向上するか確認した．検討した前処理方法では分類精度の向上が見られなかったので，前処理方法の改善点の考察と，前処理以外の分類精度を向上させる方法の考察を行った．今後の課題として，前処理方法の改善や，異なるアプローチの前処理方法を検討することが挙げられる．分類精度の向上のためには，前処理方法だけではなく，データの追加やモデルの調整，テキストベースによる分類との比較，統合等を行うことも必要である．

## 参 考 文 献


[1] He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[2] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[3] Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *CoRR*, Vol. abs/1409.1556(2015).

[4] 犀川巧: 実用的な画像に基づいた植物診断に向けた過学習抑制のための前処理, 法政大学大学院紀要. 理工学・工学研究科編, Vol. 61, 法政大学大学院理工学研究科, 2020, pp. 1–7.

[5] 塩川隼人, 春日孝秀, 韓炳材, 宇津呂武仁, 河田容英: 深層学習を用いた学術用語解説ウェブページの分かり易さ・見易さの自動評定, *DEIM Forum 2019I2-4*, 2019.