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の検討

2018SE018 堀江 傳貴　2018SE082 惣浜 英祐
指導教員：蜂巣 吉成

1　はじめに

現在，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 ページが検索結果に表示されることも多く，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ために，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．Web ページの分類は深層学習を用いて分類を行う．本研究では，利用者が Web ページを見分けている方法に近い形で分類するために，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．この学習データを画像認識タスクに優れた CNN を用いて分類を行う．

本研究における Web ページの分類では，利用者は何らかの特徴を Web ページのファーストビューから捉え判断していることから，Web ページのファーストビューのスクリーンショットと機械学習を用いて行う．歴史上の人物を検索キーとし，利用者の目的を解説ページとして Web ページのスクリーンショットを学習済みモデルで分類する．この分類では一定の精度は得られたが，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ができるが，前処理は分類対象で異なるので，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，適切な前処理について検討する．歴史上の人物を検索キーとし，利用者の目的を解説ページとした分類における，適切な前処理を加えることで，最低限の前処理を行った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ことを目指す．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検討」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について検討する．

2　関連研究

塩川らの研究 [1] は学術用語解説ウェブページに対して，「文章の分かり易さ」および「ウェブページの見易さ」の観点から人手評定を行い，深層学習を用いて，学術用語解説ウェブ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．「ウェブページの見易さ」では，ウェブページを画像化し CNN に入力することにより，ResNet-50 モデルを基盤の特徴抽出器として用いて，ウェブ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 件以内のウェブページが対象である．

3　学習済みモデルを用いた Web ページの分類

3.1　スクリーンショットを用いた Web ページ分類概要

本研究における Web ページの分類とは，利用者が調べたい事柄に関する検索キーを利用して，検索を行った結果として得られた Web ページの中からさらに，利用者の目的に合った Web ページと目的に合っていない Web ページの 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 1a，図 1b) を閲覧することだった場合，解説ページではないと判断する Web ページは，次のような Web ページである．Web ページにはそれぞれ違う特徴があるので，その特徴から分類を行う．

- アニメやゲームに関連した Web ページ　(図 1c)
- 通信販売サイト (図 1d)
- ソーシャルネットワーキングサービス (図 1e)
- 人物名が命名由来となったものの Web ページ (図 1f)

本研究における最終的な目標は，図 1 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とされてい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 [2] では，適切な

前処理を導入することで，平均精度を 12.2% 向上させている．

(a) 解説ページ 1 *1

(b) 解説ページ 2 *2

(c) アニメ関連ページ *3

(d) 通信販売サイト *4

(e) ソーシャルネットワーキングサービス *5

(f) 人物名が命名由来となったものの Web ページ *6

図 1: Web ページの分類の具体例

3.2 スクリーンショットを用いた分類の実験環境

3.2.1 データセットの作成

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 × 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．

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とされている，VGG-16[3] と ResNet50[4] を用いる．使用する学習済みモデルである 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．

4 スクリーンショットを用いた Web ページ分類の前処理の考察

4.1 前処理なしでの分類

前処理の検討と，検討した前処理の効果の確認のために，最低限の前処理を行った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 × 224(px) にリサイズした．分類の結果を表 1 に示す．

表 1: 前処理なしでの学習済みモデルによる分類の結果

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

最低限の前処理を行った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．このことから，Web ページの分類に適した前処理を検討する必要がある．

4.2 前処理の検討

前処理の検討のために次のことを行った．

- Grad-CAM[5] 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．

*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

- 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．

これらの検討材料から次の考察を行った．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若干ながら多く存在すると考察した．

以上の検討材料と，Web ページの特徴から次の前処理を検討した．

1 つ目に，エッジ検出を挙げる．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を検討する．

2 つ目に，平滑化を挙げる．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 ページの構成要素の抽象化をすることで，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．

3 つ目に，意味の少ない特徴を除去する前処理として，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を挙げる．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ことがわかる．

4 つ目に，意味の少ない特徴を除去する前処理として，スクリーンショット右側を除去する方法を挙げる．Web ページ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．

5 つ目に，過学習を抑制するために，検索結果に頻出する 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ことを挙げる．実験 1 では，ResNet50 での分類において過学習が起きていると考察した．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た．

5 検討した前処理の実験

検討した前処理を実際に適用し，前処理適用前の分類結果と前処理適用後の結果を比較し精度が向上するかどうかを確認する．表記の簡略化のため，本研究では Canny 法を Cm，Laplacian 法を Lap，平滑化を Sm，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が存在するか確認する．認識率，F 値が前処理適用前と比べて上昇しており，損失の上昇が抑制できていれば有効な前処理と判断する．

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: VGG-16 における各前処理の適用結果

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
Cm	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
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
Celim Cm	0.66	0.68	0.46	0.45
Relim Cm	0.70	0.72	0.42	0.48
Celim Drm	0.72	0.64	0.53	0.57
Relim Drm	0.71	0.64	0.48	0.54
Cm Drm	0.68	0.60	0.31	0.38
Celim Cm Drm	0.68	0.66	0.30	0.38
Relim Cm Drm	0.66	0.57	0.27	0.32

6 実験結果からの考察

本研究で検討した前処理では，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無くしてしまっている
- データ数の減少

Canny 法とラプラシアン法は空白部分，中央部分以外除

表 3: Resnet50 における各前処理の適用結果

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
Cm	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
Celim Cm	0.58	0.51	0.75	0.57
Relim Cm	0.61	0.72	0.42	0.48
Celim Drm	0.57	0.45	0.66	0.52
Relim Drm	0.62	0.49	0.62	0.53
Cm Drm	0.58	0.44	0.58	0.47
Celim Cm Drm	0.48	0.66	0.85	0.54
Relim Cm Drm	0.57	0.49	0.55	0.42

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，前処理を適用した結果，精度が下がってしまっていることから，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筆頭に挙げられる．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，そこを黒色などで塗りつぶすことで除去するといった方法がある．前者の方法と比べて正確な広告の除去が見込める一方，技術的な課題が多数存在することが欠点である．

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．データを外した結果，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てしまい，過学習が抑えられていないと考えた．特に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かったといえる．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．

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，分類を行った結果を表 4 に示す．

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

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言える．解説ページとそうでないページの分類があまり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言える．

7 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の検討を行った．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，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を検討した．検討した前処理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．検討した前処理では分類精度の向上が見られなかったため，前処理の改善点の考察と，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，前処理の改善や，異なるアプローチの前処理を検討することが必要である．また，分類精度の向上のためには，前処理部分だけではなく，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] 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．Forum 2019 I2-4 (2019).

[2] 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(2020).

[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F.-F. Li.:ImageNet large scale visual recognition challenge. CoRR, Vol. abs/1409.0575, (2014).

[4] K. He, X. Zhang, S. Ren, and J. Sun.:Deep residual learning for image recognition. In Proc. CVPR, pp. 770-778, (2016).

[5] 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, Proc. ICCV, pp. 618-626 (2017).