AI 画像認識システムを用いて植物の種を同定するシステムの構築

所属：　島根大学生物資源科学部

助成対象者：秋廣高志

共同研究者：白井匡人（島根大学総合理工学部）

概要

博物館や大学の収蔵庫には多くの植物標本が保管されている。最近になってそれらの標本がスキャナーやカメラを使ってデジタル化されている。本研究では、これらの画像約 48 万枚と AI 画像認識システムを用いて、植物の種名を自動で同定するシステムの構築を行った。80%の画像を学習用に、20%の画像をテスト用に用いて、計 9 回の実験を行い、2,179 種の植物を同定するシステムを構築した。コンピューターは上位 5 種の植物種名を予想するが、Top1 になる可能性は 95%であり、Top5 以内に入る確率は 99%であった。2,000 種を超える植物の同定システムはこれまで報告が無く、正解率が 95%を超えるシステムを構築した報告もない。

Abstract

Many plants are stored in museums and University herbarium. These specimens recently have been digitized using a scanner or camera. The current study is focusing on developing an automated identification system. A total of 480,000 images have been collected from 2,179 species of plants. Nine experiments were conducted using 80 % of these images. And 20% of images were used to confirm the efficiency of the identification. The system was able to identify 2,179 species of plants and predict the top one plant species with 95 % and top five plant species with 99 %. This method is a novel technique for automated identification system.

研究内容

【背景】

島根県はどんな植物がどこにどれだけ生育しているかを記載した「植物誌」が存在しない唯一の県である。植物誌が無い状況は、県内の植物の生物多様性を保護しようと考えた場合、どこのどの種を保護したら良いのかがわからないという、危機的な状態であるといえる。植物誌を作るためには、植物の種の同定ができる専門家や植物の採取に協力してくれるボランティアが必要であるが、島根県には5人程度しか種を同定できる専門家がおらず、今後もその数が増える可能性は低い。我々は2010年から、島根県立三瓶自然館サヒメルに収蔵されている植物標本（約6万点）の整理を開始し、県内のどこにどの植物が生えていたかを把握する作業を開始した。また同時に植物標本をデジタルスキャナーでスキャンし、標本画像を取得し、これをインターネットを介して公開する作業を行って来た。これまでに 41,242 点の標本のデジタル化に成功し、デジタル標本館にて公開している（ http://tayousei.life.shimane-u.ac.jp/ ）。これに加えて、鳥取県立博物館、陸前高田市博物館（岩手県）、福島大学共生システム理工学類生物標本室、鹿児島大学総合研究博物館の標本（合計 244,669 点）を公開している（引用 1)。これは世界で第5位の標本数で、国内では第2位の数である。世界では、中国科学院植物研究所が運営する中国デジタル植物標本館が、約600万点の画像データを取得しており、ヨーロッパではデジタル化作業を半自動化した装置も開発されており、今後多くの標本のデジタル化が進められるものと考えられている。

我々は、デジタル化した標本画像を用いて、島根県内に生育する植物のうち、イネ・カヤツリ科およびシダを除く植物を、葉および茎・枝の特徴によって分類し、植物種を同定するデジタル情報システム(iPis)の構築を行い、2016年にこれを完成させている（引用 2)。2018年には福島県内の植物を同定するシステムも作成している。どちらのシステムも一般市民でも種の同定ができるよう絵や説明が加えられているが、ある程度植物に関する知識がないと正しく種を絞り込めないことが明らかになっている。最近になり AI を用いた画像認識の技術開発が進んでおり、これを使って種の同定が行えないかを調査した。100 種について調査したところ、平均で 91%の判定精度（最低で 47%、最高で 100%）であった。2017年コスタリカの研究グループによって同様の研究報告が報告されているが、その判定精度はおよそ 84%であった（引用 3)。我々がこれまでに行った解析は、植物の分類学で用いられている一般的な分類方法を使っていないにも関わらず、平均で 91%と高い判定精度

を示しており、今後は細かく解析のパラメータを設定することでさらに判定精度が上がるものと考えられる。

【目的】

本研究では、以下の二点の解明を行った。

1. 現在判定精度が平均 91%となっているが、9%の誤判別がなぜ起こるのかを明らかにする。
2. 現在、判定が可能な種の数は約 100 種であるが、これを 2,000 種程度に引き上げ、判定精度がどのように変化するかを明らかにする。

最終的にすべての種において 95%以上の判定精度を持った検索システムを構築する。

【結果】

判定精度が 91%となっている原因を調査した。学習に使用している画像の数が少ない種ほど、正解率が低い傾向があることが明らかとなった。そこで、学習の用いる画像を増やすことを行った。本研究を始める前に保持していた画像の数は約 25 万枚であり、これに、国立科学博物館、兵庫県立人と自然の博物館、台湾大学博物館、東京都植物誌から、標本画像を取得し、合計で約 72 万枚の画像を取得した。集めた画像を種毎に集計すると、50 枚以上の画像がある種は 2179 種あることがわかった。全画像の 80%を学習に利用し、20%をテストに利用し、実験を 9 回行った。画像の入力サイズは 299×299 のカラー画像であり、各ピクセルは[0,0,0]～[255,255,255]の値で表される。全結合層の活性化関数には ReLU を用いた。出力層の活性化関数は Softmax 関数を用いた。損失関数にはクロスエントロピーを用いた。最適化関数には Adam を用い、学習率を 0.0001 とした。バッチサイズは 32 とし、選択されたデータには水平移動、垂直移動、水平反転の処理を加えるデータ拡張を起こった。epoch 数は 100 とした。全部で 9 回の実験を行い、最終的に accuracy が 0.719 から 0.957、Macro average が 0.661 から 0.913、Weighted average が 0.719 から 0.957 まで上昇させることができた。1 回目から 9 回目までに変更した点は以下の 3 点である。1. 今回収集した植物標本の中には、虫に食われたり、葉が破れて落ちてしまっている標本が含まれていので、これらを除去した。2. 写真で撮られている画像データとカメラで撮影されている画像データが混ざっていたが、それらは解像度や背景の色などが異なっており、それが原因で解析精度が下がっていることが明らかとなった。そこで、どちらの画像もできるだけ多く集め、できるだけ均一に混ざるようにした。3. 植物標本につけられている和名は専門家もしくは学生により同定されてつけられており、その同定が

間違っている可能性がある（少なくとも 10%程度間違っている可能性があると考えられている）。そこで、8 回の解析で 2 回以上 AI が間違えた画像（約 18,000 枚）を解析の対象から外した。これらの工夫をすることで、正解率が 95%程度になった（図 1）。AI は同定する際に候補を 5 種選び、それらの精度を算出するが、Top3 に正解が入る確率は 99.2%で、Top5 に入る確率が 99.7%であった。

	1回目	2回目	3回目	4回目	5回目	6回目	7回目	8回目	9回目
accuracy	0.719	0.868	0.942	0.93	0.888	0.934	0.901	0.942	0.957
macro avg	0.661	0.83	0.865	0.853	0.843	0.881	0.882	0.897	0.913
weighted avg	0.716	0.867	0.942	0.93	0.889	0.934	0.908	0.942	0.957
種数	2749	1735	1700	2008	2008	2238	2029	2030	2179
標本点数	420419	257598	318633	410374	407549	466471	484033	483311	489625

図 1　合計 9 回の解析の結果。Macro と Weighted avg は f1-score を示す。

勾配加重クラス活性化マッピング（Grad-CAM）手法を行う事で、AI が標本のどの部分を利用して同定したかを知ることができる。これらのデータを解析したところ、専門家が同定する際に見る部分と同じ部分を使って判定をしている種もあれば、専門家が見る部分とは違う部分を使って判定している種があることが分かった（図 2 左）。同定を間違えた約 5,000 枚の画像について解析したところ、画像の中に竹の定規が入っているものがあるが、その定規が植物であるために同定してしまっていることがあることが分かった（図 2 右）。

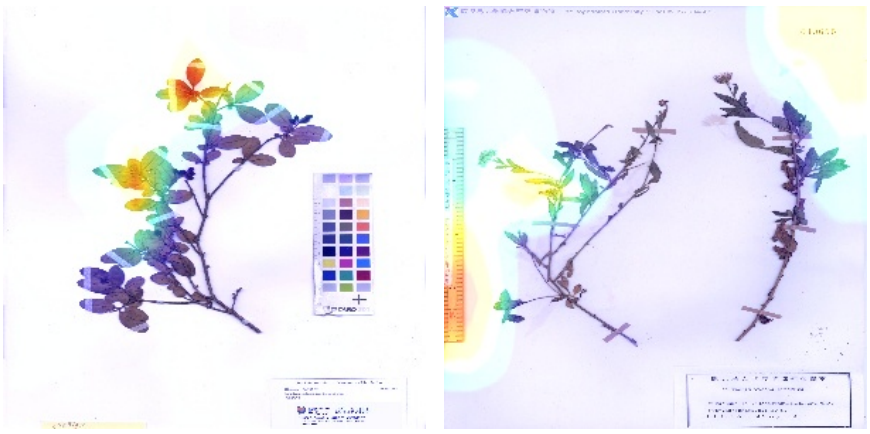

図 2　勾配加重クラス活性化マッピング（Grad-CAM）手法を用いた解析結果

解析結果を AI が出す場合は、AI が結果の予想正解率を算出するが、99.999%正解であると算出された 105 点の標本について調査してところ、105 点中 54 点は AI が間違っており、37 点が AI が正解であり、16 点が同定できないことが明らかとなった。本システムを用いることで、約 3 割の確率で同定が間違っている標本を発見できることが明らかとなった。

【今後】

今回構築したシステムを用いることで、いくつかの事が可能となる。まず 1 点目は、同定がすでにされている標本を本システムで調査し直すことで、同定が間違っている可能性のある標本を迅速に選び出せる可能性ができたこと。標本室に収蔵されている標本の同定は、10%程度間違っていると考えられているため、このシステムを用いることで同定が正しくないものを選びだすことが可能となる。既存の種と似ているが既存の種ではない新種を発見することも可能になると考えられる。2 点目は、大学生や専門家、アマチュアの人が本システムを用いて種の同定を行うことで、図鑑や専門書を用いて長い時間が掛かっていた同定にかかる時間が大幅に削減される。本システムでは、2179 種の同定が可能であるが、国内には 5000 種から 7000 種の植物が自生していると考えられている。今回は 50 枚以上標本画像があるものを検索対象としたが、今後は標本画像をさらに収集することでさらに多くの種の検索が可能なシステムを構築できると考えられた。国内の植物に限らず海外の植物の標本画像もたくさんあるので、それらを使うことで、世界中の植物の同定システムが構築できると考えられる。また、現システムでは自然界に生息している生きた植物の画像を使っても同定をすることはできない。今後は、携帯で撮影した植物の同定ができるようにシステムを改良する必要がある。本研究で AI が植物のどの部分を同定に用いているのかが明らかになった。その部分を生の植物の画像中で探すシステムを新たに開発することで、これが可能になると考えられる。ドローンにカメラを搭載してその画像からその地域にどんな植物がどれだけ生えているかを明らかにするシステムが構築できれば、生態系を保護するうえで極めて重要なツールになると考えられる。

【引用文献】

1　森口淳樹ら 分類　12(1), 41-52 (2011)

2　木戸佑子ら　分類　16(1): 63-71 (2016)

3　Carranza-Rojas et al　BMC Evolutionary Biology 17:181 (2017)

【本助成に関わる成果物】

[論文発表]

なし

[口頭発表]

なし

[ポスター発表]

なし

[その他]

なし