

識別モデルの漸次的更新に向けた インタラクションに基づくラベリング確信度推定手法の検討

安孫子 智樹^{1,a)} 岸野 泰恵^{1,2} 白井 良成²

概要: 我々は、ユーザが画像を投稿すると識別結果を返すだけでなく、投稿された画像に他のユーザがラベルを付与できる Social Networking Service (SNS) を検討している。ラベル付け可能な仕組みを SNS に導入することで、ユーザが付与したラベルに基づく識別モデルの更新を漸次的に行い、サービス向上を図ることが可能となる。本稿ではこのようなサービスの実現に向けて、ユーザのラベリング操作からユーザのラベルに対する確信度を推定する手法を提案する。不特定のユーザによって付与されたラベルは、その精度が課題となる。そこで我々は、精度と高い相関のある確信度をユーザのラベリング操作から推定可能なインタフェースの設計と実装を行った。実験の結果、ラベリング操作の時間や推薦に基づく回答の揺れなどにユーザの回答に対する確信度が表れることが分かった。

1. はじめに

画像を媒介とした Social Networking Service(以下画像媒介 SNS) は、Instagram^{*1} や Flickr^{*2} など、今日では様々なものが広く利用されている。これらの画像媒介 SNS に対し、近年成長の著しい画像識別技術を導入することで、サービスの向上やコミュニティの活性化を期待できる。例えば、植物や花好きが集まるコミュニティサービス “Green Snap”[5] では、多くのユーザが植物の写真を投稿しコミュニティを盛り上げており、画像媒介 SNS ととらえることができる。この Green Snap の公式スマートフォンアプリを利用すると、撮影した植物の名前を AI(i.e. 画像識別器)や他のユーザに問い合わせることができる。すなわち画像識別技術が画像媒介 SNS の活性化につながっている。

通常、識別するクラスが多い場合、画像識別器の識別精度の向上には、大量の訓練データが必要となる。そのため、画像媒介 SNS において、画像識別器を利用するには、あらかじめ大量の訓練データを準備しなければならない。しかし、訓練データの作成には、素材となるデータの入手や、素材に対するラベリング作業が必要であり、多くの場合準備にかなりの人的、金銭的コストがかかる。さらに、ユーザのニーズに応じて識別器を再学習する必要があるが、このためにも多くの訓練データが必要となる。

このような訓練データの作成コストを低減するために、ラベリングに必要な操作を減らすことでラベリング作業の効率化を図る方法 [1], [2], 訓練データの増やし方法 [9], 学習に寄与する可能性が高いデータに優先的にラベルを付けることで学習データの量を削減する方法 [12] などの研究がされている。また、近年では、Amazon Mechanical Turk[6] など、低コストでこのような作業を依頼できるクラウドソーシングサービスも利用可能だ。しかし、いずれの方法を用いても依然として無視できない人的、金銭的コストがかかってしまう。

そこで、我々は、このような画像媒介 SNS において、ユーザが投稿した画像を素材として、ユーザに訓練データを作成してもらうことができないかと考えている。ユーザが投稿した画像を訓練データとして利用できれば、ユーザのニーズに沿った画像識別器を構築できる。また、訓練データとして利用するためのラベリング作業もユーザに行ってもらうことができれば、その作業にかかるコストの大幅な低減が期待できる。

本論では、上記のような考えをもとに、まず、2 章において、画像媒介 SNS におけるユーザ参加型の画像識別器更新の枠組みを提案する。本枠組みでは、ユーザの自然な操作やコミュニケーションから得られたデータや、ユーザが自発的にラベリングしたデータをデータセットとして活用し、ユーザ群のニーズに合致した画像識別器を作成、更新していく。本枠組みの実現においては、データセット作成をユーザの自発的な活動に委ねるという性格上、データの質にばらつきがあり、かつ、一度に大量のデータを確保

¹ 大阪大学情報科学研究科

² NTT コミュニケーション科学基礎研究所

a) abiko.tomoki@ist.osaka-u.ac.jp

^{*1} <https://www.instagram.com/>

^{*2} <https://www.flickr.com/>

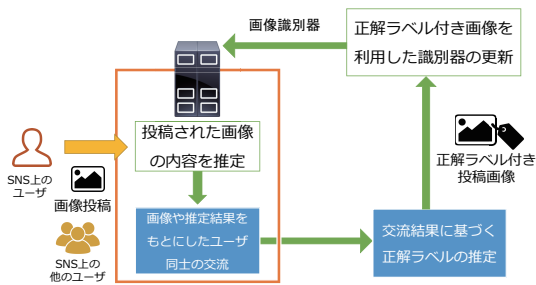


図1 ユーザ参加型識別器更新の枠組み

するのが困難という課題がある。そこで、我々は、ユーザが作成したラベリングデータの質を判定し、判定結果を利用して効率的に学習を行う手法の検討を進めている。具体的には、ユーザのラベリング操作から付与したラベルに対するユーザの確信度を推定し、その確信度を重みとして利用し、画像識別器を更新する。3章以降の章では、本手法を詳説し、予備検討の結果について述べる。

2. ユーザ参加型識別器更新の枠組み

2.1 画像媒介 SNS における識別器の漸次的更新

深層学習技術の発展に伴い、画像識別の性能は近年大幅に向上している。しかし、その性能を支えるのは、人手によって作成されたデータセットであり、その訓練データ、画像識別においてはラベル付き画像データをどのように効率的、かつ低コストに収集するかが課題となる。Human Computationの研究領域では、画像に対して正解のラベル付けをするような計算機が未だ不得意とする作業を、不特定多数の人々を対象に実施する方法に関する研究が行われている[3], [11]。金銭的な報酬による動機づけをはじめ、多くの方法が提案され、一部は実用化されている[10]。

一方、データセットを作成しても、当初は目的に対して十分な識別器が作成できない場合もある。そのような際には、識別器を運用しながら更新し、識別の精度を徐々に向上させるという考え方がある。例えば、近年注目を浴びているHuman-in-the-loop(HITL) machine learningでは、人間が機械学習の出力に応じてフィードバックを与え、識別器の向上を図る。HITLでは、Active Learning[12]の手法を用いて、識別器の向上が期待できるデータへのラベリングを優先的に行う。

ユーザ参加型の識別器更新の枠組みは、Human ComputationとHITLの概念をベースに、画像媒介SNSで利用する識別器をニーズに即して持続的に向上させていくことを低コストに実現することを狙い提案するものである。

提案する枠組みを図1に示す。画像媒介SNSでは、SNSのテーマに即した画像がユーザによって投稿され、画像をもとにしたユーザ間の交流が発生する。SNSの運用者は投稿された画像に対する自動識別機能をこれらのユーザに対する付加価値として提供する。運用者は、また、識別結果

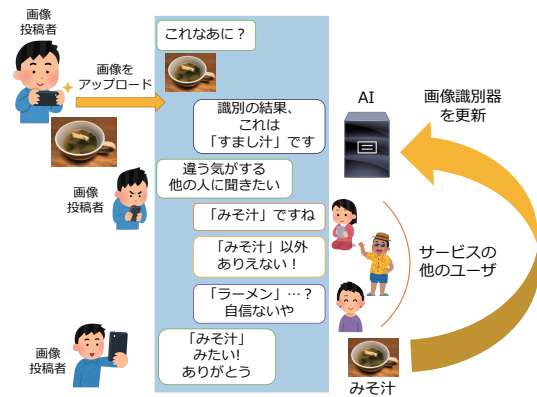


図2 料理画像識別サービスのイメージ

に対する修正機能もユーザに対して提供する。識別結果を修正したいと思ったユーザは、他のユーザとのディスカッション等を通じて識別結果の修正(i.e. 自発的なラベリング)を行う。SNSを提供するシステムは、ユーザが修正した識別結果を投稿画像の正解ラベルとして、投稿画像用識別器の再学習(追加学習)を行う。

このような仕組みを実現できれば、ユーザ同士の交流が同時に識別器の更新となり、低コストに画像に対する自動識別機能という付加価値を向上できる。また、ユーザによって投稿され、ユーザ間の交流に基づき付与されたラベリングは、SNSユーザのニーズを反映したものと考えられる。そのため、そのラベル付き画像によって訓練された識別器はユーザのニーズに即したものに成長していく。

本枠組みによって実現される画像媒介SNSの具体例を図2に示す。この例では、ユーザは料理の画像を投稿し、ユーザ間で交流を図る。あるユーザが、旅行先で食事の際に撮影した料理の写真を投稿すると、画像識別機能が識別結果を投稿者に提示するとともに、SNSのユーザに対する掲示板を開設する。投稿者を含めたユーザは、投稿された写真や画像識別機能の回答結果を見ながらメッセージを交換し、最終的に写真の料理を特定する。料理名が特定できれば、その料理の食材や調理法、提供している飲食店などの情報共有も円滑になるだろう。また、ユーザ間で最終的に特定した料理名が当初の識別結果と異なる場合には、その料理名を投稿画像の正解ラベルとして識別器を再学習し、画像識別機能の向上を図れる。

2.2 識別器の漸次的更新の実現に向けた課題

前節で述べたように、画像媒介SNSのユーザによって投稿画像のラベル付けが行えれば、低コストに画像識別器を更新できると考えられる。一方で、本枠組みではSNSのユーザが画像識別器の訓練に利用するラベリング作業を行うため、実現に向けた幾つかの課題を解消する必要がある。すなわち、本枠組みでは、当該画像に関する専門家とは限らないユーザ(群)が、自発的に付与したラベルを用いて、識別器の学習を行うため、ラベリングデータの量や質が課

題となる。本節ではユーザ参加型識別器更新の枠組みにおいて求められる技術について整理を試みる。

2.2.1 動機づけ

我々の枠組みにおいては、ユーザの投稿画像を素材として、ユーザの自発的なラベリングの結果を訓練データとして扱う。そのため、訓練データの入手が保証されない。訓練データを定期的に十分な量入手するためには、ユーザがラベリングをしたくなるような動機づけが重要となる。例えば、正しい答えを多く投稿することで、他のユーザから称賛が得られる仕組みは動機づけの一つとなるだろう。

2.2.2 ユーザインタフェース

ユーザに動機があっても、ラベリングを付与するための操作負荷が高ければ、ユーザの多くはラベリングを敬遠すると思われる。理想的にはユーザ間の交流結果が自動的にラベルとなる、すなわちユーザが意識せずにラベリングを行えるようなインタフェースが望ましいが、そうでなければできるだけユーザの負担にならないような低負荷のラベリングインタフェースを提供すべきであろう。

2.2.3 精度評価

画像媒介 SNS のユーザには、特定の画像に関して詳しい者もいればそうでない者もいる。また、ある画像に詳しいユーザも別の画像に関しては詳しくない場合もある。このようなユーザ群が行ったラベリング結果には多数の誤りが混在してしまう。また、ラベリングをユーザの自発性に委ねる場合、各ユーザのラベリング結果に対する責任感には幅があると想定される。ラベリングに誠実に取り組まないユーザはしばしば不正確なラベルを画像に付与してしまうだろう。

ラベリング結果の精度評価手法は、これまでに多くの研究がなされている。単純な多数決 (Majority Voting) による手法から、統計的アプローチによるより高精度な精度評価手法 [4], [13] までさまざまな手法が提案されている。しかしながら、これらの統計的アプローチに基づく手法は、その前提としてラベリング結果の冗長性が求められる。本枠組みではラベリング作業はユーザの自発性を期待するため、訓練データ、すなわちラベリング結果が一度に十分な量入手できるとは限らない。そのため、低い冗長度であっても精度評価ができる手法が望ましい。

2.2.4 学習方法

提案する枠組みで扱うのは、これまでに述べたように精度にばらつきのあるデータセットであり、データ量にも限りがある。そのため、精度が低いと見込まれるデータを除いて識別器の学習を行うのではなく、精度が高いと予想されるデータを重視し、精度が低いと予想されるデータは軽視するといった、限られたデータを極力活用して学習する手法がふさわしい。

2.3 ユーザ参加型画像識別器更新の実現の初期検討

ユーザ参加型識別器更新の枠組みでは、ラベリングデータの質や量にばらつきがあることが予想されるため、ユーザの入力したラベリングデータを理想的には冗長化せずに取得し、識別器の更新に活用したい。そこで本研究では、統計的なアプローチではなく、ユーザがラベリングを行う際の操作のログを利用してラベリングデータの精度を推定する手法を検討する。得られた推定の精度を元に限られたデータを極力活用した識別器の追加学習を行うことを目的としている。

また、機械学習には、データセットのクラスごとのデータ数が均衡していない時に、データ数が少ないデータは重く扱い、データ数が多いデータは軽く扱うことで、各クラス平等に識別できるように学習する重みづけ学習とよばれる手法がある。本研究では、推定したラベリングデータの精度を利用して各データの重みを決定し、識別器を追加学習することで、精度の情報のないデータセットを用いた場合よりも精度よく識別が行えることを確認した。

以降の章では、3 章においてラベリング操作ログからラベリングの精度を評価する試み、4 章において、重み付きデータの学習に関する初期検討結果について報告する。

3. インタラクションに基づくラベリング確信度推定

HCI の研究領域では、ユーザの操作履歴や振る舞いからユーザの興味や作業結果に対する確信度を推定する研究が行われてきた。例えば、文章閲覧時間からユーザの興味を推定する [17]、視線の動きから Web ページ上の注目領域を推定する [16]、マウスの軌跡からアンケートの回答に対する確信度を判定する [14]、といったものだ。

上記研究群がターゲットとする興味や確信の度合いに関する情報は、識別器の訓練データ作成におけるラベリング結果の精度評価にも利用できると我々は考えている。小山らは、ラベリングを行う作業者が自己申告した確信度がラベリングの精度と相関があることを画像分類実験において確認し、確信度を用いてラベル統合の精度を向上させるための確率的手法を提案している [15]。この実験で小山らはユーザに直接的に作業結果に対する確信度を入力させているが、ラベリング操作からラベリングに対する確信度を推定できれば、より低負荷にユーザからラベリングの確信度に関する情報を引き出せるだろう。また、小山らは、自信過剰または自信過小な作業者の存在を指摘している。自己申告ではなく、ラベリング操作ログから確信度を推定できれば、作業者自身のメタ認知能力に由来する自己申告確信度の揺れ [15] を取り除くことができると考える。

本章では、このような考えのもとに設計した画像ラベリングシステムについて述べる。

3.1 ラベリング確信度推定のためのインタフェース設計方針

ユーザ間の交流 (メッセージのやり取り) から自動的にラベルデータを生成することが本枠組みでは理想的と考えているが、メッセージからラベルを抽出し、かつ精度を推定するのは容易ではない。そこで、まずは、明示的なラベリングインタフェースをユーザが操作しラベリングを行うことを前提とする。本枠組みでは、画像を投稿した人が、画像の内容に対して他のユーザにアンケートを取る (その回答が各ユーザのラベリング作業に該当する) ようなイメージを想定している。また、画像媒介 SNS における多数のユーザを想定するため、確信度を推定するためにマウスやキーボード、もしくはタッチパネルなど、画像媒介 SNS を利用する PC やスマートフォンで利用可能なデバイスのみから得られる操作情報を利用すべきと考える。

しかしながら、画像に対して正解をラベル付けするという作業は単純である。表示された画像に対して、候補の中から正解と思われるラベルを選択する、もしくはタイプする [7] だけだ。このような単純作業における操作ログには、ユーザの確信度はほとんど表出しえないと思われる。そこで、我々は、操作にユーザの確信度がより表出しやすいと思われる 2 つの仮説を立て、インタフェースを設計することとした。以下に我々が事前に立てた仮説を示す。

比較作業 回答に対する確信度を直接聞くのではなく、他の回答候補に対してどちらがより確からしいかを答えてもらい、その結果から回答に対する確信度を推定する。相対比較形式にすることで、自らの自信を直接回答する場合に比べ自己評価に伴う揺れが低減する。

推薦による回答の揺れ 回答候補を推薦することで回答を変えたり長考したりする場合には、当初の選択に自信がないと推測できる。

また、回答に自信がある場合は素早く回答し、自信がなかったり、迷ったりする場合には回答に時間がかかると思われる。そのため、指標として操作時間も確認する。

3.2 画像ラベリングシステムのデザイン

前節で述べた仮説をもとに、ラベリング作業時の操作ログからユーザの確信度を判定できるかを検討するため 3 種類のラベリング方式を実装した実験システムを構築した。実験システムは、画像媒介 SNS を Web ブラウザで利用することを想定し node.js+PostgreSQL で実装した。

方式 1: 確信度申告ラベリング

本方式では、小山らが行った実験 [15] と同様に、ラベリングを行う際、同時にそのラベルに対する確信度を 20 段階の間隔尺度で自己申告する。小山らの実験によると、自己申告された確信度と回答の正答率には相関が見られるため、この方式で得られた確信度をラベリング作業の精度評価に利用するのはある程度有効であろう。ただし、自信過剰も

しくは自信過少な回答があるという報告もあるため、回答時間からも正答率との関係を検証することを想定し、回答結果に合わせて回答時間もログとして記録する。

方式 2: 相対比較ラベリング

本方式では、確からしいと思われるラベルをユーザに 2 つ選択してもらい、どちらがより確からしいかの相対値を入力してもらう。相対比較形式とすることで方式 1 で想定される自信過剰・自信過少といった個人の性格に由来する回答の揺れが減少することを期待して設計した。本方式では Scheffé の一対比較法における 20 段階の間隔尺度に基づく相対評点をユーザに行ってもらい。ただし、候補となるラベルの全組み合わせをユーザが実施するのは、ユーザに対する操作負荷が高いため、今回はユーザが選択した 2 つのラベルに関してのみ相対評点を依頼する形とした。

方式 3: 推薦に基づくラベリング

本方式では、システムが推薦する候補を踏まえてラベリングを行ってもらう。システム推薦によってユーザの判断が揺らぐか否かを確信度推定の材料とすることを狙い設計した。判断の変化を確認するため、本方式では、まず、ユーザはシステムが推薦する候補を見ずに画像に対応すると思われるラベルを回答する。その回答後、システムは回答にかかわらず、ラベル選択肢の中から第 1 候補、第 2 候補を表示する。ユーザには、1 回目の回答と表示された 2 つの候補を踏まえ、もう一度ラベルを選択肢から選んで回答してもらう。システムは、上記 2 回の回答結果だけでなく、それぞれの回答に要した時間も記録することとした。これは、候補の推薦によってユーザが悩んでいるか否かを検証する材料として利用することを想定した。

3.3 確信度推定に向けたラベリング取得実験

ラベリング操作ログから確信度推定が可能かを検証するため、構築した画像ラベリングシステムを用いて、ラベリング操作ログ取得実験を行った。

3.3.1 実験設計

実験では、図 2 のような画像媒介 SNS を想定し、料理画像に対するラベリング作業をタスクとして設定した。料理は多くの人にとって身近であり、一般の人にも料理画像に対するラベリングが可能である。一方でなじみのない料理も多数存在するため、各被験者が提示される刺激画像に対してすべて確信度が高いとは考えにくく、確信度の違いを検証するという観点でも適していると考えた。

被験者に対して提示される刺激画像には、UECFood-256[8] を用いた。データセットに含まれる 256 種類の料理画像から、麺類およびスープ類の料理 23 種類を選別し、実験に用いることとした。実験時に被験者に提示される画面例を図 3~5 に示す。方式 1 では、画面上部に刺激となる画像が表示され、画面中央の選択肢の中から、対応するラベルを一つ選んで回答ボタンを押してもらう。回答ボタン



図 3 確信度申告ラベリング

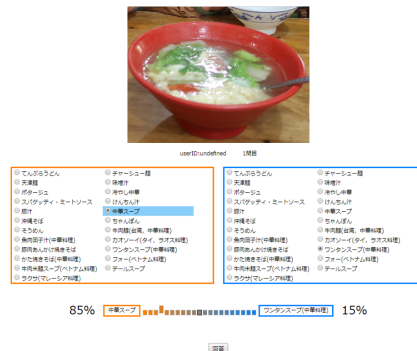


図 4 相対比較ラベリング



図 5 推薦に基づくラベリング

を押すと、正解している自信を回答するための評点ボタン (20 段階) が提示される (図 3)。被験者が自信度を選択し回答ボタンを再度押すと、回答結果と“ラベルの”回答までにかかった時間が記録され、次の刺激画像へ切り替わる。

方式 2 では、方式 1 と同様の選択肢が画面中央の左右に表示される。被験者は刺激画像に対応すると思われるラベルの候補を左右の選択肢群の中からそれぞれ選択する (重複不可)。選択したラベルは画面下部に表示される。被験者は、どちらのラベルがより確からしいかを評点ボタンで選択する (図 4)。被験者が最後に回答ボタンを押すと、一連の回答結果が記録され、次の刺激画像に切り替わる。

方式 3 においても方式 1 と同様の選択肢が画面中央に提示される。被験者が刺激画像に対応するラベルを選び回答ボタンを押すと、システムは AI による候補と称して刺激画像に対応するラベル候補を 2 つ表示する (図 5)。被験者に AI が実際に画像を分析していると認識してもらうため、システムは 1/3 の割合で、正解となるラベルを第 1 候補、もしくは、第 2 候補に提示する。すなわち、2/3 の割合で第 1 候補か第 2 候補に正解が提示される。また、それ以外の候補には、提示された画像の正解ラベルに対して間違いやすいと思われるラベル候補を予め 5 つ選別し、その中から提示することとした。被験者が、この「AI による候補」を見た後再びラベルを選択し回答ボタンを押すと、一連の結果は、1 度目の回答にかかった時間、2 度目の回答にかかった時間とともに記録され、次の刺激に切り替わる。

各被験者ごとに、方式 1 及び 2 では 138 枚の刺激画像に選択してもらった。方式 3 では 1 回目及び 2 回目の回答の正解、不正解の組み合わせごとに分析を行うことを想定し、倍の 276 個の刺激画像を用意した。また、(方式 3 の 139 枚目以降を除く) 方式 1 から 3 で表示される刺激画像は提示順も含め、全被験者を通して同一のものとした。実験の実施順は方式 1 → 方式 2 → 方式 3 の固定とした。

また、上記ログ取得実験の前に、被験者に対して、選択肢となる 23 種類の料理に関し、よく知っているから全く知らないまでを 5 段階でそれぞれ回答してもらうアンケートを実施した。これは、各料理への事前知識が、回答時間

等にどの程度影響するかを確認するためである。

3.4 実験結果

20 代の男性 14 名を対象に、2019/12/4～12/9 に被験者実験を実施した。実験において取得した操作ログの分析結果をそれぞれ以下に示す。

3.4.1 方式 1: 確信度申告ラベリング

図 6 は、方式 1 におけるラベリング結果が正解の場合、不正解の場合それぞれに対する自己申告確信度の分布を示す箱ひげ図である。縦軸は 20 段階の間隔尺度で得られた被験者の確信度申告を 0.0～1.0 に正規化している (1.0 が最も確信度が高い)。ラベリング結果が正解の場合よりも不正解の場合のほうが、相対的に自己申告の確信度は低い。また、各回答における時間のばらつきを示した図 7 からは、ラベリング結果が正解の場合に比べ、不正解の場合に回答時間が伸びる傾向が読み取れる。

3.4.2 方式 2: 相対比較ラベリング

方式 2 の実験では、被験者はラベル候補を 2 つ選択した後、どちらがより確からしいかを 20 段階の間隔尺度で指定する。方式 2 では、より確からしいとしたラベル候補を被験者の回答とし、実際の正解ラベルと比較して正誤を判定した。また、間隔尺度選択の水平に並んだ評点ボタン (図 4 下部) によってラベルの確からしさを被験者は入力するが、中央から両端に向かうそれぞれ 10 段階のこのボタンによる回答を 0.0～1.0 の値に正規化し確信度としている。結果として、方式 2 においても、ラベリング結果が正解の場合よりも不正解の場合のほうが、相対的に確信度が低いという結果が得られた (図 8)。また、本方式では、回答が正解の場合、推定された確信度の多くが 1.0 と推定されている。

3.4.3 方式 3: 推薦に基づくラベリング

方式 3 においては、システムが提示する推薦を見てから行った回答を被験者が最終的に付与したラベルとして正誤を判定した。図 9 より、正しい候補が推薦されたかどうかに関わらず、正解よりも不正解のラベリング結果のほうが回答に時間がかかっている (図 9)。また、1 回目の回答及び 2 回目の回答がそれぞれ正解 (T) か不正解 (F) かの組み

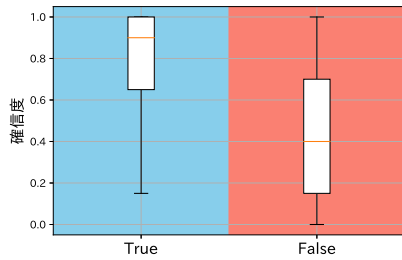


図 6 方式 1 における自己申告確信度

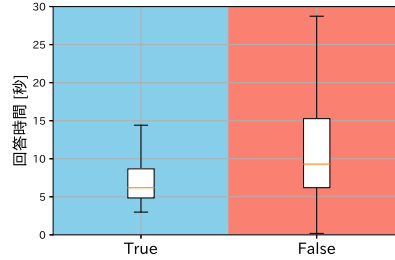


図 7 方式 1 における回答時間

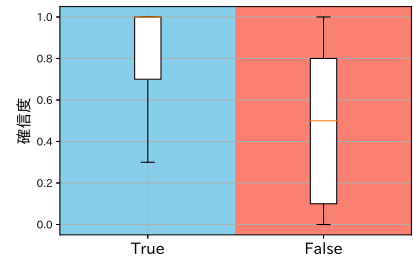


図 8 方式 2 において推定した確信度

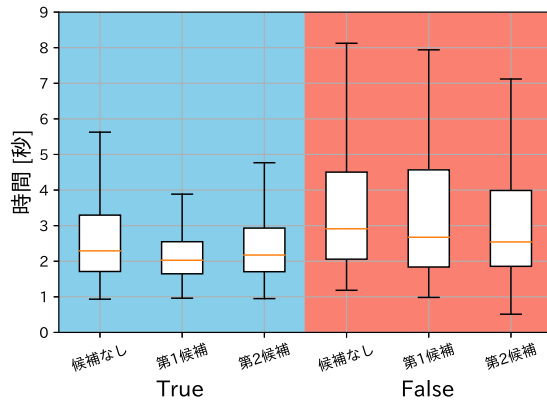


図 9 方式 3 における推薦後の回答時間。推薦の第 1 候補に正解があった場合 (第 1 候補), 第 2 候補に正解があった場合 (第 2 候補), 推薦に正解がなかった場合 (候補なし) に分けて集計。

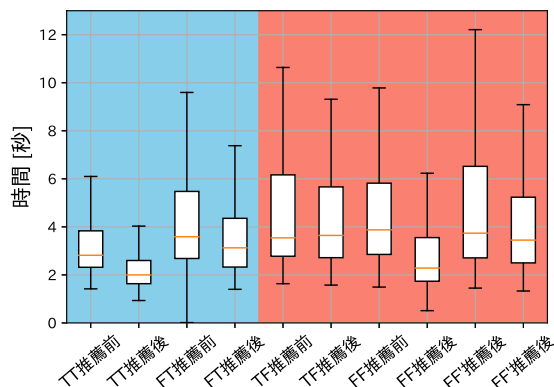


図 10 方式 3 における回答時間を推薦前後の正誤の組み合わせに基づいて集計。推薦前後で何れも誤回答の場合、誤回答の中で回答を変更した場合は FF' として集計した。

合わせ (TT, TF, FT, FF) 別に, 時間を比較したところ, T から F に変更した際には, 2 回目の回答のほうがかった時間の中央値が上昇する結果となった (図 10)。

3.5 考察

3.5.1 相対比較形式の効果

方式 1 及び方式 2 における各被験者の確信度の平均と正答率をそれぞれプロットした図を図 11 に示す。正答率に比べ確信度が高い点が自信過剰な被験者を示し, 正答率に

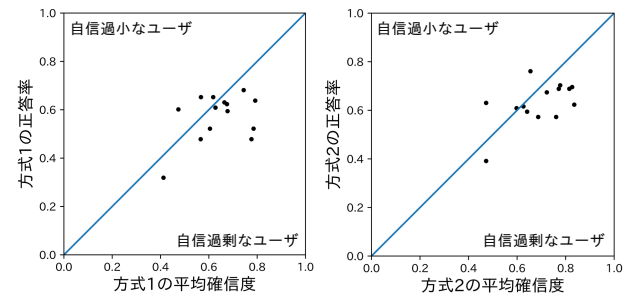


図 11 平均確信度と正答率

比べ確信度が低い点が自信過小な被験者を示す。被験者数が少ないものの, 今回の結果からは方式 2 においても方式 1 とほぼ同様の傾向がみられ, 確信度と正答率に相関は見られるものの, 相対比較形式にしたことによる大きな効果は確認できなかった。

3.5.2 推薦による回答の揺れ

方式 3 において, 回答を変更しなかった場合, した場合の正答率を確認したところ, 変更しなかった場合の正答率は 80.3% に対し, 変更した場合の正答率は 52.9% であり, 回答を変更しなかった場合のほうが顕著に正答率が高い結果となった。これは, 回答を変更した場合には被験者が回答に対して自信がなく, ラベル付けの精度も悪いという仮説を裏付ける結果と考えている。ただし, 回答を変更したことで正解となったケースも一定数存在したため, 回答を変更したか否かで一律に確信度を割り振るのはいささか乱暴であると考えている。特に, AI の推薦精度がより向上した場合, 回答の変更によって正解となる割合は上昇すると思われる。変更したかしなかったかの結果とともに, 別の確信度推定方法も組み合わせると良いと考える。

3.5.3 時間による重みづけ

方式 1 及び方式 3 の 1 回目, 2 回目の回答にかかった時間は, いずれの条件においても回答が正解の場合よりも不正解の場合により時間がかかる傾向が表れている。時間情報は, 自己申告による確信度などに比べれば正解・不正解に対して顕著な差が得られるわけではないが, 回答にかかった時間を用いて, ユーザが自己申告による確信度や推薦による回答の揺れを利用して推定した確信度を補正できるのではないかと考えている。

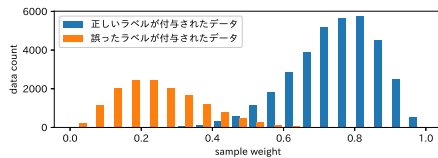


図 12 精度値割り振りパターン A

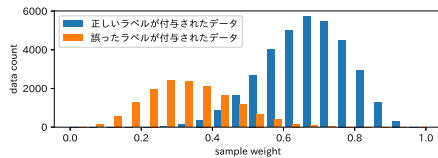


図 13 精度値割り振りパターン B

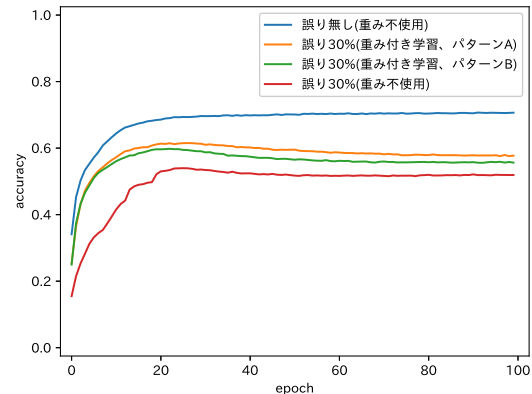


図 14 学習結果の比較

3.5.4 事前調査を利用した確信度補正

選択肢となる 23 種類の料理に対して事前に実施したアンケートの結果に基づいて、方式 1 における回答結果を当該料理を「知っている」と、「知らない」の 2 群に分類した。そして、各群における正答率を算出した結果、「知っている」群の正答率は 72.5%であり、「知らない」群の正答率は 31.5%であった。これは、例えば、画像媒介 SNS への参加登録時にラベルに対する知識を事前に確認しそれを利用することで、ラベリング結果の確信度をある程度推定できることを示唆していると考えられる。

4. 精度値を利用した重み付き学習

本論で提案する枠組みでは、潤沢にラベリングデータを取得できるとは限らないため、精度が低いデータもできるだけ無駄にせず訓練データとして利用したいと考えている。そこで、我々は、精度値を利用した重み付き学習でどの程度識別器の性能向上が可能かを確認した。前章の実験で得られたデータは、実験結果の分析を容易にするため、各被験者に同一の刺激画像を提示している。そのため、各クラスの刺激画像はそれぞれ 6 枚しかなく (方式 3 では 12 枚)、識別器の構築精度を確認するには不十分なため、今回は、cifar-10 の画像を評価に用いることとした。

4.1 実験設計

本枠組みにおいて付与されたラベルには一定の誤りが混在することを想定し、実験では、30%の誤りラベルが付与されているという想定で行うこととした。cifar-10 の 10 カテゴリの各画像データを訓練データと検証データに分け、訓練データに誤りを混在させた。具体的には、各カテゴリそれぞれ 30%の画像を、別のカテゴリに均等に割り振った。

次に訓練データには、確信度に基づく精度値が付与されていると想定し、各画像に対するラベル精度の割り振りを行った。精度がある程度明確に割り振られた場合と、そうでない場合を想定し、2 パターンの割り振りを行っている。両パターンの精度値のヒストグラムを図 12, 13 に示す。パ

ターン A では、前章の手法、もしくは他の手法によってある程度精度を明確に割り振ることができた想定で精度値の割り振りを行った。正解ラベルに関しては、精度値 0.8 付近、誤りラベルに関しては精度値 0.2 付近がピークとなるようにそれぞれベータ分布に基づく乱数を生成し精度値を割り振った (図 12)。パターン B では、パターン A よりも精度評価がうまくいかず、正解ラベルに低精度値、不正解のラベルに高精度値がある程度付与されている状況を想定した。具体的には、正解ラベルに関しては、精度値 0.7 付近、誤りラベルに関しては、精度値 0.3 付近がピークとなるようにそれぞれベータ分布に基づく乱数を生成し精度値を割り振った (図 13)。重み付き学習には keras を用い、学習時に各訓練データの sample_weight^{*3}に上記精度値を用いることで実現した。識別モデルは keras のサンプルにある一般的なものを用いた^{*4}。

4.2 実験結果

実験結果を図 14 に示す。図の横軸は epoch 数、縦軸は評価データの識別精度を示している。各学習曲線は、10 回施行した結果の平均である。比較のため、誤りラベルがない場合の学習曲線、および誤りを 30%含み、かつ精度値を用いなかった場合の学習曲線も描画した。誤りラベルを含んでいる場合、精度値を利用した重み付き学習を行うことで、重み付き学習を行わない場合よりも識別性能が向上することがわかる。ラベリング結果の精度評価が不十分な場合 (精度値割り振りパターン B) も、精度評価の精度が良い場合に比べ識別精度は若干劣るものの、重み付き学習を行わない場合に比べ良い結果が得られた。

5. 議論

5.1 インタラクションに基づく確信度推定

本論では、ユーザ参加型識別器更新の枠組みの実現に向

^{*3} <https://keras.io/ja/models/sequential/>

^{*4} https://keras.io/examples/cifar10_cnn/

けて、ラベリングの操作ログから、ラベリング結果に対する確信度を推定するための基礎調査を行った。その結果、ラベル付与時の時間や、AIの推薦に対する回答の揺れ等に、ラベリング結果に対する確信度が表れることが確認できた。一方で、より詳細な実験結果の分析は必要なものの、これらの指標は、それぞれ単一ではラベルの正誤を示す精度値としては、不十分である。各指標により明確に確信度が表出するようなインタフェースの設計や、指標同士の組み合わせによる精度の向上などを今後検討していきたいと考える。前者としては、例えば、より時間に人の迷いが反映されるようなボタンの配置や表現方法などがあげられる。後者としては、例えば、方式1において、ユーザが自己申告した確信度を回答にかかった時間で補正することでより確信度の精度を高めることができるかもしれない。また、インタラクションに基づく確信度推定は、統計的アプローチによる精度評価を否定するものではない。ある程度ラベリングの冗長化が図れる状況下においては、両者を組み合わせることでより高精度にラベリング結果の精度を評価することができると思われる。

5.2 精度値を利用した重み付き学習

本論では、ラベリング結果の精度値を利用した重み付き学習の初期検討結果を示した。4章で示した実験におけるパラメータ(誤りラベルの混入割合や確信度を想定した精度値の分布)は、提案する枠組みを想定して得られた実際のラベリングに基づいたものではない。今後実際のラベリング実験結果を利用した重み付き学習の評価を行いたいと考える。また今回は検討の端緒としてkerasにおいてsample_weightを利用した重み付き学習を行ったが、他の重み付き学習方法についても調査検討を行いたい。

5.3 ユーザ参加型識別器更新の枠組みの実現

本論で提案したインタラクションに基づく確信度推定方法は、ユーザ参加型識別器更新の枠組みにおける主に精度評価技術として検討したものである。同様に、精度値を利用した重み付き学習は、学習方法技術として利用することを想定した。一方で、提案する枠組みの実現においては、動機づけ方法や低負荷なユーザインタフェースの設計も必要となる。これらは独立に考えるべきものではなく、各視点に立ってシステムを設計すべきであろう。例えば、動機づけ方法の良し悪しは、ラベリング時の操作時間に影響を与えるかもしれない。

6. まとめ

本論では、画像媒介SNSにおける画像識別器の利用を目的として、ユーザ参加型画像識別器更新の枠組みを提案した。また、提案した枠組みの実現に向けて、インタラクションに基づくラベリング確信度推定手法の検討を行った。

参考文献

- [1] Acuna, D. et al.: Efficient Interactive Annotation of Segmentation Datasets With Polygon-RNN++, *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).
- [2] Andriluka, M. et al.: Fluid Annotation: A Human-Machine Collaboration Interface for Full Image Annotation, *Proc. MM2018*, pp. 1957–1966 (2018).
- [3] Balducci, F. and Buono, P.: Building a Qualified Annotation Dataset for Skin Lesion Analysis Through Gamification, *Proc. AVI2018*, pp. 1–5 (2018).
- [4] Dawid, A. P. and Skene, A. M.: Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 28, No. 1, pp. 20–28 (1979).
- [5] GreenSnap, Inc.: Green Snap, <https://greensnap.jp/>.
- [6] Ipeirotis, P. G.: Analyzing the Amazon Mechanical Turk Marketplace, *XRDS*, Vol. 17, No. 2, pp. 16–21 (2010).
- [7] Jafarainami, N.: Exploring the Character of Participation in Social Media: The Case of Google Image Labeler, *Proceedings of the 2012 iConference*, iConference '12, New York, NY, USA, ACM, pp. 72–79 (2012).
- [8] Kawano, Y. and Yanai, K.: Automatic Expansion of a Food Image Dataset Leveraging Existing Categories with Domain Adaptation, *Proc. of ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)* (2014).
- [9] Krizhevsky, A. et al.: ImageNet Classification with Deep Convolutional Neural Networks, *Proc. Advances in Neural Information Processing Systems*, pp. 1097–1105 (2012).
- [10] Quinn, A. J. and Bederson, B. B.: Human computation: A survey and taxonomy of a growing field, *Conference on Human Factors in Computing Systems (CHI2011)*, pp. 1403–1412 (2011).
- [11] von Ahn, L. et al.: reCAPTCHA: Human-Based Character Recognition via Web Security Measures, *Science*, Vol. 321, No. 5895, pp. 1465–1468 (2008).
- [12] Wang, M. and Hua, X.-S.: Active Learning in Multimedia Annotation and Retrieval: A Survey, *ACM Trans. Intell. Syst. Technol.*, Vol. 2, No. 2, pp. 1–21 (2011).
- [13] Whitehill, J., Wu, T.-f., Bergsma, J., Movellan, J. R. and Ruvolo, P. L.: Whose vote should count more: Optimal integration of labels from labelers of unknown expertise, *Advances in neural information processing systems*, pp. 2035–2043 (2009).
- [14] 酒巻隆治, 染矢聡: Web アンケートにおける回答者のマウスログに基づく確信度の強さの序列推定, 可視化情報学会論文集, Vol. 31, No. 8, pp. 43–43 (2011).
- [15] 小山聡ほか: クラウドソーシングにおけるワーカーの確信度を用いた高精度なラベル統合, 第27回人工知能学会全国大会論文集, pp. 1–4 (2013).
- [16] 大野健彦: IMPACT: 視線情報の再利用に基づくブラウジング支援法, インタラクティブシステムとソフトウェア WISS2000, pp. 137–146 (2000).
- [17] 梅村祥之, 増山繁: 文章閲覧における読者の操作行動からの興味推定, ヒューマンインタフェース学会論誌, pp. 435–444 (2006).