インタラクティブな画像認識システムにおける画像テキスト変換手法の活用

川辺 航[1,a)]　菅野 裕介[1,b)]

概要：インタラクティブ機械学習は，専門知識を持たないユーザによる自立的な機械学習モデルの作成を可能にする．当該分野ではシステム設計に際して分類アルゴリズムが採用される例が多いが，分類は形式が単純すぎるあまりにユーザが定義できるタスクの幅を狭めるという難点がある．また初心者ユーザにとっては，分類での定式化が可能なタスクを解く時でさえ適切にラベルを定義することは困難である．本研究は，インタラクティブな画像認識システムを題材に上記の問題の解決を試みる．具体的には，アルゴリズムとして画像分類ではなく画像テキスト変換を導入する．テキストの場合は定式化可能なタスクの種類に理論上制限がなく，さらにラベルという形式に縛られずに柔軟なアノテーションが可能となる．テキストの長所・短所の検証として，初心者ユーザに画像分類システムと画像テキスト変換システムの両者を与え多様な画像認識タスクを解いてもらう比較実験を行った．結果として，画像テキスト変換を採用した場合，ユーザ体験の質は画像分類に対して有意に優れることはないものの，ユーザは表現形式に縛られない多様なアノテーションができるようになることが判明した．

1. はじめに

近年，機械学習，とりわけ深層学習が実世界の様々な場面に適用され，ツールとしての存在感を高めている．しかしながら，ユーザが多様な目的に対して自らの手で機械学習モデルを設計・構築する機会は依然として少ない．これに対しインタラクティブ機械学習 [6] は，適切な UI やアルゴリズムの設計によって，ユーザが自身の力で機械学習モデルを作成できる状態を目指す．当該分野では，これまで多くの研究がシステムの設計や実証実験を通じたユーザの理解に取り組んできた．

先行研究で提案されたシステムは，その多くが分類のアルゴリズムに基づいている [5]．分類の枠組みにおいては，認識対象となるデータに対する正解クラスをユーザが定義し，モデルはデータとクラスのペアを学習する．分類は直感的な理解が容易いアルゴリズムである反面，複数のオブジェクトを包括する抽象的なクラスの定義はユーザにとって困難であるという問題 [14] や，回帰問題のように分類の枠組みでは定義できないタスクが存在するという難点を抱えている．よって，ユーザが個別の目的に応じて機械学習モデルを自由に設計する状況を考慮する場合，分類アルゴリズムの愚直な導入は適切でないと言える．

分類が出力とするラベルは事前に定義された正解の一つである一方で，テキストは最も汎用性の高い形式である．もしテキストを出力に採用する場合，その柔軟な記述可能性ゆえに，ユーザの解釈や多様なタスクへの記述可能性について異なる結論が得られる可能性がある．にもかかわらず，テキストを導入した際に上記の諸問題が同様に発生するかもしくは解決されるかといった検証は未だになされていない．

そこで我々は，画像を題材に，テキストを正解ラベルとするインタラクティブな認識システムをプロトタイプとして作成し，ユーザとシステムのインタラクションの詳細な解析を試みた．我々が作成したプロトタイプである画像テキスト変換システムでは，ユーザは事前に定義したクラスを画像に割り振る代わりに，各画像に紐づいたテキストを入力する．さらに，システムの内部では，画像分類モデルではなく画像テキスト変換モデルがユーザーの定義した画像とテキストのペアを学習する．

本研究では画像テキスト変換を導入した際のユーザの解釈や多様なタスクへの記述可能性を検証するべく，初心者ユーザを対象として分類システムとの比較実験を行った．実験参加者は画像テキスト変換システムと画像分類システムの双方を用いて複数の画像認識タスクを解き，我々は参加者が抱いた印象や彼らが作成した学習データ，モデルの学習状況を解析した．その結果，画像テキスト変換システ

[1] 東京大学 生産技術研究所
[a)] wkawabe@iis.u-tokyo.ac.jp
[b)] sugano@iis.u-tokyo.ac.jp

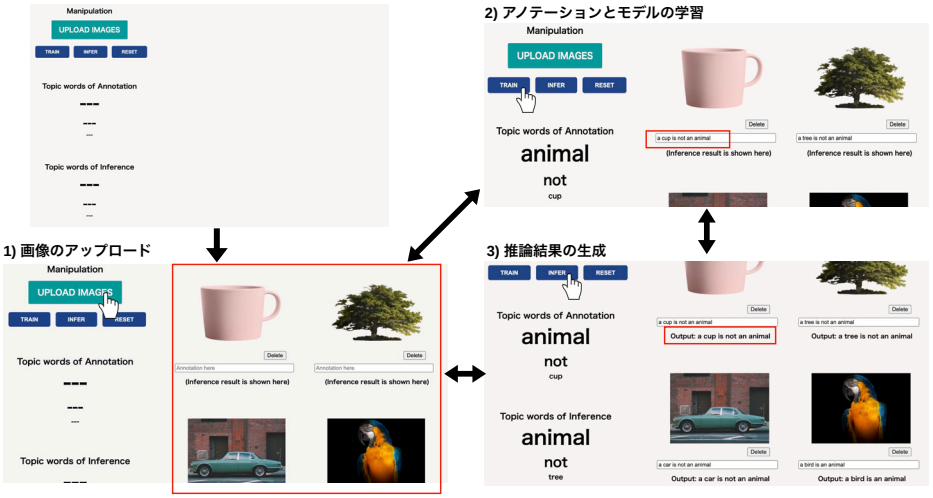

図 1 画像テキスト変換システムの操作フロー．(1) UPLOAD IMAGES を押すとローカルに保存してある画像をアップロードでき，右側のエリアに反映される．(2) 画像下のテキストボックスにアノテーションし，TRAIN を押すことで学習が実行される．(3) INFER を押すと，テキストボックスの下に推論結果が表示される．

ムと従来の分類システムの間には使いやすさの面で有意な差はないものの，かつユーザに多様なラベルの定義を可能とすることが示された．

2. 関連研究

2.1 インタラクティブな画像認識システム

機械学習の専門知識を持たないユーザによるモデル設計を可能とするシステムは，その多くが分類アルゴリズムに基づいている [11], [15], [17]．分類をベースとしたインタラクティブシステムにおいては，回帰タスクを解くことができない．ゆえに，ユーザが想起する多様な使用目的に対し柔軟に機械学習モデルを適応させることが不可能である．本研究では，テキストが複数のタスクに対する柔軟な定義を可能にするという仮説を立て，その検証のために実験参加者に多様な画像認識タスクを与えている．

分類をベースとしないインタラクティブシステムについては，画像検索 [8]，ターゲットとなる物体の追跡 [1]，データのエンベディング [15] 等がある．これらも，特定のタスクの範囲内でのモデル設計のみが想定されており，ユーザが複数のタスクに対してモデルを自由に設計できる枠組みではない．本研究では，テキストの多様なタスクへの応用可能性を検証するため，複数の画像認識タスクを同一のシステムで解くというユーザ実験を行った．

2.2 カスタマイズ性のある画像テキスト変換

画像テキスト変換モデルの出力をユーザ自身がカスタマイズする研究は存在する．iCap [10] はユーザのテキスト入力に対してシステムが動的に正解キャプションの候補を提示するシステムである．ユーザはインタラクティブにテキストを選択することが可能である反面提示された候補を逸脱したテキストの定義が許されておらず，iCap によって汎用的なキャプショニングが実現されたとは言えない．本研究で用いる画像テキスト変換プロトタイプは，テキストの汎用性を前提としており，ユーザが自由にテキストを定義する．

一方で，ドメイン適応を用いることで，事前学習に使われたデータセットとは異なるスタイルのテキストをモデルに出力させる例がある [19], [22], [23]．これらの研究は画像テキスト変換モデルにカスタマイズ性を付与した研究であると解釈されるものの，あくまで精度の向上が主な目的であり，リアルタイムなインタラクションを想定したものではない．対して本研究で用いるシステムは，GUI を通じたユーザとモデルのインタラクションが想定されており，実験の内容も参加者によるシステムを用いた機械学習モデル設計がテーマである．

3. インタラクティブな画像テキスト変換システム

テキストを機械学習モデルの出力形式として採用することの有用性を検証するべく，インタラクティブな画像テキスト変換システムの実験用プロトタイプを開発した．本プロトタイプでは，ユーザが画像に対してテキストをアノテーションすることで学習データを作成し，モデルの学習は画像とテキストのペアを用いて行われる．

3.1 インタラクションの設計

本システムは図 1 に示すような GUI を持つ Web アプリである．ユーザは (1) 画像のアップロード，(2) テキストのアノテーションおよび作成した画像テキストによる変換モデルの学習，さらに (3) 学習済みのモデルを用いた推論

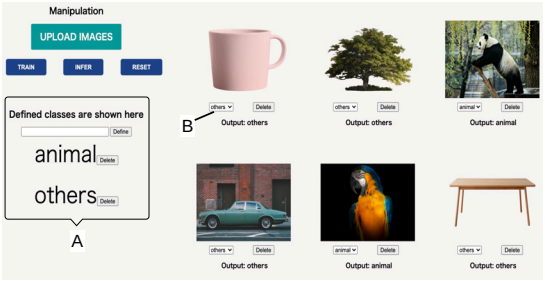

図 2 画像分類のプロトタイプの概要．（A）ユーザはラベルを事前に定義し，（B）各画像に対してラベルを割り振る．

結果の生成と確認を行う．学習時には画像とその下にタイピングされた英語のフレーズ・文がセットとなり学習データとして用いられる．また，推論時には，テキストボックスの下に推論結果が表示される．ユーザは推論結果や画面左下に表示されるアノテーション及び推論結果での頻出語を参照し，自らが目的とするモデルを作成するにあたってより効果的なテキストを検討・再入力することができる．

3.2 アルゴリズムの設計

画像テキスト変換モデルとしてはエンコーダデコーダ構造を持つ End-to-end モデル [4] が用いられている．エンコーダの CNN は画像から特徴を生成し，デコーダの Transformer Decoder は特徴から単語のシーケンスを生成する．Transformer Decoder は画像特徴とシーケンスを入力とし，そのシーケンスに続く一単語を出力する．エンコーダを ImageNet [3] で学習させた後，デコーダを MS COCO [12] で学習させる．

モデルの学習時には GPU を用いて高速化を試みるものの，依然として学習にかかる時間は無視できない．学習時間を短縮しユーザ体験の質を高めるべく，我々はモデルの学習部位を制限する工夫を導入した．Transformer Decoder は 6 つの層を有するが，そのうち 5 つを学習の対象から外し，残りの 1 つのみを最適化する．

3.3 画像分類ベースライン

比較対象として，GUI を画像テキスト変換システムに似せた画像分類システム（Fig. 2）を開発した．画像テキスト変換システムとの違いとしては，画像の下にテキストを打ち込む欄はなく（Fig. 2A），代わりにクラスを定義する欄と定義したクラスを選択するプルダウン式のボタンが用意されている（Fig 2）．

4. 実験

インタラクティブな画像認識システムにおいてテキスト変換を採用した場合と分類を用いる場合を比較するためにユーザ実験を行った．ここでは実験の詳細と実験結果に触れる．

4.1 実験手順

20 人（男性 9 人，女性 11 人，21 歳-48 歳）の被験者を雇った．共通する特性として，機械学習に関する知識を持たず，英語はネイティブではないものの辞書を用いての読み書きができる．まず，参加者は機械学習を用いた画像認識の概要とシステムの使い方について講義を受け，次にシステムを使って合計 4 つのタスクを解いた．4 つの内 2 つは画像テキスト変換システム，残り 2 つは画像分類システムを使って行われ，後述の 10 あるタスクの内 4 つがランダムに割り振られた．システムを使う順番とタスクの割り振りは参加者ごとに異なり，全体で釣り合いが取れた状態であった．

終了後は，実験やシステムに対する所感を聞くためにアンケート調査を行った．ユーザビリティに関する定量的な質問は 6 つあり，5 段階の Semantic Differential Scale を用いた．それぞれシステムの使いやすさ，直感的な操作の可能性，学習データ作成の効率性，学習データ中に必要なデータを含められたかどうか，学習方法の有効性，日常生活におけるシステムの有用性を問うている．定性的な質問は自由記述式であり，PQ1：「どのようなことを意識して学習データを作ったり学習させましたか？」と PQ2：「システムに関して感じたこと、考えたことを自由に書いてください」の 2 つである．また，メンタル負荷を NASA-TLX を用いて調査した．項目 Physical Demand は，本研究と無関係であるので削除した．

4.1.1 画像認識タスク

先述の通り分類の場合は，ユーザが抽象的なラベルを適切に定義できないという問題や，表現可能なタスクの幅が限られるという欠点がある．本実験では，テキスト出力の場合どの程度多様なタスクに対して柔軟に適応可能であるかを把握するために，分類以外の認識タスクを複数用意した．10 個の異なるタスクのうち，5 つは検出タスク，残りの 5 つは回帰タスクである（表 1）．分類をアルゴリズムとして採用したシステムの場合，検出タスクは理論上定義可能であるもののユーザにとっては困難であり [14]，回帰タスクは定義不可能である．回帰タスクについては，画像から量が目視で把握できるような指標を対象としてタスクを選定した．参加者が用いる画像セットについては，検出タスクでは元のデータセットから 40 クラスを抽出した．回帰タスクではデータセットからランダムに 200 枚を抽出した．

4.1.2 第三者によるモデルの定量的評価

各タスクのアノテーションに対する想定解は存在するものの，特定の正解と合致するか否かで一律に評価を行うと不都合が生じる可能性がある．たとえば参加者がこちらの想定を逸脱したアノテーションを行い，かつそれがタスクの要求を正しく満たした場合に，そのアノテーションを高く評価することはできない．そこで我々は，参加者の学習

表 1 事前に用意されたタスクの一覧と画像の出所であるデータセット．

ID	タスクのカテゴリ	タスクの要求	画像の出典
1	検出	陸上の生物であるかどうかを判断	Photo Art 50 [18]
2	検出	画像中の人物が何らかの運動をしているかどうかの判断	Stanford 40 Action [20]
3	検出	画像中のオブジェクトが道具として用いられるものか否かを判断	Caltech 101 [7]
4	検出	女性用のトップスが含まれているか否かを判断	DeepFashion [13]
5	検出	料理が肉料理であるかそうでないかの判断	Food 101 [2]
6	回帰	群衆がどの程度の混雑の度合いなのかを推定	CrowdHuman [16]
7	回帰	画像中の人物がどのくらい歳をとっているかを推定	UTKFace [21]
8	回帰	草や木が画像のどの程度の割合を占めるかを推定	Stanford Background [9]
9	回帰	画像中の人物の顔が向いている方向を推定	AFLW2000-3D [24]
10	回帰	船やボートが画像中のどこに存在するかを推定	MSCOCO [12]

データ作成のプロセスやモデルの学習状況の評価を第三者に委託した．評価者は参加者が作成した学習データや学習後のモデルの推論結果を参照し，(1) アノテーションがタスクの要求を満たすか，(2) 学習データの作り方は技術的観点から適切か，(3) 学習後のモデルの性能は参加者の意図通りだと思うか，の 3 つの質問に答えた．それぞれが EQ1，2，3 に対応する．

4.2 実験結果

参加者がモデルを学習させる過程や，彼らが試行を通じて得た印象を紹介する．

4.2.1 アノテーションの詳細

分類システムの場合，検出タスクでは 4 件が課題の意図を満たす抽象的なラベリングに成功していた．具体的にラベルを見ると，一単語でのラベリングが目立った（タスク 2 での “Yes”/“No” や “Sports”/“Other”）．回帰タスクでは 2 件が数値を用いたラベル付けを行った．ラベルの粒度が高く，似た意味であっても複数の言い回しで表現するケースがあった（「若い」という意味に対して “young”，“boys”，“children” など）．

テキスト変換システムでの試行を見ると，単語に限らずフレーズや文でのアノテーションも多い．検出タスクでは 7 件が抽象的ラベリングに成功しており，文中に正解ラベルを含めるアノテーションが目立った（タスク 2 では “Running is a sport”，“Playing the violin is not a sport” など）．検出タスクでは 5 件が数値を用いたラベリングをしており，ラベルの粒度自体は高くなかった．

4.2.2 ユーザーによる評価

Usability，メンタル負荷については，いずれの項目についても，Wilcoxon Signed-rank Test での検定を行った結果，有意差は見られなかった．これは，画像テキスト変換システムと画像分類システムの間に使いやすさの面で有意な差はないことを示す．

PQ1 の回答を見ると，画像分類システムに関しては，選択する画像の多様性に気を使う人がいた．「同じ種類（のラベル）であっても様々なパターンがあることを学習させようとした」．また，画像分類においてはクラスの名前自体は学習に影響を与えないものの，「重要な固有名詞を入れた」，「課題のワード（緑，運動，向きなど）を含めた」など名前の付け方に気を配ったという声も散見された．

一方で画像テキスト変換についても，選択画像の多様性やバランスに着目する参加者が存在した．また，テキストを記述する際の気遣いとして，シンプルさや統一性を挙げる参加者がいた．「明瞭な単語を用いて学習データを作った」，「単純な言葉で入力した」．

PQ2 については，両システムに対して適切な学習の難しさを訴える声があった．さらに，テキスト出力独自の性質に気づく人もいた．「画像のどこをメインにして（アノテーションを）記述するかで（推論結果の）内容が変わる」，「(アノテーションの際に）ルールが明確であれば学習させるのは容易だと感じた」．

4.2.3 第三者評価

第三者評価の結果を図 3 に示す．Mann-Whitney U Test での検定の結果，回帰タスクの EQ3 では画像テキスト変換が画像分類に比べて有意に高く評価されており（$p < 0.05$），その他の結果については有意差は見られなかった．ここから，画像テキスト変換がユーザのタスク遂行に悪影響を及ぼさないこと，また画像分類の枠組みで解決が不可能なタスクに対してテキストの場合は相対的により精度の良いモデルができていることが示唆される．

5. 議論

5.1 画像分類システムでの試行

画像分類システムを用いた試行では，分類の制約に起因する発見が目立つ．まず検出のタスクに着目すると，タスクの要求を満たす抽象的なラベルが少ない．その代わりに，画像に含まれる具体的なクラス名をそのままラベルにした例が多い．これは，画像の名称を端的にラベル化するという分類の性質に起因するのではないか．また回帰タスクを見ると，類似した表現に異なる単語をあてがうなど，粒度の高いアノテーションが散見される．さらに，数字を使ったアノテーションは避けられ，単語による端的なアノ

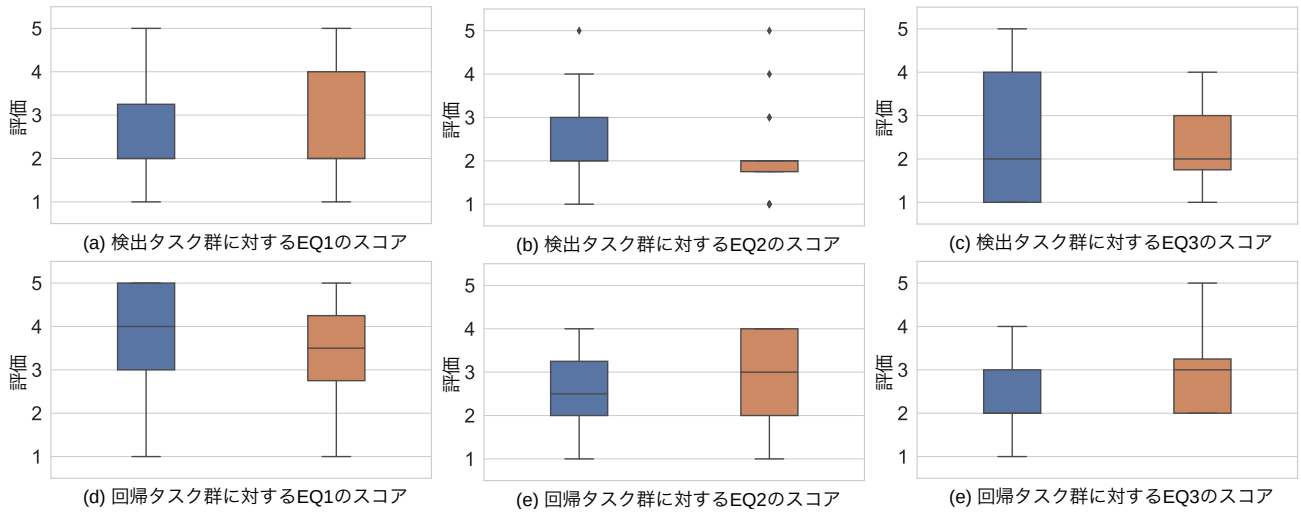

図 3 第三者評価の結果．左側青色が画像分類システム，右側オレンジ色が画像テキスト変換システムを示す．

テーションが目立つ．以上から，画像分類はラベルの簡潔な定義に適するものの，ユーザが想起する表現の幅を狭めると言える．これには，事前説明の際に単語ベースのアノテーションを提示したこと，クラスの定義という手順自体が画像の端的な説明を促していること等が原因として考えられる．

5.2 画像テキスト変換を採用する場合の長所短所

画像テキスト変換システムの場合，テキストの柔軟性に由来する多様な形式のアノテーションが観測された．検出タスクでは抽象的なラベルが画像分類の場合よりも多く，また文の形式を取るラベルも目立った．ここから，テキストによるアノテーションがより柔軟かつ的確なラベリングを促すことが示唆される．回帰タスクにおいては，アノテーションの粒度が低い傾向にあった．また，文の形式でのラベルが多いのは検出タスクと同一の傾向である．テキストによって単なる一単語，一フレーズに止まらない多様なアノテーションが実現されている．以上から，テキストの採用はユーザに多様な表現を想起させると考えられる．この柔軟性の高さはテキストの長所である反面，タスクに適したアノテーションの形式をユーザに提供できないという短所であるとも解釈できる．仮にユーザが自身の目的に対する解決策（モデル，学習方法など）を明確にイメージできている場合，それに特化したモデルを用いるのが精度面を考慮すると合理的である．テキストを用い，多様なタスクにモデルを適応させる意義としては，ユーザにとって事前知識が不要であり，タスクの区別のないアノテーションを可能にすることが挙げられる．

5.3 画像テキスト変換システムのユーザビリティ

定量的には，2 システムのユーザビリティや参加者が受ける精神的疲労に有意差はない．テキストを画像一つ一つに打ち込む作業は，クラスを事前に定義して割り振る作業に比べ手間と時間がかかるプロセスである．にもかかわらず両者に有意差が出ていないのは，テキスト出力の解釈性の高さが良い印象を与えたことが理由の一つであろう．テキストの場合，クラスと異なり，出力が正解と不正解のに明確に二分されることはない．ゆえに，不完全な出力に対して，ユーザは学習がどの程度達成されたかを主観的に判断することが可能である．この出力に対する解釈可能性は，画像分類にはない画像テキスト変換の特性である．

参加者から明示的に指摘されなかったものの，画像テキスト変換システムが抱える問題点として，アノテーションの非効率性が挙げられる．テキストを画像の枚数分打ち込む作業ではなく複数の画像に対して同時にテキストを割り振る作業等が実行可能になれば，効率面は改善されるだろう．

5.4 モデルの性能

学習データの量が学習の成否に大きく関わるものの，図 3 の (c) や (e) を見るに，画像テキスト変換において画像分類よりも学習がうまくいったと判断されている．ラベルの質（図 3 の (a) と (d)）や学習方法の質（(b) と (e)）については有意差が見られないことを加味すると，少なくとも本実験設定においては画像テキスト変換モデルの性能が画像分類のそれより優れたと結論づけられる．

6. おわりに

本研究では，画像テキスト変換をインタラクティブなシステムとして実装し，多様なタスクを用いて初心者ユーザを相手に実験を行った．結果として，テキストによる機械学習モデルのデザインツールは，ユーザの体験としては分

類システムというベースラインに対し有意な差はないが，多様な表現のアノテーションを可能にすることが確認された．

本研究の結果は実験参加者の属性やシステムデザインに大きく影響を受けることが推測されるため，詳細な考察のためには異なる状況下での検証が別途必要であろう．たとえば機械学習の知識を有する母集団を相手にした実験や，既存の分類システムを用いての比較実験などができるだろう．さらに，システムデザインが非効率的なアノテーションを強いる点については改善が必要であろう．テキストを各画像に順次打ち込んでいく作業はユーザの負担が大きいため，UI の工夫による効率化の余地があると考えられる．

参考文献

[1] Dragan Ahmetovic, Daisuke Sato, Uran Oh, Tatsuya Ishihara, Kris Kitani, and Chieko Asakawa. Recog: Supporting blind people in recognizing personal objects. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, p. 1–12, 2020.

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Proceedings of the IEEE conference on European Conference on Computer Vision*, pp. 446–461, 2014.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 248–255. Ieee, 2009.

[4] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11162–11173, 2021.

[5] John J Dudley and Per Ola Kristensson. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, Vol. 8, No. 2, pp. 1–37, 2018.

[6] Jerry Alan Fails and Dan R Olsen Jr. Interactive machine learning. In *Proceedings of the international conference on Intelligent User Interfaces*, pp. 39–45, 2003.

[7] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Conference on Computer Vision and Pattern Recognition Workshop*, pp. 178–178, 2004.

[8] James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. Cueflik: interactive concept learning in image search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 29–38, 2008.

[9] Stephen Gould, Richard Fulton, and Daphne Koller. Decomposing a scene into geometric and semantically consistent regions. In *Proceedings of the IEEE conference on International Conference on Computer Vision*, pp. 1–8, 2009.

[10] Zhengxiong Jia and Xirong Li. icap: Interactive image captioning with predictive text. In *Proceedings of the International Conference on Multimedia Retrieval*, pp. 428–435, 2020.

[11] Hernisa Kacorri, Kris M. Kitani, Jeffrey P. Bigham, and Chieko Asakawa. People with visual impairment training personal object recognizers: Feasibility and challenges. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, p. 5839–5849, 2017.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the IEEE conference on European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

[13] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[14] Yuri Nakao and Yusuke Sugano. Use of machine learning by non-expert dhh people: Technological understanding and sound perception. In *Proceedings of the Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, pp. 1–12, 2020.

[15] Meg Pirrung, Nathan Hilliard, Nancy O'Brien, Artem Yankov, Court D Corley, and Nathan O Hodas. Sharkzor: Human in the loop ml for user-defined image classification. In *Proceedings of the International Conference on Intelligent User Interfaces Companion*, pp. 1–2, 2018.

[16] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018.

[17] Justin Talbot, Bongshin Lee, Ashish Kapoor, and Desney S Tan. Ensemblematrix: interactive visualization to support machine learning with multiple classifiers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1283–1292, 2009.

[18] Qi Wu, Hongping Cai, and Peter Hall. Learning graphs to model visual objects across different depictive styles. In *Proceedings of the IEEE conference on European Conference on Computer Vision*, pp. 313–328, 2014.

[19] Min Yang, Wei Zhao, Wei Xu, Yabing Feng, Zhou Zhao, Xiaojun Chen, and Kai Lei. Multitask learning for cross-domain image captioning. *IEEE Transactions on Multimedia*, Vol. 21, No. 4, pp. 1047–1061, 2018.

[20] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *Proceedings of the IEEE conference on International Conference on Computer Vision*, pp. 1331–1338, 2011.

[21] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

[22] Wei Zhao, Wei Xu, Min Yang, Jianbo Ye, Zhou Zhao, Yabing Feng, and Yu Qiao. Dual learning for cross-domain image captioning. CIKM '17, p. 29–38. Association for Computing Machinery, 2017.

[23] Wentian Zhao, Xinxiao Wu, and Jiebo Luo. Cross-domain image captioning via cross-modal retrieval and model adaptation. *IEEE Transactions on Image Processing*, Vol. 30, pp. 1180–1192, 2020.

[24] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 146–155, 2016.