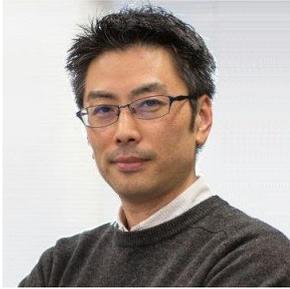


表現スタイルの異なる画像に対して自動適応する物体検出技術

Cross-Style Weakly-Supervised Object Detection



山崎 俊彦 (Toshihiko YAMASAKI, Dr. Eng.)

東京大学大学院 情報理工学系研究科 電子情報学専攻 准教授
(Associate Professor, Department of Information and Communication Engineering, The University of Tokyo)

IEEE ACM IEICE ITE IPSJ JSAI

受賞：前田嵩樹、映像情報メディア学会 2013 年度優秀研究発表賞, 2014/09/01、前田嵩樹、山崎俊彦、相澤清晴, “多クラス分類器の確信度処理による詳細な正誤判定,” 映像情報メディア学会冬期大会, 6-11, Tokyo, Dec. 17-18, 2013. / Toshihiko Yamasaki, International Workshop on Multimedia Signal Processing (MMSp) Top 10% Paper Award, Sep. 22-24, 2014. / IMPS2014 ベストポスター賞, 山崎俊彦, “One-vs-All SVM による多クラス画像認識における認識結果に対する「確信度」解析,” 2014 年度映像メディア処理シンポジウム (IMPS 2014), I-3-16, ラフォーレ修善寺, 伊豆市, 静岡, Nov. 12-14, 2014. / 電子情報通信学会 画像工学研究会 IE 賞, 2016 年 2 月 23 日, 高階 MRF によるクラス間の共起情報を用いたマルチラベル分類の精度改善, 古田諒佑・福島悠介・山崎俊彦・相澤清晴 (東大) / 山崎 俊彦, 本間 拓人, 相澤 清晴, MIRU インタラクティブ発表賞, 2016.08.04.

研究専門分野：マルチメディア パターン認識 機械学習

あらまし 最先端の物体検出モデルの学習には、インスタンスレベル (クラス名+場所) のアノテーションを伴う大規模な画像データセットを必要とするため、自然画像以外の画像への適用は難しい。例えば、スケッチや油絵において物体検出を行う際は、そのドメインでアノテーション付きのデータセットを改めて作成するのが一般的である。本稿では、検出対象のドメインにおいて、画像レベル (クラス) のアノテーションと、自然画像ドメインで学習された物体検出器の統合により、インスタンスレベルの仮想アノテーションを生成する手法を提案する。生成された仮想アノテーションを用いて検出モデルの学習を行うことで、表現スタイルの異なる画像群に対しても精度の良い物体検出器が実現される。筆者らは、検出性能評価用のデータセットを clipart と呼ばれる画像ドメインで新たに構築し、提案手法の有用性を確認した。

1. 研究の目的

物体検出はある画像の中に存在する物体のカテゴリ (クラス) と位置を検出するものである。物体検出は非常に基礎的な問題であり、近年、畳み込みニューラルネットワーク (CNN) の研究の発展により急速に精度の向上が見られている。最先端の物体検出手法 [1]-[5] では、インスタンスレベル (クラス名+場所) のアノテーションを伴う大規模な画像データセットから学習する教師あり学習によって、高い検出精度を実現している。

自然画像における物体検出では、教師あり学習により高精度な物体検出が実現されているが、自然画像以外の画像スタイルにおける物体検出は今まであまり扱われてこなかった。理由としては、インスタンスレベル (クラス名+場所) のアノテーションを伴う大規模な画像データセットを構築するのが非常に難しい事があげられる。具体的な理由としては、著作権等の関係でそもそも対象となるような画像が集めづらいこと、また、アノテーションを付与することにかかるコストが大きいこと等があげられる。

このようなデータセットの不足の元で物体検出を行う代替手法として、画像レベル (クラス名) のアノテーションを伴う画像データセットから物体検出器の学習を行う、弱教師あり学習による手法があげられる [6]-[8]。しかし、弱教師あり学習で生成した物体検出器は、物体の位置の正確な推定が難しいといった問題がある。

筆者らは、インスタンスレベルのアノテーションが利用できず、画像レベルのアノテーションのみが利用可能な、新規画像ドメインで物体検出を行うという新しい課題を取り扱う。より具体的な目的としては、自然画像ドメインで広く使われている物体検出データセットである Pascal Visual Object Classes (Pascal VOC) [9] と同じ物体クラスを検出することとする。この設定は、ある画像スタイル (ソースドメイン) で学習したモデルを他の画像スタイル (ターゲットドメイン) に適用したときでも精度が比較的落ちないようにする、ドメイン適合問題の一種とみなすことが出来る。本稿では、自然画像ドメインをソースドメインとし、新規画像ドメインをターゲットドメインと見なす。

表現スタイルの異なる画像に対して自動適応する物体検出技術

Cross-Style Weakly-Supervised Object Detection

筆者らの手法は、2つの事実に基づいた手法である。(i) 自然画像ドメインで学習された CNN ベースの物体検出器は、異なる画像ドメインでも、ある程度の精度で物体検出が可能である。(ii) ターゲットドメインにおいてインスタンスレベルの仮想アノテーションが付与された画像を用いて、ドメイン適合の為にファインチューニングを行うことで、検出器の精度は大きく向上する。仮想アノテーションの生成は、以下のようにして行われる。本手法では、斉藤らの研究[10]で提案された疑似ラベルの概念を拡張し、疑似バウンディングボックス(疑似 BB)を、画像レベルのアノテーションが付いた画像に対して追加で付与することを試みる。疑似 BB 生成は、ターゲットドメインの画像レベルのアノテーションが付与された画像群と、自然画像ドメインで学習された物体検出器によるそれらの画像への検出結果を統合することによって実行される。生成された疑似 BB と画像レベルのアノテーションを使用して、さらに仮想アノテーションを構成する。筆者らの手法は、ドメイン特有の前処理、特徴抽出、後処理などを使用しないため、任意の画像ドメインにおける物体検出に関して適用可能である。

筆者らは、本手法の妥当性を新たに収集した clipart と呼ばれる画像ドメインに関するデータセット UTclipart を使用して検証した。このデータセットは、検出器の学習の為に使われる、画像レベルのアノテーションの付いた 3,862 枚の画像群 (UTclipart-train) と、手法の精度を検証するために使われる、3,165 個のインスタンスレベルのアノテーションの付いた 1,000 枚の画像群 (UTclipart-test) から構成される。評価指標である mean average precision (mAP) において、既存手法の検出器を単体で適用した場合に 25.3% しかなく、複数の検出器のアンサンブルを用いても 28.1% しか達成していないのに対し、提案手法では 34.5% の mAP を達成した。

筆者らの本稿における貢献は、以下の三点である。

- スタイルを跨いだ物体検出の為にフレームワークを提案した。このフレームワークは、ソースドメインで教師あり学習された物体検出器と、ターゲットドメインの画像レベルのアノテーションのみを用いる。

- 各画像に様々なクラスの複数のインスタンスのアノテーションが付与された。clipart の物体検出評価用データセットを構築した。
- 提案手法は既存手法を大きく上回る性能を示す。

2. 研究背景

2.1 教師あり学習による物体検出

教師あり学習による物体検出としては、R-CNN[11], Fast R-CNN[1], Faster R-CNN[2]のように物体の候補領域からそれぞれ特徴を抽出し、分類を行う手法が最も主流であった。近年では、SSD[3], YOLOv2[4], R-FCN[5]に代表されるように、CNN の 1 回の forward のみで検出を一気に行う手法が主流になりつつある。上記の手法は、全て Pascal VOC や MSCOCO[12]に代表されるように、膨大なインスタンスレベル(クラス名+場所)のアノテーションを伴う大規模な画像データセットが学習の為に必須である。しかし、そのようなデータを収集することは、画像およびクラスの数が増えるにつれてより困難になる。Suらの研究[13]では、作業者がアノテーションを付けるのに 1 インスタンスあたり約 40 秒かかることが報告されている。対照的に筆者らの手法では、検出したいターゲットドメインの画像に対し、インスタンスレベルのアノテーション付与が不必要である。

2.2 弱教師あり学習による物体検出

弱教師あり学習による物体検出器の学習には、画像と画像レベルのアノテーションのペア(各画像内の物体のクラス名は与えられるが、位置情報は与えられない)が必要である。WSDDN[6]とそれに続く研究[7][8]では、物体らしい領域を推定するネットワークとその物体のクラスを推定するネットワークの2つの結果を統合する end-to-end な学習方法を用いることで、高精度な検出を実現している。しかし、弱教師あり学習で生成した物体検出器は、物体の位置を正確に推定することが難しいといった問題がある。

2.3 疑似ラベルとドメイン適合

画像分類では、分類器の予測を組み合わせて、ラベル(クラス名)が付与されていない大量の画像に対し

表現スタイルの異なる画像に対して自動適応する物体検出技術

Cross-Style Weakly-Supervised Object Detection

て `pseudo-label` (疑似ラベル) を付与し、それを用いて分類器の学習を再度行う `self-training` と呼ばれる手法が存在する[14]。この手法では、分類器が高い確信度で正しいと予測した分類は、実際に正しいという仮定を置いている。この仮定は単純であるが、アノテーション付き画像の数が少ない場合には、分類器の性能が大幅に向上することが知られている。`co-training` は、`self-training` を発展させ、2つの分類器の予測を組み合わせて、より正確な疑似ラベルを付与するものである[15][16]。

ドメイン適合とは、画像のドメインによらず上手く予測モデルを構築するための手法である。画像認識のための識別的な学習法は、トレーニングとテストのデータの同じドメインからサンプリングされたものを使用する際には非常に高い識別能を示すが、そうでない時には識別性能が落ちることが知られている。そのため、少量のラベル付きデータしか得られないドメインにおいて識別的な学習モデルを上手く働かせるためにはドメイン適合が不可欠である。

画像分類におけるドメイン適合では、従来は `MMD` を使った手法[17]や `domain classifier network` を使った手法[18]-[21]などが提案されている。`co-training` は、ドメイン適合の手法としても解釈することが出来ることが知られており[22]、斉藤らの研究[10]では、ターゲットドメインの画像に対して `pseudo-label` を付与して識別器を直に学習させることで、ドメイン適合を行う手法を提案しており、この手法は従来手法を大きく上回る性能を示した。

本稿では、斉藤らの研究を拡張し、物体検出におけるドメイン適合問題に対して適用する。ターゲットドメインの画像レベルのアノテーションの付いた画像に対して、`pseudo bounding-box` (`pseudo-BB`、疑似 `BB`) を付与することで、物体検出器をターゲットドメインで直に学習させるためのデータを生成する。

2.4 ドメインを跨いだ物体検出

ある画像ドメインにおいて、事前に全く学習を行わずに物体認識を行うことは非常に難しい。Wilber らの研究[23]では、CNN ベースの物体検出器は、対象画像ドメインでの学習及びファインチューンを行わない場

合、認識性能は著しく低いものであると報告している。

Wu らの研究[24]では、画像ドメインの変化に頑健な物体検出を提案し、さらに、検証用に複数ドメインの画像を含むデータセットを構築している。しかし、ここで提案された手法は、膨大なインスタンスレベルのアノテーションを必要とし、また、検出対象の画像には1インスタンスしか含まれていないという、やや現実的ではない設定下で用いられる手法となっている。

Westlake らの研究[25]では、`people` の1クラスを検出するため、`People-Art` という写真、漫画、41種の異なるスタイルの絵画からなり、インスタンスレベルのアノテーションが付与された画像群が構築された。この研究では、自然画像で教師あり学習された物体検出器を、`People-Art` を用いてファインチューニングすることにより、精度の良い物体検出を実現している。しかし、この研究も膨大なインスタンスレベルのアノテーションを必須としている。

本稿は、ドメインを跨いだ物体検出を、ターゲットドメインにおいては画像レベルのアノテーションのみを用いて実現した初めての論文である。筆者らは、各画像に複数クラス・複数インスタンスがアノテーションとして付与されているデータセットを構築したが、このデータセットは、提案手法の評価のためにしか用いられていない。

3. データセット

本稿の目的は、自然画像ドメインで教師あり学習された物体検出器をドメイン適合させることで、自然画像以外の対象画像ドメインで同じクラスを検出する事である。自然画像においては、大規模なインスタンスレベルのアノテーション付きのデータセットが利用可能であるため、自然画像ドメインで学習された物体検出器は、容易に手に入るものとしている(配布されている訓練済みの検出器を使用して学習はスキップできる)。本稿では、ターゲットドメインとして `clipart` と呼ばれる画像ドメインを用い、`Pascal VOC` に含まれる20クラスを検出対象とした。`clipart` はベクターグラフィックス、絵画、スケッチ等、様々なドメインを含んでいる。本稿で使用されているすべての `clipart` 画像は、`Openclipart` (<https://openclipart.org/>) 及び

表現スタイルの異なる画像に対して自動適応する物体検出技術

Cross-Style Weakly-Supervised Object Detection

Pixabay (<https://pixabay.com/>) というサイトから CC0 のものだけを収集し、さらに、CMPlaces[26] というデータセットからも収集するという形で構築された。

3.1 UTClipart-train

20 クラスそれぞれのクラス名を、クエリとして検索することで画像を収集した。収集された画像のほぼ全ては、1 画像につき 1 つのクラスしか含まれていなかった。筆者らは、対象となるクラスが画像内に無いもの、複数のクラスを 1 画像中にもつものを手動で削除した。これにより、作成したデータセットには、1 つのクラスのみが含まれる画像のみから構成される。ただし、同一クラスで複数のインスタンスを持つ場合は有り得る事を明記しておきたい。結果として、画像レベルのアノテーションが付与された 3,862 枚の画像群 (UTClipart-train) を収集した。UTClipart-train の例を図 1 に示す。

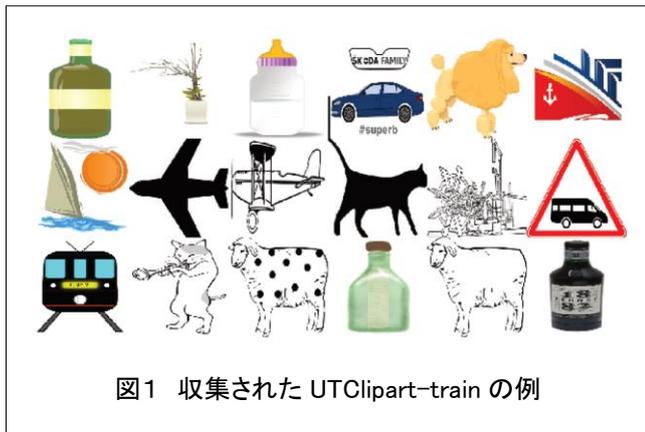


図1 収集された UTClipart-train の例

インスタンスしか含まない Photo-Art[24] より複雑で、検出難易度の高いデータセットである。収集された画像例を図 4 に示す。

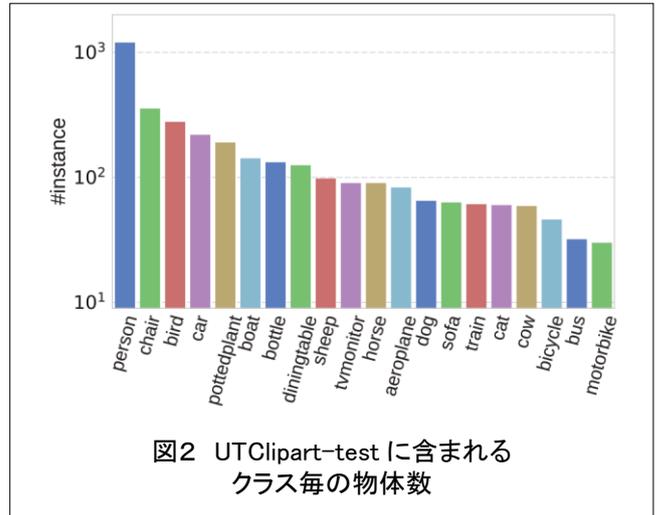


図2 UTClipart-test に含まれるクラス毎の物体数

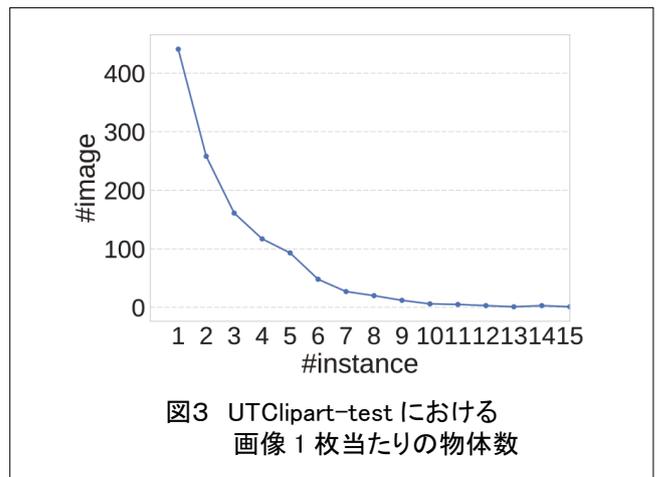


図3 UTClipart-test における画像 1 枚当たりの物体数

3.2 UTClipart-test

筆者らは、CMplaces で用いられている 205 種のシーンを表すクラス (例: pasture) のクエリを使用して画像を収集した。検出対象の 20 種のクラスのいずれか 1 つ以上を含む各画像について、インスタンスレベルのアノテーションを付与した。結果として、インスタンスレベルのアノテーションが 3,165 個含まれる 1,000 枚の画像群 (UTClipart-test) を収集した。UTClipart-test におけるクラス毎のインスタンス数を図 2 に、1 画像あたりに含まれるインスタンス数を図 3 にそれぞれ示す。UTClipart-test は、1 画像に 1 イ



図4 収集された UTClipart-test の例

表現スタイルの異なる画像に対して自動適応する物体検出技術

Cross-Style Weakly-Supervised Object Detection

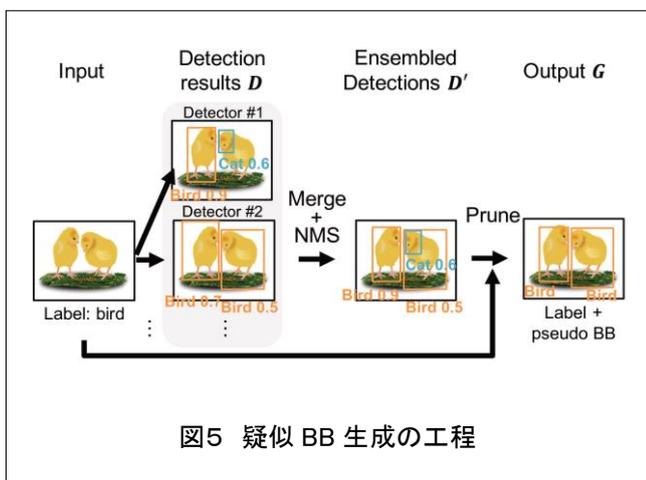
4. 提案手法

提案手法は、疑似 BB 生成と教師あり物体検出器のファインチューニングの 2 つのパートからなる。

(1) 疑似 BB 生成

$x \in \mathbb{R}^{H \times W \times 3}$ を画像とする。この時 H と W は、その画像の高さと幅をそれぞれ表す。 C は検出対象の物体クラスの集合を、 z は画像レベルのアノテーション、すなわち、画像 x に含まれている物体クラスの集合をそれぞれ表すとする。本稿で使用する UTClipart-train では、 z には必ず一つのクラスしか含まれていないが、一般には、 z には複数のクラスが含まれるものとする。このプロセスの目標は、仮想インスタンスレベルのアノテーション G を画像 x に対して生成することである。 G は、 b をバウンディングボックス、 $c \in C$ として $g = (b, c)$ からなる。

疑似 BB を生成する全工程を図 5 に示す。まず始めに、 N 個の異なる物体検出器を x に対して適用し、検出結果 $D = \{D_1, D_2, \dots, D_N\}$ を得る。 D_i は、それぞれの検出 $d = (p, b, c)$ からなる。ここで $c \in C$ であり、 p は b がクラス c である確率である。ここでは、自然画像で教師あり学習された物体検出器だけでなく、ターゲットドメインで弱教師あり学習された物体検出器を用いることも出来る。



次に、全ての D を単にあわせてうえて、冗長な検出を除くため non-maximum suppression (NMS) を行った後に残った検出結果の集合 D' を得る。NMS の詳

細は、文献[27]等に詳細に記述されているため、ここでは省略する。

次に、 p の大きい順に並び替えられた検出 $d = (p, b, c) \in D$ について、もし、 c が z に含まれる時、その d を正しい検出とみなし、 (b, c) を仮想インスタンスレベルのアノテーションの集合 G に対して加える。この操作を、全ての x から生成される G に含まれる検出の合計が定数 T に達するまで繰り返す。すなわち、提案手法では、データセット全体に対する検出のうち、最も確信度の高い T 件のみを仮想アノテーションとして採用する。

(2) 教師あり物体検出器のファインチューニング

画像 x と仮想アノテーション G のペアを用いて、教師あり学習用の物体検出器をファインチューニングする。検出器の初期パラメータとしては、自然画像で学習済みの同一モデルの検出器のパラメータをコピーして用いる。

パラメータ T は、以下に示す手順に基づいて設定した。仮に、完全な疑似 BB 生成を行うことができれば、 G のインスタンス数は実際に画像に含まれるインスタンス数に等しい。そこで、提案手法では、 T は G に含まれるインスタンス数に等しいと仮定する。本稿では、UTClipart-train を用いているが、UTClipart-train の各画像には 1 つのクラスが含まれず、かつ、含まれるインスタンスの数はほぼすべての画像で 1 である。従って、UTClipart-train に真に含まれるインスタンス数は、UTClipart-train の画像数 N と同じであると見積もることが可能である。この観測に基づいて $T \approx N$ という近似を行うことで、最終的に T を設定することが出来た。一般的には、それぞれの x に対応する z のクラス数を合計して T を得ることが出来る。

5. 実験結果

5.1 実装と評価の詳細

提案手法の妥当性を検証するための実験を 3 章で構築したデータセットを用いて行った。ファインチューニングする物体検出器としては、SSD[3]を使用した。ファインチューニングは、学習率 10^{-5} で 5,000 イテレーションを行った。この学習率は、自然画像で SSD を

表現スタイルの異なる画像に対して自動適応する物体検出技術

Cross-Style Weakly-Supervised Object Detection

学習する時の最後の学習率と同じである。疑似 BB の生成の精度を評価するため、UTClipart-train から 200 枚の画像を抽出し、インスタンスレベルのアノテーションを付与した。疑似 BB と付与されたアノテーションのバウンディングボックスの intersection over union (IoU) が 0.5 より大きい場合、疑似 BB は正しいとみなされ、それより小さい場合には、疑似 BB は正しくないとみなされる。評価指標としては、各クラスでの F1 (=precision と recall の調和平均) の平均である mF1 を使用した。提案手法の検出結果の評価に関しては、average precision (AP) とその平均値 mean AP (mAP) を用いた。UTClipart-test に対して、それぞれの手法を用いて検出を行ない、その結果の AP と mAP を比較した。

5.2 結果

表 1 に、提案手法と対抗手法によって得られた物体検出モデルを UTClipart-test に適用した際の結果を示す。比較した手法は以下の 2 つである。

(1) 単体の物体検出器

既存の物体検出器をそのまま用いた場合について検証した。2.1 章と 2.2 章で述べたように、様々な教師あり/弱教師あり物体検出の手法があるが、筆者らは、その中で最先端のモデルを幾つか選択し比較した。SSD[3]と YOLOv2[4]を教師あり学習による物体検出器として選択した。この検出器は、自然画像ドメインで学習済みである。ContextLocNet (CLNet) [7]を弱教師あり学習による物体検出器として選択した。

この検出器の学習は、UTClipart-train を用いて学習した。それぞれの検出結果は、表 1 の 4 列目に示されている。単体での検出精度は、SSD が最も高い結果となった。

(2) 複数の物体検出器のアンサンブル

画像分類においては、複数の分類器の予測結果を単に平均する (アンサンブル) と、分類の精度が向上することが知られている[28]-[30]。物体検出についても、似た処理を行うことで検出精度が向上する。単体の物体検出器による検出結果を単に足し合わせたうえで、改めて non maximum suppression (NMS) を適用することで、検出におけるアンサンブルを行う。NMS を適用する際のパラメータは、元の SSD や YOLO で用いられる物と同じ値を用いる。複数モデルをアンサンブルした検出結果は、表 1 の 5 列目に示されている。モデルをアンサンブルするほど、検出性能は向上することが確認された。提案手法の結果を表 1 の 6 列目に示す。提案手法は、疑似 BB を生成する際にベースとして組み合わせる物体検出器によらず、精度を改善することが確認された。しかし、疑似 BB 生成の際に弱教師あり学習による検出器である CLNet の結果を組み合わせると、mAP の改善度合いが小さくなることも確認された。図 6 に、UTClipart-train の画像に対して生成された仮想アノテーションの例を示す。図 6a や図 6b では、生成が非常に上手く出来ているが、図 6c のように全く仮想アノテーションが付与されなかったり、図 6d のように物体の一部だけが囲われた疑似 BB が生成されたりするケースも存在する。

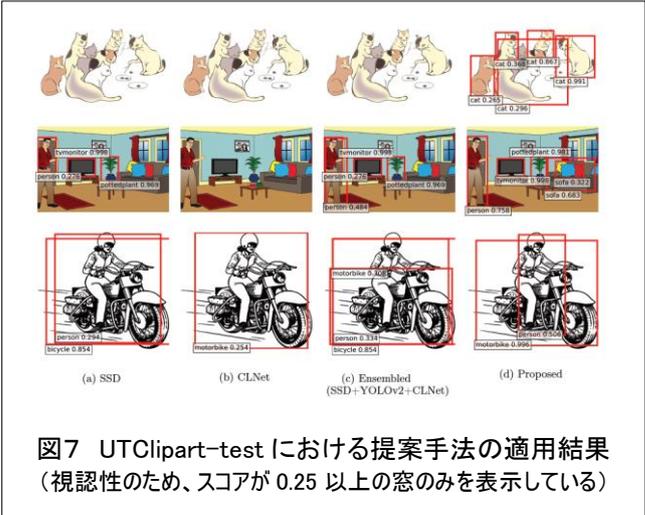
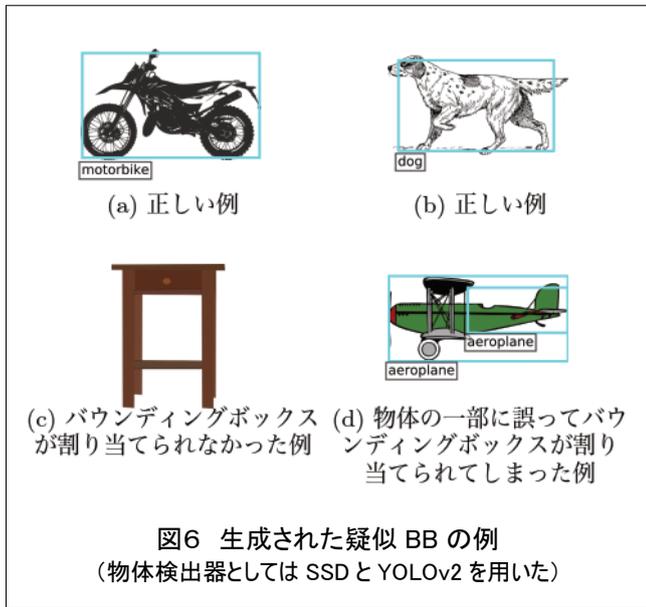
表 1 UTClipart-test における物体検出結果の比較と疑似 BB 生成結果の比較

Combination of detectors			Baseline methods		Proposed method	
FSD		WSD	Single detector mAP[%]	Ensembled detectors mAP[%]	Finetuned on pseudo-BB mAP[%]	Pseudo-BB mF1[%]
SSD [3]	YOLOv2 [4]	CLNet [7]				
✓			25.3	-	33.8	57.4
	✓		22.5	-	32.1	51.5
✓	✓		-	26.8	34.5	59.7
		✓	10.8	-	24.5	66.6
✓		✓	-	27.1	31.4	72.3
	✓	✓	-	25.0	29.4	66.6
✓	✓	✓	-	28.1	30.4	71.0

表現スタイルの異なる画像に対して自動適応する物体検出技術

Cross-Style Weakly-Supervised Object Detection

図 7 に、実際の検出結果を示す。提案手法は、様々なスタイルの画像について有効であることが示唆される。



6. まとめ、将来展望

筆者らは、画像スタイルを跨いだ物体検出手法を提案した。筆者らの手法を評価するため、画像レベルのアノテーションの付いた 3,862 枚の画像 (UTClipart-train) と、インスタンスレベルのアノテーションの付いた 1,000 枚の画像 (UTClipart-test) を、新たに clipart と呼ばれるドメインの画像で構築した。提案手法は、全ての既存手法を上回ることが確認された。今後の展望としては、自然画像以外の様々なドメインの画像の大規模データセットである BAM![23]を用いて、漫画や水彩画など、より広範な画像ドメインで提案手法の適用可能である。また、本稿で提案した仮想アノテーションを用いることで、半教師あり学習[32]と呼ばれる少量のインスタンスレベルのアノテーションと、多量の画像レベルのアノテーションから物体検出器を学習する手法への適用も検討している。

(3) 物体クラスに注目した性能解析

表 2 に、単体の物体検出器、複数の検出器のアンサンブル、提案手法のそれぞれを用いた UTClipart-test における検出結果を示す。表中の 'Proposed' は、疑似 BB 生成の際に用いる検出器としては、SSD と YOLOv2 を用いたものである。提案手法は、chair 以外の全てのクラスにおいて最高精度を達成し、特に、元々検出するのが難しい cat, dog, bottle 等でも改善を示した。

表2 Cliaprt-test における各クラスの検出結果の AP[%]

method	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
SSD [3]	25.3	16.2	57.3	16.7	11.4	7.9	39.9	31.9	4.0	35.3	18.6	21.3	9.3	20.0	49.5	41.9	32.6	7.3	28.7	32.6	23.8
YOLOv2 [4]	22.5	16.1	55.9	13.8	5.1	4.6	45.5	24.4	4.4	34.7	10.7	20.0	4.2	17.3	49.2	31.8	35.2	1.9	18.9	34.6	20.9
CLNet [7]	10.8	3.6	29.6	5.3	4.1	3.9	53.5	9.0	1.0	0.1	10.6	1.4	1.4	8.5	56.3	1.4	1.7	4.5	7.9	9.4	3.5
Ensembled (SSD+YOLOv2+CLNet)	28.1	16.8	61.0	21.7	11.7	6.6	50.6	33.1	5.4	38.5	30.9	22.5	9.5	22.0	55.8	42.4	34.8	9.7	30.1	34.5	24.2
Proposed	34.5	21.6	70.9	23.6	12.1	18.7	60.8	39.0	6.6	36.1	32.7	29.3	12.5	26.1	85.4	58.0	37.2	9.7	34.4	40.6	33.6

'Ensembled' は SSD, YOLO, CLNet をアンサンブルしたものであり、'Proposed' はファインチューニング用の疑似 BB 生成の際に用いる検出器としては SSD と YOLOv2 を用いたものである

表現スタイルの異なる画像に対して自動適応する物体検出技術

Cross-Style Weakly-Supervised Object Detection

参考文献

- [1] Ross Girshick. Fast R-CNN. In ICCV, 2015.
- [2] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.
- [3] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In ECCV, 2016.
- [4] Joseph Redmon and Ali Farhadi. YOLO9000: Better, Faster, Stronger. arXiv preprint arXiv:1612.08242, 2016.
- [5] Yi Li, Kaiming He, Jian Sun, et al. R-FCN: Object detection via region-based fully convolutional networks. In NIPS, 2016.
- [6] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In CVPR, 2016.
- [7] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. ContextLocNet: Context-aware deep network models for weakly supervised localization. In ECCV, 2016.
- [8] Ke Yang, Dongsheng Li, Yong Dou, Shaohe Lv, and Qiang Wang. Weakly supervised object detection using pseudo-strong labels. arXiv preprint arXiv:1607.04731, 2016.
- [9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. IJCV, Vol. 88, No. 2, 2010.
- [10] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. arXiv preprint arXiv:1702.08400, 2017.
- [11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.
- [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 ECCV, 2014.
- [13] Hao Su, Jia Deng, and Li Fei-Fei. Crowdsourcing annotations for visual object detection. In AAAI workshop, 2012.
- [14] Xiaojin Zhu. Semi-supervised learning literature survey. 2005.
- [15] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In COLT, 1998.
- [16] Jafar Tanha, Maarten van Someren, and Hamideh Afsarmanesh. Ensemble based co-training. In 23rd Benelux Conference on Artificial Intelligence, 2011.
- [17] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. Journal of Machine Learning Research, Vol. 13, No. Mar, 2012.
- [18] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. arXiv preprint arXiv:1409.7495, 2014.
- [19] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In ICML, 2015.
- [20] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. Journal of Machine Learning Research, Vol. 17, No. 59, 2016.
- [21] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In NIPS, 2016.
- [22] Minmin Chen, Kilian Q Weinberger, and John Blitzer. Co-training for domain adaptation. In NIPS, 2011.
- [23] Michael J Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge Belongie. BAM! the behance artistic media dataset for recognition beyond photography. arXiv preprint arXiv:1704.08614, 2017.
- [24] Qi Wu, Hongping Cai, and Peter Hall. Learning graphs to model visual objects across different de pictive styles. In ECCV, 2014.
- [25] Nicholas Westlake, Hongping Cai, and Peter Hall. Detecting people in artwork with cnns. In ECCV workshop, 2016.
- [26] Lluís Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning aligned cross-modal representations from weakly aligned data. In CVPR, 2016.

表現スタイルの異なる画像に対して自動適応する物体検出技術

Cross-Style Weakly-Supervised Object Detection

- [27] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. TPAMI, Vol. 32, No. 9, 2010.
- [28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.
- [29] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In CVPR, 2015.
- [30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.
- [31] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In ECCV, 2012.
- [32] Ziang Yan, Jian Liang, Weishen Pan, Jin Li, and Changshui Zhang. Weakly-and semi-supervised object detection with expectation-maximization algorithm. arXiv preprint arXiv:1702.08740, 2017.

この研究は、平成25年度SCAT研究助成の対象として採用され、平成26～28年度に実施されたものです。