計算工学講演会論文集 Vol. 28 (2023 年 5 月)
計算工学会

データセット作成における自己教師あり学習の有効性の検討

Investigating the effectiveness of self-supervised learning in dataset creation

塩崎雄晴 [1] 入江寿弘 [2] 小林伸彰 [3] 新宮清志 [4]
Yusei Shiozaki, Toshihiro Irie, Nobuaki Kobayashi, and Kiyoshi Shingu

[1]日本大学大学院 理工学研究科 精密機械工学専攻 大学院生
(〒 274-8501 千葉県船橋市習志野台 7-24-1, csys21001@g.nihon- u.ac.jp)
[2]博士 (工学) 日本大学 理工学部 精密機械工学科 教授 (同上, tirie@eme.cst.nihon-u.ac.jp)
[3]博士 (工学) 日本大学 理工学部 精密機械工学科 准教授 (同上, kobayashi@eme.cst.nihon-u.ac.jp)
[4]工学博士 日本大学 名誉教授 (〒 102-0072 東京都千代田飯田橋 3-10-1-1901, kshingu@ocean.cst.nihon-u.ac.jp)

Self-supervised learning(SSL) is a hopeful way to obtain features of target data with minimal human supervision in the future. There are well-known extensive datasets like COCO or ImageNet in the world that we can use to train the models. However specific manual annotation tasks have been required to make unique datasets. Therefore, we developed an application to make a dataset and present the efficiency of an SSL-based annotation that can learn features without ground truth in advance.

Key Words : *ML/AI, CV, SSL, App, Annotation, Dataset*

1. はじめに

データセット作成作業は膨大な単純作業と時間の消費を伴うため，教師あり学習における最も避けたい作業の一つである．このために以下にして効率化するか，一人当たりの作業量を減らすか等，近年に至るため様々な手法が提案されてきた．半教師あり学習や自己教師あり学習もその一つであり，教師データ作成作業を伴うことなく，入力データの特徴量を効率的に取得できる．現在では，COCO [1] や ImageNet [2] 等の大規模データセットは既に存在しており，これらを用いた学習済みモデルを使用することで下流タスクではより高いスコアを得ることができる．しかし，食品の傷みなどを判別するような大規模データセットにない特定の条件ではいまだに自身で教師データを追加してデータセット作成する必要がある．そこで，我々は自己教師あり学習を応用して事前に任意の入力データを分類することで人間のタスクを最小限に抑える手法を提案する (図- 1)

2. 背景

教師あり学習は学習手法の一つであり，多くは手作業で作成されたデータセットを参照して，正例を正しく認識できるモデルを作るために行われる．そこで，我々はこの参照データを作成することをアノテーションと呼び，モデルに識別して欲しい情報をデータに含めることを指している．特に画像領域ではアノテーションには画像に含まれる対象の大きさ，位置の指定が必要とされ，これらを一枚一枚手作業で行うことは膨大な時間と体力を要するとされている．人海戦術的に解決することも可能だが，人的資源や予算に限りがある場合は現実的でなく，この点で研究開発に遅れがでる可能性があり早急な改善が求められている．我々が提案するアプリケーション及び手法では事前にクラスターとして分類された画像群を人間が一括で名付けを行う．これによってアノテーションに要する作業量は入力データ量が多ければより効果を示すことが期待できる．

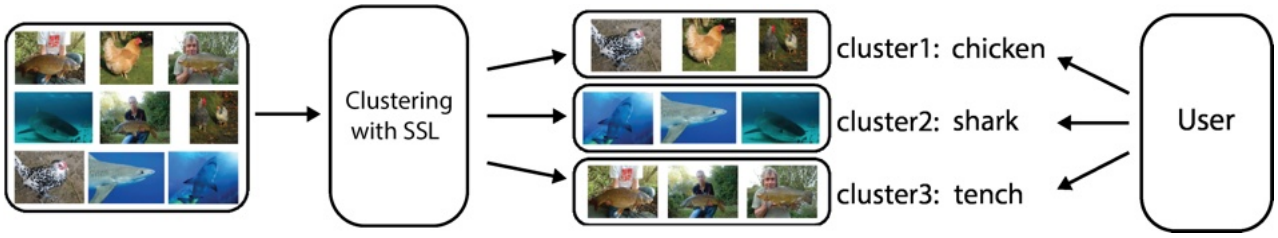
3. 関連研究

(1) 作業工程の改善

クラウドソーシングを利用し複数人でデータセット作成することで一人当たりの負荷を軽減する手法が提案されているが，作業人間のバウンディングボックスのバラツキなどが問題とされている．この問題に対処するためにアノテーション作業を三つの工程に分割することでアノテーションの精度を改善している [3].

(2) ツールの改善

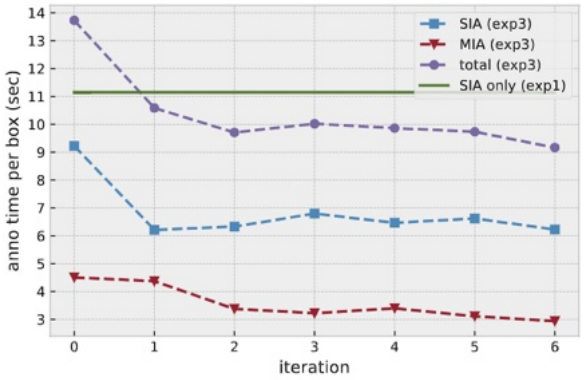
事前学習済みモデルを使用し，事前にバウンディングボックスを作成することでアノテーション作業を効率化する研究は数多くあり，この事前に作成したバウンディングボックスの正負をユーザーが指定することでさらなる作業の効率化に成功している [4]. 加えて，オブジェクトの中心点をクリックすることで物体の位置形状を推定しバウンディングボックスを生成する手法も提案されている [**?**].
LOST [5] はアノテーション作業負荷を軽減するために開発されたツールであり，Multi Image Annotation(MIA) に対応している．MIA とは，事前に画像群をまとめて分類しておくタスクのことであり， これによる作業者の作業時間を 図-2 に示す． しかし，このツールでは前もって各画像にアノテーションを行うことが必要であり，この作業は前述の通り最も削減した工程である．この工程のことを Single Image Annotation(SIA) と呼ぶ.



図–1 概要

(3) 手法やモデルの改善

自己教師あり学習ではアノテーションをすることなく良い特徴量の学習ができる．これを事前学習とすることで様々な下流タスクでより良い結果を期待できる．この特徴を利用することで我々がアプリケーションで達成したい機能を得られると期待できる．そこで，我々はこれら自己教師あり学習モデルである SwAV [6] と iBOT [7] について，比較検討を行った．



図–2 Time per box

4. 実験

SwAV と iBOT を使用した際のスコアを比較する．スコアには教師なしでのクラスタリングでは NMI, ARI, FMI を使用し，教師ありでの分類では学習させたデータの割合毎に精度を求め，比較を行った．その実験結果を表-1 から表-6 に示す．各クラスターに割り当てられたデータラベルを 図-3，図-4 に示す．

(1) クラスタリング

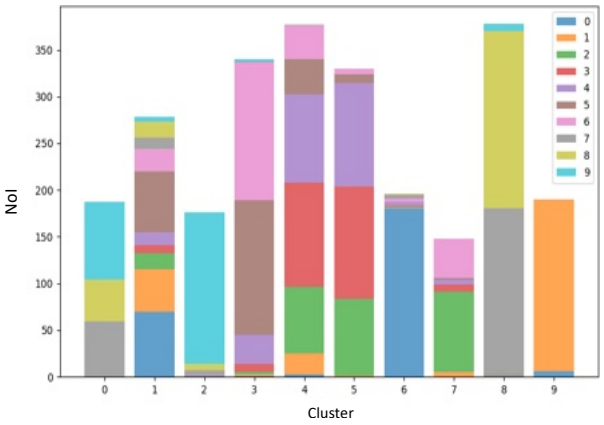
クラスタリング評価の結果を表-1 に示す．SwAV, iBOT はそれぞれ NMI において 57.9%， 71.0%を示した．これらのモデルには訓練済みのパラメータを使用した．加えて，ImageNet でのクラスタリング結果を図-3, 4 に示す．これらは各クラスターに含まれる種類が少ない方が良い結果である．これらの図を見ると，NMI の高い iBOT の方がより良くクラスタリングができていることが視覚的にも理解できる．

(2) 画像認識 (MNIST)

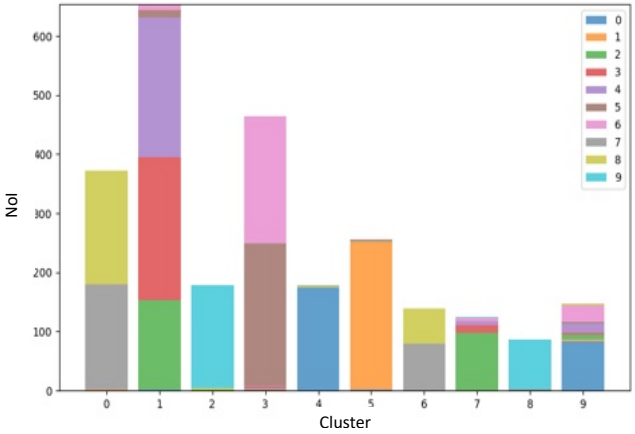
次に，MNIST を使用してデータ情報が少ない場合の SwAV と iBOT の半教師あり学習の精度を比較する．そ

表–1 クラスタリングスコア

Method	Arch	NMI [%]	ARI [%]	FMI [%]
SwAV	RN50	57.88	52.45	42.16
iBOT	ViT16/S	71.00	70.80	57.74



図–3 各クラスターに含まれるラベル (SwAV)



図–4 各クラスターに含まれるラベル (iBOT)

の実験結果を表-3 と表-4 に示す．実際の環境では事前に入力データのカテゴリ分布を揃えて入力することを想定していない．その状況を再現するためにランダムに取得したデータを入力として使用した場合とそうでない場合について実験を行った．本実験で使用したシード値は 251 とした．この場合，不均衡データの学習について考慮しなければならない点に注意する．

図–5 使用したラベルと画像

表–2 使用したラベル ID の対応表

Label	Category
n01440764	tench, Tinca tinca
n01443537	goldfish, Carassius auratus
n01484850	great white shark, white shark, man-eater, man-eating shark, carcharodon carcharias
n01491361	tiger shark, Gakeocerdo cuvieri
n01494475	hammerhead, hammerhead shark
n01496331	electric ray, crampfish, numbfish, torpedo
n01498041	stingray
n01514668	cock
n01514859	hen
n01518878	ostrich, Struthio camelus

表–3 MNIST での SwAV の各抽出割合における精度

Extraction	Epoch	1% Label	5% Label	10% Label	50% Label
Not Random	20	96.67	99.79	99.51	99.95
Not Random	200	100	100	100	100
Random Seed 251	20	57.81	93.72	97.12	99.10
Random Seed 251	200	78.93	94.42	97.61	99.31

表–4 MNIST での iBOT の各抽出割合における精度

Extraction	Epoch	1% Label	5% Label	10% Label	50% Label
Not Random	20	10.87	92.69	44.71	91.86
Not Random	200	75.77	99.24	97.71	99.27
Random Seed 251	20	9.79	20.41	92.90	97.50
Random Seed 251	200	41.42	96.76	98.71	99.54

表–5 ImageNet での SwAV の各抽出割合における精度

Extraction	Epoch	1% Label	5% Label	10% Label	50% Label
Not Random	20	52.58	64.00	77.77	92.31
Not Random	200	51.27	68.73	79.19	92.50
Random Seed 251	20	44.46	61.69	75.62	92.39
Random Seed 251	200	44.46	65.16	78.42	93.12

表–6 ImageNet での iBOT の各抽出割合における精度

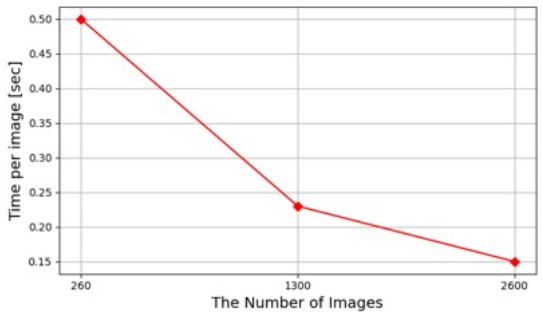
Extraction	Epoch	1% Label	5% Label	10% Label	50% Label
Not Random	20	38.96	65.42	79.65	87.35
Not Random	200	58.16	72.23	81.69	93.58
Random Seed 251	20	21.08	70.73	77.35	85.08
Random Seed 251	200	34.69	74.85	79.15	92.96

(3) 画像認識 (ImageNet)

続いて，ImageNet に含まれる 10 クラスを抽出して半教師あり学習の評価を行った．その実験結果を表-5, 6 に示す．

(4) アノテーション作業時間

ImageNet から取り出した 10 クラス 2600 枚のデータセットを下に，一枚あたりのアノテーション作業時間を計測した．この時のデータセットに使用したラベル群と画像例をそれぞれ表-2，図-5 に示す．しかし，取り出したラベルにおいて，識別が困難なものがあったため，n01484850 と n01491361 を”shark”, n01496331 と n01498041 を”ray”, n01514668 と n01514859 を”chicken”として扱う．よってラベルとしては全 7 種として実験を行った．この実験結果を図-6 に示す． この実験では iBOT のアーキテクチャをベースに最終層に k-means を使用したものを使用した．実験では画像 260 枚の作業時間は 129.9 秒であり，一枚あたりの作業時間は 0.5 秒となる．また，画像 1300 枚での作業時間は 294.7 秒，一枚あたりでは 0.23 秒,．画像 2600 枚では一枚あたり 0.15 秒という結果となった．



図–6 Time per image

5. 終わりに

本研究では，SwAV と iBOT を使用してデータセット作成における自己教師あり学習の有効性を検討した．iBOT でのクラスタリングでは NMI が 71%と高い値を示し，SwAV と iBOT は半教師あり学習において両モデルとも MNIST,ImageNet で 90%を超える高い精度を示した．さらに，SwAV は特に MNIST では 20 エポック時点では iBOT よりも高い精度を示している．このことから，SwAV は情報の少ない画像の学習において，半教師あり学習で使用する場合には一般に精度が高いことで知られている ViT モデルである iBOT よりも早く良い結果を得られると言える．また，ImageNet においても 50%ラベルにおいては 20 エポック時点では SwAV が iBOT よりも高い精度を示した．

また，ImageNet では 90%を超えるような高い精度を得るには 50%以上のデータが必要であることが，5, 6 よりわかる．しかし，iBOT の学習にはより多くの時間が必要になる．そのため，クラスタリングにおいては iBOT のような ViT モデルを使用し，作成したデータを用いた半教師あり学習では SwAV のような CNN モデルを使用することとした．実際にクラスタリングを行っ多者に対し作業者がラベル付けを行った場合にかかった時間は画像 2600 枚に対して一枚あたり 0.15 秒という結果となった (図-6).

本研究の目標はアノテーションにおける SIA プロセスを撤廃することであるため，実際には物体検出のために領域候補の予測が必要になる．特に教師なしでの領域候補の予測については 2022 年に発表があった [8]．また，生成モデルの学習において，物体の位置空間を含んだ情報を学習することで物体の位置予測が可能になることが示唆されているため，これらの手法について検討し研究を進めていく [9]．特に，大規模モデルの使用については限られた資源における実際の運用について詳しく検証を重ねる必要がある．

参考文献

[1] COCO - Common Objects in Context.
[2] ImageNet.
[3] Hao Su, Jia Deng, and Li Fei-Fei. Crowdsourcing Annotations for Visual Object Detection. Technical report.
[4] Dim P. Papadopoulos, Jasper R. R. Uijlings, Frank Keller, and Vittorio Ferrari. We don’t need no bounding-boxes: Training object class detectors using only human verification. feb 2016.
[5] Jonas Jäger, Gereon Reus, Joachim Denzler, Viviane Wolff, and Klaus Fricke-Neuderth. LOST: A flexible framework for semi-automatic image annotation. oct 2019.
[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. jun 2020.
[7] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT Pre-Training with Online Tokenizer. nov 2021.
[8] Amir Bar, Xin Wang, Vadim Kantorov, Colorado J Reed, Roei Herzig, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. DETReg: Unsupervised Pretraining with Region Priors for Object Detection. Technical report.
[9] Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. SPACE: Unsupervised Object-Oriented Scene Representation via Spatial Attention and Decomposition. jan 2020.