屋外画像に対する基盤モデルの汎化性能に関する考察

八千代エンジニヤリング㈱　正会員　○藤井 純一郎，高橋 悠太，吉田 龍人
八千代エンジニヤリング㈱　非会員　岡野 将大

１．教師あり画像認識の課題

社会インフラの点検や巡視において，画像認識の適用が進められている．例えば，コンクリートのひび割れをSemantic Segmentationで検出した事例や，河川の不法投棄をObject Detectionで検出した事例など，多くの事例が報告されている．これらの手法は，検出対象をアノテーションした教師データを学習する教師あり学習により実現するが，検出対象のバリエーションを網羅した大量の教師データが必要となる．一般的に学習時の教師データのバリエーションが十分でないと，学習していない特徴量を持つ画像が運用時に出現した際，大幅に精度が下がる．

画像認識の実用化で先行する製造業や医療分野に比べて，屋外で自然環境にさらされることが基本の土木分野では，下記の理由から画像特徴量のばらつきが大きく，それを網羅する教師データを作成することが難しい．

(1)　画像全体の特徴量のバリエーション

季節や天候により色味や植生繁茂状況が変わり，同じ地点・アングルであっても画像全体の特徴量がばらつく

(2)　背景の特徴量のバリエーション

現場ごと・地域ごとに環境が異なり，検出対象が同じであっても背景の特徴量がばらつく

(3)　検出対象の特徴量のバリエーション

土木分野で検出すべき対象は多種多様で，発生頻度が低い事象が多く存在する

教師あり学習では，これら３つのバリエーションの組み合わせを網羅した教師データセットを用意することが理想だが，それだけの画像を収集するには膨大な作業と時間を要する．そのため，土木分野での画像認識の適用は，実証実験や特定の地点では有効に機能するが，実用化するためには汎化性能の向上が課題として残されている．

２．課題の解決に向けた仮説

近年は特定の目的のための専用の教師データを自前で用意するのではなく，公開データセットやインターネット上の膨大なデータを学習した基盤モデルが注目を集めている．画像よりも特徴量空間が限定される自然言語処理や音声の基盤モデルが先行し，画像やマルチモーダルモデルにも拡大している．

先述の課題のうち，(1)(2)は基盤モデルでは汎化性能を獲得している可能性が高いと考えた．基盤モデルを用いることで(1)(2)に対応した教師データのバリエーションを確保する労力を低減できれば，(3)の土木特有の検出対象の教師データ収集に注力できるようになり，屋外でも画像認識の適用可能性が広がる．

３．空撮画像を用いた不法投棄物検出実験

4.1. 実験概要

本研究では，土木分野の教師データのバリエーション不足の課題を，基盤モデルの汎化性能により解決する可能性を探るため，不法投棄物検出を例に基盤モデルの一つであるCLIPによるZero-shot画像分類を試行した．CLIPの汎化性能を確認するため，画像の明るさや背景といった特徴量を変えた画像で，分類結果がどのように変化するか実験した．

4.2. 基盤モデルCLIP

CLIP (Contrastive Language-Image Pre-training)[1)]は，テキスト(キャプション)と画像のペアからなる，大規模な教師データで学習された画像分類モデルである．また，OpenAIから発表されて以降もオープンソースプロジェクト(Open CLIP[2)])として研究が続けられている．大規模データで学習したCLIPは，多様なデータ表現を保持していると考えられ，さまざまなタスクに合わせて追加の学習をすることなく利用が可能である(Zero-shot transfer)．

CLIPのZero-shot画像分類では，教師データを作成することなく，学習済みのモデルに対して検出対象物をテキストで指定することにより，その物体が映っている確率が返される．本研究の場合，不法投棄物として検出したい複数の物体のクラスを指定し，各画像がどのクラスに該当するかを，確率が高い順に出力した（図-1）．

本研究では，予備実験で不法投棄物検出において最も高精度だったOpen CLIPのLAION-2B ViT-bigG-14を採用した．本モデルは，事前に25億ペアのデータセットで学習されており，一般公開されている．[https://github.com/mlfoundations/open_clip]

4.3. 実験データ

土・芝・砂利と背景の異なる複数の地点に不法投棄物（ゴミ袋）を配置し，ドローンで空撮を行った．土の上にゴミ袋が映った画像については，元画像に加えて輝度・彩度・コントラストを変更して色情報を加工した画像，色情報を落としたグレースケール画像，色情報を逆転させた色反転画像を用いた．

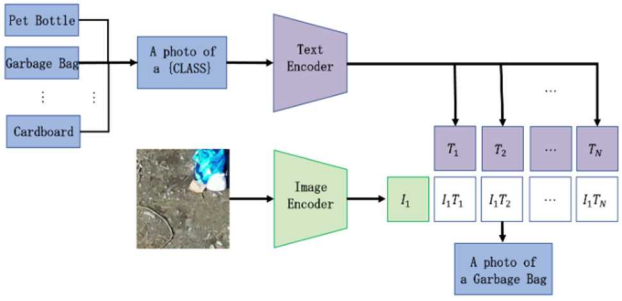

図-1　CLIP ゼロショット推論

キーワード　人工知能，深層学習，画像分類，基盤モデル
連絡先　〒111-8648　東京都台東区浅草橋 5-20-8 CS タワー　八千代エンジニヤリング㈱　TEL:03-5822-6626

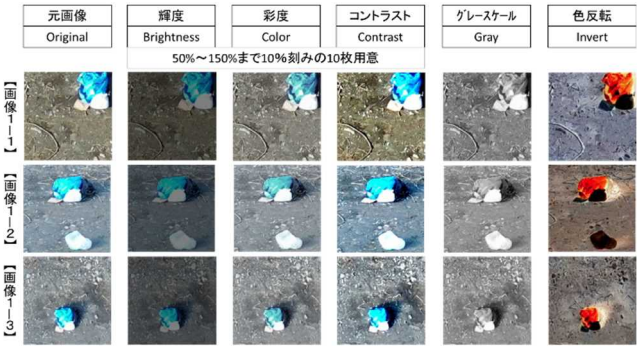

図-2　実験 1 の画像

		Top1	Top1_class	Top2	Top2_class	Top3	Top3_class
Original		91.0%	'garbage bag'	8.9%	'plastic bag'	0.0%	'bottle'
Brightness	50%	87.3%	'garbage bag'	12.6%	'plastic bag'	0.1%	'ground'
	60%	85.0%	'garbage bag'	14.8%	'plastic bag'	0.1%	'ground'
	70%	89.1%	'garbage bag'	10.8%	'plastic bag'	0.0%	'bottle'
	80%	87.8%	'garbage bag'	12.1%	'plastic bag'	0.1%	'bottle'
	90%	88.1%	'garbage bag'	11.8%	'plastic bag'	0.0%	'bottle'
	110%	76.1%	'garbage bag'	23.4%	'plastic bag'	0.2%	'ground'
	120%	57.4%	'garbage bag'	37.0%	'plastic bag'	4.2%	'ground'
	130%	49.9%	'garbage bag'	47.9%	'plastic bag'	1.6%	'ground'
	140%	50.6%	'garbage bag'	46.4%	'plastic bag'	2.5%	'ground'
	150%	53.5%	'garbage bag'	42.3%	'plastic bag'	3.4%	'ground'
Color	50%	88.6%	'garbage bag'	11.4%	'plastic bag'	0.0%	'ground'
	60%	92.6%	'garbage bag'	7.4%	'plastic bag'	0.0%	'ground'
	70%	86.0%	'garbage bag'	14.0%	'plastic bag'	0.0%	'ground'
	80%	88.0%	'garbage bag'	12.0%	'plastic bag'	0.0%	'ground'
	90%	88.9%	'garbage bag'	11.0%	'plastic bag'	0.0%	'ground'
	110%	86.6%	'garbage bag'	13.4%	'plastic bag'	0.0%	'bottle'
	120%	82.6%	'garbage bag'	17.3%	'plastic bag'	0.1%	'bottle'
	130%	83.9%	'garbage bag'	16.1%	'plastic bag'	0.0%	'bottle'
	140%	84.3%	'garbage bag'	15.6%	'plastic bag'	0.1%	'bottle'
	150%	83.4%	'garbage bag'	16.4%	'plastic bag'	0.1%	'bottle'
Contrast	50%	83.2%	'garbage bag'	16.6%	'plastic bag'	0.1%	'ground'
	60%	81.6%	'garbage bag'	18.2%	'plastic bag'	0.1%	'ground'
	70%	85.6%	'garbage bag'	14.4%	'plastic bag'	0.0%	'ground'
	80%	88.1%	'garbage bag'	11.9%	'plastic bag'	0.0%	'bottle'
	90%	86.7%	'garbage bag'	13.3%	'plastic bag'	0.0%	'bottle'
	110%	84.0%	'garbage bag'	15.9%	'plastic bag'	0.0%	'ground'
	120%	75.3%	'garbage bag'	24.0%	'plastic bag'	0.3%	'ground'
	130%	67.6%	'garbage bag'	28.1%	'plastic bag'	3.5%	'ground'
	140%	63.4%	'garbage bag'	30.9%	'plastic bag'	4.9%	'ground'
	150%	67.7%	'garbage bag'	27.8%	'plastic bag'	4.0%	'ground'
Gray		96.6%	'garbage bag'	3.3%	'plastic bag'	0.1%	'ground'
Invert		79.0%	'garbage bag'	12.4%	'ground'	8.4%	'plastic bag'

図-3　実験 1 の分類結果（画像 1-1 の例）

4.4. 実験1：画像全体の特徴量の変化

(1)　実験内容

図-2で示した画像群に対し，CLIPに不法投棄物21種(ゴミ袋/ダンボール/ペットボトル…)＋背景27種(地面/草/川…)の合計48種類のクラスを与えて分類を行った．CLIPでは1枚の画像に対して，その画像がどのクラスに該当するかの確率を，48クラスの合計が100%になるよう出力する．

(2)　実験結果

図-3に画像1-1に対するTop3までの分類結果を示す．元画像での分類結果は91%の確率でゴミ袋と分類しており，教師データなしのZero-shot画像分類で高い精度を発揮している．輝度を上げた場合はゴミ袋の確率が下がり，その分ビニール袋の確率が上がっているが，クラスの順位に変動はなく，極端な精度の低下は見られない．色情報を大幅に変えた色反転画像ではTop2が地面となり唯一Top2の順位が逆転しているが，その場合でもTop1は79%の確率でゴミ袋と判定できている．

この結果から，CLIPは画像全体の輝度・彩度・コントラストなどの特徴量の変化に対して高い汎化性能を有していると判断した．

図-4　実験 2 の画像と分類結果

4.5. 実験2：背景の特徴量の変化

(1)　実験内容

図-4で示した背景の異なる複数地点でゴミ袋が映った画像に対し，CLIPで実験1と同様の48クラスの分類を行った．画像2-1と2-2は共に背景が芝であるが，枯れ方や土の見え方が異なる．画像2-3は背景が砂利であり，大きく特徴量が異なる．

(2)　実験結果

図-4に各画像のTop3までの分類結果を示す．画像2-1では，画像中央付近の白いビニール袋に反応し，Top1は77%でビニール袋と分類しているが，画像の上部に見切れて青いビニールが映っていることで焦点が分散したためか，Top2で15.6%の確率で地面と分類している．これに対し，同じ芝が背景の画像2-2では，96.9%と高い確率でゴミ袋と分類し，地面の確率は非常に低く抑えられた．砂利が背景の画像2-3でも同様の傾向であった．

この結果から，CLIPでは背景の違いよりもゴミ袋の配置や映り方の違いが，分類結果に影響を与えていることが示唆される．そのため，背景の特徴量の変化に対して頑健な実験結果となった．

５．まとめと今後の課題

基盤モデルの一つである CLIP の Zero-shot 画像分類を用いて，一般的な不法投棄物を検出できることを確認した．日照条件や現地環境などにより色情報が異なる画像・背景が異なる屋外画像に対して，CLIP が汎化性能を有することを確認した．

将来的に，土木分野の教師あり学習で課題となっている教師データのバリエーション不足を，基盤モデルの汎化性能により解決する可能性がある．画像全体や背景の特徴量のバリエーションには基盤モデルの汎化性能で対応し，土木分野で特有の検出対象に絞って追加学習・ファインチューニングを行う手法を開発することが期待される．

参考文献

1) Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever: Learning Transferable Visual Models From Natural Language Supervision, arXiv: 2103.00020, 2021.
2) Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev: LAION-5B An open large-scale dataset for training next generation image-text models, arXiv:2210.08402, 2022.