The 37th Annual Conference of the Japanese Society for Artificial Intelligence, 2023

河川の不法投棄物検知におけるマルチモーダルモデルの適用検証

Validation of the application of a multimodal model in the detection of illegal dumping in rivers

岡野 将大[*1] 吉田 龍人[*1] 高橋 悠太[*1] 藤井 純一郎[*1] 天方 匠純[*1]
Masahiro Okano Ryuto Yoshida Yuta Takahashi Junichiro Fujii Masazumi Amakata

[*1] 八千代エンジニヤリング株式会社
Yachiyo Engineering Co., Ltd. #1

Recently the river management policy in Japan has progressed the river policy initiative toward auto-patrol using drone surveillance and computer vision technique, due to shortage of qualified workers. If we automate to detect in river patrols, the prediction errors could happen in object identification, so that the illegal judgment not always corresponds to subtle nuances stipulated in the patrol regulations.

It is the most appropriate method to detect illegal objects as riverine class labels: dumping and structures, and to classify the detected region into multiple classes. Furthermore, we can automate to assign the detected region to illegal objects according to the classified outputs.

However, it is difficult to build an illegal object detection model because we could not completely define all of classes with respect to riverine objects.

This paper proposes an application for auto-patrol task using text-image multimodal prediction based on a foundation model: the CLIP that incorporates the mapping between riverine text and drone patrol images so as to tag illegal objects. We demonstrate our method to a dataset of drone patrol images, and implement experimental studies to evaluate the test prediction.

1. はじめに

河川の異常を発見するための巡視は各自治体によって日常的に実施されているが，専門知識を有する作業員の人手を有するため、作業効率が低く，記録内容にも一貫性がないなどの問題がある。これに対して、国土交通省では革新的河川技術プロジェクト[国土交通省 2019]の一環として、ドローンと AI を組み合わせによって河川巡視を自動化する取り組みを開始し、これらの解決を目指している。

他方で人が行うことを前提にして定められた巡視規程に整合するように、AI で直接的に解決を目指すことは容易ではない。中でも、巡視規定で求められる違法に設置された看板等の工作物と違法に捨てられた投棄物の画像認識で識別するのは困難を極める。例えば川を渡すようにかけられた小さな木の板は規程上工作物として扱うが、河川上にただ置かれた茶色い段ボールは投棄物として扱うなど、AI の誤認があって然るべき微妙なニュアンスの識別が求められる。

巡視記録が人によって判断誤差があるように、空撮画像から作成する AI 用の教師ラベルにも判断誤差が生じるのは自明であり、これらを 2 クラス分類するよう学習するのは妥当ではない。さらに、不法工作物および不法投棄物の 2 クラスで分類するとその物体の詳細が明らかにならず、河川管理上危険になりうる物体を迅速に把握することができない。

この問題に対する解法として、不法投棄物および不法工作物を河川上に設置された物体として 1 クラスとして検出し、①検出領域を多クラス分類する、②分類されたクラスに応じて不法工作物および不法投棄物に割り当てるといった処理を実施することが最も妥当であると想定されるが、多岐にわたる対象物体のクラス数を定義した上で、それらの識別を行うモデルを構築することは現実的ではない。

そこで、本研究では大規模データセットによって自然言語と画像の相関性を学習したマルチモーダルモデルを利用し、河川の空撮画像に写った不法投棄物に対するタグ付けを試行し、その結果を評価する。

連絡先：岡野将大，八千代エンジニヤリング株式会社，東京都台東区浅草橋 5-20-8 CS タワー 技術創発研究所，TEL：03-5822-6028，Mail：ms-okano@yachiyo-eng.co.jp

2. 関連研究

河川巡視に深層学習を適用した事例として[高橋 2020]と[吉岡 2021]らの研究がある。[高橋 2020]らの研究では、学習データとして不足している不法投棄物が写ったドローン画像を補う手法を実験し、[吉岡 2021]らの研究ではドローン画像の撮影方法(対地高度，撮影アングル)を工夫する実験を行った。これらの研究で、ドローンで撮影した画像から深層学習を使用して不法投棄物を正確に検知することが可能となった。

教師あり学習の物体認識モデルの欠点として、新たに検出する対象が増えた場合、ラベル付きのデータを再定義・再学習するため、コストが掛かる。また、不法投棄物検知の課題として、河川ごとに検出したい対象が異なることや、河川巡視の検出対象が多様なため、河川内の検出対象すべてを対象にできないこと、アノテーションのコストを下げるためラベルを簡素にしており、不法投棄物を単一のラベルとして定義しているなどの課題がある。以上のことから、不法投棄物を詳細に識別するモデルを作ることは困難である。

近年、深層学習はスケーリング則[Kaplan 2020]に従うことでめざましい進歩を遂げており、学習ステップ、パラメータ、学習データの3要素を大きくすることで性能を飛躍的に向上させている。その中で、注目されているモデルの 1 つに CLIP[Radford 2021]がある。

CLIP (Contrastive Language–Image Pre-training) [Radford 2021]は、スケーリング則に従うモデルの 1 つであり、テキスト(キャプション)と画像のペアからなる大規模な学習データで、自己教師あり学習で学習された画像分類モデルである。同様なモデルで ALIGN[Jia 2021]と BASIC[Pham 2021]がある。どちらも CLIP をベースに改良を加え、CLIP を超える精度を達成した。しかし、この 2 つのモデルは公開されておらず、公開されている大規模データで学習されたマルチモーダルモデルは CLIP のみとなっている。また、OpenAI から発表されて以降も、[Schuhmann 2022]ら有志によりオープンソースプロジェクト(Open CLIP)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 画像分類の精度は、ImageNet を学習した ResNet101

が検証データを推論した精度と同精度に達しており、個別のタスクごとに学習したモデルと同精度または上回る精度を出している。このことから、河川巡視の不法投棄物検知のタスクにおいても CLIP の Zero-shot 画像分類を適用することで、多様な検出対象を網羅するとともに、ラベルの再定義や再学習のコストを低減できる可能性がある。

また、河川巡視に適用してきた深層学習モデルの場合、学習時に設定したクラスに推論時も限定されてしまう課題があったが、CLIP の場合、学習時に与えられている画像のペアであるテキストは、画像を説明するためのテキスト(キャプション)を与えられているため、クラスを限定することなく、タスクごとにクラス数(テキスト)を決めることができ、自由度の高い設定を可能としている。このことから、不法投棄物検知タスクで検出したい対象数が変化する場合においても推論時に自らが与えたテキストでクラス数を変更することで、欲しい結果を得ることが出来る可能性がある。

本研究では、大規模マルチモーダルモデルである CLIP が、ドローンで空撮された河川内の不法投棄物画像に Zero-shot 画像分類をすることで、詳細なタグ付け(属性情報付与)が可能か調査し、河川巡視の不法投棄物検知のタスクに適用できるか検証する。他方で入力のテキストによってモデルの精度に影響することがしられているため[Zhou 2021]、河川巡視の点検項目である不法投棄物に対してより正確なタグを付与するべく、精度の向上手法として Prompt Engineering と Prompt Ensemble を実施し、精度比較を行う。

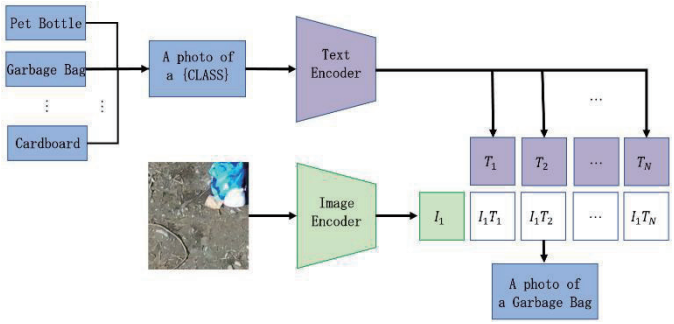

図 1　CLIP ゼロショット推論

3. 実験概要

3.1 実験データ

検証データを図 2 に示す。検証データは、河川上空からドローンで撮影した画像を用いた. この空撮画像は疑似的に再現した不法投棄物を直上から撮影したもので、不法工作物は含まれないものとした。空撮画像のサイズは 2160×3840 である。地点が変化することでの精度の変動を見るため、27 地点分計用意した。図 3 に推論時に使用する画像を示す。推論時は、CLIP の入力サイズである 224×224 にグリッド上に分割した画像を使用した。

3.2 手法

図 3 で示したドローンで空撮された河川内の不法投棄物が写っている画像に CLIP (Contrastive Language–Image Pre-training)を適用し、Zero-shot 画像分類を実施する。今回の実験で使用したモデルは、[Radford 2021]たちが公開した CLIP の ViT-L-14 モデルと[Schuhmann 2022]ら有志によりオープンソース(Open CLIP)として公開された LAION-400M の ViT-L/14、LAION-2B の ViT-L/14、LAION-2B の ViT-bigG-14 使用した。Open CLIP はスケーリング則に従い[Cherti 2022]、CLIP で使用された 4 億ペアのデータセットを超える 25 億ペアのデータセットで学習した LAION-2B の ViT-bigG-14 を公開しており、Zero-shot 画像分類の精度も公開されているモデルの中で最も高いモデルとなっている。河川巡視に適用する上で、複数の CLIP モデルの中から適切にモデルを選択する必要あるため、各モデルの精度を比較する。

図 2 で示した空撮画像は撮影時の GPS 情報が備わっているが、画像全体に対する不法投棄物の割合が少ないため、空撮画像だけでは不法投棄物のおおよその位置しか把握できない。物体認識モデルであれば不法投棄物の位置を把握することができ詳細な位置情報を取得できるが、CLIP には位置を特定する機能はない。また、空撮画像を CLIP の入力サイズである 224×224 にリサイズした場合、対象物を上手く分類することができないことが考えられる。そこで、画像を CLIPの入力サイズである 224×224 でグリッド上にクロップして、入力することで情報の損失を防ぐことができる。また、グリッド上にクロップしているため位置情報も獲得するこができる。

今回の実験で使用するクラス数は、27 地点の不法投棄物を検知するため、不法投棄物 21 種(ビニール袋、ダンボール、ペットボトル …)＋バックグラウンド 27 種(地面、草、川 …)の合計 48 種類とした。

$$p(y=i|x)=\frac{exp(\cos(T_i,\ I))}{\sum_{j=1}^{K} exp\left(\cos\left(T_j,\ I\right)\right)} \tag{1}$$

上式に CLIP のゼロショット推論時の式を示し、図 1 にアーキテクチャを示す。I と T はそれぞれ画像とテキストを Image Encoder と Text Encoder で出力した値であり、この値でコサイン類似度をとったあと、ソフトマックス関数を適応する。この計算により、分類カテゴリ確率を計算できる。

図 2　検証データ（空撮画像）

図 3　推論画像（分割画像）

3.3 精度評価

図 4 に各モデルの精度比較を示す。精度指標は、推論画像内に複数の物体が写るため、Top N Accuracy($N = 1,3,5$)で求めた。テキストはクラス名のみで推論を行った。

図 5 に各モデルのテキストをクラス名、Prompt Engineering、Prompt Ensemble の3つに変更し精度比較した結果を示す。精度指標は Top 1 Accuracy で求める。

Prompt Engineering などの手法が、個別の推論画像の精度をどれくらい向上させたかを比較するため、図 6 に精度比較した結果を示す。モデルは Open CLIP の LAION-2B ViT-bigG-14 を使用し、精度指標は分類カテゴリ確率を使用した。

3.4 精度向上に向けた検討

(1) Prompt Engineering

精度向上へ向けた検討として、Prompt Engineering を実施し、実験を行った。Prompt Engineering は、図 1 に示すアーキテクチャの Text Encoder に入力するテキストを「CLASS」から「A photo of a{CLASS}」に変更することで、テキストに説明文(Prompt)を加えモデルが読み取りしやすくする技術であり、他の大規模言語モデルでも精度の向上が報告されている[Kojima 2022]。テキストで画像を説明する CLIP においても精度が向上することが報告されており[Zhou 2022]、式(1)で示したテキストの入力 を、Prompt Engineering したテキストに変更することで、画像を最も説明できるテキストと画像のコサイン類似度が大きくなることで精度が向上する。

(2) Prompt Ensemble

精度向上へ向けた検討として, Prompt Ensemble を実施し、実験を行った。Prompt Engineering と同様、テキストに説明文(Prompt)を加えモデルが読み取りしやすくする技術であるが、異なる点として説明するテキストを複数用意し、各テキストのコサイン類似度の平均で計算される。CLIP でも同様の手法で精度が向上したことを報告している[Radford 2021]。今回、Prompt Ensemble を実施するにあたり Prompt を 12 文用意した。Prompt の選定方法は、CLIP で使用されていた ImageNet や COCO のプロンプトテンプレートからランダムに 100 文選択したあと、個別で CLIP に入力し精度を比較した上、もっとも精度が高かった Prompt を Prompt Engineering に使用し、上位 12 文を Prompt Ensemble 用に選定した。

4. 結果と考察

4.1 精度比較

図 4 で示す、各モデルの Top N Accuracy の比較では、クラス名のみの推論ではあったが LAION-2B ViT-bigG-14 が最も高い精度であり Top1 で 0.748%、Top3 では精度が 100%に達していた。Open CLIP のモデルは LAION-2B ViT-bigG-14 、LAION-2B ViT-L/14、LAION-400M ViT-L/14 の順に精度が高く、スケーリング則に従っているのがわかる。CLIP ViT-L-14 は学習データの違いから LAION-400M ViT-L/14 より精度が高かったと推察する。また、データセットは 27 地点分用意したが、地点の変化による精度の劣化はなかいことを確認した。

図 5 のテキストを変更した場合の精度は、4 つのモデル全てで、Prompt Ensemble、Prompt Engineering 、クラス名の順で精度が高かった。モデルの最高精度は LAION-2B ViT-bigG-14 で 0.948%だった。Prompt Engineering が Prompt Ensemble よりも精度が低い理由として、データの多様性が原因だと考えられる。Prompt Engineering の 1 文で画像との相関性を上げる場合、Prompt に合う画像と合わない画像があり、合わない画像の精度が下がってしまうと考えられる。Prompt Ensemble の場合、複数の Prompt で画像との相関性を上げる為、別の Prompt で合なかった画像をカバーすることができるからと考えられる。

図 6 に示した通り、個別の推論画像の精度でも、ほとんどの場合で Prompt Ensemble が最も高い精度だった。テキストがクラス名のみの場合、ゴミ袋は 0.531%の分類確率だったが、Prompt Ensemble を変更した場合、0.873%まで精度の向上がみられた。段ボールも微増ではあるが 0.072%向上した。Prompt Ensemble がもっとも精度が向上したことで選択した 12 文の Prompt は、有効に機能したと考察できる。

しかし、ビニール袋の画像では、テキストがクラス名の場合と比べ Prompt Ensemble の精度結果が 0.194%低下し、もっとも低くなっていました。図 5 の結果からデータセット全体の精度が上がっているので、選択された Prompt がほとんどの推論画像を説明できていると考えられ、ビニール袋など精度が低下している画像は上手く説明できていないからだと考えられる。

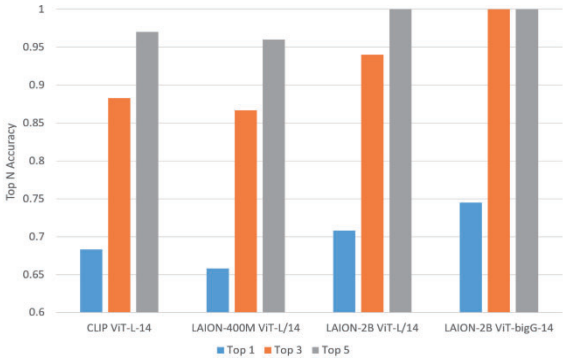

図 2 各モデルの Top N Accuracy 比較

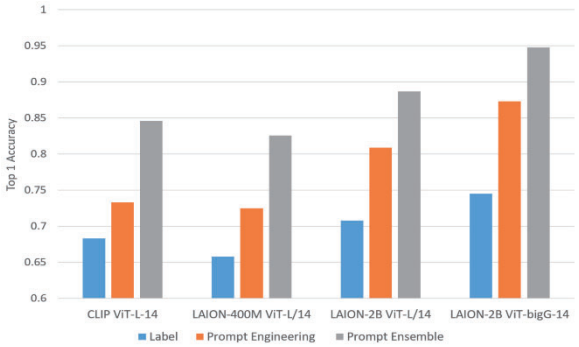

図 3 テキスト変更による各モデルの精度比較

おわりに

4.2 本研究の成果

複数ある CLIP モデルの中で LAION-2B ViT-bigG-14 がもっとも河川のデータを分類でき、テキストがラベル名のみの場合でも、74.8%の正解率を達成することが分かった。精度向上手法として Prompt Engineering と Prompt Ensemble 試行した。選択した 12 文の Prompt は LAION-2B ViT-bigG-14 で Prompt Engineering が 87.3%、Prompt Ensemble が 94.8%まで精度を向上させることができ有効な手段であることを確認した。

このことから CLIP を使用し、ドローンで空撮された河川内の不法投棄物画像に Zero-shot 画像分類をすることで、詳細なタグ付け(属性情報付与)が可能なことが実証できた。

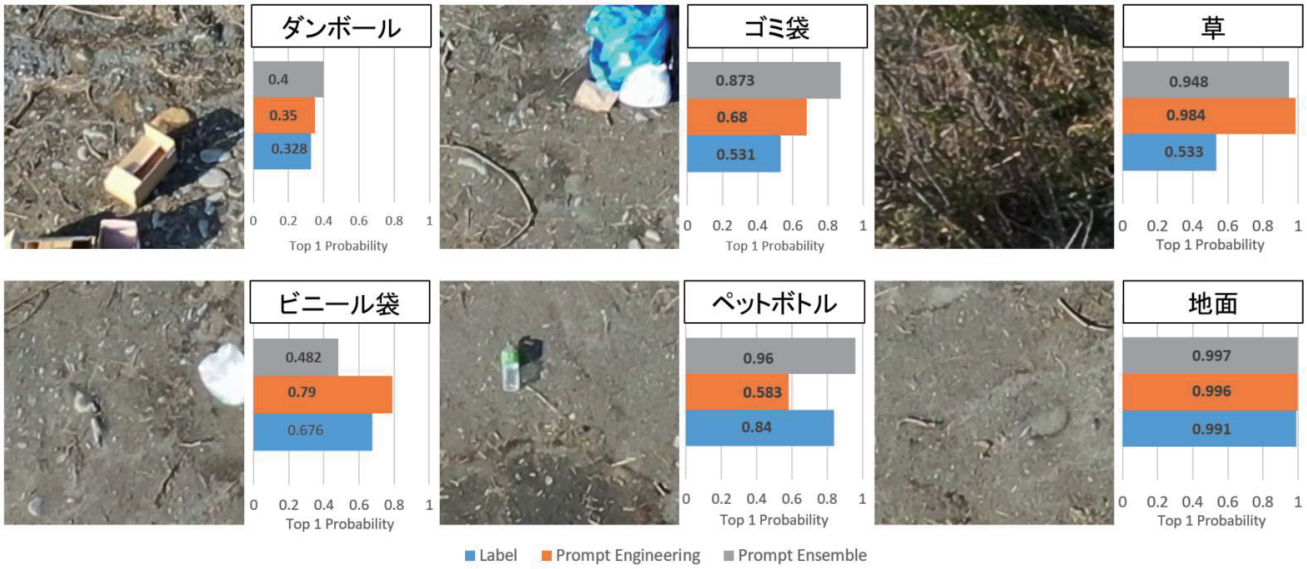

図 4 推論画像ごとのテキスト変更による精度比較

4.3 今後の課題

今後の課題を次に整理する. まず, 本研究で対象とした不法投棄物は疑似的に再現した空撮画像であるため、本物の不法投棄物で同様の精度が得られるか確認する必要がある。また、今回の研究で不法工作物は対象外としたため、不法工作物にも対応できるか検証を必要がある。

フィールドは 27 地点分を用意し地点が変更した場合に精度が変わらないことを確認したが、画角や地上解像度の違いで同様の結果が得られるか確認が必要である。

Zero-shot 画像分類の精度は、94.8%と高精度であったが、さらなる精度を目指すため転移学習を行い、精度を検証する。

今回使用した Prompt は、ImageNet や COCO の Prompt テンプレートからランダムに 100 文選んだため、空撮画像を説明するのには適当でなかった可能性がある。今後は、河川の画像に合う Prompt を作成し、精度を比較する必要がある。

謝辞

本研究でしようしたドローンの空撮画像は、国土交通省の革新的河川技術開発の一環 として実施した際に撮影したものです. 国土交通省水管理・国土保全局, より空撮データをご提供いただきました。厚く御礼申し上げ、ここに感謝の意を表します。

参考文献

[Cherti 2022] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman , Gabriel Ilharco , Cade Gordon , Christoph Schuhmann , Ludwig Schmidt , Jenia Jitsev : Reproducible scaling laws for contrastive language-image learning, arXiv:2212.07143, 2022.

[Jia 2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, Tom Duerig :Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision, ICML, 2021

[Kaplan 2020] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, Dario Amodei : Scaling Laws for Neural Language Models, arXiv:2001.08361, 2020.

[Pham 2021] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, Mingxing Tan, Quoc V. Le : Combined Scaling for Open-Vocabulary Image Classification , arXiv:2111.10050, 2021.

[Radford 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh , Gabriel Goh , Sandhini Agarwal , Girish Sastry , Amanda Askell , Pamela Mishkin , Jack Clark , Gretchen Krueger , Ilya Sutskever : Learning Transferable Visual Models From Natural Language Supervision, arXiv : 2103.00020, 2021.

[Schuhmann 2022] 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, Jenia Jitsev : LAION-5B: An open large-scale dataset for training next generation image-text models , NeurIPS, 2022.

[Zhou 2021] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, Ziwei Liu : Learning to prompt for vision-language models, arXiv preprint arXiv : 2109.01134, 2021.

[Zhou 2022] Kaiyang Zhou, Jingkang Yang Chen, Change Loy, Ziwei Liu : Conditional Prompt Learning for Vision-Language Models, CVPR, 2022.

[国土交通省 2019] 国土交通省水管理・国土保全局:革新的河川技術プロジェクト(第 5 弾), https://www.mlit.go.jp/river/gijutsu/inovative_project/project5.html , 2019.

[高橋 2020] 高橋悠太, 藤井純一郎, 天方匡純, 山下隆義:画像認識 AI による河川巡視を補う地上画像の特徴量とその利用法検討, AI・データサイエンス論文集, 1 巻 J1 号, p. 580-587, 2020.

[吉岡 2021] 吉岡小百合, 尾方浩平, 林雨亭, 下野友裕:UAV・AI を活用した河川巡視の高度化, 建設コンサルタント業務研究発表会論文集, 21 巻, p. 1, 2021.