E-070

Transformer モデルを応用したテキストベースの画像理解および分類手法の提案
An Application of Transformer Models to Text-Based Image Understanding and Classification

高橋 秀常† 小野田 弘士†
Hidetsune Takahashi Hiroshi Onoda

1. はじめに

近年リチウムイオン電池（LIB）を搭載した製品による発火事故が増加傾向にあり、分別のスマート化が喫緊の課題となっている。現状では、多くの場合、作業員の手選別により、破砕機等への流入を防いでいるが、混入による発火事故は後を絶たず、令和 5 年度には全国の市町村にて 8,543 件の関連火災が発生した[1]。これにより生じる廃棄物処理施設の稼働停止やそれに伴う社会的・経済的損失、さらには分別作業員のリスクがますます危惧されている[1]。

こうした課題を解決するために、さまざまな視点により先行研究がなされてきた。例えば、X線検出[1]や発火検出システム[2]が挙げられ、その有用性が認められて導入が進められている一方で、前者は導入費用が非常に高価であること、後者は発火自体を阻止することはできない等の課題が存在する。したがって、LIB を搭載した使用済み小型家電を、外観から自動検出するシステムへのニーズが高まっている。

筆者らはそのニーズに応える AI 開発のための方法論として、Transformer モデルを画像理解に活用することを検討している。CNN ベースのモデルは物体検出全般において非常に有用である一方、今回の事例研究では検出対象物の種類や形状が多様であるため、同モデルの補完的もしくは選択的な位置づけとして Transformer モデルを活用する。

とくに本研究においては、画像キャプション生成を用いたテキストベースでの画像理解および分類手法（以下「本手法」という）に焦点を置く。Vision Transformer (ViT) を基盤とする物体検出用途に特化した活用も視野に入れている一方、より言語的なアプローチを画像分類の段階にて扱う。近年の Vision Language Model (VLM) によるキャプション生成能力の発展がその背景あり、例えばノートパソコンのロゴや飲料の製造元などの詳細な情報も、画像のインプットに対してテキストで得られる。

本手法では、画像へのキャプションを生成して自然言語処理的アプローチにより画像分類を行う。キャプション生成には ViT と Transformer の両者を搭載した Florence2[3]を用い、テキスト分類においては BERT シリーズをはじめとした Transformer モデルを活用する。これを LIB 搭載製品に適用することで、画像理解および分類への自然言語処理的な手法を提案する。さらには、本研究で得た知見をもとに、同手法を物体検出等に応用するための方向性を提示し、CNN ベースのモデルとの補完的・選択的な活用を検討する。これらを通して、画像分野における、Transformer モデルのテキストベースでの活用を提案するとともに、本手法のさらなる発展に向けた指針を示す。

† 早稲田大学創造理工学部総合機械工学科 Department of Modern Mechanical Engineering, School of Creative Science and Engineering, Waseda University

2. システム概要

本研究では Transformer モデルを用いて、キャプション生成をもとにした画像分類を展開する。具体的には、Florence2 により画像に対する詳細な英文キャプションを生成し、2 クラス間（LIB 搭載および LIB 非搭載）のテキスト分類として扱う。分類においては、BERT やその派生を含む計 12 種の Transformer モデルを扱い、複数のハイパーパラメータ設定でファインチューニングを行い、性能を比較する。さらに、キャプションに複数の処理を加えた上で、最も高い性能を示したパラメータ設定にて、各モデルのファインチューニングを行う。

これらの開発および比較・検討を通して、画像理解および分類手法における自然言語処理的なアプローチを提案する。CNN をベースにしたモデルでの画像分類が主流になりつつある一方、キャプション生成とテキスト分類技術の向上を背景に、両者を併用することで、従来手法に対して補完的・選択的に機能する可能性が考えられる。本研究では、この独創的な視点をもとに Transformer モデルの有用性を検証するとともに、画像理解および分類手法のひとつとして提案する。

3. 検出手法の検討と開発

3.1 実験設定

3.1.1 検出対象の選定

LIB を搭載した使用済み小型家電の検出対象として、スマートフォン、ノートパソコン、タブレット PC、ワイヤレスファン、ワイヤレスイヤホン、ワイヤレスイヤホンのケース、モバイルバッテリー、電子タバコの計 8 種類を選定した。なかでもワイヤレスイヤホンは極めて小型であり、ワイヤレスファンや電子タバコは製品によって形状が多様であるため、これらの検出は相対的に難易度が高いと考えられる。

3.1.2 モデルとハイパーパラメータの選定

本稿では、Hugging Face が提供する無償で利用可能な Transformer ベースのモデル群[4]を活用した。具体的には、BERT およびその派生モデルを含む 6 種類の系列と、それらの拡張版を合わせて計 12 種類のモデルを選定し、ファインチューニングを実施した。

学習率は各モデルに応じて 3 段階に設定し、エポック数を 1 から 4 まで変化させて、それぞれの条件下で重みを得た。多くのモデルでは学習率を 1×10^{-5}、2×10^{-5}、3×10^{-5} としたが、GPT-2 系列については学習の不安定性を避けるため、より小さい値を用いた。選定したモデルおよび学習率の組み合わせを表 1 に示す。

表 1: 使用したモデルと学習率の一覧

Transformer モデル	学習率 1	学習率 2	学習率 3
BERT-base	1×10^{-5}	2×10^{-5}	3×10^{-5}
BERT-large			
ALBERT-base			
ALBERT-large			
RoBERTa-base			
RoBERTa-large			
ELECTRA-base			
ELECTRA-large			
XLNet-base			
XLNet-large			
GPT2	3×10^{-6}	5×10^{-6}	8×10^{-6}
GPT2-large	5×10^{-6}	1×10^{-5}	2×10^{-5}

3.1.3 テキスト処理の検討

Florence2 により得るキャプションは、プレーンテキストとしての使用に加えて、事前にテキスト処理を施した上でのファインチューニングにも活用した。具体的には、プレーンテキストのレンマ化、レンマ化およびストップワードと句読点の除去、名詞の抽出の 3 種類の処理を施した。これらの事前処理済みテキストを用い、先に選定した 12 種類の Transformer モデルに対して、前項において最も高い性能を示したパラメータ設定のもとでファインチューニングと評価を実施した。なお、Transformer モデルの事前学習は、通常こうした前処理を伴わずに行われているため、識別性能が低下することが懸念される。しかし、本検討は 6.2 で述べるフレーズ・グラウンディング等による応用を視野に入れたものであり、その際の方針を検討する上での参考となることを目的としている。

3.2 データセットの作成

3.2.1 学習用データと検証用データの作成

LIB 搭載製品（ラベル: *LIB*）および LIB 非搭載製品（ラベル: *non-LIB*）に関してデータを作成した。具体的には、インターネット上および既存のデータセット[5]により計 12,000 枚以上の関連画像を、ラベル間でデータの偏りがないよう取得した。検出対象の種別ごとに、インスタンス数の 10 %を検証用データとして割り当て、残りを学習用データとした。その後、学習用の画像に対して Florence2 を用いてキャプションを生成し、ラベルとともに CSV ファイルに出力した。検証用の画像に対しても同様の処理を行い、これらをもって学習用データと検証用データを構成した。

3.2.2 テストデータの作成

各ラベルに対して 320 枚、合計 640 枚の画像に対するキャプションによりテストデータを構築した。各画像におけるインスタンス数は 1 とし、背景は白、撮影距離は物体から約 45 cm に統一した。

LIB に関しては、検出対象となる 8 種類の種別ごとに 5 つの異なるインスタンスを用意し、それぞれ表面および裏面から撮影を行った。さらに、各面を 90 度ずつ回転させて全方向を撮影することで、合計 320 枚の画像を得た。*non-LIB* に関しては、LIB を搭載していないさまざまな種類の廃棄物（商品パッケージや飲料容器等）を対象として 80 インスタンス収集し、同様に 90 度ずつ 1 周回転させて撮影し 320 枚の画像を取得した。

これらの画像に対して、前項と同様に Florence2 によりキャプションを生成し、画像 ID および正解ラベルとともに CSV ファイルとして出力し、テストデータを構成した。

4. 検出結果と考察

4.1 プレーンテキストにおける適用

まずは、キャプションに特別な処理を行わない状態で、3.1.2 に示した条件のもと、各 Transformer モデルにファインチューニングを実施した。各モデルにおいて最も高い F1-Score を示した条件とスコアを抜粋し、表 2 に示す。

全モデルにおいて F1-Score が概ね 80%以上を達成し、中でも最も高いスコアを示したのは RoBERTa-base で 86.8 %であった。その他のモデルについても、ELECTRA-base が 82.1 %、BERT-base が 85.4 %と、識別性能に大きなばらつきは見られなかった。

さらに、Recall に着目すると、全てのモデルが 92 %以上となり、とりわけ GPT2-large が 99.4 %を達成した。第 1 章で述べた通り、本研究で対象とする LIB 搭載製品は、家庭ごみ等への混入を防ぐための識別が課題となっており、その点からも Recall を可能な限り 100 %に近づけることが求められる。この観点からも、本手法は同様の画像識別のニーズに対して有効なアプローチとなり得ることが示唆される。

以上を踏まえると、本研究で用いた Transformer モデルを活用する手法は、手法自体の簡便性や画像データの質を考慮すると、十分な有用性があるといえる。さらに、本研究で用いた一部の画像には、スマートフォンやノートパソコンの初期機種など、テストデータと比較して著しく古い製品が含まれていたが、そのような制約下においても、キャプション生成とテキスト分類により高い識別率を維持できた。これらより、本手法は画像理解および分類の方法論のひとつとして成立し得ると考えられる。

4.2 各処理による実験

3.1.3 で記した 3 種類の前処理（レンマ化、レンマ化およびストップワードと句読点の除去、名詞の抽出）をキャプションに適用する実験を実施した。学習データ、検証データ、およびテストデータの全てに各処理を施し、表 1 で示したモデルおよびハイパーパラメータの条件に従ってファインチューニングを行った上で、テストデータにおける識別率を算出した。表 1 の F1-Score および Recall で最も高い識別率を達成した、RoBERTa-base (F1-Score) と GPT2-large (Recall) を代表例として、結果を表 3 および表 4 に示す。

プレーンテキスト条件とその他 3 種の事前処理条件を F1-Score に基づいて比較した結果、GPT2-large におけるレンマ化適用時を除き、いずれのモデルにおいてもプレーンテキスト条件が最も高い識別性能を示した。この傾向は、3.1.3 で述べた通り、Transformer モデルはレンマ化等の事前処理を伴わない通常の文で事前学習されていることに起因すると考えられる。よってモデルがそのような入力に対して相対的に脆弱なため、事前処理をしたテキストに対して識別性能が低下するといえる。

表 2: プレーンテキスト条件下での識別性能の比較

Transformer モデル	学習率	エポック数	Accuracy	Precision	Recall	F1-Score
BERT-base	2×10^{-5}	1	0.839	0.780	0.944	0.854
BERT-large	2×10^{-5}	1	0.761	0.690	0.947	0.798
ALBERT-base	2×10^{-5}	1	0.811	0.739	0.963	0.836
ALBERT-large	1×10^{-5}	4	0.814	0.754	0.931	0.834
RoBERTa-base	3×10^{-5}	1	0.855	0.793	0.959	0.868
RoBERTa-large	2×10^{-5}	1	0.811	0.748	0.938	0.832
ELECTRA-base	2×10^{-5}	3	0.798	0.739	0.922	0.821
ELECTRA-large	2×10^{-5}	1	0.839	0.767	0.975	0.858
XLNet-base	3×10^{-5}	1	0.806	0.744	0.934	0.828
XLNet-large	3×10^{-5}	1	0.800	0.720	0.981	0.831
GPT2	1×10^{-5}	3	0.822	0.743	0.984	0.847
GPT2-large	5×10^{-6}	1	0.761	0.678	0.994	0.806

このように識別率は一定程度低下するものの、各種前処理を施した条件においても、全体として安定かつ妥当な性能が維持されていることが確認できる。例えば RoBERTa-base においては、プレーンテキスト条件下で F1-Score が 86.8 %であったのに対し、名詞抽出処理を施した条件でも 75.4 %を維持している。F1-Score において約 10 %の低下に留まっていて、他の評価指標においても概ね 70 %を上回る結果が得られているため、現段階の識別性能として妥当である。なお、本実験ではプレーンテキスト条件と同一の学習率およびエポック数を簡易的に適用しているが、3.1.2 と同様にハイパーパラメータを調整することで、さらなる性能向上が見込まれる可能性がある。これらの検討より、6.2 で述べるような、キャプションをフレーズ単位で処理することによる物体検出への応用においても、比較的有用な検出性能が期待される。

5. YOLO との併用に関する検討

本研究では簡易的に画像分類を対象としたが、筆者らは先行研究において YOLO11x を用いた物体検出も実施している[6]。両者は本質的に異なるタスクであり、直接的な比較には適さないものの、いずれの評価にも同一のテストデータを使用しており、全画像においてフレームあたりのインスタンス数は 1 に統一されている。

したがって、位置情報の特定が必要とされない状況（例えば小規模な回収ボックスにおける利用）を想定した場合には、両手法の結果を目安として比較することは妥当であると考えた。そこで本稿では一例として、テキストベースの画像分類手法の中で最も高い F1-Score を示した、RoBERTa-base のプレーンテキスト条件下での結果を比較対象として用い、YOLO11x による物体検出との比較を表 5 に示す。

比較の結果、Precision を除く 3 指標において、RoBERTa-base を用いたテキストベースの画像分類が YOLO11x による物体検出を大きく上回る結果となった。特に Recall に関しては、物体検出手法を 30 %以上も上回っており、選別漏れを抑制するという観点から本手法の有用性が示唆される。

本章で比較した両手法は、それぞれ異なる特性や制約があり、ユースケースに応じて補完的・選択的に活用することが望まれる。例えば、ベルトコンベア上を流れる廃棄物（LIB 搭載製品や家庭ごみ等）を対象とする場面では、YOLO の即時性を活かした検出を基盤としつつ、破砕機投入前の最終確認としてテキストベースの識別手法を併用でき得る。また選別工程の上流にて、処理の即応性がさほど重視されない状況では、商業施設等に設置される小規模な回収ボックスを例に、テキストベースの識別手法も実用的な選択肢になる。

Transformer モデルを用いたテキストベースの手法は、CNN ベースのモデルとはアーキテクチャや得られる情報が異なるため、画像に対する処理に対して異なる特性が生じる利点がある。例えば図 1 に示すのは、YOLO11x では検出できなかった一方で、テキストベースの分類手法では正しく識別された例である。このように両手法を選別フローにおいて併用することで、それぞれの特性を補完し合い、アルゴリズム全体としての識別性能が向上することが期待できる。

さらに、CNN ベースのモデルでは対応が難しい未学習の物体に対しても、Transformer や ViT ベースのモデルはより柔軟に対応できる可能性がある。例えば、日本製のペットボトルをもとにファインチューニングを行った YOLO は、海外製の形状が特徴的なものには対応が難しいことが想定される一方、テキストベースの手法により補完できる可能性がある。さらには現時点で予測できない、将来的に流通する廃棄物（LIB 搭載製品や家庭ごみ等）に関しても同様で、識別対象の多様化や、その変化への適応からも利点となり得る。

6. 応用の検討

6.1 フレームあたりの物体数の増加

本研究では、各フレームに含まれるインスタンスを 1 つに限定し、テキストベースの画像分類手法の有効性を検証した。一方で、例えばごみ処理施設の選別ラインのような、実際のシステムへの応用を想定した場合には、1 フレーム中に複数の物体が存在する状況でも適切に識別が行える必要がある。

そのためには、物体数が増加した場合でも、キャプションはそれぞれの物体を十分に描写してかつ、それを

Transformer モデルが適切に処理できることが前提となる。本手法がこの条件でも有効に機能するかどうかについては、本稿の検証には含まれておらず、今後の研究課題となる。

6.2 物体検出への応用

本稿は当該手法を、画像分類への方法論として提案したが、物体検出への応用も検討している。例えば事例研究である LIB 搭載製品の選別に限ると、順次投函される小規模な回収ボックスのような場面において画像分類で十分な一方で、大規模なごみ処理施設における選別レーンなど、複数の廃棄物が同時に存在する環境では、対象物の有無に加えて、その位置の特定も必要となる。

このような場合には、画像キャプション生成と併せて領域特定を行う、フレーズ・グラウンディングの適用が有効である。これを適用すると、画像に対する詳細なキャプションに加え、その各フレーズと対応するバウンディングボックスを同時に取得することが可能となる。また 4.2 で示した通り、もとのキャプション文から部分的に抽出する処理を施しても、本手法は識別性能を一定程度確保しており、この点からもフレーズ単位での有用性が十分に期待される。したがって本手法をフレーズ単位で適用し、さらにフレーズ・グラウンディングと組み合わせることで、物体検出へ応用することを検討している。

6.3 データの拡張

6.1 および 6.2 で示したような応用を試みるにあたっては、精度の向上を図るために基盤となるデータの増幅が今後の課題の一つとなる。その具体的な手法として主に 2 つ挙げられ、具体的には、画像処理と自然言語処理の両面からのアプローチが考えられる。

画像処理の観点からは、学習用および検証用データに対して画像の回転、色調やコントラストの調整を行うことで、追加の画像取得やアノテーションなしにデータを増幅することが可能である。加えて、画像の一部にマスキング処理を施すことで、物体同士が重なった状況を簡易的に再現し、より実用的な場面にも柔軟に対応できるようになる。

一方、自然言語処理的な手法としては、機械翻訳によるバックトランスレーションが有効であると考えられる。これは、例えば英語からスペイン語、さらにドイツ語を経由して再び英語に翻訳することにより、文意を維持しながら元のキャプションとは異なる表現を生成するもので、テキストデータの増幅手法として広く用いられている。

このような手法を用いることにより、新たなデータを追加取得することなく、もとのデータを増幅することができる。先述のフレームあたりのインスタンス数の増加や、物体検出への応用を視野に入れると、より多くのデータが必要になる可能性があり、そのための有効な手段となる。

表 3: RoBERTa-base の各処理条件下での識別性能

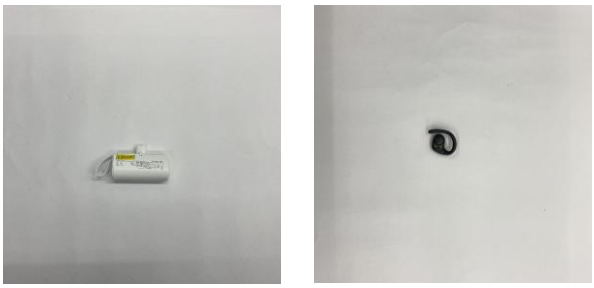
実験条件	Accuracy	Precision	Recall	F1-Score
プレーン	0.855	0.793	0.959	0.868
レンマ	0.788	0.711	0.969	0.820
レンマほか	0.745	0.739	0.759	0.749
名詞抽出	0.758	0.767	0.741	0.754

表 4: GPT2-large の各処理条件下での識別性能

実験条件	Accuracy	Precision	Recall	F1-Score
プレーン	0.761	0.678	0.994	0.806
レンマ	0.809	0.730	0.981	0.837
レンマほか	0.717	0.960	0.453	0.616
名詞抽出	0.758	0.816	0.666	0.733

表 5: YOLO と提案手法の識別能力に関する比較評価

指標	YOLO11x による物体検出	RoBERTa-base による画像分類
Accuracy	0.788	0.855
Precision	0.900	0.793
Recall	0.647	0.959
F1-Score	0.753	0.868



(a) モバイルバッテリー　　(b) ワイヤレスイヤホン

図 1　提案手法が YOLO の検出漏れを正しく識別した例

7. おわりに

本研究にて得られた知見を要約して以下に示す。

- Transformer モデルを活用したテキストベースのアプローチを、画像理解および分類に対する有効な手法のひとつとして提案した。
- 当該手法による画像分類結果を、CNN ベースのモデルによる物体検出結果と比較することにより、両者の補完的および選択的な活用を検討した。
- テキストベースの画像分類手法を物体検出などに応用することを検討し、その実現に向けた具体的な方向性を提示した。

参考文献

[1] 環境省環境再生・資源循環局, “市町村におけるリチウム蓄電池等の適正処理に関する方針と対策について（通知）”, 環循適発第 2504151 号 (2025).

[2] Tianhao Cheng et al., “Visual Identification-Based Spark Recognition System”, International Journal of Automation Technology, Vol.16, No.6 (2022).

[3] Bin Xiao et al., “Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks”, arXiv, 2311.06242 (2023).

[4] Thomas Wolf et al., “Transformers: State-of-the-Art Natural Language Processing”, Journal of ACL Anthology (2020).

[5] [Online Source], cited March 24, 2025, Available HTTP: https://www.kaggle.com/datasets/alistairking/recyclable-and-household-waste-classification.

[6] Hidetsune Takahashi, Hiroshi Onoda, Advancing End-of-Life Product Sorting Processes through Hybrid AI Approaches: A Case Study on SDA (Small Domestic Appliances) with Lithium-ion Batteries, IWEE2025 (予定).