# 震災アーカイブにおける言語モデルを用いた視覚探索

富谷, 竜一
Hascoet Tristan Erwan, Marie
高島, 遼一
滝口, 哲也

(Citation)
神戸大学都市安全研究センター研究報告,27:45-50

(Issue Date)
2023-03

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/0100482494

(URL)
https://hdl.handle.net/20.500.14094/0100482494

神戸大学都市安全研究センター
研究報告，第27号，令和５年３月

# 震災アーカイブにおける言語モデルを用いた視覚探索

Visual search using vision-language models for an earthquake disaster archive

富谷　竜一 1)
Ryuichi Tomiya
Tristan Hascoet 2)
高島　遼一 3)
Ryoichi Takashima
滝口　哲也 4)
Tetsuya Takiguchi

概要：神戸大学附属図書館は、阪神・淡路大震災に関連する画像、映像を多く保存する震災文庫デジタルアーカイブを公開している。当時の被害状況や防災意識を把握し、来るべき震災に備えるために、このような大規模なアーカイブの画像や映像の中から、人が自然に表現するような任意の自由形式のリクエストにより、特定の物体を見つける視覚探索は重要であると考えられる。近年、画像認識技術は、言語モデルを組み合わせることで、任意の自由形式のテキストの入力が可能になった。また大量データで学習を行った事前学習済みの一般モデルが公開されており、そのような画像認識の評価が可能になってきている。

本稿では、“a backpack”、“a safety cone”、“a blue tarp”、“a person sitting”、“a person wearing a helmet”、“a person riding a bike” と 6 つのリクエストを設定し、事前学習済み一般モデルを、震災アーカイブにおける視覚探索に用いたときの能力を評価する。また、リクエスト、アーカイブに特化したモデルの学習のための高効率なデータ収集方法を提案し、収集したデータでモデルのファインチューニングを行った際の結果について報告する。

キーワード：震災アーカイブ、視覚探索、言語モデル、物体検出、深層学習

## 1.　序論

現在、震災を経験したことのない世代の人達が増えている。過去に発生した震災の教訓・経験を風化させず、後世に伝え、来るべき震災に備えることが重要であると考えられる。当時の被害状況や防災意識を分かりやすく伝えるために、画像や映像を利用することは有効である。神戸大学は、阪神・淡路大震災に関連する約24,000枚の画像、約10時間の映像と多くのメディアを保存する震災文庫デジタルアーカイブを公開している。このアーカイブは、検索ツールを提供しているが、メタデータに画像や映像の詳細の部分に対しての記述がない事、表記ゆれの問題など、実際に、特定の物体が写っている画像や映像を取り出す事は難しい。目視で大規模なアーカイブを探索するとなると、非常に労力がかかってしまう。そこで、ユーザが自然に表現する自由形式のテキストで、画像や映像の一部分を精度よく取り出すシステムが必要であると考えられる。本稿では、震災文庫デジタルアーカイブの画像を対象として、人が自然に表現するリクエストの物体を見つける視覚探索を行う。“a backpack”、“a safety cone”、“a blue tarp”、“a person sitting”、“a person wearing a helmet”、“a person riding a bike”と 6 つのリクエストを設定して実験を行い、特定のアーカイブ、リクエストに対する、実際に公開されている事前学習済み一般モデルの能力を評価した。また、高効率な学習データの収集とファインチューニングの枠組みについて提案し、アーカイブ、リクエストに特化したモデルへの改善方法について検討する。

## 2. 震災アーカイブに対する取り組み

神戸大学附属図書館は、阪神・淡路大震災関係資料文庫（震災文庫）を管理している。震災文庫の収集資料は、商業出版物・私費出版物、各種団体・個人による研究報告・調査報告・統計資料・講演会等の記録・レジュメ・チラシ類などの印刷資料のほかに、電子資料（CD-ROM 等）・ビデオ・録音カセット・マイクロ資料・写真・地図など多岐にわたる。また、その一部はデジタルアーカイブとして公開されている。[1] 震災アーカイブの構築、公開は、震災が起きたあと、短期間で大量の映像や画像が撮影されるため、整理作業は膨大となり、データベース化が進まない問題がある。この問題を解決するために、メタデータ付与作業の一部を自動化する取り組みが行われている。しかし、事前に対象となる被写体の学習データを集め、学習と検証を行う必要がある。[2] 本稿では、詳細なメタデータが付与されていない画像アーカイブに対して、言語モデルを用いた画像認識技術を利用した視覚探索について検討する。

## 3. 言語モデルを用いた画像分類技術/物体検出技術

近年、画像認識技術は、言語モデルを組み合わせることで、任意の自由形式のテキストの入力が可能になった。それに加えて、従来の画像認識技術では、「自転車に乗っている人」を検出するためには、「自転車に乗っている人」の画像を大量に用意してモデルを学習しなければならなかったが、従来の技術に言語モデルを組み合わせることで、「自転車」、「人」、「乗っている」の概念を学習していれば、言語モデルの貢献によって、それに対応する画像データを用いた学習無しで検出することが可能になる。このような利点を活かした研究が多く行われていて、現在、事前学習済みの一般モデルが公開されている。本章では、本稿で利用する言語モデルを用いた画像分類技術/物体検出技術について紹介する。

### (1) 言語モデルを用いた画像分類技術「CLIP」

CLIP[3] は、カテゴリを自由に設定できる画像分類モデルである。インターネットから収集した 4 億もの画像とテキストのペアを学習データとしていて、画像と、画像に関するテキストとの類似度を近づけるという対照学習が行われている。画像エンコーダには、CNN か Vision Transformer、言語エンコーダには、Transformer を用いてモデルを構築している。画像分類の推論時には、候補となるクラスのテキストと分類したい画像を入力し、一番類似度が高くなるクラスをクラス分類の結果として出力する。特定のカテゴリを学習していないのにも関わらず、そのカテゴリの画像を理解するゼロショット理解が可能である。

### (2) 言語モデルを用いた物体検出技術「MDETR」と「GLIP」

CLIP のように、大量の画像とテキストのセットで学習を行っているフレーズベースの物体検出モデルが提案されている。

MDETR[4] は、マルチモーダルの推論システムである。MDETR では、CNN を用いて抽出した画像特徴量と、事前学習済みの自然言語モデル RoBERTa を用いて抽出した言語特徴量を、共有の埋め込み空間に投影し、連結させた後、Transformer に通し、テキストに登場している物体のバウンディングボックス、その物体について述べているテキストから予測されるラベル、信頼度（スコア）を出力する。テキスト中のフレーズと画像内の物体の間の明確なアライメントを持つ、既存のマルチモーダルデータセットを含めた 130 万の画像とテキストのペアを用いて学習を行っている。

GLIP[5] は、MDETR と同じような機能をもつ、マルチモーダルの推論システムである。GLIP は、MDETR が学習に利用しているデータセットだけでなく、固定されたクラスの検出データセットでも学習を行っている。Swin Transformer を用いて抽出した画像特徴量と、事前学習済みの自然言語モデル BERT を用いて抽出した言語特徴量を融合させて、テキストに登場している物体のバウンディングボックス、その物体について述べているテキストから予測されるラベル、信頼度（スコア）を出力する。既存のデータセットに加えて、インターネットから収集した 2,700 万の画像とテキストのペアを用いて学習を行っている。

## 4. 高効率なデータ収集とファインチューニングの枠組みの提案

3 章で紹介したモデルは、一般データセットで大量に学習が行われている。震災アーカイブを精度よく探索するためには、アーカイブに特化したモデルを構築する必要がある。そのためにファインチューニングを用い

る。それには、リクエストに対する正例と負例の学習データが必要であるが、阪神・淡路大震災デジタルアーカイブの画像には、「何がどこに写っているか」などの注釈であるアノテーションは付与されていない。学習データを収集する際にも、大規模なアーカイブから探索する必要があるため、非常に労力がかかる。そこで、高効率なデータ収集方法について提案する。アーカイブのすべての画像 24,303 枚の中から、リクエストに対する正例と負例を収集する。図 1 に学習データ収集の手順を示す。左は、完全に手動で収集した場合、右は、モデルを用いて、画像を並び替えた後に目視で確認を行い、収集した場合である。モデルを利用することで、効率良く学習データを収集することができる。検出モデル（GLIP、MDETR）を用いる場合は、ボックスがもつスコアの高い順、分類モデル(CLIP) を用いる場合は、テキストと画像の類似度が高い順に並べ替えを行う。検出モデルを用いる場合は、画像一枚につき複数のボックスをもつ場合もあるが、より高いスコアを対象に並べ替えを行い、二度同じ画像が登場しないように考慮している。

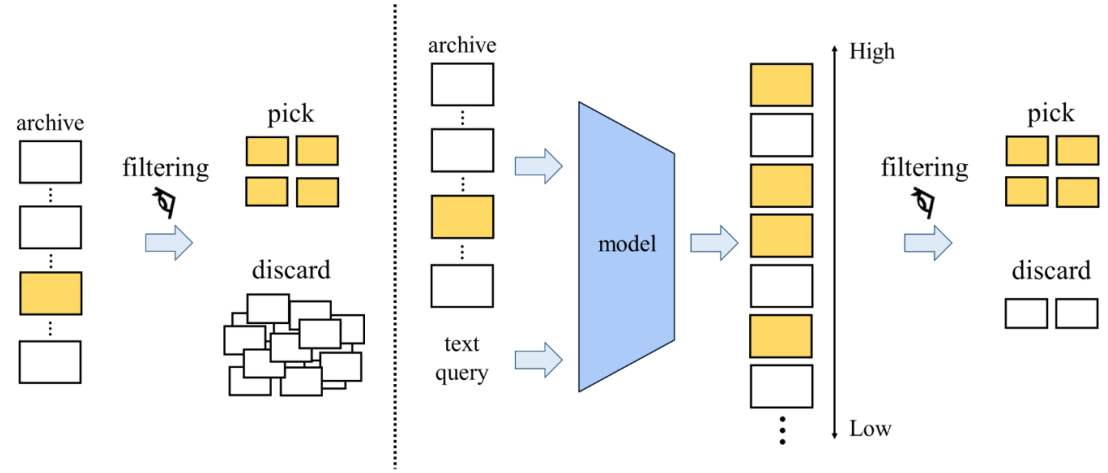


図 1:学習データ収集の手順　完全に手動で収集した場合（左）、モデルを用いて画像を並び替えたあと、目視で確認を行い収集する場合（右）

図 2 に、GLIP を用いて、リクエストを“a person wearing a helmet”としたとき、高い信頼度を持った検出例を示す。左の画像は正しく検出できている。右の画像は、“a person”は検出できているが、“a person not wearing a helmet”を高いスコアで検出してしまっている。高いスコアで誤検出を起こす“a person not wearing a helmet”の例は、他にも、帽子を被っている人など、紛らわしい物体が多かった。このように、一定数の学習データが集まるまで、目視で画像のチェックを行い、リクエストに対する正例と負例のデータ収集を行った。

図 3 に、 CLIP を用いて、リクエストを“a person sitting”としたとき、高い信頼度を示した 2 つの画像を示す。左の画像は“a person sitting”が写っているが、右の画像は椅子のみが写っている写真で“a person”は写っていない。CLIP は画像全体とテキストの類似度を計算するため、「テキストで指示した物体が写っていそうなシーンや場所の画像」と類似度が高いという特徴があるのかもしれないと考えられる。

このように、事前学習済みの一般モデルを利用し、画像を並べ替えることで、リクエストに対する正例と負例を収集することができた。


(a) 正しい検出例　　(b) 誤検出例

図2:リクエスト“a person wearing a helmet” に対するGLIPの検出例

(a) 正しい画像　　(b) 誤った画像

図 3:リクエスト“a person sitting” に対して CLIP が高い類似度を示した画像例

## 5. ファインチューニングを用いた実験

本実験では、リクエストが「写っている」か「写っていない」か、ラベル付けを行った 600 枚の小規模画像アーカイブに対して、視覚探索を行う。リクエストの物体が写っている、あらゆる画像を見つける精度について評価する。検出モデル(GLIP、MDETR)を用いる場合は、画像一枚につき、複数のボックスをもつ場合もあるため、「写っている」とラベルがついた画像は、目視チェックにより、対象の物体を検出できていると判断した最もスコアが高いボックス、「写っていない」とラベルがついた画像は予測された中で最もスコアが高いボックス、それぞれのスコアを対象として、600 枚の画像を並べ替える。分類モデル(CLIP)を用いる場合は、画像とテキストの類似度順に 600 枚の画像を並べ替える。評価指標として、各再現率レベルでの適合率の平均である、Average Precision を用いた。リクエストの物体が写っている画像すべてが上位に、写っていない画像すべてが下位にといった理想的な並べ替えができた場合、100 [%] となる。表 1 に、それぞれのリクエストに対して各モデルの Average Precision を示す。

表 1. Average Precision の比較 [%]

| | GLIP | MDETR | CLIP |
|---|---|---|---|
| “a backpack” | **20.31** | 6.07 | 13.58 |
| “a safety cone” | **84.81** | 27.59 | 28.94 |
| “a blue tarp” | **36.83** | 28.24 | 22.57 |
| “a person sitting” | **17.66** | 11.10 | 16.15 |
| “a person wearing a helmet” | 18.52 | 22.39 | **25.06** |
| “a person riding a bike” | 10.27 | 6.26 | **17.27** |

表 1 から、“a backpack”、“a safety cone”、“a blue tarp”、“a person sitting”は、GLIP を用いたときに最も精度が良いことが分かる。この 4 つのリクエストに対しては、GLIP は、リクエストが写ったあらゆる画像を探索する能力が最も優れていると考えられる。特に、“a backpack”、“a safety cone”、“a blue tarp”の単純なリクエストに対して GLIP は効果を発揮した。この理由として、MDETR、CLIP とは異なり、GLIP は、固定されたクラスの検出データセットでも学習を行っていて、リクエストに近い意味のクラスが含まれている事が考えられる。精度のよい GLIP を使い、ファインチューニングの実験を行う。

4 章に示した手順で収集したボックスと画像の情報を利用して、GLIP のファインチューニングを行った。本実験では、それぞれのリクエストごとに学習を行い、リクエストごとに特化したモデルを目指すことに専念している。2 種類のデータを用いて学習を行った。

(1) 正例データ 50 枚(スコアが高い順に 40 枚を学習データ、残り 10 枚を検証データ)で学習

(2) (1)と同じ正例データに加えて、負例データ(それぞれスコアが高い順に 40 枚を学習データ、残り 10 枚を検証データ)で学習

学習データ、検証データには、テストデータセットの 600 枚の画像を含まないようにして、バッチサイズ:2、最

適化手法:AdamW、学習率:0.0001 と設定して、学習を行った。表 2 に、それぞれのリクエストごとに、一般モデル、正例データを用いたファインチューニングを行ったモデル、正例データと負例データを用いてファインチューニングを行ったモデルの Average Precision を示す。

表 2. ファインチューニング前後の Average Precision の比較 [%]

| | 一般モデル | ファインチューニング（正例） | ファインチューニング（正例+負例） |
|---|---|---|---|
| “a backpack” | 20.31 | 35.27 | 39.14 |
| “a safety cone” | 84.81 | 80.06 | 90.68 |
| “a blue tarp” | 36.83 | 49.02 | 67.51 |
| “a person sitting” | 17.66 | 22.47 | 61.77 |
| “a person wearing a helmet” | 18.52 | 24.00 | 61.35 |
| “a person riding a bike” | 10.27 | 12.44 | 45.27 |

表 2 から、リクエストごとに、正例データをファインチューニングに用いると、“a safety cone”を除いて、精度が高くなったことが分かる。また、全てのリクエストに対して、正例データに加えて、リクエストの物体が写っていない画像を負例データとして学習に用いると、正例データのみを学習に用いる場合よりも精度がよくなった。リクエストごとに特化したモデルの構築ができたと考えられる。一般モデルが誤検出を起こしやすい画像を負例として学習させることで、より精度が上がったと考えられる。

## 6. 結論

本稿では、言語モデルを用いた画像認識技術を、自由形式のテキストをリクエストとして、震災アーカイブに対する視覚探索に利用した。検出モデル（GLIP）を用いる場合、事前学習済み一般モデルを用いて獲得したデータをファインチューニングに用いることで、リクエストごとに特化したモデルを構築することができた。これにより、さらに学習データ収集の効率を上げることができると考えられる。さらに、高効率なデータ収集とファインチューニングのサイクルを繰り返すことで、よりよい精度を目指したい。今後の課題として、ファインチューニングの効率についての分析についての調査が考えられる。

**付録**:本稿では、実験結果として、神戸大学附属図書館震災文庫から提供された写真を掲載している。
図 2(a)、2(b)、3(a) 撮影:大木本美通　　図 3(b):撮影:神戸大学附属図書館

## 参考文献


1) 神戸大学附属図書館, 震災文庫利用案内, https://lib.kobe-u.ac.jp/media/riyou_annai.pdf（参照 2023-3-27）
2) 住吉英樹, 河合吉彦, 望月貴裕, 佐野雅規, 震災アーカイブスメタデータ補完システムの開発, 映像情報メディア学会誌, 70.6, pp. J133-J141, 2016.
3) Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., Learning Transferable Visual Models From Natural Language Supervision, International conference on machine learning, PMLR, pp. 8748-8763, 2021.
4) Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N., Mdetr-modulated detection for end-to-end multi-modal understanding, Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1780-1790, 2021.
5) Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J. N., Chang, K. W., Gao, J., Grounded language-image pre-training, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10965-10975, 2022.



**筆者**:1）富谷竜一、工学部情報知能工学科、学生；2）Tristan Hascoet、大学院経営学研究科、助教；3）高島遼一、都市安全研究センター、准教授；4）滝口哲也、都市安全研究センター、教授

# Visual search using vision-language models for an earthquake disaster archive

Ryuichi Tomiya
Tristan Hascoet
Ryoichi Takashima
Tetsuya Takiguchi

Abstract

Kobe University Library disclose the Earthquake Disaster Materials Collection, which preserves a lot of images and videos related to the Great Hanshin-Awaji Earthquake. In order to understand the damage and awareness of disaster prevention at the time, and to prepare for the coming earthquake, it is considered important to conduct a visual search to find specific objects in the images and videos in such a large archive, set up with any free-form text queries that people would naturally express.

Recently, by adding natural language understanding to computer vision, we can input any free-form text. In addition, those models have been trained on large amounts of data and pre-trained models have been released.

In this paper, six text queries such as “a backpack”, “a safety cone”, “a blue tarp”, “a person sitting”, “a person wearing a helmet" and “a person riding a bike” are input into the vision-language pre-trained models. We evaluated the ability of the pre-trained model to be used for visual search in the earthquake archive. We also proposed a highly efficient data collection method for training archive specific models, and conducted fine-tuning of the models with the collected data.