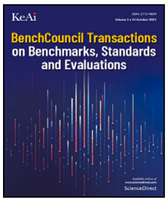
BenchCouncil Transactions on Benchmarks, Standards and Evaluations 4 (2024) 100199

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-on-benchmarks-standards-and-evaluations/



Full length article

Evaluating long-term usage patterns of open source datasets: A citation network approach

Jiaheng Peng, Fanyu Han, Wei Wang *

School of Data Science and Engineering, East China Normal University, Shanghai, 200062, China
Engineering Research Center of Big Data Management, Shanghai, China
Engineering Research Center of Blockchain Data Management (East China Normal University), Ministry of Education, China

ARTICLE INFO

Keywords:
Dataset evaluation
Open source datasets
Usage Pattern Analysis
Citation network

ABSTRACT

The evaluation of datasets serves as a fundamental basis for tasks in evaluatology. Evaluating the usage patterns of datasets has a significant impact on the selection of appropriate datasets. Many renowned Open Source datasets are well-established and have not been updated for many years, yet they continue to be widely used by a large number of researchers. Due to this characteristic, conventional Open Source metrics (e.g., number of stars, issues, and activity) are insufficient for evaluating the long-term usage patterns based on log activity data from their GitHub repositories.

Researchers often encounter significant challenges in selecting appropriate datasets due to the lack of insight into how these datasets are being utilized. To address this challenge, this paper proposes establishing a connection between Open Source datasets and the citation networks of their corresponding academic papers. By mining the citation network of the corresponding academic paper, we can obtain rich graph-structured information, such as citation times, authors, and more. Utilizing this information, we can evaluate the long-term usage patterns of the associated Open Source dataset.

Furthermore, this paper conducts extensive experiments based on five major dataset categories (Texts, Images, Videos, Audio, Medical) to demonstrate that the proposed method effectively evaluates the long-term usage patterns of Open Source datasets. Additionally, the insights gained from the experimental results can serve as a valuable reference for future researchers in selecting appropriate datasets for their work.

1. Introduction

The evaluation of datasets is a cornerstone in various domains of research, forming a critical foundation for advancing the field of evaluatology [1]. High-quality datasets serve as essential building blocks for designing experiments, validating models, and deriving insights across disciplines [2]. As the volume of data and the diversity of datasets grow exponentially, the ability to evaluate and select appropriate datasets has become a vital skill for researchers [3]. Central to this process is the understanding of dataset usage patterns, which offer insights into their practical utility, relevance, and long-term significance [4]. However, this understanding is often obscured by the limitations of conventional evaluation metrics, particularly in the context of Open Source datasets.

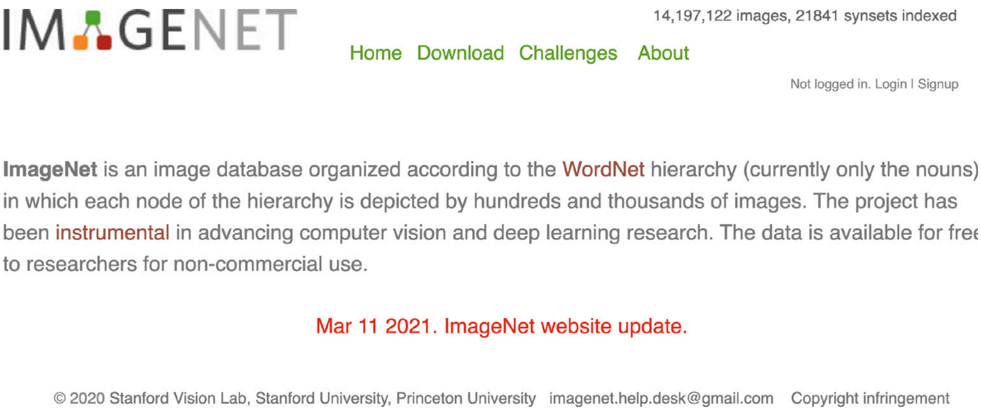
Open Source datasets have gained widespread attention for their accessibility, collaborative development, and impact on the research ecosystem. Notably, many renowned Open Source datasets maintain their prominence and continued usage over extended periods, even without frequent updates or maintenance. For example, Fig. 1 displays the official website of the well-known dataset ImageNet in the image processing domain. As shown in the figure, the website provides only limited information, such as a brief introduction to the dataset and download links. However, it does not offer any insights into the dataset's recent usage or updates. In contrast, Fig. 2 presents the corresponding GitHub repository for the ImageNet dataset. From this figure, it is evident that the repository has not been updated for over a year, suggesting that no significant activity has occurred during this period. This lack of recent logs or updates poses a challenge for us in understanding the dataset's current usage trends.

When researchers select datasets for data science tasks, their choices are often driven by personal subjective preferences, such as opting for well-known datasets they are familiar with. However, they lack factual evidence derived from behavioral log data to understand the recent and long-term usage patterns of these datasets.

Common data insight metrics are derived from the activity log data of GitHub repositories (e.g., stars, issues, forks, and activity levels on

* Correspondence to: School of Data Science and Engineering, East China Normal University, Shanghai, 200062, China.
E-mail address: wwang@dase.ecnu.edu.cn (W. Wang).

https://doi.org/10.1016/j.tbench.2025.100199
Received 6 January 2025; Received in revised form 14 March 2025; Accepted 19 March 2025
Available online 28 March 2025



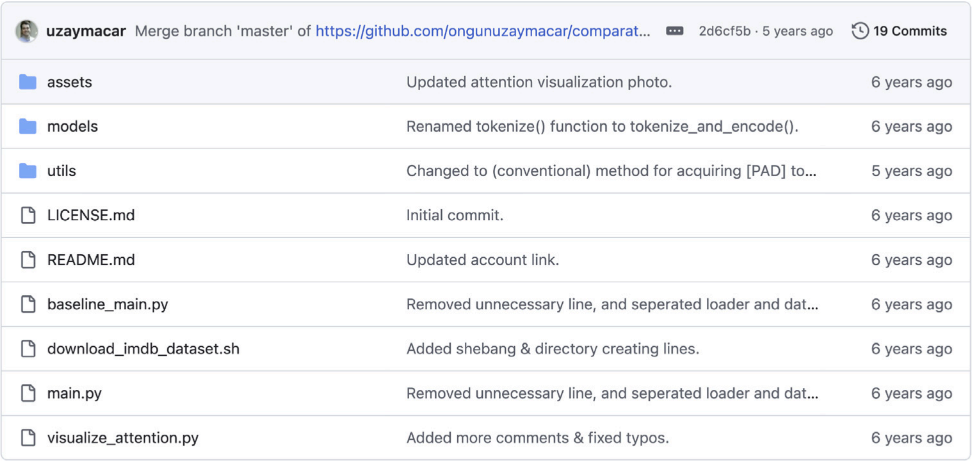
14,197,122 images, 21841 synsets indexed

ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns) in which each node of the hierarchy is depicted by hundreds and thousands of images. The project has been instrumental in advancing computer vision and deep learning research. The data is available for free to researchers for non-commercial use.

Mar 11 2021. ImageNet website update.

© 2020 Stanford Vision Lab, Stanford University, Princeton University   imagenet.help.desk@gmail.com   Copyright infringement

Fig. 1. ImageNet dataset official website.



uzaymacar Merge branch 'master' of https://github.com/ongunuzaymacar/comparat... 2d6cf5b · 5 years ago 19 Commits

assets	Updated attention visualization photo.	6 years ago
models	Renamed tokenize() function to tokenize_and_encode().	6 years ago
utils	Changed to (conventional) method for acquiring [PAD] to...	5 years ago
LICENSE.md	Initial commit.	6 years ago
README.md	Updated account link.	6 years ago
baseline_main.py	Removed unnecessary line, and seperated loader and dat...	6 years ago
download_imdb_dataset.sh	Added shebang & directory creating lines.	6 years ago
main.py	Removed unnecessary line, and seperated loader and dat...	6 years ago
visualize_attention.py	Added more comments & fixed typos.	6 years ago

Fig. 2. Github repository of the IMDB dataset.

GitHub repositories), which are used to measure the long-term popularity and developer activity of a repository. However, these metrics are heavily reliant on repository log activity data. In particular, when a repository has minimal log activity but its dataset continues to be widely used, these data insight metrics become ineffective.

For researchers, this gap presents a significant challenge. The lack of a comprehensive understanding of dataset usage patterns often results in inefficient selection processes and suboptimal utilization of resources. Without reliable indicators of long-term relevance and impact, researchers face difficulties in identifying datasets that best align with their specific needs and objectives. This limitation calls for innovative approaches to evaluate datasets that transcend traditional metrics and incorporate a more nuanced understanding of their role in the academic and research ecosystem.

In response to this challenge, this paper proposes a novel method to bridge the gap between Open Source datasets and their corresponding academic influences. We observed that most Open Source datasets are accompanied by a corresponding academic paper authored by the dataset's creators. This allows us to establish a connection between the dataset and the citation network of its associated academic paper.

Specifically, it establishes a connection between Open Source datasets and the citation networks of their associated academic papers. Academic papers often serve as a formal record of the development, application, and impact of datasets, and their citation networks offer a wealth of information. By mining and analyzing the citation networks, we can uncover critical data points such as citation counts, author contributions, collaboration patterns, and the influence of cited works. This approach leverages the inherent richness of graph-structured citation data to evaluate long-term usage patterns, providing a more comprehensive and reliable basis for dataset assessment.

This study conducts extensive experiments across five major categories of datasets — Texts, Images, Videos, Audio, and Medical — to validate the proposed approach. The experimental results demonstrate the effectiveness of utilizing citation network analysis for understanding the long-term usage and relevance of datasets. Insights derived from this evaluation not only contribute to the broader field of Open Source dataset assessment but also offer practical value to researchers. By enabling more informed decision-making in dataset selection, this work aims to improve the overall efficiency and impact of research efforts.

The contributions of this study are as follows:

- We propose an innovative approach that connects the GitHub repositories of Open Source datasets with the citation networks of their corresponding academic papers. Beyond addressing the direct challenges in existing dataset evaluation methods, this dual perspective enriches our understanding of the Open Source ecosystem. Furthermore, it provides a holistic framework for assessing datasets in a rapidly evolving research landscape, offering valuable insights into both their practical usage and academic influence over time.
- We not only analyze the usage patterns of Open Source datasets from a temporal perspective by examining citation timelines, but also explore potential collaboration patterns within the corresponding GitHub repositories by constructing various collaboration networks. These networks provide valuable insights into

the underlying reasons for the repository's development and sustained influence, shedding light on the factors driving its continued growth and relevance in the Open Source ecosystem. Open Source ecosystems and provides a holistic framework for evaluating datasets in a rapidly evolving research landscape.
- The findings presented in this work aspire to serve as a guide for researchers, dataset curators, and policymakers, fostering a deeper appreciation of the long-term value of Open Source datasets and their critical role in advancing scientific discovery.

2. Related works

The evaluation of Open Source datasets has attracted considerable attention in both academic and industrial domains, primarily due to the growing reliance on datasets for various tasks, including machine learning, data analytics, and scientific research. Existing studies on dataset evaluation can be broadly categorized into two areas: (1) methods and metrics for assessing Open Source projects and (2) citation network analysis for understanding academic influence and impact.

2.1. Open source project evaluation and dataset evaluation metrics

Metrics for evaluating Open Source projects often focus on repository-level statistics such as the number of stars, forks, issues, pull requests, and contributors. These metrics serve as proxies for popularity, community engagement, and activity levels. For example, there are tools and frameworks designed to provide insights into Open Source data, such as Open Source data insight integration plugins [5], mining collaborative patterns in Open Source communities [6], analyzing the geographical distribution of Open Source developers [7], and deriving insights from student performance in Open Source education programs [8], among others. However, when the target of analysis involves underlying collaboration networks, these tools and methods prove to be insufficient.

To address these limitations, researchers have explored more comprehensive graph-based frameworks for evaluating collaborative behaviors, such as Open Source maturity models and quality assurance metrics. For instance, influence assessment models based on contribution metrics, such as the OpenRank model [9], and collaboration pattern mining methods using OpenRank have been proposed [10]. While these models offer more effective ways to evaluate Open Source software, their applicability to dataset evaluation remains limited. This is primarily because these models lack sufficient data to capture the usage patterns and long-term relevance of datasets.

Several studies have proposed dataset-specific evaluation metrics, focusing on attributes like dataset size, diversity, annotation quality, and application domains. For instance, Schmidt et al. [11] highlighted the importance of dataset representativeness and its impact on model generalization. Similarly, Lalor et al. [12] proposed metrics for assessing the fairness and bias in datasets. While these approaches provide valuable insights into dataset quality, they do not address the longitudinal aspect of dataset usage in the research community.

2.2. Connecting datasets with citation networks: Research gaps and contributions

Citation network analysis has emerged as a powerful tool for understanding the academic influence of papers and their associated datasets. Researchers such as McLaren and Bruner [13] and Van Eck and Waltman [14] have demonstrated the potential of citation networks in identifying influential works, mapping collaboration patterns, and studying knowledge dissemination. These studies highlight the richness of citation data, which includes not only citation counts but also relationships between authors, institutions, and research domains.

Recent works have also explored the application of graph-based methods to analyze citation networks [15]. For example, Cummings and Nassar [16] utilized graph neural networks (GNNs) to predict the impact of scientific papers based on their position in the citation graph. Similarly, Liu et al. [17] and He et al. [18] studied the temporal evolution of citation networks to identify emerging research trends. These approaches underscore the value of leveraging graph-structured data to gain deeper insights into academic influence and usage patterns.

While research on Open Source project evaluation and citation network analysis has been extensive, there is a noticeable gap in connecting Open Source datasets with the citation networks of their corresponding academic papers. To date, no systematic efforts have been made to bridge this connection. Building on the insights from these related works, this paper addresses the gap by proposing a novel approach that combines Open Source dataset evaluation with citation network analysis. By leveraging the rich, graph-structured information in citation networks, this method provides a more comprehensive evaluation of long-term usage patterns. Unlike traditional metrics, it accounts for the enduring influence of datasets, offering valuable insights for researchers and dataset curators alike.

3. Methodology

In this section, we present the methodological framework employed to establish a connection between Open Source datasets and the citation networks of their corresponding academic papers. The goal of this methodology is to analyze the long-term usage patterns of datasets based on the citation activities of the academic papers associated with those datasets. The overall framework is depicted in Fig. 3.

3.1. Paper corresponding to the dataset

The first stage of the framework involves identifying the academic papers that correspond to the selected Open Source datasets. To achieve this, we leverage the paperswithcode platform. This is a widely used platform for collecting and organizing datasets along with their corresponding academic papers. We utilized this platform to obtain relevant data on Open Source datasets. The process can be divided into the following steps:

- Top-5 Selection: Using the API provided by the paperswithcode platform,[1] we retrieved the top five most popular dataset modality categories: text, image, video, audio, and medical. These five distinct modalities were selected to ensure comprehensive coverage across various data types and to capture diverse usage patterns in different research domains.
- Categorization: From each of these five modality categories, we selected representative datasets of small, medium, and large scales to ensure a balanced evaluation across different dataset sizes.
- Dataset Name Extraction: After categorization, we extract the names of the corresponding datasets. These dataset names are used as input for the next stage, which involves searching for the associated academic papers.

3.2. Citation network mining

The second stage of the framework focuses on mining the citation networks underlying their corresponding academic publications. This process is critical for evaluating the long-term impact and usage of the datasets. The following steps outline this process:

Searching via Semantic Scholar: We utilized the Semantic Scholar API[2]—an extensive academic search engine—to obtain the unique IDs of the corresponding papers by searching for the titles of the academic

[1] https://paperswithcode.com/

[2] https://api.semanticscholar.org/api-docs/graph

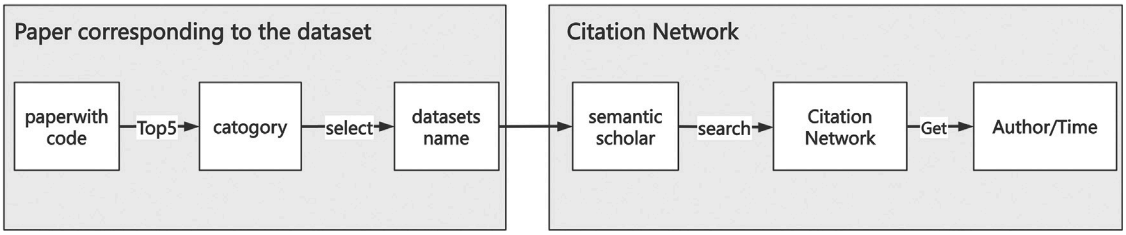


Fig. 3. Framework.

papers associated with each dataset. Using these IDs, we were able to mine the underlying citation networks and the networks of cited papers through the API. Subsequently, we retrieved all papers that cited the target papers and extracted relevant information from these citing papers.

Information Extraction: From the citation network, we extract key information, including:

- Title: The title of the cited papers.
- Authors: The authors of the cited papers and their affiliations.
- Publication Time: The publication date of the cited papers.

Below is a portion of the key code for citation network information mining using the Semantic Scholar API:

Algorithm 1 The method of citation network information mining using the Semantic Scholar API.

```
Input: paper title: $title$, optional fields: $fields$, paper ID: $paper\_id$, output file name: $output\_file$.
Output: Citation data CSV file.
$base\_url \leftarrow$ "https://api.semanticscholar.org/graph/v1/paper/search/match"
$params \leftarrow$ {"query": $title$, "fields": $fields$}
$response \leftarrow$ requests.get($base\_url, params = params, timeout = 10$)
$paper\_data \leftarrow response.json()$
$paper\_id \leftarrow paper\_data.get$("paperId")
$base\_url \leftarrow$ "https://api.semanticscholar.org/graph/v1/paper/" + $paper\_id$ + "/citations"
$citations \leftarrow []$
$offset \leftarrow 0$
while True do
  $params \leftarrow$ {"offset": $offset$, "limit": 1000, "fields": $fields$}
  $response \leftarrow$ requests.get($base\_url, params = params, timeout = 10$)
  $data \leftarrow response.json()$
  $citations.extend(data.get$("data", []))
  if $data.get$("next") then
    $offset \leftarrow data.get$("next")
  else
    break
  end if
end while
with open($output\_file$, "w", newline="", encoding="utf-8") as $file$:
  $writer \leftarrow$ csv.writer($file$)
  $writer.writerow$(["Paper Title", "Authors", "Publication Year"])
```

3.3. Evaluating long-term usage patterns

Through citation network information mining, we can leverage the obtained data to assess the long-term usage patterns of Open Source datasets.

By combining the dataset information with the citation network data, we can evaluate the long-term usage patterns of the selected Open Source datasets. The citation network provides a graph-structured representation of how the dataset's corresponding paper has influenced subsequent research over time. This approach addresses the limitations of conventional Open Source metrics by focusing on citation trends rather than repository activity alone.

Key Insights from the Framework:

- **Cumulative Citation Trend**: The total number of citations accumulated by a paper since its publication, calculated on an annual basis. This metric provides a historical perspective on the impact of the dataset-associated academic papers.
- **Annual Citation Growth Trend**: Refers to the number of new citations a paper receives each year since its publication.
- **Growth Rate Trend**: Also known as the growth speed, it represents the ratio of the increase in a data indicator to the base period data over a certain period, expressed as a percentage. This can be formulated as: $Y = \frac{X_t - X_{t-1}}{X_{t-1}} \times 100\%$ where Y denotes the growth rate, X_t and X_{t-1} represent the total number of citations in year t and t_{t-1},respectively.
- **Three Types of Collaborative Network Analysis**: Project Contribution Network analysis, Project Ecosystem Network analysis and Project Community Network analysis, all constructed via the Open Source project osgraph.[3]
  Project Contribution Network analysis: Find core project contributors based on developer activity information (Issues, PRs, Commits, CRs, etc.).
  Project Ecosystem Network analysis: Extract relationships between projects' development activities and organizations to build core project ecosystem relationships.
  Project Community Network analysis: Extract core developer community distribution based on project development activities and developer organization information.

In addition to analyzing academic papers associated with datasets that have been published for a considerable duration, the analyses of the **Annual Citation Growth Trend** and **Growth Rate Trend** also enable a clearer identification of datasets with substantial growth potential. This is particularly crucial for relatively new datasets that have been published for only one to three years, as their cumulative citation counts are typically lower. Citation networks not only provide information on the quantity and timing of dataset citations but also reveal the collaborative network structures formed around the datasets within the academic and industrial communities.

4. Experiment

4.1. Setup and datasets

For our study, we selected five distinct data modalities. Within each modality type, we established three different dataset scales. From each scale within every modality, we randomly selected one dataset to serve as the representative for that particular category and scale. Specifically, we included datasets of varying scales within each modality type — small, medium, and large. A small-scale dataset is defined as one

[3] https://github.com/TuGraph-family/OSGraph

Table 1
The specific names and categories of the selected datasets.

Category	Small-dataset	Medium-dataset	Large-dataset
Images	CityFlow (350)	Food-101 (2003)	Fashion-MNIST (7949)
Texts	FinQA (213)	CommonsenseQA (1349)	GLUE (6334)
Videos	MSVD (115)	OTB (2898)	UCF101 (5629)
Audio	XD-Violence (245)	Common Voice (1319)	Librispeech (5752)
Medical	VerSe (203)	ChestX-ray14 (2157)	MIMIC-III (6449)

Table 2
The selected dataset (with the total citation count of its corresponding academic paper).

Category	Small-dataset	Large-dataset
Images	JFT-3B (961) CityFlow (350) WildDeepfake (330)	CelebA (7959) Fashion-MNIST (7949) SVHN (6571)
Texts	CLINC150 (489) COCO (486) FinQA (213)	SST (8113) SQuAD (7686) GLUE (6334)

whose corresponding academic paper has fewer than 500 citations, a medium-scale dataset is defined as one with 500 to 5,000 citations of its corresponding paper, and a large-scale dataset is defined as one whose corresponding paper has been cited more than 5,000 times.

The specific dataset names corresponding to the five selected categories are presented in Table 1.

In addition, to conduct more comprehensive experiments and analyses, we randomly selected three datasets from each of the two domains (image and text), covering both large-scale and small-scale categories. The selected dataset names, along with their corresponding citation counts from the literature, are presented in Table 2.

4.2. Academic papers corresponding to the dataset

Table 1 and Table 2 lists the abbreviated names of each dataset. Below is a detailed description of their corresponding academic paper titles:

Fashion-MNIST [19]: A Novel Image Dataset for Benchmarking Machine Learning Algorithms

Food-101 [20]: Mining Discriminative Components with Random Forests

CityFlow [21]: A City-Scale Benchmark for Multi-Target Multi-Camera Vehicle Tracking and Re-Identification

GLUE [22]: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding

CommonsenseQA [23]: A Question Answering Challenge Targeting Commonsense Knowledge

FinQA [24]: A Dataset of Numerical Reasoning over Financial Data

UCF101 [25]: A Dataset of 101 Human Actions Classes From Videos in The Wild

OTB [26]: Object Tracking Benchmark

MSVD [27]: Collecting Highly Parallel Data for Paraphrase Evaluation

Librispeech [28]: An ASR corpus based on public domain audio books

Common Voice [29]: A Massively-Multilingual Speech Corpus

XD-Violence [30]: Not only Look, but also Listen: Learning Multimodal Violence Detection under Weak Supervision

MIMIC-III [31]: MIMIC-III, a freely accessible critical care database

ChestX-ray14 [32]: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases

VerSe [33]: A Vertebrae Labeling and Segmentation Benchmark for Multi-detector CT Images

JFT-3B [34]: Scaling Vision Transformers

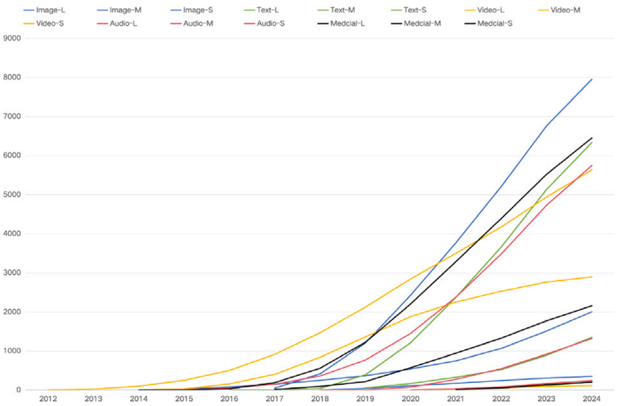


Fig. 4. The number of citations about datasets.

WildDeepfake [35]: A Challenging Real-World Dataset for Deepfake Detection.

CelebA [36]: Deep Learning Face Attributes in the Wild

SVHN [37]: Reading Digits in Natural Images with Unsupervised Feature Learning

SST [38]: Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank

CLINC150 [39]: An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction

COCO-Text [40]: COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images

4.3. The development status of datasets across different categories

Fig. 4 depicts the temporal evolution of cumulative citations for papers associated with datasets across various domains. From the perspective of cumulative citations, Image-L also stands out as the most prominent category. Its total citations have grown exponentially since 2017, far surpassing other categories by 2024. The datasets in the image domain have implicitly demonstrated their status as the most popular and highly scrutinized research area within the broader landscape of deep learning. Similarly, the cumulative citations of Text-L and Medical-L have also risen rapidly, particularly Text-L, whose growth trajectory has almost paralleled that of Image-L since 2020. This indicates that, in addition to the image domain, datasets in the text domain are also one of the focal points of researchers' attention.

In contrast, datasets in the video and audio domains (including large, medium, and small datasets) have seen slower growth in cumulative citations. Although Video-L and Audio-L have shown year-over-year increases in total citations, they still lag significantly behind the image and text domains. This may be due to the higher complexity of data processing and the more specialized application scenarios in these fields.

Overall, the trend in cumulative citations aligns with the trend in annual citation growth, where large-scale datasets — particularly in the image and text domains — continue to dominate, while medium and small-scale datasets, as well as those in the audio and video domains, have relatively lower influence and slower growth rates.

The annual citation growth trend is illustrated in Fig. 5. The annual growth trend in dataset citations clearly demonstrates the dominance of large-scale datasets. In particular, Image-L has seen a rapid increase in citations since 2016, peaking in 2022, followed by a slight decline in 2023 and 2024, while still maintaining the highest number of citations. This suggests that large-scale image datasets continue to attract significant attention from researchers and developers, despite the slowing growth in recent years.

Since 2017, the citation counts of most datasets have exhibited significant growth, particularly for Image-M and Text-M. This surge

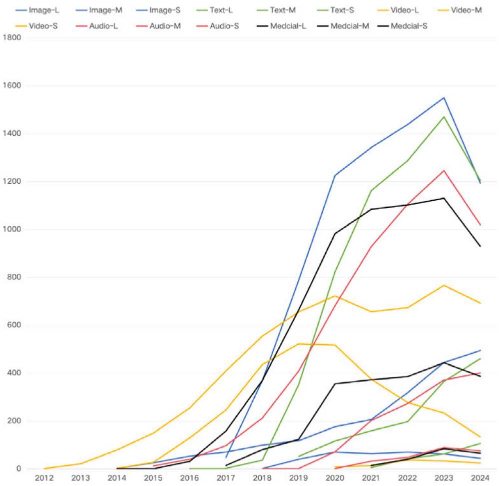


Fig. 5. The number of citations about datasets each year.

is likely attributable to the rapid development and widespread application of deep learning technologies during this period. Given that deep learning tasks in the image and text domains are among the most popular, the citation counts for datasets in these two fields have increased markedly.

Between 2020 and 2022, the growth rate peaked for most datasets, with the annual increase reaching its zenith in 2022. This peak may be associated with the heightened demand for datasets during the pandemic, as many studies shifted towards remote data collection and analysis. The increased reliance on datasets during this period likely contributed to the surge in citation counts.

Compared to large-scale datasets, medium-scale datasets exhibited less pronounced growth, possibly due to their narrower scope of applicability. The relatively slow development trend of small-scale datasets may be partly attributed to their limited application range and niche task suitability.

4.4. Evolutionary trends: Large- and small-scale datasets in image and text domains

The Fig. 6, Fig. 7 and Fig. 8 illustrate the development trends of image datasets of different scales (large and small) over the years, providing insights into their respective growth trajectories in terms of cumulative citations and annual citations.

4.4.1. Cumulative citation trends

In Fig. 6, the blue and red bars in the figure represent large-scale datasets, while the green and orange bars correspond to small-scale datasets. large-scale datasets (Image-L and Text-L) exhibit a significantly steeper growth curve compared to small-scale (Image-S and Text-S) dataset. The Fig. 6 illustrates the trends in cumulative citations for large-scale and small-scale datasets within the domains of image and text.

Around 2017, the number of citations for large-scale datasets began to increase rapidly. We posit that this surge is likely associated with the burgeoning development of deep learning. During this period, a significant number of researchers initiated work related to deep learning, which in turn led to a substantial increase in the citation counts of corresponding papers.

Beginning in 2017, the citation gap between large-scale datasets and small-scale datasets has progressively widened. By 2024, the highest citation count among the selected small-scale datasets was approximately 1000, with others exhibiting even lower citation frequencies. This trend suggests that small-scale datasets have not demonstrated robust capabilities in disseminating academic influence.

Through systematic observation and analysis, we have identified that this phenomenon can be primarily attributed to the fact that a substantial proportion of papers associated with large-scale datasets are published in top-tier conferences, particularly in premier venues such as CVPR, ICCV within the computer vision domains. These prestigious conferences, recognized as CCF-A class or Core Conference Ranking A* category, possess significant academic influence and visibility, thereby attracting greater attention from the research community and consequently generating higher citation rates.

4.4.2. Annual citation growth trends

As shown in Fig. 7, the blue and red bars in the figure represent large-scale datasets, while the green and orange bars correspond to small-scale datasets. large-scale datasets (Image-L and Text-L) exhibit a significantly steeper growth curve compared to small-scale (Image-S and Text-S) datasets. This exponential growth in cumulative citations for Image-L began around 2017. The analysis indicates that by 2020, large-scale datasets demonstrated an annual citation growth of about 1000 citations, far outpacing the growth observed in small-scale datasets. This substantial absolute increase underscores the continuing prominence and research value of large-scale image datasets, primarily due to their fundamental contributions to multiple deep learning sub-fields such as image recognition, object detection, and text classification.

Statistical evidence indicates that large-scale datasets often attain remarkable research impact, as reflected in their citation metrics, within the first three years after publication. On the contrary, small-scale datasets exhibited a markedly slower trajectory in citation growth. Typically, even after several years of availability, their cumulative citation counts remained within the range of a few hundred citations.

Consistent with the findings in Section 4.4.1, we observe that papers associated with small-scale datasets are often not published in the most prestigious academic conferences or journals. Additionally, the deep learning tasks corresponding to these datasets tend to be relatively niche, with fewer researchers engaged in related work. Consequently, the growth in citation counts for these datasets is relatively slow.

4.4.3. Growth rate trends

In addition, we conducted experiments on the annual growth rate trends of the papers corresponding to these datasets, as shown in Fig. 8. Similar to Section 4.4.1 and Section 4.4.2, the blue and red bars in the figure represent large-scale datasets, while the green and orange bars correspond to small-scale datasets. The large-scale datasets (Image-L and Text-L) exhibit a significantly steeper growth curve compared to the small-scale datasets (Image-S and Text-S).

Given that the citation growth rate of papers typically experiences an explosive increase shortly after publication — reaching up to 3700% in some cases — we truncated growth rates exceeding 200% to ensure clarity in the trend visualization. This truncation represents the "explosive growth" phase, primarily to focus on the citation growth patterns after the initial surge in popularity. This approach allows us to observe the citation dynamics once the initial fervor surrounding the publication has subsided.

As can be observed from the figure, large-scale datasets typically experience a prolonged period of "explosive growth", during which their citation counts increase rapidly. After this initial surge, the citation growth rate tends to decline gradually, yet remains relatively high. In fact, even eight to ten years after publication, the annual citation growth rate for these large-scale datasets can still exceed 20%.

In contrast, small-scale datasets exhibit a much shorter period of growth. They generally attract significant attention only in the first one to three years following publication. Their citation growth rates decline sharply thereafter, typically falling below 20% within five to seven years.

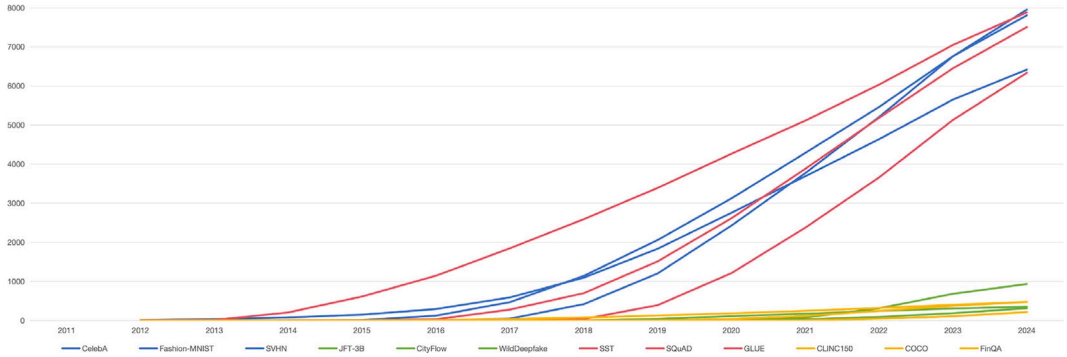


Fig. 6. Cumulative Citation Trends Across Image and Text Datasets.

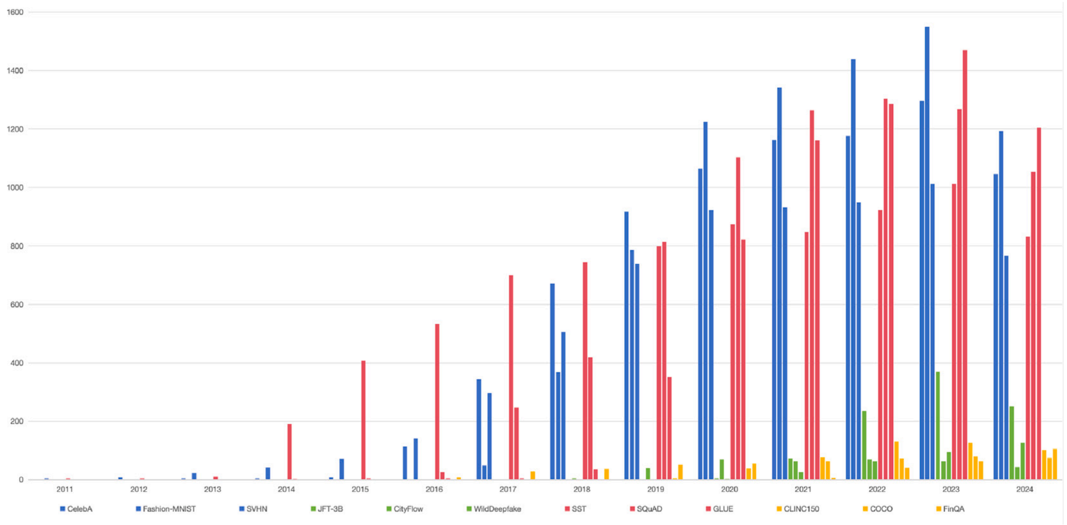


Fig. 7. Cumulative Citation Trends Across Image and Text Datasets Over the Years.

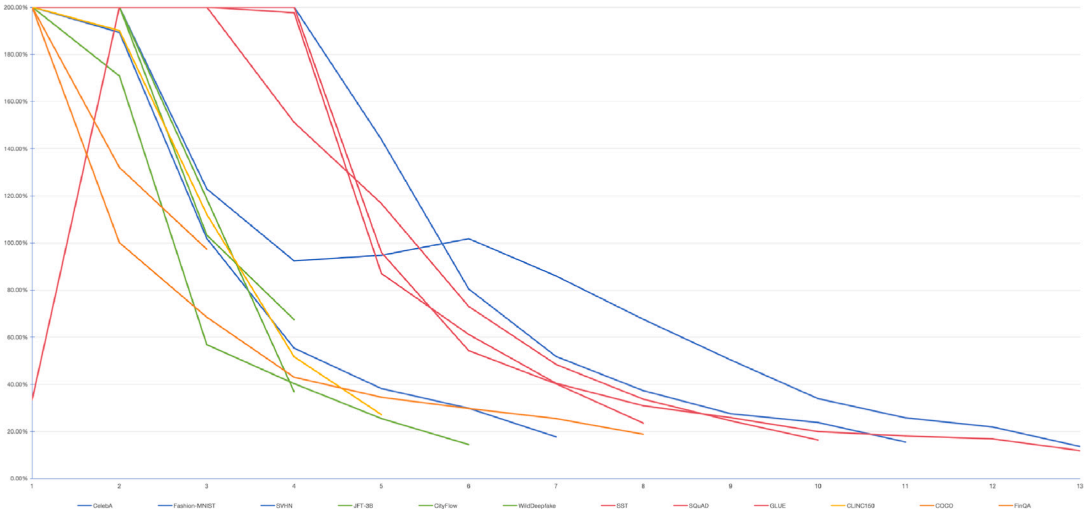


Fig. 8. Cumulative Citation Trends Across Image and Text Datasets.

4.5. Three types of collaborative network analysis

4.5.1. Project contribution network analysis

Taking the GitHub repository with the highest star count corresponding to a large-scale image dataset as an example, we conducted an in-depth analysis of its contribution collaboration network. The results reveal a highly active and diverse contributor network, comprising both individual contributors and automated bots. The visualization illustrates that the dataset has attracted numerous influential contributors, such as MarkDaoust, nealwu, and cshtjn, who have made significant contributions to the project through code reviews (CR), pull requests (PR), and issue discussions. These core contributors demonstrate the sustained interest and involvement of experienced developers in the ongoing maintenance and improvement of the dataset. (see Fig. 9).

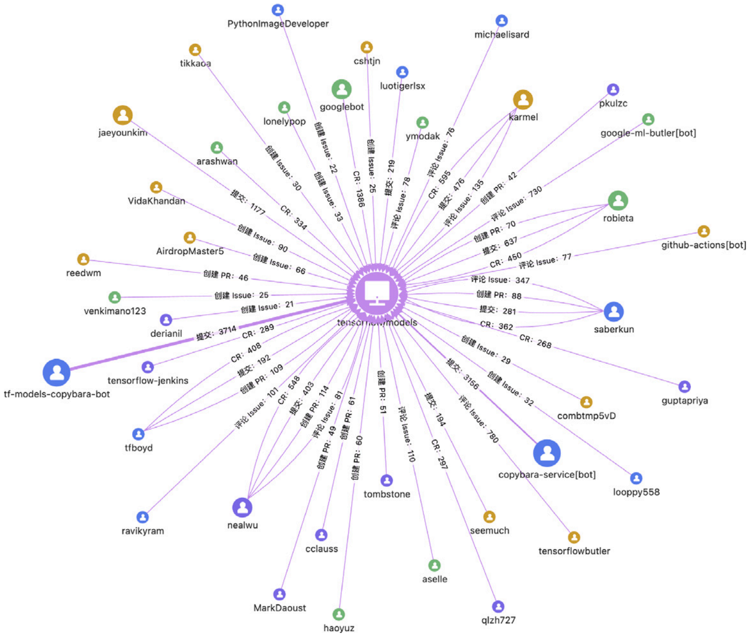


Fig. 9. Project Contribution Network.

In addition to individual contributors, the network also highlights the role of automated bots, including googlebot and tensorflowbutler, which represent well-known automated tools from major companies such as Google. These bots play a crucial role in maintaining the repository's automated workflows, indicating the importance of continuous integration/continuous deployment (CI/CD) processes in the project's lifecycle. The presence of such automation tools suggests that the repository maintains high standards of quality control, ensuring that updates and contributions are systematically reviewed and integrated.

Furthermore, the collaboration network demonstrates the involvement of a wide range of contributors across different organizational backgrounds, indicating the broad adoption and community-driven nature of the project. The combination of human contributors and automated bots highlights the hybrid nature of modern Open Source collaborations, where manual contributions are complemented by automated processes to ensure efficiency and reliability. This analysis underscores the significance of collaborative networks in maintaining large-scale Open Source datasets and the critical role of automation in facilitating seamless collaboration across distributed teams.

4.5.2. *Project ecosystem network analysis*

The project ecosystem collaboration network of the GitHub repository corresponding to a large-scale image dataset illustrates the extensive collaboration between this repository and other well-known projects in the Open Source community. As shown in the visualization, the repository attracts collaborations with several prominent repositories and organizations, including PyTorch, Microsoft's VS Code, and the Hugging Face community, along with its widely used Transformers library. These collaborations highlight the interconnectedness of major Open Source projects and demonstrate the dataset's influence across various domains of machine learning and software development.(see Fig. 10)

The network also reflects the growing importance of ecosystem-level interactions within Open Source communities. For instance, repositories such as TensorFlow, Keras, and Apache MXNet exhibit strong collaboration links with the dataset's repository, indicating shared contributions, joint development efforts, or the use of the dataset in complementary tools and frameworks. Such ecosystem interactions reinforce the dataset's role as a critical component within the broader machine learning infrastructure.

A particularly noteworthy observation is the emergence of the "rich-get-richer" effect, often referred to as the "rich club" phenomenon in network theory. The more a dataset or repository is cited and referenced within the community, the more likely it is to attract collaborations with other high-profile projects. This positive feedback loop results in widely-used datasets forming core hubs within the Open Source ecosystem, drawing further attention and engagement from influential developers and repositories. This effect underscores the importance of visibility and reputation in Open Source projects, where well-established repositories tend to attract more collaborators and maintain their central position within the ecosystem over time.

4.5.3. *Project community network analysis*

The community collaboration network of the GitHub repository corresponding to a large-scale image dataset with the highest star count demonstrates the extensive and diverse collaborations established with developers, companies, and research institutions across the globe. The network highlights significant contributions from developers and communities in countries such as China, Germany, United Kingdom, United States, and India, indicating the dataset's broad international adoption and its appeal to a wide range of contributors. (see Fig. 11)

Furthermore, the network reveals collaborations with some of the most prominent tech companies in the world, including Google, NVIDIA, and Microsoft, which play a crucial role in the development and promotion of cutting-edge machine learning technologies. The involvement of such well-established organizations suggests that the dataset is not only academically relevant but also practically significant for industry use cases. These collaborations reflect the repository's central position within the global Open Source ecosystem and its influence on both academic research and industrial applications.

This global and multi-organizational collaboration network underscores the growing importance of cross-border and cross-institutional partnerships in Open Source projects. The network demonstrates that widely-used datasets attract contributions from a diverse set of stakeholders, including independent developers, research institutions, and large technology companies. This diversity contributes to the repository's sustainability and long-term relevance by ensuring continuous improvements and the integration of new features driven by both academic and industry needs.

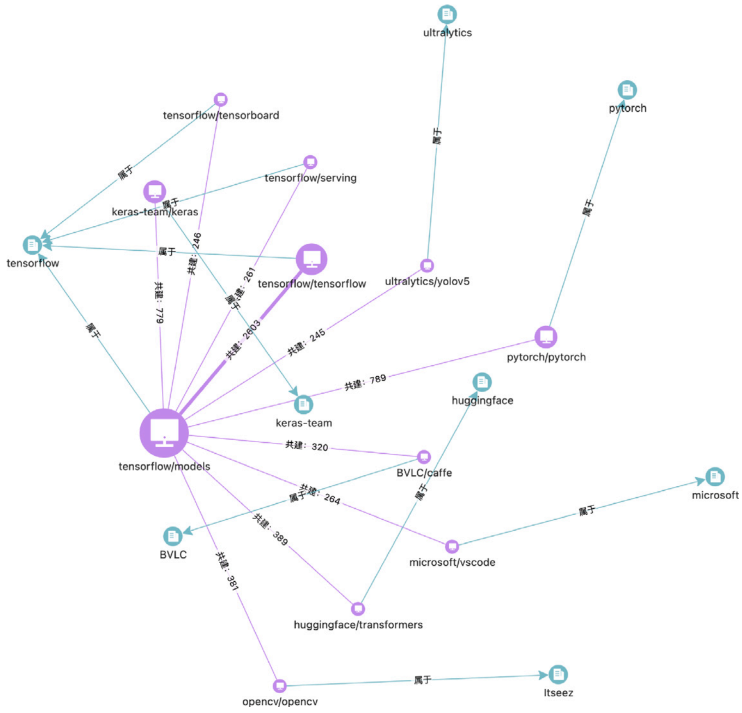


Fig. 10. Project Ecosystem Network.

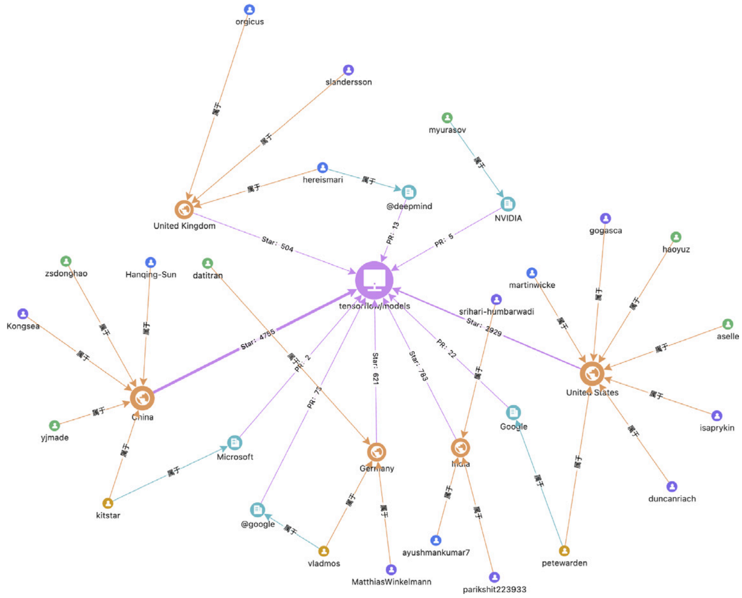


Fig. 11. Project Community Network.

Overall, the analysis of the project's community collaboration network highlights how large-scale datasets serve as focal points for global collaboration in Open Source ecosystems, driving innovation and knowledge sharing across countries and sectors.

5. Conclusion

This study proposes a novel framework for evaluating the long-term usage patterns of Open Source datasets by connecting them with the citation networks of their corresponding academic papers. Traditional data insight metrics, such as star counts and issue counts, become ineffective in the absence of GitHub log data. By mining the academic networks associated with datasets, we can indirectly analyze the long-term usage patterns of these datasets.

Through extensive experiments across five dataset modalities — text, image, video, audio, and medical — the study validates the effectiveness of the proposed method. The analysis of project contribution networks, ecosystem networks, and community networks reveals the collaborative nature of Open Source development and highlights the critical role of automated tools and global partnerships in sustaining large-scale repositories.

Overall, this research bridges the disconnect between Open Source activity metrics and academic citation analysis, laying the groundwork for a more holistic framework for assessing dataset relevance and impact. Nonetheless, several limitations remain in this study. For instance, some Open Source datasets lack corresponding published academic papers, or their associated papers have only recently been published, making citation information unavailable. In such cases, our approach is constrained. Addressing this issue is one of our future research directions. We aim to explore alternative methods to achieve a more comprehensive evaluation of Open Source datasets.

CRediT authorship contribution statement

Jiaheng Peng: Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Fanyu Han:** Writing – review & editing, Investigation. **Wei Wang:** Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation grants 62137001, 62277017, and 61977026.

References

[1] J. Zhan, A short summary of evaluatology: The science and engineering of evaluation, BenchCouncil Trans. Benchmarks Stand. Eval. (2024) 100175.

[2] J. He, S. Yang, S. Yang, A. Kortylewski, X. Yuan, J.N. Chen, S. Liu, C. Yang, Q. Yu, A. Yuille, Partimagenet: A large, high-quality dataset of parts, in: European Conference on Computer Vision, Springer, 2022, pp. 128–145.

[3] N. Meron, V. Blass, G. Thoma, Selection of the most appropriate life cycle inventory dataset: new selection proxy methodology and case study application, Int. J. Life Cycle Assess. 25 (2020) 771–783.

[4] F.A. Silva, A.C. Domingues, T.R.B. Silva, Discovering mobile application usage patterns from a large-scale dataset, ACM Trans. Knowl. Discov. Data (TKDD) 12 (5) (2018) 1–36.

[5] Y. Tang, S. Zhao, X. Xia, F. Bi, W. Wang, HyperCRX: A browser extension for insights into GitHub projects and developers, in: Proceedings of the 32nd IEEE/ACM International Conference on Program Comprehension, 2024, pp. 460–464.

[6] W. Huang, X. Xia, A. Zhou, X. Zhou, W. Wang, S. Zhao, Z. Wang, S. Bian, OSGraph: A data visualization insight platform for open source community, in: International Conference on Database Systems for Advanced Applications, Springer, 2024, pp. 476–479.

[7] L. You, J. Peng, W. Wang, Y. Xia, K. Zhou, Data driven visualized analysis: Visualizing global trends of GitHub developers with fine-grained geo-details, in: International Conference on Database Systems for Advanced Applications, 2024, pp. 498–502.

[8] W. Jie, W. Huang, Z. Shengyu, X. Xiaoya, H. Fanyu, W. Wei, Y. Zhang, OpenRank contribution evaluation method and empirical study in open-source course, J. East China Norm. Univ. (Nat. Sci.) 2024 (5) (2024) 11.

[9] S. Zhao, X. Xia, B. Fitzgerald, X. Li, V. Lenarduzzi, D. Taibi, R. Wang, W. Wang, C. Tian, Motivating open source collaborations through social network evaluation: A gamification practice from Alibaba, 2023.

[10] S. Zhao, X. Xia, B. Fitzgerald, X. Li, V. Lenarduzzi, D. Taibi, R. Wang, W. Wang, C. Tian, OpenRank leaderboard: Motivating open source collaborations through social network evaluation in Alibaba, in: Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Practice, 2024, pp. 346–357.

[11] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, A. Madry, Adversarially robust generalization requires more data, Adv. Neural Inf. Process. Syst. 31 (2018).

[12] J.P. Lalor, A. Abbasi, K. Oketch, Y. Yang, N. Forsgren, Should fairness be a metric or a model? A model-based framework for assessing bias in machine learning pipelines, ACM Trans. Inf. Syst. 42 (4) (2024) 1–41.

[13] C.D. McLaren, M.W. Bruner, Citation network analysis, Int. Rev. Sport. Exerc. Psychol. 15 (1) (2022) 179–198.

[14] N.J. Van Eck, L. Waltman, CitNetExplorer: A new software tool for analyzing and visualizing citation networks, J. Informetr. 8 (4) (2014) 802–823.

[15] L. You, J. Peng, H. Jin, C. Claramunt, H. Zeng, Z. Zhang, DRGAT: Dual-relational graph attention networks for aspect-based sentiment classification, Inform. Sci. 668 (2024) 120531.

[16] D. Cummings, M. Nassar, Structured citation trend prediction using graph neural networks, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2020, pp. 3897–3901.

[17] J. Liu, F. Xia, X. Feng, J. Ren, H. Liu, Deep graph learning for anomalous citation detection, IEEE Trans. Neural Netw. Learn. Syst. 33 (6) (2022) 2543–2557.

[18] G. He, Z. Xue, Z. Jiang, Y. Kang, S. Zhao, W. Lu, H2CGL: Modeling dynamics of citation network for impact prediction, Inf. Process. Manage. 60 (6) (2023) 103512.

[19] H. Xiao, K. Rasul, R. Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017, arXiv preprint arXiv:1708.07747.

[20] L. Bossard, M. Guillaumin, L. Van Gool, Food-101–mining discriminative components with random forests, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13, Springer, 2014, pp. 446–461.

[21] Z. Tang, M. Naphade, M.Y. Liu, X. Yang, S. Birchfield, S. Wang, R. Kumar, D. Anastasiu, J.N. Hwang, Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8797–8806.

[22] A. Wang, Glue: A multi-task benchmark and analysis platform for natural language understanding, 2018, arXiv preprint arXiv:1804.07461.

[23] A. Talmor, J. Herzig, N. Lourie, J. Berant, Commonsenseqa: A question answering challenge targeting commonsense knowledge, 2018, arXiv preprint arXiv:1811.00937.

[24] Z. Chen, W. Chen, C. Smiley, S. Shah, I. Borova, D. Langdon, R. Moussa, M. Beane, T.H. Huang, B.R. Routledge, et al., FinQA: A dataset of numerical reasoning over financial data, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 3697–3711.

[25] K. Soomro, UCF101: A dataset of 101 human actions classes from videos in the wild, 2012, arXiv preprint arXiv:1212.0402.

[26] A. Berg, J. Ahlberg, M. Felsberg, A thermal object tracking benchmark, in: 2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS, IEEE, 2015, pp. 1–6.

[27] D. Chen, W.B. Dolan, Collecting highly parallel data for paraphrase evaluation, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 190–200.

[28] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, Librispeech: an asr corpus based on public domain audio books, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2015, pp. 5206–5210.

[29] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, G. Weber, Common voice: A massively-multilingual speech corpus, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 4218–4222.

[30] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, Z. Yang, Not only look, but also listen: Learning multimodal violence detection under weak supervision, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16, Springer, 2020, pp. 322–339.

[31] A.E. Johnson, T.J. Pollard, L. Shen, L.w.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data 3 (1) (2016) 1–9.

[32] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2097–2106.

[33] A. Sekuboyina, M.E. Husseini, A. Bayat, M. Löffler, H. Liebl, H. Li, G. Tetteh, J. Kukačka, C. Payer, D. Štern, et al., VerSe: a vertebrae labelling and segmentation benchmark for multi-detector CT images, Med. Image Anal. 73 (2021) 102166.

[34] X. Zhai, A. Kolesnikov, N. Houlsby, L. Beyer, Scaling vision transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12104–12113.

[35] B. Zi, M. Chang, J. Chen, X. Ma, Y.G. Jiang, Wilddeepfake: A challenging real-world dataset for deepfake detection, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2382–2390.

[36] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3730–3738.

[37] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A.Y. Ng, et al., Reading digits in natural images with unsupervised feature learning, in: NIPS Workshop on Deep Learning and Unsupervised Feature Learning, vol. 2011, Granada, 2011, p. 4.

[38] R. Socher, A. Perelygin, J. Wu, J. Chuang, C.D. Manning, A.Y. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1631–1642.

[39] S. Larson, A. Mahendran, J.J. Peper, C. Clarke, A. Lee, P. Hill, J.K. Kummerfeld, K. Leach, M.A. Laurenzano, L. Tang, et al., An evaluation dataset for intent classification and out-of-scope prediction, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, 2019, pp. 1311–1316.

[40] A. Veit, T. Matera, L. Neumann, J. Matas, S. Belongie, Coco-text: Dataset and benchmark for text detection and recognition in natural images, 2016, arXiv preprint arXiv:1601.07140.